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?
Answer: In the Data provided there are categorical variables, they are as follows:

- Season- Defines the season (Summer,Winter,Fall,Spring) it was in that date.
- Yr- Defines the year of the date (2018,2019)
- Holiday-Defines whether the date was a public holiday or not
- Weekday- Defines the day of the week (Sun-Sat) it was on that day
- Workingday- Defines whether it was a working day or not(Holidays+Weekends)
- Weathersit-Defines the weather on that day(Clear, Light rain/snow, Mist, Heavy rain/snow)

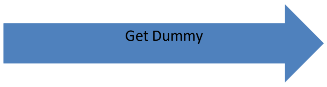
Out of all these we have found that “Yr”,“Weathersit” and “Season” have the most influence on the dependent variable. More specifically when we break it down

- Yr and, summer and winter (sub-sets of Season) have a positive correlation.
- Light rain/snow and mist (sub-sets of Weathersit) have a negative correlation.

Q2. Why is it important to use drop_first=True during dummy variable creation?
Answer: When we break down the categorical variables into dummy variables we essentially convert them into binary format of 0s and 1s , For example if we take a column that has 2 values say True and False , we can show them in a format as 0 = False and 1 = True , when we make dummy variables of this column we will get a table that contains 2 columns True and False , when the True column has a 1 then False will be 0 and when False is 1 , True will be 0 . In other words both the columns are inter dependant and can predict each other , therefore it does not make sense to keep both the columns .

We can look at 3 or more values in a column in the example below .

Sample column
Yes
No
Yes
May Be
Yes
No



Yes	**No**	**May Be**
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

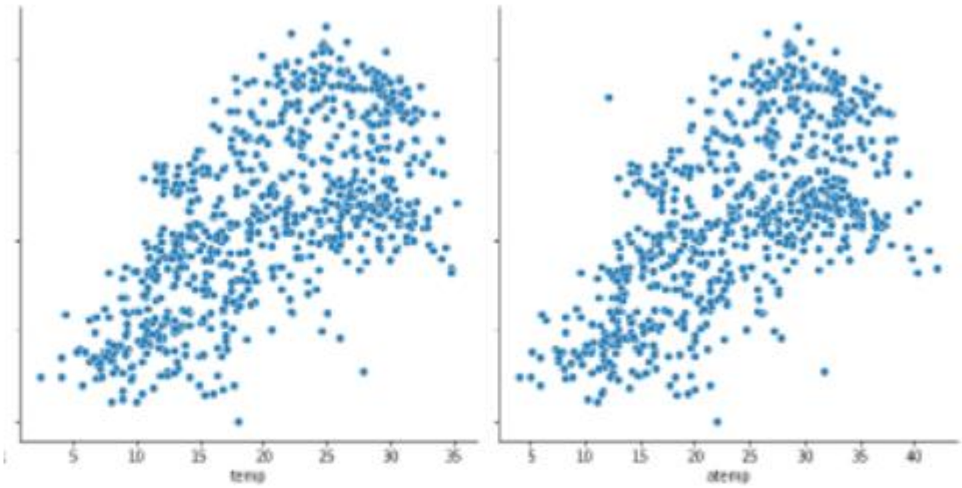
We can interpret this as :Yes= 1,0,0 No=0,1,0 Maybe=0,0,1

We can make the data more concise and less correlated by dropping one column and still retain the data as : Yes =1,0 No =0,1 and Maybe=0,0

Hence we drop one column and since we have the option to drop a column with the attribute drop_first while creating the dummies we do it during the creation itself.

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

Answer: Looking at the pair plots we can see that Temp, Atemp have the highest visual correlation to the target variable .



But we also see that both are highly related and therefore we drop atemp as it does not add any significant data along side temp.

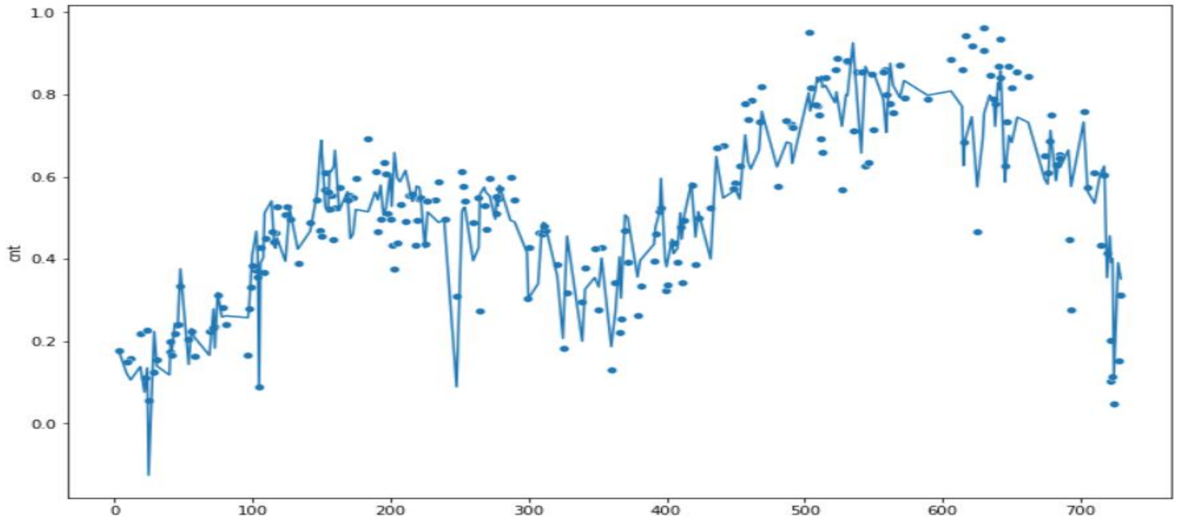
Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: The effectiveness of the model was validated by the following steps.

First we split the provided data into 2 parts using the train_test_split function in a ratio of 70:30, where 70% of the rows were used to train the model and 30% of the rows were kept aside as test or unseen data for validation of the model.

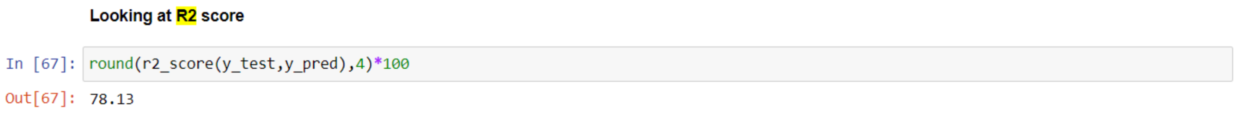
Next once the model was trained on the train set (70% of the original data) we asked the model to predict the target variable based on the unseen data or test data(30% of the data that was kept for validation) . We then check the closeness of predicted and the actual value using 2 methods.

1. Using visualization where we can see if there are any major fall out that the model fails at.



The Dots represent the actual value and the line shows us the predicted value.

2. By using the in-build function of scikitlearn called R2.



Looking at R2 score

```
In [67]: round(r2_score(y_test,y_pred),4)*100
Out[67]: 78.13
```

Here we see that the model has an accuracy of 78% i.e. It has predicted 78% of the data accurately.

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

Answer: The 3 most important features contributing towards explaining the demand of the shared bikes in decreasing order of importance (most important to least important) are as follows:

1. Temp (coef: 0.579, Positive effect)
2. Light snow/rain (subset of weathersit , coef: -0.271 , Negative effect)
3. Yr(Coef: 0.232, Positive effect)

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Linear regression is the most common and most basic machine learning algorithm .It is based on supervised learning. This model uses a single or group of independent variables to predict a single target variable. It is mostly used to find the relation between the independent variable and the target or dependant variable. Different prediction models differ based on the kind of relation between the independent and dependant variables and the number of independent variables being used .
A typical liner regression has a formula as follows :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Here Y is the dependent variable and X is the independent variable, Beta is the coefficient of X .

The strength of a linear regression model is mainly explained by R2, where R2 = 1 - (RSS / TSS)
RSS: Residual Sum of Squares
TSS: Total Sum of Squares

R2 tells us how well the model fits the data by measuring the proportion of the variability in Y that can be explained by X. For that we need the RSS and Total Sum of Squares (TSS).
Apart from R2, there is one more quantity named RSE (Residual Square Error) which is linked to RSS.

$$RSE = \sqrt{\frac{1}{n-2} RSS}$$

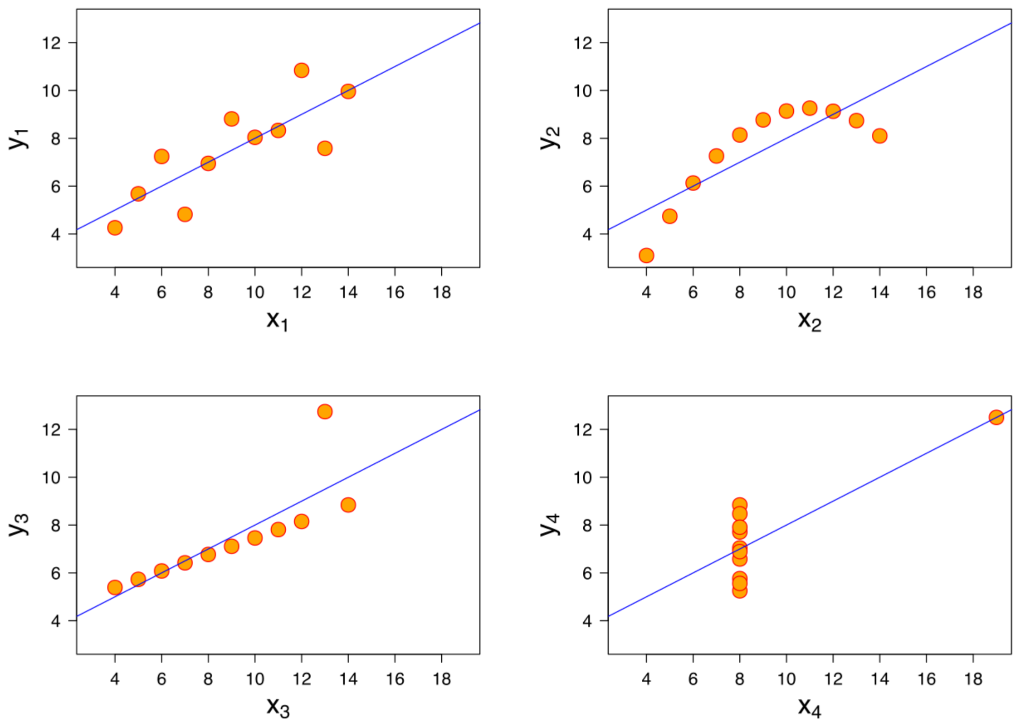
The RSE is an estimate for the standard deviation of the true regression line. This means that on average the value of y is one RSE away from the true regression line.

Linear regression models are usually used as a way to answer questions like :

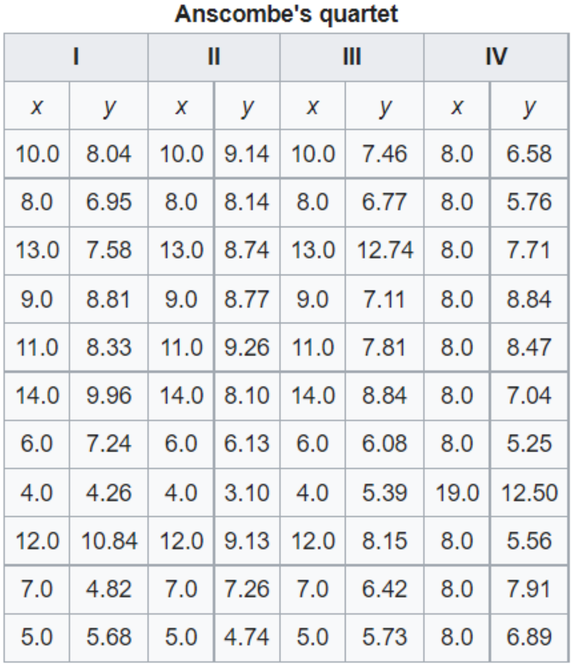
“Is the rise of the sea level connected to rising temperatures?”,
“How expensive will a house with 3 bedrooms be?”
“How many items do we sell if we increase our marketing budget by 20%?”

2. Explain the Anscombe’s quartet in detail.

Lets say we have a data set, we might want to use stats to understand the data such as its , mean , variance ,correlation coefficient and a line of best fit. Now suppose we are given 3 more sets with identical summery stats . We would assume that all the sets would look very similar when plotted but when we plot it we find the plots as below .



They don’t look similar at all. These sets have the exact same stats summery.



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

These are called the **Anscombe's quartet .** They are 4 Data sets that exactly the same in every statistical value.

It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.

The Anscombe`s Quartet tells shows us how import it is to visualize the data before we come to a conclusion on its trends and also shows us how the outliers influence the statistics of a data frame.

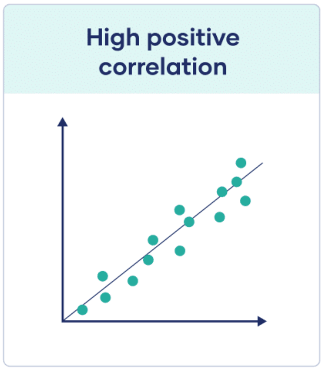
Q3.What is Pearson's R?

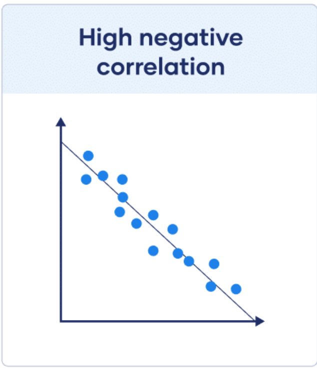
Answer: Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation. The full name is the Pearson Product Moment Correlation (PPMC). It shows the linear relationship between two sets of data. In simple terms, it answers the question, Can I draw a line graph to represent the data? Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample. This might seem and look similar to a simple linear regression model but the problem here is that this This method does not differentiate between dependant and independent variable and assumes that either can predict the other.

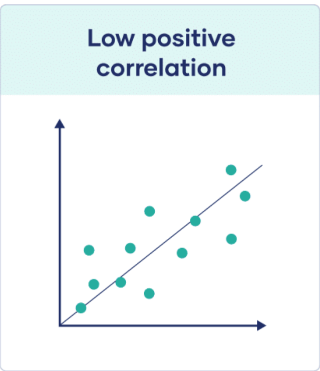
When we have a pair of variables that fit well on a line when plot, we can call it a high correlation. When the line passes though the data points exactly then the coefficient will be 1 on the other hand if the pair of variables have no or low correlation then neither can be used to predict the other.

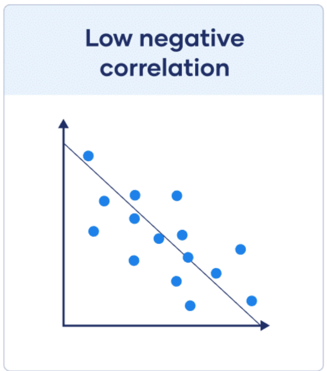
In Python we can see the correlation using the command "dataset.corr()" which gives is a matrix of all the correlations between all the features of the data set.

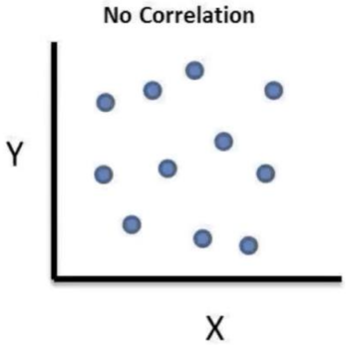
Lets look at a few examples of the same :











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

Answer: Scaling is a part of data Pre-Processing which is applied to independent variables to normalize the variable within a particular range. Scaling also helps speed up the calculations in an algorithm.

Usually, a collected data set contains features that are highly varying in magnitudes, units and range. If scaling isn't done then algorithm only takes magnitude in account and not units hence the model might be incorrect trained and give the wrong weightage to the variable. To solve this issue, we do scaling so that all the variables are brought to the same level of magnitude.

Noremalized scaling is a process of bringing all the values of a numeric variable to a range of 0 to 1. This method uses the highest value as 1 and lowest value as 0 and all the values in between are replaced as a faction of the max value for example .

Original		Normalized Value
100		1
20		0.2
35	→	0.35
55		0.55
0		0

In Python this is done by using MinMaxScaler which is a function in sklearn.preprocessing

Standardized scaling is a process of bringing the data into a normal distribution by replacing the data with its Z score centring it around the mean . The major advantage of Standardized scaling is that it does not have a bound range so if your data does have any outliers then their presence will not distort the scaling of the other values .

In Python we implement standardized scaling by using the method scale form sklearn.preprocessing.

Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbors and Neural Networks.

Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true.

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

Answer: This happens when there is a perfect correlation between the variables that is the correlation is 1.

When that is the case R2 = 1 and VIF = 1/(1-R2) . 1-R2= 0 and when we divide by 0 the VIF = infinite .

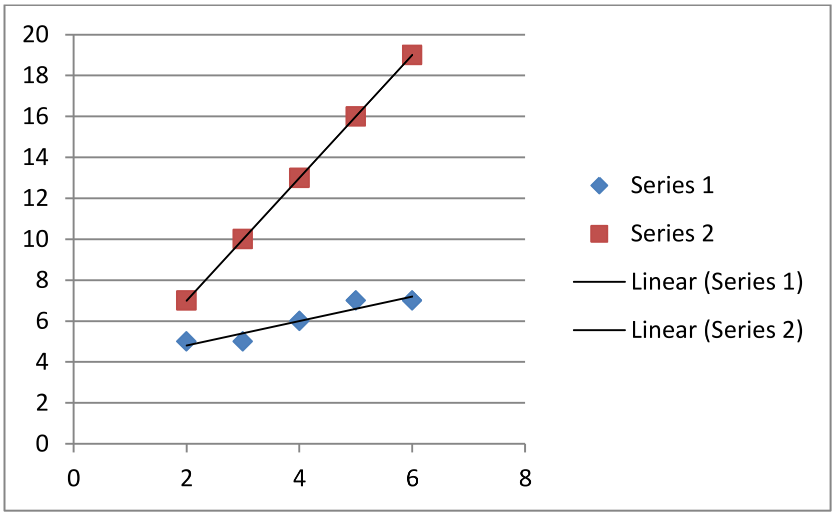
An infinite VIF means that we may be able to express the variables as a perfect linear model by them selves.

The only way to solve this problem is to drop one of the columns that show this VIF.

Example :

X	Y	Z
2	5	7
3	5	10
4	6	13
5	7	16
6	7	19

Here if we plot these :



We can clearly see the red plot(X vs Z) will have a correlation of 1 resulting in a VIF of Infinite. Whereas the VIF of X vs Y will not be Infinite as the line doesn't pass exactly though the centre of the points.

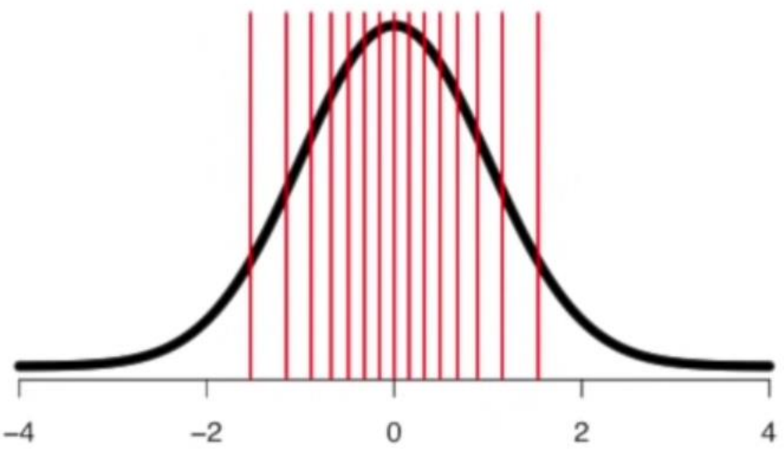
Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

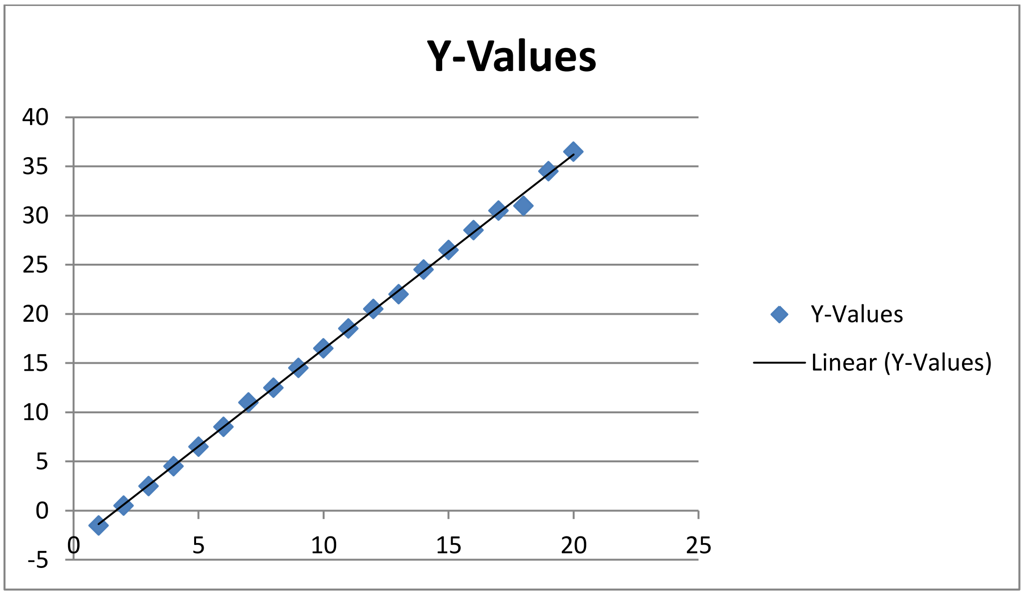
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

A Q-Q Plot can also be used to check the kind of distribution of a single data set. We can plot the quantiles of the data set against a normal distribution, or a uniform distribution and if we can draw a line through those plots cleanly then the data is distributed as per the distribution chosen . Note while using these distributions we need to divide these distributions into the same number of Qs as that in our data set and while doing that in a normally distributed set we need to draw our Qs based on equal probability meaning that the higher probability points will come in thinner Qs and as the data moves away from the centre we should take wider Qs to compensate for the lower probability .



As we can see we must have narrow Qs near the centre and as it moves away we make the Qs wider.

If our data was normally distributed then our plot would look something like this .



If our data was not normally distributed then the points who not fall close to the line.