Wrangle Report

Introdução

Projeto Data Wrangling, o projeto foi realizado em 3 fases; Coleta de dados, Avaliação dos dados e Limpeza dos dados. Após isso os dados limpos foram armazenados para uma avaliação e para gerar um relatório.

Coleta de dados

Os dados foram coletados de 3 fontes diferentes, os arquivos de twitters (twitter-archive-enhanced.csv) e o arquivo de imagens (image-predictions.tsv') foram lidos por meio do método read_csv() presente na biblioteca Pandas. Os aquivos lidos foram armazenados na estrutura do tipo DataFrame.

Para a coleta das informações adicionais sobre os twitters foi utilizado a biblioteca tweepy, que permite realizar comunicação com a API do Twitter. Para cada identificador de twitter presente no arquivo de twitters, foi buscado as informações que foram armazenadas no formato json no arquivo tweet_json.txt. As informações foram recuperadas por meio do método read_json() da biblioteca Pandas e armazenadas em um DataFrame. Após isso as informações que seriam utilizadas foram selecionadas, o número de retweets e o número de favorite de cada twitter.

Avaliação dos Dados

Para a avaliação dos dados foram utilizados os métodos dataframe.info() e dataframe.head(). Os seguintes problemas foram detectados:

Qualidade

- Nomes de cachorros que estão incorretos. Ex.: 'a', 'the', 'an', 'None'
- Há twits que são retweets e devem ser retirados, já que queremos somente os twitters originais
- A coluna timestamp, do arquivo data_twitter_archive apresenta o tipo de dado errado.
- Falta imagens para alguns twittes
- Há previsões que não são cachorros
- Falta os valores de retweets e favorite para alguns twitters
- Nas colunas doggo, floofer, pupper, puppo 'None' não é tratado como null
- Estão faltando dados de expanded_urls, há somente 2297 valores.

Arrumação

- No dataframe de twitters as colunas doggo, floofer, pupper, puppo devem ser
- agrupados em uma só coluna
- A coluna de rating deveria ser única e não dividida em duas (numerador e
- denominador)
- As tabelas deveriam fazer parte de um único dataset

- Falta uma coluna para indicar a raça mais provável

Limpeza de Dados

Para a limpeza dos dados foram utilizadas as seguintes soluções.
Problema: Há twitters que são retweets e devem ser retirados, já que queremos somente os twitters originais

- Solução: Retirar registros de retweets e retirar colunas que se referem a eles retweeted_status_id, retweeted_status_user_id e retweeted_status_timestamp

Problema: As tabelas deveriam fazer parte de um único dataset, Falta imagens para alguns twittes, Falta os valores de retweets e favorite para alguns twitters e falta dados de expanded_urls, há somente 2297 valores.

- Solução: Unir os dataset em uma tabela, garantindo assim um dataset que possui twittes que possuem imagem e a contagem de retweets e favorites.

Problema: No dataframe de twitters as colunas doggo, floofer, pupper, puppo representam o mesmo tipo de dado, nessas colunas o 'None' não é tratado como null

- Solução: Criar uma coluna para receber esses valores e tratar 'None' como null, Há mais de classificação para alguns cachorros, para resolver esse problema iremos manter a primeira classificação.

Problema: A coluna timestamp, do arquivo data_twitter_archive apresenta o tipo de dado errado.

- Solução: Realizar um cast na coluna

Problema: - Há nomes incorretos para os cachorros. Ex.: 'a', 'the', 'an' e 'None'.

- Solução: Substitui dados incorretos por NaN

Problema: A coluna de rating deveria ser única e não dividida em duas (numerador e denominador)

- Solução: Criação de uma coluna para receber os valores de rating

Problema: Há previsões que não são cachorros e falta uma coluna para indicar a raça mais provável

- Solução: Criação de uma coluna para indicar qual a raça mais provável, desconsiderado as previsões que não são cachorros.