Data Mining

October 28th 2020

Manel AGUILAR BARROSO

Daniel CANO CARRASCOSA

Oriol CATASÚS LLENA

Jesus MOLINA ROLDAN

Eduard ORTUÑO GARROTE

Adrià VENTURA HERCE

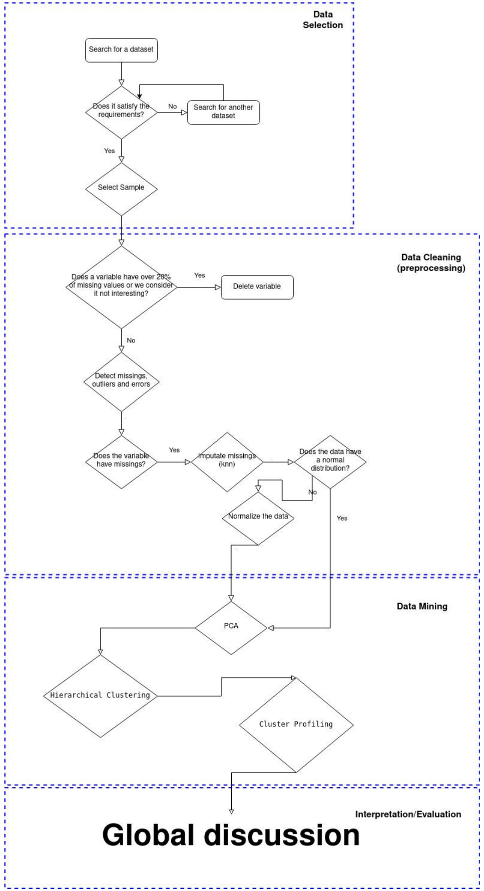
Outline of talk

- Work overview
- Data mining process
- Descriptive analysis
- Univariate descriptive analysis
- Bivariate descriptive analysis
- Preprocessing
- Scree plot
- Factorial map visualization
- Relationship among variables
- Conclusions of PCA
- Clustering process
- Tools of class interpretation used
- Profiling graphs
- Final class profiling
- PCA and Hierarchical Clustering
- Conclusions
- Original and final scheduling

Graphics Processing Units (GPUs)

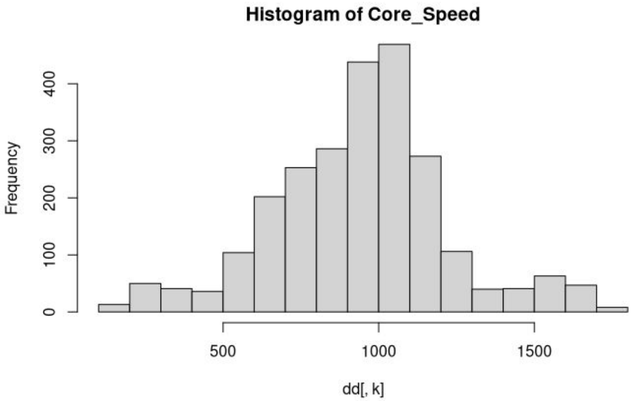
- GPU is one of the most important part of a computer
- Goal of the project:
    - Deepen our understanding about GPU
    - Learning how its different features are related with each other
- Data Overview
    - 3406 models of GPUs, from 2000 until 2017
    - 21 variables
    - About 8.03% of nulls

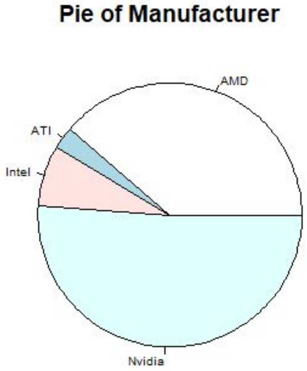
Data Mining process

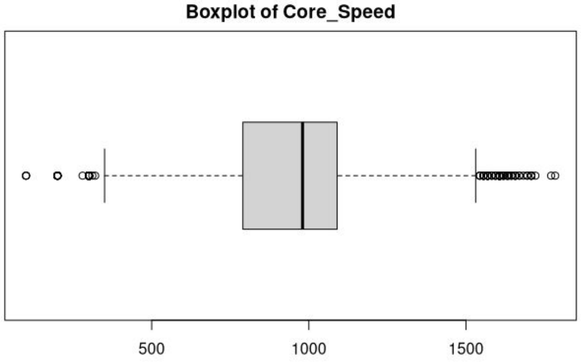

Data Selection
Search for a dataset
Does it satisfy the requirements?
No
Search for another dataset
Yes
Select Sample
Data Cleaning (preprocessing)
Does a variable have over 20% of missing values or we consider it not interesting?
Yes
Delete variable
No
Detect missings, outliers and errors
Does the variable have missings?
Yes
Imputate missings (knn)
Does the data have a normal distribution?
No
Normalize the data
Yes
Data Mining
PCA
Hierarchical Clustering
Cluster Profiling
Interpretation/Evaluation
Global discussion

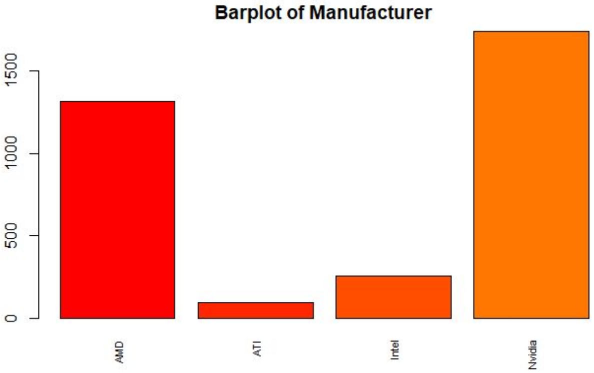
1. Data Selection
2. Data Cleaning
3. Data Mining
4. Interpretation/Evaluation

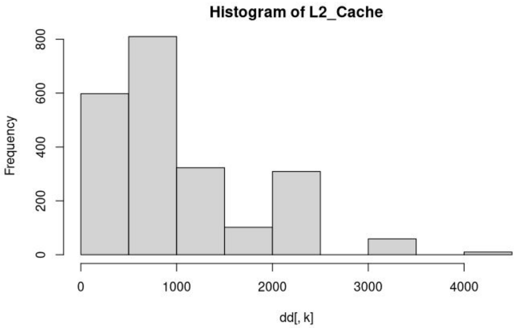
Descriptive analysis

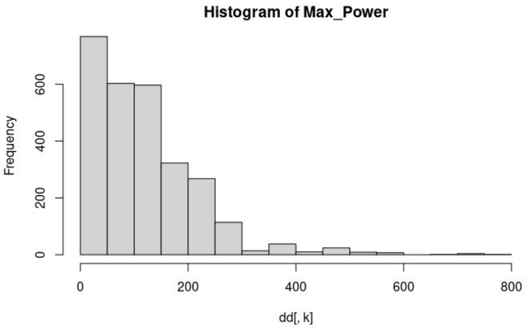


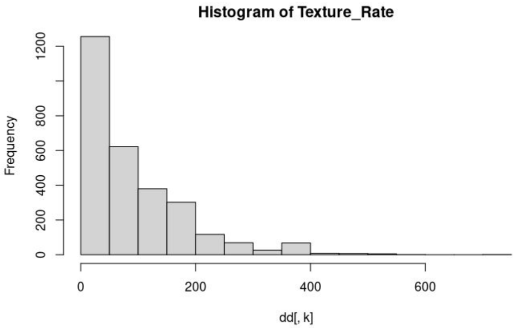


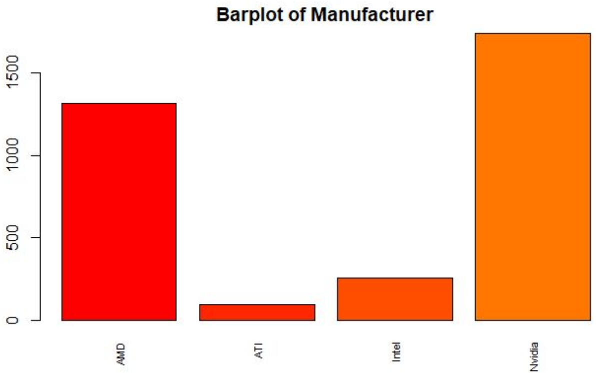


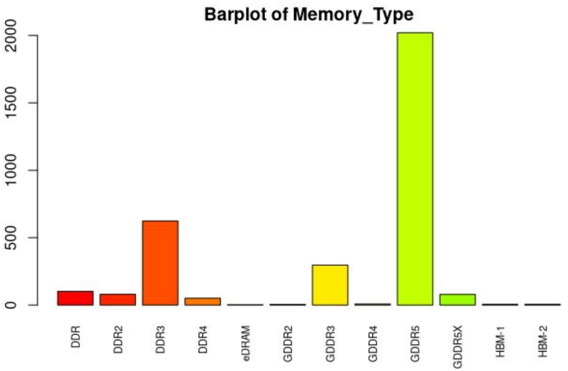



Histogram of L2_Cache
Frequency
0
200
400
600
800
0
1000
2000
3000
4000
dd[, k]

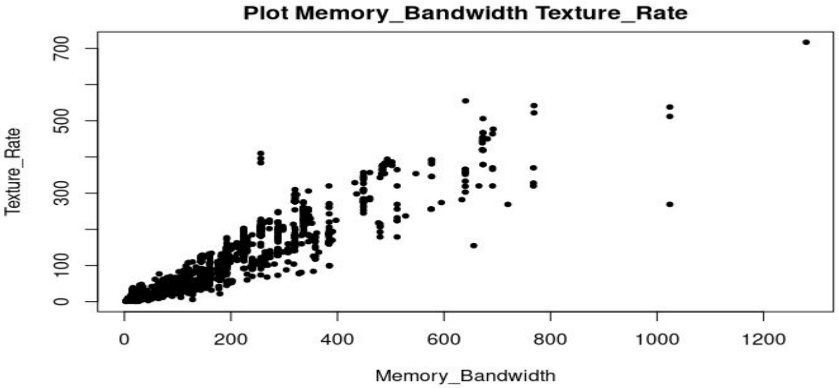

Histogram of Max_Power
Frequency
0
200
400
600
0
200
400
600
800
dd[, k]

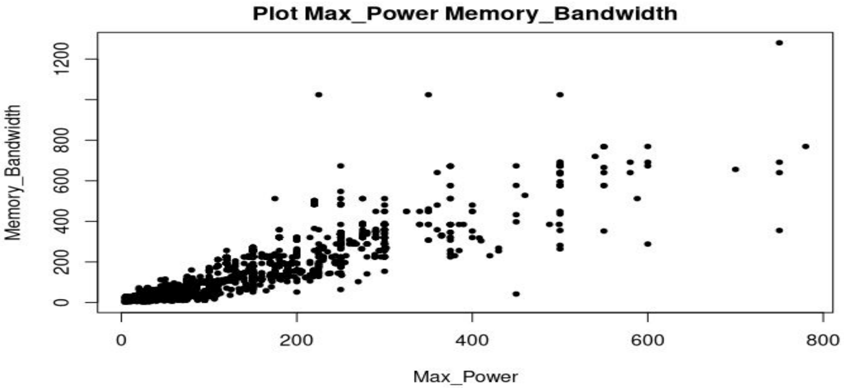

Histogram of Texture_Rate
Frequency
0
200
400
600
800
1200
0
200
400
600
dd[, k]

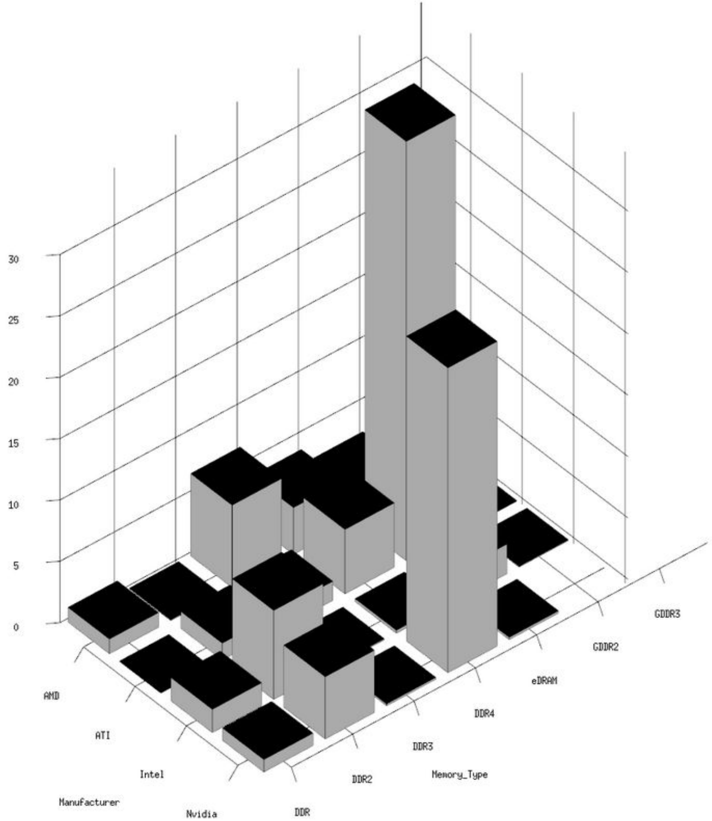

Barplot of Manufacturer
0
500
1000
1500
AMD
ATI
Intel
Nvidia


Barplot of Memory_Type
0
500
1000
1500
2000
DDR
DDR2
DDR3
DDR4
eDRAM
GDDR2
GDDR3
GDDR4
GDDR5
GDDR5X
HBM-1
HBM-2

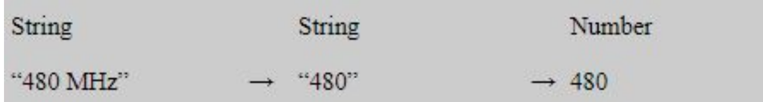
Bivariate descriptive analysis



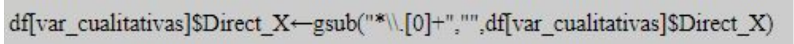


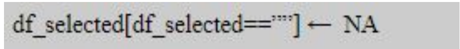


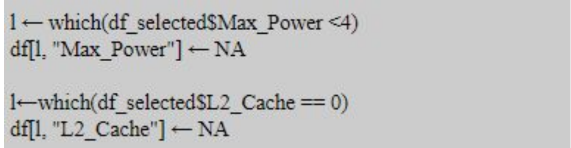
Preprocessing

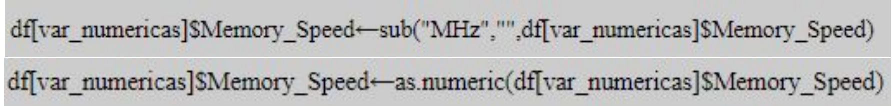

String → String → Number

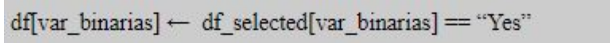
“480 MHz” → “480” → 480

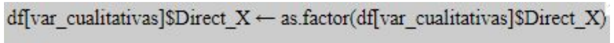

```
df[var_cualitativas]$Direct_X←gsub("*\\.[0]+","",df[var_cualitativas]$Direct_X)
```

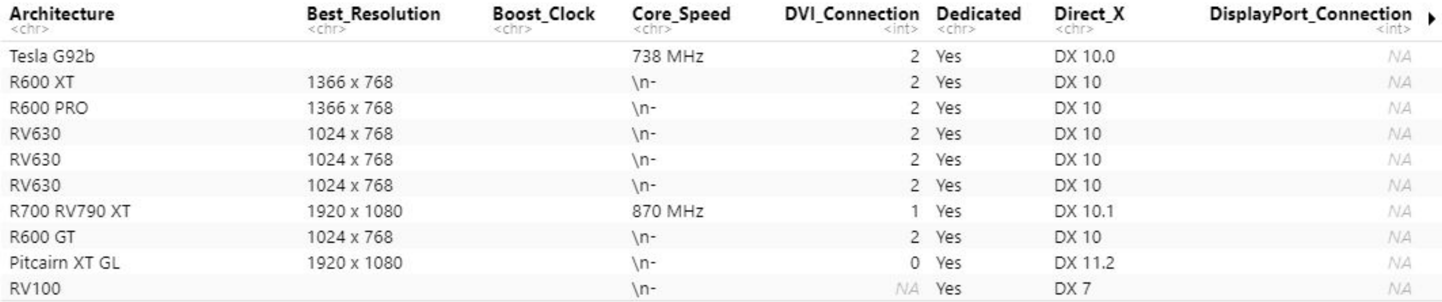

```
df_selected[df_selected==""] ← NA
```


```
l ← which(df_selected$Max_Power <4)
df[l, "Max_Power"] ← NA

l←which(df_selected$L2_Cache == 0)
df[l, "L2_Cache"] ← NA
```

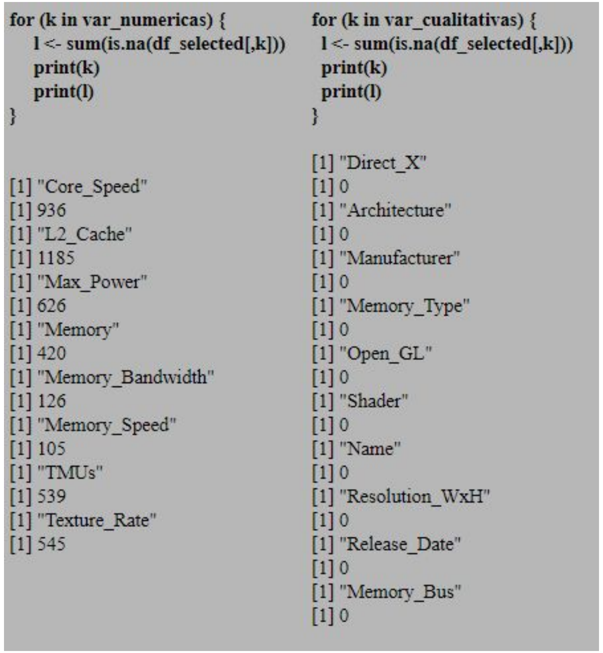

```
df[var_numericas]$Memory_Speed←sub("MHz","",df[var_numericas]$Memory_Speed)
df[var_numericas]$Memory_Speed←as.numeric(df[var_numericas]$Memory_Speed)
```


```
df[var_binarias] ← df_selected[var_binarias] == “Yes”
```


```
df[var_cualitativas]$Direct_X ← as.factor(df[var_cualitativas]$Direct_X)
```

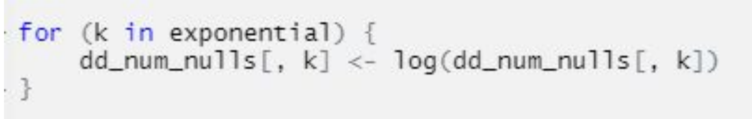


Architecture <chr>	Best_Resolution <chr>	Boost_Clock <chr>	Core_Speed <chr>	DVI_Connection <int>	Dedicated <chr>	Direct_X <chr>	DisplayPort_Connection <int>
Tesla G92b			738 MHz	2	Yes	DX 10.0	NA
R600 XT	1366 x 768		\n-	2	Yes	DX 10	NA
R600 PRO	1366 x 768		\n-	2	Yes	DX 10	NA
RV630	1024 x 768		\n-	2	Yes	DX 10	NA
RV630	1024 x 768		\n-	2	Yes	DX 10	NA
RV630	1024 x 768		\n-	2	Yes	DX 10	NA
R700 RV790 XT	1920 x 1080		870 MHz	1	Yes	DX 10.1	NA
R600 GT	1024 x 768		\n-	2	Yes	DX 10	NA
Pitcairn XT GL	1920 x 1080		\n-	0	Yes	DX 11.2	NA
RV100			\n-	NA	Yes	DX 7	NA

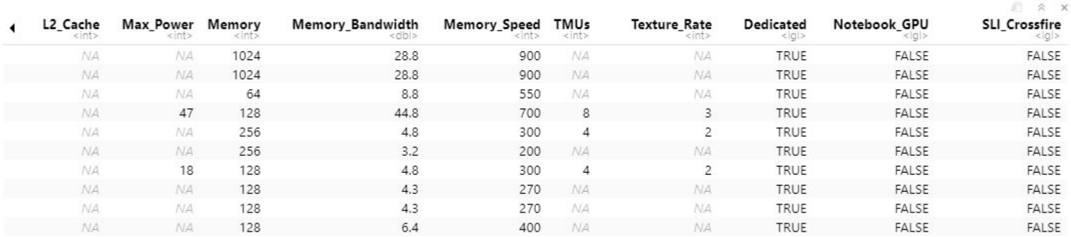


```
for (k in var_numericas) {
    l <- sum(is.na(df_selected[,k]))
    print(k)
    print(l)
}


[1] "Core_Speed"
[1] 936
[1] "L2_Cache"
[1] 1185
[1] "Max_Power"
[1] 626
[1] "Memory"
[1] 420
[1] "Memory_Bandwidth"
[1] 126
[1] "Memory_Speed"
[1] 105
[1] "TMUs"
[1] 539
[1] "Texture_Rate"
[1] 545
```

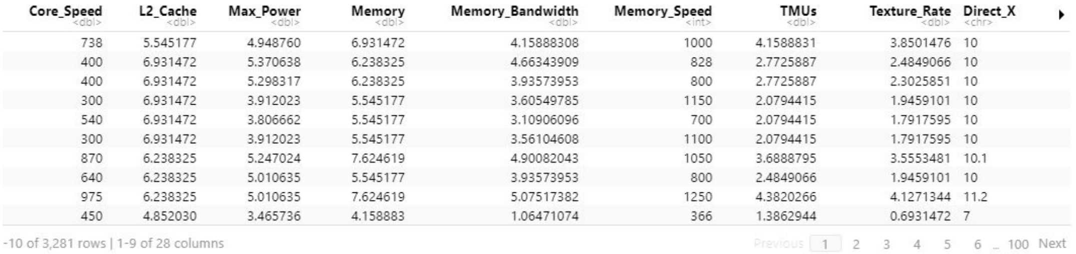
```
for (k in var_cualitativas) {
  l <- sum(is.na(df_selected[,k]))
  print(k)
  print(l)
}

[1] "Direct_X"
[1] 0
[1] "Architecture"
[1] 0
[1] "Manufacturer"
[1] 0
[1] "Memory_Type"
[1] 0
[1] "Open_GL"
[1] 0
[1] "Shader"
[1] 0
[1] "Name"
[1] 0
[1] "Resolution_WxH"
[1] 0
[1] "Release_Date"
[1] 0
[1] "Memory_Bus"
[1] 0
```



```
for (k in exponential) {
    dd_num_nulls[, k] <- log(dd_num_nulls[, k])
}
```

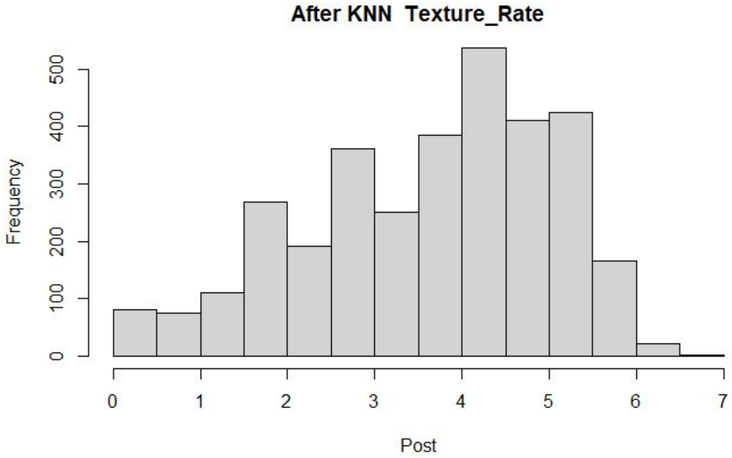


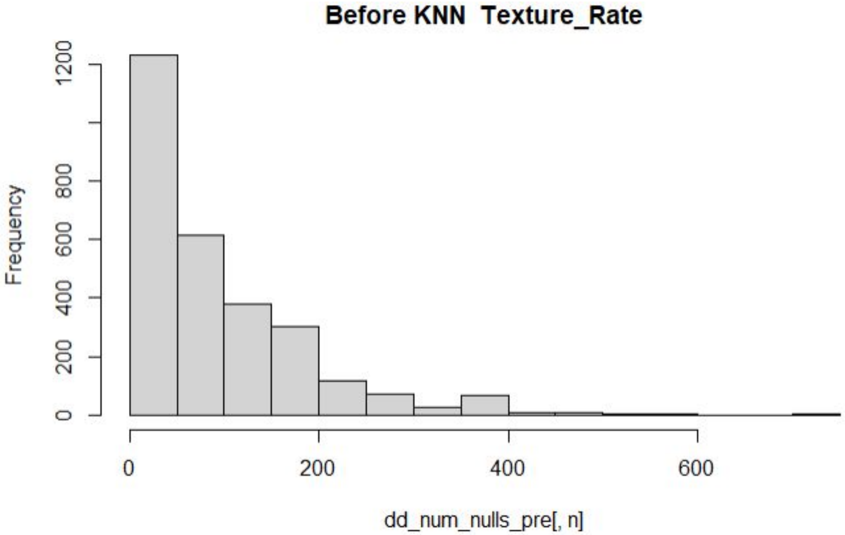
L2_Cache	Max_Power	Memory	Memory_Bandwidth	Memory_Speed	TMUs	Texture_Rate	Dedicated	Notebook_GPU	SLI_Crossfire
<int>	<int>	<int>	<dbl>	<int>	<int>	<int>	<lgl>	<lgl>	<lgl>
NA	NA	1024	28.8	900	NA	NA	TRUE	FALSE	FALSE
NA	NA	1024	28.8	900	NA	NA	TRUE	FALSE	FALSE
NA	NA	64	8.8	550	NA	NA	TRUE	FALSE	FALSE
NA	47	128	44.8	700	8	3	TRUE	FALSE	FALSE
NA	NA	256	4.8	300	4	2	TRUE	FALSE	FALSE
NA	NA	256	3.2	200	NA	NA	TRUE	FALSE	FALSE
NA	18	128	4.8	300	4	2	TRUE	FALSE	FALSE
NA	NA	128	4.3	270	NA	NA	TRUE	FALSE	FALSE
NA	NA	128	4.3	270	NA	NA	TRUE	FALSE	FALSE
NA	NA	128	6.4	400	NA	NA	TRUE	FALSE	FALSE



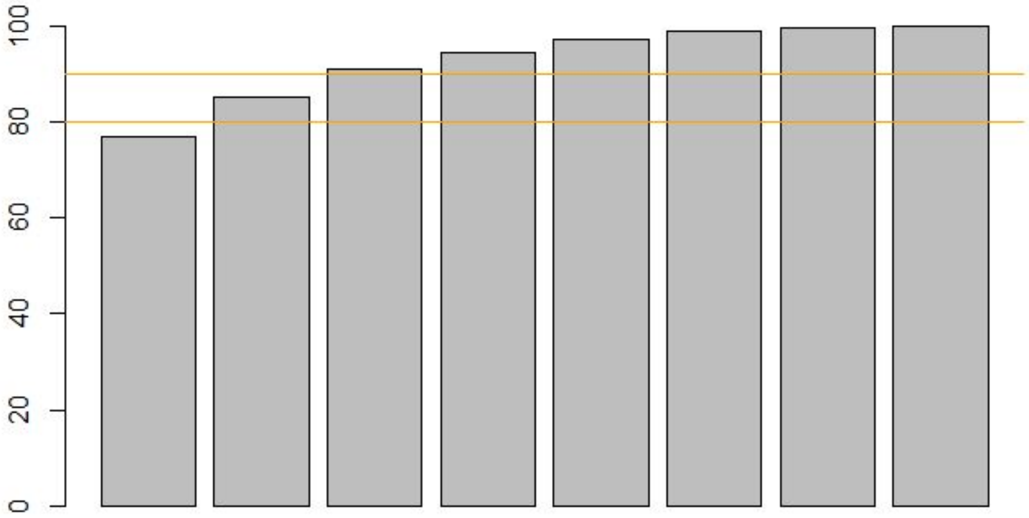
Core_Speed	L2_Cache	Max_Power	Memory	Memory_Bandwidth	Memory_Speed	TMUs	Texture_Rate	Direct_X
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<chr>
738	5.545177	4.948760	6.931472	4.15888308	1000	4.1588831	3.8501476	10
400	6.931472	5.370638	6.238325	4.66343909	828	2.7725887	2.4849066	10
400	6.931472	5.298317	6.238325	3.93573953	800	2.7725887	2.3025851	10
300	6.931472	3.912023	5.545177	3.60549785	1150	2.0794415	1.9459101	10
540	6.931472	3.806662	5.545177	3.10906096	700	2.0794415	1.7917595	10
300	6.931472	3.912023	5.545177	3.56104608	1100	2.0794415	1.7917595	10
870	6.238325	5.247024	7.624619	4.90082043	1050	3.6888795	3.5553481	10.1
640	6.238325	5.010635	5.545177	3.93573953	800	2.4849066	1.9459101	10
975	6.238325	5.010635	7.624619	5.07517382	1250	4.3820266	4.1271344	11.2
450	4.852030	3.465736	4.158883	1.06471074	366	1.3862944	0.6931472	7

-10 of 3,281 rows | 1-9 of 28 columns

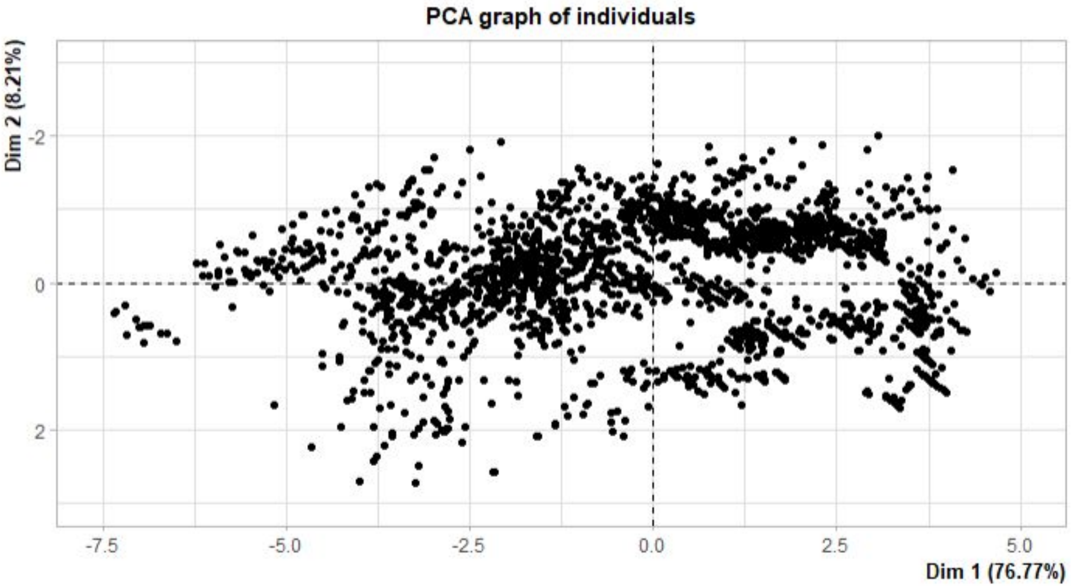

After KNN Texture_Rate
Frequency
0
100
200
300
400
500
0
1
2
3
4
5
6
7
Post

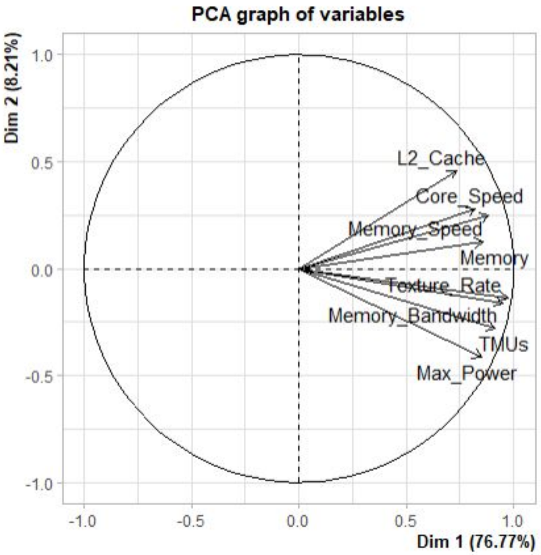

Before KNN Texture_Rate
Frequency
0
200
400
600
800
1200
0
200
400
600
dd_num_nulls_pre[, n]

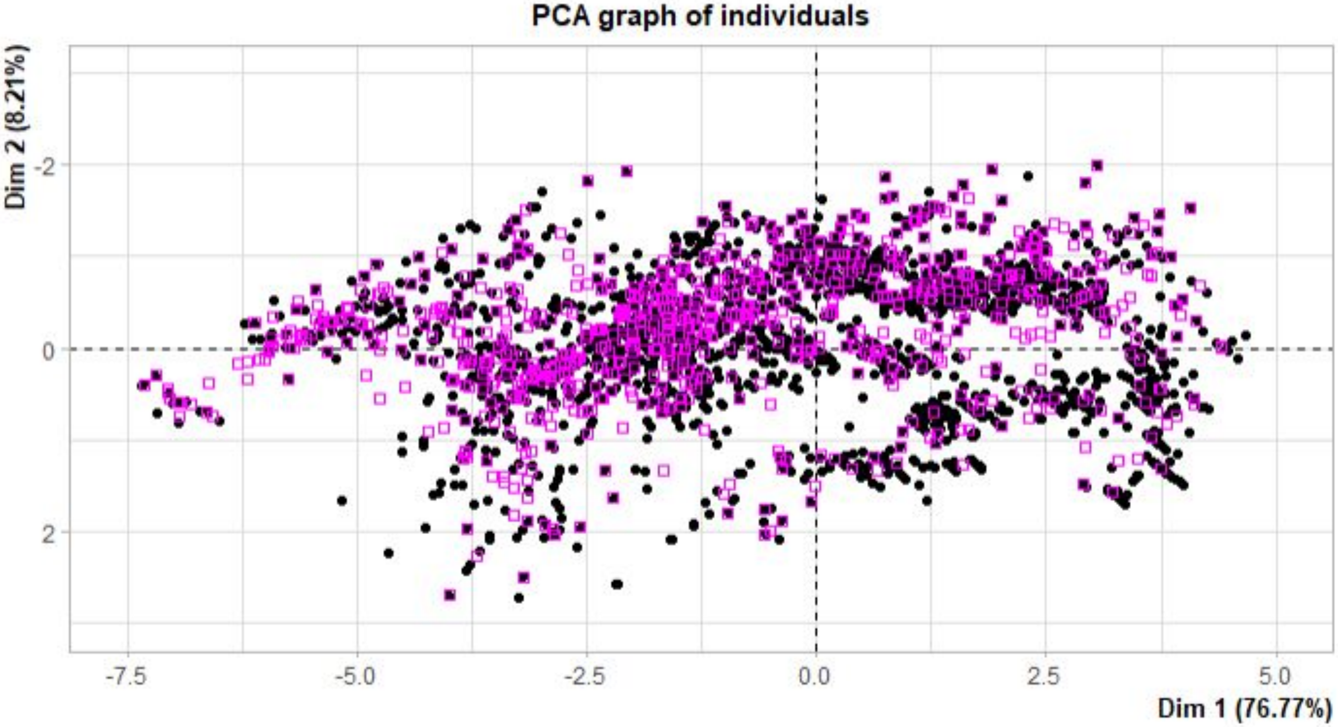
Scree Plot



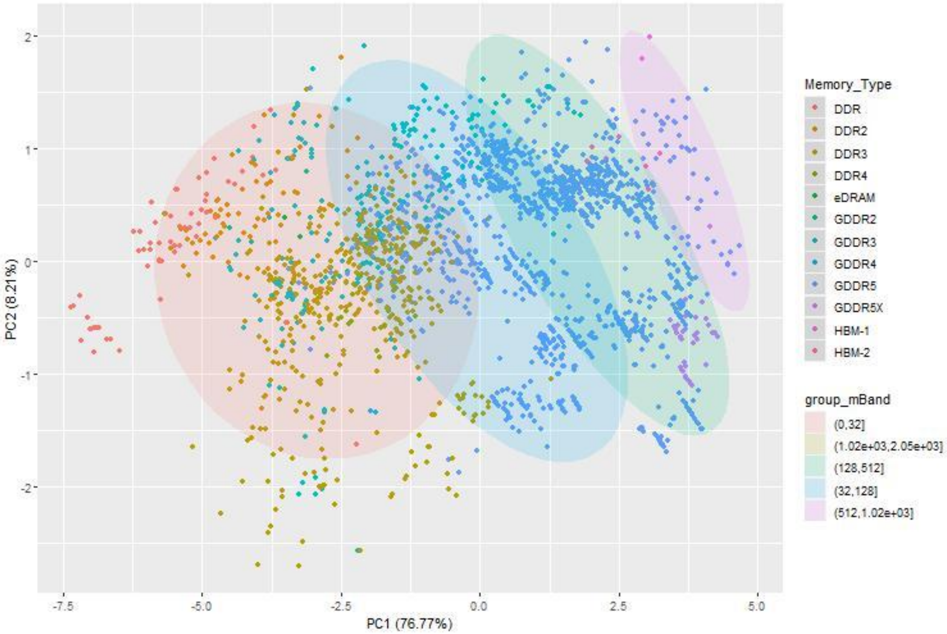
Factorial map visualization

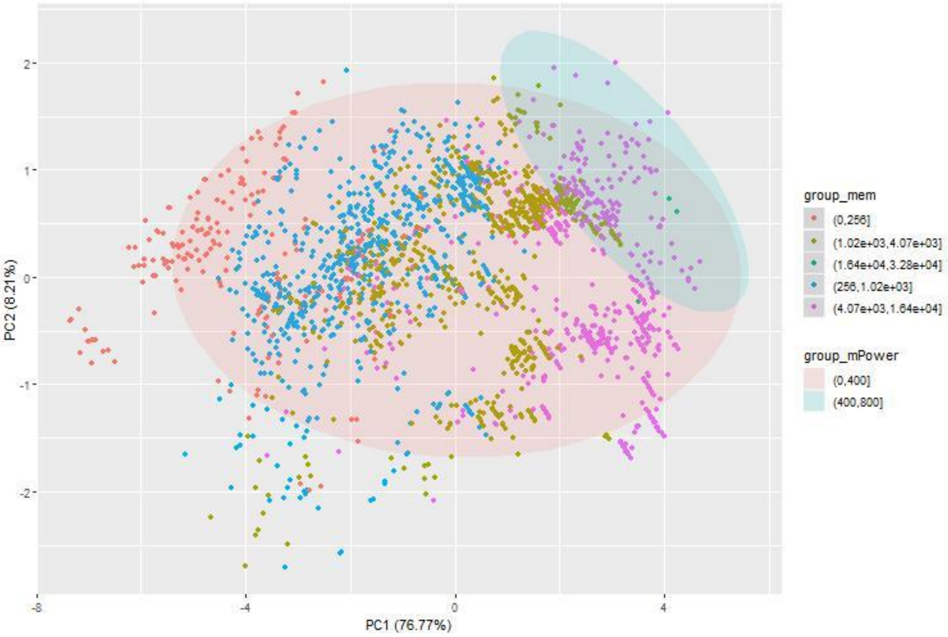


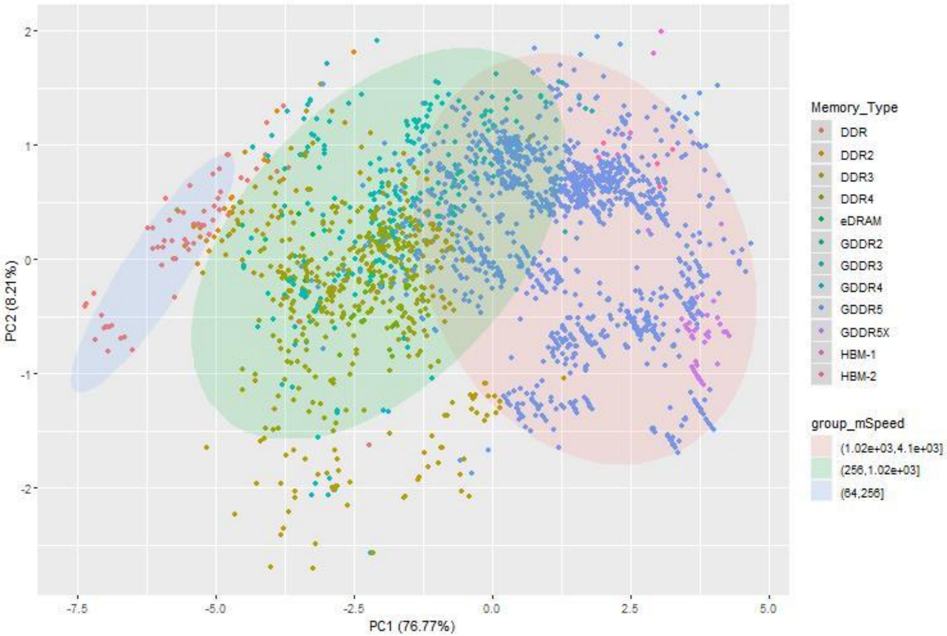


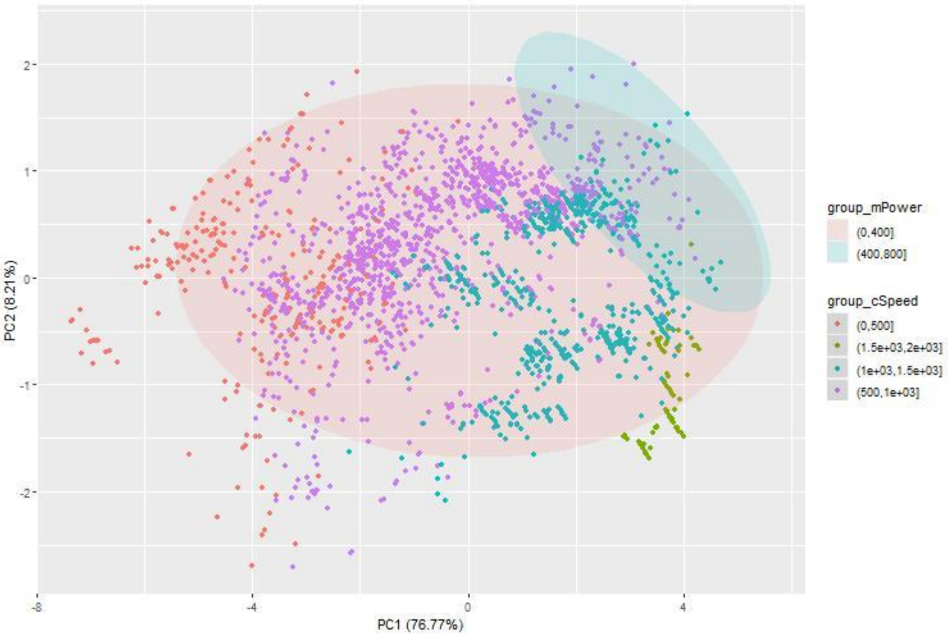

PCA graph of individuals
Dim 2 (8.21%)
-2
0
2
-7.5
-5.0
-2.5
0.0
2.5
5.0
Dim 1 (76.77%)

Conclusions of PCA



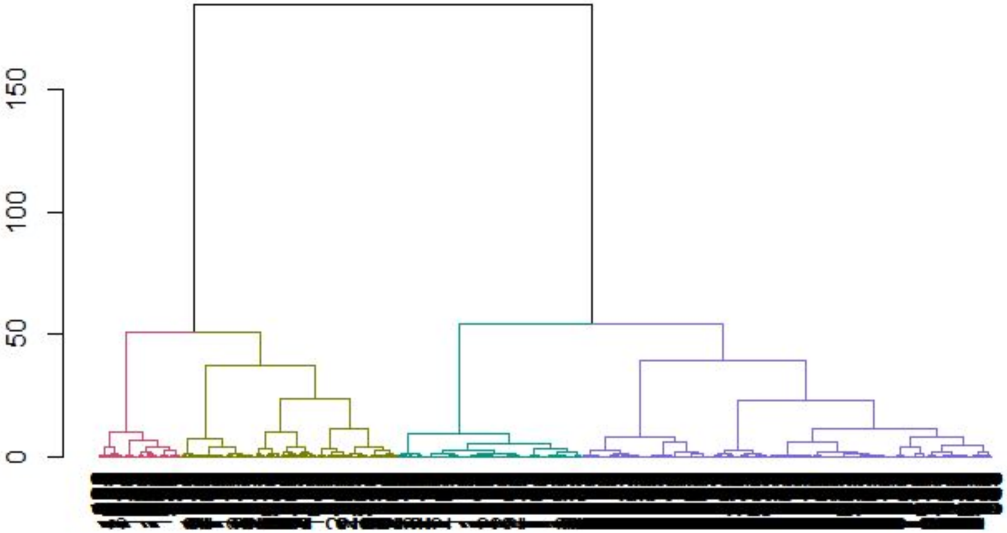



PC2 (8.21%)
2
1
0
-1
-2
-7.5
-5.0
-2.5
0.0
2.5
5.0
PC1 (76.77%)
Memory_Type
DDR
DDR2
DDR3
DDR4
eDRAM
GDDR2
GDDR3
GDDR4
GDDR5
GDDR5X
HBM-1
HBM-2
group_mSpeed
(1.02e+03,4.1e+03]
(256,1.02e+03]
(64,256]

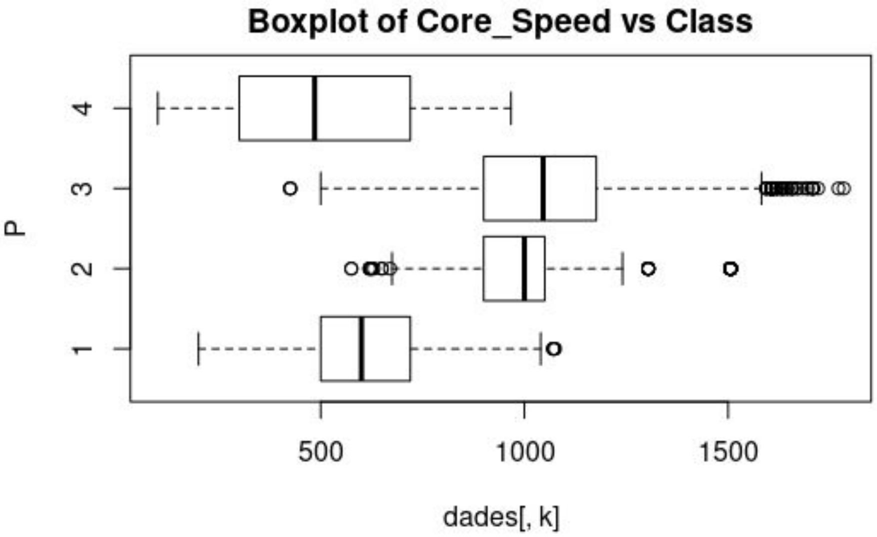

PC2 (8.21%)
2
1
0
-1
-2
-8
-4
0
4
PC1 (76.77%)
group_mPower
(0,400]
(400,800]
group_cSpeed
(0,500]
(1.5e+03,2e+03]
(1e+03,1.5e+03]
(500,1e+03]

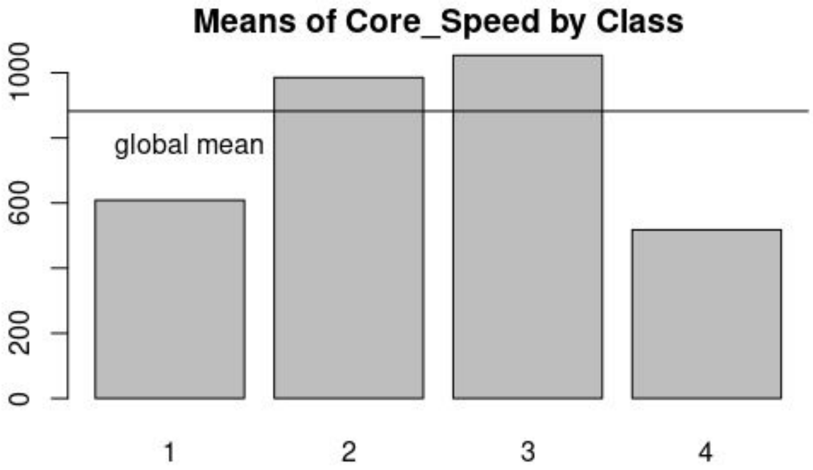
Clustering process

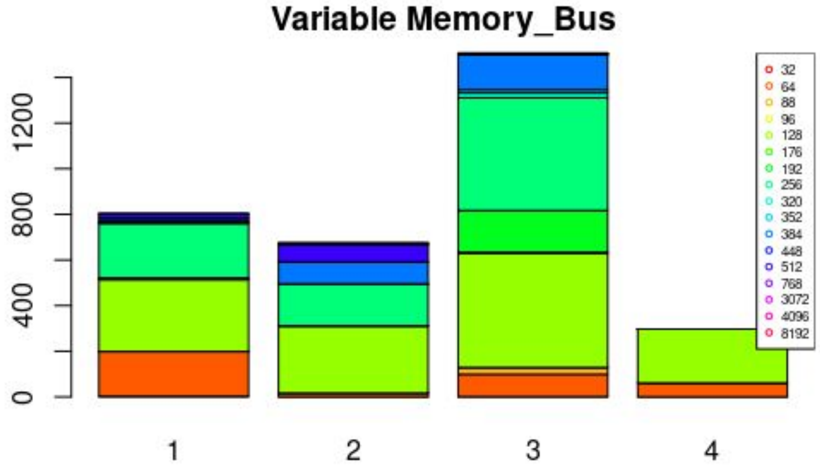
- Almost all data used
- Gower dissimilarity coefficient
- Ward.D aggregation criteria
- 4 clusters
  - Class 1: 798
  - Class 2: 675
  - Class 3: 1507
  - Class 4: 301



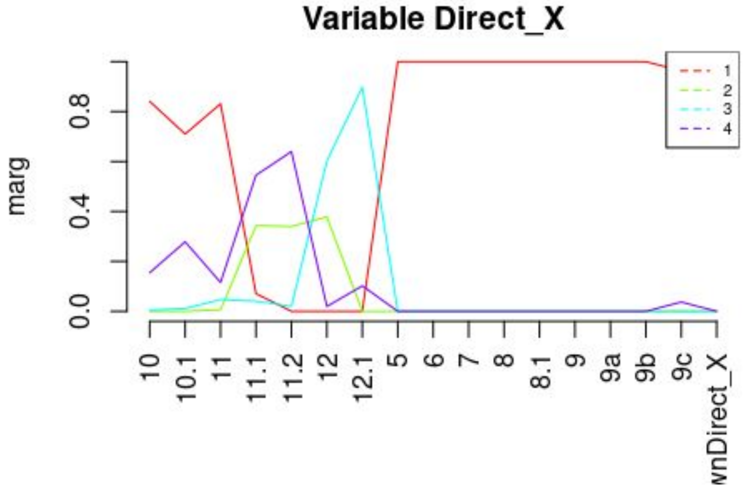
Tools of class interpretation used



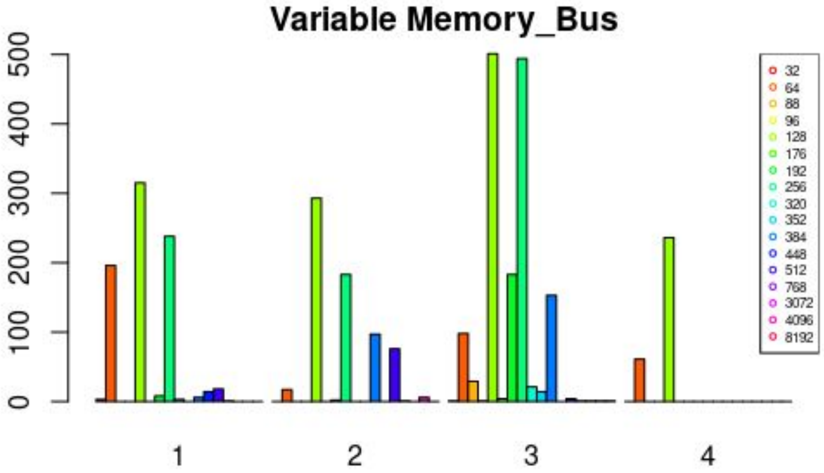




Variable Memory_Bus

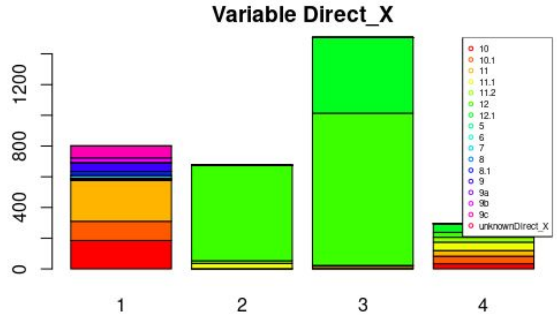


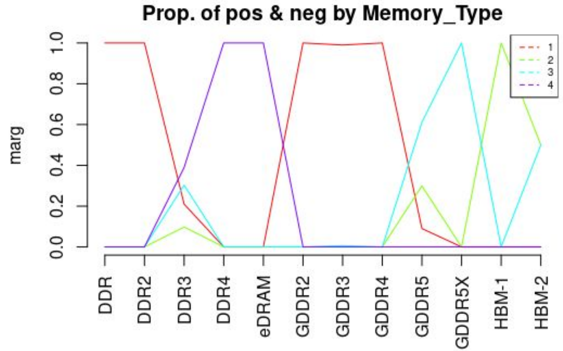
Variable Direct_X

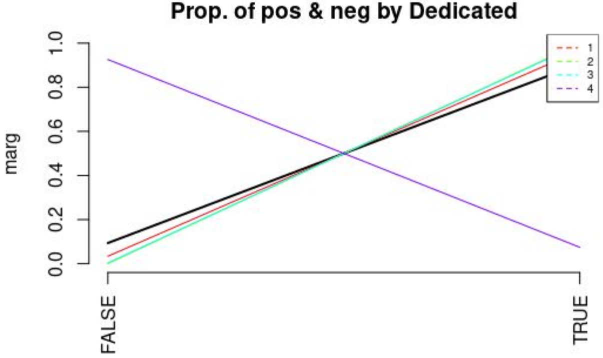


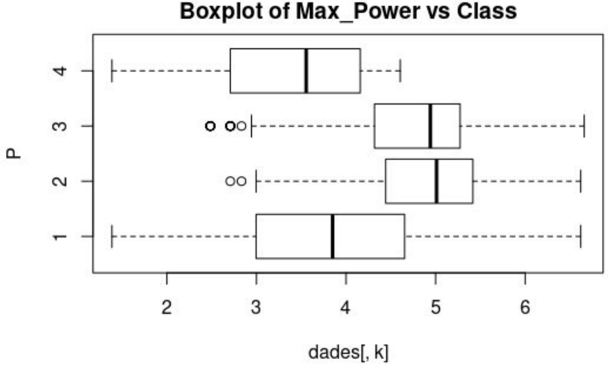
Variable Memory_Bus

Profiling graphs

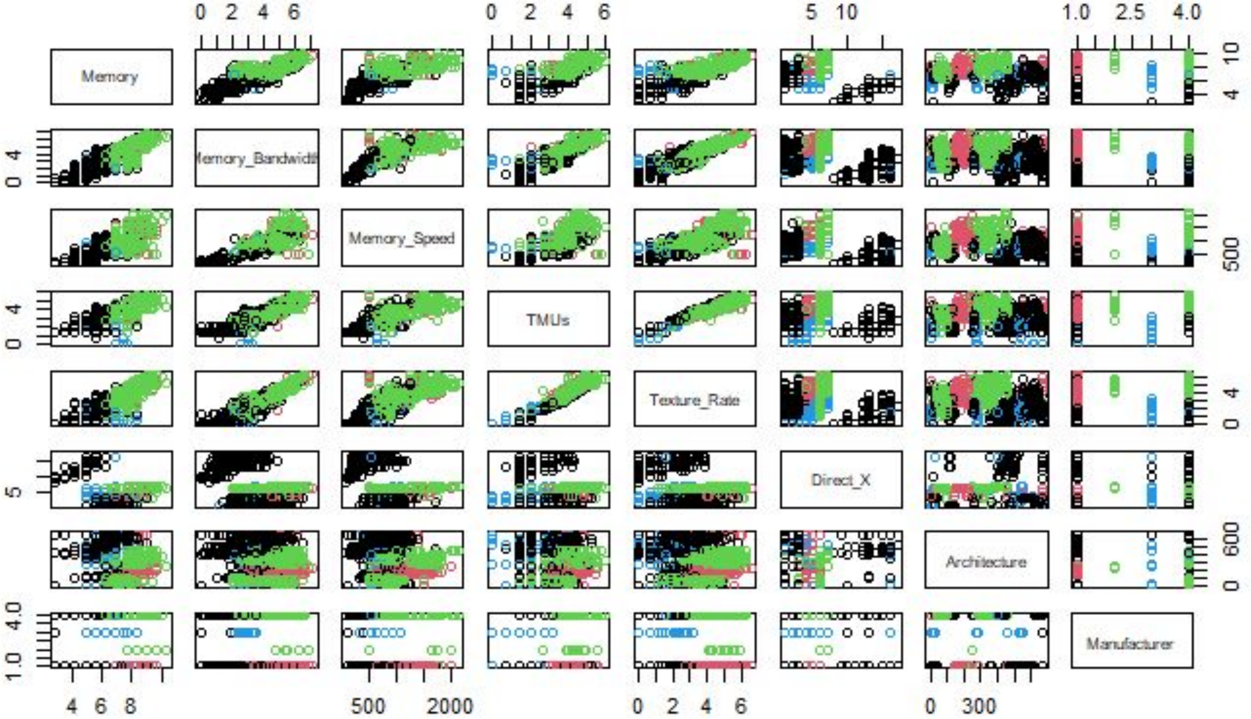




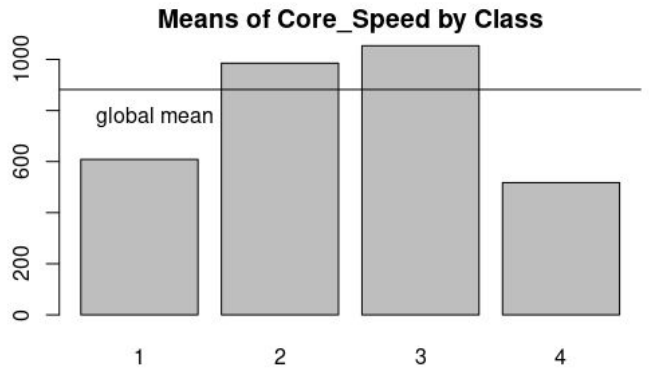


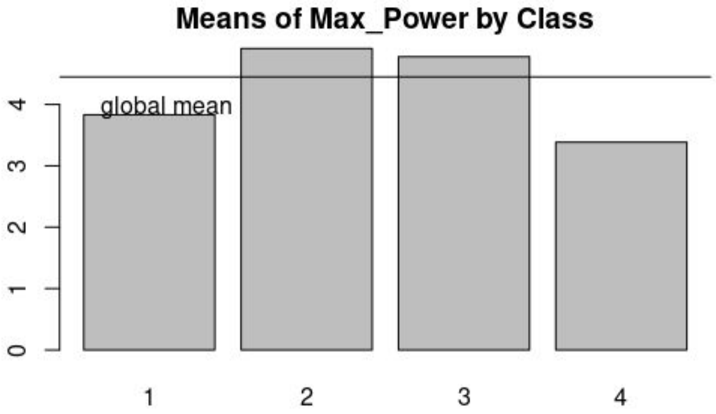


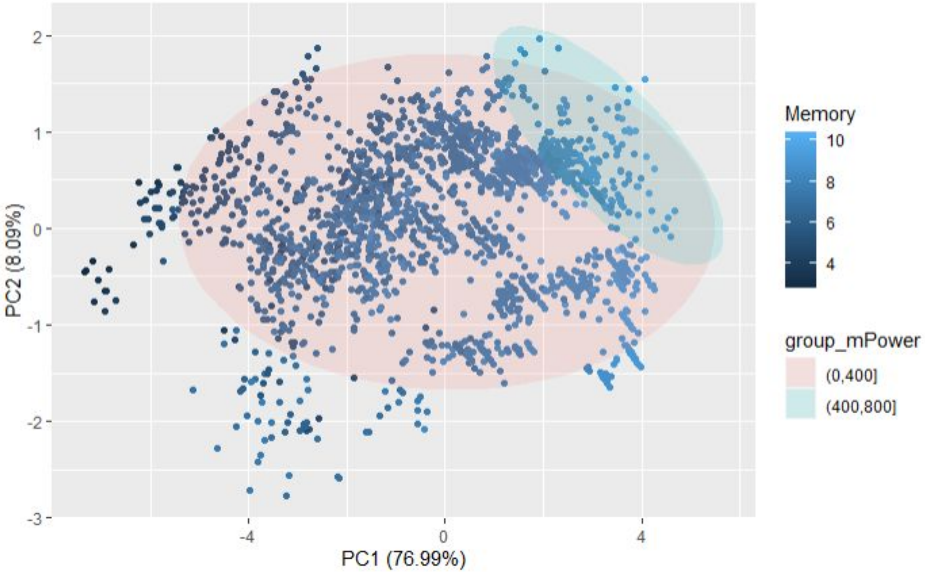
Final class profiling



PCA and Hierarchical Clustering


Means of Core_Speed by Class


Means of Max_Power by Class



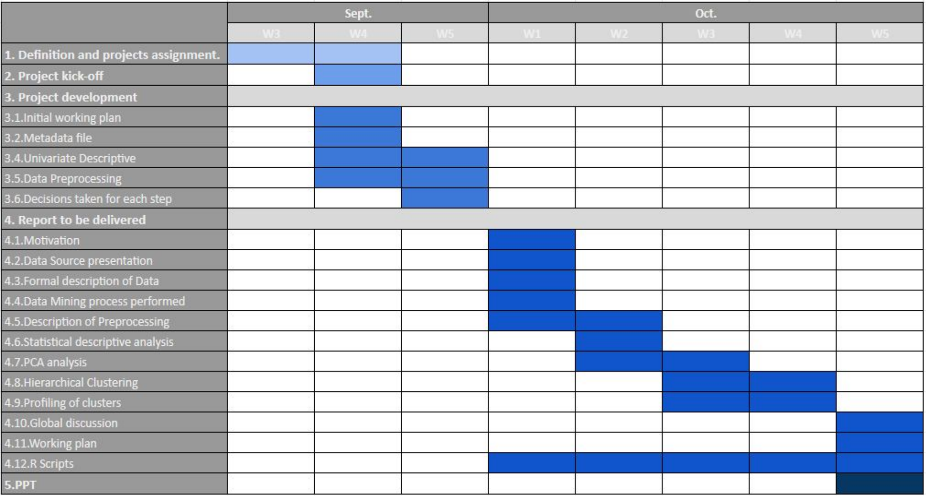
Conclusions

- Relationship between Memory Bus and Core Speed

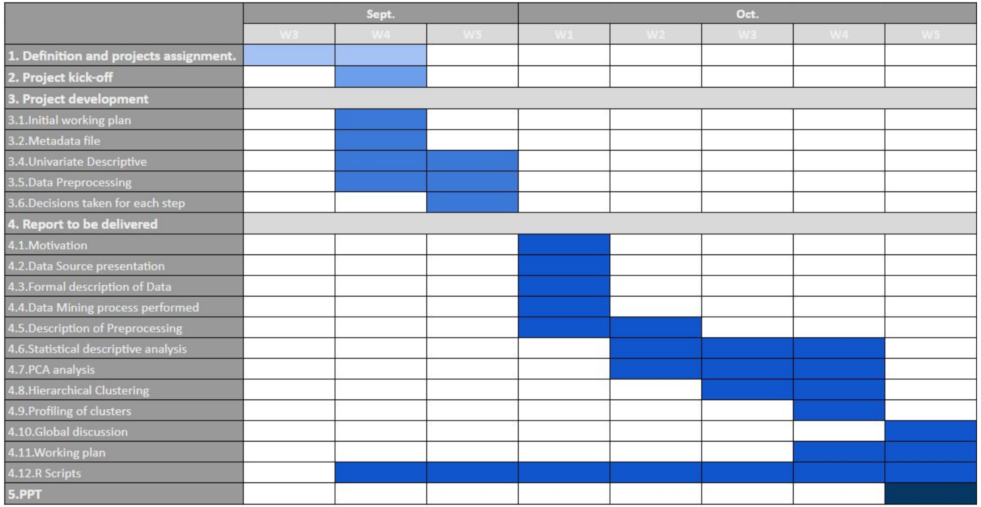
- Group of GPUs with
  - high power consumption
  - shared features

- Identification of clusters of GPUs grouped by performance

Original and final scheduling

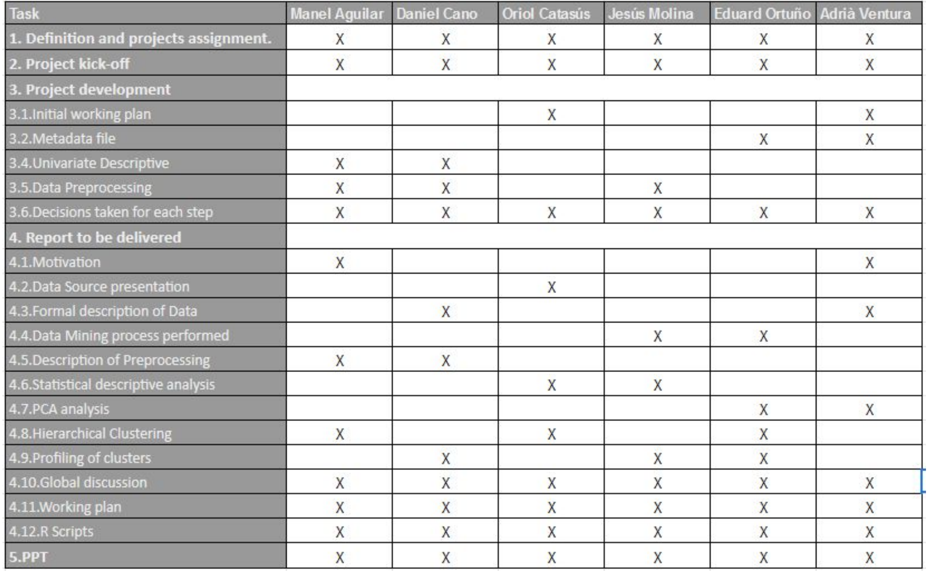


	Sept.			Oct.				
	W3	W4	W5	W1	W2	W3	W4	W5
1. Definition and projects assignment.	■	■						
2. Project kick-off		■						
3. Project development								
3.1.Initial working plan		■						
3.2.Metadata file		■						
3.4.Univariate Descriptive		■	■					
3.5.Data Preprocessing		■	■					
3.6.Decisions taken for each step			■					
4. Report to be delivered								
4.1.Motivation				■				
4.2.Data Source presentation				■				
4.3.Formal description of Data				■				
4.4.Data Mining process performed				■				
4.5.Description of Preprocessing				■	■			
4.6.Statistical descriptive analysis					■			
4.7.PCA analysis					■	■		
4.8.Hierarchical Clustering						■	■	
4.9.Profiling of clusters						■	■	
4.10.Global discussion								■
4.11.Working plan								■
4.12.R Scripts				■	■	■	■	■
5.PPT								■

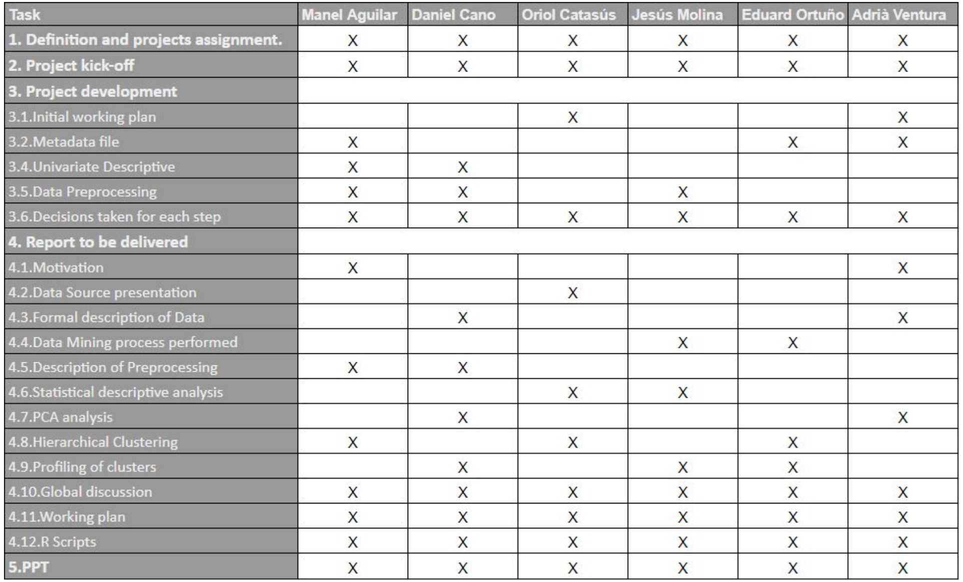


	Sept.			Oct.				
	W3	W4	W5	W1	W2	W3	W4	W5
1. Definition and projects assignment.	■	■						
2. Project kick-off		■						
3. Project development								
3.1.Initial working plan		■						
3.2.Metadata file		■						
3.4.Univariate Descriptive		■	■					
3.5.Data Preprocessing		■	■					
3.6.Decisions taken for each step			■					
4. Report to be delivered								
4.1.Motivation				■				
4.2.Data Source presentation				■				
4.3.Formal description of Data				■				
4.4.Data Mining process performed				■				
4.5.Description of Preprocessing				■	■			
4.6.Statistical descriptive analysis					■	■	■	
4.7.PCA analysis					■	■	■	
4.8.Hierarchical Clustering						■	■	
4.9.Profiling of clusters							■	
4.10.Global discussion								■
4.11.Working plan							■	■
4.12.R Scripts		■	■	■	■	■	■	■
5.PPT								■

Original and final scheduling



Task	Manel Aguilar	Daniel Cano	Oriol Catasús	Jesús Molina	Eduard Ortuño	Adrià Ventura
1. Definition and projects assignment.	X	X	X	X	X	X
2. Project kick-off	X	X	X	X	X	X
3. Project development						
3.1.Initial working plan			X			X
3.2.Metadata file					X	X
3.4.Univariate Descriptive	X	X				
3.5.Data Preprocessing	X	X		X		
3.6.Decisions taken for each step	X	X	X	X	X	X
4. Report to be delivered						
4.1.Motivation	X					X
4.2.Data Source presentation			X			
4.3.Formal description of Data		X				X
4.4.Data Mining process performed				X	X	
4.5.Description of Preprocessing	X	X				
4.6.Statistical descriptive analysis			X	X		
4.7.PCA analysis					X	X
4.8.Hierarchical Clustering	X		X		X	
4.9.Profiling of clusters		X		X	X	
4.10.Global discussion	X	X	X	X	X	X
4.11.Working plan	X	X	X	X	X	X
4.12.R Scripts	X	X	X	X	X	X
5.PPT	X	X	X	X	X	X



Task	Manel Aguilar	Daniel Cano	Oriol Catasús	Jesús Molina	Eduard Ortuño	Adrià Ventura
1. Definition and projects assignment.	X	X	X	X	X	X
2. Project kick-off	X	X	X	X	X	X
3. Project development						
3.1.Initial working plan			X			X
3.2.Metadata file	X				X	X
3.4.Univariate Descriptive	X	X				
3.5.Data Preprocessing	X	X		X		
3.6.Decisions taken for each step	X	X	X	X	X	X
4. Report to be delivered						
4.1.Motivation	X					X
4.2.Data Source presentation			X			
4.3.Formal description of Data		X				X
4.4.Data Mining process performed				X	X	
4.5.Description of Preprocessing	X	X				
4.6.Statistical descriptive analysis			X	X		
4.7.PCA analysis		X				X
4.8.Hierarchical Clustering	X		X		X	
4.9.Profiling of clusters		X		X	X	
4.10.Global discussion	X	X	X	X	X	X
4.11.Working plan	X	X	X	X	X	X
4.12.R Scripts	X	X	X	X	X	X
5.PPT	X	X	X	X	X	X