Findings

In the cell below, gather all three pieces of data for this project and load them in the notebook. Note: the methods required to gather each data are different.

Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv).

Quality issues

arch_enhance table

1.There are 6 columns which have many missing values.

2.Incorrect dogs name with "a", "an", "such","the", "very"...

3.tweet_ id, the timestap,rating_numerato and rating_denominator are not correct datatype.

4.In the column rating_numerato and rating_deminator it is needed to correct some decimals.

Predictions table

5.The dataypes of some column are not correct.

6.the datatype of twitter_id should be converted.

7.It is needed to remove '_' ''and make title of data upper in columns p1, p2, p3.

tweets_json table

8.the datatype of favorites and retweets should be converted

Tidiness issues

1.doggo, floofer, pupper, puppo in df1_clean should be in one column.

2.There are three different dataframe which need join or merge with each other.

After Data wrangling process which includes data gathering, assessing and cleaning data. I explain and analyse data with insights and visualization I found that most popular name is Charlie, however most name of dogs is unknown. And in second and third places the names Tucker and Cooper stay.
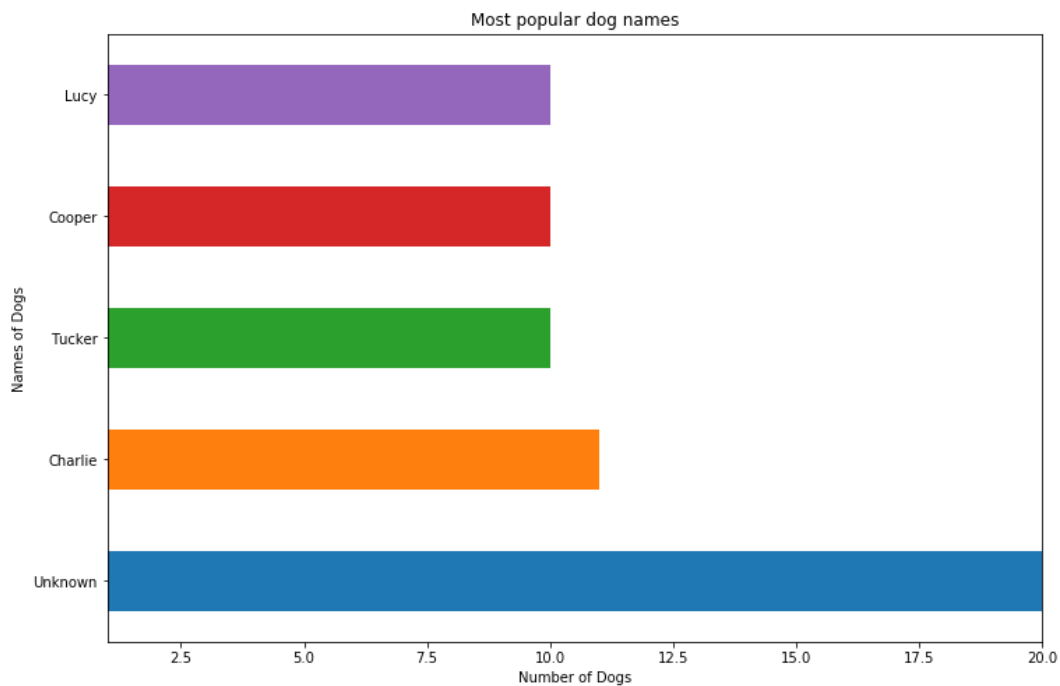
After that, Rating numerator mean is about 12. On the other I found that most Favorites and retweet dogs are Labrador retriever and Lakeland terrier.

At last I found that the golden retriever is most popular dog Breed in the data. In following places the Labrador, Pembroke and Chihuahua are the other popular Breeds.

Most popular dog names
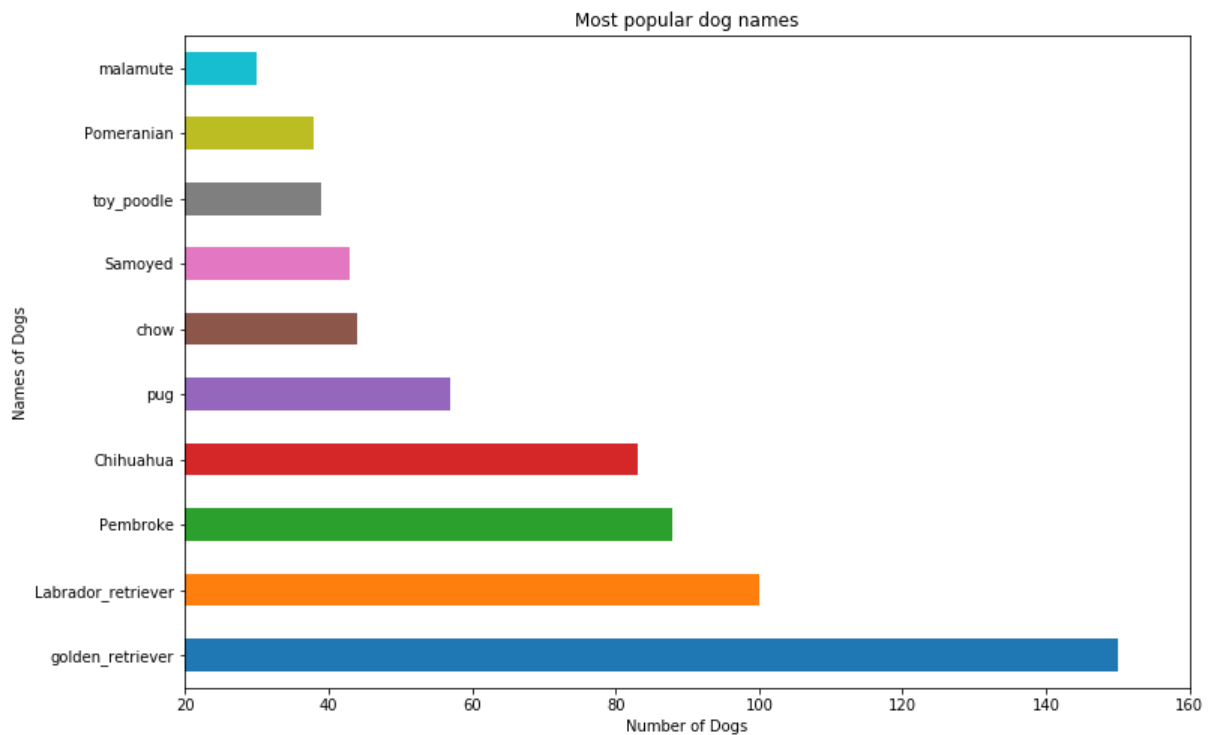
At last I can say that the interest of dogs show the data in limited area, in the future we can analyse and make visualization in most big data which can cover the world rate. But Finally we can say that our data shows Breeds of Retriever are most attractive and popular.