Wrangle Report

1. Introduction

The purpose of this project is to apply the data wrangling of datasets that given in the project overview. The wrangled data is WeRateDogs. In order to apply wrangling first, I had to apply the data wrangling steps which are described in this report.

2. Data wrangling steps:

- gather
- assess
- clean

a. gather:

The data of this project consists of three datasets which are as following:

Twitter archive enhanced file: it is csv file. The download process for this file was done manually from Udacity classroom. I read this file as csv file using pandas' library.

Image predications file: it is tsv file. The download process was done manually from Udacity classroom. I read this file as csv file using pandas' library but as tab separator.

Tweet json file: it is txt file. Since the request had been rejected the download process was done manually from Udacity classroom. The file format was in json format and I read this file as json form using pandas' library. The lines=True was used in parameter to read each line as json not the whole file as json.

b. assess:

When the gathering process was done, the three datasets were available and ready to process. In order to access the datasets and check entire contents, visual assessment and the programmatic ways were used. In this process many methods were used like:

datasetName	=> to show all contents
datasetName.head()	=> to show only first 5 rows

datasetName.shape() =>to check the number of rows and columns

datasetName.info() =>columns' names and datatypes of each column

datasetName.column.duplicated()=> to check the duplicate in column

datasetName.column.value_counts()=> show the values with counts.

At the end of assessing process, the issues were easily separated as quality and tidiness issues. The type of issue name was mentioned above the code.

c. clean:

After the assessing process, the cleaning process was started. Eight quality issues and 2 tidiness issues were found.

Quality issues like changing the datatypes, removing the < a > tag from column, replacing 'None', 'a', 'an' and unnecessarily data, separating the result of combination of columns, deleting columns, checking for duplicate, removing the brackets and 0 for text range column and renaming columns. All the cleaning parts are done with test functions to make sure the cleaning codes are done correctly.

For the tidiness issues, the three dog stages were combined in one column and three datasets merge into one master dataset.

Generally, I found solving the tidiness issues is more trick than the quality issues.

3.Storing the mastered dataset

After cleaning process, the cleaned dataset was stored as new csv form.