Data Wrangling: We Rate Dogs

Summary:

- **Introduction**
- **Data Wrangling**
  - Gather
  - Access
  - Clean
- **Storing**
- **Conclusion**

Introduction

In this Project, I used Twitter data posted by We Rate Dogs, a famous account on twitter which rates people's dogs. I Scraped some details for each tweet from twitter. Udacity provides me a neural network which predicts Dogs rate. I scraped data using API against this neural network and will compare how true predictions were at the end of this section. However, the core purpose of this project to master Data Wrangling Skills (gathering, Accessing, Cleaning). I Performed these detailed Steps in order to draw some useful insights from this data.

Data Wrangling

Here are steps I performed to Wrangle data.

Gather Data:

- Initially, I used a Dataset provided by Udacity where I have a List of tweets emailed by WeRateDogs personally to Udacty. I lodead data set using pandas function **read_csv.**
- Call Twitter API by URL and scrap tweet's details individually.
- Call API to get neural network data processed by Udacity.

Access Data:

Data Inclusion :

In project description we are asked to go with following Criteria:

- Do not include retweets

- Only tweets that have images
- I've also skipped Replies for simplicity..

Quality

Tweets Archive table

- Retweets were found in the DataSet.
- Replies were also found in the DataSet.
- Erroneous data types exist in the DataSet. i-e tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper and puppo
- Missing info in the expanded_urls column, only 2218 rows has data.
- Nulls represented as a string of "None" for columns: name, doggo, floofer, pupper, puppo
- Missing values for columns : doggo, floofer, pupper, puppo
- Some names are not proper names.
- Source column contains urls with Anchor tags (HTML <a href='') tags.
- rating_numerator and rating_denominator columns contains some misleading values.

Predictions table

- jpg_url is duplicated 66 times.
- 281 records are missing in the prediction table.
- incorrect data type for the column tweet_id

Tweet's Details table

- incorrect data type for the column tweet_id

Tidiness

Tweets Archive Table

- some columns are with same data like: doggo, floofer, pupper and puppo all contain dog types

Predictions Table

- some columns are with same data like: doggo, floofer, pupper and puppo all contain dog types

Tweet's Details

- This table contains totally different data from the Archive and Predictions Table.

Clean Data:

- First, I created copies from original data and stored them into separate variables for further processing.
- Removed these columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- Converted tweet_id into Object/String using astype function in df_archive_clean dataframe.
- Converted timestamp column to datetime object using pd.to_datetime(df)
- Removed Source column as I'll not be using it for any kind of processing.
- Counting missing values for doggo, floofer, pupper and puppo in archive table
- remove those rows where rating_denominator is misleading. (more than 10)
- Identify and remove rating_numerator rows which have misleading values. (more than 20)
- I removed Tidiness Issues, like multiple rows had the same data and data type issues in all three tables.

Save Data:

- I saved cleaned data into csv and sql database files for further use.

Conclusion:

Initially I got list of tweet's having some tweet's details with tweet id. I scrapped images, retweet_count, favourite_count etc from Twitter using tweepy library. I performed the detailed Data Wrangling steps, ASSESS data and found Quality and Tidy Issues. I cleaned these issue to produce better results. Last, I took a sample number of images from the original data set and compared what neural networks predict for these images to test how efficiently neural networks did this job.