DATA WRANGLING

Introduction

In this project (Data Wrangling) we start to do Data Gathering then we assets the Data in Data Assessing stage then the Data Cleaning to clean the data to be readable and useful after doing this the dataset will be easy to access and easy to analysis, then finally we stored it in CSV file.

Data Gathering:

We collect Data from three resources:

1\ WeRateDoges Twitter Archive that stored in (twitter-archive-enhanced.csv)

2\ The image-predictions.tsv' that we retrieve it from URL

3\ The tweet-json.txt that upload it from the Udacity site

Library Used in project:

```
import pandas as pd
import numpy as np
import requests
import os
import io
import time
import csv
import json
import os
import os.path
import re
import itertools
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Resource Steps:

Import the CSV File:

twitter_archive_df was create to store the data from twitter-archive-enhanced.csv using pandas (pd.read_csv)

Download File from Udacity Server :

Using (requests) to read the data from the link

(url='https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv') then stored it in img_predictions_df

Import json File:

json_df was create to store the data from tweet-json.txt using pandas (pd.read_json).

Data Assessing:

We list the:

- Data quality issue:

  1 some tweets doesn't contain rating we need to Drop them
  2 sources was difficult to read
  3 jason file contain many columns that unnecessary
  4 need to rename id to tweet id in jason file
  5 there is Incorrect dog names
  6 contain NaN values
  7 timestamp has not correct datatype
  8 dog names contain a and an

- Data tidiness issue.

  1 We don't need to use every column in json_df file
  2 there is 4 columns of dog staegs need to merge them in one

Data Cleaning:

Data cleaning in this stage we start to solve the issue that we found in Data Assessing stage.

There is some function we used in this project :

.timestamp

.loc

.name.str.islower()

.value_counts()

.drop()

.value_counts()

.astype()

.merge()

Data Stored:

In this stage Data Stored is the final step we took the cleaned data then saved it in twitter_archive_master.csv using the function (.to_csv).