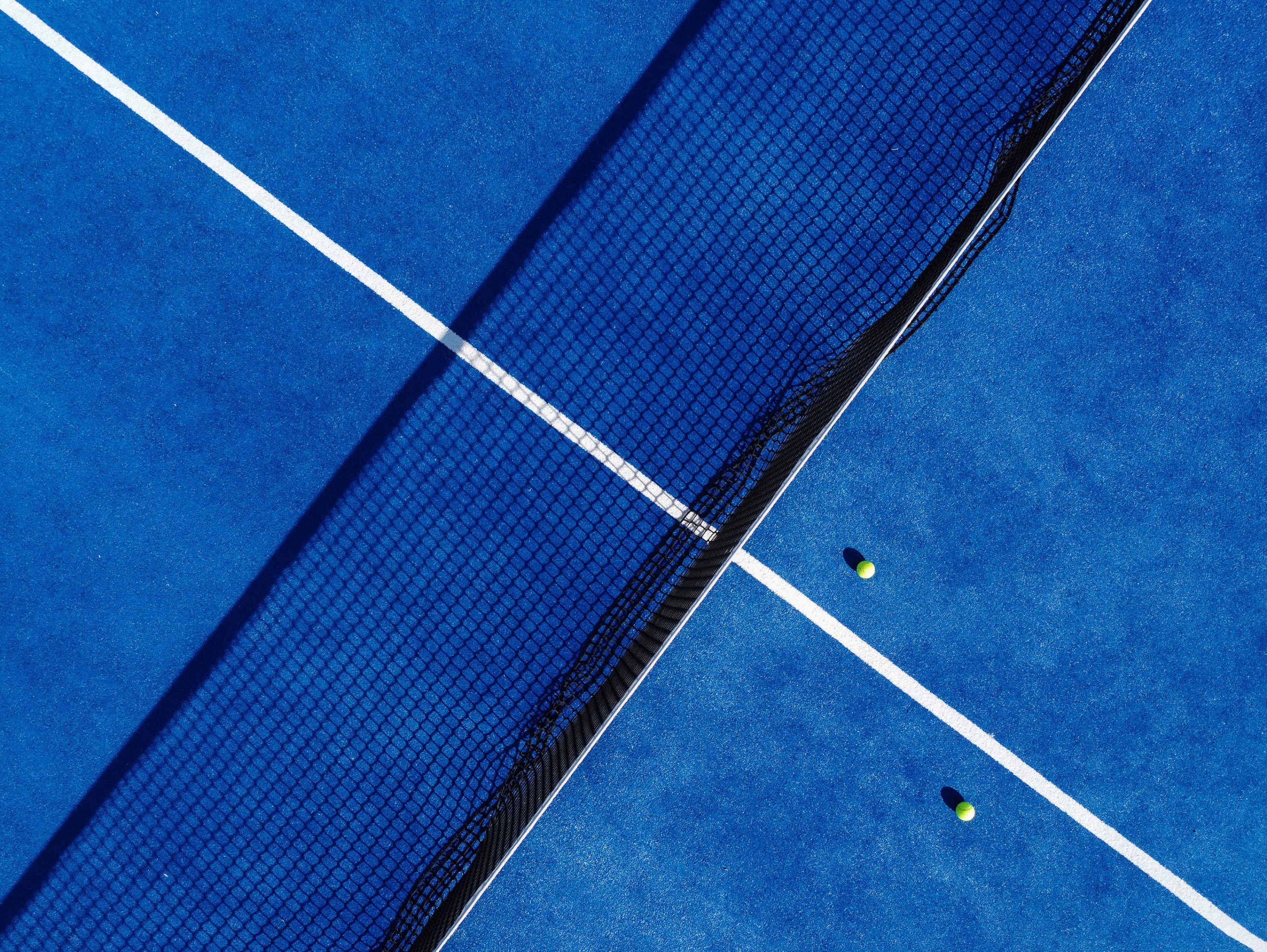
**A picture containing font, text, graphics, logo

Description automatically generated**

**PROJETO APLICADO EM CIÊNCIA DE DADOS I -**

**RELATÓRIO**

Licenciatura em Ciência de dados

**Base de Dados ATP Players -**

**Brasil**

Grupo 5:

Eliane Susso Efraim Gabriel, Nº 103303

Marco Delgado Esperança, Nº 110451

Maria João Ferreira Lourenço, Nº 104716

Umeima Adam Mahomed, Nº 992397

Docentes: Diana Aldea Mendes, Sérgio Moro

19 de abril de 2023

**Índice:**

[1. Introdução 4](#_Toc133225008)

[2. Business Understanding 4](#_Toc133225009)

[Dataset ATP 4](#_Toc133225010)

[ATP Tour 4](#_Toc133225011)

[ATP no Brasil 5](#_Toc133225012)

[Impacto do Ténis na economia, cultura e desenvolvimento do desporto no Brasil 5](#_Toc133225013)

[Questões de negócio 6](#_Toc133225014)

[Objetivos da análise de dados 6](#_Toc133225015)

[Hipóteses iniciais 6](#_Toc133225016)

[Terminologia 6](#_Toc133225017)

[3. Data Understanding 7](#_Toc133225018)

[Importação dataset 7](#_Toc133225019)

[Variáveis em estudo 7](#_Toc133225020)

[4. Data Preparation 8](#_Toc133225021)

[1. Uniformização da variável Prize 9](#_Toc133225022)

[2. Criação das variáveis DateStart e DateEnd 9](#_Toc133225023)

[3. Valores omissos 9](#_Toc133225024)

[4. Imputação dos valores omissos na variável DateEnd 9](#_Toc133225025)

[5. Substituição de valores omissos na variável Height 9](#_Toc133225026)

[6. Substituição de valores omissos na variável GameRank 9](#_Toc133225027)

[7. Renomear a variável GameRank 9](#_Toc133225028)

[8. Criação da função que associa o jogador ao seu rank por ano 9](#_Toc133225029)

[9. Criação da função que busca o valor do ranking e da coluna RankPlayer 9](#_Toc133225030)

[10. Uniformização do Score 9](#_Toc133225031)

[11. Criação da coluna NumberSets 9](#_Toc133225032)

[12. Campo \_id 9](#_Toc133225033)

[13. Criação da nova variável City a apartir da variável Location 9](#_Toc133225034)

[14. Importação do dataset Worldcities 9](#_Toc133225035)

[15. Duplicados de cidades em países diferentes 9](#_Toc133225036)

[16. Valores omissos em Country 9](#_Toc133225037)

[17. Países com escritas diferentes 9](#_Toc133225038)

[18. Formas diferentes de escrita de Brasil 9](#_Toc133225039)

[19. Uniformização da escrita para “Brasil” 9](#_Toc133225040)

[20. Valores omissos 9](#_Toc133225041)

[5. Modelling 10](#_Toc133225042)

[6. Evaluation 11](#_Toc133225043)

[7. Deployment 11](#_Toc133225044)

[8. Conclusões 13](#_Toc133225045)

[9. Bibliografia 13](#_Toc133225046)

# Introdução

Objetivo do trabalho

Como vamos fazer isso: Metodologia CRISP-DM

# Business Understanding

Com o objetivo de prever o número de sets para a conclusão de um jogo de ténis profissional no Brasil, começamos por perceber o contexto do problema e de onde vinha o dataset, uma vez que este pode condicionar toda a formulação e resolução do problema.

NÃO ESQUECER OBJETIVO/PROPÓSITO VIRADO PARA O NEGÓCIO

## Dataset ATP

O dataset presente no ficheiro atpplayers.json contém dados que foram recolhidos do site do ATP (https://www.atptour.com/) e contém informações acerca dos jogos singulares de 10361 jogadores masculinos, pertencentes ao Top 500, que jogaram de 28/03/1973 a 14/02/2022.

## ATP Tour

Antes de mais, fomos investigar em que consistia a Associação de Tenistas Profissionais (*ATP*), que é a entidade responsável pelos circuitos de ténis profissionais masculinos, o ATP Tour, ATP Challenger Tour e ATP Champions Tour.

Esta começou em setembro de 1972, devido à ação de Donald Dell, Jack Kramer e Cliff Drysdale com o objetivo de proteger os interesses dos tenistas profissionais. A denominação desta organização foi mudada ao longo dos anos, mas por fim retomou ao nome original que prevalece atualmente.

A organização está sediada em Londres, tendo as suas filiais noutros continentes: na América em Ponte Vedra Beach, no estado norte americano na Flórida, na Europa tem a sua filial no Mónaco e a internacional que cobre Ásia, África e Australásia em Sidney. Esta é gerida pelo atual diretor executivo Massimo Calvelli e uma equipa de seis membros.

Os torneiros mais prestigiosos na modalidade desportiva ténis são os de Grand Slam, ou *majors*- Australian Open, Roland Garros (ou French Open) e Wimbledon e US Open, que não são sob a alçada de ATP, no entanto, são atribuídos pontos do *ranking* ATP.

## ATP no Brasil

Relativamente aos torneios que ocorrem no Brasil, o mais famoso é o Rio Open, que teve início em 2014 e ocorre anualmente, apesar de não ter ocorrido em 2021, devido à pandemia. Este é considerado o maior torneio de tênis da América do Sul e pertence ao ATP 500, logo é um torneio bastante prestigiado a nível internacional, onde participam os grandes nomes do ténis a nível mundial.

Também podemos mencionar o Brasil Open que se realizou de 2001 a 2019, apenas para atletas masculinos e ocorreu na Costa do Sauipe, São Paulo e Rio de Janeiro e pertencia ao ATP Tour 250; o ATP São Paulo que ocorreu de 2001 a 2012, também só para atletas masculinos, sendo substituído pelo Rio Open.

As competições do tipo BRA vs. X são jogos que ocorrem na Davis Cup, nomeadamente na World Group Play-Offs, onde jogadores de países diferentes se defrontam.

Por fim, podemos referir as competições que fazem parte da *ITF Men’s Circuit*: Brazil Fx - apenas para atletas masculinos e estão divididas em diferentes níveis consoante o nível dos atletas e prémio, sendo principalmente para os jogadores ganharem experiência e melhorarem os seus rankings -, M15 São Paulo - só para atletas masculinos, cujo prémio é de 15000$ e ocorre em São Paulo, permitindo o melhoramento dos rankings e a qualificação para o ATP Challenger Tour e ATP World Tour - e M25 Rio de Janeiro - com as mesmas características que a última, só que o prémio é de 25000$ e ocorre no Rio de Janeiro.

## Impacto do Ténis na economia, cultura e desenvolvimento do desporto no Brasil

Como referido acima, o país tem competições de ténis de grande relevância tanto a nível nacional como internacional, que trazem para o país nomes conhecidos da modalidade e um grande conjunto de entusiastas e fãs do desporto, gerando grande receita para o país.

Relativamente a alguns nomes do tênis brasileiro que ganharam títulos do circuito ATP de 1973 a 2022 podemos apontar Gustavo Kuerten, Thiago Monteiro, Thomaz Belluci, Thiago Alves e Ricardo Mello. – depois ver se todos estes nomes estão no dataset

Os jogos são cobertos pelos meios de comunicação e muitas vezes transmitidos na televisão, permitindo a chegada ao povo brasileiro de forma fácil. Desta forma, é possível a divulgação do talento nacional, permitindo que os jovens talento mostrem as suas capacidades, compitam em níveis mais elevados de futuro e alcancem a internacionalização.

## Questões de negócio

## Objetivos da análise de dados

## Hipóteses iniciais

1. Questões de negócio: Identifique as questões de negócio que a análise dos dados deve abordar, como:

* Qual é o desempenho dos jogadores brasileiros nos jogos de tênis no Brasil?
* Quais são os jogos mais populares e bem-sucedidos em termos de público e resultados?
* Como os jogos no Brasil se comparam aos jogos em outros países em termos de público, resultados e qualidade do jogo?

1. Objetivos da análise de dados: Descreva os objetivos da análise de dados, incluindo o que você espera obter da análise, como insights sobre o desempenho dos jogadores, tendências no público e resultados dos jogos e possíveis oportunidades para aprimorar a experiência dos fãs de tênis.
2. Hipóteses iniciais: Liste as hipóteses iniciais que você tem sobre os dados com base nas questões de negócios e nos objetivos da análise de dados. Por exemplo, você pode hipotetizar que os jogadores brasileiros têm melhor desempenho nos jogos realizados no Brasil do que em outros países.

## Terminologia

Esta modalidade apresenta termos técnicos que têm que ser compreendidos para garantir a criação de um bom modelo que responda ao objetivo proposto, para isso estes vão ser explorados de forma breve.

Relativamente ao sistema de pontuação, temos os termos “*point*”, “*game*”, “*set*” e “*match*”. De forma simples, o primeiro jogador a ganhar quatro pontos (“points”) ganha um “*game*”, e o primeiro jogador a ganhar seis “*games*” ganha um “*set*”, dependendo do tipo de torneiro, o primeiro que ganhar dois ou três “*sets*” ganha o “*match*”.

O primeiro ponto é chamado de “15”, o seguinte de “30”, “40” e por fim “*game*”.

Referente ao tipo de piso, este será explorado futuramente, mas adiantamos que existem quatro tipos principais: relva artificial, barro/argila artificial, cimento e borracha, sendo o último considerado o melhor.

# Data Understanding

Nesta etapa começamos por importar o *dataset* e seguidamente exploramos as diversas variáveis e seus significados.

## Importação dataset

Importamos o *dataset* presente no ficheiro atpplayers.json através da função read\_json do Pandas.

## Variáveis em estudo

Relativamente às variáveis presentes no *dataset* estas podem ser divididas em dois subgrupos: as que estão relacionadas com os jogadores e com os jogos.

* **\_id:** identificador único de cada linha, que foi eliminado;
* **PlayerName:** referente ao principal jogador da partida;
* **Born:** país e /ou cidade que o jogador nasceu;
* **Height:** altura dos jogadores em cm;
* **Hand:** indica a mão dominante do jogador, que pode ser Right-Handed, se for destro, Left-Handed se for canhoto ou Ambidextro se forem as duas; e que mãos utiliza ao executar o serviço, pode ser One-Handed Backhand, se usar uma, Two-Handed Backhand, se usar 2, ou Unknown Backhand, se não se souber;
* **LinkPlayer:** é um link que nos remete para o perfil detalhado do jogador no website do ATP Tour;
* **Tournament:** nome do torneio;
* **Location:** país e cidade, onde determinado torneio se realizou;
* **Date:** indica as datas de início e fim do torneio;
* **Ground:** tipo de terreno em que o torneio foi jogado, pode ser “*Hard*” - composto por materiais uniformes e endurecidos, com uma camada superficial de acrílico –“*Grass*” - relva natural-, “*Clay*” - pequenos pedaços de pedra pouco consolidada ou que se parte facilmente pode ser de cor vermelha ou verde - ou “*Carpet*” - materiais sintéticos, como relva artificial com preenchimento de areia;
* **Prize:** prémio monetário;
* **GameRound:** fase do torneio a que pertence o jogo, pode pertencer aos jogos de qualificação - têm Qualification na designação -, e os de eliminação que têm Round of X ou Finals no nome. Para esta última categoria temos também a Round Robin que é formato das eliminatórias, em que os jogadores são separados em grupos e jogam todos contra todos, os que passarem a prova vão para as semifinais;
* **GameRank**, mais tarde apelidada de **RankOpponent:** refere-se ao rank do Opponent;
* **Opponent:** nome do adversário;
* **WL:** resultado do jogo, é W - se ganhou - ou L - se perdeu;
* **Score:** resultados do jogo por sets, sendo que o primeiro número se refere ao número de jogos que o jogador principal daquele jogo venceu e o segundo ao número de jogos que o oponente venceu;

Como também criamos algumas variáveis com o objetivo de melhor explorar o dataset, vamos proceder de seguida à descrição destas.

* **RankPlayer:** ranking do jogador principal num determinado ano;
* **City:** cidade presente na variável Location
* **DateStart:** data de início do torneio
* **DateEnd:** data de fim do torneio, que pode coincidir com a data de início do mesmo
* **NumberSets:** contagem do número de pares presentes na variável Score

1. Análise das variáveis: Para cada variável coletada, você pode incluir informações sobre a sua distribuição, valores ausentes, valores discrepantes e outras características importantes. Você pode apresentar gráficos e tabelas para ajudar a ilustrar essas informações.
2. Análise das instâncias: Para cada instância coletada, você pode incluir informações sobre as suas características, como valores de cada variável, valores ausentes e valores discrepantes. Você também pode explorar possíveis relações entre as diferentes instâncias, utilizando gráficos e tabelas para ajudar a visualizar essas informações.
3. Análise exploratória de dados: Além de analisar cada variável e instância separadamente, você também pode explorar possíveis relações entre as diferentes variáveis. Isso pode ajudar a identificar padrões e tendências nos dados, o que pode ser útil para orientar análises futuras.
4. Limitações dos dados: É importante incluir informações sobre quaisquer limitações dos dados, como possíveis fontes de erro ou vieses nos dados coletados. Isso pode ajudar a garantir que os resultados da análise sejam interpretados corretamente.

## Estrutura do dataset

O nosso *dataset*, após ser filtrado para conter apenas Country igual a Brasil, apresenta 37367 linhas e 21 colunas. Relativamente ao tipo de variáveis, estas variam, existindo catorze variáveis do tipo object, quatro do tipo float, 2 do tipo datetime e uma do tipo int.

Inicialmente, nem todas as linhas estavam completamente preenchidas, como é o caso das variáveis RankPlayer com 2620 entradas não preenchidas, Height com 11432 linhas, Prize com 367, RankOpponent com 3511 e Score com 2 linhas.

Relativamente à análise descritiva, dividimos as análises entre as variáveis categóricas e numéricas. As primeiras estão completamente preenchidas e os seus valores únicos variam de 1 no caso de Country até 2655 em Opponent. As últimas, apresentam alguns valores nulos e os seus valores variam. O número de sets varia de 1 a 5, sendo 2 e 3 os valores mais comuns. O rank do jogador e do oponente varia de 1 a 2243. Na variável Height temos claramente outliers, considerando que o valor mínimo é 0 cm e o valor máximo é 510 cm. O valor do prémio varia de 10000 a 1786690 unidades monetárias e a média é de 72149,6.

TIPO DE VAR

Os valores únicos das variáveis variam de 1 a 2655. Considerando o número de valores únicos presentes e as próprias características das variáveis, procedemos à sua classificação: Country tem 1 (….)

## Qualidade dos dados

Não verificamos a existência de duplicados para o dataset filtrado, mas fizemos essa operação inicialmente com a coluna \_id, que foi eliminada antes da criação do dataset em que nos focamos agora, mas provamos a não existência de duplicados.

Relativamente aos valores omissos, a maioria das variáveis está completamente preenchida, excepto a variável RankPlayer com 2620 entradas por preencher, Height com 11432, Prize com 367, RankOpponent com 3511 e Score com 2.

## Investigação do conteúdo

# Data Preparation

Os passos que seguimos para a limpeza dos dados seguem a seguinte ordem:

### Uniformização da variável Prize;

### Criação das variáveis DateStart e DateEnd;

### Valores omissos

### Imputação dos valores omissos na variável DateEnd

### Substituição de valores omissos na variável Height

### Substituição de valores omissos na variável GameRank

### Renomear a variável GameRank

### Criação da função que associa o jogador ao seu rank por ano

### Criação da função que busca o valor do ranking e da coluna RankPlayer

### Uniformização do Score

### Criação da coluna NumberSets

### Campo \_id

### Criação da nova variável City a apartir da variável Location

### Importação do dataset Worldcities

### Duplicados de cidades em países diferentes

### Valores omissos em Country

### Criação da variável BornCountry

### Criação da variável L\_OR\_R

### Formas diferentes de escrita de Brasil

### Uniformização da escrita para “Brasil”

A partir daqui, podemos trabalhar apenas com os dados para a o Brasil, uma vez que já obtivemos uma base de dados mais coerente a respeito do país em que nos iremos focar.

### Valores Únicos

### Investigação da qualidade dos dados

### Imputação

Na fase de preparação dos dados de um relatório sobre o dataset ATP com jogos apenas no Brasil, você deve incluir informações sobre as etapas que foram realizadas para limpar, organizar e transformar os dados para análise. Algumas informações importantes que podem ser incluídas nessa seção incluem:

1. Limpeza de dados: Descreva as etapas que foram realizadas para limpar os dados, incluindo a remoção de valores ausentes ou duplicados, correção de erros de digitação e padronização de formatos de dados.
2. Transformação de dados: Explique as etapas que foram realizadas para transformar os dados para análise, como a criação de novas variáveis ou a combinação de variáveis existentes para gerar insights mais significativos.
3. Normalização de dados: Descreva as etapas que foram realizadas para normalizar os dados, como a padronização de unidades de medida ou a transformação de dados em uma escala comum para permitir comparações mais precisas entre variáveis.
4. Seleção de variáveis: Explique as etapas que foram realizadas para selecionar as variáveis que seriam incluídas na análise. Isso pode incluir a remoção de variáveis irrelevantes ou redundantes, bem como a seleção de variáveis-chave que sejam relevantes para as questões de negócios e os objetivos da análise de dados.
5. Análise exploratória de dados: Descreva as etapas que foram realizadas para explorar os dados, como a criação de gráficos e visualizações para entender a distribuição dos dados, identificar outliers ou padrões interessantes nos dados.

|  |  |
| --- | --- |
| Variáveis para o dataset df\_brasil | Valores únicos |
| Country | 1 |
| WL | 3 |
| Ground | 4 |
| NumberSets | 5 |
| Hand | 8 |
| GameRound | 11 |
| Height | 21 |
| Prize | 40 |
| City | 51 |
| Location | 57 |
| Tournament | 148 |
| DateEnd | 577 |
| DateStart | 595 |
| Date | 599 |
| Born | 779 |
| RankPlayer | 1422 |
| RankOpponent | 1825 |
| LinkPlayer | 1900 |
| PlayerName | 1900 |
| Score | 2102 |
| Oponent | 2655 |

# 

# coia Modelling

Na fase de Modelagem de um relatório sobre o dataset ATP com jogos apenas no Brasil, é importante justificar a escolha do modelo utilizado na análise. A escolha do modelo deve estar alinhada com os objetivos da análise de dados e as perguntas de negócios que precisam ser respondidas.

Existem muitos modelos estatísticos e algoritmos de machine learning que podem ser aplicados aos dados do ATP para gerar insights significativos. No entanto, a escolha do modelo deve ser baseada em vários fatores, incluindo a natureza dos dados, a complexidade dos relacionamentos entre as variáveis e a capacidade do modelo de atender aos objetivos da análise.

No caso do dataset ATP com jogos apenas no Brasil, uma das possíveis abordagens de modelagem seria a utilização de regressão linear múltipla. Essa técnica permite modelar as relações lineares entre uma variável dependente (como a pontuação do jogo) e várias variáveis independentes (como a idade dos jogadores, ranking dos jogadores e superfície da quadra, entre outras).

Outra abordagem possível seria a utilização de algoritmos de machine learning, como árvores de decisão ou redes neurais, que podem ser úteis para explorar relações não lineares e complexas entre as variáveis do dataset.

É importante justificar a escolha do modelo selecionado, explicando como ele pode ser aplicado aos dados do ATP com jogos apenas no Brasil e como ele pode ajudar a responder às perguntas de negócios definidas no relatório. Além disso, deve-se fornecer informações sobre a qualidade do modelo, como sua precisão e desempenho em diferentes cenários, e discutir possíveis limitações e viéses que possam afetar a análise e as conclusões.

# Evaluation

Na fase de Avaliação de um relatório sobre o dataset ATP com jogos apenas no Brasil, é importante avaliar os resultados da análise e determinar se eles são consistentes com as perguntas de negócios e os objetivos definidos. Esta etapa permite identificar possíveis limitações da análise e fornecer insights adicionais que possam ser relevantes para a tomada de decisão.

Algumas informações importantes que podem ser incluídas na fase de Avaliação são:

1. Avaliação da qualidade dos resultados: Descreva como a qualidade dos resultados foi avaliada, incluindo as medidas utilizadas para avaliar a precisão, a validade e a confiabilidade dos resultados obtidos.
2. Discussão dos resultados: Discuta os resultados da análise, incluindo os principais insights obtidos, os padrões e tendências identificados e as conclusões gerais que podem ser tiradas a partir dos dados.
3. Comparação com objetivos de negócios: Compare os resultados da análise com os objetivos de negócios definidos no início do relatório. Determine se os resultados são consistentes com esses objetivos e se eles fornecem insights relevantes para a tomada de decisão.
4. Identificação de limitações e viéses: Identifique possíveis limitações e viéses que possam afetar a análise e os resultados. Forneça informações sobre como essas limitações podem afetar a interpretação dos resultados e as conclusões do relatório.
5. Sugestões para futuras análises: Fornecer sugestões para futuras análises, incluindo possíveis melhorias na coleta de dados, na escolha de modelos ou na definição de perguntas de negócios mais precisas. Isso pode ajudar a orientar futuras análises e garantir que elas sejam ainda mais relevantes para as necessidades de negócios.

# Deployment

Na fase de Implementação de um relatório sobre o dataset ATP com jogos apenas no Brasil, é importante apresentar uma proposta para a implementação das conclusões e recomendações obtidas a partir da análise dos dados.

Essa proposta deve ser baseada nas necessidades e objetivos de negócios identificados na fase de Business Understanding e nas conclusões obtidas a partir da análise dos dados.

Algumas informações importantes que podem ser incluídas na fase de Implementação são:

1. Descrição da proposta: Descreva a proposta para implementação das conclusões e recomendações obtidas a partir da análise dos dados. Essa proposta pode incluir ações específicas a serem tomadas para melhorar o desempenho dos jogadores, estratégias de marketing para aumentar a popularidade do tênis no Brasil, ou até mesmo uma estratégia para aumentar o número de torneios de tênis no país.
2. Benefícios da implementação: Liste os benefícios que a implementação da proposta pode trazer para o negócio, como aumento de receita, melhoria do desempenho dos jogadores, maior engajamento do público, entre outros.
3. Desafios e limitações: Identifique possíveis desafios e limitações que possam afetar a implementação da proposta. Isso pode incluir limitações financeiras, políticas ou tecnológicas, por exemplo. Discuta possíveis soluções para superar esses desafios e limitações.
4. Plano de ação: Descreva um plano de ação detalhado para implementação da proposta, incluindo as etapas necessárias, o cronograma de implementação, as responsabilidades e os recursos necessários.
5. Avaliação da implementação: Descreva como a implementação será avaliada, incluindo as medidas utilizadas para avaliar o sucesso da implementação e o impacto das mudanças propostas.

Ao incluir essas informações na fase de Implementação do seu relatório, você estará fornecendo uma proposta sólida e concreta para a implementação das conclusões e recomendações obtidas a partir da análise dos dados. Isso pode ajudar a orientar as decisões de negócios e garantir que as mudanças propostas sejam implementadas de forma eficaz e eficiente.

Top of Form

# Conclusões

# Bibliografia

*History: ATP tour: Tennis*. ATP Tour. (n.d.). Retrieved April 23, 2023, from https://www.atptour.com/en/corporate/history

*Senior leadership team: ATP tour: Tennis*. ATP Tour. (n.d.). Retrieved April 23, 2023, from https://www.atptour.com/en/corporate/management

Hadlich, G. (2022, January 7). *How does tennis scoring work? (by former pro)*. My Tennis HQ. Retrieved April 23, 2023, from https://mytennishq.com/tennis-scoring-rules-the-ultimate-guide-explained/

*4 types of surfaces used for tennis court*. Rubcorp. (2022, February 24). Retrieved April 23, 2023, from https://www.rubcorp.com/types-of-tennis-court-surfaces/

O Torneio – Rio Open. (n.d.). Retrieved April 16, 2023, from <https://rioopen.com/o-torneio/>

# Anexos

## Valores únicos para cada variável da df\_brasil

|  |  |
| --- | --- |
| Variáveis para o dataset df\_brasil | Valores únicos |
| Country | 1 |
| WL | 3 |
| Ground | 4 |
| NumberSets | 5 |
| Hand | 8 |
| GameRound | 11 |
| Height | 21 |
| Prize | 40 |
| City | 51 |
| Location | 57 |
| Tournament | 148 |
| DateEnd | 577 |
| DateStart | 595 |
| Date | 599 |
| Born | 779 |
| RankPlayer | 1422 |
| RankOpponent | 1825 |
| LinkPlayer | 1900 |
| PlayerName | 1900 |
| Score | 2102 |
| Oponent | 2655 |