CSci 343 Fundamentals of Data Science
Challenge 4

Submission Window Opens:
November 4, 2019

Points Available:
200 XP for a working demonstration
50 XP for readable & understandable code

Objectives:

- Learn about basic Nearest Neighbor Approximation
- Learn about migration patterns within the US
- Have fun!

Assignment:

We've been looking at a lot of maps lately. In particular, we've been reconstructing missing data in maps (also images and plots) using *nearest neighbor approximation*. This method is often used to approximate trends determined by surveys or polls. One very good example is a linguistic survey originally administered by North Carolina State University (http://www.businessinsider.com/22-maps-that-show-the-deepest-linguistic-conflicts-in-america-2013-6/). In this challenge, we will be making maps similar to these.

You have been hired as a junior data analyst for the US Census Bureau. Your first assignment is to analyze *Domestic Migration* patterns. Domestic migration is the metric used to determine if people are moving into or moving out of a particular region. You've been given a CSV file that describes the domestic migration of a sampling of counties scattered across the continental US. For this sampling, the US was divided into a uniform grid of 194x120. Each grid unit represents a 13-mile wide square. Not all counties in the US are represented in the data, and you've been asked to construct a full map that models the missing data based on the available data. The CSV file contains three columns: X grid position, Y grid position, and change in population between 2015 & 2016. You were also given a second CSV file that contains the X,Y grid positions that define the outline of the US border.

Your task is to implement the *Mean k-Nearest Neighbors Approximation* method. You are to use this method to generate a map of the US that

shows the rate of domestic migration across the country. For this assignment, you will plot your reconstruction along with the US map outline. Your program will need to prompt the user for a k-value. The k-value represents the number of neighbors you wish to sample. This value can be any integer greater than 0.

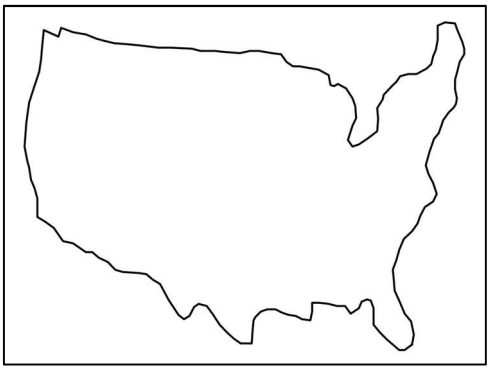
You will need to plot your reconstruction as a scatter plot using a color map. The color map you need to use is "viridis". You can earn a 10XP bonus (and your teacher's respect) if you can make your program only draw the reconstruction within the US border.

Reminder: Don't forget to use one of the more efficient methods of calculating distance! We talked about some of these in class. They are also discussed on the class wiki:
http://www.cs.olemiss.edu/~jones/doku.php?id=csci343_nearest_neighbors

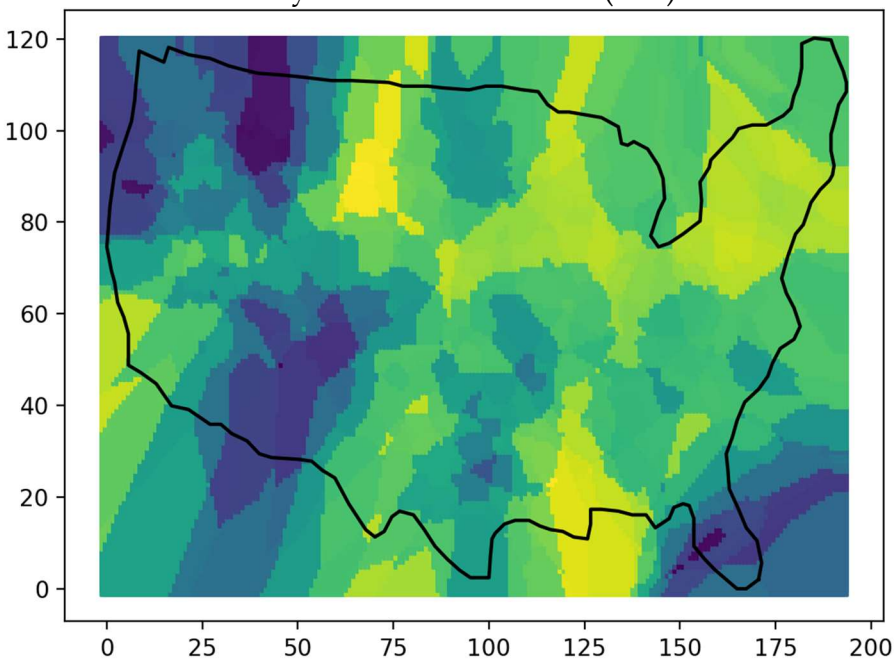
Here is an example of the input data and output image:

Reference Data

US Outline



Fully Reconstructed Data (k=5):



Deliverables

Before you upload your code to Blackboard, you MUST demo your *working* project to the TA. Once you've done this, you can upload ***all your code*** and a saved image of your final output (named "reconstructed.png") to Blackboard as a single ZIP file. Name your ZIP file s*piritAnimal*.zip, where s*piritAnimal* is your class user ID (not your webID or ID number). Be sure to name your main source file "main.py". In a comment at the top of the file, include the following information.

- Spirit Animal User ID, Date the file was last edited, Challenge Number
- Cite any sources that you used as a reference for code, data, and content (including title and URL)