



Portfolio Milestone Presentation

Randall Scott Taylor
Summer 2020
SUID: 965757884

Portfolio Milestone Presentation

Course Projects included in Portfolio:

- IST652: *Scripting for Data Analysis*
- IST659: *Database Administration Concepts and Database Management*
- IST707: *Data Analytics*
- IST718: *Big Data*

- Course Projects not included in Portfolio:
 - ACC 652 Accounting Analytics
 - IST 687 Introduction to Data Science
 - IST 664 Natural Language Processing
 - IST 623 Intro to Information Security
 - IST 719 Information Visualization
 - IST 769 Advanced Database Management

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE


Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

Portfolio Milestone Presentation

- IST 659: Database Administration. Professor *Gregory Block PhD*
 - *Law Office database project*
- IST 652: Scripting for Data Analysis Professor Deborah V. Landowski, PhD
 - *World Atlas of Language Structures*
- Data Analytics: Professor Jeremy Bolton, PhD
 - *Community Health Status Indicators*
- ST 718: Big Data Professor Jon Fox
 - *COVID-19 County Spread Predictors*

General Learning Objectives

- Methodology
 - Articulate the business question(s)
 - Data Acquisition
 - Data Cleansing, Transformation, Architecture
 - Analysis
 - Visualization
 - Interpretation
 - Summary Assessment, Actionable Steps
 - Did you answer the business question(s)
-  Iterative

Applied Data Science Program's Seven Learning Objectives

The Applied Data Science Program has seven stated learning objectives, which were achieved as this portfolio shall prove:

1. *Describe a broad overview of the major practice areas in data science.*
2. *Collect and organize data.*
3. *Identify patterns in data by way of visualizations, statistical analysis, and data mining.*
4. *Develop alternative strategies, based upon the data.*
5. *Develop a plan of action to implement the business decisions derived from analysis.*
6. *Demonstrate communication skills regarding data and its analysis for relevant professionals in their organization.*
- *Synthesize the ethical dimensions of data science practice.*

IST 659: Database Administration

- **Project Description**

- the student created the “Law Office” database, which was created to serve a newly created law office within the Financial District of Manhattan. The focus of the practice is: criminal, family, torts, and immigration law
- First: Data Normalization – to understand the data, such that it can be normalized, one must become familiar with a couple of key players:
- The stakeholders:

IST 659: STAKEHOLDERS

Stakeholder Description:

The benefit of centralizing this data and being able to aggregately track its contents is a business value added action, that will benefit the following stakeholders:

Client: The preservation of the integrity of any case or cause of action is paramount within the judicial system. From a client stakeholder perspective, the centralization of such data is in their best interest from a legality standpoint, as well as from a business decision. Proper tracking as to the schedule, assignment, and elements of their respective cases all helps to expedite a very tenuous and stressful situation.

Legal Personnel *Paralegal*: The first person that the client stakeholder usually engages with, the Paralegal has a vital role that relates to the interactions of the Client stakeholder, all the way from case intake to scheduling. The establishment of a centralized database would be a massive value added to their business processes and would assist to expedite their support role within the litigation process.

Legal Personnel *Attorney(s)*: Clients make appointments with their counsel and representation Paralegals work to sort out scheduling, judicial form generation, court room assistance, and billing. All these actions are such that the attorney can represent the interest of the client in the most productive way possible.

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

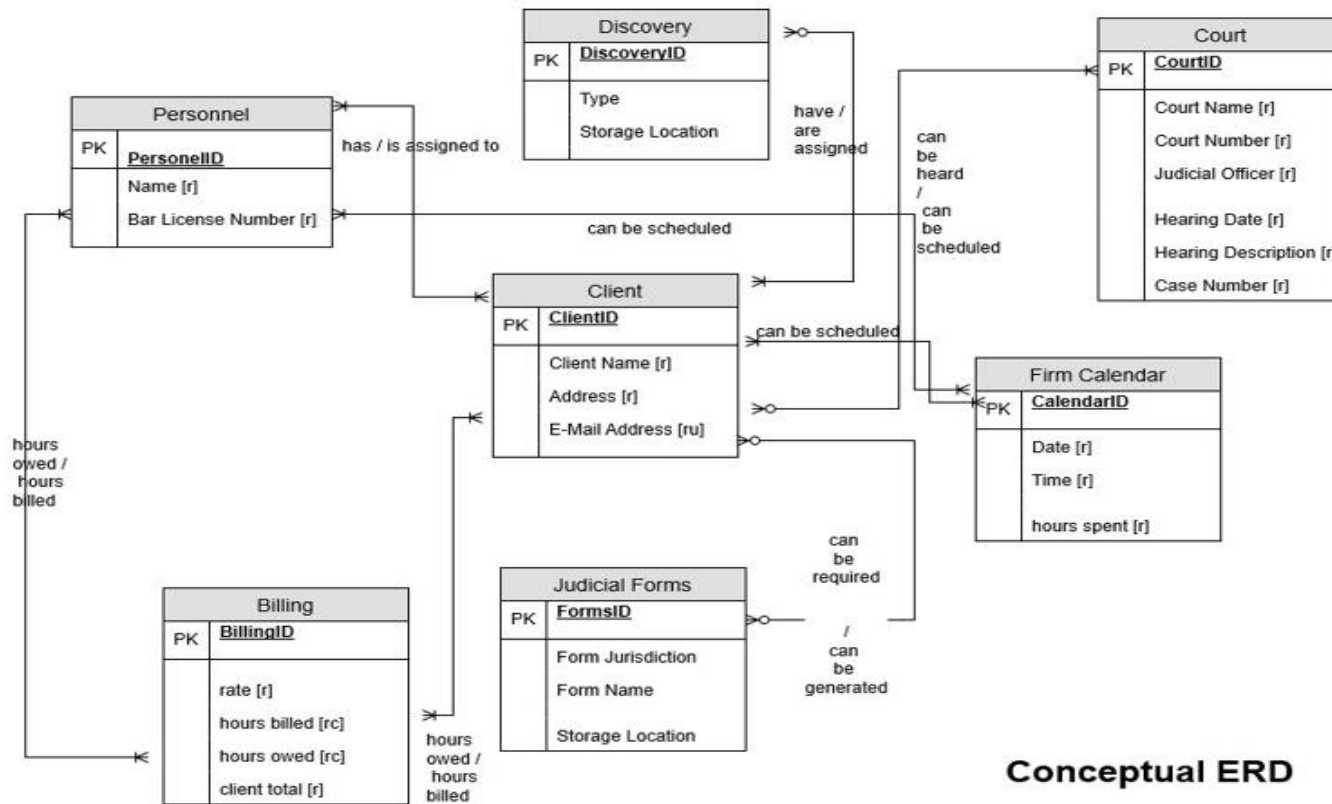
ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 659: CONCEPTUAL ERD

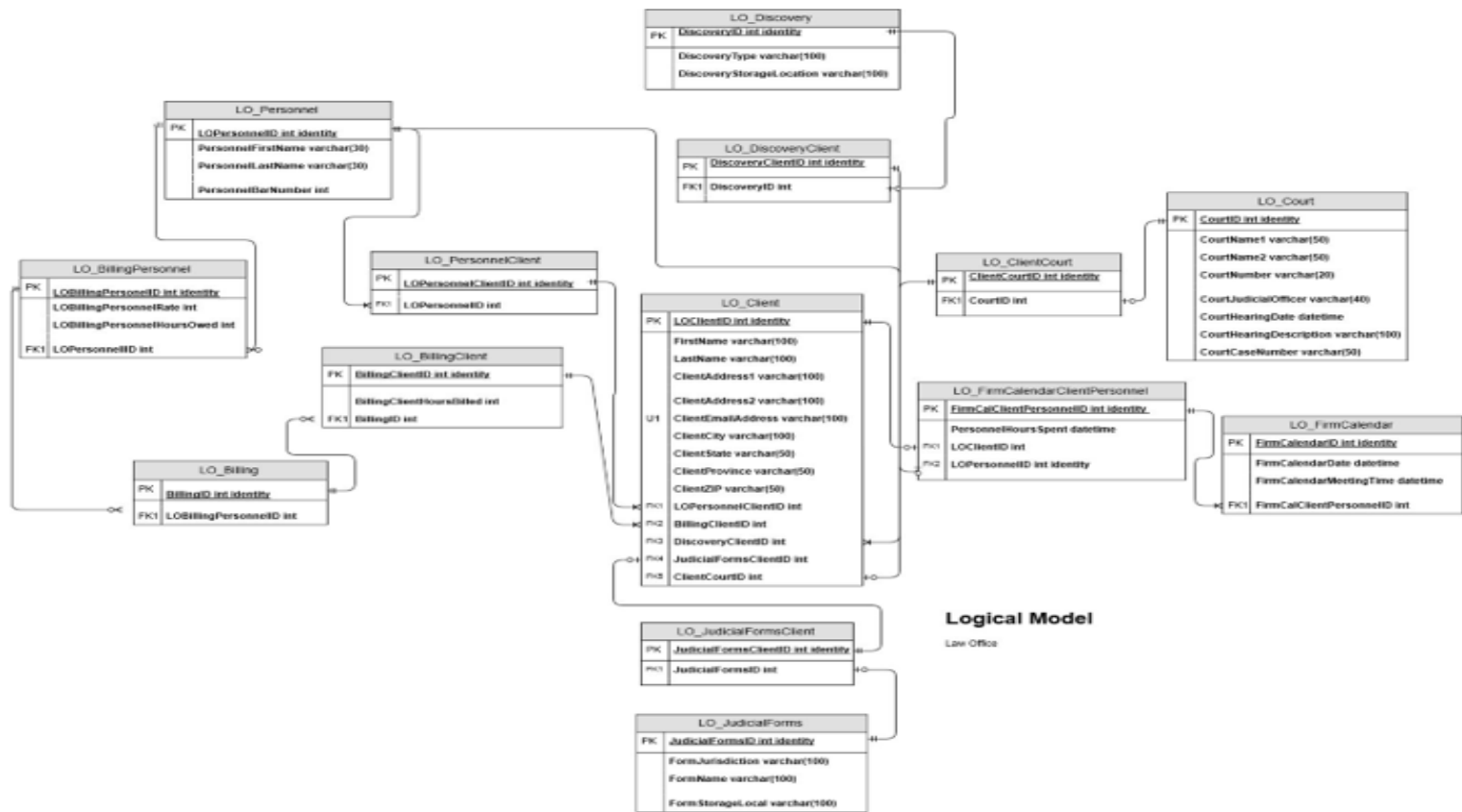


Conceptual ERD

Law Offices

IST 659: THE LOGICAL MODEL

The Logical Model: Law Office (UPDATED)



IST 659: GETTING AT THE BUSINESS QUESTION

- Based upon the logical ERD, tables were created via SQL Server Management Studio, that were normalized. There were specific data questions, goals, that were stipulated to coincide with the databases creation. For instance, “as a user how does one demonstrate a Law Firm Master Calendar schedule, such that a user can create / view TOTAL calendar personnel assignments” (Taylor,” IST 659,” 2020)? To answer this question and others like it, several functions, views, and stored procedures were created to facilitate in the addition of data, organization of data, and reporting of data, as the following examples demonstrate:

```
CREATE PROC spCreateFirmMasterCalendar
    @MasterMeetingNumber int,
    @PersonnelHoursSpent int,
    @ClientID int,
    @PersonnelID int,
    @FirmCalendarID int
AS
BEGIN
    IF EXISTS
        (SELECT * FROM FirmCalendarClientPersonnel WHERE @MasterMeetingNumber = MasterMeetingNumber )
    BEGIN
        UPDATE FirmCalendarClientPersonnel
        SET MasterMeetingNumber = @MasterMeetingNumber, PersonnelHoursSpent = @PersonnelHoursSpent, ClientID = @ClientID, PersonnelID = @PersonnelID,
        column MasterMeetingNumber(int, not null)
    END
    ELSE
    BEGIN
        INSERT INTO FirmCalendarClientPersonnel
        ( MasterMeetingNumber, PersonnelHoursSpent, ClientID, PersonnelID, FirmCalendarID)
        VALUES
        (@MasterMeetingNumber, @PersonnelHoursSpent, @ClientID, @PersonnelID, @FirmCalendarID)
    END
    RETURN @@IDENTITY
END
GO
```

IST 659: GETTING AT THE BUSINESS QUESTION

```
GO
CREATE VIEW TotalBillingHoursByClient

AS
SELECT DISTINCT Client.ClientID, Client.ClientLastName, Client.ClientFirstName, (SUM(BillingClient.ClientTotalBillingHours)) AS TotalAllBilling
FROM BillingClient
RIGHT OUTER JOIN Client
ON BillingClient.ClientID = Client.ClientID
GROUP BY BillingClient.ClientTotalBillingHours, Client.ClientID, Client.ClientLastName, Client.ClientFirstName

GO

SELECT * FROM TotalBillingHoursByClient
```

The above are a small sample to the many additional functions, stored procedures, functions, views, and reports that were created to answer the data questions that were presented further within the project. When the SQL database and Access were connected, the candidate had created a functional relational database, that would be able to serve as the basis for this new law firm in Manhattan

Client Billing Hours Report

Last Name	First Name	Total Billing Hours
Crable	Shelly	80
Crane	Icabod	20
Crane	Icabod	78
Lee	Gavin	40
Mustang	Shelby	40
Samson	Hillary	65
Thomas	Greg	40
Thomas	Greg	60
Tucker	Chris	
Ulvade	Franny	
Vincent	Edward	40
Vincent	Edward	78
Wearhouse	David	10

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 659: REFLECTION AND LEARNING GOALS

- At the beginning of this project, the candidate was new to the processes and procedures needed to create and normalize a database. The way ERD Cardinality functions, the concept of bridge tables, the logical model – these are ways of perceiving information that I had never once considered. I never considered that there were levels of normal form, three standards (that have been learned) about, let alone one standard form, had never once entered my vernacular, or mind mappings.
- The candidate's assumptions from the start of the project, having now painted a picture of my lack of understanding regarding SQL, were very incorrect. I had 'application' 'app' in my mind the whole of the beginning of the class, without an understanding as to how the tables related to one another, nor how the transacted data.

IST 652: Scripting for Data Analysis

- **PROJECT DESCRIPTION:**
- In response to the final project's requirements, and to answer the call of the questions presented, the candidate authored the following final report "WALS-Dataset." Description of the data and its source(s); the purpose of this study is the presentation of a suitable dataset, that has been extracted, transformed, and loaded into python3 for further analysis. The subject matter of this study was derived from the World Atlas of Language Structures (WALS) Online, which is a database of structural (phonological, grammatical, lexical) properties of languages gather from descriptive materials, from around the world. The database is maintained by the Max Planck Institute for Evolutionary Anthropology. The editors are: Martin Haspelmath, Matthew S.Dryer, David Gil, and Bernard Comrie.

IST 652: Scripting for Data Analysis

- **PROJECT DESCRIPTION:**

- To answer these questions, the .csv file was read into the local environment of Jupyter notebooks for further data exploratory analysis utilizing the Python 3 programming language. The methods of analysis that followed were: Data Exploration, Data Cleaning, Data Exploration of Unstructured Data, and Data Cleaning of Semi-Structure geojson. There were research questions presented and answered by the data.
- *Data Exploration* : the WALS dataset, obtained to a local machine under the title of *languages.csv*, was downloaded from the aforesaid website, and was imported into Python 3 via the Jupyter notebook's IDE. The dataset contains the release of data showing the geographical distribution of structural linguistic features years of data (from 2005 to 2008). The database is updated yearly

IST 652: Scripting for Data Analysis

- **PROJECT DESCRIPTION:**
- *Reviewing the Data*, first step in the process is to understand the dataset that is to be cleaned, modeled, and visualized. The features (columns) of the dataset are presented for exploration in the wide format, a subject's repeated responses will be in a single row, and each response is in a separate column. To give the reader a sense of the data collected with these rows, see the following demonstration: a screenshot of the raw .csv data:

	A	B	C	D	E	F	G	H	I	J
1	wals_code	iso_code	glottocode	Name	latitude	longitude	genus	family	macroare	count
2	aab			Arapesh (Abu)	-3.45	142.95	Kombio-Arapesh	Torricelli		PG

IST 652: Scripting for Data Analysis

- **PROJECT DESCRIPTION:**

- *Data Cleaning* was accomplished by importing the .csv file into python, utilizing the following libraries via import statements:

```
#In order to complete the exploratory analysis of the wals dataset, the
#researcher will require the following libraries:
import pandas as pd #for data processing, CSV file input/output
import os # directory structures access
import numpy as np # numpy arrays, linear algebra
import matplotlib.pyplot as plt # this is for plotting
from sklearn.preprocessing import StandardScaler
from mpl_toolkits.mplot3d import Axes3D
from mpl_toolkits.basemap import Basemap
```

- Unstructured Data obtained via geojson files from the wals website was also scraped to add to the study. To explore the data further the researcher has utilized the following libraries to web scrap and pull in the data from the main website, to further understand the vast data contained within the dataset.

```
#imports
import requests
import folium
import json
import pandas as pd
import numpy as np
from pymongo import MongoClient
from IPython.display import HTML
from folium.plugins import HeatMap
import matplotlib.pyplot as plt
```

IST 652: Scripting for Data Analysis

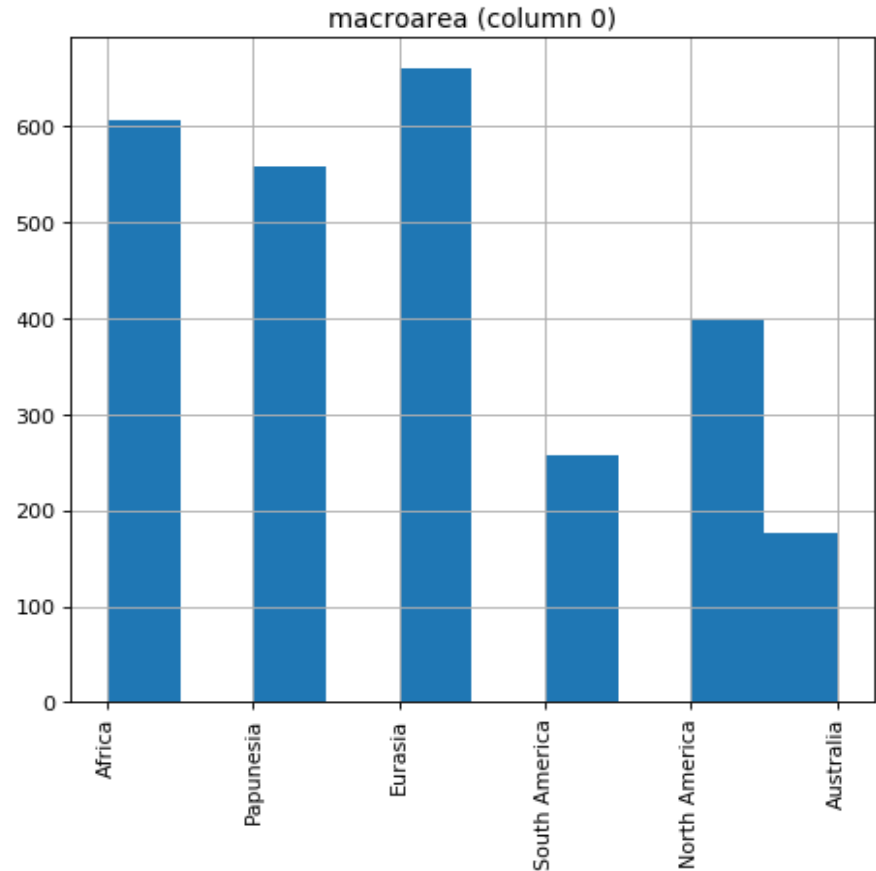
- **PROJECT DESCRIPTION:**
- To answer research questions, comparison questions about the data, the data transformation and feature extraction process enhanced the data in such as fashion as to increase its likelihood for classification algorithms. This established meaningful prediction that the data may provide. The dataset as presented here is rather sorted for the specifics of the beginning of this research project.

IST 652: Scripting for Data Analysis

- **RESEARCH QUESTIONS:**

- Exploratory research questions presented against the data:

- *What is the frequency distribution of the language's, determinate upon their macroarea?*



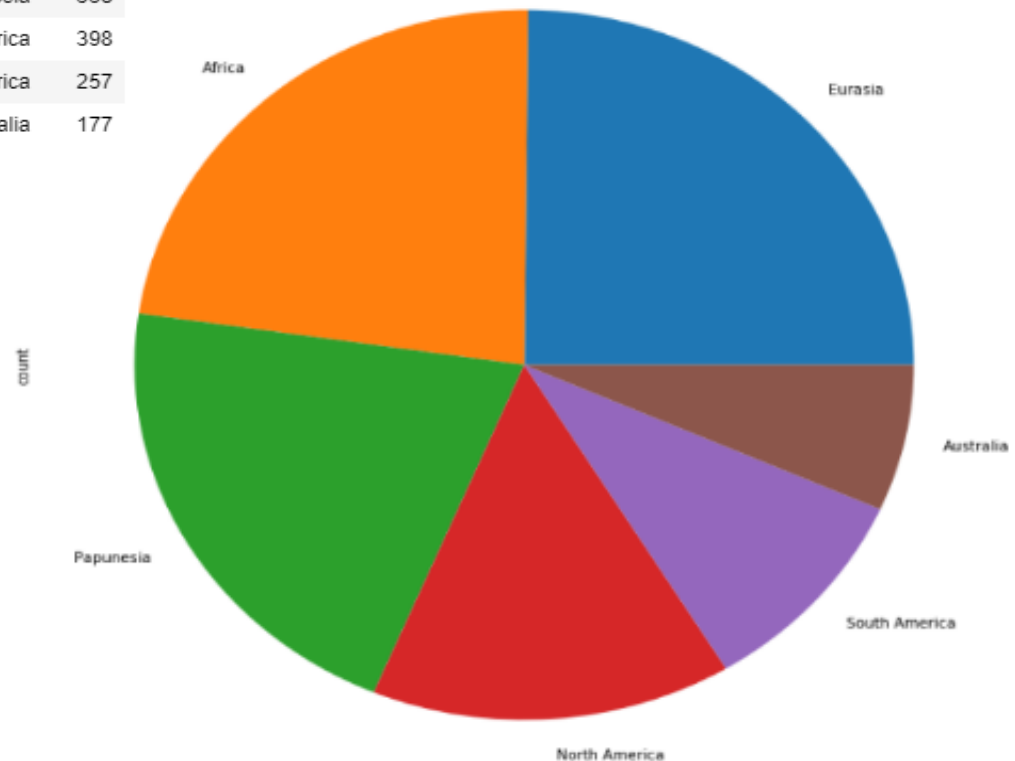
IST 652: Scripting for Data Analysis

- **RESEARCH QUESTIONS:**

- Exploratory research questions presented against the data:

- *What is the frequency distribution of the language's, determinate upon their macroarea?*

	macroarea	count
2	Eurasia	660
0	Africa	607
4	Papunesia	558
3	North America	398
5	South America	257
1	Australia	177



IST 652: Scripting for Data Analysis

- *What are the major language*

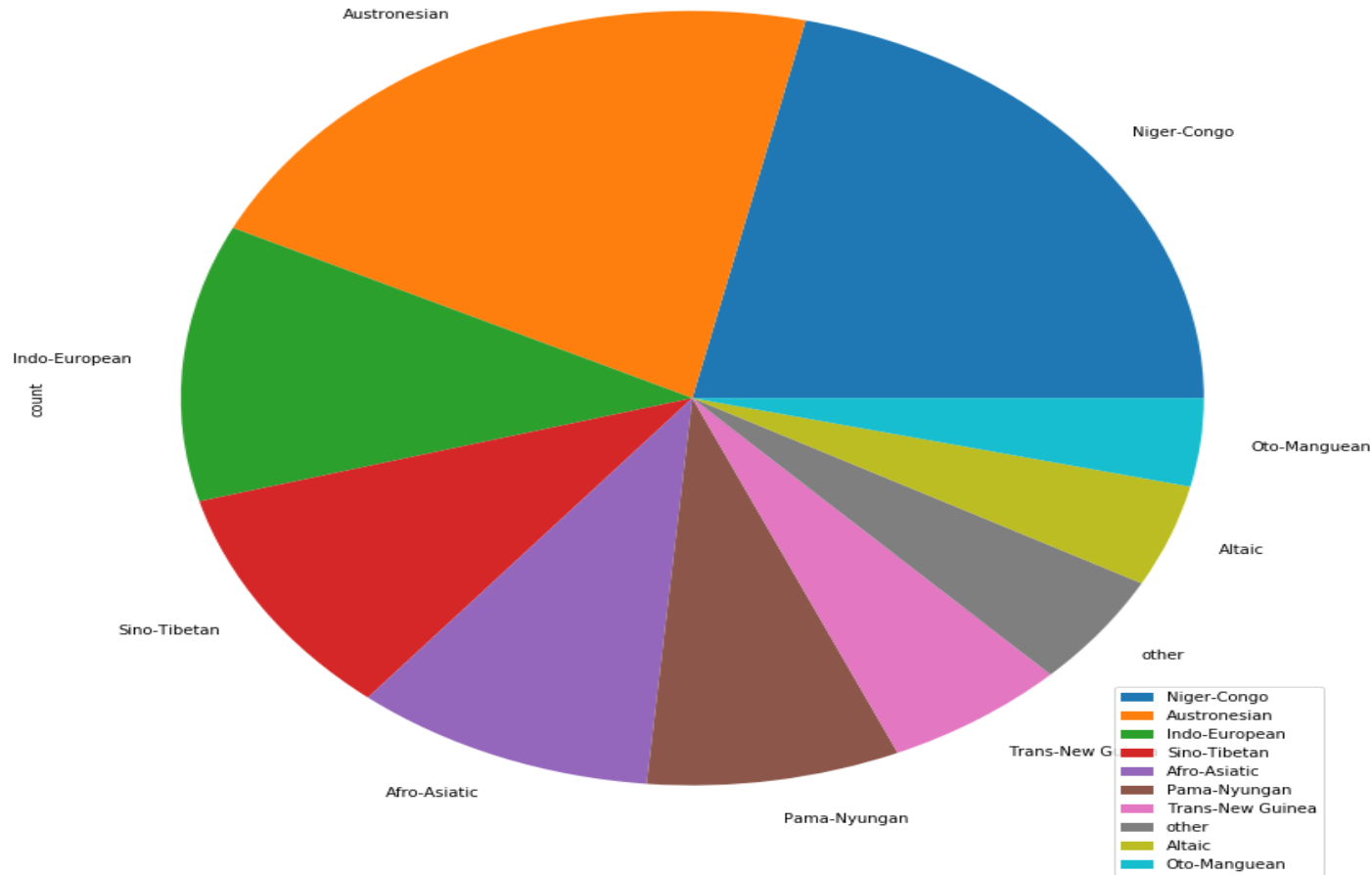
families counts of each respective individual language's family?

- The following Figures, provide the top twenty distributions of the Major Family Groupings, grouped by 'family' utilizing a groupby function, based upon the index of name= 'count', as demonstrated:

	family	count
158	Niger-Congo	327
14	Austronesian	325
86	Indo-European	176
185	Sino-Tibetan	149
0	Afro-Asiatic	145
167	Pama-Nyungan	122
215	Trans-New Guinea	88
255	other	72
5	Altaic	65
166	Oto-Manguean	56
13	Austro-Asiatic	49
62	Eastern Sudanic	47
225	Uto-Aztecan	44
138	Mayan	35
4	Algic	31
132	Mande	29
155	Nakh-Daghestanian	28
11	Arawakan	28
222	Uralic	27
37	Central Sudanic	26

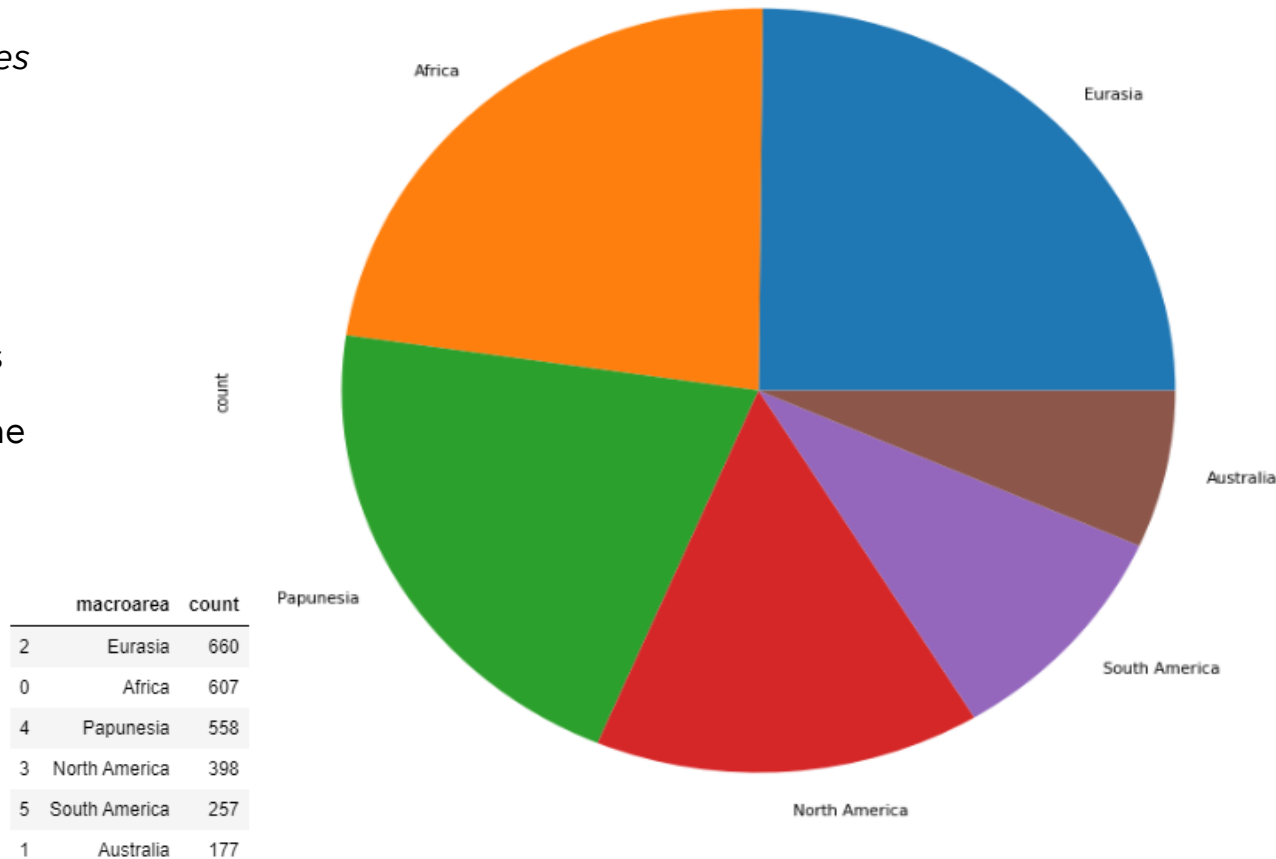
IST 652: Scripting for Data Analysis

- plot



IST 652: Scripting for Data Analysis

- *Can a Geographic spatial analysis of the top three Language Families be identified from the data?*
- Having analyzed the data and establishing what macroarea(s) contained what language family's frequencies, the next aspect of the research project is a geographic spatial analysis of the language families.

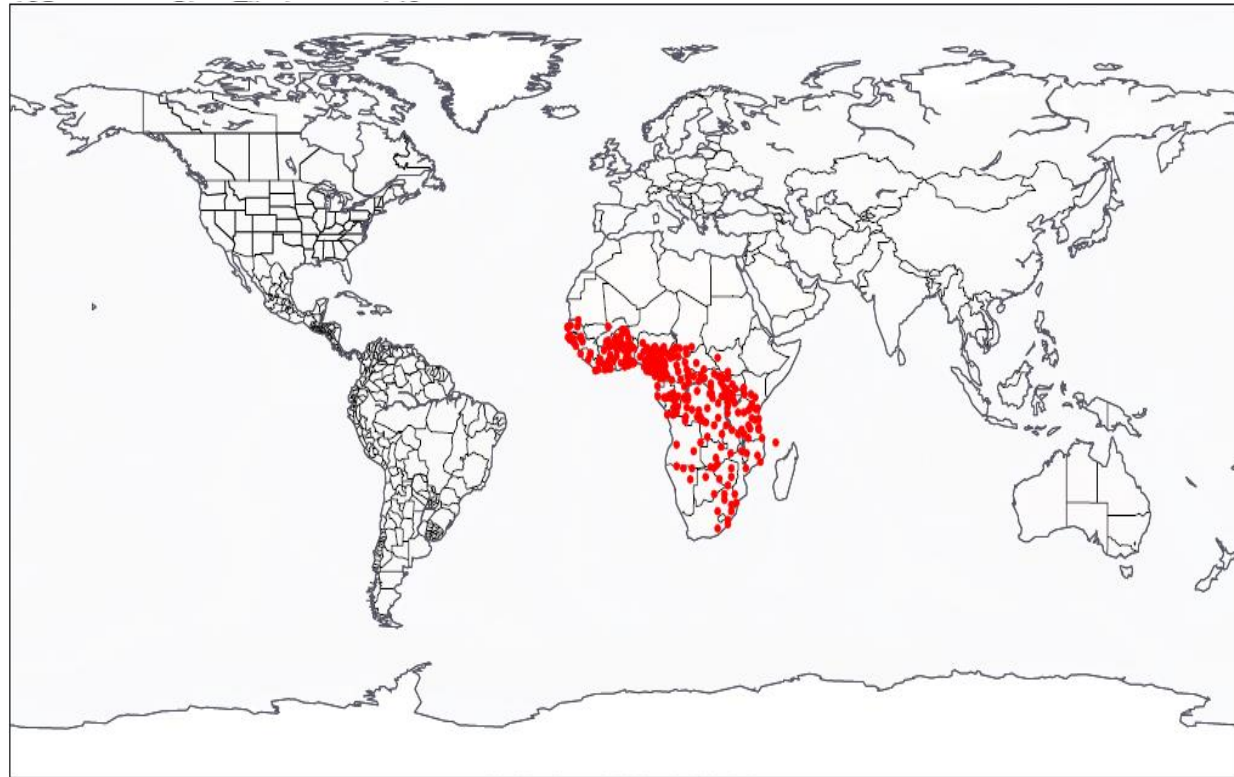


IST 652: Scripting for Data Analysis

- *No.1 Niger-Congo Language Family:*

- Geographic distribution is found throughout Africa; it is the world's third largest spoken language family, however, per the dataset, is first in the counts of individual language members to the respective language family. The Niger-Congo Language Family is first, in total individual members to their language family, which is intuitive, humans evolved within the Africa continent.

	family	count
158	Niger-Congo	327
14	Austronesian	325
86	Indo-European	176



Niger-Congo Language Family Geo-Spatial Distribution

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

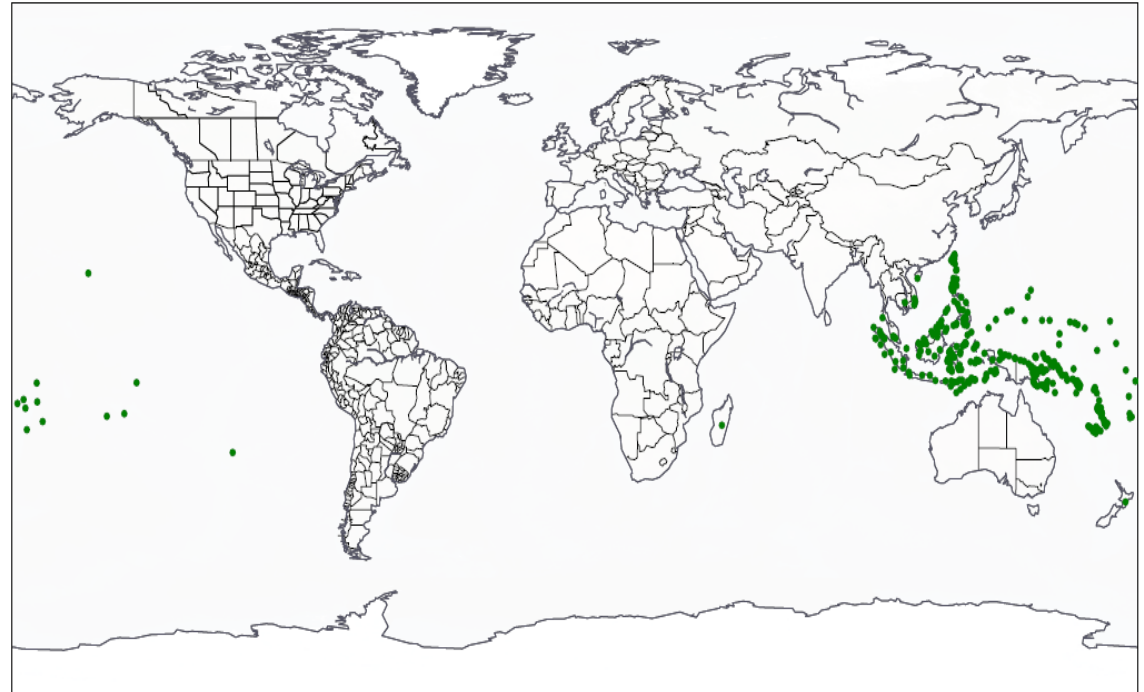
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 652: Scripting for Data Analysis

- *No.2 Austronesian Language Family:*

- Geographic distribution is found throughout: Taiwan, the Malay Peninsula, Maritime Southeast Asia, Madagascar, and the islands of the Pacific Ocean. It is the fifth-largest largest spoken language family, however, per the dataset, it is second in terms of the counts of individual language members to the respective language family. The numbers of individual languages to this language family follow close in-line with the Niger-Congo language family, which can also be seen as intuitive and interesting for the following reasons: the geographic area that this language family is distributed through is massive, thus, allowing for distance and time to create sub-dialects of the proto-language family inheritance, AND, as the first waves of human migration (DNA haplogroups F - Forward) initiated along the very migration pattern demonstrated by the languages geospatial analysis, as seen



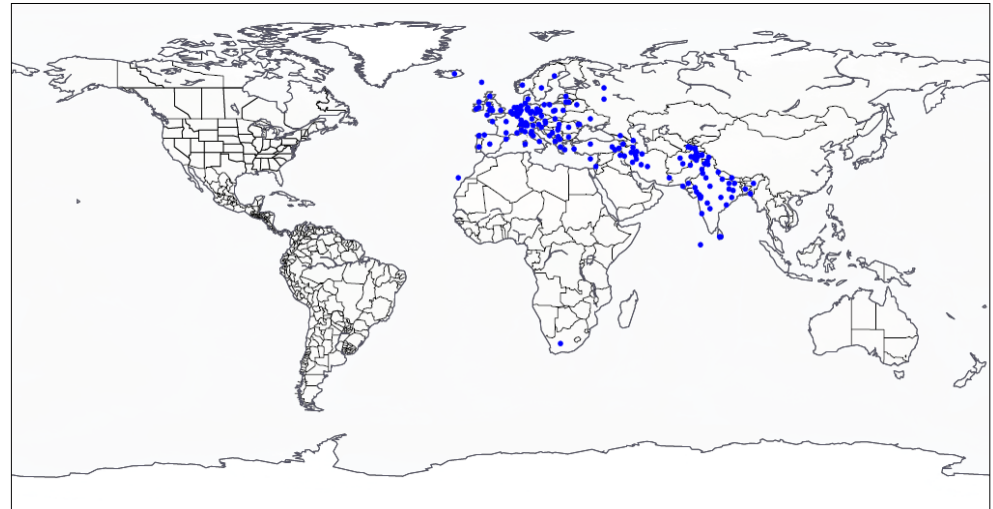
Austronesian Language Family Geo-Spatial Distribution

IST 652: Scripting for Data Analysis

•No.3 Indo-European Language Family

- Geographic distribution is found throughout: western Eurasia comprising most of the languages of Europe together with those of the Indian Subcontinent (mostly in the northern portions of the subcontinent) and the Iranian Plateau. It is the second-largest spoken language family, however, per the dataset, it is third in terms of the counts of individual language members to the respective language family. The Language families 'linguistic homeland,' is in dispute, but given the nature of similarities found throughout Latin and Sanskrit and their respective language subdivisions, anthropologist, linguist, and, archeologist are beginning to determine that the 'linguistic homeland,' of the ancient proto-Indo-European language family to be found in the Caucasus regions of southern Russia, therein radiating out westerly, and southerly through the passes of the Hindu Kush, into Northern India. Linguistic analysis of the various subdivisions within the language demonstrates this, as demonstrated

"father"	"brother"	Meaning:	Sanskrit	Latin:
<ul style="list-style-type: none"> ◦ <i>pitar</i> (Sanskrit) ◦ <i>pater</i> (Latin) ◦ <i>pater</i> (Greek) ◦ <i>padre</i> (Spanish) ◦ <i>pere</i> (French) ◦ <i>father</i> (English) ◦ <i>fadar</i> (Gothic) ◦ <i>faðir</i> (Old Norse) ◦ <i>vader</i> (German) ◦ <i>athir</i> (Old Irish--with loss of original consonant) 	<ul style="list-style-type: none"> ◦ <i>bhratar</i> (Sanskrit) ◦ <i>frater</i> (Latin) ◦ <i>phrater</i> (Greek) ◦ <i>frere</i> (French) ◦ <i>brother</i> (Modern English) ◦ <i>brothor</i> (Saxon) ◦ <i>bruder</i> (German) ◦ <i>broeder</i> (Dutch) ◦ <i>bratu</i> (Old Slavic) ◦ <i>brathair</i> (Old Irish) 	"three"	<i>trayas</i>	<i>tres</i>
		"seven"	<i>sapta</i>	<i>septem</i>
		"eight"	<i>ashta</i>	<i>octo</i>
		"nine"	<i>nava</i>	<i>novem</i>
		"snake"	<i>sarpa</i>	<i>serpens</i>
		"king"	<i>raja</i>	<i>regem</i>
		"god"	<i>devas</i>	<i>divus</i> ("divine")



Indo-European Language Family Geo-Spatial Distribution

IST 652: Scripting for Data Analysis



Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

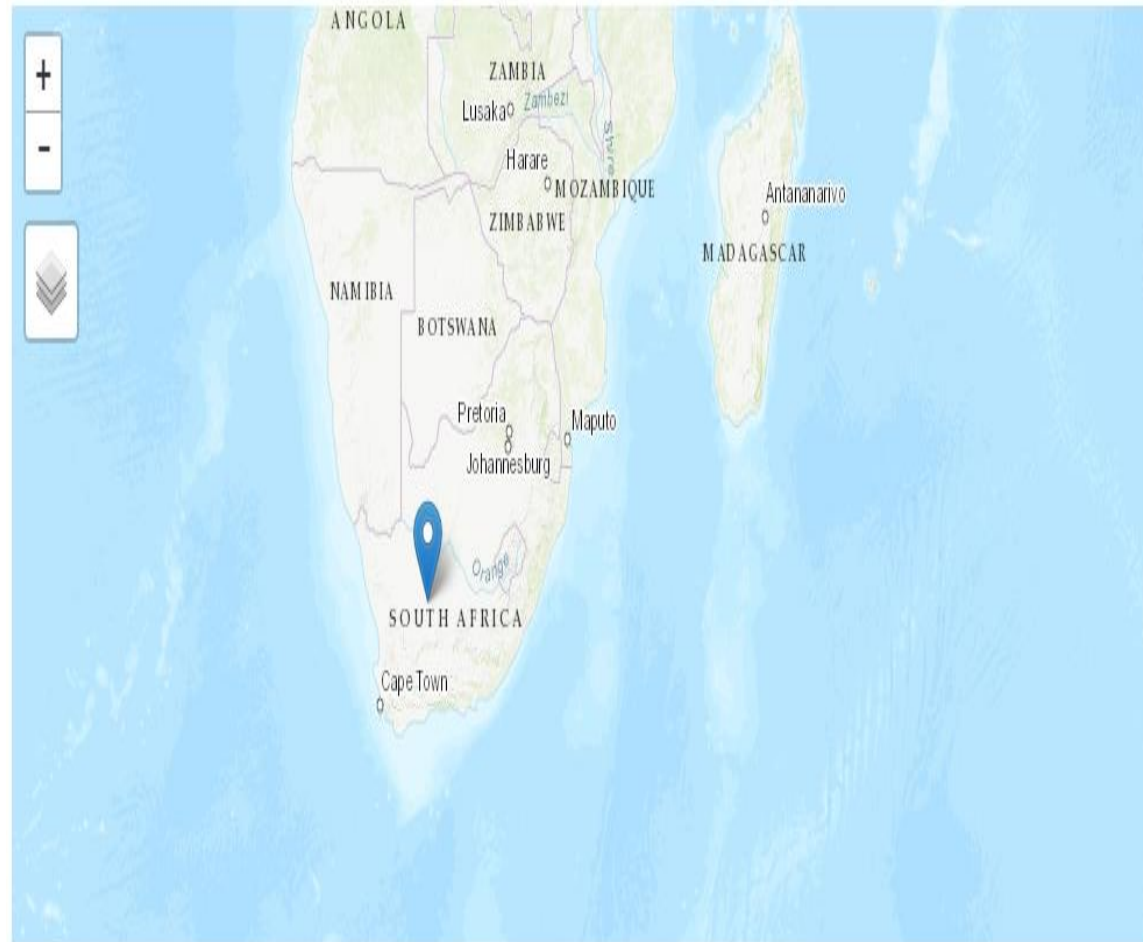
ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 652: Scripting for Data Analysis

- Interesting outlier
- ... last vestige of
- the Indo-European
- invasion sagas



Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ, TomTom, Intermap, iPC, USGS, FAO, NPS, NRCAN, GeoBase, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 652: Reflection and Learning Goals

- At the onset of the research project, the following questions were presented to be answered:
 - What is the frequency distribution of the language's, determinate upon their macroarea?*
 - What are the major language families counts of each respective individual language's family?*
 - Can a Geographic spatial analysis of the top three Language Families be identified from the data?*
- (
- From the dataset obtained, via semi-structured data and structured data, the research project was able to identify the frequency distribution of each language, determinate upon their macroarea, and the researcher was able to demonstrate the shape and size of that data. The research project was also able to identify the major language families counts of each respective individual language's family subdivisions. Finally, the dataset provided for the ability to render a geospatial analysis of the data contained within the dataset, it allowed for the researcher to plot the language families on a geographic representation of the earth, whereby the distribution of the family of languages, per their language family could be demonstrated by static and interactive mappings. In so doing, the research projects conclusion are as follows:

IST 652: Reflection and Learning Goals

•: the top three language family's totals are intuitive, given the historical, anthropological, and archeological studies that have been conducted, throughout the 19th, 20th, and 21st century. The language families, along with their specific subdivision can be visually demonstrated via geographic spatial analysis. The research projects goals of showing that distribution via the data, and expounding upon the researcher's native spoken language family, the Indo-European language family specific geo spatial analysis was accomplished.

In so much as the learning goals, per the call of the directive of the course,

- This was an individual project, based upon the work of Applied Data Science master's student Randall Scott Taylor, utilizing the brilliant work done by those at The World Atlas of Language Structures Online.
- The project began in late January 2020.
- The conclusion of this research project: March 11, 2020.

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

- :Project Description

- In response to the final project's requirements, and to answer the call of the questions presented, the candidate authored the following final report "*Final Project Report: Community Health Status Indicators.*" Description of the data and its source(s);

- CDC(Centers for Disease control and prevention) helps protect America from health, safety and security threats, both foreign and in the U.S. whether diseases start at home or abroad, are chronic or acute, curable or preventable, human error or deliberate attack, CDC fights disease and supports communities and citizens to do the same (Taylor, , "IST 707," 2020).

IST 707: Data Analytics

- :Project Description

- CDC increases the health security of our nation. As the nation's health protection agency, CDC saves lives and protects people from health threats. To accomplish our mission, CDC conducts critical science and provides health information that protects our nation against expensive and dangerous health threats and responds when these arise. Centers for Disease Control and Prevention (CDC) Community Health Status Indicators is a website that provides health profiles for all U.S. counties, including health outcomes, population health status, healthcare access and quality, health behaviors, social factors and the physical environment

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

- :Project Description

- The project objective was finding a data set in public health. We were looking for a data set that would help us better understand a broad set of health conditions, populations, and potential correlations. This project is an exercise in taking a large pool of data across a variety of metrics and generating relevant questions. Using tools and techniques learned in this course, we will use those insights, which could be used to drive actions for specific populations.

IST 707: Data Analytics

- :Project Description

- The data set we chose was the Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer that are major components of the Community Health Data Initiative. This dataset provides key health indicators for local communities and encourages dialogue about actions that can be taken to improve community health (e.g., obesity, heart disease, cancer).

- Website: <https://healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer>

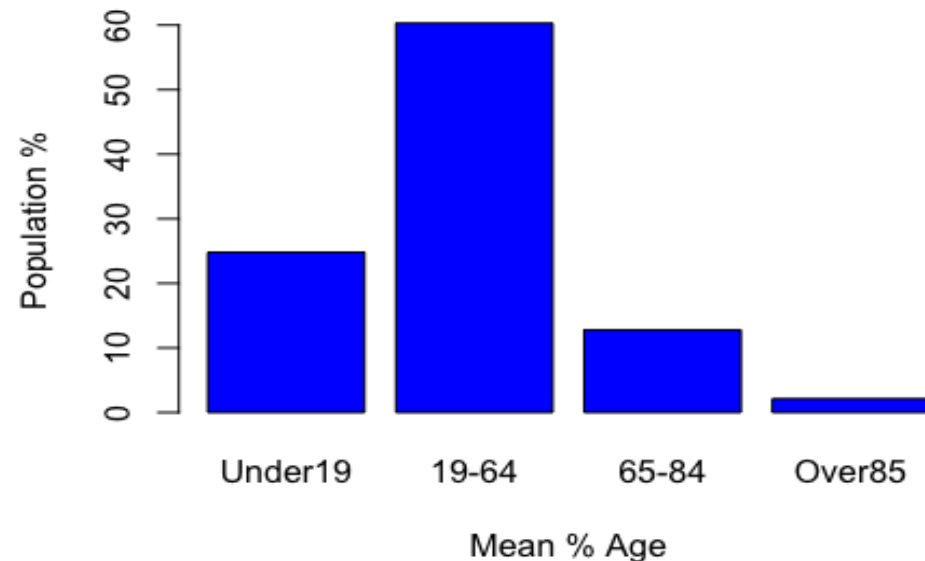
IST 707: Data Analytics

•Project Description

- To provide a robust dataset for our project, we chose a large health dataset containing 573 unique columns for every county in the United States. This broad scope of our dataset was so large that we needed to reduce it in order to focus on key health indicators.

- Next, we looked at another demographic variable, ethnicity. The combination of age/ethnicity would become important for the design of any health program or intervention that is targeted within a certain community

Figure 1 - National Age %



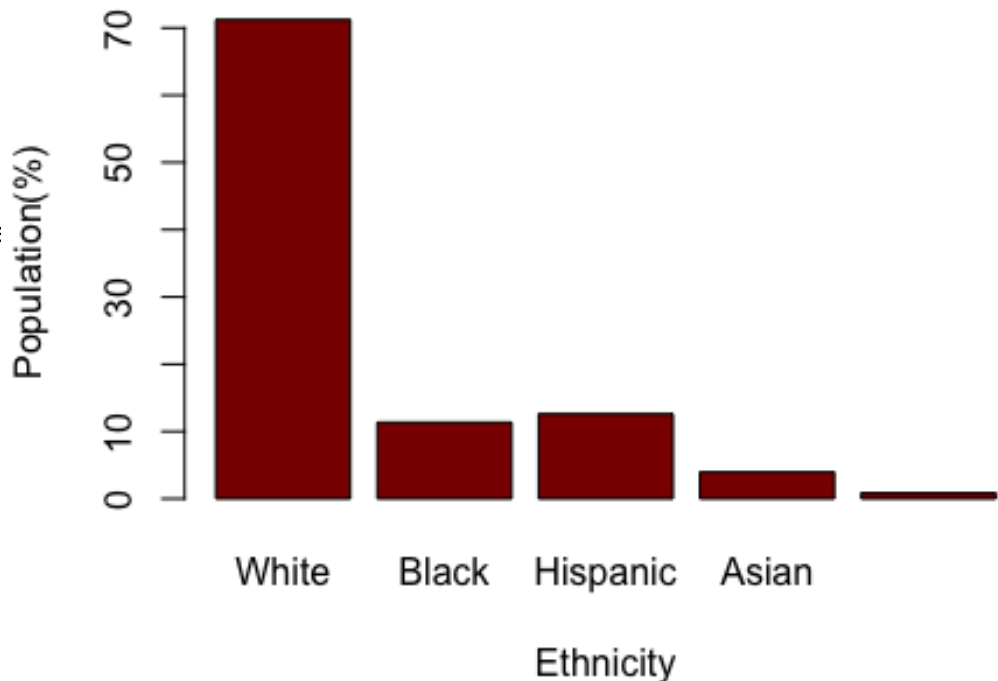
IST 707: Data Analytics

•Project Description

•To provide a robust dataset for our project, we chose a large health dataset containing 573 unique columns for every county in the United States. This broad scope of our dataset was so large that we needed to reduce it in order to focus on key health indicators.

- Next, we looked at another demographic variable, ethnicity. The combination of age/ethnicity would become important for the design of any health program or intervention that is targeted within a certain community

Figure 2 - National Mean Ethnicity



IST 707: Data Analytics

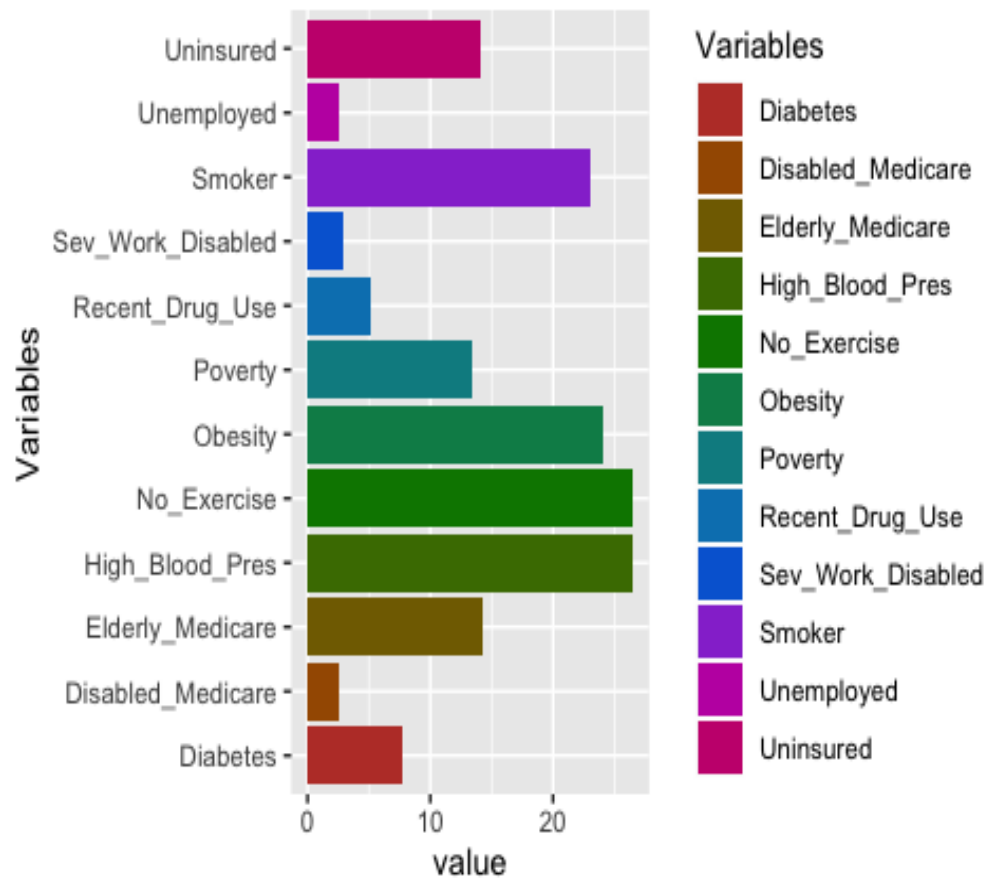
•Project Description

•Population Health Statistics

•The data set contained rich information pertaining to life expectancy and major risk factors contributing to the reduction in life expectancy. The goal is to use this data to better understand how to reduce the risk of premature death. In looking at the major contributors to premature death; High Blood pressure, No Exercise, Obesity, and Smoking led the way (Figure 3).

- It was analyzed to see what percent of the population of the death was caused due to Suicide and homicide

Figure 3 - Population Health Statistics



IST 707: Data Analytics

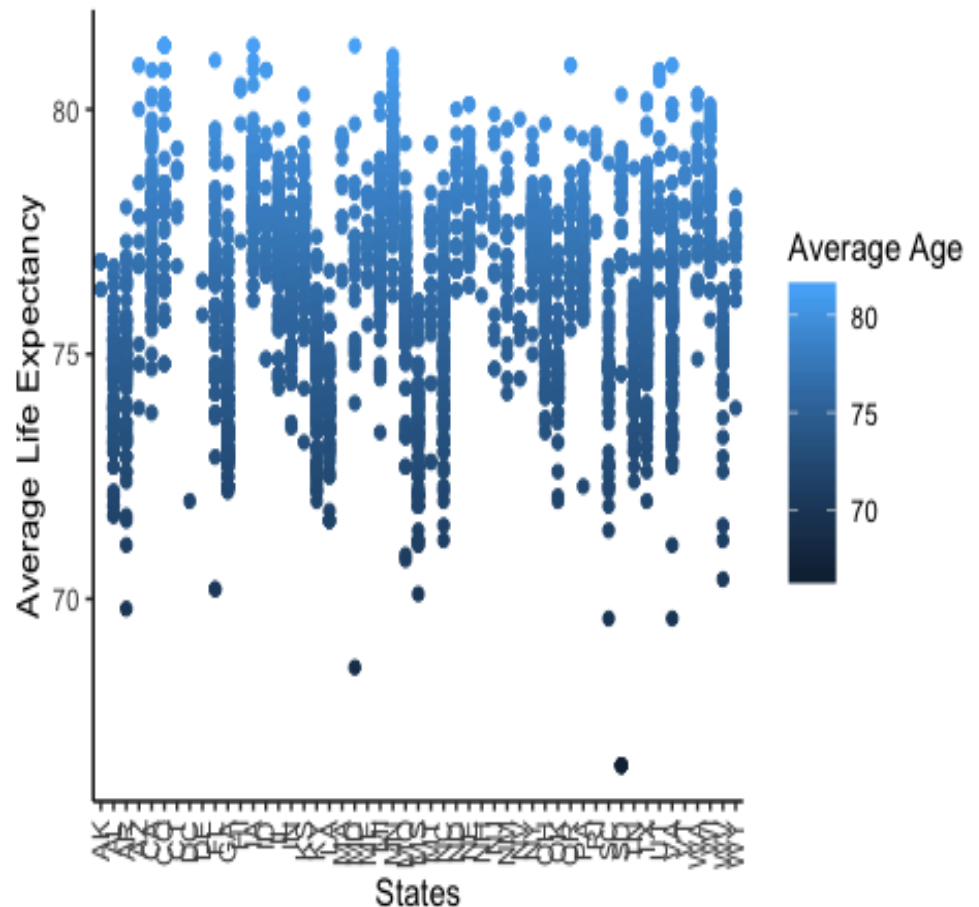
- Project Description

- Population Health Statistics

- The data set contained rich information pertaining to life expectancy and major risk factors contributing to the reduction in life expectancy. The goal is to use this data to better understand how to reduce the risk of premature death. In looking at the major contributors to premature death; High Blood pressure, No Exercise, Obesity, and Smoking led the way (Figure 3).

- It was analyzed to see what percent of the population of the death was caused due to Suicide and homicide

Figure 4 - State Average Life Expectancy

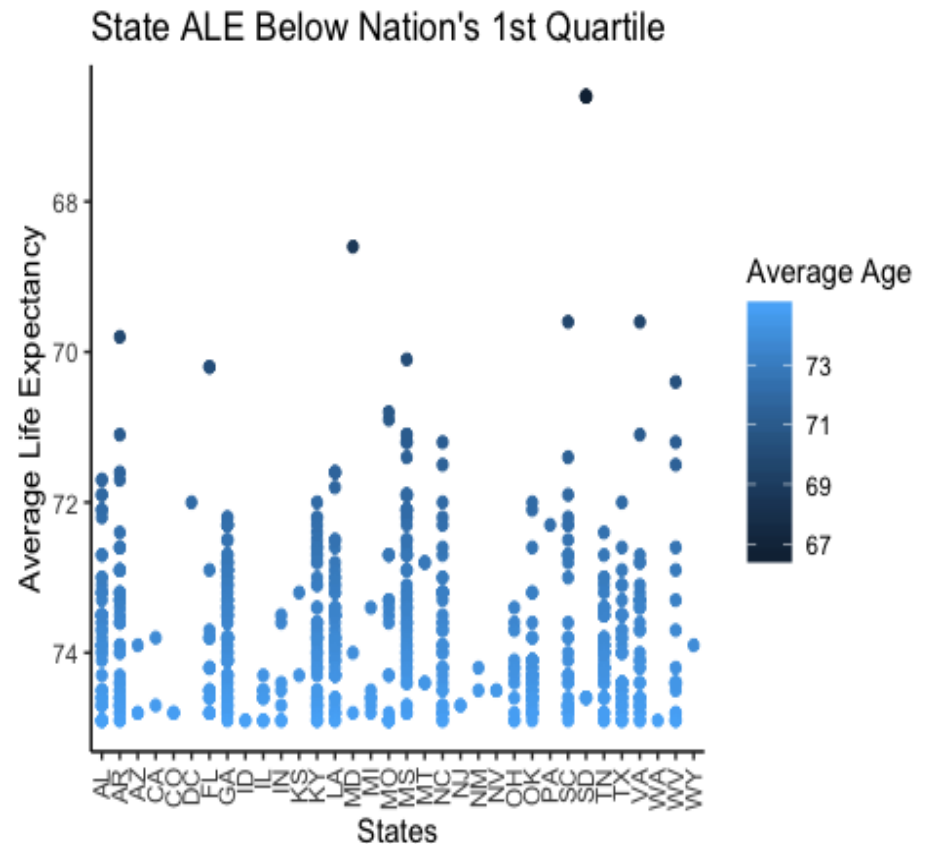


IST 707: Data Analytics

•Project Description

•Population Health Statistics

- The data set contained rich information pertaining to life expectancy and major risk factors contributing to the reduction in life expectancy. The goal is to use this data to better understand how to reduce the risk of premature death. In looking at the major contributors to premature death; High Blood pressure, No Exercise, Obesity, and Smoking led the way (Figure 3).
- It was analyzed to see what percent of the population of the death was caused due to Suicide and homicide



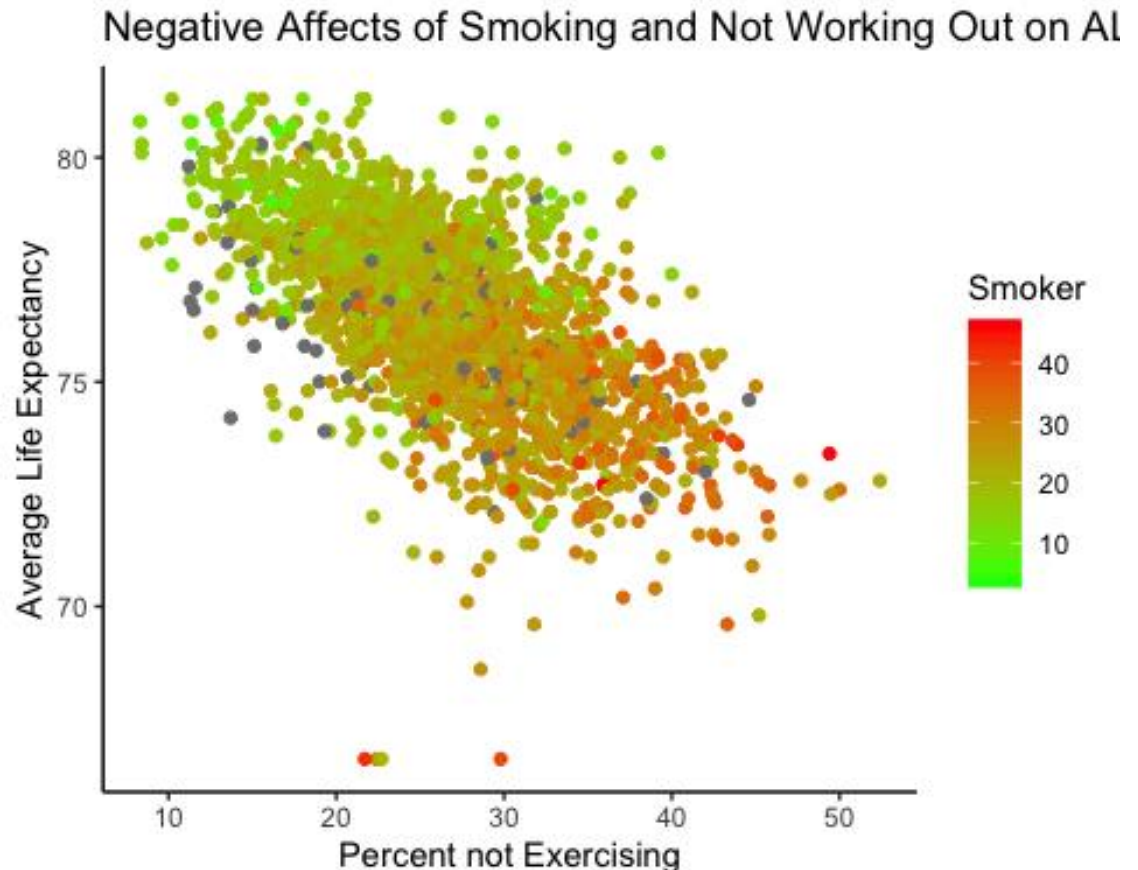
IST 707: Data Analytics

- Project Description

- Population Health Statistics

- The data set contained rich information pertaining to life expectancy and major risk factors contributing to the reduction in life expectancy. The goal is to use this data to better understand how to reduce the risk of premature death. In looking at the major contributors to premature death; High Blood pressure, No Exercise, Obesity, and Smoking led the way (Figure 3).

- It was analyzed to see what percent of the population of the death was caused due to Suicide and homicide

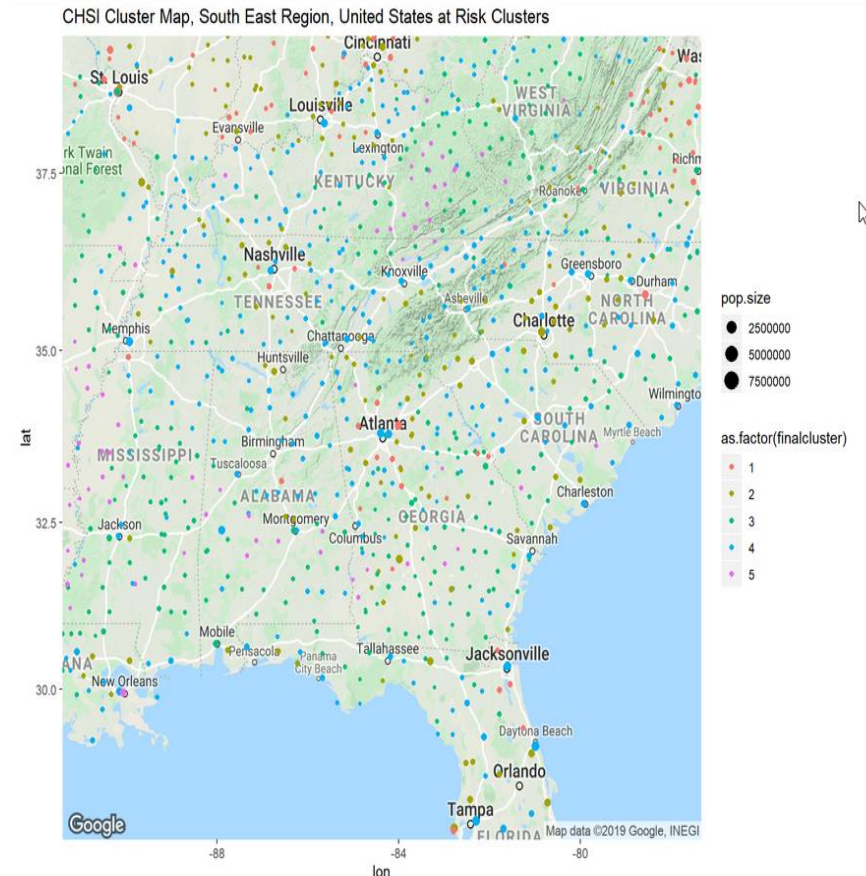


IST 707: Data Analytics

- Project Description

- Population at Risk Clusters

•To assist in the relationship of this vast amount of data, and the understanding of how the data correlates and relates to one another, K-means clustering was done, based upon geographic analysis, and, interactive maps were created by the candidate to assist with the intake of this information, as referenced below.

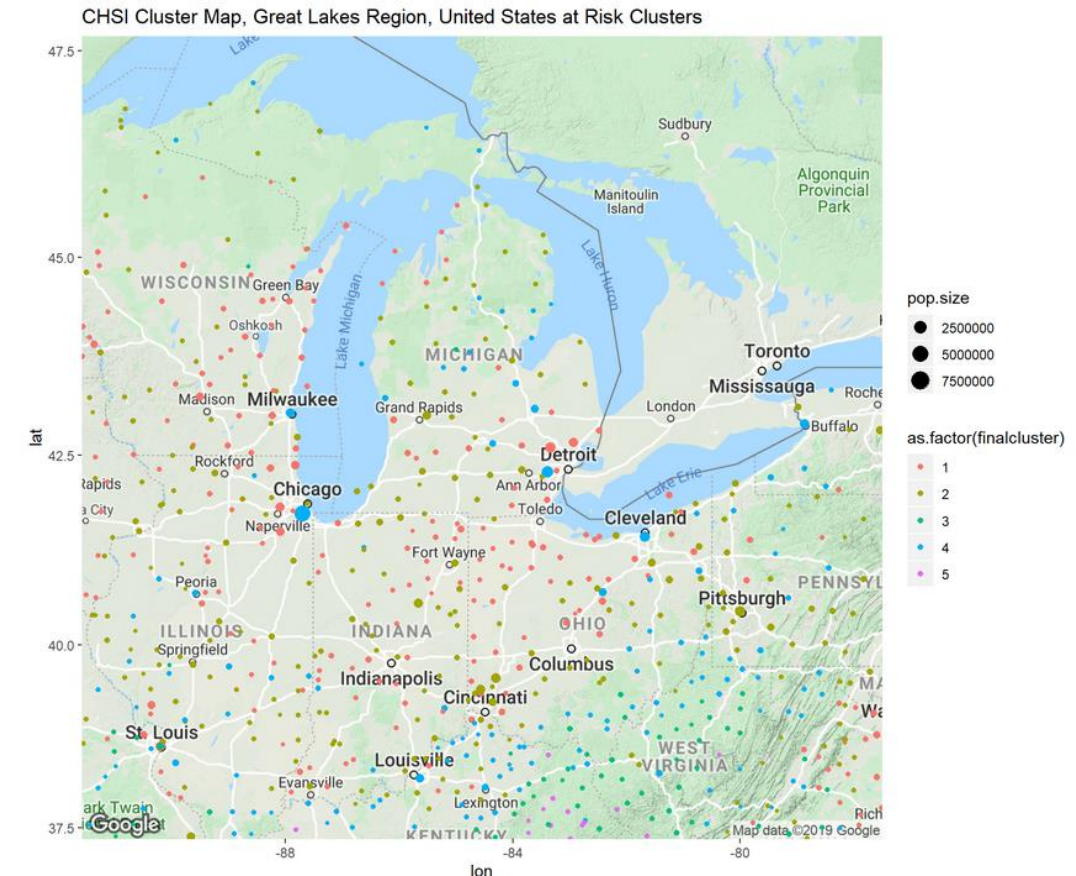


IST 707: Data Analytics

•Project Description

•Population at Risk Clusters

•To assist in the relationship of this vast amount of data, and the understanding of how the data correlates and relates to one another, K-means clustering was done, based upon geographic analysis, and, interactive maps were created by the candidate to assist with the intake of this information, as referenced below.

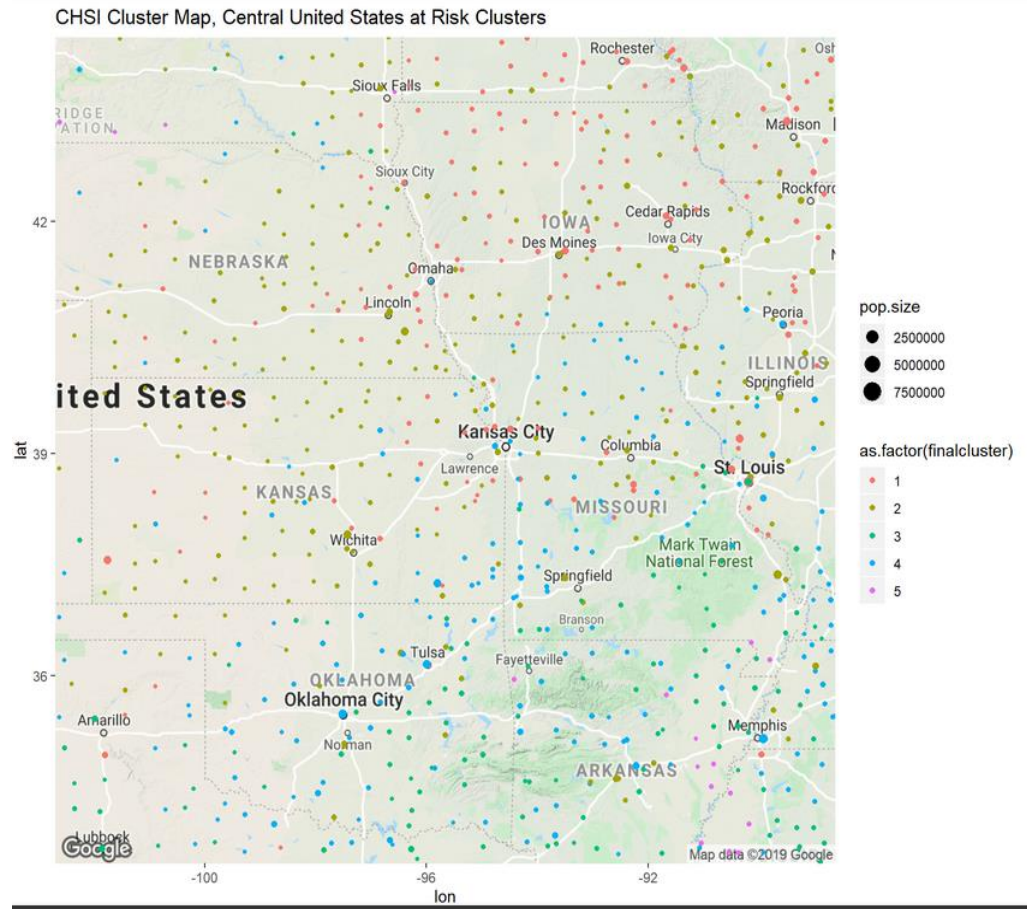


IST 707: Data Analytics

•Project Description

•Population at Risk Clusters

•To assist in the relationship of this vast amount of data, and the understanding of how the data correlates and relates to one another, K-means clustering was done, based upon geographic analysis, and, interactive maps were created by the candidate to assist with the intake of this information, as referenced below.

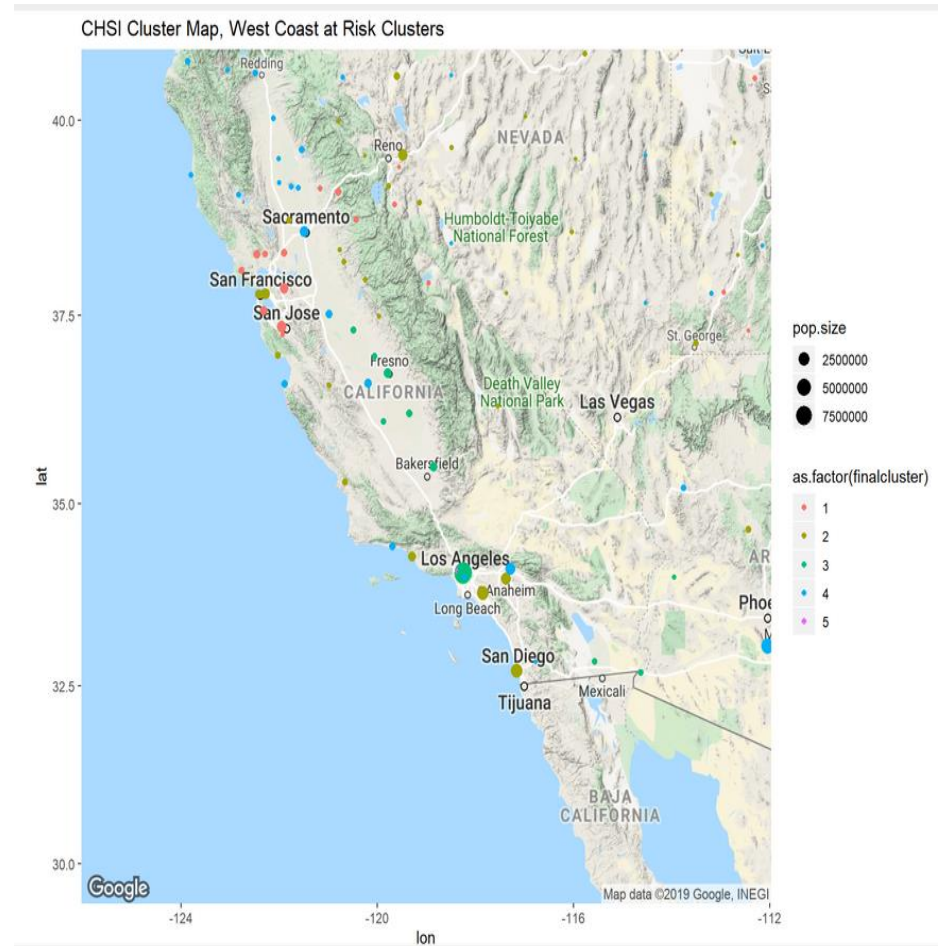


IST 707: Data Analytics

- Project Description

- Population at Risk Clusters

•To assist in the relationship of this vast amount of data, and the understanding of how the data correlates and relates to one another, K-means clustering was done, based upon geographic analysis, and, interactive maps were created by the candidate to assist with the intake of this information, as referenced below.

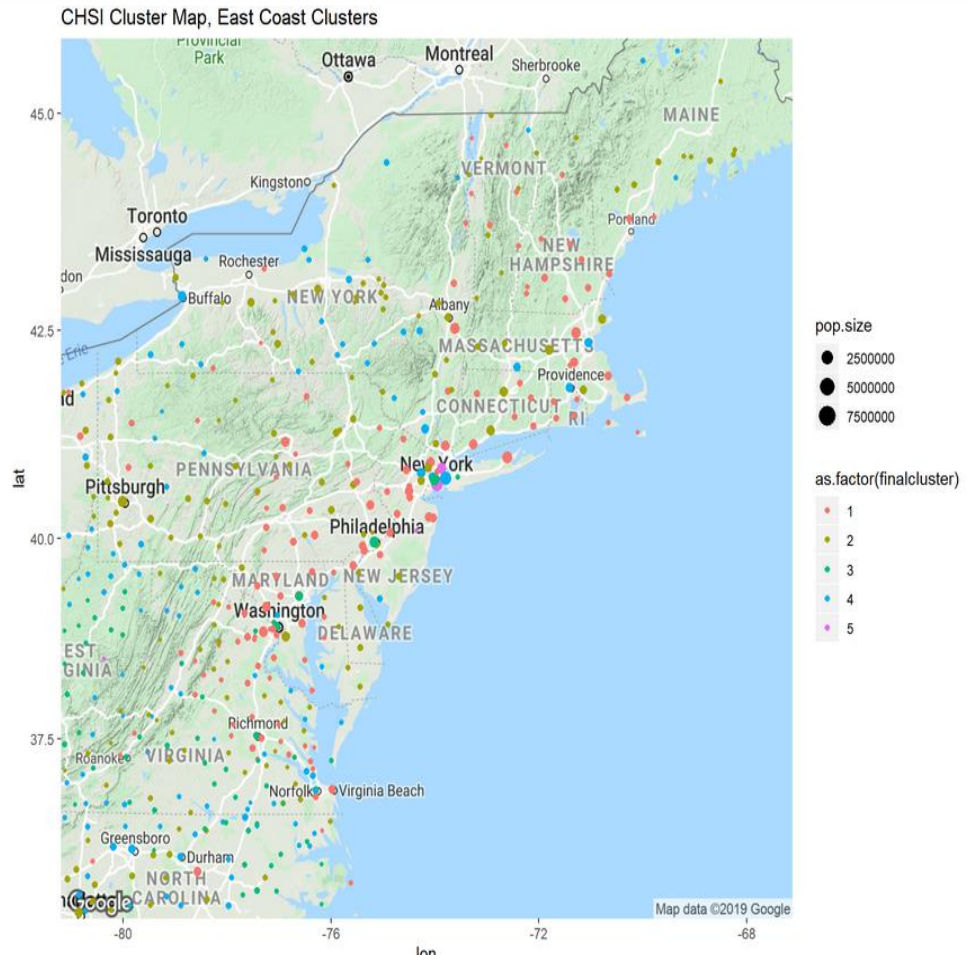


IST 707: Data Analytics

•Project Description

•Population at Risk Clusters

•To assist in the relationship of this vast amount of data, and the understanding of how the data correlates and relates to one another, K-means clustering was done, based upon geographic analysis, and, interactive maps were created by the candidate to assist with the intake of this information, as referenced below.

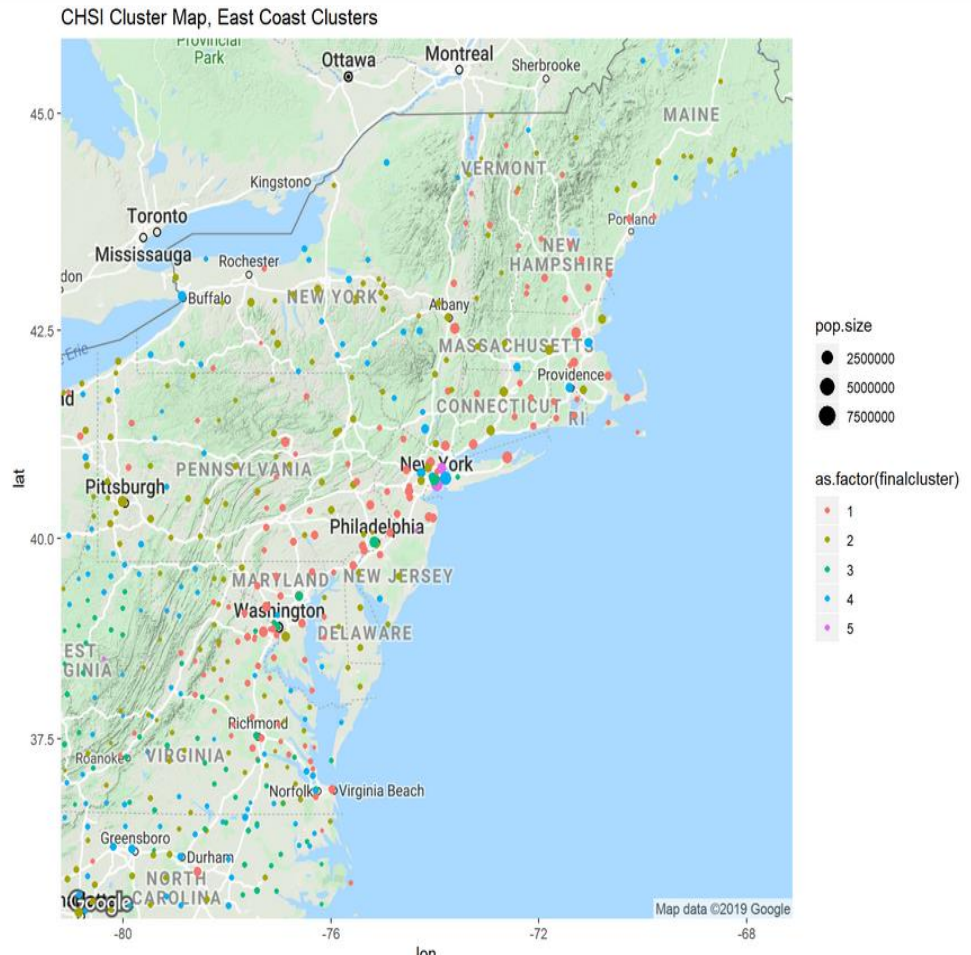


IST 707: Data Analytics

•Project Description

•Population at Risk Clusters

•To assist in the relationship of this vast amount of data, and the understanding of how the data correlates and relates to one another, K-means clustering was done, based upon geographic analysis, and, interactive maps were created by the candidate to assist with the intake of this information, as referenced below.



IST 707: Data Analytics

•Project Description

•Reviewing the assignment step, the algorithm computes the new mean value of each cluster. The term cluster “centroid update” is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means. The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration

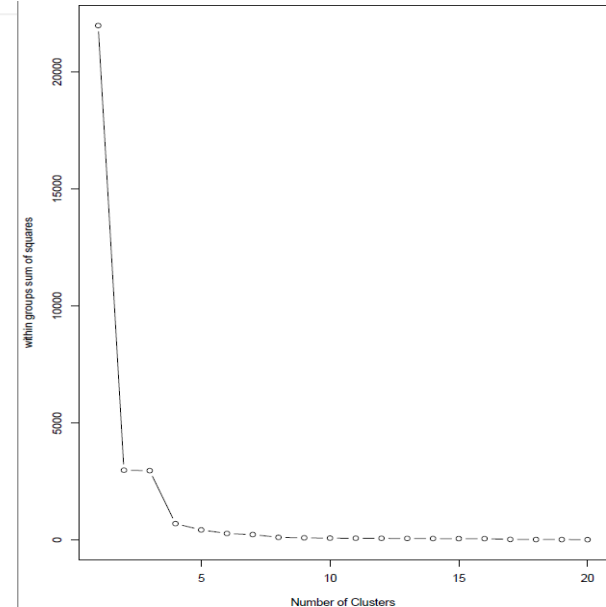
AK 11 7 5 3 1

```
1 2 3 4 5
AK 11 7 5 3 1
AL 1 6 20 32 8
AR 0 4 36 27 8
AZ 0 2 6 6 1
CA 11 17 9 21 0
CO 17 24 8 14 1
CT 6 2 0 0 0
DC 0 0 1 0 0
DE 1 2 0 0 0
FL 6 19 15 27 0
GA 10 32 68 41 8
HI 0 3 0 2 0
IA 56 41 0 2 0
ID 3 17 0 24 0
IL 29 49 3 21 0
IN 43 44 0 5 0
KS 19 65 0 21 0
KY 4 19 29 50 18
LA 0 5 33 14 12
MA 7 5 0 2 0
MD 13 7 2 2 0
ME 3 9 0 4 0
MI 21 43 1 18 0
MN 57 27 0 3 0
MO 16 31 19 48 1
MS 1 1 43 20 17
MT 0 18 12 23 3
NC 4 24 24 48 0
ND 12 30 1 8 2
NE 22 60 1 10 0
NH 9 1 0 0 0
NJ 12 6 0 3 0
NM 1 2 18 7 5
NV 2 11 0 4 0
NY 8 28 1 23 2
OH 39 32 2 15 0
OK 0 14 28 35 0
OR 4 11 1 20 0
PA 21 38 1 7 0
RI 4 0 0 1 0
SC 0 10 13 21 2
SD 5 36 4 12 9
TN 3 18 12 60 2
TX 10 36 92 95 21
UT 9 13 1 6 0
VA 43 33 25 33 0
VT 6 7 0 1 0
WA 4 16 2 17 0
WI 45 25 1 1 0
WV 0 8 23 21 3
WY 5 13 0 5 0
```

=====
randIndex: measure between state and cluster partitions.

Note: values vary between -1 to 1

=====
ARI
0.02352161



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

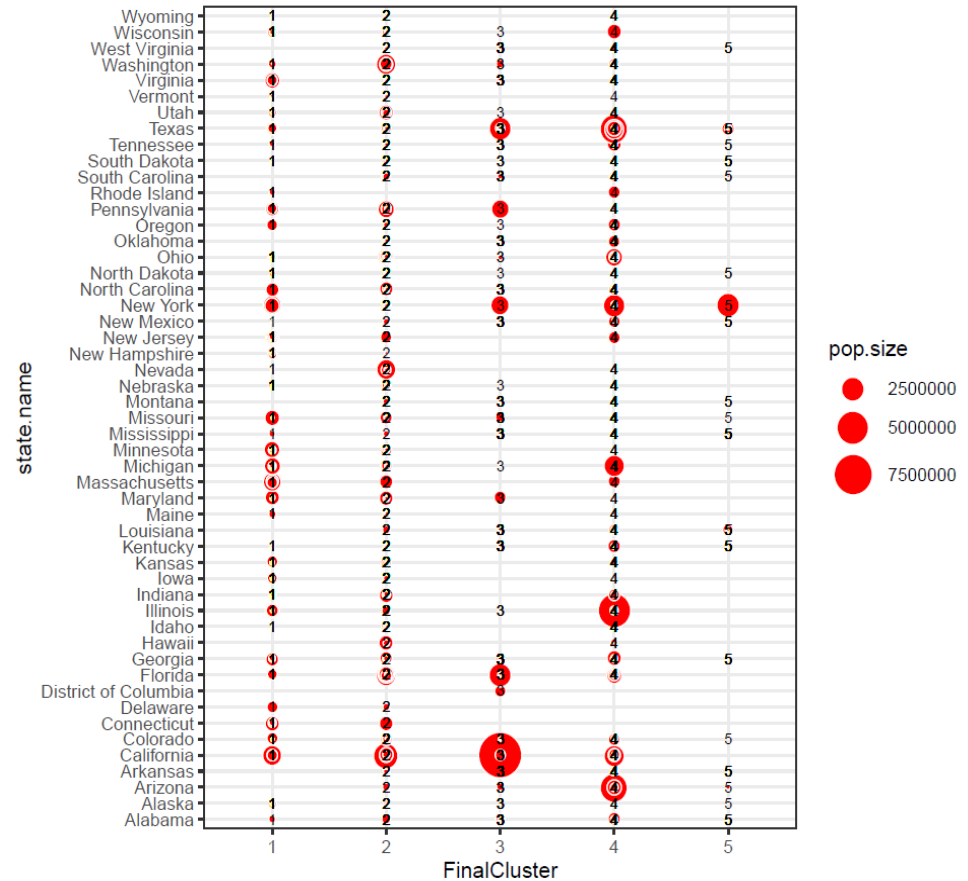
iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

•Project Description

- Reviewing the assignment step, the algorithm computes the new mean value of each cluster.

The term cluster “centroid update” is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means. The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration



IST 707: Data Analytics

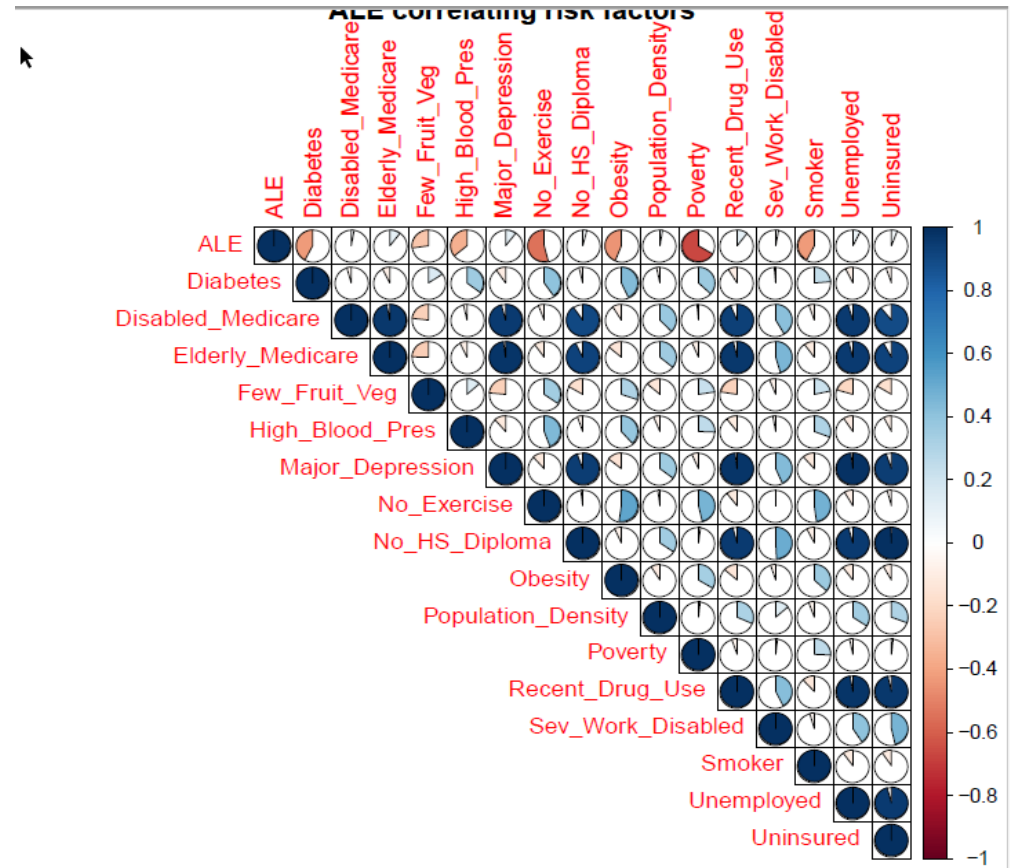
- Project Description

- Data Models

- Linear regression

•The first step that was taken in the process of transition from “EDA,” by the researcher(s) was to identify: the basic health variables within the data that had to due with negative correlations to life expectancy, to identify and therein see if there was any basic linearity between them.

The researcher(s) chose the following correlation matrix visualization to demonstrate the chosen correlation groups; from the CORRPLOT package:



IST 707: Data Analytics

•Reflection & Learning Goals

•The results from linear regression model gave an understanding that 71.2% of the variability in ALE could be explained by the below factors which were used to build a model. The maximum impact being in the order of Ethnicity, Poverty, HIV, Obesity and Blood Pressure.

•Correlation analysis revealed that having No_highschool diploma is highly related to depression, being uninsured, unemployed, use of drugs that impact ALE strongly. Kmeans Result helped us understand segments in our population data Decision Tree models provided almost the same accuracy of ~92% with major decision-making attributes being Poverty, Ethnicity, Exercising and Smoking.

Poverty	-1.7E-01
White	5.6E-02
Black	2.3E-02
Native_American	6.8E-02
Asian	1.1E-01
Hispanic	3.4E-02
Recent_Drug_Use	1.6E-07
Toxic_Chem	-2.5E-09
No_Exercise	-4.2E-02
Few_Fruit_Veg	-1.2E-02
Obesity	-2.1E-02
High_Blood_Pres	-2.5E-02
Smoker	-5.0E-02
Diabetes	-4.6E-02
HIV	-1.7E-02
E_HeartDis	-9.9E-03
F_HeartDis	-2.2E-03

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Reflection & Learning Goals

•Conclusions:

•Conclusions

•1. The data set we chose really does fit its intended purpose, which was to assist local health agencies with assessing the needs of their communities. In addition, armed with this data they would be able to create programs and services that would directly impact the overall health of their communities.

•2. Irrespective of how you cut the data, we saw that a lack of education (defined as no HS diploma) had the single largest impact on overall health. Though it wasn't directly significant in the linear model, it had a high correlation to things like unemployment, drug use, and depression. These in turn contribute to a lower Average Life Expectancy (ALE). Programs that target education and/or gainful employment, especially in rural counties, would seem to have the largest impact.

•3. In addition, communities with adverse behavioral or lifestyle choices (most notably those who don't exercise, those who eat few fruits/vegetables, and those who smoke) are statistically more at risk for premature death. These correlations (negative correlative value to life expectancy) are within the individual's control and would benefit from additional support within the community.

•4. A general observation of the project is that while we chose a data set that allowed for each individual to learn something or probe in a different direction (e.g. some thinking about cancer, some looking at mental health and others suicide rates) it created a challenge in focusing in on a cohesive data story. The team was often caught between applying things we had learned in class (tools - ensure we "check all the right boxes") and really understanding what the data was telling us. It was a great exercise to highlight the challenges in translating business needs (what do you want to know, how do you want to use the data) and the data side (coding) to ensure they are aligned.

Poverty	-1.7E-01
White	5.6E-02
Black	2.3E-02
Native_American	6.8E-02
Asian	1.1E-01
Hispanic	3.4E-02
Recent_Drug_Use	1.6E-07
Toxic_Chem	-2.5E-09
No_Exercise	-4.2E-02
Few_Fruit_Veg	-1.2E-02
Obesity	-2.1E-02
High_Blood_Pres	-2.5E-02
Smoker	-5.0E-02
Diabetes	-4.6E-02
HIV	-1.7E-02
E_HeartDis	-9.9E-03
F_HeartDis	-2.2E-03

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

IST 718: Big Data

•Project Description

- Covid-19 has created a global health crisis unlike any since the flu pandemic of 1918. In addition to the record number of deaths across the globe, the virus has caused society to change socially and economically through social distancing measures. To better understand this global impact, this project will aim to answer several data questions.
- Can we predict whether a county is more at risk for Covid-19 infection? Can this be clustered and analyzed geospatially?
- Can a time series data be used for modeling and forecasting to predict the trajectory of cases and deaths in the future?*
- How is the American lifestyle being affected by Covid-19 and what is being done about it?*
- What has been published about ethical and social science considerations regarding Covid-19?*
- What is being done in terms of research/study to understand and combat this virus?*

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

- :Project Description

- Population Health Statistics*

- Randall Taylor** - *Utilizing New York Times Covid-19 Data and CDC CHSI data can we predict whether a county is more at risk for Covid-19 infection? Can this be clustered and analyzed geospatially?*

MEMBERS OF THE TEAM	CONTRIBUTION
Patricia A. Mills	sWhat has been published about ethical and social science considerations regarding Covid-19?
Jose Conrado T Reyes	Can time series data from the CDC, JHU, and the COVID Tracking Project be used to forecast the trajectory of Covid-19 cases and deaths in the future?
Thomas Bahng	How has Covid-19 impacted us socially and the Academic research to address the virus?
Randall Scott Taylor	Utilizing New York Times Covid-19 Data and CDC CHSI data can we predict whether a county is more at risk for Covid-19 infection? Can this be clustered and analyzed geospatially?

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•:Project Description

- At the time of the completion of this paper, the current final project is the last two weeks of culminating into a final presentation. The candidate will discuss the current work and the findings, to date.
- The libraries that were utilized to rendered analysis are many:

```
#import packages for analysis and modeling

import pandas as pd # dataframe operations#####
from pandas.io.json import json_normalize#####
import numpy as np # arrays and math functions#####
from scipy.stats import uniform # for training and test split#####
import statsmodels.api as sm # statistical models (regression)#####
import statsmodels.formula.api as smf # for R likened specifications####
#####
import addfips # for the import of proper fips coding IMPORTANTE#####

import matplotlib
from heatmap import heatmap, corrplot

matplotlib.use('Agg')
matplotlib.style.use('ggplot')

import matplotlib.pyplot as plt # 2D plotting (very 2010 )
import seaborn as sns #provides trellis, small multiple plotting (not my favorite)
from scipy import stats
from statsmodels.formula.api import ols
import scipy.stats as stats
from sklearn import linear_model
from shapely.geometry import Point, Polygon
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import decimal
from fuzzywuzzy import fuzz
from fuzzywuzzy import process
from string import ascii_letters
#import chart_studio.plotly as py -- next time (keep as a template install )
#import plotly.graph_objs as go4 -- next time (keep as a template install )
```

IST 707: Data Analytics

•Project Description

•This is required for the interactive nature of the visualizations that follow throughout the extraction, transformation, loading of the initial datasets. The further exploratory data analysis – this is where the keen use of interactive plotting can expedite the analysis of the data, initially, to allow for better decisions to be made – regarding the direction of UNSUPERVISED MACHINE LEARNING techniques, to a furtherance of SUPERVISED MACHINE LEARNING techniques.

```
#geospatial analysis
import plotly.figure_factory as ff
import plotly.express as px
import plotly.graph_objects as go
from plotly.figure_factory._county_choropleth import create_choropleth
from plotly.offline import iplot
plt.style.use('fivethirtyeight')

#import plotly-geo
from sklearn.preprocessing import StandardScaler
#from mpl_toolkits.mplot3d import Axes3D
#from mpl_toolkits.basemap import Basemap
from geopandas import GeoDataFrame
from shapely.geometry import Point
#from ipyleaflet import *
#from ipyleaflet import Map, GeoData, basemaps, LayersControl
#import geopandas
import folium
#from ipyleaflet import Map, GeoData, basemaps, LayersControl
import geopandas
import json
import urllib.request
from urllib.request import Request, urlopen
from urllib.request import urlopen as req
import addfips
from bs4 import BeautifulSoup as soup
#import csv
#from urllib.request import Request, urlopen
#from urllib.request import urlopen as req
#from bs4 import BeautifulSoup as soup
#from autoplotter import run_app #GUI_Based EDA

#pull in datasets GitHub Repository
!git clone https://github.com/randallscott25/BigData
!git clone https://github.com/nytimes/covid-19-data/
```

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

- :Project Description

- Three datasets

- Fips

- New York Times git

- CDC CHSI

	County_FIPS_Code	State_FIPS_Code	CHSI_County_Name	CHSI_State_Name	CHSI_State_Abbr	Population_Size	Population_Density	Poverty	Age_19_Under	Age_19_64	Age_65_84	Age_85_and_Over	White	Black	Native_American	Asian
0	1	1	Autauga	Alabama	AL	48612	82.0	10.4	26.9	62.3	9.8	0.9	80.7	17.3	0.5	0.6
1	3	1	Baldwin	Alabama	AL	162586	102.0	10.2	23.5	60.3	14.5	1.8	88.4	9.9	0.5	0.4
2	5	1	Barbour	Alabama	AL	28414	32.0	22.1	24.3	62.5	11.6	1.6	52.2	46.8	0.4	0.3
3	7	1	Bibb	Alabama	AL	21516	35.0	16.8	24.6	63.3	10.9	1.2	76.8	22.5	0.3	0.1
4	9	1	Blount	Alabama	AL	55725	86.0	11.9	24.5	62.1	12.1	1.3	97.1	1.5	0.5	0.2
...
3136	37	56	Sweetwater	Wyoming	WY	37975	4.0	8.6	26.6	65.1	7.4	0.9	95.5	1.1	1.1	1.0
3137	39	56	Teton	Wyoming	WY	19032	5.0	5.6	18.8	73.3	7.5	0.4	97.9	0.2	0.4	0.8
3138	41	56	Uinta	Wyoming	WY	19939	10.0	10.6	29.1	63.1	7.0	0.8	97.5	0.1	1.1	0.3
3139	43	56	Washakie	Wyoming	WY	7933	4.0	11.1	23.5	59.5	14.6	2.3	97.2	0.2	0.8	0.7
3140	45	56	Weston	Wyoming	WY	6671	3.0	9.9	20.1	63.2	14.4	2.4	97.6	0.1	1.4	0.2

9141 rows x 16 columns

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

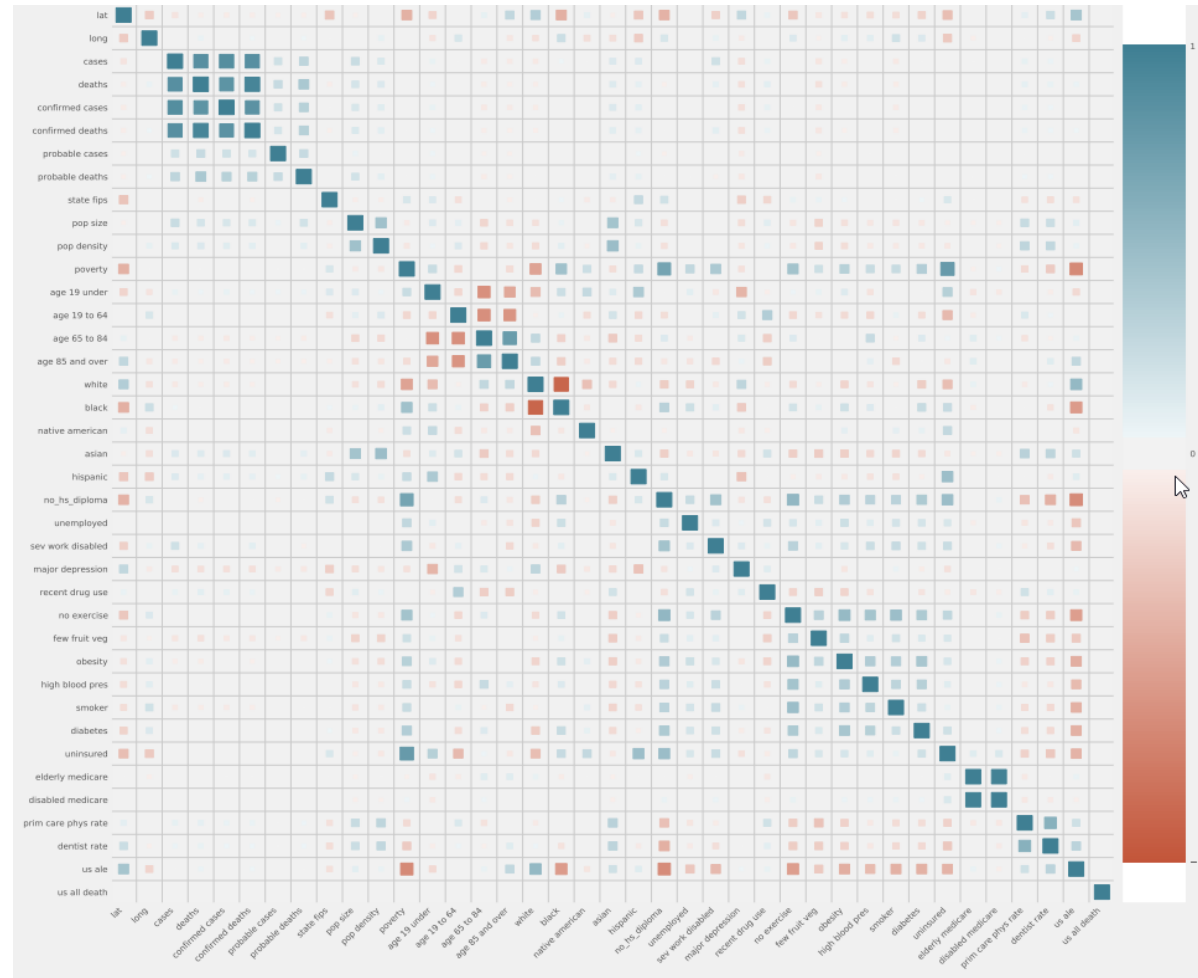
ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

- :Project Description
- All three datasets were combined into one dataset to enable visualization of the exploratory process of the dataset, and to further assist in modeling. The following correlation plot was created, for all potential features within the dataset:



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

- Project Description
- Initial visualizations of the linearity between the chosen features, and the target variable, cases, where create to begin to direct the candidate as to how to approach modeling, and what features to select in that model:



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

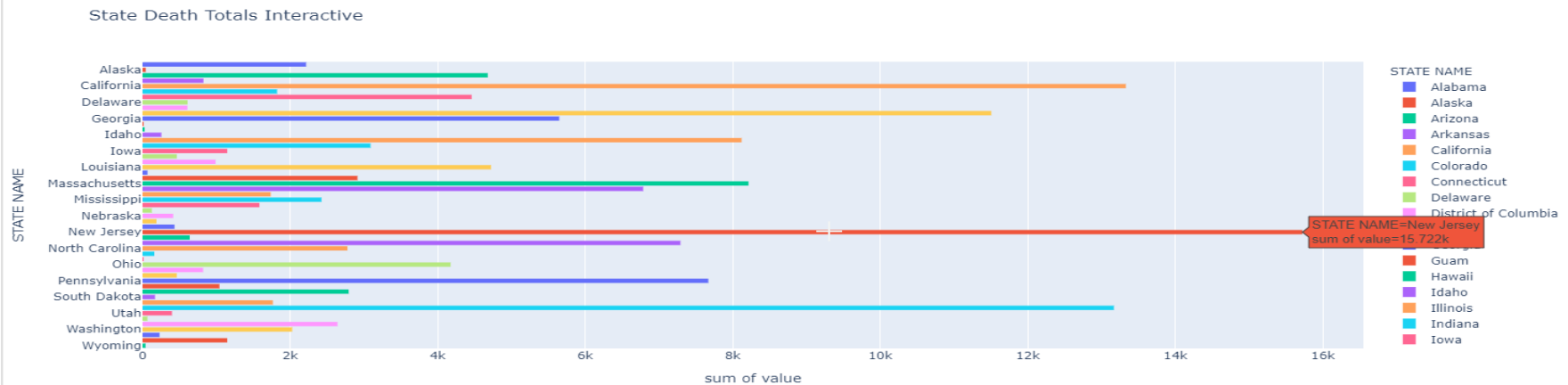
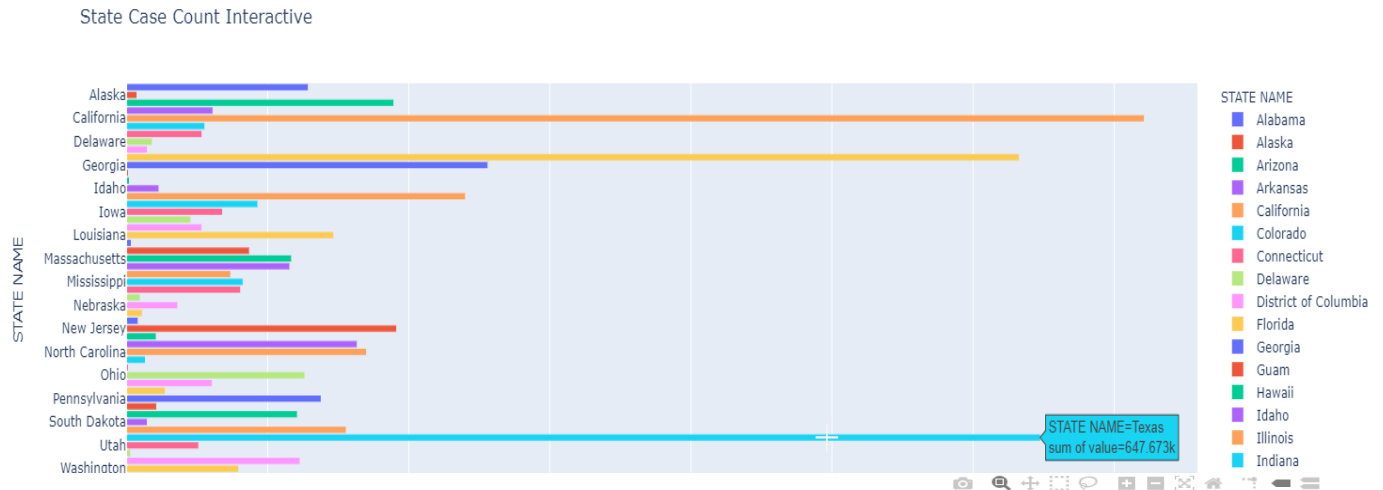
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•All three datasets were combined into one dataset to enable visualization of the exploratory process of the dataset, and to further assist in modeling. The following correlation plot was created, for all potential features within the dataset:



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

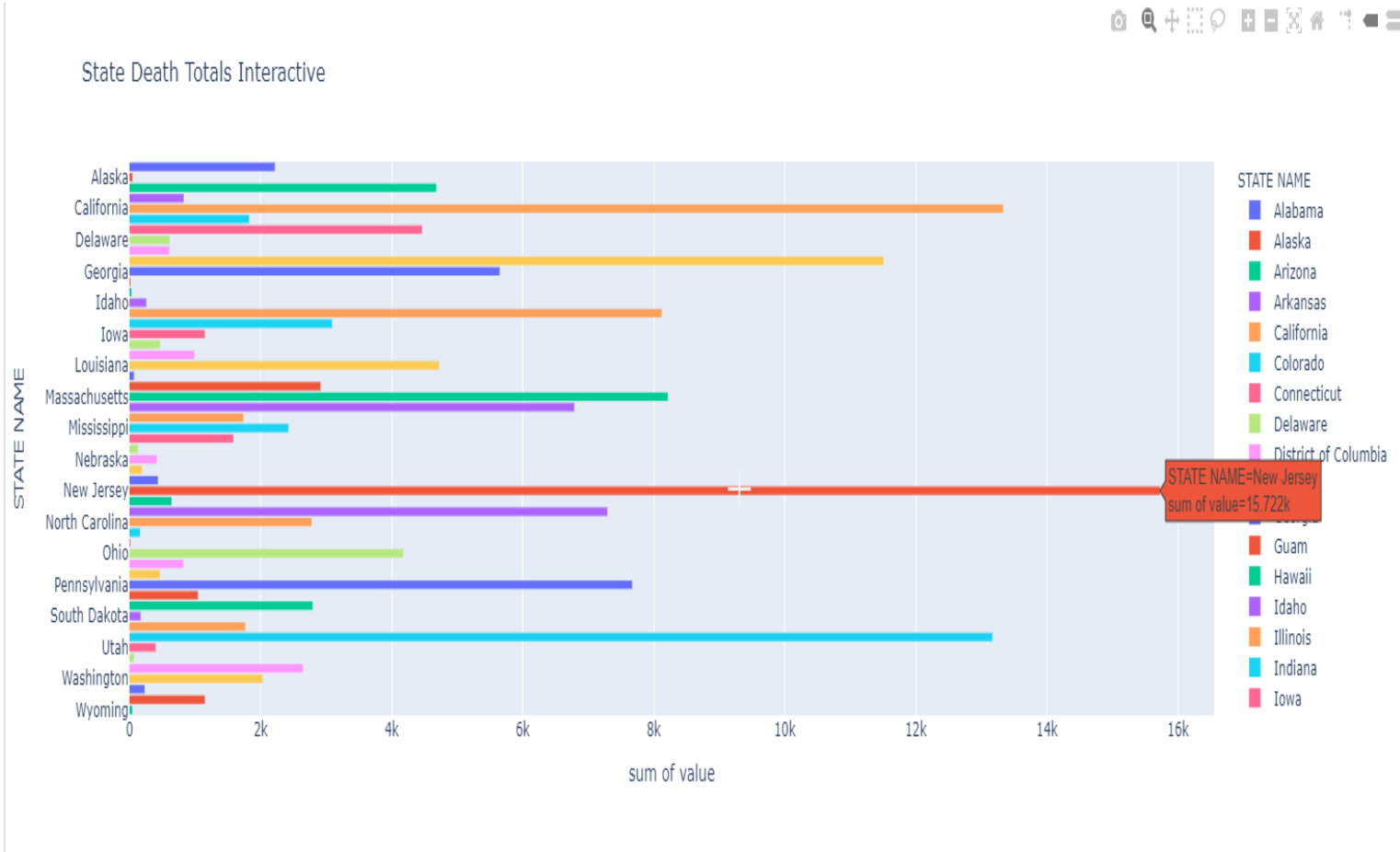
Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

•Project Description

•All three datasets were combined into one dataset to enable visualization of the exploratory process of the dataset, and to further assist in modeling. The following correlation plot was created, for all potential features within the dataset:



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

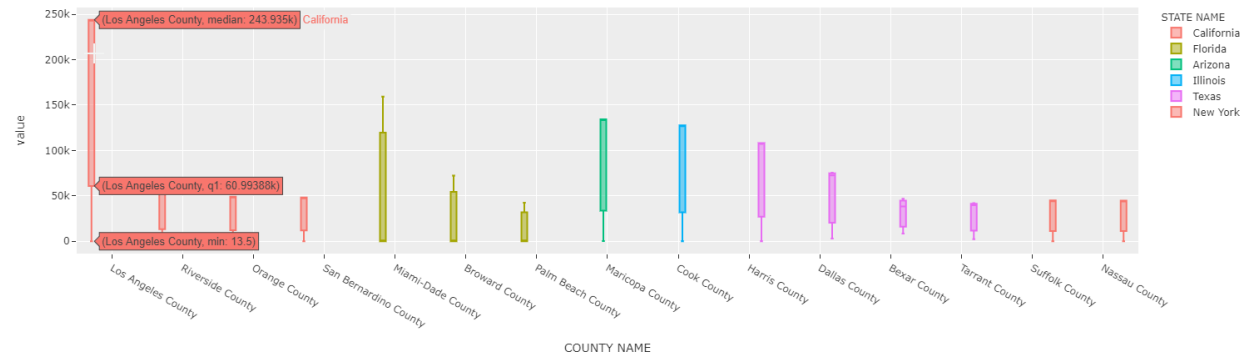
iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

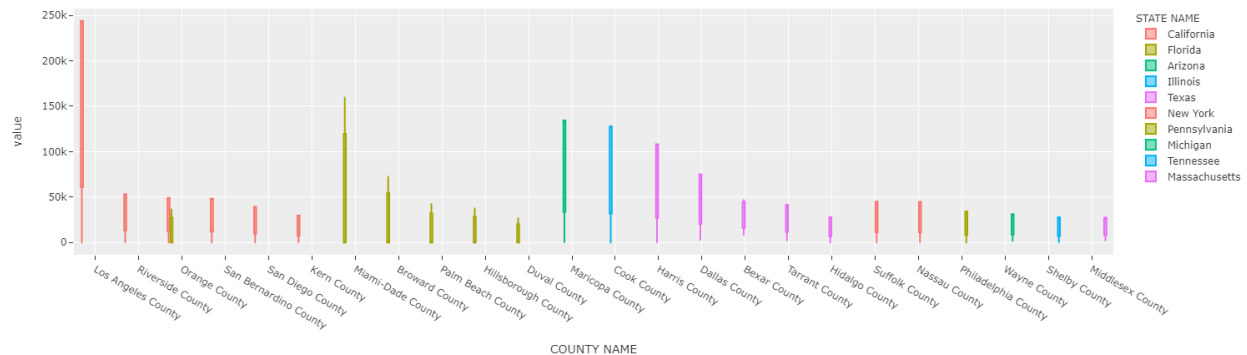
•:Project Description

•To better understand the distribution of COVID-19 around the nation, an interactive box plot of the top fifteen metro areas was created, followed by a top twenty five metro areas interactive boxplot, as shown:

BoxPlot All Cases, Top 15 Metro Areas Interactive



BoxPlot All Cases, Top 25 Metro Areas Interactive



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

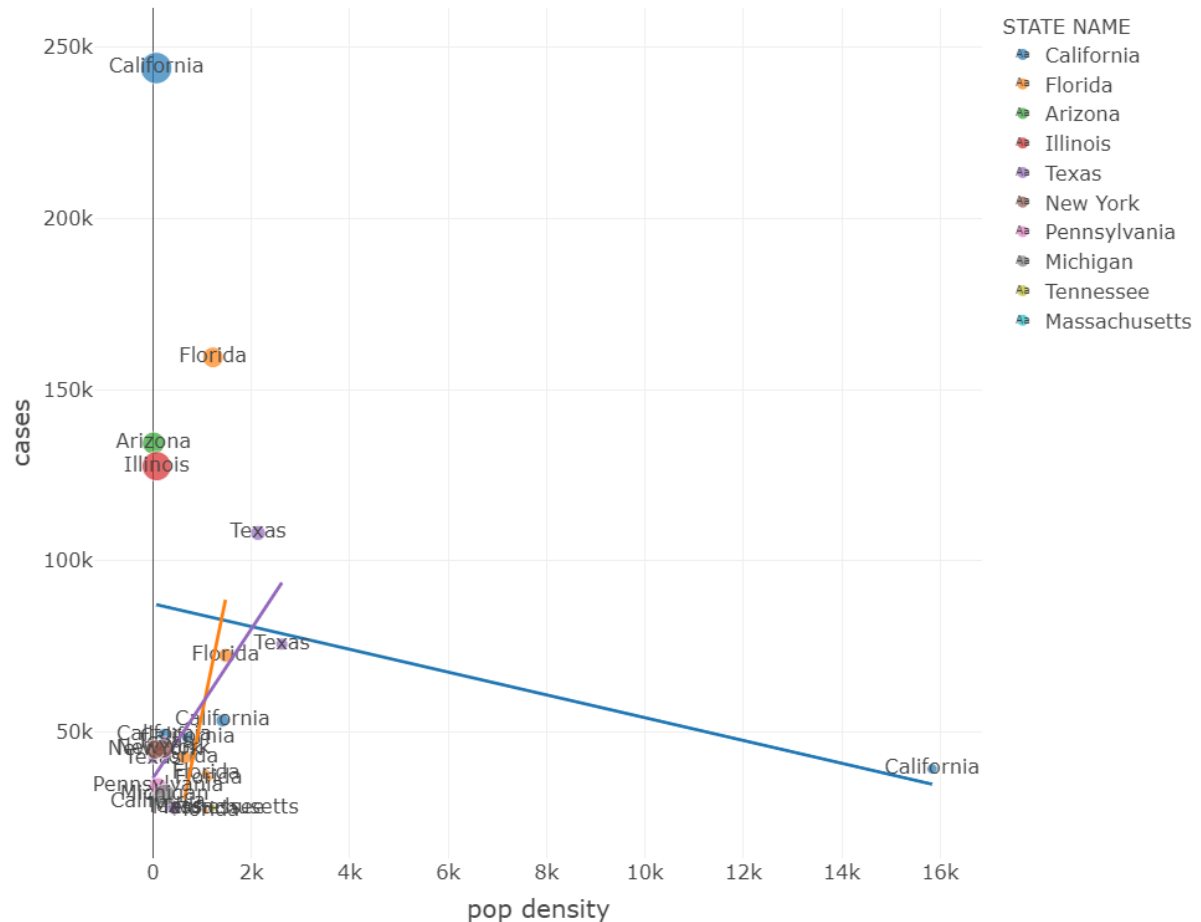
ARCHIVE

Get the full lifecycle view.

iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

Population Density, Distribution of highest cases by state



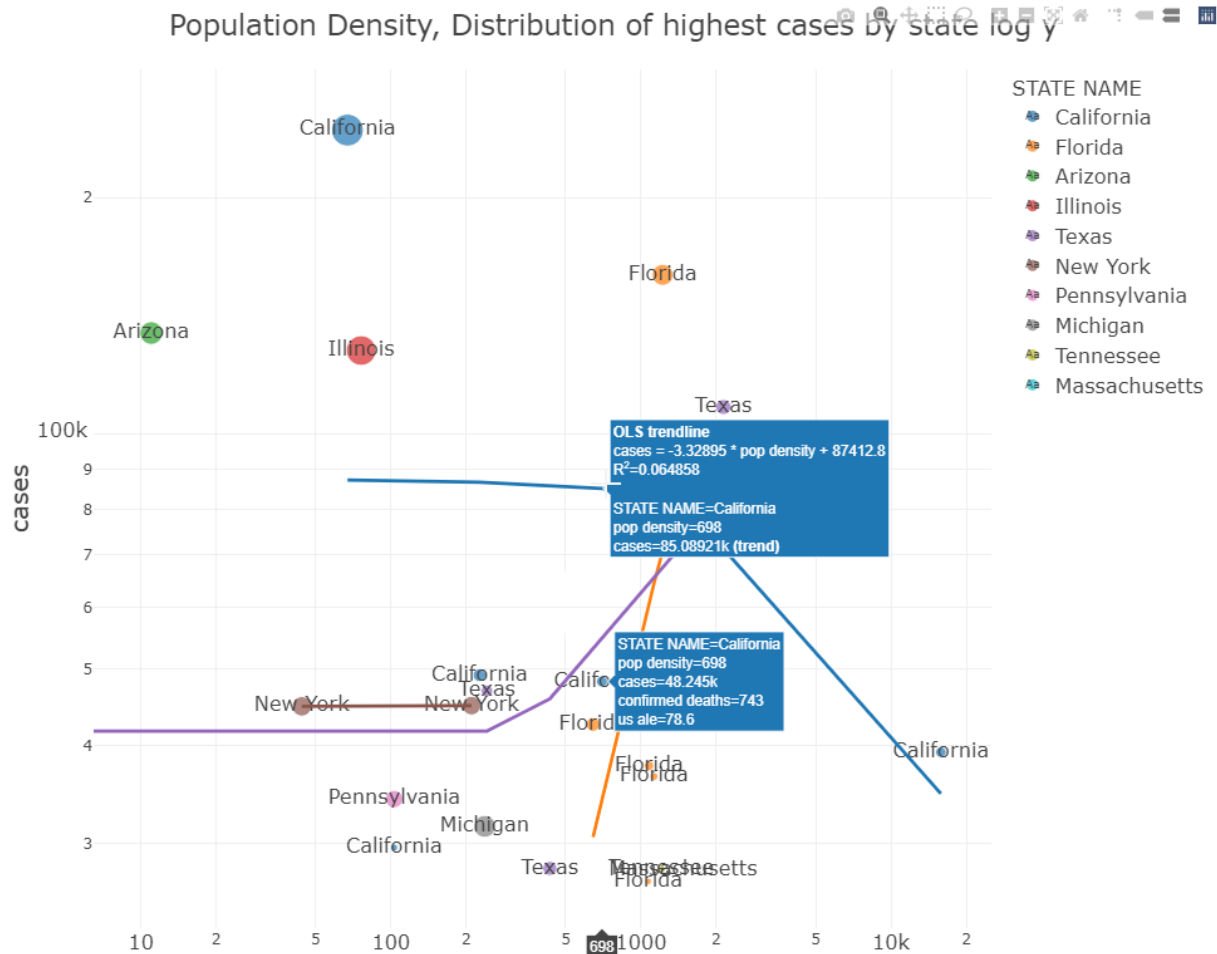
•Project Description

•Further investigation lead to the follow interactive, that demonstrates that there is indeed linearity between our target variable, cases, and one of the targeted key dependent variables 'features,' population density

IST 707: Data Analytics

•Project Description

•Further investigation lead to the follow interactive, that demonstrates that there is indeed linearity between our target variable, cases, and one of the targeted key dependent variables 'features,' population density



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

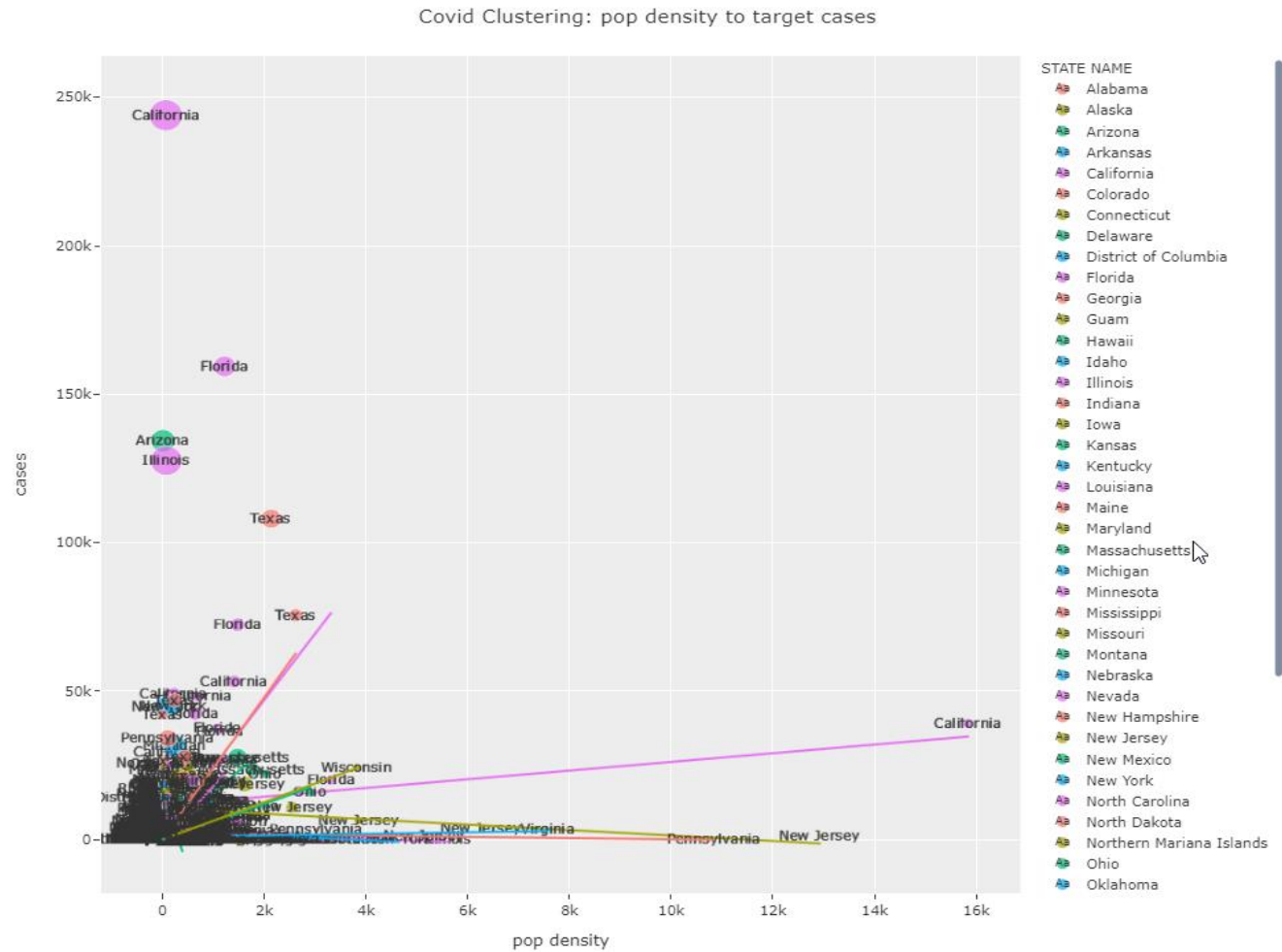
iSCHOOL.SYR.EDU/BIGDATA

IST 707: Data Analytics

•:Project

Description

•Cluster plot of All
states



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

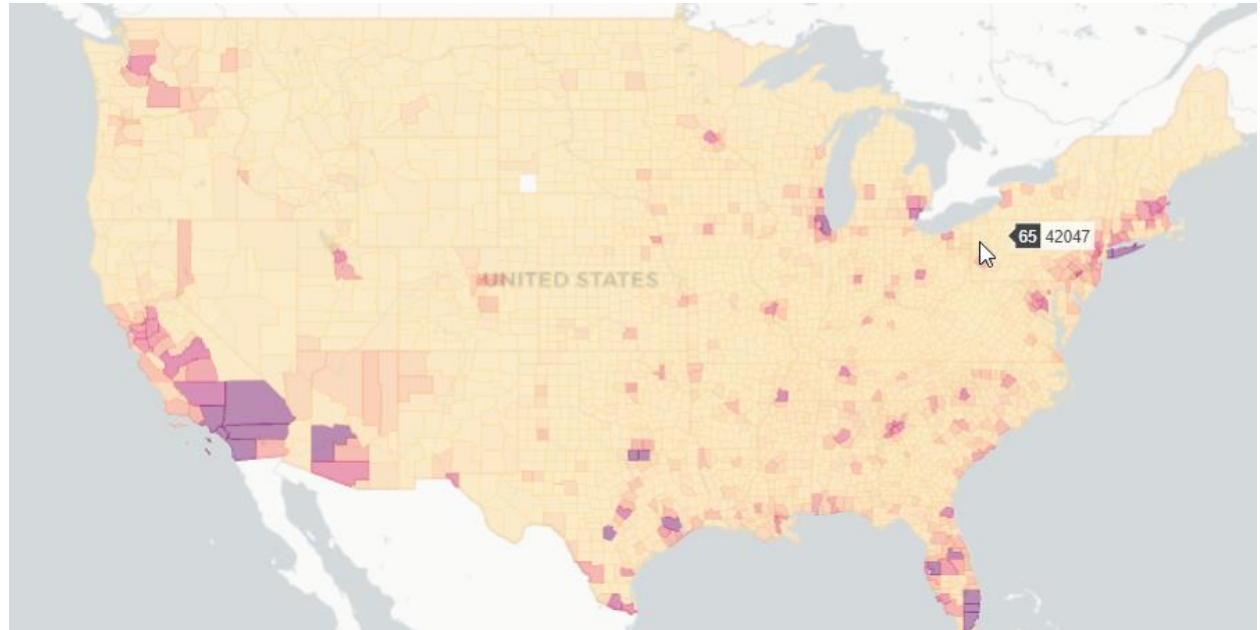
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•To further the exploratory data analysis, the candidate began to geospatially plot the data as to help researchers understand the implications of the findings through pre-modeling exercises to determine correlation and linearity between the variables, a choropleth interactive heat map was created to demonstrate the distribution of COVID-19



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

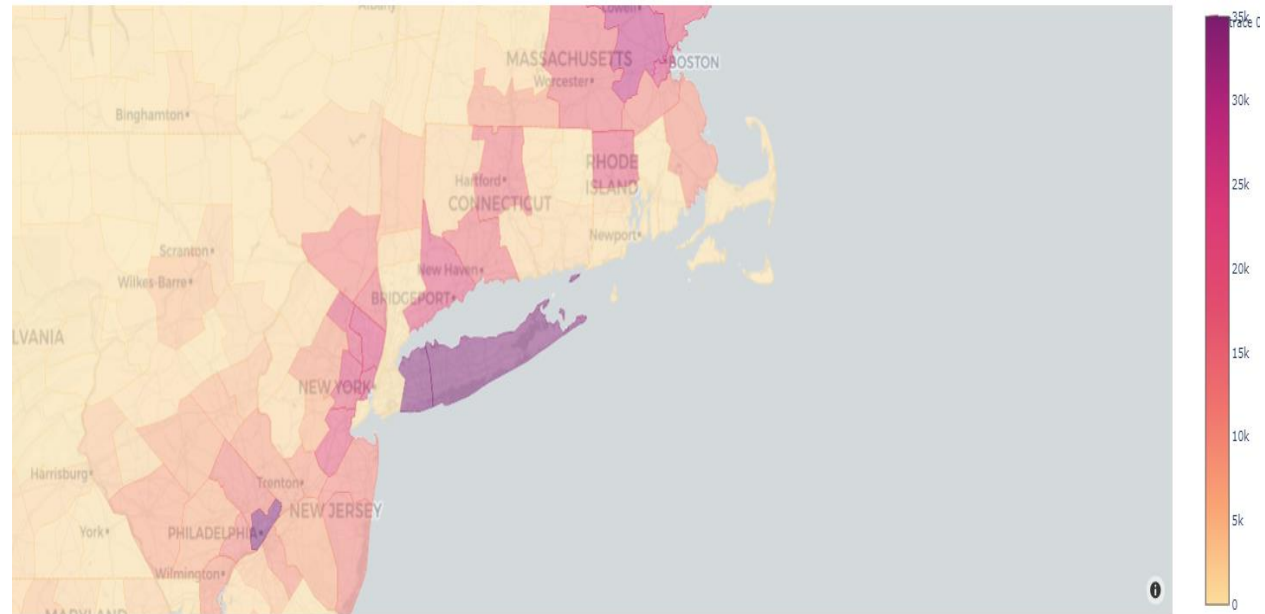
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•To further the exploratory data analysis, the candidate began to geospatially plot the data as to help researchers understand the implications of the findings through pre-modeling exercises to determine correlation and linearity between the variables, a choropleth interactive heat map was created to demonstrate the distribution of COVID-19



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

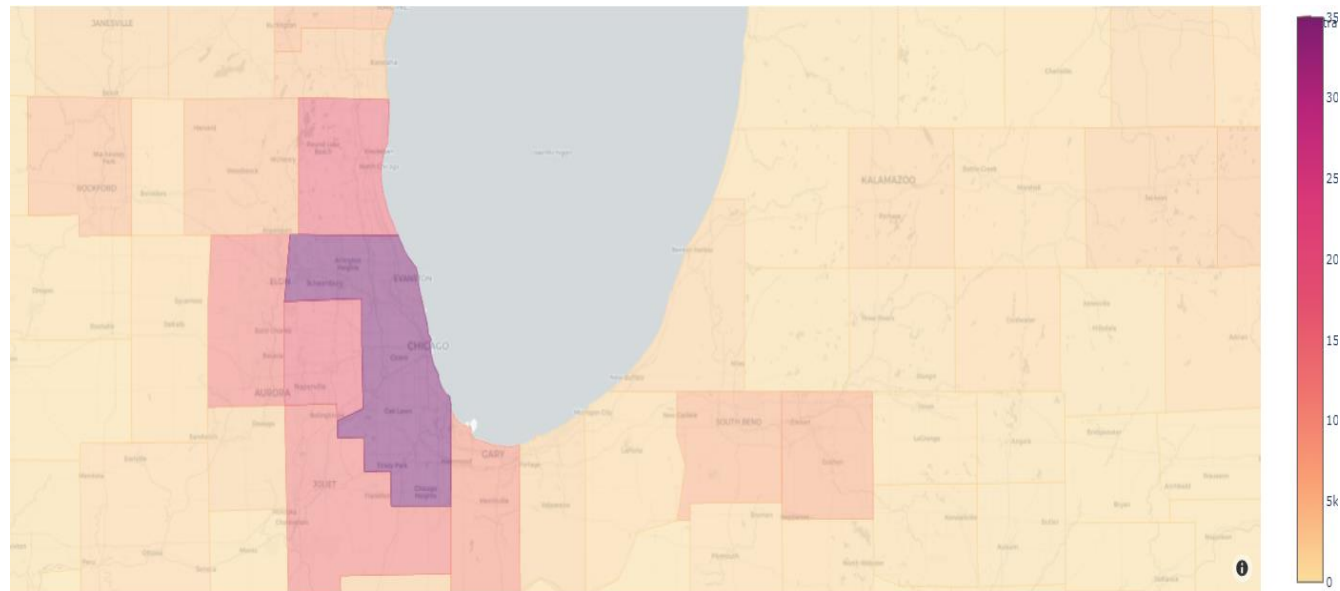
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•To further the exploratory data analysis, the candidate began to geospatially plot the data as to help researchers understand the implications of the findings through pre-modeling exercises to determine correlation and linearity between the variables, a choropleth interactive heat map was created to demonstrate the distribution of COVID-19



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

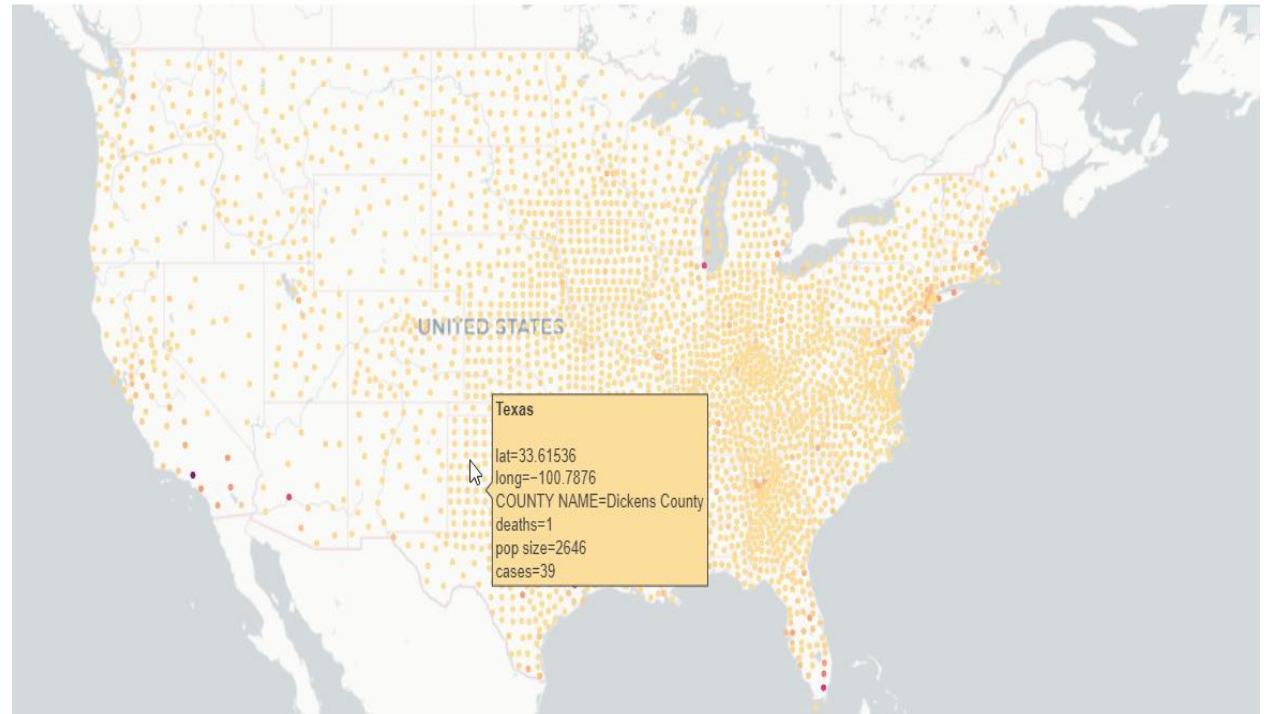
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•To further explore the data from a geospatial perspective, the candidate created another interactive continental map to demonstrate the COVID-19 distribution, however, this was based on the latitude and longitude coordinate, not FIPS coding, and further demonstrated a degree level of coloring, based upon the number of 'cases,' as demonstrated



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

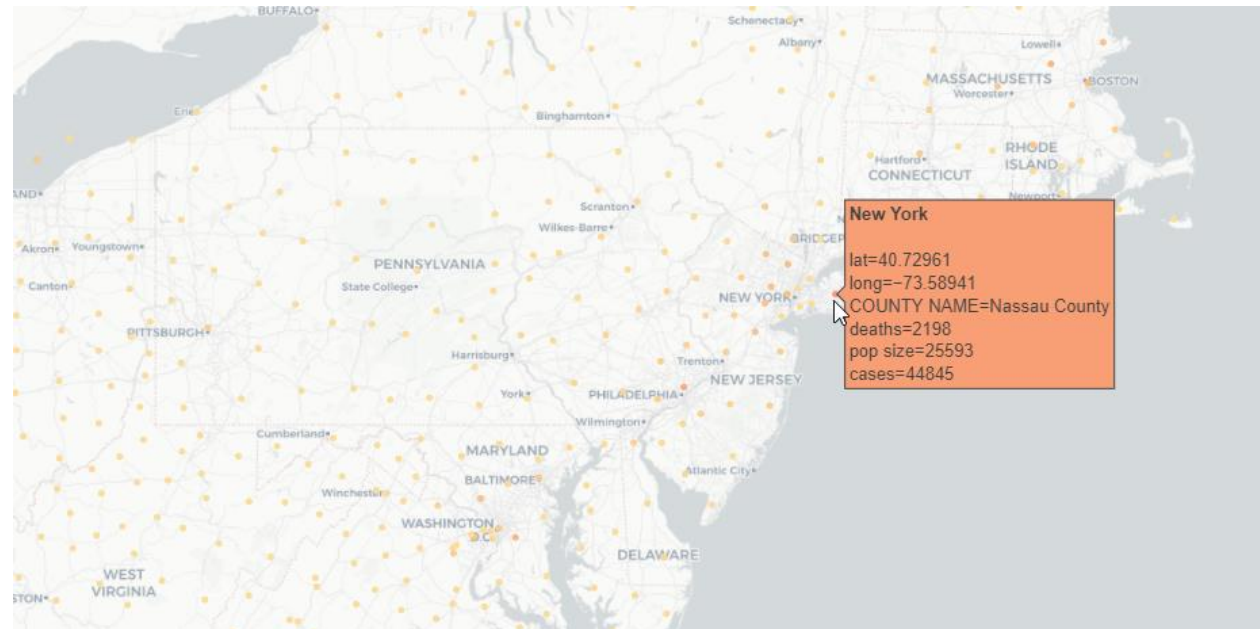
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•To further explore the data from a geospatial perspective, the candidate created another interactive continental map to demonstrate the COVID-19 distribution, however, this was based on the latitude and longitude coordinate, not FIPS coding, and further demonstrated a degree level of coloring, based upon the number of 'cases,' as demonstrated



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

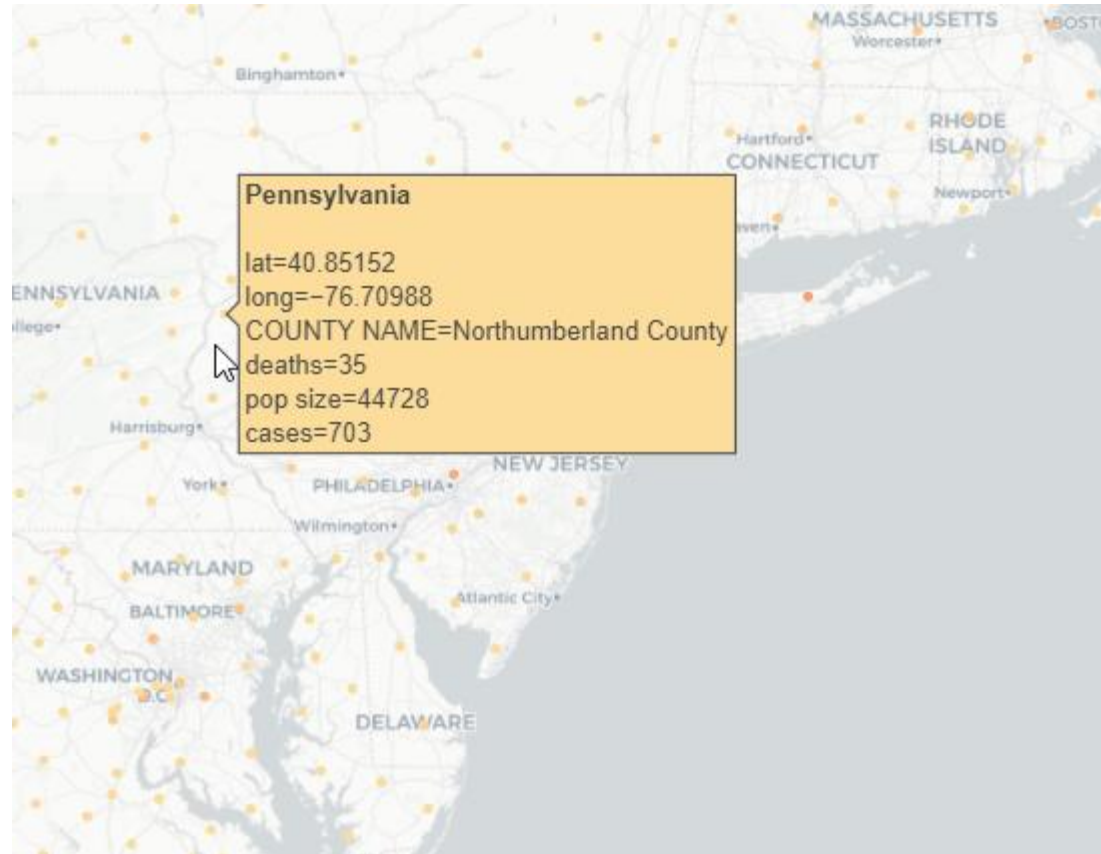
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•To further explore the data from a geospatial perspective, the candidate created another interactive continental map to demonstrate the COVID-19 distribution, however, this was based on the latitude and longitude coordinate, not FIPS coding, and further demonstrated a degree level of coloring, based upon the number of 'cases,' as demonstrated



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

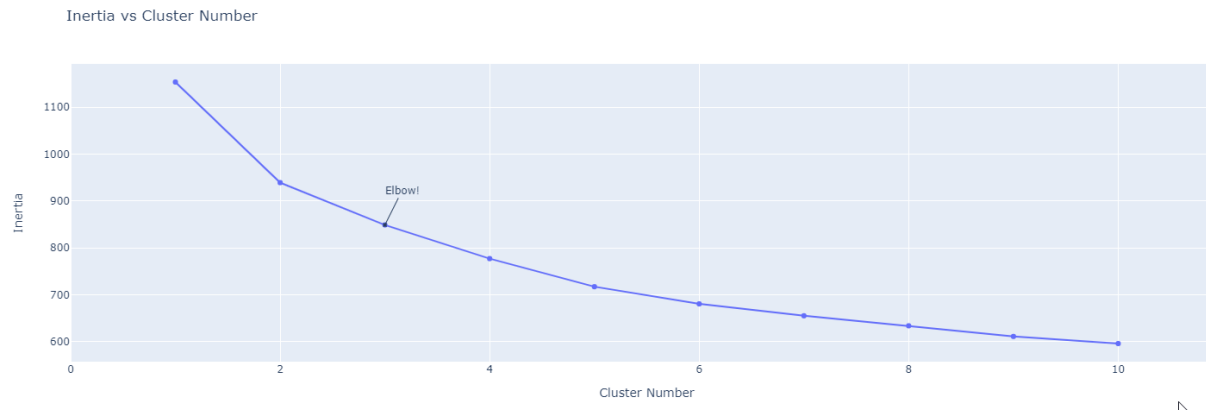
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

- Approaching the deadline for the submission of the portfolio will not allow for the SUPERVISED MACHINE LEARNING outcomes to be published within the paper, however, the candidate can share the UNSUPERVISED MACHINE LEARNING outcomes, via Kmeans Analysis.
- K-Means is a distance-based algorithm. Because of that, it's super important to normalize, standardize, or to choose any other option in which the distance has some comparable meaning for all the columns. MinMaxScaler, it's an excellent tool for it.
- After scaling our dataset, we can evaluate our inertia on different cluster numbers. If we see the chart, we could say that the elbow is on 3 or 4. For simplicity, we will use 3.



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

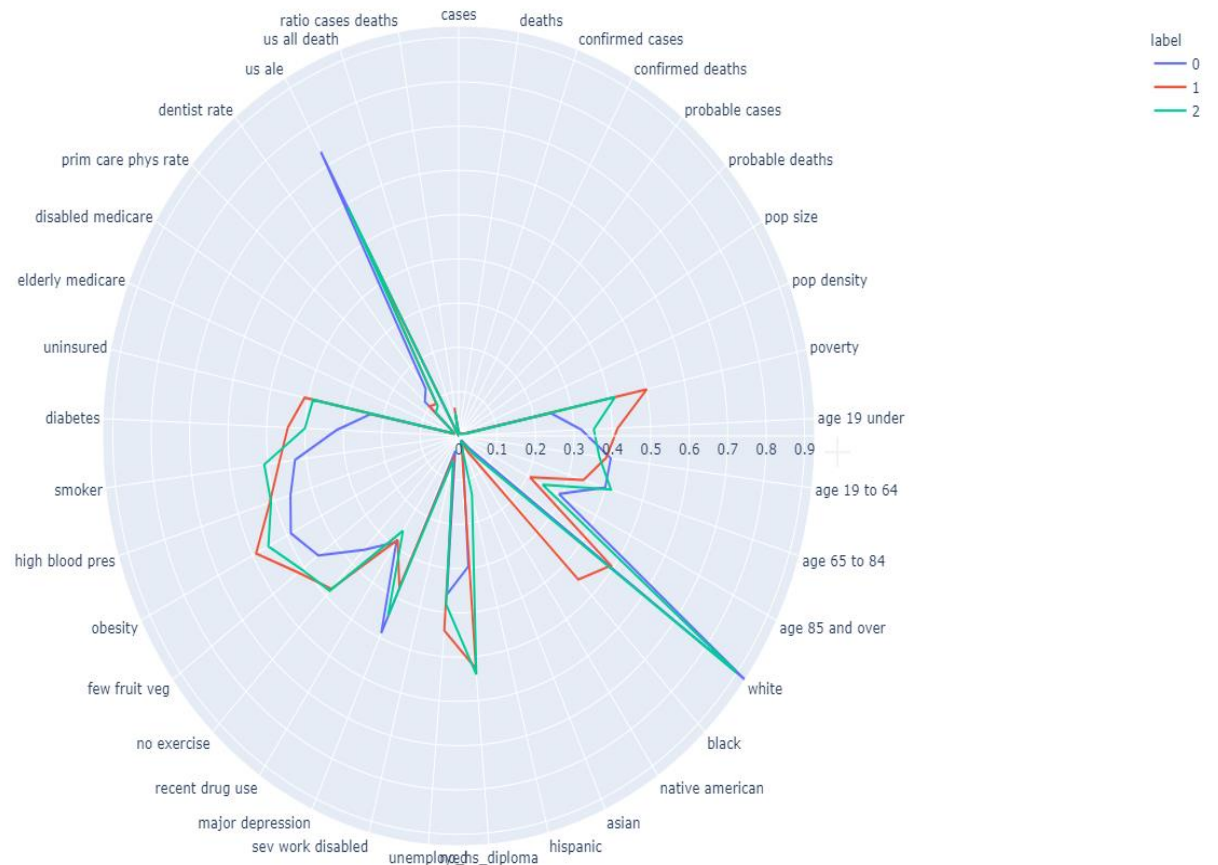
[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project

Description

•Line_polar



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

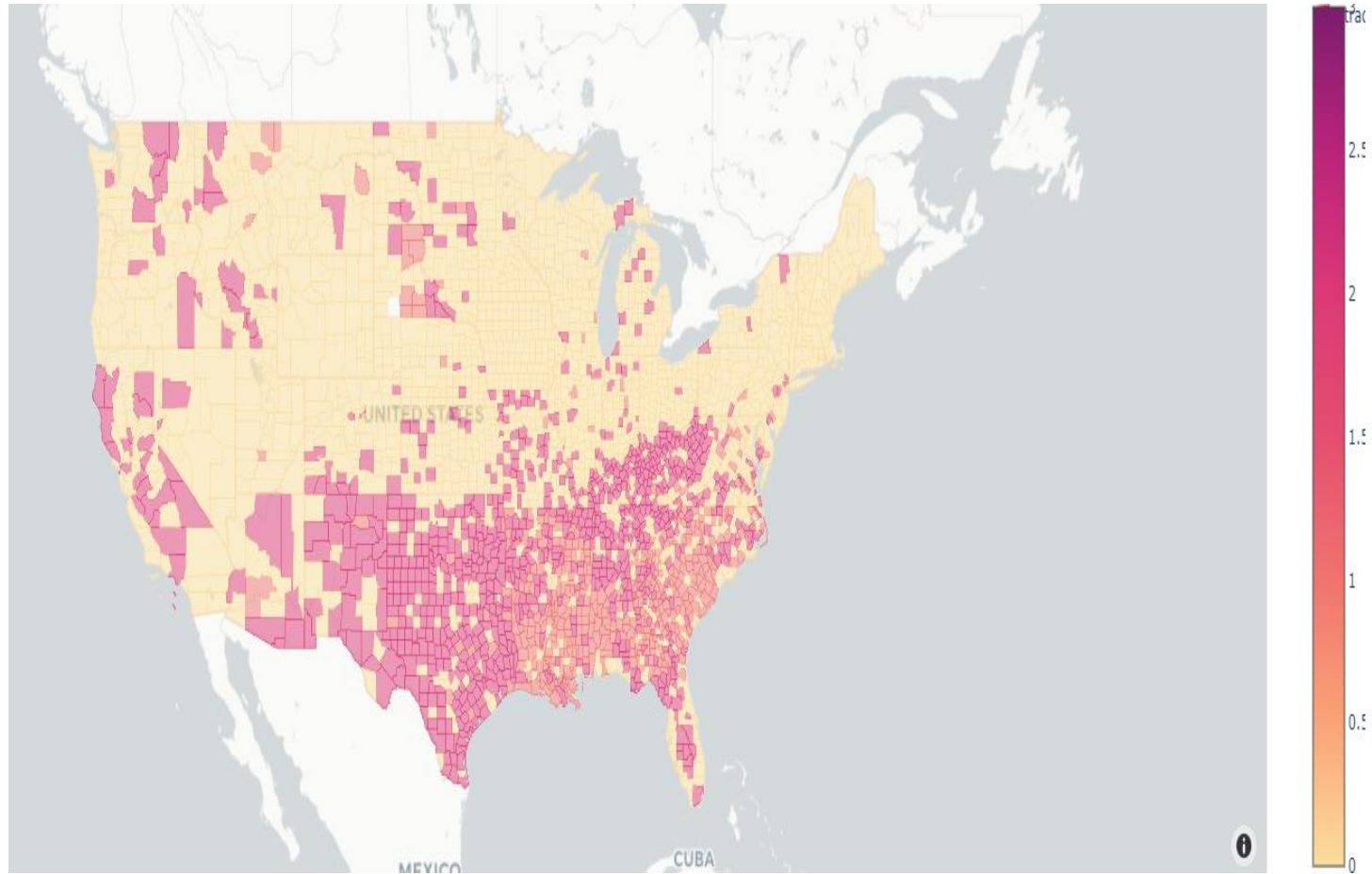
Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

•Project Description

•The Kmeans clustering, three cluster in total, were then appended back to their original dataset, such, that a geospatial rendering of the clusterings around the country could be visualized, as demonstrated:



DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

IST 707: Data Analytics

a. :Learning Goals

- The learning goals of the Big Data COVID-19 project, from the candidate's contribution to that on-going project, is to try to determine some predictability related to distribution of the pandemic, per county. Given the various at-risk factors that already existed pre-pandemic, it is the candidates goal to establish predicted model based upon those at-risk factors, and the current growth of COVID-19, per county.

DATA SCIENCE AT THE iSCHOOL AT SYRACUSE UNIVERSITY

ARCHITECTURE

ACQUISITION

ANALYTICS

ARCHIVE

Get the full lifecycle view.

[iSCHOOL.SYR.EDU/BIGDATA](https://ischool.syr.edu/bigdata)

Conclusion

This portfolio has been compiled to testify to the successful implementation of these learning objectives, and the mastery of the major practice areas within Data Science by the master's Candidate, Randall Scott Taylor. In the four demonstrated projects, data was collected, via standard .csv, web scrapping, and application of programming interfaces in conjunction with databasing solutions. All to be utilized in the endeavor to analyze the data, using statistical methods and data mining techniques for tasks ran against selected features, for such tasks as, regression, classification, or clustering. All of these works were done for the betterment of understanding, and to provide meaningful analysis and interpretation therein, to assist decision makes, and humanity by examining some of the most interesting produce data, within the human experience, the law, healthcare, and anthropological studies. It is with great joy that the candidate chooses these subject matters, as they are very close and dear to the human experience

Conclusion

- The candidate communications skills were further developed and displayed in the delivery of insights, the organizations of projects and the leadership of data collection, wrangling, multiple linear regression models, K-means clustering, Decision Trees, and geospatial analysis. The candidate was particularly forward thinking in the expression of the chosen packages and methods utilized to visualize the information, and to analyze large data sets with geographic representations to assist decision makers in their attempt to quickly make the hard decisions.

Conclusion

- Syracuse University's School of Information Studies provided the candidate, as they provide every student, the opportunity to learn and grown within the new advance field of Data Science. Skills learned in the program have cultivated the candidate to providing a multifaceted analytical approach to the needs of future organizations, and their stakeholders and business professionals.
- Thank you