# Assignment-based Subjective Question and Answers

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 “season”, “month”, ”weathersit”, ”workingday”, ”weekday” all have a significant effect on the dependent variable.

Month and season show the same story, saying on a particular season like summer, and fall the demand of the bike will be high and with the help of working day, we can say that from Monday to Saturday the demand of the bikes will be less compared to weekends.

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

**Ans:** While creating the dummy variables we should be using **drop_first = True** as it helps to reduce the extra columns which will get created during the creation of the dummy variable and that results in a reduction of correlation among columns.

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

**Ans:** By looking at the Pair plot we can say that **temp** and **temp** have the 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:** We can validate the **Linear and Additive** assumption by looking for residual vs fitted value plots.

We can validate the **Autocorrelation** assumption by looking at the Durbin – Watson (DW) statistic it must lie between 0 to 4 when the DW is greater than 0 (0 < DW) and lesser than 2 (DW < 2) this is positive autocorrelation, and when the DW is greater than 2 (DW> 2) and lesser than 4 (DW < 4) this is negative autocorrelation.

We can validate the **Multicollinearity** assumption with the help of scatter plot to visualize the correlation effect among variables or we can use VIF.

We can validate the **Heteroskedasticity** assumption by using residual vs fitted values plot.

We can validate the **Normal Distribution of error terms** assumption by using the QQ plot or histogram.

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

**Ans:** Based on the final model, the top 3 features are **year**, **August**(month), **September**(month).

## General Subjective Questions and Answers

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

**Ans:** We use Linear regression for the prediction of the continuous variables. Linear regression is a part of ML in which we train the model using the historical data to predict the trend.

Linear regression also helps us to find the effect of any input variable on the target variable.

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

**Ans:** Anscombe's quartet comprises of 4 data sets, which is an almost identical simple descriptive statistic, but still has very different distribution and looks completely different when graphed.

Each of the data-set contains 11 (x,y) points, we use a scatter plot to look at this dataset.

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

**Ans:** Person's R is a summary of numerical variables, by which we can understand the linear strength between two variables. We can say if both the variables move together up and down in a pattern that means the correlation coefficient will be positive, and if the two variables move in the opposite direction one having low value and one having high value then we can say, the correlation coefficient will be negative.

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

**Ans:** Scaling is a technique which we use in ML models. We use it to set a standard range between 1 to 0 for the independent variables. This technique helps us to overcome the problem of outlines in categorical data.

Scaling is performed to standardize the independent variables.

In normalized scaling, independent variables are standardized between the range from 1 to 0, and in standardized scaling, variables are re-scaled so it has distribution with 0 mean value and variance equal to 1.

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

**Ans:** Sometimes VIF becomes infinite because it's an indication that says the corresponding variable may be expressed exactly by a linear combination of the other variables.

In other words, it shows a perfect correlation between two independent variables, which means we will get R2 = 1. We can solve this by dropping any of the one variables which is causing this perfect multicollinearity.

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

**Ans:** Q-Q plot (Quantile-Quantile plots), these plots are plotted using two quantiles against each other. Q-Q plots are used to find out if the two sets of data come from the same distribution or not.

In linear regression, this can be used when the train and the test data sets are collected. Separately but from the same population distribution to find out this, we can use the Q-Q plot.