Wrangle Report

Wrangle and Analyze
Data Project

By: Mina Ghabbour

- Project Details:

Real-world data rarely comes clean. By 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. Here I will show my wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using different libraries in Python.

The dataset we are wrangling 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.

- Project Steps:

The tasks in this project are as below:

* Gathering data

* Assessing data

* Cleaning data

* Storing, analyzing, and visualizing the wrangled data

- Gathering Data:

Data maily was gathered from 3 different sources:

1) The enhanced twitter archive file 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.

- Assessing Data:

After gathering the data, comes the process of assessing data

and that was done using the following methods:

Visual Assessment and Programmatic Assessment

- The functions below helped assessing the data accurately

  using python:

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

- Cleaning Data:

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 one column “dog_stage”
- 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 dropped

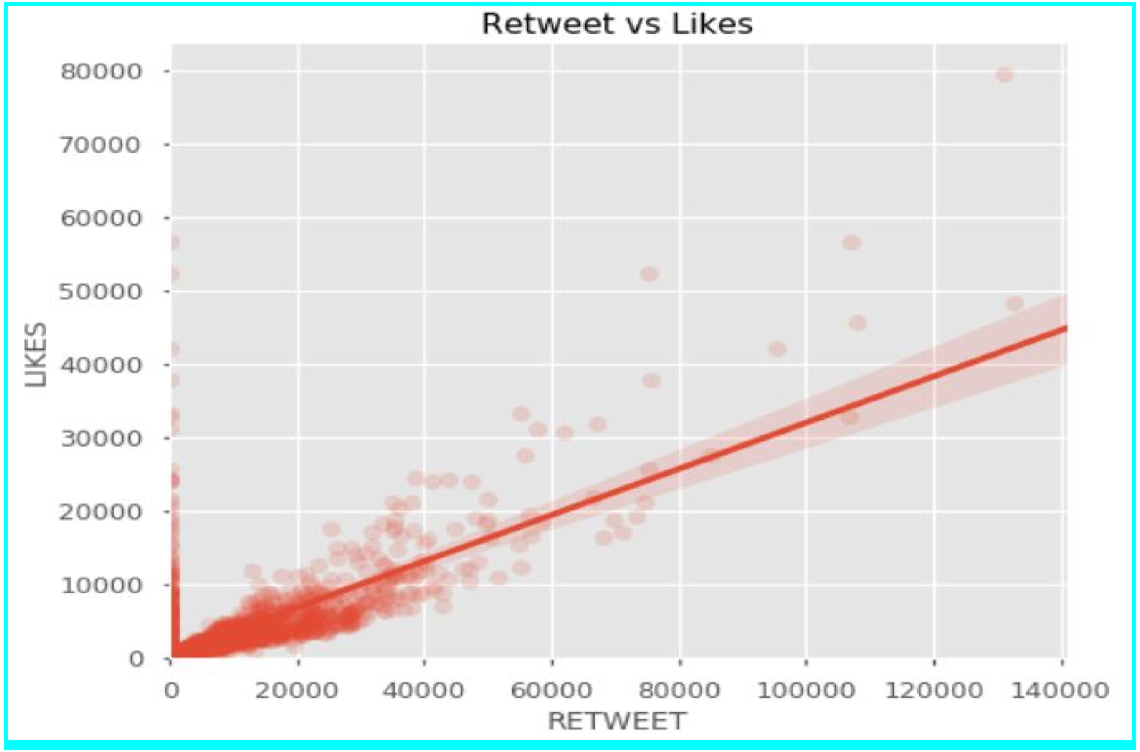
The methods below used to code and test:

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

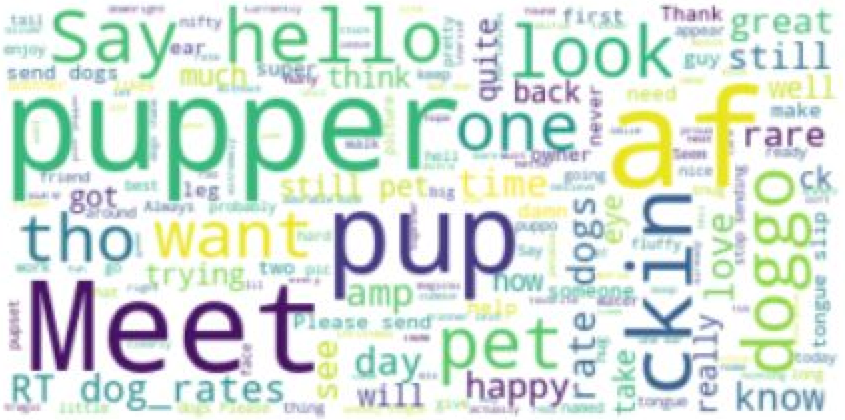
Eventually I merged the data in one table and saved in “twitter-archive-new.csv

- Analysis and Visualization:

- In the plots below we can see the relationship between number of likes and number of retweets
- We can find that when the retweets increases the likes also increases


Retweet vs Likes
LIKES
80000
70000
60000
50000
40000
30000
20000
10000
0
RETWEET
0
20000
40000
60000
80000
100000
120000
140000

- Wordcloud:



pupper
af
Meet
pup
Say hello
look
one
tho
want
pet
ck
kin
doggo
RT dog_rates
rate dogs
still
great
rare
happy
know
day
amp
time
still pet
trying
love
really
back
quite
think
much
will
see
take
now
got
send dogs
Please send
super
two
well
make
need
first
Thank
guy
eye
someone
owner
never
leg
tail
nifty
ear
enjoy
fluffy
puppo
damn
nice
tongue slip
today
little
thing