PROJECT: DATA WRANGLING(REPORT)

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, and 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.

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.

Context

My goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

Data

In this project, I worked on the following three datasets.

- Enhanced Twitter Archive
- Additional Data via the Twitter API
- Image Predictions File

Data wrangling consist of three steps: data gathering, data assessing and data cleaning. We now start our data wrangling process with the first step: data gathering.

STEP 1: Data Gathering

To gather the three datasets to be used in this project, the following steps will be followed:

- Download and upload the *twitter_archive_enhanced.csv* and read it into a pandas dataframe.
- Download the *image_prediction.tsv* from the provided this url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv using the requests library in python.
- Query the tweet's retweet count and favorite("like") count using the tweapylibrary and store the data in *tweet_json.txt*.
- Read the *tweet_json.txt* line by line into a pandas dataframe with tweetID, retweet count, and favorite count.

STEP 2 : Data Assessment

Quality issues

Image predictions Table

- the dataset contains duplicated values on the jpg_url column
- Erroneous datatype on the tweet_id (int instead of float)

twitter_archive_enhanced Table

- has null columns
- inconsistency in values of the name column

Twitter archive table

- drop columns not needed for our analysis
- Erroneous datatypes in these columns (tweet_id, rating_denominator,rating_numerator, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- Missing values in 'name' and dog stages represented as 'None'
- Some records have more than on dog stage
- Missing URLs in expanded_urls
- Error in dog names (e.g a,an,actually) are not a dog's name.
- text column includes a text and a short link.

Tidiness issues

Image predictions Table

- Contact column in patients table should be split into phone number and email
- Three variables in two columns in treatments table (treatment, start dose and end dose)

twitter_archive_enhanced

- The four spacies of dog(doggo, floofer, pupper, puppo) should be joined in a single column named species

tweet_json table

- Join twitter_archive_enhanced, Image predictions and tweet_json tables

STEP 3: Data Cleaning

- remove duplicated values on the image prediction dataset
- remove the HTML formating on the source column in twitter_archive_clean table
- change the Datatype of the tweet_id, for the three datasets
- change the timestamp on the twitter_archive_clean to datetime

From the twitter_archive_clean

- Drop retweeted_status_timestamp", "retweeted_status_user_id","retweeted_status_id", "in_reply_to_user_id", "in_reply_to_status_id"
- convert the four columns containing the stages of dog into one single column (doggo, floofer, pupper, puppo)
- Drop the four columns and use the newly created column (stage)
- correct the name errors and replace none with Nan

Clean off the HTML formating on th source column

- change the ID datatype from int to Object

Conclusion

- Some dogs have Zero rating
- Most tweets have just one image
- From the source column, most tweet came from twitter iphone
- Most tweeted dogs were on the pupper stage