Wrangle and Analyze Data Project

WRANGLE REPORT

- OMNIA

#Project Overview

Real-world data rarely comes clean. Using Python and its libraries, we can gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. The task is to document our wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) . The dataset that we are wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog

#Step Involved

which regions are involved in Survey?#4 Which regions are involved in Survey? Project Details The tasks in this project are as follows: Data wrangling, which consists of:

1.Gathering data

2.Assessing data

3.Cleaning data

4.Storing, analyzing, and visualizing the wrangled data

#GATHERING DATA

Data was gathered from 3 different sources:

1) The enhanced twitter archive file was provided and downloaded manually which includes various variables for each tweet including tweet id, timestamp, text, rating numerator and denominator, name, etc.

2) Additional data, including favorite count and retweet count, were gathered using Twitter API.

3) The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers.

#ASSESSING DATA

After the data was gathered, assessment was performed using the following methods: Visual Assessment and Programmatic Assessment

- .head()
- .tail()
- .info()
- .value_counts()

Tidiness issues that were cleaned:

- Combining all data frames together as they all contained information about the same tweets
- Combining 4 variables about dog type into 1 column “dog_stage”

#ASSESSING DATA

- Name contained various inaccuracies which were regular lowercase words
- Rating numerators which contained decimals were incorrected exported
- Numerator and Denominator ratings are present differently , combined standard rating need to be provided
- Undesired columns present

#CLEANING DATA

- The three step Process deployed

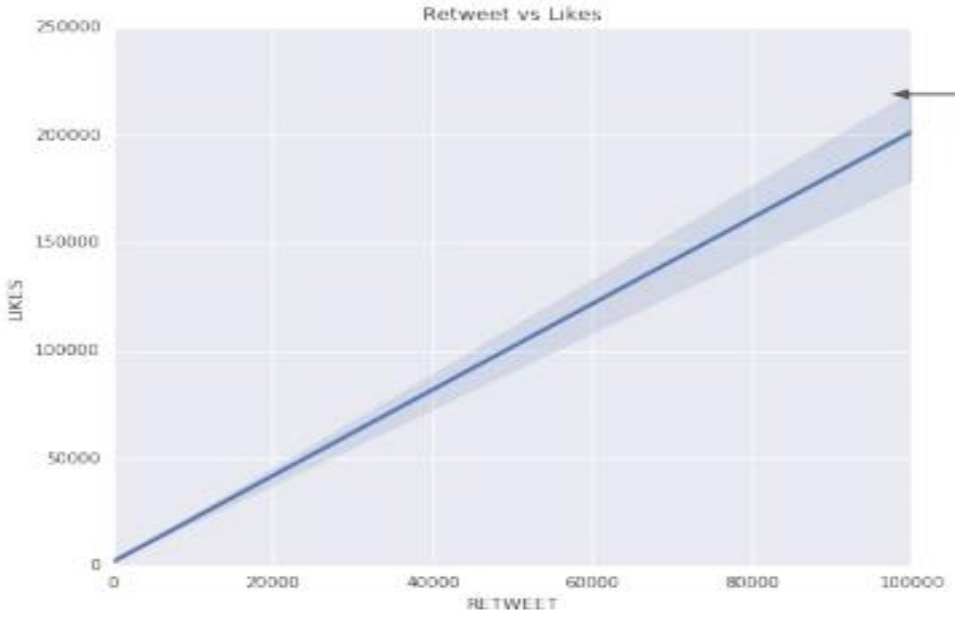
 - DEFINE → CODE → TEST

Used following methods to code and test:

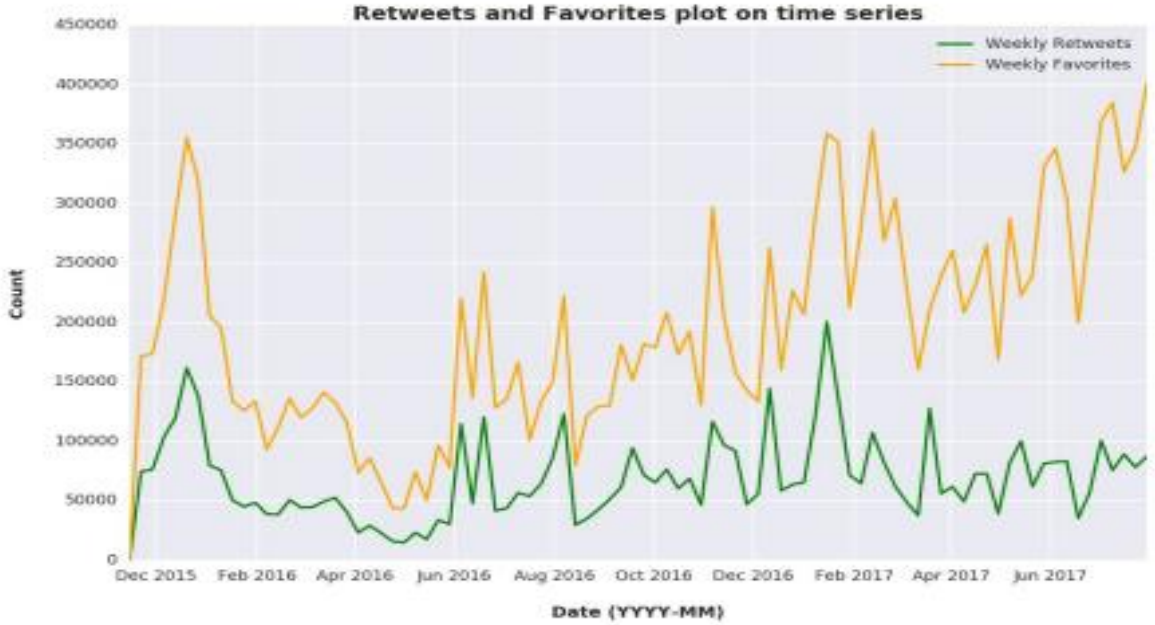
.unique(), .capitalize(), .drop(), .replace(), .merge(), regex,loops, .info(), .head(), .value_counts(), .rename()

After that merged the data in one table and saved in “twitter-archive-master.csv” And then performed few visualization and analyses.

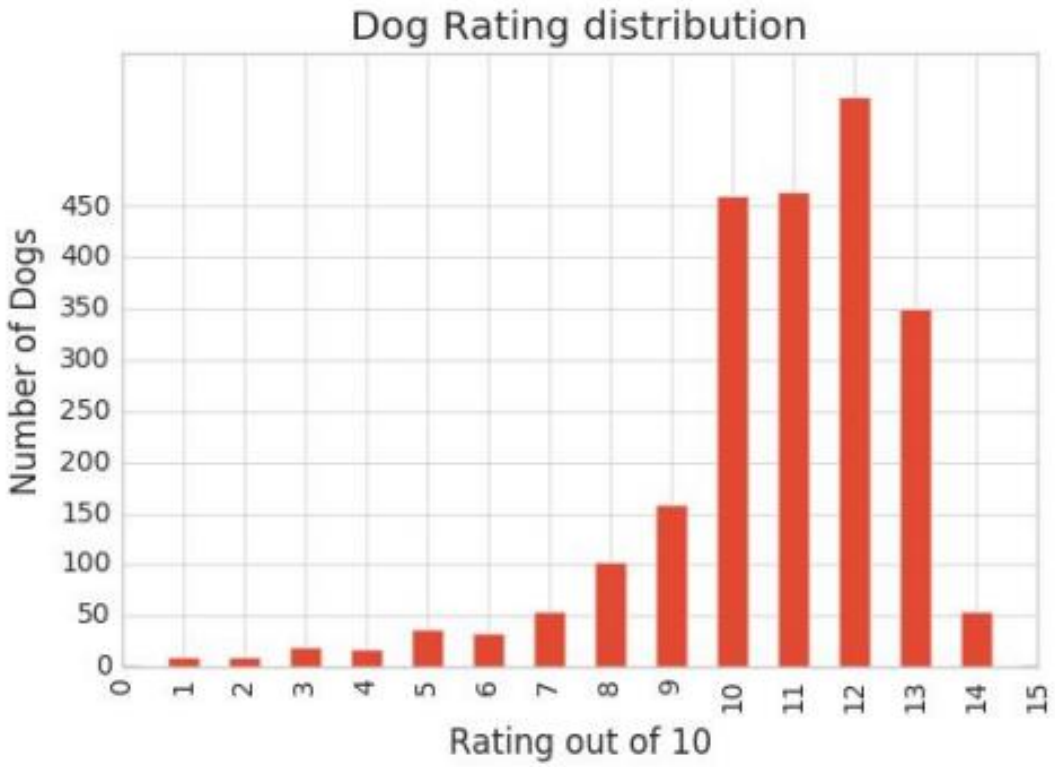
#ANALYSIS AND VISUALS



As the retweets increases,so is the likes and vice versa. There is linear correlation here.



#ANALYSIS AND VISUALS



Most of the dogs are rated on 12 here and the 2nd most rating is 11 and 3rd obviously is 10 as can be visualized here.