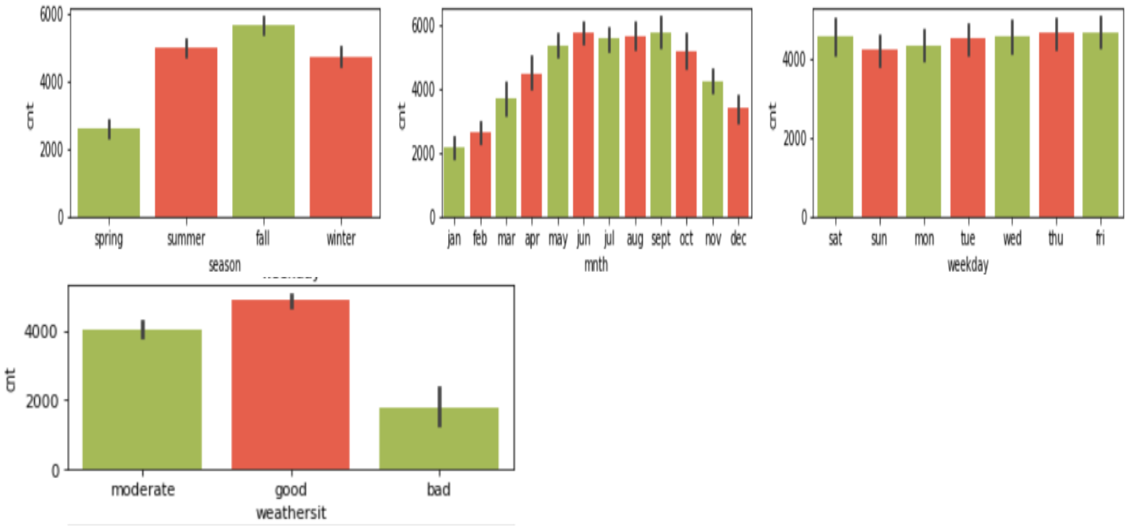
Assignment-based Subjective Questions

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

   Ans:



- Demand is spiked up for every fall and summer season
- Greater demand in 2019 than 2018
- Slight higher when the day is working day
- May to Sept there is increase in demand for bikes
- All days of week show almost similar behaviour on demand of bikes
- When the weather is good the demand is clearly way higher as it is convienent to ride a bike in good weather

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

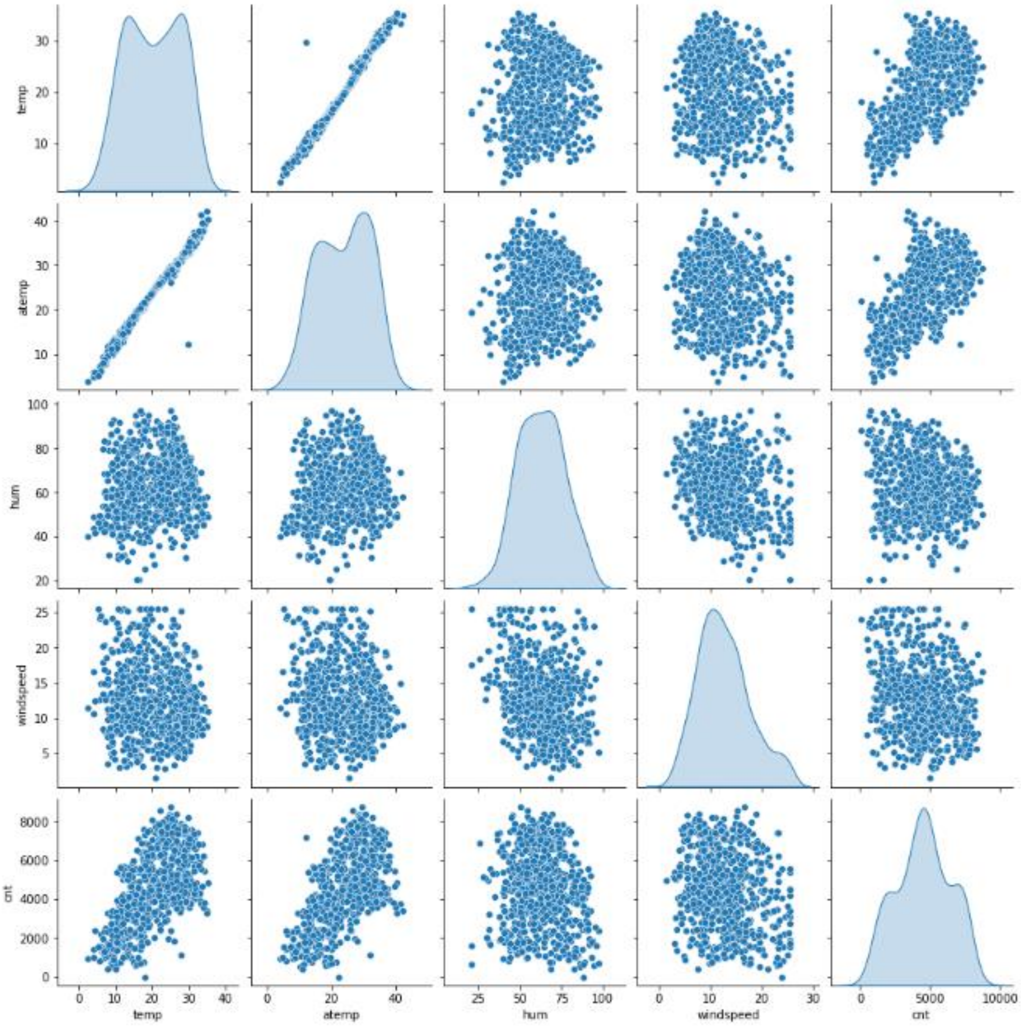
   Ans :

   get_dummies created dummy variables to all categorical variables which increases feature list.We know if there are n values in feature n-1 is sufficient to represent all variation. Most importantly we do drop_first=True this reduces correlation created among dummy variables.

   Suppose we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B , then It is obvious C. So we don't need 3rd variable to identify the C.

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

   Ans :

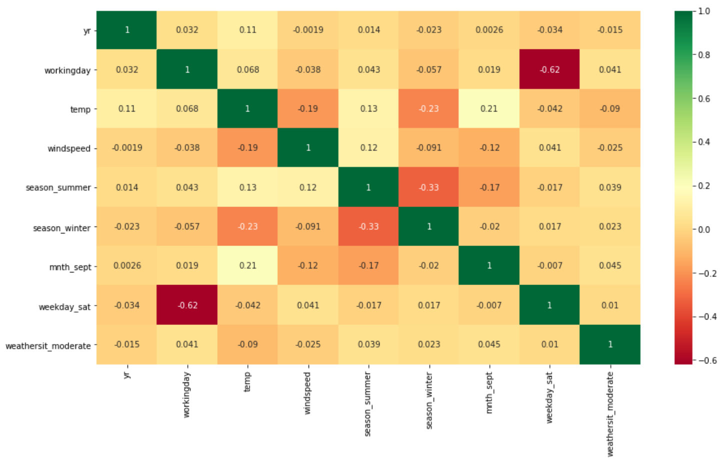


   Clearly shows atemp and temp are most correlated with cnt (target)

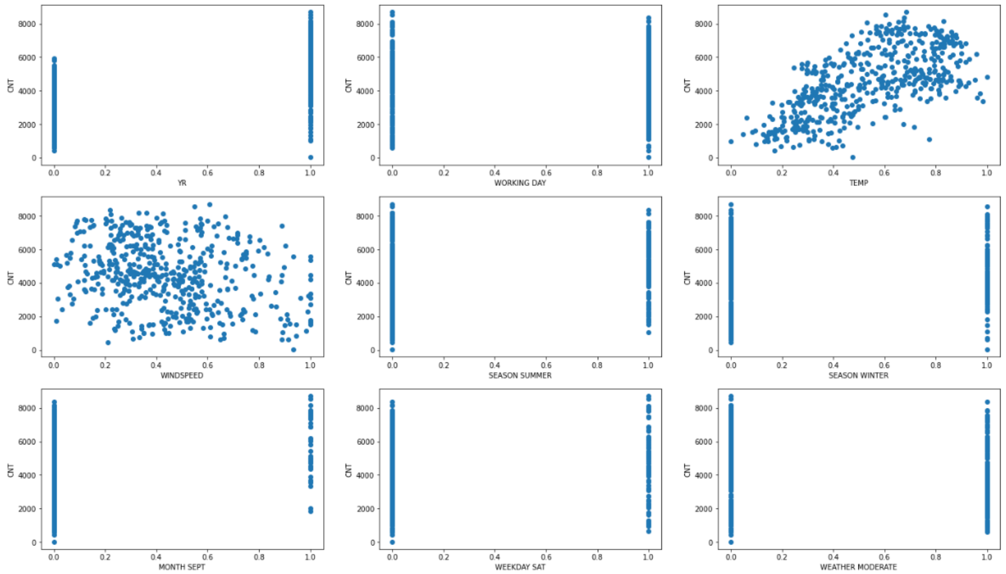
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   Ans :

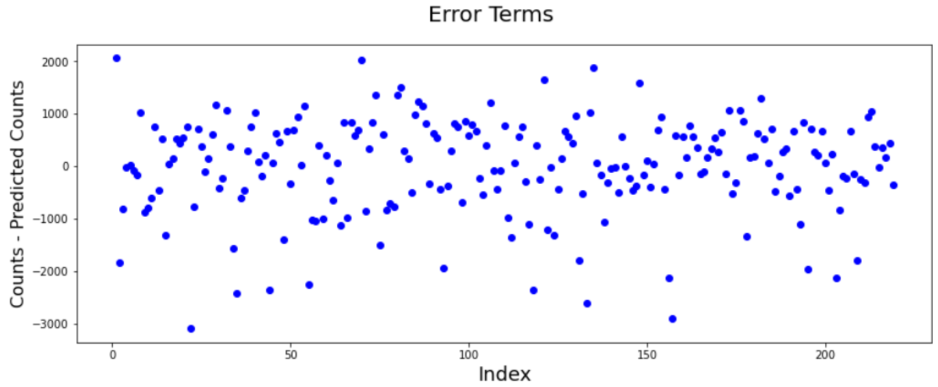
   1. The independent variables are not highly correlated



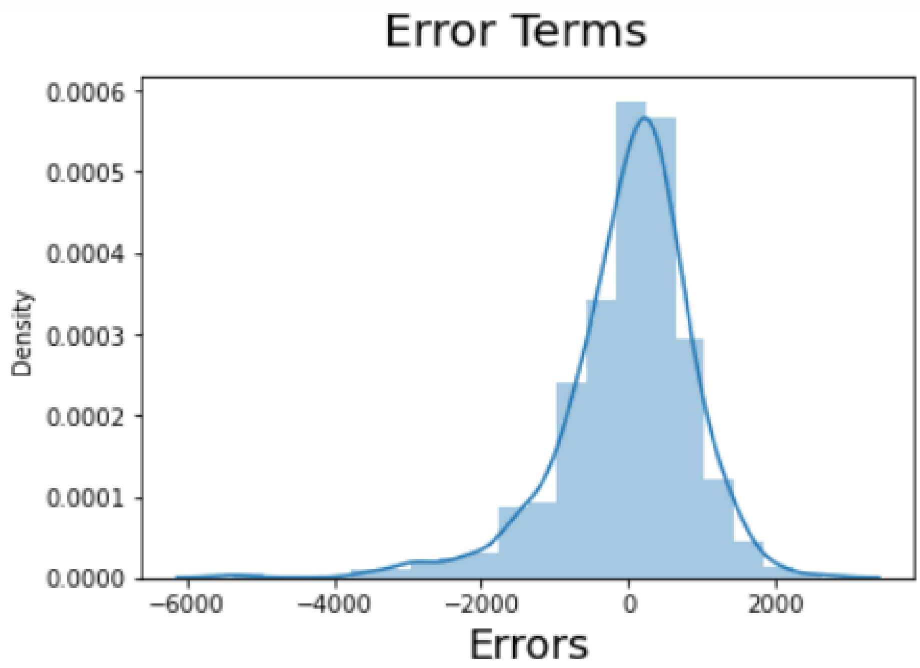
2. Linear relationship between the dependent and independent variables.



3. Variance of the residuals is constant



4. Error terms should be 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)

Ans:

The top 3 influence on the bike booking /demand

- temp : A coefficient value of 4794.5384 indicated that a unit increase in temp variable increases the bike demand by same number
- yr : A coefficient value of 2075.5328 indicated that a unit increase in yr variable increases the bike demand by same number
- season_winter : A coefficient value of 1022.4238 indicated that a unit increase in yr variable increases the bike demand by same number

General Subjective Questions

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

Ans:

Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.

The value of independent variables change (decrease/increase) or generally effect the dependent variable (decrease /increase)

Mathematically the relationship can be represented with the help of following equation –

Y = mX + c

Here, Y is the dependent variable – The predict variable

X is the independent variable – using which we make prediction / in general what influences dependent y – influencer variable .

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept.

A linear relationship will be called positive if both independent and dependent variable increases. – Positive relation

A linear relationship will be called Negative if both independent and dependent variable decreases. – Negative relation

Assumptions:

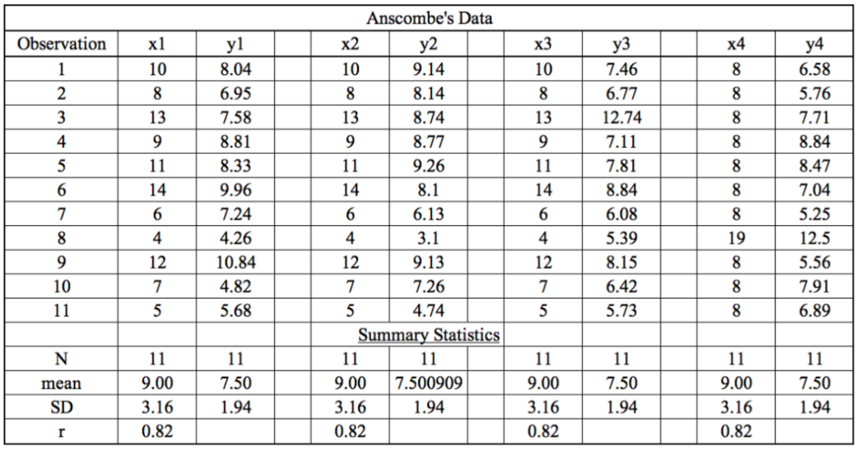
- Error terms should be normally distributed
- Variance of the residuals is constant
- The independent variables are not highly correlated
- Linear relationship between the dependent and independent variables.

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

Ans:

Anscombe's Quartet - developed by statistician Francis Anscombe.

- Has four datasets, each containing eleven (x, y) pairs.
- The datasets share the same descriptive statistics.



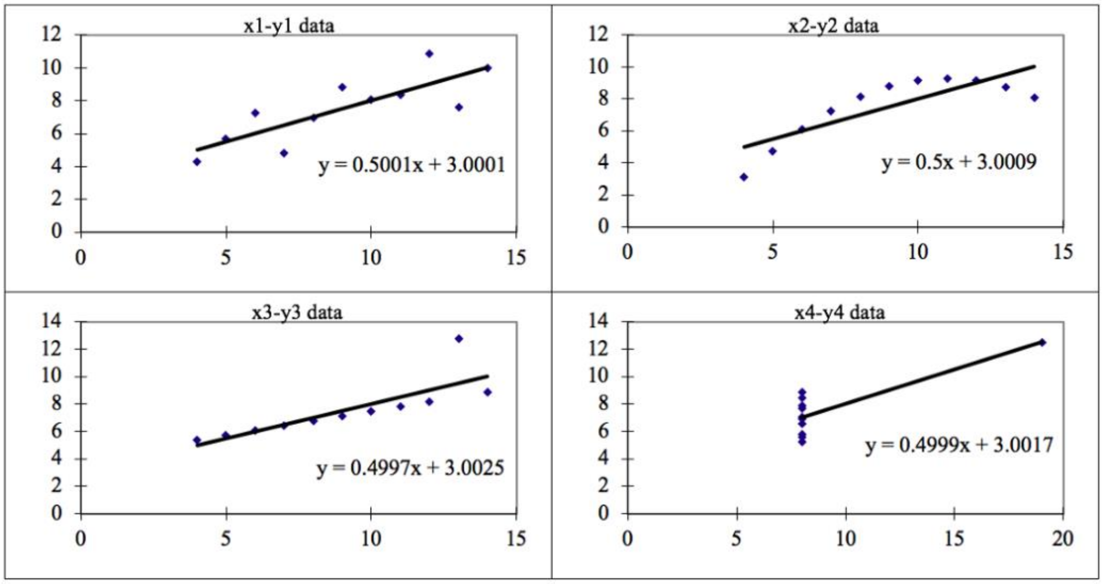
Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

Note : descriptive statistics is identical across dataset

Next step,
When we get the regression line plot against data , it shows that regression line obtained doesn't fit all the datasets at all

Not interpretable by any regression algorithm



Observation:

1. **Dataset 1:** this **fits** the linear regression model pretty well.

2. **Dataset 2:** this **could not fit** linear regression model on the data quite well as the data is non-linear.

3. **Dataset 3:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

4. **Dataset 4:** shows the **outliers** involved in the dataset which **cannot be handled** by linear regression model

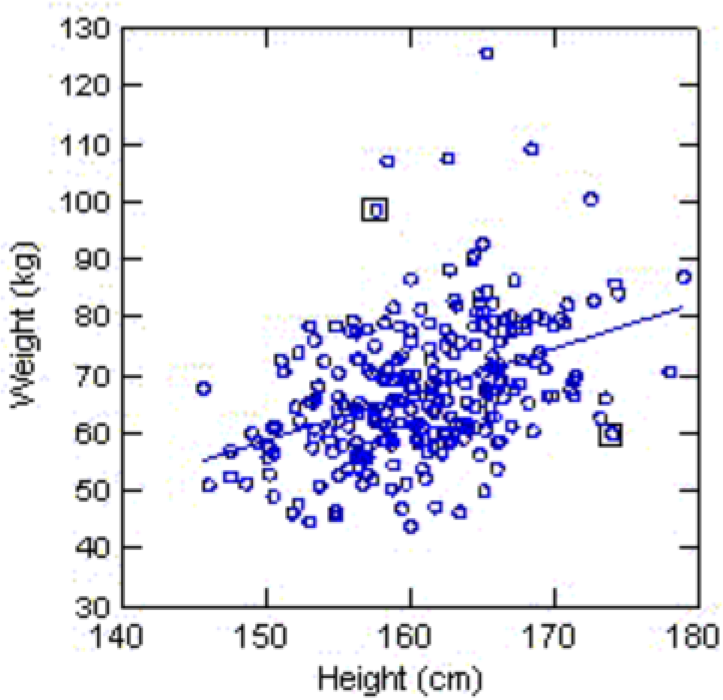
This quartet says **about importance of visualization in Data Analysis**. See and observing data reveals a lot of the structure and a clear picture of the dataset.

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

Ans:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

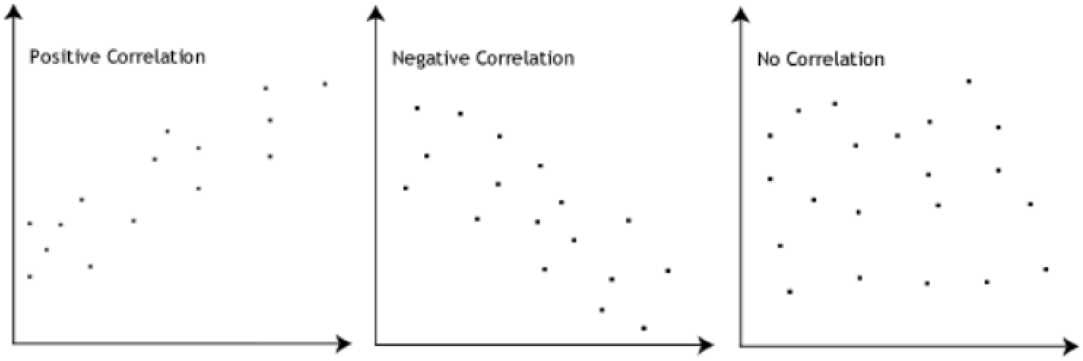
"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows. The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate



The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

- A value of 0 indicates that there is no association between the two variables.
- A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
- A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

This is shown in the diagram below:



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

Ans:

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

For eg : Suppose you have column distance_school_home = 150m for few and for few distance_school_home =50km in same dataset as the kids were given a form where they fill distance in (meter,km).

Now if you don't feature scale then 150m will be consider greater than 50km, which is wrong. This lowers accuracy of model. Hence its important to bring all values to same magnitude for better comparison.

Normalised Scaling	Standardised Scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range
It is really affected by outliers.	It is much less affected by outliers.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

Ans:

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).

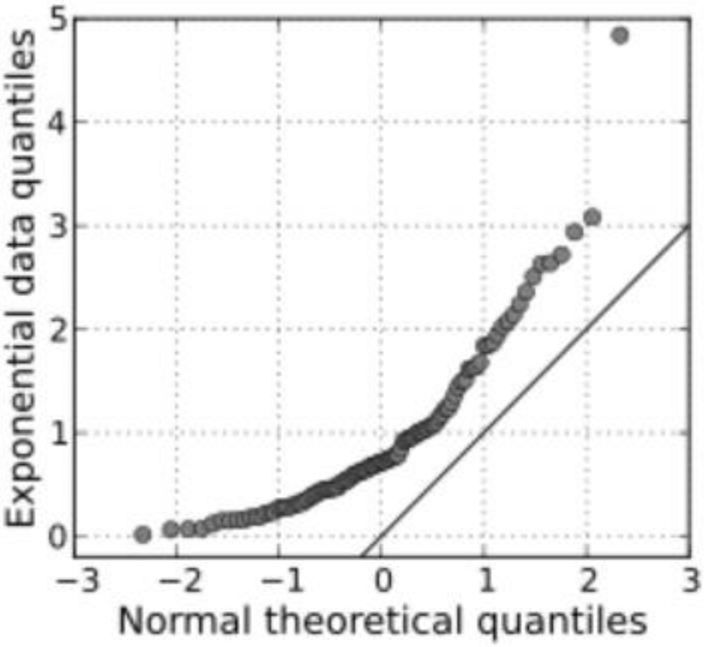
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

Where this is used:

Q-Q plot can also be used to test distribution amongst 2 different datasets.

For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.