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?
   Anwer:
   a. There were 5 categorical variables: months, Weekday, Year, Season and Weather situation.
   b. Counts of the number of bike users was the dependent variable.
   c. In case of weekday, the distribution was nearly uniform.
   d. In case of year, there was a significant increase in the count from year 2018 to 2019.
   e. There was seasonal pattern in case of months and seasons. For months, the demand was low at the beginning of the year i.e., in the month of Jan and would increase and peak by the month of June and July. After September the demand would again fall in November and December. This pattern was visible for both the years 2018 and 2019, hence proving seasonality. Similar pattern was visible in case of seasons as well. The demand was low during spring time, would increase during summer and fall and the demand would decrease in winter.
   f. In case of weather situation, though the data dictionary mentioned 4 different kinds of weather situations, the dataset had only 3. There were fewer data points for category 3 i.e., Light rain or snow. The demand was higher for Partly cloudy as compared to Mist weather condition.

2. Why is it important to use drop_first=True during dummy variable creation?
   Answer:
   a. drop_first=True drops the first column during dummy variable creation. Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".
   b. 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?
   Answer:
   a. Based on the pair-plot of the numerical variables, temp(temperature) and atemp(actual temperature has highest correlation with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   Answer:
   We validate the assumptions in the following ways:
   a. Linear relationship between dependent and independent variables.
      We check this by plotting a regression line for each variables to check this.
   b. Error terms are normally distributed.
      We predict the dependent variable using the trained model on the training set. We calculate

the residuals by subtracting the actual value from the predicted value.
We plot a histogram for the residuals and this histogram should be a normal distribution.

c. Homoscedacity: Error terms should have constant variance.
We validate this by plotting a scatter plot of the Error terms. The error terms should have constant variance.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer:
The top 3 features of the final model are:
    a. Whether it is a spring season (negative correlation)
    b. Whether it's the year 2019 (positive correlation)
    c. The weather situation is Light Rain or Light Snow. (Negative correlation)

General Subjective Questions

1. Explain the linear regression algorithm in detail.
Answer:
    a. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
    b. It estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).
    c. For example: The equation of a line with 1 variable is: y = mX+c
    Here y is the dependent variable, X is the independent variable. Given these variables we try to find the slope or coefficient(m) and intercept of constant (c).
    d. By finding m and c, we try to find the generalize the data and use the values to predict the dependent variable based on independent variables.

2. Explain the Anscombe's quartet in detail.
Answer:
    1. 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 fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

2. The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?
   Answer:
   1. The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset.
   2. It describes the strength and direction of the linear relationship between two quantitative variables.
   3. The range of the Pearson's correlation is from 0 to 1.
   4. 0 means there does not exist any linear relationship between the quantitative values being compared.
   5. As the value decreases from 0 to -1, it means the strength of the negative correlation is increasing and -1 being maximum or strong negative correlation.
   6. Similarly as the value increases from 0 to 1, it means the strength of the positive correlation is increasing and 1 being the maximum or strong positive correlation.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer:
   1. Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range.
   2. It also helps in speeding up the calculations in an algorithm.
   3. Most of the times, collected data set contains features highly varying in magnitudes, units and range.
   4. Scaling brings all the values having different range to a common scale making it easier to compare. If the variables are not scaled, the interpretation of the resultant model will be incorrect making the model unusable for prediction.
   5. 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.
   6. 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.
      a. MinMax Scaling: x=x-min(x)/max(x)-min(x)
   7. 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).
      a. Standardization: x= x-mean(x)/sd(x)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer:
   1. 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.
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.
   Answer:
   1. 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.
   2. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.
   3. 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