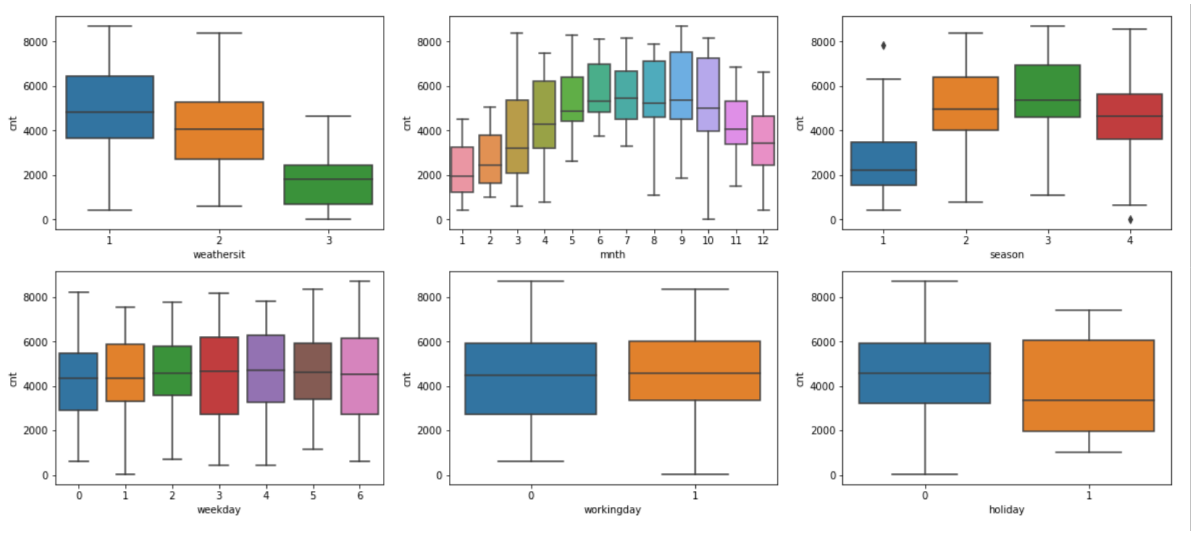
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)



Weathersit =
1: Clear, Few clouds, Partly cloudy, Partly cloudy
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

So in “Clear, Few clouds, Partly cloudy” weather setting – bikes count were between 4000-6000 also maximum range is for this season .starting from approx. 100 -8000
In “Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds” lesser bikes were rented however the lower range is still the same , median is little lower approx. 4000
In “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds” the least no of bikes were rented between 100 to 2000.
in” Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog”no bike seems to have been taken for obvious reasons

Month – demand for bikes increases towards summer months, it seems the data is of probably winter setting since people are renting more bikes in summer to enjoy summer

Season : season (1:spring, 2:summer, 3:fall, 4:winter)
as analysed earlier the demand for bikes increases with summer and remains same or slight increase with fall , but it drops as winter sets in and drops considerably in spring – probably due to high wind speed as seen in excel data sheet.

Weekday – the median lies between 4000-5000 for all days – so cant really make a call how much this would have a dpenedcy on the target variable

Working day – more bikes are taken on the day when its not working ,rather than when its working . However the difference is marginal , cannot make any direct inference.

On holidays – generally people don't go for renting bikes .

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

Because we can save space by encoding the first dummy as 000..upto n levels, also it reduces the co-relation among dummy variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

if we don't drop the registered and casual variables then highest correlation will be with registered , however since we do drop the column , the next higher relationship would be with temp or atemp both are approximately same.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

validation is based on the scatter plot of predicted value and actual values of the test data set. R^2 also gives a fair notion of the linearity of the model

Assumption 1:<u>The Dependent variable and Independent variable must have a linear relationship.</u>
How to Check?
A simple pairplot of the dataframe can help us see if the Independent variables exhibit linear relationship with the Dependent Variable.

Assumption 2:<u>No Autocorrelation in residuals.</u>
How to Check?
Use Durbin-Watson Test.
DW = 2 would be the ideal case here (no autocorrelation)
0 < DW < 2 -> positive autocorrelation
2 < DW < 4 -> negative autocorrelation
statsmodels' linear regression summary gives us the DW value amongst other useful insights.

Assumption 3:<u>No Heteroskedasticity</u>.
How to Check?
Residual vs Fitted values plot can tell if Heteroskedasticity is present or not.
If the plot shows a funnel shape pattern, then we say that Heteroskedasticity is present.
Assumption 4:<u>No Perfect Multicollinearity.</u>
In case of very less variables, one could use heatmap, but that isn't so feasible in case of large number of columns.
Another common way to check would be by calculating VIF (Variance Inflation Factor)

values.
If VIF=1, Very Less Multicollinearity
VIF<5, Moderate Multicollinearity
VIF>5 , Extreme Multicollinearity (This is what we have to avoid)

Assumption 5: Residuals must be normally distributed.
How to Check?
Use Distribution plot on the residuals and see if it is normally distributed.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

temperature year, and season_4

General Subjective Questions

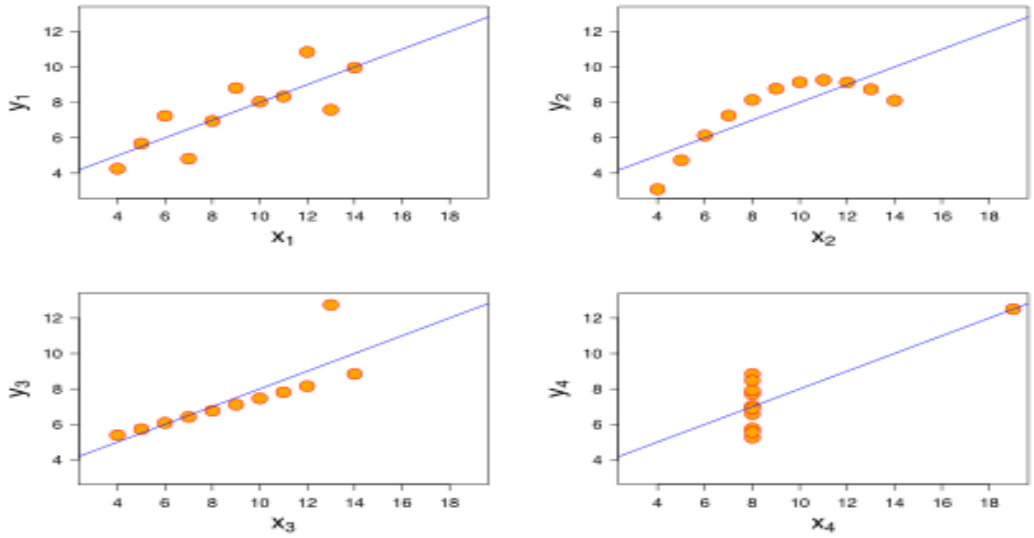
1. Explain the linear regression algorithm in detail. (4 marks)

1. Loading the Data- first we load the data , drop all the redundant features
2. Exploring the Data- categoricals and Numerics for each we perform different kind of approach for – categorical we basically use boxplot, numerical variables are mostly seen through heatmap by seaborne , apart from that we can also take general idea of target variable through simple histogram.
3. Scaling The Data- since most of the numerical data is not in congruence with other data – what we do is scale them to same proportions so that they can be easily compared with other data.
4. Train and Split Data- first we split the data into test set and train set then we start training our train data through selecting features one by one or eliminating one by one if we are using recursive approach, elimination is done through generating an Linear regression summary which gives some statistics such as p-value , F-statistic , AIC , Condition No which if high means there exists multicollinearity between the variables and target variables need to be removed, other method for checking mullticollineaity among the variables themselves is the variance inflation factor that tells us how much one of the input vraibke is dependent upon other input variable
5. Generate The Model – after removing the multicollinearity we may have linear regression model
6. Evaluate The accuracy- accuracy of the model is evaluated by plotting a scatter plot between the predicted target values and actual values of the test set.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset

consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

The datasets are as follows. The *x* values are the same for the first three datasets.

ANSCOMBE'S QUARTET

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y

10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places

Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

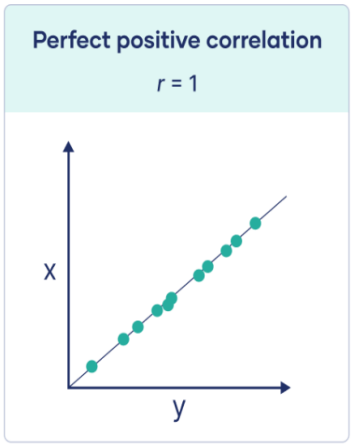
3. What is Pearson's R? (3 marks)

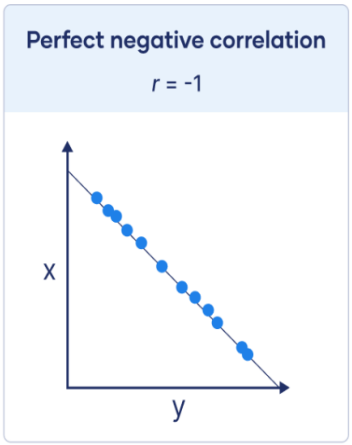
The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

When to use the Pearson correlation coefficient

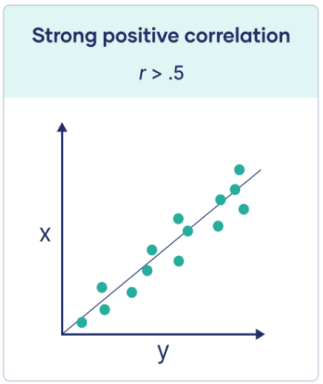
The Pearson correlation coefficient (*r*) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when **all** of the following are true:

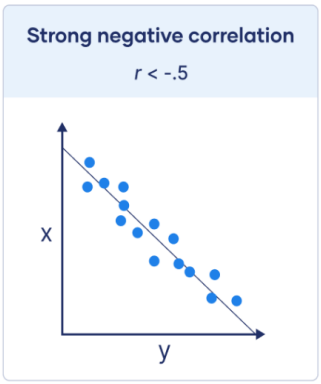
- **Both variables are quantitative:** You will need to use a different method if either of the variables is qualitative.
- **The variables are normally distributed:** You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- **The data have no outliers:** Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- **The relationship is linear:** "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear.

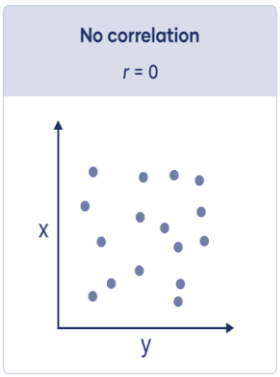




When *r* is greater than .5 or less than –.5, the points are close to the line of best fit:







4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

It 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. **sklearn.preprocessing.MinMaxScaler** helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

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**).

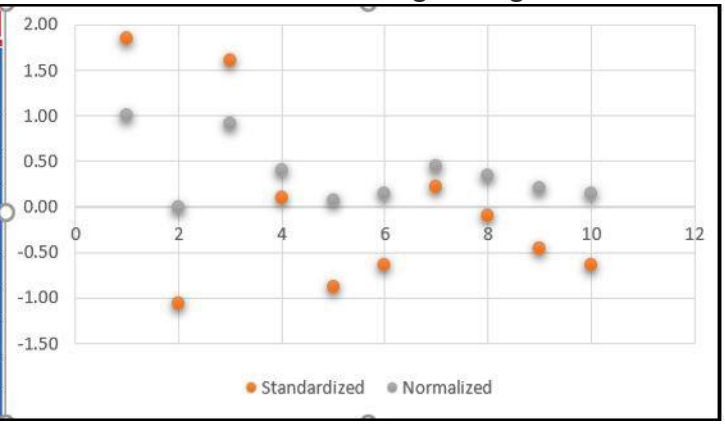
$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

Example:

Below shows example of Standardized and Normalized scaling on original values.

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

(3 marks)

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 (which show an infinite VIF as well).

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

(3 marks)

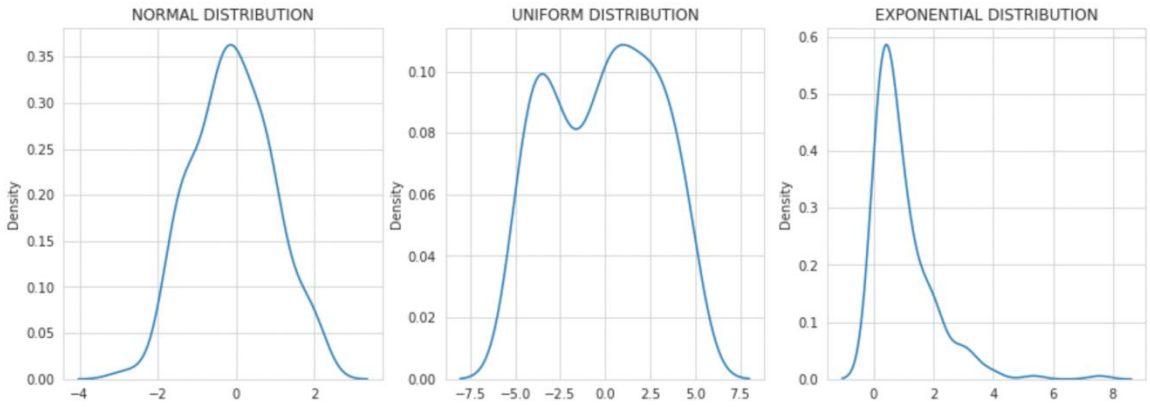
Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential.

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential. Before we dive into the Q-Q plot, let's discuss some of the probability distributions.

What are probability distributions?



In probability distributions, we represent data using charts where the x-axis represents the possible values of the sample and the y-axis represents the probability of occurrence.

There are various probability distribution types like Gaussian or Normal Distribution, Uniform distribution, Exponential distribution, Binomial distribution, etc.

In this blog, we will be looking into three types of distributions namely Normal, Uniform, and Exponential, and how we can identify them using a QQ plot.

- Normal distributions are the most popular ones. They are a probability distribution that peaks at the middle and decreases at the end of the axis. It is also known as a bell curve or Gaussian Distribution. As normal distributions are central to most algorithms, we will discuss this in detail below.

- Uniform distribution is a probability distribution type where the probability of occurrence of x is constant. For instance, if you throw a dice, the probability of any number is uniform.
- Exponential distributions are the ones in which an event occurs continuously and independently at a constant rate. It is commonly used to measure the expected time for an event to occur.

Why are probability distribution types important?

Probability distributions are essential in data analysis and decision-making. Some machine learning models work best under some distribution assumptions. Knowing which distribution we are working with can help us select the best model.

Hence understanding the type of distribution of feature variables is key to building robust machine learning algorithms.

Normal distributions

We regularly make the assumption of normality in our distribution as we perform statistical analysis and build predictive models. Machine learning algorithms like linear regression and logistic regression perform better where numerical features and targets follow a Gaussian or a uniform distribution.

It's an important assumption as normal distribution allows us to use the empirical rule of 68 – 95 – 99.7 and analysis where we can predict the percentage of values and how far they will fall from the mean.

In regression models, normality gains significance when it comes to error terms. You want the mean of the error terms to be zero. If the mean of error terms is significantly away from zero, it means that the features we have selected may not actually be having a significant impact on the outcome variable. It's time to review the feature selection for the model.

How Q-Q plots can help us identify the distribution types?

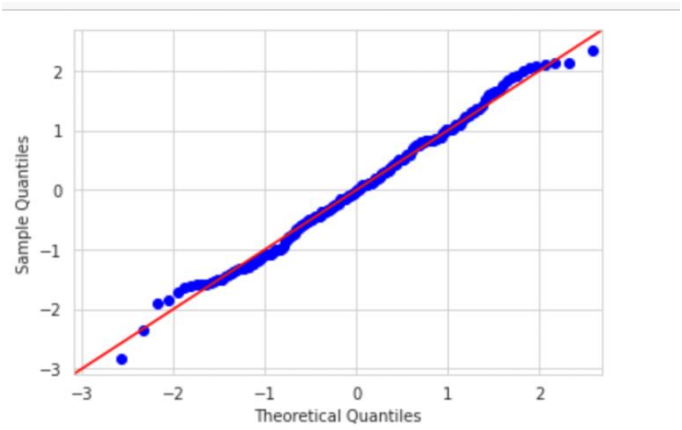
The power of Q-Q plots lies in their ability to summarize any distribution visually.

QQ plots is very useful to determine

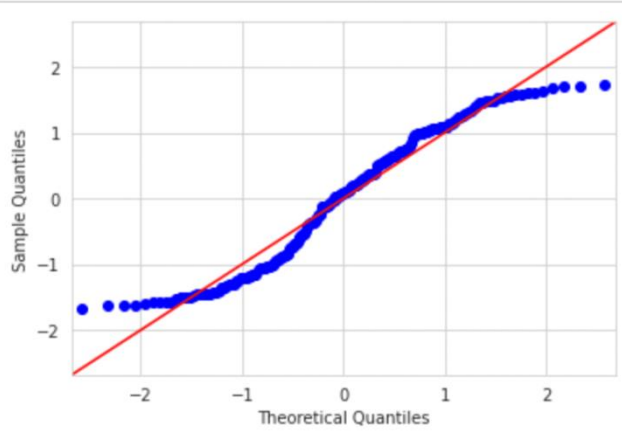
- If two populations are of the same distribution
- If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
- Skewness of distribution

  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.

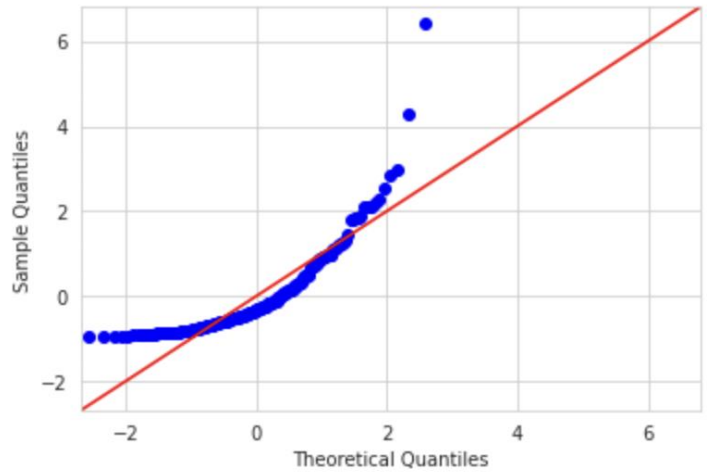


Uniform distrubution



in the above Q-Q plot since our dataset has a uniform distribution, both the right and left tails are small and the extreme values in the above plot are falling close to the center. In a normal distribution, these theoretical extreme values will fall beyond 2 & -2 sigmas and hence the S shape of the Q-Q plot of a uniform distribution.

Exponential Distribution

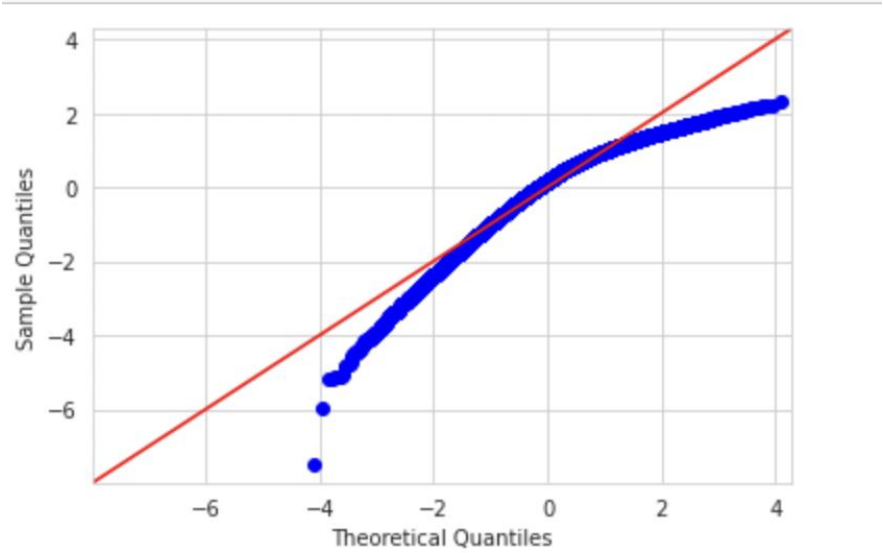


If we plot a variable with exponential distribution with theoretical normal distribution, the graph would look like below

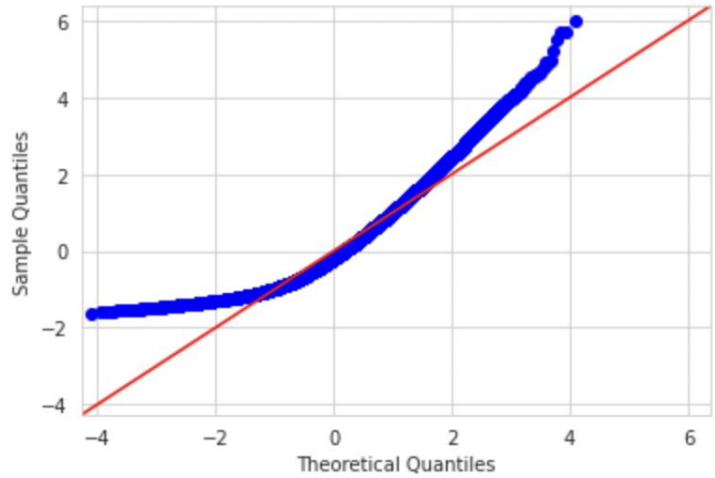
Q-Q plots and skewness of data

Q-Q plots can be used to determine skewness as well. If the see the left side of the plot deviating from the line, it is left-skewed. When the right side of the plot deviates, it's right-skewed

a left-skewed distribution



Similarly, a right-skewed distribution would look like below



As we build your machine learning model, we have to check the distribution of the error terms or prediction error using a Q-Q plot. If there is a significant deviation from the mean, we need to check the distribution of the feature variable and transform them into a normal shape