EM 623 – Data Science and Knowledge Discovery

EXERCISE 04

Rushabh Barbhaya | Observation Report | 3/4/2018

Contents

List of Figures

Objective

We were given a data set of Wine and a complete freedom to perform CRISP-DM on the data set. So my objective was to clean the data i.e. perform missing value detection & outlier detection. Then I decided to normalize the data for ease in comparing the data. After performing all the data cleaning tools I added nodes to visualize it.

SOFTWARES USED

- Knime
- Rattle with Rstudio

Understanding the data

'Wine_Header' the dataset has many components in it. To understand these components a 'wines_metadata' file was given. The metadata explained the source of the file along with the contributors and contact information. The dataset has 14 dimensions to it. A sample of dataset can be seen in Figure 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	WineType	Alcohol	Malic acid	Ash	Alcalinity o	Magnesiun	Total phen	Flavanoids	Nonflavanc	Proanthocy	Color inten	Hue	OD280/OD315 of diluted wines	Proline
2	1	14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
3	1	13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
4	1	13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
5	1	14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
6	1	13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
7	1	14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
8	1	14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
9	1	14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
10	1	14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
11	1	13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045
12	1	14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510
13	1	14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280
14	1	13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320

Image 1 Sample of dataset

Data Cleaning

It's an important and a necessary step. We don't want to include values that are partially or totally missing parameters. It would affect our statistics and model of the whole system. We also don't want to include data that is miscalculated. Image 2 shows the data cleaning process in Knime.

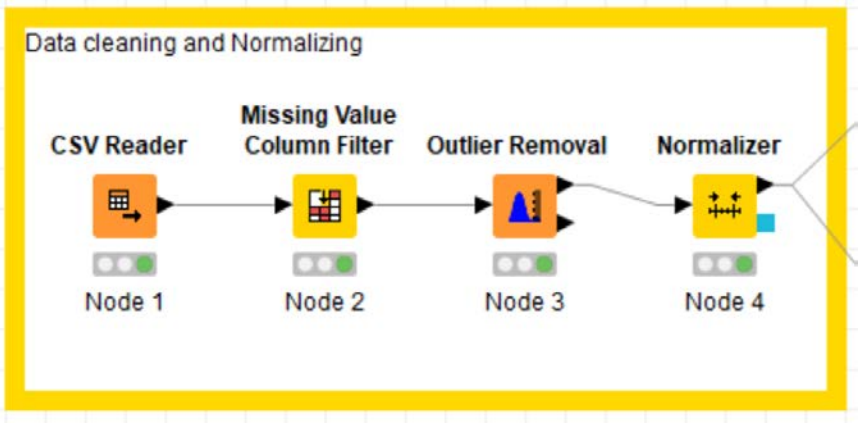

Image 2 Data cleaning

MISSING VALUES

Missing value filter detects and passes only those data which are complete. We can set the tolerance value/threshold value for missing entries. I have set the threshold to 95%. To do that I configured the value in 'Missing Value Column Filter' [shown in image 3]

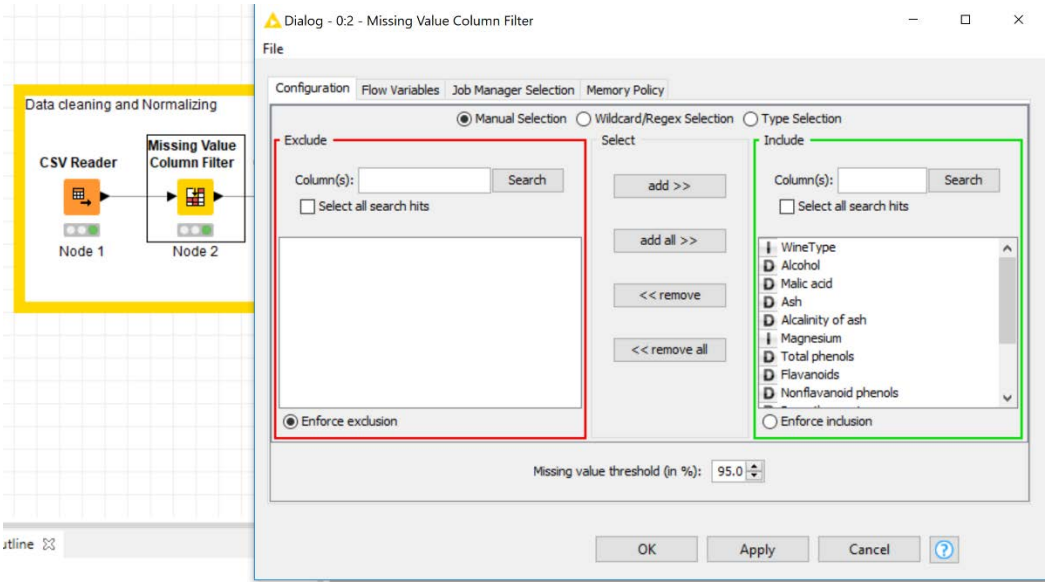

Image 3 Missing Value Filter Configuration

OUTLIERS

Our dataset could contain some outliers, due to human error in recording or some faulty equipment values. To remove those an outlier removal node is added. Its configuration is set to remove any data which doesn't fall under ±3 standard deviations from the mean value. Image 4 shows the configuration in Knime.

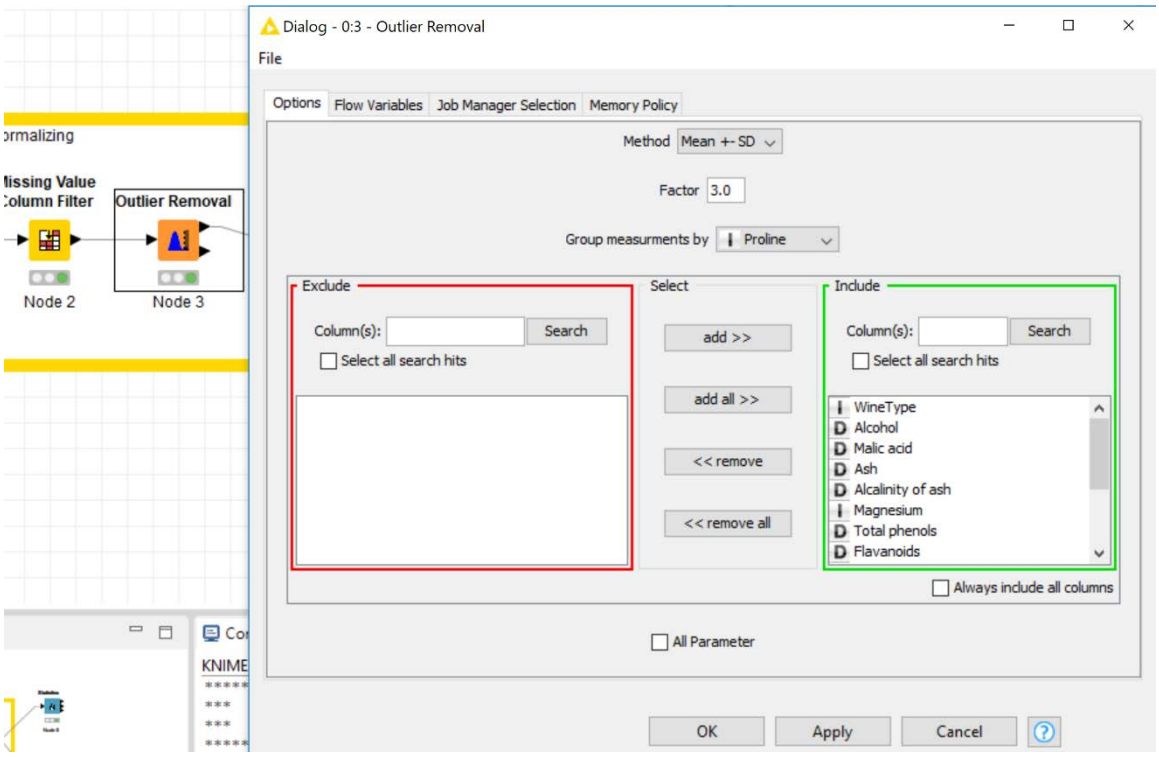

Image 4 Outlier Remover Configuration

NORMALIZING

Normalizing the values makes visualization easier. Values in the columns become smaller and are easier to compare and plot on a graph. That's why before sending data to visual parameters I passed them through a normalizer. In the configuration options the values were set to a minimum of 0 and a maximum of 1. This process is also called min-max transformation.

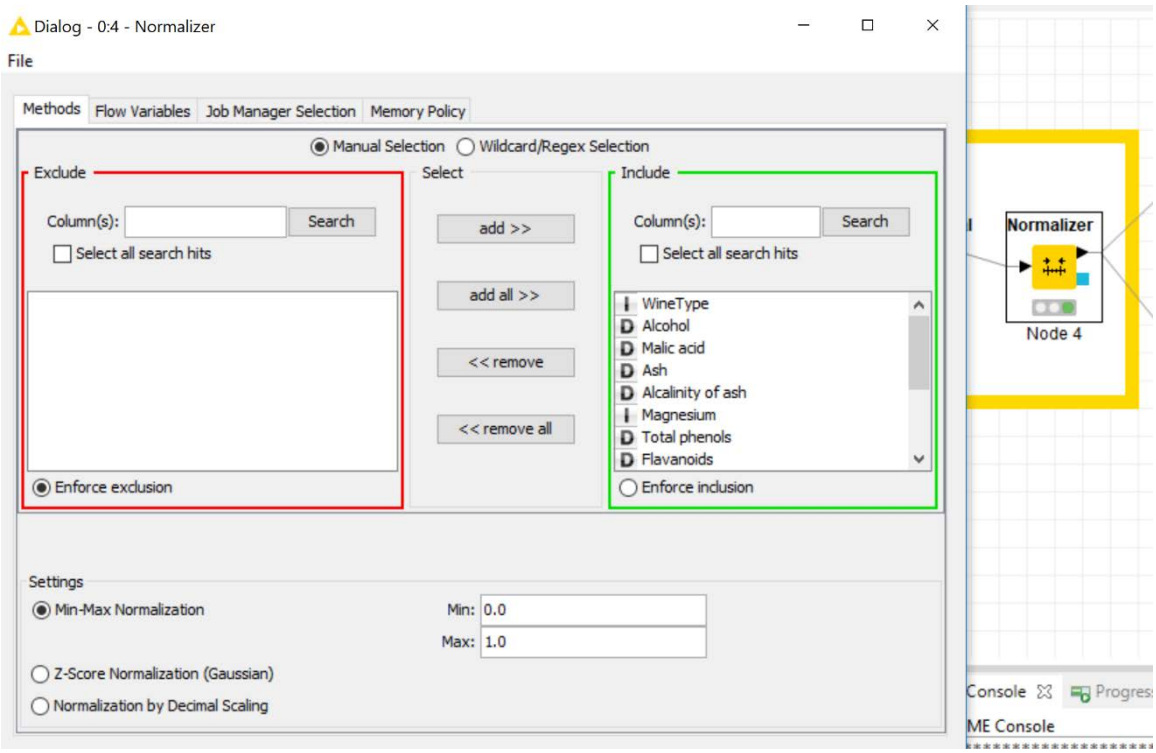

Image 5 Normalization Configuration

Statistics

For statistics I prefer to use Rattle. To get statistics from rattle. First load the csv file in rattle and execute. The dataset will be visible as shown in the image below.

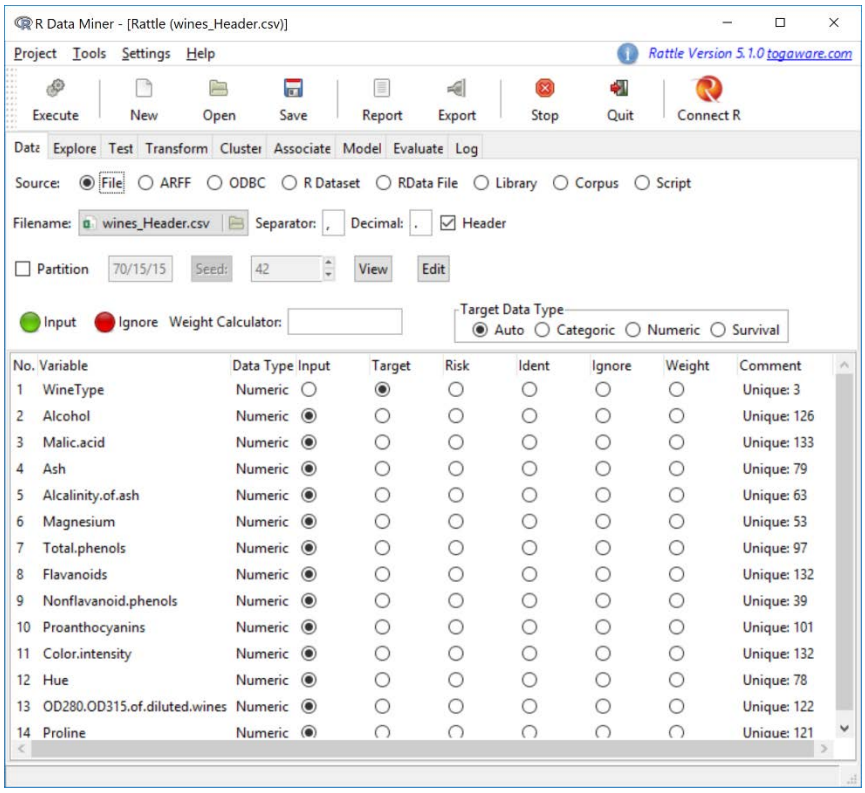

Image 6 Loading Dataset

Over in the 'explore' tab select 'distribution' and select the columns for which a histogram is to be plotted.

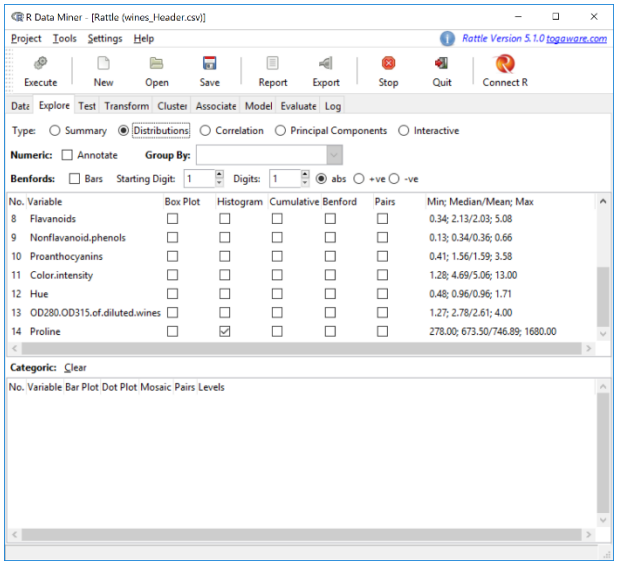

Image 7 Histogram Plot path

After executing the selection over to the rstudio plot pane the histogram will be visible. And they look like this.

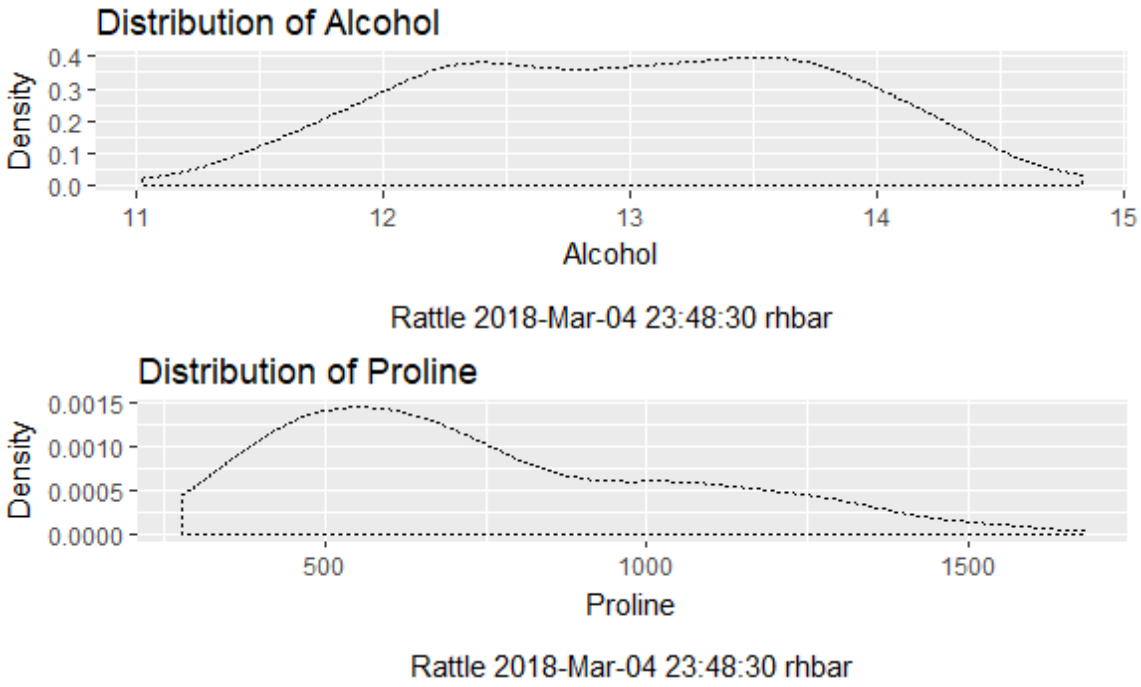

Image 8 Histogram

From the plot we can observe that the alcohol content in the wine are mostly in the range of 12 to 14 with a smooth slope on both the ends. On the other hand, the proline saturation is mostly concentrated below 1000.

Correlation

The correlation analysis shows how each parameter is dependent on others. Some parameters are highly correlated i.e. one parameter grows, other also grow. Whereas, some are inversely correlated i.e. if one parameter grown the other reduces. There could be some which are not correlated. The correlation of these 14 dimensions are shown in the image below.

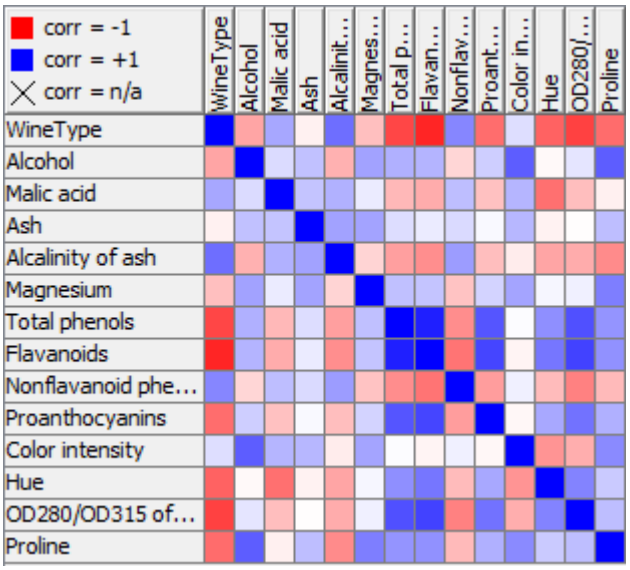

Image 9 Correlation Matrix

From this matrix it can be stated that the 'praline' parameter is inversely related with the 'alcohol' parameter & the 'color intensity' is directly correlated with 'alcohol', 'flavonoids' and 'praline'

To get this correlation matrix. Add a 'rank correlation' node and select the 'view: correlation matrix' from its options.

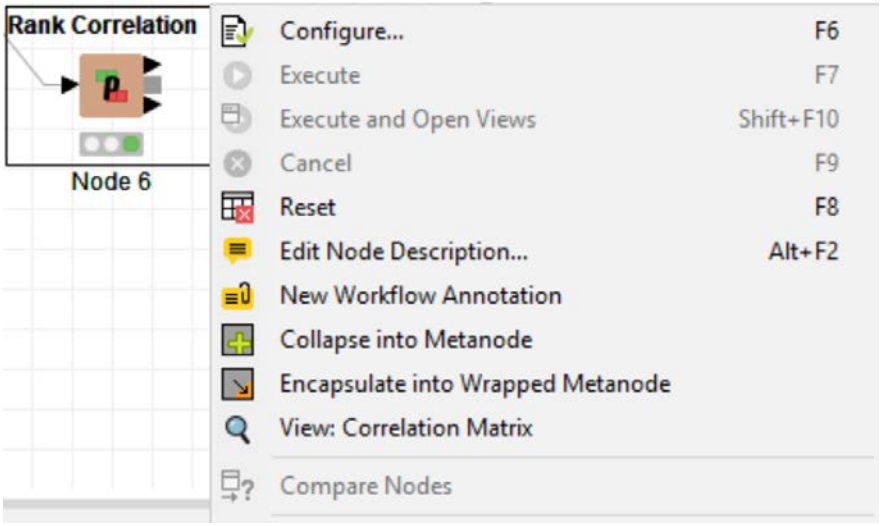

Image 10 Correlation Matrix Option

K-Means Clustering

We can also form clusters of data and compare it. K-means is a method of clustering where a multiple group of clusters can be formed and displayed. These clusters revolve around a centroid value. I performed this operation using Rattle.

First we need to determine how many cluster we need to form. To know that we select the 'cluster' tab and iterate clusters. When we execute that we get a graph.

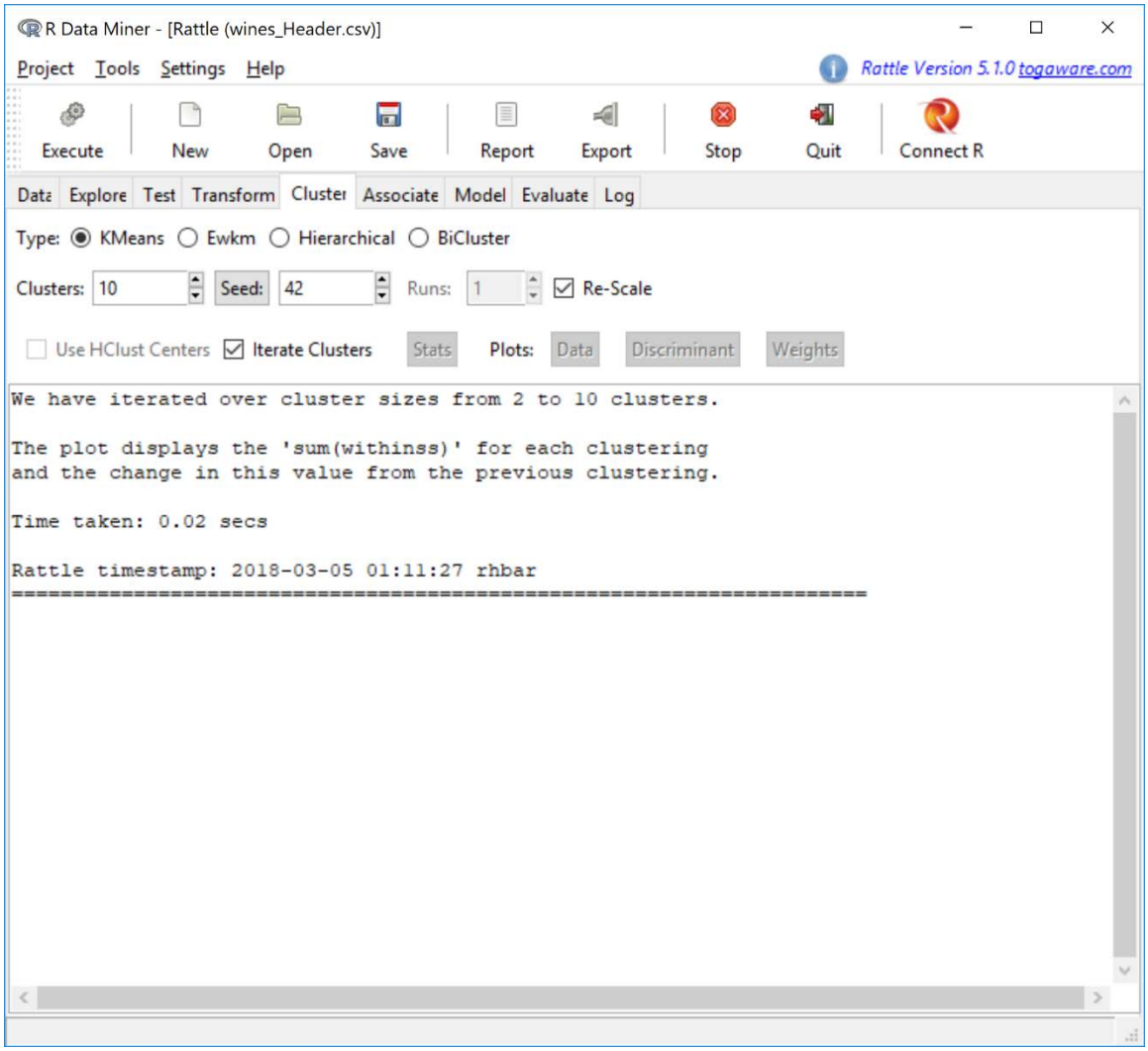

Image 11 Iteration Function

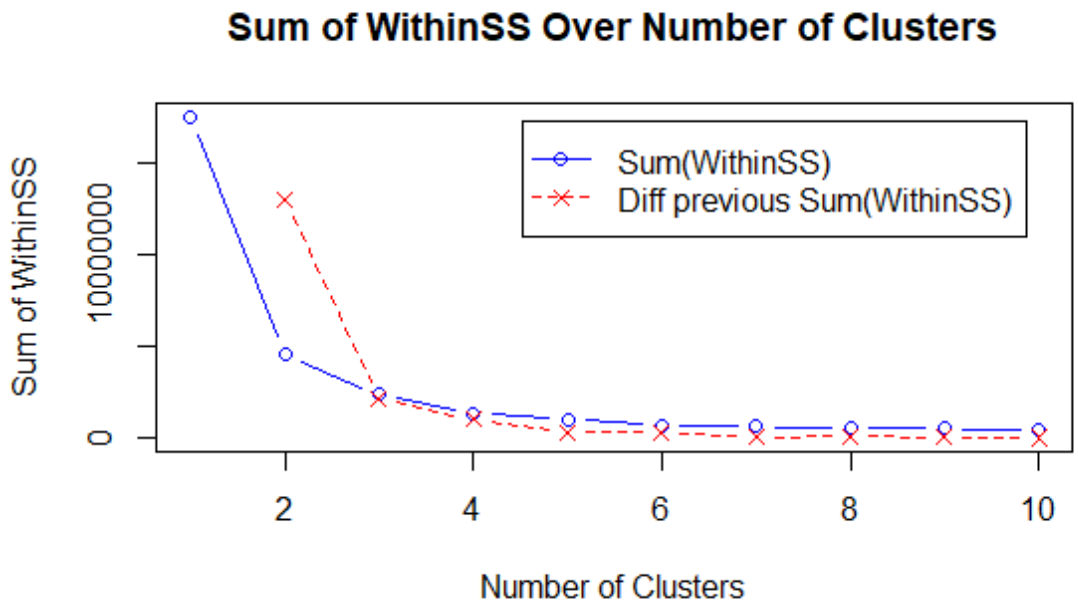

Image 12 Clusters

We select the number of clusters where the difference with previous sum (red) changes the least. In this case I chose 5. When I entered that data in cluster group and executed that. A graph was plot.

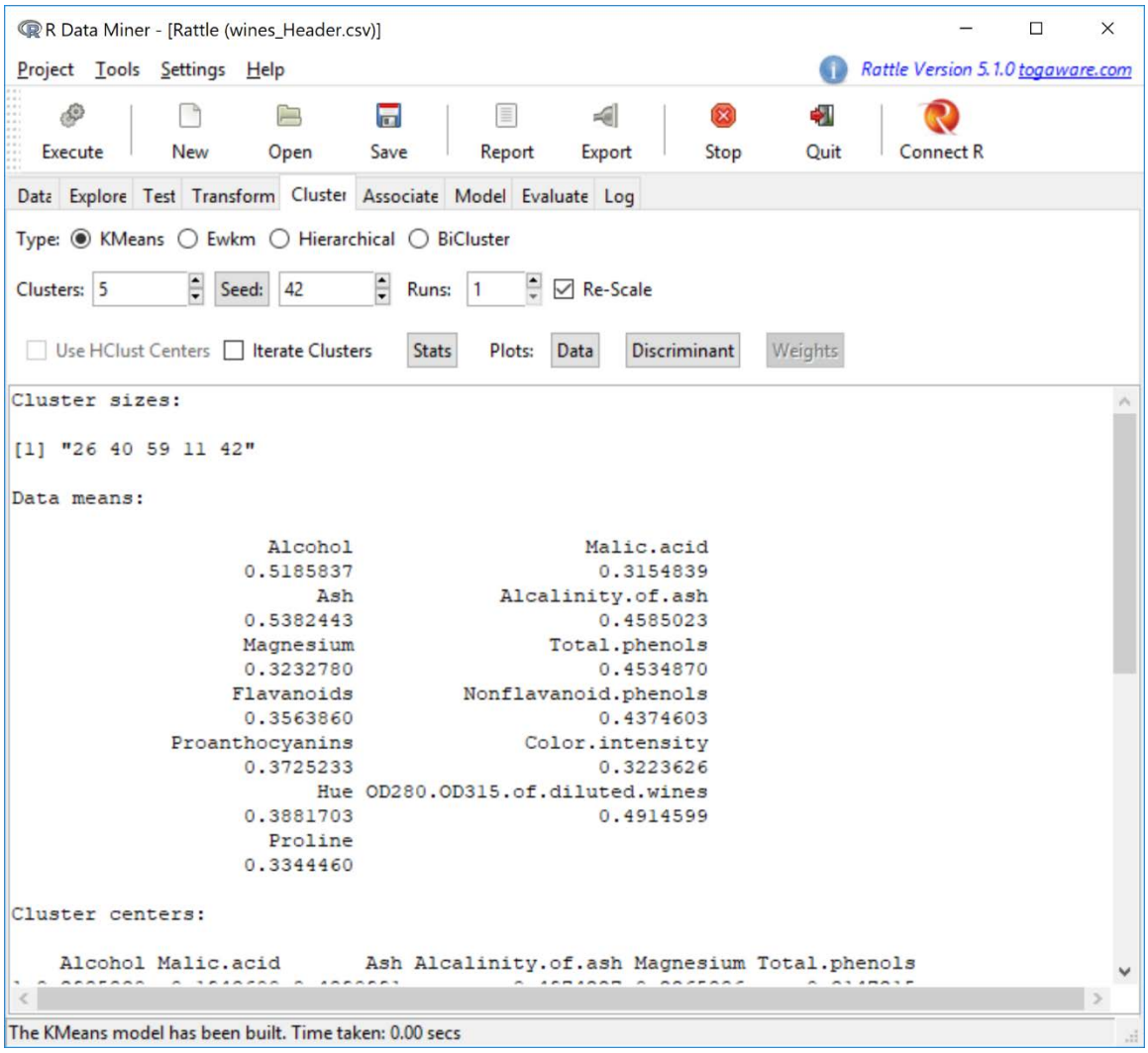

Image 13 Cluster data

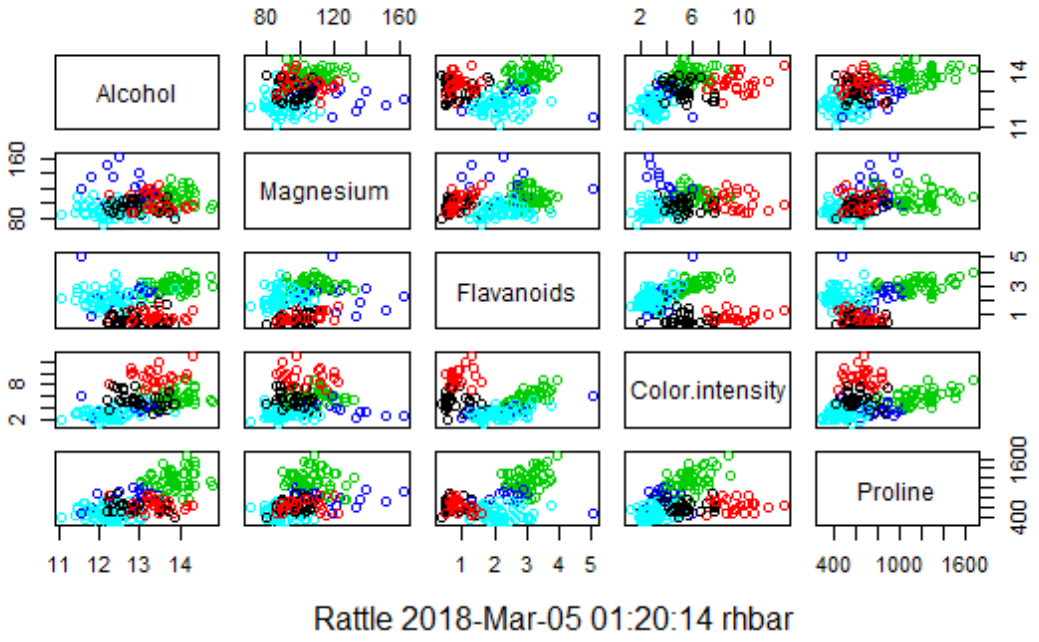

Image 14 K-Means Cluster

I selected the data of 'Alcohol', 'Magnesium', 'Flavanoids', 'Color Intensity' and 'Proline' for clusters. I ignored all other values in the 'data' tab.