**Wrangle report**

This project was part of the data wrangling section of the Udacity Data Analyst Nanodegree program and is primarily focused on wrangling data from the [WeRateDogs](https://twitter.com/dog_rates) Twitter account using Python, documented in a Jupyter Notebook (wrangle\_act.ipynb). This Twitter account rates dogs with humorous commentary. The rating denominator is usually 10, however, the numerators are usually greater than 10. This aspect was not cleaned as it is part of the humor and popularity of WeRateDogs.

For this project, we only wanted original ratings (no retweets) that have images. Not all of the original tweets in the dataset are dog ratings and some are retweets.

Fully assessing and cleaning the entire dataset would require exceptional effort so only a subset of its issues (eight quality issues and two tidiness issues at minimum) needed to be assessed and cleaned.

The tasks for this project were:

* Data wrangling, which consisted of:
  + Gathering data
  + Assessing data
  + Cleaning data
* Storing, analyzing, and visualizing the wrangled data
* Reporting on my data analyses and visualizations (act\_report.pdf)

**The Data**

WeRateDogs provided their Twitter archive (which included tweets through August 1, 2017) of basic tweet data (tweet ID, timestamp, text, etc.) for use with this project. The "enhanced" csv file provided by Udacity (twitter\_archive\_enhanced.csv) also contains columns which were extracted programatically: the rating numerator, rating denominator, dog's name, and dog stages (doggo, floofer, pupper, and puppo). These columns needed to be assessed and cleaned as the extraction process wasn't perfect.

The provided Twitter archive lacked some useful information: retweet count and favorite count. I used the tweet IDs to query the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet\_json.txt. I then read the txt file line by line into a pandas DataFrame only including the desired variables; retweet count and favorite count.

Udacity also provided a link to image\_predictions.tsv which I downloaded programatically using the Requests library.

In this project, we had to conduct data wrangling on twitter data of ‘WeRateDogs’. This is a twitter account which basically rates dog images with humorous content. The rating denominator is usually 10, however, the numerators are usually greater than 10. This aspect was not cleaned as it is part of the humor and popularity of WeRateDogs

The tasks for this project were:

* Data wrangling comprising of following steps:
* Gathering data
* Assessing data
* Cleaning data
* Storing, analyzing, and visualizing the wrangled data
* Creation of report on the analyses and visualizations of the wrangled data
* **Data wrangling**

1. Gather

Gathering data for this project was focused around three files. We were provided with two files, twitter archive file containing data of 2356 tweets and second file, consisted of the image prediction data.

For the third file, we had to query the twitter API for each tweet’s JSON data. This was carried out using Python’s Tweepy library and the output of this was stored in a text file.

Now, these are the three files on which next steps, assessing and cleaning was to be carried out.

1. **Assess**

Assessment on the gathered data had to be done visually and programmatically.

Visual assessment of these three files helped finding the three tidiness issues and a few of the quality issues. Programmatic assessment helped filtering the issues.

Quality

Quality issues were detected on the basis of completeness, validity, accuracy and consistency. Ten issues were detected based on these parameters.

Tidiness

Tidiness revolves around structural issues. Here, we could find 3 issues regarding to the structure.

During programmatic assessment, I checked for data types, null entries, duplicate values, value counts etc. As for structural issues, I have combined the data in three files and stored in to a single file. Also, some columns were condensed into a single column and few columns were removed.

1. **Cleaning**

This is where, all the issues found in step 2 had to be fixed. Process for cleaning is define, code and test. Each issue was fixed with these three steps.

Analysing and visualizing

The master data file was then analysed to find basic results like which is the most favourited ? etc. Visualization helped in summarizing the trends over the years.

Conclusion

Data wrangling helped finding insights into the data which normally would have been misleading, This was a tough yet an interesting exercise. The issues which were solved here are only some of the issues in the dataset. There are many issues that can help refining the data further. I hope to carry those out in the future.

**ACT\_REPORT**

**WeRateDogs data wrangling project**

WeRateDogs is a twitter account that rates people’s dogs with a humorous comment about the dog. It was started in 2015 by college student Matt Nelson, and has since received international media coverage. It has also been endorsed by a lot of famous personalities including author J.K. Rowling.

WeRateDogs asks people to send photos of their dogs, then tweets selected photos rating and a humorous comment. Dogs are rated on a scale of 1 -10 but are invariably given ratings in excess of the maximum, such as "13/10". Popular posts are re-posted on Instagram and Facebook.

As of end of April,2018 the twitter account has nearly 6.72M followers with 6,889 tweets and 133K likes.

I have analyzed this twitter account and spotted some interesting revelations. For this analysis, data considered was the tweet archive provided by Udacity which contains basic tweet data of this account as of 1 Aug,2017.

First and Foremost, I really wanted to know which is the most retweeted and favorited tweet on this account. Also, questions like which is the most common dog name or dog type, what is the average rating, which is the lowest rated dog etc. popped up. Lets check out the inferences below.

I discovered that the most retweeted tweet (77659 times) is actually a 30 second video of a “doggo” trying to float in a pool. This is at the below link.

<https://twitter.com/dog_rates/status/744234799360020481/video/1>

Is the most retweeted tweet also the most favorite one? The answer is No. The most favorited tweet (143871 times) is an image of a puppo. Below is that image:



Link of the tweet : <https://twitter.com/dog_rates/status/822872901745569793/photo/1>

The average rating of a dog is 12.27/10. Talking about the highest rated dog, interestingly, it has been rated 1776/10! Below is the image of the dog.

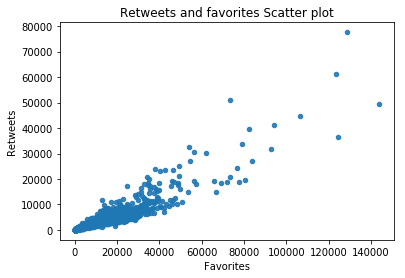


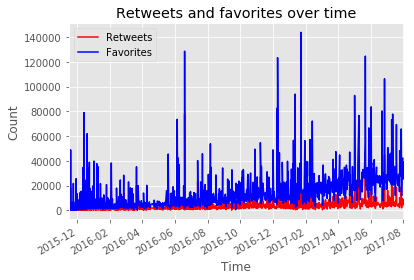
Link of the tweet : <https://twitter.com/dog_rates/status/749981277374128128/photo/1>

The least rated dog is sadly 1/10. There are 4 such dog images that have been rated this way.

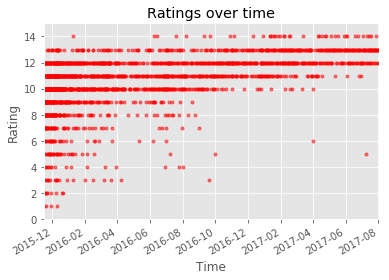
The most common dog type is a “pupper” and people seem to favor the name “Lucy” and “Charlie” more than other dog names.

Going deeper into the analysis, I carried out some visualizations. These produced further insights.

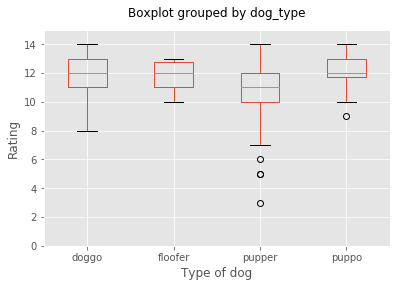




The two plots shows the positive relation between retweet and favorites count and that the retweets and favorites have increased over time. This proves a gradual increase in popularity and following of this twitter acoount.



Now, interestingly, the ratings given to dogs have also increased over time.



As mentioned before, pupper is the most commonly rated dog. However, the ratings shows a contrasting pattern. This is the least rated dog on an average. The highest rated dog is generally ’puppo’.

**Conclusion**

WeRateDogs twitter account has gradually increased in popularity and with time, the dog ratings and also, retweeting and favorite count of tweets have also increased. Pupper is the most commonly rated dog, however, if ratings are to go by, puppo seems to be the more popular choice.

**Act report**