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

- We see that fall season have brought in more booking. The booking count also increased.
- During clear weather we see more bookings.
- During Thursday, Friday, Saturday and Sunday we find more bookings.
- More bookings are during the month of May, June, July, August, and September and October.
- We see during model building on inclusion of categorical features like year and season we can see a significant growth in value of R squared and adjusted R Squared. This helps us know that categorical features can help us in explaining a great proportion of variance in a dataset.

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

Use drop first=True while creating dummy values (dummy encoding), else we will obtain a redundant feature.

As this data is of the categorical nominal type, we must map the seasons of spring, summer, fall, and winter before we can use these column values to create dummy variables.

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

‘registered’ has the highest correlation with the target variable.

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

By plotting the residuals' distribution against the dependent variable's levels. The divergence of the residuals from a normal distribution can be determined using a QQ-plot.

It may point out a problem if the resulting curve is distorted and not normal.

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

temp, winter, sep are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A linear model, such as one that assumes a linear relationship between an input variable (x) and a single output variable, is known as linear regression (y). More specifically, the linear combination of the input variables can be used to calculate that "y" (x).

The process is known as a simple linear regression when there is just one input variable, "x". Numerous linear regression is the term used in statistical literature to describe the technique when there are multiple input variables.

The linear regression equation can be prepared and trained using a variety of strategies, one of which is known as Ordinary Least Squares. The straightforward linear model would be depicted by

y = B0 + B1*x

2.Explain the Anscombe’s quartet in detail.

Anscombe’s quartet emphasizes the significance of plotting data to verify the accuracy of the model fit.

The Pearson correlation between the variable's 'x' and 'y' values in a single panel is the same, r =. 816. Moreover, the mean and variance values of the variable’s "x" and "y" in the four independent data sets are identical.

3.What is Pearson’s R?

Between +1 and -1 are the possible values for the Pearson correlation coefficient, or r.

It is a statistical model that evaluates two variables in bivariate form.

As long as the variables under study are both quantifiable, Pearson's correlation can be employed to assess the associative research hypothesis.

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

Scaling is the process of measuring and allocating numbers to items in accordance with predetermined guidelines.

It is a data pre-processing procedure used to normalize data within a specific range by applying independent variables.

When normalized Scaling is based on the minimum and maximum value of the characteristics.

Scaling is performed using the mean and standard deviation.

When characteristics are of different scales, normalization is utilized.

When we wish to guarantee a zero mean and a single standard deviation, we utilize standardization.

Normalisation the range [0, 1] or [-1, 1] for the scale.

The scope of standardization is not restricted.

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

VIF = infinity if there is a perfect correlation. A high VIF score denotes a strong connection between the variables. The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4.

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

A graphical method for determining if two data sets originate from populations with a common distribution is the quantile-quantile (q-q) plot.

Determine whether the collection of data may reasonably be attributed to the theoretical distribution using the Q-Q plot. exponential, uniform, or normal distribution.