Wrangle Report

WeRateDogs Twitter Data

Udacity Project 2

Data Analysis using Python

Introduction

The dataset being wrangled (and analyzed and visualized) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

The project aimed at getting data from web based sources, identifying tidiness and quality issues and cleaning the data.

Steps taken in Wrangling Process

1. Gathering the Data
2. Assessing the Data
3. Cleaning the data
4. Storing the Data
5. Analyzing the Data
6. Reporting the Data

Gathering the Data

The data came in three (3) forms, a csv file, a web link and tweets that needed to be scrapped. I loaded the csv file into the jupyter notebook using pandas read_csv function. I then used request library to download the tweet image prediction data then used the tweepy library to scrape tweets from Twitter's API. The processes were relative easy although the scraping of tweets took a lot of time to run.

Assessing the Data

In this section, I detected and documented **quality issues** and **tidiness issue**. Using both **visual assessment programmatic assessment** to assess the data.
Only original ratings were used. out of the 5000+ tweets in the data containing dog ratings were used.
One of the quirks of the WeRateDogs twitter page was in the ratings of the dogs, the numerators were grater than the denominators. It was quite interesting to see and analyze.

Visual assessment was done by visually inspecting the dataset for quality and tidiness issues.

Programmatic assessment was done using python functions and methods such as .info(), .nunique(), .describe()

Quality Issues Identified

- Twitter archived data

- Some inaccurate names of dogs -accuracy issues
- 181 Retweeted data present -validity issues
- Some numerator are more than the denominator but that would not be corrected because some people rate their dogs more than the highest rating -validity issues
- Timestamp is an object datatype instead of a datetime datatype

- Image prediction data

- Non-descriptive column names such as p1_dog, p2, p3, etc
- 66 Jpg_url are duplicated. This means 66 rows have the same image but different Tweet Id
- Inconsistency in the name of the dogs

- Extended tweets data

- Many unusable columns in the dataset, filled with NaN or is just not relevant
- Non-original data is present
- Id and id_str are duplicated columns in the extended tweets data
-

- Combined problem

- Source data has unnecessary characters present in both Twitter archived and Extended tweets

Tidiness Issues Identified

- Doggo, Floofer, Pupper and Puppo are dog classifications that could be under one column
- Timestamp can be split into Year, Month and Day
- The prediction and confidence level could be reduced into two columns, keeping the most likely in the Image Prediction data
- All tables can be merged into one Dataset

Cleaning the Data

In this section, I cleaned all the issues documented while assessing.
The result was high-quality and tidy master pandas DataFrame data according to the rules of tidy data

- Twitter Archive Data

Issue #1: Inaccurate names of dogs
From the assessment of the unique names of dogs, there are values that start with lowercase letters and don't seem to be names of dogs. The solution was to replace the inaccurate data with a none value.

Issue #2: Keeping only original ratings
181 observations are retweets and as such were removed from the analysis

Issue #3: Inconsistent datatype for timestamp
Timestamp is an object datatype instead of a datetime datatype. This was easily solved by converting it to datetime datatype

Issue #4: Dog stages are separated into different columns
The Dog stages are divided into doggo, floofer, pupper and puppo. They were combined under a dog stage column

- Twitter Archive Data

Issue #5: Unnecessary columns in the dataset

There are multiple columns that are not necessary to the dataset. These columns were dropped.

Issue #6: Source column contains unnecessary data

The source column still contains the html href tag. This was removed using regular expressions

- Image Prediction Data

Issue #1: Duplicated image links

There are some links for images that are duplicated in the jpg_url column. This can be resolved by removing the duplicates.

Issue #2: Multiple prediction and confidence column

The prediction and confidence level was reduced into two columns, keeping the most likely.

Issue #3: Inconsistent dog breed names

The names of the dogs are inconsistent with an underscore(_) separating them. This was resolved by replacing the underscore with a space and making each starting letters capital

- Extended Tweets data

Issue #1: There are some unoriginal tweets included

Some tweets are retweeted or in reply to other tweets. The analysis only requires original tweets. This was solved by removing all the tweets that are not original

Issue #2: Most columns are not needed for the analysis

There are many included columns that would not be needed for analysis. The solution was to drop such columns.

Storing the Data

After cleaning the datasets, I merged them together under one master dataset using a left merge. I then saved the cleaned master dataset into a csv file 'twitter_archive_master.csv'.

Analyzing the Data

In this section I analyzed and visualized the dataset to draw out some insights

- 2016 had the highest number of tweets
- The month with the highest number of tweets is **December,** while the month with the lowest number of tweets is **August**

- The most popular dog breed is Golden Retriever
- Charlie and Lucy were the most popular dog names
- **Twitter for iPhone** is the most popular tweet source
- The stage with the highest number of dogs is **Pupper**
- There is a correlation between number of retweets and number of favourite

Conclusion

In this wrangle act, data was gathered, inspected for quality and tidiness issues, wrangled, cleaned and visualized. Despite the uniqueness of WeRateDogs twitter page, the project was carried out in a way as to account for such uniqueness. The insights gathered above are important in showing the admin of WeRateDogs twitter page the general performance, liked and dislikes of their followers and their dogs. WeRateDogs have a lot of potential as a brand and as a twitter page.