**St. Francis Institute of Technology, Mumbai-400 103**

**Department Of Information Technology**

**A.Y. 2024-2025**

**Class: TE-ITA/B, Semester: VI**

**Subject: Business Intelligence Lab**

**Experiment – 7: To implement K-means clustering algorithm using open source tools, WEKA & ORANGE**

1. **Aim: :** To implement K-means clustering algorithm using open source tools, WEKA & ORANGE
2. **Objectives:** After study of this experiment, the students will be able to

                              Implement K Means

1. **Outcomes:** After study of this experiment, the students will be able to

**CO 4:** Design andImplement various clustering data mining techniques such as Partitioning methods, Hierarchical Methods, Density - Based methods along with identification and analysisof outliers.

1. **Prerequisite:** Introduction to all the three clustering algorithms & Problem solving approach.
2. **Requirements:** Personal Computer, Windows XP operating system/Windows 7, Internet Connection, Microsoft Word, WEKA tool, ORANGE tool.
3. **Theory:**
4. **Explain K means (graph) algorithm.**

**ANS:**

K-Means is a centroid-based clustering algorithm that partitions data into KKK clusters based on similarity. The goal is to minimize the variance within clusters.

#### Steps of K-Means Algorithm:

1. Initialize Centroids: Randomly select KKK initial cluster centers (centroids).
2. Assign Points to Clusters: Each data point is assigned to the nearest centroid using the Euclidean distance.
3. Update Centroids: Compute the mean of all points in each cluster and set it as the new centroid.
4. Repeat: Steps 2 and 3 are repeated until centroids do not change significantly or a stopping condition is met.

Graph Representation of K-Means:

1. Data points are plotted in a coordinate space.
2. Cluster centroids are marked and updated iteratively.
3. Boundaries between clusters emerge as centroids shift.

#### Limitations of K-Means:

1. Sensitive to outliers.
2. Requires specifying KKK beforehand.
3. Assumes clusters are spherical and of similar sizes.

**2. Explain K-medoids algorithm.**

**ANS:**

K-Medoids is a clustering algorithm similar to K-Means but is more robust to noise and outliers. Instead of using the mean, it selects actual data points (medoids) as cluster centers.

#### Steps of K-Medoids Algorithm:

1. Initialize Medoids: Choose KKK random data points as medoids.
2. Assign Points to Medoids: Each data point is assigned to the closest medoid.
3. Update Medoids:
   1. For each cluster, swap the medoid with a non-medoid data point and compute the cost (sum of distances).
   2. If the new medoid reduces the cost, update the medoid.
4. Repeat: Continue iterating until no further improvement occurs.

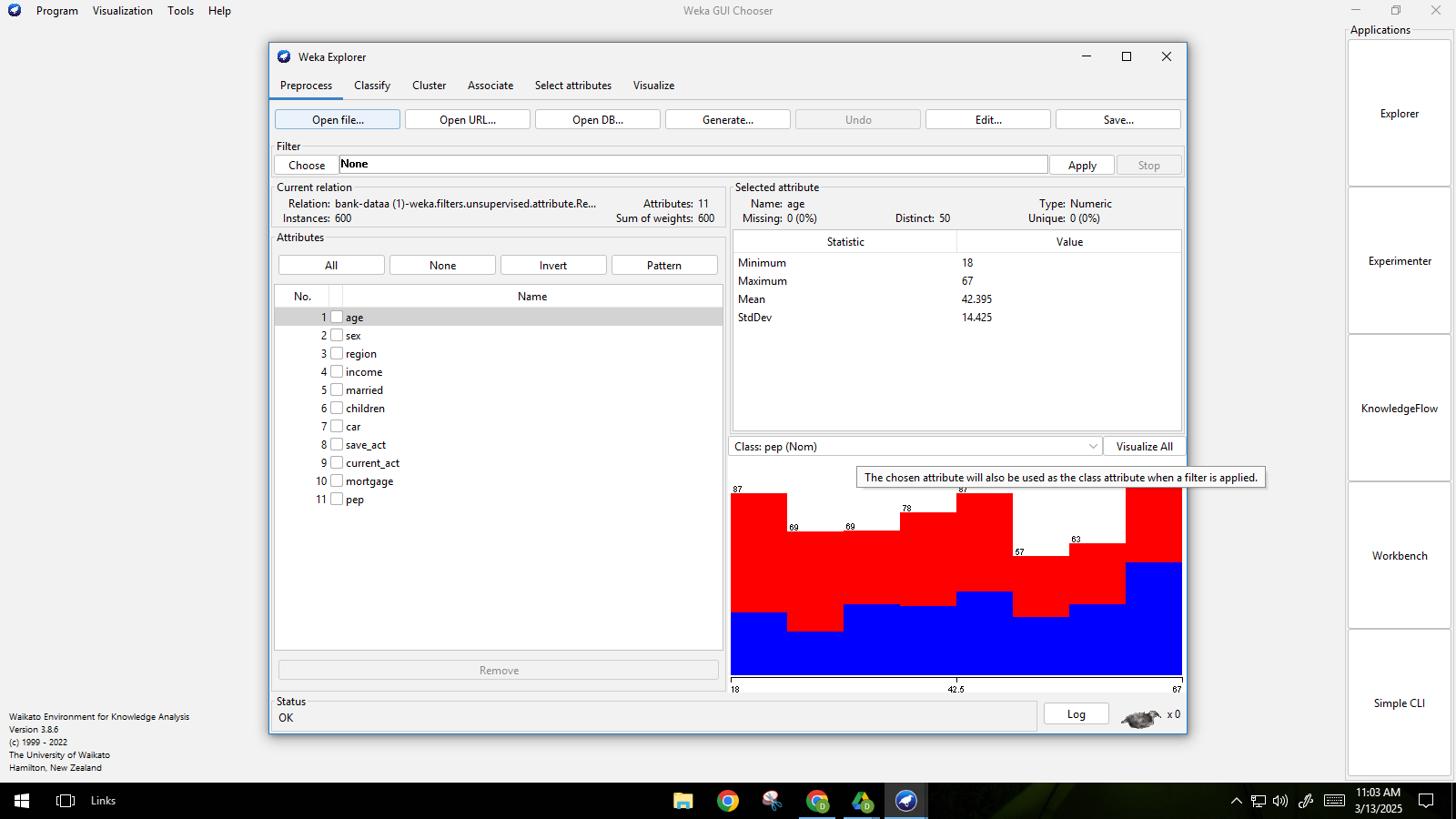
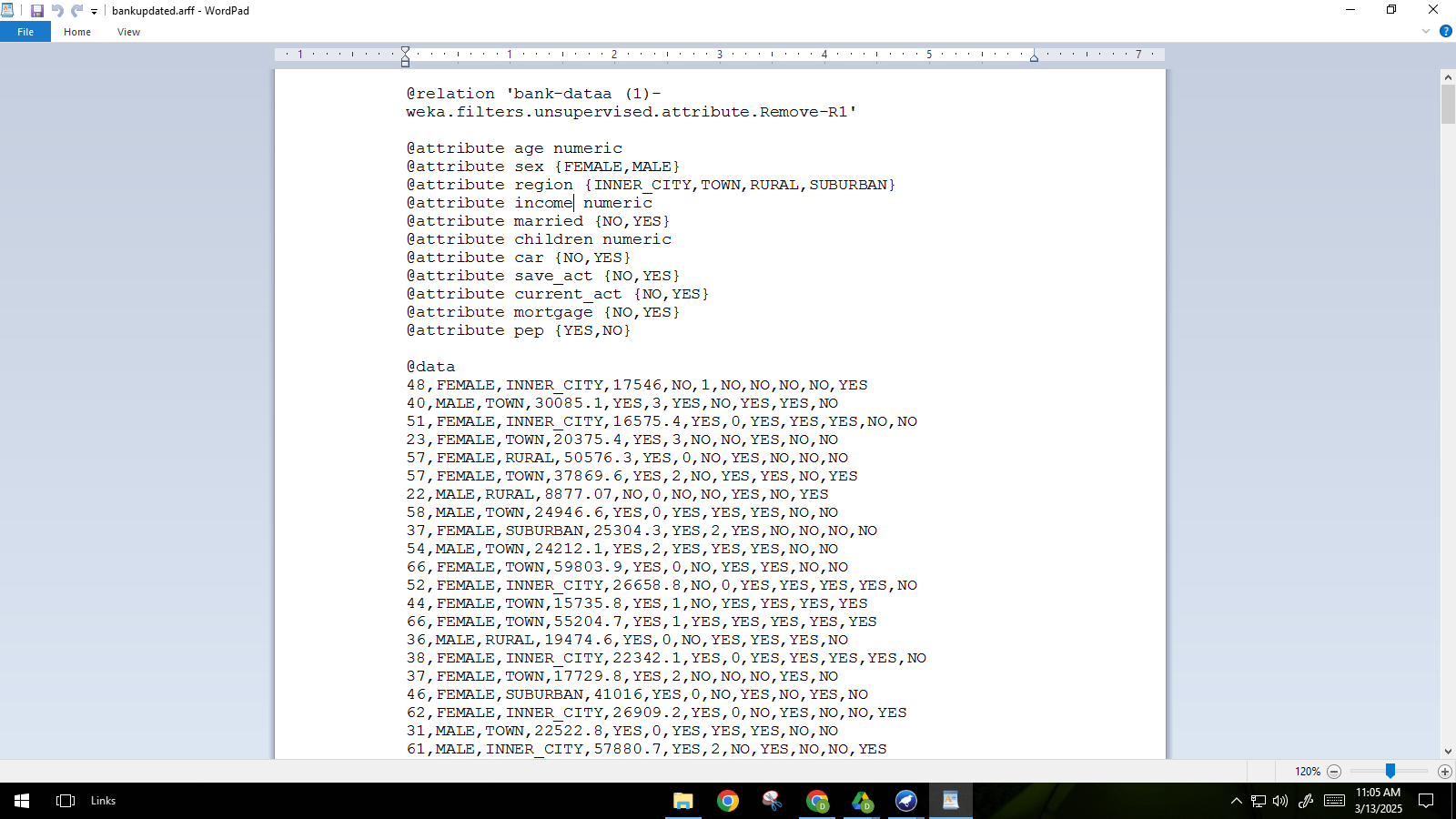
#### Advantages of K-Medoids Over K-Means:

1. More robust to outliers and noise.
2. Does not require the assumption of spherical clusters.
3. Can handle non-convex clusters better.

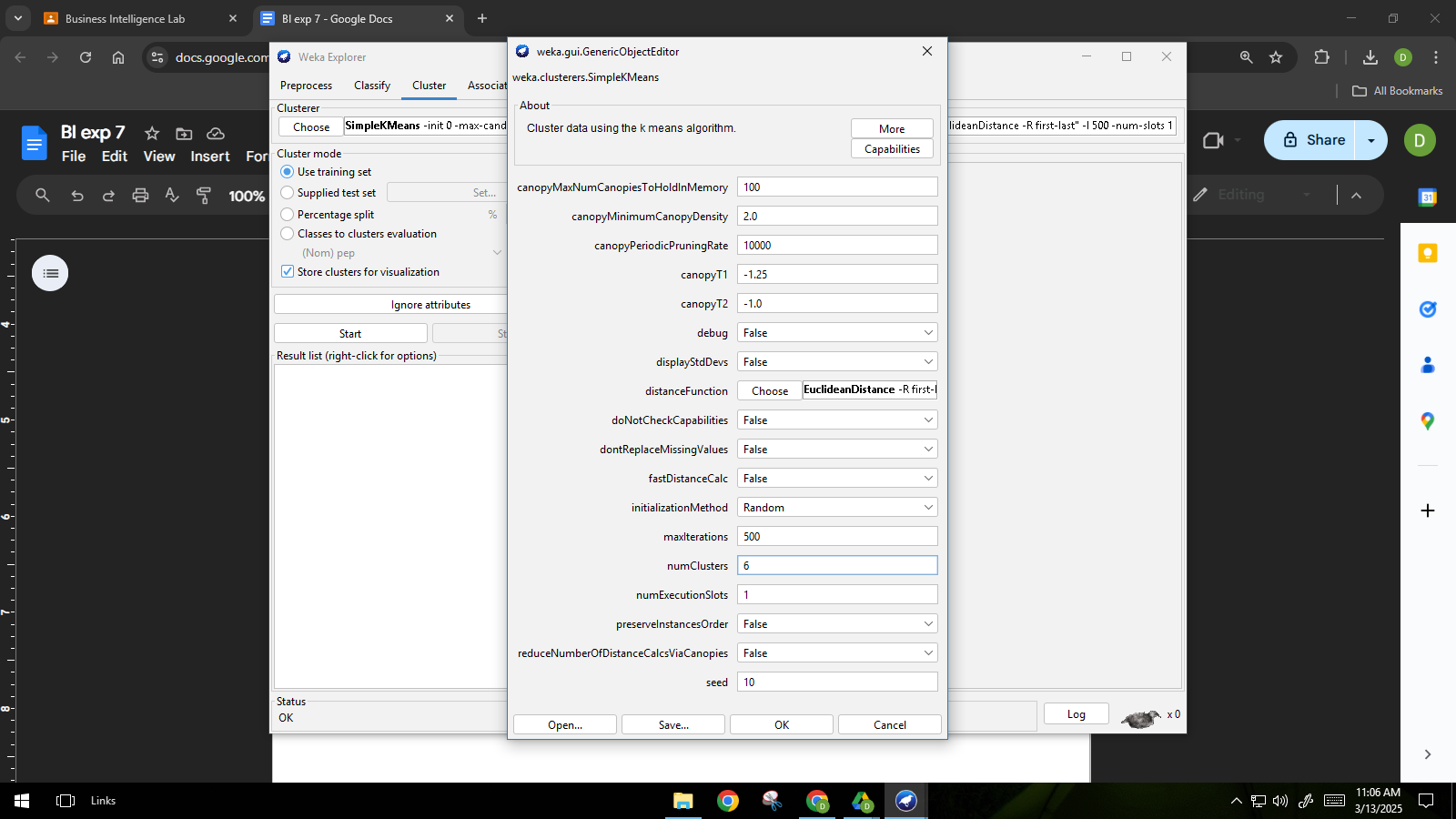
#### Limitations:

1. Computationally more expensive than K-Means.
2. Convergence may be slow for large datasets.
3. **Laboratory Exercise:** Implementation of K means clustering algorithm using WEKA & ORANGE along with screenshots.
4. **Post-Experiments Exercise**
5. **Questions:**
   * Explain advantages and disadvantages of K means
   * K means (graph) solved numerical
6. **Conclusion:**
   * Summary of Experiment
   * Importance of Experiment

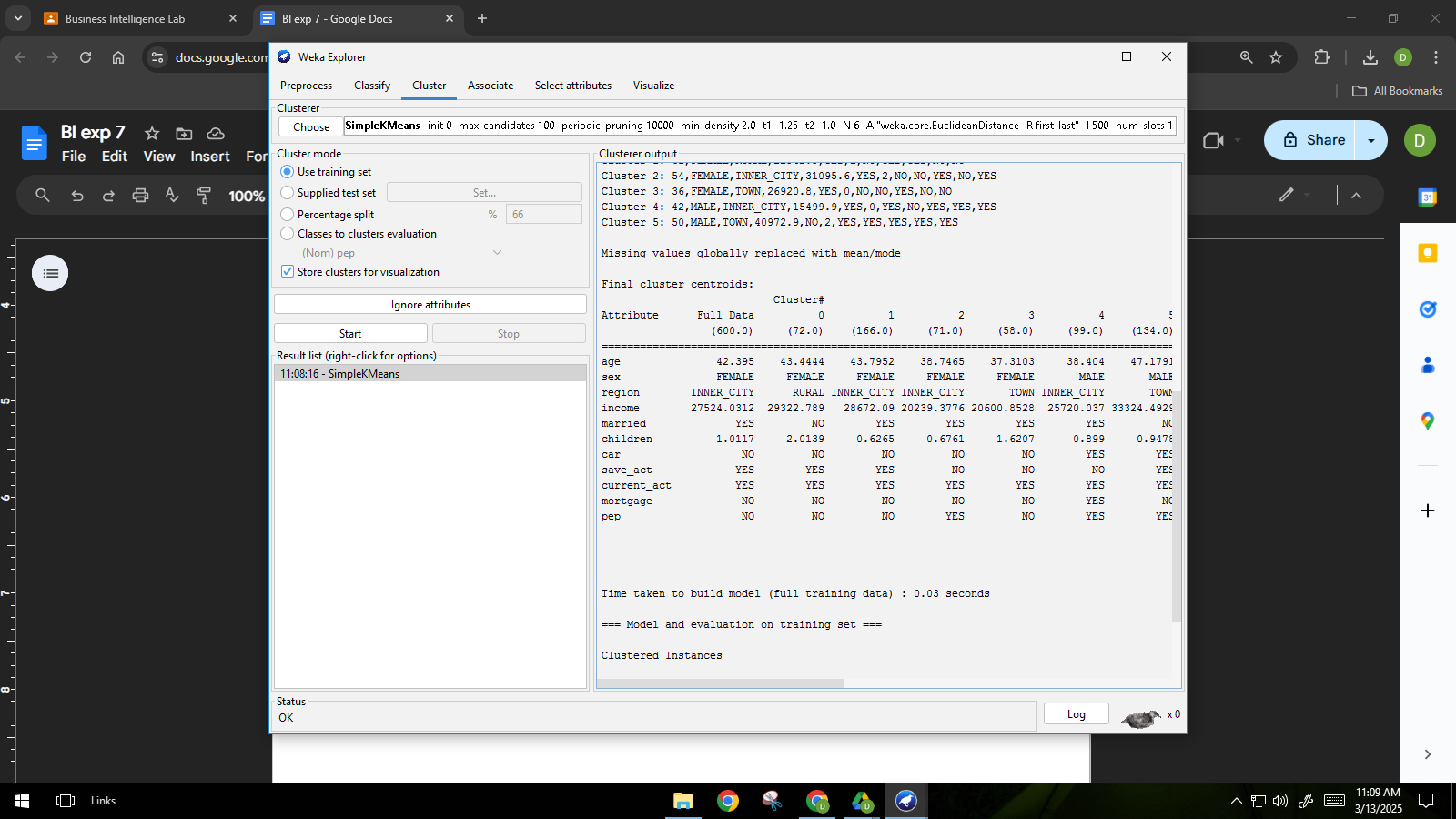
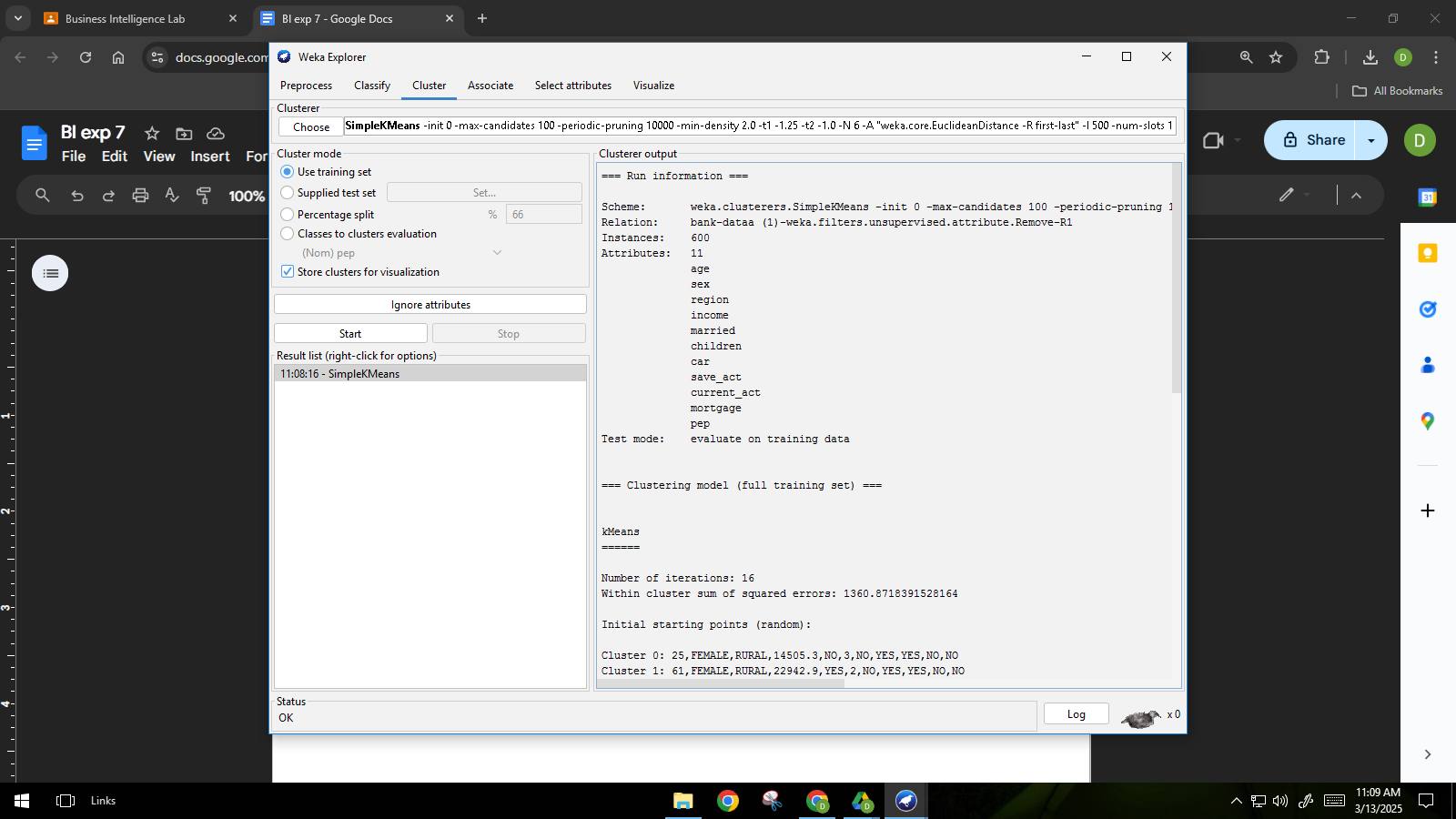
opening file in weka pre-processed dataset

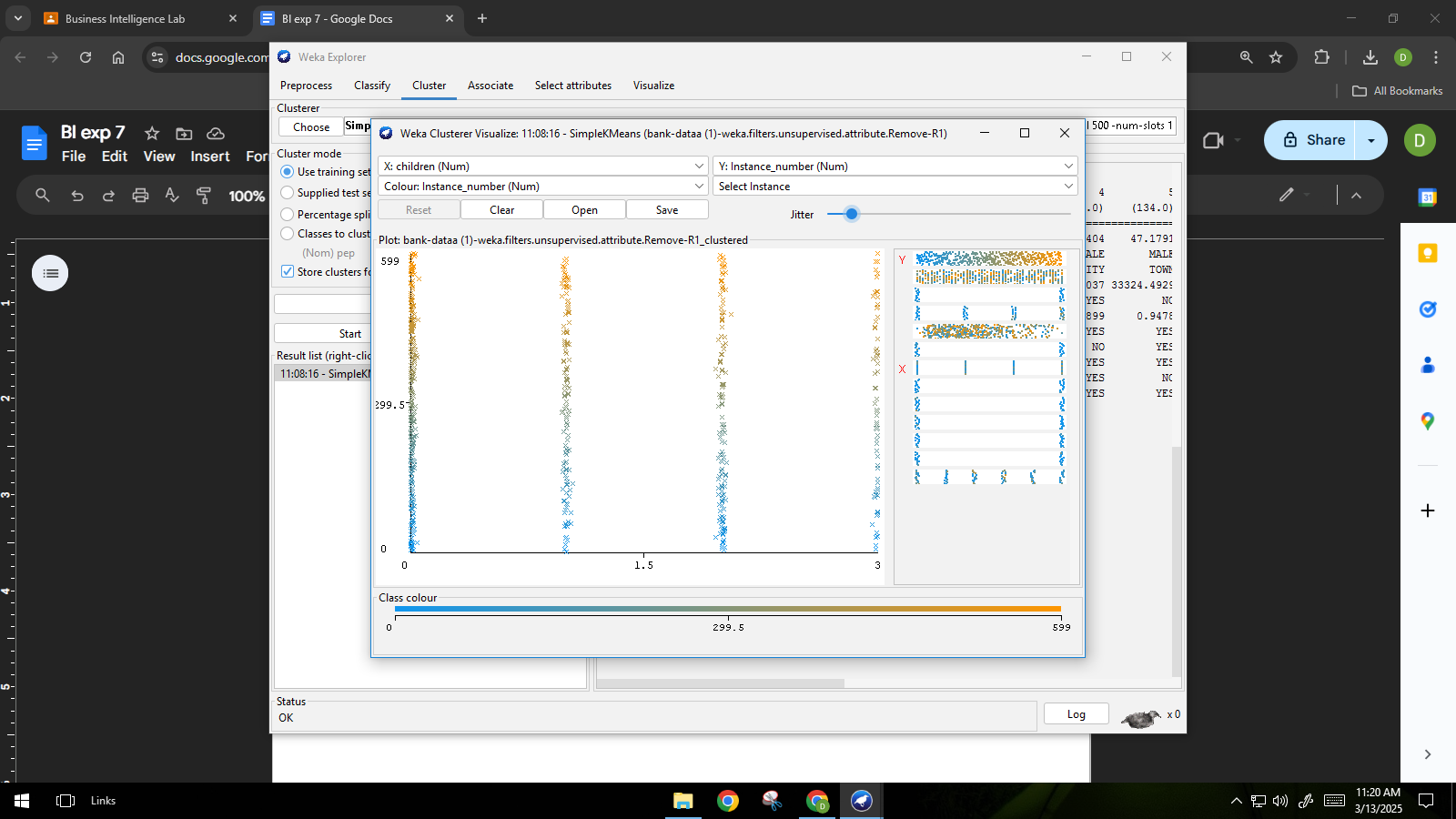
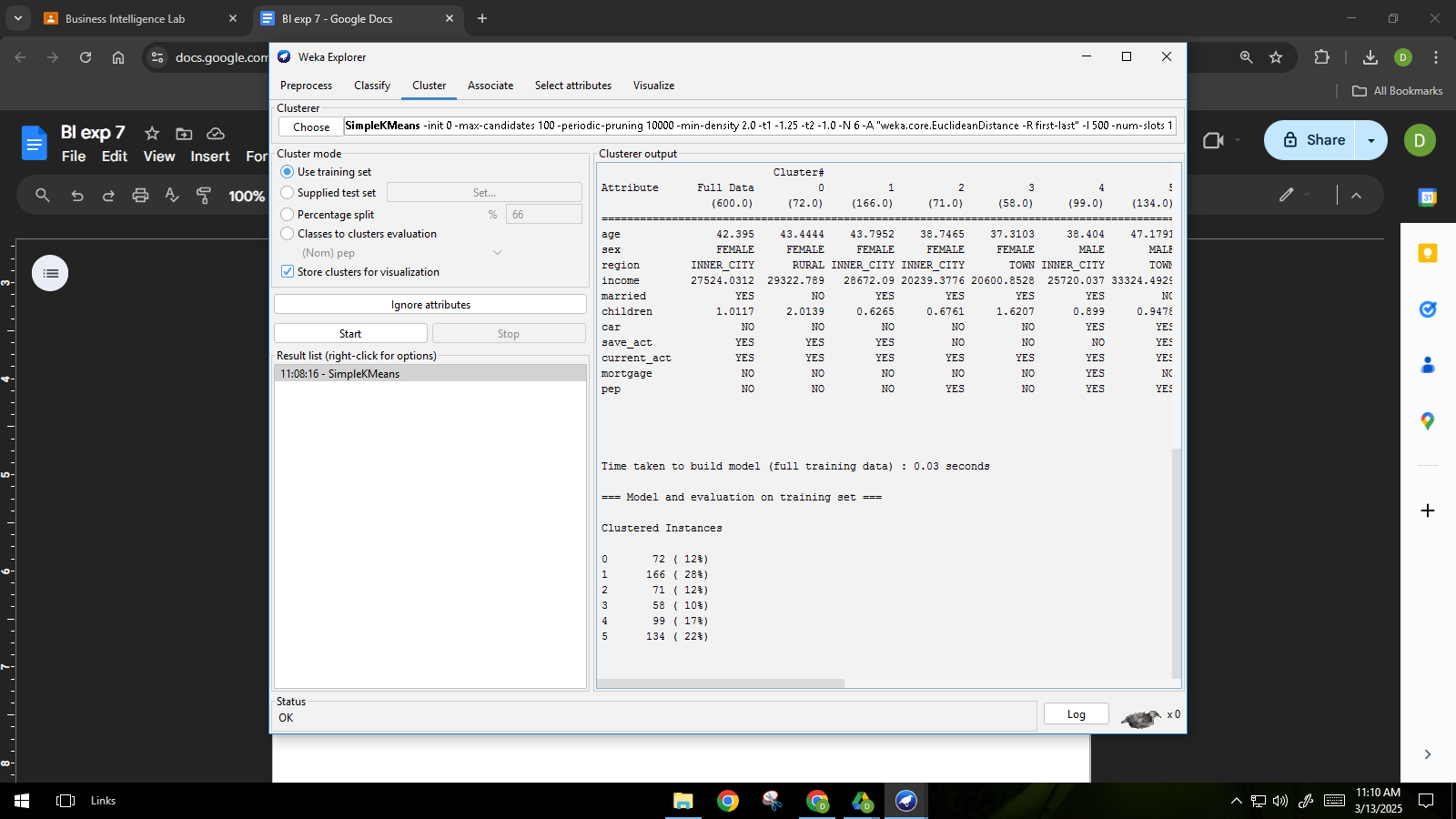
applying k-means on the preprocessed dataset & number of clusters set as 6.



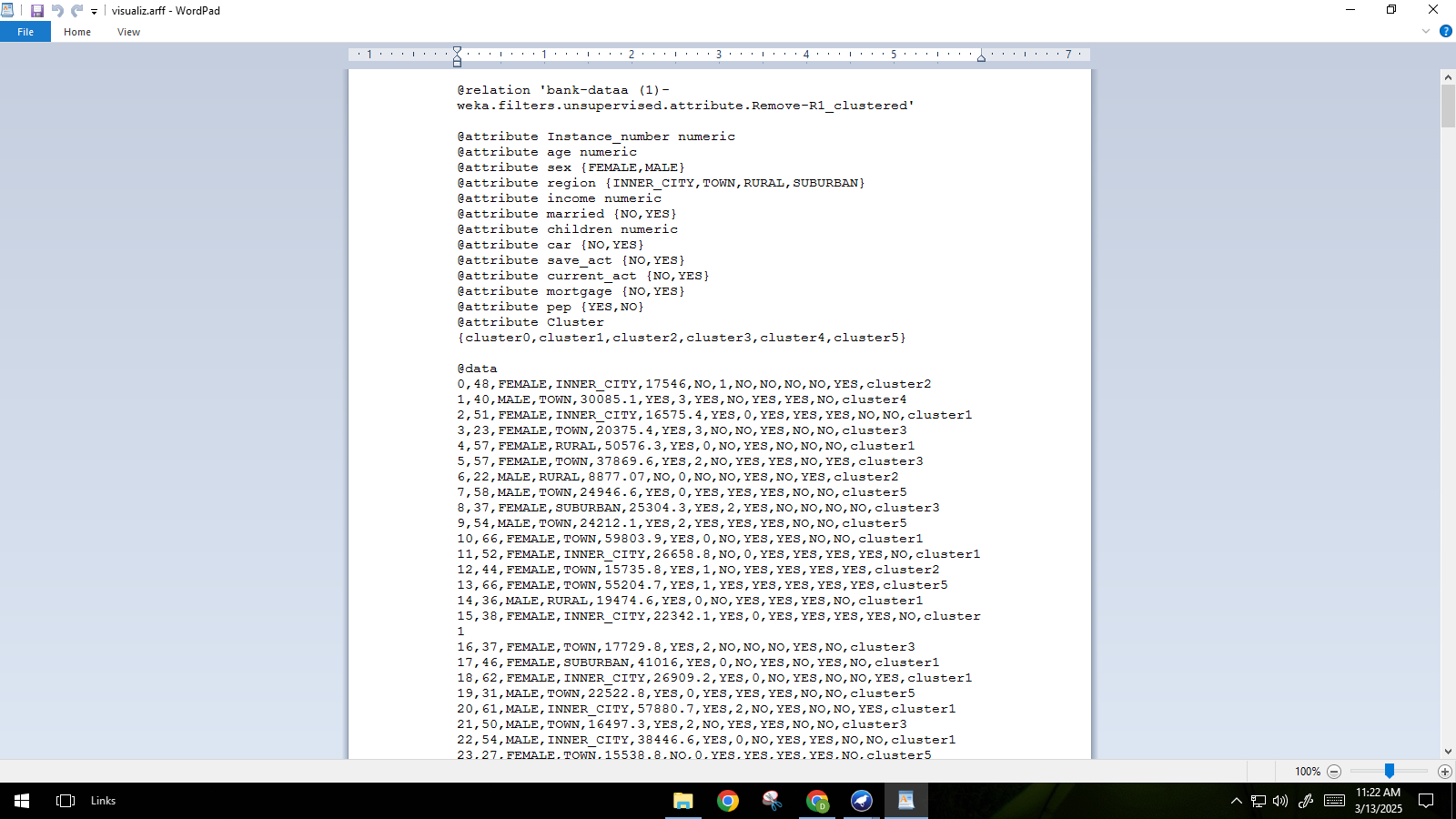
Result of k-means in WEKA:



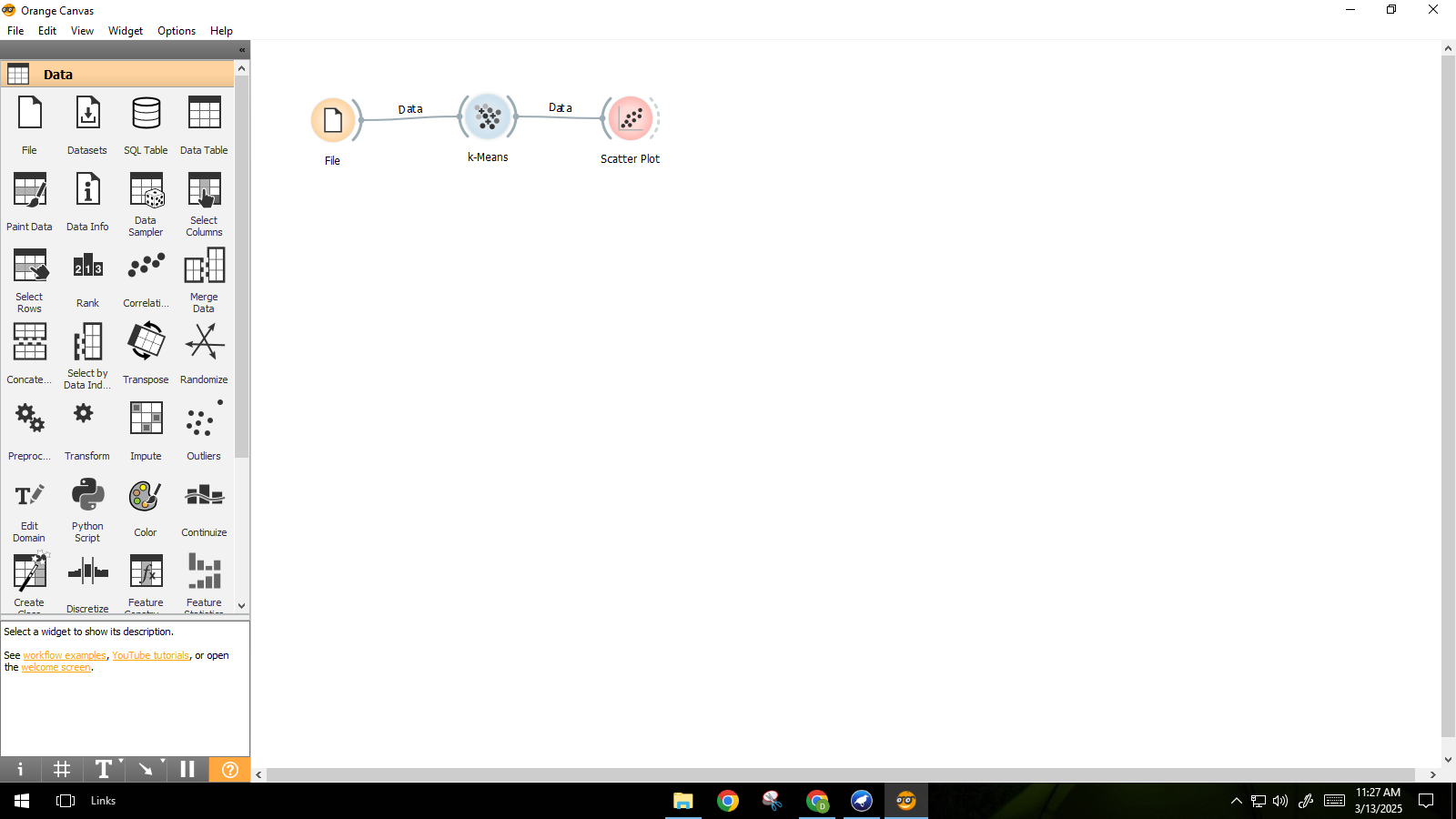
Visualization using tool



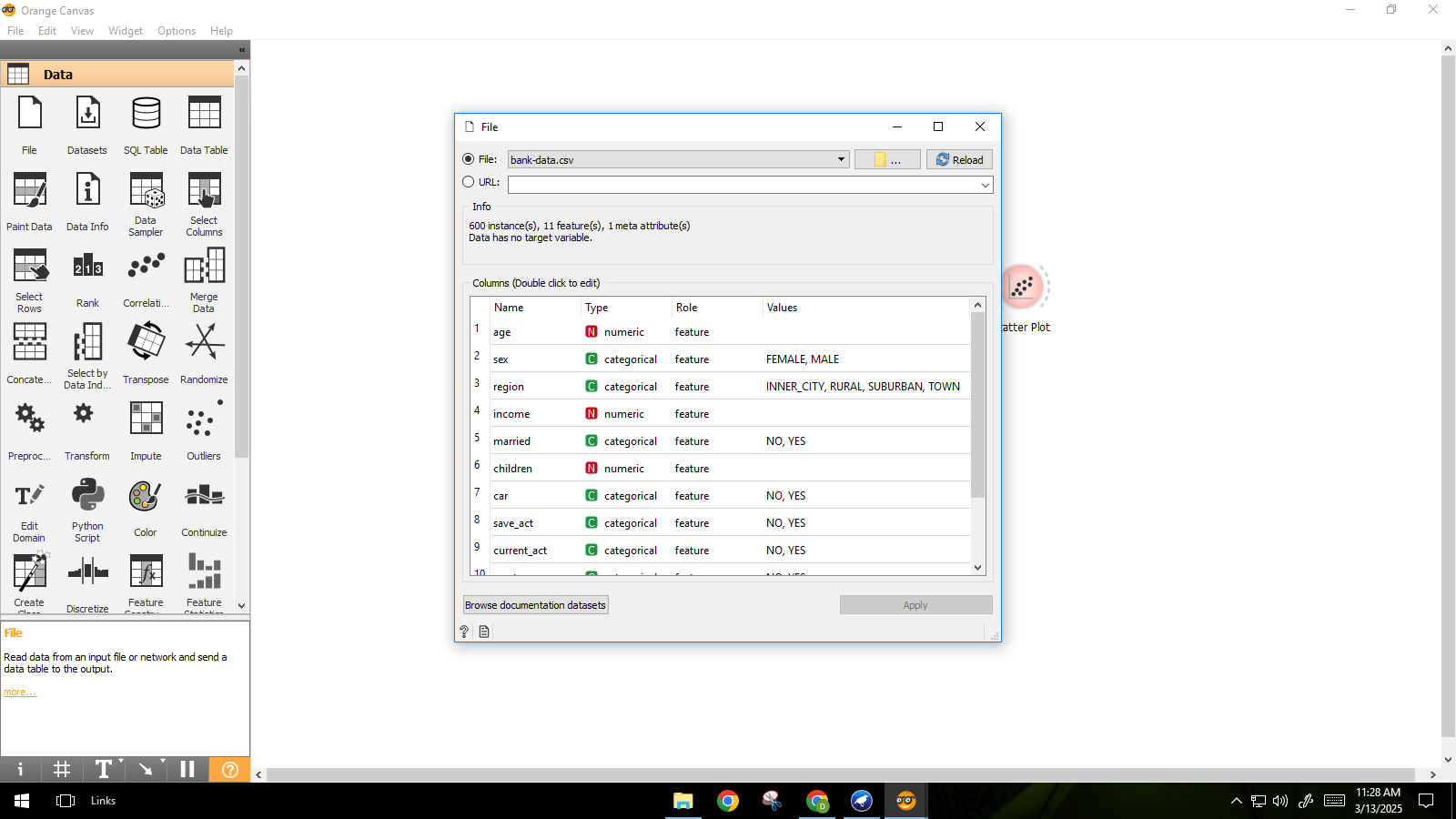
Clustered dataset:



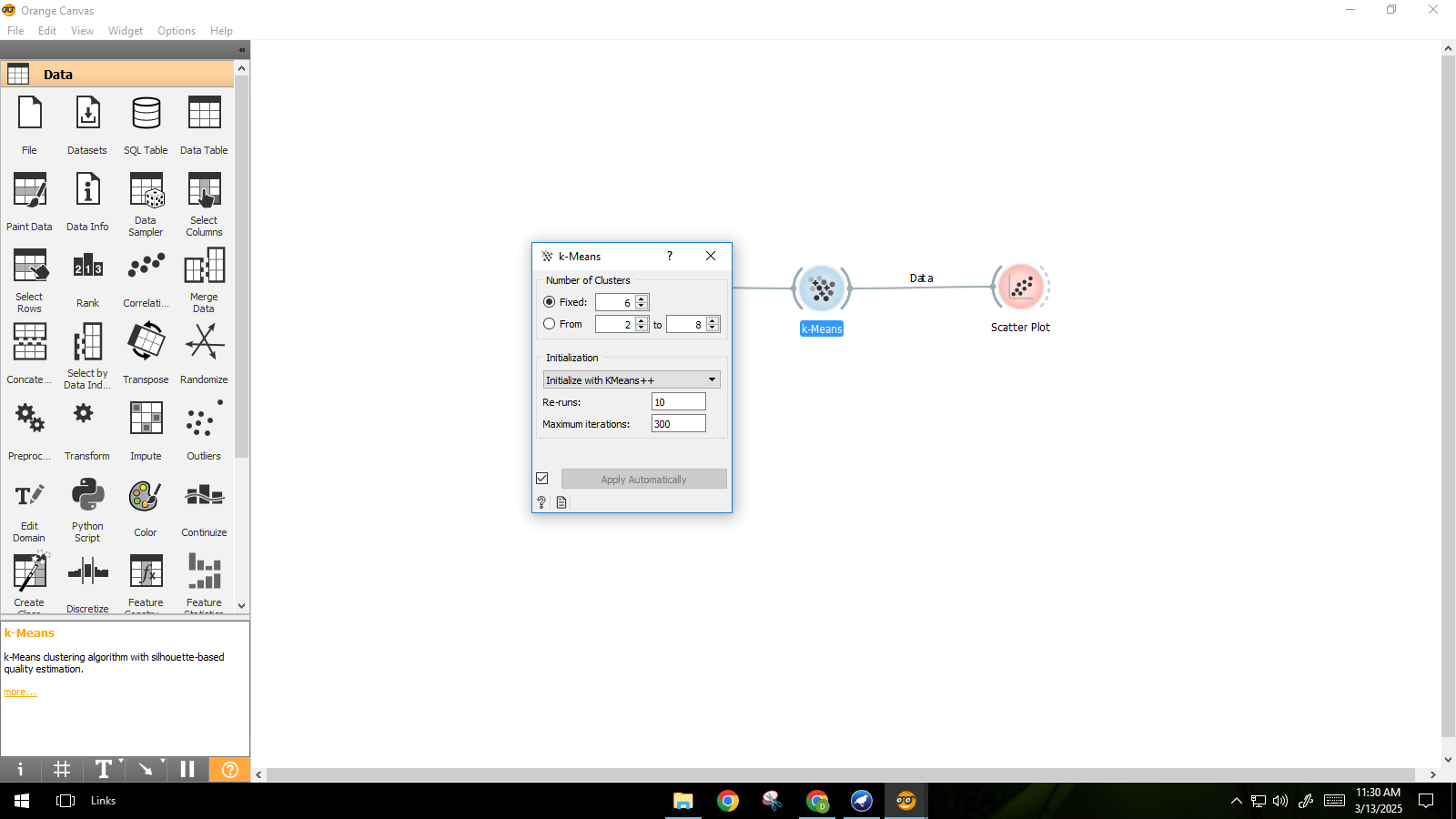
Applying K-means clustering in ORANGE:



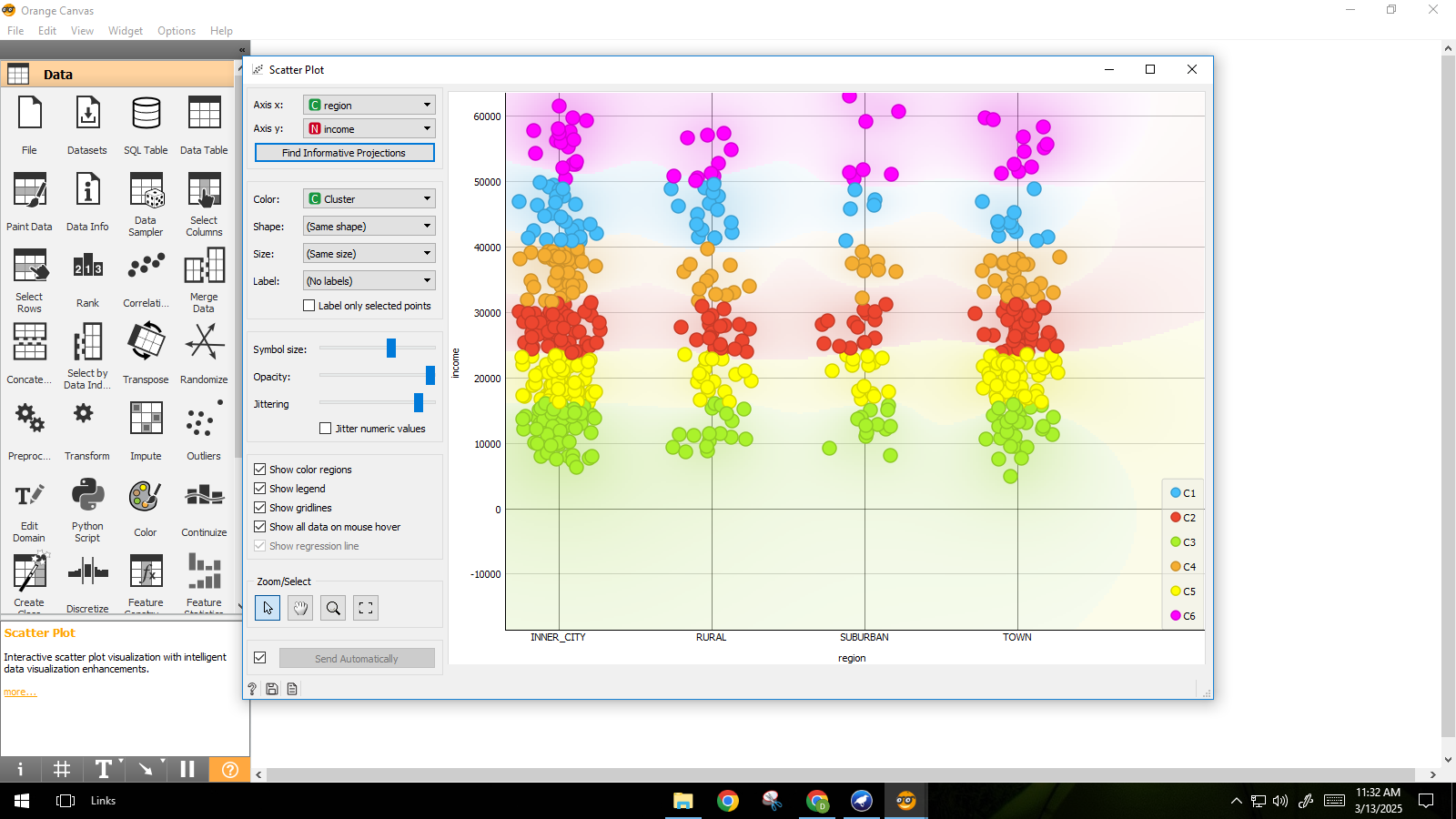
Loading the dataset in ORANGE:



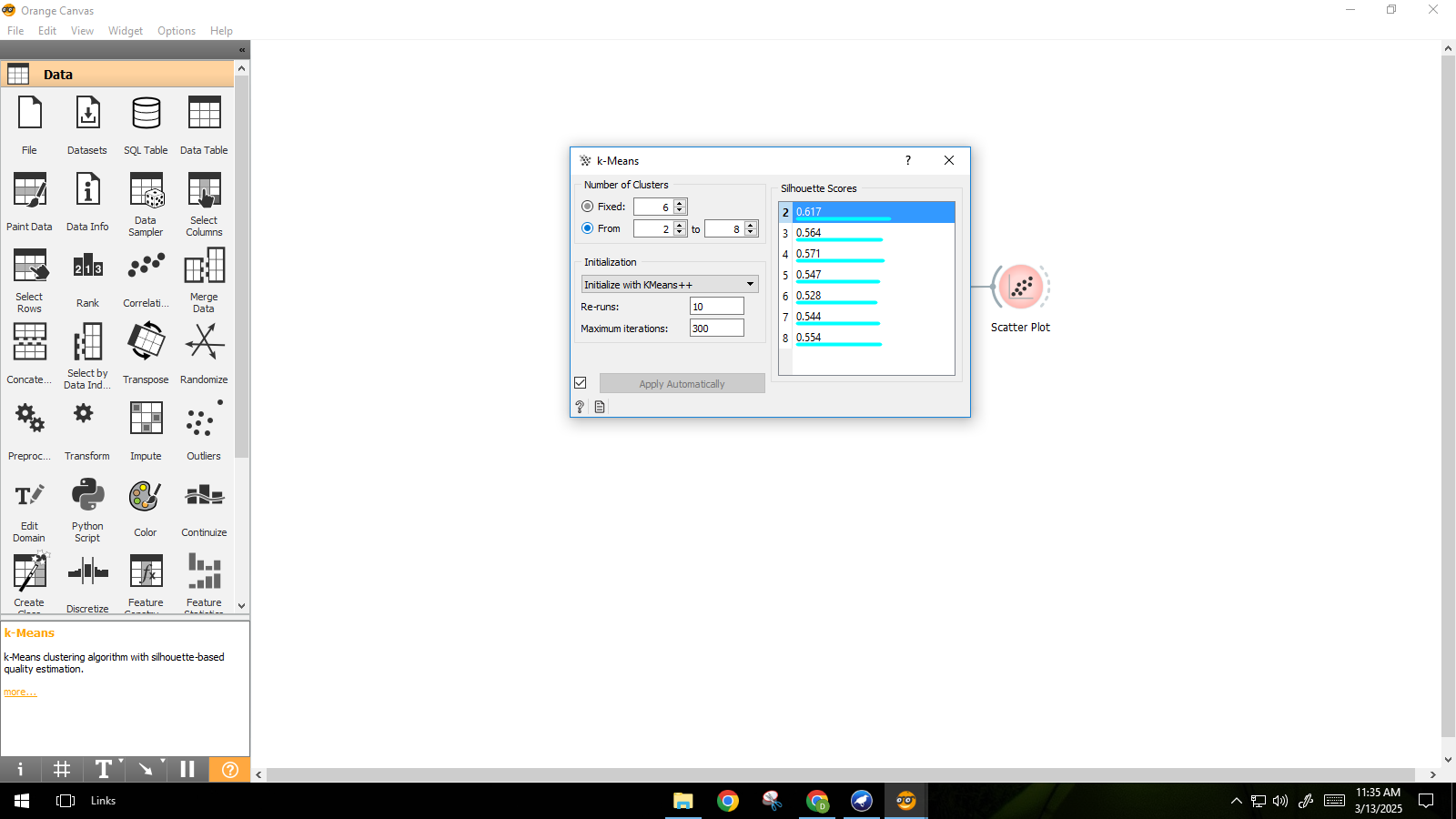
Applying K-means clustering with number of clusters as 6 in orange.



Observing the 6 clusters through scatter plot.



Deciding on the K-value based on silhouette score:



| Observing the 2 clusters generated in the scatter plot: | Applying K-means on painted data: |
| --- | --- |
| Painting data in Paint Widget: | Applying K-means for this painted data:    Unable to cluster the points correctly for  a non-spherical shape: |