Assignment-based Subjective Quest

1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer :

- There is significant rise for bike in the year 2019
- When the season is ‘fall’ there is huge demand for bikes.
- Demand for the bikes on all weekdays remains same.
- Not much significance for bike demand when it’s working day or non working day.
- Demand for bike is high when weather is ‘cloudy’ or ‘partly cloudy’
- There is a significant rise for bike demand after March month till October.

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

Answer :

- It is important to achieve n-1 dummy variables as it can be used to delete extra column while creating dummy variables.
- We can avoid multicollinearity by using drop_first = true.

3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Answer : atemp and temp have highest correlation with cnt, which is 0.65 and 0.64

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

Answer :

- Linearity of relationship between response and predictor variables.
- Normal distribution of error terms.
- Constant Variance of errors.
- Low VIF between features

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

Answer :

- Temp
- Summer season
- September month

General Subjective Questions

1.Explain the linear regression algorithm in detail?

Answer :

Linear Regression is supervised machine learning method and is used to find a linear equation between target variable and independent variables.

The following is an example of linear regression equation

y= m1x1 + m2x2 + m3x3 + ………. + m(n) x(n) + c

y is the target variable and x1,x2.. are predictor variables.

m1,m2,m3.. and c are the two unknowns.(m is the intercept and c is constant)

We have to find values of m and c such that we have minimum error and find the best fit line which has the minimum error.

In linear regression, when the error is calculated using the sum of squared error, this type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression.

Error function is explained by 'e = - y', and error depends on the values of 'm' and 'c'. Our aim is to build an algorithm which can minimize the error.

2.Explain the Anscombe's quartet in detail?

Answer :

Anscombe's quartet is a group of four data sets that are nearly identical in descriptive statistics but there is a peculiarity that fools the linear regression once you plot each data set individually.

The four data sets will have similar descriptive statistics like mean,standard deviation but when you plot the these data set all of them will look different.

Anscombe's quartet tells us the importance of visualizing the data sets before using them to build the model. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.)

When you plot the four data sets which have very identical descriptive Statistics, the result might look like below.

Data Set 1: fits the linear regression model pretty well.
Data Set 2: cannot fit the linear regression model because the data is non-linear.
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

3.What is Pearson’s R?

Answer :

Pearson’s R measures the Correlation between two variables whose value lies between -1 and +1.

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

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

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction) Example : If X increases then Y also increases.

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions) Example : If X increases then Y decreases and vice versa.

r = 0 means there is no linear association. X and Y doesn’t have any Correlation.

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

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

Answer :

Scaling is a data preprocessing Step which is performed on the Independent Variables to Normalize the data within a Particular range.

Why Scaling?

In the datasets the all the independent Variables will not have a same range of values, one or the other column will have much higher value compared to other Variables, in this case we need to Scale such variables in order to get comparable scale. If we do not have comparable scale then we might end up getting higher/lower coefficient compared to other Variables, So this might create a bigger problem during our model evaluation.

Example :

X	Y	M	N
0	1	0	8765
1	1	0	7890
1	1	1	5689
0	0	1	3456

In the above table we have four variables, as we can see the Values in each Variables we can see that Values of N are really higher compared to other 3, so we need to Scale these variable inorder to have a comparable scale.

Difference between normalized scaling and standardized scaling?

Normalized Scaling	Standardized Scaling
Uses Minimum and Maximum value to Rescale the variables.	Uses Mean and Standard deviation to rescale the variables
This can be used when the variables are on Various Scales.	This can be used when variables mean and standard deviation are set to 0.
Values on the scale falls between [0,1] and [-1,1]	Values on the scale will not fall on to a particular range
When feature distribution is unclear it is helpful.	When the feature distribution is consistent it is helpful.

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

Answer :

VIF Standas for Variance Inflation Factor,which is a measure of multicollinearity in regression analysis.

We usually delete all the Varibales which have a VIF value higher than 5 through some process.

VIF is calculates using below formula

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

We sometimes get VIF value of infinity, the reason for this is listed below.

- It’s because the two independent variables have a perfect correlation.
- This usually happens when R2 is equal to 1.
- In this case we will have to remove one such variable from dataset.

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

Answer:

Q-Q Plots stands for Quantile-Quantile Plots, it’s a plot for Quantiles of Sample Distribution against Quantiles of theoretical Distribution.

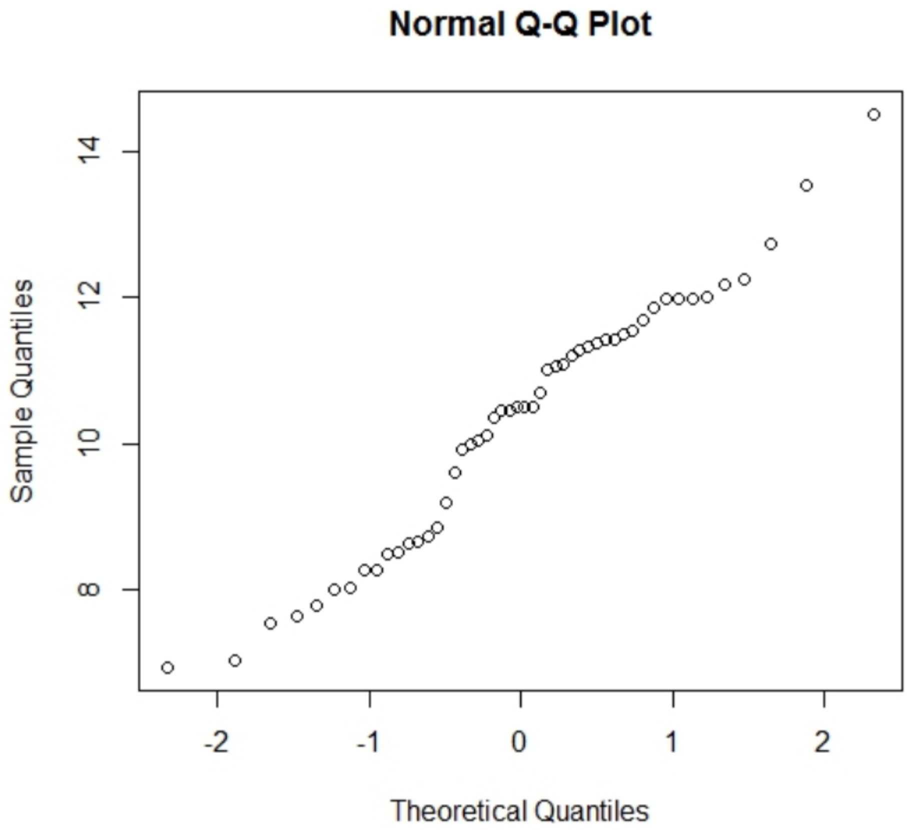
This plot helps us to identify if it follows any particular type of probability distribution like normal, uniform, exponential.

The Q-Q Plot helps to answer the below Questions.

1.Do two datasets come from populations with a common distribution.

2. Do two datasets have common location and scale.

3. Do two datasets have similar distributional shapes.


Normal Q-Q Plot
Sample Quantiles
8
10
12
14
-2
-1
0
1
2
Theoretical Quantiles