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)

In 2018 less bikes were rented and in 2019 more the main reson may be because of popularity gain

Most bikes were rented on working days instead of holidays

Median of fall is higher than any other season and in spring the median is least i.e in season fall most bikes were rented and in spring season less bikes were rented

People rented more bike when the weather condition is very good

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

Ans)

To deal with categorical features we use dummy variables represented by 0 and 1.A feature with n categorical levels can be represented by n-1 dummy variables for example if a feature has 5 different categories then we create 4 dummy variables because if none of the 4 dummy variables are 1 then we can say that 5th categorie is true so we only have 4 dummy variables.To drop one dummy variable we use drop_first=true while creating dummy variables.

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

Ans)

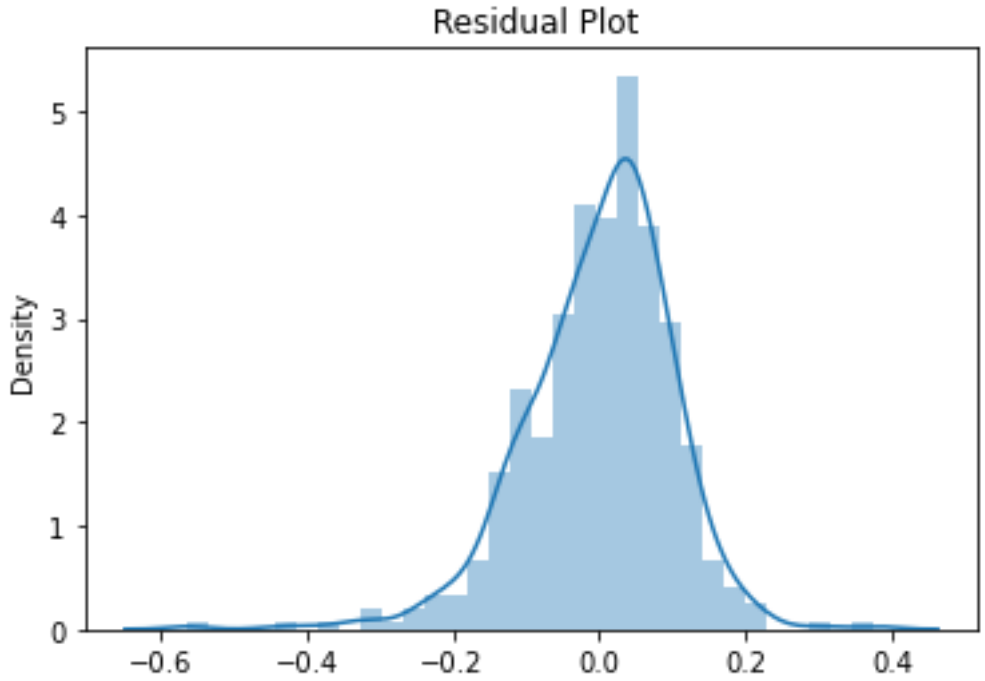
By looking at the pair-plot, variable temp and atemp which highly correlated among themself have the highest correlation with the taget variable cnt the correlation between them is 0.63. and second highest correlated vriables is yr with 0.57 correlation with cnt.

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

Ans)

One of the assumptions of Linear Regression is error terms should be normally distributed ie residuals should be distributed normally

To validate the assumptions of linear regression i ploted a distplot on residuals(y_train-y_train_pred)



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

Ans)

Top 3 features contibuting significantly towards explaining the demand of the shared bikes are:

a.yr with 0.2495 coef on cnt

b.weathersit(2) with 0.2259 coef on cnt

c.weathersit(1) with 0.3113 coef on cnt

General Subjective Questions

1. Explain the linear regression algorithm in detail?

Ans)

Linear regression is a type of supervised learning

Linear regression attempts to model the relationship between two variables by fitting a linear equation (a straight line) to the observed data. target variable is considered to be an dependent variable, and other variables are considered to be a independent variables.

There are two types of linear regression:

<u>Simple linear regression</u>**:**

In this we will only have one independent variable

Regression line:

y=b0+b1x

Multiple linear regression**:**

In this we will have more than one independent variable

Regression line**:**

y=b0 + b1x1 + b2x2 +.....+ bnxn

Having multiple indepented variables in linear regression will definetly increase r-squared value

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 , but there are some abnormalities 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 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.There are 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 r is also called as Pearson correlation coefficient. It is a measure of linear correlation between two set of variables. It is the covariance of two variables, divided by the product of their standard deviations. Covariance is the measure of relationship between two variables. Covariance doesn't give strength of relation ship it can only say if the relation is positive or negative but Person correlation give strength of relationship between variables

P(x,y)=cov(x,y)/(standard deviation of X) *(standard deviation of Y)

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.

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 becomes infinity. If VIF is infinity it means that there is a perfect correlation between two independent variables. In this case of perfect correlation, we get R^2 =1, which lead to 1/(1-R^2) to be 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(Quantile-Quantile plot) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For

example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.