1.) Perform Clustering for the crime data and identify the number of clusters formed and draw inferences.

Data Description:

Murder -- Muder rates in different places of United States

Assualt- Assualt rate in different places of United States

UrbanPop - urban population in different places of United States

Rape - Rape rate in different places of United States

Solution:-

**Business Problem:-** To perform clustering for the crime data and identify the number of clusters formed and draw inferences.

**Datasets:-** crime\_data.csv

Variables :- X – State names

Murder -- Muder rates in different places of United States

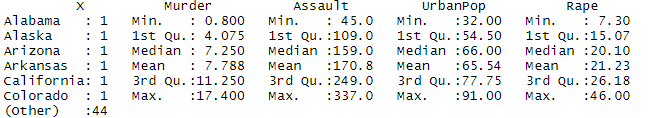
Assualt- Assualt rate in different places of United States

UrbanPop - urban population in different places of United States

Rape - Rape rate in different places of United States

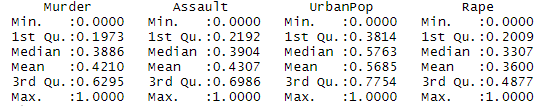
**EDA:-**

**Summary :-**



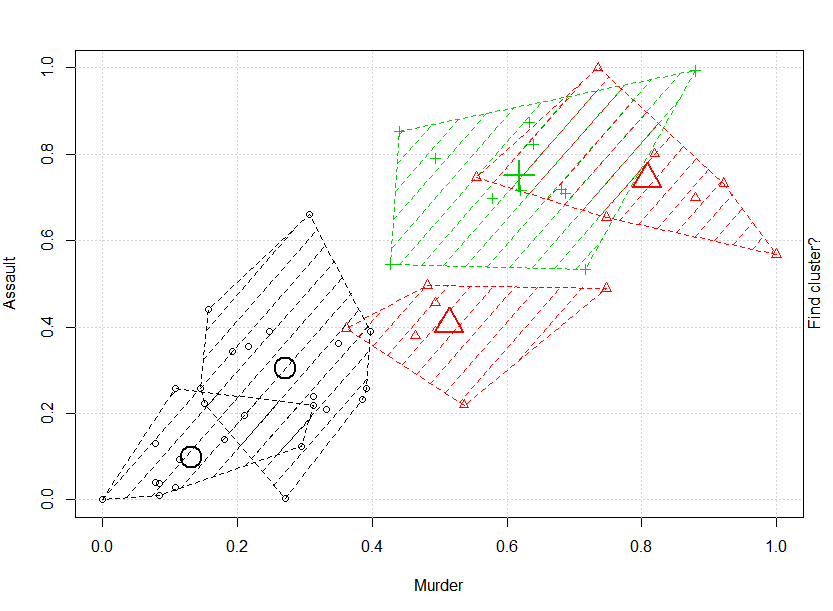
**Normalization : -** To make the data unit free and scale free.

Summary of Normalized Data



Clustering

After Applying k-means non-hierarchical (k=5) Divided into five clusters



Clusters of the data points

5 5 3 2 3 3 4 4 3 5 4 1 3 4 1 4 2 5 1 3 4 3 1 5 2 1 4 3 1 4 3 3 5 1 4 4 4 4 4 5 1 2 3 4 1 2 4 1 1 2

Centers of the 5 clusters

Murder Assault UrbanPop Rape

0.1314348 0.09806974 0.3251156 0.1061781

0.5140562 0.40639269 0.4576271 0.3591731

0.6177437 0.75031133 0.7981510 0.6542166

0.2710843 0.30410959 0.7242938 0.2949182

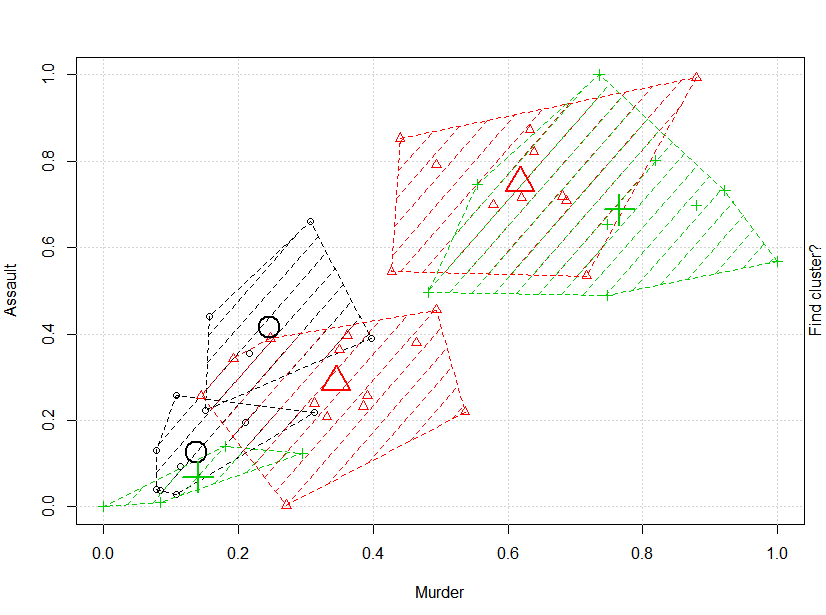
0.8080895 0.74315068 0.3510896 0.4366925

Membership- This will summarize which data points belongs to which cluster

|  |
| --- |
| fit\_crime\_data.cluster X Murder Assault Urban Pop Rape  1 5 Alabama 13.2 236 58 21.2  2 5 Alaska 10.0 263 48 44.5  3 3 Arizona 8.1 294 80 31.0  4 2 Arkansas 8.8 190 50 19.5  5 3 California 9.0 276 91 40.6  6 3 Colorado 7.9 204 78 38.7  7 4 Connecticut 3.3 110 77 11.1  8 4 Delaware 5.9 238 72 15.8  9 3 Florida 15.4 335 80 31.9  10 5 Georgia 17.4 211 60 25.8  11 4 Hawaii 5.3 46 83 20.2  12 1 Idaho 2.6 120 54 14.2  13 3 Illinois 10.4 249 83 24.0  14 4 Indiana 7.2 113 65 21.0  15 1 Iowa 2.2 56 57 11.3  16 4 Kansas 6.0 115 66 18.0  17 2 Kentucky 9.7 109 52 16.3  18 5 Louisiana 15.4 249 66 22.2  19 1 Maine 2.1 83 51 7.8  20 3 Maryland 11.3 300 67 27.8  21 4 Massachusetts 4.4 149 85 16.3  22 3 Michigan 12.1 255 74 35.1  23 1 Minnesota 2.7 72 66 14.9  24 5 Mississippi 16.1 259 44 17.1  25 2 Missouri 9.0 178 70 28.2  26 1 Montana 6.0 109 53 16.4  27 4 Nebraska 4.3 102 62 16.5  28 3 Nevada 12.2 252 81 46.0  29 1 New Hampshire 2.1 57 56 9.5  30 4 New Jersey 7.4 159 89 18.8  31 3 New Mexico 11.4 285 70 32.1  32 3 New York 11.1 254 86 26.1  33 5 North Carolina 13.0 337 45 16.1  34 1 North Dakota 0.8 45 44 7.3  35 4 Ohio 7.3 120 75 21.4  36 4 Oklahoma 6.6 151 68 20.0  37 4 Oregon 4.9 159 67 29.3  38 4 Pennsylvania 6.3 106 72 14.9  39 4 Rhode Island 3.4 174 87 8.3  40 5 South Carolina 14.4 279 48 22.5  41 1 South Dakota 3.8 86 45 12.8  42 2 Tennessee 13.2 188 59 26.9  43 3 Texas 12.7 201 80 25.5  44 4 Utah 3.2 120 80 22.9  45 1 Vermont 2.2 48 32 11.2  46 2 Virginia 8.5 156 63 20.7  47 4 Washington 4.0 145 73 26.2  48 1 West Virginia 5.7 81 39 9.3  49 1 Wisconsin 2.6 53 66 10.8  50 2 Wyoming 6.8 161 60 15.6  Elbow chart |
|  |
|  |

Now from the elbow chart I could k=6, is required no. of clusters are required

After applying k-means for, k=6



Cluster of the data points

6 6 2 6 2 2 4 4 2 6 5 1 2 5 1 5 5 6 1 2 4 2 1 6 5 1 1 2 1 4 2 2 6 3 5 5 5 5 4 6 3 6 2 5 3 5 5 3 1 5

Centres

Murder Assault UrbanPop Rape

0.1370482 0.12500000 0.4427966 0.13888889

0.6177437 0.75031133 0.7981510 0.65421658

0.1400602 0.06849315 0.1355932 0.07364341

0.2457831 0.41438356 0.8474576 0.17467700

0.3447637 0.28819810 0.6232073 0.35738422

0.7650602 0.68759513 0.3578154 0.43095033

Membership

fit\_crime\_data.cluster X Murder Assault UrbanPop Rape

1 6 Alabama 13.2 236 58 21.2

2 6 Alaska 10.0 263 48 44.5

3 2 Arizona 8.1 294 80 31.0

4 6 Arkansas 8.8 190 50 19.5

5 2 California 9.0 276 91 40.6

6 2 Colorado 7.9 204 78 38.7

7 4 Connecticut 3.3 110 77 11.1

8 4 Delaware 5.9 238 72 15.8

9 2 Florida 15.4 335 80 31.9

10 6 Georgia 17.4 211 60 25.8

11 5 Hawaii 5.3 46 83 20.2

12 1 Idaho 2.6 120 54 14.2

13 2 Illinois 10.4 249 83 24.0

14 5 Indiana 7.2 113 65 21.0

15 1 Iowa 2.2 56 57 11.3

16 5 Kansas 6.0 115 66 18.0

17 5 Kentucky 9.7 109 52 16.3

18 6 Louisiana 15.4 249 66 22.2

19 1 Maine 2.1 83 51 7.8

20 2 Maryland 11.3 300 67 27.8

21 4 Massachusetts 4.4 149 85 16.3

22 2 Michigan 12.1 255 74 35.1

23 1 Minnesota 2.7 72 66 14.9

24 6 Mississippi 16.1 259 44 17.1

25 5 Missouri 9.0 178 70 28.2

26 1 Montana 6.0 109 53 16.4

27 1 Nebraska 4.3 102 62 16.5

28 2 Nevada 12.2 252 81 46.0

29 1 New Hampshire 2.1 57 56 9.5

30 4 New Jersey 7.4 159 89 18.8

31 2 New Mexico 11.4 285 70 32.1

32 2 New York 11.1 254 86 26.1

33 6 North Carolina 13.0 337 45 16.1

34 3 North Dakota 0.8 45 44 7.3

35 5 Ohio 7.3 120 75 21.4

36 5 Oklahoma 6.6 151 68 20.0

37 5 Oregon 4.9 159 67 29.3

38 5 Pennsylvania 6.3 106 72 14.9

39 4 Rhode Island 3.4 174 87 8.3

40 6 South Carolina 14.4 279 48 22.5

41 3 South Dakota 3.8 86 45 12.8

42 6 Tennessee 13.2 188 59 26.9

43 2 Texas 12.7 201 80 25.5

44 5 Utah 3.2 120 80 22.9

45 3 Vermont 2.2 48 32 11.2

46 5 Virginia 8.5 156 63 20.7

47 5 Washington 4.0 145 73 26.2

48 3 West Virginia 5.7 81 39 9.3

49 1 Wisconsin 2.6 53 66 10.8

50 5 Wyoming 6.8 161 60 15.6

Aggregation

Clusters Murder Assault UrbanPop Rape

1 10.844444 267.1111 80.33333 33.94444

2 5.275000 180.0000 83.25000 14.80000

3 5.457143 119.7143 69.78571 19.47857

4 14.500000 291.6667 45.66667 18.56667

5 12.333333 224.0000 62.00000 26.84444

6 3.618182 77.0000 49.90909 11.53636

**Interpretation**

Now the states belong to cluster 4th is the highest in murder and assault cases,

compared to others states, even though the urban population is less there.

Rape cases are more in the states of cluster 1. Less crime occur in the states of cluster 6.

2.) Perform clustering (Both hierarchical and K means clustering) for the airlines data to

obtain optimum number of clusters.

**Business Problem:** To perform clustering

**Datasets:** EastWestAirlines.xlsx

**EDA:**

**Summary**

Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles

Min. : 0 Min. : 0.0 Min. :1.00 Min. :1.000 Min. :1.000

1st Qu.: 18528 1st Qu.: 0.0 1st Qu.:1.00 1st Qu.:1.000 1st Qu.:1.000

Median : 43097 Median : 0.0 Median :1.00 Median :1.000 Median :1.000

Mean : 73601 Mean : 144.1 Mean :2.06 Mean :1.015 Mean :1.012

3rd Qu.: 92404 3rd Qu.: 0.0 3rd Qu.:3.00 3rd Qu.:1.000 3rd Qu.:1.000

Max. :1704838 Max. :11148.0 Max. :5.00 Max. :3.000 Max. :5.000

Bonus\_miles Bonus\_trans Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll

Min. : 0 Min. : 0.0 Min. : 0.0 Min. : 0.000 Min. : 2

1st Qu.: 1250 1st Qu.: 3.0 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.:2330

Median : 7171 Median :12.0 Median : 0.0 Median : 0.000 Median :4096

Mean : 17145 Mean :11.6 Mean : 460.1 Mean : 1.374 Mean :4119

3rd Qu.: 23801 3rd Qu.:17.0 3rd Qu.: 311.0 3rd Qu.: 1.000 3rd Qu.:5790

Max. :263685 Max. :86.0 Max. :30817.0 Max. :53.000 Max. :8296

Award.

Min. :0.0000

1st Qu.:0.0000

Median :0.0000

Mean :0.3703

3rd Qu.:1.0000

Max. :1.0000

**Normalization**

**Summary of normalized data**

Balance Qual\_miles cc1\_miles cc2\_miles

Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.000000

1st Qu.:0.01087 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.000000

Median :0.02528 Median :0.00000 Median :0.0000 Median :0.000000

Mean :0.04317 Mean :0.01293 Mean :0.2649 Mean :0.007252

3rd Qu.:0.05420 3rd Qu.:0.00000 3rd Qu.:0.5000 3rd Qu.:0.000000

Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.000000

cc3\_miles Bonus\_miles Bonus\_trans Flight\_miles\_12mo

Min. :0.000000 Min. :0.00000 Min. :0.00000 Min. :0.00000

1st Qu.:0.000000 1st Qu.:0.00474 1st Qu.:0.03488 1st Qu.:0.00000

Median :0.000000 Median :0.02720 Median :0.13953 Median :0.00000

Mean :0.003063 Mean :0.06502 Mean :0.13491 Mean :0.01493

3rd Qu.:0.000000 3rd Qu.:0.09026 3rd Qu.:0.19767 3rd Qu.:0.01009

Max. :1.000000 Max. :1.00000 Max. :1.00000 Max. :1.00000

Flight\_trans\_12 Days\_since\_enroll Award.

Min. :0.00000 Min. :0.0000 Min. :0.0000

1st Qu.:0.00000 1st Qu.:0.2807 1st Qu.:0.0000

Median :0.00000 Median :0.4936 Median :0.0000

Mean :0.02592 Mean :0.4963 Mean :0.3703

3rd Qu.:0.01887 3rd Qu.:0.6979 3rd Qu.:1.0000

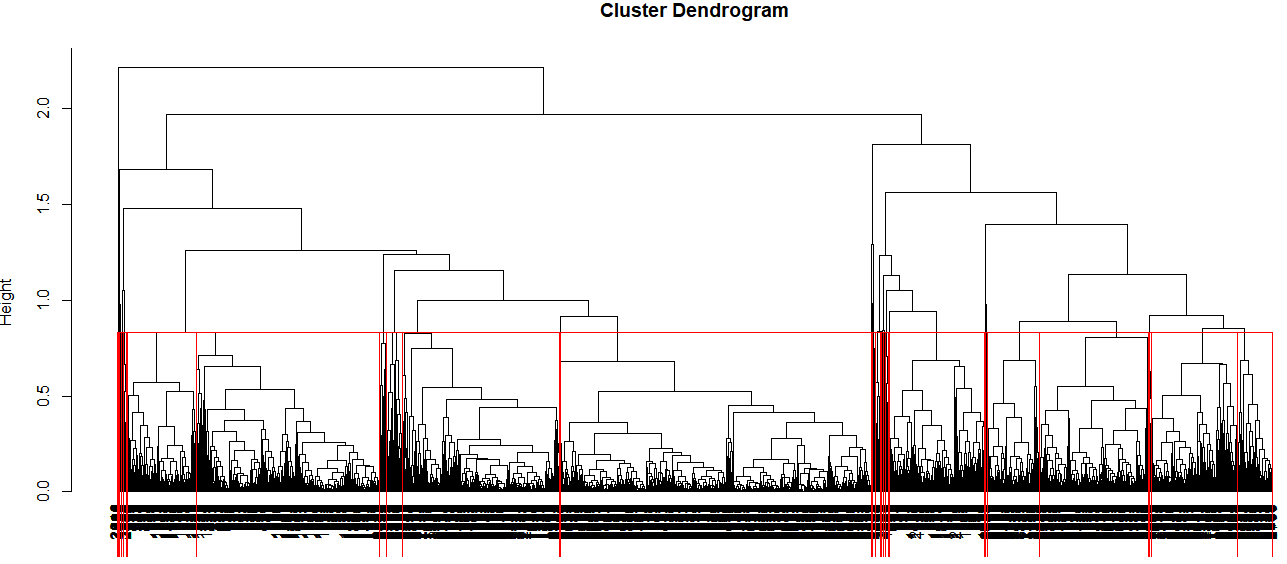
Max. :1.00000 Max. :1.0000 Max. :1.0000

**Hierarchical Clustering**

Calculating Euclidean distance

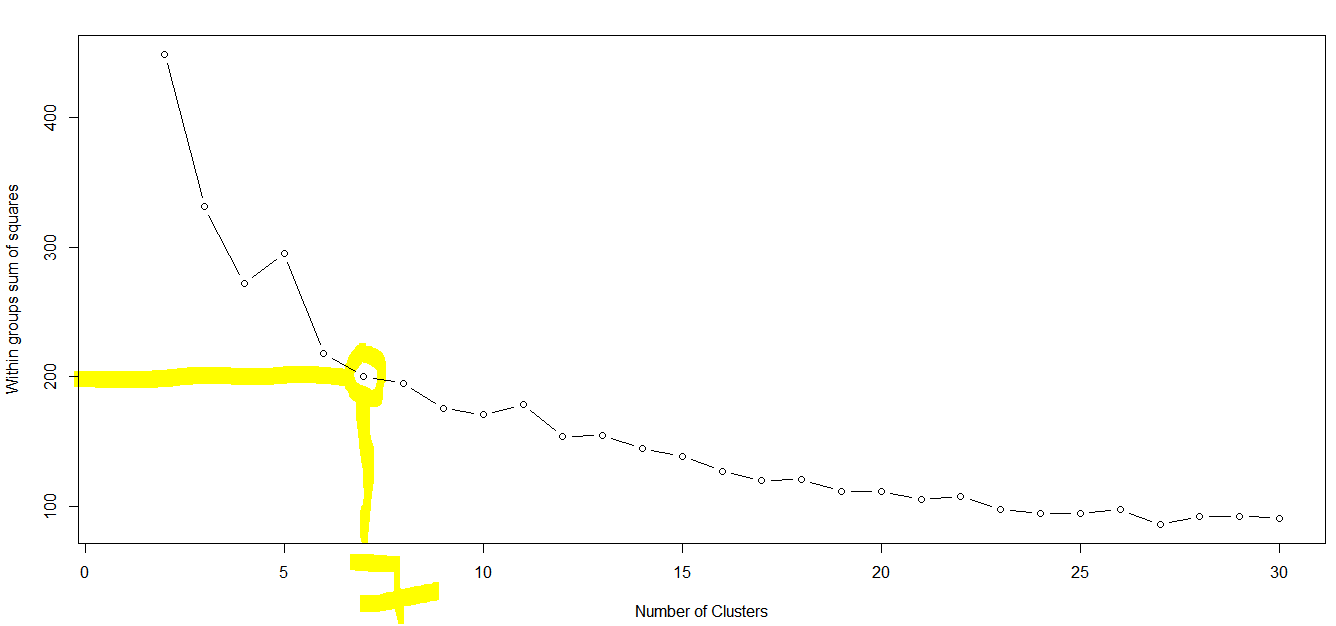
And then cluster summary

Dendogram for 30 clusters

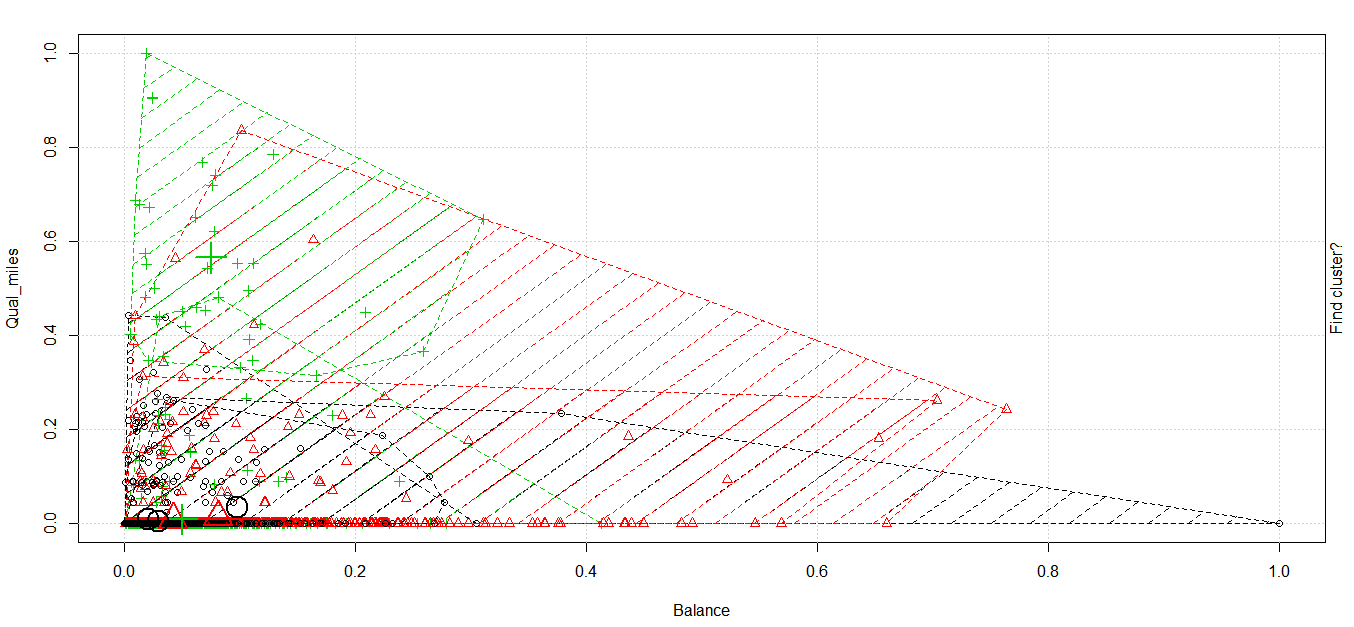


K-Means

Elbow Chart



This chart shows 7 no. of clusters



The plot represents the seven clusters

Aggregation

Group.1 Balance Qual\_miles cc1\_miles cc2\_miles cc3\_miles Bonus\_miles Bonus\_trans

1 1 34859.30 79.92998 1.084813 1.015779 1.003945 4089.945 6.560158

2 2 71569.82 83.51980 1.150165 1.018152 1.011551 5150.622 7.803630

3 3 84976.99 92.31306 3.507418 1.001484 1.013353 31120.889 17.925816

4 4 165638.65 377.80769 1.307692 1.096154 1.000000 22592.538 31.115385

5 5 137703.73 138.67655 4.018868 1.001348 1.035040 47023.350 19.623989

6 6 127154.50 6296.70588 1.470588 1.000000 1.000000 11221.118 11.235294

7 7 49288.43 52.22463 1.111745 1.027366 1.003421 4413.482 7.265678

Flight\_miles\_12mo Flight\_trans\_12 Days\_since\_enroll

1 266.5059 0.7642998 1730.811

2 335.6964 1.0643564 6674.135

3 351.1395 1.0326409 3018.283

4 7696.2308 22.6346154 4038.981

5 563.1199 1.6954178 6180.633

6 1114.6765 3.3529412 4321.500

7 311.8449 0.9441277 4211.220