Wrangle Report

Introduction

The goal of this project to utilize the data wrangling concept that I have learned so far in the Udacity Data Analysis program in addition to creating reliable analysis and visualizations. The dataset that I was working on is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs which is a Twitter account that rates people's dogs. These ratings almost always have a denominator of 10.

This report briefly describes my wrangling efforts.

Project details

The tasks of this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

The dataset for this project consists of three different parts that were obtained as following:

- **Twitter archive file:** the *twitter_archive_enhanced.csv* that contains specific data about the tweets, and it was downloaded manually.

- **Image predictions file**: this basically classifies what breed of is present in each tweet according to a neural network. This file *image_predictions.tsv* was downloaded programmatically using the Requests library and URL information.

- **Twitter API & JSON file:** using the tweet IDs provide in the twitter archive file, I queried the Twitter API and obtained JSON data using Python's Tweepy library. This data was stored in *tweet_json.txt* file.

Assessing data

Once the whole dataset has been obtained I loaded them into separate dataframes and began to asses them as follows:

- I printed the dataframes in the Jupyter Notebook and looked at the records in order to find an anomaly or issues.
- I looked at the csv file provided to give me a better judgement on the data.
- I also used tools in python to help me asses the dataset programmatically using functions such as info, sample, head , tail ,and value_counts.

This assessment helped me to find **ten** quality issues and **two** tidiness issues. All

these issues are separated and documented in details in the *wrangle_act..ipynb*.

Cleaning data

This part of the project was the most challenging for me as it required creative approaches to solve some issues discovered in the previous step.

At the beginning I took a copy of the original dataframes in case an error happens during cleaning.

I then started with basic cleaning such as: removing retweets, removing rows that had missing values, removing duplicates and removing unimportant columns.

Some of the cleaning steps were challenging and required a lot of research. One of them was combining multiple columns into one column in the twitter archive data.

After finishing the cleaning process , I merged the three dataframes together and stored the resultant in *twitter_archive_master.csv.*

Conclusion

Data wrangling is a concept and skill that should be available in every data scientist or analyst as it is the only way to make raw data (such as the twitter archive) meaningful and useful. It also saves time for data analysts and let them focus on analysis and reporting resulting in a better and more accurate decision making based on the data given.