Relatório 2

Luciano Belo - 3897

02/09/2021

Introdução

A *estatística* se refere à matemática e às técnicas com as quais entendemos os dados. Como dito por Mark Twain: "Os fatos são teimosos, mas as estatísticas são mais maleáveis".

A estatística é um ramo de grande importância,já que desenvolve técnicas como a coleta de dados e sua organização, interpretação, análise e representação. O uso da matemática para a tomada de decisões vem acompanhando nossa história desde o início das grandes civilizações.

E como seria se a estatística não existisse? Bom, essa pergunta é muito complexa, apesar da dificuldade em perceber, muitas das decisões que tomamos na vida cotidiana são baseadas em estatísticas.

Desta forma, este relatório tem como objetivo relacionar e analisar dados comuns no nosso dia, neste caso um censo feito com os alunos, com auxílio de ferramentas *estatísticas* .

```
# Importação do dataframe
df <- readRDS("data/censo.Rds")
```

Análises

Nesta seção iremos analisar cada uma das variáveis presentes em nosso *dataframe* (tabela de uma base de dados, em que cada linha corresponde a um registo - linha - da tabela e cada coluna corresponde às propriedades - campos - a serem armazenadas para cada registo da tabela).

Variáveis

Antes de começar a fazer as análises precisamos a priori, entender o que são variáveis e quais são as suas classificações. Variável é uma característica de interesse que é medida em cada elemento da amostra ou população em estudo. As variáveis podem ter valores numérico ou não podendo ser classificadas da seguinte forma:

Variáveis Numéricas ou Quantitativas

- Quantitativas Discretas: Se tratam de características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente são o resultado de contagens
- Quantitativas Contínuas: Se tratam de características que assumem valores em escala continua (na reta real), para que os valores fracionais (números com virgula) façam sentido. Usualmente devem ser medidas através de algum instrumento

Variáveis Categóricas ou Qualitativas

São as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos estudados e são divididas em:

- Qualitativas Nominais: Não existe ordenação dentre as categorias. Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.
- Qualitativas Ordinais: Existe uma ordenação entre as categorias. Exemplos: escolaridade (1 grau, 2 graus, 3 graus), estágio da doença (inicial, intermediário, terminal), mês de observação (janeiro, fevereiro, dezembro).

Gráficos

Gráficos são a tentativa de se expressar visualmente dados ou valores numéricos, de maneiras diferentes, assim facilitando a sua compreensão. Existem vários tipos de gráficos, entretando devemos saber quais são os mais adequados a partir dos tipos de variáveis. No caso deste relatório temos dois tipos de variáveis: Categóricas nominal e Numérica Discreta.

Sendo assim, para as variáveis Categóricas nominal nós usaremos gráficos de barra e gráficos de setores. Já para as variáveis Numéricas discretas usaremos Histograma, Polígono de frequência com histograma e Boxplot.

Gráficos de barra

Os gráficos em barras são comumente usados para exibir distribuições de frequências de variáveis qualitativas, como por exemplo a variável “Sexo”.

Gráficos de setores

O gráfico em setores é comumente utilizado para representar parte de um todo, geralmente em percentagens, e é bastante apropriado para mostrar frequências de ocorrências de variáveis qualitativas.

Histograma

O histograma é um gráfico de barras contíguas, e é apropriado para representar distribuições de frequências tando de variáveis quantitativas contínuas como discretas com muitos valores possíveis. Se a variável é quantitativa discreta, mas contém poucos valores possíveis, então os gráficos utilizados para apresentar a distribuição de frequência de variáveis qualitativas podem ser usados para representar sua frequência nos dados observados.

Polígono de frequência

Após a construção do histograma, é importante observar tendência de alturas das barras, para isso é interessante construir o polígono de frequência sobre as barras, que é um gráfico de linhas obtido ligando os pontos médios dos topos de cada barra.

Boxplot

O Boxplot ou box plot é um diagrama de caixa construído utilizando as referências de valores mínimos e máximos, primeiro e terceiro quartil, mediana e outliers da base de dados. O boxplot tem como objetivo estudar as medidas estatística do conjunto de dados, como propriedades de locação, variabilidade, média, e outliers.

Análises Dataframe

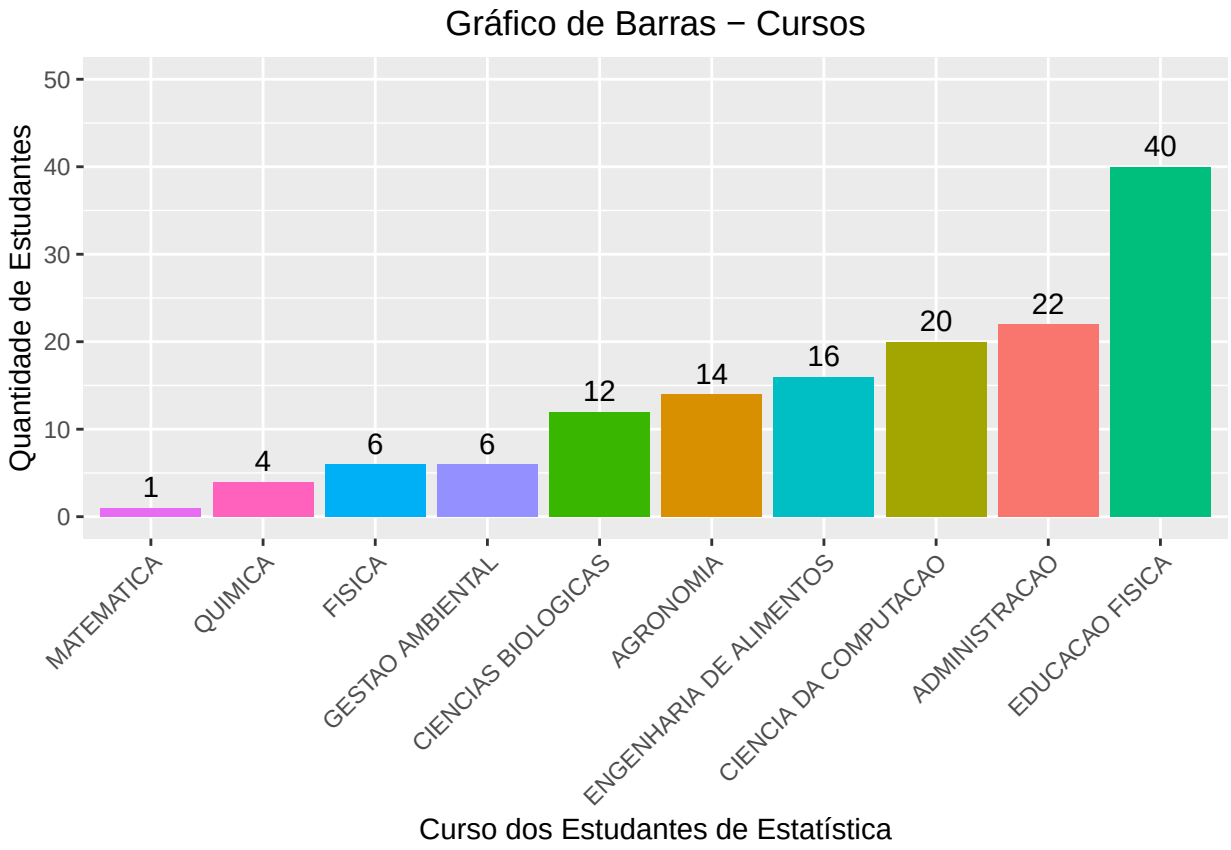
Agora já definidos os conceitos, iremos iniciar as análises. Dentre as variáveis que serão analisadas temos:

- Curso
- Provedor
- Idade
- Residência
- Estado
- Semestre
- Primeira vez na matéria
- Facilidade em Exatas
- Opinião sobre os períodos remotos
- Turma

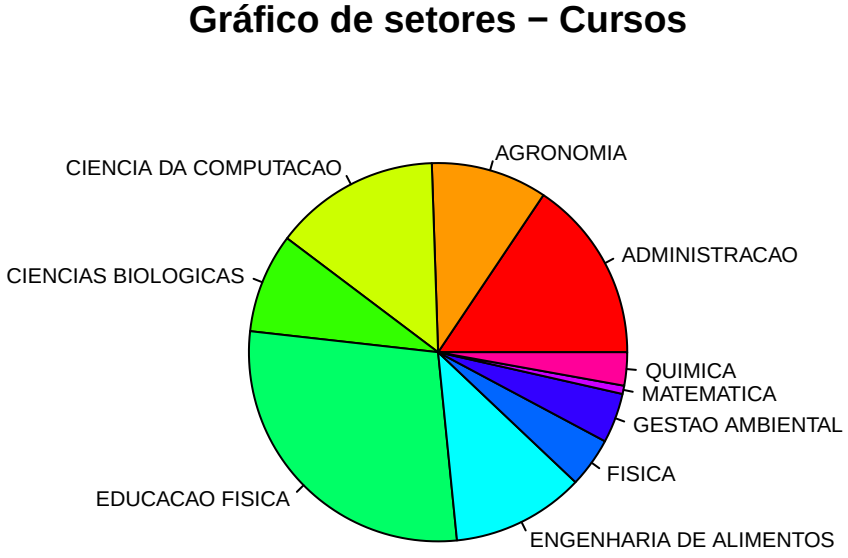
Curso

Nesta subseção será analisado os dados referentes aos Cursos. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

```
curso <- df %>%
  group_by(CURSO) %>%
  summarise(Total=n())

curso %>% ggplot(aes(reorder(CURSO,Total), Total, fill=CURSO)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), vjust=-0.5)+
  ylim(0,50)+
  theme_gray()+
  xlab("Curso dos Estudantes de Estatística")+
  ylab("Quantidade de Estudantes")+
  ggtitle("Gráfico de Barras - Cursos")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
tabela <- table(df$CURSO)
pie(tabela,col=rainbow(10),cex=0.7, main="Gráfico de setores - Cursos")
```

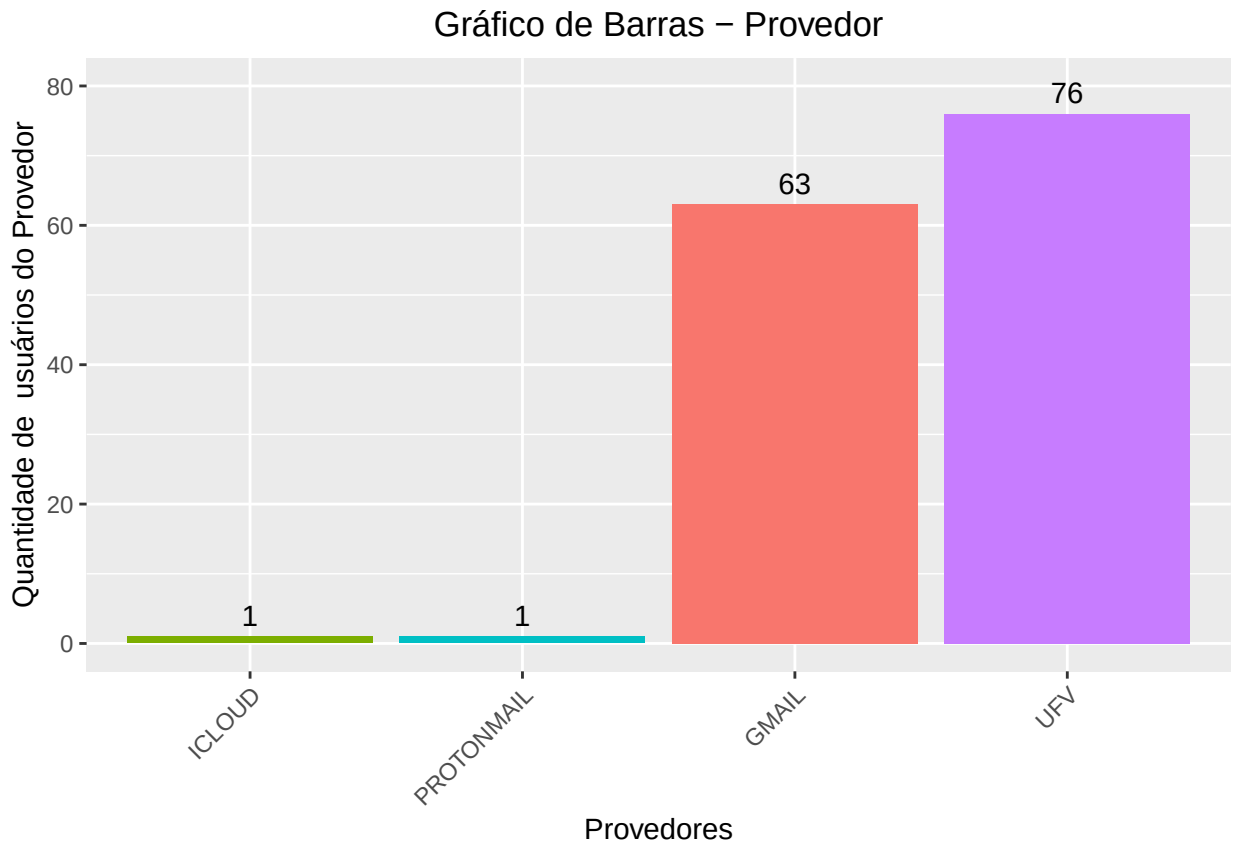


É perceptível a partir dos gráficos que o curso de maior expressão é Educação Física e o de menor matemática. É notável também que os cursos de Ciências Biológicas, Agronomia, Engenharia de Alimentos, Ciência da Computação e Administração possuem quantidades bem próximas de alunos.

Provedor

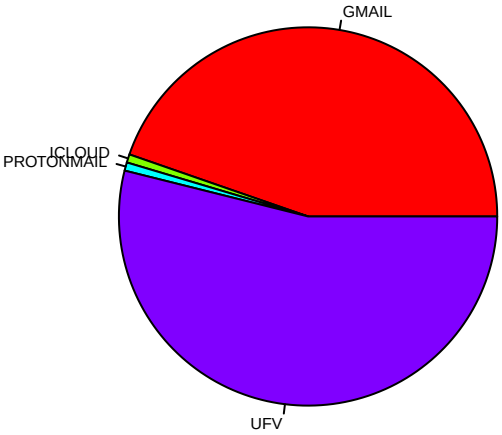
Nesta subseção será analisado os dados referentes aos Provedores de Email. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

```
provedor <- df %>%
  group_by(PROVEDOR) %>%
  summarise(Total=n())
provedor %>% ggplot(aes(reorder(PROVEDOR,Total), Total, fill=PROVEDOR)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), vjust=-0.5)+
  ylim(0,80)+
  theme_gray()+
  xlab("Provedores")+
  ylab("Quantidade de  usuários do Provedor")+
  ggtitle("Gráfico de Barras - Provedor")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
tabela <- table(df$PROVEDOR)
pie(tabela,col=rainbow(4),cex=0.5, main="Gráfico de setores - Provedores")
```

Gráfico de setores − Provedores

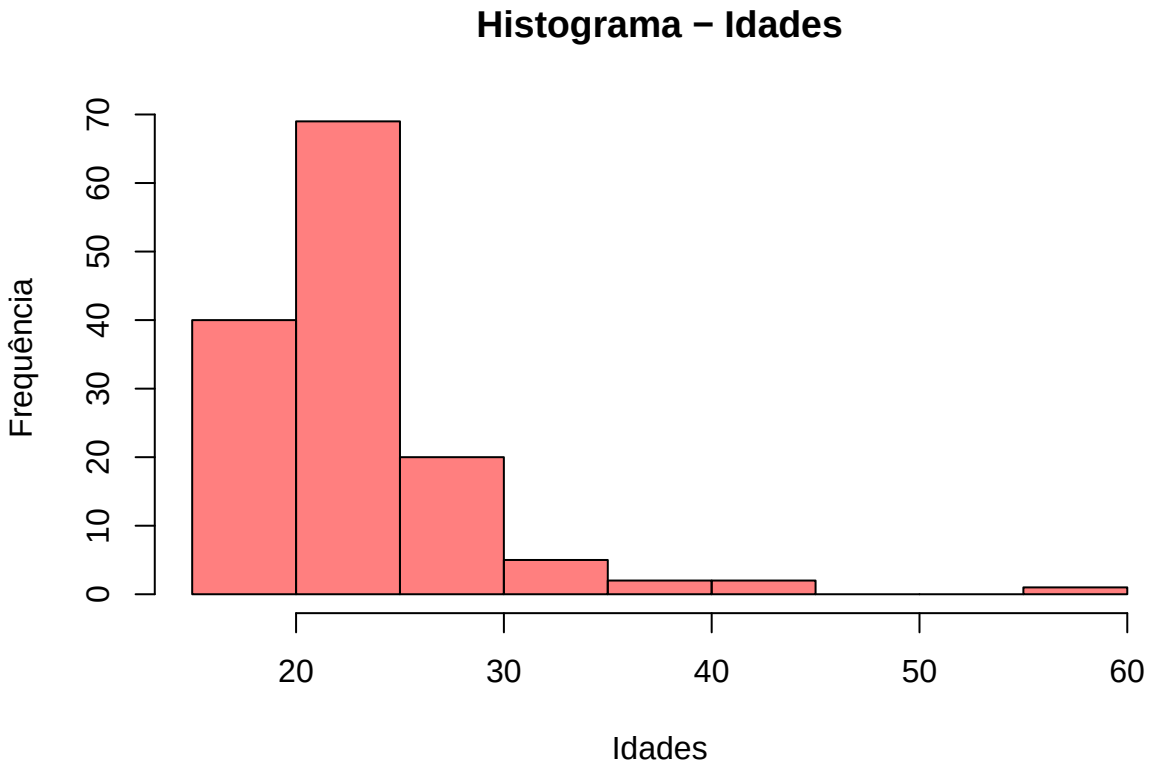


Notamos a partir dos gráficos que os provedores Gmail e UFV são mais expressivos. Vale ressaltar que o gráfico de setores é de grande auxílio quanto a interpretação, já que visualmente conseguimos mensurar os dados como parcelas.

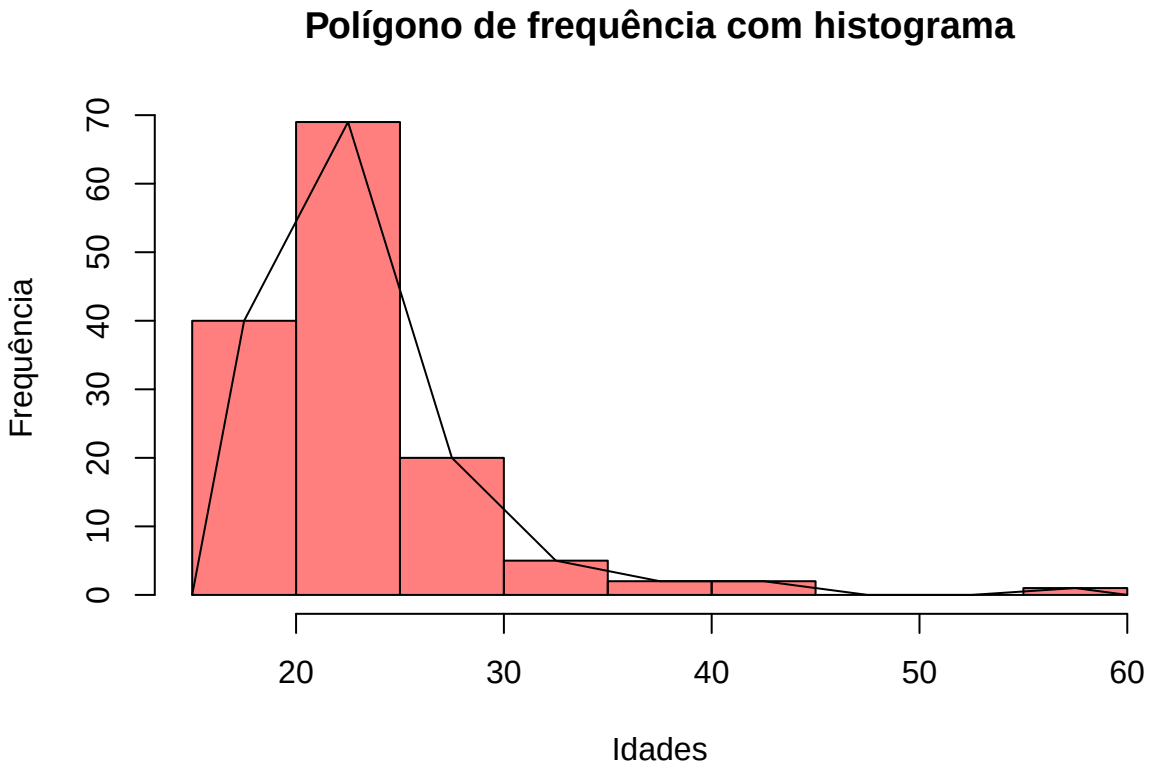
Idade

Nesta subseção será analisado os dados referentes as Idades. Esta coluna é uma variável Numérica Discreta, logo, teremos Histograma, Polígono de frequência com histograma e Boxplot.

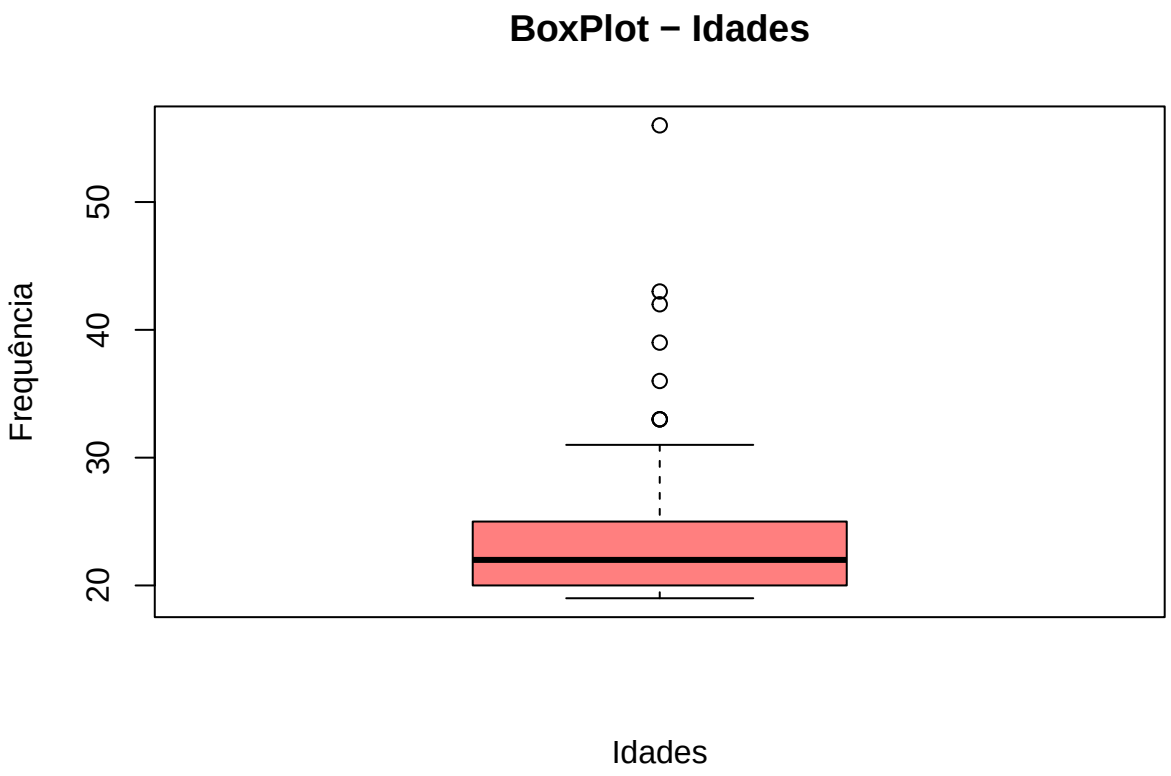
```
hist(df$IDADE,
     main="Histograma - Idades",
     xlab="Idades",
     ylab = "Frequência",
     border="black",
     col=rgb(1,0,0,0.5))
```




```
h=hist(df$IDADE,
       main="Polígono de frequência com histograma",
       xlab="Idades",
       ylab="Frequência",
       border="black",
       col=rgb(1,0,0,0.5))
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0), type = "l")
```




```
boxplot(
  df$IDADE,
  main="BoxPlot - Idades",
  xlab="Idades",
  ylab = "Frequência",
  col=rgb(1,0,0,0.5))
```



Abaixo temos os resultados obtidos a partir da função summary que é capaz de resumir vários tipos de objetos em uma única função. Dentre esses objetos encontram-se o primeiro e o terceiro quartil, sendo que o segundo quartil é dado indiretamente através da mediana.

```
summary(df$IDADE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   19.00   20.00   22.00   23.34   25.00   56.00       2
```

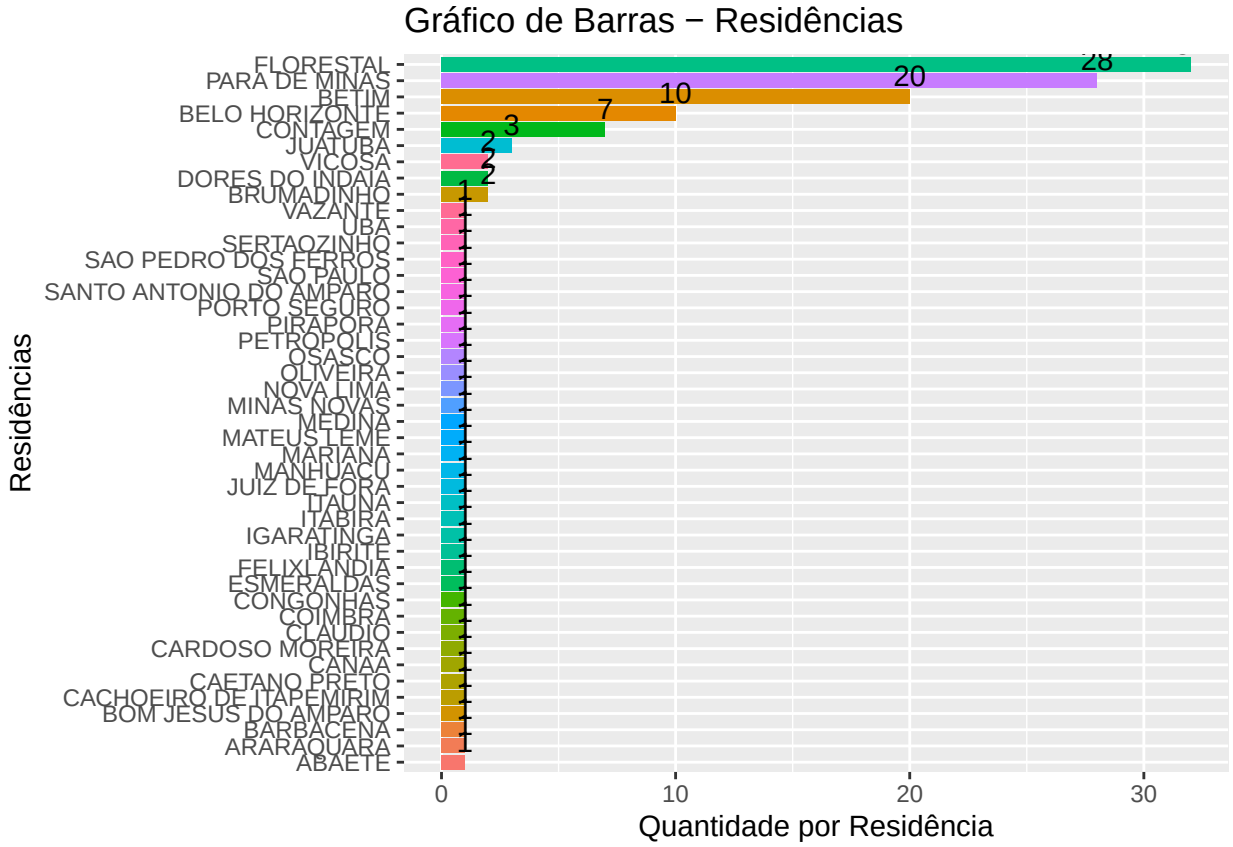
Podemos notar a partir das informações anteriores, tanto dos gráficos quanto da função *summary* que as idades são em média 23 anos. Podemos notar também a presença de outliers (os outliers são dados que se distanciam radicalmente de todos os outros São pontos fora da curva normal, valores que fogem da normalidade e que podem causar desequilíbrio nos resultados obtidos), ja que possuímos valores acima da soma Q3 + FIQ * 1,5, em que FIQ = Q3 - Q1. Além disso, podemos notar um valor alto para a amplitude (diferença entre valores máximo e mínimo) o que nos indica uma grande dispersão ou variabilidade dos dados para esta variável.

Residência

Nesta subseção será analisado os dados referentes as Residências. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

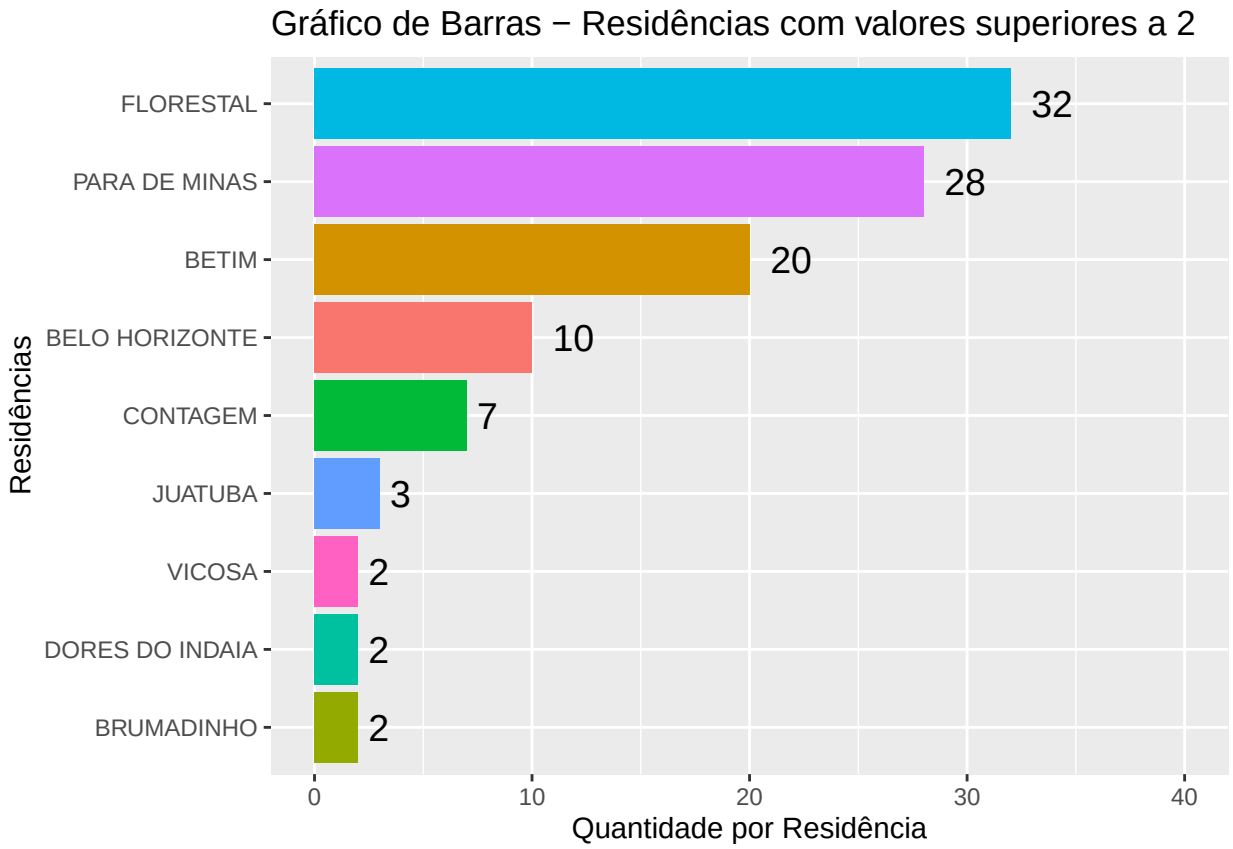
```
df$RESIDENCIA <- stringr::str_trim(df$RESIDENCIA)

residencia <- df %>%
  group_by(RESIDENCIA) %>%
  summarise(Total=n())
```

```
residencia %>% ggplot(aes(reorder(RESIDENCIA, Total), Total, fill=RESIDENCIA)) +
  geom_col(show.legend = FALSE) +
   theme(legend.position = 'top',
         legend.spacing.x = unit(10.0, 'cm'),
         legend.text = element_text(margin = margin(t = 10))) +
  theme_gray()+
  xlab("Residências")+
  ylab("Quantidade por Residência")+
  ggtitle("Gráfico de Barras - Residências")+
  geom_text(aes(label=Total), vjust=-0.5)+
  coord_flip()
```




```
###################################
residencia <- df %>%
  group_by(RESIDENCIA) %>%
  summarise(Total=n())

residencia <- residencia %>% filter(Total>=2)

residencia %>% ggplot(aes(reorder(RESIDENCIA, Total), Total, fill=RESIDENCIA)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), hjust=-0.5, size=5)+
    theme_gray()+
  xlab("Residências")+
  ylab("Quantidade por Residência")+
  ggtitle("Gráfico de Barras - Residências com valores superiores a 2")+
  coord_flip()+
```

```
  ylim(0,40)
```



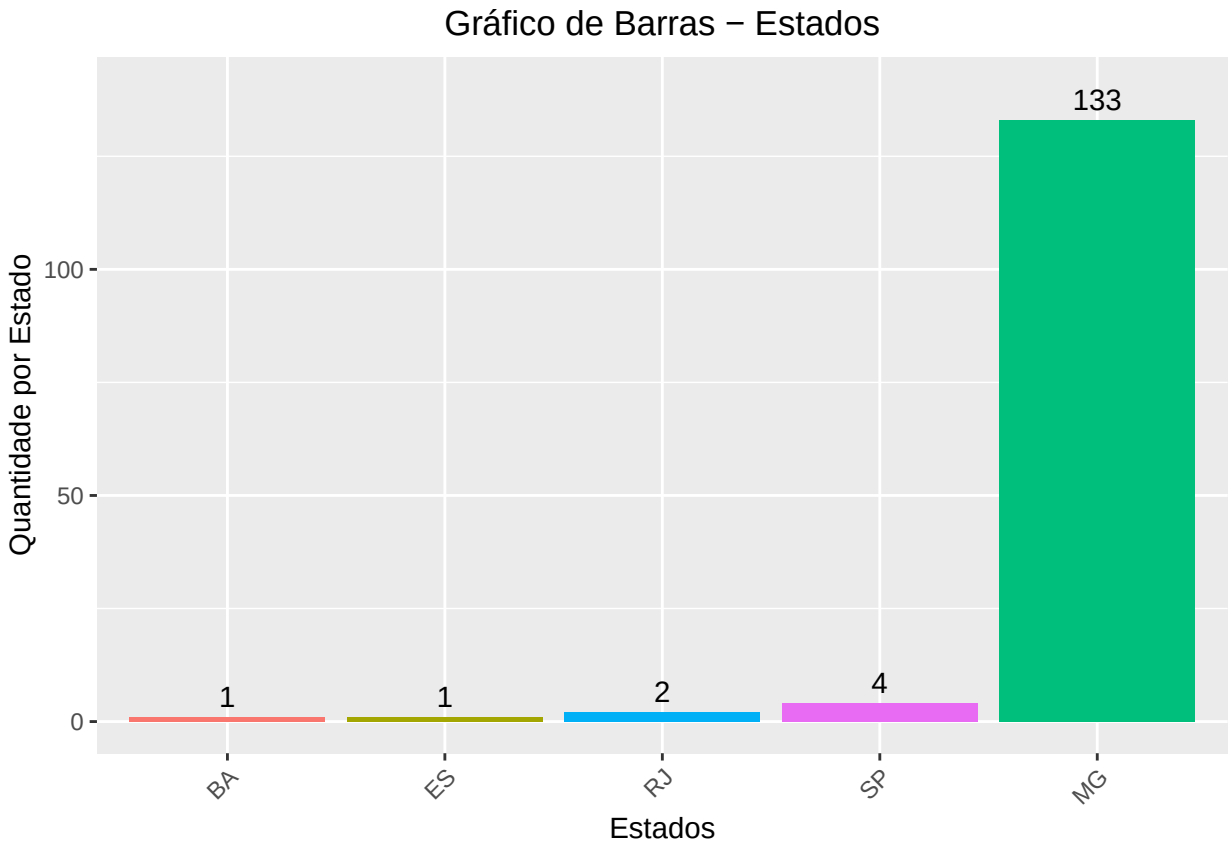
Nesta seção foram feitos gráficos de barras, sendo que o primeiro contém todas as residências da coluna e o segundo residências com valores superiores a 2. Como nesta variável há uma quantidade expressiva de dados distintas, percebe-se que o primeiro gráfico torna-se de difícil leitura, já o segundo como foi feita a filtragem, é mais agradável de ser visto. Notamos também, que as cidades de Florestal e Pará de Minas possuem os maiores valores.

Estado

Nesta subseção será analisado os dados referentes aos Estados. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

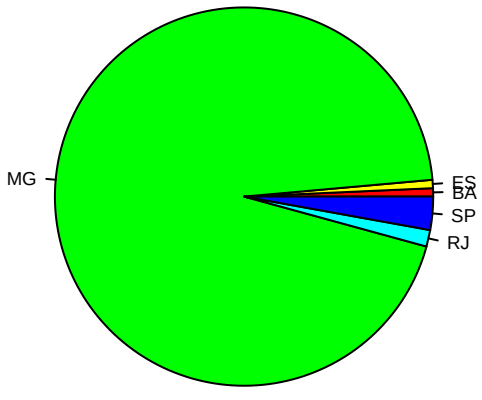
```
estado <- df %>%
  group_by(ESTADO) %>%
  summarise(Total=n())

estado %>% ggplot(aes(reorder(ESTADO,Total), Total, fill=ESTADO)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), vjust=-0.5)+
  ylim(0,140)+
  theme_gray()+
  xlab("Estados")+
  ylab("Quantidade por Estado")+
  ggtitle("Gráfico de Barras - Estados")+
```

```
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
tabela <- table(df$ESTADO)
pie(tabela,col=rainbow(6),cex=0.6, main="Gráfico de setores - Estados")
```

Gráfico de setores − Estados



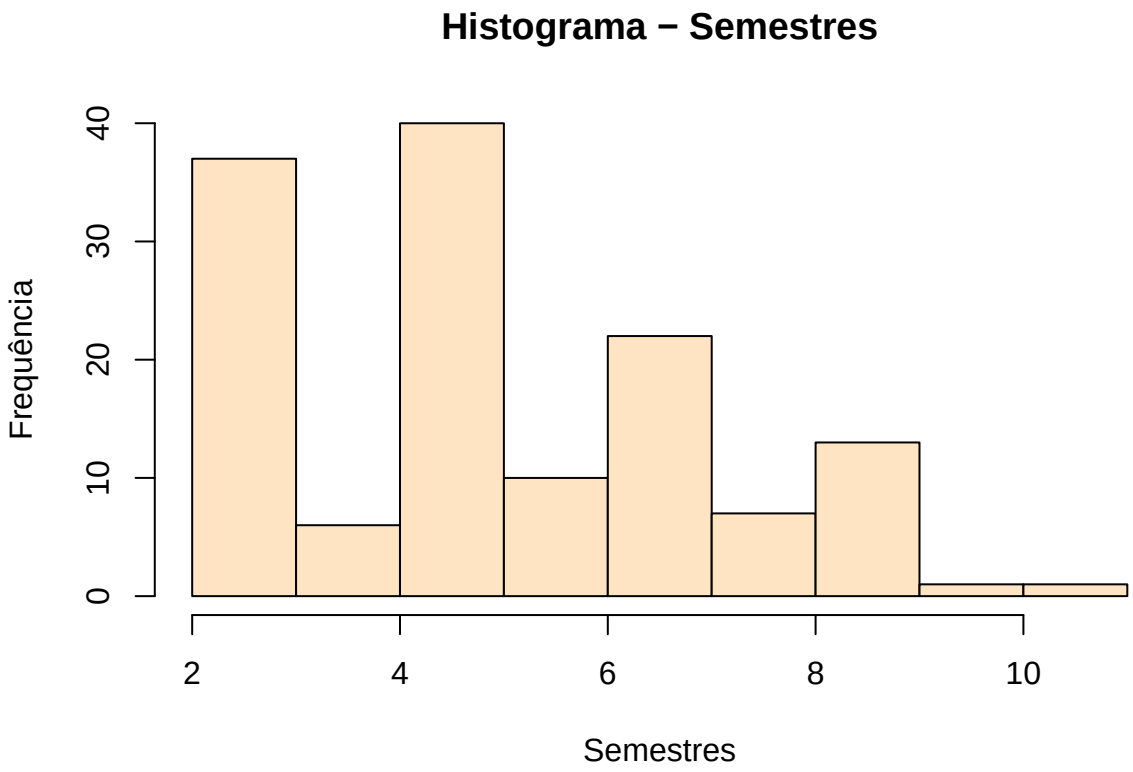
Notamos a partir dos gráficos que Minas Gerais é o estado mais expressivo. Vale ressaltar que assim como o gráfico de setores para a variável de Provedores, neste caso também é de grande auxílio quanto a interpretação,

já que visualmente conseguimos mensurar os dados como parcelas.

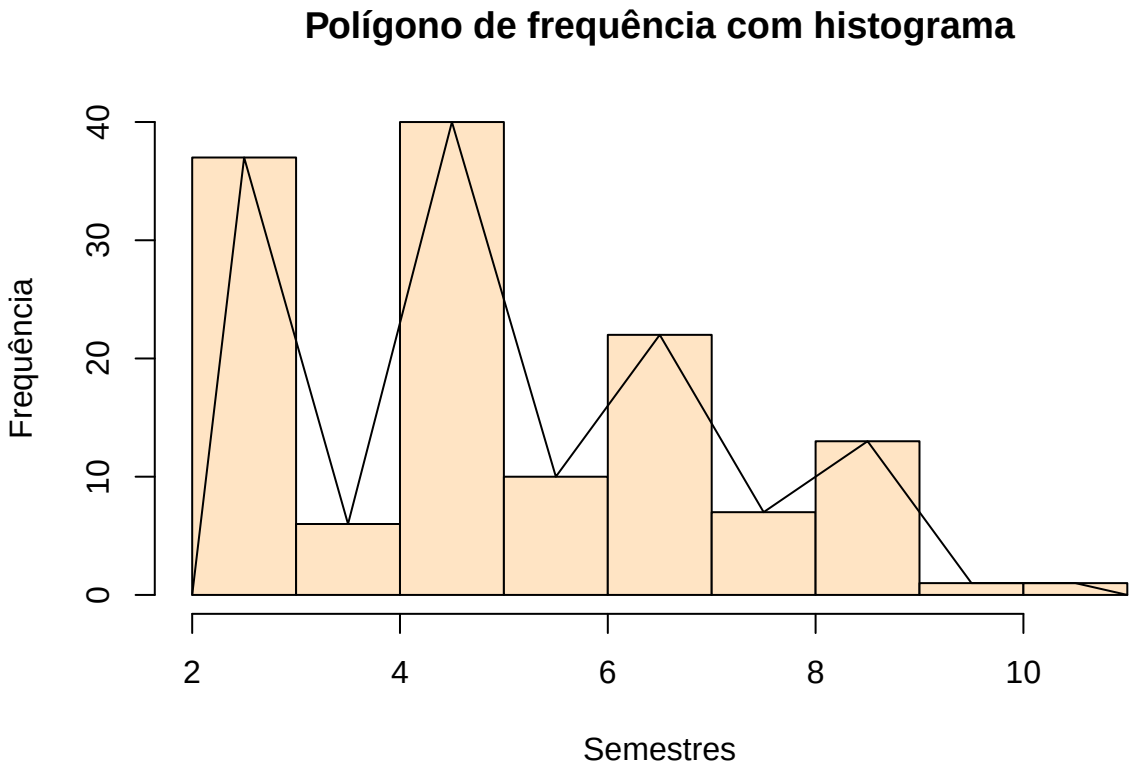
Semestre

Nesta subseção será analisado os dados referentes aos Semestres. Esta coluna é uma variável Numérica Discreta, logo, teremos Histograma, Polígono de frequência com histograma e Boxplot.

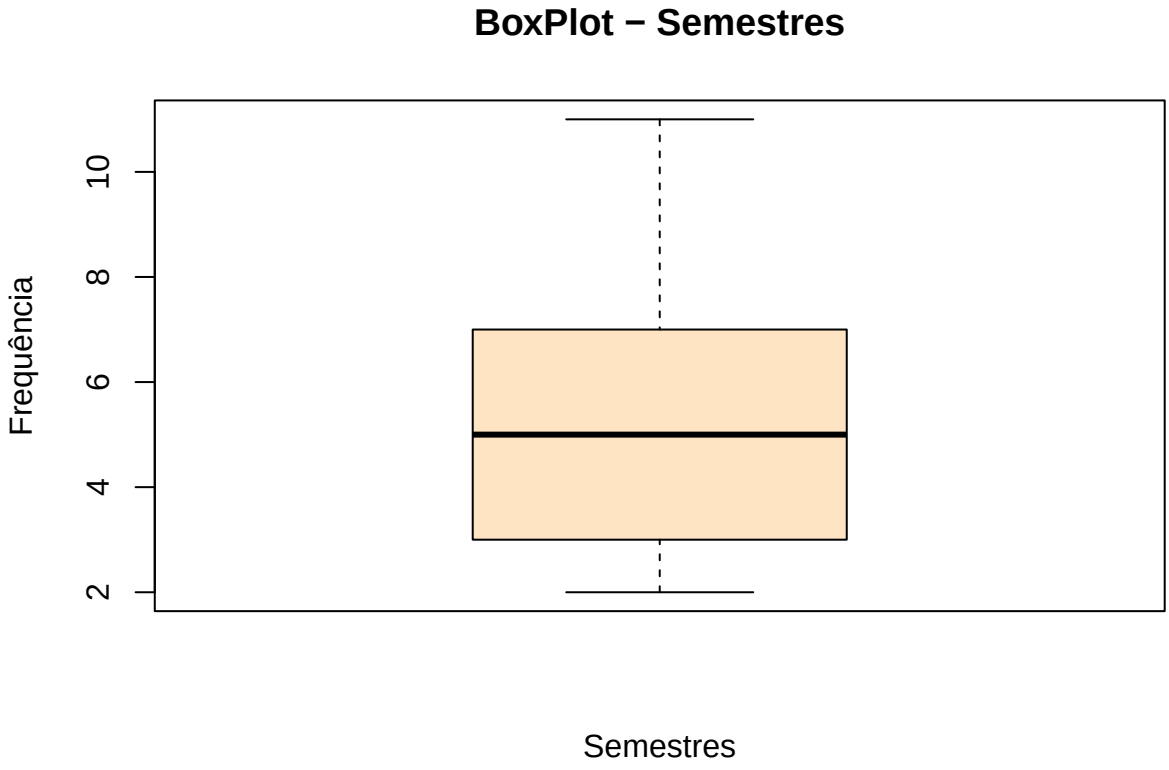
```
hist(df$SEMESTRE,
     main="Histograma - Semestres",
     xlab="Semestres",
     ylab = "Frequência",
     border="black",
     col="bisque")
```




```
h=hist(df$SEMESTRE,
       main="Polígono de frequência com histograma",
       xlab="Semestres",
       ylab="Frequência",
       border="black",
       col="bisque")
lines(c(min(h$breaks), h$mids, max(h$breaks)), c(0,h$counts, 0), type = "l")
```




```
boxplot(
  df$SEMESTRE,
  main="BoxPlot - Semestres",
  xlab="Semestres",
  ylab = "Frequência",
  col="bisque")
```



Abaixo temos os resultados obtidos a partir da função summary que como explicado anteriormente, encontram-se o primeiro e o terceiro quartil, sendo que o segundo quartil é dado indiretamente através da mediana.

```
summary(df$SEMESTRE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.000   3.000   5.000   5.401   7.000  11.000       4
```

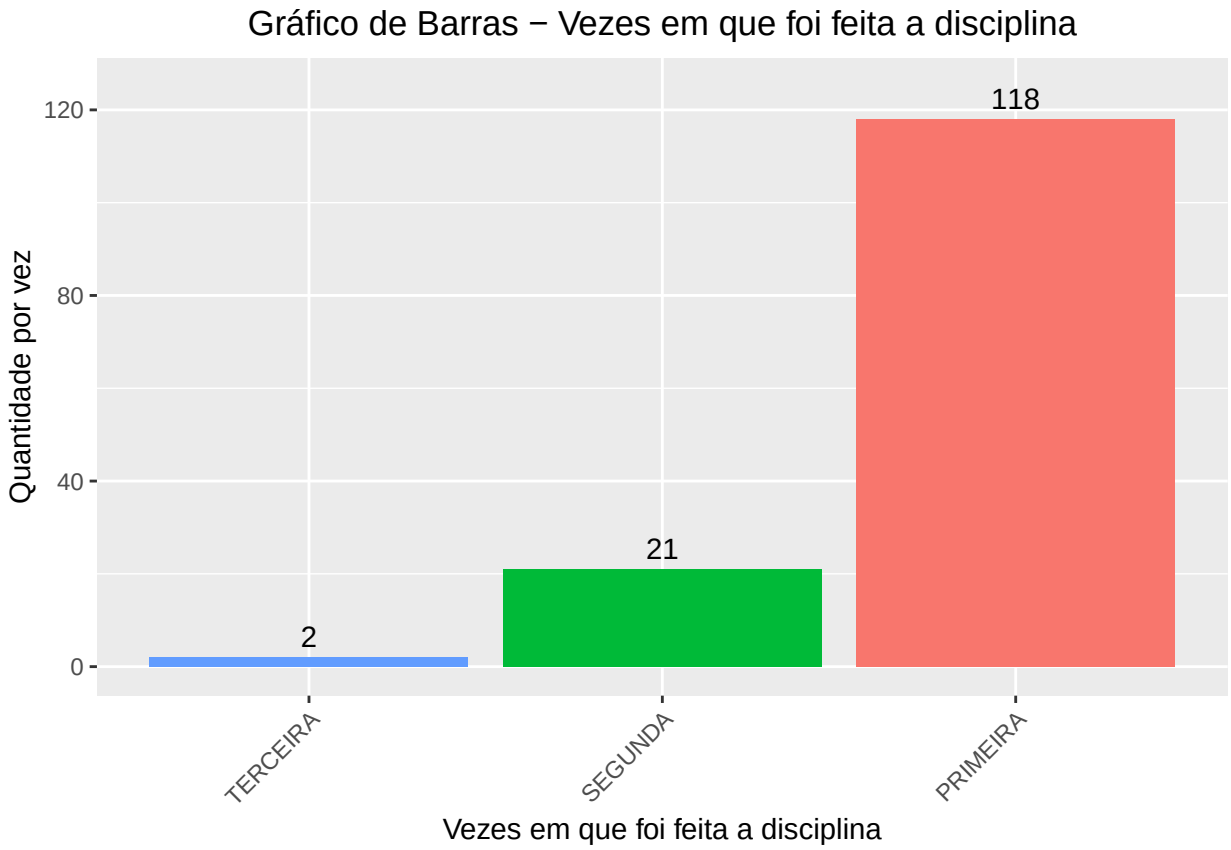
Podemos notar a partir das informações anteriores, tanto dos gráficos quanto da função *summary* que as pessoas estão em média no 5º semestre. Podemos notar também que ao contrário da variável Idades não há presença de outliers, mostrando então melhor precisão.

Primeira vez na matéria

Nesta subseção será analisado os dados referentes as vezes em que a matéria foi cursada. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

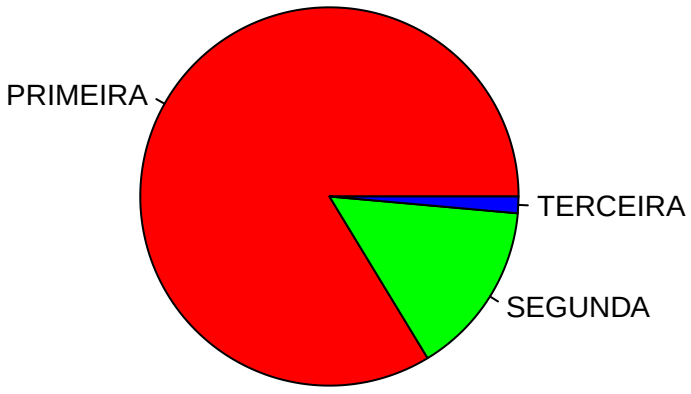
```
primeira_vez <- df %>%
  group_by(PRIMEIRAVEZ) %>%
  summarise(Total=n())

primeira_vez %>% ggplot(aes(reorder(PRIMEIRAVEZ,Total), Total, fill=PRIMEIRAVEZ)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), vjust=-0.5)+
  ylim(0,125)+
  theme_gray()+
  xlab("Vezes em que foi feita a disciplina")+
  ylab("Quantidade por vez")+
```

```
  ggtitle("Gráfico de Barras - Vezes em que foi feita a disciplina")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
tabela <- table(df$PRIMEIRAVEZ)
pie(tabela,col=rainbow(3),cex=0.9, main="Gráfico de setores - Vezes em que foi feita a disciplina")
```

Gráfico de setores − Vezes em que foi feita a disciplina



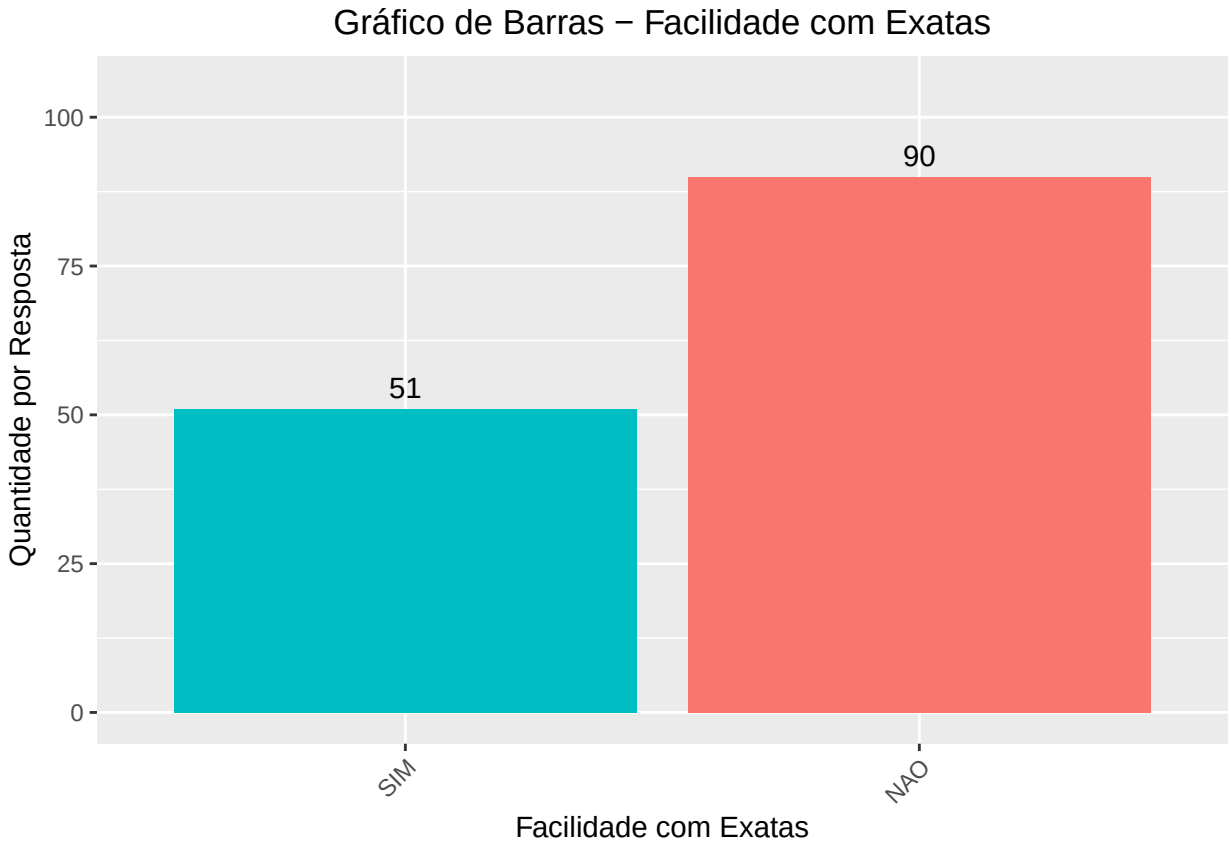
Notamos a partir dos gráficos que a quantidade mais expressiva é de alunos que estão fazendo a disciplina

pela primeira vez. Vale ressaltar que assim como o gráfico de setores para a variável de Provedores e Estados, neste caso também é de grande auxílio quanto a interpretação.

Facilidade em Exatas

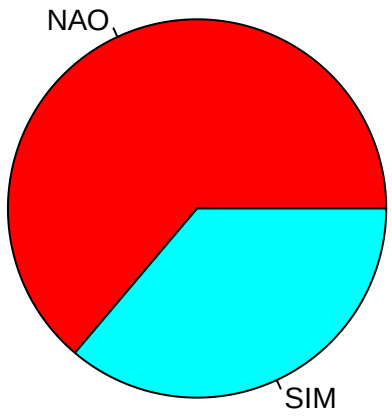
Nesta subseção será analisado os dados referentes a Facilidade em Exatadas. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

```
facilidade_exatas <- df %>%
  group_by(FACILIDADEEXATAS) %>%
  summarise(Total=n())

facilidade_exatas %>% ggplot(aes(reorder(FACILIDADEEXATAS,Total), Total, fill=FACILIDADEEXATAS)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), vjust=-0.5)+
  ylim(0,105)+
  theme_gray()+
  xlab("Facilidade com Exatas")+
  ylab("Quantidade por Resposta")+
  ggtitle("Gráfico de Barras - Facilidade com Exatas")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
tabela <- table(df$FACILIDADEEXATAS)
pie(tabela,col=rainbow(2),cex=0.9, main="Gráfico de setores - Facilidade com Exatas")
```

Gráfico de setores − Facilidade com Exatas

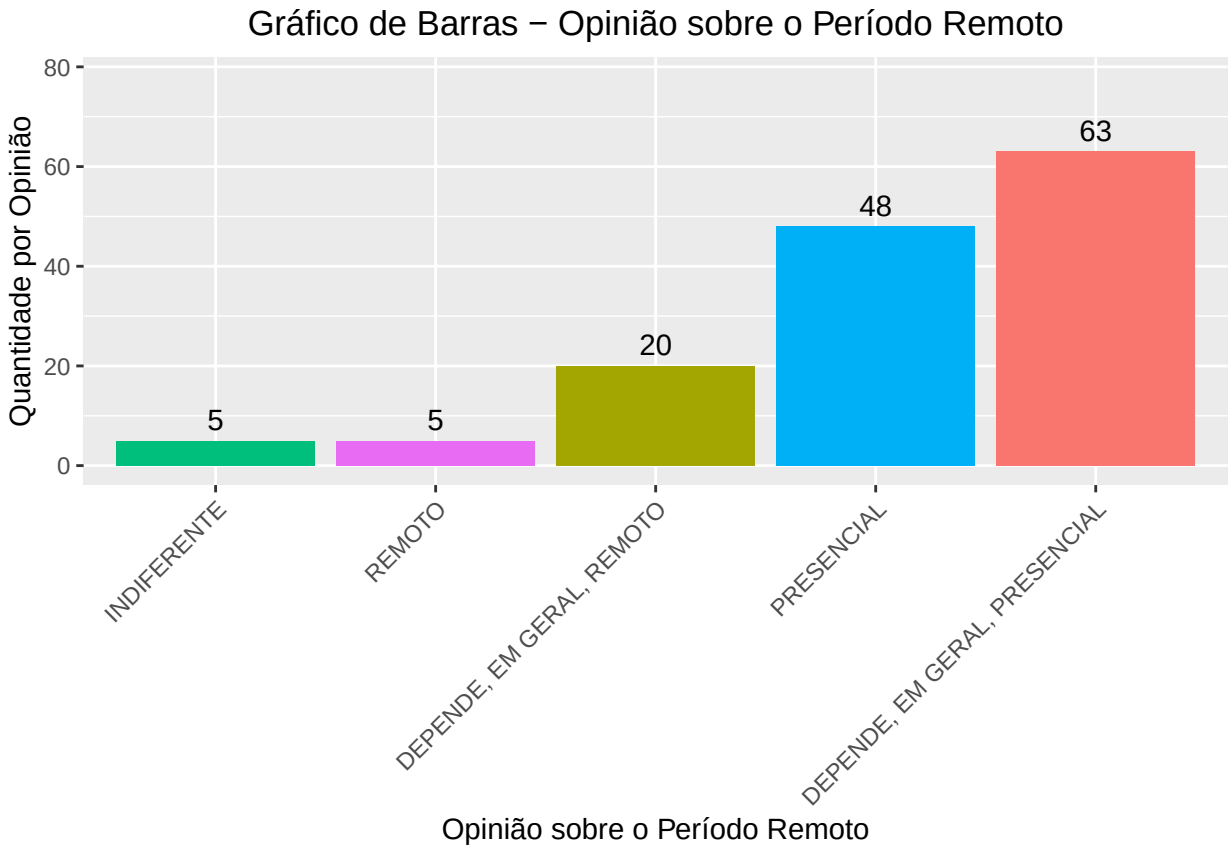


Notamos a partir dos gráficos que a quantidade mais expressiva é de alunos que não possuem facilidade em exatas, sendo que estes valores chegam a ser mais de 60% do valor total.

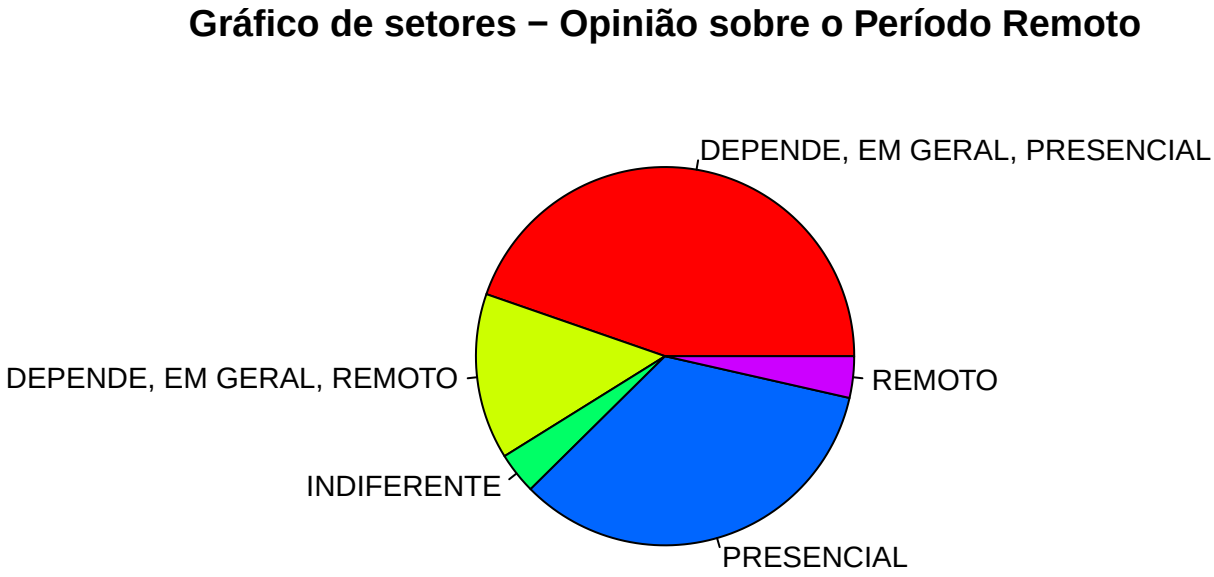
Opinião sobre os períodos remotos

Nesta subseção será analisado os dados referentes as opiniões sobre os períodos remotos. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

```
opiniao_periodos_remotos <- df %>%
  group_by(OPINIAOPERIODOSREMOTOS) %>%
  summarise(Total=n())

opiniao_periodos_remotos %>% ggplot(
    aes(reorder(OPINIAOPERIODOSREMOTOS,Total),
        Total,
        fill=OPINIAOPERIODOSREMOTOS)
    ) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), vjust=-0.5)+
  ylim(0,78)+
  theme_gray()+
  xlab("Opinião sobre o Período Remoto")+
  ylab("Quantidade por Opinião")+
  ggtitle("Gráfico de Barras - Opinião sobre o Período Remoto")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
tabela <- table(df$OPINIAOPERIODOSREMOTOS)
pie(tabela,col=rainbow(5),cex=0.9, main="Gráfico de setores - Opinião sobre o Período Remoto")
```

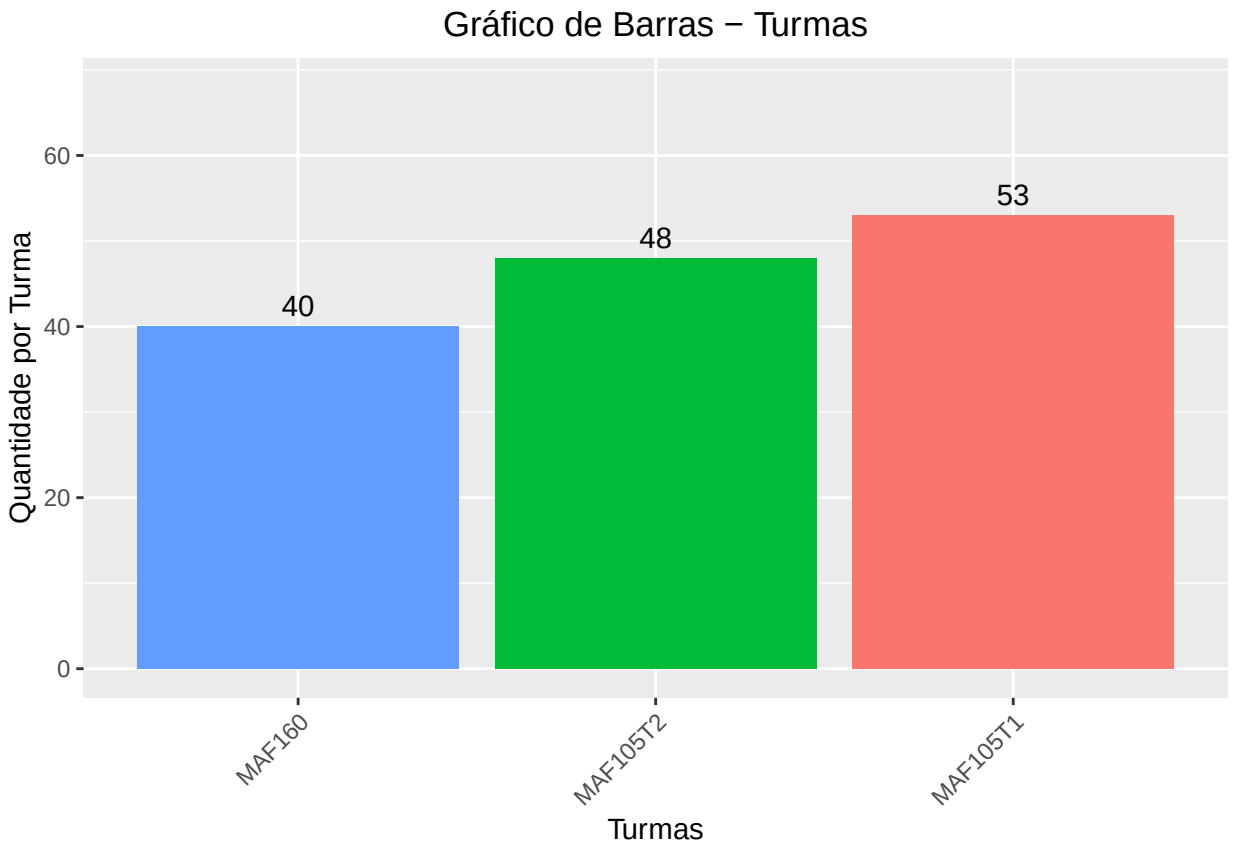


Podemos interpretar a partir das informações anteriores, que os alunos de modo geral preferem as aulas presenciais, enquanto a quantidade de alunis que preferem o ensino remoto é a menor.

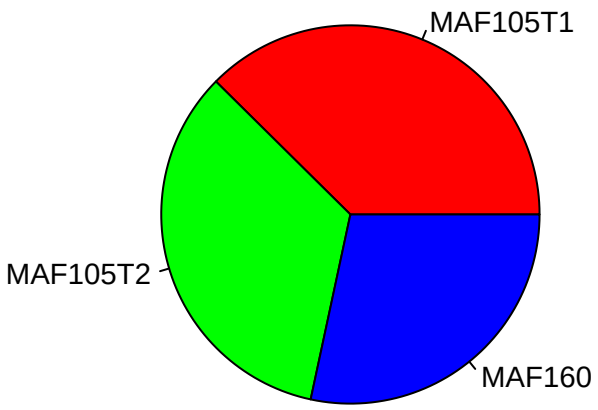
Turma

Nesta subseção será analisado os dados referentes as Turmas. Esta coluna é uma variável Categórica nominal, logo, teremos gráficos de barra e gráficos de setores.

```
turma <- df %>%
  group_by(TURMA) %>%
  summarise(Total=n())

turma %>% ggplot(aes(reorder(TURMA,Total), Total, fill=TURMA)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label=Total), vjust=-0.5)+
  ylim(0,68)+
  theme_gray()+
  xlab("Turmas")+
  ylab("Quantidade por Turma")+
  ggtitle("Gráfico de Barras - Turmas")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```




```
tabela <- table(df$TURMA)
pie(tabela,col=rainbow(3),cex=0.9, main="Gráfico de setores - Turmas")
```



Podemos interpretar a partir das informações anteriores, que os valores de cada turma são bem próximos e em contraste a disciplina de MAF105 tanto da turma 1 quanto a 2 a disciplina de MAF160 é a de menor expressão.

Conclusão

Conclui-se que este relátorio apresentou resultados muito interessantes, por relacionar os tipos de variáveis e os mais variados tipos de gráficos. Por conseguinte podemos ressaltar que a linguagem R nos permite criar gráficos interativos a partir dos resultados das análises de dados. Os gráficos podem ser usados para obter insights significativos durante todo o processo de análise de dados ou podem ser exportados em um relatório como neste caso.

A análise de dados permite, ao mesmo tempo, ter um panorama geral do dataframe e riqueza de detalhes em cada variável. Ademais, a junção de análise e uso de gráficos, corrobora para melhor entendimento da disciplina de Estatística, sendo assim, conseguimos ter um melhor aprendizado da matéria.

Referências

O que é a Estatística

Importância da Estatística

O que é um Dataframe

Variáveis e Classificação

Gráficos

Quando usar determinado Gráfico

Boxplot