Atividade Avaliativa

Estatística descritiva para Data Science

Ana Maria Alves da Silva

2024-08-14

Manipulção Preliminar e Tratamento de Dados

Observação: Antes de realizar qualquer manipulação nos dados para responder o item solicitado é necessário carregar os dados e realizar a limpeza desses dados que serão utilizados ao longo dessa atividade. Como eles estão em formato csv, iremos utilizar a função *read_csv*.

1. Carregando os dados:

```
setwd <- "/Users/anamaria/especializacao/modulo_4/atividade"
df <- read.csv("SG_UFGO_16_07_24.csv", sep = ";")
dim(df)
```

```
## [1] 255647     64
```

2. Verificando se há dados duplicados:

```
duplicados <- duplicated(df)
ha_duplicados <- any(duplicados)
print(ha_duplicados)
```

```
## [1] FALSE
```

3. Note que o conjunto de dados contém 255.647 observações e 64 variáveis e não há dados duplicados. No entanto, após análise dos itens solicitados e de acordo com do dicionário de dados verificamos que precisamos apenas das colunas:

- profissionalSaude
- racaCor
- codigoRecebeuVacina
- sexo
- classificacaoFinal
- idade
- dataNotificação
- sintomas
- evolucaoCaso

Além disso, precisaremos apenas dos dados correspondentes ao ano de 2024 pois ao análisar o conjunto de dados previamente verificamos que há datas de notificação que não correspondem ao ano de 2024.

```r
# Filtragem dos dados para apenas considerar o ano de 2024
library(lubridate)
library(dplyr)
df <- df %>% mutate(dataNotificacao = ymd(dataNotificacao))
df_2024 <- df %>% filter(year(dataNotificacao) == 2024)
dim(df_2024)
```

```
## [1] 172713     64
```

```r
# Filtragem para obtermos apenas as colunas seleciondas com os dados de 2024
var_selecionadas <- c('profissionalSaude', 'racaCor', 'codigoRecebeuVacina',
                      'sexo', 'classificacaoFinal', 'idade',
                      'dataNotificacao', 'sintomas', 'evolucaoCaso')


df_2 <- df_2024[, var_selecionadas]
dim(df_2)
```

```
## [1] 172713      9
```

Dessa forma, nosso conjunto de dados ficou com 172713 observações e 9 colunas.

4. Visualizando um resumo dos dados:

```r
summary_df <- summary(df_2)
print(summary_df)
```

```
##  profissionalSaude    racaCor          codigoRecebeuVacina     sexo
##  Length:172713      Length:172713      Min.   :1.000       Length:172713
##  Class :character   Class :character   1st Qu.:1.000       Class :character
##  Mode  :character   Mode  :character   Median :1.000       Mode  :character
##                                        Mean   :1.136
##                                        3rd Qu.:1.000
##                                        Max.   :3.000
##                                        NA's   :17423
##  classificacaoFinal     idade       dataNotificacao        sintomas
##  Length:172713      Min.   : 0.00   Min.   :2024-01-01   Length:172713
##  Class :character   1st Qu.:18.00   1st Qu.:2024-02-09   Class :character
##  Mode  :character   Median :30.00   Median :2024-02-28   Mode  :character
##                     Mean   :28.84   Mean   :2024-03-14
##                     3rd Qu.:42.00   3rd Qu.:2024-04-12
##                     Max.   :54.00   Max.   :2024-07-14
##                     NA's   :47790
##  evolucaoCaso
##  Length:172713
##  Class :character
##  Mode  :character
##
##
##
##
```

Note que através do resumo dos dados podemos ver que as colunas codigoRecebeuVacina e idade são númericas e ambas possuem valores ausentes que serão tratados posteriormente. Além disso podemos validar que na coluna dataNotificacao temos apenas dados referente ao ano de 2024 pois a mesma possui valor mínimo de 2024-01-01 e valor máximo de 2024-07-14.

5. Criando os fatores de algumas da variáveis (colunas) categóricas e adicionando os respectivos labels. Para isso, usaremos a função *factor*.

- Profissional da Saúde

```
table(df_2$profissionalSaude)
```

```
##
##    Não    Sim
## 170911   1802
```

```
df_2$profissionalSaude <- factor(df_2$profissionalSaude, levels = c("Não", "Sim"),
                                 labels = c("Não", "Sim"),
                                ordered = FALSE)
table(df_2$profissionalSaude)
```

```
##
##    Não    Sim
## 170911   1802
```

- Recebeu Vacina

```
table(df_2$codigoRecebeuVacina)
```

```
##
##      1      2      3
## 134246  21003     41
```

```
df_2$codigoRecebeuVacina <- factor(df_2$codigoRecebeuVacina, levels = c(1, 2, 3),
                                   labels = c("Sim", "Não", "Ignorado"),
                                   ordered = FALSE)
table(df_2$codigoRecebeuVacina)
```

```
##
##      Sim      Não Ignorado
##   134246    21003       41
```

- Classificação Final

```
table(df_2$classificacaoFinal)
```

```
##
##                                  Confirmado Clínico-Epidemiológico
##                            63361                               853
```

```
##         Confirmado Clínico-Imagem           Confirmado Laboratorial
##                                24                             51550
##   Confirmado por Critério Clínico                         Descartado
##                              1033                             15977
##  Síndrome Gripal Não Especificada
##                             39915
```

```
df_2$classificacaoFinal <- factor(df_2$classificacaoFinal,
    levels = c("Confirmado Laboratorial", "Síndrome Gripal Não Especificada",
               "Confirmado Clínico-Epidemiológico", "Confirmado por Critério Clínico",
               "Confirmado Clínico-Imagem", "Descartado"),
    labels = c("Confirmado Laboratorial", "Síndrome Gripal Não Especificada",
               "Confirmado Clínico-Epidemiológico", "Confirmado por Critério Clínico",
               "Confirmado Clínico-Imagem", "Descartado"),
    ordered = FALSE)
table(df_2$classificacaoFinal)
```

```
##
##           Confirmado Laboratorial  Síndrome Gripal Não Especificada
##                             51550                             39915
## Confirmado Clínico-Epidemiológico   Confirmado por Critério Clínico
##                               853                              1033
##         Confirmado Clínico-Imagem                        Descartado
##                                24                             15977
```

- Evolução do Caso

```
table(df_2$evolucaoCaso)
```

```
##
##                                    Cancelado                       Cura
##                     83493                4419                      69340
## Em tratamento domiciliar              Ignorado                  Internado
##                      3028                12299                         50
##         Internado em UTI                 Óbito
##                        7                    77
```

```
df_2$evolucaoCaso <- factor(df_2$evolucaoCaso,
    levels = c("Cancelado", "Ignorado", "Em tratamento domiciliar",
               "Internado em UTI", "Internado", "Óbito", "Cura"),
    labels = c("Cancelado", "Ignorado", "Em tratamento domiciliar",
               "Internado em UTI", "Internado", "Óbito", "Cura"),,
    ordered = FALSE)
table(df_2$evolucaoCaso)
```

```
##
##               Cancelado               Ignorado Em tratamento domiciliar
##                    4419                  12299                      3028
##         Internado em UTI              Internado                     Óbito
##                       7                     50                        77
##                    Cura
##                   69340
```

Itens - Parte I

Item 1 - Faça uma tabela de frequências e responda à seguinte pergunta: Qual a porcentagem dos profissionais da saúde com notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás? Utilize duas casas decimais.

Solução: Para responder essa pergunta, será necessário analisar a quantidade de profissionais da saúde com suspeita de covid em relação a quantidade total de pessoas com suspeita de covid, observe que na construção do nosso dataframe já relizamos a filtragem referente ao dados do ano de 2024 no Estado de Goiás. Para isso, será necessário criarmos tabela de frequência absoluta no R, cuja função utilizada é a *table* e a tabela de frequência relativa no R, a função utilizada é a *prop.table* que deve ser aplicada à uma tabela de frequência absoluta.

```
# Verificando se há valores ausentes:
tamanho <- length(df_2$profissionalSaude)
soma <- sum(table(df_2$profissionalSaude))
print(tamanho - soma)
```

```
## [1] 0
```

```
#Frequencia Absoluta:
freq_saude <- round(table(df_2$profissionalSaude),4)
print(freq_saude)
```

```
##
##    Não    Sim
## 170911   1802
```

```
#Frequencia Relativa:
freq_saude_relativa <- round(prop.table(table(df_2$profissionalSaude)),4)
print(freq_saude_relativa)
```

```
##
##    Não    Sim
## 0.9896 0.0104
```

Logo, a porcentagem dos profissionais da saúde com notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás é de 1,04%.

Ps. Optei por fazer o arredondamento com 4 casas decimais para que a porcentagem tivesse 2 casas decimais. Caso usassemos o arredondamento com 2 casa, a porcentagem equivalente seria 1% referente a sim para Recebeu Vacina e 99% para não na variável Recebeu Vacina.

Item 2 - Faça uma tabela de frequências e responda à seguinte pergunta: Qual é o sexo que apresenta o maior número de notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás?

Solução:

A solução é similar ao que foi feito no item anterior, no entanto será necessário calcular apenas a tabela de frequência absoluta.

```r
# Verificando se há valores ausentes:
tamanho_s <- length(df_2$sexo)
soma_s <- sum(table(df_2$sexo))
print(tamanho_s - soma_s)
```

```
## [1] 0
```

```r
#Frequencia Absoluta:
freq_sexo <- round(table(df_2$sexo),2)
print(freq_sexo)
```

```
##
##  Feminino Masculino
##    101776     70937
```

O Sexo com o maior número de notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás é o sexo Feminino, com 101776 suspeitas.

Item 3 - Faça uma tabela de frequências e responda à seguinte pergunta: Qual a raça que apresenta a maior proporção de notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás? Utilize quatro casas decimais.

Solução:

Similarmente ao que foi feito no item 1, será necessário calcular a tabela de frequencia absoluta e relativa referente a coluna racaCor em nosso dataframe mas será necessário realizar um ajuste na função *round* para termos 4 casas decimais.

```r
# Verificando se há valores ausentes:
tamanho_s <- length(df_2$racaCor)
soma_s <- sum(table(df_2$racaCor))
print(tamanho_s - soma_s)
```

```
## [1] 0
```

```r
#Frequencia Absoluta:
freq_racaCor <- round(table(df_2$racaCor),4)
print(freq_racaCor)
```

```
##
##  Amarela   Branca Ignorado Indigena    Parda    Preta
##    21489    44375    24220       27    76813     5789
```

```r
#Frequencia Relativa:
freq_racaCor_relativa <- round(prop.table(table(df_2$racaCor)),4)
print(freq_racaCor_relativa)
```

```
##
##  Amarela   Branca Ignorado Indigena    Parda    Preta
##   0.1244   0.2569   0.1402   0.0002   0.4447   0.0335
```

Portanto, a Raça que apresenta a maior proporção de notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás é a Parda, com 44.47% das suspeitas.

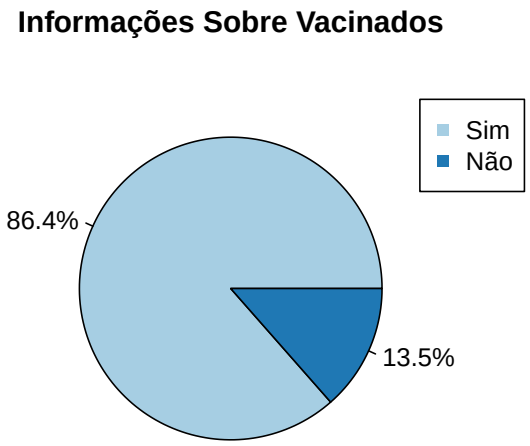
Item 4 - Faça um gráfico em setores para a variável "recebeu vacina" (codigoRecebeuVacina) das notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás. Utilize a frequência relativa em porcentagem. Além disso, responda: Qual é a categoria mais frequente nas notificações? Para o gráfico, adicione os nomes das categorias e as porcentagens como rótulos. Utilize uma casa decimal.

Solução:

Observe que no início desta atividade realizamos algumas limpezas e análises inicias nos dados. Em uma dessas etapas realizamos a categorização das variáveis categoricas e percebemos que a variável "recebeu vacina" não possui informações referente a categoria Ignorado, portanto, para realizar a construção do gráfico de setores, consideraremos apenas os casos no qual os dados são diferentes de ignorados.

Para realizar a construção do gráfico solicitado, usaremos a função *table* e *round*, já calculando a porcentagem referente a cada categoria e usaremos a função *pie* para construir o gráfico solicitado.

Para a esssa questão, optei em não exibir o código mas o mesmo se encontra no arquivo markdown que gerou este pdf.

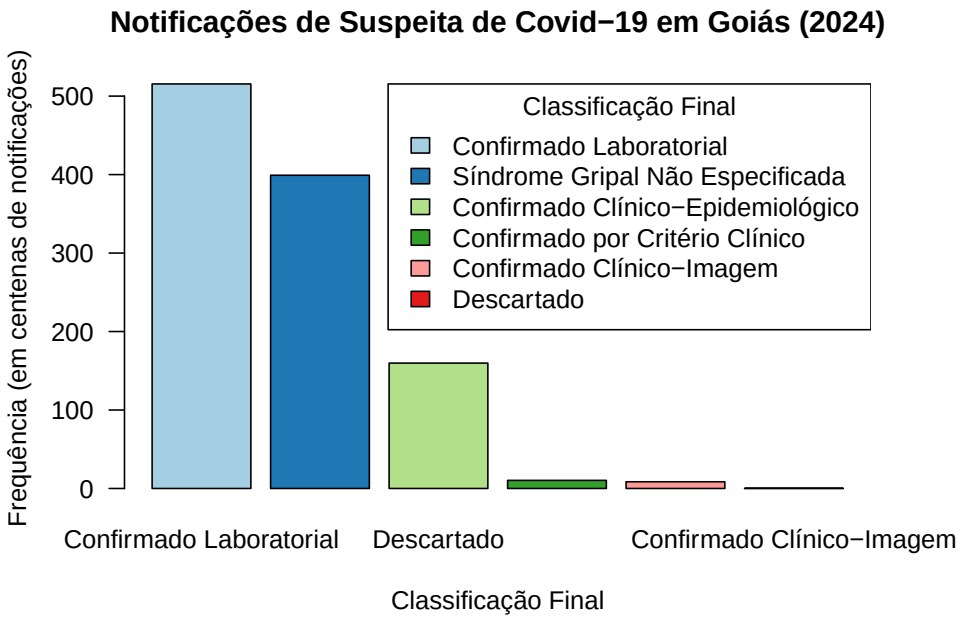


Portanto, a categoria sim - que significa que a pessoa recebeu vacina - é a mais frequente, com 86,4% das observações, nos dados de suspeita de Covid-19 no ano de 2024 em Goiás.

Item 5 - Faça um gráfico em barras para a variável "classificação final" (classificacaoFinal) das notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás. Utilize a frequência absoluta (em centenas de notificações). Além disso, analisando o gráfico, responda: Qual é a classificação mais frequente?

Solução:

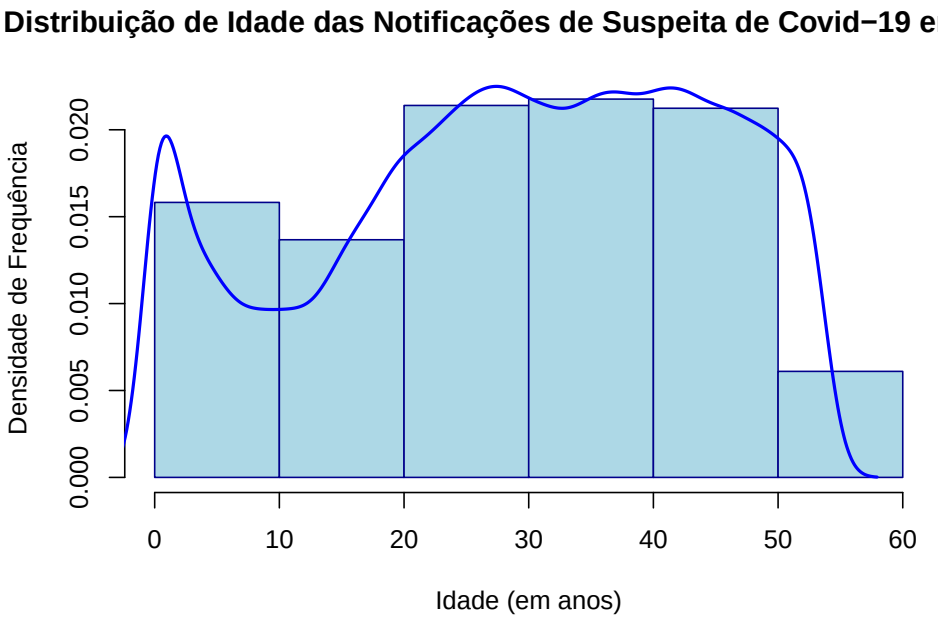
Usaremos a função *barplot* para criar esse gráfico. Além disso, também usaremos a função *brewer* para mudar as cores do gráfico em questão, a função *sort* para ordenar em ordem decrescente. Optaremos por omitir o código que gerou a imagem.



Portanto, a categoria Confirmado Laboratorial é a mais frequente nos dados de suspeita de Covid-19 no ano de 2024 em Goiás.

Item 6 - Faça um histograma para a idade (em anos) das pessoas que tiveram notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás. Utilize a densidade de frequência e classes de amplitude 10 (com a primeira iniciando na idade 0). Além disso, analisando o gráfico, responda: Qual é a faixa etária mais frequente?

Solução: Usaremos a função *hist* para criar o histograma solicitado, *seq* para definir os limites das classes e *lines* para inserir uma linha de desnsidade.




```
## A idade mais frequente é 35 anos.
```

Itens - Parte II

Item 7 - Faça um sumário com as principais medidas resumo (média, mediana, mínimo, máximo, primeiro e terceiro quartil) da variavél idade. Além disso, calcule o desvio padrão e o coeficiente de variação para a variável idade.

Solução:

Inicialmente, podemos obter algumas das principais medidas solicitadas usando a funçao *summary*.

```
resumo_idade <- summary(df_2$idade)
print(resumo_idade)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   18.00   30.00   28.84   42.00   54.00   47790
```

Observe que há vários valores ausentes, optaremos por omiti-los usando a função *na.omit*.

```
idade_na <- na.omit(df_2$idade)
resumo_idade_2 <- summary(idade_na)
print(resumo_idade_2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   30.00   28.84   42.00   54.00
```

Agora, vamos calcular o desvio padrão e o coeficiente de variação. Para o desvio padrão, utilizaremos a função *sd*, mas poderiamos calcular o desvio padrão através da raiz quadrada da variancia. Optei por realizar um arredondamento de 2 casas decimais.

```r
desvio_idade <- round(sd(idade_na),2)
cat("Usando a função sd: ",desvio_idade, "\n")
```

```
## Usando a função sd:  15.26
```

```r
checking_desvio <- round(sqrt(var(idade_na)),2)
cat("Usando a raiz quadrada da variancia: ",checking_desvio)
```

```
## Usando a raiz quadrada da variancia:  15.26
```

Embora não haja uma função específica para o cálculo do coeficiente de variação no R sabemos que o coeficiente de variação é uma medida de dispersão relativa que é obtido dividindo o desvio padrão pela média e multiplicando por 100, para que o valor fique em porcentagem. Vamos fazer esse cálculo.

```r
cv_idade <- round((desvio_idade / mean(idade_na))*100, 2)

cat("O Coeficiente de Variação é: ", cv_idade,"%.")
```

```
## O Coeficiente de Variação é:  52.91 %.
```

Item 8 - Qual foi o sintoma mais frequente (moda da variável sintoma) registrado entre as notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás?

Solução: Para obtermos a moda, basta utilizarmos a função *mfv* do pacote modeest.

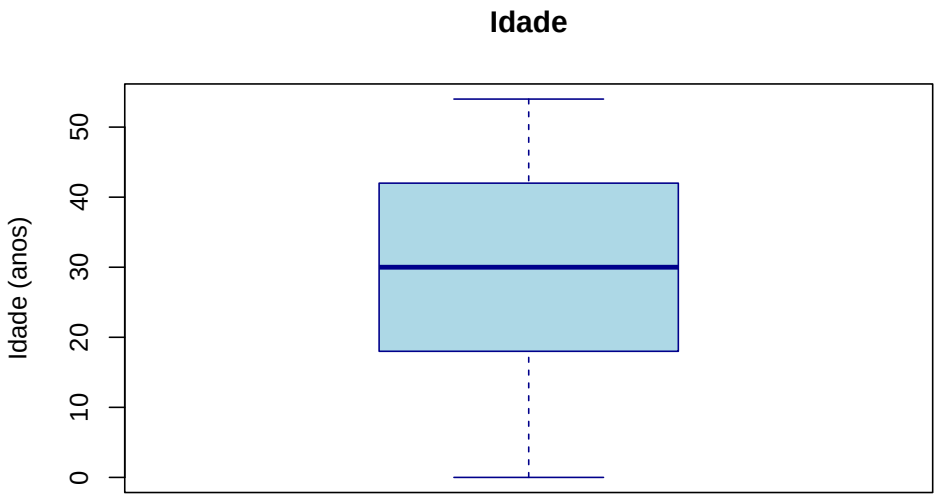
```r
library(modeest)
moda_sintoma <- mfv(df_2$sintomas, na_rm = TRUE)
cat("O sintoma mais frequente foi:", moda_sintoma)
```

```
## O sintoma mais frequente foi: Assintomático
```

Item 9 - Faça um boxplot da variável idade e responda: existem idades discrepantes (outliers) na amostra?

Solução: Utilizaremos a função *boxplot* e a variável idade desconsiderando os valores ausentes.

```r
boxplot(idade_na,
        main='Idade',
        ylab='Idade (anos)',
        col = "lightblue",
        border = "darkblue")
```



O boxplot acima mostra uma distribuição de dados centrada em 30 anos, com a maioria dos dados variando entre 20 e 40 anos. A distribuição parece simétrica pois a mediana está aproximadamente no centro do retângulo (30 anos), e não há sinais de outliers significativos. Observe que os dados estão de acordo com o resumo que fizemos no item 7 dessa atividade.

Item 10 - Faça uma tabela cruzada para as variáveis "evolução do caso" (evolucaoCaso) e "recebeu vacina" (codigoRecebeuVacina), utilizando frequências relativas pelo total da coluna (utilize 4 casas decimais para a proporção). Exclua valores ausentes (NA's) e Ignorados de ambas as variáveis. Responda: Do total de pessoas que tomaram vacina, qual a porcentagem de pessoas que foram internadas em UTI? E considerando o total de pessoas que não tomaram vacina, qual porcentagem foram internadas em UTI?

Solução:

Para fazer a tabela cruzada, usaremos a biblioteca gmodels e a função *CrossTable*. Antes, iremos filtrar os dados para remover os Na's e Ignorados. E como queremos o total de pessoas que tomaram vacina, utilizaremos prop.c = TRUE pois retorna a proporção em relação as colunas.

```
library(gmodels)
dados_filtrados <- subset(df_2,
                          !is.na(evolucaoCaso) & !is.na(codigoRecebeuVacina) &
                          evolucaoCaso != "Ignorado" & codigoRecebeuVacina != "Ignorado")

# Criar a tabela cruzada com frequências relativas
CrossTable(dados_filtrados$evolucaoCaso, dados_filtrados$codigoRecebeuVacina,
           prop.r = FALSE,
           prop.t = FALSE,
           prop.chisq = FALSE,
           prop.c = TRUE,
           digits = 4)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  66843
##
##
##                               | dados_filtrados$codigoRecebeuVacina
## dados_filtrados$evolucaoCaso |       Sim |       Não | Row Total |
## -----------------------------|-----------|-----------|-----------|
##                     Cancelado |      3343 |       662 |      4005 |
##                               |    0.0601 |    0.0591 |           |
## -----------------------------|-----------|-----------|-----------|
##     Em tratamento domiciliar |      2258 |       239 |      2497 |
##                               |    0.0406 |    0.0213 |           |
## -----------------------------|-----------|-----------|-----------|
##             Internado em UTI |         5 |         1 |         6 |
##                               |    0.0001 |    0.0001 |           |
## -----------------------------|-----------|-----------|-----------|
##                     Internado |        20 |        14 |        34 |
##                               |    0.0004 |    0.0013 |           |
## -----------------------------|-----------|-----------|-----------|
##                        Óbito |        60 |         8 |        68 |
##                               |    0.0011 |    0.0007 |           |
## -----------------------------|-----------|-----------|-----------|
##                         Cura |     49960 |     10273 |     60233 |
##                               |    0.8978 |    0.9175 |           |
## -----------------------------|-----------|-----------|-----------|
##                 Column Total |     55646 |     11197 |     66843 |
##                               |    0.8325 |    0.1675 |           |
## -----------------------------|-----------|-----------|-----------|
##
##
```

Logo, do total de pessoas que tomaram vacina, 0,01% foram internadas em UTI e considerando o total de pessoas que não tomaram vacina 0,01% foram internadas em UTI.

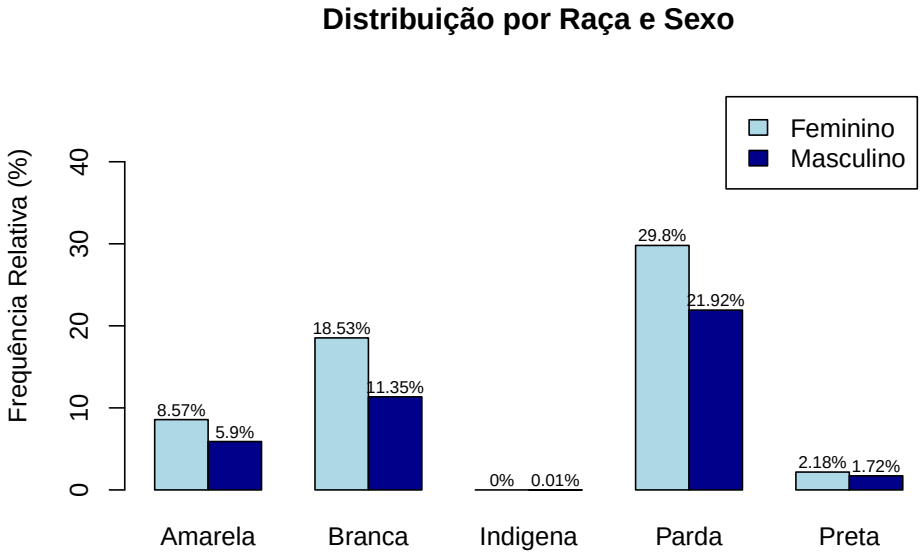
Item 11 - Faça um gráfico de colunas justapostas para as variáveis raça (racaCor) e sexo. Exclua valores ausentes (NA's) e Ignorados de ambas as variáveis. Utilize a frequência relativa porcentagem para o eixo y do gráfico. Responda: qual a categoria com maior frequência conjunta dentre as notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás?

Solução: Com base no gráfico abaixo, a categoria com maior frequência conjunta dentre as notificações de suspeita de Covid-19 no ano de 2024 no Estado de Goiás é a Parda.

Para gerar o gráfico filtramos para remover os "Ignorados", usamos a função *table* e *prop_table* para gerar as frequencias. Após essas tratativas, usamos a função *barplot* para criar o gráfico de colunas justapostas solicitados. Optamos por omitir o código.

```
##
##             Amarela Branca Indigena Parda Preta
##   Feminino    12723  27518        7 44258  3234
##   Masculino    8766  16857       20 32555  2555
```

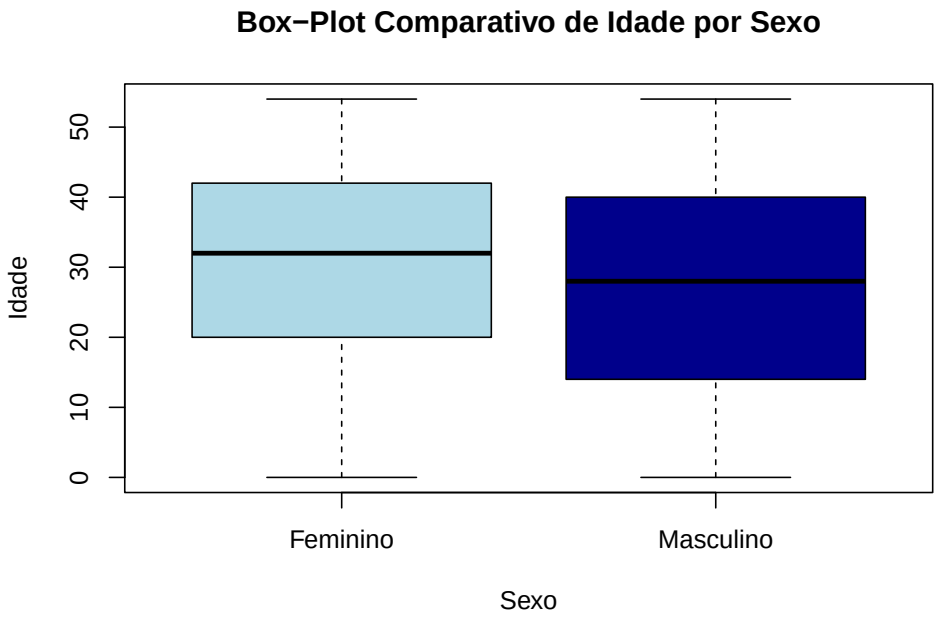
```
##
##             Amarela Branca Indigena Parda Preta
##   Feminino     8.57  18.53     0.00 29.80  2.18
##   Masculino    5.90  11.35     0.01 21.92  1.72
```



Item 12 - Faça um box-plot comparativo das variáveis idade e sexo e comente a respeito da dispersão e da idade mediana dos dois grupos.

Solução:

Optamos por omitir o código que gerou o gráfico abaixo. Fizemos a filtragem dos dados para remover os dados correspondentes a "Ignorados" e utilizamos a função *boxplot*.



Com base no gráfico acima, observamos que não há outliers significativos, o que sugere que as idades dos indivíduos dentro do sexo Feminino e Masculino estão relativamente próximas entre si. Além disso, observamos que a mediana do sexo feminino, superiormente próxima de 30 anos, é ligeiramente maior do que a do sexo Masculino que está ligeiramente abaixo de 30 anos.