

# Portfólio Gabriel Maciel

Gabriel Maciel

2022-05-31



# Contents

<b>1</b>	<b>Sobre Mim</b>	<b>5</b>
1.1	Quem sou eu? . . . . .	5
1.2	Formação . . . . .	6
1.3	Experiência Profissional . . . . .	6
1.4	Habilidades . . . . .	8
<b>2</b>	<b>Projetos</b>	<b>9</b>
2.1	Banco Iris . . . . .	9
2.2	Data Science e Credit Scoring . . . . .	16



# Chapter 1

## Sobre Mim



Olá, bem vindo ao meu portfólio de ciência de dados, nele eu vou falar um pouco sobre minha relação com estatística e ciência de dados, mostrar minhas habilidades e experiências profissionais.

E-mail: [gabrielmacieldias@hotmail.com](mailto:gabrielmacieldias@hotmail.com)

Celular: (31) 98905-9541

### 1.1 Quem sou eu?

Meu nome é Gabriel Maciel Dias, tenho 23 anos e moro na cidade de Belo Horizonte.

Ingressei no curso de Estatística na UFMG em 2017 através do Enem, onde tive o primeiro contato com o mundo dos dados, desde então venho aprendendo sobre ciência de dados e como ela pode resolver problemas de diversas áreas.

Durante minha trajetória na graduação passei por estágios em diferentes áreas, comecei com uma rápida passagem no Hospital Odilon Behrens, no qual fazia análises voltada para saúde, m seguida fiz a transição para a própria UFMG, trabalhando como estagiário na Prograd fazendo grandes relatórios internos e públicos para avaliar qual é performace dos docentes da UFMG, bem como o perfil dos ingressantes pelo SiSU.

Depois desses estágios comecei minha carreira na área de crédito, iniciei no Banco BDMG onde vi os primeiros modelos de risco, modelagem de perda e estudos de fraude voltados para pessoa jurídica. Trabalhei também no Banco BS2, vendo novamentes os modelos de crédito para pessoa jurídica e também para pessoa física, lá tive mais contato com os desenvolvedores que criavam as API's e engenheiros de dados que preparavam os bancos de dados para as análises.

Atualmente trabalho como Assistente de Crédito no Banco Semear, no qual posso ter ainda mais contato com as esteiras e modelos de créditos, aprendendo diariamente sobre a parte operacional e teórica do crédito.

## 1.2 Formação

Bacharel em Estatística na Universidade Federal de Minas Gerais (UFMG) 2017 - 2021

## 1.3 Experiência Profissional

Desde que comecei na graduação passei por alguns estágios em áreas diferentes. Atualmente trabalho como Assistente de Crédito no Banco Semear e também como cientista de dados na Startup Sua Rua.

### 1.3.1 Hospital Odilon Behrens

Cargo: Estagiário de Estatística - Abril/2018 à Outubro/2018

Responsável pela confecção de relatórios mensais, além do atendimento de demanda de dados, fazendo a retirada deles em um banco específico do hospital. Durante esse período de trabalho elaborei alguns códigos na linguagem de programação R com o objetivo de aperfeiçoar relatórios e análises de dados.

### 1.3.2 Prograd UFMG

Cargo: Estagiário no Setor de Estatística da Prograd UFMG - Novembro/2018 à Junho/2020

Responsável pela elaboração de relatórios com análises descritivas e modelos de regressão, com finalidade de analisar o perfil e o desempenho dos estudantes da UFMG e divulgar o resultado em sites da universidade. Além disso, eram atendidos demandas de dados feitas por funcionários da universidade. Neste período os relatórios foram feitos usando o software R em conjunto com o R Markdown, Latex e excel.

### 1.3.3 Banco de Desenvolvimento de Minas Gerais - BDMG

Cargo: Estagiário de Crédito – Julho/2020 à Dezembro/2020

Responsável pela conferência de planilhas de classificação de risco e criação de programa em R para automatizar determinadas conferências de dados. Além disso, também faço trabalhos com simulações (em Python e R) de carteiras de crédito para prever possíveis perdas. Fiz a criação de um programa em R que fazia a leitura e interpretação de PDF's que verificavam possíveis anomalias, além disso, trabalhei diretamente com análise de fraude.

### 1.3.4 Banco BS2

Cargo: Estagiário de Crédito – Janeiro/2021 à Agosto/2021

Responsável pelos atendimentos de ordens de serviço relacionado a dúvidas sobre análise de crédito além verificar como está o andamento de uma solicitação de crédito dentro do sistema do banco. Além disso, também trabalho na confecção de relatórios Backtest em R em conjunto com R Markdown com base de dados retiradas do SQL Server para verificar como a esteira de crédito está funcionando e se ela está tomando boas decisões de previsão.

### 1.3.5 Banco Semear

Cargo: Assistente de Crédito - Setembro/2021 - Atualmente

Responsável por criação de dashboards para tomada de decisões usando dados de carteira e esteira de crédito. Participo de comitês de crédito, reuniões com fornecedores de dados e de modelos e reuniões com lojistas do banco. Faço também rateio e pagamento de fornecedores, estudo de base de dados para pré aprovado, bem como sugestões de análises que podem ajudar na resolução de problemas de negócio. Todas as análises faço através do software R em conjunto com o Shiny e Excel.

### 1.3.6 Sua Rua

Cargo: Cientista de Dados - Agosto/2021 - Atualmente

Nessa startup trabalho com construção de modelos de dados geográficos, trabalhando com R e QGis, além disso atuo coletando dados e criando robôs para fazer Web Scrapping.

## 1.4 Habilidades

Programação avançada em R

Criação de relatórios com R Markdown

Criação de Dashboards com Shiny

Inglês intermediário

Pacote Office



## Chapter 2

# Projetos

Abaixo tenho alguns projetos análises de dados que fiz para treinar algumas habilidades.

### 2.1 Banco Iris

Para primeira análise não tinha como ser diferente, vou começar com o dataset Iris, que é uma das bases de dados mais utilizadas para fazer análises básicas.

Para verificar a estrutura do banco de dados iris podemos usar a função “head”:

```
head(iris)
```

##	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
## 1	5.1	3.5	1.4	0.2	setosa
## 2	4.9	3.0	1.4	0.2	setosa
## 3	4.7	3.2	1.3	0.2	setosa
## 4	4.6	3.1	1.5	0.2	setosa
## 5	5.0	3.6	1.4	0.2	setosa
## 6	5.4	3.9	1.7	0.4	setosa

Usando a função “head” podemos ver as primeiras linhas da base de dados e suas colunas.

Segundo o UCI Machine Learning, a base do Iris contém ao todo 150 linhas com 5 colunas, sendo:

- Sepal.Length = comprimento das sépalas (em cm);
- Sepal.Width = largura das sépalas (em cm);

- Petal.Length = comprimento das pétalas (em cm);
- Petal.Width = largura das pétalas (em cm);
- Species = espécie das plantas;

Para saber com qual análise posso começar vou verificar antes as classes das variáveis usando a função “glimpse” do pacote “dplyr”.

```
# Carregando a função dplyr
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
glimpse(iris)
```

```
## Rows: 150
## Columns: 5
## $ Sepal.Length <dbl> 5.1, 4.9, 4.7, 4.6, 5.0, 5.4, 4.6, 5.0, 4.4, 4.9, 5.4, 4.~
## $ Sepal.Width <dbl> 3.5, 3.0, 3.2, 3.1, 3.6, 3.9, 3.4, 3.4, 2.9, 3.1, 3.7, 3.~
## $ Petal.Length <dbl> 1.4, 1.4, 1.3, 1.5, 1.4, 1.7, 1.4, 1.5, 1.4, 1.5, 1.5, 1.~
## $ Petal.Width <dbl> 0.2, 0.2, 0.2, 0.2, 0.2, 0.4, 0.3, 0.2, 0.2, 0.1, 0.2, 0.~
## $ Species <fct> setosa, setosa, setosa, setosa, setosa, setosa, setosa, s~
```

Sabendo que as classes são numéricas vou utilizar o “summary” para uma visão geral.

```
summary(iris[,1:4])
```

```
##   Sepal.Length   Sepal.Width   Petal.Length   Petal.Width
## Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
## 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
## Median :5.800   Median :3.000   Median :4.350   Median :1.300
## Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
## 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
## Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Ao analisar os resultados pode-se notar que o comprimento das sépalas e pétalas é maior do que a largura dos mesmos.

Após observar brevemente o banco Iris gostaria de responder as seguintes perguntas:

- O tamanho de sépala e as espécies influenciam no tamanho de pétala?
- Se sim, qual é essa relação?

### 2.1.1 Análise descritiva

Para responder essa pergunta vou começar com análise descritivas, que podem ser simples mas que em muitos casos já conseguem identificar informações valiosas

Vou partir do pressuposto que a média entre comprimento e largura de sépalas e pétalas é um bom indicador do seu tamanho, então farei isso no R.

```
iris <- iris %>%
  mutate(media_petala = (Petal.Length+Petal.Width)/2, # Média das pétalas
         media_sepala = (Sepal.Length+Sepal.Width)/2) %>% # Média das sépalas
  select(media_petala,media_sepala,Species)

head(iris, 5)
```

```
##   media_petala media_sepala Species
## 1         0.80         4.30  setosa
## 2         0.80         3.95  setosa
## 3         0.75         3.95  setosa
## 4         0.85         3.85  setosa
## 5         0.80         4.30  setosa
```

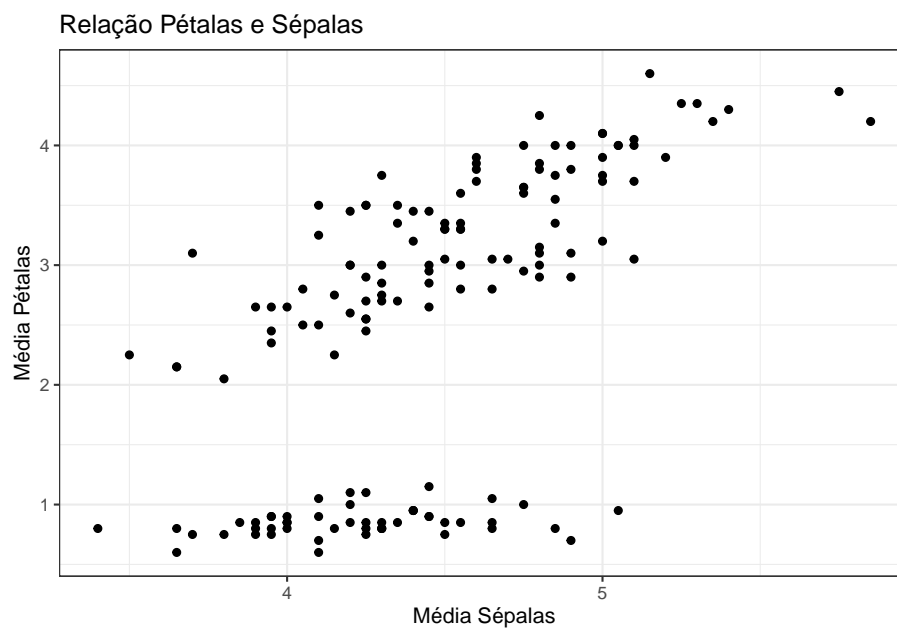
Podemos ver acima como ficou o banco de dados agora com apenas 3 variáveis.

Primeiramente farei um gráfico para verificar visualmente se há uma possível relação entre a média de comprimento e largura das pétalas e sépalas.

```
# Carregando a função do ggplot2

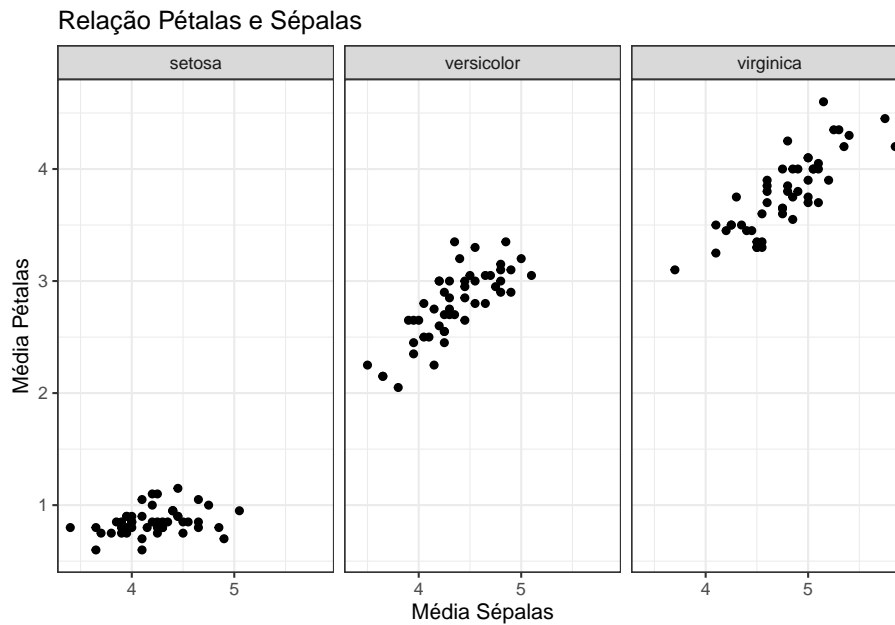
library(ggplot2)

ggplot(iris) +
  geom_point(aes(x = media_sepala,y = media_petala)) +
  labs(title = "Relação Pétalas e Sépalas",
       x = "Média Sépalas",
       y = "Média Pétalas") +
  theme_bw()
```



É possível notar que parece existir alguma relação, porém o efeito das espécies não nos permite enxergar isso da melhor forma, então vou segmentar o gráfico de acordo com a espécie.

```
ggplot(iris) +  
  geom_point(aes(x = media_sepala, y = media_petala)) +  
  labs(title = "Relação Pétalas e Sépalas",  
       x = "Média Sépalas",  
       y = "Média Pétalas") +  
  theme_bw() +  
  facet_grid(~ Species)
```



Agora podemos notar que as espécies parecem influenciar no tamanho das pétalas e sépalas, visto que os pontos se encontram em locais distintos dos gráficos. Além disso, nota-se possível relação positiva entre tamanho de sépalas e pétalas nas espécies versicolor e virginica, já a setosa não apresenta graficamente essa relação.

### 2.1.2 Análise de Regressão

Para avaliar se realmente existe essa relação estatística entre as pétalas, sépalas e espécie de planta eu vou usar a técnica de regressão linear.

```
# Usando a função lm para ajustar o modelo

fit <- lm(media_petala~media_sepala+Species, data = iris)

# Usando a função summary para avaliar resultados dos ajustes

summary(fit)

##
## Call:
## lm(formula = media_petala ~ media_sepala + Species, data = iris)
##
## Residuals:
```

```
##      Min      1Q   Median      3Q      Max
## -0.50872 -0.13092  0.00096  0.14311  0.61936
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.33611    0.18761  -7.122 4.49e-11 ***
## media_sepala     0.51935    0.04398  11.809 < 2e-16 ***
## Speciesversicolor  1.86837    0.04055  46.073 < 2e-16 ***
## Speciesvirginica   2.64209    0.04716  56.025 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2005 on 146 degrees of freedom
## Multiple R-squared:  0.9749, Adjusted R-squared:  0.9744
## F-statistic: 1893 on 3 and 146 DF,  p-value: < 2.2e-16
```

Acima temos alguns resultados interessantes, primeiramente é possível notar que o tamanho médio das sépalas e a espécie influencia no tamanho médio das pétalas, chego nessa conclusão pois o p-valor do modelos para cada um das variáveis é menor que 0,05.

A cada unidade que aumentamos no tamanho médio das sépalas, 0,51 é adicionado no tamanho médio das pétalas. Já quanto as espécies, o tamanho das pétalas na espécie Versicolor é 1,86 vezes maior que na espécie Setosa (que está oculto nos resultados) enquanto as pétalas na espécie Virginica são 2,64 vezes maiores que na Setosa.

Além disso, o  $R^2$  foi de 0,9744, isso indica que aproximadamente 97% da variação do tamanho das pétalas é explicada pelo tamanho das sépalas e espécie.

Antes de chegar a uma conclusão final dos resultados devemos verificar algumas suposições. Para o modelo de regressão linear é necessário normalidade e homocedasticidade dos resíduos.

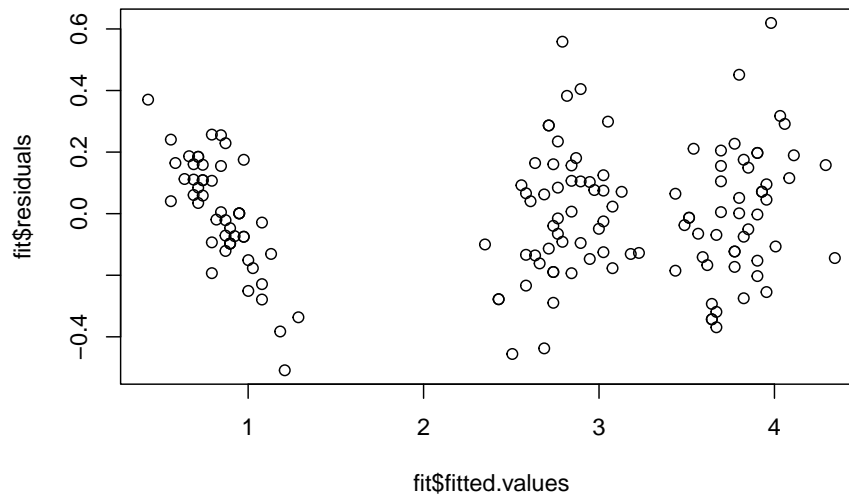
```
# Teste de normalidade de Shapiro Wilk
```

```
shapiro.test(fit$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit$residuals
## W = 0.99395, p-value = 0.7868
```

O p-valor do teste de Shapiro Wilk foi de 0,7868, dessa forma concluímos que os resíduos seguem a distribuição normal.

```
# Verificação de homocedasticidade  
plot(fit$fitted.values,fit$residuals)
```



Acima temos o gráfico dos resíduos pelo valores ajustados, a dispersão dos resíduos ao longo do eixo dos valores ajustados parece ser a mesma, esse é um grande indicativo de homocedasticidade dos resíduos, ou seja, variância constante.

### 2.1.3 Conclusão

Com os pressupostos verificados posso chegar a uma conclusão respondendo as perguntas feitas no início.

O tamanho de sépala e tipo de espécie influenciam sim no tamanho das pétalas, e essa relação acontece da seguinte forma:

- O tamanho de sépala tem uma relação positiva com o tamanho das pétalas, ou seja, quanto maior são as sépalas, maiores vão ser as pétalas.
- As espécies Virginica e Versicolor apresentam tamanho de pétala maior em relação a espécie Setosa

## 2.2 Data Science e Credit Scoring

Como grande entusiasta de modelos estatísticos e de crédito, busco sempre reter mais conhecimento sobre isso, nas últimas semanas venho lendo o livro Credit Scoring do Abharam Laredo. Nele o autor passa toda sua experiência sobre esse tema e no final fornece alguns bancos de dados e um “problema” que pode ser utilizado para colocar em prática o que é aprendido com a leitura. Aproveitei essa oportunidade e usei o problema passado para desenvolver um modelo e uma política de crédito que solucione o caso específico.

Atividade proposta:

“Livraria Dorela é uma cadeia de livrarias que tem quiosques nos principais supermercados das grandes capitais brasileiras. A Dorela passou a fazer o financiamento de livros, de acordo com um score definido de forma subjetiva, a taxa de rejeição era de 30% e a taxa aplicada era muito baixa, visto isso, os resultados não eram satisfatórios.

O novo diretor de crédito da Dorela, que havia atuado como gestor de crédito de uma grande cadeia varejista de moda e tinha experiência no uso de modelos estatísticos decidiu desenvolver um modelo para esse caso.

Ele coletou uma amostra aleatória de 3.000 clientes cujo financiamento foi realizado no período de julho de 2007 a junho de 2008, ele considerou a performance do cliente nos 6 meses seguintes, e classificou como mau cliente aquele que teve qualquer atraso acima de 30 dias, caso contrário era classificado com bom cliente.”

Com esse banco de dados disponível vou iniciar o estudo e propor ao final um modelo e política que atenda a necessidade da livraria Dorela.

### 2.2.1 Banco de Dados

```
# Lendo dataset

data <- readxl::read_xls("351.xls")

# Selecionando as colunas de interesse

data <- data %>%
  dplyr::select(IDADE, UNIFED, FONE, INSTRU, CARTAO, RESTR, RESID, FICCAO, NAOFICCAO, AUTOAJUDA)

# Tipo das colunas da base de dados

glimpse(data)
```



```
## Rows: 3,000
## Columns: 12
## $ IDADE      <dbl> 26, 43, 33, 39, 43, 40, 39, 50, 51, 45, 67, 34, 49, 59, 41, ~
## $ UNIFED     <chr> "SP", "SP", "OUTROS", "RJ", "OUTROS", "RJ", "SP", "SP", "OUT~
## $ FONE       <chr> "SIM", "SIM", "SIM", "SIM", "SIM", "SIM", "SIM", "SIM", "SIM~
## $ INSTRU     <chr> "PRIM & SEC", "SUP", "SUP", "MV", "PRIM & SEC", "PRIM & SEC"~
## $ CARTAO     <chr> "SIM", "NAO", "NAO", "SIM", "SIM", "SIM", "MV", "SIM", "SIM"~
## $ RESTR      <chr> "SIM", "SIM", "SIM", "NAO", "NAO", "NAO", "NAO", "NAO", "SIM"~
## $ RESID      <chr> "PROP", "ALUG", "PROP", "PROP", "PROP", "PROP", "PROP", "PRO~
## $ FICCAO     <chr> "SIM", "NAO", "SIM", "SIM", "SIM", "SIM", "SIM", "SIM", "SIM"~
## $ NAOFICCAO  <chr> "NAO", "NAO", "NAO", "NAO", "NAO", "NAO", "NAO", "NAO", "NAO"~
## $ AUTOAJUDA  <chr> "SIM", "SIM", "NAO", "NAO", "NAO", "NAO", "NAO", "NAO", "NAO"~
## $ CATEG      <dbl> 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, ~
## $ STATUS     <chr> "MAU", "MAU", "BOM", "BOM", "BOM", "BOM", "MAU", "BOM", "BOM"
```

Primeiro passo foi abrir a base de dados, selecionar as colunas que são do meu interesse e aplicar o glimpse para entender as variáveis.

Entre as variáveis, temos:

- Idade: idade do cliente em anos;
- UNIFED: estado do cliente (SP, RJ e Outros);
- FONE: indica se há presença de telefone fixo ou não;
- INSTRU: nível de escolaridade;
- CARTAO: tem cartão de crédito;
- RESTR: apresenta algum restritivo no mercado;
- RESID: mora em residência própria ou alugada;
- FICCAO: comprou livro de ficção;
- NAOFICCAO: comprou livro de não ficção;
- AUTOAJUDA: comprou livro de autoajuda;
- CATEG: comprou dois livros ou mais;
- STATUS: variável resposta, bom ou mau cliente.

Com isso já é possível mapear a classe de cada variável, se é numérica, string ou fator, essa diferença entre as classes pode interferir na análise descritiva e na modelagem.

Ainda há alguns pontos para analisar antes de passar para a parte descritiva, vou verificar se há algum dado faltante ou vazio que ainda não está mapeado.

```
# Conferindo dados faltantes

# NA

table(is.na(data))
```

```
##
## FALSE
## 36000
```

```
# Vazio
```

```
table(data == "")
```

```
##
## FALSE
## 36000
```

Nenhuma das colunas apresentaram dados faltantes (NA) ou em branco.

### 2.2.2 Análise Descritiva

Como estou trabalhando com uma base de dado de variável resposta binária é importante ver qual é a proporção dessa variável, nesse caso, de bons e maus clientes.

```
# Status
```

```
prop.table(table(data$STATUS))
```

```
##
## BOM MAU
## 0.8 0.2
```

Vemos então que de todos os 3000 clientes, 80% são bons e apenas 20% apresentaram atraso maior do que 30 dias.

Agora vou partir para a análise das variáveis preditoras, começando com a variável numérica idade

```
# Descrição com dados simples
```

```
data %>%
  filter(STATUS == "BOM") %>%
  summarise(summary(IDADE)) %>%
  mutate(metrica = c("min", "1 quartil", "mediana", "media", "3 quartil", "max")) %>%
  select(metrica, everything()) %>%
  rename("Bom" = "summary(IDADE)") %>%
  left_join(data %>%
```

```

filter(STATUS == "MAU") %>%
  summarise(summary(IDADE)) %>%
  mutate(metrica = c("min", "1 quartil", "mediana", "media", "3 quartil", "max")) %>%
  select(metrica, everything()), by = "metrica") %>%
  rename("Mau" = "summary(IDADE)")

```

```

## # A tibble: 6 x 3
##   metrica Bom      Mau
##   <chr>   <table> <table>
## 1 min      18.00  18.00000
## 2 1 quartil 34.00  30.00000
## 3 mediana  41.00  37.00000
## 4 media    42.72  38.28333
## 5 3 quartil 49.25  45.00000
## 6 max      88.00  70.00000

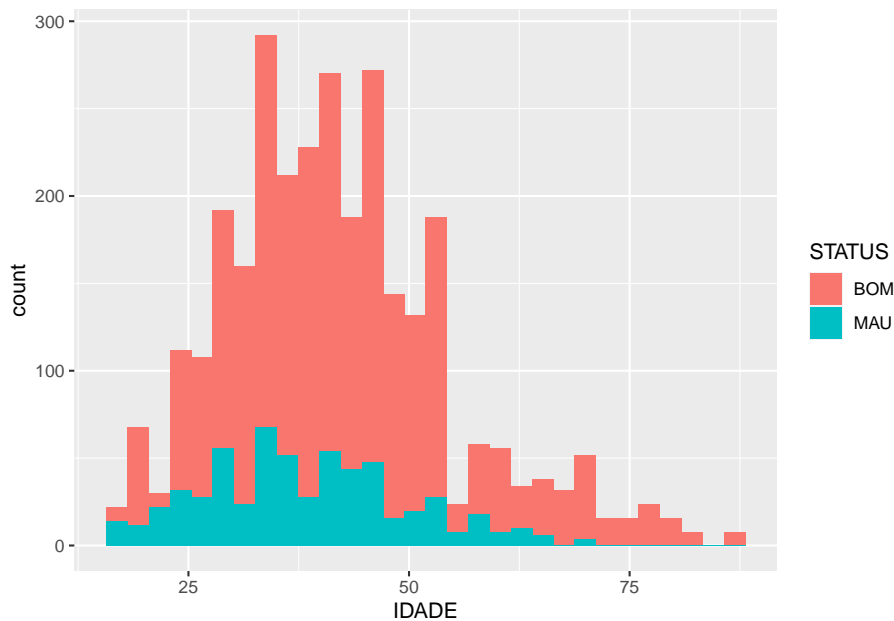
```

```

data %>%
  ggplot(aes(x = IDADE, fill = STATUS)) +
  geom_histogram()

```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



É possível notar acima que parece haver pouca diferença de idade entre as classes de bons e maus pagadores, a média de idade dos bons pagadores é 4 anos maior que a dos maus pagadores.

Para analisar as variáveis categóricas eu criei uma pequena função que cria tabelas com a proporção de bons e maus pagadores em cada uma das classes das variáveis. Veja abaixo a função utilizada e a aplicação para montar a tabela.

```
freq.tab <- function(data,coluna){

  #coluna <- all_of(coluna)

  tabela <- data %>%
    rename("variavel" = coluna) %>%
    group_by(variavel,STATUS) %>%
    filter(STATUS == "BOM") %>%
    summarise(n_bom = n()) %>%
    ungroup() %>%
    select(variavel,n_bom) %>%
    left_join(data %>%
      rename("variavel" = coluna) %>%
      group_by(variavel,STATUS) %>%
      filter(STATUS == "MAU") %>%
      summarise(n_mau = n()) %>%
      ungroup() %>%
      select(variavel,n_mau), by = c("variavel"))

  tabela <- tabela %>%
    mutate(total = n_bom + n_mau,
           perc.bom = paste0(round(n_bom/total,4)*100,"%"),
           perc.mau = paste0(round(n_mau/total,4)*100,"%"),
           perc.total = "100%") %>%
    select(variavel,perc.bom,perc.mau,perc.total)

  colnames(tabela)[1] <- c(coluna)

  return(tabela)

}
```

Utilizando a função para criar as tabelas vou começar analisando as variáveis de UNIFED, FONE, INSTRU e CARTAO.

```
# UNIFED

tabela_unifed <- freq.tab(data,"UNIFED")
```

```

# Fone
tabela_fone <- freq.tab(data,"FONE")

# Instru
tabela_instru <- freq.tab(data,"INSTRU")

# Cartão
tabela_cartao <- freq.tab(data,"CARTAO")

grid.arrange(tableGrob(tabela_unifed, rows = NULL),
              tableGrob(tabela_fone, rows = NULL),
              tableGrob(tabela_instru, rows = NULL),
              tableGrob(tabela_cartao, rows = NULL),
              top = textGrob(c("Análise variáveis"),
                             nrow = 2, ncol = 2)

```

Análise variáveis

UNIFED	perc.bom	perc.mau	perc.tot	FONE	perc.bom	perc.mau	perc.total
OUTROS	72.63%	27.37%	100%	NÃO	81.69%	18.31%	100%
RJ	88.33%	11.67%	100%	SIM	79.82%	20.18%	100%
SP	76.81%	23.19%	100%				

INSTRU	perc.bom	perc.mau	perc.	CARTAO	perc.bom	perc.mau	perc.tota
MV	82.83%	17.17%	100%	MV	71.79%	28.21%	100%
IM & SEC	70.82%	29.18%	100%	NAO	74.39%	25.61%	100%
SUP	83.61%	16.39%	100%	SIM	81.91%	18.09%	100%