**UDACITY DATA ANALYSIS COURSE**

**PROJECT 2**

**DATA WRANGLING**

**By**

**TIJANI MOHAMMED**

**JUNE 2022**

**INTRODUCTION**

As part of the course requirement, I am supposed to undertake several projects as part of my learning process. In this second project, my task was to wrangle data from 3 different sources all centered on data from @weratedogs twitter page.

**MOTIVATION**

Objective is to wrangle @WeRateDogs Twitter data to create interesting and "*Wow!*"- worthy analyses and visualizations.

**DATA SOURCES**

Three data sources were used”:

1. **Enhanced Twitter Archive.** This was provided to be manually downloaded.
2. **Tweets Data.** Additional Data via the Twitter API.
3. **Image Predictions File.** Data consisting of probability of dog predictions using 3 different neural networks.

**DATA WRANGLING**

**Step 1. Gather**

1. I manually download the [twitter\_archive\_enhanced.csv](https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv) and uploaded into my workspace.
2. I programmatically downloaded the image\_predictions.tsv) using this url: <https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv>
3. I then gathered data from the Twitter API by using tweet\_id in my twitter\_archive\_enhanced dataframe created.

**Step 2. Assess**

**Visual assessment.** I exported files into excel to explore and displayed samples in my notebook to have a snapshot of the kind of cleaning to be done.

**Programmatic Assessment.** I used pandas built-in functions like info(), head(), tail(), describe(), dtypes, sort\_values(), value\_counts(), duplicated(), isna(), amongst others to do my assessment.

After my assessment of the 3 dataframes, the issues were found in areas of Duplicate records, Incorrect data types, Missing data, Structural problems, irrelevant column amongst others as indicated below:

***Quality issues:***

1. Delete duplicated rows and records with same tweet id.
2. Delete all retweets rows
3. Identify missing and wrong dog names
4. Delete all rows without images
5. Fix nwrong ratings
6. Remove urls from source colum
7. Extract ratings from text colum and update into respecctive columns
8. Need to convert 'timestamp' column into datestamp
9. Need to correctly identify real dogs with corresponding breeds for analysis

**Tidiness issues**

1. Delete irrelevant columns.
2. Rename some columns headers.
3. Need to melt last four columns into dog\_stage and categorize it
4. Need to merge dfs to make analysis simpler

**Step 3. Cleaning**

The following cleaning were done on each of the dataframes. The cleaning was done iteratively even after, merging and even during visualization.

**For Twitter archived data:**

1. Deleted all rows identified as retweets and reply using in\_reply\_to\_user\_id column.
2. Created a single column ‘stage’ and melted other columns used to derive the stage column.
3. Cleaned the stage name after the melting and also excluded floofer as a stage, replace missing stages with None.
4. Renamed the following columns as indicated:

{'timestamp':'date\_time',

'expanded\_urls':'links',

‘rating\_numerator’:’rating\_num’,

'rating\_denominator':'rating\_denom',},

1. The columns arrangements were also changed to allow easy assessment.

**For image prediction data:**

1. Defined a function to select best prediction amongst all three predictors and created a breed column.
2. All irrelevant columns were deleted.
3. Renamed columns as: {'jpg\_url':'img\_url','p1\_dog':'is\_dog'}

**For Twitter API data for tweets:**

1. Deleted all irrelevant columns
2. Renamed id to tweet\_id for normalization and merging.

**For merged data**

1. All dogs without images were deleted
2. The following datatypes were changed as indicated:

{"date\_time":'datetime64',

"stage":'category',

"retweet\_count":'int64',

"favorite\_count":'int64',

"rating\_num":'float',

"rating\_denom":'float',

"is\_dog":'boolean'}

1. Removed all urls to extract only source names from source columns.
2. Source column was converted to category
3. Missing names and incorrect names were replaced with None.
4. Wrong ratings were corrected by extracting from text column.
5. The denominator column was used to identify groups of dogs and a dog\_count column was created.
6. Group rating was made the average for each group to be in accepted rating range.
7. Some ratings were cleaned visually.
8. Outrageous ratings were also deleted.
9. Breeding names were also cleaned by removing underscore and capitalizing the names. Missing names were replaced with None.
10. Text data was also cleaned of urls with https links.
11. Final cleaning was done to remove certain columns which won’t be needed for our analysis.

**Storage.** Our cleaned data was stored into a csv file twitter\_archive\_master.csv file.

**CONCLUSION**

Below is a snapshot of the information on our cleaned data ready for visualization with a total record of 1959 and 12 columns.

**Graphical user interface

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**REFERENCES**

* Udacity Notes
* Stackoverflow.com
* Geegforgeeks.com
* Pandas Official Documentation