26th August 2022

Report

On Data Wrangling, Analysis and Insights

By:

Thomas Oluwasegun Olubunmi

Introduction:

In the real world, data hardly come clean. Cleaning is a fundamental step in data analysis process as it greatly increases the integrity of the data. Good data analysis results rely heavily on the reliability and integrity of the data. Data analyst clean data in order to ensure the accuracy and reliability of the data.

Usually, dirty data can be as a result of duplicate data, outdated data, incomplete data, inconsistent data, incorrect or inaccurate data and many more.

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

This project is about using Python and its libraries, to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

Gather:

The data for this project were gathered from the following sources:

- The WeRateDogs Twitter archive. The ‘twitter-archive-enhanced-2.csv’ file was provided by Udacity.
  This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- The Tweet image prediction, i.e., what breed dogs (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students (Like me).
- Twitter API and Python’s Tweepy library to gather each tweet’s retweet count and favourite count.

Assess:

Once the data is available, I then assessed the data both visually and programmatically to discover quality and tidiness issues.

Few of the quality’s issues are inconsistency, incompleteness, inaccuracy.

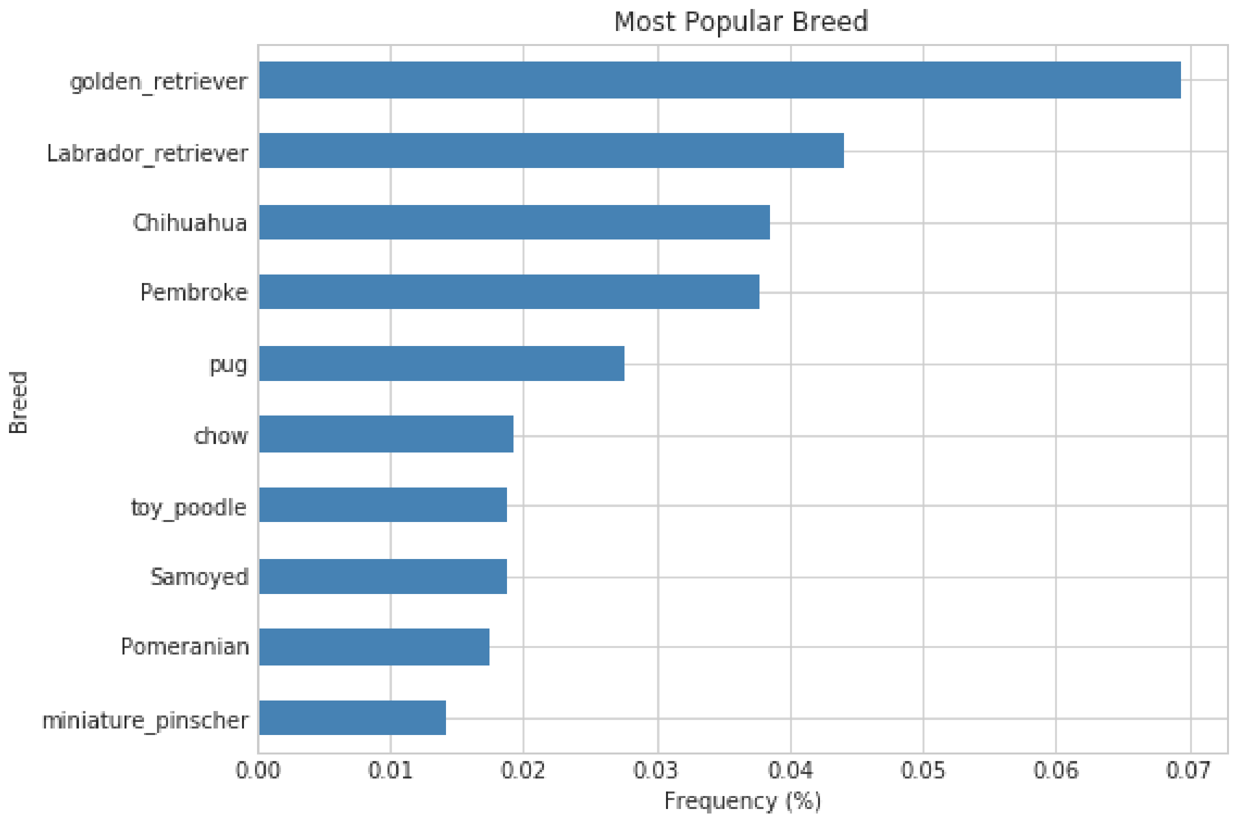
Clean:

Data cleaning is the most difficult phase of data wrangling process. Data analyst spend a lot of time here and the process Is highly iterative. In this project I followed the define, code and test process of cleaning data.

Analysis and Visualization

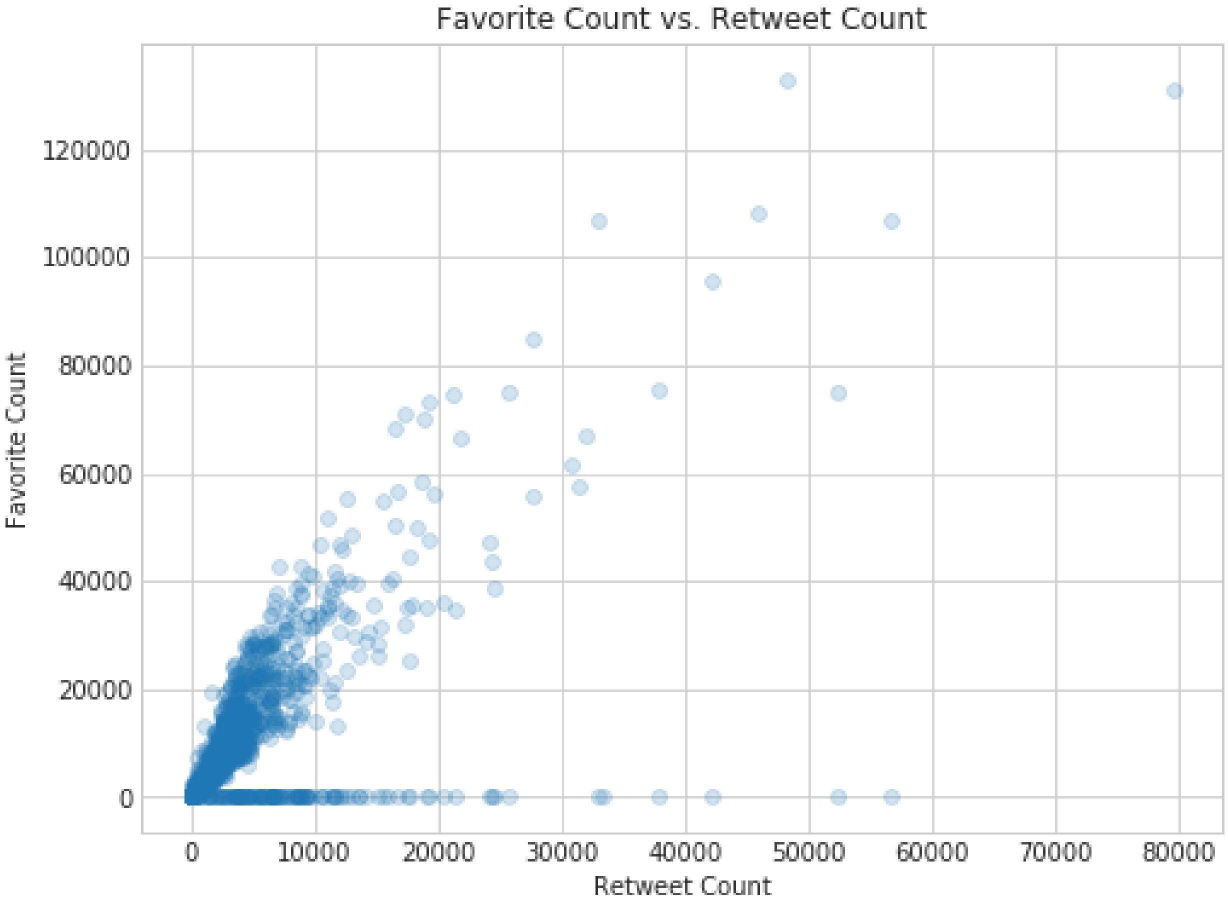
1. **Most Popular Dog Breed:**

   Golden retriever is the most popular dog breed. Labrador retriever is the second most popular breed. Chihuahua retain the third spot. The page owner could use this information to create targeted marketing efforts for certain breed that aren't so popular to increase their popularity and can also take the advantage of the popular breed to drive user traffic to the page.



2. Correlation between Favourite count and Retweet count:

It was observed that a positive correlation exists between favourite count and retweet count. This relationship is vital for the owner of WeRateDogs twitter account owner to identify what need to be done to draw more user traffic to the page.


Favorite Count vs. Retweet Count
Favorite Count
Retweet Count