ANA-522-OL1 Spring 2022
Mod03 week05 HW: Data Preparation
Due: Sunday February 13th at midnight

Explore the Dataset of your interests The purpose of the data preparation homework is to utilize what we have learned, mainly in pandas, to explore data in limited but practical ways. You are to find a dataset at a topic of your interest, to understand its data fields, and to apply cleaning and preparation techniques learned in the text reading (Chapter 7) of this week. The deliverable items to be submitted including:

0. Acquire a dataset of your interest.
   - Since we are practicing data cleaning and preparation, "look for" missing values in your dataset collection.
1. Write up the Abstract of your homework topic and a concise description the dataset to be analized. (20%)
   - Title and context of the data topic.
   - Resource links and the download address of the dataset.
   - A copy of the dataset file used is to be submitted with the Jupyter Notebook report.
2. Create a full list of fields and the description for each attribute (20%)
   - See Covid daily dataset as an example in the appendix of this instruction.
3. Handle Missing Data (20%)
   - Apply techniques covered in the text reading of this week to exam the data, filter out and fill in missing data, remove duplicates, update with proper values/names to meet the context of the data topic.
   - Provide informative statements and comments on the data cleaning opertions performaned at your work in the context of the data topic.
4. Handle Outliners (20%)
   - Apply techniques covered in the text reading of this week to exam the data, filter out or transform outliners to meet the context of the data topic.
   - Provide informative descriptions and comments to the outliner data transformation opertions performaned in the context of the data topic.
5. Summary and Conclusion (20%)
   - Use binning and sampling techniques to summarize the data and give conclusions, after applying data preparation steps above.

Turn in your Jupyter Notebook file (.ipynb) along with the dataset file used to upload on the Blackboard for submission.

Appendix Sample for COVID-19 Overview and Field List

COVID-19 Data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the

Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) is available on https://github.com/CSSEGISandData/COVID-19. Within the repository, there are daily reports for Covid-19 cases around the world. The field description for daily reports dataset can be found here and copying in the following:

- FIPS: US only. Federal Information Processing Standards code that uniquely identifies counties within the USA.
- Admin2: County name. US only.
- Province_State: Province, state or dependency name.
- Country_Region: Country, region or sovereignty name. The names of locations included on the Website correspond with the official designations used by the U.S. Department of State.
- Last Update: MM/DD/YYYY HH:mm:ss (24 hour format, in UTC).
- Lat and Long_: Dot locations on the dashboard. All points (except for Australia) shown on the map are based on geographic centroids, and are not representative of a specific address, building or any location at a spatial scale finer than a province/state. Australian dots are located at the centroid of the largest city in each state.
- Confirmed: Counts include confirmed and probable (where reported).
- Deaths: Counts include confirmed and probable (where reported).
- Recovered: Recovered cases are estimates based on local media reports, and state and local reporting when available, and therefore may be substantially lower than the true number. US state-level recovered cases are from COVID Tracking Project.
- Active: Active cases = total cases - total recovered - total deaths.
- Incident_Rate: Incidence Rate = cases per 100,000 persons.
- Case_Fatality_Ratio (%): Case-Fatality Ratio (%) = Number recorded deaths / Number cases.