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:

1. The demand of bike is less in the month of spring when compared with other seasons
2. In months Jul and Sep bike demand is high
3. Bike demand is less in holidays in comparison to not being holiday
4. There is no significant change in bike demand with working day and non working day
5. The bike demand is low when weather is bad

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

Ans:

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?

Ans:

Temp 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:

After verifying linearity, homoscedasticity, absence of multicollinearity, independence and normality of errors.

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

Ans:

Temp, Weathersit, Year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

An example is running a sales promotion and expecting a certain number of count of customers to be increased. Now what we can do is we can look the previous promotions and plot if over on the chart and then try to see whether there is an increment into the number of customers to the rate of promotions. With the help of the previous historical data try we can figure it out or try to estimate what will be the count or what will be the estimated count of current promotion. This will give an idea to do the planning in a much better way about how many numbers of stalls maybe needed or how many increase number of employees needed to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

Linear regression is used to predict a quantitative response Y from the predictor variable X. Mathematically, we can write a linear regression equation as:

$$Y = a+bx$$

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

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. It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties. These four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

3. What is Pearson's R?

Ans:

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ. Given a pair of random variables the formula for ρ

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

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 step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. 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.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.

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).

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

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

Ans:

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 .

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

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.

c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.

d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis