WRANGLE REPORT

FOR PROJECT

" WRANGLE AND ANALYZE DATA "

_This project is one of Udasty's projects in the field of data analysis…..

Prepared by : shuaib elamrity

Date : 4 june 2021

Introduction:

Data wrangling is the process of cleaning and unifying messy and complex **data** sets for easy access and analysis. ... This process typically includes manually converting and mapping **data** from one raw form into another format to allow for more convenient consumption and organization of the **data**.

For this project :

Using Python and its libraries, I 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. I will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL

Project Details :

The tasks of this project as follows :

1. **Gathering data**
2. **Assessing data**
3. **Cleaning data**

1.Gathering data :

The data for this project consist on three different dataset that were obtain as following:

1. **Twitter archive file:** the twitter archive enhanced.csv was provided by Udacity.

2. **Twitter API &JSON :** by using the tweet ids in the weratedods twitter archives , and python's tweepy library to gather each tweet's retweet count and favorite count

3. The tweet image predictions :, i.e., what breed of dog is present in each tweet according to a neural network

This file was provided to udacity student

2.ASSESSING DATA:

-After gathering each of the above pieces of data,I will assess them visually and programmatically for quality and tidiness issues

_ Tidiness issues

1. convert timestamp to datetime and have a columns for year and moth
2. columns [doggo , floofer , pupper , puppo] must be one columns name stage
3. drop some rows for 'retweet' and 'replay_to_state' to not may skew the analysis
4. remove some columns that will not be needed for analysis [in_reply_to_status_id ,in_reply_to_user_id , source , ...]
5. there some rows need to remove Because the picture does not belong to a dog

-Quality issues

1. Some columns need to change the data type like [tweet_id , timestamp ,stage]
2. Unify the denominator in the rating to become 10 and create a new column containing one number for the rating
3. rename text column to be tweet
4. column name have a name "a" need to change and convert Nane to np.nan .
5. there some rows need to remove Because the picture does not belong to a dog
6. change the type data for tweet id to str
7. In dog names, some names start with a capital letter and some start with a lowercase letter
8. There a duplicated on jpg_url column ..This column must not contain duplicate rows

Libraries that are used in the assessing and cleaning processes

1. **Pandas**
2. **Numpy**
3. **Requests**
4. **Json**
5. **tweepy**

3.Cleaning Data :

This part of the data wrangling divided in three parts "Define Code Test " these three steps were on each issues described In the assess section

First step always is create a copy of the three original of dataframe If there was an error , I could create a new copy from the original