Project 4

Wrangle and Analyze Data
Act Report

By: Alhanoof Alnajashi

Goal of project:

Wrangle and cleaning data from Twitter account @dog_rates to get the most dog images that has the highest retweet and favorites.

Project is divided into 3 part:

- **Gathering**
- **Assessing**
- **Cleaning**

Gathering Data:

Done in 3 formats TSV, CSV, and from JSON file.

Assessing Data:

Identify some issues in quality and tidiness to clean the data set and try to get some insight about it.

Data Visualization and Insights:

I get three insights in the final clean dataset which are:

1- Number of dogs in each breed.
2- Correlation between retweets and dog's stage.
3- The highest source interaction.

From *Fig1* the most five breeds are: (golden retriever – Labrador retriever – Pembroke – Chihuahua – Pug) and differ from almost 150 until 50 dogs in each breed then it gets more decrease in the other types of the breed.

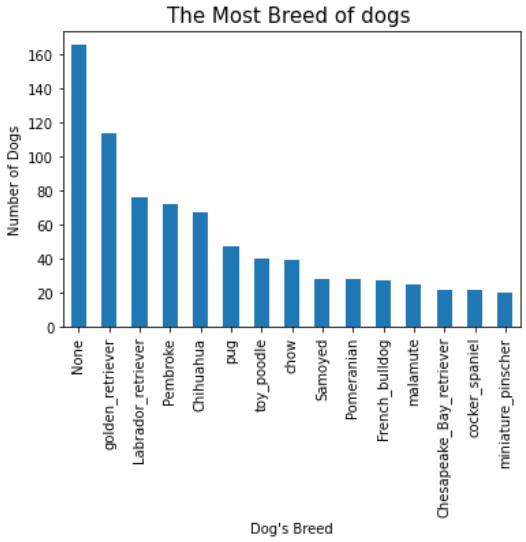

Figure 1

The second insight from *Fig2* is the correlation between number of retweets and dog stage (Doggo - Puppo – Pupper - Floofer) from the scatter we know that the doggo and puppo are the most common stages, the number of dogs in doggo almost 10000 dogs and the puppo almost 50000 to 60000 dogs. There is a huge number of dogs are None and we can't get their stages since some of them the algorithms cannot figure them all.

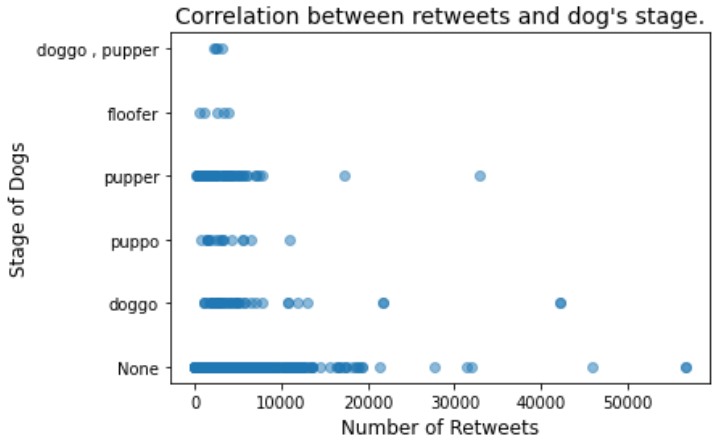

Figure 2

The final insight in *Fig3* is about the interaction and the source of that interaction there are tree sources and the most common one is Twitter from iPhone.

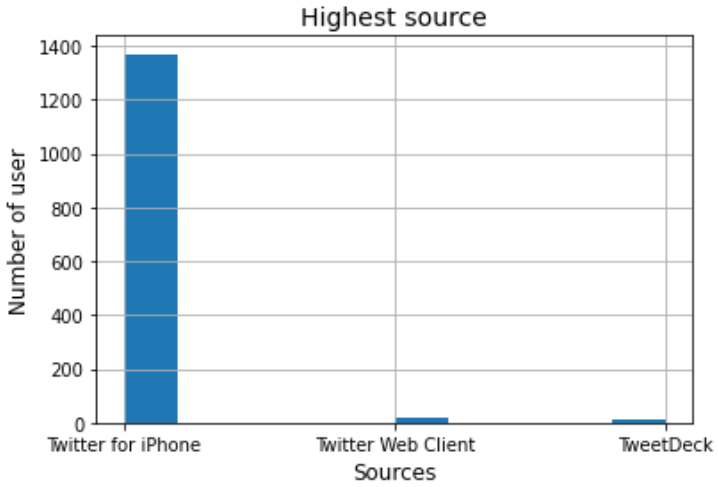
Highest source
1400
1200
1000
800
600
400
200
0
Number of user
Twitter for iPhone
Twitter Web Client
TweetDeck
Sources

Figure 3