Data Wrangling Report

This report briefs about the data wrangling efforts involved in completing WeRateDogs project as a part of Udacity Data Analyst Nanodegree assignment. Data Wrangling typically involves below steps:

- Data Gathering
- Assessing the data
- Data Cleaning
- Storing Data

Our assignment majorly revolves around the steps mentioned above. We'll go through in breief what are the steps in each of theses phases and what we have done in them.

1. Data Gathering

Gathering data for this project involves fetching data from three different sources. Complete data for our analysis is available in 3 parts from below sources.

- First file called 'twitter_archive_enhanced.csv' consists of 2356 records of data. This file has to be downloaded manually and then has to be read using pandas read_csv function.
- Second file needs to be downloaded programmatically using python's 'request' library by hitting the specified url. This file consists of image predictions of the breed of the dog coming from a neural network on some of the tweets downloaded from the archive file. File downloaded will be in '.tsv' format and needs to be handled using pandas 'read_csv' function with 'sep' (separator) attribute set to '\t'.
- Third file is a bit tough to create. This file can't be downloaded directly. We'll need to be query json content from Twitter's API using a python library called 'Tweepy'. For this we need to obtain a few keys by creating an application in Twitter's developer account. Then we have to pass these keys to the Twitter API and fetch json content for each tweet and store them to a file. This file isn't in a proper format, so, we'll need to create a csv from this json file. This file consists of information about retweet counts and favourite count for all the tweets.

2. Assessing the Data

Three files created in Data Gathering section are have been looked manually using excel workbook. Also, they have been looked programmatically using Jupyter Notebook which allows us to do a few investigations about the data. Sample functions used in programmatic assessment were info(), describe(), sample(), head(), value_counts() functions etc.

Data in the three datasets were assessed for 2 kinds of issues:

```
- Quality Issues
- Tidiness Issues
```

Quality Issues refers to the issues observed with the content in the datasets. This category of issues is also called as dirty data. Quality Issues majorly checks for below criteria.

```
- Completeness
- Validity
- Accuracy
- Consistency
```

8 issues (missing data in tables, invalid datatypes for a few columns, inaccurate data in 'ratings', 'source' columns etc) around these criteria were observed and documented in 'Assess Data' section of 'wrangle_act.ipnyb' file.

Tidiness Issues refers to the issues with structure of how the data is organized in the table. It basically looks for the following criteria.

```
- Each observation forms a record
- Each variable should form a column
- Each type of observational unit forms a row
```

4 tidiness issues were identified (i.e unpivoting columns, removing unneeded columns, merging tables) and documented in 'Assess Data' section of 'wrangle_act.ipnyb' file.

3. Data Cleaning

This is the final step in Data Wrangling process where the issues documented in 'Assessing the Data' section are actually acted upon. It consists of three phases.

```
- Define - Tells how you are going to resolve the issue
- Code - Where you actually resolve the issue by writing code
- Test - Write some code for programmatic or visual assessment to check if the iss
ue is resolved
```

All the quality and tidiness issues observed in previous phase are fixed in this phase so that we'll have a clean dataset which helps us in making best analysis. I have clearly mentioned my thoughts on how i am going to fix the issue in 'Define' section, written proper comments in 'Code' and 'Test' phases for the code snippets wherever necessary. All the issues were fixed programmatically using the 'pandas' library. In the end, all the three datasets were merged into a single table following the Tidiness rules.

4. Storing Data

Why do we need to store data?

This phase is about storing the clean data we just worked upon. Output of the 'Data Cleaning' phase in this project is a single table that helps us in analysis. If in case you want to analyze things, you shouldn't be be running all the code of "Data Gathering", "Assessing the Data" and "Data Cleaning" phases. Instead, your starting point should be cleaned data. So, for this purpose, we sore the output of 'Data Cleaning' phase to a file which can be read directly whenever we want to perform analysis.

How did i store data?

I have used pandas "to_csv()" function to store the dataframe to a csv file.

In this project, we have stored this cleaned table into a file named 'twitter_archive_master.csv'. This file can now be shared with others so that anyone can perform analysis on the cleaned dataset.