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?**

   Analysis of the categorical variables is done using plots. Below are few points we can infer from the visualization.
   a) Fall season seems to have attracted more bookings.
   b) Clear weather attracted more booking.
   c) Saturday and Sunday have a greater number of bookings.

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

   Drop_first=True 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?**

   'Temp' variable has 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**?

   Validation is done as following:

   a) The training and testing accuracy are nearly equal hence there is no Overfit/Underfit situation.

   b) The predicted values have linear relationship with the actual values.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   a) Temp
   b) Yr
   c) winter

General Subjective Questions

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

   It is a machine learning algorithm where we train a model to predict the behaviour of your data based on some variables. It has two variables which are on the x-axis and y-axis should be linearly correlated.

   Mathematically, we can write a linear regression equation as:

   $$y = a + bx$$

   Where a and b given by the formulas:

   $$b(slobe) = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

   $$a(intercept) = \frac{n \sum y - b(\sum x)}{n}$$

   Here, x and y are two variables on the regression line.

   b = Slope of the line

   a = y-intercept of the line

   x = Independent variable from dataset

   y = Dependent variable from dataset

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

   four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

   Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. Those 4 sets of 11 data-points are given below.

   After that, the council analysed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

```
+-------+--------+-------+-------+-------+-------+-------+------+
|   I            |       II      |      III       |      IV     |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
----+---------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

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

It a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations. In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

- r = 0 means there is no linear association

- r > 0 < 5 means there is a weak association

- r > 5 < 8 means there is a moderate association

- r > 8 means there is a strong association

Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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

It 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.

Why 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.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead 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.

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

Q-Q Plots (Quantile-Quantile plots) are plots 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.

The purpose of Q Q plots is to find out if two sets of data come from the same distribution. 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.

A Q Q plot showing the 45 degree reference line: