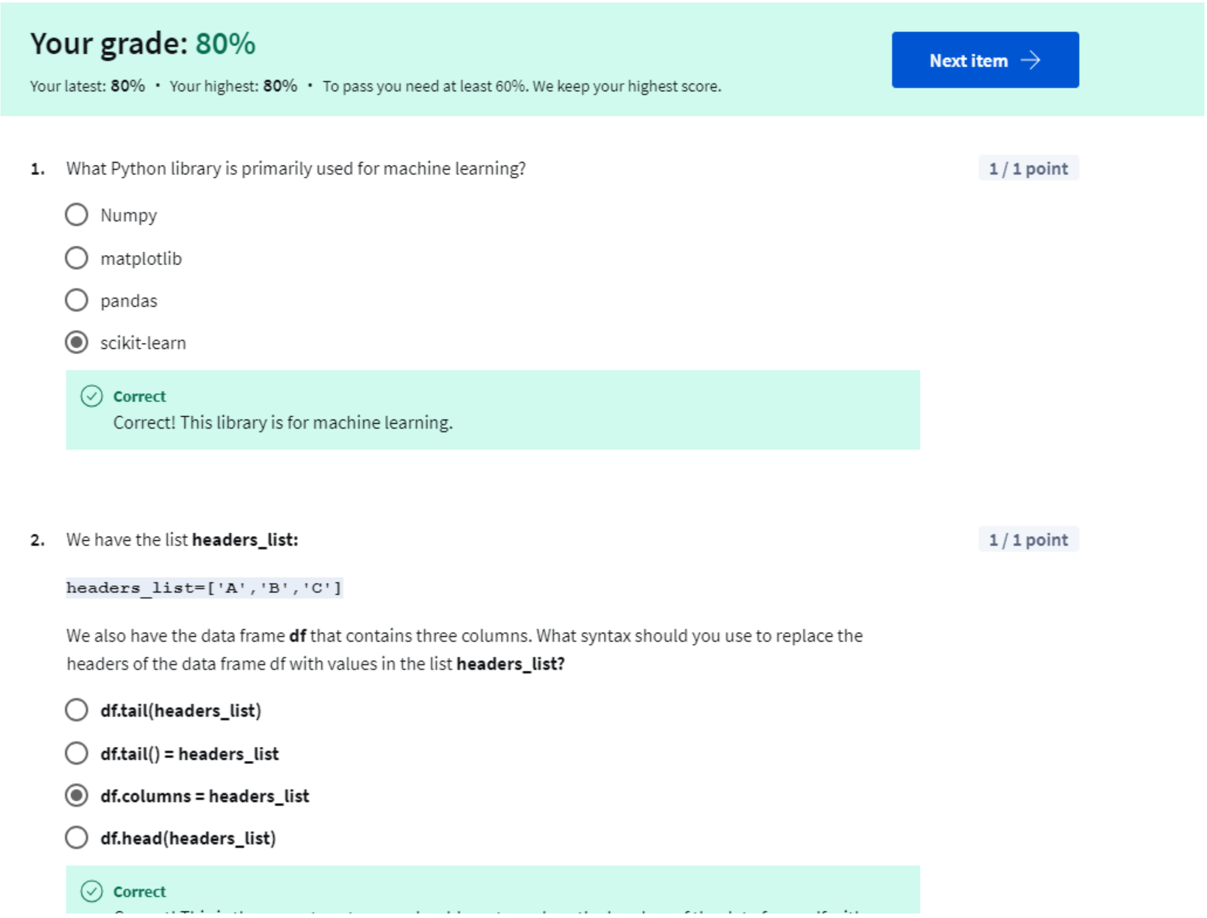
Graded Quiz: Importing Data Sets



Your grade: 80%

Your latest: **80%** • Your highest: **80%** • To pass you need at least 60%. We keep your highest score.

1. What Python library is primarily used for machine learning? 1 / 1 point

○ Numpy

○ matplotlib

○ pandas

● scikit-learn

✓ Correct
Correct! This library is for machine learning.

2. We have the list **headers_list:** 1 / 1 point

```
headers_list=['A','B','C']
```

We also have the data frame **df** that contains three columns. What syntax should you use to replace the headers of the data frame df with values in the list **headers_list?**

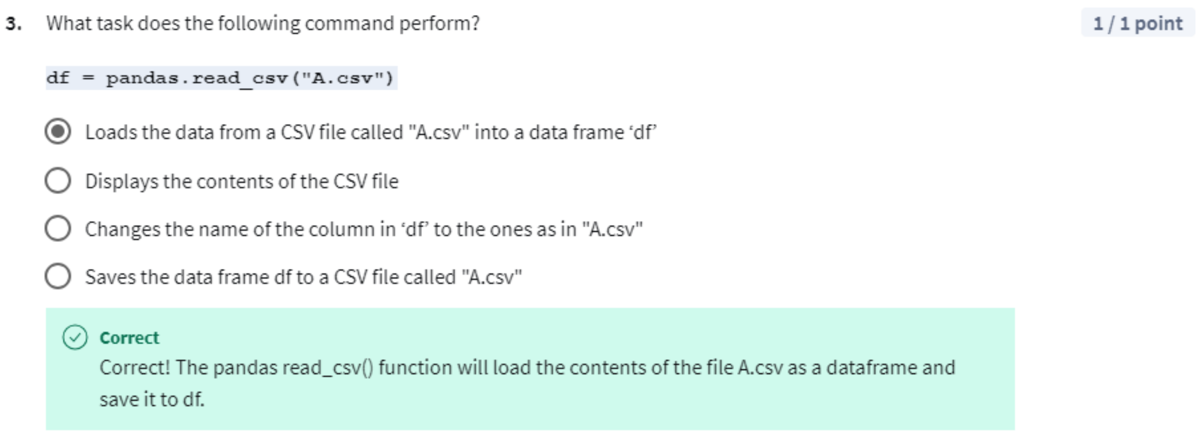
○ **df.tail(headers_list)**

○ **df.tail() = headers_list**

● **df.columns = headers_list**

○ **df.head(headers_list)**

✓ Correct



3. What task does the following command perform? 1 / 1 point

```
df = pandas.read_csv("A.csv")
```

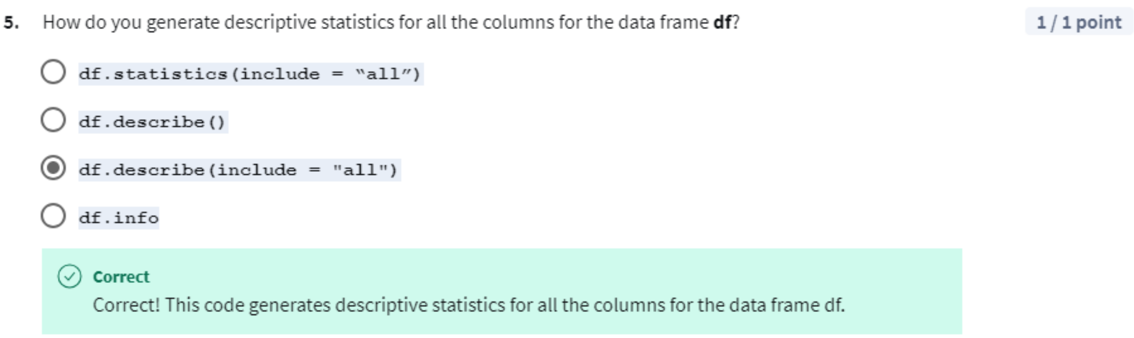
● Loads the data from a CSV file called "A.csv" into a data frame 'df'

○ Displays the contents of the CSV file

○ Changes the name of the column in 'df' to the ones as in "A.csv"

○ Saves the data frame df to a CSV file called "A.csv"

✓ Correct
Correct! The pandas read_csv() function will load the contents of the file A.csv as a dataframe and save it to df.



5. How do you generate descriptive statistics for all the columns for the data frame **df**? 1 / 1 point

○ `df.statistics(include = "all")`

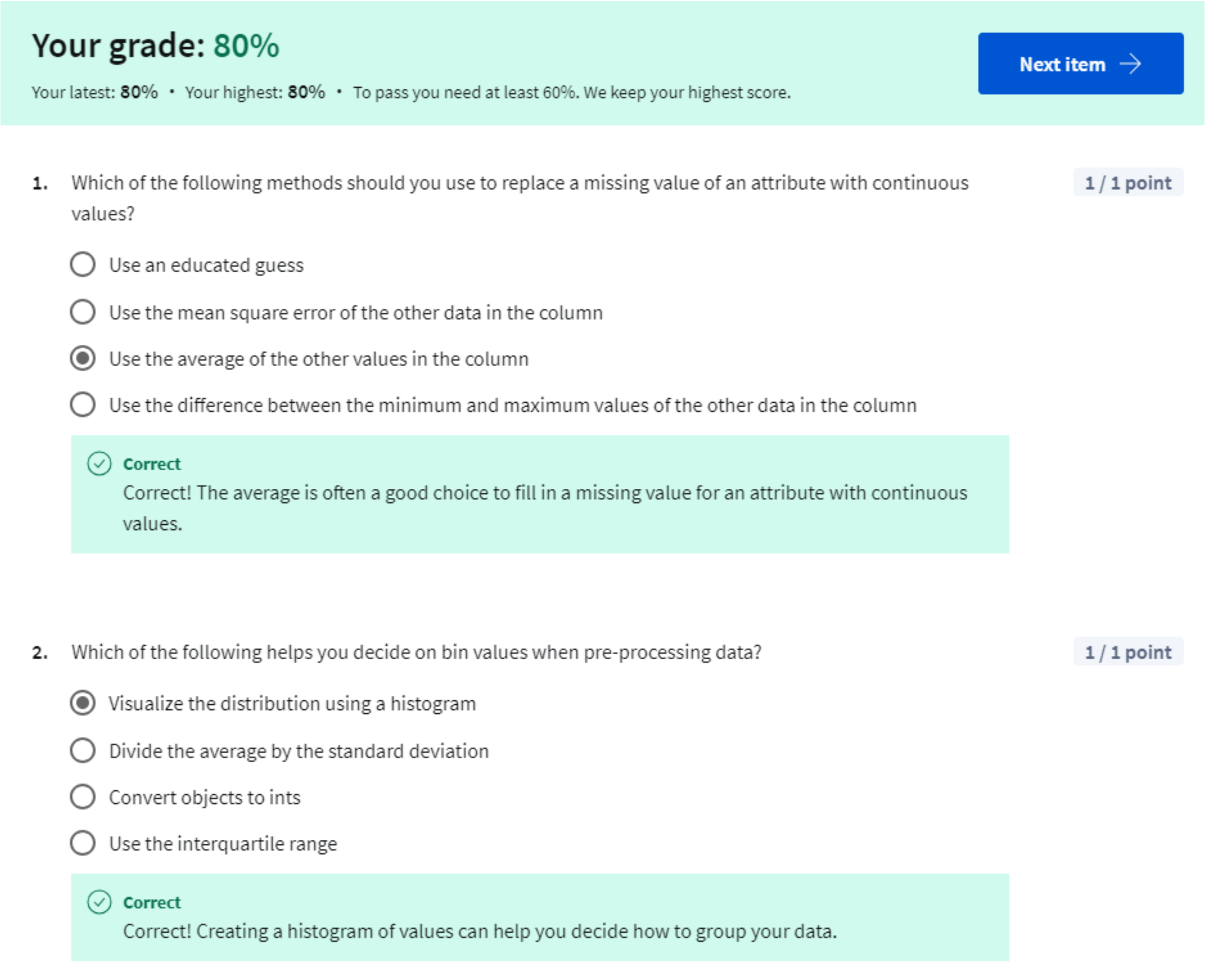
○ `df.describe()`

● `df.describe(include = "all")`

○ `df.info`

✓ Correct
Correct! This code generates descriptive statistics for all the columns for the data frame df.

Graded Quiz: Data Wrangling



Your grade: 80%

Your latest: **80%** • Your highest: **80%** • To pass you need at least 60%. We keep your highest score.

1. Which of the following methods should you use to replace a missing value of an attribute with continuous values?

1 / 1 point

○ Use an educated guess

○ Use the mean square error of the other data in the column

● Use the average of the other values in the column

○ Use the difference between the minimum and maximum values of the other data in the column

✓ Correct

Correct! The average is often a good choice to fill in a missing value for an attribute with continuous values.

2. Which of the following helps you decide on bin values when pre-processing data?

1 / 1 point

● Visualize the distribution using a histogram

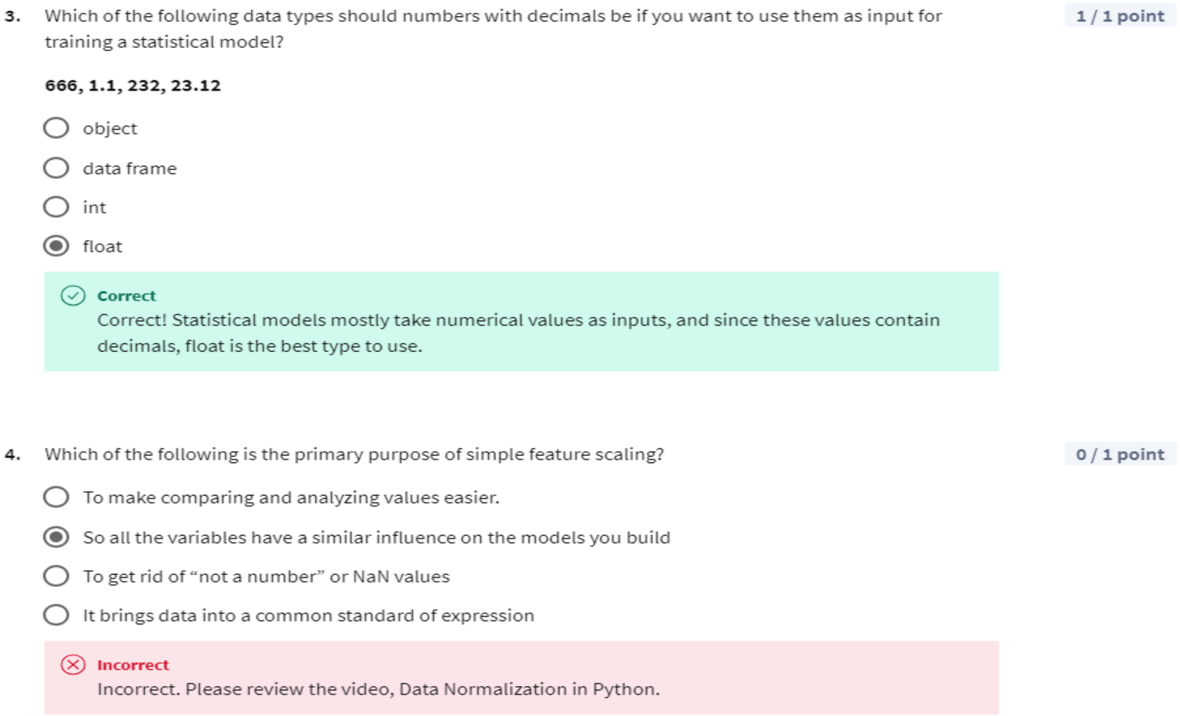
○ Divide the average by the standard deviation

○ Convert objects to ints

○ Use the interquartile range

✓ Correct

Correct! Creating a histogram of values can help you decide how to group your data.



3. Which of the following data types should numbers with decimals be if you want to use them as input for training a statistical model?

1 / 1 point

666, 1.1, 232, 23.12

○ object

○ data frame

○ int

● float

✓ Correct

Correct! Statistical models mostly take numerical values as inputs, and since these values contain decimals, float is the best type to use.

4. Which of the following is the primary purpose of simple feature scaling?

0 / 1 point

○ To make comparing and analyzing values easier.

● So all the variables have a similar influence on the models you build

○ To get rid of "not a number" or NaN values

○ It brings data into a common standard of expression

⊗ Incorrect

Incorrect. Please review the video, Data Normalization in Python.

5. Which of the following is the primary purpose of the get_dummies() method?

1 / 1 point

● Converts categorical values into numerical ones

○ Converts numerical values into categorical ones

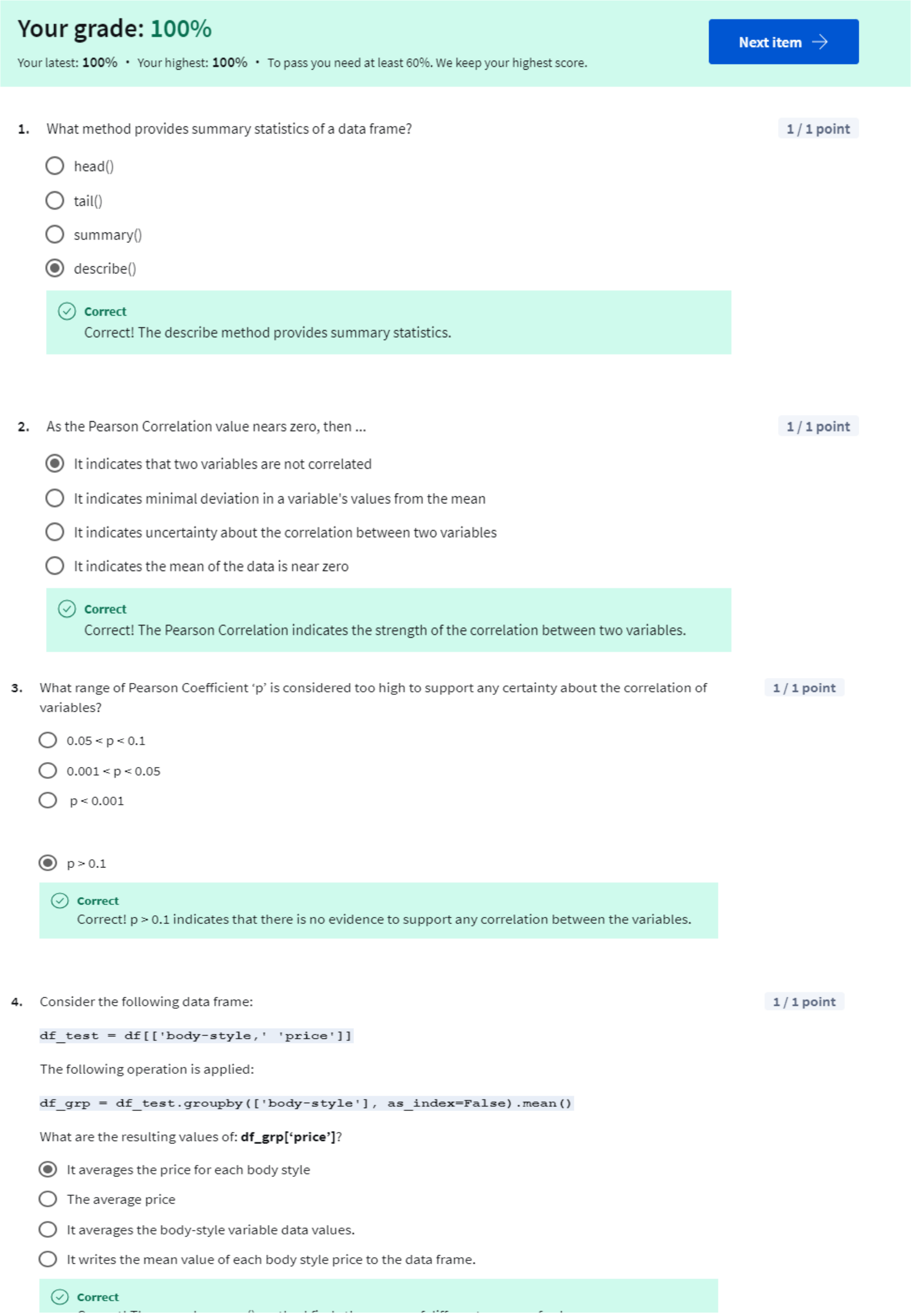
○ To help you group your data into bins

○ Converts the data's data type

✓ **Correct**

Correct! It creates a separate column with names as the entries of the variable's categorical values. It assigns numerical values to each column based on the value of the actual attribute.

Graded Quiz: Exploratory Data Analysis



Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 60%. We keep your highest score.

1. What method provides summary statistics of a data frame?

1 / 1 point

○ head()

○ tail()

○ summary()

● describe()

Correct

Correct! The describe method provides summary statistics.

2. As the Pearson Correlation value nears zero, then ...

1 / 1 point

● It indicates that two variables are not correlated

○ It indicates minimal deviation in a variable's values from the mean

○ It indicates uncertainty about the correlation between two variables

○ It indicates the mean of the data is near zero

Correct

Correct! The Pearson Correlation indicates the strength of the correlation between two variables.

3. What range of Pearson Coefficient ‘p’ is considered too high to support any certainty about the correlation of variables?

1 / 1 point

○ 0.05 < p < 0.1

○ 0.001 < p < 0.05

○ p < 0.001

● p > 0.1

Correct

Correct! p > 0.1 indicates that there is no evidence to support any correlation between the variables.

4. Consider the following data frame:

1 / 1 point

```
df_test = df[['body-style,' 'price']]
```

The following operation is applied:

```
df_grp = df_test.groupby(['body-style'], as_index=False).mean()
```

What are the resulting values of: **df_grp[‘price’]**?

● It averages the price for each body style

○ The average price

○ It averages the body-style variable data values.

○ It writes the mean value of each body style price to the data frame.

Correct

Correct! The groupby mean() method finds the mean of different groups of val

5. What is the Pearson Correlation between two variables if the input variable is equal to the output variable?

1 / 1 point

○ Between -1 and 0

● 1

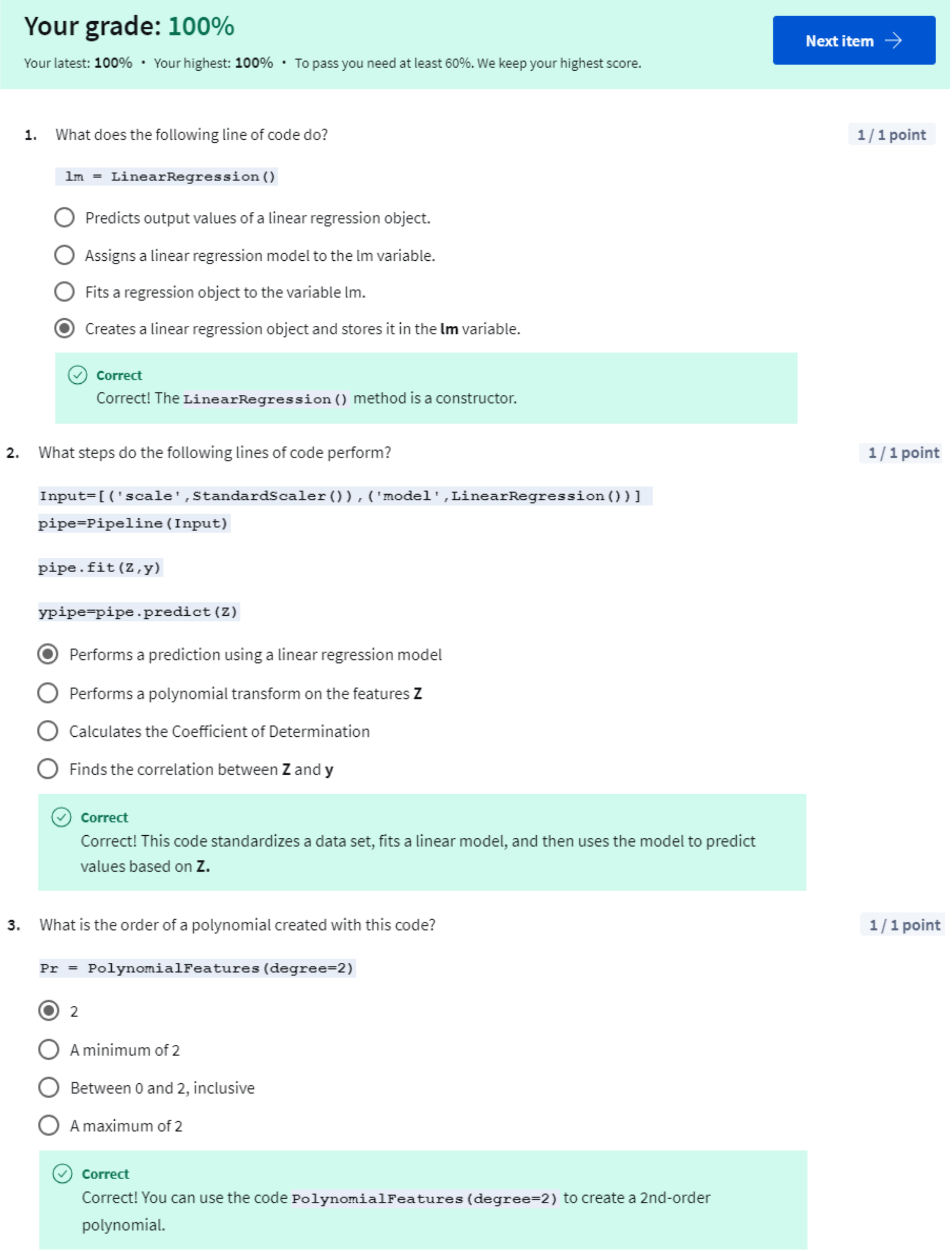
○ Between 0 and 1

○ -1

Correct

Correct! The closer the Pearson Correlation is to 1, the stronger the correlation between input and output. If the values are equal, then 1 indicates the strongest relationship possible.

Graded Quiz: Model Development



Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 60%. We keep your highest score.

1. What does the following line of code do?

1 / 1 point

```
lm = LinearRegression()
```

○ Predicts output values of a linear regression object.

○ Assigns a linear regression model to the lm variable.

○ Fits a regression object to the variable lm.

● Creates a linear regression object and stores it in the **lm** variable.

✓ Correct

Correct! The `LinearRegression()` method is a constructor.

2. What steps do the following lines of code perform?

1 / 1 point

```
Input=[('scale',StandardScaler()),('model',LinearRegression())]
pipe=Pipeline(Input)

pipe.fit(Z,y)

ypipe=pipe.predict(Z)
```

● Performs a prediction using a linear regression model

○ Performs a polynomial transform on the features **Z**

○ Calculates the Coefficient of Determination

○ Finds the correlation between **Z** and **y**

✓ Correct

Correct! This code standardizes a data set, fits a linear model, and then uses the model to predict values based on **Z.**

3. What is the order of a polynomial created with this code?

1 / 1 point

```
Pr = PolynomialFeatures(degree=2)
```

● 2

○ A minimum of 2

○ Between 0 and 2, inclusive

○ A maximum of 2

✓ Correct

Correct! You can use the code `PolynomialFeatures(degree=2)` to create a 2nd-order polynomial.

4. Which statement about **R^2**, the coefficient of determination, is true?

1 / 1 point

○ Its value can be any positive number.

● Its value can be between 0 and 1 inclusive.

○ Its value can be in the range of -1 to 1, inclusive.

○ Its value can be either 0 or 1.

✓ **Correct**
Correct! The coefficient of determination can be a minimum of 0 and a maximum of 1.

5. Consider the following equation:

1 / 1 point

$$y = b_0 + b_1 x$$

The variable y is _________?

○ The predictor or independent variable

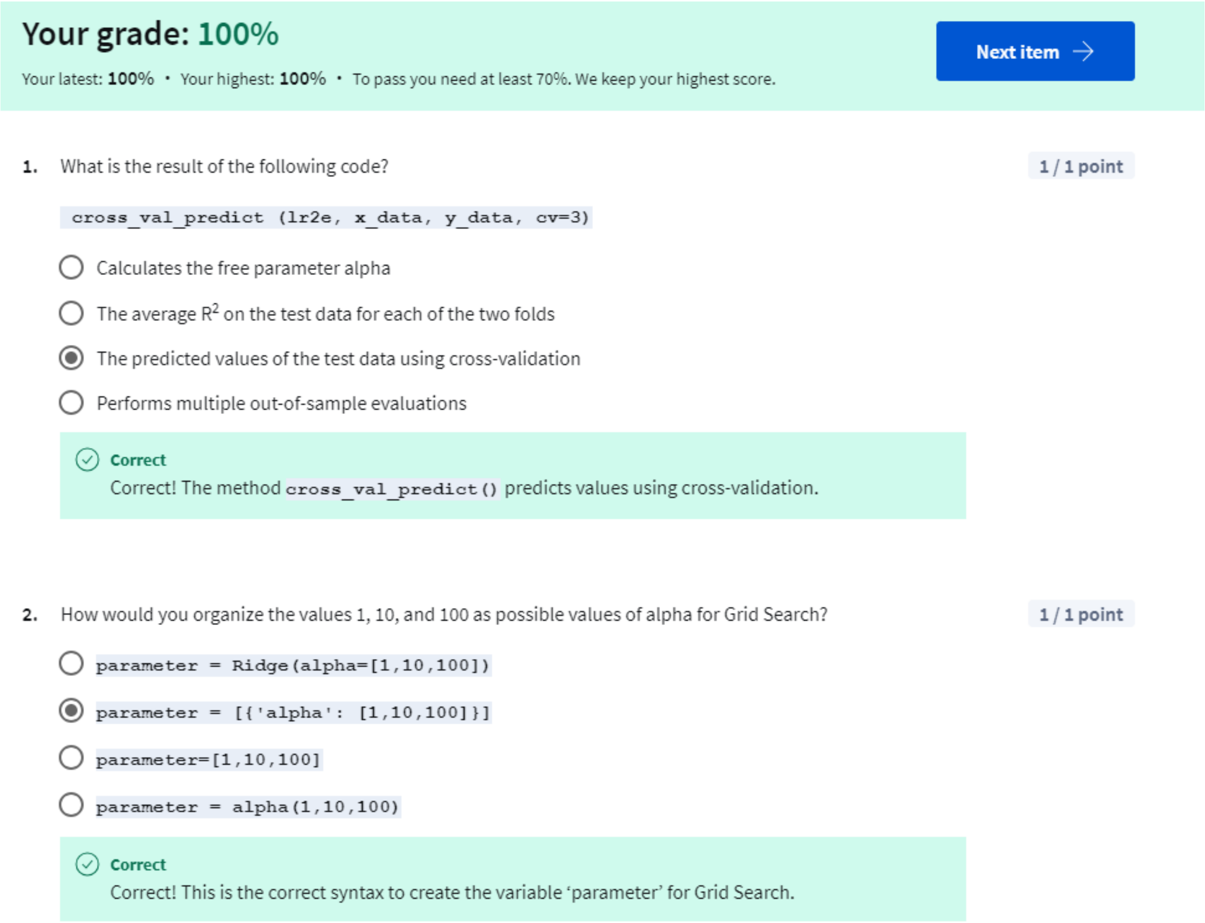
● The target or dependent variable

○ The intercept

○ The degree of the polynomial

✓ **Correct**
Correct! The variable y is the output variable, which depends on the values of the other variable x and parameters b_0 and b_1.

Graded Quiz: Model Evaluation and Refinement



Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 70%. We keep your highest score.

1. What is the result of the following code?

```
cross_val_predict (lr2e, x_data, y_data, cv=3)
```

○ Calculates the free parameter alpha

○ The average R^2 on the test data for each of the two folds

● The predicted values of the test data using cross-validation

○ Performs multiple out-of-sample evaluations

✓ **Correct**
Correct! The method `cross_val_predict()` predicts values using cross-validation.

2. How would you organize the values 1, 10, and 100 as possible values of alpha for Grid Search?

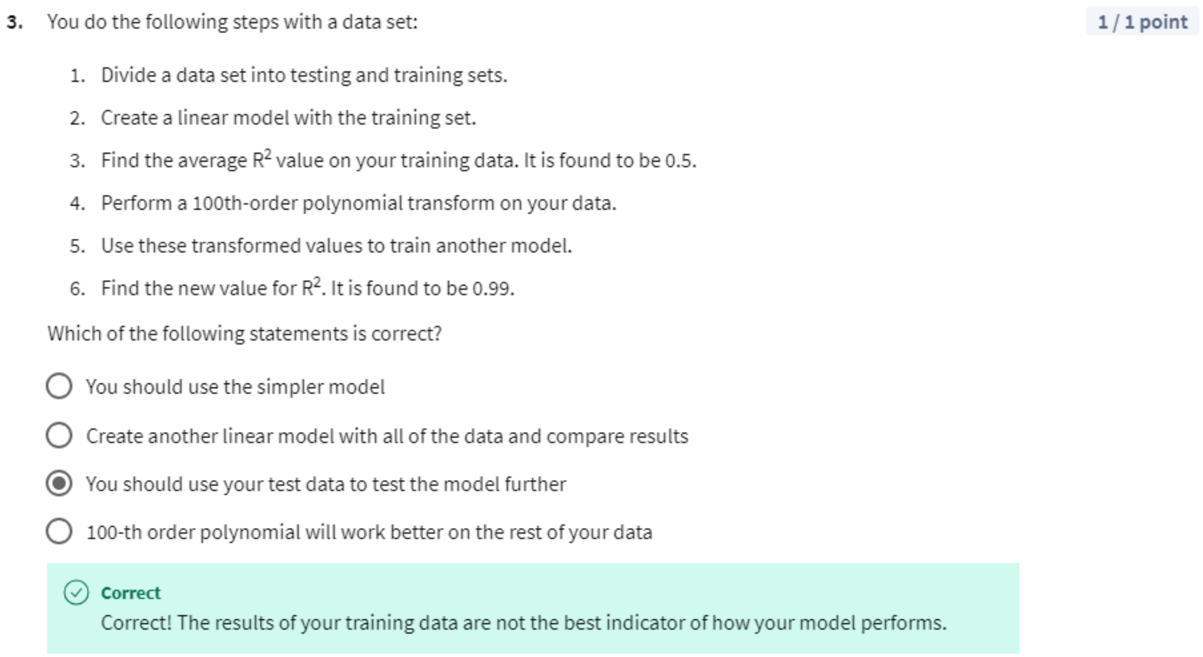
○ `parameter = Ridge(alpha=[1,10,100])`

● `parameter = [{'alpha': [1,10,100]}]`

○ `parameter=[1,10,100]`

○ `parameter = alpha(1,10,100)`

✓ **Correct**
Correct! This is the correct syntax to create the variable 'parameter' for Grid Search.



3. You do the following steps with a data set:

1. Divide a data set into testing and training sets.
2. Create a linear model with the training set.
3. Find the average R^2 value on your training data. It is found to be 0.5.
4. Perform a 100th-order polynomial transform on your data.
5. Use these transformed values to train another model.
6. Find the new value for R^2. It is found to be 0.99.

Which of the following statements is correct?

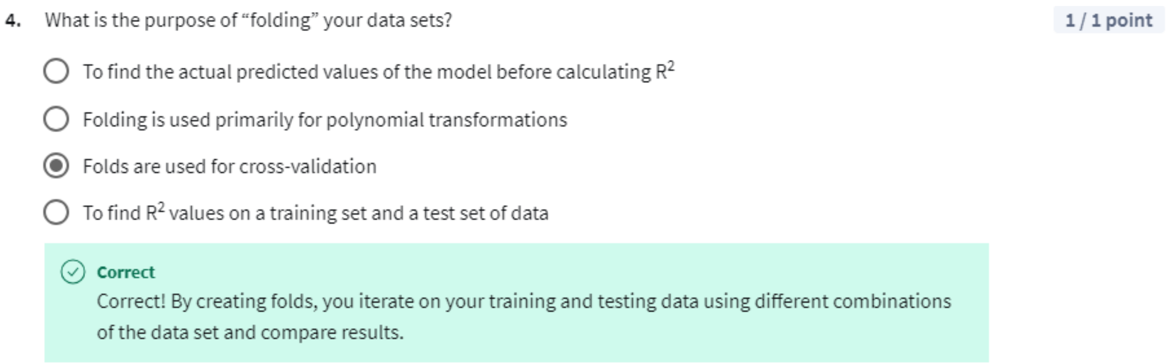
○ You should use the simpler model

○ Create another linear model with all of the data and compare results

● You should use your test data to test the model further

○ 100-th order polynomial will work better on the rest of your data

✓ **Correct**
Correct! The results of your training data are not the best indicator of how your model performs.



4. What is the purpose of “folding” your data sets?

1 / 1 point

○ To find the actual predicted values of the model before calculating R^2

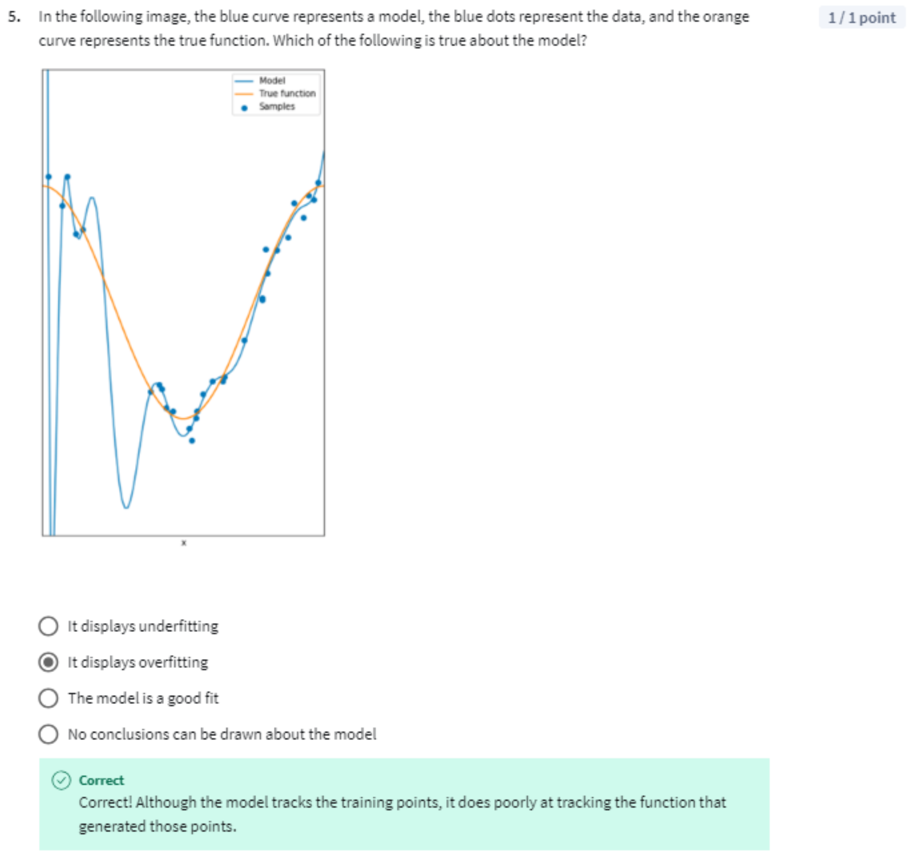
○ Folding is used primarily for polynomial transformations

● Folds are used for cross-validation

○ To find R^2 values on a training set and a test set of data

✓ **Correct**

Correct! By creating folds, you iterate on your training and testing data using different combinations of the data set and compare results.



5. In the following image, the blue curve represents a model, the blue dots represent the data, and the orange curve represents the true function. Which of the following is true about the model?

1 / 1 point

○ It displays underfitting

● It displays overfitting

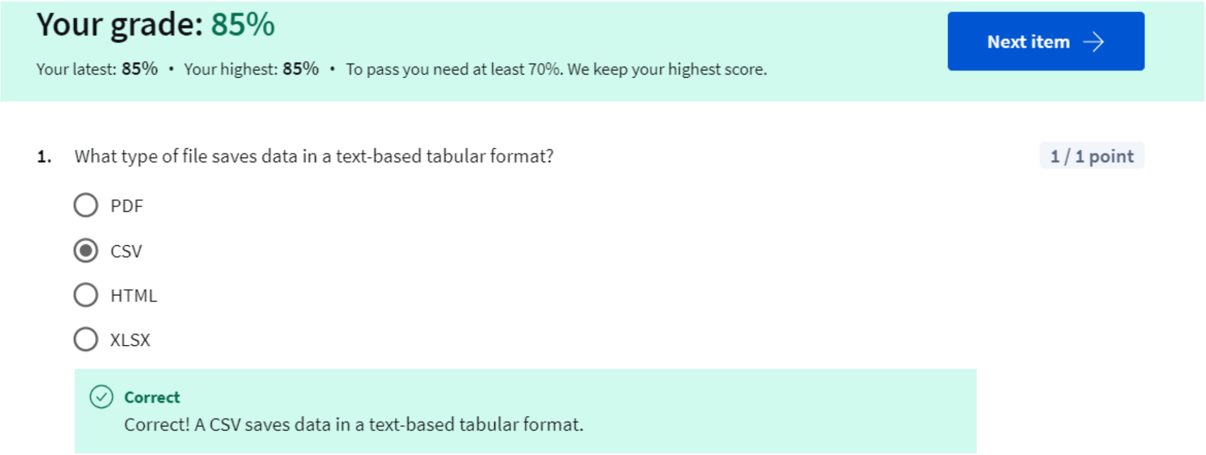
○ The model is a good fit

○ No conclusions can be drawn about the model

✓ **Correct**

Correct! Although the model tracks the training points, it does poorly at tracking the function that generated those points.

Final Exam



Your grade: 85%

Your latest: **85%** • Your highest: **85%** • To pass you need at least 70%. We keep your highest score.

1. What type of file saves data in a text-based tabular format?

1 / 1 point

○ PDF

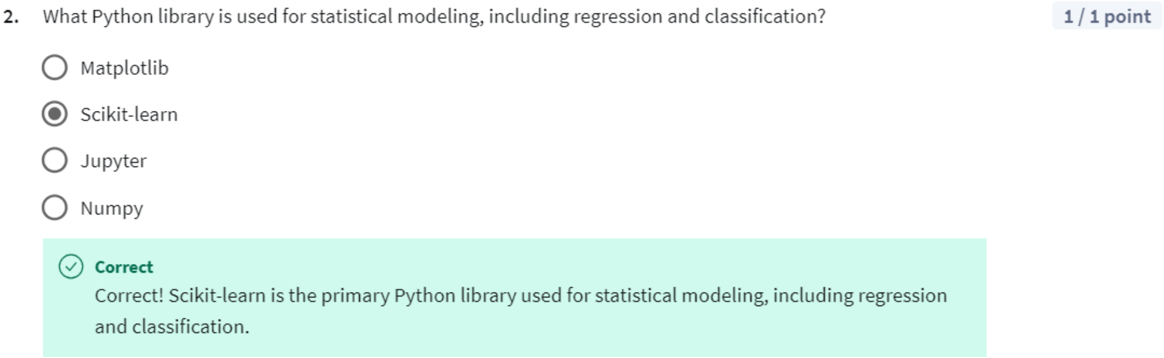
● CSV

○ HTML

○ XLSX

✓ **Correct**

Correct! A CSV saves data in a text-based tabular format.



2. What Python library is used for statistical modeling, including regression and classification?

1 / 1 point

○ Matplotlib

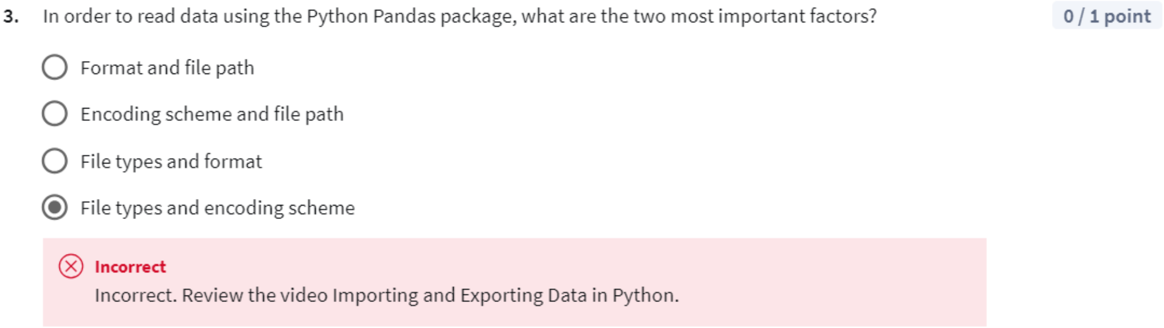
● Scikit-learn

○ Jupyter

○ Numpy

✓ **Correct**

Correct! Scikit-learn is the primary Python library used for statistical modeling, including regression and classification.



3. In order to read data using the Python Pandas package, what are the two most important factors?

0 / 1 point

○ Format and file path

○ Encoding scheme and file path

○ File types and format

● File types and encoding scheme

ⓧ **Incorrect**

Incorrect. Review the video Importing and Exporting Data in Python.

4. What attribute or function returns the data types of each column of a data frame?

1 / 1 point

● `dtypes`

○ `head()`

○ `tail()`

○ `datatypes`

Correct
Correct! The **`dtypes`** attribute returns the data types of each column.

5. What is a header?

1 / 1 point

○ The name of the rows

● The name of the columns

○ The first value in a column

○ The first value in a row

Correct
Correct! The header refers to the names of the columns.

6. The Matplotlib library is mostly used for what?

1 / 1 point

○ Statistical modeling

○ Machine learning algorithms

● Data visualization

○ Data analysis

Correct
Correct! The Matplotlib library is mostly used for data visualization.

7. What is the output of the following code segment of the data frame **df**?

1 / 1 point

```
df.tail(10)
```

● It returns the last 10 rows of the data frame

○ It returns the header of the data frame

○ It returns the first 10 rows of the data frame

○ It returns all of the rows of the data frame

Correct
Correct! The code **`df.tail(10)`** returns the last 10 rows of the data frame.

8. What task does the following code segment do to a data frame **df**?

1 / 1 point

```
mean = df["price"].mean() df["price"].replace(np.nan, mean)
```

○ It drops rows that contain missing values

○ It calculates the mean of the data in the column "price"

● It replaces the missing values in the column "price" with the mean values of that column

○ It replaces the data in the column "price" with normalized values

✓ **Correct**

Correct! This line of code replaces the missing values in the column "price" with the mean values of that column.

9. Which statement about binning is true?

1 / 1 point

○ It is primarily used to calculate descriptive statistics.

● It is primarily used to gain a better understanding of the data distribution.

○ It is primarily used to normalize the data.

○ It is primarily used to format the data.

✓ **Correct**

Correct! Binning is primarily used to gain a better understanding of the data distribution.

10. What is the primary purpose of standardizing a set of values?

1 / 1 point

○ To find how well a data set fits a model.

○ So you can see the spread of the data set and identify outliers.

○ To see how many standard deviations each value is from the mean.

● It places different variables on the same scale, allowing you to compare them more easily.

✓ **Correct**

Correct! Standardizing values serves to place different variables on the same scale, allowing you to compare them more easily.

11. What is the primary purpose of *one-hot encoding*?

1 / 1 point

○ To convert numeric variables into categorical ones

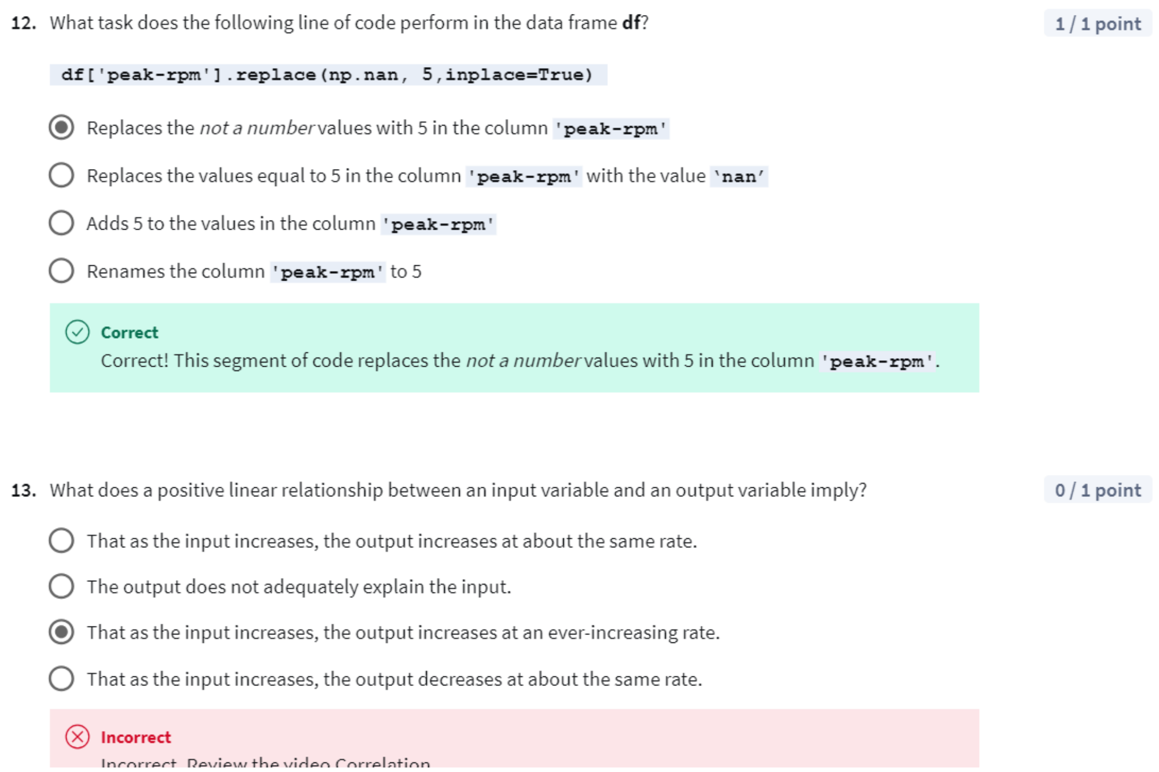
● To convert categorical variables into numeric ones

○ To convert numeric data types into object data types

○ To convert object data types into numeric data types

✓ **Correct**

Correct! *One-hot encoding* converts categorical variables into numeric ones.



12. What task does the following line of code perform in the data frame **df**?

1 / 1 point

```
df['peak-rpm'].replace(np.nan, 5,inplace=True)
```

● Replaces the *not a number* values with 5 in the column `'peak-rpm'`

○ Replaces the values equal to 5 in the column `'peak-rpm'` with the value `'nan'`

○ Adds 5 to the values in the column `'peak-rpm'`

○ Renames the column `'peak-rpm'` to 5

✓ **Correct**
Correct! This segment of code replaces the *not a number* values with 5 in the column `'peak-rpm'`.

13. What does a positive linear relationship between an input variable and an output variable imply?

0 / 1 point

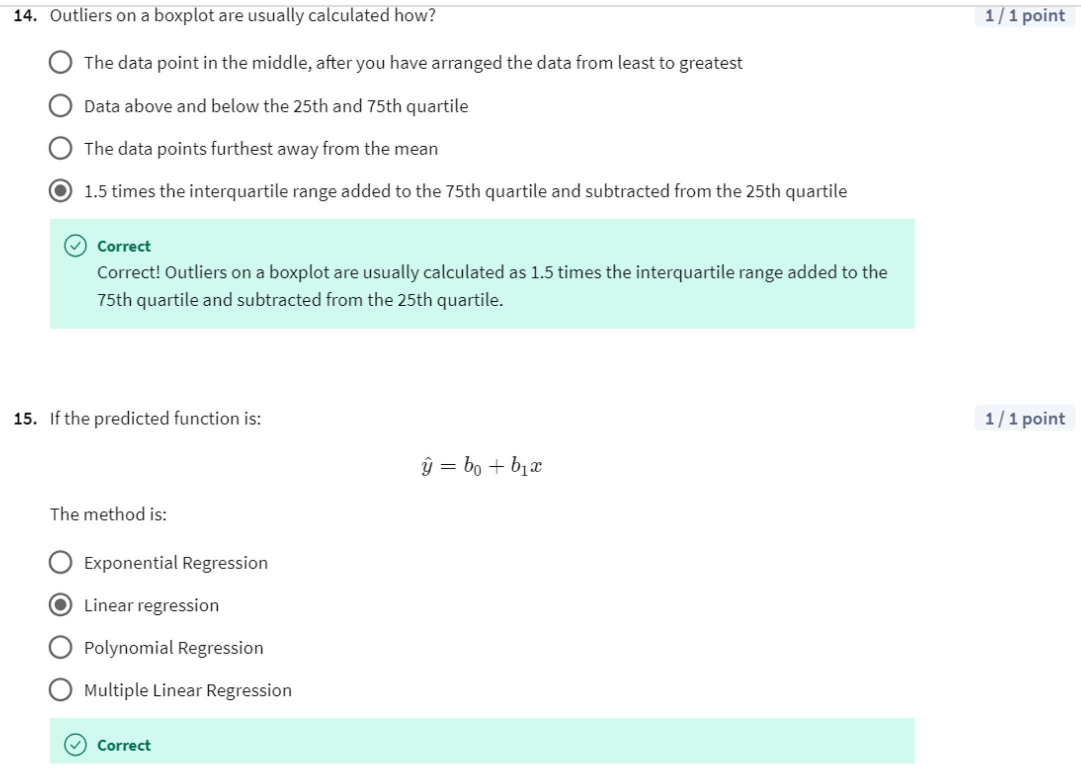
○ That as the input increases, the output increases at about the same rate.

○ The output does not adequately explain the input.

● That as the input increases, the output increases at an ever-increasing rate.

○ That as the input increases, the output decreases at about the same rate.

⊗ **Incorrect**
Incorrect. Review the video Correlation



14. Outliers on a boxplot are usually calculated how?

1 / 1 point

○ The data point in the middle, after you have arranged the data from least to greatest

○ Data above and below the 25th and 75th quartile

○ The data points furthest away from the mean

● 1.5 times the interquartile range added to the 75th quartile and subtracted from the 25th quartile

✓ **Correct**
Correct! Outliers on a boxplot are usually calculated as 1.5 times the interquartile range added to the 75th quartile and subtracted from the 25th quartile.

15. If the predicted function is:

1 / 1 point

$$\hat{y} = b_0 + b_1 x$$

The method is:

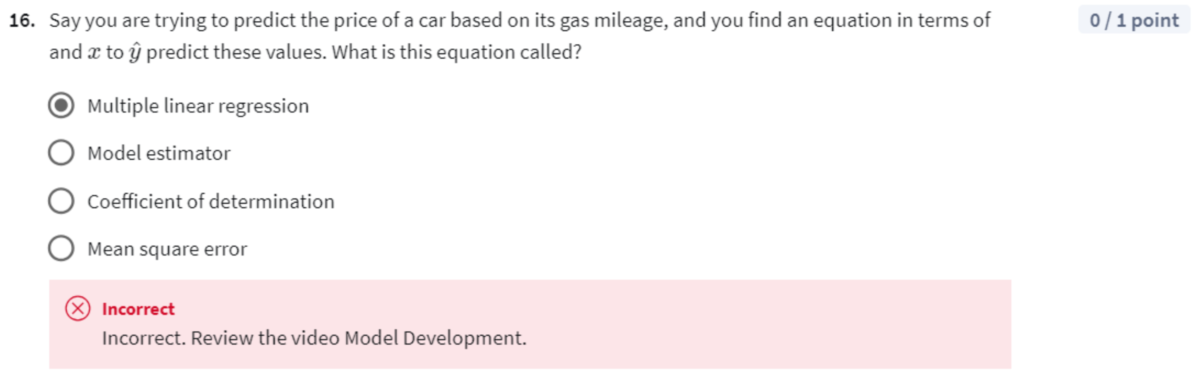
○ Exponential Regression

● Linear regression

○ Polynomial Regression

○ Multiple Linear Regression

✓ **Correct**



16. Say you are trying to predict the price of a car based on its gas mileage, and you find an equation in terms of and x to $\hat{y}$ predict these values. What is this equation called?

0 / 1 point

● Multiple linear regression

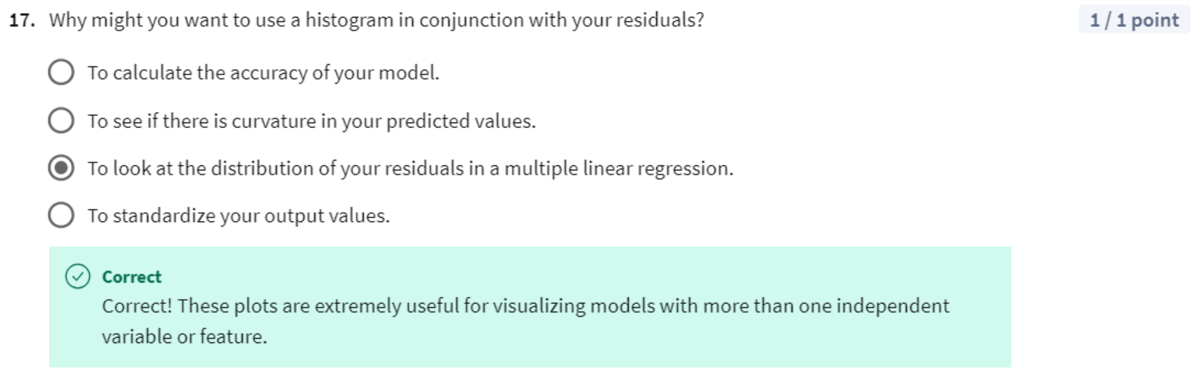
○ Model estimator

○ Coefficient of determination

○ Mean square error

ⓧ **Incorrect**

Incorrect. Review the video Model Development.



17. Why might you want to use a histogram in conjunction with your residuals?

1 / 1 point

○ To calculate the accuracy of your model.

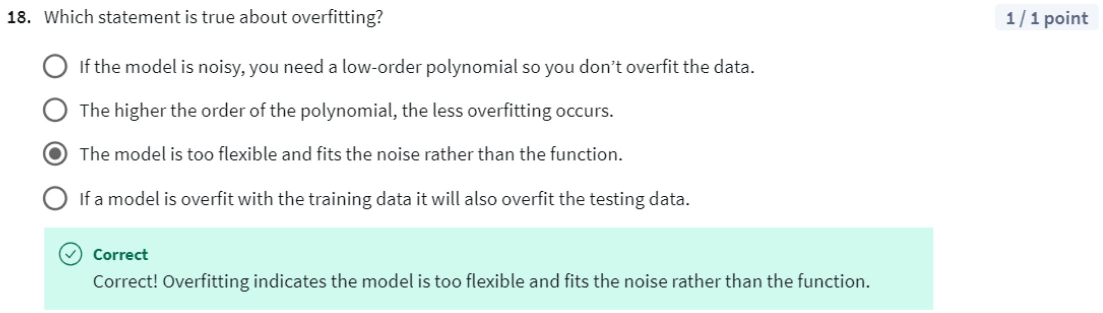
○ To see if there is curvature in your predicted values.

● To look at the distribution of your residuals in a multiple linear regression.

○ To standardize your output values.

✓ **Correct**

Correct! These plots are extremely useful for visualizing models with more than one independent variable or feature.



18. Which statement is true about overfitting?

1 / 1 point

○ If the model is noisy, you need a low-order polynomial so you don't overfit the data.

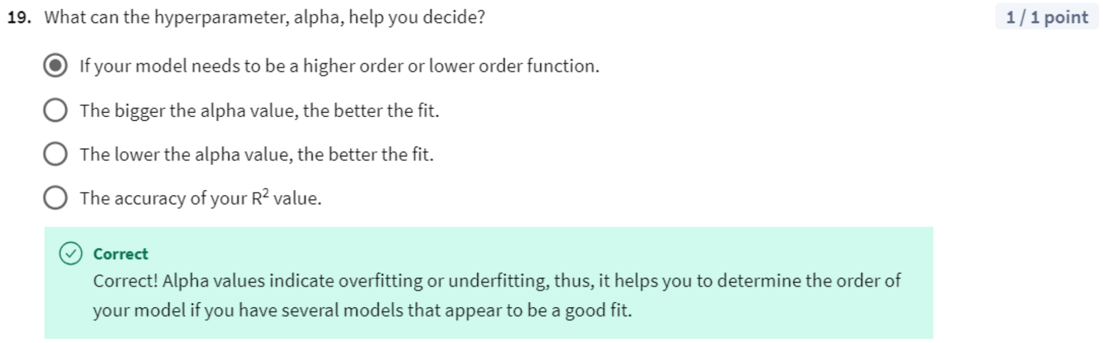
○ The higher the order of the polynomial, the less overfitting occurs.

● The model is too flexible and fits the noise rather than the function.

○ If a model is overfit with the training data it will also overfit the testing data.

✓ **Correct**

Correct! Overfitting indicates the model is too flexible and fits the noise rather than the function.



19. What can the hyperparameter, alpha, help you decide?

1 / 1 point

● If your model needs to be a higher order or lower order function.

○ The bigger the alpha value, the better the fit.

○ The lower the alpha value, the better the fit.

○ The accuracy of your R^2 value.

✓ **Correct**

Correct! Alpha values indicate overfitting or underfitting, thus, it helps you to determine the order of your model if you have several models that appear to be a good fit.

20. What does the `GridSearchCV()` method do?

1 / 1 point

○ It’s another way to cross-validate your data set.

○ It selects the appropriate hyperparameters for your model.

○ It gives you R2 values for different orders of polynomial models.

● It iterates over hyperparameters using cross-validation.

✓ **Correct**
Correct! The `GridSearchCV()` iterates over hyperparameters using cross-validation.