UNIT II

CHAPTER 2

Basic Data Analytic Methods

Syllabus Topics

Statistical Methods for Evaluation- Hypothesis testing, difference of means, wilcoxon rank–sum test, type 1 type 2 errors, power and sample size, ANNOVA. Advanced Analytical Theory and Methods: Clustering- Overview, K means- Use cases, Overview of methods, determining number of clusters, diagnostics, reasons to choose and cautions.

Syllabus Topic : Statistical Methods for Evaluation

2.1 Statistical Methods for Evaluation

Q. 2.1.1 Explain statistical methods for evaluation. (Refer section 2.1) (8 Marks)

- During the Data analytic lifecycle the statistics are very important as compared to visualization because the statistics are used in the whole lifecycle whereas visualization is used only for exploring and presenting data.
- The statistical methods are used in the early phase of data exploration and data preparation, model building, estimation of the final models, and calculation of how the new models progress the circumstances when deployed in the field.
- In precise, with the help of statistics following questions for data analytics can be resolved:
  1. Model Constructing and Planning
  2. What are the best input variables for the model?
  3. Can the model forecast the result for the given input?

Model Evaluation

1. Is the model perfect?
2. Does the model implement superior than an obvious guess?
3. Does the model implement superior than another candidate model?

Model Deployment

1. Is the prediction sound?
2. Does the model have the expected influence?

Syllabus Topic : Hypothesis Testing

2.1.1 Hypothesis Testing

Q. 2.1.2 Explain hypothesis testing in brief. (Refer section 2.1.1) (4 Marks)

- **Hypothesis testing** or **significance testing** is a method for testing an assertion or assumption about a parameter in a population, with the help of data measured in a sample.
- In this technique, we check some hypothesis by determining the probability that a sample statistic could have been selected, if the hypothesis regarding the population parameter were true.

- The hypothesis testing is nothing but forming a claim and check it with data.
- In this technique we are going to check the results either they are valid or not by considering the odd results which are generated during the testing.
- When carrying out hypothesis tests, the common hypothesis is that there is no difference among two samples.
- This hypothesis is used as the default position for constructing the test or conducting a scientific experiment. Statisticians called this hypothesis as null hypothesis (H_0).
- Another hypothesis is alternative hypothesis (H) which contains the difference between two samples.
- Consider an example of recognizing the teaching effect of Teacher T1 and Teacher T2 on students.
- The corresponding null hypothesis and alternative hypothesis will be as follow:
  - **H0 :** Teaching of Teacher T1 and Teacher T2 have the same effect on the students.
  - **HA :** Teaching of Teacher T1 has a better effect than Teacher T2 on the students.
- For performing hypothesis test it is required to describe the null and alternative hypothesis.
- This is because hypothesis test is depending upon the accepting or rejecting the null and alternative hypothesis against each other.
- The following table shows the example of null and alternative hypotheses that should be answered throughout the analytic lifecycle.

Application	Null Hypothesis	Alternative Hypothesis
Accuracy Forecast	Model A does not forecast better than the present model.	Model A predicts better than the present Model.
Recommendation Engine	Algorithm B does not creates better recommendations than the existing algorithm being used.	Algorithm B creates better recommendations than the existing algorithm being used.
Regression Modeling	This variable does not affect the outcome because its coefficient is zero.	This variable affects outcome because its coefficient is not zero.

- When a model is constructed over the training data, it desired to be estimated over the testing data to see if the proposed model forecast is improved than the existing model presently being used.
- The null hypothesis is used to indicate that the existing model is better than the proposed model whereas the alternative hypothesis indicates vice versa.
- In accuracy forecast, the result of any entity is same for previous and upcoming operations. The hypothesis test desires to assess if the proposed model offers a better prediction.
- In the recommendation engine, the null hypothesis indicates that the current algorithm crates better recommendation whereas the alternative hypothesis indicates that the new algorithm creates better recommendation.
- The regression analysis asks for the regression coefficient for a variable for hypothesis test.
- If the coefficient is zero then it states that the variable does not have influence on the result however.

Syllabus Topic : Difference of Means

2.1.2 Difference of Means

Q. 2.1.3 Write note on difference of means. (Refer section 2.1.2) **(8 Marks)**

- Hypothesis testing is a common method to draw implications on whether or not the two populations, indicated by P1 and P2, are different from each other.
- Consider a simple example in which we are taking two samples which are arbitrarily drawn from both populations for performing the hypothesis test for finding their means.

- For this hypothesis test the null and alternative hypothesis is given below:

$H_0: \mu_1 = \mu_2$ $H_A: \mu_1 \neq \mu_2$

Here the μ_1 and μ_2 represents the means for population P1 and P2 correspondingly.

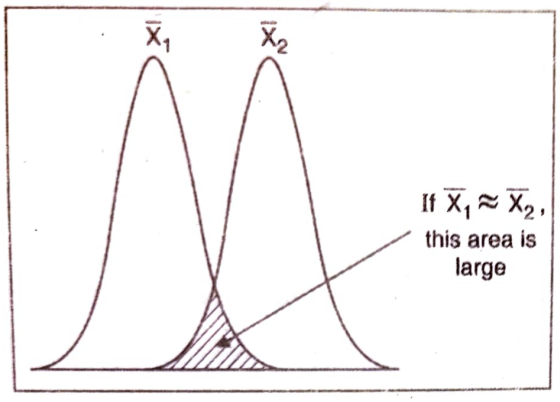

Fig. 2.1.1 : Overlap of the two distributions is larger if $\bar{X}_1 \approx \bar{X}_2$

- In this diagram the two samples X_1 and X_2 are compared with each other. If the values of them are almost same then they overlap and follow the null hypothesis.
- Whereas if they there is difference between their values then they do not follow the null hypothesis and take the help of Student's t-test or the Welch's t-test for testing.

2.1.2(A) Student's t-test

Q. 2.1.4 What is student's t-test ? (Refer section 2.1.2(A)) **(4 Marks)**

- The t-test can be defined as it is a kind of hypothesis test which can be executed by following the t-distribution underneath null hypothesis.
- When the scaling terms of two populations are known and they are following the normal distribution then the t-test are applied however if the scaling terms of two populations are unknown the student's t-test cannot be applied.
- Assume S1 and S2 samples are arbitrarily and independently chosen from two populations, P1 and P2, correspondingly.
- If every population is normally distributed with the same mean (M1 = M2) and with the same scaling term, then T given in following equation follows a t-distribution with S + S2 - 2 degrees of freedom (df).

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S_p^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}$$

- The shaping of t-distribution looks like the same way as the normal distribution.
- Actually, as the degrees of freedom approaches thirty or more, the t-distribution is approximately identical to the normal distribution.
- Since the numerator of T is the difference of the sample means, if the examined value of T is far away from zero such that the probability of observing such a value of T is one would reject the null hypothesis that the population means are equal.
- Therefore, for a small probability, such as $\alpha = 0.05$, T* is calculated such that $P(|T| \geq T^*) = 0.05$.
- When all the samples are gathered and the observed value of T is estimated as per the above formula the null hypothesis ($\mu_1 = \mu_2$) is rejected if $|T| \geq T^*$.
- During the hypothesis testing, the smallest probability says "n" is called as the significance level of the test.
- The significance level which is carrying out throughout the test is nothing but the possibilities of rejecting the null hypothesis when they are true in reality.
- This can be illustrated in simple way as for n = 0.05, if the means from the two populations are actually equal, then in constant arbitrary sampling, the observed magnitude of T would go above T* 5% of the time.
- A main supposition while making the use of Student's t-test is that the population variances are equal.
- In the prior example, the t .test ( ) function call contains var.equal = TRUE to state that equality of the variances should be supposed.
- If that supposition is not suitable, then Welch's t-test should be used.

2.1.2(B) Welch's t-test

Q. 2.1.5 What is Welch'st-test? (Refer section 2.1.2(B)) **(2 Marks)**

- Justification of equal population variance assumption is not satisfied after performing Student's t-test for the difference of means then Welch's t-test can be used based on T as shown in following formula.

$$T_{welch} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Where, X_i = sample mean

S_i^2 = sample variance

And n_i = sample size.

- Observe that Welch's t-test makes the use of the sample variance S^2_i for every population rather than the pooled sample variance.
- In Welch's test, under the remaining suppositions of arbitrary samples from two normal populations with the same mean, the distribution of T is estimated with the help of the t-distribution.

Syllabus Topic : Wilcoxon Rank-Sum Test

2.1.2(C) Wilcoxon Rank-Sum Test

Q. 2.1.6 Explain Wilcoxon rank-sum test. (Refer section 2.1.2(C)) **(4 Marks)**

- At-test signifies a parametric test where it creates suppositions about the population distributions from which the samples are drawn.
- Whereas the nonparametric test is used when the populations are not supposed to follow a normal distribution.
- The Wilcoxon rank-sum test is a kind of nonparametric hypothesis test that validates either the two populations is identically distributed or not.
- It is required to assume that the arrangements of sampled observation in one population are consistently mixed with each other while supposing that the two populations are identically distributed.
- For instance, in ordering the observations, one would not expect to see a large number of observations from one population clustered together, specifically at the start or the end of ordering.
- Consider the two populations P1 and P2, with independently random samples of size S1 and S2 correspondingly.
- The total number of observations is

$$N = S1 + S2$$

- The first step of the Wilcoxon test is to give ranks to the set of observations from the two clusters as if they are originated from one big cluster.
- The ranks can be given to observations are as the largest observation gets the last rank of total ranks observation whereas the smallest observation gets the first ranks of it.
- If the two observations are similar then they gets the number that is average of the total number observations.
- The test uses ranks rather than numerical outcomes to avoid particular expectations about the shape of the distribution.
- Once giving ranks to all the observations, the assigned numbers are summed for at least one population's sample.
- If the distribution of P1 is shifted to the right of the other distribution, the rank-sum equivalent to P1's sample should be larger than the rank-sum of P2.
- The Wilcoxon rank-sum test controls the implication of the observed rank-sums.
- The below R code executes the test on the same dataset used for the previoust-test.
- Wilcox.test(x, y, conf.int TRUE)

```
Wilcoxon rank sum test
Data x and y
W=55, Probability_val = 0.04903
```

```
Alternative hypothesis: true location shift is ≠0
95 percent confidential interval:
-6.2596774 -0.1240618
Sample estimates:
Difference in location
-3.417658
```

- The wilcox. test() function ranks the observations, decides the relevant rank-sums according to every population's sample, and then decides the probability of such rank-sums of such magnitude being examined supposing that the population distributions are identical.
- In the above given example, the probability is given by the Probability_val of 0.04903.
- Therefore, the null hypothesis would be rejected at a 0.05 significance level.
- The reader is warned in contradiction of interpreting that one hypothesis test is undoubtedly improved than another test based exclusively on the examples given in this section.
- Since the Wilcoxon test does not suppose anything about the population distribution, it is usually considered more vigorous than the t-test.
- In other words, there are smaller numbers of suppositions to disrupt.
- On the other hand, when it is reasonable to suppose that the data is normally distributed, Student's or Welch's t-test is a suitable hypothesis test to consider.

Syllabus Topic : Type I and Type II Errors

2.1.3 Type I and Type II Errors

Q. 2.1.7 Explain type I and type II errors in detail. (Refer section 2.1.3) **(4 Marks)**

- Depending upon the acceptance or rejection of null hypothesis there are two types of errors produced during the test.
- These errors are called as type I and type II errors.
- The type I error is occurred when the null hypothesis is true however it is rejected.
- The probability of this kind of error is indicated with the help of Greek letter α.
- In other words, a type I error is to falsely assume the presence of something that is not there.
- The type II error is occurred when the null hypothesis is false however it is accepted.
- The probability of this kind of error is indicated with the help of Greek letter β.
- In other words a type II error is to falsely assume the non-existence of something that is present.
- The following table shows the four possible states of a hypothesis test, containing the two types of errors.

Table 2.1.1 : Type I and Type II Error

	H_o is true	H_A is true
H_o is accepted	Correct outcome	Type II Error
H_o is accepted	Type I Error	Correct outcome

- The significance level, as stated in the Student's t-test, is corresponding to the type I error.
- For a significance level for example $\alpha = 0.05$, if the null hypothesis ($\mu_1 = \mu_2$) is TRUE, there is a five percent chance that the observed T value based on the sample data will be big sufficient to reject the null hypothesis.
- By choosing a suitable significance level, the possibility of promising a type I error can be defined prior of any data is gathered or analyzed.
- The possibility of promising a Type II error is somewhat harder to decide.
- If two population means are truly not equal, the possibility of promising a type II error will depend on how far apart the means truly are.
- To decrease the probability of a type II error to a reasonable level, it is frequently required to raise the sample size.

Syllabus Topic: Power and Sample Size

2.1.4 Power and Sample Size

Q. 2.1.8 Write a note on: Power and sample size. (Refer section 2.1.4) **(4 Marks)**

- The power of a test can be defined as a possibility of rejecting the null hypothesis.
- It is signified with the help of 1- β, where β is the possibility of a type II error.
- When the sample size increases, the power of test is also improves.
- The power is used to decide the required sample size.
- When we concern about the difference of means, the power of a hypothesis test is based on the true difference of the population means.
- In other words, for a fixed significance level, a greater sample size is mandatory to discover a minor difference in the means.
- Universally, the degree of the difference is called as the effect size.
- As the sample size grow more and more, it is easier to discover a given effect size δ as demonstrated in Fig. 2.1.2.

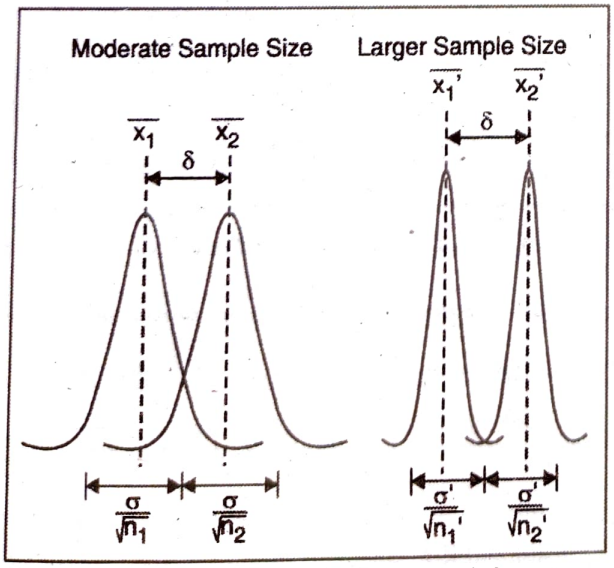

Fig. 2.1.2 : A large size better recognizes a fixed effect size

- With a huge sufficient sample size, almost any effect size can look statistically important.
- On the other hand, a very minor effect size may be unusable in a practical point of view.
- It is vital to consider a suitable effect size for the problem in hand.

Syllabus Topic : ANNOVA

2.1.5 ANNOVA

Q. 2.1.9 What is ANNOVA? Explain with example. (Refer section 2.1.5) **(8 Marks)**

- Till now we observed that the hypothesis tests discussed so far are good for examining means among two populations.
- Now the situation is that if there are more than two populations exist then what actions need to take?
- Let's take an example of testing the selectiveness of people while shopping on 5 different candidates who are between the age of 10 and 80.
- Here the goal is to decide which person's selection criterion is most efficient.
  1. Person 1 only chooses white color clothes.
  2. Person 2 only chooses less priced clothes.
  3. Person 3 chooses white color and good quality cloths.
  4. Person 4 chooses less priced and good quality cloths.
  5. Person 5 chooses white color, good quality and less priced cloths.
- To perform the test numbers of t-test have to be applied on every pair of selection criterion.
- In the above example, the person 1 is compared with the person 2, 3, 4, or 5.
- In the same way, the person 2 is compared with that of the next 3 groups.
- Hence, a total of 10 t-tests would be performed.
- On the other hand, numerous t-tests may not perform well on number of populations for following two reasons:
  1. The number of t-tests grows as the number of groups grows; analysis with the help of several t-tests becomes more problematic.

2. Performing various analyses, the possibility of promising at least one type I error somewhere in the analysis significantly increases.

- To solve these problems an Analysis of Variance (ANNOVA) is designed.
- ANNOVA is a generalization of the hypothesis testing which perform the testing of difference of two population means.
- ANNOVA tests if any of the population means vary from the other population means.
- In ANNOVA the null hypothesis is considered as all population means are equal whereas alternative hypothesis is considered as at least one pair of the population means is not equal.
- In other words,

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_n$$

$$H_A: \mu_i \neq \mu_j \text{ for at least one pair of i, j}$$

- As we know in the Difference of Means every population is supposed to be normally spread with the same variance.
- The main thing to compute for the ANNOVA is the test statistic.
- Fundamentally, the aim is to test whether the clusters made by every population are more tightly grouped than the spread across all the populations.
- Now consider the total number of populations be k. The total number of samples N is arbitrarily divided into k groups.
- The number of samples in the i^{th} group is signified as n_i, and the mean of the group is X_1 where i€[1,k].
- The mean of all the samples is represented as X_0
- **The between-groups mean sum of square** (S_B^2) is an approximation of the between-groups variance.
- It measures how the population means differ as per the grand mean or the mean distributed across all the populations.
- This can be represented with the help of following formula.

$$S_B^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i \cdot (\overline{X}_1 - \overline{X}_0)^2$$

- **The within-group mean sum of squares,** (S_W^2), is an approximation of the within-group variance.
- It calculates the spread of values inside groups. This can be represented with the help of following formula.

$$S_W^2 = \frac{1}{n-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$$

- If (S_B^2) is considerably bigger than (S_W^2) then some of the population means are different from each other.
- The ratio of the between-groups mean sum of squares and the within group mean sum of squares is called as F-test statistic.
- This can be represented with the help of following formula.

$$F = \frac{S_B^2}{S_W^2}$$

- The F-test statistic in ANNOVA can be supposed as a measure of how different the means are comparative to the inconsistency inside every group.
- The greater the observed F-test statistic, superior the probability that the differences among the means are because of something other than chance alone.
- The F-test statistic is used to test the hypothesis that the observed effects are not due to chance-that is, if the means are significantly different from one another.

Syllabus Topic : Advanced Analytical Theory and Methods

2.2 Advanced Analytical Theory and Methods

- In this sub-section we are going to study about the cluster and k-means.
- Now first we see what is mean by cluster and then k-means, use cases, methods as well as diagnostics and reasons to choose.

Syllabus Topic : Overview of Clustering

2.2.1 Overview of Clustering

Q. 2.2.1 What is mean by clustering? (Refer section 2.2.1) (2 Marks)

- Basically clustering is nothing but the grouping of things.
- Clustering is defined as grouping of same kind of objects which are gathered by the use of unsupervised method.
- When we concern with machine learning, the term unsupervised indicates the problem of discovering unknown structure inside the unlabeled data.
- Here unsupervised in Clustering methods says that the data scientist does not decide the labels earlier to apply to the clusters.
- The structure of the data defines the objects of interest and decides how better is to group the objects.
- Consider an example, based on customers' choices in the cloths; it is open to split the customers into multiple groups depending on randomly selected values.
- The customers could be split into following groups:
  - Customer that chooses kids wear
  - Customer that chooses women's wear
  - Customer that chooses men's wear
- In this case, the wears were selected somewhat individually based on easy-to-communicate points of description.
- On the other hand, such groupings do not specify a natural similarity of the customers inside every group.
- This means the persons choosing kids wear are not behaving different than persons who choose women's or men's wear.
- As extra dimensions are bring together by inserting additional variables about the customers, the job of discovering meaningful groupings becomes more difficult.
- For example, assume variables like size, color, price, type of material were considered together with the wear variable.
- What are the ordinary arising groupings of customers? This is the kind of question that clustering analysis can help answer.
- Clustering is a technique frequently used for investigative analysis of the data.
- In clustering, there are no forecasting can be made.
- Somewhat, clustering techniques discover the resemblances among objects corresponding to the object attributes and group the same kind of objects into clusters.
- Clustering techniques are used in marketing, economics, and a number of branches of science.
- A common clustering method is k-means.

Syllabus Topic : K-Means

2.2.2 K- Means

Q. 2.2.2 Define K-Means. (Refer section 2.2.2) (2 Marks)

- Given a group of objects all with n computable attributes, *k-means* is an analytical method that, for a selected value of k, recognizes k clusters of objects based on the objects closeness to the center of the k groups.
- The center is estimated by the arithmetic average of every cluster's n-dimensional vector of attributes.
- This section describes the algorithm to decide the k means along with how better to apply this technique to number of use cases.
- Fig. 2.2.1 demonstrates 3 clusters of objects having 2 attributes.
- In Fig. 2.2.1 the small dots are the object color-coded to the nearest big dot which indicates the cluster.

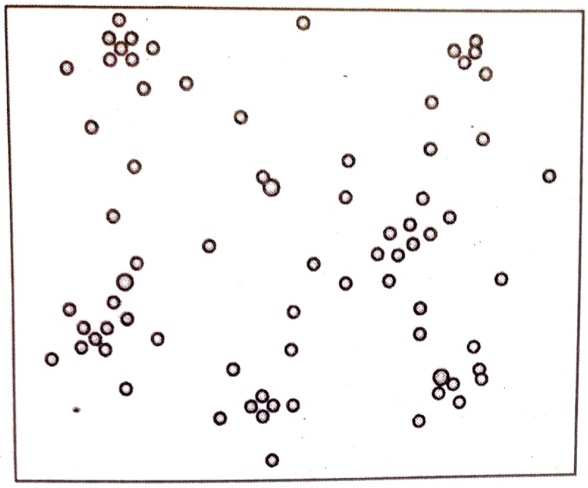

Fig. 2.2.1

Syllabus Topic : Use Cases

2.2.3 Use Cases

Q. 2.2.3 Write applications of K-Means. (Refer section 2.2.3) **(4 Marks)**

- Clustering is frequently used as an opener to classification.
- As soon as the clusters are recognized, labels can be applied to each and every cluster to organize every group based on its characteristics.
- Clustering is mainly an investigative method to find out the unknown structures of the data, maybe as an introduction to further focused analysis or decision processes.
- Following are the **applications** of k means:

Image Processing

- One of the best example of increasing volumes of unstructured data being gathered is video.
- K-means analysis is used for recognizing the object present in the video.
- This can be done for each and every frame of video.
- For every frame, the job is to decide which pixels are most related to each other.
- Each pixel consists of various attributes such as brightness, color, location, and the x and y coordinates in the frame.
- In the video images having some kind of security, the sequential frames are analyzed to find out is there any changes to the clusters.
- These freshly recognized clusters may show unauthorized access to a facility.

Medical

- Each patient has the number of attributes like height, weight, age, blood pressures, cholesterol level, sugar level which can recognize naturally occurring clusters.
- With the help of such clusters it is easy to target persons for particular anticipatory measures or clinical experimental involvement.
- Like a medical field, a clustering is also helpful in biology for the organization of plants and animals along with the field of human genetics.

Customer Segmentation

- Marketing and sales groups use k-means to recognize customers who have same kind of behaviors and payment patterns.
- Consider an example of a wireless provider who sees the customer attributes such as bill on month basis, the count of text messages, data volume used up, minutes used in several everyday periods, and years as a customer.
- The wireless company could then explore the naturally arising clusters and consider strategies to raise sales or decrease the customer *churn rate,* the proportion of customers who end their connection with a specific company.

Syllabus Topic : Overview of the Methods

2.2.4 Overview of the Method

Q. 2.2.4 Define Centroid. (Refer section 2.2.4) **(2 Marks)**

Q. 2.2.5 Explain K-Means algorithm. (Refer section 2.2.4) **(4 Marks)**

- To demonstrates the method to discover k clusters from a group of M objects having n attributes, the two dimensional case (n = 2) is analyzed.
- The reason behind the use of two dimensional is that the k-means method is very easy to visualize.
- In the following example we are going to use two attributes of every object so we have to use two point x and y and i =1, 2...M.
- For a given cluster of m points (m ~ M), the point that relates to the cluster's mean is known as *centroid.*
- The term centroid is referred as a point that relates to the center of mass for an object in mathematic field.
- The following steps defining the k-means algorithm to discover k clusters:

1. First step in k-means algorithm is to select the value of k and the k primary suppositions for the centroids

- In our example, k = 3, and the initial centroids are specified with the help of points which carry the colors red, green and blue in Fig. 2.2.2.

Fig. 2.2.2 : Initial starting points for the centroids

2. Calculate the distance from every data Point (x_i, y_i) to all centroid. Allocate every point to the nearby centroid.

- This relationship describes the first k clusters.
- In two dimensions, the distance, *d,* among any two points, (x1, y1) and (x2, y2), in the Cartesian plane is normally stated with the help of Euclidean distance measure given as below:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

3. Calculate the centroid, the center of mass, of every freshly described cluster which was resulted from Step 2.

- The following formula is used to calculate the centroid (x_c, y_c) of points in a k-means cluster in two dimensions.

$$(x_c, y_c) = \left(\frac{\sum_{i=1}^{m} x_1}{m}, \frac{\sum_{j=1}^{m} y_1}{m} \right)$$

- Therefore, (x_c, y_c) is the ordered pair which contains the arithmetic means of points in the cluster.
- During step 3, a centroid is calculated for every k clusters.

4. Repeat Steps 2 and 3 till the algorithm meets to an answer.

(a) Allocate every point to the nearby centroid calculated during Step 3.

(b) Calculate the centroid of recently described clusters.

(c) Repeat until the algorithm reaches the desired answer.

- Convergence is reached when the calculated centroids do not alter or the centroids and the allocated points are in decisive backward and forward from one iteration to the next iteration.
- *To* simplify the previous algorithm to n dimensions, assume there are M objects, where every object is defined with the help of n attributes or property values $(p_1, P_2 \bullet\bullet\bullet p_n)$. Then object *i* is defined with the help of $(p_{i1}, P_{i2} \bullet\bullet\bullet p_{in})$ for *i* = 1, 2... M.
- To enlarge the previous process to discover the k clusters from two dimensions to n dimensions, the following equations offer the formulas for computing the distances and the locations of the centroids for n≥1.

- For a given point, p_i, *at* (p_{i1}, P_{i2}, ••• p_{in}) and a centroid, q, positioned at (q_1,q_2, .. • q_n) the distance, *d*, among p_i and *q*, is given in following formula:

$$d(p_i,q) = \sqrt{\sum_{j=1}^{n} (p_{ij} - q_i)^2}$$

- The centroid, *q*, of a cluster of m points(p_{i1}, P_{i2}, ••• p_{in}) is computed as given below:

$$(q_1, q_2, \ldots, q_a) = \left(\frac{\sum_{i=1}^{m} Pi_1}{m}, \frac{\sum_{i=1}^{m} Pi_2}{m}, \ldots, \frac{\sum_{i=1}^{m} Pi_n}{m} \right)$$

Syllabus Topic : Determining the Number of Clusters

2.2.5 Determining the Number of Clusters

Q. 2.2.6 Explain how to determine the number of clusters ? (Refer section 2.2.5) **(4 Marks)**

- Using the algorithm which is discussed in previous section, k clusters can be recognized in a given dataset, but what value of k should be chosen?
- The value of k can be chosen based on a prediction or some already defined requirements.
- Still, it is good to know how much it is better or worse having k clusters against k - 1 or k + 1 clusters in explaining the structure of the data.
- Afterward, a heuristic with the help of Within Sum of Squares (WSS) metric is observed to decide a reasonably optimal value of k.
- By making the use of **distance function** an WSS is defined as given below:

$$WSS = \sum_{i=1}^{M} d(p_i,q^{(i)})^2$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{n} (p_{ij} - q_j^{(i)})^2$$

In simple words, WSS is the sum of the squares of the distances among each and every data point and the centroid which are closest to it.

- The term $q^{(i)}$ specifies the closest centroid that is related with the i^{th} point.
- The WSS is small when the points which are near to the corresponding centroids are small.
- The addition of cluster does not give an extra advantage when k + 1 cluster do not significantly decrease the value of WSS having only k clusters.

Using R to Perform a K-means Analysis

- To demonstrate how to use the WSS to decide a suitable number, k, of clusters, the below example uses R to perform a k-means analysis.
- The job is to group 620 school students based on their grades in three subject areas: Marathi, Geography, and Hindi.
- The following R code takes the essential R libraries and imports the CSV file which contains the grades.

```
library (plyr)
library(ggplot2)
library(cluster)
library(lattice)
library(graphics)
library(grid)
library(gridExtra)
grade_data=
as.data.frame(read.csv("D:/gradedata/grades_km_data.csv"))
```

- The data file which is imported includes the data in tabular format.
- It contains 4 columns such as Student_ID, Marathi, Geography, and Hindi respectively.
- As the student_ID is not used in the clustering analysis, it is omitted from the k-means input matrix, *kmdata*.

```
kmdata_original = as. matrix (grade_data [, c ("Student"
,"Marathi ",
"Geography " ," Hindi ")])
kmdata<- kmdata_original[,2:4]
kmdata[1 :10,)
```

	Marathi	Geography	Hindi
[1,]	99	96	97
[2,]	99	96	97
[3,]	98	97	97
[4,]	95	100	95
[5,]	95	96	96
[6,]	96	97	96
[7,]	100	96	97
[8,]	95	98	98
[9,]	98	96	96
[10,]	99	99	95

- To decide an suitable value for k, the k-means algorithm is used to recognize clusters for k = 1, 2, ..., 15.
- For all values of k, the WSS is computed.
- The below R code loops through numerous k-means examines for the number of centroids, k, changing from 1 to 15.
- For every k, the option *n start* = *25*states that the k-means algorithm will be repeated 25 times, all starting with *k* arbitrary initial centroids.
- The wss vector is used to hold the respective value of WSS for every k-mean analysis.

```
wss<- numeric (15)
for (kin 1 :15) wss[k) < - sum (kmeans (kmdata, centers=k,
nstart=25 ) $withinss )
```

- With the help of basic R plot function, every WSS is plotted against the particular number of centroids, 1 through 15.
- This plot is give in Fig. 2.2.3.

```
plot(1 : 15, wss, type=" b ", xlab="Number of Clusters" ,
ylab="Ili t hin Sum ofSquares" )
```

- As can be seen, the WSS is significantly decreased when k increases from one to two.
- Another considerable decrease in WSS happens at k = 3.
- On the other hand, the enhancement in WSS is fairly linear for k > 3.

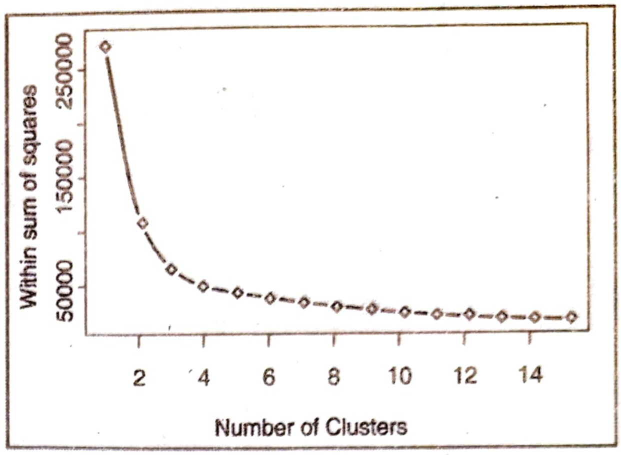

Fig. 2.2.3 : WSS of the student grade data

- Thus, the k-means analysis will be carried out for k = 3.
- The process of recognizing the suitable value of k is called as discovering the "elbow" of the WSS curve.

km = kmeans(kmdata,3, nstart=25) km

- K-means clustering having 3 clusters of sizes 158, 218, 244

Cluster means

```
Marathi Geography Hindi
1 97.21519 93.37342 94.86076
2 73.22018 64.62844 65.84862
3 85.84426 79.68033 81.50820
Clustering vector:
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
[41] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
[81] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
[121] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
3 3 3 3 3 3 3 3 3 3
[161] 3 3 3 3 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 1 1 3 3 1 3 3 3 1 3
3 3 3
3 3 1 3 3 3 3 3 3
[201] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3
333 3 33 3 33 3
```

```
[241] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
333 3 33 3 33 3
[281] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
333 3 33 3 33 3
[321] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
333 3 33 3 33 3
[361] 333 3 33 3 333 3 3 3 3 3 3 2 2 2 2 2 2 2 3 2 3 2 2 2 2
333
2 2 2 2 3 3 2 2 2 2
[401] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2
2 2 2 2 2 2 2 2 2 2
[441] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2
2 2 2 2 2 2 2 2 3 2
[481] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2
2 2 2 2 2 2 2 2 2 2
[521] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2
2 2 2 2 2 2 2 2 2 2
[561] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2
2 2 2 2 2 2 2 2 2 2
[601] 3 32 2 3 33 31 1 3 3 3 2 2 3 2 3 3
```

Within cluster sum of squares by cluster:

```
[1] 6692.589 34806.339 22984.131
(between_SS / total_SS 76.5 %)
```

Available components:

```
[1] "cluster" "centers" "totss"
[6] "betweenss" "size" "iter" "withinss" "ifault" "tot.withinss"
```

- The demonstrated contents of the variable *km* encompasses following things:
  1. The position of the cluster means
  2. A clustering vector that describes the relationship of every student to respective cluster 1, 2, or 3
  3. The WSS of every cluster
  4. A list of all the available k-means components
- The reader can discover details on these components and with the help of k-means in R by using the help facility.
- The reader may have speculated whether the k-means results stored in *km* are equal to the WSS results achieved prior in producing the plot.
- The below check validates that the results are truly equal.

```
c(wss[3] , sum(km$withinss)
[1] 64483.06 64483.06
```

Syllabus Topic : Diagnostics

2.2.6 Diagnostics

Q. 2.2.7 Write a short note on diagnostics. (Refer section 2.2.6) **(4 Marks)**

- The heuristic with the help of WSS can offer at least a number of potential k values to use.
- When the number of attributes is comparatively small, a common method to make more perfect selection of k is to plot the data to decide how distinct the recognized clusters are from all other.
- While doing this the following questions should be taken into account:
- Are the clusters well detached from each other?
- Do any of the clusters have only a small number of points?
- Do any of the centroids look to be too close to each other?
- Now consider the following figure where four clusters are plotted having n=2.
- In the four recognized clusters an enough space is maintained between them.
- The Fig. 2.2.4 shows that clusters may be adjacent to each other, and the difference may not be so understandable.

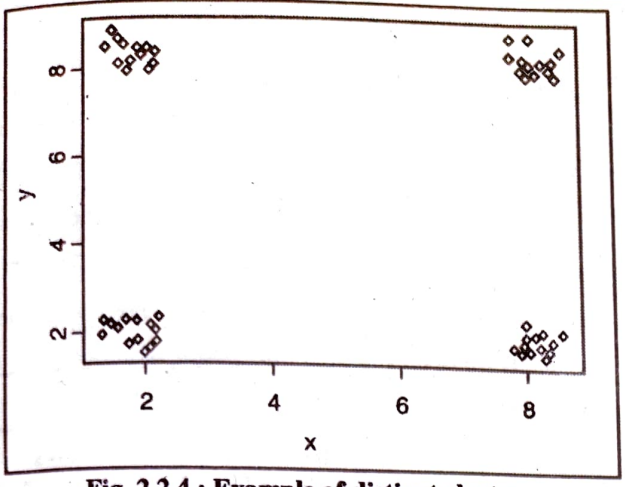

Fig. 2.2.4 : Example of distinct clusters

- In such cases, it is essential to find out that because of using number of clusters it may effects the results or may not.

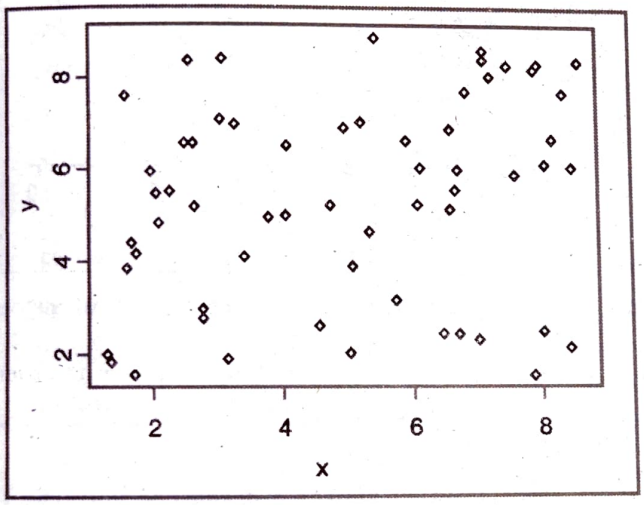

Fig. 2.2.5 : Less obvious cluster

- For instance, following figure makes use of 6 clusters to define the same dataset which is defined in the Fig. 2.2.5.

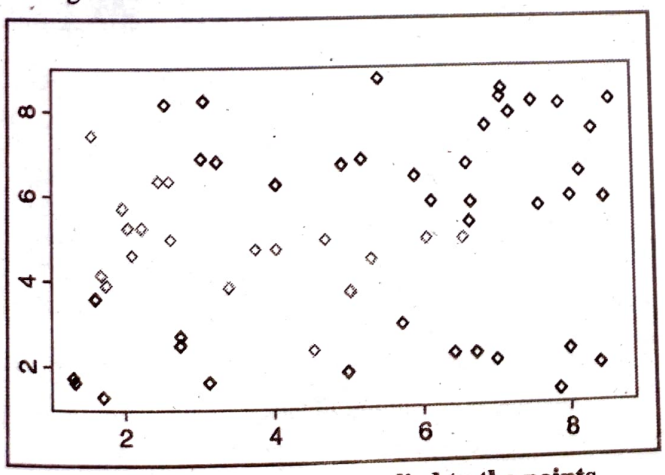

Fig. 2.2.6 : Six clusters applied to the points

- When we use the number of clusters it is sometime not easy to distinguish, whereas using less clusters will results in well distinguished groups.

Syllabus Topic : Reasons to Choose and Cautions

2.2.7 Reasons to Choose and Cautions

Q. 2.2.8 Write a note on :

1) Object attributes
2) Units of measure
3) Rescaling
4) Additional consideration.

(Refer section 2.2.7) **(8 Marks)**

- Up till now we have seen that describing the clusters a K-means is a simple and direct method.
- As soon as clusters and their related centroids are recognized, it is easy to allocate new objects to a cluster which are depend on the object's distance from the nearby centroid.
- As the method is unsupervised, use of k-means assists to remove subjectivity from the analysis.
- There are a number of decisions that the practitioner must make even though k-means is assumed as unsupervised method:
- Which object attributes should be involved in the analysis?
- Which unit of measure should be used for every attribute?
- Do the attributes need to be rescaled so that one attribute does not have an inconsistent effect on the results?
- What other thoughts might apply?

☞ Object Attributes

- Concerning which object attributes to use in the analysis, it is required to know what attributes will be needed during assignment of new object to a cluster.
- Consider an example, information on present customers' fulfilment or purchase occurrence may be available, but such information may not be exists for possible customers.

- The Data Scientist can have the more choices concerning the attributes to use in the clustering analysis.
- Whenever possible and based on the data, it is best to decrease the numerous attributes to the amount possible.
- In addition lots of attributes can reduce the impact of the most significant variables.
- Also, the use of number of same kinds of attributes can place extreme significance on single type of attribute.
- For example, if 5 attributes associated to personal wealth are involved in a clustering analysis, the wealth attributes control the analysis and probably covers the significance of other attributes, like age.
- While dealing with the problem of number of attributes, one valuable method is to recognize any highly associated attributes and use only one or two of the associated attributes in the clustering analysis.
- In the Fig. 2.2.7 a scatter plot matrix is used which is valuable tool to imagine the pair-wise relationships among the attributes.
- The robust relationship is detected to be among Attribute3 and Attribute7.
- If value on any one of these attribute is identified then value of other attribute is identified with near certainty.
- Another option to decrease the numerous attributes is to merge number of attributes into single measure.

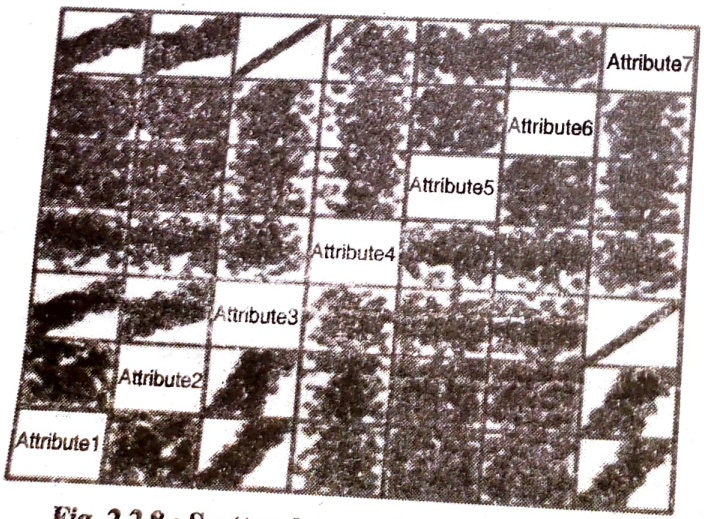

Fig. 2.2.8 : Scatterplot matrix for 7 attributes

Units of Measure

- From a computational point of view, the k-means algorithm is slightly unconcerned to the units of measure for a specified attribute.
- On the other hand, the algorithm will recognize various clusters based on the selection of the units of measure.
- For instance, assume that k-means is used to cluster patients based on age in years and height in centimetres.
- For k = 2, Fig. 2.2.9 demonstrates the two clusters that would be determined for a specified dataset.

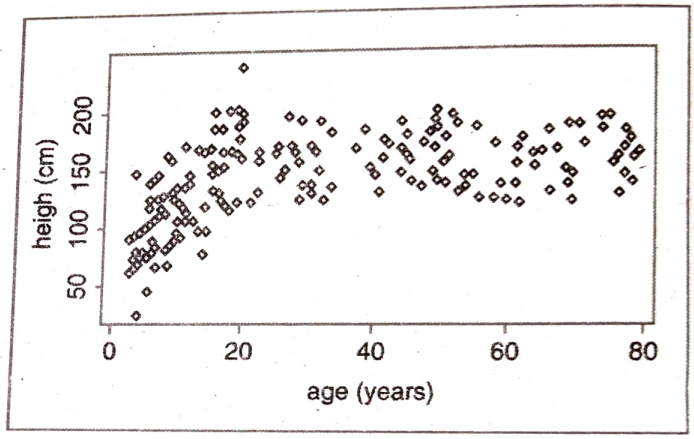

Fig. 2.2.9 : Clusters with height expressed in centimetres

- When the height is rescaled from centimetres to meters the resulting clusters would be somewhat different, as shown in Fig. 2.2.10.

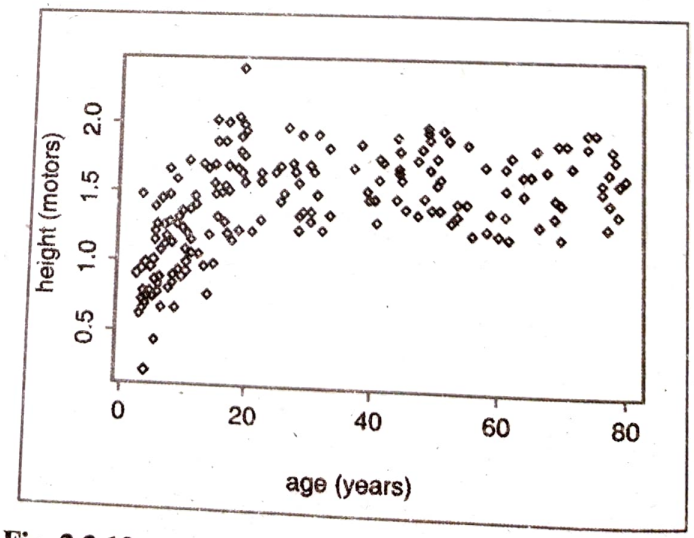

Fig. 2.2.10 : Clusters with height expressed in meters

- When the height is displayed in meters, the magnitude of the ages controls the distance computations among two points.

- The height attribute provides just that much of square among the difference of the maximum height and the minimum height or (2.0 - 0)2 = 4 to the radicand.

Rescaling

- Attributes that are represented in dollars are common in clustering analyses and can vary in magnitude from the other attributes.
- Let's say, if personal income is stated in dollars and age is stated in years, the income attribute, often more than $15,000, can easily control the distance calculation with ages normally below the 100 years.
- While some amendments could be made by saying the income in thousands of dollars, an easiest method is to divide every attribute by the attribute's standard deviation.
- The resultant attributes will have a standard deviation which is = 1 and having no units.
- Now coming back to the age and height example, the standard deviations are 25.2 years and 38.6 cm, correspondingly.
- Dividing every attribute value by the suitable standard deviation and carrying out the k-means analysis yields the result shown in Fig. 2.2.11.

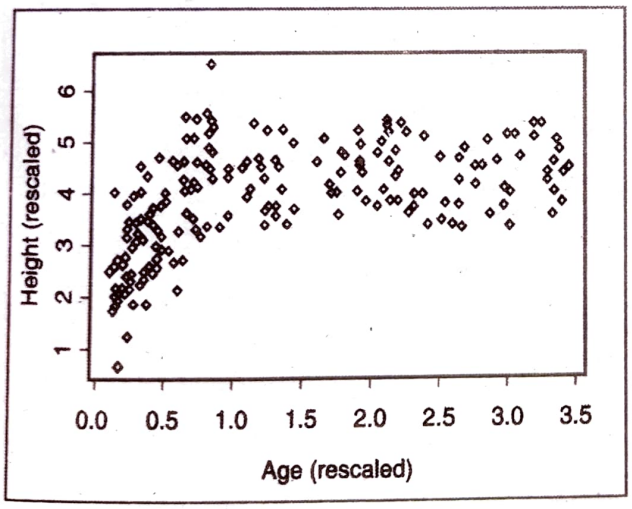

Fig. 2.2.11 : Clusters with rescaled attributes

- Using the rescaled attributes for age and height, the border of the resulting clusters at this moment collapse somewhere among the two earlier clustering analyses.
- Such an incidence is not unexpected based on the magnitudes of the attributes of the prior clustering attempts.
- There are number of practitioners that can deduct the means of the attributes to centre the attributes around zero.
- On the other hand, this step is needless since the distance formula is only sensitive to the scale of the attribute, not its location.
- In several statistical analyses, it is normal to convert normally skewed data like income, with long tails using the logarithm of the data.
- This kind of transformation can also be applied ink-means, however the Data Scientist have to be prepared for the effects of that transformation.

Additional Considerations

- The k-means algorithm is complex to the beginning positions of the initial centroid.
- Therefore, it is essential to rerun the k-means analysis number of times for a specific value of k to make sure the cluster results offer the completely least WSS.
- A Euclidean distance function is used for assigning the points to the closest centroids.
- For doing above task other possible functions can also be used such as cosine similarity and the Manhattan distance functions.
- The cosine similarity function is frequently selected to compare two documents which are depend on the occurrence of every word that present in every document.
- The following function is Manhattan distance function for two points p and q which are ranging from p_1 and q_1 to p_n and q_n respectively.

$$d_1(p, q) = \sum_{j=1}^{n} |p_j - q_j|$$

- The Manhattan distance function is similar to the distance covered by a car in a city, where the streets are positioned in a rectangular grid.

- The Euclidean distance, uses the straight line for measurement.
- From an optimization point of view, if it is essential to use the Manhattan distance for a clustering analysis, the median is a good choice for the centroid instead of using mean.
- K-means clustering is related to objects that can be defined with the help of attributes that are numerical with a significant distance measure.

❑❑❑