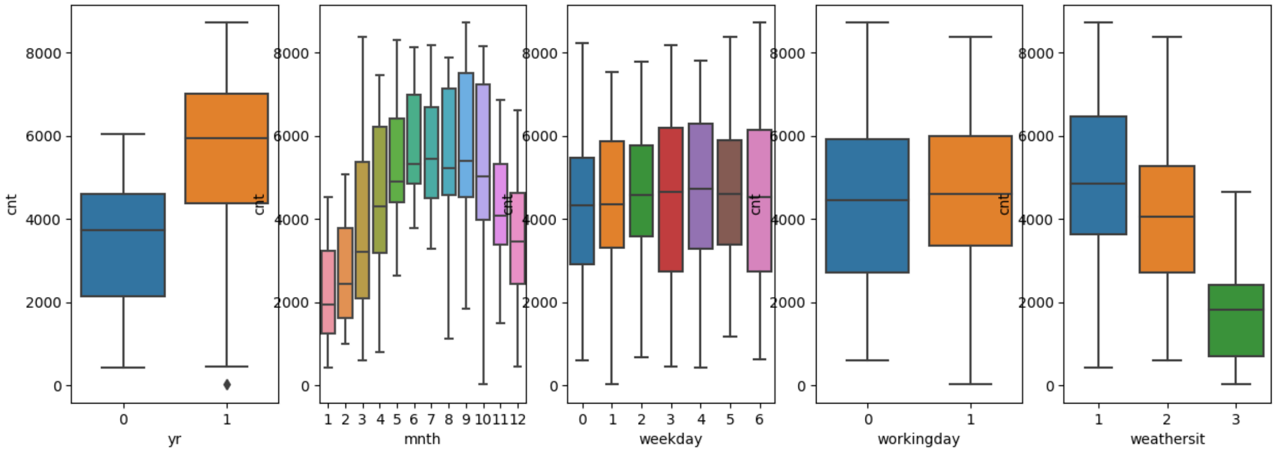
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

A. The following observations could be made:



- Month and weather seem to have a strong correlation with cnt.

- Clear and cloudy weather were good for the bikers.

- Bike rentals grew gradually from January through July and dropped considerably from October throughDecember. This is probably due to the strong snow and rainfall in those months.

- 2019 had significantly more sales than 2018.

- There is slight rise through the bike rentals from Monday through Friday and these drop through the weekend.

2. **Why is it important to use drop_first=True during dummy variable creation?**

A. In dummy variable creation we are creating new features in the dataset to accommodate the categorical variables in the model. In such a case, it is possible to create these features in n-1 columns instead of n, because n-1 columns suffice the data without any loss. Here n is the number of possible values in the categorical feature.

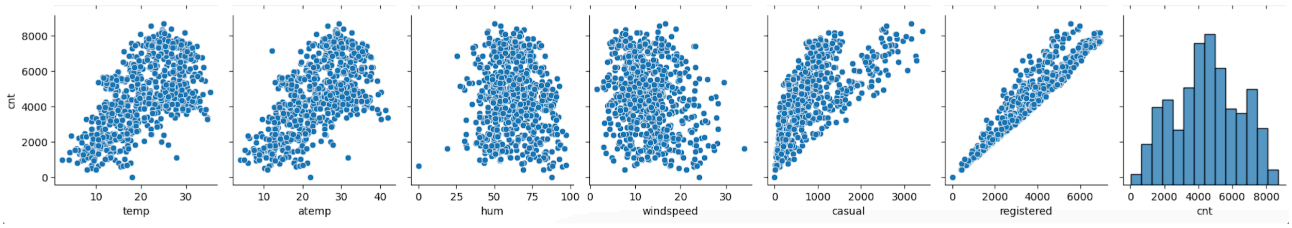
For example, if a categorical variable is ‘workingday’ and its possible values are just True and False. We can convert this to one (n-1) dummy variable such as:

- 0 - False
- 1 - True

So we can make do with just one dummy variable instead of two.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

A. Among the numerical variables, temperature seems to have a linear relation with the bike rentals.



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

A. The following are the assumptions of linear regression:

1. Linearity: The relationship between the independent variable(s) and the dependent variable is linear.

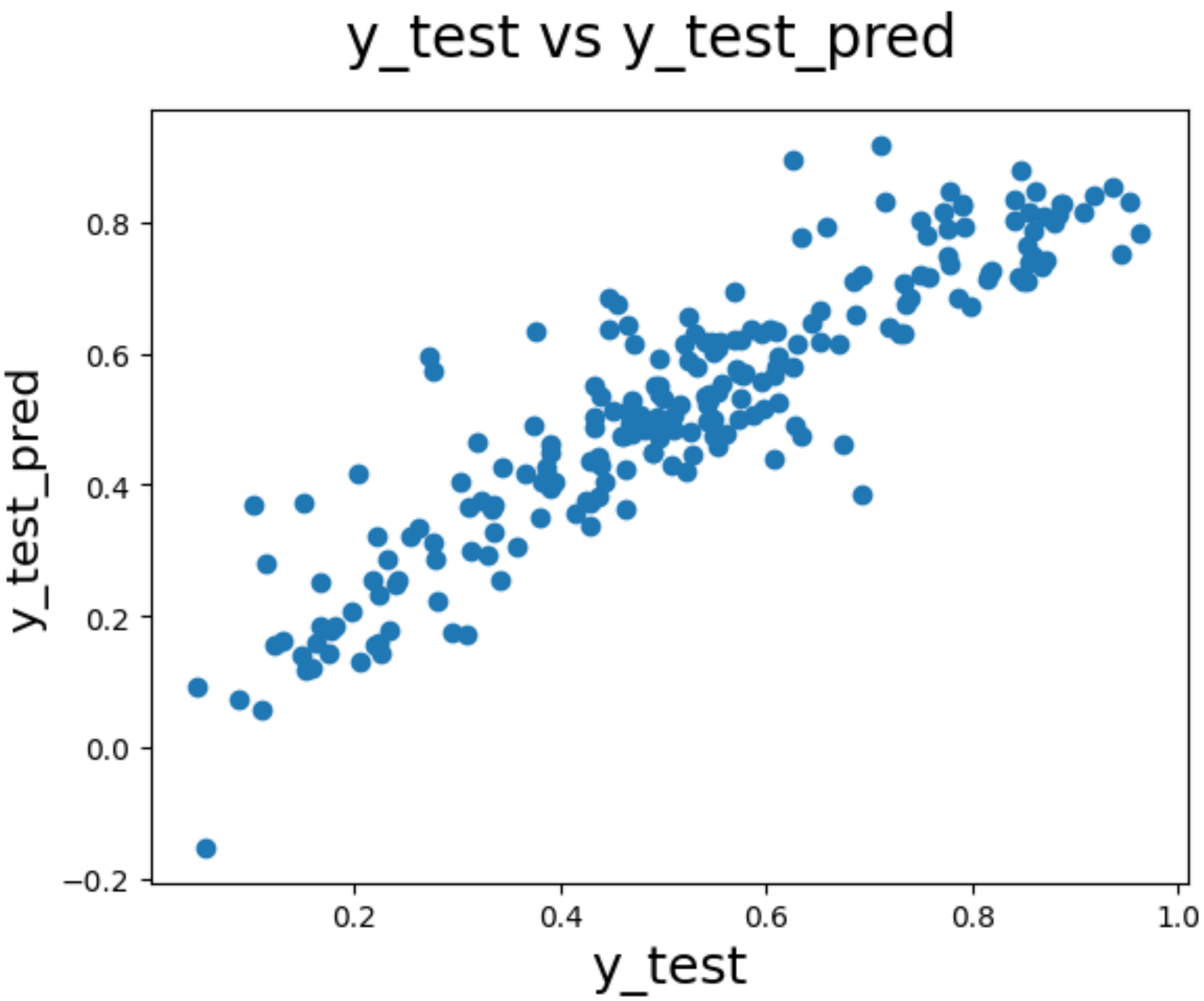
   A. Verified using EDA that there is linear relationship between some of the independent variables such as temp, weather, weekday with the target variable cnt.

2. Independence: The observations are independent of each other.

   A. Each record in the dataset represents an individual day's bike rental sales. They are independent of each other.

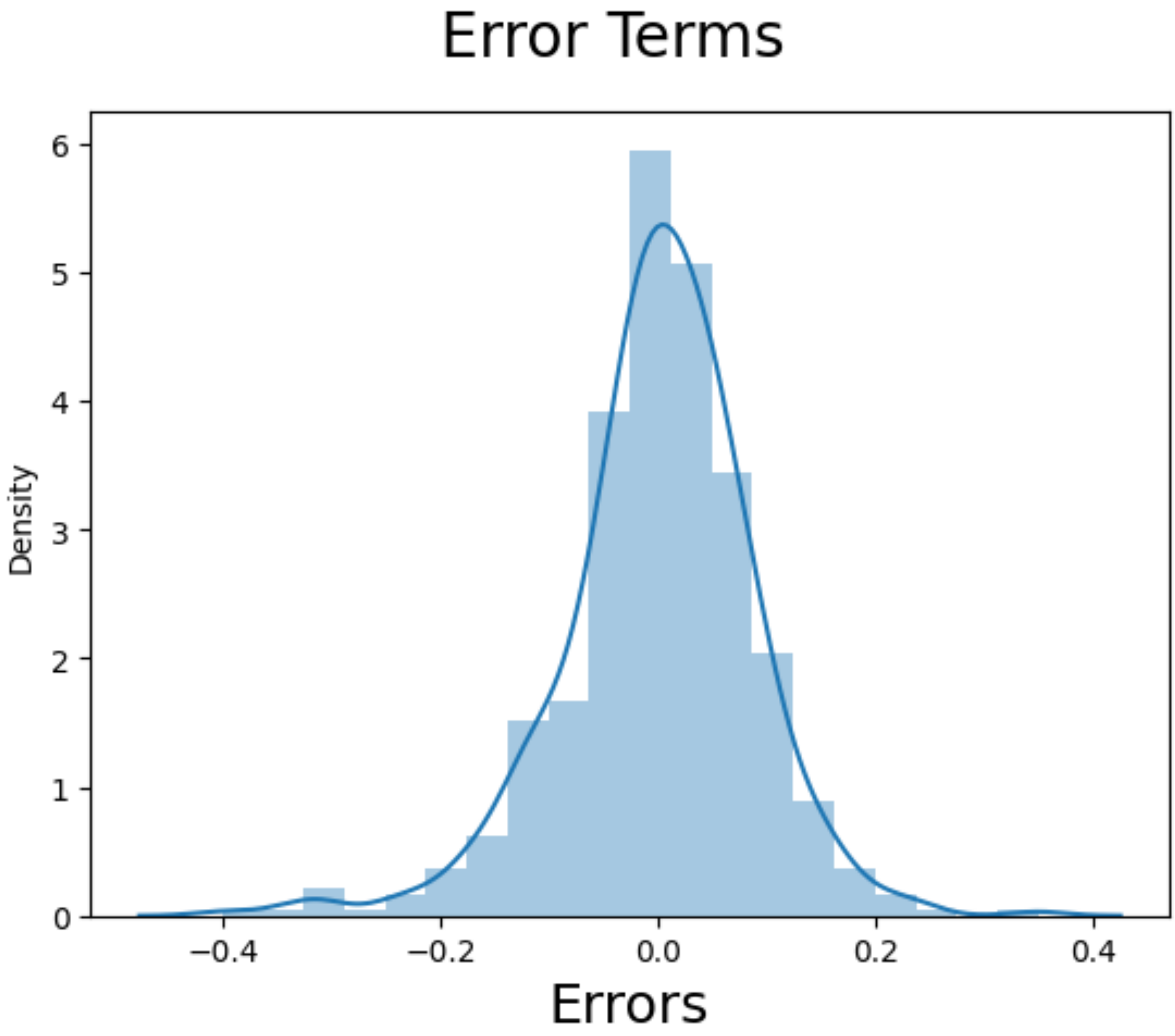
3. Homoscedasticity: The variance of the errors is constant across all levels of the independent variable(s).

   A. Verified this by plotting the final scatter plot between y_test and y_test_pred and the spread is consistent.



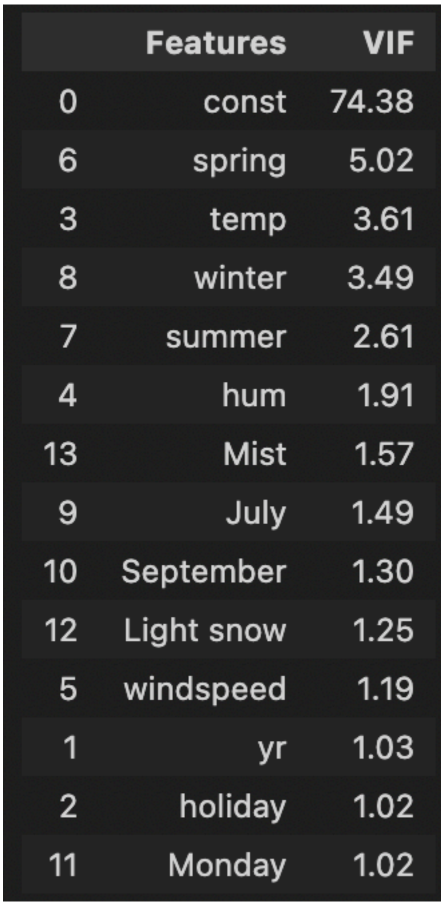
4. Normality: The errors are normally distributed.

   A. Plotted the Dist plot of the residuals to verify that they are normally distributed.



5. No multicollinearity: The independent variables are not highly correlated with each other.

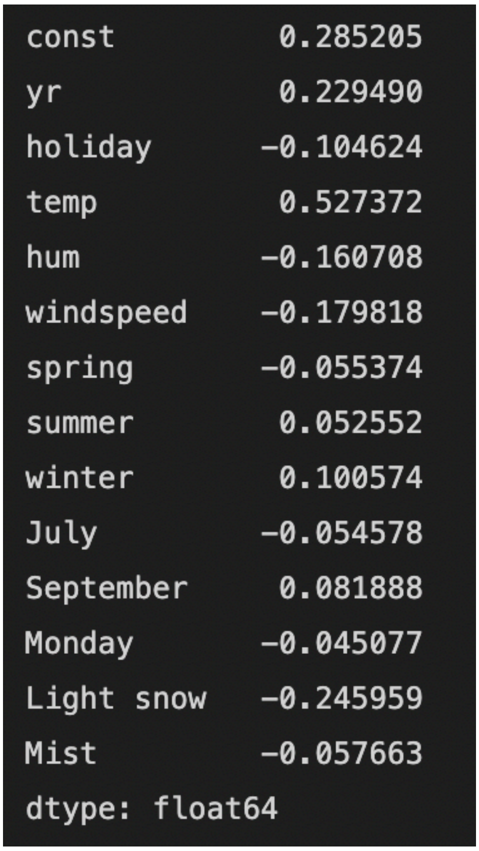
   A. This can be verified through the Variation Inflation Factor in the selected variables by the model. Below is the final list of features selected and their VIF. As we know, VIF less than 5 indicates no or negligible multicollinearity.



	Features	VIF
0	const	74.38
6	spring	5.02
3	temp	3.61
8	winter	3.49
7	summer	2.61
4	hum	1.91
13	Mist	1.57
9	July	1.49
10	September	1.30
12	Light snow	1.25
5	windspeed	1.19
1	yr	1.03
2	holiday	1.02
11	Monday	1.02

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

The coefficients for the final model are as below. From the numbers we can say that ‘cnt’ has positive impact by ‘temp’ and ‘yr’, and negative impact by ‘Light snow’, ‘windspeed’, and ‘hum’.



```
const         0.285205
yr            0.229490
holiday      -0.104624
temp          0.527372
hum          -0.160708
windspeed    -0.179818
spring       -0.055374
summer        0.052552
winter        0.100574
July         -0.054578
September     0.081888
Monday       -0.045077
Light snow   -0.245959
Mist         -0.057663
dtype: float64
```

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear Regression is the study of the relationship among different independent factors on a target variable.

For example, how the weather affects the bike rental sales. How the day of the week affects the bike rental sales. Is there any relationship between day's weather and the month of the year.

We try to answer such questions to study the impact of different factors.

If only one variable impacts the target variable, we can build a simple linear regression model. If more than one variable impacts the target variable (which is usually the case), we need to build a multiple linear regression model.

Linear regression has the following assumptions:

- Linearity: The relationship between the independent variable(s) and the dependent variable is linear.
- Independence: The observations are independent of each other.
- Homoscedasticity: The variance of the errors is constant across all levels of the independent variable(s).
- Normality: The errors are normally distributed.
- No multicollinearity: The independent variables are not highly correlated with each other.

The following steps are performed in the linear regression algorithm to achieve the study:

- **Importing and understanding data**
  - This is the EDA on the given data set to understand what features are available and in what shape and form.
  - We figure out the numerical vs categorical features and identify the steps needed to sanitise the data if needed.

  - We identify the target and the independent variables on which we are going to build the model.

- **Preparing data**

  - here we deal with the missing values either by imputation or by dropping the columns depending on their business use case in the final algorithm.

  - we create dummy variables for the categorical features so they can be used in their numerical form in the model. We then drop the categorical features.

  - we then split data into test and train sets. We build the model on the training set, and test it on the test set. Usually this is a 70-30 split.

  - we then scale and transform the training data set so all the numbers in the data set are on the same comparable scale. We can use the Normalisation or the Standardisation methods to do this.

  - we are now ready to build the model.

- **Building model**

  - the idea of model building is to keep as many of the significant features in the model as possible.

- we can build the model in two ways.

  - Bottom up —> we start with one variable and analyse the model. Then we keep adding one variable at a time and decide if to keep the new variable or drop it. This is feasible when the number of variables are less in the data set.

  - Top down —> build the model with all the variables and drop one variable at a time.

- To achieve the top down approach, we can use the Recursive Feature Elimination (RFE) technique. RFE is available in sklearn package.

- start with recursive feature elimination (RFE), the RFE function eliminates the most insignificant variables one by one. The catch here is that, we need to provide the final number of features where the RFE should stop at. So it makes sense to experiment with different outputs of RFE and arrive at a final feasible number.

- We then build model using statsmodel with RFE features and analyse further if there are any variables to be removed. We use the p-value from the summary of the stats model and the VIF to decide if any feature needs to be dropped.
  - ***Thumb rule for dropping a variable***
    - High p, High VIF --> drop
    - High p, Low VIF --> drop first
    - Low p, High VIF --> drop these after the ones above
    - Low p, Low VIF --> keep them
- It's important to remember to drop only one variable at a time as this could change the p-values and the VIF of the remaining variables. So it's a bit of trial and error.
- The number we are finally looking at is the adjusted R-squared value from the summary of the stats model. When we run the model on the test set, we need the R-squared value to be closer to the model's adjusted R-squared value.
- **Residual analysis**
  - Once we have a model, we need to verify that it's adhering to the linear regression assumptions.
  - For this, plot the error terms in a Dist plot and verify if it's normally distributed.
- **Prediction on the test set**
  - We can now run the model on the test set to verify if the R-squared value on the test set is close to the model's adjusted R-squared value.
- **Model Evaluation**
- we plot the final spread of the test data vs predicted test data to understand the spread.
- From here, we could always make the model better by maybe adding new derived variables in the feature set.

2. **Explain the Anscombe’s quartet in detail.**

A. Anscombe's quartet is a set of four datasets that have almost identical summary statistics, but are vastly different when visualized. The four datasets in Anscombe's quartet are as follows:

1. Dataset I: Two variables, x and y, are closely correlated and have a linear relationship.

2. Dataset II: Two variables, x and y, are closely correlated but have a non-linear relationship.

3. Dataset III: Two variables, x and y, are not closely correlated, but have the same variance.

4. Dataset IV: Two variables, x and y, are not closely correlated and have different variances.

5. All four datasets have the same mean, variance, correlation, and linear regression line. However, when plotted, they reveal very different shapes and patterns, highlighting the importance of visualizing data before analyzing it.

3. **What is Pearson’s R?**
   Pearson's R, is a measure of the linear association between two variables. It is a value between -1 and 1 that indicates the strength and direction of the correlation. A value of 1 indicates a perfect positive correlation, meaning that as the value of one variable increases, the value of the other variable also increases. A value of -1 indicates a perfect negative correlation, meaning that as the value of one variable increases, the value of the other variable decreases. A value of 0 indicates no correlation.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

A. Scaling is the method used to bring all the numerical values in the data set to be comparable to each other. We do this step to keep the data set on the same numerical scale. Otherwise, the coefficients we get after model building can be very varied and can lead to confusion.

For example, if a variable is in 000’s of $, and another is in micro seconds, the coefficients for each of these variables can have a huge variance. This makes model evaluation difficult.

To perform scaling, two methods are commonly used:

Normalisation (Min Max scaling) - this scales the values between 0 and 1. The formula for scaling is given as below.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}.$$

Standardisation - this scales the values to that the result data set has mean of 0 and standard deviation of 1.

In summary, normalization helps to scale data to a given range, while standardization helps to center the data at zero while maintaining the original distribution.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

A. VIF is infinite, meaning there is a very high (almost a perfect) correlation between one or more variables in the model. This happens when two or variables in the model have a strong positive or negative relationship with each other.

Such variables need to be dropped one at a time and the model needs to be reevaluated each time.

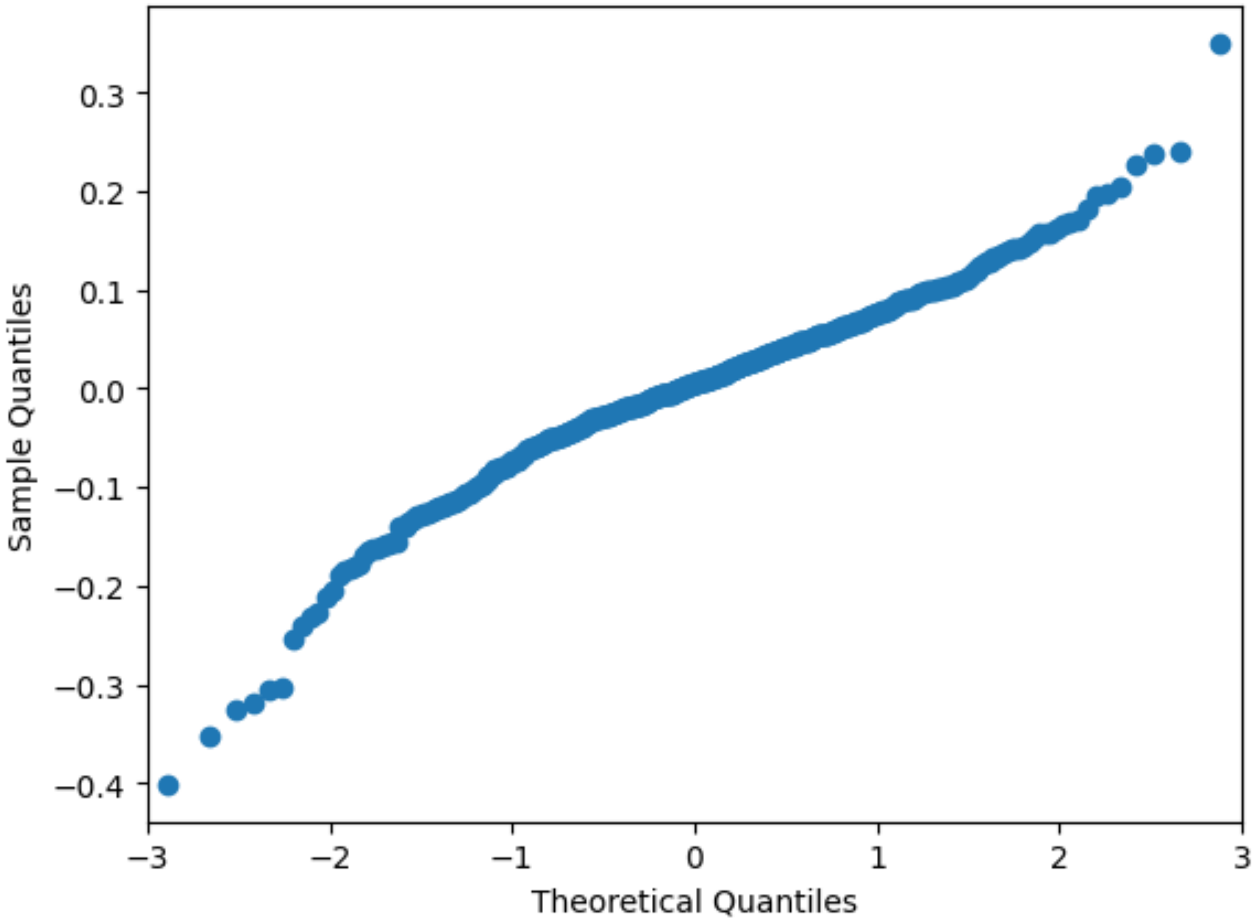
6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

A. QQ Plot is used to check if two data sets have similar characteristics. It is a plot of the quantiles of the first data set against the quantiles of the second data set.

The following characteristics are verified:

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior

In the plot, if the data points are close to a 45-degree line, then the above characteristics are met. The below chart is the qq-plot from the bike-sharing assignment with the training and test data sets.


Sample Quantiles
Theoretical Quantiles