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 tweepy

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). We download it from Udacity server because the tweeter server is not work with me.

Data Assessing:

- ## Quality Issues

Issue 1 : some tweets doesn't contain rating we need to Drop them

Issue 2 : Dataset contains retweets

Issue 3 : jason file contain many columns that unnecessary

Issue 4 : contain characters after '&'

Issue 5 : there is Incorrect dog names

Issue 6 : contain NaN values

Issue 7 : timestamp has not correct datatype

Issue 8 : dog names contain a and an

- ## Tidiness Issues

Issue 1 : Merge the datasets

Issue 2 : We don't need to use every coloumn in json_df file

Issue 3 : 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).