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?

A)

- The demand of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018
- April to October having bike demand so high, Very low demand for months January, February
- There is almost same bookings on Working days/holidays or any week days
- Demand is very high, and climate is mist, clear and almost no demand when climate is light snow

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

A) drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. So, 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?

A) Feature temp has highest positive correlation with target variable cnt.

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

A) Performing Residual analysis is one of the important step to execute to validate assumptions of Linear regression. Conclusion should be error terms are normally distributed.

   The training and testing accuracy is nearly equal

   The Predicted values have linear relationship with 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) Based on the final model, top 3 features contributing are

- temp, 0.4089
- weathersit, -0.2936(Light Snow: -0.2353,Mist: -0.0583)
- Year, 0.2332

General subjective Questions

1) Explain the linear regression algorithm in detail

A) Linear regression is one of the algorithm of machine learning where model is trained to predict the data based on independent variables. Here the variable on x-axis and y-axis should be correlated linearly

- Formula for linear regression is y=mx+c
  - m=slope of the line
  - c = intercept of the line on y-axis
  - x = independent variable
  - y=dependent variable

2) Explain the Anscombe's quartet in detail

A) Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different on graph. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3) What is Pearson's R?

A) Pearson's R is the bivariate correlation, is 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. Thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

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

Formula:

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

Here,

- r =correlation coefficient
- x_i =values of the x-variable in a sample
- $\bar{x}$ =mean of the values of the x-variable
- y_i =values of the y-variable in a sample
- $\bar{y}$ =mean of the values of the y-variable

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

A) Scaling is a pre-processing step which is applied on independent features to normalize the data within some range. Scaling also helps to speed up the calculations as the magnitude of the number is not very high

Most of the times, data set contains features highly varying in units and range. If scaling is not executed, then algorithm only considers magnitude and final output will be incorrect modelling. So, to avoid this inconsistency scaling should be performed.

Normalized scaling rescales the values into a range of (0,1)

Ex: MinMax Scaling, ***sklearn.preprocessing.MinMaxScaler***

Standardized scaling rescales the data to have mean of 0 and a standard deviation of 1.

Ex: Standardisation, ***sklearn.preprocessing.scale***

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

A) VIF becomes infinite when independent variables is perfectly correlated. In the case of high correlation value of R2 becomes 1 and VIF becomes infinity. To overcome this issue, drop one of the correlated features.

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

A) If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.