Homework N1

The "crime_dataset.csv" reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy. (source)

Instructions:

1. Use the provided crime dataset (CSV format) for the tasks.
2. Complete each task in both Python (using pandas) and R (using dplyr).
3. Submit your cleaned dataset and code files (.ipynb for Python, .R for R).
4. Write explanations or comments for your code where you feel necessary.
5. Submit your homework on GitHub and provide the link in Moodle.

Part 1: Data Cleaning and Exploration (Python & R) *10 points per problem: 5 for Python and 5 for R*

1. Load the dataset. Check the first 5 rows.
2. Identify columns with missing values and their respective counts. Drop columns where more than 50% of the data is missing (store this version as a new dataset).
3. Convert the *DATE OCC* column to a datetime format. Extract the year, month, and day into separate columns. Create a new column for the hour using the *TIME OCC* column.
4. Filter the dataset for crimes that occurred in 2023. Further filter crimes with the description *BURGLARY* in the *Crm Cd Desc* column.
5. Group the data by *AREA NAME* and calculate the total number of crimes and the average victim age. Sort the results by total crimes in descending order.

Part 2: Further Exploration (Python only) *5 points per problem*

1. Find the top 3 most frequent *Crm Cd Desc* values.
2. Group the data by *Hour* and count the number of crimes.
3. Group the data by *Vict Sex* and calculate: Total crimes, Average victim age.

Part 3: Further Exploration (R only) *5 points per problem*

1. Group the data by *Month* and count the number of crimes.
2. Count the number of crimes where a weapon was used (*Weapon Used Cd* is not null).
3. Group the data by *Premis Desc* and count the number of crimes.

Part 4: Advanced Analysis (Python & R) *20 points per problem: 10 for Python and 10 for R*

1. Create a new column, *Severity Score*, based on the following rules:

- Assign 5 points if a weapon was used (*Weapon Used Cd* is not null).
- Assign 3 points for crimes under *BURGLARY*.
- Assign 1 point for all other crimes.
- Group by *AREA NAME* and find the total severity score for each area.

Bonus Part: *5 points*

Use the *LAT* and *LON* columns to identify crimes that occurred within a specific latitude-longitude bounding box (e.g., downtown area).