WeRateDog Data wrangling project

INTRODUCTION :

Data wrangling is about gathering the right pieces of data, assessing your data's quality and structure, then modifying your data to make it clean.

Data wrangling process consists of three main steps :

1.Gathering data

2.Assessing data

3.Cleaning data

In this project, I will gather, assess, and clean data of @dog_rates account’s tweet archive of Twitter user , then act on it through analysis, visualization and/or modeling .

DATA GATHERING :

Gathering Data for this Project is Obtained from 3 different sources

1.Getting data from an existing file (twitter-archive-enhanced.csv) Reading from csv file using pandas

2.Downloading a file from the internet (image-predictions.tsv) Downloading file using requests

3.Querying an API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy

Then importing these data into our programming environment (Jupyter Notebook)

DATA ASSESSING :

In the assessing step I went through the data to check if it follows these rules below :

1- Quality: issues with content. Low quality data is also known as dirty data.

2- Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:
   - Each variable forms a column.
   - Each observation forms a row.
   - Each type of observational unit forms a table.

I used two types of Assessment, manually using MS- Excel and programmatically using python .

DATA CLEANING :

There are two types of cleaning:
1-Manual (not recommended unless the issues are single occurrences and that type I didn't use.)
2-Programmatic(which I used in my cleaning process)

After I identified the quality and tidiness issues in the assessment step, I should fix them by coding in form of DEFINE ..CODE ..TEST

CONCLUSION

Data is a very precious things that we could wrangle and discover very Interesting insights and trends , visualize them to help the top management in the decision making process.