Data Mining project

Q1-Autumn-21-22

COVID IMPACT

Group 6

28-10-2021

Felipe Castro, Théo Fuhrmann, Xavier Gordillo,
Javier Rivera, Armando Rodríguez, Hasnain Shafqat

Outline

1. Dataset
2. Data mining schema
3. Analysis
    a. One numerical and qualitative
    b. Synthesize univariate
    c. Issues found
4. Preprocessing
5. PCA
    a. Specifications
    b. First factorial plane
    c. Conclusions
6. Clustering
    a. Process
    b. Class interpolation
    c. Eventual Profiling graphs
    d. Final class profiling
7. Conclusions
8. Scheduling

Dataset

The dataset contains information about hospital utilization in all hospitals in the United States registered with centers for Medicare and Medical Services. Specifically talking about the impact that Covid-19 pandemic made.

A	B	C	D	E	F	G	H	I
state	hospital_name	city	zip	hospital_subtype	fips_code	metro	t_b_7da	a_adult_b_7da
AK	Elmendorf	Elmendorf AFB	99506	Short Term	7450	FALSO	101.0	100.7
NV	Elite Medical Ce	Las Vegas	89109	Short Term	29940	FALSO	37.0	37.0
PR	HealthproMed	Vieques	775	Short Term	72147	FALSO	10.0	12.0
PR	Hospital San An	Mayagüez	680	Short Term	72097	FALSO	48.3	35.0
LA	Crescent City Su	Metairie	70118	Short Term	22051	FALSO	12.1	12.1
LA	Alexandria Emer	Alexandria	71303	Short Term	51510	FALSO	14.0	12.0
PR	Hospital Industri	San Juan	935	Short Term	72127	FALSO	50.6	49.1
FL	Reception And M	LAKE BUTLER	32054	Short Term	37650	FALSO	120.0	91.7

Updated
October 18, 2021 6:13 PM CEST

Data Last Updated
October 18, 2021 6:13 PM CEST

Metadata Last Updated
October 18, 2021 6:10 PM CEST

Date Created
May 5, 2021 8:30 PM CEST

HealthData.gov

Data Provided by
U.S. Department of Health & Human Services

Dataset Owner
HHS Office of the Chief Data Officer

What's in this Dataset?

Rows	Columns	Each row is a
315K	105	Aggregated Facility Report for a Given Collection Week

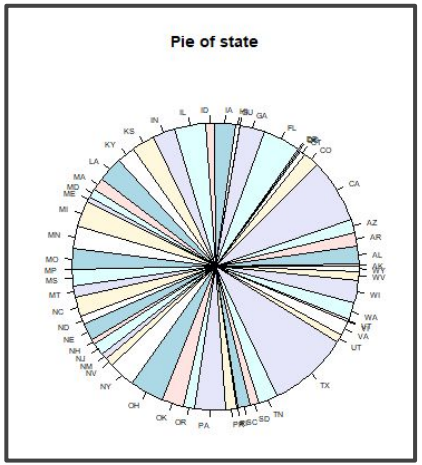
Dataset url: [COVID-19 Reported Patient Impact and Hospital Capacity by Facility]

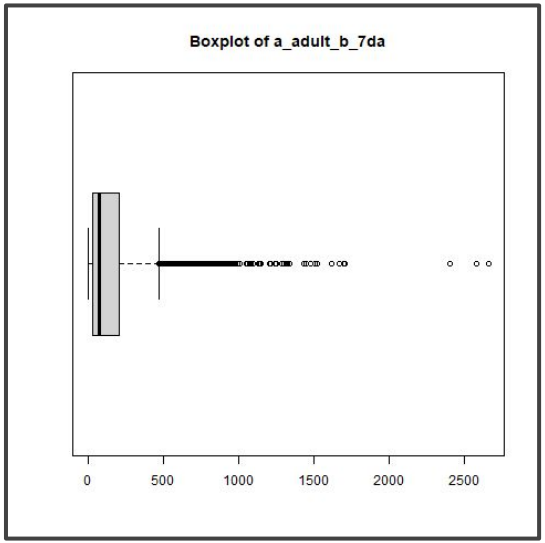
Data mining schema

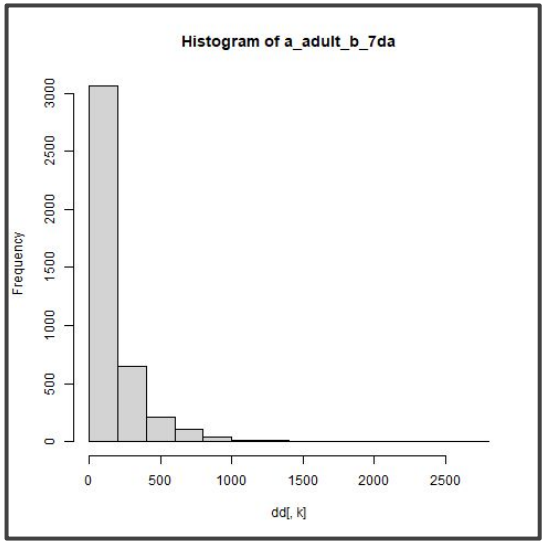
Analysis: One numerical and qualitative

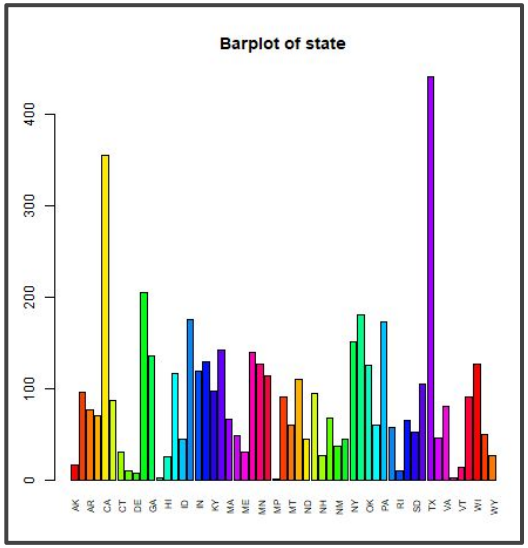
Numerical a_adult_b_7da

Qualitative State





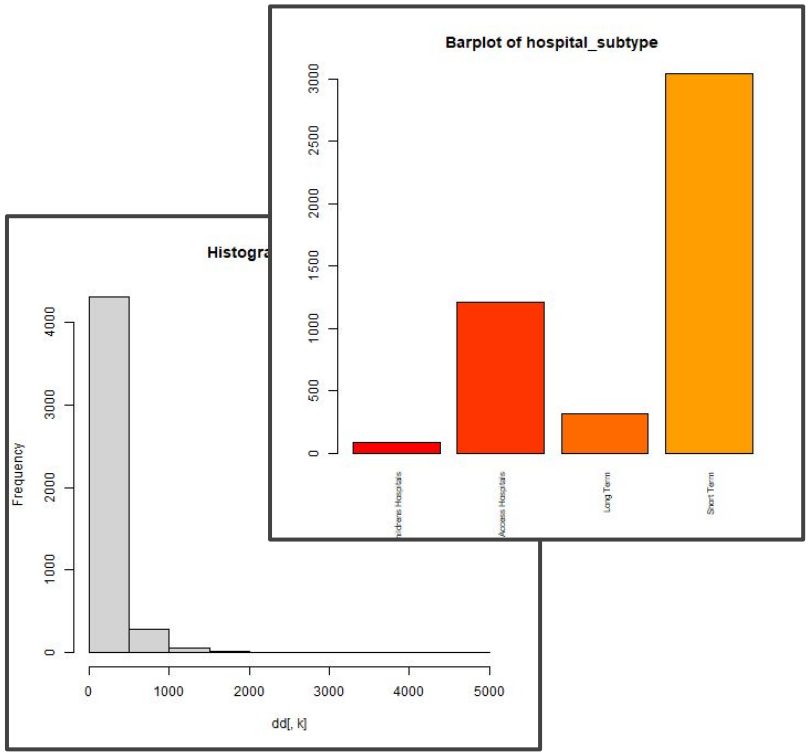




Analysis: Synthesize univariate

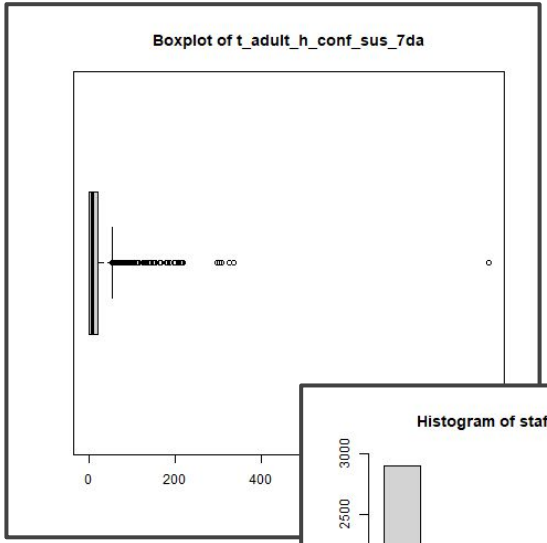
- Categorical variables
  - What is expected
  - In general, distributed quite equally

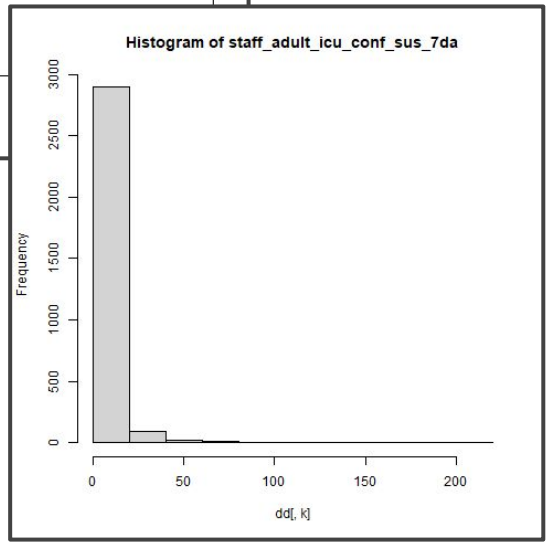
- Numerical variables
  - All with long tails
  - Complicated to visualize due to outliers
  - “Center of mass” near zero
  - Not gaussian distributions



Analysis: Issues

- Really long tails
- Possible outliers
- Not gaussian distributions
- Many values being 0 in every variable
- Many missings that should be correctly imputed





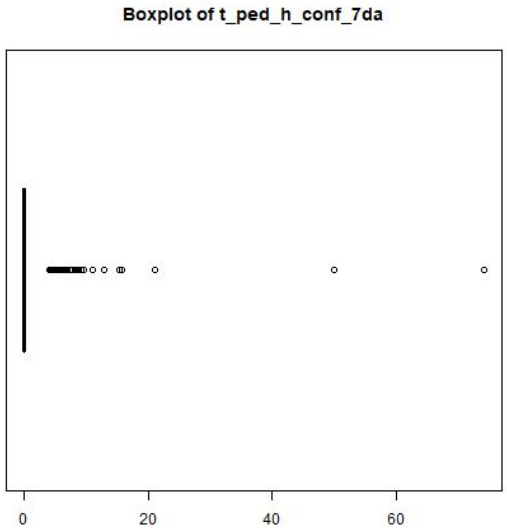
Preprocessing steps

1. First sight

- Records
  - 315.000 records
  - More than one record per hospital
  - We chose one week
- Columns:
  - Variable with no relevant data
  - NA coded as numbers
  - Categorical values coded with strange values
  - Numerical values treated as dates
  - Variable names were too long

Preprocessing steps

1. First sight
2. Outliers



Preprocessing steps

1. First sight
2. Outliers
3. Refining the data

- We searched for irrelevant data
  - Variables without many data
  - No relevant instances
  - Almost all the records with zeros
- Keep variables and records with information that can be studied

Preprocessing steps

1. First sight
2. Outliers
3. Refining the data
4. Missings

- Type of missings found:
  - Most of them related to numerical variables
  - Records with just missings
  - Some missings in important info data
  - Variables with more than 20% missing records
- How we treated them:
  - Numerical variables: Knn method
  - Important data filled by hand
  - Removed variables and records with a high percentage of missings

Preprocessing steps

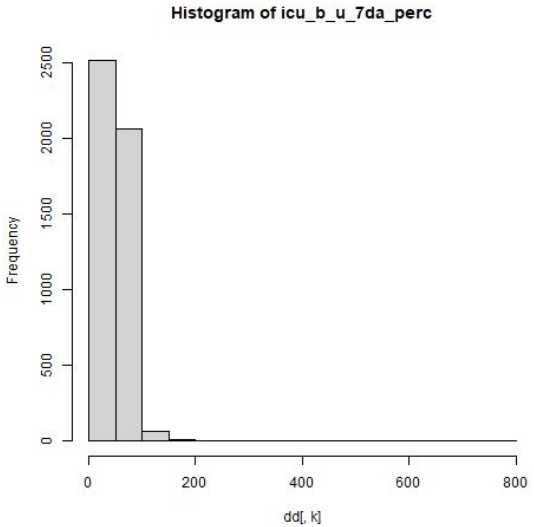
1. First sight
2. Outliers
3. Refining the data
4. Missings
5. New Variables

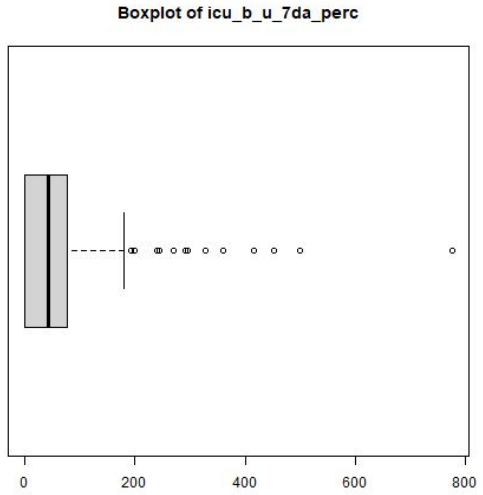
- **inp_b_u_7da_perc:** average percentage of beds used by inpatients in 7 days
- **icu_b_u_7da_perc:** average percentage of beds used by inpatients in UCI in 7 days
- **t_adult_h_conf_7da_perc:** average percentage of inpatients that have Covid-19 confirmed by laboratory

Analysis showed same behaviour in boxplots and histograms

Preprocessing: New Variables

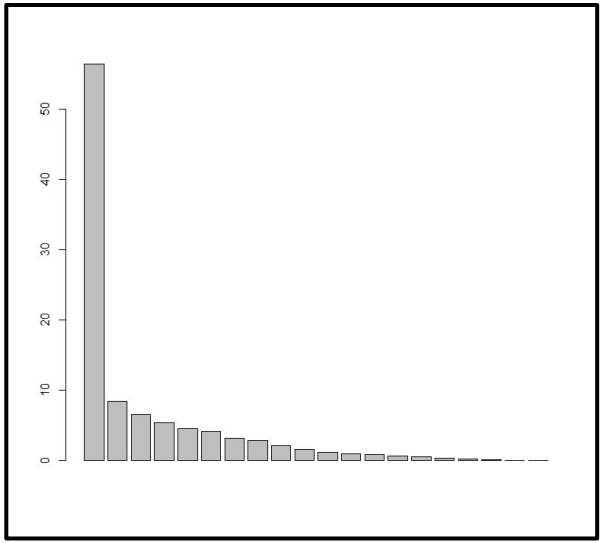
ICU beds used 7 days percentage

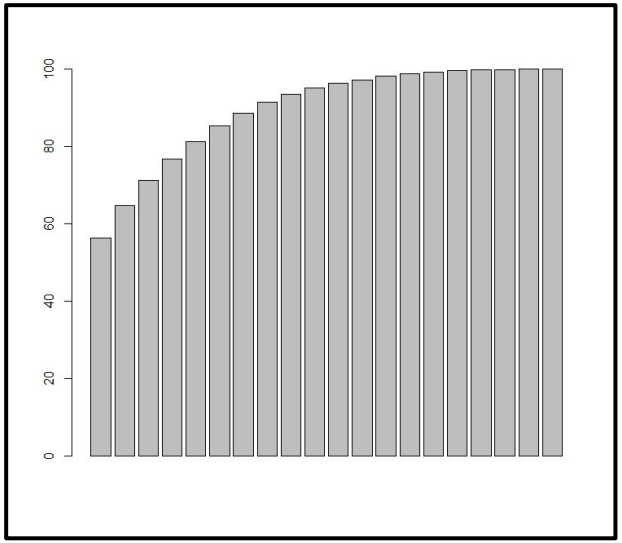




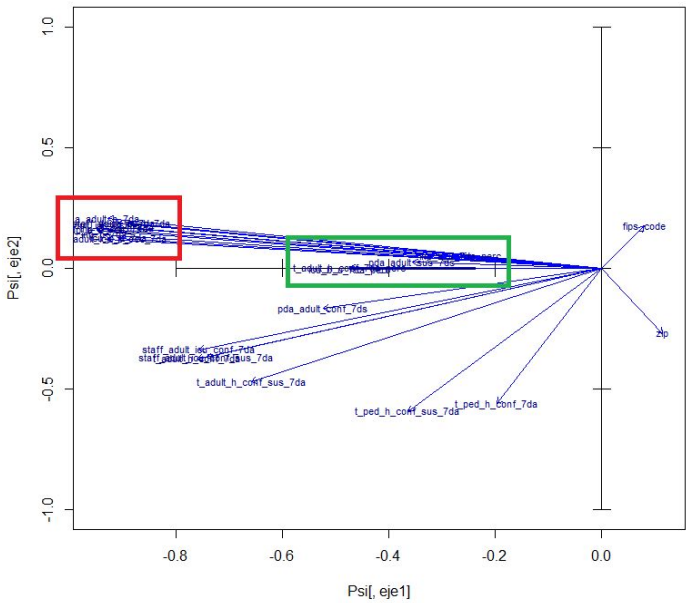
PCA: Specifications

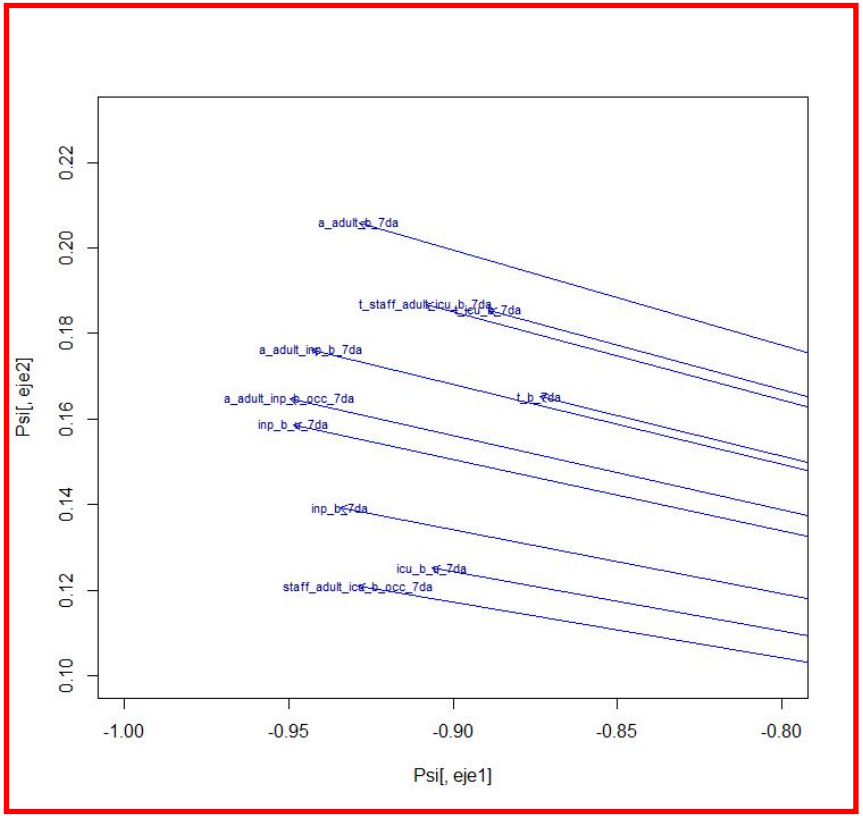
- First axis describes almost 60% of the data variance
- Selected 5 axis achieving more than 80% of the variance



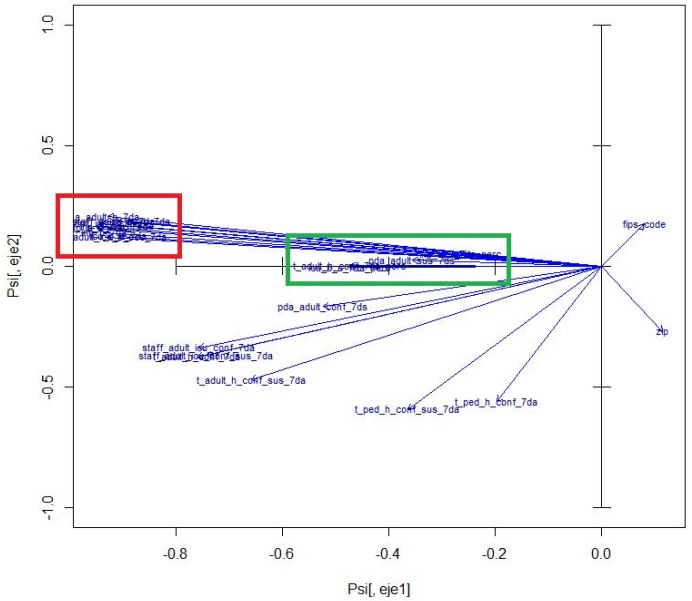


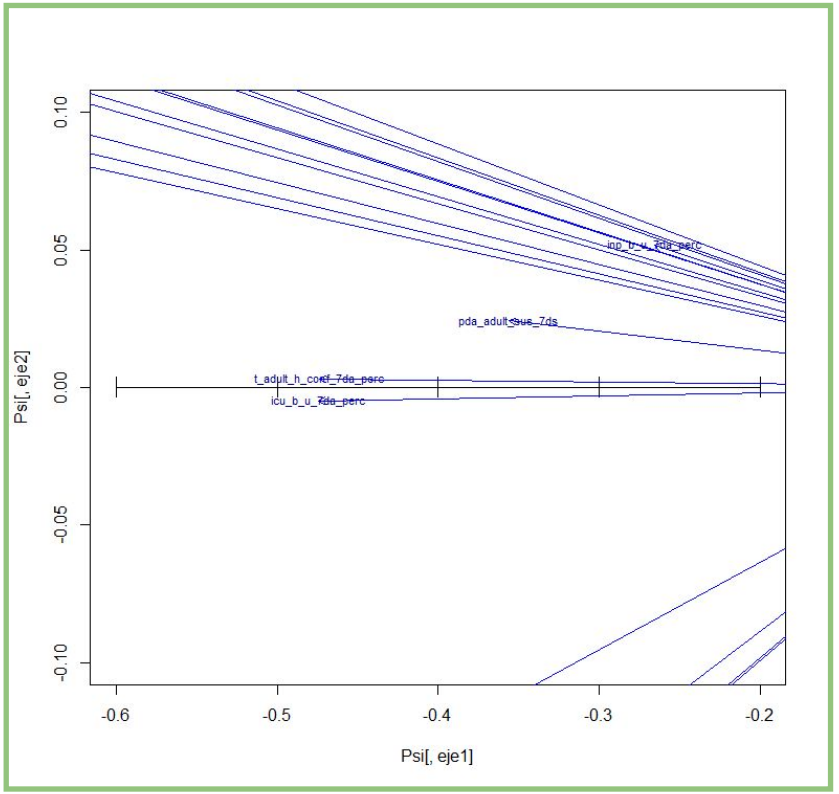
PCA: First factorial plane



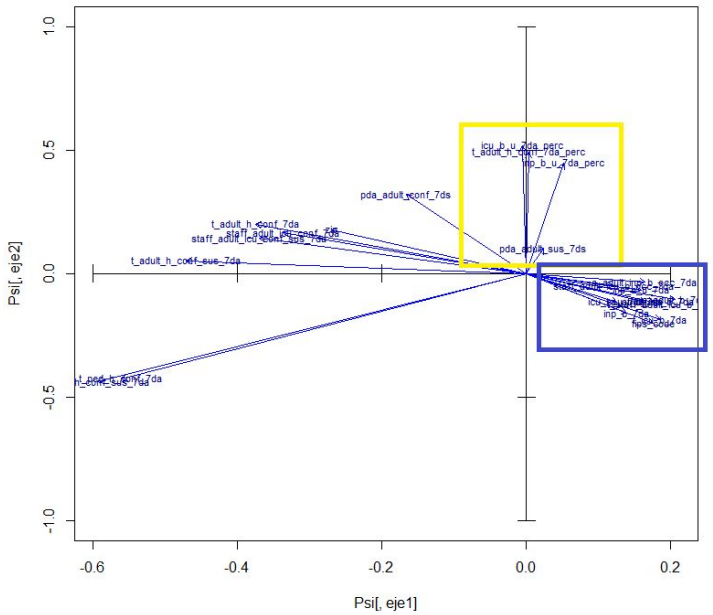


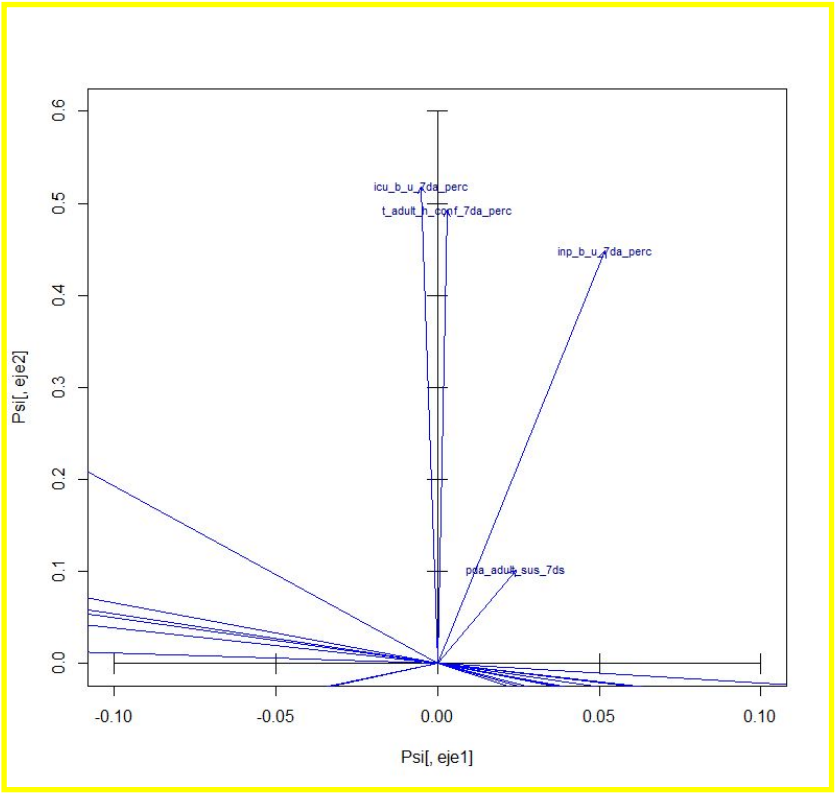
PCA: First factorial plane



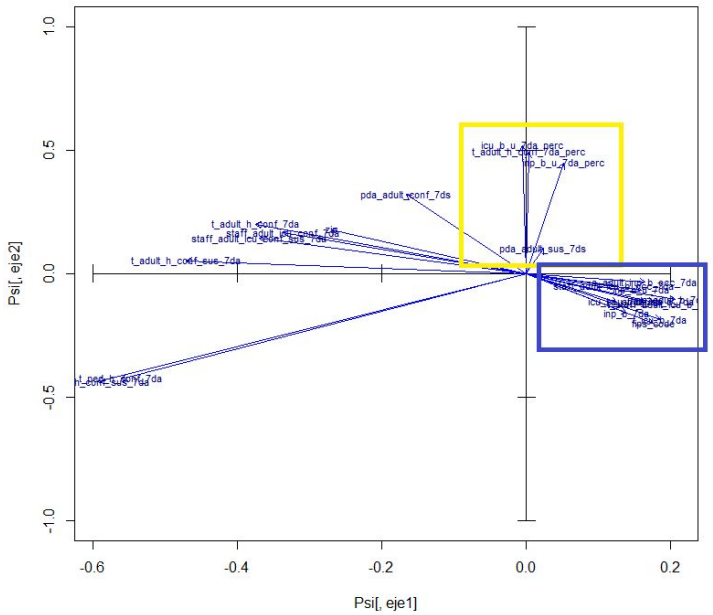


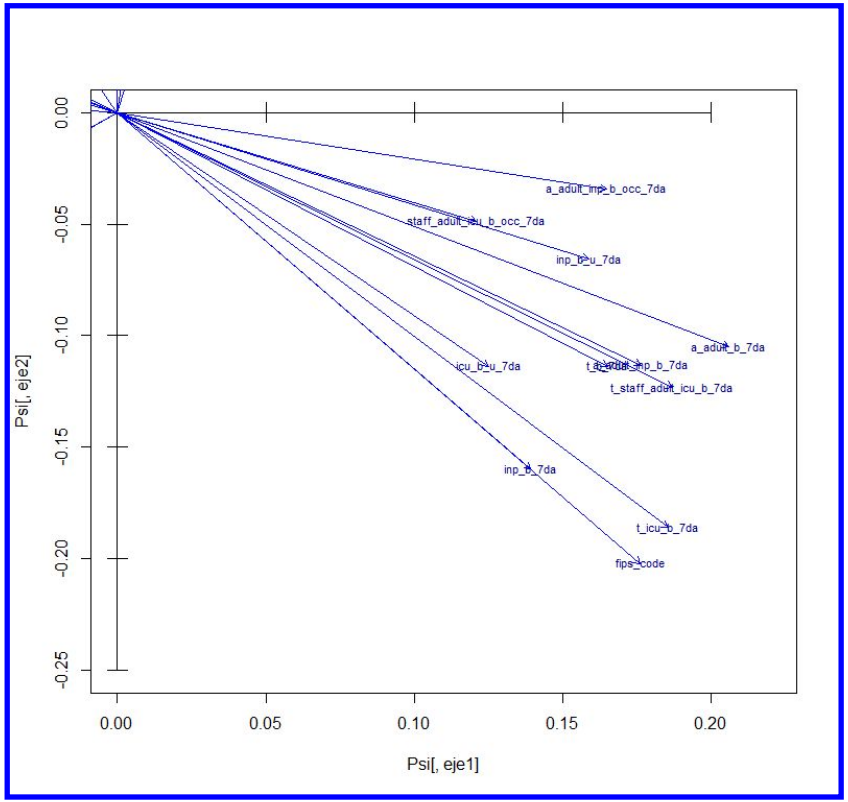
PCA: First factorial plane





PCA: First factorial plane





PCA: Conclusions

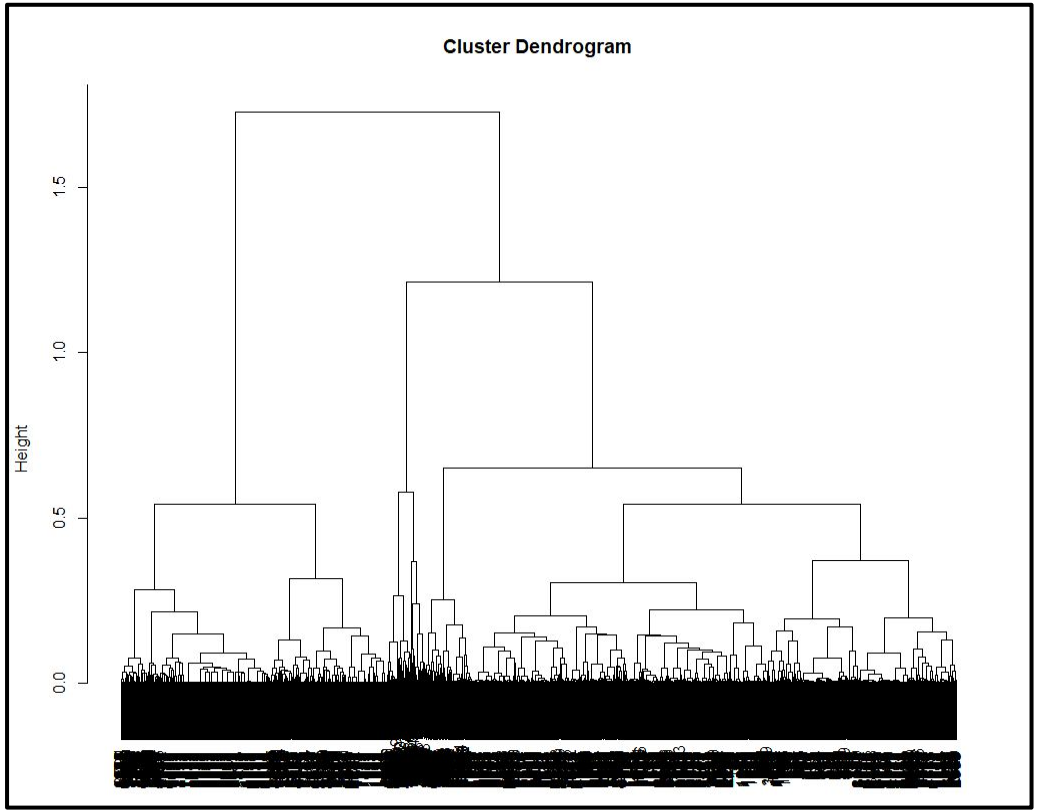
- Strong correlation between most of the variables in the first axis
    - Most of the variables are directly related to the spread of the virus
- The second-third axis analysis reaffirmed the variable correlations
- Useful and rational plot conclusions overall

Clustering: Process

- Every column except of hospital_name
- Gower dissimilarity to the square metric
- Ward.D2 aggregation criteria
- Number of clusters: 3 or 7

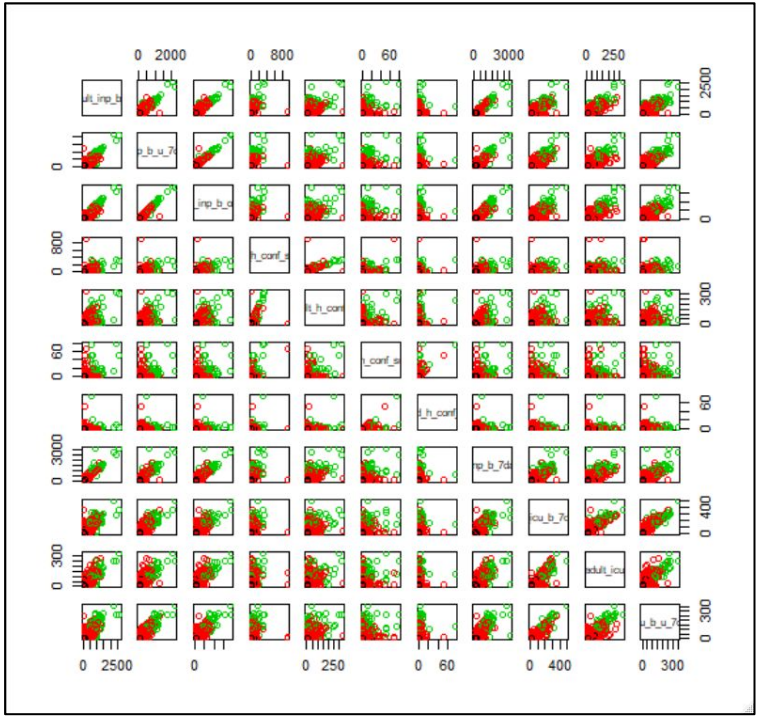
3 classes	1	2	3
Size	1580	2993	84

7 classes	1	2	3	4	5	6	7
Size	847	1639	733	263	84	949	142

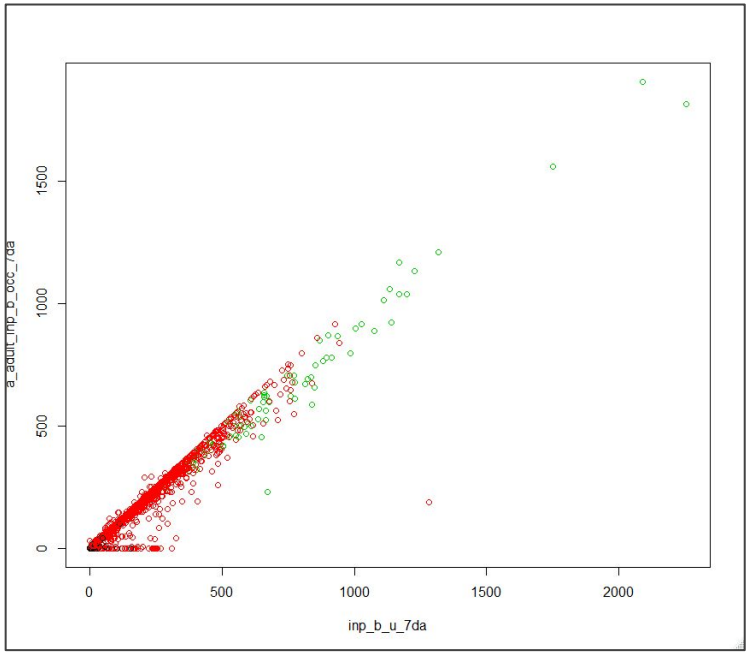


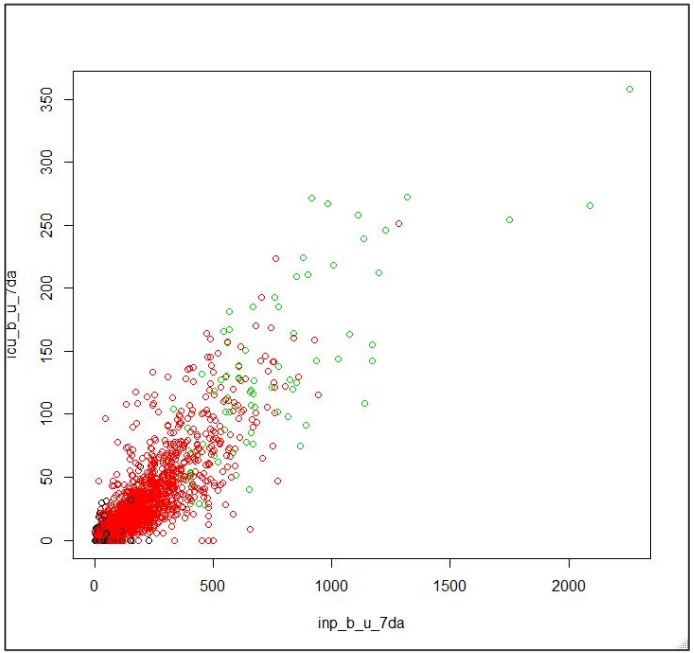
Clustering: Class interpretation

- Conceptual interpretation of the clases
- Graphical Methods
- Pairs plot between variables
- Colored the hospitals by the class obtained
- Select the most representative plots

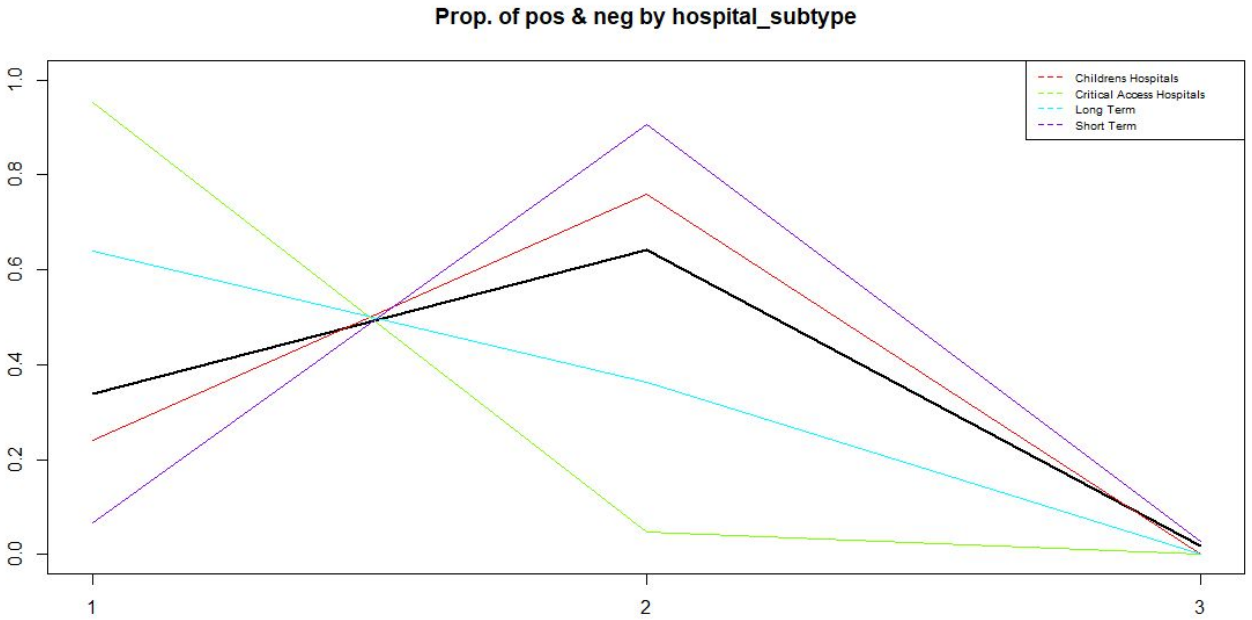


Clustering: Class interpretation

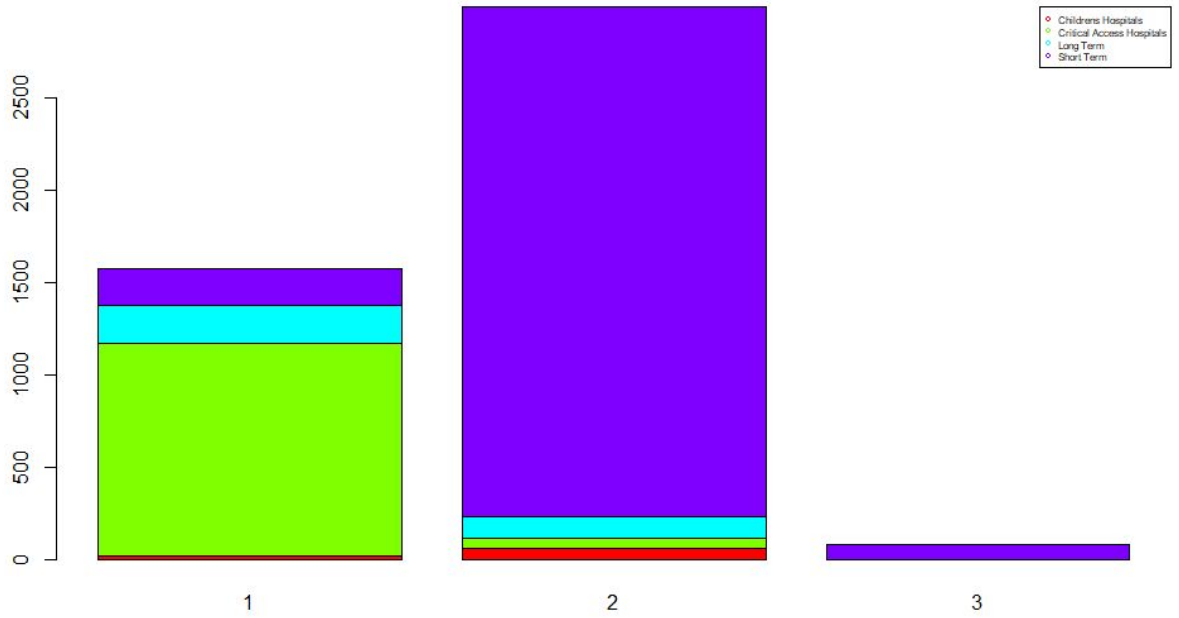




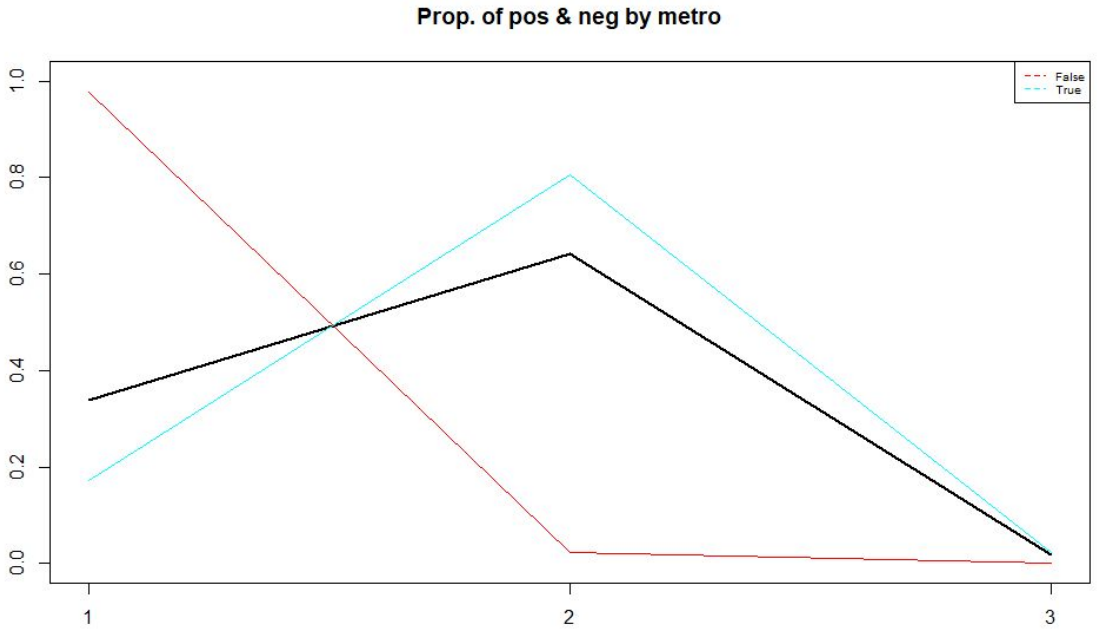
Clustering: Profiling tests



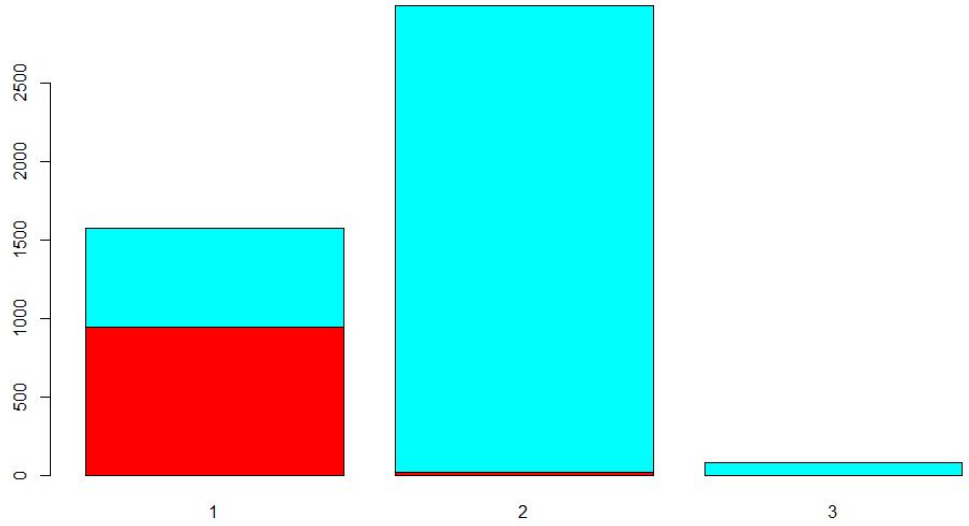
Clustering: Profiling tests



Clustering: Profiling tests



Clustering: Profiling tests



Clustering: Final class profiling

- Our clustering partition splitted the data quite correctly
- We could even define templates for all the three classes:
  - `Class 1 with size 1580 = **Severe** covid incidence`
  - `Class 2 with size 2993 = **Medium** covid incidence`
  - `Class 3 with size 84   = **Low** covid incidence`
- We are happy with that results because we could distribute our data into clear subgroups which could be easier to study in a further project.

PCA & Clustering conclusions

PCA

- Principal axis with most of the variance of the data
- Many correlated variables “coincide” with that axis
- We already know that in some way
- Interesting results

Clustering

- Easy method implementation
- Clear partition at 3 clusters with hierarchical clustering
- Fancy profiling correlations
- Easy descriptors/templates of each cluster

Both are useful at their own objectives

General conclusions

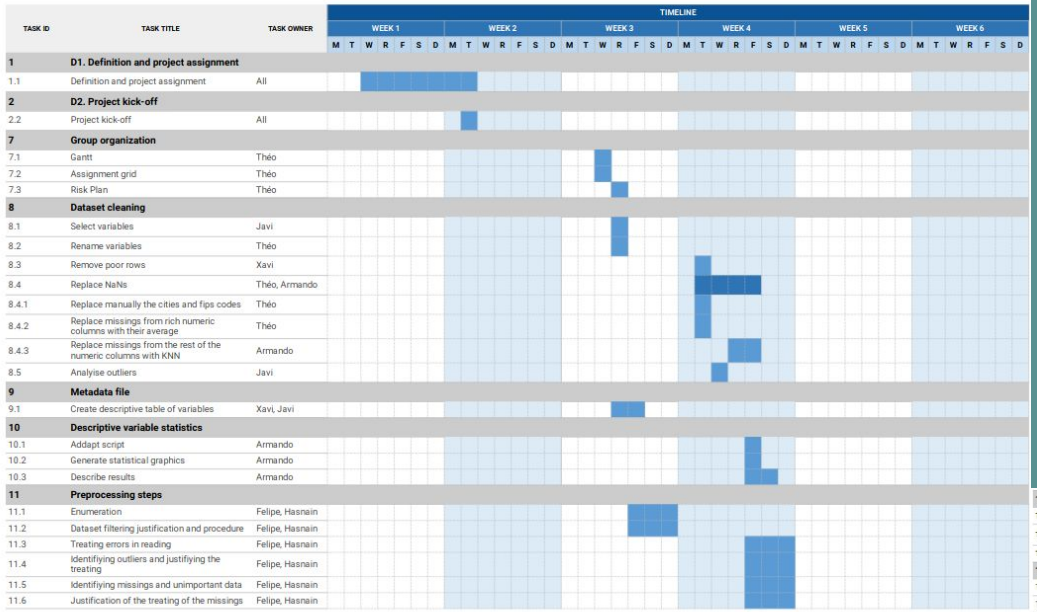
- We learned that the algorithms used in the project weren’t simply a machine that you can feed and pick up the outcome no matter how it looks.
- The algorithms used and the obtained plots were able to give us greater insight of the data patterns in a way that made us understand even better the COVID-19 situation.
- All of the information provided by the different project phases came in harmony towards the end.
- The project gave us more experience regarding data mining and team work.

Scheduling

- Respected the assigned tasks for each member
- Most of the tasks have been completed in their respective scheduled frames.
- More time consuming than expected tasks were managed by including more members in their due
  - We followed the the risk plan
- No unexpected risks appeared in the project
  - Completely stable development.

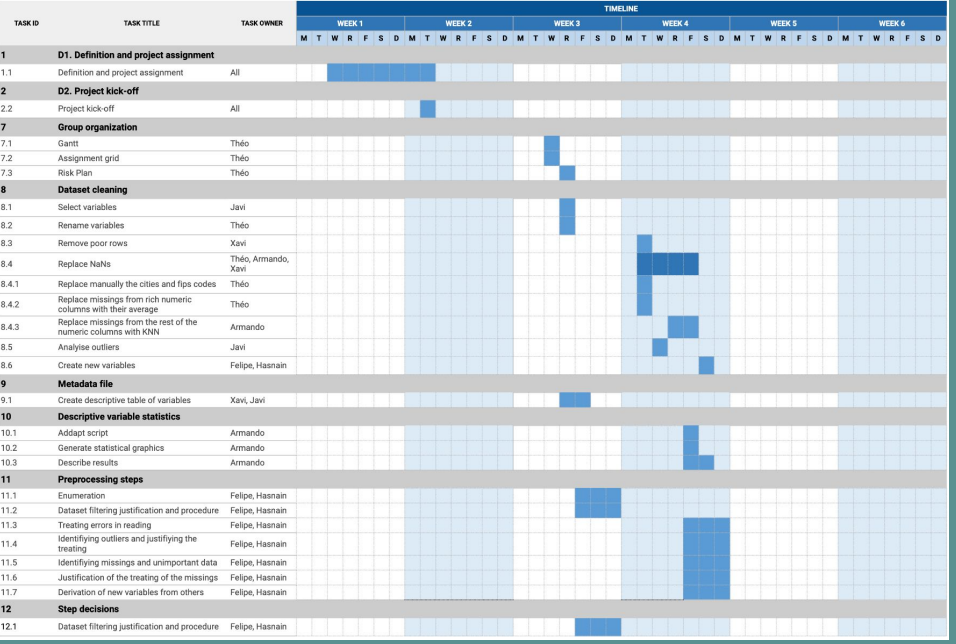
Original scheduling

TASK ID	TASK TITLE	TASK OWNER
1	**D1. Definition and project assignment**	
1.1	Definition and project assignment	All
2	**D2. Project kick-off**	
2.2	Project kick-off	All
7	**Group organization**	
7.1	Gantt	Théo
7.2	Assignment grid	Théo
7.3	Risk Plan	Théo
8	**Dataset cleaning**	
8.1	Select variables	Javi
8.2	Rename variables	Théo
8.3	Remove poor rows	Xavi
8.4	Replace NaNs	Théo, Armando
8.4.1	Replace manually the cities and fips codes	Théo
8.4.2	Replace missings from rich numeric columns with their average	Théo
8.4.3	Replace missings from the rest of the numeric columns with KNN	Armando
8.5	Analyise outliers	Javi
9	**Metadata file**	
9.1	Create descriptive table of variables	Xavi, Javi
10	**Descriptive variable statistics**	
10.1	Addapt script	Armando
10.2	Generate statistical graphics	Armando
10.3	Describe results	Armando
11	**Preprocessing steps**	
11.1	Enumeration	Felipe, Hasnain
11.2	Dataset filtering justification and procedure	Felipe, Hasnain
11.3	Treating errors in reading	Felipe, Hasnain
11.4	Identifiying outliers and justifiying the treating	Felipe, Hasnain
11.5	Identifiying missings and unimportant data	Felipe, Hasnain
11.6	Justification of the treating of the missings	Felipe, Hasnain



12	**Step decisions**	
12.1	Dataset filtering justification and procedure	Felipe, Hasnain
12.2	Identifiying outliers and justifiying the treating	Felipe, Hasnain
12.3	Justification of the treating of the missings	Felipe, Hasnain
13	**Description of changes**	
13.1	New variables	Felipe, Hasnain
13.2	Modified variables	Felipe, Hasnain
14	**Motivation of the work**	
14.1	General description of the problem	Armando, Javi
15	**Datamining process perfomance**	
15.1	Description	Théo, Xavi
15.2	Workflow	Théo, Xavi
16	**Final scope of the study**	
16.1	Description	Felipe, Hasnain
17	**PCA analysis for numerical variables**	
17.1	Scree plot	Felipe, Xavi, Javi
17.2	Factorial map visualisation	Felipe, Xavi, Javi
18	**Hierarchical clustering on original data**	
18.1	Description of data used	Théo, Armando, Hasnain
18.2	Clustering method used, metrics and aggregation criteria used	Théo, Armando, Hasnain
18.3	Dendrogram	Théo, Armando, Hasnain
18.4	Table with description of cluster size	Théo, Armando, Hasnain
19	**Profiling for clusters**	
19.1	Profiling graphs, CPGs, etc.	Felipe, Théo, Xavi
19.2	Add specific profiling tests to relevant variables	Felipe, Théo, Xavi
19.3	Synthesize the result of the classes' interpretation process into templates	Felipe, Théo, Xavi
20	**Global discussion and conclusions of the whole work**	
20.1	Description	Javi, Armando, Hasnain
20.2	Analysis of coincidences and divergences between ACP, AMC, Clustering	Javi, Armando, Hasnain
21	**PTT preparation**	
21.1	Script	All
21.2	Slides	All

TASK ID	TASK TITLE	TASK OWNER
1	**D1. Definition and project assignment**	
1.1	Definition and project assignment	All
2	**D2. Project kick-off**	
2.2	Project kick-off	All
7	**Group organization**	
7.1	Gantt	Théo
7.2	Assignment grid	Théo
7.3	Risk Plan	Théo
8	**Dataset cleaning**	
8.1	Select variables	Javi
8.2	Rename variables	Théo
8.3	Remove poor rows	Xavi
8.4	Replace NaNs	Théo, Armando, Xavi
8.4.1	Replace manually the cities and fips codes	Théo
8.4.2	Replace missings from rich numeric columns with their average	Théo
8.4.3	Replace missings from the rest of the numeric columns with KNN	Armando
8.5	Analyise outliers	Javi
8.6	Create new variables	Felipe, Hasnain
9	**Metadata file**	
9.1	Create descriptive table of variables	Xavi, Javi
10	**Descriptive variable statistics**	
10.1	Addapt script	Armando
10.2	Generate statistical graphics	Armando
10.3	Describe results	Armando
11	**Preprocessing steps**	
11.1	Enumeration	Felipe, Hasnain
11.2	Dataset filtering justification and procedure	Felipe, Hasnain
11.3	Treating errors in reading	Felipe, Hasnain
11.4	Identifying outliers and justifiying the treating	Felipe, Hasnain
11.5	Identifiying missings and unimportant data	Felipe, Hasnain
11.6	Justification of the treating of the missings	Felipe, Hasnain
11.7	Derivation of new variables from others	Felipe, Hasnain
12	**Step decisions**	
12.1	Dataset filtering justification and procedure	Felipe, Hasnain
12.2	Identifying outliers and justifiying the treating	Felipe, Hasnain
12.3	Justification of the treating of the missings	Felipe, Hasnain
13	**Description of changes**	
13.1	New variables	Felipe, Hasnain
13.2	Modified variables	Felipe, Hasnain
14	**Motivation of the work**	
14.1	General description of the problem	Armando, Javi
15	**Datamining process perfomance**	
15.1	Description	Théo, Xavi
15.2	Workflow	Théo, Xavi
16	**Final scope of the study**	
16.1	Description	Felipe, Hasnain
17	**PCA analysis for numerical variables**	
17.1	Scree plot	Javi
17.2	Factorial map visualisation	Javi
18	**Hierarchical clustering on original data**	
18.1	Description of data used	Armando
18.2	Clustering method used, metrics and aggregation criteria used	Armando
18.3	Dendrogram	Armando
18.4	Table with description of cluster size	Armando
19	**Profiling for clusters**	
19.1	Profiling graphs, CPGs, etc.	Felipe, Hasnain
19.2	Add specific profiling tests to relevant variables	Felipe, Hasnain
19.3	Synthesize the result of the classes' interpretation process into templates	Armando
20	**Global discussion and conclusions of the whole work**	
20.1	Description	Xavi, Théo
20.2	Analysis of coincidences and divergences between ACP, AMC, Clustering	Xavi, Théo
21	**PTT preparation**	
21.1	Script	All
21.2	Slides	All



Final scheduling

Data Mining project

Q1-Autumn-21-22

COVID IMPACT

Group 6

28-10-2021

Felipe Castro, Théo Fuhrmann, Xavier Gordillo,
Javier Rivera, Armando Rodríguez, Hasnain Shafqat