Important

1. Due Date: November 26th, 2018
2. This homework is graded out of 100 points.
3. This is an Individual Assignment. You may collaborate with other students in this class. Collaboration means talking through problems, assisting with debugging, explaining a concept, etc. Students may only collaborate with fellow students currently taking CS 1301, the TA's and the lecturer. You should not exchange code or write code for others. For individual assignments, each student must turn in a unique program. Your submission must not be substantially similar to another student's submission. Collaboration at a reasonable level will not result in substantially similar code.
4. For Help:
    - TA Helpdesk (Schedule posted on class website.)
    - Email TA's or use Piazza Forums Notes:
    - How to Think Like a Computer Scientists [http://openbookproject.net/thinkcs/python/english3e/]
    - CS 1301 Python Debugging Guide [http://www.cc.gatech.edu/classes/AY2016/cs1301_spring/CS-1301-Debugging-Guide/index.html]
5. Don’t forget to include the required collaboration statement (outlined on the syllabus). Failing to include the Collaboration Statement will result in no credit.
6. Do not wait until the last minute to do this assignment in case you run into problems.
7. **Read the entire specifications document before starting this assignment.**
8. **IF YOUR CODE CANNOT RUN BECAUSE OF AN ERROR, IT IS A 0%.**
9. **IF YOUR DELIVERABLES ARE NOT NAMED EXACTLY AS STATED IN THE INSTRUCTIONS IT IS A 0%.**

CS 1301 Data Analytics Project

Introduction

In modern computing, Data Analytics is one of the most lucrative practices and engagements of mathematics, engineering, and computer science. It blends logic and reasoning, mathematical integrity, and creativity in order to solve problems. Some of the biggest problems today that Data Analytics is solving are traffic reduction, flight logistics, healthcare system economies, package delivery logistics, smart policing, modern medicine, ridesharing, etc.

The goal of this lab is to further understand the process of data collection and analysis. Carefully read the instructions below and refer to the rubric to see the components required for points. You have until **November 26th, 2018** to complete this assignment.

Instructions

Choose one CSV file and one JSON file from the list of approved data sets below that contain data that interests you. For each data set, you should come up with 3 research questions that you will answer by writing python functions to read and perform calculations on your data sets. After answering each question, use Microsoft Excel or Google Sheets (or Matplotlib for extra credit) to create a figure (bar graph, chart, scatterplot, etc.) to display your findings. You will then update the data.html page that you created for Lab01 with the questions you chose to investigate, the answers (supported with figures), and a short analysis of your findings.

Approved Data Sets

- CSV - choose one of the following:
    - Automobile data
        - https://archive.ics.uci.edu/ml/machine-learning-databases/autos/
        - Download link: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.data
        - More information about the dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/autos/imports-85.names
    - Airplane accident data
        - https://catalog.data.gov/dataset/accidents-fatalities-and-rates-1995-through-2014-u-s-general-aviation
        - Download link: http://www.ntsb.gov/investigations/data/Documents/datafiles/table10_2014.csv
    - Forest fires data
        - https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/

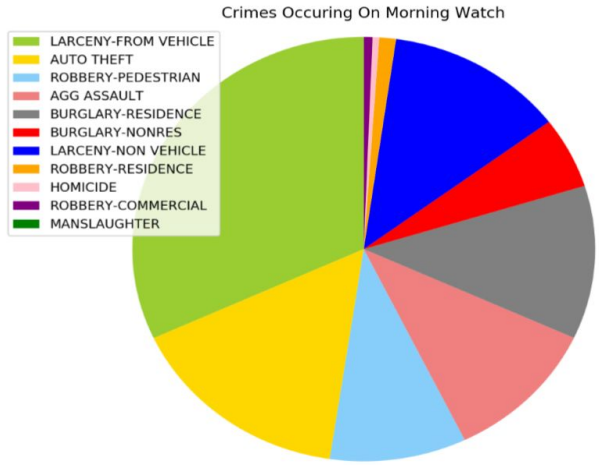
Rohan Bawa (rbawa@gatech.edu) and Roshini Damodoran (rdamodaran@gatech.edu)

        - Download link: https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv
        - More information about the dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.names
- JSON - choose one of the following (the files can be found on Canvas):
    - Iris data
        - This dataset includes sepal and petal size data for 150 iris flowers, as well as the species of each specimen
        - There are 3 possible values for the "class" column: "setosa", "versicolor", or "virginica"
    - Facebook interactions data
        - This dataset includes data about 500 posts from one cosmetic brand's Facebook page
        - There are 4 possible values for the "type" column: "Link", "Photo", "Status", or "Video"
        - This study describes the dataset in more detail: [http://www.math-evry.cnrs.fr/_media/members/aguilloux/enseignements/m1mint/moro2016.pdf]
    - Wholesale customers data
        - This dataset includes data about 440 wholesale customers in Portugal
        - There are 2 possible values for the "industry" column: "hotel/restaurant/cafe" or "retail"
        - There are 3 possible values for the "region" column: "Lisbon", "Oporto", or "other region"

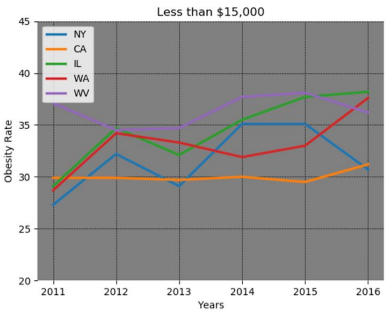
Some examples of appropriate analysis questions (do not copy these exactly!):

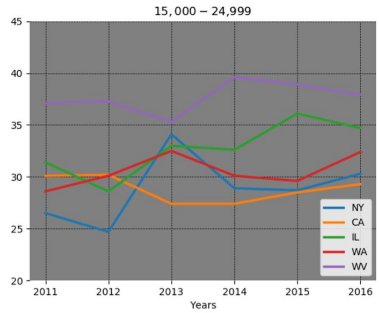
- For airplane accident data: Determine the average number of accidents per year from 1995 to 2014 and show the decreasing trend on a scatterplot.
- For airplane accident data: Find the two years from 1995 to 2014 in which the most and least fatalities occurred. Create a bar graph displaying fatality data for each year and indicate the maximum and minimum years.
- For Facebook interactions data: In which month do posts receive the largest number of likes? What about the smallest number of likes? Create a bar graph displaying average likes for posts published in each month and indicate the maximum and minimum months.

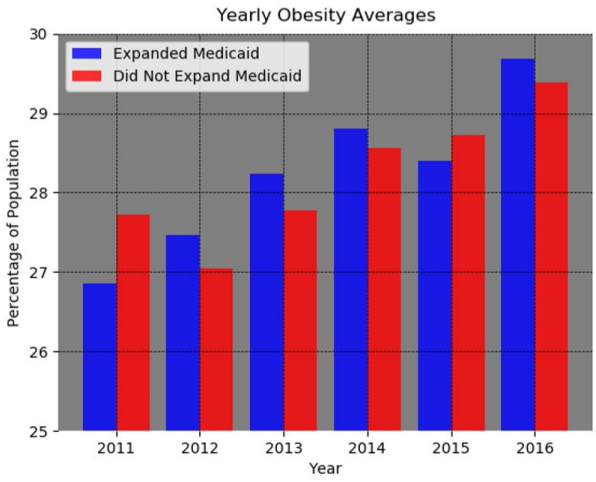
Examples of Graphs



States that Expanded Medicaid







Rohan Bawa (rbawa@gatech.edu) and Roshini Damodoran (rdamodaran@gatech.edu)

Interpreting JSON files in Python

We have provided 3 JSON files for you to choose from. One easy way to use these JSON files in Python is to import the **json** module. The **json** module contains a method called **json.loads()** which takes in a JSON string as a parameter and returns a Python dictionary object with all of the data provided from your JSON string.
EXAMPLE: Say you have a file called “example.json”. Here’s how you would turn it into a dictionary object in Python for easy usage.

```
import json
my_file = open("example.json", "r")
json_string = my_file.read()
# .read() returns your entire example.json file as one string
my_dict = json.loads(json_string)
my_file.close()
```

Now you can use **my_dict** like you would any other dictionary.

Analysis

For each data set, respond to the following:

- Why did you choose this data set / why is this data important?

For each of the research questions, respond to the following:

- State the question you chose to investigate here.
- Briefly describe the Python function you wrote to answer this question. Which columns/attributes in the data set helped you answer the question?
- Explain the graph or chart you created to display your findings, and point out some interesting things about your findings.

Important notes

For your data analysis calculations, you may ***only use*** Python. You ***may not*** use Excel to do any calculations. You ***may*** use Excel to create your charts and graphs. You may look online for ideas and help, but do not copy any code from them. DO NOT copy any code or ideas from the example provided.

Extra Credit

You may receive up to 15 points of extra credit on this lab for going above and beyond what is required. Bonus points will be awarded at the discretion of the TA’s. Here are some of the possible ways to earn extra credit:

- Using matplotlib, pandas, plotly, and/or numpy
- Using additional libraries or modules

- Analyzing more data sets in addition to the 1 required JSON and 1 required CSV
- Going above and beyond on the design of your HTML page
- Choosing thought-provoking and interesting research questions

Matplotlib (extra credit):

Matplotlib is a plotting library for Python. It allows you to display data and generate plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with just a few lines of code.

Use the following link as a reference and for tutorials: https://matplotlib.org/tutorials/index.html

Deliverables:

Have all of the following in a zipped folder called Lab02.zip.

- All your files from lab01 (lab1.html, data.html, style.css, etc.)
- lab2.py file containing all your functions
- Chosen CSV data set
- Chosen JSON data set

Grading Rubric

- CSV data set analysis
    - Explain why you chose your specific research questions - 3 pts
    - Data analysis question 1
        - Python function - 7 pts
        - Graph/figure - 3 pts
        - Explanation of function & analysis - 3 pts
    - Data analysis question 2
        - Python function - 7 pts
        - Graph/figure - 3 pts
        - Explanation of function & analysis - 3 pts
    - Data analysis question 3
        - Python function - 7 pts
        - Graph/figure - 3 pts
        - Explanation of function & analysis - 3 pts
- JSON data set analysis
    - Explain why you chose your specific research questions - 3 pts
    - Data analysis question 1
        - Python function - 7 pts
        - Graph/figure - 3 pts
        - Explanation of function & analysis - 3 pts
    - Data analysis question 2
        - Python function - 7 pts
        - Graph/figure - 3 pts
        - Explanation of function & analysis - 3 pts
    - Data analysis question 3
        - Python function - 7 pts
        - Graph/figure - 3 pts
        - Explanation of function & analysis - 3 pts
- data.html page
    - Validates - 6 pts
    - Includes all the questions with answers, graphs/figures - 10 pts
- Extra Credit - 15 points
    - Awarded at your grading TA's discretion