WeRateDogs

WRANGLE _ACT REPORT

By Hadiya Usman

Introduction

The dataset I used for this 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 comments about the dog. WeRateDogs has over 4 million followers and has received international media coverage.

To become a great data wrangler I have to learn and master how to integrate information from multiple data sources and identify and resolve data cleaning and quality issues. In this project, I used Python and its libraries to gather data from a

variety of sources and in a variety of formats, assess its quality and tidiness issues then clean it.

I document the wrangling efforts and showcase them through analyses and visualizations using Python and its libraries.

Steps that I follow in this project are:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing, and visualizing data

Step 6: Reporting

I worked on the following three datasets.

Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for 5000+ of WeRateDogs tweets, but not everything. One column the archive does contain each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and q) to make this Twitter archive enhanced. Of the 5000+ tweets, I have filtered for tweets with ratings only.

Data via the Twitter API

The Twitter archives: retweet count and favorite count are two notable columns omissions. The additional data was gathered by querying Twitter's API to gather the valuable data.

DATA GATHERING

The WeRateDogs Twitter Archive

Downloaded the `twitter_archive_enhanced.csv`. file manually. After it is downloaded, i upload it and read the data into a pandas DataFrame.

The Tweet Image Predictions

The (`image_predictions.tsv`) is hosted on Udacity's servers i downloaded it programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Additional Data from the Twitter API

I gather **each tweet's retweet count** and **favorite ("like") count**. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file.

Each tweet's JSON data was written to its own line. Then read the .txt file line by line into a pandas DataFrame with **tweet ID, retweet count, and favorite count**.

ASSESSING DATA

In this step of Data Wrangling, I visually and programmatically assessed the data gathered for quality and tidiness issues. I identify and document 8 quality and 3 tidiness issues.

CLEANING DATA

In this step of data wrangling which is also the last step, I cleaned all the quality and tidiness issues I earlier identify for analysis and visualization. Before taking steps to clean the issues, I made a copy of the data using the .copy() pandas method, should in case I might need the original data later on.

I followed the **define** → **code** → **test** framework and clearly document it then merge the different data pieces into one single dataframe (table), which is according to the rules of tidy data.