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?

    i) The demand for the bike is higher in the year of 2019 then 2018. It seems like the demand is increasing w.r.to year.
    ii) The demand is higher in the Fall season and in the middle of the Months.
    iii) Clear weather condition is also playing a vital role in increasing the demand.
    iv) Non-Registered users are higher in number on the weekend and it is quite opposite on the weekdays.

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

    i) We should drop the first column, because it is completely defined by other variables, if all other variables are zero then it quite obvious it is first one so it does not add any unique information to our model.
    ii) Dummy variable trap: It is impossible for our model to tease out whether the effect came from the first variable equalling 1 or from the other variables equalling 0.

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

    i) Temperature variable has the highest correlation with the target variable count and Correlation of the temperature variable with count variable is 0.64 and it is positively correlated.

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

    i) Plotted a distplot with the residuals (Actual Y value – Predicted Y value) and verified whether the error terms are uniformly distributed and the mean is at zero.
    ii) Plotted scatter plot with residuals and verified if the variance of error term is constant and it is not following any pattern.

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

    i) Temperature, Year and Winter (Count = 0.261(Const) + 0.238(Year) + 0.400(Temp) - 0.153(windspeed) - 0.082(Spring) + 0.037(Summer) + 0.104(Winter) - 0.057(Dec) - 0.050(Jan) + 0.034(Jun) - 0.069(Nov) + 0.059(Sep) - 0.029(Mon) - 0.041(Sun) - 0.272(Light Rain) - 0.078(Mist)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

   Linear regression algorithm is used to predict the value of dependent variable with the independent variables. In simple linear regression the algorithm fits a straight line where in case of multiple linear regression it fits a surface that minimize the difference between the predicted and actual output values. To find the best fit line or surface gradient descent algorithm is used to reduce the cost function (calculation of error between predicted value and actual value)

2. Explain the Anscombe's quartet in detail.

   Different datasets with exactly same summary statistics such as mean, variance, corelation coefficient and line of best fit does not mean that they look exactly similar when visualised. The effect of outliers and curvature might drastically throw off our summary statistics and this is called Anscombe's quartet and demonstrates how important it is to always plot your data rather than relying on summary statistics alone.

3. What is Pearson's R?

   Pearson's R measures the strength of the linear relationship between two variables and it is always between -1 to 1. R would be 1 if there is a perfect positive linear relationship (i.e., Y increases exactly with increase in X) and R would be -1 if there is a perfect negative linear relationship (i.e., Y decreases exactly with increase in X).

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

   Scaling is process that bring all the variables (dependant and independent) to the same scale. We perform scaling for ease interpretation and faster convergence for gradient descent method. Standardized scaling basically brings all the data into a standard normal distribution with mean zero and standard deviation one whereas normalized scaling or minmax scaling brings all the data in the range of 0 and 1.

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

   VIF shows the correlation between the independent variables and the VIF Formula is 1/(1-R2) where R2 is a correlation between the variable. If the variable is perfectly positively correlated then the R2 value would be 1 and the VIF formula would become infinite as the denominator will be zero. To solve this problem we need to drop one of the variables from the dataset.

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

   Q-Q plot (Quantile-Quantile) plot helps us to identify how the data in the dataset are distributed. A Q-Q plot has two axes like a scatter plot in which we plot distribution quantile (i.e., uniform, normal etc) vs data quantile. If the Q-Q plot shows linear relationship then the distribution with which we plotted the Q-Q is same as that of the data is distributed else it is not.