Data Mining (MD)

Bachelor Degree on Informatics Engineering

Final report
COVID IMPACT
Grup 6

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

28/10/2021

Q1-Autumn-2021/2022

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Summary

Work motivation and general problem description

The initial idea of this report is to get people to know about the impact that the crisis of Covid-19 produced in our daily lives. More specifically, the impact in the most affected, obviously, area of this crisis, the hospitals. We selected an interesting dataset about the Covid impact in the hospitals of the United States. In this report we will show you about the idea of extracting useful information from a complex dataset and visualizing and obtaining results and conclusions from different types of variable data and their possible interrelations.

We arrived at some fancy conclusions at the end, so you should follow us on our journey through the interesting world of Data Mining and the impact that covid made in our society.

Data source presentation

The data source of our database is a web page of the government of the United States where they upload all the data related to Covid-19 Impact.

Source:
https://healthdata.gov/Hospital/COVID-19-Reported-Patient-Impact-and-Hospital-Capa/uqq2-txqb

Data description

What is the data about?

Our database is related to the current situation in which we are living, the Covid-19 pandemic. Most precisely it is about the patient impact and the hospital's capacity in the United States. We observed that we have a lot of entries on our database because there are a lot of hospitals in the United States. So basically, if we want to resume what is about our database, we can simply say that it is about the impact of the Covid-19 crisis in the United States hospitals and the type of patients that are dealing with it. The reason for the existence of this database is, obviously, because of Covid-19. Because of the pandemic, the government believed that it was mandatory to have a database where we can analyse the impact of the Covid-19, and how to react to that situation. The dataset started reporting in July of 2020. Finally, most of the data comes from two main sources: HHS TeleTracking and reporting provided directly to HHS Protect by state/territorial health departments. HHS stands for Health and Human Services Department.

Data structure

The original dataset was too big for the purposes of this course so we made a small selection of the data in order to make our work more doable. In the original dataset there were records about almost every week since July of 2020, so we just took the record of the

day 03/09/2021 from every hospital. In total we have 5004 records, and for every record we have 59 variables. 27 of them are categorical, for example one of them is hospital subtype. We only have 2 binary variables, we had to convert one of the numerical ones into a binary variable. And finally we have 30 numeric variables counting for example the average number of ICU beds occupied in the last 7 days. It is possible in the future we combine several variables into more complex ones. We have merged the type, number and percentage of missing values they have with the metadata file that can be found in the annex.

Metadata file

We have included a chart that briefly mentions all relevant information about our dataset's variables. The chart includes information about the variable's name, meaning, range, measuring unit, the encoding of missing variables and more. To view the full chart, see the Annex (Annex_COVIDImpact6).

Guide

We've shortened the variable names for the sake of simplicity, here are the initial meanings:

7da: 7 day average	h: hospitalized
7ds: 7 day sum	inp: inpatient
a: all	occ: occupied
b: beds	ped: pediatric
conf: confirmed	sus: suspected

Including and excluding criteria

The first step of preprocessing we took was to define which variables and records we will keep, and which we will delete. So, first of all, we decided that we will reduce the number of records because we had 295k records. We observed that we had different weeks for every hospital, hence we chose just one. We just chose one week because within one week we can do a pretty good study. And in case we didn't have representative samples in that week, we can always change the week.

Once we reduced the number of rows in our database, we looked for columns (variables) we could remove. And we observed that there were variables like hospital address, which we thought was irrelevant to our study, so we removed it. We just kept the essential variables that give us interesting information to do a study on. Once we had treated our input data matrix to just have important information, we tried to read it.

Data mining process

In order to carry out this project, we have followed a set of steps that have allowed us to properly obtain the necessary data, easily manipulate it and draw conclusions from it. The whole data mining process which we followed in this project has been the following:

Obtaining the dataset

We searched multiple public data repositories until we found a dataset that met all our requirements; the data was reliable, relevant in the present, representative of the real world, easy to understand and came from a trusted source.

Analyzing and preprocessing

Once we gathered the data, we analyzed it in order to identify the most important information, as well as the information that was not relevant for our study. We eliminated a very significant amount of columns that we considered irrelevant and that only added more noise and unnecessary dimensionality to the problem. We also removed a big proportion of the rows in the dataset, striving to keep only the relevant information without introducing any bias in the data. Finally, we identified which values were missing and applied the most suitable imputation techniques for each case, making an effort to keep the dataset representative and to not bias the data.

Performing Principal Component Analysis (PCA)

Once we had our data preprocessed and ready to be analyzed, we performed PCA in order to identify which features were the most relevant in our dataset. We rendered a scree plot to visualize the variance and accumulated variance in our dataset relative to the number of axes used. After finding the axes that describe the most variance in our dataset, we plotted our variables along those axes in order to analyze more features about them, such as their correlation or how much variance each introduces to our data.

Applying clustering techniques

We applied a hierarchical clustering algorithm to our data in order to analyze and visualize how our data might be subdivided into distinct classes. We applied the clustering techniques that fit our goals the best given our dataset. After applying the clustering algorithm, we visualized the results using a variety of plots, such as a dendogram, which allowed us to

determine the optimal number of classes for our dataset, or a set of pair plots that allowed us to visualize the clusters in two-dimensional space given a set of pairs of variables.

Drawing conclusions

After analyzing our dataset, we had all the necessary tools to draw conclusions about both the data and the correlation that it has to the current world situation. We were able to work with and perform analysis on a piece of real world data, while following a complete and rigorous set of procedures from beginning to end.

Preprocessing

Defining the scope of our dataset

The first step of preprocessing we took was to define which variables and records we will keep, and which we will delete. So, first of all, we decided that we will reduce the number of records because we had 295k records. We observed that we had different weeks for every hospital, hence we chose just one. We just chose one week because within one week we can do a pretty good study. And in case we didn't have representative samples in that week, we can always change the week.

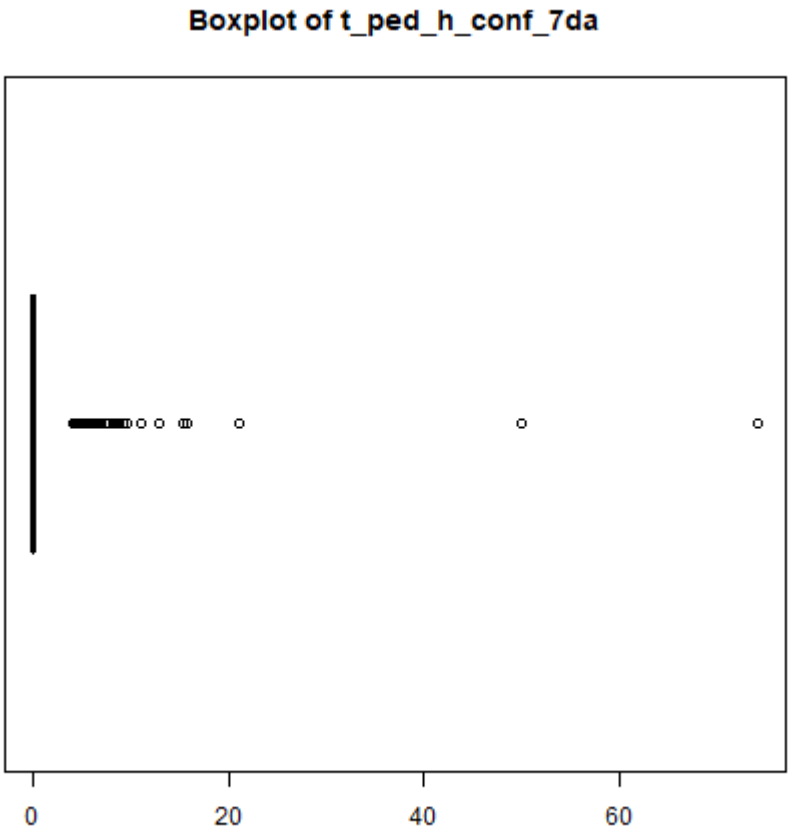
Once we reduced the number of rows in our database, we looked for columns (variables) we could remove. And we observed that there were variables like hospital address, which we thought was irrelevant to our study, so we removed it. We just kept the essential variables that give us interesting information to do a study on. Once we had treated our input data matrix to just have important information, we tried to read it.

Reading the data and solving data formatting problems

When we got the dataset and started reading and declaring the data we encountered some minor problems. Luckily our dataset was completely in English so we didn't have any problem with multiple languages nor special characters. The main problems that we encountered were that the missings we coded differently from NA, some of them were -99999 in numerical variables, and in categorical variables were strange values. Also, the categorical variables were coded using numbers and we had to rename the rows in the document with those variables. Finally, because of an error in the script, some numerical values were treated as dates instead of numbers, causing a lot of trouble understanding the plots.

Outliers

Part of the preprocessing consists of treating outliers. At first, we didn't execute a specific statistical test because we can't assume without further analysis that some variable has a specific probability distribution. So in order to treat them, we did some graphical representation of the data.



As we can see in the previous figure, it is clear that we have some outliers in our data, for example in those plots we can see very strange values, differing from most of the data. In theory class, we studied that using boxplots to distinguish outliers can be dangerous because boxplots can overlap points and it can be hard to distinguish certain points, but in this case, we think it's okay to suppose that those values are real outliers, because of their distance with respect to the other points. Once the detection of outliers is done, the next step is treating them.

Doing it we made several decisions concerning how the treatment will be. Because all the values seemed reasonable so we decided to leave them as they are. We decided to do so because eliminating them could lead to the data not being representative.

Refining the data

Once we had treated the outliers in our database, we looked for irrelevant information in our database. So, we searched and we noticed that some variables looked interesting but a lot of their records had just zeros and there was no information. We did some plots and spotted that these kinds of variables were unbalanced and didn't have relevant instances. Hence, we removed these columns. We kept looking and kept removing until we just had variables with good data. We deleted these columns because they had a high percentage of missings, and almost all of the data were zeros, so it was absurd to study them. So the best thing to do

was to delete them. With these steps done, we had a dataset with relevant and interesting samples. Once we got to this point, we had to treat the missings on our records.

Missings

When we were doing the treatment of the missings, we observed that almost all of the missings were related to numerical variables. And all of those variables had the same topic, the average. So, our first idea we found that the appropriate way to fill those missings was the average of those columns. But finally we decided that the best option was to use the knn method because for similar hospitals, we will fill them with the same kind of information. With this method we can avoid having incongruencies between variables. We also had variables that had a high level of missing values, reaching up to 20% of missing values in some cases. So we considered that it was absurd to fill those missing values. And the best way was to delete these kinds of columns.

Another kind of missings we had were the rows that had just minimum information like hospital name and some other variables. So, we thought that these kinds of variables are irrelevant because we will fill them with “imaginary” data, hence, we decided to remove them from the database. We put a condition that checks if a row has more than 50% missing values then it removes the row.

We also had missing values like state name, fips codes, and other types of important data. So, doing analysis, we observed that there were not many of them. So, we considered that we could fill them using real data that could be found in google.

We also had records where we just had the state and the hospital name. So we removed all records of this type. The reason for this solution is that we didn’t have any important information about that record. So, filling all the missings of the row was a worse option than removing the record from the database.

Feature selection

Once we had treated the missings of our variables, we tried to do a feature selection and we noticed that we already had done it before treating the missings. When we did instance selection, we just deleted a lot of non-interesting variables which had a lot of instances with no interesting information. We tried to look if we got some other variables that were non-interesting, but we didn’t find any more. So we proceeded to look if we could create some new variables or if we could convert some types to other types of variables.

New variables

Once done all the previous preprocessing steps, we could create new variables to make the understanding of the model and further analysis easier. We created a binary variable using another column of our data, but then we realized that the binary data was more useful than the actual data so we decided to remove the actual data from our dataset. It was more useful because it was about the child that had covid-19 confirmed in every hospital, but the values were most of them 0, about 1 non-zero value for 500 values viewed. And the actual values

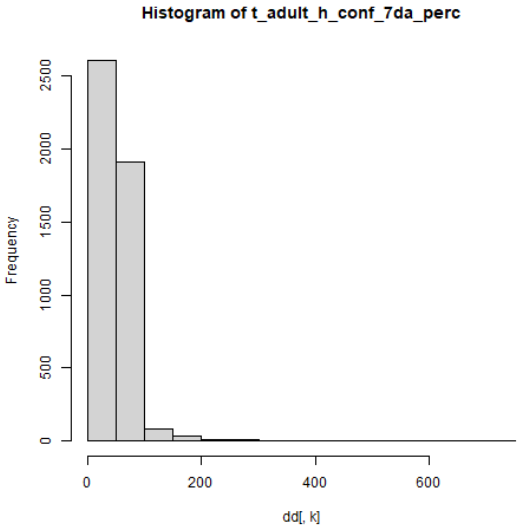
were all of them in between 1 to 10. So we created the binary variable if there were children with covid-19 diagnosed.

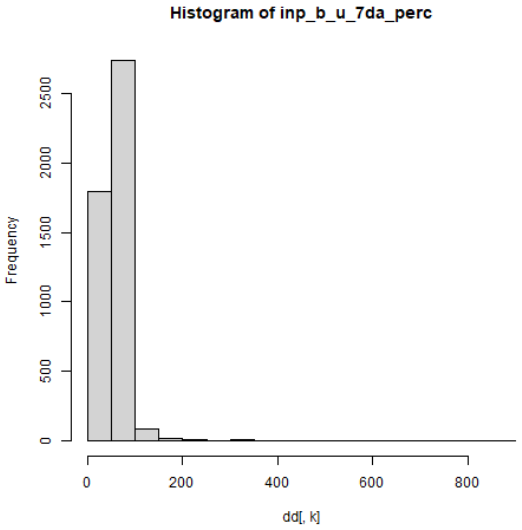
We didn't create new variables because at the start of the project we started with 104 variables. That number didn't fit the purpose of our project so we reduced them to 26.

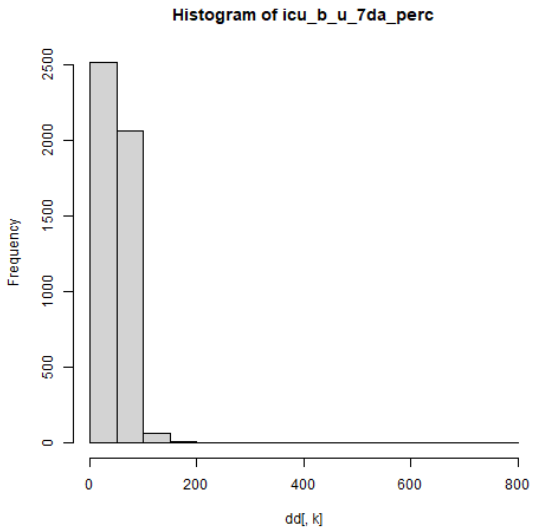
We also created 3 more numerical variables using 2 columns for each one. So, we were combining two variables into one variable. This is because we considered that data in percentages is more understandable than the number of beds. Another reason for this was for better comparison between different hospitals. We can compare the percentage of used beds between hospitals, but not the number of beds used because every hospital has a different number of beds depending on the state they are in and the population in that state. The variables that we created were the followings:

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

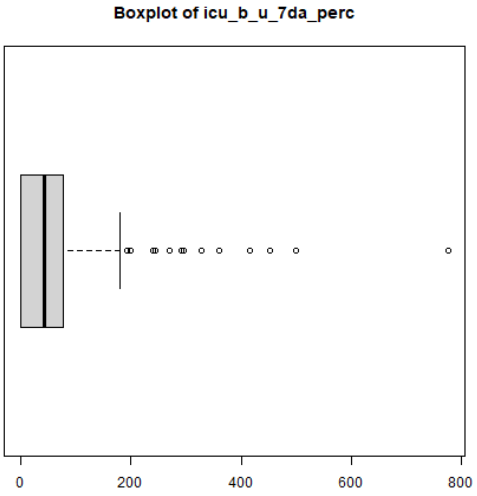
Descriptive statistics of the new variables

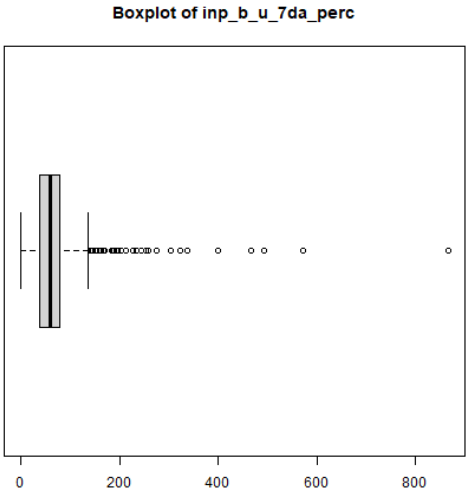


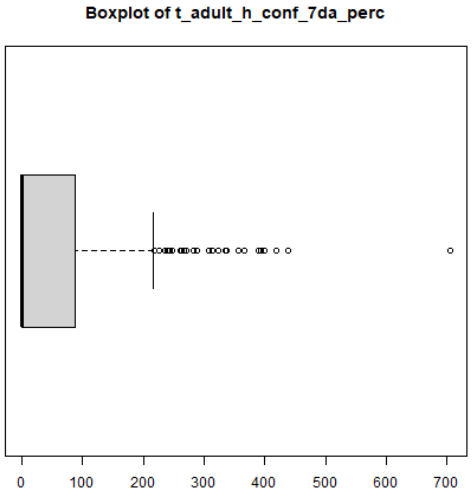




As we can see from the previous plots, they are pretty similar. We can expect that the most of the data stays in the range of 0 to 100 because we are talking about percentages. It could be possible that there exists some values higher than 100% because we did the calculations using a 7 day average of our data, so it's possible to have in some cases more beds occupied than the total of beds. But it's also possible that some errors exist in our data. We could see them at extreme values of 800%. Those are real outliers and have to be corrected. We can see better those cases in our boxplots below.







As we have said earlier, most of the data is confined in the range 0 to 100 and exist some values higher than that, we assume that is because of the use of the average and also because during Covid-19 some weird cases could have appeared, such as having more beds occupied because some of them were donated from another hospital, but they weren't counted in the total beds. Also we assume that the most extreme values as 800% are errors in our data and we might treat them as so

Basic statistical descriptive analysis

In this section we are going to, firstly, describe in a basic way the original (without being treated) dataset variables. Then we will show some side by side comparison between some pairs of interesting variables. Finally we will make some conclusions about the data.

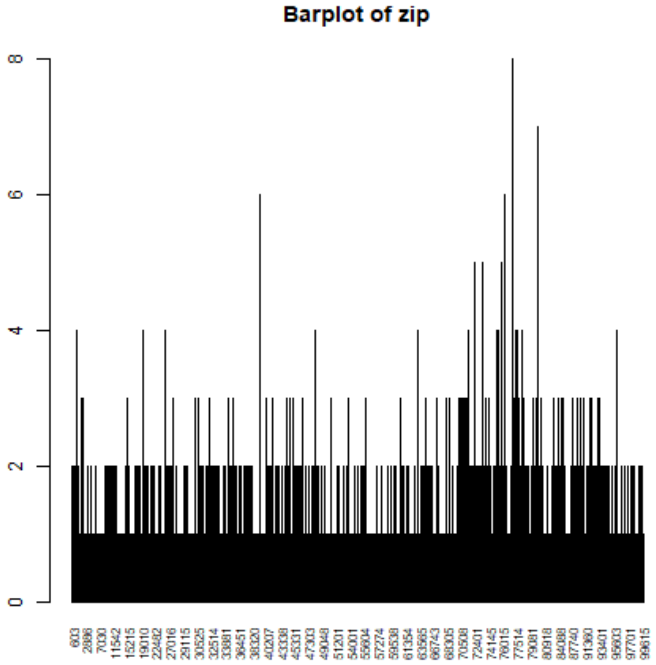
But to start we will describe the categorical variables showing the barplots and the pie graphics. Nevertheless, some of these categorical variables have a huge amount of categories and we won't show the pie graphic because they won't give us any useful information.

Univariate

Categorical variables

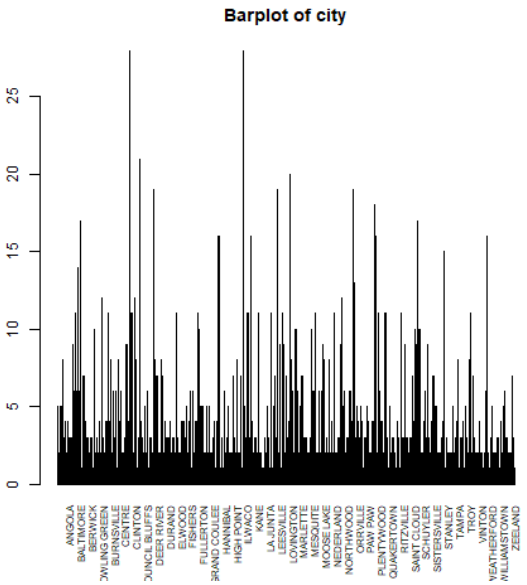
Zip

In this first variable we can see only the barplot. We can see that we have a lot of repeated (at least 2) zip codes among hospitals. Indeed, there are up to 8 hospitals with the same zip code. This is interesting information because we could check here the proximity among hospitals with different zip codes or even study the impact in some zip code zones separately.



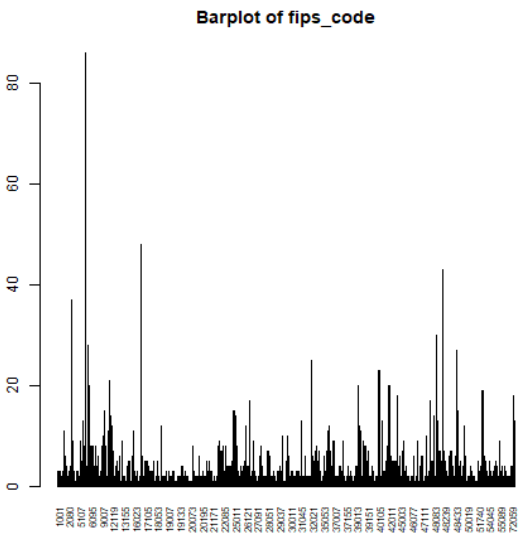
City

The city variable is also really interesting because we could do some studies about the impact in some isolated cities. We can see in the barplot that the samples are from a lot of different cities, which are good news to make sure that we are taking into account different localities. Also we can see that there are a lot of hospitals from the same city and, furthermore, there are, in some cases, more than 25 hospitals coinciding in the same city.



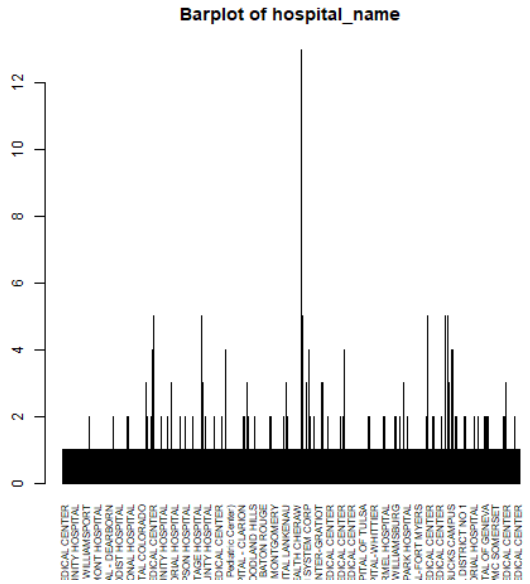
Fips_code

Here we have another locator variable and we can see, as seen in the previous locating variables, that there are quite a lot of hospitals from the same region defined by this variable. In this case we have more than 80 hospitals that are geolocated into the same zone as we can see in this barplot.



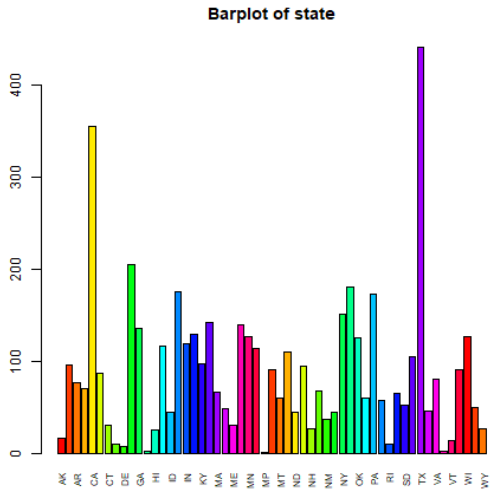
Hospital name

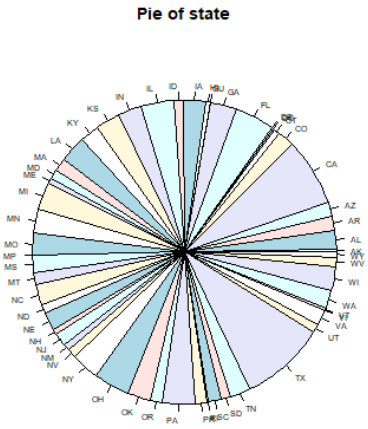
This variable expresses the name of the hospital. We can see that almost all of them are different but there are a few ones repeated. There are even a few that are repeated more than two times achieving up to 13 hospitals with the same name.



State

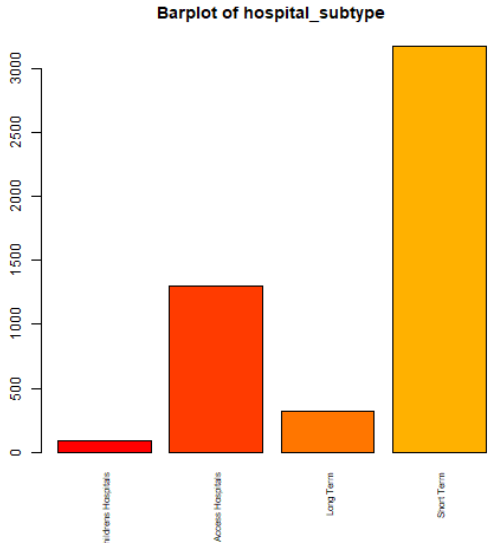
In this variable we introduce the pie graphics where we can see that the hospitals are quite evenly distributed among states but we have two major states (the one in yellow, CA, and purple, TX) where there are quite more samples. We don't think that's an issue because we could maybe retrieve only the samples on one of the states and check properties of it.

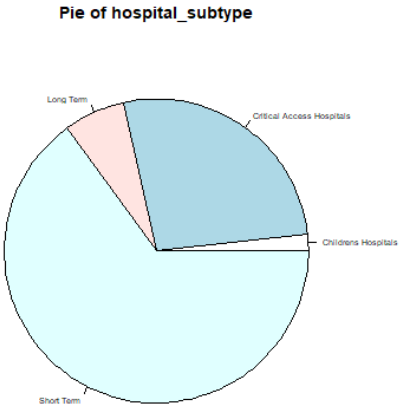




Hospital subtype

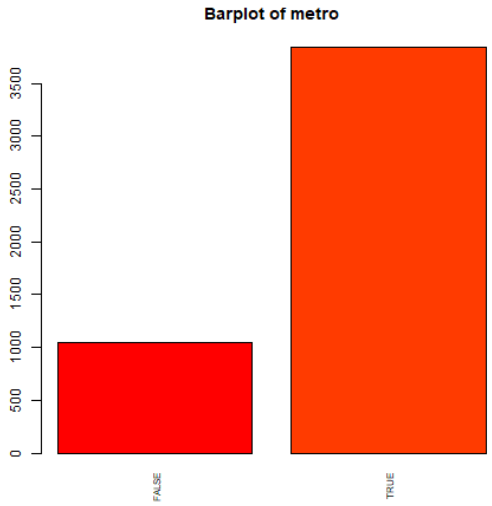
Here we find our last categorical (not binary) variable. We can clearly see that there are only 4 possible hospital subtypes. The ones which have fewer samples are the children's hospitals and the long term ones. On the other hand, almost 65% of our samples are about short term hospitals. Nevertheless we were shocked to see that many samples were from critical access hospitals.

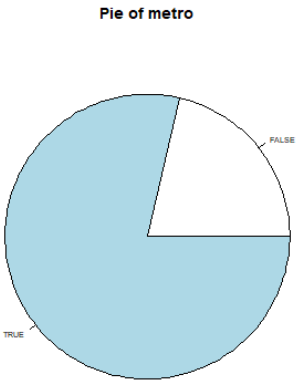




Metro

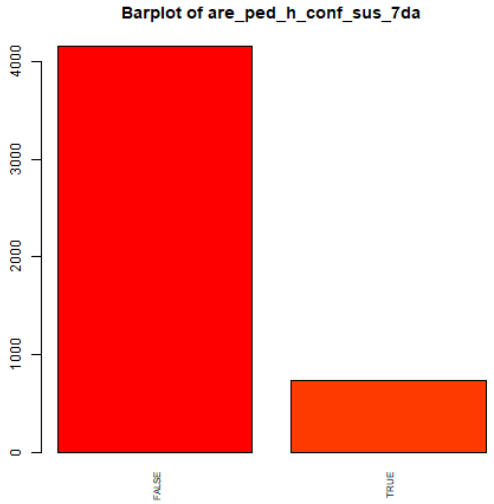
This is our first binary variable which indicates whether the hospital belongs to a metropolitan city or not. We can see that more than 75% do. Even so, we think that this information is not a problem because like this we can have access to different city structures impacts and see how the results were from a global perspective.

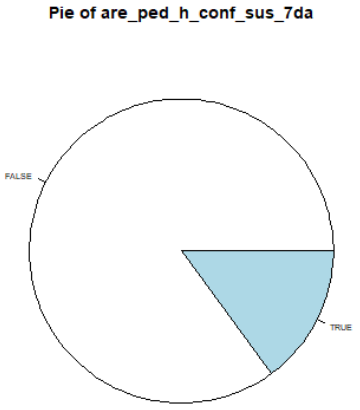




Are_ped_h_conf_sus_7da

This is our second binary variable which is about having pediatric patients which were suspected and confirmed with covid-19. We can see that more than 80% of the hospitals declared that they didn't have any of those in the whole week. We think that this is quite strange, nevertheless we decided to trust the source.



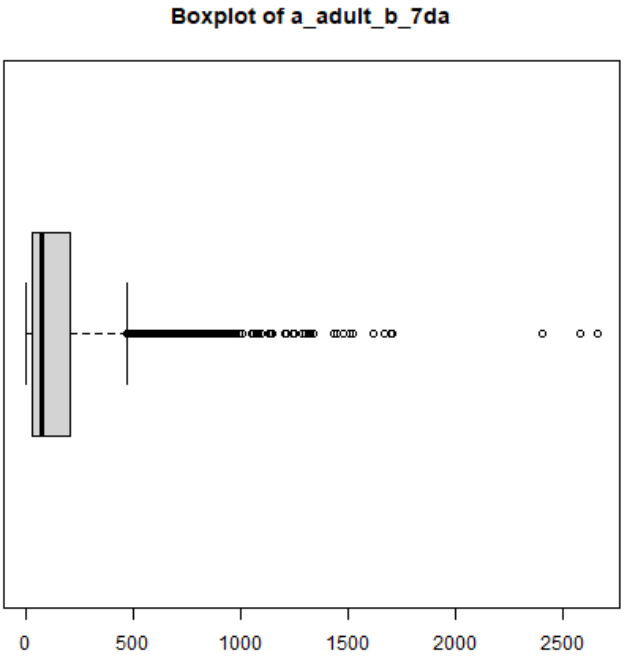


Now we are going to describe the numerical variables. In the original dataset we obtained quite bad graphics as we will see now. Some numerical values were negative, we had many nan's, etc. We produced boxplots and histograms of all the variables. Even so, we will only reference the boxplots because the histograms aren't very representative. This is because all the variables have a great amount of zeros and null values and, with the outliers we will now see, the histograms obtained were not really fancy.

Numerical variables

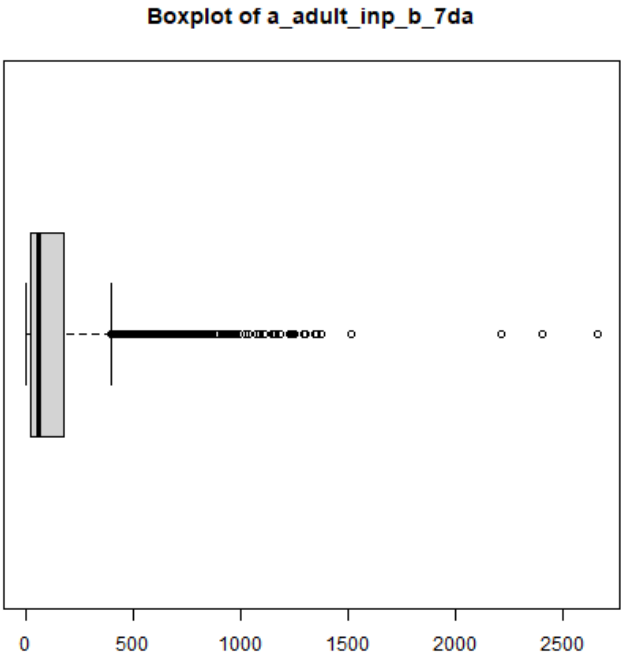
A adult b 7da

In this first numerical variable, as well as we will see in the following ones, we obtained a long boxplot with the "center of mass" located around small values (0-100) and many outliers achieving up to 2500 in some points. As we will comment later, those samples could be from hospitals that were really affected from the covid crisis so we can't delete them by considering that they are errors or missings. More or less all the following boxplot graphics looks similar to this one with many values around zero and many outliers exceeding the cuartiles.

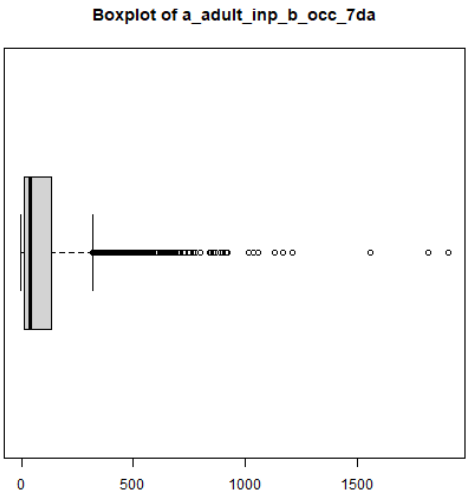


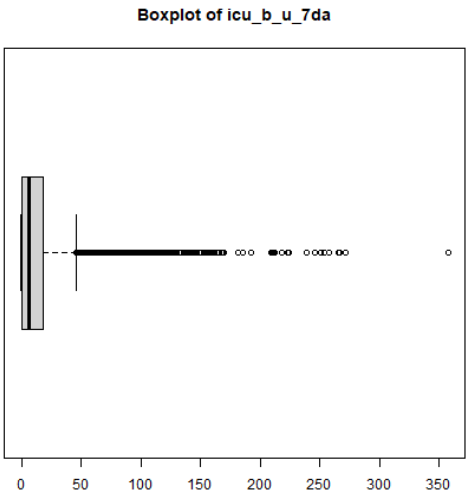
A adult inp b 7da

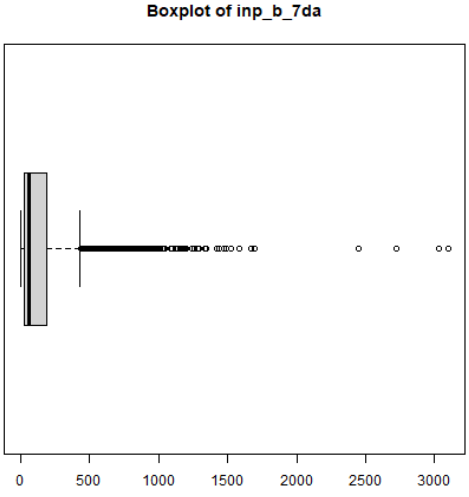
In this box plot we can see a similar result to the previous one. We have many variables creating a “center of mass” around small values and a large tail with some outliers to be studied and maybe treated.

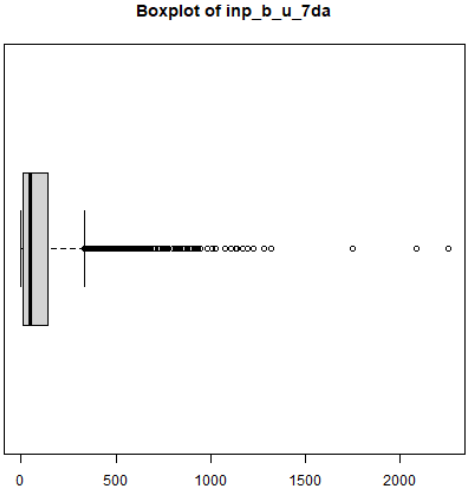


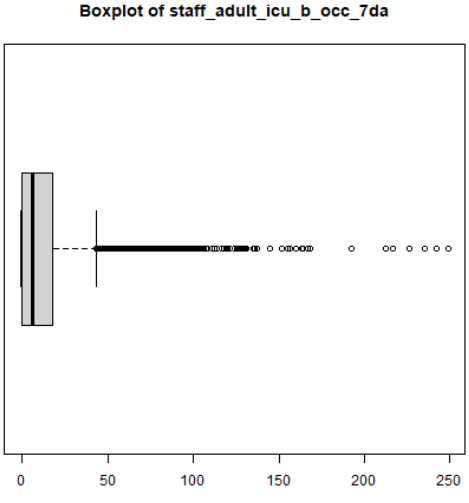
Similar variables

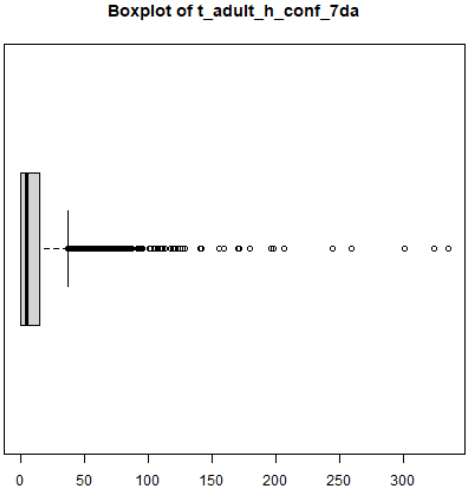


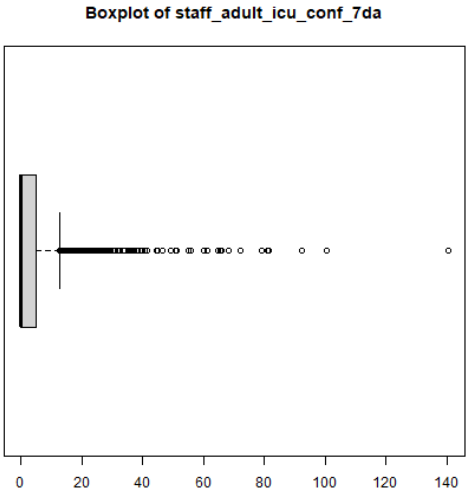


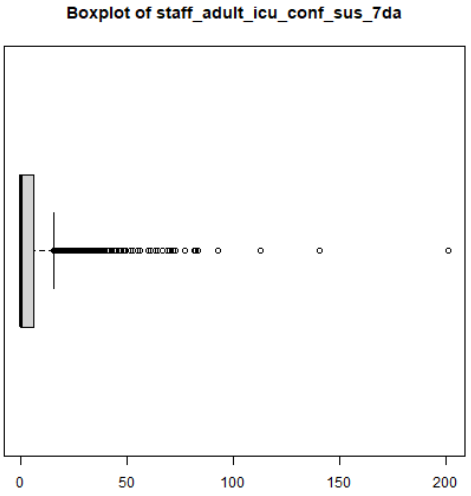


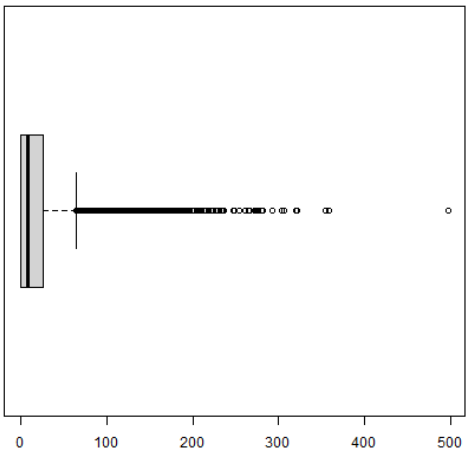


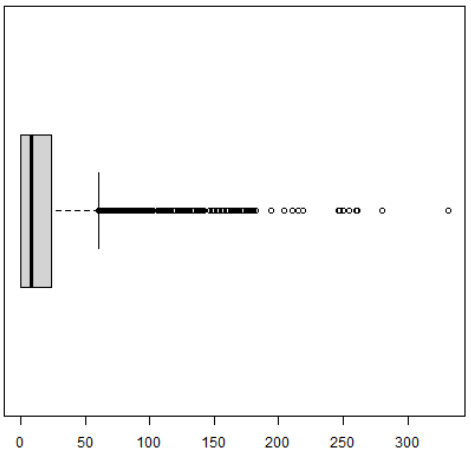


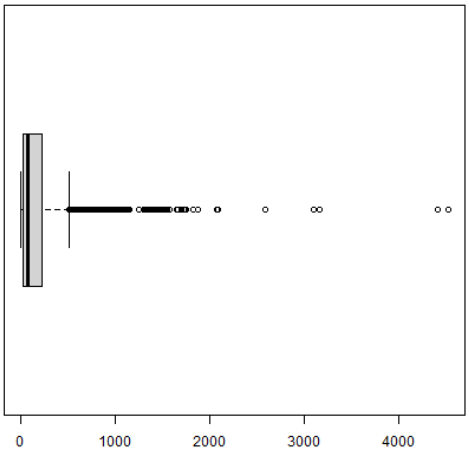


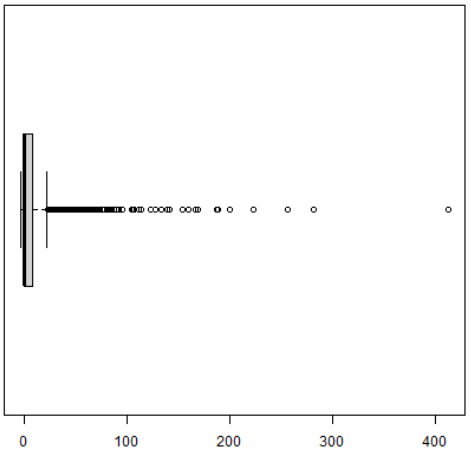

Boxplot of staff_adult_icu_conf_7da
0
20
40
60
80
100
120
140


Boxplot of staff_adult_icu_conf_sus_7da
0
50
100
150
200


Boxplot of t_icu_b_7da
0
100
200
300
400
500

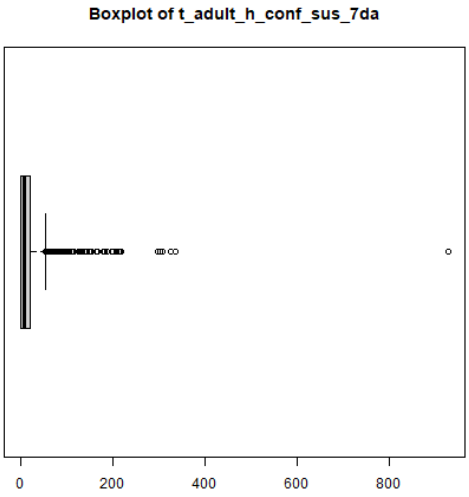

Boxplot of t_staff_adult_icu_b_7da
0
50
100
150
200
250
300

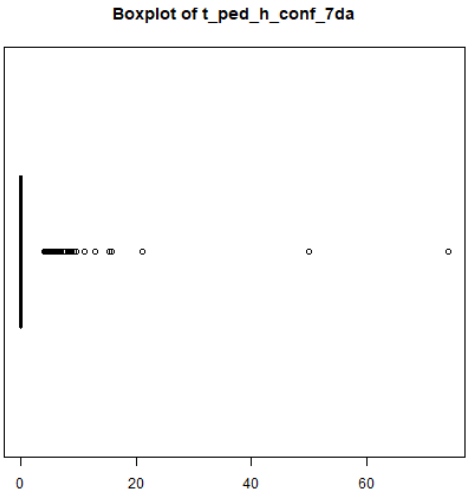

Boxplot of t_b_7da
0
1000
2000
3000
4000

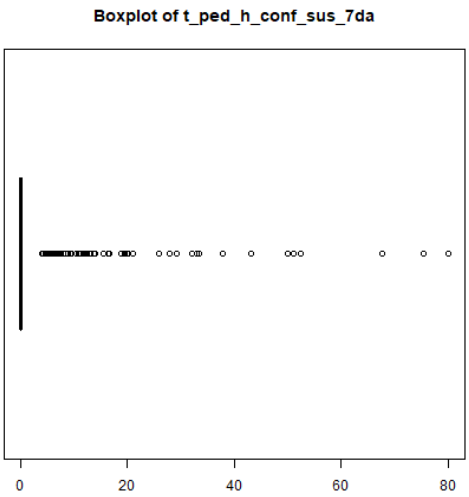

Boxplot of pda_adult_conf_7ds
0
100
200
300
400

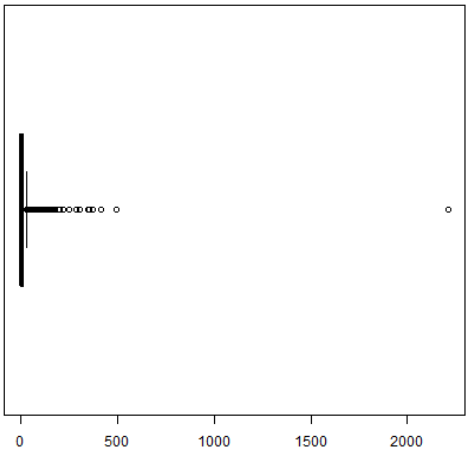
Even worse variable boxplots

In this cases we can see that the “box” it’s only represented as a line because a great amount of the data has zero or null values. Even though we can see that we have many outliers, as happens before, “destroying” the equilibrium of the plots. In the first (top left) and last (bottom right) one we can see a single sample being the outlier. In this case we could try to delete this sample but we will see that in the following section and study if we can do it or not to don’t skew the sample.





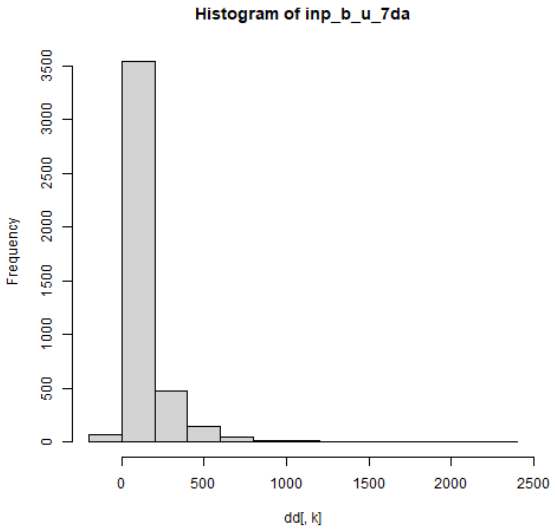


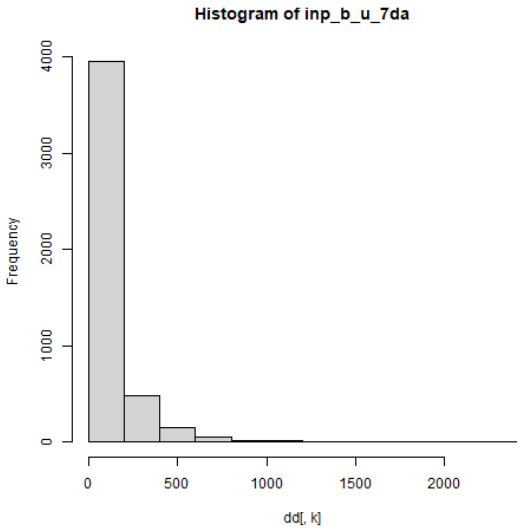


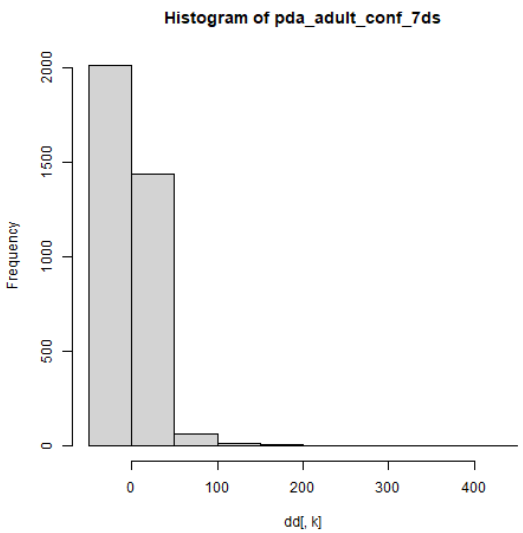
Preprocessing changes

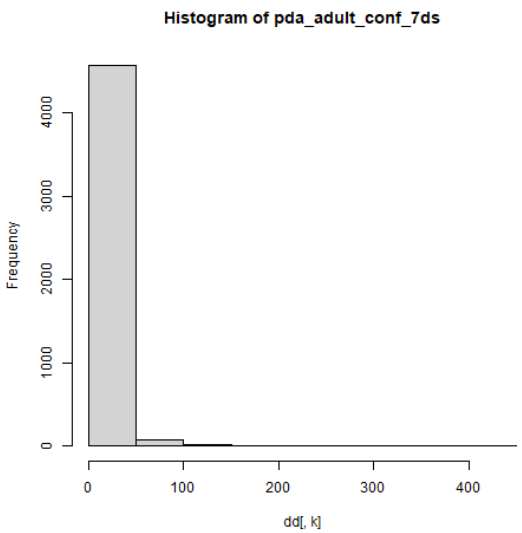
In this section we will see some examples of preprocessing resulting in a much better result. In both following pair cases we can find the histogram of a numerical variable (inp_b_u_7da and pda_adult_conf_7ds) before and after suffering the preprocessing with tha KNN method.

We can clearly see that both variables had some (or many in the second case) with negative value. That's not possible because the variables are about averaging and adding the total week values. As we classified them as missing values then we just imputed them using KNN resulting in the results of the right that didn't contain any negative value. We can also clearly see that the zone that increases his height the most is the nearest to the zero or negative values which is exactly how KNN should work by imputing using the value of their nearest neighbor.









Conclusion: How is our data?

To conclude the analysis part we can say that our data is not the best. Or this is what it looks like at first sight. Our numerical variables have really long tails with many possible outliers (which are not as we see in the preprocessing section). It's also convenient to add that the data isn't gaussian at all, this isn't surprising at all since the information it contains describes a very volatile subject.

Nevertheless we have to make sure that we understand what the data is about. Remember that covid pandemic affected differently in many places. We can see in every box plot that we have some hospitals with severe affectations and many hospitals that aren't that bad. Furthermore, we can see that there are a lot of zeros in our dataset which shows evidence of having many controlled hospitals without severe affectations.

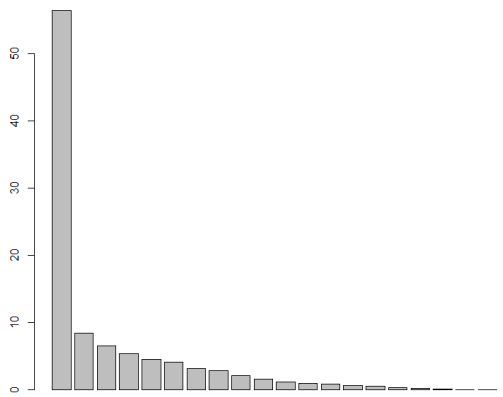
However, with the PCA study and, specifically, with the clustering section, we will see how our data behave and we will be able to achieve interesting conclusions.

PCA

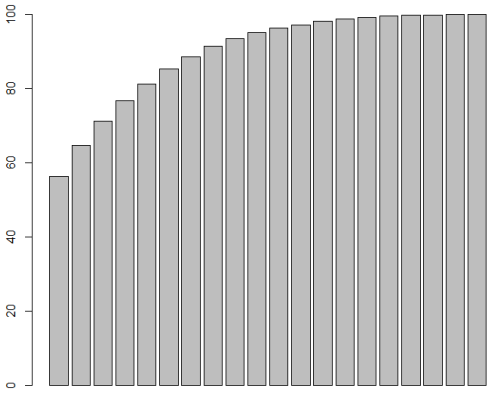
Specifications

Looking at the scree plot and the accumulative scree plot, we can see that the first axis of the principal components describes almost 60% of the variance of our data. We will consider that selecting 5 axis is enough due to the fact that with 5 axis we can describe more than 80% of the variance in our observed data as we see in the accumulative scree plot.

Although the best case scenario would have been a scree plot with two or three axes accumulating 80%+ of the variance, the distribution of the variance description in ours is still decent. The two figures below are telling us that one axis will clearly have the highest weight when it comes to determining any type of variance or correlation between our variables.



Scree plot



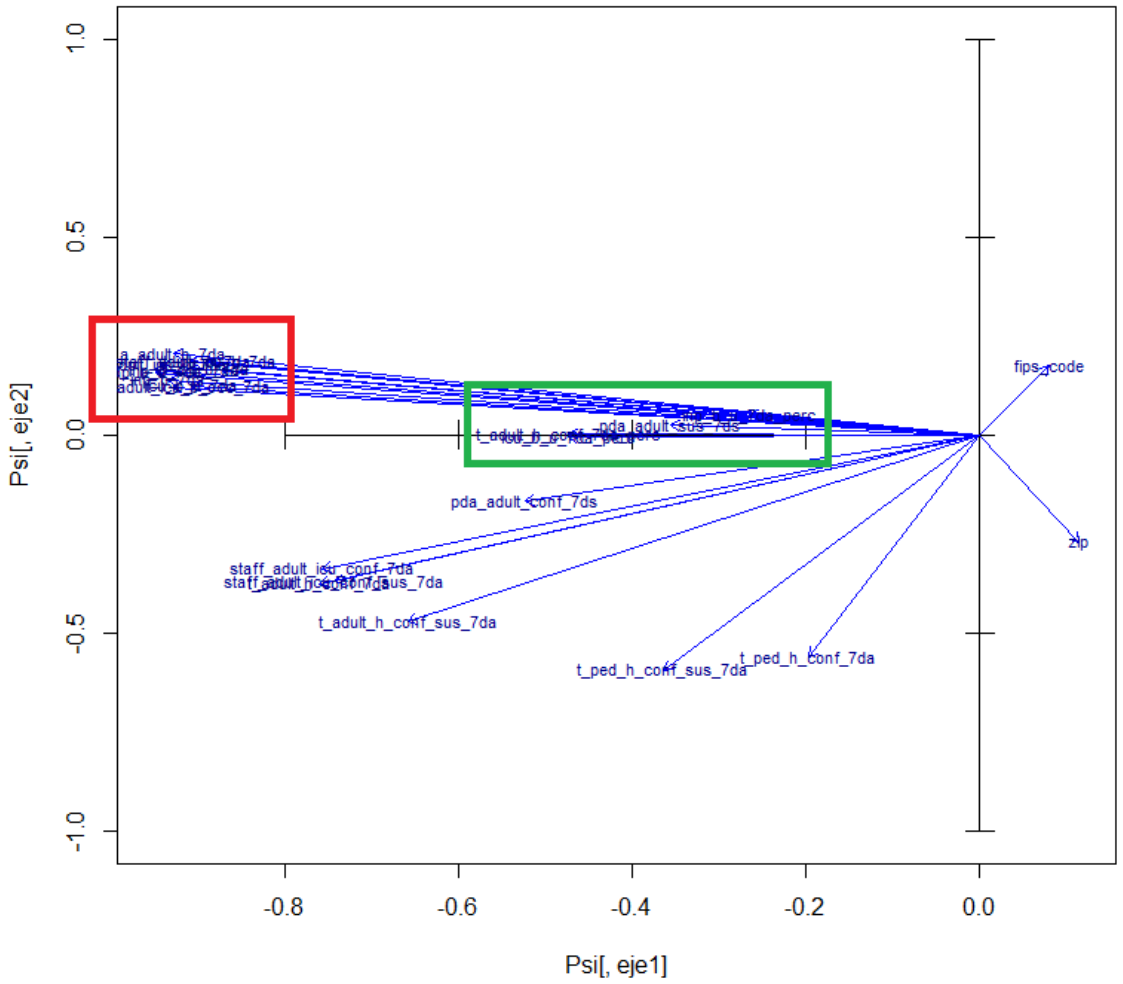
Accumulative scree plot

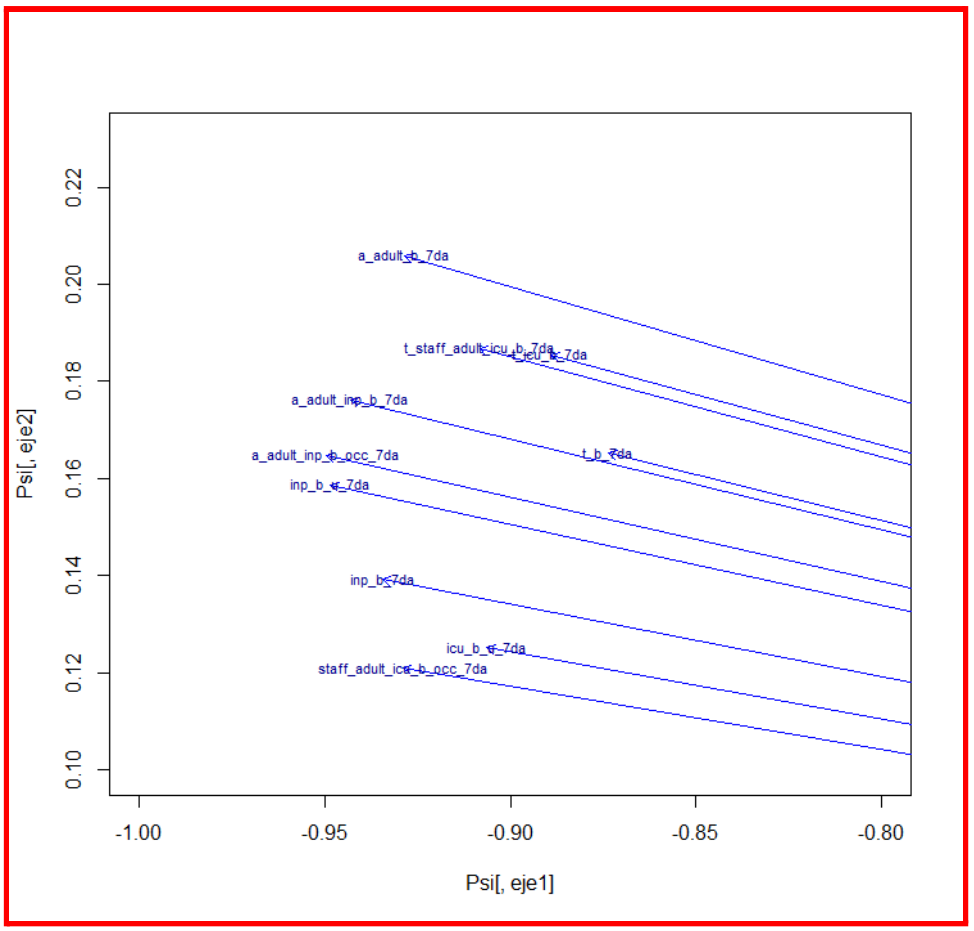
Now that we have analyzed the plots, we will jump right into the factorial planes to determine whether the notable contrast between the first axis and the four following ones can still provide us some interesting results.

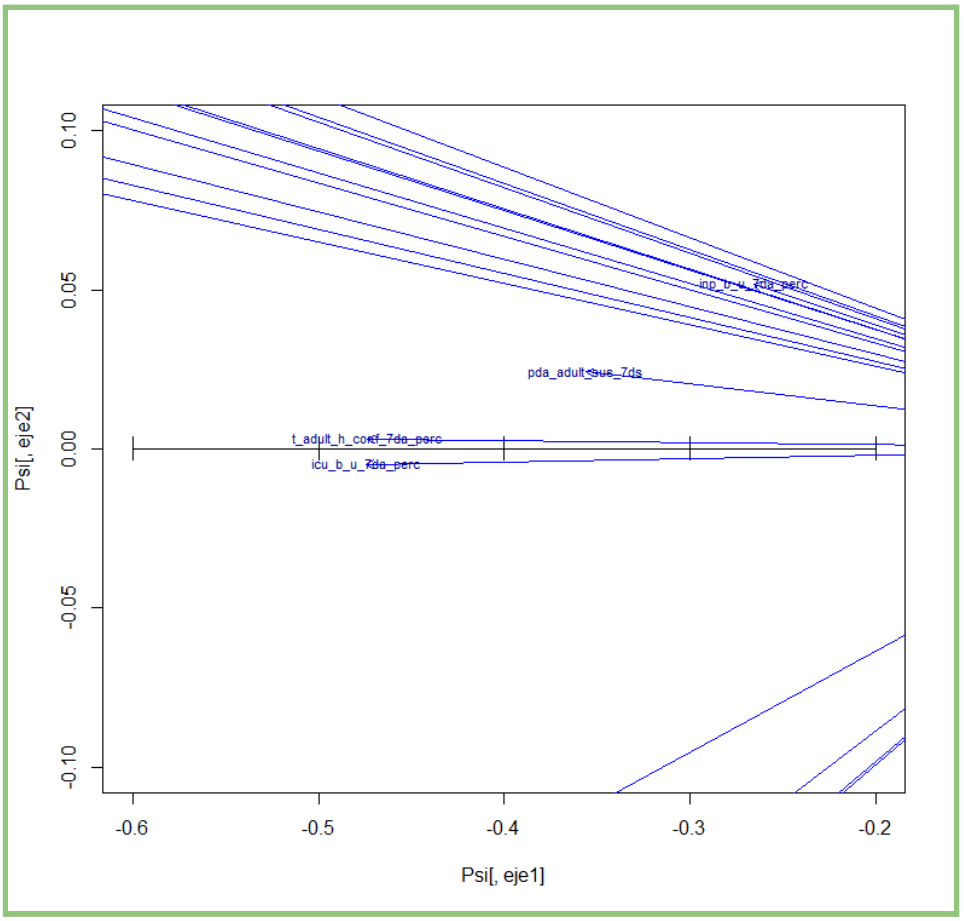
First factorial plane

Therefore, to visualize the data we will plot the numerical variables along with the Principal Components to see if we can detect which variables are more determinant describing the variance of our data or any correlation.

First of all, we will plot the numerical variables along the first and the second axis that describe more variance (we can see it in the scree plot).




Psi[, eje2]
0.10
0.12
0.14
0.16
0.18
0.20
0.22
-1.00
-0.95
-0.90
-0.85
-0.80
Psi[, eje1]
a_adult_b_7da
t_staff_adult_icu_b_7da
t_icu_b_7da
a_adult_inp_b_7da
t_b_7da
a_adult_inp_b_occ_7da
inp_b_u_7da
inp_b_7da
icu_b_u_7da
staff_adult_icu_b_occ_7da

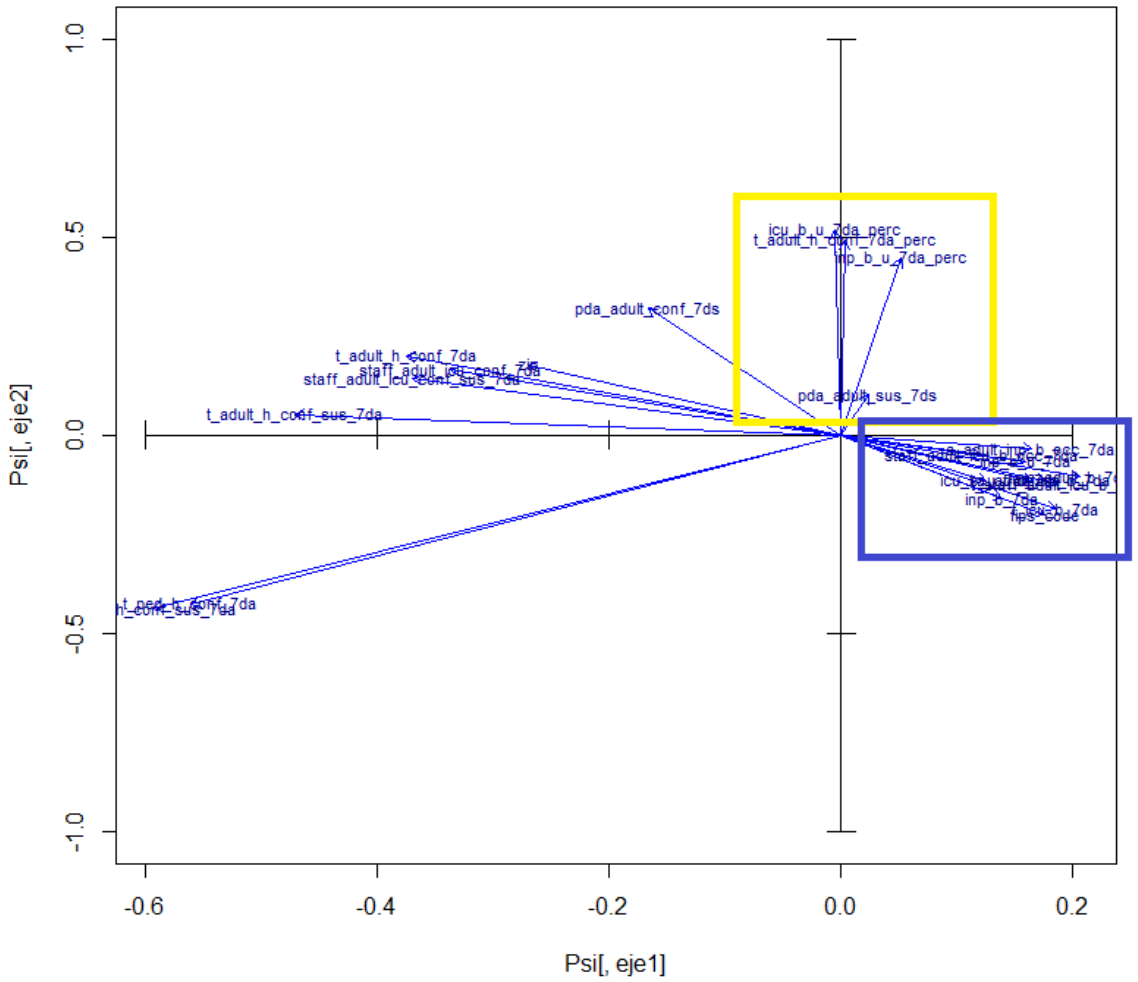

Psi[, eje2]
-0.10
-0.05
0.00
0.05
0.10
-0.6
-0.5
-0.4
-0.3
-0.2
Psi[, eje1]
inp_b_u_7da_perc
pda_adult_sus_7ds
t_adult_h_conf_7da_perc
icu_b_u_7da_perc

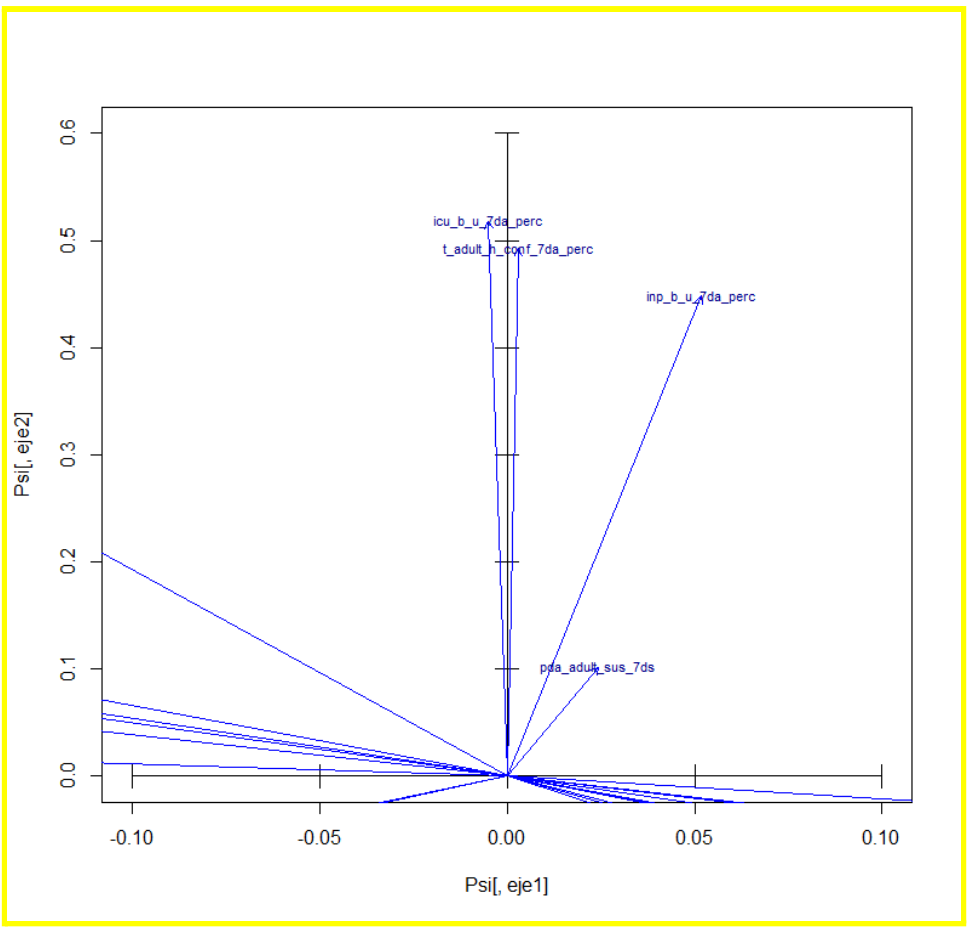
As we could observe, the horizontal axis is related to the variables that describe the percentage of ICU beds occupied and the percentage of inpatients that have COVID-19 confirmed. This relationship is inversely proportional because of the orientation of the arrows. This shows us that this axis is the one that represents the hospital burden.

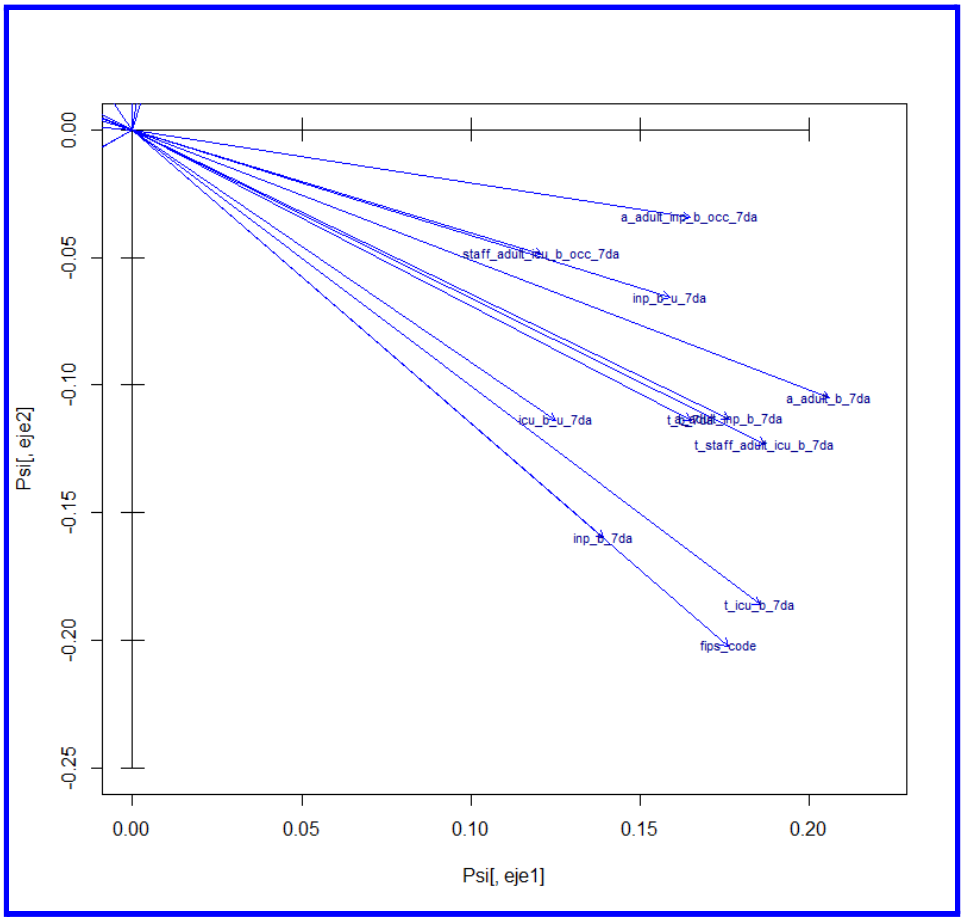
It can also be observed that similar variables related as number of beds occupied and total number of adult inpatients are close. That happens because they are correlated to each other.

The vertical axis is barely related to any variable. This is due to the fact that, as we saw in the scree plot, the first axis describes a lot more the variance of the data than the others, so that is why the implication of the variables to this axis tends to be greater than in any other direction.

In order to dilute this effect, we have plotted the numerical variables along the second and the third axis, because they have less difference between the variance that they describe.




icu_b_u_7da_perc
t_adult_h_conf_7da_perc
inp_b_u_7da_perc
pda_adult_sus_7ds
Psi[, eje2]
Psi[, eje1]
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.10
-0.05
0.00
0.05
0.10


a_adult_inp_b_occ_7da
staff_adult_icu_b_occ_7da
inp_b_u_7da
a_adult_b_7da
icu_b_u_7da
t_staff_adult_icu_b_7da
inp_b_7da
t_icu_b_7da
fips_code
Psi[, eje2]
Psi[, eje1]
0.00
-0.05
-0.10
-0.15
-0.20
-0.25
0.00
0.05
0.10
0.15
0.20

As it is shown in this second factorial map, with axis 2 and 3, the vertical axis is described by the percentage of ICU beds occupied and the percentage of inpatients that have COVID-19 confirmed. On the other hand, the horizontal axis is described by the average of hospitalized adult patients who have confirmed or suspected COVID-19 during the 7-day period and other variables that describe hospital occupation such as the average number of staffed inpatient adult beds that are occupied during the 7-day period.

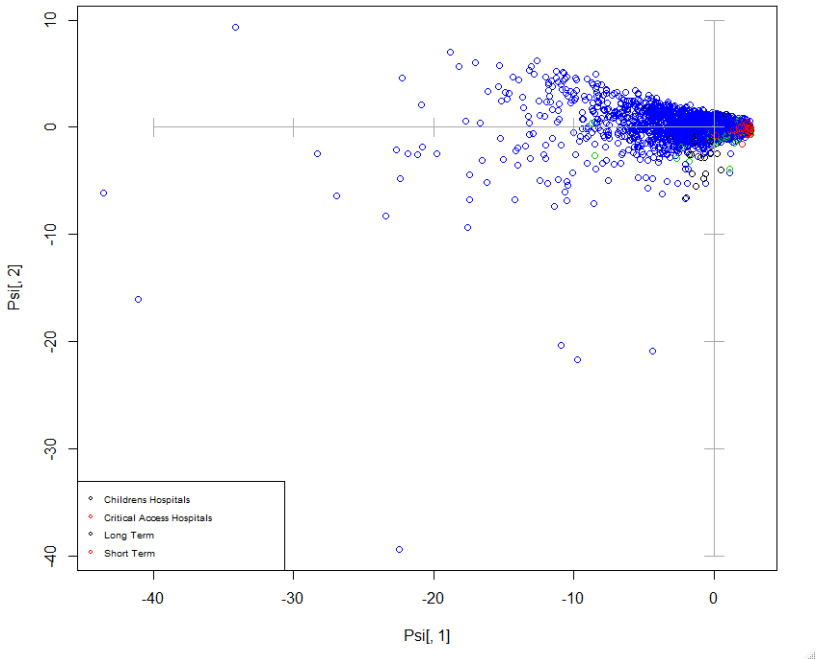
Qualitative Variables

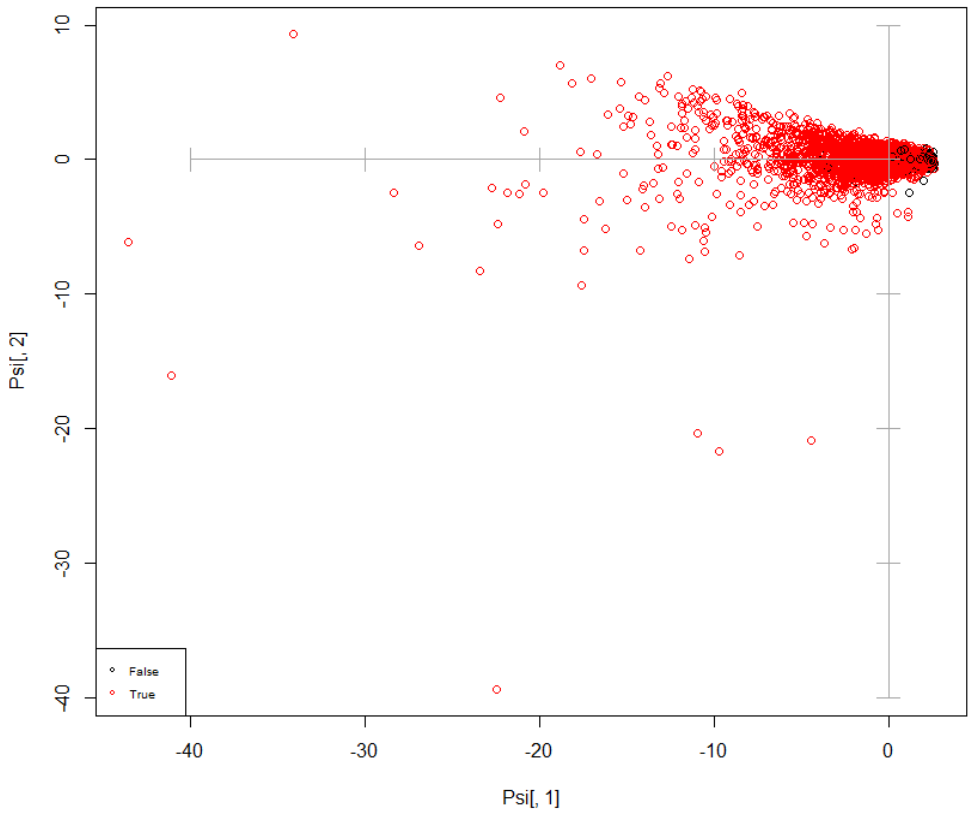
To analyze the qualitative variables in the PCA, the first and second axes components were used. As it was described before, using the first and second components, horizontal axis describe the hospital burden with an inversely proportional relationship.

We tried to plot the qualitative variables state, zip and city along the factorial plane, but it was messy and no information could be obtained.

Hereafter, the next two figures describe the metropolitan and the hospital subtype along the factorial plane. As can be seen in the plots, for the vast majority of hospitals they are metropolitan and Long Term hospitals, and they tend to be more present in the left side of the first axis (with more hospital burden) than the other type of hospitals. It is interesting to see that non metropolitan hospitals and Critical Access Hospitals are found at the right side of the first axis (less hospital burden). It makes sense since Critical Access Hospitals are commonly hospitals from rural areas, far away from big cities; and since if you are in critical condition due to COVID-19 or any kind of disease, you will be transported to a metropolitan larger hospital.

We can also see that Children's Hospitals have less hospital burden than the Long Term Hospitals.





Conclusions

To finalize with the study of the PCA we will try to summarize the principal ideas and give some final conclusions. Firstly, talking about the scree plot, we can say that our data variance can be described, almost at the 60%, with only one axis. This shows that our data is quite correlated within itself. This seems reasonable as long as almost all the variables are directly related with the hospital burden and having, for example, more people confirmed of Covid will be directly correlated with the number of beds occupied in the whole hospital.

Secondly, talking about the first factorial plane axis, we can see clearly that the first axis, the one that describes a great amount of the variance of the data, has almost all the variables relationed when drawn with the second axis. This is directly related to the fact that it's the axis that describes most of the data and the axis corresponding to the hospital burden.

Thirdly, it is interesting to see the results of the second and third axis because they didn't describe much data but they described almost the same amount. This gives us a factorial plane with the vertical axis being again the hospital burden but the horizontal one being the one that describes the capacity and the hospitalized number of people in the hospital.

Finally we can conclude that the study of the PCA was really interesting and reinforced the idea that many variables are correlated representing the hospital burden. Even so, there are others, that are described in other axes, that are also interesting to see and comprehend.

Hierarchical clustering

First of all, as the number of classes in the dataset are unknown, a hierarchical method as the dendrogram should be used in order to adjust the number of classes that distributes more balanced the number of hospitals in each class.

Data used

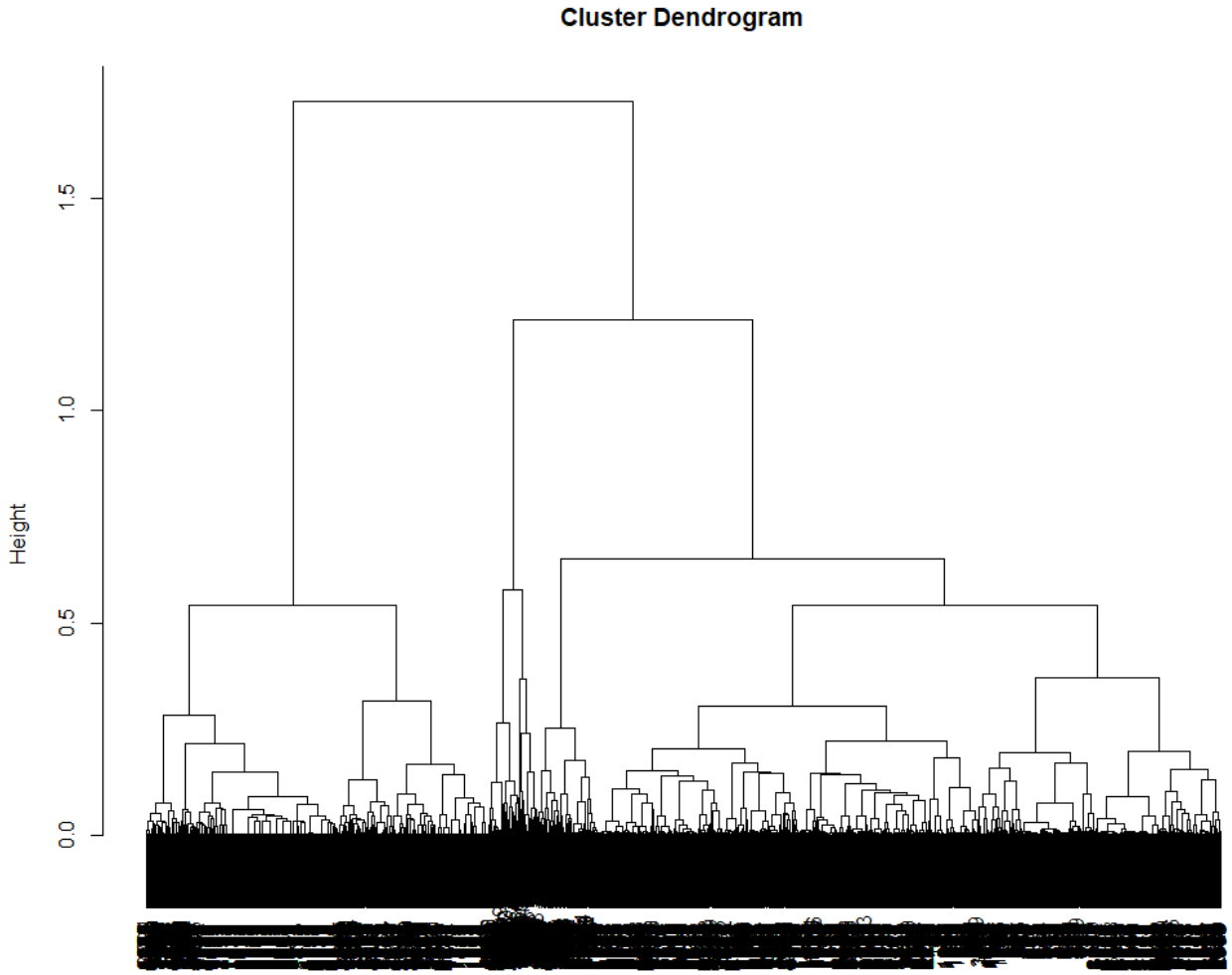
In this part of the hierarchical clustering we decide which data is going to be used. As we know, we can only choose the variables that are not identifiers. Knowing that we excluded the second column which is the variable hospital_name as we considered quite an identifier categorical variable. That's the only variable we excluded since it's the only one that behaves almost as an identifier.

Specifications

In this subsection we will talk about the metric and the aggregation criteria used to compute the hierarchical clustering dendrogram. First of all we will say that we used the Gower dissimilarity coefficient to the square as the metric due to is the one recommended for messy data as ours. We standardized the data in the *daisy* function by using the parameter *stand* setting it to *True*.

To complement this metric we used the Ward.D2 aggregation criteria, which is the most recommended one, to obtain the dendrogram which can be seen in the following section.

Dendrogram



Discussion

In the dendrogram, we can see different optimal partitions with different numbers of clusters. The most clear partition is the one that contains only 3 classes. Even so, we can see that this partition creates one really small class which won't be a problem as we will see in the profiling section. However, we also checked another partition with 7 clusters which is also a plausible partition as long as it's quite clear that can be made in the dendrogram. Before checking the results we will see which are the sizes of the clusters using 3 or 7 classes.

Table of cluster sizes

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

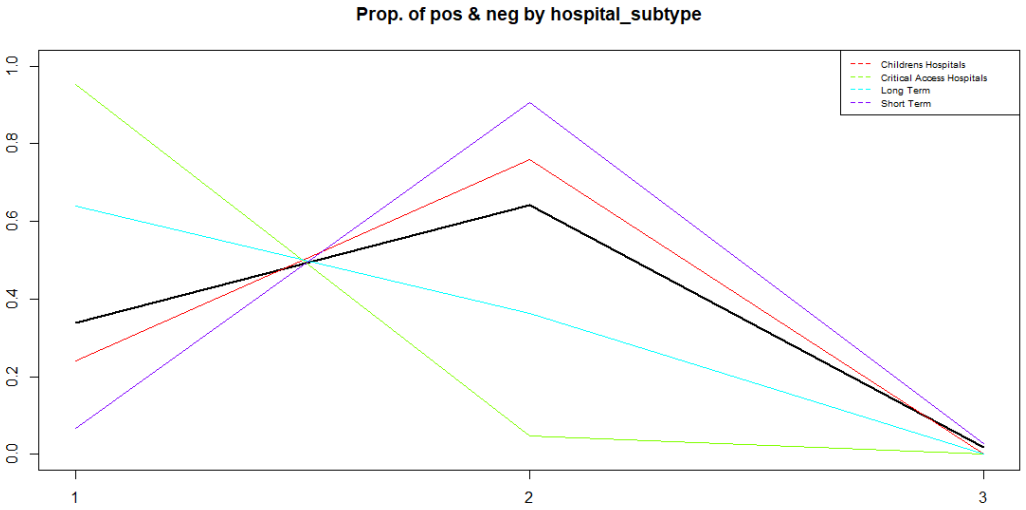
We can see that the 7 classes partitions have equally distributed size values. Nevertheless we will see in the profiling section that we didn't need that great amount of classes and that we can obtain good conclusions with only 3 classes even having a quite small one.

So we finally choose 3 as the number of clusters to partitionate.

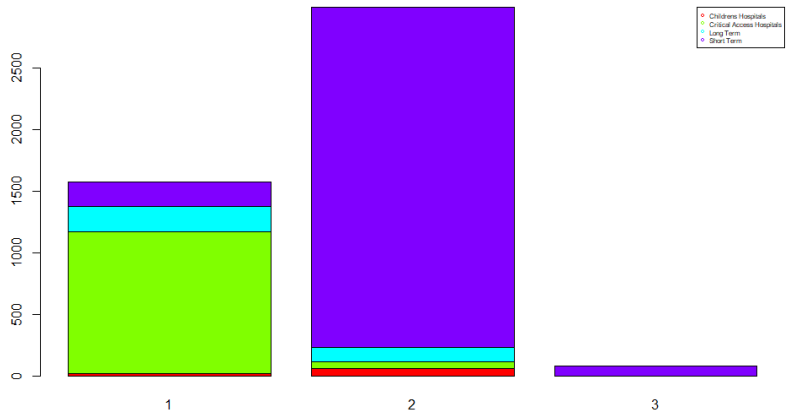
Profiling

Profiling graphs

In the following graphs we will discuss the distribution of univariate variables in the different clusters. The first variable we will comment on is the hospital subtype. The hospital subtype indicates various types of hospitals based on which kind of patient they treat.

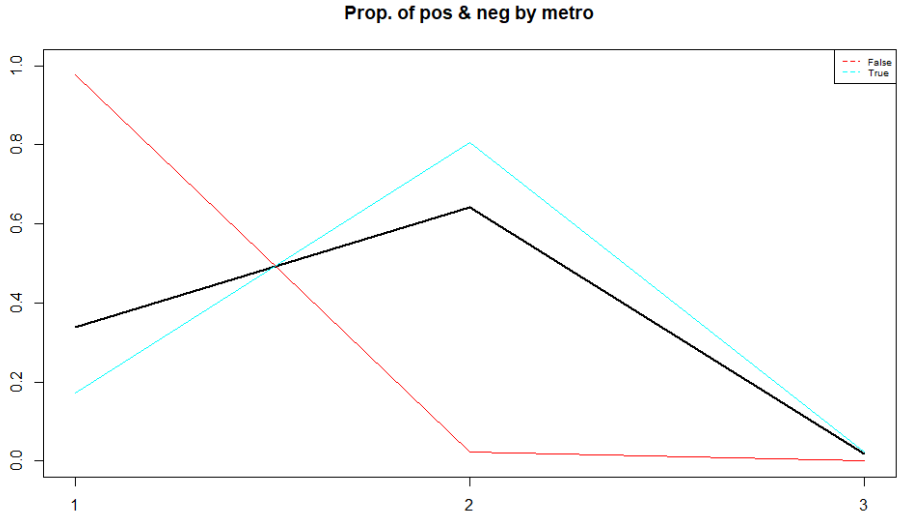


We can appreciate that most of the critical access hospitals are in the cluster one. Consequently, we can see that there is almost no hospital of this type in the other 2 clusters. The Short term type hospitals are in the second cluster. Then, the children's hospitals are distributed between clusters 1 and 2. Finally, we can appreciate that 80% of the Long Term type hospitals are in cluster 2, and 20 % in cluster 1. At this point, you will notice that there is almost no hospital type in the 3rd cluster. We get to the conclusion that the reason for this was because there are almost no records on cluster 3. On the next graph you will notice the difference between the sizes of the three clusters.

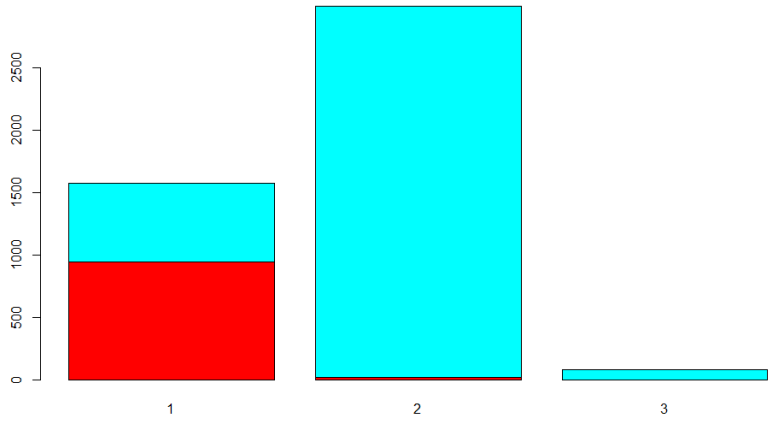


As we have said before, cluster 3 is nearly nothing compared to the other two clusters, which is reasonable as we will talk before about the size of different clusters. As we can see, this graph represents the same information discussed in the previous graph. We can clearly see that the majority of short term hospitals are in the second cluster And the same goes for the Critical access hospitals which are in the first one. If we look at the big picture, we could say that cluster one represents the critical access hospitals and cluster two the short term which will mean that cluster ones containe severe covid incidence and the second one not that severe. With nothing more to say, now we will discuss the graphs of another variable.

In the following graph we can see the distribution of the metro variable in the different clusters. The metro variable means if the hospital is in a metropolitan area. We have suggested that the first cluster contains the hospitals with critical access, so it is reasonable that those hospitals aren't in a metropolitan area. But we can also see that there are a few hospitals within the metropolitan area in this cluster. This could be because they are far from the main city or some other casualty. But not all of them are hospitals with critical access, which makes a lot of sense with what we have seen in the previous paragraphs.



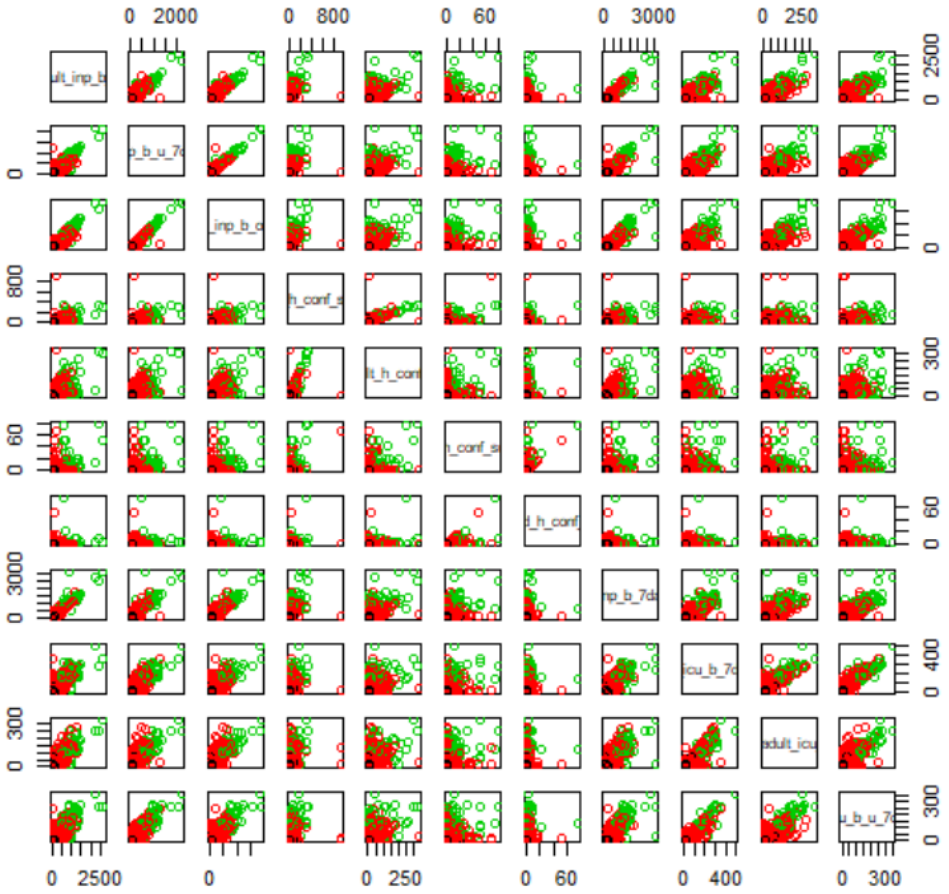
The next plot contains the same information but in a different display. We can observe that the cluster 2 are mostly hospitals who are in a metropolitan area. Finally we can observe that the third cluster doesn't provide much information, it is due to the fact that there aren't many hospitals in this cluster. In conclusion, cluster 1 contains a mix of the hospitals that aren't in a metropolitan area, about 40%, and those who are. Cluster 2 and cluster 3 contain only hospitals that are in a metropolitan area, and most of those are in cluster 2 which is the biggest cluster.



In conclusion, we have seen that cluster 1 contains hospitals with critical access, and it really makes sense that about 60% percent of them don't reside in a metropolitan area. We have also seen that cluster 2 contains hospitals that are in a metropolitan area and most of them are short term hospitals. Cluster 3 is very hard to distinguish from the others due to it containing a smaller amount of records.

Profiling tests

To see which pair of variables could show interesting relations with the selected classes, a ‘pairs plot’ was done.

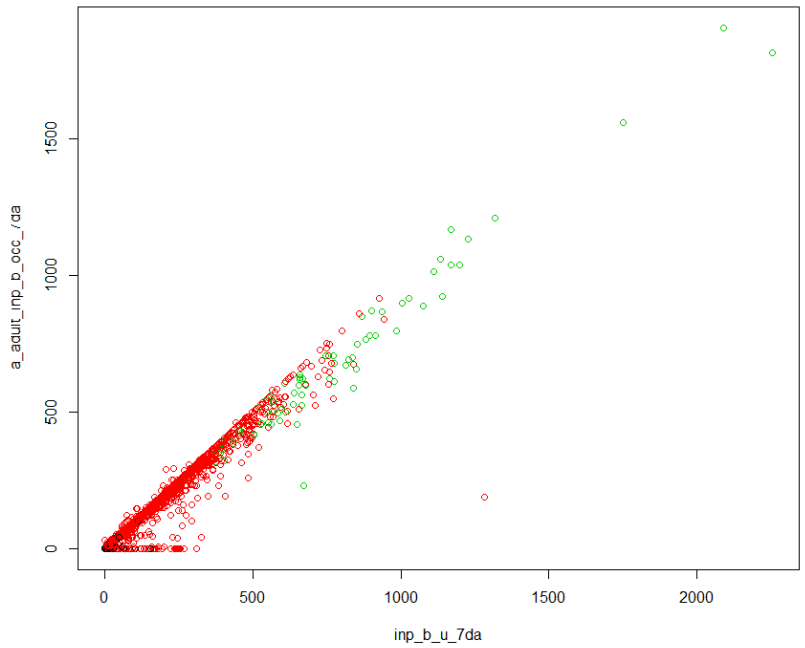


As seen in this figure, some of the variables do not seem to show the clusters separated, whereas others do and also show interesting direct relationships between variables.

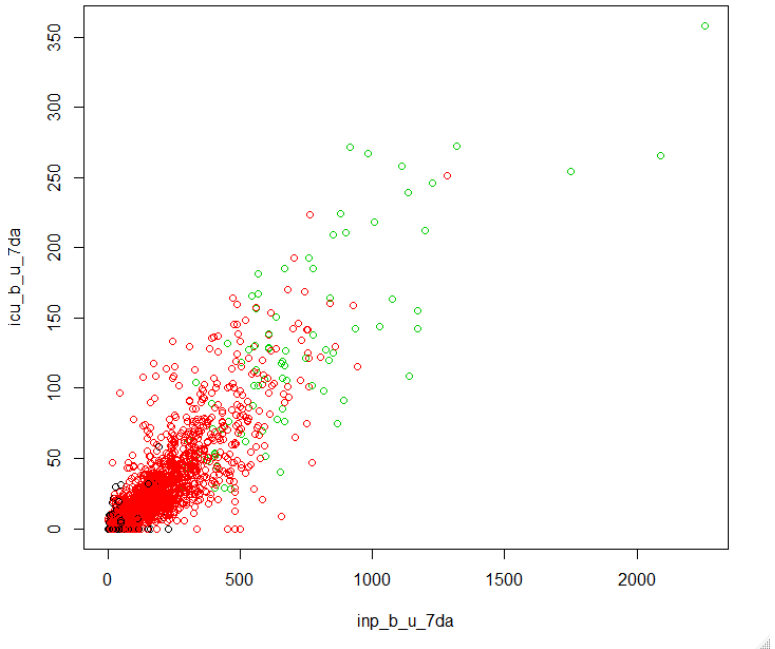
The next pairs were chosen:

- Average of total number of staffed inpatient beds occupied during the 7-day period vs Average of total number of inpatient adult beds occupied during the 7-day period
- Average of total number of staffed inpatient beds occupied during the 7-day period vs Average number of total staffed inpatient ICU beds occupied during the 7-day period

In the first pair of variables shown in the next figure, there is a strong direct relationship that is reasonable because if the number of total beds occupied increases, the number of adults occupying a bed will increase. Also it could be noticed that there is a black class that are in low values of both axis that seems to be the class formed by the hospitals that do not have almost any incidence; a red class that describes the hospitals with a low to medium incidence, and finally, a green class that describes the hospitals with worse conditions.



In the second pair of variables, shown in the following figure, there is also a strong direct relationship, but not as strong as the previous one that almost perfectly fitted a straight ascendent line. It is also reasonable that this direct relationship exists, due to the fact that if the number of occupied beds increases, some of these patients could show a greater virulence and occupy an ICU bed. The same classes as before seems to appear in this pair of variables: A black class that its formed by the hospitals with almost no incidence; a red class that are formed by the hospitals with a low to medium incidence, and a green class, composed by the hospitals with most indicidence.



Templates

As we had seen in the last section we can define a template for each one of the classes obtained. The result would be the following:

3 classes	**1**	**2**	**3**
Size	1580	2993	84
Template	Severe covid incidence	Medium covid incidence	Low covid incidence

With this template classification we can describe quite correctly all the data of our dataset and conclude that the hierarchical clustering method used, obtaining the dendrogram, was quite useful and the resulting division of clusters was correctly approximated.

Global discussion and general conclusions

Throughout the whole process of finding, manipulating and analyzing our data we have been able to draw some knowledge on how to carry out a complete and thorough study on an arbitrary dataset.

As far as information retrieval goes, the source of the data has been one of the main focuses. We had to make sure that the data was reliable, complete and trustworthy, while also providing valuable and relevant information about the real world.

Despite choosing a dataset that was already processed and contained plenty of useful information, the preprocessing stage was very important in order to carry out the rest of the work, since the dataset was still flawed and needed a treatment before we could start diving deeper into the information it contained.

By plotting the variables and analyzing their features individually as well as collectively, we were able to trim the data down to the most essential parts. We also learnt that irrelevant data does matter a lot when it comes to synthesising the results, since it adds unnecessary noise and overheads to the whole process, while also distracting from the relevant information in the dataset. Despite this, the whole preprocessing had to be done with care, since removing data arbitrarily could bias the dataset and lead us to the wrong conclusions.

We also 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. We had to make many decisions in order to maximize the accuracy and the logical meaning of the given results. We had to properly understand all of our variables in order to synthesize the obtained plots throughout the data analysis and we also had to be constantly aware of the data limitations and noise in order to interpret and fathom the outcome of every researching process.

Looking back at all of the descriptive statistics and plots, it has become apparent that we were able to extract significant information out of the raw data we came across at the beginning of the whole procedure. We can positively claim that the algorithms used were able to give us greater insight of the data patterns in a way that made us understand even better the COVID-19 situation in some of the US hospitals. Thanks to the illustration of the variables correlation we started gaining deeper knowledge for future pattern recognitions in the clustering phase, which gave a general classification of the different levels of hazard around the country by intertwining the samples that could be related in terms of health jeopardy.

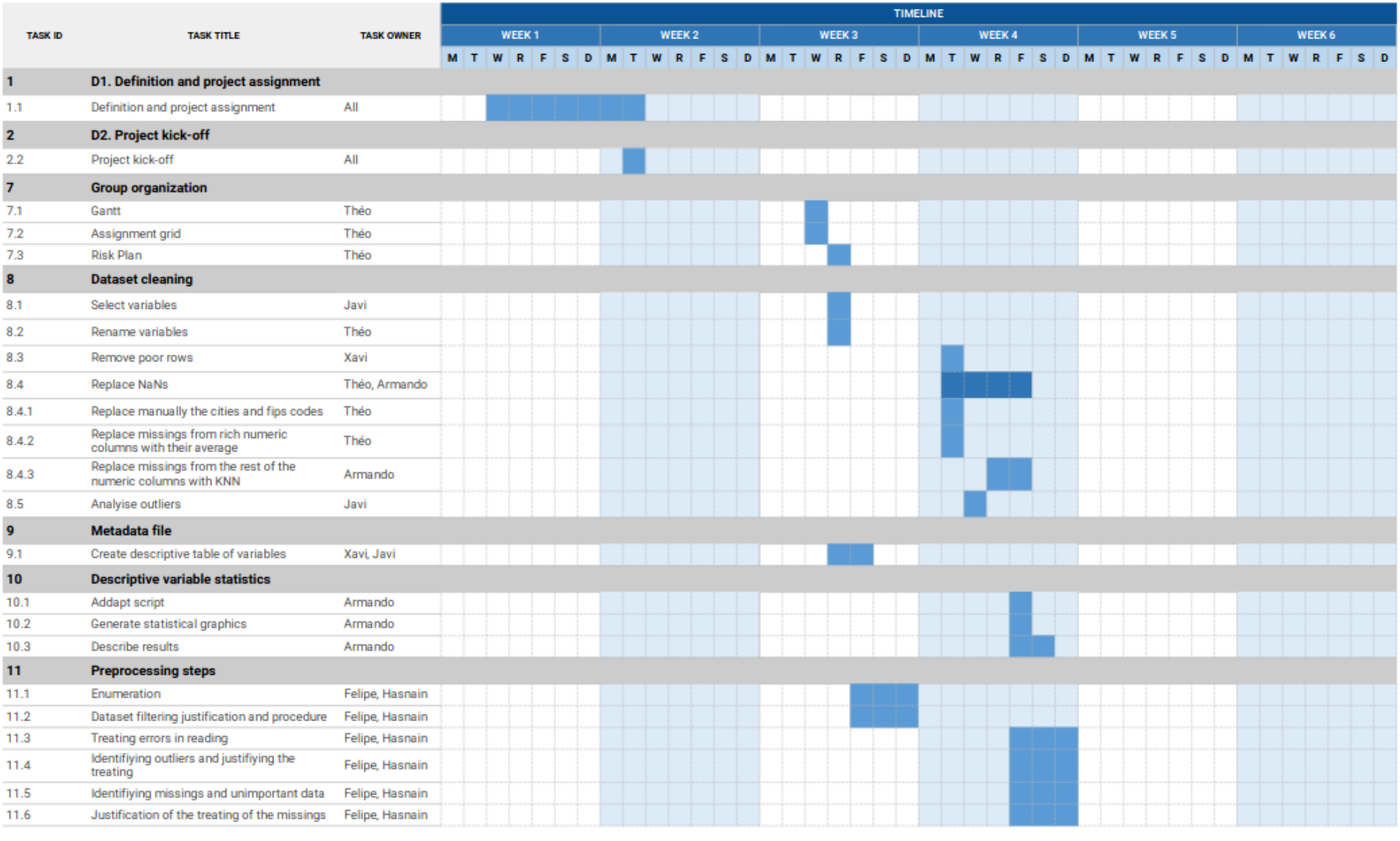
All of the information provided by the different project phases (mainly PCA, Clustering and profiling) came in harmony towards the end, considering they were all pointing towards a similar direction, they all reasonably complemented each other.

We can firmly report that this whole project helped us gain knowledge and experience in the datamining realm and was also very useful to improve our team skills as a whole, working with a bigger team and keeping track of all of the progress was challenging but rewarding.

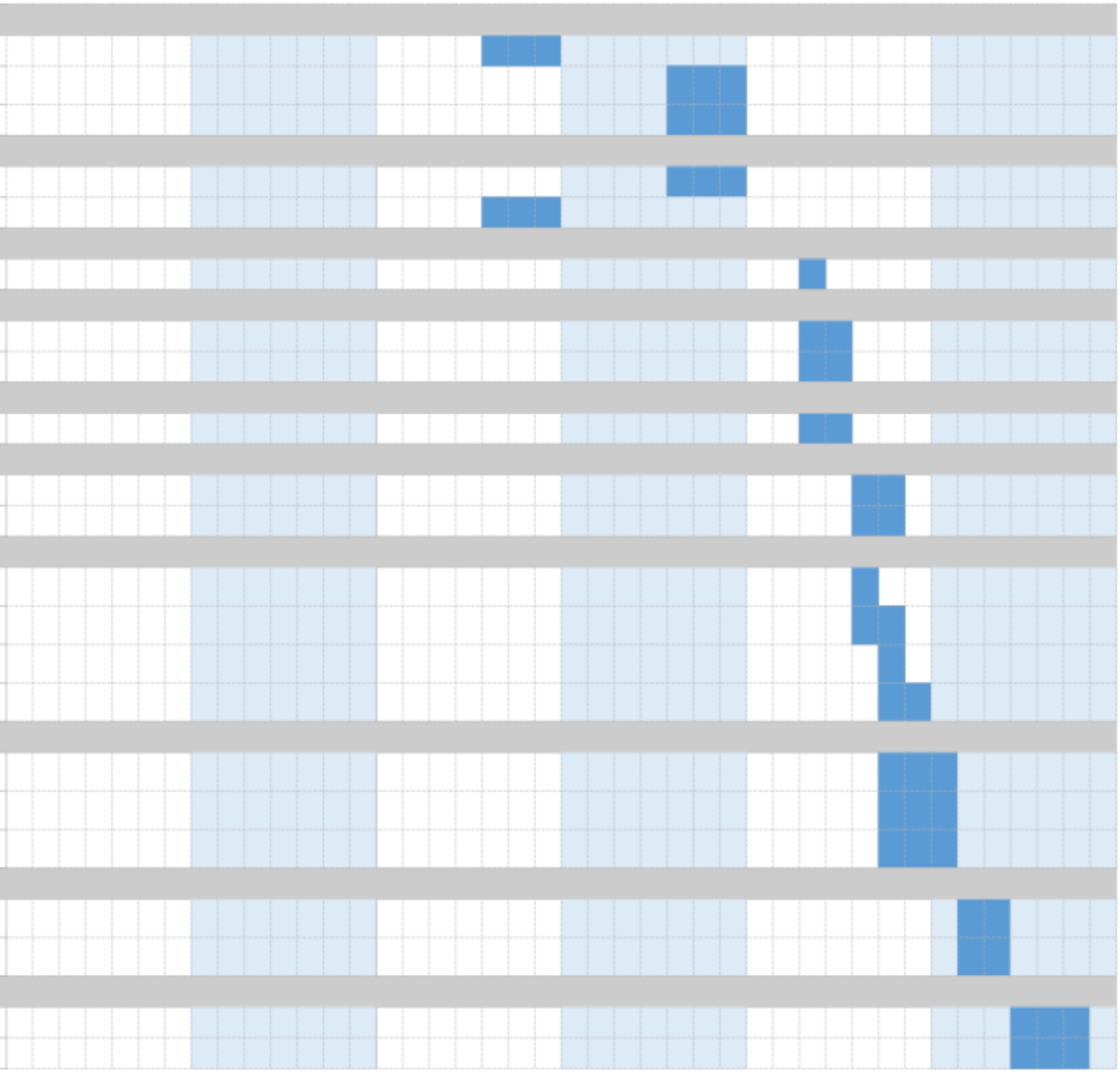
Working plan

Initial Gantt

TASK ID	TASK TITLE	TASK OWNER	TIMELINE
			WEEK 1 (M T W R F S D) · WEEK 2 (M T W R F S D) · WEEK 3 (M T W R F S D) · WEEK 4 (M T W R F S D) · WEEK 5 (M T W R F S D) · WEEK 6 (M T W R F S D)
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	



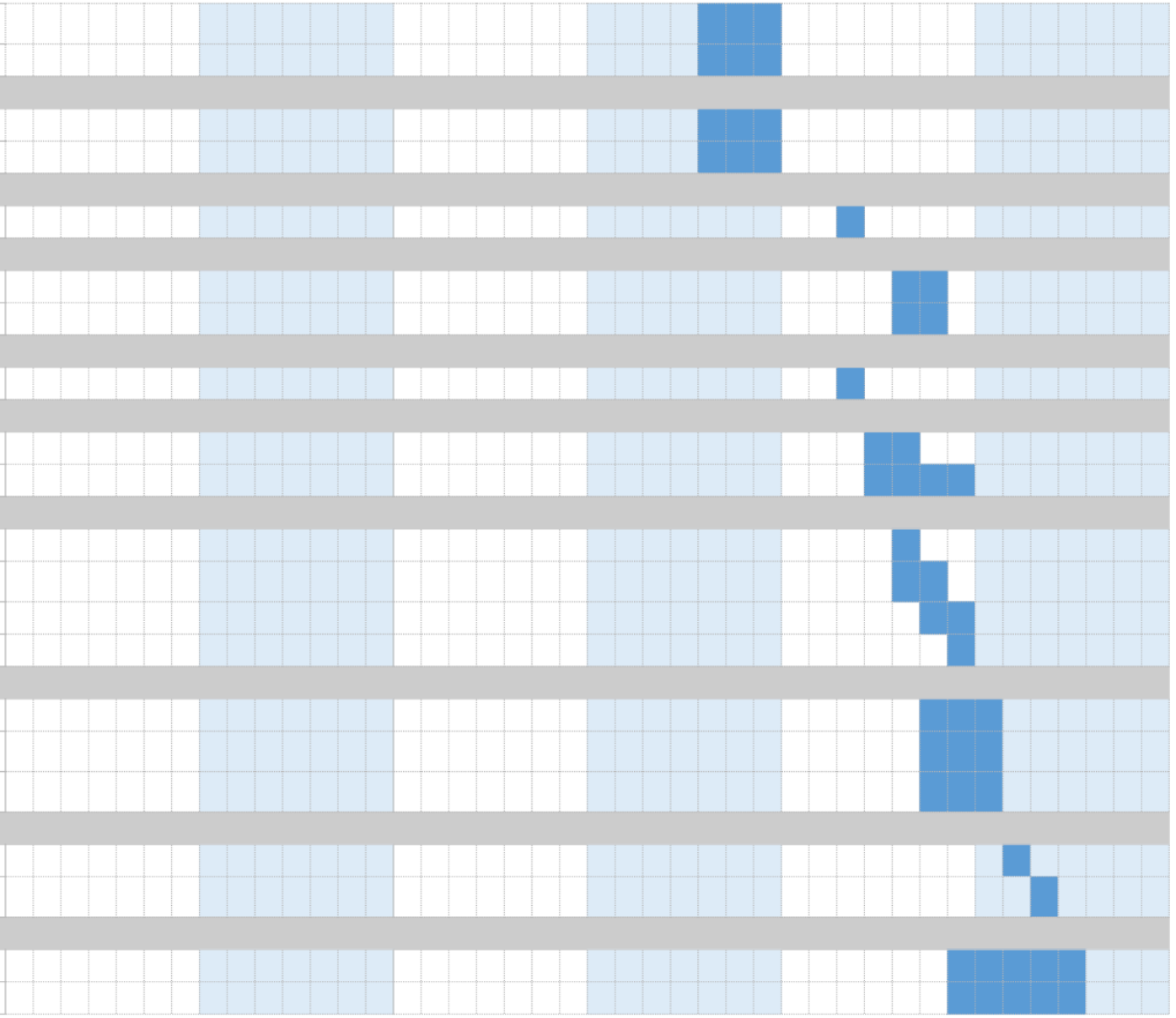
12	**Step decisions**		
12.1	Dataset filtering justification and procedure	Felipe, Hasnain	
12.2	Identifiying outliers and justifiying the treating	Felipe, Hasnain	
12.3	Justification of the treating of the missings	Felipe, Hasnain	
13	**Description of changes**		
13.1	New variables	Felipe, Hasnain	
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	



Final Gantt

TASK ID	TASK TITLE	TASK OWNER	TIMELINE
			WEEK 1: M T W R F S D · WEEK 2: M T W R F S D · WEEK 3: M T W R F S D · WEEK 4: M T W R F S D · WEEK 5: M T W R F S D · WEEK 6: M T W R F S D
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	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	
11.7	Derivation of new variables from others	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	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



Assignment grid

	Felipe	Théo	Xavier	Armando	Javier	Hasnain
Group organization		**X**				
Dataset cleaning		X	**X**		X	
Metadata File			X		**X**	
Descriptive variable statistics				**X**		
Preprocessing steps	X			X		**X**
Step decisions	**X**					X
Description of new/modified variables	**X**					X
Motivation of the work and general description of the problem to be analyzed				X	**X**	
Complete Data Mining process performed		X	**X**			
Final scope of the study		X	**X**			
PCA analysis for numerical variables	X		X		**X**	
Hierarchical Clustering on original				X	**X**	

data						
Profiling of clusters	X					**X**
Global discussion and general conclusions of the whole work		**X**	X			
PPT preparation	X	X	**X**	X	X	X
Coordinator				**X**		

Risk plan

Risk	Prevention	Management
A team member leaves the course	Everybody is aware of all tasks	Even task redistribution
A new quarantine begins	This cannot be prevented	Creation of a discord server to keep up with telematic work
The dataset has too many missings and isn't usable	Backup dataset	Exploit as much as possible the old dataset treatment methods
A substitution method for NaNs doesn't perform correctly on the analysis	List of other substitution methods that also suit the variables	Implementation of another substitution method that also makes sense
A team member is failing to reach a deadline	The gantt chart is made so that there's always another team member available to help	The available team member(s) help the struggling one to finish the task before the deadline
A team member has failed to reach a deadline	The gantt chart is made to prevent leaving a task for the last day	Other members help to finish the task within a day or more (minimum time)

A team member is underperforming	Everybody is aware of the overall progress of the project	The remaining members help the underperforming member to get back on track

Critical discussion

After finishing the project, we have concluded that we have overall respected the assigned tasks for each member and most of the tasks have been completed in their respective scheduled frames. Few tasks have been done earlier/later than their original set time frame, this has been changed in order to keep the workflow as fluid as possible.

After the initial gantt distribution we changed the members assigned in some of the remaining tasks in order to work properly, after the previous delivery we had a wider vision on how every member worked and we decided to take advantage of each other's virtues.

The tasks that have turned out to be more time consuming than expected have been managed correctly, as we have included more members in their due (as originally planned in the risk plan), these tasks such as the PCA analysis, have taken one or two more days to complete than the originally accorded time but none of it have corrupted our workflow in any harmful way.

Thankfully no unexpected risks appeared in the project, which enabled us to develop it in a completely stable way.

Scripts

The scripts used, which have been done with Python and R, are attached in the folder scripts of the project.