DATA SCIENCE MATERIALS

Formative Assessments

#1

The table below contains data collected by New York City's law enforcement over the years 2000 to 2021. The data lists what the City's law enforcement considers as seven major felonious crimes. Use the data below to create a **data frame** for **five consecutive years** to implement what was modeled in class by following the steps below.

1. Reading the NYC Law Enforcement Crime Data from Eurostat
    a. Data downloaded and stored as: crime_major_Felony_Data.csv
    b. Steps:
        Download in same directory as Python. Open file, then read file
    c. Viewing top and bottom of data frame
    d. Viewing summary of the data frame
2. Selecting a subset of data from the data frame
    a. Selecting one column
    b. Selecting a subset of rows
    c. Selecting a subset of columns and rows
3. Use the data below to create a data frame for five consecutive years in Python as shown in class.

    Data Source:
    https://www1.nyc.gov/assets/nypd/downloads/pdf/analysis_and_planning/historical-crime-data/seven-major-felony-offenses-2000-2021.pdf

#2

The table below gives the heights of fathers and their sons, based on a famous experiment by Karl Pearson around 1903. The number of cases is 1078. Random noise was added to the data to produce heights to the nearest 0.1 inch.

https://www.kaggle.com/datasets/abhilash04/fathersandsonheight

#3

Study of animal abundance in forests

https://portal.edirepository.org/nis/codeGeneration?packageId=ecotrends.1.2&statisticalFileType=py

#4

NYC Health Survey Data

https://www1.nyc.gov/site/911health/researchers/health-data-tools.page

#5

High School & Beyond Data

https://nces.ed.gov/surveys/hsb/

#6

An Interesting dataset of your choice. It must contain at least four series, along with missing data, and at least 30 rows.

Summative Assessments

Project

Select one of the six suggested options for your project. Your selection cannot be one of the data sources you used for classwork. Your data must contain four series, along with missing data, and at least 30 rows (records) if you choose option #6.

Instructions	Grade

- **Part 1: First weekend after lesson #5** — **30%**
  - One paragraph description of dataset
  - Statistical summary of dataset
  - All code used to execute procedures from lessons #1 to 5
  - Check in: next Monday
- **Part 2: Second weekend after lesson #10** — **30%**
  - All code used to execute procedures from lesson #6 to 10
  - Output for bottom 5 and top 5 rows of processed dataset
  - Statistical summary of data set
  - Check in: next Monday
  - Submission: Thursday