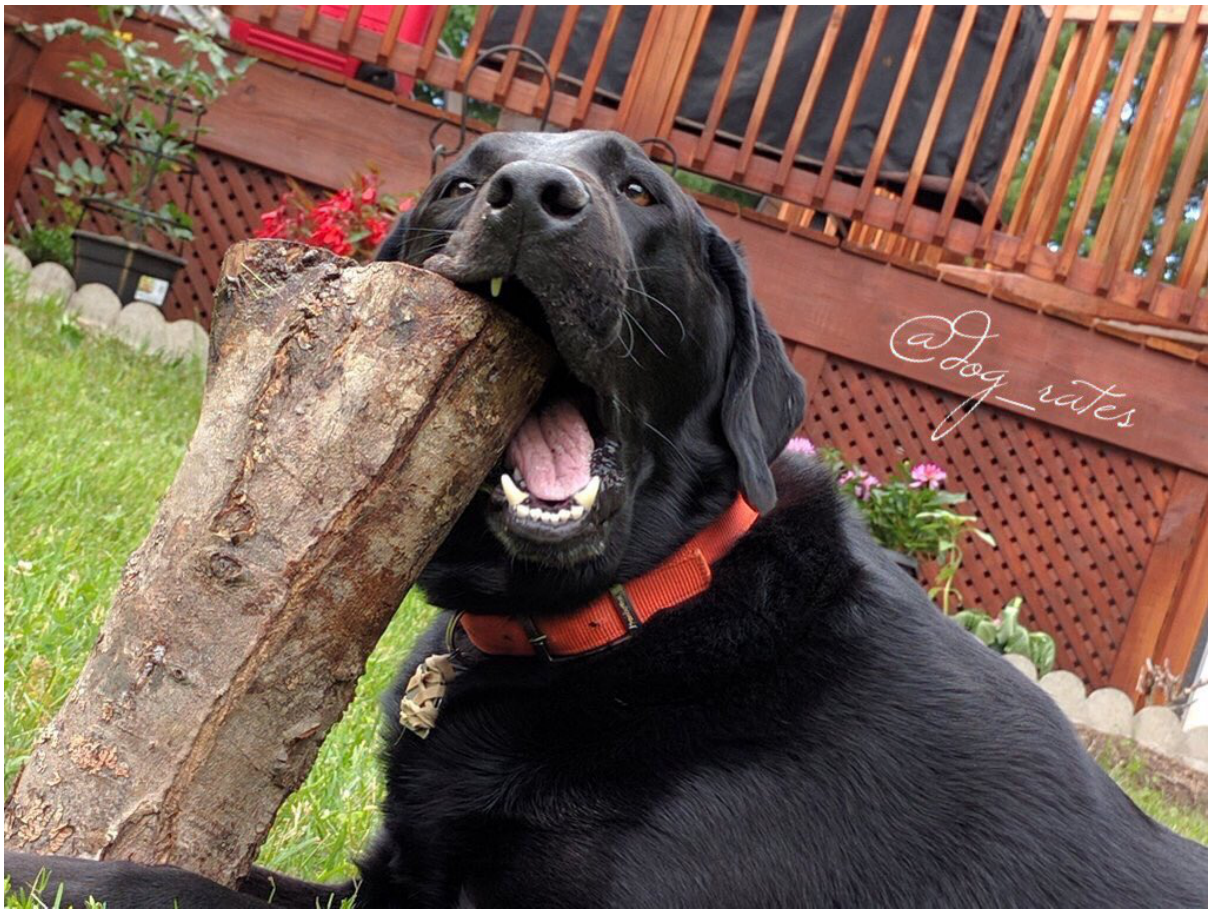
TWITTER DATA ANALYSIS

INTRODUCTION

This project is about using Data to help reach informed and/or scientific opinions for decision making using Data. As well known, good quality data directly implies more accurate, well informed outcomes in the decision making process can be achieved. In most cases, this Data is not readily available in great quality, and often needs to be put through a process of cleaning referred to herein in as Data Wrangling. As part of the Data Analytics course outcomes for the UDACITY'S Nanodegree program, the Data Wrangling Project will explore some main concepts about Data Wrangling to better test the student's ability to Gather, Assess, and Present this Data.



Data Wrangling

This is the process of cleaning and transforming raw data into a much more meaningful, high quality Data framework that can be used for analytical purposes in order to draw accurate and meaningful inferences from that Data. It is often the most critical point in the whole chain of Data analysis, as the decisions or inferences drawn thereof are dependent on the quality of the Data used.

GATHERING THE DATA

The first step in this process will be to import all the relevant libraries that will be used for this project. This helps avoids repetitive coding.

We will gather data from 3 sources that will be used for the purposes of this project. The source names of this data are:

1. Twitter_Archive_Enhanced.csv
2. Image_predictions.tsv
3. Tweet_json.txt

ASSESSING THE DATA

The three types of the gathered data will now be assessed in this segment of the project. Herein, the Data will be assessed visually and programmatically. The issues found during the assessment of the Data will be listed at the end of each Data Type Assessment in their respective headers.

Data Tidiness Issues

1. Dog types are not properly classified. A column will be created to represent all dog type classes

2. The timestamps values are displayed as strings. They should be converted into Date Format.
3. The Columns "Doggo,Floofer,Pupper, and Puppo" should be combined into one New Column Dog Type that will contain the entries "Doggo,Floofer,Pupper, and Puppo" as dog types.

Quality Issues

1. The ratings were not controlled in the data gathering stage, therefore there are ratings that are greater than 10 (The prescribed total rating benchmark) for the poll. Some ratings will be discarded (Specifically ratings above 25 in the rating numerator and denominator columns).
2. Delete unnecessary Columns, that will not be useful for this project.
3. Missing Values are represented by the entry None. Replace these entries with NaN entry.
4. Some tweets have no images. This is because they are often retweets. Only Original tweets will be retained and Retweets will be deleted.
5. The Column text which contains the actual tweets has a misleading name. This column should be renamed to accurately represent the corresponding entries.
6. Some of the entries in the Column Name have *a* which is not a name. They should be recognised as an empty field. Therefore, correct and/or convert these entries.
7. The Column name format is not consistent as it has both lower and Upper cases in the first letter of the name. Make all names have the first letter in Caps format.

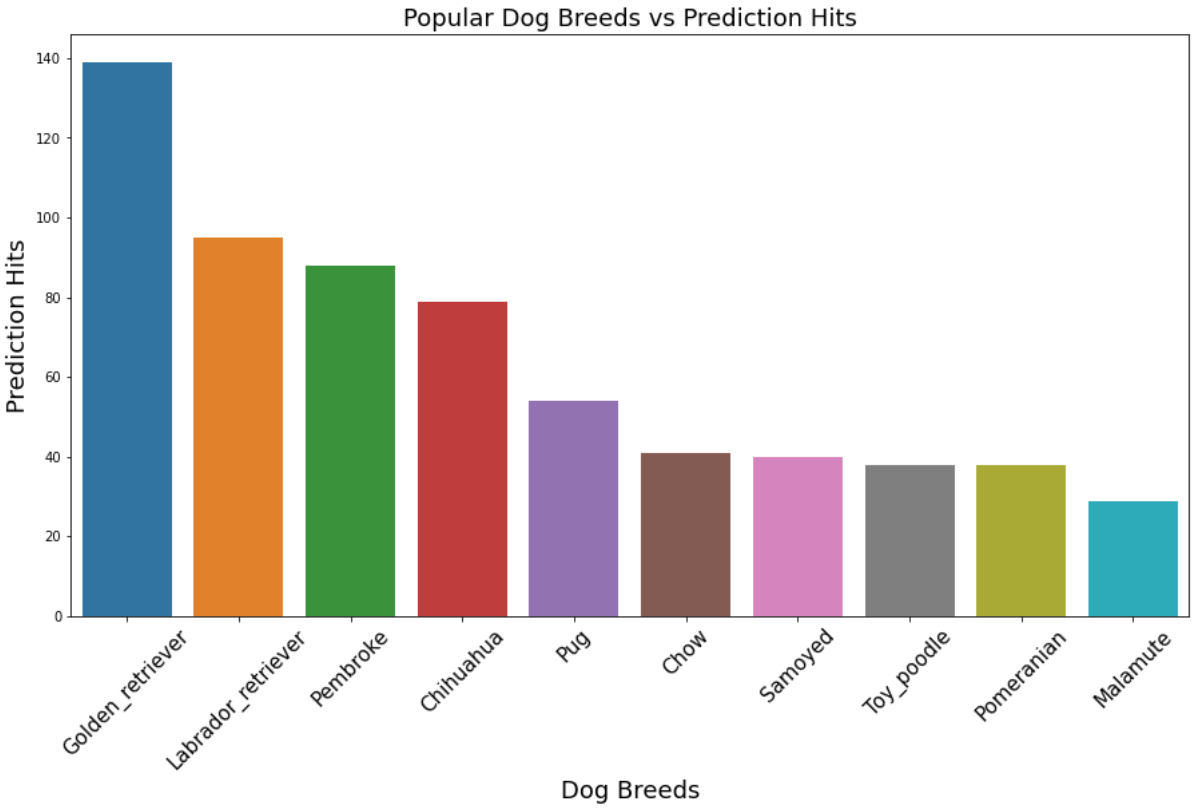
8. There are duplicated images in the jpg_url. Duplicated entries should be dropped.
9. Columns names with "p1_conf, p2_conf, and p3_conf" do not provide sensible information regarding the type of data entries collected in those respective columns.
10. Dog name predictions contain inconsistent formatting (lower and Upper cases in the first letter of the name). Make all names have the first letter in Caps format.

DATA CLEANING

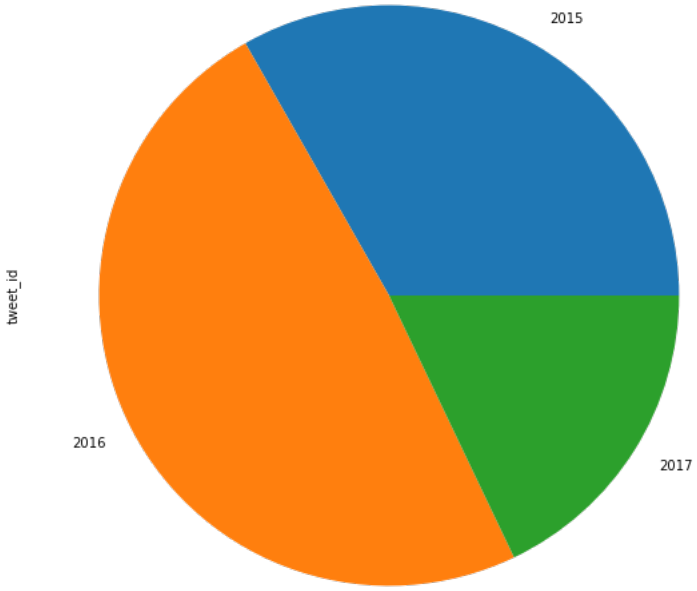
The three Data sources will be cleaned separately before finally combining these Data sources for Data Visualisation purposes. In order to have the original Data Source, in case mistakes are made during the cleaning process, Copies of each Data source will be created below.

DATA VISUALIZATION

Here, we will now look at our 'twitter.csv' data to get a picture of the polling results regarding the vote ratings of dogs to draw meaningful inferences.



The above bar graph shows the popularity of the Golden Retriever breed, which has the highest vote predictions with the highest confidence.



The Pie Chart above shows that the majority of the Votes were achieved in the year 2016. This can be further investigated to find out why votes reduced in 2017.

Conclusions

1. By wrangling the provided data sets, one clearly sees the impact and importance of methodological processes of cleaning data and documentation of each step to allow the reviewer to follow logically the steps and methods employed by the Data Analyst
2. It is clear, that data not always useful in its raw form, until it is manipulated without altering the rawness of the data to get a picture of what the data is trying to paint.