Survey of various clients: Cleaning, NA`s, Linear Regression and Plots

Miguel Angel Bustos Sáez

12/7/2021

HarvardX Capstone PH125.9X

Table of Contents

Introduction……………………………………………………………………………………………………………….3

[1. Structure people and create people1 data set 4](#_Toc89937252)

[Structure 4](#_Toc89937253)

[People1 4](#_Toc89937254)

[2. Variables data set classes and modeling some information 5](#_Toc89937255)

[Working with NA´s 5](#_Toc89937256)

[Working with outliers 5](#_Toc89937257)

[Eliminate the registers that are grater than or equal to 100 ages and plot it 6](#_Toc89937258)

[Users that compleated age in minus cero, minus 18 and plot 7](#_Toc89937259)

[3. NA´s Global Exploratory 8](#_Toc89937260)

[Creating a new variable called “Education Level” 8](#_Toc89937261)

[Working with the 467 NA´s observations of people1 data set 8](#_Toc89937262)

[4.NA´s replacement with Critical Thinking 9](#_Toc89937263)

[The mean and the median, central tendency measures 9](#_Toc89937264)

[Delete NA´s 9](#_Toc89937265)

[Without answer data 12](#_Toc89937266)

[5. Inference, new variable products and plots 13](#_Toc89937267)

[Who answered the survey? 13](#_Toc89937268)

[What`s the income of “Employees” and “Do not work”? 14](#_Toc89937269)

[Creating variables magazines and price to offer to survey people 15](#_Toc89937270)

[6. Linear regression and plots 17](#_Toc89937271)

[Linear models 17](#_Toc89937272)

[Linear model chosen 18](#_Toc89937273)

[Data set survey concentration 21](#_Toc89937274)

[Main Businesses Periscope 30](#_Toc89937275)

[7. CONCLUSION AND SALES RECOMENDATION 32](#_Toc89937276)

# Introduction

This is the second of two Capstone Projects, after approving eight R programming language courses, corresponding to finish the program of EDX, and to obtain the HarvardX Professional Certificate in Data Science.

This report, it´s about the problem of cleaning data and lead with NA´s observations, in a simulation of a Survey that answered people of different countries. Linear regression and RMSE or Root-Mean Square of Error are good topics in this data and some plots that shows a good treatment in the data set.

The treatment of NA´s is usually a frustrated situation for data analysts or data scientist and they must take decisions to remove or replace the null or NA observations, it´s very important have a critical thinking in a realistic way to do things, because if the elimination of NA`s could change the result of all data set and in consequences, in different departments in the organization, the good manage the information it´s critical because the information it´s probably one of the most important assets that have a company: Clients, invoices, financial reports, data bases, sales, and in this project, a survey, what will be the correct treatment, make inferences, linear regression, RMSE and take decisions that who will be the selection of the first clients that must receive the magazine and the information of the company.

Virtual content; videos or read services, it`s a great market niche, it because doesn´t have a strong starter capital to make a start up, is more, design the video or the magazine, and make management in social media, videos, content, and the good inference and management of client information, and in that last point, this project is focused on, receiving the data, cleaning the data set, make inferences, plotting changes, make linear regression and identify the main market niche and clients.

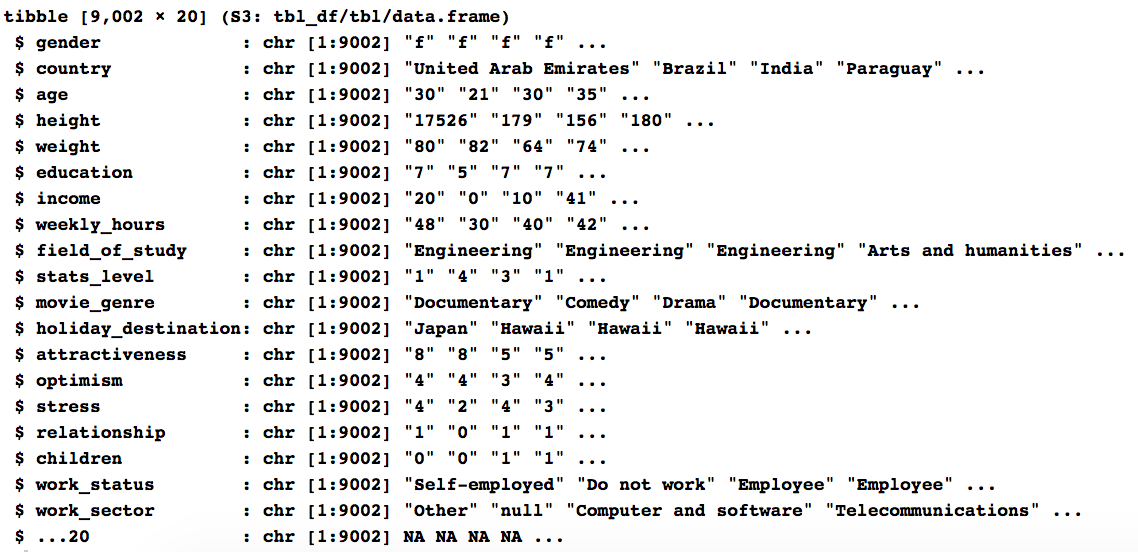
In companies, the main objective, is reach good clients and have a good sales and incomes, for that objective, make inferences in the clients and view the data in plots, it is the best way, before take any decision and the data analysis it`s critical for the future of any company.

# 1. Structure people and create people1 data set

### Structure

f for female, m for male, NA´s and a lot of outliers were found in the survey

str(people)

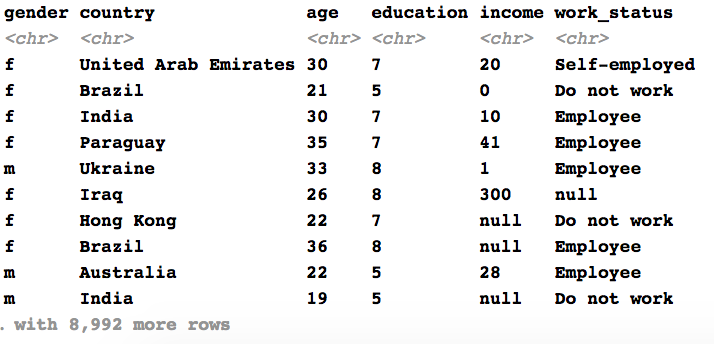


### 

### People1

Creating people1 data set and the essentials variables that must be cleaned

people1 <- select(people, "gender", "country", "age", "education", "income", "work\_status")



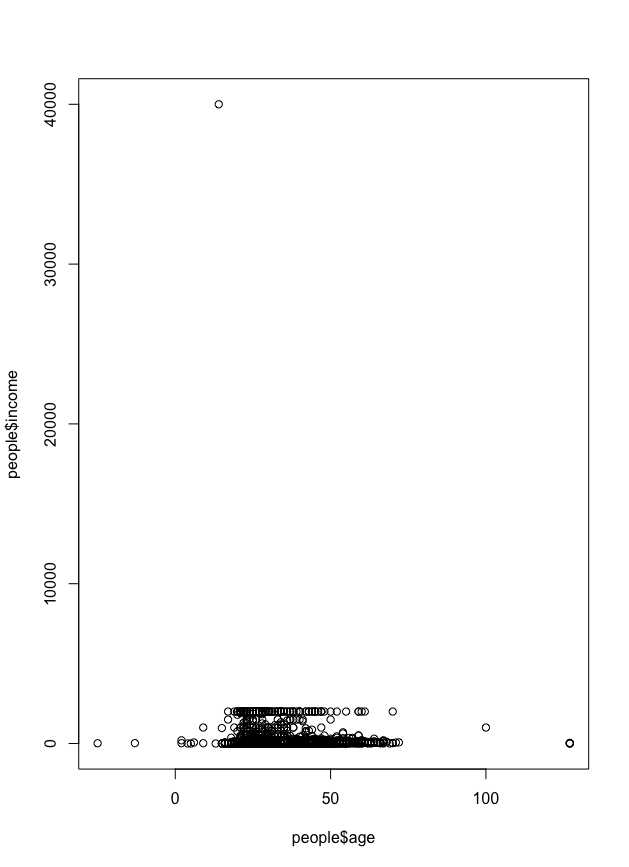
# 2. Variables data set classes and modeling some information

### Working with NA´s

|  |  |
| --- | --- |
| people1$age <- as.numeric(as.character(people1$age)) people1$education <- as.numeric(as.character(people1$education)) people1$income <- as.numeric(as.character(people1$income)) sapply(people1, class) mean(people1$age, na.rm = TRUE)  sapply(people1, class)  plot(people1$age, people1$income) |  |

### Working with outliers

plot(people1$age, people1$income)



which.max(people1$income)   
people1[9002,]

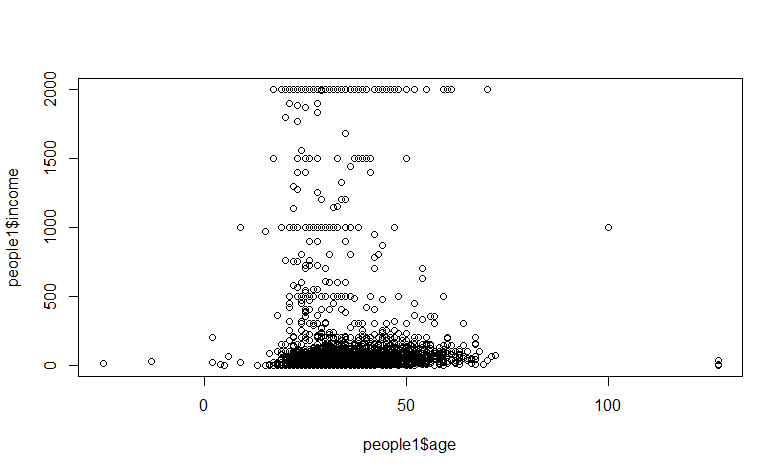


people1 <- people1[-c(9002),]

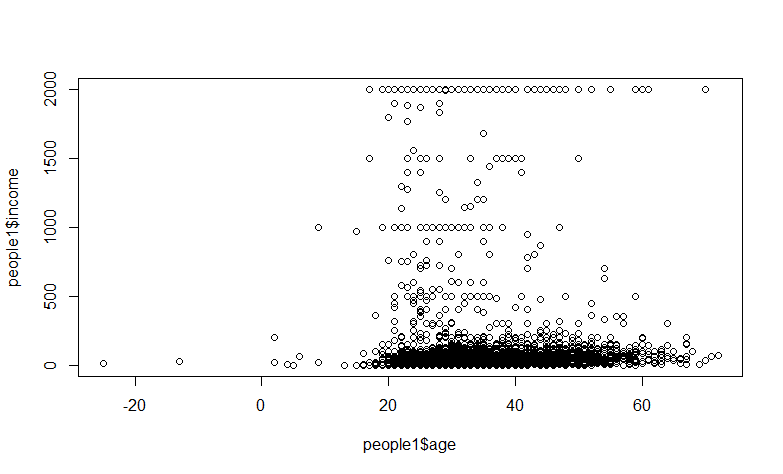
### Eliminate the registers that are grater than or equal to 100 ages and plot it

plot(people1$age, people1$income)

Removing 9002 row, it`s the way to remove the outlier at the top. But it has in the left and in the right:



which(people1$age >= 100)  
people1[c(212, 1548, 1601, 2331, 4278, 7898),]  
people1 <- people1[-c(212, 1548, 1601, 2331, 4278, 7898),]  
plot(people1$age, people1$income)



### Users that completed age in minus cero, minus 18 and plot

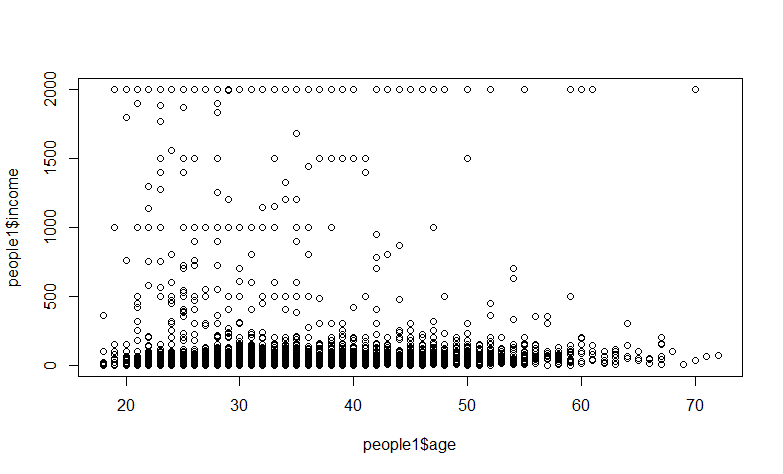
Because the magazines are for business men and woman, the decision is focusing in more than 18 years old clients:

which(people1$age < 0)  
people1[c(2906, 5013, 5814, 5990, 6572, 8749),]

Eliminate registers that are under 18 years old and plot

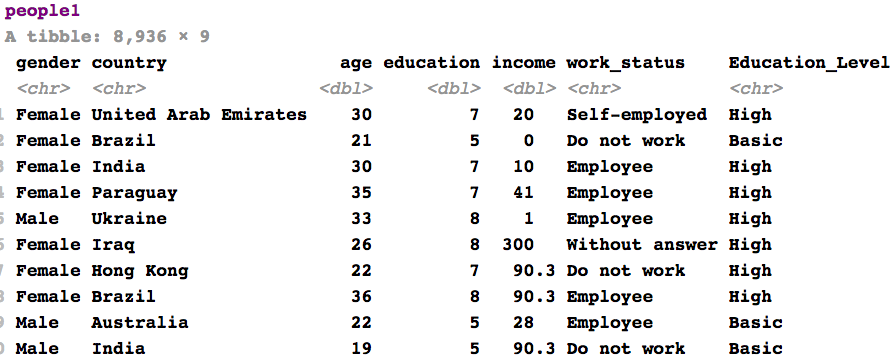
which(people1$age < 18)  
people1[c(223, 305, 440, 1247, 1397, 1415, 1466, 1507, 1713, 1718, 2049, 2593, 2703, 2739, 2859, 2870, 2906, 2955, 3268, 3396, 3434, 3575, 3837, 4031, 4350, 4450, 4738, 4888, 4922, 4987, 5013, 5137, 5298, 5388, 5695, 5733, 5814, 5890, 5904, 5990, 6026, 6080, 6375, 6455, 6572, 6785, 6930, 7072, 7229, 7509, 7712, 7875, 8076, 8224, 8580, 8643, 8749, 8751, 8896),]  
  
people1 <- people1[-c(223, 305, 440, 1247, 1397, 1415, 1466, 1507, 1713, 1718, 2049, 2593, 2703, 2739, 2859, 2870, 2906, 2955, 3268, 3396, 3434, 3575, 3837, 4031, 4350, 4450, 4738, 4888, 4922, 4987, 5013, 5137, 5298, 5388, 5695, 5733, 5814, 5890, 5904, 5990, 6026, 6080, 6375, 6455, 6572, 6785, 6930, 7072, 7229, 7509, 7712, 7875, 8076, 8224, 8580, 8643, 8749, 8751, 8896),]

plot(people1$age, people1$income)



### Creating a new variable called “Education Level”

people1 <- people1 %>%  
 mutate(Education\_Level = case\_when(  
 education == 0 ~ "No education",  
 education >= 1 & education <= 6 ~ "Basic",  
 education >= 7 ~ "High"))



# 3. NA´s Global Exploratory

### Working with the 467 NA´s observations of people1 data set

which(is.na(people1))   
sum(is.na(people1))

colSums(is.na(people1))



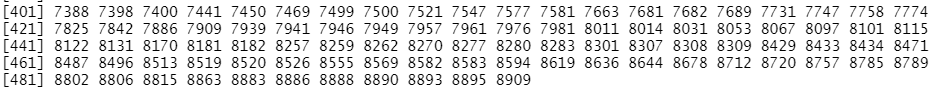
# 4.NA´s replacement with Critical Thinking

### 

### The mean and the median, central tendency measures

491 NA´s were found in age, without these NA`s, age variable has a mean of 30.13061 and a median of 28

which(is.na(people1$age))



mean(people1$age, na.rm = TRUE)

median(people1$age, na.rm = TRUE)

### 

### Delete NA´s

All 491 NA´s were delete with this code:

people1$age[is.na(people1$age)] <- mean(people1$age, trim = 0, na.rm = TRUE)

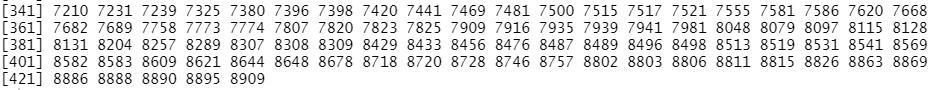
And the result is: integer(0) because were eliminated the NA´s

which(is.na(people1$age))



425 NA´s were found in education, the mean is 8.265656 and the median is 7

which(is.na(people1$education))



mean(people1$education, na.rm = TRUE)

median(people1$education, na.rm = TRUE)

All 425 NA´s were delete with this code:

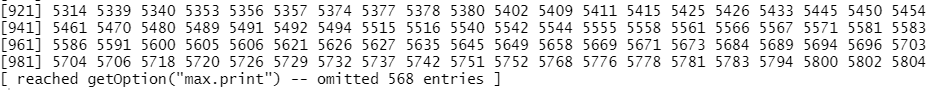
people1$education[is.na(people1$education)] <- mean(people1$education, na.rm = TRUE)

which(is.na(people1$education))



1568 NA´s were found in income variable, the mean is 90 and the median is 23

which(is.na(people1$income))



mean(people1$income, na.rm = TRUE)

median(people1$income, na.rm = TRUE)

All 1568 NA´s were delete with this code:

people1$income[is.na(people1$income)] <- mean(people1$income, na.rm = TRUE)

which(is.na(people1$income))

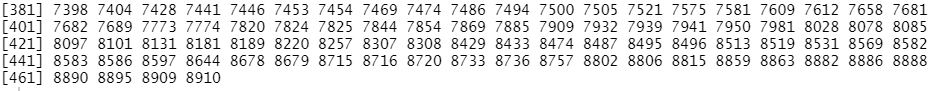


In gender variable, 0 NA were found, but has m and f values, let`s change these for male and female

|  |  |
| --- | --- |
| which(is.na(people1$gender)) #0 NA´s were found people1$gender[people1$gender == "m"] <- "Male" people1$gender[people1$gender == "f"] <- "Female" people1$gender[people1$gender == "null"] <- "Prefer not to say" |  |

464 NA`s were found in country variable

which(is.na(people1$country))



Setting all NA`s will be categorized as “World”:

people1$country[people1$country == "NA"] <- "World"  
people1$country[is.na(people1$country)] <- "World"

And now we have zero NA in country:

which(is.na(people1$country))



### 

### Without answer data

The variable work\_status has a lot of null expressions, all these expressions will be change by “Without Answer”

which(is.na(people1$work\_status))

0, 1, 2, 3, 4, 6, 11 These were changed by others, and “null” by “Without answer”:

people1$work\_status[people1$work\_status == "null"] <- "Without answer"  
people1$work\_status[people1$work\_status == 0] <- "Others"  
people1$work\_status[people1$work\_status == 1] <- "Others"  
people1$work\_status[people1$work\_status == 2] <- "Others"  
people1$work\_status[people1$work\_status == 3] <- "Others"  
people1$work\_status[people1$work\_status == 4] <- "Others"  
people1$work\_status[people1$work\_status == 6] <- "Others"  
people1$work\_status[people1$work\_status == 11] <- "Others"

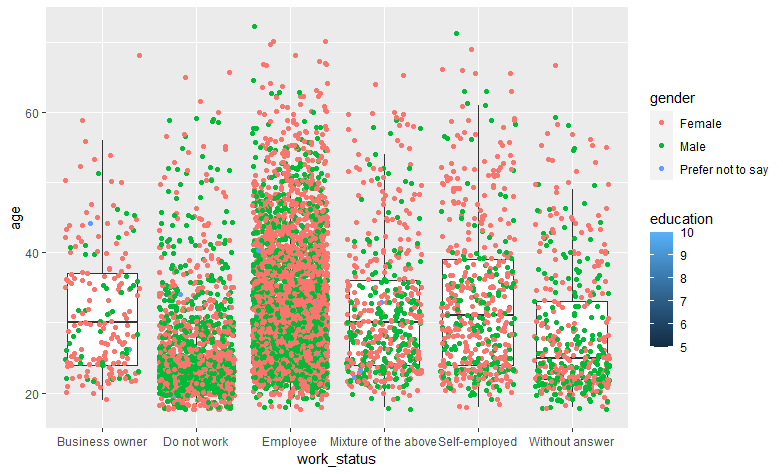
# 

# 5. Inference, new variable products and plots

### Who answered the survey?

“Employees” and “Do not work” people:

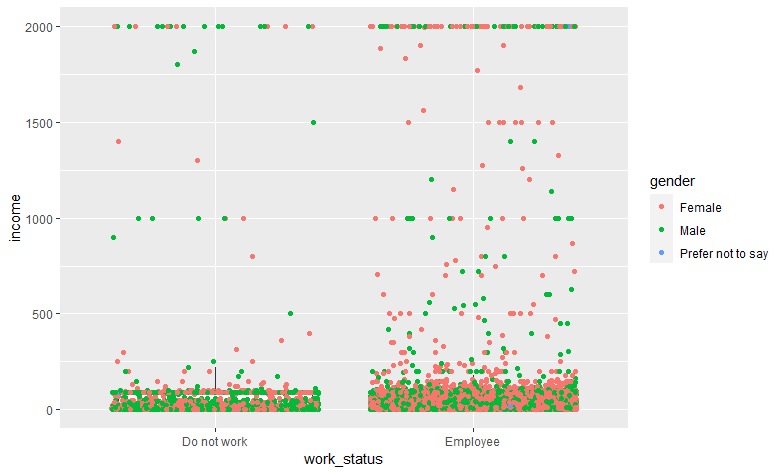
people1 %>% filter(age %in% 18:100)%>%   
 ggplot(aes(x = work\_status, y = age, fill = education))+  
 geom\_boxplot(outlier.shape = NA)+  
 geom\_point(position = "jitter", aes(color = gender))



### 

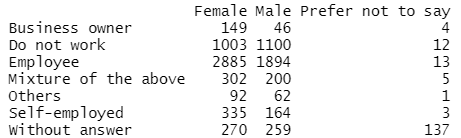
### What`s the income of “Employees” and “Do not work”?

people1 %>% filter(age %in% 18:100 & work\_status %in% c("Employee", "Do not work"))%>%   
 ggplot(aes(x = work\_status, y = income))+  
 geom\_boxplot(outlier.shape = NA)+  
 geom\_point(position = "jitter", aes(color = gender))



Sales focus on: “Employees” What course we can sale to them? Most of them are female:

table(people1$work\_status, people1$gender)



### Creating variables magazines and price to offer to survey people

The Sales Manager decision was prepare Business Magazines to people that answered the survey

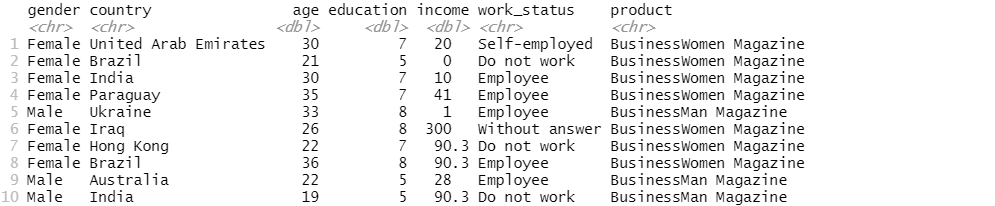
sample1 = people1 %>% filter(age %in% 30:35 & work\_status %in% c("Business owner", "Employee", "Self-employed"))

BM <- c("BusinnessMen Magazine")  
BMP <- c(250)  
Business\_Men\_Magazine <- data.frame(BM, BMP, stringsAsFactors = FALSE)

BW <- c("BusinessWomen Magazine")  
BWP <-c(200)  
Business\_Women\_Magazine <- data.frame(BW, BWP, stringsAsFactors = FALSE)

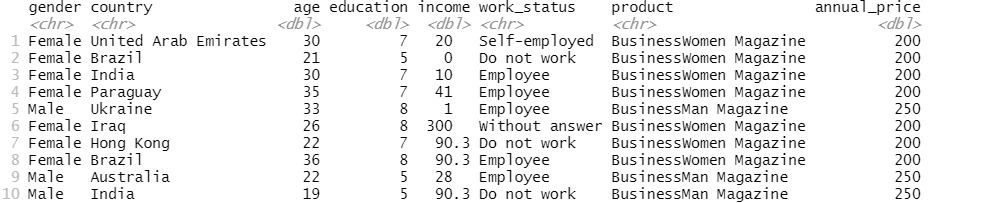
Now the Sales Manager wants to sale two magazines, one for women and another for men Here are a new variable column, with a magazine for women and men:

people1 <- people1 %>%   
 mutate(product = case\_when(  
 gender == "Female"~ "BusinessWomen Magazine",  
 gender == "Male" ~ "BusinessMan Magazine"))



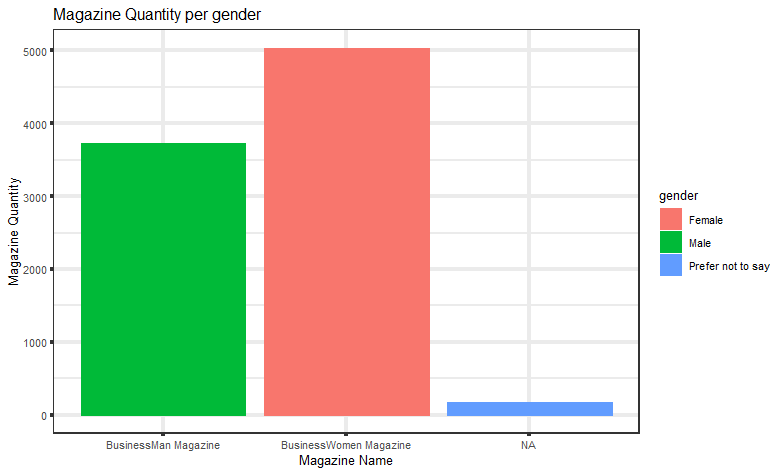
And the price of the Magazines will be USD 250 for Female and 200 dollars for Male, annually digital subscription:

people1 <- people1 %>%   
 mutate(annual\_price = case\_when(  
 gender == "Female" ~ 200,  
 gender == "Male" ~ 250)  
)



Now this is the quantity of Magazine for genre:

people1 %>% ggplot(aes(x = product)) +  
 geom\_bar(mapping = aes(x = product, color = gender, fill = gender))+  
 theme\_bw(base\_size = 10, base\_rect\_size = 1, base\_line\_size = 1.5)+  
 labs(y = "Magazine Quantity", x = "Magazine Name", title = "Magazine Quantity per gender")



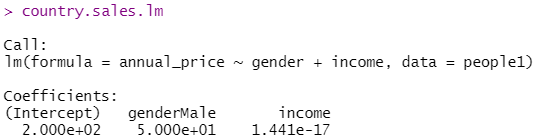
# 6. Linear regression and plots

Testing some lineal models country.sales.lm, country.sales.lm2 and country.sales.lm3, and choose just one, country.sales.lm3 because it shows country information.

### Linear models

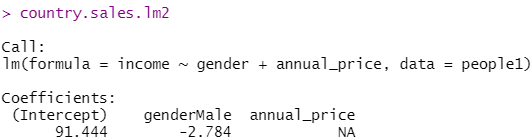
linear model for annual\_price as a function of gender and income:

country.sales.lm <- lm(annual\_price ~ gender + income, data = people1)



linear model for income as a function of gender and annual\_price:

country.sales.lm2 <- lm(income ~ gender + annual\_price, data = people1)



### 

### Linear model chosen

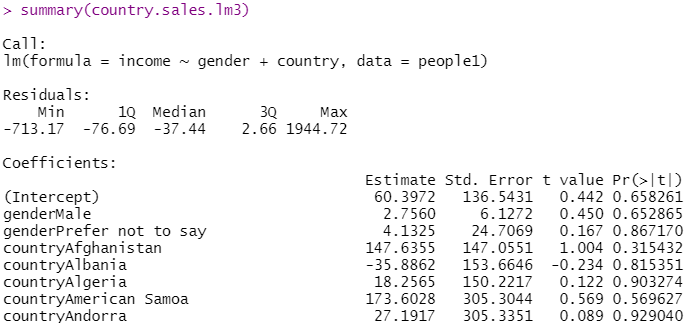
Now this linear model has been chosen, linear model for income as a function of gender and country:

country.sales.lm3 <- lm(income ~ gender + country, data = people1)



Tis summary of the country.sales.lm3 linear model shows residuals and coefficients per country:

summary(country.sales.lm3)



Based on the latest three tests, we can see that lm3, shows a relationship with the income, gender and country:

summary(country.sales.lm3)

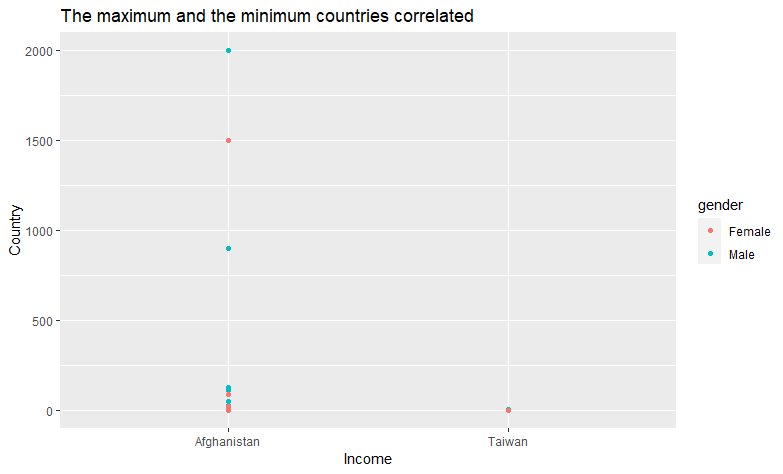
Interpretation and difference between Afghanistan and Taiwan, this example of these two countries because both of them has the biggest difference in the income, that has this data set people1:

The estimation of gender and country in Afghanistan is 147 and Taiwan is -55 This means that for every 1% of the gender there`s a correlated of 147% in Afghanistan in incomes and for everyone 1% in Taiwan there´s a correlated of -55.

The Standard error in Afghanistan is 147 and in Taiwan it is minus 142, notorious gap in both countries, and all others has over 100. The T-statistics or T-values are all in -0 and 1, exceptionally Madagascar that has 3.562 The p-values reflects these errors all over zero and there is almost zero probability that this effect is due to chance.

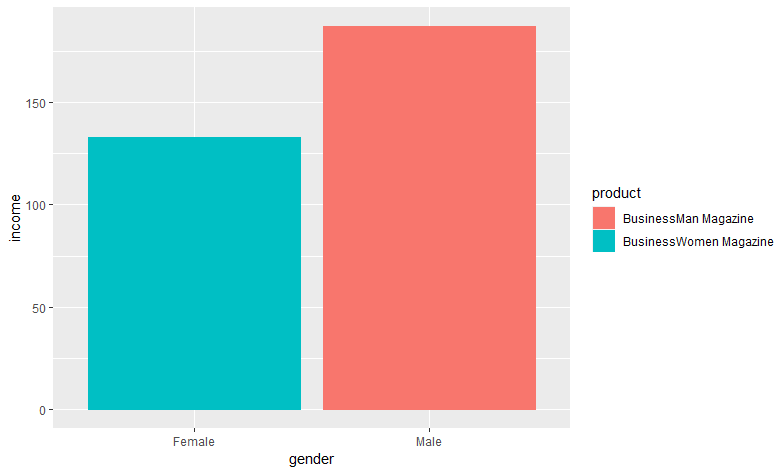
Afghanistan and Taiwan, here in this plot, we can see the enormous difference in income:

lm3plot <- people1 %>% select(country, income, gender) %>% filter(country %in% c("Afghanistan", "Taiwan"))  
lm3plot %>% ggplot(aes(country, income))+  
 geom\_point(aes(country, income, color = gender))+  
 labs(y = "Country", x = "Income", title = "The maximum and the minimum countries correlated")



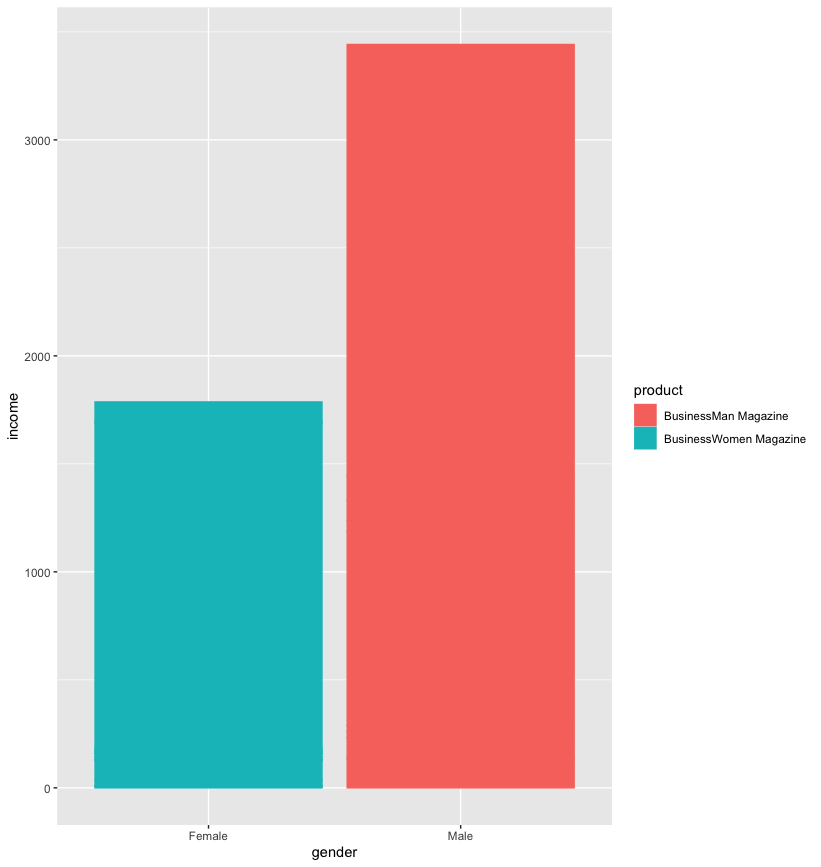
Taiwan Income, gender and product:

people1 %>% filter(country == "Taiwan") %>%   
 ggplot()+  
 geom\_bar(mapping = aes(x = gender, y = income, color = product, fill = product), stat = "identity")

Income until 150

Afghanistan Income, gender and product:

people1 %>% filter(country == "Afghanistan") %>%   
 ggplot()+  
 geom\_bar(mapping = aes(x = gender, y = income, color = product, fill = product), stat = "identity")

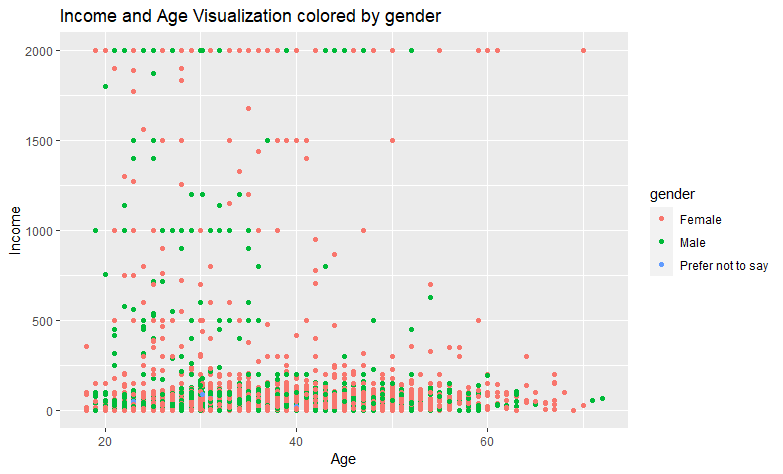
Income over 3000

### Data set survey concentration

In all people1 data set, the maximum concentration it`s less than 500 income between 20 and 40 years old.

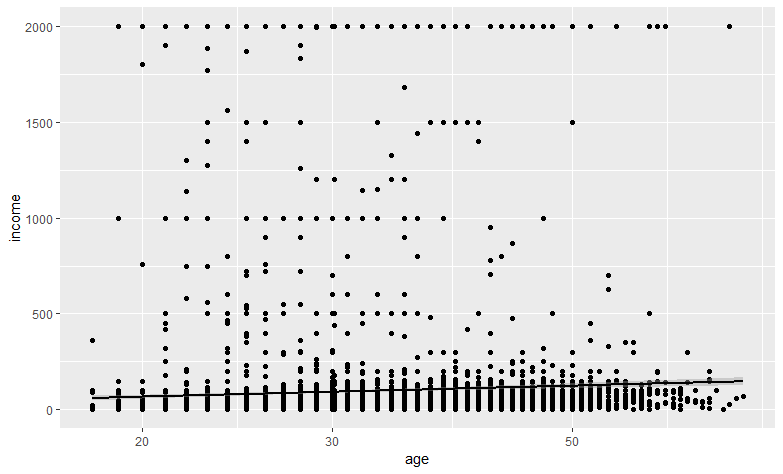
Let`s plot this:

ggplot(people1, aes(x = age, y = income))+  
 geom\_point(aes(x = age, y = income, color = gender))+  
 labs(y = "Income", x = "Age", title = "Income and Age Visualization colored by gender")



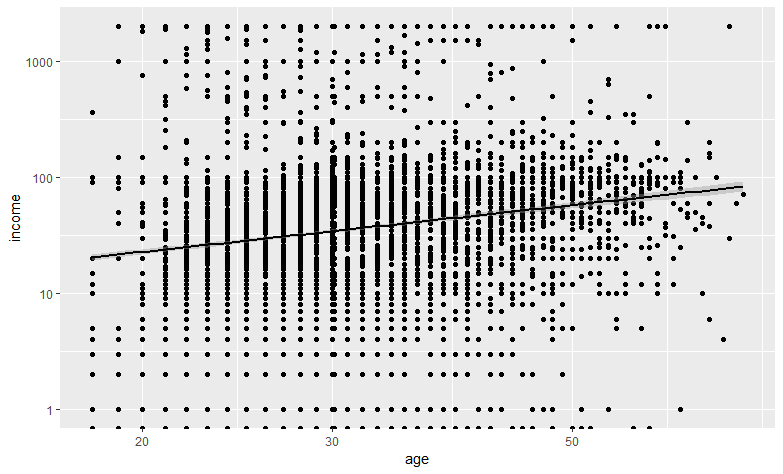
More age, more income, and the line:

lm3plot2 <- ggplot(people1, aes(age, income))+geom\_point()  
graphlm3 <- lm3plot2 + geom\_smooth(method = "lm", col = "black") + scale\_x\_log10()  
graphlm3



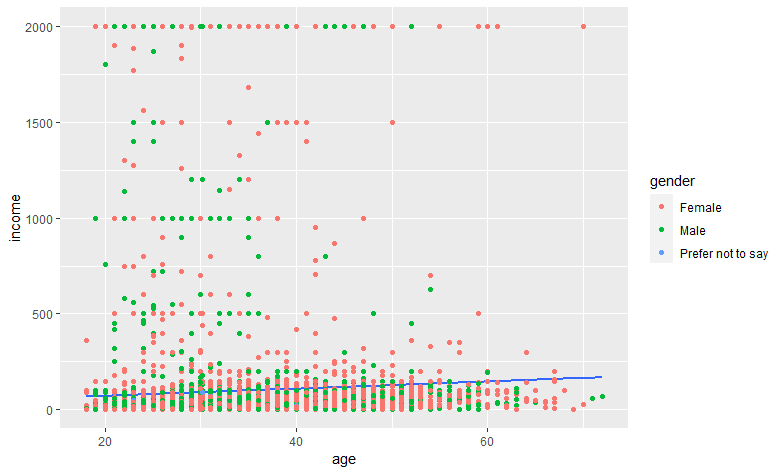
Obviously, y axis in this line is disproportionate, but anyway we can see the line better, this because has a transformation scale in the code:

graphlm3 <- lm3plot2 + geom\_smooth(method = "lm", col = "black") + scale\_x\_log10() + scale\_y\_log10()



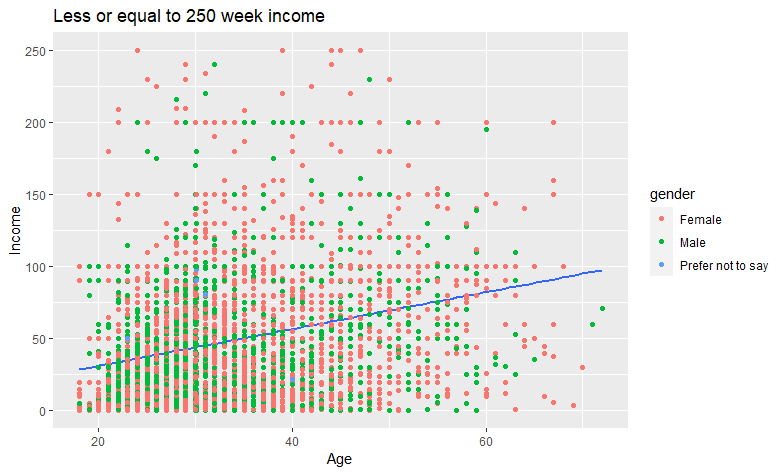
or colored but not with the transformation scale:

ggplot(data = people1, aes(age, income))+  
 geom\_smooth(method = lm, se = FALSE)+  
 geom\_point(aes(color = gender))



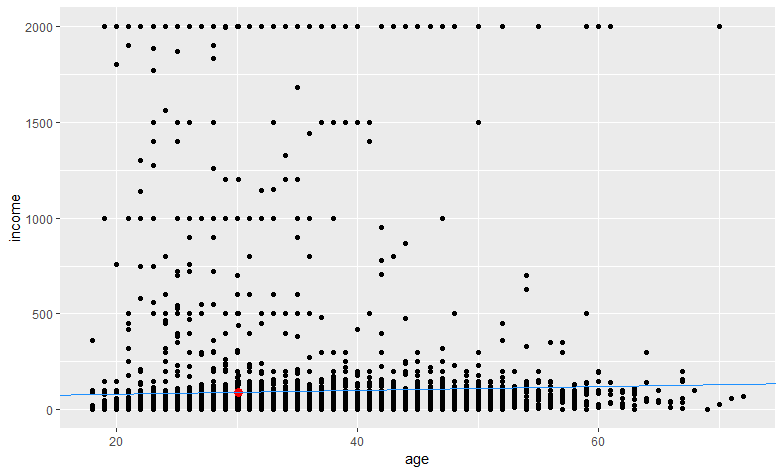
Much better if we take a sample more or equal to 250, the blue line has a better visualization:

people1 %>%   
 filter(income <= 250) %>% # Less or equal to 250 of income  
 ggplot(aes(x = age, income))+  
 geom\_smooth(method = lm, se = FALSE)+  
 geom\_point(aes(color = gender))+  
 labs(x = "Age", y = "Income", title = "Less or equal to 250 week income")



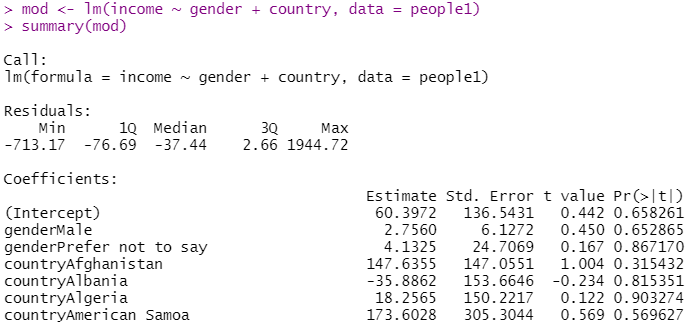
Now the linear regression line and a red point that intersect the mean average and the mean age:

add\_line <- function(slope\_people){  
 people1\_summary <- people1 %>%   
 summarize(N = n(), r = cor(age, income),   
 mean\_age = mean(age), mean\_income = mean(income),  
 sd\_age = sd(age), sd\_income = sd(income)) %>%   
 mutate(true\_slope = r \* sd\_age / sd\_income,  
 true\_intercept = mean\_income - true\_slope\*mean\_age) # This vector shows means, slope, sd and intercept  
 p <- ggplot(data = people1, aes(x = age, y = income)) +  
 geom\_point()+  
 geom\_point(data = people1\_summary,  
 aes(x = mean\_age, y = mean\_income), # This scatter plot intercept mean age and mean income  
 color = "red", size = 3)   
 my\_dat <- people1\_summary %>%   
 mutate(slope\_people = slope\_people,  
 my\_intercept = mean\_income - slope\_people \* mean\_age)  
 p + geom\_abline(data = my\_dat,  
 aes(intercept = my\_intercept, slope = slope\_people), color = "dodgerblue")  
 }  
  
add\_line(slope\_people = 1.0)



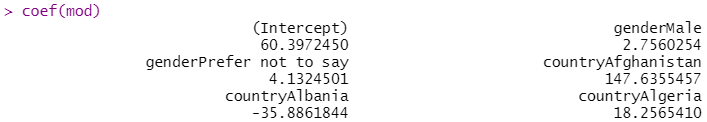
About the model, here it is the summary and the residuals have a mean of (residuals(mod)) [1] -4.7963, the residuals are the difference between observed and predicted data:

mod <- lm(income ~ gender + country, data = people1)  
summary(mod)



And here are the coefficients, describing the relationship the predictor variable and the response, that shows the beta coefficients and their statistical significance, because are a lot of countries, this imagen shows just the intercept, these intercepts are beta coefficients between country and income.

coef(mod)



And show the full output, this is the mean or average of all the residuals:

mean(residuals(mod))

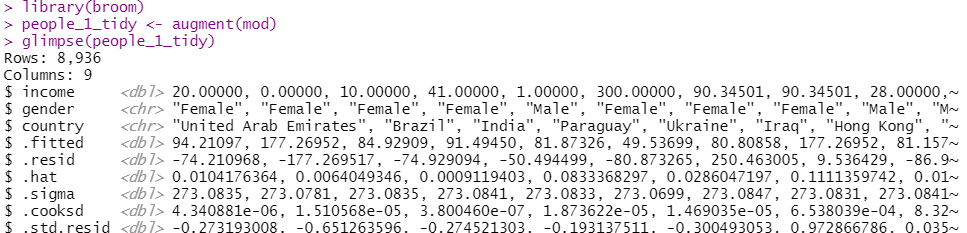


Now in R we have the augment function, it takes the object and gives the data that adjust to the model, with other information (residuals and others)

library(broom)

This people\_1\_tidy vector with augment() shows data that fixed to the model: .fitted, .residuals, .hat, .sigma , . cooksd, std.residual

people\_1\_tidy <- augment(mod)  
glimpse(people\_1\_tidy)  
mean(residuals(mod))

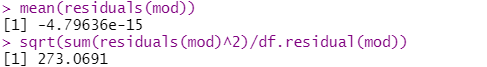


Let`s take RMSE, and after this, with .residuals we will take R2

Now RMSE:

mean(residuals(mod))  
sqrt(sum(residuals(mod)^2)/df.residual(mod))

Root Mean Square Error of 273.0691, this RMSE is very high.



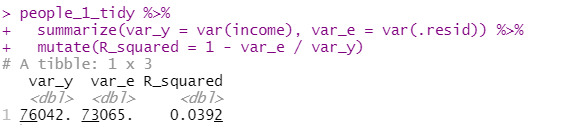
And now R2 with .residuals, that were obtained from glipse(people\_1\_tidy)

Simple linear regression consists of generating a regression model (equation of a line) that allows us to explain the linear relationship that exists between two variables. The dependent or response variable is identified as Y and the predictor or independent variable as X

And with .residuals information, it is possible to create a column called R\_squared and we get R2 = 0.0392

summary(mod)  
people\_1\_tidy %>%   
 summarize(var\_y = var(income), var\_e = var(.resid)) %>%   
 mutate(R\_squared = 1 - var\_e / var\_y)

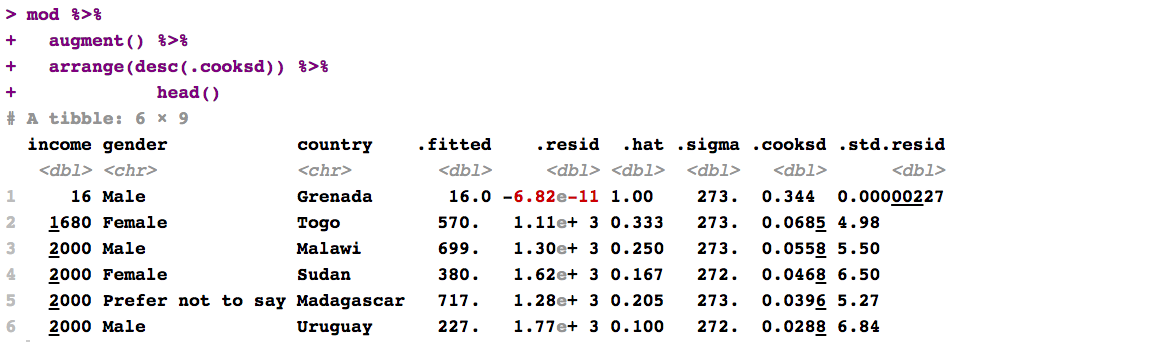
The .resid value is in the second last picture, this is in the glimpse(people\_1\_tidy), there are residuals or .resid



This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. This result of 0.0392 of R Sqared means that has a low relationship between income and country.

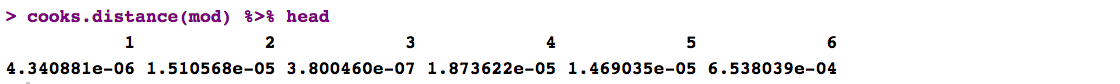
Other value is .cooksd that were consider in the line glimpse(people\_1\_tidy), the measurement of influence, this is used to identify some outliers:

mod %>%   
 augment() %>%   
 arrange(desc(.cooksd)) %>%   
 head()

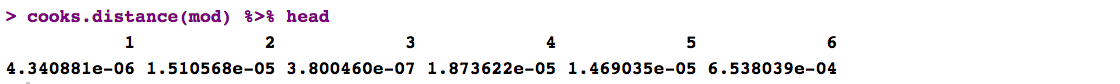


Or with this code

cooks.distance(mod) %>% head



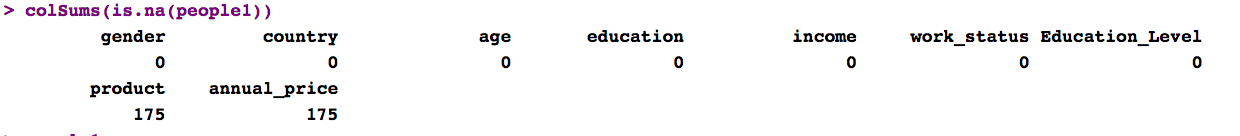
cooks.distance(mod) %>% tail

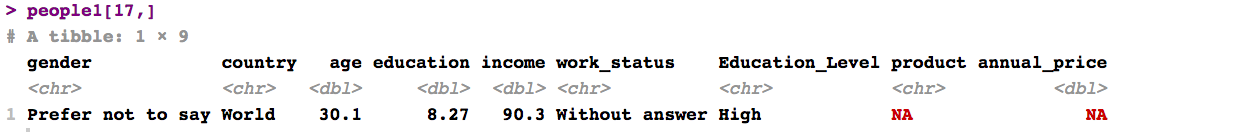


Anyway, it is not necessary remove outliers, because, the NAs were removed. At the beginning, this project removed the outliers and reduced the NAS values. But now more NAS were appear:

These NAS are in product and annual\_price, lets remove them and again check NAS presence:

which(is.na(people1$product))   
people1[17,]



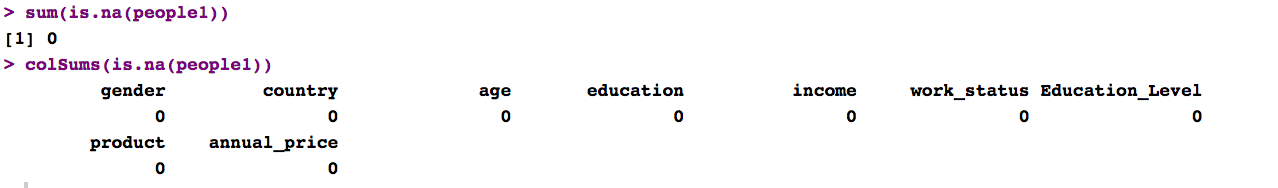


The sales strategy is offer the two Magazines to all these clientes that did not answer the gender in the form, so the client can accept one or both magazines, and to offer any magazine just by 200 dollars per year

people1$product[people1$product == "NA"] <- "Offertwomagazines"  
people1$product[is.na(people1$product)] <- "Offertwomagazines"  
people1$annual\_price[people1$annual\_price == "NA"] <- 200  
people1$annual\_price[is.na(people1$annual\_price)] <- 200

And now we don`t have NAS:

sum(is.na(people1))   
colSums(is.na(people1))



### Main Businesses Periscope

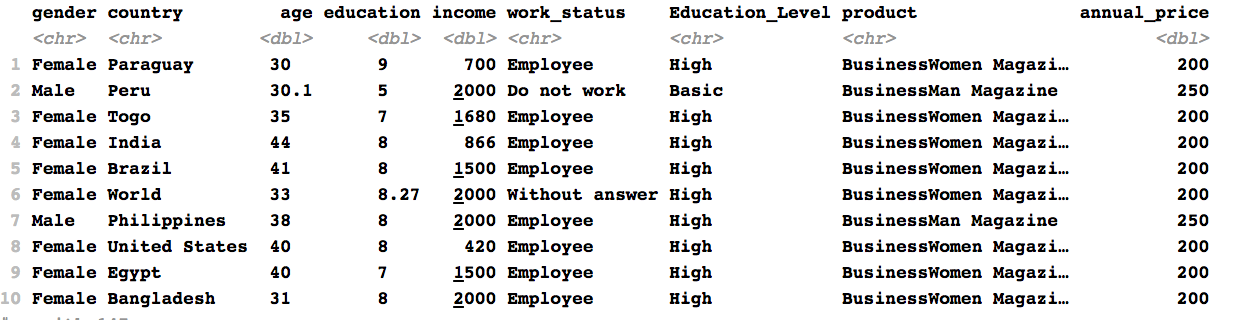
And now that definitely don´t have NAS, and we know that we´ll have more alternative to make selling’s to over 30 years old people, let`s see the countries were people have more income, we have 366 people that is more or equal to 30 years old and has an income of more or equal tan 120 dollars per week, these people will be the focus at the beginning, because as they have more income, they could buy the magazine more quickly:

This code will show the 179 countries

unique(people1$country)

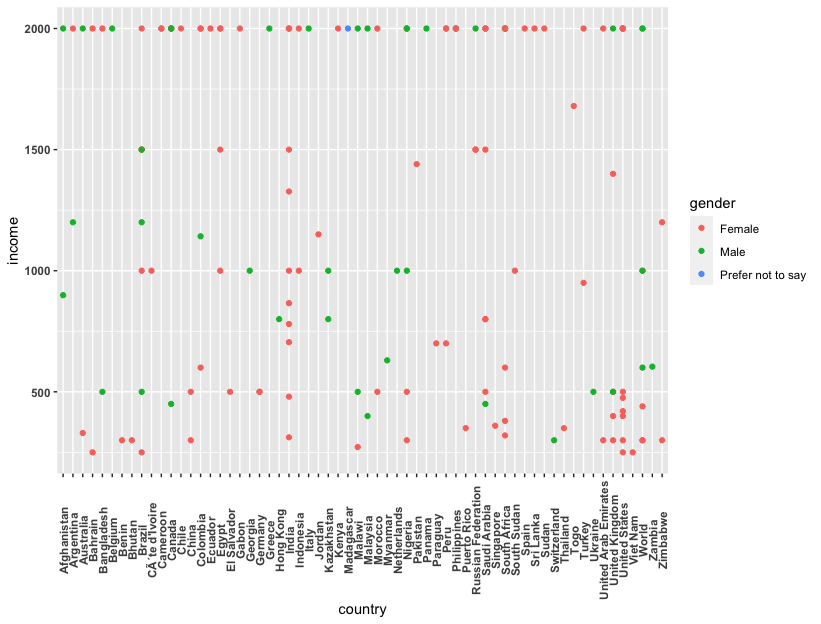
But the main business periscope is in people older than 30 and more than 250 dollars of income per week:

topsales <- people1 %>% filter(age >= 30 & income >= 250)



Here it is the main business periscope, 61 countries:

topsales %>% ggplot(aes(x = country, y = income, color = gender)) +   
 geom\_point(aes(x = country, y = income))+  
 theme(axis.text.x = element\_text(angle = 90, vjust = 0.6, face = "bold"))+  
 theme(axis.text.y = element\_text(face = "bold"))



At the beginning, just 61 countries will be the main focus of the magazine business

unique(topsales$country)



# 7. CONCLUSION AND SALES RECOMMENDATIONS

The linear model based on country and income, gives a disperse RMSE and very disperse relationship between income and country, but a positive relationship between Age and Income, where the blue line and the median of these two variables, represented by the red point, gives information to the Sales Manager and the CEO, that the magazine must be sell to more age or older clients, because older more income they have.

In the survey, 175 people answered “Prefer not to say” in the gender or sex option, female or male. The recommendation for these clients is offer at the minimum price of 200 dollars per year, the subscription on one of any magazines, so it is important for them offer the two magazines.

OK, this project had a good treatment a lot of NA information and take this survey, make a linear model in age, income and country variables and shows important information to take good decision to offer the magazines to over 30 years old, because it`s the mean of age and 90 dollars weekly income because it is the median of the income, and there is the red point in this linear model:

