Wrangle Report

Introduction

The main goal of this project is to put all skills I learned about wrangling and analyzing data into practice. I will use Python Jupyter Notebook and its libraries to gather data from different types of sources, assess its quality and tidiness, clean it, and derive insights from it. This project will include some interesting data analysis and visualizations.

Project Details

- Gathering Data
- Assessing Data from both Quality and Tidiness sides
- Cleaning Data
- Storing, Analyzing, and Visualization

Gathering Data

In this project, data were collected from three sources.

- Twitter_Archive_enhanced.csv: I manually download from Udacity and loaded it into workbook.
- Image_Predictions.tsv: what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. I programmatically download from udacity server using the url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Twitter API & JSON: For gathering this dataset, we have to use 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. Then we transfer txt file into dataframe.

Assessing Data

In this stage, I mainly use df.info() to check the data types and see whether if there are some missing values. Value_counts is also being frequently used to check the extreme value. I find some quality issues:

- Twitter_Archive: there are 181 rows have retweets which should be removed. Some columns have large amount of missing value, we should consider remove these columns if they are not needed in later data analysis. Also, we find that time stamp should be timedate instead of object. What's more, the rating(rating_denominator / rating_numerator) should be around 10, but in this dataset, there are some extreme values like 1776. Also, there are some decimal numerators, we have to correct them.
- Image_predictions: there are 66 duplicates in column jpg_url. And it is strange that some predictions of dog type are box-turtle and ice-bear. We should create new columns prediction and confidence level to replace p1, p2, p3, p1_conf, p2_conf and p3_conf to make it more readable.
- Tweet_json: the number of rows should be consistent with other two dataset.

There are also two tideness issues:

- Columns 'doggo', 'floofer', 'pupper' and 'puppo' should be included in one column dog_type instead of four columns.
- All three dataset should be from same dataset and have same number of rows, thus we have to merge three tables into one.

Cleaning data

In this stage, I followed udacity instruction to use Define, Code, and Test to correct and revise each data quality and tidiness issue.

extreme value of denominator and numerator are standardized manually or programmatically. Columns that are not needed are dropped. 181 retweets were dropped because they are duplication data.

also, I dropped 66 duplicates in jpg_url and create new columns for dog type prediction and confidence level.

the last but not least, I melted 4 dog types into 1 and merged three dataframe into a comprehensive one.

Storing, analyzing and visualizing

In this stage, we first stored the comprehensive dataset as twitter_archive_master.csv. then we started to derive insights from data. My first visualization showing the distribution of dog types and find that the golden retriever, labrador and Pembroke are the most common rated types of dog. Then I compared the rating and retweets. Overall, number of retweets goes up when rating goes up, however, the highest rating does not have the highest retweets counts. In the next visualization, I find that favorites have similar trend with retweets when compared to ratings. Last but not least, retweets and favorites have positive trend, I think number of retweets can be used as a predictor of favorites in linear regression model.