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?

Ans:

- Fall season seems to have attracted more booking overall. Adding to that in each season the booking count has increased drastically from 2018 to 2019.

- Most of the bookings have been done during the month of May, June, July, August, September and October. Trend seems to be increasing from starting of the year till mid of the year peak is June and then it started to decrease as we approach the end of year. Number of booking for each month seems to have increased from 2018 to 2019.

- Booking is higher during non-holidays.

- Thursday, Friday, Saturday has more number of bookings as compared to the start of the week.

- On a working day we seem to have a bit increased booking compared to holiday/weekend. As usual we see drastic increase in booking in 2019 when compared to 2018.

- Clear weather attracted more booking. In comparison to 2018 booking increased for each weather situation in 2019.

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

Ans:

If you set drop_first = True , then it will drop the first category. So if you have K categories, it will only produce K – 1 dummy variables

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

Ans:

Temp and atemp has highest correlation.

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

Ans:

Validate the assumptions using VIF, Error Terms, and Residual analysis.

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

Ans:

Top 3 features are temp, atemp, Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. After reading and understanding the data, train and test slit is done using that we check R-value, P-value, VIF, and Collinearity.

2. Explain the Anscombe’s quartet in detail.

Ans:

A regression model is not always necessarily an exact one; it can also be fooled by some data! In certain cases, there are multiple datasets which are completely different but after training, the regression model looks the same. A group of four such datasets having identical descriptive statistics but with some peculiarities is the Anscombe's quartet.

3. What is Pearson’s R?

Ans:

Pearson's correlation coefficient, also known as Pearson's R, is a measure of the strength of correlation between two variables. It is commonly used in linear regression. The value of Pearson's R always lie between -1 and +1.

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

Ans.

Scaling is necessary for a model to be functional with the appropriate range of coefficients. There are two types of scaling:

Normalized scaling: This scaling is done to make the distribution of data into a Gaussian one. It doesn't have a preset range. Typically used in neural networks broadly.

Standardized scaling: The example given above is of standardized scaling. Here, the value of variable(s) is/are compressed into a specific range to suit the model.

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

Ans:

If there is a perfect correlation between the dependent variable and independent variable(s), the R-squared value comes out to be 1. Hence VIF, which is (1/ (1-R^2)) turns out to approach infinity.

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

Ans:

Q-Q plot is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model.