Data Wrangling:

Data wrangling is a three-stage process of gathering the data, assessing the quality and structure of the data and cleaning the data.

Problems with Data:

1. Missing data
2. Duplicate data
3. Inaccurate data
4. Problems related to its structure.

We carry out data wrangling in order to clean up the existing inconsistencies or inaccuracies that occur during the sample gathering process by people who record inaccurate data, simply miss recording required information or just make mistakes during recording.

Introduction to the Project:

The dataset that we have is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. We need to clean this data in order to get some meaningful insights.

Gathering of Data:

We have the following datasets for our disposal after gathering process:

1. twitter_archived_enhanced.csv

We have downloaded the above file manually.

2. tweet_json.txt is constructed via twitter API

For this file we have gathered it by twitter API.

We only have columns in this dataset that are relevant to our analysis.

We ignore the other columns.

Due to some problems related to API, I have directly used this file given by udacity.

3. image_predictions.tsv is the file we have downloaded programmatically.

This file is downloaded from a url given to us by using pythons request library.

Assessing the Data:

We assess the data in order to maintain its quality so that it does not hinder our data analysis process and we can also get some inaccurate data analysis results due to it.

We assess the datasets because these datasets may have the following problems:

1. Missing data
2. Inconsistent data
3. Inaccurate data
4. Invalid data

The above are called as Quality issues.

Other the Quality issues we may have some structural issues also known as messy data.

Structural issues are more related to the structure of the tables in terms of its columns and rows and any multiple types of observation found in same table.

The process of resolving the structural issues is also known as tidying the data.

The following should be taken account in tidying of the data:

1. Column headers are values, not variable names.
2. Multiple variables are stored in one column.
3. Variables are stored in both rows and columns.
4. Multiple types of observational units are stored in the same table.
5. A single observational unit is stored in multiple tables.

How do we assess the quality and structural issues?

We assess the above by doing visual assessment and programmatic assessment in pandas.

Some common problems we can find during assessment are duplicate values, null values, inaccurate names, inaccurate data types, multiple columns for a particular group, multiple tables, not required data and columns.

In the current project we found the following issues in the data sets:

Quality Issues:

Images Table:

- tweet_id is a float not a string.

- p1,p2,p3 are not in title case.

Tweets Table:

- tweet_id is a float not a string.

Tweeter archive enhanced table:

- tweet_id is a float not a string.

- timestamp is a string not datetime datatype.

- retweeted_status_timestamp is a string not datetime datatype.

- Inaccurate dog name called 'a' found. This is not a standard name.

- Dog name have None. Need to replace this by NaN. Lower case names are not proper nouns. That should be replaced by NaN

- We only want original ratings. Retweets have no ratings. So not null data in retweeted_status_user_id not required.

- in retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp are to be removed.

- source column has unnecessary tags.

- in_reply_to_status_id,in_reply_to_user_id are floats. Also they are not required.

Tidiness Issues:

- doggo,floofer,pupper,puppo should be clubbed together in one column. Remove the original four columns.

- Join tables together to tidy data(To satisfy one of the condition of tidiness(A single observational unit is stored in multiple tables).

Cleaning of the data:

The final stage of data wrangling is the cleaning of the data.

After specifying the quality and tidiness issues we start correcting these defects by cleaning the data.

We do this by using different pandas methods.

After cleaning the data we perform the necessary testing in order to verify if the data is cleaned or not.

Summary:

Data wrangling is a very important part in Data Analysis. Meaningful insights can only be derived by having proper data wrangling. After cleaning and testing the issues we found we have a dataset on which we can perform data analysis and derive some conclusions that can be useful for businesses or individuals.