# Assignment-based Subjective Questions and Answers

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:**

   a. There is an increase in the bike rental count in spring and summer seasons and then there is a decrease in the bike rental count in fall and winter season.
   b. The demand for rental bikes increased in the year 2019 compared to year 2018.
   c. Month June to September is the period when bike demand is high. January is the lowest demand month.
   d. Bike demand is less in holidays in comparison to non-holidays.
   e. The demand for rental bikes is almost similar throughout the weekdays.
   f. There is no significant change in bike demand with working days and non-working days.
   g. The bike rental count
      a. is the highest during clear, partly cloudy weather
      b. second-highest during misty cloudy weather,
      c. 3rd highest, during light snow and light rain weather.

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

   **Ans:**

   drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

   **Ans:**

   a. The “temp” and “atemp” variables are highly positively correlated to each other, it means that both are carrying the same information.
   b. The total_count, casual and registered are highly positively correlated to each other.

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

   **Ans:** The original dataset has been split into two datasets called training and validation/testing. The training and test data split was done as 80:20 ratio. Scikit-Learn’s “train_test_split” function has been used to split the data. Linear Regression model is then trained with training dataset. Later is validated with test dataset. R2 Score is used for evaluation.

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:**

   Based on coefficients of Linear Regression model, following are top 3 features:

   a. Year
   b. Temp
   c. Weather situation

# General Subjective Questions and Answers

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

   **Ans:**

   a. Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

   b. When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.

   c. Similarly, when there is only one dependent variable, it is considered Univariate Linear Regression, while when there are more than one dependent variables, it is known as Multivariate Regression.

   d. Simple Linear Regression: The equation for simple linear regression is:
   $Y=c+mX$
   where:
      I. Y is the dependent variable
      II. X is the independent variable
      III. c is the intercept
      IV. m is the slope

   e. Multiple Linear Regression: The equation for multiple linear regression is:
   $Y=c+m_1X_1+m_2X_2+m_3X_3.....+m_nX_n$
   where:
      I. Y is the dependent variable
      II. $X_1, X_2, \ldots, X_n$ are the independent variables
      III. c is the intercept
      IV. $m_1, m_2, m_3, \ldots, m_n$ are the slopes

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

   **Ans:**

   Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fool the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly the same statistical observations, which provide the same statistical information that involves variance, and mean of all x,y points in all four datasets.

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

   **Ans:**

   The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

I. Pearson's *r*
II. Bivariate correlation
III. Pearson product-moment correlation coefficient (PPMCC)
IV. The correlation coefficient

The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

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

**Ans:**

I. 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.
II. Most of the times, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the 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. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.
III. Normalized / Min-Max scaling vs Standardized scaling:
    a. Normalized: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.
    b. A Min-Max scaling is typically done via the following equation

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

    c. 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). sklearn.preprocessing.scale helps to implement standardization in Python.

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

Ans:

If there is perfect correlation, then VIF (Variance Inflation Factor) = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) = infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other 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)**

   Ans:

   a. Q-Q plot (Quantile-Quantile plots) is a plot of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
   b. The purpose of Q-Q plot is to find out if two sets of data come from the same distribution.
   c. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.