Udacity: Wrangle and Analyze Data

Author: Yogesh Chaudhari

Date: 1/16/2020

Data Wrangling

This document is one of the artifacts for the Data Wrangling exercise. This document addresses one of the requirements for the project. This document describes the various steps during data wrangling viz. a) Data Gathering, b) Data Assessment, c) Data Cleaning and d) Storage and Analysis of the data. These activities are programmed in jupyter notebook provided in Udacity workspace.

Listed below are the various steps taken during data wrangling process:

a) **Data Gathering:** This project required collection of three disparate data sources as listed below:

1) twitter_archive_enhanced.csv – This data is provided by the course instructor which consists of data points such as tweet_id, tweet text, tweet source, timestamp of tweet, rating information etc.

2) tweet metadata – The project required gathering additional data directly from twitter using twitter API. Necessary steps were taken to fetch required metadata for each tweet using tweet id. The twitter API response for each tweet was stored in the text file tweet_json.txt on separate line. This data was later was parsed using python's json library. A dataframe was created using these responses which includes tweet data such as retweet count, favorite count, tweet source.

3) image_prediction.tsv – This is a tab delimited file provided by course instructor which consists of "dog breed" attribute predicted from the images uploaded to twitter. The file was downloaded from the link provided by the instructor.

These 3 datasets were consumed in pandas dataframe and subsequently used for assessment, cleaning and finally used for deriving insights.

b) **Data Assessment:** The 3 datasets above were assessed visually and programmatically to identify cleanliness (quality) and tidiness issues. Listed below are the issues identified for each dataset:

Quality Issues

Twitter Archive

1) Timestamp column is in string format
2) Source field can be cleaned to have distinct name of source e.g. iPhone, Web Client etc.
3) Subset of records have Numerator on higher end (1776)
4) Subset of records have Denominator not equal to 10
5) Remove columns that are not needed for analysis
6) Convert tweet_id to string

Tweet Metadata

7) Convert tweet_id to string

Image Prediction

8) Convert tweet_id to string

Tidiness Issues

Twitter Archive

1) The columns doggo, floofer, pupper and puppo can be collapsed into one column to just represent type of dogs
2) Replace None with null (NaN)

Tweet Metadata

3) Source column is duplicated, as its available in Twitter Archive as well, let's drop it

Image Prediction

4) Fix column names such as p1, p2, p3 to have legitimate names
5) Select only required fields
6) Use only the rows where dog = True

c) **Data Cleaning:** With various techniques learned during the coursework supplemented by proper google search, various steps were taken to address quality issues and tidiness issues listed above. Detailed steps/code is available in accompanied jupyter notebook.

d) **Storage and Analysis:** After cleanup, all 3 datasets were joined using key tweet_id to form a final dataset for analysis. This file was saved as csv in program workspace as twitter_archive_master.csv. This data was further used for visualization and identifying insights.