Covid Pre-Condition

Rafael Pereira

2/1/2021

Introduction

This algorithm use data to predict if a covid positive patient will need hospitalization or not based in his/her pre-conditions. The data used for this is COVID-19 patient pre-condition dataset (https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset) acquired from the Mexican government. The variable that will be predicted is the type of patient (variable's name: patient_type), 1 for outpatient (a patient who receives medical treatment without being admitted to a hospital) and 2 for inpatient (a patient who's been admitted to hospital for medical treatment).The pre-conditions (predictors) used for this are:sex of the patient (1 for female and 2 for male, variable's name: sex),the age of the patient (variable's name: age). In the next variables 1 indicates that the patient has it and 2 that the patient doesn't have it: pneumonia (variable's name: pneumonia), diabetes(variable's name: diabetes),chronic obstructive pulmonary disease (variable's name: copd),asthma (variable's name: asthma),immunosuppression (variable's name: inmsupr),hypertension (variable's name: hypertension),other diseases (variable's name: other_disease), cardiovascular diseases(variable's name:cardiovascular), obesity (variable's name: obesity),chronic kidney disease (variable's name: renal_chronic) and smoking habits (variable's name: tobacco). Only the covid positive patients will be used for the models(variable's name: covid_res) For the algorithm are used five models: linear discriminant analysis (LDA), generalized linear model (LGM), quadratic discriminant analysis (QDA), classification and regression tree (RPART) and an ensemble of the other four models.

Analisis

The first step is the installation of the libraries that will be used:

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-project.org")
if(!require(gridExtra)) install.packages("gridExtra", repos ="http://cran.us.r-project.org")

library(tidyverse)
library(caret)
library(data.table)
library(ggplot2)
library(gridExtra)
```

The data is loaded from the url in cvs format, after that is read and stored in a data frame:

```
url <- "https://github.com/RPereira98/Covid-Pre-condition/raw/main/covid.zip"
dl <- tempfile()
download.file(url, dl)
```

```
unzip(dl,"covid.csv")
covid_dat <- read_csv("covid.csv")
covid_dat<-data.frame(covid_dat)
```

Only the covid positive patients are important for the models:

```
ind<- which(covid_dat$covid_res==1)
covid_dat<-covid_dat[ind,]
```

Some variables will not be used for the analysis (dates,ID of patients, pregnancy, intubation, ICU):

```
covid_dat<-covid_dat[,-c(1,4,5,6,7,10,21,22,23)]
```
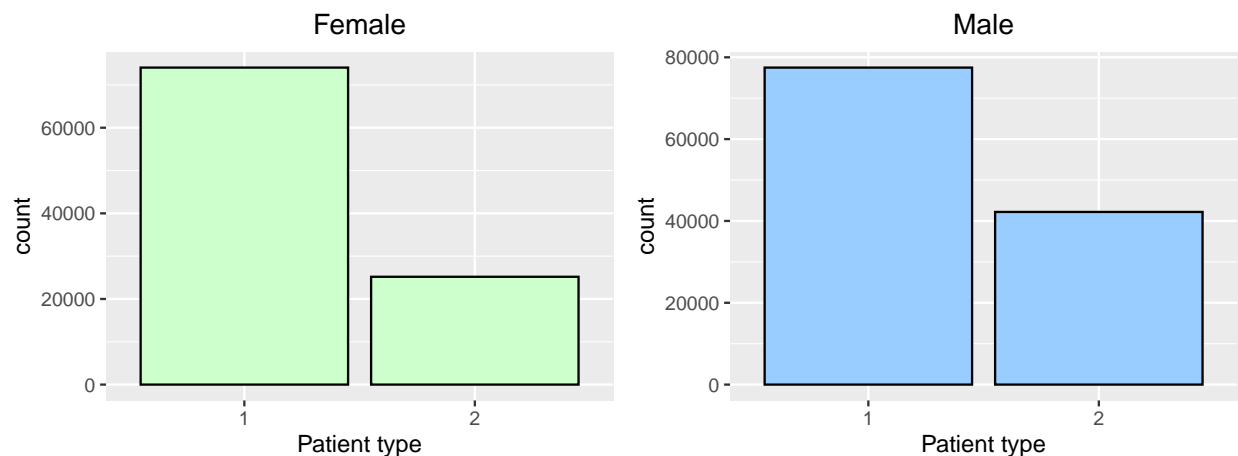
The values 97, 98 and 99 are NAs, not useful:

```
ind<- which((covid_dat$pneumonia%in%c(97,98,99))|(covid_dat$diabetes%in%c(97,98,99))|(covid_dat$copd%in%
covid_dat<-covid_dat[-ind,]
```

There are more outpatients than inpatients, the proportion of outpatients is:

```
mean(covid_dat$patient_type==1)
```

```
## [1] 0.6921728
```

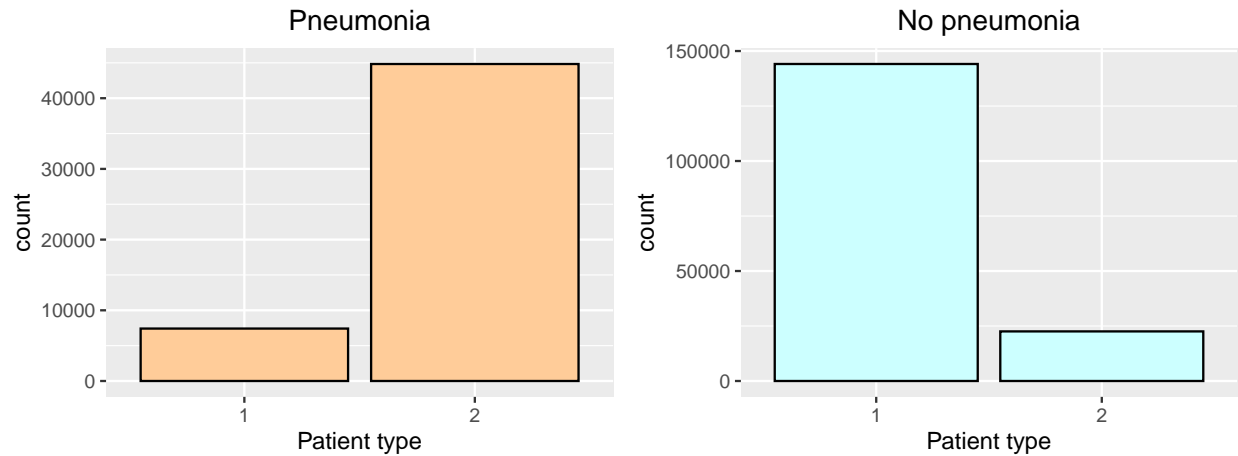Doing a visual inspection of the data: First parameter to analyze: Sex

Both sexes have more outpatients than inpatients,with men having a slightly higher percentage of inpatients. The correlation of sex and patient type is:

```
## [1] 0.1065695
```

A small correlation value. An interpretation could be that sex doesn't influence much in the patient type.
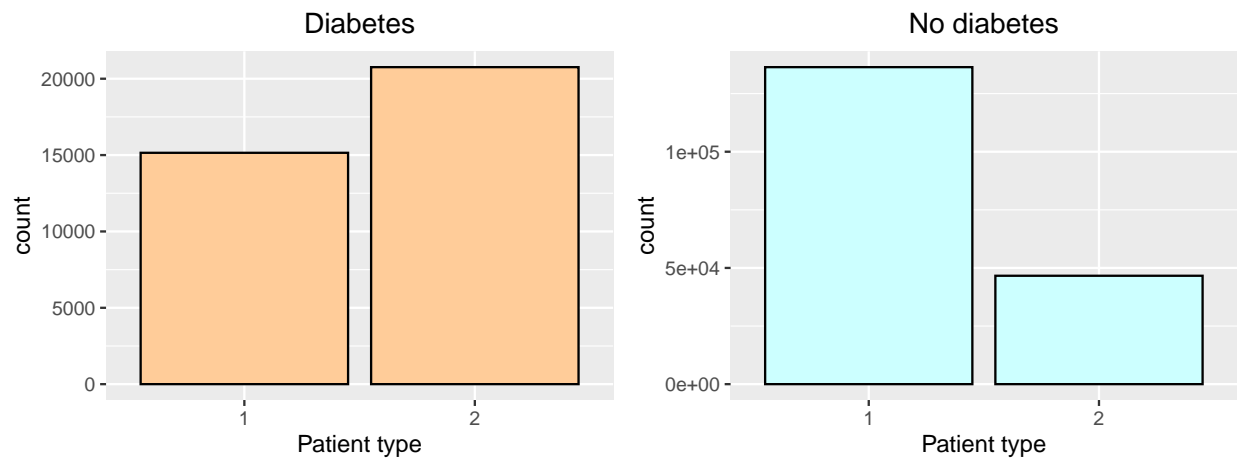
Parameter: Pneumonia

2

More patients with pneumonia are inpatients than outpatients, and more patients without pneumonia are outpatients than inpatients. The correlation value is:

```
## [1] -0.6675009
```

The correlation value is big and negative, indicating that patients with pneumonia are more susceptible to need hospitalization.
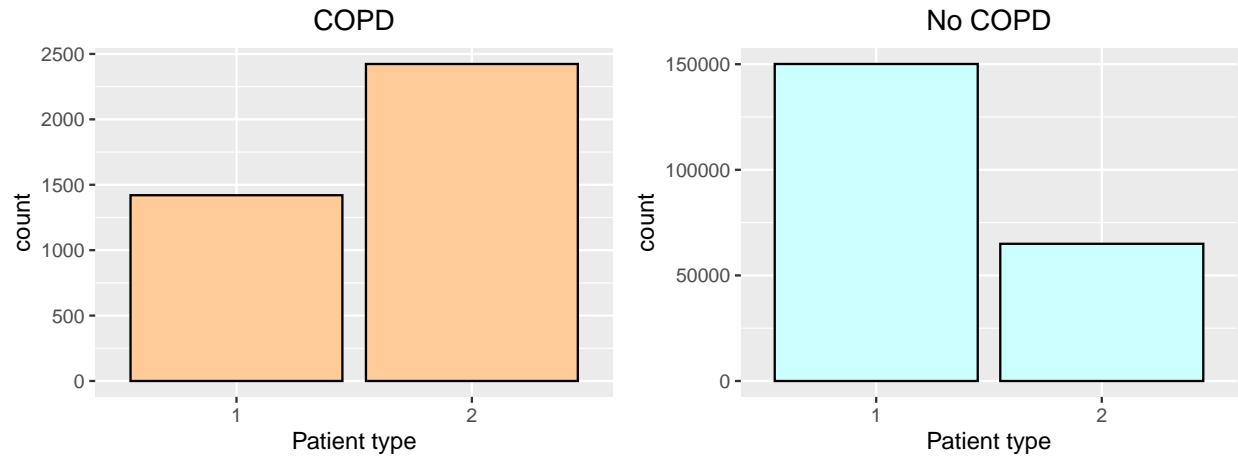
Parameter: Diabetes

There are slightly more patients with diabetes that are inpatients than outpatients, and there are more patients without diabetes that are outpatients than inpatients. The correlation value is:

```
## [1] -0.2593493
```

A small negative value, maybe the diabetes influence a little in the patient type.

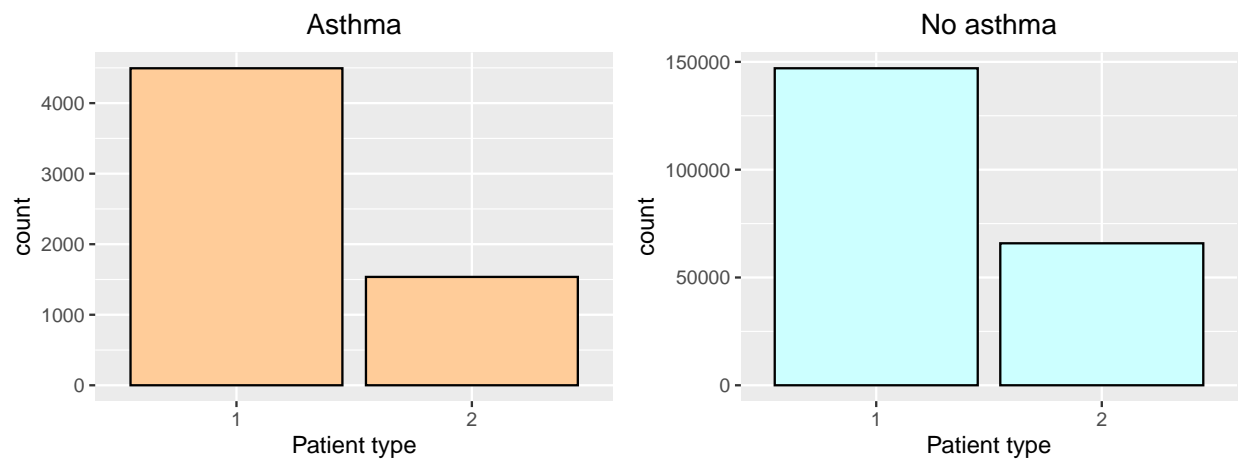Parameter: Chronic obstructive pulmonary disease

There are more patients with COPD that are inpatients than outpatients and there are more outpatients without COPD than inpatients without COPD. The correlation value is:

```
## [1] -0.09344453
```

A very small negative correlation value, may be due to the small quantity of patients with COPD.
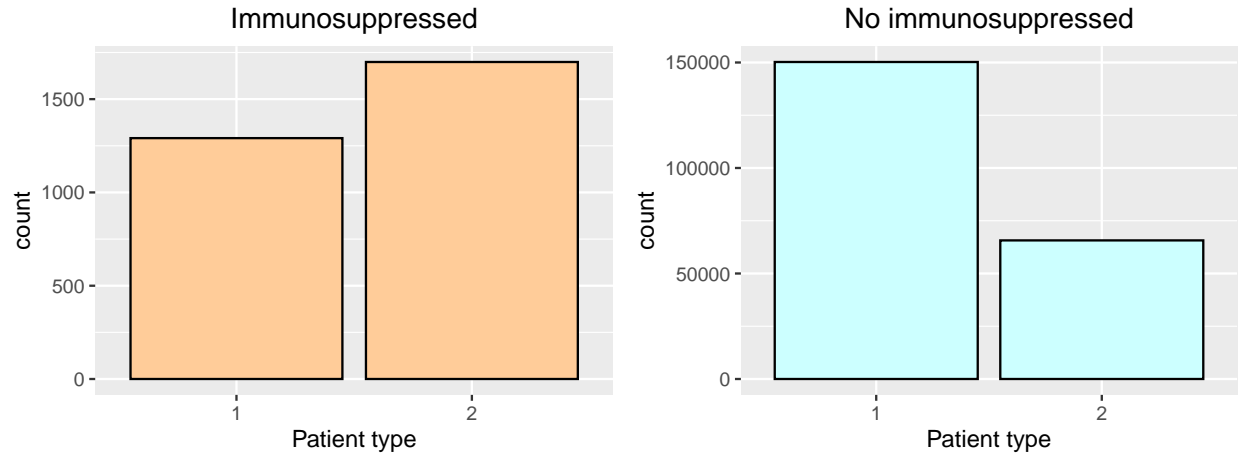
Parameter: Asthma

Both patients with and without asthma have more outpatients than inpatients. The correlation value is:

```
## [1] 0.01936156
```

A very small positive value, maybe asthma doesn't influence much in the need of hospitalization or not.
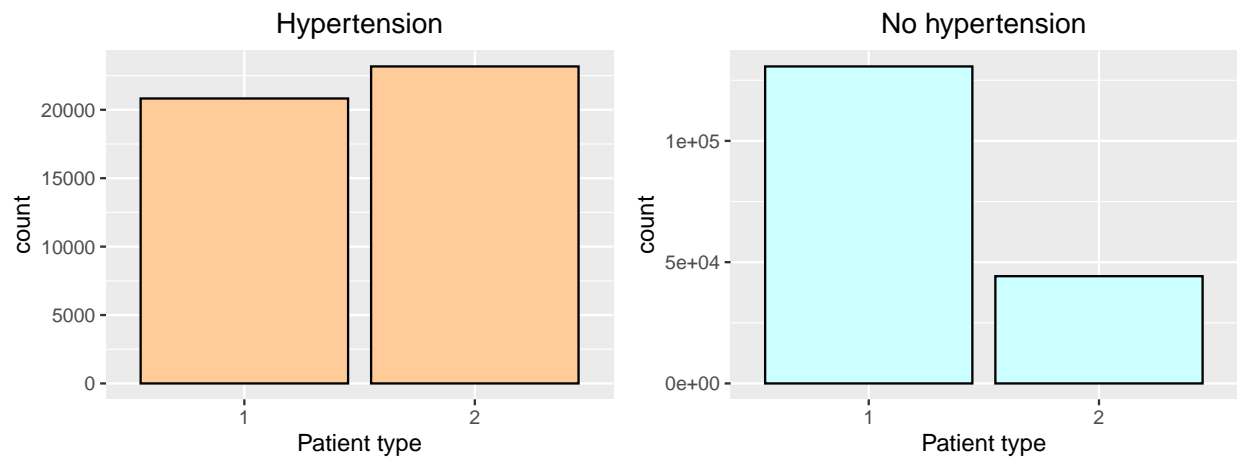
Parameter: Immunosuppression

There are slightly more patients with immunosuppression that are inpatient than outpatient and there are more outpatients without immunosuppression than inpatients. The correlation value is:

```
## [1] 0.01936156
```

A very small negative value, maybe due to the small quantity of patients with immunosuppression.
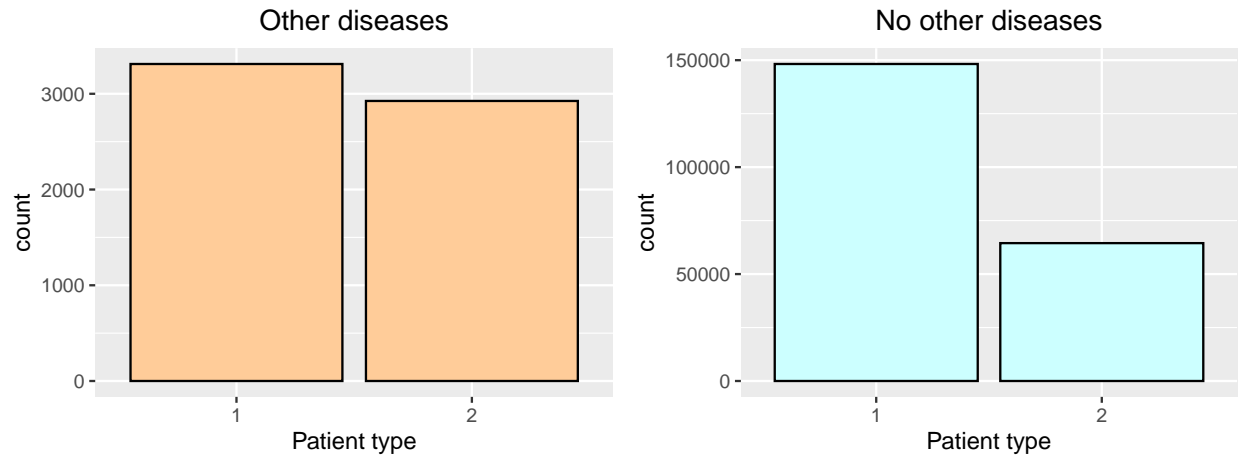
Parameter: Hypertension

There are slightly more patients with hypertension that are inpatients than outpatients and more patients without hypertension that are outpatients than inpatients. The correlation value is:

```
## [1] -0.237701
```

A small negative value, maybe hypertension doesn't influence a lot in the hospitalization.
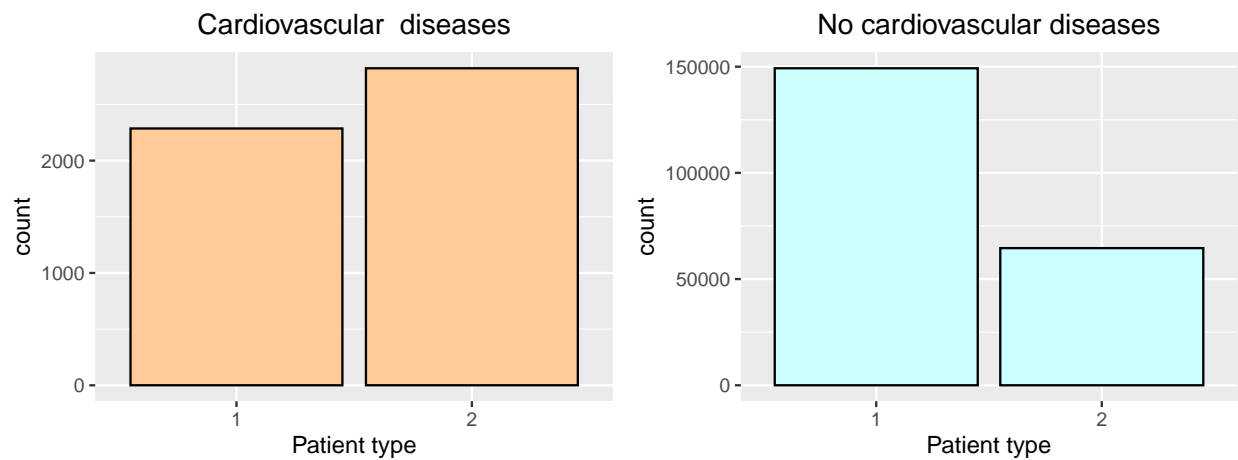
Parameter: Other diseases

There are slightly more outpatients with other diseases than inpatients and there are more outpatients without other diseases than inpatients. The correlation value is:

```
## [1] -0.05980967
```

A very small negative value, may be due to the small quantity of patients with other diseases.

Parameter: Cardiovascular diseases

There are slightly more inpatients with other cardiovascular diseases than outpatients and there are more outpatients without cardiovascular diseases than inpatients. The correlation value is:

```
## [1] -0.08192066
```

A very small negative value, may be due to the small quantity of patients with cardiovascular diseases.
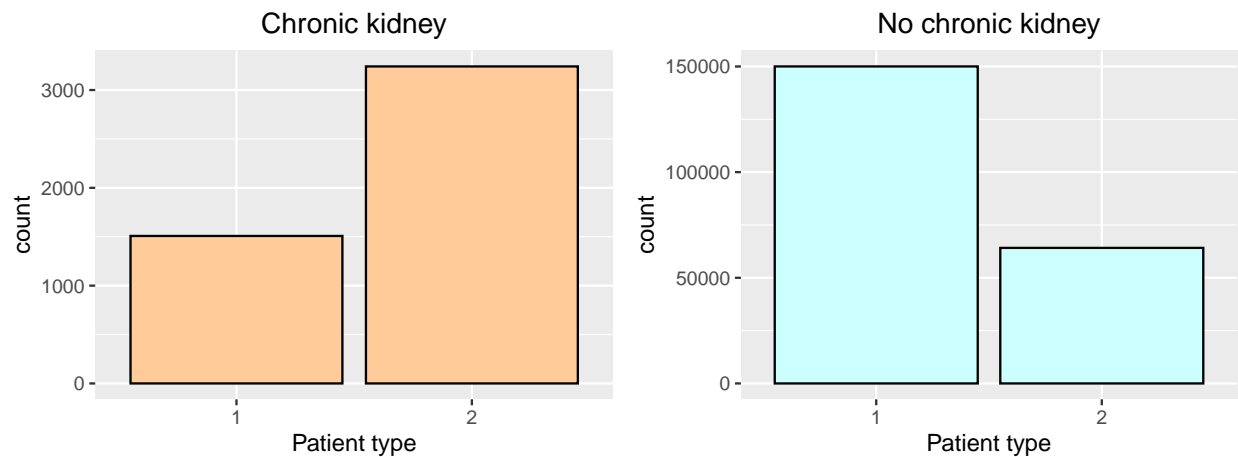
Parameter: Obesity
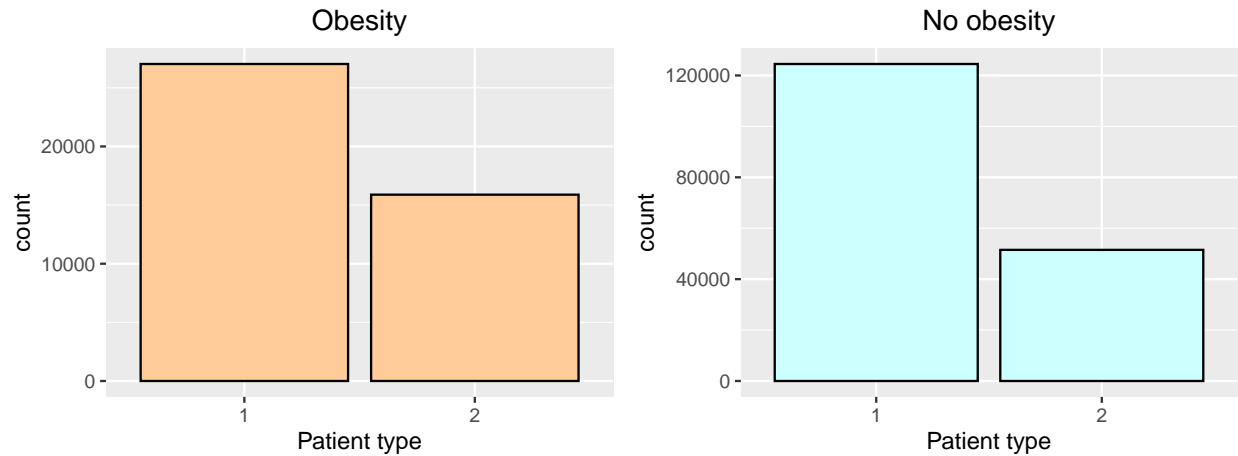
There are slightly more outpatients with obesity than inpatients and there are more outpatients without obesity than inpatients. The correlation value is:

```
## [1] -0.08192066
```

A very small negative value, maybe obesity doesn't influence much in the type of patient.

Parameter:Chronic kidney disease

There are more inpatients with chronic kidney disease than outpatients and more outpatients without chronic kidney disease than inpatients. The correlation value is:

```
## [1] -0.1208602
```

A small negative value. The small value may be due to the small quantity of patients with the disease,but is bigger than other parameters with small quantities of patients with the disease, maybe influence considerably the type of patient.

Parameter: Smoking habit
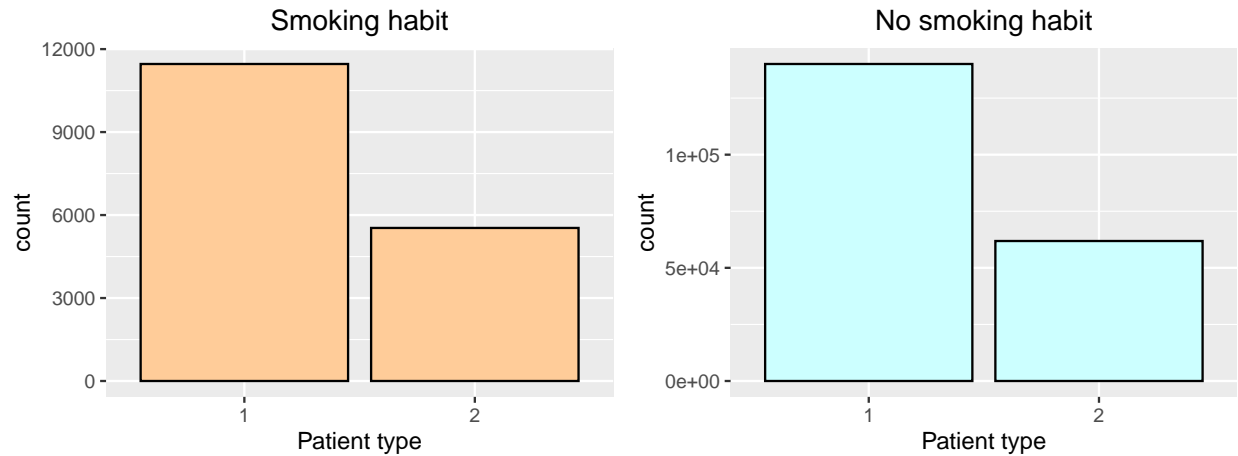
\end{figure}

With or without smoking habits the number of outpatients is bigger than the number of inpatients. The correlation value is:

```
## [1] -0.01118649
```

A small negative value, maybe smoking habit doesn't influence a lot in the type of patient.

Parameter: Age

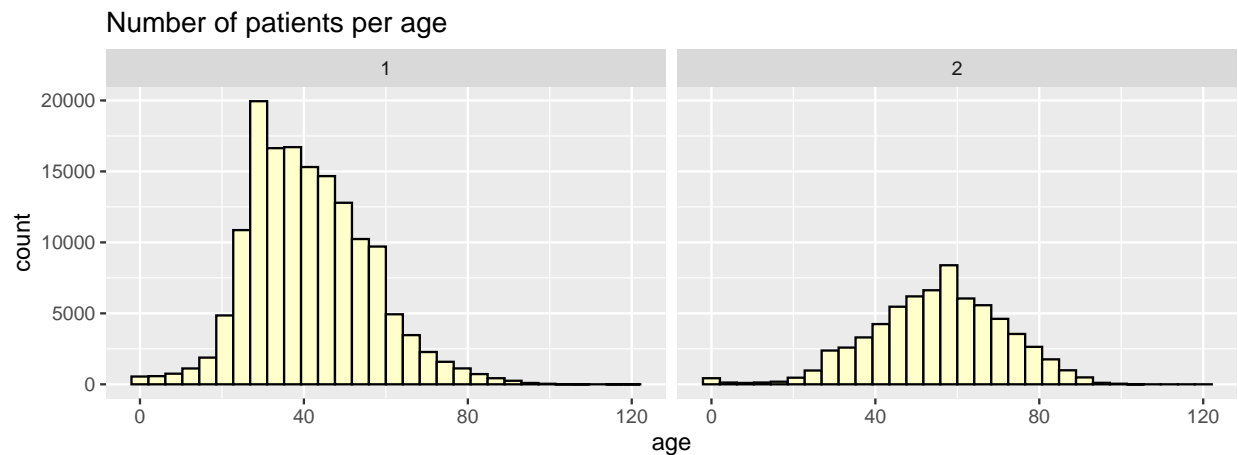The proportion of inpatients change with age, it grows with the age. The correlation value is:

```
## [1] 0.3940115
```

A not very big but considerable positive value. Maybe age influence considerably in the patient type.

Splitting the data in a training and test sets. The test set will be 20% of the original data set.

```
set.seed(1, sample.kind="Rounding")
test_index <- createDataPartition(y = covid_dat$patient_type, times = 1, p = 0.2, list = FALSE)
train_set<-covid_dat[-test_index,]
test_set<-covid_dat[test_index,]
```

The first model will be linear discriminatory analysis.

```
train_lda<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
data =train_set,method="lda")
lda_pred<-predict(train_lda,test_set)
cm<-confusionMatrix(table(as.numeric(lda_pred), test_set$patient_type))
lda_ac<-cm$overall[["Accuracy"]]
```

The second model will be a generalized linear model.

```
train_glm<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
method="glm",data =train_set)
glm_pred<-predict(train_glm,test_set)
cm<-confusionMatrix(table(as.numeric(glm_pred), test_set$patient_type))
glm_ac<-cm$overall[["Accuracy"]]
```

The third model will be a quadratic discriminatory analysis.

```
train_qda<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
data =train_set,method="qda")
qda_pred<-predict(train_qda,test_set)
cm<-confusionMatrix(table(as.numeric(qda_pred), test_set$patient_type))
qda_ac<-cm$overall[["Accuracy"]]
```

The fourth model will be a classification tree.

```
train_rpart<-train(as.factor(patient_type)~as.factor(sex)+as.factor(pneumonia)+age+
as.factor(diabetes)+as.factor(copd)+as.factor(asthma)+as.factor(inmsupr)+
as.factor(hypertension)+as.factor(other_disease)+as.factor(cardiovascular)+
as.factor(obesity)+as.factor(renal_chronic)+as.factor(tobacco),
data = train_set,method="rpart")
rpart_pred<-predict(train_rpart,test_set)
cm<-confusionMatrix(table(as.numeric(rpart_pred), test_set$patient_type))
rpart_ac<-cm$overall[["Accuracy"]]
```

The fifth model will be an ensemble of the other four models.If the majority of the models predict inpatient, the ensemble will predict inpatient. If the majority of models predict outpatient it will predict outpatient. If there is a tie, the ensemble will predict an outpatient, because there are more outpatients.

```
ensemble<-data.frame(LDA=as.numeric(lda_pred),
                     QDA=as.numeric(qda_pred),
                     GLM=as.numeric(glm_pred),
                     RPART=as.numeric(rpart_pred))
```

```
ensemble_pred<-ifelse(rowMeans(ensemble)<=(1*2+2*2)/4,1,2)
cm<-confusionMatrix(table(ensemble_pred, test_set$patient_type))
ensemble_ac<-cm$overall[["Accuracy"]]
```

Finally a accuracy data.frame is created:

```
accuracy<-data.frame(row.names = c("LDA","GLM","QDA","RPART","ENSEMBLE"),
                     Accuracy=c(lda_ac,glm_ac,qda_ac,rpart_ac,ensemble_ac))
```

Results

The accuracy of the models are:

```
##           Accuracy
## LDA      0.8618579
## GLM      0.8621777
## QDA      0.8423289
## RPART    0.8637308
## ENSEMBLE 0.8626345
```

The ensemble has a better performance than LDA,GLM and QDA models but inferior than the RPART model's performance. All the models have an accuracy around of 86%, only the QDA method is inferior to that with an accuracy of 0.842. The accuracy of the others models doesn't vary much, only changing from the third decimal. One of the motives that the QDA model has a considerably low accuracy in comparison to the other models is that there are 13 predictors and QDA doesn't perform well with many predictors.

Conclusion

The analysis of the plots and correlation values are based on the sample. The negative correlation value indicates (except for sex and age) that a patient with the pre-condition is more susceptible to be hospitalized (inpatient), a positive value indicate that a patient with the pre-condition is more susceptible to be an outpatient. For sex the positive value indicate that a male is more susceptible to be hospitalized, but the value is very small, so maybe doesn't influence much. For age the positive value indicate that the older the patient the more susceptible to need a hospitalization is. For confirmation of these conclusions more studies are needed. The models could help to prioritize the medical attention of patients that are more susceptible to need a hospitalization based on the pre-conditions. None of the models uses pregnancy, dates, intubation or UCI to make predictions, using these or more parameters (others pre-conditions as diet,physical activity, consumption of alcohol,etc.) could help to improve the performance of the models.