JANUARY 4, 2020

PROJECT #4: WRANGLE AND ANALYZE DATA

ACT REPORT

MASHAEL ALSAADAN

Introduction

Real-world data rarely comes clean. Using Python and its libraries, I gathered data from a variety of sources and in a variety of formats, assess its quality and tidiness, then cleaned it in this project.

The dataset that I have wrangled, 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.

WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Figure 1: Image via Boston Magazine

This project was done in a jupyter notebook using the following libraries:

- Pandas
- NumPy
- Tweepy
- Json
- Re
- Matplotlib

Wrangling Process

The process consists of the following:

1. **Gathering Data**

   Data was gathered from three different resources:

   a. The WeRateDogs Twitter Archive file that I was given.
   b. The Tweet image predictions file which is hosted on Udacity's server.
   c. Using python's library tweepy and twitter API I will retrieve each tweets' retweet and favorite counts.

2. **Assessing Data**

   The assessment was done both visually and programmatically to find the following:

   - Quality issues: which are content issues, categories of these issues (quality dimensions):
     - Completeness: missing data
     - Validity: data don’t conform to defined schema
     - Accuracy: wrong data
     - Consistency: referring to piece of data in multiple ways

   - Tidiness issues: structural issues, tidy dataset meets these requirements:
     - Each variable is a column
     - Each observation is a row
     - Each type of observational unit is a table

3. **Cleaning Data**

   This involves the following process:

   - Define: defining what to be cleaned and how (usually they are our assessment findings)
   - Code: code to clean the dataset
   - Test: evaluating the code to make sure the issue was solved

Analyzing and Visualizing

In the analysis process, I have focused on analyzing the account's activity over time and what are the most occurred dog breeds in the account and is it the top favorited breeds to people.

1. Analyzing WeRateDogs activity:

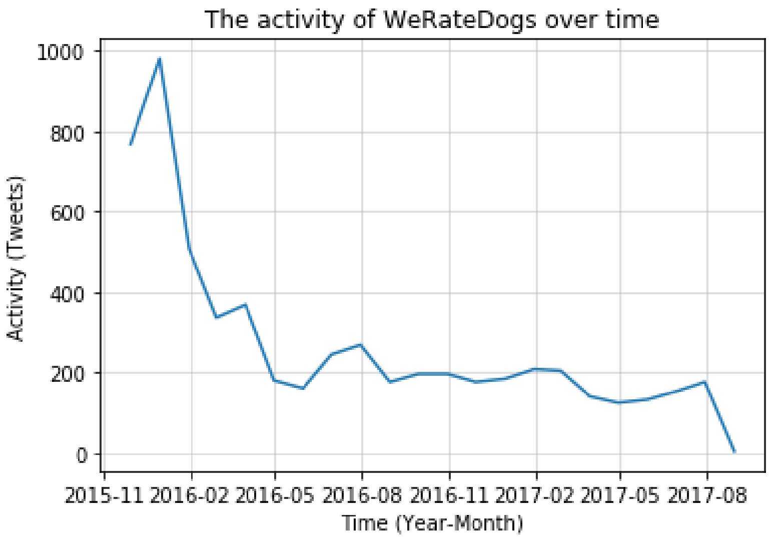
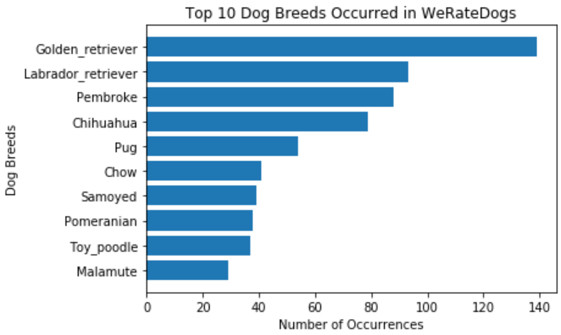


Figure 2: WeRateDogs' activity overtime

The above chart illustrates the activity of WeRateDogs' account over time. We can see that its activity decreases over time. It was at its peak in 2015-12-31 and at its bottom in 2017-8-31 in this dataset.

2. Analyzing top 10 dog breeds' occurrences in the account and favorite counts



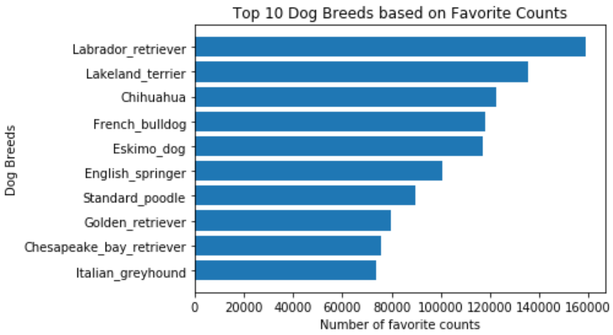
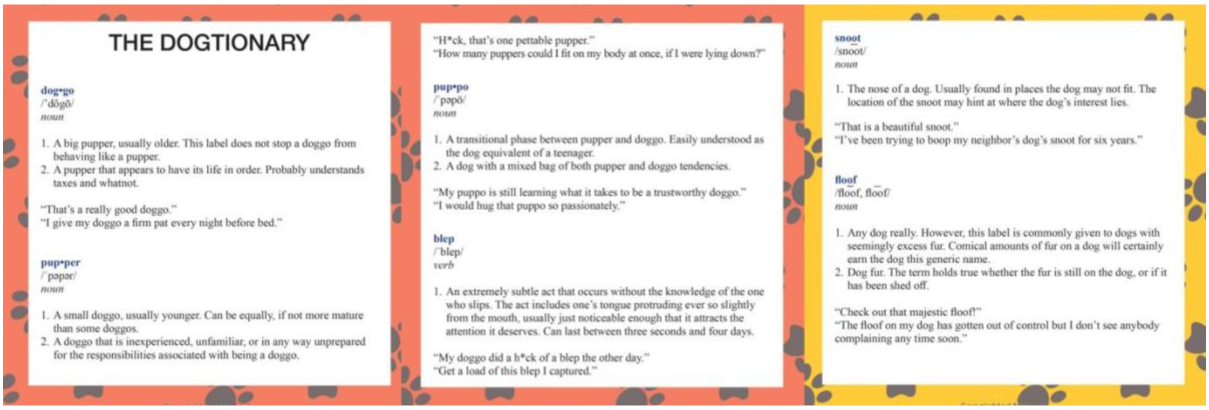


Figure 3: Top 10 dog breeds based on occurrence

Figure 4: Top 10 dog breeds based on favorite counts

Based on the above charts, we can see that there are two breeds that people like and occurrs alot in WeRateDogs account which are: Laborador_retriever and Chihuahua dogs.

3. After some investigation in the dataset, I found that the most occurred dog stage (that was successfully identified) is pupper, which is based on a dogtionary it is a small doggo. I find it quite convincing that it is the most occurred stage as the majority of people tend to like puppies more.



THE DOGTIONARY

dog•go
/ˈdôgō/
noun

1. A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
2. A pupper that appears to have its life in order. Probably understands taxes and whatnot.

"That's a really good doggo."
"I give my doggo a firm pat every night before bed."

pup•per
/ˈpəpər/
noun

1. A small doggo, usually younger. Can be equally, if not more mature than some doggos.
2. A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

"H*ck, that's one pettable pupper."
"How many puppers could I fit on my body at once, if I were lying down?"

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

"My puppo is still learning what it takes to be a trustworthy doggo."
"I would hug that puppo so passionately."

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one's tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

"My doggo did a h*ck of a blep the other day."
"Get a load of this blep I captured."

snoot
/sno͞ot/
noun

1. The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog's interest lies.

"That is a beautiful snoot."
"I've been trying to boop my neighbor's dog's snoot for six years."

floof
/flo͞of, flo͞of/
noun

1. Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
2. Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

"Check out that majestic floof!"
"The floof on my dog has gotten out of control but I don't see anybody complaining any time soon."

Figure 5: Dogtionary

Conclusion

To conclude, this report covered the activities done in data wrangling and some analysis and visualization. The dataset is rich with information and can derive much more insights from it. Finally, data analysis is powerful where you can gain lots of information and insights with a little data.