WE RATE DOGS TWITTER ARCHIVE DATASET

OBJECTIVE

Put int practice data wrangling skills using the WeRateDogs, a Twitter page that ranks dogs with a unique metric. Consideration was made with regards to most rated dogs, various stages of development of the dogs and common breeds that are rated together with their names.

PYTHON PACKAGES USED

NumPy, pandas, Matplotlib, Jupyter Notebook, JSON, Requests

STEPS FOLLOWED

- Examined three datasets, df_archive, df_predictions, df_status
- Used Python for wrangling and analysis
- Created visualizations to communicate observations and draw insights

INTRODUCTION

Business data should always be cleaned. This is an iterative process. In this report, we will be wrangling, analyzing, and visualizing the WeRateDogs Twitter archive data. The API approval for elevated access was not approved and thus, the data used was directly downloaded from the classroom notes.

THE DATA WRANGLING PROCESS

Data gathering: Data was obtained from three sources as enumerated below,

- The Twitter archive #WeRateDogs, a csv file that containing the tweets, rating, dog name, and dog 'stage' in life (such as doggo).
- An 'image prediction' file determined by s neural network approximation. The image prediction file was downloaded programmatically from classroom servers using the Requests library.
- Twitter's API data containing retweet count and favorite count. The data was downloaded from classroom notes due to delays in twitter elevated access approval

Data Assessment. The data was assessed on the basis of quality and tidiness.

- Dirty data with issues such as missing, invalid, inaccurate, and inconsistent data was noted for cleaning in the next stage

- Untidy data with structural issues. Was noted. In this case, we checked for columns that need to be combined to form an observation. The datasets were later merged to form a master dataset

Data Cleaning.

- Each assessment was corrected using the define, code and test methodology. The data was then merged to form a master dataset and stored as csv.

CONCLUSION

The following observations were made from the final dataset through merging of th df_archive, df_predictions and df_status datasets.

- The highest probable rating was 14
- Most dogs in the Twitter feed are puppies (or 'puppers').
- Most people post from their iPhone app.
- People prefer to 'favorite' than retweet