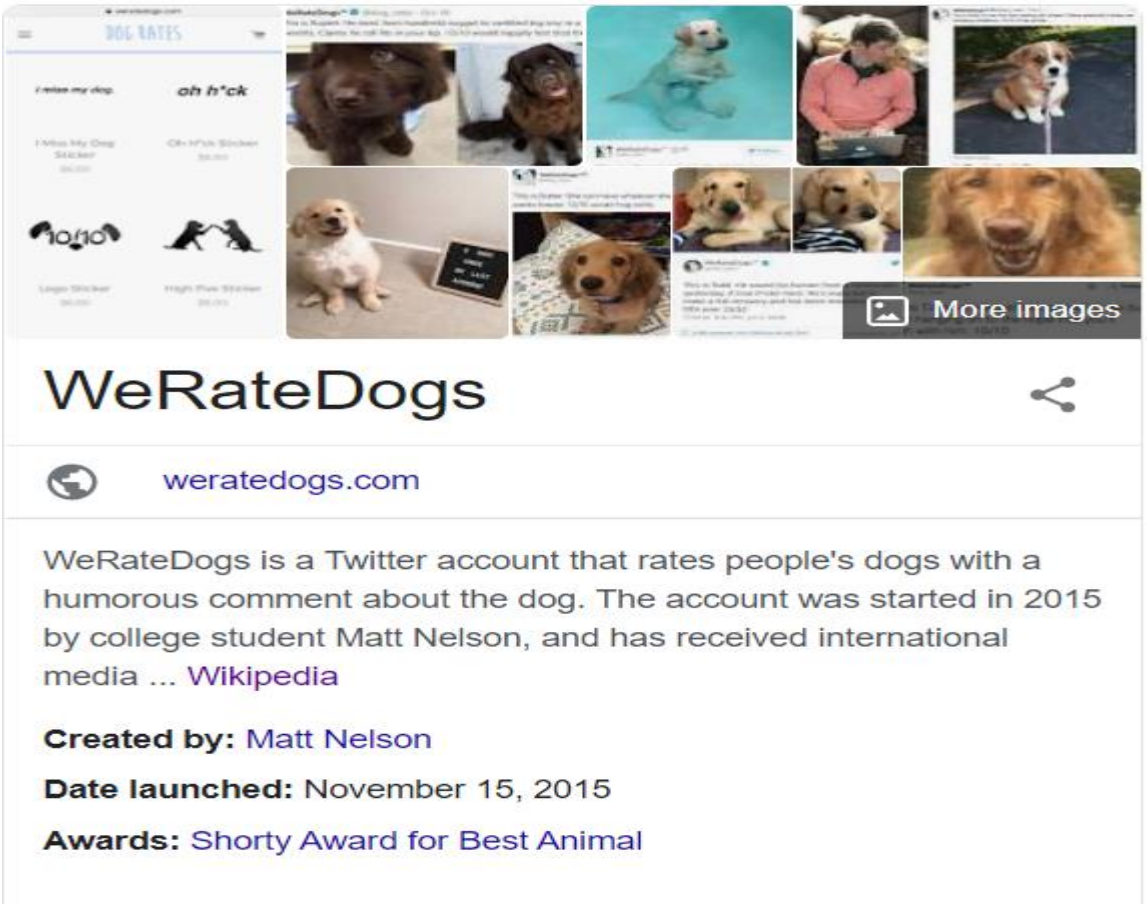
Data Analysis Professional Nanodegree Program

WRANGLE AND ANALYZE DATA



Introduction

Real-world data rarely comes clean. Using Python and its libraries, we will 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. we will document our wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

MATERIALS

1- The WeRateDogs Twitter archive (twitter-archive-enhanced.csv)

2- The tweet image predictions (image_predictions.tsv)

3-Twitter API (tweet_json.txt)

PROCEDURE

- Data wrangling, which consists of:
    - Gathering data
    - Assessing data
    - Cleaning data

DATA

- df1 represents (twitter-archive-enhanced.csv)

- df2 represents (image_predictions.tsv)

- df3 represents (tweet_json.txt)

Gathering data

DATA FRAME	TYPE	LIBRARY USED
df1	File on hand	Pandas
df2	Hosted on Udacity's servers	Requests
df3	Tweet's JSON data	Tweepy - Json

Assessing data

DATA FRAME	ISSUE	ISSUE TYPE
df1	-removing rows that have non-empty values in [in_reply_to_status_id - in_reply_to_user_id - retweeted_status_id - retweeted_status_user_id -retweeted_status_timestamp	Quality
	- Nan in [in_reply_to_status_id - in_reply_to_user_id - retweeted_status_id - retweeted_status_user_id - retweeted_status_timestamp] Which can't be estimated	Quality
	-‘timestamp’ column has +0000 in it.	Quality
	-‘source’ column is not clear.	Quality
	-'tweet_id' is an int not a string	Quality

	- 'timestamp' is a float not a datetime.	Quality
	- One variable in Four columns ['doggo', 'floofer', 'pupper', puppo] and each variable should form a column.	Tidiness
	extract the rating numerator from the text.	Quality
	Now we need to change the rating numerator data type into float.	Quality
df2	- We have Three prediction rates we can get the first one that will refer to True. Each variable forms a column.	Tidiness
	- We have Three breed predictions we can get the one who meets with the first prediction rate that refers to True.each variable forms a column.	Tidiness
	- Some problems with the breed predictions column, some names are capitalized and the other are in the lower case.	Quality
df3	- get just the needed columns.	Quality
Df1 & df2 & df3	- finally, we need to merge the Three data frames.	Tidiness

Cleaning data

We cleaned every mentioned issue with the proper code as shown in the notebook attached in the project file.

CONCLUSION

We need to wrangle our data for good outcomes, otherwise there could be consequences. If we analyze, visualize, or model our data before we wrangle it, our consequences could be making mistakes, missing out on cool insights, and wasting time. So best practices say wrangle. Always.

REFERENCES

1- Wikipedia

2- Data Wrangling course pages on Udacity