Report on Data Wrangling Steps

Wrangling Report

Wrangling process is applied on this project on a data set from twitter archive of account known as WERATEDOGS.WERATEDOGS rates people's dog by making comment about the dog.

The wrangling project goals included:

- Wrangling data from following steps:
  - Gathering Data
  - Assessing Data
  - Cleaning Data
- Sorting, Analyzing and Visualization the cleaned data.
- Report on the data wrangling efforts ,data analysis and visualization

Gathering Data:

The project gathered data from 3 different sources:

- Twitter archive " Tiwtter-archive-enhanced.cs" was provided by Udacity. This file contain basic data about tweets itself like tweet id , timestamp, text, etc. this file contains about 5000 tweets of their tweet.
- Image prediction file, , this data was downloaded Programmatically from file provided by Udacity.it contain data like what breed of dog based on neural network.
- Tweet-Json to gather each tweet retweet count and favorite count and additional data.

Assessing Data:

After gathering data I started to detect things should be cleaned and edited in both way visually and programmatically for both quality and tidiness issues.

Quality Issues:

- **Completeness**:
  Missing data in many columns.

- **Validity:**
  Name column has invalid data for dogs like (a,an,the).
- **Accuracy:**
  Timestamp represented as object.

Tidiness Issues:

- In twitter archive enhanced puuper,floofer,doggo,puppo should merge in one table.
- Merge twitter archive enhanced,image prediction,json in one table that represent all data about tweets, dog and rating.
- Drop unwanted columns.

Cleaning Data:

It was done through 3 stages:

- Define: Determine what need to clean and how.
- Code: Apply what was determined and it should make Programmatically to clean code.
- Test: Ensure that the data set is cleaned.

I cleaned the following issues:

- **Quality Issues:**
  - there are many columns have instuitable datatype like (time_stamp,p1_conf, p1, p1_dog).
  - Nan values should be replace with ''None''.
  - clean source column
  - extract numerator and denominator from text col
  - column name has invalid name like(a-an-all-etc)
  - numerator and denominator column has invalid number like(152,8,etc) should be fixed.
  - keep orginal tweets
  - drop duplicated rows.

- **Tidiness Issues:**
  - combine 3 sources in one dataframe.
  - drop unwanted columns like(p2,p3,....)
  - combine puppo , floofer ,doggo and pupper in one column