Wrangle Report

Projeto We Rate Dogs - Twitter

Leandro Baruch

Coleta

Com o objetivo de realizar as atividades de Data Wrangling do projeto em questão, inicialmente eu coletei as informações que estavam em arquivos CSV e TSV, criando seus respectivos DataFrames. Na sequência, para coletar as informações restantes, utilizei o API Tweepy do Twitter, criando um terceiro DataFrame.

Análise

Como parte da Análise, realizei verificações visuais e programáticas para encontrar erros de arrumação e qualidade, anotei os erros encontrados para realizar a Limpeza. Ainda na fase da Análise, acessei o Twitter para verificar algumas informações "suspeitas". Como se trata de dados de uma conta que tem como propósito o entretenimento de seus seguidores, a falta de padrão existe em inúmeros lugares. Notas com numeradores maiores que denominadores, valores que aparentemente não fazem sentido me fizeram vasculhar mais de perto. Ao visitar alguns tuítes pude perceber que por exemplo, a nota 20/16 tinha como motivo a foto de alguns cachorros dentro de uma estátua do número 2016. Levando isso em consideração, fui flexível no momento de "corrigir" os dados. Aqueles que faziam sentido com o conteúdo, não foram considerados como erros. Outro ponto importante, encontrei bastante erros de datatypes dos valores. Valores de ID como float, ou integer foram bem comuns, dentre outros. Pude perceber diversos valores faltantes e também informações trazidas erradas. Segue a lista dos erros encontrados:

Arrumação

- As colunas `doggo`, `floofer`, `pupper` e `puppo` deveriam ser uma única coluna preenchida com sua respectiva classificação.
- A maior parte das informações estão no DataFrame `twitter_archive`, porém existem informações necessárias que estão no DataFrame `tweet_status`.

Qualidade

- Como motivação do projeto, só é necessário classificações originais, que não sejam retweets, portanto, retirar a linhas que são retweets de acordo com a coluna `retweeted_status_id`.
- Alguns cachorros estão com o nome None, `a`, `an`.
- Alguns denominadores estão com valores diferentes de 10.
  Coluna `rating_denominator`
- Alguns numeradores estão com valores suspeitos. Por ser uma conta do Twitter que tem uma proposta de entretenimento através de classificar cachorros, podemos entender alguns valores acima de 10, no entanto temos alguns casos de 1 ou 2 ocorrências que parecem errados. Coluna `rating_numerator`.

- A coluna timestamp está com o datatype errado. Atualmente está como string.
- Valores em branco na coluna expanded_urls.
- As colunas in_reply_to_status_id, in_reply_to_user_id, estão com o datatype errado. Atualmente estão como float.
- As colunas favorite_counts e retweet_counts possuem uma linha em branco.
- As colunas rating_numerator, favorite_counts e retweet_counts estão com o datatype errado. Atualmente estão como float.
- A coluna tweet_id está com o datatype errado. Atualmente está como integer.

Limpeza

Durante a parte da Limpeza, detectei alguns outros erros que eu não havia detectado anteriormente e seguindo a boa prática, para cada erro eu realizei o ciclo de Definir, Codificar e Testar. Pelo fato da qualidade dos dados estar muito ruim, e a equipe We Rate Dogs não seguir nenhum padrão para realizar suas classificações e escrever seus textos, dificultou algumas tarefas. Impossibilitou a automação de alguns ajustes usando for loops, como ajustar os nomes de alguns cães por exemplo, se tornou uma tarefa manual.