INTRODUCTION

This project is the second project of the Udacity Data Analyst Nanodegree program and it primary focus on data wrangling. The data for the project is gotten from twitter WeRateDogsaccount. Python is the tool use for this project with some other libraries like pandas, matplotlib ,request and seaborn. the gathering and cleaning process is documented in a jupyter notebook.

The Data Gathering

Three different data which was later merge into one was used for the project The first data is the WeRateDogs twitter-archive-enhanced.csv which was provided by Udacity in csv format and contain the folowing columns;tweet_id,in_reply_to_status_id,in_reply_to_user_id,timestamp,source,text,retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp,expanded_urls,rating_numerator,rating_denominator,nam

Although not all the columns are valid as some have high percentage of null values,some columns need to be tidied , some are not important.

The second data is the image-predictions.tsv ,the url link was provided by Udacity was programmatically downloaded using request library, it contains the following columns;tweet_id,jpg_url,img_num,p1,p1_conf,p1_dog,p2,p2_conf,p2_dog,p3,p3_conf,p3_dog

The third data is from the twitter API but since i couldnt get developer access on twitter i use the json file provided by udacity which i read line by line into pandas dataframe,the line i read are;id,retweet_count,favorite_count because thats what i require

Data assessing

The data was asses visually by scrolling through each of it in excel and programmatically using pandas and methods such as .head(),.info() etc from assessing the data,some issues were discovrered from the dataset,the issues were seperated into tidyness issues and quality issues

Data Cleaning

The first and important step in data cleaning is making a copy of the original dataset ,its the copied dataset that the cleaning is performed on so that the original data remain intact

The datasets were cleaned in three steps which are the

define;here the solution to the problem is defineed

code; the code to effect the cleaning is done at this stage

test; the issue is re-assesed to see if its still there

one of the cleaning challenge i had is combining the pupper,floofer,pupperer and puppo column into one

Conclusion

Data wrangling is an important and useful skill in data analysis because it can be use to extract data from different source and it better for big data compared to excel

In [ ]: