Wrangle Report - YX

Introduction

Wrangle and Analyze Data project gathers data from various sources associated with a tweeter account, WeRateDogs. After colleting all the information, quality and tidiness issues were determined and then cleaned the data. As a last step, the finalized data were analyzed, four visualizations were created, and significant insights were obtained from the image.

Gathering Data

Data was gathered from 3 different sources:

1. The twitter_archive_enhanced.csv was provided and can be downloaded manually. This data contains information regarding the tweet id, timestamp, text, rating, name, etc. [df_twi_enhan]
2. The tweet image info is downloaded programmatically by using the Udacity's request library servers. The dog's breed info was obtained based on the images. [df_breed]
3. Favorite counts and retweet counts info can be assessed through Twitter API. [df_tweets]

Assessing Data

After gathering the data, data duplication, data inconsistency, missing values were checked. Quality and tidiness issues were determined:

Quality:

1. df_twi_enhan contains retweets, where retweeted_status_id has a number instead of NaN.
2. "Timestamp is a string" is a wrong datatype.
3. Dogs'name start with lowercase letter which are invaid infomation, ex. "a", "actually", "all", etc.
4. doggo, floofer, pupper, and puppo are using the string "None" instead of NaN
5. tweet_id, in_reply_to_status_id, in_reply_to_user_id should be changed to string.
6. Rating scale is not accurate.
7. df_breed dataset contains data not related to dogs.
8. The p1, p2, and p3 contents are not consistent, some are capitalized, some contain underscores.

Tidiness:

9. Some info are not useful, such as source column in df_twi_enhan table and img_num in df_breed table.
10. There are three separate data sources instead of one giant table, since they all talking about the same tweet.
11. doggo, floofer, pupper, and puppo are all talking about dogs characteristics, can combine them into one column, and drop them afterall.

Cleaning Data

All the issues found in Assessing Data section were cleaned by using expression such as:

drop(), rename(), replace(), extract(), astype(), value_counts(), head(), info(), for..in.., query(), capitalize(), merge(), etc.

Analyzing and Visualizing Data

There are four visualizations (bar chart, pie chart, etc.) were created, and significant insights were obtained from the image in this section. Organizing method includes groupby(), sort_values(), etc.

Conclusion

In real world, majority data will come from different sources, and are all pieces of information. This project helps to learn how to use different tools to gather data, define issues and clean the data before any data analysis can be performed.