# 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

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)
```

```
tibble [9,002 × 20] (S3: tbl_df/tbl/data.frame)
 $ gender             : chr [1:9002] "f" "f" "f" "f" ...
 $ country            : chr [1:9002] "United Arab Emirates" "Brazil" "India" "Paraguay" ...
 $ age                : chr [1:9002] "30" "21" "30" "35" ...
 $ height             : chr [1:9002] "17526" "179" "156" "180" ...
 $ weight             : chr [1:9002] "80" "82" "64" "74" ...
 $ education          : chr [1:9002] "7" "5" "7" "7" ...
 $ income             : chr [1:9002] "20" "0" "10" "41" ...
 $ weekly_hours       : chr [1:9002] "48" "30" "40" "42" ...
 $ field_of_study     : chr [1:9002] "Engineering" "Engineering" "Engineering" "Arts and humanities" ...
 $ stats_level        : chr [1:9002] "1" "4" "3" "1" ...
 $ movie_genre        : chr [1:9002] "Documentary" "Comedy" "Drama" "Documentary" ...
 $ holiday_destination: chr [1:9002] "Japan" "Hawaii" "Hawaii" "Hawaii" ...
 $ attractiveness     : chr [1:9002] "8" "8" "5" "5" ...
 $ optimism           : chr [1:9002] "4" "4" "3" "4" ...
 $ stress             : chr [1:9002] "4" "2" "4" "3" ...
 $ relationship       : chr [1:9002] "1" "0" "1" "1" ...
 $ children           : chr [1:9002] "0" "0" "1" "1" ...
 $ work_status        : chr [1:9002] "Self-employed" "Do not work" "Employee" "Employee" ...
 $ work_sector        : chr [1:9002] "Other" "null" "Computer and software" "Telecommunications" ...
 $ ...20              : chr [1:9002] NA NA NA NA ...
```

## People1

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

```
people1 <- select(people, "gender", "country", "age", "education", "income",
"work_status")
```

| gender | country | age | education | income | work_status |
|---|---|---|---|---|---|
| <chr> | <chr> | <chr> | <chr> | <chr> | <chr> |
| f | United Arab Emirates | 30 | 7 | 20 | Self-employed |
| f | Brazil | 21 | 5 | 0 | Do not work |
| f | India | 30 | 7 | 10 | Employee |
| f | Paraguay | 35 | 7 | 41 | Employee |
| m | Ukraine | 33 | 8 | 1 | Employee |
| f | Iraq | 26 | 8 | 300 | null |
| f | Hong Kong | 22 | 7 | null | Do not work |
| f | Brazil | 36 | 8 | null | Employee |
| m | Australia | 22 | 5 | 28 | Employee |
| m | India | 19 | 5 | null | Do not work |

. with 8,992 more rows

## 2. Variables data set classes and modeling some information

### Working with NA´s

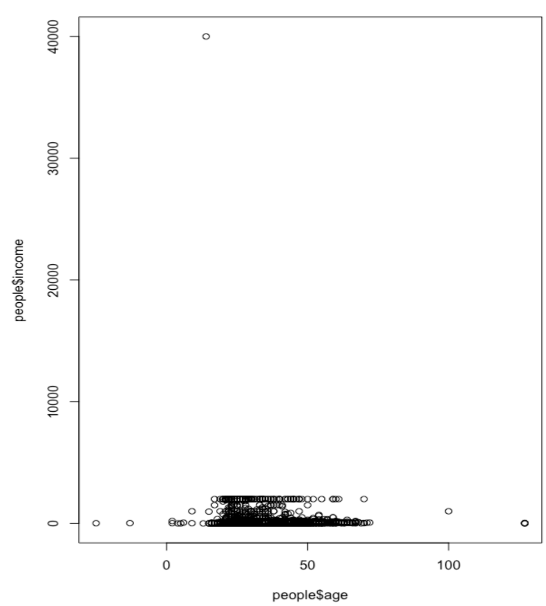
```
people1$age <- as.numeric(as.character(people1$age))
people1$education <- as.numeric(as.character(people1$edu
cation))
people1$income <- as.numeric(as.character(people1$income
))
sapply(people1, class)
mean(people1$age, na.rm = TRUE)


sapply(people1, class)
plot(people1$age, people1$income)
```

```
> sapply(people1, class)
         gender
    "character"
        country
    "character"
            age
      "numeric"
      education
      "numeric"
         income
      "numeric"
    work_status
    "character"
Education_Level
    "character"
        product
    "character"
   annual_price
      "numeric"
```

### Working with outliers

```
plot(people1$age, people1$income)
```



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

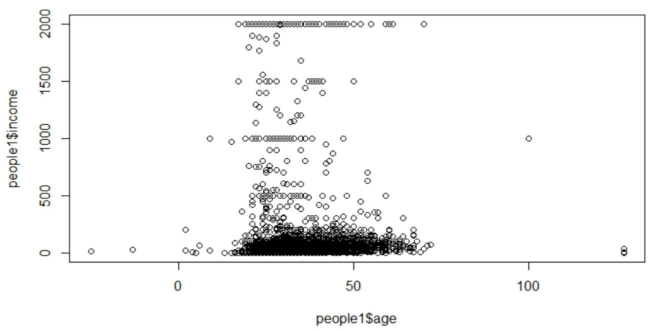
```
 gender country   age education income work_status
 <chr>  <chr>    <dbl>     <dbl>  <dbl> <chr>
 m      UK          14         2  40000 working
```

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

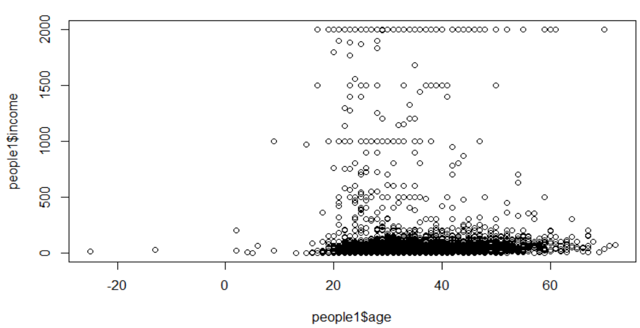
## Eliminate the registers that are grater than or equal to 100 ages and plot it

```
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:

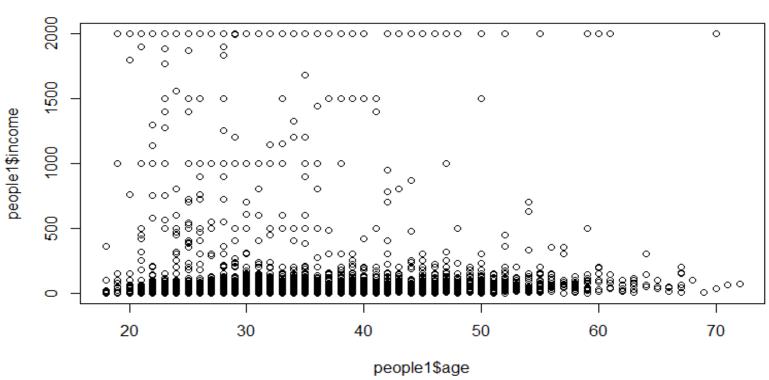
```
which(people1$age < 0)
people1[c(2906, 5013, 5814, 5990, 6572, 8749),]
```

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, 435
0, 4450, 4738, 4888, 4922, 4987, 5013, 5137, 5298, 5388, 5695, 5733, 5814, 58
90, 5904, 5990, 6026, 6080, 6375, 6455, 6572, 6785, 6930, 7072, 7229, 7509, 7
712, 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, 383
7, 4031, 4350, 4450, 4738, 4888, 4922, 4987, 5013, 5137, 5298, 5388, 5695, 57
33, 5814, 5890, 5904, 5990, 6026, 6080, 6375, 6455, 6572, 6785, 6930, 7072, 7
229, 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"))
```

```
people1
A tibble: 8,936 × 9
 gender country                age education income work_status    Education_Level
 <chr>  <chr>                <dbl>     <dbl>  <dbl> <chr>          <chr>
 Female United Arab Emirates    30         7   20   Self-employed  High
 Female Brazil                  21         5    0   Do not work    Basic
 Female India                   30         7   10   Employee       High
 Female Paraguay                35         7   41   Employee       High
 Male   Ukraine                 33         8    1   Employee       High
 Female Iraq                    26         8  300   Without answer High
 Female Hong Kong               22         7   90.3 Do not work    High
 Female Brazil                  36         8   90.3 Employee       High
 Male   Australia               22         5   28   Employee       Basic
 Male   India                   19         5   90.3 Do not work    Basic
```

# 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))
```

```
colSums(is.na(people1)) #Shows all columns and the quantity of NA`s that have these columns
   gender      country          age    education       income work_status
        0          464          491          425         1568           0
```

## 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))
```

```
[401] 7388 7398 7400 7441 7450 7469 7499 7500 7521 7547 7577 7581 7663 7681 7682 7689 7731 7747 7758 7774
[421] 7825 7842 7886 7909 7939 7941 7946 7949 7957 7961 7976 7981 8011 8014 8031 8053 8067 8097 8101 8115
[441] 8122 8131 8170 8181 8182 8257 8259 8262 8270 8277 8280 8283 8301 8307 8308 8309 8429 8433 8434 8471
[461] 8487 8496 8513 8519 8520 8526 8555 8569 8582 8583 8594 8619 8636 8644 8678 8712 8720 8757 8785 8789
[481] 8802 8806 8815 8863 8883 8886 8888 8890 8893 8895 8909
```

```
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))
```

```
> which(is.na(people1$age))
integer(0)
```

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

```
which(is.na(people1$education))
```

```
[341] 7210 7231 7239 7325 7380 7396 7398 7420 7441 7469 7481 7500 7515 7517 7521 7555 7581 7586 7620 7668
[361] 7682 7689 7758 7773 7774 7807 7820 7823 7825 7909 7916 7935 7939 7941 7981 8048 8079 8097 8115 8128
[381] 8131 8204 8257 8289 8307 8308 8309 8429 8433 8456 8476 8487 8489 8496 8498 8513 8519 8531 8541 8569
[401] 8582 8583 8609 8621 8644 8648 8678 8718 8720 8728 8746 8757 8802 8803 8806 8811 8815 8826 8863 8869
[421] 8886 8888 8890 8895 8909
```

```
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))
```

```
> which(is.na(people1$education))
integer(0)
```

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

```
which(is.na(people1$income))
```

```
[921] 5314 5339 5340 5353 5356 5357 5374 5377 5378 5380 5402 5409 5411 5415 5425 5426 5433 5445 5450 5454
[941] 5461 5470 5480 5489 5491 5492 5494 5515 5516 5540 5542 5544 5555 5558 5561 5566 5567 5571 5581 5583
[961] 5586 5591 5600 5605 5606 5621 5626 5627 5635 5645 5649 5658 5669 5671 5673 5684 5689 5694 5696 5703
[981] 5704 5706 5718 5720 5726 5729 5732 5737 5742 5751 5752 5768 5776 5778 5781 5783 5794 5800 5802 5804
[ reached getOption("max.print") -- omitted 568 entries ]
```

```
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))
```

```
> which(is.na(people1$income)) # 1568 NA´s values
integer(0)
```

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
"
```

```
> people1
# A tibbl
   gender
   <chr>
 1 Female
 2 Female
 3 Female
 4 Female
 5 Male
 6 Female
 7 Female
 8 Female
 9 Male
10 Male
# ... witl
```

464 NA`s were found in country variable

```
which(is.na(people1$country))
```

```
[381] 7398 7404 7428 7441 7446 7453 7454 7469 7474 7486 7494 7500 7505 7521 7575 7581 7609 7612 7658 7681
[401] 7682 7689 7773 7774 7820 7824 7825 7844 7854 7869 7885 7909 7932 7939 7941 7950 7981 8028 8078 8085
[421] 8097 8101 8131 8181 8189 8220 8257 8307 8308 8429 8433 8474 8487 8495 8496 8513 8519 8531 8569 8582
[441] 8583 8586 8597 8644 8678 8679 8715 8716 8720 8733 8736 8757 8802 8806 8815 8859 8863 8882 8886 8888
[461] 8890 8895 8909 8910
```

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))
```

```
> which(is.na(people1$country))
integer(0)
```

## 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”:

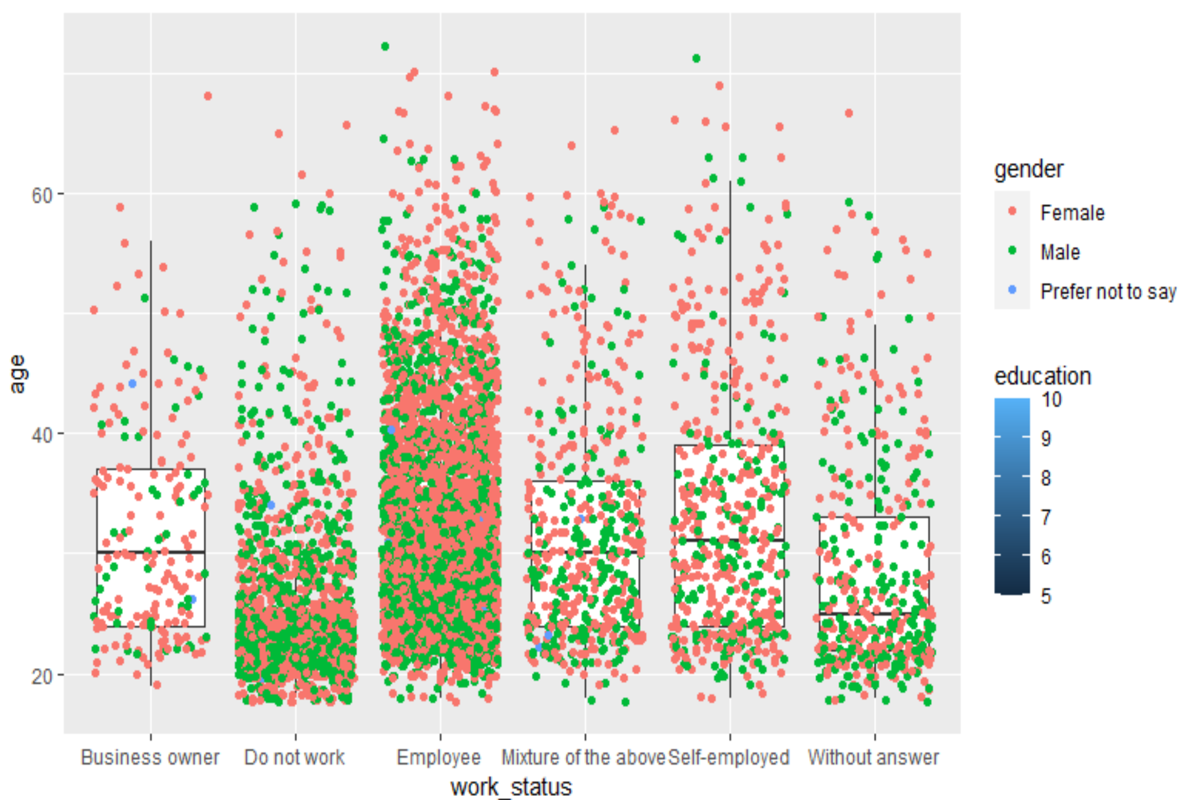
```
people1$work_status[people1$work_status == "null"] <- "Without answer"
people1$work_status[people1$work_status == 0] <- "Others"
people1$work_status[people1$work_status == 1] <- "Others"
people1$work_status[people1$work_status == 2] <- "Others"
people1$work_status[people1$work_status == 3] <- "Others"
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?

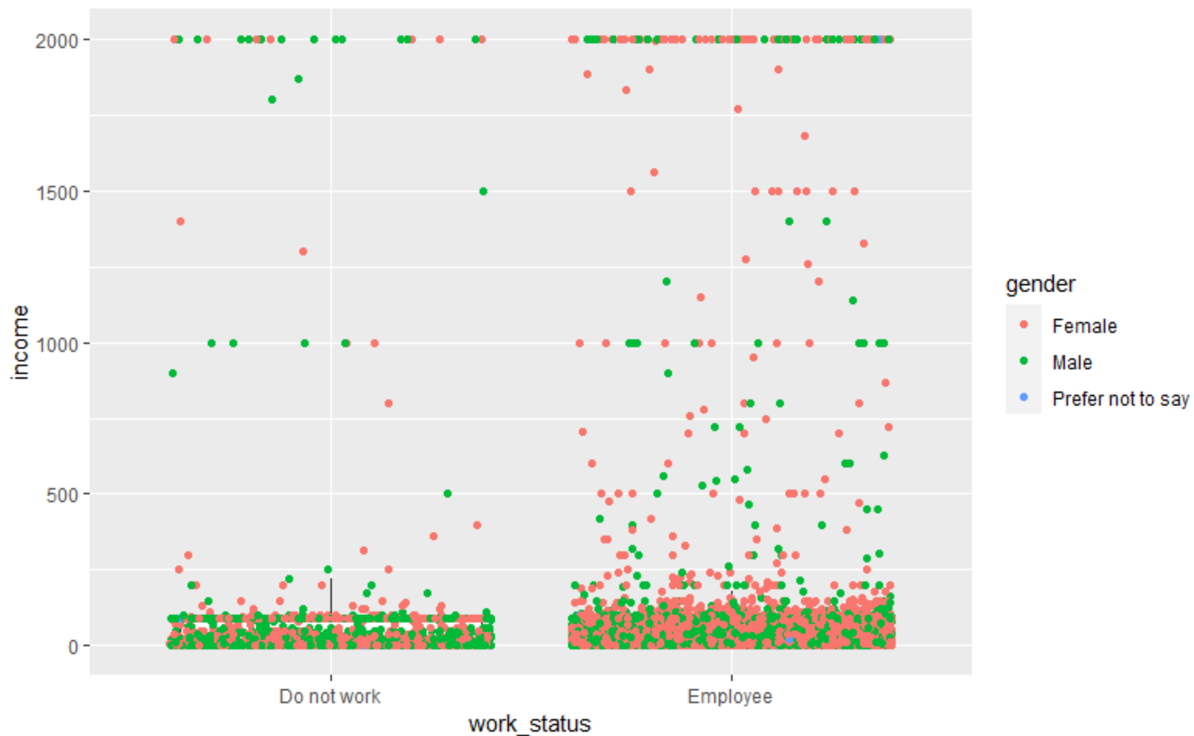
“Employees” and “Do not work” people:

```
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 w
ork"))%>%
    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)
```

```
                        Female Male Prefer not to say
  Business owner          149   46                 4
  Do not work            1003 1100                12
  Employee               2885 1894                13
  Mixture of the above    302  200                 5
  Others                   92   62                 1
  Self-employed           335  164                 3
  Without answer          270  259               137
```

## 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 ow
ner", "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"))
```

| | gender | country | age | education | income | work_status | product |
|---|---|---|---|---|---|---|---|
| | <chr> | <chr> | <dbl> | <dbl> | <dbl> | <chr> | <chr> |
| 1 | Female | United Arab Emirates | 30 | 7 | 20 | Self-employed | BusinessWomen Magazine |
| 2 | Female | Brazil | 21 | 5 | 0 | Do not work | BusinessWomen Magazine |
| 3 | Female | India | 30 | 7 | 10 | Employee | BusinessWomen Magazine |
| 4 | Female | Paraguay | 35 | 7 | 41 | Employee | BusinessWomen Magazine |
| 5 | Male | Ukraine | 33 | 8 | 1 | Employee | BusinessMan Magazine |
| 6 | Female | Iraq | 26 | 8 | 300 | Without answer | BusinessWomen Magazine |
| 7 | Female | Hong Kong | 22 | 7 | 90.3 | Do not work | BusinessWomen Magazine |
| 8 | Female | Brazil | 36 | 8 | 90.3 | Employee | BusinessWomen Magazine |
| 9 | Male | Australia | 22 | 5 | 28 | Employee | BusinessMan Magazine |
| 10 | Male | India | 19 | 5 | 90.3 | Do not work | BusinessMan Magazine |

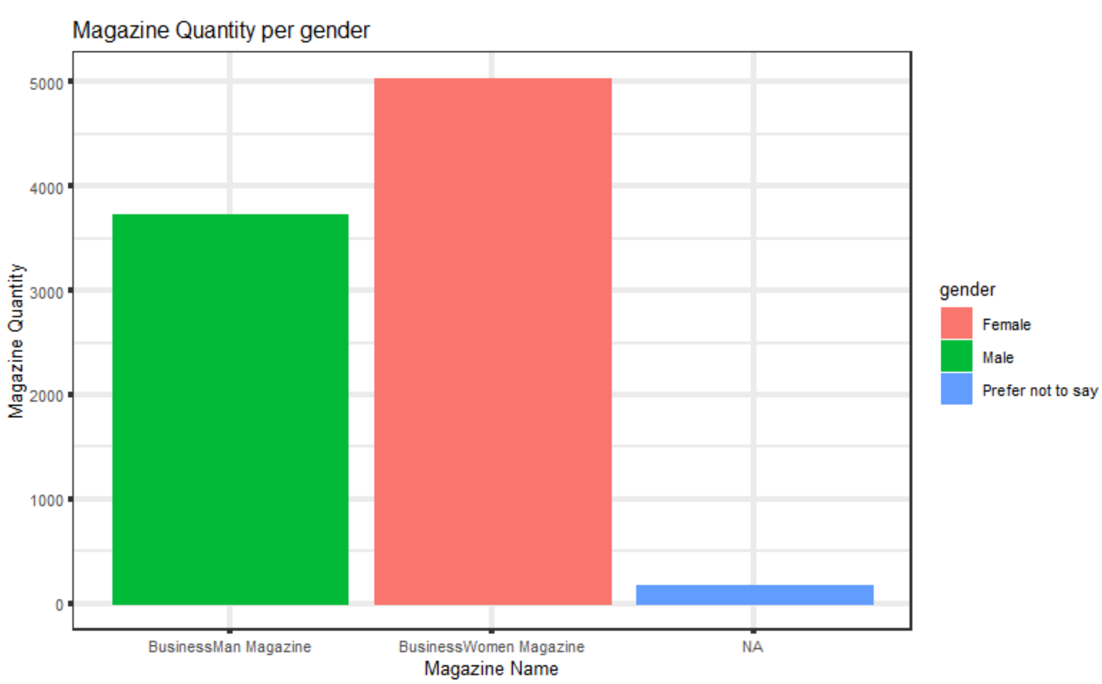
And the price of the Magazines will be USD 250 for Female and 200 dollars for Male, annually digital subscription:

```
people1 <- people1 %>%
  mutate(annual_price = case_when(
  gender == "Female" ~ 200,
  gender == "Male" ~ 250)
)
```

| | gender | country | age | education | income | work_status | product | annual_price |
|---|---|---|---|---|---|---|---|---|
| | <chr> | <chr> | <dbl> | <dbl> | <dbl> | <chr> | <chr> | <dbl> |
| 1 | Female | United Arab Emirates | 30 | 7 | 20 | Self-employed | BusinessWomen Magazine | 200 |
| 2 | Female | Brazil | 21 | 5 | 0 | Do not work | BusinessWomen Magazine | 200 |
| 3 | Female | India | 30 | 7 | 10 | Employee | BusinessWomen Magazine | 200 |
| 4 | Female | Paraguay | 35 | 7 | 41 | Employee | BusinessWomen Magazine | 200 |
| 5 | Male | Ukraine | 33 | 8 | 1 | Employee | BusinessMan Magazine | 250 |
| 6 | Female | Iraq | 26 | 8 | 300 | Without answer | BusinessWomen Magazine | 200 |
| 7 | Female | Hong Kong | 22 | 7 | 90.3 | Do not work | BusinessWomen Magazine | 200 |
| 8 | Female | Brazil | 36 | 8 | 90.3 | Employee | BusinessWomen Magazine | 200 |
| 9 | Male | Australia | 22 | 5 | 28 | Employee | BusinessMan Magazine | 250 |
| 10 | Male | India | 19 | 5 | 90.3 | Do not work | BusinessMan Magazine | 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 Quanti
ty 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)
```

```
> country.sales.lm

Call:
lm(formula = annual_price ~ gender + income, data = people1)

Coefficients:
(Intercept)   genderMale       income
  2.000e+02    5.000e+01    1.441e-17
```

linear model for income as a function of gender and annual_price:

```
country.sales.lm2 <- lm(income ~ gender + annual_price, data = people1)
```

```
> country.sales.lm2

Call:
lm(formula = income ~ gender + annual_price, data = people1)

Coefficients:
 (Intercept)    genderMale  annual_price
      91.444        -2.784            NA
```

## 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)
```

```
> country.sales.lm3

Call:
lm(formula = income ~ gender + country, data = people1)

Coefficients:
                  (Intercept)                     genderMale
                      60.3972                         2.7560
         genderPrefer not to say            countryAfghanistan
                       4.1325                       147.6355
               countryAlbania                 countryAlgeria
                     -35.8862                        18.2565
       countryAmerican Samoa                  countryAndorra
                     173.6028                        27.1917
                countryAngola      countryAntigua and Barbuda
                     -15.3307                       -63.1533
             countryArgentina                 countryArmenia
                     124.2007                       -25.2578
                 countryAruba               countryAustralia
                      39.6028                        18.0044
               countryAustria              countryAzerbaijan
                      -4.3743                       -24.3639
               countryBahamas                 countryBahrain
                     -25.1533                       171.8359
            countryBangladesh                countryBarbados
                      50.4159                       -20.9228
               countryBelarus                 countryBelgium
                     -28.3780                        83.1976
                countryBelize                   countryBenin
                      29.9478                        90.6028
```

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

```
summary(country.sales.lm3)
```

```
> summary(country.sales.lm3)

Call:
lm(formula = income ~ gender + country, data = people1)

Residuals:
    Min      1Q  Median      3Q     Max
-713.17  -76.69  -37.44    2.66 1944.72

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                             60.3972   136.5431   0.442 0.658261
genderMale                               2.7560     6.1272   0.450 0.652865
genderPrefer not to say                  4.1325    24.7069   0.167 0.867170
countryAfghanistan                     147.6355   147.0551   1.004 0.315432
countryAlbania                         -35.8862   153.6646  -0.234 0.815351
countryAlgeria                          18.2565   150.2217   0.122 0.903274
countryAmerican Samoa                  173.6028   305.3044   0.569 0.569627
countryAndorra                          27.1917   305.3351   0.089 0.929040
```

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

```
summary(country.sales.lm3)
```

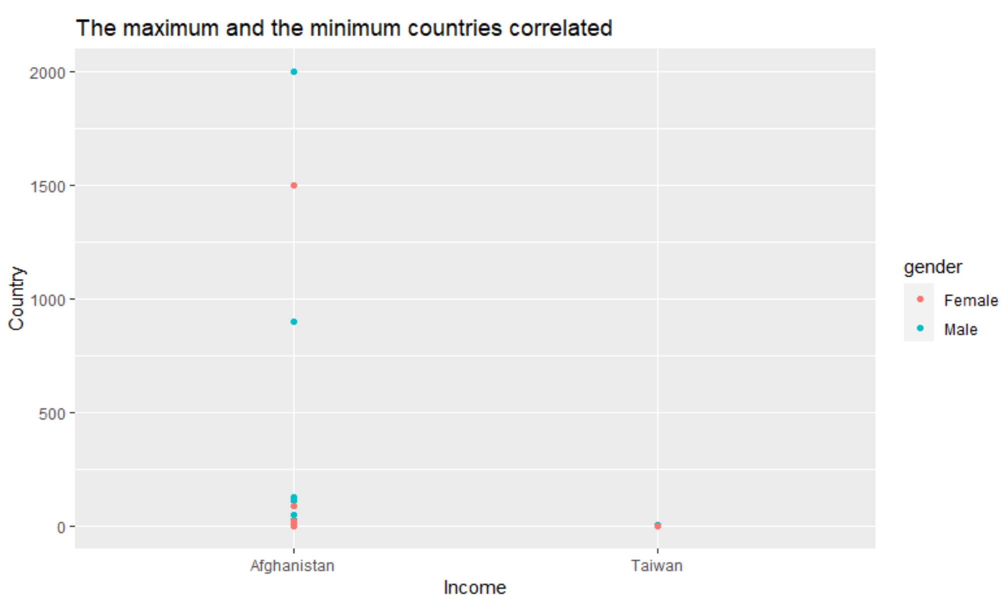
Interpretation and difference between Afghanistan and Taiwan, this example of these two countries because both of them has the biggest difference in the income, that has this data set people1:

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.

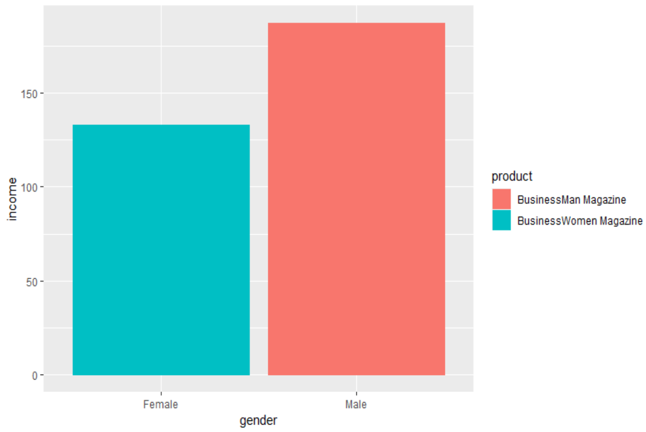
Afghanistan and Taiwan, here in this plot, we can see the enormous difference in income:

```
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 coun
tries correlated")
```



Taiwan Income, gender and product:

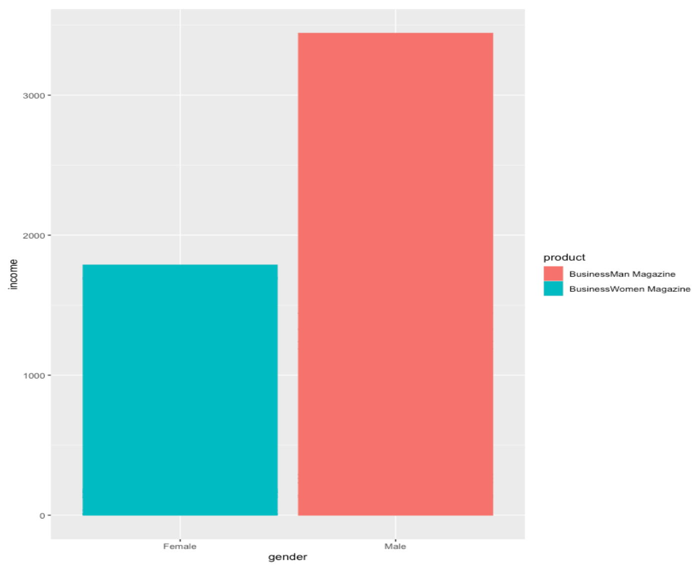
```
people1 %>% filter(country == "Taiwan") %>%
  ggplot()+
  geom_bar(mapping = aes(x = gender, y = income, color = product, fill = prod
uct), stat = "identity")
```




Income until 150

Afghanistan Income, gender and product:

```
people1 %>% filter(country == "Afghanistan") %>%
  ggplot()+
  geom_bar(mapping = aes(x = gender, y = income, color = product, fill = prod
uct), stat = "identity")
```



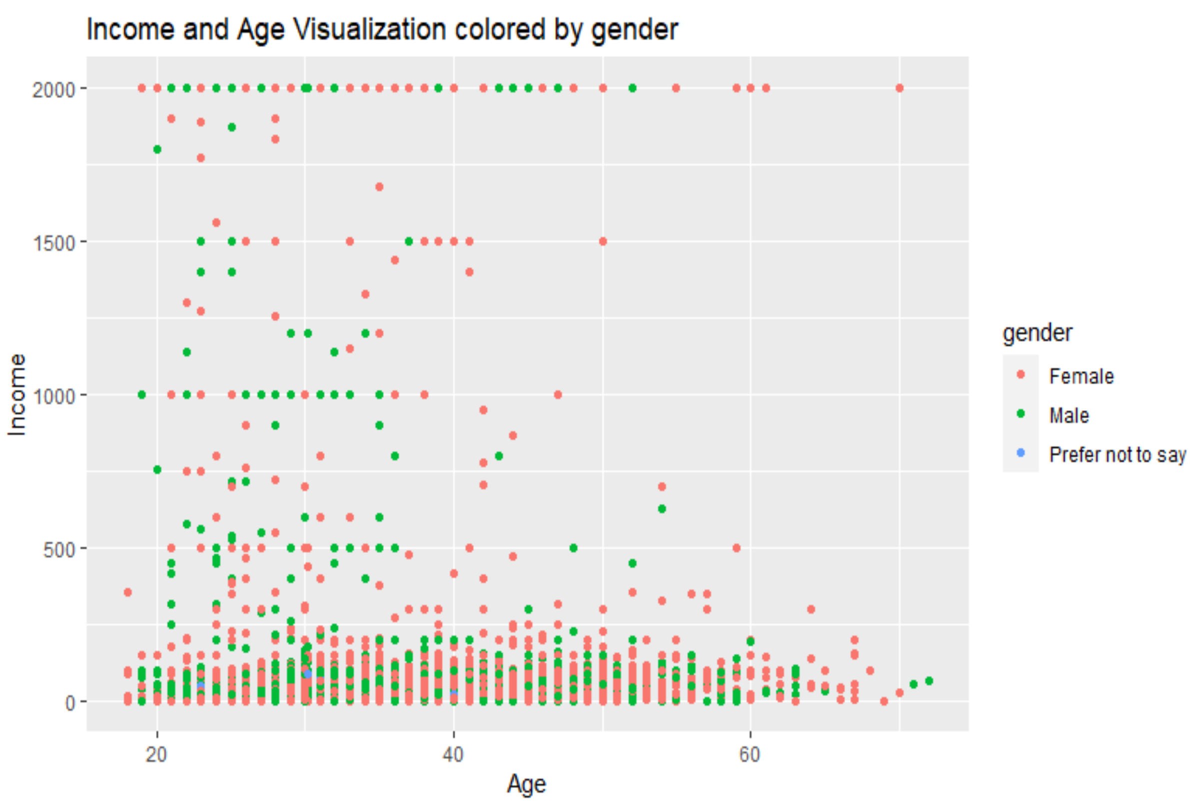

Income over 3000

## Data set survey concentration

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

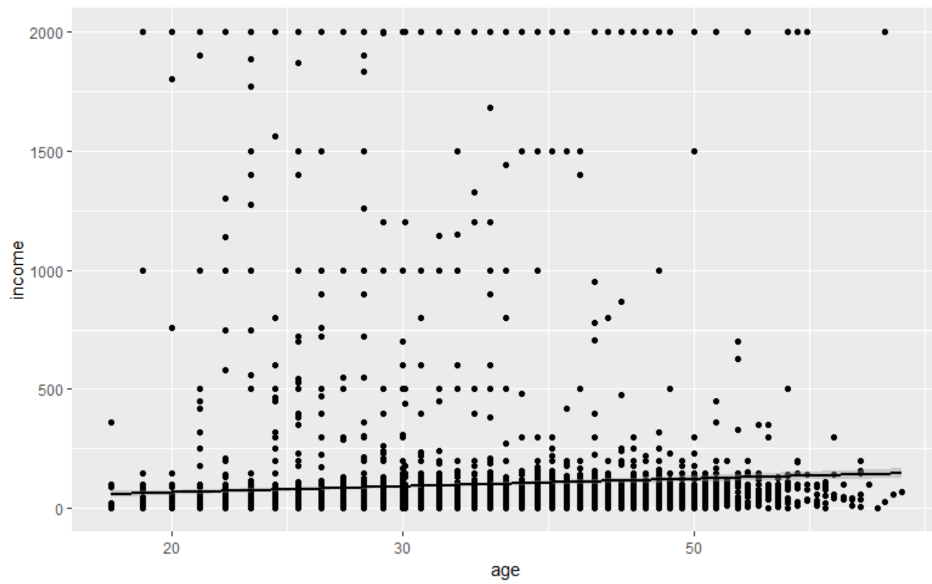
Let`s plot this:

```
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")
```



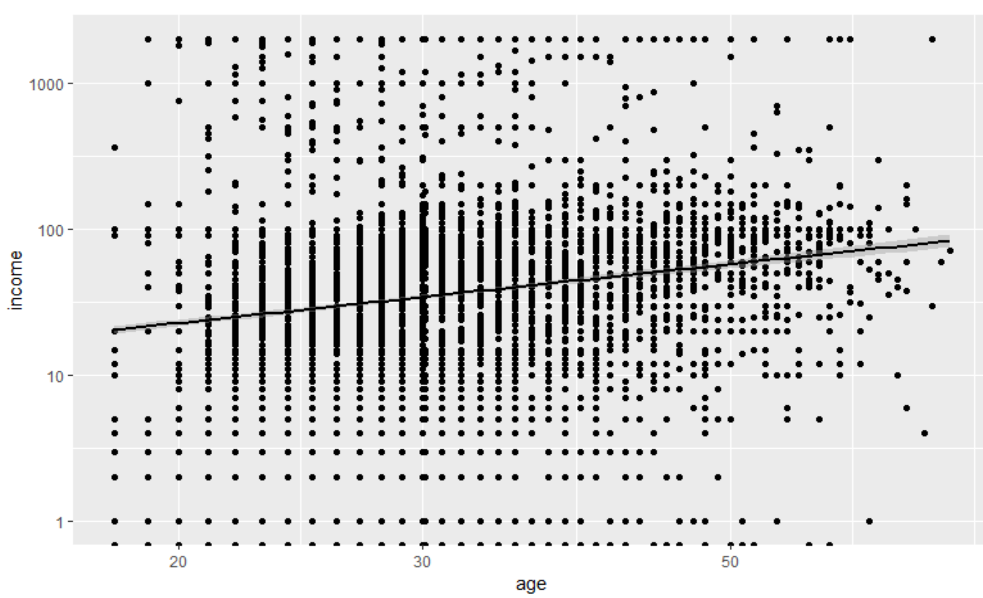
More age, more income, and the line:

```
lm3plot2 <- ggplot(people1, aes(age, income))+geom_point()
graphlm3 <- lm3plot2 + geom_smooth(method = "lm", col = "black") + scale_x_lo
g10()
graphlm3
```



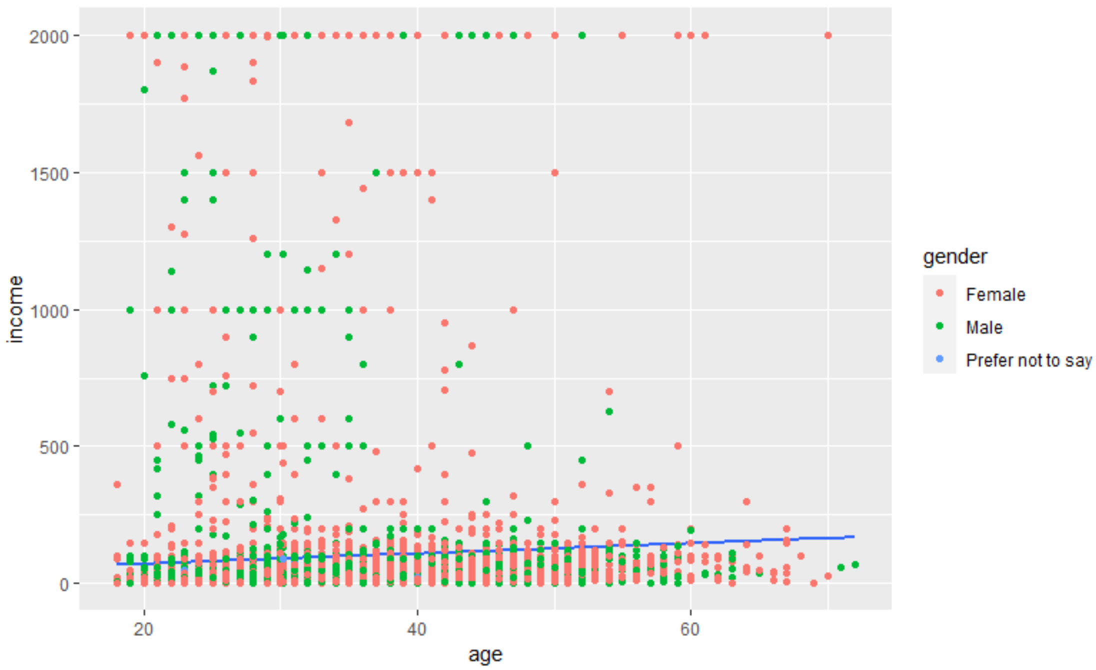

Obviously, y axis in this line is disproportionate, but anyway we can see the line better, this because has a transformation scale in the code:

```
graphlm3 <- lm3plot2 + geom_smooth(method = "lm", col = "black") + scale_x_lo
g10() + scale_y_log10()
```



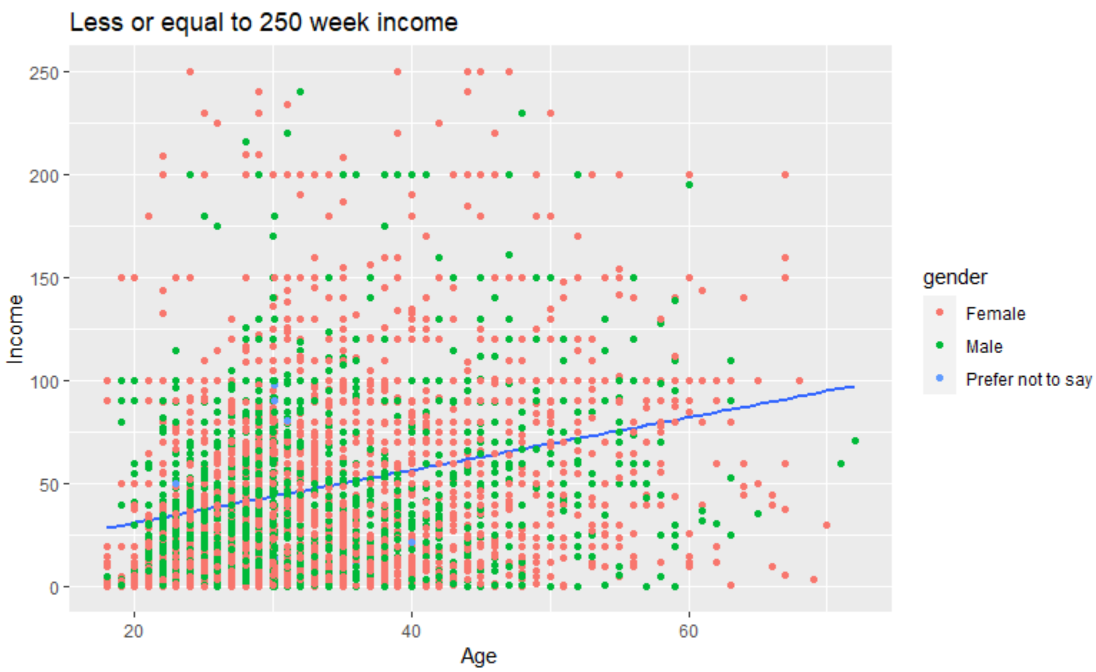
or colored but not with the transformation scale:

```
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:

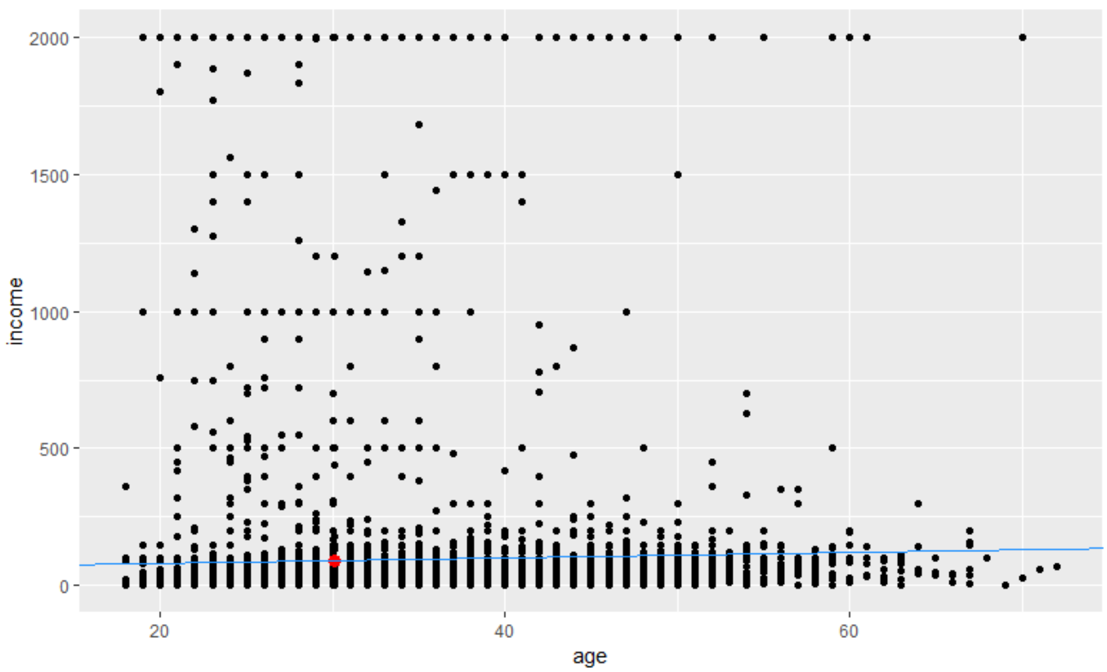
```
people1 %>%
  filter(income <= 250) %>%  # Less or equal to 250 of income
  ggplot(aes(x = age, income))+
  geom_smooth(method = lm, se = FALSE)+
  geom_point(aes(color = gender))+
  labs(x = "Age", y = "Income", title = "Less or equal to 250 week income")
```




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 interc
ept 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)
```

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

Call:
lm(formula = income ~ gender + country, data = people1)

Residuals:
    Min      1Q  Median      3Q     Max
-713.17  -76.69  -37.44    2.66 1944.72

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              60.3972   136.5431   0.442 0.658261
genderMale                                2.7560     6.1272   0.450 0.652865
genderPrefer not to say                   4.1325    24.7069   0.167 0.867170
countryAfghanistan                      147.6355   147.0551   1.004 0.315432
countryAlbania                          -35.8862   153.6646  -0.234 0.815351
countryAlgeria                           18.2565   150.2217   0.122 0.903274
countryAmerican Samoa                   173.6028   305.3044   0.569 0.569627
```

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)
```

```
> coef(mod)
                (Intercept)                  genderMale
                 60.3972450                   2.7560254
     genderPrefer not to say          countryAfghanistan
                  4.1324501                 147.6355457
              countryAlbania              countryAlgeria
                -35.8861844                  18.2565410
```

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

```
mean(residuals(mod))
```

```
> mean(residuals(mod))
[1] -4.79636e-15
```

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))
```

```
> library(broom)
> people_1_tidy <- augment(mod)
> glimpse(people_1_tidy)
Rows: 8,936
Columns: 9
$ income     <dbl> 20.00000, 0.00000, 10.00000, 41.00000, 1.00000, 300.00000, 90.34501, 90.34501, 28.00000,~
$ gender     <chr> "Female", "Female", "Female", "Female", "Male", "Female", "Female", "Female", "Male", "M~
$ country    <chr> "United Arab Emirates", "Brazil", "India", "Paraguay", "Ukraine", "Iraq", "Hong Kong", "~
$ .fitted    <dbl> 94.21097, 177.26952, 84.92909, 91.49450, 81.87326, 49.53699, 80.80858, 177.26952, 81.157~
$ .resid     <dbl> -74.210968, -177.269517, -74.929094, -50.494499, -80.873265, 250.463005, 9.536429, -86.9~
$ .hat       <dbl> 0.0104176364, 0.0064049346, 0.0009119403, 0.0833368297, 0.0286047197, 0.1111359742, 0.01~
$ .sigma     <dbl> 273.0835, 273.0781, 273.0835, 273.0841, 273.0833, 273.0699, 273.0847, 273.0831, 273.0841~
$ .cooksd    <dbl> 4.340881e-06, 1.510568e-05, 3.800460e-07, 1.873622e-05, 1.469035e-05, 6.538039e-04, 8.32~
$ .std.resid <dbl> -0.273193008. -0.651263596. -0.274521303. -0.193137511. -0.300493053. 0.972866786. 0.035~
```

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.

```
> mean(residuals(mod))
[1] -4.79636e-15
> sqrt(sum(residuals(mod)^2)/df.residual(mod))
[1] 273.0691
```

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

```
> people_1_tidy %>%
+   summarize(var_y = var(income), var_e = var(.resid)) %>%
+   mutate(R_squared = 1 - var_e / var_y)
# A tibble: 1 x 3
   var_y  var_e R_squared
   <dbl>  <dbl>     <dbl>
1 76042. 73065.    0.0392
```

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()
```

```
> mod %>%
+   augment() %>%
+   arrange(desc(.cooksd)) %>%
+            head()
# A tibble: 6 × 9
  income gender            country    .fitted    .resid  .hat .sigma .cooksd .std.resid
   <dbl> <chr>             <chr>        <dbl>     <dbl> <dbl>  <dbl>   <dbl>      <dbl>
1     16 Male              Grenada       16.0 -6.82e-11 1.00    273.  0.344  0.00000227
2   1680 Female            Togo         570.   1.11e+ 3 0.333   273.  0.0685 4.98
3   2000 Male              Malawi       699.   1.30e+ 3 0.250   273.  0.0558 5.50
4   2000 Female            Sudan        380.   1.62e+ 3 0.167   272.  0.0468 6.50
5   2000 Prefer not to say Madagascar   717.   1.28e+ 3 0.205   273.  0.0396 5.27
6   2000 Male              Uruguay      227.   1.77e+ 3 0.100   272.  0.0288 6.84
```

Or with this code

```
cooks.distance(mod) %>% head
```

```
> cooks.distance(mod) %>% head
           1            2            3            4            5            6 
4.340881e-06 1.510568e-05 3.800460e-07 1.873622e-05 1.469035e-05 6.538039e-04 
```

```
cooks.distance(mod) %>% tail
```

```
> cooks.distance(mod) %>% head
           1            2            3            4            5            6 
4.340881e-06 1.510568e-05 3.800460e-07 1.873622e-05 1.469035e-05 6.538039e-04 
```

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,]
```

```
> colSums(is.na(people1))
         gender         country             age       education          income     work_status Education_Level 
              0               0               0               0               0               0               0 
        product    annual_price 
            175             175 
```

```
> people1[17,]
# A tibble: 1 × 9
  gender          country   age education income work_status    Education_Level product annual_price
  <chr>           <chr>   <dbl>     <dbl>  <dbl> <chr>          <chr>           <chr>          <dbl>
1 Prefer not to say World    30.1      8.27   90.3 Without answer High            NA                NA
```

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))
```

```
> sum(is.na(people1))
[1] 0
> colSums(is.na(people1))
         gender         country             age       education          income     work_status Education_Level 
              0               0               0               0               0               0               0 
        product    annual_price 
              0               0 
```

## 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)
```

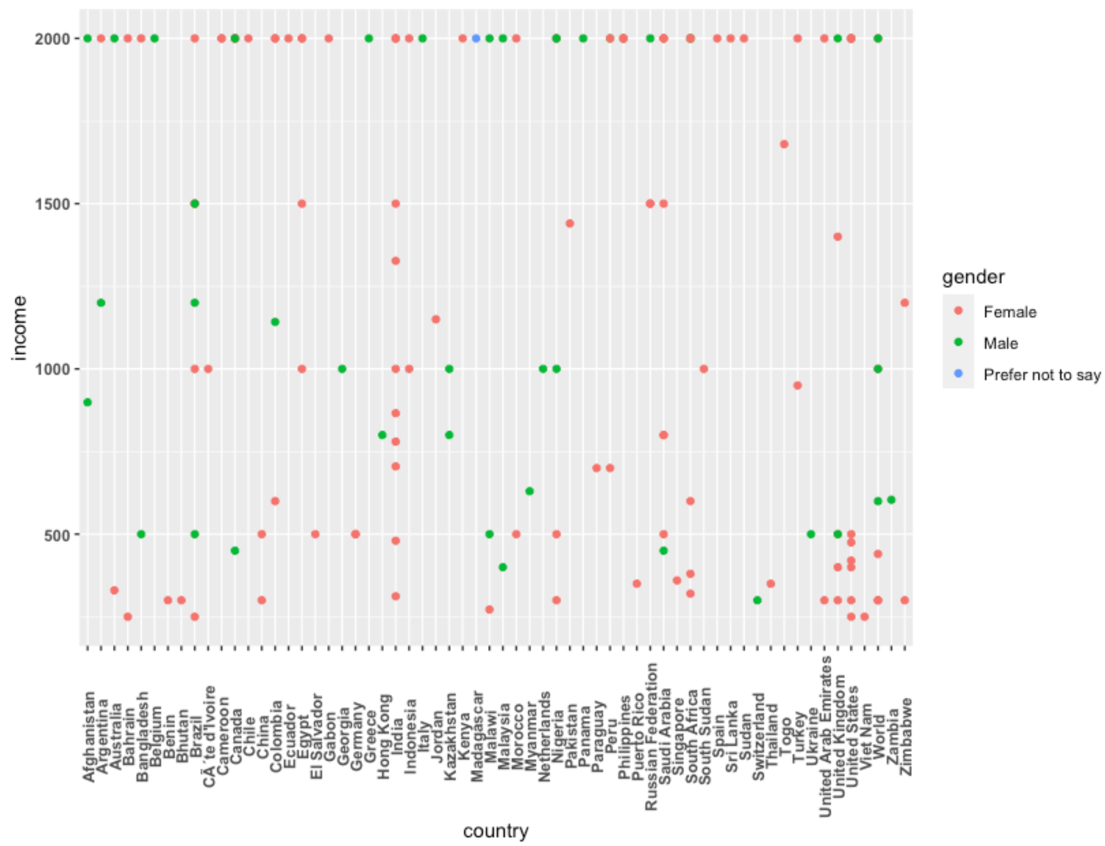
But the main business periscope is in people older than 30 and more than 250 dollars of income per week:

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

| | gender | country | age | education | income | work_status | Education_Level | product | annual_price |
|---|---|---|---|---|---|---|---|---|---|
| | <chr> | <chr> | <dbl> | <dbl> | <dbl> | <chr> | <chr> | <chr> | <dbl> |
| 1 | Female | Paraguay | 30 | 9 | 700 | Employee | High | BusinessWomen Magazi… | 200 |
| 2 | Male | Peru | 30.1 | 5 | 2000 | Do not work | Basic | BusinessMan Magazine | 250 |
| 3 | Female | Togo | 35 | 7 | 1680 | Employee | High | BusinessWomen Magazi… | 200 |
| 4 | Female | India | 44 | 8 | 866 | Employee | High | BusinessWomen Magazi… | 200 |
| 5 | Female | Brazil | 41 | 8 | 1500 | Employee | High | BusinessWomen Magazi… | 200 |
| 6 | Female | World | 33 | 8.27 | 2000 | Without answer | High | BusinessWomen Magazi… | 200 |
| 7 | Male | Philippines | 38 | 8 | 2000 | Employee | High | BusinessMan Magazine | 250 |
| 8 | Female | United States | 40 | 8 | 420 | Employee | High | BusinessWomen Magazi… | 200 |
| 9 | Female | Egypt | 40 | 7 | 1500 | Employee | High | BusinessWomen Magazi… | 200 |
| 10 | Female | Bangladesh | 31 | 8 | 2000 | Employee | High | BusinessWomen Magazi… | 200 |

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)
```

```
> unique(topsales$country)
 [1] "Paraguay"             "Peru"          "Togo"          "India"
 [5] "Brazil"               "World"         "Philippines"   "United States"
 [9] "Egypt"                "Bangladesh"    "Kenya"         "Russian Federation"
[13] "China"                "Netherlands"   "Belgium"       "United Kingdom"
[17] "South Africa"         "Australia"     "Saudi Arabia"  "Panama"
[21] "Indonesia"            "Canada"        "Nigeria"       "Sri Lanka"
[25] "Cameroon"             "El Salvador"   "Madagascar"    "Jordan"
[29] "Singapore"            "Benin"         "Morocco"       "CÃ´te d'Ivoire"
[33] "Gabon"                "Colombia"      "Afghanistan"   "Malaysia"
[37] "United Arab Emirates" "Zimbabwe"      "Turkey"        "Hong Kong"
[41] "Bahrain"              "Viet Nam"      "Malawi"        "Pakistan"
[45] "Italy"                "Myanmar"       "Chile"         "Puerto Rico"
[49] "Ukraine"              "Kazakhstan"    "Zambia"        "Argentina"
[53] "Germany"              "Switzerland"   "Thailand"      "Georgia"
[57] "Greece"               "Sudan"         "Bhutan"        "Ecuador"
[61] "South Sudan"          "Spain"
```

# 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: