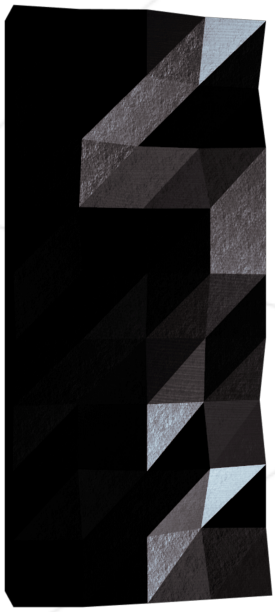


Training Piscine datascience - 2

Data Viz

Summary: Today, you will display the data

Version: 1.00

Contents

I	**General rules**	**2**
II	**Instructions**	**3**
III	**Exercise 00**	**4**
IV	**Exercise 01**	**5**
V	**Exercise 02**	**7**
VI	**Exercise 03**	**9**
VII	**Exercise 04**	**10**
VIII	**Exercise 05**	**11**
IX	**Submission and peer-evaluation**	**14**

Chapter I

General rules

- You have to render your modules from a computer in the cluster either using a virtual machine:
    - You can choose the operating system to use for your virtual machine
    - Your virtual machine must have all the necessary software to realize your project. This software must be configured and installed.
- Or you can use the computer directly in case the tools are available.
    - Make sure you have the space on your session to install what you need for all the modules (use the goinfre if your campus has it)
    - You must have everything installed before the evaluations
- Your functions should not quit unexpectedly (segmentation fault, bus error, double free, etc) apart from undefined behaviors. If this happens, your project will be considered non functional and will receive a `0` during the evaluation.
- We encourage you to create test programs for your project even though this work **won't have to be submitted and won't be graded**. It will give you a chance to easily test your work and your peers' work. You will find those tests especially useful during your defence. Indeed, during defence, you are free to use your tests and/or the tests of the peer you are evaluating.
- Submit your work to your assigned git repository. Only the work in the git repository will be graded. If Deepthought is assigned to grade your work, it will be done after your peer-evaluations. If an error happens in any section of your work during Deepthought's grading, the evaluation will stop.
- By Odin, by Thor ! Use your brain !!!

Chapter II

Instructions

The role of the data analyst is to understand the past and present data, to make graphs to explain to the teams the "numbers" (reporting), and to be a force of proposal to bring solutions, to create tools to help the decision making.

They often use tools such as techno R, Excel, Power BI, Jupyter Notebook ...

It is up to you to find the tools that suit you. You are free to use any language of your choice for this module.

Be careful with this "piscine". Even if you manage to validate a module, you may be stuck later if you haven't cleaned up or stored your data properly.

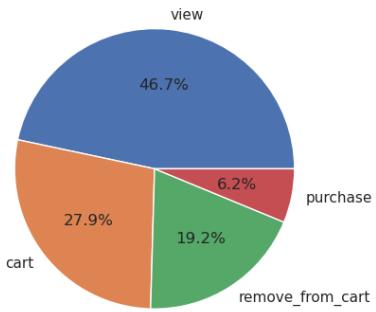
All the graphs are done without the February data, you have to redo them all with the new data

Chapter III

Exercise 00

	Exercise 00
Exercice 00 : American apple Pie	
Turn-in directory : *ex*00/	
Files to turn in : `pie.*`	
Allowed functions : `All`	

- Make your own pie chart to understand what people do on the site
- You have to connect to your Data Warehouse of module 01

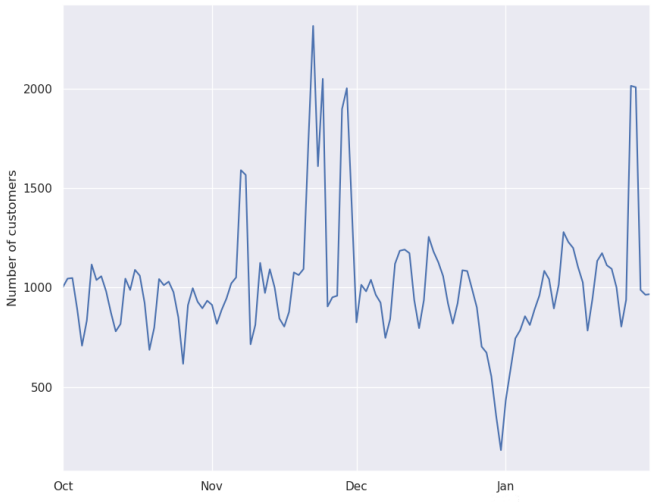


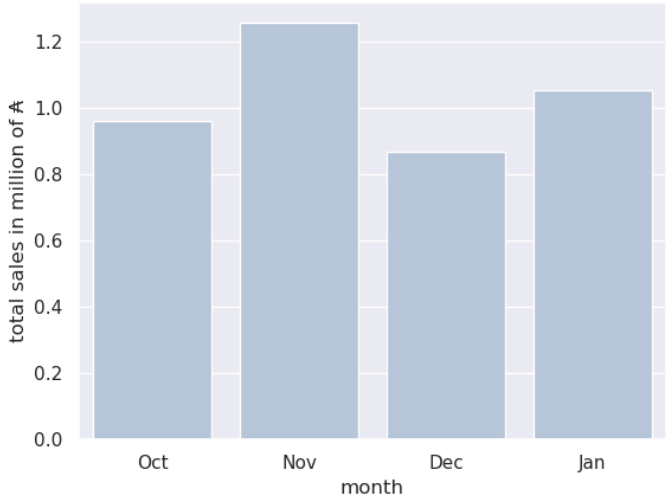
Chapter IV

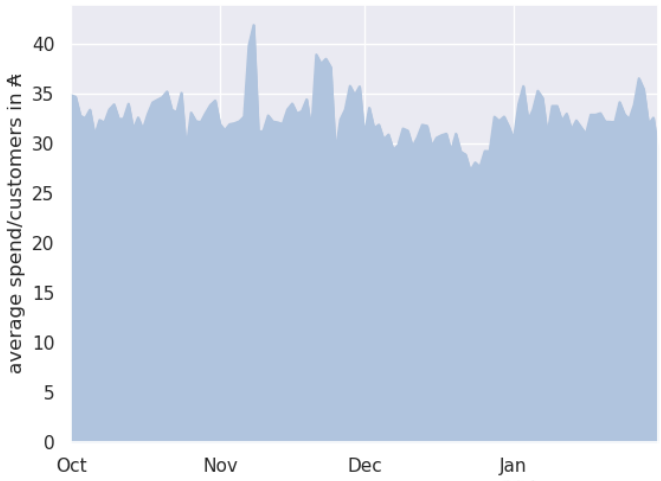
Exercise 01

	Exercise 01
Exercice 01 : initial data exploration	
Turn-in directory : $ex01/$	
Files to turn in : `chart.*`	
Allowed functions : `All`	

- Keep only the "purchase" data of "event_type" column
- All prices are in Altairian Dollars, you know the money that is your wallet ?
- You have to create 3 charts from the beginning of October 2022 to the end of February 2023 like the ones below



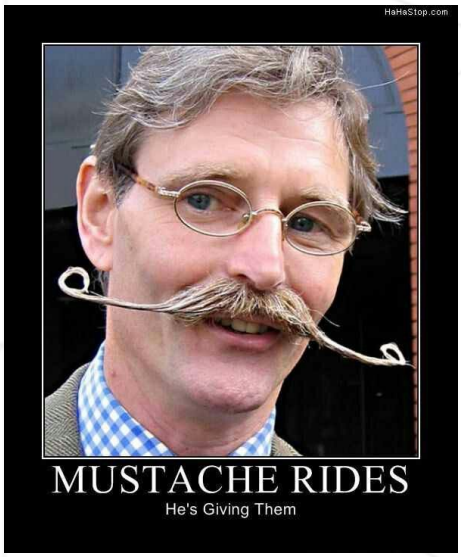

1.2
1.0
0.8
0.6
0.4
0.2
0.0
total sales in million of ₳
Oct
Nov
Dec
Jan
month


40
35
30
25
20
15
10
5
0
average spend/customers in ₳
Oct
Nov
Dec
Jan

Chapter V

Exercise 02

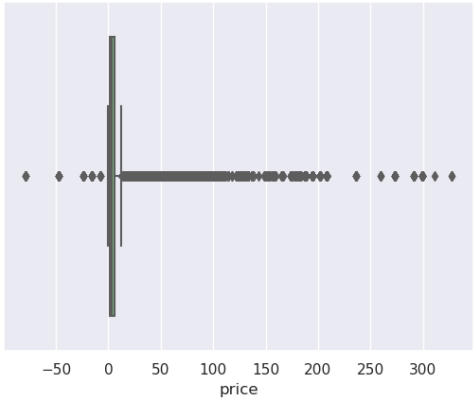
	Exercise 02
Exercice 02 : My beautiful mustache	
Turn-in directory : $ex02/$	
Files to turn in : `mustache.*`	
Allowed functions : `All`	

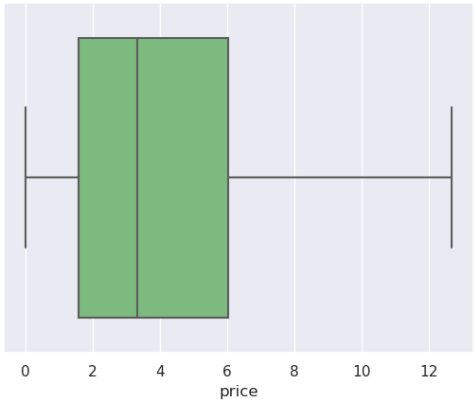


- Print the mean, median, min, max, first, second and third quartile of the price of the items purchased
- Make box plots that display the price of the items purchased

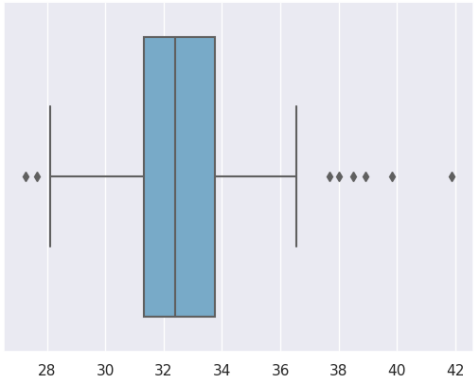
```
count   741644.000000
mean         5.575068
std         10.264594
min        -79.370000
25%          1.590000
50%          3.330000
75%          6.030000
max        327.780000
```

Expected output:





Then a box plot with the average basket price per user

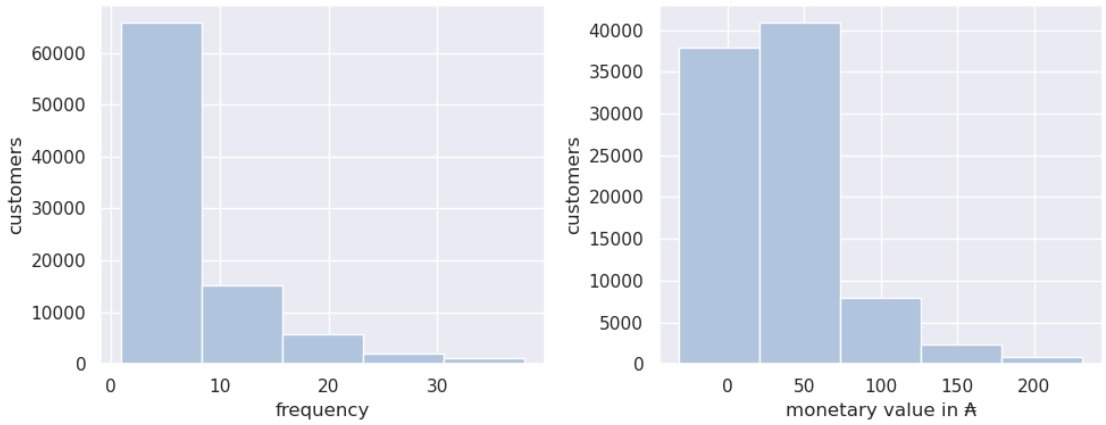


Chapter VI

Exercise 03

	Exercise 03
Exercice 03 : Highest Building	
Turn-in directory : $ex03/$	
Files to turn in : `Building.*`	
Allowed functions : `All`	

- made a bart chart with the number of orders according to the frequency
- made a bart chart the Altairian Dollars spent on the site by customers



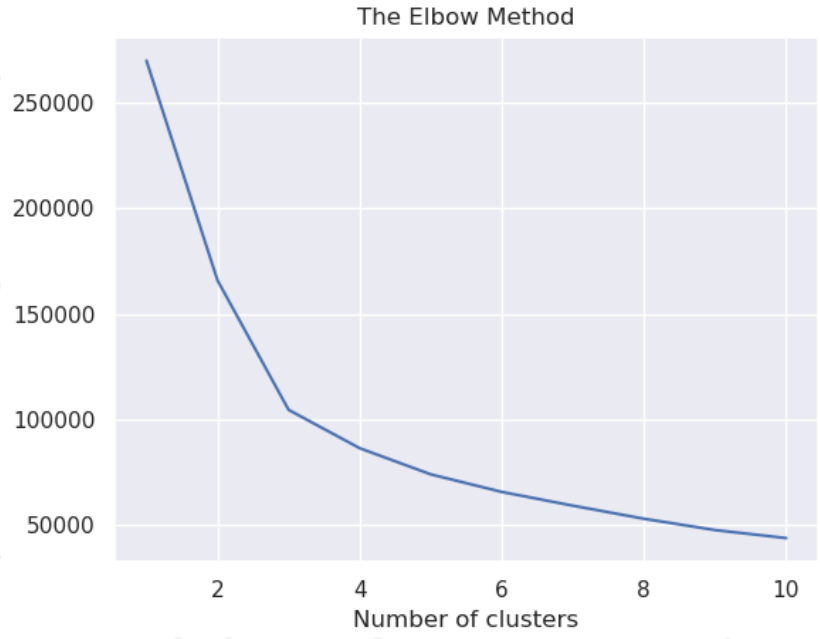
Chapter VII

Exercise 04

	Exercise 04
Exercice 04 : Elbow	
Turn-in directory : $ex04/$	
Files to turn in : `elbow.*`	
Allowed functions : `All`	

Your Boss wants to make groups of customer type to make a commercial targeting with offers by e-mails

- Make an Elbow Method to understand the 'optimal' number of clusters to make
- You have to be able to explain how many clusters you choose and why



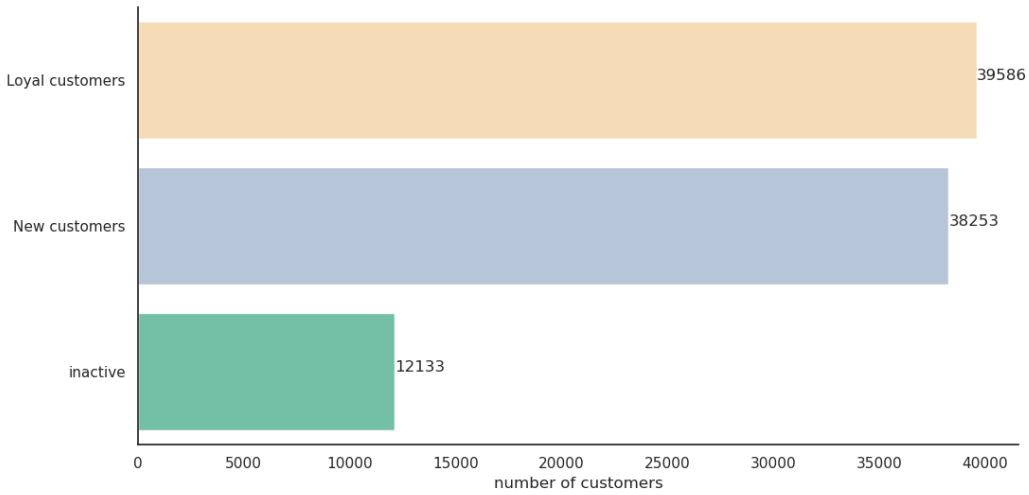
Chapter VIII

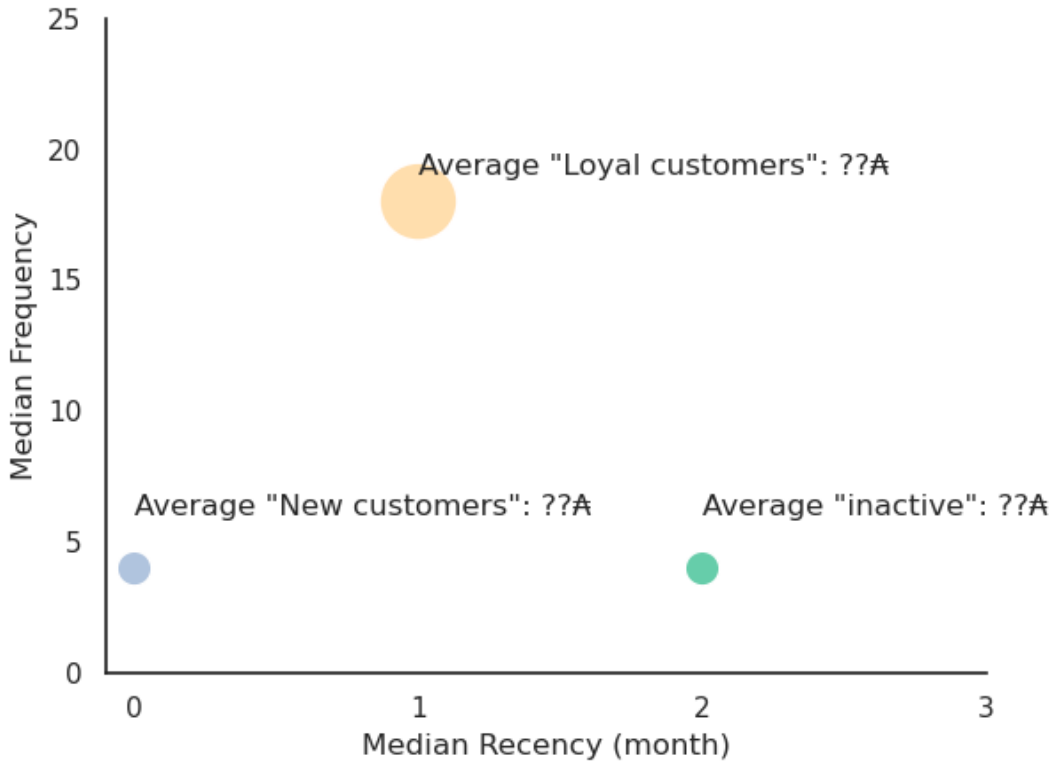
Exercise 05

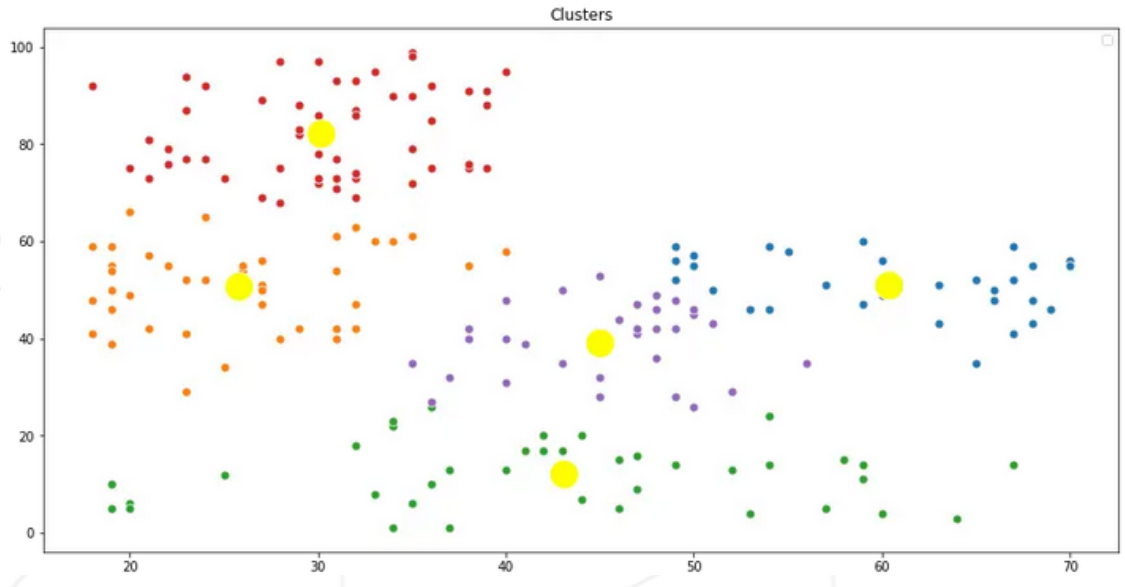
	Exercise 05
Exercice 05 : Clustering	
Turn-in directory : $ex05/$	
Files to turn in : `Clustering.*`	
Allowed functions : `All`	

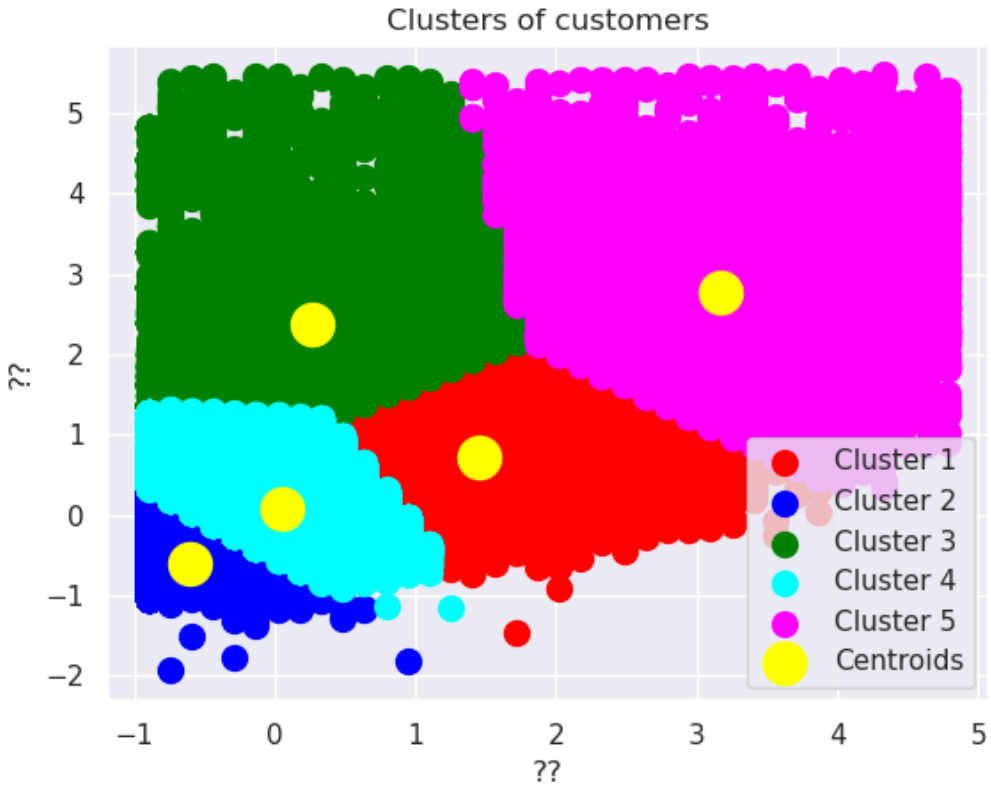
Your boss wants to make groups of type of customer to make a commercial targeting with offers by e-mails (welcome offers for new customers, coupon to bring back old customers, special status for loyal customers like: gold, silver, platinum ...)

- Make at least 4 groups (new customer, inactive customer, loyalty status: gold + silver + platinum ...)
- You have to use a Clustering Algorithms
- Make graphic representations of the groups (minimum 2)




25
20
15
10
5
0
Median Frequency
Average "Loyal customers": ??₳
Average "New customers": ??₳
Average "inactive": ??₳
0
1
2
3
Median Recency (month)


Clusters
100
80
60
40
20
0
20
30
40
50
60
70


Clusters of customers
??
??
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Centroids

Chapter IX

Submission and peer-evaluation

Turn in your assignment in your `Git` repository as usual. Only the work inside your repository will be evaluated during the defense. Don't hesitate to double check the names of your folders and files to ensure they are correct.

`The evaluation process will happen on the computer of the evaluated group.`