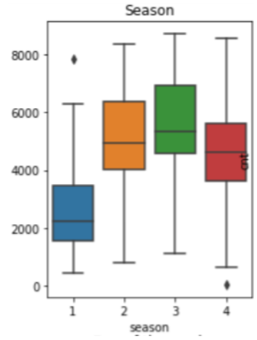
Assignment-based Subjective Questions

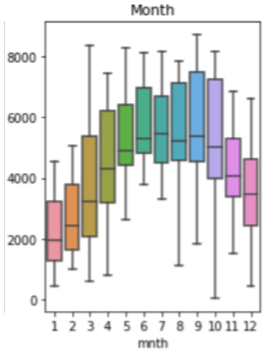
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: The categorical variables from the dataset are: **season, mnth, weekday, weathersit, holiday and workingday**.

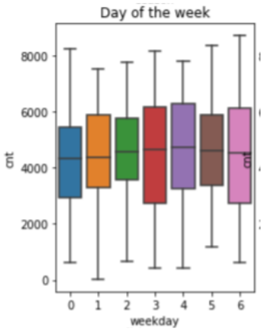
- **Season**: Most of the bike booking from BoomBikes happen in rainy season with a Median above 5000 count. The second highest number of bookings happen in Summer close to 5000 count. Spring has least number of bookings. Season is a good predictor variable.



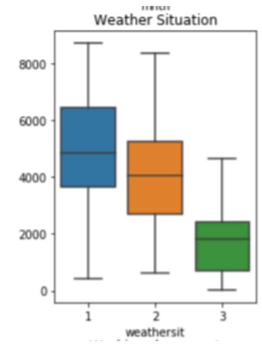
- **mnth**: Most of the booking happen from May to September close to 5000 booking per month. April, October and November months have a booking of more than 4000.



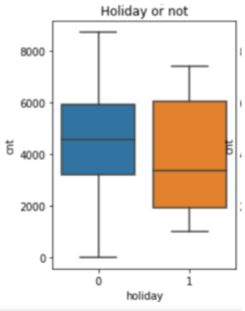
- **weekday (Day of the week):** 0 indicates Tuesday and 6 indicates Monday here. All the days have same median around 4300 to 5000 approximately. The model shows Monday would be better day.



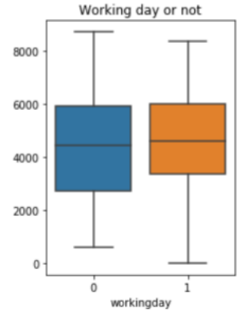
- **weathersit**: When the weather situation is cloudy or Mist (value = 1), there are a greater number of bookings happening on the BoomBikes. The median is close to 5000 bookings



- **holiday**: The plot shows median more than 4500 counts, when it is not a holiday.



- **workingday**: Workingday has value 1 for a working day. Working day have higher number of bookings.



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

Answer: Dummy variables are created for categorical variables. If there are ‘k’ number of values in a categorical variable, then we need to represent it in (k-1) dummy variables and then drop the original column. The drop_first=True will achieve in creating (k-1) dummy variables and mapping the values accordingly.

E.g., Four seasons are indicated by 3 dummy variables:

Dummy variables for Season column

Season	Season_summer	Season_rain	Season_winter
1: spring	0	0	0
2: summer	1	0	0
3: fall	0	1	0
4: winter	0	0	1

(i) Season (1:spring) when Season_summer=0, Season_rain=0, Season_winter=0

(ii) Season (2:summer) when Season_summer=1, Season_rain=0, Season_winter=0

(iii) Season (3:fall) when Season_summer=0, Season_rain=1, Season_winter=0 indicated by Season_rain

(iv) Season (4:winter) when Season_summer = 0, Season_rain=0, Season_winter=1 indicated by Season_winter

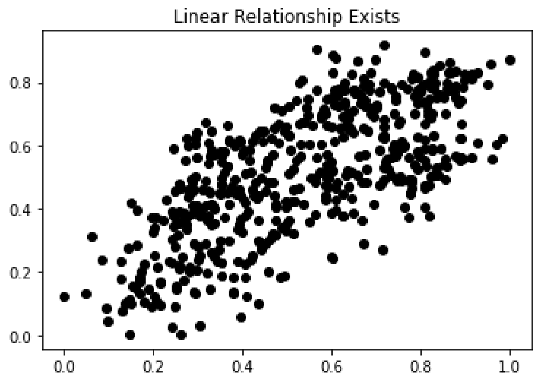
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Temp variable has highest correlation

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

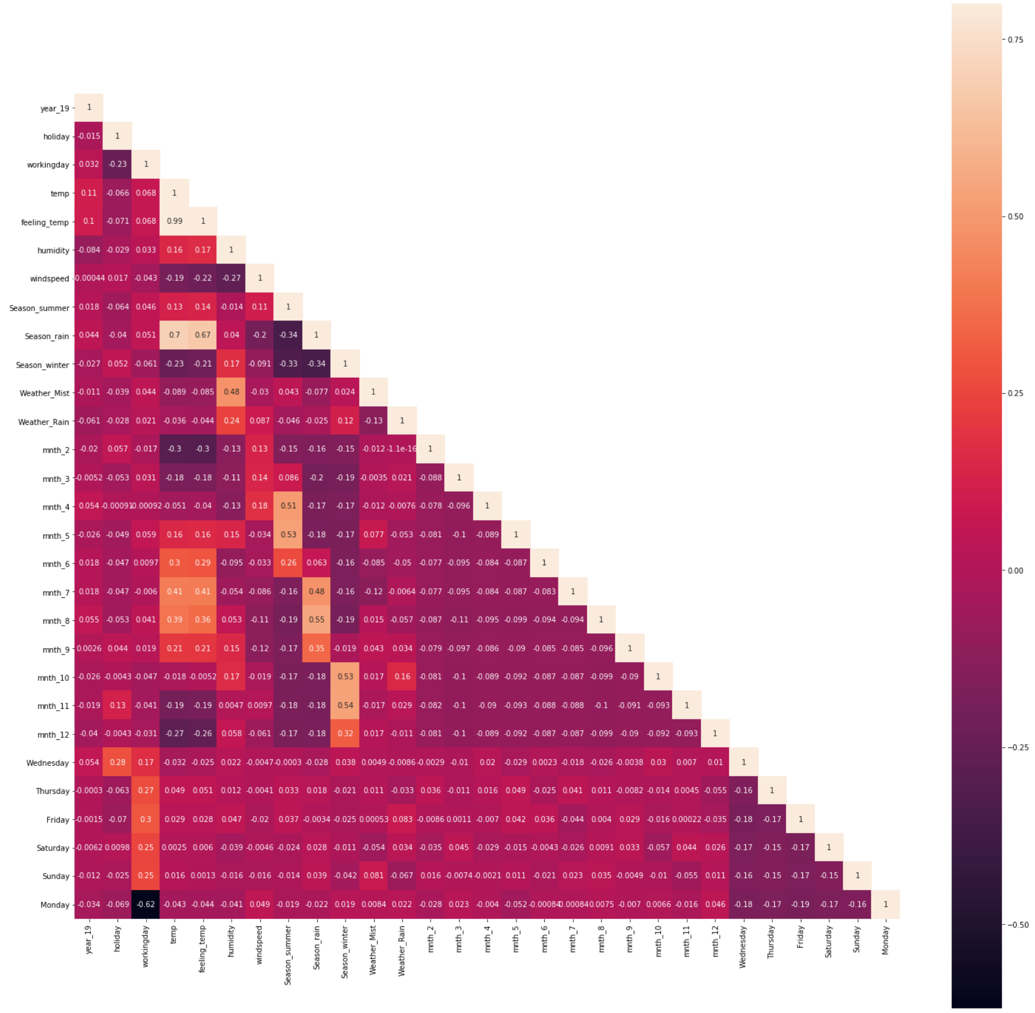
a) Linear Relationship between the features and the target

- Linear relationship exists with 'temp' feature



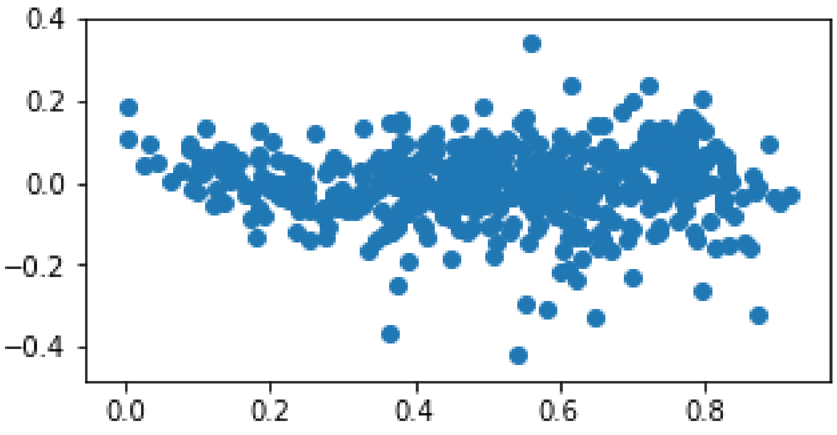
b) Little or no Multicollinearity between the features

- We can note that there are no multi collinearity



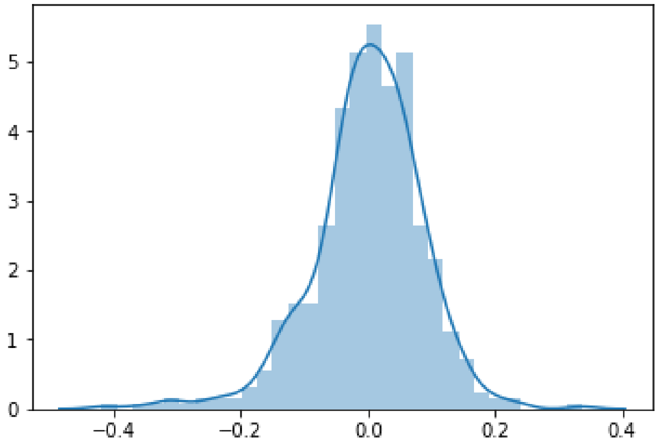
c) Homoscedasticity Assumption

We are unable to findout any pattern with the noise, hence it is a Homoscedasticity



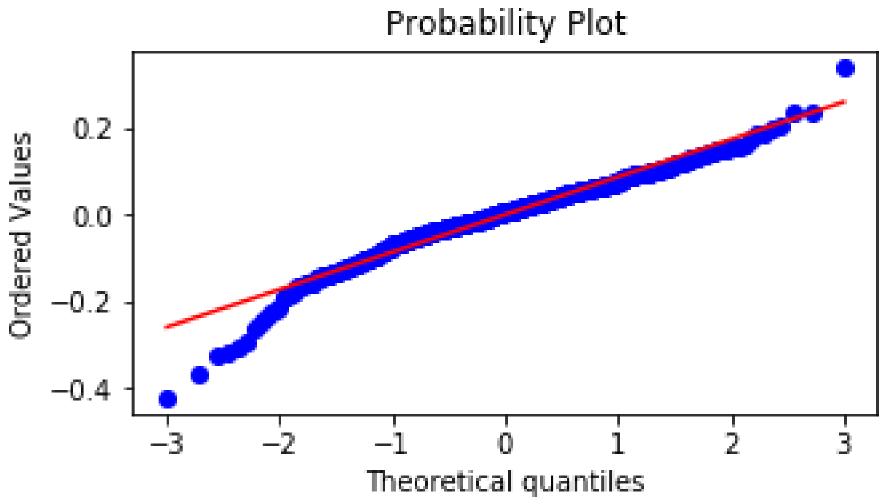
d) Error terms are normally distributed with mean zero

- We can notice almost normal distribution with residuals



e) Variables follow a Normal Distribution

- We can notice almost normal distribution on the q-q plot



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

- **temp**: The bike demand in BoomBikes increases by 0.5309 times when there is raise in temperature.
- **year_19**: In the year 2019, there has been an increase in demany in BoomBikes by a ratio of 0.2292 when compared to 2018. As the years increases, the demand will also increase
- **Season_winter**: The demand for Bike sharing is high in winter by 0.1349 times

General Subjective Questions

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

Answer: Linear Regression algorithm is based on Supervised Learning methods. It is used to predict the target variable. There are two types of linear regression:

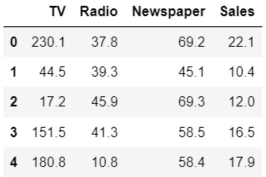
- Simple linear regression
- Multiple linear regression

a) **Simple linear regression**: This is basic type of regression model. This explains relationship between dependent variable and one independent variable and is depicted by a straight line. Usually, a scatter plot is plotted, and straight line is obtained from the data points.

There are four steps in Simple Linear Regression:

1) Reading and understanding the data:

○ Let us consider the example of advertising data set.



	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0
3	151.5	41.3	58.5	16.5
4	180.8	10.8	58.4	17.9

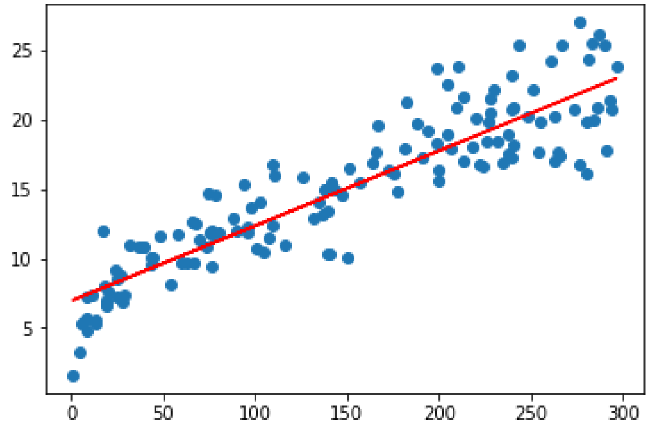
The target variable is Sales (which is y). We will predict using one independent variable TV. In this case, y = c + m1 X TV m is model coefficients or model parameters, c is intercept and y is the response.

We will then split the model in 70:30 or 80:20 as Train:Test data set respectively

2) Training the model

○ We use the Train data set and build the model. We will then use the model and verify with the test data set. We will get a straight line as shown below.

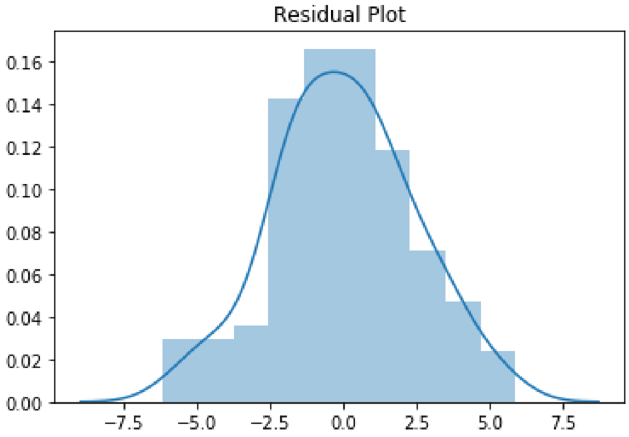
○ We will do the prediction



-

3) Residual Analysis - whether the residuals are normally distributed or not

- We will subtract y_train data with y_prediced value. If we plot the residuals, we should get a normally distributed curve which indicates that the model that we have built is good.



4) Predicting on the test set and we evaluate the model

- We will use same model and test it with our Test dataset.

R2 Statistics: The accuracy of the model is evaluated by R2 value. It is a number and value lies between 0 and 1. The higher the value, the better is the model.

b) Multiple Linear Regression:

This technique shows the relationship between one dependent variable with several independent variable. The aim of multiple regression is to figure out best value of variable y, when multiple independent variables on X have been selected.

It is noted that whenever a new feature or an independent variable is added, the better R-squared value is obtained. This shows that instead of using one single variable or feature, combination of multiple features will give better predictions.

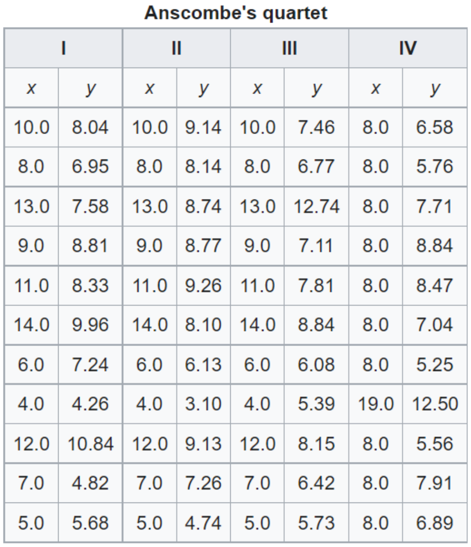
Sometimes adding more variables unnecessarily or all the variables would give rise to issues like 'overfit' and 'multicollinearity'. Hence it is important to select the best features and eliminate rest of them from the model. In multiple linear regression, this is achieved through various approaches iteratively. Each step is validated through R-squared calculation and p-value for each feature. This can also be validated through the use of VIFs (Variance Inflation Factors).

Steps to create Multiple Linear Regression model:

- Reading, Understanding and Visualizing the data
- Preparing the data for modelling (train-test split, scaling etc)
  - Encoding:
    - Converting binary categorical variables to 1s and 0s
    - Other categorical variables to dummy variables
  - Splitting into train and test
  - Rescaling of variables
- Training the model
- Residual analysis
- Predictions and evaluations on the test set

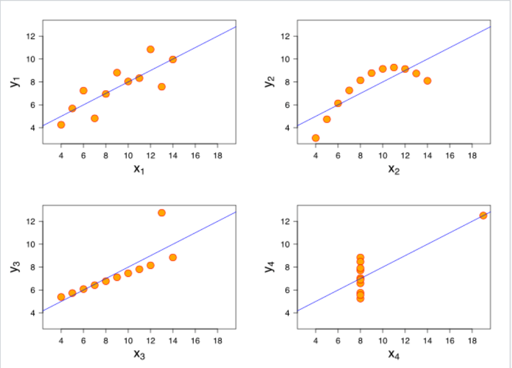
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: Francis Anscombe built Anscombe's quartet in 1973. With this, he tried to explain the importance of plotting the data before doing the analysis. He used four data sets. Statistical observation on these data sets would give same mean and variance. But, when we plot the data sets, it looks different from one another.



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



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Anscombe used the above data set. If we derive statistical information, it will be the same for all the 4 datasets. Mean for x will be 9.00 and for y will be 7.50. Similarly the Standard deviation for x will be same for all the four datasets with value 3.16 and for y SD is 1.94.

However, if we examine the scatter plot, we can note that first graph fits the linear regression. Second graph is non-linear in nature and is curved. Third and fouth shows outliers. Hence does not fit linear regression model.

3. What is Pearson’s R? (3 marks)

Answer: The Pearson Correlation coefficient is denoted by Pearson’s r.

Pearson’s r = (covariance of two variables) / (product of their standard deviations)

The values will always be between -1 and +1. If the value is exactly 1, it means that a linear equation with perfect relationship exists between X and Y and all the points in the data set lies on the line. A zero means that there is no relationship exists between X and Y. Similarly -1 indicates perfect linear but negative relationship between X and Y.

If Pearson’s r is greater than zero, it means a positive association exists. If the value of one variable increases, then the other tend to increase.

Similarly, if Pearson’s r is less than zero, it means a negative association exists. If the value of one variable increases, then the other will decrease.

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

Answer:

Scaling: It is a process of normalizing the data of numeric variables in the data set. It is generally performed during the data pre-processing step.

When we have to perform Linear Regression, the numeric features will have values with different scale or with different range and comparing relationship with other variables will not make sense. It is important to normalize the data before we start with start building the model and it is called Scaling. E.g., if we have different columns like age, price, salary etc with different range, scaling will help them to normalize and to have the values between 0 and 1.

Normalized Scaling is also called as Min-Max scaling.

The formula used is [x – min(x)] / [max(x) – min(x)]

Here min(x) is minimum value of column and max(x) is maximum value.

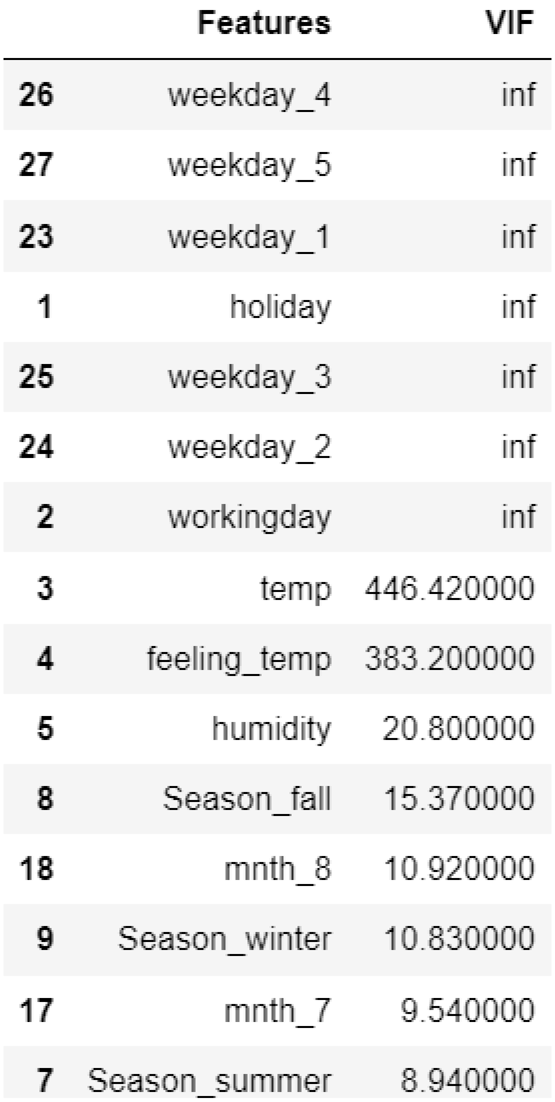
Standardized Scaling ensures that the data in the selected column will have a mean = 0 and variance = 1.

It is calculated by ratio of distribution mean over standard deviation of that feature.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: In the day.csv bike assignment dataset, we can note that few of the features will have infinite VIF when calculates using all the columns.

This shows that there is a perfect correlation exists between these variables.



	Features	VIF
26	weekday_4	inf
27	weekday_5	inf
23	weekday_1	inf
1	holiday	inf
25	weekday_3	inf
24	weekday_2	inf
2	workingday	inf
3	temp	446.420000
4	feeling_temp	383.200000
5	humidity	20.800000
8	Season_fall	15.370000
18	mnth_8	10.920000
9	Season_winter	10.830000
17	mnth_7	9.540000
7	Season_summer	8.940000

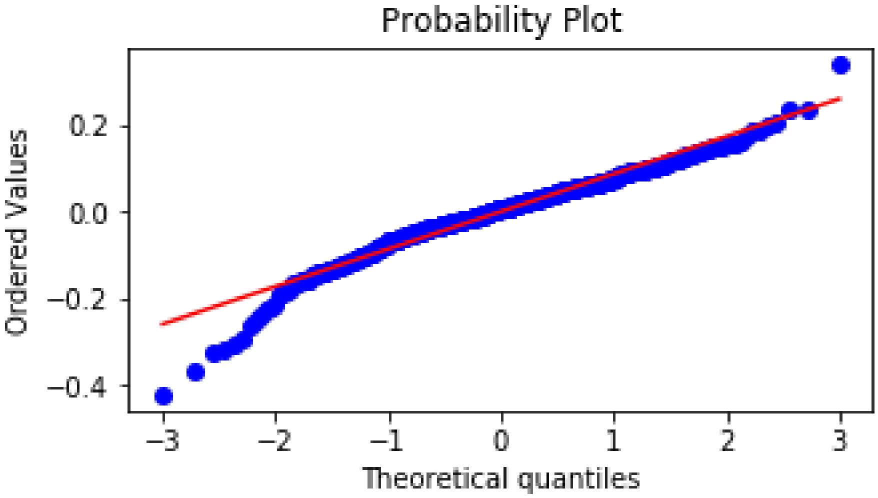
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: The Q-Q plot is the quantile-quantile plot.

One of the assumptions in linear regression analysis is the normal distribution of error terms. To test this normality, we can use Q-Q plot.

If the error terms are not normally distributed, the confidence interval becomes unstable and it will not be a better model.

The Q-Q plot from the assignment is given below:


Probability Plot
Ordered Values
0.2
0.0
−0.2
−0.4
−3
−2
−1
0
1
2
3
Theoretical quantiles