Project Name:

WeRateDogs Wrangle and Analyse Data.

Project Motivation:

Wrangle Provided WeRateDogs Twitter data and using techniques for gathering, assessing and cleaning data to be able to provide analysis, insights and visualizations about the cleaned data.

What are WeRateDogs??

As they describe themselves on twitter, they are source for professional dog rating on Instagram and twitter. They started their twitter account at November 2015 and now they have more than 8 million followers on twitter.

They depend on their unique rating system which you can read about here:

https://knowyourmeme.com/memes/theyre-good-dogs-brent

so, it's not wrong to find rating more than 10/10 based on their system.

Report For:

Communicates the insights and displays the visualization(s) produced from wrangled data.

After completing the three phases of wrangling data:

1. **Gathering Data**
    - ➢ Enhanced Twitter Archive
      Basic Tweets data
    - ➢ Additional Data via the Twitter API
      Data scraped from twitter using there api
    - ➢ Image Predictions File
      Data generated from running photos from twitter archive in neural networks to predict do breed using three different algorithms.
2. **Assessing Data**

    - ➢ Detecting Quality Issues
      Issues with content. Low quality data is also known as dirty data.
      Based on these dimensions:
        - ▪ Completeness

- Accuracy
- Availability
- Consistency

> Detecting Tidiness Issue.
  issues with structure that prevent easy analysis. Untidy data is also known
  as messy
  Based on these rules:
  - Each variable forms a column.
  - Each observation forms a row.
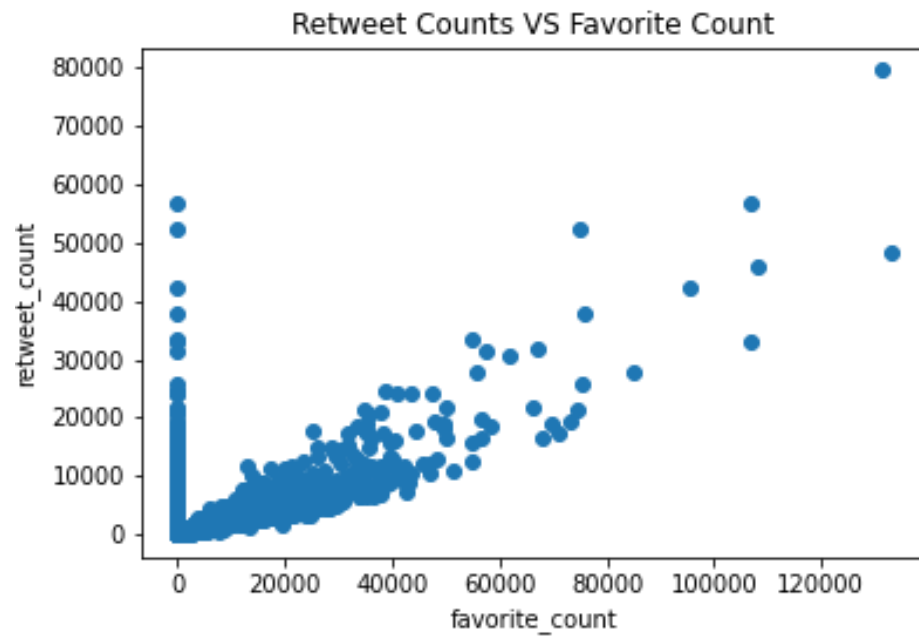 - Each type of observational unit forms a table.

3. **Cleaning Data**
   Basically, running through the previous detected issues and trying to fix it in three
   steps Define, Code and Test

So, these are some insights and visuals about the WeRateDogs Twitter data

Insights:

- **There is positive correlation between retweet_count and favorite_count**

  > It makes sense that there should be positive correlation between number of
    retweet_count and number of favorite_cout as tweets that are most liked
    should be most retweeted.

  > Although the tweet that has max retweet_count is not the same that has max
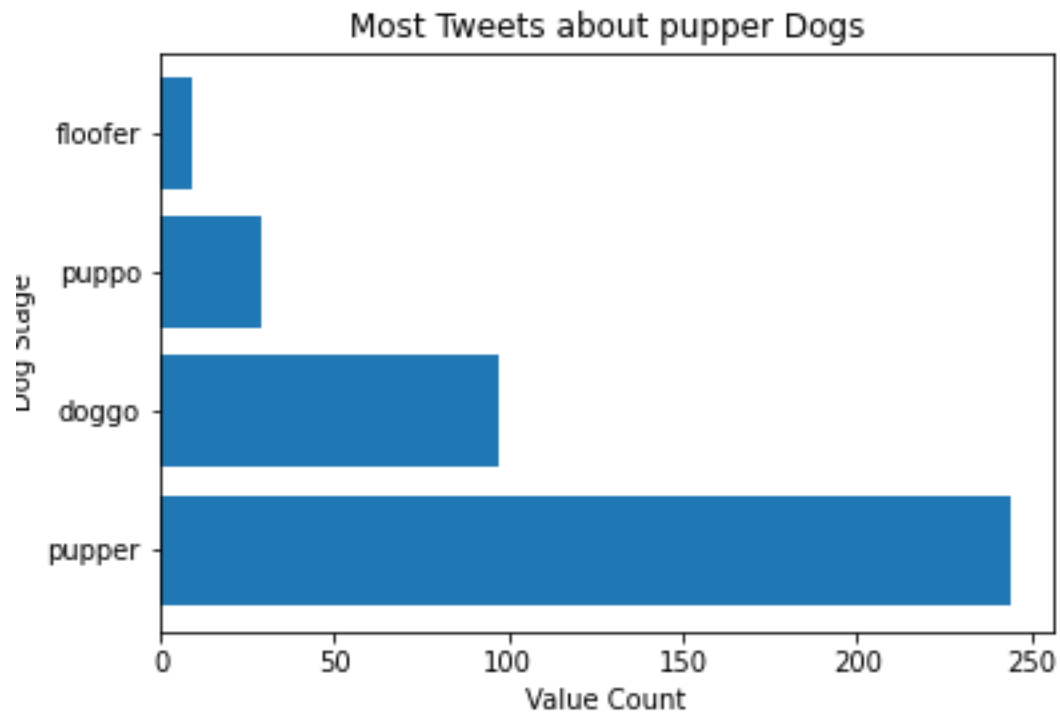   favorite_count

Retweet Counts VS Favorite Count

- **Most tweets about pupper dogs**

  ➢ After seeing stats about dog stage, it pops out that most tweets for pupper compared to other dog stages.

```
tweets_clean['dog_stage'].value_counts()
```

```
pupper     244
doggo       97
puppo       29
floofer      9
Name: dog_stage, dtype: int64
```

Most Tweets about pupper Dogs

- **Top 10 most common names for dogs**

  ➢ Also it appears that regardless the wide diversity in names for dog , there are some names that appears more than the rest, this is the list of top 10 most common names:
  Charlie
  Lucy
  Oliver
  Cooper
  Lola
  Tucker
  Penny
  Bo
  Winston
  Sadie

- **Predict with heighest number of tweets belongs to golden_retriever**

  ➢ After analyzing and combing the twitter data and predictions data for the highest prediction confidence, it shows that all predicts are 215 predictions made but the highest on that shows in most tweets is golden_retriever

  ➢ Here is the list for the top ten:

1. 'golden_retriever'
2. 'Labrador_retriever'
3. 'Pembroke',
4. 'Chihuahua'
5. 'pug'
6. 'chow'
7. 'Samoyed'
8. 'toy_poodle'
9. 'Pomerania'
10. 'cocker_spaniel'