WeRateDogs

A Data Wrangling Report

> By Bukola Ajayi

Introduction

> This project was focused on data gathering, wrangling, and visualisation analysis from WeRateDogs Twitter.

I thoroughly assessed and cleaned this the entire dataset, this required exceptional effort, so only a subset of its issues (eight quality issues and two tidiness issues at minimum) are reported and communicated.

During this project, data was collected from various sources and in a variety of formats.

The wrangling process consisted of:

1. Data Gathering : The three pieces of data were gathered and represented as pandas dataframes, which included the following steps.

• The WeRateDogs Twitter archive (file on hand, manual download of ' twitter-archive_enhanced.csv '). I downloaded this file programmatically using the Requests library from a provided URL. It included tweets from November, 2015 through August 1, 2017 of basic tweet data (tweet ID, timestamp, text, etc.) that contained columns that were extracted programatically: the rating numerator, rating denominator, dog's name, and dog stages (doggo, floofer, pupper, and puppo).

•Image predictions for tweets ('image-predictions.tsv'): This consists of the tweet IDs, image urls, the dog confidence and prediction algorithm.

• The Twitter API and Python's Tweepy library were used to save each tweet's entire set of JSON data (with a minimum of tweet ID, retweet count, and favorite count) in a file called "tweet_json.txt."Each tweet's JSON data was written and then read to text file line by line into a Pandas DataFrame only including the desired variables; tweet IDS, retweet count, and favorite count.

2. Data Assessing

3. Data Cleaning

4. Storing, analyzing, and visualizing the wrangled data in order to draw out different insights about the dogs' information.

5. Visualising and reporting my data analyses.

6. Documenting for data wrangling steps and genarating the insights I discovered throught the entire process.