# 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? (3 marks)

**A:** There are some categorical variables which have some significant effect on the dependable variables. Which are:

Month categories where July and September were effected, with increase of negative coefficient, which is decrease in bike hires and similarly in September there was certain increase of bike hires. The some other factors are Light_rainsnow, Misty, Summer and winter where there was some variance in the bike hiring.

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

**A:** drop_first=True is really important command or code to use because it helps in reducing the extra columns that were created during creating the dummy variables. So it is used to reduce the correlations within the dummy variables created.

## 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

**A:** Among the numerical variables the “temp” i.e., temperature has the highest correlation with the target variable i.e, “cnt” which is count of bike hires.

## 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

**A:** To validate the assumptions of linear regression after building the model on the training set is to perform some steps they are:

The Linear relationship, which is a scatter plot which was plotted between independent and dependent variable one on one each other respectively, a straight line passing through the points could be observed. Later on Homo scedasiticity will observe the variance of error terms and concludes the error in variance is constant. Absence of multicollinearity is on heatmap and VIF was used.

Durbin Watson test was conducted for independence  of residuals. And the histograms and Q-Q plots were also used for normality of errors

### 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

**A:** The top three features that contributing significantly towards explaining the demand of shared bikes were:

1. Temperature
2. Year
3. Light_rainsnow

## General Subjective Questions

### 1. Explain the linear regression algorithm in detail. (4 marks)

**A:** The Linear regression attempts to model the best relationship between variables by fitting a linear equation to observed data. In this model one variable is called as the dependent variables and all others are the explanatory variables.

There are some steps in the linear regression algorithm. They are:

1. Analysis and conversion of variables, In this the variables must be converted to required format like categorical variables  and also should understand the correlation and the directionality of the data.
2. Dividing the model into the test and train datasets, In this we should divide the train and test datasets from the given whole data mostly 70% of data was divided into training dataset and remaining 30% was used to test the model of dataset.
3. Estimating the model which is of fitting the line, In this model is estimated which has the best representation off maximum points in a straight line i.e, linear line. Later on we check the assumptions of linear regression model to know the purpose of model

4. Evaluating the validity and accuracy of the model, In this the model is set to run and test the testing dataset to obtain the $R^2$ and some other factors.

## 2. Explain the Anscombe’s quartet in detail. (3 marks)

**A:** It comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions when they were graphed.

There are 4 dataset plots which have nearly same statistical observations, which provides same statistical information that involves variance, mean of all x, y points in all 4 datasets. This explains about the importance of visualising the data before applying the various algorithms to build models

## 3. What is Pearson’s R? (3 marks)

**A:** It is the correlation coefficient and it is also known as the best method of measuring the association between variables of interest because of the method of covariance. It is a test statistics that measures the statistical relationship or association between 2 continuous variables. This gives us a multi dimensional information- the magnitude of the association and correlation and the direction of relationship.

## 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

**A:** The scaling is the part of data pre-processing which is applied to independent variables to normalize the data within the particular range, it also helps in spending up the calculations in the algorithm. This is done when the collected dataset contains variables with highly varying magnitudes, units and ranges. If scaling is not done correctly then the algorithm takes only the magnitude in account and not units hence incorrect modelling, this brings all variables to the same level of magnitude. Normalization scales a variable to have a value between 0 to 1. While standardization transforms data to have a mean of zero and standard deviation of 1.

## 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

**A:** An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables which are the variables which show an infinite VIF as well

## 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

**A:** The quantile-quantile plot which is Q-Q plot in general called is a graphical technique for determining if 2 datasets come from populations with a common distribution. A Q-Q plot of the quartiles of the first dataset against the quantiles of second dataset. By a quantile we mean the fraction of points below the given values. That is the 0.3 or 30% quantile is the point at which 30% of data fall below and 70% fall above that value. The purpose of Q-Q plot is to find out if 2 sets of data come from the same distribution, a 45 degree angle is plotted on the Q-Q plot, if 2 datasets come from a common distribution then the points will fall on that reference line.