DATA ANALYTICS REPORT

Damien Matias - Lagrenaudie Victor

Cleaning of data

First thing, with our dataset we needed to clean it because there were some wrong values among the excel/csv file. For this purpose we implemented a cleaning function 'cleandata', within the cleaning.R file.

To begin, we cleaned the the weight column. Indeed, sometimes the suffix 'kg' was written, so we made it disappear.

```
new_weight <- c()
for (weights in mydata$How.much.do.you.weigh...kg.) {
  if (endsWith(weights, "kg")) {
    withspaces <- substr(weights, 0, nchar(weights)-2)
    weights <- gsub(" ", "", withspaces, fixed = TRUE)
  }
  new_weight <- append(new_weight, strtoi(weights))
}
mydata$How.much.do.you.weigh...kg. <- new_weight
```

Then we moved to the height. In the dataset we erased data with the suffix 'cm' , and when the height was below 100 cm we assumed it's an error and add 100cm to have a '"human" height.

```
new_height <- c()
for (heights in mydata$What.is.your.height...cm.) {
  if(!is.na(heights)) {

    if (endsWith(heights, "cm")) {
      withspaces <- substr(heights, 0, nchar(heights)-2)
      heights <- gsub(" ", "", withspaces, fixed = TRUE)
    }
    if(heights=="1.60") {
      heights <- 160
    }

    heights <- strtoi(heights)
    #print(heights)
    if(heights < 100) {
      heights <- heights + 100
    }
  }
```

We then corrected the value among the column relative to the number of smoke each day. Indeed, the excel file understood the value 11-20 as the 20th of November, so we corrected it.

```
#How.many.cigarettes.do.you.smoke.each.day.
new_howmanycig <- c()
for (number in mydata$How.many.cigarettes.do.you.smoke.each.day.) {
  if (number == "nov-20") {
    number <- "11-20"
  }
  new_howmanycig <- c(new_howmanycig, number)
}
mydata$How.many.cigarettes.do.you.smoke.each.day. <- factor(new_howmanycig)
```

Then, for the phones, some of them had wrong syntaxe, so we considered every phone beginning with an "I" would be iPhone, the ones beginning with A will be Android and the other regrouped into a category "Other". For the monthly salary column, we corrected the missing value followed by P such as it would be understood as "Prefer not to say".

```
#What.type.of.phone.do.you.have.
new_phone <- c()
for (phone in mydata$What.type.of.phone.do.you.have.) {
  if (startsWith(phone, "i")) {
    phone <- "iPhone"
  }
  else if (startsWith(phone, "A")) {
    phone <- "Android"
  }
  else {
    phone <- "Other"
  }
  new_phone <- c(new_phone, phone)
}
mydata$What.type.of.phone.do.you.have. <- factor(new_phone)

#How.much.salary.do.you.earn.each.month.
new_salary <- c()
for (salary in mydata$How.much.salary.do.you.earn.each.month.) {
  if (startsWith(salary, "P")) {
    salary <- 'Prefer not to say'
  }

  new_salary <- c(new_salary, salary)
}
mydata$How.much.salary.do.you.earn.each.month. <- factor(new_salary)
```

We categorized health description in 3 groups : healthy, minor problem and major problem, depending on the beginning of the answer.

```
#How.do.you.describe.your.health.
new_health <- c()
for (health in mydata$How.do.you.describe.your.health.) {
  if (startsWith(health, "I am")) {
    health <- 'Healthy'
  }
  else if (startsWith(health, "I have major")) {
    health <- 'Major problem'
  }
  else if (startsWith(health, "I have minor")) {
    health <- 'Minor problem'
  }

  new_health <- c(new_health, health)
}
mydata$How.do.you.describe.your.health. <- factor(new_health)
```

We applied the same thing for the “reduce smoking” feature, same thing for education.

```
#Do.you.prefer.to.quit.or.to.reduce.smoking.
new_quit <- c()
for (quit in mydata$Do.you.prefer.to.quit.or.to.reduce.smoking.) {
  if (startsWith(quit, "I would like to q")) {
    quit <- 'Quit smoking'
  }
  else if (startsWith(quit, "I would like to r")) {
    quit <- 'Reduce smoking'
  }
  else if (startsWith(quit, "I am happy")) {
    quit <- 'Don t change anything'
  }
  else {
    quit <- ''
  }

  new_quit <- c(new_quit, quit)
}
mydata$Do.you.prefer.to.quit.or.to.reduce.smoking. <- factor(new_quit)
```

```r
#Education
new_education <- c()
for (edu in mydata$Education) {
  if (startsWith(edu, "Graduate degree")) {
    edu <- 'Graduate degree'
  }
  else if (startsWith(edu, "High school")) {
    edu <- 'High school'
  }
  else if (startsWith(edu, "Undergraduate degree")) {
    edu <- 'Undergraduate degree'
  }
  else {
    edu <- ''
  }

  new_education <- c(new_education, edu)
}
mydata$Education <- factor(new_education)
```

We rounded up the BMI (to have 21 instead of 21,324 or 23 instead of 22,785 etc…)

```r
#BMI
mydata$BMI <- round(mydata$How.much.do.you.weigh...kg./((mydata$What.is.your.height...cm./100)**2))
```

We renamed the column to have clearer and shorter labels.

```r
#Renaming columns
names(mydata)[names(mydata)=="How.do.you.describe.your.health."] <- "Health.category"
names(mydata)[names(mydata)=="How.much.do.you.weigh...kg."] <- "Weight"
names(mydata)[names(mydata)=="What.is.your.height...cm."] <- "Height"
names(mydata)[names(mydata)=="Do.you.have.or.have.you.had.any.of.the.below.health.conditions...sel
names(mydata)[names(mydata)=="At.what.age.you.started.to.smoke.regularly."] <- "Age.started.smokin
names(mydata)[names(mydata)=="How.many.cigarettes.do.you.smoke.each.day."] <- "Cigarettes.each.day
names(mydata)[names(mydata)=="How.soon.after.you.wake.up.do.you.smoke.your.first.cigarette."] <- "
names(mydata)[names(mydata)=="When.did.you.last.try.to.quit.smoking."] <- "When.last.try"
names(mydata)[names(mydata)=="What.method.did.you.try.to.quit.smoking.before...select.all.that.app
names(mydata)[names(mydata)=="Did.you.manage.to.quit.smoking.using.that.method."] <- "Success.quit
names(mydata)[names(mydata)=="How.would.you.categorize.your.friends."] <- "Categorize.friends"
names(mydata)[names(mydata)=="How.would.you.categorize.your.family."] <- "Categorize.family"
names(mydata)[names(mydata)=="What.is.the.brand.of.your.cigarettes."] <- "Brand.cigarettes"
names(mydata)[names(mydata)=="Which.type.of.cigarettes.box.do.you.buy."] <- "Type.cigarettes"
names(mydata)[names(mydata)=="How.important.is.having.your.own.lighter.in.your.smoking.process.exp
names(mydata)[names(mydata)=="What.type.of.phone.do.you.have."] <- "Type.phone"
names(mydata)[names(mydata)=="Where.do.you.live."] <- "Where.living"
names(mydata)[names(mydata)=="How.much.salary.do.you.earn.each.month."] <- "Salary"
names(mydata)[names(mydata)=="Do.you.prefer.to.quit.or.to.reduce.smoking."] <- "Quit.reduce"
names(mydata)[names(mydata)=="Why.do.you.want.to.reduce.quit.smoking."] <- "Why.quit"
```

Some statistics and associations

After all this cleaning, we are now able to display some interesting value of our dataset, such as the mean age, some statistics about it, same thing for value depending on the whole dataset or the age.

```
source("cleaning.R")
mydata=read.csv("Dataset.csv",header=TRUE, sep=";" ,na.strings=" ")
cleaned=cleandata(mydata)
summary(cleaned)

#Some computation
age_mean = mean(cleaned$Age)
age_sd = sd(cleaned$Age)
women = cleaned[cleaned$Gender=="Female",]
men = cleaned[cleaned$Gender=="Male",]

summary(cleaned)
summary(women)
summary(men)
```

For the women we have :

```
      Age                       Education  Family.status      Weight          Height     
 Min.   :20.00   Graduate degree     :15   Married:27    Min.   :42.00   Min.   :145.0  
 1st Qu.:29.00   High school         : 9   Single :42    1st Qu.:55.00   1st Qu.:160.0  
 Median :35.00   Undergraduate degree:45                 Median :62.00   Median :165.0  
 Mean   :38.12                                           Mean   :62.64   Mean   :164.7  
 3rd Qu.:47.00                                           3rd Qu.:67.00   3rd Qu.:168.0  
 Max.   :62.00                                           Max.   :95.00   Max.   :183.0
```

and the men we have :

```
      Age                       Education  Family.status      Weight          Height     
 Min.   :18.00   Graduate degree     :64   Married:101   Min.   : 59.00   Min.   :160.0  
 1st Qu.:29.00   High school         :31   Single : 88   1st Qu.: 76.00   1st Qu.:173.0  
 Median :37.00   Undergraduate degree:94                 Median : 85.00   Median :178.0  
 Mean   :37.84                                           Mean   : 85.84   Mean   :177.7  
 3rd Qu.:45.00                                           3rd Qu.: 95.00   3rd Qu.:182.0  
 Max.   :73.00                                           Max.   :135.00   Max.   :196.0
```

Then for our association rules we loaded the arules library, which allows us for instance to use the “apriori” function.

Our goal in our first try, is to find association rules with the health category in mind. We choose several category based on what we thought could have an influence on the health.

```r
# Association rules
library(arules)

#1st try (Major problem)
first <- cleaned[,c("Age.cat", "Age.started.smoking.cat", "Cigarettes.each.day", "First.cigarette", "

rules1 <- apriori(first, control = list(verbose=T), parameter = list(minlen=2, supp=0.005, conf=0.8),
quality(rules1) <- round(quality(rules1), digits = 3)
rules1.sorted <- sort(rules1, by="lift")
inspect(rules1.sorted)

#2nd try (Healthy)
second <- cleaned[,c("Age.cat", "Age.started.smoking.cat", "Cigarettes.each.day", "First.cigarette",

rules2 <- apriori(second, control = list(verbose=T), parameter = list(minlen=2, supp=0.005, conf=0.8,
quality(rules2) <- round(quality(rules2), digits = 3)
rules2.sorted <- sort(rules2, by="lift")
inspect(rules2.sorted)

#3rd try (Phones)
third <- cleaned[,c("Gender", "Education","Type.phone", "Salary", "BMI.cat")]

rules3 <- apriori(third, control = list(verbose=T), parameter = list(minlen=2, supp=0.005, conf=0.8,
quality(rules3) <- round(quality(rules3), digits = 3)
rules3.sorted <- sort(rules3, by="lift")
inspect(rules3.sorted)
```

We obtained 210 associative rules and the one that we thought interesting was the only one that converged to the “Health.category => major problem”. In fact, there is very low number of people with major problem. This is why it doesn’t appear a lot in the rules.

```
     lhs                                          rhs                          support confidence   lift count
[1]  {Age.cat=[54.7,73.0],
      Age.started.smoking.cat=[ 0,21),
      First.cigarette=5-30 minutes,
      When.last.try=Over 5 years ago}          => {Health.category=Major problem}   0.008      1.000 14.333     2
```

We can see that there is a huge lift so it means that this rules is very important. This type a person is smoking since he’s a teenager, he’s old and he starts to smoke very early in the day.

For the next one we added the BMI and the salary to see if this two features have an influence on the health.

Here’s a glimpse of what we found :

```
[139] {Age.cat=[18.0,36.3),
       Salary=$1,000-$5,000,
       BMI.cat=Normal}                         => {Health.category=Healthy}         0.058      0.938 1.531    15
[140] {Cigarettes.each.day=10 or fewer,
       BMI.cat=Normal}                         => {Health.category=Healthy}         0.050      0.929 1.516    13
[141] {Age.started.smoking.cat=[ 0,21),
       Cigarettes.each.day=10 or fewer,
       BMI.cat=Normal}                         => {Health.category=Healthy}         0.047      0.923 1.507    12
[142] {Age.cat=[36.3,54.7),
       Cigarettes.each.day=10 or fewer}        => {Health.category=Healthy}         0.043      0.917 1.497    11
```

So basically if you’re young, have enough money and in good shape you have a good chance to be healthy.

The last was more of a funny one, we wanted to see what type of people have an iPhone, here's the results :

```
{Education=Undergraduate degree,Salary=Below $1,000,BMI.cat=Normal}          => {Type.phone=iPhone}  0.016   1.000      1.583  4
{Gender=Female,Education=Graduate degree,Salary=$5,000-$10,000}              => {Type.phone=iPhone}  0.016   1.000      1.583  4
{Education=High school,Salary=$10,000 and more,BMI.cat=Normal}               => {Type.phone=iPhone}  0.008   1.000      1.583  2
{Education=Graduate degree,Salary=$10,000 and more,BMI.cat=Obesity}          => {Type.phone=iPhone}  0.008   1.000      1.583  2
{Education=Graduate degree,Salary=$10,000 and more,BMI.cat=Overweight}       => {Type.phone=iPhone}  0.035   1.000      1.583  9
{Education=Undergraduate degree,Salary=$10,000 and more,BMI.cat=Normal}      => {Type.phone=iPhone}  0.019   1.000      1.583  5
{Education=Undergraduate degree,Salary=$10,000 and more,BMI.cat=Overweight} => {Type.phone=iPhone}  0.019   1.000      1.583  5
```

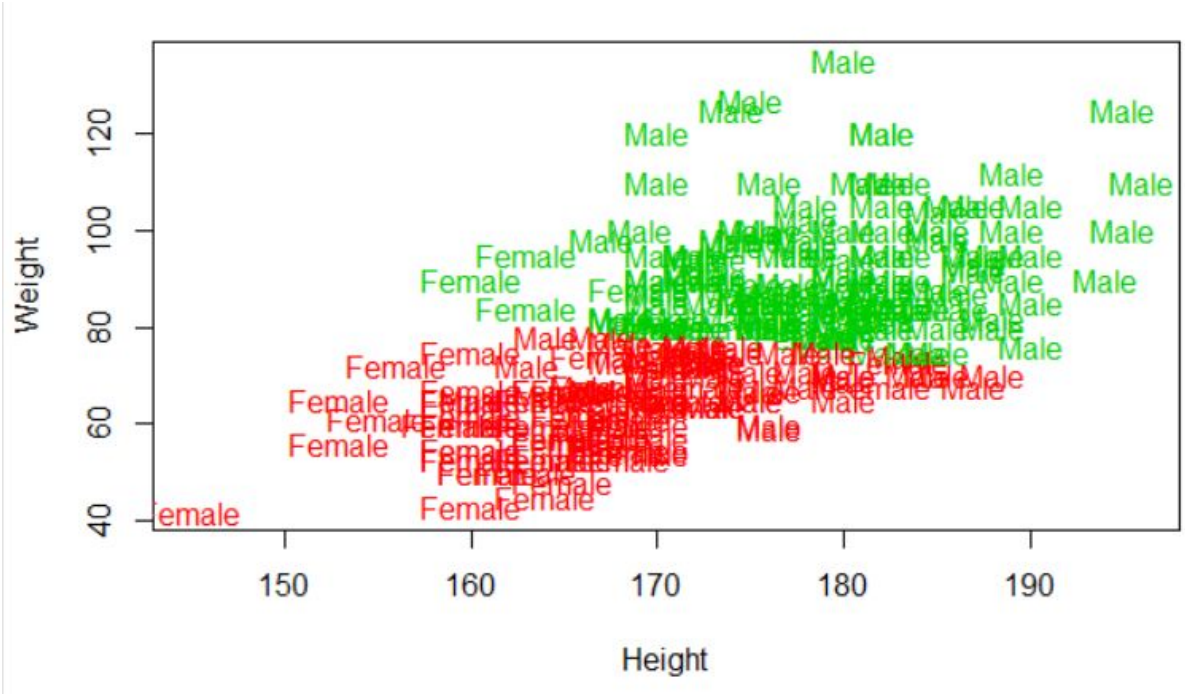
We can say that if you're at least in overweight, earn a lot of money and have a minimum of education

Clustering

To finish, we implement some clustering. We tried several clustering like you can see below.

```
source("cleaning.R")
mydata=read.csv("Dataset.csv",header=TRUE, sep=";" ,na.strings=" ")
cleaned=cleandata(mydata)
summary(cleaned)

clustered <- kmeans(cleaned[,c("Height","Weight")], centers=2, nstart=10)
plot(cleaned$Height, cleaned$Weight, type="n", xlab="Height", ylab="Weight")
text(x=cleaned$Height, y=cleaned$Weight, labels = cleaned$Gender, col = clustered$cluster+1)
```



We tried some clustering but the rendering was not so good :