DATACLYSMIC

101 COLLECTION

NUMERICAL HIERARCHICAL CLUSTER & MALL CASE

R CODING
VERSION 01

THE EXPERIMENT

TITLE:	Numerical hierarchical cluster & Mall case
VERSION:	01
DESCRIPTION:	Use of clusters for better understating a mall base customer.
DATA SOURCE:	https://www.kaggle.com/datasets/akram24/mall-customers
CODE:	PENDING!
COMPLETE POST:	

101 COLECCTION – About

This Dataclysmic collection explore a series of experiments where different chapters of BI are developed with applied cases.
This is the final report, the reader also can have access to the web post and the code that support this document, see the links above.

The 101 COLLECTION seeks to inspire the reader to find meaning in the tools that data and code allows, to think beyond the screen and wonder about the value, to be creative and think over real application of this capabilities.

EXECUTIVE SUMARY

This experiment uses the Hierarchical Cluster in order to understand the characteristic of the principal groups of customers of a given mal (the Mall) in order to find insights and propose actions.

EXPERIMENT NOTES

The process
We run a process of average Hierarchical cluster using the Euclidean method for a data base that describe different characteristics for the customer of the Mall. Age, Annual income, gender, and Spending Score, (Spending.Score..1.100.).

The result
With the data, 5 clusters of customers were identified as you can see on the right.

Initial notes
The power of each cluster (Power Cluster) was determined by the size of the cluster (the proportion among the sample) and the Spending Score. That is, that the biggest clusters with the top Spending Score has more power and are the best performance cluster for the Mall

Main Insights

- There are 3 top performance clusters the 2nd, 3rd, and 4th. (check on Power_cluster section in the right chart)
- The 4th one has the best overall Spending score 82.1, but 2nd, and 3rd, are bigger clusters, that is how the power of this clusters makes this top performers.
- The 5th cluster has an important size 17%, and a attractive profile with a good income but does not has a good Spending Score.
- Gender did not show significance change on the behavior of profile.

KEY TEAKE AWAYS
Identify different clusters allows to recollect different insights for each profile, even more allows to propose craft strategies for each cluster according to the particular context and the company objectives.

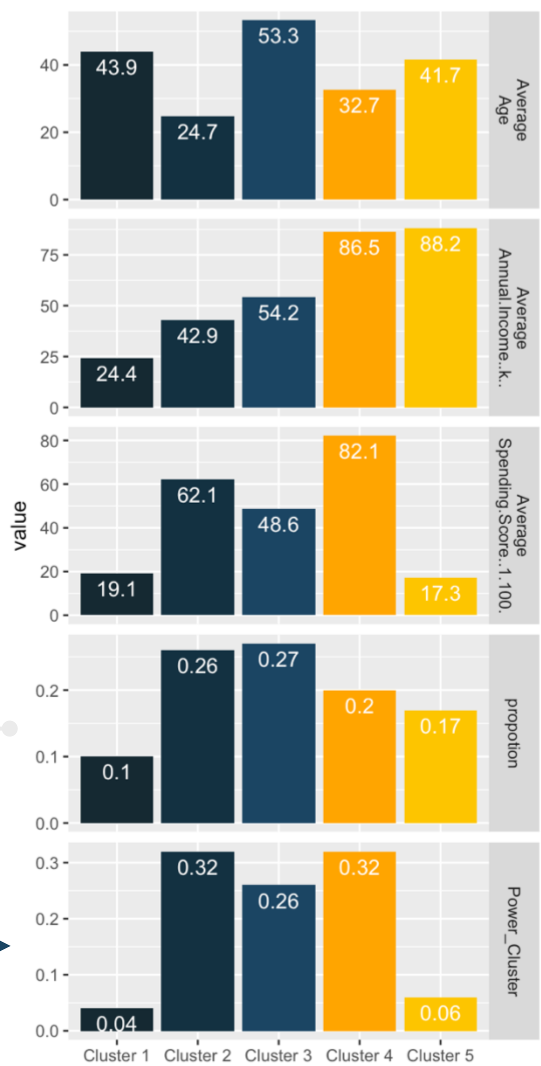
Some of those:

- The 4th cluster has a good spending score but is not the biggest in the mall, this can be challenge for externa communication. Once these clients are in the Mall, they tend to spend, but need to be attracted to the site.
- The 5th cluster shows an attractive profile, and has an important share of the customers, but they don't spend as much. The challenge here can be about the

CLUSTER CHARACTERISTICS

Variables	Type
Gender	character
Age	integer
Annual.Income..k..	integer
Spending.Score..1.100.	integer

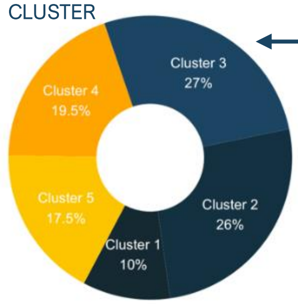
CLUSTERS SUMMARY



POWER CLUSTER ANALISIS

This considers the average wight of the spending score in relation of the size (Proportion) of each cluster.

PROPORTION OF THE CLUSTER



INDEX

1. INTRODUCTION

This document pretends to show the process of a Cluster experiment in a applied case. The link for the R code can be find in the first page.

As always, the interest in Dataclysmic is to conjugate different elements to show value in the of analytics in a feasible context.

THE CASE

We want to target better the mall efforts in marketing, product development and how the mall can focus on efforts of the actual merchants or future ones. The we can clusters can give insights for a former optimization.

2. FIRS LOOK OF THE DATA

Always know your data first. The data set is composed by 4 variables as follows:

	class_summary
Gender	character
Age	integer
Annual.Income..k..	integer
Spending.Score..1.100.	integer

The Spending Score would be our main variable of inters for the case.

3. FIRST CHECK OF OUTLIERS

Age	Annual.Income..k..	Spending.Score..1.100.
Mode :logical	Mode :logical	Mode :logical
FALSE:200	FALSE:200	FALSE:200

All the results turn in to false in this first verification of outliers

4. SCALE THE DATA

The data is scaled to avoid large numbers to dominate the model, you can see the difference in range between the two summary tables

ORIGINAL DATA

Age	Annual.Income..k..	Spending.Score..1.100.
Min. :18.00	Min. : 15.00	Min. : 1.00
1st Qu.:28.75	1st Qu.: 41.50	1st Qu.:34.75
Median :36.00	Median : 61.50	Median :50.00
Mean :38.85	Mean : 60.56	Mean :50.20
3rd Qu.:49.00	3rd Qu.: 78.00	3rd Qu.:73.00
Max. :70.00	Max. :137.00	Max. :99.00

SCALED DATA

This is the summary of the scaled data set. Would be used to build the model and avoid the bias of the large values.

Age	Annual.Income..k..	Spending.Score..1.100.
Min. :-1.4926	Min. :-1.73465	Min. :-1.905240
1st Qu.:-0.7230	1st Qu.:-0.72569	1st Qu.:-0.598292
Median :-0.2040	Median : 0.03579	Median :-0.007745
Mean : 0.0000	Mean : 0.00000	Mean : 0.000000
3rd Qu.: 0.7266	3rd Qu.: 0.66401	3rd Qu.: 0.882916
Max. : 2.2299	Max. : 2.91037	Max. : 1.889750

5. BULD THE CLUSTERS

The for the construction of the cluster the method use is the average Euclidean method

<u>The best partition in is determinate by 5 clusters</u>

6. THE CLUSTERS PROFILE

COMPOSITION OF EACH CLUSTER

Each cluster is composed by 4 variables:

Variables	Type
Gender	character
Age	integer
Annual.Income..k..	integer
Spending.Score..1.100.	integer

SUMMARY OF THE CLUSTERS

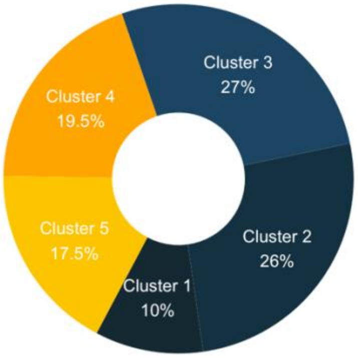
The first statistics and parameters are summarized bellow

Cluster	Count	propotion	Average Age	Average Annual.Income..k..	Average Spending.Score..1.100.	Female	Male
Cluster 1	20	0.100	43.9	24.4	19.1	60.0%	40.0%
Cluster 2	52	0.260	24.7	42.9	62.1	61.5%	38.5%
Cluster 3	54	0.270	53.3	54.2	48.6	59.3%	40.7%
Cluster 4	39	0.195	32.7	86.5	82.1	53.8%	46.2%
Cluster 5	35	0.175	41.7	88.2	17.3	42.9%	57.1%

The proportion field reflects the size of the cluster in the sample.

7. CLUSTER ANALISIS

The objective in the case is to identify insights about the customers, punctually the analysis would be focus on the Spending Score ("Spending.Score..1.100"), this will indicate the performance and interest of action for each cluster"



SIZE OF THE CLUSTERS

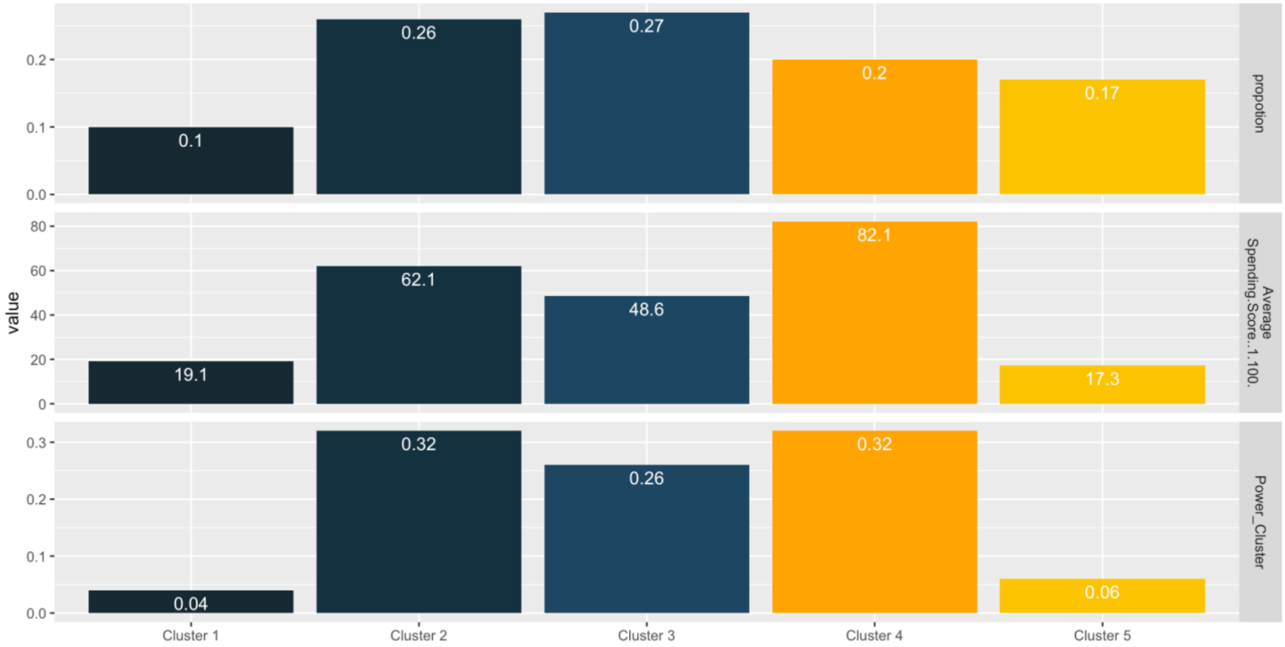
This has impact in given the decisions that can be taken by the mall having in to account the size of the clusters

THE POWER CLUSTER ANALISIS

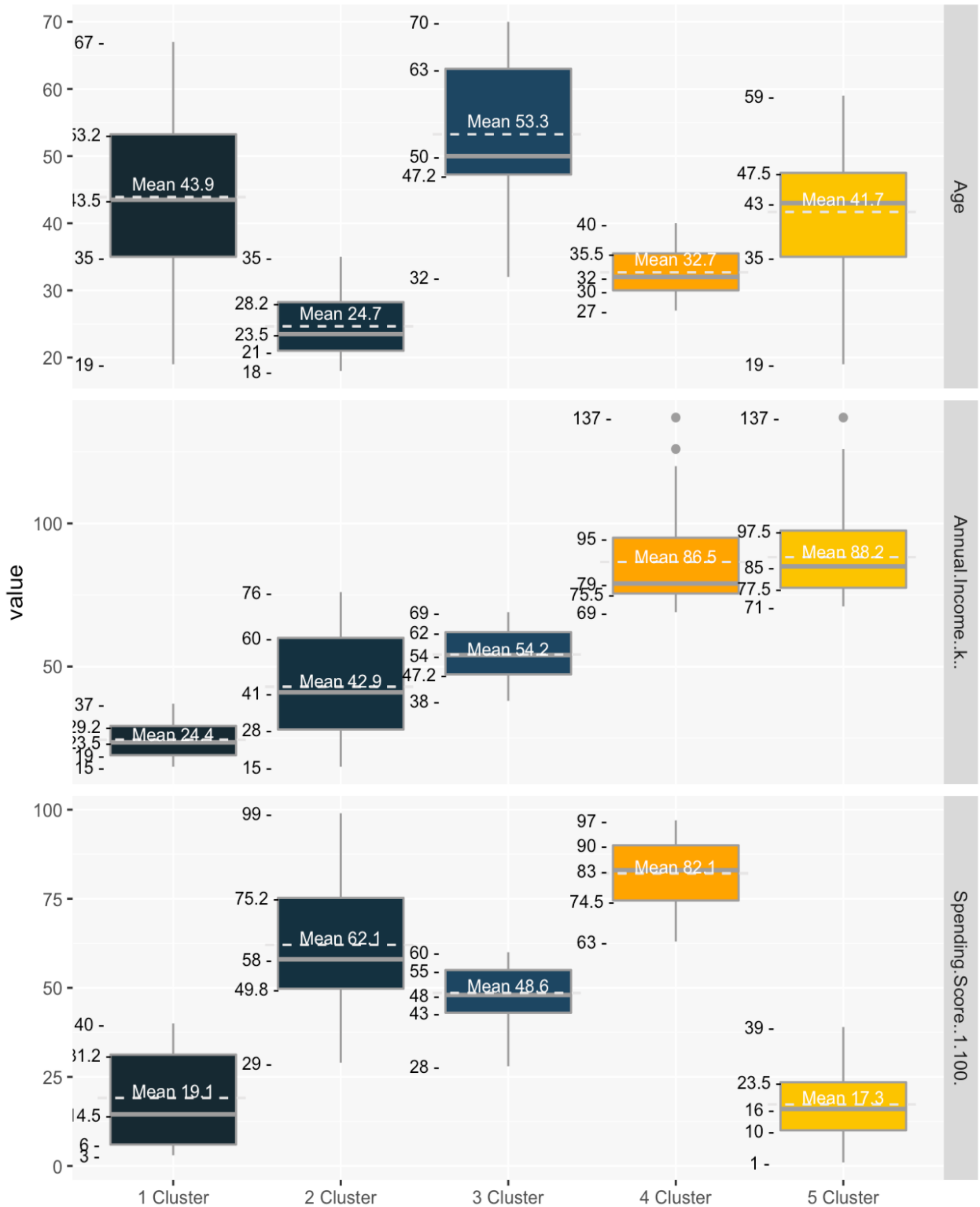
The Power cluster shows a weighted average of the Spending Score. This show, as the name says, the power of each cluster when you consider the size and the willing to spend.

The best performance cluster is the 4th one with an Average spending Score of 82,1. Nevertheless the consideration the proportion of the clusters the 2,3 adds the 53% of the compositions, that result in the 32% and 26% respectively, of the weighted averages of the spending score. As it is shown in the Power_Cluster section bellow.

This bring up front that the main clusters are 2,3 and 4 in spending impact for the Mall, and not just the 4th one.


value
propotion
Average Spending.Score..1.100.
Power_Cluster
0.1
0.26
0.27
0.2
0.17
19.1
62.1
48.6
82.1
17.3
0.04
0.32
0.26
0.32
0.06
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5

8. THE CLUSTERS PROFILE



TAKE AWAYS ABOUT THE EXPERIMENT

With the the previews information, there are a few things that can be said, among many others and in a different and dipper analysis that goes beyond this document.

The main point is that cauterization allows companies and department to identify profiles - not just for customers, think about speech (NLP) or human work force (think about application in RRHH), - and from there construct different strategies for each segment that is according to the particular context and the objects of the organizations.

Below you can find some illustrative examples.

OPORTUNITIES: ABOUT THE WORST PERFORMERS

The worst performer clusters (the lowest Spending Score) are 1 and 5.

Cluster 1:
Could be because of the intuitive reason of a low income and the consequences that this has. In the case that this is a strategic segment according to the Mall business model it would be worth to analyze the services and products that the mall is offering to fulfill accessible goods for adults between 35 and 53. Probably, because of the range of age, families looking for less expensive experiences.

However, this cluster is the 10% of the base customers of the Mall, perhaps not making it for a significance participation, and risking a conservative return against the efforts of the administration. Finance calculations should join these efforts to calculate the respective indicators, risk, and projections.

Cluster 5:
Different from the 1st cluster, this segment has a better income, and represent the 17% of the participation of the Mall costumer's base. Here the challenge is that the Mall already has the customers, but they are not expending as much as the top performers in Spending Score. They have the market, the "right" profile, just they are not buying the other clusters.

If the Mall, considering that already has the costumers, should thin in strategies to offer more experiences and products for this Cluster. Should, analyze if the brans and merchants fulfill the necessities for adults between 35 and 47 with a good income: ¿They are missing luxury brands? ¿Do the services that they look are not in the Mall? Or figure out is something in the customer experiences is going wrong. ¿How is the Mall VOC program?

OPORTUNITIES: ABOUT THE BEST PERFORMERS

The previous does not mean to neglect the other clusters, even more, the top performers (best size and spend score) can be motivated for attract other segments or customers with propositions, discounts, and so on that involves interaction. Marketing can explore some strategies.

This clusters can develop into ambassadors and help to grow not just they own cluster but the other segments as well if the correct strategies are executed.

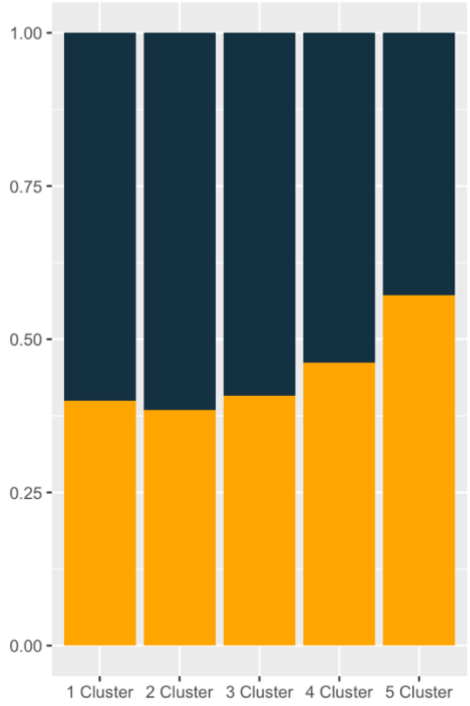
About Cluster 04, that sems a good profile of consumer, the participation in the total share of the customer is not highest, that meaning that more costumers like this can be attracted to the company. Once this segment is in the Mall, the spend. Said that it can be more a external communication problem. Marketing could check the interaction with this type of customers.

9. COMPLEMENTARY ANALISIS

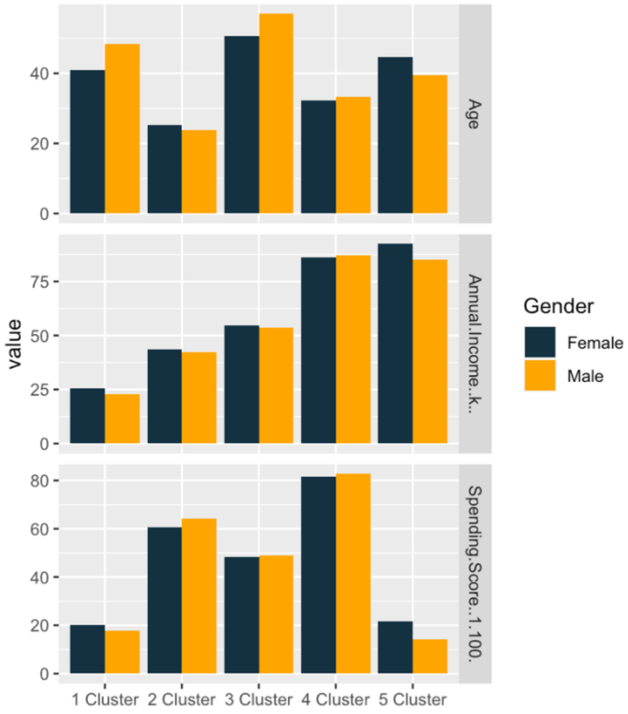
More insights and conclusion can be extracted from the dataset and the case context, in this section some of those are presented.

CATERGORICAL ANALISIS: GENDER BEHAVIOR

CLUSTER GENDER COMPOSITION



BEHAVIOR OF GENDERS AMONG CATHEGORIES

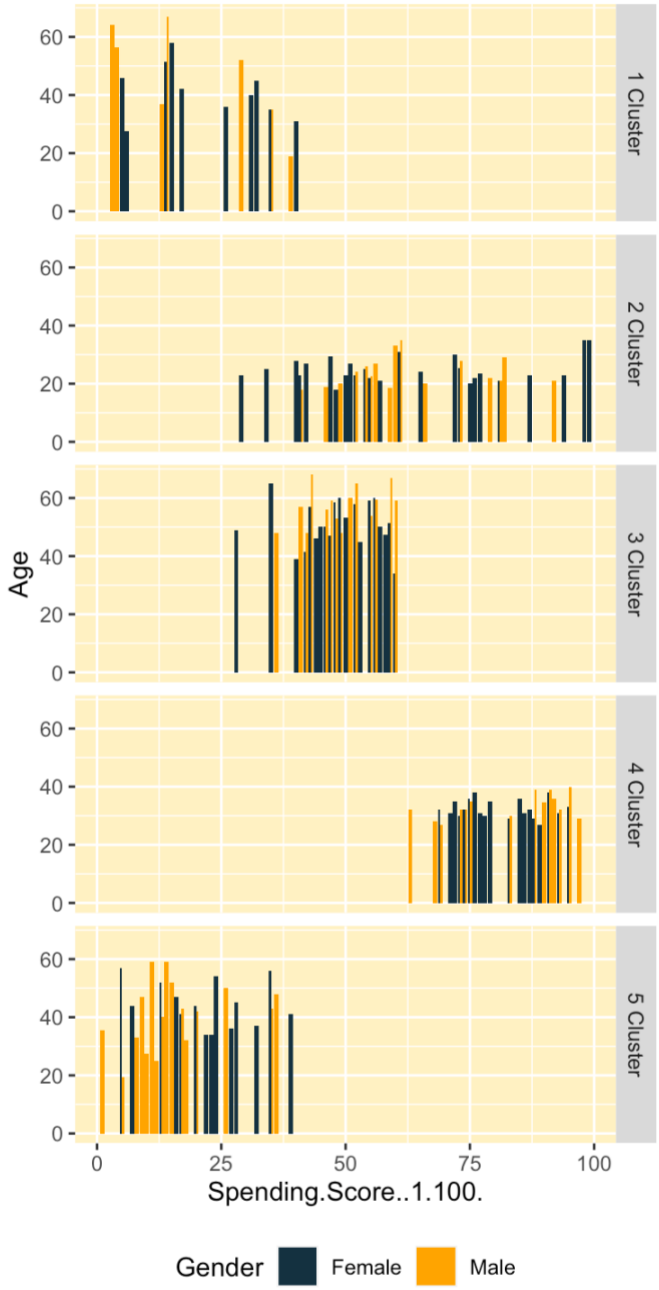


There is not much variable behavior or distribution among the genders compared with the other variables.

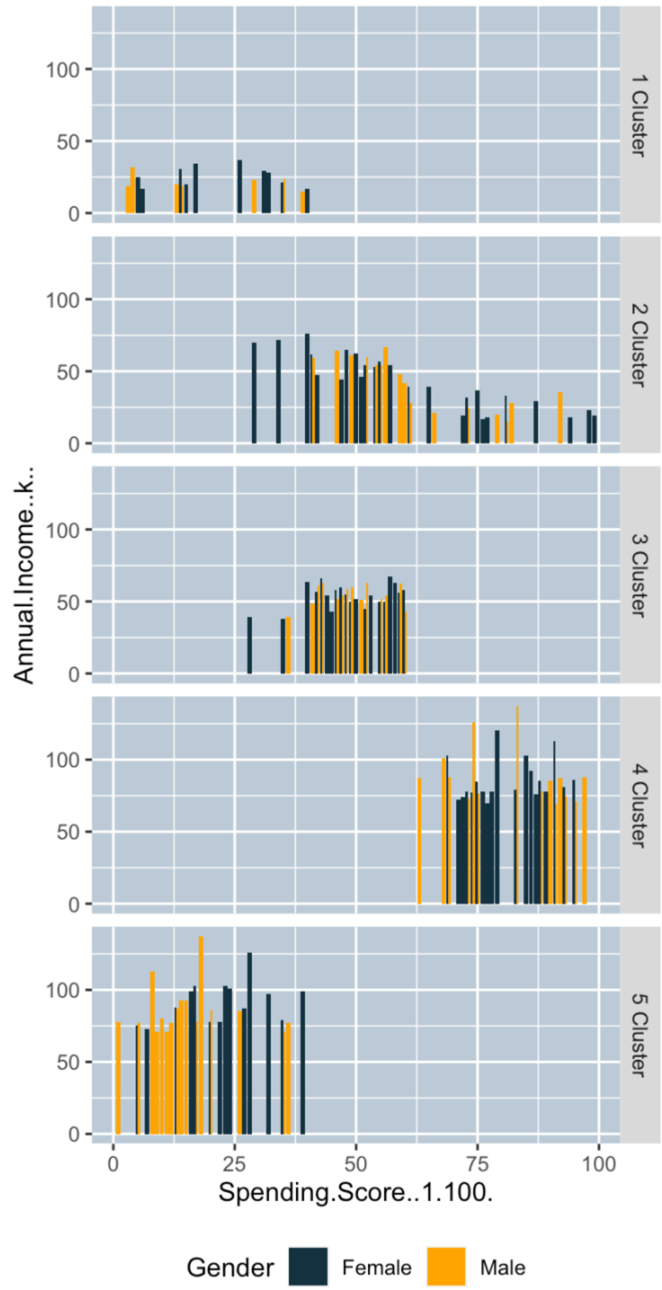
VARIAABLES AGAINST THE SPENDING SCORE

This allows to identify patrons in a 2D comparation helps by the cluster classification. At first sight seem that the more disperse is the 2nd Cluster.

AGE Vs. SPENDING SCORE



INCOME Vs. SPENDING SCORE



10. FINAL NOTES AND DISCLAIMER

Some more considerations should be considered for a formal development. Pay attention to the rigor of the math, statistics, and data cleaning in further incursions. E.g., the boxplot showed some probable outliers that should be investigated.

ALAWAYS LOOK FOR VALUE

Finally, it must be reenforce that the idea es how all this creates value in the real world. That is the challenge in this collection, inspire to look real value, make the reader question about the purpose of all this mare magnum of data and technologies, and the why behind it.

DATACLYSMIC

CREATIVITY + DATA + CODE + BUSINESS

WWW.DATACLYSMIC.CO