Data Wrangling Report

Intoduction

The purpose of this project is to practice data wrangling data .The dataset wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The data wrangling process include three main steps:

- Gathering.
- Assessing.
- Cleaning.

This report briefly describes my wrangling efforts.

Gathering

Data was gathered from three different sources:

1. From WeRateDogs Twitter archive given by Udacity in csv format:
   Using panda's method 'read_csv', I managed to read the data stored in the file 'twitter-archive-enhanced.csv'. I stored it in a DataFrame called 'twitter_archive_df '. The data has many issues that will be cleaned and resolved later.
2. Image prediction file downloaded programmatically using Requests library and the URL provided by Udacity in tsv format :
   Using Requests library and 'get' method, data was downloaded in a file 'image_predictions.tsv'. Then, the content was stored in a DataFrame called 'image_predictions_df ' using pandas' method 'read_csv()'.
3. Data retrieved by querying Twitter's APIs and using Tweepy library:
   I didn't receive a confirmation from twitter developer so I used the file 'Tweet-Json' which provided by udacity .I uploaded the file to the notebook then read it line by line .Then I loaded it into a DataFrame called 'Twitter_API_df' and saved the DataFrame in csv file called 'tweet_json.csv' padnas method 'to_csv()'.

Assessing

After gathering the data and storing them in DataFrames, the following step was assessing the data for quality and tidiness. Data were assessed programmatically and visually.

Quality

It includes issues such as completeness, validity, Accuracy and consistency in data

twitter_archive_clean

- Remove tweets that are actually retweets bacause they are not original.
- Drop columns with retweet information.

- Remove tweets that are actually replys bacause they are not original.
- Drop columns with replys information.
- Change tweet_id from number to string.
- Change timestamp from string to date time and make separate columns for date and time.
- Repalce invalid name values like ['not','old','one','the','a','o','an','all','by','this','getting','BO','AI','actually','incredibly',just','my','unacceptable','very','officially','not','quite','space','such','infuriating'] with None values
- Fill missing values in expanded_urls with None values
- Change retweeted_status_timestamp from string to date time

image_predictions_clean

- Change tweet_id from number to string.
- P1 , P2 and P3 have inconsisitent capital words
- Drop duplicate jpg_url.
- P1 , p2 and P3 have unnessary underscore instead of space.

Twitter_API_clean

- Rename id to tweet_id so can merge later.
- Change tweet_id from number to string

Tidiness

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types. In tidy data:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

issues

- combining dog stages to one column.
- Newly created Date and time column needed to change from object(string) to date time format.
- perform inner join between three data frame as they all have data for same tweet.

Cleaning

It is the process of fixing and resolving issues identified in the Cleaning process. The (define, code, and test) steps were used in the cleaning process. First, copies of the DataFrames were created before cleaning. Then, the steps of cleaning were applied iteratively on all issues.

Storing

The final DataFrame called df_all_cleaned ' contains 1990 rows and 23 columns with the correct data types. The dataset is then stored in a csv file called 'df_all_cleaned.csv' . At this point, the data was successfully wrangled and therefore ready for analysis and visualization.

Analysis & Visualization

These steps are not part of data wrangling process. However, it cannot reflect correct and accurate insights without performing data wrangling first. Visualizations and insights are provided in ‘act_report.pdf