WeRateDogs Data Wrangling Project Report

Introduction

In this report, I outline the steps taken to Wraggle the Twitter archive data for the user WeRateDogs with username @dog_rates. This archive consists of 2356 basic tweet data of the 5000+ data collected between the November of 2015 and 2017.

WeRateDogs is a Twitter account that rates peoples' dogs with a humorous comment about the dog.

Data Wrangling is a process that involves gathering data, assessing, cleaning and unifying messy / complex datasets for easy analysis and visualization.

The objective of this project is to demonstrate the steps taken in gathering data from different sources, assess it, clean, store and analyze to draw conclusions for reports.

Data Gathering

Three different sets of data were gathered from different sources;

1. **WeRateDogs Twitter Enhanced Archive;** This was provided and was downloaded manually as a csv from the Udacity servers. It consisted of 2356 basic tweet data.

2. **Image predictions;** This was downloaded programmatically from the Udacity servers with the help of the Python's Requests library. This file (image_predictions.tsv) contained tweet image predictions, breed of dogs present in each tweet as according to a neural network classifier.

3. **Tweets Json & Twitter API;** This data was gathered by querying the Twitter API using Tweepy (Python Request library for Twitter) for tweet data based on the tweet Ids available in the Twitter Enhanced Archive. As a Twitter API Use

Policy, before one queries this data from the Twitter Server, One has to signup for a Twitter developer account which is later reviewed before approval. I was successful in my request for the Twitter Elevated access and this was used to generate the tokens for querying the data.

Data Assessment

This is the second step in the Data wrangling process. In this step, we inspect the datasets earlier collected for issues in terms of quality and tidiness.

Quality issues are issues that have to do with the contents of the data. These may include; duplicates, NaN values, inconsistent data types among others. While;

Tidiness issues are issues that involve the structure of the data.

In this step, both Visual and Programmatic assessment methods were used to identify the various issues in the data.

In Visual assessment, data was loaded in the Excel sheets where identification of the issues was done through scrolling the data. Some of the issues identified with this method included;

- **Tidiness issues:** Columns 'doggo', 'floofer', 'pupper', and 'puppo' in the twitter enhanced archive needed to be in one column which could be named dog_stage and
- **Quality Issues:** Some of the columns in the image prediction dataset were not descriptive i.e. p1, p1_conf, among others.

In Programmatic Assessment, Pandas methods such as pd.info, pd.describe, pd.name.value_counts, among others were used to identify the issues in the three datasets. A number of quality and tidiness issues were spotted in the three datasets.

Cleaning

Before cleaning the datasets, copies of the datasets were made to avoid messing with the original datasets.

Cleaning was done and most of the issues found in the three datasets were addressed one at a time. This cleaning was made based on the principles of **Define**, **Code** and **Test.**

Data Storage

Upon successful cleaning process, a master dataset was created by merging the three datasets together. This master dataset was later stored for later use and this was done with the help of pandas method pd.to_csv.

Analysis and Visualization

A selective analysis was made on the final master dataset to draw conclusions on;

1. What was the rating distribution?
2. What were the common dog stages?
3. What was the correlation between Retweets and Favorite counts?

Limitations

- Given the limited time, a partial assessment and cleaning was carried out on the datasets. More work can be done to make the final dataset cleaner
- A partial analysis also made, deeper analysis and visualizations can be drawn from the dataset
- The querying of the data from the Twitter api returned errors to some of the tweet Ids, this might have altered the final outcome of the process

Thanks;

Author: Wafula Abdalah Hassani, ALX-DAND 2022 Student

Github: https://github.com/Garande