Análise de dados Campeonato Brasileiro 2003 a 2022

Nicolas Souza Ferreira, Eryck da Silva, Luciano Moraes
FACSENAC-DF, Brasília-DF
erycksr@gmail.com,
lucianomoraes1980@gmail.com,
nicksfbsb@gmail.com

Abstract:. This research project aims to conduct a comprehensive analysis of data from the Brazilian Football Championship over two decades, exploring detailed statistics to unveil patterns and trends. The investigation will encompass aspects such as overall team performance, individual player statistics, efficiency in shots and passes, on-field discipline, and comparative analysis between specific periods. By highlighting geography as an influential factor, the project seeks correlations between team performance and their location. The objective is to provide valuable insights into successful strategies, challenges faced by teams, and the evolution of football in Brazil, contributing to an in-depth understanding of the sport in the country.

Resumo.
Esta análise aprofundada dos dados do Campeonato Brasileiro busca revelar nuances cruciais sobre o desempenho das equipes nas últimas duas décadas. Exploraremos estatísticas detalhadas, desde a eficiência nos chutes e passes até a disciplina em campo, identificando padrões e tendências ao longo do tempo. A análise individual dos jogadores e a comparação entre equipes proporcionarão insights sobre estratégias bem-sucedidas e desafios enfrentados. Além disso, a abordagem geográfica poderá revelar correlações entre o desempenho das equipes e sua localização. Esta jornada visa desvendar os segredos e histórias ocultos nos números do esporte mais amado do Brasil, contribuindo para uma compreensão mais profunda da evolução do futebol no país.

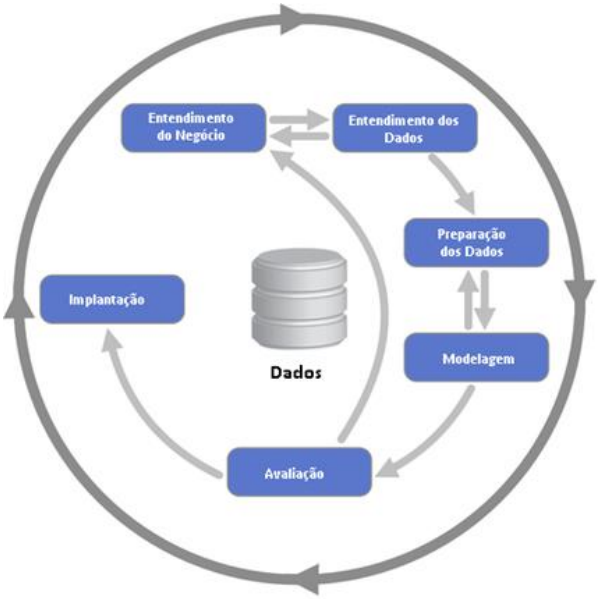
1. Introdução

Em uma era cada vez mais orientada por dados, os esportes, incluindo o futebol, destacam-se como uma área cativante para exploração. Além da emoção visceral das partidas, as competições de futebol abrigam um tesouro de informações valiosas prontas para serem estudadas e analisadas, revelando tendências, estratégias e insights sobre o desempenho tanto das equipes quanto dos jogadores. O Campeonato Brasileiro de Futebol, como uma potência entre os eventos globais de futebol, serve como uma fonte abundante de dados. Este artigo mergulha no extenso conjunto de dados do Campeonato Brasileiro de Futebol, abrangendo o período de 2003 a 2022. Esses dados, meticulosamente selecionados e validados a partir de fontes confiáveis, agora estão acessíveis em um repositório público. Nosso objetivo é realizar uma análise aprofundada das estatísticas de cada partida, desde o número de chutes e precisão de passes até as faltas cometidas e cartões distribuídos. Além disso, buscamos rastrear o desempenho dos clubes ao longo do tempo, identificando tendências e padrões que ofereçam insights sobre estratégias bem-sucedidas e desafios enfrentados pelas equipes. No cerne desse empreendimento está a convicção de que o conjunto de dados do Campeonato Brasileiro de Futebol representa um tesouro de informações para analistas, pesquisadores e entusiastas fervorosos do futebol. Ao aprofundarmos a compreensão das narrativas tecidas por esses números, aspiramos enriquecer nosso entendimento da dinâmica evolutiva do esporte no Brasil ao longo dessas duas décadas. Junte-se a nós nesta jornada esclarecedora enquanto navegamos pela intrincada paisagem de dados do Campeonato Brasileiro de Futebol.

2. Metodologia e Referencial teórico

A CRISP-DM (Cross-Industry Standard Process for Data Mining) é uma metodologia amplamente reconhecida para projetos de mineração de dados. Seus principais objetivos são proporcionar uma abordagem estruturada e sistemática para o desenvolvimento de projetos de mineração de dados, garantindo a eficiência e a eficácia do processo. A metodologia é composta por seis fases principais: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implementação. Cada fase aborda aspectos específicos do ciclo de vida do projeto, desde a compreensão dos objetivos de negócios até a implementação e monitoramento do modelo desenvolvido. A CRISP-DM fornece uma estrutura flexível que pode ser adaptada a diferentes contextos e tipos de projetos de mineração de dados, sendo amplamente utilizada na indústria e na academia.

2.1. CRISP-DM



MySQL Workbench: Uma ferramenta de desenvolvimento integrado para bancos de dados MySQL, proporcionando funcionalidades de modelagem, administração e consulta.

Power BI: Plataforma da Microsoft especializada em visualização de dados, transformando informações brutas em insights visuais interativos.

Notion: Ferramenta versátil de organização e colaboração, utilizada para criar documentos, gerenciar projetos e tomar notas de maneira eficiente.

Google Colab: Plataforma baseada em nuvem para execução de códigos em Python, especialmente útil para análise de dados e implementação de algoritmos de aprendizado de máquina.

Python: Linguagem de programação versátil e popular, amplamente empregada em ciência de dados, análise estatística e desenvolvimento de aplicações.

Matplotlib: Biblioteca em Python utilizada para criar visualizações estáticas, gráficos e plots, contribuindo para a apresentação eficaz de dados.

Pandas: Biblioteca em Python que fornece estruturas de dados flexíveis e eficientes para manipulação e análise de dados, facilitando a limpeza e transformação de conjuntos de dados.

Seaborn: Baseada em Matplotlib, Seaborn é uma biblioteca Python que facilita a criação de gráficos estatísticos atraentes e informativos.

Essas tecnologias são essenciais para a realização de projetos abrangentes de ciência de dados, desde a gestão de bancos de dados até a visualização e interpretação eficaz dos resultados. A combinação destas ferramentas oferece uma abordagem completa para explorar, analisar e comunicar informações a partir de conjuntos de dados complexos.

3. Resultados

Este estudo emprega a metodologia CRISP-DM para analisar o desempenho no Campeonato Brasileiro de Futebol. Iniciando com a compreensão do negócio e entendimento dos dados, avançamos para a preparação e modelagem, aplicando um modelo de classificação. A avaliação dos resultados oferece insights valiosos sobre estratégias e desafios enfrentados pelas equipes. Essa abordagem estruturada busca assegurar a reprodutibilidade do estudo, permitindo que profissionais possam replicar e validar os resultados, promovendo transparência e confiabilidade no processo analítico.

3.1. Compreensão do Negócio

A fase de Compreensão do Negócio neste estudo desempenha um papel inicial ao direcionar nossa análise específica para o desempenho dos clubes e jogadores no contexto do Campeonato Brasileiro de Futebol. Em sintonia com o foco deste trabalho, nossa abordagem busca não apenas entender, mas aprofundar-se nas métricas essenciais e nos fatores determinantes que moldam o cenário competitivo ao longo de duas décadas de intensa competição.

Ao mergulhar nos objetivos intrínsecos do campeonato, almejamos identificar padrões consistentes e discrepâncias que definem o sucesso ou os desafios enfrentados por equipes e jogadores. A análise meticulosa das dinâmicas do desempenho, considerando variáveis como gols marcados, assistências, disciplina tática e outros indicadores relevantes, proporciona um panorama abrangente que serve como base sólida para nossas investigações subsequentes.

A compreensão refinada do negócio não se limita apenas à interpretação de estatísticas; ela se estende à captura da essência competitiva do Campeonato Brasileiro. Compreender as nuances dos momentos cruciais, as estratégias de sucesso e as adversidades

enfrentadas pelos clubes e jogadores contribui para uma análise mais holística do desempenho no cenário futebolístico nacional.

Nossa abordagem busca, assim, ir além dos números, desvendando histórias e tendências que pintam um retrato mais vívido do esporte no Brasil. Ao centrar nossa compreensão do negócio nas particularidades do desempenho no Campeonato Brasileiro, buscamos oferecer uma contribuição significativa para o entendimento aprofundado do cenário esportivo brasileiro ao longo das últimas duas décadas.

3.2. Entendimento de Dados

Este estudo se propõe a realizar uma análise aprofundada do desempenho no Campeonato Brasileiro de Futebol, utilizando três conjuntos de dados fundamentais. Os arquivos CSV "campeonato-brasileiro-full.csv", "campeonato-brasileiro-estatisticas-full.csv" e "campeonato-brasileiro-gols.csv" oferecem uma visão abrangente, cobrindo desde os aspectos gerais da competição até estatísticas específicas de jogadores e detalhes sobre os gols marcados. Vamos explorar cada conjunto de dados para extrair insights valiosos sobre a evolução do futebol brasileiro ao longo das temporadas.

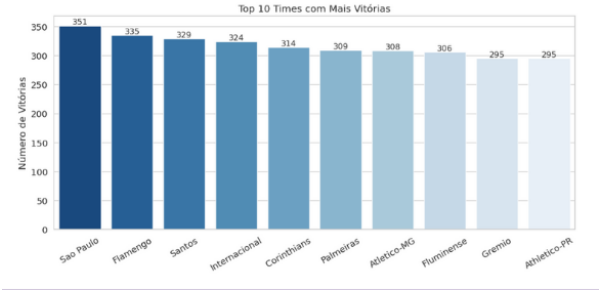
campeonato-brasileiro-full.csv:
Este arquivo serve como o ponto de partida, proporcionando uma visão global do Campeonato Brasileiro. Contém informações cruciais sobre as equipes participantes, resultados das partidas, pontuações e outros elementos essenciais para entender a dinâmica da competição ao longo do tempo.

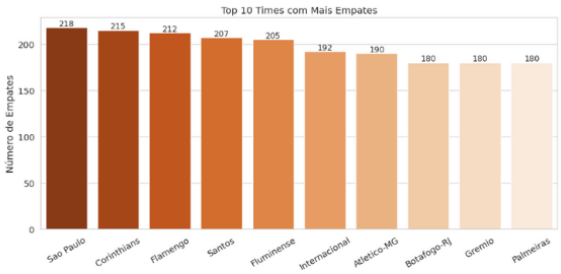
campeonato-brasileiro-estatisticas-full.csv:
Este conjunto de dados oferece uma camada mais detalhada, fornecendo estatísticas específicas relacionadas ao desempenho de jogadores e equipes. Com dados como chutes a gol, passes certos e faltas cometidas, essa fonte enriquece nossa análise, permitindo uma compreensão mais refinada do rendimento em campo.

campeonato-brasileiro-gols.csv:
Adicionando um componente crucial à análise, este arquivo traz informações específicas sobre os gols marcados ao longo das temporadas. Permite uma investigação mais aprofundada dos aspectos ofensivos do campeonato, destacando padrões de pontuação e identificando jogadores e equipes mais prolíficos na arte de marcar gols.

Ao integrar esses conjuntos de dados, nosso objetivo é desvendar não apenas a história do Campeonato Brasileiro, mas também oferecer insights significativos sobre estratégias de sucesso e desafios enfrentados pelas equipes e jogadores ao longo do tempo. Esta abordagem analítica busca proporcionar uma compreensão abrangente e contextualizada do cenário do futebol brasileiro.

Foram feitos alguns gráficos de primeiramente para melhorar entender os dados e saber como busca-los posteriormente de forma mais fácil e eficiente. A seguir segue alguns desses gráficos feitos a partir do Google Colab em Phyton para a visualização dos dados.





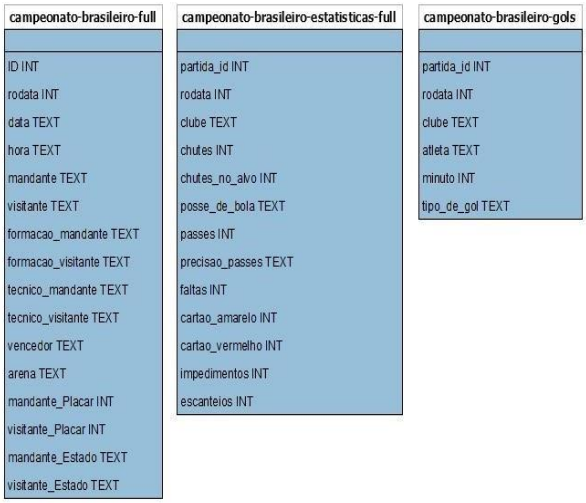
3.3. Preparação de Dados

A preparação dos dados é uma etapa fundamental em projetos de análise, envolvendo a limpeza, transformação e organização de conjuntos de dados brutos. Nesse processo, dados inconsistentes, ausentes ou redundantes são tratados para garantir a qualidade e coesão das informações. A preparação é essencial para criar conjuntos de dados prontos para análise, facilitando a extração de insights e a construção de modelos precisos.

3.3.1 Elaboração do Staging Area

A staging area, ou área de preparação, é uma etapa intermediária em processos de integração de dados. Nesse espaço temporário, os dados são consolidados, limpos e transformados antes de serem carregados no destino final. Essa abordagem permite uma preparação eficaz, garantindo que apenas dados de qualidade sejam movidos para a próxima fase do processo.

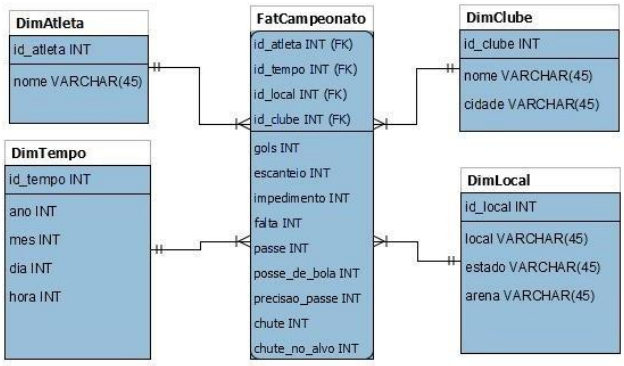
(Github dos scripts SQL)



3.3.2 Modelo Multidimensional

Um modelo multidimensional é uma abordagem de design de banco de dados que organiza os dados em torno de conceitos fundamentais, como dimensões e fatos. Essa estrutura facilita análises complexas ao permitir a representação visual e hierárquica dos dados. As dimensões representam as categorias pelos quais os dados são analisados, enquanto os fatos são as métricas quantificáveis. Essa modelagem é comumente usada em sistemas OLAP (Online Analytical Processing) para suportar eficientemente consultas analíticas em grandes conjuntos de dados.

(Github dos scripts SQL)



3.3.3 Higienização dos dados

Higienização de dados refere-se ao processo de limpar, corrigir e organizar conjuntos de dados, visando garantir a consistência e a qualidade das informações. Isso envolve a identificação e correção de dados ausentes, inconsistentes ou duplicados, além da padronização de formatos para facilitar análises. A higienização é crucial para preparar dados de maneira confiável, promovendo resultados mais precisos em análises e modelagem.

Higienização dos dados feitos nos três arquivos, mostrando abaixo apenas um como exemplo. Dados em forma de texto foram padronizados para melhor entendimento e estudo dos dados e informações que estavam nulas foram devidamente mostradas de forma mais fácil como NaN

Codigo feito para a higienização:

```python
import pandas as pd
import nltk
from nltk.corpus import stopwords
import spacy
from sklearn.feature_extraction.text import TfidfVectorizer

df = pd.read_csv('campeonato-brasileiro-estatisticas-full.csv')

df = df.apply(lambda x: x.astype(str).str.lower())

df.to_csv('campeonato-brasileiro-estatisticas-full.csv', index=False)

nltk.download('stopwords')
stop_words = set(stopwords.words('portuguese'))

def remove_stopwords(text):
    words = text.split()
    filtered_words = [word for word in words if word.lower() not in stop_words]
    return ' '.join(filtered_words)

df = df.applymap(remove_stopwords)

df.to_csv('campeonato-brasileiro-estatisticas-full.csv', index=False)
```

ANTES:

ID	rodata	data	hora	mandante	visitante	formacao_mandante	formacao_visitante	tecnico_mandante	tecnico_visitante
1	1	29/3/2003	16:00	Guarani	Vasco				
2	1	29/3/2003	16:00	Athletico-PR	Gremio				
3	1	30/3/2003	16:00	Flamengo	Coritiba				
4	1	30/3/2003	16:00	Goias	Paysandu				
5	1	30/3/2003	16:00	Internacional	Ponte Preta				
6	1	30/3/2003	16:00	Criciuma	Fluminense				
7	1	30/3/2003	16:00	Juventude	Sao Paulo				
8	1	30/3/2003	16:00	Fortaleza	Bahia				
9	1	30/3/2003	16:00	Cruzeiro	Sao Caetano				
10	1	30/3/2003	16:00	Vitoria	Figueirense				

DEPOIS:

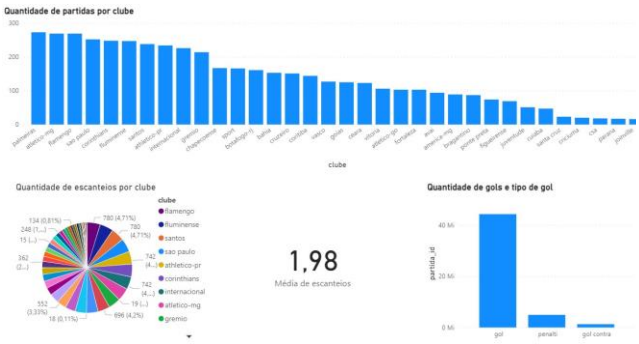
ID	rodata	data	hora	mandante	visitante	formacao_mandante	formacao_visitante	tecnico_mandante	tecnico_visitante
1	1	29/3/2003	16:00	guarani	vasco	nan	nan	nan	nan
2	1	29/3/2003	16:00	athletico-pr	gremio	nan	nan	nan	nan
3	1	30/3/2003	16:00	flamengo	coritiba	nan	nan	nan	nan
4	1	30/3/2003	16:00	goias	paysandu	nan	nan	nan	nan
5	1	30/3/2003	16:00	internacional	ponte preta	nan	nan	nan	nan
6	1	30/3/2003	16:00	criciuma	fluminense	nan	nan	nan	nan
7	1	30/3/2003	16:00	juventude	sao paulo	nan	nan	nan	nan
8	1	30/3/2003	16:00	fortaleza	bahia	nan	nan	nan	nan
9	1	30/3/2003	16:00	cruzeiro	sao caetano	nan	nan	nan	nan
10	1	30/3/2003	16:00	vitoria	figueirense	nan	nan	nan	nan

3.4. Modelagem

A modelagem de dados é o processo de criar representações estruturadas e abstratas dos dados de uma organização. Ela envolve a definição de entidades, relacionamentos e atributos para organizar as informações de forma eficiente. Esses modelos podem ser elaborados por meio de diferentes abordagens, como modelo relacional para bancos de dados ou modelos conceituais para compreensão inicial. A modelagem de dados é fundamental para projetar sistemas de informação coesos e facilitar a análise, armazenamento e recuperação de dados de maneira organizada.

(GithHub com o script do dashboard)

3.4.1 Dashboard



3.5. Avaliação

Este projeto representou uma incursão abrangente e meticulosa no universo do Campeonato Brasileiro de Futebol ao longo de duas décadas, utilizando a metodologia CRISP-DM como guia. Inicialmente, uma profunda compreensão do negócio foi estabelecida, focalizando a análise específica do desempenho de clubes e jogadores.

A fase de higienização dos dados desempenhou um papel crucial, garantindo consistência e uniformidade nos conjuntos de dados. A padronização e o tratamento de valores nulos permitiram uma preparação eficiente para análises subsequentes. A modelagem multidimensional emergiu como uma ferramenta valiosa, proporcionando uma representação visual e hierárquica dos dados, facilitando análises complexas.

Ao explorar os conjuntos de dados, notadamente "campeonato-brasileiro-full.csv," "campeonato-brasileiro-estatisticas-full.csv," e "campeonato-brasileiro-gols.csv," obtivemos uma visão rica e diversificada do torneio. Cada arquivo contribuiu com camadas distintas de informação, permitindo insights profundos sobre estratégias adotadas e desafios enfrentados pelas equipes.

A preparação efetiva dos dados foi fundamental para criar conjuntos robustos, enquanto a modelagem multidimensional ofereceu uma abordagem eficaz para representar a complexidade do cenário futebolístico. A higienização de dados garantiu a qualidade e a coesão necessárias, e a integração dos conjuntos proporcionou uma visão holística do desempenho no Campeonato Brasileiro.

No entanto, reconhecemos que este é um ponto de partida, e a busca pela compreensão aprofundada do futebol brasileiro é contínua. A iteratividade e refinamento constante são essenciais para garantir que as análises permaneçam relevantes e proporcionem uma contribuição significativa para o entendimento do cenário esportivo nacional.

4. Conclusão

Ao término deste projeto, emerge uma compreensão profunda e rica sobre o desempenho no Campeonato Brasileiro de Futebol ao longo de duas décadas. A aplicação da metodologia CRISP-DM proporcionou uma estrutura robusta, iniciando com uma análise minuciosa do negócio, focalizando especialmente o rendimento de clubes e jogadores.

A higienização dos dados revelou-se como uma etapa essencial, transformando conjuntos brutos em fontes de informação confiáveis e acessíveis. A padronização, correção de valores nulos e uniformização de formatos contribuíram para a consistência e qualidade dos dados, preparando o terreno para análises significativas.

A modelagem multidimensional, adotada na representação estruturada dos dados, possibilitou uma compreensão visual e hierárquica, facilitando interpretações complexas. Essa abordagem foi crucial para identificar padrões, tendências e nuances no desempenho das equipes ao longo do tempo.

Exploramos três conjuntos de dados essenciais: "campeonato-brasileiro-full.csv," "campeonato-brasileiro-estatisticas-full.csv," e

"campeonato-brasileiro-gols.csv." Cada arquivo proporcionou uma perspectiva única, desde dados gerais sobre o torneio até estatísticas detalhadas de jogadores e informações específicas sobre gols marcados. Essa diversidade de fontes enriqueceu nossa análise, oferecendo uma visão abrangente e multifacetada do cenário futebolístico brasileiro.

A preparação eficaz dos dados, a modelagem multidimensional e a higienização dos dados culminaram em uma integração coesa e significativa. No entanto, reconhecemos que este projeto é um ponto de partida, e a evolução contínua é imperativa. A dinâmica do futebol brasileiro é fluida, exigindo uma abordagem iterativa e uma constante busca por refinamento para manter a relevância das análises.

Este projeto não apenas fornece uma visão aprofundada do desempenho no Campeonato Brasileiro, mas também destaca a importância da análise de dados no contexto esportivo. A capacidade de extrair insights significativos sobre estratégias, desafios e tendências oferece um caminho promissor para uma compreensão mais rica e informada do amado esporte brasileiro. Este é um convite para continuar explorando, questionando e aprimorando nossa compreensão do futebol, sempre buscando decifrar os segredos ocultos nos números e nas histórias que eles contam.

5. Referências

MySQL Workbench:
- Oracle Corporation. (Ano). MySQL Workbench. https://www.mysql.com/products/workbench/

Power BI:
- Microsoft Corporation. (Ano). Power BI. https://powerbi.microsoft.com/

Notion:
- Notion Labs, Inc. (Ano). Notion. https://www.notion.so/

Google Colab:
- Google. (Ano). Google Colab. https://colab.research.google.com/

Python:
- Python Software Foundation. (Ano). Python Programming Language. https://www.python.org/

Matplotlib:
- Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

Pandas:
- McKinney, W. (2011). Pandas: a Foundational Python Library for Data Analysis and Statistics. Python for High Performance and Scientific Computing, 14.

Seaborn:
- Waskom, M. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.

Lembre-se de ajustar as informações conforme necessário, incluindo o ano de acesso ou publicação, dependendo da disponibilidade da informação.

APÊNDICE A
Arquivos de dados

Rótulos: Dataset 1: campeonato-brasileiro-full Dataset 2: campeonato-brasileiro-gols Dataset 3: campeonato-brasileiro-estatisticas-full

Tipos de Dados:

1. **Dataset 1: campeonato-brasileiro-full** • ID : numero inteiro • Rodada : numero inteiro • Data : data • Horario : hora • Dia : texto • Mandante : texto • Visitante : texto
2. **Dataset 2: campeonato-brasileiro-gols** • partida_ID – numero inteiro • Rodada – numero inteiro • Clube - texto • Atleta - texto • Minuto – texto • Tipo_de_gol - texto
3. **Dataset 3: campeonato-brasileiro-estatisticas-full** • partida_ID – numero inteiro • Rodada – numero inteiro • Clube - texto • Chutes - texo • Chutes a gol - texto • Posse de bola – numero inteiro • Passes – numero inteiro • precisao_passes – numero inteiro • Faltas - texto • cartao_amarelo - texto • cartao_vermelho - texto • Impedimentos - texto • Escanteios – numero inteiro

Quantitativos: Dataset 1: campeonato-brasileiro-full Registros: 8026 Campos: 7 **Dataset 2: campeonato-brasileiro-gols** Registros: 7987 Campos: 6 **Dataset 3: campeonato-brasileiro-estatisticas-full** Registros: 16051 Campos: 13 Número de Datasets: número total de datasets: 3

Relacionamentos: Temos relacionados entre os datasets algumas colunas: partida_id, rodada e clube

Formato de dados: Os três datasets estão em arquivo csv em excell

Dicionario de dados: Dataset 1: campeonato-brasileiro-full ID - ID da partida Rodada : Rodada que aconteceu a partida Data : Data que ocorreu a partida Horario : Horario que ocorreu a partida Dia : Dia da semana que ocorreu a partida Mandante : Clube mandante Visitante : Clube Visitante formacao_mandante: Formacao do mandante formacao_visitante: Formacao do visitante tecnico_mandante: tecnico do mandante tecnico_visitante: tecnico do visitante Vencedor : Clube vencedor da partida. Arena : Arena que ocorreu a partida Mandante Placar : Gols que o clube mandante fez na partida Visitante Placar : Gols que o clube visitante fez na partida Estado Mandante : Estado do clube mandatorio Estado Visitante : Estado do clube visitante Estado Vencedor : Estado do clube vencedor.

Dataset 2: campeonato-brasileiro-estatisticas-full partida_ID - ID da partida Rodada - Rodada da partida Clube - Nome do clube Chutes - Finalizacoes Chutes a gol - Finalizacoes na direcao do gol Posse de bola - Percentual da posse de bola Passes - Quantidade de passes que o clube deu na partida precisao_passes - Percentual da precisao de passe Faltas - Quantidade de faltas cometidas na partida cartao_amarelo - Quantidade de cartoes amarelos para o clube na partida cartao_vermelho - Quantidade de cartoes vermelhos para o clube na partida Impedimentos - Quantidade de impedimentos para o clube na partida Escanteios - Quantidade de escanteios para o clube na partida

Dataset 3: campeonato-brasileiro-gols partida_ID - ID da partida Rodada - Rodada da partida Clube - Nome do clube Atleta - Nome do atleta que fez o gol Minuto - Minuto na partida em que o gol foi marcado