WRANGLE REPORT

INTRODUCTION:

- The goal of this project is to gather the data from different sources and formats i.e,csv,tsv,url and json files.
- After gathering,we need to assess it's quality and tidiness and then clean it.
- Later we need to analyse and visualise the cleaned data.

GATHERING DATA:

Data is gathered from three sources:

- twitter_archive_enchanced.csv,is downloaded manually.
- image_predictions.tsv is downloaded programmatically using Requests library from Udacity's server.
- Additional data is gathered from Twitter API and is stored in json.txt file.

ASSESSING DATA

- In this step we assess the data to find the tidiness and quality of the data.
- Dirty data/low quality data have quality issues.
- Messy data/untidy data have structural issues.
- There are two types of assessment:
  1.Visual assessment:opening the csv file in excel or google sheets and scrolling the data.
  2.Programmatic assessment:we use functions such as head,tail,value_counts,sample,info,describe methods by importing pandas library.

QUALITY ISSUES:

- tweed_id has the incorrect data type.
- timestamp has the incorrect data type.
- Name column contains inaccurate name likes a,the,by etc.
- The name O'Malley was incorrectly represented as O.
- Unnecessary columns are present.
- Inappropriate rating numerators and rating denominators are present.

TIDINESS ISSUES:

- Merging all dataframes into a single dataframe.
- Combining four variables of dog type into a single column.

CLEANING DATA:

- Copy of each dataframe is stored in an other dataframe.
- For each quality/tidiness issue,perform programmatic data cleaning in three steps-Define,Code and test.
- We use functions like drop,merge,islower,astype,replace etc.
- The cleaned data frames are merged into a single dataframe.

STORING DATA:

- After the completion of cleaning process,the dataframe is stored in twitter_archive_master.csv.

CONCLUSION:

When performing data analysis the data is rarely tidy.

- In this case if you analyze the data without proper wrangling then, your results are going to be incorrect.
- In this project using Python and its libraries.
- We gathered data from different sources and cleared all its quality and tidiness issues.

.