ML project

Pietro

2/24/2022

Two aricles where these data come from:

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5#Tab1

Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA (2017) Survival analysis of heart failure patients: A case study. PLoS ONE 12(7): e0181001. https://doi.org/10.1371/journal.pone.0181001

These are the data we have about the survavial from

Gender, Smoking, Diabetes, BP and Anaemia are factors that have been saved as numerical values. To start, we decided to transform them into categorical variables

DESCRIPTION OF THE DATA

... IMPORTANT: explain difference btw usual data with only classification of the event and this problem, that requires survival models.

Before looking at the data, the split the data in one set for the train and one for the test. In this way, we will assure that the observations we will do and the decisions that we will make based on them will not be biased by the knowledge acquired from the test set. What follow is based on the training set.

EXPLORATIVE ANALYSIS

We started with the explorative analysis of the data. The aim of this first section is to reach a general understanding of the data we have, their distribution, the correlation between variables and in general all the aspects that concern the descriptive analysis of the observations. Only after a rigorous analysis we can start to model the data, since before that we will not have a complete comprehension of the data.

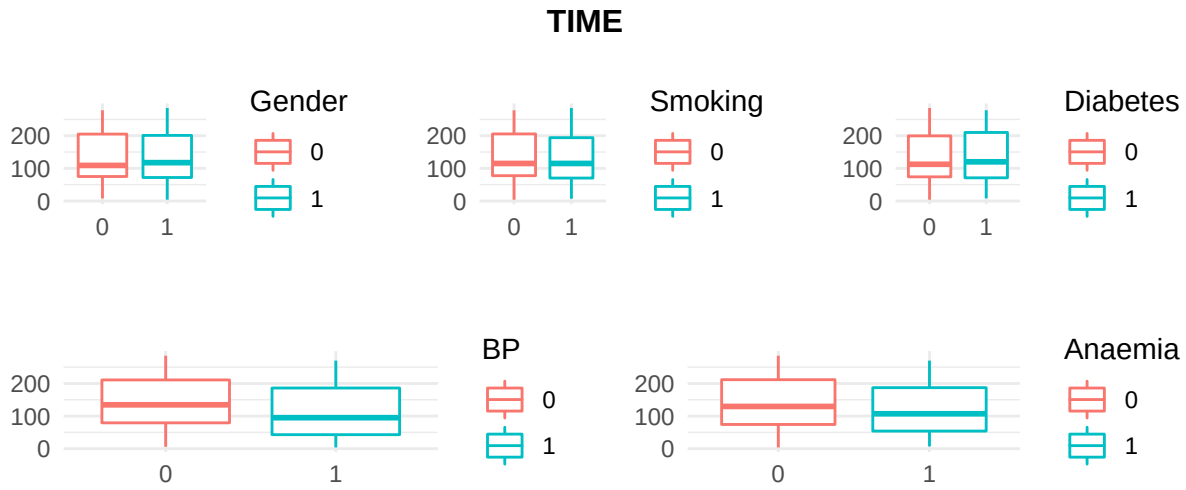
Pairs of continous variables

First, we are interested to see the joint and disjoint distribution of each variable. These are the distribution of each numerical variable and of the combination of all of them

The correlation is significantly different from zero only for the couples Creatinine-Age, Creatinine-Sodium and Sodium-Ejection.Fraction, and in any case is very low.

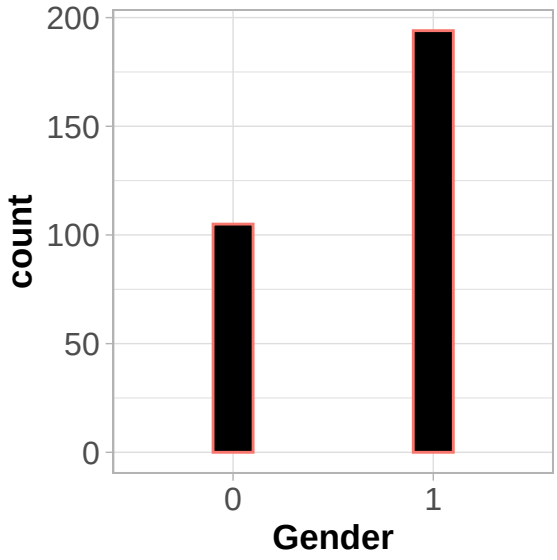
Time and factors

An important thing to see how the time under observation is related to the other variables. The following boxplots show that there is not relation between the time and the binomial variables. This result is interesting, since it tell us that the experiment that generated he data was unbiased by differences between the persons under observation. In other words, we can trust these data and use them to provide results that will not be biased by a lack of strictness during the data collection, since who collected them randomized the sample.



In addition, we can also have a look at the distributions of the persons inside the groups defined by different levels of the dichotomous variables. As shown in these histograms, they are basically even distributed.

We have to consider aside the boxplot about the time of observation for the group of people on which the event has been observed. Comparing this one with the distribution of the other group, it is clear that the time under observation for people that have presented an hart attack is lower than the the one of the ones that did not. As expected, this indicates that the heart attack looks to reduce the time under observation. It can seems obvious, but not having this result would have lead us to conclude that the data were not able to explain the relation between the explicative variables and the answer variable.



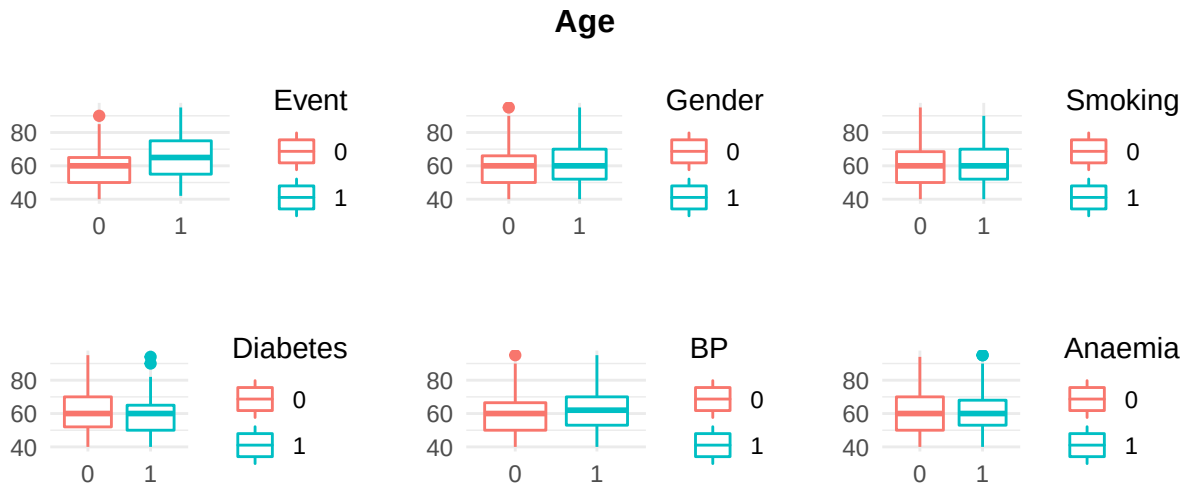
Countinous variables vs Event or cathegorical variables

Then, we looked at the relation between the pairs of cathegorical and continous variables, included the Event.

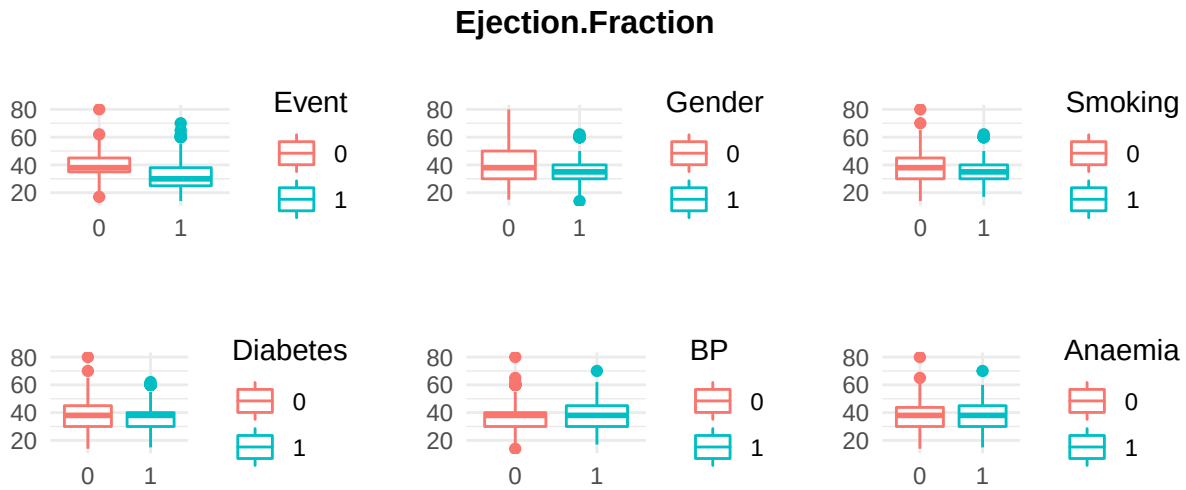
The scope of the following analysis is double:

a) to understand if the fact that the event of having an heart attack defines different distributions of the continuous variables

b) to understand if the categorical variables are related to the continuous ones

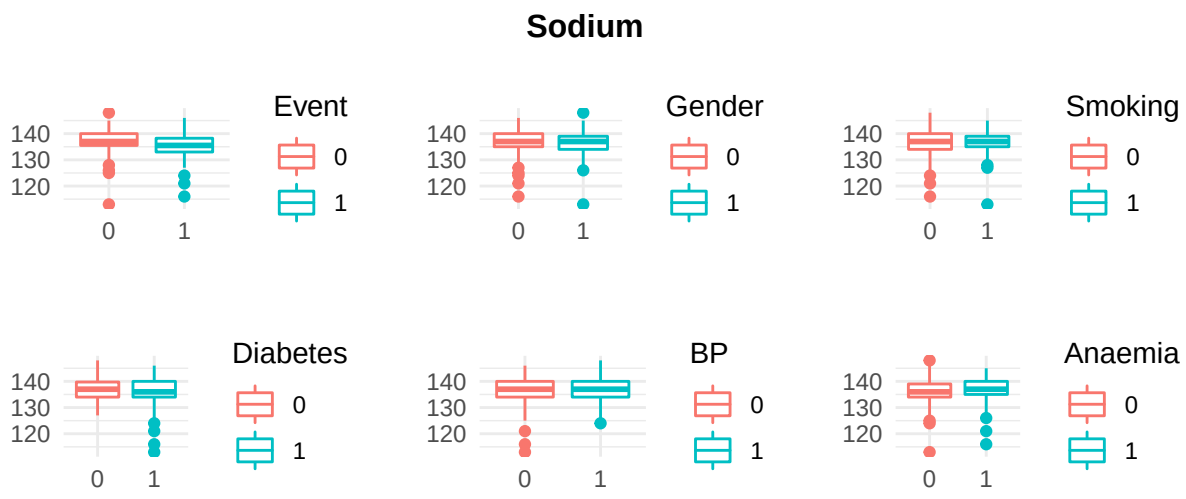


It seems that the population on which the event has occurred is older than the one that did not observed it. It is reasonable, since heart attack are generally more frequent in older persons. For all the other variables, there is basically not difference between the groups. Some points are over 1.5 times the interquartile distance and quartile, but they do not look to be to far from the median. There is not enough evidence to consider them as outliers, but it is of interest to start noticing them and keeping that in mind. The following analysis will tell us how to procede under this point of view. It should be noticed that, as said at the beginning, the youngest patients that partecipated at this research were fourty years old.

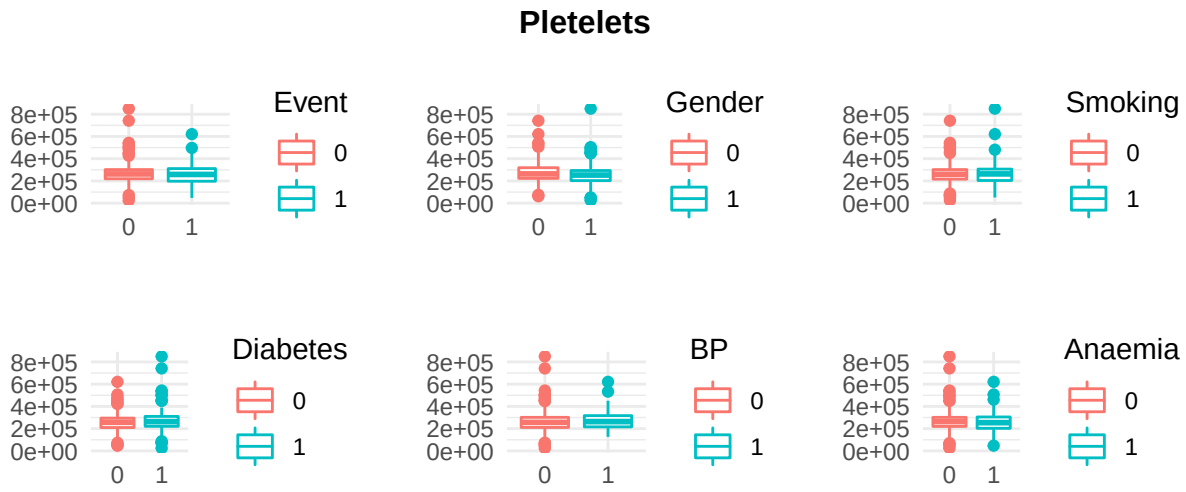


It seems that the population on which the event has occurred has lower values of ejection fraction. That means that this variable can be useful to describe the distribution of the vent over the population.

About the relation between this variable and the categorical ones, we can say something similar to what we said for the age. For men, the ejection fraction looks to be lower as for persons with diabetes. The remaining ones are not really related to this continous variable

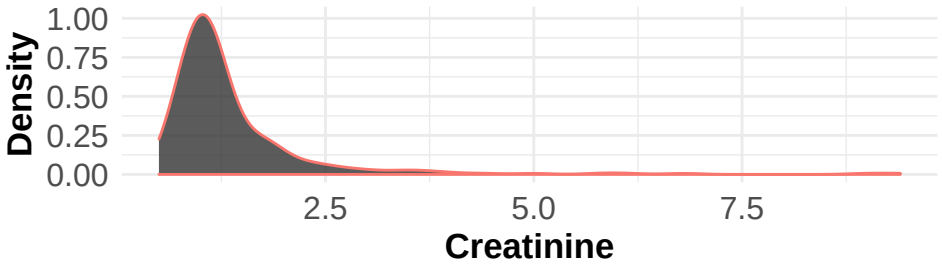


In this situation data look to be less regular. There are multiple points that are enough far from the median to allow us to say that a deeper analysis of outliers could be performed. In particular, some data presents an unusual low value of Sodium. Talking about differences between the distribution of the amount of sodium in the two groups, there is not difference to be noticed. On the other hand, this variable does not look very useful in term of explaining the distribution of the event, since the level of sodium is almost the same for both the groups of people having had an heart attack while being under observation and people who didn't.



The platelets' value has not distribution changes between the people having had the heart attack and the others. This suggest that this variable is not really useful for our scopes. Furthermore, we can notice that there is not difference in distribution between this continuous variable and all the others.

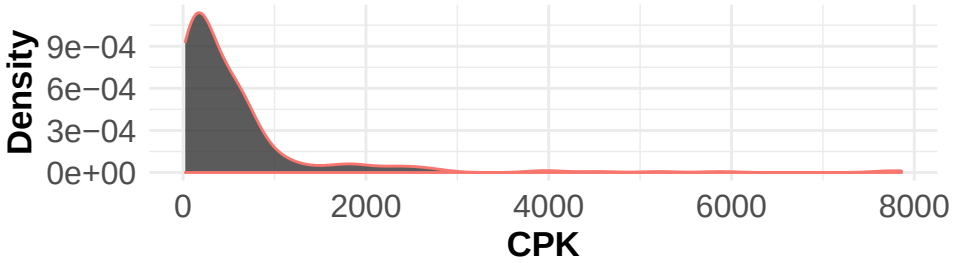
Before looking at how Creatinine variable is related to the categorical ones, we can look at the distribution of it:



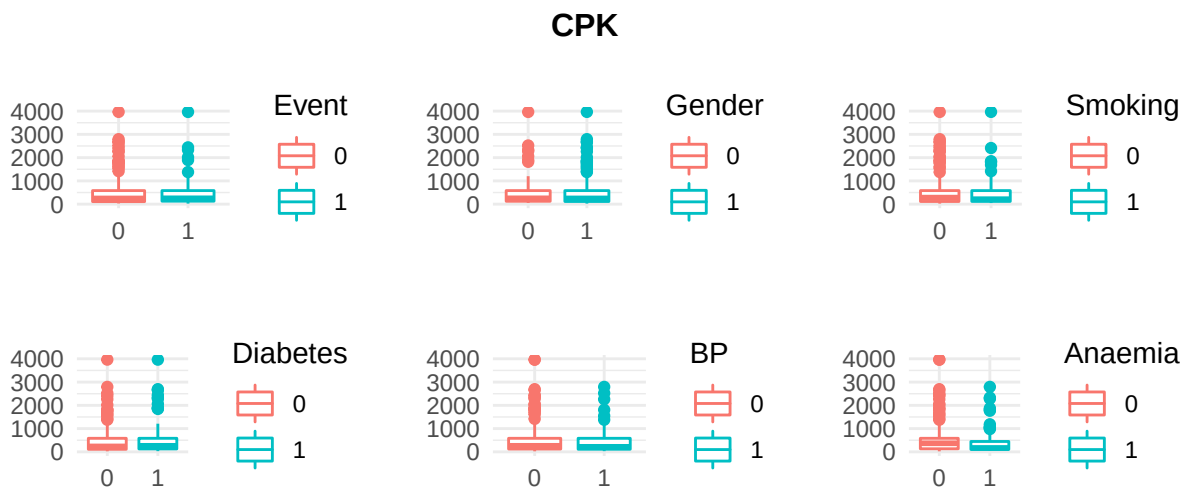
The distribution has a long right tail. In order to provide a useful visualization, we removed those very high value. More specifically, we removed from the visualization observation with a level of Creatinine higher than 5.

As saw before, the distribution of the Creatinine variable is uneven, even removing the larger values. Anyway, it seems that this variable can be considered between the ones useful to forcast the event, since people having observed the event during the observation time seems to have higher value of Creatinine. On the other hand, the other categorical variables do not present differences in distribution in the two groups identify by the two levels of each categorical variable.

Before analyzing the level of enzyme CPK, we had to do the same procedure we used with the Creatinine variable. In fact, its distribution is unbalanced:



This curve is very similar to the distribution of the Creatinine variable. This relation between the two of them can also been seen



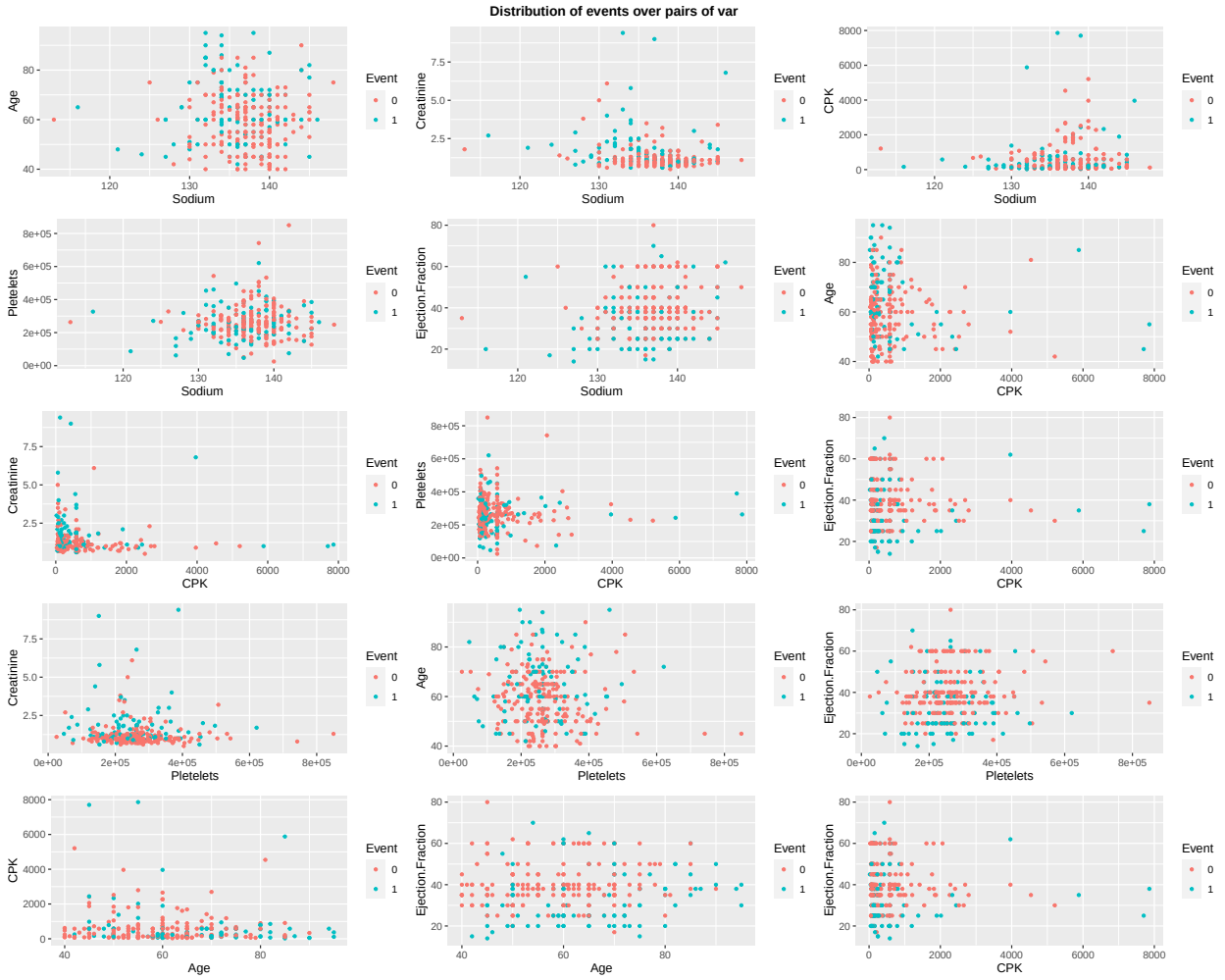
Even the distribution of CPK is unbalanced to the rigth, with a lot of big values. Unfortunatly, the CPK/Event boxplot shows that the former variable is probably not really useful to explain the latter one. The relation between the level of CPK and the other factors is the same of the Creatinine variable.

From this latter graphs, we can conclude that the relation between these two variables (Creatinine and CPK) is be very strong. The difference between the two of them is only in the effect that they have on the distribution of the events. While the Creatinine level looks to have an effect on the heart attacks, the CPK does not.

We should take into account this observation while performing variable selection.

Pairs of continuous variables and Event

The next step is to look for patterns between pairs of continuous variables and the fact that the event was observed or not



There are not natural clusters defined by these composition of variables.

Chategorical variables and Event

An other interesting thing to see is if the binomial variables have impact over the Event. To perform this check we compared dived in two groups, defined by the levels of the factors, the people that had an heart attack while being under observation. The results follow:

```
##       Gender
## Event         0         1
##   Yes 0.6761905 0.6804124
##   No  0.3238095 0.3195876

##       smoking
## Event         0         1
##   Yes 0.6748768 0.6875000
##   No  0.3251232 0.3125000

##       Diabetes
## Event         0         1
```

```
##   Yes 0.6781609 0.6800000
##   No  0.3218391 0.3200000
```

```
##      BP
## Event         0         1
##   Yes 0.7061856 0.6285714
##   No  0.2938144 0.3714286
```

```
##      Anaemia
## Event         0         1
##   Yes 0.7058824 0.6434109
##   No  0.2941176 0.3565891
```

The columns of each of these tables represent the frequencies of people that have or have not experienced the event. The results show that the distribution of this variable is not influenced by the levels of the other categorical ones.

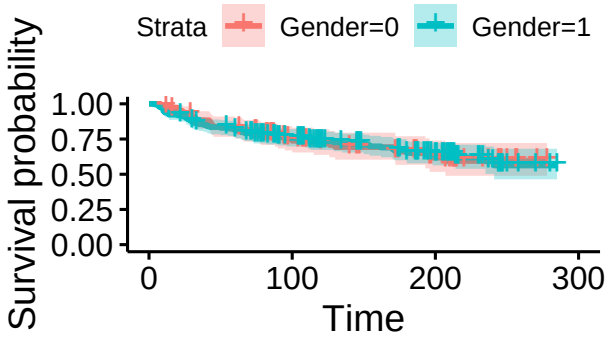
Conclusions of typical explorative analysis

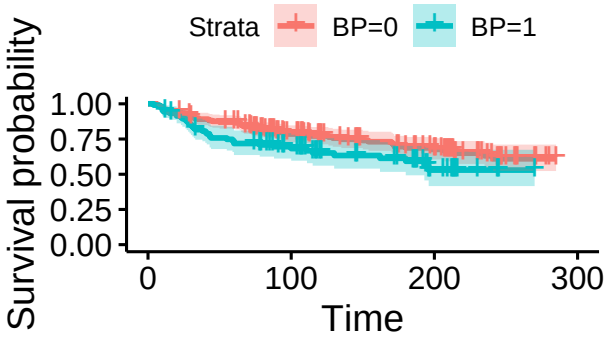
CHANGE THE WORLD TYPICAL!!!!!! To summarize the result of this explorative analysis, we can say that: * there is a strong relation between variables CPK and Creatinine; * the level of Creatinine looks to have an high impact; * the level of Sodium, the Age and the Ejection Fraction look to have a low impact on the Event variable; * the number of Pletelets in the blood and the level of CPK have no impact on the Event variable. * There are not natural clusters defined by pairs of countinous variables

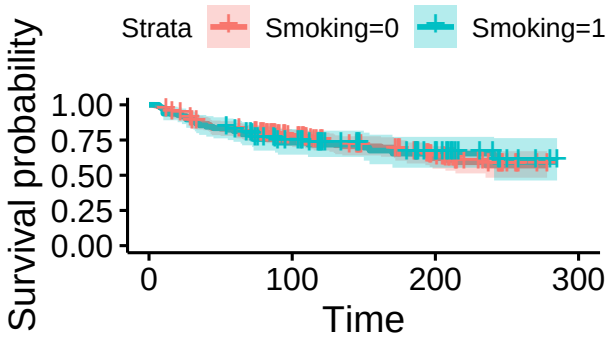
Survival models

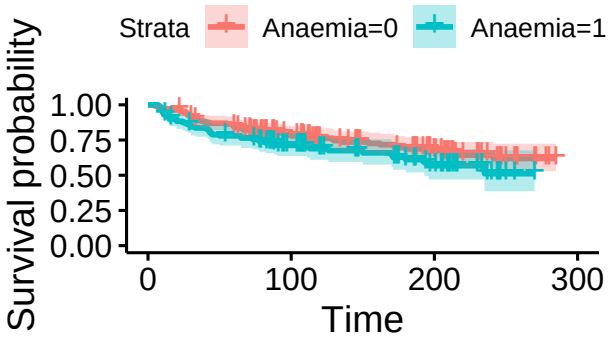
As said in the introduction to the data, this dataset has a specificity regarding the way we the variable Event has been collected. More precisely, each patient was under observation for a certain number of days (variable TIME). That means that the Event could be observed only under the period of observation of each person. This considerations leaded us to use the traditional approach to data of this kind, represented by survival models. In particular, we decided to use a combination of these common tools and of more modern approaches.

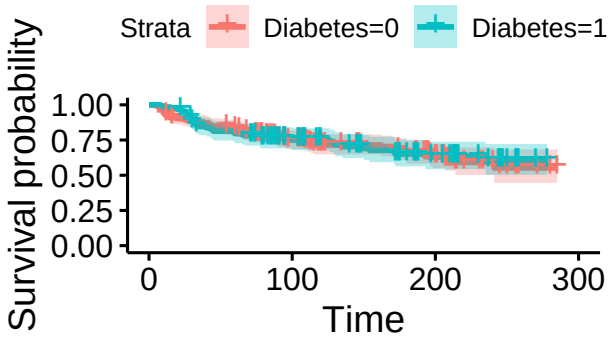
To perform our explorative analysis was then necessary to look also at the survival curves, in orther to compare them and decide which variables where really useful. From the previous section, we knew that we had to take care of the impact of Sodium, Age, Ejection Fraction, Pletelets and CPK











Each pair of survival curves defined by different levels of the categorical variables have confidence intervals that overlaps, as it can be seen throw the fact that the red area and th e light blue one always overlap. This means that, from a descriptive point of view, these variables do not have effects on the answer.

We can perform a general test to see if the combination of all these variables, together, has significant effect on the survival curve.

Before, we saw the comparison between survival curves of groups of people individuated by each of the factors, taken individually. From those graphs, we can notice that the lines generally does not intersect. Then, we can reasonably use a logarithmic rank test for the hypothesis that all the survival curves defined by all the combination of the possible values of the binomial variables are equal. The overall p-value of this test is 0.0137005. That means that, even if individually the variables does not look to be very useful, their combination can be used to explain the behavior of the heart attacks among the population of interest.

To conclude, we can say that: * every the categorical variable does not seem to have impact on the event if individually considered, * the combination of all the categorical variables seems to have impact on the event.

These results imply that, while doing variables selection and, in general, building our models, we will have to pay attention in finding the set of variables which joints effect is explicative, even if the single ones are

not.

FEATURES SELECTION

CPK and Creatinine

In the previous analysis we saw that CPK and Creatinine are highly related. On the other hand, the correlation between the two is very low: $Cor_{Creatinine,CPK} =$
-0.016

We applied a generalized linear model to see if the variables were useful if jointly used to explain the behavior of the Event variable. To look at this kind of linear dependence we have generated two models:

- $logit(Event) = \beta_0 + \beta_1 CPK + \beta_2 Creatinine + \epsilon$
- $logit(Event) = \beta_0 + \beta_1 CPK + \epsilon$

The significance of the parameters of the first model is, respectively: 0.18, 0. It can be concluded that CPK can be dropped from this model. In our context, it means that CPK is not useful in inferring the Event of interest if combined with Creatinine.

Before definitely discharging CPK as a variable for the classification of the level of Event, we performed two other tests to see even over the linear dependency. In order to do that, we built the following model: $logit(Event) = \beta_0 + \beta_1 CPK + \beta_2 CPK^2 + \beta_3 CPK^3 + \epsilon$
. All these coefficients where not significantly different from 0. In fact, their p-values are respectively: 0.55, 0.32, 0.24.

We then ddecided to discard the CPK variable.