Hospital Data

Max van Esso Castellet - 73539

8 de abril de 2016

This report attemps to be a data visualization exercise on a data set containing information about a Hospital in the city of Barcelona.

The data set is structured in a matrix of 43 variables and 39.265 observations.

The variables are the following:

Patient Data: Age, Sex, City, Region, Country the patient lives, Zip Code, Basic Health Area, Country the patient was borned, Medical Insurance.

General Medical Data: Main Diagnostic Group, Other Diagnostic Group, Comorbidities number, CondClinEspec (The patient has specific clinical conditions like severe COPD, diabetes, heart failure, loss of weight, depression, asthma, anemia (yes/no), MentalDisorder (yes/no), AlcoholDrugs (The patient is addicted to alcohol or other drugs (yes/no)), Neoplasia, Retardation, Respiratory disease, Diabetes, Cardiac Deficiency, Loss of weight, Depression, Anemia, Complex Cronical Patient (yes/no), Advanced cronical disease (yes/no).

Admission Data: Date of Admission, Time of admission, Day of the week, Date of discharge, Time of discharge, Day of the week of the discharge, Length of stay.

Admission Clinical Data: Medical Specialty, Oncology service (yes/no), Origin of the patient (home, emergency. . .), Number of procedures during admission, Number of laboratory procedures during admission.

Usage Data Before Admissions: Number of admissions during last year, Number of emergency visits during last year, Number of visits during last year.

Usage Data After Discharge: Number of days since discharge until readmission, Readmission type1, Readmission type2.

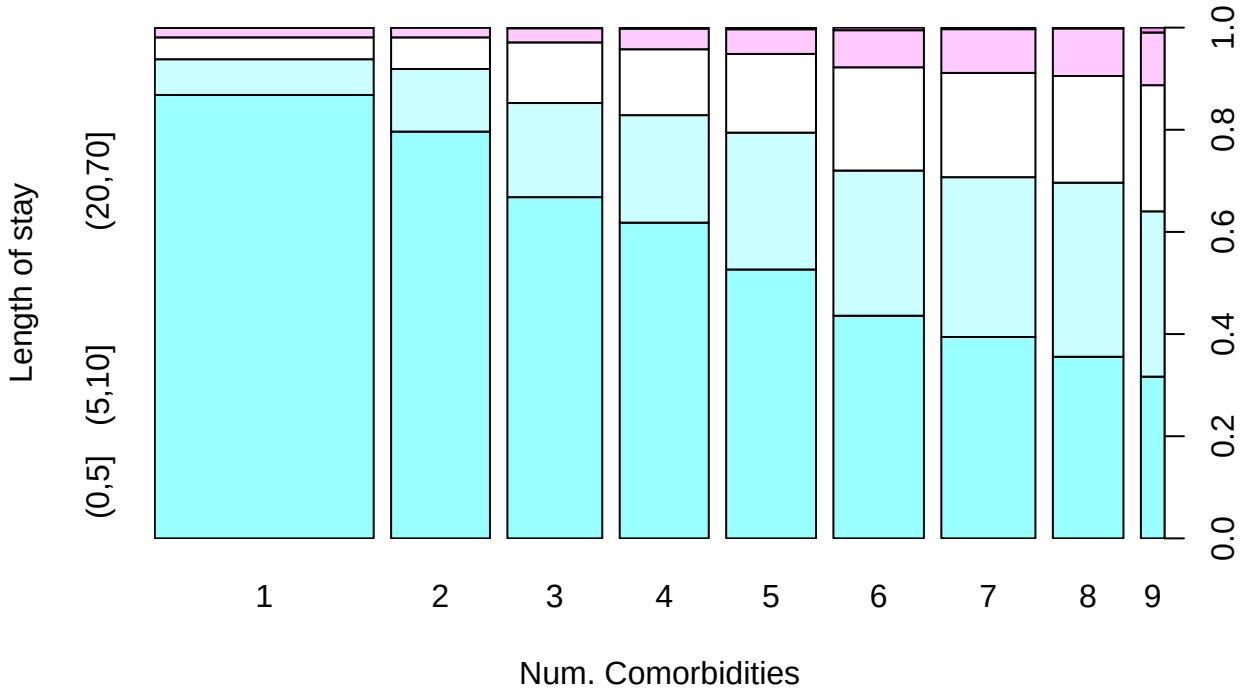
After cleanning the data of *NA* and missing values I ended up with 39.080 observations. The dataset is divided into categorical, binary and discrete variables, including also some time variables such as the exact time of admission or discharge that I choose not to use in the visualization due to the lack of interes and predictability power in the study of the relation between the variables and the 3 responses (Length of stay, Type1 Readmission and Type2 Readmission).

In order to visualize the data and fully grasp the information that the variables have to offer about the profile of the patients and the responses, we divided the analysis in 3 parts:

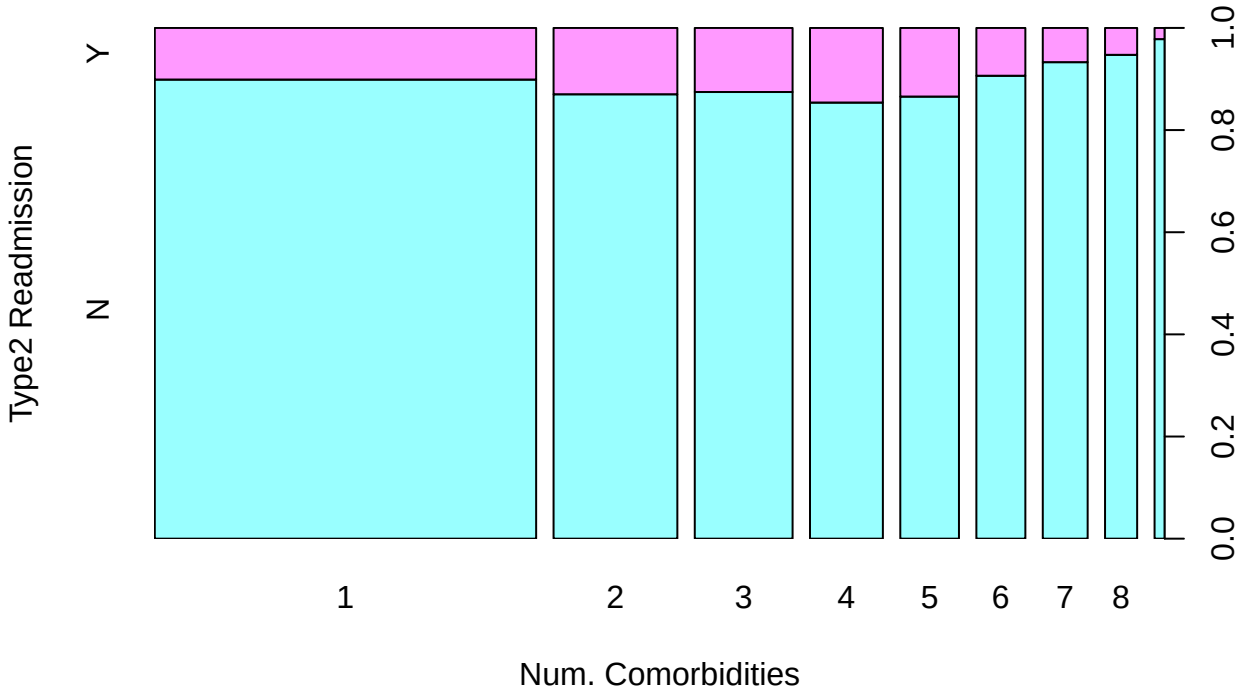
- A first part to show the relationship between some variables and the response variables.
- A second part to get an idea of the profile of the user via plotting how are they distributed by location, insurance company, diagnostic group etc.
- A third part to see the relation between the variables.

1.- Variables and responses

The first 3 plots show how the Number of Comorbidities (NoC) are related to either the Length of Stay of the patient in the hospital or the Type of Readmissions.

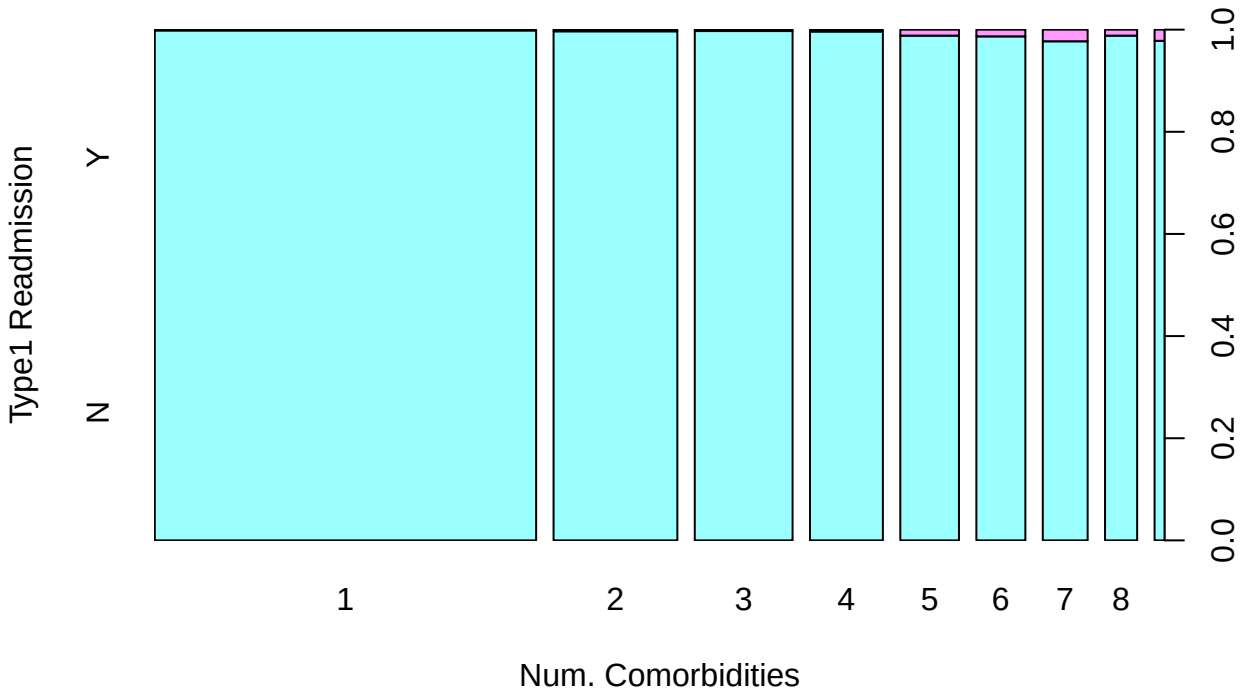


On this plot we can see a clear relation between the NoC and the difference in length of stay. As the NoC increases, the percentage of patients with longer stays increases significantly, also notice how the percentage of patients staying between 0 and 5 days (the shortest interval that we have divided the length of stays with) decreases from 80% when NoC = 1 to aproximately a 30% when NoC = 9.



The Type2 Readmission (T2R) means that the patient returned to the hospital within 30 days from the

discharge. In this case we can't observe a positive relation between the NoC and the T2R, if anything it seems like there is a negative relationship.

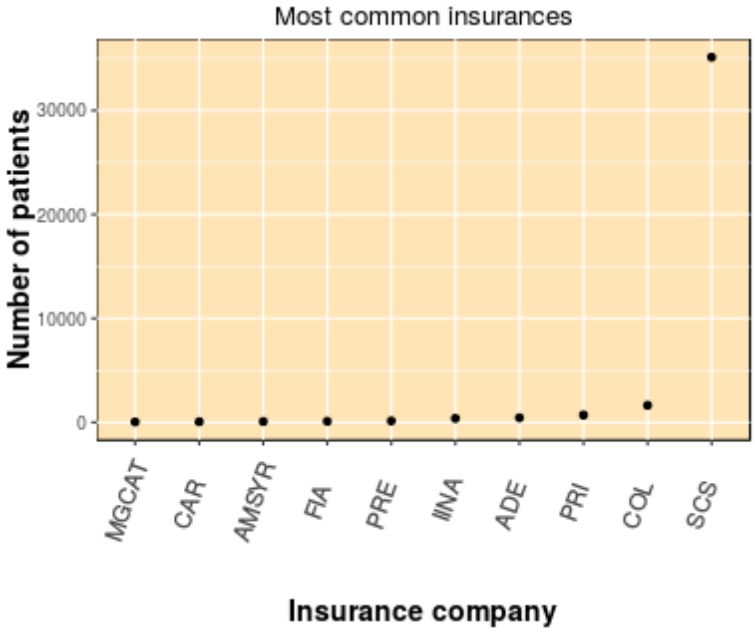


The Type1 Readmission (T1R) means that the patient returned to the hospital within 30 days from the discharge for the same reason he/she was originally admited or a disease of the same "family". In this case we can, as opposed to the T2R, observe a positive relationship between NoC and T1R.

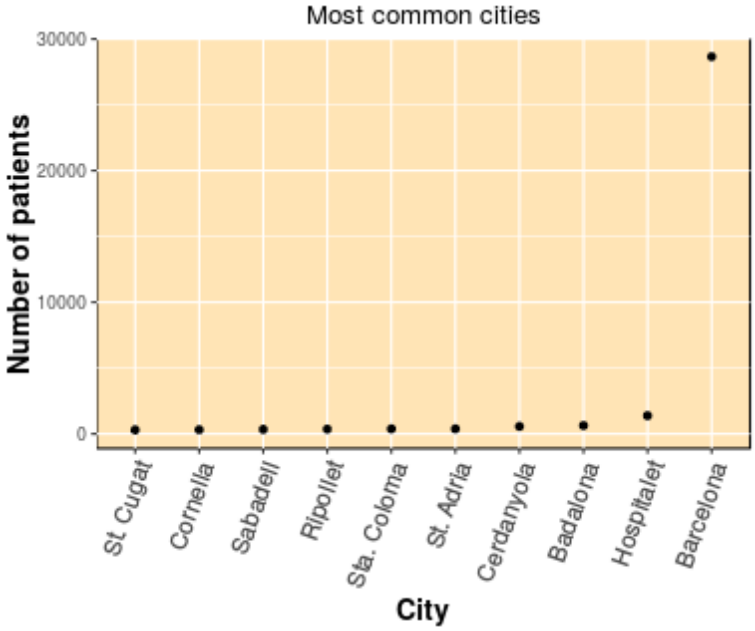
It will be left for the statistical and medical analysis of the data to try to assess why this inverse relationship happens between NoC-T2R and NoC-T1R.

2.- User characteristics and clustering

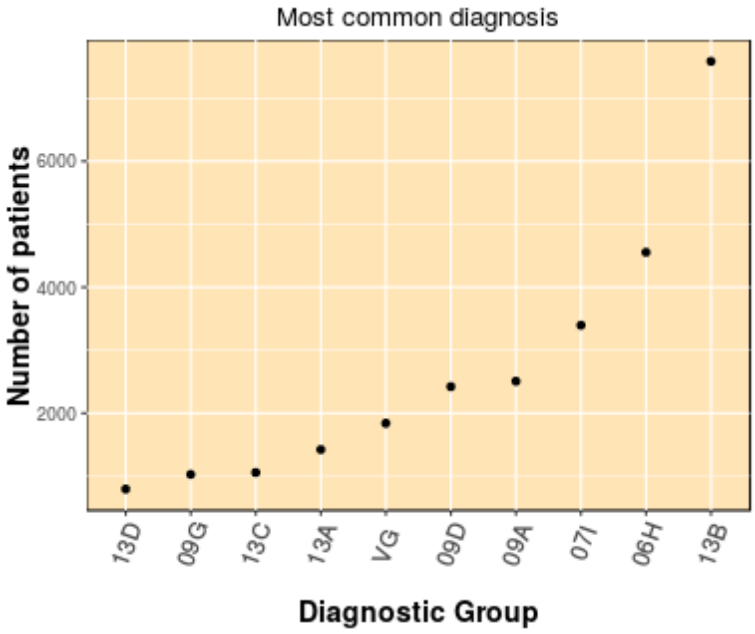
The following 5 plots visualize how the patients we have data for are grouped according to some of the more interesting categorical variables.



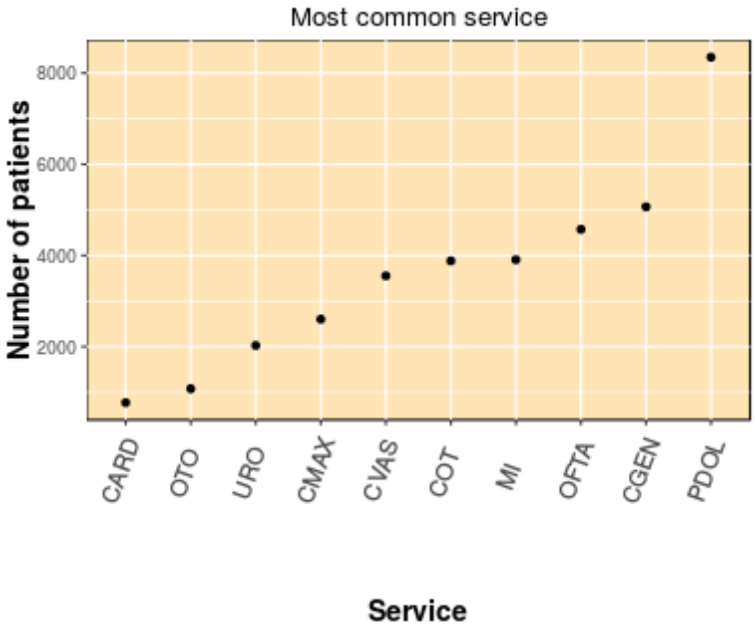
Out of 63 different insurance companies we can see how basically SCS is the one that captures most of the patients, remember that the total number of observations is 39.080 and SCS has more than 35.000 patients. This could be due to the various reasons, for example that most of the doctors of the hospital have better deals for the patients of that insurance company, or that most of the patients live in the same area and therefore they receive more advertising from that insurance policy.



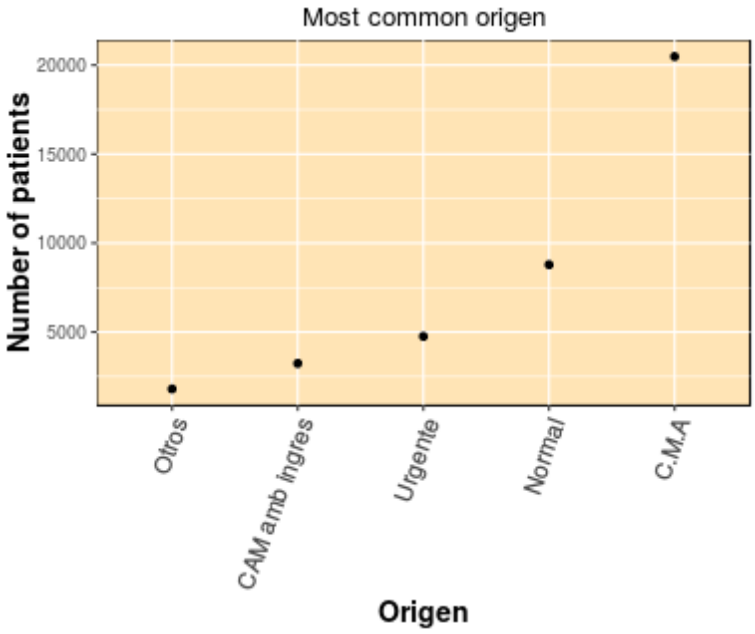
Out of 406 different cities there is a clear winner, Barcelona. If we look at the top 6, they are all cities in the metropolitan area of Barcelona (except for Cerdanyola, that anyway it is not far away), this is reasonable since the data comes from a hospital in Barcelona and therefore the distance is an important factor.



Out of 105, the top 10 diagnosis cover roughly 70% of the total. It is a fair and expected result because some diagnosis are easier than others and because rare diagnosis are associated with rare diseases which are not that recurrent.



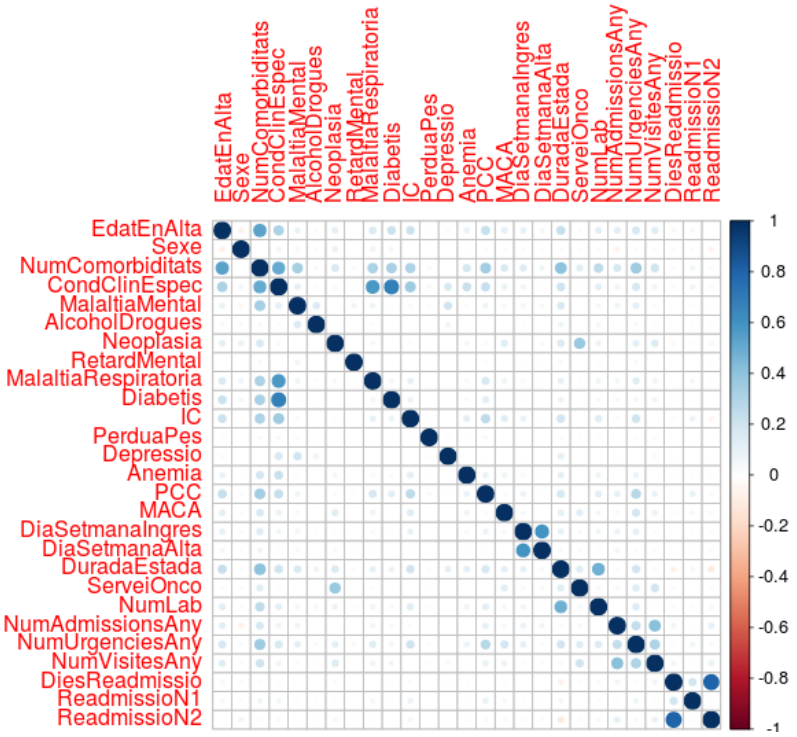
Out of 23 different services we can see that the top 8 cover more than 85% of the total number of patients. Podology is the most used service, followed by Gyneacology and Oftalmology, which make sense because there is a big number of patients that go to the hospital for "normal" aches.



This is the only case where we can plot all the different types of origens (at least for the classification the hospital works with). We can notice how the first cause of admission is C.M.A (Major Surgery) followed by Normal and Emergencies.

3.- Variable correlation

This correlogram shows how little the variables are correlated with each other. Nevertheless we can see a few interesting correlations, for example:



- Between the T2R and the number of days since discharge until readmission there is a positive correlation. It is logical to see such a correlation if we think of the definition of T2R because this includes the possibility of being readmited for a different disease, and it is fair to think that those will apear at the end of the readmission window (the 30 days for which an admission is counted as readmission) much more frecuently that a relapse from the same disease.
- There is also a positive correlation between Diabetes and Respiratory Disease and Especial Clinical Condition which can be due to the fact that the first two are a subset of the third.
- Also is logical to interpret the positive correlation between the Length of Stay and whether the patient has been in the Oncologic Service or not.
- Lastly it is pretty straight forward to see why the NoC is positively correlated with Age, Special Clinical Condition and Length of Stay.

4.- Conclusions

To conclude I must say that:

- The data has a lot of information an useful variables to explore to help predicting the response variables.
- The variables themselfs are not particularly highly correlated although there are some particular correlations that help to understand the synergies of the data.
- The user information of the categorical variables is clustered in a particular way. For some variables there is a huge difference between the first and the second leading category (meaning there is a category that encapsulates most of the observations) and even with the first top 5-10, out of 20 to 400 different possible categories, it is possible to capture more than 80% to 95% of the information. That means that even if the total number of different profiles is huge, the majority of the users share the same characteristics, which obviously will have to be taken into account in order to analitically try to explain and predict the responses.