

## D398 Data Wrangling Project

**Introduction:** The purpose of this project is to gather data from three different sources, remediate issues of tidiness and quality, then combine all 3 into one data frame and then validate this file by reporting insights.

- **Process**
- Gather data, the data sources will be:
  - twitter\_archive\_enhanced.csv
  - image\_predictions.tsv
  - tweet\_json.txt
- Import data and create DataFrames.
- Inspect data, report findings for quality and tidiness deficiencies.
- perform remediations on items reported above.
- Merge the 3 DataFrames into 1.
- Report insights on this DataFrame.

Thoughts, insights and findings:

Thoughts:

I found this to be a valuable exercise as it gave me the opportunity to work with the following items / concepts:

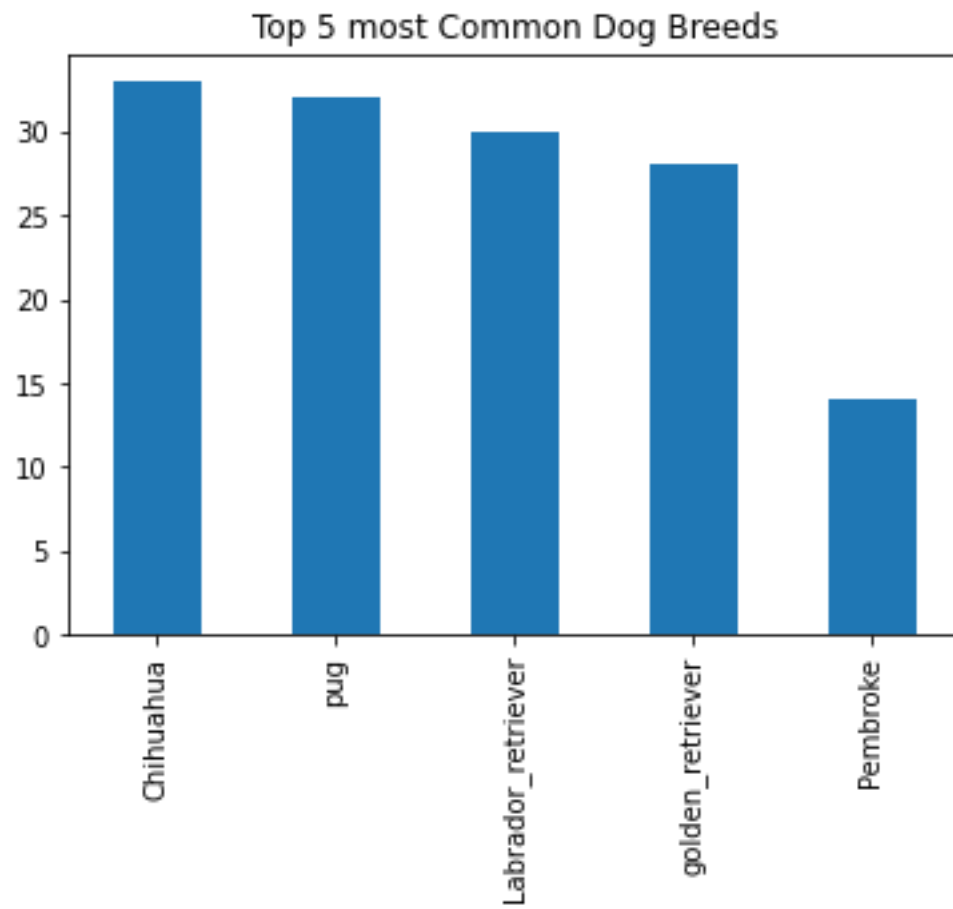
- Merging data frames. As I was familiar with joins from SQL this was not a new idea for me; however, I could easily see the value for future work in Python / Pandas.
- Pandas pivot table functions! Highly informative, the pivot tables are a valuable tool for aggregating data. Doing visualizations for this project gave me a good introduction to using pivot tables with dataFrames. This an excellent method to associate aggregate data
- I am not certain the value of the twitter data, maybe it is because I am not much of a Twitter user but I was unable to see the value this brought to the project.
- Some of the data presented seemed pointless such as img\_num in the image predictions file – was this the number of images or the image sequence number, regardless of answer this seemed pointless.

Insights and findings:

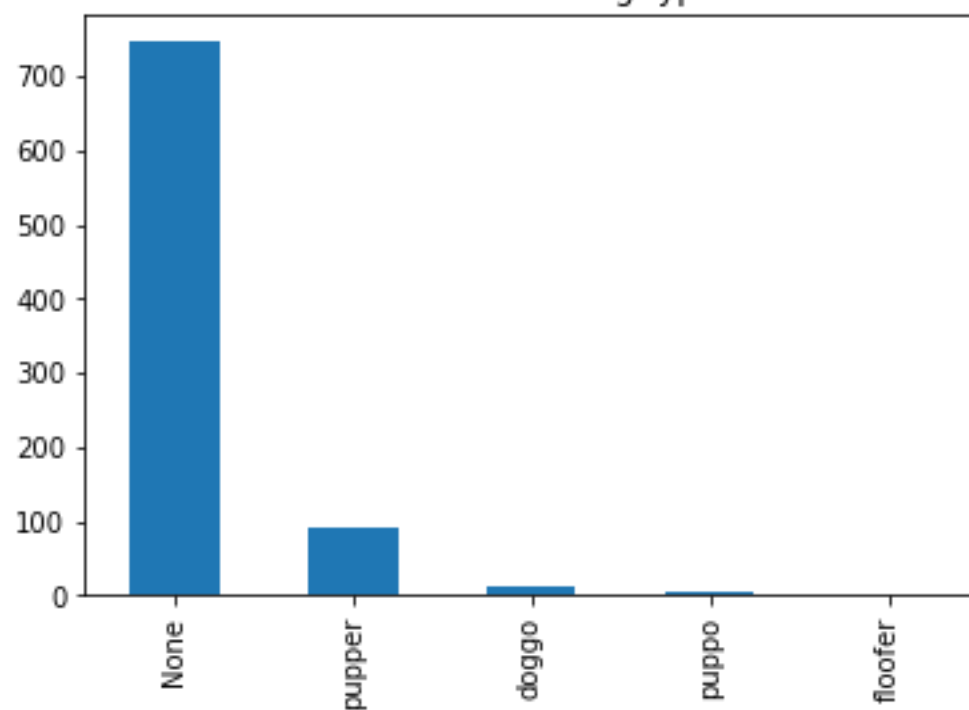
From the merged DataFrame I was able to determine 3 things from the provided data (see visualizations below):

- 5 most common dog breeds.
- Most common category by count
- Category ratings ( I was able to use the pivot table function to relate an aggregate to a category)

## Visualizations for findings



Most Common Dog types



Highest rated Catagory

