Wrangle and Analyze Data Report

Project 2

Project Details

My tasks in this project are as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

1- The WeRateDogs Twitter archive, manually downloaded twitter-archive-enhanced.csv and read the file in jupyter notebook

2- The tweet image predictions, giving the URL https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv I used **requests.get(url)** to download it programmatically and save it in my path folder

3- Twitter API for each tweet's JSON data, by creating tweeter's developer account to generate the Consumer API keys, and the Access Token and Access Token Secret that you will need for authentication, then using **tweet_id** in **twitter_archive_enhanced** file to query the Twitter API for each tweet's JSON data and store each tweet's entire set of JSON data in a file called **tweet_json.txt** file as required using **api.get_status(tweet_id, tweet_mode='extended')** and finally read **tweet_json.txt** file line by line into a pandas DataFrame

4- (I used Stackoverflow to help me writing this part of code https://stackoverflow.com/questions/47612822/how-to-create-pandas-dataframe-from-twitter-search-api , I added this part from Stackoverflow '**parser = tweepy.parsers.JSONParser()** ' in '**api = tweepy.API(auth)**' as I had this issue '**object of type status is not JSON serializable**'

Assessing data

1- Visually, by reading the gathered three tables in jupyter notebook and in Excel files
2- Programmatically, using .info(), .head(), .duplicated(), .sample(), .value_counts()

After assessing the data I could separate between quality issues and tidiness issues which are

Quality

1- twitter_archive_enhanced table
  - twitter_archive_enhanced contains retweets
  - timestamp column has date and time in the same column
  - wrong datatype for rating_numerator and rating_denominator
  - rating_denominator column has values not equal to 10
  - wrong dog names starting with lowercase characters like a, an

2- image_predictions table
  - jpg_url column has duplicated data
  - Replace '_' with ' ' in p1, p2 and p3 names
  - dogs types names in p1, p2 and p3 are in lowercase sometimes

3- tweet_json table
  - tweet_json, tweet_id column is str data

Tidiness

1- doggo, floofer, pupper and puppo columns should be merged into one column named "dog_stage"
2- the three dataframes should be merged as they are part of the same observational unit

Cleaning data

1- Defining, after assessing data now we defined the issues to clean twitter_archive_enhanced table

- Remove all rows that have values (not blank or non-null) in retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
- Separate between time and date
- Convert datatype for rating_numerator and rating_denominator into float
- Correct rating_denominator for values not equal to 10
- Replace dogs name starting with lowercase characters like a, an with NaN
- Drop retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp
- Merge doggo, floofer, pupper and puppo columns into one column named "dog_stage"

image_predictions table

- Drop duplicate data in jpg_url
- Replace '_' with ' ' in p1, p2 and p3 names
- Convert first character of each word of dogs types names to uppercase and remaining to lowercase

tweet_json table

- Convert tweet_id, in_reply_to_status_id, in_reply_to_user_id columns to object type

2- Coding, before cleaning we keep the original data and make a cope of to clean it, using .copy(), .astype(int), .fillna(0), .drop(), pd.melt(), .drop_duplicates()

3- Testing, test every piece of cleaned data to make sure from cleaning using, sum(), .info(), .head(), .value_counts(),