PROJECT: WE RATE DOGS : DATA WRANGLING & ANALYZE

Starting with a clean slate & an intimidating blank jupyter notebook, I took some time to visualize the analysis & outcomes at the end to begin the process of data wrangling.

I downloaded the twitter archive & image predictions data and had a look at it using excel. I inferred that the bulk of the outcomes were reliant on rating_numerator & timestamp. I also assumed the dog category/stage would also play a huge role in the analysis part which in hindsight was not to be, due to the scale of this particular data.

Starting with the gathering process, I faced the first hurdle in twitter api data due to not being registered in twitter. To circumvent it I used the code & existing json file as instructed in project details. Took some time to organize the data from the json text file but completed with inputs from knowledge home.

Moved to assess phase of the process

- Straightway observed that the Retweeted_status_id columns with values showed there are retweets and structurally columns based on dog_stage was to be collapsed into single column
- For assessing, Focused primarily on the key variables of rating_numerator. Found there were several different issues based on criteria such as denominator > 10 and < 10. Similarly for numerator > 14 & <10. Post feedback from the mentor observed that there were few more issues to be found. Invalid names & multiple dog stages in same picture
- Other issues identified were related to none strings,redundant columns, data type changes, renaming columns & null values.

I started cleaning up the above mentioned issues one by one in the clean phase. I changed the order of the issues (as identified) being cleaned to get to more concise data set first and cosmetic changes at the last

- Used regex patterns to identify few issues with numerator. Came across a very useful site for playing around regex and arriving at the correct one during this exercise. https://regex101.com/
- During this cleanup I missed one glaring error when I tried to fold dog stages to one column where none values in those columns resulted in duplication of rows with more than 8000 rows. Was not aware of this error until the end of the cleaning phase. After identifying this, few code rearrangements had to be made to make it work right.
- After fixing numerator issues, other quality issues were quickly fixed.
- When merging the tables of twitter archives, image predictions & tweet extracts, I removed several columns as they seem to be not required for final analysis.
- From image predictions, I took the first predictions (p1) to proceed further, though I believed that the correct method is to find true values of any of the 3 predictions with reasonable confidence intervals.

Moving on to the analysis part

- I focused heavily on dog ratings , retweet & favourite as that show the rating behaviour of the twitter handle and its popularity among its followers
- Used the timestamp to arrive at these trends over the years
- I separated the data into rows where image predictions were true for dogs and false.
- For dogs = true, I filtered the data frame into new one with dog breeds having frequency of at least 17. This is to get a clearer heat map visualizations that followed
- For dogs = false i tried to see if there are any high ratings which might mean that some of the image predictions could be incorrect. Turns out the intuition was right. The paper tower prediction seems to be way too harsh on this cute one !!