Table of Contents

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?

- Fall and summer are having max cnt followed by winter.
- cnt is more when there is no holiday, so clearly holiday column cannot be good predictor.
- Working day for both 0's and 1's median is almost same. And max cnt showing nearly 6000 so it can be a good predictor.
- Weathers it showing the good trend for the clear followed by misty. So, it can be good predictor.
- Months from April to Oct showing good trend so mnth can be good predictor.
- Weekdays median for all the days showing over 4000. so, at this point not sure about it can be good predictor.
- year 2019 showing the good trend.

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

⇨ If we are not using drop_first= true then we will get a redundant feature i.e. dummy variables might be correlated because the first column becomes a reference group during dummy encoding. Therefore we are using drop_first= True.

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

⇨ The numerical variable ‘registered’ has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, when we drop registered due to multicollinearity the numerical variable ‘atemp’ has the highest correlation with the target variable ‘cnt’

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

- ⇨ The Assumptions of the linear regression are: -
- ⇨ Validating the assumption of Linear Regression Model:
- ⇨ - Linear Relationship
- ⇨ - Homoscedasticity
- ⇨ - Absence of Multicollinearity
- ⇨ - Independence of residuals
- ⇨ - Normality of Errors

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

=> Top 3 predictor variables of the model are:

1. temp: A coefficient value of '0.564438' indicated that a temperature has significant impact on bike rentals

2. weathersit: A coefficient value of '-0.307082' indicated that the light snow and rain deters people from renting out bikes

3. yr: A coefficient value of '0.230252' indicated that a unit increase in yr variable, increases the bike hire numbers by 0.230252 units.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

- ⇨ Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

- ⇨ Mathematically the relationship can be represented with the help of following equation
  $Y = mX + b$

- ⇨ Here, Y is the dependent variable we are trying to predict

- ⇨ X is the independent variable we are using to make predictions.

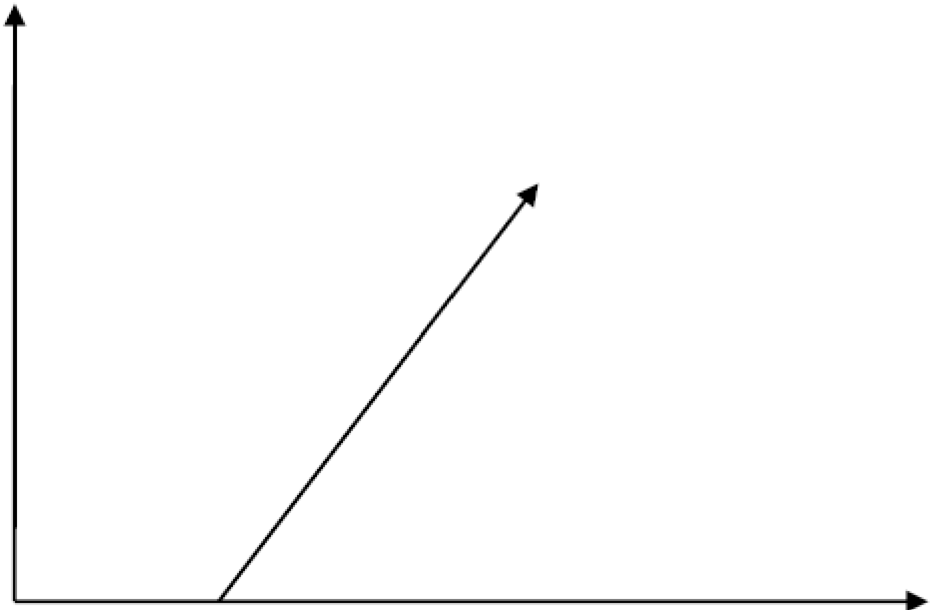
- ⇨ m is the slop of the regression line which represents the effect X has on Y

- b is a constant, known as the Y-intercept. If X = 0,Y would be equal to b.

- Furthermore, the linear relationship can be positive or negative in nature.

Positive Linear Relationship

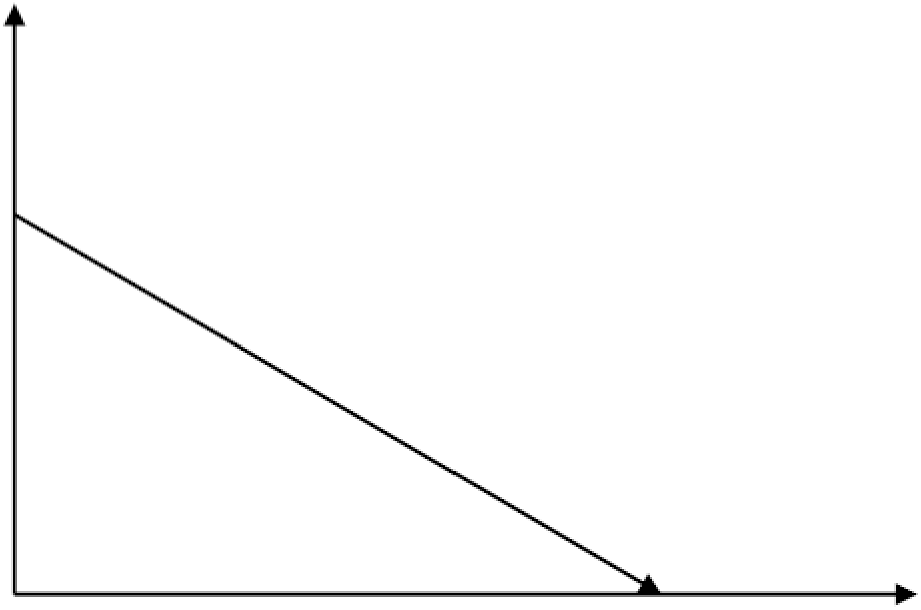
A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



Positive Linear Relationship

Negative Linear relationship

A linear relationship will be called negative if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Negative Linear Relationship

Types of Linear Regression

Linear regression is of the following two types –

- **Simple Linear Regression:** Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:
- One variable, denoted x is regarded as the predictor, explanatory, or independent variable.
- The other variable, denoted y, is regarded as the response, outcome, or dependent variable.
- **Multiple Linear Regression:** Multiple linear regression is a statistical analysis technique used to predict a variable's outcome based on two or more variables. It is an extension of linear regression and known as multiple regression. When there are two or more independent variables used in the regression analysis, the model is not simply linear but a multiple regression model. The variable to be predicted is the dependent variable, and the variables used to predict the value of the dependent variable are known as independent or explanatory variables.

Assumptions of Linear Regression

- **Linear relationship:** One of the most important assumptions is that a linear relationship is said to exist between the dependent and the independent variables.

- **No auto-correlation or independence:** The residuals (error terms) are independent of each other. In other words, there is no correlation between the consecutive error terms of the time series data.

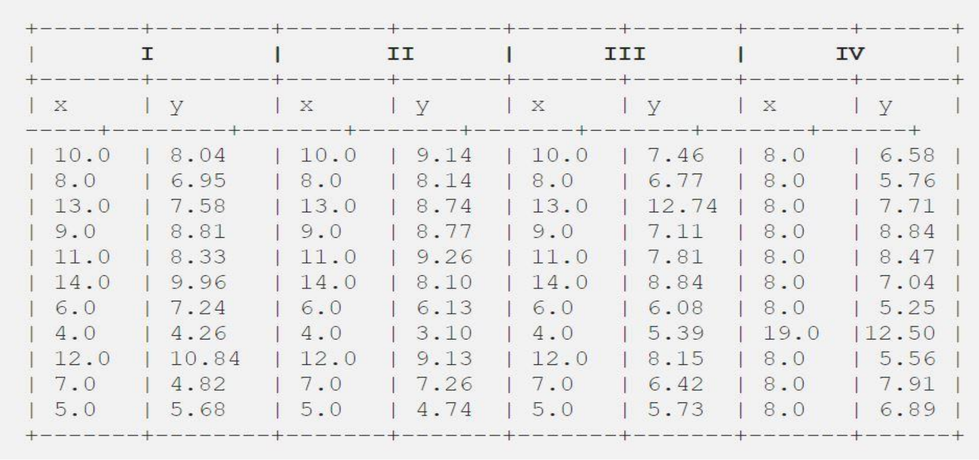
- **No Multicollinearity** :The independent variables shouldn't be correlated. If multicollinearity exists between the independent variables, it is challenging to predict the outcome of the model.

- **Homoscedasticity**: Homoscedasticity means the residuals have constant variance at every level of x. The absence of this phenomenon is known as heteroscedasticity.

- **Normal distribution of error terms:** The last assumption that needs to be checked for linear regression is the error terms' normal distribution.

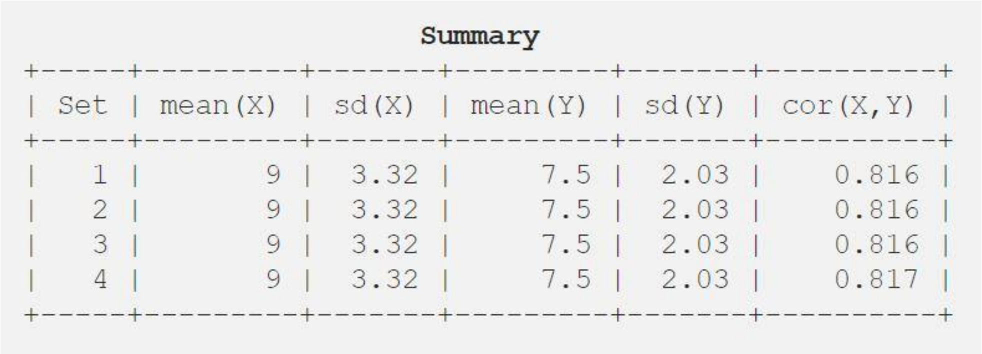
Q2. Explain the Anscombe's quartet in detail.

- 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.
- Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points and plot those points. Those 4 sets of 11 data-points are given below.



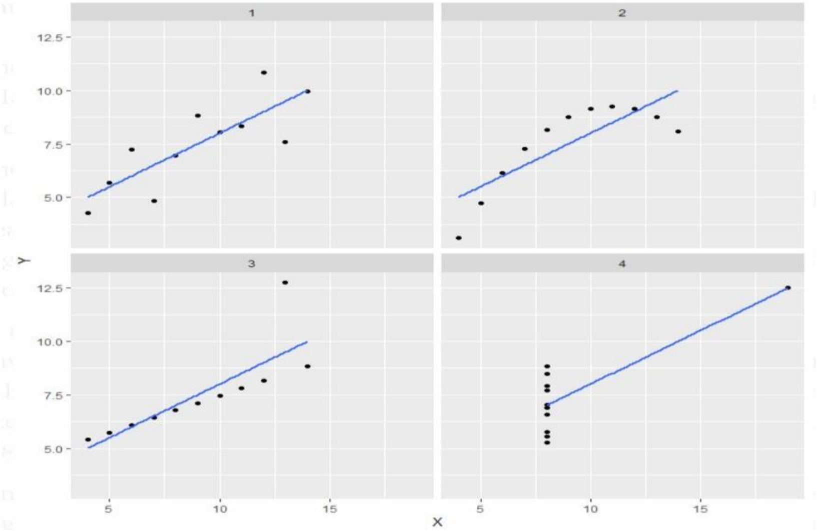
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

- After analysing them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y as:



```
                         Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

⇨ After plotting these points in scatter plots:



⇨ It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

Explanation of this output:

⇨ In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.

⇨ In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.

⇨ In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.

⇨ Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Application:

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

Q3. What is Pearson’s R?

- It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

- In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation.

- Generally, Pearson's correlation coefficient, capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

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

- Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

- Using the formula proposed by Karl Pearson, we can calculate a linear relationship between the two given variables.

- we can calculate the relationship between these two variables by obtaining the value of Pearson's Correlation Coefficient r. There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where,

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

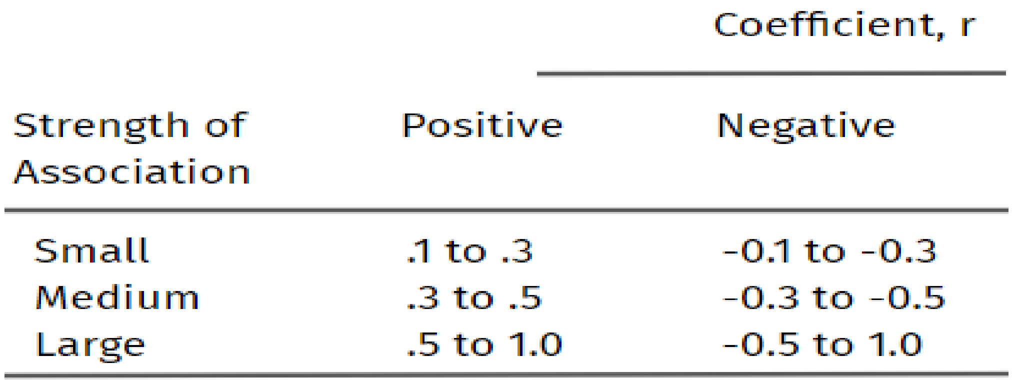
Σy2 = the sum of squared y scores

Determining the strength of the Pearson product-moment correlation coefficient

As we have learned from the definition of the Pearson product-moment correlation coefficient, it measures the strength and direction of the linear relationship between two variables.

The more inclined the value of the Pearson correlation coefficient to -1 and 1, the stronger the association between the two variables.

Below, we have shown the guidelines to interpret the Pearson coefficient correlation :



	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

A notable point is that the strength of association of the variables depend on the sample size and what you measure.

Strength: Strength implies the relationship connection between the two given factors. It implies how reliably one variable will change because of the adjustment in the other. Qualities that are near +1 or - 1 show a solid relationship.

Direction: The direction of the line demonstrates a positive direct or negative straight connection between factors. On the off chance that the line has an upward slant, the factors have a positive relationship. This implies an expansion in the estimation of one variable will prompt an increment in the estimation of the other variable. A negative relationship portrays a descending slant. This implies an expansion in the measure of one variable prompts a lessening in the estimation of another variable.

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

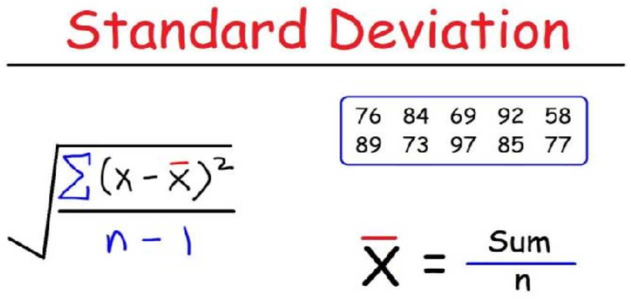
- In Statistics, the variables or numbers are defined and categorised using different scales of measurements. Each level of measurement scale has specific properties that determine the various use of statistical analysis.
- Feature Scaling is one of the important pre-processing that is required for standardizing/normalization of the input data. When the range of values are very distinct in each column, we need to scale them to the common level. The values are brought to common level and then we can apply further machine learning algorithm to the input data.

The Feature scaling is required because:

- Regression Coefficients are directly influenced by scale of Features
- Features with higher scale dominates over lower scale features
- Gradient Descent can be achieved easily if we have scaled values
- Some of the Algorithms would reduce time of execution, if scaled.
- Some Algorithms are based on Euclidean Distances, Euclidean distances are very sensitive to the feature scales.

Different Feature Scaling Techniques

- We can use different Scaling Techniques in order to scale the input dataset. We can apply either of the following:
- Standardization: Standardization is based out of Standard Deviation. It measures the spread of value in the features. This is one of the most used.

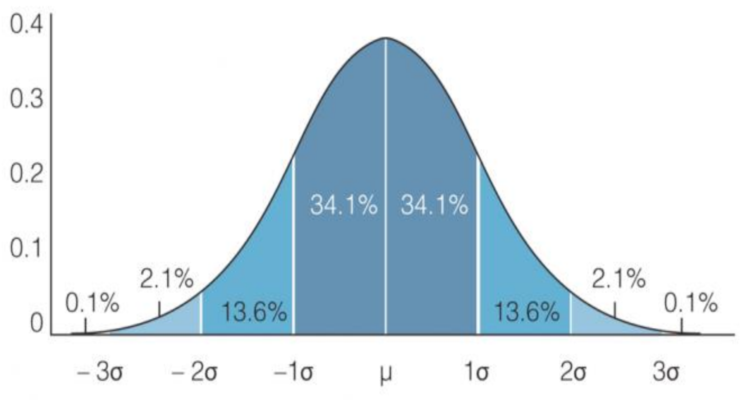

Standard Deviation
76 84 69 92 58
89 73 97 85 77
$\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
$\bar{X} = \frac{\text{Sum}}{n}$

During standard scaling, we shift the mean of the features to value 0 and have standard deviation as 1. When the standard scaler is applied, we get values in the range of -3 to 3

- Centres the variable at 0 and sets the variance to 1.

$$\text{Z-score} = \frac{X - mean(X)}{Std(X)}$$

⇨ When Standard deviation is applied over the features values, 99.7% of values in the feature set ranges between -3 SD (Standard Deviation) to 3 SD(Standard Deviation).



⇨ Normalization: Normalization is the concept of scaling the range of values in a feature between 0 to 1. This is referred as Min-Max Scaling.

⇨ In the above equation:

Xmax and Xmin is Maximum and Minimum Value of the feature column

⇨ The value of X, is always between Minimum and Maximum Value

Difference between Normalisation and Standardisation

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

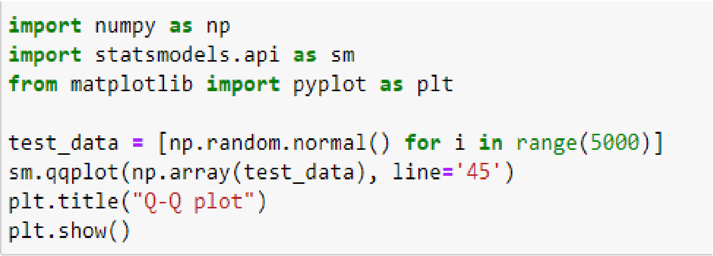
⇨ If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this 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).

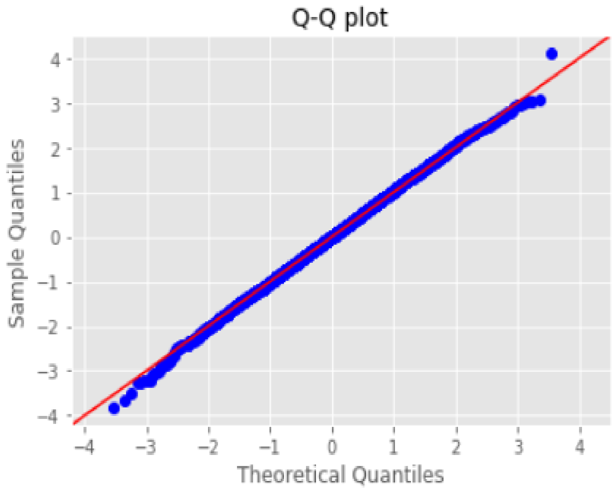
⇨

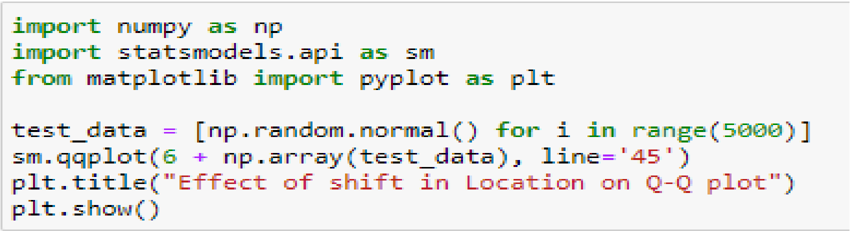
Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- ⇨ The quantile-quantile plot or Q-Q plot is a graphical tool to validate if two datasets are coming from populations with common distribution.
- ⇨ Very often we assume given data to be normally distributed for ease of inferring useful information. One way to assess our assumption's correctness is to use Q-Q plot. Not just Normal distribution, we can test for other distributions (for eg. uniform distribution etc.) as well.
- ⇨ Quantiles are the breakpoints that divide the ordered numerical data into equal sized bins.
- ⇨ Percentiles are a type of quantiles that divide the data into 100 equal bins, quartiles divide the data into 4 equal parts and so on.
- ⇨ Q-Q plot compares the quantiles of 2 datasets. We can make Q-Q for any 2 datasets as long as the quantiles can be calculated for both of them.
- ⇨ Importance of Q-Q plot in Linear Regression:
- ⇨ 1. Two datasets/sample can be of different size.
- ⇨ 2. Q-Q plot can detect outliers, shifts in scale, location, symmetry etc. simultaneously.
- ⇨ 3. One of the important assumptions of Linear Regression is that the residual of the model is normally distributed. This can be assessed using Q-Q plot.
- ⇨ Example of Q-Q plot:
- ⇨ Here, first 5000 normally distributed random points are generated.
- ⇨ Then the random points are fed into the Q-Q plot. The blue dots representing the random points are aligning with 45 degree reference straight line in red. This re-confirms the test_data is actually normally distributed.
- ⇨ As test_data is shifted in location, the blue dots have shifted on the left side of the 45 degree reference line. Thus Q-Q plot can show different statistical aspects.

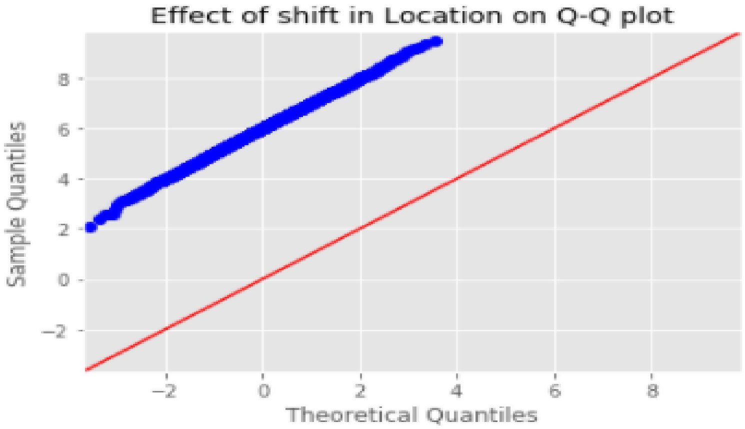


```
import numpy as np
import statsmodels.api as sm
from matplotlib import pyplot as plt

test_data = [np.random.normal() for i in range(5000)]
sm.qqplot(np.array(test_data), line='45')
plt.title("Q-Q plot")
plt.show()
```





```
import numpy as np
import statsmodels.api as sm
from matplotlib import pyplot as plt

test_data = [np.random.normal() for i in range(5000)]
sm.qqplot(6 + np.array(test_data), line='45')
plt.title("Effect of shift in Location on Q-Q plot")
plt.show()
```



⇨ The q-q plot is used to find out the following:

⇨ Whether the two data sets come from populations with a common distribution

⇨ Whether the two data sets have a common location and scale?

⇨ Whether the two data sets have similar distributional shapes?

⇨ Whether the two data sets have similar tail behaviour?