DAND- Udacity

Wrangle and Analyze Data - #WeRateDogs

Wrangle Report

Introduction

Udacity's Nanodegree program teaches us the data wrangling process. It includes the following:

- Gathering data from different sources(csv, tsv, json files or others).
- How to assess the gathered data visually and programmatically and writing down quality and tidiness issues.
- Cleaning data by solving those quality and tidiness issues.

After these steps we start analyzing and visualizing the clean data and get conclusions.

This project is about a Twitter account called @dog_rates, we got data from this profile and we started seeing the interactions, ratings of different dog types and dog stages. These ratings were made by users logged in to Twitter from different sources.

1- Gathering Data

There were three sources to gather data from:

- **Enhanced twitter archive:** which is a csv file that we read using pandas.
- **Image predictions file** which is extracted from url using requests and a tsv file.
- **Twitter API data** which is in form of a json file.

2- Assessing data:

Assessing data both visually and programatically in order to get to know the quality and tidiness issues. Dirty data is what has problems in content which is called quality issues. While messy data is what has tidiness issues. Assessment contains two steps:

- **Visual assessment:** which means scrolling through data to find some issues.
- **Programmatic assessment:** using different methods to find issues in the dataset such as (.describe(), .info(), .head, .tail … etc.).

The highlighted issues are:

Quality:

- Change in data type is needed in (tweet_id, source, timestamp, img_num and dog stages).
- Dog name needs to be edited and to be accurate. (some dogs are named 'a', or 'an').
- There are 183 retweets to be deleted.
- Sources need to be clearly defined.
- Some tweets don't include images.
- Delete duplicated jpg_url entries.
- Some denominators are of value other than 10.
- There are decimal numbers in ratings.

Tidiness:

- Merge the three datasets into one dataset using tweet_id.
- Delete unnecessary columns in all datasets.
- Dog stages need to be in one column instead of four.

3- Cleaning data:

In the cleaning steps we used the three steps (Define, Code and then Test) to make sure the issue was solved. We used different methods such as (drop, merge, loc, extract, ...etc.). Then we saved gathered clean data into a new csv file.

Conclusion:

Data wrangling is a core skill that any person working with data needs to know. In the whole mentioned above steps I used Python and its different packages like pandas, tweepy, numpy and others. Python has many good pros such as being good to deal with large data sets, and it can be run every period of time to see how data changes and findings change as well.

Resources:

- https://github.com/latinacode/Wrangle-and-Analyze-Data
- https://github.com/kdow/WeRateDogs
- https://github.com/auroredupontd/dand-p8-wrangle-analyze-data
- https://github.com/auroredupontd/dand-p8-wrangle-analyze-data
- https://stackoverflow.com/questions/11882393/matplotlib-disregard-outliers-when-plotting
- https://medium.com/python-pandemonium/data-visualization-in-python-bar-graph-in-matplotlib-f1738602e9c4
- https://python-for-multivariate-analysis.readthedocs.io/a_little_book_of_python_for_multivariate_analysis.html
- https://www.geeksforgeeks.org/python-pandas-melt/
- https://www.datacamp.com/community/tutorials/joining-dataframes-pandas
- https://www.geeksforgeeks.org/python-pandas-series-str-lower-upper-and-title/