Wrangle Report

In this project, we mainly focused on data wrangling part. This data wrangling process was divided in 3 parts:

- Data Gather
- Data Assess
- Data Clean

In the data gathering part, which is the first step in data wrangling, we have collected data from 3 different sources:

- **Twitter_archive_data** – directly downloaded from website in csv format.
- **Image_predictions_data** – used Python's Requests library to download that programmatically in tsv format.
- **Tweet_json.txt** – created twitter account and setup Developer portal to generate API access, secrete keys. Stored the twitter json data in txt format with help of tweepy.

In the later part of gather phase, loaded these files using pandas and inspected the data to make sure it has loaded successfully.

In the next part, the focus was on to assess the data visually and programmatically. The pandas functions like head (), tail () were used for visual assessment and functions like info () and describe () were used for programmatic assessment. Pandas duplicated function also helped to find out duplicate data. This data assessment helped in finding out the quality and tidiness issues in all 3 datasets. The idea here was to find out quality and tidiness issues, such as, non-descriptive column headers, missing values, messy data in these datasets. These issues then cleaned out in next phase, cleaning the data.

All the quality and tidiness issues listed in assess phase were addressed in the cleaning data phase. The programmatic data cleaning process was defined here for cleaning purpose. The process was divided into define, code and test phase.

- **Define** – converted the assessments into defined cleaning tasks.
- **Code** – converted the above definitions to code and ran that code.
- **Test** – tested the dataset to make sure the cleaning worked correctly.

The reassessment and iteration were performed after the cleaning process to crosscheck the progress made on data wrangling steps.

The summary of the data wrangling part for this project is as below:

Gathering

- Explored the data sources in various formats (JSON, web scraping and HTML, APIs)
- Evaluated the structure of each file format

- Converted data files using Python and Python libraries

Assessing

- Visual and programmatic assessment were performed to view specific portions and summaries of the data
- Addressed the issues with content as quality
- Addressed the issues with structure as tidiness

Cleaning

- A systematic process was designed to perform cleaning operation on issues addressed in assessing phase.