Introdução: Este relatório apresenta uma análise exploratória dos dados da Fórmula 1, com o objetivo de responder algumas perguntas que julgamos interessantes a partir desses dados. As informações foram obtidas através do site Kaggle, de uma base de dados da categoria que abrangem diversos aspectos relacionados aos pilotos, equipes, corridas e campeonatos da Fórmula 1.

Filtro: Foi realizado um filtro dentro dos arquivos CSV, para que sejá importado apenas as colunas necessárias para fazer as consultas SQL e realizar a resposta das perguntas. Com isso, no momento da criação das perguntas, separamos quais CSV's e quais colunas que iriamos utilizar.

Integrantes: Mateus Paulo da Silva, Julia Machado, Heitor Assad Farad, Henrique Quintão e Franz Blauth

Descrição das tabelas dentro do banco:

1. **circuits**
    1. **circuitId** -> Representação do ID do circuito como chave única para cada um deles
    2. **name** -> Nome do circuito
    3. **country** -> País que é sediado o circuito
2. **constructor_standings**
    1. **constructorStandingsId** -> Primary Key da tabela como chave única
    2. **raceId** -> Foreign Key do ID da corrida que vem da tabela races
    3. **constructorId** -> Foreign Key do ID da construtora que vem da tabela constructors
    4. **wins** -> Quantidade de vitórias que a construtora teve
3. **constructors**
    1. **constructorID** -> Primary Key que fornece o ID da construtora (que é usada como Foreign Key em outras tabelas, com a constructor_standings)
    2. **name** -> Nome da construtora
4. **drivers**
    1. **driverID** -> Primary Key que fornece o ID do piloto (que é usada como Foreign Key em outras tabelas)
    2. **forename** -> Primeiro nome do piloto
    3. **surname** -> Segundo nome do piloto
    4. **dob** -> Data de nascimento do piloto
    5. **nationality** -> Nacionalidade do piloto
5. **lap_times**
    1. **raceID** -> Foreign Key do ID da corrida que vem da tabela races
    2. **driverID** -> Foreign Key do ID do piloto que vem da tabela drivers
    3. **lap** -> Volta que o piloto fez determinada volta rápida
    4. **time** -> O tempo que o piloto obteve em uma determinada volta
6. **races**
    1. **raceID** -> Primary Key que fornece o ID da corrida (que é usada como Foreign Key em outras tabelas como, lap_times e constructor_standings)
    2. **circuitID ->** Foreign Key do ID do circuito que vem da tabela circuits
    3. **name** -> Nome do circuito, caso necessite de algum comparação do nome do circuito da tabela circuits
7. **results**
    1. **resultId** -> Primary Key da tabela como chave única
    2. **raceId** -> Foreign Key do ID da corrida que vem da tabela races
    3. **driverId** -> Foreign Key do ID do piloto que vem da tabela drivers
    4. **position** -> Posição que o piloto finalizou em determinada corrida

Perguntas realizadas pelo grupo com base no tema:

1. Qual o top 10 de pilotos mais jovens da Fórmula 1?
2. Qual a relação dos pilotos e suas vitórias na Fórmula 1?
3. Qual é a distribuição do número de corridas por país na história da Formula 1?
4. Qual a relação de vitórias das construtoras e suas vitórias na Fórmula 1?
5. Qual é a relação entre o tempo de volta de um piloto e a posição final no circuito de interlagos?

Estimativas e representações visuais: Nesse tópico você irá encontrar algumas estimativas e representações visuais, através de histogramas, boxplot, gráficos e tabelas, de cada pergunta que foi realizada pelo grupo acerca do tema.

Qual o top de 10 de pilotos mais jovens da Fórmula 1

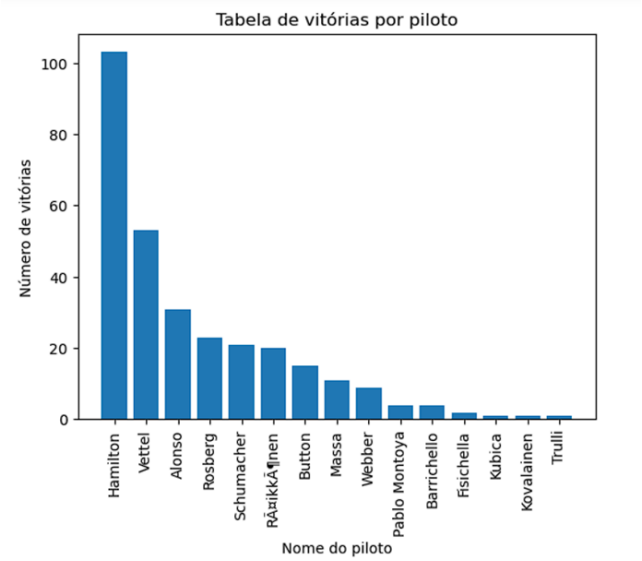
	driverId	forename	surname	dob	nationality	age
0	67	SÃ©bastien	Buemi	1988-10-31	Swiss	34
1	20	Sebastian	Vettel	1987-07-03	German	35
2	12	Nelson	Piquet Jr.	1985-07-25	Brazilian	37
3	3	Nico	Rosberg	1985-06-27	German	37
4	6	Kazuki	Nakajima	1985-01-11	Japanese	38
5	1	Lewis	Hamilton	1985-01-07	British	38
6	9	Robert	Kubica	1984-12-07	Polish	38
7	32	Christian	Klien	1983-02-07	Austrian	40
8	26	Scott	Speed	1983-01-24	American	40
9	16	Adrian	Sutil	1983-01-11	German	40

Tabela(s) necessária(s): Para a criação da representação da querie, em forma de tabela, que lista os 10 pilotos mais jovens da fórmula 1, foi necessário a utilização da tabela drivers, que contém todos os dados necessários para mostrar a idade do piloto e suas informações. Além disso, esse csv com o nome de **"drivers.csv"**, que está na pasta **"csv_files_formula_one"** dentro do repositório. Por fim, para saber mais sobre como foi executado a consulta, abra o arquivo **analise_exploratoria_formula_one.ipynb** que está dentro da pasta **jupyter_notebook_code**
Descrição das colunas:

- **driverId**: Representa o ID do piloto na tabela dimensão.
- **Forename**: Primeiro nome do piloto
- **Surname**: Segundo nome do piloto
- **Dob**: Data de nascimento do piloto
- **Nationality**: Nacionalidade do piloto
- **Age**: Idade do piloto com base na data atual

Análise: Com base nos dados que recuperamos, não é possível fazer uma análise tão precisa e valiosa, contudo é perceptível que, com esses 10 pilotos, nenhum deles nasceu antes de 1983 e a idade média dos pilotos é de 37 anos.

Qual a relação dos pilotos e suas vitórias na fórmula 1?



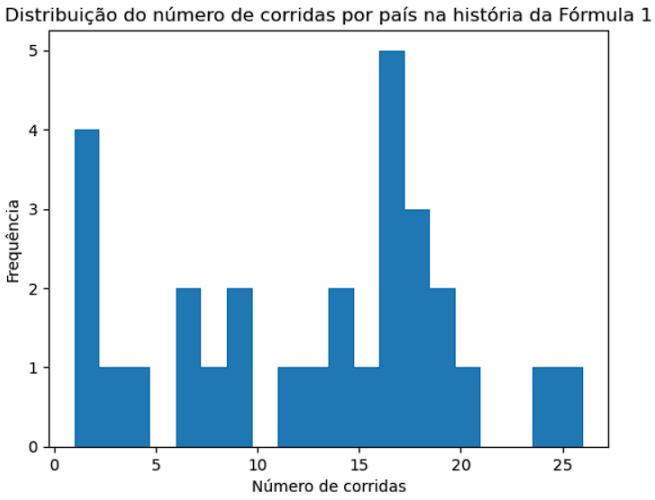
Tabela(s) necessária(s): Para a criação da representação da querie, em forma de um gráfico de barras, que lista os pilotos e suas respectivas vitórias, utilizamos de duas tabelas dimensões (drivers e results), pois com elas tínhamos como fazer uma comparação do piloto(drivers) com os resultados(results) e criar uma contagem de suas vitórias em todos os anos que correu na fórmula 1. Além disso, esses csvs podem ser encontrados pelo nome de **"drivers.csv"** e **"results.csv"**, que está na pasta **"csv_files_formula_one"** dentro do repositório. Por fim, para saber mais sobre como foi executado a consulta, abra o arquivo **analise_exploratoria_formula_one.ipynb** que está dentro da pasta **jupyter_notebook_code**

Descrição dos itens

- **Eixo Y**: Nesse eixo temos a quantidade do número de vitórias de cada piloto
- **Eixo X**: Nome dos pilotos

Análise: Com os dados fornecidos, podemos perceber que temos um outlier que seria as vitórias do piloto Lewis Hamilton, tendo um número de vitórias destoante de todo o resto do grid da fórmula 1. Além disso, percebe-se que Hamilton obteve maior sucesso na história da fórmula 1 com base nas suas vitórias.

Qual é a distribuição do número de corridas por país na história da Fórmula 1?



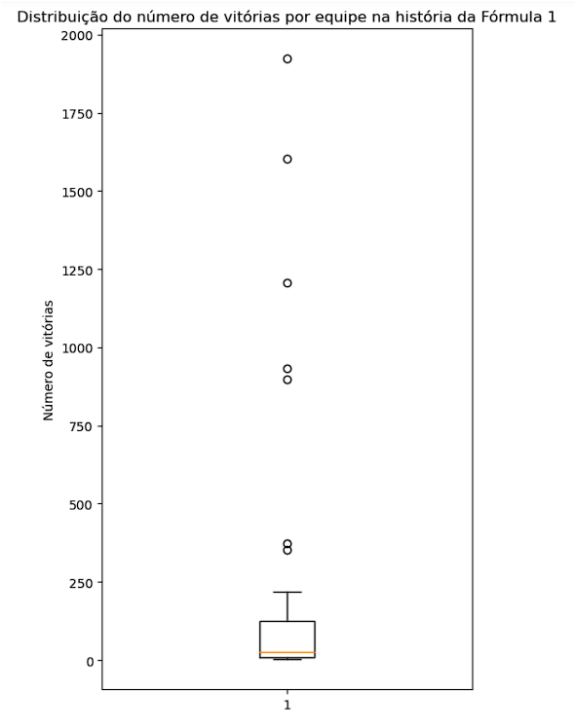
Tabela(s) necessária(s): Para a criação da representação da querie, em forma de um histograma, que lista a frequência de corridas em um determinado país, utilizamos de três tabelas dimensões (circuits, races e results). Além disso, esses csvs podem ser encontrados pelo nome de **"circuits.csv"**, **"races.csv"** e **"results.csv"**, que está na pasta **"csv_files_formula_one"** dentro do repositório. Por fim, para saber mais sobre como foi executado a consulta, abra o arquivo **analise_exploratoria_formula_one.ipynb** que está dentro da pasta **jupyter_notebook_code**

Descrição dos itens

- **Eixo Y:** Frequência/número de países que têm um determinado número de corridas dentro de um determinado intervalo
- **Eixo X:** Número de corridas

Análise: É possível identificar que, em toda a história da fórmula 1, a maioria dos pilotos são italianos. Além disso, é possível notar que temos pilotos de diferentes nacionalidades, o que reflete o caráter internacional do esporte.

Qual a relação de vitórias das construtoras e suas vitórias na Fórmula 1?



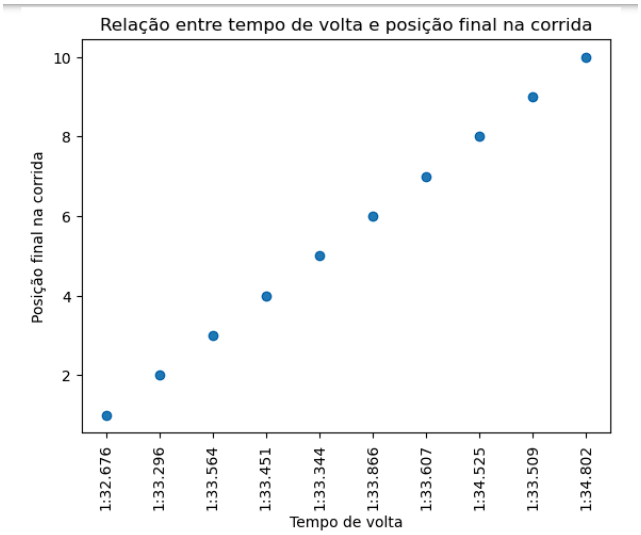
Tabela(s) necessária(s): Para a criação da representação da querie, em forma de um boxplot, que lista a quantidade de vitórias de cada construtora, utlizamos de duas tabelas dimensões (constructor e constructor_standings). Além disso, esses csvs podem ser encontrados pelo nome de **"constructors.csv"** e **"constructor_standings.csv"**, que está na pasta **"csv_files_formula_one"** dentro do repositório. Por fim, para saber mais sobre como foi executado a consulta, abra o arquivo **analise_exploratoria_formula_one.ipynb** que está dentro da pasta **jupyter_notebook_code**

Descrição dos itens

- **Eixo Y:** Número de vitórias por equipe

Análise: É possível perceber que a caixa do boxplot é pequena pelo fato de que existem muitos outliers, tendo um valor máximo de mais ou menos 250, portanto existe uma discrepância muito grande de vitórias entre as equipes, principalmente entre equipes de porte maior em comparação com as equipes de porte menor.

Qual é a relação entre o tempo de volta de um piloto e a posição final no circuito de interlagos?



Tabela(s) necessária(s): Para a criação da representação da querie, em forma de um boxplot, que lista a volta mais rápida no circuito de interlagos e a posição final que o piloto terminou, utilizamos de duas tabelas dimensões (results e lap_times). Além disso, esses csvs podem ser encontrados pelo nome de **"results.csv"** e **"lap_times.csv"**, que está na pasta **"csv_files_formula_one"** dentro do repositório. Por fim, para saber mais sobre como foi executado a consulta, abra o arquivo **analise_exploratoria_formula_one.ipynb** que está dentro da pasta **jupyter_notebook_code**

Descrição dos itens

- **Eixo Y:** Posição final na corrida
- **Eixo X:** Tempo de volta na corrida

Análise: Uma análise clara nesses dados é que quanto menor o tempo de volta, mais rápido o piloto é, e a chance de ele ganhar é mais alta. Além disso, vale ressaltar que a diferença de tempo entre cada piloto depende também, além da performance do piloto, de um bom carro.