1 Data Wrangling Report

1.1 Data set description

The data set object of this partial analysis exercise is the result of merging two separated data sets provided by UDACITY as final project for the Dada Wrangling section of the Data Analyst Nano Degree. The data sets where limited to the offline version tweeters “WeRateDogs”, due to Tweeter not approved the access to its API.

The data sets: - twitter-archive-enhanced.csv - image-predictions.tsv

1.2 Data Insights and Visualizations

Exploring the cleaned and merged dataset. (Exploratory Data Analisys - EDA)

1.2.1 [Insight 1] Dogs Names and Dog Stage sparcity is the highest rates for the dataset

The sparcity is known as the Missing Values ratio. It can be calculated by total number of lines divided by the number of missing values for each column.

- Dog Stage Nan ratio: 0.8441994247363375
- Dog Names Nan ratio: 0.31255992329817833

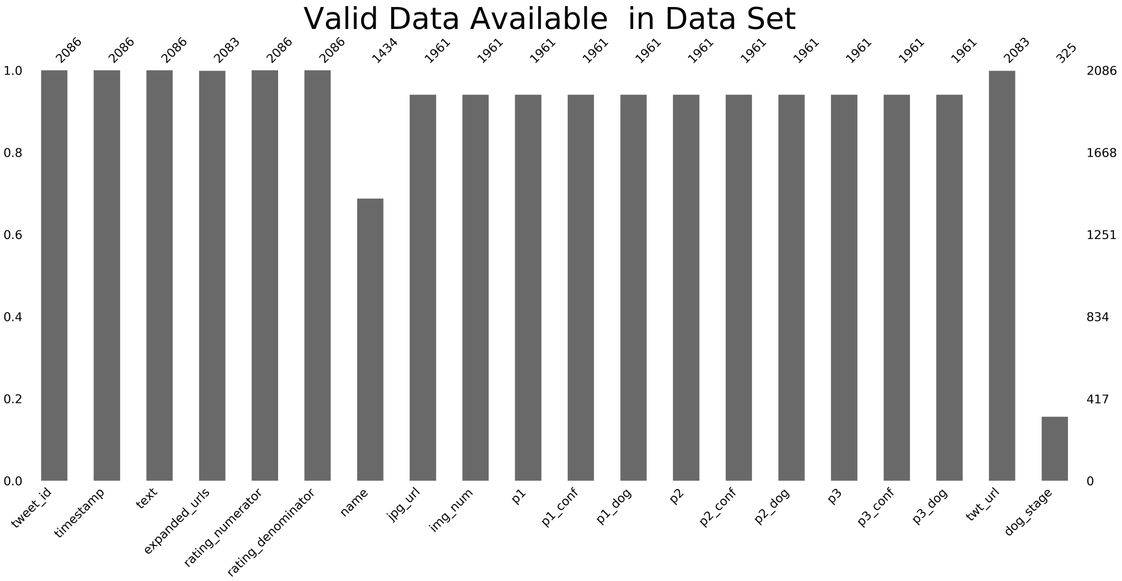


fig 01 - Valid Data for the cleaned data set.

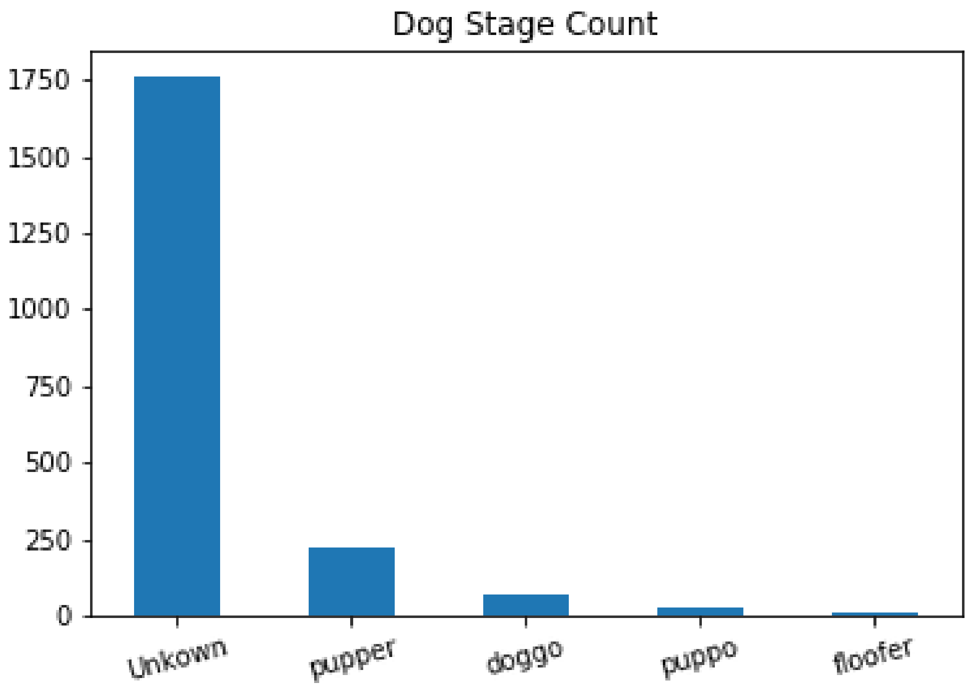
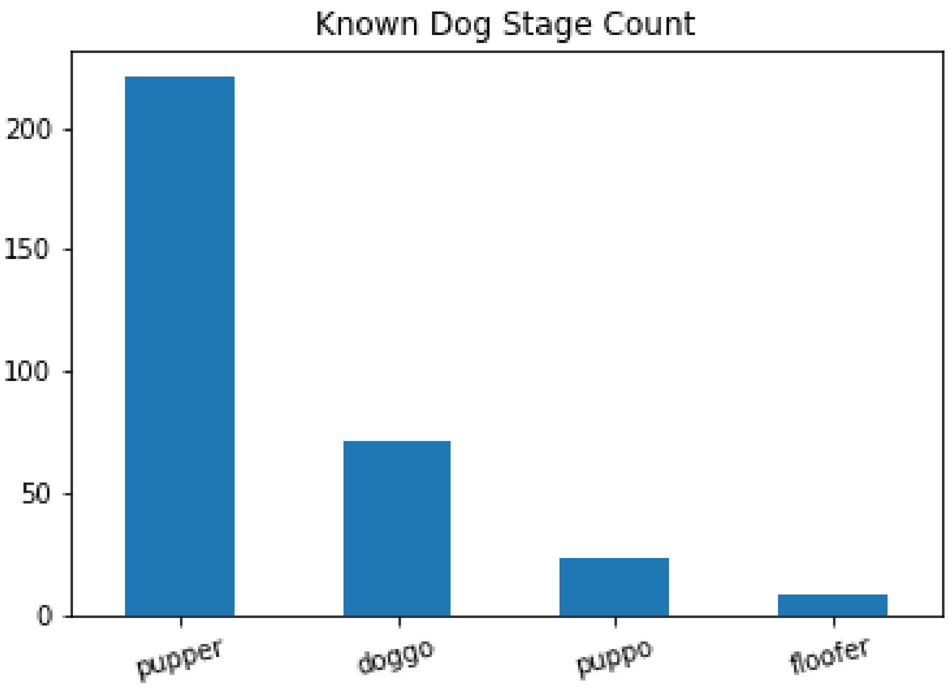


fig 02 - Dog Stages counting.

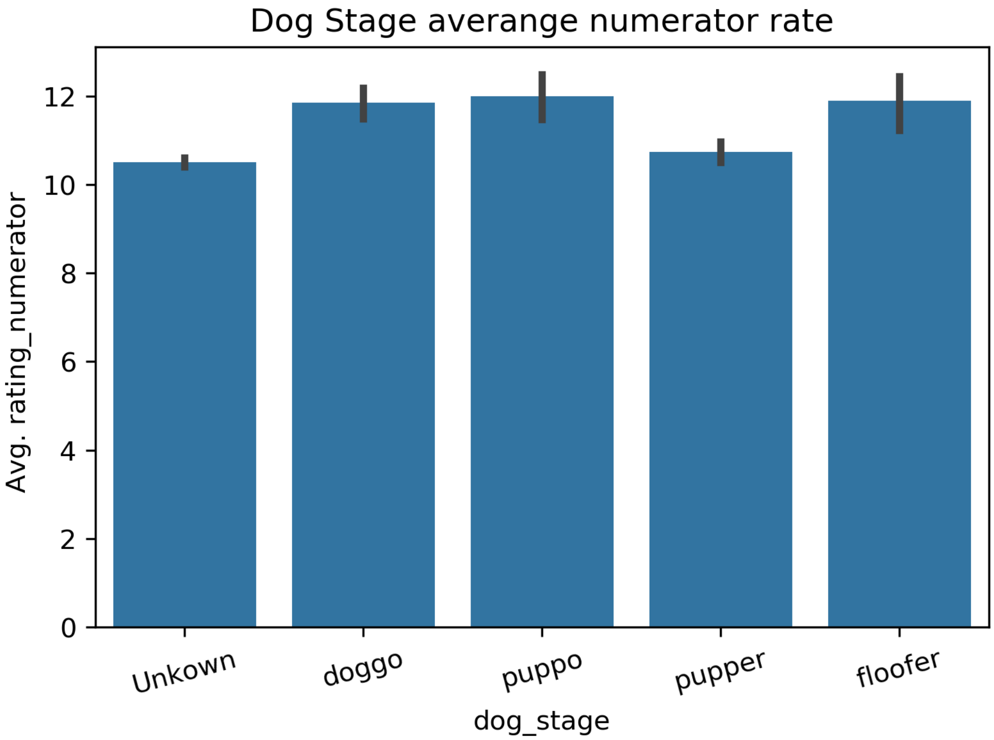
1.2.2 [Insight 2] For the tweets with known valid dog stage entries, the most common is the pupper followed by doggo, puppo and floofer .

Exploring known dog stages and counting unique dog stages for these entries is possible to identify that the most common dog stage classinfication in the data set is pupper followed by doggo, puppo and floofer .



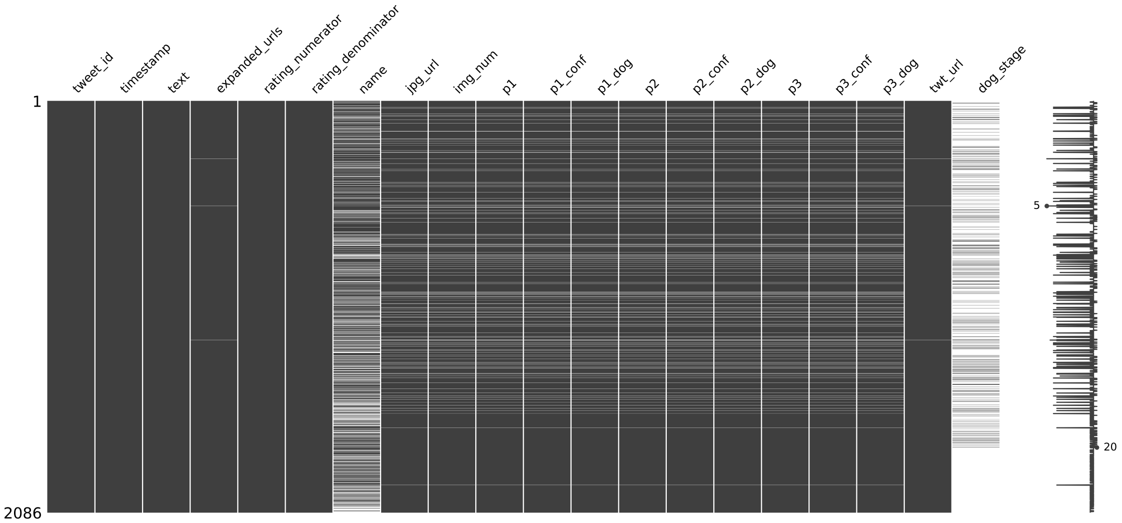
1.2.3 [Insight 3] Puppers dog has the less numerator rating averange.

Exploring the rates for dogs based in the stage it is possible to depict that pupper has the less numerator rate averange.



1.3 Visualization

Visualizing cleaned dataframe sparsity can help directing how the missing values will be treated. Specific for this project, due to high sparsity of Dog__stage column and moderate at name column a good approach is fill missing values with “unknown”.



1.4 Conclusions

Further exploration analysis must be conducted using the data set, it is possible that different strategies for dataframe merging or missing values treatment can rise. Nevertheless, this is why data wrangling is considered an iterative process.