Introduction

By Fabi Daniel O.

Wrangle & Analyze Data

The project was part of the requirement of the data wrangling section of the Udacity Data Analyst Nanodegree program. The section was focused on wrangling data and the project is also focused on reinforcing all the techniques and tools needed to accomplish the gathering, assessing and cleaning of data from the WeRateDogs Twitter account using Python and libraries

Project Summary

Using Python and its Libraries, I was able to gather data from 3 different sources, assess and identify various quality and tidiness issues and ultimately take control of the wild data through data cleaning. Other tasks include storing the cleaned data, completing the analysis, presenting at least a visual representation of the insights and also writing a report.

Gathering the Data

As earlier stated, data was gathered from three different sources and all these were done programmatically to allow for reproducibility of the process:

Files on hand

- Twitter Archive File: Using python library pandas, I was able to read and store this file into a dataframe named *twitter_archive_df*

Web Scraping

- Image Predictions File: Using python requests, I was able to send an http request to Udacity Servers and download the file (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image- predictions.tsv) and used another library to find and read the contents of the returned object into a dataframe *image_predictions_df*

API

- Twitter API & Tweet JSON File: this was the most tasking in the whole gathering phase. I started with applying and getting my customer and access details from twitter. I used the details to query twitter api for the tweets status using their ids. The response was a get_status object, which I had to *filter for the _json part and later was used in populating the dataframe. tweet_api_df*

Assessing the Data

This process was done both virtually and programmatically. It involves checking the datasets for quality (content) and tidiness issues (Structural) and this formed the basis of our classification. Using these methods, I was able to check the datasets for completeness, validity, accuracy, and consistency. With virtual assessment, I was able to identify inconsistences in values of a certain column, also discover that some values are being represented as variables. Programmatic Assessment was much more effective in revealing most of the data defects addressed in the project.

Cleaning the Data

As advised, the first thing I did was to make a copy of all the datasets, this is important because of the trial-and-error nature of data cleaning and also for data recovery.

The issues were assessed & documented per datasets and they were treated as per the same format. For each dataset, quality issues came first, missing & surplus values to avoid repetition and in total eight issues were treated.

In the first dataframe, some of the issues include, datatype conversion to datetime, standardization of a column with different cases etc. In the image prediction dataframe, I was able rename the columns and standardize some columns. Third column cleaning duties include deleting retweets and conversion from string to datetime/

As per the requirements, only two tidiness issues should be attended to and in my case, I was able to concat four columns into one and standardize a column across table to facilitate joins.

Analyzing the Data

At this stage, the dataset was merged into a master file and stored programmatically. It is safe to say that the dataset was good enough to carry out a mini analysis. I was able to identify the following

- I was able to identify the most prominent dog breeds
- I was able to identify the most prominent dog stage
- I was able to identify the most prominent source(app)

Conclusion

I was able to complete this project and go through the three stages of data wrangling. I'm now very comfortable with integrating information from multiple data sources, checking for structural and content issues, treating these issues, all programmatically. Together with some python libraries, I was able to meet the requirements for the project. All the activities that occurred in each stage have been duly communicated and summarized in the above paragraphs