Name: Kolla Neeraja

Machine Learning Assignment Subjective Questions

Assignment based Subjective Questions

Q1- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

1. Company should focus on business when hum and windspeed are high by providing safety

2. Based on data it expected to have a boom in number of users once situation comes back to normal.

3 . There would be less bookings during weathersit_Mist_cloudy, they could probably use this time to service the bikes without having business impact.

4 . Hence when the situation comes back to normal, the company should come up with new offers during spring when the weather is pleasant and also advertise a little for September as this is when business would be at its best.

4. When temperature is good there are high no of rental bikes

5 . Hum and windspeed is high count is less

6 . sat and has high sales of bike rental

7 . temp has high correlation and taking bike rental demand is going high are good at that time

7 . spring , mnth _nov and mnth_dec has there are less demand

8. mnth_sep, season winter has bit demand of bikes

. Significant variables to predict the demand for shared bikes

. Working day

. temp

. hum

. windspeed

. Season (Spring, Winter)

. Month (Dec, Sep, Nov)

. Weatherist_mist_cloudy

Q2- Why is it important to use drop_first=True during dummy variable creation?

Answer:

Dummy variables will be created with one hot encoding and each attribute will have a value of either 0 or 1, representing the presence or absence of that attribute.

Trap of Dummy variable :

The dummy variable trap is a scenario where there are attributes that are highly correlated (multi collinear) and one variable predicts the value of others.

When we use one _hot encoding for handling the categorical data, then one dummy variable (attribute) can be predited with the bhelp of other dummy variables.

Hence, one dummy variable is highly correlated with other dummy variables. Using all dummy variables for regression models leads to a dummy variable trap.

So the regression models should be designed to exclude one dummy variable that's why we use .drop inplace = true

Lets have an example:

Let's consider the case of gender having two values male (0 or 1) and female (1 or 0).

Including both the dummy variables can cause redundancy because if a person is not a male in such case that person is a female, hence we don't need to use both the variables in regression models.

This will protect us from the dummy variable trap

Q3- Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

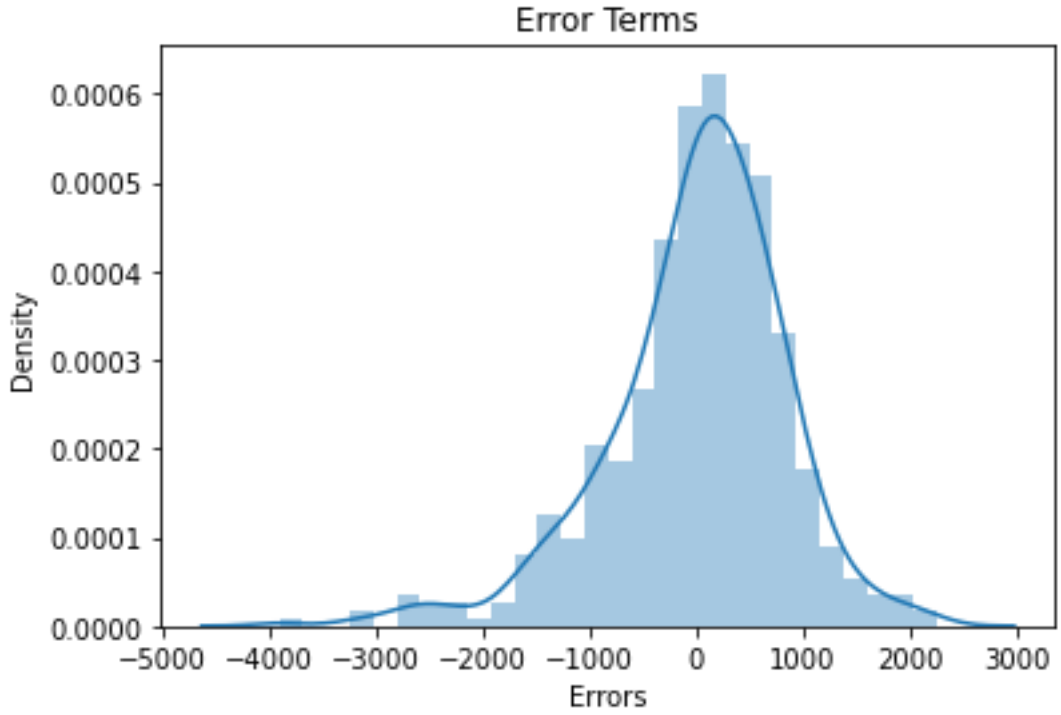
Answer:

. temp and atemp are highly correlated with target variable ("cnt)

Q4- How did you validate the assumptions of Linear Regression after building the model on the training set?

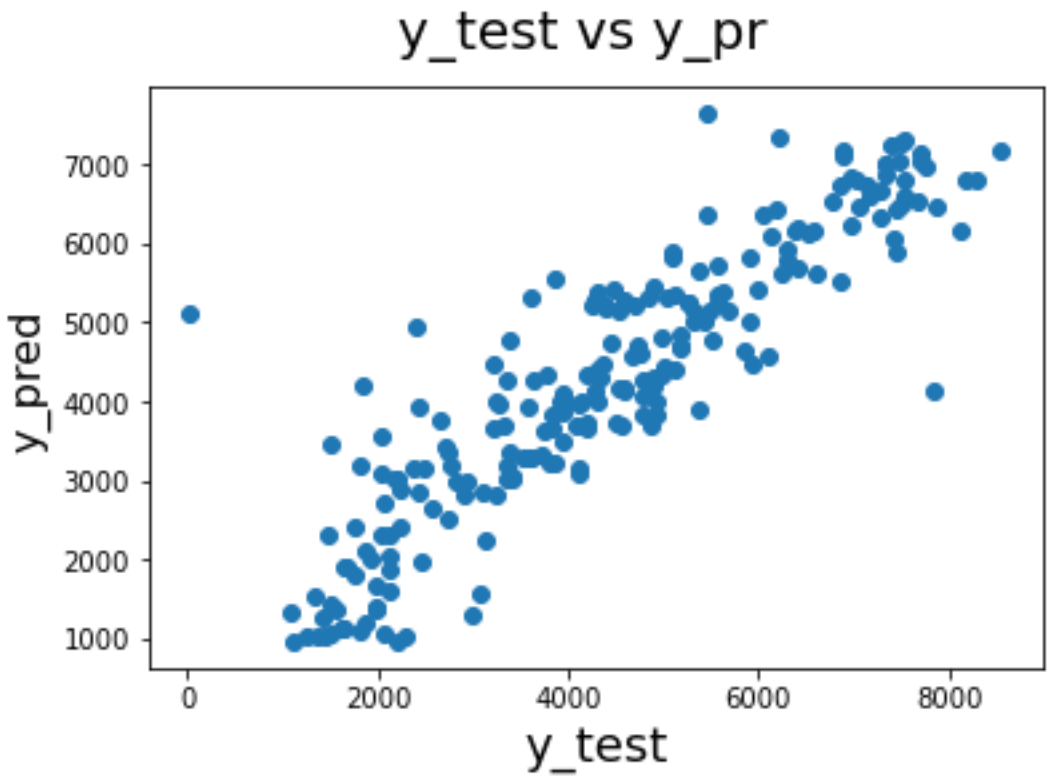
Answer:

By using residual analysis and by checking the error terms are normally distributed with mean Zero using distplot.

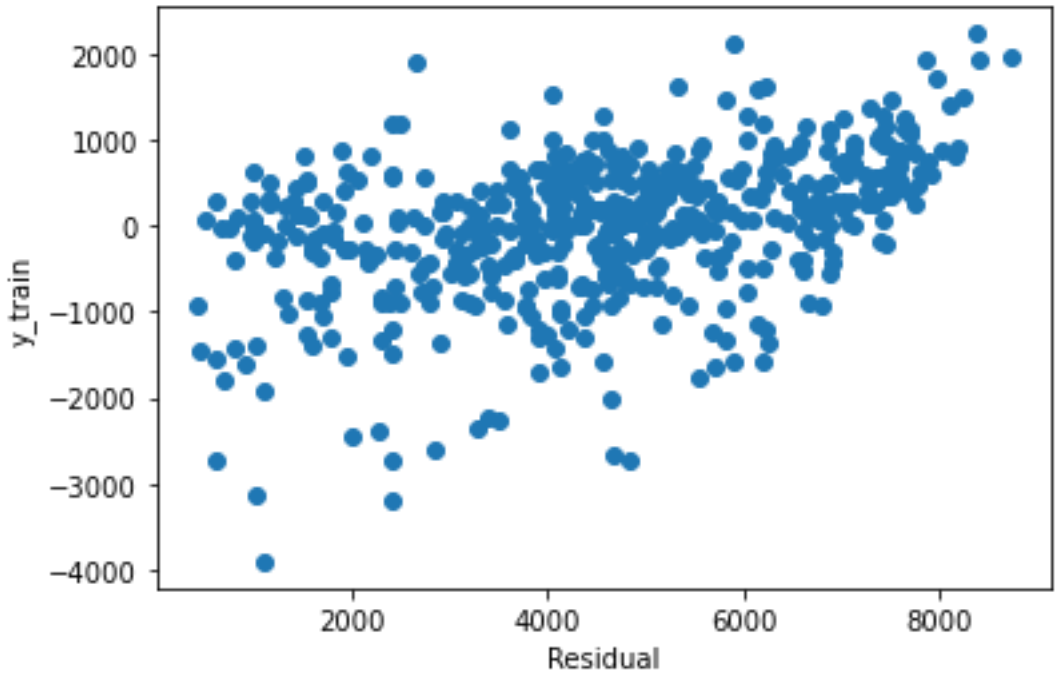


 # Homoscedacity:

Calculated based on error terms for training set and testing set and checked the variance error terms must have constant variance at different data points.



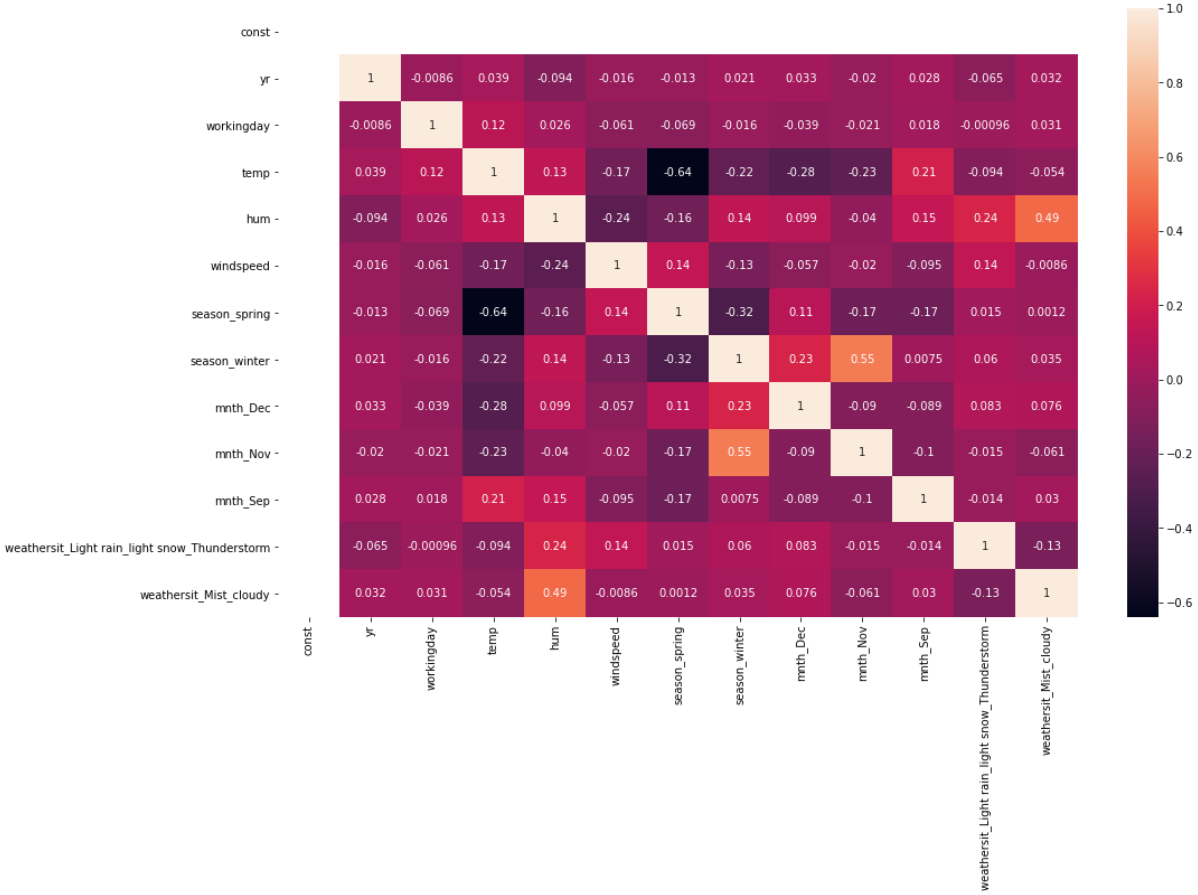
Linear Relationship: Perform the linearity check plotted actual and predicted to test the linearity.



Q5 Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Positive and negative :

Temp, hum, yr are the top 3 features contributing significally towards demand of bikes



General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Answer:

In order to understand linear regression algorithm we need to understand below concepts

Regression

Linear Regression

Cost function

Gradient Descent

Assumptions

Regression:

Regression analysis is nothing but a predictive modelling methodology that aims to investigate the relation that exists between independent variables or predicators and dependent or targets

This is done by fitting a line or curve to different data points in a way that we can minimise the difference in data point distance from the line or the curve

Regression shows a line or curve that passes though all data points on a target predictor graph in such way that the vertical distance between the data points and the regression line is minimum.

Linear Regression:

Linear regression shows the linear relationship between the independent variable x and dependent carriable y .

If we only have one independent variable then it will be a simple linear regression problem and if we have multiple independent variables then it will be a multiple linear regression function.

For EG: we can observe a linear relationship between TV an sales will be like

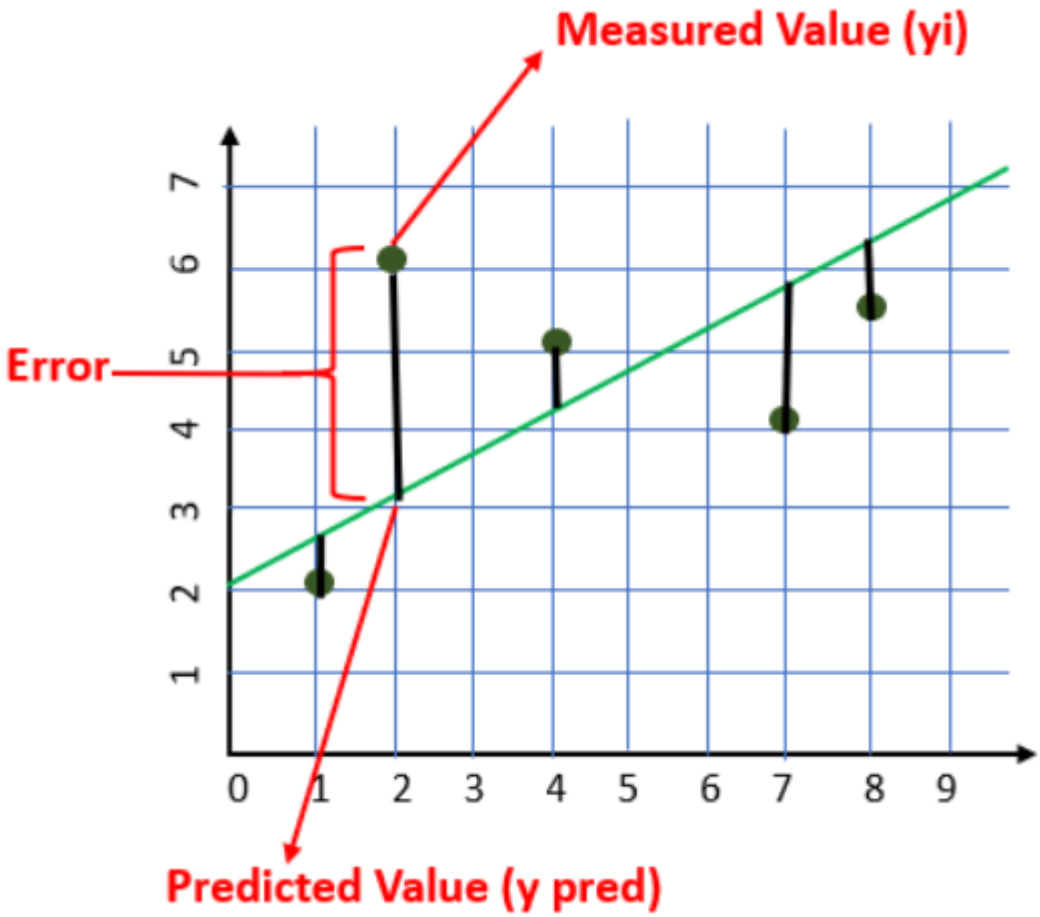
Simple linear regression salesy = Bo + B1 * Tv

Multiple regression: salesy = B0 + b1*Tv+B2*radio

Goal: Our goal is to find the bes fit line minimum the cost function

Cost Function

The cost is the error in our predicted value. We will use the Mean Squared Error function to calculate the cost.



$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - y_{i\,pred})^2$$

Replace $y_{i\,pred}$ with $mx_i + c$

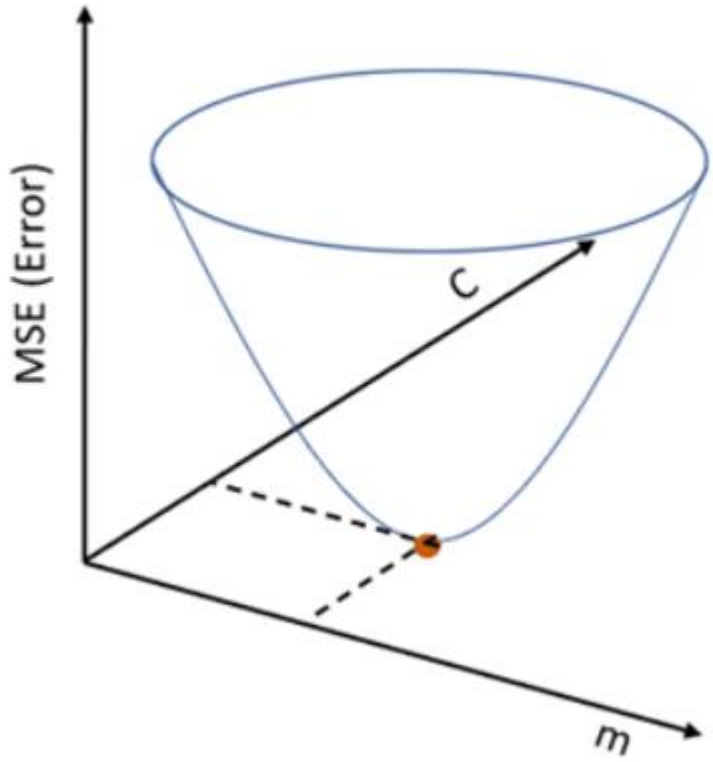
$$Cost\ Function(MSE) = \frac{1}{n}\sum_{i=0}^{n}(y_i - (mx_i + c))^2$$

Our goal is to minimize the cost as much as possible in order to find the best fit line. We are not going to try all the permutation and combination of m and c (inefficient way) to find the best-fit line. For that, we will use Gradient Descent Algorithm.

Gradient Descent Algorithm:

Gradient Descent is an algorithm that finds the best-fit line for a given training dataset in a smaller number of iterations.

If we plot m and c against MSE, it will acquire a bowl shape (As shown in the diagram below)



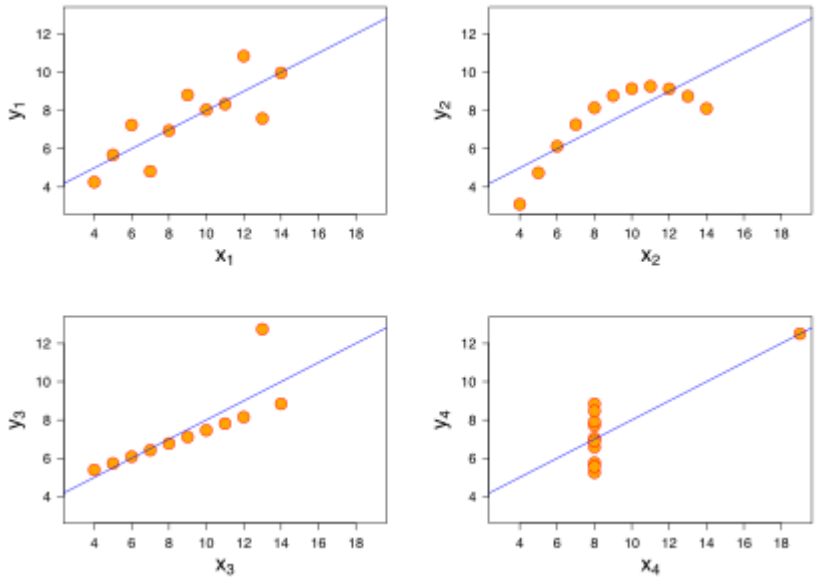
For some combination of m and c, we will get the least Error (MSE). That combination of m and c will give us our best fit line.

The algorithm starts with some value of m and c (usually starts with m=0, c=0). We calculate MSE (cost) at point m=0, c=0. Let say the MSE (cost) at m=0, c=0 is 100. Then we reduce the value of m and c by some amount (Learning Step). We will notice a decrease in MSE (cost). We will continue doing the same until our loss function is a very small value or ideally 0 (which means 0 error or 100% accuracy).

Q2. Explain the Anscombe's quartet in detail.

comprises four datasets that have nearly identical simple statistical properties yet have very different distributions and appear very different when have graphed Each dataset consists of eleven points(X,Y).

They were constructed in 1973 by the statistician francis ansconb'e to demonstrate both the importance of graphing data before analyzing it, and the effect of outliers and other on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough. It has been rendered as an actual quarter



Application:

The quarter is still often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to particular type of relationship and the inadequacy of basic statistic properties for describing realistic datasets.

Q3. What is Pearson's R?

In statistics the Pearson correlation coefficient(pcc) also referred to aspearson's

Definition

Let's focus on some statistical explanation of it. Pearson's Correlation coefficient is represented as 'r', it measures how strong is the linear association between two continuous variables using the formula:

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

Where,

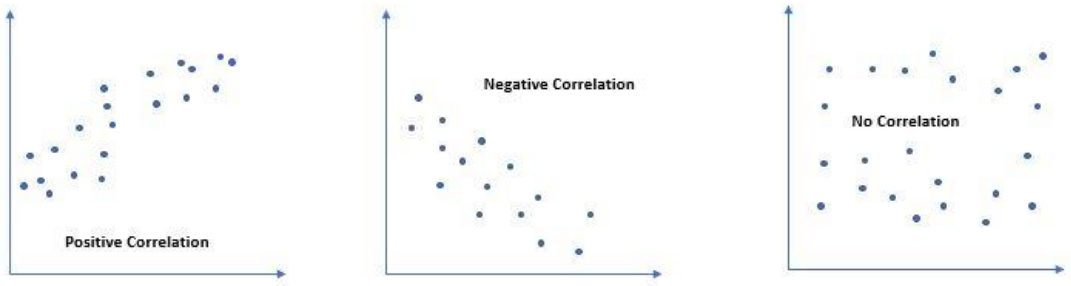
r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

Values of Pearson's Correlation are:

Value of 'r' ranges from '-1' to '+1'. Value '0' specifies that there is no relation between the two variables. A value greater than '0' indicates a positive relationship between two variables where an increase in the value of one variable increases the value of another variable. Value less than '0' indicates a negative relationship between two variables where an increase in the value of one decreases the value of another variable.



Pearson correlation attempts to draw a line of best fit through the spread of two variables. Hence, it specifies how far away all these data points are from the line of best fit. Value of ‘r’ equal to near to +1 or -1 that means all the data points are included on or near to the line of best fit respectively. Value of ‘r’ closer to the ‘0’ data points is around the line of best fit.

Considering the same example of the car price, let’s find out the ‘r’ value using ‘pearsonr’ function in python.

As stated earlier, the value of Pearson correlation for Price vs Curbweight is 0.835 and as there is no correlation between Price and Carheight, hence the Pearson Correlation value between Price & Carheight is near to 0 which is 0.12.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling is a technique of bringing the values of all independent feature of our dataset on the same scale .

Feature selection helps to do calculations in algorithms very quickly . it is important stage of data preprocessing.

If we didn't do feature scaling then the machine learning model gives higher weightage to higher values and lower weightage to lower values.

Also, takes a lot of time for training the machine learning model.

Why scaling:

When we have a lot of independent variables in a mode; a lot of them be on very different scales which will lead a model with very wired co-efficient that might be different to interpret because of 2 reasons

1) Ease of interpretation
2) Faster convergence for gradient descent methods

Normalization:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0

- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

μ

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Answer:

If there is perfect correlation , then VIF = infinity. This shows a perfect correlation between two independent variables.

In this case of perfect correlation, we get R@ = 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 values indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which shows an infinite VIF as well)

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

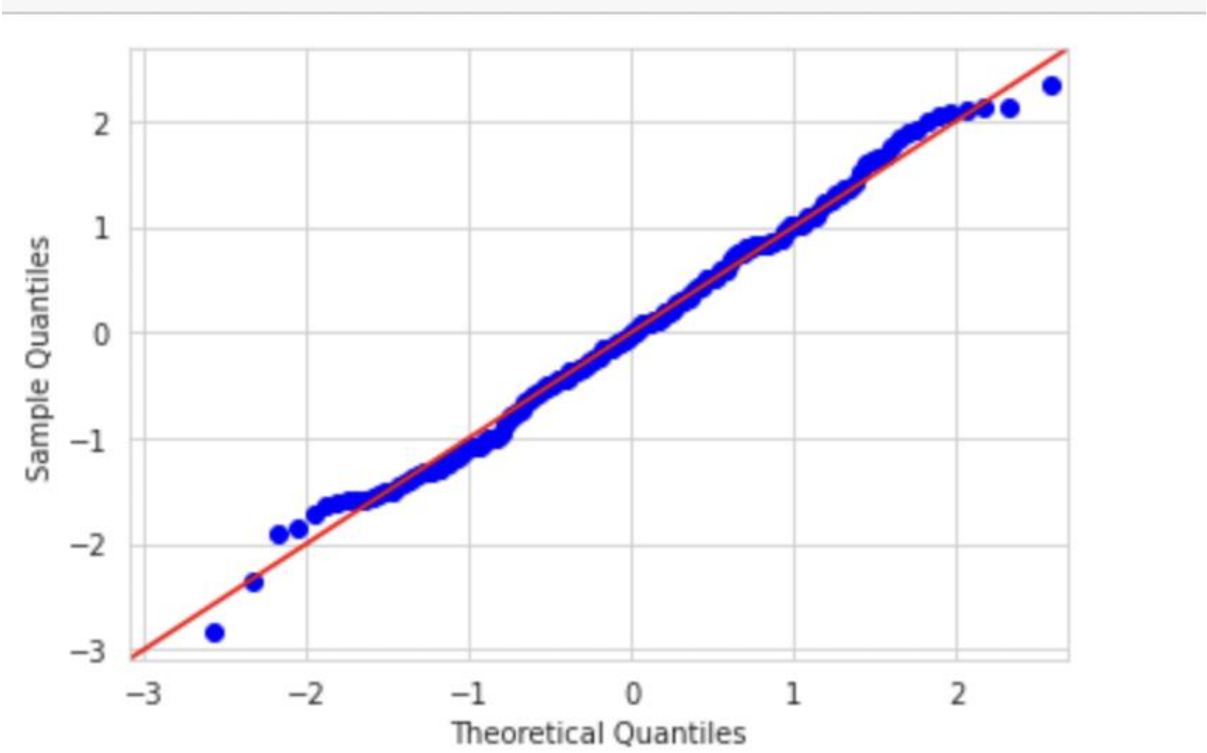
The power of Q-Q plots lies in their ability to summarize any distribution visually.

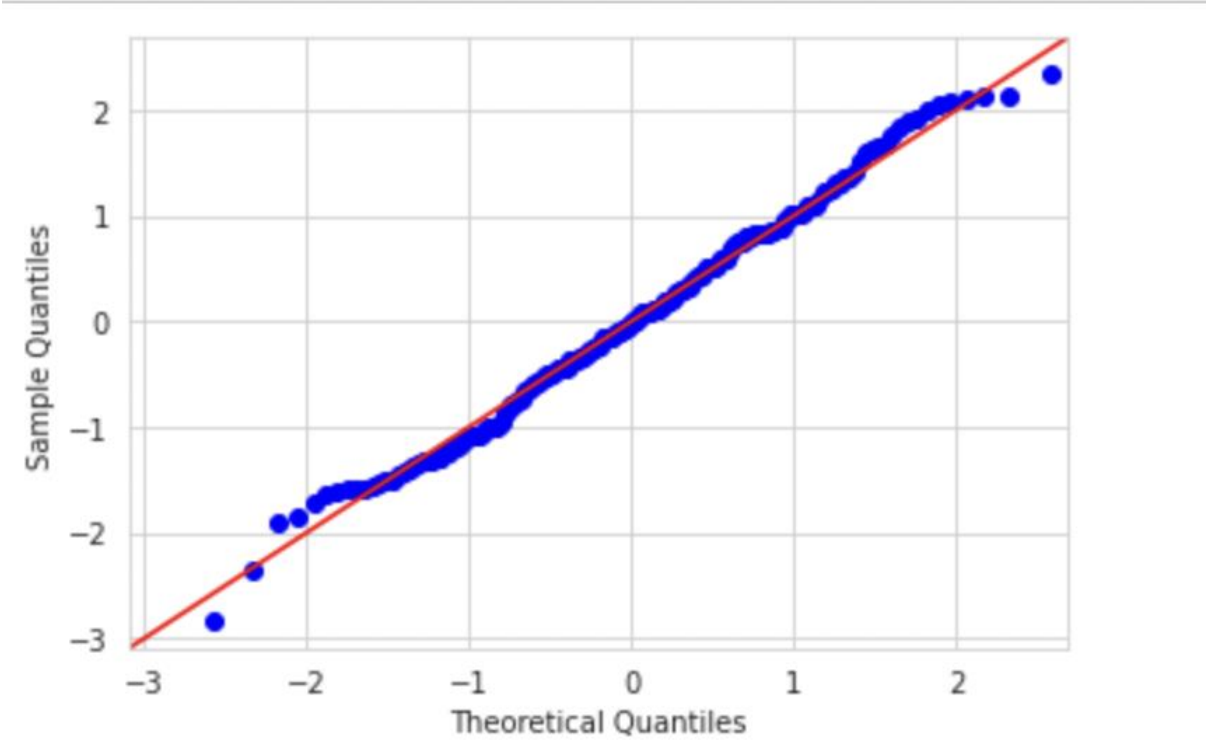
QQ plots is very useful to determine

- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.

If the datasets we are comparing are of the same type of distribution type, we would get a roughly straight line. Here is an example of normal distribution.





As you build your machine learning model, ensure you check the distribution of the error terms or prediction error using a Q-Q plot. If there is

a significant deviation from the mean, you might want to check the distribution of your feature variable and consider transforming them into a normal shape. As you build your machine learning model, ensure you check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, you might want to check the distribution of your feature variable and consider transforming them into a normal shape.
