Healthcare Accessibility and Risk Analysis for Stroke Mortality in Missouri

Alhassan Sahad[1], Ajay Karera[2], Sai Abhishek Talabattula[2], Conlin Steinert[2], Dr. Timothy L. Haithcoat[3], Dr. Kate Trout[3]

[1]Geospatial Specialization

[2]High Performance Computing Specialization

[3]University of Missouri Faculty

Abstract

Stroke is one of the leading causes of death in the United States as a whole, and there are many factors that can either raise or lower the risk of mortality from it. One of these, having been proven, is accessibility to healthcare facilities that can treat stroke. Outside of large-scale studies, there haven't been any that have tried to quantify the exact effect that lack of accessibility has on stroke mortality. This lack of quantification is likely due to the necessary qualifications of those researching, needing to have some background in both geospatial and healthcare data, or at least have team members having those qualifications. It is also notoriously difficult to quantify an issue like this; due to the overwhelming amounts of factors that factor into an issue like stroke mortality, separating out the impact that one factor has on the issue requires an immense amount of work.

Using data from the Centers for Disease Control (CDC) and other agencies, such as County Health Rankings and Road-network maps and the Joint Commission, we were able to create a model that can access the impact that accessibility has on stroke mortality for the state of Missouri. We also created various indices that will be useful for continued research into this subject.

Audience, Project Definition, Background, and Aims

Our goal for this capstone is to assess the impact healthcare accessibility to stroke care centers has on stroke mortality in Missouri.

BUSINESS LOGIC AND TARGET AUDIENCE:

- Domain Problems and Questions:
    - Resource allocation for Healthcare Services (Understanding Healthcare Service Coverage and Gaps in Service)
    - Identifying underlying factors affecting Healthcare Accessibility
    - Identifying populations at risk (Lack of access and susceptibility to health conditions due to societal behaviors and comorbidities)
    - Determining the state of healthcare accessibility across the state of Missouri
- Target Audience:

- Healthcare Services
- Policy Makers
- General Public

The aim of this study is to:

1) Determine health care accessibility across the state of Missouri using stroke as a primary metric.

2) Produce a model for classifying/predicting stroke mortality across Missouri.

3) Make recommendations on areas within Missouri that require urgent health care facilities or policy interventions.

By the completion of this capstone, we seek to accomplish the following.

•Improved resource allocation for healthcare facilities.
•Enhanced decision-making for Policy makers.
•Reduced healthcare provision disparities.
•Improved healthcare accessibility improves public health, further enhancing the functioning of both communities and the state.

When it comes to healthcare, any additional information on best practices helps save lives. This can be anything from how best to treat a patient with a myriad of symptoms and disease, to the best way to transport a patient between different hospitals. In our capstone, our attempt at helping patients is to determine issues concerning accessibility for stroke patients, especially concerning distance from hospitals and any demographic disparities that we can find. To do so, we must gather data and process it through models so we can reach a conclusion concerning the best direction to take with stroke care facilities.

Our capstone is aimed at state policymakers and hospital administrators to improve public health and resource allocation. A hope for our capstone is that it will be used to determine the best locations for opening new healthcare facilities to improve healthcare accessibility throughout the state.

Literature Review

The landscape of healthcare accessibility in Missouri presents a complex challenge influenced by socio-economic, geographical, and systemic factors. This literature review synthesizes findings from various studies to explore these challenges, with a particular focus on stroke care, revealing how insurance coverage, geographic location, socio-economic status, and healthcare system capabilities play critical roles in healthcare access.

Studies have shown that in Missouri, healthcare accessibility issues are not only about the lack of insurance but also about the myriads of obstacles insured individuals face, including high out-of-pocket costs and limited-service availability. For example, families often struggle with co-pays and deductibles, making it difficult to afford necessary healthcare services despite having insurance. Geographic challenges are particularly pronounced in rural areas where healthcare facilities are sparse, exacerbating the difficulties for insured individuals who might lack reliable transportation or the means to travel long distances for healthcare.

DeVoe et al. (2007) discuss the intricate relationship between insurance coverage and actual healthcare access, emphasizing the need for a multi-pronged policy approach that expands insurance coverage, improves affordability, and enhances the geographic distribution of healthcare services. Their findings suggest that increasing funding for healthcare facilities in underserved areas and investing in transportation infrastructure could significantly improve access.

Similarly, Hammond et al. (2020) highlight the disparities in stroke care between urban and rural areas in the United States, showing that patients in rural Missouri are less likely to receive advanced stroke treatments and have higher in-hospital mortality rates. Addressing these disparities requires enhancing healthcare infrastructure in rural areas, possibly through incentives for healthcare professionals and investments in telemedicine and emergency medical transportation.

Menon et al. (1998) focus on the demographic influences on stroke care accessibility, noting significant delays in hospital arrival times for women and Hispanic Americans. This indicates a need for targeted public health campaigns and education programs in Missouri to raise awareness about stroke symptoms and ensure language accessibility in healthcare communications.

The use of Geographic Information Systems (GIS) to analyze accessibility to emergency care, as discussed by Pedigo and Odoi (2010), is particularly relevant for Missouri's varied geography.

This methodology could identify areas with limited access to emergency stroke and myocardial infarction care, guiding strategic placement of new healthcare facilities to maximize accessibility.

Freising et al. (2018) explore the critical period from the emergency call to hospital admission using GIS to model travel times, which is crucial in rural Missouri where longer travel times can delay critical care for stroke patients. Strategies to improve accessibility might include enhancing emergency medical services and establishing more stroke care centers, especially in underserved areas.

El Khoury et al. (2012) provide a framework for understanding the complexities of stroke care accessibility, pointing towards potential areas of improvement across prehospital, hospital, legislative, and economic domains. For instance, raising public awareness and establishing specialized stroke units could significantly enhance stroke care services in Missouri.

The experience of stroke survivors and caregivers, as explored by Pindus et al. (2018), underscores the need for greater integration of care and comprehensive support for stroke survivors, suggesting strategies such as expanding in-home care options and supporting caregivers through education and respite care.

Finally, Zachrison et al. (2019) evaluate the accessibility of acute stroke care and tele stroke services, highlighting the potential of telemedicine to extend the reach of stroke specialists to remote or rural areas, ensuring timely and expert consultation.

Enhancing healthcare infrastructure, leveraging technology like telemedicine, and ensuring equitable resource allocation across different regions and demographic groups. By addressing these multi-dimensional challenges, Missouri can move towards more equitable and effective healthcare access for all its residents.

While previous studies have looked at the accessibility of stroke care within the State of Missouri, there is still a gap between conducting the study at the county level and using fine population data at the block level. Therefore, this study seeks to bridge that gap.

We needed ways to determine what factors we consider for stroke morbidities, so we started looking at research papers to determine what the wider consensus was. One of these papers used was “Stroke – Incidence, Mortality, Morbidity, and Risk” by Dr. Timothy Ingall (2004). This paper includes multiple parts that were outside the scope of our project, including statistics on reoccurrence and the effects of different drugs on stroke recovery and treatment. What we were primarily concerned about was the morbidities brought up in the paper. One of the primary

predictors for stroke was a history of diabetes, which closely aligns with data that the CDC Interactive Heart Atlas noted, which improved our confidence about the types of data that the Atlas collected which we were able to use in our program.

“Incidence, Short-term Outcome, and Spatial Distribution of Stroke Patients in Ludhiana, India” by Jeyaraj Pandian et al. (2016), was a very similar survey into stroke incidence and results in the city of Ludhiana, India, using heatmaps and hospital statistics to discover similarities between stroke cases that turned fatal and those that did not. The creation of heatmaps and concerning themselves with hospital locations are very similar to what we are doing for our project, but there are key differences between our points. The paper generally looks at overall effects that ordinary hospitals have on stroke mortality, while our data generally is focused on Stroke Care Centers, which have certification for advanced stroke care facilities. The paper noted that there was increased stroke mortality in locations that were primarily served by public hospitals in poor areas, which are places that logically make sense, as resources are strained generally in those locations. With this, we are more confident in looking at Stroke Care Centers as the hotspots for our network analysis, as there is an obvious difference between locations that can take care of stroke and those that do not.

“Spatial Analysis of Service Areas for Stroke Centers in a City with High Traffic Congestion,” by Ivan Pradilla et al. (2020), investigates the effect that different times have on accessibility when it comes to locations that have high congestion. The main takeaway from this study was that high levels of congestion in areas lead to smaller than expected catchment areas for each individual hospital, with the idea that a patient has to arrive within an hour of stroke manifesting, causing strains on the resources of each hospital within the area. Though much more compact in scope, and outside the reach that we will be using for our project, as we will not be looking into congestion as a root cause of healthcare inaccessibility, it is still important to note how they looked into their geospatial data and how we can learn for our own project, as we both want to create catchment areas, which we are calling service areas, and how better we can make our programs for such a thing.

“Younger Age of Stroke in Low-Middle Income Countries is Related to Healthcare Access and Quality” by Mohammad Rabar et al. (2021), looks into the effect that healthcare accessibility has on stroke mortality, especially when considering younger ages. The most interesting point brought by this study was how lower healthcare accessibility and quality, which they used the HAQ index

for, accounted for almost all the difference in stroke mortality between the countries, as once that was brought out, the rate of stroke mortality between high income and low to medium income countries by age had little significant difference. This is important for our study as it shows that healthcare accessibility and quality of care are the most significant differences for stroke mortality rate, something we wish to showcase on a smaller scale than a comparison between countries.

“Healthcare Resource Availability, Quality of Care, and Acute Ischemic Stroke Outcomes,” by Dr. Emily O’Brien et al. (2017), takes a closer look at the impact that healthcare quality has on stroke outcomes, looking more specifically at how the resource levels of the hospital where the care was provided affects the quality of care and post-treatment outcomes. What was found is that, if a hospital follows stroke guidelines set up by the American Heart Association, even though there is significant regional variability in resource allocation, the eventual result will be similar throughout all hospitals. This is important for our study as it shows that looking at the quality of resources is not important to look into, as long as the hospital follows strict guidelines for stroke care.

Amongst the Geospatial analysis techniques discussed and used, Geographically Weighted Regression (GWR) proved to be the most effective means to model healthcare accessibility at county level. We also realized that making our study more refined and localized would help us in comparing the urban and rural areas better. With such localized studies, we can theorize behavioral patterns, living conditions, geographical influence, etc. so that we can establish trends and patterns for urban and rural areas.

Factors affecting stroke incidence and post-care behaviors were also investigated in these studies to get an overall picture of stroke treatment cases, which further emphasized the importance of psychological, social, and functional needs.

With this literature review, we identified and understood all the principal factors essential to accurately model healthcare accessibility across the state and which additional considerations were to be acknowledged.

Research Questions

Healthcare accessibility is a very broad topic, and as such, to study it, we must pare it down to a reasonable size. For our capstone, the primary metrics that we have decided to focus on are

distance and time-based accessibility, with the understanding that other metrics, such as demographic differences, are important to consider when looking at stroke mortality rates.

Our primary question concerns itself with the spatial and temporal aspects of stroke care accessibility, with the spatial aspect being in regards with commute distance, spatial behavioral patterns, and topology of the areas of interest to the stroke care facility, and the temporal aspect is about seeing how the opening and closure of certified stroke care facilities affects the stroke mortality rate of the surrounding areas and how the healthcare accessibility in general changed over time. To do this, we require data on stroke mortality rate and certified hospitals over a series of years to study these effects.

Secondarily, we look at how different demographics affect this measure. The socioeconomic and demographic impact on healthcare accessibility is vaguely known and needs further scrutiny in our study to determine the interplay of these factors to better represent healthcare accessibility model.

Data Sources and Data Provenance

Our primary data source for our project is the CDC dataset called *Rates and Trends in Heart Disease and Stroke Mortality Among US Adults (35+) by County, Age Group, Race/Ethnicity, and Sex – 2000-2019.* This dataset, as noted in its title, is a dataset of the mortality rate of every county in the United States for heart disease and stroke, separated by race/ethnicity, sex, and age group, with the age groups being 35-64 and 65+. Much of the data ended up being suppressed due to low amounts of the population in each individual county, and in the end, we only ended up using the data for ages 65+, but this was all that was needed for our project. The mortality rates were temporally smoothed over three years for each year and normalized into a rate of deaths attributed per 100,000 people. The deaths were also counted for the county that a person lived in, not where they ended up dying, which is important to note for the clusters of high-profile research hospitals that exist within the city of St. Louis, which could have had an outsized impact on the mortality rate for that county alone.

Our secondary data sources were much more numerous, beginning with the data surrounding comorbidities. One location we got this data from was the CDC, specifically their diabetes database. This dataset had the comorbidities for stroke from diabetes, obesity, and lack of physical activity, which we have already established are some of the most important comorbidities

for stroke. The main unfortunate issue with this dataset was its age grouping, which was only for people 65 and older, due to using Medicare data to achieve its results. Otherwise, this data only went back to 2004, and the comorbidities were only percentage based, but for getting a clear identity of the individual counties, how they have changed, and which ones were the most at risk from these comorbidities, this was more than enough for our project.

The American Community Survey (ACS) is a survey create by the US Census Bureau that looks at data from the previous five years to create a socioeconomic demographic overview of census block groups. This data was used in our project to help understand the outside circumstances that could lead to an increase in stroke, including economic indicators like poverty and multigenerational housing and the overall racial profile of the area as a whole, which is important as data has shown that certain minority groups have an increased chance of stroke when compared on equal grounds otherwise. Unfortunately, the ACS was created in 2005, and as such is our limitation when it comes to years looked in to for our project.

County Health Rankings and Roadmap is a set of demographic rankings similar to both the American Community Survey and the CDC diabetes dataset, created by the University of Wisconsin-Madison. Our primary use of this dataset is looking at any comorbidities that the diabetes dataset does not talk about, such as smoking and drinking. The main issues with this dataset were how the data was formatted, using changing markers for the same data over the years that it was available, which only went back to the year 2011.

ArcGIS was our primary source for our geospatial data, where we plotted the hospitals and created their service areas. The reason we put ArcGIS in our data provenance part of our report is that we used their road network data to create the network analysis vital for our project.

RUCC, standing for Rural-Urban Continuum Codes, are a series of codes created by the United States Department of Agriculture (USDA) Economic Research Service (ERS) to designate how urban a county is. This was primarily used in our project to see how different rural and urban areas are when it comes to stroke mortality and healthcare accessibility.

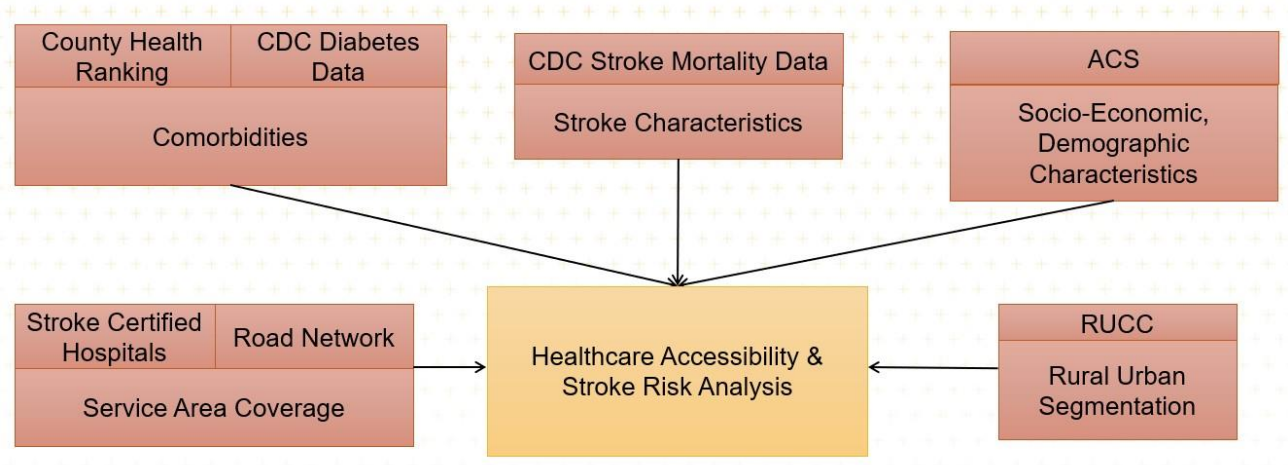

Figure 1: Data Sources

Primary Data

CDC Stroke Mortality Data

Our primary dataset is the dataset called- "**Rates and Trends in Heart Disease and Stroke Mortality Among US Adults (35+) by County, Age Group, Race/Ethnicity, and Sex – 2000-2019**." provided by CDC As quoted by the data.gov website, "This dataset documents rates and trends in heart disease and stroke mortality. Specifically, this report presents county (or county equivalent) estimates of heart disease and stroke death rates in **2000-2019** and trends during two intervals (2000-2010, 2010-2019) by age group (ages 35–64 years, ages 65 years and older), race/ethnicity (non-Hispanic American Indian/Alaska Native, non-Hispanic Asian/Pacific Islander, non-Hispanic Black, Hispanic, non-Hispanic White), and sex (women, men). The rates and trends were estimated using a Bayesian spatiotemporal model and are smoothed over space, time, and demographic group. Rates are age-standardized in 10-year age groups using the 2010 US population. **Data source: National Vital Statistics System.**"

The parts of this that we are most interested in are already separated by race and gender, as well as the different years that are shown. It is important to realize the data is already temporally smoothed, meaning the data for every year is an average of three years for each year and county.

Data Dimensions: 104,856 Rows (All counties in the United States), 21 columns.

Keys: State, County Name/ Code and Year (**Primary**), Stroke per 100K population with break down by Age Group, Race/Ethnicity, and Sex (**Secondary**), Geographic Level: County

Secondary Data

ACS (American County Survey) Data

The ACS 5-Year Data Profiles have been gathered from the **US Census Bureau**, with the dataset containing **Demographics, Social, Housing and Economic** factors. Primarily, we will be using these datasets to collect a detailed understanding of the population in the state of Missouri, looking at age, sex, socio-economic, racial, and housing differences between the counties. The Estimates in ACS data are based on prior 5-year aggregate smoothed estimates. The US Census Bureau provides data from 2009 till 2022; Therefore, we will have to combine 2009 5-year profile with Stroke dataset of years (2000-2009) which creates bias.

Data Dimensions (for each year): 115(Rows) Counties in Missouri, 300 + Variables Comprising of Estimates, Percentages, Margin of errors for each category.

Keys: County Name/ Code and Year (**Primary**), Demographics, Social, Housing and Economic factors (**Secondary**), Geographic Level: County

Geospatial Road Network Data

The data contains a road network of Missouri and 30 miles outside of Missouri's borders, and the speed limit of every road. This dataset, therefore, is the basis for our project when looking at accessibility, as people who are too far away from health centers have no accessibility to it, simply due to distance.

Data Dimensions: (385695,9) - Rows represent all the primary road segments within the state, columns represent the attributes of the roads

Keys: RoadID (**Primary**), Name (**Secondary**)

The key attributes of this dataset include the IDs of the road segments which uniquely identifies the segment of the roads, the names of each road segment, length, the speed the road segment the time for travelling along each road the segments of the road. Geographic Level: State

Census Block Data

Census Block data comes from the MSDIS, or the Missouri Spatial Data Information Service. Our use for this data will be in conjunction with the geospatial data, using the census block populations to both get an understanding of the total population each hospital serves, as well as the percentage of each individual county's population that is contained within the hospital's service area.

Data Dimensions: 253632(Rows) census blocks in Missouri in Missouri, 18 columns

The key attributes of this data set include blockID, name, number of housing units within a block, and the population of each block.

Keys: County Name/ Code and Year (**Primary**), population and houses (**Secondary**), Geographic Level: Block

Hospital Certifications and Locations Data (MO)

Hospital Certifications comes from the **Joint Commission**, the data contains the level of stroke care each hospital provides and the exact location of each hospital. Combining this information with our geospatial data allows us to run network analysis from the point data that the hospitals are to determine time-based distances from said hospitals for the reason of looking at accessibility.

Data Dimensions: (68,10)

Keys: Hospital Name/ID (**Primary**), Location and certification detail **(Secondary)**

Hospital Closures Data

Rural Hospital Closures comes from the **UNC Shep Research Center,** providing point data on rural hospital closures across the nation from **2005 to present**. This data can be used to see the effects closing hospitals have on stroke mortality. As such, this data will also be used with our geospatial data to look at accessibility.

Data Dimensions: (109 Hospital, 14 Variables) All States, (20 Hospital, 14 Variables) MO

Keys: Hospital Name/ID (**Primary**)

Comorbidity Data

- **CDC (Centers for Disease Control) and County Health Ranking**

The Comorbidity data sources are from the CDC and multiple other sources for County Health Ranking data for different measures. This data is formatted like the stroke mortality data; Therefore, similar data structuring is needed to integrate it into our database. We are still currently looking at research papers to determine the comorbidities that have the most impact on mortality rate in conjunction with stroke, and each of their predispositions for stroke occurrence.

Data Dimensions: 115(Rows) Counties in Missouri, Variables to be selected.

Keys: County Name/ Code and Year (**Primary**), Geographic Level: County

Data Shaping and Carpentry

While examining the stroke mortality rates startified by the age brackets (35-64 and 64+), Sex , Race, the rates in age group 35-64 year old have very low mortality rates (20 per 100K) over the years (2005-2019) with no variation. Similar case is observed in Sex and in by Race there were more than 50% suppressed values due to low population density. Therefore, our focus group for analysis is fixed as overall stroke data within age group 64 and older.

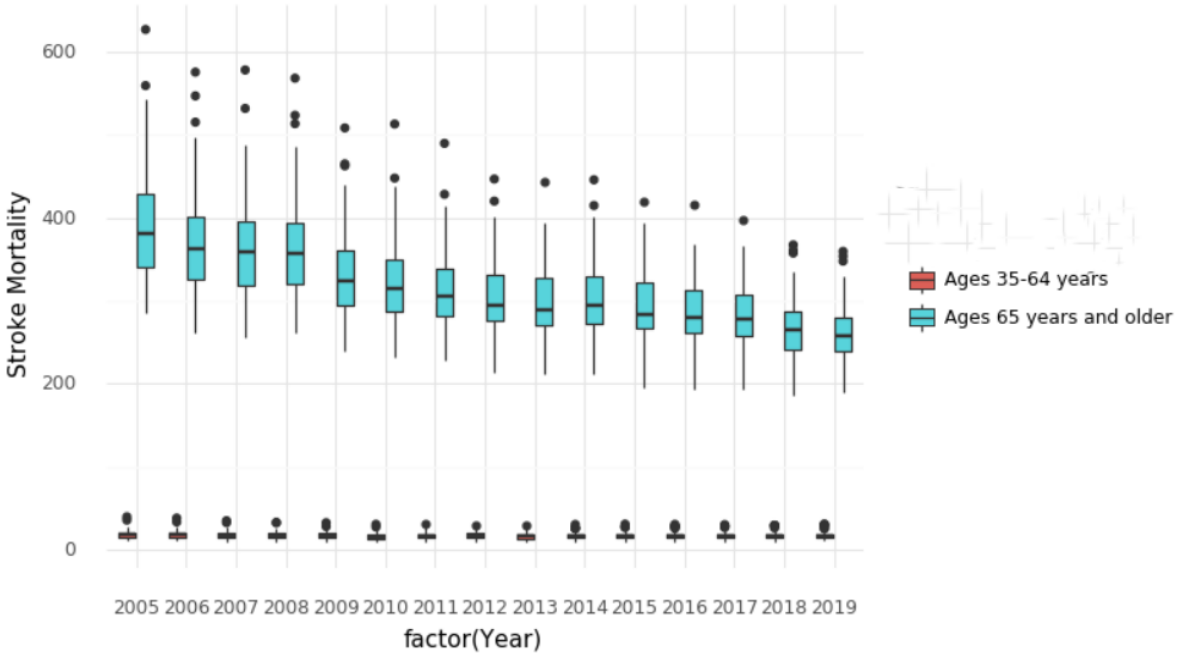

Figure 2: Stroke Mortality Rate – Age Stratified – Temporal Trend

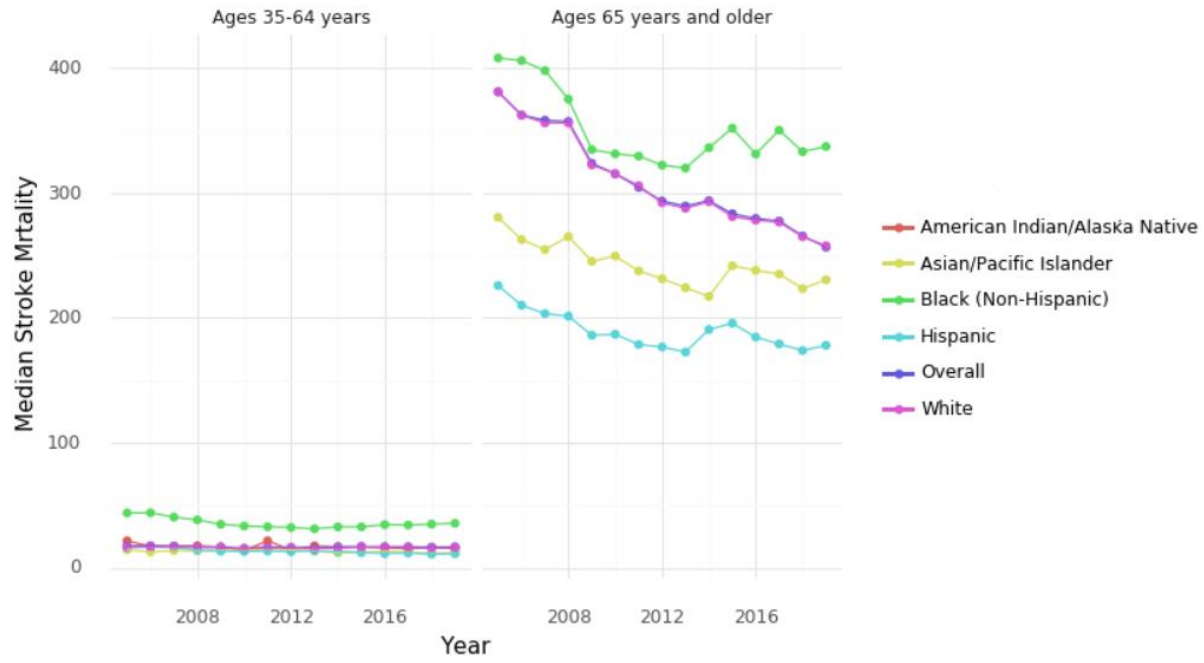

Figure 3: Stroke Mortality Rate – Race Stratified – Temporal Trend

The primary preparatory work that we did for our project was standardization and minimization of the data that we were able to collect. The data from both the CHR (County Health Ranking) and ACS 5-Year Data Profile from the US Census Bureau are extensive which narrowed down by choosing varibles regarding Education , Poverty Rate , Income and Benefits , Crime Rate , Comorbidities , Life style choices. These Attributes had variation in name description over the years which had to be changed to maintain consistency. With the datasets including data from other states and mortality rates for deaths caused by diseases other than stroke were excluded. Paring down these datasets, allowed them to be combined into datasets that could be used for both exploratory data analysis and predictive modelling.

The government data is assumed to be as high quality as possible, due to both other papers that are published using similar data, as well as the fact that these two agencies are highly prestigious for their data acquisition, and if they have faulty data, that will reflect negatively on the US government as a whole.

Unfortunately, the data for certifications is too different from the data for stroke mortality rate and other demographics; as such, they were required to have their own separate dataset. The most that was required for the certification data when it came to data shaping and carpentry was minimizing the certifications to only the state of Missouri.

Methodology

Stroke Mortality is one of the leading causes of death in the state of Missouri, and the best chance for treatment of stroke comes when someone can get treatment in a timely manner. To receive timely healthcare, it depends upon the service coverage ranges and the ability of the individual to attain healthcare services.

But, for a Stroke to occur, there must be a risk for stroke, and certain lifestyle choices lead to comorbidities such as hypertension, obesity, and diabetes, implying that there could be geographical pattern linked to socio-economic factors. These medical conditions and diseases vary in their dispositions to contribute to stroke risk.

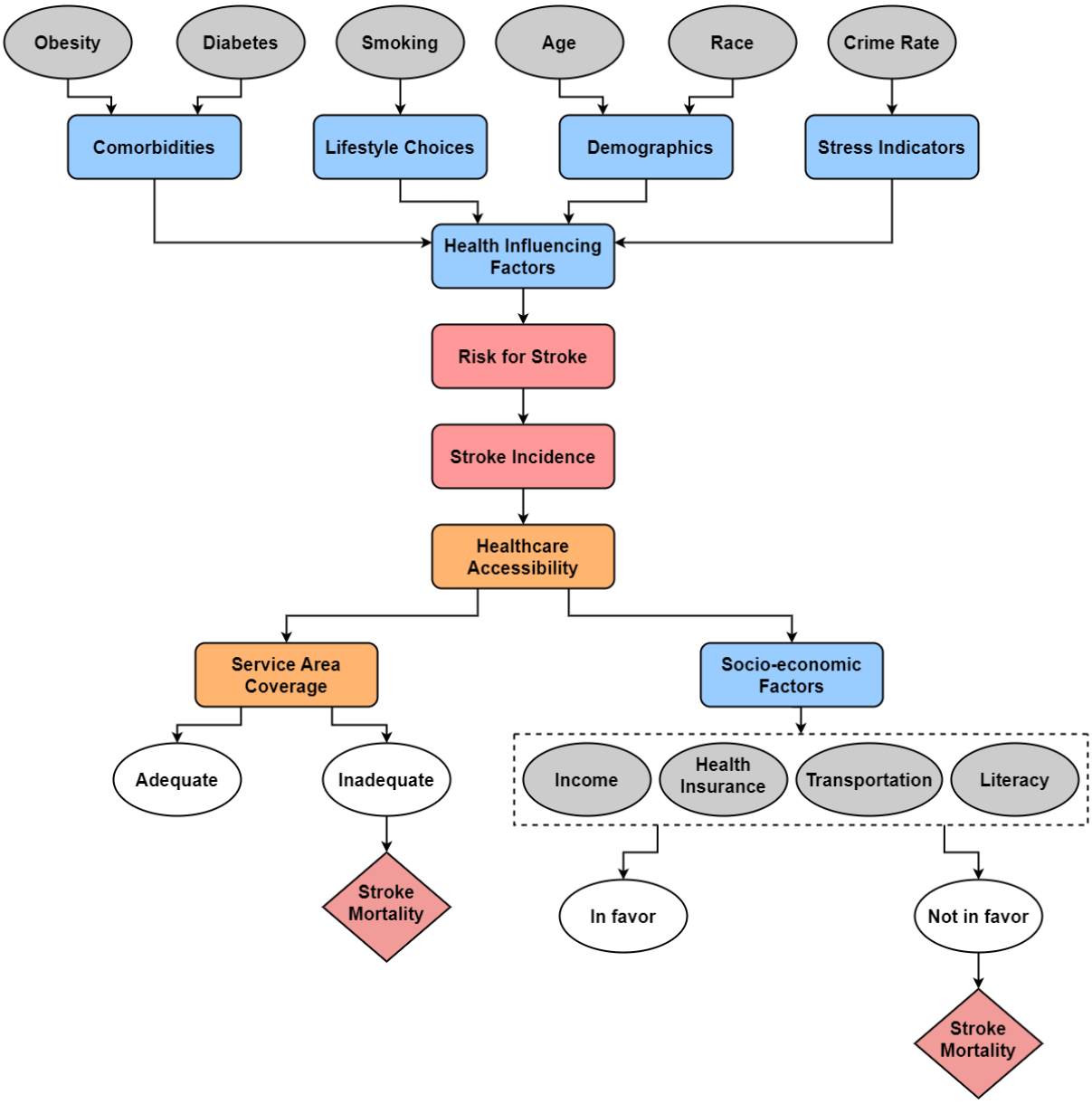

Figure 4: Problem Space Flowchart

The stroke mortality is a result of multiple factors, and for each county, with its unique combination of characteristics, we compare the stroke mortality to understand and determine the risk associating factors for stroke death. Amongst them, healthcare accessibility plays a vital role because it represents the 'supply' aspect of the problem by providing stroke care at varied levels. Whereas the demand is in the places where the risk of stroke incidence is high and where stroke mortality is high.

Even in cases where hospitals are accessible and can provide timely care, financial constraints may prevent individuals from receiving any treatment, leading to increased stroke mortality rates. So, it is important to understand the influence of socio-economic factors over access to healthcare services and how it translates into stroke mortality and service provision.

To understand the factors influencing stroke mortality, we built a model that predicts stroke mortality. For this, we needed a holistic view of the problem, and everything associated with it, to build a representative model. To do so, we conducted preliminary research to familiarize ourselves with the domain, which subsequently helped in identifying and acquiring relevant datasets. Using these datasets, we then conducted Exploratory Data Analysis (EDA) to find trends and patterns.

The healthcare accessibility can also vary among socio-economic factors such as literacy level, income, housing conditions, access to transportation, health insurance coverage, etc. We explore further to check whether these socioeconomic factors aggregate into societal behaviors. Therefore, we then theorize such latent features by further research and exploration.

EDA

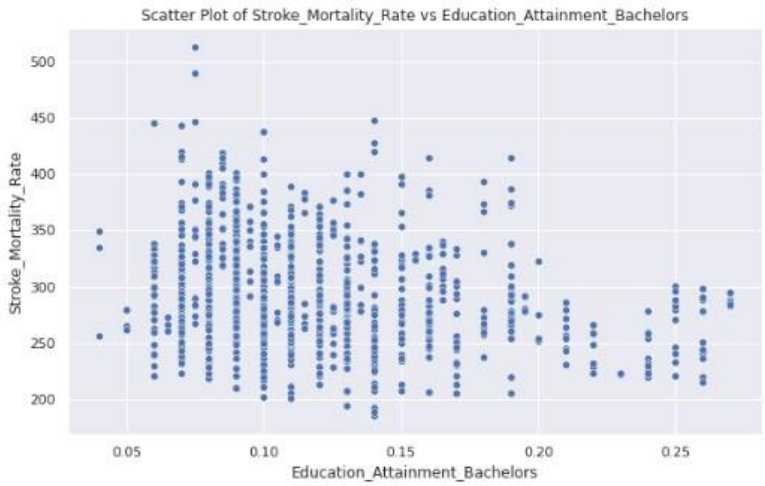

Figure 5: Stroke Mortality Rate v Education Attainment Bachelors

The variable has a linear relationship with the Stroke Mortality Rate.

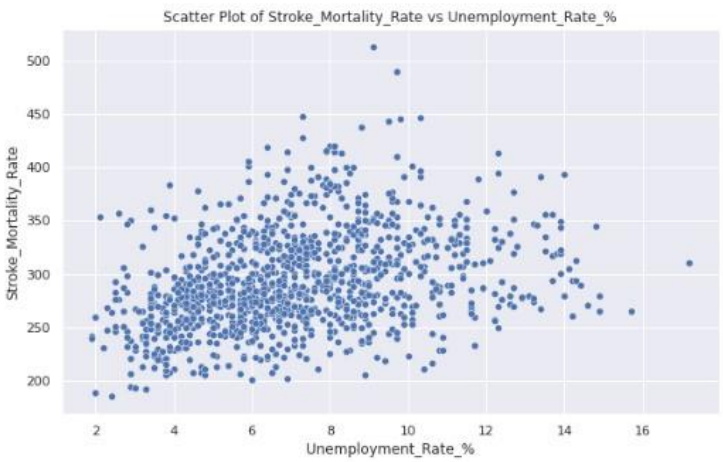

Figure 6: Stroke Mortality Rate v Unemployment Rate %

The variable has a linear relationship with the Stroke Mortality Rate.

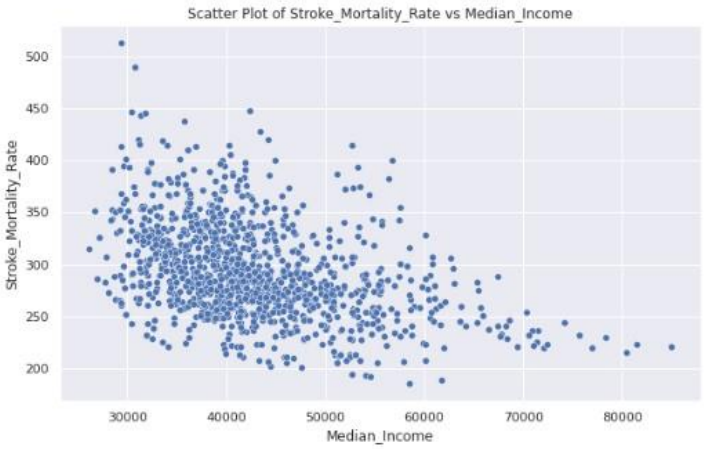

Figure 7: Stroke Mortality Rate v Median Income

The variable has a linear relationship with the Stroke Mortality Rate.

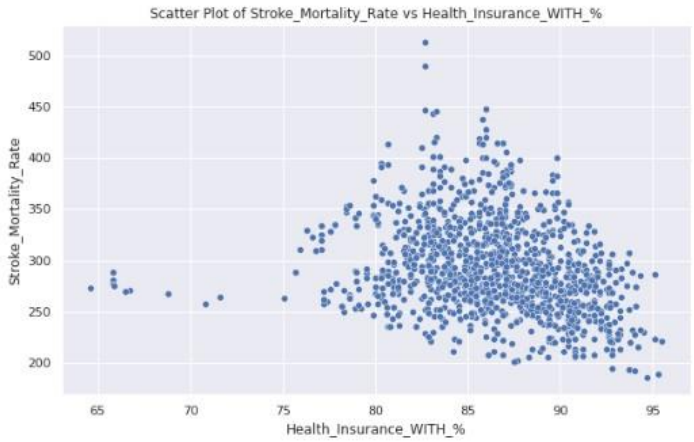

Figure 8: Stroke Mortality Rate v WITH Health Insurance %

The variable has a linear relationship with the Stroke Mortality Rate.

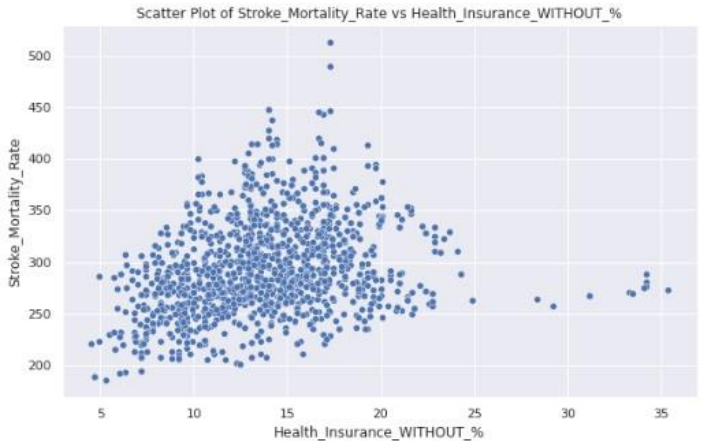

Figure 9: Stroke Mortality Rate v WITHOUT Health Insurance %

The variable has a linear relationship with the Stroke Mortality Rate.

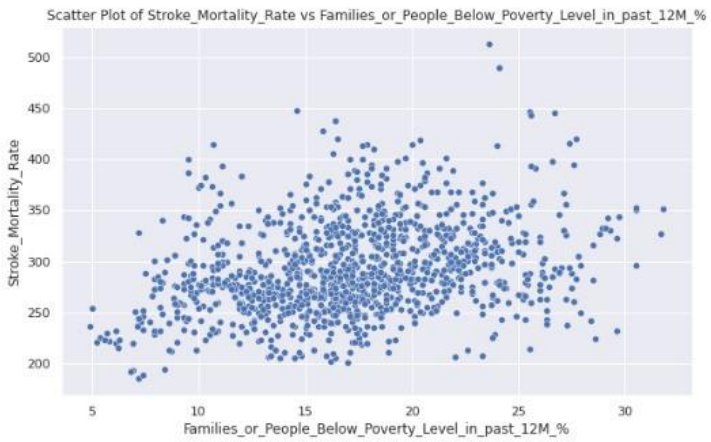

Figure 10: Stroke Mortality Rate v People Below Poverty in Past 12M %

The variable has a linear relationship with the Stroke Mortality Rate.

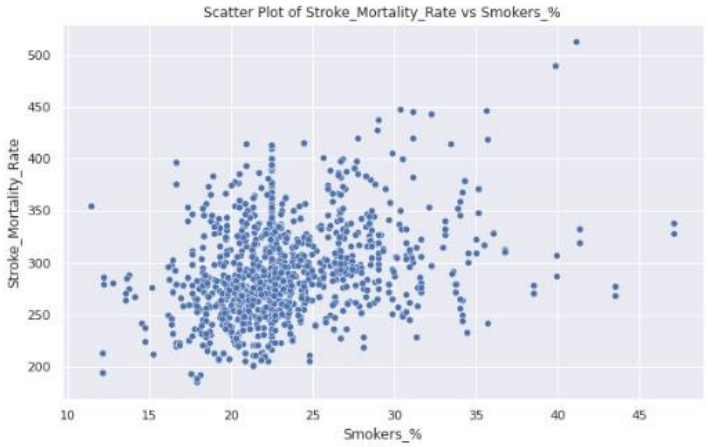

Figure 11: Stroke Mortality Rate v Smokers %

The variable has a linear relationship with the Stroke Mortality Rate.

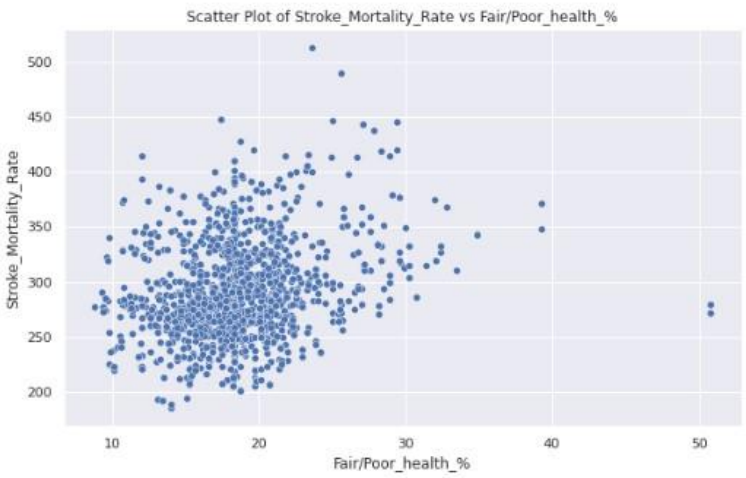

Figure 12: Stroke Mortality Rate v Fair/Poor Health %

The variable has a linear relationship with the Stroke Mortality Rate.

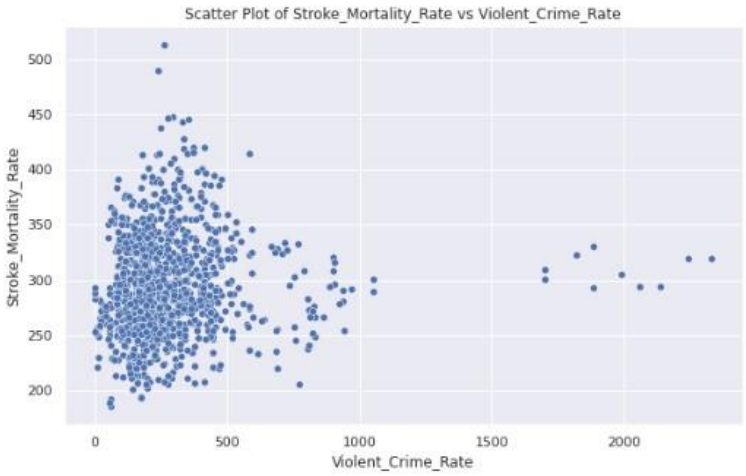

Figure 13: Stroke Mortality Rate v Violent Crime Rate

The variable has somewhat of a linear relationship with the Stroke Mortality Rate.

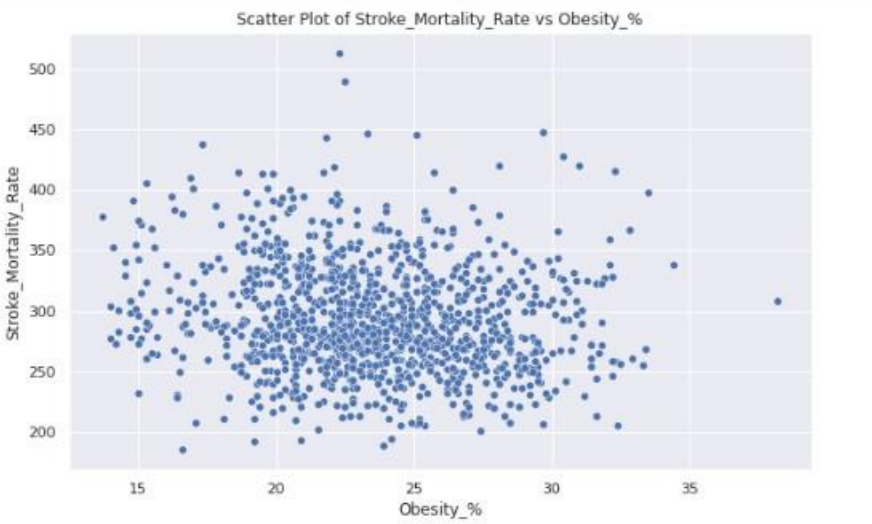

Figure 14: Stroke Mortality Rate v Obesity %

The variable has a non-linear relationship with the Stroke Mortality Rate.

Figure 15: Stroke Mortality Rate v Diabetes %

The variable has a non-linear relationship with the Stroke Mortality Rate.

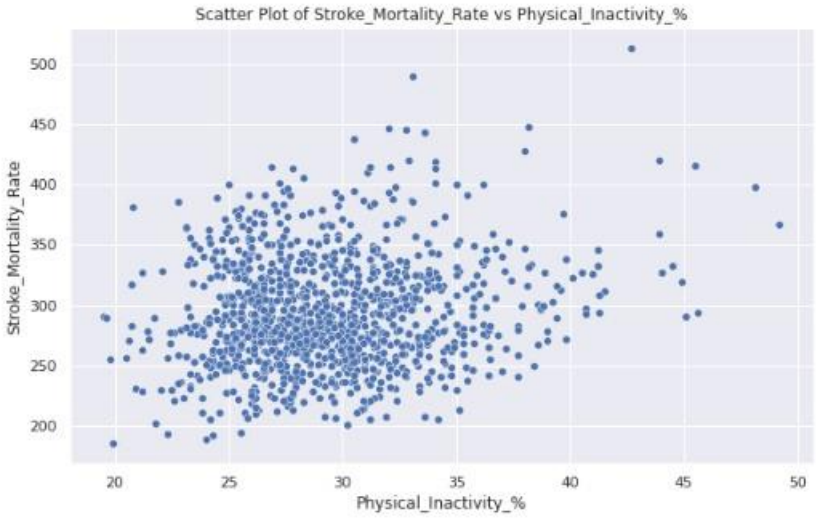

Figure 16: Stroke Mortality Rate v Physical Inactivity %

The variable has somewhat of a linear relationship with the Stroke Mortality Rate.

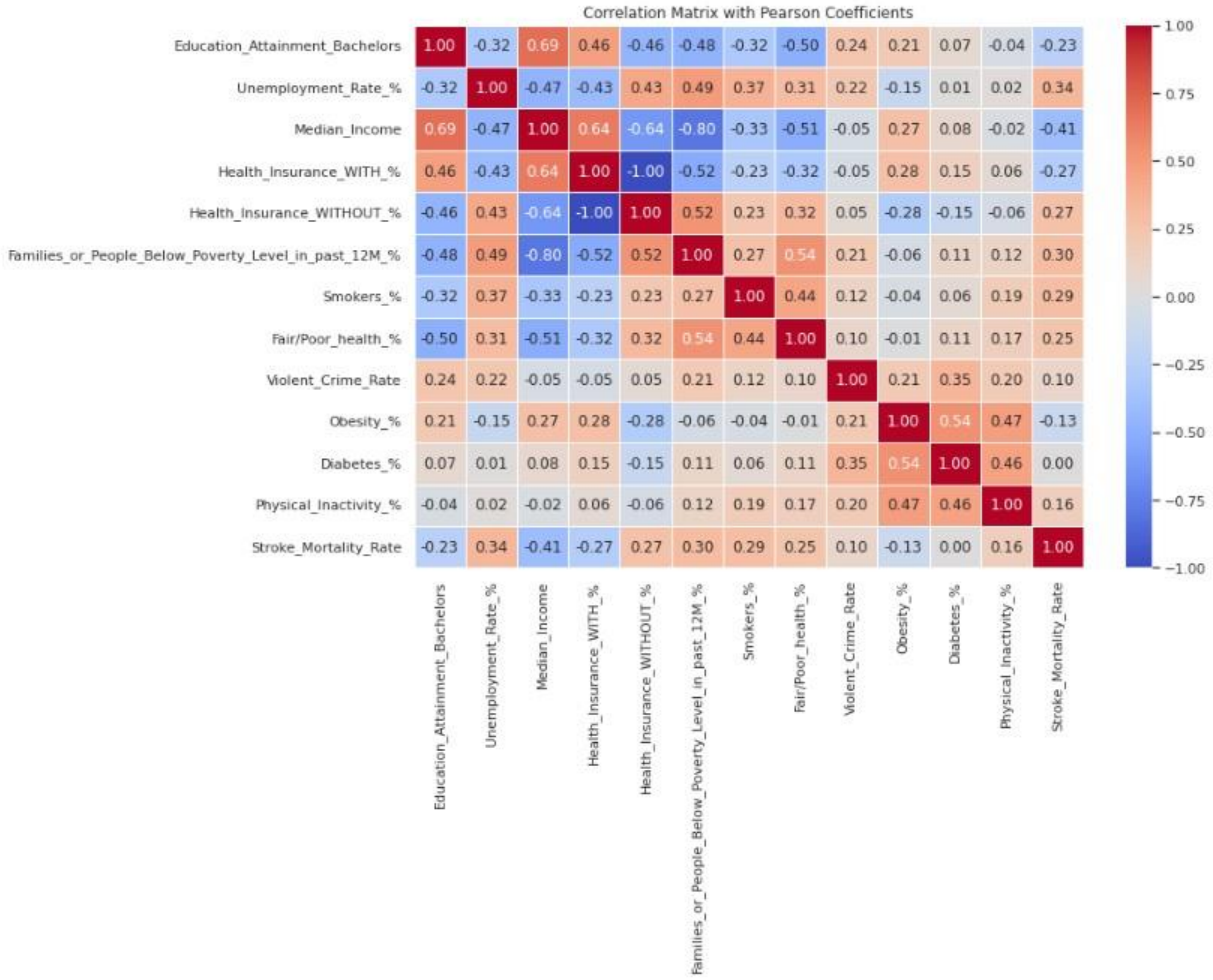

Figure 17: Correlation Matrix

Significance Testing

Stroke Mortality Rates were tested for gaussian distribution using Shapiro-Wilk test and with p value(0.0) suggest that data does not follow gaussian distribution.

Additionally, Independent variables were also tested for significance through Pearson correlation. All varibles tested negative but the following variables have a significant level greater than 0.05. Diabetes, which is considered a comorbidity for Stroke mortality has no significance association and population with education levels School graduate, some College or no Degree were removed from further study.

Educational_Attainment_High_School_Graduate	Co-efficient - 0.020565	p-value - 0.485993	Not Significant
Educational_Attainment_Some_College_No_Degree	Co-efficient - 0.026031	p-value - 0.377810	Not Significant
Diabetes_%	Co-efficient - 0.003089	p-value - 0.916665	Not Significant

Service Area Analysis

This study used service area analysis to ascertain the accessibility of stroke healthcare in Missouri. A Service area refers to a region or area that can be serviced by a point location within a specified time or distance. This study used time-based service area. The time brackets we used for this analysis are: 0-15, 15-30, 30-45, 45-60 mins. We categorize everything else that is beyond 60 mins. as gap areas/60+ mins. We could add more definition to the dataset by including information about the hospitals such as the level of stroke care certification. Lack of resources, workforce, and accommodation denies access to patients even for those within the accessibility range. As such, it is important to make note of such healthcare serviceability factors. However, these factors play a more prevalent role during global events such as the COVID-19 pandemic.

To conduct the service area, we used two main sources of data: 1) The road network data from ESRI road network API; 2) the point locations of all the stroke certified hospitals within the State of Missouri. The service area was done using the ArcGIS online platform, this was chosen because it has the road network data incorporated into it and therefore does not require a separate road network dataset.

The following are the steps to follow for the creation of service areas:

1) The point locations of the stroke certified hospitals are uploaded into the ArcGIS online platform.
2) The Analysis and use proximity geoprocessing tools are selected.
3) The desired time ranges for the service area are selected.
4) The dissolved option is selected to combine service areas for all facilities within the same travel time.

The generated feature classes are then downloaded and loaded into ArcGIS Pro where they are converted into shapefiles.

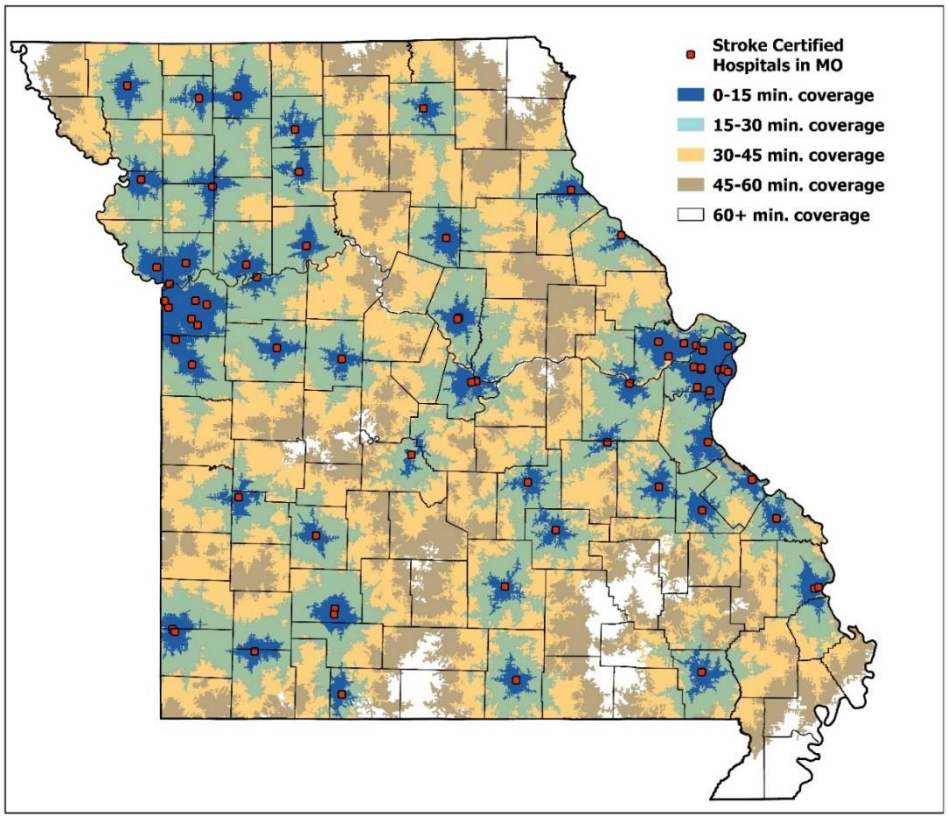

Figure 18: Service Coverage Area Map of the State of Missouri

Data Extraction from Service Areas

The extraction of the data into the format needed for further analysis was done using principally the GeoPandas library via the Jupiter notebook platform. The main data extracted from the service areas is the percentage of people in the various counties who live within the various travel distances: 0-15, 15-30, 30-45, 45-60 mins. and 60+ mins. travel distances. To get the population, we used the 2020 population census data at the block level. This population data was overlayed with the counties shapefile to obtain the percentages of people within each county who live within the various service areas.

This process is repeated for all the travel times to obtain the percentages of coverage for them as well. The result is a data frame with it row being a county and the columns representing the percentage coverage areas.

Using this data frame, we create the Healthcare Accessibility Index to use it as a proxy/quantifiable measure of Accessibility, which is later used in the modelling to understand its effects on Stroke Mortality Rate.

Instead of primary road network data, if service area analysis is conducted using street-level road network data, we will get more accurate and realistic results. By categorizing the population based on insurance coverage types and by including the influence of healthcare services of adjacent states on the patients originating in the state of our study, Missouri, we are producing a more realistic healthcare accessibility representation. If we do not consider such an external factor, we would not be accounting for the deaths that are avoided due to timely access to adjacent state healthcare services. However, certain types of health insurance limit the patient's healthcare to their own state. Also, because there are only a few healthcare services just outside the state's borders, the count of such avoided deaths would be low for Missouri.

Since the healthcare accessibility for different health conditions varies from each other mostly based on targeted response times and since the influence of the demographics and socio-economic factors remains the same, we can reproduce this methodology with minor adjustments to produce similar solutions or recommendations. With this ideology and methodology, we can even conduct similar studies for any state in America.

RUCC Binned

To embed the information about rurality into the model, the RUCC codes were used. But, the RUCC codes spanning 9 continuum codes, were clubbed together into 3 classes – 1,2,3, based on Metro and Non-metro classes and their adjacency to urban counties, to create more concentrated classes. The following map is a visualization of these new classes of RUCC.

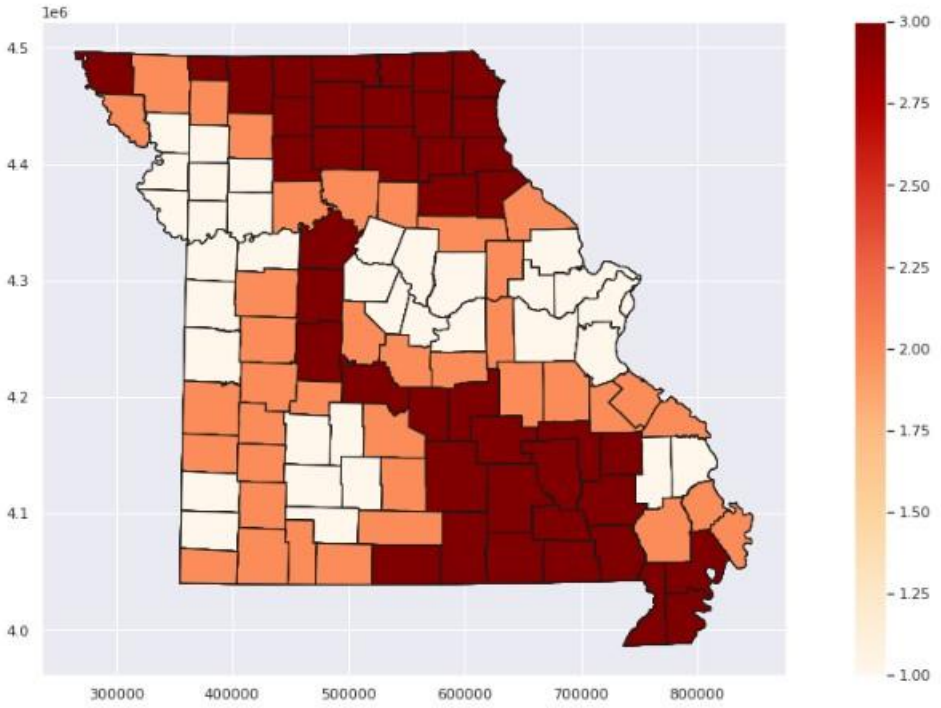

Figure 19: RUCC Binned Map

Analytical Pipeline

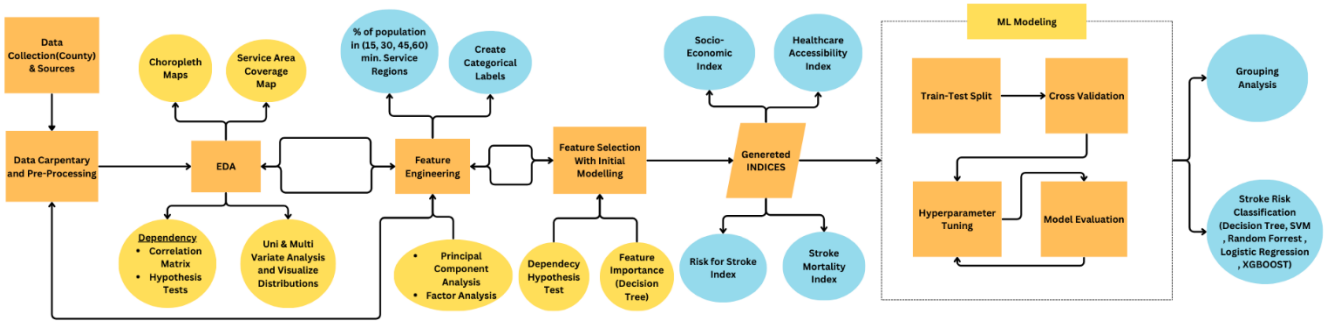

Figure 20: Flowchart of our Analytical Pipeline

LISA (Local Indicator of Spatial Association)

LISA is a concept used in spatial analysis to identify clusters of similar values within a geographic area, resulting from spatial autocorrelation of a particular variable. Using this, the Hot-spots and Cold-spots as clusters, for each variable are derived. While using LISA, several statistical measures help in describing the spatial autocorrelation and significance of the clusters formed. One of the statistical measures is Moran's I, which basically explains the spatial autocorrelation. A negative Moran's I indicate that the observations are surrounded by dissimilar values – resulting in no clusters, and rather is deemed as outliers. A positive Moran's I indicate that the observations are surrounded by similar values, but the nature of the clusters – Hot-spot or Cold-spot, is dictated by the magnitude of Moran's I. A low-positive Moran's I, indicates a Cold-spot and a high-positive Moran's I, indicates a Hot-spot. As for the significance of the spatial patterns observed, we do significance testing to determine whether the spatial patterns are due to randomness or not. The null hypothesis for LISA significance testing states that there is no spatial autocorrelation, implying that the observed spatial patterns are random. If the p-value is less than 0.05, we can reject the null hypothesis, and say that the spatial clusters formed are significant, and are not due to randomness. The Hot-spots and Cold-spots are given HH and LL labels respectively. Aside from the HH and LL labels, there are HL and LH labels which signify unit-regions which are High surrounded by Low unit-regions and unit-regions which are Low surrounded by High unit-regions. These HL and LH labels are outliers and can be discarded from our analysis. Only the HH and LL labels are retained after conducting LISA for each variable. Since the spatial pattern of each variable is to be extracted based on their original counterparts, we conduct LISA for each variable, for all the years separately. This ensures that granular information is retained.

LISA on Stroke Mortality Rate:

2010:

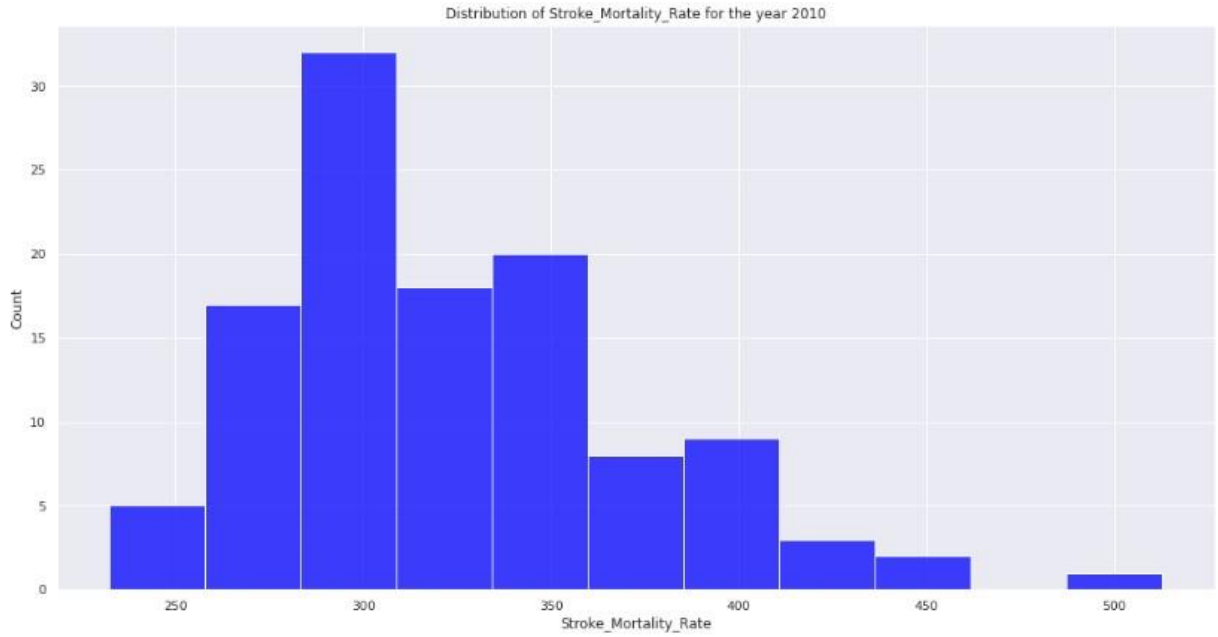

Figure 21: Stroke Mortality Rate Distribution 2010

The following visualization represents the HH, LL, HL, LH clusters for Stroke Mortality Rate for the year 2010:

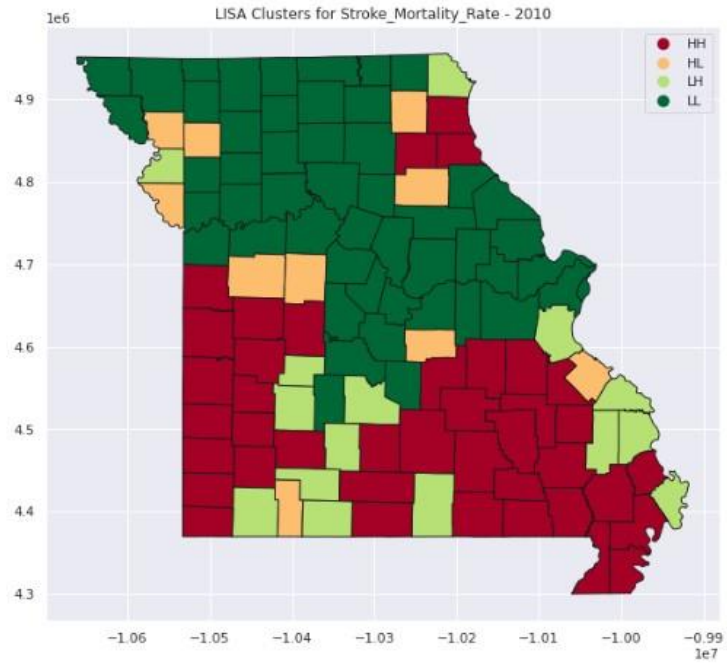

Figure 22: Stroke Mortality Rate LISA Clusters 2010

2019:

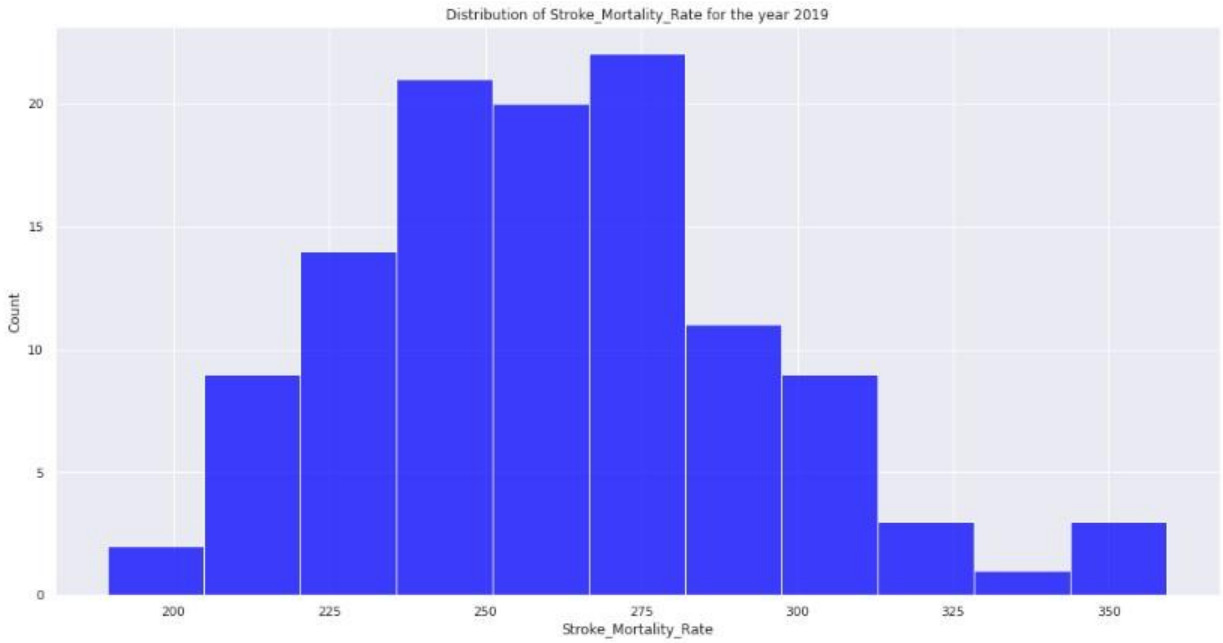

Figure 23: Stroke Mortality Rate Distribution 2019

The following visualization represents the HH, LL, HL, LH clusters for Stroke Mortality Rate for the year 2019:

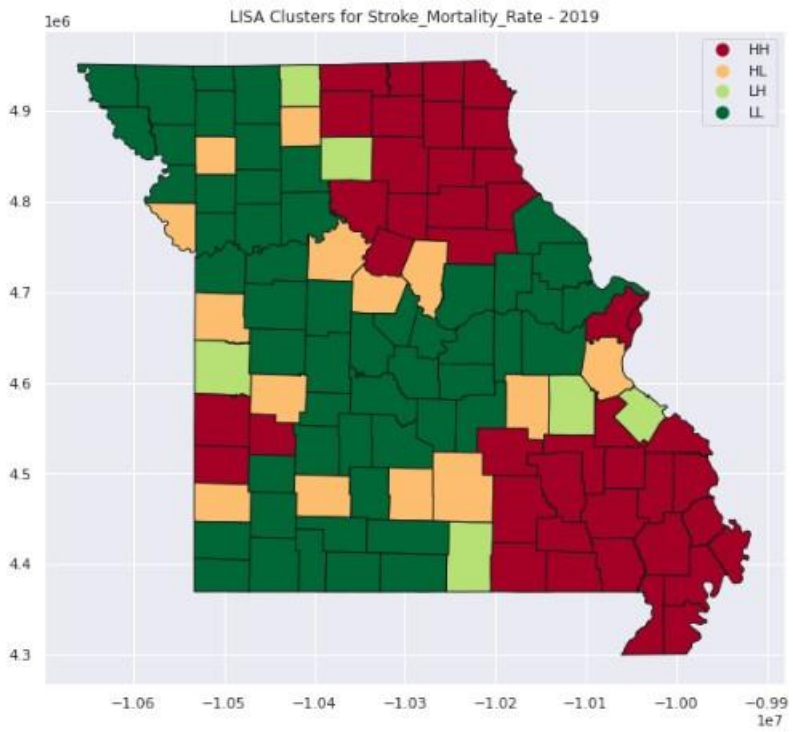

Figure 24: Stroke Mortality Rate LISA Clusters 2019

The variation in a variable is generally small for a few years. However, over several years, like a decade, things might change. As depicted in the 2 visualizations for Stroke Mortality Rate, the HH Clusters have shifted spatially over a decade.

Similarly, to extract this sort of additional information, LISA was conducted for every variable for each year.

Modelling

Vanilla/Baseline Model Results (Model with original variables): (DV – HH & LL Stroke Mortality Class Labels)

```
Random Forest Accuracy: 0.81
Classification Report for Random Forest:
              precision    recall  f1-score   support

          HH       0.83      0.77      0.80        93
          LL       0.80      0.85      0.83       100

    accuracy                           0.81       193
   macro avg       0.81      0.81      0.81       193
weighted avg       0.81      0.81      0.81       193
```

```
Feature Importances for Random Forest:
Median_Income                                          0.153203
Families_or_People_Below_Poverty_Level_in_past_12M_%   0.114834
Education_Attainment_Bachelors                         0.100896
Health_Insurance_WITHOUT_%                             0.089513
Violent_Crime_Rate                                     0.088837
Health_Insurance_WITH_%                                0.082470
Unemployment_Rate_%                                    0.078298
Fair/Poor_health_%                                     0.074575
Smokers_%                                              0.054851
Physical_Inactivity_%                                  0.045362
Obesity_%                                              0.044705
Diabetes_%                                             0.039510
RUCC_3                                                 0.014038
RUCC_2                                                 0.009552
RUCC_1                                                 0.009356
dtype: float64
```

Figure 25: Vanilla Model RF Results

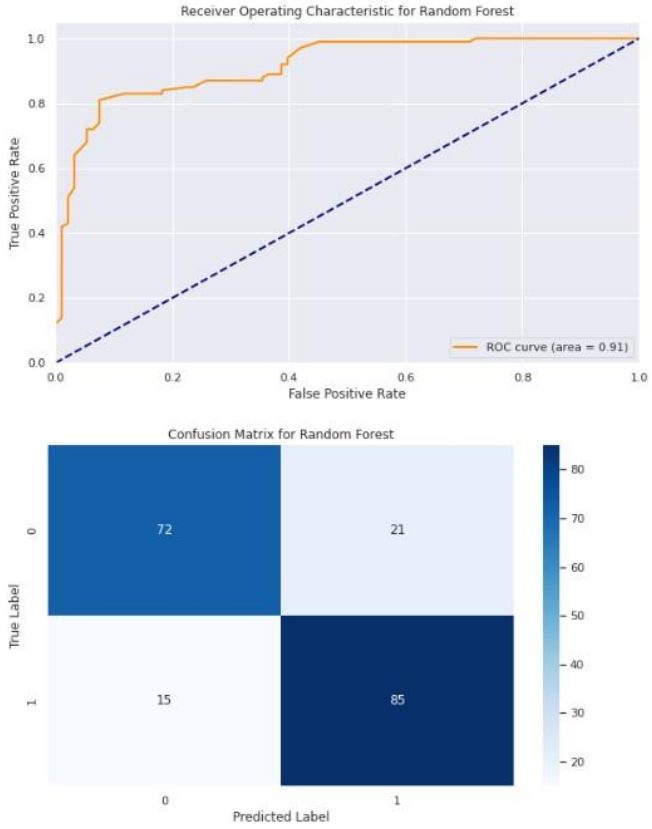

Figure 26: Vanilla Model RF Results

```
Decision Tree Accuracy: 0.75
Classification Report for Decision Tree:
              precision    recall  f1-score   support

          HH       0.74      0.72      0.73        93
          LL       0.75      0.77      0.76       100

    accuracy                           0.75       193
   macro avg       0.75      0.75      0.75       193
weighted avg       0.75      0.75      0.75       193


Feature Importances for Decision Tree:
Median_Income                                          0.288409
Violent_Crime_Rate                                     0.102021
Unemployment_Rate_%                                    0.088438
Health_Insurance_WITH_%                                0.084278
Fair/Poor_health_%                                     0.082710
Families_or_People_Below_Poverty_Level_in_past_12M_%   0.073045
Education_Attainment_Bachelors                         0.061554
Smokers_%                                              0.057808
Physical_Inactivity_%                                  0.034849
Diabetes_%                                             0.030778
Health_Insurance_WITHOUT_%                             0.029312
RUCC_2                                                 0.028949
Obesity_%                                              0.027398
RUCC_1                                                 0.010451
RUCC_3                                                 0.000000
dtype: float64
```

Figure 27: Vanilla Model DT Results

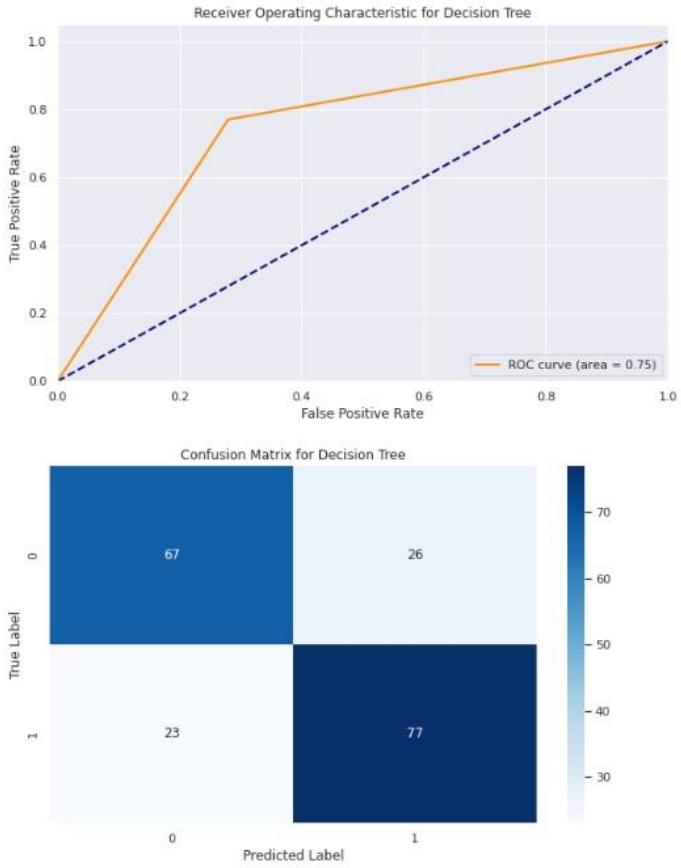

Figure 28: Vanilla Model DT Results

```
Cross-validation scores: [0.89690722 0.90625    0.94791667 0.85416667 0.89583333 0.875
 0.90625    0.82291667 0.75       0.73958333]
Mean accuracy: 0.8594823883161512
Standard deviation: 0.06548689293774776

RandomForestClassifier(random_state=43)
```

Figure 29: Vanilla Model RF CV Results

Model with Geo-Labels from LISA: (DV – HH & LL Stroke Mortality Class Labels)

```
Random Forest Accuracy: 0.83
Classification Report for Random Forest:
              precision    recall  f1-score   support

          HH       0.84      0.80      0.82        93
          LL       0.82      0.86      0.84       100

    accuracy                           0.83       193
   macro avg       0.83      0.83      0.83       193
weighted avg       0.83      0.83      0.83       193
```

```
Feature Importances for Random Forest:
Median_Income_Geolabels_HH                                          0.068282
Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_HH   0.056960
Education_Attainment_Bachelors_Geolabels_LL                         0.056834
Health_Insurance_WITH_%_Geolabels_HH                                0.054752
Health_Insurance_WITHOUT_%_Geolabels_LL                             0.050463
Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_LL   0.050282
Median_Income_Geolabels_LL                                          0.046841
Health_Insurance_WITHOUT_%_Geolabels_HH                             0.046754
Health_Insurance_WITH_%_Geolabels_LL                                0.045693
Fair/Poor_health_%_Geolabels_LL                                     0.042367
Unemployment_Rate_%_Geolabels_LL                                    0.042112
Diabetes_%_Geolabels_LL                                             0.038199
Violent_Crime_Rate_Geolabels_HH                                     0.037773
Physical_Inactivity_%_Geolabels_LL                                  0.036153
Violent_Crime_Rate_Geolabels_LL                                     0.032271
Obesity_%_Geolabels_LL                                              0.030938
Education_Attainment_Bachelors_Geolabels_HH                         0.030219
Unemployment_Rate_%_Geolabels_HH                                    0.027830
Physical_Inactivity_%_Geolabels_HH                                  0.026915
Smokers_%_Geolabels_HH                                              0.026693
Fair/Poor_health_%_Geolabels_HH                                     0.026335
Obesity_%_Geolabels_HH                                              0.025034
RUCC_3                                                              0.023767
Smokers_%_Geolabels_LL                                              0.021583
RUCC_2                                                              0.021464
Diabetes_%_Geolabels_HH                                             0.018104
RUCC_1                                                              0.015382
dtype: float64
```

Figure 30: Geo-labels Model RF Results

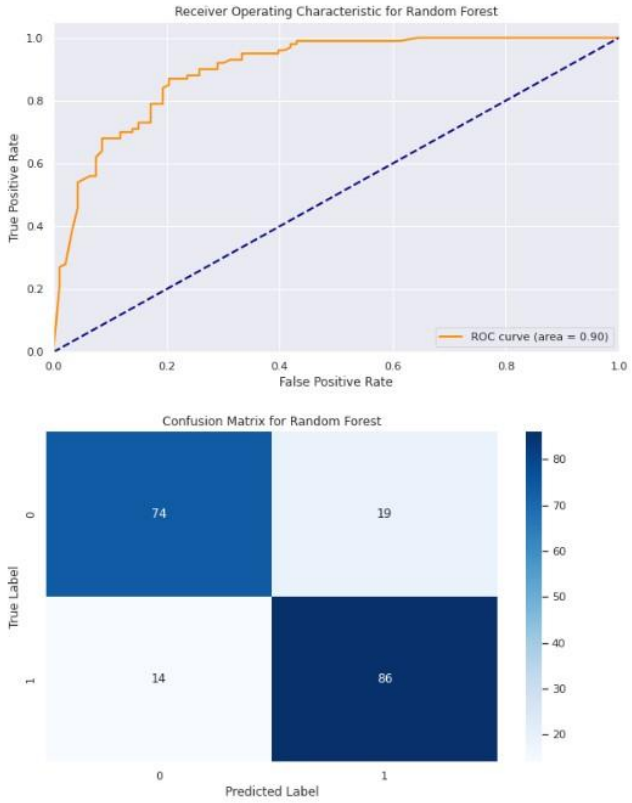

Figure 31: Geo-labels Model RF Results

```
Decision Tree Accuracy: 0.80
Classification Report for Decision Tree:
              precision    recall  f1-score   support

          HH       0.77      0.83      0.80        93
          LL       0.83      0.77      0.80       100

    accuracy                           0.80       193
   macro avg       0.80      0.80      0.80       193
weighted avg       0.80      0.80      0.80       193


Feature Importances for Decision Tree:
Median_Income_Geolabels_HH                                          0.336426
Unemployment_Rate_%_Geolabels_LL                                    0.101949
Violent_Crime_Rate_Geolabels_HH                                     0.063807
Diabetes_%_Geolabels_LL                                             0.038774
Education_Attainment_Bachelors_Geolabels_LL                         0.035884
Fair/Poor_health_%_Geolabels_HH                                     0.035112
Fair/Poor_health_%_Geolabels_LL                                     0.033176
Obesity_%_Geolabels_LL                                              0.031625
Physical_Inactivity_%_Geolabels_LL                                  0.031495
Violent_Crime_Rate_Geolabels_LL                                     0.028768
RUCC_3                                                              0.024663
Physical_Inactivity_%_Geolabels_HH                                  0.023038
RUCC_1                                                              0.022842
Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_LL   0.022705
Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_HH   0.021356
Education_Attainment_Bachelors_Geolabels_HH                         0.018371
Diabetes_%_Geolabels_HH                                             0.017559
RUCC_2                                                              0.016803
Obesity_%_Geolabels_HH                                              0.013785
Health_Insurance_WITH_%_Geolabels_LL                                0.013484
Health_Insurance_WITHOUT_%_Geolabels_LL                             0.012514
Smokers_%_Geolabels_HH                                              0.012178
Median_Income_Geolabels_LL                                          0.011897
Health_Insurance_WITHOUT_%_Geolabels_HH                             0.010475
Smokers_%_Geolabels_LL                                              0.008509
Unemployment_Rate_%_Geolabels_HH                                    0.008152
Health_Insurance_WITH_%_Geolabels_HH                                0.004655
dtype: float64
```

Figure 32: Geo-labels Model DT Results

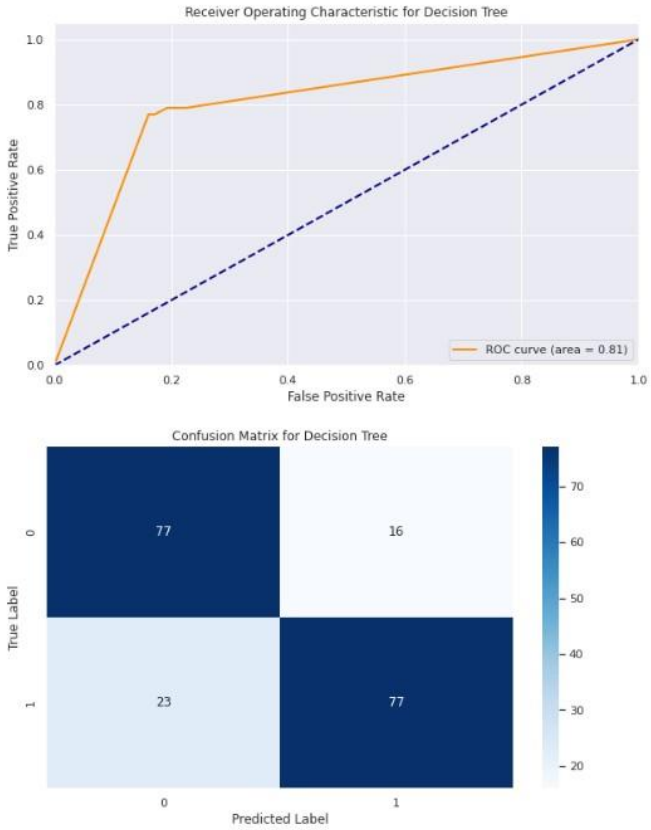

Figure 33: Geo-labels Model DT Results

```
Cross-validation scores: [0.80412371 0.86458333 0.85416667 0.9375     0.91666667 0.88541667
 0.88541667 0.82291667 0.73958333 0.6875    ]
Mean accuracy: 0.8397873711340207
Standard deviation: 0.074318680640939

RandomForestClassifier(random_state=43)
```

Figure 34: Geo-labels Model RF CV Results

Both the models – Vanilla model and the model with Geo-Labels, have similar performance results.

Indices

Indices are used to operationalize the concepts/factors which each group of variables represents. The Indices are deemed as latent features resulting from amalgamation of their respective buckets of variables. By creating the Indices as new engineered features, we can use them as conceptual shorthand or means to quantify the themes they represent.

Based on Feature Importance:

The Feature Importance Scores from the Random Forest Classifier were used as weights to create a weighted sum of the scaled input variables, which results in an Index. So, by doing this, 2 Indices were created – Socioeconomic Vulnerability Index and Risk for Stroke Index. The Feature Importance Scores are used as weights to create a weighted sum because those scores are given to each variable in terms of how important those variables are in predicting the measurable outcome which is Stroke Mortality in our case. The meaning or representation of an Index varied based on what Dependent Variable it predicts or has an influence over. So, in terms of creating a Socio-economic themed Index which influences Stroke Mortality, the resulting Index would be a representation of Socio-economic Vulnerability. Saying that the variation in Socio-economic Vulnerability translates into variation in Stroke Mortality. Similarly, Comorbidities' themed set of variables with a disposition for Stroke results in a Risk for Stroke Index.

```
Random Forest Accuracy: 0.83
Classification Report for Random Forest:
              precision    recall  f1-score   support

          HH       0.84      0.80      0.82        93
          LL       0.82      0.86      0.84       100

    accuracy                           0.83       193
   macro avg       0.83      0.83      0.83       193
weighted avg       0.83      0.83      0.83       193


Feature Importances for Random Forest:
Median_Income_Geolabels_HH                                        0.068282
Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_HH 0.056960
Education_Attainment_Bachelors_Geolabels_LL                       0.056834
Health_Insurance_WITH_%_Geolabels_HH                              0.054752
Health_Insurance_WITHOUT_%_Geolabels_LL                           0.050463
Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_LL 0.050282
Median_Income_Geolabels_LL                                        0.046841
Health_Insurance_WITHOUT_%_Geolabels_HH                           0.046754
Health_Insurance_WITH_%_Geolabels_LL                              0.045693
Fair/Poor_health_%_Geolabels_LL                                   0.042367
Unemployment_Rate_%_Geolabels_LL                                  0.042112
Diabetes_%_Geolabels_LL                                           0.038199
Violent_Crime_Rate_Geolabels_HH                                   0.037773
Physical_Inactivity_%_Geolabels_LL                                0.036153
Violent_Crime_Rate_Geolabels_LL                                   0.032271
Obesity_%_Geolabels_LL                                            0.030938
Education_Attainment_Bachelors_Geolabels_HH                       0.030219
Unemployment_Rate_%_Geolabels_HH                                  0.027830
Physical_Inactivity_%_Geolabels_HH                                0.026915
Smokers_%_Geolabels_HH                                            0.026693
Fair/Poor_health_%_Geolabels_HH                                   0.026335
Obesity_%_Geolabels_HH                                            0.025034
RUCC_3                                                            0.023767
Smokers_%_Geolabels_LL                                            0.021583
RUCC_2                                                            0.021464
Diabetes_%_Geolabels_HH                                           0.018104
RUCC_1                                                            0.015382
dtype: float64
```

Figure 35: Geo-labels Model RF Results for Feature Importance Scores

SOCIO-ECONOMIC VULNERABILITY INDEX:

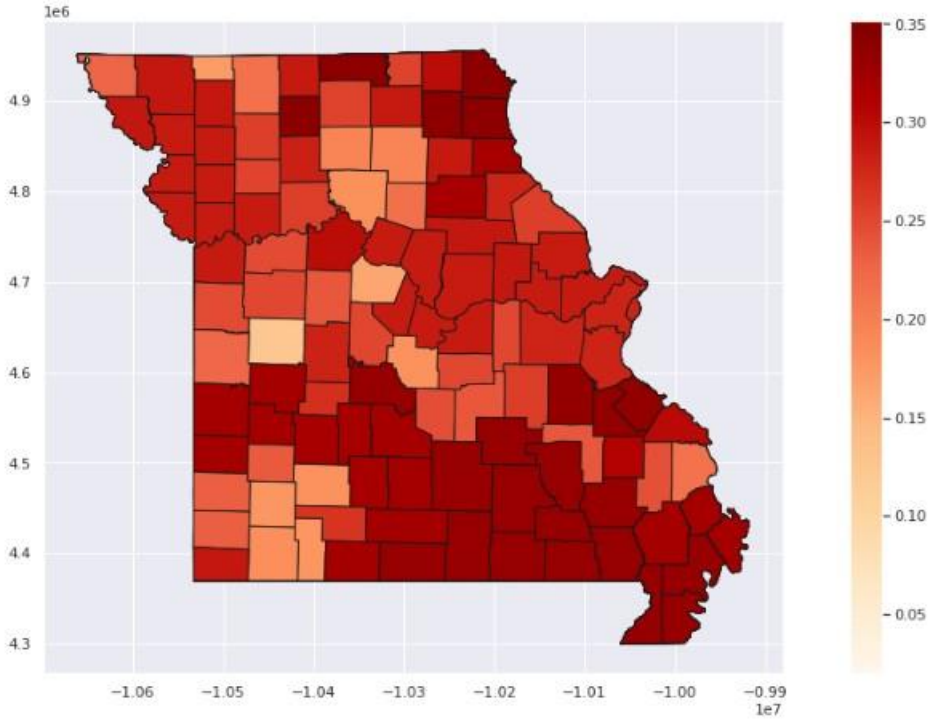

Figure 36: Socio-economic Vulnerability Index Map

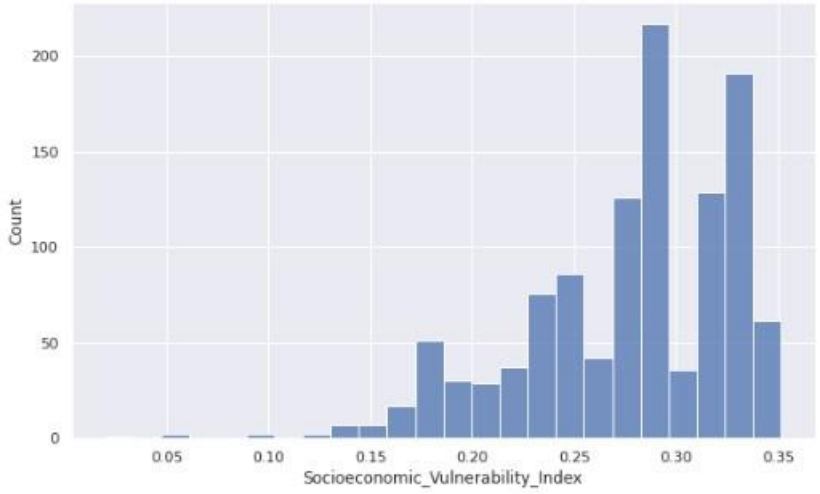

Figure 37: Socio-economic Vulnerability Index Distribution

RISK FOR STROKE INDEX:

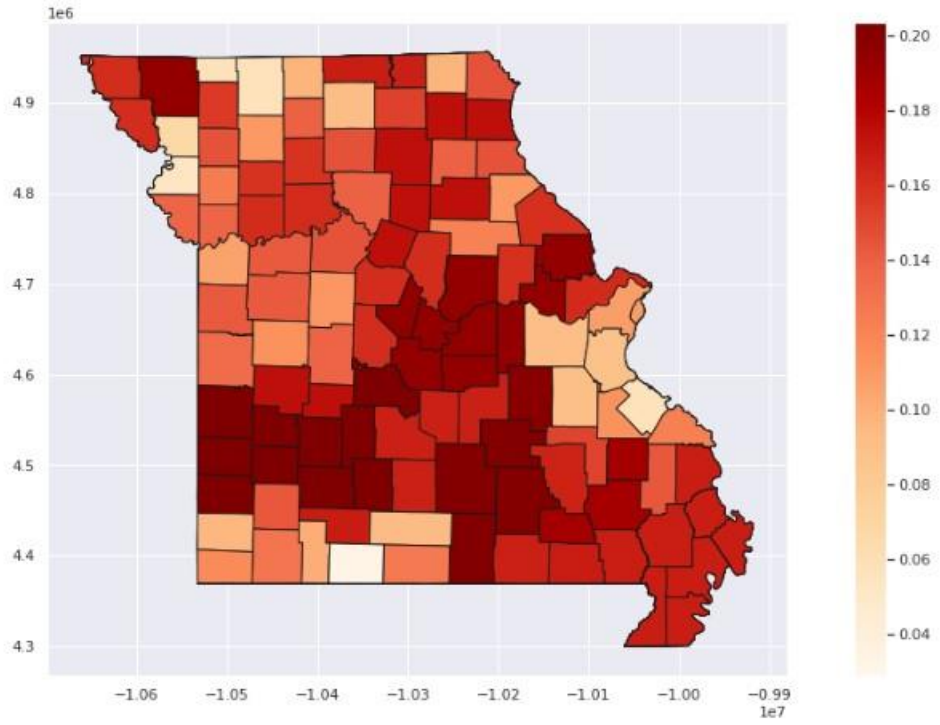

Figure 38: Risk for Stroke Index Map

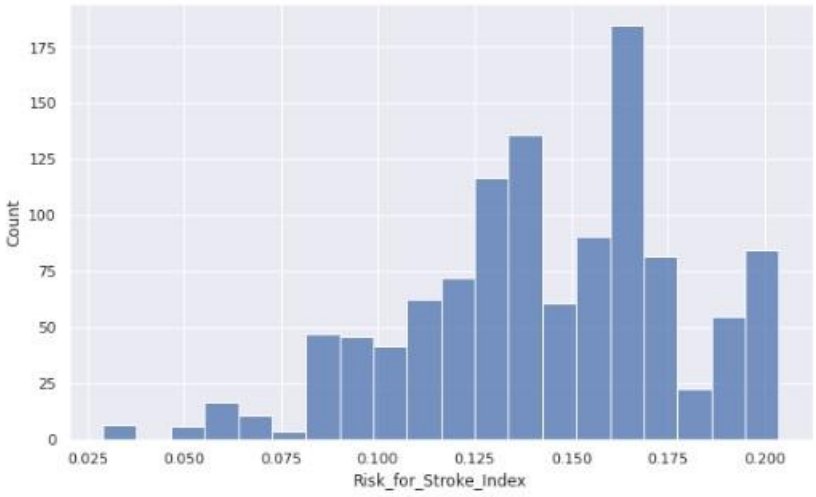

Figure 39: Risk for Stroke Index Distribution

Based on PCA:

Principal Component Analysis (PCA) is a statistical method that keeps as much of the dataset's variation as feasible while reducing the dimensionality of a dataset with many connected variables. To do this, the original variables are converted into a new set of variables that are just linear combinations of the original variables. The primary components, the new variables, are arranged in an orthogonal (uncorrelated) manner, with the first few retaining most of the variation found in all the original variables.

Factor loadings can be extracted - Factor loadings are coefficients that indicate the proportion of an original variable's variance that can be accounted for by a principal component. These loadings can be thought of as weights that represent each variable's significance or contribution to a specific primary component. Factor loadings from the first principal component (or another component of interest) are used as weights when transposing them. The first component often explains most of the variance, and its loadings indicate the relative contribution of each variable to the component.

The transposed factor loadings from the PCA are then used to compute a weighted sum of the scaled variables. An index is created by multiplying each initial variable by the matching loading from the selected primary component, then adding the results. Based on the variance explained by the principal component, this index offers a score for each observation in the dataset that reflects the fundamental idea that the principal component analysis (PCA) was designed to quantify.

Unlike Feature Importance, the Indices created represent the absolute latent feature rather than in terms of a Dependent Variable. So, the Index created reflects the type of variables chosen to amalgamate. Using this, 3 Indices were created – Healthcare Accessibility Index, Socioeconomic Vulnerability Index and Risk for Stroke Index.

Additionally, Jenks Natural Breaks Algorithm was used to create clusters/groups based on the groupings in the frequency distribution. Jenks Natural Breaks Algorithm creates these groups such that there is least variance in the groups and most variance amongst the groups created. By using this algorithm, we create thresholds for creating different classes for the Indices.

HEALTHCARE ACCESSIBILITY INDEX:

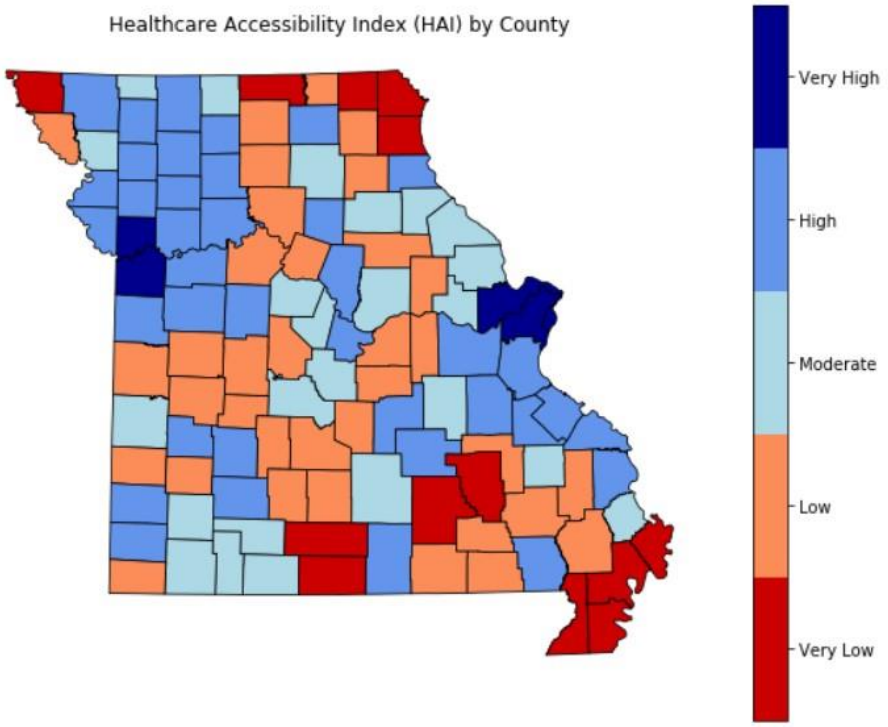

Figure 40: Healthcare Accessibility Index Map

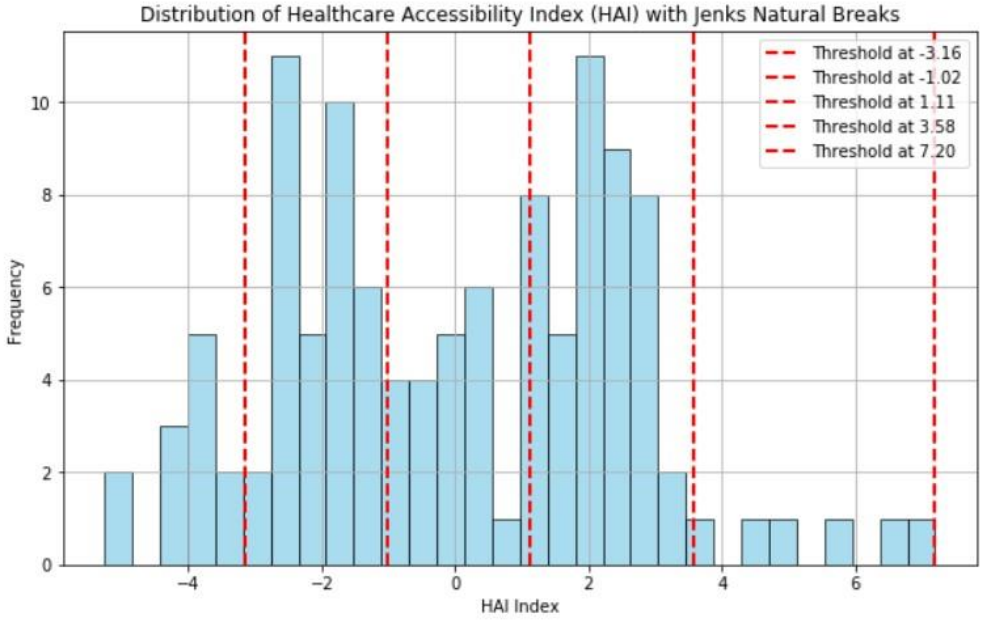

Figure 41: Healthcare Accessibility Index Distribution

SOCIO-ECONOMIC INDEX:

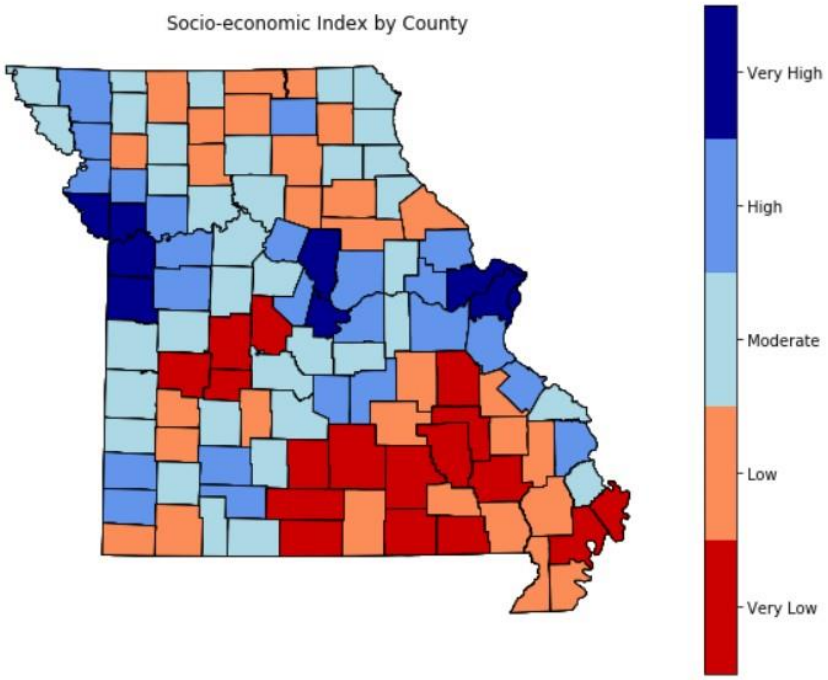

Figure 42: Socio-economic Index Map

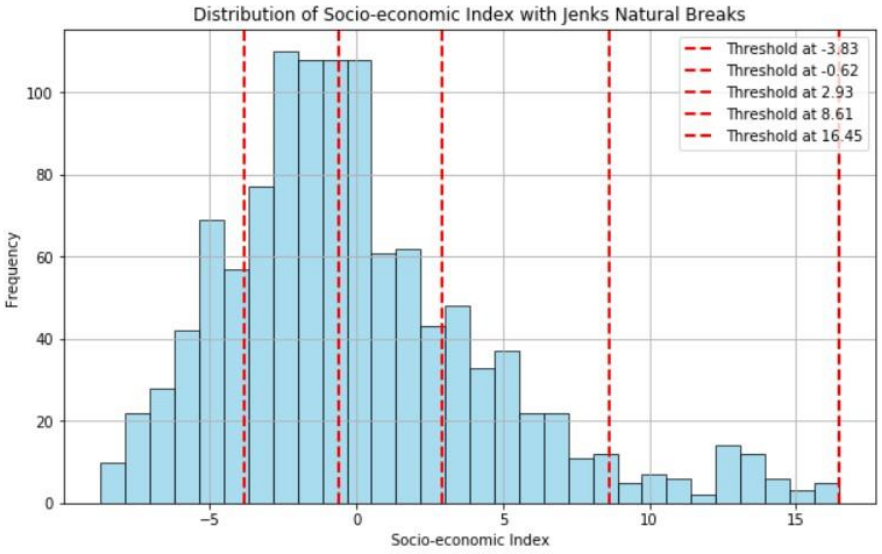

Figure 43: Socio-economic Index Distribution

RISK FOR STROKE INDEX:

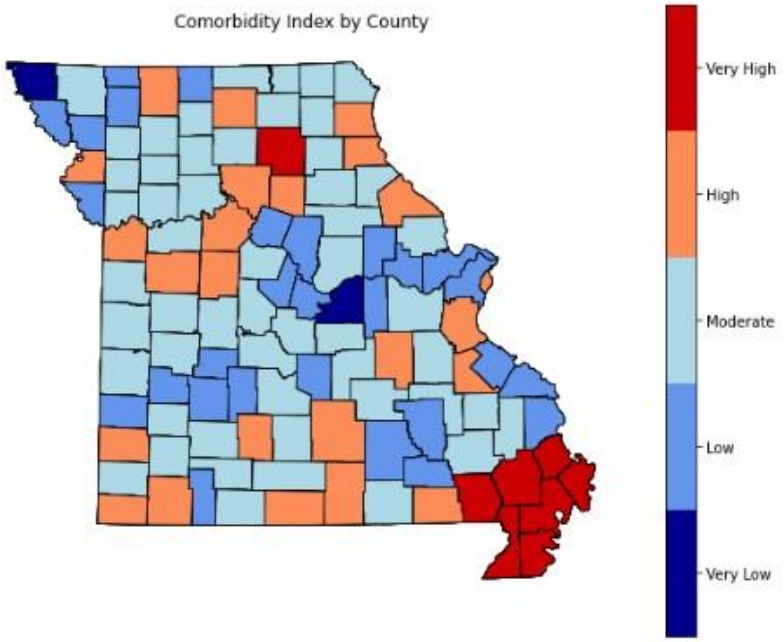

Figure 44: Risk for Stroke Index Map

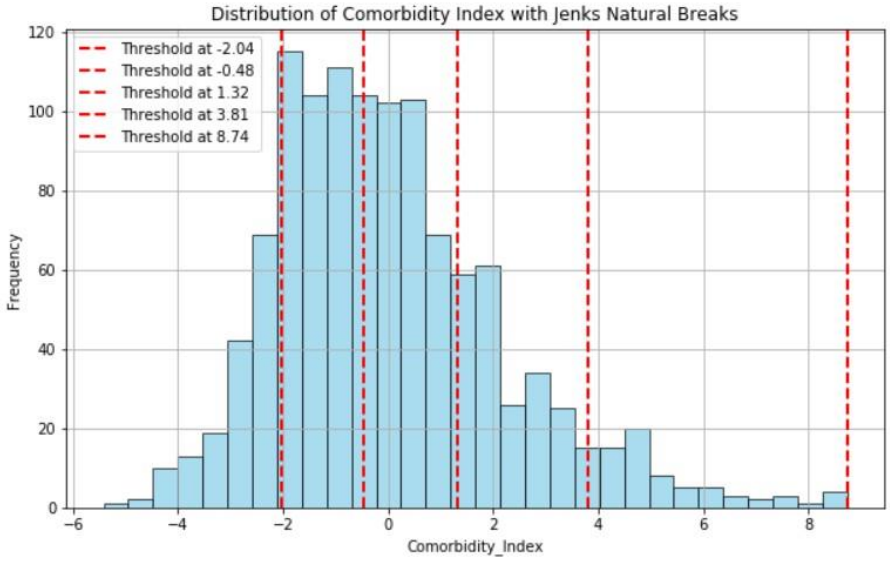

Figure 45: Risk for Stroke Index Distribution

Inferences from maps

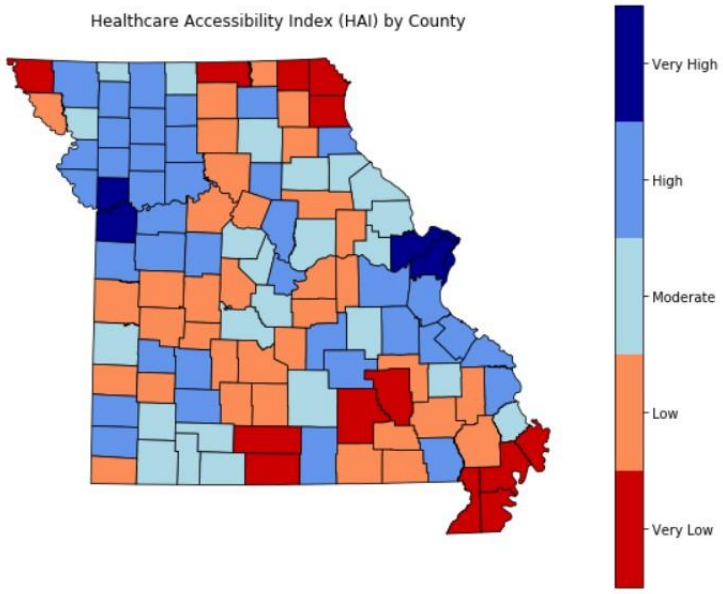

Figure 46: Healthcare Accessibility Index Map

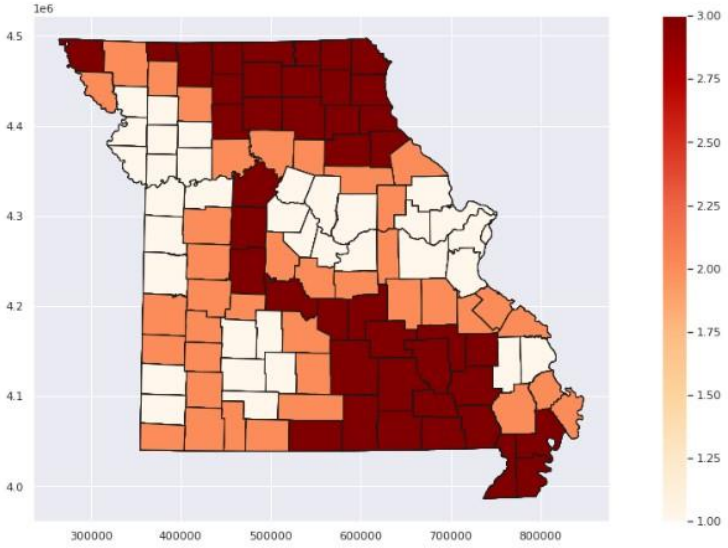

Figure 47: RUCC Binned Map

By comparing these 2 choropleth maps of Healthcare Accessibility Index (HAI) and RUCC Binned, we can see that the High HAI regions are correlated with urban centers.

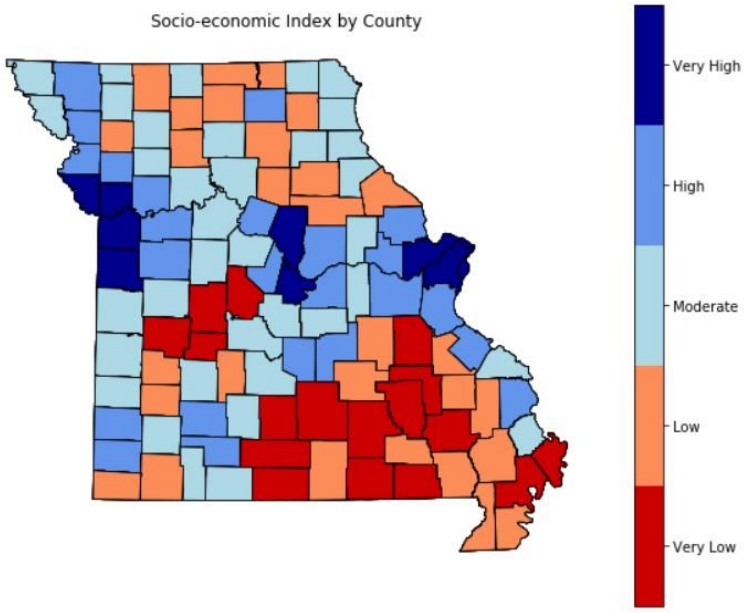

Figure 48: Socio-economic Index Map

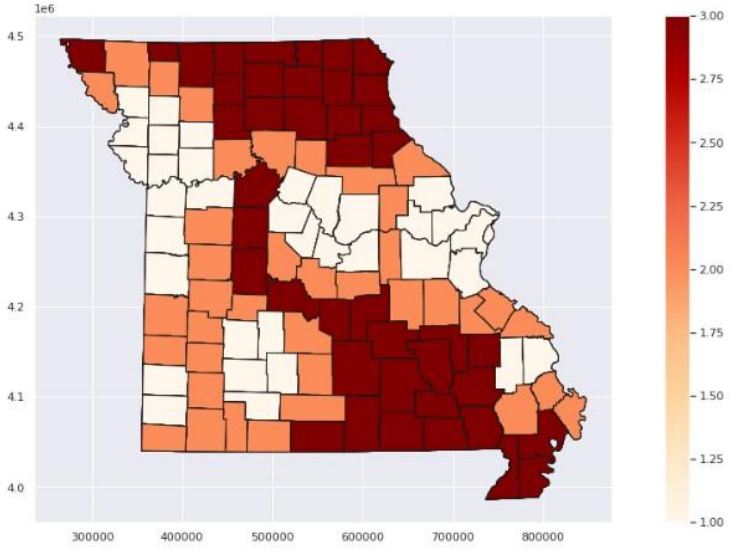

Figure 49: RUCC Binned Map

Similarly, when comparing these 2 choropleth maps of Socioeconomic Index and RUCC Binned, we can see that the High Socioeconomic regions are correlated with urban centers. This should hold true because the level of urbanization should be a part of the socioeconomic factor.

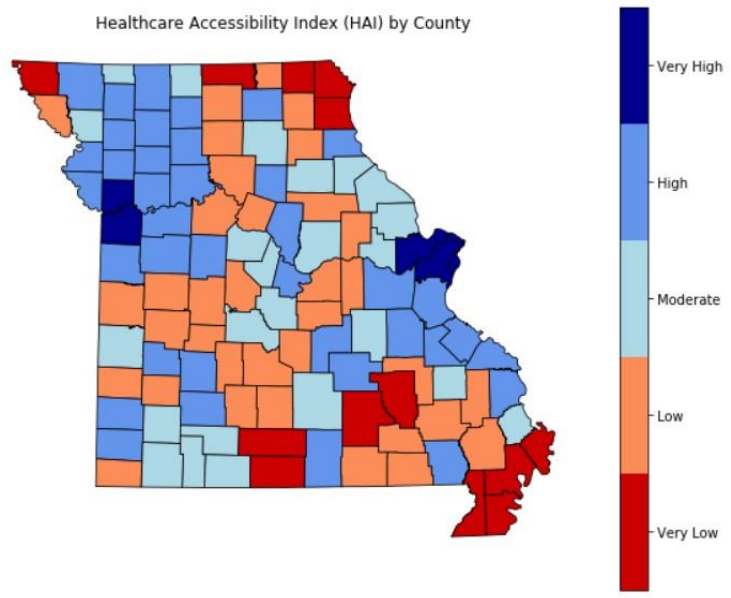

Figure 50: Healthcare Accessibility Index Map

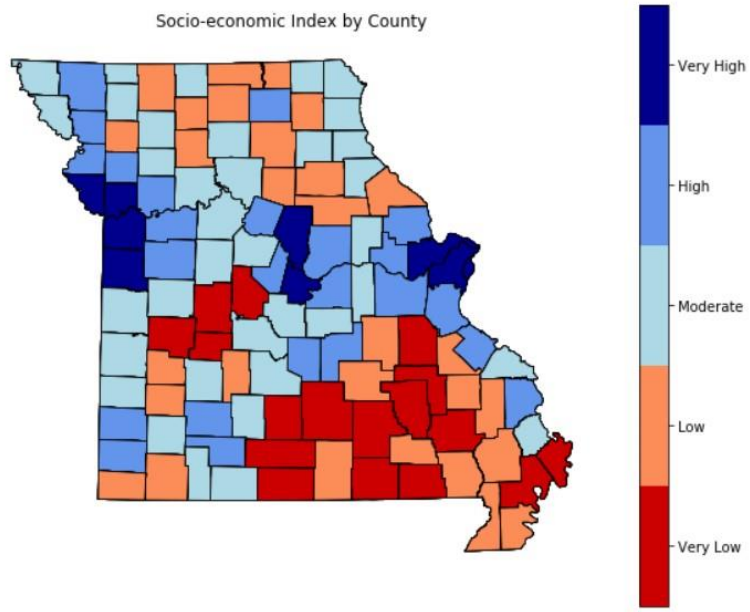

Figure 51: Socio-economic Index Map

Even when comparing the HAI map with Socioeconomic Index, we see that the Highs and Lows of each Index are correlated with each other respectively. This is mostly due to correlation with urbanization, which was inferred from earlier maps.

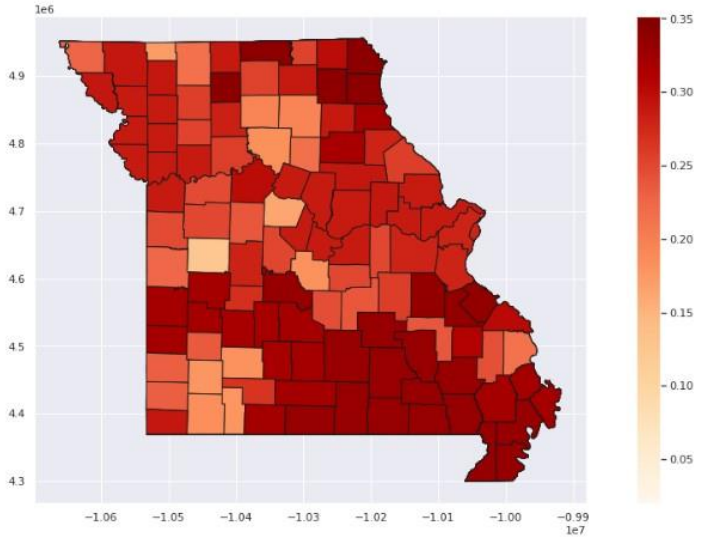

Figure 52: Socio-economic Vulnerability Index Map

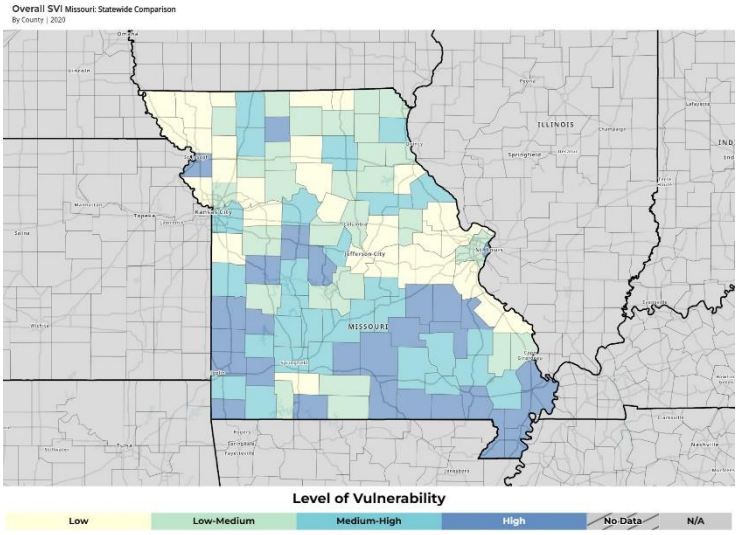

Figure 53: Social Vulnerability Index (SVI) Map

When the engineered Socioeconomic Vulnerability Index is compared with readily available Social Vulnerability Index (SVI), we can see that the engineered Socioeconomic Vulnerability Index has similarities with Social Vulnerability Index (SVI), which reinforces the credibility of the engineered feature.

Modelling using Indices

```
Random Forest Accuracy: 0.84
Classification Report for Random Forest:
              precision    recall  f1-score   support

          HH       0.86      0.81      0.83        93
          LL       0.83      0.88      0.85       100

    accuracy                           0.84       193
   macro avg       0.85      0.84      0.84       193
weighted avg       0.85      0.84      0.84       193


Feature Importances for Random Forest:
Socioeconomic_Vulnerability_Index    0.422631
NEW_HAI_RAW                          0.311870
Risk_for_Stroke_Index                0.265499
dtype: float64

Decision Tree Accuracy: 0.85
Classification Report for Decision Tree:
              precision    recall  f1-score   support

          HH       0.83      0.86      0.85        93
          LL       0.87      0.84      0.85       100

    accuracy                           0.85       193
   macro avg       0.85      0.85      0.85       193
weighted avg       0.85      0.85      0.85       193


Feature Importances for Decision Tree:
Socioeconomic_Vulnerability_Index    0.536395
NEW_HAI_RAW                          0.252095
Risk_for_Stroke_Index                0.211510
dtype: float64
```

Figure 54: Model with Indices Results

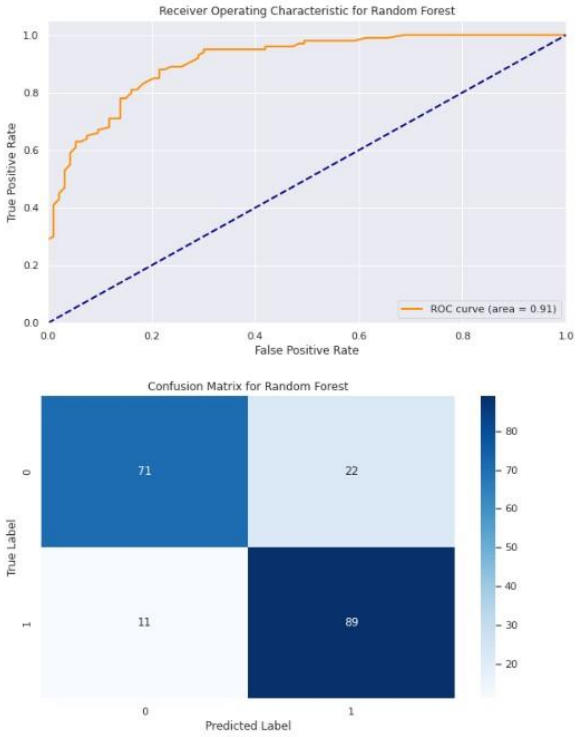

Figure 55: Model with Indices RF Results

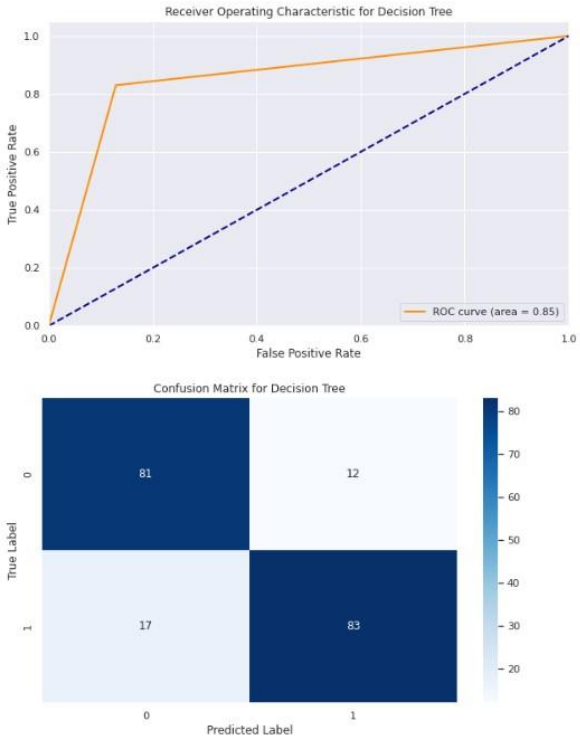

Figure 56: Model with Indices DT Results

```
Cross-validation scores: [0.84536082 0.86458333 0.91666667 0.92708333 0.875      0.89583333
 0.875      0.8125     0.76041667 0.70833333]
Mean accuracy: 0.8480777491408935
Standard deviation: 0.06590530501921066

RandomForestClassifier(random_state=43)
```

Figure 57: Model with Indices RF CV Results

We understand the importance/significance of each Index based on the Feature Importance scores derived by modelling these Indices with respect to Stroke Mortality Rate's High and Low Cluster Classes. By deriving the level of importance of each Index, we are getting an understanding of which concepts or factors are important in determining the Stroke Mortality Rates for a particular region.

The Feature Importance Scores of the Indices indicate that the Socio-economic vulnerability factor is the most influential factor in determining the Stroke Outcome. The second most important factor for determining Stroke Outcome is Healthcare Accessibility and is followed by Risk for Stroke in importance.

County Rankings based on Indices – Socioeconomic Index, Comorbidities Index, Healthcare Accessibility Index, in relation with Stroke Mortality

The Indices are fed into the model to get their Feature Importance Scores, using which a weighted sum based on these Index values is created, which is then binned based on K Bins Discretizer to create new groups/classes of counties. This binning strategy creates classes for counties which have varied levels of risks in losing Healthcare Accessibility.

The counties are classified and ranked as:

Very High Risk - Need New hospitals,

High Risk - Covered by hospitals from neighbour counties,

Moderate Risk - High Priority counties which have only have 1 hospital - Not to be closed,

Low Risk - Has adequate number of hospitals,

Very Low Risk - Has surplus amount of hospitals

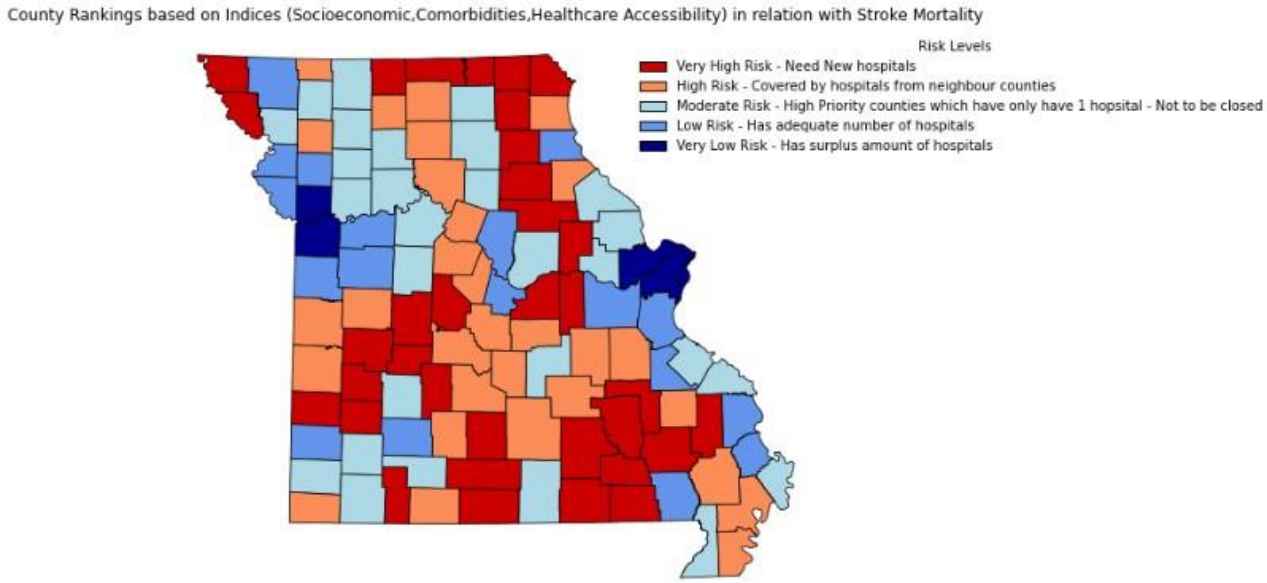

Figure 58: County Rankings based on engineered Indices in relation with Stroke Mortality

Regionalization

Once the Indices are engineered as new features, which each represent a concept/theme, Regionalization-Clustering algorithm is used to create new groups amongst the counties which all share consistent statistics or values for each Index. These new clusters/groups help us in creating more appropriate and targeted policy interventions, since it is easier to treat multiple counties at once, where all the members of these clusters are similar.

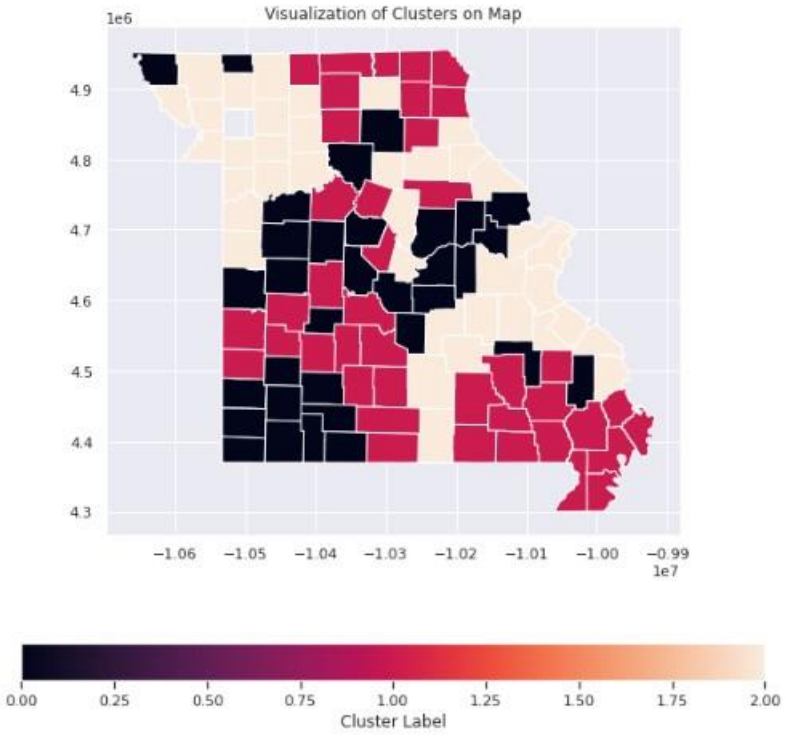

Figure 59: Regions – Groups with commonalities

```
                        Socioeconomic_Vulnerability_Index  \
INDICES_Cluster_Labels
0                                                0.222114
1                                                0.315347
2                                                0.282179

                        Risk_for_Stroke_Index  NEW_HAI_RAW
INDICES_Cluster_Labels
0                                    0.142975    64.166593
1                                    0.150374    50.843555
2                                    0.134814    80.355701
1    373
2    308
0    280
Name: Cluster_Size, dtype: int64
                        Socioeconomic_Vulnerability_Index  \
INDICES_Cluster_Labels
0                                                0.222114
1                                                0.315347
2                                                0.282179

                        Risk_for_Stroke_Index  NEW_HAI_RAW  Cluster_Size
INDICES_Cluster_Labels
0                                    0.142975    64.166593           280
1                                    0.150374    50.843555           373
2                                    0.134814    80.355701           308
```

Figure 60: Regions' Description

Socioeconomic Vulnerability Index:

'RUCC_1', 'RUCC_2', 'RUCC_3',

'Education_Attainment_Bachelors_Geolabels_HH',

'Education_Attainment_Bachelors_Geolabels_LL',

'Unemployment_Rate_%_Geolabels_HH', 'Unemployment_Rate_%_Geolabels_LL',

'Median_Income_Geolabels_HH', 'Median_Income_Geolabels_LL',

'Health_Insurance_WITH_%_Geolabels_HH',

'Health_Insurance_WITH_%_Geolabels_LL',

'Health_Insurance_WITHOUT_%_Geolabels_HH',

'Health_Insurance_WITHOUT_%_Geolabels_LL',

'Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_HH',

'Families_or_People_Below_Poverty_Level_in_past_12M_%_Geolabels_LL'

Risk for Stroke Index:

'Smokers_%_Geolabels_HH', 'Smokers_%_Geolabels_LL',

'Fair/Poor_health_%_Geolabels_HH', 'Fair/Poor_health_%_Geolabels_LL',

'Violent_Crime_Rate_Geolabels_HH', 'Violent_Crime_Rate_Geolabels_LL',

'Obesity_%_Geolabels_HH', 'Obesity_%_Geolabels_LL',

'Diabetes_%_Geolabels_HH', 'Diabetes_%_Geolabels_LL',

'Physical_Inactivity_%_Geolabels_HH',

'Physical_Inactivity_%_Geolabels_LL'

Since, we know what groups are formed based on the clustering of the Indices, and their commonly shared statistics/value; and know which variables make up these Indices along with

their importances, we can pinpoint the areas of interest and devise appropriate policy interventions.

Projection of Risk Groups

Once the factors indicative of being influencing factors for stroke mortality are ascertained, we note these factors or variables to monitor their trends and project the risks they pose. Based on these projected risks, we can even predict the risk for stroke mortality.

Data Story

When it comes to healthcare, every possible way to improve both the lifespan and the livelihood of the general population is one that needs to be explored. In our project's case, we looked at the effect healthcare accessibility has on stroke mortality and found some interesting things. It has already been shown on the scale of countries that an increase of accessibility to healthcare institutions decreases stroke mortality when everything else is equal, so we want to show that on a smaller scale, looking specifically at counties.

By initially taking data from government sources, primarily the Centers for Disease Control, as well as road network maps and the addresses of Stroke Care Centers, we were able to construct a service area map of the state of Missouri, with which we could compare socio-economic and stroke mortality maps. Using these, we were able to construct indices related to the location of each issue and use them in the creation of our risk models.

Once we had our indices, we were able to do some real magic. Putting all the different variables - e.g. percentage of population of a county covered by a stroke care center, percentage of a county having diabetes, etc. – into a machine learning model, we were able to create a rudimentary relationship between stroke care access and stroke mortality.

Reflection

The scope of this project kept growing and growing, to the extent that we as a group lost sight of the initial goal near the middle of the semester. This led to a scramble at the end as we attempted

to get our models situated, and some aspects were forgotten about in this mad scramble. Issues began to crop up, and we began to look for a “silver bullet” solution to stroke mortality, which was outside the initial scope.

However, the envisioned product and outcomes from our Capstone Project were achieved to a decent degree, although intended timeline was deterred due to a few setbacks. As part of one of our deliverables, we wanted to do Hospitals Closure Impact Analysis, but couldn’t come out with fruitful results in that analytical arm, although efforts were made. If we had to redo this project to match all our envisioned deliverables – which includes Hospitals Closure Impact Analysis, we would be doing our geo-spatial analysis on a geo-database; which would allow us to create more comprehensive data-frames, which subsequently would allow us to complete our Hospitals Closure Impact Analysis.

Why is this Data Science?

Collecting several datasets from reliable sources, then fine-tuning the information to guarantee its accuracy and usefulness. Analyzing the data to find underlying trends and linkages, then creating indices to measure stroke risk and socioeconomic susceptibility to improve the dataset's predictive capacity. Using Random Forest classifiers—a type of machine learning technique—to examine how different factors affect the risk of stroke. This includes weighting input variables in the indices with feature relevance scores that come from the models. By the use of Geographic Information Systems (GIS), geographical data may be visualized and analyzed to reveal geographic differences in healthcare access. By using service area analysis, it is possible to determine how easily accessible stroke care is in various parts of Missouri. Putting the Indices into useful groups using this statistical technique, which will help with the application and interpretation of the study results. In order to promote policy-making for better healthcare infrastructure and to highlight locations with essential requirements for healthcare services, the results will be synthesized and communicated through intricate maps and data visualizations.

Looking at unique solutions to problems and using the data that is created for a multitude of reasons together in a unique way` is quintessential data science.

The fact that data can be transformed multiple times using multiple different methods, and be interpreted differently in each state, encapsulates data science.

In our case,

Raw data > Interaction Terms > Features > Indices > Clusters/Groups

Same data, but different forms and interactions.

References

Adeoye, O., Albright, K. C., Carr, B. G., Wolff, C., Mullen, M. T., Abruzzo, T., Ringer, A., Khatri, P., Branas, C., & Kleindorfer, D. (2014). Geographic access to acute stroke care in the United States. Stroke, 45(10), 3019-3024. https://doi.org/10.1161/STROKEAHA.114.006293

Awlachew Dejen, Sandeep Soni, Fisha Semaw, Spatial accessibility analysis of healthcare service centers in Gamo Gofa Zone, Ethiopia through Geospatial technique, Remote Sensing Applications: Society and Environment, Volume 13, 2019, Pages 466-473, ISSN 2352-9385, https://doi.org/10.1016/j.rsase.2019.01.004.

Bagheri, N., Holt, A., & Benwell, G. L. (2009). Using Geographically Weighted Regression to Validate Approaches for Modelling Accessibility to Primary Health Care. Applied Spatial Analysis, 2, 177–194. https://doi.org/10.1007/s12061-009-9021-0

Busingye, D., Pedigo, A., & Odoi, A. (2011). Temporal changes in geographic disparities in access to emergency heart attack and stroke care: are we any better today? Spatial and Spatiotemporal Epidemiology, 2(4), 247-263. https://doi.org/10.1016/j.sste.2011.07.010

Chaohui Yin, Qingsong He, Yanfang Liu, Weiqiang Chen, Yuan Gao. (2018). Inequality of public health and its role in spatial accessibility to medical facilities in China. Applied Geography, 92, 50-62. https://doi.org/10.1016/j.apgeog.2018.01.011

Comber, A. J., Brunsdon, C., & Radburn, R. (2011). A spatial analysis of variations in health access: linking geography, socio-economic status, and access perceptions. International Journal of Health Geographics, 10(44). https://doi.org/10.1186/1476-072X-10-44

DeVoe, J. E., Baez, A., Angier, H., Krois, L., Edlund, C., & Carney, P. A. (2007). Insurance + Access ≠ Health Care: Typology of Barriers to Health Care Access for Low-Income Families. Annals of Family Medicine, 5(6), 511-518. https://doi.org/10.1370/afm.748

El Khoury, R., Jung, R., Nanda, A., Sila, C., Abraham, M. G., Castonguay, A. C., & Zaidat, O. O. (2012). Overview of Key Factors in Improving Access to Acute Stroke Care. Neurology, 79(Suppl 1), S26–S34. https://doi.org/10.1212/WNL.0b013e3182695a2a

Freyssenge, J., Renard, F., Schott, A.M. *et al.* Measurement of the potential geographic accessibility from call to definitive care for patient with acute stroke. *Int J Health Geogr* **17**, 1 (2018). https://doi.org/10.1186/s12942-018-0121-4

Fujiwara, K., Osanai, T., Kobayashi, E., Tanikawa, T., Kazumata, K., Tokairin, K., Houkin, K., & Ogasawara, K. (2018). Accessibility to Tertiary Stroke Centers in Hokkaido, Japan: Use of Novel Metrics to Assess Acute Stroke Care Quality. Journal of Stroke and Cerebrovascular Diseases, 27(1), 177-184. https://doi.org/10.1016/j.jstrokecerebrovasdis.2017.08.013

Gao, F., Languille, C., Karzazi, K., et al. (2021). Efficiency of fine scale and spatial regression in modelling associations between healthcare service spatial accessibility and their utilization. International Journal of Health Geographics, 20, 22. https://doi.org/10.1186/s12942-021-00276-y

Garber, K., Fox, C., Abdalla, M., Tatem, A., Qirbi, N., Lloyd-Braff, L., Al-Shabi, K., Ongwae, K., Dyson, M., & Hassen, K. (2020). Estimating access to health care in Yemen, a complex humanitarian emergency setting: a descriptive applied geospatial analysis. Lancet Global Health, 8(11), e1435-e1443. https://doi.org/10.1016/S2214-109X(20)30359-4

Gonzales, S., Mullen, M. T., Skolarus, L., Thibault, D. P., Udoeyo, U., & Willis, A. W. (2017). Progressive rural-urban disparity in acute stroke care. Neurology, 88(5), 441-448. https://doi.org/10.1212/WNL.0000000000003562

Hammond, G., Luke, A. A., Elson, L., Towfighi, A., & Joynt Maddox, K. E. (2020). Urban-Rural Inequities in Acute Stroke Care and In-Hospital Mortality. Stroke, 51(2131-2138). https://doi.org/10.1161/STROKEAHA.120.029318

Ingall T. (2004). Stroke--incidence, mortality, morbidity and risk. *Journal of insurance medicine (New York, N.Y.)*, *36*(2), 143–152.

Kapral, M. K., Hall, R., Gozdyra, P., Yu, A. Y. X., Jin, A. Y., Martin, C., Silver, F. L., Swartz, R. H., Manuel, D. G., Fang, J., Porter, J., Koifman, J., & Austin, P. C. (2020). Geographic Access to Stroke Care Services in Rural Communities in Ontario, Canada. Canadian Journal of Neurological Sciences, 47(3), 301-308. https://doi.org/10.1017/cjn.2020.9

Lawal, Olanrewaju & Anyiam, Felix. (2019). Modelling Geographic Accessibility to Primary Health Care Facilities: Combining Open Data and Geospatial analysis. Geo-spatial Information Science. 1-11. https://doi.org/10.1080/10095020.2019.1645508

Menon, S. C., Pandey, D. K., & Morgenstern, L. B. (1998). Critical factors determining access to acute stroke care. Neurology, 51(2), 427-432. https://doi.org/10.1212/WNL.51.2.427

Mullen, M. T., Wiebe, D. J., Bowman, A., Wolff, C. S., Albright, K. C., Roy, J., Balcer, L. J., Branas, C. C., & Carr, B. G. (2014). Disparities in accessibility of certified primary stroke centers. Stroke, 45(11), 3381-3388. https://doi.org/10.1161/STROKEAHA.114.006021

O'Brien, E. C., Wu, J., Zhao, X., Schulte, P. J., Fonarow, G. C., Hernandez, A. F., Schwamm, L. H., Peterson, E. D., Bhatt, D. L., & Smith, E. E. (2017). Healthcare Resource Availability, Quality of Care, and Acute Ischemic Stroke Outcomes. *Journal of the American Heart Association*, *6*(2), e003813. https://doi.org/10.1161/JAHA.116.003813

Pandian, Jeyaraj & Singh, Gagandeep & Kaur, Paramdeep & Bansal, Rajinder & Paul, Birinder & Singla, Monika & Singh, Shavinder & Samuel, Clarence & Verma, Shweta & Moodbidri, Premjeeth & Mehmi, Gagandeep & Sharma, Amber & Arora, Om & Sobti, Manoj & Sehgal, Harish & Kaur, Mohanjeet & Grewal, Sarvpreet & Jhawar, Sukhdeep & Sharma, Meenakshi. (2016). Incidence, short-term outcome, and spatial distribution of stroke patients in Ludhiana, India. Neurology. 86. 10.1212/WNL.0000000000002335.

Pedigo, A. S., & Odoi, A. (2010). Investigation of Disparities in Geographic Accessibility to Emergency Stroke and Myocardial Infarction Care in East Tennessee Using Geographic Information Systems and Network Analysis. Annals of Epidemiology, 20, 924–930. https://doi.org/10.1016/j.annepidem.2010.06.013

Pindus, D. M., Mullis, R., Lim, L., Wellwood, I., Rundell, A. V., Aziz, N. A. A., & Mant, J. (2018). Stroke survivors' and informal caregivers' experiences of primary care and community healthcare services – A systematic review and meta-ethnography. PLoS ONE, 13(2), e0192533. https://doi.org/10.1371/journal.pone.0192533

Pradilla, Ivan & Macea-Ortiz, Jaiver & Polo-Pantoja, Paola & Palacios-Ariza, María & Díaz, Andrés & Velasquez-Torres, Alejandro. (2020). Spatial Analysis of Service Areas for Stroke Centers in a City with High Traffic Congestion. Spatial and Spatio-temporal Epidemiology. 35. 100377. 10.1016/j.sste.2020.100377.

R. Shanmathi Rekha, Shayesta Wajid, Nisha Radhakrishnan, Samson Mathew. (2017). Accessibility Analysis of Health care facility using Geospatial Techniques. Transportation Research Procedia, 27, 1163-1170. https://doi.org/10.1016/j.trpro.2017.12.078

Rahbar, M. H., Medrano, M., Diaz-Garelli, F., Gonzalez Villaman, C., Saroukhani, S., Kim, S., Tahanan, A., Franco, Y., Castro-Tejada, G., Diaz, S. A., Hessabi, M., & Savitz, S. I. (2022). Younger age of stroke in low-middle income countries is related to healthcare access and quality. *Annals of clinical and translational neurology*, *9*(3), 415–427. https://doi.org/10.1002/acn3.51507

Shoff, C., Yang, T. C., & Matthews, S. A. (2012). What has geography got to do with it? Using GWR to explore place-specific associations with prenatal care utilization. GeoJournal, 77(3), 331-341. https://doi.org/10.1007/s10708-010-9405-3

Zachrison, K. S., Onnela, J.-P., Reeves, M. J., Hernandez, A. F., Camargo Jr, C. A., & Matsouaka, R. A. (2022). Estimated Population Access to Acute Stroke and Telestroke Centers in the US 2019. JAMA Network Open, 5(3), e2145824. https://doi.org/10.1001/jamanetworkopen.2021.45824