University of Utah
School of Computing

CS 6965	**Project #2**	**Spring 2018**

Due Feburary 20, 2018 at the start of class

Project 2 is the 2nd mini project, with a total of 15 possible points (15% of the final grade).

Project description:

Using a combination of data analysis and data visualization techniques to solve the Data Challenge problem from Utah Department of Workforce Services.

1 Getting Started

To start, please read the PDF "DataChallenge.pdf" that describes the project (the project description); together with the data file "UT_asu_exampleData.csv" and "tl_2016_49_tract.shp".

All 3 files can be downloaded via Assignment 2 on Canvas.

2 Expected Outcome

Recommended programming language: Python and D3.js.

This project can be considered as a real research project where minimal guidance is given. For instance, you are expected to learn to process a shape file "*.shp"; or if you have no prior visualization experience, learn some basic visualization using D3.js.

The project will be graded based on two parts:

- Create an algorithm that try to "identify the set of n Utah Census Tracts in k contiguous clusters (having a minimum cluster unemployment rate of 6.5 percent) that maximizes the total population across all n Census Tracts." At the minimal, implement the naive algorithm described in the project description.

- Create a visualization that, at a minimal, visualizes census tracks in a map visualization (and clusters of them) that satisfy the minimum unemployment requirement.

3 Submission

Project 2 is to be submitted via Canvas. Please provide the following in a ZIP file (contains a PDF and another ZIP):

- Report: A PDF that describes your algorithm and visualization design and implementation (the two parts described above). (5 points; 3 points for algorithm description; 2 points for visualization design description.)

- Source code: A ZIP file that contains the source codes for both the algorithm and visualization, and a REAME on how to run the code. The program is expected to run properly. Please include a file that includes the list of Census tracks IDs (or clusters of the IDs in consecutive regions) and their corresponding (combined) unemployment rates. There is no partial credit if the program does not run. (However, if you describe your implementation efforts in the report, you will get partial credits in the report portion). (5 points)

- A link to your deployed D3.js visualization, or a video that showcase the visualization in action, or a captured image of your visualization. (5 points)

4 Some Ideas on Algorithm

Here are some ideas as what you might want to do with this project. Some ideas are not guaranteed to work, but probably worth trying.

It is recommended you implement your algorithm in Python (given your exposure to Python vis scikit-learn). Other programming language is fine as well.

- Learning Geospatial Analysis with Python

- Read some research paper with the topic of "Spatial Scan Statistics"

- Read some research paper with the topic of "Geometric range queries"

- Implement the baseline (greedy) algorithm, that is, find the "seed" tracts and expand from the seed tracts

- Think about whether you could apply existing clustering algorithms to solve this problem (e.g. hierarchical clustering, k-means, etc.)

- You might not find the optimal algorithm; but you should try to find a solution that is better than the solution provided by the baseline algorithm

- Think about creating an adjacency graph among all tracts

- Can dimensionality reduction help in this project?

- What about mapper algorithm?

5 Some Ideas on the Visualization

Once you have obtained a solution, that is, a list of census tracts, visualize the census tracts in a map visualization similar to Page 5 of the project description. Additionally, you could consider the following visual features:

- Think about what type of interaction you could do with the map visualization.

- Imagine that you want to include "human-in-the-loop", where you enable the users to select consecutive tracts and compute the unemployment rate and population on the fly.

If you would like to use something other than D3.js to visualize the census tracts, please discuss with the instructor ahead of time.

6   Recommended Reading

Interactive Data Visualization for the Web, 2nd Ed. Scott Murray, 2017: `http://alignedleft.com/work/d3-book-2e`

Also other related D3.js materials online.