Para o trabalho, foi escolhido um dataset da Fórmula 1 que se encontrea disponível na plataforma Kaggle. (<https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/data>).

O Data Set consiste em todas as informações sobre corridas de Fórmula 1, pilotos, construtores, qualificação, circuitos, tempos de volta, pit stops, campeonatos de 1950 até a última temporada de 2023.

O conjuto de dados é constituído por 14 cvs, mas foram selecionados apenas 4 deles.

O races.csv, tem as informações, de data, ano circuito, horas de começo da corrida e primeiro treino, de cada uma das 1120 corridas que se realizaram até ao final da presente temporada de 2023. Alguns valores como as horas de começo da corrida e do primeiro treino estão em falta, nomeadamente de GrandPrix mais antigos.

O constructers.csv, contém o nome, designações e nacionalidades de cada equipa que participa ou já participou em algum GrandPrix.

O drivers.csv é constituído por uma referência do piloto, assim como o seu código de três silabas, o primeiro e último nome, a data de nascimento, nacionalidade e em alguns casos o número que usam.

A tabela results.csv é um ponto de encontro das três tabelas indicadas anteriormente, pelo que contém a referência através da foreign keys das chaves primárias de outras tabelas. A tabela result tem então informação como o id da corrida, o id do piloto, o id do construtor para qual o piloto corria nessa corrida, o número do carro, a posição inicial do piloto na corrida, a posição final do piloto (não tendo/ tendo em conta desqualificações e corridas incompletas), e os pontos que o piloto acumulou naquela corrida.

É importante frisar que existem informações nas tabelas que não acrescentam valor ao trabalho, como por exemplo o url da página do Wikipédia de cada corrida que aconteceu ao longo dos anos.