Clustering

K –means

Assignment 1

BP: Analyse the information given in the following ‘Insurance Policy dataset’ to create clusters of persons falling in the same type.

PROCEDURE:

STEP 1: First we have to Exploratory Data Analysis which can be done by plotting scattered plot, box plots and summary.

summary(Insurance\_Dataset)

Premiums Paid Age Days to Renew Claims made Income

Min. : 2800 Min. :23.00 Min. : 1.0 Min. : 1978 Min. : 28000

1st Qu.: 6975 1st Qu.:34.00 1st Qu.: 56.0 1st Qu.: 5221 1st Qu.: 65125

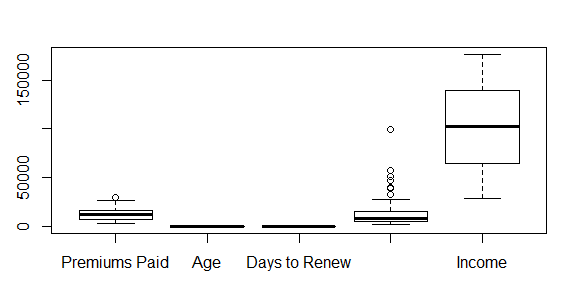
Median :11825 Median :45.00 Median : 89.0 Median : 8386 Median :102250

Mean :12542 Mean :46.11 Mean :120.4 Mean :12579 Mean :102250

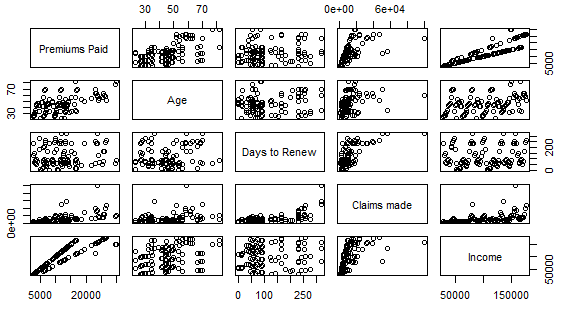
3rd Qu.:15475 3rd Qu.:54.50 3rd Qu.:186.5 3rd Qu.:14671 3rd Qu.:139375

Max. :29900 Max. :82.00 Max. :321.0 Max. :99677 Max. :176500

From the summary we can see that in column claims made the difference between mean and max is large so it may be right skewed it can my confirmed by using Box Plot .



Scatter Plot for the Following data frame is:



STEP 2: To determine the value of k, we have to do code for elbow curve :

for (i in 1:5) { wss <- c(wss,kmeans(nd,centers=i)$tot.withins)

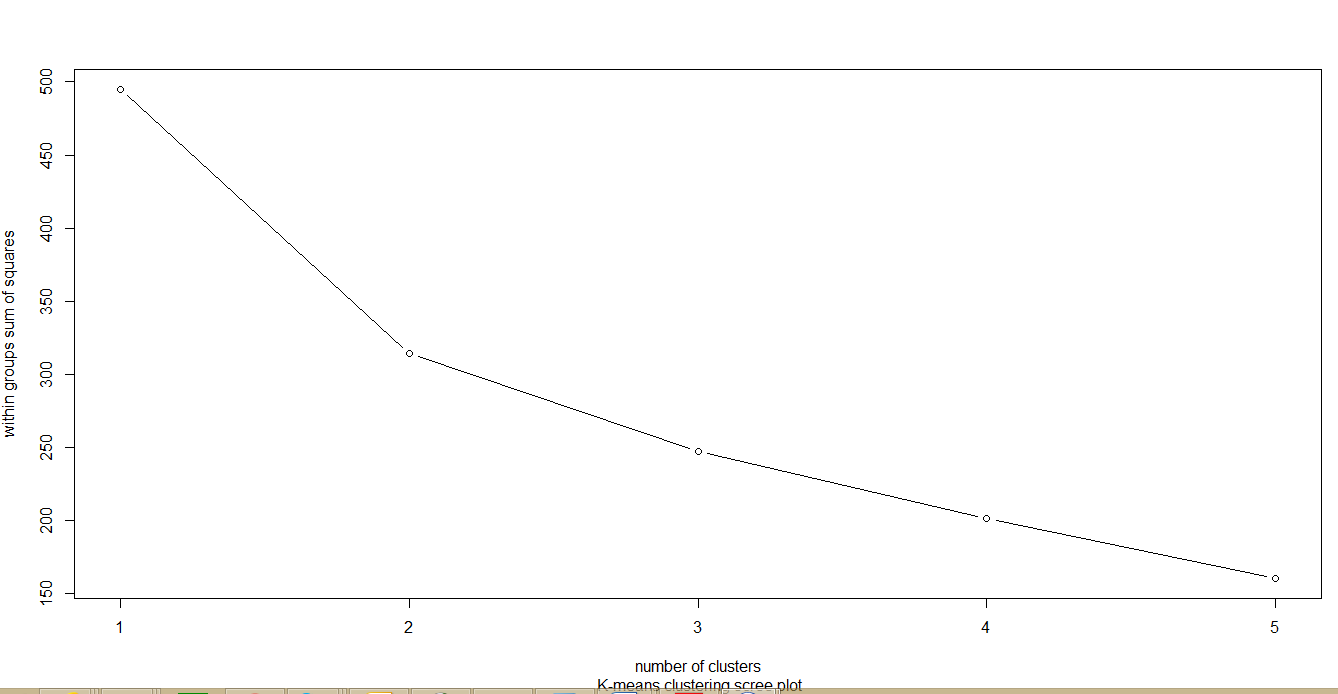
}

> wss

[1] 495.0000 313.9598 246.8452 201.2060 159.9849

> plot(wss, type = "b" , xlab ="number of clusters", ylab = "within groups sum of squares")

> title(sub = " K-means clustering scree plot")



From the elbow curve we can assume that we can form 2 clusters , to confirm our assumption we can use the auto selection of k by code :

|  |
| --- |
| k <- kselection(Insurance\_Dataset , parallel = TRUE , max\_centers=12)  > k  f(k) finds 2 clusters |
|  |
| |  | | --- | | Thus we can conclude that we can form 2 clusters . | |

STEP 3: Now we have to cluster it now:

clust <- kmeans(nd,2)

> str(clust)

List of 9

$ cluster : int [1:100] 2 2 2 2 2 2 2 2 2 2 ...

$ centers : num [1:2, 1:5] 1.307 -0.436 1.072 -0.357 0.602 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:2] "1" "2"

.. ..$ : chr [1:5] "Premiums Paid" "Age" "Days to Renew" "Claims made" ...

$ totss : num 495

$ withinss : num [1:2] 121 193

$ tot.withinss: num 314

$ betweenss : num 181

$ size : int [1:2] 25 75

$ iter : int 1

$ ifault : int 0

- attr(\*, "class")= chr "kmeans"

> clust$centers

Premiums Paid Age Days to Renew Claims made Income

1 1.3070388 1.0721764 0.6018913 1.1064252 0.9933994

2 -0.4356796 -0.3573921 -0.2006304 -0.3688084 -0.3311331

> clust$cluster

[1] 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 1 1 1 2 2 2

[40] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2

[79] 1 2 1 1 2 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1

> final <- data.frame(clust$cluster , Insurance\_Dataset)

> View(final)

> final

clust.cluster Premiums.Paid Age Days.to.Renew Claims.made Income

1 2 2800 26 233 3890.076 28000

2 2 2950 27 130 2294.444 29500

3 2 3100 28 144 2564.545 31000

4 2 3250 30 65 1978.261 32500

5 2 3400 32 56 2009.091 34000

6 2 3550 35 89 2349.455 35500

7 2 3700 44 95 2503.346 37000

8 2 3850 45 48 2217.405 38500

9 2 4000 46 76 2527.778 40000

10 2 6225 56 200 6908.232 41500

11 2 6450 67 211 7672.549 43000

12 2 6675 69 245 10208.824 44500

13 2 6900 70 261 12192.233 46000

14 2 4750 34 278 10052.326 47500

15 2 4900 44 182 4900.000 49000

16 2 7575 45 60 4535.033 50500

17 2 5200 23 12 2688.636 52000

18 2 8025 53 2 4034.669 53500

19 2 5500 48 1 2757.576 55000

20 2 5650 49 14 2938.000 56500

21 2 5800 41 17 3042.075 58000

22 2 5950 42 65 3621.739 59500

23 2 9150 50 56 5406.818 61000

24 2 6250 26 89 4136.364 62500

25 2 6400 27 95 4330.112 64000

26 2 6550 28 48 3772.468 65500

27 2 6700 30 76 4234.028 67000

28 2 6850 32 39 3836.000 68500

29 2 7000 35 34 3860.606 70000

30 2 7150 44 57 4238.762 71500

31 2 7300 45 85 4762.007 73000

32 2 7450 46 165 6813.568 74500

33 2 11400 56 234 15960.000 76000

34 1 11625 67 256 19590.278 77500

35 1 11850 69 233 16463.359 79000

36 1 12075 70 321 51108.140 80500

37 2 8200 34 233 11392.366 82000

38 2 12525 44 130 9741.667 83500

39 2 12750 45 144 10547.727 85000

40 2 12975 23 65 7897.826 86500

41 2 13200 53 56 7800.000 88000

42 2 8950 48 89 5923.273 89500

43 2 9100 49 95 6156.877 91000

44 2 9250 41 48 5327.532 92500

45 2 9400 42 76 5940.278 94000

46 2 14325 50 123 10818.050 95500

47 2 9700 26 156 8487.500 97000

48 2 9850 27 245 15064.706 98500

49 2 10000 28 261 17669.903 100000

50 2 10150 30 278 21480.233 101500

51 2 10300 32 182 10300.000 103000

52 2 10450 35 60 6256.250 104500

53 2 10600 44 12 5480.682 106000

54 2 10750 45 2 5404.696 107500

55 2 10900 46 1 5465.014 109000

56 2 16575 56 14 8619.000 110500

57 2 16800 67 17 8811.527 112000

58 1 17025 69 65 10363.043 113500

59 1 23000 70 56 13590.909 115000

60 2 11650 34 89 7710.182 116500

61 2 11800 44 95 7983.643 118000

62 2 11950 45 48 6882.595 119500

63 2 12100 23 76 7646.528 121000

64 2 18375 53 39 10290.000 122500

65 2 12400 48 34 6838.788 124000

66 2 12550 49 57 7440.065 125500

67 2 12700 41 85 8284.588 127000

68 2 12850 42 165 11752.261 128500

69 1 19500 50 234 27300.000 130000

70 2 13150 32 256 22160.185 131500

71 2 13300 34 233 18477.863 133000

72 1 13450 36 321 56927.907 134500

73 2 13600 39 65 8278.261 136000

74 2 13750 42 56 8125.000 137500

75 2 13900 44 89 9199.273 139000

76 2 14050 45 95 9505.948 140500

77 2 14200 48 48 8178.481 142000

78 2 14350 49 76 9068.403 143500

79 1 21750 54 39 12180.000 145000

80 2 14650 32 34 8079.697 146500

81 1 29600 77 57 17547.883 148000

82 1 29900 82 85 19504.660 149500

83 2 15100 34 165 13810.050 151000

84 1 15250 56 234 21350.000 152500

85 1 23100 63 256 38927.778 154000

86 1 23325 62 233 32405.725 155500

87 1 23550 59 321 99676.744 157000

88 1 23775 58 233 33030.916 158500

89 1 24000 52 130 18666.667 160000

90 1 16150 45 144 13360.455 161500

91 1 24450 54 65 14882.609 163000

92 2 16450 36 56 9720.455 164500

93 1 16600 82 89 10986.182 166000

94 2 16750 34 95 11332.714 167500

95 1 25350 56 48 14600.316 169000

96 1 25575 63 76 16161.979 170500

97 1 25800 62 166 23715.152 172000

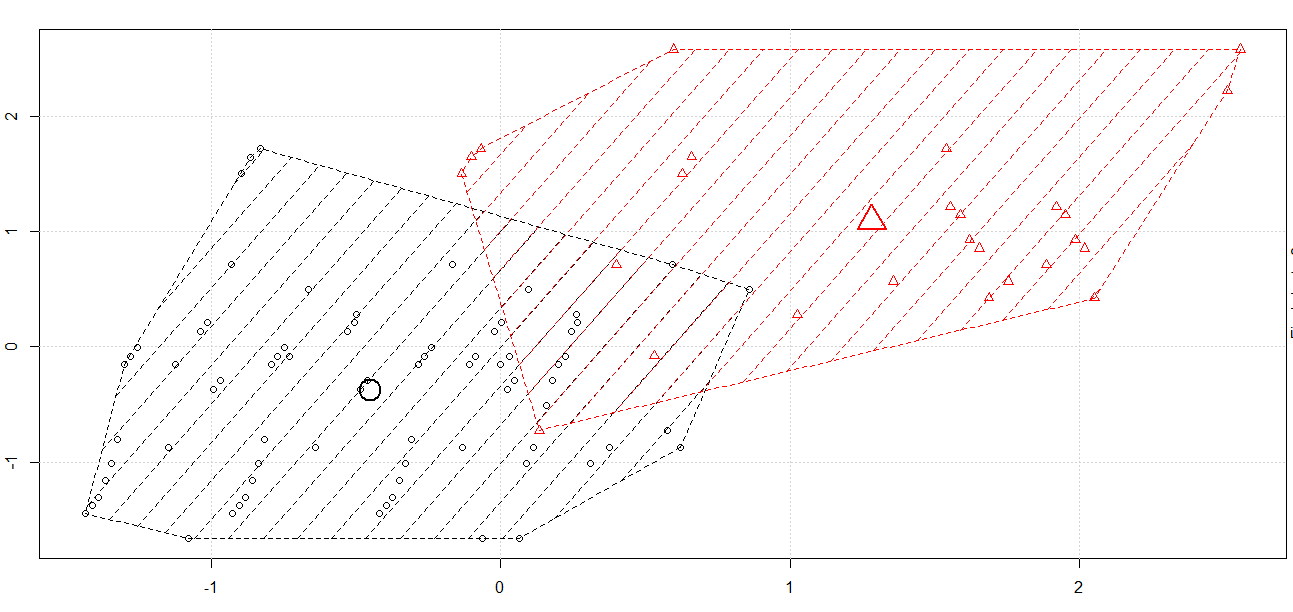
98 1 26025 59 167 24043.401 173500

99 1 26250 58 245 40147.059 175000

100 1 26475 52 261 46781.068 176500

STEP 4: Now we can visualize the clusters in graphical form by:

km <- kmeans.ani(nd,2)



Therefore, out of 100 rows from the data frame with k clustering method we have formed them into two clusters.

The inferences which we can make from this is that cluster 1 consists of the people who have premiums paid , age and the claims made are on the higher side whereas in cluster 2 comparatively these factors are relatively low.