WeRateDogs Wrangle Report

This project aimed to better understand and analyze the account of WeRateDogs, I used three different data sources to get my data:

1. Twitter API data
2. enhanced_twitter_extract
3. NLP analysis data from the photos

all the three data sources are joined via tweet_id, I initially cleaned the Twitter API data throughout multiple steps :

1. keeping only original tweets [excluding retweets and replies]
2. extracting, cleaning and normalizing the rating of each tweet.
3. extracting dog name of each tweet.
4. extracting tweet source (Web, iPhone or other)

then, I joined the clean data with the NLP data (by taking the 1st breed) and joining with the enhanced_twitter_extract data to get the dog type (doggo, fluffo, etc) where I clean as well by pivoting the output in one column.

afterwards, I visualized the output on multiple graphs and the main highlights are :

1. number of tweets deteriorated significantly since 2Q2015.
2. more than 95% of the tweets are published via iPhone.
3. many dogs types couldn't be identified, but the second largest group is "Pupper"
4. more than 60% of the dogs have a rating >8, which confirms the hypothesis of bais-ness of the ratings.
5. Golden and Labrador retrievers are the most dogs being tweeted about.
6. Cooper and Tooker are the most common dogs names on the cleaned data.