DUMPSBOSS.

CompTIA Data Science Certification Exam

CompTIA DY0-001

Version Demo

Total Demo Questions: 10

Total Premium Questions: 85

Buy Premium PDF

https://dumpsboss.co

support@dumpsboss.co

support@dumpsboss.co
dumpsboss.co

QUESTION NO: 1 - (SIMULATION)

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

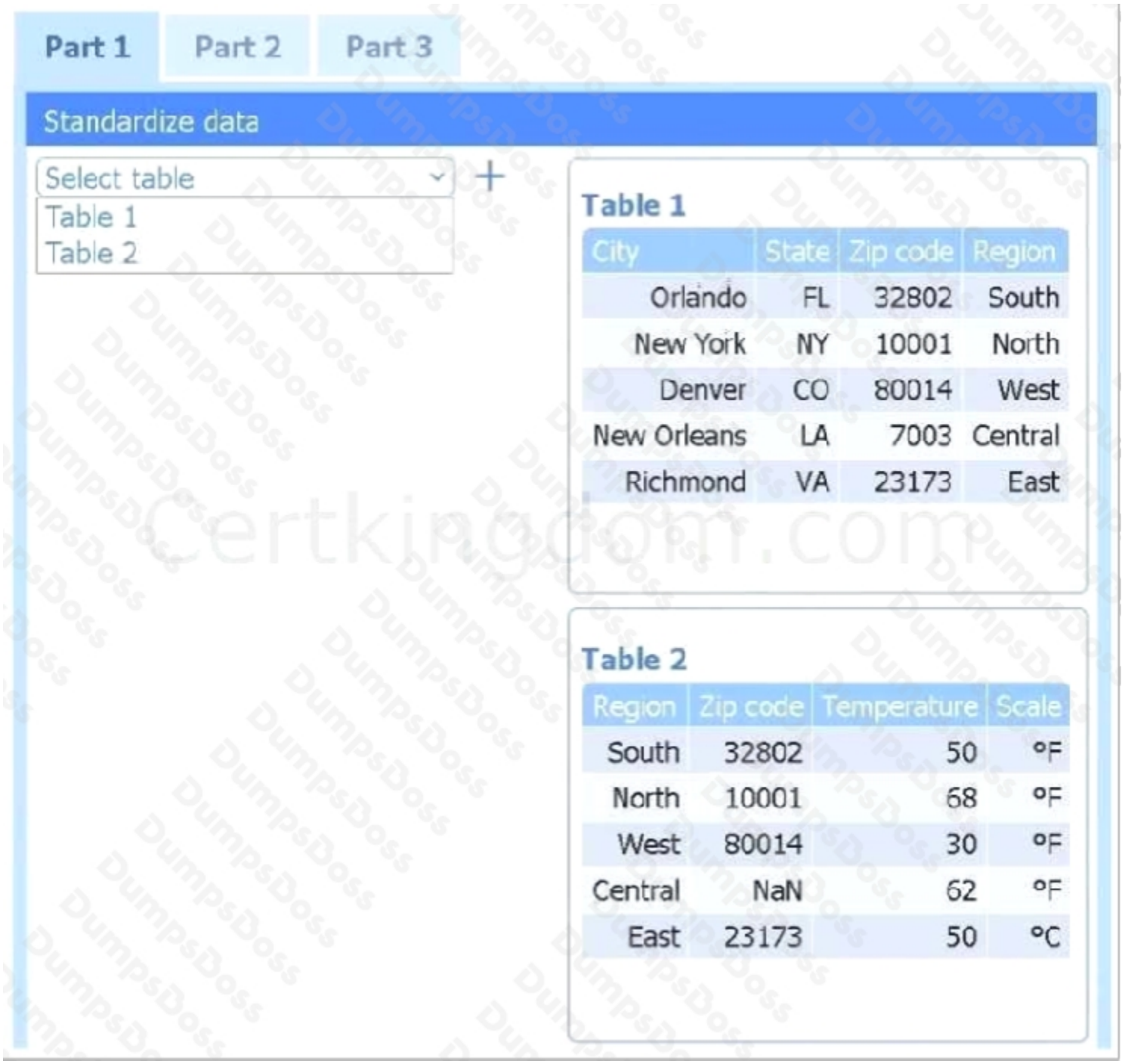
Review the charts provided and use the drop-down menu to select the most appropriate way to standardize the data.

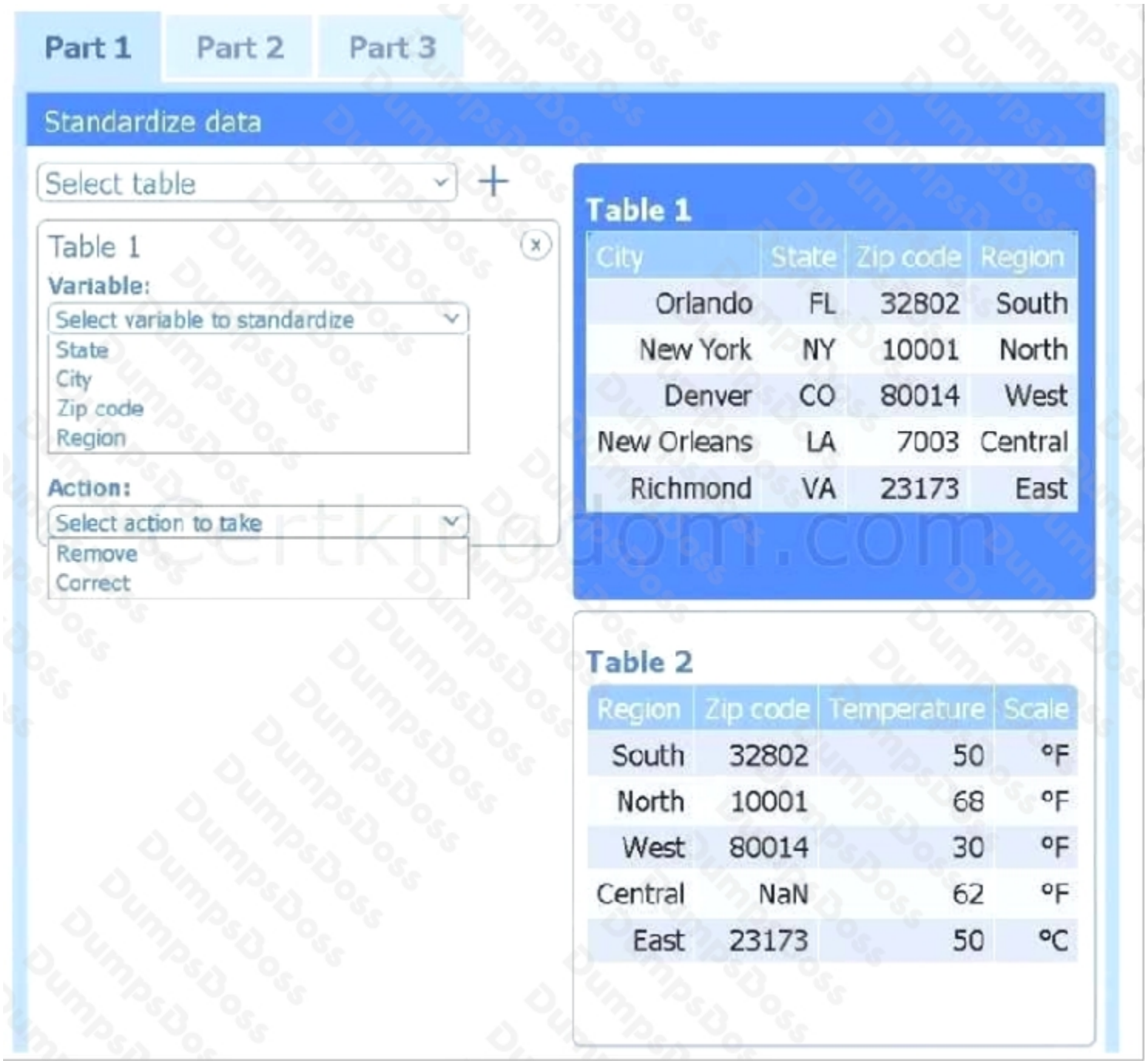
Part 2

Answer the questions to determine how to create one data set. Part 3

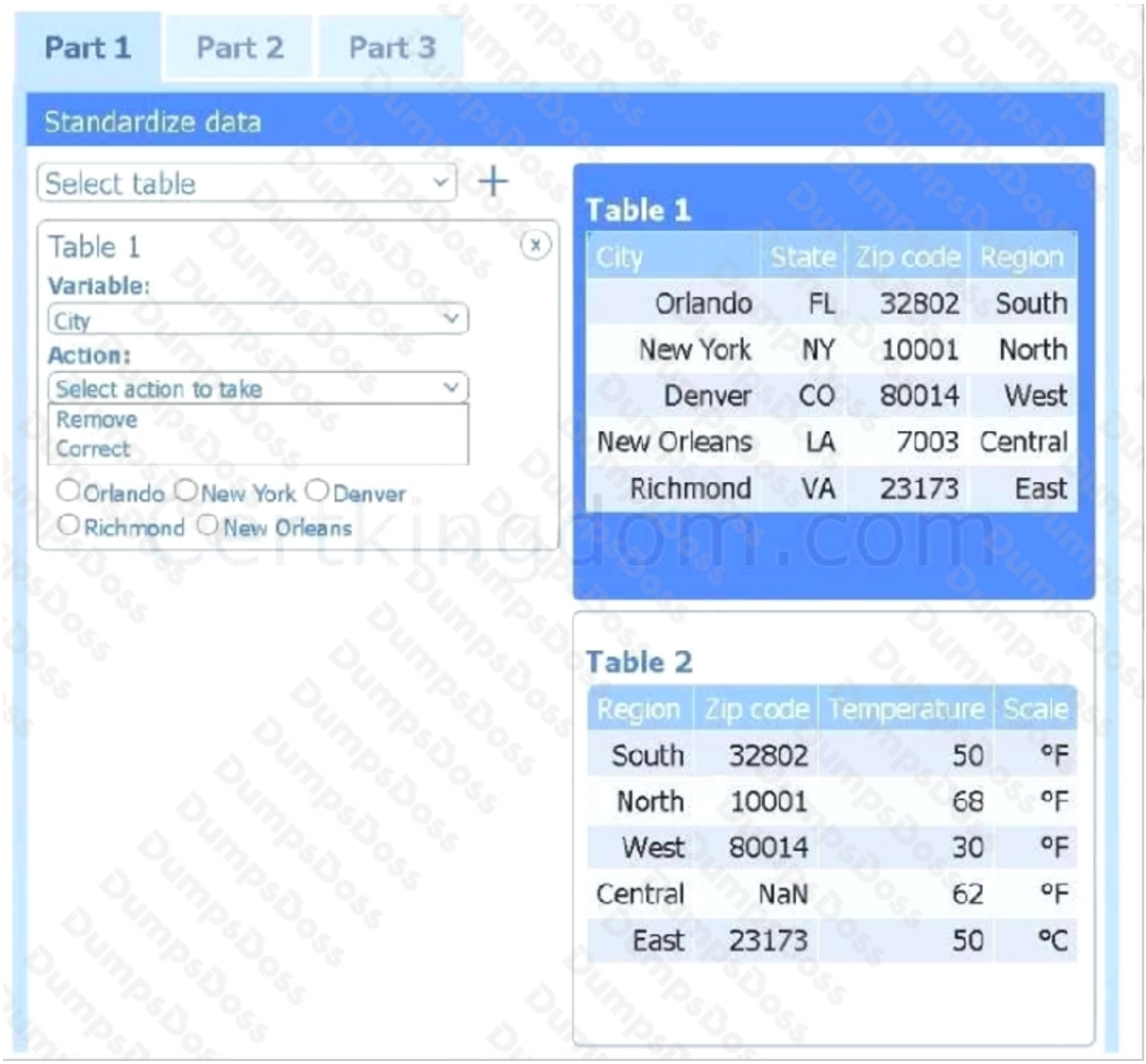
Select the most appropriate visualization based on the data set that represents what the client is looking for.

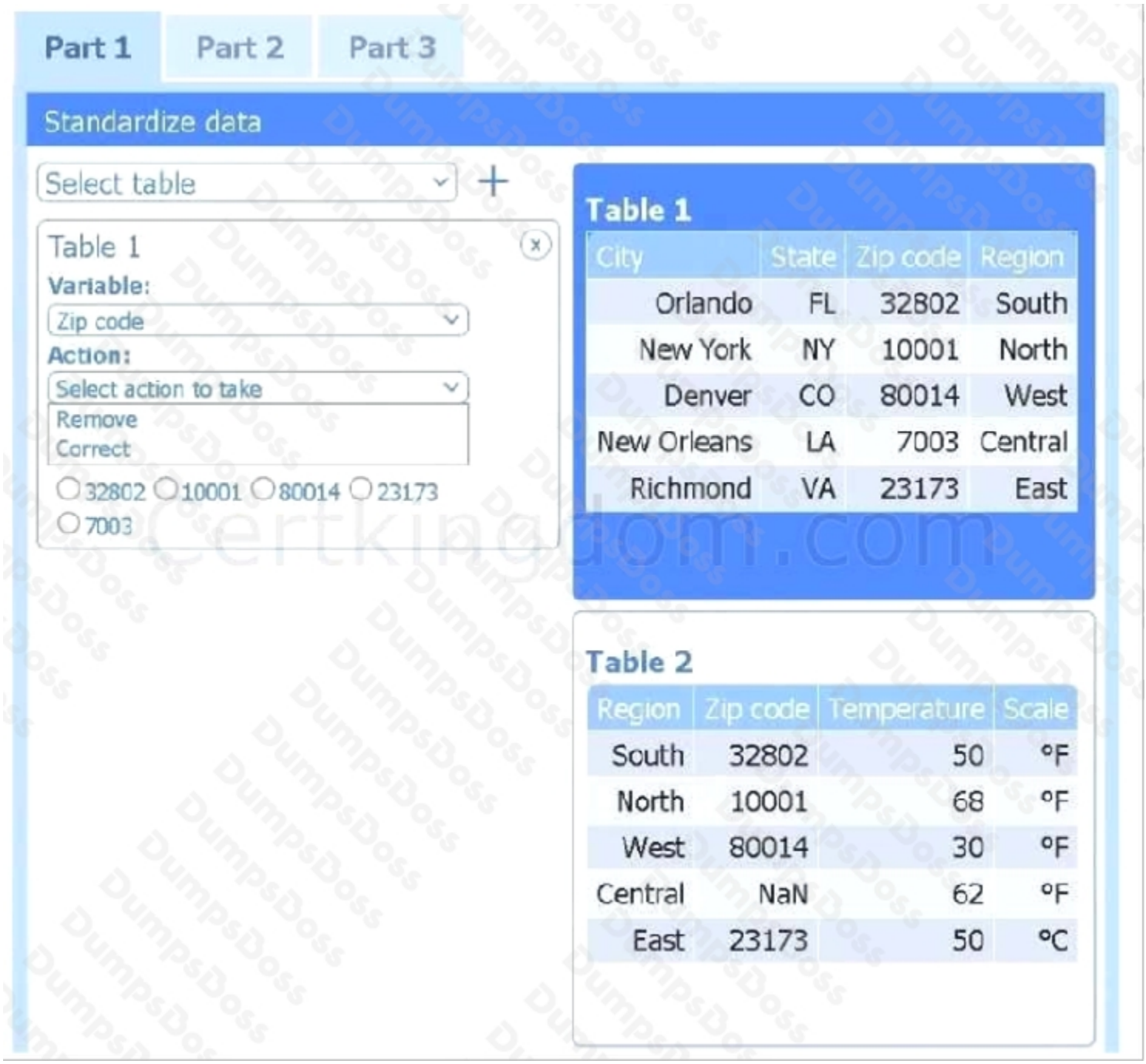
If at any time you would like to bring back the initial state of the simulation, please click the Reset All button.

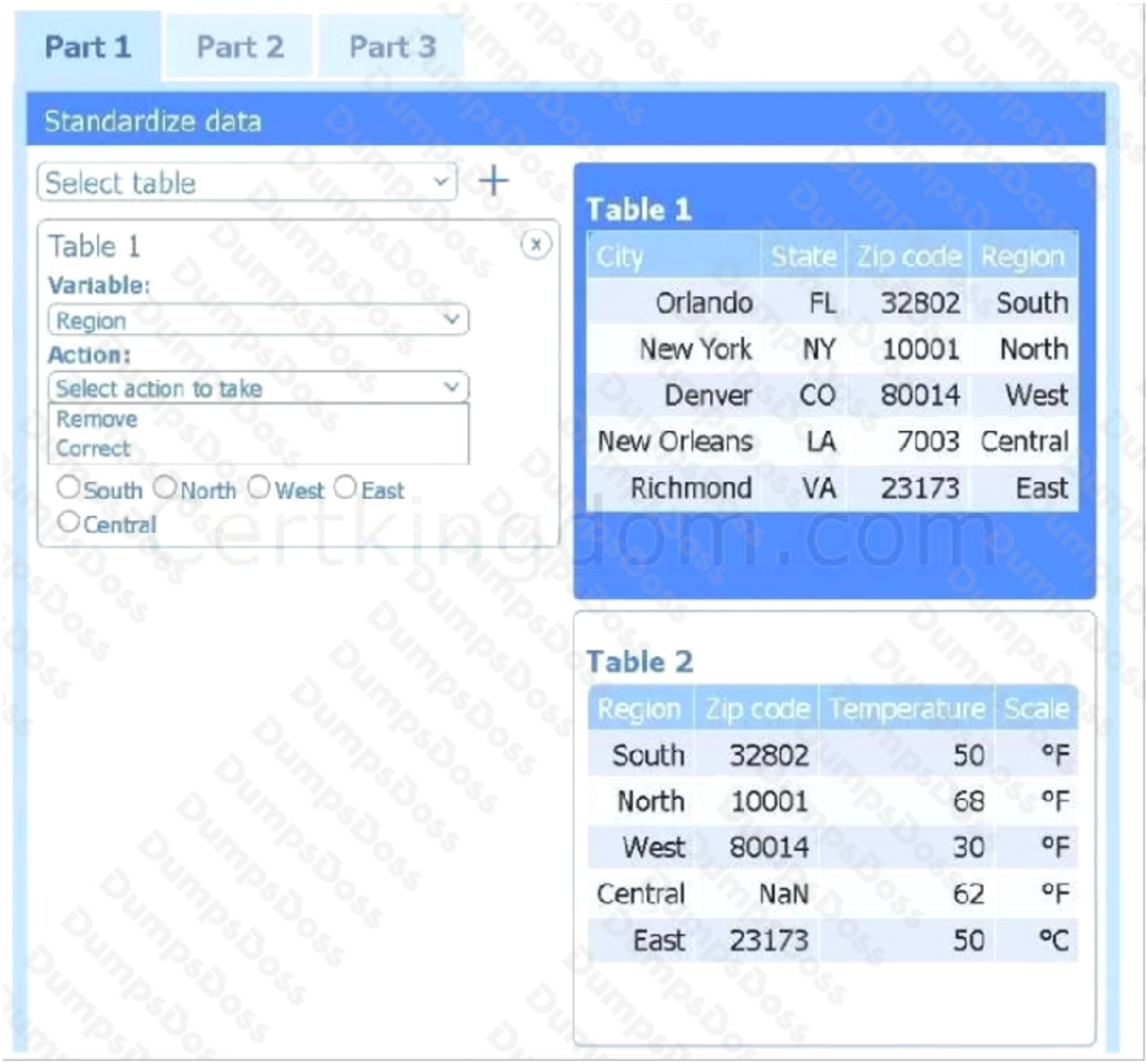

Part 1
Part 2
Part 3
Standardize data
Select table
Table 1
Table 2
Table 1
City State Zip code Region
Orlando FL 32802 South
New York NY 10001 North
Denver CO 80014 West
New Orleans LA 7003 Central
Richmond VA 23173 East
Table 2
Region Zip code Temperature Scale
South 32802 50 °F
North 10001 68 °F
West 80014 30 °F
Central NaN 62 °F
East 23173 50 °C

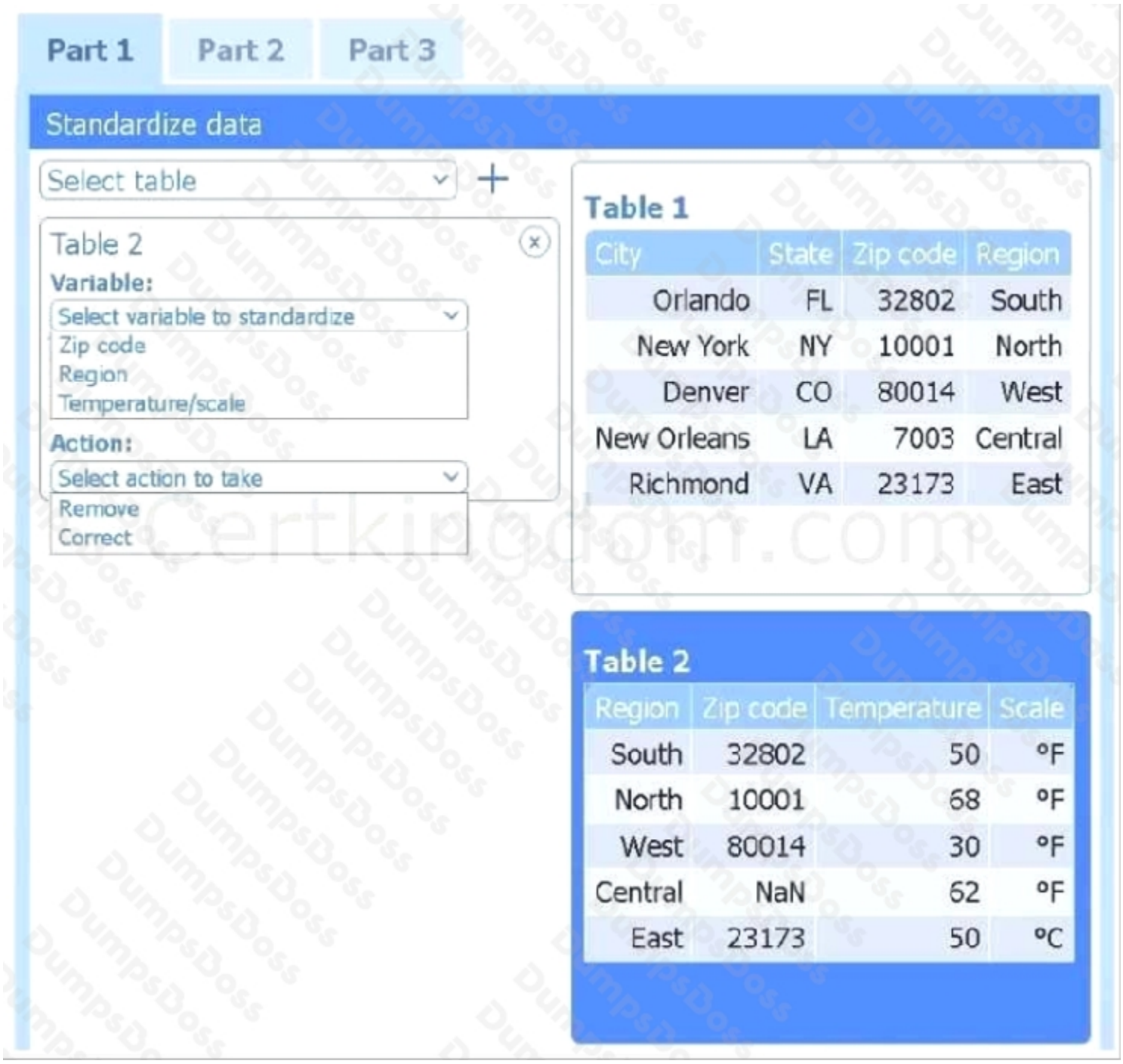

Part 1
Part 2
Part 3
Standardize data
Select table
Table 1
Variable:
Select variable to standardize
State
City
Zip code
Region
Action:
Select action to take
Remove
Correct
Table 1
City State Zip code Region
Orlando FL 32802 South
New York NY 10001 North
Denver CO 80014 West
New Orleans LA 7003 Central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

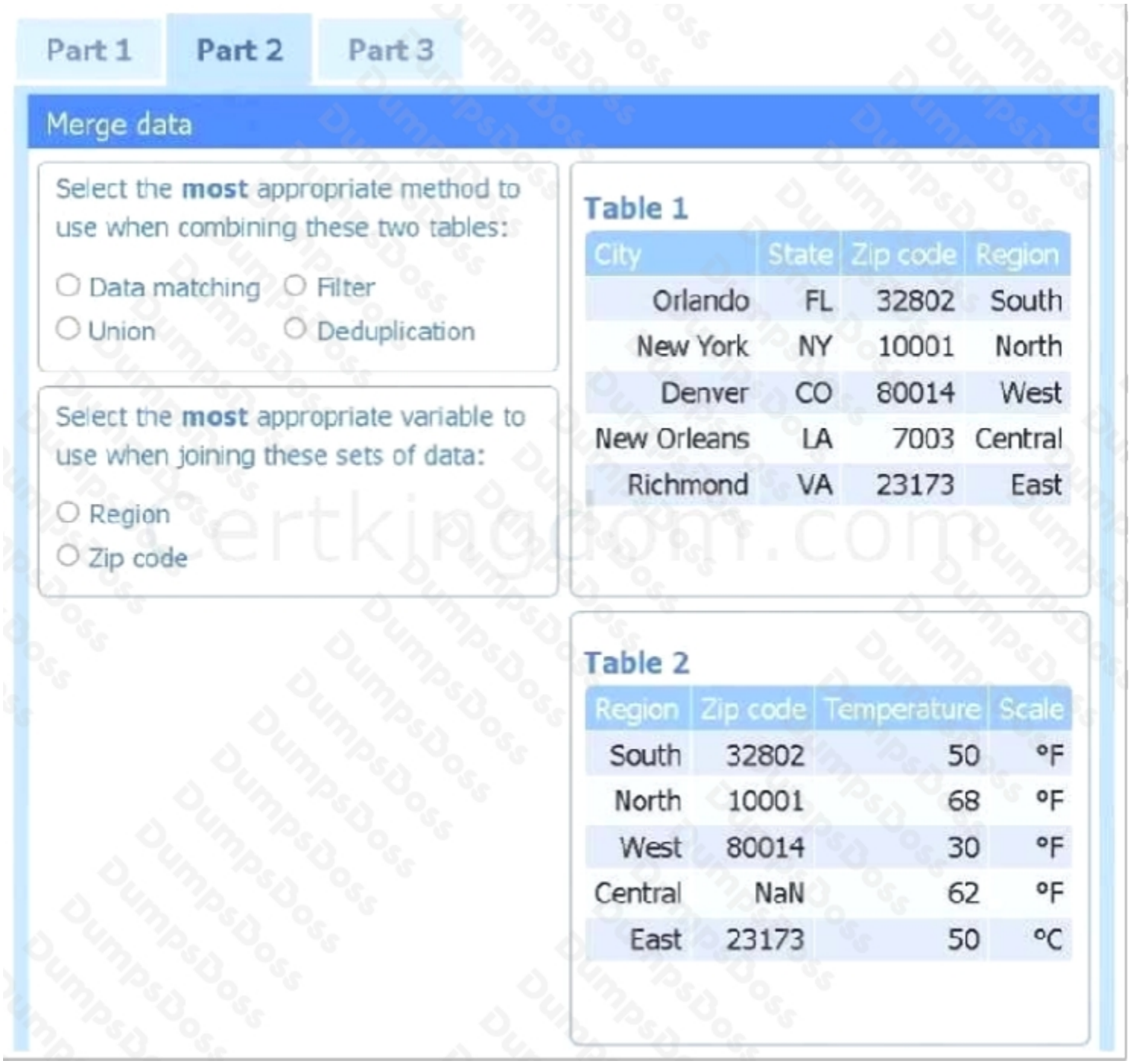

Part 1
Part 2
Part 3
Standardize data
Select table
Table 1
Variable:
Region
Action:
Select action to take
Remove
Correct
South
North
West
East
Central
Table 1
City State Zip code Region
Orlando FL 32802 South
New York NY 10001 North
Denver CO 80014 West
New Orleans LA 7003 Central
Richmond VA 23173 East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

- ○ Data matching
- ○ Filter
- ○ Union
- ○ Deduplication

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

- ○ Region
- ○ Zip code

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

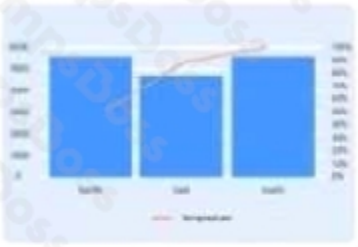
Region	Zip code	Temperature	Scale
South	32802	50	°F
North	10001	68	°F
West	80014	30	°F
Central	NaN	62	°F
East	23173	50	°C

Part 1 | Part 2 | Part 3

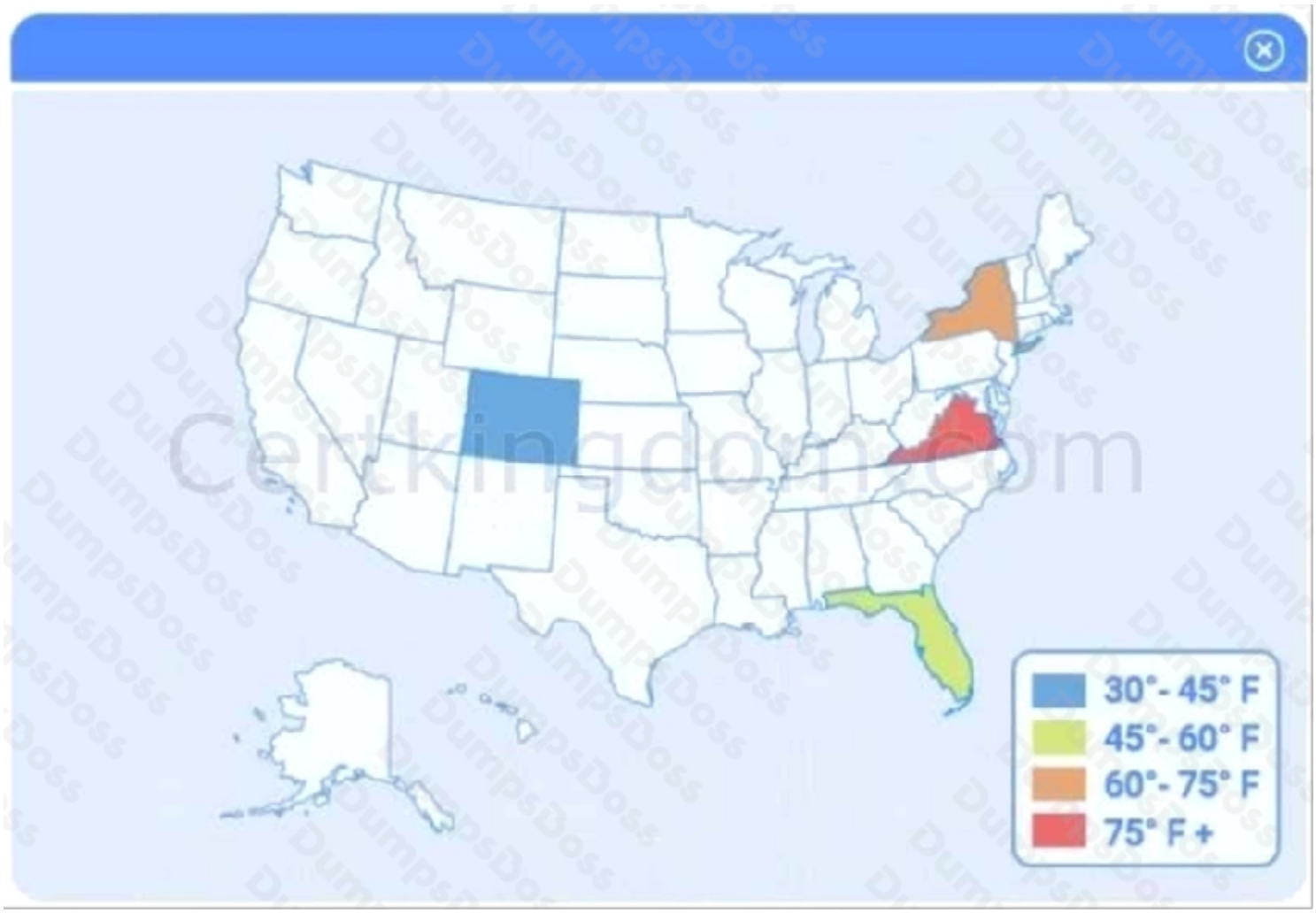
Visualization

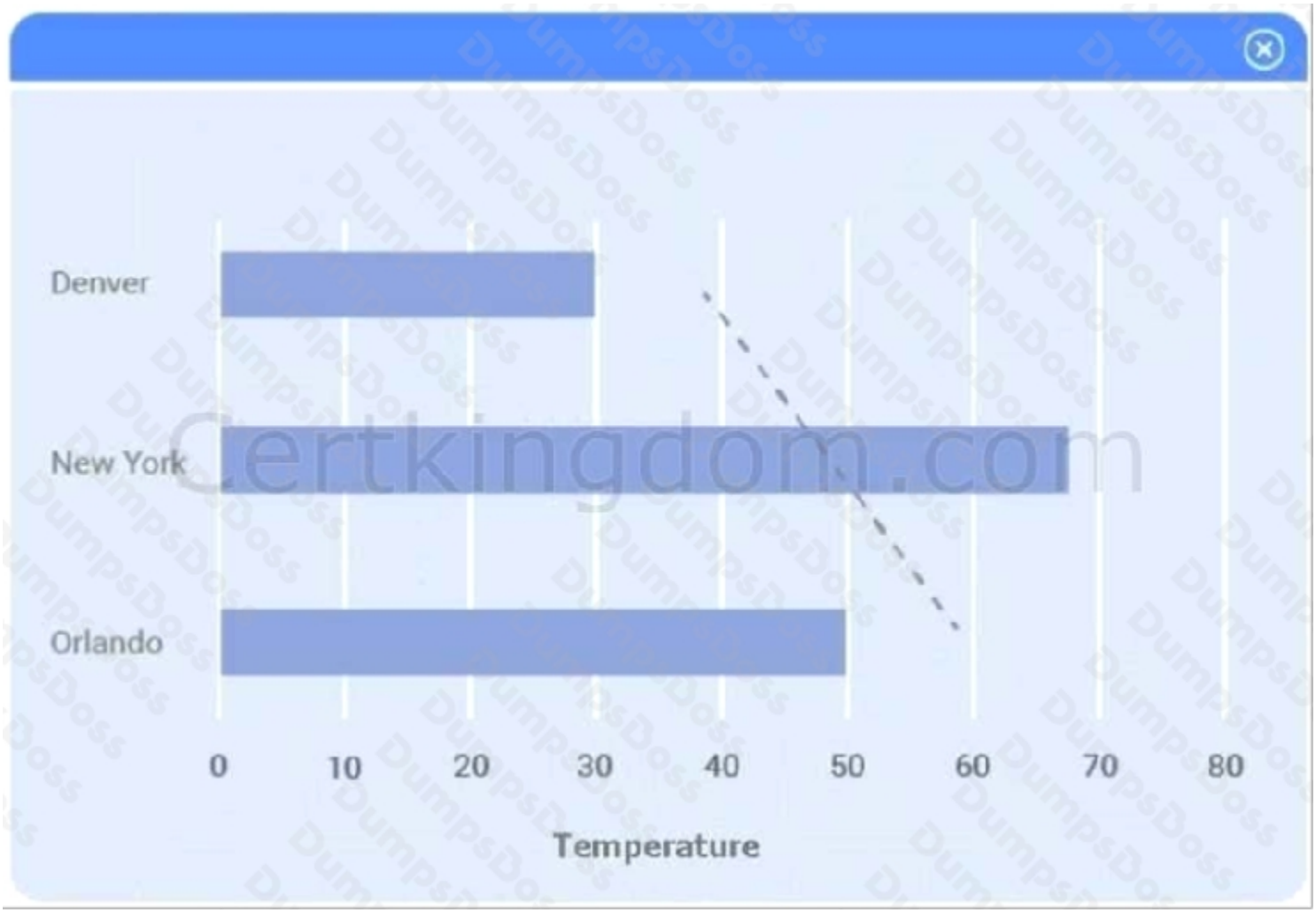
Select the **most** appropriate visualization based on the data set which represents what the client is looking for:

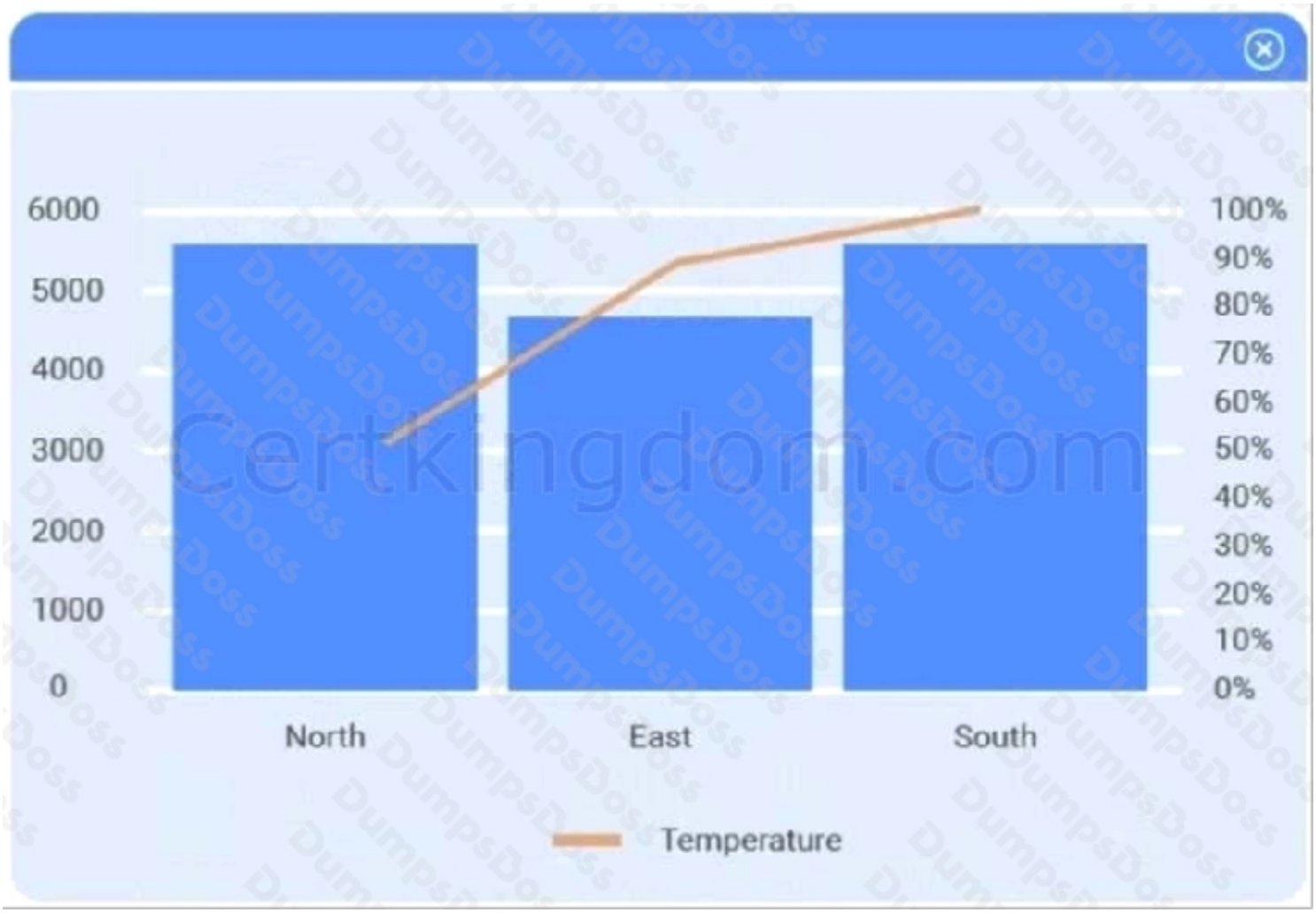


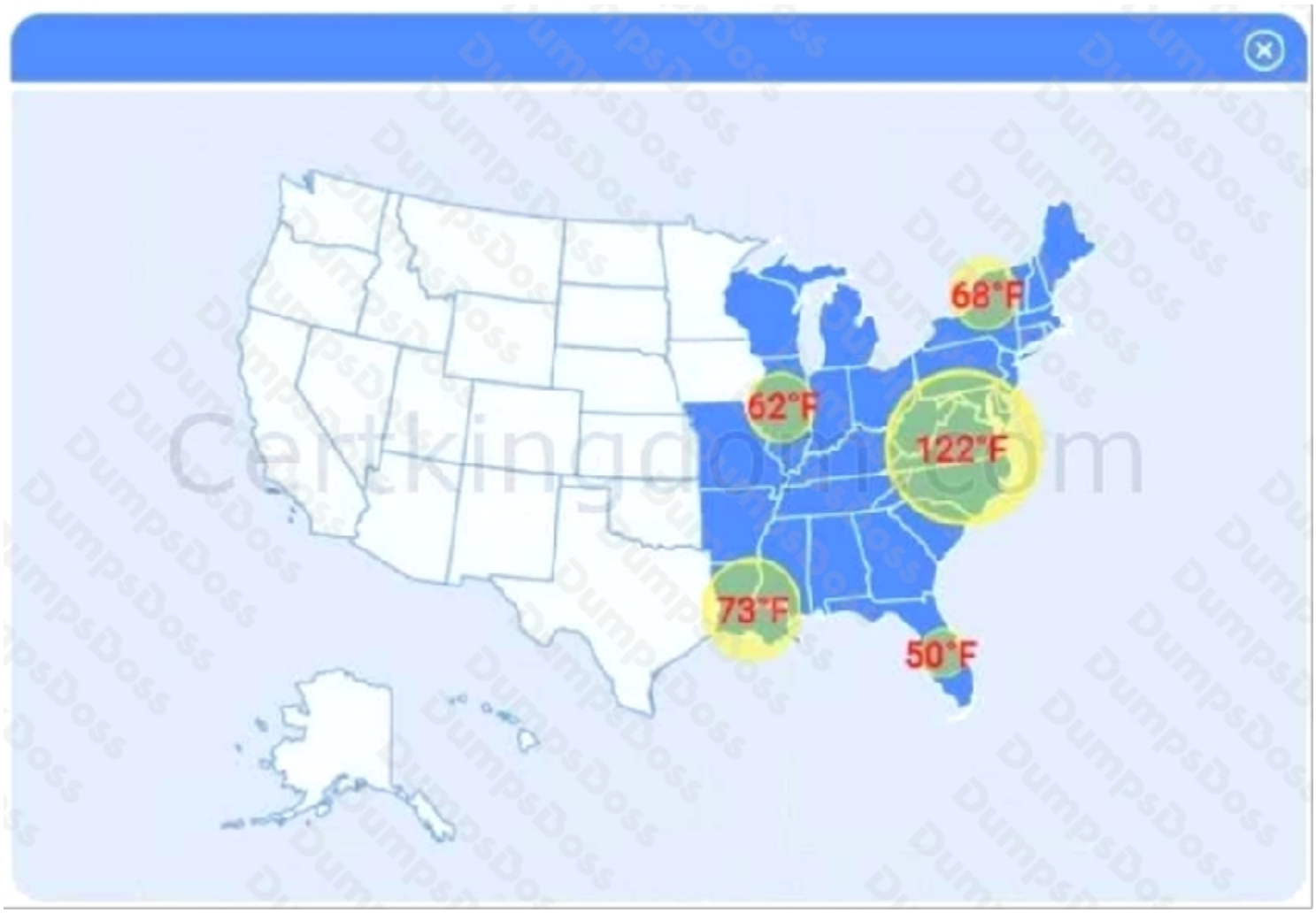


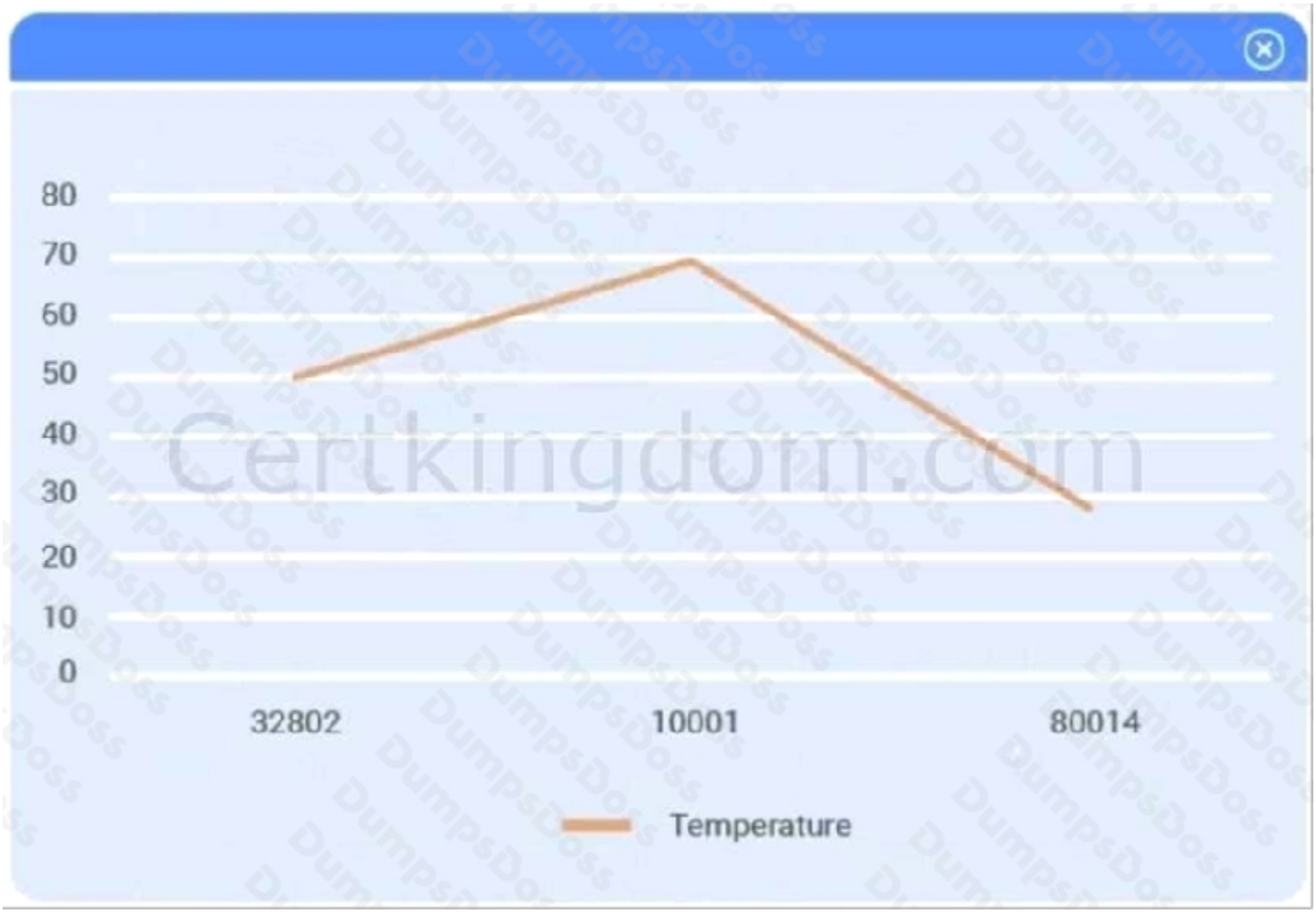
Region	City	State	Zip code	Temperature	Scale
South	Orlando	FL	32802	50	°F
North	New York	NY	10001	68	°F
West	Denver	CO	80014	30	°F
Central	New Orleans	LA	7003		
East	Richmond	VA	23173	50	°C
Central			NaN	62	°F


30°- 45° F
45°- 60° F
60°- 75° F
75° F +






68°F
62°F
122°F
73°F
50°F



ANSWER: See explanation below.

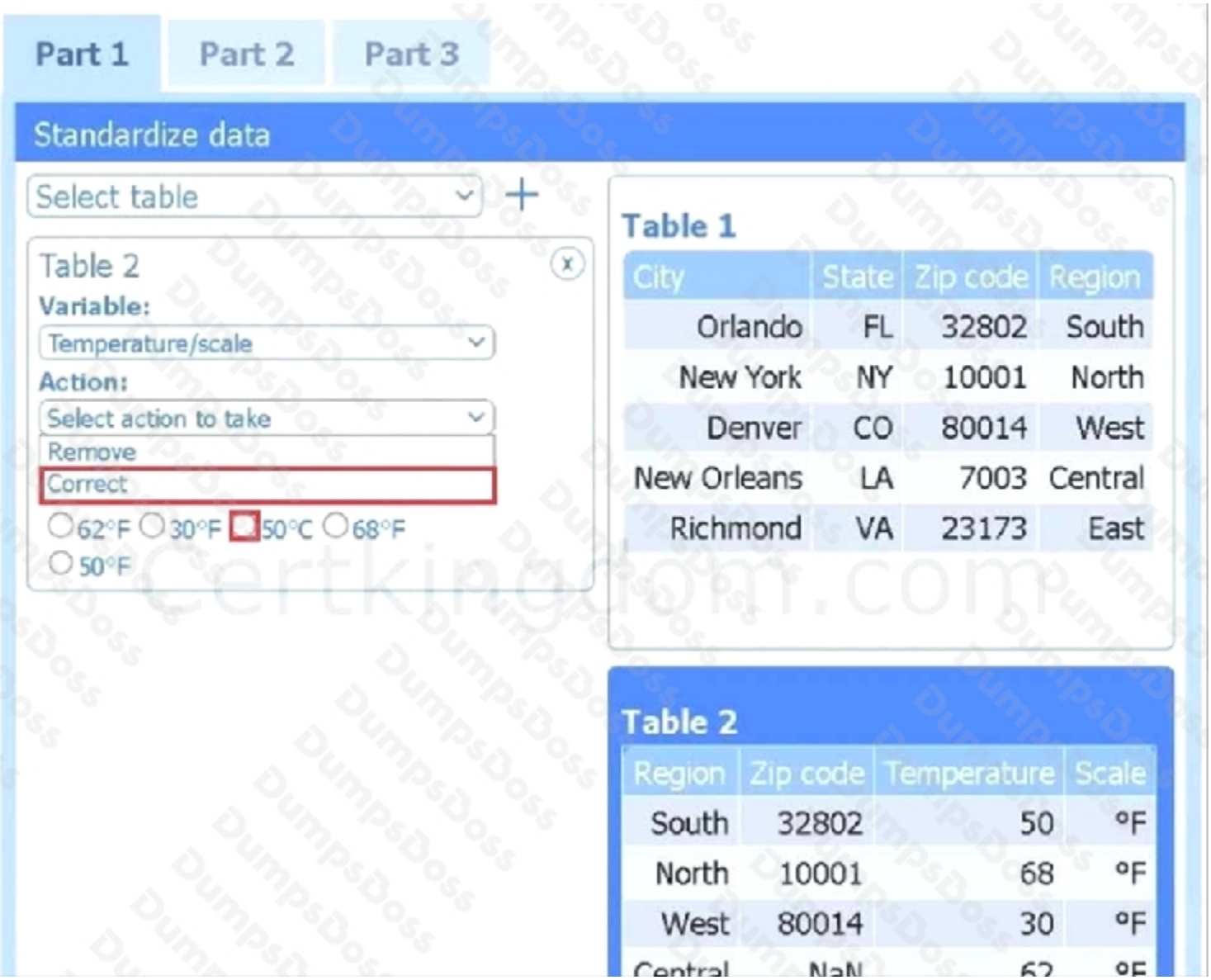
Explanation:

Part 1

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

before visualization. Variable: Temperature/scale Action: Correct

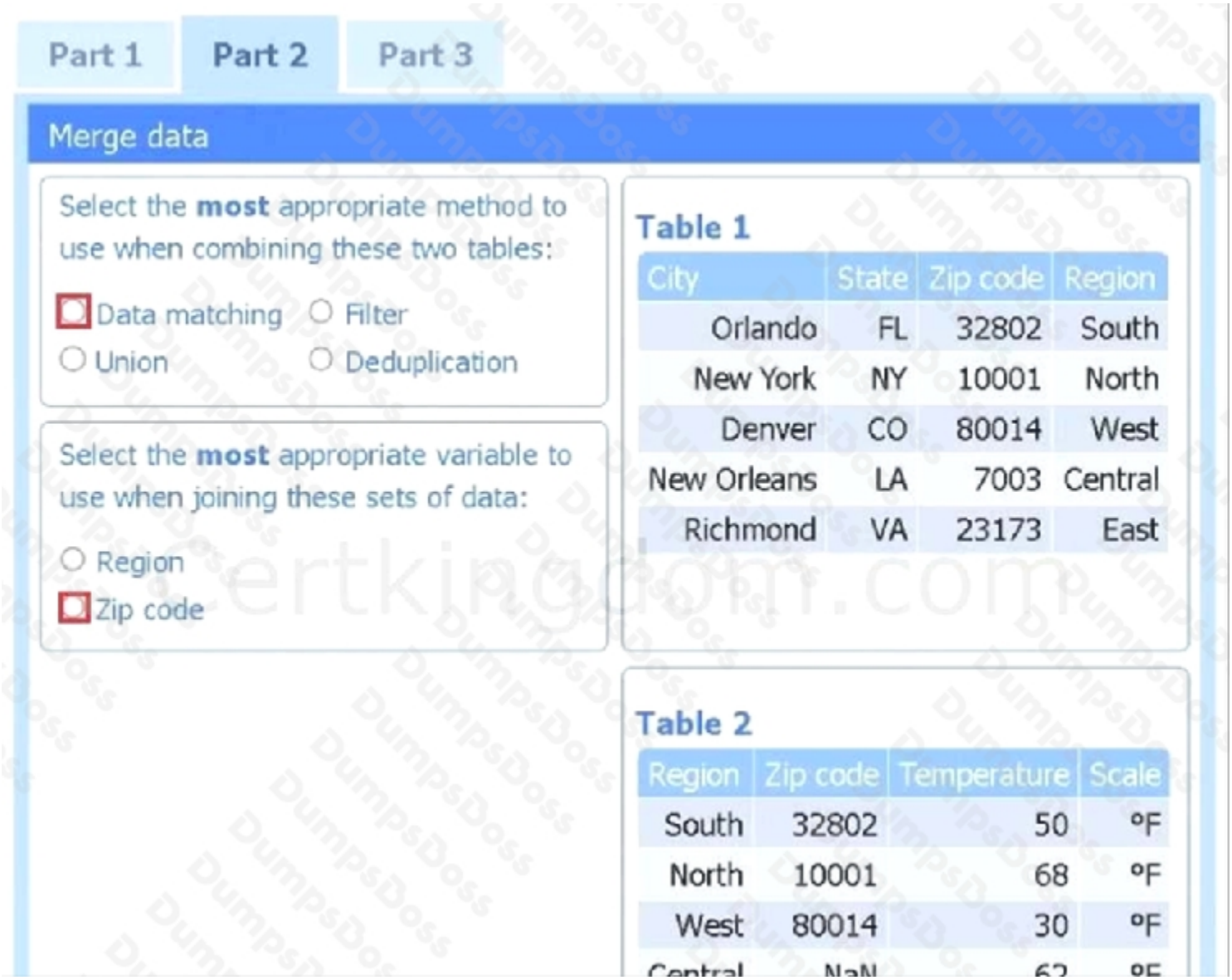
Value to correct: 50 Â°C



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 regions weather reading to its city.



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.

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		

QUESTION NO: 2

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:

In LASSO (Least Absolute Shrinkage and Selection Operator) linear regression, the objective function is the ordinary least squares (OLS) loss plus an L1 regularization term on the coefficients. The key point is that the *residual/loss term* remains the same as OLS: it minimizes the *sum of squared residuals* (i.e., squared errors). LASSO changes the optimization by adding a penalty proportional to the sum of the absolute values of the coefficients (the L1 norm), which encourages sparsity (some coefficients become exactly zero), but it does not replace the squared-error residual term with absolute error.

Therefore, the residual term being minimized is best represented as e^2 (squared residual), not |e| or e. Option C (0) is not a description of a residual minimization term; while the goal is to drive residuals toward zero, the minimized quantity in the loss function is not literally "0" but the squared residuals aggregated across observations. Option A (|e|) corresponds more closely to least absolute deviations (L1 loss) regression, which is a different loss function than standard LASSO uses.

References: scikit-learn: Lasso, Wikipedia: Lasso (statistics).

QUESTION NO: 3

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:

The best answer is **adding nodes to a cluster deployment** because distributed computing frameworks (e.g., Spark, Dask, Ray, distributed training) can increase the total available memory by scaling out. When a job is memory-constrained, adding worker nodes increases aggregate RAM and often increases the number of executors/workers that can hold partitions, model state, or intermediate results in memory. This is the most direct "environmental" change that addresses memory pressure without requiring a redesign of the model logic.

Option A (containerizing) changes packaging and isolation, but it doesn't inherently increase available memory; containers still run on the same underlying hosts and are limited by host resources and configured limits. Option B (migrating to cloud) can help only if you also choose larger instances or scale out; simply moving environments doesn't automatically fix memory constraints. Option C (edge deployment) typically reduces available resources compared to a data center/cluster and is more likely to worsen memory issues.

References: Apache Spark Cluster Overview, Kubernetes: Managing Resources for Containers.

QUESTION NO: 4

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:

The best choice is **REST** because RESTful APIs are widely supported across languages and platforms and are typically consumed using standard HTTP methods (GET/POST/PUT/DELETE) with lightweight payloads such as JSON. That combination usually means other departments can integrate quickly using common tooling (curl, Postman, browser-based clients, standard HTTP libraries) without needing specialized frameworks or strict contracts. This aligns well with the requirement for broad access and minimal development effort.

SOAP is generally heavier-weight: it relies on XML envelopes and often WSDL-based contracts, which can increase integration complexity and tooling requirements. **RPC** (as a general style) can be efficient but often requires tighter coupling to specific protocols or client stubs, which can raise the barrier for diverse teams. **JSON** is not an API type by itself; it's a data interchange format commonly used *within* REST APIs, so it doesn't directly answer the question about which API to use.

References: Red Hat: What is a REST API?, MDN: HTTP request methods

QUESTION NO: 5

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

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:

To build a linear regression model that explains the most variation in the dependent variable, the data scientist should maximize **Rb2 (coefficient of determination)**. Rb2 measures the proportion of variance in the outcome that is explained by the predictors in the model. An Rb2 closer to 1 indicates the model accounts for more of the variability in the data, which directly matches the goal stated in the question.

Accuracy and **AUC** are classification metrics, not regression metrics, so they dont apply to evaluating a linear regression models explained variance. AUC (area under the ROC curve) specifically evaluates how well a binary classifier separates classes across thresholds. The **p-value** is used for hypothesis testing (e.g., whether a coefficient is statistically different from zero) and is not a measure of overall explained variation; smaller p-values can indicate statistical significance, but they dont necessarily mean the model explains more variance.

Therefore, maximizing Rb2 is the correct choice when the objective is to explain the most variation in a linear regression context.

References: https://en.wikipedia.org/wiki/Coefficient_of_determination, https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd44.htm

QUESTION NO: 6

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:

The best choice is **LASSO** because it performs *embedded feature selection*. LASSO (Least Absolute Shrinkage and Selection Operator) adds an **L1 regularization** penalty to the loss function. As the penalty increases, LASSO not only shrinks coefficients, but can drive some coefficients to **exactly zero**. Coefficients that become zero effectively remove those predictors from the model, leaving only the predictors with the strongest influence on the outcome—exactly what the question is asking for.

OLS (ordinary least squares) fits all predictors and does not include any built-in mechanism to drop less-influential variables. **Ridge** regression uses **L2 regularization**, which shrinks coefficients but typically does *not* set them to zero, so it reduces variance but doesn't truly select a subset of predictors. **Weighted least squares** is used when observations have non-constant variance (heteroscedasticity) and addresses weighting, not feature selection.

References: scikit-learn: Lasso, Wikipedia: Lasso (statistics).

QUESTION NO: 7

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:

Correct answer: C. A linear regression

The dependent variable is explicitly described as a real-number (continuous) outcome, and there is a strong relationship visible in a scatter plot between the single predictor and the outcome. For an inferential, easily interpretable model in this setting, *simple linear regression* is typically the best fit: it provides a straightforward coefficient (slope) that quantifies the expected change in the dependent variable per unit change in the predictor, along with confidence intervals and hypothesis tests that support inference. The note that the predictor is approximately normally distributed with few outliers further supports using a standard linear model (while normality of the predictor isn't a strict requirement, fewer outliers reduces leverage issues and improves stability of estimates).

Option A (logistic regression) and option D (probit regression) are designed for *binary* or categorical outcomes (modeling probabilities), not a continuous real-valued dependent variable. Option B (exponential regression) could be appropriate if the relationship is clearly nonlinear and exponential in shape, but the prompt emphasizes interpretability and does not indicate an exponential pattern—linear regression is the default interpretable inferential choice for a strong, approximately linear relationship.

References: https://en.wikipedia.org/wiki/Linear_regression, https://en.wikipedia.org/wiki/Logistic_regression

QUESTION NO: 8

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:

The correct answer is **Euclidean**. Euclidean distance is the classic "straight-line" (as-the-crow-flies) distance between two points in Euclidean space. In KNN, when features are represented as coordinates in an n-dimensional space, Euclidean distance measures the length of the direct line segment connecting two observations, which matches the question's description.

Manhattan distance is not straight-line; it measures distance along axis-aligned paths (like navigating a city grid), summing absolute differences across dimensions. **Cosine** is also not straight-line distance; it measures the angle (similarity in direction) between vectors and is often used for text embeddings or high-dimensional sparse data. **Radial** is not a standard named distance metric for KNN in the way the others are; "radial basis" typically refers to kernel functions (e.g., RBF) rather than a primary distance metric description.

References: https://en.wikipedia.org/wiki/Euclidean_distance, https://en.wikipedia.org/wiki/Taxicab_geometry

QUESTION NO: 9

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 (PCA) is primarily used for **dimensionality reduction**. It takes a dataset with potentially many correlated numeric features and transforms it into a smaller number of new features (principal components) that are *uncorrelated* and ordered by how much variance they explain. This is useful when you want to reduce feature count to speed up training, mitigate multicollinearity, reduce noise, or enable visualization (e.g., projecting high-dimensional data into 2D/3D) while retaining most of the information (variance) in the original data.

Option A is correct because it directly matches PCA's core purpose: reducing dimensionality while preserving as much variance as possible. Option B (Classification) and Option C (Regression) are supervised learning tasks; PCA is not a classifier or regressor, though it can be used as a preprocessing step before those models. Option D (Recommendation) is typically handled with collaborative filtering, matrix factorization, or embedding-based methods; while PCA-like techniques can sometimes be applied to user-item matrices, "recommendation" is not the standard or most appropriate primary use case description for PCA in this context.

References: scikit-learn PCA documentation, Wikipedia: Principal component analysis.

QUESTION NO: 10

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:

The best choice here is **XGBoost** because it supports *continued training* (often described as incremental training / warm start) by adding additional boosting rounds to an existing model using new data. That capability aligns with the requirement that the model "automatically adjust its parameters" as the underlying reporting-agency data changes over time. In practice, you can load an existing booster and continue training so the ensemble adapts without starting from scratch each time, which is a common approach for handling evolving patterns (concept drift) in tabular credit-risk style data.

A **decision tree** is typically trained in one shot; updating it with new data usually means rebuilding the tree. A **random forest** similarly consists of many independently trained trees; most standard implementations don't support true incremental updates of the existing forest (you generally retrain or build a new forest). **Linear discriminant analysis (LDA)** is primarily a classification technique with strong distributional assumptions; while parameters can be recomputed, it's not the best fit for

ongoing adaptive learning in this context and is not commonly used for credit score prediction compared to gradient-boosted trees.

References: XGBoost documentation (saving/loading and continued training), scikit-learn ensemble methods overview.