Mini Project 1

Full Name: Miguel Ángel Benítez Alguacil
Student ID: 1900853

Åbo Akademi University

Data Analytics Software

Index

1. Brief description of the problem

The aim of this project is to choose one of the two projects given and submit a report including:

- a short explanation on what the data is about and how you are planning to use it in doing the task of the project
- what are the steps you are taking in order to perform the task (include screenshots of each step)
- interpreting the output.

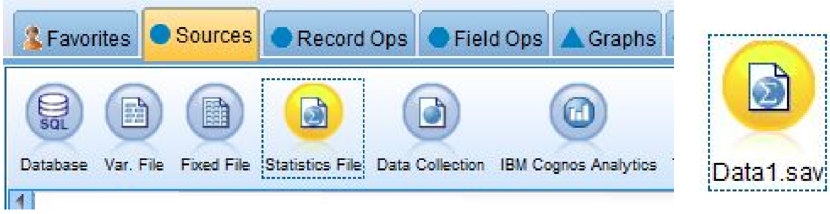
I have chosen the first project (**Project 1: K-Mean Cluster Analysis**), which problem is the following: “A telecommunications provider who wants to segment its customer base by service usage patterns. If customers can be classified by usage, the company can offer more attractive packages to its customers. Use clustering to find subsets of "similar" customers.”

In the next pages I will explain how I have solved the problem using IBM SPSS Modeler.

2. Solution to the Problem

2.1. Get the Data

The first thing I had to do was getting the data from the file “Data1.sav” and import it in IBM SPSS Modeler. To do so, I used a Statistics File, clicking on **Sources** → **Statistics File**.

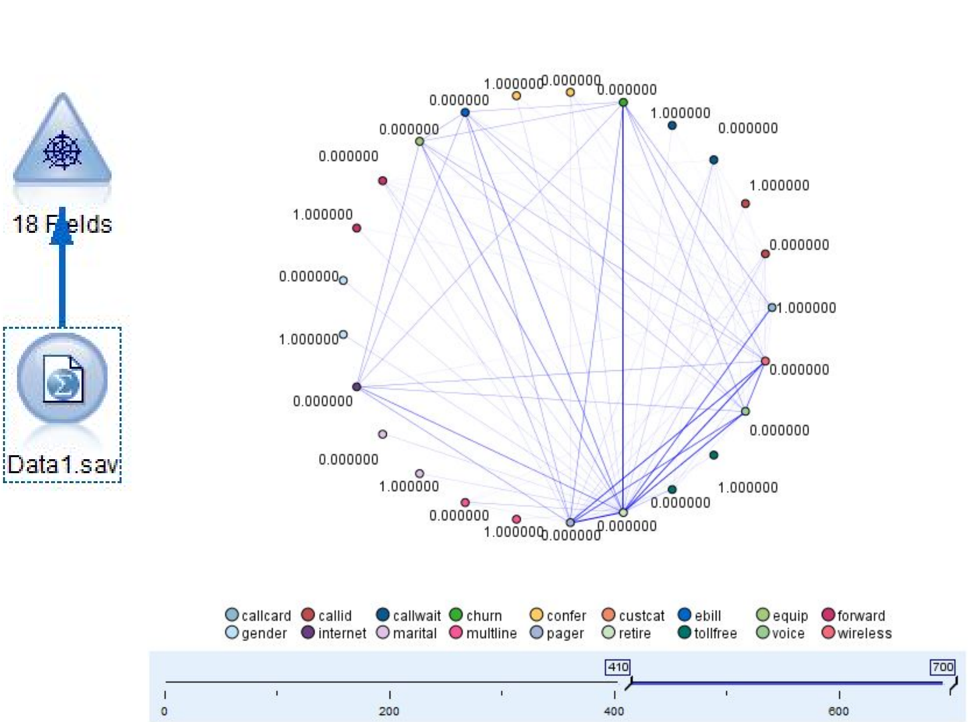


Once we have it, we can double click on him and **preview** the data. If we do so, we obtain the following table with the data (it will only show the 10 first rows of data):

	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	tollfree	equip	callcard	wireless	longmon	tollmon	equipmon	cardmon	wiremon	multline	voice	pager	intern
1	2.000	13.000	44....	1.000	9.000	64.000	4....	5.000	0.000	0.000	2.000	0.000	0.000	1.000	0.000	3.700	0.000	0.000	7.500	0.000	0.000	0.000	0.000	0.0
2	3.000	11.000	33....	1.000	7.000	136.000	5....	5.000	0.000	0.000	6.000	1.000	0.000	1.000	1.000	4.400	20.750	0.000	15.250	35.700	0.000	1.000	1.000	0.0
3	3.000	68.000	52....	1.000	24.000	116.000	1....	29.000	0.000	1.000	2.000	1.000	0.000	1.000	0.000	18.150	18.000	0.000	30.250	0.000	0.000	0.000	0.000	0.0
4	2.000	33.000	33....	0.000	12.000	33.000	2....	0.000	0.000	1.000	1.000	0.000	0.000	0.000	0.000	9.450	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
5	2.000	23.000	30....	1.000	9.000	30.000	1....	2.000	0.000	0.000	4.000	0.000	0.000	0.000	0.000	6.300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
6	2.000	41.000	39....	0.000	17.000	78.000	2....	16.000	0.000	1.000	1.000	1.000	0.000	1.000	0.000	11.800	19.250	0.000	13.500	0.000	0.000	0.000	0.000	0.0
7	3.000	45.000	22....	1.000	2.000	19.000	2....	4.000	0.000	1.000	5.000	0.000	0.000	1.000	0.000	10.900	0.000	0.000	8.750	0.000	1.000	0.000	0.000	1.0
8	2.000	38.000	35....	0.000	5.000	76.000	2....	10.000	0.000	0.000	3.000	1.000	1.000	1.000	1.000	6.050	45.000	50.100	23.250	64.900	1.000	1.000	1.000	1.0
9	3.000	45.000	59....	1.000	7.000	166.000	4....	31.000	0.000	0.000	5.000	1.000	0.000	1.000	0.000	9.750	28.500	0.000	12.000	0.000	1.000	0.000	0.000	0.0
10	1.000	68.000	41....	1.000	21.000	72.000	1....	22.000	0.000	0.000	3.000	0.000	0.000	1.000	0.000	24.150	0.000	0.000	16.500	0.000	1.000	0.000	0.000	0.0

2.2. Work with the Data

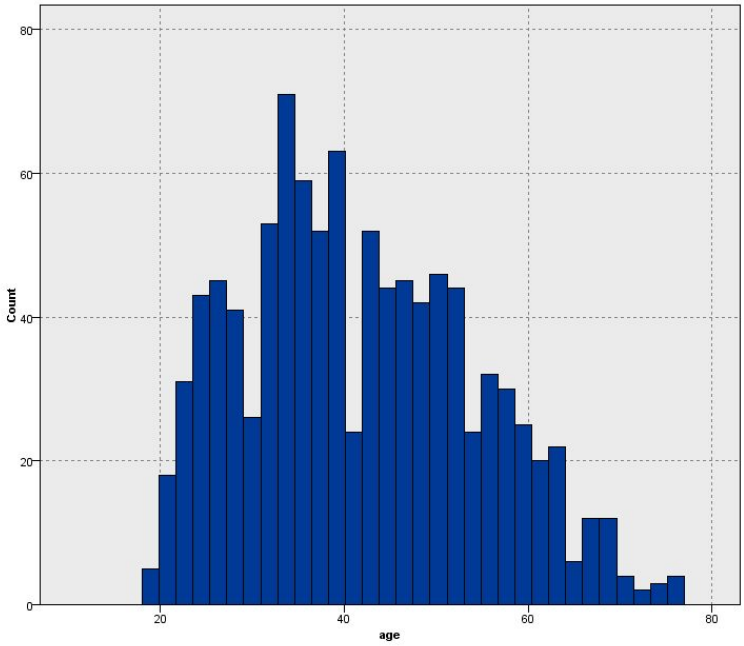
The first thing I did, as I saw in the videos, was create a **web graph** and **data audit**. To do so, first for the web, I click on **Graphs** → **Web**, and select only those which data is 1 or 0, there are 18 fields like that. And the result is the following:



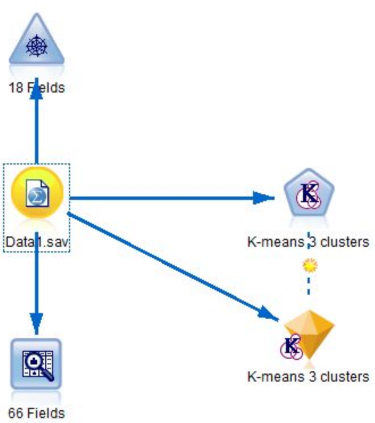
Links	Field 1	Field 2
699	pager = "0.000000"	retire = "0.000000"
682	churn = "0.000000"	retire = "0.000000"
667	retire = "0.000000"	wireless = "0.000000"
658	retire = "0.000000"	voice = "0.000000"
652	pager = "0.000000"	wireless = "0.000000"
637	callcard = "1.000000"	retire = "0.000000"
623	pager = "0.000000"	voice = "0.000000"
619	wireless = "0.000000"	voice = "0.000000"
591	ebill = "0.000000"	retire = "0.000000"
590	internet = "0.000000"	retire = "0.000000"
574	equip = "0.000000"	retire = "0.000000"
562	churn = "0.000000"	pager = "0.000000"
543	internet = "0.000000"	pager = "0.000000"
543	callcard = "1.000000"	churn = "0.000000"
538	churn = "0.000000"	wireless = "0.000000"
538	churn = "0.000000"	voice = "0.000000"
532	equip = "0.000000"	pager = "0.000000"
531	ebill = "0.000000"	pager = "0.000000"
530	internet = "0.000000"	wireless = "0.000000"
529	equip = "0.000000"	internet = "0.000000"
523	ebill = "0.000000"	internet = "0.000000"
520	ebill = "0.000000"	equip = "0.000000"
520	equip = "0.000000"	wireless = "0.000000"
517	ebill = "0.000000"	wireless = "0.000000"
517	churn = "0.000000"	internet = "0.000000"
516	internet = "0.000000"	voice = "0.000000"
512	churn = "0.000000"	ebill = "0.000000"
512	churn = "0.000000"	equip = "0.000000"
507	ebill = "0.000000"	voice = "0.000000"
506	retire = "0.000000"	tollfree = "0.000000"
505	equip = "0.000000"	voice = "0.000000"
503	multline = "0.000000"	retire = "0.000000"
498	callid = "0.000000"	retire = "0.000000"
496	callwait = "0.000000"	retire = "0.000000"
487	forward = "0.000000"	retire = "0.000000"
487	gender = "1.000000"	retire = "0.000000"
481	confer = "0.000000"	retire = "0.000000"
480	marital = "1.000000"	retire = "0.000000"

With this web we can see the relation between each field, so the darker lines are the services that are usually bought or not bought together. If we click on **Show web output summary and control** we can see it clearer. For example, if we focus on the **retired** people, we can see that they do **not** usually bought a **paging service**; so maybe it would be a better idea to offer them other services such as **calling card** and exclude the paging one.

The second thing was to create the **data audit**. For this I used the 66 fields. This generated several graphs with useful information. If we see the **Histogram of age**, we can see that the people who are between **30** and **40 years old** usually buy more services than the others because the count in this range is always above 50.

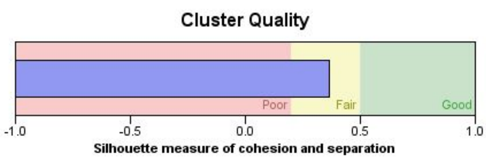


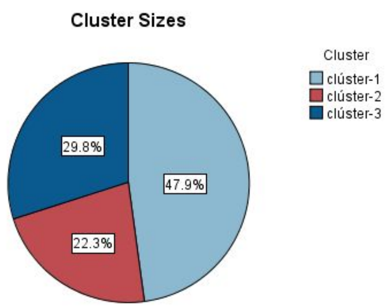
The final step and the most important one is to use clustering to find subsets of “similar” customers. To do so, we have to use the **K-means** tool in **Modeling**, once we connect the data and run it **using predefined roles** (this will take the 66 fields of the data) and **3 clusters.**



Model Summary

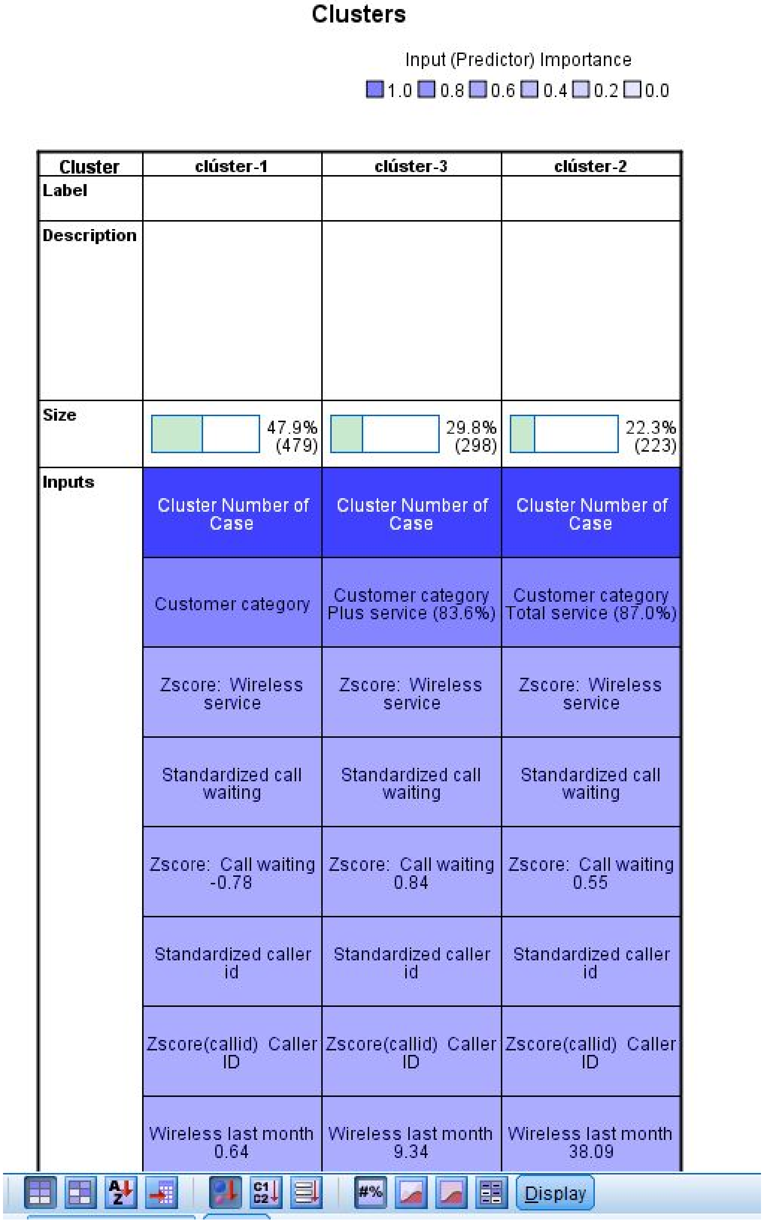
Algorithm	K-Means
Inputs	66
Clusters	3





Size of Smallest Cluster	223 (22.3%)
Size of Largest Cluster	479 (47.9%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	2.15

As we can see, the **Cluster Quality** is not really good, just a **0.4**. If we select the **Clusters view** instead, we can see that there is a lot of fields that **do not have** any **importance**, such as the **Geographic indicator**.



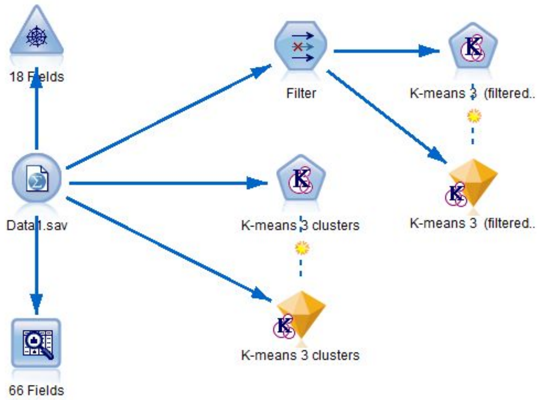
Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	clúster-1	clúster-3	clúster-2
Label			
Description			
Size	47.9% (479)	29.8% (298)	22.3% (223)
Inputs	Cluster Number of Case	Cluster Number of Case	Cluster Number of Case
	Customer category	Customer category Plus service (83.6%)	Customer category Total service (87.0%)
	Zscore: Wireless service	Zscore: Wireless service	Zscore: Wireless service
	Standardized call waiting	Standardized call waiting	Standardized call waiting
	Zscore: Call waiting -0.78	Zscore: Call waiting 0.84	Zscore: Call waiting 0.55
	Standardized caller id	Standardized caller id	Standardized caller id
	Zscore(callid) Caller ID	Zscore(callid) Caller ID	Zscore(callid) Caller ID
	Wireless last month 0.64	Wireless last month 9.34	Wireless last month 38.09

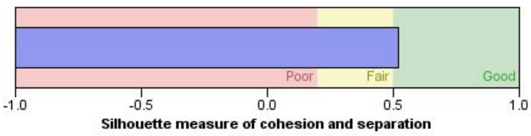
To solve this, I take the **20 fields more important** of the data using a **Filter (Field Ops → Filter)** and generate again the K-means, this is the result:



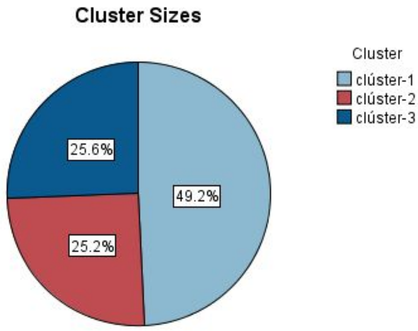
Model Summary

Algorithm	K-Means
Inputs	20
Clusters	3

Cluster Quality



Cluster Sizes



Size of Smallest Cluster	252 (25.2%)
Size of Largest Cluster	492 (49.2%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.95

The **Cluster Quality is now 0.5** which means that is **good** enough, and we have 3 groups that we can distinguish from one another.

Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	clúster-1	clúster-3	clúster-2
Label	Non Profitable Customers	Medium Customers	Best Customers (More Profitable)
Description	This group is the one that bought less services	This is a medium group, the people in here did buy services but with moderation	This group is the one that bought more services
Size	49.2% (492)	25.6% (256)	25.2% (252)
Inputs	Wireless last month 1.63	Wireless last month 3.02	Wireless last month 39.73
	Zscore: Wireless last month	Zscore: Wireless last month	Zscore: Wireless last month
	Zscore: Wireless service	Zscore: Wireless service	Zscore: Wireless service
	Standardized caller id	Standardized caller id	Standardized caller id
	Zscore(callid) Caller ID	Zscore(callid) Caller ID	Zscore(callid) Caller ID
	Standardized call waiting	Standardized call waiting	Standardized call waiting
	Zscore: Call waiting -0.79	Zscore: Call waiting 0.73	Zscore: Call waiting 0.80
	Standardized paging -0.45	Standardized paging -0.44	Standardized paging 1.33

Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	clúster-1	clúster-2	clúster-3
Label	Non Profitable Customers	Best Customers (More Profitable)	Medium Customers
Description	This group is the one that bought less services	This group is the one that bought more services	This is a medium group, the people in here did buy services but with moderation
Size	49.2% (492)	25.2% (252)	25.6% (256)
Inputs	Wireless last month 1.63	Zscore: Wireless service	Standardized call waiting
	Zscore: Wireless last month	Standard m **Zscore: Call waiting Importance = 0.80 Mean: 0.73**	waiting
	Standardized caller id	Zscore(voice) Voice mail	Standardized call forwarding
	Zscore(callid) Caller ID	Standardized paging 1.33	Zscore(forward) Call forwarding
	Standardized call waiting	Zscore(pager) Paging service	Zscore: Wireless last month
	Zscore: Call waiting -0.79	Wireless last month 39.73	Wireless last month 3.02
	Zscore: Toll free service	Zscore: Wireless last month	Standardized caller id
	Standardized call forwarding	Standardized caller id	Zscore(callid) Caller ID

By looking into the **means** of the first group we can see that their values are really low, even negative, this means that there are **few people using the services**, which is really bad because is the biggest cluster. The **third cluster** is the one of the people that bought more services, so is **the most profitable one**.

Also if we sort the input by within cluster importance we discover that the **second cluster** prioritize the most the **wireless service** and the **third** one, the **calling service**.

3. Conclusion

Using the clusters can make differentiate groups of clients and learn about their needs. The better the quality of the cluster is, the most accurate the information provided would be. Also, as we saw before, clustering is not the only tool we have (even though is a very powerful one), we can also use Web graphs and Data audit, as many others.
I think that the **most important thing** is to know which part of the data is useful and which data is not, because using non relevant data can lead into errors and get information not as accurate as it should be.

4. Link to the Source

The file is upload in moodle but also in my Google Drive:
https://drive.google.com/open?id=1xGhD8FOWDd13rXNg93sYLuIlYIsMtobA