Tarefa da Semana 1

Leonardo da Silva Correa

20/04/2021

Mental health in tech survey

```
df
```

```
## # A tibble: 1,259 x 27
##    Timestamp             Age Gender Country   state self_employed family_history
##    <dttm>              <dbl> <chr>  <chr>     <chr> <chr>         <chr>
##  1 2014-08-27 11:29:31    37 Female United S~ IL    <NA>          No
##  2 2014-08-27 11:29:37    44 M      United S~ IN    <NA>          No
##  3 2014-08-27 11:29:44    32 Male   Canada    <NA>  <NA>          No
##  4 2014-08-27 11:29:46    31 Male   United K~ <NA>  <NA>          Yes
##  5 2014-08-27 11:30:22    31 Male   United S~ TX    <NA>          No
##  6 2014-08-27 11:31:22    33 Male   United S~ TN    <NA>          Yes
##  7 2014-08-27 11:31:50    35 Female United S~ MI    <NA>          Yes
##  8 2014-08-27 11:32:05    39 M      Canada    <NA>  <NA>          No
##  9 2014-08-27 11:32:39    42 Female United S~ IL    <NA>          Yes
## 10 2014-08-27 11:32:43    23 Male   Canada    <NA>  <NA>          No
## # ... with 1,249 more rows, and 20 more variables: treatment <chr>,
## #   work_interfere <chr>, no_employees <chr>, remote_work <chr>,
## #   tech_company <chr>, benefits <chr>, care_options <chr>,
## #   wellness_program <chr>, seek_help <chr>, anonymity <chr>, leave <chr>,
## #   mental_health_consequence <chr>, phys_health_consequence <chr>,
## #   coworkers <chr>, supervisor <chr>, mental_health_interview <chr>,
## #   phys_health_interview <chr>, mental_vs_physical <chr>,
## #   obs_consequence <chr>, comments <chr>
```

```
# NÚMERO DE ATRIBUTOS
ncol(df)
```

```
## [1] 27
```

```
# NÚMERO DE INSTÂNCIAS
nrow
```

```
## function (x)
## dim(x)[1L]
## <bytecode: 0x5556552e2700>
## <environment: namespace:base>
```

```
# TIPOS DOS ATRIBUTOS
str(df)
```

```
## spec_tbl_df[,27] [1,259 x 27] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Timestamp                : POSIXct[1:1259], format: "2014-08-27 11:29:31" "2014-08-27 11:29:37" .
##  $ Age                      : num [1:1259] 37 44 32 31 31 33 35 39 42 23 ...
##  $ Gender                   : chr [1:1259] "Female" "M" "Male" "Male" ...
##  $ Country                  : chr [1:1259] "United States" "United States" "Canada" "United Kingdom"
##  $ state                    : chr [1:1259] "IL" "IN" NA NA ...
##  $ self_employed            : chr [1:1259] NA NA NA NA ...
##  $ family_history           : chr [1:1259] "No" "No" "No" "Yes" ...
##  $ treatment                : chr [1:1259] "Yes" "No" "No" "Yes" ...
##  $ work_interfere           : chr [1:1259] "Often" "Rarely" "Rarely" "Often" ...
##  $ no_employees             : chr [1:1259] "6-25" "More than 1000" "6-25" "26-100" ...
##  $ remote_work              : chr [1:1259] "No" "No" "No" "No" ...
##  $ tech_company             : chr [1:1259] "Yes" "No" "Yes" "Yes" ...
##  $ benefits                 : chr [1:1259] "Yes" "Don't know" "No" "No" ...
##  $ care_options             : chr [1:1259] "Not sure" "No" "No" "Yes" ...
##  $ wellness_program         : chr [1:1259] "No" "Don't know" "No" "No" ...
##  $ seek_help                : chr [1:1259] "Yes" "Don't know" "No" "No" ...
##  $ anonymity                : chr [1:1259] "Yes" "Don't know" "Don't know" "No" ...
##  $ leave                    : chr [1:1259] "Somewhat easy" "Don't know" "Somewhat difficult" "Somewha
##  $ mental_health_consequence: chr [1:1259] "No" "Maybe" "No" "Yes" ...
##  $ phys_health_consequence  : chr [1:1259] "No" "No" "No" "Yes" ...
##  $ coworkers                : chr [1:1259] "Some of them" "No" "Yes" "Some of them" ...
##  $ supervisor               : chr [1:1259] "Yes" "No" "Yes" "No" ...
##  $ mental_health_interview  : chr [1:1259] "No" "No" "Yes" "Maybe" ...
##  $ phys_health_interview    : chr [1:1259] "Maybe" "No" "Yes" "Maybe" ...
##  $ mental_vs_physical       : chr [1:1259] "Yes" "Don't know" "No" "No" ...
##  $ obs_consequence          : chr [1:1259] "No" "No" "No" "Yes" ...
##  $ comments                 : chr [1:1259] NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Timestamp = col_datetime(format = ""),
##   ..   Age = col_double(),
##   ..   Gender = col_character(),
##   ..   Country = col_character(),
##   ..   state = col_character(),
##   ..   self_employed = col_character(),
##   ..   family_history = col_character(),
##   ..   treatment = col_character(),
##   ..   work_interfere = col_character(),
##   ..   no_employees = col_character(),
##   ..   remote_work = col_character(),
##   ..   tech_company = col_character(),
##   ..   benefits = col_character(),
##   ..   care_options = col_character(),
##   ..   wellness_program = col_character(),
##   ..   seek_help = col_character(),
##   ..   anonymity = col_character(),
##   ..   leave = col_character(),
##   ..   mental_health_consequence = col_character(),
##   ..   phys_health_consequence = col_character(),
##   ..   coworkers = col_character(),
##   ..   supervisor = col_character(),
##   ..   mental_health_interview = col_character(),
##   ..   phys_health_interview = col_character(),
##   ..   mental_vs_physical = col_character(),
##   ..   obs_consequence = col_character(),
##   ..   comments = col_character()
##   .. )
```

```
# SUMÁRIO
numericColumns<-select(df,where(is.numeric))
summary(numericColumns)
```

```
##       Age
##  Min.   :-1.726e+03
##  1st Qu.: 2.700e+01
##  Median : 3.100e+01
##  Mean   : 7.943e+07
##  3rd Qu.: 3.600e+01
##  Max.   : 1.000e+11
```

```
#VALOR MÍNIMO
df %>% summarize_if(is.numeric, min)
```

```
## # A tibble: 1 x 1
##     Age
##   <dbl>
## 1 -1726
```

```
#VALOR MÁXIMO
df %>% summarize_if(is.numeric, max)
```

```
## # A tibble: 1 x 1
##           Age
##         <dbl>
## 1 99999999999
```

```
#MÉDIA
df %>% summarize_if(is.numeric, mean)
```

```
## # A tibble: 1 x 1
##         Age
##       <dbl>
## 1 79428148.
```

```
#MEDIANA
df %>% summarize_if(is.numeric, median)
```

```
## # A tibble: 1 x 1
##     Age
##   <dbl>
## 1    31
```

```
#DESVIO PADRÃO
df %>% summarize_if(is.numeric, sd)
```

```
## # A tibble: 1 x 1
##           Age
##         <dbl>
## 1 2818299443.
```

```
#ATRIBUTOS CATEGÓRICOS
categoricalColumns<-select(df,where(is.character))
categoricalColumns %>% count(Gender)
```

```
## # A tibble: 47 x 2
##    Gender                 n
##    <chr>              <int>
##  1 A little about you     1
##  2 Agender                1
##  3 All                    1
##  4 Androgyne              1
##  5 Cis Female             1
##  6 cis male               1
##  7 Cis Male               2
##  8 Cis Man                1
##  9 cis-female/femme       1
## 10 Enby                   1
## # ... with 37 more rows
```

```
categoricalColumns %>% count(Country)
```

```
## # A tibble: 48 x 2
##    Country                    n
##    <chr>                  <int>
##  1 Australia                 21
##  2 Austria                    3
##  3 Bahamas, The               1
##  4 Belgium                    6
##  5 Bosnia and Herzegovina     1
##  6 Brazil                     6
##  7 Bulgaria                   4
##  8 Canada                    72
##  9 China                      1
## 10 Colombia                   2
## # ... with 38 more rows
```

```
categoricalColumns %>% count(state)
```

```
## # A tibble: 46 x 2
##    state     n
##    <chr> <int>
##  1 AL        8
##  2 AZ        7
##  3 CA      138
##  4 CO        9
##  5 CT        4
##  6 DC        4
##  7 FL       15
##  8 GA       12
##  9 IA        4
## 10 ID        1
## # ... with 36 more rows
```

```
categoricalColumns %>% count(self_employed)
```

```
## # A tibble: 3 x 2
##   self_employed     n
##   <chr>         <int>
## 1 No             1095
## 2 Yes             146
## 3 <NA>             18
```

```
categoricalColumns %>% count(family_history)
```

```
## # A tibble: 2 x 2
##   family_history     n
##   <chr>          <int>
## 1 No               767
## 2 Yes              492
```

```
categoricalColumns %>% count(treatment)
```

```
## # A tibble: 2 x 2
##   treatment     n
##   <chr>     <int>
## 1 No          622
## 2 Yes         637
```

```
categoricalColumns %>% count(work_interfere)
```

```
## # A tibble: 5 x 2
##   work_interfere     n
##   <chr>          <int>
## 1 Never            213
## 2 Often            144
## 3 Rarely           173
## 4 Sometimes        465
## 5 <NA>             264
```

```
categoricalColumns %>% count(remote_work)
```

```
## # A tibble: 2 x 2
##   remote_work     n
##   <chr>       <int>
## 1 No            883
## 2 Yes           376
```

```
categoricalColumns %>% count(tech_company)
```

```
## # A tibble: 2 x 2
##   tech_company     n
##   <chr>        <int>
## 1 No             228
## 2 Yes           1031
```

```
categoricalColumns %>% count(benefits)
```

```
## # A tibble: 3 x 2
##   benefits       n
##   <chr>      <int>
## 1 Don't know   408
## 2 No           374
## 3 Yes          477
```

```
categoricalColumns %>% count(care_options)
```

```
## # A tibble: 3 x 2
##   care_options     n
##   <chr>        <int>
## 1 No             501
## 2 Not sure       314
## 3 Yes            444
```

```
categoricalColumns %>% count(wellness_program)
```

```
## # A tibble: 3 x 2
##   wellness_program     n
##   <chr>            <int>
## 1 Don't know         188
## 2 No                 842
## 3 Yes                229
```

```
categoricalColumns %>% count(seek_help)
```

```
## # A tibble: 3 x 2
##   seek_help      n
##   <chr>      <int>
## 1 Don't know   363
## 2 No           646
## 3 Yes          250
```

```
categoricalColumns %>% count(anonymity)
```

```
## # A tibble: 3 x 2
##   anonymity      n
##   <chr>      <int>
## 1 Don't know   819
## 2 No            65
## 3 Yes          375
```

```
categoricalColumns %>% count(leave)
```

```
## # A tibble: 5 x 2
##   leave                  n
##   <chr>              <int>
## 1 Don't know           563
## 2 Somewhat difficult   126
## 3 Somewhat easy        266
## 4 Very difficult        98
## 5 Very easy            206
```

```
categoricalColumns %>% count(mental_health_consequence)
```

```
## # A tibble: 3 x 2
##   mental_health_consequence     n
##   <chr>                     <int>
## 1 Maybe                       477
## 2 No                          490
## 3 Yes                         292
```

```
categoricalColumns %>% count(phys_health_consequence)
```

```
## # A tibble: 3 x 2
##   phys_health_consequence     n
##   <chr>                   <int>
## 1 Maybe                     273
## 2 No                        925
## 3 Yes                        61
```

```r
categoricalColumns %>% count(coworkers)
```

```
## # A tibble: 3 x 2
##   coworkers        n
##   <chr>        <int>
## 1 No             260
## 2 Some of them   774
## 3 Yes            225
```

```r
categoricalColumns %>% count(supervisor )
```

```
## # A tibble: 3 x 2
##   supervisor       n
##   <chr>        <int>
## 1 No             393
## 2 Some of them   350
## 3 Yes            516
```

```r
categoricalColumns %>% count(mental_health_interview)
```

```
## # A tibble: 3 x 2
##   mental_health_interview     n
##   <chr>                   <int>
## 1 Maybe                     207
## 2 No                       1008
## 3 Yes                        44
```

```r
categoricalColumns %>% count(phys_health_interview)
```

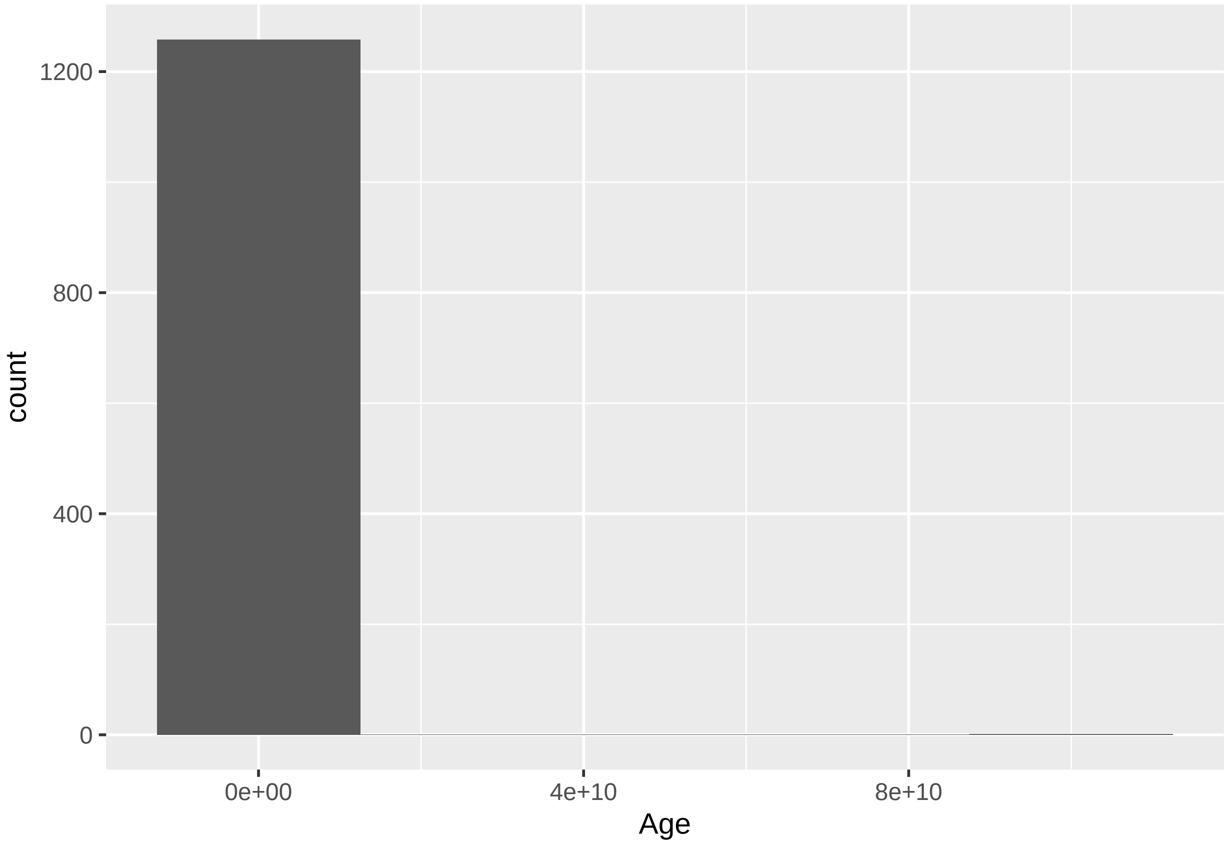
```
## # A tibble: 3 x 2
##   phys_health_interview     n
##   <chr>                 <int>
## 1 Maybe                   557
## 2 No                      500
## 3 Yes                     202
```

```r
categoricalColumns %>% count(mental_vs_physical)
```

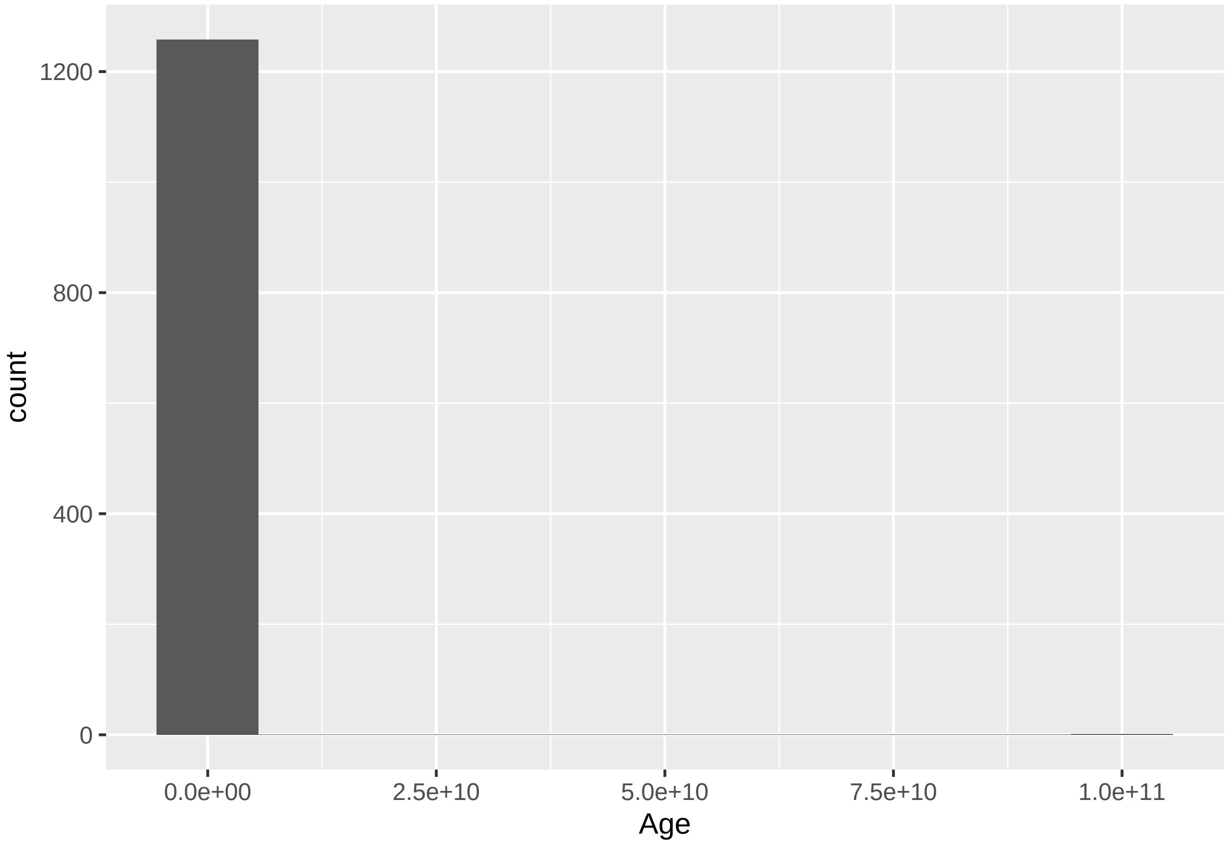
```
## # A tibble: 3 x 2
##   mental_vs_physical     n
##   <chr>              <int>
## 1 Don't know           576
## 2 No                   340
## 3 Yes                  343
```

Representações Gráficas

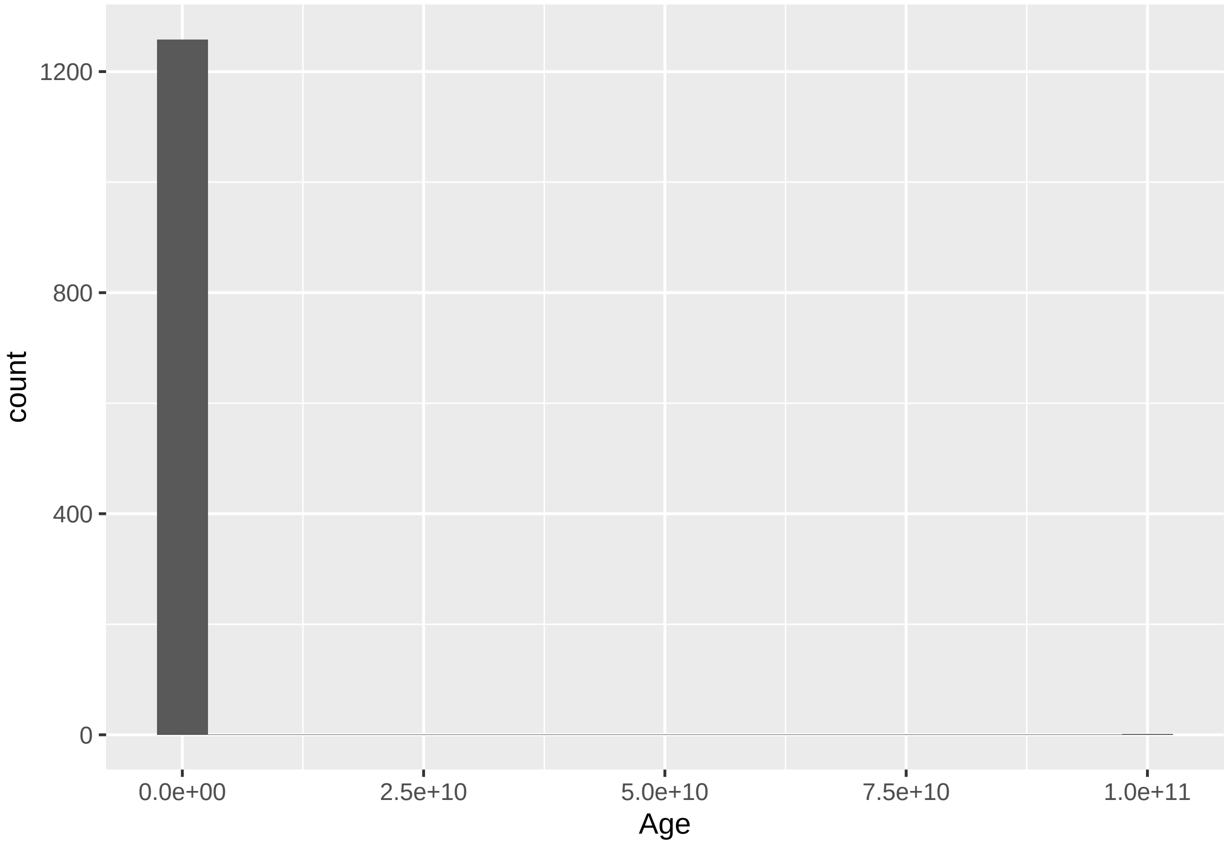
```r
# HITOGRAMA COM DIFERENTES FAIXAS DE VALORES (DADOS ORIGINAIS)
ggplot(df, aes(Age)) + geom_histogram(bins = 5)
```



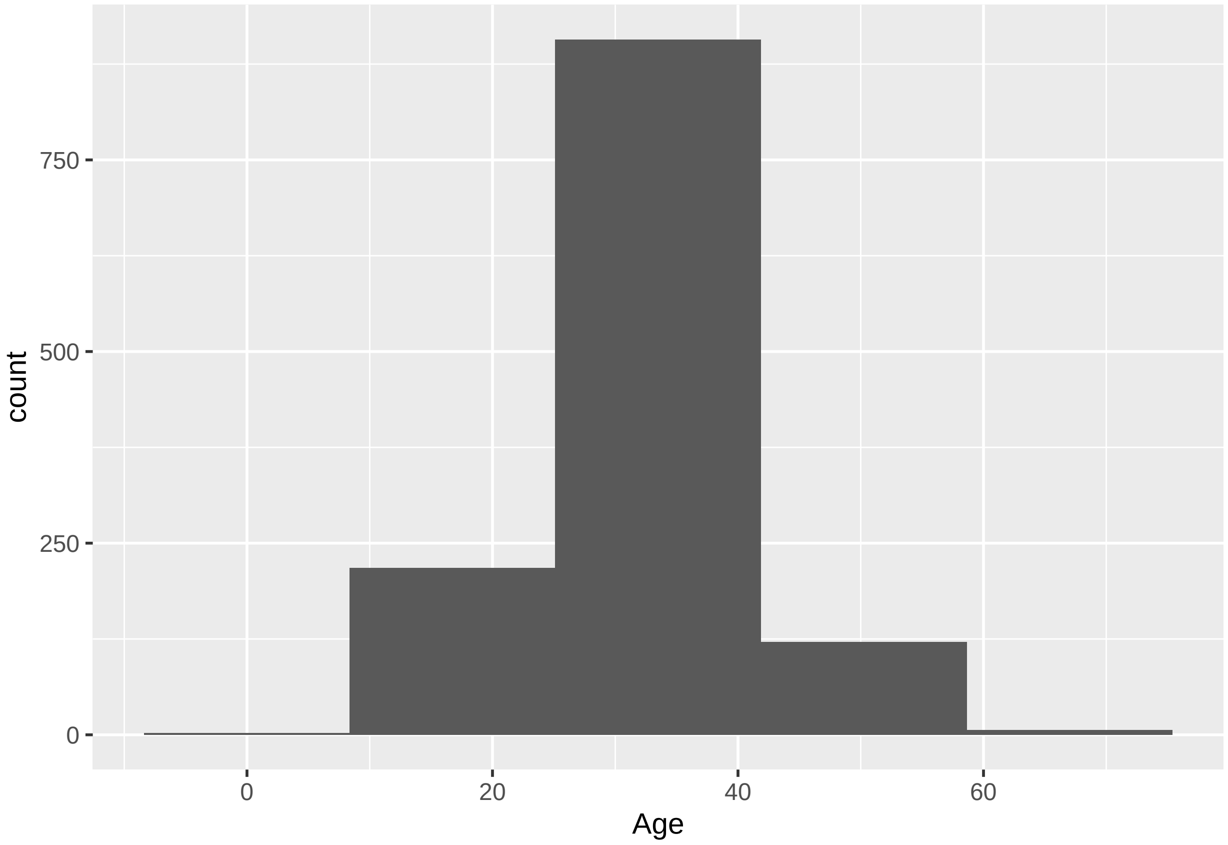
```
ggplot(df, aes(Age)) + geom_histogram(bins = 10)
```



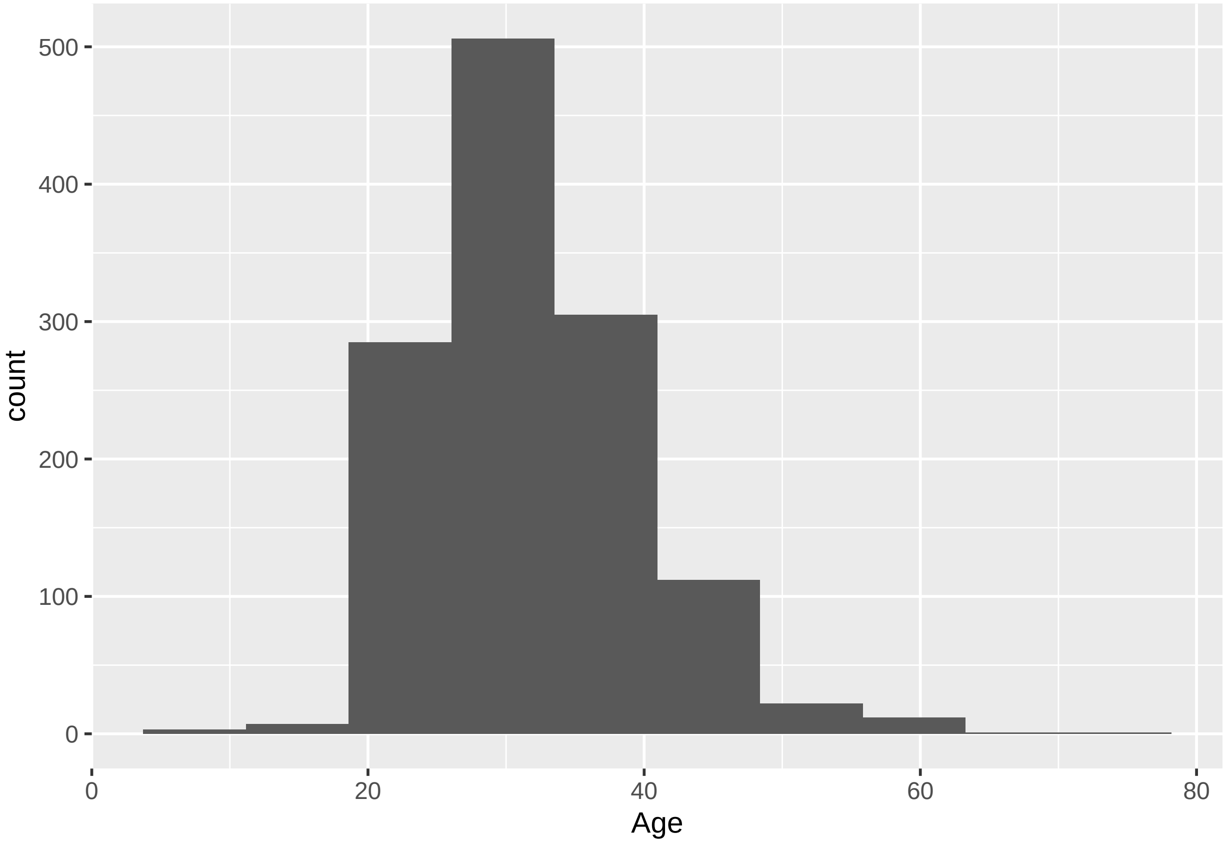
```
ggplot(df, aes(Age)) + geom_histogram(bins = 20)
```



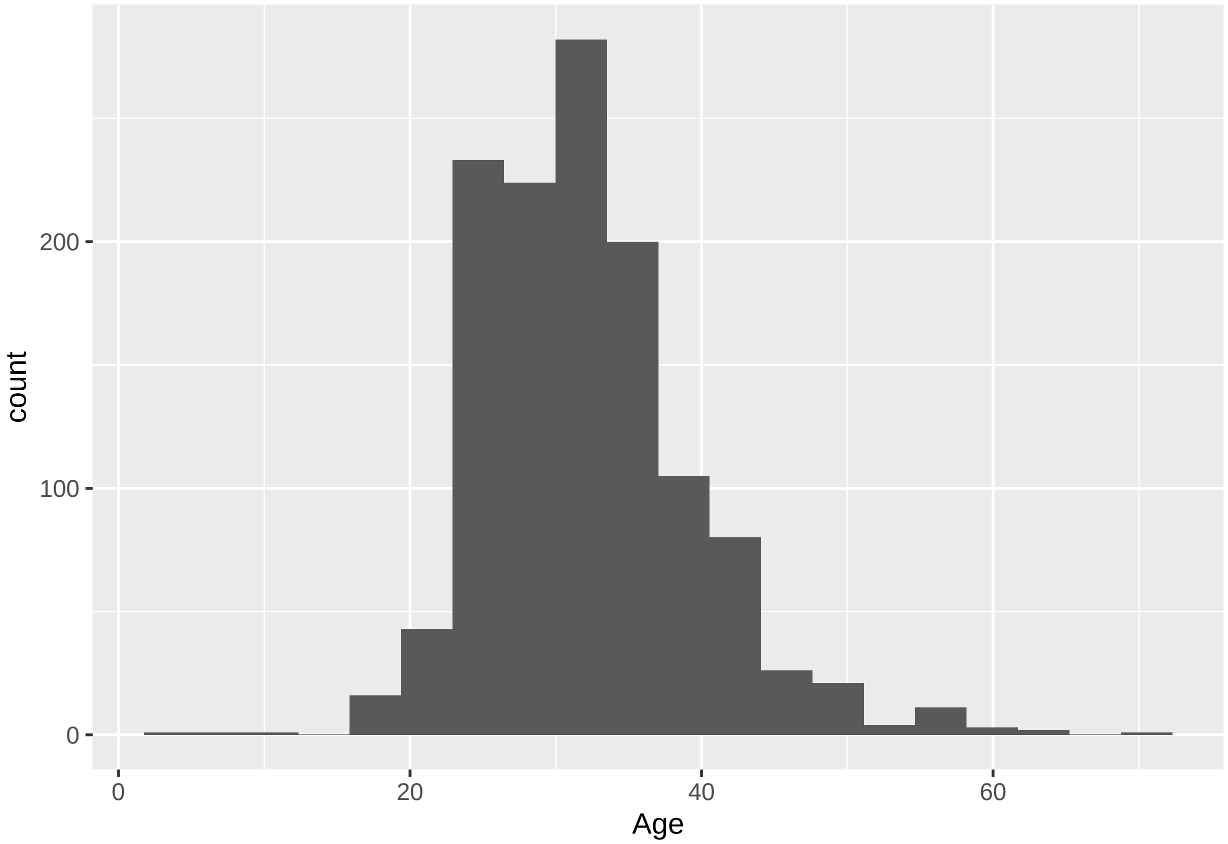
```
# HITOGRAMA COM DIFERENTES FAIXAS DE VALORES (DADOS FILTRADOS / SEM OUTLIERS)
dfCleaned <-filter(select(df, Age),Age >0 & Age <150)
ggplot(dfCleaned, aes(Age)) + geom_histogram(bins = 5)
```



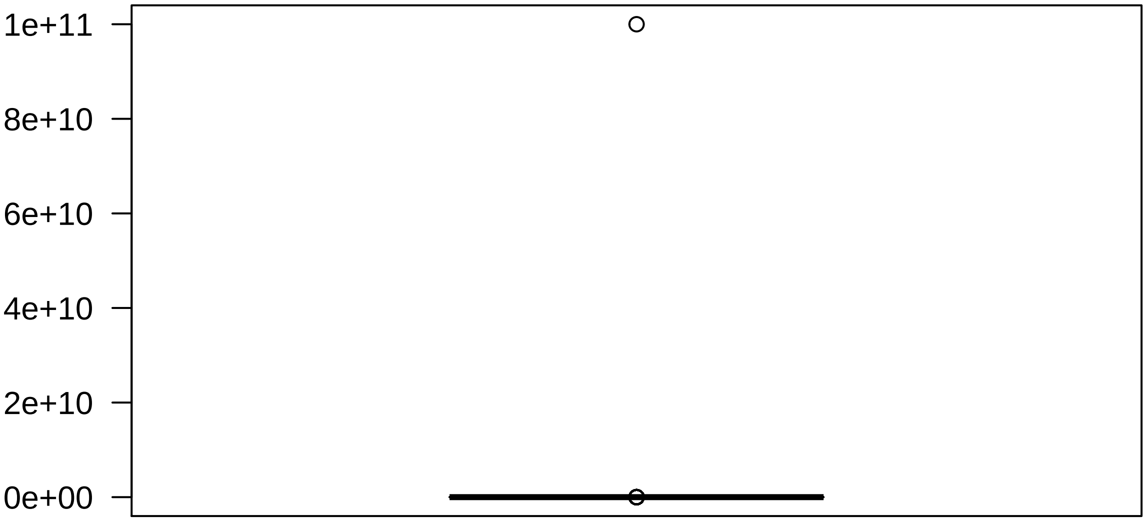
```
ggplot(dfCleaned, aes(Age)) + geom_histogram(bins = 10)
```



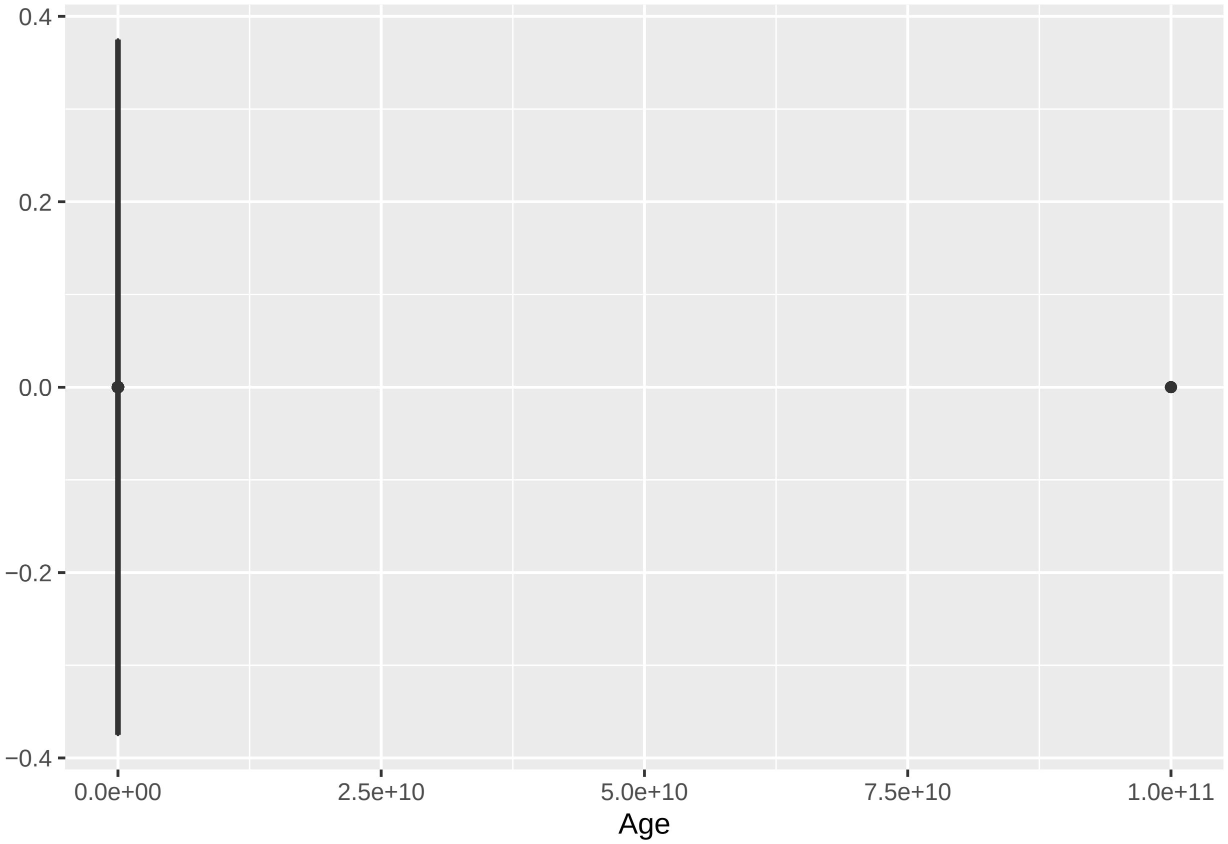
```
ggplot(dfCleaned, aes(Age)) + geom_histogram(bins = 20)
```



```
#BOXPLOT
boxplot(select(df, Age),las=2)
```



```
ggplot(select(df, Age), aes(Age)) + geom_boxplot()
```

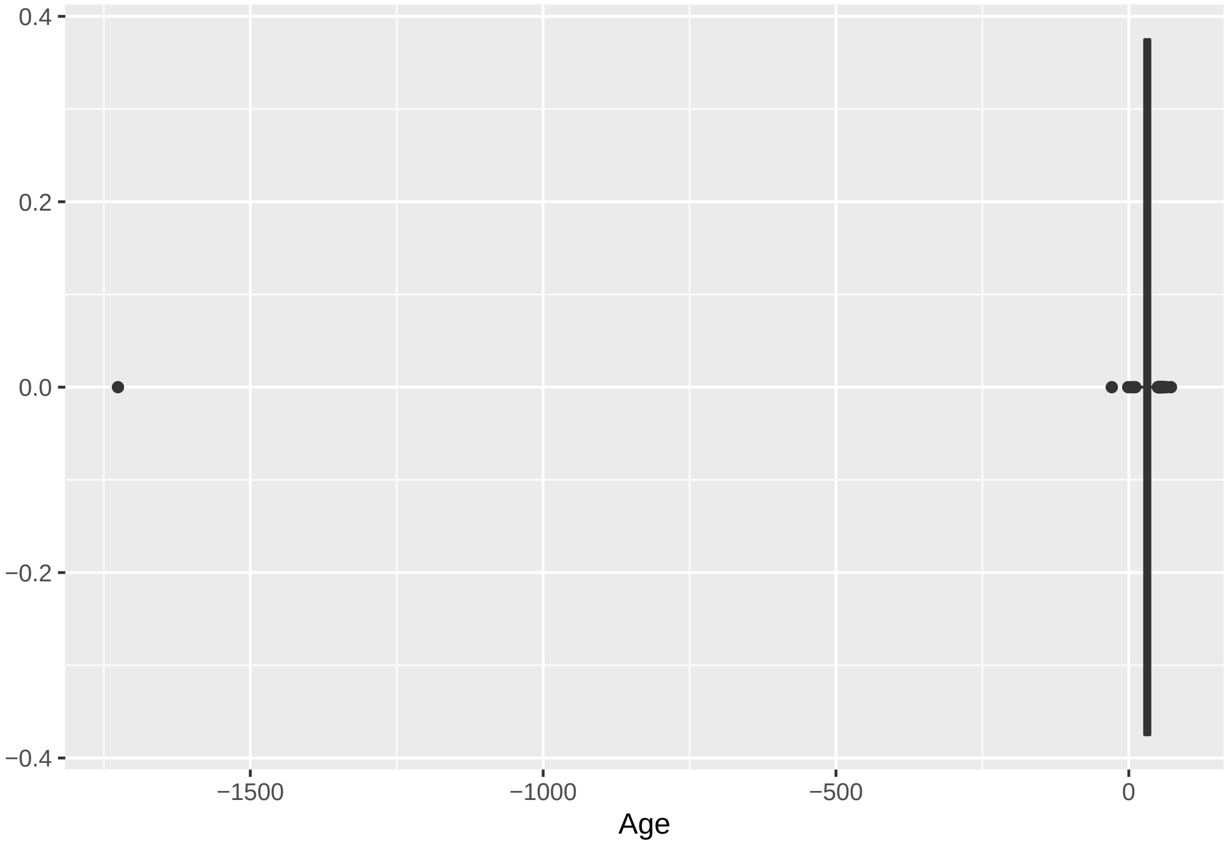


Mostra-se que existe um valor muito alto com OUTLIERS. Abaixo um BOXPLOT eliminando esses valores muito altos.

```
#IDADES ACIMA DE 100 ANOS
dfAge <-filter(select(df, Age),Age >100)
dfAge
```

```
## # A tibble: 2 x 1
##           Age
##         <dbl>
## 1         329
## 2 99999999999
```

```
dfAge <-filter(select(df, Age),Age <100)
ggplot(select(dfAge, Age), aes(Age)) + geom_boxplot()
```

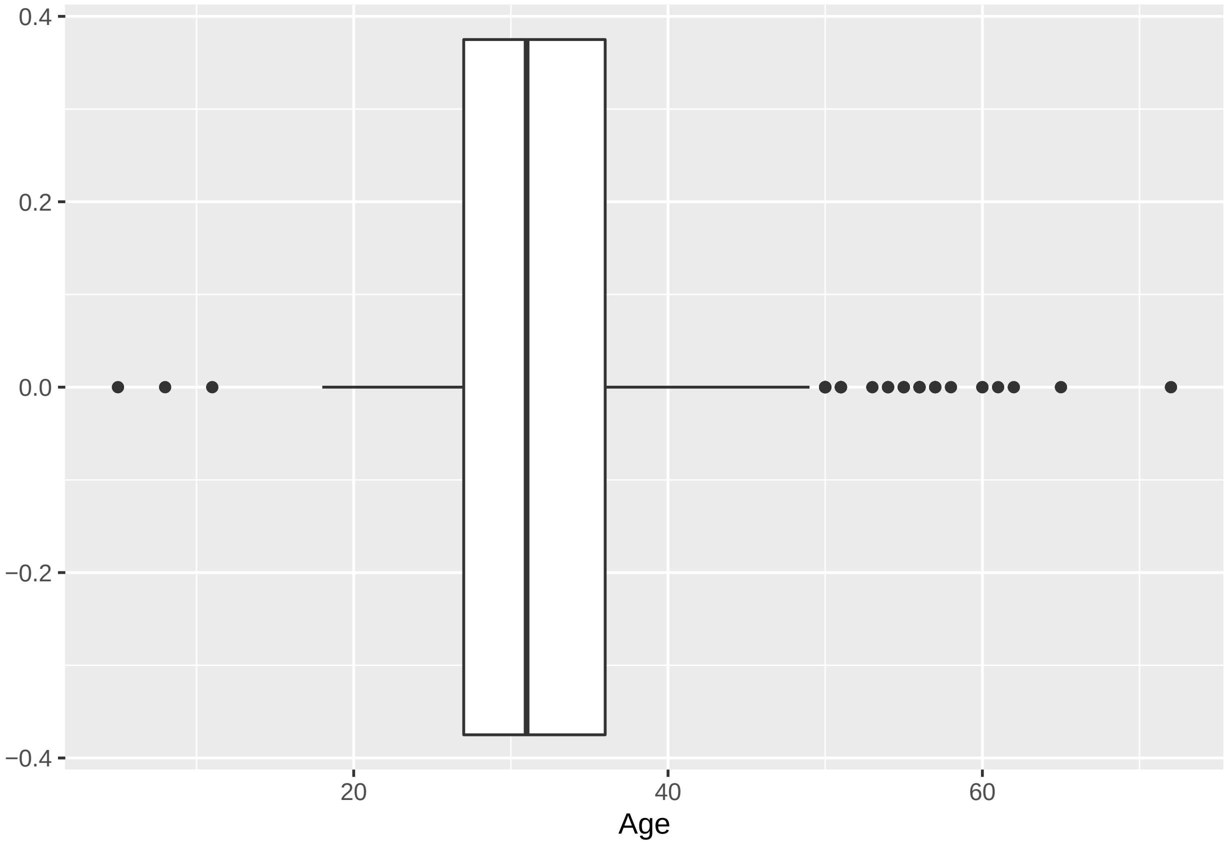


Após remover OUTLIERS altos, verificamos que existem OUTLIERS negativos. Abaixo também mostraremos como ficaria um BOXPLOT eliminando estes OUTLIERS baixos também.

```
#OUTLIERS
dfAge <-filter(select(df, Age),Age < 0 | Age > 100)
dfAge
```

```
## # A tibble: 5 x 1
##           Age
##         <dbl>
## 1         -29
## 2         329
## 3 99999999999
## 4       -1726
## 5          -1
```

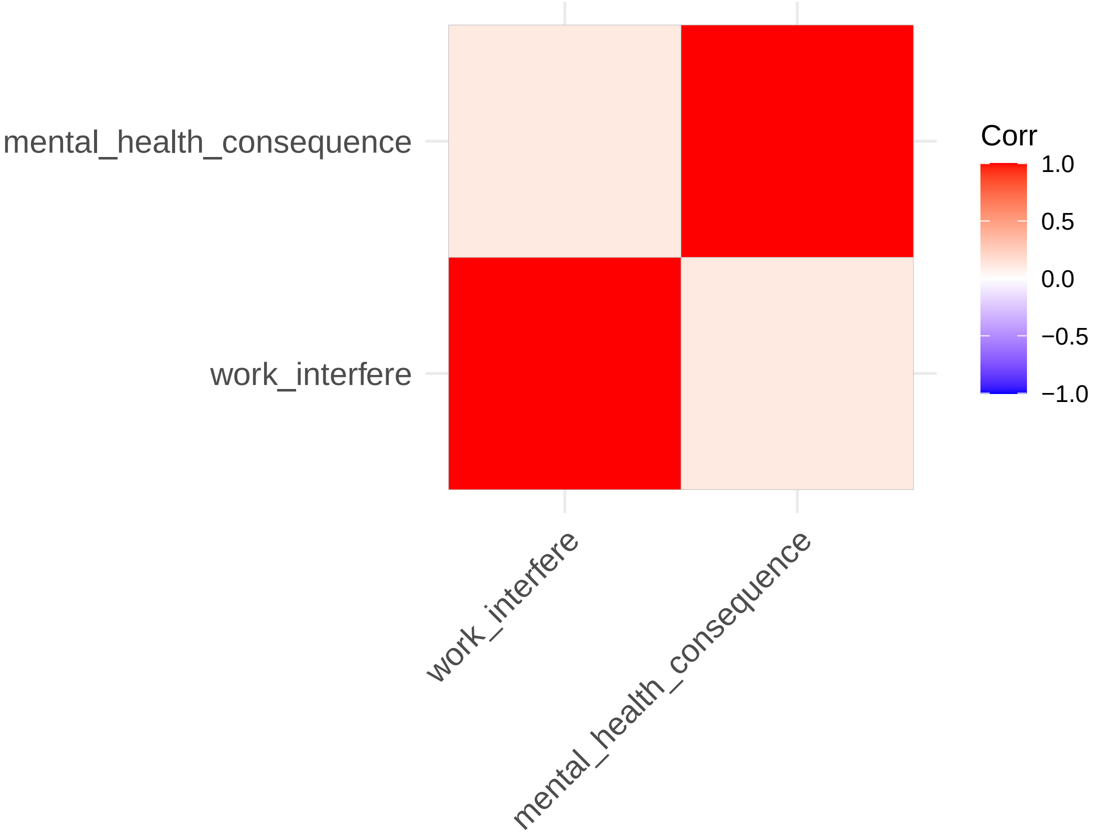
```
dfAge <-filter(select(df, Age),Age >0 & Age <150)
ggplot(select(dfAge, Age), aes(Age)) + geom_boxplot()
```



Matrizes

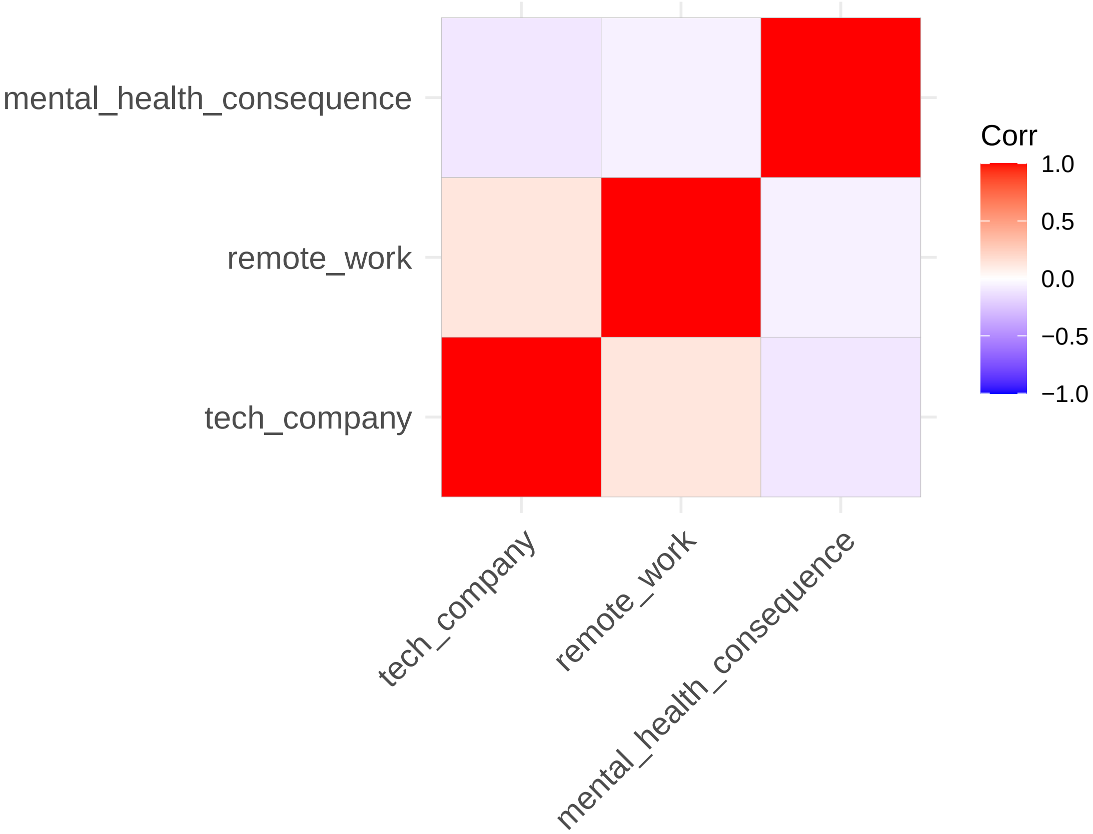
Correlacionando interferências no trabalho com consequências na saúde mental

```
df$work_interfere[is.na(df$work_interfere)] = 0
df$work_interfere[df$work_interfere == "Never"] = 1
df$work_interfere[df$work_interfere == "Rarely"] = 2
df$work_interfere[df$work_interfere == "Sometimes"] = 3
df$work_interfere[df$work_interfere == "Often"] = 4
df$work_interfere = as.numeric(df$work_interfere)
df$mental_health_consequence[df$mental_health_consequence == "No"] = 0
df$mental_health_consequence[df$mental_health_consequence == "Yes"] = 1
df$mental_health_consequence[df$mental_health_consequence == "Maybe"] = 2
df$mental_health_consequence = as.numeric(df$mental_health_consequence)
cm1 <- df %>% select(work_interfere,mental_health_consequence) %>% as.matrix %>% cor()
ggcorrplot(cm1)
```



Outro correlacionamento é Trabalho Remoto em empresas de tecnologia possuem consequências na saúde mental

```
df$tech_company[df$tech_company == "No"] = 0
df$tech_company[df$tech_company == "Yes"] = 1
df$tech_company = as.numeric(df$tech_company)
df$remote_work[df$remote_work == "No"] = 0
df$remote_work[df$remote_work == "Yes"] = 1
df$remote_work = as.numeric(df$remote_work)
cm1 <- df %>% select(tech_company,remote_work,mental_health_consequence) %>% as.matrix %>% cor()
ggcorrplot(cm1)
```



Outros Gráficos

```
groups <- filter(df,Age > 0 & Age < 100) %>% group_by(Country,tech_company)
groups <- groups %>% summarise(Age = mean(Age))
```

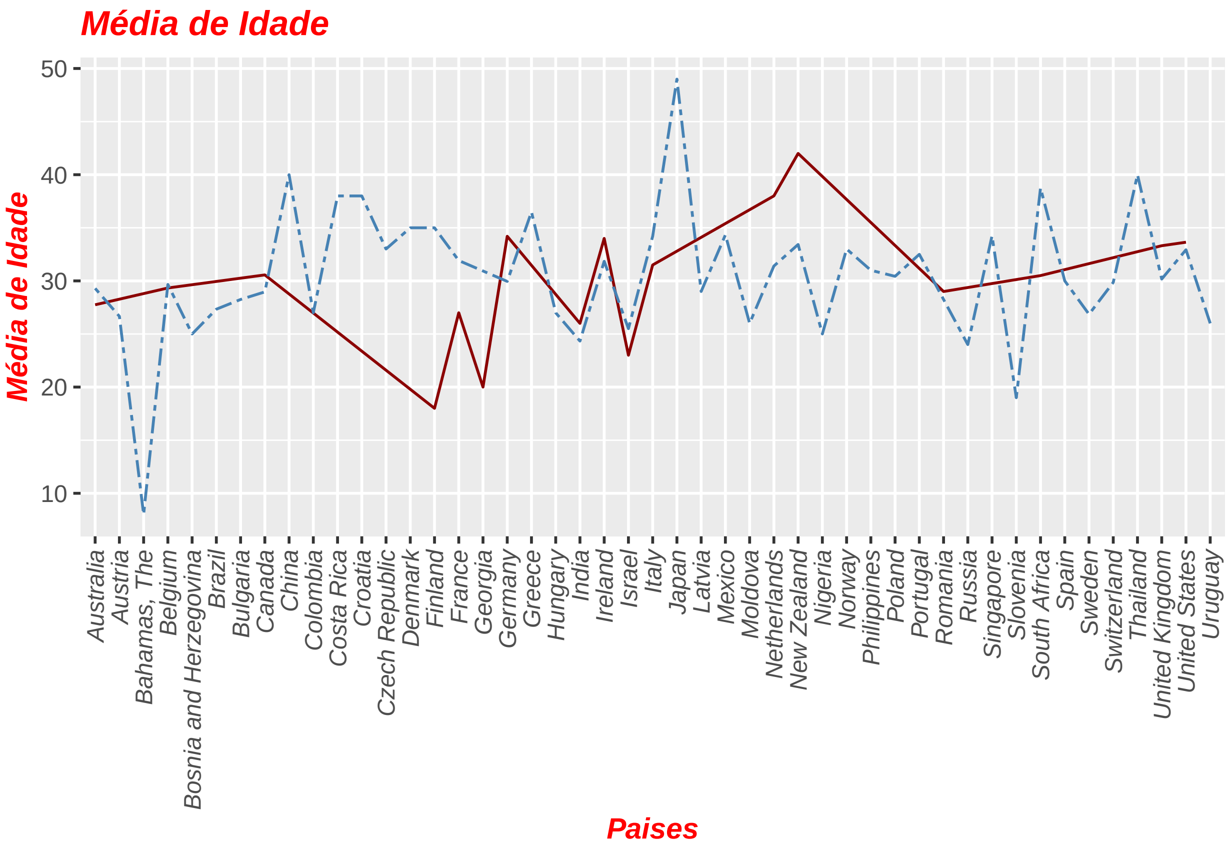
```
## `summarise()` has grouped output by 'Country'. You can override using the `.groups` argument.
```

```
groups
```

```
## # A tibble: 62 x 3
## # Groups:   Country [47]
##    Country                tech_company   Age
##    <chr>                         <dbl> <dbl>
##  1 Australia                         0  27.8
##  2 Australia                         1  29.3
##  3 Austria                           1  26.7
##  4 Bahamas, The                      1   8
##  5 Belgium                           0  29.3
##  6 Belgium                           1  29.7
##  7 Bosnia and Herzegovina            1  25
##  8 Brazil                            1  27.3
##  9 Bulgaria                          1  28.2
## 10 Canada                            0  30.6
## # ... with 52 more rows
```

```
#filter(select(groups,Country,Age),tech_company==0,Country=="United States")

grafico <- ggplot() +  geom_line(data=filter(groups,tech_company==0), aes(x=Country, y=Age, group=1), c
grafico.labs <- grafico + labs(title = "Média de Idade", x = "Paises", y = "Média de Idade")
red.bold.italic.text <- element_text(face = "bold.italic", color = "red")
grafico.labs + theme(title = red.bold.italic.text, axis.title = red.bold.italic.text, axis.text.x = elem
```


Média de Idade
Média de Idade
Australia
Austria
Bahamas, The
Belgium
Bosnia and Herzegovina
Brazil
Bulgaria
Canada
China
Colombia
Costa Rica
Croatia
Czech Republic
Denmark
Finland
France
Georgia
Germany
Greece
Hungary
India
Ireland
Israel
Italy
Japan
Latvia
Mexico
Moldova
Netherlands
New Zealand
Nigeria
Norway
Philippines
Poland
Portugal
Romania
Russia
Singapore
Slovenia
South Africa
Spain
Sweden
Switzerland
Thailand
United Kingdom
United States
Uruguay
Paises