Report: act_report

1. Introduction

In this project, I wrangled, analyzed and visualized data from the archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

To be more specificly, I gathered, assessed, cleaned the provided data, stored the data in a csv file, and analyzed the data with some visualization.

2. Issues found in the assessing step

Quality issues

Twitter Archive table (df)

1. Missing data in columns including "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp".
2. Retweets are included in the dataset.
3. Source columns have HTML tags.
4. Timestamp and retweeted_status_timestamp is an object.
5. Dogs name have 'None', or 'a', or 'an.' and some only has lower case as names.
6. Multiple dog stages occurs such as 'doggo puppo', 'doggo pupper', 'doggo floofer'.
7. The ratings for dogs are nor standardized.

Tweet image predictions (image_df):

8. Dog breeds are not consistently in p1,p2,p3 columns.

Tweet-json dataframe (tweet_json):

9. Retweet and favorite information is not available for all tweets and cannot be retrieved.

Tidiness issues

Twitter Archive table (df)

1. There are multiple columns containing the same type of data, e.g. doggo, floofer, pupper and puppo all contain dog types.

Tweet image predictions (image_df):

2. This dataset is part of the same observational unit as the data in the previous dataframe(df).

Tweet-json dataframe (tweet_json):

3. This dataset is part of the same observational unit as the data in the previous dataframe (df).

3. Insights and visualization

Insights:

1. From the provided data, Top 10 frequent breeds for dogs are:

golden_retriever, labrador_retriever, pembroke, chihuahua, pug,chow, Samoyed, Pomeranian, toy_poodle, and malamute.

2. From the provided data, Top 5 names for dogs are:

Oliver, Charlie, Cooper, Tucker, and Lucy.

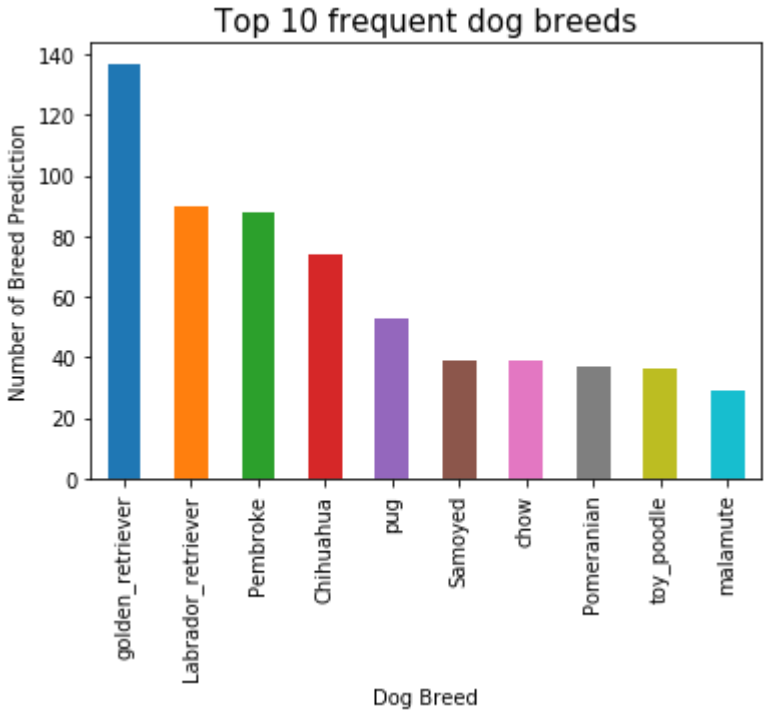
3. From the provided data, the most frequent stage for dogs is pupper.

More details about how I came up with the insight could be found in wrangle_act.ipynb.

visualization

In [4]:
```
import pandas as pd
import matplotlib.pyplot as plt
# read main_df.csv
new_df = pd.read_csv('twitter_archive_master.csv')
```

In [5]:
```
# Top 10 frequent frequent breeds for dogs
new_df['breed_prediction'].value_counts()[0:10].sort_values(ascending=False)
plt.ylabel('Number of Breed Prediction')
plt.title('Top 10 frequent dog breeds', size=15)
plt.xlabel('Dog Breed')
plt.plot();
```


Top 10 frequent dog breeds
Number of Breed Prediction
0
20
40
60
80
100
120
140
golden_retriever
Labrador_retriever
Pembroke
Chihuahua
pug
Samoyed
chow
Pomeranian
toy_poodle
malamute
Dog Breed