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:-

The following categorical variables played an significant role in Bike Rental:

1]Month 2]Holiday 3] Season 4]Weather Situation

Following are the effects on the dependant variable cnt observed from the boxplot:

1] Month- The month of September showed the highest no of bike rentals. December showed the lowest no of bike rentals.

2] Holiday-There is significant reduction in bike rentals during Holidays.

3] Season-There is increase in cnt value in summer season.

4] Weather Situation- December showed the lowest no of bike rentals due to the snowy weather.

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

Ans:-

During Dummy Variable Creation, some extra columns are created. In order to reduce those extra columns, it is significant to use drop_first=True.

For Example: If there is a situation giving you success or failure. Dummy variable is created for the situation giving you either success or failure. These dummy variable requires very less memory & can be beneficial in proper modeling function.

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

Ans:-

As observed from the pairplot, temp shows 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:-

i] Residuals are nothing but error terms. To draw proper insights the residuals must follow normal distribution. This assumption is validated by using distplot of residuals in order to check whether the residuals are normally distributed or not.

ii] The linearity between the dependant variables & their predictors is validated using the pairplot for the dataframes.

iii] Multicollinearity absence: This is validated by finding out the correlation between the variables using heatmaps.

iv] Independent errors terms is validated by using DW statistics.

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

Ans: Variable Variable Coefficient

i] temp=0.472772

ii] weathersit_Light Snow & Rain= -0.286377

iii] yr = 0.234197

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Ans:-

i] Linear Regression algorithm is the ML algorithm which is based on supervised learning. In linear regression algorithm, the machine is being trained to predict the data using variables.

ii] Linear Regression establishes a relation between predictor variables & the outcome variables.

iii] Linear Regression uses a slope intercept form:

$$y=mx+c$$

where y= independent variable

x= independent variable

iv] Linear Regression gives us the best value of a0 and a1 which is used to find the best fit line. This best fit line must have least error.

v]In short it predicts the values in a specific continuous range.

Types Of Linear Regression:

i]Simple Linear Regression: If the value of Single Numerical dependant variable is predicted by using single independent variable, then the algorithm is termed as the simple linear Regression.

ii]Multiple Linear Regression: If the value of Numerical dependant variable is predicted by using more than one independent variable, then the algorithm is termed as the multiple linear Regression.

2. Explain the Anscombe's quartet in detail.

Ans:

i]Anscombe's quartet contains 4 datasets that have similar simple statistical properties. But when we plot a graph of these 4 datasets, we can notice the difference in between them.

ii]Eleven datapoints are consisted in each of the dataset.

iii]In the year 1973, Anscombe's quartet was discovered by Francis Anscombe.

iv]This quartet gives the importance of plotting the data graphically before starting any analysis of data.

v]In short Anscombe's quartet focuses on visualizing the data & in turn gives the confirmation for the model fit validity.

3. What is Pearson's R?

Ans:

i] Pearson's R is also known as the Pearson's Coefficient.

ii] It is denoted by r

iii] Strength of relationship of two variables is measure by using Pearson's R.

iv]The numerical coefficient value lies in between +1 & -1.

v] a] If r=1, data is linear & has a +ve slope.

b] If r=-1, data is linear & has a-ve slope.

c] If r=0, data is not linear.

v] The formula for Pearson's R is given by:

$$r= (N \Sigma XY -(\Sigma X)(\Sigma Y)) /(\sqrt{} [N(\Sigma X)^2-(\Sigma X)^2] [N(\Sigma Y)^2-(\Sigma Y)^2])$$

where,

N=no of pair

ΣXY= Sum of products of pairs

(ΣX)=Sum of X

(ΣY)=sum of Y

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

Ans:

Scaling: Scaling is a process of transforming raw values into some standard values within specified range. It is also termed as data normalization.

Reason behind scaling: i] A lot of times, many datasets contains features which highly vary in magnitude, units, etc. Since these features may have different magnitudes & units, it becomes difficult to compute the Euclidean distance between 2 data points.

ii] Hence to make sure all these features are under same level & range, we need to perform scaling.

Difference between normalized scaling and standardized scaling:

i]Normalized Scaling is used to scale the feature under same level, while standardized scaling is done by transforming the feature by subtracting from mean & dividing it by std Deviation.

ii]]Normalized Scaling is used in the case of no outliers, while standardized scaling is used in the case of data following the Guassian Distribution.

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

Ans: VIF stands for variance inflation factor. If VIF is inifinity, it indicates the variables are perfectly correlated. A linear combination of other variables can be established for the corresponding variable if the value of VIF is infinity.

Formula for VIF is given by:

=1/(1-R2)

In this case R2=1, substituting these value in the formula we get VIF=Infinity.

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

Ans: Q-Q plot stands for Quantile- Quantile plot. It is the representation of quantiles in the form of plots. It plots the first quantile of dataset against the second quantile of dataset.

Use & Importance of Q-Q plot:

i]The main purpose of using the Q-Q plot is to determine whether the two datasets come from the same distribution or not.

ii]The Q-Q plots helps in visualizing the comparison sample quantiles & the theoretical quantiles.

iii] Q-Q plot can be used if the datasets have common location & scale.

iv]If the two datasets have identical tail behavior, Q-Q plot can be used to visualize the dataset in this case.