ATMIC NETWORKS
YOUR GUIDE TO GLOBAL CERTIFIATION

"Please note that these files may not be up to date. However, the questions will help you understand the exam format and typical question patterns."

www.atmicnetworks .com

Warning: Keep connected with our support team for latest updates

Question: 1

SIMULATION

A client has gathered weather data on which regions have high temperatures. The client would like a visualization to gain a better understanding of the data.

INSTRUCTIONS

Part 1

Review the charts provided and use the drop-down menu to select the most appropriate way to standardize the data.

Part 2

Answer the questions to determine how to create one data set.

Part 3

Select the most appropriate visualization based on the data set that represents what the client is looking for.

If at any time you would like to bring back the initial state of the simulation, please click the Reset All button.

Part 1 Part 2 Part 3

Standardize data

Select table ᵛ] +

Table 1

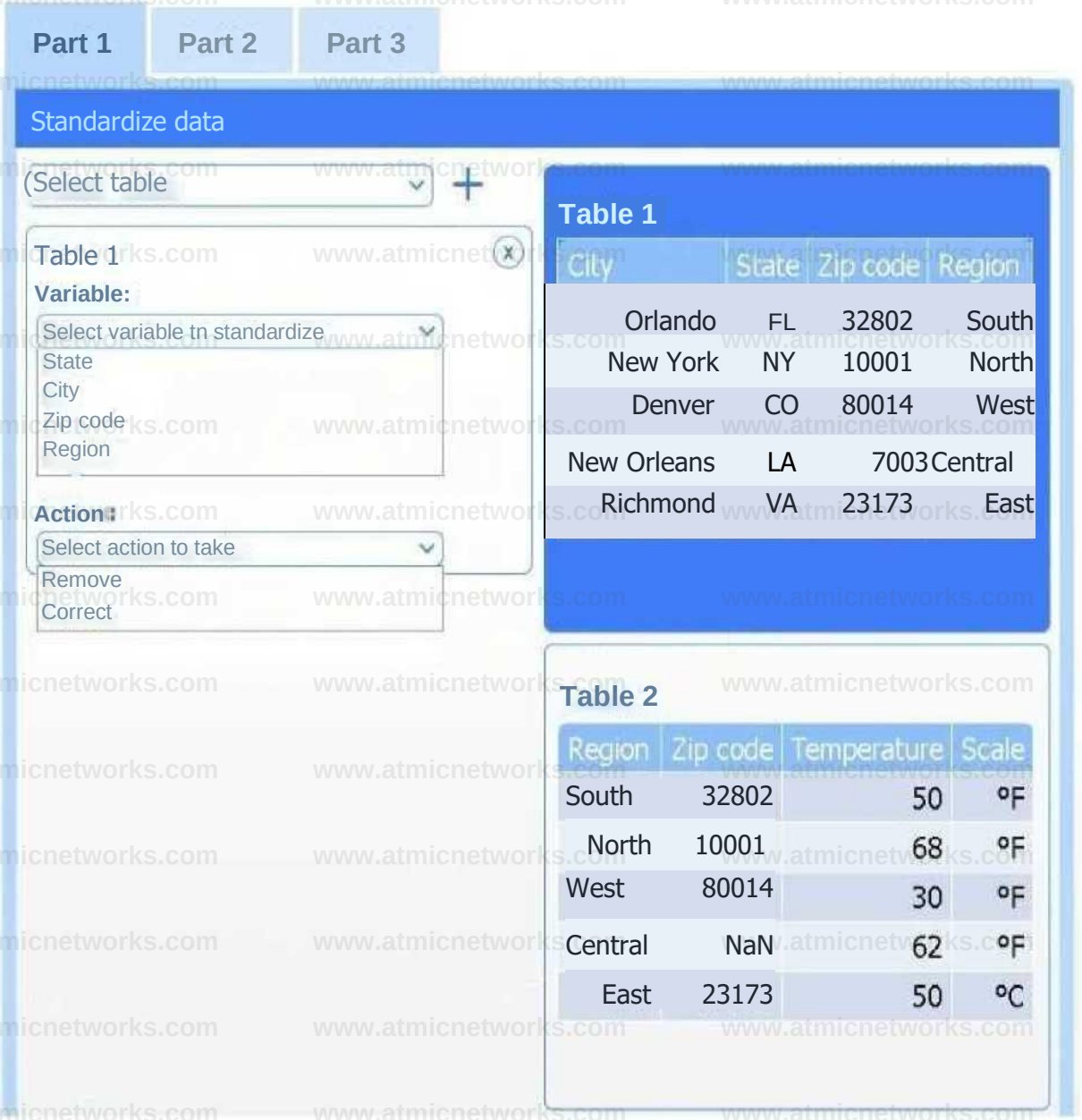
Table 2

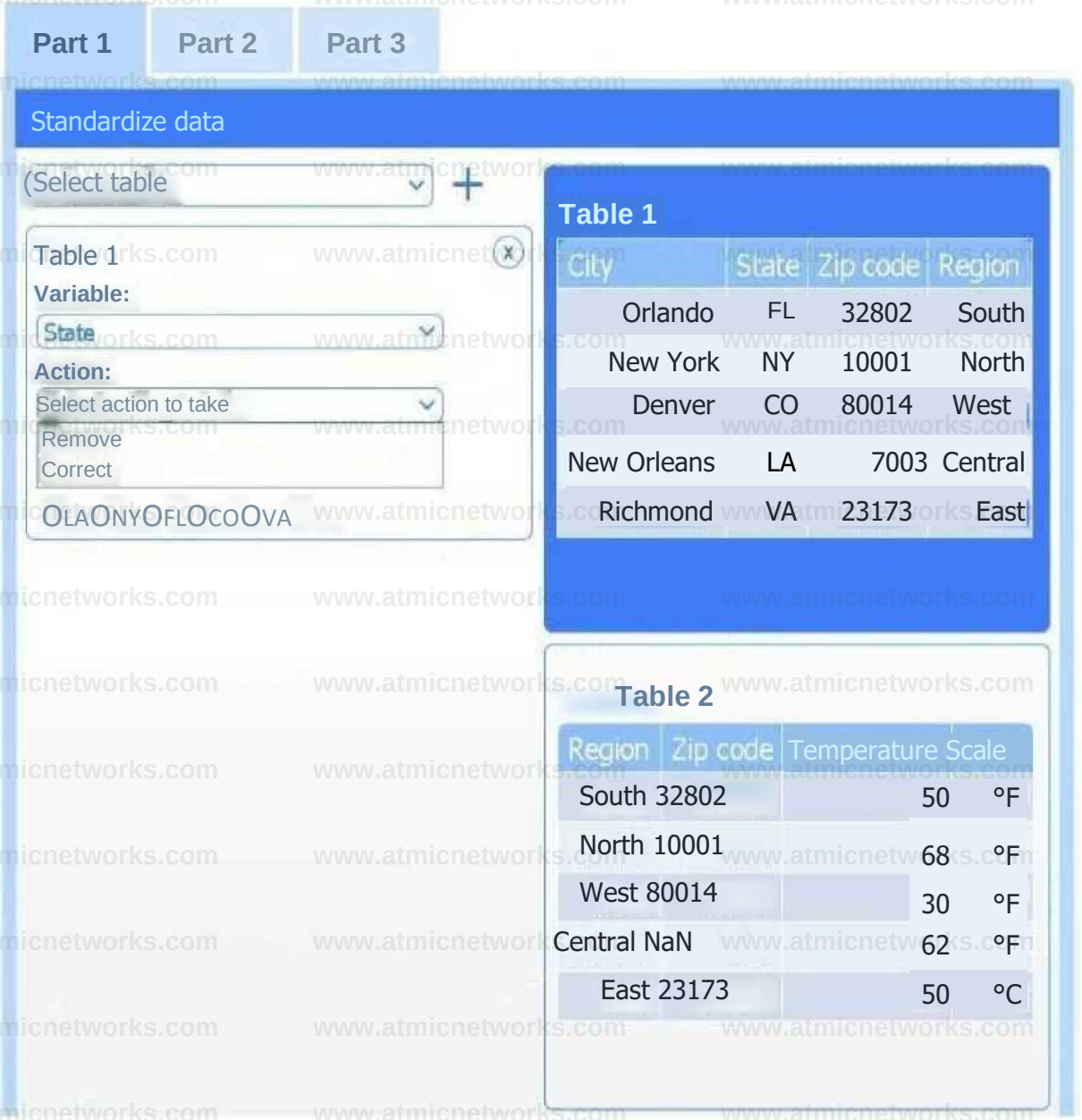
Table 1

Orlando FL 32802 South
New York NY 10001 North
Denver CO 80014 West
New Orleans LA 7003 Central
Richmond VA 23173 East

Table 2

I Region Zip code	Temperature	Scale I
South 32802	50	°F
North 10001	68	°F
West 80014	30	°F
Central NaN	62	°F
East 23173	50	°C





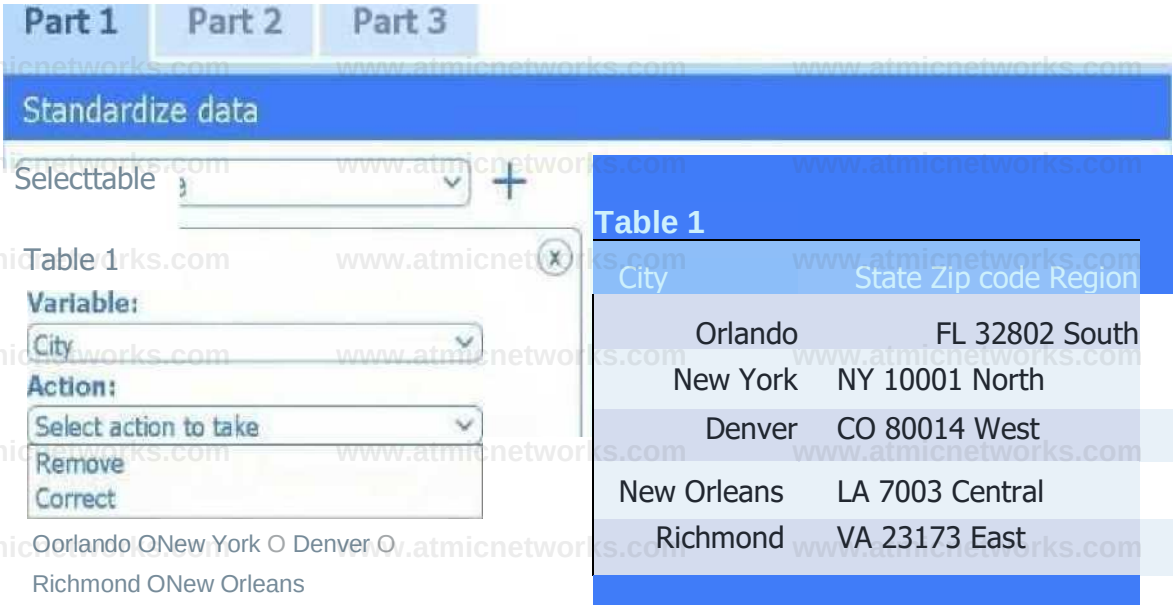


Table 1

City	State	Zip code	Region
Orlando	FL	32802	South
New York	NY	10001	North
Denver	CO	80014	West
New Orleans	LA	7003	Central
Richmond	VA	23173	East

Table 2

		Temperature	Scale
South	32802	50	°F
North	10001	68	°F
West	80014	30	°F
Central	NaN	62	°F
East	23173	50	°C


Part 1
Part 2
Part 3
Standardize data
(Select table
Table 1
Variable:
I Zip code
Action:
Select action to take
Remove
Correct
0 32802 O10001 080014 0 23173
07003
Table 1
City State Zip code Region
Orlando FL 32802 South
New York NY 10001 North
Denver CO 80014 West
New Orleans LA 7003 Central
Richmond VA 23173 East
Table 2
Region Zip code Temperature Scale
South 32802 50
North 10001 68 °F
West 80014 30 °F
Central NaN 62 °F
East 23173 50 °C

Part 1 Part 2 Part 3

Standardize data

i Select table

Table 1

Variable:

(Region

Action:

Select action to take

Remove

Correct

O South J North ! J West East

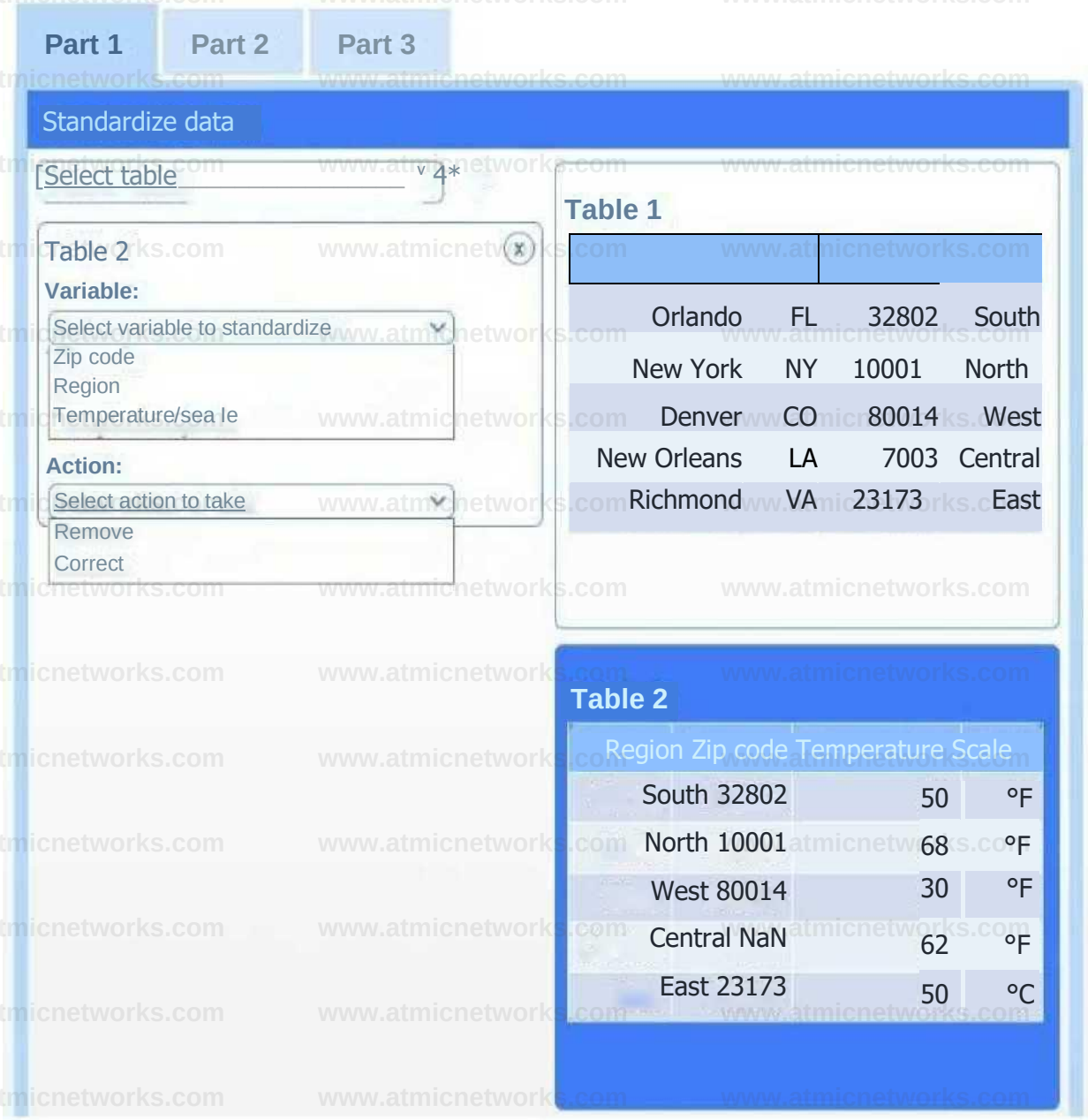
O Centra l

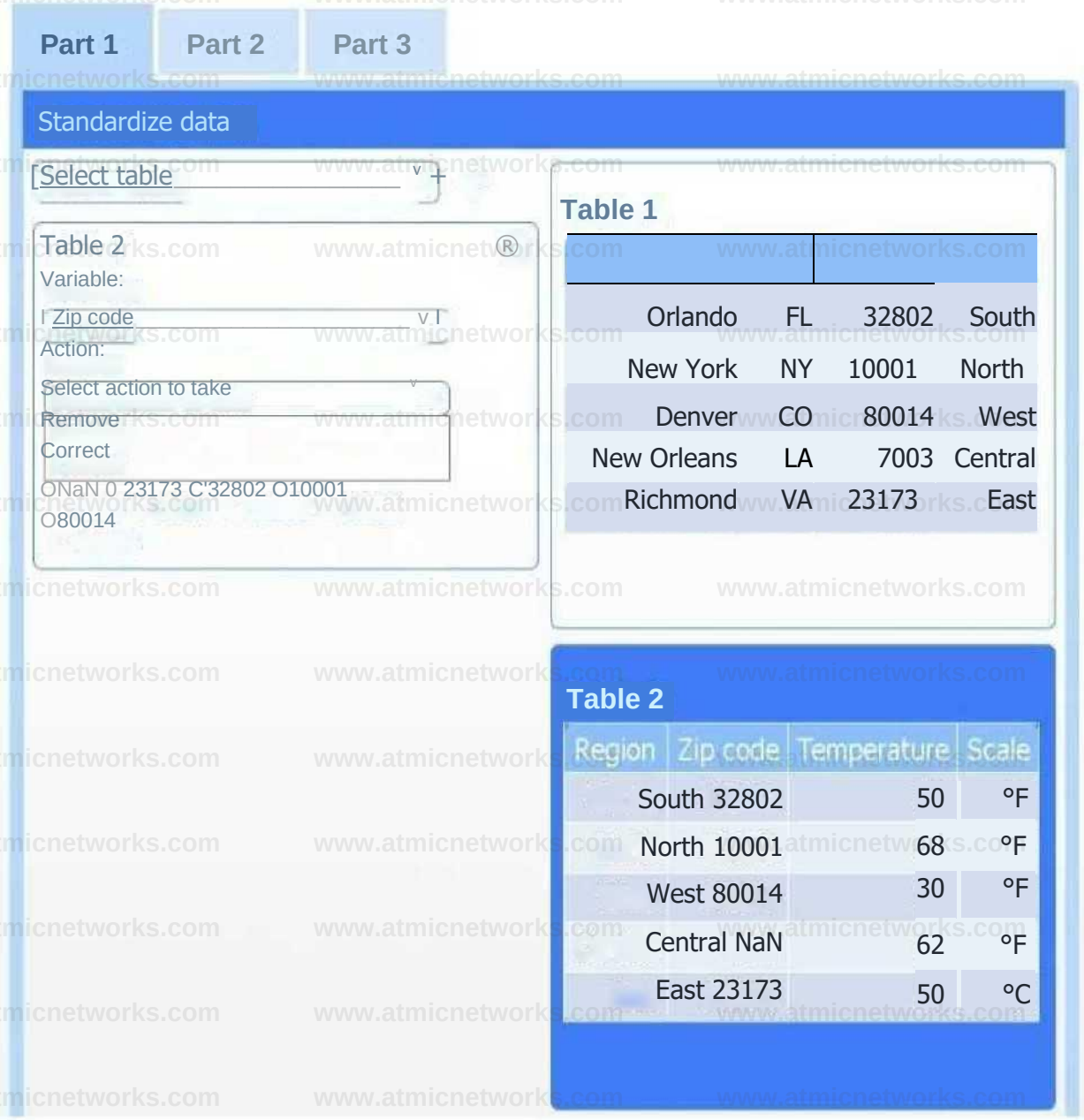
Table 1

City	State	Zip code	Region
Orlando	FL	32802	South
New York	NY	10001	North
Denver	CO	80014	West
New Orleans	LA	7003	Central
Richmond	VA	23173	East

Table 2

		Temperature	Scale
South	32802	50	°F
North	10001	68	°F
West	80014	30	°F
Central	NaN	62	°F
East	23173	50	°C





Part 1 Part 2 Part 3

Standardize data

[Select table ˅ 4"

Table 2 ®

Variable:

I Region ˅]

Action:

Select action to take ˅

Remove

Correct

South - North '-'West -'East

O Central

Orlando	FL	32802	South
New York	NY	10001	North
Denver	CO	80014	West
New Orleans	LA	7003	Central
Richmond	VA	23173	East

Table 1

Table 2

Region	Zip code	Temperature	Scale
South	32802	50	°F
North	10001	68	°F
West	80014	30	°F
Central	NaN	62	°F
East	23173	50	°C

Table 2

Region	Zip code	Temperature	Scale
South	32802	50	°F
North	10001	68	°F
West	80014	30	°F
Central	NaN	62	°F
East	23173	50	°C

Part 1 | **Part 2** | Part 3

Merge data

Select the **most** appropriate method to use when combining these two tables:

0 Data matching O Filter

O Union O Deduplication

Select the **most** appropriate variable to use when joining these sets of data:

O Region

O Zip code

Table 1

Orlando	FL	32802	South
New York	NY	10001	North
Denver	CO	80014	West
New Orleans	LA	7003	Central
Richmond	VA	23173	East

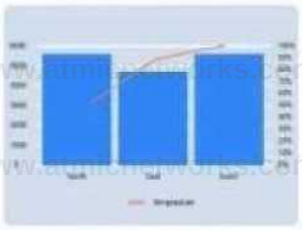
Table 2

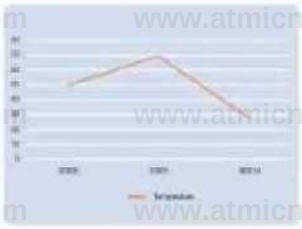
Region	Zip code	Temperature	Scale
South	32802	50	°F
North	10001	68	°F
West	80014	30	°F
Central	NaN	62	°F
East	23173	50	°C

Part 1 | Part 2 | **Part 3**

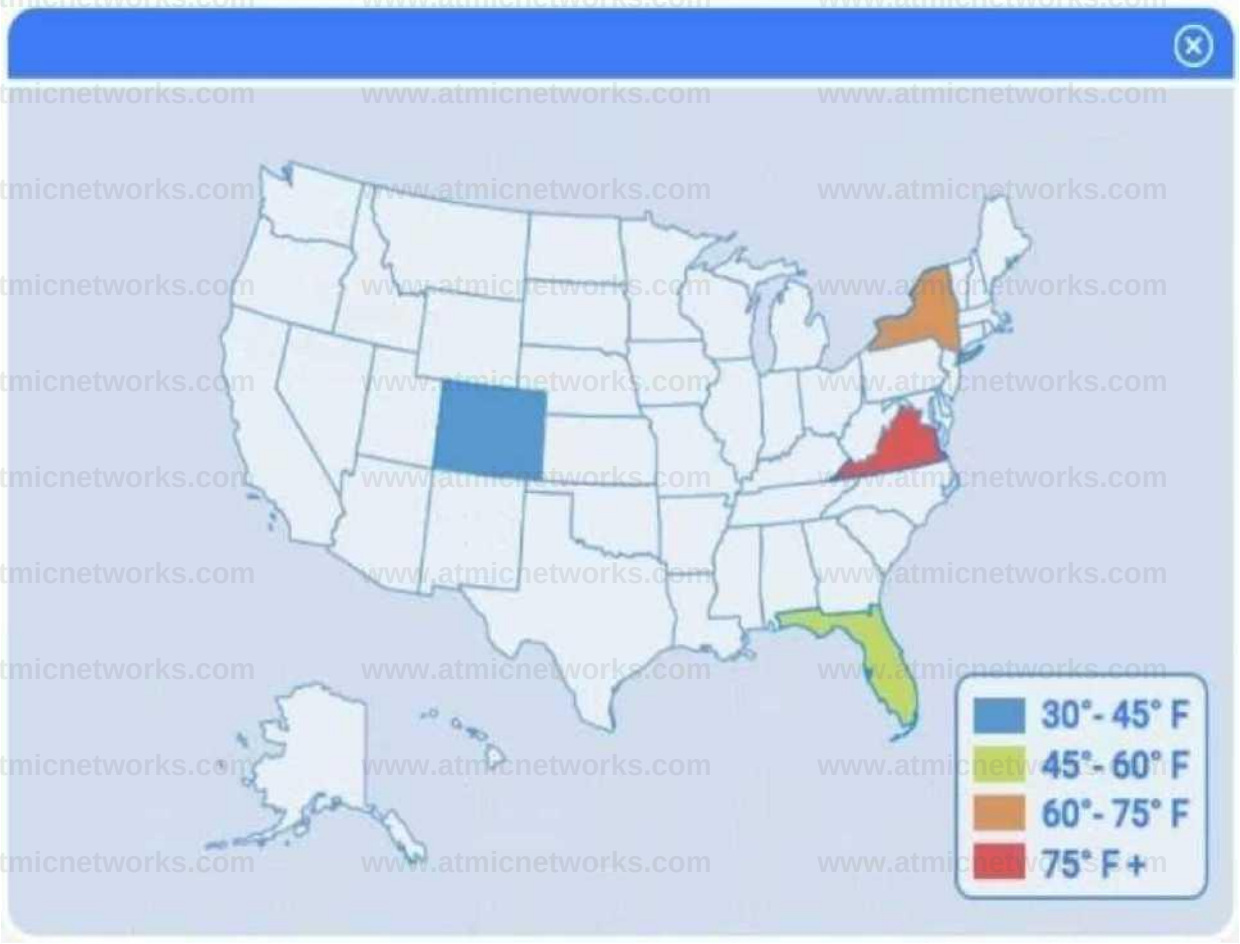
Visualization

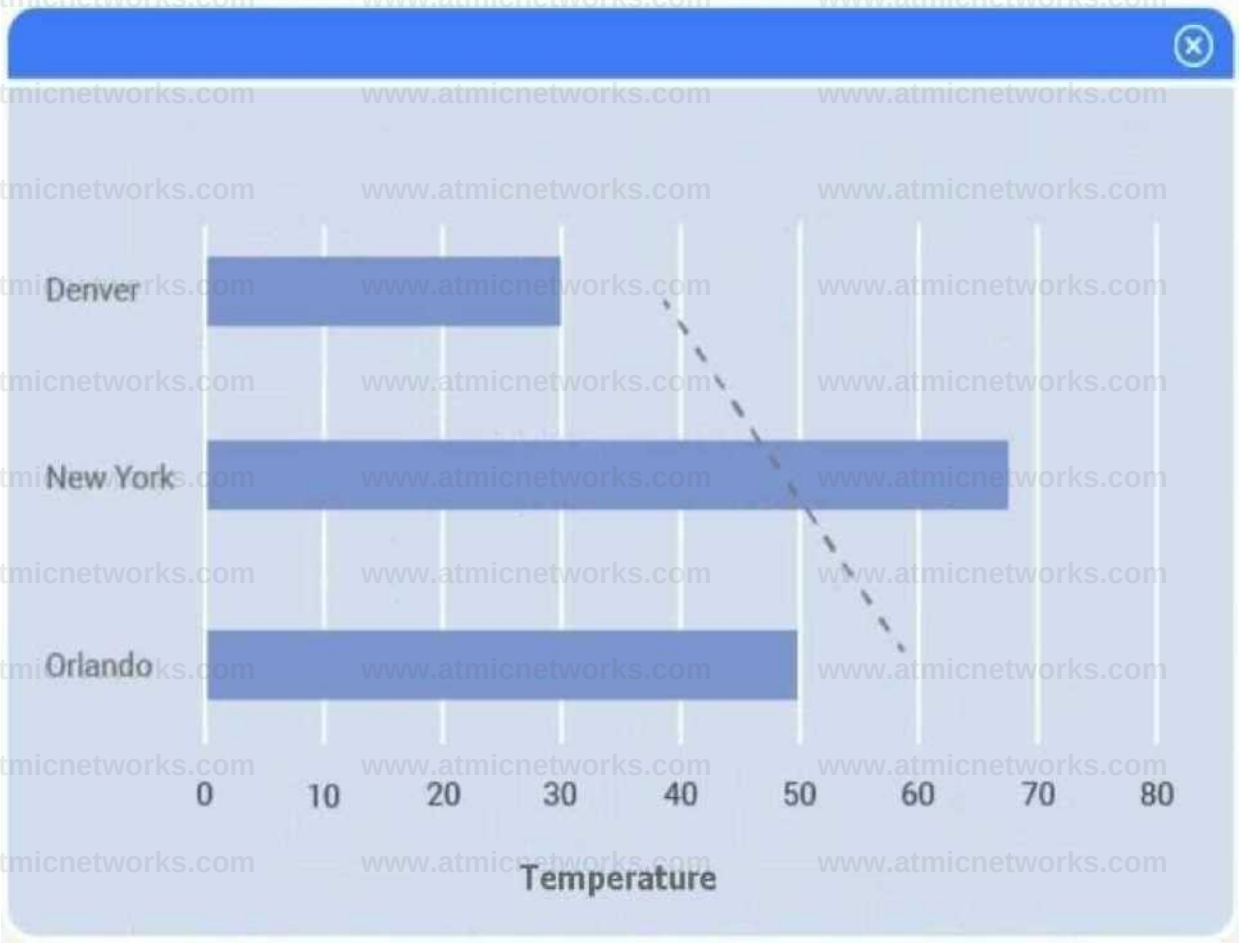
Select the **most** appropriate visualization based on the data set which represents what the client is looking for:

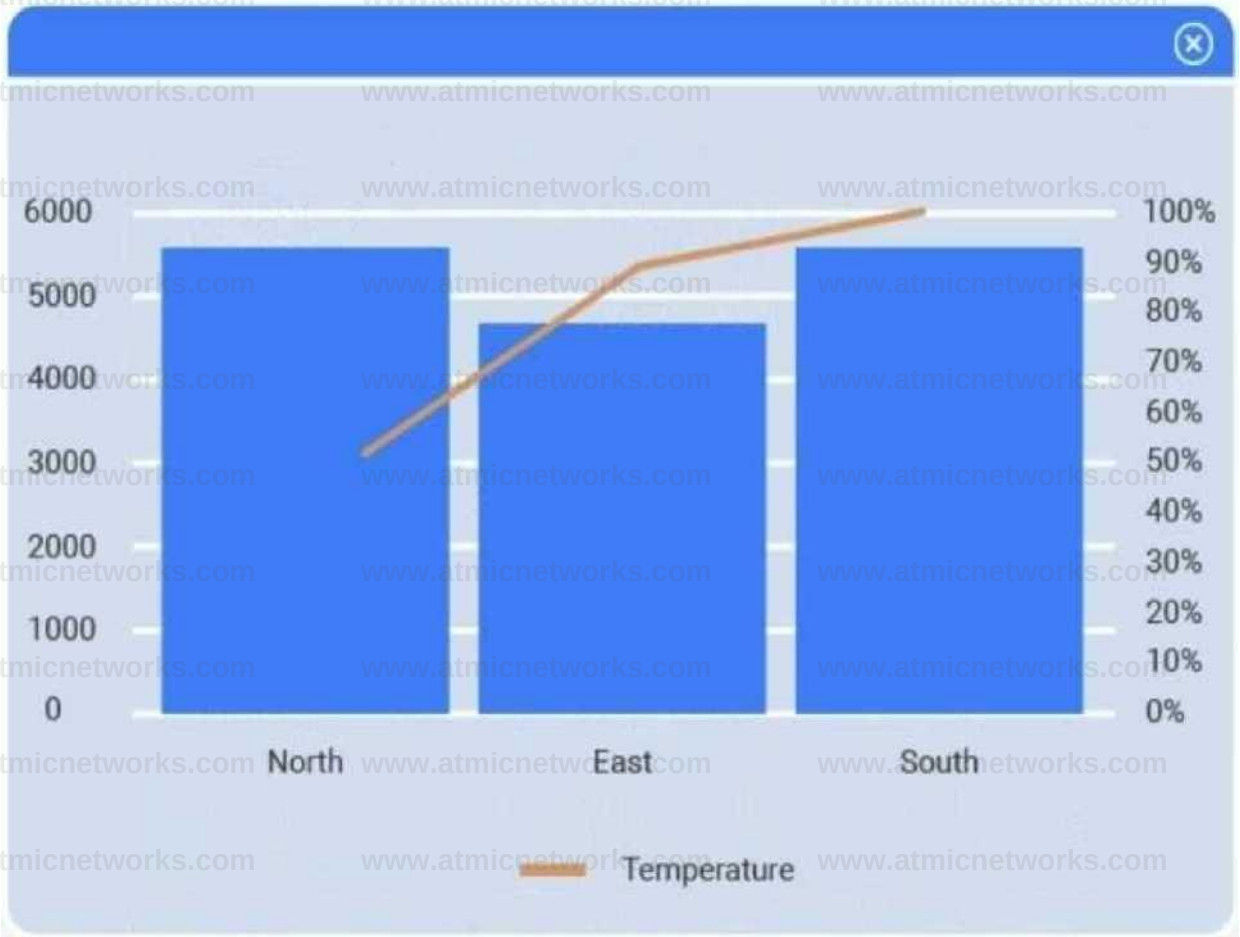




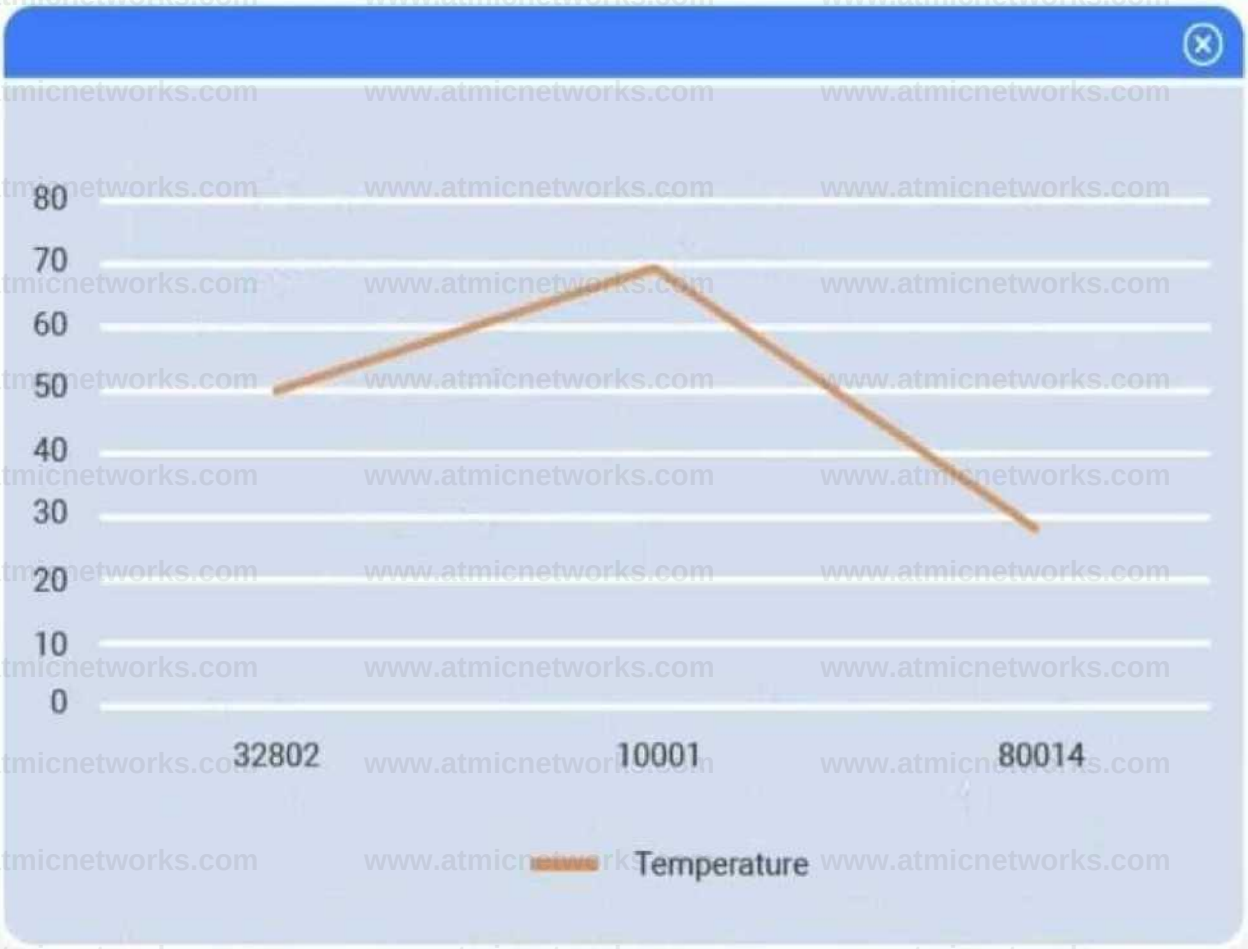
Region	City	State	Zip code	Temperature	Scale
South	Orlando	FL	32802	50	°F
North	New York	NY	10001	68	°F
West	Denver	CO	80014	30	°F
Central	New Orleans	LA	7003		
East	Richmond	VA	23173	50	°C
Central			NaN	62	°F




Denver
New York
Orlando
0
10
20
30
40
50
60
70
80
Temperature


6000
5000
4000
3000
2000
1000
0
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
North
East
South
Temperature


68°F
62°F
122°F
73°F
50°F



Answer: See explanation below.

Explanation:

Explanation:

Part 1

Select Table 2. Table 2 contains mixed temperature scales (°F and °C) that must be standardized before visualization.

Variable: Temperature/scale

Action: Correct

Value to correct: 50 °C

Part 1 Part 2 Part 3

Standardize data

(Select table ᵛ] +

Table 2 ®

Variable:

I Temperature/scale ᵛ I

Action: Select action to take ᵛ

Remove

Zorrect 062 F O

30 'F QSO C O68°F 050 F

Table 1

Orlando	FL	32802	South
New York	NY	10001	North
Denver	CO	80014	West
New Orleans	LA	7003	Central
Richmond	VA	23173	East

Table 2

South 32802	50	°F
North 10001	68	°F
West 80014	30	°F
Central NaN	62	°F
East 23173	50	°C

Part 2

Method: Data matching

Join variable: Zip code

You need to merge the two tables by aligning matching records, which is a data-matching (join) operation, and ZIP code is the shared, uniquely identifying field linking each region's weather reading to its city.

Part 1 Part 2 Part 3

Merge data

Select the **most** appropriate method to use when combining these two tables:

OData matching 0 Filter

0 Union 0 Deduplication

Select the **most** appropriate variable to use when joining these sets of data:

0 Region

□ Zip code

Table 1

Orlando	FL	32802	South
New York	NY	10001	North
Denver	CO	80014	West
New Orleans	LA	7003	Central
Richmond	VA	23173	East

Table 2

South	32802	50	°F
North	10001	68	°F
West	80014	30	°F
Central	NaN	62	°F
East	23173	50	°C

Part 3

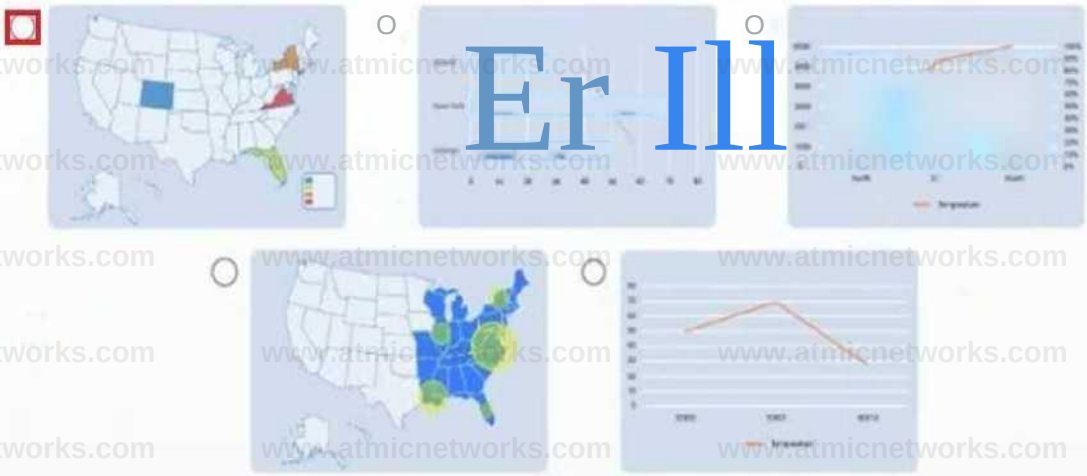
Choose the choropleth map (the first option).

A choropleth map best shows geographic variation in temperature by coloring each state (or region) according to its recorded value. This lets the client immediately see where the highest and lowest temperatures occur across the U.S. without distracting elements like bubble size or combined chart axes.

Part 1 | Part 2 | Part 3

Visualization

Select the **most** appropriate visualization based on the data set which represents what the client is looking for:



Region	City	State	Zip code	Temperature	Scale
South	Orlando	FL	32802	50	°F
North	New York	NY	10001	68	°F
West	Denver	CO	80014	30	°F
Central	New Orleans	LA	7003		
East	Richmond	VA	23173	50	°C
Central			NaN	62	°F

Question: 2

SIMULATION

A data scientist needs to determine whether product sales are impacted by other contributing factors. The client has provided the data scientist with sales and other variables in the data set.

The data scientist decides to test potential models that include other information.

INSTRUCTIONS

Part 1

Use the information provided in the table to select the appropriate regression model.

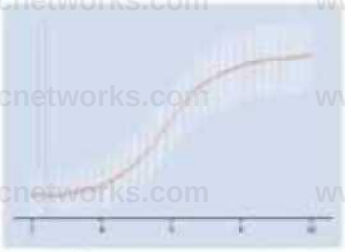
Part 2

Review the summary output and variable table to determine which variable is statistically significant.

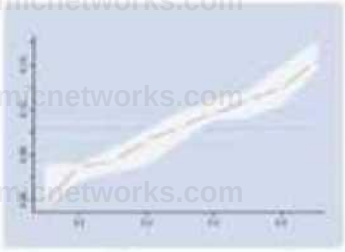
If at any time you would like to bring back the initial state of the simulation, please click the Reset All button.

Part 1 Part 2

Given the R^2 values, which of the following regression models best fits the relat variables^



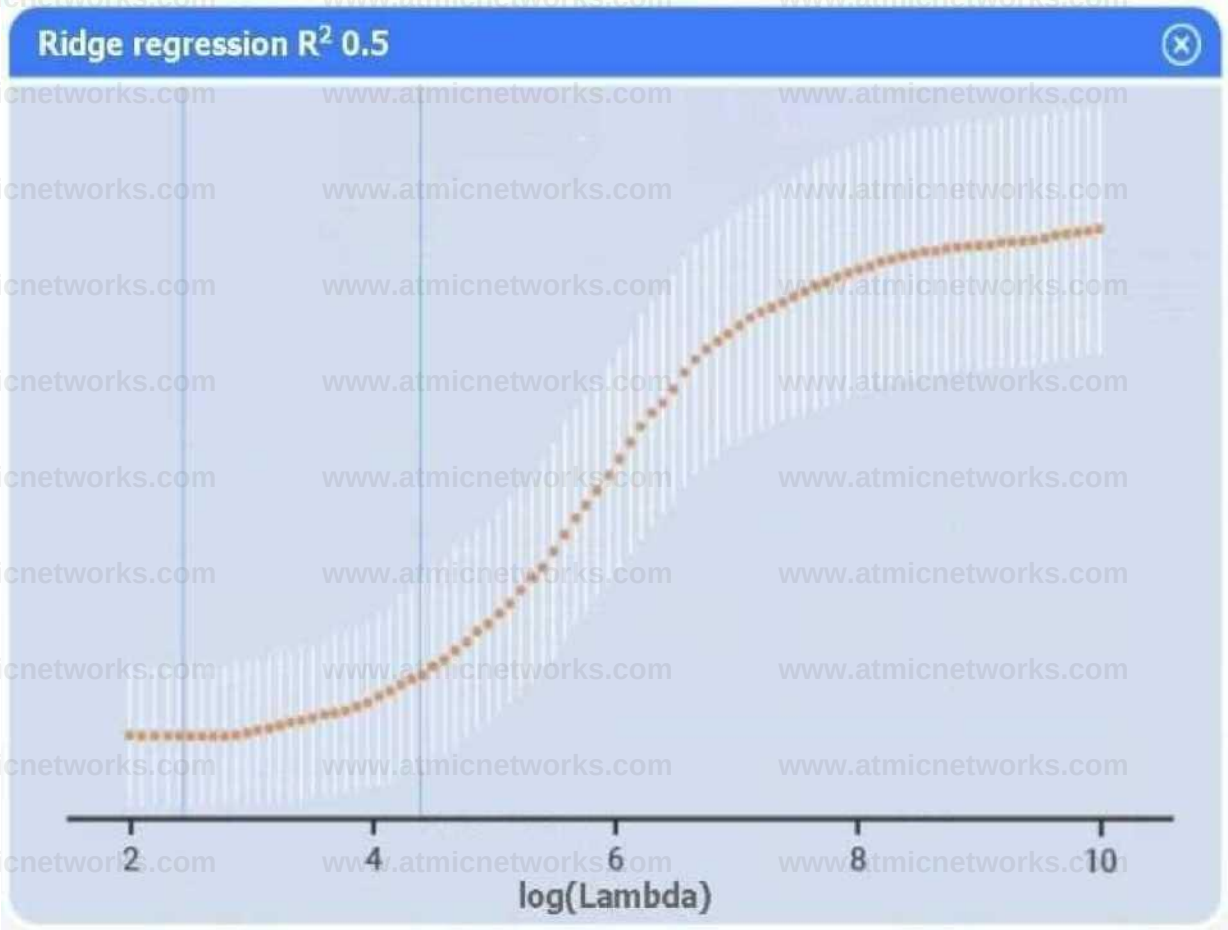
Ridge regression
R^2 0.5

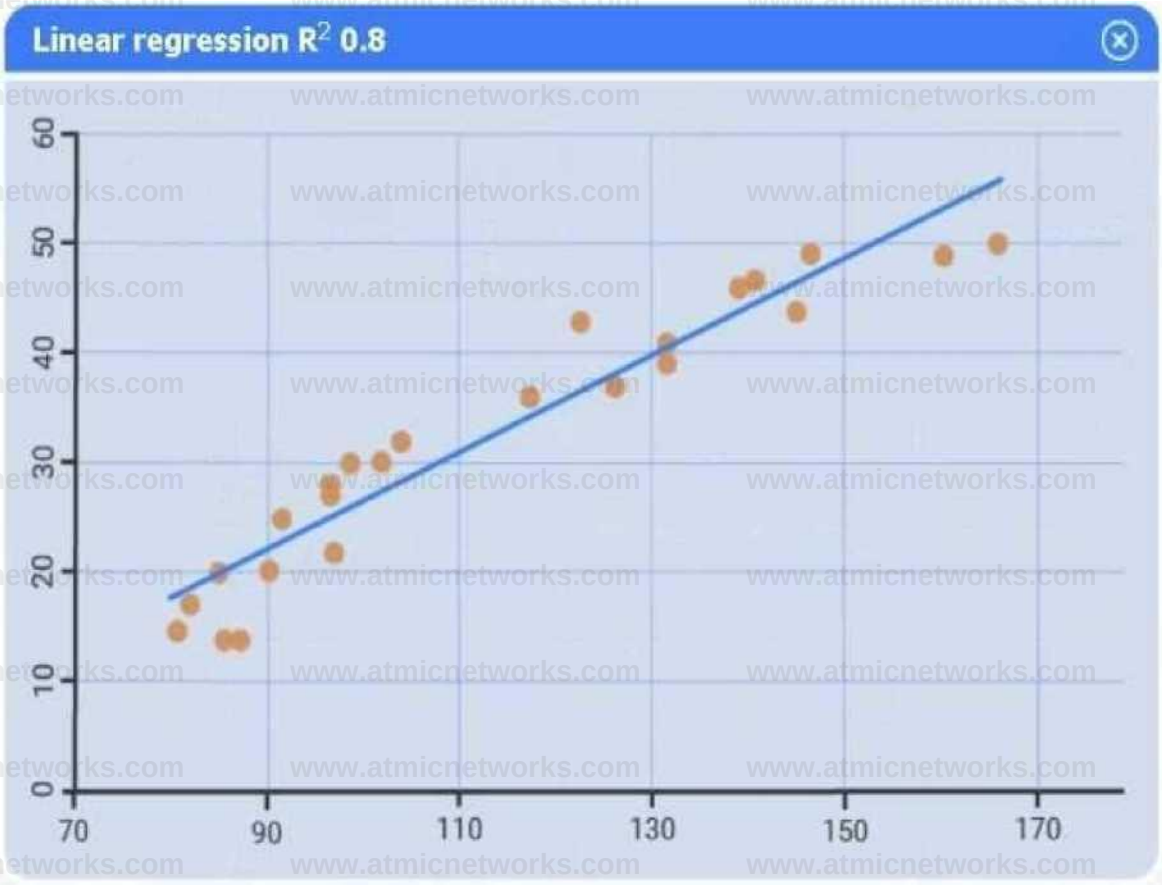


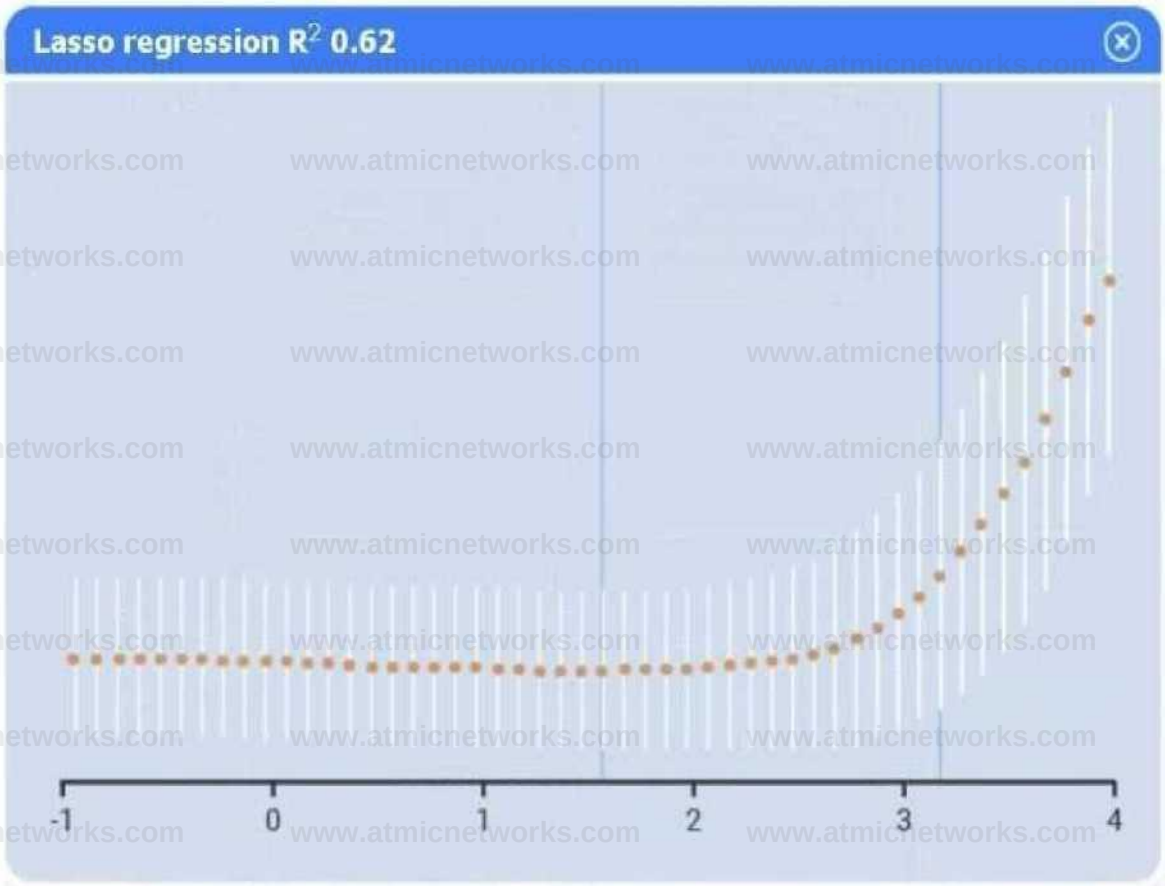
Quantile regression
R^2 0.6

Linear regression
R^2 0.8

1	3.118026935	6%
2	4.823728572	11%
3	7.149131157	18%
4	2.173859679	5%
5	3.519662597	9%
6	5.98246748	12%
7	8.495414141	14%
8	3.678906129	7%
9	3.539605808	6%

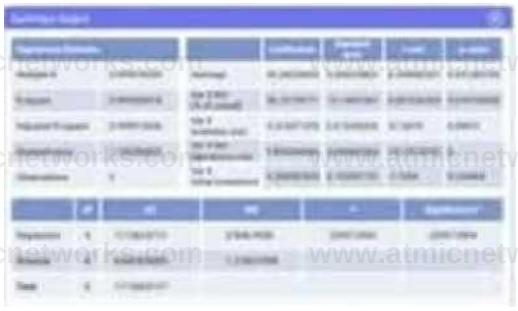






Part 1 Part 2

Time	[illegible] (in millions)	Var 2 ROI (% of overall)	Var 3 Inventory cost oper	Var
1	326.311584	16%	58	
2	507.9584031	8%	57	
3	232.5685962	5%	53	
4	117.3342091	7%	50	
5	242.866515	7%	60	
6	359.6300247	14%	50	
7	119.384542	19%	56	
8	372.064584	5%	56	
9	320.0212452	18%	51	



View summary output

Which of the following additional variable include in the new model?

O Var 5 Initial investment
O Var 4 Net opi
O Var 3 Inventory cost
O None of the i

Summary output

Multiple R 0.999976259		Intercept	30.24229003	9.306229821
R square 0.999956518		Var 2 ROI (% of overall)	50.72139711	13.14967361
Adjusted R square 0.999913036		Var 3 Inventory cost	-0.315571292	2.013342425
Standard error 1.100286825		Var 4 Net operations cost	9.854244454	0.049842563
Observations	9	Var 5 Initial investment	-0.268287655	0.103591751

Regression	4	111363.9712	27840.9928	22997.0<
Residual	4	4.842524393	1.210631098	
Total	8	111368.8137		

Answer: See explanation below.

Explanation:

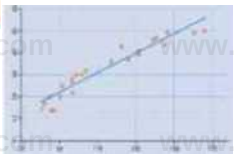
Part 1

Linear regression.

Of the four models, linear regression has the highest R^2 (0.8), indicating it explains the greatest proportion of variance in sales.

Part 1 Part 2

Given the R^2 values, which of the following regression models best fits the relationship between the variables?



Ridge regression	Quantile regression	Linear regression	Lasso regression
$R^2$0.5	R^2 0.6	$R^2$0.8	R^2 0.62

1	3.118026935	6%
2	4.823728572	11%
3	7.149131157	18%
4	2.173859679	5%
5	3.519662597	9%
6	5.98246748	12%
7	8.495414141	14%
8	3.678906129	7%
9	3.539605808	6%

Part 2

Var 4 – Net operations cost.

Net operations cost has a p-value of essentially 0 (far below 0.05), indicating it is the only additional predictor statistically significant in explaining sales. Neither inventory cost ($p \approx 0.90$) nor initial investment ($p \approx 0.23$) reach significance.

Part 1 Part 2

1	326.311584	16%	58	32
2	507.9584031	8%	57	50
3	232.5685962	5%	53	23
4	117.3342091	7%	50	11
5	242.866515	7%	60	24
6	359.6300247	14%	50	35
7	119.384542	19%	56	11
8	372.064584	5%	56	37
9	320.0212452	18%	51	31

View summary output

Which of tine following additional variables should the dat include in the new model?

O Var 5 Initial investment . ‘

Var 3 Inventory cost

QVar 4 Net operations cost

O None of the variables should be

Question: 3

A data scientist is building an inferential model with a single predictor variable. A scatter plot of the independent variable against the real-number dependent variable shows a strong relationship between them. The predictor variable is normally distributed with very few outliers. Which of the following algorithms is the best fit for this model, given the data scientist wants the model to be easily interpreted?

A. A logistic regression
B. An exponential regression
C. A linear regression

D. A probit regression

Answer: C

Explanation:

Question: 4

A data scientist wants to evaluate the performance of various nonlinear models. Which of the following is best suited for this task?

A. AIC
B. Chi-squared test
C. MCC
D. ANOVA

Answer: A

Explanation:

Question: 5

Which of the following is the layer that is responsible for the depth in deep learning?

A. Convolution

B. Dropout
C. Pooling
D. Hidden

Answer: D

Explanation:

Question: 6

Which of the following modeling tools is appropriate for solving a scheduling problem?

A. One-armed bandit
B. Constrained optimization
C. Decision tree
D. Gradient descent

Answer: B

Explanation:
Scheduling problems require finding the best allocation of resources subject to constraints (e.g., time slots, resource availability), which is precisely what constrained optimization algorithms are designed to handle.

Question: 7

Which of the following environmental changes is most likely to resolve a memory constraint error when running a complex model using distributed computing?

A. Converting an on-premises deployment to a containerized deployment

B. Migrating to a cloud deployment

C. Moving model processing to an edge deployment

D. Adding nodes to a cluster deployment

Answer: D

Explanation:

Increasing the number of nodes in your cluster directly expands the total available memory across the distributed system, alleviating memory-constraint errors without changing your code or deployment paradigm. Containerization or edge deployments don't inherently provide more memory, and migrating to the cloud alone doesn't guarantee additional nodes unless you explicitly scale out.

Question: 8

A data analyst wants to save a newly analyzed data set to a local storage option. The data set must meet the following requirements:

Be minimal in size

- Have the ability to be ingested quickly
- Have the associated schema, including data types, stored with it

Which of the following file types is the best to use?

A. JSON

B. Parquet

C. XML

D. CSV

Answer: B

Explanation:

Parquet is a columnar storage format that automatically includes schema (data types), uses efficient compression to minimize file size, and enables very fast reads for analytic workloads.

Question: 9

Which of the following is a key difference between KNN and k-means machine-learning techniques?

A. KNN operates exclusively on continuous data, while k-means can work with both continuous and categorical data.

B. KNN performs better with longitudinal data sets, while k-means performs better with survey data sets.

C. KNN is used for finding centroids, while k-means is used for finding nearest neighbors.

D. KNN is used for classification, while k-means is used for clustering.

Answer: D

Explanation:

KNN is a supervised algorithm that assigns labels based on the closest labeled examples, whereas k- means is an unsupervised method that partitions data into clusters by finding centroids without using any pre-existing labels.

Question: 10

A data scientist needs to:

Build a predictive model that gives the likelihood that a car will get a flat tire.

Provide a data set of cars that had flat tires and cars that did not.

All the cars in the data set had sensors taking weekly measurements of tire pressure similar to the sensors that will be installed in the cars consumers drive. Which of the following is the most immediate data concern?

A. Granularity misalignment

B. Multivariate outliers

C. Insufficient domain expertise

D. Lagged observations

Answer: D

Explanation:

Because tire-pressure sensors report only weekly measurements, you risk missing the critical pressure drop immediately preceding a flat. Those stale ("lagged") readings may not reflect the condition just before failure, undermining your model's ability to learn the true precursors to a flat tire.

Question: 11

The term "greedy algorithms" refers to machine-learning algorithms that:

A. update priors as more data is seen.

B. examine even/ node of a tree before making a decision.

C. apply a theoretical model to the distribution of the data.

D. make the locally optimal decision.

Answer: D

Explanation:

Greedy algorithms build the solution iteratively by choosing at each step the option that appears best at that moment, without reconsidering earlier choices.

Question: 12

A data scientist is deploying a model that needs to be accessed by multiple departments with minimal development effort by the departments. Which of the following APIs would be best for the data scientist to use?

A. SOAP
B. RPC
C. JSON
D. REST

Answer: D

Explanation:

RESTful APIs use standard HTTP methods and lightweight data formats (typically JSON), making them easy for diverse teams to integrate with minimal effort and without heavy tooling.

Question: 13

Which of the following compute delivery models allows packaging of only critical dependencies while developing a reusable asset?

A. Thin clients
B. Containers
C. Virtual machines
D. Edge devices

Answer: B

Explanation:

Containers encapsulate just the application and its critical dependencies on a lightweight runtime, making the resulting asset portable and reusable without bundling an entire operating system.

Question: 14

A data analyst is analyzing data and would like to build conceptual associations. Which of the following is the best way to accomplish this task?

A. n-grams

B. NER

C. TF-IDF

D. POS

Answer: A

Explanation:

n-grams capture contiguous sequences of words, revealing which terms co-occur and form meaningful multi-word concepts. By analyzing these frequent word combinations, you directly uncover conceptual associations in the text.

Question: 15

Which of the following belong in a presentation to the senior management team and/or C-suite executives? (Choose two.)

A. Full literature reviews

B. Code snippets

C. Final recommendations

D. High-level results

E. Detailed explanations of statistical tests

F. Security keys and login information

Answer: C

Explanation:

Senior leaders need actionable insights and the overarching outcomes, not the implementation details, so you present your key recommendations alongside a summary of results at a high level.

Question: 16

During EDA, a data scientist wants to look for patterns, such as linearity, in the dat

a. Which of the following plots should the data scientist use?

A. Violin

B. Box-and-whisker

C. Scatter

D. Q-Q

Answer: C

Explanation:

Scatter plots display pairs of numeric values on two axes, letting you visually assess relationships and patterns, such as linear trends, between variables.

Question: 17

Which of the following distribution methods or models can most effectively represent the actual arrival times of a bus that runs on an hourly schedule?

A. Binomial

B. Exponential

C. Normal

D. Poisson

Answer: C

Explanation:

Scheduled buses tend to arrive around a fixed time with random delays that cluster symmetrically around the hour. A normal distribution effectively models those continuous, bell-shaped deviations from the exact schedule.

Question: 18

A data scientist has constructed a model that meets the minimum performance requirements specified in the proposal for a prediction project. The data scientist thinks the model's accuracy should be improved, but the proposed deadline is approaching. Which of the following actions should the data scientist take first?

A. Continue collecting data.

B. Request additional funding.

C. Consult the key project stakeholder.

D. Test additional model specifications.

Answer: C

Explanation:

Since the model already meets the agreed-upon requirements and the deadline is near, the first step is to confirm with the stakeholder whether pursuing further accuracy gains is worth the additional time and resources. This ensures you align with business priorities before collecting more data, requesting funding, or tweaking the model further.

Question: 19

Which of the following best describes the minimization of the residual term in a ridge linear regression?

A. |e|

B. e

C. e2

D. 0

Answer: C

Explanation:

Ridge regression extends ordinary least squares by adding an L2 penalty on the coefficients, but it still minimizes the sum of squared residuals (e^2) as its loss term.

Question: 20

A statistician notices gaps in data associated with age-related illnesses and wants to further aggregate these observations. Which of the following is the best technique to achieve this goal?

A. Label encoding

B. Linearization

C. Binning

D. Imputing

Answer: C

Explanation:

Binning groups continuous age values into discrete intervals (e.g., age ranges), filling gaps by aggregating observations into broader categories. This directly addresses uneven or sparse age data by creating consistent age groups.

Question: 21

A data scientist needs to analyze a company's chemical businesses and is using the master database of the conglomerate company. Nothing in the data differentiates the data observations for the different businesses. Which of the following is the most efficient way to identify the chemical businesses' observations?

A. Ingest the data from all of the hard drives and perform exploratory data analysis to identify which business is responsible for chemical operations.

B. Perform analysis on all of the data and create a summary report on the results relevant to chemical operations.

C. Consult with the business team to identify which sites are responsible for chemical operations and ingest only the relevant data for analysis.

D. Ingest data from the hard drive containing the most data and present sample results on the chemical operations.

Answer: C

Explanation:
Engaging the business team leverages domain expertise to pinpoint which records pertain to chemical operations, allowing you to extract and analyze just the relevant subset. This avoids the time and resource waste of ingesting and sifting through unrelated data.

Question: 22

Which of the following distance metrics for KNN is best described as a straight line?

A. Radial

B. Euclidean

C. Cosine

D. Manhattan

Answer: B

Explanation:

Euclidean distance measures the straight-line distance between two points in space, matching the geometric "as-the-crow-flies" notion of distance.

Question: 23

A data scientist is building a forecasting model for the price of copper. The only input in this model is the daily price of copper for the last ten years. Which of the following forecasting techniques is the most appropriate for the data scientist to use?

A. Autoregressive

B. Moving average

C. Dynamic time warping

D. Relative strength

Answer: A

Explanation:

An autoregressive model uses past values of the series itself (here, historical daily copper prices) as predictors for future values, making it the most suitable technique when only the time-series history is available.

Question: 24

An analyst wants to show how the component pieces of a company's business units contribute to the company's overall revenue. Which of the following should the analyst use to best demonstrate this breakdown?

A. Box-and-whisker chart

B. Sankey diagram

C. Scatter plot matrix

D. Residual chart

Answer: B

Explanation:

A Sankey diagram visualizes flows from individual business units into the total, with the width of each flow proportional to its revenue contribution, making it ideal for showing how each component feeds the overall total.

Question: 25

Which of the following does k represent in the k-means model?

A. Number of model tests

B. Number of data splits

C. Number of clusters

D. Distance between features

Answer: C

Explanation:

In k-means clustering, the parameter k directly defines how many clusters the algorithm will partition the data into.

Question: 26

Which of the following techniques enables automation and iteration of code releases?

A. Virtualization

B. Markdown

C. Code isolation

D. CI/CD

Answer: D

Explanation:

Continuous Integration/Continuous Deployment pipelines automate the building, testing, and delivery of code, enabling rapid, repeatable, and iterative releases with minimal manual intervention.

Question: 27

In a modeling project, people evaluate phrases and provide reactions as the target variable for the model. Which of the following best describes what this model is doing?

A. Sentiment analysis

B. Named-entity recognition

C. TF-IDF vectorization

D. Part-of-speech tagging

Answer: A

Explanation:

The model predicts people's reactions (e.g., positive, negative, neutral) to given phrases, which is the core of sentiment analysis.

Question: 28

A computer vision model is trained to identify cats on a training set that is composed of both cat and dog images. The model predicts a picture of a cat is a dog. Which of the following describes this error?

A. Error due to reality

B. False positive error

C. Sampling error

D. Type II error

Answer: D

Explanation:

Classifying an actual cat (positive instance) as a dog (negative prediction) is a false negative, which corresponds to a Type II error.

Question: 29

Which of the following JOINS would generate the largest amount of data?

A. RIGHT JOIN

B. LEFT JOIN

C. CROSS JOIN

D. INNER JOIN

Answer: C

Explanation:

A CROSS JOIN produces the Cartesian product of the two tables (every row from the first paired with every row from the second), yielding far more rows than any of the other join types.

Question: 30

A data scientist built several models that perform about the same but vary in the number of features. Which of the following models should the data scientist recommend for production according to Occam's razor?

A. The model with the fewest features and highest performance

B. The model with the fewest features and the lowest performance

C. The model with the most features and the lowest performance

D. The model with the most features and the highest performance

Answer: A

Explanation:

According to Occam's razor, when models perform equivalently, you choose the simplest one - in this case, the model that achieves the needed performance with the fewest features.

Question: 31

A data analyst wants to use compression on an analyzed data set and send it to a new destination for further processing. Which of the following issues will most likely occur?

A. Library dependency will be missing.

B. Server CPU usage will be too high.

C. Operating system support will be missing.

D. Server memory usage will be too high.

Answer: B

Explanation:

Compression and decompression are CPU-intensive operations; on large data sets, the extra processing load can significantly spike CPU utilization. Memory, OS support, or library dependencies are far less likely to be the primary bottleneck in a standard compression workflow.

Question: 32

The most likely concern with a one-feature, machine-learning model is high error due to:

A. bias

B. dimensionality.

C. variance.

D. probability.

Answer: A

Explanation:

A model with only one feature is unlikely to capture the true complexity of the data's underlying relationships, leading to systematic underfitting - i.e., high bias.

Question: 33

Given matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \\ 3 & 2 & 1 \end{bmatrix}$$

Which of the following is AT?

A)

```
'3 21'
  213
.1 23.
```

B)

```
'12 3'
  213
-3 2 1-
```

C)

```
1 123"
  212
  33 1.
```

D)

```
'33 1'
212
-12 3-
```

A. Option A

B. Option B

C. Option C

D. Option D

Answer: C

Explanation:

Transposing swaps rows and columns, so the (i, j) entry becomes the (j, i) entry.

Question: 34

A data scientist is clustering a data set but does not want to specify the number of clusters present.
Which of the following algorithms should the data scientist use?

A. DBSCAN

B. k-nearest neighbors

C. k-means

D. Logistic regression

Answer: A

Explanation:

DBSCAN discovers clusters based on density without requiring you to predefine the number of clusters, automatically finding arbitrarily shaped groups and identifying noise points.

Question: 35

A data analyst wants to find the latitude and longitude of a mailing address. Which of the following is the best method to use?

A. One-hot encoding

B. Binning

C. Geocoding

D. Imputing

Answer: C

Explanation:

Geocoding is the process of converting a postal address into geographic coordinates (latitude and longitude), making it the appropriate method.

Question: 36

Which of the following describes the appropriate use case for PCA?

A. Dimensionality reduction

B. Classification

C. Regression

D. Recommendation

Answer: A

Explanation:

Principal Component Analysis transforms correlated features into a smaller set of uncorrelated components that capture most of the variance, making it ideal for reducing dimensionality before modeling or visualization.

Question: 37

A data scientist observes findings that indicate that as electrical grids in a country become more and more connected over time, the frequency of brownouts and blackouts in total decrease, and the frequency of major brownouts and blackouts increase.

Which of the following distribution metrics could best be identified?

A. Scale axis magnitudes

B. Kurtosis

C. Skewness

D. Normality

Answer: B

Explanation:

Kurtosis quantifies how heavy or light the tails of a distribution are. In this case, fewer overall events but more extreme (major) brownouts/blackouts indicates heavier tails over time. This is exactly what an increasing kurtosis would reveal.

Question: 38

A data scientist is merging two tables. Table 1 contains employee IDs and roles. Table 2 contains employee IDs and team assignments. Which of the following is the best technique to combine these data sets?

A. inner join between Table 1 and Table 2

B. left join on Table 1 with Table 2

C. right join on Table 1 with Table 2

D. outer join between Table 1 and Table 2

Answer: A

Explanation:

An INNER JOIN merges records only where the employee ID exists in both tables, yielding a single combined table of each employee's role paired with their team assignment.

Question: 39

Which of the following is a classic example of a constrained optimization problem?

A. The cold start problem

B. The traveling salesman

C. Calculating local maximum

D. Calculating gradient descent

Answer: B

Explanation:

The traveling-salesman problem seeks the shortest possible route that visits each city exactly once and returns to the start, making it a textbook example of optimization under explicit constraints.

Question: 40

A data scientist wants to digitize historical hard copies of documents. Which of the following is the best method for this task?

A. Word2vec

B. Optical character recognition

C. Latent semantic analysis

D. Semantic segmentation

Answer: B

Explanation:

OCR converts scanned images of text into machine-readable characters, making it the appropriate tool for digitizing printed or handwritten historical documents.

Question: 41

A data scientist trained a model for departments to share. The departments must access the model using HTTP requests. Which of the following approaches is appropriate?

A. Utilize distributed computing.

B. Deploy containers.

C. Create an endpoint.

D. Use the File Transfer Protocol.

Answer: C

Explanation:

Exposing the model behind an HTTP endpoint (for example, a REST API) allows other departments to send requests and receive predictions directly over HTTP. The other options don't inherently provide a request–response interface for sharing a model.

Question: 42

Given the following:

$$X_t = \delta + \phi_1 X_{t-1} + \omega_t \text{ where } \omega_t \sim N(0, \sigma_\omega^2)$$

Which of the following time series models best represents this process?

A. ARIMA(1,1,1)

B. ARMA(1,1)

C. SARIMA(1, 1, 1) x (1, 1, 1)1

D. AR(1)

Answer: D

Explanation:

The model has a single autoregressive term and only white-noise errors, matching the definition of an AR(1) process.

Question: 43

Which of the following methods should a data scientist use just before switching to a potential replacement model?

A. A/B testing
B. Performance monitoring
C. CI/CD
D. Containerization

Answer: A

Explanation:

A/B testing lets you compare the current model against the candidate in parallel, measuring performance on live data, before fully switching to the new model.

Question: 44

A data scientist is presenting the recommendations from a monthslong modeling and experiment process to the company's Chief Executive Officer. Which of the following is the best set of artifacts to include in the presentation?

A. Methods, data overview, results, recommendations, and charts

B. Results, recommendations, justifications, and clear charts

C. Recommendation charts justifications code reviews and results

D. Methodology, code snippets, findings, data tables, and p values

Answer: B

Explanation:

Executive audiences need concise, high-level insights: what you found (results), what you suggest (recommendations), why it matters (justifications), and visual summaries (clear charts). Detailed methods, code, or raw data aren't appropriate at the C-suite level.

Question: 45

A data scientist is developing a model to predict the outcome of a vote for a national mascot. The choice is between tigers and lions. The full data set represents feedback from individuals representing 17 professions and 12 different locations. The following rank aggregation represents 80% of the data set:

Survey rank	Profession	Location	Voter preference
1	Data scientist	4	Tigers
2	Data scientist	3	Tigers
3	Data analyst	4	Tigers

Which of the following is the most likely concern about the model's ability to predict the outcome of the vote?

A. Interpolated data

B. Extrapolated data

C. In-sample data

D. Out-of-sample data

Answer: D

Explanation:

The aggregated feedback covers only 80% of respondents, mostly from a few professions and locations, so the model hasn't "seen" the remaining 20% (and those underrepresented groups). Its performance on those unseen subsets (out-of-sample data) is therefore the primary concern for how well it will predict the actual vote.

Question: 46

A data scientist is working with a data set that covers a two-year period for a large number of machines. The data set contains:

- Machine system ID numbers
- Sensor measurement values
- Daily time stamps for each machine

The data scientist needs to plot the total measurements from all the machines over the entire time period. Which of the following is the best way to present this data?

A. Scatter plot

B. Line plot

C. Histogram

D. Box-and-whisker plot

Answer: B

Explanation:

Summing measurements across all machines for each day produces a time series, and a line plot is the standard way to visualize how that daily total evolves over the two-year period.

Question: 47

A data scientist has built an image recognition model that distinguishes cars from trucks. The data scientist now wants to measure the rate at which the model correctly identifies a car as a car versus when it misidentifies a truck as a car. Which of the following would best convey this information?

A. Confusion matrix

B. AUC/ROC curve

C. Box plot

D. Correlation plot

Answer: A

Explanation:

A confusion matrix directly shows true positives (cars correctly identified) and false positives (trucks misidentified as cars), giving you exactly the rates you're interested in.

Question: 48

A data analyst wants to generate the most data using tables from a database. Which of the following is the best way to accomplish this objective?

A. INNER JOIN

B. LEFT OUTER JOIN

C. RIGHT OUTER JOIN

D. FULL OUTER JOIN

Answer: D

Explanation:
A full outer join returns every row from both tables, matched where possible and unmatched rows filled with NULLs, yielding at least as many (and typically more) rows than any other join type.

Question: 49

A data scientist is building a model to predict customer credit scores based on information collected from reporting agencies. The model needs to automatically adjust its parameters to adapt to recent changes in the information collected. Which of the following is the best model to use?

A. Decision tree

B. Random forest

C. Linear discrimination analysis

D. XGBoost

Answer: D

Explanation:

XGBoost supports “warm-start” incremental training, continuing to refine the existing ensemble with new data, so it can automatically update its parameters as new agency information arrives. The other methods require full retraining to incorporate recent changes.

Question: 50

A data scientist is creating a responsive model that will update a product's daily pricing based on the previous day's sales volume.

Which of the following resource constraints is the data scientist's greatest concern?

A. Deployment time
B. Training time
C. Development time
D. Data collection time

Answer: B

Explanation:

Because the model must be retrained every day on yesterday's sales data to set today's prices, the time it takes to train the model becomes the critical bottleneck in a responsive, daily-update workflow.

Question: 51

A data scientist wants to predict a person's travel destination. The options are:

- Branson, Missouri, United States
- Mount Kilimanjaro, Tanzania
- Disneyland Paris, Paris, France
- Sydney Opera House, Sydney, Australia

Which of the following models would best fit this use case?

A. Linear discriminant analysis

B. k-means modeling

C. Latent semantic analysis

D. Principal component analysis

Answer: A

Explanation:

You need a supervised multiclass classification model to predict one of the four labeled destinations. Linear Discriminant Analysis is designed for such tasks, finding the linear boundaries that best separate the known destination classes.

Question: 52

A data scientist is working with a data set that has ten predictors and wants to use only the predictors that most influence the results. Which of the following models would be the best for the data scientist to use?

A. OLS

B. Ridge

C. Weighted least squares

D. LASSO

Answer: D

Explanation:

LASSO regression uses an L1 penalty that drives less-important feature coefficients to exactly zero, effectively selecting only the predictors that most influence the outcome.

Question: 53

A data scientist uses a large data set to build multiple linear regression models to predict the likely market value of a real

estate property. The selected new model has an RMSE of 995 on the holdout set and an adjusted R2 of .75. The benchmark model has an RMSE of 1,000 on the holdout set. Which of the following is the best business statement regarding the new model?

A. The model should be deployed because it has a lower RMSE.

B. The model's adjusted R2 is exceptionally strong for such a complex relationship.

C. The model fails to improve meaningfully on the benchmark model.

D. The model's adjusted R2 is too low for the real estate industry.

Answer: C

Explanation:

Although the new model's RMSE is technically lower (995 vs. 1,000), the five-point improvement on holdout data is negligible in most real-estate contexts and unlikely to produce meaningful business value over the existing benchmark.

Question: 54

Which of the following layer sets includes the minimum three layers required to constitute an artificial neural network?

A. An input layer, a pooling layer, and an output layer

B. An input layer, a convolutional layer, and a hidden layer

C. An input layer, a hidden layer, and an output layer

D. An input layer, a dropout layer, and a hidden layer

Answer: C

Explanation:

By definition, an artificial neural network requires at least these three fundamental layers: the input layer to receive data, one or more hidden layers to perform transformations, and the output layer to produce predictions. Pooling, convolutional, and dropout layers are useful in specialized architectures (e.g., CNNs) but aren't part of the minimal ANN structure.

Question: 55

Which of the following best describes the minimization of the residual term in a LASSO linear regression?

A. |e|

B. e

C. 0

D. e2

Answer: D

Explanation:

LASSO regression retains the ordinary least squares loss by minimizing the sum of squared residuals (e^2), with an added L1 penalty on the coefficients, but the residual term itself remains squared.

Question: 56

A data scientist is building a proof of concept for a commercialized machine-learning model. Which of the following is the best starting point?

A. Literature review

B. Model performance evaluation

C. Hyperparameter tuning

D. Model selection

Answer: A

Explanation:

Before diving into selecting or tuning models, a literature review grounds the proof of concept in existing research and best practices, ensuring the approach aligns with state-of-the-art methods and the problem's domain requirements.

Question: 57

Which of the following explains back propagation?

A. The passage of convolutions backward through a neural network to update weights and biases

B. The passage of accuracy backward through a neural network to update weights and biases

C. The passage of nodes backward through a neural network to update weights and biases

D. The passage of errors backward through a neural network to update weights and biases

Answer: D

Explanation:

Back propagation computes the gradient of the loss (error) with respect to each weight by propagating the error signal backward through the network, then uses those gradients to adjust weights and biases.

Question: 58

A data scientist is standardizing a large data set that contains website addresses. A specific string inside some of the web addresses needs to be extracted. Which of the following is the best method for extracting the desired string from the text data?

A. Regular expressions

B. Named-entity recognition

C. Large language model

D. Find and replace

Answer: A

Explanation:

Question: 59

A model's results show increasing explanatory value as additional independent variables are added to the model. Which of the following is the most appropriate statistic?

A. Adjusted R2

B. p value

C. x2

D. R2

Answer: A

Explanation:

Adjusted R^2 accounts for the number of predictors in the model, only increasing when a new independent variable adds genuine explanatory power beyond what random chance would predict. In contrast, plain R^2 will always rise (or stay the same) as you add more variables, regardless of their true relevance.

Question: 60

A team is building a spam detection system. The team wants a probability-based identification method without complex, in-depth training from the historical data set. Which of the following methods would best serve this purpose?

A. Logistic regression

B. Random forest

C. Naive Baves

D. Linear regression

Answer: C

Explanation:

Naive Bayes directly computes class probabilities using simple frequency counts under the independence assumption, requiring minimal training complexity and no iterative optimization— ideal for fast, probability-based spam detection.

Question: 61

A data scientist is using the following confusion matrix to assess model performance:

	Actually fails	Actually succeeds
Predicted to fail	80%	20%
Predicted to succeed	15%	85%

The model is predicting whether a delivery truck will be able to make 200 scheduled delivery stops. Every time the model is correct, the company saves an hour in planning and scheduling of maintenance work. Every time the model is wrong, the company loses four hours of delivery time for the truck. Which of the following is the net model impact for the company?

A. 25 hours lost

B. 25 hours saved

C. 165 hours lost

D. 165 hours saved

Answer: A

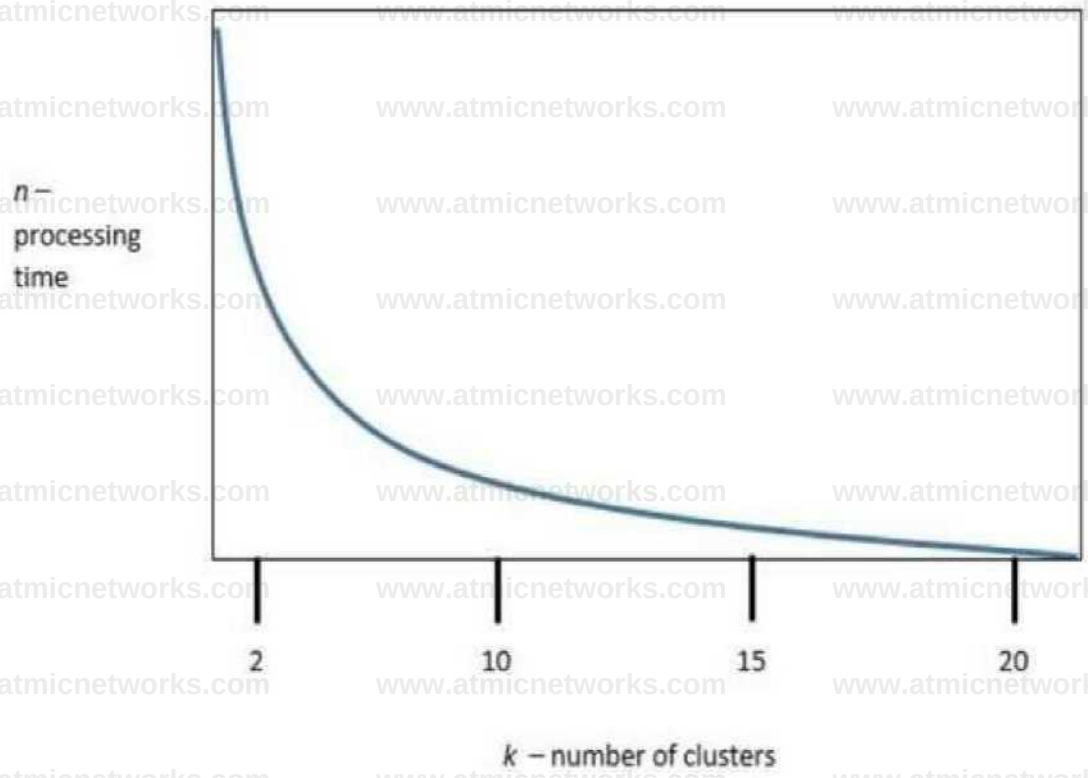
Explanation:

Treat each "predicted-to-fail" and "predicted-to-succeed" row as coming from 100 cases apiece (200 total).

- Predicted-fail & actually-fail: 80 → saves 80 hr
- Predicted-succeed & actually-succeed: 85 → saves 85 hr
- Total saved = 80 + 85 = 165 hr

Question: 62

The following graphic shows the results of an unsupervised, machine-learning clustering model:



k is the number of clusters, and n is the processing time required to run the model. Which of the following is the best value of k to optimize both accuracy and processing requirements?

A. 2

B. 10

C. 15

D. 20

Answer: B

Explanation:

The curve shows a steep drop in processing time up to about k = 10, after which gains in speed taper off. Choosing 10 clusters balances sufficient model complexity with reasonable computational cost.

Question: 63

Under perfect conditions, E. coli bacteria would cover the entire earth in a matter of days. Which of the following types of models is the best for explaining this type of growth?

A. Linear

B. Logarithmic

C. Polynomial

D. Exponential

Answer: D

Explanation:

Under ideal conditions, each E. coli cell divides into two in a fixed time interval, causing the population to double repeatedly - classic exponential growth.

Question: 64

Which of the following problem-solving approaches is a set of guidelines to handle highly variable and not fully apparent situations?

A. Schedule

B. Plan

C. Heuristic

D. Algorithm

Answer: C

Explanation:

Heuristics are rule-of-thumb strategies that guide problem solving in complex, uncertain situations where a fixed algorithm or plan isn't feasible.

Question: 65

A data analyst is examining the correlation matrix of a new data set to identify issues that could adversely impact model performance. Which of the following is the analyst most likely checking for?

A. Undersampling

B. Multicollinearity

C. Oversampling

D. Overfitting

Answer: B

Explanation:

Examining a correlation matrix helps identify predictors that are highly correlated with each other, which can inflate variance in coefficient estimates and degrade model reliability - i.e., multicollinearity.

Question: 66

A data scientist is designing a real-time machine-learning model that classifies a user based on initial behavior. The run times of these models are provided in the following table:

Model	Run time	Accuracy
Artificial neural network	12 minutes	95%
Decision trees	10 minutes	92%
Random forest	1 minutes	88%
XGBoost	5 minutes	90%

Which of the following models should the data scientist recommend for deployment?

A. XGBoost

B. Random forest

C. Decision trees

D. Artificial neural network

Answer: B

Explanation:

For a real-time application, inference latency is critical. Although its accuracy (88%) is slightly lower than the others, the random forest's 1-minute run time is by far the fastest, making it the only model capable of meeting real-time responsiveness.

Question: 67

A movie production company would like to find the actors appearing in its top movies using data from the tables below. The resulting data must show all movies in Table 1, enriched with actors listed in Table 2.

Table 1: Top Movies

■D	Movie	Year
1	Movie 1	2000
2	Movie 2	2010
3	Movie 3	2015
4	Movie 4	1990

Table 2: Actors

ID	Actor	Acted_In
W	Smith	Movie 3
11	Johnson	Movie 5
30	Taylor	Movie 1
50	Smith	Movie 7

Which of the following query operations achieves the desired data set?

A. Perform an INNER JOIN between Table 1 using column Movie, and Table 2 using column Acted_In.

B. Perform a UNION between Table 1 using column Movie, and Table 2 using column Acted_In.

C. Perform an INTERSECT between Table 1 using column Movie, and Table 2 using column Acted_In.

D. Perform a LEFT JOIN on Table 1 using column Movie, with Table 2 using column Acted_In.

Answer: D

Explanation:

A LEFT JOIN returns every row from Table 1 (all top movies) and brings in matching actors from Table 2 where the Movie = Acted_In, leaving NULLs for movies without listed actors.

Question: 68

A data scientist is preparing to brief a non-technical audience that is focused on analysis and results. During the modeling process, the data scientist produced the following artifacts:

- Charts and dashboards
- Model performance statistics (accuracy, precision, recall, F1 score, etc.)
- Mathematical descriptions of clustering algorithms included in the selected model
- Model selection, justification, and purpose
- Code documentation

- Data dictionary

Which of the following artifacts should the data scientist include in the briefing? (Choose two.)

A. Final charts and dashboards

B. Model selection, justification, and purpose

C. Code documentation

D. Mathematical descriptions of clustering algorithms included in the selected model

E. Model performance statistics (accuracy, precision, recall, F1_ score, etc.)

F. Data dictionary

Answer: A

Explanation:

For a non-technical audience centered on results, polished visualizations (charts and dashboards) and clear, high-level performance metrics (accuracy, precision, recall, F1 score) best convey the key takeaways. The deeper technical details, code docs, data dictionaries, and algorithm math, should be omitted at this level.

Question: 69

A data scientist has built a model that provides the likelihood of an error occurring in a factory. The historical accuracy of the model is 90%. At a specific factory, the model is reporting a likelihood score of 0.90. Which of the following explains a confidence score of 0.90?

A. Running this model for all known factory issues, it is expected the model will identify 90 out of 100 known factory issues.

B. Running this model on 100 samples of factories, a certain model performance is expected for 90 out of the 100 samples.

C. Running this model 100 times on a factory, it is expected the model will predict 90 out of 100 factory errors.

D. Running this model 100 times within a factory it is expected the model will predict error 90 out of 100 times the model is ran.

Answer: D

Explanation:

A confidence score of 0.90 is a probabilistic estimate, interpreted as the model assigning a 90% chance of an error on that particular factory instance, which in the long run corresponds to predicting "error" in about 90 out of every 100 identical runs.

Question: 70

An analyst is examining data from an array of temperature sensors and sees that one sensor consistently returns values that are much higher than the values from the other sensors. Which of the following terms best describes this type of error?

A. Synthetic

B. Systematic

C. Heteroskedastic

D. Idiosyncratic

Answer: B

Explanation:

A sensor that consistently reads higher than the others exhibits a repeatable bias, which is characteristic of a systematic error.

Question: 71

Which of the following is the naive assumption in Bayes' rule?

A. Normal distribution

B. Independence

C. Uniform distribution

D. Homoskedasticity

Answer: B

Explanation:
Naive Bayes assumes that all predictor variables are conditionally independent of each other given the class label, dramatically simplifying the joint probability calculation in Bayes' rule.

Question: 72

Which of the following types of machine learning is a GPU most commonly used for?

A. Deep learning/neural networks

B. Clustering

C. Natural language processing

D. Tree-based

Answer: A

Explanation:

GPUs excel at the massive parallelism required for the matrix and tensor operations at the heart of deep neural network training and inference, making them the go-to hardware for deep learning workloads.

Question: 73

A data scientist is attempting to identify sentences that are conceptually similar to each other within a set of text files. Which of the following is the best way to prepare the data set to accomplish this task after data ingestion?

A. Embeddings
B. Extrapolation
C. Sampling
D. One-hot encoding

Answer: A

Explanation:

Generating embeddings transforms each sentence into a dense numerical vector in a semantic space, where conceptually similar sentences lie close together, enabling straightforward similarity calculations (e.g., cosine similarity) to group or identify related sentences.

Question: 74

Which of the following distributions would be best to use for hypothesis testing on a data set with 20 observations?

A. Power law
B. Normal
C. Uniform
D. Student's t-

Answer: D

Explanation:

k

With only 20 observations and an unknown population variance, the t-distribution (with – 1 degrees of freedom) properly accounts for the extra uncertainty in the standard error when performing hypothesis tests.

Question: 75

Which of the following types of layers is used to downsample feature detection when using a convolutional neural network?

A. Pooling

B. Input

C. Output

D. Hidden

Answer: A

Explanation:

Pooling layers (such as max pooling or average pooling) reduce the spatial dimensions of the feature maps by summarizing local neighborhoods, effectively downsampling the detected features and controlling overfitting.

Question: 76

Which of the following image data augmentation techniques allows a data scientist to increase the size of a data set?

A. Clipping

B. Cropping

C. Masking

D. Scaling

Answer: B

Explanation:

By taking multiple crops from each original image (e.g., random or sliding-window crops), you generate distinct new training examples, directly increasing the dataset size.

Question: 77

A data scientist receives an update on a business case about a machine that has thousands of error codes. The data scientist creates the following summary statistics profile while reviewing the logs for each machine:

Number of machines observed	3.000,000
Number of unique error codes observed	19,000
Median number of unique codes observed per machine	7
Median number of error transactions observed per machine	45

Which of the following is the most likely concern with respect to data design for model ingestion?

A. Sparse matrix

B. Granularity misalignment

C. Insufficient features

D. Multivariate outliers

Answer: A

Explanation:

With 19,000 possible error-code features and each machine reporting only a handful (median of 7), your feature matrix will be extremely sparse (most entries zero) which can negatively impact both storage and model performance unless you address it (e.g., via sparse data structures or dimensionality reduction).

Question: 78

A company created a very popular collectible card set. Collectors attempt to collect the entire set, but the availability of each card varies, with because some cards have higher production volumes than others. The set contains a total of 12 cards. The attributes of the cards are below:

Card number	Wrapper color	Wrapper shape	Animal	Habitat
1	Red	Diamond	Dog	Land
2	Red	Triangle	Whale	Sea
3	Red	Diamond	Fish	Sea
4	Red	Triangle	Shark	Sea
5	Red	Diamond	Elephant	Land
6	Red	Triangle	Squid	Sea
	Black	Diamond	Bird	Land
3	Black	Triangle	Horse	Land
9	Black	Diamond	Octopus	Sea
10	Black	Triangle	Clam	Sea
11	Black	Diamond	Bear	Land
12	Black	Triangle	Lion	Land

A data scientist is provided a historical record of cards purchased, which was acquired by a local collectors' association. The data scientist needs to design an initial model iteration to predict whether or not the animal on the card lives in the sea or on land given the provided attributes. Which of the following is the best way to accomplish this task?

A. ARIMA

B. Linear regression

C. Association rules

D. Decision trees

Answer: D

Explanation:

You have categorical inputs (wrapper color, shape, animal) and a binary target (sea vs. land). A decision tree natively handles categorical features and yields clear, rule-based splits that predict habitat, making it the most appropriate choice.

Question: 79

A data scientist would like to model a complex phenomenon using a large data set composed of categorical, discrete, and continuous variables. After completing exploratory data analysis, the data scientist is reasonably certain that no linear relationship exists between the predictors and the target. Although the phenomenon is complex, the data scientist still wants to maintain the highest possible degree of interpretability in the final model. Which of the following algorithms best meets this objective?

A. Artificial neural network

B. Decision tree

C. Multiple linear regression

D. Random forest

Answer: B

Explanation:

Decision trees capture complex, nonlinear relationships with a transparent, rule-based structure. They remain highly interpretable (each split can be visualized and explained) unlike ensembles (random forests) or neural networks, and they don't rely on linear assumptions.

Question: 80

Which of the following is best solved with graph theory?

A. Optical character recognition

B. Traveling salesman

C. Fraud detection

D. One-armed bandit

Answer: B

Explanation:

The traveling-salesman problem is a prototypical graph theory challenge, finding the shortest tour through a graph's nodes, whereas the other tasks rely on different domains (OCR on image processing, fraud detection often on statistical/anomaly methods, bandit problems on sequential decision theory).

Question: 81

Given these business requirements:

Needs to most efficiently move 3,000 boxes across a river

Has one boat that holds eight boxes, travels at ten nautical miles per hour, and has a fuel economy of six nautical miles per gallon

- Has another boat that holds two boxes, travels at 50 nautical miles per hour, and has a fuel economy of 18 nautical miles per gallon
- The river is one nautical mile wide
- The data scientist only has access to 125 gallons of fuel

Which of the following is the most likely optimization technique a data scientist would apply?

A. Constrained

B. Unconstrained

C. Non-iterative

D. Iterative

Answer: A

Explanation:

You must optimize boat trips subject to strict resource limits (fuel, boat capacity, travel distance), making this a constrained optimization problem (e.g., solvable via linear programming).

Question: 82

A data scientist is analyzing a data set with categorical features and would like to make those features more useful when building a model. Which of the following data transformation techniques should the data scientist use? (Choose two.)

A. Normalization

B. One-hot encoding

C. Linearization

D. Label encoding

E. Scaling

F. Pivoting

Answer: B

Explanation:

One-hot encoding creates binary indicator columns for each category, allowing models to treat nominal categories without implying any order.

Label encoding maps categories to integer labels, which can be useful for tree-based models or when you need a single numeric column (though you must ensure the algorithm can handle treated ordinality appropriately).

Question: 83

Which of the following measures would a data scientist most likely use to calculate the similarity of two text strings?

A. Word cloud

B. Edit distance

C. String indexing

D. k-nearest neighbors

Answer: B

Explanation:

Edit distance quantifies how many single-character insertions, deletions, or substitutions are needed to transform one string into another, making it a direct measure of their similarity.

Question: 84

Which of the following issues should a data scientist be most concerned about when generating a synthetic data set?

A. The data set consuming too many resources

B. The data set having insufficient features

C. The data set having insufficient row observations

D. The data set not being representative of the population

Answer: D

Explanation:

If synthetic data don't accurately mirror the real-world distributions and relationships, any models trained on them will perform poorly in deployment. Representativeness is the critical concern when generating synthetic data.

Question: 85

A data scientist is performing a linear regression and wants to construct a model that explains the most variation in the dat

a. Which of the following should the data scientist maximize when evaluating the regression

performance metrics?

A. Accuracy

B. R2

C. p value

D. AUC

Answer: B

Explanation: