## 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)

**Ans:** Categorical variables are as follows: Season,Year,weathersit,Month,weekday and working day.

- Season : High demands for rentals seem to be in summer and fall season . Higher targets can be planned in Summer and fall .
- Month : Bike rentals seen in high demans from June to October months.
- Year : 2 Year data is available and there is high demand in 2019 as compared to 2018
- Weathersit : Clear Weathersit conditions seems to be more favorable for bike rental demands.
- Working day: Bike demands seems to be slightly higher on working day as compared to weekend or holiday.
- Higher demands on Thursday and Friday as compared to other days

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

Ans:

- It is important to use drop_first = True as it will help reduce the extra column created during creation of dummy variable.
- When we create dummy variables , new columns gets generated. If we don't use drop_firs = True, n dummy variables will be created which are highly correlated in themselves.

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

**Ans:** From the pair plot of numerical variables and the target variable (cnt), temp is highly correlated with the target variable followed by atemp.

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

   Ans: Validated the assumptions by checking below assumptions in Model preparation.
   Also plotted the scatter plot which confirms the same.

   - Error terms have normal distribution – Residual analysis
   - Error terms have a constant variance.
   - and by checking Linear relationship between dependent and independent variables

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

**Ans: Top 3 features** contributing significantly towards explaining the demand of the shared bikes are weathersit, year and season.

- **Weathersit :** Temperature is the most significant feature which affects the bike demand positively.The weather conditions as rain,cloudy weather and humidity affects it negatively.
- **Yr:**Bike demands increased for the current year as compared to last year

- Season: Summer and Fall season has highest bike demands.

## General Subjective Questions

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

**Ans:** Linear regression is a supervised machine learning technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. This algorithm finds a linear equation that best describes the correlation of independent variable with dependent variable. This is achieved by fitting a line to the data plotted using least squares technique.

The best fit line tries to minimize the sum of the squares of the residuals. The residual is the distance between the line and the actual value of the explanatory variable. Finding the line of best fit is an iterative process.

The standard equation of the regression line is given by the following expression:

$$Y = \beta_0 + \beta_1.X$$

Where, $\beta_0$ is the intercept, $\beta_1$.is the slope or gradient, X is the independent variable and Y is the dependant variable that we are trying to predict.

Linear Regression are of two types:
1. Simple Linear Regression
2. Multiple Linear Regression

A linear regression model helps in predicting the value of a dependent variable, and it can also help explain how accurate the prediction is. This is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the explanatory variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0.

3. Explain the Anscombe's quartet in detail. (3 marks)

**Ans**:

**Anscombe's quartet** comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

**Purpose of** Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

Anscombe's quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

4. What is Pearson's R? (3 marks)

**Ans:**

The Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value $r = 1$ means a perfect positive correlation and the value $r = -1$ means a perfect negataive correlation.

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

**Ans**:

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

**Need for Scaling:**

Most of the times, collected 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 hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

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

Ans: The VIF formula clearly signifies when the R square will be 1. If the R square will be 1, then VIF is infinite. R square will be exactly 1 when there is a perfect correlation between 2 independent variables.

6.What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Ans:
Q-Q plots are graphical tools that help you assess the validity of some assumptions in regression models, such as normality, linearity, and homoscedasticity. A Q-Q plot, short for quantile-quantile plot, is a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution.

**Importance of Q-Q plot:**
A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, you can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. You can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.