Relatório Atividade 1

Liz Alexandrita de Souza Barreto

20/10/2021

Universidade de Franca
RGM: 21125066

Preparação do Ambiente e Análise dos dados

Este relatório foi feito com o Rmarkdown!

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Leitura do arquivo e entendimento dos campos

```
pacientes <- read_csv('pacientes.csv')
```

```
## Rows: 27846 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): ID, Idade do Segurado, Código do Procedimento Principal
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(pacientes)
```

```
## Rows: 27,846
## Columns: 4
## $ ID                                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, ~
## $ `Idade do Segurado`               <dbl> 38, 44, 60, 37, 66, 67, 48, 76, 57,~
## $ `Código do Procedimento Principal` <dbl> 31101569, 31401155, 10102019, 31307~
## $ `Valor Total Liberado`            <dbl> 52950.00, 66759.96, 30240.78, 58959~
```

Qualidade dos dados

```
pacientes %>% summary
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
##        ID        Idade do Segurado Código do Procedimento Principal
##  Min.   :    1   Min.   :  0.00    Min.   :   20010
##  1st Qu.: 6962   1st Qu.: 37.00    1st Qu.:10102019
##  Median :13924   Median : 55.00    Median :10104020
##  Mean   :13924   Mean   : 53.03    Mean   :17261453
##  3rd Qu.:20885   3rd Qu.: 71.00    3rd Qu.:30804132
##  Max.   :27846   Max.   :107.00    Max.   :84520604
##                                    NA's   :96
##  Valor Total Liberado
##  Min.   :   490.9
##  1st Qu.: 56620.5
##  Median : 71872.5
##  Mean   : 86512.1
##  3rd Qu.: 98470.8
##  Max.   :980639.4
##  NA's   :1
```

Expurgo de NA e limpeza de tipo de dado

```
pacientes <- drop_na(pacientes)
pacientes$ID <- as.factor(pacientes$ID)
pacientes$`Código do Procedimento Principal` <- as.factor(pacientes$`Código do Procedimento Principal`)
```

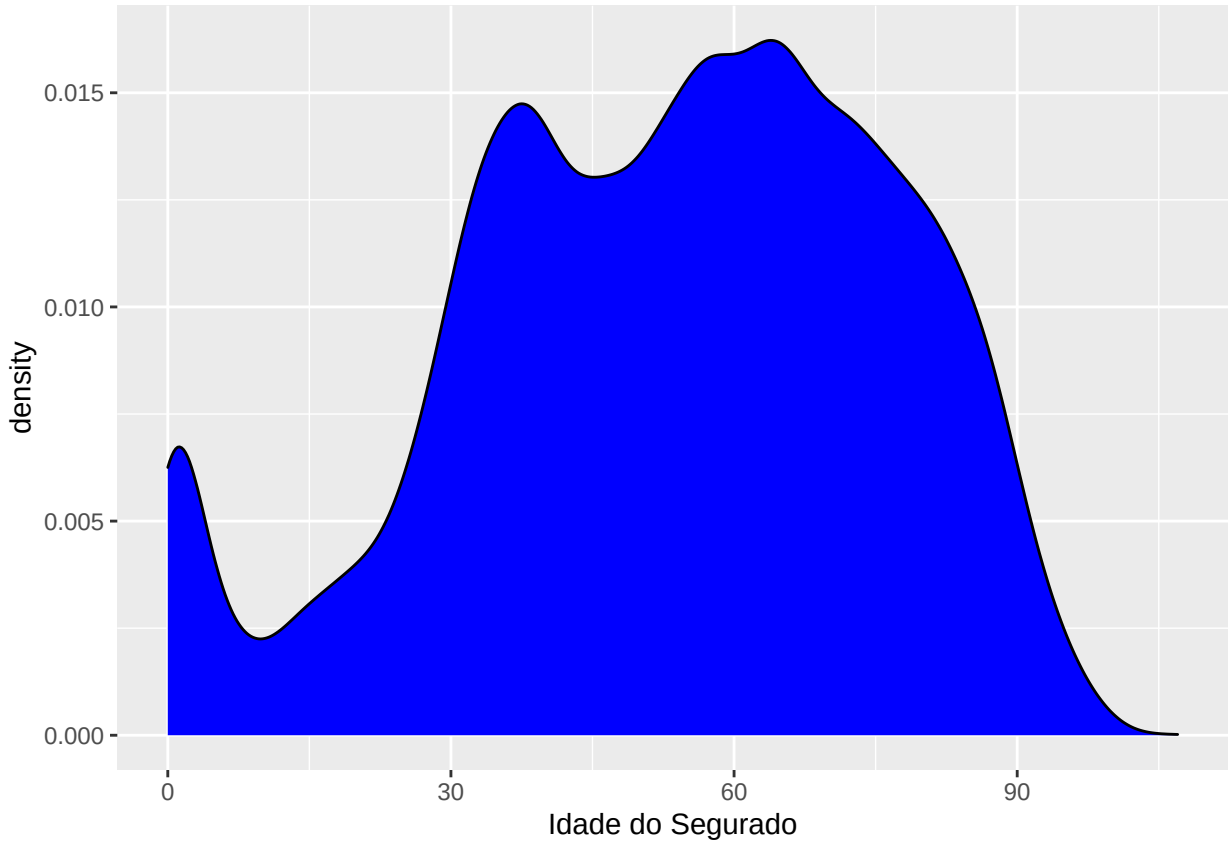
Estatísticas da Idade do Segurado

Vamos estudar algumas estatísticas descritivas da Idade:

```
pacientes %>% summarize(
  min = min(`Idade do Segurado`),
  max = max(`Idade do Segurado`),
  mean = mean(`Idade do Segurado`),
  median = median(`Idade do Segurado`),
  var = var(`Idade do Segurado`),
  sd = var(`Idade do Segurado`) ^ (1/2))
```

```
## # A tibble: 1 x 6
##     min   max  mean median   var    sd
##   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0   107  53.0     55  536.  23.2
```

Gráfico de Densidade da Idade



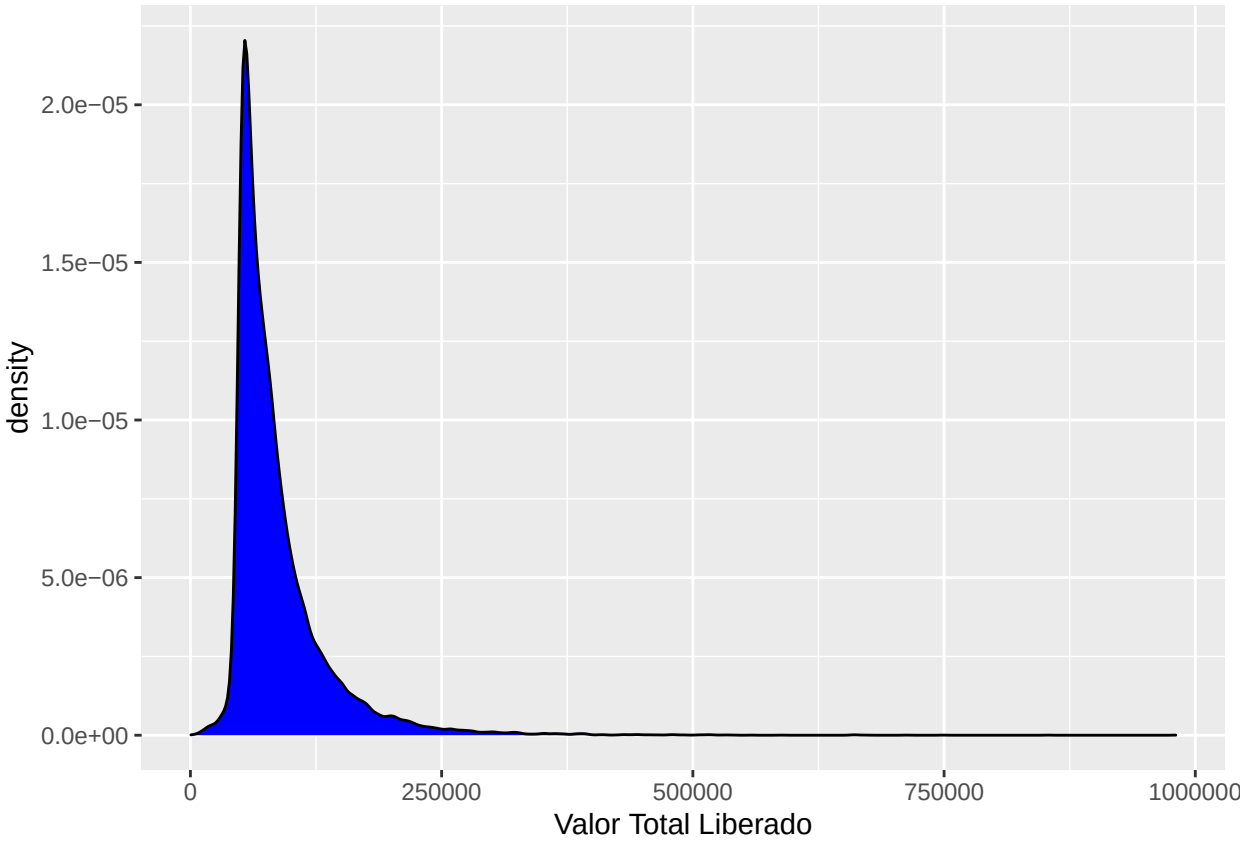
Estatísticas do Valor Total Liberado

Vamos estudar algumas estatísticas do Valor Liberado:

```
pacientes %>% summarize(
  min = min(`Valor Total Liberado`),
  max = max(`Valor Total Liberado`),
  mean = mean(`Valor Total Liberado`),
  median = median(`Valor Total Liberado`),
  var = var(`Valor Total Liberado`),
  sd = var(`Valor Total Liberado`) ^ (1/2))
```

```
## # A tibble: 1 x 6
##     min     max   mean median         var     sd
##   <dbl>   <dbl>  <dbl>  <dbl>       <dbl>  <dbl>
## 1  491. 980639. 86509. 71866. 2403486794. 49025.
```

Gráfico de Densidade do Valor



Estudo de Interações entre Valor e Idade do Segurado

```
pacientes %>% filter(`Valor Total Liberado` == max(`Valor Total Liberado`))
```

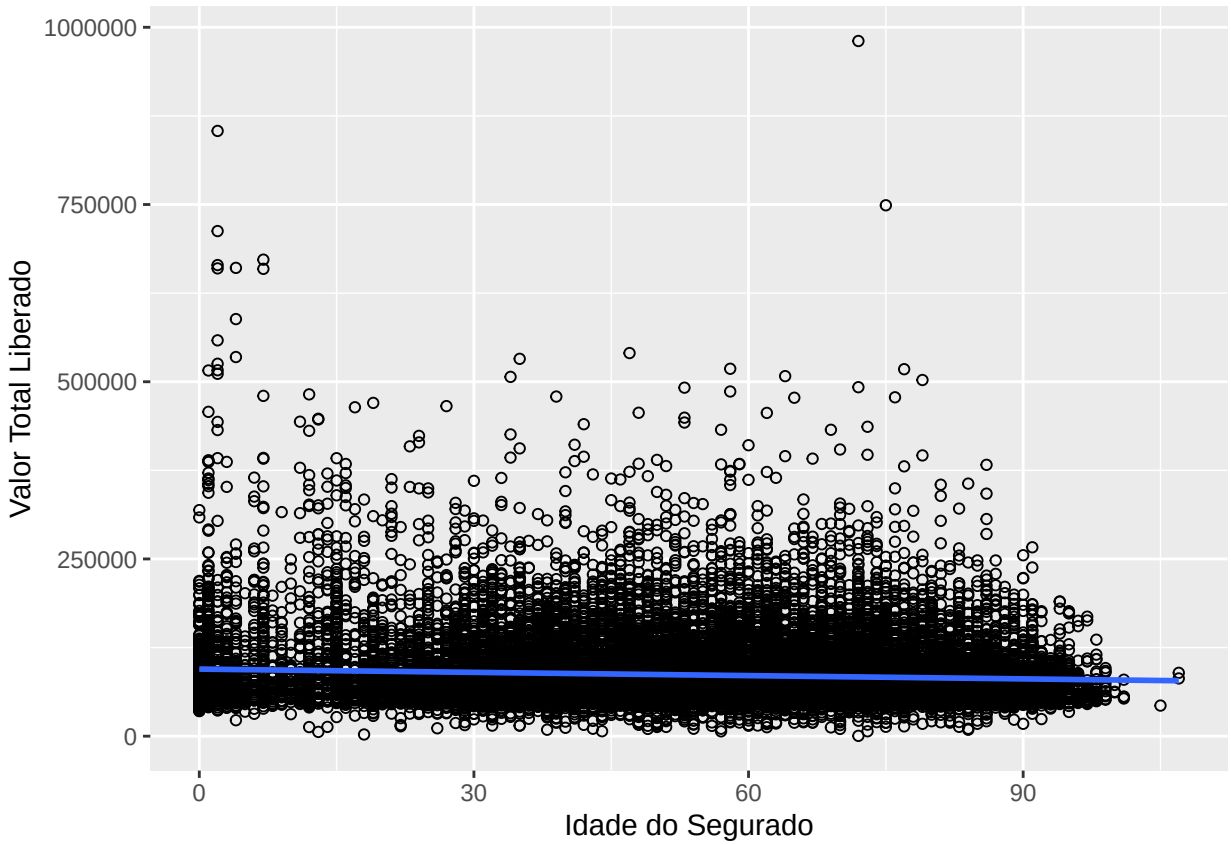
```
## # A tibble: 1 x 4
##   ID    `Idade do Segurado` `Código do Procedimento Principal` `Valor Total Lib~
##   <fct>               <dbl> <fct>                                          <dbl>
## 1 8133                   72 20010                                        980639.
```

```
pacientes %>% filter(`Idade do Segurado` == max(`Idade do Segurado`))
```

```
## # A tibble: 2 x 4
##   ID    `Idade do Segurado` `Código do Procedimento Principal` `Valor Total Lib~
##   <fct>               <dbl> <fct>                                          <dbl>
## 1 2435                  107 20010                                         89387.
## 2 2436                  107 20010                                         81479.
```

Estudo gráfico

```
## `geom_smooth()` using formula 'y ~ x'
```



Conclusões

A variável etária não segue uma distribuição normal e também não segue a pirâmide etária demográfica, porém a maior parte de seus dados cai no intervalo entre os 30 e os 76 anos, segundo a análise gráfica e estatística (53 anos de média +- 23 de desvio padrão) Já a variável de valor liberado parece seguir uma distribuição normal porém com uma longa cauda à direita (skewness positiva) e pequeno desvio padrão em relação ao tamanho da sua cauda. A variância dos dados tanto de idade, mas principalmente de Valor é muito grande (justamente pela cauda da distribuição), o que significa que esse dataset pertenceria, por exemplo a uma seguradora que não tem um nicho etário pra além de adultos pagantes ou em idade economicamente ativa nem de preços de procedimentos. Contra intuitivamente, não há uma correlação positiva entre valor liberado e idade segundo a análise de interação dessas duas variáveis, a tendência é praticamente uniforme puxando para levemente negativa segundo análise gráfica. Apesar disso, o máximo de valor liberado individualmente (foi para paciente de 72 anos) e o total dos valores liberados de pacientes do último quartil (107 anos) serem acima da média. Temos como definir o código dos procedimentos mais liberados em termos de valores, e os mais caros para pacientes mais velhos, no entanto não conseguimos identificá-los através do código.

Código em R na íntegra

```
# Preparação do ambiente
#@title Atividade 1 PIC
#install.packages("tidyverse")
#install.packages("rmarkdown")
library(tidyverse)
library(rmarkdown)
# Ler arquivo e analisar campos
pacientes <- read_csv('pacientes.csv')
```

```r
glimpse(pacientes)
pacientes$ID <- as.factor(pacientes$ID)
pacientes$`Código do Procedimento Principal` <- as.factor(pacientes$`Código do Procedimento Principal`)
# Estatísticas e Gráficos
pacientes %>% ggplot(aes(`Idade do Segurado`)) +
  geom_density(fill = "blue")
summary(pacientes)
pacientes %>% summarize(
  min = min(`Idade do Segurado`),
  max = max(`Idade do Segurado`),
  mean = mean(`Idade do Segurado`),
  median = median(`Idade do Segurado`),
  var = var(`Idade do Segurado`),
  sd = var(`Idade do Segurado`) ^ (1/2))
pacientes %>% ggplot(aes(`Valor Total Liberado`)) +
  geom_density(fill = "blue")
range(pacientes$`Valor Total Liberado`)
pacientes <- drop_na(pacientes)
pacientes %>% summarize(
  min = min(`Valor Total Liberado`),
  max = max(`Valor Total Liberado`),
  mean = mean(`Valor Total Liberado`),
  median = median(`Valor Total Liberado`),
  var = var(`Valor Total Liberado`),
  sd = var(`Valor Total Liberado`) ^ (1/2))
pacientes %>% filter(`Valor Total Liberado` == max(`Valor Total Liberado`))
pacientes %>% filter(`Idade do Segurado` == max(`Idade do Segurado`))
pacientes %>% ggplot(aes(`Valor Total Liberado`)) +
  geom_density(fill = "blue")
pacientes %>% ggplot(aes(`Valor Total Liberado`)) +
  geom_histogram(fill = "blue")
valor_proced <- pacientes %>% group_by(`Código do Procedimento Principal`) %>% summarise(valor_total = s
glimpse(valor_proced)
valor_proced %>% summarize(
  min = min(valor_total),
  max = max(valor_total),
  mean = mean(valor_total),
  median = median(valor_total),
  var = var(valor_total),
  sd = var(valor_total) ^ (1/2))
valor_proced %>% filter(valor_total == max(valor_total))
```