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

Image via Boston Magazine

Wrangle and Analyze Data 🤠

Report on the wrangling steps taken to clean the data

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) 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. 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 downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Wrangling Endeavours

1. Getting the data from Twitter

The rubric expected us to get the data from the twitter API. I followed the instructions and set up a twitter app, to allow me to get access to the tweets.

I loaded the twitter archive which was provided to me and started to iterate over the *tweet_ids*. It was here that I noticed that some of the *tweet_ids* we actually invalid, I think, this might be because they were deleted at some point after the csv was created.

I then saved all of the twitter data to a csv and moved to step 2.

2. Accessing the data

A direct link to the image predictions csv was provided, I downloaded the CSV using requests library.

I then went on to load all of the other CSVs which were provided to me in their respective data frames.

This was when I decided that it made more sense to combine all of the data frames into a single larger one, and so this is just what I did. I used the merge function in pandas to create this table. I called this the *df_outer* table, and move to the next step.

3. Analyzing the data for issues

I detected several - quality and tidiness issues in the data, they are summarized below.

Quality Issues

1. Dataset has an unnamed column
2. Retweets need to be removed
3. Replies need to be removed
4. Columns associated with retweets and replies need to be dropped
5. tweet_id is an int it needs to be a string
6. timestamp column needs to be converted to DateTime
7. name column has some serious quality issues and should be dropped
8. There are fewer rows in the image predictions table than there are in the twitter archive table

9. doggo, floofer, pupper, and puppo could be made boolean
10. Add proper punctuation to the predictions
11. Column names are not readable

Tidiness Issues

1. source column in the dataset is unreadable
2. doggo, floofer, pupper, and Puppo columns need to be a categorical variable

4. Solving the issues

All the steps that I took to resolve the issues are in wrangle_act.ipynb.

5. Problems faced

It was really difficult to figure out how to convert the doggo, floofer, etc. columns into a single column of a categorical variable, I discovered that I needed to use **isnan** to fix this.

I also had to spend some time figuring out how the twitter api worked.

6. Storing the cleaned data

I contemplated using storing the data in an SQLite database but later decided to just simply store it in a csv. I had to remember to set **index=False** to prevent the unwanted column.

I also tested if the stored data frame could be reloaded, and it worked.

Conclusion

Thus we conclude from all of the data that - people love Golden Retrievers! And all dogs in general. We also understand that consistent posting on twitter will lead to bigger followings overtime