Data Wrangling Report

Executive Summary

This document outlines the steps that have been undertaken in the data wrangling process for project 5 of Udacity's Data Analytics Nanodegree.

Data Wrangling

The data wrangling comprised three steps, gathering, assessing and cleaning. This was an iterative process. Further assessing and cleaning one analysis was started. The steps undertaken are described in the following sections.

Data Gathering

The data from this project came in three parts. The first of which was provided as a CSV file. This was manually downloaded and then loaded into a pandas dataframe using the '*pd.read_csv'* function.

The second file, a .tsv was programmatically downloaded from a given url and read into a pandas dataframe using the function mentioned above.

The third dataset came from querying Twitters tweet API. Taking the Tweet IDs from the csv file provided, I queried the API to download the stored data in JSON format to a .txt file. This file was then read and converted into a pandas dataframe ready for the assessing process.

Data Assessing

The data assessing process was undertaken in two stages. Firstly the data was visually assessed within MS Excel. Any obvious problems were recorded to be addressed in the following section.

Secondly, programmatic assessment was carried out. In this stage the following add-ins within python were used. Similarly, data quality issues and data tidiness issues were recorded to be cleaned in the following section.

- .info()
- .describe()
- .value_counts()

Data Cleaning

To complete the data cleaning process I worked through each of the quality or tidiness issues recorded in the data assessing section one by one. When cleaning each item I followed the *Define, Code, Test* structure. First documenting the intended way of addressing the issue, the writing the code followed before finally testing.

Data Analysis

Within the data analysis stage further data wrangling, in particular the later stages, assessing & cleaning were undertaken were appropriate.