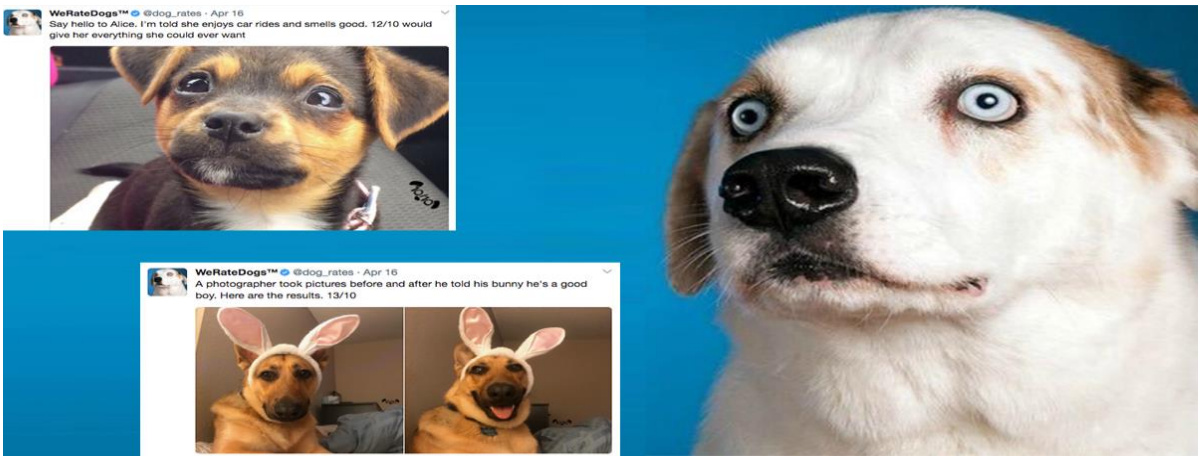
WRANGLE REPORT

#WeRateDogs

INTRODUCTION

This project involves wrangling (and analysing and visualising) 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.



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

After scrapping the data, quality and tidiness issues were assessed and then cleaned. Finally, this cleaned data was used to produce visualizations and make some conclusions about the dataset which could be found in the act_report.pdf file.

Gathering Data

The data set for this project comprises of three different datasets that were gathered as following:

- The enhanced twitter archive file was provided and downloaded manually. It contains basic tweet data for all 5000+ of their tweets.
- The tweet's image predictions file was downloaded programmatically using the Requests library from Udacity's servers. This file contains the information on what breed of dog is present in each tweet's picture according to a neural network.
- The tweet Ids in the twitter archive file were used to query the Twitter API for each tweet's JSON data using Python's Tweepy library and each tweet's entire set of JSON data was stored in a .txt file and then this .txt file was read line by line into a pandas dataframe with 'id', 'created_at', 'favorite_count' and 'retweet_count'.

Assessing Data

After the data gathering process, assessment was performed on the data to identify quality and tidiness issues.

Quality issues pertain to the content of data. Low quality data is also known as dirty data. There are four dimensions of quality: completeness, validity, accuracy and consistency.

The following quality issues were identified:

- There were 2075 rows in the image_predictions dataframe compared to 2356 rows in the twitter_archive dataframe. This is due to tweets with no images and retweets included.
- Several columns such as in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls have empty values.
- The name column has a lot of non-name values. The most common name is 'a' which is not actually a name.
- The numerator and denominator columns have wacky values.
- The timestamp type is an object, not a timestamp.
- The text column could be parsed to include gender.
- The text column could also be parsed to include hashtags.
- In several columns null objects are non-null.

Tidiness issues pertain to the structure of data. These structural problems generally prevent easy analysis. Untidy data is also known as messy data.

The following tidiness issues were identified:

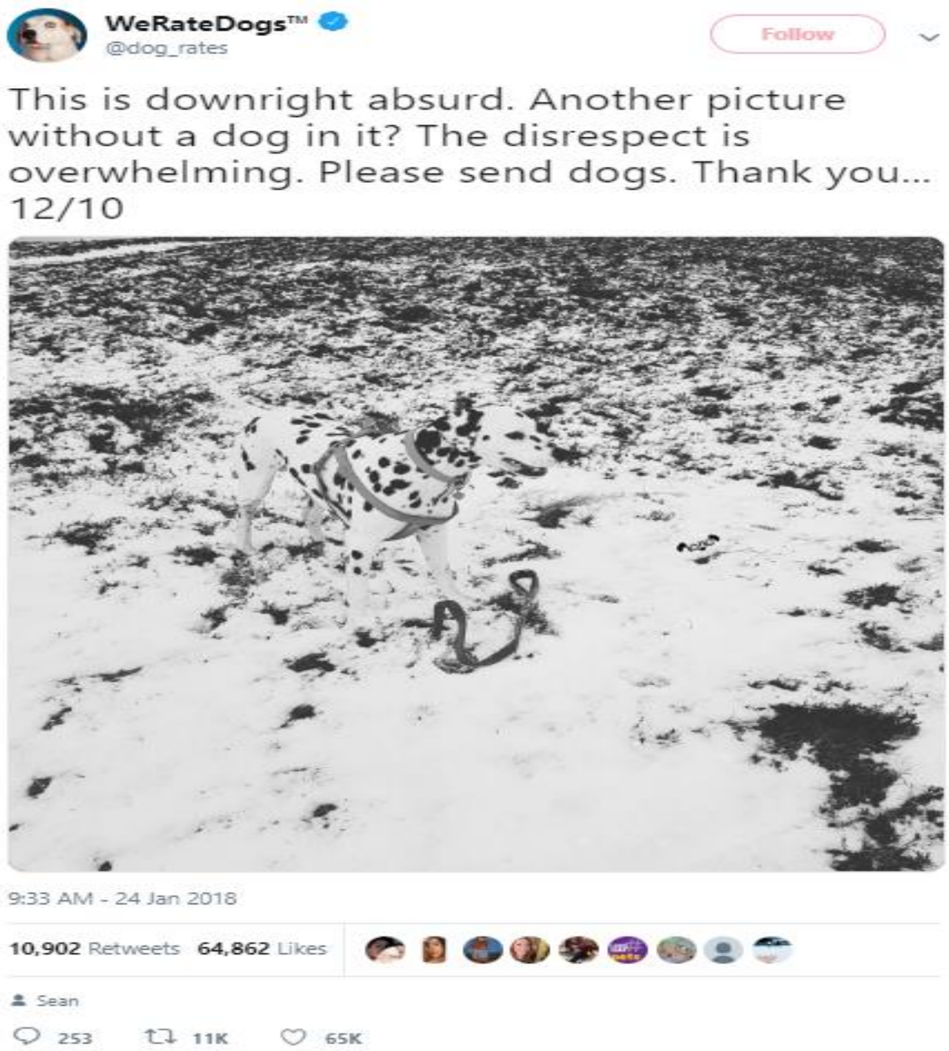
- The columns predicting the dog breed could be condensed.
- The dog 'stages' have values as columns, instead of one column filled with the values.

Cleaning Data

All of the issues while assessing were cleaned. This cleaning was performed in wrangle_act.ipynb and all the steps (Define, Code, Test) of the cleaning process were clearly documented. The resulting was a high quality and tidy master pandas DataFrame i.e, twitter_archive_master.csv

Conclusion

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, assessed its quality and tidiness, then cleaned it for “Wow!”- worthy analyses and visualizations. This is called data wrangling. I documented the wrangling efforts in a Jupyter Notebook, plus showcased them through interesting and trustworthy analyses and visualizations using Python (and its libraries).



WeRateDogs™
@dog_rates

This is downright absurd. Another picture without a dog in it? The disrespect is overwhelming. Please send dogs. Thank you... 12/10

9:33 AM - 24 Jan 2018

10,902 Retweets 64,862 Likes

Sean

253 11K 65K