Introduction:

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset that I'll use for wrangle is the tweet 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.

Project Details:

The task in the project is to complete the following:

- Data wrangling, which consists of:
  - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
  - Assessing data
  - Cleaning data

Gathering the data:

Gathering is the first step in the data wrangling process

The data was gathered from the following sources:

- Twitter archive (this is a csv file and was downloaded manually, provided in the classroom - twitter_archive_enhanced.csv)
- Tweet image predictions (this is a tsv file – image_predictions.tsv, hosted on udacity server and was downloaded programmatically using 'Request library')
- Twitter API & JSON (this is a txt file - tweet_json.txt, and was downloaded manually provided in the classroom)

Assessing the data:

Assessing is the second step in the data wrangling process. The data is assessed for:

- Quality: issues with content:
  The quality issues here were missing values, inconsistent data, incorrect names (a, an, the ..), source column having HTML tags, extra columns (doggo, floofer, pupper, puppo) though related to the same variable (name).

- Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:
  - Each variable forms a column.
  - Each observation forms a row.
  - Each type of observational unit forms a table.

  The two tidiness issues that were observed are:
  - Merging the datasets as they all contain details about same tweets.
  - Converting the data type of the column tweet_id for smooth merging.

Types of assessment:

- Visual assessment:
  scrolling through the data in Excel and printing the entire data frame in the jupyter notebook.

- Programmatic assessment:
  using code to view specific portions and summaries of the data (pandas' 'head', 'tail', 'valu_counts', 'sample', and 'duplicated' methods).

Cleaning the data:

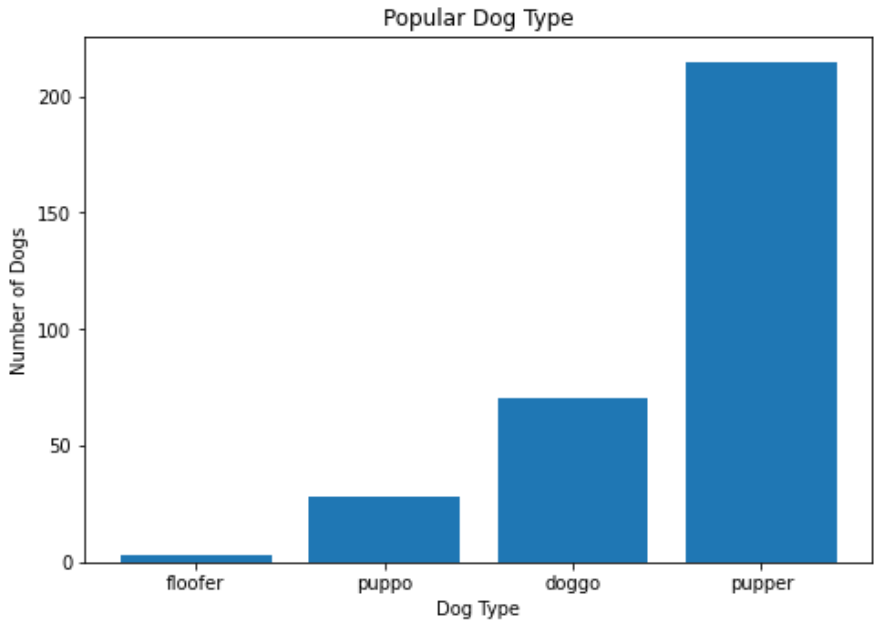
Cleaning is the third step in the data wrangling process. After assessing the gathered data, copy of the merged dataset was made to fix the quality and tidiness issues. During the cleaning process we first "define" what we are going to do. Second, we "code" to fix the issues programmatically. Third we "test" to make sure cleaning operations have worked.

Analyzing and Visualizing the Data:

After cleaning the dataset and storing it in a new file name called 'twitter_archive_master.csv', I've made my analysis on the following points:

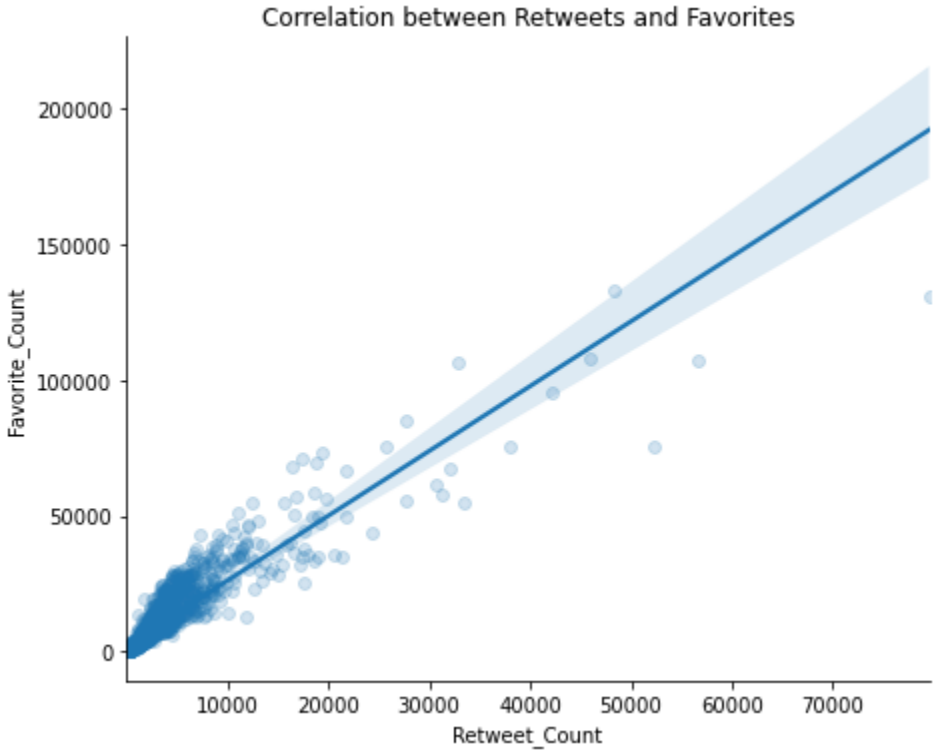
- Most popular dog types
- Is there any correlation between retweet counts and favorite counts?
- What's the common rating number given?
- What's the correlation between each variable?

Most popular dog types:



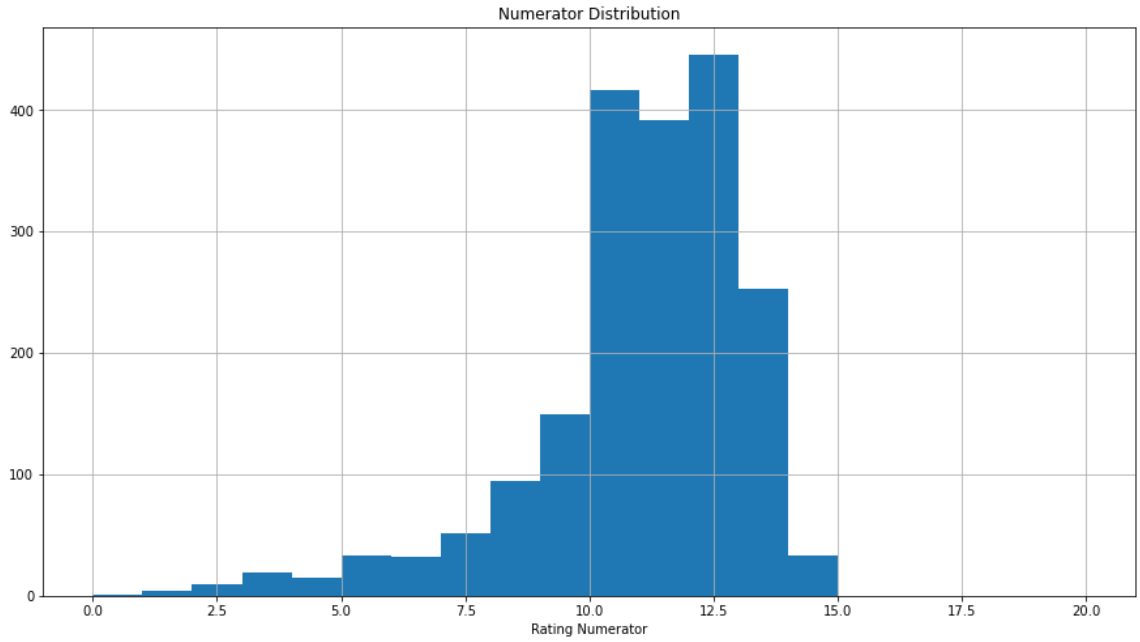
From the above plot, it can be said that the "pupper" is the most popular type of dog.

Is there any correlation between retweet counts and favorite counts?



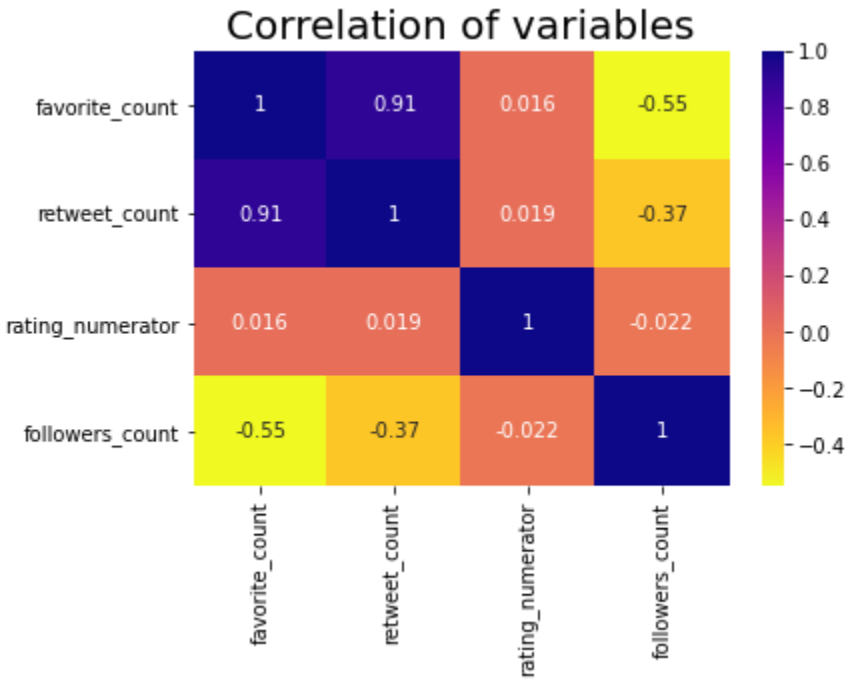
From the plot, it can be said that there's positive correlation between retweets and favorites.

What's the common rating number given?



From the above plot, it can be said that the rating_numerator lies between 10 - 12.5 and 12.5 being the popular rating number given.

What's the correlation between each variable?



From the above plot it can be said that there's positive correlation between favorite_count and retwee_count (0.91) & negative correlation between retwee_count and followers_count (-0.37)