Exam Questions DA0-001

CompTIA Data+ Certification Exam

https://www.2passeasy.com/dumps/DA0-001/

NEW QUESTION 1
Taylor wants to investigate how manufacturing, marketing, and sales expenditures impact overall profitability for her company.
Which of the following systems is the most appropriate?

A. OLTP.
B. OLAP.
C. Data warehouse.
D. Data mart.

Answer: C

Explanation:

A Data mart is too narrow, because Taylor needs data from across multiple divisions. OLAP is a broad term for analytical processing, and OLTP systems are transactional and not ideal for the task. Since Taylor is working with data across multiple different divisions, she will work with a Data warehouse.

NEW QUESTION 2
A financial institution is reporting on sales performance to a company at the account level. Due to the sensitive nature of the government the does il with, some account information is not shown. Which of the following fields should be masked?

A. Sales volume
B. Start date
C. Product name
D. Customer name

Answer: D

Explanation:
Customer name is the field that should be masked, because it contains sensitive information that could identify the government accounts that the financial institution deals with. Masking is a technique that replaces or obscures sensitive data with dummy or random data, such as asterisks or hashes. Masking can help protect the privacy and security of the data, while still allowing for some analysis and reporting. Therefore, the correct answer is D. References: [Data Masking | Definition, Techniques & Examples - Talend], [Data masking - Wikipedia]

NEW QUESTION 3
A data set was recorded using multimedia technology. Which of the following is a necessary step on the way to interpretation?

A. Structural equation modeling
B. Transcription
C. Sequential analysis
D. Sampling

Answer: B

Explanation:
The correct answer is B. Transcription.
Transcription is a necessary step on the way to interpretation when a data set was recorded using multimedia technology. Multimedia technology refers to the use of various forms of media, such as audio, video, images, and text, to capture and present information1 Transcription is the process of converting multimedia data into written or textual form, which can then be analyzed using various methods and tools2 Transcription can help to make the data more accessible, searchable, and manageable, as well as to preserve the data for future use.
Structural equation modeling is not correct, because it is a statistical technique that tests the causal relationships between multiple variables using observed and latent variables. Structural equation modeling is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data.
Sequential analysis is not correct, because it is a method of analyzing the order and timing of events or behaviors in a data set. Sequential analysis is not a necessary step on the way to interpretation, but rather an optional method that can be applied to certain types of data. Sampling is not correct, because it is the process of selecting a subset of data from a larger population for analysis. Sampling is not a necessary step on the way to interpretation, but rather a preliminary step that can be done before collecting or analyzing the data.

NEW QUESTION 4
Which of the following concepts should be applied if a data set with 40 fields needs to be pared down to 20 fields and contains similar data across multiple fields?

A. Duplication
B. Consolidation
C. Compliance
D. Standardization

Answer: B

Explanation:
Consolidation is the process of combining multiple elements into a single, more effective or coherent whole. In the context of data analytics, consolidation would involve merging similar fields to reduce the overall number of fields in a dataset. This is particularly useful when a dataset contains redundant or similar data across multiple fields, as it helps to simplify the data structure and improve efficiency. Techniques such as dimensionality reduction are often applied to achieve this, where the goal is to retain the most informative and representative features of the data while reducing the number of total features. References:
? Applied Dimensionality Reduction — 3 Techniques using Python1.
? Seven Techniques for Data Dimensionality Reduction2.
? Best practices when working with datasets3.
? Effectively Handling Large Datasets4.

NEW QUESTION 5
You are working with a professional statistician to perform an analysis and would like to use a statistics package.
Which one of the following would be the most appropriate?

A. Rapid Miner.
B. QLIK.
C. Power BI.
D. Minitab.

Answer: D

Explanation:

Minitab is statistical analysis software. It can be used for learning about statistics as well as statistical research. Statistical analysis computer applications have the advantage of being accurate, reliable, and generally faster than computing statistics and drawing graphs by hand.

NEW QUESTION 6
An analysts building a monthly report for production and wants to ensure the audience is aware of its once-a-month cadence. Which of the following is the MOST important to convey that information?

A. The date of the dashboard build
B. The data refresh date
C. A report summary
D. Frequently asked questions

Answer: A

Explanation:
This is because the date of the dashboard build is the most important component to convey that information, which is the once-a-month cadence of the monthly report for production. The date of the dashboard build can convey that information by indicating when the dashboard was created or updated, as well as showing the frequency or interval of the dashboard creation or update. For example, the date of the dashboard build can convey that information by displaying a date format that includes the month and year, such as January 2020, February 2020, etc., or by displaying a text format that includes the word ??monthly??, such as Monthly Report for Production - January 2020, Monthly Report for Production - February 2020, etc. The other components are not the most important components to convey that information. Here is why:
? The data refresh date is a component that indicates when the data on the dashboard was refreshed or retrieved from the source or system, such as a database, a cloud service, or a web application. The data refresh date does not convey that information, but rather conveys how current or up-to-date the data on the dashboard is.
? A report summary is a component that provides an overview or a highlight of the main findings or insights from the dashboard, such as key metrics, indicators, or trends. A report summary does not convey that information, but rather conveys what the dashboard is about or what it shows.
? Frequently asked questions is a component that provides answers or explanations to common or expected questions from the audience or users of the dashboard, such as how to use or interpret the dashboard, what are the assumptions or limitations of the dashboard, etc. Frequently asked questions does not convey that information, but rather conveys how to understand or interact with the dashboard.

NEW QUESTION 7
A sales team wants visibility of current sales numbers, pipeline, and team performance. The team would also like to see calculations of individuals?? earned commissions and projected commissions based on sales, but they want that information to be kept confidential. Which of the following would be the BEST way to provide this visibility?

A. Create a dashboard displaying a data refresh date so users know the current sales numbers and configure permissions to control access.
B. Create a dashboard for sales numbers, pipeline, and team and individual performance for the management team.
C. Create a dashboard with filters for the overall team, individuals, and managemen
D. Users can filter to see the data they want.
E. Create a dashboard with views for team, individuals, and managemen
F. Configure permissions to control access.

Answer: D

Explanation:
Create a dashboard with views for team, individuals, and management. Configure permissions to control access. This is because a dashboard is a type of visualization that displays multiple charts or graphs on a single page, usually to provide an overview or summary of some data or information. A dashboard can be used to provide visibility of current sales numbers, pipeline, and team performance by showing different metrics and indicators related to these aspects. By creating a dashboard with views for team, individuals, and management, the analyst can customize the content and layout of the dashboard for different audiences and purposes. By configuring permissions to control access, the analyst can ensure that the confidential information, such as individuals?? earned commissions and projected commissions based on sales, is only visible to the authorized users. The other ways are not the best way to provide this visibility. Here is why:
Creating a dashboard displaying a data refresh date so users know the current sales numbers and configuring permissions to control access would not be sufficient to provide visibility of pipeline and team performance, as well as individuals?? earned commissions and projected commissions based on sales. The dashboard would only show the current sales numbers and the date when the data was updated, which would not give a comprehensive or detailed view of the sales situation.
Creating a dashboard for sales numbers, pipeline, and team and individual performance for the management team would not be appropriate to provide visibility for the sales team, as they would not have access to the dashboard or the information they need. The dashboard would only be available for the management team, which would limit the transparency and collaboration among the sales team members.
Creating a dashboard with filters for the overall team, individuals, and management would not be secure to provide visibility of confidential information, such as individuals?? earned commissions and projected commissions based on sales. The dashboard would allow users to filter and see the data they want, which could expose sensitive or personal information to unauthorized users.

NEW QUESTION 8
A site reliability team wants to monitor the stability of their website. so they can proactively diagnose issues when they occur Which of the following deliverables would best suit their needs?

A. A self-serve dashboard of website performance that updates in real time

B. A weekly log report of site visits and user actions
C. A portal that is refreshed daily and reports errors classified by type
D. A daily summary email indicating website outages for the previous day

Answer: A

Explanation:
The best deliverable that would suit the site reliability team??s needs is A. A self-serve dashboard of website performance that updates in real time.
A self-serve dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance. A self-serve dashboard of website performance that updates in real time would allow the site reliability team to easily and quickly access the information they need about the stability of their website, such as uptime, response time, error rate, traffic volume, etc. A self-serve dashboard would also enable the team to proactively diagnose issues when they occur, by providing alerts, notifications, or drill-down options. A self-serve dashboard would also be more interactive and engaging than a report or an email.
A weekly log report of site visits and user actions would not be a good deliverable for the site reliability team??s needs, because it would not provide timely or relevant information about the stability of their website. A weekly log report would be too infrequent and delayed to monitor and diagnose issues when they occur. A weekly log report would also focus on the behavior and actions of the users, rather than the performance and functionality of the website.
A portal that is refreshed daily and reports errors classified by type would not be a good deliverable for the site reliability team??s needs, because it would not provide real-time or comprehensive information about the stability of their website. A portal that is refreshed daily would be too slow and outdated to monitor and diagnose issues when they occur. A portal that reports errors classified by type would be too narrow and limited to capture the full picture of the website performance.
A daily summary email indicating website outages for the previous day would not be a good deliverable for the site reliability team??s needs, because it would not provide real-time or actionable information about the stability of their website. A daily summary email would be too late and retrospective to monitor and diagnose issues when they occur. A daily summary email indicating website outages would also be too passive and generic to help the team resolve or prevent issues in the future.

NEW QUESTION 9
Which of the following is an example of a discrete data type?

A. 8in (20cm)
B. 5 kids
C. 2.5mi (4km)
D. 10.7lbs (4.9kg)

Answer: B

Explanation:
A discrete data type is a data type that can only take on a finite number of values, such as integers or categories. An example of a discrete data type is the number of kids, as it can only be a whole number. The other options are examples of continuous data types, as they can take on any value within a range. The length in inches or centimeters, the distance in miles or kilometers, and the weight in pounds or kilograms are all continuous data types. Reference: CompTIA Data+ (DA0-001) Practice Certification Exams | Udemy

NEW QUESTION 10
Jenny wants to study the academic performance of undergraduate sophomores and wants to determine the average grade point average at different points during an academic year.
What best describes the data set she needs?

A. Sample.
B. Observation.
C. Variable.
D. Population.

Answer: A

Explanation:
Correct answer A. Sample.
Jenny does not have data for the entire population of all undergraduate sophomores. While a specific grade point average is an observation of variable, jenny needs sample data.

NEW QUESTION 10
Which of the following data cleansing issues will be fixed when a DISTINCT function is applied?

A. Missing data
B. Duplicate data
C. Redundant data
D. Invalid data

Answer: B

Explanation:
This is because duplicate data refers to data that is repeated or copied in a data set, which can affect the quality and validity of the analysis. A DISTINCT function is a type of function that removes duplicate values from a column or a table, leaving only unique values. For example, a DISTINCT function in SQL that can achieve this is:

```
SELECT DISTINCT column_name FROM table_name;
```

The other data cleansing issues will not be fixed by applying a DISTINCT function. Here is why:
Missing data refers to data that is absent or incomplete in a data set, which can affect the accuracy and reliability of the analysis. A DISTINCT function does not

help with missing data, because it does not fill in or impute the missing values.
Redundant data refers to data that is unnecessary or irrelevant for the analysis, which can affect the efficiency and performance of the analysis. A DISTINCT function does not help with redundant data, because it does not remove or filter out the redundant values.
Invalid data refers to data that is incorrect or inaccurate in a data set, which can affect the validity and reliability of the analysis. A DISTINCT function does not help with invalid data, because it does not validate or correct the invalid values.

NEW QUESTION 12
An analyst modified a data set that had a number of issues. Given the original and modified versions:

Original data:

Var001	Var002	Var003	Var004
1	0	0	0
0	1	0	1
1	1	1	2
0	0	0	1

Modified data:

Var001	Var002	Var003	Var004
Yes	Absent	No payment	No
No	Present	No payment	Yes
Yes	Present	Payment	Maybe
No	Absent	No payment	Yes

Which of the following data manipulation techniques did the analyst use?

A. Imputation
B. Recoding
C. Parsing
D. Deriving

Answer: B

Explanation:
The correct answer is B. Recoding.
Recoding is a data manipulation technique that involves changing the values or categories of a variable to make it more suitable for analysis. Recoding can be used to simplify or group the data, to correct errors or inconsistencies, or to create new variables from existing ones12
In the example, the analyst used recoding to change the values of Var001, Var002, Var003, and Var004 from numerical to textual form. The analyst also used recoding to assign meaningful labels to the values, such as ??Absent?? for 0, ??Present?? for 1, ??Low?? for 2, ??Medium?? for 3, and ??High?? for 4. This makes the data more understandable and easier to analyze.

NEW QUESTION 14
A sales director has requested a report for individual team members within the division be developed. The director would like the report to be shared with all team members, but individual team members should not be identifiable within the report Which of the following access requirements would support the director's needs?

A. Create an acceptable use policy for the sales data.
B. Release the report as user-group-based access and include data masking.
C. Get a data use agreement from the individual team members.
D. Provide the report based on role and include data encryption.

Answer: B

NEW QUESTION 19
A marketing analytics team received customer transaction data from two different sources. The data is complete and accurate; however, the field names appear to be inconsistent. Given the following tables:

Online transactions:

Customer_ID	Channel	Segment	Amount ($)
001	Online	Existing	3,000
002	Online	Existing	4,000
003	Online	New	1,500

Store transactions:

Customer_ID	Source	Segment	Amount ($)
001	In-store	New	1,000
004	In-store	Existing	4,000
005	In-store	New	3,500

Which of the following is considered best practice if the team wants to consolidate the files and conduct further analysis?

A. Standardize the field names.
B. Recode the data values.
C. Overwrite the field names in one of the tables.
D. Edit the field names in the data dictionary.

Answer: A

Explanation:
When consolidating data from different sources, it is crucial to standardize field names to ensure consistency across datasets. This process involves aligning the field names so that they are the same in both tables, which simplifies the merging of data and subsequent analysis. Standardizing field names helps in maintaining data integrity and avoids confusion that may arise from having different names for the same data point. Recode the data values (B) would not be necessary unless the data values themselves are inconsistent or in different formats. Overwriting the field names in one of the tables © could lead to loss of information or confusion. Editing the field names in the data dictionary (D) is helpful, but it does not address the immediate need to harmonize the field names in the actual datasets.
References:
? Best practices in data management.
? Principles of data integration and consolidation.

NEW QUESTION 21
A data analyst wants to create "Income Categories" that would be calculated based on the existing variable "Income". The "Income Categories" would be as follows:
Income category 1: less than $1.
Income category 2: more than $1 and less than $20,000. Income category 3: more than $20,001 and less than $40,000. Income category 4: more than $40,001.
Which of the following data manipulation techniques should the data analyst use to create "Income Categories"?

A. Data merge
B. Derived variables
C. Data blending
D. Data append

Answer: B

Explanation:

The correct answer is B: Derived variables Derived variables are variables that you create by calculating or categorizing variables that already exist in your data set.
Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set. Data blending is incorrect.
Data blending involves pulling data from different sources and creating a single, unique, dataset for visualization and analysis.
Data append is incorrect. A data append is a process that involves adding new data elements to an existing database.

NEW QUESTION 26
Which of the following data manipulation techniques is an example of a logical function?

A. WHERE
B. AGGREGATE
C. BOOLEAN
D. IF

Answer: D

Explanation:
This is because an IF function is a type of logical function that returns a value based on a condition or a set of conditions. An IF function can be used to manipulate data by applying different actions or calculations depending on whether the condition is true or false. For example, an IF function in Excel that can achieve this is:
=IF (condition, value_if_true, value_if_false)
The other data manipulation techniques are not examples of logical functions. Here is why:
? WHERE is a type of clause that filters data based on a condition or a set of conditions. A WHERE clause can be used to manipulate data by selecting only the rows that satisfy the condition(s). For example, a WHERE clause in SQL that can achieve this is:

```
SELECT column_name FROM table_name WHERE condition;
```

? AGGREGATE is a type of function that performs a calculation on a group of values, such as sum, average, count, etc. An AGGREGATE function can be used to manipulate data by summarizing or aggregating the values in a column or a table. For example, an AGGREGATE function in SQL that can achieve this is:

```
SELECT AGGREGATE(column_name) FROM table_name;
```

? BOOLEAN is a type of data type that represents two possible values: true or false.
A BOOLEAN data type can be used to manipulate data by storing or returning logical values based on a condition or a set of conditions. For example, a BOOLEAN data type in Python that can achieve this is:

```
boolean_variable = condition
```

NEW QUESTION 27
Given the following data:

Name	Gender	Age	Annual income
Ralph	M	27	$75,000
Jessie	F	3	$75,000
Monica	F	31	$125,000
Carlos	M	53	$75
Sara	F	43	$0

Which of the following BEST describes the data set?

A. There is data bias.
B. The data is incomplete.
C. The data is inconsistent.
D. The data is outliers.

Answer: C

Explanation:
This is because inconsistency is a type of data quality issue that occurs when the data does not follow a common format, structure, or rule across different sources or systems, which can affect the efficiency and performance of the analysis or process. Inconsistency can be caused by having different spellings, punctuations, capitalizations, or abbreviations for the same or similar values in a data set, such as ??M??, ??m??, ??Male??, or ??male?? for gender in this case. Inconsistency can be eliminated or reduced by using data cleansing techniques, such as standardizing or normalizing the data values. The other options are not correct descriptions of the data set. Here is why:
? Data bias is a type of data quality issue that occurs when the data is not representative or proportional of the population or the parameter, which can affect the validity and reliability of the analysis or process. Data bias can be caused by having a sample that is too small, too large, or too skewed for the population or the parameter, such as having only male customers for a product that targets both genders in this case. Data bias can be eliminated or reduced by using sampling techniques, such as stratified or cluster sampling.
? The data is incomplete is a type of data quality issue that occurs when the data is
absent or missing in a data set, which can affect the accuracy and reliability of the analysis or process. The data is incomplete can be caused by various factors, such as human error, system error, or non-response. The data is incomplete can be addressed by using various methods, such as replacing or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression.
? The data is outliers is a type of data quality issue that occurs when the data has
values that are unusually high or low compared to the rest of the data set, which can affect the quality and validity of the analysis or process. The data is outliers can be caused by various factors, such as measurement error, natural variation, or extreme events. The data is outliers can be addressed by using various methods, such as removing or filtering out the outliers, or using robust statistics that are less sensitive to outliers, such as median, interquartile range, or box plot.

NEW QUESTION 30
A data analyst received the information in the table below from a recently completed marketing campaign:

Channels	Clicks	Orders
Display	580	55
PPC	800	100
Social	1,200	220
Mobile	300	60
SEO	620	85

Which of the following is the total order conversion rate?

A. 13.2%
B. 14.8%
C. 22.3%
D. 85.2%

Answer: B

Explanation:
The correct answer is A. 13.2%.
The total order conversion rate is the ratio of the total number of orders to the total number of clicks, expressed as a percentage. To calculate the total order conversion rate, we need to sum up the clicks and orders from all the channels, and then divide the orders by the clicks and multiply by 100.
Using the data from the table, we can do the following:
? Total clicks = 580 + 800 + 1,200 + 300 + 620 = 3,500
? Total orders = 55 + 100 + 220 + 60 + 85 = 520
? Total order conversion rate = (520 / 3,500) x 100 = 14.857%
? Rounding to one decimal place, we get 14.9% Therefore, the total order conversion rate is 14.9%.

NEW QUESTION 34
An analyst is building a new dashboard for a user. After an initial conversation with the user. the analyst created a mock-up of the dashboard. Which of the following best explains why the analyst created the mock-up?

A. To identify the dimensions and measures
B. To send to the client after deploying the dashboard to production
C. To confirm important details before dashboard development begins
D. To receive client approval for the final dashboard design

Answer: C

Explanation:
Answer C. To confirm important details before dashboard development begins.
A dashboard mockup is a prototype of a finished dashboard directly in the product. It is a way to visualize the layout, design, and functionality of the dashboard before it is built with real data and code. A dashboard mockup can help the analyst to confirm important details
with the user, such as the business objectives, the key performance indicators, the data sources, the filters, the charts, and the interactivity. By creating a dashboard mockup, the analyst can get immediate feedback and validation from the user, and avoid wasting time and resources on developing a dashboard that does not meet the user??s expectations or needs1.

NEW QUESTION 38
Which of following is a non-relational database?

A. Neo4j
B. SQLite
C. MySQL
D. PostgreSQL

Answer: A

Explanation:
Neo4j is a type of non-relational database that uses a graph model to store data. A graph database is a database that represents data as nodes and edges, where nodes are entities and edges are relationships between them. A graph database can store complex and diverse data that is not easily structured in tables. A graph database can also perform fast and efficient queries on the data by traversing the connections between the nodes

NEW QUESTION 40
An analyst collected data that includes primary account numbers, expiration dates, and service codes. Which of the following data governance classifications is used to describe this data?

A. PI I
B. PCI
C. PBI
D. PHI

Answer: B

NEW QUESTION 41
Which of the following will MOST likely be streamed live?

A. Machine data
B. Key-value pairs
C. Delimited rows
D. Flat files

Answer: A

Explanation:
Machine data is the most likely type of data to be streamed live, as it refers to data generated by machines or devices, such as sensors, web servers, network devices, etc. Machine data is often produced continuously and in large volumes, requiring real-time processing and analysis. Other types of data, such as key-value pairs, delimited rows, and flat files, are more likely to be stored in databases or files and processed in batches.

NEW QUESTION 42
An analyst is reviewing the following data: Car IDSpeed
123155
566436
564418
650567
546436
645638
Which of the following should the analyst include in the measures of central tendency for speed?

A. Mode = 38 Range = 31 Mean = 42.5
B. Range = 49 Max = 67 Min = 18
C. Mode = 36 Max = 67 Min = 18
D. Mode = 36 Median = 37 Mean = 41.5

Answer: D

Explanation:
The measures of central tendency include the mode, median, and mean. The mode is the value that appears most frequently in a data set. In this case, the speed of 36 appears twice, making it the mode. The median is the middle value when a data set is ordered from least to greatest; for these speeds, when ordered (18, 36, 36, 38, 55, 67), the median is the average of the two middle numbers, which is (\frac{36 + 38}{2} = 37). The mean is the average of all values, calculated as (\frac{55 + 36 + 18 + 67 + 36 + 38}{6} = 41.7). References:
? The calculation of the mode, median, and mean is based on standard statistical
formulas and definitions.
The measures of central tendency for speed include the mode, median, and mean. To calculate these, we first need to organize the data:
? Speeds in ascending order: 18, 36, 36, 38, 55, 67
? Mode is the value that appears most frequently, which is 36, as it appears twice.
? Median is the middle value when the data is ordered. Since we have an even number of observations, we take the average of the two middle values (36 and 38), resulting in 37.
? Mean is the sum of all values divided by the number of values. (18+36+36+38+55+67)/6=41.5(18+36+36+38+55+67)/6=41.5.
Thus, the correct option is D, which includes Mode = 36, Median = 37, and Mean = 41.5. The range, maximum, and minimum values, although useful in understanding data dispersion, are not measures of central tendency and are therefore not relevant to this specific question.

NEW QUESTION 47
When analyzing the values of two variables, you decide to convert both variables so they are on a scale of 0 to 1.
What term describes this action?

A. Filtering.
B. Normalization.
C. Transposition.
D. Aggregation.

Answer: B

Explanation:

Normalization is the process of reorganizing data in a database so that it meets two basic requirements: There is no redundancy of data, all data is stored in only one place. Data dependencies are logical, all related data items are stored together.
Put simply, data normalization ensures that your data looks, reads, and can be utilized the same way across all of the records in your customer database. This is done by standardizing the formats of specific fields and records within your customer database.

NEW QUESTION 49
An analyst reviews the following data: 7
3
5
2
3

7
7
10
Which of the following is the value of the mode?

A. 3
B. 5
C. 7
D. 10

Answer: C

Explanation:
The mode is the value that appears most frequently in a data set. In the provided data set, the number 7 appears three times, which is more than any other number. Therefore, the mode of this data set is 7.
? 3 appears twice, but less frequently than 7.
? 5 and 10 each appear only once, so they cannot be the mode.
References:
? Mode in Statistics - Definition and Examples1
? Understanding Measures of Central Tendency2
? Mode (statistics) - Wikipedia3

NEW QUESTION 52
Which of the following data types would a telephone number formatted as XXX-XXX-XXXX be considered?

A. Numeric
B. Date
C. Float
D. Text

Answer: D

Explanation:
A telephone number formatted as XXX-XXX-XXXX would be considered a text data type, as it is composed of alphanumeric characters and symbols. A numeric data type is composed of only numbers, such as integers or decimals. A date data type is composed of values that represent dates or times, such as YYYY-MM-DD or HH:MM:SS. A float data type is composed of numbers with fractional parts, such as 3.14 or 0.5. Reference: Guide to CompTIA Data+ and Practice Questions - Pass Your Cert

NEW QUESTION 56
A database consists of one fact table that is composed of multiple dimensions. Each dimension is represented by a denormalized table. This structure is an example of a:

A. non-relational schema.
B. galaxy schema.
C. snowflake schema.
D. star schema.

Answer: D

Explanation:
A star schema is a type of database schema that consists of one fact table and multiple dimension tables. The fact table contains the measures or metrics of the business process, such as sales, orders, or transactions. The dimension tables contain the attributes or characteristics of the business entities, such as products, customers, or locations. The fact table is connected to the dimension tables by foreign keys that reference the primary keys of the dimension tables. The fact table is located at the center of the schema, while the dimension tables are located at the edges, forming a star-like shape1.
A star schema is an example of a denormalized schema, which means that the dimension tables are not normalized and may contain redundant or repeated data. This is done to improve the performance and simplicity of queries, as there are fewer joins and tables involved. A star schema is suitable for data warehouses and business intelligence applications that require fast and efficient data retrieval2.

NEW QUESTION 57
Given the following table:

Code	New_Measure	Old_Measure
A	10	12
B	14	12
C	5	12
D	9	12

Which of the following methods is the best way to describe the changes in the values in the table?

A. Average
B. Range
C. Standard deviation
D. Median

Answer: B

NEW QUESTION 60
What SQL command is used to delete an entire table from a database?

A. DROP.
B. MODIFY.
C. DELETE.
D. ALTER.

Answer: A

NEW QUESTION 61
A junior web developer is developing a new application where users can upload short videos. The first task is to create a homepage that shows the headline "Upload Your Short Videos" and a clickable button that says "upload now".
Which of the following HTML commands would help the developer to complete the task successfully?

A. < span >Upload Your Short Videos< /span >< button >upload now< /button >
B. < p >Upload Your Short Videos< /p >< p >upload now< /p >
C. < hl >Upload Your Short Videos< /h1 >< button >upload now< /button >
D. < hl >Upload Your Short Videos< /h1 >< hl >upload now< /h1 >

Answer: C

Explanation:
The HTML commands that would help the developer to complete the task successfully are
<h1>Upload Your Short Videos</h1> and <button>upload now</button>. The <h1> tag defines a heading level 1, which is the largest and most important heading on a webpage. The <button> tag defines a clickable button that can perform some action when clicked. The other options are not suitable for the task, as they either use the wrong tags or do not create a clickable button. The tag defines a section of text with no specific meaning or formatting. The <p> tag defines a paragraph of text. The <hl> tag does not exist in HTML. Reference: HTML Tags - W3Schools

NEW QUESTION 63
A county in Illinois is conducting a survey to determine the mean annual income per household. The county is 427sq mi (2.65q km). Which of the following sampling methods would MOST likely result in a representative sample?

A. A stratified phone survey of 100 people that is conducted between 2:00 p.
B. and 3:00 p.m.
C. A systematic survey that is sent to 100 single-family homes in the county
D. Surveys sent to ten randomly selected homes within 5mi (8km) of the county??s office
E. Surveys sent to 100 randomly selected homes that are reflective of the population

Answer: D

Explanation:
Surveys sent to 100 randomly selected homes that are reflective of the population. This is because a random sample is a type of sample that is selected by using a random method, such as a lottery or a computer-generated number, which ensures that every element in the population has an equal chance of being selected. A random sample can result in a representative sample, which means that the sample reflects the characteristics and diversity of the population. By sending surveys to 100 randomly selected homes that are reflective of the population, the analyst can ensure that the sample is representative of the county??s households and their income levels. The other sampling methods are not likely to result in a representative sample. Here is why:
A stratified phone survey of 100 people that is conducted between 2:00 p.m. and 3:00 p.m. would result in a biased sample, which means that the sample favors or excludes certain groups or elements in the population. By conducting the survey only between 2:00 p.m. and 3:00 p.m., the analyst would miss out on people who are not available or reachable at that time, such as those who are working or sleeping. This could affect the representativeness and generalizability of the sample.
A systematic survey that is sent to 100 single-family homes in the county would result in an unrepresentative sample, which means that the sample does not reflect the characteristics and diversity of the population. By sending surveys only to single-family homes, the analyst would ignore other types of households, such as apartments, condos, or mobile homes. This could affect the accuracy and reliability of the sample.
Surveys sent to ten randomly selected homes within 5mi (8km) of the county??s office would result in a small sample, which means that the sample size is too low to capture the variability and diversity of the population. By sending surveys only to ten homes within a limited area, the analyst would miss out on many households that are located in different parts of the county. This could affect the precision and confidence of the sample.

NEW QUESTION 67
An analyst wants to create a historical data set for the past five years with each year in its own data set. Which of the following methods is the best way to create this historical data set?

A. Data transpose
B. Data concatenation
C. Data append
D. Data normalization

Answer: B

NEW QUESTION 72
A data analyst needs to create a master file that includes customer information from the tables below:

Table 1: Online Transactions

Order_ID	Customer_ID	Date	Amount	Quantity
002A	002	03/01/2020	$800	109
001B	001	02/01/2020	$400	14
001B	001	02/01/2020	$400	14
001B	001	02/01/2020	$400	14
004C	004	06/01/2020	$700	52
003D	003	05/01/2020	$900	20

Table 2: In-store Transactions

Order_ID	Customer_ID	Date	Amount	Quantity
006A	006	04/01/2020	$200	59
007B	007	03/01/2020	$500	54
008C	008	02/01/2020	$600	15
009D	009	05/01/2020	$800	18
001E	001	07/01/2020	$300	50
003F	003	08/01/2020	$200	55

Table 3: Customer Table

Customer_ID	Segment	Region
001	New	BC
002	Existing	ON
003	New	MB
004	New	ON
005	Existing	AT
006	Existing	MB
007	New	QC
008	New	QC
009	Existing	BC

Given the three tables above, the analyst wants to filter down the information prior to joining it together. In which of the following orders should this data manipulation bo approached for the most efficient result?

A. Merge, append, deduplicate
B. Merge, deduplicate, append
C. Deduplicate, append, merge
D. Append, deduplicate, merge

Answer: B

Explanation:
For efficient data manipulation, the ideal order would be to first merge related tables to create a comprehensive set of records, then deduplicate to remove any redundant information. Lastly, appending additional data, such as from another source or table, ensures that all relevant data is included without redundancy before the final analysis. This order prevents unnecessary duplication of effort, such as deduplicating both before and after appending, which would be less efficient.
In the context of the tables provided, merging would likely involve combining customer information from the online and in-store transaction tables with the customer table. Deduplication would remove any redundant customer records that may exist across these tables. Finally, appending would involve adding any additional transaction records to the master file, ensuring a complete dataset for analysis.

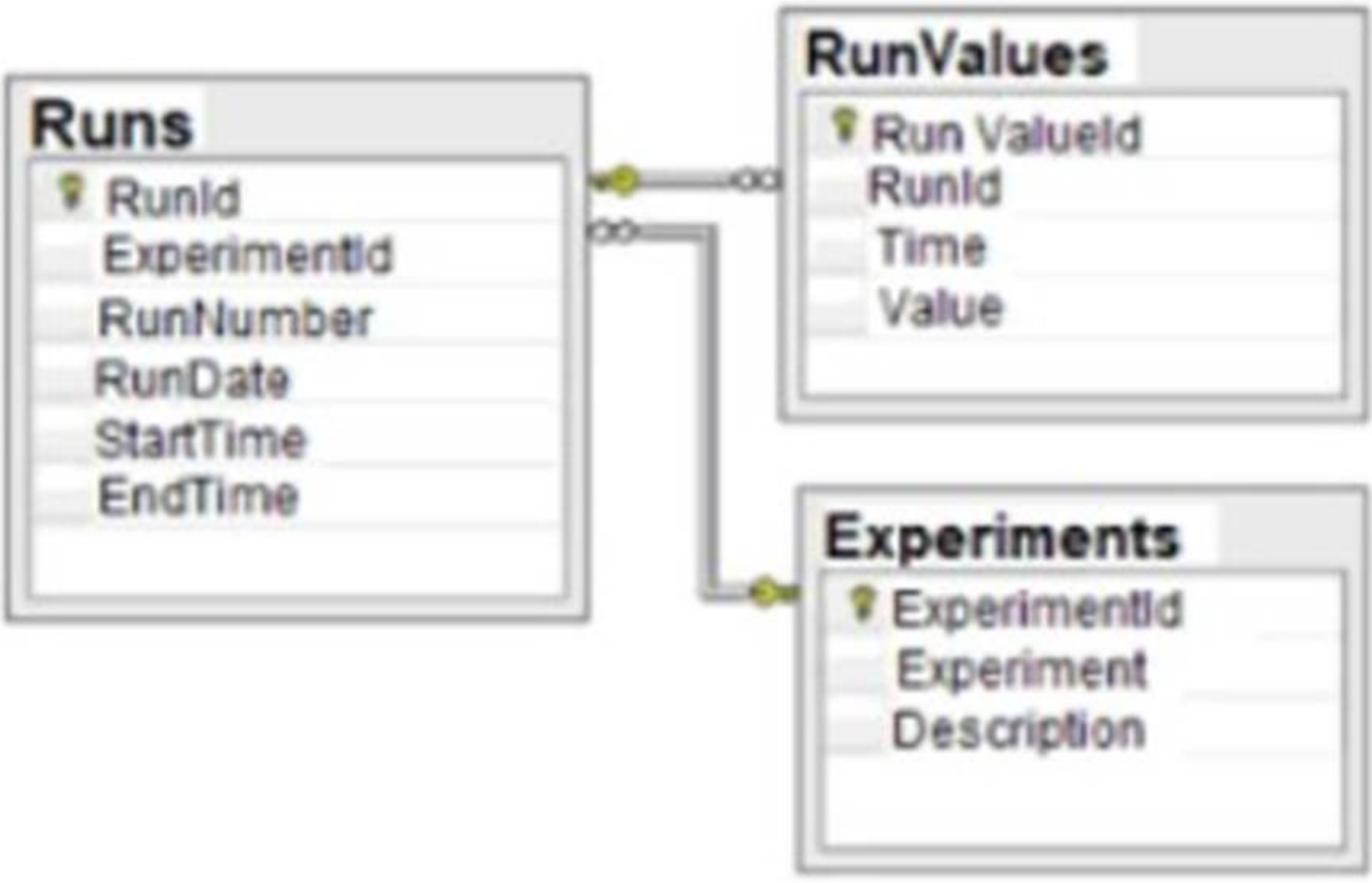
NEW QUESTION 74
Which of the following best describes the law of large numbers?

A. As a sample size decreases, its standard deviation gets closer to the average of the whole population.
B. As a sample size grows, its mean gets closer to the average of the whole population
C. As a sample size decreases, its mean gets closer to the average of the whole population.
D. When a sample size double
E. the sample is indicative of the whole population.

Answer: B

Explanation:
The best answer is B. As a sample size grows, its mean gets closer to the average of the whole population.
The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population. This is due to the sample being more representative of the population as it increases in size. The law of large numbers guarantees stable long-term results for the averages of some random events1
* A. As a sample size decreases, its standard deviation gets closer to the average of the whole population is not correct, because it confuses the concepts of standard deviation and mean. Standard deviation is a measure of how much the values in a data set vary from the mean, not how close the mean is to the population average. Also, as a sample size decreases, its standard deviation tends to increase, not decrease, because the sample becomes less representative of the population.
* C. As a sample size decreases, its mean gets closer to the average of the whole population is not correct, because it contradicts the law of large numbers. As a sample size decreases, its mean tends to deviate from the average of the whole population, because the sample becomes less representative of the population.
* D. When a sample size doubles, the sample is indicative of the whole population is not correct, because it does not specify how close the sample mean is to the population average. Doubling the sample size does not necessarily make the sample indicative of the whole population, unless the sample size is large enough to begin with. The law of large numbers does not state a specific number or proportion of samples that are indicative of the whole population, but rather describes how the sample mean approaches the population average as the sample size increases indefinitely.

NEW QUESTION 76
Given the diagram below:



Which of the following data schemas shown?

A. Key-value pairs
B. Online transactional processing
C. Data Lake
D. Relational database

Answer: D

Explanation:
A relational database is a type of database that organizes data into tables, where each table has a fixed number of columns and a variable number of rows. Each row in a table represents a record or an entity, and each column represents an attribute or a property of that entity. The tables are linked by common fields, called keys, which enable the database to establish relationships between the data. A relational database schema is a diagram that shows the structure and organization of the tables, columns, keys, and constraints in a relational database. The diagram given in the question is an example of a relational database schema, as it shows two tables: ??Runs?? and ??Experiments??, with their respective columns, data types, and primary keys. The ??Runs?? table also has a foreign key that references the ??ExperimentId?? column in the ??Experiments?? table, indicating a relationship between the two tables. Therefore, the correct answer is D.
References: What is a database schema? | IBM, Database Schema - Javatpoint

NEW QUESTION 78
Which of the following report types is most appropriate for a high-level, year-end report requested by a Chief Executive Officer?

A. Dynamic
B. Recurring
C. Ad hoc
D. Self-service

Answer: B

Explanation:
For a high-level, year-end report requested by a Chief Executive Officer (CEO), a recurring report type is most appropriate. Recurring reports are regular, scheduled reports that typically summarize information over a set period, such as a fiscal year. They provide a consistent format for executives to track performance over time, and their standardized nature makes them suitable for high-level analysis and decision-making. Since CEOs need to monitor performance and make strategic decisions, a recurring report that provides a comprehensive overview of the year's activities and outcomes would be valuable. This allows the CEO to evaluate the company's performance against its goals and objectives systematically.
Dynamic reports (A) are more interactive and typically used for in-depth analysis where users can drill down into the data. Ad hoc reports (C) are one-time, usually unscheduled reports tailored for specific questions, which may not be as comprehensive as a year-end report requires. Self-service reports (D) allow users to create their reports on demand, which may not be the formal, synthesized view a CEO would need for a year-end report.

NEW QUESTION 81
A customer list from a financial services company is shown below:

Name	Number of credit cards	Age	Income
Sean	0	27	$60,000
Angela	4	31	$50,000
Terry	3	40	$170,000
Paula	1	25	$70,000
Malcolm	3	28	$150,000

A data analyst wants to create a likely-to-buy score on a scale from 0 to 100, based on an average of the three numerical variables: number of credit cards, age, and income. Which of the following should the analyst do to the variables to ensure they all have the same weight in the score calculation?

A. Recode the variables.
B. Calculate the percentiles of the variables.
C. Calculate the standard deviations of the variables.
D. Normalize the variables.

Answer: D

Explanation:
Normalizing the variables means scaling them to a common range, such as 0 to 1 or -1 to 1, so that they have the same weight in the score calculation. Recoding the variables means changing their values or categories, which would alter their meaning and distribution. Calculating the percentiles of the variables means ranking them relative to each other, which would not account for their actual magnitudes. Calculating the standard deviations of the variables means measuring their variability, which would not make them comparable. References: CompTIA Data+ Certification Exam Objectives, page 10

NEW QUESTION 83
Which of the following would be considered non-personally identifiable information?

A. Cell phone device name
B. Customer??s name
C. Government ID number
D. Telephone number

Answer: A

Explanation:
Non-personally identifiable information (non-PII) is any data that cannot be used to identify, contact, or locate a specific individual, either alone or combined with other sources. Non-PII can include aggregated statistics, anonymous data, device identifiers, IP addresses, cookies, and other types of information that do not reveal the identity or location of a person. Cell phone device name is an example of non-PII, as it does not reveal any personal information about the owner or user of the device. Therefore, the correct answer is A. References: What is Non-Personally Identifiable Information (Non-PII)? | Definition and Examples, What is Personally Identifiable Information (PII)? | Definition and Examples

NEW QUESTION 87
Which of the following types of analyses should be used to evaluate the connections and anomalies in a data set when either known patterns are being violated or new patterns are emerging?

A. Correlation
B. Descriptive
C. Graph
D. Regression

Answer: C

NEW QUESTION 91
A Chief Executive Officer (CEO) is requesting more up-to-date sales data for improved visibility prior to month-end. An analyst must determine the frequency of a sales report that was previously distributed on an as-needed basis. Which of the following would be the most appropriate frequency for this report?

A. Monthly
B. Quarterly
C. Weekly
D. Every other month

Answer: C

Explanation:
The most appropriate frequency for the sales report is weekly, as this will provide the CEO with more up-to-date sales data for improved visibility prior to month-end. A weekly sales report can show the sales performance, trends, and issues of the sales team on a regular basis, and help the CEO to monitor and evaluate the progress and results of the sales activities. A weekly sales report can also help the CEO to identify and address any problems or opportunities that may arise during the month, and to make timely and informed decisions.

NEW QUESTION 95
A data scientist wants to see which products make the most money and which products attract the most customer purchasing interest in their company.
Which of the following data manipulation techniques would he use to obtain this information?

A. Data append
B. Data blending
C. Normalize data
D. Data merge

Answer: B

Explanation:

The correct answer is B: Data blending.
Data blending is combining multiple data sources to create a single, new dataset, which can be presented visually in a dashboard or other visualization and can then be processed or analyzed. Enterprises get their data from a variety of sources, and users may want to temporarily bring together different datasets to compare data relationships or answer a specific question. Data append is incorrect. Data append is a process that involves adding new data elements to an existing database. An example of a common data append would be the enhancement of a company's customer files. A data append takes the information they have, matches it against a larger database of business data, allowing the desired missing data fields to be added. Normalize data is incorrect.
Data normalization is the process of structuring your relational customer database, following a series of normal forms. This improves the accuracy and integrity of your data while ensuring that your database is easier to navigate. Data merge is incorrect. Data merging is the process of combining two or more data sets into a single data set.

NEW QUESTION 99
Alex wants to use data from his corporate sale, CRM, and shipping systems to try and predict future sales.
Which of the following systems is the most appropriate? Choose the best answer.

A. Data mart.
B. OLAP.
C. Data Warehouse.
D. OLTP.

Answer: C

Explanation:

Correct Answer: C. Data Warehouse.
Data warehouse bring together data from multiple systems used by an organization. A data mart is too narrow, as Alex needs data from across multiple divisions. OLAP is a broad term of analytical processing, and OLTP systems are transactional and not ideal for this task.

NEW QUESTION 103
Which of the following best describes the process of examining data for statistics and information about the data?

A. Cleansing
B. search
C. Profiling
D. Governance

Answer: C

Explanation:
Data profiling is the process of examining data for statistics and information about the data, such as the structure, format, quality, and content of the data. Data profiling can help to understand the characteristics, patterns, relationships, and anomalies of the data, as well as to identify and resolve any errors, inconsistencies, or missing values in the data. Data profiling can be done using various tools and methods, such as spreadsheets, databases, or programming languages12.

NEW QUESTION 107
A research analyst collects ten data points from 1.000 specimens. The analyst will not need any additional data to complete the analysis and will not need to retrieve information by specifier. Which of the following is the best data structure for the analyst to use?

A. NoSQL
B. Flat file
C. JSON
D. Relational database

Answer: B

Explanation:
A flat file is a type of data structure that stores data in a plain text format, such as CSV, TSV, or TXT. A flat file consists of one or more records, each containing one or more fields, separated by a delimiter, such as a comma, tab, or space. A flat file does not have any hierarchical or relational structure, and does not support any complex queries or operations1.
A flat file may be the best data structure for the analyst to use in this scenario, because:
? The analyst collects ten data points from 1,000 specimens, which means the data is relatively small and simple, and can be easily stored and processed in a flat file.
? The analyst will not need any additional data to complete the analysis, which means the data is static and does not require any updates or modifications.
? The analyst will not need to retrieve information by specifier, which means the data
does not require any indexing or searching by key or value.

NEW QUESTION 112
A development company is constructing a new Init in its apartment complex. The complex has the following floor plans:

Unit name	Sq. Ft.	Price	$/Sq. Ft.
Jasmine	1,000	$345,000	$345
Orchid	1,100	$425,000	$386
Azalea	1,300	$460,000	$354
Tulip	1,640	$525,000	$320
Rose	2,000		

Using the average cost per square foot of the original floor plans. which of the following should be the price of the Rose Init?

A. $640,900
B. $690,000
C. $705,200
D. $702,500

Answer: D

Explanation:
The correct answer is D. $702,500.
To find the price of the Rose unit, we need to use the average cost per square foot of the original floor plans. The average cost per square foot is calculated by dividing the price by the square footage of each unit type. Using the data from the table, we can do the following:
? Jasmine: $345,000 / 1,000 = $345 per square foot
? Orchid: $525,000 / 2,000 = $262.5 per square foot
? Azalea: $375,000 / 1,500 = $250 per square foot
? Tulip: $450,000 / 1,800 = $250 per square foot
The average cost per square foot of the original floor plans is the mean of these four values, which is ($345 + $262.5 + $250 + $250) / 4 = $276.875 per square foot.
To find the price of the Rose unit, we need to multiply the average cost per square foot by the square footage of the Rose unit. The Rose unit has a square footage of 2,535, according to the table. Therefore, the price of the Rose unit is $276.875 x 2,535 = $702,421.875.
Rounding to the nearest whole number, we get $702,500 as the price of the Rose unit.

NEW QUESTION 114
While reviewing survey data, a research analyst notices data is missing from all the responses to a single question. Which of the following methods would BEST address this issue?

A. Replace missing data.
B. Remove duplicate data.
C. Replace redundant data.
D. Remove invalid data.

Answer: A

Explanation:
This is because missing data is a type of data quality issue that occurs when data is absent or incomplete in a data set, which can affect the accuracy and reliability of the analysis or process. Missing data can be caused by various factors, such as human error, system error, or non-response. Missing data can be addressed by using various methods, such as replacing missing data, which means filling in or imputing the missing values with some reasonable estimates, such as mean, median, mode, or regression. The other methods are not used to address missing data. Here is why:
? Remove duplicate data is a type of method that eliminates or reduces duplicate data, which is a type of data quality issue that occurs when data is repeated or copied in a data set. Removing duplicate data does not address missing data, but rather affects the quantity and validity of the data.
? Replace redundant data is a type of method that eliminates or reduces redundant data, which is a type of data quality issue that occurs when data is unnecessary or irrelevant for the analysis or purpose. Replacing redundant data does not address missing data, but rather affects the efficiency and performance of the analysis or process.
? Remove invalid data is a type of method that eliminates or reduces invalid data, which is a type of data quality issue that occurs when data is incorrect or inaccurate in a data set. Removing invalid data does not address missing data, but rather affects the validity and reliability of the analysis or process.

NEW QUESTION 117
An analyst is required to run a text analysis of data that is found in articles from a digital news outlet. Which of the following would be the BEST technique for the analyst to apply to acquire the data?

A. Web scraping
B. Sampling
C. Data wrangling
D. ETL

Answer: A

Explanation:
This is because web scraping is a technique that allows the analyst to extract data from web pages, such as articles from a digital news outlet. Web scraping can be done using various tools and methods, such as Python libraries, browser extensions, or online services. The other techniques are not suitable for acquiring data from web pages. Here is why:
Sampling is a technique that involves selecting a subset of data from a larger population, usually for statistical analysis or testing purposes. Sampling does not help the analyst to acquire data from web pages, but rather to reduce the amount of data to be analyzed. Data wrangling is a technique that involves transforming and cleaning data to make it suitable for analysis or visualization. Data wrangling does not help the analyst to acquire data from web pages, but rather to improve the quality and usability of the data.
ETL stands for Extract, Transform, and Load, which is a process that involves moving data from one or more sources to a destination, such as a data warehouse or a database. ETL does not help the analyst to acquire data from web pages, but rather to store and organize the data.

NEW QUESTION 120
Which of the following can be used to translate data into another form so it can only be read by a user who has a key or a password?

A. Data encryption.
B. Data transmission.
C. Data protection.
D. Data masking.

Answer: A

Explanation:

Data encryption can be used to translate data into another form so it can only be read by a user who has a key or a password. Data encryption is a process of transforming data using an algorithm or a cipher to make it unreadable to anyone except those who have the key or the password to decrypt it. Data encryption is a common method of protecting data from unauthorized access, modification, or theft. Reference: Guide to CompTIA Data+ and Practice Questions - Pass Your Cert

NEW QUESTION 122
An analyst has written the following code: SELECT *
FROM Cust_table
WHERE age > 60 AND City = "New York"
Which of the following criteria is the analyst retrieving?

A. All customers older than age 60 in New York state
B. All customers aged 60 and older in New York state
C. All customers older than age 60 in New York City
D. All customers younger than age 60 in New York City

Answer: C

Explanation:
The SQL query provided is selecting all records from the Cust_table where the age column has values greater than 60 and the City column matches ??New York??. The > operator selects values that are strictly greater than the comparison value, so it does not include customers aged exactly 60. The term ??New York?? in the context of a city database typically refers to New York City, not the state of New York. Therefore, the correct answer is that the analyst is retrieving data for all customers older than age 60 in New York City.
References:
? The use of the > operator in SQL is to select values greater than the specified value1.
? Understanding the WHERE clause in SQL and its use in filtering records based on specified conditions2.
? Clarification on the distinction between city and state names in database records3.

NEW QUESTION 126
Joseph is interpreting a left skewed distribution of test scores. Joe scored at the mean, Alfonso scored at the median, and gaby scored and the end of the tail. Who had the highest score?

A. Joseph
B. Joe
C. Alfonso
D. Gaby

Answer: C

Explanation:

Alfonso had the highest score. A left skewed distribution is a distribution where the tail is longer on the left side than on the right side, meaning that most of the values are clustered on the right side and there are some outliers on the left side. In a left skewed distribution, the mean is less than the median, which is less than the mode. Therefore, Joseph, who scored at the mean, had the lowest score, Gaby, who scored at the end of the tail, had the second lowest score, and Alfonso, who scored at the median, had the highest score. Reference: Skewness - Statistics How To

NEW QUESTION 129
The current date is July 14, 2020. A data analyst has been asked to create a report that shows the company??s year-over-year Q2 2020 sales. Which of the following reports should
the analyst compare?

A. A Q2 2020 and Q4 2019
B. YTD 2020 and YTD 2019
C. Q2 2020 and Q2 2019
D. Q2 2020 and Q2 2021

Answer: C

Explanation:
To create a report that shows the company??s year-over-year Q2 2020 sales, the analyst should compare the sales data from Q2 2020 and Q2 2019. Year-over-year (YoY) analysis is a method of comparing the performance of a business or a financial instrument over the same period in different years. It helps to identify trends, growth patterns, and seasonal fluctuations. Q2 refers to the second quarter of a year, which is usually from April to June. Therefore, the correct answer is C. References: YoY - Year over Year Analysis - Definition, Explanation & Examples, What is an Annual Sales Report: Definition, metrics, and tips - Snov.io

NEW QUESTION 133
Emma is working in a data warehouse and finds a finance fact table links to an organization dimension, which in turn links to a currency dimension that not linked to the fact table.
What type of design pattern is the data warehouse using?

A. Star.
B. Sun.
C. Snowflake.
D. Comet.

Answer: C

Explanation:
Correct answer C. Snowflake.
Since the dimension links to a dimension that isn't connected to the fact table, it must be a Snowflake, with a Star, all dimensions link directly to the fact table, Sun and Comet are not data warehouse design patterns.

NEW QUESTION 138
A data analyst is helping a retail store categorize its customers into five different groups based on the following information:
• How recently the customers made purchases
• How frequently the customers made purchases
• How much the customers spent Given the following information:

Customer_ID	Channel	Order_Date	Quantity	Territory	Amount ($)
1001	Online	2/11/2020	12	North	1,250
2001	Store	2/10/2020	31	East	5,000
4001	Online	2/09/2020	24	West	2,500
3001	Online	2/11/2020	51	South	6,000
1001	Store	3/10/2020	22	North	2,000
1001	Online	1/09/2020	87	North	8,400
1001	Store	2/09/2020	23	North	2,000

Which of the following would be most important for the analysis?

A. CustomerJ
B. Channel, Order_Date
C. CustomerJD, Territor
D. Amount
E. CustomerJD, Order_Dat
F. Amount
G. CustomerJ
H. Quantity, Amount

Answer: C

NEW QUESTION 142
An analyst has generated a report that includes the number of months in the first two quarters of 2019 when sales exceeded $50,000:

Month	Sales	Sales_indicator
January 2019	$52,005	Exceeded $50,000
February 2019	$48,687	Not exceeded $50,000
March 2019	$50,255	Exceeded $50,000
April 2019	$38,924	Not exceeded $50,000
June 2019	$57,076	Exceeded $50,000
July 2019	$51,035	Exceeded $50,000

Which of the following functions did the analyst use to generate the data in the Sales_indicator column?

A. Aggregate
B. Logical
C. Date
D. Sort

Answer: B

Explanation:
This is because a logical function is a type of function that returns a value based on a condition or a set of conditions. A logical function can be used to generate the data in the Sales_indicator column by comparing the values in the Sales column with a threshold of $50,000 and returning either ??Exceeded $50,000?? or ??Not exceeded $50,000?? accordingly. For example, a logical function in Excel that can achieve this is:

```
=IF(Sales>50000,"Exceeded $50,000","Not exceeded $50,000")
```

The other functions are not suitable for generating the data in the Sales_indicator column. Here is why:
Aggregate is a type of function that performs a calculation on a group of values, such as sum, average, count, etc. An aggregate function cannot generate the data in the Sales_indicator column because it does not compare the values in the Sales column with a threshold or return a text value based on a condition.
Date is a type of function that manipulates or extracts information from dates, such as year, month, day, etc. A date function cannot generate the data in the Sales_indicator column because it does not use the values in the Sales column or return a text value based on a condition.
Sort is a type of function that arranges the values in a column or a range in ascending or descending order. A sort function cannot generate the data in the Sales_indicator column because it does not create a new column or return a text value based on a condition.

NEW QUESTION 144
Consider this dataset showing the retirement age of 11 people, in whole years: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60
This tables show a simple frequency distribution of the retirement age data.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

A. 56

B. 55
C. 57
D. 54

Answer: D

Explanation:

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.
There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.
What is the mode?
The mode is the most commonly occurring value in a distribution.
The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

NEW QUESTION 149
A data analyst needs to perform a full outer join of a customer's orders using the tables below:

Sales_table

Cust_id	Order_id	Order_qty
Tc - 5858	Od - 9800	50
Tc - 5833	Od - 9801	68
Tc - 5890	Od - 9802	103

Order_table

Order_id	Order_qty
Od - 9803	102
Od - 9800	50
Od - 9802	103
Od - 9805	80
Od - 9804	70

Which of the following is the mean of the order quantity?

A. 73.5
B. 76.5
C. 78.8
D. 81.5

Answer: D

Explanation:
The correct answer is D. OUTER JOIN, seven rows.
An OUTER JOIN is a type of SQL join that returns all the rows from both tables, regardless of whether there is a match or not. If there is no match, the missing side will have null values. An OUTER JOIN can be either a LEFT JOIN, a RIGHT JOIN, or a FULL JOIN, depending on which table??s rows are preserved1
Using the example tables, a FULL OUTER JOIN query would look like this:
SELECT Cust_id, Order_id, Order_qty FROM Sales_table FULL OUTER JOIN Order_table ON Sales_table.Order_id = Order_table.Order_id;
The result of this query would be:
Cust_id | Order_id | Order_qty --------??---------??--------- 1 | 1 | 100 2 | 2 | 50 3 | 3 | 25 4 | 4 |
75 NULL | 5 | 10 NULL | 6 | 20 NULL | 7 | 15
As you can see, the query returns seven rows, one for each order in either table. The orders that are not in the Sales_table have null values for the Cust_id column.
To find the mean of the order quantity, we need to sum up the order quantities and divide by the number of rows. In this case, the mean is (100 + 50 + 25 + 75 + 10 + 20 + 15) / 7 = 42.14. Rounding to one decimal place, we get 42.1 as the mean of the order quantity.

NEW QUESTION 152
An analyst is reporting on the average income for a county and is reviewing the following data:

Name	Address	Yearly income
Jessica Jones	145 Stonebridge Avenue	$634,900
Spencer James	1567 Watercress	$135,000
Olivia Baker	456 Harvard Road	$95,000
Layla Harding	5674 Yarding Street	$37,000

Which of the following is the reason the analyst would need to cleanse the data in this data set?

A. Data completeness
B. Data outliers
C. Duplicate data
D. Missing values

Answer: B

NEW QUESTION 155
A report is scheduled to run and be distributed at the end of business each day. On Mondays, one of the recipients opens the previous week's reports and combines them to calculate the weekly totals and projections for the coming week. This is a tedious process, and the recipient asks an analyst for help. Which of the following should the analyst recommend?

A. Add calculation fields to the daily report so the totals are built in.
B. Create a new report with weekly totals set to run at the end of business on Friday.
C. Provide a daily summary to the report with totals to save the user the effort of manual calculations.
D. Reduce the frequency of the report to once a week and change the date range.

Answer: B

Explanation:
Creating a new report that automatically calculates weekly totals would streamline the process for the recipient. By setting this report to run at the end of business on Friday, it would provide the recipient with the necessary information for the entire week in one consolidated document. This eliminates the need for manual calculations and combines the previous week??s data into one report, making it more efficient and less time- consuming.
References:
? Best practices in business analytics suggest automating repetitive tasks and consolidating reports where possible to improve efficiency and reduce the potential for human error.

NEW QUESTION 159
A data analyst must fulfill a request for information that is needed weekly and should be automatically emailed to a specific set of users. Which of the following types of reports should the analyst recommend?

A. A self-service report
B. A research report
C. An ad hoc report
D. An operational report

Answer: D

Explanation:
An operational report is the most suitable type of report for information that needs to be sent out on a regular, scheduled basis, such as weekly. Operational reports are designed to provide ongoing insights into the performance of an organization??s operations and are typically automated to be distributed at set intervals. This automation can include scheduling the reports to be emailed to a specific list of recipients, making it an efficient solution for the analyst??s requirement.
Operational reports are often generated from data that is continuously updated, ensuring that the recipients receive the most current information at the time of the report??s distribution. This contrasts with ad hoc reports ©, which are usually created as needed and are not scheduled. Self-service reports (A) require users to generate the report themselves, which is not the requirement here. Research reports (B) are generally more detailed and are not typically used for regular operational updates.
References:
? The guidelines on writing email reports suggest that for regular, scheduled information dissemination, structured reports like operational reports are preferred1.
? Best practices in reporting also recommend automated and scheduled reports for consistent and timely updates, which operational reports provide2.

NEW QUESTION 161
A data analyst is creating a dashboard and trying to identify the type of information that should be included. Which of the following should the analyst consider first?

A. Data refresh rate
B. Consumer types
C. Access permissions
D. Data sources and attributes

Answer: D

Explanation:
The answer is D. Data sources and attributes.
Short Explanation: The data analyst should consider the data sources and attributes first when creating a dashboard, because they determine what kind of information can be
included and how it can be displayed. The data sources and attributes define the origin, quality, format, and structure of the data that will be used for the dashboard. They also affect the data refresh rate, the consumer types, and the access permissions of the dashboard12
* A. Data refresh rate is not the first thing to consider, because it depends on the data sources and attributes. The data refresh rate is how often the data in the dashboard is updated or refreshed to reflect the latest changes. The data refresh rate can vary depending on the type, frequency, and availability of the data sources1
* B. Consumer types are not the first thing to consider, because they depend on the data sources and attributes. The consumer types are the intended audiences or users of the dashboard, who may have different needs, preferences, and expectations for the dashboard. The consumer types can influence the design, layout, and functionality of the dashboard. However, the consumer types cannot be determined without knowing what kind of data is available and relevant for them1
* C. Access permissions are not the first thing to consider, because they depend on the data sources and attributes. The access permissions are the rules or policies that govern who can view, edit, or share the dashboard. The access permissions can protect the confidentiality, integrity, and availability of the data in the dashboard. However, the access permissions cannot be set without knowing what kind of data is involved and who needs to access it1

NEW QUESTION 162
While reviewing survey data, an analyst notices respondents entered ??Jan,?? ??January,?? and ??01?? as responses for the month of January. Which of the following steps should be taken to ensure data consistency?

A. Delete any of the responses that do not have ??January?? written out.
B. Replace any of the responses that have ??01??.
C. Filter on any of the responses that do not say ??January?? and update them to ??January??.
D. Sort any of the responses that say ??Jan?? and update them to ??01??.

Answer: C

Explanation:
Filter on any of the responses that do not say ??January?? and update them to ??January??. This is because filtering and updating are data cleansing techniques that can be used to ensure data consistency, which means that the data is uniform and follows a standard format. By filtering on any of the responses that do not say ??January?? and updating them to ??January??, the analyst can make sure that all the responses for the month of January are written in the same way. The other steps are not appropriate for ensuring data consistency. Here is why:
Deleting any of the responses that do not have ??January?? written out would result in data loss, which means that some information would be missing from the data set. This could affect the accuracy and reliability of the analysis.
Replacing any of the responses that have ??01?? would not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??Jan?? and ??January??. This could cause confusion and errors in the analysis. Sorting any of the responses that say ??Jan?? and updating them to ??01?? would also not solve the problem of data inconsistency, because there would still be two different ways of writing the month of January: ??01?? and ??January??. This could also cause confusion and errors in the analysis.

NEW QUESTION 165
Given the information in the following tables:

Online transactions:

Customer ID	Channel	Segment	Amount ($)
001	Online	Existing	3,000
002	Online	Existing	4,000
003	Online	New	1,500

In-store transactions:

Customer ID	Channel	Segment	Amount ($)
001	In-store	New	1,000
004	In-store	Existing	4,000
005	In-store	New	3,500

Which of the following describes merging these tables to create a master file that includes all transactions for both online and in-store sales?

A. Data audit
B. Data completeness
C. Data validation
D. Data consolidation

Answer: D

Explanation:
Merging tables to create a master file that includes all transactions for both online and in- store sales is best described as data consolidation. This process involves

combining data from various sources into a single, unified dataset. Data consolidation is essential for providing a comprehensive view of all transactions, which can be used for analysis, reporting, and decision-making purposes.
References: The answer is based on standard data management practices and the definition of data consolidation. No specific external documents were referenced for this response.

NEW QUESTION 170
A stakeholder wants to see daily sales targets organized in a dashboard by country, state, city, and ZIP Code. Which of the following delivery considerations must a data analyst take into account when creating the dashboard?

A. Variable formatting
B. Drill-down capability
C. Saved searches
D. Access permissions

Answer: B

NEW QUESTION 171
Which of the following data types best describe 4Ac1? (Select two).

A. Alphanumeric
B. Symbolic
C. Numeric
D. Float
E. Boolean
F. String

Answer: AF

Explanation:
The term ??4Ac1?? is a combination of numbers and letters, which fits the definition of an alphanumeric string. Alphanumeric refers to a character set that contains both letters and numbers. In data analytics and programming, such a value is typically treated as a string, which is a sequence of characters. Strings can include letters, digits, and various other symbols.
A numeric data type would only include numbers, and a float is a specific kind of numeric data type that includes decimal points, neither of which applies to ??4Ac1??. A boolean data
type represents one of two values: true or false. Since ??4Ac1?? does not represent a true or false value, it cannot be classified as boolean. Lastly, symbolic is not a standard data type in the context of programming and data analytics.
References:
? Understanding Python 3 data types1.
? Basic Data Types in Python2.
? Java Data Types3.

NEW QUESTION 172
Which one the following is not considered an aggregate function?

A. SUM
B. MIN
C. SELECT
D. MAX

Answer: C

Explanation:
The option that is not considered an aggregate function is SELECT. An aggregate function is a function that performs a calculation on a set of values and returns a single value. Examples of aggregate functions are SUM, MIN, MAX, AVG, COUNT, etc. SELECT is not an aggregate function, but a SQL command that is used to select data from a table or a query. Reference: SQL Aggregate Functions - W3Schools

NEW QUESTION 177
A data analyst needs to calculate the mean for Q1 sales using the data set below:

Product	Q1 sales
Ground beef	$2,667.60
Crab meet	$1,768.41
Swiss cheese	$3,182.40
Broccoli	$1,509.60
Vegetable spread	$3.202.87

Which of the following is the mean?

A. $2,466.18
B. $2,667.60
C. $3,082.72
D. $12,330.88

Answer: C

Explanation:
The mean is the average of all the values in a data set. To calculate the mean, we add up all the values and divide by the number of values. In this case, the mean for Q1 sales is ($2,000 + $3,000 + $4,000 + $2,500 + $3,500) / 5 = $3,082.72 References: CompTIA Data+ Certification Exam Objectives, page 9

NEW QUESTION 178
Under which of the following circumstances should the null hypothesis be accepted when a
= 0.05?

A. When p is 0.00003
B. When p is 0.001
C. When p is 0.04
D. When p is 0.06

Answer: C

Explanation:
The null hypothesis should be accepted when the p-value is greater than the alpha level, which is the significance level of the test. The p-value is the probability of obtaining a test statistic at least as extreme as the one observed in the sample, assuming that the null hypothesis is true. The alpha level is the probability of rejecting the null hypothesis when it is true, which is also known as a type I error12.
In this case, the alpha level is 0.05, which means that there is a 5% chance of rejecting the null hypothesis when it is true. Therefore, to reject the null hypothesis, the p-value must be less than or equal to 0.05, which indicates that the test statistic is very unlikely to occur by chance under the null hypothesis. Conversely, to accept the null hypothesis, the p-value must be greater than 0.05, which indicates that the test statistic is not very unlikely to occur by chance under the null hypothesis.
Among the four options, only option D has a p-value that is greater than 0.05 (p = 0.06). Therefore, option D is the correct answer. When p = 0.06, it means that there is a 6% chance of obtaining a test statistic at least as extreme as the one observed in the sample, assuming that the null hypothesis is true. This probability is not very low, and therefore does not provide enough evidence to reject the null hypothesis.

NEW QUESTION 183
A web developer wants to ensure that malicious users can't type SQL statements when they asked for input, like their username/userid.
Which of the following query optimization techniques would effectively prevent SQL Injection attacks?

A. Indexing.
B. Subset of records.
C. Temporary table in the query set.
D. Parametrization.

Answer: D

Explanation:

The correct answer is D: Parametrization. Parameterized SQL queries allow you to place parameters in an SQL query instead of a constant value. A parameter takes a value only when the query is executed, allowing the query to be reused with different values and
purposes. Parameterized SQL statements are available in some analysis clients, and are also available through the Historian SDK.
For example, you could create the following conditional SQL query, which contains a parameter for the collector's name: SELECT* FROM ExamsDigest WHERE coursename=? ORDER BY tagname SQL Injection is best prevented through the use of parameterized queries.

NEW QUESTION 188
Which of the following best describes how discrete data differs from continuous data?

A. Discrete data cannot create a sloped line.
B. Discrete data can only be a finite number of values.
C. Discrete data can have decimal points.
D. Discrete data applies only to numbers.

Answer: B

Explanation:
Discrete data are data that can only assume specific values that are countable and distinct. For example, the number of books, the number of heads in a coin toss, or the number of patients in a hospital are discrete data. Discrete data cannot have fractional or decimal values, and there are clear spaces between the possible values12. Continuous data are data that can assume any value within a range and can be meaningfully divided into smaller parts. For example, the weight, height, length, time, or temperature are continuous data. Continuous data can have fractional or decimal values, and there are infinite numbers of possible values between any two points12.

NEW QUESTION 191
A data analyst has been asked to create one table that has each employee's first name, last name, sales, and address. The sales and addresses are listed in the tables below:

Table 1

First name	Last name	Sales
John	Knox	$30
John	Johnson	$10
John	Sinclair	$70
Bob	Sinclair	$100

Table 2

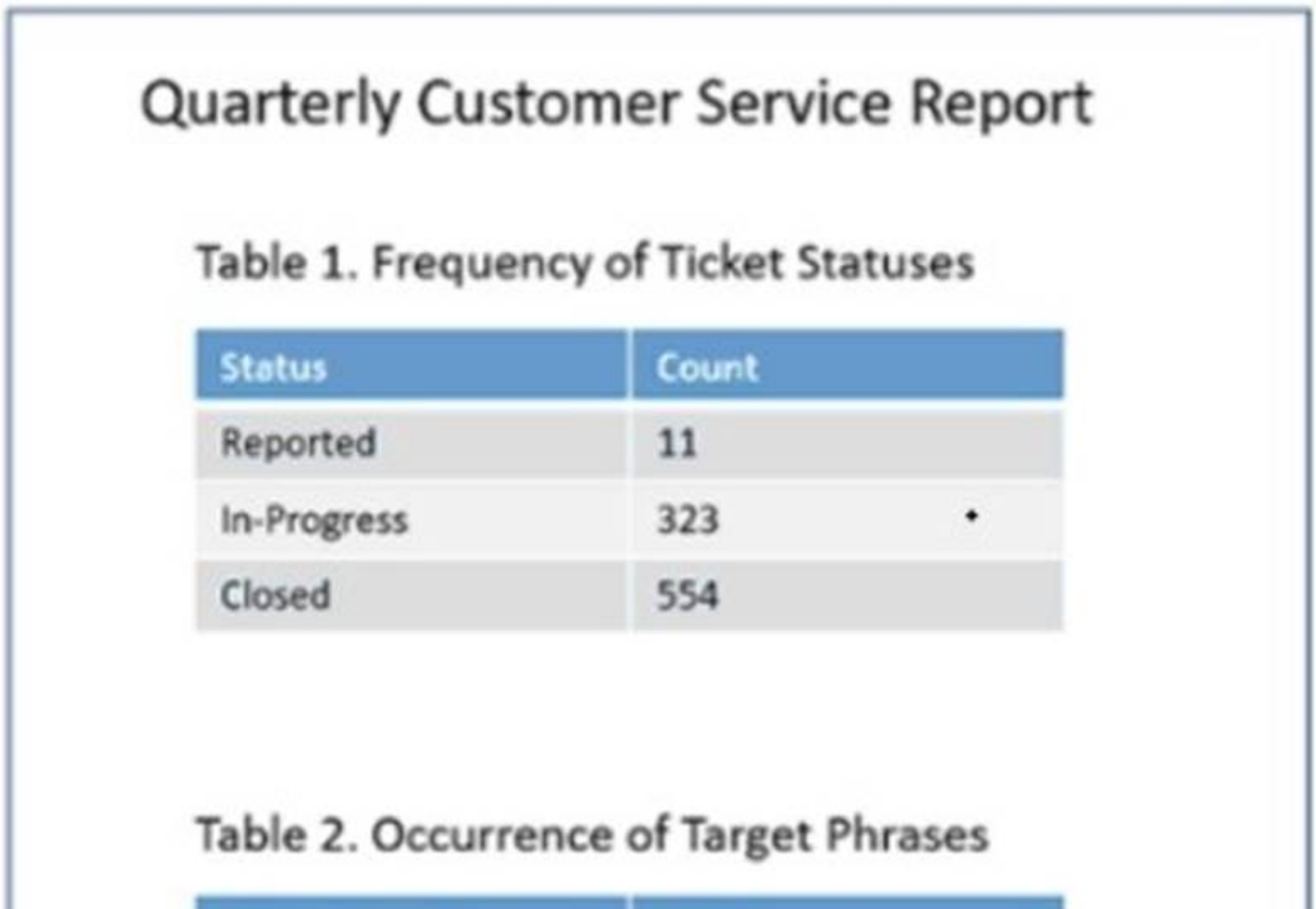
First name	Last name	Address
John	Knox	2851 N. Southport
John	Johnson	457 Bridle Ridge
John	Sinclair	1067 Windwood Lane
Bob	Sinclair	71 S. Wacker Drive

Which of the following steps should the analyst take to create the table?

A. Transpose the first name and last name in both table
B. Use lookup to pull the address field from Table 2 into Table 1.
C. Use lookup with the first name or first name to pull the address field from Table 2 into Table 1.
D. Use the append formula in both tables for the first name and last nam
E. Use lookup topull the address field from Table 2 into Table 1.
F. Create a column that concatenates the first name and last name in each tabl
G. Use concatenate and lookup to bring the address field into Table 1.

Answer: D

NEW QUESTION 193
Given the following report:



Quarterly Customer Service Report

Table 1. Frequency of Ticket Statuses

Status	Count
Reported	11
In-Progress	323
Closed	554

Table 2. Occurrence of Target Phrases

Target Phrases	Count
Have a great day!	1200
It is my pleasure to assist you.	70
Can you please hold?	7352

Most tickets are being addressed soon after being reported. Asking customers to hold is the most commonly used target phrase.

Which of the following components need to be added to ensure the report is point-in-time and static? (Choose two.)

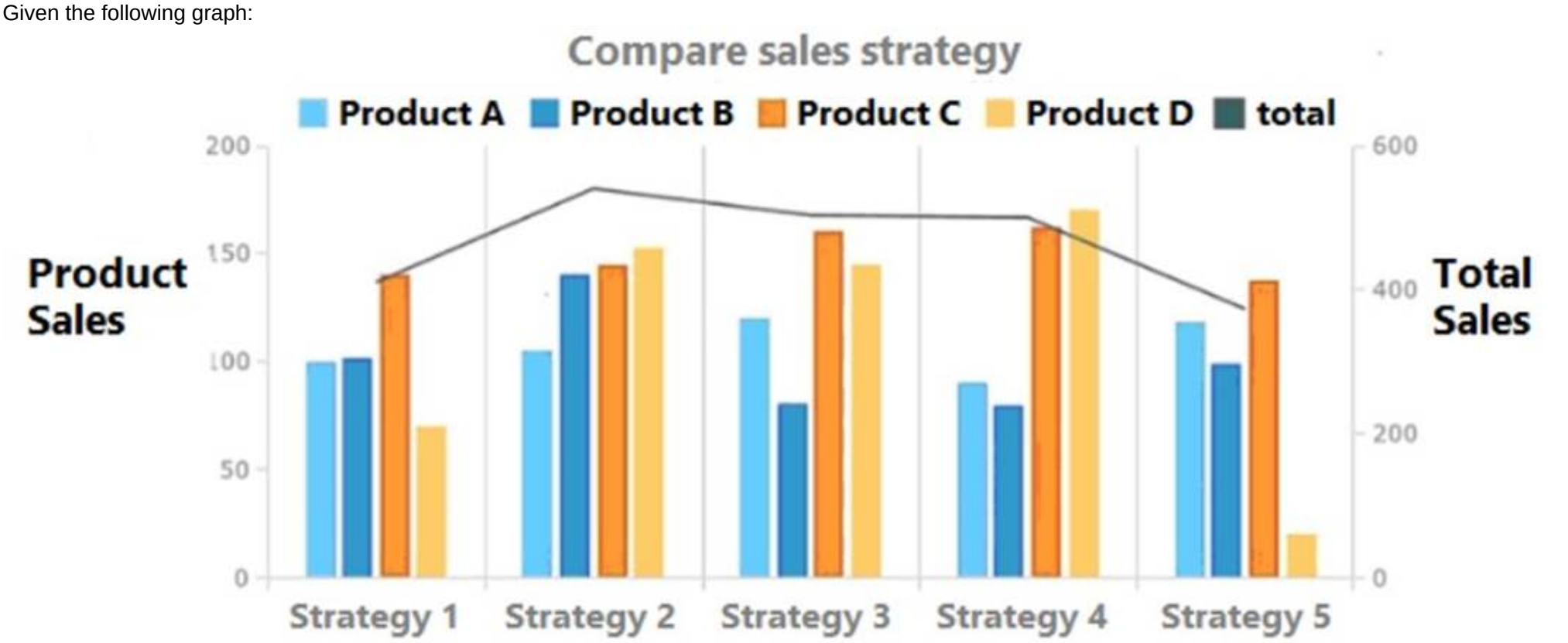
A. A control group for the phrases
B. A summary of the KPIs
C. Filter buttons for the status
D. The date when the report was last accessed
E. The time period the report covers
F. The date on which the report was run

Answer: E

Explanation:
The date on which the report was run. This is because the time period the report covers and the date on which the report was run are two components that need to be added to ensure the report is point-in-time and static, which means that the report shows the data as it was at a specific moment or interval in time, and does not change or update with new data. By adding the time period the report covers and the date on which the report was run, the analyst can indicate when and for how long the data was collected and analyzed, as well as avoid any confusion or ambiguity about the currency or validity of the data. The other components do not need to be added to ensure the report is point-in-time and static. Here is why:
A control group for the phrases is a type of group that serves as a baseline or a reference for comparison with another group that is exposed to some treatment or

intervention, such as a target phrase in this case. A control group for the phrases does not need to be added to ensure the report is point-in-time and static, because it does not affect the time frame or the stability of the data. However, a control group for the phrases could be useful for evaluating the effectiveness or impact of the target phrases on customer satisfaction or retention.
A summary of the KPIs is a type of document that provides an overview or a highlight of the key performance indicators (KPIs), which are measurable values that indicate how well an organization or a process is achieving its goals or objectives. A summary of the KPIs does not need to be added to ensure the report is point-in-time and static, because it does not affect the time frame or the stability of the data. However, a summary of the KPIs could be useful for communicating or presenting the main findings or insights from the report.
Filter buttons for the status are a type of feature or function that allows users to select or deselect certain values or categories in a column or a table, such as ticket statuses in this case. Filter buttons for the status do not need to be added to ensure the report is point-in- time and static, because they do not affect the time frame or the stability of the data. However, filter buttons for the status could be useful for exploring or analyzing different aspects or segments of the data.

NEW QUESTION 197
Given the following graph:



Which of the following summary statements upholds integrity in data reporting?

A. Sales are approximately equal for Product A and Product B across all strategies.
B. Strategy 4 provides the best sales in comparison to other strategies.
C. While Strategy 2 does not result in the highest sales of Product D, over all products it appears to be the most effective.
D. Product D should be promoted more than the other products in all strategies.

Answer: B

Explanation:
Strategy 4 provides the best sales in comparison to other strategies. This is because the total sales for Strategy 4 are the highest among all the strategies, as shown by the black line. The other statements are not accurate or do not uphold integrity in data reporting. Here is why:
Statement A is false because sales are not approximately equal for Product A and Product B across all strategies. For example, in Strategy 1, Product A has more sales than Product B, while in Strategy 3, Product B has more sales than Product A.
Statement C is misleading because it does not account for the difference in scale between the products. While Strategy 2 has the highest total sales among all products, it does not necessarily mean that it is the most effective for each product. For instance, Product D has very low sales in Strategy 2 compared to other strategies.
Statement D is biased because it does not provide any evidence or justification for why Product D should be promoted more than the other products in all strategies. It also ignores the fact that Product D has the lowest sales among all products in most of the strategies.

NEW QUESTION 200
An analyst wants to combine two data sets into a single spreadsheet. Column names from
the first spreadsheet are listed in rows in the second spreadsheet. Which of the following is the first step the analyst should take to combine the data sets?

A. Blend
B. Merge
C. Concatenate
D. Transpose

Answer: C

NEW QUESTION 204
Five dogs have the following heights in millimeters: 300, 430, 170, 470, 600
Which of the following is the mean height for the five dogs?

A. 394mm
B. 405mm
C. 493mm
D. 504mm

Answer: B

Explanation:

The mean height for the five dogs is 405mm. The mean, or average, is a measure of central tendency that represents the sum of all values divided by the number of values. To calculate the mean height for the five dogs, we can use the following formula: Mean = (300 + 430 + 170 + 470 + 600) / 5 = 2020 / 5 = 404
We can round up the result to the nearest millimeter, which is 405mm. The other options are not correct, as they are either too high or too low than the actual mean. Reference: [Mean - Math is Fun]

NEW QUESTION 209
Given the below:

		Conclusion from statistical analysis	
		Accept the null hypothesis	Reject the null hypothesis
The true state of nature	Null hypothesis is true	1	3
	Null hypothesis is false	2	4

Which of the following numbers represents a Type I error?

A. 1
B. 2
C. 3
D. 4

Answer: C

NEW QUESTION 213
Which of the following types of analysis is used when comparing last week's sales to the previous week's sales?

A. Trend analysis
B. Exploratory analysis
C. Prescriptive analysis
D. Link analysis

Answer: A

NEW QUESTION 214
An analyst needs to provide a chart to identify the composition between the categories of the survey response data set:

Favorite color	Responses
Red	15
Blue	35
Green	25
Yellow	25
Total	100

Which of the following charts would be BEST to use?

A. Histogram
B. Pie
C. Line
D. Scatter pot
E. Waterfall

Answer: B

Explanation:
A pie chart is the best choice to show the composition between the categories of the survey response data set. A pie chart represents the whole with a circle, divided by slices into parts. Each slice shows the relative size of each category as a percentage of the total. A pie chart is useful when the categories are mutually exclusive and add up to 100%. The table shows the favorite color and the number of responses for each color, which can be easily converted into percentages. A pie chart can show how each color contributes to the total number of responses.
Option A is incorrect because a histogram is used to show how data points are distributed along a numerical scale. The survey response data set is not numerical,

but categorical. Option C is incorrect because a line chart is used to show trends or changes over time. The survey response data set does not have a time dimension.
Option D is incorrect because a scatter plot is used to show the relationship between two numerical variables. The survey response data set does not have two numerical variables. Option E is incorrect because a waterfall chart is used to show how an initial value is increased or decreased by a series of intermediate values. The survey response data set does not have an initial value or intermediate values.
References:
? How to Choose the Right Chart for Your Data - Infogram
? How to Choose the Right Data Visualization | Tutorial by Chartio
? Find the Best Visualizations for Your Metrics - The Data School
? How to choose the best chart or graph for your data

NEW QUESTION 219
A company wants to know how its customers interact with an e-commerce website based on clicks over items. Which of the following is the primary requirement for this report?

A. Data content
B. Frequency
C. Filtering
D. Views

Answer: B

NEW QUESTION 221
Daniel is using the structured Query language to work with data stored in relational database.
He would like to add several new rows to a database table. What command should he use?

A. SELECT.
B. ALTER.
C. INSERT.
D. UPDATE.

Answer: C

Explanation:
INSERT
The INSERT command is used to add new records to a database table.
The SELECT command is used to retrieve information from a database. It's the most commonly used command in SQL because it is used to pose queries to the database and retrieve the data that you're interested in working with.
The UPDATE command is used to modify rows in the database.
The CREATE command is used to create a new table within your database or a new database on your server.

NEW QUESTION 224
Each month an analyst needs to execute a data pull for the two prior months. Which of the following is the most efficient function for the analyst to use?

A. Logical
B. Date
C. Aggregate
D. System

Answer: B

Explanation:
The most efficient function for an analyst to execute a data pull for the two prior months would be the Date function. This function allows for the manipulation and formatting of date values within a database. Using Date functions, an analyst can dynamically calculate the start and end dates for the previous two months, ensuring that the data pull is accurate and automated without manual intervention.
For example, SQL functions like DATEADD and DATEDIFF can be used to determine the exact range of dates needed for the data pull. These functions can calculate the first and
last day of the previous months relative to the current date, which is essential for monthly reporting and analysis.
References:
? Discussions on Stack Overflow suggest using SQL date functions
like DATEADD and DATEDIFF to dynamically extract data for previous months, which supports the use of Date functions12.
? The use of Date functions is also recommended for ensuring that the data pull is
not only efficient but also accurate, as it avoids potential errors associated with manual date entry3.

NEW QUESTION 229
A data analyst has been asked to organize the table below in the following ways: By sales from high to low -
By state in alphabetic order -

First_name	Last_name	Address	City	State	Sales
Ed	Edens	2851 N. Southport	Chicago	IL	$125,689
Pat	Mudd	710 Bridle Ridge Road	Eagan	MN	$101,259
Katie	Hofstad	2851 S. Windwood Lane	Rosemount	NY	$105,779
Edward	Frank	281 S. Northport	Chicago	IL	$456,231
Rachel	Newman	305 Big Timber Trail	Wheaton	CO	$99,876
Kaylyn	Korth	332 Richfield Drive	Lakeview	MN	$166,874

Which of the following functions will allow the data analyst to organize the table in this manner?

A. Conditional formatting
B. Grouping
C. Filtering
D. Sorting

Answer: D

Explanation:
Sorting is the function that will allow the data analyst to organize the table in the desired manner. Sorting means arranging the data in a specific order, such as ascending or descending, based on one or more criteria. Sorting can be applied to any column in the table, such as sales or state. References: CompTIA Data+ Certification Exam Objectives, page 11

NEW QUESTION 232
Which of the following is the most likely reason for a data analyst to optimize a query using parameterization?

A. To return a subset of records
B. To insert a temporary table
C. To prevent SQL injections
D. To increase the query speed

Answer: C

Explanation:
Parameterization in SQL queries is a technique used to prevent SQL injection, which is a common security vulnerability that allows an attacker to interfere with the queries that an application makes to its database. By using parameterized queries, the database can distinguish between code and data, regardless of the input received. This method ensures that an attacker cannot change the intent of a query, even if SQL commands are inserted by the attacker. While parameterization can also affect performance by enabling consistent query execution plans, its primary purpose is to enhance security.
References:
? Medium article on SQL Query Optimization1.
? MSSQLTips on SQL Query Performance2.
? Blog post on SQL Performance Optimization3.
? SQL Easy guide on improving SQL Query Performance4.
? LearnSQL.com on SQL for Data Analysis5.

NEW QUESTION 237
An analyst wants to extract data from a variety of sources and store the data in a cloud- based environment prior to cleaning. Which of the following integration techniques should the analyst use?

A. ETL
B. API
C. SQL
D. ELT

Answer: A

NEW QUESTION 242
A data analyst needs to write a SOL query measuring last month's website visits and distribute a summary report to the marketing team. Which of the following is the analyst creating?

A. Date range
B. Distribution list
C. Data content
D. Report view

Answer: D

NEW QUESTION 245
A data engineer is creating a database field to capture whether a customer likes vanilla ice cream. Which of the following data types is the best to capture this information?

A. Integer
B. Boolean

C. Categorical
D. Numeric

Answer: B

NEW QUESTION 249
Which one of the following values will appear first if they are sorted in descending order?

A. Aaron.
B. Molly.
C. Xavier.
D. Adam.

Answer: C

Explanation:
The value that will appear first if they are sorted in descending order is Xavier. Descending order means arranging values from the largest to the smallest, or from the last to the first in alphabetical order. In this case, Xavier is the last name in alphabetical order, so it will appear first when sorted in descending order. The other names will appear in the following order: Molly, Adam, Aaron. Reference: Sorting Data - W3Schools

NEW QUESTION 253
The duration of a phone call in milliseconds is an example of:

A. ordinal data.
B. nominal data.
C. boolean data.
D. continuous data.

Answer: D

Explanation:
The correct answer is D. Continuous data.
Continuous data is a type of quantitative data that can take any value within a range and can be measured with infinite precision. Continuous data can be expressed as fractions, decimals, or percentages. Examples of continuous data are height, weight, temperature, time, speed, etc12
The duration of a phone call in milliseconds is an example of continuous data, because it can take any value within a range (from zero to infinity) and can be measured with infinite precision (up to milliseconds or even smaller units). The duration of a phone call in milliseconds can also be expressed as fractions, decimals, or percentages of a larger unit (such as seconds, minutes, or hours).
Ordinal data is not correct, because ordinal data is a type of qualitative or categorical data that can be ordered or ranked according to some criterion. Ordinal data can have a logical order, but the intervals between the values are not equal or meaningful. Examples of ordinal data are grades, ratings, ranks, etc12
Nominal data is not correct, because nominal data is a type of qualitative or categorical data that can be labeled or named without any order or ranking. Nominal data can have a finite number of categories or classes, but the categories have no intrinsic value or hierarchy. Examples of nominal data are gender, color, nationality, etc12
Boolean data is not correct, because boolean data is a type of binary data that can have only two possible values: true or false. Boolean data can be used to represent logical statements, conditions, or outcomes. Examples of boolean data are yes/no, on/off, 1/0, etc.

NEW QUESTION 257
Andy is a pricing analyst for a retailer. Using a hypothesis test, he wants to assess whether people who receive electronic coupons spend more on average. What should Andy's null hypothesis be?

A. People who receive electronic coupons spend more on average.
B. People who receive electronic coupons spend less on average.
C. People who receive electronic coupons do not spend more on average.
D. People who do not receive electronic coupons spend more on average.

Answer: C

Explanation:

The null hypothesis presumes the status quo. Andy is testing whether or not people who receive an electronic coupon spend more on average, so, the null hypothesis states that people who receive the coupon do spend more on average.

NEW QUESTION 259
A data analyst is designing a dashboard that will provide a story of sales and determine which site is providing the highest sales volume per customer. The analyst must choose an appropriate chart to include in the dashboard. The following data is available:

Site	Customers	Sales volume	Average sales per customer
A1	2236	$3,415,372.00	$1,527.45
A2	885	$1,405,437.00	$1,588.06
A3	333	$952,723.00	$2,861.03
B1	483	$4,871,380.00	$10,085.67
B2	2969	$780,381.00	$262.84
B4	2357	$4,917,436.00	$2,086.31
C1	1524	$1,135,204.00	$744.88
C2	878	$614,964.00	$700.41
C3	1925	$4,035,100.00	$2,096.16

Which of the following types of charts should be considered?

A. Include a line chart using the site and average sales per customer.
B. Include a pie chart using the site and sales to average sales per customer.
C. Include a scatter chart using sales volume and average sales per customer.
D. Include a column chart using the site and sales to average sales per customer.

Answer: C

Explanation:
A scatter chart using sales volume and average sales per customer is the best type of chart to include in the dashboard. A scatter chart is a type of chart that displays the relationship between two numerical variables using dots or markers. A scatter chart can show how one variable affects another, how strong the correlation is between them, and how the data points are distributed. In this case, a scatter chart can show the story of sales and determine which site is providing the highest sales volume per customer by plotting the sales volume on the x-axis and the average sales per customer on the y-axis. Each dot on the chart will represent a site, and the analyst can easily compare the sites based on their position on the chart. A site with a high sales volume and a high average sales per customer will be in the upper right quadrant, indicating a high performance. A site with a low sales volume and a low average sales per customer will be in the lower left quadrant, indicating a low performance. A site with a high sales volume and a low average sales per customer will be in the lower right quadrant, indicating a high volume but low value. A site with a low sales volume and a high average sales per customer will be in the upper left quadrant, indicating a low volume but high value. A scatter chart can also show if there is a positive or negative correlation between the two variables, or if there is no correlation at all. A positive correlation means that as one variable increases, so does the other. A negative correlation means that as one variable increases, the other decreases. No correlation means that there is no relationship between the two variables.
The other types of charts are not as suitable for this purpose. A line chart is a type of chart that displays the change of one or more variables over time using lines. A line chart can show trends, patterns, and fluctuations in the data. However, in this case, there is no time variable involved, so a line chart would not be appropriate. A pie chart is a type of chart that displays the proportion of each category in a whole using slices of a circle. A pie chart can show how each category contributes to the total and compare the relative sizes of each category. However, in this case, there are two numerical variables involved, so a pie chart would not be able to show their relationship. A column chart is a type of chart that displays the comparison of one or more variables across categories using vertical bars. A column chart can show how each category differs from each other and rank them by size. However, in this case, a column chart would not be able to show the relationship between sales volume and average sales per customer, as it would only show one variable for each site.

NEW QUESTION 261
An analyst is preparing a report that contains weather data. The temperatures are shown in Fahrenheit. but they must be reported in Celsius. Which of the following should the analyst do to fix this issue?

A. Normalize the data.
B. Standardize the data.
C. Rescale the data.
D. Aggregate the data.

Answer: C

Explanation:
The analyst should rescale the data to fix this issue. Rescaling is a process of transforming data from one scale to another, such as changing the units of measurement. In this case, the analyst needs to rescale the temperatures from Fahrenheit to Celsius, which are two different scales for measuring temperature. To do this, the analyst can use the following formula:
Celsius = (Fahrenheit - 32) * 5/9
This formula converts each temperature value from Fahrenheit to Celsius by subtracting 32
and multiplying by 5/9. For example, if the temperature is 68??F, the rescaled value in Celsius is:
Celsius = (68 - 32) * 5/9 Celsius = 20??C
Rescaling the data can help the analyst to report the temperatures in a consistent and accurate way, and to avoid any confusion or errors that may arise from using different scales. Rescaling can also make the data more comparable and compatible with other data sources or standards that use the same scale12.

NEW QUESTION 263
Which of the following is the best description of discrete data types?

A. Non-numeric data used to describe attributes of a population sample
B. The frequency of the number of times each value occurs by using whole numbers

C. Numeric values that can be measured on a continuous scale
D. Non-numeric data used to describe attributes of a population sample ranked in a specific order

Answer: B

NEW QUESTION 267
An organization would like to add a secondary email field to its customer database in order to enrich the customer profiles. Which of the following data manipulation techniques should the analyst use to add this information?

A. Blend
B. Merge
C. Append
D. Aggregate

Answer: C

NEW QUESTION 271
Which of the following is the correct data type for text?

A. Boolean
B. String
C. Integer
D. Float

Answer: B

Explanation:
The correct data type for text is string. A string is a data type that represents a sequence of characters, such as letters, numbers, symbols, or spaces. A string can be enclosed by single quotes (?? ') or double quotes (" ") in most programming languages. For example, ??Hello??, ??World??, and ??123?? are all strings. The other options are not data types for text, but for other kinds of values. A boolean is a data type that represents a logical value, either true or false. An integer is a data type that represents a whole number, such as 1, 0, or -5. A float is a data type that represents a number with a fractional part, such as 3.14, 0.5, or -2.7.
Reference: Data Types - W3Schools

NEW QUESTION 275
A data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a review of the business??s performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. Which of the following report types should the data analyst create?

A. Static
B. Real-time
C. Self-service
D. Dynamic

Answer: A

Explanation:
A dynamic report is a type of report that shows data that changes or updates automatically based on certain criteria or parameters. A dynamic report can allow users to interact with the data, filter it, drill down into it, or visualize it in different ways. A dynamic report is suitable for situations where the data changes frequently or where real-time or near-real-time data is needed for decision making or analysis. In this case, the data analyst is asked to create a sales report for the second-quarter 2020 board meeting, which will include a review of the business??s performance through the second quarter. The board meeting will be held on July 15, 2020, after the numbers are finalized. This means that the data analyst does not need to show real-time or dynamic data, but rather a fixed and accurate view of the sales data for the second quarter. Therefore, a static report would be the best way to meet this stakeholder requirement. Therefore, the correct answer is A.
References: [What are Dynamic Reports? | Sisense], Static vs Dynamic Reports - What??s The Difference? | datapine

NEW QUESTION 277
Which of the following data protection methods provides confidentiality for data in transit?

A. De-identification
B. Encryption
C. Masking
D. Anonymization

Answer: B

NEW QUESTION 281
A data analyst is performing a data merge within a spreadsheet using the tables below:
https://www.bing.com/images/blob?bcid=S1XCF9p02M4GjpbGxHj0lrIaj9sw.....4c

Table 1

Last name	Sales
Knox	$30
Johnson	$10
Sinclair	$70

Table 2

Last name	Address
Knox	2851 N. Southport
Johnson	467 Bridle Ridge
Sinclair	1067 Windwood Lane

The analyst is attempting to pull the addresses from Table 2 into Table 1 using the last names and is receiving an error message. Which of the following steps can the analyst perform to fix the error?

A. Use concatenate to combine the tables.
B. Ensure the formula is pulling from right to left.
C. Sort the data by the last name field.
D. Review the spelling and data type.

Answer: D

Explanation:
The error in merging data from Table 2 into Table 1 using last names could be due to discrepancies in spelling or data type between the two tables. It is essential to ensure that the last names are spelled consistently and that the data types are compatible for a successful merge. Option D suggests reviewing these aspects, which can potentially resolve the error, ensuring that each last name in Table 1 accurately corresponds to the same last name in Table 2, allowing for a successful data pull of addresses.
References: This answer is based on general data analytics practices and does not reference a specific document.

NEW QUESTION 284
A research analyst wants to determine whether the data being analyzed is connected to other datapoints. Which of the following is the BEST type of analysis to conduct?

A. Trend analysis
B. Performance analysis
C. Link analysis
D. Exploratory analysis

Answer: C

Explanation:
This is because link analysis is a type of analysis that determines whether the data being analyzed is connected to other datapoints, such as entities, events, or relationships. Link analysis can be used to identify and visualize the patterns, networks, or associations among the datapoints, as well as measure the strength, direction, or frequency of the connections. For example, link analysis can be used to determine if there is a connection between a customer??s purchase history and their loyalty program status. The other types of analysis are not the best types of analysis to conduct to determine whether the data being analyzed is connected to other datapoints. Here is why:
? Trend analysis is a type of analysis that determines whether the data being analyzed is changing over time, such as increasing, decreasing, or fluctuating. Trend analysis can be used to identify and visualize the patterns, cycles, or movements in the data points, as well as measure the rate, direction, or magnitude of the changes. For example, trend analysis can be used to determine if there is a change in a company??s sales revenue over a period of time.
? Performance analysis is a type of analysis that determines whether the data being
analyzed is meeting certain goals or objectives, such as targets, benchmarks, or standards. Performance analysis can be used to identify and visualize the gaps, deviations, or variations in the data points, as well as measure the efficiency, effectiveness, or quality of the outcomes. For example, performance analysis can be used to determine if there is a gap between a student??s test score and their expected score based on their previous performance.
? Exploratory analysis is a type of analysis that determines whether there are any insights or discoveries in the data being analyzed, such as patterns, relationships, or anomalies. Exploratory analysis can be used to identify and visualize the characteristics, features, or behaviors of the data points, as well as measure their distribution, frequency, or correlation. For example, exploratory analysis can be used to determine if there are any outliers or unusual values in a dataset.

NEW QUESTION 286
An analyst is working with a data set that lists individuals' first and last names in separate columns. Which of the following processes should the analyst use to combine the first and last names into a single spreadsheet cell?

A. Transpose
B. Blend
C. Concatenate
D. Merges

Answer: C

NEW QUESTION 288
Which of the following query optimization techniques involves examining only the data that is needed for a particular task?

A. Making a temporary table
B. Creating a flat file
C. Indexing documents
D. Creating an execution plan

Answer: C

Explanation:
The correct answer is C. Indexing documents.
Indexing documents is a query optimization technique that involves creating a data structure that allows faster access to the data in the documents. Indexing documents can reduce the amount of data that needs to be scanned for a particular query, thus improving the performance and efficiency of the query. Indexing documents can also help with searching, sorting, filtering, and aggregating the data in the documents12

NEW QUESTION 292
Which of the following would a data analyst look for first if 100% participation is needed on survey results?

A. Missing data
B. Invalid data
C. Redundant data
D. Duplicate data

Answer: A

Explanation:
Missing data is a type of data quality issue that occurs when some values in a data set are not recorded or available. Missing data can affect the validity and reliability of survey results, especially if the missing values are not random or ignorable. Missing data can also reduce the sample size and the statistical power of the analysis12
If 100% participation is needed on survey results, a data analyst would look for missing data first, because missing data would indicate that some participants did not complete or submit the survey, or that some responses were not recorded or transmitted correctly. A data analyst would need to identify the causes and patterns of missing data, and apply appropriate methods to handle or prevent missing data, such as imputation, deletion, weighting, or follow-up12

NEW QUESTION 297
Which one of the following would not normally be considered a summary statistic?

A. z-score.
B. Mean.
C. Variance.
D. Standard deviation.

Answer: A

Explanation:

Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is. A z-score can be placed on a normal distribution curve.

NEW QUESTION 298
An analyst in a consumer bank department wants to showcase the concentration of accounts opened in the United States by ZIP Code to describe the effectiveness of the bank's marketing campaigns. Which of the following would be the best way to visualize the data?

A. A stacked chart
B. A tree map
C. A waterfall chart
D. A geographic map

Answer: D

NEW QUESTION 302
The ACME Corporation hired an analyst to detect data quality issues in their Excel documents. Which of the following are the most common issues? (Select TWO)

A. Apostrophe.
B. Commas.
C. Symbols.
D. Duplicates.
E. Misspellings.

Answer: DE

Explanation:

* 1. Duplicates
* 2. Misspellings
The most common data quality issues are difficult to resolve in Excel because of their rigidity. It forces analysts to do a ton of manual work, which results in a high probability of an error being introduced to the data set. Those common issues include:
- Blanks
- Nulls
- Outliers
- Duplicates
- Extra spaces

- Misspellings
- Abbreviations and domain-specific variations
- Formula error codes

When introduced, these errors can skew or even invalidate the resulting analysis. A smart tool would minimize the possibility of error by automating the manual work. In Excel, you might look for data quality issues in one of two ways. First, you might use auto filters on specific columns to scan for anomalies and blanks or you might use a pivot table to find gaps and discrepancies.
In either case, you're scanning for the anomalies yourself. Suffice it to say that's not a very efficient process. It also means accuracy is only as good as the analyst's eye, so the probability of error varies throughout the day.

NEW QUESTION 307
Which of the following best describes a business analytics tool with interactive visualization and business capabilities and an interface that is simple enough for end users to create their own reports and dashboards?

A. Python
B. R
C. Microsoft Power Bl
D. SAS

Answer: C

Explanation:
The best answer is C. Microsoft Power BI.
Microsoft Power BI is a business analytics and business intelligence service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards. Power BI can connect to multiple data sources, clean and transform data, create custom calculations, and visualize data through charts, graphs, and tables. Power BI can be accessed through a web browser, mobile device, or desktop application and integrated with other Microsoft tools like Excel and SharePoint12
Python is not correct, because Python is a general-purpose programming language that can be used for various applications, including data analysis and visualization. However, Python is not a dedicated business analytics tool, and it requires coding or programming skills to create reports and dashboards.
R is not correct, because R is a programming language and software environment for statistical computing and graphics. R can be used for data analysis and visualization, but it is not a specialized business analytics tool, and it requires coding or programming skills to create reports and dashboards.
SAS is not correct, because SAS is a software suite for advanced analytics, business intelligence, data management, and predictive analytics. SAS can provide interactive visualizations and business capabilities, but it does not have an interface that is simple enough for end users to create their own reports and dashboards. SAS also requires coding or programming skills to use its features.

NEW QUESTION 310
Which of the following is the first step an analyst should perform upon receiving a business request for analysis?

A. Determine the data needs and sources for analysis.
B. Initiate the analysis for exploratory data analysis.
C. Review the business questions to understand the scope.
D. Finalize the methodology to solve the problem.

Answer: C

Explanation:
Answer C. Review the business questions to understand the scope.
The first step an analyst should perform upon receiving a business request for analysis is to review the business questions to understand the scope of the problem, the objectives, and the expected outcomes. This will help the analyst to define the analytical approach, identify the data needs and sources, and plan the analysis process. Reviewing the business questions will also help the analyst to communicate with the stakeholders and clarify any assumptions or ambiguities1.
Option A is incorrect, as determining the data needs and sources for analysis is not the first step, but rather a subsequent step that depends on the business questions and the analytical approach.
Option B is incorrect, as initiating the analysis for exploratory data analysis is not the first step, but rather a part of the analysis process that involves examining and summarizing the data, identifying patterns and outliers, and testing hypotheses.
Option D is incorrect, as finalizing the methodology to solve the problem is not the first step, but rather a later step that involves selecting and applying the appropriate analytical techniques, tools, and models to answer the business questions.

NEW QUESTION 313
You are working with a dataset and need to swap the values in rows with those in columns. What action do you need to perform?

A. Recording
B. Filtering.
C. Aggregation.
D. Transposition.

Answer: D

Explanation:

Transpose creates a new data file in which the rows and columns in the original data file are transposed so that cases (rows) become variables and variables (columns) become
cases. Transpose automatically creates new variable names and displays a list of the new variable names.
Transposing data is useful for data analysis. At times, we have to pull data from various files with different formats for analysis and preparing reports. In such circumstances, we may have to transpose some data from one file to the other. In excel, we can transpose data in multiple ways.

NEW QUESTION 316
Which of the following statements would be used to append two tables that have the same number of columns?

A. UNION ALL
B. MERGE

C. GROUP BY
D. JOIN

Answer: A

Explanation:
The correct answer is A. UNION ALL.
UNION ALL is a SQL statement that appends two tables that have the same number of columns and compatible data types. UNION ALL preserves all the rows from both tables, including any duplicates12
* B. MERGE is not correct, because MERGE is a SQL statement that combines the data of two tables based on a common column. MERGE can perform insert, update, or delete operations on the target table depending on the matching or non-matching rows from the source table34
* C. GROUP BY is not correct, because GROUP BY is a SQL clause that groups the rows of a table based on one or more columns. GROUP BY is often used with aggregate functions, such as SUM, AVG, COUNT, etc., to calculate summary statistics for each group56
* D. JOIN is not correct, because JOIN is a SQL clause that combines the data of two tables based on a common column or condition. JOIN can produce different results depending on the type of join, such as INNER JOIN, LEFT JOIN, RIGHT JOIN, etc.

NEW QUESTION 320
.....

THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual DA0-001 Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the DA0-001 Product From:

https://www.2passeasy.com/dumps/DA0-001/

Money Back Guarantee

DA0-001 Practice Exam Features:

* DA0-001 Questions and Answers Updated Frequently

* DA0-001 Practice Questions Verified by Expert Senior Certified Staff

* DA0-001 Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* DA0-001 Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year