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. By analysing each categorical variable with the target variable i.e., ‘cnt’, I made following inferences:

   - ● Year- The count of bike rentals in year 2019 has significantly increased as compared to year 2018. In coming years we can expect the similar trends.
   - ● Holiday- The median of counts is more for the non-holidays as compared to the holidays.
   - ● Month- September , October and November has high bike rentals. In January bike rentals is lowest.
   - ● Season- The fall season has high bike rent demands while it is least in spring .
   - ● Weathersit- The bike rentals are high when weather is clear and it is low in bad whether.
   - ● Weekday- Since median is almost constant. there is almost no variation of count of bike rentals with week days.

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

   Ans. It reduces the extra column created during dummy variable creation. Hence it reduces the correlations 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:

   - Plotting distribution of error terms to confirm that they are normally distributed.
   - Homoscedasticity can be validated by scatter plot of residual values vs. predicted values.
   - Also, we verified with VIF calculations that no variable has value more than 5 showing that there is no multicollinearity between independent variables in the final model.

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

   I. Temperature

   II. weathersit_Light Snow & Rain

   III. Year

General Subjective Questions

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

Ans. linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables. linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Mathematically the relationship can be represented with the help of following equation –
Y = mX + b

- Y is the dependent variable we wish to predict
- X is the set of dependent variables we are using to make predictions
- m is the slope of the regression line which represents the effect X has on Y
- b is a constant, known as the Y-intercept, If X = 0, Y would be equal to b.

The linear relationship can be positive or negative in nature.
Simple Linear Regression (SLR): It is the most basic version of linear regression which predicts a response using a single feature. The assumption in SLR is that the two variables are linearly related.
Multiple Linear Regression (MLR): It is the extension of simple linear regression that predicts a response using two or more features.

Following are the steps involved in Linear Regression problem:
- First get the data and perform quality checks on it.
- Identify the dependent variables and the target variable.
- Find relations between dependent variables on target variable using EDA.
- For Categorical variables, make dummy variables.
- Drop all the irrelevant columns from the data
- Split the dataset into train and testing data
- For training data, scale the numerical variables to the appropriate ranges.
- Build model on training data.
- Check model performance R^2 score, p value of columns and VIF for correlation between independent columns. Perform iterative steps to remove collinearity and insignificant columns. Keep on repeating this step till you get an acceptable model.
- Do residual analysis on the final model and check if all the assumptions of Linear regression hold true.
- Evaluate the model using test data.

2. **Explain the Anscombe's quartet in detail.**

Ans. Anscombe’s quartet comprises of **four datasets that have nearly identical simple statistical properties, yet appear very different when graphed**. Each dataset consists of eleven (x,y) points. It shows the importance of graphing data before analysing it and the effect of outliers on statistical properties

3. **What is Pearson’s R?**
   Ans. Pearson's correlation (also called Pearson's R) is a correlation coefficient. It is a measure of linear correlation between two sets of data.

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.

   The two most common ways of rescaling are:

   - **Min-Max scaling**: It brings all of the data in the range of 0 and1.
   - **Standardisation**: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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.** It is used to compare the shapes of distributions, providing a graphical view of how properties are similar or different in the two distributions. It is is a scatterplot created by plotting two sets of quantiles against one another. It mainly helps us to understand whether two datasets are similar or not.