**BUSINESS UNDERSTANDING**

Data on customers of an Automotive business have been recorded and the source’ business owner needs customer segmentation to be done to the dataset.

‘Due to the rise of income levels in many African countries and the emergence of a middle class, Deloitte regards the continent as the final frontier for the global automotive industry. Given Africa’s population size and its positive economic outlook, automotive companies will be able to gain a competitive advantage by adopting a medium- to long-term view towards the continent. ´- Deloitte Africa Automotive Insights

The customer needs a competitive edge to a very cutthroat market which is the current situation to which the client’s business resides. This shall be done in the following ways.

**OBJECTIVES**

* To help segment the customers of the client’s business**.**
* To help the client identify customers to retain.
* Decide how to relate to customers in each segment to maximize the value of each customer to the business.
* Figure out which groups (or segments) within the market are the best fit for particular products.

**BUSINESS SUCCESS CRITERIA**

The project shall be deemed successful if the objectives above can be achieved.

* This is there are distinct segments for the customers are derived.
* There are distinct criteria to which customers can be retained using.
* There are guidelines to which value of customers can be determined.
* Determine which product is best for different segments.

**ASSESS THE CURRENT SITUATION**

**Inventory of resources.**

Listed are the resources that were at the disposal of execution of this project:

Team of Data Scientist

Data

Required software: R-Studio

**Requirements, assumptions and constraints**

The assumptions to the undertaking of the project are:

Financing is primary to automotive sales

Income is a factor to consider when assessing customers to retain

The data is integral

The factors that affect the customers’ purchase behaviour is wholly as a result of the recorded attributes within the dataset.

**PROJECT PLAN**

Project plan

1. Acquisition of the data
2. Understanding the data
3. Preparing the data
4. Building a model
5. Challenging the model
6. Testing the model

**Initial assessment of tools and techniques**

**DATA UNDERSTANDING**

**Data collection Report**

The data was acquired from the site link <https://www.kaggle.com/yashgupta011/customer-segmentation-dataset?select=Customer\_Segmentation.csv>

**Description of the data**

The data acquired has the following attributes:

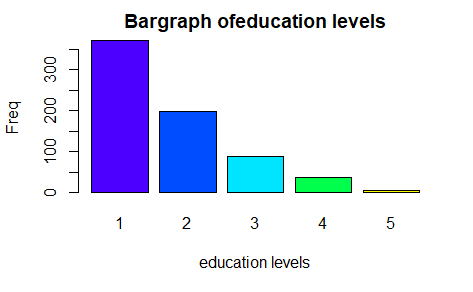
|  |  |
| --- | --- |
| **Column** | **Description** |
| Customer ID | Identification of the customer within the clients' data |
| Age | Age of the client |
| Edu | Education level of the |
| Years Employed | The years that the customer has been employed |
| Income | The income of the customers |
| Card Debt | The debt accrued by the customers credit card |
| Other Debt | Debt accrued by the customer through other means but credit cards |
| Defaulted | The state of payment of debt by the custome based on the terms of debt |
| Address | Address of the customer |
| DebtIncomeRatio | The ratio of the income to the debt |

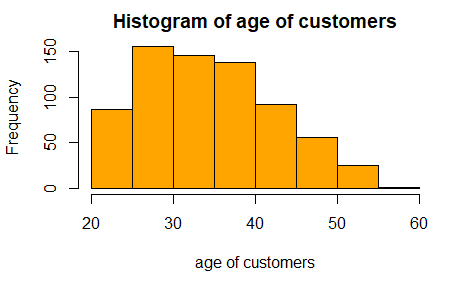
**Explore the data**

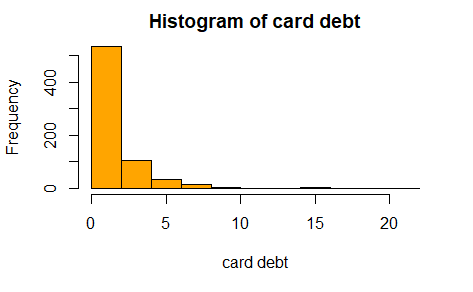
**Univariate analysis**

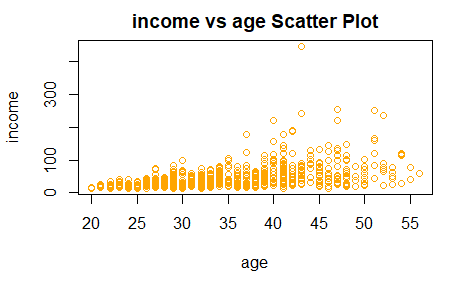
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Mean** | **Mode** | **Standard Deviation** | **Mediam** | **Variance** | **Minimum** | **Maximum** | **Range** | **Kurtosis** | **Skewness** |
| **Age** | 34.86 | 29.00 | 8.00 | 34.00 | 63.96 | 20.00 | 56.00 | 36.00 | 2.39 | 0.36 |
| **Edu** | 1.72 | 1.00 | 0.93 | 1.00 | 0.86 | 1.00 | 5.00 | 4.00 | 3.73 | 1.20 |
| **Years Employed** | 8.39 | 0.00 | 6.66 | 7.00 | 44.33 | 0.00 | 31.00 | 31.00 | 3.22 | 0.83 |
| **Income** | 45.60 | 25.00 | 36.81 | 34.00 | 1355.29 | 14.00 | 446.00 | 432.00 | 28.97 | 3.85 |
| **Card Debt** | 1.55 | 0.69 | 2.12 | 0.86 | 4.48 | 0.01 | 20.56 | 20.55 | 24.81 | 3.89 |
| **Other Debt** | 3.06 | 1.07 | 3.29 | 1.99 | 10.81 | 0.05 | 27.03 | 26.99 | 13.25 | 2.72 |
| **Defaulted** | 0.26 | 0.00 | 0.44 | 0.00 | 0.19 | 0.00 | 1.00 | 1.00 | 2.18 | 1.09 |
| **DebtIncomeRatio** | 10.26 | 4.50 | 6.83 | 8.60 | 46.61 | 0.40 | 41.30 | 39.90 | 4.20 | 1.09 |

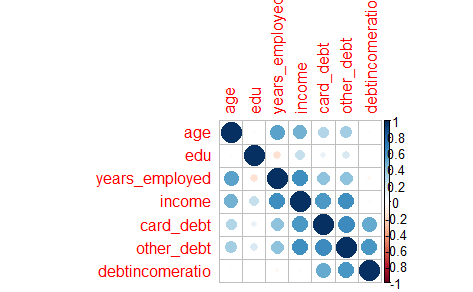
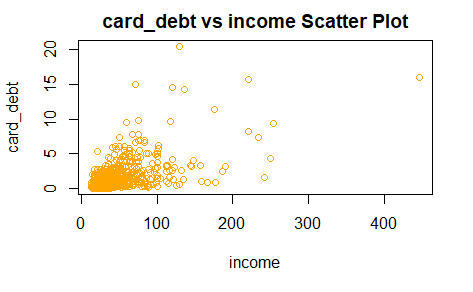
From the univariate analysis nothing out of the ordinary can be noted.











**Verifying the data quality**

An area of contention arose due to the lack of 150 records within the defaulters attributes. As a result of vigorous discussion, a conclusion was arrived at that the records be dropped which will retain 82.36% of the original dataset.

The decision above was arrived at by the argument that defaulting on debt payment is primarily a choice as opposed to dependent on the other attributed within a dataset. That is; a hypothetical situation of a high-income low debt consumer may default on rent because they wanted to do something on their bucket list.

**Data Quality repost**

After the EDA was done the outliers, duplicates and missing values had been handled the data was ready for modelling.

**DATA PREPARATION**

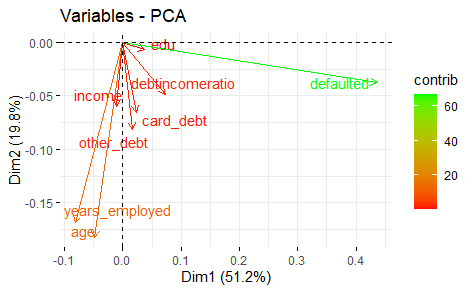
Due to the decisions made while understanding the data; the following conclusion was arrived at:

1. No derivative data will be required for the modelling
2. The columns customer\_id and address will not be used within the data modelling.
3. Clustering will be used for the modelling thus reliant on the numerical attributes of the data thus maintenance of the both the categorical data and Boolean data as numerical data.

**MODELLING**

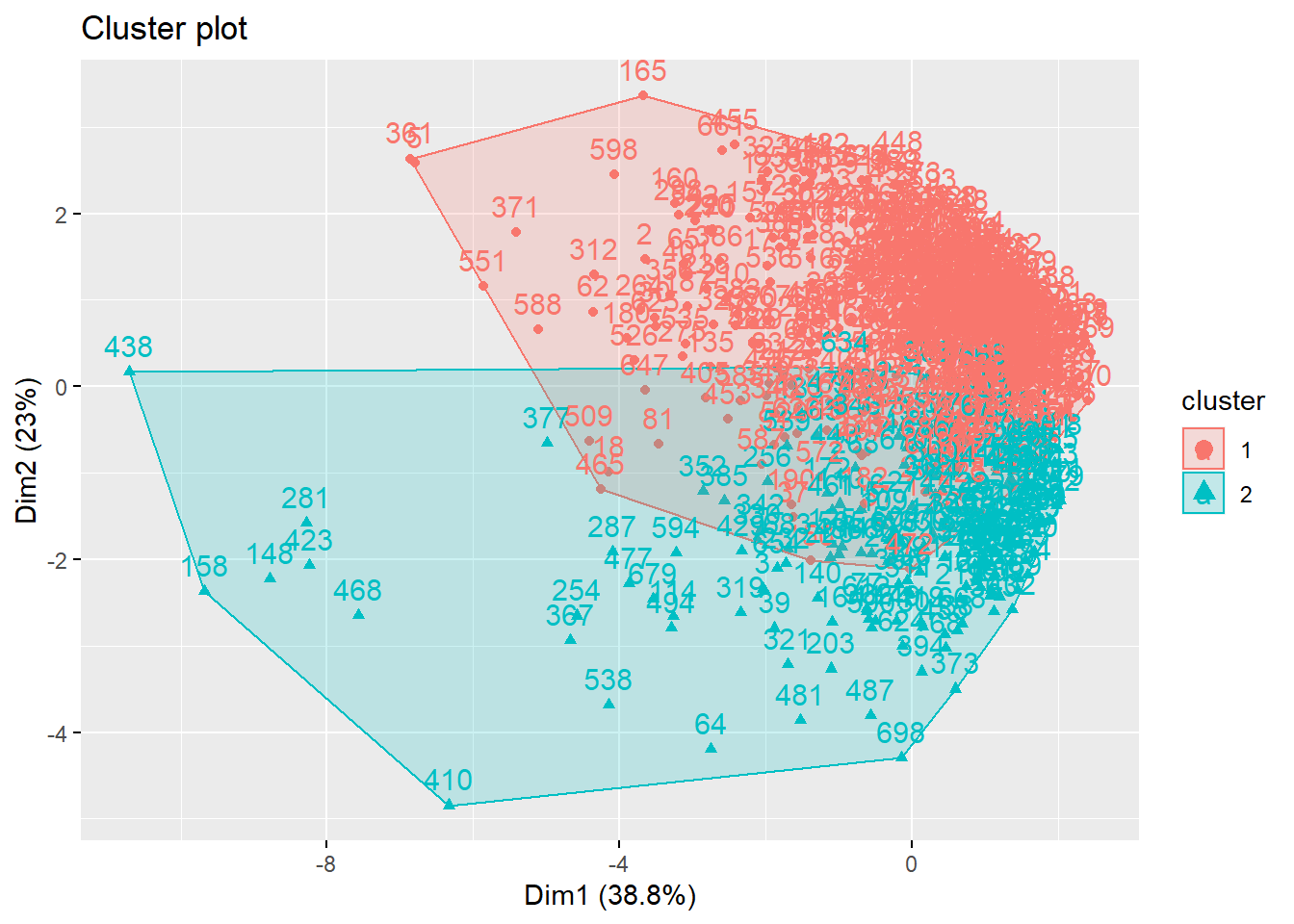
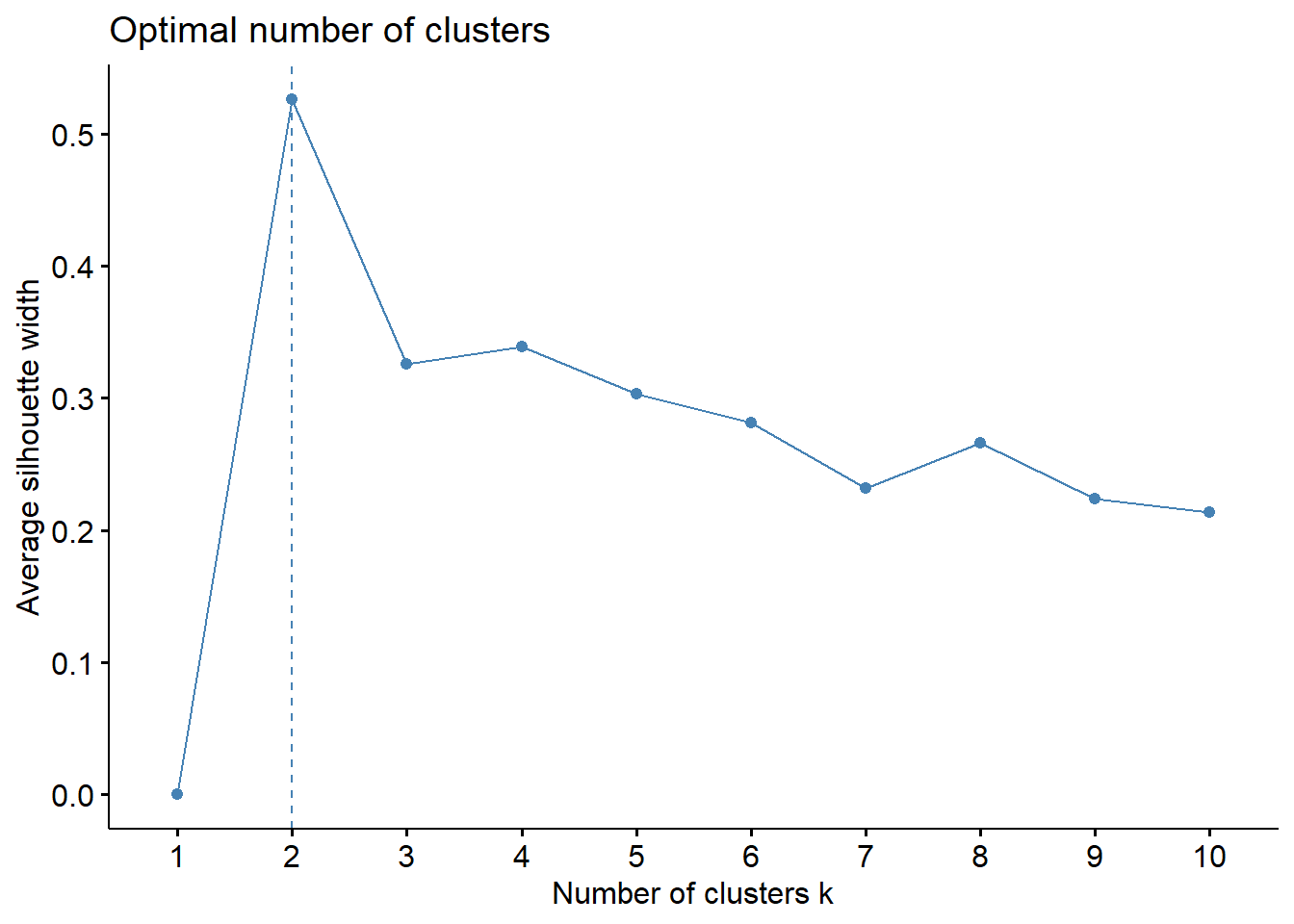
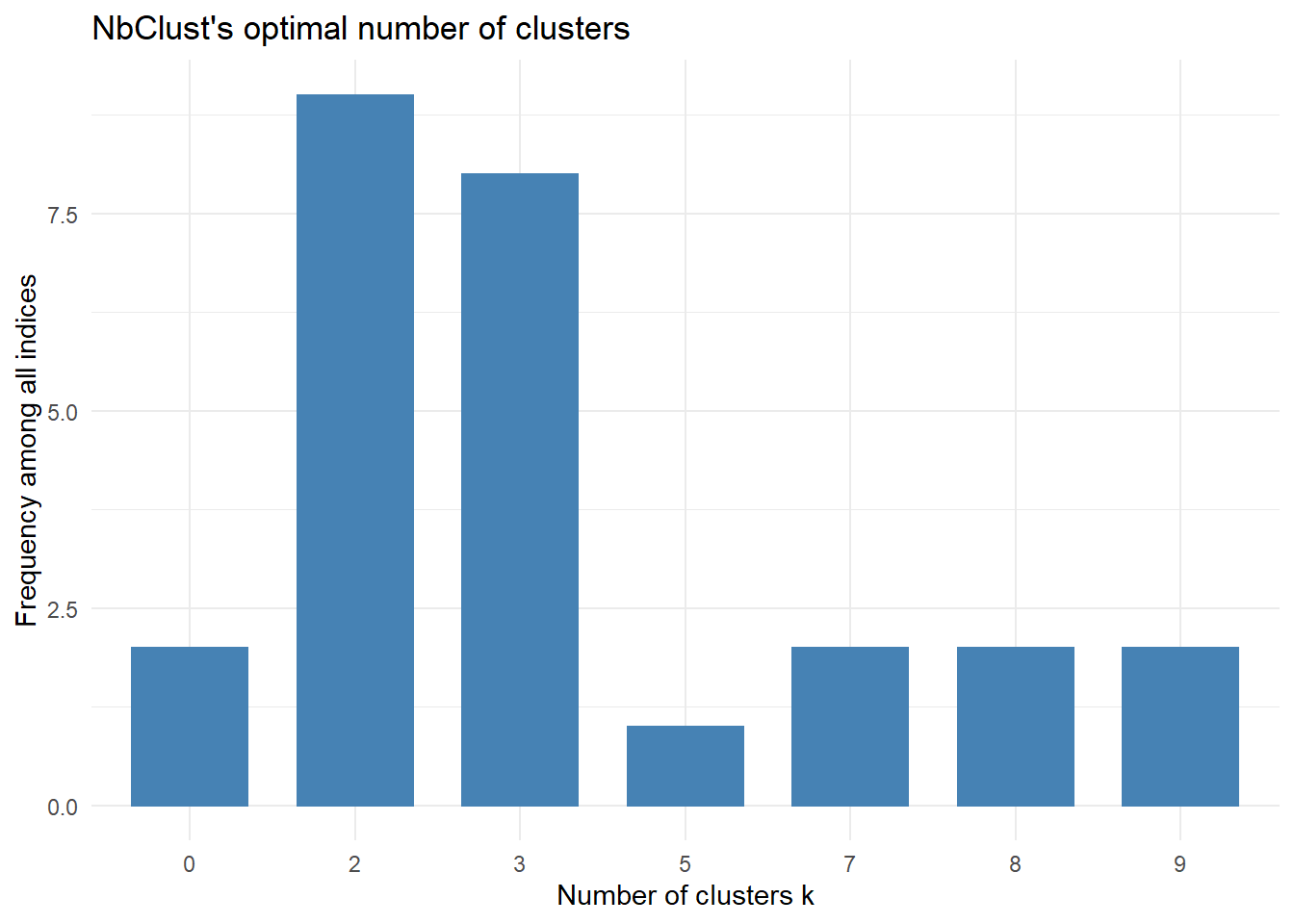
First; the data was normalized so that no particular attribute had more impact on the clustering algorithm. The result is summarised in the markdown document accompanied by this document.

Secondly, Principal Component Analysis was done for dimensionality reduction before the use of the algorithm. As observed above with the correlation plot, the data dimension needs to be reduced while all of the contributing attributes that make up the data maintained.



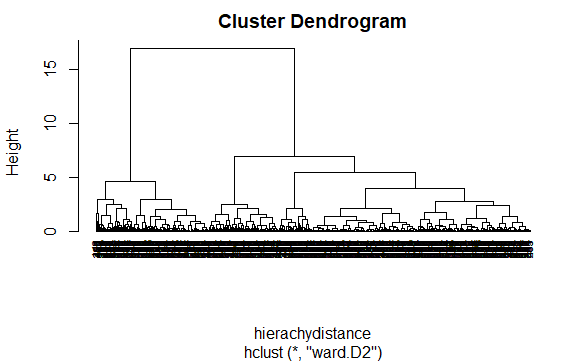
Third; Clustering

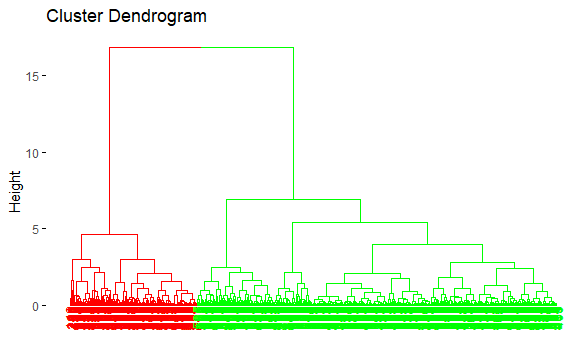
Both the elbow method and the silhouette method were looked into while determining the number of clusters to which the dataset is to be divided to.



**EVALUATION**

**Hierarchical clustering is used to challenge the solution that has been gotten from the modelling process. A dendogram is used to graphically display the data and the clustering that has been developed in the modelling stage of the project.**





**DEPLOYMENT**