WRANGLE AND ANALYZE DATA

Act Report

JUNE 20, 2019

5TH PROJECT

Udacity Data Analysis Nanodegree

Contents

1. Abstract 2
2. Introduction 2
3. Data Visualization 3
   3.1 Most Sources Used 3
   3.2 Number of Tweets Over Time 3
   3.3 Top 10 Popular Dogs 4
   3.4 Retweet Vs Favorite Count 5
4. References 6

Figures

Figure 1 Most Sources are Used 3
Figure 2: Number of Tweets Over Time 4
Figure 3: Top 10 Popular Dogs 5
Figure 4: Retweet vs Favorite Count 5

1. Abstract

A lot of data around the word some of those data aren't cleared , structured, or organized, based on **DOMO** site *"Over 2.5 quintillion bytes of data are created every single day, and it's only going to grow from there. By 2020, it's estimated that 1.7MB of data will be created every second for every person on earth."*[1].

In this project I will practice what I learned in data wrangling Udacity's Data Analysis Nanodegree program. The dataset that I will be wrangling is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**.

After I gathered, assessed, cleaned data (See wrangle_report for more details) now we have a cleaned data and I will analyze the data (tweets) then visualization my insights. I created three insights with visualization, and I will explain those insights in this report.

2. Introduction

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."*[2]. So, in the wrangle_report I finish the first steps (gathered, assessed, cleaned) now I will analysze those tweets by pandas library, some of techniques, and methods then I will use matplotlib and seaborn libraries to visualization my data. In this report the data analyzing consists of:

- Most sources used based on our final dataframe.
- Number of tweets over time.
- Top 10 popular dogs.
- Retweet vs favorite count based on 5 popular dogs

The most tools, libraries and programming language used in this project are:

- Python
- Pandas Library
- Numpy Library
- Requests Library
- Tweepy Library
- Json Library
- Matplotlib Library
- Seaborn Library
- Jupyter Notebook
- Twitter's API

3. Data Visualization

"Data visualization is viewed by many disciplines as a modern equivalent of visual communication"[3]

3.1 Most Sources Used

There are three sources in our dataset:

a. Twitter for iPhone

b. Vine - Make a Scene

c. Twitter Web Client

d. TweetDeck

After cleaned the data remained just three sources a, b and d. (See Figure 1)

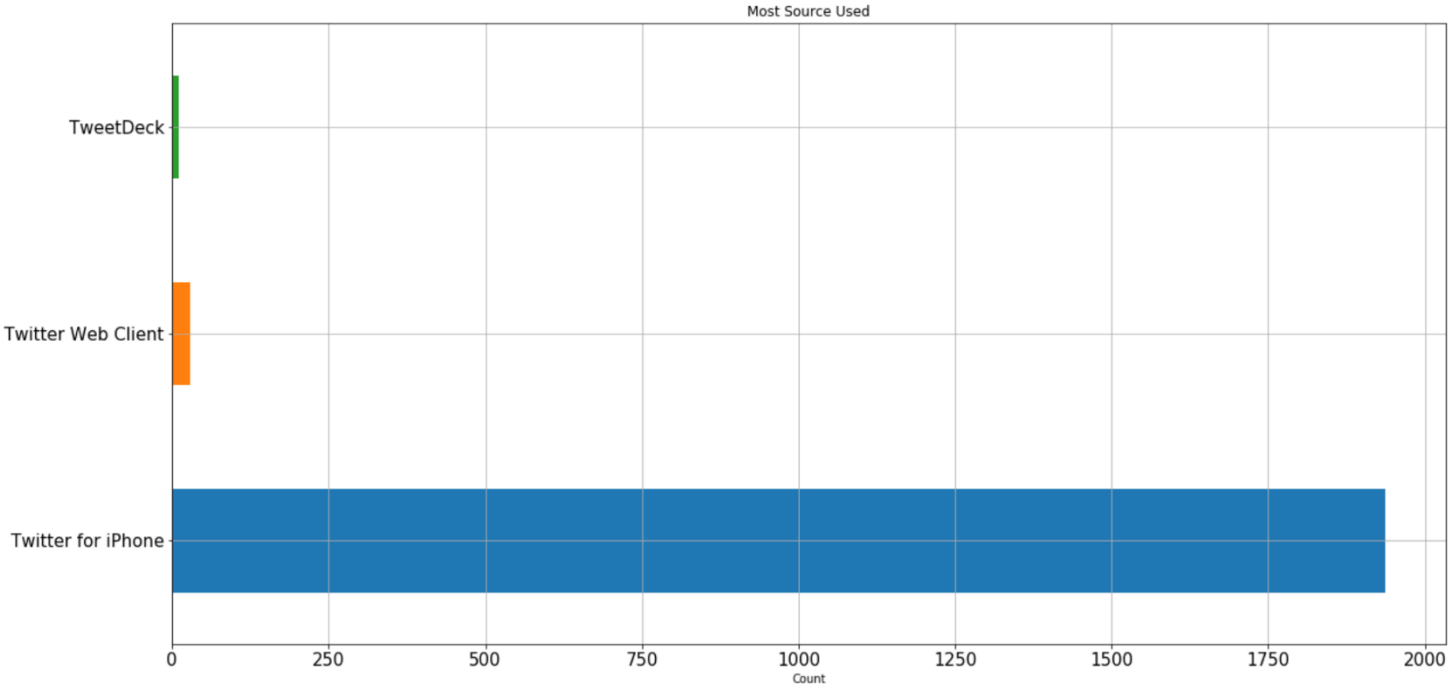

Figure 1 Most Sources are Used

3.2 Number of Tweets Over Time

It is show number tweet monthly from November 2015 to August 2017. On November 2015 there is about 300 tweets, after one month the rate of tweeting increased about 360

tweets. After December we can see the rate of tweeting was decreased over time. (See Figure 2)

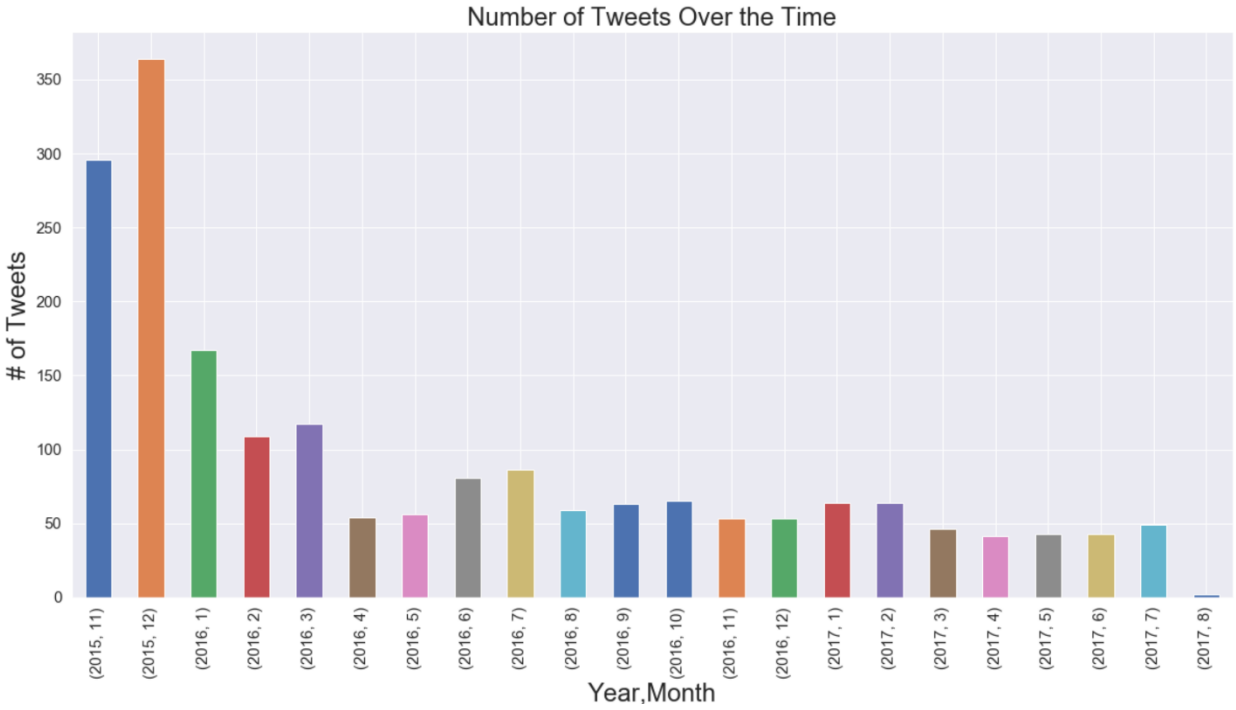

Figure 2: Number of Tweets Over Time

Hint: I used bar chart because line chart doesn't show the date (x-axes).

3.3 Top 10 Popular Dogs

I extract the top 10 popular dog's names by using value_counts() function that give me the top 10 dogs (See Figure 3)

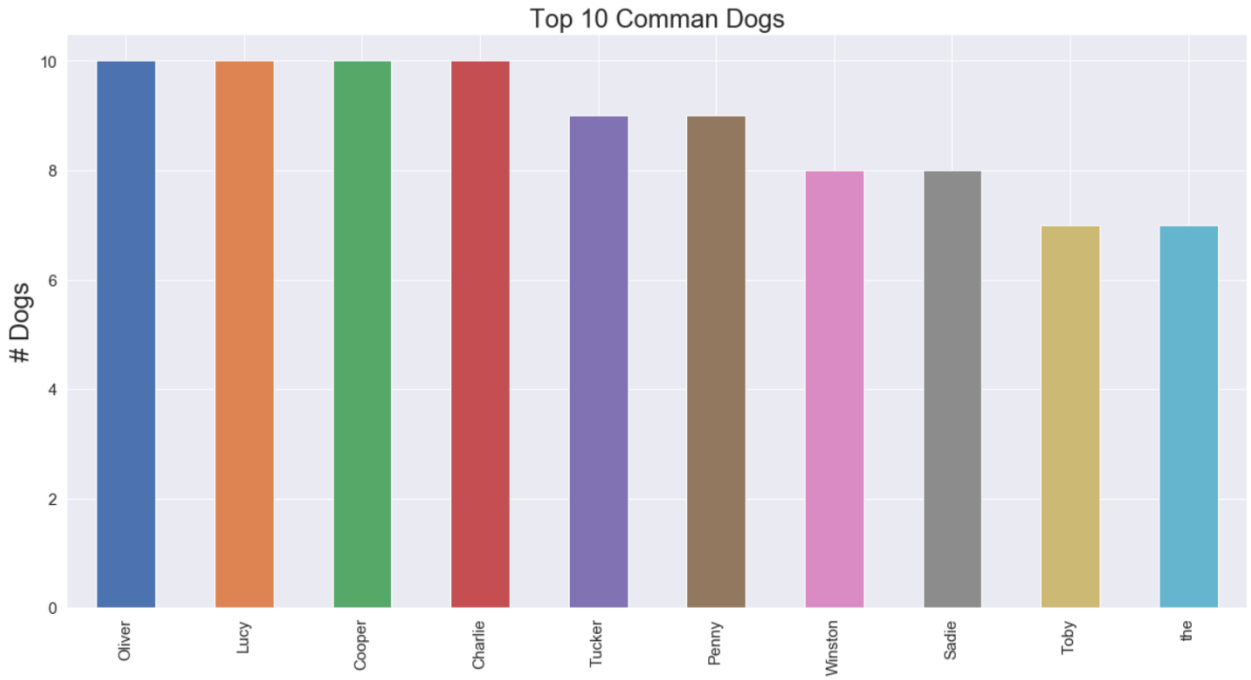

Figure 3: Top 10 Popular Dogs

3.4 Retweet Vs Favorite Count

Retweet vs. favorite count based on top 5 popular dogs, I used new library called seaborn and I used lmplot() function that show the relationship between two variables. It is showing the relationship between retweet and favorite for the top 5 popular dogs. (See Figure 4)

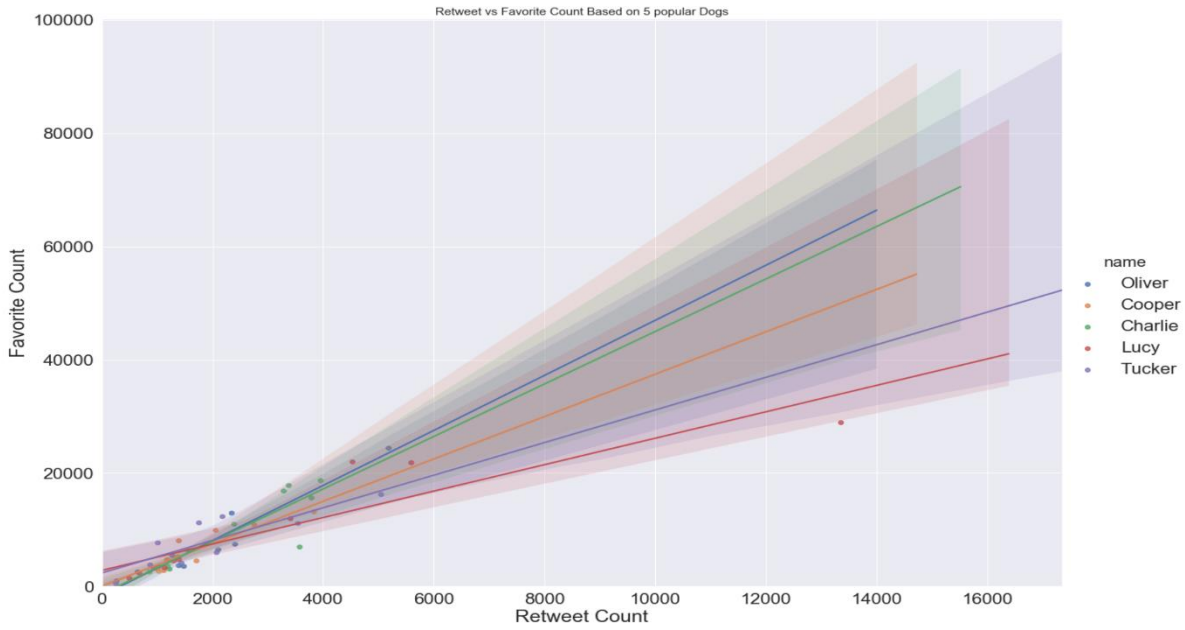

Figure 4: Retweet vs Favorite Count

4. References

1- https://www.domo.com/solution/data-never-sleeps-6
2- Udacity: Data Wrangling section > Wrangle and Analyze Data Project > Project Overview
3- https://en.wikipedia.org/wiki/Data_visualization