Wrangle Report

Introduction

In this report, I will document the wrangling efforts in the section of wrangling WeRateDogs

Data Wrangling Process:
- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data

Gathering Data for this process took place in 3 stages

1. The WeRateDogs Twitter archive. I am giving this file to you, so imagine it as a file on hand. Download this file manually by clicking the following link: `twitter_archive_enhanced.csv`
2. The tweet image predictions, i.e., 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 hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. 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.

Gathering Summary

Gathering was the first step in the data wrangling process. We could finish the high-level gathering process: - Obtaining data - Getting data from an existing file (twitter-archive-enhanced.csv) Reading from csv file using pandas - Downloading a file from the internet (image-predictions.tsv) Downloading file using requests - Querying an API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy - Importing that data into our programming environment (Jupyter Notebook).

Accessing

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues was our next step. We could detect and document the following quality issues and tidiness issues.

Quality

Completeness, Validity, Accuracy, Consistency.

- Content issues archive dataset - in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be intergs instead of float.
- retweeted_status_timestamp, timestamp should be datetime instead of object (string).
- In several columns null objects are non-null (None to NaN).
- Missing values from images dataset (2075 rows instead of 2356).
- A separate column for gender of dogs

Tidiness

- No need to all the informations in images dataset, (tweet_id and jpg_url what matters).
- Various stages of dogs in columns instead of rows archives dataset.
- All tables should be part of one data set.

Cleaning

Cleaning our data is the third step in data wrangling. It is where we fixed the quality and tidiness issues that we identified in the assess step. We used the two types of cleaning, the manual and programmatic even the manual not recommended but the issues were one-off occurrences. Our process was Define, Code and Test and we were always making a copy of that dataset even we made the copy in file to test the change before applying to the main dataset. We didn't spot all the quality and tidiness assessments at the assessing data section, so we have been iterating and revisiting assessing to add these assessments to our notes.

Conclusion

Data wrangling indeed is a core skill that everyone who works with data should be familiar with since so much of the world's data isn't clean. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. We couldn't be able to make some of the visualization without wrangling.