WERATEDOGS

A Report on Data Analysis Effort

Mohamed Khaled Ahmed Nour

ABSTRACT

“Valuable insights comes from hard wrangling” Throughout this report we will review the back scene circumstances, steps, actions, encountered challenges and the result during doing the “Data Analysis Process” for “WeRateDogs” Twitter account

Udacity Student

Study Objective

Wrangling and analyzing WeRateDogs Twitter account's tweet data (tweet ID, timestamp, text, etc.) for 5000+ of their tweets as they stood on August 1, 2017.

We will work on:

- Original ratings only (no retweets and/o no replies).
- Tweets that have images only.

Our goal is to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information

Challenges:

1) Messy data with low quality. That requires additional hard gathering, then assessing and cleaning is required for "Wow!"- worthy analyses and visualizations. ¬ Overcoming data Challenge techniques:

1. Take a good time in knowing the datasets and WeRateDogs Twitter account.

2. An organized mindset.

3. Start small-end big way.

4. A lot of practice with "try-error".

2) Twitter Approval Cycle.

Creating a Twitter Developer Account and getting the approval to use Twitter API. It takes a lot of time with a provability to be declined.

- **Overcoming Approval Cycle Challenge techniques:**

a. Start approval cycle early.

b. Demonstrate my request purpose clearly with integrity.

c. Answer every piece of question from Twitter regarding my request.

3) Data Cleaning Complex Code.

Some cleaning actions may require a complex code.

- **Overcoming Complex Code Challenge techniques:**

a. Search at Udacity FWD Community (Discourse).

b. Use google search.

c. Search at (stackoverflow.com, geeksforgeeks.org, github.com).

Background

Study Context

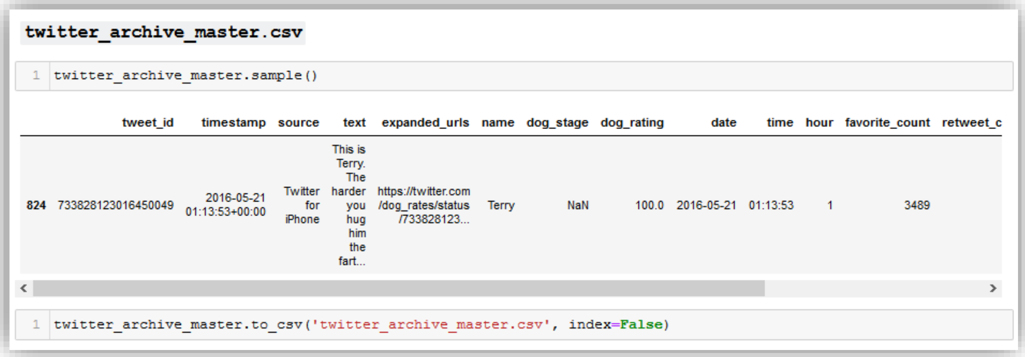
- dog_rates, also known as WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.
- These **ratings** almost always have **a denominator of 10**. The numerators, though? Almost always greater than 10. 11/10,12/10, 13/10, etc
- Nowadays, WeRateDogs has over 8 million followers and has received international media coverage.
- WeRateDogs profile URL: https://twitter.com/dog_rates?s=20

Datasets

WE started with using 3 datasets as "data input", processed them by making required assessing and cleaning actions, finally ended with 2 datasets in hand as " data output". Our analysis done using output datasets.

Output Datasets

Output_File #1: twitter_archive_master.csv



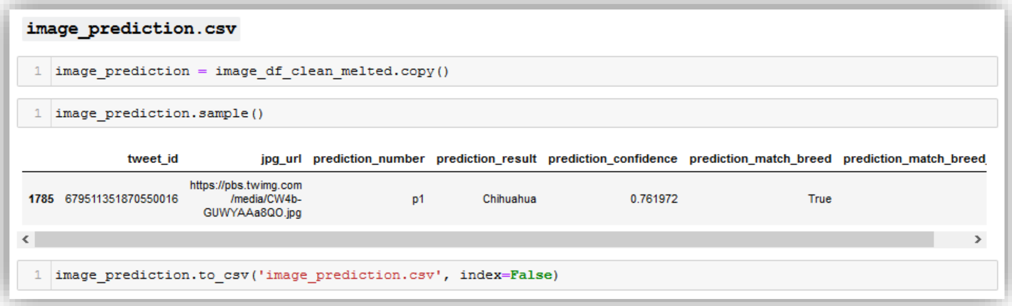
twitter_archive_master.csv

```
twitter_archive_master.sample()
```

	tweet_id	timestamp	source	text	expanded_urls	name	dog_stage	dog_rating	date	time	hour	favorite_count	retweet_c
824	733828123016450049	2016-05-21 01:13:53+00:00	Twitter for iPhone	This is Terry. The harder you hug him the fart...	https://twitter.com/dog_rates/status/733828123...	Terry	NaN	100.0	2016-05-21	01:13:53	1	3489	

```
twitter_archive_master.to_csv('twitter_archive_master.csv', index=False)
```

Output_File #2: imagee_prediction.csv



image_prediction.csv

```
image_prediction = image_df_clean_melted.copy()
```

```
image_prediction.sample()
```

	tweet_id	jpg_url	prediction_number	prediction_result	prediction_confidence	prediction_match_breed	prediction_match_breed
1785	679511351870550016	https://pbs.twimg.com/media/CW4b-GUWYAAa8QO.jpg	p1	Chihuahua	0.761972	True	

```
image_prediction.to_csv('image_prediction.csv', index=False)
```

Data Analysis Result

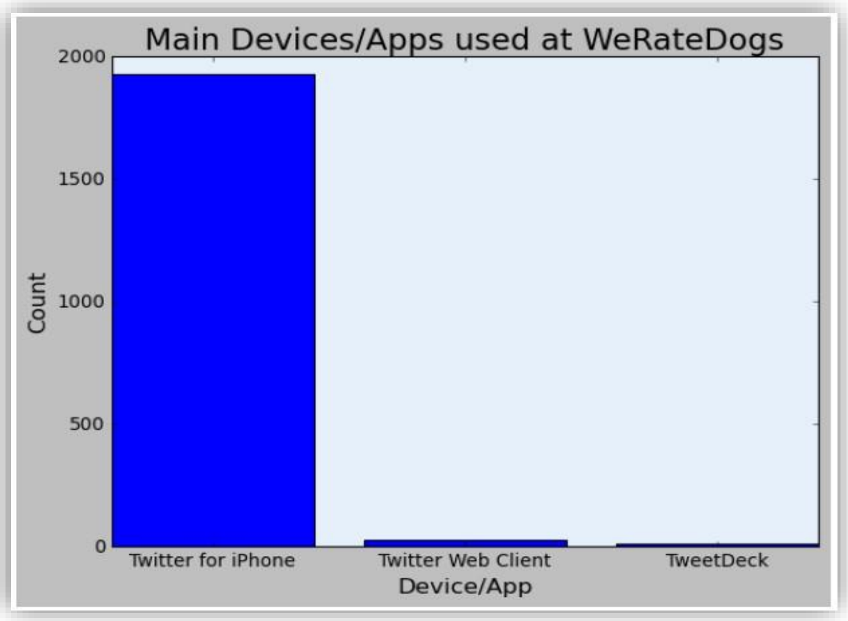
1.Research Questions

Below are what we tried to find out...

Q1 What are the main devices/apps that WeRateDogs' users use?
Q2 Is there a relationship between dog rates and retweet count?
Q3 Is there a relationship between dog rates and favorite count?
Q4 Is there a relationship between favorite count retweet counts?
Q5 What time that most of tweets are tweeted at?
Q6 Is high confidence prediction meet reality more than low ones?

2.Insights and Conclusion

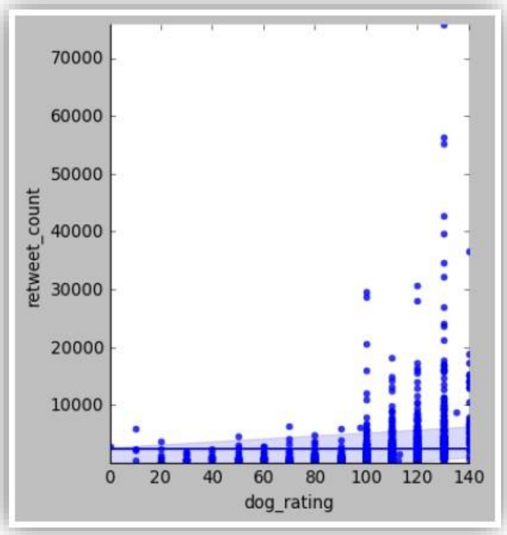
Q1: What are the main devices/apps that WeRateDogs' users use? Chart



Conclusion

Most of WeRateDogs' users are using 'Twitter for iPhone'.

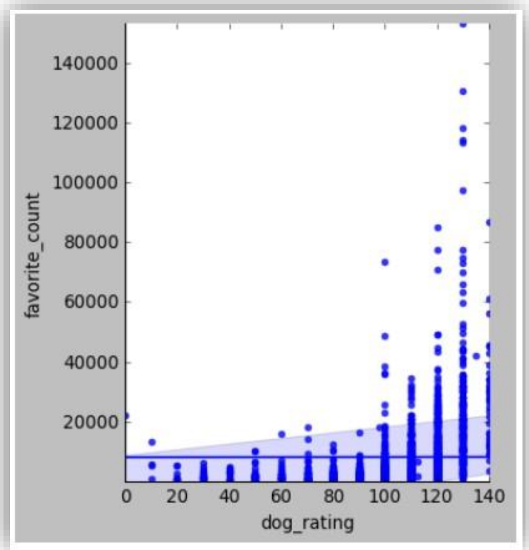
Q2: Is there a relationship between dog rates and retweet count? Chart



Conclusion

It is clear that there is a Positive relationship between dog_rating & retweet_count.

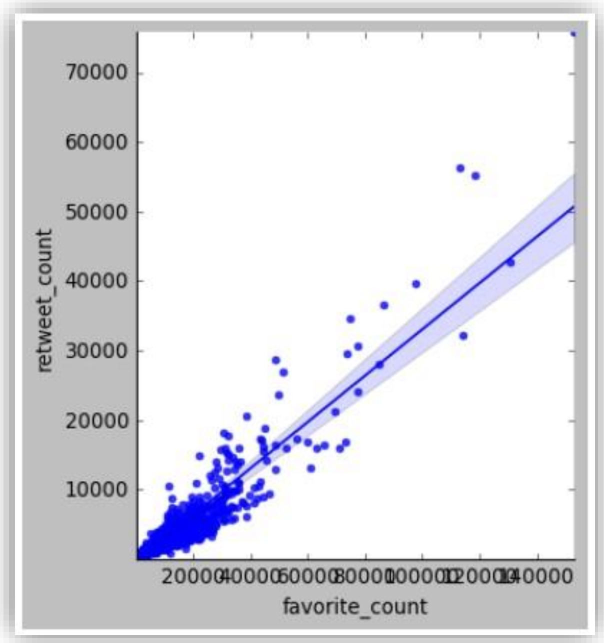
Q3: Is there a relationship between dog rates and favorite count? Chart



Conclusion

It is clear that there is a Positive relationship between dog_rating & favorite_count.

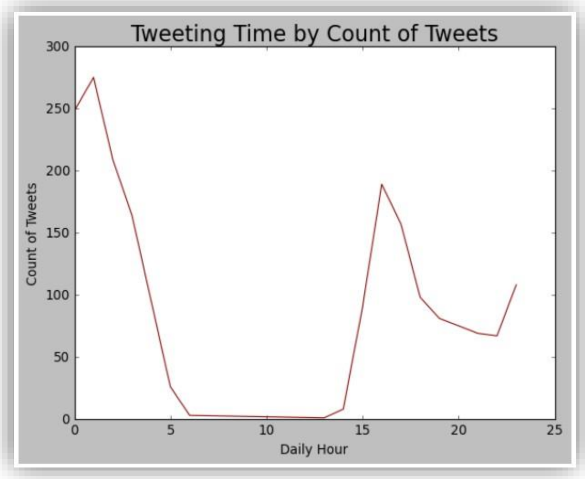
Q4: Is there a relationship between favorite count retweet counts? Chart



Conclusion

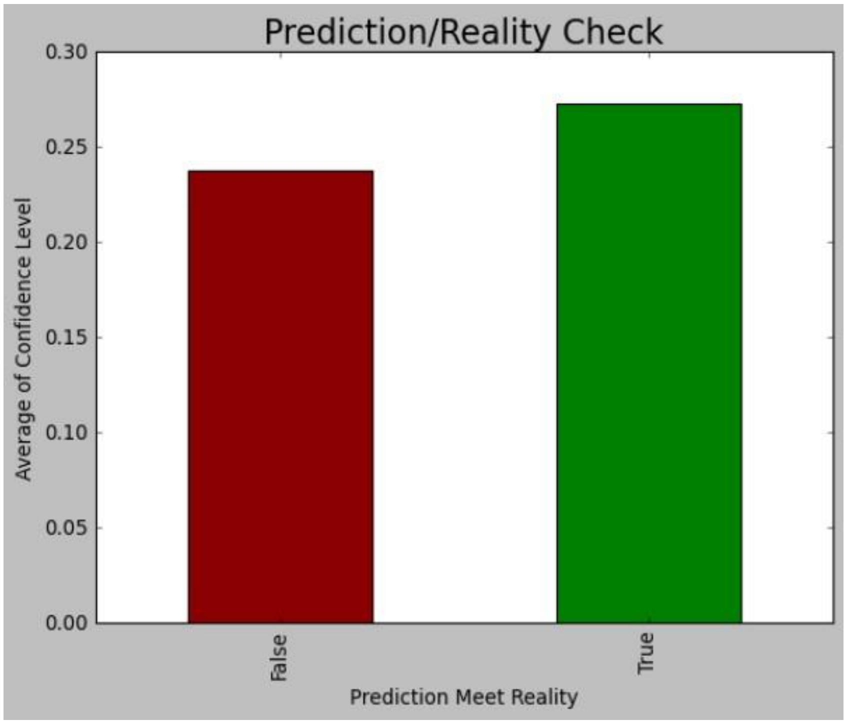
It is clear that there is a Strong Positive relationship between favorite_count & retweet_count.

Q5: What time that most of tweets are tweeted at? Chart



Conclusion

Q6: Is high confidence prediction meet reality more than low ones? Chart

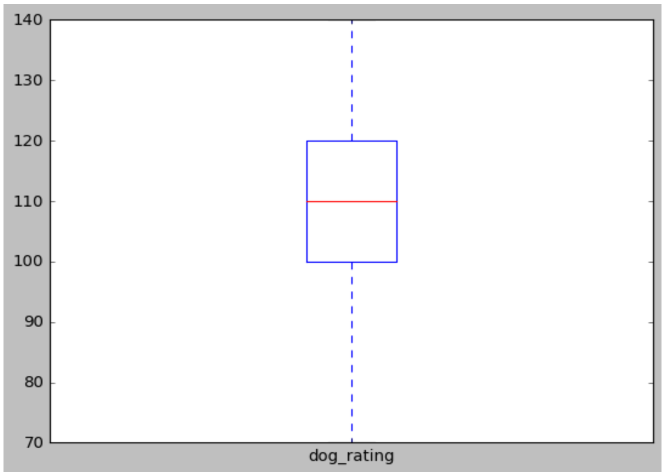


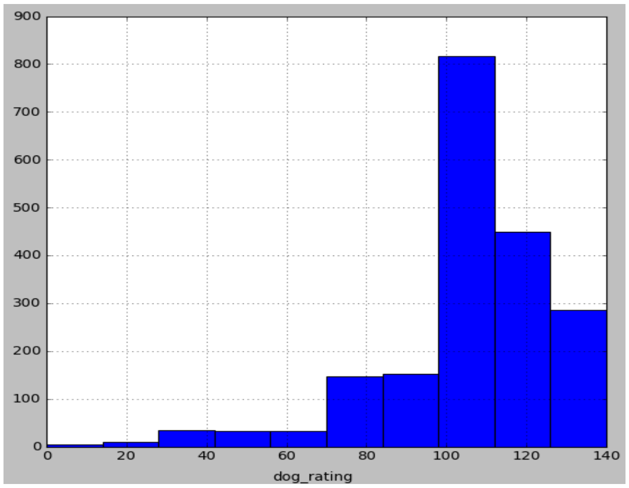
Conclusion

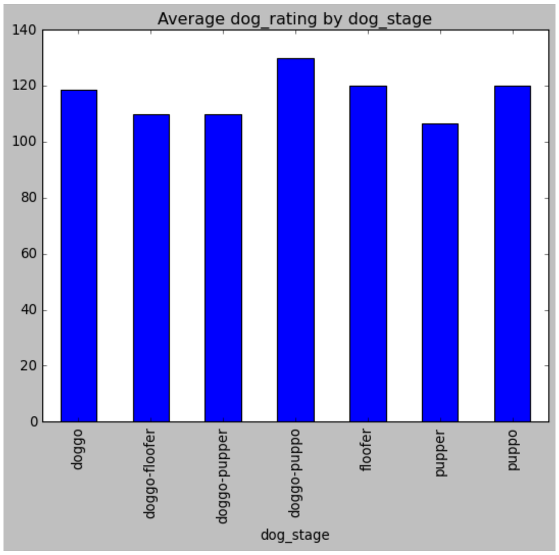
It seems that High prediction confidence level usually has good True reality than Low confidence.

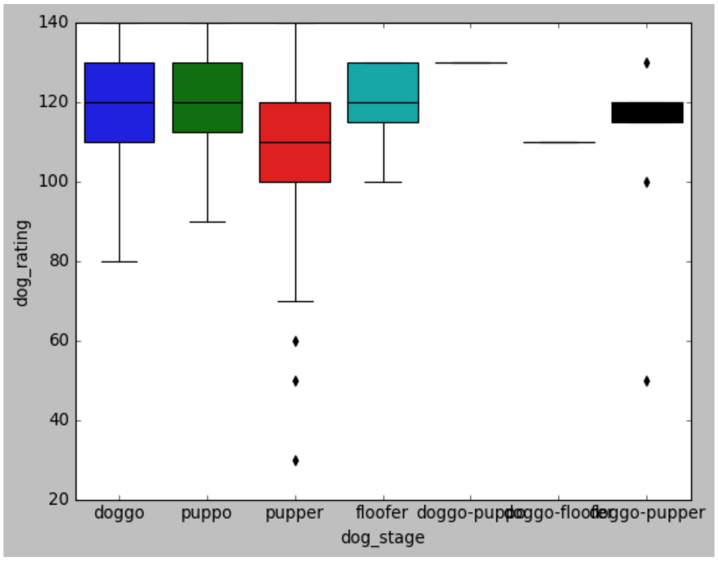
That proves the efficiency of the image prediction model.

3. Additional EDA Figures


dog_rating


dog_rating


Average dog_rating by dog_stage
doggo
doggo-floofer
doggo-pupper
doggo-puppo
floofer
pupper
puppo
dog_stage


dog_rating
doggo
puppo
pupper
floofer
doggo-puppo
doggo-floofer
doggo-pupper
dog_stage