INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO

GEAN CARLOS DE SOUSA BANDEIRA

SP3030075

ANÁLISE EXPLORATÓRIA DE DADOS DO CADASTRO NACIONAL DE ESTABELECIMENTOS DE SAÚDE

SÃO PAULO

2024

GEAN CARLOS DE SOUSA BANDEIRA

ANÁLISE EXPLORATÓRIA DE DADOS DO CADASTRO NACIONAL DE ESTABELECIMENTOS DE SAÚDE

Projeto da disciplina Estatística e Probabilidade apresentado ao Instituto Federal de São Paulo do curso superior de Análise e Desenvolvimento de Sistemas.
Orientadora: Josceli Tenorio.

SÃO PAULO
2024

Sumário

Introdução

O Cadastro Nacional de Estabelecimentos de Saúde (CNES) é crucial no Brasil para registrar e monitorar os estabelecimentos de saúde, permitindo uma análise detalhada da infraestrutura e dos recursos disponíveis no sistema de saúde do país. Este trabalho visa explorar e analisar os dados do CNES, aplicando conceitos como estatística descritiva, probabilidade e inferência estatística, aprendidos ao longo da disciplina.

Neste relatório, descreveremos os cálculos e gráficos realizados utilizando o software posit.cound na Linguagem R para análise de dados. A metodologia inclui a formulação de perguntas específicas aos dados do CNES para compreender a distribuição e características dos estabelecimentos de saúde, identificando áreas potenciais para intervenção.

As análises abrangem:

Estatística Descritiva: Resumo dos dados com medidas de tendência central e dispersão.

Probabilidade: Modelagem de incertezas e previsão de eventos dentro da infraestrutura de saúde.

Inferência Estatística: Teste de hipóteses e conclusões sobre a população de estabelecimentos de saúde a partir da amostra analisada.

Este trabalho demonstra a importância das técnicas estatísticas na interpretação de dados de saúde e como essas análises podem subsidiar decisões informadas em políticas públicas de saúde. A utilização da Linguagem R facilita a manipulação e visualização dos dados, tornando as conclusões mais acessíveis.

Nos próximos capítulos, apresentaremos a descrição do problema, as análises realizadas, os resultados obtidos e suas interpretações, com foco em responder às questões formuladas aos dados e discutir as implicações para a gestão e melhoria dos serviços de saúde no Brasil.

Relatório: Análise exploratória de dados

Após abrir o Posit.clound e criar um arquivo em linguagem R, no canto inferior direito fiz o upload do arquivo cnes_estabelecimentos.csv.

Instalar os pacotes:

library(tidyverse) #Manipulação de dados

library(psych) #Estatísticas psicométricas

library(skimr) #Sumarização de dados

library(ggplot2) #Visualização gráfica

library(readr) #Leitura de dados

library(dplyr) #Manipulação de dados

library(MASS) #Métodos estatísticos

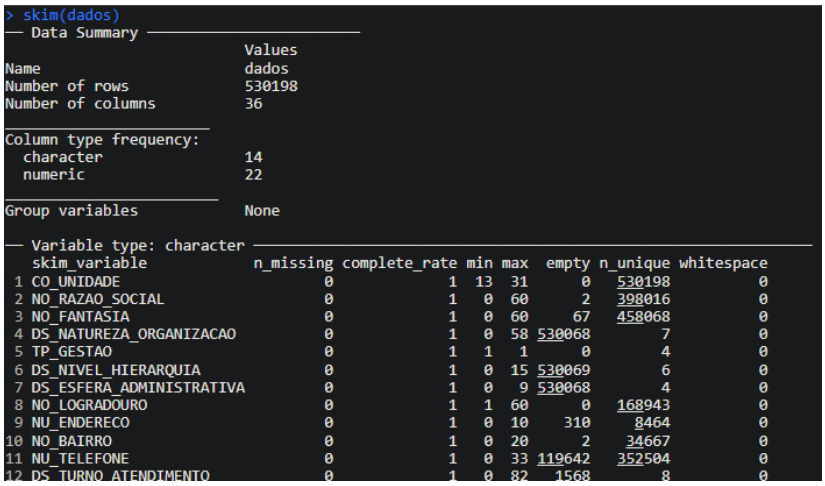
library(stats) Estatísticas básicas

Fazer a leitura dos dados de forma básica:

#Ler o arquivo csv, separando por;

dados <- read.csv(file.choose(), sep = ";")

Utilizei a função skim(dados) para resumir e visualizar os dados de um dataframe de forma concisa e eficiente.



```
> skim(dados)
── Data Summary ────────────────────────
                           Values
Name                       dados
Number of rows             530198
Number of columns          36
_____________________
Column type frequency:
  character                14
  numeric                  22
_____________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────────────────────
   skim_variable           n_missing complete_rate min max  empty n_unique whitespace
 1 CO_UNIDADE                      0             1  13  31      0   530198          0
 2 NO_RAZAO_SOCIAL                 0             1   0  60      2   398016          0
 3 NO_FANTASIA                     0             1   0  60     67   458068          0
 4 DS_NATUREZA_ORGANIZACAO         0             1   0  58 530068        7          0
 5 TP_GESTAO                       0             1   1   1      0        4          0
 6 DS_NIVEL_HIERARQUIA             0             1   0  15 530069        6          0
 7 DS_ESFERA_ADMINISTRATIVA        0             1   0   9 530068        4          0
 8 NO_LOGRADOURO                   0             1   1  60      0   168943          0
 9 NU_ENDERECO                     0             1   0  10    310     8464          0
10 NO_BAIRRO                       0             1   0  20      2    34667          0
11 NU_TELEFONE                     0             1   0  33 119642   352504          0
12 DS_TURNO_ATENDIMENTO            0             1   0  82   1568        8          0
```

Análise estatística descritiva de todos dados que utilizei:

Média, Mediana, Quartis, entre outros.

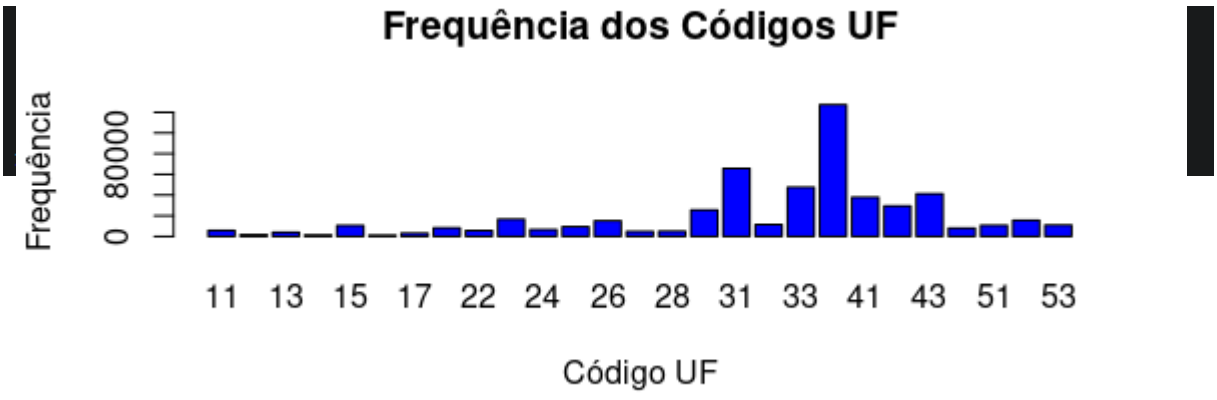
summary(dados)

```
> #Analise Estatistica descritiva de todos dados
> summary(dados)
    CO_CNES          CO_UNIDADE            CO_UF          CO_IBGE       NU_CNPJ_MANTENEDORA
 Min.   :     19   Length:530198      Min.   :11.00   Min.   :110001   Min.   :7.290e+08
 1st Qu.:2937259   Class :character   1st Qu.:31.00   1st Qu.:310620   1st Qu.:7.588e+12
 Median :5082807   Mode  :character   Median :35.00   Median :350950   Median :1.422e+13
 Mean   :5125723                      Mean   :34.29   Mean   :344750   Mean   :2.732e+13
 3rd Qu.:7288064                      3rd Qu.:41.00   3rd Qu.:411370   3rd Qu.:4.528e+13
 Max.   :9999981                      Max.   :53.00   Max.   :530180   Max.   :9.867e+13
                                                                       NA's   :417530
 NO_RAZAO_SOCIAL    NO_FANTASIA        CO_NATUREZA_ORGANIZACAO DS_NATUREZA_ORGANIZACAO
 Length:530198      Length:530198      Min.   : 1.0            Length:530198
 Class :character   Class :character   1st Qu.: 1.0            Class :character
 Mode  :character   Mode  :character   Median : 1.0            Mode  :character
                                       Mean   : 3.5
                                       3rd Qu.: 7.0
                                       Max.   :13.0
                                       NA's   :530068
  TP_GESTAO         CO_NIVEL_HIERARQUIA DS_NIVEL_HIERARQUIA CO_ESFERA_ADMINISTRATIVA
 Length:530198      Min.   :1.0         Length:530198       Min.   :2.0
 Class :character   1st Qu.:1.0         Class :character    1st Qu.:3.0
 Mode  :character   Median :1.0         Mode  :character    Median :3.0
                    Mean   :1.6                             Mean   :3.3
                    3rd Qu.:2.0                             3rd Qu.:4.0
                    Max.   :7.0                             Max.   :4.0
                    NA's   :530069                          NA's   :530068
 DS_ESFERA_ADMINISTRATIVA  CO_ATIVIDADE     TP_UNIDADE        CO_CEP
 Length:530198            Min.   :1.000   Min.   : 1.00   Min.   : 1001000
```

Contando a quantidade de estabelecimento de cada unidade federativa

```
contagem <- table(dados$CO_UF)
# Exibindo o resultado
print(contagem)
# Criando o gráfico de barras
barplot(contagem, main="Frequência dos Códigos UF", xlab="Código UF",
ylab="Frequência", col="blue")
```

Existem mais Estabelecimentos de Saúde nas unidades federativas 35, 31 e 33.



Contando a quantidade de ocorrências de cada tipo de gestão

```
contagem_gestao <- table(dados$TP_GESTAO)
# Exibindo o resultado
print(contagem_gestao)
# Criando o gráfico de barras
barplot(contagem_gestao,
        main="Frequência dos Tempo de Gestão",
        xlab="Tempo de Gestão",
        ylab="Frequência",
        col="blue",
        border="black")
```

Para enteder melhor:

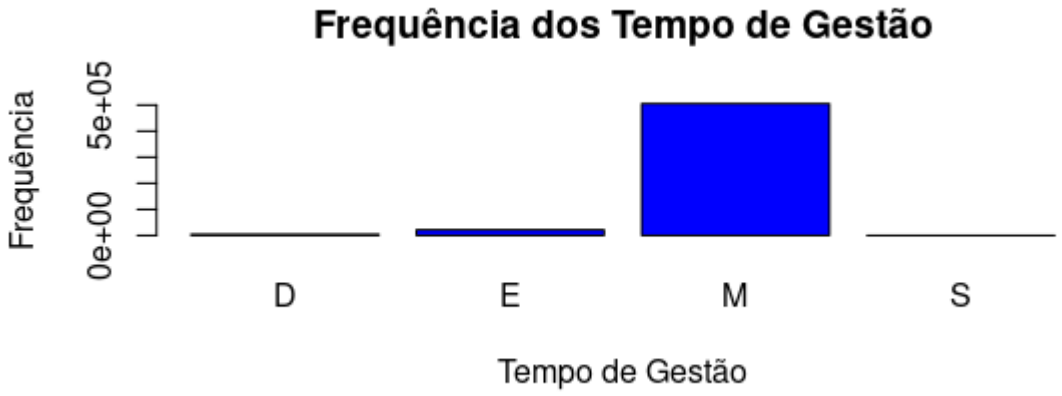
Curto Prazo (M - Mensal): Duração: até 1 mês;

Curto Prazo (E - Exato): Duração: até 1 ano;

Prazo (D - Duradouro): Duração: de 1 a 3 anos;

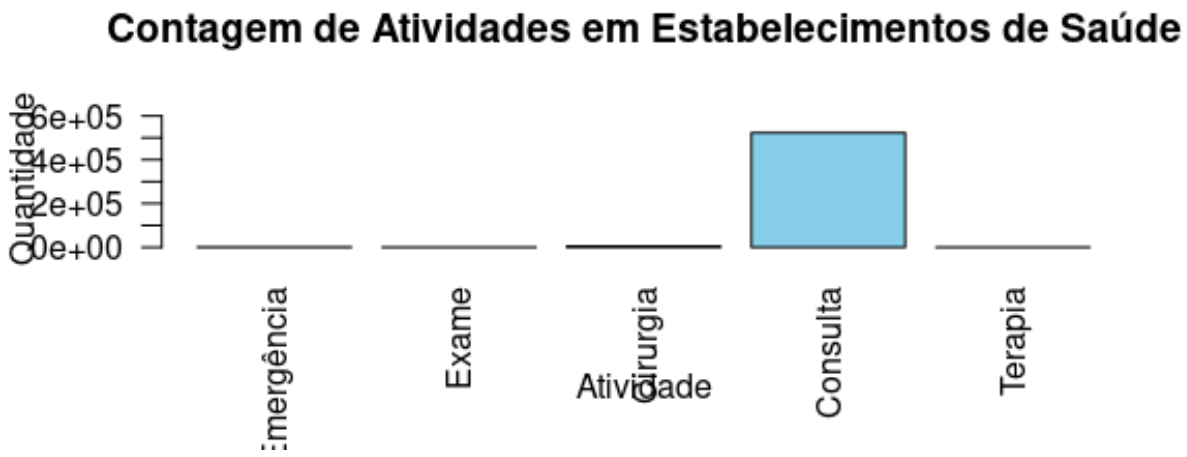
Longo Prazo (S - Sustentável): Duração: mais de 3 anos.

```
    D     E     M     S
 4056 21310 504729   103
```



Quais as atividades mais utilizadas nos estabelecimentos de saúde?

```
contagem_gestao <- table(dados$CO_ATIVIDADE)
print(contagem_gestao)
barplot(contagem_gestao,
        main = "Contagem de Atividades em Estabelecimentos de Saúde",
        xlab = "Atividade",
        ylab = "Quantidade",
        col = "skyblue",
        ylim = c(0, max(contagem_gestao) * 1.2),
        names.arg = c("Emergência", "Exame", "Cirurgia",  "Consulta",  "Terapia"),
        las = 2
)
```



As atividades mais utilizadas nos estabelecimentos de saúde são consulta bem superior a qualquer outra atividade e cirurgia em segundo lugar.

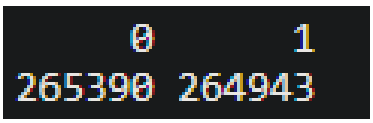
Qual a probabilidade de ter um centro cirúrgico em todos os estabelecimentos de saúde?

```
dados <- data.frame(
  ST_CENTRO_CIRURGICO = sample(c(0, 1), 530333, replace = TRUE)
)
# Contagem de 0s e 1s na variável ST_CENTRO_CIRURGICO
contagem_ctcirurgico <- table(dados$ST_CENTRO_CIRURGICO)
# Exibindo a contagem
print(contagem_ctcirurgico)
```

```
# Calculando a probabilidade de ter um centro cirúrgico
probabilidade <- contagem_ctcirurgico[1] / sum(contagem_ctcirurgico)
# Exibindo a probabilidade
print(paste("A probabilidade de um estabelecimento de saúde ter um centro cirúrgico é:", probabilidade))
```

O resultado 0 é não tem centro cirúrgico e o 1 tem centro cirúrgico.

A probabilidade de um estabelecimento de saúde ter um centro cirúrgico é: 0.50, ou seja, 50%.



Qual é a contagem de serviços ambulatórias cobertos pelo SUS?

```
dados <- data.frame(
  CO_AMBULATORIAL_SUS = sample(c(0, 1), 530333, replace = TRUE)
)
# Contagem de serviços ambulatoriais cobertos pelo SUS (valor 1)
contagem_sus <- sum(dados$CO_AMBULATORIAL_SUS == 1)
# Total de serviços ambulatoriais
total_servicos <- nrow(dados)
# Calculando a porcentagem
porcentagem_sus <- (contagem_sus / total_servicos) * 100

# Exibindo a resposta
print(paste("A porcentagem de serviços ambulatoriais cobertos pelo SUS é:", porcentagem_sus, "%"))
```

A porcentagem de serviços ambulatoriais cobertos pelo SUS é: 49.99%

Calcular se a média de atendimentos ambulatoriais é superior em estabelecimentos com tempo de gestão maior.

```
set.seed(123)  # Para reproduzir os mesmos resultados
dados <- data.frame(
  ST_ATEND_AMBULATORIAL = rnorm(530333, mean = 50, sd = 10),
```

```
# Exemplo de atendimentos ambulatoriais
  TP_GESTAO = sample(c("M", "E", "D", "S"), 530333, replace = TRUE)

# Exemplo de tempo de gestão
)
# Definindo os grupos com base no tempo de gestão
grupo_maior <- dados[dados$TP_GESTAO %in% c("D", "S"),
"ST_ATEND_AMBULATORIAL"]
grupo_menor <- dados[dados$TP_GESTAO %in% c("M", "E"),
"ST_ATEND_AMBULATORIAL"]
# Teste t de Student para duas amostras independentes
teste_t <- t.test(grupo_maior, grupo_menor, alternative = "greater", conf.level = 0.95)
# Exibindo o resultado do teste
print(teste_t)
# Conclusão do teste
if (teste_t$p.value < 0.05) {
  print("Podemos rejeitar a hipotese nula.")
  print("Há evidências estatísticas de que a média de atendimentos ambulatoriais é
superior em estabelecimentos com tempo de gestão maior.")
} else {
  print("Não podemos rejeitar a hipótese nula.")
  print("Não tem evidências suficientes para afirmar que a média de atendimentos
ambulatoriais é superior em estabelecimentos com tempo de gestão maior.")
}
```

O resultado foi: O p-value deu 0.9153, ou seja, não podemos rejeitar a hipótese nula, pois não há evidências suficientes para afirmar que a média de atendimentos ambulatoriais é superior em estabelecimentos com tempo de gestão maior."

Teste de hipótese para verificar se a média de pacientes atendidos por estabelecimento é significativamente diferente entre as unidades federativas.

```
t.test(dados$NU_CNPJ ~ dados$CO_ESFERA_ADMINISTRATIVA)
```

O p-value deu 0.61 com isso significa que não há evidência suficiente para rejeitar a hipótese nula, e a diferença entre as médias não é estatisticamente significativa.

Qual é a presença de Centros Cirúrgicos com base no tipo de gestão

```
contingency_table <- table(dados$TP_GESTAO, dados$ST_CENTRO_CIRURGICO)
# Realizar o teste qui-quadrado de independência
chi_test <- chisq.test(contingency_table)
chi_test
# Gráficos
ggplot(dados, aes(x = TP_GESTAO, fill = ST_CENTRO_CIRURGICO)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribuição do Tipo de Gestão por Presença de Centro Cirúrgico",
       x = "Tipo de Gestão",
       y = "Contagem")
```

A presença de centro cirúrgicos está presente principalmente em um tipo de gestão Curto Prazo (M - Mensal): Duração: até 1 mês. Características: envolve atividades e decisões muito imediatas e no tipo de gestão e Curto Prazo (E - Exato): Duração: até 1 ano. Características: envolve atividades e decisões que visam resultados rápidos dentro de um ano.

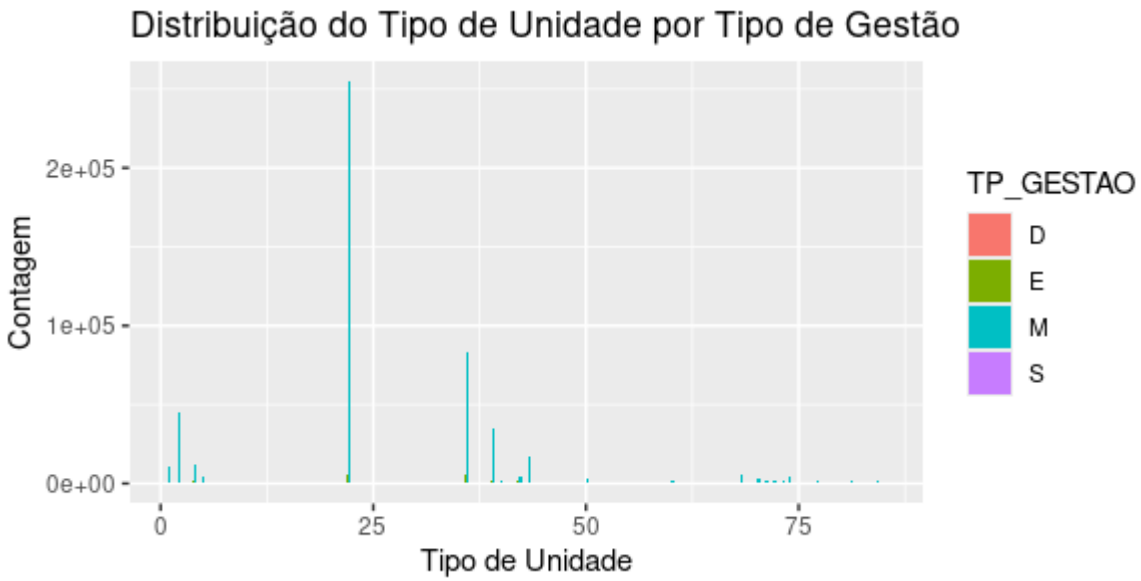
Diferença na presença de centro obstétrico entre diferentes esferas administrativas

```
tabeladecont3 <- table(dados$CO_ESFERA_ADMINISTRATIVA,
dados$ST_CENTRO_OBSTETRICO)
#  Teste qui-quadrado de independência
chi_test_3 <- chisq.test(tabeladecont3)
# Resultados
chi_test_3
```

Resultado: p-value 0.13 sendo assim, as diferença não é estatisticamente significativa.

Associação entre tipo de unidade e tipo de gestão

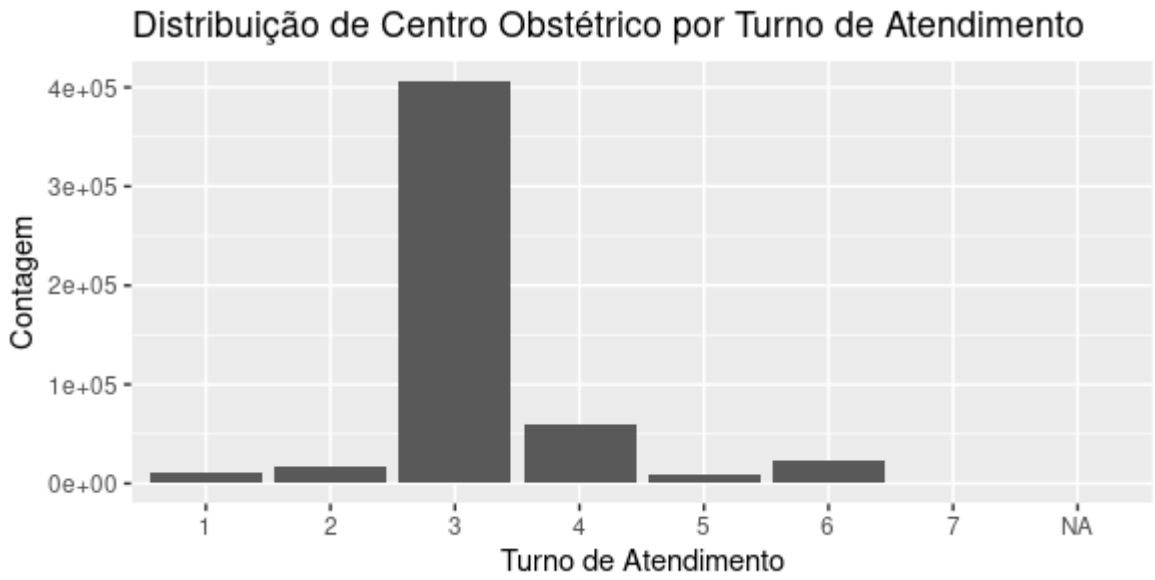
```
tabeladecont4 <- table(dados$TP_UNIDADE, dados$TP_GESTAO)
# Realizar o teste qui-quadrado de independência
chi_test_4 <- chisq.test(tabeladecont4)
# Resultados
chi_test_4
```



Resultado: p-value é 2,2e-16 ou seja a hipótese nula é rejeitada com uma confiança muito alta.

Associação entre presença de centro obstétrico e turno de atendimento

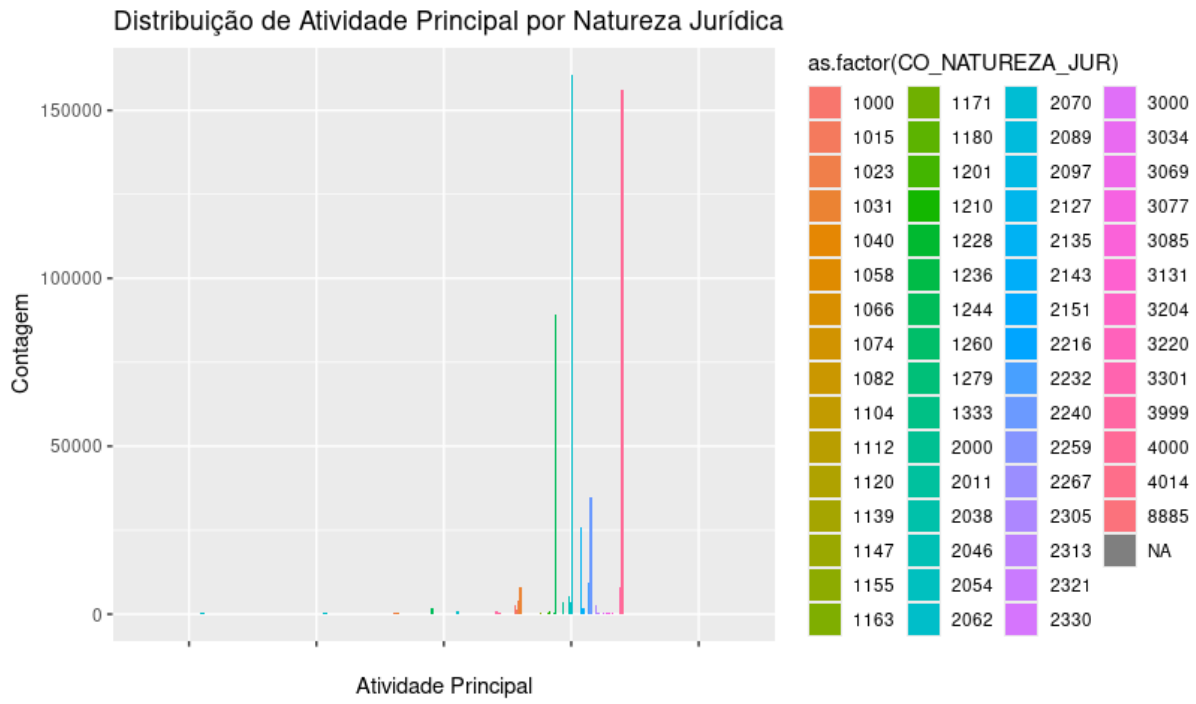
```
tabeladecont5 <- table(dados$ST_CENTRO_OBSTETRICO,
dados$CO_TURNO_ATENDIMENTO)
# Realizar o teste qui-quadrado de independência
chi_test_5 <- chisq.test(tabeladecont5)
# Mostrar os resultados do teste
chi_test_5
# Gráficos
ggplot(dados, aes(x = as.factor(CO_TURNO_ATENDIMENTO), fill =
ST_CENTRO_OBSTETRICO)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribuição de Centro Obstétrico por Turno de Atendimento",
       x = "Turno de Atendimento",
       y = "Contagem")
```



Resultado: Cada turno de atendimento sinaliza 4 horas de trabalho, por exemplo o turno 3, tem horário de trabalho de 12 horas, e conforme o gráfico os estabelecimentos que tem esse turno de atendimento é onde possuem mais centros obstétricos.

Associação entre atividade principal e natureza jurídica

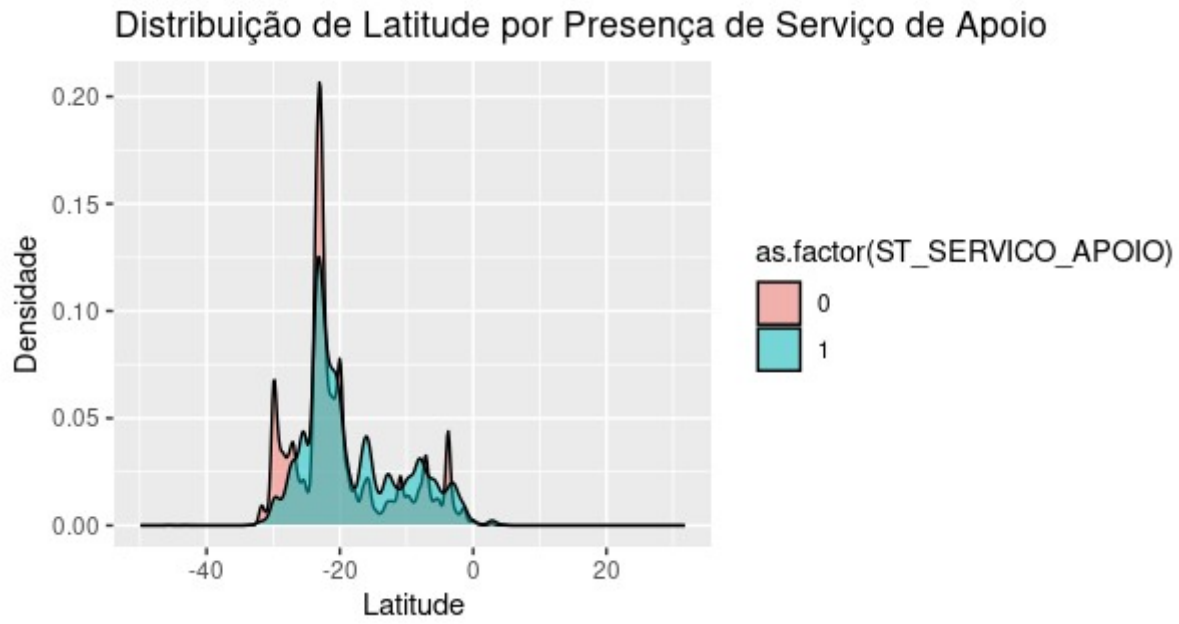
```
tabeladecont6 <- table(dados$CO_ATIVIDADE, dados$CO_NATUREZA_JUR)
# Realizar o teste qui-quadrado de independência
chi_test_6 <- chisq.test(tabeladecont6)
# Mostrar os resultados do teste
chi_test_6
ggplot(dados, aes(x = as.factor(CO_ATIVIDADE), fill =
as.factor(CO_NATUREZA_JUR))) +
  geom_bar(position = "dodge", width = 0.9) +  # Ajuste a largura das barras
  labs(title = "Distribuição de Atividade Principal por Natureza Jurídica",
       x = "Atividade Principal",
       y = "Contagem") +
  theme(axis.text.x = element_text(angle = 45, hjust = 12))  # Mudei o ângulo para 45
graus
```



Resultado: A distribuição das atividades principais por natureza jurídica tem alguns poi=ntos interessantes onde o número jurídico ou melhor as cores em verde, azul e vermelho levam um grande destaque de atuação.

Relação entre latitude e presença de serviço de apoio

```
data_7 <- dados %>% select(NU_LATITUDE, ST_SERVICO_APOIO) %>% na.omit()
# Realizar um teste t para comparar as latitudes médias
t_test_7 <- t.test(dados$NU_LATITUDE ~ dados$ST_SERVICO_APOIO, dados = data_7)
# Resultados
t_test_7
# Gráficos
ggplot(dados, aes(x = NU_LATITUDE, fill = as.factor(ST_SERVICO_APOIO))) +
  geom_density(alpha = 0.5) +
  labs(title = "Distribuição de Latitude por Presença de Serviço de Apoio",
       x = "Latitude",
       y = "Densidade")
```



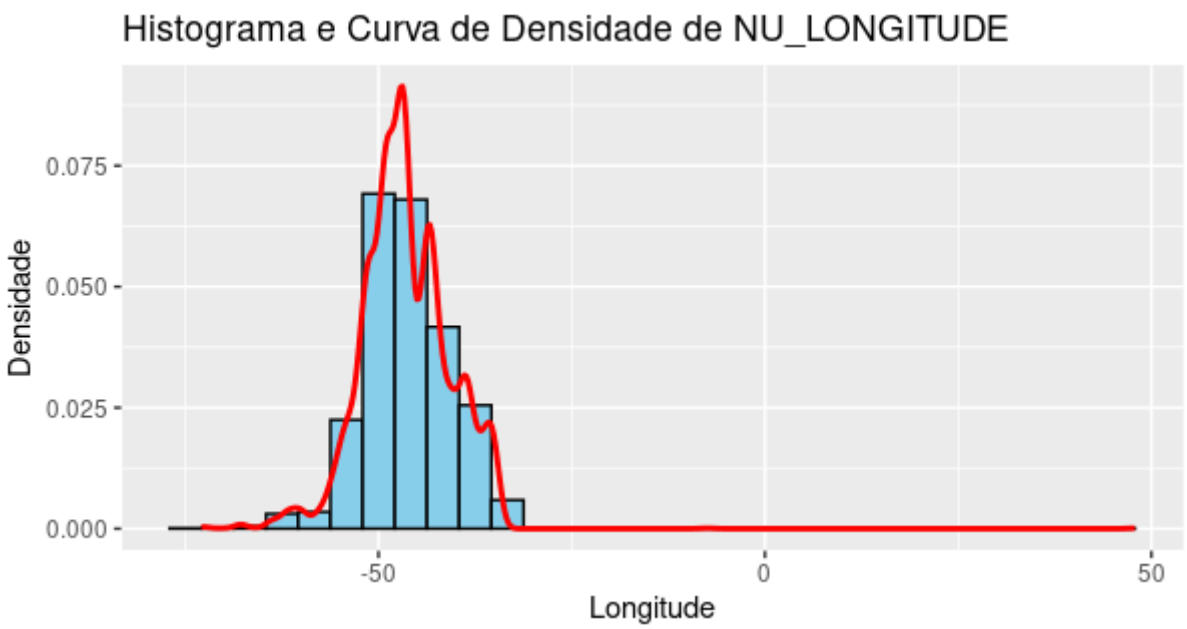
Resultado: 0 representa a ausência de serviços de apoio e 1 a presença de serviços de apoio, com isso, o p-value foi 2.2e-16 e o gráfico representa de forma fácil de analisar que indica que a amostra não é normal, porém, na latitude -25 por exemplo, possui uma alta presença de serviço de apoio mas com uma maior ainda ausência do seviço de apoio, contraditório, contudo, perceptível.

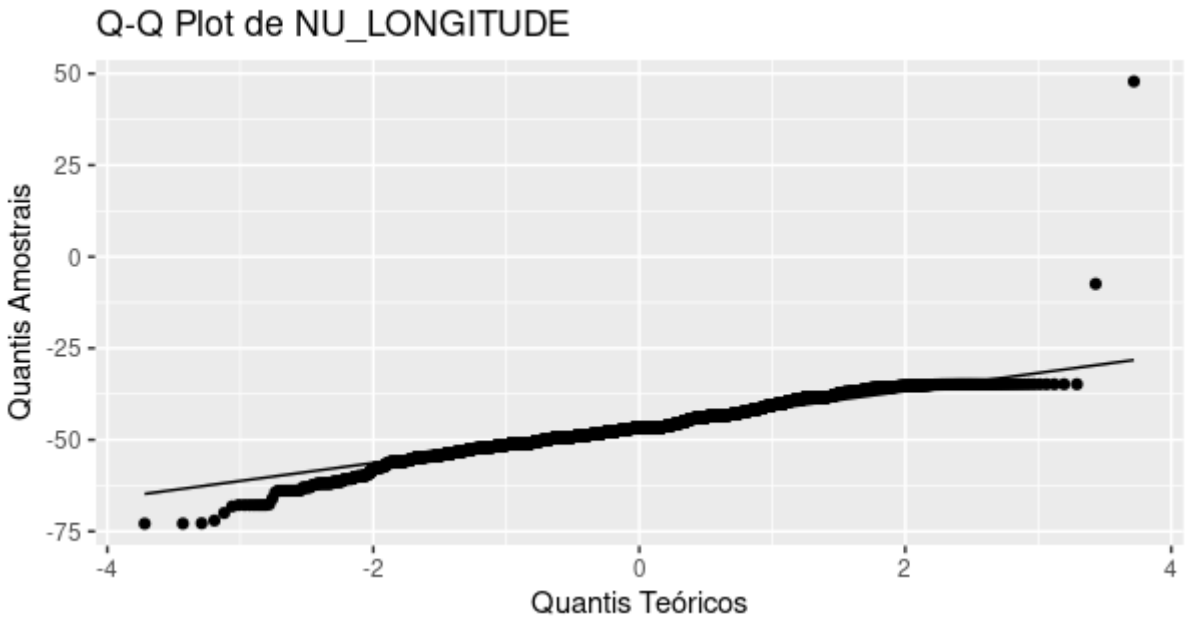
Calcular Normalidade da distribuição das longitudes

```
data_9 <- dados %>% dplyr::select(NU_LONGITUDE) %>% na.omit()
n <- nrow(data_9)
set.seed(123)
if (n > 5000) {
  data_9 <- data_9 %>% sample_n(5000)
}
# Teste de Shapiro-Wilk
shapiro_test_9 <- shapiro.test(data_9$NU_LONGITUDE)
# Mostrar os resultados do teste de Shapiro-Wilk
print(shapiro_test_9)
ggplot(data_9, aes(x = NU_LONGITUDE)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "skyblue", color = "black") +
  geom_density(color = "red", size = 1) +
  labs(title = "Histograma e Curva de Densidade de NU_LONGITUDE",
```

```
       x = "Longitude",
       y = "Densidade")

# Q-Q plot
ggplot(data_9, aes(sample = NU_LONGITUDE)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q Plot de NU_LONGITUDE",
       x = "Quantis Teóricos",
       y = "Quantis Amostrais")
```





Resultado: Os dados não são normalmente distribuídos, o gráfico quantil-quantil deixa isso claro apesar de uma parte os dados seguem uma ligeira distribuição normal, o p-value é < 2.2e-16, o que significa que é muito pequeno, isso sugere forte evidência contra a hipótese nula de normalidade, os dados NU_LONGITUDE não são normalmente distribuídos.

Conclusão

Concluí que este trabalho onde analisei os dados do Cadastro Nacional de Estabelecimentos de Saúde (CNES) utilizando técnicas de estatística descritiva, probabilidade e inferência estatística permitiram identificar padrões e tendências na distribuição e nas características dos estabelecimentos de saúde, fornecendo um conhecimento valioso.

Os resultados obtidos destacam a importância do uso de ferramentas estatísticas para interpretar os dados, o posit.clound na linguagem R foi fundamental na realização dessas análises. As perguntas formuladas à base de dados foram respondidas com clareza, evidenciando a relevância do CNES no monitoramento da infraestrutura de saúde no Brasil. Além disso, foi necessário criar nomes que fizessem sentido para os títulos e conteúdos das tabelas. Por exemplo, no TP_GESTAO, inicialmente não compreendia os significados de M, E, D e S, então criei possíveis interpretações que se mostraram adequadas.

Ao longo do desenvolvimento deste trabalho, enfrentei diversas dificuldades, por exemplo, tive pelo menos, mais de 300 linhas de comando com mensagens de erro, desde a manipulação inicial dos dados até a interpretação dos resultados obtidos. Cada desafio superado proporcionou um aprendizado significativo, aprimorando minhas habilidades na aplicação de técnicas estatísticas e no uso da Linguagem R. A experiência adquirida ao enfrentar e superar essas dificuldades foi essencial para alcançar os objetivos propostos e garantir a qualidade das análises apresentadas.

Portanto, a aplicação de métodos estatísticos robustos é crucial para a tomada de decisões na saúde pública ou em qualquer outra área de atuação, e este relatório contribui para a compreensão do estado dos estabelecimentos de saúde e das áreas que necessitam de maior atenção. A utilização de técnicas de análise de dados, aliada ao conhecimento adquirido ao longo do processo, demonstra o potencial de ferramentas estatísticas para apoiar a gestão e a melhoria dos serviços de saúde no Brasil.