Wrangle Report

Introduction:

This Report summarises for part wrangle on the WeRateDogs Twitter Analysis. WeRateDogs is a Twitter account that rates people`s dogs with humorous comment about the dog .The three steps of the wrangle report are as follows:
Gathering , Assessing ,Cleaning.

Gathering Data:

Data were collected from three different sourctes .

1-witter-archive-enhanced.csv file was provided by Udacity , downloaded manually then was loaded and read using pandas data frame.

2- image-predictions.tsv file is hosted on Udacity servers,was extracted programmatically using python request library f

3- Twitter API file contains tweet id , favorite count and retweet count .Data was provided by Udacity, Downloaded manually then was loaded from the tweet-json.txt

Assessing Data:

After gathering the data , I assessed the data visually and programmatically . for assess the data I used various methods Such as info(), head(), sample(),value_counts(). I was able to spot 10 quality issues and 2 tidiness issues listed below:

Quality and Tidiness issues:

1 timestamp column must be datetime type instead of the object

2- there is tweet not original which should not be present for analysis

3 having None string in columns (doggo,floofer,pupper,puppo) instead 'NaN'

4 dog stages are categorized over 4 columns and should be in one column. This is tidiness issue

5 some columns we do not need and should not be in the master dataset

6 there is invalid dog name such : a,o,the,an

7 some rating numerator less than 10 and denominator not equal 10

8 there is source column html link should be change to actual source

9 p1 ,p2, p3 column names are not clear should be change to meaningful names

10 there is some value in the p_1 dog ,p2_dog,p3_dog columns not actual dog breeds

11 image_predictions table and df_ API table should be combine with df_enhanced table to make one clean dataset.This is Tidiness issue

12 tweet id column must be str type instead of the int type

Cleaning Data:

Data cleaning process consists three step: Define,Code and Test . was cleaned each of the issues documented while assessing , the first step was fixed is convert timestamp to datetime type using to _datetime() method, then fixed tweet id column by drop retweets that no need it , also extact to columns (doggo,floofer,pupper,puppo) into new stage column, Drop some columns that not be present for analysis, also cleaned dog name column from invalid name using islower(), change p1, p2,p3 name columns by rename() to make more clear , replace html link in source column to actual source ,also some other issues was cleaned to get cleaning dataset that we can analysis and extact insights .