

Project Report:

The screenshot shows the RapidMiner Studio interface. The main process canvas contains three operators: 'Retrieve Iris', 'Clustering', and 'SVD (Singular Value Decomposition)'. The 'SVD' operator is selected, and its parameters are visible on the right: 'dimensionality reduc...' is set to 'fixed number', and 'dimensions' is set to '2'. The bottom status bar shows the system time as 12:53 on 16/12/2015.

1. Describe the original example set in terms of statistics and visualization
 - a. For the statistics, use the relevant statistics generated by RapidMiner

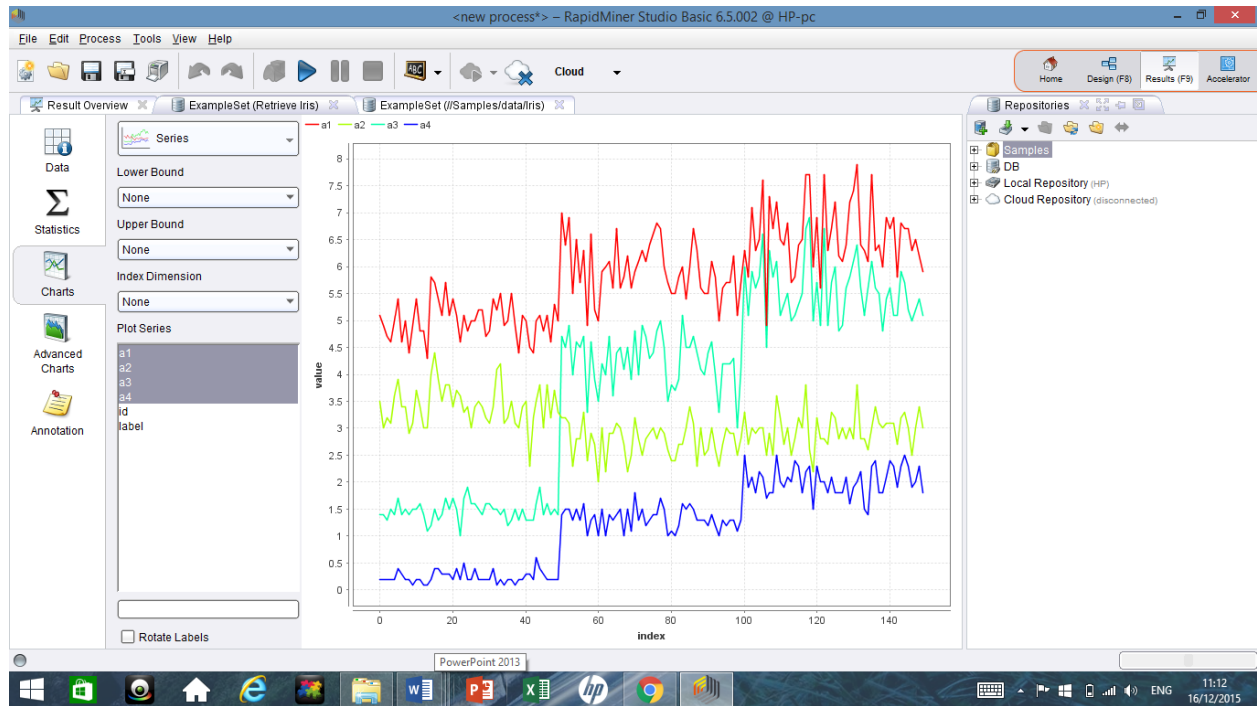
Examples = 150; Special Attributes = 2; Regular Attributes = 4

The screenshot shows the 'Result Overview' window in RapidMiner Studio. It displays a table of statistics for the Iris dataset. The table has columns for Name, Type, Miss., Statistics, and Values. The first row shows a Nominal attribute with 0 missing values and a list of values. The second row shows a Nominal attribute with 0 missing values and a bar chart. The following three rows show Real attributes with their respective Min, Max, Average, and Deviation values.

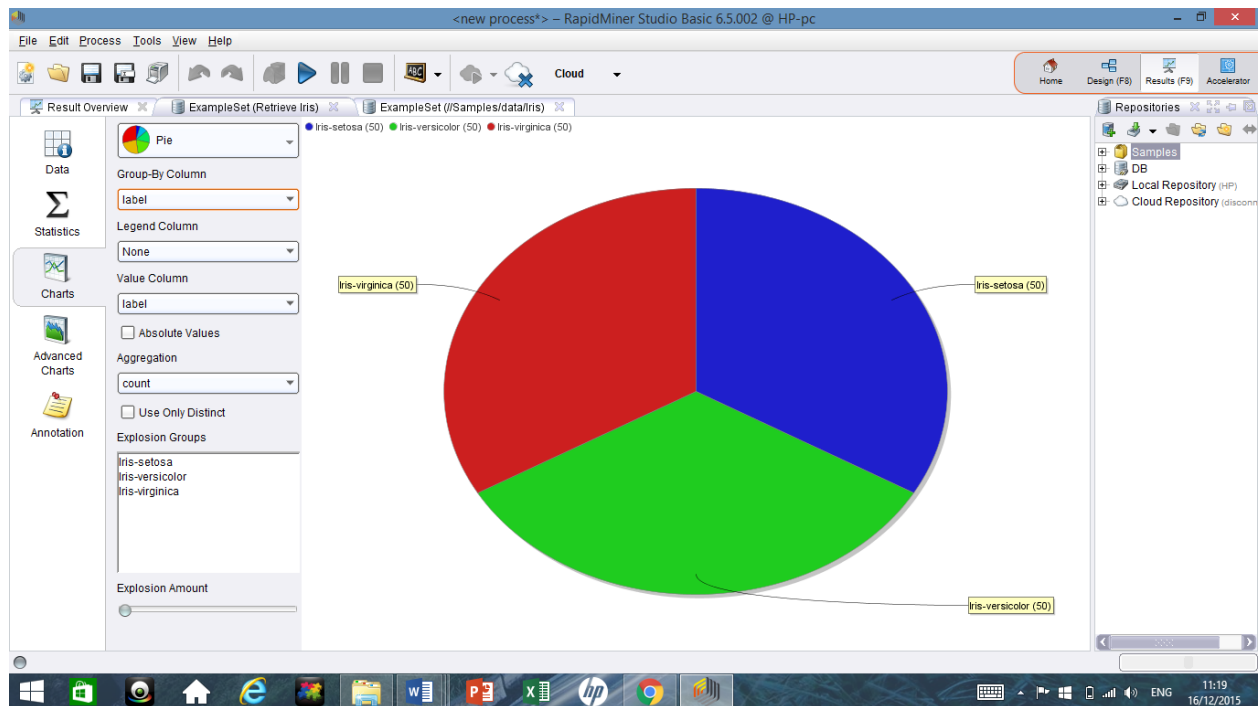
Name	Type	Miss.	Statistics	Values
Nominal	0	Least id_99 (1)	Most id_1 (1)	id_1 (1), id_10 (1), ...[148 more]
Nominal	0		Least Iris-virginica (50)	Most Iris-setosa (50)
Real	0	Min 4.300	Max 7.900	Average 5.843
Real	0	Min 2	Max 4.400	Average 3.054
Real	0	Min 1	Max 6.900	Average 3.759
Real	0	Min 0.100	Max 2.500	Average 1.199

Showing attributes: 1 - 6 Examples: 150 Special Attributes: 2 Regular Attributes: 4

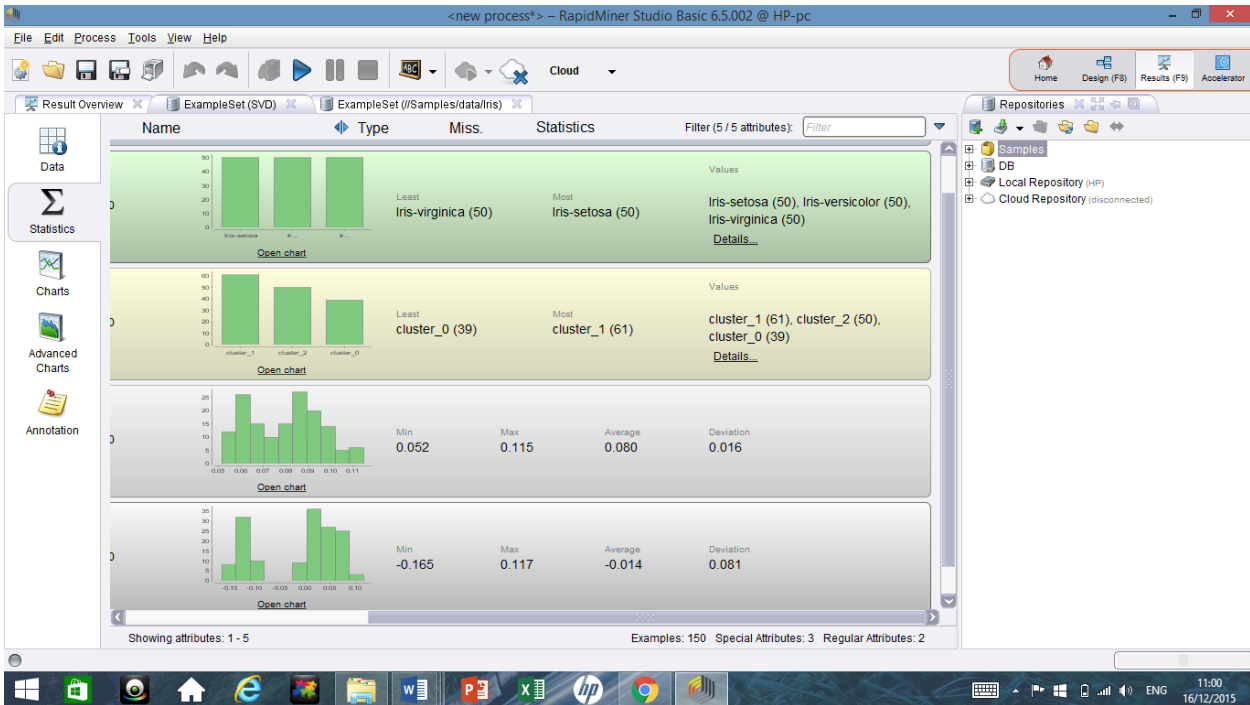
- b. For the visualization,
 - i. Draw a series plot of the four attributes (a1, a2,a3,a4)



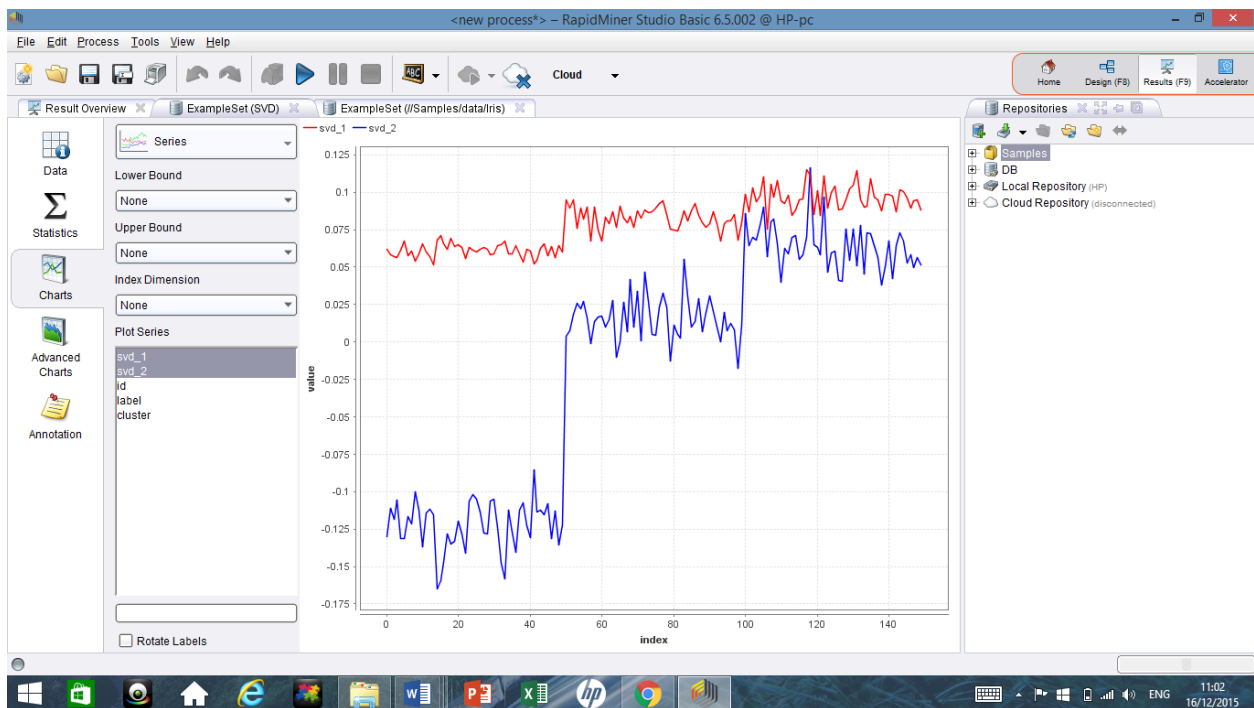
- ii. Draw a pie chart to show how many examples fall in each Iris plant group(type)



2. Describe the example set after using the SVD operator in terms of statistics and visualization
 - a. For the statistics, use the relevant statistics generated by RapidMiner



- b. For the visualization draw a series plot of the current dimensions.



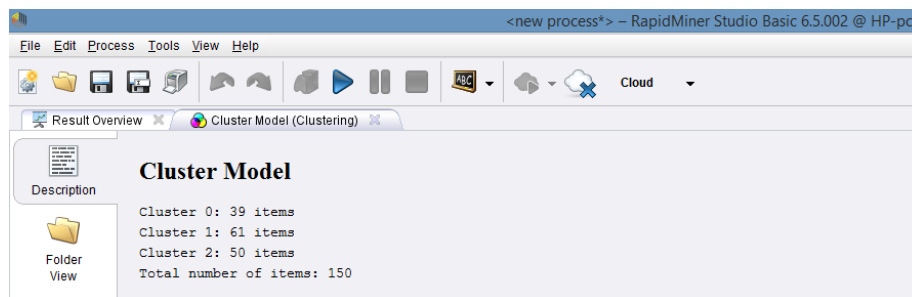
c. Comment on the graphs and statistics, describing what the SVD has done.

It simplified the DataSet by reducing the 4 attributes (a1, a2, a3, and a4) to only 2 attributes (svd_1, and svd_2). So, this reduction removes unnecessary attributes which are linearly dependent. Also, we can notice that there was some redundancy in those attributes since some of the attributes are correlated with another which was clear from the series plot of original data. So, this reducing will reduce the observed attributes and will enhance the total process.

3. Describe the clustering results of:

a. Clustering results of original data before using SVD

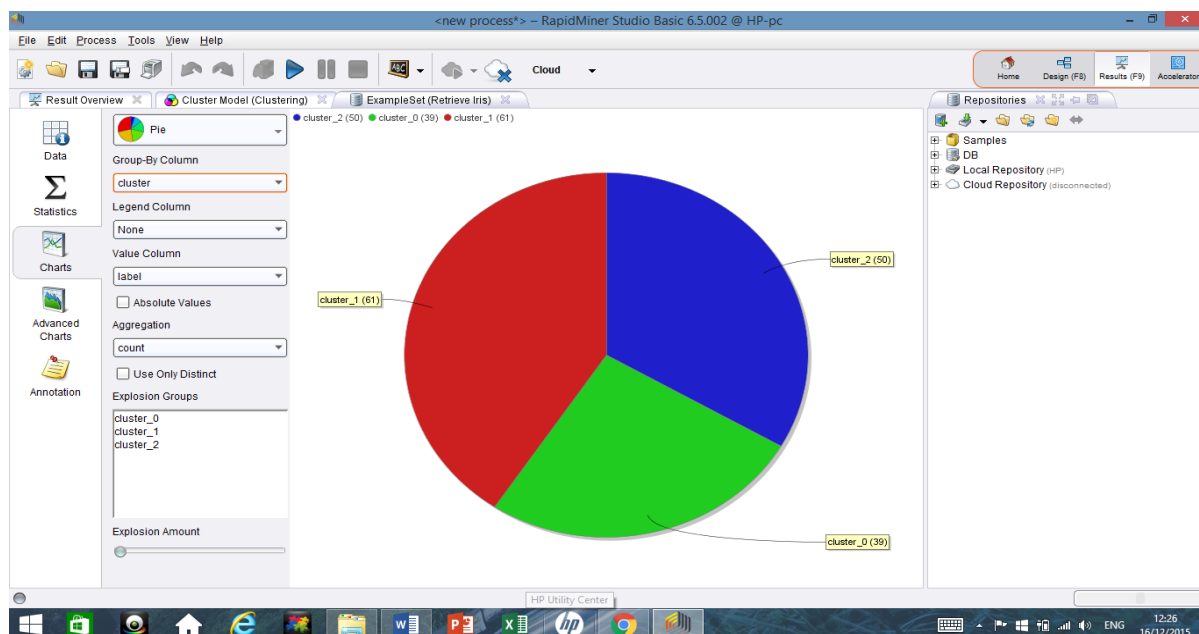
i. Statistics (eg. Nb of examples per cluster, Centroid Table)

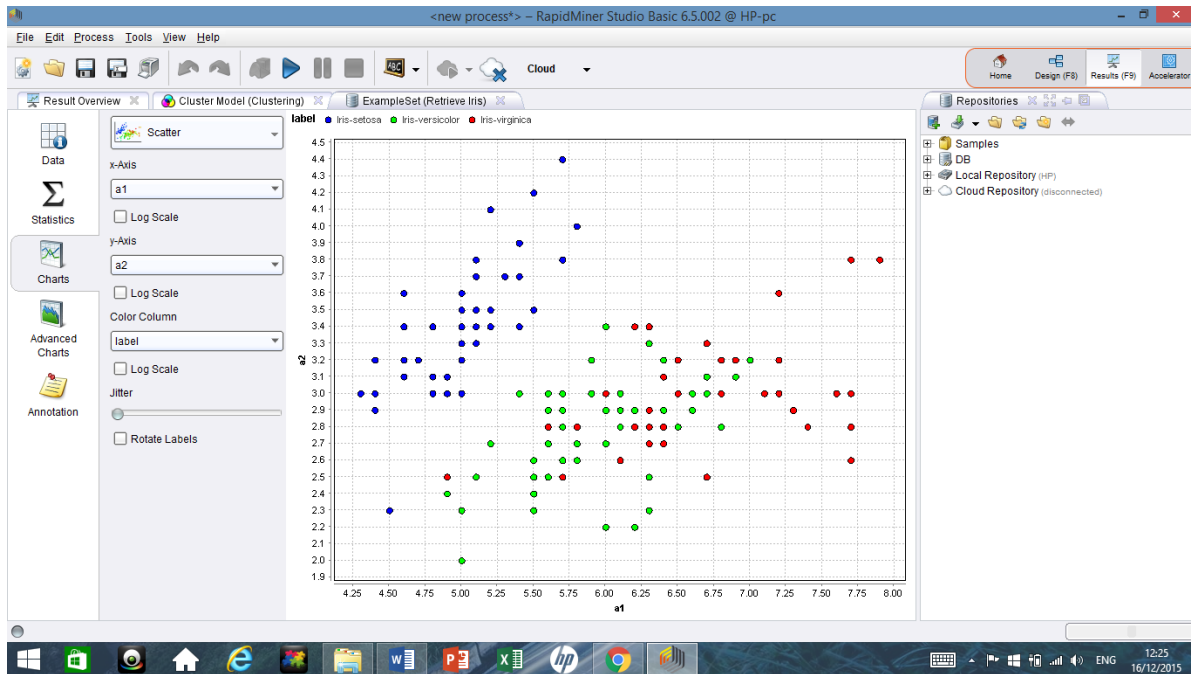


Centroid Table:

Attribute	cluster_0	cluster_1	cluster_2
a1	6.854	5.884	5.006
a2	3.077	2.741	3.418
a3	5.715	4.389	1.464
a4	2.054	1.434	0.244

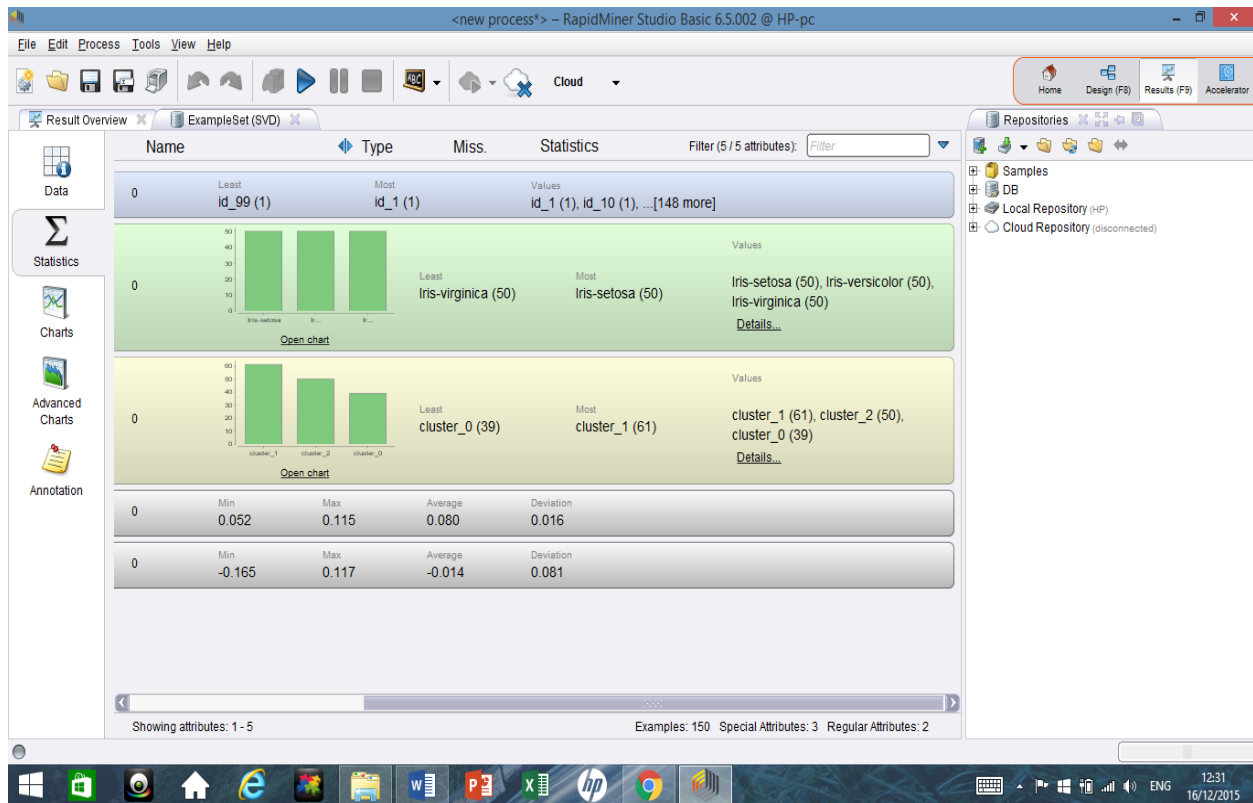
ii. Visualization (Pie chart and Scatter plots)



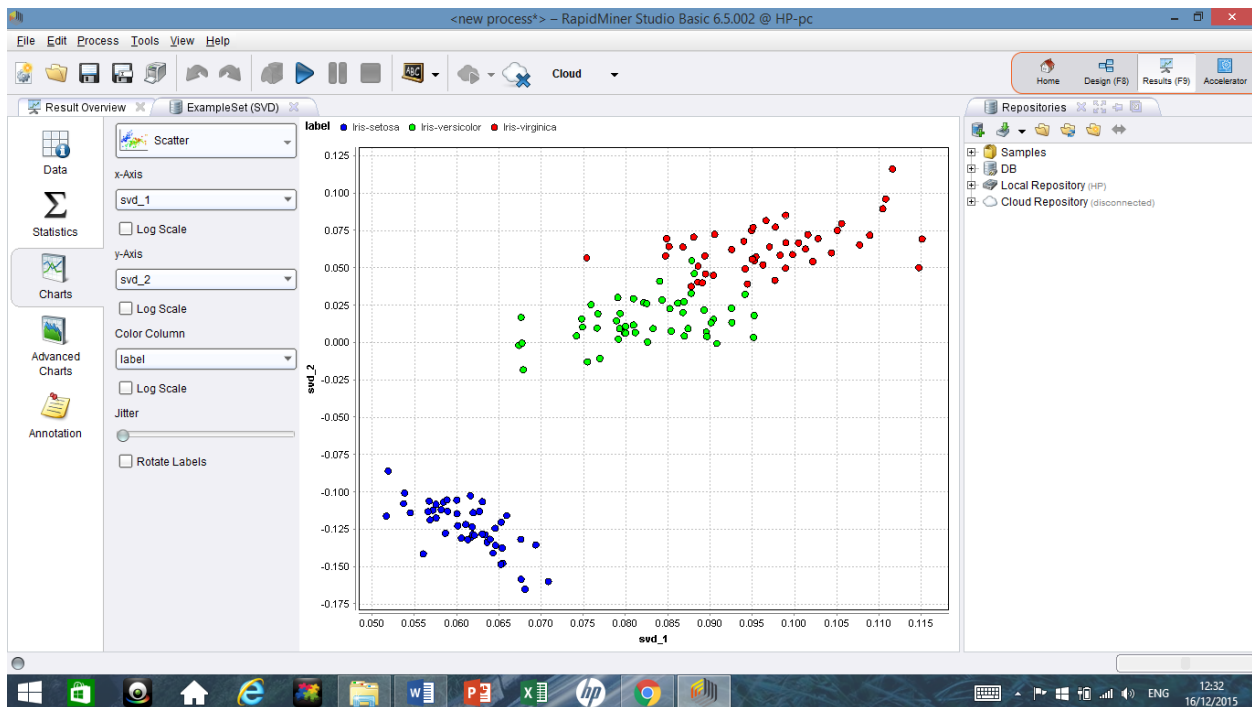
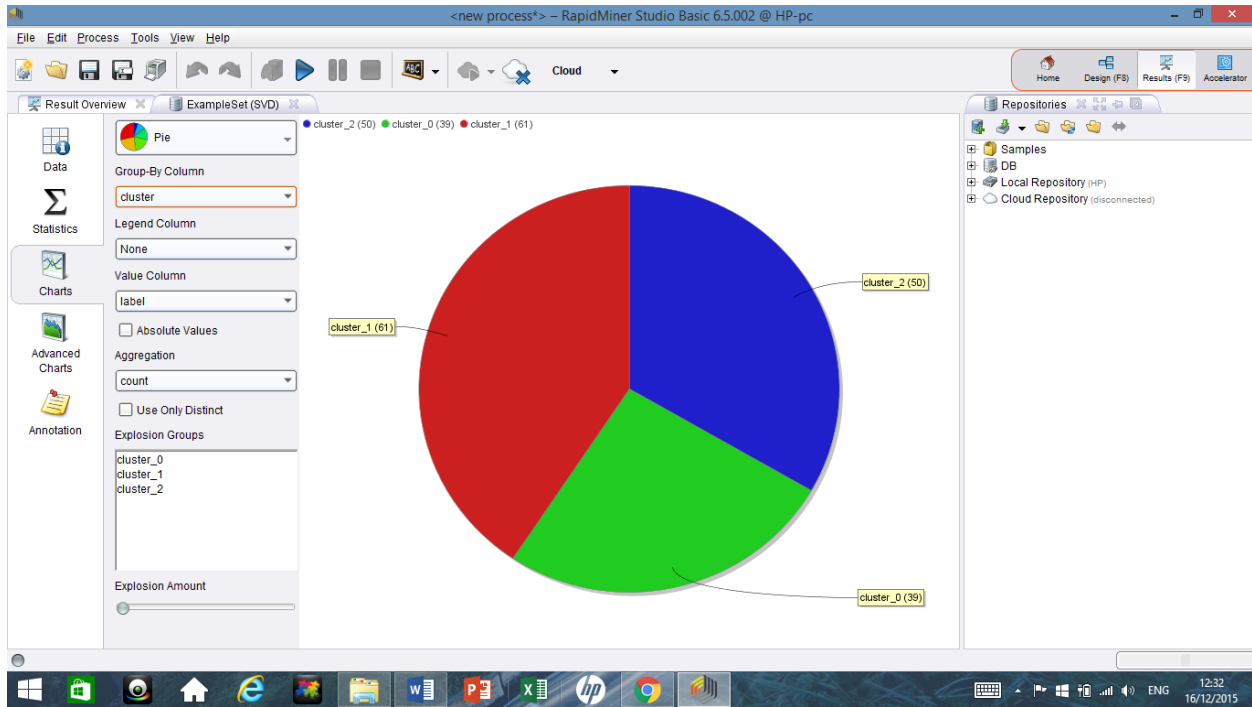


b. Clustering results after using SVD

i. Statistics (Nb of examples per cluster)



ii. Visualization (Pie chart and Scatter plot)



4. Comment on the Clustering results and how accurate they are, given the facts that you know about the data. You can also add any statistics or charts you wish, that you think would support your evaluation.

The clustering results are same whether before or after applying SVD. Results didn't change, there are 150 total examples as the original dataset with 50 examples in each class of Iris Plant, which are all clustered into 3 clusters:

Cluster_0 = 39 examples

Cluster_1 = 61 examples

Cluster_2 = 50 examples

By looking to the scatter plot, we can notice that one class is linearly separable from the other two, besides, the other two are not separable from each other. So, this means that the fact which is given at the question didn't change even after applying clustering model and after applying dimension reduction SVD.

=====