The MONICA (Multinational MONItoring of trends and determinants in CArdiovascular disease) Analysis

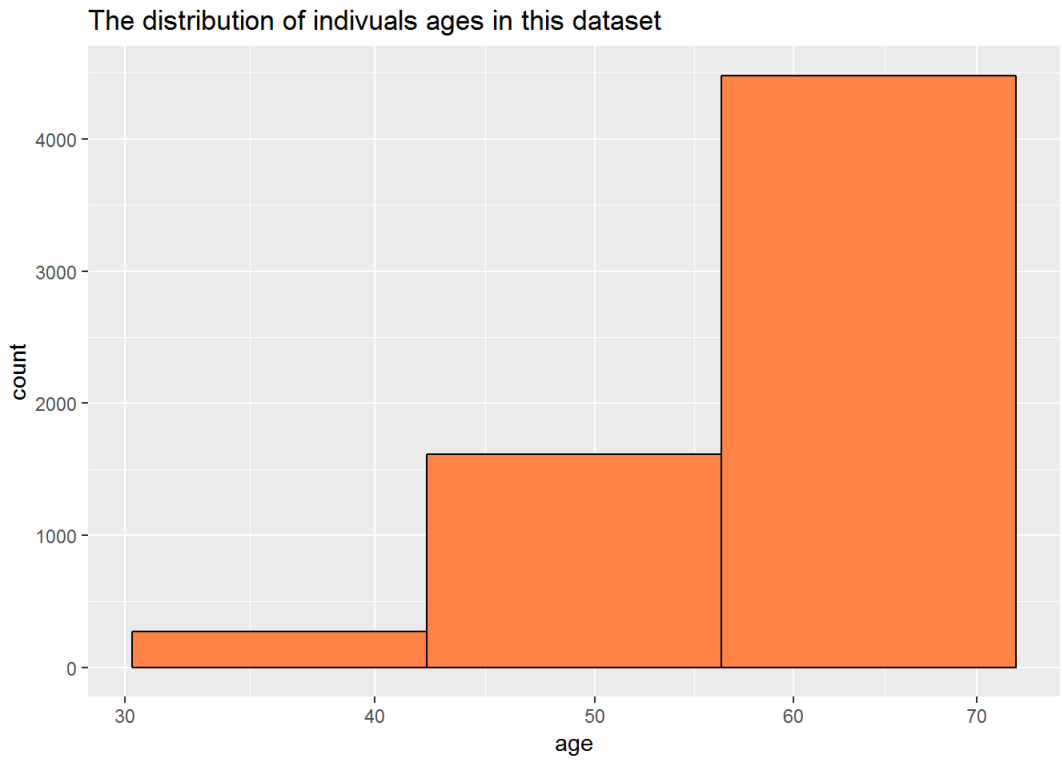
By Soumia Zohra El Mestari

The monica dataset (Multinational MONItoring of trends and determinants in CArdiovascular disease) is a WHO (World Health Organization) dataset, the main purpose behind this project was to monitor trends in cardiovascular diseases, and to relate these to risk factor changes in the population over a ten year period. It was set up to explain the diverse trends in cardiovascular disease mortality which were observed from the 1970s onwards. The dataset contains 6357 rows and 12 following variables :
outcome :mortality outcome, a factor with levels live, dead
age : age at onset
sex : m = male, f = female
hosp : y = hospitalized, n = not hospitalized
yronset :year of onset
premi :previous myocardial infarction event, a factor with levels y, n, nk not known.
smstat :smoking status, a factor with levels c current, x ex-smoker, n non-smoker, nk not known **diabetes** :a factor with levels y, n, nk (not known)
highbp : high blood pressure, a factor with levels y, n, nk (not known)
hichol : high cholesterol, a factor with levels y, n nk (not known)
angina : a factor with levels y, n, nk angina is described by the American Heart Association as, "…chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood"
stroke : is the patient had or hadn't a stroke a factor with levels y, n, nk (not known)

Univariate Plots Section

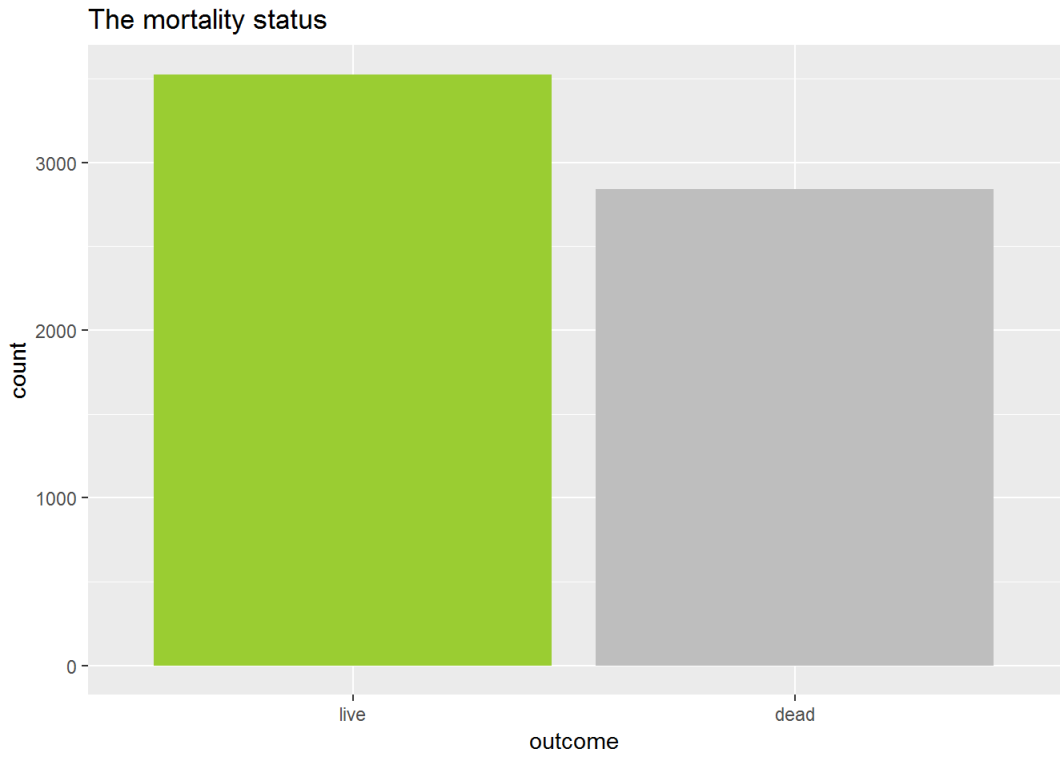
An overview about the variables

1st Exploration : Let's explore the ages of people in this dataset



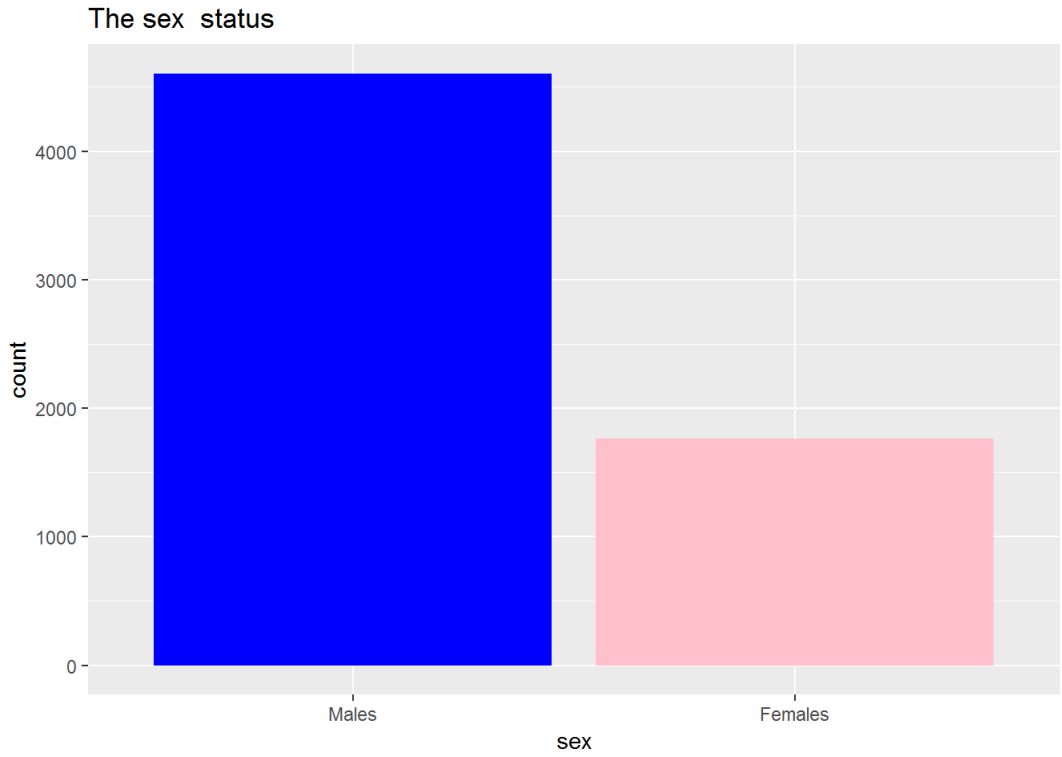
> People in this dataset are all adults ranging from 35 to 69 years old ,with the majority of them beyond 60 years old , (statistically the distribution is negatively skewed)

2nd Exploration let's explore the number of the mortality status



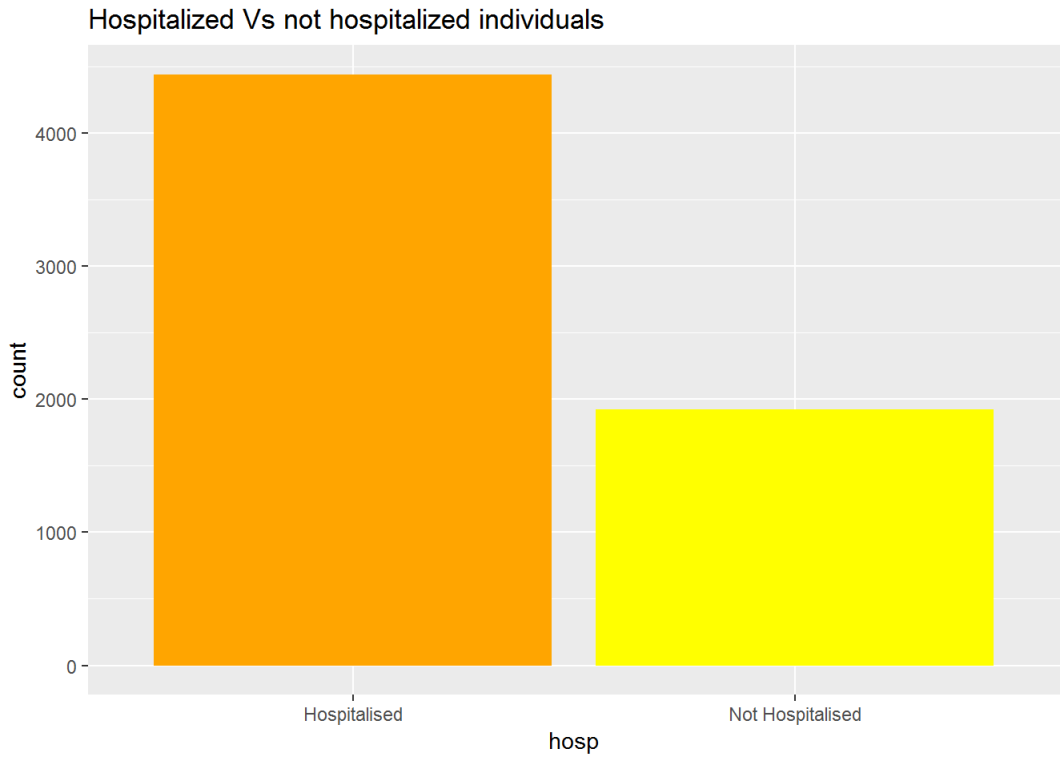
> The number of death cases is less than the live cases .

3rd exploration : sex distribution in our dataset



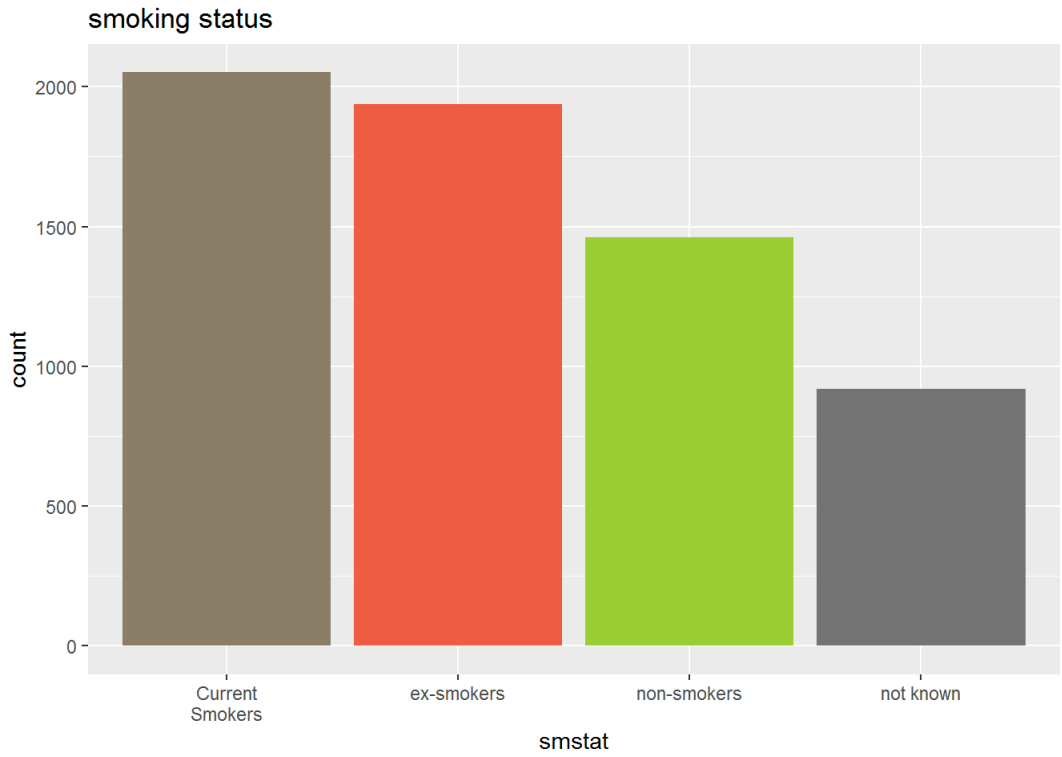
> It's to note that we have more males than females in this dataset so maybe the sex can influence the outcome (we will explore this possibility latter

4th exploration : hospitalized vs not hospitalized cases .



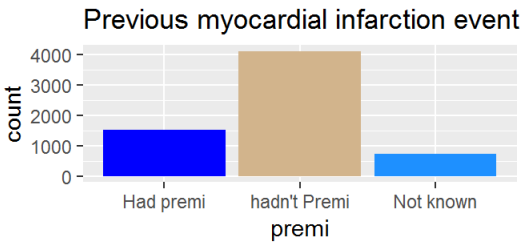
> Most people were hospitalized in this dataset

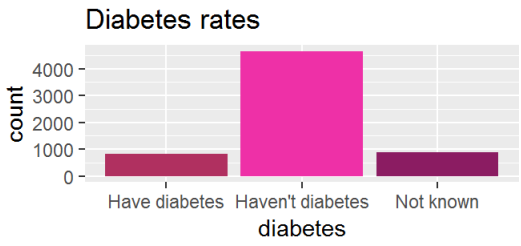
5th variable : smoking status

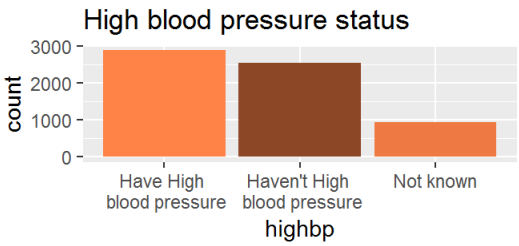


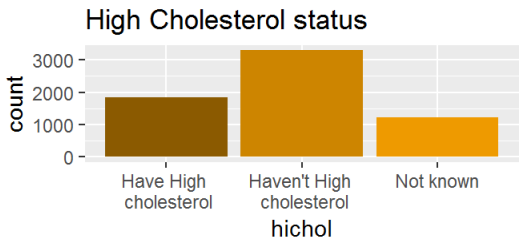
> The most people here had a smoking experience either a pervious or a current one.

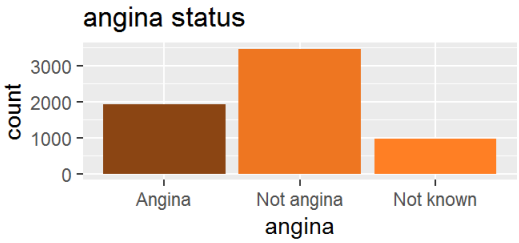
6th exploring medical status : In the Following I will explore the ratios of diffrent diseases and events : previous myocardial infarction event , diabetes , high blood pressure , high cholesterol , angina , and the stroke rates .

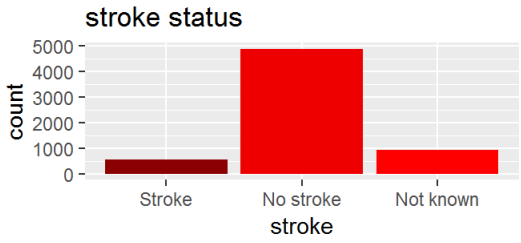












> Generally speaking we have a high rate of high blood pressure in our population while diabetes and stroke rates are low .

Univariate Analysis

What is the structure of your dataset?

The dataset contains 6357 records or cases with 12 variables (outcome,age,sex,hosp,yronset,premi,smstat,highbp,hichol,angina,stroke)
* The data collection was from 1989 to 1993
* All the idividuals were adults and the most of them were beyond 60 years old.
* The number of males is roughly 4 times the number of females in this dataset.
* Most of our cases had a smoking experience either previous or current
* nearly 70% of the individuals were hospitalized

What is/are the main feature(s) of interest in your dataset?

The main features in this dataset are : the mortality rate (outcome) , age . I want to determine the best features in predicting the life expectancy for cardiovascular patients.

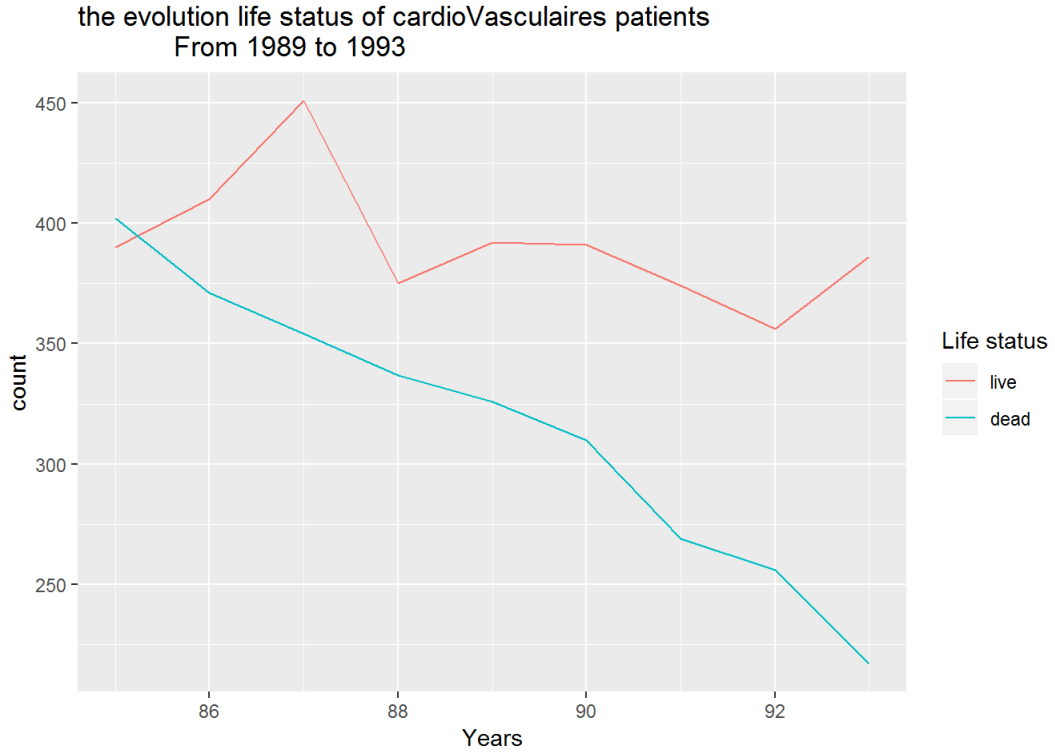
What other features in the dataset do you think will help support your
All the other features (stroke , angina ……) may influence the life expectancy.

Did you create any new variables from existing variables in the dataset?

No I haven't created any new variables
Of the features you investigated, were there any unusual distributions? I scaled the age with sqrt to help determine and emphazise the most frequent age of these individuals.

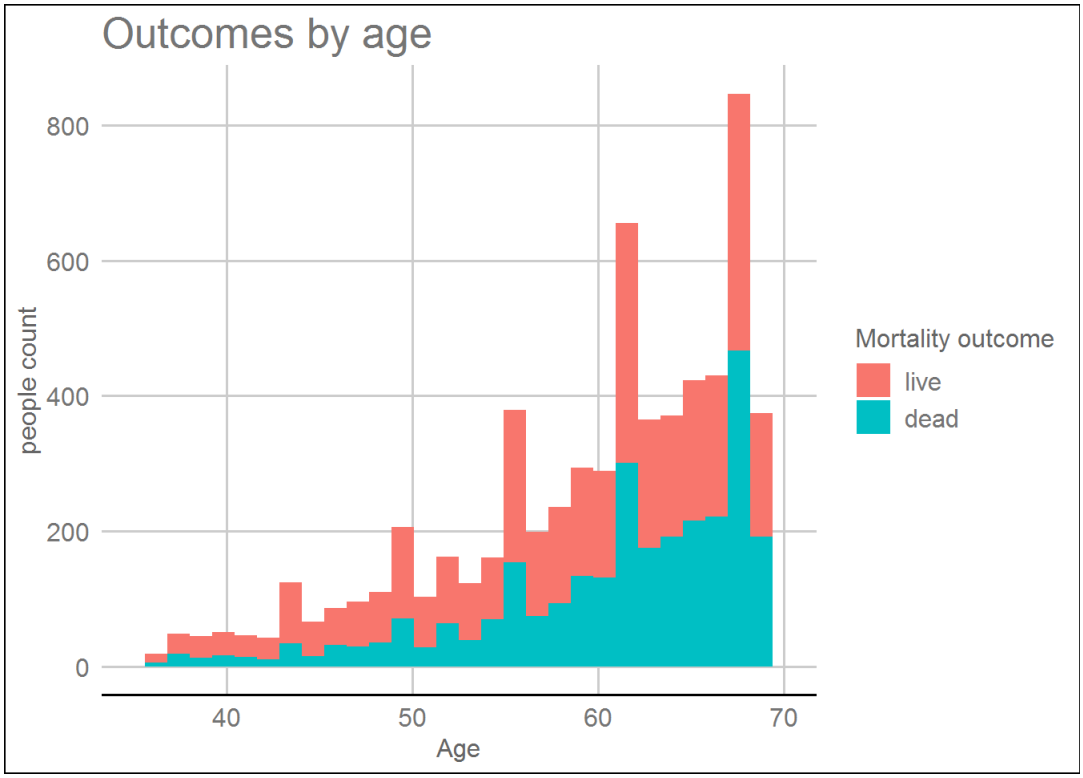
Bivariate Plots Section

> **1st life status over years** : In this section we will explore the number of deaths and survival cases per years .



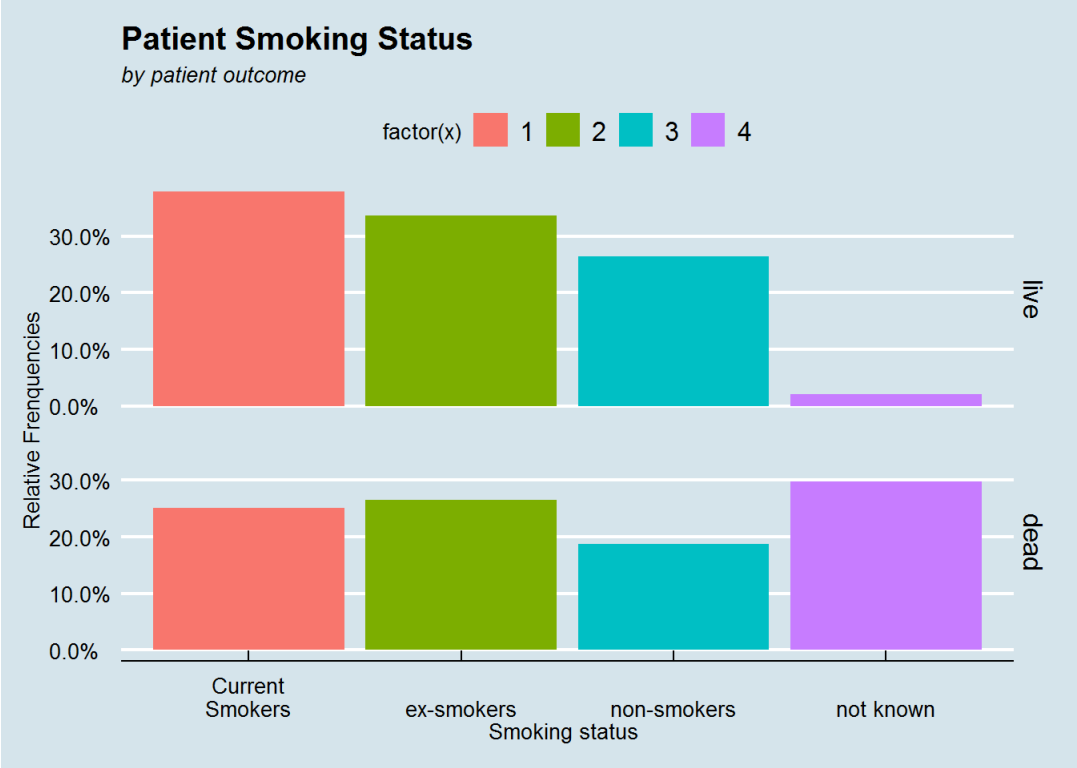
> We can See that in general the number of deaths tended to decrease which is a good indicator

2nd Outcome by age



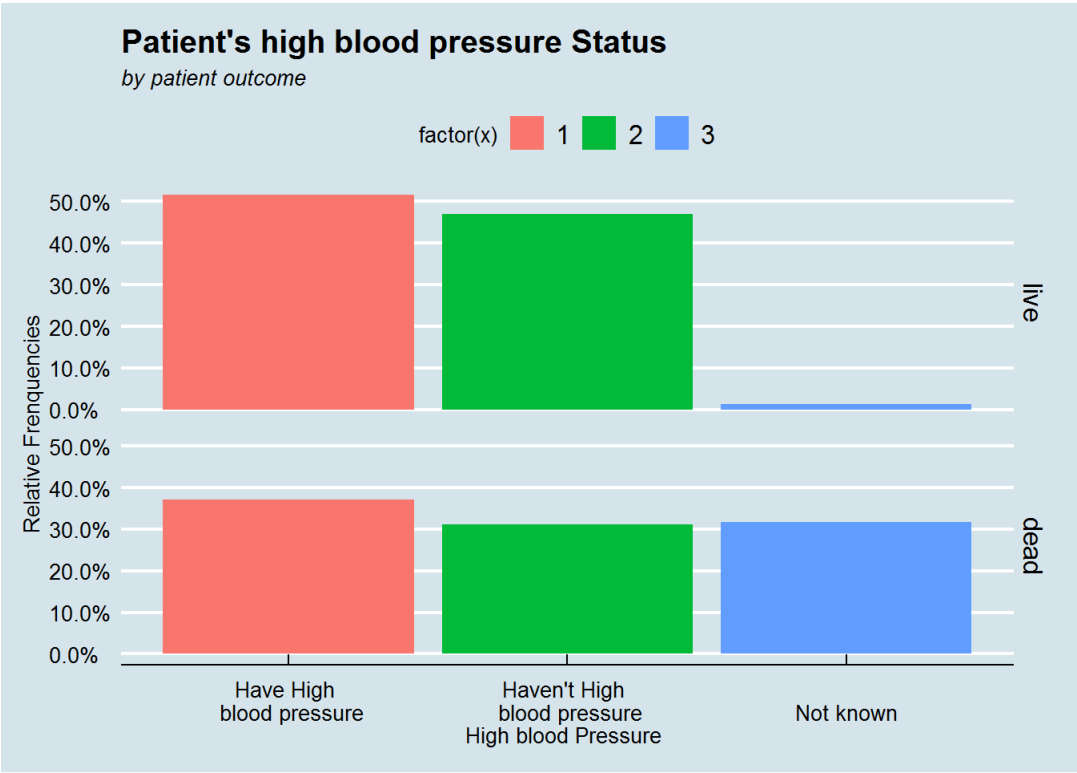
> There no general trend here but we can see that for individuals under 55 years old more people survived generally , and after 55 years old the the rate of deaths increases .

3rd Smoking Vs outcome



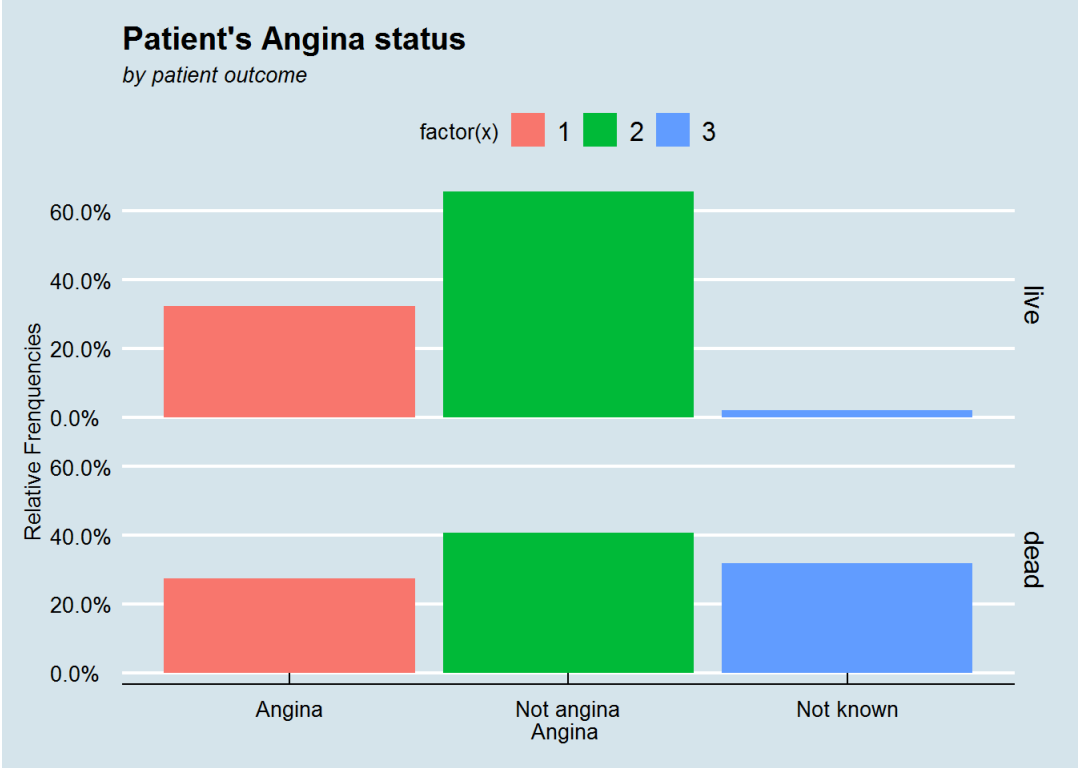
> There is a large proportion of missing values (not known) in the deceased group, so with high level of missing data we are having a misrepresented group (deaths) So this variable won't play a significant role in discovering trends.

4th High blood pressure Vs outcome



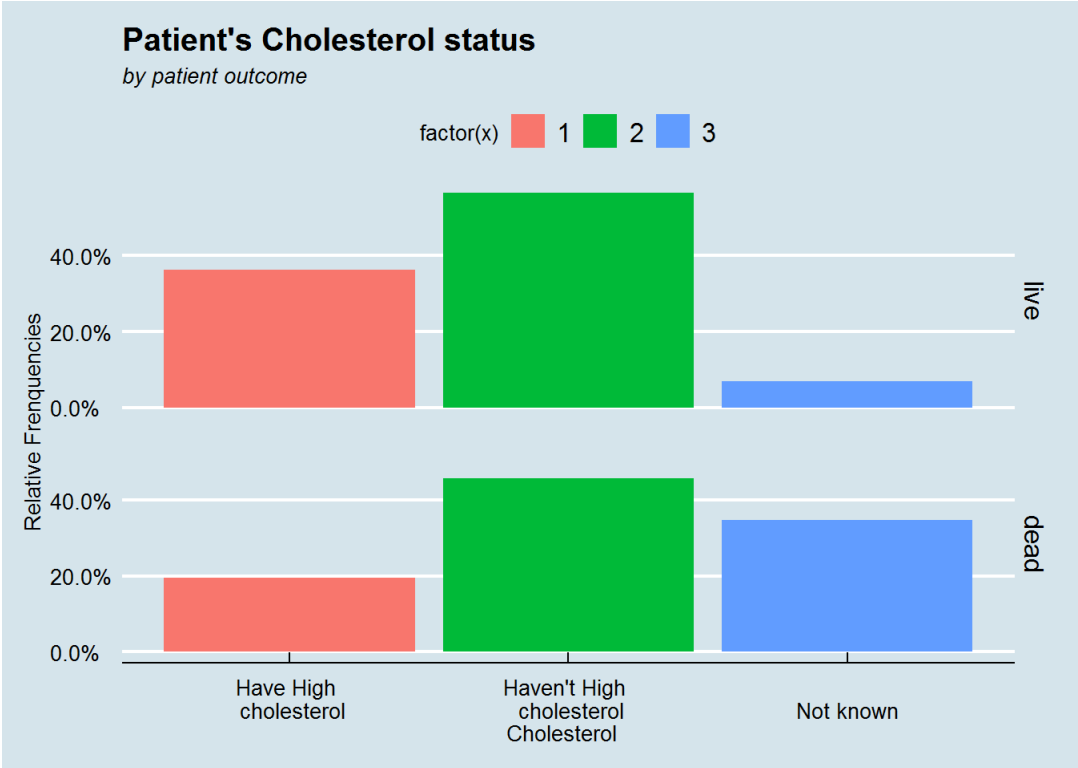
> First of all just like with the smoking status we have a huge portion of missing data in of the groups (deceased group) this make it difficult to say whether the high blood pressure plays a significant role here or not , So this variable needs to be investigated in more depth.

5th Angina Vs outcome



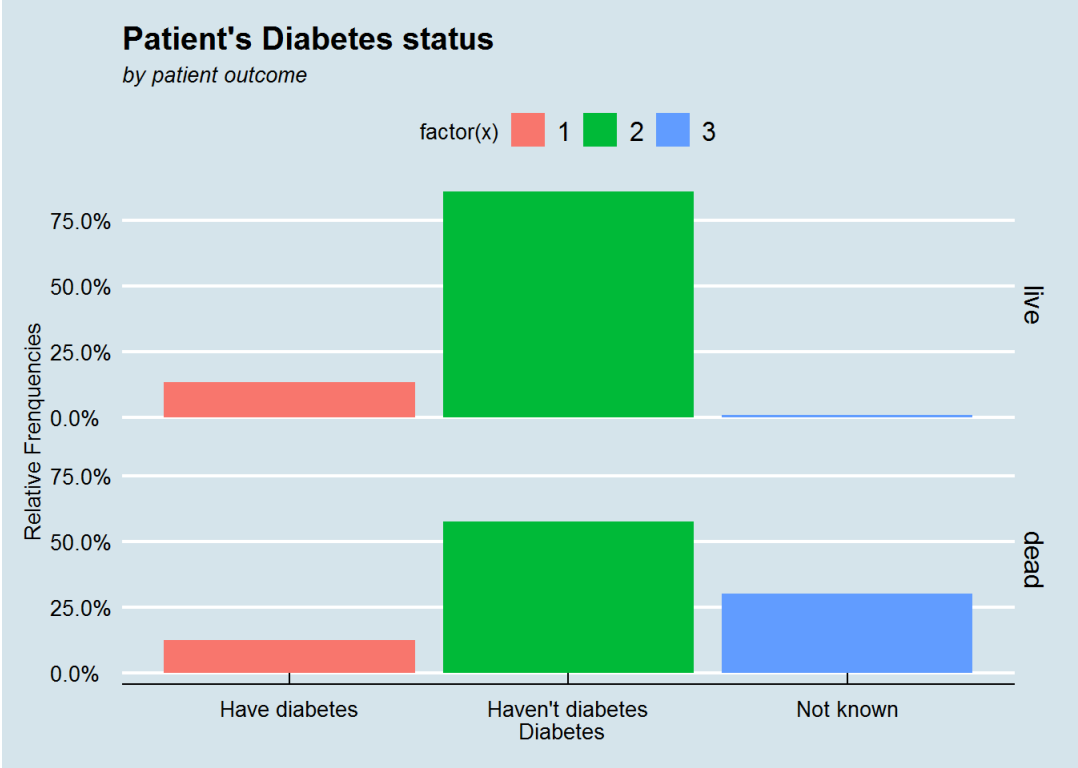
> in the living Group those who hadn't Angina survived more , So this variable maybe useful . It's to note that like the previous variables the deceased group suffers from a huge miss of data.

6th Cholesterol Vs outcome



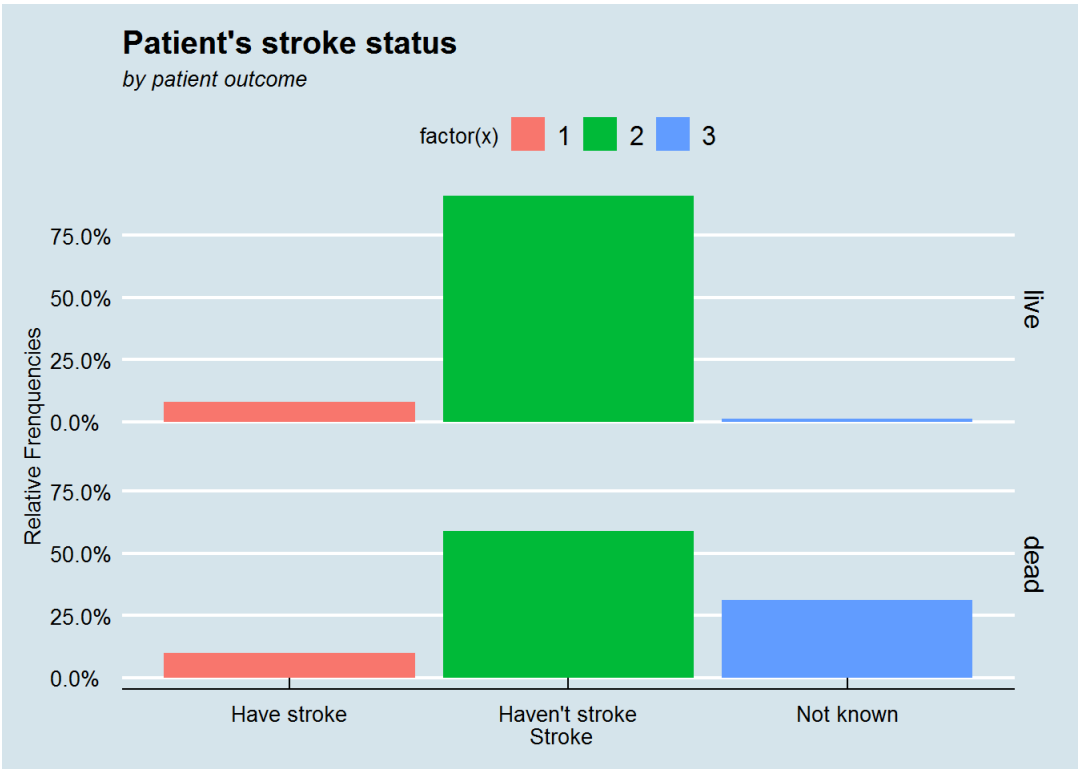
> We can see that people not having High cholesterol status are heavily represeted in both groups Plus having a big portion of missing data in the deaths group So this makes it hard to tell whether this variable is influencing or not .

7 Diabetes Vs outcome



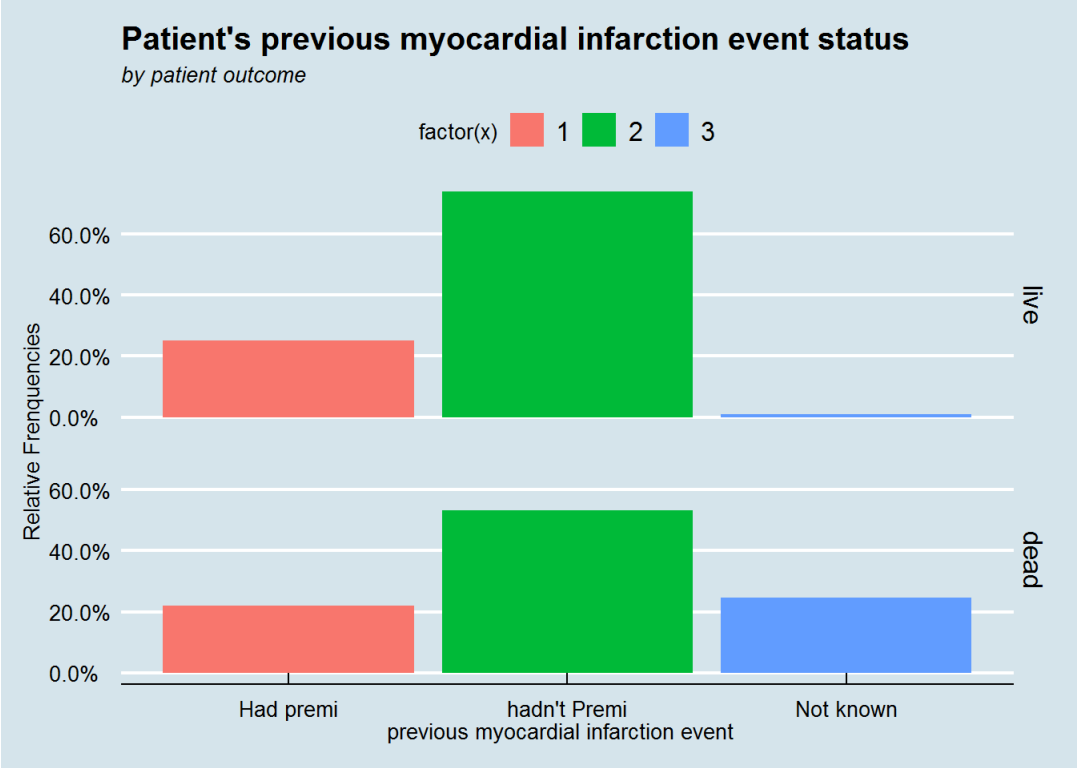
> In both groups those who have not diabetes are heavily represented Plus having a good portion of missing data in the deaths group , This variable may not be significant in discovering trends

8th Stroke Vs outcome



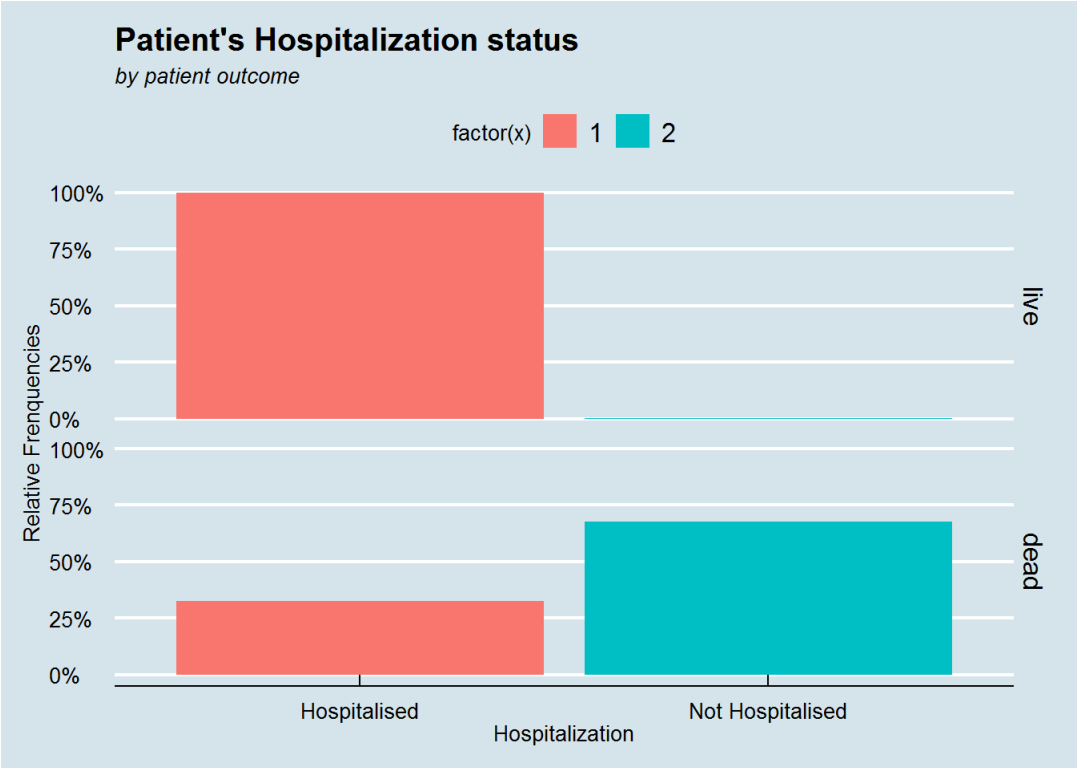
> The proportion of survivors who had a stroke is roughly similar to those who died .With as a huge missing data points in the deaths group.

9th previous myocardial infarction event Vs outcome



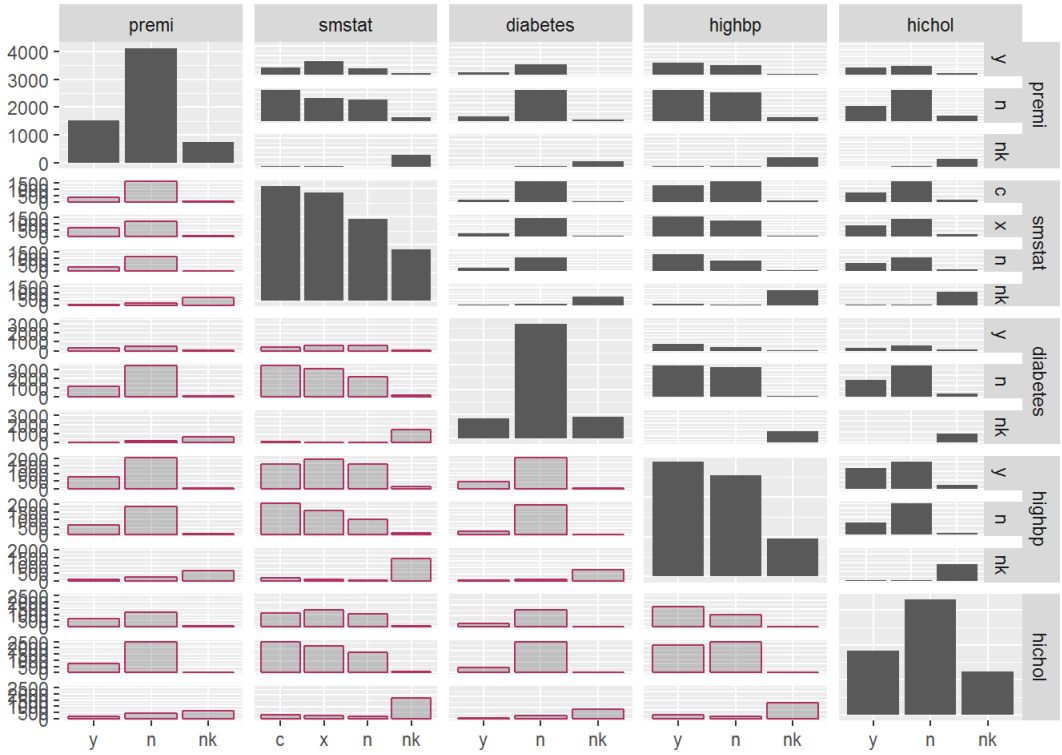
> the biggest portion of survivors hadn't previous myocardial infarction event , when it comes to the deaths group it hard to tell because we have a good portion of missing data .

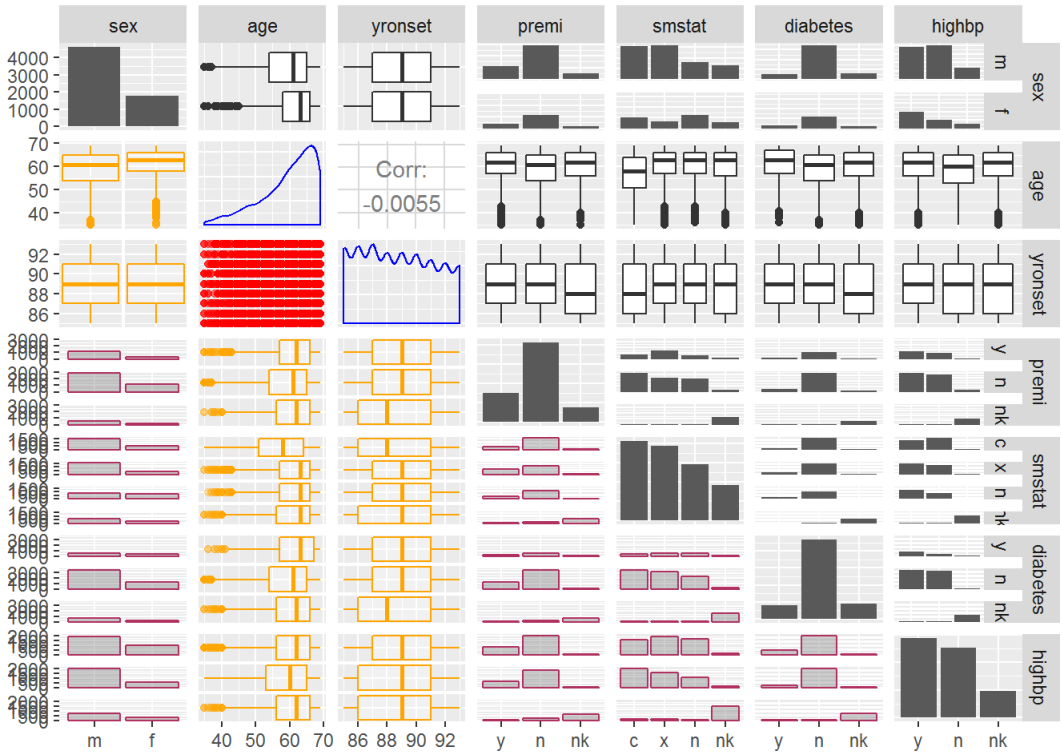
10th Hospitalization Vs outcome



> Here it can easily be seen that From those who didn't survive a large portion were hospitalized and also all the survivors were hospitalized, so this variable is significant .

11th pairs correlations between variables





Bivariate Analysis

Talk about some of the relationships you observed in this part of the
* The general trend of surving is increasing throughout the years and decreasing for the deaths which is a good sign .

- There is a relationship between hospitalization and the outcome.
- Angina may have a relashionship with the outcome but as the remain of variables the missing data makes it hard to tell

Did you observe any interesting relationships between the other features
From the scatter plot matrix we can see :

- Possible positive relationships between High blood pressure and each of : diabetes , previous myocardial infarction event
- More females have high Blood pressure .

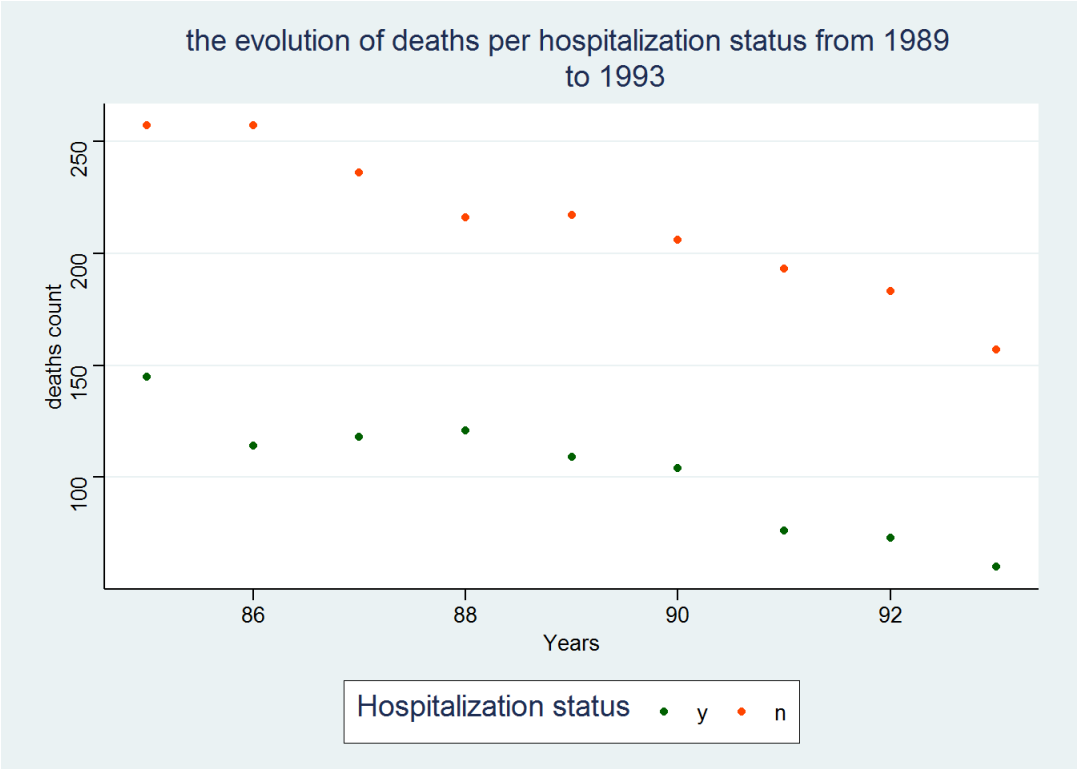
- More males have diabetes .
- Males tend to have more smoking experiences (either current or previous)

What was the strongest relationship you found?

The hospitalization of cardiovascular patients plays a principal role in their survival , so the outcome is highly correlated with the hospitalization and the age of the patient. The hospitalization of cardiovascular patients plays a principal role in their survival , so the outcome is highly correlated with the hospitalization .

Multivariate Plots Section

Deaths Vs hospitalization status :

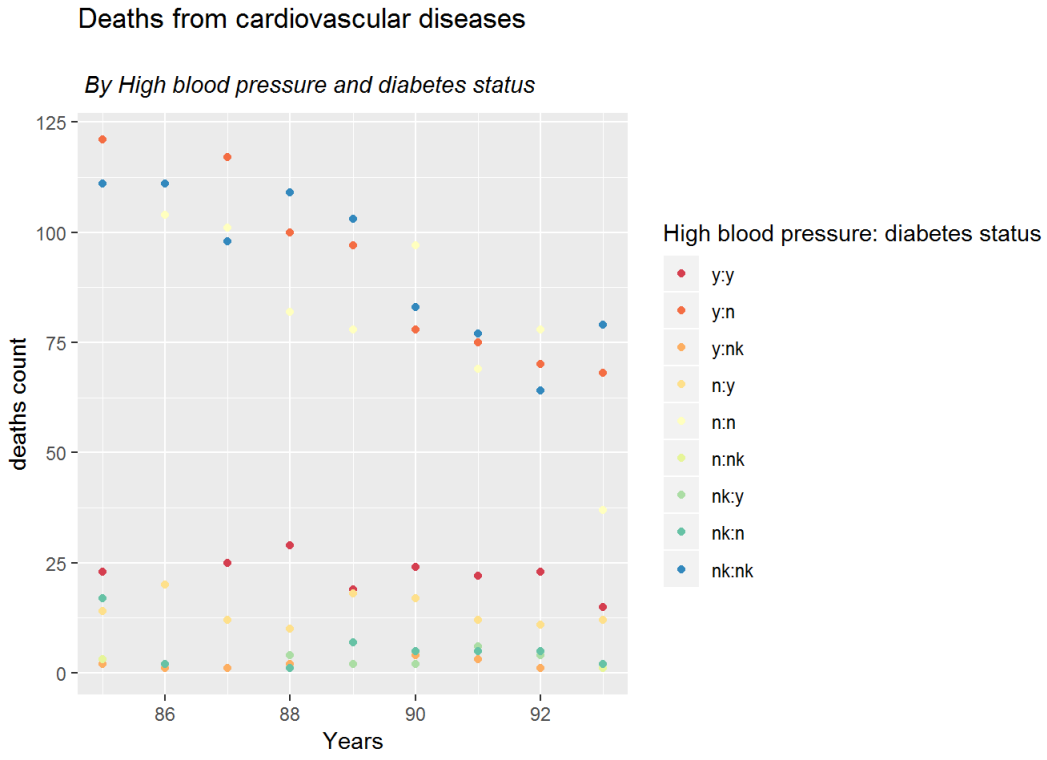


> It can be seen that throghout the years the deaths number decreased rapidly when the patient was hospitalized.

Deaths Vs (Angina and hospitalization):

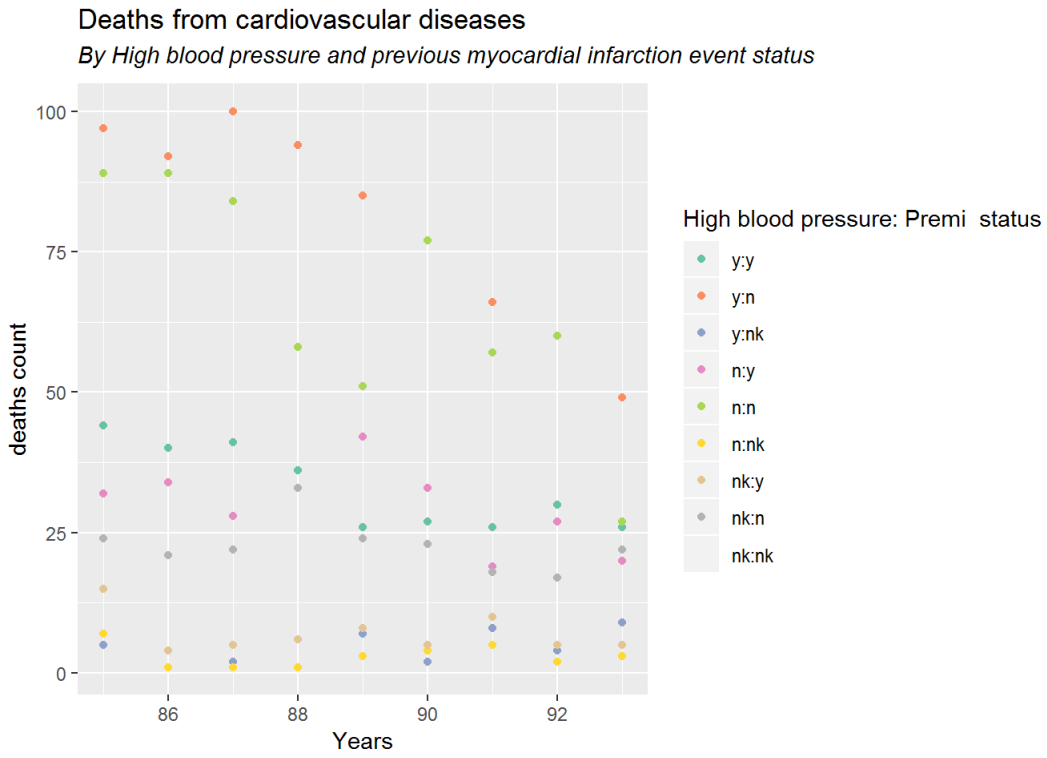
> So far not having Angina and being hospitalized released in less deaths per years , while having Angina and not being hospitalized released in more deaths.

Let's explore the deaths for patients having high blood pressure and/or diabetes :



> It can be seen that having a high blood pressure only is influencing more on deaths than having diabetes and high blood pressure together.

Let's explore the deaths for patients having high blood pressure and/or previous myocardial infarction event :



> It can be seen that having a high blood pressure only is influencing more on deaths than having Premi and high blood pressure together which interstingly scored less deaths than having none of them ! So there must be a stronger influencer on deaths (hospitalization status).

Multivariate Analysis

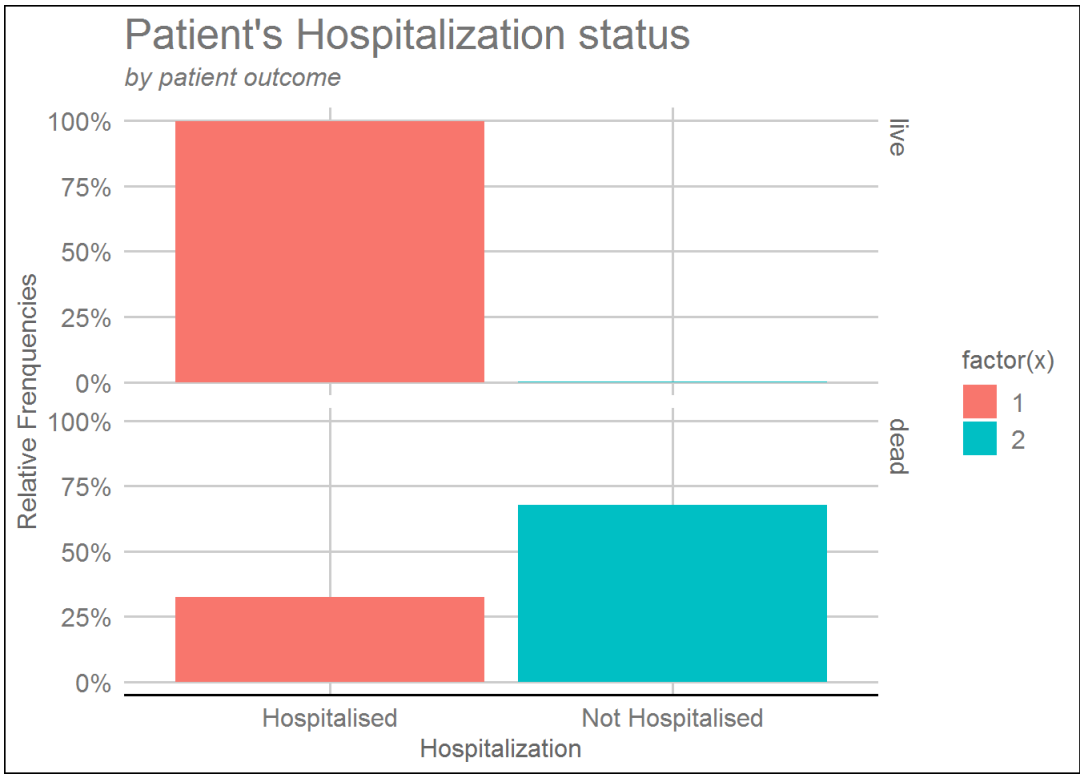
Talk about some of the relationships you observed in this part of the
The hospitazation plays a key role in decreasing the number of deaths , also having a high blood pressure increased the deaths .

Were there any interesting or surprising interactions between features?

The most suprising feature is that having high blood pressure and previous myocardial infarction event scored less deaths than not having any of them.

Final Plots and Summary

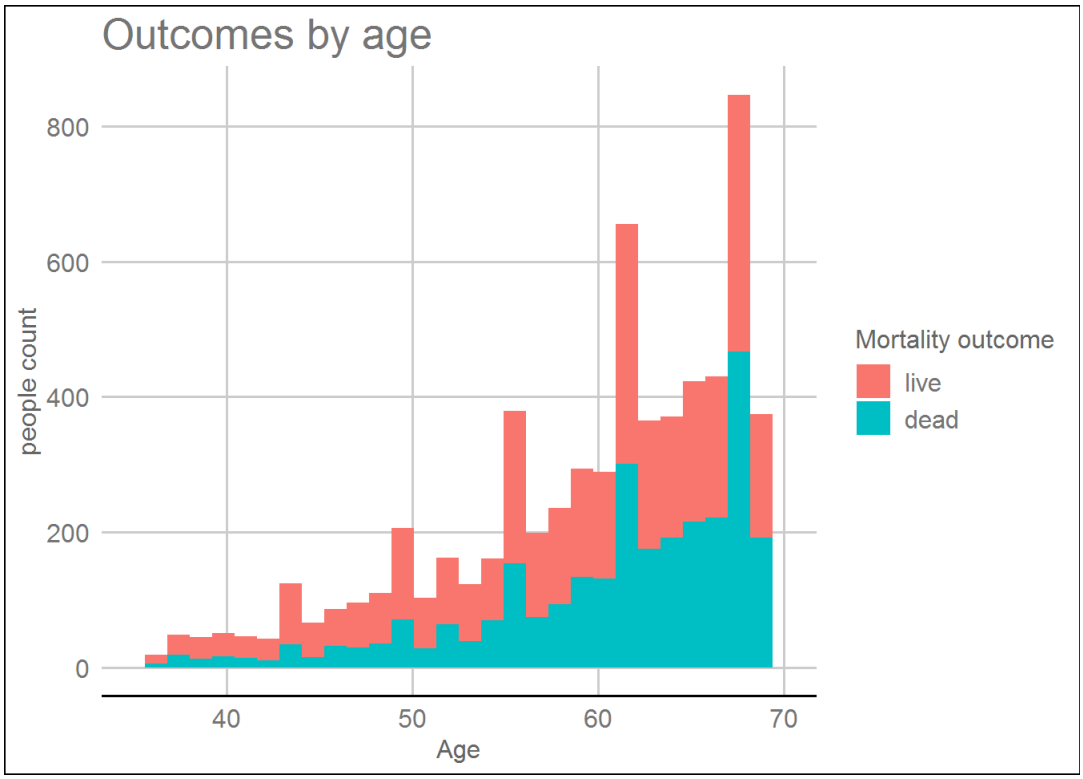
Plot One



Description One:

The plot shows clearly a strong relationship between hospitalization and the survival rate of cardiovascular patients, all of those who survived were hospitalised and most of those who died were not hospitalized.

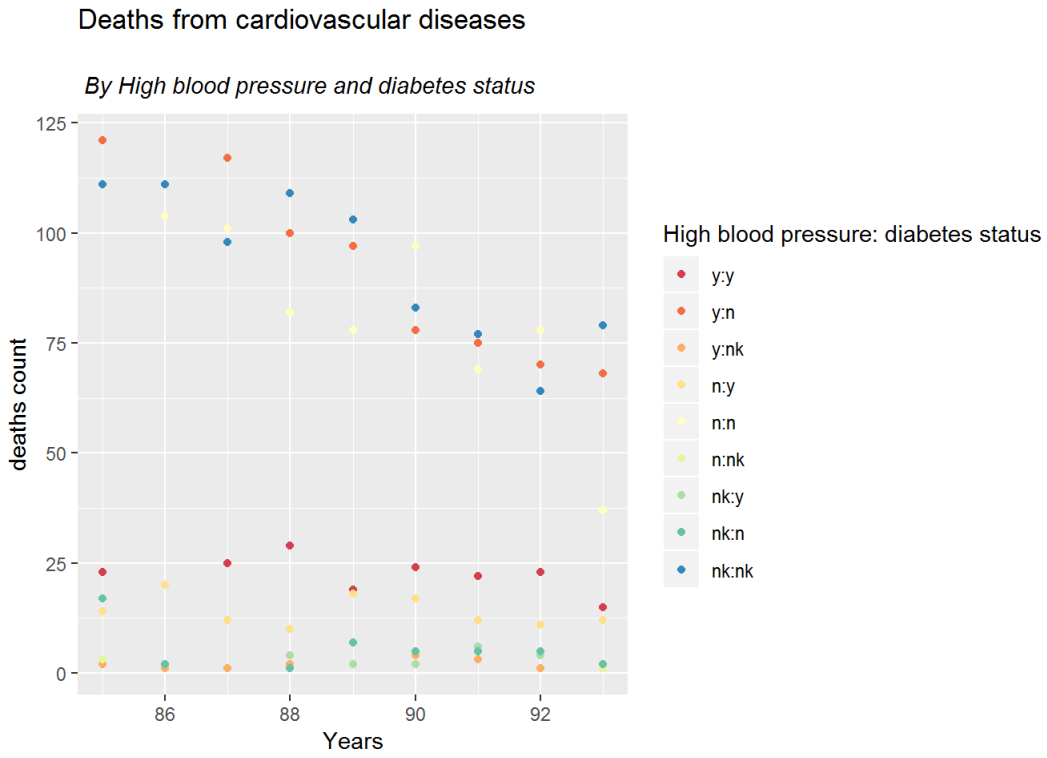
Plot Two:



Description Two:

The age of the patient also played a role in decreasing the number of deaths , the older the patient is the higher the risk of death becomes .

Plot Three:



Description Three:

Generally people havinf Dibetes have a higher chance to have high blood pressure which leds to cardiovascular issues. The plot shows an interesting and surprising fact : high blood pressure only is way more influencing on deaths than having diabetes and high blood pressure together. This discovery leds us to think that there must be some kind of relationship between Diabetes and High blood pressure from a side and the deaths caused by cardiovascular diseases from another . Or maybe having both raised the chance of being diagnosed at least by one of them and as a result receive some kind of treatments that helped reduces the risks and seriousness of the cardiovascular events , however these are just assumptions and since we don't have enough data to support them , they will still only hypothesis under the scope of this study.

Reflection:

Issues with this Analysis : The main issue was the huge amount of missing data for many variables which made it difficult to identify which variables really influence the outcome , this also made it hard to tell which variables were really correlated to each other, and Also the nature of variables (factors) it would be more interesting if we had high blood pressure measurements instead of a categorical variable only.

Surprises Some of the well known causes of deaths in that period like diabetes didn't have much effect in this data .

Conclusions : This dataSet was challenging and surprising at the same time , in the beginning I was thinking that some variables like Premi , Angina and high cholesterol will play a significant role in causing deaths for cardiovascular patients , but when taking a look at the sources of data and the nationalities of patients (see : https://thl.fi/publications/monica/coredb/table1.htm) ,These countries were having poor economics by then So the consumption of fatty meals was not frequent which is the first cause of cholesterol ; So this fact can explain the weak relationship between cardiovascular patients deaths and their cholesterol levels.
More interesting Investigation could've been done if we had numerical data instead of categorical variables , models could've been created for each variable's trends .
The key factors in this investigation were the age of the patients , their high blood pressure levels and also the hospitalization status were all improuving the rate of survivals for these patients .