2018

UDACITY REPORT

PREPARED BY

Volodymyr Kovalchuk

Lausanne, Switzerland
volodymyr.kovalchuk2@gmail.com

PREPARED FOR

WeRateDogs Udacity

www.udacity.com
WeRateDogs™
@dog_rates

INTRODUCTION

This particular analysis was conducted for WeRateDogs Twitter account as well as Udacity community and based on the information provided by the WeRateDogs and additional information gathered through various channels and data sources. The whole working process included two core stages which are the following:

- **Data wrangling**
- **Data analysis**

DATA WRANGLING

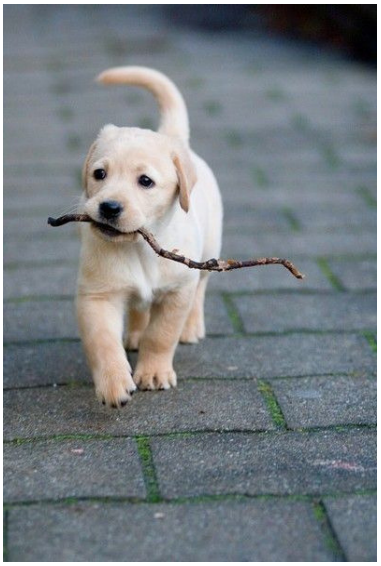
1.1 Gathering data

All of the required data for this particular analysis was gathered through the following sources:

- The WeRateDogs twitter archive that consists of 17 columns and 2356 tweets. This dataset was stored in ***twitter-archive-enhanced.csv*** file. The twitter archive was downloaded manually from the link provided by Udacity instructors.

- The second dataset was gathered by using a neural network developed by Udacity community. The results were stored in ***image_predictions.tsv*** file. This file was hosted on Udacity's servers and downloaded programmatically using the Requests library and the URL provided. The dataset includes top 3 predictions of a dog breed for every tweet. This information is formed into a table of 12 columns and 2075 rows.

- The third source of data was Twitter API from where each tweet's retweet count and favorite ("like") count was gathered using Python's Tweepy library. Every tweet's entire set of JSON data was stored in the file called ***tweet_json.txt***. Then this file was read line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count. The API dataset was successfully stored and consists of 3 columns and 2342 rows. However, while extracting data from API, 14 errors occurred, namely "No status found with that ID". The reason is that those tweets were deleted.



DATA WRANGLING

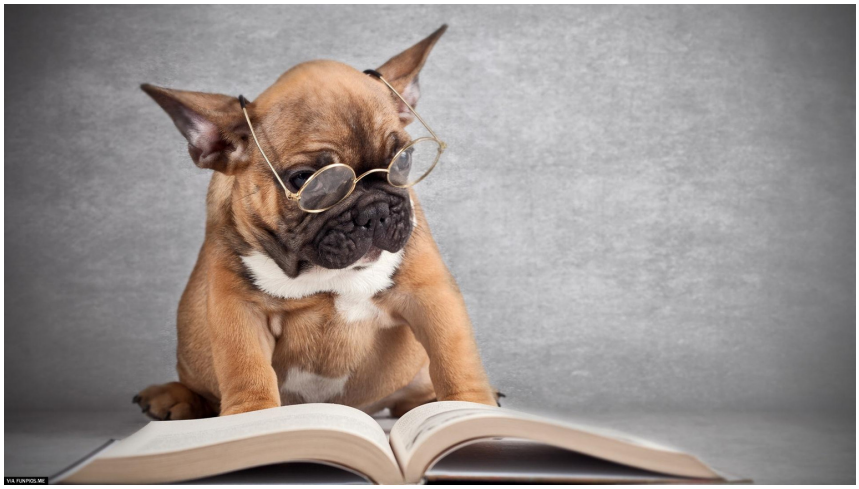
1.2 Assessing data

The next step was to assess all the information that was gathered in the previous step visually and programmatically. At this stage several data issues were detected. These issues were divided into two groups: quality and tidiness issues.

Data quality issues are issues related to content and can be considered at four dimensions: completeness, validity, accuracy, and consistency. On the other hand, tidiness issues are specific structural issues which are not aligned with the characteristics of the tidy data:

- Each variable forms a column
- Each observation forms a row
- Each observational unit forms a table

All of the issues more precisely were described in the Jupyter Notebook.



DATA WRANGLING

1.3 Cleaning data

The last step in the data wrangling process was cleaning our data. At this stage, all of the issues described in the previous section were solved and cleaned. The following structure was used “Define – Code – Test” in order to make this process smooth and avoid possible mistakes. The data was cleaned using both manual and programmatic cleaning as some of the issues were one-off occurrences. The final clean dataset was stored in the file called ***twitter_archive_master.csv***.

All of the work described above was done in the Jupyter Notebook called ***wrangle_act.ipynb***.

DATA ANALYSIS

What would be an ideal tweet for WeRateDogs?

The next step was to analyze our clean data and provide professional visualization of our key findings. The analysis aimed to find answers to the following questions:

- What is the WeRateDogs' account performance over time?
- What are the TOP 10 most popular breeds?
- What are the most popular dogs' names?
- What are the most rated breeds among the most popular?
- Is there a relationship between ratings and number of likes/retweets?
- What is the average number of likes and retweets per dog stage?
- What is the average number of likes and retweets per breed?
- What is the best time to make a tweet?

Therefore, by answering the questions above, we found some interesting insights how would an ideal tweet for WeRateDogs look like.

This part of work was done in the Jupyter Notebook called ***analyze_act.ipynb***. In addition to that, all of the insights are described and explained in the ***act_report.pdf*** file.