



**CENTRO UNIVERSITÁRIO INSTITUTO DE
EDUCAÇÃO SUPERIOR DE BRASÍLIA**

**Bacharelado em
Ciência de Dados e Inteligência Artificial**

***Projeto Aplicado Integrador de Disciplinas –
PAID***

Jonathan Ferreira Costa

Marley Abe Silva

Maycon Moriy Abe Machado

Vinicius de Paula Ribeiro

***Implementação do
Banco de Dados das Eleições Gerais de 2022
e
Análise dos Resultados***

**Brasília
Dezembro de 2022**

Resumo

No projeto que se segue, foi desenvolvido uma análise descritiva dos resultados das eleições estaduais ordinárias do estado do Rio de Janeiro para os cargos de Presidente, Governador, Senador, Deputado Federal e Estadual, utilizando-se as ferramentas SAS e Python. Juntamente com a análise estatística também foi desenvolvido um banco de dados relacional, devidamente normalizado, por meio do Sistema de Gerenciamento de Banco de Dados (SGBD) PostgreSQL, que serviu como base para as análises dos dados coletados junto ao Tribunal Superior Eleitoral (TSE). O objetivo do projeto é proporcionar aos alunos a oportunidade de praticar os conhecimentos adquiridos durante as disciplinas de: Estatística Aplicada, Tecnologias de Big Data e Introdução a Programação, ampliando suas experiências práticas e seus portfólios para ingresso no mercado de trabalho/estágio.

Palavras-chaves: Ciência de Dados, SQL, Python, SAS Guide Enterprise, Estatística Descritiva, Estatística Univariada, Normalização, Banco de Dados.

ABSTRACT

In the project that follows, a descriptive analysis of the results of the ordinary state elections of the state of Rio de Janeiro for the positions of President, Governor, Senator, Federal and State Deputy, using the SAS and Python. Along with the statistical analysis, a database was also developed. relational, duly normalized, through the Database Management System Data (DBMS) PostgreSQL, which served as the basis for the analysis of the collected data before the Superior Electoral Court (TSE). The aim of the project is to provide students the opportunity to practice the knowledge acquired during the disciplines of: Applied Statistics, Big Data Technologies and Introduction to Programming, expanding their practical experiences and their portfolios for entering the job/internship market.

Keywords: Data Science, SQL, Python, SAS Guide Enterprise, Descriptive Statistics Descriptive Statistics, Univariate Statistics, Normalization, Database.

Lista de ilustrações

VOTACAO_SECAO (NR_ZONA, NR_SECAO, CD_ELEICAO, SQ_CADIDATO, NR_VOTAVEL, CD_CARGO, CD_MUNICIPIO, ANO_ELEICAO, CD_TIPO_ELEICAO, NM_TIPO_ELEICAO, NR_TURNO, DS_ELEICAO, DT_ELEICAO, TP_ABRANGENCIA, SG_UF, SG_UE, NM_UE, NM_MUNICIPIO, DS_CARGO, MN_NOTAVEL, QT_VOTOS, NR_LOCAL_VOTACAO, NM_LOCAL_VOTACAO, DS_LOCAL_VOTACAO_ENDERECO)

VOTACAO (NR_ZONA, NR_SECAO, CD_ELEICAO(fk), SQ_CADIDATO(fk), NR_VOTAVEL(fk), CD_CARGO(fk), CD_MUNICIPIO(fk), TP_ABRANGENCIA, SG_UF, SG_UE(fk), NM_UE, QT_VOTOS, NR_LOCAL_VOTACAO, NM_LOCAL_VOTACAO, DS_LOCAL_VOTACAO_ENDERECO)

ELEICAO (CD_ELEICAO, DS_ELEICAO, DT_ELEICAO, CD_TIPO_ELEICAO, NM_TIPO_ELEICAO, ANO_ELEICAO, NR_TURNO)

MUNICIPIO (CD_MUNICIPIO, NM_MUNICIPIO)

CARGO (CD_CARGO, DS_CARGO)

VOTAVEL (SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO(fk), SG_UE(fk), MN_NOTAVEL)

VOTACAO (NR_ZONA, NR_SECAO, CD_ELEICAO(fk), SQ_CANDIDATO(fk), NR_VOTAVEL(fk), CD_CARGO(fk), CD_MUNICIPIO(fk), TP_ABRANGENCIA, SG_UF, QT_VOTOS, SG_UE(fk), NR_LOCAL_VOTACAO, NM_LOCAL_VOTACAO, DS_LOCAL_VOTACAO_ENDERECO)

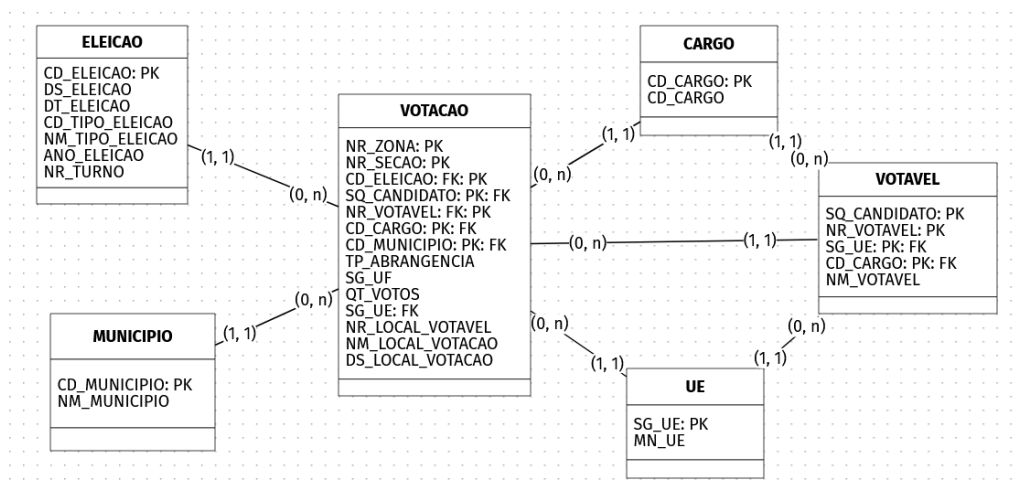
ELEICAO (CD_ELEICAO, DS_ELEICAO, DT_ELEICAO, CD_TIPO_ELEICAO, NM_TIPO_ELEICAO, ANO_ELEICAO, NR_TURNO)

MUNICIPIO (CD_MUNICIPIO, NM_MUNICIPIO)

CARGO (CD_CARGO, DS_CARGO)

VOTAVEL (SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO(fk), SG_UE(fk), MN_NOTAVEL)

UNIDADE ELEITORAL (SG_UE, NM_UE)



```
CREATE TABLE "cargo"
(
    CD_CARGO char(1) not null,
    DS_CARGO varchar(20) not null,

    constraint pk_cargo PRIMARY KEY (CD_CARGO)
);
```

```
CREATE TABLE "municipio"
(
    CD_MUNICIPIO varchar(5) not null,
    NM_MUNICIPIO varchar(32) not null,

    constraint pk_municipio PRIMARY KEY (CD_MUNICIPIO)
);
```

```
CREATE TABLE "eleicao"
(
    CD_ELEICAO varchar(3) not null,
    DS_ELEICAO varchar(30) not null,
    DT_ELEICAO date not null,
    CD_TIPO_ELEICAO varchar(1) not null,
    NM_TIPO_ELEICAO varchar(17) not null,
    ANO_ELEICAO varchar(4) not null,
    NR_TURNIO varchar(1) not null,

    constraint pk_eleicao PRIMARY KEY (CD_ELEICAO)
);
```

```
CREATE TABLE "ue_eleitoral"
(
    SG_UE varchar(2) not null,
    NM_UE varchar(14),

    constraint pk_ue_eleitoral PRIMARY KEY (SG_UE)
);
```

```

CREATE TABLE "votavel"
(
    SQ_CANDIDATO varchar(12) not null,
    NR_VOTAVEL varchar(5) not null,
    MN_NOTAVEL varchar(51) not null,
    CD_CARGO char(1) not null,
    SG_UE varchar(2) not null,

    constraint PK_VOTAVEL PRIMARY KEY (SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, SG_UE)
    CONSTRAINT FK_CARGO FOREIGN KEY(CD_CARGO) REFERENCES cargo(CD_CARGO)
    CONSTRAINT FK_UE FOREIGN KEY(SG_UE) REFERENCES ue_eleitoral(SG_UE)
);

```

```

CREATE TABLE "votacao"
(
    NR_ZONA varchar(3) not null,
    NR_SECAO varchar(4) not null,
    CD_ELEICAO varchar(3) not null,
    SQ_CANDIDATO varchar(12) not null,
    NR_VOTAVEL varchar(5) not null,
    CD_CARGO varchar(1) not null,
    CD_MUNICIPIO varchar(5) not null,
    TP_ABRANGENCIA varchar(100) not null,
    SG_UF varchar(2) not null,
    QT_VOTOS int not null,
    SG_UE varchar(2) not null,
    NR_LOCAL_VOTACAO varchar(4) not null,
    NM_LOCAL_VOTACAO varchar(100) not null,
    DS_LOCAL_VOTACAO_ENDERECO varchar(100) not null,

    constraint pk_votacao PRIMARY KEY
(NR_ZONA, NR_SECAO, CD_ELEICAO, SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, CD_MUNICIPIO),
    constraint FK_ELEICAO FOREIGN KEY(CD_ELEICAO) REFERENCES eleicao(CD_ELEICAO),
    constraint FK_VOTAVEL FOREIGN KEY(SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, SG_UE)
REFERENCES votavel(SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, SG_UE),
    constraint FK_CARGO FOREIGN KEY(CD_CARGO) REFERENCES cargo(CD_CARGO),
    constraint FK_MUNICIPIO FOREIGN KEY(CD_MUNICIPIO)
REFERENCES municipio(CD_MUNICIPIO),
    constraint FK_UE FOREIGN KEY(SG_UE) REFERENCES ue_eleitoral(SG_UE)
);

```

```

eleicoes_df = pd.read_csv("votacao_secao_2022_RJ.csv", sep=';', encoding='cp1252')

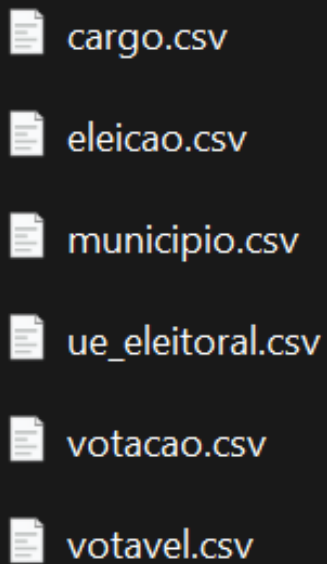
```

```

votacao = eleicoes_df[[Attr_votacao]]
eleicao = eleicoes_df[[Attr_eleicao]].drop_duplicates()
municipio = eleicoes_df[[Attr_municipio]].drop_duplicates()
cargo = eleicoes_df[[Attr_cargo]].drop_duplicates()
votavel = eleicoes_df[[Attr_votavel]].drop_duplicates()
ue_eleitoral = eleicoes_df[[Attr_ue_eleitoral]].drop_duplicates()

```

```
for df in ['votacao', 'eleicao', 'municipio', 'cargo', 'votavel', 'ue_eleitoral']:  
    eval(df+'.to_csv("C:/Users/marle/Desktop/eleicoes/tabelas_normalizadas/'+df+'.csv", \\  
        header=True, index=False, sep=";")')
```



cargo.csv

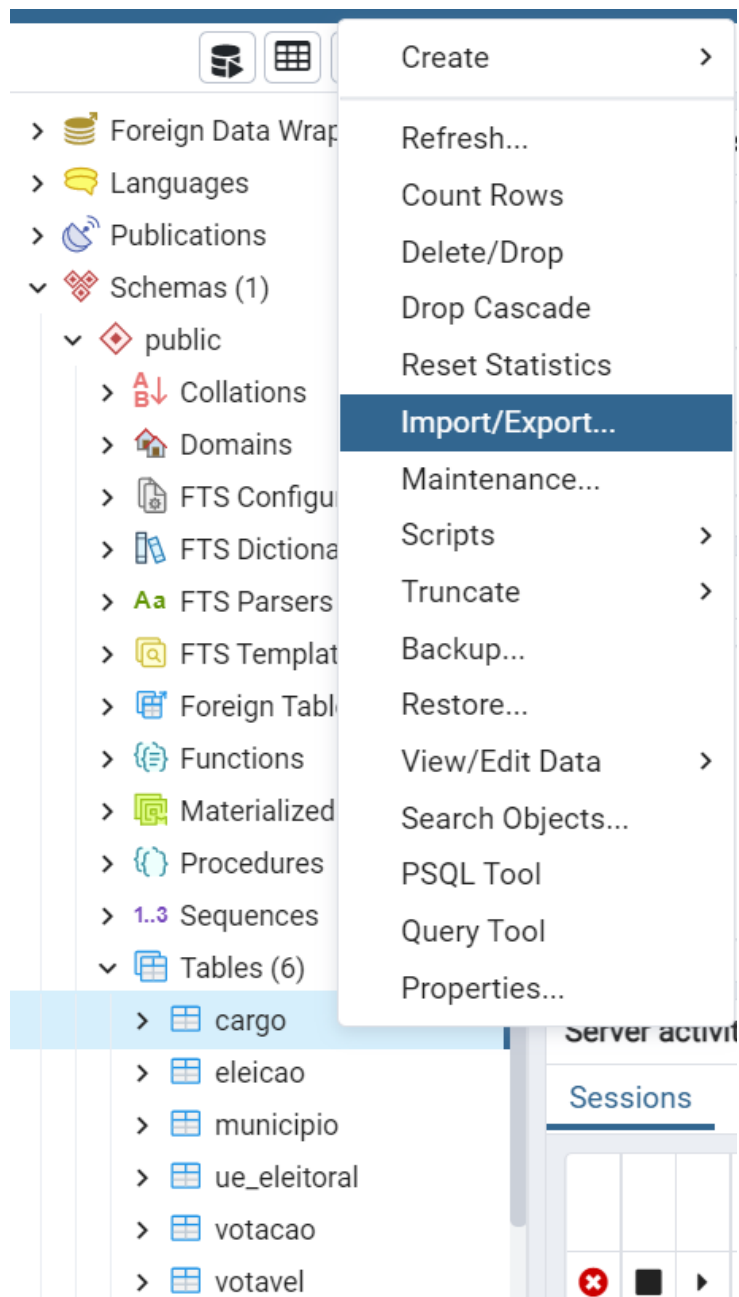
eleicao.csv

municipio.csv

ue_eleitoral.csv

votacao.csv

votavel.csv



Import/Export data - table 'cargo'

OptionsColumns

Import/Export

Import

File Info

Filename

C:\Users\marle\Desktop\eleicoes\tabelas_normalizadas\cargo.csv

...

Format

csv

▼

Encoding

UTF8

×

▼

Miscellaneous

OID

No

Header

Yes

Delimiter

;

▼

Specifies the character that separates columns within each row (line) of the file. The default is a tab character in text format, a comma in CSV format. This must be a single one-byte character. This option is not allowed when using binary format.

✕ Cancel

✓ OK

Query Editor

Query History

Scratch Pad

1

select * from cargo

Data Output

Explain

Messages

Notific

	cd_cargo	ds_cargo
	[PK] character (1)	character varying (20)
1	1	PRESIDENTE
2	6	DEPUTADO FEDERAL
3	7	DEPUTADO ESTADUAL
4	3	GOVERNADOR
5	5	SENADOR

Variable Number	Name	Type	Format	Label	Length
1	CD_CARGO	Character	\$CHAR		1
2	DS_CARGO	Character	\$CHAR		17

Variable Number	Name	Type	Format	Label	Length
1	CD_MUNICIPIO	Character	\$CHAR		5
2	NM_MUNICIPIO	Character	\$CHAR		31

Variable Number	Name	Type	Format	Label	Length
1	CD_ELEICAO	Character	\$CHAR		3
2	DS_ELEICAO	Character	\$CHAR		32
3	DT_ELEICAO	Numeric	DATE		8
4	CD_TIPO_ELEICAO	Character	\$CHAR		1
5	NM_TIPO_ELEICAO	Character	\$CHAR		20
6	ANO_ELEICAO	Numeric	BEST		8
7	NR_TURNO	Numeric	BEST		8

Variable Number	Name	Type	Format	Label	Length
1	SG_UE	Character	\$CHAR		2
2	NM_UE	Character	\$CHAR		14

Variable Number	Name	Type	Format	Label	Length
1	SQ_CANDIDATO	Character	\$CHAR		12
2	NR_VOTAVEL	Character	\$CHAR		5
3	NM_VOTAVEL	Character	\$CHAR		51
4	CD_CARGO	Character	\$CHAR		1
5	SG_UE	Character	\$CHAR		2

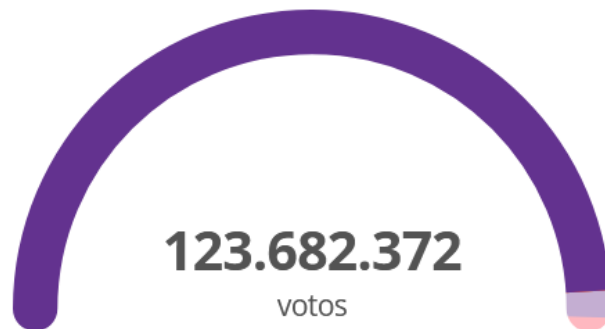
Variable Number	Name	Type	Format	Label	Length
1	SQ_CANDIDATO	Character	\$CHAR		12
2	NR_VOTAVEL	Character	\$CHAR		5
3	NM_VOTAVEL	Character	\$CHAR		51
4	CD_CARGO	Character	\$CHAR		1
5	SG_UE	Character	\$CHAR		2

```
3 select sum(qt_votos) from votacao where(cd_eleicao = '544')
```

Data Output Explain Messages Notifications

	sum bigint	
1	123682372	

Votação



Votos a candidatos concorrentes · 95,59%

■ 118.229.719
Votos Válidos

■ 0
Anulados

■ 0
Anulados Sub
Judice

■ 3.487.874 ·
2,82%
Nulos

■ 1.964.779 ·
1,59%
Em Branco

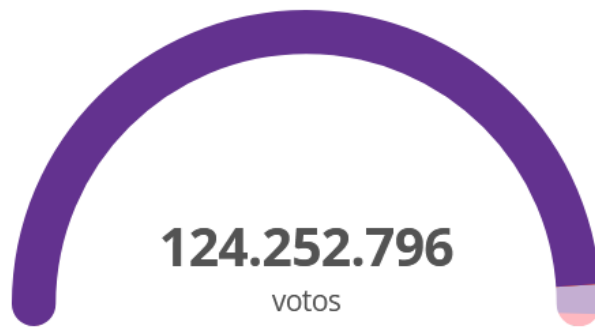
■ 0 · 0,00%
Anulados e
apurados em
separado

```
3 select sum(qt_votos) from votacao where(cd_eleicao = '545')
```

Data Output Explain Messages Notifications

	sum bigint	
1	124252796	

Votação



Votos a candidatos concorrentes · 95,41%

118.552.353
Votos Válidos

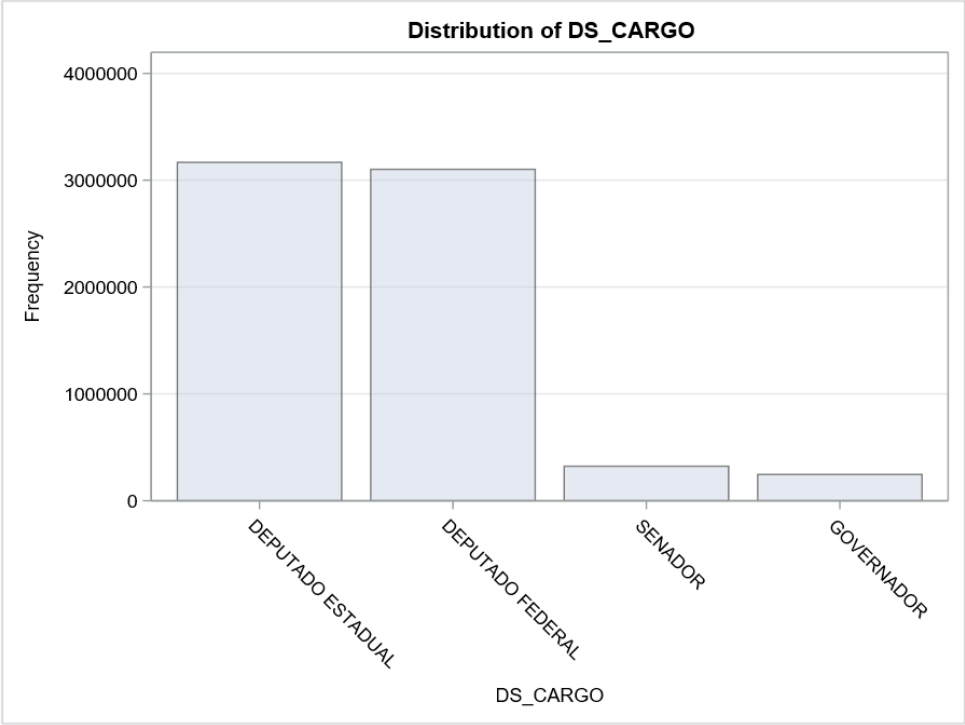
0
Anulados

0
Anulados Sub
Judice

3.930.765 ·
3,16%
Nulos

1.769.678 ·
1,43%
Em Branco

0 · 0,00%
Anulados e
apurados em
separado



Analysis Variable : QT_VOTOS												
DS_CARGO	N Obs	Mean	Std Dev	Minimum	Maximum	Range	N	Lower Quartile	Median	Upper Quartile	90th Pctl	95th Pctl
SENADOR	323116	30.6195236	27.1256294	1.0000000	266.0000000	265.0000000	323116	10.0000000	25.0000000	43.0000000	68.0000000	84.0000000
GOVERNADOR	246574	40.1244981	50.5043772	1.0000000	366.0000000	365.0000000	246574	8.0000000	19.0000000	48.0000000	129.0000000	161.0000000
DEPUTADO FEDERAL	3101337	3.1901267	5.9810683	1.0000000	206.0000000	205.0000000	3101337	1.0000000	1.0000000	3.0000000	6.0000000	11.0000000
DEPUTADO ESTADUAL	3167243	3.1237445	7.2460267	1.0000000	246.0000000	245.0000000	3167243	1.0000000	1.0000000	2.0000000	6.0000000	11.0000000

Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0000	Observations	1642
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 20:26:31	Observation Length	88
Last Modified	04/12/2022 20:26:31	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0000		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

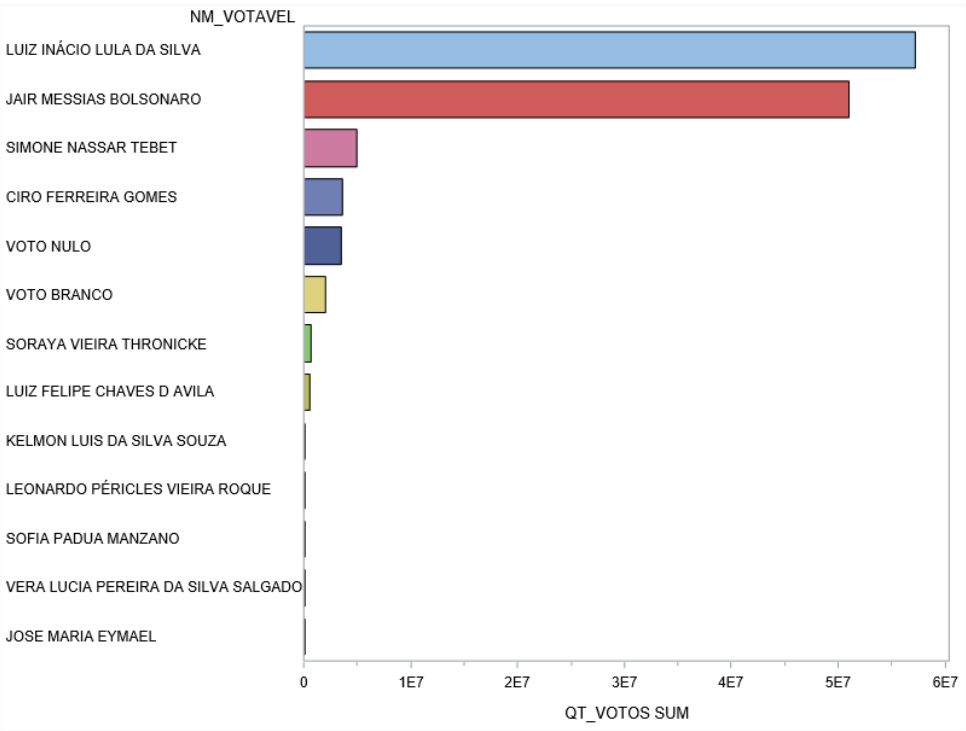
Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0001	Observations	1086
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 20:45:44	Observation Length	88
Last Modified	04/12/2022 20:45:44	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0000		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

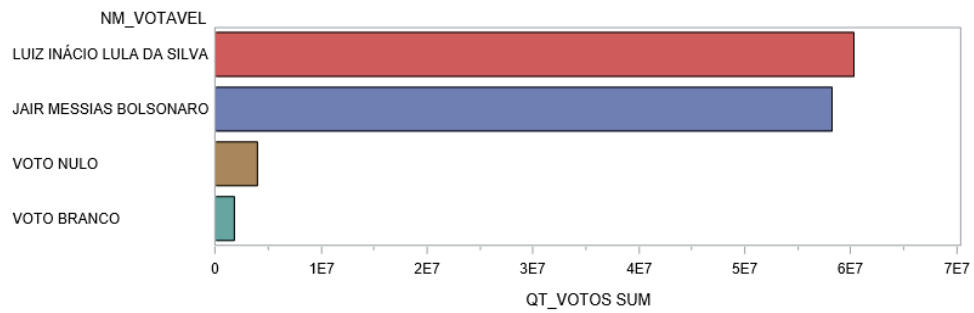
Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0002	Observations	15
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 20:51:06	Observation Length	88
Last Modified	04/12/2022 20:51:06	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0001		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0003	Observations	11
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 21:01:08	Observation Length	88
Last Modified	04/12/2022 21:01:08	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0002		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

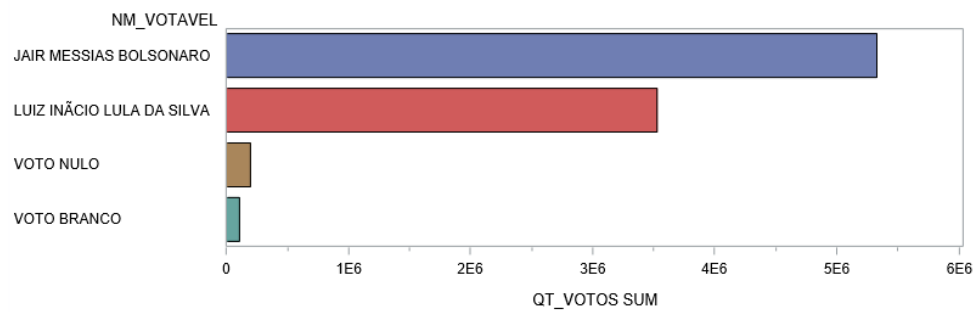
Analysis Variable : COUNT Frequency Count				
Mean	Minimum	Maximum	Sum	N
206.4727273	36.0000000	417.0000000	34068.00	165

	NM_VOTAVEL	QT_VOTOS_Sum
1	VOTO BRANCO	6E5
2	VOTO NULO	6E5
3	MARCIO CORREIA DE OLIVEIRA	2E5
4	DOUGLAS RUAS DOS SANTOS	2E5
5	RENATA DA SILVA SOUZA	2E5
6	Partido Liberal	1E5
7	ROSENVERG REIS DE OLIVEIRA	1E5
8	Partido dos Trabalhadores	1E5
9	SÉRGIO LUIZ COSTA AZEVEDO FILHO	1E5
10	GUILHERME JANDRE DELAROLI	1E5

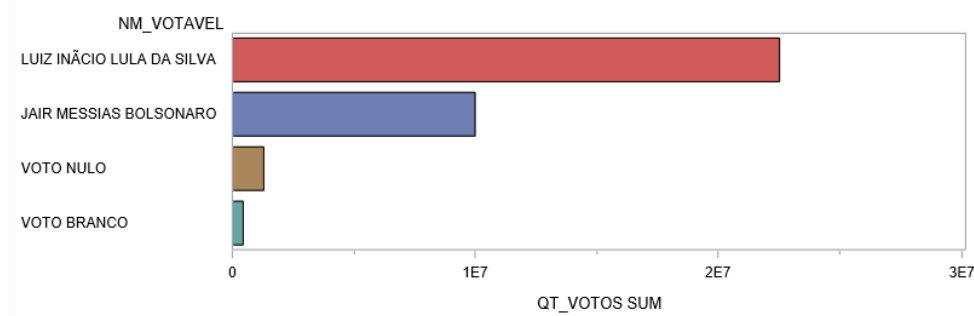


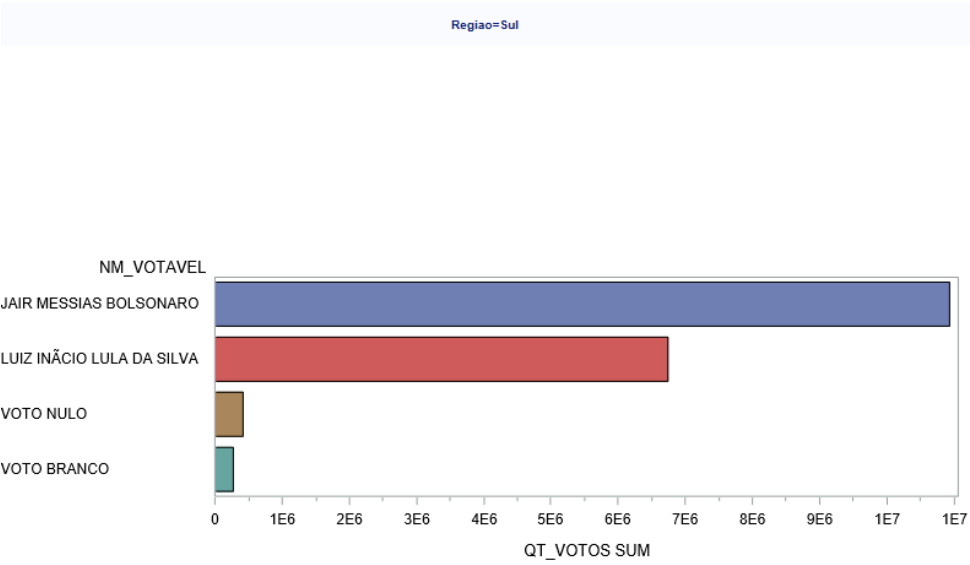
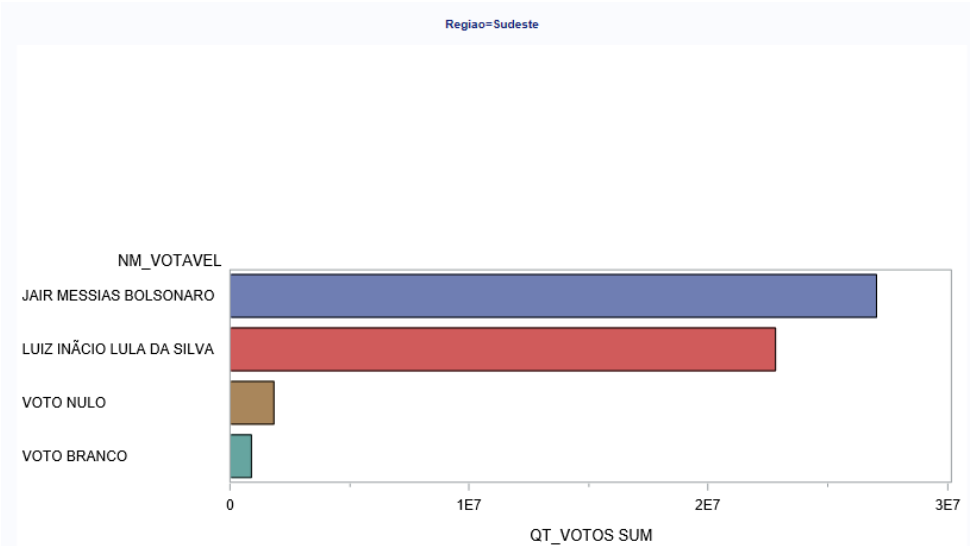
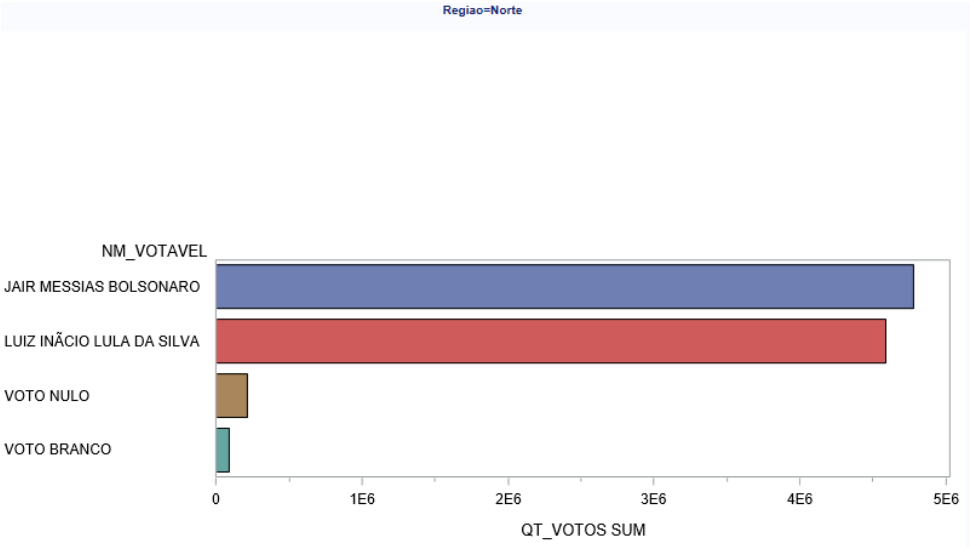


Regiao=Centro-Oeste

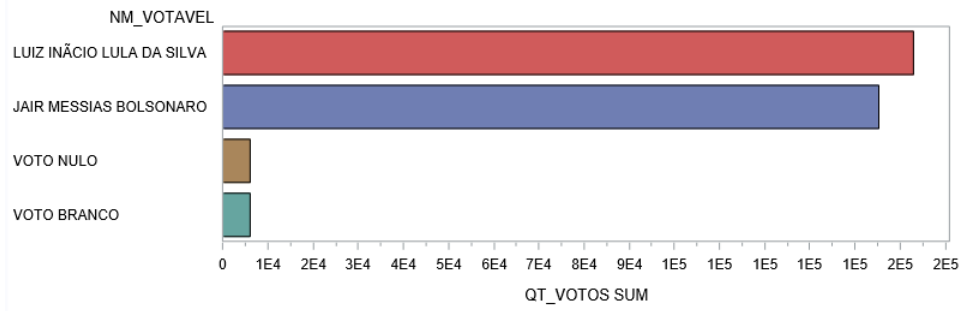


Regiao=Nordeste





Regiao=Exterior



Lista de tabelas

DS_CARGO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DEPUTADO ESTADUAL	3167243	46.32	3167243	46.32
DEPUTADO FEDERAL	3101337	45.35	6268580	91.67
GOVERNADOR	246574	3.61	6515154	95.27
SENADOR	323116	4.73	6838270	100.00

Analysis Variable : QT_VOTOS	
DS_CARGO	Sum
DEPUTADO ESTADUAL	9893658.00
DEPUTADO FEDERAL	9893658.00
GOVERNADOR	9893658.00
SENADOR	9893658.00

Analysis Variable : QT_VOTOS				
DS_CARGO	NM_VOTAVEL	N Obs	Sum	N
DEPUTADO ESTADUAL	VOTO BRANCO	34061	644435.00	34061
	VOTO NULO	34052	602675.00	34052
DEPUTADO FEDERAL	VOTO BRANCO	34060	616380.00	34060
	VOTO NULO	34056	600046.00	34056
GOVERNADOR	VOTO BRANCO	34058	591576.00	34058
	VOTO NULO	34068	894384.00	34068
SENADOR	VOTO BRANCO	34066	752864.00	34066
	VOTO NULO	34067	972167.00	34067

	A	B
1	NM_VOTAVEL	QT_VOTOS_Sum
2	VOTO BRANCO	644435
3	VOTO NULO	602675
4	MARCIO CORREIA DE OLIVEIRA	181274
5	DOUGLAS RUAS DOS SANTOS	175977
6	RENATA DA SILVA SOUZA	174132
7	Partido Liberal	144794
8	ROSENVERG REIS DE OLIVEIRA	131308
9	Partido dos Trabalhadores	126155
10	SÉRGIO LUIZ COSTA AZEVEDO FILHO	123739
11	GUILHERME JANDRE DELAROLI	114155

1	NM_VOTAVEL	QT_VOTOS_Sum
2	VOTO BRANCO	616380
3	VOTO NULO	600046
4	DANIELA MOTÉ DE SOUZA CARNEIRO	213432
5	EDUARDO PAZUELLO	205324
6	TALÍRIA PETRONE SOARES	198548
7	LUIZ ANTONIO DE SOUZA TEIXEIRA JUNIOR	190071
8	ALTINEU CORTES FREITAS COUTINHO	167512
9	TARCÍSIO MOTTA DE CARVALHO	159928

Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
CLÁUDIO BOMFIM DE CASTRO E SILVA	4930288.00
CYRO GARCIA	12627.00
EDUARDO GONÇALVES SERRA	10852.00
JULIETE PANTOJA ALVES	27344.00
LUIZ EUGÊNIO HONORATO	1844.00
MARCELO RIBEIRO FREIXO	2300980.00
PAULO GUSTAVO GANIME ALVES TEIXEIRA	447051.00
RODRIGO NEVES BARRETO	672291.00
VOTO BRANCO	591576.00
VOTO NULO	894384.00
WILSON JOSÉ WITZEL	4892.00

Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
ALESSANDRO LUCCIOLA MOLON	1731786.00
ANDRÉ LUIZ CECILIANO	986676.00
ANTONIO HERMANO LEMME	1198.00
BENEVENUTO DACIOLO FONSECA DOS SANTOS	285037.00
BÁRBARA DEL PENHO SINEDINO PINHEIRO	18222.00
CLARISSA GAROTINHO BARROS ASSED MATHEUS DE OLIVEIRA	1145413.00
DANIEL LUCIO DA SILVEIRA	1566352.00
HELVIO COSTA DE OLIVEIRA TELLES	7036.00
HIRAN ROEDEL	5120.00
MARCELO ZATURANSKY NOGUEIRA ITAGIBA	18224.00
RAUL BITTENCOURT PEDREIRA	7299.00
ROMÁRIO DE SOUZA FARIA	2385181.00
SUED HAIDAR NOGUEIRA	11933.00
VOTO BRANCO	752864.00
VOTO NULO	972167.00

Regiao=Nordeste

Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
LUIZ INÁCIO LULA DA SILVA	22534987.00
JAIR MESSIAS BOLSONARO	9982947.00
VOTO NULO	1275547.00
VOTO BRANCO	415203.00

Regiao=Sul

Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
LUIZ INACIO LULA DA SILVA	6750374.00
JAIR MESSIAS BOLSONARO	10940158.00
VOTO NULO	409071.00
VOTO BRANCO	274488.00

Analysis Variable : QT_VOTOS	
Regiao	Sum
Sudeste	52542866.00
Nordeste	34188864.00
Sul	18374089.00
Norte	9675082.00
Centro-Oeste	9161947.00
Exterior	310148.00

Lista de abreviaturas e siglas

CD	Ciência de Dados (<i>Data Science</i>)
TSE	Tribunal Superior Eleitoral
CSV	(<i>Comma Separated Values</i>)
SQL	(<i>Structured Query Language</i>)
DataFrame	(<i>Data Structured</i>)
SGBD	(<i>Sistema de Gerenciamento de Bando de Dados</i>)
postgres	(<i>SGBD</i>)
pgAdmin	(<i>Software to use postgres</i>)
SAS Enterprise Guide	(<i>Software para análise estatística</i>)

Sumário

1	INTRODUÇÃO	27
2	MOTIVAÇÃO	29
3	OBJETIVOS E ORGANIZAÇÃO DO TRABALHO	31
3.1	Objetivos	31
3.2	Organização do Trabalho	32
3.3	Cronograma do Trabalho	33
4	CONSTRUÇÃO DO BANCO DE DADOS	34
4.1	Fontes de Dados	34
4.1.1	Definição das Fontes de Dados	34
4.1.2	Descrição dos Dados nas Fonte de Dados	35
4.2	Modelagem Conceitual, Lógica e Física do Banco de Dados	37
4.2.1	Modelo de Entidades e Relacionamentos	37
4.2.2	Normalização das Entidades	38
4.2.3	Modelo Lógico do Banco de Dados	40
4.2.4	SQL de Criação do Banco de Dados	40
4.2.5	Descrição da Carga dos Dados no Banco de Dados	43
4.2.6	Dicionário de Dados	47
4.2.7	Resultados da Carga no Banco de Dados	48
5	ANÁLISE DOS DADOS DAS ELEIÇÕES	51
5.1	Estatísticas Descritivas	51
5.1.1	Eleição Governador, Senador, Deputado Federal e Deputado Estadual	51
5.1.2	Eleição Presidencial	60
6	CONCLUSÕES	67
	REFERÊNCIAS	69
	ANEXOS	71
	ANEXO A – DICIONÁRIO DE DADOS	72
	ANEXO B – CÓDIGOS DOS PROGRAMAS/NOTEBOOKS PYTHON	73

1 INTRODUÇÃO

A **Ciência de Dados** é uma área nova de ciência, multidisciplinar, que se baseia na integração das áreas de matemática, estatística e computação. Em termos simples, resume-se na exploração e análise de dados visando à extração de informações e conhecimento a partir dos dados. A Ciência dos Dados em muitos aspectos é uma consequência da necessidade de analisar grandes bancos de dados (Big Data), até pouco tempo conhecidos como Very Large Database (VLDB), que possuem características de grandes volumes de dados, alta velocidade de geração e armazenamento de novos dados, variedade de tipos de dados, alta qualidade nos dados e o processamento de seus dados agregam valor nas organizações. A Ciência de Dados é um campo interdisciplinar que exige a formação de um novo profissional que possua as habilidades e competências no uso da matemática, da estatística e da computação, aplicados na extração de informações e conhecimentos dos dados ([HAN; KAMBER; PEI, 2012](#)).

A **Inteligência Artificial (IA)** é uma tecnologia que usa um mecanismo de processamento que simula a inteligência humana. As vezes chamada de inteligência de máquina, é a inteligência demonstrada por máquinas, em contraste com a inteligência natural exibida por humanos. Algumas das atividades para as quais foi projetado são reconhecimento de fala, aprendizagem, reconhecimento de imagens, processamento de linguagem natural, planejamento e solução de problemas, entre tantas outras atividades. Visto que a Robótica é o campo relacionado com a conexão da percepção com a ação, a Inteligência Artificial deve ter um papel central na Robótica se a conexão deve ser inteligente. A Inteligência Artificial aborda as questões cruciais de: qual conhecimento é necessário em qualquer aspecto do pensamento; como esse conhecimento deve ser representado; e como esse conhecimento deve ser usado. A robótica desafia a Inteligência Artificial, forçando-a a lidar com objetos reais no mundo real ([LEVINE; STEPHAN; SZABAT, 2016](#)).

O **Aprendizado de Máquina** tornou-se um dos tópicos mais importantes dentro das organizações de desenvolvimento que buscam maneiras inovadoras de aproveitar ativos de dados para ajudar a empresa a obter um novo nível de compreensão. Com os modelos de aprendizado de máquina apropriados, as organizações têm a capacidade de prever continuamente as mudanças nos negócios para que possam prever o que está por vir. Como os dados são adicionados constantemente, os modelos de aprendizado de máquina garantem que a solução seja atualizada constantemente. O valor é direto: se você usar as fontes de dados mais adequadas e em constante mudança no contexto do aprendizado de máquina, terá a oportunidade de prever o futuro ([GRUS, 2016](#)).

O **Aprendizado de Máquina** é uma forma de IA que permite que um sistema aprenda com os dados em vez de por meio de programação explícita. No entanto, o

aprendizado de máquina não é um processo simples. Ele usa uma variedade de algoritmos que aprendem iterativamente com os dados para melhorar, descrever os dados e prever resultados. À medida que os algoritmos ingerem dados de treinamento, é possível produzir modelos mais precisados com base nesses dados. Um modelo de aprendizado de máquina é a saída gerada quando você treina seu algoritmo de aprendizado de máquina com dados. Após o treinamento, ao fornecer uma entrada a um modelo, você receberá uma saída. Por exemplo, um algoritmo preditivo criará um modelo preditivo. Então, ao fornecer dados ao modelo preditivo, você receberá uma previsão com base nos dados que treinaram o modelo. O aprendizado de máquina agora é essencial para criar modelos analíticos.

2 MOTIVAÇÃO

Eleição é todo processo pelo qual uma sociedade seleciona um ou mais de um de seus cidadãos para ocupar um cargo por meio de votação. Na Democracia Representativa brasileira, ela é o processo que consiste na escolha dos indivíduos que irão exercer o poder Executivo e Legislativo, tanto do Governo Federal, tanto dos Governos Estaduais e Municipais concedido pelo povo através do voto, devendo estes, assim, exercerem o seu devido papel pré-estabelecido na constituição brasileira de 1988.(Link de Referência: 1)

Apenas os brasileiros naturalizados podem participar das eleições que são obrigatórias para os maiores de 18 anos e menores de 70 anos e facultativa para os analfabetos, para os maiores de 16 anos e menores de 18 anos e maiores de 70 anos. As eleições no Brasil são realizadas sempre em anos pares. Os mandatos de vereadores, prefeitos, deputados estaduais, federais, governadores e do presidente da República duram quatro anos; o dos senadores duram oito anos. A legislação brasileira determina que todas as eleições ocorram no primeiro domingo de outubro dos anos em que serão realizadas, no horário das 8 horas até as 17 horas.

No Código Eleitoral Brasileiro, de 1965, são definidos três sistemas eleitorais distintos: as eleições proporcionais para a Câmara dos Deputados, tanto das esferas estadual e municipal, onde prevalece o chamado sistema de lista aberta, no qual a proporção de cadeiras parlamentares ocupada por cada partido é diretamente determinada pela proporção de votos obtida por ele, assim, os eleitores votam em partidos e na ordem dos candidatos numa lista pré determinada por esses partidos.

Para o cálculo do número de vagas de cada partido utiliza-se um método conhecido como quociente eleitoral, definido como o total de votos válidos dividido pelo número de vagas, e outro conhecido como distribuição das sobras para ocupar as cadeiras não preenchidas pelo quociente eleitoral. (Link de Referência: 2)

Os outros sistemas são as eleições majoritárias com um ou dois eleitos para o Senado Federal e eleições majoritárias em dois turnos para presidente e demais chefes do executivo nas outras esferas. Nessas eleições prevalecem o sistema de maioria absoluta, onde o eleito precisa obter mais de 50% dos votos válidos, desconsiderados os brancos e nulos, para ser eleito.(Links de Referência: 3, 4 e 5)

Pode-se perceber que a utilização de técnicas estatísticas e matemáticas estão presentes de forma bastante concisa no processo eleitoral brasileiro , visto que as eleições de 2022, segundo o presidente do Tribunal Superior Eleitoral (TSE), ministro Alexandre de Moraes, foi registrado o maior número de votos em candidatos da história brasileira, tanto percentualmente quanto em termos absolutos. (Link de Referência: 6) Com isso,

é de extrema valia a elaboração de qualquer projeto que ajude a sociedade brasileira a entender e compreender melhor o resultados das eleições de 2022, vista a relevância que essa eleição tivera em comparação as eleições passadas.

3 Objetivos e Organização do Trabalho

Neste capítulo são apresentados o objetivo geral do projeto, assim como os seus respectivos objetivos específicos e hipóteses que irão ajudar a complementar as análises dos resultados das eleições, assim como as etapas e datas que foram cumpridas para o desenvolvimento final do projeto.

3.1 Objetivos

O objetivo geral do Projeto é implementar um banco de dados relacional com os dados das eleições de 2022 do TSE do estado do Rio de Janeiro e descrever, por meio de técnicas de estatística descritiva, os resultados encontrados neste estado específico, afim de compreender melhor as características únicas dos sistemas eleitorais utilizados para eleger os candidatos aos poderes executivo e legislativo no Brasil e também identificar possíveis preferências do eleitorado carioca quanto a partidos ou candidatos. Quanto aos objetivos específicos presentes no trabalho:

- Implementar um banco de dados relacional com todos os dados necessários para as análises descritivas dos resultados, com todos os requerimentos técnicos aprendidos em sala de aula.

- Identificar nas bases estaduais e municipais a quantidade necessária de votos para que um partido possa indicar candidatos a deputados e identificar os partidos ou coligações que mais tiveram votos e calcular quantos candidatos cada um desses partidos tiveram direito de indicar e comparar com quantos cada um dos partidos mais votados realmente indicaram para os cargos, segundo o TSE.

- Identificar nas bases estaduais os candidatos ao Senado mais e menos votados e os candidatos a Governador mais e menos votados e comparar com os candidatos realmente nomeados para os cargos, segundo o TSE.

- Identificar nas bases nacionais os candidatos a presidência mais e menos votados e comparar com o candidato vencedor, segundo o TSE.

Para chegar ao objetivo final de forma mais concisa, antes do início das análises estatísticas, utilizando-se das informações pré-apuradas a respeito dos sistemas eleitorais brasileiros, foram desenvolvidas as seguintes hipóteses que serão avaliadas após a análise dos resultados:

- Os Deputados Federais e Estaduais indicados pelos partidos estão entre os deputados mais votados em comparação a todos os deputados? E em comparação aos deputados dos mesmos partidos?

- Os partidos que mais tiveram votos para Deputados Federais e Estaduais também tiveram mais votos para Senador, Governador e Presidente?
- O Presidente, Governador e Senador mais votados são do mesmo partido? Existem municípios ou zonas eleitorais onde isso acontece?

3.2 Organização do Trabalho

Para a implementação da base de dados, foram definidas as seguintes etapas:

Etapa 1: Coletar os dados relacionados a eleição estadual e Presidencial do Rio de Janeiro junto ao TSE.

Etapa 2: Construir um modelo conceitual, por meio do Diagrama de Entidade e Relacionamento (DER), para enxergar de forma clara as entidades, os atributos e os relacionamentos que estarão presentes dentro das bases de dados coletadas.

Etapa 3: Desenvolver uma lista de entidades do DER desenvolvido na etapa 1 e aplicar as regras de normalização (primeira, segunda e terceira) nas bases de dados coletadas e, se necessário, gerar novas entidades e/ou relacionamentos e desenvolver novamente o Diagrama de Entidades e Relacionamentos, após a normalização.

Etapa 4: Desenvolver o modelo lógico do banco de dados com todas as entidades e atributos devidamente normalizadas indicando quais serão as chaves primárias, estrangeiras e possíveis chaves alternadas do banco de dados.

Etapa 5: Implementar a base de dados utilizando o SGBD Postgres e a linguagem de consulta SQL.

Para as análises descritivas dos resultados será utilizado o software SAS, onde serão desenvolvidas as seguintes etapas do projeto:

Etapa 6: Selecionar os dados onde estão presentes os votos apenas para os cargos de Deputados Estaduais e Federais e analisar a distribuição de frequências das variáveis qualitativas presentes e as medidas-resumo das variáveis quantitativas.

Etapa 7: Após uma primeira análise descritiva, descobrir a quantidade de votos necessários para que cada partido tenha direito de indicar um candidato (quociente eleitoral) dividindo a quantidade total de votos pela quantidade total de cargos para as vagas de Deputado Estadual e Federal do Estado do Rio de Janeiro.

Etapa 8: Descobrir quantos candidatos cada partido terá direito de indicar para esses cargos somando os votos para os candidatos a Deputado Federal e Estadual de cada partido, e dividindo-os pelo quociente eleitoral.

Etapa 9: Analisar por meio de medidas-resumo a quantidade de votos que tiveram cada um dos candidatos a Senador, Governador e Presidente da República no Brasil.

3.3 Cronograma do Trabalho

A seguir segue o cronograma para o desenvolvimento deste trabalho:

Etapa 1 até Etapa 3; Data de início: 01/12/2022 Data de termino: 08/12/2022

Etapa 4 e Etapa 5; Data de início: 08/12/2022 Data de termino: 12/12/2022

Etapa 6 e Etapa 7; Data de início: 08/12/2022 Data de termino: 12/12/2022

Etapa 8 e Etapa 9; Data de início: 12/12/2022 Data de termino: 15/12/2022

4 Construção do Banco de Dados

Neste capítulo será apresentada as etapas de construção do Banco de Dados das Eleições Gerais de 2022.

4.1 Fontes de Dados

Criada em 24 de fevereiro de 1932, como uma das consequências da revolução de 30, o Código Eleitoral brasileiro representa a regulamentação da Justiça Eleitoral do Brasil, está determinada nos artigos 118 a 121 da Constituição Federal de 1988 que é de competência apenas da União legislar sobre o Direito Eleitoral e, ainda, que: "disporá sobre a organização e competência dos tribunais, dos juízes de direito e das juntas eleitorais."

Esta é apenas uma das muitas normas que concedem ao TSE poderes característicos do Poder Executivo e Poder Legislativo.

São funções da Justiça Eleitoral do Brasil:

A regulamentação do processo eleitoral

A administração completa de todo o processo eleitoral

A vigilância para o fiel cumprimento das normas jurídicas que regem o período eleitoral

Expedir instruções para execução da lei eleitoral;

Responder consultas sobre matéria eleitoral;

Com isso, o TSE é o único Órgão Federal responsável pela coleta, organização, registro, análise e divulgação tanto dos dados quanto dos resultados de todas as eleições desde 1932, e por isso será a única fonte de dados utilizada nesse trabalho.

4.1.1 Definição das Fontes de Dados

O TSE disponibiliza de forma gratuita e acessível acesso aos dados de todas as eleições desde o ano de 1933, de forma digital em formato CSV. O portal de dados abertos do TSE (link: <https://dadosabertos.tse.jus.br/>) veio para substituir o antigo Repositório de Dados Eleitorais, descontinuado em janeiro de 2022. Este portal disponibiliza à sociedade os dados gerados ou custodiados pelo TSE, de forma a garantir o acesso a informações e aprimorar a cultura de transparência. Os dados aqui disponíveis podem ser livremente acessados, utilizados, modificados e compartilhados por qualquer pessoa. Para utilizar-se do portal, é necessário primeiramente entrar na página inicial do site e encontrar uma caixa de busca para pesquisar pelo conjunto de dados desejado. É possível clicar

diretamente no menu Grupos, para verificar os temas disponíveis, ou no menu Conjuntos de dados para ir direto à relação completa dos conjuntos de dados. Nessa relação, cada conjunto de dados é exibido com o seu nome, um extrato de sua descrição e os formatos disponibilizados. Nessa página você também poderá aplicar os seguintes filtros disponíveis no lado esquerdo da tela: Organização, Tema, Etiqueta, Formato e Licença.

Os dados catalogados no portal estão organizados utilizando-se estruturas de conjuntos de dados e recursos. Os conjuntos de dados são os elementos principais retornados a partir das buscas. Cada conjunto de dados possui uma descrição, um ou mais recursos, e uma série de outros metadados, como periodicidade de atualização e unidade gestora. Para o TSE, são exemplos de conjuntos: coleções de tabelas relacionadas entre si, dados extraídos de um mesmo sistema de informações, ou ainda uma API de dados abertos. Um conjunto de dados deve possuir, pelo menos, um recurso que seja dado aberto. Os conjuntos de dados agrupam recursos oriundos da mesma base de dados ou que possuam metadados em comum, facilitando a busca e o entendimento de seu conteúdo.

Todas as bases utilizadas no projeto podem ser baixadas no [link:https://dadosabertos.tse.jus.br/dataset/resultados-2022/resource/f509562b-3b7f-487d-ad61-145a7ae6b96f](https://dadosabertos.tse.jus.br/dataset/resultados-2022/resource/f509562b-3b7f-487d-ad61-145a7ae6b96f)

4.1.2 Descrição dos Dados nas Fonte de Dados

A seguir, seguem os nomes das variáveis presentes nos bancos de dados, juntamente com suas descrições:

DT_GERACAO: Data da extração dos dados para geração do arquivo.

HH_GERACAO: Hora da extração dos dados para geração do arquivo com base no horário de Brasília.

ANO: Ano de referência da eleição para geração do arquivo.

CD_TIPO_ELEICAO: Código do tipo de eleição. Pode assumir os valores: 1 - Eleição Suplementar, 2 - Eleição Ordinária e 3 - Consulta Popular.

NM_TIPO_ELEICAO: Nome do tipo de eleição.

NR: Número do turno da eleição.

CD_ELEICAO: Código único da eleição no âmbito da Justiça Eleitoral.

DS_ELEICAO: Descrição da eleição.

DT_ELEICAO: Data em que ocorreu a eleição.

TP_ABRANGENCIA: Abrangência da eleição. Pode assumir os valores: Municipal, Estadual e Federal.

SG_UF: Sigla da Unidade da Federação em que ocorreu a eleição

SG_UE: Sigla da Unidade Eleitoral em que a candidata ou o candidato concorre na eleição. A Unidade Eleitoral representa a Unidade da Federação ou o Município em que a candidata ou o candidato concorre na eleição e é relacionada à abrangência territorial desta candidatura. Em caso de abrangência Federal (cargo de Presidente e Vice-Presidente) a sigla é BR. Em caso de abrangência Estadual (cargos de Governador, Vice-Governador, Senador, Deputado Federal, Deputado Estadual e Deputado Distrital) a sigla é a UF da candidatura. Em caso de abrangência Municipal (cargos de Prefeito, Vice-Prefeito e Vereador) é o código TSE de identificação do município da candidatura.

NM_UE: Nome de Unidade Eleitoral da candidata ou candidato (em caso de eleição majoritária é o nome da UF que o candidato concorre e em caso de eleição municipal é o nome do município).

CD_MUNICIPIO: Código TSE do município onde ocorreu a eleição.

NM_MUNICIPIO: Nome do município onde ocorreu a eleição.

NR_ZONA: Número da zona onde ocorreu a eleição.

NR_SECAO: Número da seção em que ocorreu a eleição.

CD_CARGO: Código do cargo da candidata ou candidato.

DS_CARGO: Descrição do cargo da candidata ou candidato.

NR_VOTAVEL: Número do votável. Pode assumir os valores: - número da candidata ou candidato, quando voto nominal; - número do partido, quando voto em legenda; - número 95, quando voto em branco; - número 96, quando voto nulo; - número 97, quando voto anulado e apurado em separado.

NM_VOTAVEL: Nome do votável. Pode assumir os valores: - nome do candidato, quando voto nominal ou voto anulado; - nome do partido, quando voto em legenda; - "Voto em branco", quando voto em branco; - "Voto nulo", quando voto nulo; - "Voto anulado e apurado em separado", quando voto anulado e apurado em separado.

QT: Quantidade de votos recebidos pelo votável naquele município, zona e seção.

NR_LOCAL_VOTACAO: Número do local de votação da eleitora ou eleitor.

NM_LOCAL_VOTACAO: Nome do local de votação da eleitora ou eleitor.

DS_LOCAL_VOTACAO_ENDERECO: Descrição do endereço do local de votação da eleitora ou eleitor.

4.2 Modelagem Conceitual, Lógica e Física do Banco de Dados

4.2.1 Modelo de Entidades e Relacionamentos

O Modelo Entidade Relacionamento é um modelo conceitual utilizado para descrever os objetos(entidades) envolvidos em um domínio de negócios, com suas características(atributos) e como elas se relacionam entre si.

Este modelo representa de forma abstrata a estrutura que possuirá o banco de dados da aplicação.

Os objetos, também chamados de entidades podem ser classificados como físicos ou lógicos. As entidades físicas são aquelas tangíveis, existentes e visíveis no mundo real, como um cliente ou um produto.

As entidades lógicas são aquelas que existem geralmente em decorrência da interação entre ou um com entidades físicas. As entidades são nomeadas com substantivos concretos ou abstratos que representem de forma clara sua função dentro do domínio.

Tipos de entidades:

- Entidades Fortes: São aquelas cuja existência independe de outras entidades, ou seja, por si só elas já possuem total sentido de existir.
- Entidades Fracas: As entidades fracas são aquelas que dependem de outras entidades para existirem, pois individualmente elas não fazem sentido.
- Entidades Associativas: São aquelas que não possuem atributos, mas que servem para relacionar outras entidades.

Depois que as entidades são definidas, é necessário definir os atributos que elas possuirão.

- Relacionamento 1-1: Um para um, ou seja, uma entidade pode estar relacionada com apenas uma outra entidade.
- Relacionamento 1-N: Um para muitos, ou seja, uma entidade pode estar relacionada com várias outras entidades.
- Relacionamento N-N: Muitos para muitos, ou seja, uma entidade pode estar relacionada com várias outras entidades e uma outra entidade pode estar relacionada com várias outras entidades.

Os relacionamentos em geral são nomeados com verbos ou expressões que representam a forma como as entidades interagem, ou a ação que uma exerce sobre a outra.

Os atributos descrevem as características de uma entidade, ou seja, são as propriedades que ela possui. Os atributos podem ser classificados como:

- **Descritivos:** São aqueles que descrevem a entidade, como por exemplo, o nome de um cliente.
- **Identificadores:** São aqueles que identificam a entidade, como por exemplo, o CPF de um cliente.
- **Compostos:** São aqueles que são formados por outros atributos, como por exemplo, o endereço de um cliente, que é composto por rua, número, bairro, cidade, estado e CEP.
- **Nominativos:** São aqueles que são formados por outros atributos, mas que não possuem sentido por si só, como por exemplo, o nome completo de um cliente, que é composto pelo nome e sobrenome.
- **Referenciais:** Representam a ligação de uma entidade com outra em um relacionamento. Por exemplo, uma venda possui o CPF do cliente, que a relaciona com a entidade cliente.
- **Simples:** um único atributo define uma característica da entidade. Exemplos: nome, peso.

4.2.2 Normalização das Entidades

O processo de normalização compreende o uso de um conjunto de regras, chamados de formas normais. Ao analisarmos o banco de dados e verificarmos que ele respeita as regras da primeira forma normal, então podemos dizer que o banco está na “primeira forma normal”. Caso o banco respeite as primeiras três regras, então ele está na “terceira forma normal”. Mesmo existindo mais conjuntos de regras para outros níveis de normalização, a terceira forma normal é considerada o nível mínimo necessário para grande parte das aplicações.

A tabela foi obtida da seguinte forma abaixo, e não tendo a chave primária, e para normalizar um banco de dados é necessário um chave primária, para isso é preciso encontrar um atributo ou conjunto de atributos que consiga identificar uma instância de forma unívoca dentro da tabela. Seguindo essa regra foi encontrada as seguintes variáveis que obedeciam essa definição, sendo elas: NR_ZONA, NR_SECAO, CD_ELEICAO(fk), SQ_CADIDATO, NR_VOTAVEL), CD_CARGO, CD_MUNICIPIO.

Isto porque NR_ZONA e NR_SECAO consegue separar cada urna por uma zona e por uma seção, CD_ELEICAO faz parte por cada eleição ter um código único, SQ_CANDIDATO e NR_VOTAVEL separam os candidatos por urnas que são identificadas nas duas primeiras variáveis, CD_CARGO porque um governador pode ter o mesmo número de um presidente, CD_MUNICIPIO entra na chave primária por ter mesmo números de candidatos e mesmos identificadores de urnas em municípios diferentes.

Após encontrar a chave primária da tabela pode-se começar o processo de normalização. A tabela obedece a regra da primeira forma normal, sendo assim ela não tem nenhuma tabela aninhada, ou seja, não existe atributos multivalorados.

```
VOTACAO_SECAO (NR_ZONA, NR_SECAO, CD_ELEICAO, SQ_CADIDATO, NR_VOTAVEL,  
CD_CARGO, CD_MUNICIPIO, ANO_ELEICAO, CD_TIPO_ELEICAO, NM_TIPO_ELEICAO,  
NR_TURNO, DS_ELEICAO, DT_ELEICAO, TP_ABRANGENCIA, SG_UF, SG_UE, NM_UE,  
NM_MUNICIPIO, DS_CARGO, MN_NOTAVEL, QT_VOTOS, NR_LOCAL_VOTACAO,  
NM_LOCAL_VOTACAO, DS_LOCAL_VOTACAO_ENDERECO)
```

A tabela não estão seguindo a regra da segunda forma normal, ela tem dependências parciais da chave primária, ou seja, alguns atributos dependem apenas de uma parte de chave primária. Depois de aplicar a segunda forma normal obtemos as seguintes entidades e suas listas de atributos:

```
VOTACAO (NR_ZONA, NR_SECAO, CD_ELEICAO(fk), SQ_CADIDATO(fk), NR_VOTAVEL(fk),  
CD_CARGO(fk), CD_MUNICIPIO(fk), TP_ABRANGENCIA, SG_UF, SG_UE(fk), NM_UE,  
QT_VOTOS, NR_LOCAL_VOTACAO, NM_LOCAL_VOTACAO, DS_LOCAL_VOTACAO_ENDERECO)  
ELEICAO (CD_ELEICAO, DS_ELEICAO, DT_ELEICAO, CD_TIPO_ELEICAO, NM_TIPO_ELEICAO,  
ANO_ELEICAO, NR_TURNO)  
MUNICIPIO (CD_MUNICIPIO, NM_MUNICIPIO)  
CARGO (CD_CARGO, DS_CARGO)  
VOTAVEL (SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO(fk), SG_UE(fk), MN_NOTAVEL)
```

As entidades na segunda forma normal é possível analisa-las para aplicar a segunda forma normal, dessa forma é preciso tirar todas as dependências transitivas das variáveis das tabelas, isso significa que precisamos tirar atributos que dependem de outros atributos que não fazem parte da chave primária.

```

VOTACAO (NR_ZONA, NR_SECAO, CD_ELEICAO(fk), SQ_CANDIDATO(fk), NR_VOTAVEL(fk),
CD_CARGO(fk), CD_MUNICIPIO(fk), TP_ABRANGENCIA, SG_UF, QT_VOTOS, SG_UE(fk),
NR_LOCAL_VOTACAO, NM_LOCAL_VOTACAO, DS_LOCAL_VOTACAO_ENDERECO)

ELEICAO (CD_ELEICAO, DS_ELEICAO, DT_ELEICAO, CD_TIPO_ELEICAO, NM_TIPO_ELEICAO,
ANO_ELEICAO, NR_TURN0)

MUNICIPIO (CD_MUNICIPIO, NM_MUNICIPIO)

CARGO (CD_CARGO, DS_CARGO)

VOTAVEL (SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO(fk), SG_UE(fk), MN_NOTAVEL)

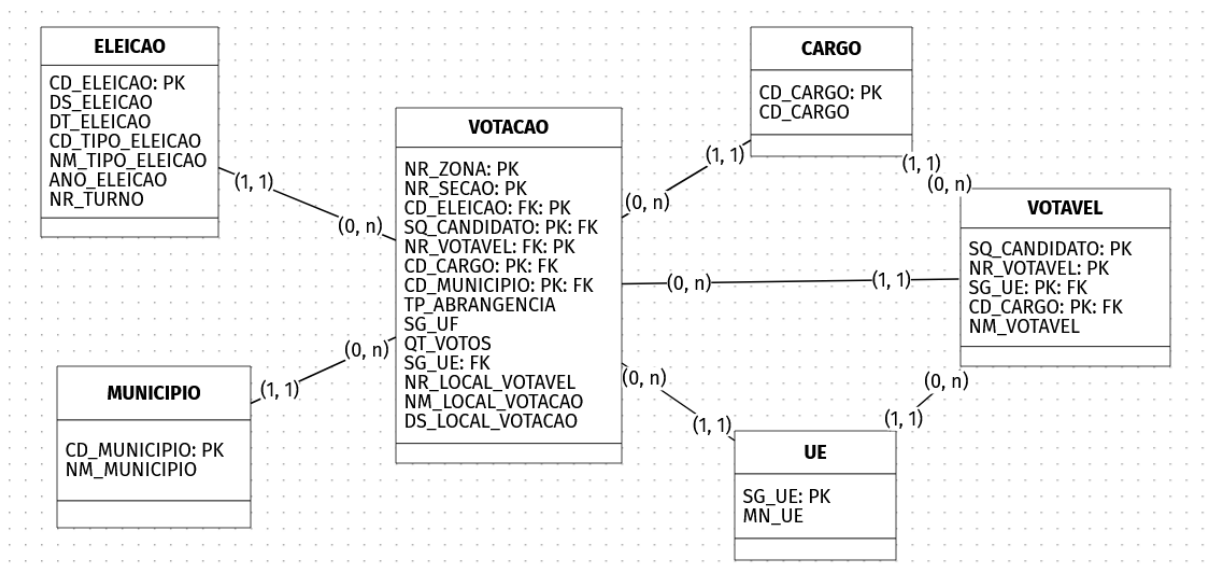
UNIDADE ELEITORAL (SG_UE, NM_UE)

```

4.2.3 Modelo Lógico do Banco de Dados

Um modelo de dados lógico estabelece uma estrutura de dados que pode ser usada para representar um conjunto de dados. Um modelo de dados lógico é um modelo de dados que é independente de qualquer sistema de armazenamento de dados específico. Ele pode ser usado para representar dados em um banco de dados relacional, em um banco de dados orientado a objetos, em um banco de dados de documentos ou em qualquer outro sistema de armazenamento de dados.

O modelo lógico da banco de dados utilizado é a seguinte:



4.2.4 SQL de Criação do Banco de Dados

SQL é uma linguagem de programação utilizada para trabalhar com banco de dados relacionais, por ser uma linguagem declarativa não é preciso um grande conhecimento em programação para se usar SQL, apenas de um conhecimento de sintaxe. Para esse projeto foi utilizado o SGBD postgres hospedado em um servidor

Para criar tabelas em SQL é preciso usar o comando `CREATE TABLE "tabela"`, após isso é preciso definir as variáveis que pertencem as tabelas e o tipo de cada uma, isso é identificado logo a frente do nome de cada variável.

Existem diferentes tipo de variáveis, `VARCHAR`, por exemplo, é um tipo de texto, isso significa que ela aceita qualquer carácter que for adicionado a ela, mas existe um limite definido pelo próprio usuário para a quantidade de variáveis. Existem também os tipos numéricos, o `INT` é um exemplo, no caso do Postgres o limite é de 2147483647 positivo e negativo.

É necessário também colocar se uma variável pode aceitar `NULL` ou não, ou seja, se uma variável pode ter um valor vazio pra que um instância possa ser carregada dentro do banco de dados, se for definido '`NOT NULL`' isso significa que aquela variável precisa de um valor para poder ser cadastrada.

Existem também as constraints, elas são regras que a tabela deve seguir, no caso desse projeto apenas as constraints de FK e PK foram usadas, mas existem diversas regras, por exemplo se uma variável numérico pode ou não ser maior que um determinado número.

Os códigos SQL utilizados dentro do software pgAdmin foram os seguintes:

```
CREATE TABLE "cargo"
(
    CD_CARGO char(1) not null,
    DS_CARGO varchar(20) not null,

    constraint pk_cargo PRIMARY KEY (CD_CARGO)
);
```

```
CREATE TABLE "municipio"
(
    CD_MUNICIPIO varchar(5) not null,
    NM_MUNICIPIO varchar(32) not null,

    constraint pk_municipio PRIMARY KEY (CD_MUNICIPIO)
);
```

```
CREATE TABLE "eleicao"
(
    CD_ELEICAO varchar(3) not null,
    DS_ELEICAO varchar(30) not null,
    DT_ELEICAO date not null,
    CD_TIPO_ELEICAO varchar(1) not null,
    NM_TIPO_ELEICAO varchar(17) not null,
    ANO_ELEICAO varchar(4) not null,
    NR_TURNO varchar(1) not null,

    constraint pk_eleicao PRIMARY KEY (CD_ELEICAO)
);
```

```
CREATE TABLE "ue_eleitoral"
(
    SG_UE varchar(2) not null,
    NM_UE varchar(14),

    constraint pk_ue_eleitoral PRIMARY KEY (SG_UE)
);
```

```
CREATE TABLE "votavel"
(
    SQ_CANDIDATO varchar(12) not null,
    NR_VOTAVEL varchar(5) not null,
    MN_NOTAVEL varchar(51) not null,
    CD_CARGO char(1) not null,
    SG_UE varchar(2) not null,

    constraint PK_VOTAVEL PRIMARY KEY (SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, SG_UE)
    CONSTRAINT FK_CARGO FOREIGN KEY(CD_CARGO) REFERENCES cargo(CD_CARGO)
    CONSTRAINT FK_UE FOREIGN KEY(SG_UE) REFERENCES ue_eleitoral(SG_UE)
);
```

```
CREATE TABLE "votacao"
(
    NR_ZONA varchar(3) not null,
    NR_SECAO varchar(4) not null,
    CD_ELEICAO varchar(3) not null,
    SQ_CANDIDATO varchar(12) not null,
    NR_VOTAVEL varchar(5) not null,
    CD_CARGO varchar(1) not null,
    CD_MUNICIPIO varchar(5) not null,
    TP_ABRANGENCIA varchar(100) not null,
    SG_UF varchar(2) not null,
    QT_VOTOS int not null,
    SG_UE varchar(2) not null,
    NR_LOCAL_VOTACAO varchar(4) not null,
    NM_LOCAL_VOTACAO varchar(100) not null,
    DS_LOCAL_VOTACAO_ENDEREÇO varchar(100) not null,

    constraint pk_votacao PRIMARY KEY
    (NR_ZONA, NR_SECAO, CD_ELEICAO, SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, CD_MUNICIPIO),
    constraint FK_ELEICAO FOREIGN KEY(CD_ELEICAO) REFERENCES eleicao(CD_ELEICAO),
    constraint FK_VOTAVEL FOREIGN KEY(SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, SG_UE)
    REFERENCES votavel(SQ_CANDIDATO, NR_VOTAVEL, CD_CARGO, SG_UE),
    constraint FK_CARGO FOREIGN KEY(CD_CARGO) REFERENCES cargo(CD_CARGO),
    constraint FK_MUNICIPIO FOREIGN KEY(CD_MUNICIPIO)
    REFERENCES municipio(CD_MUNICIPIO),
    constraint FK_UE FOREIGN KEY(SG_UE) REFERENCES ue_eleitoral(SG_UE)
);
```

Para aplicar dentro do banco de dados foi preciso seguir a ordem abaixo por causa da dependências das FKs:

CARGO → MUNICIPIO → ELEICAO → UE_ELEITORAL → VOTAVEL → VOTACAO

4.2.5 Descrição da Carga dos Dados no Banco de Dados

Para dar carga no banco de dados é necessário separar os dados da tabela sem estar normalizada que foi baixada do tse, para isso foi utilizado a biblioteca pandas do python para separar em arquivos normalizados o arquivo com todos os dados. Primeiro é preciso importar a biblioteca:

```
import pandas as pd
```

Depois de importar a biblioteca pode-se importar o arquivo de dados para ser utilizado no programa:

```
eleicoes_df = pd.read_csv("votacao_secao_2022_RJ.csv", sep=';', encoding='cp1252')
```

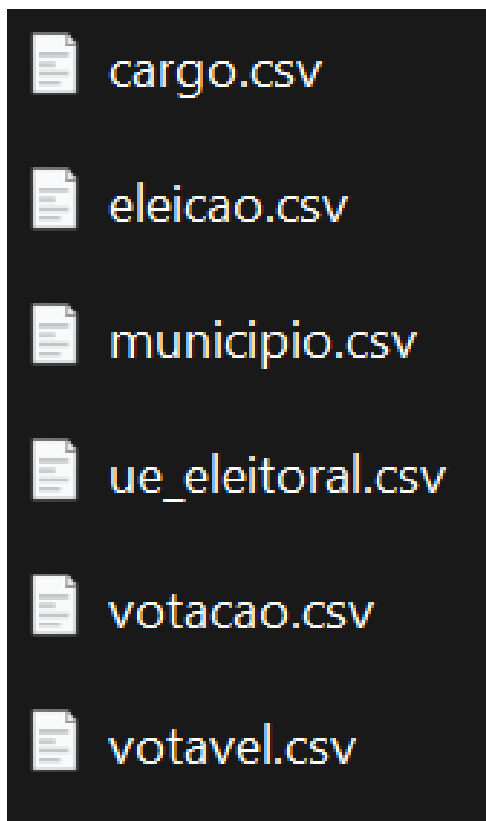
Depois do arquivo ser importando com sucesso é possível separar cada uma das tabelas normalizadas na terceira forma normal, o seguinte código foi utilizado:

```
votacao = eleicoes_df[[Attr_votacao]]
eleicao = eleicoes_df[[Attr_eleicao]].drop_duplicates()
municipio = eleicoes_df[[Attr_municipio]].drop_duplicates()
cargo = eleicoes_df[[Attr_cargo]].drop_duplicates()
votavel = eleicoes_df[[Attr_votavel]].drop_duplicates()
ue_eleitoral = eleicoes_df[[Attr_ue_eleitoral ]].drop_duplicates()
```

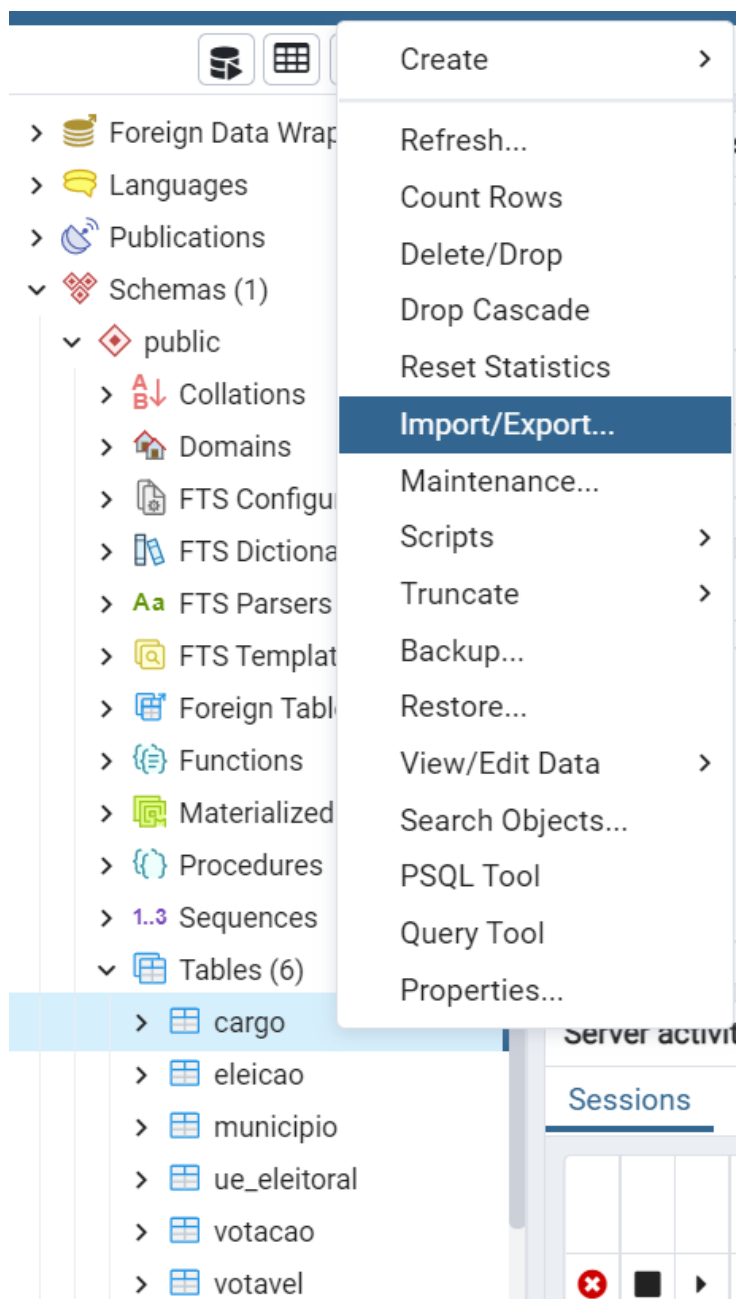
Após ter as tabelas normalizadas é possível transformar os dataframes do pandas em arquivos csv, para isso foi usado o código:

```
for df in ['votacao', 'eleicao', 'municipio', 'cargo', 'votavel', 'ue_eleitoral']:
    eval(df+'.to_csv("C:/Users/marle/Desktop/eleicoes/tabelas_normalizadas/'+df+'.csv", \
        header=True, index=False, sep=";")')
```

O resultado desse código são os seguintes arquivos:



Após todos os arquivos das tabelas normalizadas é possível importar eles para dentro do banco de dados criado anteriormente. Utilizando o pgAdmin existe a opção “Import/Export” que é a ferramenta utilizada para carregar o banco.



Dentro da ferramenta é possível configurar as opções básicas do arquivo, como por exemplo, o enconder e o delimitador, nesses arquivos foi utilizado o “utf-8” e “;” respectivamente. Também é preciso mostrar ao programa onde o arquivo está armazenado localmente.

Import/Export data - table 'cargo'

Options Columns

Import/Export **Import**

File Info

Filename: C:\Users\marle\Desktop\eleicoes\tabelas_normalizadas\cargo.csv

Format: csv

Encoding: UTF8

Miscellaneous

OID: No

Header: Yes

Delimiter: ;

Specifies the character that separates columns within each row (line) of the file. The default is a tab character in text format, a comma in CSV format. This must be a single one-byte character. This option is not allowed when using binary format.

Cancel OK

Para verificar se os dados foram importados com sucesso pode-se fazer uma requisição SQL para ver os dados dentro da tabela, para isso foi utilizado o seguinte código e teve o seguinte resultado para essa tabela:

Query Editor Query History Scratch Pad

```
1 select * from cargo
```

Data Output Explain Messages Notific

	cd_cargo [PK] character (1)	ds_cargo character varying (20)
1	1	PRESIDENTE
2	6	DEPUTADO FEDERAL
3	7	DEPUTADO ESTADUAL
4	3	GOVERNADOR
5	5	SENADOR

4.2.6 Dicionário de Dados

Cargo:

Variable Number	Name	Type	Format	Label	Length
1	CD_CARGO	Character	\$CHAR		1
2	DS_CARGO	Character	\$CHAR		17

Município:

Variable Number	Name	Type	Format	Label	Length
1	CD_MUNICIPIO	Character	\$CHAR		5
2	NM_MUNICIPIO	Character	\$CHAR		31

Eleicao:

Variable Number	Name	Type	Format	Label	Length
1	CD_ELEICAO	Character	\$CHAR		3
2	DS_ELEICAO	Character	\$CHAR		32
3	DT_ELEICAO	Numeric	DATE		8
4	CD_TIPO_ELEICAO	Character	\$CHAR		1
5	NM_TIPO_ELEICAO	Character	\$CHAR		20
6	ANO_ELEICAO	Numeric	BEST		8
7	NR_TURNO	Numeric	BEST		8

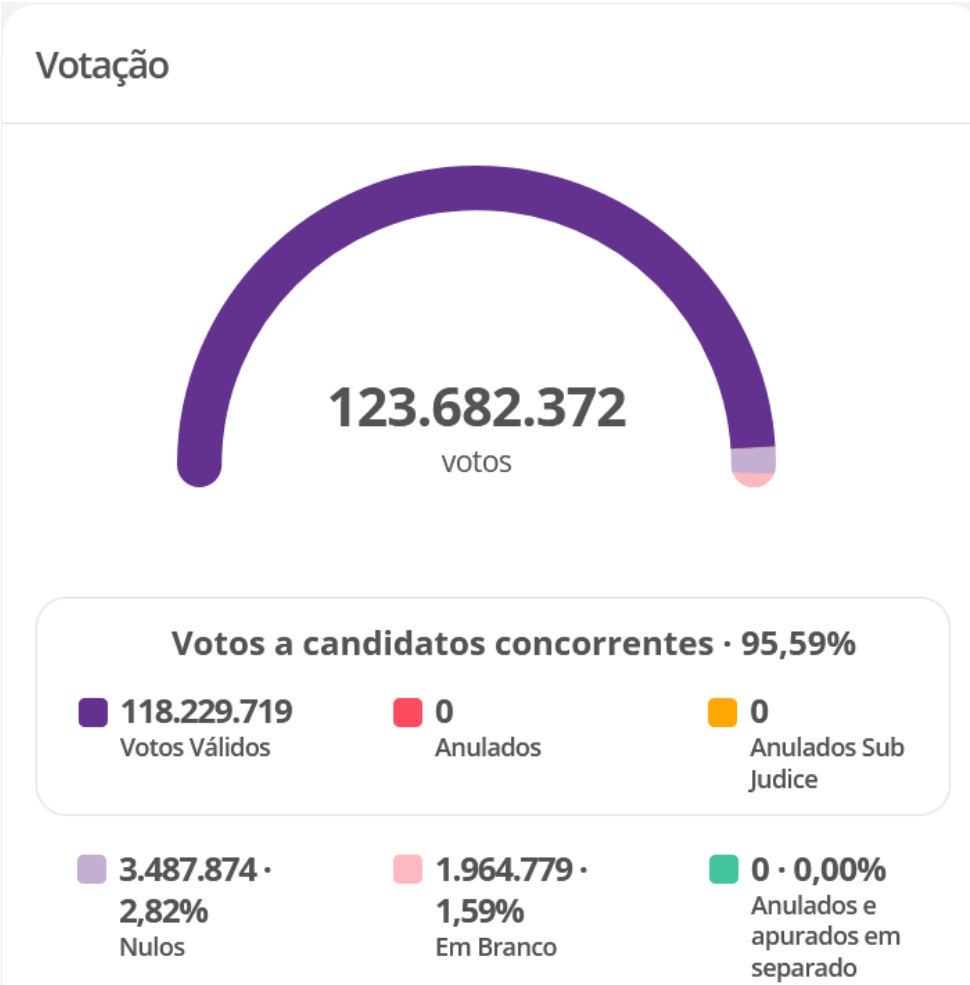
UE Eleitoral:

Variable Number	Name	Type	Format	Label	Length
1	SG_UE	Character	\$CHAR		2
2	NM_UE	Character	\$CHAR		14

Votavel:

Variable Number	Name	Type	Format	Label	Length
1	SQ_CANDIDATO	Character	\$CHAR		12
2	NR_VOTAVEL	Character	\$CHAR		5
3	NM_VOTAVEL	Character	\$CHAR		51
4	CD_CARGO	Character	\$CHAR		1
5	SG_UE	Character	\$CHAR		2

Votacao:



Para o segundo turno presidencial temos as seguintes quantidades de dados

```
3 select sum(qt_votos) from votacao where(cd_eleicao = '545')
```

Data Output Explain Messages Notifications

	sum bigint	
1	124252796	

No banco de dados a quantidade de votos para o segundo turno presidencial é de 124.252.796.



As quantidade de votos no banco de dados e no site do tse é a mesma, dessa forma está comprovada que todos os dados foram importados com sucesso para dentro do banco de dados

5 Análise dos Dados das Eleições

Neste capítulo será apresentada as Estatísticas Descritivas da análise dos Dados das Eleições Gerais 2022. Serão realizadas análises da eleição para presidente, em dois turnos, e para o estado do, para os cargos de Governador, Senador, Deputado Federal e Deputado Estadual.

5.1 Estatísticas Descritivas

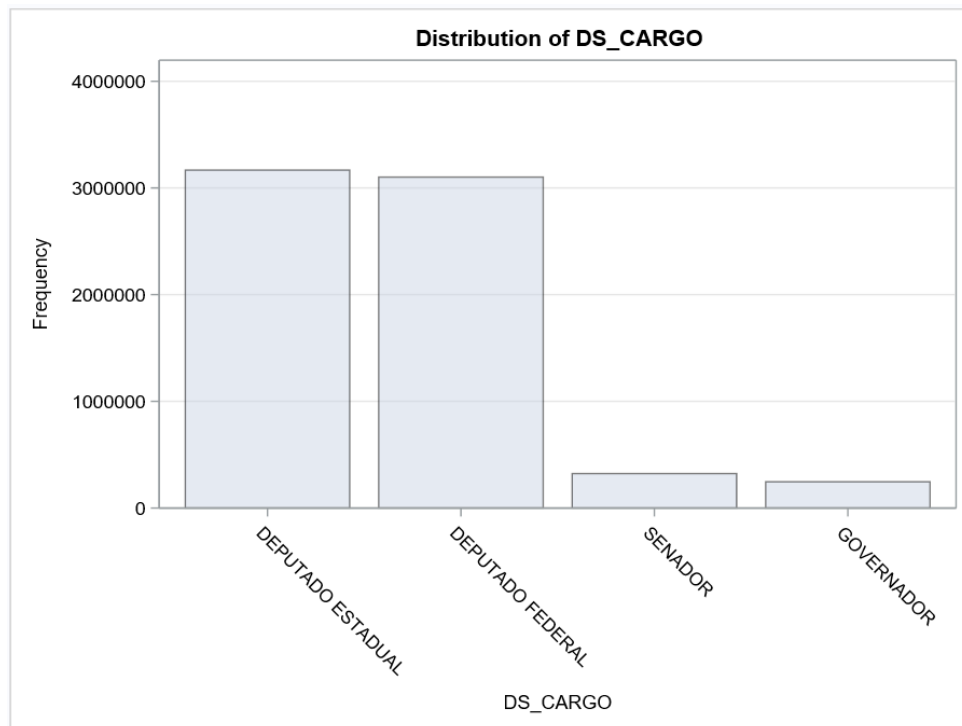
5.1.1 Eleição Governador, Senador, Deputado Federal e Deputado Estadual

Para as eleições gerais 2022, no estado do Rio de Janeiro, segundo o TSE, 119 vagas para cargos públicos estavam em disputa, entre elas: 46 vagas para Deputados Federais, 70 vagas para Deputados Estaduais, 1 vaga para Senador, Governador e Presidente da República. Pode-se confirmar isso verificando o domínio da variável "DS_CARGO" na base de dados onde estão os votos estaduais, gerando uma Tabela de Distribuição de Frequências da variável. No SAS, para realizar tal tarefa, basta clicar na opção Tasks, depois selecionar a opção Browse, abrir o menu de opções Describe e escolher a opção One-Way Frequencies. Depois de escolhidas as configurações desejadas basta clicar em Run. Após realizar este último passo foi gerada a seguinte tabela:

DS_CARGO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
DEPUTADO ESTADUAL	3167243	46.32	3167243	46.32
DEPUTADO FEDERAL	3101337	45.35	6268580	91.67
GOVERNADOR	246574	3.61	6515154	95.27
SENADOR	323116	4.73	6838270	100.00

Na imagem acima pode-se analisar que existem 5 colunas e 5 linhas. Na primeira coluna se encontra o nome da variável juntamente com suas 4 categorias, que no caso são as descrições dos cargos relacionados. Como se trata de uma base de dados com votos apenas para cargos estaduais, isto é, cargos que só podem receber votos de um único estado no domínio da variável "DS_CARGO" aparecem apenas essas 4 categorias. Já nas 4 colunas que se seguem pode-se analisar a presença das colunas; "Frequency": que nos informa a frequência absoluta das categorias, isto é, o número de ocorrências de cada categoria; também temos a coluna "Percent", que é a frequência relativa de cada valor absoluto das categorias, isto é, relativa à frequência absoluta. Por fim, as colunas "Cumulative Frequency" e "Cumulative Percent", que nos informa, respectivamente, a soma de todas as ocorrências até o elemento analisado e a porcentagem relativa à frequência acumulada.

Com isso, pode-se analisar que o número de ocorrências para as categorias “DEPUTADO ESTADUAL” e “DEPUTADO FEDERAL” representou 91,67% das ocorrências totais, pode-se analisar essa diferença também no gráfico de barras a seguir, que ilustra a frequência absoluta das categorias:



Essa diferença aconteceu por consequência do número de vagas e candidatos para os dois primeiros cargos ser muito maior do que para os dois últimos. Pode-se perceber essas e outras diferenças também entre esses números gerando uma “Task” com o sumários estatístico, utilizando cada uma das categorias da variável “DS_CARGO” como variáveis classificatórias e a variável “QT_VOTOS” como variável quantitativa. Um sumário estatístico é um resumo de atributos de determinadas variáveis de uma tabela. Neste caso, as categorias da variável “DS_CARGO”, como mostrado na tabela a seguir:

Analysis Variable : QT_VOTOS												
DS_CARGO	N Obs	Mean	Std Dev	Minimum	Maximum	Range	N	Lower Quartile	Median	Upper Quartile	90th Pctl	95th Pctl
SENADOR	323116	30.6195236	27.1256294	1.0000000	266.0000000	265.0000000	323116	10.0000000	25.0000000	43.0000000	68.0000000	84.0000000
GOVERNADOR	246574	40.1244981	50.5043772	1.0000000	366.0000000	365.0000000	246574	8.0000000	19.0000000	48.0000000	129.0000000	161.0000000
DEPUTADO FEDERAL	3101337	3.1901267	5.9810683	1.0000000	206.0000000	205.0000000	3101337	1.0000000	1.0000000	3.0000000	6.0000000	11.0000000
DEPUTADO ESTADUAL	3167243	3.1237445	7.2460267	1.0000000	246.0000000	245.0000000	3167243	1.0000000	1.0000000	2.0000000	6.0000000	11.0000000

Como mostrado anteriormente, é possível perceber que o número de ocorrências de Deputado Federal e Estadual é maior do que de Senador e Governador, mas além disso é possível analisar também que a média de votos para Senador e Governador é maior do que para os cargos de Deputado. Isso acontece pois, apesar do número de ocorrências ser menor, veremos mais a frente que a soma total de votos para cada um dos cargos é a mesma para ambos os cargos. Ou seja, a única diferença entre a media para cargos de deputados, governador e senador é o número total de observações ao qual será dividido com o total de votos, por isso a média para Senador e Governador é maior. Outro número que é possível analisar também é o Desvio-padrão. É possível perceber que o desvio-padrão

dos votos para senador e governador é maior do que para deputado estadual e federal, isso significa que a variabilidade, isto é, a diferença de votos contabilizados por sessão para cada uma das ocorrências de senador e governador estão mais distantes entre si em comparação com os votos nas ocorrências para Deputado Estadual e Federal. É possível perceber isso também analisando os quartis e percentis das ocorrências, que nos indicam que 50% das ocorrências para cargos de Senador e Governador que menos tiveram votos, respectivamente, tiveram até 25 e 19 votos, enquanto que essas mesmas 50% de ocorrências para Deputados Estaduais e Federais não passaram de 1 voto.

Outra maneira de compreender a diferença de ocorrências das categorias da variável “DS_CARGO” é estimando a quantidade de candidatos para cada um dos cargos criando novas tabelas com as variáveis “NM_VOTAVEL” e “DS_CARGO” aplicando filtros na tabela original selecionando apenas as linhas onde ocorre cada uma das categorias da variável “DS_CARGO”, isto é, gerar quatro novas tabelas apenas com as variáveis “NM_VOTAVEL” e “DS_CARGO”, onde cada uma das tabelas terá registros de apenas uma das 4 categorias diferentes dessa última variável. Para realizar tal tarefa utilizando o SAS basta, primeiramente, ir na opção “Tasks”, depois em “Browse”, selecione a opção “Data”, depois clique em “Filter and Sort”. Na aba “Variables” selecione as variáveis “NM_VOTAVEL” e “DS_CARGO” e na aba “Filter” selecione a variável com a descrição dos cargos, selecione a opção “Equal to” e depois preencha o terceiro campo com o nome da categoria a ser realizada o filtro. Após a filtragem, basta gerar uma nova análise de distribuição de frequência, da mesma maneira que no exemplo acima, mas dessa vez utilizando as tabelas oriundas das filtragens. Após isso, será necessário transformar os dados das Tabelas de Distribuição de Frequências em novas tabelas, para que seja possível realizar novas análises. Depois que as quatro tabelas oriundas das Tabelas de Distribuições estiverem prontas, basta então gerar uma análise de atributos dessas tabelas, para descobrirmos o número total de linhas de cada uma, subtraindo-se duas, que são as ocorrências de votos nulos e brancos, o resultado será o número total de indivíduos que se candidataram e que receberam pelo menos um voto somados aos votos de legenda. Após todo esse processo, serão geradas as seguintes observações:

Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0000	Observations	1642
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 20:26:31	Observation Length	88
Last Modified	04/12/2022 20:26:31	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0000		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0001	Observations	1086
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 20:45:44	Observation Length	88
Last Modified	04/12/2022 20:45:44	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0000		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0002	Observations	15
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 20:51:06	Observation Length	88
Last Modified	04/12/2022 20:51:06	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0001		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Data Set Name	WORK.ONEWAYFREQOFNM_VOTAVELINFIL_0003	Observations	11
Member Type	DATA	Variables	5
Engine	V9	Indexes	0
Created	04/12/2022 21:01:08	Observation Length	88
Last Modified	04/12/2022 21:01:08	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label	Cell statistics for NM_VOTAVEL analysis of WORK.FILTER_FOR_VOTACAO_SECAO_20_0002		
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Pode-se perceber nas tabelas acima que o número de ocorrências únicas da variável “NM_VOTAVEL” para os votos a candidatos ao cargo de Deputado Estadual foram de 1642, enquanto para candidatos a Deputado Federal foi de 1086. Já para Senador e Governador, foi de 15 e 11, respectivamente. Mais uma vez, vale ressaltar que esta é apenas uma estimativa, não é possível afirmar que estes são os valores exatos de candidatos aos cargos pois dentro das ocorrências ainda constam os votos nulos, brancos e de legenda.

Para descobrirmos a quantidade exata de votos que cada partido precisava para indicar candidatos a Deputados Federais e Estaduais é necessário calcular, primeiramente, o quociente eleitoral de cada vaga em disputa. Para isso, como primeiro passo é necessário saber qual foi a quantidade total de votos recebidos para cada cargo, somando todos os votos de legenda e votos para candidatos únicos e subtraindo-se os votos nulos e brancos. Para realizar essa tarefa no SAS é necessário utilizar a “Task” chamada “Summary Statistics”, e configurá-la para nos mostrar a soma da quantidade de votos. Para isso, levando em consideração que nós precisaremos apenas adicionar a coluna com a quantidade de votos nas tabelas já filtradas e utilizadas para realizar as análises de frequências, devemos seguir os seguintes passos, um por um: clicar em “Tasks”, selecionar “Describe” e depois “Summary Statistics”, configurar como variável quantitativa a variável “QT_VOTOS”, e como variável classificatória “DS_CARGO” e selecionar na aba “Statistics” a opção “sum”, depois é só executar. Após realizar esse processo os seguintes parâmetros irão ser calculados:

Analysis Variable : QT_VOTOS	
DS_CARGO	Sum
DEPUTADO ESTADUAL	9893658.00
DEPUTADO FEDERAL	9893658.00
GOVERNADOR	9893658.00
SENADOR	9893658.00

É possível perceber que a soma de votos para cada um dos cargos é exatamente a mesma, isso aconteceu porquê todos que votaram em algum candidato para Deputado Estadual, por exemplo, também precisaram votar para os outros cargos, nem que o voto fosse nulo ou branco ou em algum partido. Já era de se esperar esse resultado. Mas, segundo o TSE, os votos nulos e brancos não podem ser contabilizados para o cálculo do quociente eleitoral, assim, para descobrir a quantidade de votos necessária para um partido indicar alguém, precisaremos subtrair da soma total de votos as ocorrências de votos nulos e brancos. Para realizar tal tarefa é necessário, primeiramente, somar os votos nulos e brancos contabilizados em cada um dos cargos. Para realizar tal tarefa no SAS será necessário criar uma nova tabela utilizando a “Task” “Filter and Sort”, apenas com as variáveis “DS_CARGO”, “NM_VOTAVEL” e “QT_VOTOS” e configurar como regras de filtragem apenas as ocorrências de votos nulos e brancos na variável “NM_VOTAVEL”. Após realizado esse processo, será necessário utilizar a “TASK” “Summary Statistics”, utilizando como variável quantitativa “QT_VOTOS” e como variáveis classificatórias “DS_CARGO” e “NM_VOTAVEL”, e selecionar a métrica “Sum”. Após finalizado todo esse processo, as seguintes observações serão geradas:

Analysis Variable : QT_VOTOS				
DS_CARGO	NM_VOTAVEL	N Obs	Sum	N
DEPUTADO ESTADUAL	VOTO BRANCO	34061	644435.00	34061
	VOTO NULO	34052	602675.00	34052
DEPUTADO FEDERAL	VOTO BRANCO	34060	616380.00	34060
	VOTO NULO	34056	600046.00	34056
GOVERNADOR	VOTO BRANCO	34058	591576.00	34058
	VOTO NULO	34068	894384.00	34068
SENADOR	VOTO BRANCO	34066	752864.00	34066
	VOTO NULO	34067	972167.00	34067

É possível perceber na tabela acima que o número total de observações de votos brancos e nulos para cada cargo é muito parecido. Isso acontece pois a base de dados utilizada para essas análises, exatamente a mesma utilizada pelo TSE, contabiliza em cada linha de registro a quantidade de votos de um único candidato, para um único cargo em uma única sessão dentro de uma única zona eleitoral. Ou seja, em uma única zona eleitoral, é possível ter várias sessões eleitorais, e em cada sessão eleitoral nós teremos contabilizados vários votos para vários candidatos aos 4 cargos possíveis (desconsiderando os votos para presidente, no momento estamos analisando apenas votos para cargos estaduais) incluindo os votos brancos, nulos e de legenda. É possível identificar exatamente a quantidade de sessões únicas do estado do Rio de Janeiro utilizando o SAS criando uma nova tabela por meio da opção “Query Builder”, para isso basta selecioná-la, escolher as variáveis “NR_ZONA” e “NR_SECAO” e selecionar também a opção “Select Distinct rows only”, para que apenas combinações diferentes apareçam na nova tabela. Após realizar esse processo, basta agora criar uma Tabela de Distribuição de Frequências utilizando a variável “NR_ZONA” e o número que aparecerá na coluna “Frequency” é exatamente a quantidade de sessões contidas dentro daquela zona. Visto que não será possível registrar toda a tabela de distribuição de frequências por conta de seu tamanho, nós iremos transformar as distribuições das frequências em uma nova tabela analisar esses números por meio da “Task” “Summary Statistics” para que seja possível observarmos essas características:

Analysis Variable : COUNT Frequency Count				
Mean	Minimum	Maximum	Sum	N
206.4727273	36.0000000	417.0000000	34068.00	165


A tabela acima nos informa que a média de sessões únicas por zona eleitoral é de, aproximadamente 207 sessões. Enquanto que a zona eleitoral que contabiliza o menor número de sessões possíveis tem exatamente 36 sessões, já a zona com o maior número de

sessões possíveis, têm 417 sessões. Além disso é possível perceber que o número de zonas eleitorais únicas no estado do Rio de Janeiro é 165, enquanto a soma de todas as sessões dessas zonas é 34068. Não por coincidência, o número total de sessões únicas no estado do Rio de Janeiro é exatamente igual ao número de sessões únicas com votos nulos para o cargo de Governador, isso significa que em todas as sessões possíveis, nós tivemos ao menos um voto nulo para governador, e apesar de muito próximos, nenhuma ocorrência de sessões que contiveram votos nulos ou brancos é maior do que as ocorrências de votos nulos para governador, visto que o número máximo de sessões únicas no estado do Rio de Janeiro é 34068. Voltando ao cálculo do quociente eleitoral, por meio dos números vistos na tabela 9, possível observar também, além do número de sessões, a soma total de votos brancos e nulos para cada um dos cargos, exatamente, 1.247.110 para Deputado Estadual, 1.216.426 para Deputado Federal, 1.485.960 para Governador e 1.725.031 para Senador. Agora para descobrirmos o quociente eleitoral para os cargos de Deputado Estadual e Federal, basta subtrair o total de votos de cada um dos cargo, números já apurados neste projeto, com o total de votos nulos e brancos de cada um e dividir o resultado pelo número total de cargos disponíveis para cada um, com isso, chegaremos aos seguintes valores:

- Quociente Eleitoral p/ Dep. Estadual: 123.522,11 (Aprox. 123.523 votos) Fórmula: $(9893658 - 1.247.110) / 70$

- Quociente Eleitoral p/ Dep. Federal: 188.635,47 (Aprox. 188.636 votos) Fórmula: $(9893658 - 1.216.426) / 46$

Para os cargos de Governador e Senador, como só existem 1 vaga para cada, prevalece o sistema majoritário, ou seja, vence quem tiver a maior quantidade de votos. Infelizmente, não será possível dizer exatamente quantos candidatos cada partido de fato pôde indicar para os cargos de Deputado, pois não temos informações sobre o partido de cada um dos candidatos, mas é possível perceber quais foram os candidatos para os cargos de Deputados que mais conseguiram puxar votos para o seu partido:

	 NM_VOTAVEL	 QT_VOTOS_Sum
1	VOTO BRANCO	6E5
2	VOTO NULO	6E5
3	MARCIO CORREIA DE OLIVEIRA	2E5
4	DOUGLAS RUAS DOS SANTOS	2E5
5	RENATA DA SILVA SOUZA	2E5
6	Partido Liberal	1E5
7	ROSENVERG REIS DE OLIVEIRA	1E5
8	Partido dos Trabalhadores	1E5
9	SÉRGIO LUIZ COSTA AZEVEDO FILHO	1E5
10	GUILHERME JANDRE DELAROLI	1E5

Na tabela acima é possível analisarmos os 10 candidatos a Deputados Estaduais que mais receberam votos no estado do Rio de Janeiro. Como o SAS não nos permite analisar o número exato de votos desses candidatos, utilizaremos a estimativa que consta na tabela. Em primeiro e segundo, é possível perceber que os votos nulos e brancos foram os que mais apareceram na eleição. Para que fosse possível descobrir exatamente a quantidade de votos que cada candidato recebeu foi necessário abrir a tabela acima do excel, pois o SAS não nos permitiu realizar o ordenamento e analisar a exata quantidade de votos, assim, foram geradas as seguintes tabelas utilizando o excel:

	A	B
1	NM_VOTAVEL	QT_VOTOS_Sum
2	VOTO BRANCO	644435
3	VOTO NULO	602675
4	MARCIO CORREIA DE OLIVEIRA	181274
5	DOUGLAS RUAS DOS SANTOS	175977
6	RENATA DA SILVA SOUZA	174132
7	Partido Liberal	144794
8	ROSENVERG REIS DE OLIVEIRA	131308
9	Partido dos Trabalhadores	126155
10	SÉRGIO LUIZ COSTA AZEVEDO FILHO	123739
11	GUILHERME JANDRE DELAROLI	114155

Na tabela acima é possível perceber que apenas as 9 primeiras ocorrências conseguiram a quantidade suficiente de votos para indicar um candidato a Deputado Estadual. Dessas 9, 4 foram ou votos de legenda ou votos brancos e nulos, ou seja, apenas 5 candidatos conseguiram sozinhos a quantidade de votos necessárias para se candidatar. Já para o Cargo de Deputado Federal:

1	NM_VOTAVEL	QT_VOTOS_Sum
2	VOTO BRANCO	616380
3	VOTO NULO	600046
4	DANIELA MOTÉ DE SOUZA CARNEIRO	213432
5	EDUARDO PAZUELLO	205324
6	TALÍRIA PETRONE SOARES	198548
7	LUIZ ANTONIO DE SOUZA TEIXEIRA JUNIOR	190071
8	ALTINEU CORTES FREITAS COUTINHO	167512
9	TARCÍSIO MOTTA DE CARVALHO	159928

É possível perceber na tabela acima que apenas os 7 primeiros candidatos conseguiram conquistar a quantidade correta de votos para se candidatar, sendo que desses, os dois primeiros foram votos nulos e brancos.

Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
CLÁUDIO BOMFIM DE CASTRO E SILVA	4930288.00
CYRO GARCIA	12627.00
EDUARDO GONÇALVES SERRA	10852.00
JULIETE PANTOJA ALVES	27344.00
LUIZ EUGÊNIO HONORATO	1844.00
MARCELO RIBEIRO FREIXO	2300980.00
PAULO GUSTAVO GANIME ALVES TEIXEIRA	447051.00
RODRIGO NEVES BARRETO	672291.00
VOTO BRANCO	591576.00
VOTO NULO	894384.00
WILSON JOSÉ WITZEL	4892.00

Já na tabela acima, com a quantidade de votos que cada um dos candidatos a governador, é possível perceber que o vencedor foi Cláudio Bomfim de Castro.

Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
ALESSANDRO LUCCIOLA MOLON	1731786.00
ANDRÉ LUIZ CECILIANO	986676.00
ANTONIO HERMANO LEMME	1198.00
BENEVENUTO DACIOLO FONSECA DOS SANTOS	285037.00
BÁRBARA DEL PENHO SINEDINO PINHEIRO	18222.00
CLARISSA GAROTINHO BARROS ASSED MATHEUS DE OLIVEIRA	1145413.00
DANIEL LUCIO DA SILVEIRA	1566352.00
HELVIO COSTA DE OLIVEIRA TELLES	7036.00
HIRAN ROEDEL	5120.00
MARCELO ZATURANSKY NOGUEIRA ITAGIBA	18224.00
RAUL BITTENCOURT PEDREIRA	7299.00
ROMÁRIO DE SOUZA FARIA	2385181.00
SUED HAIDAR NOGUEIRA	11933.00
VOTO BRANCO	752864.00
VOTO NULO	972167.00

Na na tabela acima, com a quantidade de votos que cada um dos candidatos a

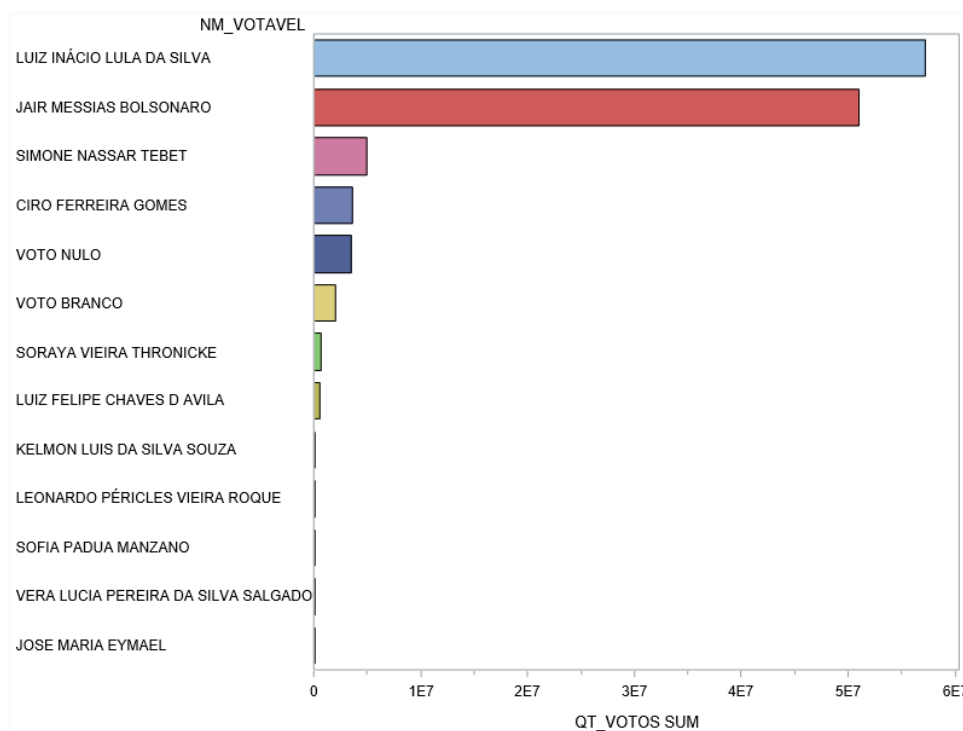
senador, é possível perceber que o vencedor foi Romário de Souza, com a maior quantidade de votos.

5.1.2 Eleição Presidencial

Nas eleições brasileiras de 2022, tivemos 11 candidatos concorrendo ao cargo de Presidente da República, sendo eles: Luiz Inácio Lula Da Silva, Jair Messias Bolsonaro, Simone Nassar Tebet, Ciro Ferreira Gomes, Soraya Vieira Thronicke, Luiz Felipe Chaves D Avila, Kelson Luis Da Silva Souza, Leonardo Péricles Vieira Roque, Sofia Padua Manzano, Vera Lucia Pereira Da Silva Salgado, José Maria Eymael.

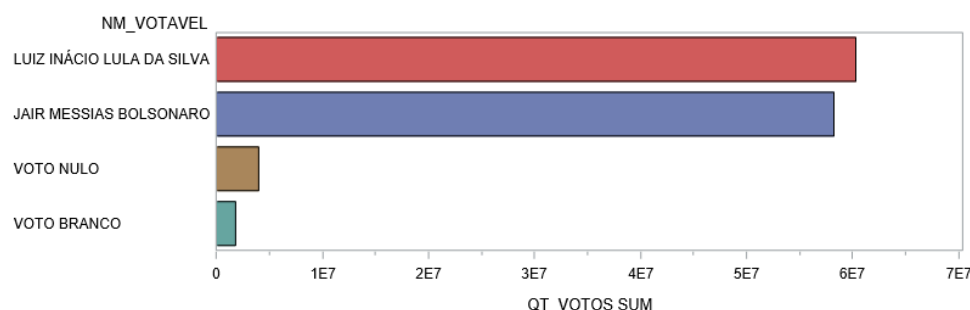
No 1º turno, foram computados 123.682.372 de votos, sendo eles 118.229.719 votos válidos, 1.964.779 brancos, 3.487.874 nulos, além de 32.770.982 abstenções. A divisão de votos foi feita da seguinte forma: Luiz Inácio Lula Da Silva - 57.259.504 votos; Jair Messias Bolsonaro - 51.072.345 votos; Simone Nassar Tebet - 4.915.423 votos; Ciro Ferreira Gomes - 3.599.287 votos; Soraya Vieira Thronicke - 600.955 votos; Luiz Felipe Chaves D Avila - 559.708 votos; Kelson Luis Da Silva Souza - 81.129 votos; Leonardo Péricles Vieira Roque - 53.519 votos; Sofia Pádua Manzano - 45.620 votos; Vera Lucia Pereira Da Silva Salgado - 25.625 votos; José Maria Eymael - 16.604 votos.

Luiz Inácio Lula Da Silva obteve 48,43% dos votos e Jair Messias Bolsonaro 43,20%, com isso, destacaram-se diante dos demais, obtendo juntos 91,63% dos votos, demonstrando grande disparidade de concorrência em relação aos outros candidatos. Observou-se também, que além deles, apenas a candidata Simone Nassar Tebet e o candidato Ciro Ferreira Gomes, conseguiram mais votos do que a quantidade de votos nulos e brancos, como apresentado no gráfico a seguir:

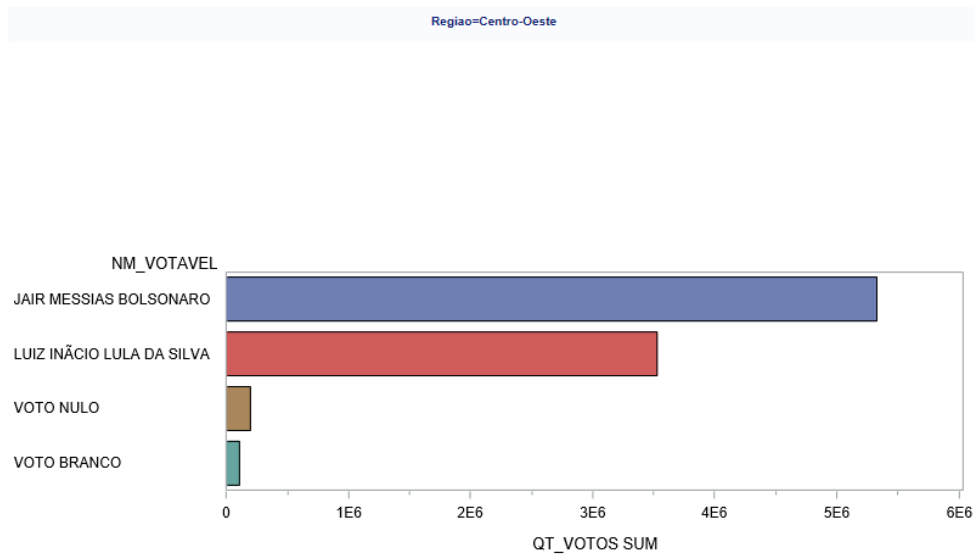


Como previsto nos artigos 28º e 29º da Constituição Federal de 1988, para que haja um eleito, faz-se necessário a obtenção de mais da metade dos votos válidos (excluídos os votos em branco e os votos nulos) para ser eleito, em 1º ou em 2º turno. Portanto, as eleições de 2022 tiveram dois turnos, visto que os candidatos Luiz Inácio Lula Da Silva e Jair Messias Bolsonaro não obtiveram a quantidade necessária para serem eleitos já no primeiro turno.

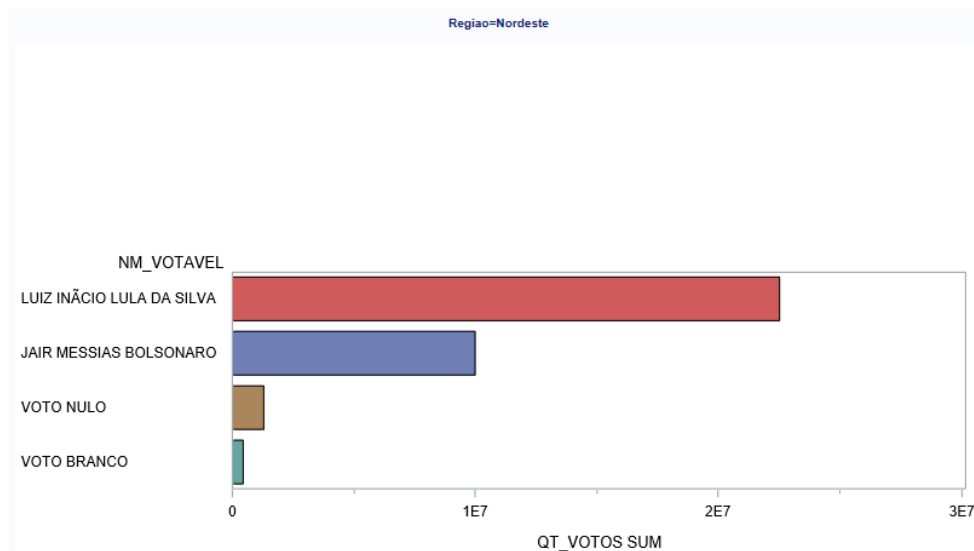
No 2º turno, foram computados 124.252.796 de votos, sendo eles 118.552.353 votos válidos, 1.769.678 brancos, 3.930.765 nulos, além de 32.200.558 abstenções. A divisão de votos foi feita da seguinte forma: Luiz Inácio Lula Da Silva - 60.345.999 votos, equivalente a 50,90% do total de votos; Jair Messias Bolsonaro - 58.206.354 votos, equivalente a 49,10% do total de votos.



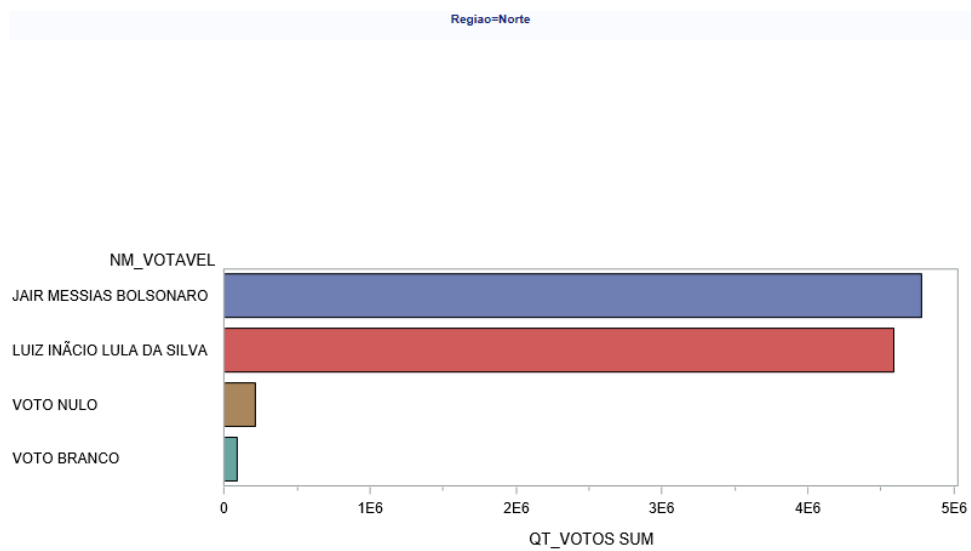
Na região Centro-Oeste, o candidato Jair Messias Bolsonaro obteve 58% dos votos, enquanto o candidato Luiz Inácio Lula da Silva obteve 38%. O desempenho dos candidatos em cada estado foi: Goiás: Bolsonaro - 58,71%; Lula: 41,29%. Mato Grosso: Bolsonaro - 65,0%; Lula - 34,92%. Mato Grosso do Sul: Bolsonaro - 59,49%. Lula - 40,51% e no Distrito Federal foi: Bolsonaro - 58,81%; Lula - 41,19%.



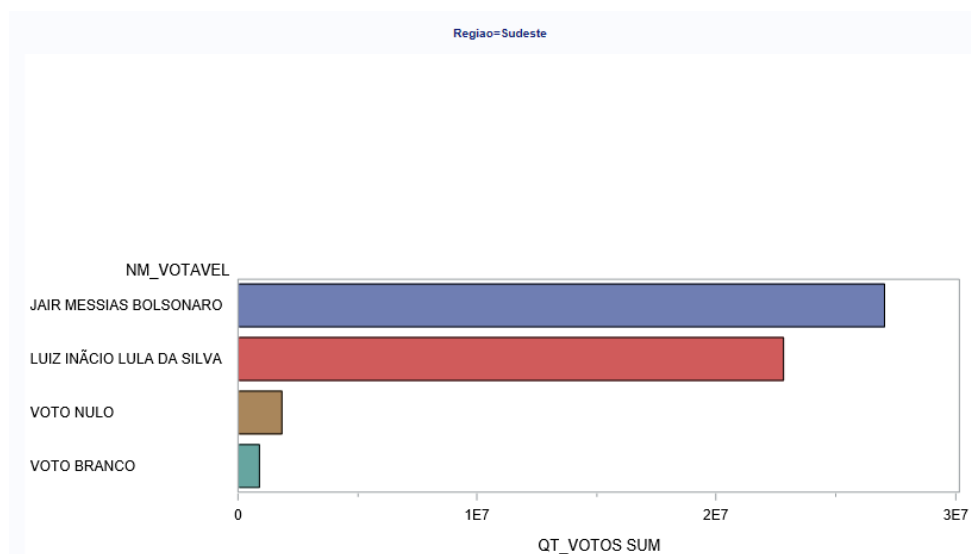
Na região Nordeste, o candidato Jair Messias Bolsonaro obteve 29% dos votos, enquanto o candidato Luiz Inácio Lula da Silva obteve 65%. O desempenho dos candidatos em cada estado foi: Alagoas: Bolsonaro - 41,32%; Lula - 58,68%. Bahia: Bolsonaro - 27,88% ; Lula - 72,12%. Ceará: Bolsonaro - 30,03% ; Lula - 66,62%. Maranhão: Bolsonaro - 28,86%; Lula - 71,14%. Paraíba: Bolsonaro - 33,38%; Lula - 66,62%. Pernambuco: Bolsonaro - 33,07%; Lula - 66,93%. Piauí: Bolsonaro - 23,14%; Lula - 76,86%. Rio Grande do Norte: Bolsonaro - 34,90%; Lula - 65,10%. Sergipe: Bolsonaro - 32,79%; Lula - 67,21%.



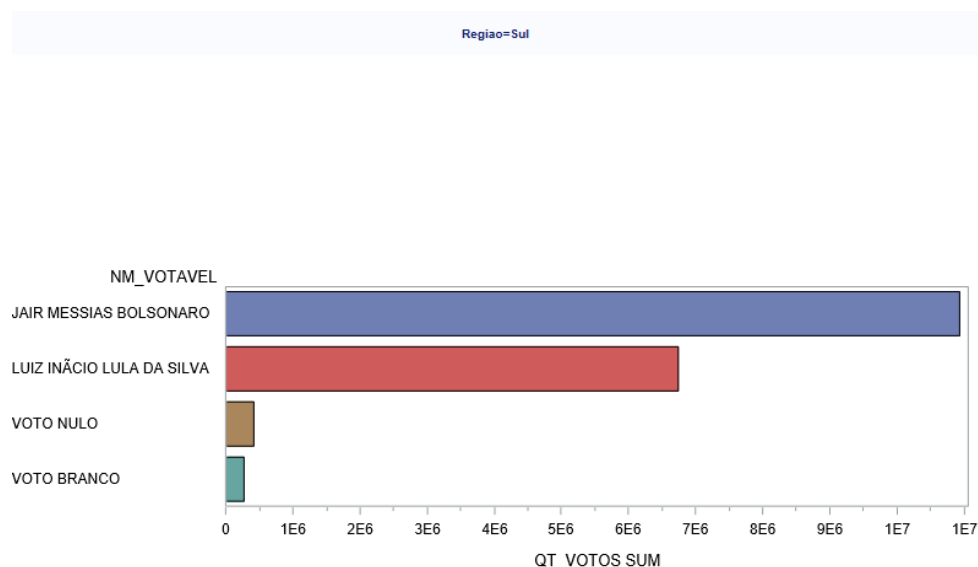
Na região Norte, o candidato Jair Messias Bolsonaro obteve 49% dos votos, enquanto o candidato Luiz Inácio Lula da Silva obteve 47%. O desempenho dos candidatos em cada estado foi: Acre: Bolsonaro - 70,30%; Lula - 29,70%. Amazonas Bolsonaro - 48,90%; Lula - 51,10%. Amapá: Bolsonaro - 51,36%; Lula - 48,64%. Pará: Bolsonaro - 45,25%; Lula - 54,75%. Rondônia: Bolsonaro - 70,66%; Lula - 29,34%. Roraima: Bolsonaro - 76,08%; Lula - 23,92%. Tocantins: Bolsonaro - 48,64%; Lula - 51,36%.



Na região Sudeste, o candidato Jair Messias Bolsonaro obteve 51% dos votos, enquanto o candidato Luiz Inácio Lula da Silva obteve 43%. O desempenho dos candidatos em cada estado foi: Rio de Janeiro: Bolsonaro - 56,53%; Lula 41,96%- . São Paulo: Bolsonaro - 55,24%; Lula - 44,76%. Espírito Santo: Bolsonaro - 58,04%; Lula - 41,96%. Minas Gerais: Bolsonaro - 50,20%; Lula - 49,80%.

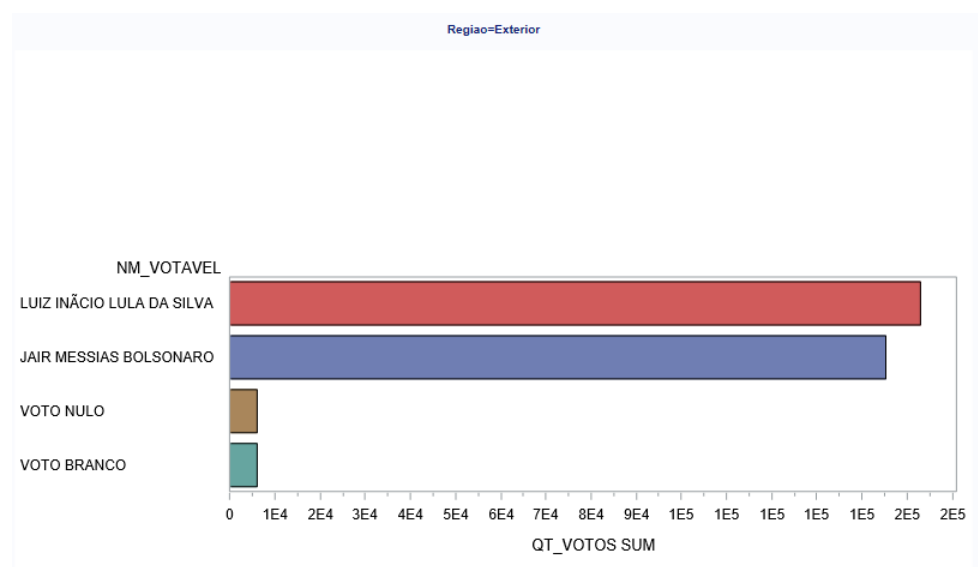


Na região Sul, o candidato Jair Messias Bolsonaro obteve 59% dos votos, enquanto o candidato Luiz Inácio Lula da Silva obteve 36%. O desempenho dos candidatos em cada estado foi: Rio Grande do Sul: Bolsonaro - 56,35%; Lula - 43,65%. Paraná: Bolsonaro - 62,40%; Lula - 37,60%. Santa Catarina: Bolsonaro - 69,27%; Lula - 30,73%.



Analisando cada região pode-se observar que o candidato Luiz Inácio Lula da Silva obteve mais sucesso nos estados de Alagoas, Amazonas, Bahia, Ceará, Maranhão, Minas Gerais, Pará, Paraíba, Pernambuco, Piauí, Rio Grande do Norte, Sergipe e Tocantins, totalizando 13 estados. Enquanto o candidato Jair Messias Bolsonaro obteve sucesso nos estados do Acre, Amapá, Distrito Federal, Espírito Santo, Goiás, Mato Grosso, Mato Grosso do Sul, Paraná, Rio de Janeiro, Rio Grande do Sul, Rondônia, Roraima, Santa Catarina e São Paulo, totalizando 13 estados e o Distrito Federal.

Além disso, houve também a apuração dos votos dos cidadãos que se encontravam em outro país no momento da eleição, onde o candidato Luiz Inácio Lula da Silva conquistou 51,28% dos votos e Jair Messias Bolsonaro conquistou 48,72% dos votos, sendo um total de 310.148 votos.



Em análise a estes resultados, observou-se uma grande equiparidade de votos entre ambos os candidatos durante toda a eleição, sendo os estados do nordeste, os principais decisores para que o candidato Luiz Inácio Lula da Silva fosse eleito o novo presidente

do Brasil, a diferença percentual entre os dois candidatos na região foi de 36 pontos percentuais a favor do candidato Lula.

Regiao=Nordeste	
Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
LUIZ INACIO LULA DA SILVA	22534987.00
JAIR MESSIAS BOLSONARO	9962947.00
VOTO NULO	1275547.00
VOTO BRANCO	415203.00

Em contrapartida, no sul a diferença entre os dois candidatos foi de 23 pontos percentuais a favor do Jair Messias Bolsonaro, sendo esta a região com maior apoio relativo ao candidato.

Regiao=Sul	
Analysis Variable : QT_VOTOS	
NM_VOTAVEL	Sum
LUIZ INACIO LULA DA SILVA	6750374.00
JAIR MESSIAS BOLSONARO	10940158.00
VOTO NULO	409071.00
VOTO BRANCO	274486.00

Somado a esta diferença de votos, se analisarmos a quantidade de votos por região, percebemos que o Nordeste é a segunda região mais populosa do Brasil, correspondendo a 27,5% do total de votos do país com 34.188.664 votos. Enquanto o estado do Sul representa 14% do número de votos computados, tendo um total de 18.374.089 votos

Analysis Variable : QT_VOTOS	
Regiao	Sum
Sudeste	52542866.00
Nordeste	34188864.00
Sul	18374089.00
Norte	9675082.00
Centro-Oeste	9161947.00
Exterior	310148.00

6 Conclusões

Resultados

No início do projeto foram definidos alguns objetivos, entre eles:

- Implementar um banco de dados relacional com todos os dados necessários para as análises descritivas dos resultados, com todos os requerimentos técnicos aprendidos dentro de sala de aula.

- Identificar nas bases nacionais e estaduais a quantidade necessária de votos para que um partido possa indicar candidatos a deputados e identificar os partidos ou coligações que mais tiveram votos e calcular quantos candidatos cada um desses partidos tiveram direito de indicar e comparar com quantos cada um dos partidos mais votados realmente indicaram para os cargos, segundo o TSE.

- Identificar nas bases estaduais os candidatos ao Senado mais e menos votados e os candidatos a Governador mais e menos votados e comparar com os candidatos realmente nomeados para os cargos, segundo o TSE.

- Identificar nas bases nacionais os candidatos a presidência mais e menos votados e comparar com o candidato vencedor, segundo o TSE.

Durante o projeto foram utilizadas diferentes ferramentas para chegar ao objetivo final, dentre elas estão o SAS Enterprise Guide, python, sql e pgAdmin todas elas colaboram para que todos os dados fossem analisados, normalizados e armazenados dentro de um banco de dados gerenciado pelo postgres. Todas essas ferramentas são utilizadas no mercado de trabalho e essa experiência é importante para que novos conhecimentos sejam adquiridos com a utilização das mesmas.

Infelizmente, não foi possível realizar todas as análises antes planejadas, pois seriam necessários implementar mais bases de dados com registros de partidos e etc. Mas, mesmo assim foi de extrema importância a elaboração deste projeto para que os alunos pudessem compreender melhor os sistemas eleitorais presentes no estado Brasileiro e também para que fosse possível colocar em prática todos os conhecimentos adquiridos durante o semestre nas aulas de Estatística Descritiva, Lógica de programação e Banco de Dados.

Desde o início da elaboração do projeto, todos os alunos do grupo se comprometeram a realizar da melhor maneira as análises e a implementação da bases de dados. Vários desafios foram postos devido a complexidade e a falta de experiência dos alunos envolvidos, mas acreditamos que o trabalho realizado tem muito potencial para ajudar os alunos a se posicionar melhor no mercado de trabalho e também a desenvolver melhor a noção de como seremos cobrados e o quê realmente as empresas e órgãos públicos que estão de mandando

Cientistas de Dados esperam de seus futuros profissionais. Apesar de utilizarmos apenas técnicas de estatística descritiva, o trabalho mostrou de forma extremamente realista a dificuldade que é lidar com grandes bases de dados, e de como pode ser complexo o trabalho de juntar essas bases em um único Sistema de Gerenciamento.

No futuro, pode ser interessante adicionar dados relacionados a outros estados e também relacionados aos partidos e gastos de campanhas dos candidatos, e compreender melhor a correlação entre esses dados, com certeza resultarão em análises mais contundentes com informações além dos resultados já publicados.

Referências

- GRUS, J. *Data Science do Zero. Primeiras Regras com o Python*. 1. ed. Rio de Janeiro: Alta Books, 2016. Citado na página [27](#).
- HAN, J.; KAMBER, M.; PEI, J. *Data Mining and Machine Learning*. 3. ed. Waltham, MA, USA: Morgan Kaufmann, 2012. Citado na página [27](#).
- LEVINE, D. M.; STEPHAN, D. F.; SZABAT, K. A. *Estatística - Teoria e Aplicações - Usando Microsoft Excel*. 7. ed. Rio de Janeiro: LTC, 2016. Citado na página [27](#).

Anexos

ANEXO A – Dicionário de Dados

A seguir estão descritos todos os bancos de dados, suas tabelas e variáveis

Pasta com arquivos csv: (https://drive.google.com/drive/folders/1Nu1wLavmZwnVFm108yWmT8JiXgTdHfm?usp=share_link)

ANEXO B – Códigos dos Programas/Notebooks Python

A seguir estão descritos todos os códigos de programas desenvolvidos ou todos os notebooks Python utilizados para gerar as análises e resultados deste trabalho

Notebook Python: https://drive.google.com/file/d/1DEfxrU5teLnUkSIRPnyCPPZHPpitfFGu/view?usp=share_link