Assignment-based Subjective Questions

Q1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1: 1. Holiday is expected to have less effect on cnt since most of the peaople will be at homes
2. It is observed that on the months of Dec, Jan & Feb usually bike rentals will be low because of christmas days
3. Usually Spring will have low bookings and other seasons are expected to have good bookings

Q2: Why is it important to use drop_first=True during dummy variable creation?

A2: if we use drop_first = True, it will not create a separate column for the first column and helps in not having more columns in the dataframe

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A3: temp' variable has highest correlation with the target variables

Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

A4: 1. R2_score - which will explain how well the model is considering the variance of the data
2. Probability - Prob score obtianed for the each variables is less than 5 % or not, generally expected to have this value less than 0.05 for all the variables
3. VIF - Variance Inflation Factor will also helps in removing the insignificant variables
3. F-statistic - The smaller the value of the this, the better the model

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5: These are the top features which are contributing significantly to the model:
1. temp : with coefficient value of ‘0.4884’ is the variables which contributing at first place as this variables increases bookings also increases
2. precipitation : with coefficient value of ‘-0.2737’ is the variables which contributing at second place as this variables increases bookings decreases since its negative
3. yr : with coefficient value of ‘0.2529’ is the variables which contributing at first place as this variables increases bookings also increases

General Subjective Questions

Q1: Explain the linear regression algorithm in detail.

A1: Linear regression is a supervised machine learning algorithm that is used to predict a continuous value (such as price, salary, or height) based on one or more independent variables (such as age, education, or experience).The linear regression algorithm assumes that there is a linear relationship between the independent variables and the dependent variable.
This means that if we increase/decreases the value of one of the independent variables by a certain amount, the value of the dependent variable will also increase/decreases by a certain amount.

The linear regression algorithm finds the best linear relationship between the independent variables and the dependent variable by minimizing the sum of the squared residuals. The residuals are the differences between the actual values of the dependent variable and the predicted values of the dependent variable.The linear regression algorithm can be expressed in the following equation:

$y = mx + b$

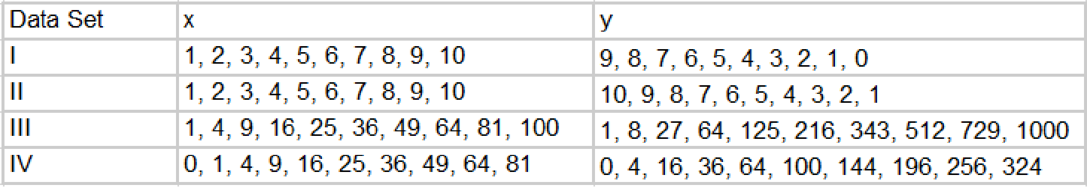
where:
'y' is the dependent variable
'm' is the slope of the line
'b' is the y-intercept

'x' is the independent variable
The slope of the line (m) tells us how much the dependent variable changes for every unit change in the independent variable. The y-intercept (b) tells us the value of the dependent variable when the independent variable is equal to 0.

Q2: Explain the Anscombe’s quartet in detail.

A2: Anscombe's quartet is a set of four data sets that were created by the statistician Francis Anscombe in 1973. The four data sets have the following properties:

They have the same mean, variance, correlation coefficient, and linear regression line.
They look very different when plotted on a scatter plot.
The purpose of Anscombe's quartet is to demonstrate the importance of visualizing data before performing statistical analysis. The four data sets have the same summary statistics, but they have very different distributions. This means that they may not behave the same way when used in a linear regression model.



Data Set	x	y
I	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	9, 8, 7, 6, 5, 4, 3, 2, 1, 0
II	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	10, 9, 8, 7, 6, 5, 4, 3, 2, 1
III	1, 4, 9, 16, 25, 36, 49, 64, 81, 100	1, 8, 27, 64, 125, 216, 343, 512, 729, 1000
IV	0, 1, 4, 9, 16, 25, 36, 49, 64, 81	0, 4, 16, 36, 64, 100, 144, 196, 256, 324

As you can see, the four data sets have the same mean, variance, correlation coefficient, and linear regression line. However, they look very different when plotted on a scatter plot. Data Set I is a perfect linear relationship, Data Set II is a quadratic relationship, Data Set III is a cubic relationship, and Data Set IV is a sinusoidal relationship.

This shows that summary statistics alone are not enough to understand the properties of a data set. It is important to visualize the data before performing statistical analysis.

Anscombe's quartet is a classic example of why it is important to visualize data before performing statistical analysis. It is also a reminder that summary statistics can be misleading.

Q3: What is Pearson’s R?

A3: Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear correlation between two variables. It is a number between -1 and 1, where:

A value of +1 indicates a perfect positive linear relationship, where the variables increase and decrease together.
A value of -1 indicates a perfect negative linear relationship, where the variables increase and decrease in opposite directions.
A value of 0 indicates no linear relationship between the variables.
The closer the value of Pearson's R is to 1 or -1, the stronger the linear relationship between the variables. A value of 0.5 indicates a moderate linear relationship, while a value of 0.3 indicates a weak linear relationship.

Pearson's R is calculated by finding the covariance of two variables and dividing it by the product of their standard deviations. The covariance measures how much two variables vary together, while the standard deviation measures how much each variable varies from its mean.

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A4: Scaling is a data preprocessing technique that is used to transform the values of features or variables in a dataset to a similar scale. This is done to ensure that all features have a similar range of values and that no single feature dominates the learning process.

There are two main types of scaling:

Normalization: Normalization is a scaling technique that transforms the values of features to a range of [0, 1]. This is done by subtracting the minimum value of each feature from all of its values and then dividing by the difference between the maximum and minimum values.
Standardization: Standardization is a scaling technique that transforms the values of features to a mean of 0 and a standard deviation of 1. This is done by subtracting the mean of each feature from all of its values and then dividing by the standard deviation of the feature.

Scaling is performed for a number of reasons, including:

To improve the performance of machine learning algorithms. Many machine learning algorithms are sensitive to the scale of the features, and scaling can help to improve their performance.
To make features comparable. When features have different scales, it can be difficult to compare them to each other. Scaling can help to make features comparable, which can improve the interpretability of machine learning models.
To reduce the impact of outliers. Outliers are data points that are significantly different from the rest of the data. Scaling can help to reduce the impact of outliers, which can improve the accuracy of machine learning models.
The main difference between normalized scaling and standardized scaling is the range of values that they produce. Normalized scaling produces values in the range of [0, 1], while standardized scaling produces values with a mean of 0 and a standard deviation of 1.

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5: The variance inflation factor (VIF) is a measure of multicollinearity in a linear regression model. Multicollinearity occurs when two or more independent variables are highly correlated. When two variables are perfectly correlated, their VIF is infinite.

There are a few reasons why the VIF might be infinite. One reason is that the independent variables are perfectly correlated. This can happen if the variables are measuring the same thing or if they are derived from the same source.

Another reason why the VIF might be infinite is that the dataset is too small. If there are not enough observations in the dataset, the VIF will be inflated.

Finally, the VIF might be infinite if there are errors in the data. If the data is not clean, the VIF will be inflated.

If the VIF is infinite, it means that there is perfect multicollinearity in the model. This can cause problems with the model, such as:

The standard errors of the coefficients will be inflated.
The t-statistics of the coefficients will be reduced.
The p-values of the coefficients will be increased.
The model will be unstable and may not be able to generalize to new data.
If the VIF is infinite, it is important to take steps to reduce the multicollinearity in the model. This can be done by:

Removing one of the correlated variables from the model.
Combining the correlated variables into a single variable.
Using a different type of regression model, such as ridge regression or lasso regression.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6: A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool used to compare two probability distributions. It is a scatter plot of the quantiles of one distribution against the quantiles of another distribution.

In linear regression, a Q-Q plot can be used to assess whether the residuals (the difference between the actual values of the dependent variable and the predicted values of the dependent variable) are normally distributed. If the residuals are normally distributed, then the Q-Q plot will be a straight line.

A Q-Q plot can be used to identify a number of problems with the residuals, including:

Non-normality: If the Q-Q plot is not a straight line, then the residuals are not normally distributed. This can be a sign of problems with the model, such as outliers or heteroskedasticity.
Outliers: Outliers are data points that are significantly different from the rest of the data. Outliers can cause the Q-Q plot to deviate from a straight line.
Heteroskedasticity: Heteroskedasticity is a violation of the assumption of homoskedasticity, which means that the variance of the residuals is constant. Heteroskedasticity can cause the Q-Q plot to have a curved shape.
If the Q-Q plot shows any of these problems, then it is important to investigate the issue and to take steps to correct it. This may involve:

Removing outliers from the dataset.
Using a different type of regression model, such as robust regression.
Including a variable to control for heteroskedasticity.
By using a Q-Q plot, you can assess whether the residuals are normally distributed and identify any problems with the residuals. This can help you to improve the accuracy and reliability of your linear regression model.