Wrangle report

Data collection:

In the phase of data collection, we had to gather data from three different sources, Udacity gave the first one, which contains the archive of tweets of WeRateDogs from 2015 to 2017, the data is about tweets of dogs rated on a particular scale. The second source was from the Twitter API in JSON format, this data set contains the number of tweets and favorites that the tweets got, And the last one is from an AI that predicts the race of dogs based on the image input from the tweets.

Assessing Data:

We looked for quality issues in the data gathered, the first problem found was the retweets and replies that needed to be deleted with the goal of eliminating duplicates and not original tweets. The second problem is that the columns that are not necessary for the analysis need to be deleted to simplify the manipulation of the data after.
The third one is the data types like time is stored as a string, we looked at rows that have a rate less than 10 and found that when the rate is less or equal to 5 the tweets are generally not about dogs, and when the rate is greater than 20 the tweets rated an image that contains multiple dogs or the rate is a double value and needed to be fixed but some tweets are greater than 20 because the author found the dog is so cute.
The tidiness issues found are there is no reason to have three data sets one on enough.

Cleaning Data:

We start cleaning the data after making a copy of the data frames, and we fixed each issue listed above and we stored the data on a new file in CSV format colled **twitter_archive_master.csv**

Data Visualization:

After the data cleaning process, we start our analysis to answer three questions :

1. What is the most popular race?
2. What is the race that gets the most retweets?
3. What is the race that gets the most likes?

And with visualization, we found the dog that has the most retweets and favorites