Wrangle Report

By
Mohamed Sebaie Sebaie Youssef

This report describes the wrangling efforts involved in completing the "WeRateDogs" project as a part of Udacity's Data Analysis Nanodegree.

The Data Wrangling process consists of:

1. Gathering the data
2. Assessing the data
3. Cleaning the data

1. Gathering

We gather the data from three different Sources:

I. Twitter Archive as a given csv file
II. Tweet Image Predictions from a given URL
III. Twitter API but the ids are expired and provides errors, So I used the given file (tweet-jason.txt)

- **Twitter Archive** is the first dataset that's loaded from csv file (twitter_archive_enhanced.csv) from the Udacity Materials and after loading as DataFrame it is consists from 2356 Rows and 17 Columns.

- **Tweet Image Predictions** is the second dataset as it is downloaded from url then loaded as DataFrame .

- **Twitter API** is the final dataset but by using tweepy library and the code provided in the materials, the code gives errors for all tweet ids though I freez the cell as comment.

These are the steps I folowed to create the tweet data:

1. I used the given file (tweet-json.txt) because the code of Tweepy Api gives errors as the ids are expired.
2. I read the (tweet-json.txt) and selected the required date ('id', 'favorite_count', 'retweet_count').
3. I created dictionary then convert it to pandas DataFrame and Save a txt and csv copies.

2. Assessing

- I assessed the data **visually** and **programmatically** to identify:
  - Data quality issues
  - Tidiness issues

Visual Assessment:

By looking through the data in Jupyter Notebook and the Excel

Programmatica Assessment:

By using Methods and Functions to summarize the data like **df.info()**, **df.sample()**, **df.describe()**, **df.column.value_counts().**

The datasets were accessed by quick scanning through the rows and use of filters to identify areas for more detailed investigation and programmaticly inside jupyter with pandas using the functions.

The assessment is under two criteria, **quality and tidiness**. Then any issue I defined, I documented for the following cleaning step.

Quality refers to issues related to the content of the data and is also known **as dirty data**. Data quality dimensions help guide your thought process while assessing and also cleaning. The four main data quality dimensions are: completeness, validity, accuracy, and consistency of the data were used to identify quality issues. These issues were listed in the assessment section of the "Wrangle_Act.ipynb" Jupyter Notebook.

Tidiness refers to issues related to the structure of the data and is also known **as messy data**. Tidy data requirements: each variable forms a column, each observation forms a row and each type of observational unit forms a table. The issues that make the 3 dataset untidy are listed in the "Wrangle_Act.ipynb" Jupyter Notebook.

3. Cleaning

The final step in the data wrangling process is cleaning the data for quality and tidiness issues.
The programmatic data cleaning process:
Define: convert our assessments to defined cleaning tasks.
Code: convert those definitions into code and run the code.
Test: test our dataset, visually or with code, to make sure our cleaning operations worked.

Most of the cleaning process was performed using programmatic tools, like pandas functions (merge, melt etc) or def functions .

Analyzing the Data

Through charts and by answering insight questions as we are going to see within the `act_report.`

Conclusion

Data wrangling process provides a clean DataFrame for future analysis and visualization, Though we sorted our clean data to "twitter_archive_master.csv'. This file could even be shared with others without having to wrangle the data and head to the subsequent step direct.