INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

CAMPUS SÃO PAULO

GABRIEL DE AZEVEDO CAMARGO

GUSTAVO DELBON SOUZA

HUGO VINICIUS MATOS DA SILVA

Análise Exploratória de Dados:
Febre Amarela em Humanos e Primatas Não-Humanos - 1994 a 2021

SÃO PAULO - SP

2023

Para o nosso projeto, optamos por analisar a base de dados "Febre Amarela em Humanos e Primatas Não-Humanos - 1994 a 2021", concentrando-nos na investigação dos diversos fatores que podem impactar a taxa de mortalidade associada a essa enfermidade.

Para realização dos testes, foi utilizado o seguinte teste (que também se encontra disponível no GitHub):

```
# Instale o tidyverse se ainda não tiver instalado
# install.packages("tidyverse")
install.packages("tidyverse")
library(tidyverse)

# Carregar dados
casos_humanos <- fa_casoshumanos_1994_2021_1_
class(casos_humanos)
epizootias <- fa_epizpnh_1999_2021

str(casos_humanos[2154,])

str(epizootias)

# Remover linhas com valores ausentes
casos_humanos <- casos_humanos %>% drop_na()
epizootias <- epizootias %>% drop_na()

str(casos_humanos)

str(epizootias)

library(ggplot2)
library(dplyr)
casos_humanos <- head(casos_humanos, 550)
epizootias <- head(epizootias,550)

# Substitua 'Assintom�tico' por NA em 'DT_IS'
casos_humanos$DT_IS[casos_humanos$DT_IS == 'Assintom�tico'] <- NA

# Filtra linhas onde 'DT_IS' e 'DATA_OCOR' são datas válidas
dados_combinados_comb2 <- inner_join(
  casos_humanos %>% filter(!is.na(DT_IS)),
  epizootias %>% filter(!is.na(DATA_OCOR)),
  by = c("COD_MUN_LPI" = "COD_MUN_OCOR")
)

# Crie um gráfico de barras empilhadas
ggplot(dados_combinados_comb2, aes(x = DATA_OCOR, fill = factor(COD_MUN_LPI)))
+
```

```r
  geom_bar(position = "stack") +
  labs(x = "Data de Ocorrência", y = "Quantidade de Casos") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_discrete(name = "Município LPI")



# Excluindo linhas com 'OBITO' igual a 'IGN'
casos_humanos <- casos_humanos[casos_humanos$OBITO != "IGN", ]

# Convertendo 'OBITO' para fator
casos_humanos$OBITO <- factor(casos_humanos$OBITO, levels = c("SIM", "NÃO"))

# Histograma
histograma_idade <- ggplot(casos_humanos, aes(x = IDADE, fill = OBITO)) +
  geom_histogram(binwidth = 5, position = "identity", alpha = 0.7) +
  labs(x = "Idade", y = "Contagem", fill = "Óbito") +
  ggtitle("Distribuição de Idade por Óbito")

# Barplot
barplot_idade <- ggplot(casos_humanos, aes(x = OBITO, fill = OBITO)) +
  geom_bar() +
  labs(x = "Óbito", y = "Contagem", fill = "Óbito") +
  ggtitle("Contagem de Óbito e Sobrevivências") +
  theme_minimal()

# Visualizando os gráficos
print(histograma_idade)
print(barplot_idade)


# Suponha que você tenha um dataframe chamado 'seu_dataset'
# com colunas 'SEXO' e 'OBITO'

# Excluindo linhas com 'OBITO' igual a 'IGN'
casos_humanos <- casos_humanos[casos_humanos$OBITO != "IGN", ]

# Criando um histograma
histograma <- table(casos_humanos$SEXO, casos_humanos$OBITO)

# Plotando o histograma
barplot(histograma, beside = TRUE, legend = rownames(histograma), col =
c("blue", "red"), main = "Relação Sexo x Óbito", xlab = "Óbitos", ylab =
"Contagem")


# Localização x Óbito
casos_humanos %>%
  ggplot(aes(x = UF_LPI, fill = OBITO)) +
```

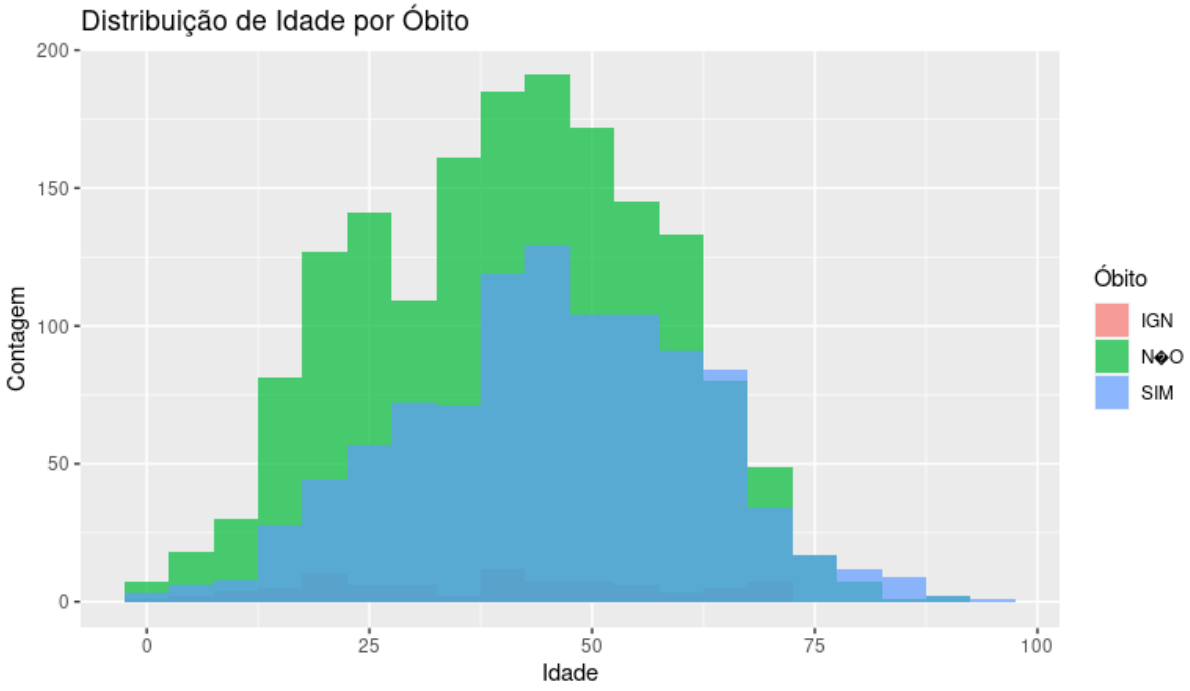
```r
  geom_bar(position = "dodge") +
  labs(x = "UF Completo (Mortes e sobrevivencias)", y = "Contagem") +
  scale_fill_manual(values = c("blue", "red"))

# Histogramas
# Histograma por Idade
ggplot(casos_humanos, aes(x = IDADE)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(x = "Idade", y = "Número de Casos")

# Histograma por Estado
ggplot(casos_humanos, aes(x = UF_LPI)) +
  geom_bar(fill = "blue", color = "black") +
  labs(x = "UF", y = "Contagem")

# Histograma por Semana Epidemiológica
ggplot(casos_humanos, aes(x = SE_IS)) +
  geom_bar(fill = "blue", color = "black") +
  labs(x = "Semana Epidemiológica", y = "Contagem")

# Histograma por Semana Epidemiológica
ggplot(casos_humanos, aes(x = MACRORREG_LPI)) +
  geom_bar(fill = "blue", color = "black") +
  labs(x = "MACRO REGIÃO", y = "Contagem")

# Instale o pacote GGally se ainda não estiver instalado
# install.packages("GGally")

install.packages("GGally")

# Instale os pacotes se ainda não estiverem instalados
# install.packages(c("GGally", "dplyr"))

library(GGally)
library(dplyr)
casos_humanos <- head(casos_humanos, 2399)

# Selecione as variáveis relevantes nos dataframes
variaveis_casos_humanos <- casos_humanos %>%
  select(COD_UF_LPI, COD_MUN_LPI)

variaveis_epizootias <- epizootias %>%
  select(COD_UF_OCOR, COD_MUN_OCOR)

# Combine os dataframes
dados_combinados <- cbind(variaveis_casos_humanos, variaveis_epizootias)

# Crie a matriz de dispersão
scatter_matrix <- ggpairs(dados_combinados)
```

```r
# Visualize a matriz de dispersão
print(scatter_matrix)

'''
library(GGally)
library(dplyr)
casos_humanos <- head(casos_humanos, 2399)
# Substitua 'Assintom�tico' por NA em 'DT_IS'
casos_humanos$DT_IS[casos_humanos$DT_IS == 'Assintom�tico'] <- NA

library(GGally)
library(dplyr)

# Substitua 'Assintom�tico' por NA em 'DT_IS'
casos_humanos$DT_IS[casos_humanos$DT_IS == 'Assintom�tico'] <- NA

# Filtra linhas onde 'DT_IS' e 'DATA_OCOR' são datas válidas
dados_combinados_comb2 <- inner_join(
  casos_humanos %>% filter(!is.na(DT_IS)),
  epizootias %>% filter(!is.na(DATA_OCOR)),
  by = c("COD_MUN_LPI" = "COD_MUN_OCOR")
)

# Crie a matriz de dispersão
scatter_matrix_comb2 <- ggpairs(dados_combinados_comb2, cardinality_threshold
= 1000)

# Visualize a matriz de dispersão
print(scatter_matrix_comb2)


# Crie a matriz de dispersão
scatter_matrix_comb2 <- ggpairs(dados_combinados_comb2)

# Visualize a matriz de dispersão
print(scatter_matrix_comb2)
'''

# Exemplo de teste t para amostras independentes
resultado_teste_idade <- t.test(casos_humanos$IDADE, casos_humanos$OBITO,
na.action = na.omit)

# Exibir os resultados do teste
print(resultado_teste_idade)

# Exemplo de teste Qui-Quadrado
tabela_contingencia <- table(casos_humanos$SEXO, casos_humanos$OBITO)
resultado_teste_qui_quadrado <- chisq.test(tabela_contingencia)
```

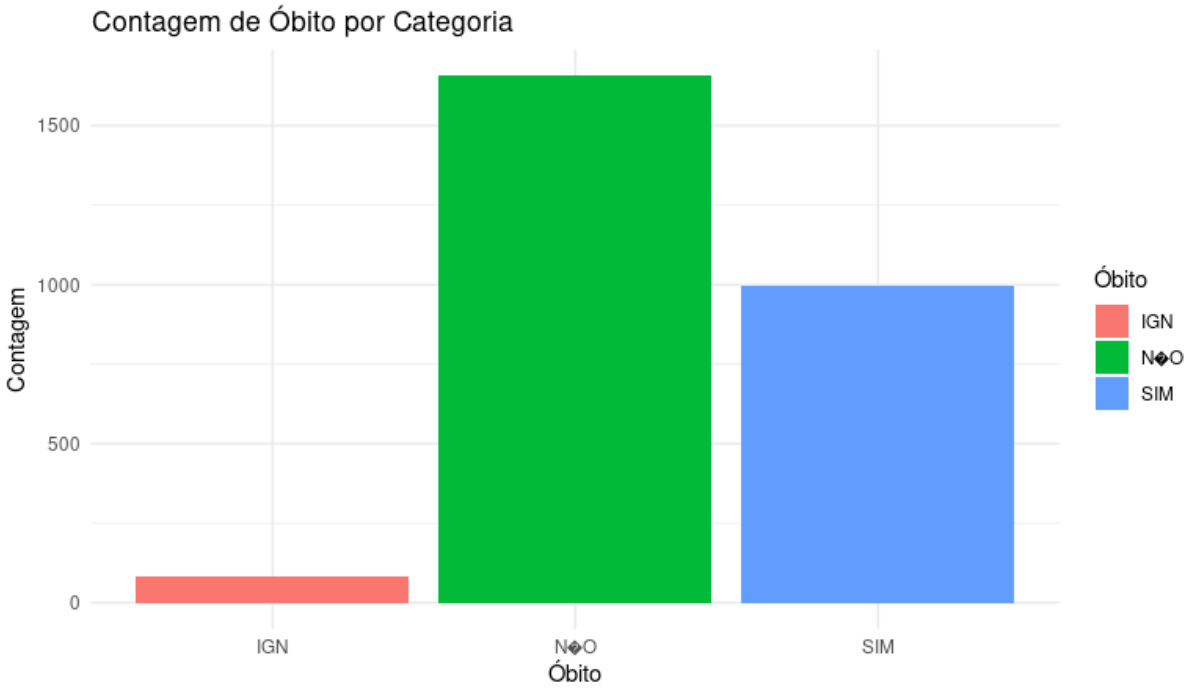
```r
# Exibir os resultados do teste
print(resultado_teste_qui_quadrado)

# Exemplo de teste de Kruskal-Wallis
resultado_teste_kruskal_wallis <- kruskal.test(IDADE ~ OBITO, data =
casos_humanos)

# Exibir os resultados do teste
print(resultado_teste_kruskal_wallis)

# Exemplo de teste de Kruskal-Wallis
resultado_teste_kruskal_wallis <- kruskal.test(SE_IS  ~ OBITO, data =
casos_humanos)
print(resultado_teste_kruskal_wallis)

# Supondo que você tenha os conjuntos de dados casos_humanos e epizootias
# Certifique-se de substituir isso pelos nomes reais dos seus conjuntos de
dados

# Instalando e carregando os pacotes, caso ainda não tenham sido
instalados/carregados
if (!requireNamespace("tidyverse", quietly = TRUE)) {
  install.packages("tidyverse")
}
library(tidyverse)

# Exemplo de dados (substitua pelos seus dados reais)
casos_humanos <- head(fa_casoshumanos_1994_2021_1_, 2399)
epizootias <- fa_epizpnh_1999_2021

# Realizando o teste de Kruskal-Wallis
resultado_teste_kruskal_wallis <- kruskal.test(SE_IS ~
factor(epizootias$DATA_OCOR), data = casos_humanos)
print(resultado_teste_kruskal_wallis)


# Supondo que você tenha os conjuntos de dados casos_humanos e epizootias
# Certifique-se de substituir isso pelos nomes reais dos seus conjuntos de
dados

# Exemplo de teste de Kruskal-Wallis
resultado_teste_kruskal_wallis <- kruskal.test(UF_LPI ~ OBITO, data =
casos_humanos)

# Exibir os resultados do teste
print(resultado_teste_kruskal_wallis)

# Exemplo de teste de Kruskal-Wallis
```

```
resultado_teste_kruskal_wallis <- kruskal.test(MACRORREG_LPI ~ OBITO, data =
casos_humanos)

# Exibir os resultados do teste
print(resultado_teste_kruskal_wallis)
```

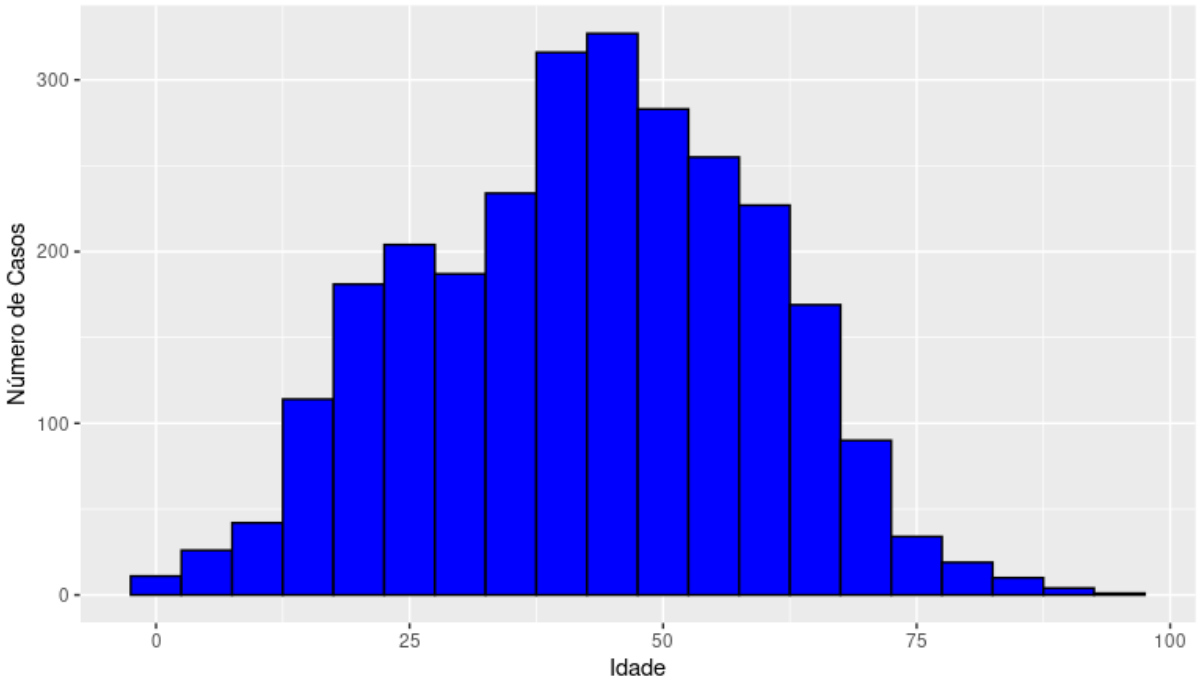
Mediante um levantamento da distribuição de idade em relação aos óbitos, apresentamos o código utilizado e o histograma resultante abaixo. Inicialmente, ao examinarmos o histograma, observamos uma tendência de maior incidência de óbitos em indivíduos com 60 anos ou mais. Essa observação é respaldada pelo fato de que o número de óbitos supera significativamente o número de não óbitos nessa faixa etária específica.



Ao examinarmos a correlação entre o número de casos e o número de óbitos, torna-se evidente que esta doença exibe uma taxa de mortalidade significativamente elevada, aproximando-se dos 40%. Essa constatação sublinha a gravidade da enfermidade em questão.

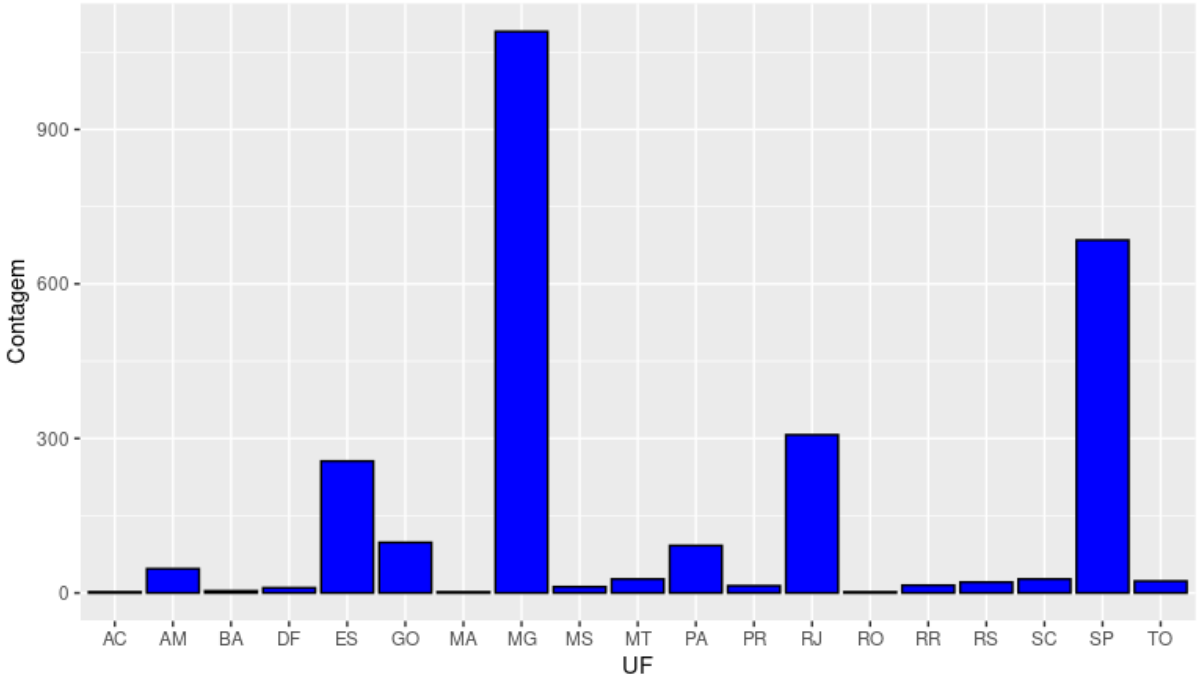


Ao analisarmos a distribuição do número de casos por faixa etária, observamos que a doença tem uma maior incidência entre pessoas na fase adulta, principalmente entre os 35 e 45 anos.



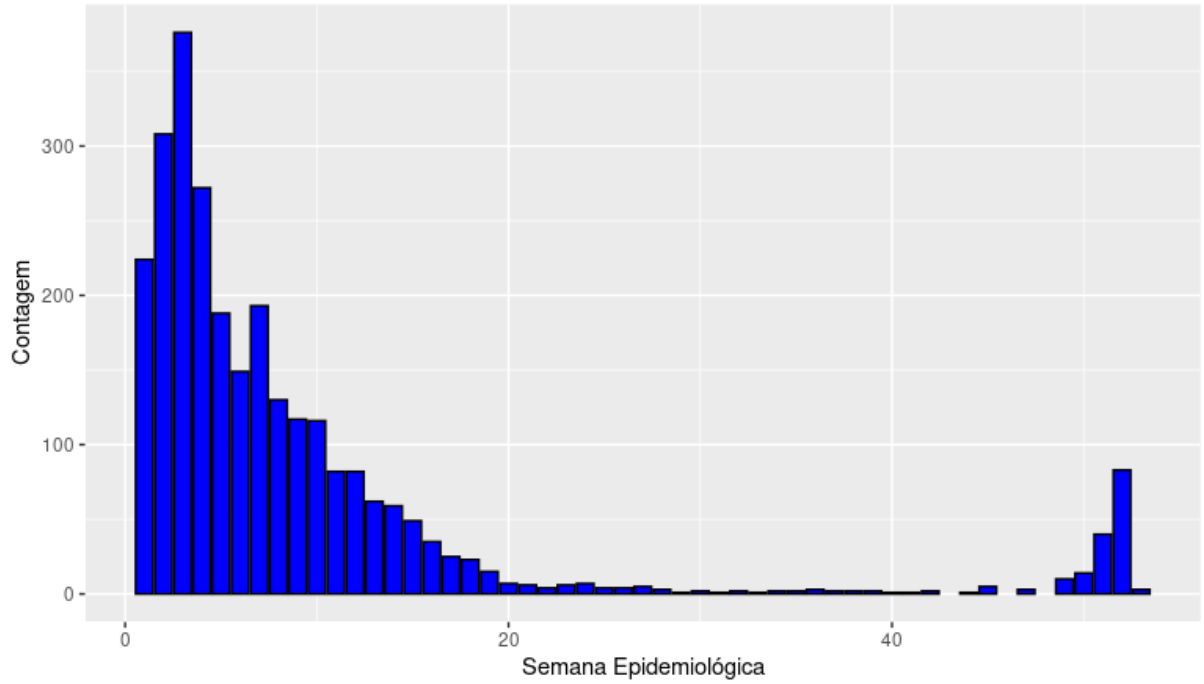
Ao examinarmos a relação entre o número de casos e os estados, constatamos que Minas Gerais, São Paulo e Rio de Janeiro emergem como os três mais

impactados pela doença. Coincidentemente, esses estados representam também os três mais populosos do Brasil. Essa correlação sugere uma possível influência da densidade populacional na propagação da enfermidade.

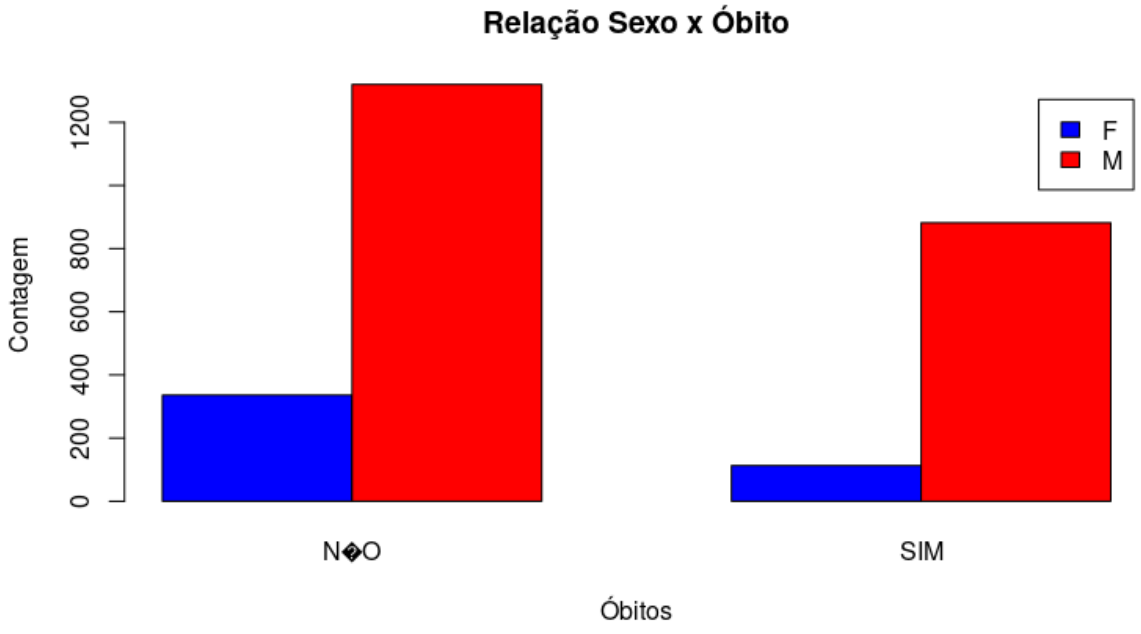


A semana epidemiológica serve como uma unidade temporal fundamental em epidemiologia, frequentemente empregada para a análise e relato da incidência de doenças ao longo do tempo. Essa medida padroniza a apresentação dos dados, simplificando a comparação e a análise temporal. O monitoramento semanal possibilita a identificação de tendências, surtos e padrões sazonais. Conforme ilustrado no gráfico acima, que representa o número de casos por semana epidemiológica analisada, observa-se que o pico de incidência se situou, por exemplo, entre 350 e 400 casos.

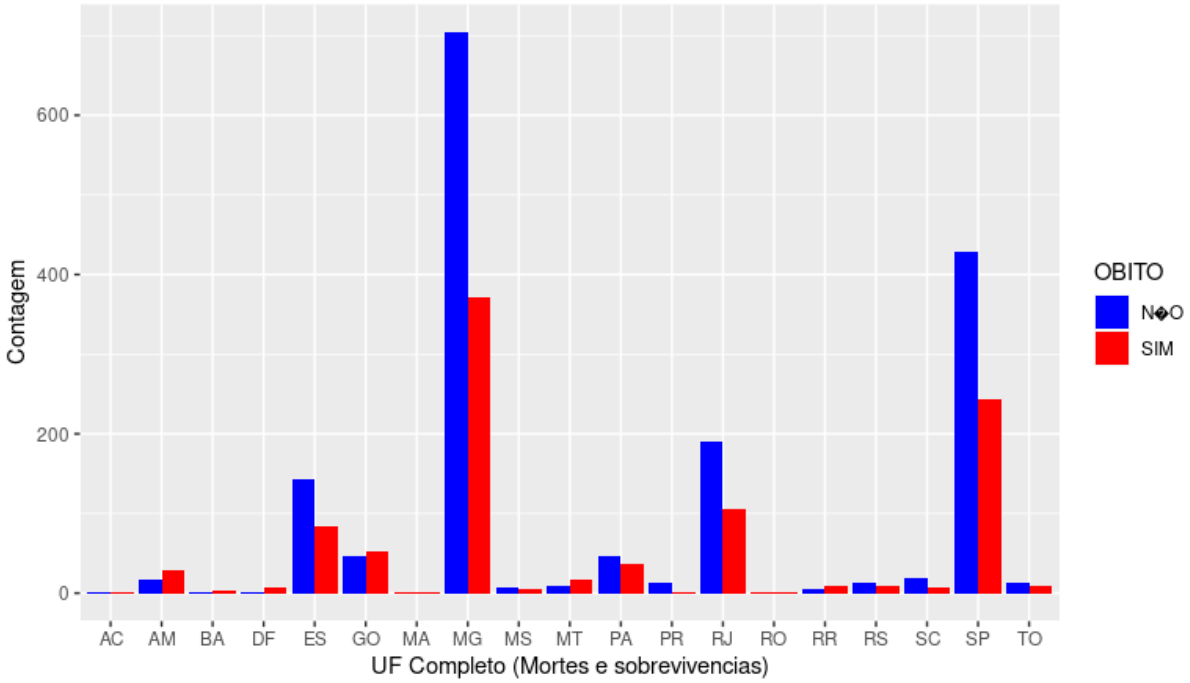
Ao examinarmos a relação entre o número de infectados e a semana epidemiológica, concluímos que a propensão dessa doença é se disseminar mais nas primeiras semanas do ano. Esses períodos coincidem com a estação do verão no Brasil, evidenciando uma possível associação entre a sazonalidade climática e a incidência da doença.



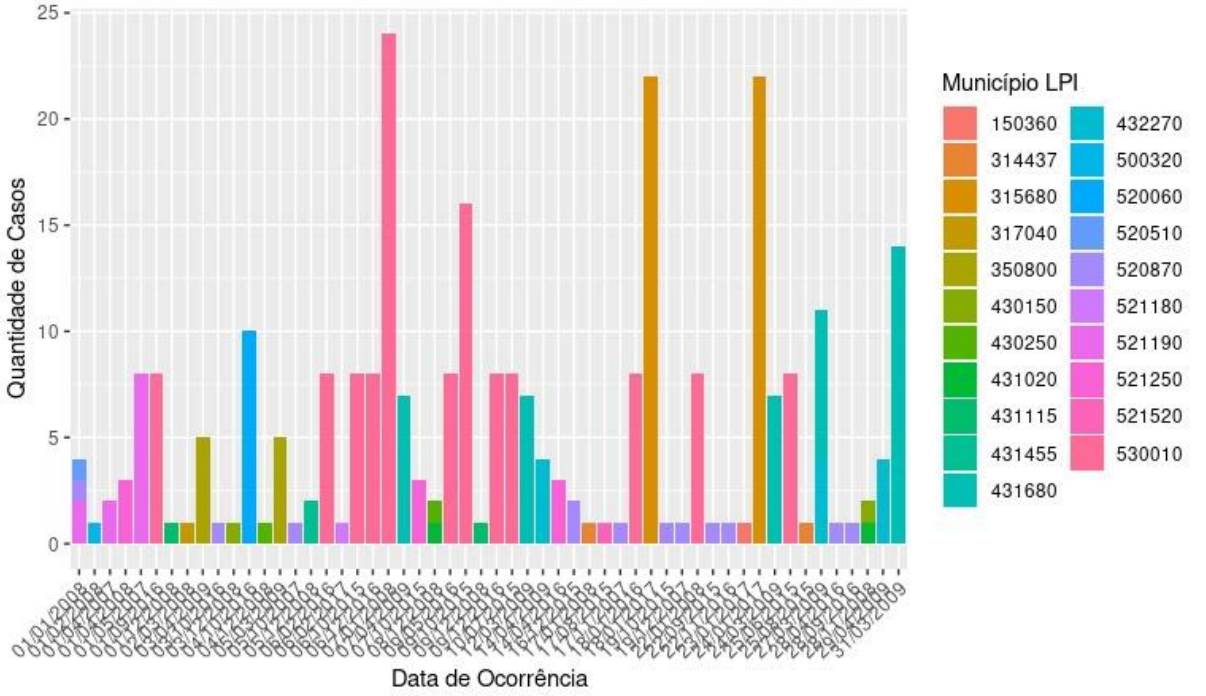
Ao analisarmos a relação entre a taxa de mortalidade e o sexo, constatamos que o gênero não exerce uma influência significativa na taxa de óbito da doença. Isso se deve ao fato de que, embora mais homens venham a óbito devido à doença, a incidência dela também é mais alta entre os homens. Portanto, proporcionalmente, as taxas de mortalidade entre ambos os sexos são bastante semelhantes.



Ao examinarmos a inter-relação entre o número de casos, os estados e os óbitos, notamos que, embora estados como São Paulo, Rio de Janeiro, Minas Gerais e Espírito Santo apresentem uma elevada taxa de contágio da doença, isso não se traduz necessariamente em uma alta taxa de mortalidade proporcional. Curiosamente, estados como Amazonas, Mato Grosso e Goiás demonstram uma taxa de sobrevivência consideravelmente menor, visto que o número de óbitos supera o de sobreviventes. Essa análise destaca a importância de considerar não apenas a incidência da doença, mas também sua letalidade em diferentes regiões.



A correlação entre a ocorrência de infecções em primatas e humanos está intrinsecamente ligada à localização geográfica, especificamente no que diz respeito ao município em que os macacos e humanos são afetados. A análise revela que a quantidade de casos aumenta proporcionalmente à coincidência temporal da ocorrência de ambos, com 550 casos sendo o ponto de referência. Essa interdependência ressalta que a incidência de infecções em macacos influencia diretamente o número de casos em humanos, destacando a importância da análise conjunta desses fatores para uma compreensão mais abrangente do cenário epidemiológico.



Por fim, fizemos os testes de resultado_teste_qui_quadrado e Kruskal-Wallis e estes foram os resultados:

```
        Kruskal-Wallis rank sum test

data:  IDADE by OBITO
Kruskal-Wallis chi-squared = 49.232, df = 2, p-value = 2.039e-11

> # Exemplo de teste Qui-Quadrado
> tabela_contingencia <- table(casos_humanos$SEXO, casos_humanos$OBITO)
> resultado_teste_qui_quadrado <- chisq.test(tabela_contingencia)
> # Exibir os resultados do teste
> print(resultado_teste_qui_quadrado)

        Pearson's Chi-squared test

data:  tabela_contingencia
X-squared = 37.803, df = 2, p-value = 6.182e-09

> # Exemplo de teste de Kruskal-Wallis
> resultado_teste_kruskal_wallis <- kruskal.test(IDADE ~ OBITO, data = casos_humanos)
> # Exibir os resultados do teste
> print(resultado_teste_kruskal_wallis)

        Kruskal-Wallis rank sum test

data:  IDADE by OBITO
Kruskal-Wallis chi-squared = 49.232, df = 2, p-value = 2.039e-11

> # Exemplo de teste de Kruskal-Wallis
> resultado_teste_kruskal_wallis <- kruskal.test(SE_IS  ~ OBITO, data = casos_humanos)
> print(resultado_teste_kruskal_wallis)

        Kruskal-Wallis rank sum test

data:  SE_IS by OBITO
Kruskal-Wallis chi-squared = 83.082, df = 2, p-value < 2.2e-16

> # Instalando e carregando os pacotes, caso ainda não tenham sido instalados/carregados
```

```
+ }
> library(tidyverse)
> # Exemplo de dados (substitua pelos seus dados reais)
> casos_humanos <- head(fa_casoshumanos_1994_2021_1_, 2399)
> epizootias <- fa_epizpnh_1999_2021
> # Realizando o teste de Kruskal-Wallis
> resultado_teste_kruskal_wallis <- kruskal.test(SE_IS ~ factor(epizootias$DATA_OCOR), data = casos_humanos)
Error in model.frame.default(formula = SE_IS ~ factor(epizootias$DATA_OCOR),  :
  variable lengths differ (found for 'factor(epizootias$DATA_OCOR)')
> print(resultado_teste_kruskal_wallis)

        Kruskal-Wallis rank sum test

data:  SE_IS by OBITO
Kruskal-Wallis chi-squared = 83.082, df = 2, p-value < 2.2e-16

> # Exemplo de teste de Kruskal-Wallis
> resultado_teste_kruskal_wallis <- kruskal.test(UF_LPI ~ OBITO, data = casos_humanos)
> # Exibir os resultados do teste
> print(resultado_teste_kruskal_wallis)

        Kruskal-Wallis rank sum test

data:  UF_LPI by OBITO
Kruskal-Wallis chi-squared = 33.358, df = 2, p-value = 5.707e-08

> # Exemplo de teste de Kruskal-Wallis
> resultado_teste_kruskal_wallis <- kruskal.test(MACRORREG_LPI ~ OBITO, data = casos_humanos)
> # Exibir os resultados do teste
> print(resultado_teste_kruskal_wallis)

        Kruskal-Wallis rank sum test

data:  MACRORREG_LPI by OBITO
Kruskal-Wallis chi-squared = 37.327, df = 2, p-value = 7.846e-09
```