Data Wrangling Report

Introduction

Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. The goal of data wrangling is to assure quality and useful data. Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.
The process of data wrangling may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses. Data wrangling typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.

Project Details

In this project we have used following Data Files for data wrangling -

Data Download from URL provided by Udacity:

- twitter-archive-enhanced.csv
- image-prediction.tsv

Data Downloaded using Tweepy, Twitter API (python):

- tweet_json.txt (JSON format)

We will analyze the data in 3 files mentioned, clean up the data, remove unnecessary columns and then join the three files to perform meaningful data analysis to provide answers to following question(s),

- What are the top 10 most mentioned dog breeds?
- What are the top 10 most favourite dog breeds?
- Top dog breed in terms of mentions & favourite count in Social Media

Quality issues with Data

Document the data quality issues with the 3 files mentioned in the project details section and we are going to also document tidiness and perform cleanup steps as part of this project.

- Twitter Archive: Timestamp column is of object datatype instead of datetime or timestamp
- Twitter Archive: We are only interested in Original Tweets which are only 2097 out of 2356. 181 tweets are retweets and 78 are replies
- Twitter Archive: Rating Denominator should be 10 but out of 2097 original tweets 17 tweets have a denominator that is not 10
- Twitter Archive: Rating Numerator should be greater than 10 but there are 855 original tweets where it is not
- Twitter Archive: There are 4 columns for displaying the dog stage - doggo, floofer, pupper, and puppo. These can be part of only one column as this is categorical data
- Twitter Archive: Some of the dogs have invalid names like, "a", "None" etc.
- Image Prediction: Total number of tweets in the dataframe is 2075 which is 281 tweets less than twitter archive
- Json Data: 23 of all the tweets provided in Archive are deleted (Tweepy Exception: 22 File Not Found Error and for one tweet I am not authorized to download)

Tidiness

- Twitter Archive: Following 4 columns can be made as single columns (as categorical) - doggo, floofer, pupper, and puppo
- Twitter Archive: Rating Denominator can be dropped when value remain static as 10
- Twitter Archive: We only need original tweets, and following columns can be dropped as they don't provide any value as such - in_reply_to_status_id, in_reply_to_user_id, source, text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls
- Twitter Archive: timestamp column datatype change and we can add Year and Month as separate column which can be used for various analysis
- Image Prediction: Get the final dog_breed using p1_dog, p2_dog and p3_dog. Choose the breed based on first True value of these 3 columns
- Json Data: This can be combined with Twitter Archive with following columns - retweet_count, favorite_count. Key used to combine is id (tweet_id)

Cleanup

The quality and tidiness issues were cleaned using programmatic techniques such as:

• Dropping unnecessary columns from the tables
• Removing rows that consisted of retweets
• Removal of rows with duplicate information
• Deleted rows that did not have any dog predictions at all
• Combining all three data frames into a single data frame

Storing and Acting on Wrangled Data

- Save the master dataset, which is combined data of the three files, to a CSV file, twitter_archive_master.csv
- The master dataset is analyzed using Pandas in the Jupyter Notebook and at least three separate insights are produced
- 3 labelled visualization is produced in the Jupyter Notebook using Matplotlib and Seaborn plotting libraries