Introduction:
In this project, you will work with Comma-Separated Value (CSV) files. Working with CSV files is an essential skill for data analysts due to the amount of data stored in this format.

Data Science & Social Justice:
Data science is a difficult concept to define; it can refer to anything from cleaning Excel sheets to creating machine learning models. 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 social justice by highlighting inequity and inequality in society.

Data Description:
One data source for this assignment is the SAT Suite of Assessments for 2021**.** We extracted the race and ethnicity data listing the number of exam takers for the SAT exam. The columns of the data are of the different race and ethnicity groups that were measured, while the rows are of the different regions.

The other data is from the United States 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.

Each file is in CSV format and is already cleaned. The instruction, the starter code, the SAT data, the Census data, and the 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. The first row of data for both the SAT and Census data is the header information.

Assignment:

You will create the following five functions, four corresponding test functions, and main to load, analyze, and store the data.

1. ***read_csv("filename")***

   read_csv will take a filename to read from as a string. It will return a dictionary of dictionaries in which a region is a key. 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 region as the values. You must convert the numbers from strings to integers.

test_read_csv tests ***read_csv***

Example output:
When run on the SAT data it should produce a dictionary like this:
{“west”: {“AMERICAN INDIAN/ALASKA NATIVE”: 2091, “ASIAN”: ...},...}
When run on the Census data it should produce a dictionary like this:
{“west”: {“AMERICAN INDIAN/ALASKA NATIVE”: 1253113, “ASIAN”: ...}….}

2. ***get_percent(dict)***
get_percent 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 region’s population.

    For the Census data, include each demographic’s regional population percent. For the SAT data, include each demographic’s regional test-taker percent. Remember that each inner dictionary value is the count of that demographic in that data set. Round your percentages to two decimal points.

    An example of this in general terms is:
    (White population of region/Region Totals) * 100 = percentage of the region’s population that is White

    test_get_percent tests ***get_percent***

3. ***get_difference(dict1, dict2)***
get_difference will take two arguments. The first is a dictionary with the SAT data while the second is a dictionary with the Census data. For each demographic category in each region (excluding the region totals), you will calculate the absolute value of the difference between the two datasets by subtracting the second dict from the first dict. This will produce a double nested dictionary which contains the difference between each “cell” of the data. As a reminder, the SAT data contains a column that won’t be found in the Census data (“NO RESPONSE”) so you’ll have to ignore that. Round your percentages to two decimal points.

An example of this in general terms is:
Absolute value (% of population of region that is white - % of test takers of SAT that are white) = the percentage difference between the population and test takes for that demographic.

test_get_difference tests ***get_difference***

4. ***csv_out(dict, "filename")***
   csv_out will take two arguments. The first is the dictionary that was produced through ***get_difference*** 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 regions 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.

5. ***max_min(dict)***
   max_min will take the argument of a dictionary. Use the provided ***max_min_mutate*** function to reformat the data into an easier-to-sort format. Your goal is to create a triple nested dictionary that will contain region with the largest and smallest differences between their demographics and the demographics of SAT test takers for each demographic. It will look like this:

   ```
   {"max": {"demographic": {"region": value}, ...},
     "min": {"demographic": {"region": value}, ...}...}
   ```

   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.

   test_max_min tests ***max_min***

6. **Questions:**
   Once you've completed the coding portion of this assignment, use the data to answer the following questions. Data scientists think critically about how to turn data into actionable information and may ask the following questions. Turn in your answers to these questions as well as your code.
   a. What story can you tell with this information? Does this story differ from your expectations? Why or why not?
   b. How could this data without the context of systemic racism be used for misinformation?

c. Think about the data sources for this information. Are there any limitations to the information based on these data sources (College Board and the US Census Bureau)? What information may be missing or is too generalized in these datasets?

d. Think about potential audiences for your story (perhaps College Board, state/federal education departments, journalists, and more). How can you use this information to advocate for individual, organizational, and/or policy change?

7. **Extra Credit:**

   Create a pair of functions to calculate the national percentages for each dataset and compute the difference between them.

   nat_percent(dict, col_list)

   Note: *col_list refers to the headers in the original spreadsheets loaded at the beginning of the homework. Most of these column names can be extracted from dict.*

   nat_percent will take in a dict (sat_data or census_data) and then calculate the demographic percentages at the national level.

   nat_difference(dict1, dict2)

   nat_difference will take in the two percentage dictionaries you created with the ***nat_percent*** function and calculate the difference between each value in them.

Rubric:

Item	Percentage
read_csv + ***test_read_csv***	25% + 5%
csv_out	20%
get_percent + ***test_get_percent***	8% + 2%
get_difference + ***test_get_difference***	8% + 2%
max_min + ***test_max_min***	12% + 3%
Reflection shows critical thought	15%
Extra credit	10%