ACT REPORT

Data Wrangling and Assessing insights from Final Data

By

Ebunoluwa Alabi

29-11-22

Introduction

Real-world data rarely come clean. Using Python and its libraries, you will gather data from various sources and formats, assess its quality and tidiness, and then clean it. This is called data wrangling.

This project walks through cleaning data from different sources and combining them to draw critical insights.

Data Gathering

The data was collected from the following sources;

1. The WeRateDogs Twitter archive provided in the project overview by Udacity, The “twitter-archive-enhanced.csv” contains basic information like tweet-id, text, timestamp, etc.
2. The image prediction file was supplied by Udacity as well and includes the dog breeds and the level of prediction accuracy.
3. Twitter API (Tweepy) was used to collect data like the latest tweets from the @dog_rates handle and other information I found interesting.

Assessing Data

With the data collected, I started to assess the quality of the data collected;

1. Quality dimensions
    a. **Completeness**; there is a lot of missing data from the different sources
    b. **Validity**; some data pieces weren’t necessary so they were dropped
    c. **Accuracy**; Some of the data accuracy is hard to verify
    d. **Consistency**; some of the rating data wasn't consistent
2. Tidiness requirements
    a. Each column should carry a single data type
    b. Each row should contain all relevant data

c. Each type of observation unit forms a table

Cleaning Data

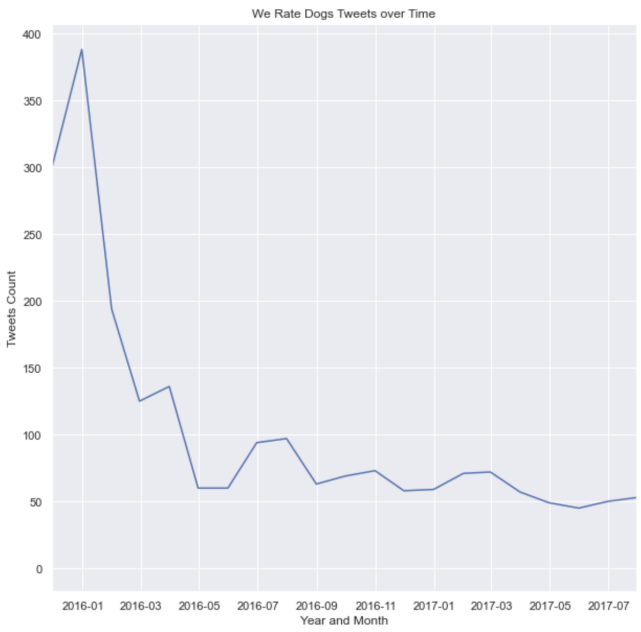
In cleaning the data, the following steps were followed;

1. Define; I decided exactly what needed to be cleaned
2. Code; Wrote the code to clean it
3. Test; Then tested to make sure it was clean

Analysing and Visualization

The following insights were drawn

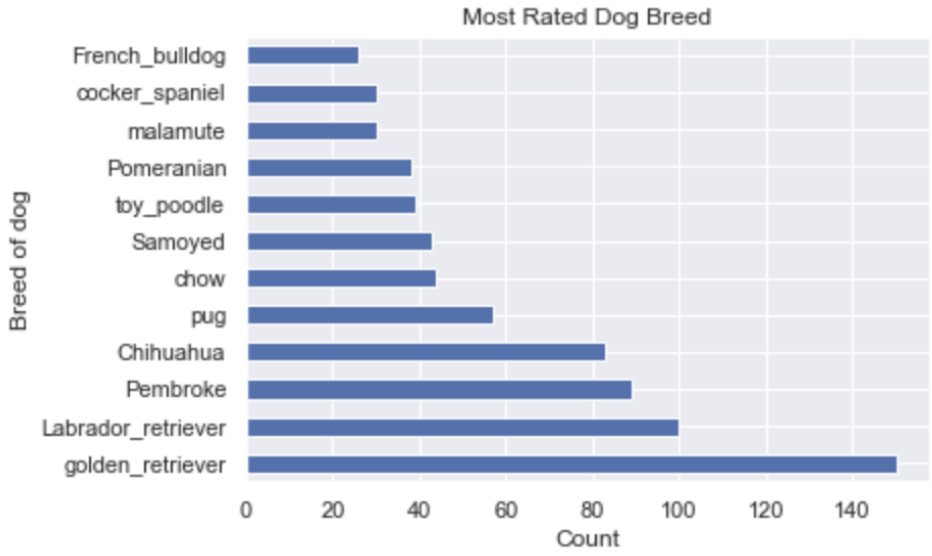
1. Tweets over time



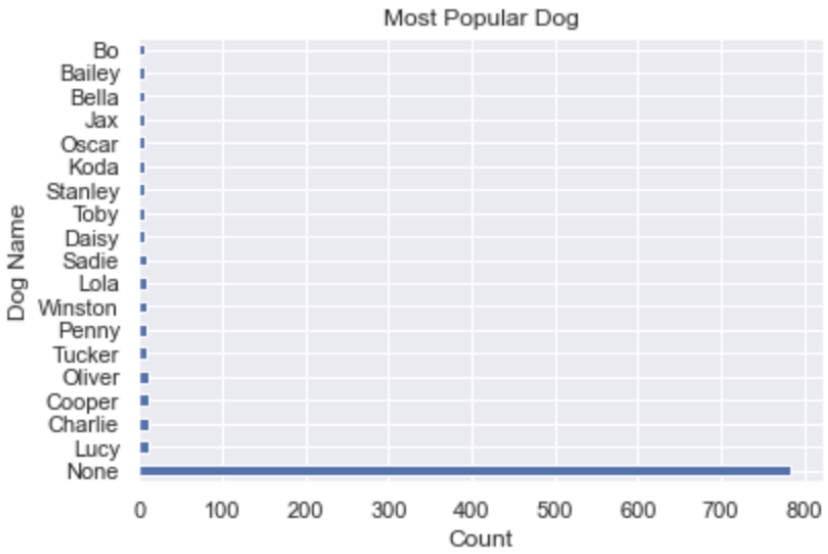
From Jan 2016, the number of tweets kept declining for the next year, suggesting that the general interest in dogs could have declined in this time or the weratedogs activity declined over time.

2. Most rated dog breeds

The most popular dog breed appears to be the Golden Retriever, the Labrador Retriever follows quickly after suggesting that these dogs might be in high demand. This information can help the @dogrates drive more traffic using these names.



3. Most popular dog names



The most popular dog name appears to be Lucy followed by Charlie (ignoring the None label). Suggesting dog owners prefer this name and users are likely to engage more with these names.

Conclusion

This report documents the data wrangling to insight process, it's important to note that a lot more was done and is provided in the wrangling report.