Wrangle Report

My workflow in this project were as follows:

- **Data wrangling, which consists of:**
  - **Gathering data**
  - **Assessing data**
  - **Cleaning data**
- **Storing, analyzing, and visualizing wrangled data (act_reprot)**

Data Gathering

I start gathering date from three different sources :
1- Enhanced Twitter Archive
File in csv format , i read it into pandas dataframe using read_csv function.
2- Image Predictions File
Url provide by udacity i use it to download the file programmatically using requests library and then save into tsv file and i read into pandas dataframe using read_csv function
3- Additional Data via the Twitter API.
Here i applied for twitter developer account and started to access twitter api through my personal keyes and tokens i installed tweepy library and then i used tweet_id from Enhanced Twitter Archive file to iterate the api and stored the data on jason file and then i red it into pandas dataframe using pandas read_json and i calculated len of the fails between two data
Output:
Now i have three data frames from three different sources

1 - archive_enhanced_df

2- image_prediction_df

3- twitter_api_df

Data Assessment:

I have done this assessment through :

Visual assessment:

I used Enhanced Twitter Archive csv file to view it in google sheet as i am very familiar with this tool.

I transferred image_prediction_df and twitter_api_df into a csv file in order to view it in google sheet as well.

programmatic assessment:

I used jupyter notebook and pandas methods & functions like DataFrame
.info
.sample
.head
.shape
.describe
.tail
.dtype
.value_counts
To assess the data

Output:

1 - archive_enhanced_df
I found tidiness and quality issues

Tidness issues :

- Four columns of dog stage [doggos,floofer, pupper and puppo]should be one column (dog_satge)
- No column for rating
- text column in archive_enhanced table should be split into tweet , rating and url_link.

quality issues :

Erroneous data types (consistency)

as per twitter data dictionary The integer representation of the unique identifier for this Tweet. This number is greater than 53 bits and some programming languages may have difficulty/silent defects in interpreting it.

- Erroneous data types of id's columns as int
- Erroneous data types of rating_numerator as int
- Erroneous data types of rating_denominator as int
- Erroneous data types of timestamp , retweeted_status_timestamp column as object.

completeness & validation

- a , an , None 'all' ,'not','by' ,'O' in column name
- Some tweets are actually retweets and replies not original tweets that have to be deleted as per the data wrangling
- missing values of images difference of tweets between image and archive data
- none value in name , doggos,floofer, pupper and puppo columns
- Four columns of dog stage [doggos,floofer, pupper and puppo]should be one column (dog_satge) and then sorted and removing duplicates

2- image_prediction_df

3- twitter_api_df

quality and tidiness issues

Tidness

- Archive , image prediction and twitter api tables must be in one dataframe
- tweet_id duplicated in three tables and source column is duplicated in enhanced table and twitter_api table
- Column headers in the image are values, not variable names.

Quality

Erroneous data types as per twitter data dictionary The integer representation of the unique identifier for this Tweet. This number is greater than 53 bits and some programming languages may have difficulty/silent defects in interpreting it.

- Erroneous data types of id & id_str columns in api dataframe as int
- Erroneous data types of tweet_id in image prediction dataframe as int
- Erroneous data types of favorite_count and retweet_count completeness

- 66 jpj url duplicated in image table
- missing values
- coordinates, place,contributors and geo are null values
  consistency
- Replies tweets included with tweets in the data frame

Data Cleaning:

- The structure of the process i used : Define, code, test
- Created copy for each table and start to clean :
- Datattypes were the most common issue need to be cleaned for the purpose of analysing and visualising
- Handling errors in timestamp and converting it into data and time datatype and extract year column for the purpose of visualising
- Handel name errors in the name column by using replace method
- Duplicated values : remove it by sorting and using remove duplicates method and keeping last value
- Handling more than one value into multiple columns by using melt function and fixing issues of duplication and extra column resulted from the process
- Validation like retweets and replies i used boolean indexing and query function to query data that have null values in retweet and reply id to remove retweets and replies
- Creating column for rating by dividing rating num over rating dominant
- Merging archive with image on image table to remove data with no imaged
- Creating master DataFrame for the three data tables since they all related to each other.

Output

Master data frame cleaned

Archive data frame cleaned copy

Image_predictin data frame cleaned cop

Twitter api data frame cleaned copy

Image_predicton csv file

Twitter_api csv file

Images of visualizing (showing in the act_report)