Data Wrangling

Introdução

O objetivo deste projeto é colocar em prática todo aprendizado que obtivemos no curso Data Wrangling do Nanodegree de Fundamentos em Data Science II, da Udacity.

O dataset que realizamos o data wrangling é um arquivo de tuítes da conta @dog_rates.

Este documento mostra o processo de limpeza e organização dos dados, visualizações e insigths realizados.

Detalhes do Projeto

- Data wrangling, que consiste em:
  - Coletar dados
  - Avaliar dados
  - Limpar dados
- Armazenar, analisar e visualizar dados wrangled
- Elaborar relatórios sobre:
  - 1) seus esforços de data wrangling;
  - 2) suas análises e visualizações de dados

Coletando dados para o projeto

1. twitter_archive_enhanced: O arquivo WeRateDogs fornecido pela Udacity;
2. image_predictions: arquivo que baixei programaticamente e que contém as previsões das imagens dos tuítes feita pelo algoritmo de redes neurais da Udacity;

3. tweet_json.txt: arquivo que criamos através da API tweepy para coletar a contagem de retweets e favoritos (curtidas) dos tuítes fornecidos. Com ele criamos um dataframe para realizar as análises

Acessando os Dados

Uma vez que coletamos os dados, transformamos todos em dataframe's do pandas para iniciar os trabalhos de limpeza e análise;

Limpeza dos Dados

Problemas de Arrumação

- Juntar as tabelas 'twitter_archive' e 'df_tweet_json' em um dataframe único
- As colunas 'doggo', 'pupper','puppo' do DF twitter_archive devem ser uma coluna só (stage)

Problemas de Qualidade

twitter_archive table

1. Não vamos utilizar as colunas: "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp". Podemos descartá-las;
2. A coluna "expanded_urls" possui valores faltantes. Devemos descartar estes registros uma vez que ratings sem imagens não faz sentido para nossa avaliação;
3. datatype da coluna 'rating_numerator' deve ser em decimais (float);
4. datatype da coluna 'timestamp' esta incorreta;

5. o padrão do rating_denominator é 10. outros valores podem ser erros;
6. rating_numerator com valores extremos devem ser descartados
7. alguns nomes com letras minúsculas. Colocar todos em letras maiúscula
8. Alguns nomes dos cachorros estão errados, como por exemplo: "a", "the", "an", "quite"

image_predictions table

- nomes das raças dos cachorros fora do padrão
- renomear nomes das colunas para que fiquem legíveis

tweet_json table

- nada para ser arrumado ou limpo