Wrangle Report

By: Lamia Alshawi

Introduction:

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

Tools used:

- Jupyter notebook
- pandas
- NumPy
- requests
- json
- Excel

The Wrangling Process:

Gathering data:

- Enhanced Twitter Archive, which was provided for me as a .csv file
- Tweet Image Predictions I downloaded this data as a .tsv from a URL with the Requests library
- WeRateDogs Twitter Archive, I used the json-tweet zip file that was provided

Assessing data:

- After gathering the data, I visually assessed the three data frames listed above for any issues. Then, I assessed programmatically in pandas to find any further issues.
- Each table was assessed initially using the describe(), sample() and info() commands to view a sample of the data and the structure of the table.
- value_counts() also helped in visually assessing the data.
- I ended up mostly looking through the data in excel to get a better picture.
- On excel looked through and filtered the data.
- Issues found included:
  - Missing data in some tables
  - Duplicated data
  - Inaccurate data types
  - Unnecessary columns

Cleaning data:

A clean copy of each table was created at the beginning of the cleaning stage to have a separate data set to work with and keep the originals.

The programmatic data cleaning process:

- Define: convert the assessments into defined cleaning tasks.
- Code: convert those definitions to code and run that code.
- Test: test the dataset, visually or with code, to make sure the cleaning operations worked.

Cleaning of the data included:

- modifying values using pandas functions,
- merging tables

Finally, the cleaned data was saved into a file called "twitter_archive_master.csv".