Introdução

Esse é um relatório que irá descrever o que foi feito durante o processo do data Wrangling do projeto WeRateDogs. Esse documento terá 3 seções que em resumo são os principais topicos em um Data wrangling, que consiste em:

• Coletar dados

• Avaliar dados

• Limpar dados

Coleta

O processo de coleta desse projeto foram divididos nas seguintes etapas:

• O arquivo WeRateDogs. Baixado manualmente pelo link fornecido na página do projeto da Udacity

• O arquivo image_predictions.tsv baixado programaticamente usando a url hospetada https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

• O arquivo tweet_json.txt baixado programaticamente usando a API Tweepy.

Avaliação

A avaliação foi feita usando comandos básicos como .head() e .info() para entender o que tinha nos arquivos e identificar visualmente e programaticamente os problemas, foi possível encontrar tanto problemas de qualidade e arrumação.

Qualidade

• Columa "name" estar imcompleta com "Nome"

• Coluna "name" tem nomes errados

• Alguns são retweets

• Alguns não tem imagem

• Na Columna "rating_numerator", quando tem decimal,so estar pegando a parte decimal do número

• Nas Columnas "rating_numerator", "rating_denominator" tem alguns erros, alguns dados não são realmente notas

• A Columna "source" não estar legível

• Aguns data type esyão errados

Arumação

• Nas colunas "doggo", "floofer", "pupper" e "puppo" deveriam ser apenas uma coluna

• Todas as tabelas devem juntar em uma

• Algumas colunas não são apropriadas para análise

Limpeza

A limpeza foi realizada de forma programátia, seguindo a metodologia de 3 passos, Definir, Codificar e Testar, a definição foi feita na maior parte em forma de comentários no código. Antes de iniciar a limpeza foram feito copias dos arquivos para não ocorrer o risco de modificar permanentemente e ter que realizar o download novamente.

Outro detalhe importante foi que antes de acatar os problemas de qualidade, foi resolvidos os problemas de arrumação, assim os cogidos que objetivaram resolver os problemas de qualidade foram mais efetivos.

Algum problemas foram de qualidade foram descobertos durante o processo de limpeza, forçando a interação dos processos de Data Wrangling.

Conclusão

Data wrangling é umas das principais atividades de quem trabalha com dados, sem ela a analise fica completamente impressiva e podendo resultar em tomadas de decisão erradas e perda de tempo. Se a visualização e a analise fossem feitas antes do wrangling provavelmente as conclusões seriam completamente diferentes.