Project #5: Wrangle and Analyze Data

Date: Oct26-2019
Name: Sami Adham
Program: Data Analysis Nanodegree

Project Task:

The project is to show abilities to wrangling and analyzing. My tasks in this project are as follows:

- Data wrangling, which consists of:
  - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

Goals:

1. Gathering dataset from CSV file using pandas.
2. Gathering dataset from html in Udacity server.
3. Gathering dataset from twitter API by creating twitter developer account. (Unfortunately, I cannot created because twitter reject my request twice.) so I download JSON file manually from project section then read it as JSON.
4. Assessing all datasets
5. Clean both tidiness issue and quality issues
6. Storing wrangled data as CSV
7. Analyzing and visualizing data
8. Create reports

Tools:

1. python through Jupyter notebook (ANACONDA) to take a chance to practice in this first project for future projects.
2. Upload work in github." https://github.com/SamiAdham/TMDb-movie-data/blob/master/investigate-a-dataset-%5BSami%20Adham%5D.ipynb"
3. Microsoft Word to prepared 'PDF'report

Project Steps:

Gathering Data: Data was taken from 3 different sources

1. Download `twitter_archive_enhanced.csv` then read it directly using pandas
2. Pull dataset for img_predection fron Udacity server using 'Request library'

3. Using API to gather dataset as json file(but we gather it manually as we could not create twitter developer account)

Assessing Data:

Using some function that can help us in assessing data such as:

- .info()
- .head()
- .sample()
- .value_counts()
- .shape()

Then we come up with quality issues and Tidiness issue as following:

- Quality Issues
  - Twitter_archive
    1. Remove Retweets
    2. Remove all column in json file except [id,favorite_count,retweet_count] since we need it.
    3. Some inaccurate names usually start with lowercase
    4. Fix names with Unfamiliar name by convert it to nan
    5. Calculate Rating_Score rather than numerator and denominators
    6. Fix Rating Numerator
    7. Rename all column to make it easy to read and remove extra columns as need it
    8. change data type.

  - img_predications(2075)
    missing Value
  - tweet_json(2354)
    missing Value
- Tidiness
  1. Join all Data frames togather in One DataFrame which is twitter_archive using tweet_id.
  2. Doge Stages should not be in column.

Cleaning Data:

We used a punch of methods to help us cleaning such as:

.duplicate()	.rename()	.to_datetime()	.extract()
.isnull()	.merge()	.islower()	for loop
.drop()	.astype()	.replace()	def to create function
.melt()			

Conclusion:

In sum, we have cleaned data from our 3 sources and we merge it together by using tweet_id. Also, we have to assess our data to remove unused features from our merged dataset. In this stage we setup our wrangled data ready to EDA and visualization process.

Limitations:

1. The difficulty to create twitter developer account to use API

References:

GitHub:
https://github.com/SamiAdham/TMDb-movie-data/blob/master/investigate-a-dataset-%5BSami%20Adham%5D.ipynb"