Wrangle and Analyze Data Report

Manar albogami (Data Analyst Nanodegree)

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs

What is Data Wrangling?
Data Wrangling is a process where first data is gathered from different sources,then the quality of the data is assessed and finally the data is cleaned to create a dataset on which exploratory data analysis could be performed.

1. **Data Gathering**
   Gathering :
   - import CSV for WeRateDogs Twitter archive using the provided twitter-archive-enhanced.csv file
   - Programmatically download the image-predictions.tsv file through the Requests library
   - Download data JSON file called tweet_json.txt
2. **Assessing Data**

   Assessing is the second step in the data wrangling process:

   Programmatic assessment: using code to view specific portions and summaries of the data (pandas' `head`, `value_counts()`, `sample()` and `info` methods .

   Twitter_archive dataset

   First:
   - **Quality:**
   - Null values recorded as None and NaN(missing values)
   - tweet_id type is int64 ,i will chanage tweet_id data type to string
   - convert timestamp to be datetime and rename the column into tweet_date
   - source mixed html tag
   - We have some columns that contain unnecessary data: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp ,... delet unneed **column**

   Second:
   - **Tidiness:**
   - The 4 different columns doggo, floofer, pupper and puppo, combine in one columns represent stages_of_dogs"

First :

Image_predictions dataset

- **Quality:**
- tweet_idf should be string type
- The types of dogs in columns p1, p2, and p3 had some uppercase \ lowercase letters.

Second:

- **Tidiness:**
- extract breed of dog from columns p, p_conf and p_dog

data_tweet dataset

First:

- **Quality:**
- Missing values in some cloumns
- id column should named 'tweet_id' as the others data have ,and dtype should be string , sourse change data type to category
- We have some columns that contain unnecessary data (created_at,full_text, Truncated,display_text_range,entities,extended_entities,in_reply_to_status_id,in_reply_to_status_id_str,in_reply_to_user_id,in_reply_to_user_id_str, In_reply_to_screen_name,user,geo,coordinates,favorited,retweeted,possibly_sensitive,possibly_sensitive_appealable,lang,retweeted_status,quoted_status_id, quoted_status_id_str,quoted_status,place,contributors,'is_quote_status
- source mixed html tag,Rewrite the tweet source, from iphone,..etc

Second:

- **Tidiness:**
- Merge twitter_archive_copy ,data_tweet_copy and image_predictions_copy to merge_df dataframe

3. Cleaning Data

What is Cleaning Data?

Cleaning is the third step in the data wrangling process The programmatic data cleaning process:

Define: convert our assessments into defined cleaning tasks. These definitions also serve as an instruction list so others (or yourself in the future) can look at your work and reproduce it.

Code: convert those definitions to code and run that code.

Test: test your dataset, visually or with code, to make sure your cleaning operations Worked.

It's most important part of this project to clean the data and store it in new and proper form to make it useable , Here I programmatically clean the data to change some types, drop some cloumns that is not useful, correct some errors in data then merge data frame

4.Store Data
I stored all the cleaned into csv file : twitter_archive_master_.csv

5. Analyzing and Visualizing Data

What is the most used source?
I use countplot to view the most source of the tweet

- What is the most stage of doges?
  I use pie chart to view the most stage of doges
- What is the 6 frequent bread?
  I use bar chart to view 6 frequent dog breed
- compare the favorite counts & retweet counts
  I use scatter to compare the favorite & etweet counts
- Number of Tweets per month?
  I use plot line to view number of tweets per month