Diabetes

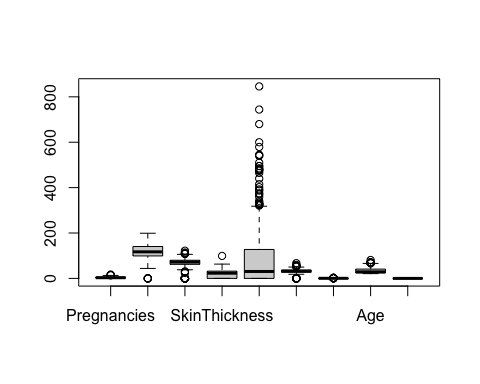
Beatriz Gámez

2024-07-15

## DIABETES DATASET

Data from pregnant women.

diabetes= read.csv("../datasets/diabetes.csv")  
View(diabetes)  
boxplot(diabetes)



## Data exploration

str(diabetes) # we can see all the variables are int or numeric. The outcome does not make sense to be integer.

## 'data.frame': 768 obs. of 9 variables:  
## $ Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ...  
## $ Glucose : int 148 85 183 89 137 116 78 115 197 125 ...  
## $ BloodPressure : int 72 66 64 66 40 74 50 0 70 96 ...  
## $ SkinThickness : int 35 29 0 23 35 0 32 0 45 0 ...  
## $ Insulin : int 0 0 0 94 168 0 88 0 543 0 ...  
## $ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...  
## $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...  
## $ Age : int 50 31 32 21 33 30 26 29 53 54 ...  
## $ Outcome : int 1 0 1 0 1 0 1 0 1 1 ...

diabetes$Outcome= as.factor(diabetes$Outcome)  
  
summary(diabetes) # Now summary for outcome makes sense where we can see the number of events and not the average and so on.

## Pregnancies Glucose BloodPressure SkinThickness   
## Min. : 0.000 Min. : 0.0 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 1.000 1st Qu.: 99.0 1st Qu.: 62.00 1st Qu.: 0.00   
## Median : 3.000 Median :117.0 Median : 72.00 Median :23.00   
## Mean : 3.845 Mean :120.9 Mean : 69.11 Mean :20.54   
## 3rd Qu.: 6.000 3rd Qu.:140.2 3rd Qu.: 80.00 3rd Qu.:32.00   
## Max. :17.000 Max. :199.0 Max. :122.00 Max. :99.00   
## Insulin BMI DiabetesPedigreeFunction Age   
## Min. : 0.0 Min. : 0.00 Min. :0.0780 Min. :21.00   
## 1st Qu.: 0.0 1st Qu.:27.30 1st Qu.:0.2437 1st Qu.:24.00   
## Median : 30.5 Median :32.00 Median :0.3725 Median :29.00   
## Mean : 79.8 Mean :31.99 Mean :0.4719 Mean :33.24   
## 3rd Qu.:127.2 3rd Qu.:36.60 3rd Qu.:0.6262 3rd Qu.:41.00   
## Max. :846.0 Max. :67.10 Max. :2.4200 Max. :81.00   
## Outcome  
## 0:500   
## 1:268   
##   
##   
##   
##

table(is.na(diabetes))

##   
## FALSE   
## 6912

table(is.null(diabetes))

##   
## FALSE   
## 1

We can see max.age is 81. That is probably a mistake.Consider.

Also variables as Insulin, Blood pressure, Glucose have several 0 values. That does not make sense.

## Questions

# What is the average number of pregnancies for the diabetic women? And for the non diabetic?  
  
mean(diabetes$Pregnancies[diabetes$Outcome == 1])

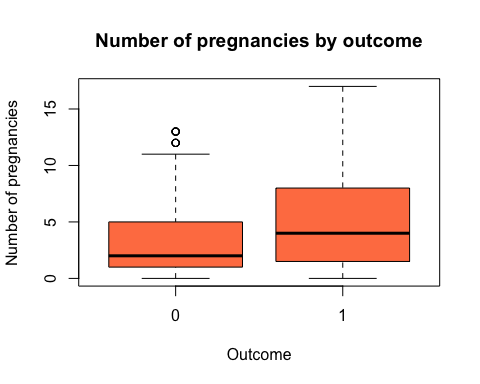
## [1] 4.865672

mean(diabetes$Pregnancies[diabetes$Outcome == 0])

## [1] 3.298

Looks like the more pregnancies you have, the more chances to have diabetes.Is that significant?

# Is the number of pregnancies on every diabetes outcome significantly different?  
  
boxplot(diabetes$Pregnancies~ diabetes$Outcome, main="Number of pregnancies by outcome", font.main =2, ylab="Number of pregnancies", xlab="Outcome", family="Palatino", col="coral")



diabetes$Outcome= factor(diabetes$Outcome, levels = c(0,1),  
labels = c("Diabetic", "No Diabetic"))   
str(diabetes)

## 'data.frame': 768 obs. of 9 variables:  
## $ Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ...  
## $ Glucose : int 148 85 183 89 137 116 78 115 197 125 ...  
## $ BloodPressure : int 72 66 64 66 40 74 50 0 70 96 ...  
## $ SkinThickness : int 35 29 0 23 35 0 32 0 45 0 ...  
## $ Insulin : int 0 0 0 94 168 0 88 0 543 0 ...  
## $ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...  
## $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...  
## $ Age : int 50 31 32 21 33 30 26 29 53 54 ...  
## $ Outcome : Factor w/ 2 levels "Diabetic","No Diabetic": 2 1 2 1 2 1 2 1 2 2 ...

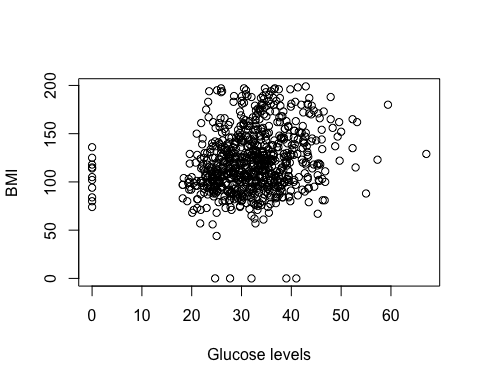
t.test(diabetes$Pregnancies~ diabetes$Outcome, data = diabetes) # perform t test

##   
## Welch Two Sample t-test  
##   
## data: diabetes$Pregnancies by diabetes$Outcome  
## t = -5.907, df = 455.96, p-value = 6.822e-09  
## alternative hypothesis: true difference in means between group Diabetic and group No Diabetic is not equal to 0  
## 95 percent confidence interval:  
## -2.089219 -1.046125  
## sample estimates:  
## mean in group Diabetic mean in group No Diabetic   
## 3.298000 4.865672

t.test(diabetes$Pregnancies~ diabetes$Outcome, data = diabetes)$p.value # get p value

## [1] 6.821926e-09

# ¿Is BMI correlated with glucose levels?  
  
# Correlation plot  
plot(diabetes$BMI, diabetes$Glucose,xlab="Glucose levels", ylab="BMI")



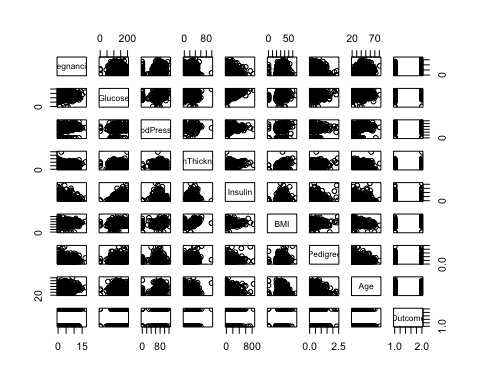
# Correlation test  
cor(diabetes$BMI, diabetes$Glucose)

## [1] 0.2210711

cor.test(diabetes$BMI, diabetes$Glucose, method="pearson") # there is significant positive correlation (p-value=5.891e-10, corr 0.2210711).

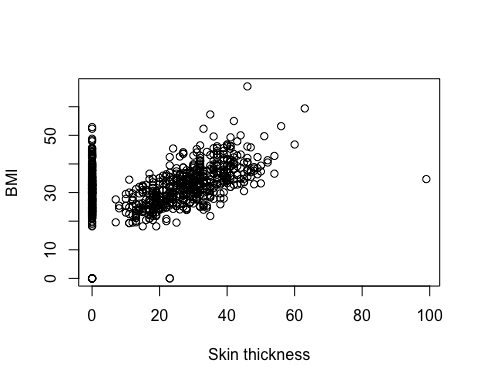
##   
## Pearson's product-moment correlation  
##   
## data: diabetes$BMI and diabetes$Glucose  
## t = 6.2737, df = 766, p-value = 5.891e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1527152 0.2873218  
## sample estimates:  
## cor   
## 0.2210711

# We can perform correlation plots for all the columns as an overview  
pairs(diabetes)



We can see skin thikness and BMI looks like they correlate too. Let us have a look.

plot(diabetes$SkinThickness, diabetes$BMI,xlab="Skin thickness", ylab="BMI") # A few 0 values that can´t be correct. Let us remove these.



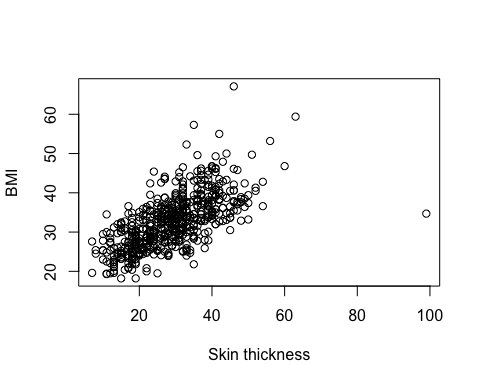
# Convert 0s to NA  
diabetes[diabetes==0]= NA  
  
# Subset dataset to skin thikness and BMI  
diabetes2= diabetes[,c(4,6)]  
diabetes2= diabetes2[complete.cases(diabetes2),]  
dim(diabetes)

## [1] 768 9

dim(diabetes2) # We go from 768 patients to 539

## [1] 539 2

# Let us plot again  
plot(diabetes$SkinThickness, diabetes$BMI,xlab="Skin thickness", ylab="BMI") # A few 0 values that can´t be correct. Let us remove these.



# Correlation test  
cor(diabetes2$SkinThickness, diabetes2$BMI)

## [1] 0.6482139

cor.test(diabetes2$SkinThickness, diabetes2$BMI, method="pearson") # there is significant positive correlation (p-value< 2.2e-16, corr 0.6482139).

##   
## Pearson's product-moment correlation  
##   
## data: diabetes2$SkinThickness and diabetes2$BMI  
## t = 19.727, df = 537, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.5964087 0.6946414  
## sample estimates:  
## cor   
## 0.6482139

# Do woman with just one pregnancy have higher incidence of Diabetes compared to more than one pregnancy?  
  
# We can just do:  
summary(diabetes$Outcome[diabetes$Pregnancies ==1])

## Diabetic No Diabetic NA's   
## 106 29 111

summary(diabetes$Outcome[diabetes$Pregnancies != 1])

## Diabetic No Diabetic NA's   
## 321 201 111

# Are these ratio significantly different?  
# it looks like there are more diabetic cases among women with just one pregnancy.