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. Categorical Variables such as year, good weather has a positive affect on dependent variable and variables such as bad weather, holiday has negative affect on dependent variable.

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

Ans. It is important to use drop_first=True because it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Ans. Looking at the pair-plot among the numerical variables, temp has highest correlation with the target variable.

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

Ans. I validated the assumptions of Linear Regression Model by checking error and whether they are normally distributed. Also we saw with VIF calculations that there is no correlation between independent variables.

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

Ans. The top 3 features are

1. Temperature
2. Year
3. Bad Weather

General Subjective Questions

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

Ans. Algorithm is as follows

- First get the data. Check its quality and do necessary steps to improve its quality.
- Identify the target variable
- Check the relation between other variables on target variable.

- Identify if there are any categorical variable. If there are then convert them to nominal variables using one – hot encoding or dummy variable creation method
- Remove all the unnecessary columns from the data
- Create train and testing data
- For training data, scale the numerical variables to the appropriate ranges.
- Build the model on training data.
- Check the model performance r2 score, p value of columns and VIF for correlation between independent columns. Do necessary steps to remove collinearity by removing/changing such columns and keep on repeating this step till you get acceptable model
- Do residual analysis on the final model and check if all the assumptions of Linear regression holds true.
- Evaluate the model using test data

2. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. The mainly demonstrates that how graphing is important.

3. What is Pearson's R?

Ans. Pearson's correlation (also called Pearson's R) is a correlation coefficient commonly used in linear regression. when anyone refers to the correlation coefficient, they are usually talking about Pearson's.

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

Ans. It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. Data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units which can lead to an important variable getting overshadowed by less important one. hence incorrect modelling.

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

Ans. VIF infinite means a perfect correlation. This becomes infinite when the variables shows perfect linear combination which lead R2 = 1. Therefore VIF = 1/(1-1) which is infinite.

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

Ans. Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. It mainly helps us to understand whether two datasets are similar or not.