Projeto-Netflix

Ana Rodrigues (2020143716), Ana Ferreira (2020128914), Beatriz Andrade (2020139743) e Daniel Madeira (2020145968)

Os dados usados para este projeto constam num ficheiro CSV, que podemos encontrar no website Kaggle (https://www.kaggle.com/shivamb/netflix-shows). Trata-se de uma fonte confiável, os dados estão atualizados, e o documento CSV que usámos foi atualizado há 4 meses. Após alguma pesquisa por outros documentos que poderiam servir para este trabalho, demos conta de que a Netflix não disponibiliza tantos dados quanto pensávamos, e isso contribuiu para a escolha deste pacote de dados.

Com a análise dos dados no nosso projeto, percebemos que existem valores em falta, podendo comprometer os resultados em certos pontos da investigação, em particular os valores em falta na coluna “country”. De resto, os valores “ausentes” não afetam o projeto, por estarem em falta em colunas irrelevantes para a análise que queríamos fazer.

Verificámos que não existem valores duplicados, não afetando a análise.

Os campos aos quais demos mais importância foram: a coluna “type”, onde conseguimos ver o tipo de programa (filme ou programa de TV); a coluna “release_year”, que mostra o ano de lançamento; “country”, que nos permite saber em que país foi produzido o filme ou programa de televisão; e “duration”, onde podemos ver a duração do filme ou programa de televisão, neste caso o que nos interessou foi a duração dos filmes.

No ponto em que vimos a duração média dos filmes, foi necessário retirar as palavras (“min”), para conseguir analisar apenas os dados numéricos e, assim, poder calcular a média. Esta foi a única correção que tivemos de fazer aos dados.

Não fizemos qualquer rejeição de dados, por isso, neste aspeto o nosso projeto não foi afetado.