SI206 Project 1
Winter 2025

Introduction:
In this project, you will work with Comma-Separated Value (CSV) files and nested dictionaries. CSV files are very common and these foundational skills will enable more complex data analysis.

On data science and social justice: Data scientists use data to create actionable insights. Social justice is a broad term for several movements based on furthering equality and ending socioeconomic oppression. We can use data science in the pursuit of social justice by uncovering inequity so that we can act to correct it.

The instructions, starter code, data, and data dictionary are included when you clone the GitHub repository. These files can also be found on the Canvas site under Files > Projects > Project 1.
Reminder: The first row in each data file is a header that describes the data.

Assignment:
The year is 1976, and you have been hired as a data scientist by a law firm specializing in civil rights and discrimination cases. Your most recent assignment involves gathering data to support the claim in the DeGraffenreid v. General Motors case. In this lawsuit, five Black women have filed a lawsuit against General Motors, alleging that the company's layoff policy perpetuated discrimination against Black women.

General Motors (GM) asserts that they did not engage in discrimination against the plaintiffs. They argue that their seniority-based layoff policy, which prioritizes laying off the most recently hired employees first, did not discriminate against individuals identifying as Black or as women at GM. GM also contends that rewarding longer tenure with greater job security is a fair practice.

The Black women, on the other hand, argue that GM's seniority-based layoff policy had a disproportionate impact on employees who specifically identified as both Black *and* female. They point out that due to discriminatory hiring practices, GM did not hire any Black women until after 1964. Consequently, when the early 1970s recession-driven layoffs occurred, the recently hired Black female employees were the first to be let go under the "last hired, first fired" seniority system. This resulted in all Black women at GM losing their jobs.

In the discovery for the case, GM provided historical employment data to you in a CSV format. It is up to you to verify your clients' claims using data. You will create the following five functions and four corresponding test functions, and run them in *main()* to load, store, and analyze this simulated data. For testing and validation, a truncated version of the dataset has also been provided.

1. ***csv_reader(filename)***

   csv_reader takes one argument: a string that represents the name of a file. The function returns a dictionary in which each key is the employee id and each value is another dictionary. Each inner dictionary will use the demographic categories or year hired as keys and their corresponding data as values. Hire year should be converted into an integer before it is added to the dictionary.

   Example output:
   When run on the GM data it should produce a dictionary like this:

   ```
   {
     'employee_1': {'gender': 'Female', 'race': 'White', 'hire_year': 1960},
     'employee_2': {'gender': 'Male', 'race': 'Black', 'hire_year': 1965},
     ...
   }
   ```

   test_csv_reader()
   tests ***csv_reader***
   Write a test case that checks for the length of the outer dictionary.
   Write a test case that checks for the length of the inner dictionary.

2. ***split_by_hire_year(employees, split_year)***

   GM's layoff policy is called "last hired, first fired." This is a common method of choosing people to lay off at a company that persists to this day. Due to the 1970s recession, the CEO decided to use this policy to shrink the company. So, we need to determine who was hired prior to a year, and who was hired after that year.

   split_by_hire_year takes two arguments: a dictionary of dictionaries, and an integer representing a hire year to split by. The function will iterate through that dictionary of dictionaries in order to return two dictionaries of dictionaries in a tuple: one in which the hire year is all before the split_year, and one in which the hire year is all after and including the split_year.

Example input and output with year 1964:

```
split_by_hire_year(employees, 1964) → ({
    'employee_1': {'gender': 'Female', 'race': 'White', 'hire_year': 1960},
    'employee_3': {'gender': 'Male', 'race': 'Other', 'hire_year': 1955}
}, {
    'employee_2': {'gender': 'Male', 'race': 'Black', 'hire_year': 1965},
    'employee_4': {'gender': 'Male', 'race': 'White', 'hire_year': 1970}
})
```

test_split_by_hire_year()
Test that the function correctly separates employees hired before 1964 from those hired in 1964 or later.
Test that the function correctly separates employees hired before 1970 from those hired in 1970 or later.

3. ***count_race_or_gender(employees)***

GM stated that their policy did not end up discriminating by race *or* by gender. We need to verify this claim with data.

count_race_or_gender takes one argument: a dictionary of dictionaries. The function should accurately count the number of employees belonging to each race and gender category. The output should be a dictionary containing two keys: 'race' and 'gender'. Under each key, there should be sub-dictionaries with race or gender categories as keys and their corresponding counts as values.

Example output:

```
{
    'race': {'White': 2, 'Black': 1, 'Other': 1},
    'gender': {'Female': 1, 'Male': 3}
}
```

test_count_race_or_gender()
Test that there are only two keys in the returned dictionary
Test that the function accurately counts the number of employees belonging to each race and gender category.

4. ***count_race_and_gender(employees)***

   The Black women claim that their policy discriminated by race *and* gender. We need to verify this claim with data.

   count_race_and_gender takes one argument: a dictionary of dictionaries. The function will return the number of employees within each combination of race and gender in a dictionary. The keys should be represented by the following format: “Race_Gender”. The dictionary should also be sorted in descending order by each key’s value.

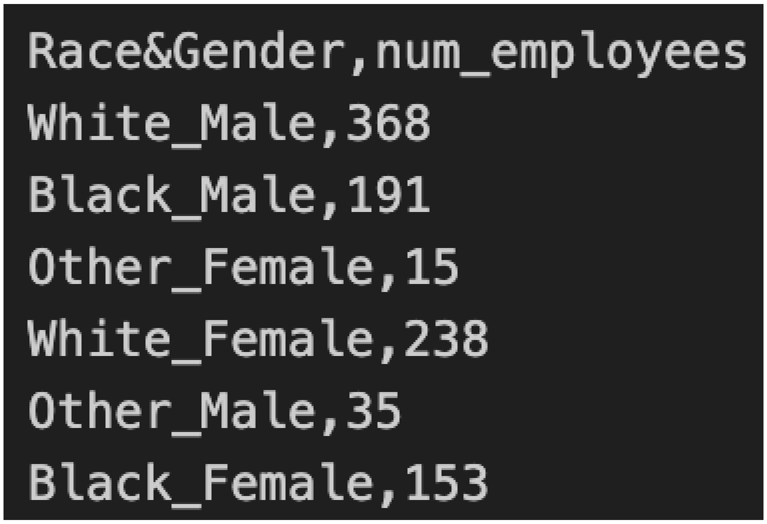
   Example output:
   {'Black_Female': 4, 'Black_Male': 2, 'White_Male': 1, …}

   test_count_race_and_gender()
   Test that there is the correct number of keys in the dictionary representing each combination of race and gender in this dataset.
   Test that the function correctly counts the number of employees within each combination of race and gender.
   Test that the dictionary is correctly sorted in descending order.

5. ***csv_writer(dict, filename)***
   csv_writer will take two arguments. A dictionary that was produced by *count_race_and_gender()*, and a string containing the location of a file. The function will write the data from the dictionary into a csv file.

   The first column should contain the combinations of demographics and the second column should be integers representing the number of employees with that combination of demographics.
   The first line of the file should be the header information and each row of data should be on a new line. **The header should follow the formatting in the example output below.**

   Example output in csv file:



```
Race&Gender,num_employees
White_Male,368
Black_Male,191
Other_Female,15
White_Female,238
Other_Male,35
Black_Female,153
```

(this is an example, your numbers and order may be different)

6. **Questions:**
   Once you’ve completed the coding portion of this assignment, use the data you produced to answer the following questions. Data scientists think critically about how to turn data into actionable information – the programming and quantitative pieces are only part of the job. **Turn in your answers to these questions as a PDF file in your Github repo along with your code and output**. A few sentences for each question is fine.
   a. What are the potential advantages and disadvantages of seniority-based layoff policies, from the company owner’s perspective, the senior employee’s perspective, and from the perspectives of the employees involved in the DeGraffenreid case? How might these stakeholders weigh these pros and cons differently?

   b. This homework project is based on one of the law cases which Dr. Kimberlé Crenshaw analyzed in order to describe how the lived experiences of Black women differ from the lived experiences of White women and simultaneously differ from those of Black men. Black women exist in a space where the realities of race and gender overlap. Within the American social structure, it is at times a toxic place where racism *and* sexism exist simultaneously. Professor Crenshaw named the place “intersectionality” [cite]. Her exploration of intersectionality brought to light complexities that would have remained hidden without listening to and incorporating the perspectives of Black women. In the context of data science, how can we apply this lesson to ensure that we aren't missing deeper narratives within our datasets? In other words, what are some limitations of just looking at data?

c. Data science is a powerful tool for uncovering information about the world, but it often grapples with imperfect data. In the context of this project, what limitations did you encounter regarding the data or analytical methods? Were there noticeable gaps in the data's representation of individuals' identities and experiences? How might these limitations impact the conclusions and insights derived from the data analysis?

7. **Extra Credit - CHOOSE ONE**

Extra Credit Option 1

In the context of the DeGraffenreid v. General Motors case, the layoff decisions were based on “last hired, first fired.” However, there could have been other ways to address these layoffs.

Imagine you are hired by GM to address the board of directors mandate of a significant cost reduction in the company’s payroll expenses. Develop your own algorithm that reduces payroll costs. You might decide to layoff employees, modify salaries, or a combination of both. You might consider factors like employee role, salary, race, gender, and gendered race. **Use the extra credit dataset for this problem, it has more headers, so you will need to create a new *csv_reader function*.**

reduce_company_costs(employees, target_reduction)

reduce_company_costs takes the employees dictionary as input and returns a new dictionary in the same format as the employees dictionary, however, the total payroll expenses has been reduced by the amount in the variable target_reduction. Any employee in the returned dataset with a salary of 0 will be laid off. Keep in mind, minimum wage is $5000.

test_reduce_company_costs()
Test when target_reduction is set to 5000000 (five million) the returned dictionary from ***reduce_company_costs*** does not contain any employee that has a salary between 0 and 5000.

Test when target_reduction is set to 5000000 (five million) the returned dictionary from ***reduce_company_costs*** reduces the cost of payroll by at least five million dollars.
Test when target_reduction is set to 6000000 (six million) the returned dictionary from ***reduce_company_costs*** reduces the cost of payroll by at least six million dollars.

In addition to the code above, answer the following questions in the same PDF in addition to the questions in part 6

1. What were the primary factors you considered when designing your layoff/salary reduction algorithm? Justify why you chose these factors over others.
2. Reflecting on this exercise, what are your thoughts on using algorithms for impactful decisions? Consider the benefits and risks of algorithmic decision-making, and discuss whether you believe it's appropriate to use algorithms in this context.

Extra Credit Option 2

In recent years, companies have increasingly turned to data-driven approaches for making complex business decisions, including layoffs. The use of algorithms in determining layoffs has both benefits and limitations.

Imagine you are an employee at GM who has been tasked to develop the layoff algorithm described in Extra Credit Option 1. However, after careful consideration, you have decided that using an algorithm for this decision is ethically problematic.

In a one-page memo (at least 500 words), write a persuasive argument addressed to your boss explaining why you will not develop the layoff algorithm. Your essay should critically examine the ethics of using algorithms in business decisions, such as layoffs.

Your essay should:

1. Be approximately 500 - 1,000 words.
2. Use a formal business tone appropriate for communicating with a superior
3. Present a clear and coherent argument against the use of algorithms for layoff decisions.

4. Propose alternative approaches to handling the company's financial challenges.

As you craft your argument, you might consider some of the following points:

- The human impact of layoffs and the importance of empathy in this process
- The potential for algorithmic bias and its consequences in this context
- The complexity of factors involved in layoff decisions that may not be quantifiable

Item	Points
csv_reader + ***test_csv_reader***	50 + 10
csv_writer	40
split_by_hire_year + ***test_split_by_hire_year***	14 + 6
count_race_or_gender + ***test_count_race_or_gender***	14 + 6
count_race_and_gender + ***test_count_race_and_gender***	24 + 6
Reflection shows critical thought	30
Extra credit (code + justification OR essay)	20