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?
   Ans: From the analysis of the categorical variables from the dataset it could be inferred the bike rental rates are likely to be higher in summer and the fall season, are more prominent in the months of September and October, more so in the days of Sat, Wed and Thurs and in the year of 2019.
   Holiday consumption of bikes if compared within "registered" and "casual" users then the observation is "casual" users are using bikes more on holiday.
   Weathersit – Most favourable weather condition is the clean/few clouds days. Registered users count is comparatively high even on the light rainy days, so the assumption can be drawn that the bikes are being used for daily commute to the workplace.

2. Why is it important to use drop_first=True during dummy variable creation?
   Ans: drop_first=True helps in reducing the extra column created during the dummy variable creation and hence avoid redundancy of any kind.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
   Ans: The 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?
   Ans: Validated the assumptions of linear regression by checking the VIF, error distribution of residuals and linear relationship between the dependent variable and a feature variable.
   - *Error terms are independent of each other* – We can see there is no specific Pattern observed in the Error Terms with respect to Prediction, hence we can say Error terms are independent of each other
   - *Linear relationship between independent and dependent variables* – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.
   - *Error terms are normally distributed*: Histogram and distribution plot helps to understand the normal distribution of error terms along with the mean of 0. The figure below clearly depicts the same.
   - *Error terms have constant variance (homoscedasticity):* We can see Error Terms have approximately a Constant Variance, hence it follows the Assumption of Homoscedasticity.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Ans: The top 3 features contributing significantly towards the demand of the shared bikes are the
   - temperature,
   - the year and
   - the season variables (Aug and September)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Ans: Linear Regression is an ML algorithm used for supervised learning. It helps in predicting a dependent variable(target) based on the given independent variable(s). The regression technique tends to establish a linear relationship between a dependent variable and the other given independent variables. There are two types of linear regression- simple linear regression and multiple linear regression. Simple linear regression is used when a single independent variable is used to predict the value of the target variable. Multiple Linear Regression is when multiple independent variables are used to predict the numerical value of the target variable. A linear line showing the relationship between the dependent and independent variables is called a regression line. A positive linear relationship is when the dependent variable on the Y-axis along with the independent variable in the X-axis. However, if dependent variables value decreases with increase in independent variable value increase in X-axis, it is a negative linear relationship.

   - Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
   - The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
   - There are 2 types of linear regression algorithms
     - Simple Linear Regression – Single independent variable is used.
       - $Y = \beta 0 + \beta 1X$ is the line equation used for SLR.
     - Multiple Linear Regression – Multiple independent variables are used.
       - $Y = \beta 0 + \beta 1X1 + \cdots + \beta pXp + \in$ is the line equation for MLR.
     - $\beta 0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
     - $\beta 1, \beta 2, \dots , \beta p = Slope\ or\ the\ gradient.$
   - Cost functions – The cost functions help to identify the best possible values for the $\beta 0$, $\beta 1$, $\beta 2$, …, βp which helps to predict the probability of the target variable. The minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches
     – Unconstrained and constrained.
       - Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
         - The straight-line equation is $Y = \beta 0 + \beta 1X$
         - The prediction line equation would be $Ypred = \beta 0 + \beta 1xi$ and the actual Y is as Yi.
         - $Now\ the\ cost\ function\ will\ be\ J(\beta 1, \beta 0\ ) = \sum(yi - \beta 1xi - \beta 0\ )$^2
       - The unconstrained minimization is solved using 2 methods
         - Closed form
         - Gradient descent
   - While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used.
   - $ei = yi - ypred$ is provides the error for each of the data point.

- OLS is used to minimize the total e2 which is called as Residual sum of squares.
- RSS = = $\sum (yi - ypred)$ 2 n i=1

- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. **Explain the Anscombe's quartet in detail.**
Ans: Anscombe's quartet consists of four data sets that have nearly identical simple descriptive statistics but have very different distributions and appear very different when presented graphically. Each dataset consists of eleven points. The primary purpose of Anscombe's quartet is to illustrate the importance of looking at a set of data graphically before beginning the analysis process as the statistics merely does not give an accurate representation of two datasets being compared.

3. **What is Pearson's R?**
Ans: Pearson's Correlation Coefficient is used to establish a linear relationship between two quantities. It gives an indication of the measure of strength between two variables and the value of the coefficient can be between -1 and +1.

The interpretation of the coefficients is:

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

$r = n(\Sigma x * y) - (\Sigma x) * (\Sigma y) / \sqrt{}[n\Sigma x\ 2 - (\Sigma x)\ 2] * [n\Sigma y\ 2 - (\Sigma y)\ 2]$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
Ans: Scaling is a technique performed in pre-processing during building a machine learning model to standardize the independent feature variables in the dataset in a fixed range. The dataset could have several features which are highly ranging between high magnitudes and units. If there is no scaling performed on this data, it leads to incorrect modelling as there will be some mismatch in the units of all the features involved in the model. The difference between normalization and standardization is that while normalization brings all the data points in a range between 0 and 1, standardization replaces the values with their Z scores.

Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

- $MinMaxScaling: x = x - \min(x) / \max(x) - \min(x)$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

- $Standardization: x = x - mean(x) \,/\, sd(x)$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   Ans: The value of VIF is infinite when there is a perfect correlation between the two independent variables. The R�squared value is 1 in this case. This leads to VIF infinity as VIF equals to 1/(1-R2). This concept suggests that is there is a problem of multi-collinearity and one of these variables need to be dropped in order to define a working model for regression.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   Ans: The quantile-quantile (Q-Q) plot are used to plot quantiles of a sample distribution with a theoretical distribution to determine if any dataset concerned follows any distribution such as normal, uniform or exponential distribution. It helps us determine if two datasets follow the same kind of distribution. It also helps to find out if the errors in dataset are normal in nature or not.

   • Interpretations
      - Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis. O
      - Y values < X values: If y-values quantiles are lower than x-values quantiles.
      - X values < Y values: If x-values quantiles are lower than y-values quantiles.
      - Different distributions – If all the data points are lying away from the straight line.
   - Advantages
      - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
      - The plot has a provision to mention the sample size as well.