Data Wrangling Project

Data wrangling project is about gathering assesing and cleaning data of dogs rating and specially from WeRateDogs twitter page from different sources , which are :

CSV file and TSV file which were downloaded manually from udacity classroom

And tweepy to access the data in we rate dogs using twitter ids in the WeRateDogs twitter archive in csv then extract retweets and favourites and storing it to tweet_json.txt file .

Then reading all those data we gathered and exploring them manually and automatically using (describe , info , some times loops , value_counts .. etc)

After all of this i started in cleaning data which was some how hard part for me

I was first so nervous about how to start and from where , so i kept assessing then cleaning and looping through this process untill i put it all together .

The hardest part was how to extrect numerators and denominators from text and i finally did it phew , consequently started to extract all the data i noticed and cleaning them using different libraries , but the most common library which helped me alot was re , i found that it has different ways to deal with it like search , sub (substitute) , findall so it was so helpful in extracting , i tried to extract some data using sql from spark library using select form where like methods but unfortunately it didn't get the needed output with me so i just took step back and returned to using python only .

Finally i gathered all data in one data fram (dog_wrangle) and visualized some of it's columns using pandas and matplotlib

Conclusion

I learned alot about different libraries and methodes to use in python

Data wrangling is literally a great part in data analysis

Jupiter notebook can handle more than a programming language like sql for example which happen by importing it's libraries