Portfolio Milestone

Randall Scott Taylor

SUID: 965757884

Github; https://github.com/randallscott25/MSADS_Portfolio

CONTENTS

Introduction:

Syracuse University's MS in Applied Data Science is a 'practitioners' degree, in that there is strong emphasis placed upon the hands-on approach to learning. The Applied Data Science program within the iSchool provides graduate students the opportunity to collect, manage, analyze, and develop data insights, utilizing various tools and techniques. The student will demonstrate mastery of these tools and techniques, by way of a report of four specific course works and their respective projects. These courses are as follows: IST 659: Database Administration (Taylor, "IST 659," 2020), IST 652: Scripting for Data Analysis (Taylor, "IST 652," 2020), IST 707: Data Analytics (Taylor, "IST 707," 2020), and IST 718: Big Data (Taylor, "IST 718," 2020). The following are an example of the many skills sets that have been developed at the School of Information Studies, the goal is to furnish future data scientist with the ability to generate value within their respective organizations and to produce meaningful analysis and recommendations.

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.*
7. *Synthesize the ethical dimensions of data science practice*.

IST 659: Database Administration. Professor *Gregory Block PhD*

a. Project Description

The Database Administration course chosen by the candidate was facilitated through adjunct professor Dr. Gregory Block. The project deliverable for the course was to create a SQL database that:

> "the purpose of this Project is to convince stakeholders, in non-technical and technical demonstrations, that there is a need to create a centralized Database" (Taylor, "IST 659," 2020).

In response to the call of the question: 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. The systems being utilized at the time were not centralized, nor was any of the data normalized which is important to efficient databasing, as James Le points out in his article "An Introduction to Big Data: Data Normalization," stating:

> There are various reasons to normalize the data, among those are: (1) Our database designs may be more efficient, (2) We can

reduce the amount of redundant data stored, and (3) We can avoid anomalies when updating, inserting, or deleting data (Le, "An Introduction to Big Data: Data Normalization," 2019).

To normalize the data, we must first become familiar with a couple of key players, namely our stakeholders, as exampled:

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.

Fig 1: *Stakeholder Description,* (Taylor, "IST 659," 2020).

Working with these key players, the student was able to develop the beginnings of moving from the first normal form of the data, by establishing parameters by which to govern some of the key tables and their future relationship to one another. The following example demonstrates that effort, by establish a base reference by which the candidate was able to establish those entities and their relationships, a **glossary,** as exampled:

Glossary:

Contained below is a glossary of *Agents* and *Resources*: required attributes and their respective relationships, entity to entity.

Client: The *Client* entity will require: client name, client address, and client e-mail address. One or many client(s) are assigned to Personnel. One to many Clients can owe hours or be billed (Billable) hours. Zero or one client can require Judicial Forms. One or many client have or, are assigned zero or many Discovery. Zero or many client can be scheduled to one or many Courts. Finally, One or many Client can be scheduled to one or many Firm Calendar entries, assignments, case numbers and descriptions, calendar entries, total hours billed, discovery, and judicial forms.

Personnel: The *Personnel* entity will require: Name, State Bar License Number, and hours owed. One or many Personnel is assigned or has one or many Clients. One or many Personnel can be scheduled to Firm Calendar. One or many hours owed Billing to one or many Personnel.

Court: The *Court* entity will require: Court Name, Court Number, Judicial Officer assigned, the hearing date, Hearing Description, and Case Number. One or Many Courts can be assigned to zero or many clients.

Firm Calendar: The *Firm Calendar* entity will require: Date and time entries. It will also require for an hour spent entry to be made on the corresponding date and time. One or many Firm Calendars can be scheduled to one or many Clients. One or many Firm Calendars can be scheduled to one or many Personnel.

Billing: The *Billing* entity will require: rate (of personnel) along with hours billed (client) and hours owed(personnel). There will also be a client total hours entry required. One or many Billing hours can be billed to one or many client. One or many hours, at required specific rate, can be owed to one or many personnel.

Discovery: The *Discovery* entity will require: Type of discovery and discovery storage location entry. Zero or many Discovery are assigned to One or many clients.

Judicial Forms: The *Judicial Forms* entity will require: Form Jurisdiction (what Court is the form from Civil or Criminal) the Form Name and the form storage location. Zero or many Judicial Forms can be required for zero or many Client.

Fig 2: *Glossary,* (Taylor, "IST 659," 2020).

The collection and the organization of the data to be considered for addition to the database, the interview with the stakeholders, and the establishment of the terms by which we would organize our normalized relational database

tables, allowed for the project to move forward into the conceptual model for the building of the database, as exampled below:

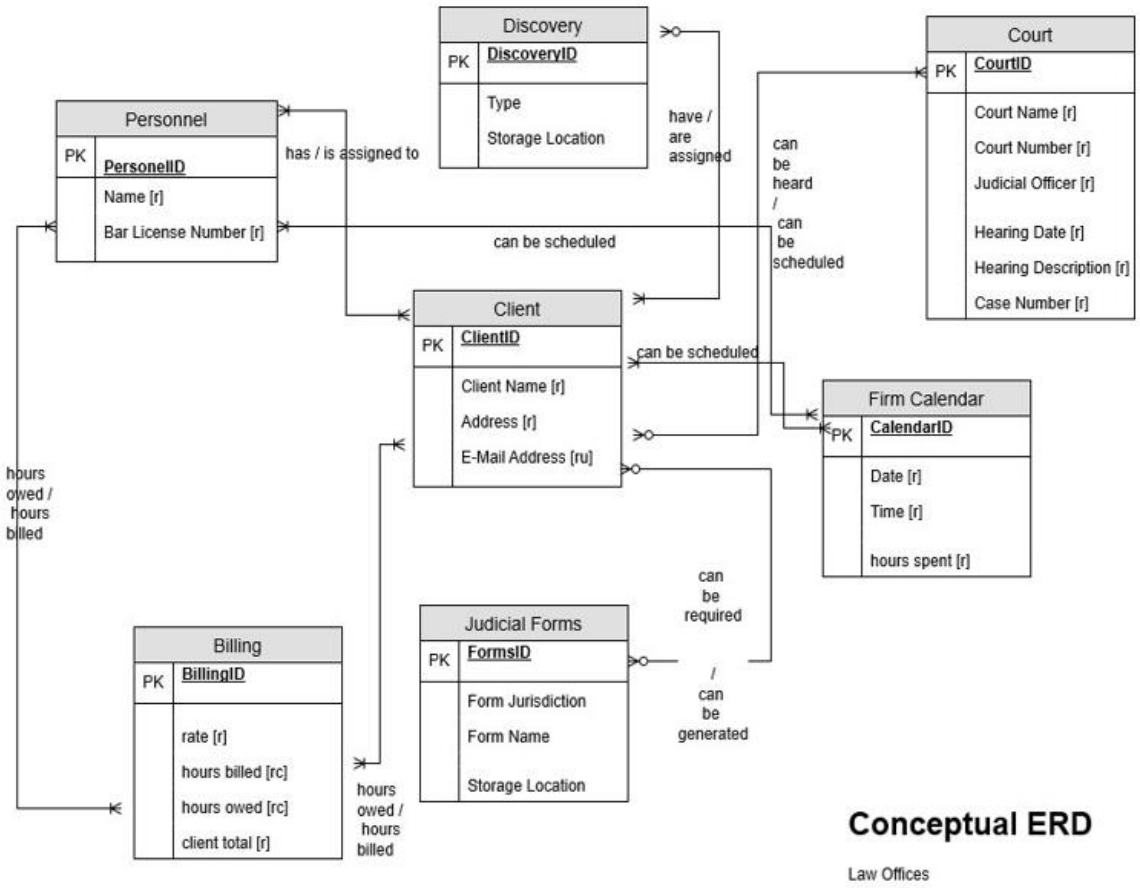


Fig 3: *Conceptual ERD,* (Taylor, "IST 659," 2020).

The development of the conceptual database allowed for the further normalizing of the data, namely into the third normal form, which is the third step in the normalization of data, as Le states:

> Third normal form (3NF) is the third step in normalizing a database and it builds on the first and second normal forms, 1NF and 2NF. 3NF states that all column reference in the referenced

> data that are not dependent on the primary key should be removed. 3NF was designed to: eliminate undesirable data anomalies; reduce the need for restructuring over time; make the data model more informative; make the data model neutral to different kinds of query statistics (Le, "An Introduction to Big Data: Data Normalization," 2019).

In layman's terms, only foreign key columns should be used to reference another table, an no other columns from the parent table should exist in the referenced table. Ensuring that the data was formed in the second normal form and contained no transitive functional dependencies, allowed for the establishment of the logical model, as exampled below:

The Logical Model: Law Office (UPDATED)

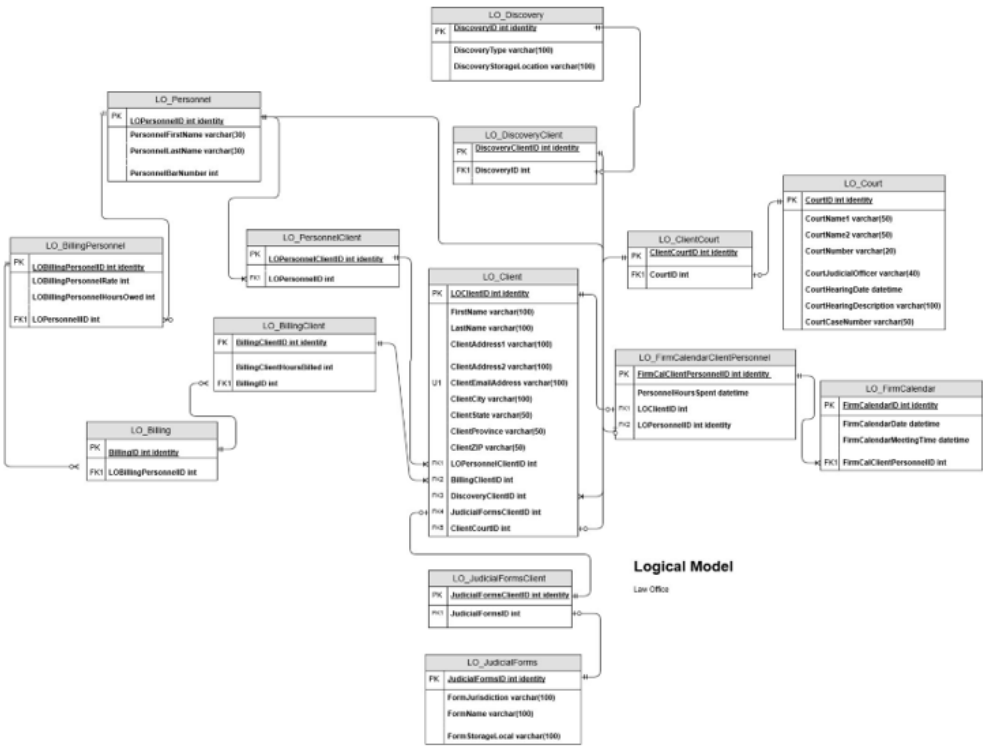


Fig 4: *Logical ERD,* (Taylor, "IST 659," 2020).

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
```

Fig 5: *stored procedure,* (Taylor, “IST 659,” 2020).

```
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
```

Fig 6: *view,* (Taylor, "IST 659," 2020).

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

Fig 7: *Microsoft Access report built from fig 6 view,* (Taylor, "IST 659," 2020).

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.

Law Office Master Calendar

Master Meeting Number	Personnel ID #	Case Hours	Mediation Bill Total $	Firm Client ID#	Date
					4/4/2019 2:30:00 AM
					4/6/2019 2:30:00 AM
					4/7/2019 2:30:00 AM
					4/9/2019 2:30:00 AM
					4/11/2019 2:30:00 AM
101	3	11	240	7	4/8/2019 2:30:00 AM
102	3	23	480	2	4/2/2019 2:30:00 AM
103	2	55	1440	9	4/10/2019 2:30:00 AM
106	4	21	480	9	4/2/2019 2:30:00 AM
107	3	42	1200	4	4/5/2019 2:30:00 AM
108	3	53	1440	1	4/1/2019 4:30:00 AM

Fig 8: *Microsoft Access: Law Office Master Calendar,* (Taylor, "IST 659," 2020).

b. Reflection & Learning Goals.

"I had no idea what I was doing, nor, what I was getting into … third day Into this class," (Taylor, "IST 659," 2020). Which, for the most part, was the truth, or is the best way one can summarize in a simple statement. 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 data transacted.

This project contributed to the successful application of the learning goals through the exercise of collecting and managing data, as well as the identification of patterns using statistical analysis. These combined observations were leveraged to bring meaningful insight into the organization of a database for a small New York law firm.

IST 652: Scripting for Data Analysis Professor *Deborah V. Landowski, PhD*

a. Project Description

The Scripting for Data Analysis course chosen by the candidate was facilitated through course professor Deborah V. Landowski, PhD. Through the utilization of python scripting, and various data mining techniques, machine learning technique and geospatial analysis, the candidate created a final project. The final project deliverable for the course was to create a data source from several sources format and unformatted data and to demonstrate the following:

> "For this assignment, you are to make an initial plan for a project. In the final project you will demonstrate your ability to write Python scripts to access and amass data from fields in one or more of the three types of data studied in the course and to

> prepare and use data to produce data summaries, lists and other structures." (Taylor, "IST 652," 2020).

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. The researcher finds it paramount to distinguish these contributors, and the works therein, before commenting further upon the study. In response to the end all driven results for the project, the purpose was defined:

> The purpose of this study of the above described data will be to provide insight and understanding to the dataset, the story the dataset contains and to answer further research question identified, within the document. The goal of this final project report will be to identify anything interesting within the distribution of the data, and the families of the various languages. Specifically, from the raw data can the genus family's groupings be shown within the data, and, are there any geospatial inferences

> that can be made, from visual display of those findings" (Taylor, "IST 652," 2020).

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. *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

Fig 9: *raw WALS dataset,* (Taylor, "IST 659," 2020).

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
```

Fig 10: *python Data Cleaning,* (Taylor, "IST 659," 2020).

During the cleaning process, the columns that were chosen needed to be identified and the missing values addressed, as demonstrated below:

```
In [45]: #Examination of the Columns contained within DataSet
         lang_df.columns

Out[45]: Index(['wals_code', 'iso_code', 'glottocode', 'Name', 'latitude', 'longitude',
                'genus', 'family', 'macroarea', 'countrycodes',
                ...
                '137B M in Second Person Singular', '136B M in First Person Singular',
                '109B Other Roles of Applied Objects',
                '10B Nasal Vowels in West Africa',
                '25B Zero Marking of A and P Arguments',
                '21B Exponence of Tense-Aspect-Mood Inflection',
                '108B Productivity of the Antipassive Construction',
                '130B Cultural Categories of Languages with Identity of 'Finger' and 'Ha
        nd'',
                '58B Number of Possessive Nouns',
                '79B Suppletion in Imperatives and Hortatives'],
               dtype='object', length=202)
```

Fig 11: *python Data Cleaning,* (Taylor, "IST 659," 2020).

```
In [44]: nRowsRead = 2679 #specify a specific number if the researcher doesn't
         #want to read in the entire language.csv file.

         #read in the language.csv as a pd
         lang_df = pd.read_csv('language.csv', delimiter=',', nrows = nRowsRead)
         lang_df.dataframeName = 'language.csv'
         nRow, nCol = lang_df.shape
         print(f'{nRow} rows and {nCol} columns: \nWill be utilized from the WALS-Dataset

         2679 rows and 202 columns:
         Will be utilized from the WALS-Dataset

In [38]: #Take a quick look at the data thus far
         lang_df.head(10)
```

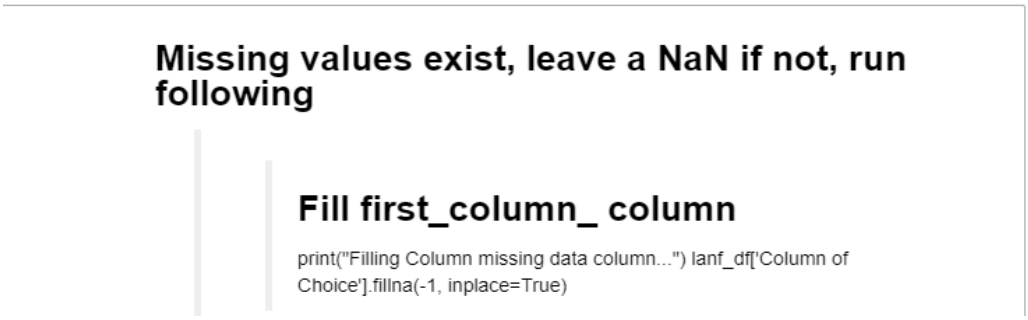


Fig 11, 12: *python Data Cleaning,* (Taylor, "IST 659," 2020).

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
```

Fig 13: *python web scraping,* (Taylor, "IST 659," 2020).

The geojson data was pulled into the local system by way of python and stored into an instance of MongoDB, or to be more precise pymongo, as demonstrated below:

Fig 14: *pymongo database,* (Taylor, "IST 659," 2020).

Storing the data in the document database allow for us to quickly poll the data by utilizing a loop to inspect the keys within the dictionary entries. Data

Cleaning of the semi-structured geojson was required, as there are quite several missing values. This is no exception regarding the unstructured data pulled via the geojson. To attempt to start to make sense of the data, the researcher first established a document collection,

And then, from this collection the researcher created a 'list' named array as demonstrated by the following:

Fig 14: *semi-structured database,* (Taylor, "IST 659," 2020).

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.

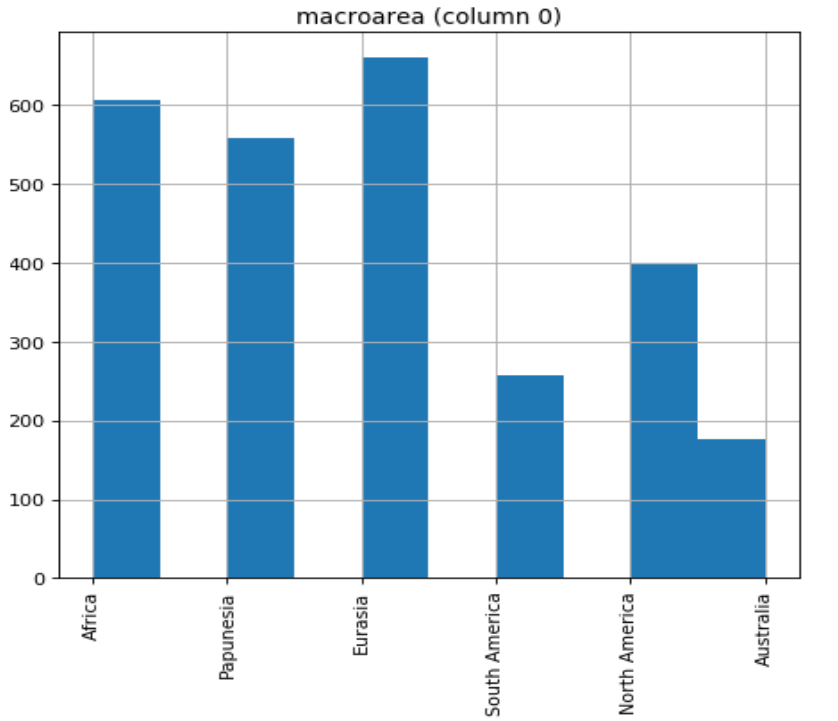
Exploratory research questions presented against the data:

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

The following frequency histogram display and code answer this first question:

```
In [49]: #Distribution graphs (hisogram/bargraph) of Column Data:

def plotPerColumnDistribution(df, nGraphShown, nGraphPerRow):
    nunique = df.nunique()
    df = df[[col for col in df if nunique[col] > 1 and nunique[col] < 50]] # For displaying purposes, pick columns that have bet
    nRow, nCol = df.shape
    columnNames = list(df)
    nGraphRow = (nCol + nGraphPerRow - 1) / nGraphPerRow
    plt.figure(num = None, figsize = (6 * nGraphPerRow, 8 * nGraphRow), dpi = 80, facecolor = 'w', edgecolor = 'k')
    for i in range(min(nCol, nGraphShown)):
        plt.subplot(nGraphRow, nGraphPerRow, i + 1)
        columnDf = df.iloc[:, i]
        if (not np.issubdtype(type(columnDf.iloc[0]), np.number)):
            valueCounts = columnDf.value_counts()
            valueCounts.plot.bar()
        else:
            columnDf.hist()
        plt.ylabel('counts')
        plt.xticks(rotation = 90)
        plt.title(f'{columnNames[i]} (column {i})')
    plt.tight_layout(pad = 1.0, w_pad = 1.0, h_pad = 1.0)
    plt.show()
```

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



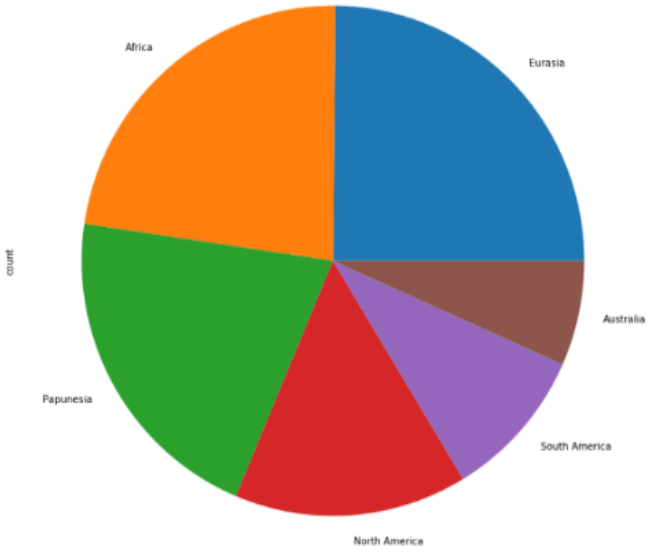


Fig 15: *frequency distribution of the language's,* (Taylor, "IST 652," 2020).

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 below:

What are the major language families counts of each respective individual language's family?

```
In [66]: # establish 'major families' in order to further understand the data to answer
         #What are the major language families counts of each respective
         #individual language's family?

         #top 20
         major_families_20 = lang_df.groupby('family').size().reset_index(name='count').sort_values(by='count', ascending=False).head(20)
```

	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

As demonstrated by Figure 10, the Major Family of Languages dispersion has now been established within the data frame, and this particular grouping of the data will assist in further analysis of the dataset in future, such as the geographic dispersion of languages, globally, as demonstrated

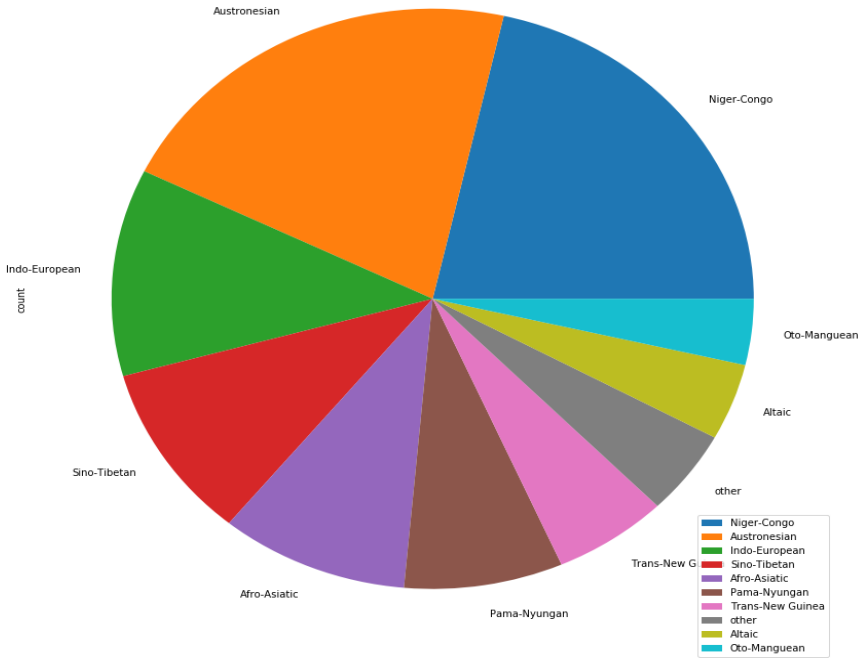
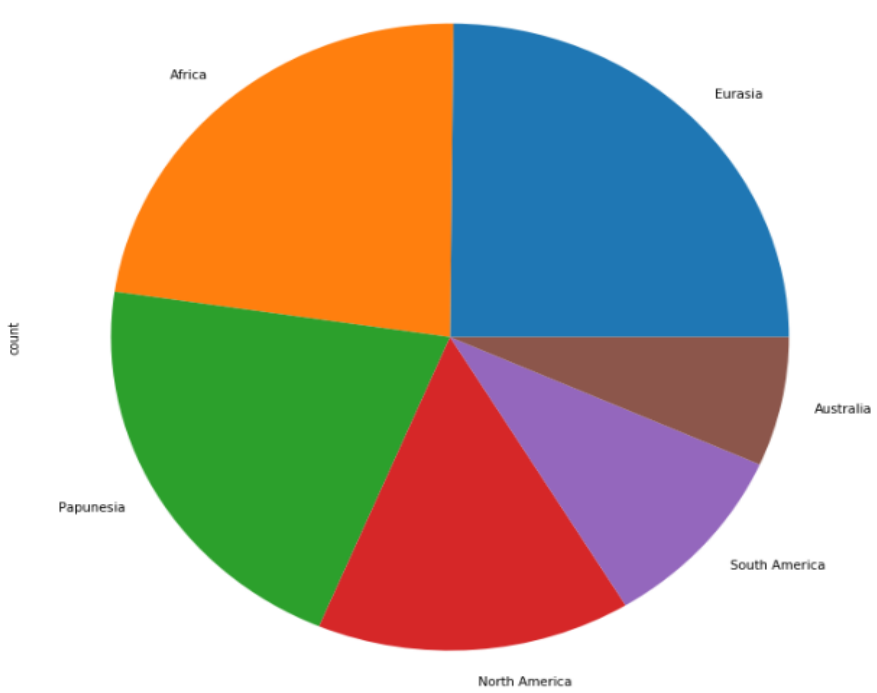


Fig 16: *major language families counts,* (Taylor, "IST 652," 2020).

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.

To approach this, code was written to establish where the counts of languages were, per 'macroarea,' which has been filter and displayed via their respective continent, as demonstrated:



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

Fig 17: *geographic distribution,* (Taylor, "IST 652," 2020).

Combination of the macroarea count findings, and the top-ranking language families, will allow for the geospatial analysis of the distribution of the top three language family groups, as demonstrated by the frequency count analysis:

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

No.1 Niger-Congo Language Family:

Geographic distribution is found throughout Africa; it is the worlds 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.

As demonstrated by the following: we can conduct a Geographic spatial analysis of Niger-Congo Language Family from the data:

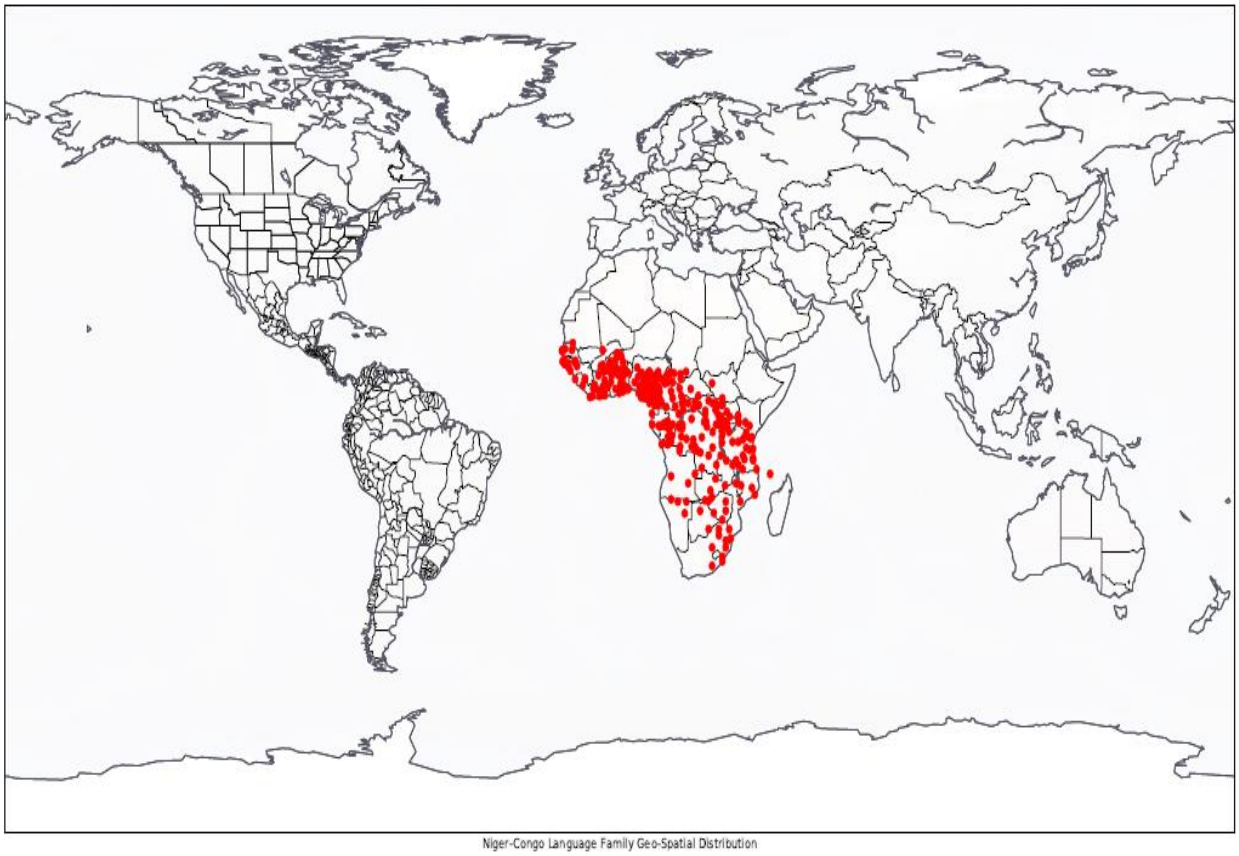


Fig 18: *Niger-Congo Language Family counts,* (Taylor, "IST 652," 2020).

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 below:

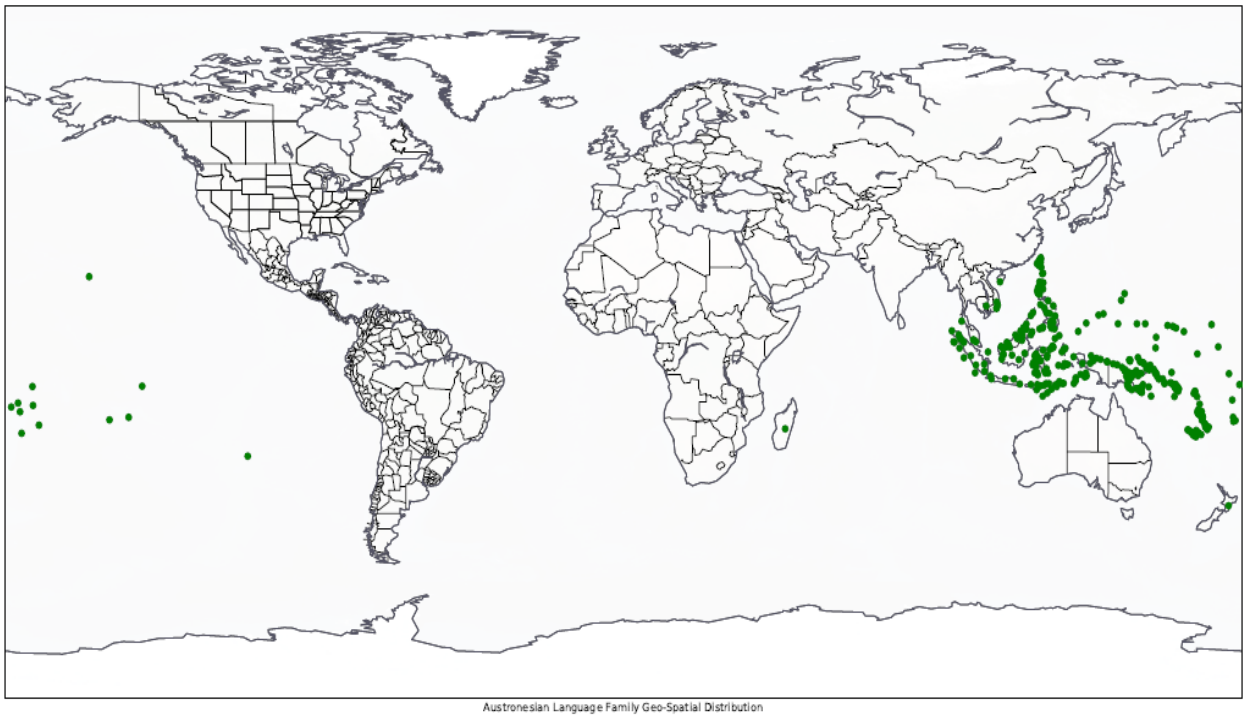


Fig 19: *Austronesian Language Family counts,* (Taylor, "IST 652," 2020).

No.3 Indo-European Language Family

Geographic distribution is found throughout: western Eurasia compromising 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 Caucus regions of southern Russia, therein radiating out westerly, and southernly through the passes of the Hindu Kush, into Northern India. Linguistic analysis of the various sub-divisions within the language demonstrates this, as demonstrated by the following:

"father"	"brother"
◦ *pitar* (Sanskrit) ◦ *pater* (Latin) ◦ *pater* (Greek) ◦ *padre* (Spanish) ◦ *pere* (French) ◦ *father* (English) ◦ *fadar* (Gothic) ◦ *faðir* (Old Norse) ◦ *vader* (German) ◦ *athir* (Old Irish--with loss of original consonant)	◦ *bhratar* (Sanskrit) ◦ *frater* (Latin) ◦ *phrater* (Greek) ◦ *frere* (French) ◦ *brother* (Modern English) ◦ *brothor* (Saxon) ◦ *bruder* (German) ◦ *broeder* (Dutch) ◦ *bratu* (Old Slavic) ◦ *brathair* (Old Irish)

Meaning:	Sanskrit	Latin:
"three"	*trayas*	*tres*
"seven"	*sapta*	*septem*
"eight"	*ashta*	*octo*
"nine"	*nava*	*novem*
"snake"	*sarpa*	*serpens*
"king"	*raja*	*regem*
"god"	*devas*	*divus* ("divine")

Linguistic analysis considered, geospatial analysis of the Indo-European language family shows the nature of its common distribution from its proto homeland, as seen below:

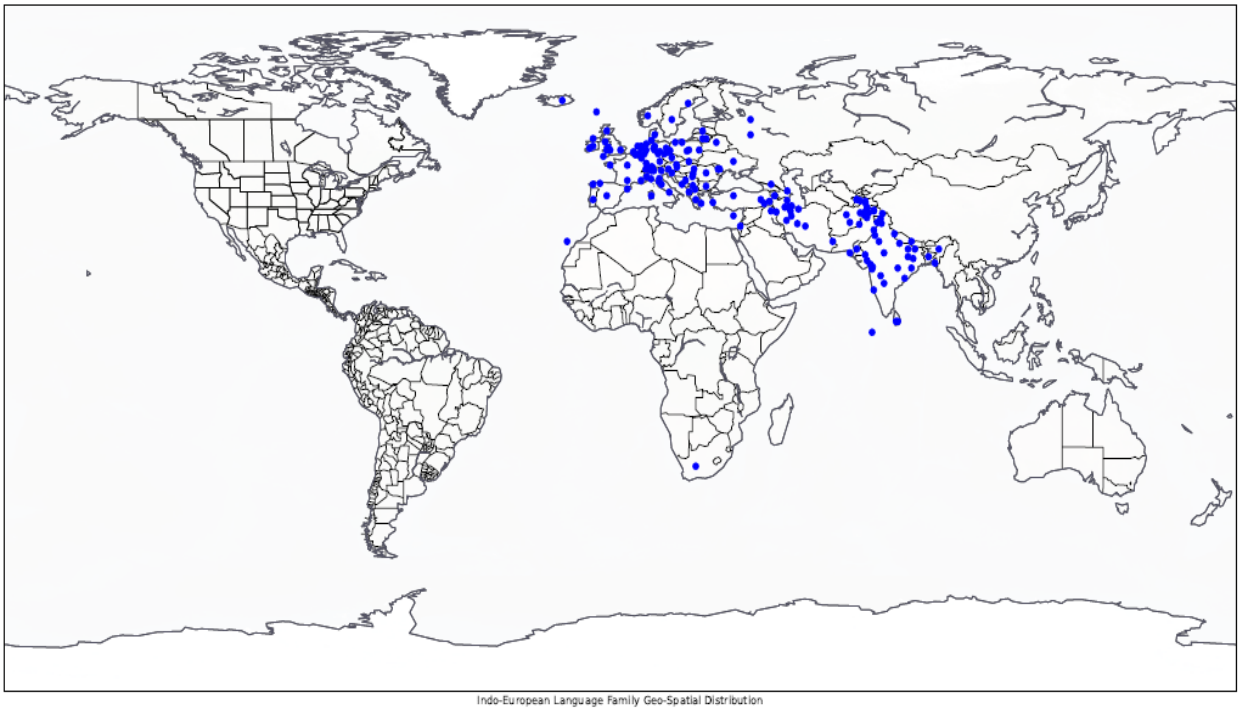


Fig 20: *Indo-European Language Geospatial analysis,* (Taylor, "IST 652," 2020).

Interesting observation; the vestige of one of the last Indo-European invasions

b. Reflection & 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: 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.

Python Program

The python program written to complete the analysis of this dataset was compiled within the Anaconda suites, Jupyter Notebooks. The following are the libraries utilized within the program:

#In order to complete the exploratory analysis of the wals dataset, the
#researcher will require the following libraries:

```
#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
from geopandas import GeoDataFrame
from shapely.geometry import Point
from ipyleaflet import *
from ipyleaflet import Map, GeoData, basemaps, LayersControl
import geopandas
import requests
import urllib.request
import folium
import json
from pymongo import MongoClient
from IPython.display import HTML
from folium.plugins import HeatMap

%matplotlib inline
```

Fig 21: *code,* (Taylor, "IST 652," 2020).

IST 707: Data Analytics: Professor Jeremy Bolton, PhD

a. Project Description

The Scripting for Data Analysis course chosen by the candidate was facilitated through course professor Jeremy Bolton, PhD. Through the utilization of R code scripting, and various data mining techniques, machine learning technique and geospatial analysis, the candidate created a final project. The final project deliverable for the course was to create a data

source from several sources format and unformatted data and to demonstrate the following:

> The objective of the project is to use the main skills taught in this class to solve a real data mining problem. Students can choose to work individually or pair up with another student. For this project, you must choose your own dataset. It can be one that you created yourself or found from other resources, such as the Kaggle competitions and the UCI repository (http://archive.ics.uci.edu/ml/). The problem may use one or more of the types of data mining algorithms that we have studied this semester: Classification, Clustering and Association Rules, in an investigation of the solution to the problem (Taylor, "IST 707," 2020).

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).

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, as stated:

> CDC's Role: * Detecting and responding to new and emerging health threats * Tackling the biggest health problems causing death and disability for Americans * Putting science and advanced technology into action to prevent disease * Promoting healthy and safe behaviors, communities and environment * Developing leaders and training the public health workforce, including disease detectives * Taking the health pulse of our nation (Taylor, "IST 707," 2020).

CDC produces Community Health Status Indicators (CHSI) for all 3,143 counties in the United States. Each profile includes key indicators of health outcomes, which describes the population health status of a county and factors that have the potential to influence health outcomes, such as health care access and quality, health behaviors, social factors, and the physical environment.

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.

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).

The data set has a broad array of health metrics as well as statistics around vulnerable populations, life expectancy and death rates. In reviewing the raw data we felt the goal was to give local public health agencies a set of tools that could help improve the health of their community by identifying root causes and at-risk populations. Website: https://healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer

Data Acquisition, Cleaning, Transformation

The file was a zip file that contained several CSV files:

* DATA_ELEMENT_DESCRIPTION.csv defines each data element and indicates where its description is found in Data Sources, Definitions, and Notes.

* DEFINED_DATA_VALUE.csv defines the meaning of specific values (such as missing or suppressed data).

* HEALTHY_PEOPLE_2010.csv identifies the Healthy People 2010 Targets and the U.S. Percentages or Rates.

* DEMOGRAPHICS.csv identifies the data elements and values in the Demographics indicator domain.

* LEADING_CAUSES_OF_DEATH.csv identifies the data elements and values in the Leading Causes of Death indicator domain.

* SUMMARY_MEASURES_OF_HEALTH.csv identifies the data elements and values in the Summary Measures of Health indicator domain.

* MEASURES_OF_BIRTH_AND_DEATH.csv identifies the data elements and values in the Measures of Birth and Death indicator domain.

* RELATIVE_HEALTH_IMPORTANCE.csv identifies the data elements and values in the Relative Health Importance indicator domain.

* VULNERABLE_POPS_AND_ENV_HEALTH.csv identifies the data elements and values in the Vulnerable Populations and Environmental Health indicator domain.

* PREVENTIVE_SERVICES_USE.csv identifies the data elements and values in the Preventive Services indicator domain.

RISK_FACTORS_AND_ACCESS_TO_CARE.csv identifies the data elements and values in the Risk Factors and Access to Care indicator domain.

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. As such, we looked to better understand the national averages and distribution before we dove into a region or county (Figure 2).

So far our statistical analysis has highlighted a largely white population in the age range of 19-64, which is not all that surprising, as exampled below:

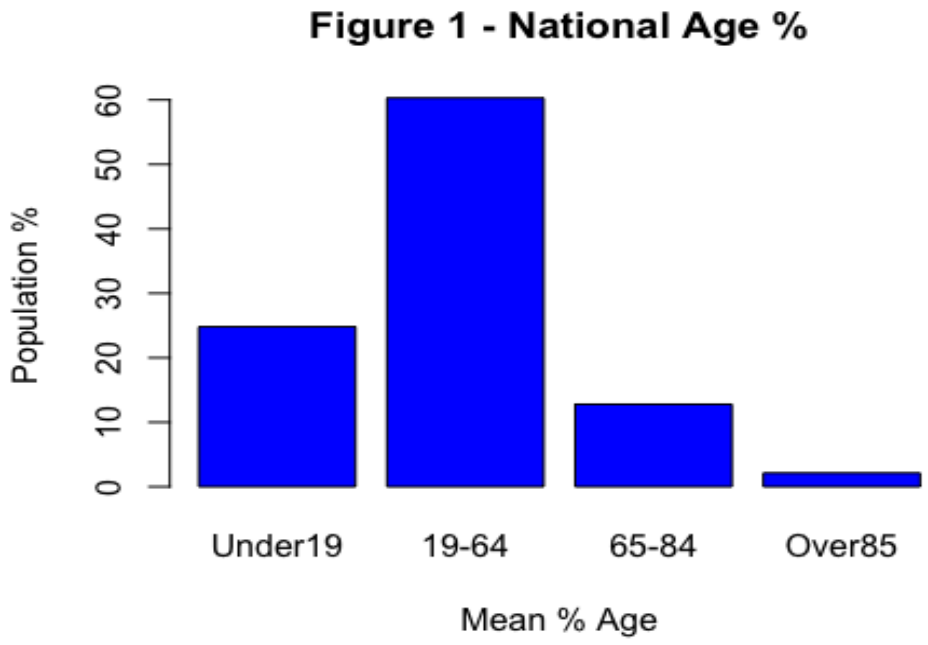


Fig 22 *age distribution.* (Taylor, "IST 707," 2020).

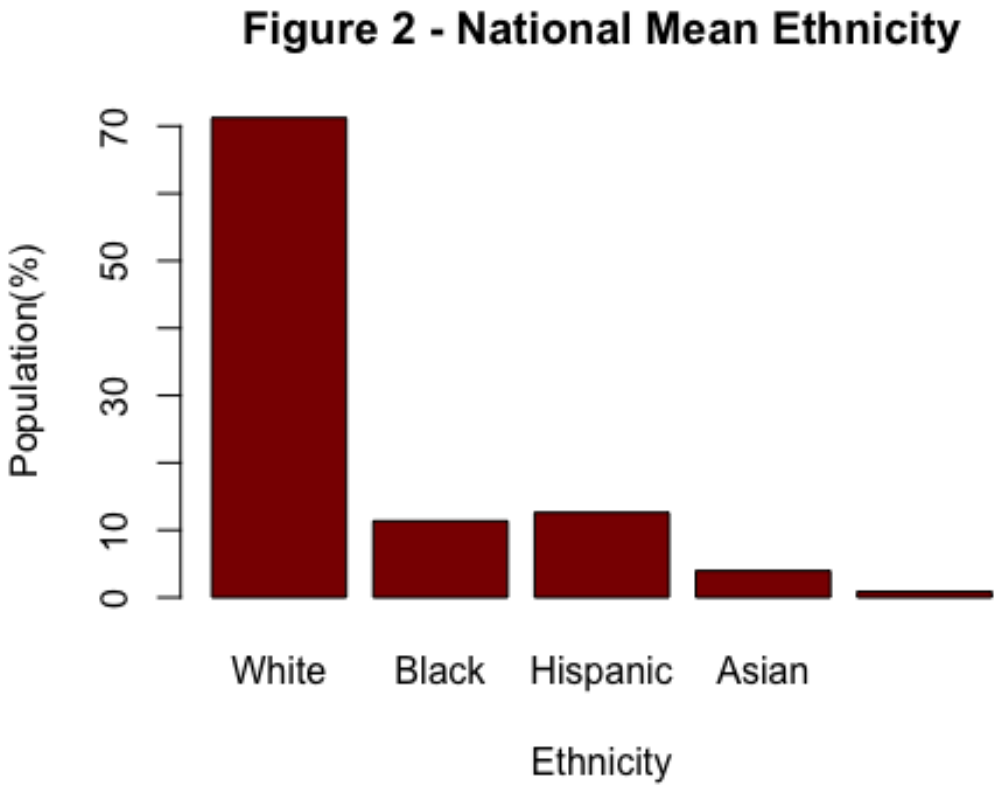


Fig 23 *ethnicity distribution.* (Taylor, "IST 707," 2020).

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 5). This shows the mental state of

the population to a certain extent. ~25% of death is caused due to suicide and homicide which is a high number of the population. Below states seem to have a higher rate in 2015 * Texas * New England * Montana * Colorado

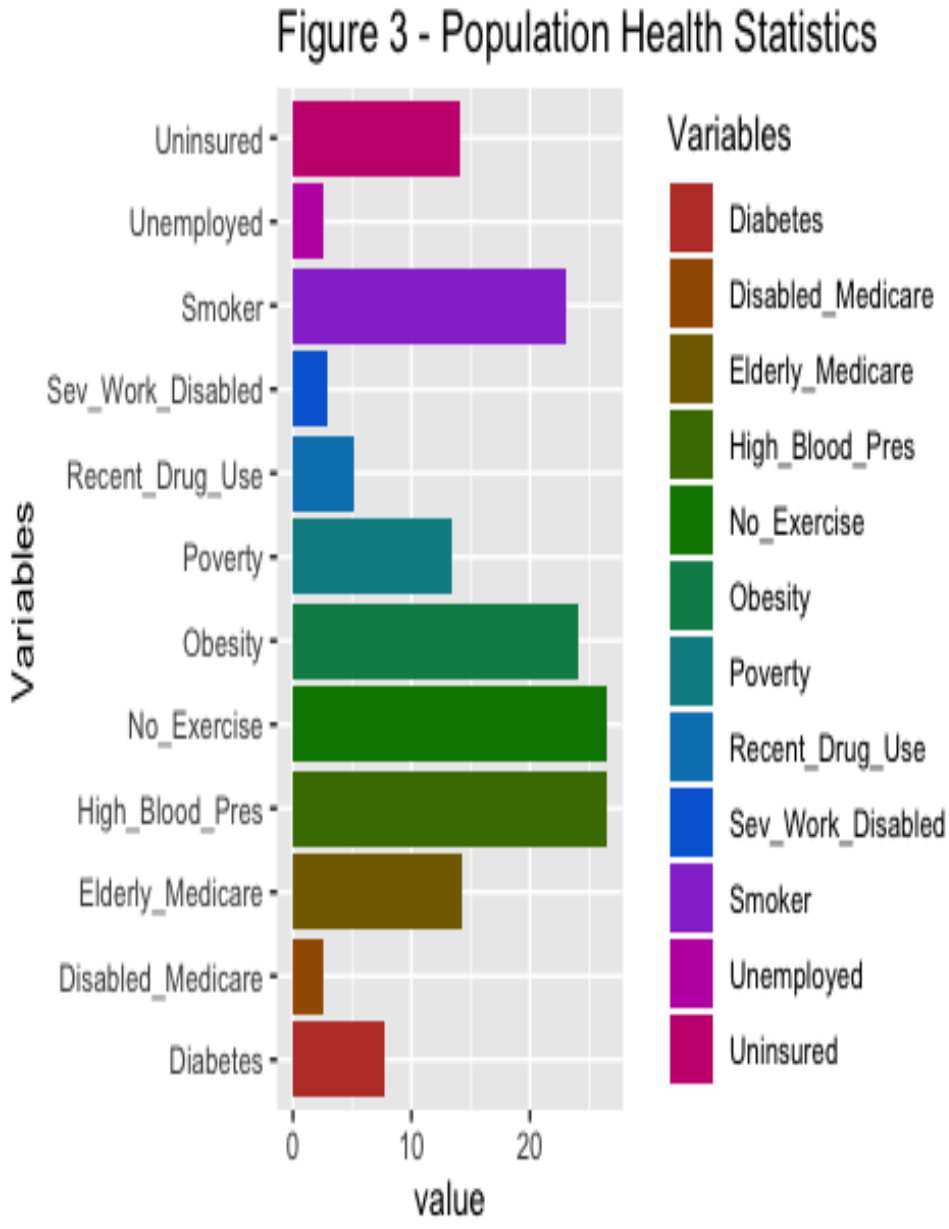


Fig 24 *at risk health distribution.* (Taylor, "IST 707," 2020).

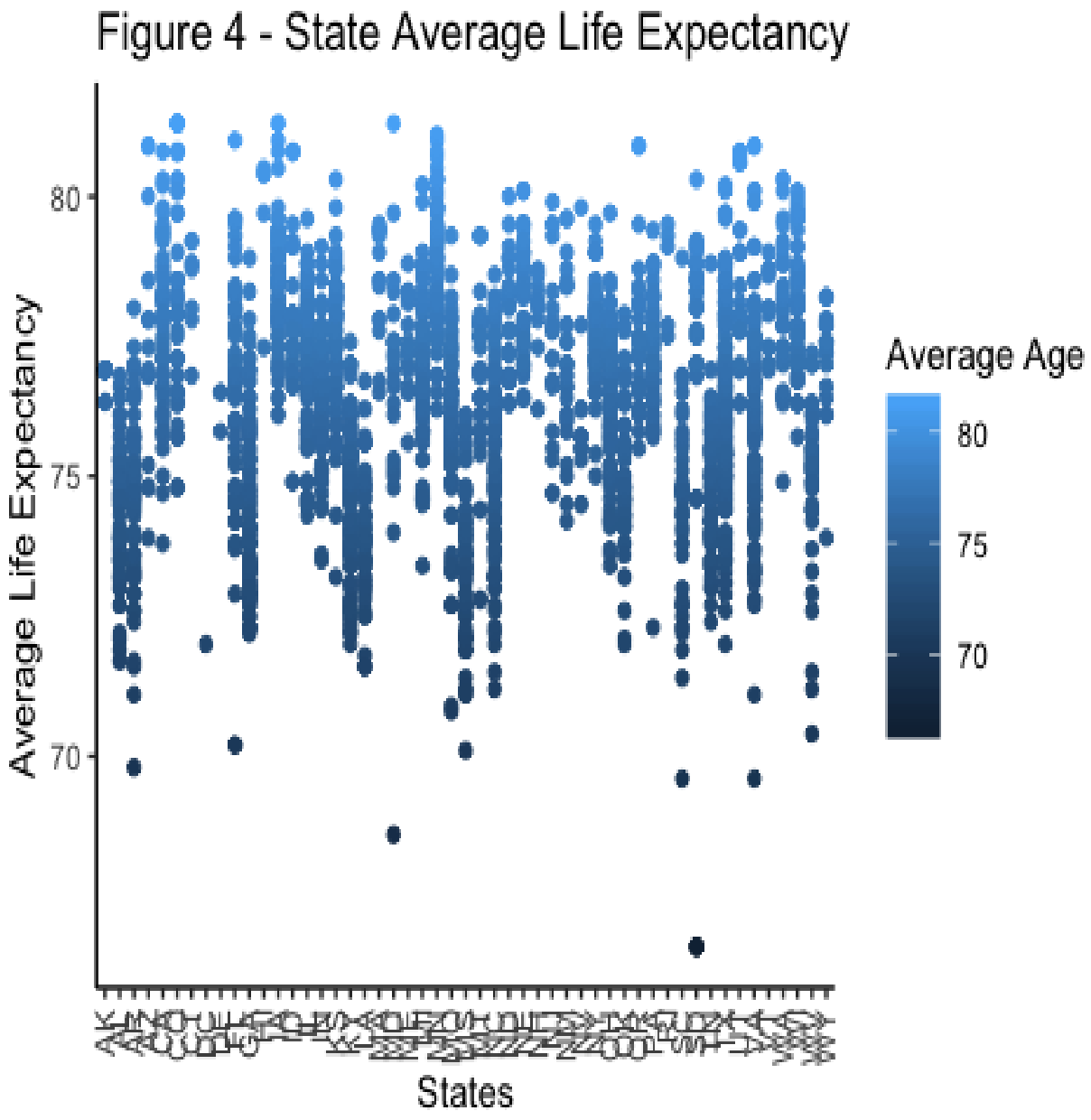


Fig 25 *average life expectancy distribution.* (Taylor, "IST 707," 2020).

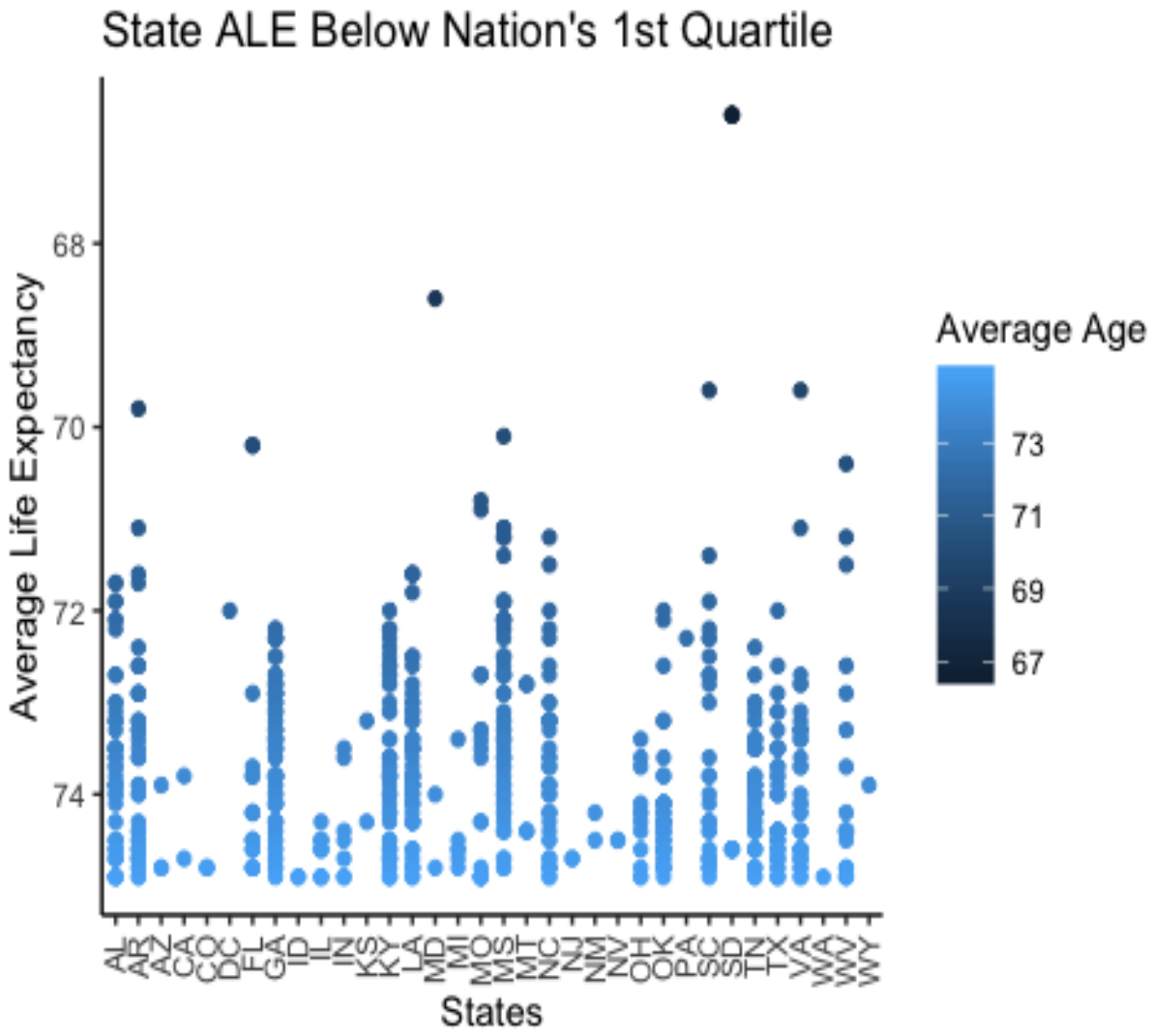


Fig 26 *average life expectancy distribution.* (Taylor, "IST 707," 2020).

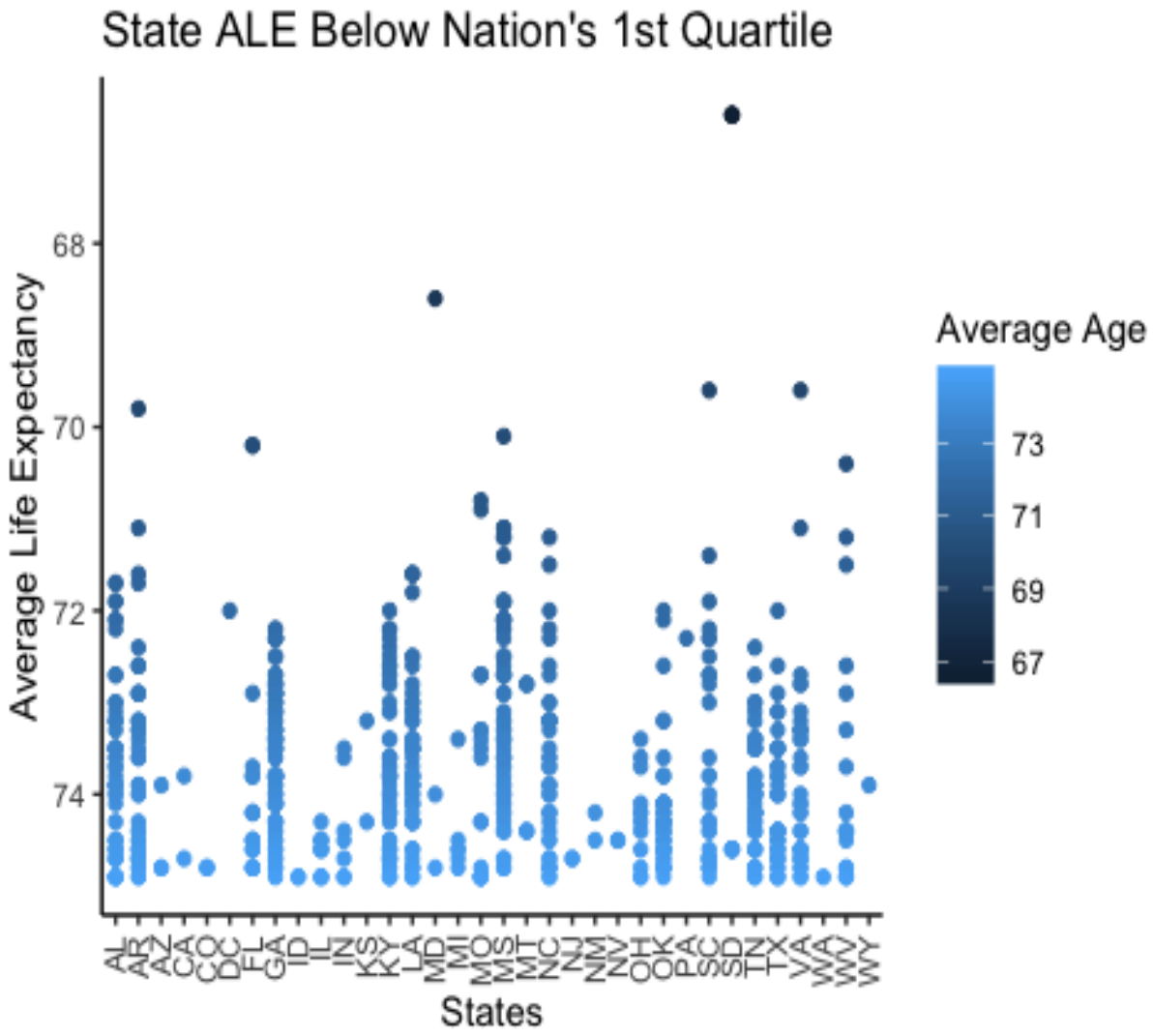


Fig 27 *average life expectancy distribution.* (Taylor, "IST 707," 2020).

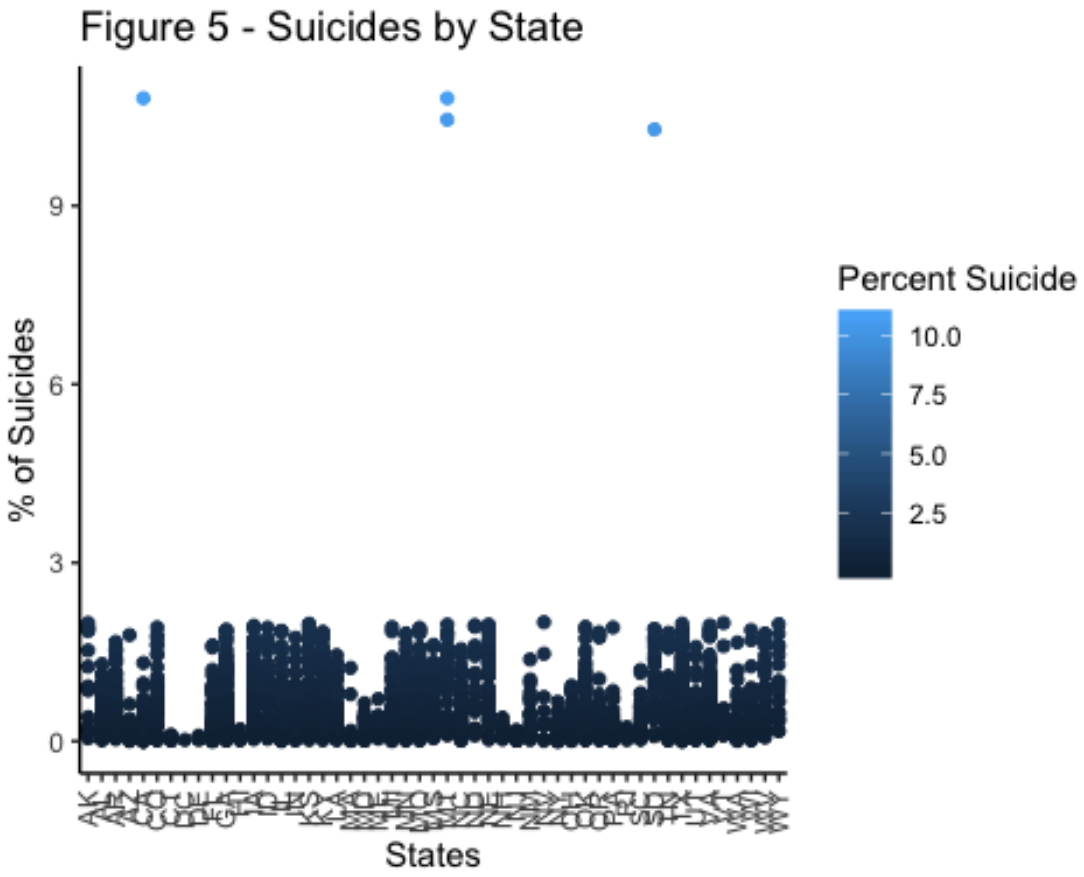


Fig 28 *suicide distribution.* (Taylor, "IST 707," 2020).

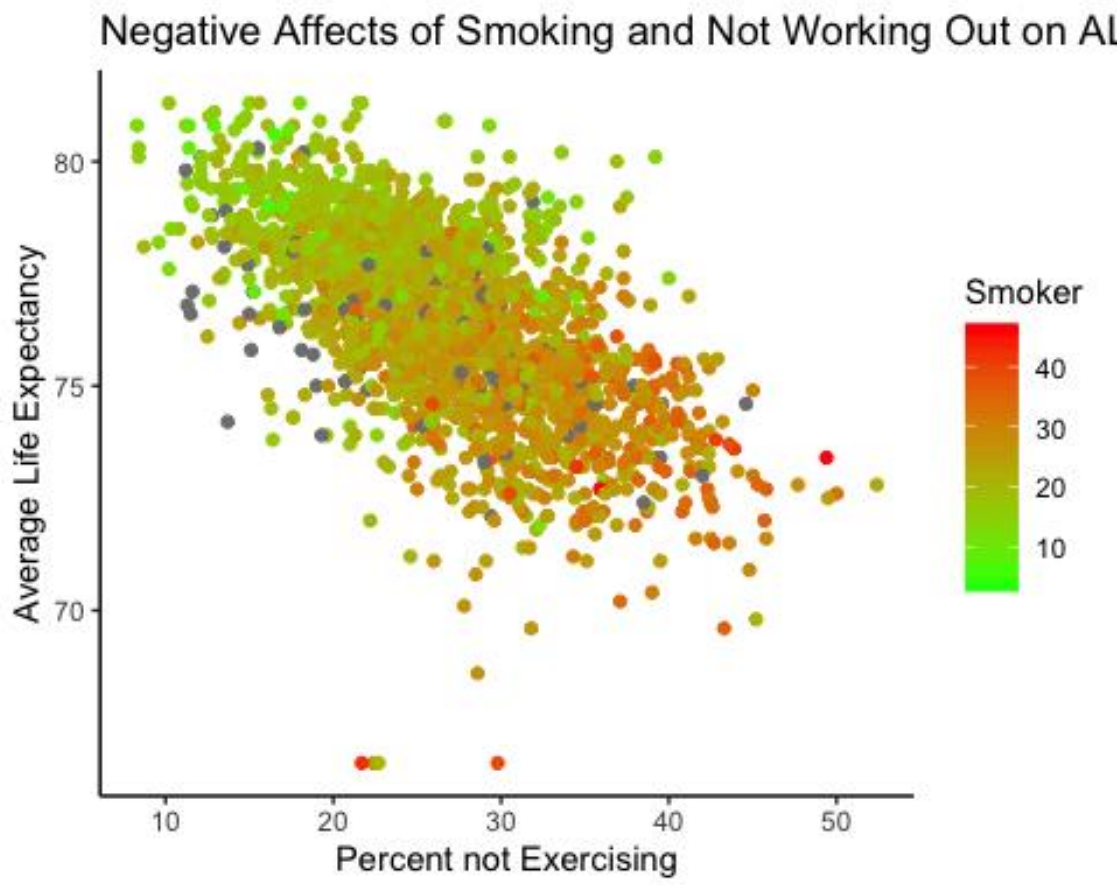
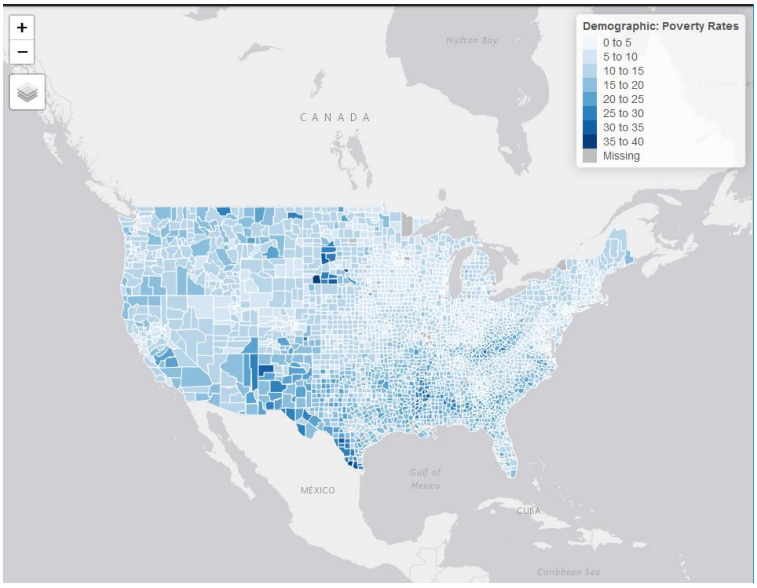


Fig 29 *smoking cluster, note regression distribution.* (Taylor, "IST 707," 2020).

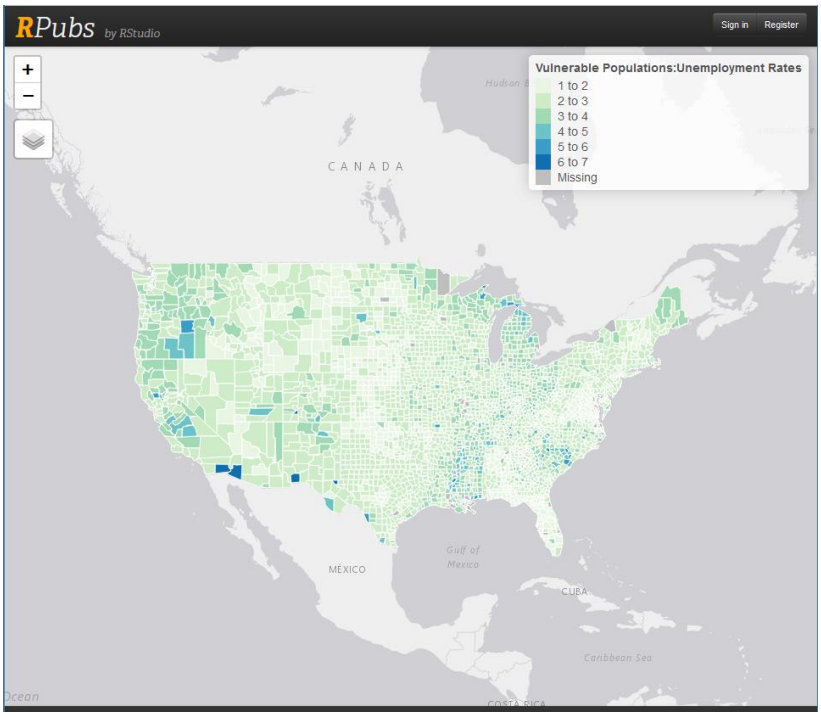
To assist in the relationship of this vast amount of data, and the understanding of how the data correlates and relates to one another, K-means clustering was done, based upon geographic analysis, and, interactive maps were created by the candidate to assist with the intake of this information, as referenced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

AT Risk Demographic, Categories from EDA with similarity clustering

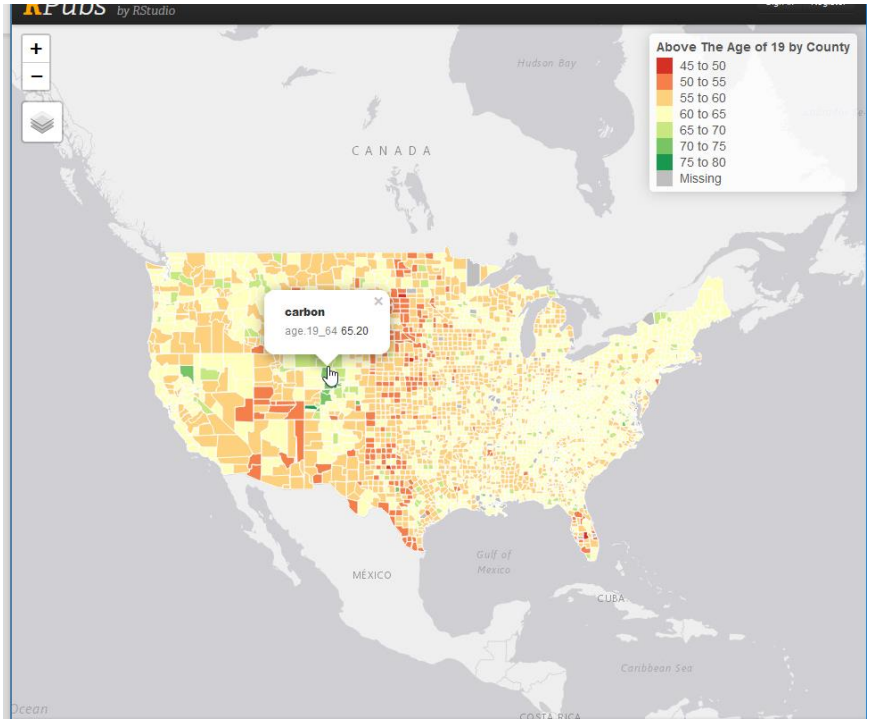
http://rpubs.com/randallscott25/chsiDemoPOV



http://rpubs.com/randallscott25/chsiVULNunemploy



http://rpubs.com/randallscott25/chsiDemoABV19



http://rpubs.com/randallscott25/526493

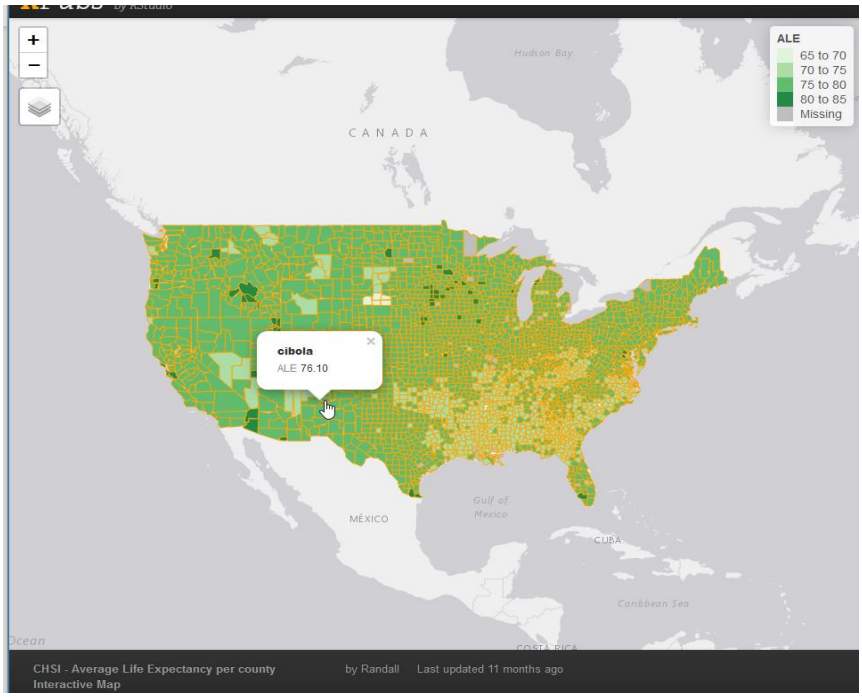


Fig 35 *interactive maps distribution.* (Taylor, "IST 707," 2020).

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:

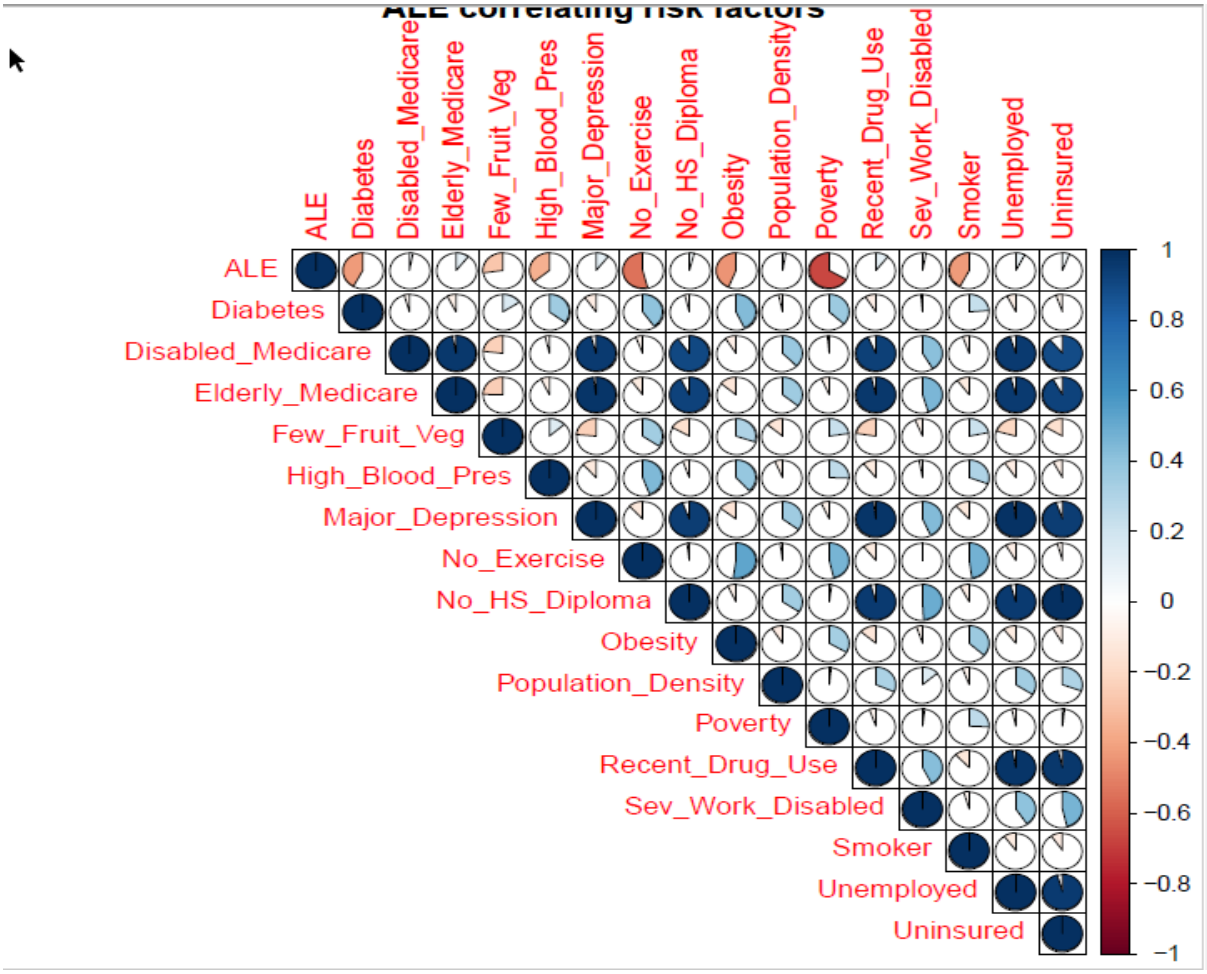


Fig 36 *correlation matrix.* (Taylor, "IST 707," 2020).

A linear model was built using variables like Age, Demographics, Drug Usage, Use of Toxic Chem, No Exercise, Fruit intake, Blood Pressure, Diabetes and an R-squared value of 71.2% was obtained that explained the variability on Average life expectancy in a county.

* Ethnicity does not impact ALE much but it is seen that impact of longer ALE is in the order of Native American, White, Hispanic, Black and Asian It is seen that as the number of people having HIV, Cancer, Diabetes, Heart Disease increases ALE decreases. The order of negative impact on ALE is Heart disease, Smoking, Diabetes, BP, Obesity, HIV

* It is seen that as the number of people having HIV, Cancer, Diabetes, Heart Disease increases ALE decreases. The order of negative impact on ALE is Heart disease, Smoking, Diabetes, BP, Obesity, HIV

* A strong relationship is seen between Unemployed, Drug Usage and Uninsured Being Uninsured has a strong relationship with unemployment, drug usage, receiving elderly Medicare, and disabled Medicare

* Diabetes is having a moderate to strong relationship with obesity, No-exercise, Blood Pressure

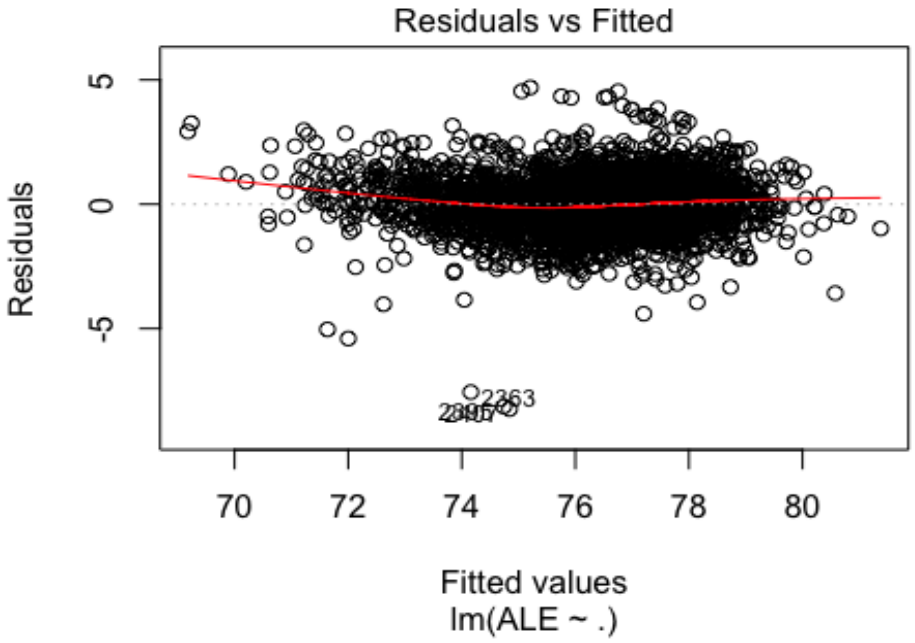
* Major Depression and Drug Use are very strongly related. It is noticed that less Fruit intake is moderately related to no-exercise and obesity

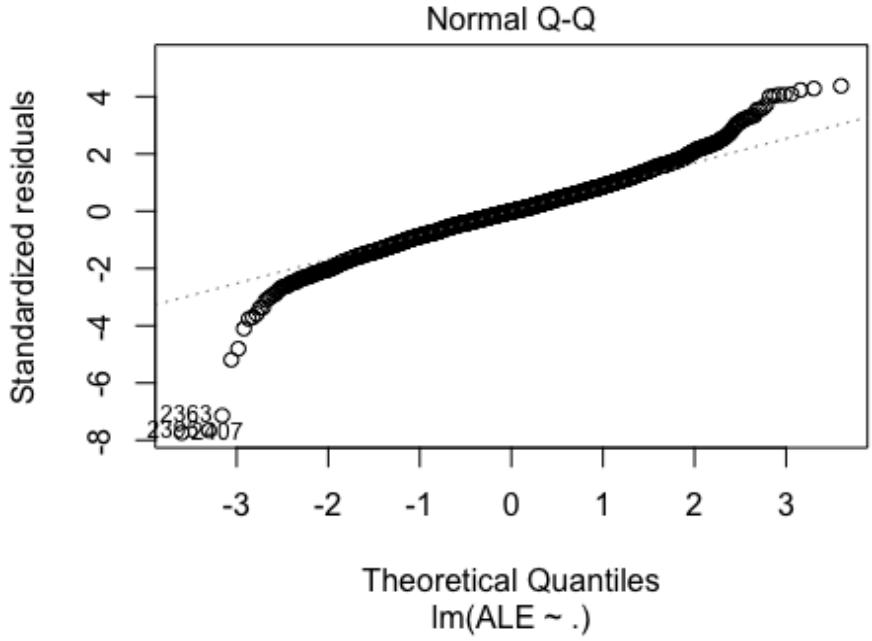
Linear Regression Plots

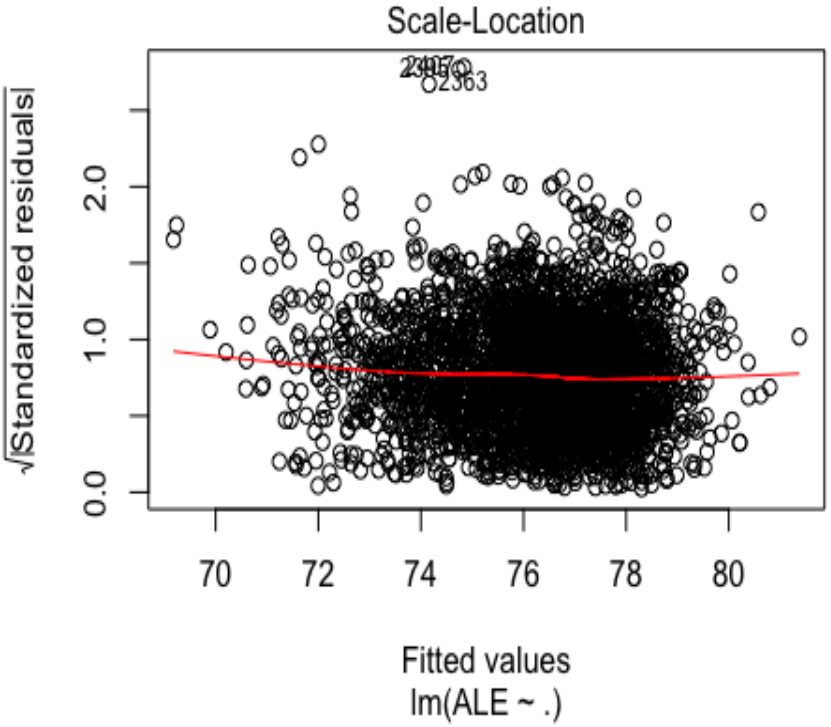
1. Residual vs. Fitter Plot - We see there is no specific pattern in the residuals predicted by our model so we can eliminate the possibility of heteroskedasticity and also possible outliers.

2. QQ Plot - Though most of the points seem to fall on the line which indicates that our residuals come from a normal distribution, there are some points that stray from the line in the lower and upper quantiles of the plot. It is possible that these points do not come from a normal distribution, but most of our points seem to come from a normal distribution so there is not a lot to worry about here.

3. Leverage Plot- This plot graphs the standardized residuals against their leverage. It also includes the Cook's distance boundaries. Any point outside of those boundaries would be an outlier in the x direction. Since we cannot even see the boundaries on our plot, we can conclude that we have no outliers.




Normal Q-Q
Standardized residuals
4
2
0
-2
-4
-6
-8
2363
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(ALE ~ .)


Scale-Location
√|Standardized residuals|
2.0
1.0
0.0
2363
70
72
74
76
78
80
Fitted values
lm(ALE ~ .)

K-means Cluster Analysis

Cluster analysis is a popular classification technique frequently used to analyze market research data which divides the data into groups. Data appears in rows, purchase intent scores for example, and columns, sales concepts for instance. Rows can then be clustered with respect to columns or columns with respect to rows. For example, clustering techniques can be used to identify demographic or psychographic characteristics of consumers with similar purchasing histories, or to isolate differences between groups of products. To understand the data story better, our research team chose to understand the data points relationship to one another utilizing unsupervised K-means clustering. To find these clusters, we utilized Lloyd's Algorithm in the following manner: we start out with k random centroids. A centroid is simply a datapoint around which we form a cluster. For each centroid, we find the datapoints that are closer to that centroid than to any other centroid. We call that set of datapoints its cluster, as demonstrated below:

Then we take the mean of the cluster and let that be the new centroid. We repeat this process (using the new centroids to form clusters, etc.) until the algorithm stops moving the centroids.

```
Cluster sizes:

[1] "644 309 433 501 311"

Data means:

   diabetes no.exercise     obesity     poverty      smoker
  0.3618741   0.4352160   0.5543834   0.3113270   0.4502348

Cluster centers:

   diabetes no.exercise   obesity   poverty    smoker
1 0.4029115   0.4268866 0.5810342 0.2222742 0.4479542
2 0.4664020   0.6463567 0.6838661 0.4284614 0.5904513
3 0.2459209   0.2805143 0.4302326 0.1877765 0.3447293
4 0.2890275   0.4380441 0.5389260 0.3248473 0.4721674
5 0.4518314   0.4535139 0.5683005 0.5295878 0.4272040

Within cluster sum of squares:

[1] 18.60534 19.75330 20.76333 14.39275 14.19160
```

We do this in order to minimize the total sum of distances from every centroid to the points in its cluster — that is our metric for how well the clusters split up the data.

```
========================================================================
General cluster statistics:

$n
[1] 2198

$cluster.number
[1] 5

$cluster.size
[1] 644 309 433 501 311

$min.cluster.size
[1] 309

$noisen
[1] 0

$diameter
[1] 28.02858 34.25872 30.77295 26.02153 40.75496

$average.distance
[1]  7.624265 11.758470 10.488787  7.809530  8.543890

$median.distance
[1]  7.300710 11.123848 10.000578  7.623948  7.990162

$separation
[1] 0.200000 1.414214 1.284523 0.100000 0.100000
```

General cluster statistics, identified by R analysis

Distance Measure:

The choice of distance measures is especially important, as it has a strong influence on the clustering results. For most common clustering software, the default distance measure is the Euclidean distance.

Within R it is simple to compute and visualize the distance matrix using the functions get_dist and fviz_dist from the factoextra R package. This starts to illustrate which states have large dissimilarities (red) versus those that appear to be similar (teal).

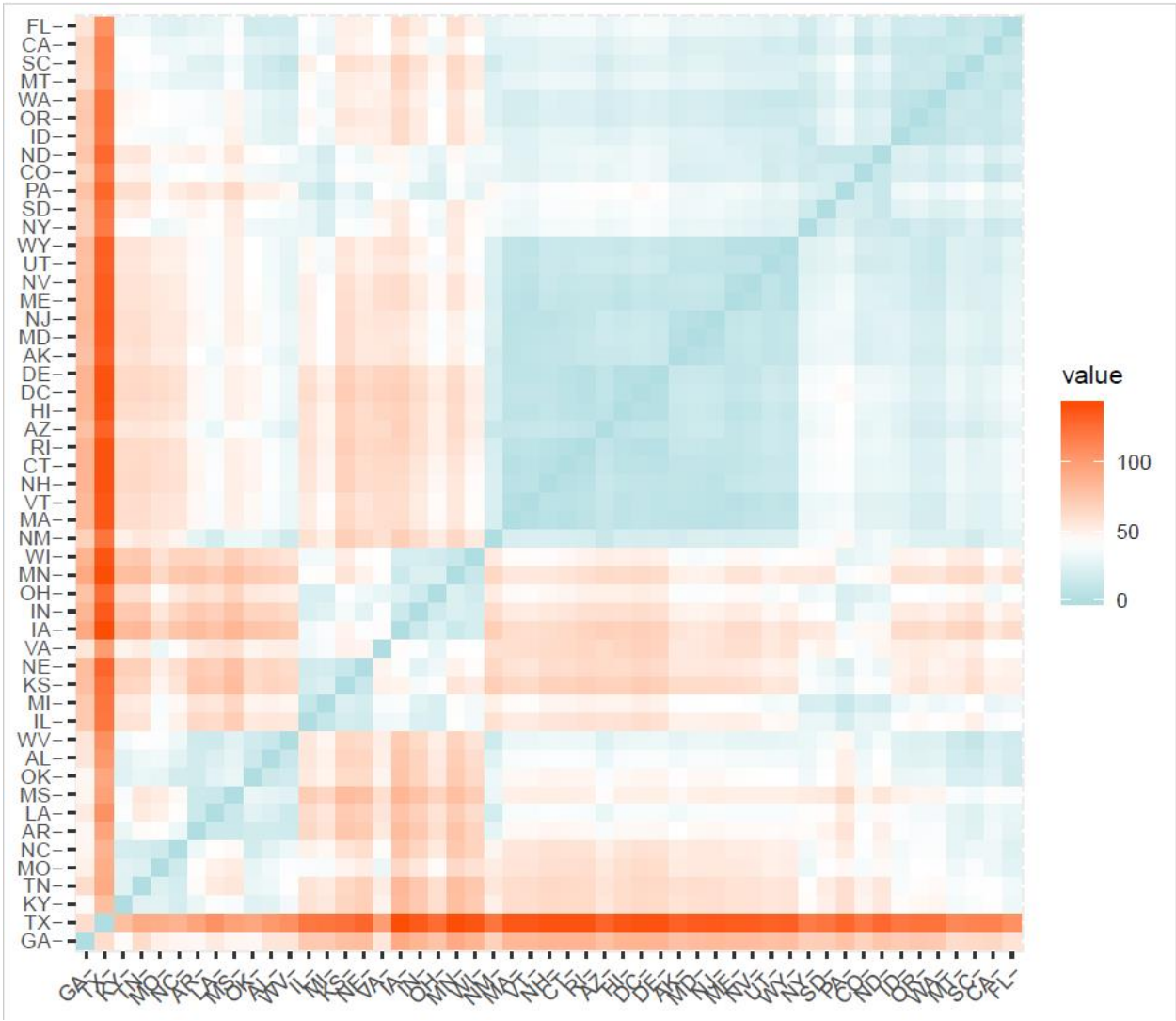


Fig 37 *correlation matrix.* (Taylor, "IST 707," 2020).

Finding 'k': number of clusters using the elbow method

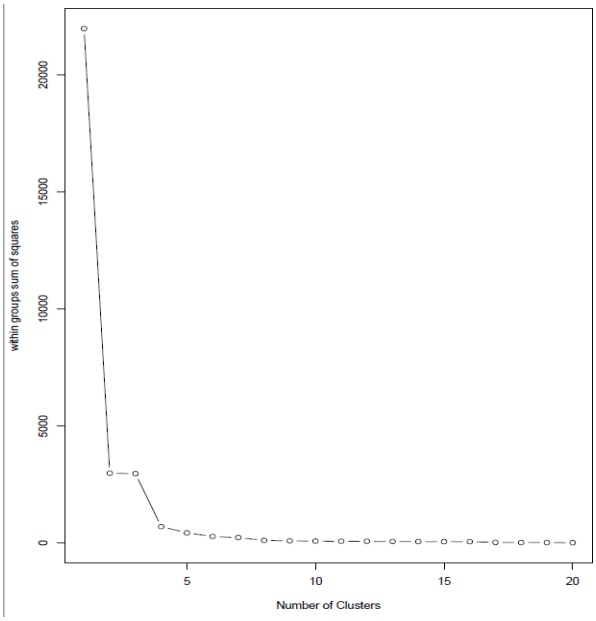
The initial cluster analysis was done utilizing a tool within R, the program rattle. The research team then needed to validate the clusters, to ensure that the initial research regarding cluster ability was sound.

To validate the total clusters of 5, the research team started to conduct validation utilizing the following method:

Chosen variables from the dataset were established as atRisk, and created into a data.matrix. The atRisk data having been established as a data.matrix was then scale the data, to normalize the data by subtracting the mean and dividing by the standard deviation, to take out the effect of different variables, being measured on different scales, and to eliminate the occurrence of NA in our dataset, as this will interfere with the ability of the algorithm to set, and measure, centroids in the k-means clustering.

Having scaled the data we run an initial, quick function

```
wssplot = function(Test1, nc=20, seed=123){
  wss = (nrow(Test1)-1)*sum(apply(Test1, 2, var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] = sum(kmeans(Test1, centers = i)$withinss)}
 plot(1:nc, wss, type='b', xlab='Number of Clusters', ylab = "within groups sum of squares")}
```

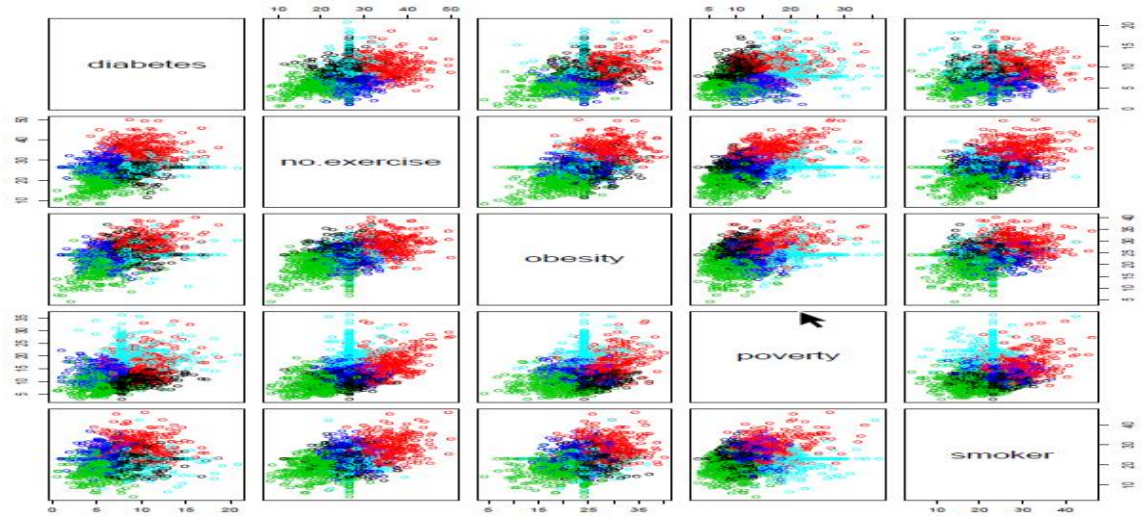


Three optimization plots: Gap-stat, within-groups sums of squares, and the silhouette method were chosen for optimization.

K-means analysis

Having utilized the above three methods, the research team decide to choose 5 as the total amount of clusters to express in the k-means analysis. Given the

previously ran correlation matrix, and initial linear regressions ran; the following visualization shows the initial cluster visualizations ran in R:

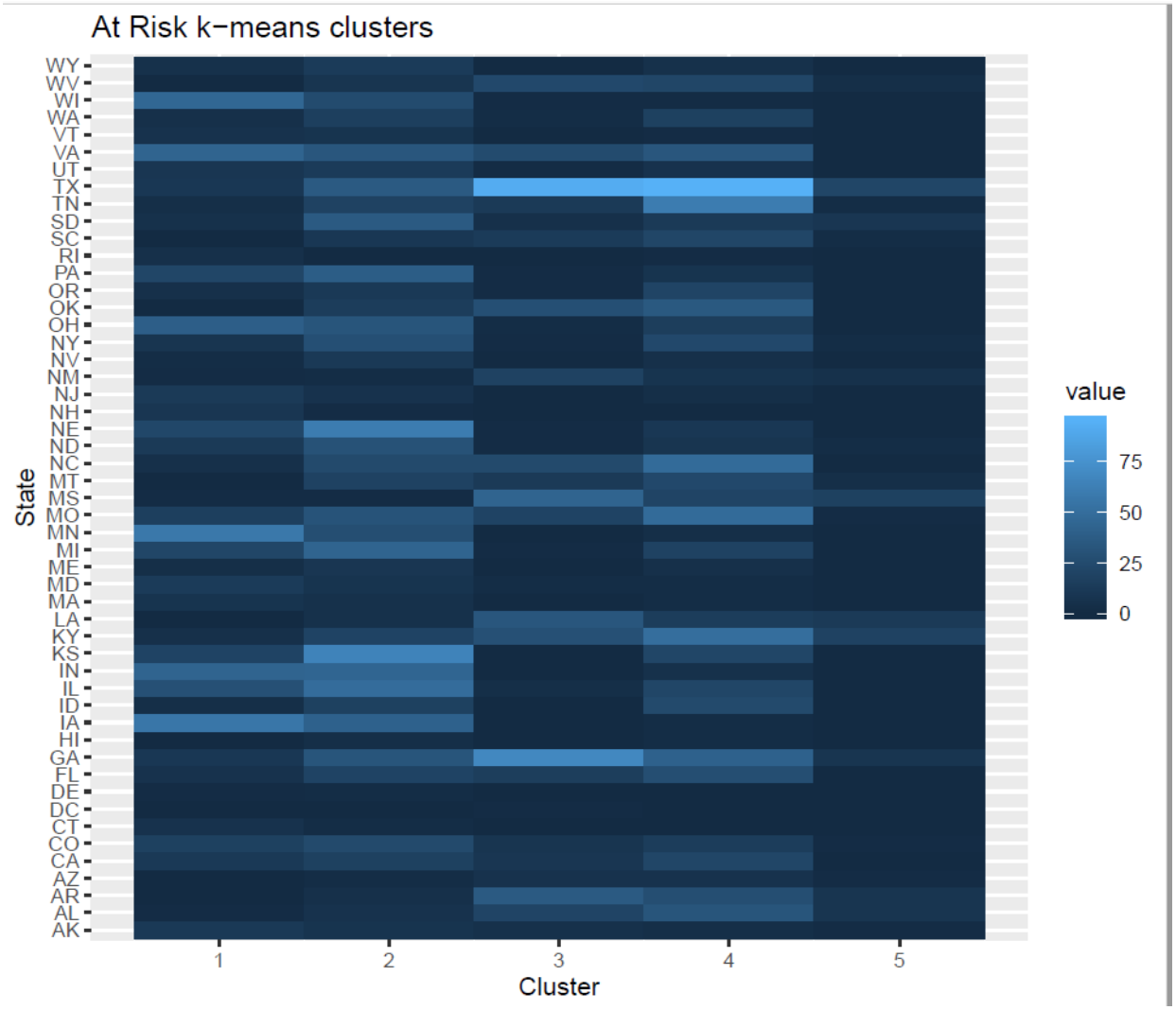



```
      1  2  3  4  5
AK 11  7  5  3  1
AL  1  6 20 32  8
AR  0  4 36 27  8
AZ  0  2  6  6  1
CA 11 17  9 21  0
CO 17 24  8 14  1
CT  6  2  0  0  0
DC  0  0  1  0  0
DE  1  2  0  0  0
FL  6 19 15 27  0
GA 10 32 68 41  8
HI  0  3  0  2  0
IA 56 41  0  2  0
ID  3 17  0 24  0
IL 29 49  3 21  0
IN 43 44  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
```

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.

Computing k-means clustering in R

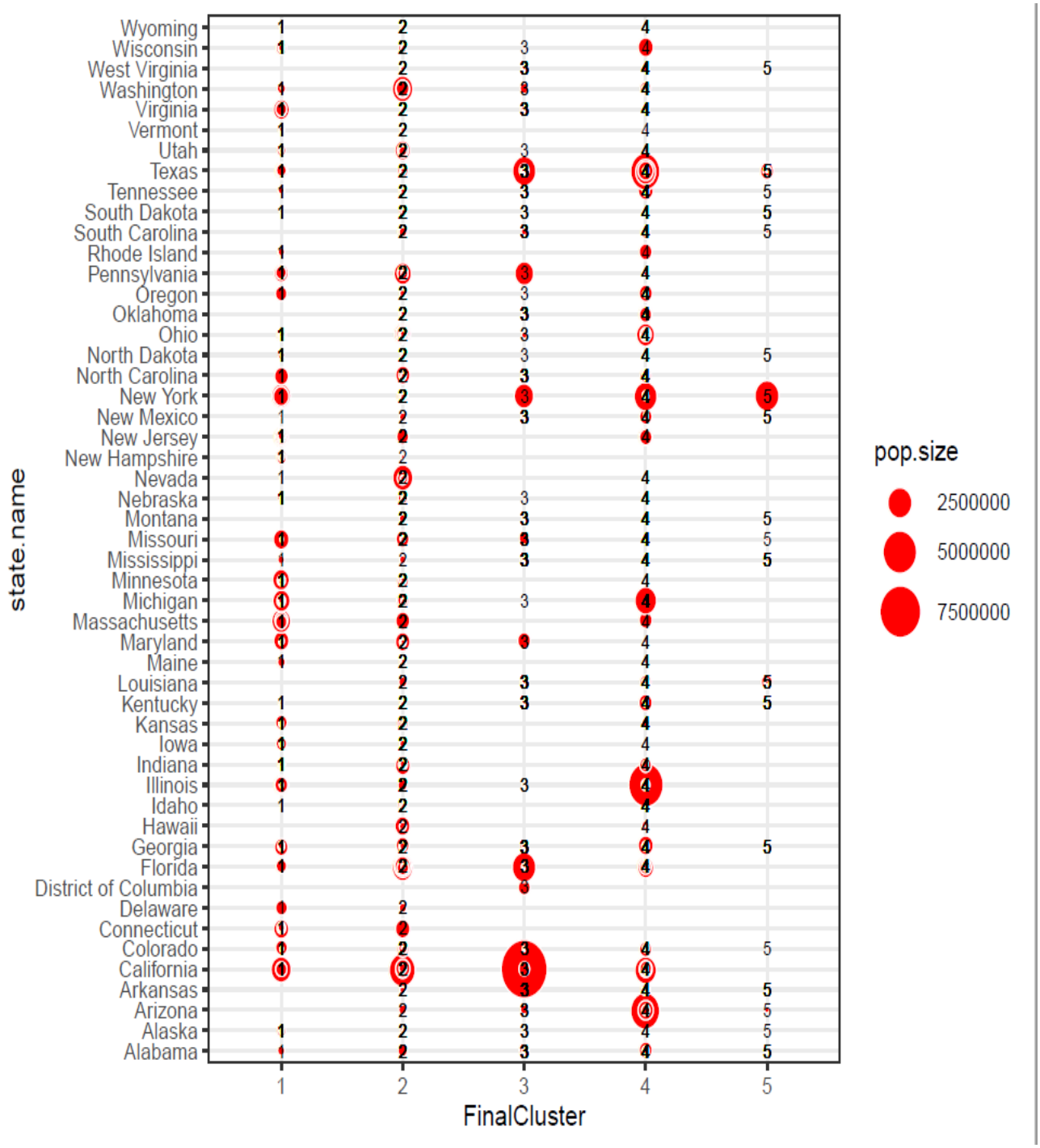
We can compute k-means in R with the kmeans function. Here will group the data into two clusters (centers = 5). The kmeans function also has an nstart option that attempts multiple initial configurations and reports on the best one. For example, adding nstart = 20 will generate 20 initial configurations



Extracting Results

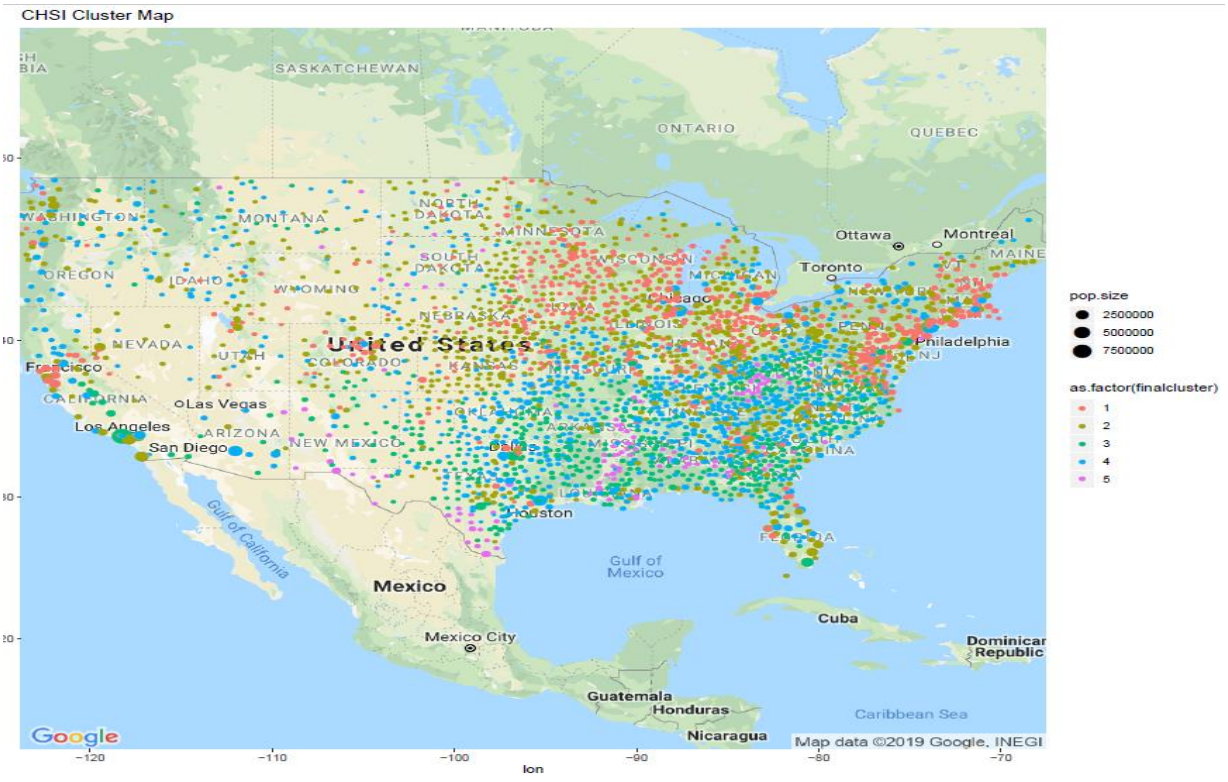
With most of these approaches suggesting 5 as the number of optimal clusters, we can perform the final analysis and extract the results using 5

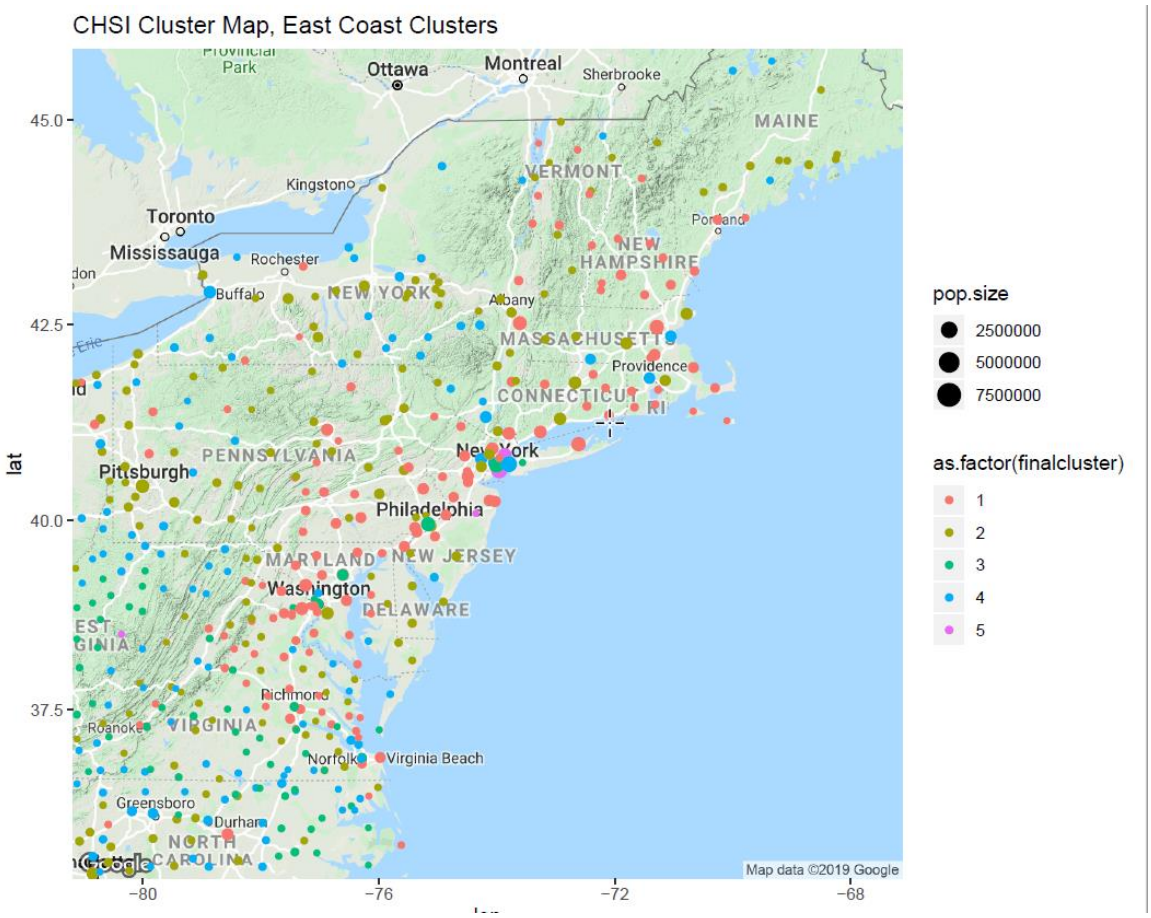
clusters. In order to better understand these clusters, the research team utilized the fviz_cluster package, the ggplot2 package, and ggmap r packages for visualization.

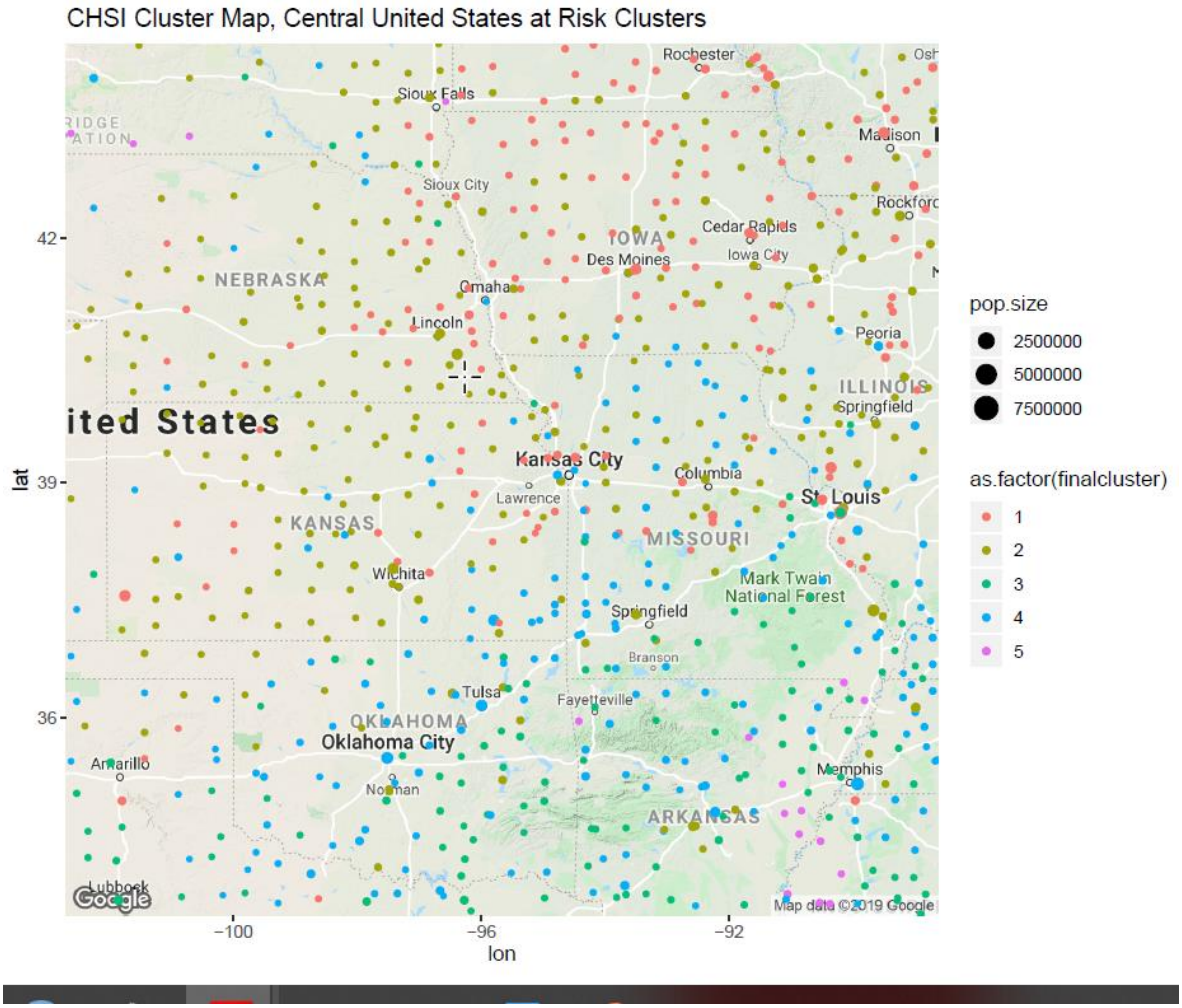


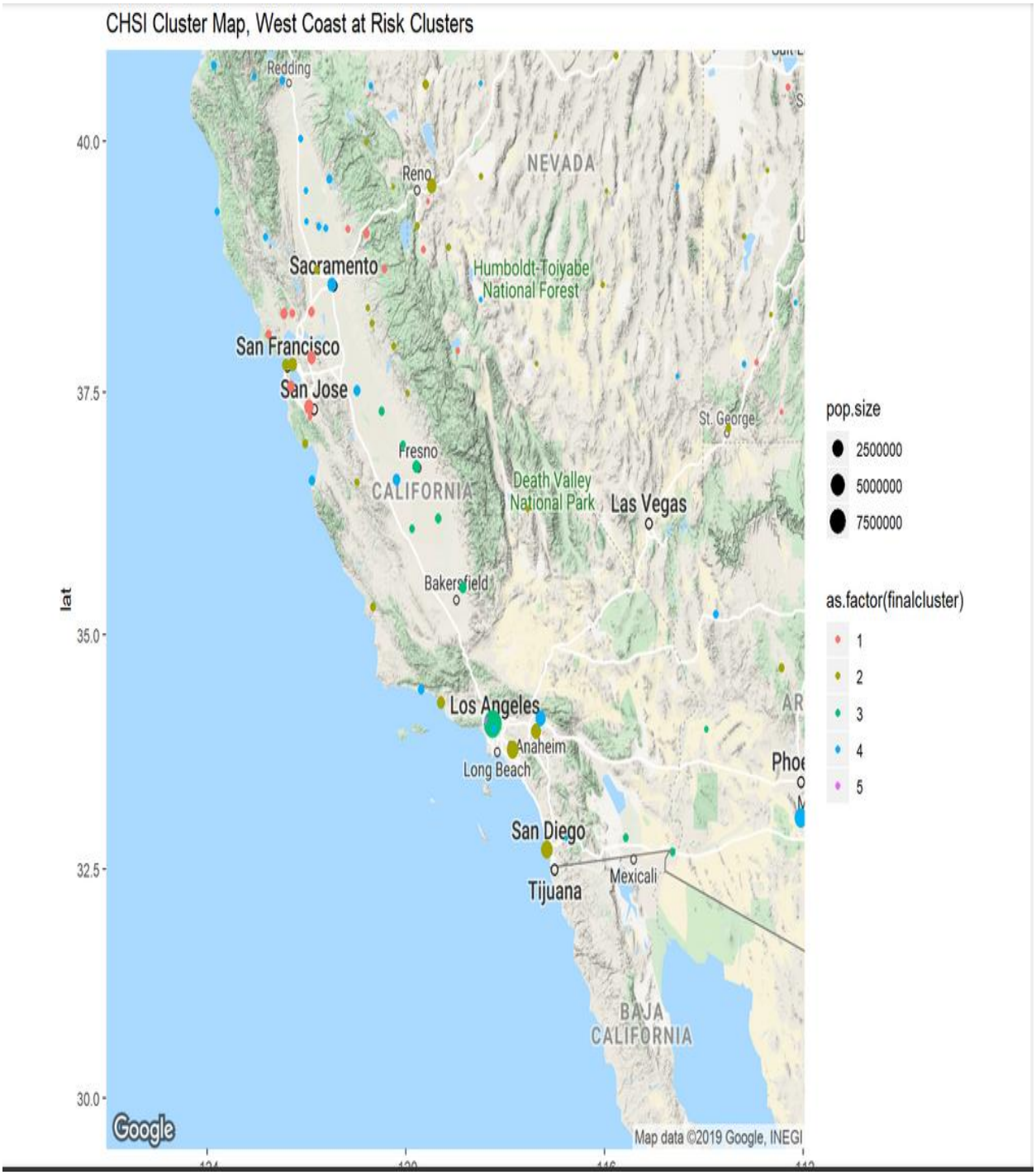
Final cluster chart, At Risk factors, CHSI

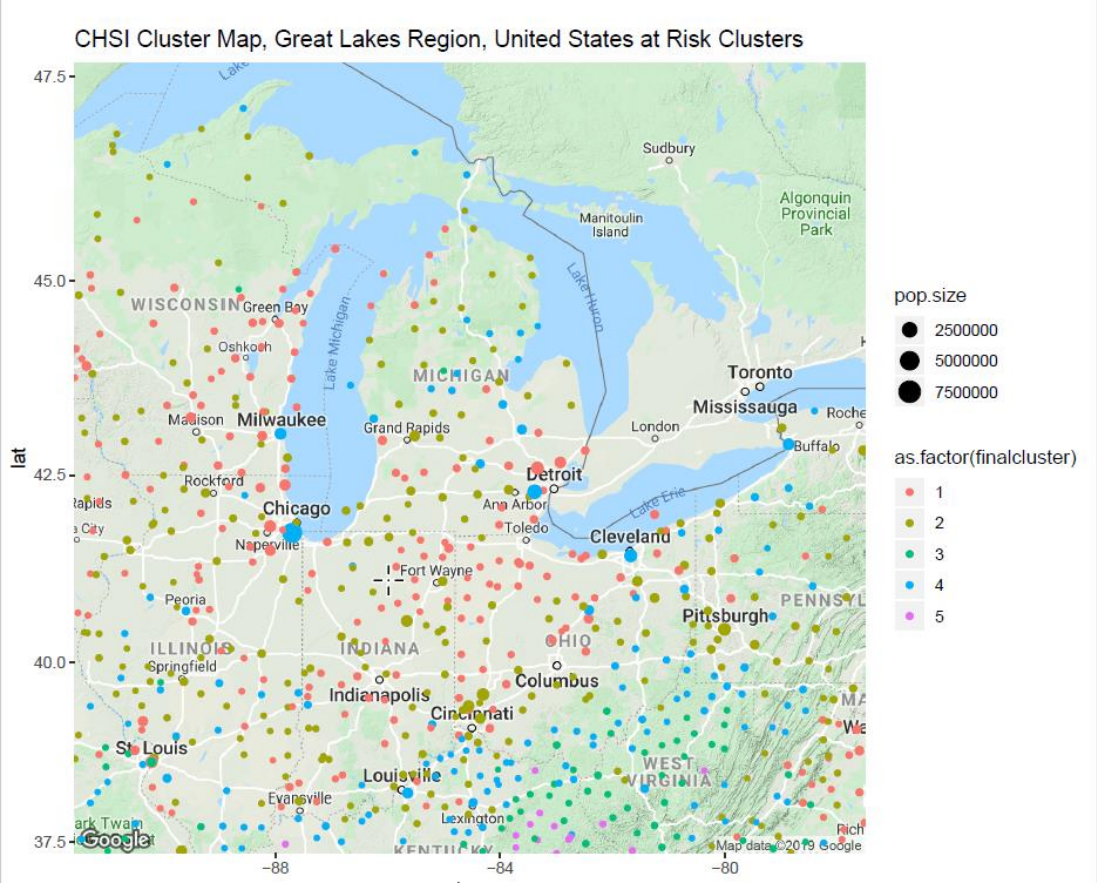
Given that the nature of our dataset is based upon a geographic plane, the research team found it expedient to demonstrate the 5 clusters of the data set on the map, as to better understand the distribution of these clusters via there geographic location, and to determine whether or not there was any explanation to that distribution in this unsupervised method. Utilizing the ggmap package; the following maps demonstrate the 5 clusters distribution; Whole, East Coast, West Coast, Great Lakes, South East Region, New York City Metro area.

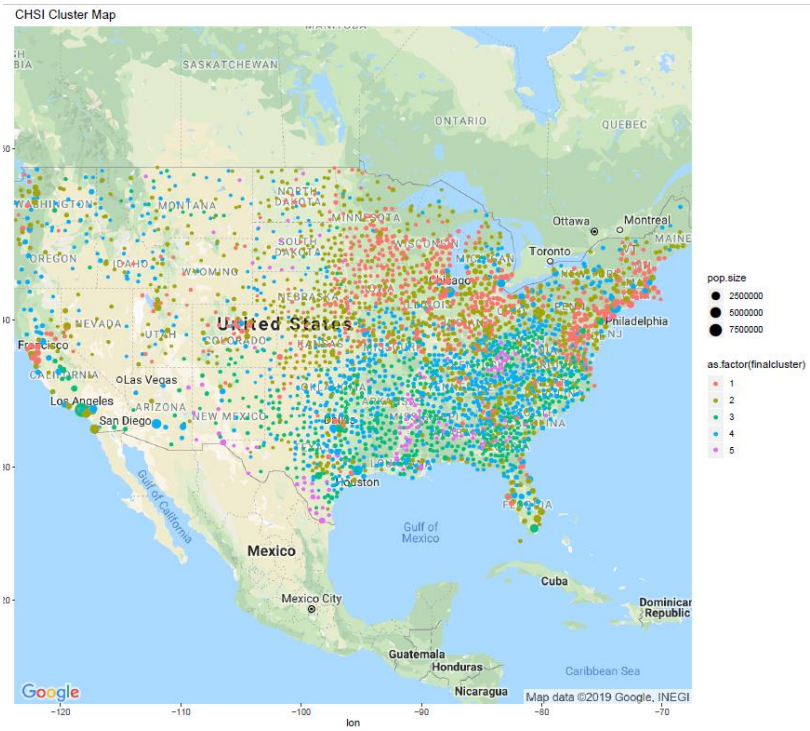



CHSI Cluster Map, East Coast Clusters
Ottawa
Montreal
Sherbrooke
MAINE
VERMONT
Kingston
Toronto
Mississauga
Rochester
Portland
NEW HAMPSHIRE
Buffalo
NEW YORK
Albany
MASSACHUSETTS
Providence
CONNECTICUT
RI
PENNSYLVANIA
New York
Pittsburgh
Philadelphia
MARYLAND
NEW JERSEY
Washington
DELAWARE
Richmond
Roanoke
VIRGINIA
Norfolk
Virginia Beach
Greensboro
Durham
NORTH CAROLINA
Map data ©2019 Google
lat
45.0
42.5
40.0
37.5
−80
−76
−72
−68
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5

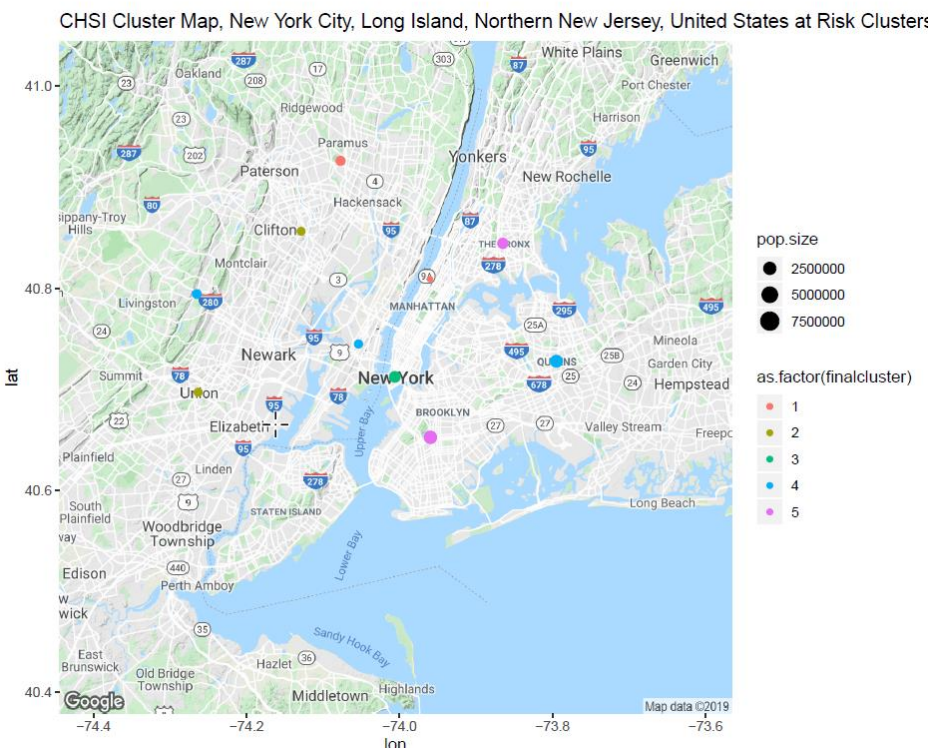

CHSI Cluster Map, Central United States at Risk Clusters
Rochester
Sioux Falls
Madison
Sioux City
Rockford
Cedar Rapids
IOWA
Des Moines
Iowa City
NEBRASKA
Omaha
Lincoln
Peoria
ILLINOIS
Springfield
ited States
Kansas City
Columbia
St. Louis
Lawrence
KANSAS
MISSOURI
Wichita
Mark Twain
National Forest
Springfield
Branson
Tulsa
Fayetteville
OKLAHOMA
Oklahoma City
Amarillo
Norman
Memphis
ARKANSAS
Lubbock
Google
Map data ©2019 Google
lat
42
39
36
−100
−96
−92
lon
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5


CHSI Cluster Map, West Coast at Risk Clusters
Redding
Reno
NEVADA
Sacramento
Humboldt-Toiyabe
National Forest
San Francisco
San Jose
Fresno
CALIFORNIA
Death Valley
National Park
Las Vegas
St. George
Bakersfield
Los Angeles
Anaheim
Long Beach
San Diego
Tijuana
Mexicali
BAJA
CALIFORNIA
Google
Map data ©2019 Google, INEGI
lat
40.0
37.5
35.0
32.5
30.0
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5



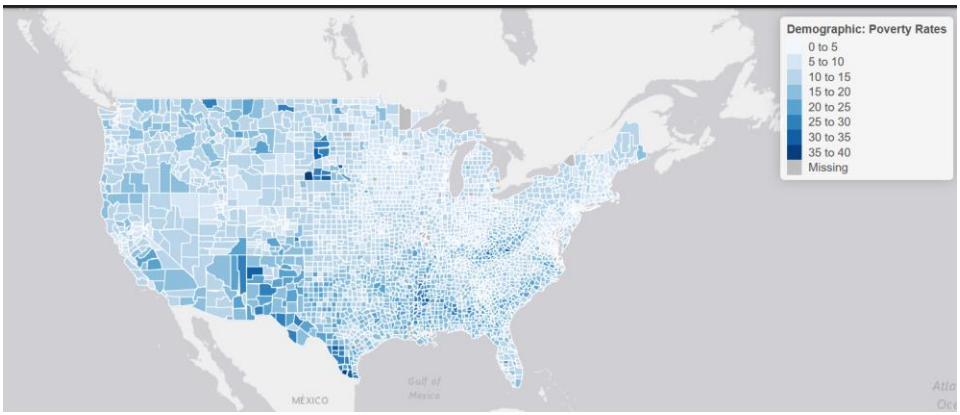


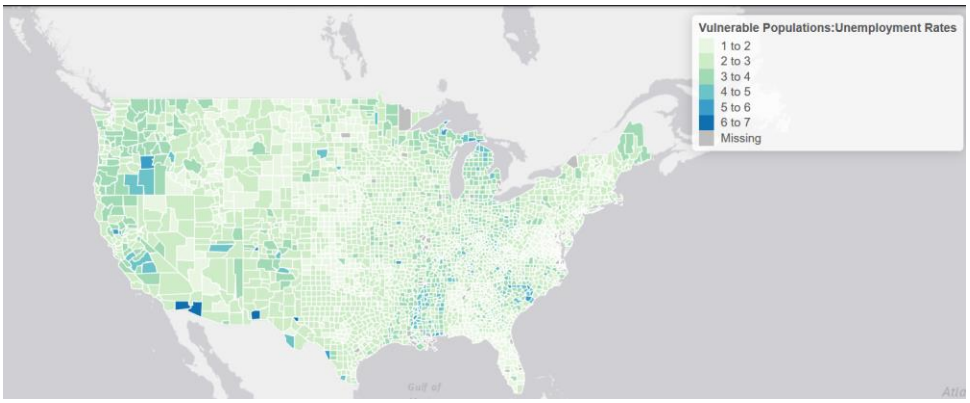
There is a patten that has been identified within the data. The at risk factors that contribute negatively to average life expectancies have cohesion of the data points within the clusters and separation between clusters.

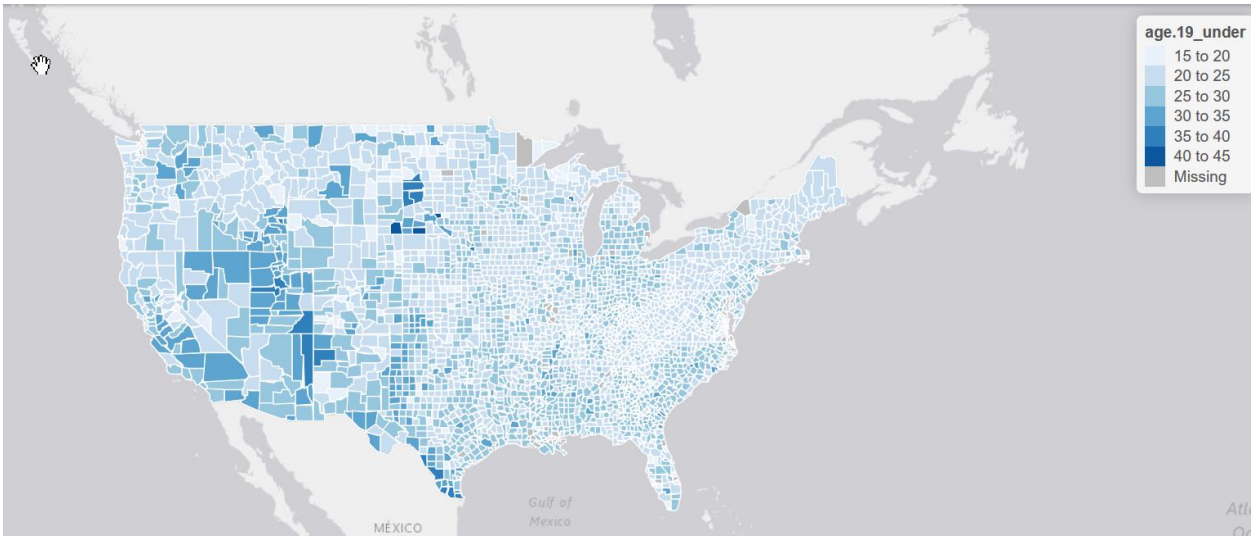


The nature of the clustering is not being determined by urban versus rural areas. In examination of the cluster's distribution around the nation, high urban areas like the west coast and east coast, the clusters contained within areas such as San Jose, San Francisco, and, Los Angeles – all belong to different cluster groups. Indeed, in examination of the New York City Metro area, this is expressed profoundly in the cluster differences found in Manhattan, compared to all the other boroughs and metropolitan surrounding areas, and indeed could be a benchmark as to the interpretation of the results. Look at the below example: Manhattan is listed in Cluster 3, whereas The Bronx and Brooklyn are listed in Cluster 5. Queens is listed within Cluster 4. Northern New Jersey communities of Paramus and Clifton are all listed in different clusters than New York City, with the highland New Jersey Region near Livingston, having a correlation in clustering with Queens.

Finally, given the nature of the clustering, and, that one of the at risk community factors to the diminishment of Average Life Expectancy is Poverty, the research team found some interesting similarities in the nature of the distribution of poverty within the nation, versus, the nature of the clustering algorithm as determined by the unsupervised k-means analysis upon the data, as shown by a couple of at risk and demographic maps:



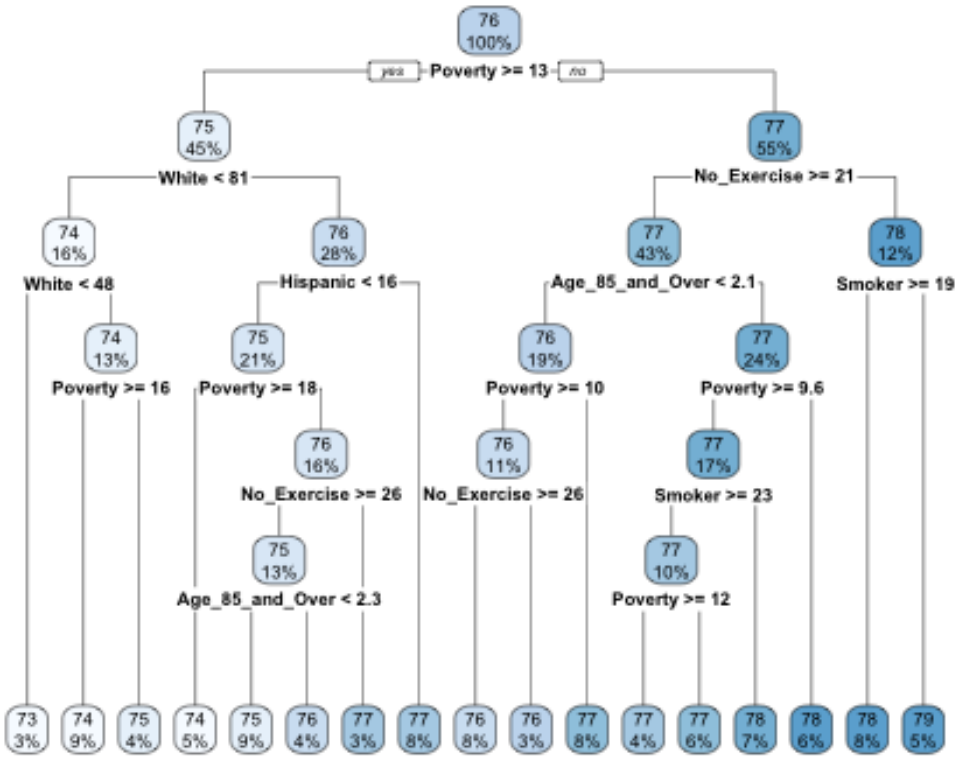




Decision Tree

A decision tree model was built using different rpart parameters like maxdepth and minsplit. A model with maxdepth of 7,minsplit of 200 yields the best result with least error of 8.2% and an accuracy of 91.8%

```
#install.packages('rpart')
#install.packages('rpart.plot')
library(rpart)
library(rpart.plot)
set.seed(50)
dt <- rpart(ALE~.,trainset1,method  = "anova", control = rpart.control(cp = 0,
maxdepth = 7,minsplit = 200))
rpart.plot(dt)

pred_dt <- predict(dt, testset1[,-38])
library(caret)
conf.dt <- data.frame(pred,testset1$ALE)
sqrt((mean(pred_dt-testset1$ALE)^2))

## [1] 0.08064625
```



b. 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.

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

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.

Conclusions

1. The data set we chose really does fit it is 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.

IST 718: Big Data Professor Jon Fox

Big Data, the final course chosen by the candidate was facilitated through course professor Jon Fox, PhD. Through the utilization of python scripting techniques, and various data mining techniques, machine learning technique and geospatial analysis, the candidate is in the concluding stages a final project. The portion assigned to the candidate will be shared in this paper. The final project deliverable for the course is as follows:

> This assignment provides an opportunity to demonstrate your ability to work on a project before the final week of the semester. Overall, the course project allows the student an opportunity to demonstrate progress (or mastery) of learning objectives 1, 2, 3, 4, 5, and 6:
>
> * Obtain data and explain data structures and data elements.
> * Scrub data by applying scripting methods, to include debugging, for data manipulation in Python, R, or other languages.
> * Explore data by analyzing using qualitative techniques including descriptive statistics, summarization, and visualizations.
> * Model relationships between data using the appropriate analytical methodologies matched to the information and the needs of clients and users.
> * Interpret the data, model, analysis, and findings, and communicate the results in a meaningful way.
> * Select an applicable analytical methodology for real problems in areas such as business, science, and engineering
>
> (Taylor, “IST 718,” 2020).

a. 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?

How has Covid-19 impacted us socially?

The global pandemic affects us all and identifying the answers to these questions will allow us to better respond, acclimate, and curb the negative effects of this ongoing crisis.

Data Sources

Data from a variety of sources will be utilized to answer these questions.

Twitter API / COVID-19 Twitter chatter dataset for open scientific research

New York Times Covid-19 Data Set

CDC Community Health in Action Data Set

Johns Hopkins University (JHU) Covid-19 Data Set

The COVID Tracking Project Data Set

The Covid-19 Open Research Data Set

Basic Dataset Metrics

CDC Covid-19 Death Counts in the United States by County

Deaths with confirmed or presumed COVID-19, coded to ICD–10 code U07.1. Counties included in this table have 10 or more COVID-19 deaths at the time of analysis. Number of deaths reported in this table are the total number of deaths received and coded as of the date of analysis and do not represent all deaths that occurred in that period. Data during this period are incomplete because of the lag in time between when the death occurred and when the death certificate is completed, submitted to NCHS and processed for reporting purposes.

769 rows x 8 columns

CDC 2010 Community Health Status Indicators (CHSI)

Today, CDC released the updated Community Health Status Indicators (CHSI) online tool that produces public health profiles for all 3,143 counties in the United States. Each profile includes key indicators of health outcomes, which describes the population health status of a county and factors that have the potential to influence health outcomes, such as health care access and quality, health behaviors, social factors, and the physical environment.

3141 rows x 55 columns

COVID-19 Twitter chatter dataset for open scientific research

100M+ cleaned tweets regularly updated through 7/26/2020.

The Covid-19 Open Research Data Set

195K scholarly articles on the topic of Covid-19

Team Roles

The project data questions will be divided amongst the team members.

Thomas Bahng - *How has Covid-19 impacted us socially and the Academic research to address the virus?*

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?***

Patty Mills - *What has been published about ethical and social science considerations regarding Covid-19?*

Jose 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?*

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?

Project Description (Randall Taylor)

Task Lists

- [x] ETL
- [x] EDA
- [x] kmeans
- [ ] linear modeling
- [x] interactive plotting

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 )
```

```
#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/
```

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.

First dataset brought into the python 3 environment (Colab); sourced from GitHub, this data set simply contains the ‘fips’, ‘state abbr’, ‘county name’, ‘lat’, and ‘long’ into python to map each of the 3,142 counties to a place on a map.

The second dataset brought into the python 3 environment (Colab); sourced from the New York Times via a live source, these number update many times during the day. The columns within this dataset are:

```
date                datetime64[ns]
county                      object
state                       object
fips                        object
cases                        int64
deaths                     float64
confirmed_cases            float64
confirmed_deaths           float64
probable_cases             float64
probable_deaths            float64
```

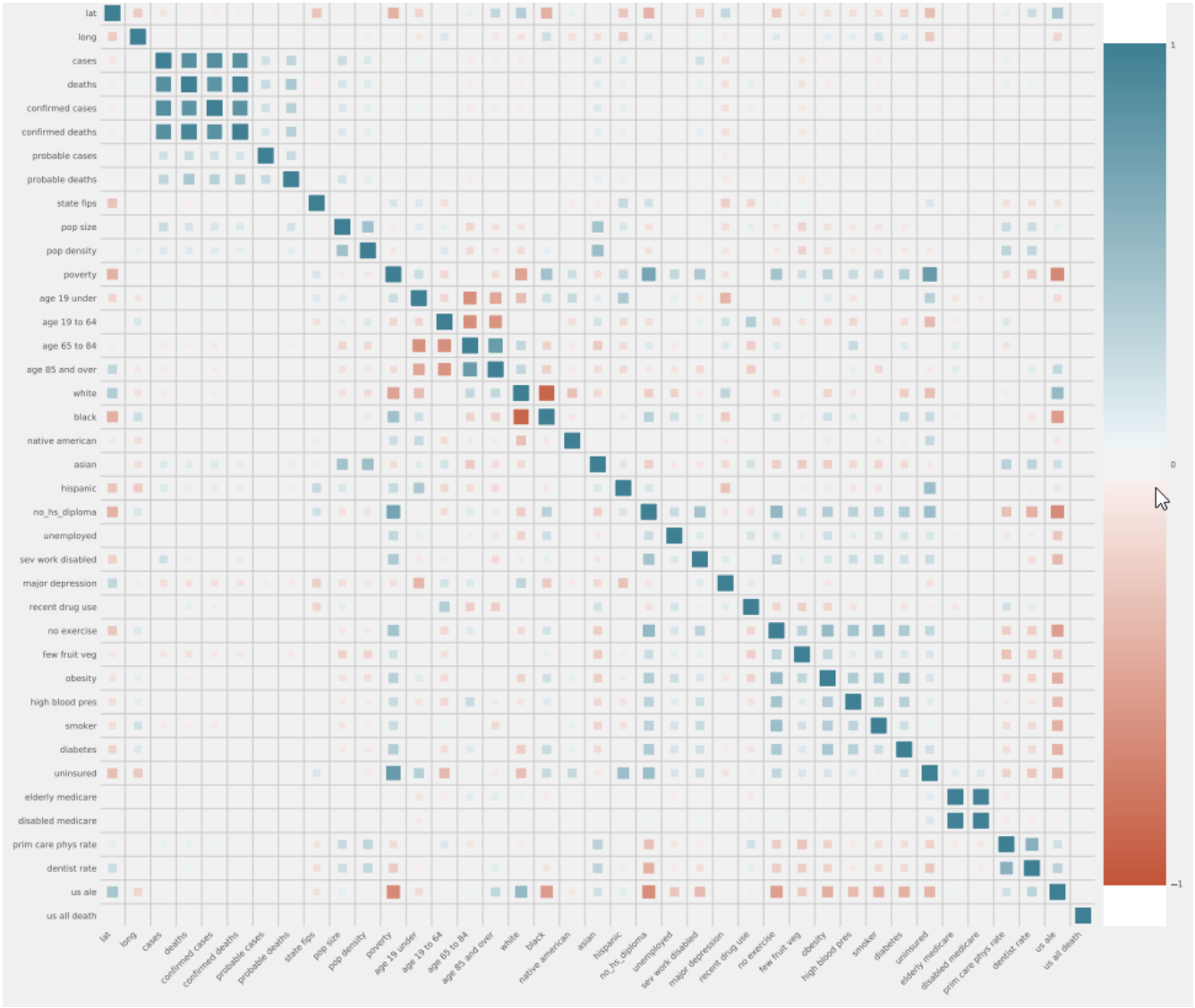
Extraction completed, transformation of the data required some applications of median values, to fill the ‘na’ entries. Median was chosen in an attempt to keep outliers from being considered in the data.

The third dataset brought into the python 3 environment (Colab); sourced from the CDC’s Community Health Status Indicators (CHSI) extensive library, of community at risk factors. The various csv were brought together via dfsql packages in R, to create an initial combined dataset, and then, the dataset was brought into the python 3 environment as a csv with 3141 rows (counties) and 35 columns of at risk factors and demographics:

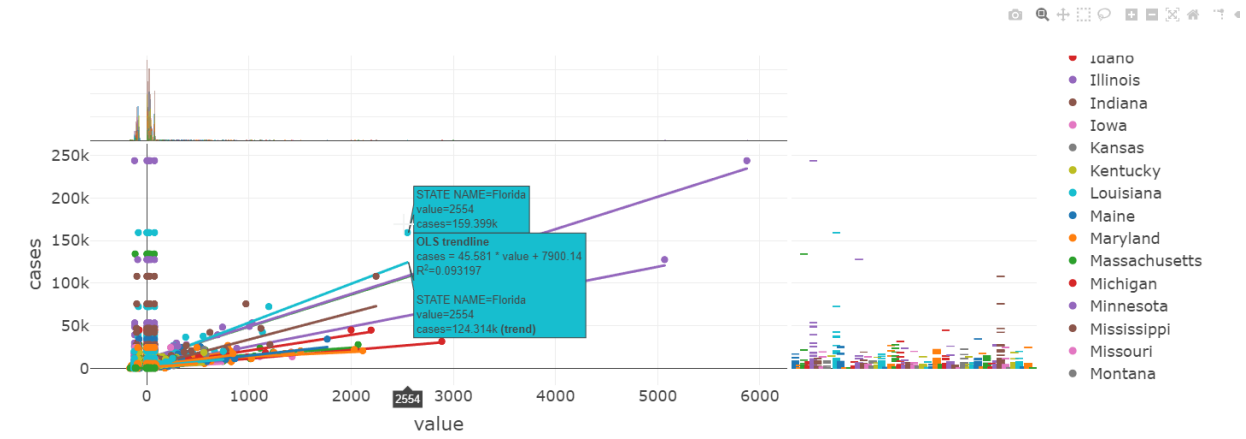
	County_FIPS_Code	State_FIPS_Code	CHSI_County_Name	CHSI_State_Name	CHSI_State_Abbr	Population_Size	Population_Density	Poverty	Age_19_Under	Age_19_64	Age_65_84	Age_85_and_Over	White	Black	Native_American	Asian
0	1	1	Autauga	Alabama	AL	48612	82.0	10.4	26.9	62.3	9.8	0.9	80.7	17.3	0.5	0.6
1	3	1	Baldwin	Alabama	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

3141 rows × 35 columns

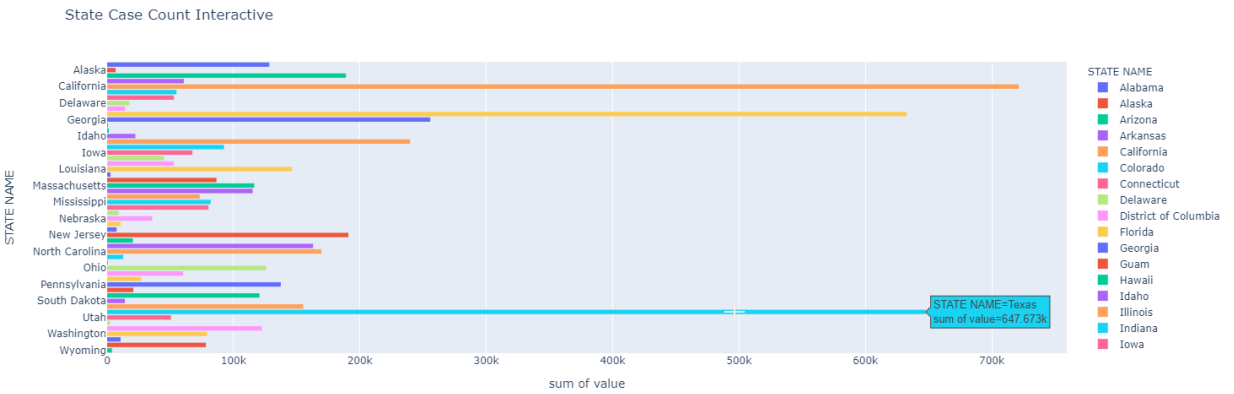
All three datasets were combined into one dataset to enable visualization of the exploratory process of the dataset, and to further assist in modeling. The following correlation plot was created, for all potential features within the dataset:



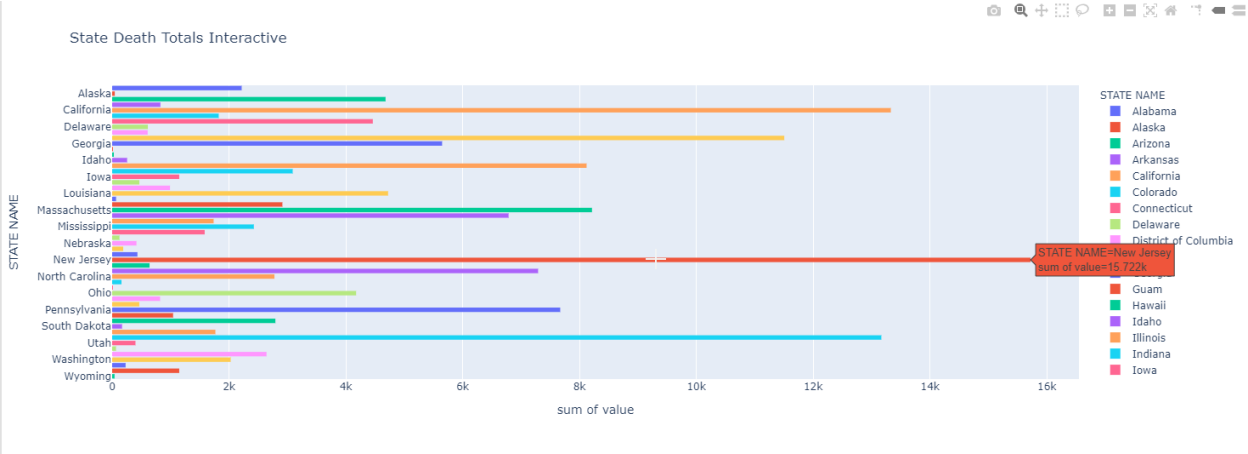
Initial visualizations of the linearity between the chosen features, and the target variable, cases, where create to begin to direct the candidate as to how to approach modeling, and what features to select in that model:



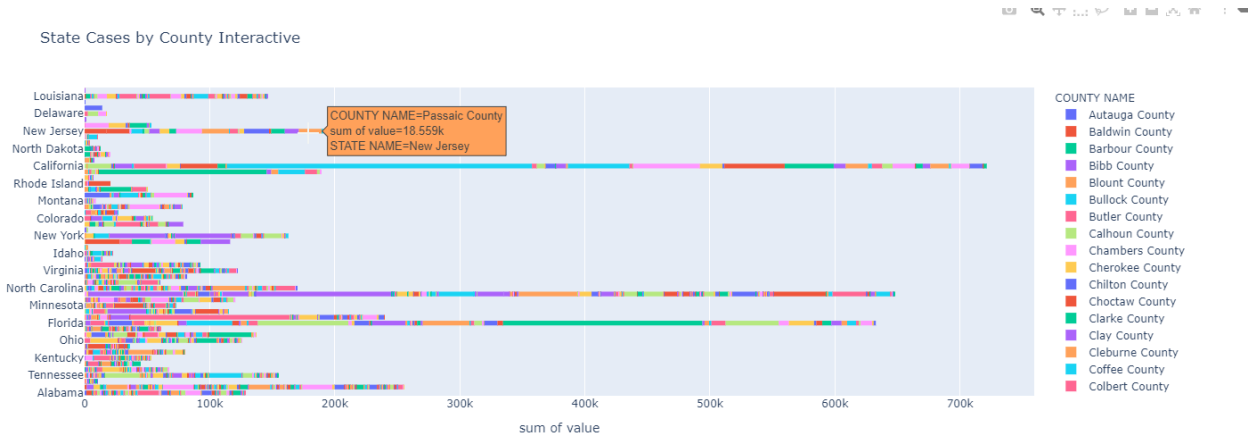
Further exploratory data analysis of the dataset allowed for visualization of the case counts, per state, in an interactive manner:



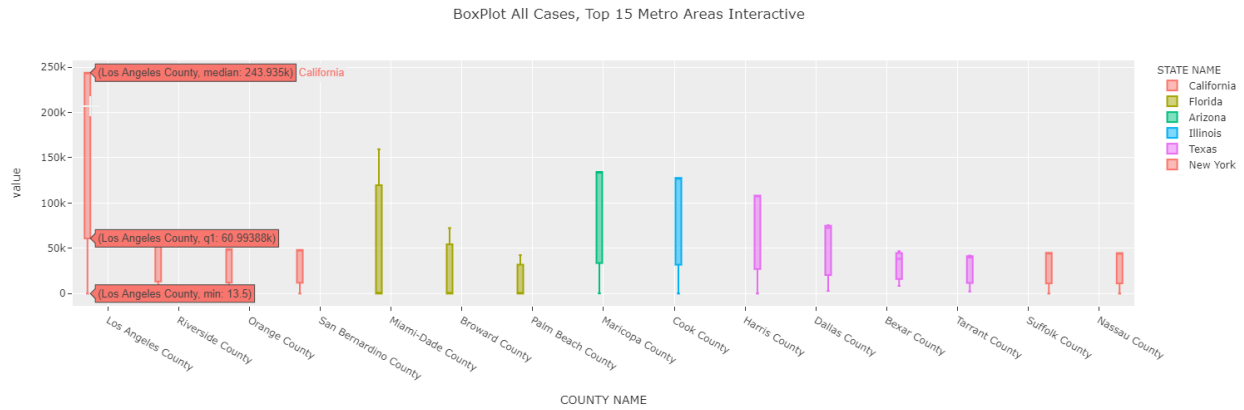
These are the totals of deaths related from Covid-19, per state:

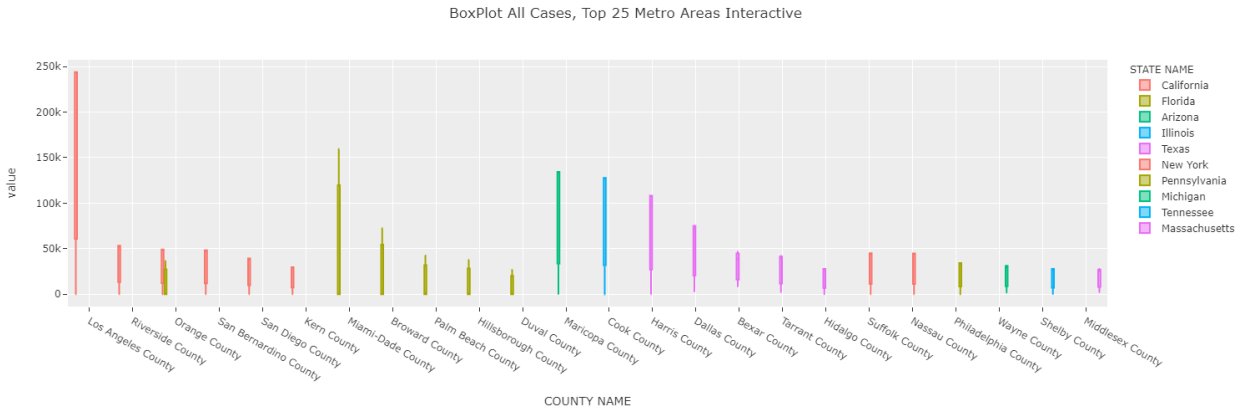


The following visualization demonstrates the totals of COVID-19 related cases count, per county by state:

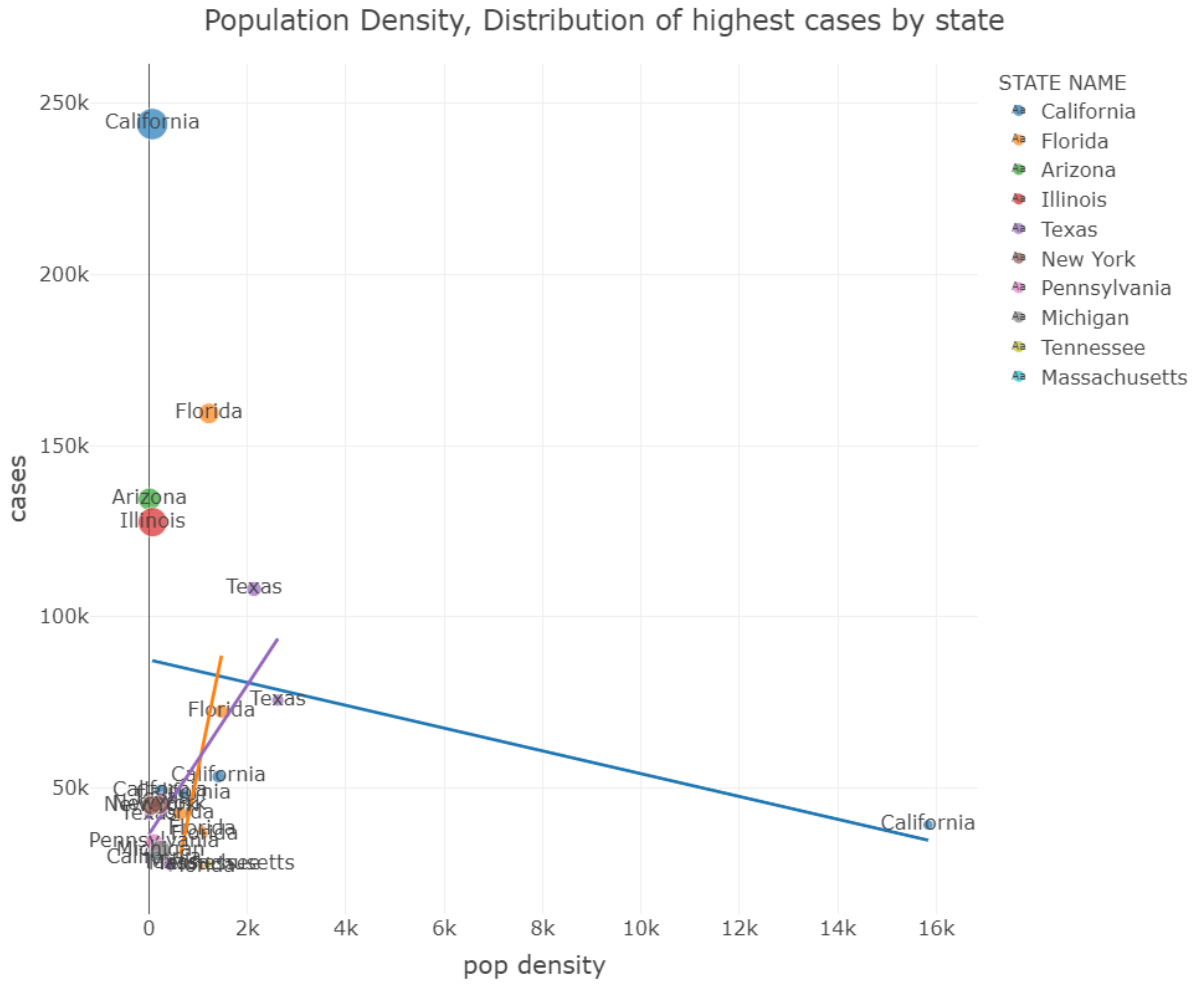


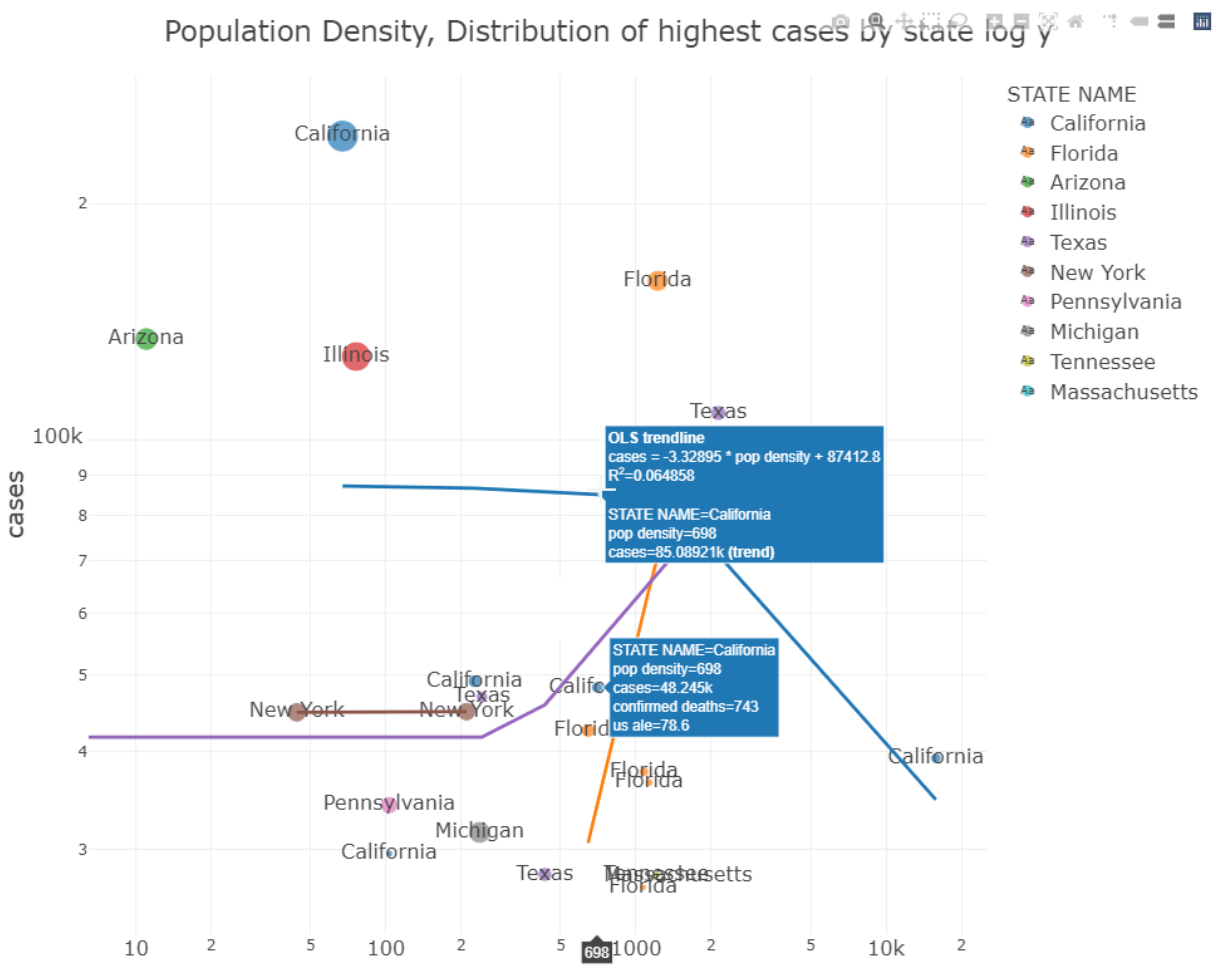
To better understand the distribution of COVID-19 around the nation, an interactive box plot of the top fifteen metro areas was created, followed by a top twenty five metro areas interactive boxplot, as shown:



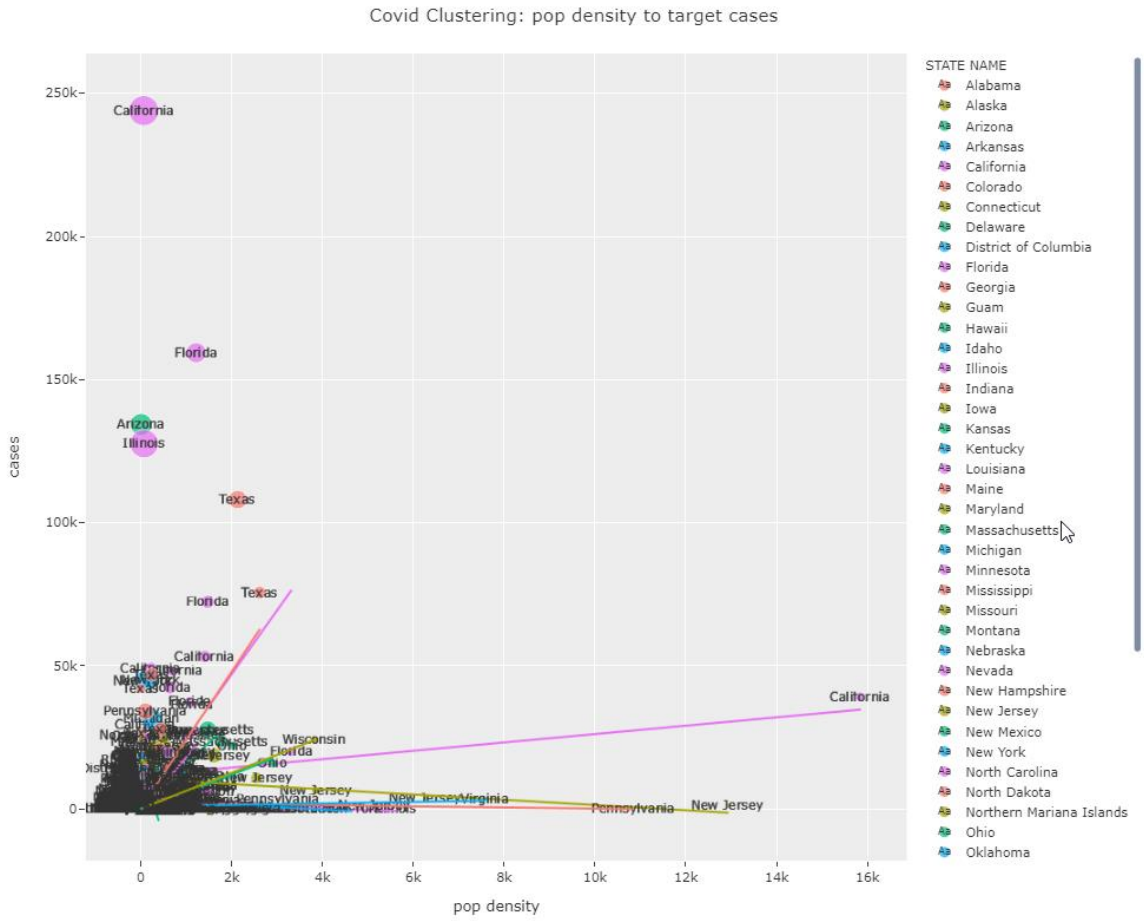


Further investigation lead to the follow interactive, that demonstrates that there is indeed linearity between our target variable, cases, and one of the targeted key dependent variables ‘features,’ population density:

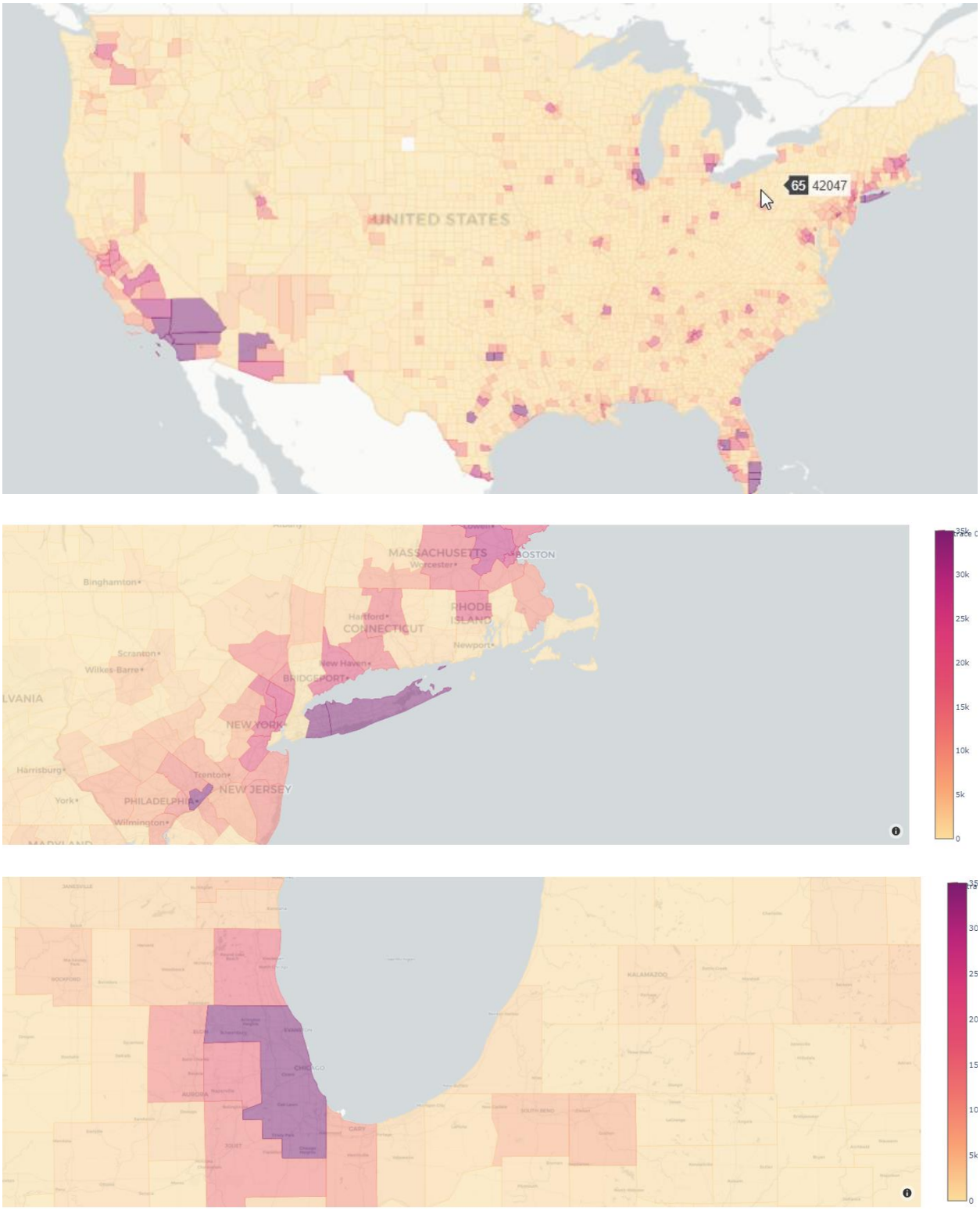

Population Density, Distribution of highest cases by state
STATE NAME
California
Florida
Arizona
Illinois
Texas
New York
Pennsylvania
Michigan
Tennessee
Massachusetts
250k
200k
150k
100k
50k
cases
0
2k
4k
6k
8k
10k
12k
14k
16k
pop density



Taking all of the respective states into account, and demonstrating the data therein, in the same fashion as above, it can be plainly seen that there is correlative linearity between the feature of population density and the target variable, cases:

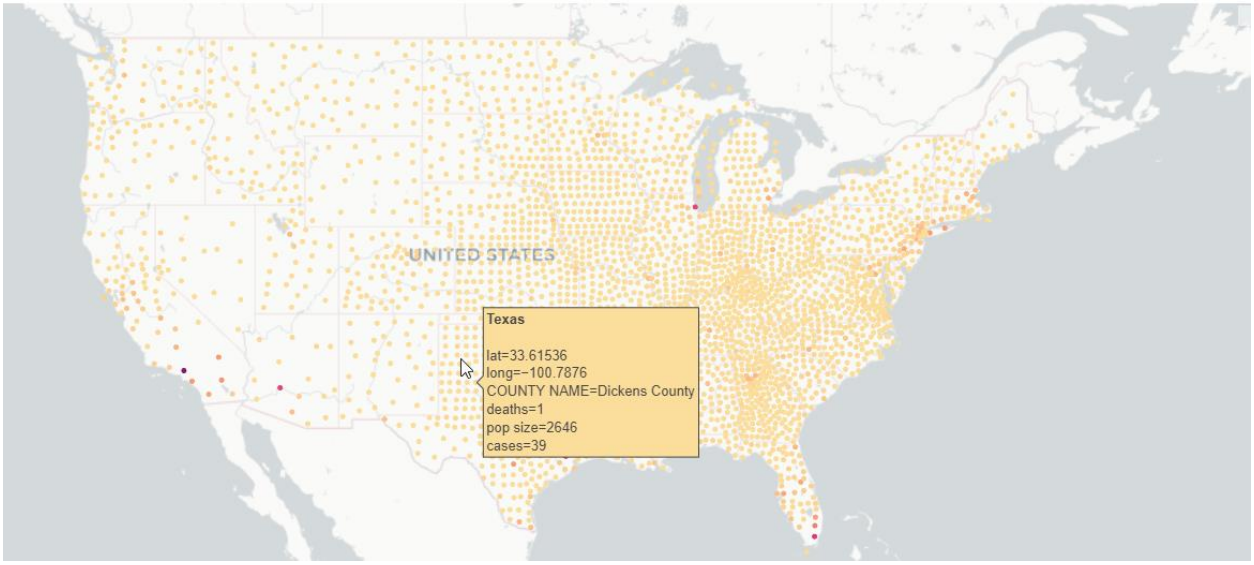


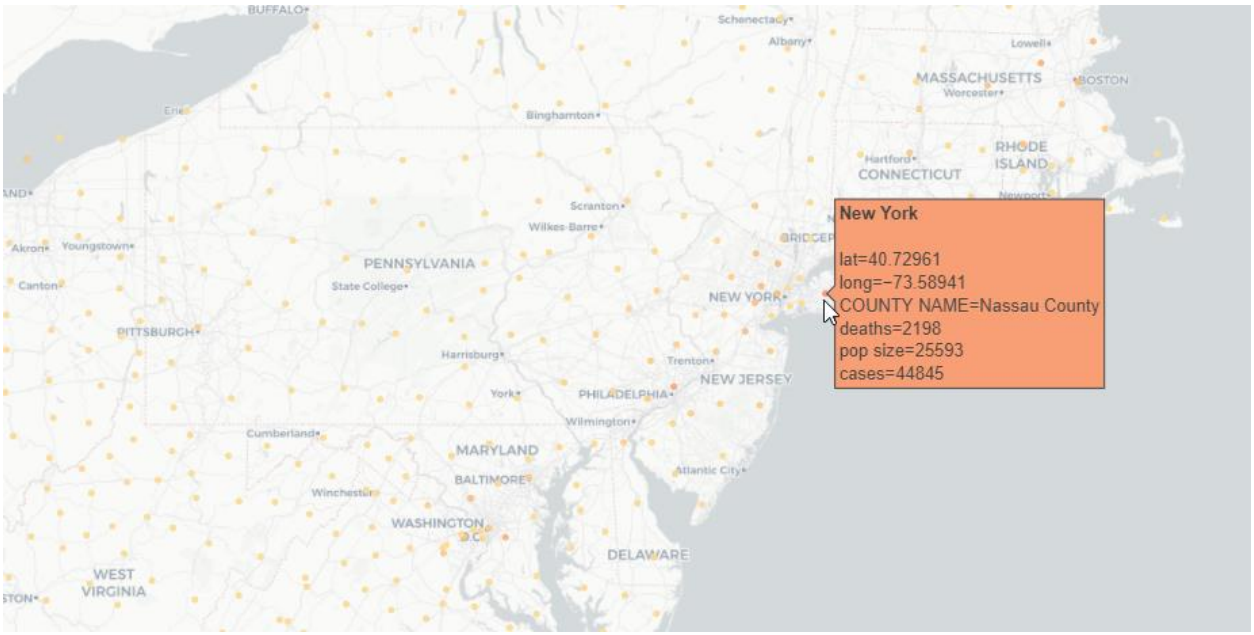
To further the exploratory data analysis, the candidate began to geospatially plot the data as to help researchers understand the implications of the findings through pre-modeling exercises to determine correlation and linearity between the variables, a choropleth interactive heat map was created to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



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:



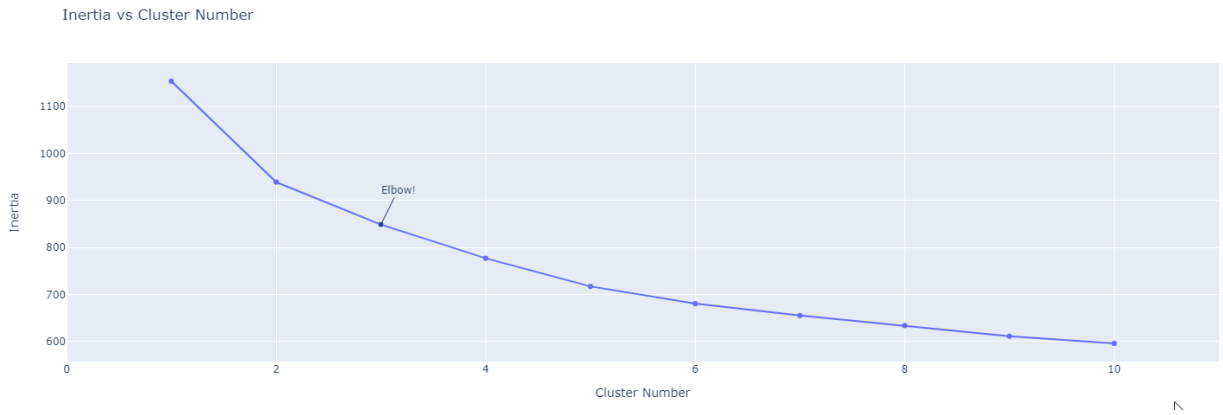


Models

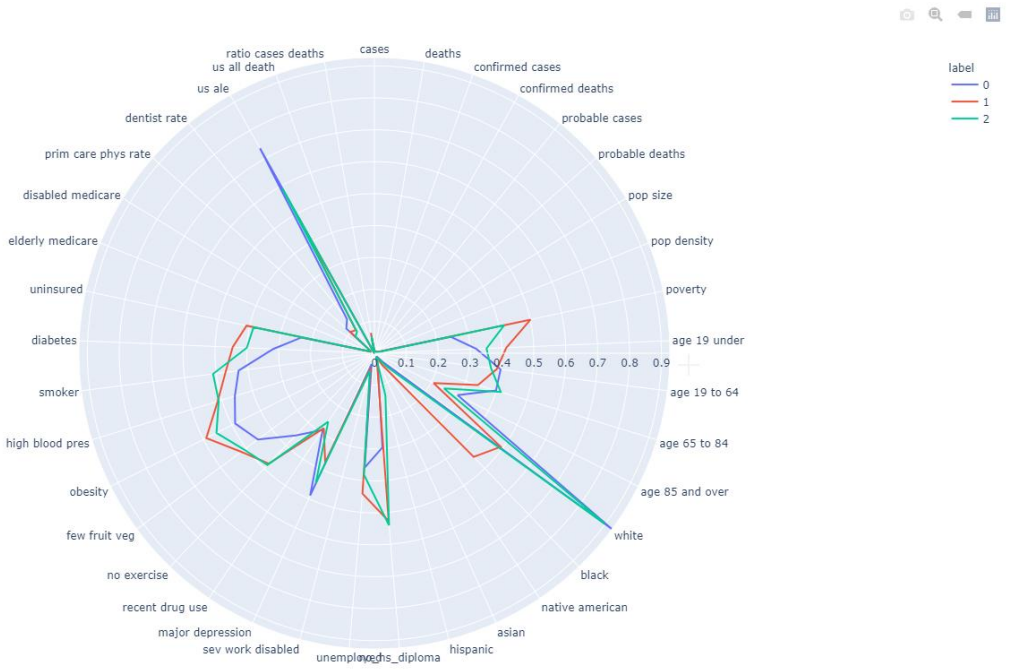
Approaching the deadline for the submission of the portfolio will not allow for the SUPERVISED MACHINE LEARNING outcomes to be published within the paper, however, the candidate can share the UNSUPERVISED MACHINE LEARNING outcomes, via Kmeans Analysis.

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.

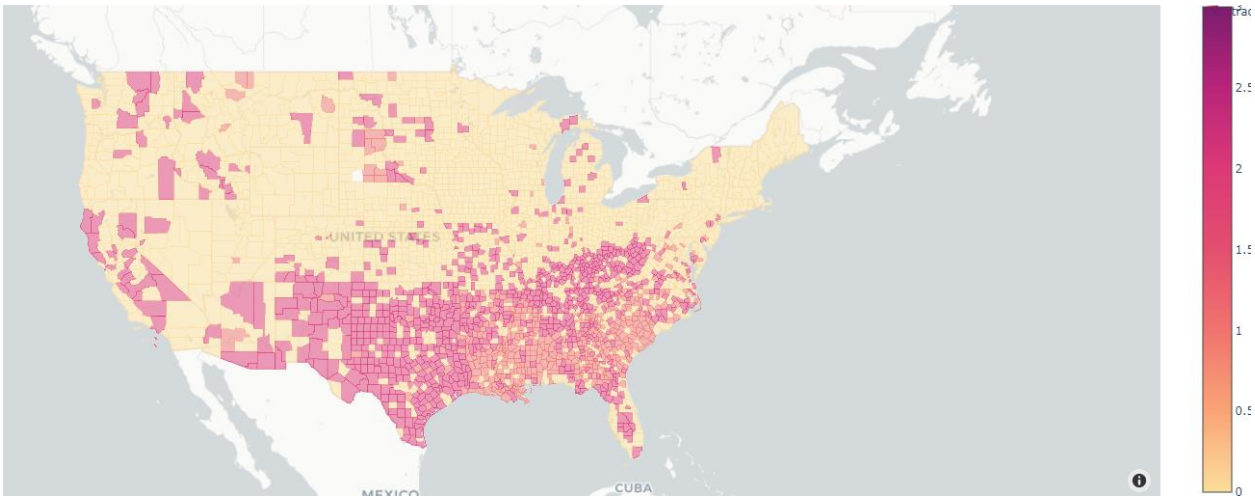


Line_polar is perfectly suited to the task of Kmeans clustering, because a circle can handle as many variables as you want, and the lines inside are pretty intuitive. To keep it simple, we can just understand our variables by the mean of their characteristics, their relationships to one another, and knowing that every variable is still scaled.



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:

Three distinct clusters



b. 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.

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.

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.

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.

References:

Le, James. (2019). An Introduction to Big Data: Data Normalization *Cracking The Data Science Interview,* Toward Data Science https://medium.com/cracking-the-data-science-interview/an-introduction-to-big-data-data-normalization-b72311f134b7

Taylor, R.S.(2020) IST 652: MSADS Portfolio Scripting for Data Analysis https://github.com/randallscott25/MSADS_Portfolio/tree/master/IST652_ScriptingForDataAnalysis

Taylor, R.S.(2020) IST 659: MSADS Portfolio Database Administration https://github.com/randallscott25/MSADS_Portfolio/tree/master/IST659_DatabaseAdminstration

Taylor, R.S.(2020) IST 707: MSADS Portfolio Data Analytics https://github.com/randallscott25/MSADS_Portfolio/tree/master/IST707_DataAnalytics

Taylor, R.S.(2020) IST 718: MSADS Portfolio Big Data https://github.com/randallscott25/MSADS_Portfolio/tree/master/IST718_BigData