Act report

Project-4

Udacity Data Analyst Nanodegree

By: Jahnavi Pinnamraju

Data Gathering:

The data was collected from twitter-archive-enhanced which is a csv file. This CSV file contains 2356 columns and 17 rows. It consists of columns like tweet_id, timestamp, source, text, etc. The file named "image-predictions.tsv" was used to analyze the type of dogs present in tweets. The data such as retweet count, friends count, followers count, and retweet count was gathered using Twitter API and JSON library.

Data Assessment:

Data was assessed for the data quality and tidiness. I have identified several issues in the dataset.

Data Cleaning:

I have considered 11 significant issues (9 quality issues and 2 tidiness issues) and solved them. I have defined the issue, solved using the code and tested for the change.

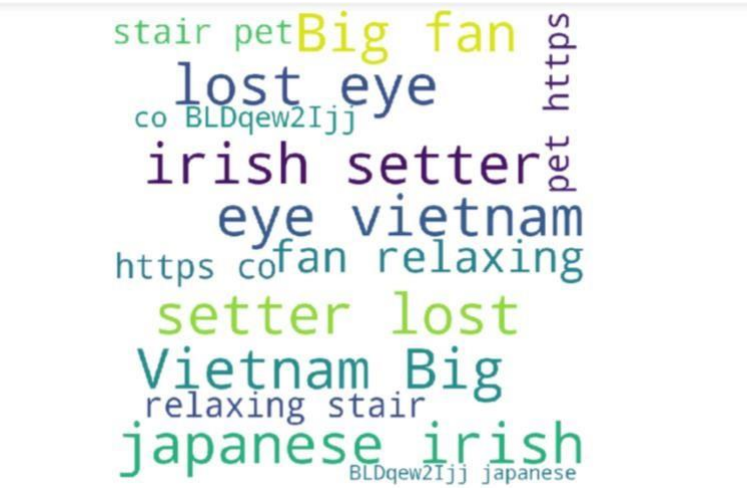
Data Analysis and Visualization:

(i) Correlation Matrix

I have found the correlation between various data elements in the data frame. I used df.corr() method for this purpose. I observed that there is a strong correlation between the tweet_id and in_reply_to_status_id(0.940568).

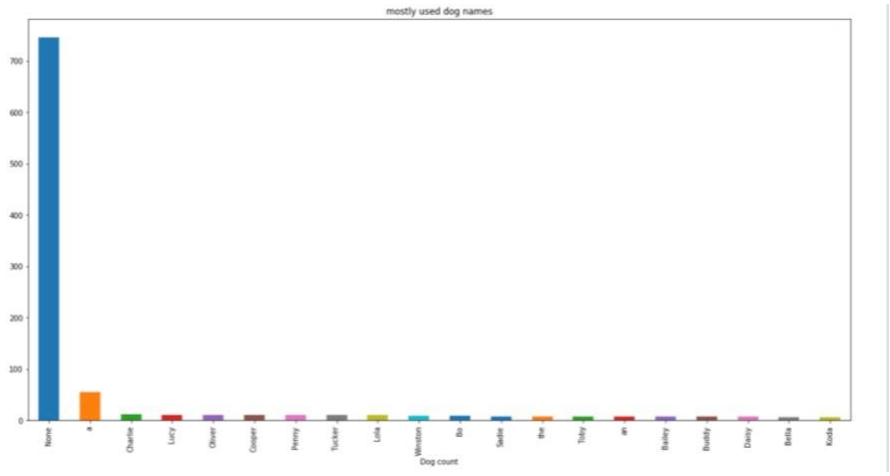
	tweet_id	in_reply_to_status_id	in_reply_to_user_id	retweeted_status_id	retweeted_status_user_id	rating_numerator	rating_denominator
tweet_id	1	0.940568	0.135083	0.744733	0.144644	0.0494963	-0.0282727
in_reply_to_status_id	0.940568	1	0.136589	nan	nan	0.265525	-0.109931
in_reply_to_user_id	0.135083	0.136589	1	nan	nan	-0.0345929	-0.019973
retweeted_status_id	0.744733	nan	nan	1	0.168284	0.17193	-0.037949
retweeted_status_user_id	0.144644	nan	nan	0.168284	1	0.0166741	-0.00966819
rating_numerator	0.0494963	0.265525	-0.0345929	0.17193	0.0166741	1	0.150388
rating_denominator	-0.0282727	-0.109931	-0.019973	-0.037949	-0.00966819	0.150388	1

(ii) Word Cloud



With the help of wordcloud, I have extracted mostly repeated words in the tweets. Japanese, Irish, Bigfan, Setter and lost were some of the mostly used words.

(iii) Barplot for mostly used dog names



From the above barplot, we can observe that a, Charlie, Lucy, Oliver, Cooper, Penny was the mostly used dog names.

(iv). Image(url = 'https://pbs.twimg.com/media/CYN_-6iW8AQhPu2.jpg')

A dog image downloaded using above url.

(v).

	tweet_id	timestamp	source	text	rating_numerator	rating_denominator	
0	892420643555336193	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	13	10	P
1	892177421306343426	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	13	10	
2	891815181378084864	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	12	10	
3	891689557279858688	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	13	10	
4	891327558926688256	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	12	10	F
5	891087950875897856	2017-07-29 00:08:17	<a href="http://twitter.com/download/iphone"	Here we have a majestic	13	10	

Image showing tweet_id along with the tweet, timestamp and source.