Introduction:
In this project you will be reading from and writing to Comma-Separated Values (CSV) files. Working with CSV files is an important skill in data management and analysis. There is a ton of data available on the internet and that data is often in a CSV format.

You will be looking at AP exam race and ethnicity data to compare it, state by state, with state race and ethnicity data.

Data Science & Social Justice:
Data science is a difficult concept to pin down; it can mean anything from working with excel to machine learning. The Oxford Reference defines social justice as *"The objective of creating a fair and equal society in which each individual matters, their rights are recognized and protected, and decisions are made in ways that are fair and honest."* (https://bit.ly/3xZrlY2). Data science can be used to further the goals of social justice by highlighting inequity and inequality in society.

Data Description:
One of the sources for this assignment is the AP exams for 2020**.** Secondary students can take Advanced Placement (AP) courses for college credit and/or placement. We pulled out just the race and ethnicity data that lists the number of exam takers for the Computer Science A exams. The columns of the data are of the different race and ethnicity groups that were measured, while the rows are of the different states. The format of the data is shown below

State	AMERICAN IND	ASIAN	BLACK	HISPANIC/LATI	NATIVE HAWAII	WHITE	TWO OR MORE	OTHER	NO RESPONSE	State Totals
Alabama	1	61	17	20	0	192	12	0	7	310

The other data is from the Census Bureau. We took the data from the American Community Survey. The exact query and resulting data table can be found here. We reduced the data to just the columns you need.

State	AMERICAN IND	ASIAN	BLACK	HISPANIC/LATI	NATIVE HAWAII	OTHER	State Totals	TWO OR MORE	WHITE
Alabama	25565	66270	1299048	208626	2238	70662	4876250	92220	3194929

In both cases, the files are CSV files that have already been cleaned. The AP data can be found here and the Census data can be found here. The data is also included when you clone the github repository. The data dictionary for the data can be found here. The first row of data for both files is the header information.

Assignment:

You will create the five functions that will read a csv into a dictionary, calculate the demographic percentages for each state in each dataset, calculate the absolute value of the difference between those percentages, and write data to a csv.

1. **read_csv("filename") -> dict**

   read_csv will take a filename to read from as a string. It will return a dictionary which has states as the keys and dictionaries as the values. The inner dictionary will use the demographic categories as the keys and either the number of exam takers or number of people of that category in that state as the values. You will want to convert the numbers from strings to integers (hint: use *int* to convert from string to integer).

   Example output:
   When run on the AP data it should produce a dictionary like this:
   {"Alabama": {"AMERICAN INDIAN/ALASKA NATIVE": 1, "ASIAN": 61,...},...}
   When run on the Census data it should produce a dictionary like this:
   {"Alabama": {"AMERICAN INDIAN/ALASKA NATIVE": 25565, "ASIAN": 66270,...}

2. **pct_calc(dict) -> dict**
   pct_calc will take a dictionary of dictionaries. The function will iterate through that dictionary (dict) of dictionaries (dicts) and return a dict of dicts where the inner key values are the proportion that each demographic category is of the state population.

   For the census data, this will be the percentage that each demographic is of the state's population. For the AP data, this will be the percentage that each category makes up of the total population of test takers in the state. Remember that each inner dict value is the count of that demographic in that data set. We recommend that you round your percentages to two decimal points.

   An example of this in general terms is:
   (White Population of state/State Totals) * 100 = percentage of the state's population that is white

Expected output:
When run on the AP data, it should produce a dictionary like as follows
{'Alabama': {'AMERICAN INDIAN/ALASKA NATIVE': 0.32, 'ASIAN': 19.68,...}...}
When run on the Census data, it should a dictionary like as follows
{'Alabama': {'AMERICAN INDIAN/ALASKA NATIVE': 0.52, 'ASIAN': 1.36,...}...}

3. **pct_dif(dict1, dict2) -> dict**
   pct_dif will take two arguments. The first is a dictionary with the AP data while the second is a dictionary with the census data. For each demographic category in each state (excluding the state totals of course), you will calculate the absolute value of the difference between the two data sets. This will produce a doubly nested dictionary which contains the difference between each "cell" of the data. As a reminder, the AP data contains a column that won't be found in the census data ("NO RESPONSE") so you'll have to ignore that. We recommend you round your percentages to two decimal points.

   An example of this in general terms is:
   Absolute value (% of population of state that is white - % of test takers of AP CS A exam that are white) = the percentage difference between the population and test takes for that demographic

   Expected output:
   When run on the two dictionaries, the resulting dictionary should look as follows
   {'Alabama': {'AMERICAN INDIAN/ALASKA NATIVE': 0.2, 'ASIAN': 18.32,...}...}

4. **csv_out(dict, "filename")**
   csv_out will take two arguments. The first is the dictionary that was produced through **pct_dif_calc** and the second is the name for the output file ("proj1-yourlastname.csv"). The function will write the data from the dictionary into a csv file. The first column should contain the states and the rest of the columns the percentages for each respective demographic category separated by commas. The first line of the file should be the header information and each row of data should be on a new line.

   # ex: Alabama,0.2,18.32,21.16,2.17,0.05,3.58,1.98,1.45

5. **max_min(dict, col_list)**
   max_min will take a dictionary and a list of columns (column headers from the input CSV file). Your goal is to create a triple nested dictionary that will contain states with the 5 largest and 5 smallest differences between their demographics and the demographics of AP exam test takers for each demographic. You will print and return that dictionary (Hint: use sorted. It will make your life easier.) If you choose to do the extra credit, the print statement won't be included in this function.

   Your returned dict & print statement (assuming you don't do the extra credit) should look like this:
   ...{'max': {'AMERICAN INDIAN/ALASKA NATIVE': {'Alaska': 14.89, 'South-Dakota': 8.75, 'New-Mexico': 7.78, 'Oklahoma': 7.04, 'Montana': 6.36}, 'ASIAN': {'South-Dakota': 42.99, 'California': 35.82, 'New-Jersey': 32.25, 'Virginia': 32.08, 'New-Hampshire': 31.24}, 'BLACK': {'District-of-Columbia': 23.92, 'Georgia': 23.9, 'Louisiana': 23.86, 'South-Carolina': 21.28, 'Alabama': 21.16}, 'HISPANIC/LATINO': {'California': 25.12, 'New-Mexico': 24.9, 'Texas': 18.81, 'Arizona': 15.25, 'Colorado': 13.51}, 'NATIVE HAWAIIAN/OTH PACF ISL': {'Hawaii': 7.07, 'Alaska': 1.25, 'Utah': 0.89, 'Oregon': 0.4, 'Washington': 0.37}, 'WHITE': {'New-Hampshire': 38.14, 'Washington': 27.87, 'South-Dakota': 26.4, 'Iowa': 25.26, 'Virginia': 23.5}, 'TWO OR MORE RACES': {'Mississippi': 11.15, 'North-Dakota': 5.3, 'Utah': 5.15, 'District-of-Columbia': 5.11, 'Kansas': 4.71}, 'OTHER': {'California': 13.95, 'Nevada': 10.26, 'New-York': 8.66, 'New-Mexico': 8.63, 'Arizona': 6.53}}, 'min': {'AMERICAN INDIAN/ALASKA NATIVE': {'Tennessee': 0.0, 'Massachusetts': 0.02, 'Texas': 0.03, 'Delaware': 0.04, 'Maryland': 0.05}, 'ASIAN': {'North-Dakota': 1.18, 'Alaska': 3.07, 'District-of-Columbia': 3.48, 'Utah': 8.02, 'Maine': 8.04}, 'BLACK': {'Nebraska': 0.03, 'Idaho': 0.09, 'Utah': 0.1, 'Vermont': 0.27, 'North-Dakota': 0.28}, 'HISPANIC/LATINO': {'Kentucky': 0.02, 'Hawaii': 0.04, 'Georgia': 0.13, 'Ohio': 0.29, 'Tennessee': 0.38}, 'NATIVE HAWAIIAN/OTH PACF ISL': {'Texas': 0.0, 'Massachusetts': 0.0, 'New-York': 0.01, 'West-Virginia': 0.02, 'New-Jersey': 0.02}, 'WHITE': {'Colorado': 0.56, 'South-Carolina': 0.87, 'Alaska': 2.16, 'Montana': 2.8, 'North-Dakota': 2.86}, 'TWO OR MORE RACES': {'Michigan': 0.04, 'Vermont': 0.15, 'Maine': 0.27, 'New-York': 0.5, 'Washington': 0.7}, 'OTHER': {'Maine': 0.27, 'Vermont': 0.39, 'West-Virginia': 0.44, 'New-Hampshire': 0.56, 'Montana': 0.67}}}

6. **Questions:**
   Once you've completed the coding portion of this assignment, think about the data and answer the following questions. Turn in your answers to these questions as well as your code.

a. Was there anything surprising in the data? Why or why not were you surprised?
b. Think about how you might attempt to convey this information visually. What sort of graph would you use and why?
c. Do you think the data is missing anything? Why or why not?
d. Can you tell a broader story with this data? If so, what broader story can you tell with this data? If not, what limits you?

7. **Extra Credit:**
   Create a pair of functions to calculate the national percentages for each dataset and compute the difference between them.

   nat_pct(data_dict, col_list)
   nat_pct will take in a data dict (ap_data or census_data) and then calculate the demographic percentages at the national level.

   The output should look like this for the ap_data:
   {'AMERICAN INDIAN/ALASKA NATIVE': 0.29, 'ASIAN': 33.45, 'BLACK': 3.46, 'HISPANIC/LATINO': 11.11, 'NATIVE HAWAIIAN/OTH PACF ISL': 0.09, 'OTHER': 0.0, 'State Totals': 65000, 'TWO OR MORE RACES': 4.81, 'WHITE': 42.07}

   And like this for the census_data:
   {'AMERICAN INDIAN/ALASKA NATIVE': 0.85, 'ASIAN': 5.52, 'BLACK': 12.7, 'HISPANIC/LATINO': 18.01, 'NATIVE HAWAIIAN/OTH PACF ISL': 0.18, 'OTHER': 4.94, 'State Totals': 324697795, 'TWO OR MORE RACES': 3.32, 'WHITE': 60.7}

   nat_dif(data_dict1, data_dict2)
   nat_dif will take in the two percentage dictionaries you created with the nat_pct function and calculate the difference between each value in them.

   The resulting dictionary should look like this:
   {'AMERICAN INDIAN/ALASKA NATIVE': -0.56, 'ASIAN': 27.93, 'BLACK': -9.24, 'HISPANIC/LATINO': -6.9, 'NATIVE HAWAIIAN/OTH PACF ISL': -0.09, 'OTHER': -4.94, 'TWO OR MORE RACES': 1.49, 'WHITE': -18.63}

Rubric:

Item	Percentage

read_csv has correct output	30%
csv_out has correct output	20%
pct_calc has correct output	10%
pct_dif has correct output	10%
max_min has correct output	15%
Reflection shows critical thought	15%
Extra Credit:	10%