# Projeto 2 - Data Wrangling e Analyzing no twitter WeRateDogs

Curso: Fundamentos de Data Science II
Aluno: Filipe Pegollo

## Relatório de Data Wrangling

## Coleta de Dados

Os dados deste projeto foram obtidos através de três arquivos:

- twitter-archive-enhanced.csv: Este arquivo contém dados básicos de milhares de tweets do usuário do Twitter @dog_rates, também conhecido como WeRateDogs. Foi fornecido pela Udacity e foi baixado manualmente.
- image-predictions.tsv: Foi gerado através de um algoritmo de rede neural que analisou cada imagem e gerou previsões de classificações. Este arquivo está hospedado em um servidor da Udacity e foi baixado programaticamente utilizando a biblioteca requests.
- tweet_json.txt: Através da API Tweepy foi feita uma conexão com Twitter para baixar dados adicionais dos tweet's, cada tweet foi armazenado em uma linha, foram recebidos em formato JSON. E então estas informações foram salvas em um arquivo TXT.

## Avaliação

Após a avaliação dos dados visualmente e programaticamente, foi possível identificar alguns problemas que precisavam ser corrigidos antes da fase de análise.

### Problemas de Qualidade

#### Tabela twitter-archive

1. Nomes na coluna 'name' com letras minúsculas são incoerentes
2. Tipo de dados incorreto na coluna 'timestamp'
3. A coluna 'source' contém códigos HTML onde só deveria haver o texto contendo a fonte
4. Existem dados de classificações com retweets, precisamos apenas de publicações originais
5. Como não serão avaliados retweets as colunas 'retweeted_status_id', 'retweeted_status_user_id' e 'retweeted_status_timestamp' não são necessárias.
6. Não serão feitas operações aritiméticas com a coluna 'tweet_id' por isso o tipo de dados deve ser string
7. A coluna 'rating_numerator' possui alguns valores incompatíveis
8. Existem valores diferentes de '10' na coluna 'rating_denominator' que podem atrapalhar as análises

#### Tabela image-predictions

9. Não serão feitas operações aritiméticas com a coluna 'tweet_id' por isso o tipo de dados deve ser string

**Tabela tweet_json**

10. Não serão feitas operações aritiméticas com a coluna 'tweet_id' por isso o tipo de dados deve ser string

### Problemas de Arrumação

1. As colunas 'doggo', 'floofer', 'pupper','puppo' mostram apenas uma variável
2. Os dados são referentes aos mesmos tweets mas estão em tabelas separadas

## Limpeza

### Problemas de Qualidade

**Problema**

1. Nomes na coluna 'name' com letras minúsculas são incoerentes.

**Solução**

Para corrigir este problema os nomes foram substituídos por 'None', que posteriormente foram substituídos por Nulo.

**Problema**

2. Tipo de dados incorreto na coluna 'timestamp'.

**Solução**

O tipo de dado da coluna foi convertido de string para data.

**Problema**

3. A coluna 'source' contém códigos HTML onde só deveria haver o texto contendo a fonte.

**Solução**

Foi utilizado um comando para remover as tags HTML.

**Problema**

4. Existem dados de classificações com retweets, precisamos apenas de publicações originais.

**Solução**

O dataframe foi filtrado utilizando os comandos ‘isnull’ e ‘notnull’ para remover retweets e registros sem ID.

**Problema**

5. Como não serão avaliados retweets as colunas 'retweeted_status_id', 'retweeted_status_user_id' e 'retweeted_status_timestamp' não são necessárias.

**Definição**

As colunas que não fariam parte do projeto foram excluídas.

**Problema**

6. Não serão feitas operações aritiméticas com a coluna 'tweet_id' por isso o tipo de dados deve ser string

**Definição**

Os dados da coluna foram convertidos de int para string

**Problemas**

7. A coluna 'rating_numerator' possui alguns valores incompatíveis
8. Existem valores diferentes de '10' na coluna 'rating_denominator' que podem atrapalhar as análises

**Definição**

Para corrigir estes dois problemas foram excluídos numeradores maiores ou iguais a 15 e denominadores diferentes de 10. Depois foi criado uma coluna de classificação calculada com base nas duas colunas, chamada rating e por fim foram excluídas as colunas que não seriam mais necessárias.

**Problema**

9. Na tabela image-predictions não serão feitas operações aritiméticas com a coluna 'tweet_id' por isso o tipo de dados deve ser string

**Definição**

Os dados da coluna foram convertidos de int para string

**Problema**

10. Na tabela tweet_json não serão feitas operações aritiméticas com a coluna 'tweet_id' por isso o tipo de dados deve ser string.

**Definição**

Os dados da coluna foram convertidos de int para string.

**Problemas de Arrumação**

**Problema**

1. As colunas 'doggo', 'floofer', 'pupper','puppo' mostram apenas uma variável.

**Definição**

Os dados foram unidos em uma única coluna, depois as colunas que se tornaram obsoletas foram excluídas.

**Problema**

2. Os dados são referentes aos mesmos tweets mas estão em tabelas separadas.

**Definição**

Os dados de diferentes dataframes foram unidos em uma única tabela utilizando o comando ‘merge’.