

Q12

1 Point

A linear model has been fitted to training data and evaluated on the test data. Use the model to predict the output on new_data.

Complete the code below accordingly.

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

new_data = np.array([[8, 9]]).reshape(-1,1)

predictions = model. write code here ( write code here )
print(predictions)
```

Q12.1

0.5 Points

What should come to the first blank?

predict

Q12.2

0.5 Points

What should come to the second blank?

new_data

Q11

0.5 Points

You have been asked to build a model to predict the success of university grant applications. The grant application status can be classified as either "successful" or "unsuccessful". Possible inputs for the model could include the Category of the Grant, The Grant Value, Research Field and Department.

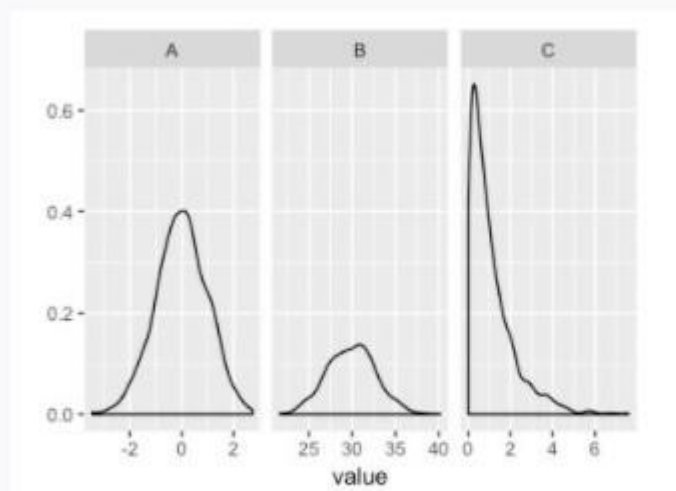
What type of variable is the grant application status?

- ☒ A dependent variable
- ☐ An independent variable
- ☐ A predictor variable

Q10

1 Point

Consider the three variables in the plots below. For which variable should transformation be considered?



☐ A

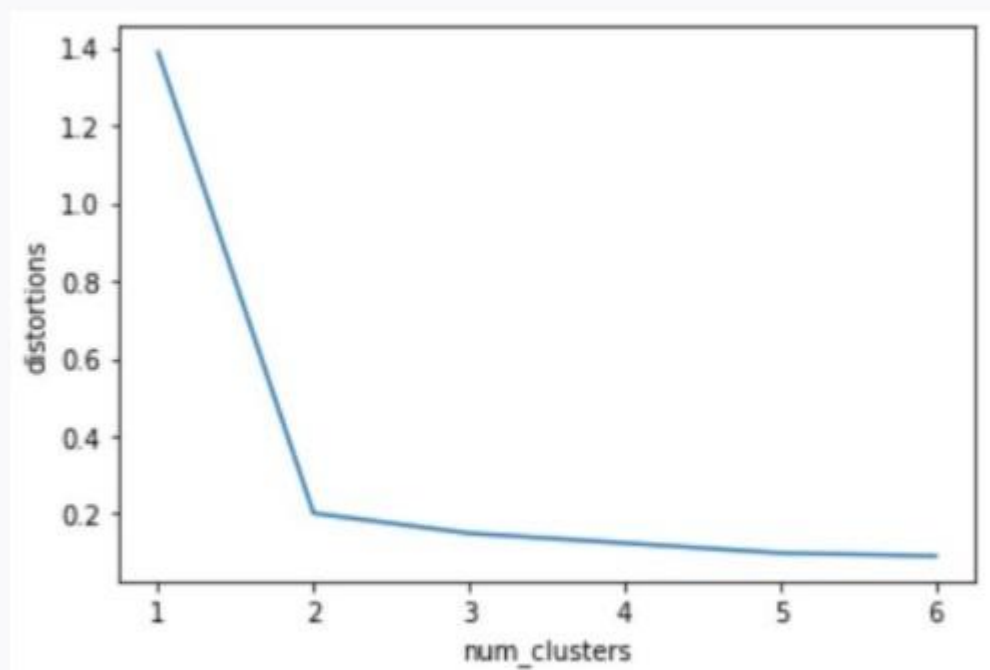
☐ B

☒ C

Q9

1 Point

Consider the plot below. What is the optimal number of clusters for this data?



☒ 2

☐ 3

☐ 4

☐ 5

Q8

1 Point

Which one of the following statements best describes k-means clustering and its utility?

- ☐ K-means clustering reduces the amount of variables in the model to determine which features have the most variability.
- ☐ K-means clustering is a supervised learning algorithm that predicts continuous variables as a function of categorical variables (called clusters).
- ☒ K-means clustering groups data into relatively distinct groups by using a pre-determined number of clusters and iterating cluster assignments.
- ☐ K-means clustering is a supervised learning algorithm that predicts categorical variables (called clusters) as a function of continuous variables.

Q7

1 Point

x and y are Numpy arrays available in this session. Use them to fit a linear model, stored in the reg variable.

Complete the code below accordingly.

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

reg = write code here ()
reg. write code here (x, y)

print("Regression coefficients: {}".format(reg.coef_))
print("Regression intercept: {}".format(reg.intercept_))
```

Q7.1

0.5 Points

What should come to the first blank?

LinearRegression

Q7.2

0.5 Points


What should come to the first blank?

fit

Q5

1 Point

Which one of the following statements describes an unsupervised learning problem?

- ☐ A machine learning where we predict a categorical variable as a function of both continuous and categorical variables.
- ☐ A machine learning problem where we fit a model a response variable as a function of a set of predictor variables.
- ☐ A machine learning problem where we seek to understand whether observations fit into distinct groups based on their similarities.
- ☒ A machine learning problem that deals with the importing and cleaning of a dataset, such that it is tidy. 

Q6

1 Point

Why is it considered good practice to withhold a portion of the data from model building?

- ☐ To use for tuning model parameters.
- ☐ To improve the performance of the model.
- ☒ To verify the performance of the model.
- ☐ To validate assumptions made about the data.

Q3

1 Point

You have built a regression model to predict future crops of tomato based on a range of growing conditions. The model seems to fit well to the training data, but when any new data is tested it performs very poorly. How would you typically describe a model where this occurs?

- ☐ This is described as a well performing model. The model can't be expected to get it right all the time.
- ☐ This is described as prediction error. The model happens to have got it wrong for all of the test data.
- ☒ This is described as underfitting. The model hasn't picked out enough information to perform well yet. **X**
- ☐ This is described as overfitting. The model has picked out details in the data that don't generalize.

Q4

1 Point

Before we fit a model to our data we should consider centering and scaling the data so that:

- ☐ It's easier to interpret the model output because the variables are the same values
- ☒ We can remove outliers from the data because they will all be on the same range **X**
- ☐ A feature does not have more influence on the model because of larger or smaller values
- ☐ We can use both the original and scaled version of the variables in our model

Q2

0.5 Points

Which of the following situations would be most suited to a logistic regression?

- ☐ Predicting the amount of time a customer will take to pay for goods they buy
- ☐ Predicting the number of users arriving at your online store on a given day
- ☐ Predicting the total revenue on a given day based on number of customers
- ☒ Predicting whether a customer will buy a product that you are selling