Wrangle Report

Introduction:

- The Wrangle and Analyse project are assigned by the Udacity’s Data Analyst Nanodegree. Here I am going to implement what I have learned so far in the Data Wrangling part. The data used for this project is twitter’s WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage. Where dogs have been rated greater than 10/10.

➢ Gathering Data:

Gathering data is the most import part of the Data wrangling.

Here in this project we have gathered three different sources:

1. Twitter-archived-enchanced.csv file.
2. Image-predictions.tsv file.
3. Twitter-json.txt file.
4. Twitter-json.csv file.

Twitter-archived-enchanced.csv file:

This file was downloaded from the Udacity portal by clicking on the link provided by the Udacity.

Image-predictions.tsv file:

This file was programmatically downloaded from the Udacity portal using requests library.

Twitter-json.txt file:

This file was programmatically downloaded from the Udacity portal using requests library. For my project I did not use the Twitter Api because I came out through many errors and problems.

Twitter-json.csv file:

This file consists of data that was extracted from the file called twitter-json.txt programmatically. This is the file I used for cleaning and merging the data with image-predictions.tsv file.

➢ Accessing:

The Accessing is the next step of the Data Wrangling. Because here we will be accessing the data that we gathered, finding out the missing data, finding the null values, describing the data, finding out the duplicate values etc.

The steps I used for the Accessing part is:

1. Accessing all the datasets that we gathered.
2. Head () used to view the first 5 rows of the data.
3. Shape used to view the number of rows and columns in the data.
4. Info () is used to know the data types of each column and viewing the number of non-null values.
5. Isnull () is used to views the number of null values in the dataset.
6. Duplicated () used to find the duplicate values in the dataset.

And many other python methods are used to access the dataset.

During accessing data there were some interesting errors found from the data. The errors include quality and tidiness.

After many coding practice the errors were solved.

➢ Cleaning:

The final step for the Data Wrangling is the Cleaning process. In the cleaning process we will solve the problems that were faced from the dataset.

The issues found during the assessment process were cleaned and tested.

The steps that I took in the cleaning process:

1. Making a copy of original data.
2. For every issue I have divided into three steps Define, Code and Test.
3. Finally, I merged the tweet_json_clean and image_predict_clean dataset into one data set called twitter-archive-master.csv file.

➢ Conclusion:

The data doesn’t come from only one source, actually the data are scraped from many sources with different formats. This project thought me the whole process of data wrangling as it helped out to understand how a data is and how to work on it.