1. COLETA DE DADOS

- image_predictions.tsv foi baixado programaticamente;
- twitter-archive-enchanced.tsv foi baixado manualmente;
- A API do twitter, Tweepy, foi utilizada para baixar a quantidade de favoritos e retweets dos tweets cujo ID estava presente no arquivo twitter-archive-enchanced.tsv. As informações coletadas foram armazenadas em tweet_json.txt.

2. ANÁLISE DE DADOS

De forma geral, para cada tabela, imprimiu-se 30 linhas aleatórias, informações sobre quantidade de valores nulos em cada coluna e seus *data types*, bem como dados estatísticos das colunas com valores numéricos, quantidade de linhas presentes e de duplicatas.

A função value_counts() foi usada em colunas onde julgou-se necessário e fez-se essencial, junto à função sort_index(), para a percepção de alguns nomes impróprios para os cães avaliados.

Verificou-se que os IDs contidos no arquivo twitter-archive-enchanced.tsv não foram integralmente acessados pela tweepyAPI e que a tweepyAPI não pôde acessar todos os IDs contidos no arquivo twitter-archive-enchanced.tsv e a ausência de uma fração desses IDs no arquivpo image_predictions.tsv.

Testou-se, também, se a soma das porcentagens das raças previstas para cada imagem de cão era superior a 100%. Uma única linha falhou no teste, mas por apresentar uma porcentagem insignificantemente superior, foi justificada como arredondamente natural dos floats pelo Python.

2.1 SÍNTESE DAS QUESTÕES ENCONTRADAS

2.1.1 Questões de Qualidade

- IDs não mais acessíveis.

TODAS AS TABELAS

- *Datatypes* errôneos (coluna: tweet id).

TABELA twitter-archive-enchanced.tsv

- *Datatypes* errôneos (colunas: timestamp, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id);
- Algumas vezes a URL se repete (coluna: expanded_urls);
- HTML tags na coluna source;
- Valores presentes na coluna name não são nomes reais de cães. Esses valores começam com letra minúscula.

TABELA image_predictions.tsv

- Os nomes das raças são separados com underline ou um sinal de menos. Algumas vezes, a primeira letra é maiúscula e, em outras vezes, minúscula;
- Não se analisou as imagens de todos os tweets.

2.1.2 Questões de Arranjo

- A variável fases do cão estão dispostas em 4 colunas em vez de em uma única;
- Existem três tabelas para apenas duas formas de unidade observacional (tweet e cão).

3. LIMPEZA DOS DADOS

- Os IDs não mais acessíveis pela tweepyAPI foram removidos das tabelas;
- As colunas que listavam IDs foram transformadas dos *datatypes* int/float para string (pandas object), enquanto a coluna timestamp da tabela twitter-archive-enchanced.tsv foi transformada do *datatype* de string (pandas object) para datetime;
- URLs, quando duplicadas, tiveram suas duplicatas removidas.
- Apenas as string entre tags foram mantidas na coluna source da tabela twitter-archive-enchanced.tsv;
- Nomes de cães inválidos foram substituidos pela string (pandas Object) 'None';
- Nome das raças do cães passaram a ser separadas apenas por *whitespaces* e ter seu primeiro caractére maiúsculo e o restante minúsculo;
- Preencheu-se a tabela image_predictions.tsv com os IDs ausentes. O restante das colunas para esses IDs foram preenchidos com NaN;
- As colunas que representavam as fases do cão (doggo, floofer, pupper, puppo) foram unidas em uma única coluna 'stage';
- Parte da tabela twitter-archive-enchanced.tsv (colunas: tweet_id, text, source, expanded_urls, timestamp, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) foi fundida com a tabela tweet_json.txt, formando o arquivo twitter_archive_master.csv. Enquanto a tabela image_predictions.tsv foi fundida com o restante da tabela twitter-archive-enchanced.tsv, formando o arquivo tweet_dogs.csv.