

introduction

The tasks for this project were:

- Data wrangling, which consists of:
 - Gathering data
 - Assessing data
 - Cleaning data
- Storing, analyzing, and visualizing our wrangled data
- Reporting on 1) our data wrangling efforts and 2) our data analyses and visualizations

Gathering data

Download the dataset `twitter_archive_enhanced.csv` provided for this project.

Gather the tweets data set using twitter api tweepy.

Gather the `image_predictions.tsv` provided for this project.

Assessing data

I assessed the data using both visual and programmatic assessment:

- Visually, I assessed data by printing the three entire data frames separate in Jupyter Notebook
- Programmatically, I used different methods (e.g. `info`, `value_counts`, `uplicated`, etc)

Cleaning data

I made a copy of 3 data sets and start clean quality and tidy problems in wrangling notebook

- **quality**

1- The numerator & denominator column is int we must correct to be float

2- The name column has many entries not look like names such as "a", which is not a name.

3- Delete columns that won't be used for analysis

- 4- columns that null values are not treated from null values.
- 5- The timestamp column is an object. It has to be a datetime object.
- 6- wrong rating numeratores were extracted from the text column
- 7- breeds of the dog is inaccurate
- 8- No need to denominator column (we add scale to numerator column header)
 - **tidiness**
- 10- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo
- 11- there are three data sets all tables should be part of one dataset