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- Effect of categorical variable on dependent variable “cnt” -

season: Almost 30% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

mnth: Almost 12% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 65% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday: weekday variable shows very close trend (between 13%-15% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable.

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

Ans- To drop the first dummy variable for each set of dummies created. (To get k-1 dummies out of k categorical levels by removing the first level.)

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

Ans- 'temp’ has the highest correlation with target variable ‘cnt’.

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

Ans- By Residual Analysis of training data. If Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

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-

1. Temperature (temp)
2. Weather Situation 3 (weathersit_3)
3. Year (yr)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans- Linear regression in machine learning is a supervised learning technique. If there is a "linear relationship" between two or more variables, then we can use historical data to find out the "routine" between the variables and build an effective model to predict the future variable results.

2. Explain the Anscombe’s quartet in detail.

Ans- The Anscombe quartet is comprised of four scatterplots that have nearly identical correlations, as well as means and standard deviations, but disparate shapes. These graphs show the crucial role that data visualization plays in developing a sensible statistical model. The negative option is included to produce the mirror image graph.

3. What is Pearson’s R?

Ans- The correlation between two variables reflects the degree to which the variables are related. When computed in a sample, it is designated by the letter "r" and is sometimes called Pearson's r.

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

Ans- Scaling is a personal choice about making the numbers feel right, e.g. between zero and one, or one and a hundred. For ex: converting data given in millimeters to meters because it's more convenient, or imperial to metric.

Normalization : Scaling some data to a confined range.

Standardization: we can transform the data into a range such that the new population has mean (average) = 0 and standard deviation = 1.

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

Ans- If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

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

Ans- qqplot(x) displays a quantile-quantile plot of the quantiles of the sample data x versus the theoretical quantile values from a normal distribution. If the distribution of x is normal, then the data plot appears linear.

qqplot plots each data point in x using plus sign ('+') markers and draws two reference lines that represent the theoretical distribution. A solid reference line connects the first and third quartiles of the data, and a dashed reference line extends the solid line to the ends of the data.