UDACITY

Project 4: Data Wrangling & Analysis (wrangle_report)

By

MOHAMMED IBRAHEEM ALSHONEEFI

Introduction:

This project (Data Wrangling & Analysis) is a part of Udacity Data Analyst Nano Degree. Moreover, this report involves multiple stages consist of gathering data, assessing data, and clean the data. The first step includes gathering data from various sources – the sources required are from twitter- need to be merged. The next step responsible for accessing the data by take snapshots, build assumptions, and conclude results. Finally, cleaning the data is the last stage that include lots of effort and creative work. Starts with defining the qualities, and tidiness issues. Then start solves each issue separately.

Gathering Data:

The data used in this project downloaded manually from three datasets:

1- Twitter archive file this dataset includes various variables tweet_id, timestamp, source, rating, dog name, and the stage.
2- Image prediction file contains the learning machine results to find out the correct dog stage. This file downloaded programmatically using Udacity Request library from Udacity server
3- Tweets_clean this file has extra information about the tweets such as retweet count, and favorite counts and it use the Twitter API to gather it.

Assessing Data:

All the datasets have its own quality, and tidiness issues.

Quality Issues:

- timestamp is object, should be datetime instead of object.
- several columns have wrong null object, i.e. None instead of NAN
- in_reply_to_status_id,tweet_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id have wrong data type, should be integers or string instead.
- delete columns that won't used in analysis
- fix the the wrong ratings and standrize the values
- dog_stage has duplicated values should be handled
- 66 duplicated jpg_url
- missing values from images dataset (2075 rows instead of 2354)
- rename id to tweet_id for merging purpose

Tidiness Issues:

- Campine the 3 datasets all in one
- merge dog stages into one column

Cleaning Data:

The issues founded from the assessing stage have been fixed at this level and been reorganized here some of the functions used at this phase:

- rename()
- merge()
- drop()
- extract()
- loc[]
- astype()
- Regular expression
- Value_counts ()
- Info()
- Head()
- Loops
- ……

Storing Data:

After completing the phases, I saved the new dataset into `'twitter_archive_master.csv'`.

Conclusion:

Wrangling data is a core skill for each data analyst. I have used Python to do the wrangling and all the analysis steps that return on me a lot advantages to improve my python programming skills and qualify me to enroll in more advanced courses in the future thank you Udacity and Misk academy for this chance.