Wrangle report

Real-world data rarely comes clean but data science can make it much cleaner! Here, we have built interesting and trustworthy analyses and visualizations for We Rate Dogs. with dedicated efforts and deep thinking, we were able to have additional gathering, then assessing and cleaning our data, eventually inferring meaningful and interesting insights and visualizations.
Wrangling we rate dogs can be summarized into 7 phases. The first phase was importing the needed libraries which were determined throughout all wrangling procedures. Secondly, gathering the data files which is very advantageous in python by requesting a certain file through the hosted website's url programmatically. We have gathered 3 files namely: (twitter_archive_enhanced.csv), (image_predictions.tsv) and (tweet_json.txt). these files were different in many things in rows and columns. But we were able to extract the intersected tweets(rows) and add much more columns than a sole one dataset has. We got a comprehensive master database!
Thirdly, the assessing procedure which is the most challenging phase. The master dataset was really suffering from many quality issues and tidiness. By the visual and programmed methods of assessing we write many notes, however we picked some of them to clean. Fourthly, we used loops and many value_counts and find function (loc and conditions) in the cleaning phase in order to get to a better master data with less quality and tidiness issues. In this phase we defined the issues, cleaned and tested them. The cleaning procedures we made is summarized in the following table:

The Cleaning Procedures	Type
1.1 Define Non Descriptive Columns	Quality
1.2 Correct Incorrect Columns Types	Quality
1.3. Fix Source Column's Data to Be Accurately Written.	Quality
1.4 Deleting Tweets That Has No Images	Quality
1.5. Deleting Retweets.	Quality
1.6. Detecting Tweets Has Float Ratings By Searching For Them In The Text And Then Correcting Them To Their Original Values.	Quality
1.7 Deal With Incorrectly Stage Values By Searching In The Text Column.	Quality
1.8 Rating_Denominator Hits Its Thershold (10) And Some Tweets Has Less Than 10. This Is Inconsistent.	Quality
1.9 Get The Dogs Predictions That Their Breed Is Really Dog. This Is By Filtering The Data Frame Based On At Least One Dog Breed Is True From The 3 Algorithms.	Quality
2.1. The 3 Data Frames Will Be Merged Into One Master Dataframe For The Same Observational Unit "Tweet".	Tidiness
2.2. Columns Doggo, Floofer, Pupper, Puppo Shall Be Merged To One Column Called "Stage".	Tidiness

By the end of the cleaning procedure, we have ended up with 1685 x 27 dataframe about we rate dogs tweet that is original not retweet and has image. These retweet has 27 different characteristics and the prediction method revert to dogs(at least one) with valid rating accepting float and with one column defining the dog's stage instead of 3. So were ready to the fifth phase of saving the data into a csv file with panads. Eventually, we could proceed to the most interactive part which is the analyzing and

visualization phase. In this phase, we knew how stage varies, the breeds of the top 25% rating dogs and the relationship between ratings and retweets and favourits.