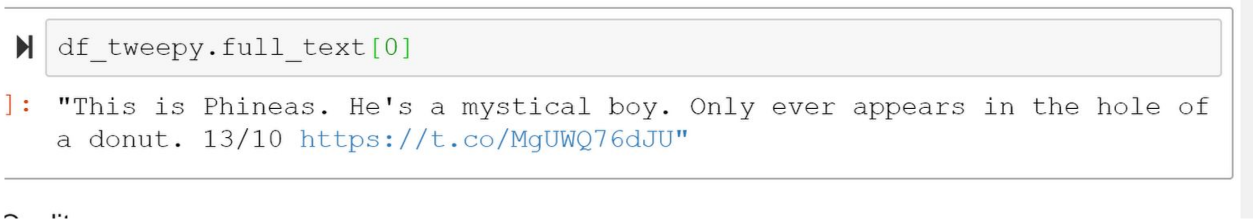
In this project I had to gather, assess, clean and analyze a dataset of tweets about dogs, breeds, their names, scores and number of retweets and favorites. This dataset consisted of three data sets:

1. tweeter archive, provided by Udacity. I had to read that csv file by pd.read_csv()
2. tweeter dog's ranking and image prediction dataset. The url with the file was provided by Udacity, I had to gather it by using Python request library
3. To query Twitter Api using tweepy library to collect extended archive with data about number of retweets and favorites. This one I querried using list of tweet_id extracted from the first archive and the code, provided by Udacity, saved to tweet_json.txt and then read line by line into dataframe.

The average tweet text looks like:

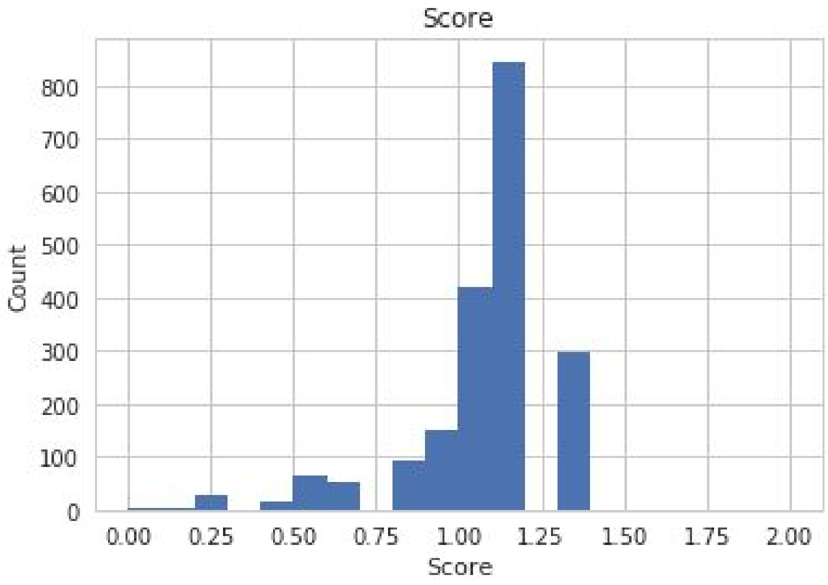


```
df_tweepy.full_text[0]
```

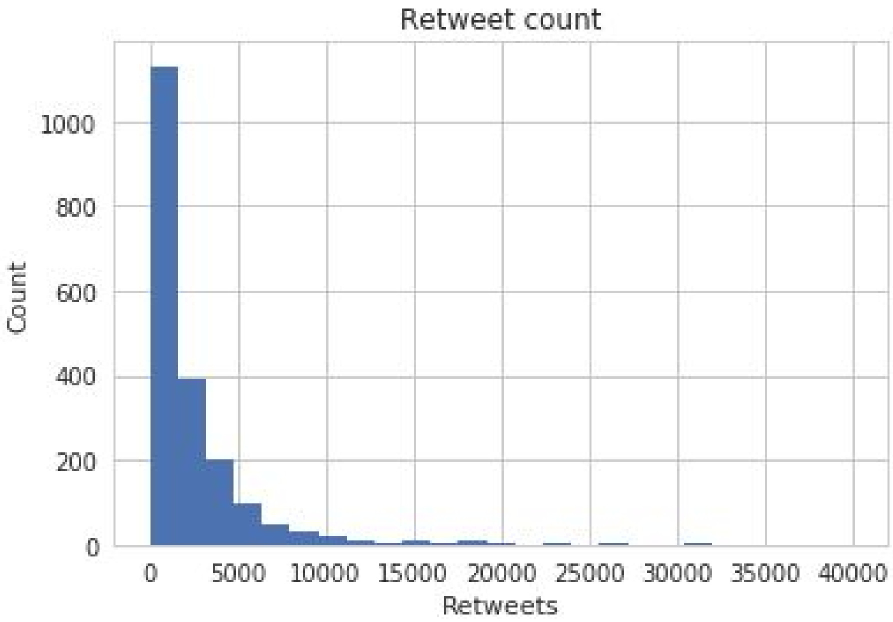
```
]: "This is Phineas. He's a mystical boy. Only ever appears in the hole of
   a donut. 13/10 https://t.co/MgUWQ76dJU"
```

After gathering, assessing and cleaning the three datasets, I've made next conclusions and visualizations:

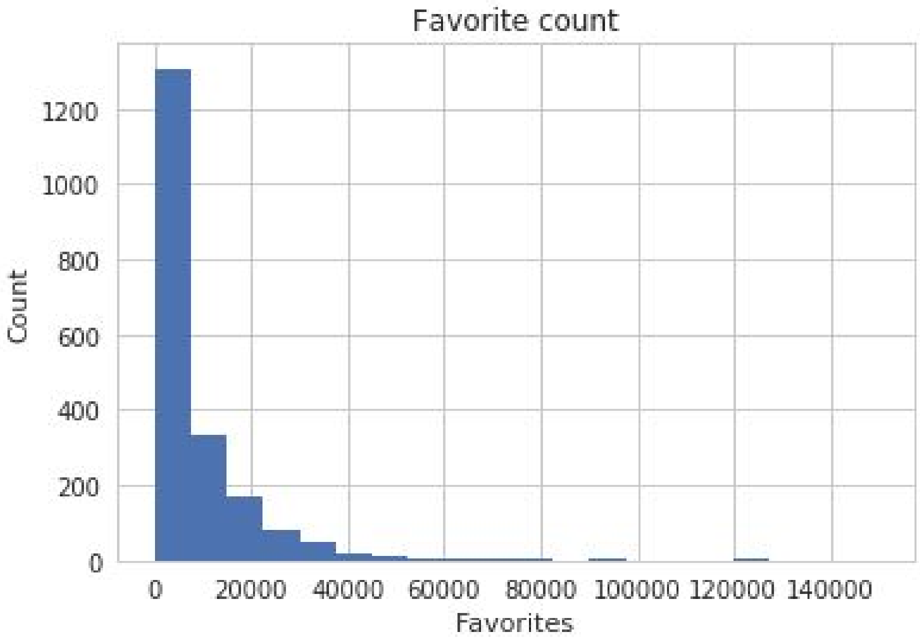
- histogram of scores distribution



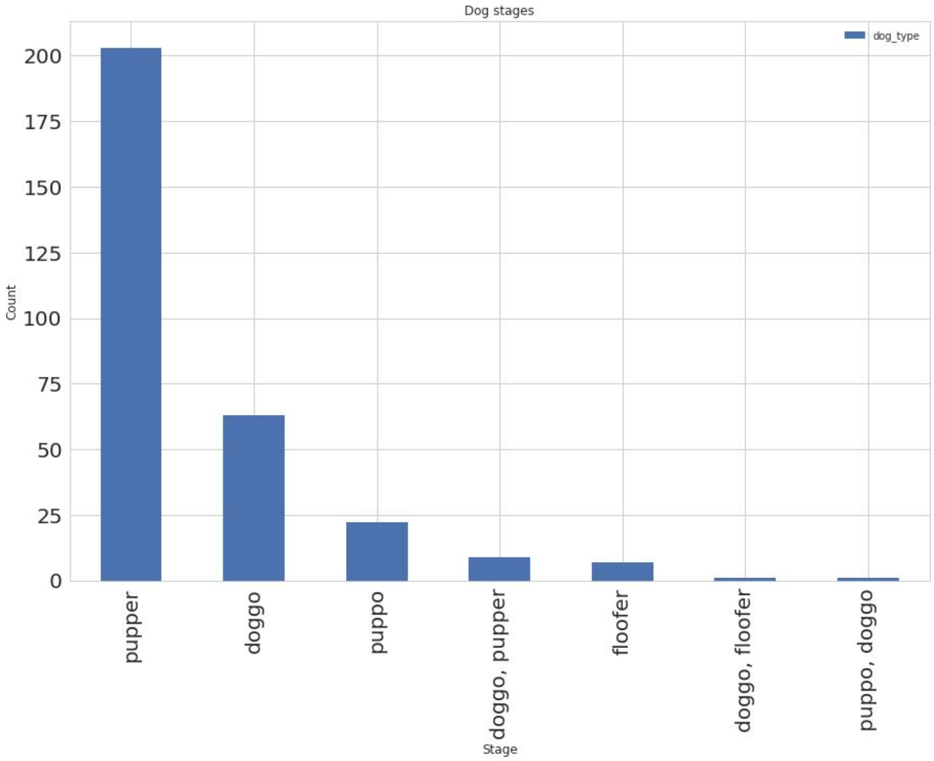
- histogram of retweets count distribution



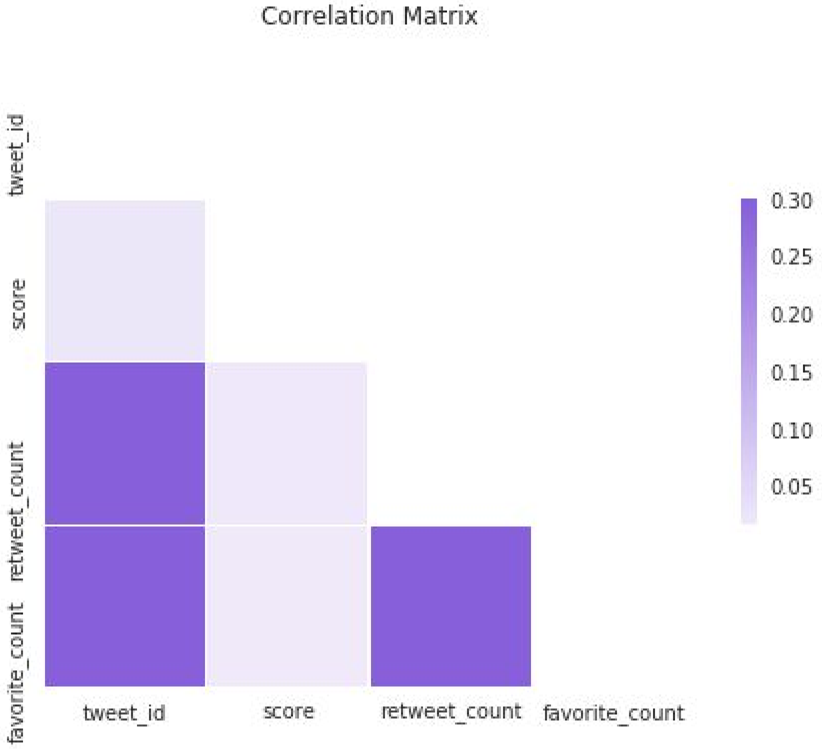
- histogram of favorite count distribution



- bar chart of dog stages distribution



- heatmap of correlation between retweet counts, favorite counts, scores



- as we see there is very strong 0.93 correlation between `retweet_count` and `favorite_count`, which makes sense

- there is almost no correlation between `retweet_count' and the `score`

- there is almost no correlation between `favorite_count' and the `score`

- the most 5 popular dog names are:

Lucy	9
Cooper	9
Oliver	8
Tucker	8
Penny	8

- the most 5 popular breeds are:

golden_retriever	139
Labrador_retriever	95
Pembroke	88
Chihuahua	79
pug	54

Which is true as retriever is considered everywhere as the most popular dog in US.

- the most popular dog stages are:

pupper	203
doggo	73
puppo	23
floofer	7

As we see, we have only 306 stages identified with 1975 unidentified which means
the actual result could be different

- Score:
  - the distribution of scores is left skewed.
  - The range is from 0 till 1.5,
  - the mean is 1.17 and mode is between 1 and 1.25

- Retweet_count:
  - the distribution is significantly right skewed with few outlier which significantly affected the mean.
  - range is from 12 to 83 604
  - mean is 2 649 and mode is in range 0-1000
  - 75% is 3032

```
 - Favorite_count:
   - the distribution is significantly right skewed with few outlier which significantly affected the mean.
   - range is from 78 to 164 220
   - mean is 8724 and mode is in range 0-5000
   - 75% is 10880

"""
```