NOVA
IMS
Information
Management
School

BUSINESS CASE 1

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

WONDERFUL WINES OF THE WORLD

Group PN:

Mafalda Garcia, number: 20210763

Simão Pereira, number: 20210250

Tiago Santos, number: 20210548

Rui Ribeiro, number: 20211017

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. Introduction

One of the best ways for a company to expand in the industry is through reaching new customers and improve the experience of existing customers. This reach will allow the company to have contact with the most differentiated styles that in turn will inspire the company to commercialize other products or even improve its business strategies, thus improving its final profit. The process of finding new customers is based on learning about how existing customers behave and being able to identify various groups within them not only to be able to satisfy their wishes but also improve the targeting of prospective customers.

To carry out this project we followed a CRISP-DM methodology. Firstly, we start by reconnaissance of the business, analyzing its objectives, which materials we have access to and defining the data mining goals for this project. Next, we will perform a recognition of the data itself, seeing its composition in detail and we will start the necessary steps for the data preparation, making all the necessary changes in the date so that it is possible to apply the methods of modeling it.

2. Business Understanding

2.1 Business Objectives

Wonderful Wines of the World started to develop its database 4 years ago, however it is quite simple and not very efficient, given the variety of products it offers. So, in order to keep up with the market and be able to make a better differentiation between the tastes and preferences of the customers, the company feels that it is time to upgrade their database. WWW want to have a database and programs focused on the characteristics and history of the customer choices that automatically associate them with a certain type of wine or related product. This improvement will give the board, marketing, operations, and clients departments a better insight to make smart and strategic decisions.

2.2 Assess Situation

WWW has provided a sample of 10,000 customers aged between 18 and 78, with different incomes, who made purchases in the last 18 months. Each customer is identified by an ID. The database provided contains information regarding the level of education, number of days as a customer, number of purchases made, number of days since the last purchase and total value of sales of the client. The database also comes with information on the products purchased from the last 18 months, the percentage of purchases he made at a discount, the percentage relative to the different types of wine (dry red, semi-dry red, white, semi-dry/ sweet white, dessert or unusual wines) and finally the percentage of purchases and visits per month made through the website.

Part of this information was deleted as this was not relevant for the improvement of the business.

2.3 Data Mining Goals

Our main goal is focus on analyzing the given data which has information about previous purchases of the clients as well as some personal information. From these two points we want to find a pattern in groups of clients that have the same characteristics (cluster) using different clustering models.

From our final model we will be able to make some business approaches for each cluster and for the business in general.

3. Data Understanding

3.1 Collect Initial Data

The dataset was present in an excel document given by the WWW company which was loaded to a Jupyter Notebook using the panda's library.

3.2 Description of data

The dataset has a total of 10001 records each representing a unique client (the last record that represented the average value of each attribute) and 17 features, all of them are numerical features.

In these features we can find some related with personal information of each customer (e.g: Age and income) and others represent the consumer preferences of each client (e.g: recency; favorite wine).

We think that the dataset was complete and had really useful information about the clients, this will help us to perform a good clustering method.

3.3 Exploring data

To start exploring the data we performed the info() method to acquire information about the dataset as number of features, type of the variables, memory usage, null values, etc. To better understand the data, we analyzed statistical information about all the features, and it matched with the previous information we had about these attributes (e.g: mean, max and min values). We also used visualization tools to identify possible outliers and to better understand the distribution of the variables and their correlation with each other.

3.4 Verify data quality

In terms of quality, this data was pretty clean, we only had one null value in the customer id. This null value was a consequence of the last row of the dataset that represented an average value of each feature and for an ID attribute that does not make sense. We ended up removing this last row and we kept 10000 rows from the original data. There were no missing values, no wrong data types, no duplicated rows and no incoherent values, so the data had high quality.

4. Data Preparation

4.1 Selection and Cleaning of Data

As mentioned above the dataset we had in our hands had high quality so there were not many changes. The first change made was the one that we already talked about, removing the last row (average values) of the data set because it had a null value. After this change and due to the high correlation with two variables we decided to eliminate from the model the feature "Frequency" because we think that it was giving us redundant information.

Note: Regardless of the high correlation between age and income we think they are both important for this particular business case.

We also removed the feature "Education" because we think that has not relevant information for this particular business case and to finalize, we did not find any clear outlier values so it`s fair to say that we kept pretty much most of the records and features of the initial data.

4.2 Formatting the data

To finish the data preparation, and since we only had numerical variables, we standardized all the values so that the model could be more accurate.

For that we used the Robust Scaler method that also deals well with the presence of outliers and because we did not remove any we thought this method might be helpful in the case we did miss some outlier.

After these steps our data was ready to be used in our modelling methods.

5. Modelling

5.1 Selecting modelling technique

5.1.1 K-means & Hierarchical Clustering

From all the cluster modeling techniques available we decided to use K-means and Hierarchical Clustering.

We decided to use K-means because the mean of the variables does not change with time and because of this we considered our as normal distributed. Also chose to use Hierarchical Clustering in order to confirm the number of clusters we should apply with K-means.

5.2 Building the model

Before using any clustering technique, we segmented our data into two segments, value and behavior, based on the variables of our data set.

The value segment contains the variables: “Dayswus”, “Income”, “Monetary” and “LTV”. This segment describes the characteristics of the customers.

The behavior segment contains the variables: “Perdeal”, “Dryred”, “Sweetred”, “Drywh”, “Sweetwh”, “Dessert”, “Exotic”, “WebPurchase”, “WebVisit” and “Recency”. This segment describes the buying behavior of each customer.

With our data divided into value and behavior segments we built our cluster models for both segmentations. In the end we merged these two solutions into a final one with all the variables.

In order to choose the right number of clusters for each K-means used we applied the elbow plot, the average silhouette scores in the different scenarios and confirmed the results with the Hierarchical Clustering.

When using Hierarchical Clustering for each of the segments we found that the best solution for the value segment was to use 4 clusters and for the behavior segment the best option was to use 5 clusters. These results were chosen after analysing the dendograms and the R2 plots for both segmentations.

After analysing the elbow plot, the average silhouette scores and the Hierarchical Clustering we ended up deciding to use 3 clusters for both value and behavior segmentation using the K-means Method.

The last step to do at this stage was to merge the best solutions. In order to define the best merge possible we used the R2 scores to help us decide.

The optimal solution we found was made after merging both segmented K-means solutions. Our final solution has 5 clusters. Each cluster characterizes a different niche of customers and because of this we named each cluster based on the main distinctions between them.

The different segments are characterized by the following:

GEN X (5850 Customers)

Value characteristics: Middle aged people, middle class, standard consumer (inside the average).

Behavior characteristics: Don't get really affected by discounts even though some amount of purchases are made on discounts, buy mainly red wine (dry red in fact), also like dry white and exotic wines and half of the people tend to buy from the website.

GEN Y1 (1698 Customers)

Value characteristics: Young adults, lower income, lower amount of purchases recently and really low LTV.

Behavior characteristics: More than half of the purchases are made on discount, buy all kinds of wine with a slightly preference for exotic wines (best costumers for this type of wine), prefer sweet wines, consume more white wine than red wine and half of the people tend to buy from the website.

AVG BOOMER (625 Customers)

Value characteristics: Older people, higher income, higher amount purchases lately and high LTV.

Behavior characteristics: Usually buy wine at the standard price (don't need discount to buy), buy all kinds of wine and consume pretty much the same amount of red and white wine and usually don't buy the wines on the website.

GEN Y2 (307 Customers)

Value characteristics: Young adults/Middle age, Lower income, don't buy wine for a long time ago (more than a year) and negative LTV.

Behavior characteristics: More than half of the purchases are made on discount, fans of the dry wines, also tend to buy exotic wines and half of the people tend to buy from the website.

BOOMERS DRY (1520 Customers)

Value characteristics: Older people, higher income, higher amount of recent purchases and high LTV.

Behavior characteristics: Usually buy wine at the standard price (don't need discount to buy), costumers tend to buy dry wines, usually don't buy the wines on the website and below average website visits.

After the modelling we conclude that we ended up with clusters that make a good segmentation of our customers based on their characteristics. The main difficulties we found on this modeling stage was to find the best clustering technique to use on this case in order to get results that made sense based of the information we had.

6. Evaluation

6.1 Evaluate Results

Through the use of the optimal model we worked with, we managed to achieve the distinct group of clients and understand what differentiates them from each other. This will help the board of the company to better understand their customers. This can be of high importance, especially when it comes to strategic decisions, such as marketing approaches and the distribution of a budget.

Given these considerations, when it comes to the objectives of the service we were hired to provide, we consider them to have been successfully achieved.

6.2 Review Process

When it comes to the complete working process, we didn’t identify any specific issues that could undermine the results. We consider the optimal model to be well built and that we used every significant attribute that was available in the Data Base.

6.3 The next steps

The next step for this team would be to try and find a model that better fits the data and better segments the groups of customers.

Although this could be an interesting endeavor, the budget impact that the search for a better model would have, and the fact that we consider the final model to have fitted very well in the data, made us disregard this next possible step.

7. Deployment

Considering we have evaluated our model as optimal and couldn’t find any improvements that made sense, in this case, we won’t proceed with the Deployment phase of the CRISP-DM process.

Although we would like to suggest the following approaches based on the cluster analysis:

Firstly, for the Gen X we suggest offering one dry wine after 5 purchases (within 3 months) and create a combo package of exotic, dry white and dry red wines.

For the younger public, the GEN Y1, we think that 50% discount on an exotic wine (up to 20€) every time they buy a wine (from 7€ on) would be a nice approach and also use influencers to do online campaigns of different exotic wines on social media.

Thirdly, we have the Average Boomer, which represents an older public that are not used to online shopping so the best option is to make a good quantity discount in store for high quality wines of any kind because this group of clients buy a lot of times and want more expensive wines normally. Could be a nice strategy as well create packs with wines and some kind of gift such as cork openers, glasses, etc.

For the GEN Y2, we suggest sending emails offering a 20% discount for a package of three specific wines (Dry red; Dry white and Exotic) and use promo stands in physical stores with wine tasting sessions.

Lastly, for the Boomers Dry, an older public with a high income, we assume that type of clients want new luxurious wines more often, so we advise WWW to acquire a package of new high-quality wines and contact directly these customers in order for them to try the new arrivals. Also offering a sampling of dry Premium Wines in physical stores would be a good strategy.

8. Conclusion

In today's market, the correct use of internet in favor of the business is key.

During our work, we realized that, even with the younger population, the Wonderful Wines of the World website sales are not in the recommended amount. Given this, our main recommendation is that the company promotes and improves their website, in order to make it better known and more user friendly.

Appendix

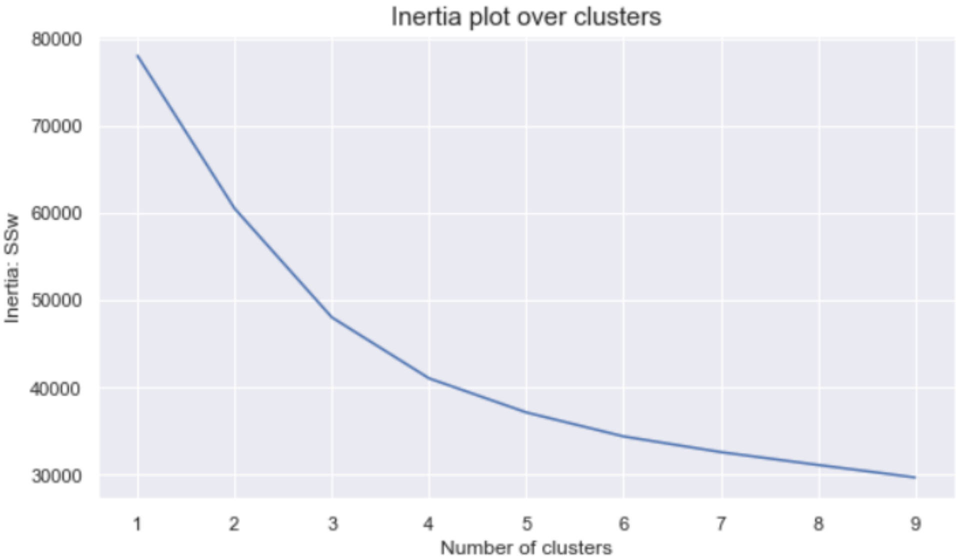
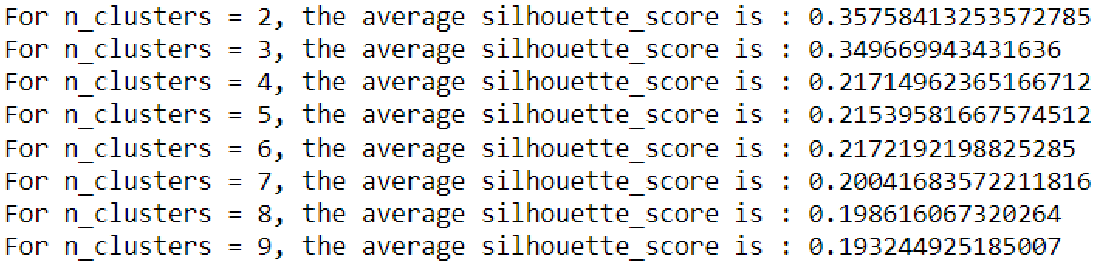


Fig.1 - Inertia K-means Behavior



```
For n_clusters = 2, the average silhouette_score is : 0.35758413253572785
For n_clusters = 3, the average silhouette_score is : 0.349669943431636
For n_clusters = 4, the average silhouette_score is : 0.21714962365166712
For n_clusters = 5, the average silhouette_score is : 0.21539581667574512
For n_clusters = 6, the average silhouette_score is : 0.2172192198825285
For n_clusters = 7, the average silhouette_score is : 0.20041683572211816
For n_clusters = 8, the average silhouette_score is : 0.198616067320264
For n_clusters = 9, the average silhouette_score is : 0.193244925185007
```

Fig.2 - Silhouette K-means Behavior

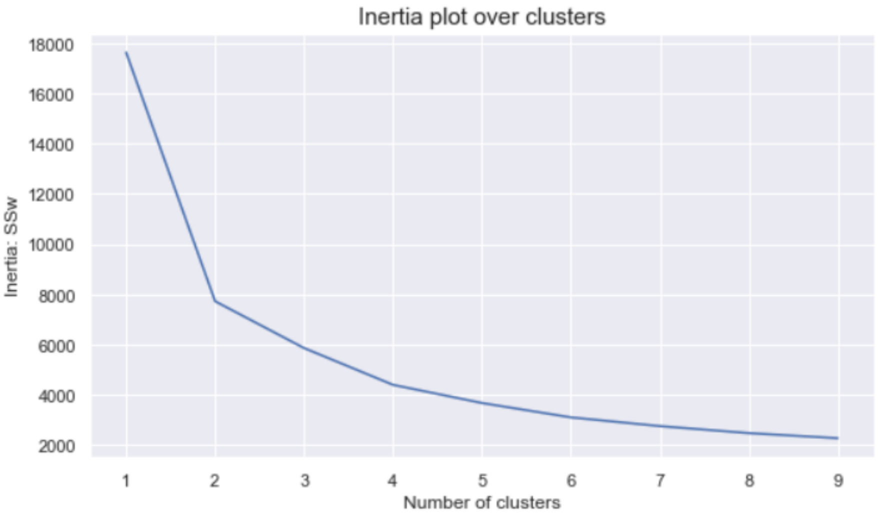
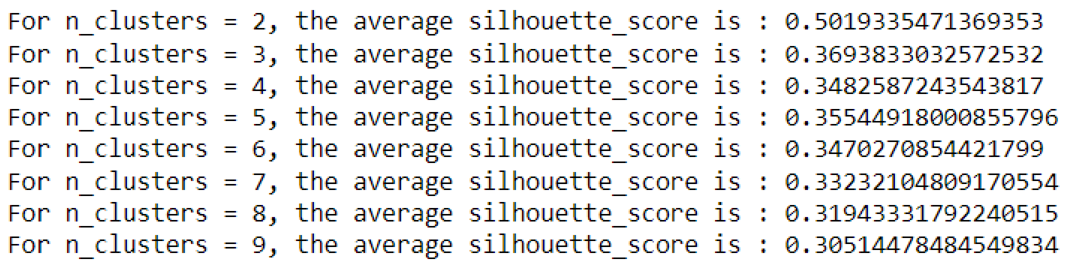


Fig.3 - Inertia K-means Value



```
For n_clusters = 2, the average silhouette_score is : 0.5019335471369353
For n_clusters = 3, the average silhouette_score is : 0.3693833032572532
For n_clusters = 4, the average silhouette_score is : 0.3482587243543817
For n_clusters = 5, the average silhouette_score is : 0.35544918000855796
For n_clusters = 6, the average silhouette_score is : 0.3470270854421799
For n_clusters = 7, the average silhouette_score is : 0.33232104809170554
For n_clusters = 8, the average silhouette_score is : 0.31943331792240515
For n_clusters = 9, the average silhouette_score is : 0.30514478484549834
```

Fig.4 - Silhouette K-means Value

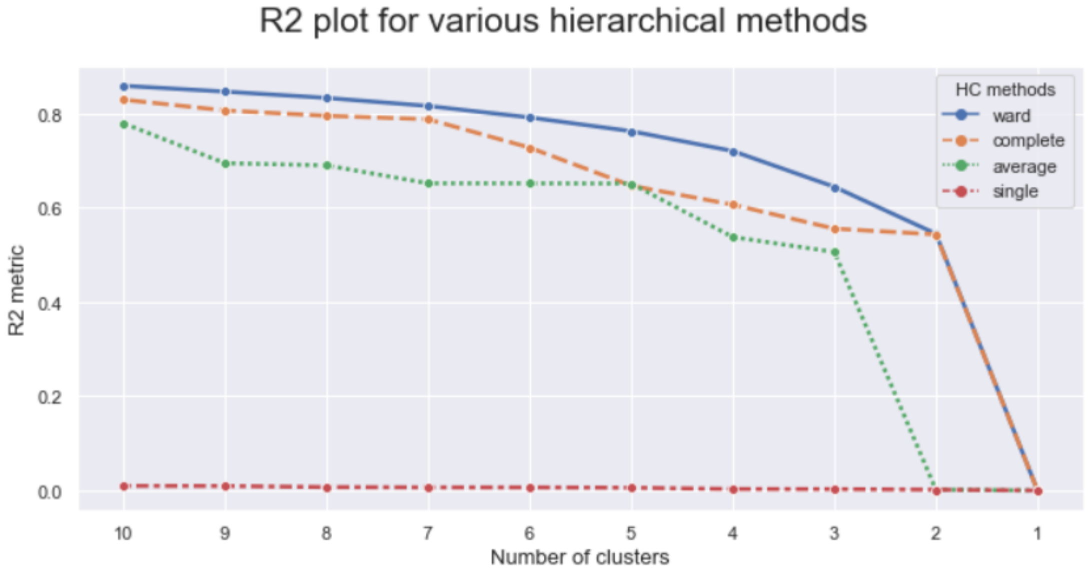


Fig.5 - HC Value R2 Plot for Various Methods

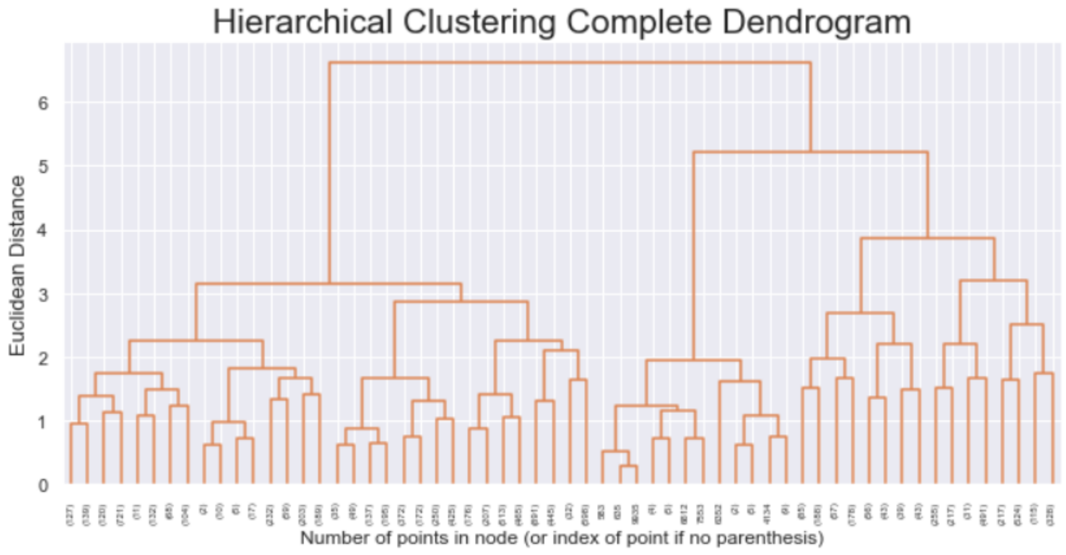


Fig.6 - HC Value Dendogram

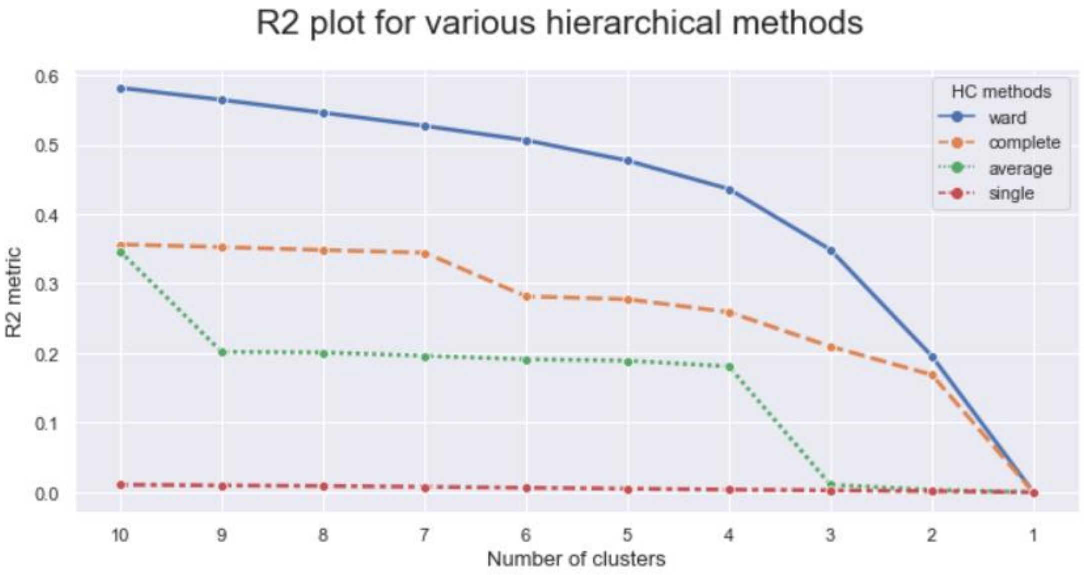


Fig.7 - HC Behavior R2 Plot for Various Methods

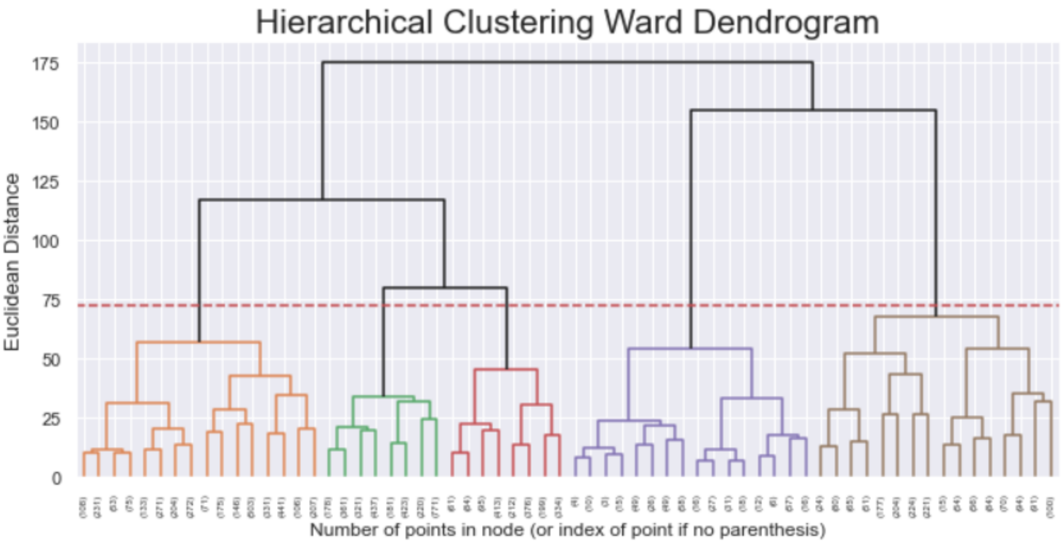


Fig.8 - HC Behavior Dendogram

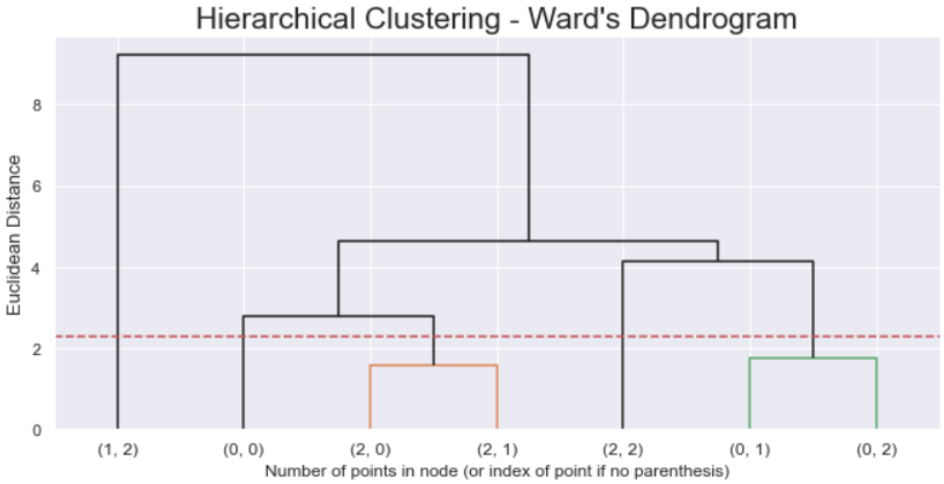


Fig.9 - Merge of K-means clusters – Dendograms

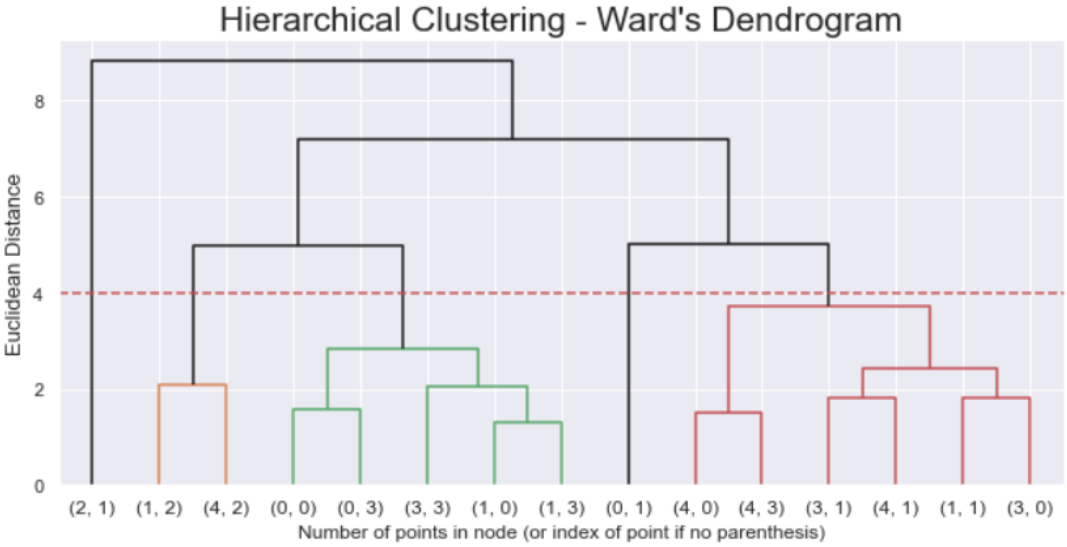
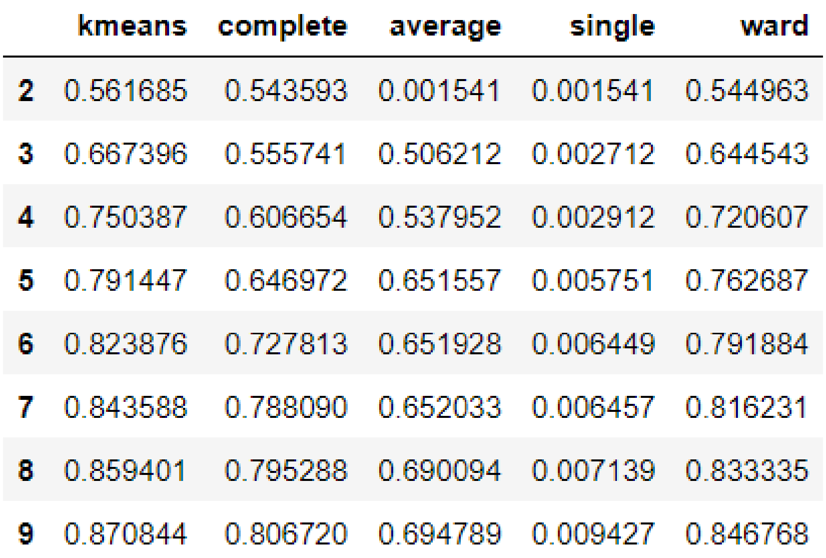
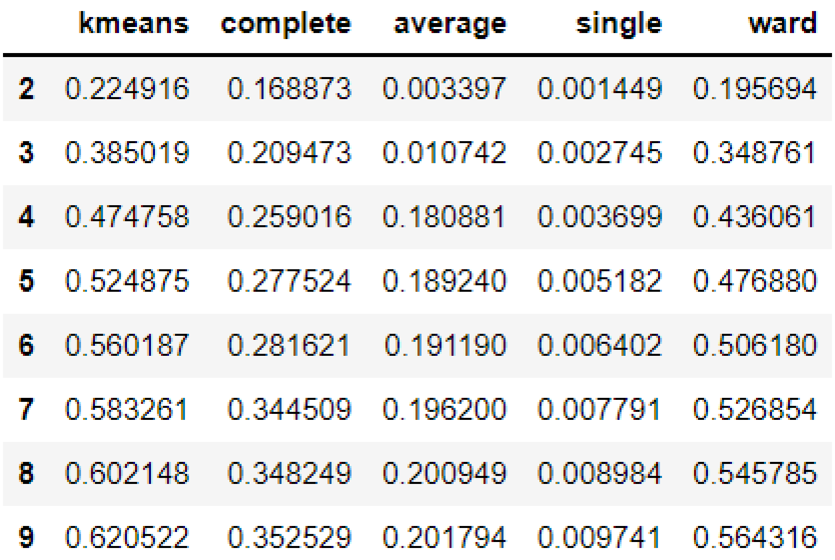


Fig.10 - Merge of HC clusters – Dendograms



	kmeans	complete	average	single	ward
2	0.561685	0.543593	0.001541	0.001541	0.544963
3	0.667396	0.555741	0.506212	0.002712	0.644543
4	0.750387	0.606654	0.537952	0.002912	0.720607
5	0.791447	0.646972	0.651557	0.005751	0.762687
6	0.823876	0.727813	0.651928	0.006449	0.791884
7	0.843588	0.788090	0.652033	0.006457	0.816231
8	0.859401	0.795288	0.690094	0.007139	0.833335
9	0.870844	0.806720	0.694789	0.009427	0.846768

Fig.11 – R2 Value



	kmeans	complete	average	single	ward
2	0.224916	0.168873	0.003397	0.001449	0.195694
3	0.385019	0.209473	0.010742	0.002745	0.348761
4	0.474758	0.259016	0.180881	0.003699	0.436061
5	0.524875	0.277524	0.189240	0.005182	0.476880
6	0.560187	0.281621	0.191190	0.006402	0.506180
7	0.583261	0.344509	0.196200	0.007791	0.526854
8	0.602148	0.348249	0.200949	0.008984	0.545785
9	0.620522	0.352529	0.201794	0.009741	0.564316

Fig.12 – R2 Behavior