DATA ANALYST NANODEGREE

UDACITY

PROJECT 5 : Wrangle & Analyze Data

Fatimah Ahmed Al-Matar

Riyadh, Saudi Arabia

Contents

List of Figures

Abstract

Data are generated from everywhere and in various formats. To understand what that data is all about and to have some clear insights, data should be gathered, cleaned and analyzed. Here comes the role of the Data Analyst; where he/she will gather the needed data from different and multiple resources, clean it if necessary and then apply the required analyses and draw some conclusions.

In this project, I am applying one of Data Analysis approaches known as “Data Wrangling” on data from Twitter about dogs, I will start this process by first gathering the required data, assessing and finding problems in the structure and quality of that data, solve and clean those problems, store that data and finally draw some conclusions and insights about this dog data.

Introduction

“WeRateDogs”, a Twitter account created in November 2015. The purpose of this account is to let people share photos about their dogs and let other people rate them. Usually, the rating is out of 10, but most of the time people rate the dogs with more than 10/10. However, in this project, I will apply the Data Wrangling three steps: Gathering, Assessing and Cleaning on this dog data and try to figure out some interesting outcomes from these lovely dogs.

I have used Jupyter Notebook and Python codes to make all of this happen. Some necessary Python libraries have been used like; Pandas, Numpy, Matplotlib, etc. Most importantly, Tweepy library has been installed as well so we can use the Twitter API to fetch in some more data directly from Twitter.

Analyzing & Visualizing Data

In this part of the project, I have visualized the wrangled data and provided some interesting insights about the dogs in the WeRateDogs Twitter account.

Following are the 4 visualizations that I have come up with:

1- The relationship between the "retweet count" and "favorite count"

After finding the relationship between the two attributes, I have found that there is a linear relationship with a positive correlation between the retweets and the favorite counts.

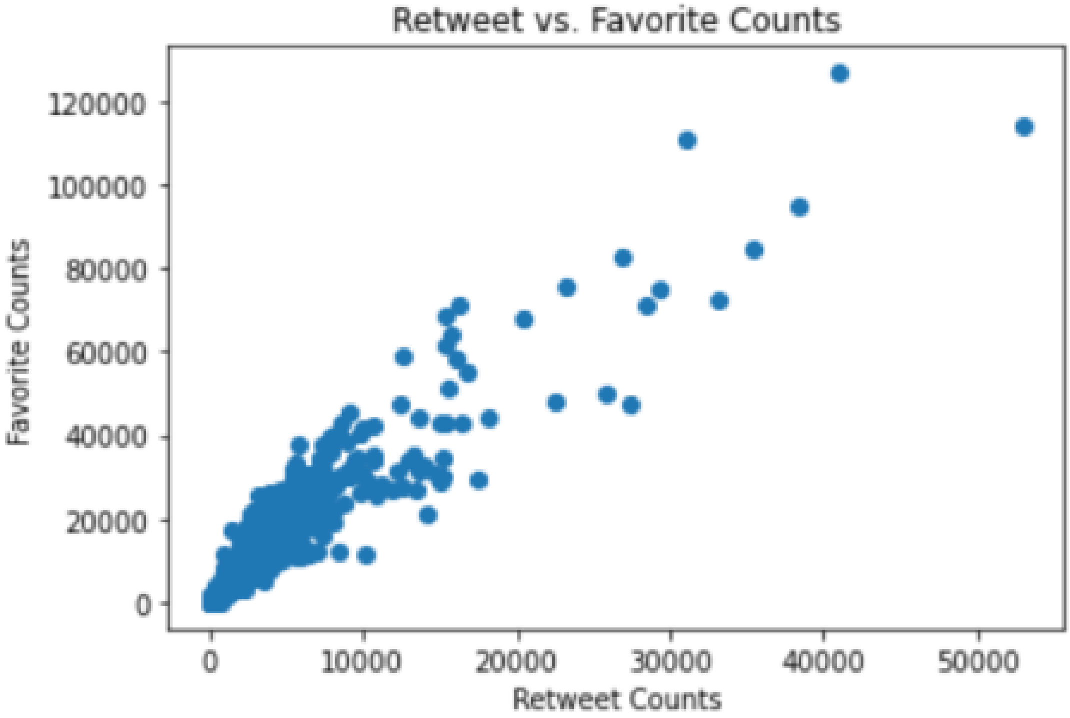

Figure 1: Retweet VS Favorite Counts

1- Most Source used to Tweet

In this visualization, I am trying to find which source was the most one used by people to tweet on the WeRateDog account. I have found that iPhone is the one that is most used to tweet.

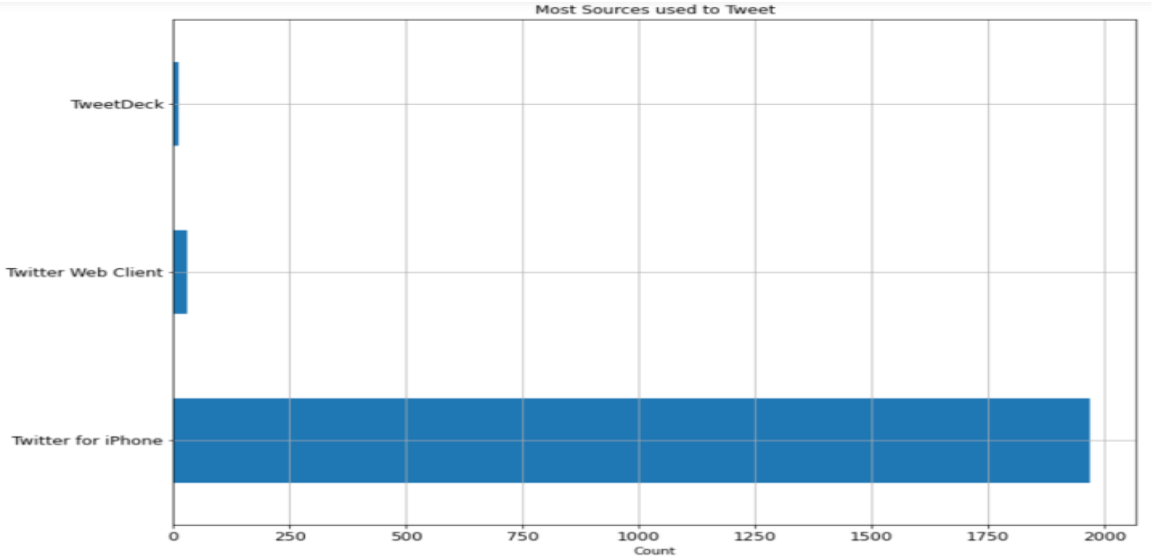

Figure 2: Most Source used to Tweet

2- The Percentage of the Dog Stages

There are 4 dog stages, Pupper, Doggo, Puppo, and Floofer. In the following chart, i t is clear the Pupper has the highest appearance among other stages, following by Doggo, Puppo and finally Floofer.

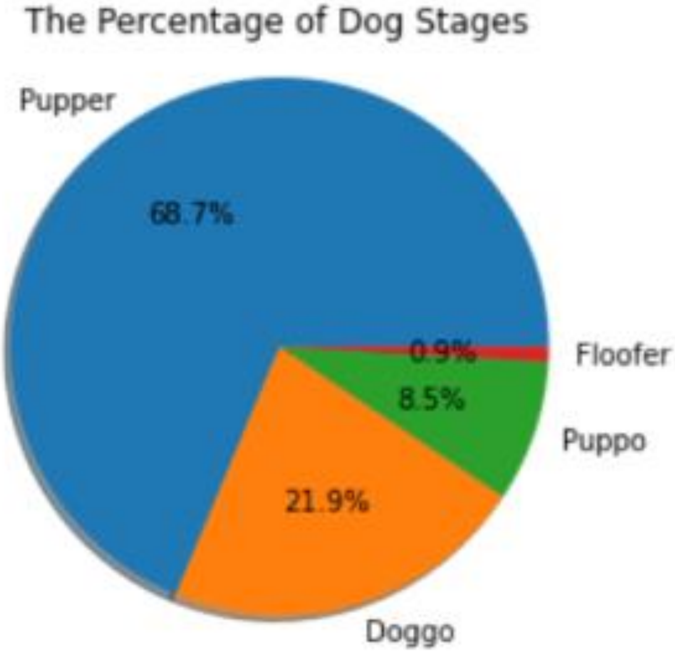

Figure 3: Percentage of Dog Stages

3- What are the Top 10 Popular Dogs on WeRateDogs?

Here, I am finding which the top 10 popular dogs are. Lucy, Cooper and Oliver were the most popular ones!

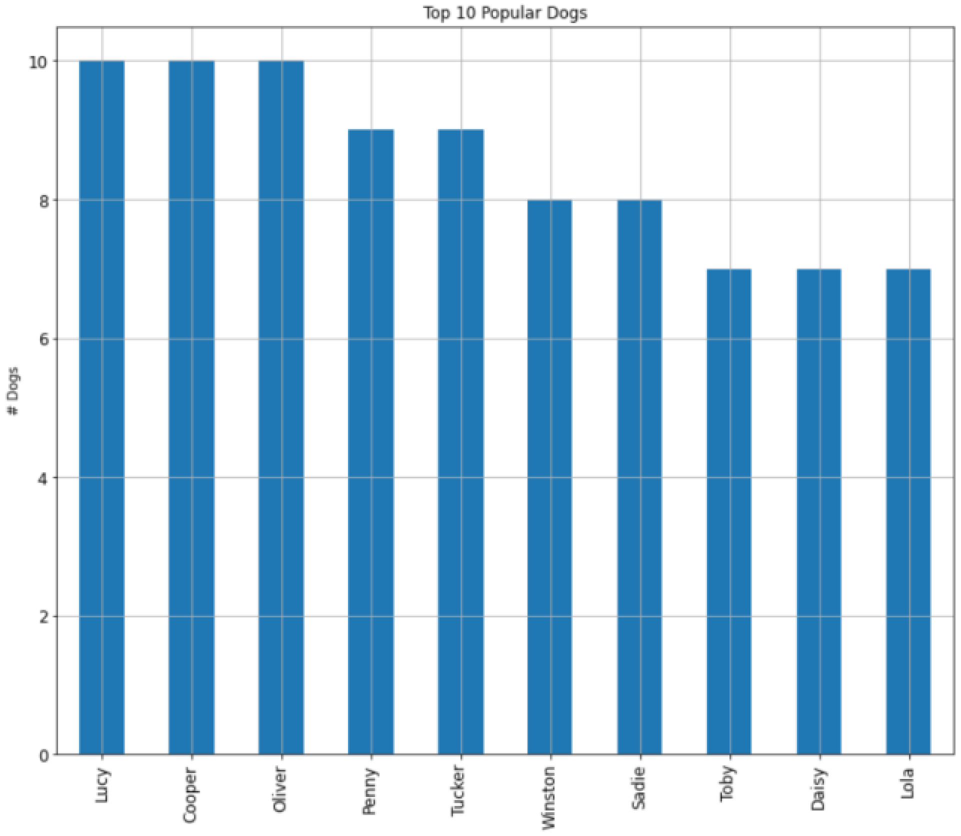

Figure 4: Top 10 Popular Dogs