Data Wrangling

Coleta

Para o projeto era necessária a coleta de três diferentes fontes de dados:

1.Arquivo ‘WeRateDogs’
Disponibilizado por um link na descrição do projeto. Após o download, foi carregado em um data frame chamado ‘df_we_rate_dogs’.

2.Arquivo com as previsões de imagens
O link para o arquivo foi disponibilizado para realizar o download e foi feito de forma programática. O arquivo foi carregado no data frame ‘’df_image_prediction’.

3.Informações adicionais (retweet e favorite count)
Após realizar uma consulta usando o API do twitter, os dados foram armazenados no arquivo ‘tweet_json.txt’ e depois, carregados no data frame ‘df_retweet_favorite’.

Avaliação

Após feita a coleta, os dados foram acessados a fim de avaliar possíveis problemas de qualidade e organização.
Usando a avaliação visual, foi notado muitos valores ausentes em algumas colunas como ‘in_reply_to_status_id’ e ‘in_reply_to_user_id’, algumas inconsistências na coluna ‘name’ e muitos valores ‘None’ nos estágios.
Na avaliação programática foi possível verificar melhor os problemas como diferença considerável entre o número de dados em cada arquivo.
Abaixo segue os problemas identificados:

Qualidade

Tabela we rate dogs

- Faltando informação em "in reply" (in_reply_to_status_id, in_reply_to_user_id)
- Faltando informação em "retweeted" (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- Faltando informação em "expanded url" (expanded_urls)
- Tipos errados (retweeted_status_timestamp, timestamp)
- Registros com 'tweet_id' deletado (2356 vs. 2338 in retweet_favorite_count table)
- Como um dos pontos chave era que apenas classificações originais com imagens importavam, alguns registros devem ser desconsiderados (2356 vs. 2075 in image_prediction table).

- Inconsistência de valores na coluna name

Tabela image_prediction

- Tipos errados (p1, p2, p3, img_num)
- Lowercase (p1, p2, p3)
- Valores Nan (retweet_count and favorite_count columns) em image_prediction merged table

Tabela retweet_favorite_count

- Como um dos pontos chave era que apenas classificações originais com imagens importavam, alguns registros devem ser desconsiderados (2338 vs. 2075 in image_prediction table).

Organização

- Coluna tweet_id em retweet_favorite_count table está duplicado em we rate dogs e image_prediction
- Coluna rating_denominator em we rate dogs table tem quase que apenas um valor (10).
- As colunas doggo, floofer, pupper e puppo devem ser unificadas.
- Colunas expand_urls e tweet_id em we rate dogs possuem a mesma informação.

Limpeza

Após feita a avaliação e documentação dos erros verificados, foi feita a limpeza dos dados. O primeiro passo foi realizar a cópia dos data frames.
As colunas com pouca informação foram descartadas, os tipos foram corrigidos, as colunas com os estágios foram unificadas em apenas uma, colunas sem padrão de letra maiúscula e minúscula foram tratadas e no final foi realizada uma junção para termos apenas uma tabela com os dados coletados.