Programming Assignment 6

Answering Our Data Question

Learning goals

- Use nested for loops to iterate through list of dictionaries and a dataset
- Use the `CSE8ACSV` library (pre-written code) to load dataset and get access to the values in each field
- Get comfortable with data processing and data analysis
    - Develop questions to ask of the data, analyze the data using Python code, and answer those questions

Background: The CSE8ACSV library and the datasets

In this project, you will write code to conduct data processing. The CSE8ACSV library is equipped with a few functions that help with loading data from CSV (Comma Separated Values) files. Your job is to write computations based on the provided functionality to do fun things with data.

As part of the starter code, we have provided you with the library in CSE8ACSV.py and the six data sets already loaded into proj2.py.

You do not need to understand the code in CSE8ACSV.py, but you do need to understand what each function does (so you should understand the description of each function). Here is the documentation of all of the functions available in the CSE8ACSV library:

```
get_column_data(my_data, col_name)
```

This function extracts all data from a specific column in a dataset, where `my_data` is the dataset (a list of dictionaries, so each “row” in the dataset is a dictionary), and `col_name` is the name of the column to extract as a string. The function **returns** a list containing all values from the specified column.

For information on the datasets, please go to the Dataset Info pages on Edstem! There you will find all the relevant information on the datasets, as well as the question that needs to be answered!

Task – Pick a data set and write code to answer a data processing question provided

In proj2.py, you will choose **one** of the six datasets that we have provided and write code to answer a provided question for that dataset. **You only need to choose one dataset!**

After having submitted your code, make sure to ALSO COMPLETE THE QUIZ WHICH HAS A TOTAL OF 1 QUESTION! We want to see how you got to the answer that you did.

Dataset Info: BLM

BLM Dataset

The following information is about the following dataset: `blm_state.csv`

How to get the data?

In order to get the data for your project, we provided the following function from `CSE8ACSV`:

```
get_blm_data(filename)
```

This function reads a CSV file of Black Lives Matter protest data by U.S. state and returns a list of dictionaries.

Data Format

Each dictionary contains:

- `State` (String): U.S. state name.
- `BlackPop` (Float): Total Black population.
- `BlackPoverty` (Float): % of Black people below poverty line.
- `AsianPop` (Float): Total Asian population.
- `AsianPoverty` (Float): % of Asian people below poverty line.
- `HispanicPop` (Float): Total Hispanic population.
- `HispanicPoverty` (Float): % of Hispanic people below poverty line.
- `WhitePop` (Float): Total White population.
- `WhitePoverty` (Float): % of White people below poverty line.
- `TotalProtests` (Float): Number of protests.
- `TotalAttendance` (Float): Total protest attendance.

Provided Question

Which state had the highest attendance per protest?

Ex: If state A had 2 protests with a total 50 people attending,and state B had 3 protests also with a total of 50 people attending. Then between those two states protests at state A had more participants per protest compared to state B. Question 1 should analyze attendance per protest across all states in the dataset and find the highest attendance per protest.

Dataset Info: Tech Diversity

Tech Diversity Dataset

The following information is about the following dataset: `tech_diversity.csv`

How to get the data?

In order to get the data for your project, we provided the following function from CSE8ACSV:

```
get_tech_diversity_data(filename)
```

This function reads a CSV file on diversity in hiring in tech companies and returns a list of dictionaries.

Data Format

Each dictionary contains:

- `company` (String): Tech company name
- `year` (String): Year of data (e.g., "2016")
- `race` (String): Racial/ethnic category (e.g., "Asian", "Black_or_African_American")
- `gender` (String): "Male" or "Female"
- `job_category` (String): Job category (e.g., "Executives", "Professionals")
- `count` (Integer or "na"): Number of employees, or "na" if not available

Provided Question

What are the top 3 companies for Latine representation for executives?

Dataset Info: Incarceration

Incarceration Dataset

The following information is about the following dataset: `incarceration_data.csv`

How to get the data?

In order to get the data for your project, we provided the following function from CSE8ACSV:

```
get_incarcerated_data(filename)
```

This function reads a CSV file of incarceration admissions and releases by state and date and returns a list of dictionaries.

Data Format

Each dictionary contains:

- `date` (String): Date (format YYYY-MM-DD)
- `state` (String): U.S. state name
- `admissions_total` (Integer): Total number of admissions on the given date
- `admissions_white` (Integer): Number of admissions for individuals identified as White
- `admissions_black` (Integer): Number of admissions for individuals identified as Black
- `admissions_hispanic` (Float): Number of admissions for individuals identified as Hispanic
- `admissions_amerind` (Float): Number of admissions for individuals identified as Indigenous Americans
- `admissions_asian` (Float): Number of admissions for individuals identified as Asian
- `admissions_other` (Float): Number of admissions for individuals identified as Other
- `releases_total` (Integer): Total number of releases on the given date
- `releases_white` (Integer): Number of releases for individuals identified as White
- `releases_black` (Integer): Number of releases for individuals identified as Black
- `releases_hispanic` (Float): Number of releases for individuals identified as Hispanic
- `releases_amerind` (Float): Number of releases for individuals identified as Indigenous Americans
- `releases_asian` (Float): Number of releases for individuals identified as Asian
- `releases_other` (Float): Number of releases for individuals identified as Other

Provided Question

How did the pandemic (~year 2020–2022) impact the number of released incarcerated adults compared to previous years?

Dataset Info: Mobile Usage

Mobile Usage Dataset

The following information is about the following dataset: `mobile_usage.csv`

How to get the data?

In order to get the data for your project, we provided the following function from CSE8ACSV:

```
get_mobile_usage_data(filename)
```

This function reads a CSV file of mobile app usage and user demographics, returning a list of dictionaries.

Data Format

Each dictionary contains:

- `User_ID` (Integer): A unique identifier for each user in the dataset
- `Age` (Integer): The age of the user in years
- `Gender` (String): The gender of the user. This dataset only reports on using "Female" and "Male" values
- `Total_App_Usage_Hours` (Float): The total time (in hours) the user spent using apps during the recorded period
- `Daily_Screen_Time_Hours` (Float): The average number of hours the user spends on their screen daily
- `Number_of_Apps_Used` (Integer): The total number of apps the user interacted with during the recorded period
- `Social_Media_Usage_Hours` (Float): The time (in hours) the user spent on social media apps
- `Productivity_App_Usage_Hours` (Float): The time (in hours) the user spent on productivity-related apps
- `Gaming_App_Usage_Hours` (Float): The time (in hours) the user spent on gaming apps
- `Location` (String): The city where the user is located

Provided Question

For people living in New York, what takes up most of their daily screen time: social media apps, productivity apps, or gaming apps?

Dataset Info: Electric Vehicles

Electric Vehicle Dataset

The following information is about the following dataset: `ev_data.csv`

How to get the data?

In order to get the data for your project, we provided the following function from `CSE8ACSV`:

```
get_ev_data(filename)
```

This function reads a CSV file containing electric vehicle charging data and returns a list of dictionaries.

Data Format

Each dictionary contains:

- `User ID` (String): A unique identifier for the user associated with the charging session.
- `Vehicle Model` (String): The make and model of the vehicle being charged.
- `Battery Capacity (kWh)` (Float): The total energy storage capacity of the vehicle's battery, measured in kilowatt-hours.
- `Charging Station ID` (String): A unique identifier for the charging station used.
- `Charging Station Location` (String): The city or location of the charging station.
- `Charging Start Time` (String): The start time of the charging session, formatted as YYYY-MM-DD HH:MM:SS.
- `Charging End Time` (String): The end time of the charging session, formatted as YYYY-MM-DD HH:MM:SS.
- `Energy Consumed (kWh)` (Float): The amount of energy consumed during the charging session, measured in kilowatt-hours.
- `Charging Duration (hours)` (Float): The total time duration of the charging session, measured in hours.
- `Charging Rate (kW)` (Float): The average charging power during the session, measured in kilowatts.
- `Charging Cost (USD)` (Float): The total cost of the charging session in U.S. dollars.
- `Time of Day` (String): The general time period of the charging session (e.g., Morning, Evening).
- `Day of Week` (String): The day of the week when the charging session occurred.
- `State of Charge (Start %)` (Float): The battery's state of charge as a percentage at the start of the charging session.

- `State of Charge (End %)` (Float): The battery's state of charge as a percentage at the end of the charging session.
- `Distance Driven (since last charge) (km)` (Float): The distance driven by the vehicle since the last charging session, measured in kilometers.
- `Temperature (°C)` (Float): The temperature at the time of the charging session, measured in degrees Celsius.
- `Vehicle Age (years)` (Float): The age of the vehicle, measured in years.
- `Charger Type` (String): The type of charger used (e.g., Level 1, Level 2, DC Fast Charger).
- `User Type` (String): The category of the user, such as "Commuter," "Casual Driver," or "Long-Distance Traveler."

Provided Question

Calculate the difference of the cost of charging during the time of day when the charging cost is highest vs lowest.

Dataset Info: Gym Tracking

Gym Tracking Dataset

The following information is about the following dataset: `gym_data.csv`

How to get the data?

In order to get the data for your project, we provided the following function from `CSE8ACSV`:

```
get_gym_tracking_data(filename)
```

This function reads a CSV file of gym workout logs and returns a list of dictionaries.

Data Format

Each dictionary contains:

- `Age` (Integer): The age of the individual in years.
- `Gender` (String): The gender of the individual. This dataset only includes “Female” and “Male” values.
- `Weight (kg)` (Float): The weight of the individual in kilograms.
- `Height (m)` (Float): The height of the individual in meters.
- `Max_BPM` (Integer): The maximum beats per minute (BPM) recorded during the workout session.
- `Avg_BPM` (Integer): The average beats per minute (BPM) recorded during the workout session.
- `Resting_BPM` (Integer): The individual's resting heart rate, measured in beats per minute (BPM).
- `Session_Duration (hours)` (Float): The duration of the workout session in hours.
- `Calories_Burned` (Float): The total number of calories burned during the workout session.
- `Workout_Type` (String): The type of workout performed (e.g., Yoga, HIIT, Cardio, Strength).
- `Fat_Percentage` (Float): The body fat percentage of the individual.
- `Water_Intake (liters)` (Float): The amount of water consumed by the individual during a day, measured in liters.
- `Workout_Frequency (days/week)` (Integer): The number of workout sessions the individual performs per week.
- `Experience_Level` (Integer): The individual's experience level in fitness, on a scale where higher numbers indicate more experience.
- `BMI` (Float): The Body Mass Index (BMI) of the individual.

Provided Question

Which workout type has a shorter duration AND burns more calories?

Quiz

Question

For the provided question, explain what you did to answer this question?