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. Following are the inferences from the analysis of the categorical variables:

- Season 3 has shown the highest demand of bike rentals.
- Bike rentals have increased in the year 2019 compared to 2018.
- Demand has grown month over month until June. September has been the highest and following which demand has decreased.
- Demand is less on holidays and more on non-holidays.
- Weekday is not providing any clear picture of any trends.
- Demand is more on a 'Clear' weather.

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

Ans. It is important to use drop_first = True to reduce the extra columns created during dummy variable creation. It helps in reducing correlation between dummy variables.

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

Ans. The variables 'temp','atemp' and 'registered' have shown the most correlation with target variable 'cnt' when compared with others.

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

Ans. The assumptions of Linear Regression was validated by plotting a distplot of the residuals and analysing to see if it is a normal distribution and has a mean = 0.

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 are:

- Temperature (temp) - A coefficient value of '4063.7596' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
- Weather Situation Bad (weathersit_bad) - A coefficient value of '-2341.6591' indicated that, a unit increase in Weathersit_bad variable decreases the bike hire numbers.
- Year (yr) - A coefficient value of '1995.3913' indicated that a unit increase in yr variable increases the bike hire numbers.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans. Linear regression predicts the relationship between a dependent (target variable) and independent variable (predictors) by assuming a linear connection between them. It finds how the value of the dependent variable is changing according to the value of the independent variable. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

The linear regression model gives a sloped straight line describing the relationship within the variables. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

2. Explain the Anscombe's quartet in detail.

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 fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.

3. What is Pearson's R?

Ans. In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

Ans. Scaling is the process of normalizing the range of features in a dataset. Real-world datasets often contain features that are varying in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale, we need to perform feature scaling.

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.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

Ans. The greater the VIF, the higher the degree of multicollinearity. In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

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

Ans. The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Importance of QQ Plot in Linear Regression: In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.