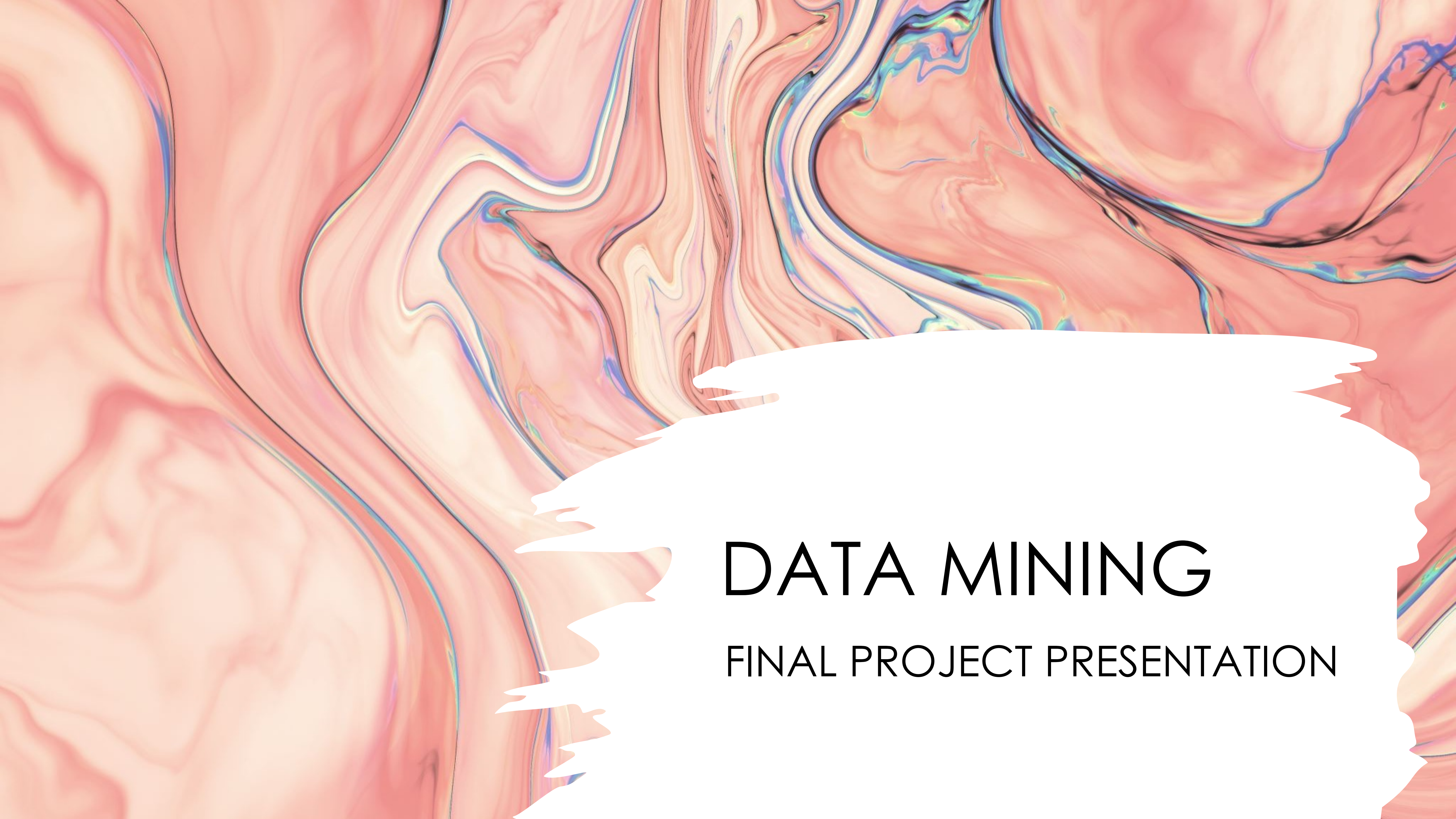


DATA MINING

FINAL PROJECT PRESENTATION

A presentation on Final Project.

- Data Mining
  (ALY 6040)

- **Guided by:**
  Prof. Sean P. Cornelius

- **Submitted by**:
  Pratikkumar Malaviya
  Taiye Murtala
  Benedict Eze

- **Date of submission :**
  26th October’2022

Northeastern University

AGENDA

DATASET

- Information of Dataset

DATA ANALYSIS

- Techniques
- Insights
- Interpretation

SUMMARY

DATASET

- Employee attrition is referred to as **the natural process of employees leaving the workforce** in the dataset.

- It contains thirty-five **(35) columns** and one thousand four hundred and seventy columns **(1470) rows** comprising only integer and character datatypes.

- Explore significant questions like, "Show me a breakdown of distance from home by job function and attrition," or "Compare average monthly pay by education and attrition" to learn the causes of employee attrition.

- The dataset contains the **age, education levels, job satisfaction, distance from home, employee performance rating, and work-life balance,** among other factors that lead to employee attrition.

- Our reason for selecting this dataset was to **understand the reason for employee turnover** based on the factors contained in the dataset and make possible suggestions as to how the company can retain its employees.

Source : https://www.kaggle.com/datasets/whenamancodes/hr-employee-attrition

TECHNIQUES

01

Clustering

- Finding groups of things known as clusters is the process of grouping objects so that they are related to one another and distinct from one another or from other groups.
- In this type of exploratory data analysis, observations are classified into groups based on traits they have in common.

02

Support Vector Machine

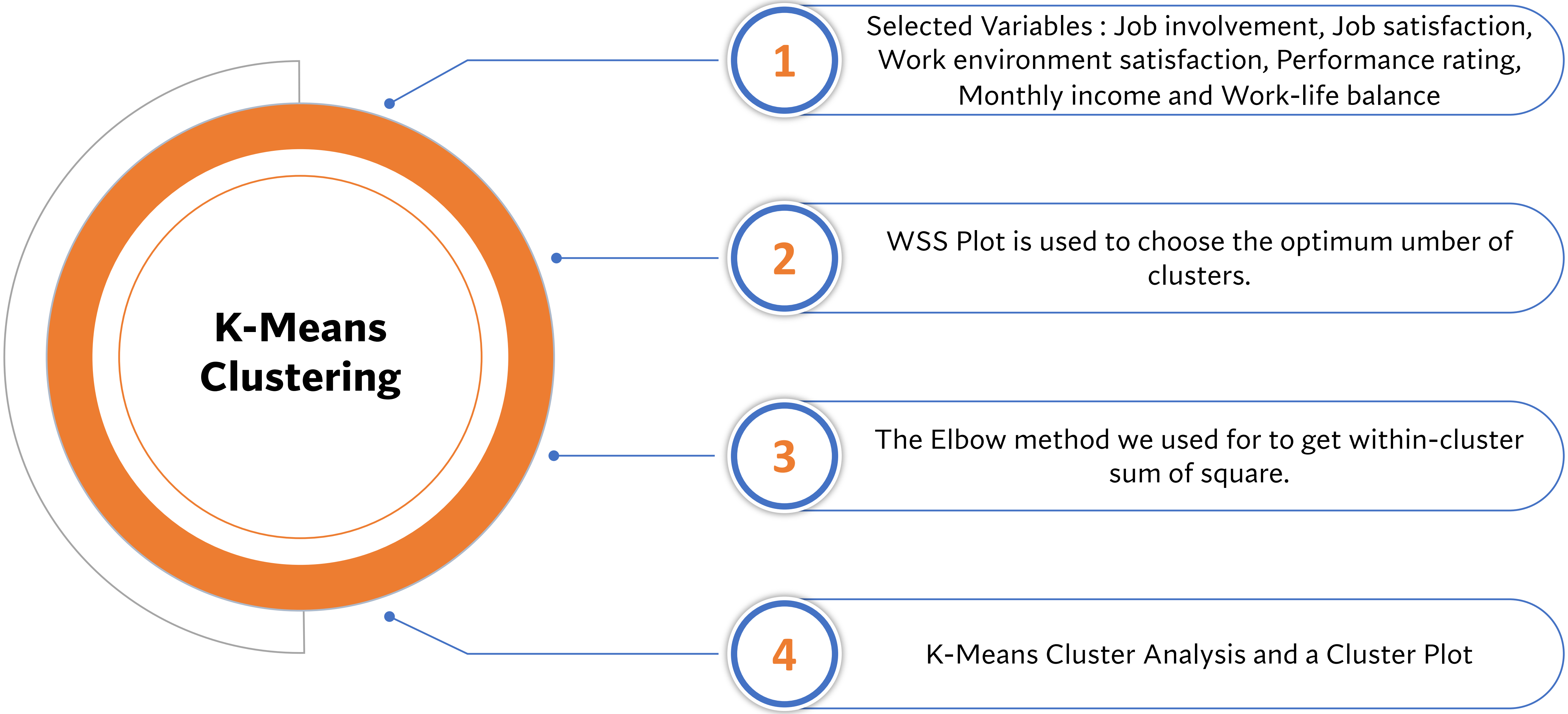
- Support Vector Machine (SVM) is a Supervised Machine Learning algorithm that is used for regression and classification.
- It is more preferred for classification but is sometimes very useful for regression as well.

03

Hierarchical Clustering

- The ideal number of clusters is established using hierarchical clustering. The dendrogram can be used to identify this ideal number of clusters.

TECHNIQUE -1



K-Means Clustering

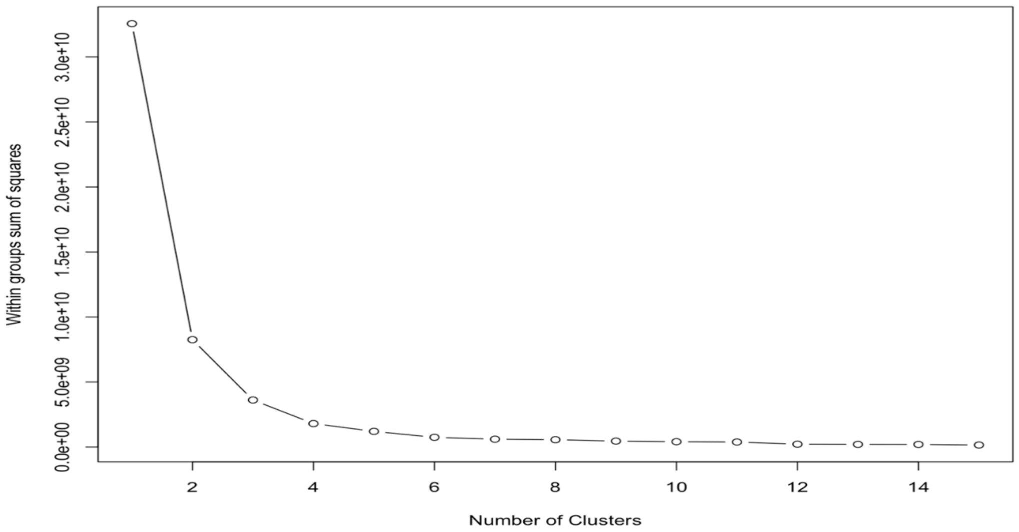
1 Selected Variables : Job involvement, Job satisfaction, Work environment satisfaction, Performance rating, Monthly income and Work-life balance

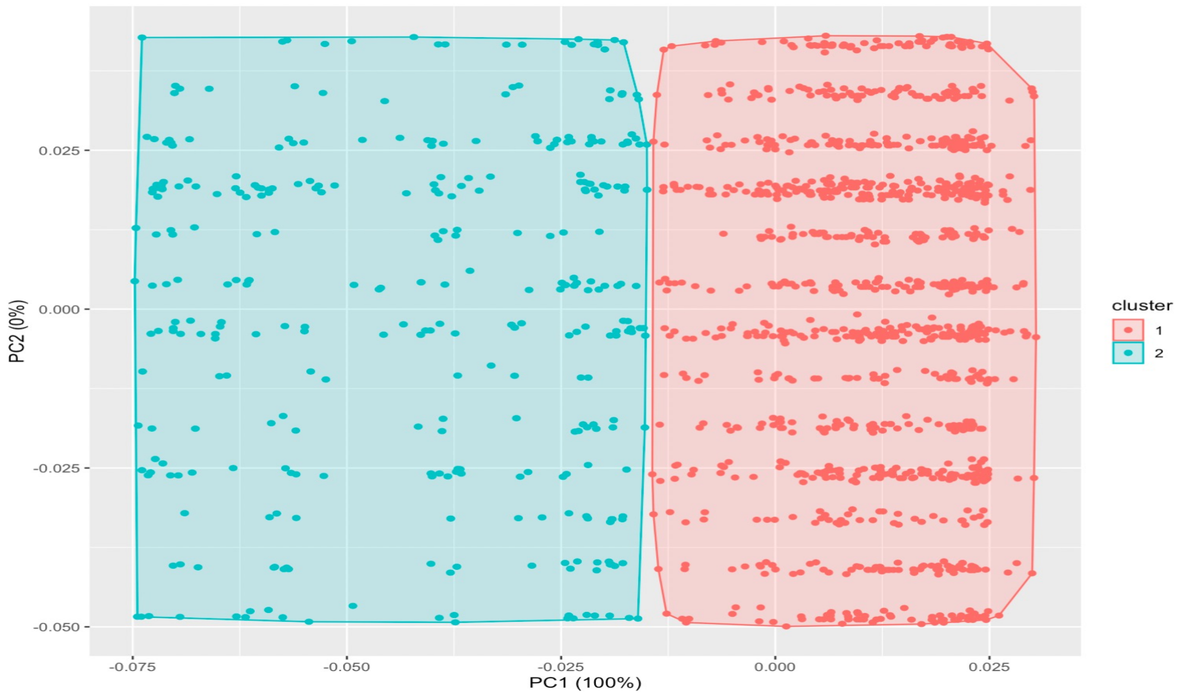
2 WSS Plot is used to choose the optimum umber of clusters.

3 The Elbow method we used for to get within-cluster sum of square.

4 K-Means Cluster Analysis and a Cluster Plot

INTERPRETATION





The optimum number is the point where there is an elbow shape in the plot. From the plot besides, the elbow shape is visible when k=2.

After obtaining the optimum number of clusters (2). I performed a k-means analysis of my unlabelled dataset with the optimum number of clusters (2).

Cluster 1 is the red dots and 2 is the blue dots. It is evident that these two clusters are distinct.

There isn’t any overlapping hence, the cluster analysis has been successfully deployed.

Employees in cluster 2 earn more as compared to employees in cluster 1 although they all showed similarities in performance indicators

TECHNIQUE -2

Support Vector Machine (SVM)

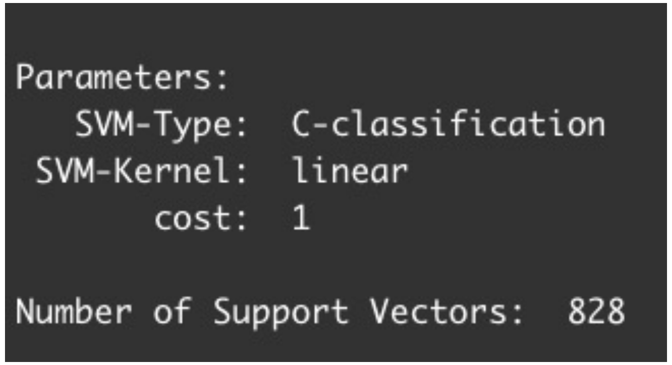
1 Selected Variables : Entire Dataset

2 In order to prepare model, the dataset has been divided into two portion Training and Testing.

3 Here, the Training set is for developing model and Testing dataset is to test the performance of model.

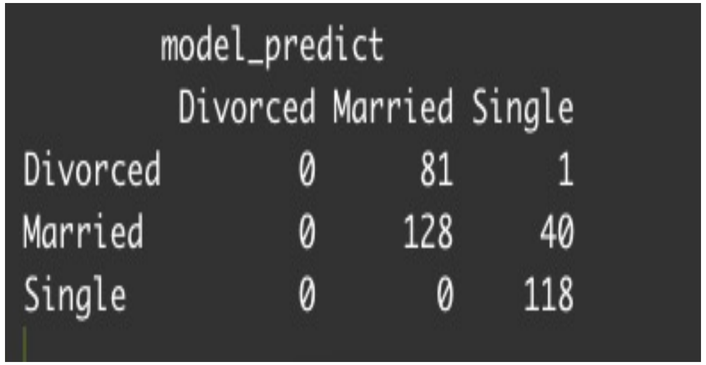
4 Training : 70% and Testing 30%

INTERPRETATION

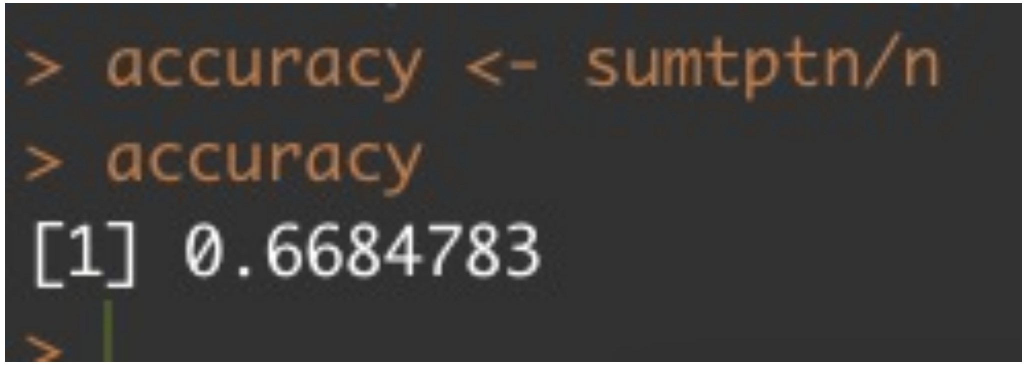


```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  1

Number of Support Vectors:  828
```



```
        model_predict
         Divorced Married Single
Divorced        0      81      1
Married         0     128     40
Single          0       0    118
```



```
> accuracy <- sumtptn/n
> accuracy
[1] 0.6684783
>
```

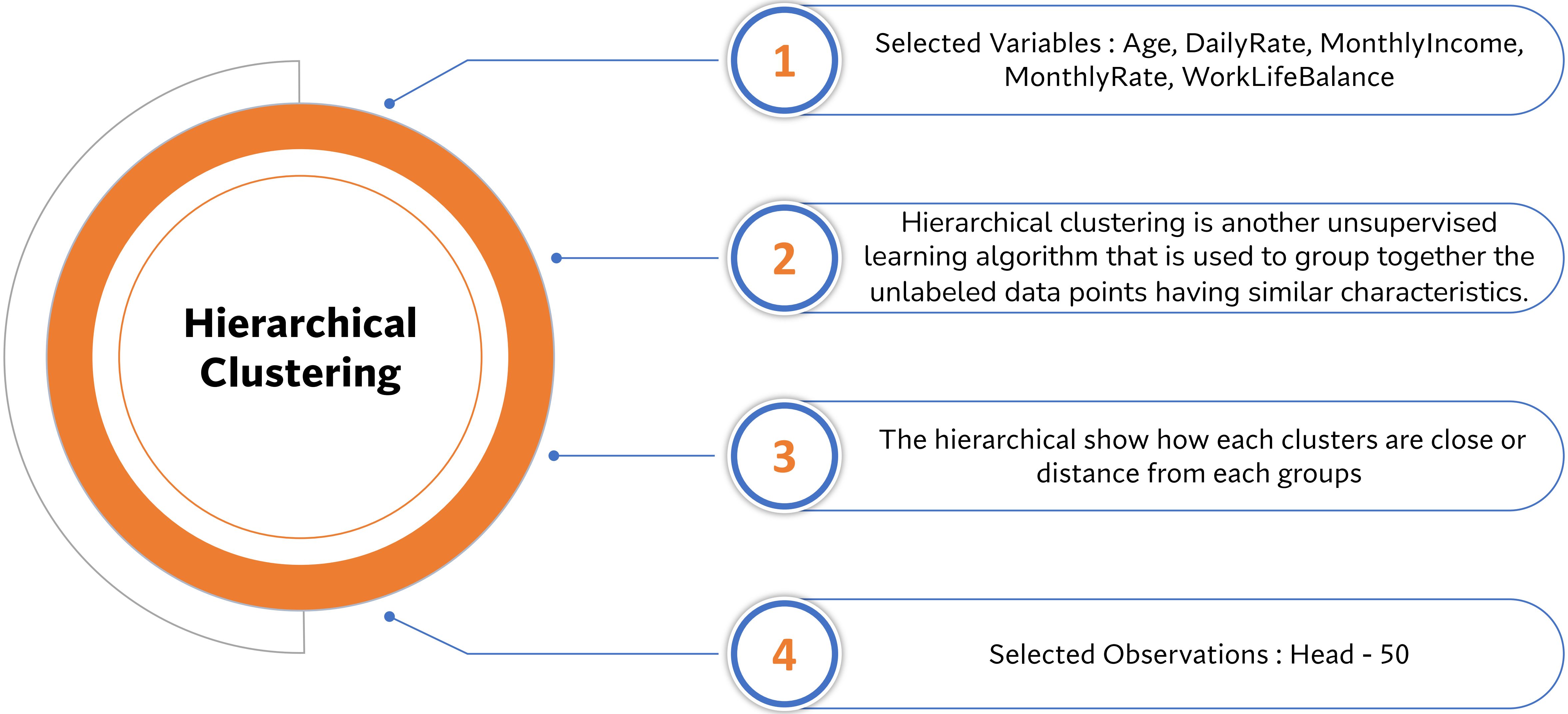
This model was created using **Marital Status** as the dependent variable and the **entire dataset independent variables.**

The model employed the type as **“C-classification” and “linear”Kernel.**

Confusion Matrix :
We then moved to develop a confusion matrix to define the performance of the classification algorithm.

Finally, we got model accuracy which is **66.84%**

TECHNIQUE -3



Hierarchical Clustering

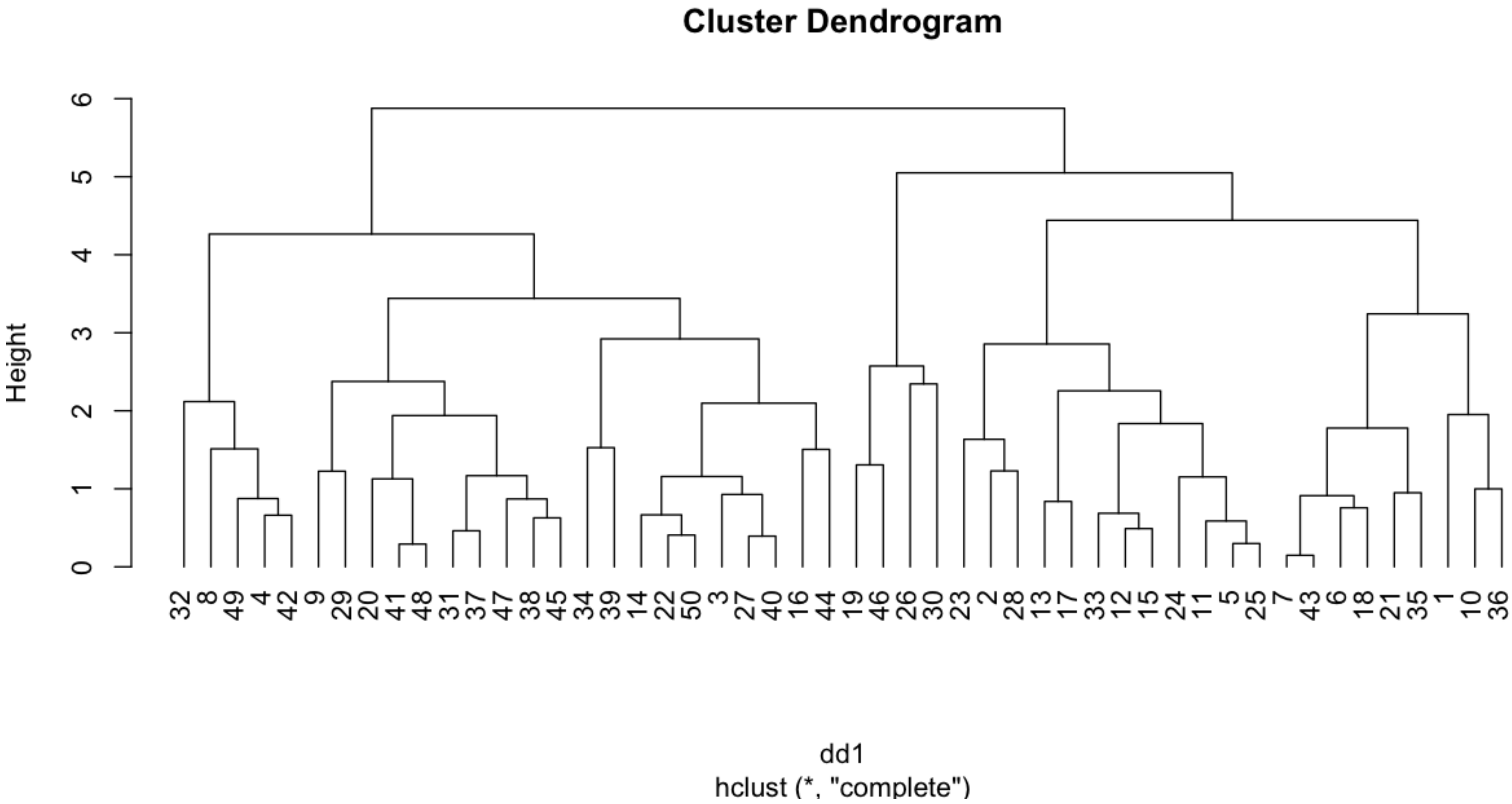
1. Selected Variables : Age, DailyRate, MonthlyIncome, MonthlyRate, WorkLifeBalance

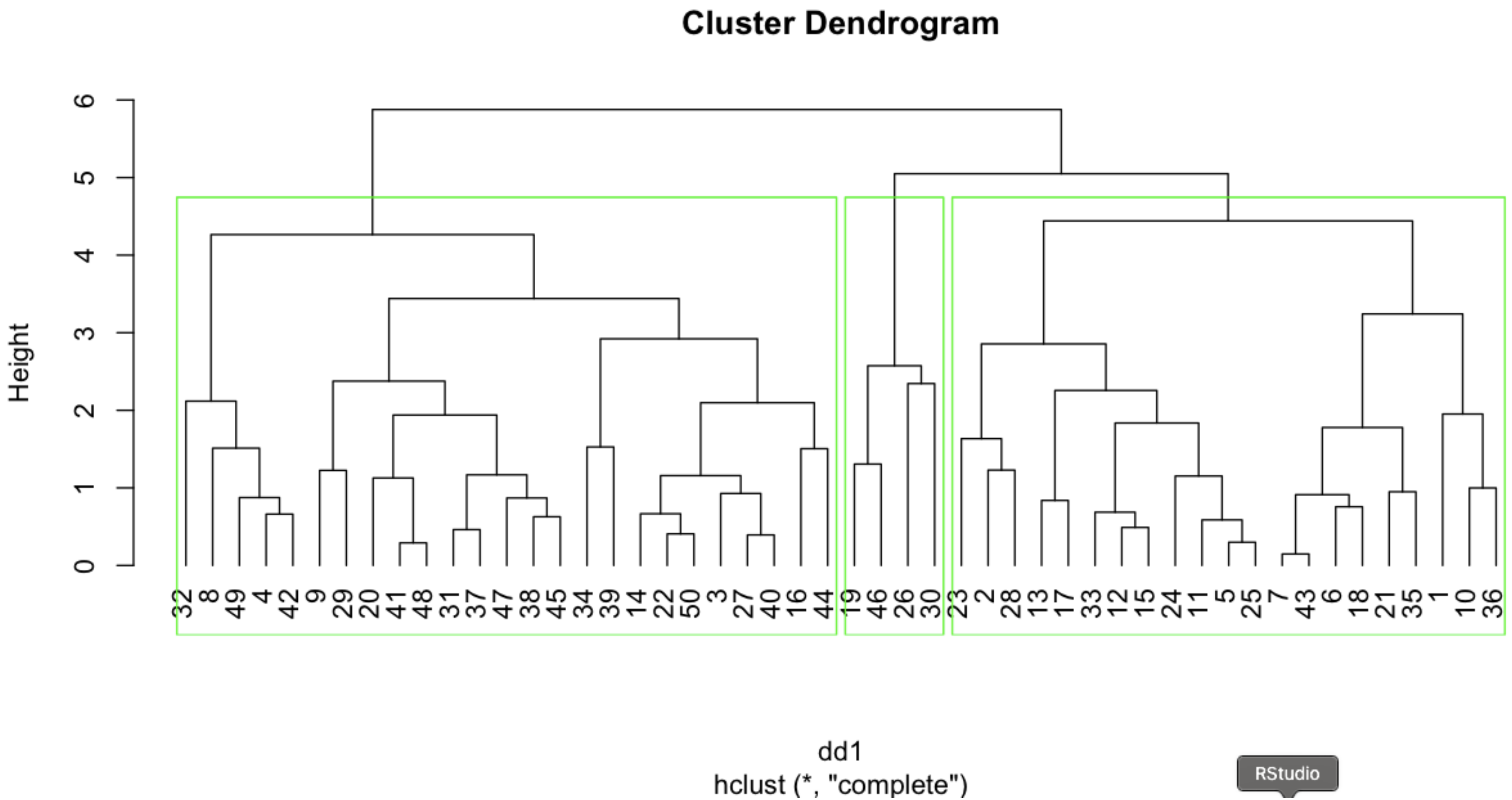
2. Hierarchical clustering is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics.

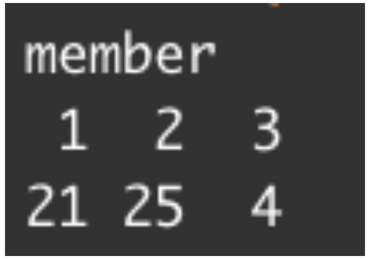
3. The hierarchical show how each clusters are close or distance from each groups

4. Selected Observations : Head - 50

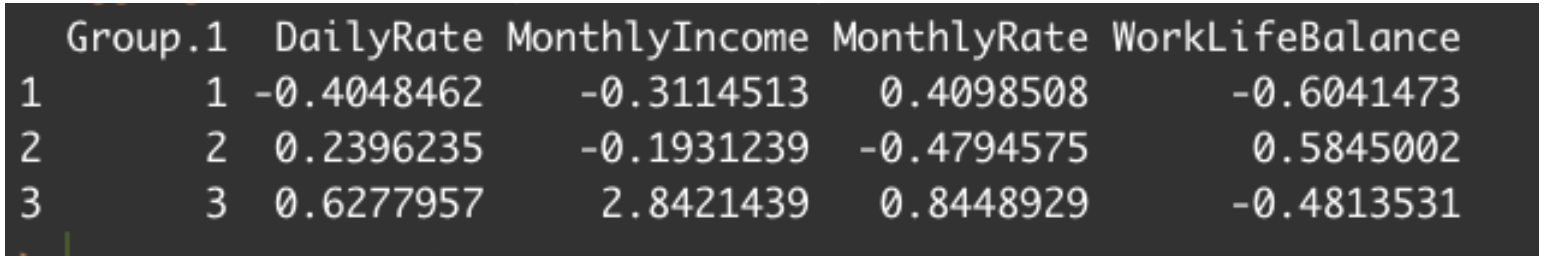
INTERPRETATION




Cluster Dendrogram
Height
0
1
2
3
4
5
6
32 8 49 4 42 9 29 20 41 48 31 37 47 38 45 34 39 14 22 50 3 27 40 16 44 19 46 26 30 23 2 28 13 17 33 12 15 24 11 5 25 7 43 6 18 21 35 1 10 36
dd1
hclust (*, "complete")
RStudio



```
member
 1  2  3
21 25  4
```



```
  Group.1  DailyRate MonthlyIncome MonthlyRate WorkLifeBalance
1       1 -0.4048462    -0.3114513   0.4098508      -0.6041473
2       2  0.2396235    -0.1931239  -0.4794575       0.5845002
3       3  0.6277957     2.8421439   0.8448929      -0.4813531
```

This is normalize averages for 3 clusters . We try to see which variable are contributing more and which are contributing less.

Variable that contribute more will have bigger averages differences while variable that contribute less will have lesser differences.

From the above interpretation we can say monthly income contribute more to employee attritation when compared to daily rate.

SUMMARY

In this presentation we covered applicable **Data Mining techniques**

After **K-Means clustering** analysis both clusters showed similarities in averages for all performance indicators selected..

For regards to the Support Vector Machine, where model we calculated accuracy is **66%** which partially good, we can say.

The last hierarchical clustering shows point at which each **cluster are been separated**.