PROJETO EM CIÊNCIA DE DADOS

SUMÁRIO

SEMESTRE	2024/2
PROJETO	Escolaridade x Popularidade
COMPONENTES DO GRUPO	Alice Cestari Colares Mykelly Costa Barros Paulo Eduardo Carvalho Mansano

Breve descrição do problema

O objetivo desta análise é entender a relação entre a escolaridade dos candidatos e sua popularidade, medida pelo número de votos recebidos nas eleições para cargos como vereador e prefeito no Rio Grande do Sul, nos anos de 2012, 2016 e 2020. A questão central é investigar se o nível de escolaridade tem influência direta nos resultados eleitorais de Porto Alegre, e como esse fator se relaciona com características como gênero, cor/raça e outros atributos dos candidatos.

Breve descrição da solução proposta

A solução proposta para este problema é realizar uma análise exploratória dos dados eleitorais, combinando múltiplas fontes de dados, incluindo resultados de votos, informações sobre candidatos e variáveis demográficas. Utilizando a metodologia CRISP-DM, vamos integrar os dados, limpar e preparar as bases, e gerar visualizações que mostrem a relação entre escolaridade e popularidade.

Resumo do que foi concluído até o momento

Até o momento, foi realizada a limpeza e preparação dos dados, com a remoção de duplicatas e o tratamento de valores ausentes. Foram gerados gráficos que mostram a evolução da participação de candidatos por gênero, grau de instrução e cor/raça ao longo dos anos de 2012, 2016 e 2020. A análise inicial sugere uma correlação positiva entre o grau de instrução e a chance de ser eleito

RELATÓRIO

Descrição das Bases de Dados

As bases de dados utilizadas possuem muitos campos, e consequentemente alguns não foram utilizados na nossa pesquisa. Utilizamos os campos DS_CARGO, NM_CANDIDATO, NM_UE e CD_GRAU_INSTRUCAO das tabelas dos candidatos para filtrar quais dos candidatos serão analisados. Já nas tabelas dos resultados, utilizamos os campos QT_VOTOS e NM_VOTAVEL para contabilizar quantos votos um candidato conseguiu, existe mais de um campo por candidato nas tabelas, o que significa que teremos de fazer somas, mas isso não prejudicará a acurácia das análises.

Justificativa

As bases de dados utilizadas neste projeto abrangem as eleições municipais do Rio Grande do Sul nos anos de 2012, 2016 e 2020, contendo informações sobre candidatos e seus desempenhos eleitorais. A base de candidatos inclui dados como nome, partido, cargo, gênero e grau de instrução, enquanto a base de resultados regista o número de votos, situação no turno e local de votação. Para contabilizar corretamente os votos de cada candidato, é necessário somar os valores de múltiplos registos relacionados ao mesmo candidato

1. Compreensão do Negócio

Background

O projeto busca entender a relação entre o grau de escolaridade dos candidatos e seus desempenhos nas eleições em Porto Alegre, utilizando bases de dados do TSE para análise histórica de tendências e padrões.

Objetivos de negócio e critérios de sucesso

Avaliar como o nível de escolaridade impacta as chances de sucesso eleitoral dos candidatos com foco nas eleições de 2012, 2016 e 2020. Critérios de sucesso incluem identificar correlações ou padrões que possam ser usados para prever tendências futuras.

Inventário de recursos

- Bases de dados do TSE (CSV de candidatos e resultados eleitorais)
- Software de análise de dados (Python, Excel e Power BI)
- Equipe com habilidades em ciência de dados e estatísticas

Requisitos, suposições e restrições

- Filtragem para Porto Alegre
- Dados precisos sobre escolaridade e resultados
- Restrições: possíveis inconsistências nos dados e limitação de informações históricas

Terminologia

- Grau de escolaridade: nível educacional dos candidatos
- Candidatos: indivíduos que concorreram a cargos políticos

Eleição: processo de votação para cargos públicos

Objetivos de mineração e critérios de sucesso

Utilizar técnicas de mineração de dados para extrair insights sobre a correlação entre o grau de

escolaridade e o sucesso eleitoral. Critérios de sucesso incluem a validação dos padrões identificados nos dados de Porto Alegre.

Plano de Projeto

1. Coleta e filtragem dos dados do TSE para Porto Alegre.
2. Limpeza e pré-processamento dos dados.
3. Análise exploratória para identificar padrões iniciais.
4. Modelagem estatística para identificar correlações e tendências.
5. Avaliação e validação dos resultados.

Avaliação inicial de técnicas e ferramentas

- Técnicas: análise de regressão, correlação, visualização de dados
- Ferramentas: Python (pandas, matplotlib), Excel, Power BI ou Notepad++ para visualização

2. Compreensão dos Dados

Esta seção descreve a compreensão inicial sobre os dados. Desde sua coleta inicial, passando por uma análise exploratória até uma avaliação de sua qualidade.

Coleta dos dados

Os dados utilizados foram adquiridos do portal de Dados Abertos do Tribunal Superior Eleitoral (TSE), na aba de Assessoria de Gestão Eleitoral (AGEL), abrangendo as eleições municipais de 2012, 2016 e 2020. Foram coletadas informações sobre candidatos e resultados eleitorais para essas três eleições. Os links para download dos datasets são:

- **Candidatos**: 2012, 2016, 2020
- **Resultados**: 2012, 2016, 2020

O principal problema enfrentado foi o tamanho dos datasets, que eram muito grandes e, muitas vezes, travavam o computador ao serem abertos no Excel, dificultando a manipulação inicial dos dados.

Descrição dos dados

*1. Base de dados dos **Candidatos**:*

Coletamos três bases de dados contendo informações dos candidatos dos anos 2020, 2016 e 2012 do estado do Rio Grande do Sul. A seguir, apresentamos todas as colunas presentes nesses datasets, acompanhadas de suas respectivas descrições e tipos de dados.

Eleição: processo de votação para cargos públicos

Objetivos de mineração e critérios de sucesso

Utilizar técnicas de mineração de dados para extrair insights sobre a correlação entre o grau de escolaridade e o sucesso eleitoral. Critérios de sucesso incluem a validação dos padrões identificados nos dados de Porto Alegre.

Plano de Projeto

6. Coleta e filtragem dos dados do TSE para Porto Alegre.
7. Limpeza e pré-processamento dos dados.
8. Análise exploratória para identificar padrões iniciais.
9. Modelagem estatística para identificar correlações e tendências.
10. Avaliação e validação dos resultados.

Avaliação inicial de técnicas e ferramentas

- Técnicas: análise de regressão, correlação, visualização de dados
- Ferramentas: Python (pandas, matplotlib), Excel, Power BI ou Notepad++ para visualização

3. Compreensão dos Dados

Esta seção descreve a compreensão inicial sobre os dados. Desde sua coleta inicial, passando por uma análise exploratória até uma avaliação de sua qualidade.

Coleta dos dados

Os dados utilizados foram adquiridos do portal de Dados Abertos do Tribunal Superior Eleitoral (TSE), na aba de Assessoria de Gestão Eleitoral (AGEL), abrangendo as eleições municipais de 2012, 2016 e 2020. Foram coletadas informações sobre candidatos e resultados eleitorais para essas três eleições. Os links para download dos datasets são:

- **Candidatos**: 2012, 2016, 2020
- **Resultados**: 2012, 2016, 2020

O principal problema enfrentado foi o tamanho dos datasets, que eram muito grandes e, muitas vezes, travavam o computador ao serem abertos no Excel, dificultando a manipulação inicial dos dados.

Descrição dos dados

*1. Base de dados dos **Candidatos**:*

Coletamos três bases de dados contendo informações dos candidatos dos anos 2020, 2016 e 2012 do estado do Rio Grande do Sul. A seguir, apresentamos todas as colunas presentes nesses datasets, acompanhadas de suas respectivas descrições e tipos de dados.

Campo	Descrição	Tipo de Dado
DT_GERACAO	Data de extração dos dados	Data
HH_GERACAO	Hora da extração dos dados (horário de Brasília)	Hora
ANO_ELEICAO	Ano da eleição de referência	Numérico
CD_TIPO_ELEICAO	Código do tipo de eleição	Numérico
NM_TIPO_ELEICAO	Nome do tipo de eleição	Texto
NR_TURNO	Número do turno da eleição	Numérico
CD_ELEICAO	Código único da eleição	Numérico
DS_ELEICAO	Descrição da eleição	Texto
DT_ELEICAO	Data da eleição	Data
TP_ABRANGENCIA_ELEICAO	Abrangência da eleição (Municipal, Estadual, Federal)	Texto
SG_UF	Sigla da unidade da federação	Texto
SG_UE	Sigla da unidade eleitoral	Texto
NM_UE	Nome da unidade eleitoral	Texto
CD_CARGO	Código do cargo disputado	Numérico
DS_CARGO	Descrição do cargo disputado	Texto
SQ_CANDIDATO	Identificação sequencial do candidato	Numérico
NR_CANDIDATO	Número do candidato na urna	Numérico
NM_CANDIDATO	Nome completo do candidato	Texto
NM_URNA_CANDIDATO	Nome do candidato na urna	Texto
NM_SOCIAL_CANDIDATO	Nome social (se aplicável)	Texto
NR_CPF_CANDIDATO	CPF do candidato	Texto
DS_EMAIL	E-mail do candidato	Texto
CD_SITUACAO_CANDIDATURA	Código da situação de candidatura	Numérico
DS_SITUACAO_CANDIDATURA	Descrição da situação de candidatura	Texto
TP_AGREMIACAO	Forma de candidatura (Partido, Coligação, Federação)	Texto
NR_PARTIDO	Número do partido	Numérico
SG_PARTIDO	Sigla do partido	Texto
NM_PARTIDO	Nome do partido	Texto
NR_FEDERACAO	Número da federação	Numérico
NM_FEDERACAO	Nome da federação	Texto
SG_FEDERACAO	Sigla da federação	Texto
DS_COMPOSICAO_FEDERACAO	Composição da federação (Siglas dos partidos)	Texto
SQ_COLIGACAO	Número sequencial da coligação	Numérico
NM_COLIGACAO	Nome da coligação	Texto
DS_COMPOSICAO_COLIGACAO	Composição da coligação	Texto

SG_UF_NASCIMENTO	Sigla do estado de nascimento	Texto
DT_NASCIMENTO	Data de nascimento do candidato	Data
CD_GENERO	Código do gênero	Numérico
DS_GENERO	Descrição do gênero	Texto
CD_GRAU_INSTRUCAO	Código do grau de instrução	Numérico
DS_GRAU_INSTRUCAO	Descrição do grau de instrução	Texto
CD_ESTADO_CIVIL	Código do estado civil	Numérico
DS_ESTADO_CIVIL	Descrição do estado civil	Texto
CD_COR_RACA	Código da cor/raça	Numérico
DS_COR_RACA	Descrição da cor/raça	Texto
CD_OCUPACAO	Código da ocupação	Numérico
DS_OCUPACAO	Descrição da ocupação	Texto
CD_SIT_TOT_TURNO	Código da situação no turno	Numérico
DS_SIT_TOT_TURNO	Descrição da situação no turno	Texto

*Base de dados dos **Resultados** das eleições:*

Coletamos três bases de dados contendo informações dos resultados das votações dos anos 2020, 2016 e 2012 do estado do Rio Grande do Sul. A seguir, apresentamos todas as colunas presentes nesses datasets, acompanhadas de suas respectivas descrições e tipos de dados.

Campo	**Descrição**	**Tipo de Dado**
DT_GERACAO	Data de extração dos dados	Data
HH_GERACAO	Hora de extração dos dados (horário de Brasília)	Hora
ANO_ELEICAO	Ano da eleição de referência	Numérico
CD_TIPO_ELEICAO	Código do tipo de eleição	Numérico
NM_TIPO_ELEICAO	Nome do tipo de eleição	Texto
NR_TURNO	Número do turno da eleição	Numérico
CD_ELEICAO	Código da eleição	Numérico
DS_ELEICAO	Descrição da eleição	Texto
DT_ELEICAO	Data em que ocorreu a eleição	Data
TP_ABRANGENCIA	Tipo de abrangência da eleição (Municipal, Estadual, Federal)	Texto
SG_UF	Sigla da Unidade da Federação	Texto
SG_UE	Sigla da Unidade Eleitoral	Texto
NM_UE	Nome da Unidade Eleitoral	Texto
CD_MUNICIPIO	Código TSE do município	Numérico
NM_MUNICIPIO	Nome do município	Texto
NR_ZONA	Número da Zona Eleitoral	Numérico
CD_CARGO	Código do cargo	Numérico
DS_CARGO	Descrição do cargo	Texto
SQ_CANDIDATO	Identificação sequencial do candidato	Numérico
NR_CANDIDATO	Número do candidato na urna	Numérico

NM_CANDIDATO	Nome completo do candidato	Texto
NM_URNA_CANDIDATO	Nome do candidato na urna	Texto
NM_SOCIAL_CANDIDATO	Nome social do candidato (se aplicável)	Texto
CD_SITUACAO_CANDIDATURA	Código da situação do registro de candidatura	Numérico
DS_SITUACAO_CANDIDATURA	Descrição da situação do registro	Texto
CD_DETALHE_SITUACAO_CAND	Código do detalhe da situação do registro	Numérico
DS_DETALHE_SITUACAO_CAND	Descrição do detalhe da situação	Texto
TP_AGREMIACAO	Tipo de agremiação (Coligação, Partido Isolado)	Texto
NR_PARTIDO	Número do partido	Numérico
SG_PARTIDO	Sigla do partido	Texto
NM_PARTIDO	Nome do partido	Texto
SQ_COLIGACAO	Sequencial da coligação	Numérico
NM_COLIGACAO	Nome da coligação	Texto
DS_COMPOSICAO_COLIGACAO	Composição da coligação	Texto
CD_SIT_TOT_TURNO	Código da situação de totalização no turno	Numérico
DS_SIT_TOT_TURNO	Descrição da situação de totalização no turno	Texto
ST_VOTO_EM_TRANSITO	Indicativo de voto em trânsito	Texto
QT_VOTOS_NOMINAIS	Quantidade total de votos nominais recebidos pelo candidato	Numérico

Dados Especiais

- **Campos Nulos**: Preenchidos com #NULO (valor textual) ou -1 (valor numérico).
- **Dados Não Existentes**: Representados por #NE (para valores textuais) ou -3 (para valores numéricos).
- **Campos Não Divulgáveis**: Em campos textuais, aparecem como "NÃO DIVULGÁVEL" e, para numéricos, como -4.

Análise exploratória dos dados

Esta análise exploratória abrange as eleições municipais de Porto Alegre nos anos de 2012, 2016 e 2020, com foco nos candidatos, cargos disputados, partidos, coligações e desempenho eleitoral. O objetivo é identificar padrões na distribuição de votos, explorando a concentração em poucos candidatos e o impacto das coligações e partidos no desempenho eleitoral.

Foram calculadas média, mediana, desvio padrão, valores mínimo e máximo dos votos. Observamos que em todos os anos existem candidatos com 0 votos, algo comum em eleições proporcionais, mas que será tratado posteriormente. A diferença significativa entre média e mediana sugere uma concentração de votos em um pequeno grupo de candidatos.

Dados estáticos dos datasets de resultados:

Ano	Média	Mediana	Desvio Padrão	Valor Mínimo	Valor Máximo
2012	253,22	26	2.286,48	0	62.890
2016	356,24	31	2.489,40	0	48.953
2020	224,83	18	1.981,28	0	46.224

Para simplificar a análise, mantivemos apenas as colunas essenciais: ANO_ELEICAO, CD_TIPO_ELEICAO, NR_TURNO, CD_CARGO, DS_CARGO, NR_CANDIDATO, NM_CANDIDATO, SQ_CANDIDATO, CD_SITUACAO_CANDIDATURA, SG_PARTIDO, NM_PARTIDO, CD_SIT_TOT_TURNO, QT_VOTOS_NOMINAIS. Foram identificados valores faltantes em DT_NASCIMENTO (2012 e 2016) e NM_URNA_CANDIDATO (2020), o que impacta análises relacionadas à idade e identificação, mas será tratado com imputação adequada.

Durante o pré-processamento, colunas consideradas irrelevantes, como CD_MUNICIPIO, NR_ZONA e SQ_CANDIDATO, serão removidas para simplificar o conjunto de dados. Candidatos com zero votos serão excluídos das análises principais. Não será necessária a normalização dos votos em QT_VOTOS_NOMINAIS, pois a escala dos valores já é adequada. As comparações entre os anos serão feitas utilizando proporções de votos para garantir consistência na análise.

Verificação de qualidade dos dados

Para verificar a qualidade dos dados, realizamos uma análise manual no Excel, aplicando filtros nas colunas para identificar e isolar células vazias. Embora esse processo tenha garantido uma identificação precisa de registros incompletos, ele se mostrou bastante demorado. No futuro, pretendemos automatizar essa verificação com Python para torná-la mais eficiente e menos suscetível a erros manuais.

Datasets de Candidatos (2012, 2016 e 2020)

- Valores Faltantes:
  - 2012: Na análise manual no Excel, encontramos algumas células vazias na coluna DT_NASCIMENTO, totalizando 38 registros ausentes.
  - 2016: Apenas 1 registro ausente em DT_NASCIMENTO e NR_IDADE_DATA_POSSE.
  - 2020: 1 registro ausente em NM_URNA_CANDIDATO, que pode ser corrigido usando o nome completo do candidato.
- Erros de Tipagem:
  - O dataset de 2012 apresentou tipos de dados mistos, indicando

inconsistência na entrada de dados.

- Registros Duplicados:
  - Não encontramos duplicatas até o momento, mas será feita uma revisão final para garantir a consistência.

Datasets de Resultados (2012, 2016 e 2020)

- Valores Faltantes:
  - Nenhum valor ausente identificado até o momento

Preparação dos Dados

A etapa de preparação dos dados foi complexa e desafiadora. Nosso grupo optou por focar na filtragem das linhas dos arquivos CSV, realizando o processo em duas etapas:

1. Filtragem dos Candidatos Votados em Porto Alegre (Tabela de Resultados)

Utilizamos um código simples em Python para identificar e extrair apenas os candidatos que foram votados na cidade de Porto Alegre. O código verificava o campo "NM_MUNICIPIO" e, caso o valor fosse "Porto Alegre", a linha correspondente era adicionada a um novo arquivo. Ao final do processo, um arquivo CSV contendo apenas os candidatos votados em Porto Alegre foi gerado.

2. Filtragem das Informações dos Candidatos (Tabela de Candidatos)

Nesta segunda etapa, o código Python utilizado foi mais complexo devido a problemas encontrados nos arquivos dos candidatos de 2020. O objetivo era selecionar apenas os candidatos de Porto Alegre utilizando a variável "SQ_CANDIDATO" e salvar as informações correspondentes em um novo arquivo CSV.

Durante o processo, encontramos uma inconsistência nos dados do dataset de 2020. Nos datasets de 2012 e 2016, o campo "SQ_CANDIDATO" estava entre aspas duplas, o que permitiu tratá-lo como uma string. No entanto, em 2020, esse mesmo campo estava armazenado como um número inteiro, o que causou erros na execução do código.

Limpeza dos dados

Segue abaixo o detalhamento do processo de limpeza de dados e as decisões tomadas durante essa fase, incluindo a seleção de features, o tratamento de dados faltantes e as transformações realizadas.

1. Transformação para UTF-8

A primeira etapa foi garantir que todos os arquivos CSV fossem lidos corretamente, independentemente da codificação original. Durante a análise, verificamos que diferentes arquivos vinham com codificações distintas, o que poderia causar problemas de leitura, especialmente com caracteres especiais ou acentuação. Para resolver isso, realizamos a conversão de todos os arquivos para o formato **UTF-8**, com o intuito de assegurar que os dados textuais, como nomes de candidatos e cargos, sejam lidos corretamente.

2. Seleção de Colunas Relevantes

Nessa etapa a ideia foi manter apenas as informações que realmente contribuiriam para os objetivos da análise preditiva, garantindo que o modelo fosse eficiente e preciso. As **features** escolhidas para o dataset final foram:

- **SQ_CANDIDATO**: Identificador único de cada candidato.
- **ANO_ELEICAO**: Ano da eleição, uma variável importante para segmentar os dados por ano e comparar os desempenhos dos candidatos ao longo do tempo.
- **NM_CANDIDATO**: Nome completo do candidato, necessário para identificação.
- **DS_CARGO**: Cargo que o candidato estava disputando (ex: Vereador, Prefeito).
- **DS_GENERO**: Gênero do candidato (Masculino, Feminino), que foi utilizado como uma variável categórica na modelagem.
- **DS_GRAU_INSTRUCAO**: Grau de instrução do candidato (ex: Superior Completo, Ensino Médio), uma variável relevante para o desempenho eleitoral e análise de perfil.
- **QT_VOTOS_NOMINAIS**: Quantidade de votos nominais recebidos, a variável central para avaliar o desempenho do candidato.

Foram removidas colunas irrelevantes ou que não agregavam valor à análise, como dados auxiliares ou informações que não eram necessárias para a análise preditiva, como o nome da zona eleitoral ou dados de votação em zonas que não eram relevantes para Porto Alegre.

3. Tratamento de Valores Nulos

Durante a limpeza, dados faltantes foram identificados nas colunas essenciais, como o número de votos (**QT_VOTOS_NOMINAIS**) ou o nome do candidato (**NM_CANDIDATO**). Como esses dados são fundamentais para a modelagem, tomamos a decisão de **remover as linhas** que apresentavam valores nulos nessas colunas essenciais. A remoção foi preferida em vez de tentar imputar valores, pois a ausência de informações cruciais (como o nome de um candidato ou o total de votos) comprometeria a análise.

Em contrapartida, variáveis secundárias que não impactavam diretamente a análise, mas que apresentavam dados faltantes, foram tratadas de forma diferente, dependendo do contexto. Por exemplo, se uma variável não era essencial para o modelo ou se a quantidade de dados faltantes era mínima, essa feature foi mantida, mas com valores imutáveis (como "Desconhecido" ou uma média da coluna, se aplicável).

4. Tratamento de Valores Negativos

Ao inspecionar os dados, identificamos valores negativos na coluna de votos (QT_VOTOS_NOMINAIS). Como isso não faz sentido (um candidato não pode ter votos negativos), tomamos a decisão de substituir todos os valores negativos por zero. Essa transformação foi feita para garantir que os dados fossem válidos e que valores atípicos ou errôneos não impactassem a análise. Para as demais colunas, valores fora do esperado também foram tratados de maneira similar, garantindo a consistência dos dados.

5. Conversão para Maiúsculas

A padronização textual foi outro passo importante para evitar discrepâncias nas variáveis categóricas. Constatamos que algumas variáveis textuais, como o nome dos candidatos e a

descrição de cargo, apresentavam variações de capitalização (ex: "Carlos" vs. "carlos"). Para evitar esse tipo de inconsistência, todas as strings foram convertidas para maiúsculas.

6. Filtragem para Porto Alegre e Cargos Específicos

Como a análise tinha como foco Porto Alegre, uma filtragem dos dados foi realizada para incluir apenas os registros dos candidatos dessa cidade que estavam concorrendo aos cargos de Vereador ou Prefeito. Esse passo foi essencial, pois dados de outros municípios e cargos não contribuíam para a análise específica que estava sendo realizada. Dessa forma, as informações irrelevantes foram descartadas, e garantimos que o dataset fosse direcionado e específico.

7. Remoção de Duplicatas

Durante a verificação de consistência, foi detectado que alguns candidatos estavam registrados várias vezes, especialmente em contextos de diferentes zonas eleitorais. Como não faria sentido manter múltiplos registros para o mesmo candidato, foi realizada a remoção de duplicatas. Mantivemos apenas a primeira ocorrência de cada candidato no dataset, o que garantiu que cada linha representasse de forma única um candidato, sem duplicação de informações.

8. Integração dos Dados de Candidatos e Votação

Por fim, a integração dos **dados de candidatos e votação** foi essencial para consolidar todas as informações relevantes em um único dataset. Utilizando o identificador único **SQ_CANDIDATO**, combinamos os dados de votos com as informações demográficas e de cargo de cada candidato. Essa integração permitiu a construção de um dataset completo e com informações suficientes para a realização da modelagem preditiva.

Criação de atributos e registros

Durante o processo de análise e preparação dos dados, foram criados atributos e registros para enriquecer as informações e possibilitar análises mais detalhadas. A seguir, estão descritas as etapas de criação de atributos e o raciocínio por trás de cada um deles.

Criação de Atributos

1. Percentual de Votos (PERC_VOTOS)

Objetivo: Calcular o percentual de votos que cada candidato obteve em relação ao total de votos válidos na eleição.

Método: O percentual de votos foi calculado pela fórmula:

$$PERC\_VOTOS = total\_votos(QT\_VOTOS\_NOMINAIS) \times 100$$

Onde QT_VOTOS_NOMINAIS representa o número de votos recebidos pelo candidato, e total_votos é a soma total de votos válidos na eleição. Esse cálculo foi realizado para cada ano (2012, 2016, 2020), criando o atributo PERC_VOTOS.

2. Classificação por Desempenho Eleitoral (CLASSIFICACAO)

Objetivo: Classificar os candidatos de acordo com o número de votos recebidos.
Método: A classificação foi realizada utilizando o método rank() do pandas. Os candidatos com maior número de votos receberam a classificação mais alta (rank 1, rank 2, etc.). O parâmetro method="min" foi utilizado para garantir que candidatos com votos iguais recebessem a mesma classificação.

3. Ano da Eleição (ANO_ELEICAO)

Objetivo: Identificar a qual eleição cada linha dos dados pertence.
Método: A coluna **ANO_ELEICAO** foi adicionada manualmente para cada conjunto de dados (2012, 2016, 2020), permitindo que o dataset consolidado fosse segmentado por ano.

4. Candidatos Eleitos

Objetivo: Filtrar os candidatos que foram eleitos em cada eleição.
Método: A coluna **DS_SIT_TOT_TURNO** foi utilizada para filtrar apenas os candidatos que estavam na situação de "ELEITO", "ELEITO POR MÉDIA" ou "ELEITO POR QP". Isso permitiu analisar exclusivamente os candidatos eleitos.

5. Distribuição de Votos por Gênero e Grau de Instrução

Objetivo: Analisar a distribuição dos votos entre os candidatos eleitos com base em gênero e grau de instrução.
Método: As variáveis **DS_GENERO** (gênero) e **DS_GRAU_INSTRUCAO** (grau de instrução) foram usadas para agrupar os dados e somar os votos de cada grupo. Isso permitiu identificar padrões relacionados ao gênero e ao grau de instrução dos candidatos eleitos.

6. Proporção de Eleitos por Grau de Instrução

Objetivo: Analisar a relação entre o grau de instrução dos candidatos e o sucesso eleitoral.
Método: O percentual de eleitos por grau de instrução foi calculado com base no total de candidatos e no número de eleitos de cada nível educacional. Esse cálculo ajudou a observar qual grau de instrução tem uma maior taxa de sucesso eleitoral.

Criação de Registros (Instâncias de Dados)

Não houve a criação de novos registros ou instâncias de dados a partir de dados externos. No entanto, ao integrar os datasets de candidatos e votação para os três anos (2012, 2016, 2020), as instâncias existentes foram enriquecidas com informações adicionais, como a quantidade de votos e percentual de votos, criando uma instância única para cada candidato, mas sem a criação de novos dados.

Integração de dados

Durante o processo de integração dos dados, enfrentamos um desafio relacionado à duplicação de registros. As bases de dados de resultados das eleições (2012, 2016 e 2020) apresentavam o mesmo identificador único para os candidatos (**SQ_CANDIDATO**), o que resultava na repetição de registros. Isso ocorreu porque os candidatos eram registrados múltiplas vezes nas diferentes zonas eleitorais nas quais concorriam, gerando duplicação de informações nos dataframes.

Esse comportamento era esperado, dado que um mesmo candidato pode atuar em mais de uma zona eleitoral em uma eleição, o que faz com que o candidato apareça em múltiplos registros dentro das bases de resultados.

Desafios Encontrados

Duplicação de registros:

A principal dificuldade foi a repetição de registros de candidatos,uma vez que cada vez que um candidato concorria em uma zona eleitoral, um novo registro era gerado nas bases de resultados.

Impacto nas análises:

A duplicação de registros poderia comprometer a integridade das análises subsequentes, uma vez que isso afetaria a contagem total de votos e a geração de métricas agregadas, como o percentual de votos. Além disso, poderia prejudicar a precisão dos modelos preditivos, ao fazer com que um mesmo candidato fosse representado múltiplas vezes.

Estratégia de Resolução

Para tratar a duplicação de registros, foi utilizado o SQ_CANDIDATO como chave primária para o agrupamento dos dados. A estratégia consistiu em somar os votos nominais de cada candidato, de modo que, ao final do processo, cada candidato tivesse apenas um registro consolidado com o total de votos recebidos.

O processo de agregação foi implementado conforme os seguintes passos:

- Agrupamento das bases de dados pela coluna SQ_CANDIDATO, identificando registros duplicados.
- Soma dos votos nominais (QT_VOTOS_NOMINAIS) para cada candidato, consolidando a informação de todas as zonas eleitorais nas quais ele concorreu.

Esse procedimento garantiu que, após a agregação, cada candidato fosse representado por uma única linha no dataset final, com a soma total de seus votos.

```
cand_votos_2012 = cand_2012.merge(votos_2012.groupby("SQ_CANDIDATO")["QT_VOTOS_NOMINAIS"].sum().reset_index(), on="SQ_CANDIDATO", how="inner")
cand_votos_2016 = cand_2016.merge(votos_2016.groupby("SQ_CANDIDATO")["QT_VOTOS_NOMINAIS"].sum().reset_index(), on="SQ_CANDIDATO", how="inner")
cand_votos_2020 =
```

```
cand_2020.merge(votos_2020.groupby("SQ_CANDIDATO")
["QT_VOTOS_NOMINAIS"].sum().reset_index(),  on="SQ_CANDIDATO",
how="inner")
```

Além disso, a integração das bases de dados de candidatos e de resultados de votação foi realizada utilizando a coluna **SQ_CANDIDATO** como chave de junção, o que assegurou que as informações dos candidatos e seus respectivos votos fossem corretamente combinadas. A operação foi executada por meio do método merge do pandas, proporcionando uma junção eficiente dos datasets.

Reaproveitamento de Features e Eliminação de Redundâncias

Durante a integração, algumas features presentes nas duas bases (candidatos e resultados de votação) foram reaproveitadas, incluindo:

- **SQ_CANDIDATO**: Identificador único do candidato.
- **ANO_ELEICAO**: Ano da eleição.
- **NM_CANDIDATO**: Nome do candidato.
- **DS_CARGO**: Cargo do candidato.
- **QT_VOTOS_NOMINAIS**: Quantidade de votos recebidos pelo candidato.

Após a junção, as colunas duplicadas foram tratadas, e mantivemos apenas uma instância de cada coluna no dataset final, eliminando redundâncias. Além disso, a duplicação de registros foi resolvida pela soma dos votos, garantindo que cada candidato tivesse apenas um registro com o total de votos recebidos, independentemente das zonas eleitorais em que concorreu.

Descrição do Dataset Final

Após a conclusão das etapas de pré-processamento e integração, o dataset final contém os dados consolidados para a análise das eleições de 2012, 2016 e 2020. Este dataset foi preparado para ser utilizado emdeanálises exploratórias mais detalhadas.

Estrutura do Dataset Final

O dataset final inclui as seguintes colunas:

- **ANO_ELEICAO**: Ano da eleição (2012, 2016, 2020).
- **SQ_CANDIDATO**: Identificador único do candidato.
- **NM_CANDIDATO**: Nome do candidato.
- **DS_CARGO**: Cargo do candidato.
- **DS_GENERO**: Gênero do candidato (ex: "MASCULINO", "FEMININO").
- **DS_COR_RACA**: Cor/Raça do candidato (ex: "BRANCA", "PRETA").
- **DS_GRAU_INSTRUCAO**: Grau de instrução do candidato (ex: "SUPERIOR COMPLETO", "ENSINO FUNDAMENTAL").
- **DS_SIT_TOT_TURNO**: Situação do candidato no turno (ex: "ELEITO", "SUPLENTE").
- **QT_VOTOS_NOMINAIS**: Quantidade de votos nominais recebidos pelo candidato.
- **PERC_VOTOS**: Percentual de votos recebidos pelo candidato em relação ao total de votos válidos da eleição.

- **CLASSIFICACAO**: Classificação do candidato, baseada no número de votos recebidos.

Estado do Dataset Final Após Pré-processamento

O dataset final está preparado para a etapa de modelagem preditiva ou análise estatística com as seguintes características:

- Dados completos, sem valores nulos nas colunas essenciais.
- Atributos derivados: Foi calculado o PERC_VOTOS, a CLASSIFICACAO dos candidatos, e identificados os candidatos eleitos.
- Redundâncias eliminadas: A duplicação de registros foi resolvida, garantindo que cada candidato tenha um único registro com o total de votos recebidos.
- Formato adequado para análise e modelagem: O dataset contém tanto variáveis numéricas (votos, percentuais) quanto categóricas (gênero, grau de instrução, cargo), facilitando sua utilização em modelos preditivos e outras análises quantitativas.

Visualização de Dados

A visualização de dados possibilitou a análise dos padrões nos dados eleitorais de Porto Alegre (2012, 2016 e 2020), dentro do modelo CRISP-DM. Essa etapa focou em variáveis como gênero, grau de instrução, cargo e votos nominais, facilitando a interpretação dos resultados e a preparação para as fases seguintes.

Distribuição de Votos Nominais por Ano (Boxplot)

- O boxplot mostrou a variabilidade dos votos nominais ao longo dos anos, identificando outliers e mudanças no padrão de votos.

Distribuição de Votos por Gênero (Gráfico de Barras)

- A visualização dos votos entre os candidatos eleitos, segmentados por gênero, revelou como a participação feminina evoluiu ao longo das eleições.

Distribuição de Candidatos por Cor/Raça (Gráfico de Contagem)

- O gráfico de contagem apresentou a diversidade racial entre os candidatos, destacando as variações na representatividade ao longo dos anos.

Top 5 Candidatos com Mais Votos por Ano (Gráfico de Barras)

- A análise dos 5 candidatos mais votados por ano forneceu uma visão clara sobre os principais candidatos para os cargos de Vereador e Prefeito.

Top 5 Candidatos com Mais Votos e Grau de Instrução (Gráfico de Barras)

- A relação entre grau de instrução e votos mostrou que candidatos com nível superior tendem a receber mais votos, indicando uma possível influência da educação no sucesso eleitoral.

Distribuição de Votos por Ano e Cargo (Boxplot)

- O boxplot comparou a distribuição de votos entre os cargos de Vereador e Prefeito,

destacando as diferenças nas dinâmicas eleitorais para cada cargo.

Percentual de Votos dos Candidatos Eleitos

- A visualização do percentual de votos revelou os candidatos mais bem sucedidos em cada eleição, proporcionando uma visão sobre a competitividade eleitoral.

Proporção de Eleitos por Grau de Instrução

- A análise da proporção de eleitos por grau de instrução mostrou que candidatos com ensino superior completo tiveram maior taxa de sucesso eleitoral

Avaliação

Nesta seção, os resultados obtidos são avaliados em relação aos objetivos de negócio definidos na fase inicial do projeto. A análise é realizada com base na comparação entre os critérios de sucesso previamente estabelecidos e os resultados observados nos dados.

Avaliação dos Resultados

Os dados indicaram que o nível de escolaridade é um fator determinante para o sucesso eleitoral, o que valida um dos objetivos principais do projeto. A análise revelou, além disso, disparidades de gênero e raça, que não foram inicialmente previstas, mas que se mostraram relevantes para a compreensão das desigualdades no processo eleitoral. Essas variáveis forneceram uma perspectiva adicional sobre a representatividade, reforçando a importância de outros fatores sociais na dinâmica eleitoral.

De maneira geral, os resultados confirmaram os objetivos do projeto, especialmente no que diz respeito à relação entre escolaridade e êxito eleitoral. A identificação das desigualdades de gênero e raça contribuiu para uma análise mais abrangente dos determinantes do sucesso nas eleições, representando um achado adicional significativo.

Revisão do Processo e Conclusões

Os objetivos iniciais foram alcançados com sucesso, porém, a análise revelou variáveis não previstas, como as disparidades de gênero e raça, que indicam a necessidade de revisar os critérios de negócio para incorporar essas novas dimensões. A inclusão de políticas mais inclusivas pode ser um objetivo futuro, visando a redução dessas desigualdades e a promoção de maior representatividade.

Para projetos subsequentes, recomenda-se a expansão do escopo da análise, incluindo outros fatores sociais e econômicos que possam influenciar os resultados eleitorais. Além disso, a modelagem preditiva poderia ser aprimorada, utilizando um conjunto mais amplo de variáveis para estimar com maior precisão as probabilidades de sucesso eleitoral de novos candidatos

Autocrítica

O trabalho seguiu a metodologia CRISP-DM e atingiu com sucesso as etapas de entendimento do negócio, dados, pré-processamento e análise exploratória. Identificamos padrões como a relação entre escolaridade e sucesso eleitoral e exploramos desigualdades de gênero e raça nas eleições. As visualizações de dados

foram eficazes e trouxeram contribuições importantes. Embora o modelo preditivo não tenha sido implementado, ele está planejado para etapas futuras.

Nota: 9.5