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?**

- Fall has the highest demand for bikes than any other season
- Demand for bikes has grown during 2019
- The month of September has the highest demand for bikes than other months
- Demand for bikes is continuously growing from the month of Jan to the month of June
- We do not get a clear picture of demand during the weekdays
- Good weathersit has highest demand for bikes
- Bike sharing is highest in September. But it falls during beginning and ending months of the year before that

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

- We are dropping the first column as p-1 can explain the p-categories of dummies in season, month and weekday
- We can lose the info about severe weather situation, hence we have not used the drop_first in weathersit

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

Temp and atemp has the highest correlation with the target variable cnt.

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

I validated the assumptions of linear regression after building the model on the training set in the following ways.

- Carried out residual analysis to check if the error terms are normally distributed and their mean is 0

- By drawing the heatmap and checking correlation, I was able to check if there was any significant linear relationship between the target variable cnt and independent variables in the test set.
- Drew the scatter plot to check for linearity between actual test and predicted test set. This helped me find out if there is homoscedasticity. By plotting error terms in a scatter plot, I found out that there was no any pattern between error terms of predicted and actual test set.
- Made a line chart comparing the error terms of actual and predicted test set. This helped me see if there were any deviations between both of them.
- Finally, I found out the r-squared for test set and it was closer to the training set. This helped me validate the assumptions of linear regression.

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

The most essential and top 3 features that contribute significantly to the demand of bikes are:

- Temperature
- Season
- Month

General Subjective Questions

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

   Linear regression algorithm is a Machine Learning algorithm that is majorly used in supervised learning. This algorithm helps us find out if there is a linear relationship between the dependent and independent variables.

   Pros

   Linear regression algorithms help us find out or predict the crucial drivers that contribute to the change of a target variable. This helps us make measures to those drivers and improve the performance of our target variable.

Cons

Linear regression algorithms cannot help us to forecast what sales will be in the future. It cannot run analysis on the time-series data. Outliers are the major factor that affects linear regression algorithms. Sometimes it oversimplifies the relationship between real-world variables.

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

The most crucial task before building any model is to plot dataset and check the descriptive statistics. Anscombe's quartet is a situation where four datasets have similar or identical descriptive statistics measures. But they vary in distributions when we visualize them as scatter plots. It emphasizes on visualizing a given dataset before we train and test the model. By visualizing the data sample distributions, we can spot the anomalies and rectify them before building the model.

3. **What is Pearson's R?**

Pearson correlation coefficient or Pearson's r, helps us find out the if there is a linear relation or association between two variables. Being the most commonly used correlation measure, it conveys how strong two variables are related to each other. It helps us know whether we can draw a line graph to represent a data.

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

Variables in the model may vary in magnitude. For example, an independent variable income would be in thousands. But the revenue of the company being another independent variable would be in crores. These varying measures among independent variables would lead to extremely high or extremely low beta coefficient values. Scaling helps us standardize the independent variables in a fixed range. It affects just the coefficients and not t-test, p-value or r-square values.

There are two types of scaling performed. They are normalized scaling (min-max scaling) and standardized scaling.

Standardisation brings all the data into a standard normal distribution with mean 0 and standard deviation 1. MinMax scaling, on the other hand, brings all the data in the range of 0-1. In real-world, we majorly use the minmax scaling or normalization for scaling the independent variables.

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

If there is a perfect correlation between two independent variables, then the Variance Inflation Factor or VIF becomes infinite. It also occurs when corresponding independent variables are expressed by linear combination of other independent variables.

When we get r2 value as 1, it would result in following infinity.

VIF = 1/1-r2

When r2 is 1, 1/1-1 = 1/0.

The value is 1/0 is infinite. Hence, the best way to counter this is to remove one of the variables that causes this multicollinearity.

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

Q-Q plots are quantile-quantile plots that help us intuitively decide whether a variable is normally distributed or not. It plots the quantiles of theoretical distribution and quantiles of sample distribution to check the normality of a variable distribution. Quantiles are data divided into four breakpoints or buckets (0-25%, 25-50%, 50-75%, 75-100%).

Use and importance of Q-Q plot in linear regression

1. It helps us to measure the skewness of the distribution
2. This plot helps us find out if error terms or residuals follow a normal distribution. It is a crucial assumption of linear regression models.

3. It enables us to decide the sampling if two populations are of the same distribution.