Relatório sobre esforços de Wrangling

1. Coleta de dados

Para a obtenção do `twitter_archive_enhanced.csv` bastou realizar o download através do link disponibilizado, o que não apresentou dificuldades para ser realizado.
No caso do `image_predictions.tsv`, tive que realizar uma pesquisa para relembrar como, fazer o download de uma arquivo programaticamente. Isto foi resolvido utilizando a função **requests.get(url)**.
O maior problema encontrado foi para carregar dados usando tweepy, onde em um primeiro momento tive que realizar o cadastro na API do twitter. Posteriormente consegui fazer o donwload no twitter das informações de cada tweet_id e salvar os dados do json em um arquivo. Após isso, carreguei os arquivos e realizei o procedimento para fazer a conversão de json para dataframe.

2. Avaliação dos dados

Para realizar a visualização dos dados utilizei tanto a análise visual quanto a programática, ao realizar esta, cheguei aos seguintes problemas:

- Qualidade
    - problemas de qualidade no df_tweets:
        - Na coluna nome, a falta de informação está sendo preechida com outros valores além de None("the", "a")
        - As colunas retweeted_status_id,retweeted_status_user_id,in_reply_to_status_id e in_reply_to_user_id estão com o tipo float64 causando a perda de informação do valor do id
        - A coluna source está trazendo toda a informação de uma div
        - Os valores nulos da coluna expanded_urls estão sendo representados por nan
        - Na coluna expanded_urls o mesmo url está sendo repetido várias vezes em um mesmo registro
        - Existem tweets em que o texto começa com RT idicando um retweet
        - A coluna timestamp não está representando um valor timestamp
    - problemas de qualidade no df_tweets_data
        - Registro faltantes(1457 de 2356)

- problemas estruturais
    - As colunas retweeted_status_id,retweeted_status_user_id e retweeted_status_timestamp do df_tweets não são úteis já que não queremos retweets.
    - Os dados do dataframe df_tweets_data fazem parte do df_tweets
    - As colunas doggo, floofer, pupper e puppo respresentam a mesma informação que é o estagio do cão.
    - O data frame `df_images_predictions` deve ser uma coluna no df_tweets indicando a raça do cão caso seja possível

3. Limpeza de dados

Para realizar a limpeza dos dados, foi imprescindível percorrer os dataframes para arrumar o problemas encontrados. Funções para remover colunas e alterar o tipo de colunas também foram necessárias.