Project Technical Report

Pushing P Breakers

Group 1

Parth, Ahad, Luis, Sharif, Marjea

New Jersey's Uninsured

Introduction

The economic downturn caused by the coronavirus pandemic has renewed attention on health insurance coverage as millions have lost their jobs and potentially their health coverage. The Affordable Care Act (ACA) sought to address the gaps in our healthcare system that left millions of people without health insurance by extending Medicaid coverage to many low-income individuals and providing subsidies for marketplace coverage for individuals below 400% of poverty. Following the ACA, uninsured non-elderly Americans declined by 20 million, dropping to a historic low in 2016. However, beginning in 2017, the number of uninsured non-elderly Americans increased for three straight years, growing by 2.2 million from 26.7 million in 2016 to 28.9 million in 2019. The uninsured rate increased from 10.0% in 2016 to 10.9% in 2019.

With the rising uninsured population, it becomes essential to keep track of those changes and be able to map a specific number of uninsured individuals to a specific geographic area. This importance is dictated by the need for private insurance entities to target particular groups for their market segmentation and for public insurance entities to focus their programs and policies that aim to reduce and eliminate the number of uninsured in the target area. While that type of information is available on the web and other resources, it may not necessarily be well organized, compact, and analyzed to be readily available and consumed by those entities. Therefore, the ultimate **objective** of this research work is to locate resources, sort and organize information into a database, analyze and provide predictions regarding the uninsured population within a specific geographic area and define a typical uninsured group given a set of variables.

There may also be recommendations regarding what areas and groups those entities need to zero in on.

Exploratory Questions

To conduct this type of research work, we developed some questions and hypotheses that guided us through our work, helping to stay focused on our subject:

1. Which city has the highest uninsurance rate?
2. How many counties have a population of uninsured people of 8% or more?
3. How does income change the amount of people being uninsured?
4. Is there a race that has a higher uninsured population?
5. Which sex has a higher population of uninsured people?
6. What age range has the most uninsured people?
7. Does employment affect the amount of the uninsured population?
8. Hypothesis test: the larger the population, the higher the uninsured rate.
9. What area and demographic group can Prudential target for insurance sales in NJ?

Dataset Introduction

We used several resources and datasets in this research, including census data. Below is the list of those datasets:

1. Small Area Health Insurance Estimates 2019 (SAHIE)
    a. This dataset contains demographic information about health insurance coverage and demographics in the counties by a single year in the United States.
2. NJ Uninsured
    a. Contains data about the uninsured population in New Jersey broken down by age, race, and sex down to the county and city location
3. NJ Unemployed
    a. This dataset has information about the unemployment rate in each county and city location in New Jersey
4. NJ Income

a. Shows the median household income in every county and city in New Jersey

5. Cartographic Boundary Files
    a. Contains files to show county boundaries for selected geographic areas
6. Coverage for the Household Population by States
    a. Shows the populations for the state of New Jersey
7. Unemployment by Counties
    a. Gives a table that shows the characteristics of the unemployment population in each county in New Jersey
8. Census Tract in NJ
    a. Shows the boundaries for the county borders in New Jersey

Research Process and Discoveries

We downloaded all datasets as CSV files and stored them in our database storage account. After conducting an exploration of our datasets, we did some ETL to get all necessary variables into one SQL database using Azure databricks and data factory and set up a pipeline. We also used ML algorithms to conduct our predictions and presented our findings through visualizations using the Dash platform.

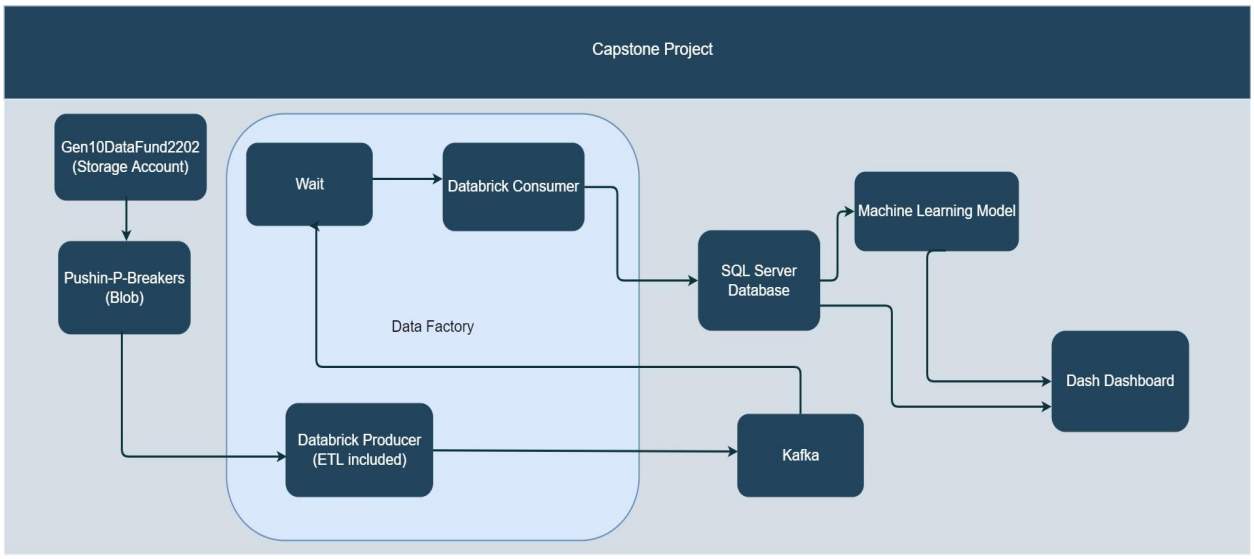


Figure 1. Data Platform

After having all our necessary data ready in our database we did the following analysis and discovered interesting results:

Regarding Q1.

One of the first criteria that assist in understanding the problem of uninsurance is to identify the geography of uninsured residents on several levels. We start by locating the top 10 cities with the highest number of uninsured residents:

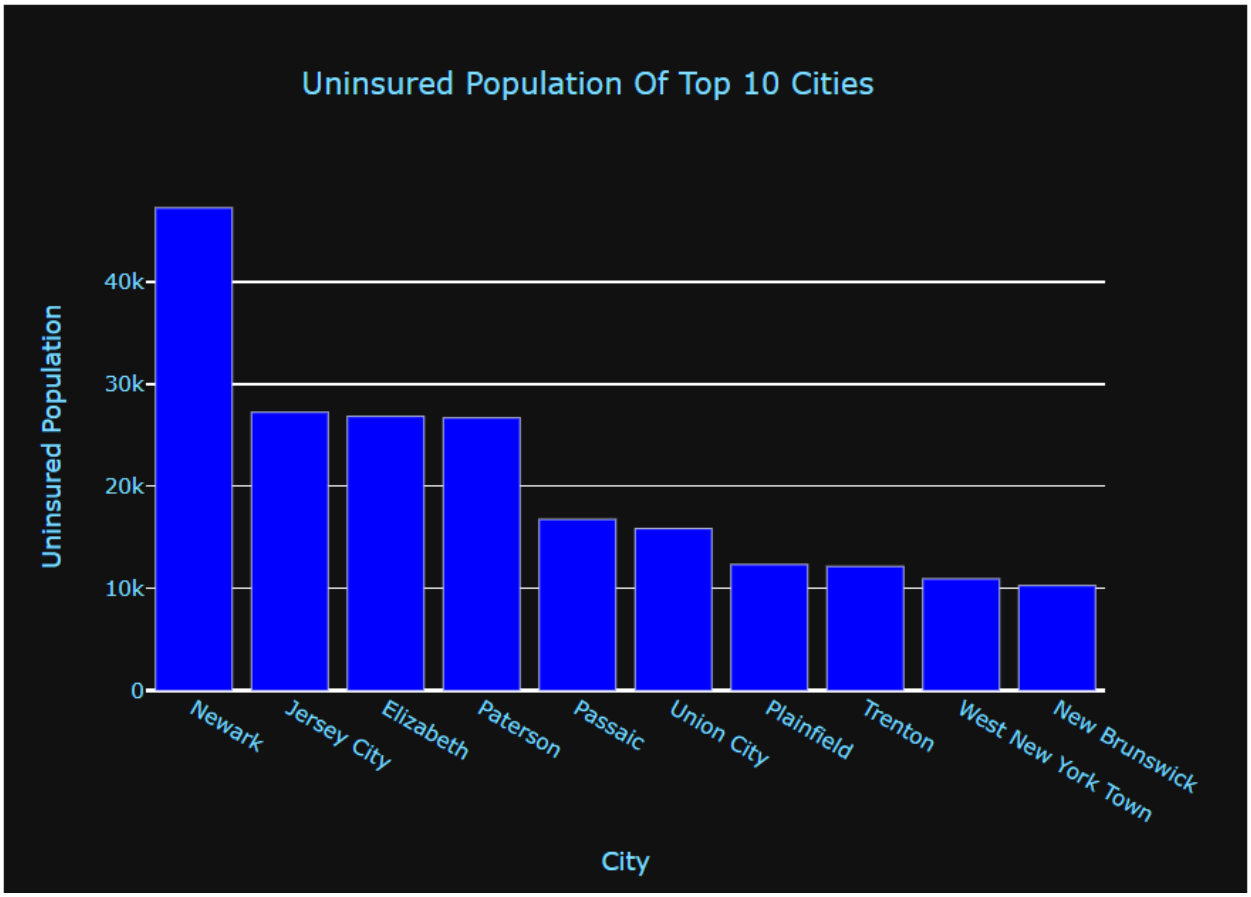


Figure 2. Top 10 cities of NJ with the highest number of uninsured residents

As seen from *Figure 2* almost all cities are located within the north-eastern counties of NJ with Newark having almost double the number of uninsured residents compared to its closest peer cities. This can be explained by two factors: a) most of those cities have the closest proximity to New York City, and b) because of the proximity to New York City, they are among the highest-density cities in NJ. We can safely make our first observation which is that the

number of uninsured residents of an area is impacted by how dense the population is within that area.

Another geographic level to look into is counties. During our research, we discovered that the national average rate for those uninsured is 8%. Thus, we wanted to identify counties that fall above the national average:

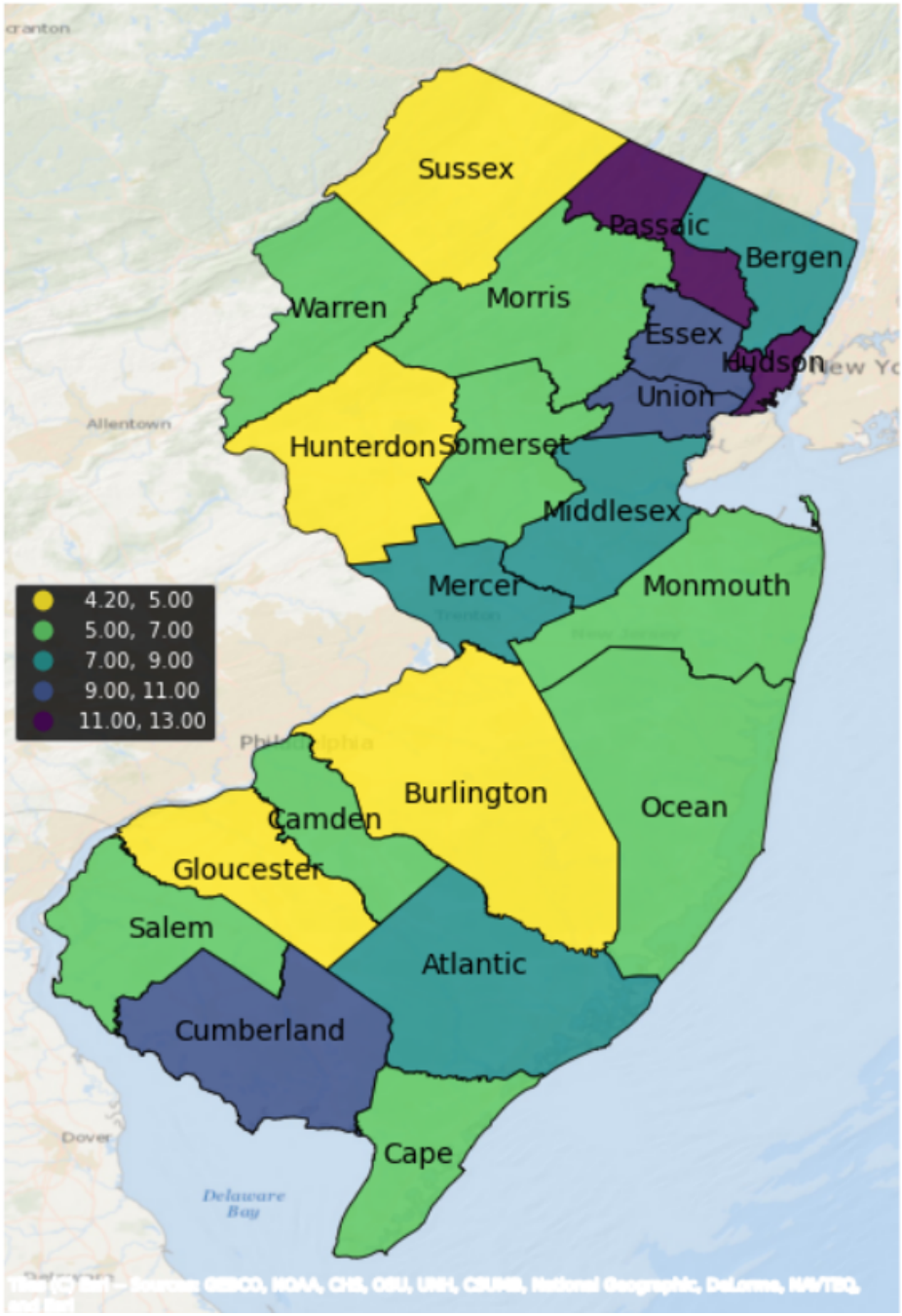


Figure 3. Counties of NJ broken by uninsurance rate

The first observation in *Figure 3* is that counties that are closer to the center of metro areas have the highest rate of uninsured residents. Primarily, northeastern counties like Hudson, Passaic, Essex, and Union are among the top 5 counties with the highest y insurance rate, and at the same time, the closest counties to the centers of the metro area. On the other side of the map, in the south and south-eastern part of the state, we see the counties Cumberland and Atlantic have uninsurance rates above the national average. These counties are also located closest to the center of the metro area, in this case, Philadelphia. These counties, primarily those located in the northeastern part of NJ are some of the densest counties in the state. This observation strongly supports our findings in question 1.

Another criterion we decided to look into is income. The rational assumption would have it that the higher the income level of a geographic area, the lower the uninsurance rate, as residents with better income and jobs would be able to get coverage through their employer's plan or buy one on their own. To check this relationship we created the following figure:

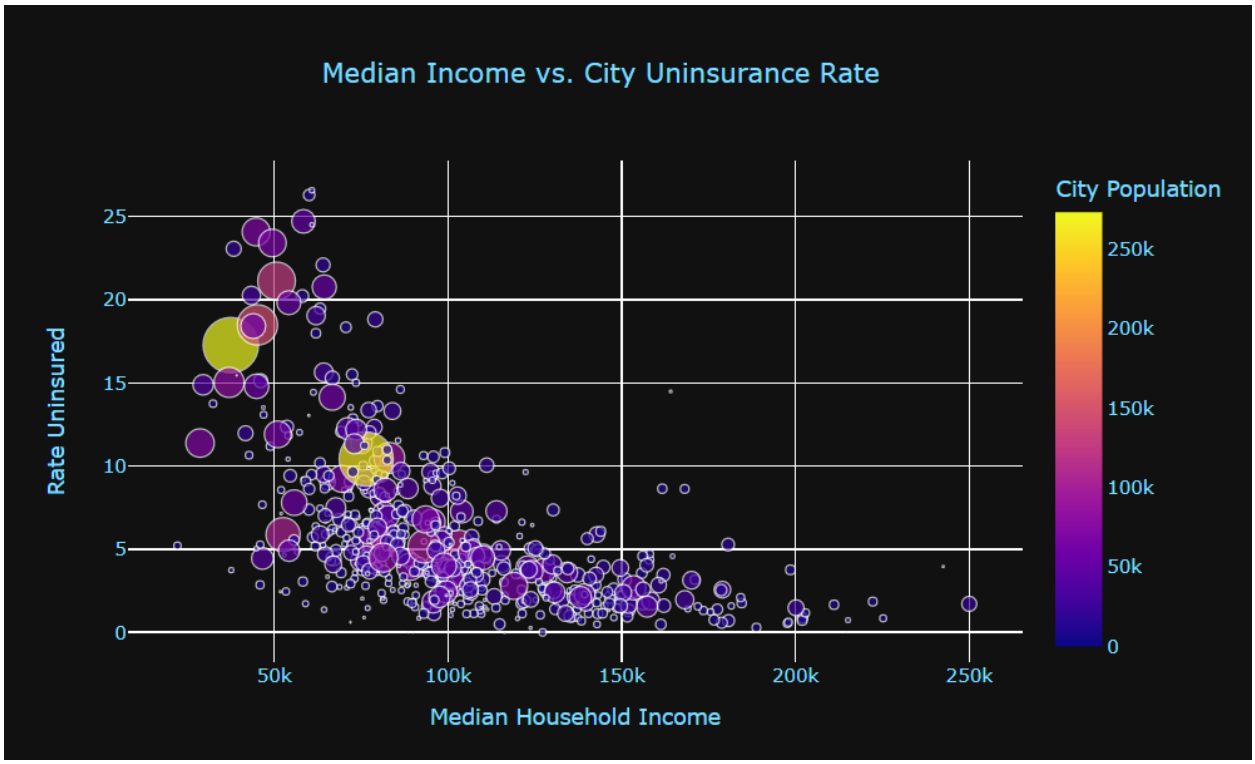


Figure 4. Scatter plot of the relationship between income and uninsurance rate of different cities (city population sizes are provided for reference to previous questions)

The first observation from *Figure 4* is that there is a negative correlation between income and the uninsurance rate of cities. The higher the income the lower the rate tends to be, which to a degree confirms the assumption we had. However, this relationship only proves itself with income starting at around $40,000. How about residents with incomes lower than $40,000? Well, as we all know, families and individuals who live below the federal poverty line can count on state coverage. This is why families and individual residents of cities that fall within those federal poverty lines do not really show as uninsured in *Figure 4*. Lastly, it seems higher populated cities tend to have higher uninsurance rates. This also can be explained by the geographically close proximity of those cities to centers of metro areas and their population density.

Among the demographic characteristics of the uninsured population in NJ, race, age, and gender stand out. Are the uninsured groups evenly distributed among the population or is there a significant difference? If so, what causes it? We start by looking into the race category:

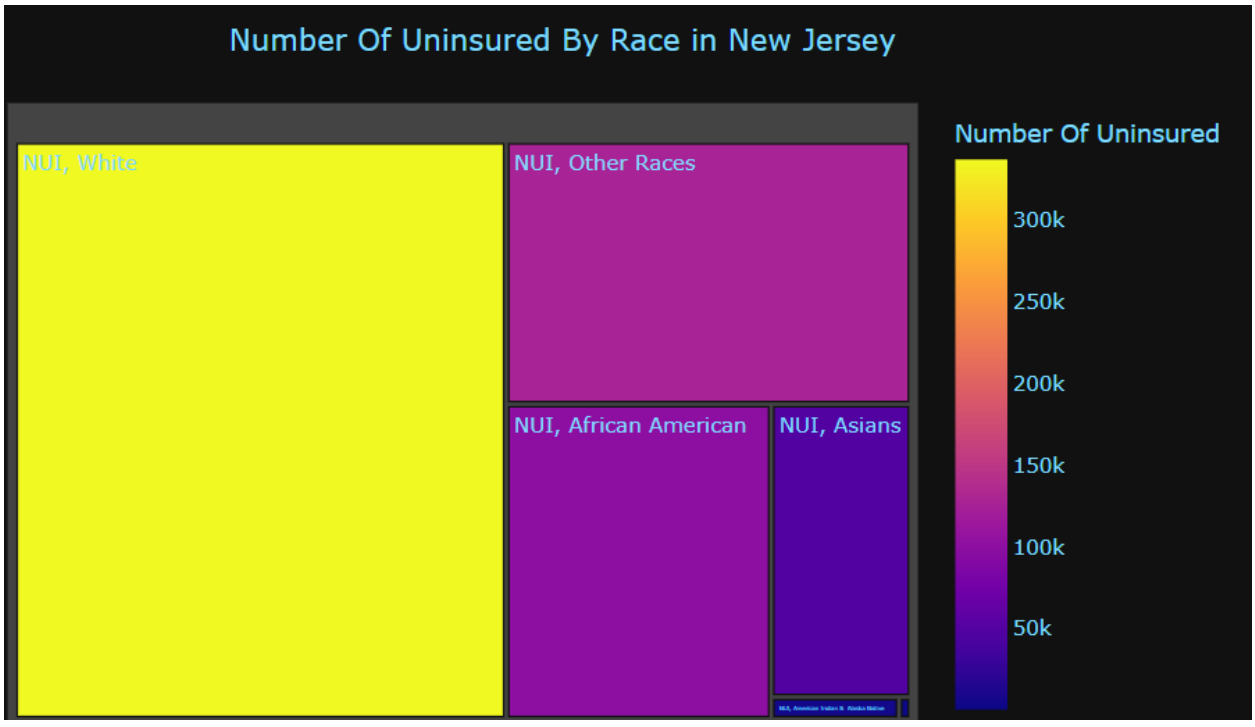


Figure 5. Uninsured population by race

After a quick glimpse over *Figure 5,* we can conclude that the majority of the uninsured population in NJ is white. However, this type of distribution is not primarily dependent on the

uninsurance factor, but rather on the race distribution of the total state population. In NJ, over 70% of the population is white, followed by about 15% being African-American, 10% Asian and the remaining being a mix of other races. *Figure 5* roughly represents the same distribution.

Another characteristic is uninsured population distribution by gender:

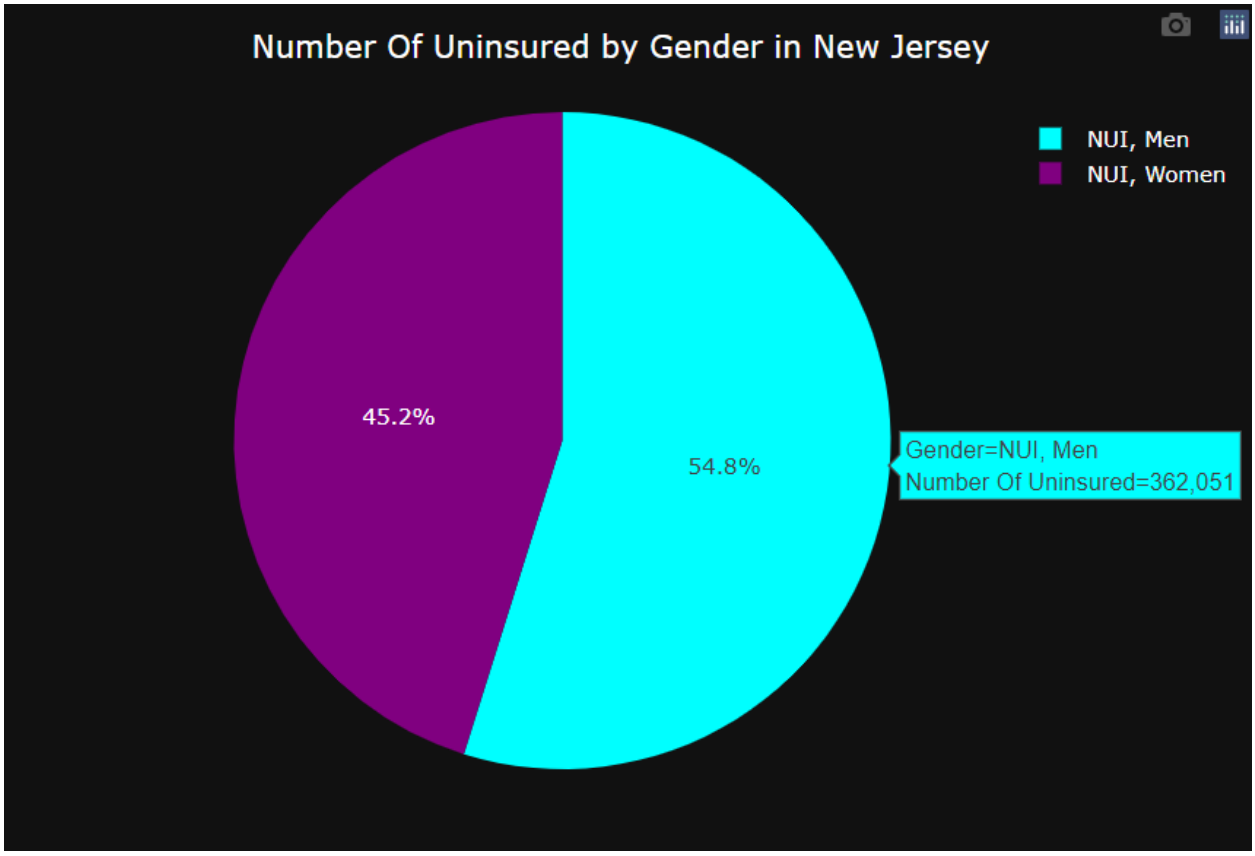


Figure 6. Uninsured population by gender

While the total population of females in NJ is higher than males, here we realize that within the uninsured population the number of males that are uninsured is slightly higher than females. One of the reasons the results are as such is that in NJ, female adults have higher chances of getting state coverage compared to same-age male adults. This is primarily because of pregnancy conditions. Another reason is that some families with tight incomes may choose to have only a part of the family, predominantly females, have coverage, leaving male members with little to no coverage.

Age is also another important characteristic to look into, since the uninsured population may be impacted by age and health conditions age ranges may cause:

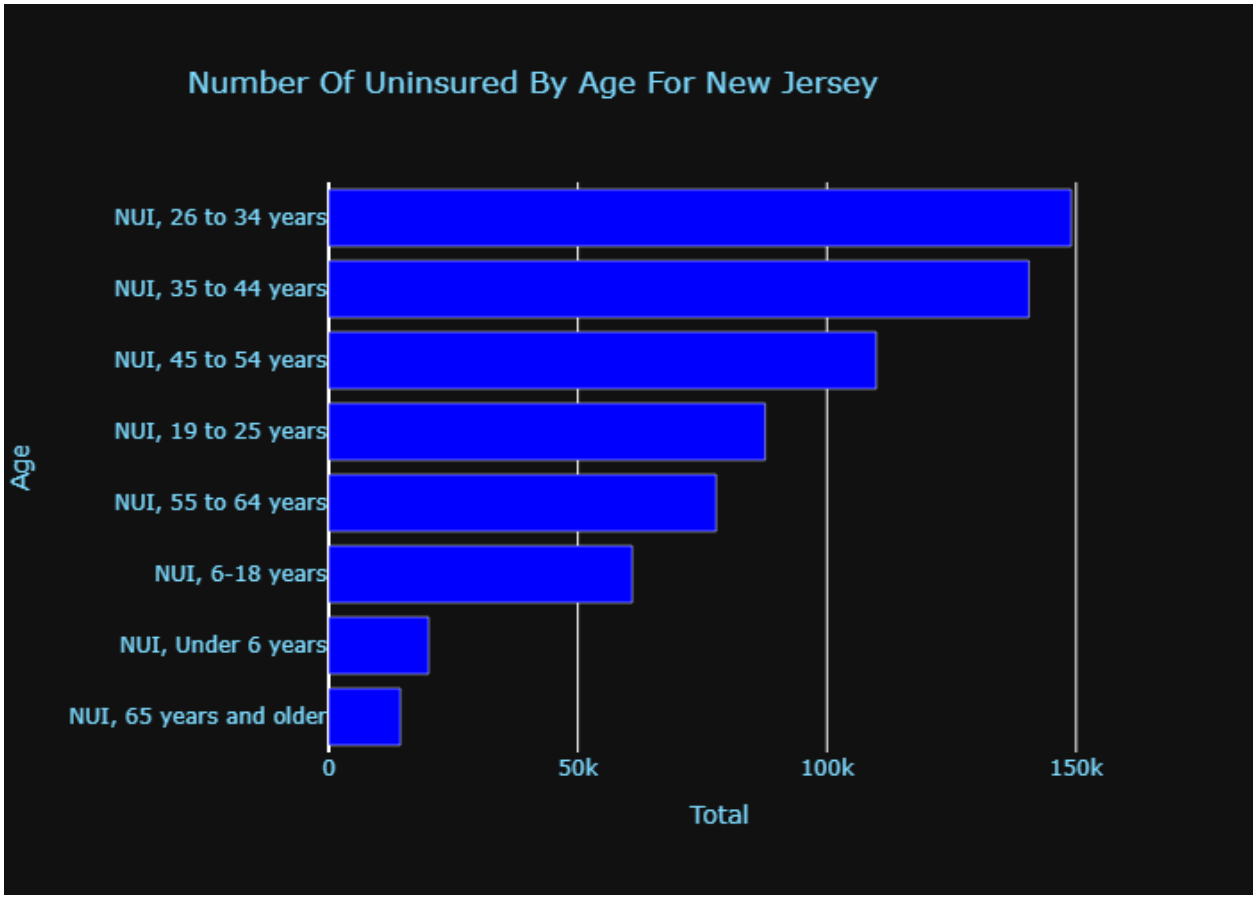


Figure 7. Uninsured population by age group

Population groups between the ages of 26-54 are more likely to be uninsured compared to the other age groups. When individuals are young, they are likely to be covered under their parents' coverage or state coverage. Once they reach 26, they are no longer eligible to be dependent on parent coverage, nor are they usually eligible for state coverage. This is a period when many face the challenge of affording coverage and there are not many options for low-income residents. Once they reach the age of 55 and over, they become Medicare eligible, which by this age, increases the probability of an individual being insured.

Lastly, we considered the employment state of the uninsured population. Our rational assumption was that within the uninsured population, the majority would be unemployed. To check this hypothesis we created *Figure 8*:

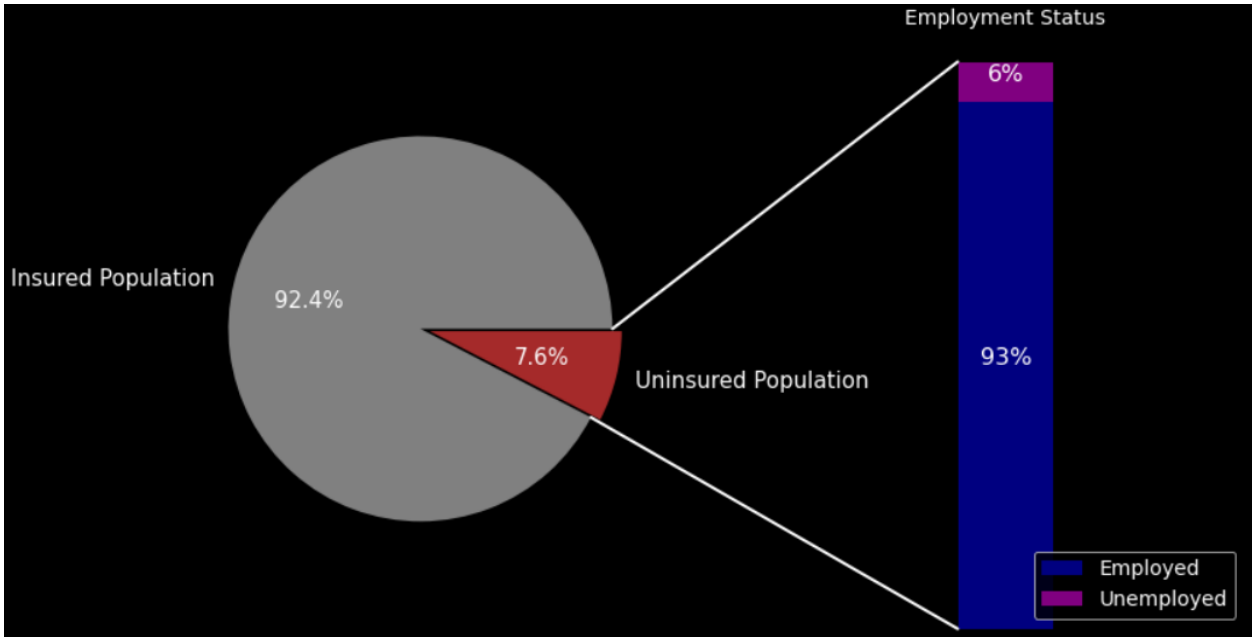


Figure 8. Uninsured population by employment status.

As seen in *Figure 8* within the uninsured population over 90% are employed individuals, and only 6% are unemployed. This is very interesting, not only because it rejects our rational assumption, but it gives insight into how employer plans lack affordability for even their own employees. The employed category also represents business owners and employees of small companies that don't offer coverage at all. However, the majority represent employees of companies that do offer some type of coverage.

Conclusion

After our research and analysis, we built a list of evidence-based findings from the uninsured population in NJ:

1. **Close proximity to the centers of metro areas (location):** The closer the location to the centers of metro areas, the higher the number of uninsured.

2. **Population density:** The higher the population density the higher number of uninsured.
3. **Income level:** Generally, higher income leads to lower uninsurance rates.
4. **Race:** The majority of the uninsured population is white, but that reflects the proportion of the white population from the total state population.
5. **Gender:** Males represent a slightly higher uninsured population than females.
6. **Age:** Those within the age range of 26-54 are at more risk of being uninsured compared to younger or older age ranges.
7. **Employment status:** Employment status doesn't seem to have a significant influence on the number of uninsured, where over 93% are employed.

So, who is a typical uninsured person in New Jersey? **This is an individual that tends to be a white male 26-54 years old who is employed and with an income of around $40,000.**

What area do uninsured individuals tend to live in? **Close to the centers of metro areas and/or urban areas, and within high-density populations.**

Machine learning

With the goal of identifying the uninsured population within a given geographic area, we are using the following Machine Learning Algorithm to predict the goal:

Spatial Interpolation

Spatial interpolation is a technique for estimating values in unknown locations that employ geocoded sample points with values. One of the most efficient ways to map unknown values at unsampled places is to use the geostatistical approach. It makes highly accurate predictions. The process of creating estimates from a source set of polygons to an overlapping but incongruent set of destination polygons is known as areal interpolation. A basic areal interpolation is the most straightforward approach accessible. Variations from the source data are weighted based on their overlap with the target polygons, then reaggregated to match the target polygon geometries.

“Unemployment Rate (16 and Over)” overlaps and makes extensive use of data. We utilized both African Americans and Caucasians. African Americans have a smaller dataset than whites, who have the greatest. The accuracy of the larger dataset is higher.

After building the model, we tried to predict the uninsurance rate of NJ census geographical tract areas and compared our results with the actual numbers:

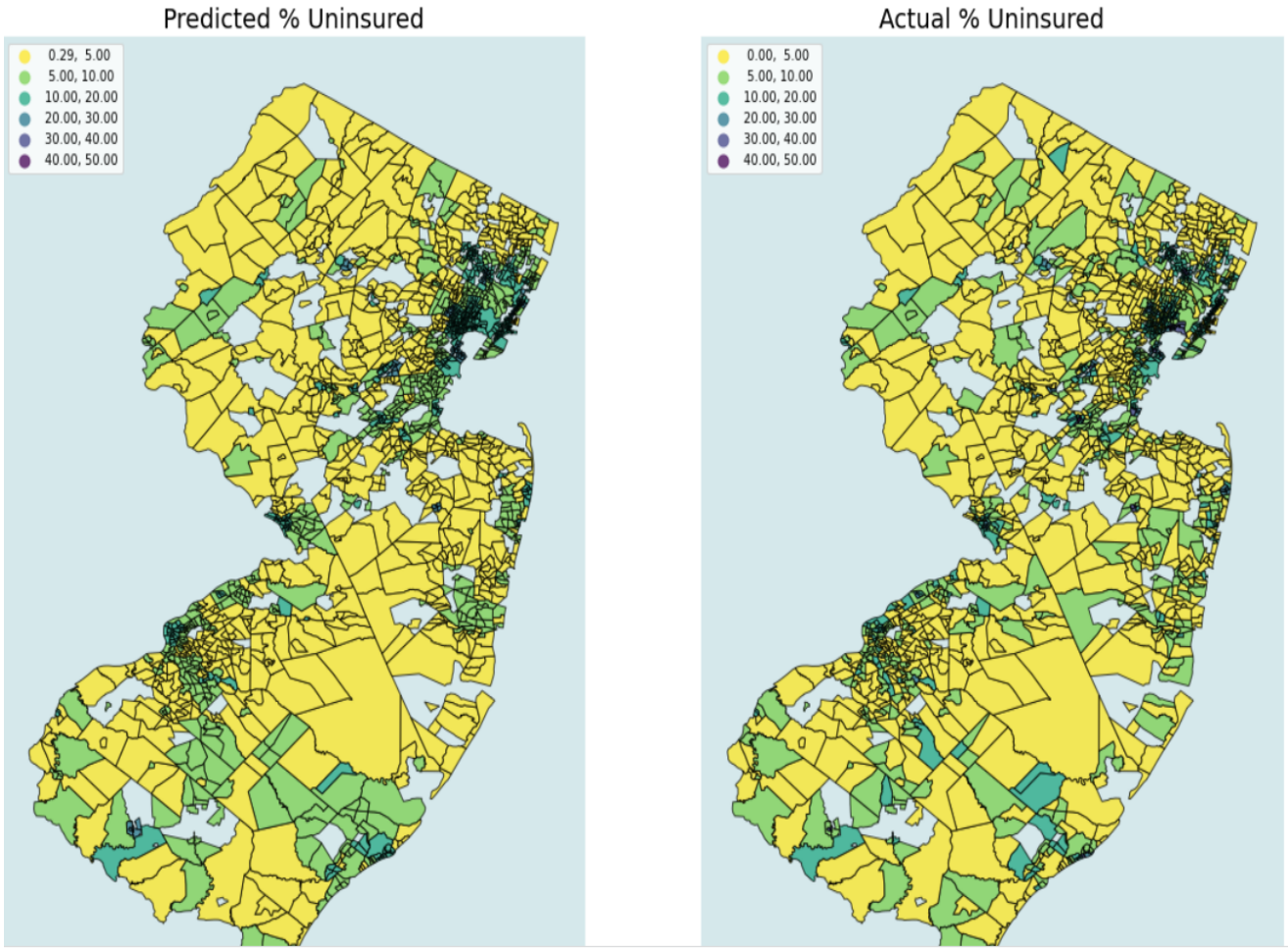
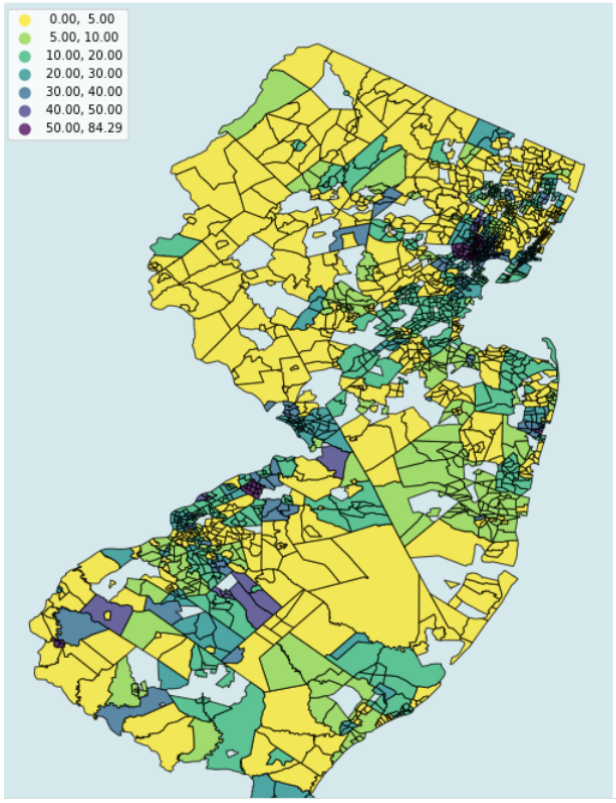


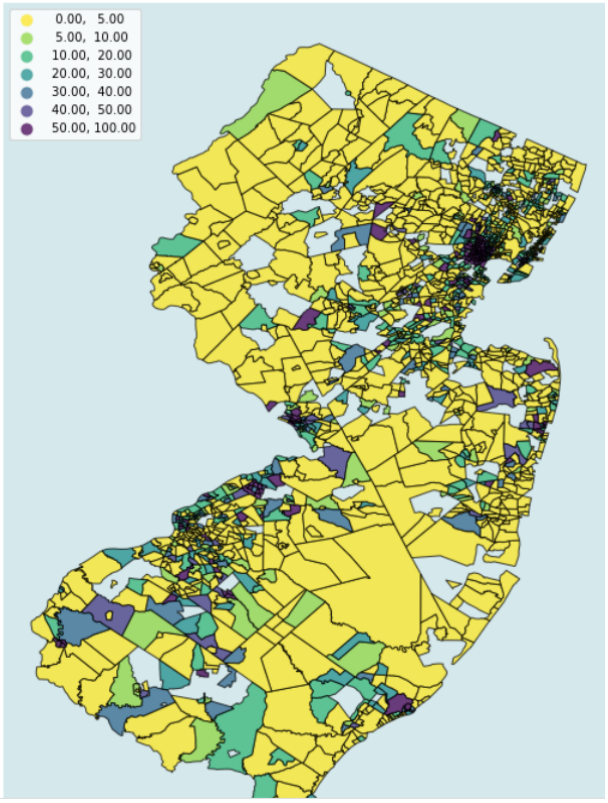
Figure 9. Actual uninsurance rate vs. predicted (NJ census tracts)

As seen in *Figure 9* our model was able to predict the uninsurance rate of given geographic areas (tracts in this case) with a strong confidence rate. Then, we went even further by predicting uninsurance rates of areas based on the demographic characteristic of race. The results are below:

Predicted % Uninsured African Americans



Actual % Uninsured African Americans



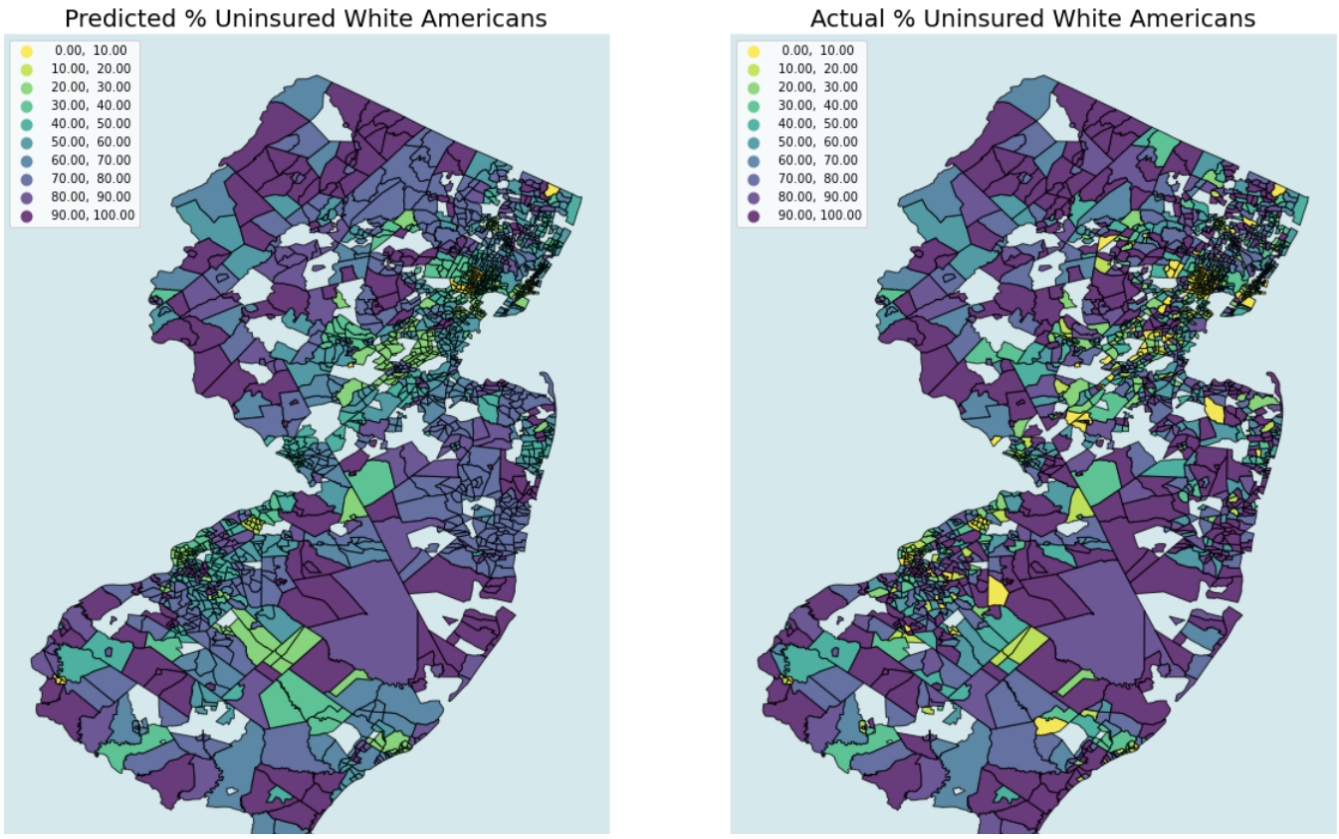


Figure 10. Actual uninsurance rate vs. predicted within NJ census tracts based on race

It becomes apparent that with larger datasets the result of our prediction model improves. Since there is more data for whites, the model was able to produce better results.