Wrangle and Analysis Data Project

Wrangle Report by Amani Alharbi

Introduction:

This is the project 4 report for Wrangle and Analysis part which is the part 4 of Udacity data Analyst Nanodegree. This project is about wrangling a real-world data which is rarely cleaning by using Python libraries.
The requirements of project are:

- Gathering data
- Assessing data
- Cleaning data
- Storing data
- Visualization and analysis Data
- Wrangle report
- Act report

This report is part of requirements which explain the wrangle process **Gathering, Assessing, Cleaning and Storing.**

Gathering Data

In this part, I have to gather dataset from three resources.

First, the twitter archive enhances csv file.

I download manually and import it by using Pandas library. The twitter archive enhance file includes 17 columns and 2356 records.

Second, the image prediction tsv file. I need to download it programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv . The image predictive file includes 12 columns and 2075 records.

Third, API Twitter & JSON file.

I use twitter API to write a JSON text file by python library(tweepy). Then, I store the data in json_text file. After that, I read json file and query (id, full_text, favorite_count, retweet_count, retweeted). The json_text file includes 6 columns and 2354 records.

Assessing Data

I assess data for quality and tidiness issues in tow steps:

1. Visual Assessing:
   I printed the three datasets in Jupyter Notebook and look at records. Also, I open the Excel files.
2. Programmatically Assessing:
   I use the pandas to find most quality and tidiness issues like (value_counts, info, describe, head, isnull, duplicated)

The issues that I find are:

Tidiness¶

- unused columns
- merge "doggo", "floofer", "pupper", "puppo" to one column
- merge 3 tables to one dataset

Quality¶

- tweet_id datatype
- unacceptable dogs names in tweet_df like "a", "None", "the", "just"
- lower case for dogs names in tweet_df
- None in "doggo", "floofer", "pupper", "puppo"
- Incorrect values in rating numerators
- lower case name in "p1", "p2", and "p3" in image_prediction
- underscores in in "p1", "p2", and "p3" in image_prediction
- missing data
- text column has retweets

Cleaning Data

First, I made a copy for each table. then, I cleaned each issue by using pandas. More one than issue clean in one step like (missing data and merge datasets). After merging three tables I deal with one dataset instead of three.

Storing

I stored the dataset after cleaning to twitter_archive_master.csv file.

Resources

- https://classroom.udacity.com/nanodegrees/nd002-connect/parts/f3fb7339-2000-47dd-a21e-29480cdd8166/modules/14d9f5f1-9e7b-4bfb-97f3-bcdbf4a3699c/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/5919f3b1-899f-4295-80f1-17f091eb4df6
- https://classroom.udacity.com/nanodegrees/nd002-connect/parts/f3fb7339-2000-47dd-a21e-29480cdd8166/modules/14d9f5f1-9e7b-4bfb-97f3-bcdbf4a3699c/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/28d4643b-3785-4700-bdee-4e5fc9963576
- https://www.datacamp.com/community/tutorials/json-data-python
- https://stackabuse.com/reading-and-writing-json-to-a-file-in-python/
- https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html
- https://s3.amazonaws.com/video.udacity-data.com/topher/2018/November/5be5fb4c_twitter-api/twitter-api.py
- https://www.geeksforgeeks.org/python-startswith-endswidth-function/
- https://www.datacamp.com/community/blog/python-pandas-cheat-sheet