DUMPSARENA

CompTIA Data Science Certification Exam

CompTIA DY0-001

Version Demo

Total Demo Questions: 8

Total Premium Questions: 126

Buy Premium PDF

https://dumpsarena.co

sales@dumpsarena.co

sales@dumpsarena.co
dumpsarena.co

Topic Break Down

Topic	No. of Questions
Topic 1, Data Concepts and Environments	17
Topic 2, Data Mining	28
Topic 3, Data Analysis	61
Topic 4, Visualization	12
Topic 5, Data Governance, Quality, and Controls	7
Topic 6, Mix Questions	1
Total	126

QUESTION NO: 1 - (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.

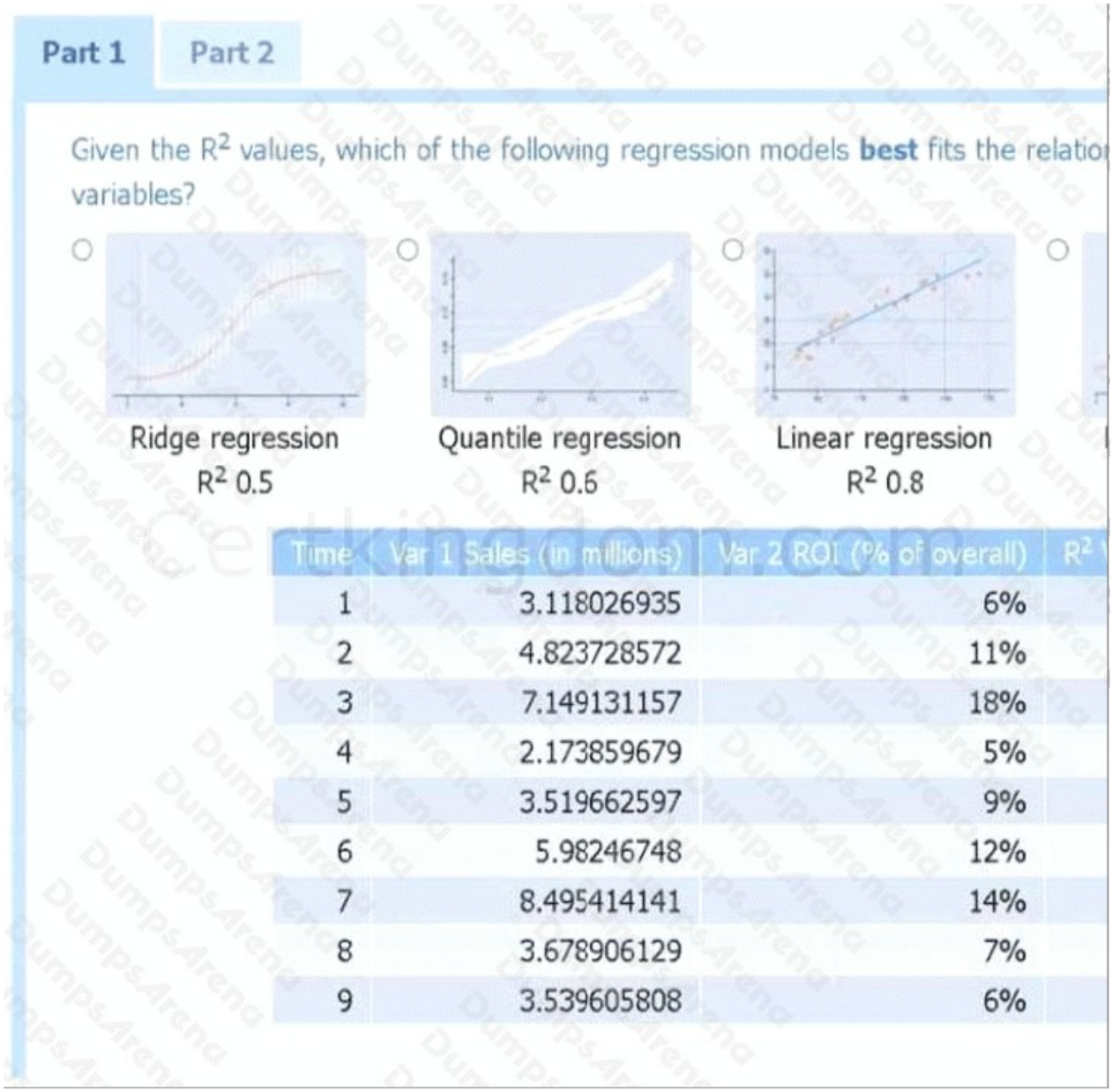
INSTRUCTIONS

Part 1

Use the information provided in the table to select the appropriate regression model. Part 2

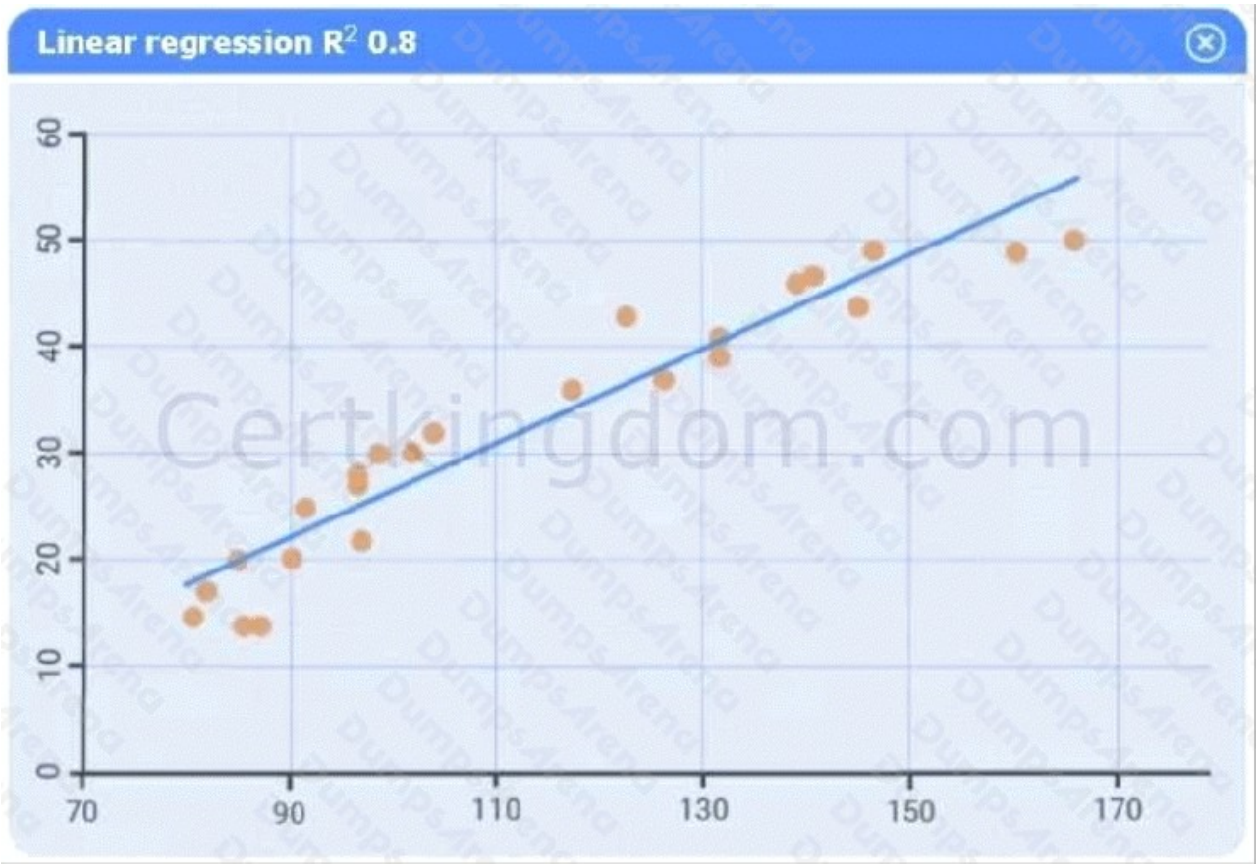
Review the summary output and variable table to determine which variable is statistically significant.

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



Time	Var 1 Sales (in millions)	Var 2 ROI (% of overall)	R^2
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%	

Certkingdom.com


Linear regression R^2 0.8
0
10
20
30
40
50
60
70
90
110
130
150
170

Lasso regression R^2 0.62
-1
0
1
2
3
4

Part 1 | Part 2

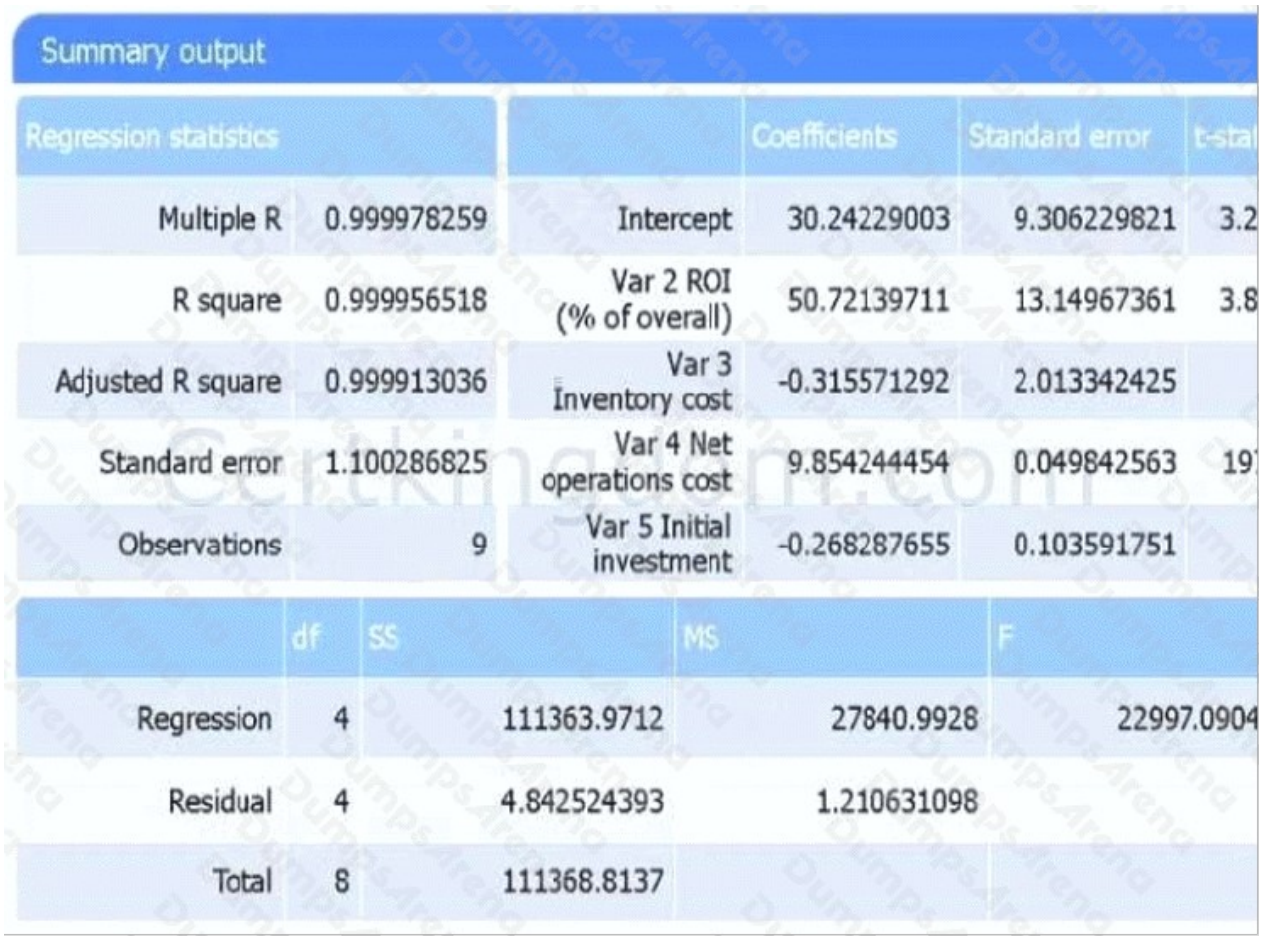
Time	Var 1 Sales (in millions)	Var 2 ROI (% of overall)	Var 3 Inventory cost	Var 4 operat
1	326.311584	16%	58	
2	507.9584031	8%	57	
3	232.5685962	5%	53	
4	117.3342091	7%	50	
5	242.866515	7%	60	
6	359.6300247	14%	50	
7	119.384542	19%	56	
8	372.064584	5%	56	
9	320.0212452	18%	51	

View summary output

Which of the following additional variables s include in the new model?

- ○ Var 5 Initial investment
- ○ Var 4 Net opera
- ○ Var 3 Inventory cost
- ○ None of the va

Certkingdom.com



Summary output

Regression statistics			Coefficients	Standard error	t-stat
Multiple R	0.999978259	Intercept	30.24229003	9.306229821	3.2
R square	0.999956518	Var 2 ROI (% of overall)	50.72139711	13.14967361	3.8
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	19
Observations	9	Var 5 Initial investment	-0.268287655	0.103591751	

	df	SS	MS	F
Regression	4	111363.9712	27840.9928	22997.0904
Residual	4	4.842524393	1.210631098	
Total	8	111368.8137		

ANSWER: See the explanation for the answer

Explanation:

Part 1: Select **Linear regression**.

Part 2: Select **Var 4 — Net operations cost**.

Final selections: `Linear regression; Var 4 Net operations cost.`

QUESTION NO: 2

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:

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 B

Explanation:

Final charts and dashboards and **Model selection, justification, and purpose** are appropriate for a non-technical briefing centered on analysis and results. Clear visualizations translate analytical findings into an accessible narrative: they allow stakeholders to see important trends, comparisons, predictions, and business outcomes without requiring them to interpret implementation details. A dashboard or final chart should emphasize the measures and insights that support decisions, rather than the mechanics used to generate them.

Communicating the model's purpose and the rationale for selecting it is also important because it gives stakeholders the context needed to use the results responsibly. The audience should understand the business problem the model addresses, the intended use of its outputs, and why the selected approach is suitable for that purpose. This supports transparency and helps connect the analytical work to organizational objectives. CompTIA's DataX certification overview emphasizes communicating data insights and applying data-science work to business needs, while its exam-objectives resources describe the skills assessed for current certifications. See CompTIA DataX and CompTIA Exam Objectives.

QUESTION NO: 3

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 are the appropriate compute delivery model because they package an application or reusable data-science asset with the specific runtime, libraries, configuration, and other dependencies required for it to operate. Unlike a full operating-system image, a container shares the host operating system kernel, so the deployable image includes only the components critical to the workload. This produces a comparatively lightweight and portable artifact that can be versioned, stored in an image registry, and run consistently across development, test, and production environments. For a data scientist, this helps make notebooks, model-serving APIs, feature-processing jobs, and other workloads reproducible when moved between local systems, cloud platforms, and deployment pipelines. Container images also support reuse: a standardized image can be instantiated repeatedly without rebuilding the environment manually for each use. Docker describes containers as isolated processes that package application code and dependencies, while Kubernetes documentation notes that container images bundle the files needed to run an application. See Docker: What is a container? and Kubernetes: Images.

QUESTION NO: 4

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 discriminant analysis

D. XGBoost

ANSWER: D

Explanation:

XGBoost is the best choice because it is a gradient-boosting implementation that builds an ensemble sequentially. At each boosting iteration, it evaluates the current model's errors through an objective function and uses gradient information to fit the next tree, thereby updating the overall model's parameters to improve predictive performance. This iterative optimization is useful for credit-score prediction, where relationships between reporting-agency attributes and outcomes may be complex, nonlinear, and subject to change as newer observations are incorporated into the training process.

In practice, adapting to recently collected information requires a controlled retraining or continued-training workflow; it is not an unsupervised change that should occur without monitoring. XGBoost supports training continuation by supplying an existing model to the training process, allowing a data science pipeline to update a prior model with additional boosting rounds as new labeled credit data becomes available. Its regularization controls and scalable tree-based learning also make it a strong practical choice for structured tabular data such as credit-report variables. XGBoost documents the iterative gradient-boosting approach and continuation from an existing model in its model training documentation. CompTIA's DataX exam objectives include selecting and applying appropriate machine-learning algorithms for business problems; see the CompTIA DataX certification page.

QUESTION NO: 5

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 R^2

B. p value

C. χ^2

D. R^2

ANSWER: A

Explanation:

Adjusted R^2 is the most appropriate statistic when assessing a regression model whose explanatory value appears to increase as independent variables are added. Ordinary R^2 measures the proportion of variation in the dependent variable explained by the predictors, but it cannot decrease when further predictors are included—even when a new variable contributes little meaningful predictive information. Adjusted R^2 modifies that measure by accounting for both sample size and the number of predictors in the model. This complexity adjustment makes it more useful for judging whether additional independent variables improve the model sufficiently to justify their inclusion. In practical model selection, adjusted R^2 supports comparison of regression models that contain different numbers of predictors, helping distinguish a meaningful improvement in fit from an apparent improvement caused solely by adding variables. A higher adjusted R^2 indicates that the expanded model has improved explanatory performance after applying its penalty for added model complexity. This aligns with standard regression practice and the model-evaluation concepts covered in the CompTIA DataX objectives. See the NIST/SEMATECH e-Handbook discussion of regression model selection and Penn State STAT 462 material on adjusted R-squared.

QUESTION NO: 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:

Constrained optimization is appropriate because scheduling requires selecting assignments and timings that optimize a measurable objective while satisfying operational rules. A scheduling model may assign employees to shifts, jobs to machines, deliveries to routes, or project activities to time periods. Its objective can be to minimize total completion time, lateness, cost, or idle time, or to maximize throughput and resource utilization. At the same time, the solution must honor hard constraints, such as limited resource capacity, worker availability, task precedence, time windows, required coverage, and restrictions on assigning work concurrently.

Constrained optimization represents these objectives and rules explicitly in a mathematical or constraint-based model. Depending on the problem, practitioners commonly use linear optimization, mixed-integer optimization, or constraint programming. These methods can enforce discrete choices—such as whether a job is assigned to a particular machine or whether a shift is staffed—while searching for a feasible schedule with the best objective value. Google's scheduling guidance demonstrates these concepts through employee scheduling and job-shop scheduling models, including constraints and optimization objectives. See Google OR-Tools Scheduling and IBM ILOG CPLEX Optimization Studio scheduling documentation.

QUESTION NO: 7

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:

Random forest is the appropriate deployment recommendation because the run-time information identifies it as the model with the shortest execution time among the available choices. In a real-time classification workflow, the model must produce a prediction quickly enough to act on a user's initial behavior while that interaction is still relevant. Low inference latency is therefore a central operational requirement, rather than merely a model-development preference.

A random forest produces a prediction by aggregating the results of multiple decision trees. Its trained model can be used directly for scoring, and its prediction process is well suited to many structured-data classification applications. When the supplied measurements show that this model has the fastest run time, selecting it minimizes response delay and helps the deployed system meet its real-time service objective. The result is a practical balance: the system can classify users promptly without waiting for a longer-running model to complete.

For additional background on how random forests combine tree-based estimators for classification and prediction, see the scikit-learn ensemble forest documentation. Guidance on evaluating and monitoring latency as part of production machine-learning systems is also available in the Google Cloud MLOps architecture guidance.

QUESTION NO: 8

A data scientist is using gradient descent to train a model. The loss decreases very slowly, and training requires an impractically large number of iterations. Which of the following hyperparameters should the data scientist increase first?

A. Learning rate

B. Number of output classes

C. Train-test split ratio

D. Number of target variables

ANSWER: A

Explanation:

The learning rate should be increased first because it controls the size of the parameter updates made during gradient descent. When the learning rate is too small, the optimizer takes very small steps toward a minimum, causing slow convergence and excessive training iterations.

The data scientist should increase the value carefully and continue monitoring training behavior. An excessively large learning rate can cause the loss to oscillate or diverge rather than converge. TensorFlow describes the learning rate as a parameter that determines the step size used by an optimizer during training. See TensorFlow optimizer documentation.