Report on Data Wrangling – We Rate Dogs

By-Shefali Luley

Wrangling Report

The dataset that I 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.

Data Wrangling

Data Wrangling consists of :

1. Gathering data : Gathering data is the first step in data wrangling. Before gathering, we have no data, and after it, we do.Gathering data varies from project to project. Sometimes you're just given data, or pointed to it like I've done for you throughout this course. Sometimes you need to search for the right data for your project. Sometimes the data you need isn't readily available, and you need to generate it yourself somehow. When you do find your data, it's not unusual for it to be spread across several different sources and file formats, which makes things tricky when organizing the data in your programming environment.

2. Assessing Data : Assessing your data is the second step in data wrangling. When assessing, you're like a detective at work, inspecting your dataset for two things: data quality issues (*i.e. content issues*) and lack of tidiness (*i.e. structural issues*).Assessing is the precursor to cleaning. You can't clean something that you don't know exists!

3. Cleaning data : **Cleaning** your data is the third step in data wrangling. It is where you fix the quality and tidiness issues that you identified in the assess step.The cleaning process is further divided into 3 steps :

   Define- Defining a data cleaning plan in writing
   Code-Translating these definitions to code and executing them
   Test- Testing our dataset, often using code to make sure changes are done

4. Storing, Analyzing and Visualization the wrangled data

5. Report on the data wrangling efforts and visulization

Gathering Data

My data gathering efforts consists of gathering from all the below sources:

1)Twitter Archive file: Download this file manually or from the resources tab of Udacity's server: twitter_archive_enhanced.csv

2)Tweet Image Prediction:The tweet image predictions,that is what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is available on resource's tab of Udacity's servers and should be downloaded programmatically using the Requests library.

3)Twitter API & JSON: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. 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 should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.m

Assessing Data

After gathering each of the above dataset, I assessed them visually and programmatically for tidiness and quality issue.You will be able to observe 8 quality issues and 2 tidiness issue in the Jupyter notebook respectively.

Following are the Quality issues:

1)Twitter Archive table:

-tweet_id is integer not string,

-timestamp is not in datetime format,

-name column should be in string,

-Remove individual columns:doggo, floofer, pupper, puppo

-Standardize dog ratings

2)Image prediction table:

-img_num column should be in string

3)Tweet json table:

-some tweets contain more than 2 rating

-Delete retweets

-Correct naming issues

-retweet_count should be integer

-favorite_count should be integer

-remove columns with missing values and Remove columns which are not required :retweeted_status_user_id,,retweeted_status_timestamp,in_reply_to_user_id,in_reply_to_status_id

Follwing are the Tidiness Issues:

-Merge the clean versions of twitter_archive dataframe, image_prediction datframe, and tweet_json dataframes

-Create one column for the various dog types: doggo, floofer, pupper, puppo

Cleaning Data

After assessing the data, it’s cleaned in the following manner: Define,code and Test

Here’s few example,

Tidiness Issue

Merge the twitter_archive dataframe, image_prediction dataframe, and tweet_json dataframes

1)Define : Using python command "CONCAT" we can merge the datasets

Code

Test : To check whether all dataset have merged

Quality Issue

Removing column's with missing values and which are not required.

1)Define: Deleting the column's using DROP function, for column Axis = 1

Code

Test: To test, we display the dataset using .HEAD() function

Quality Issue

Create one column for the various dog types : doggo, floofer, pupper, puppo

2)Define: We can extract the data using .EXTRACT()

Code

Test : To test, we display the dataset using .HEAD() function

Storing ,Analyzing and Visualizing the Wrangled Data

Storing the clean Dataframe in a csv file with the main ne named twitter_archive_master.csv. If any additional files exist due to multiple tables required for tidiness, those files are named properly.In addition to this, you may store the cleaned data in SQLite database(which is to be submitted as well if you do)Analyzed and visualized the wrangled data in wrangle_act.ipynb jupyter notebook . As per the rubrics it should have 3 insights and 1 visualization must be produced .