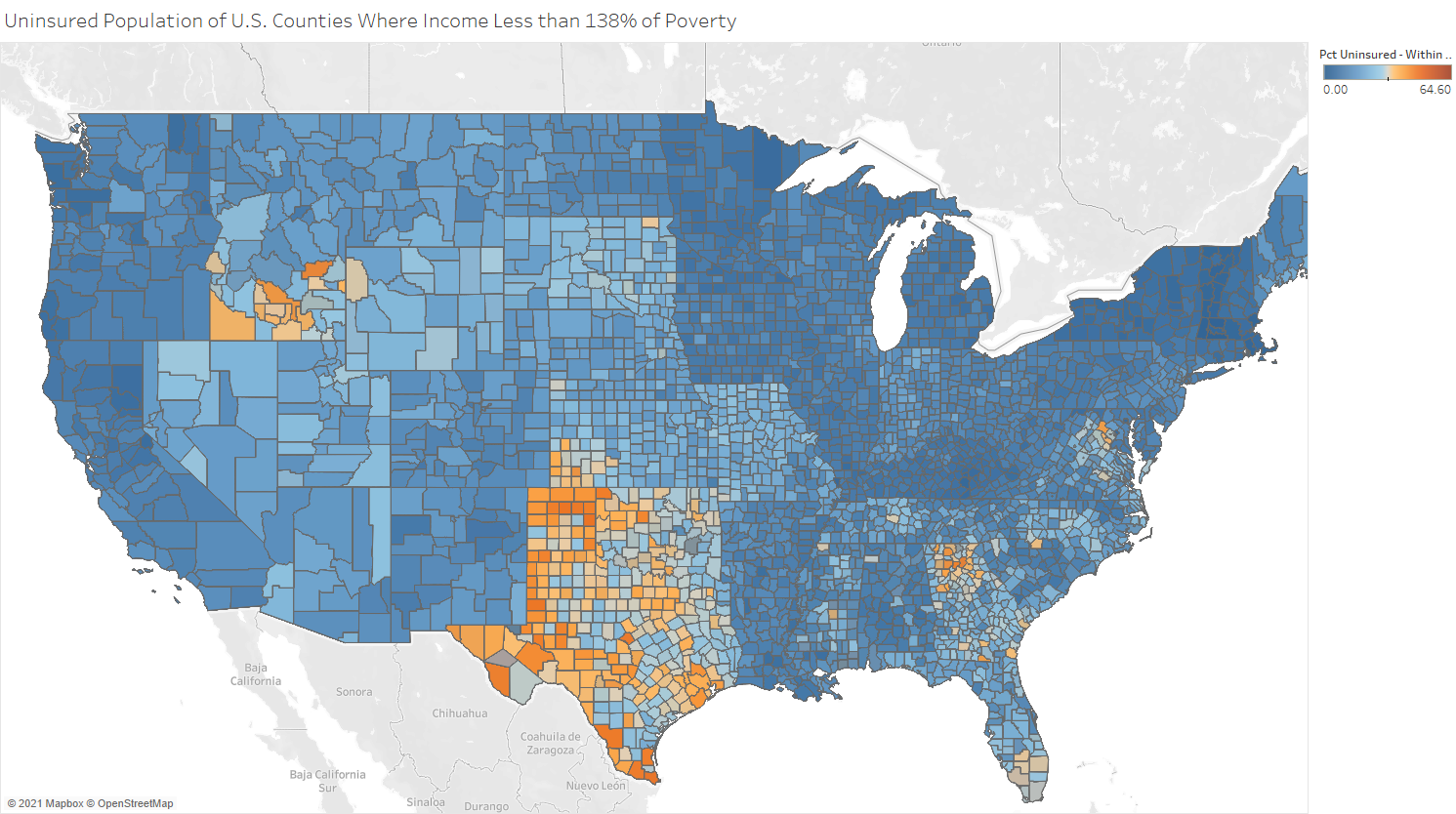
**Healthcare Accessibility Data Science Report**

Explanations for DS-Files Diagrams (to include in report)

Our solution addresses health accessibility for the county-average uninsured population whose income is less than 138% of the poverty line (or ~$36,000 for a family of 4). Though the target population is proportionally small, it still amounts to over 100 million people in the U.S.

As an end-to-end solution that supports more than just the uninsured population, the goal is to utilize user-data and a clustering algorithm to recommend medical insurance policies and a compatible hospital. The model uses the user’s county-level demographic data in combination with some extra datasets found on cms.gov for 2020 insurance policy data.

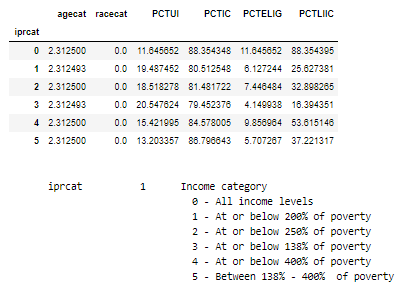
Future development would utilize the ‘USA Hospital Bed Capacity’ dataset by fitting the hospital type with the client’s newly attained healthcare plan.



Los Angeles (22,059- population uninsured), CA - pct uninsured within income demographic - 12.6%

Harris county (21,776 - population uninsured), tx - pct uninsured within income demographic - 38.6%

Uninsured patients seem to cluster in three key areas: Texas is a hotbed of uninsured persons, with many counties in the state having a high percentage of uninsured patients, likely due to Texas’ poor management of state insurance systems. Idaho and the rural states of the Northwest also present some number of counties with high uninsured persons, suggesting some correlation of uninsured persons with rural counties. And Georgia/the Deep South have the third cluster of uninsured patients, again likely due to state opposition to public insurance systems. Conversely, well-developed, urban areas like Los Angeles have a relatively low number of uninsured patients though it is a wide range of income levels and minority communities; this increased level of insurance accessibility could be due to California’s improved handling of health insurance which reduces the complexity and financial constraints of choosing a plan.



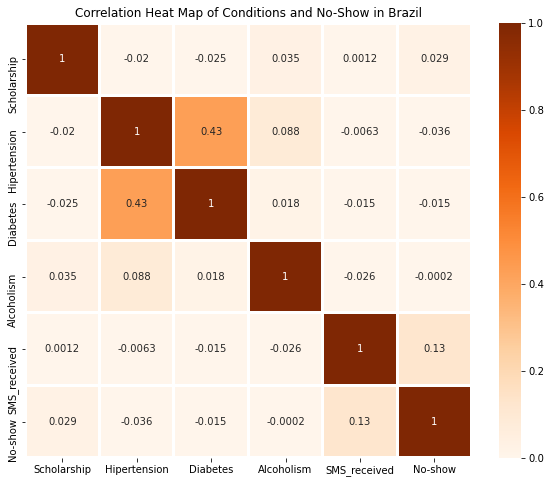
PCTUI Percent uninsured in demographic group for <income category>

PCTIC Percent insured in demographic group for <income category>

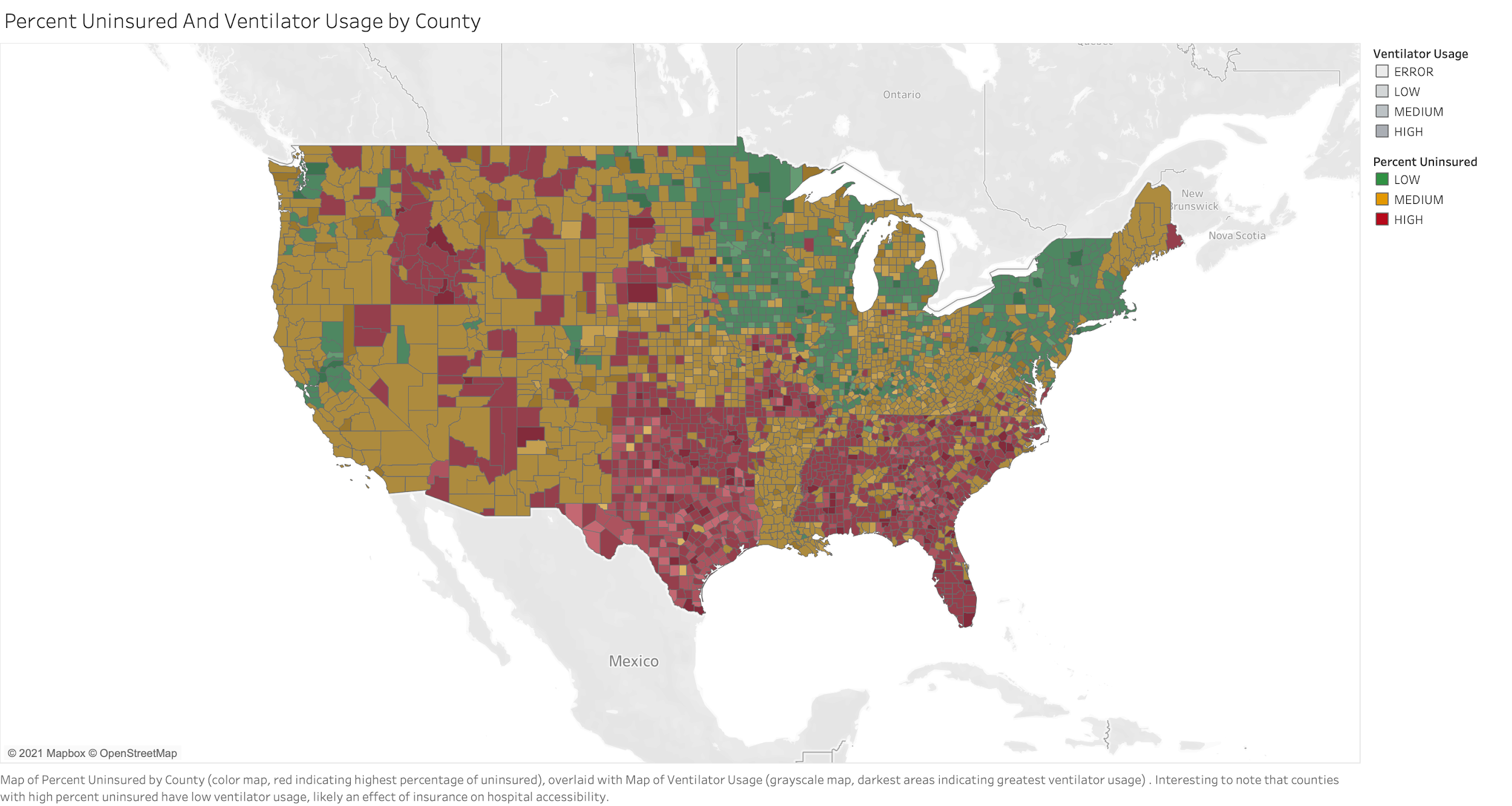
PCTELIG Percent uninsured in demographic group for all income levels

PCTLIIC Percent insured in demographic group for all income levels

The above chart looks at the percent uninsured populations for various income levels. Moving from the income categories in order from 4-3-2-1 (in order of increasing levels of poverty), we can see that the percentages of uninsured peoples certainly increase, which makes sense given that financial concerns are one of the largest limiting factors for access to insurance.



Just as a way to assess the datasets we are not focusing on for their relevance to insurance accessibility, we plotted a correlation matrix of the features in the “appointments no-show” dataset, paying special attention to their correlation with the “no\_show” target column. We can see that receiving an SMS has the most positive correlation with a patient showing up to their appointment, and that a patient having hypertension and diabetes are negatively correlated with a patient showing up to their appointment. Generally, this matrix suggests that a patient’s medical history does influence whether they show up for their appointment, meaning that pre-existing conditions are a value we need to ask the user of our app.



Map showing the relationship between the percent uninsured in each county (indicated by the color, with red being counties with the greatest percentage of uninsured people), along with the average ventilator usage in each county (which given that this is 2020 data, is positively correlated with the average covid load faced by each county, with darker areas being counties hit hardest by covid). It’s interesting to note that counties with the greatest number of uninsured patients are counties with generally the lowest ventilator usage (and corresponding covid load), which seems counterintuitive. However, a possible explanation is that uninsured patients who could need ventilators are simply not accessing hospitals due to insurance constraints, which is artificially lowering the ventilator usage rate. In other words, the ventilator usage rate is biased against counties with uninsured populations, and we can see that inverse correlation best in areas like Texas, Georgia and the Deep South, and Missouri, which have high numbers of uninsured but only low ventilator usage, even though they are well-populated areas.

Lastly, the scree plot for K-means below is a more specific plot depicting the strength of our KMeans clustering model. Inertia, or how well clusters are segregated, is the metric that we are trying to minimize, and which seems to reduce with a higher number of clusters. A larger number of clusters does seem to improve the specificity of our model in finding clusters of patients; however, these clusters are not the most intuitive because of the dimensionality reduction applied before this model which groups features together into a less intuitive combination.

