Executive Summary

Marjea, Ahad, Luis, Parth, Sharif

Our team is aimed to explore data sets regarding the uninsured population and its components with the goal of predicting likelihood of an individual being uninsured based on a series of categories. We will gather data, process data, run it through a data pipeline and store it in SQL database. We, then, will create a Machine Learning algorithm based on the Poisson Regression model to get predictions and test it. To wrap it up, we will set up a dashboard with key metrics and visualizations that will be updated periodically. To do these all, we will use following data sets:

1. SAHIE (Small area health insurance estimates)
2. 2013 Rural Urban Continuum Codes
3. County Population Totals: 2010-2019
4. Unemployment and median household income for the United States, States, and counties, 2000-20

To structure our research and exploratory process, we developed a set of questions and hypotheses to focus our attention to:

1. Which state has the highest likelihood of an uninsured person?
2. How many states have a population of uninsured people of 8% or more?
3. What are the main factors that contribute to increased likelihood of a person being uninsured?
4. Hypothesis test: The higher the income, the greater the probability of insurance coverage.
5. Hypothesis test: larger population leads to higher rate of uninsured
    a. Urban counties typically have a higher rate of uninsured.
    b. Rural counties will have lower rate of uninsured compared to urban
6. What state/counties can insurance companies target for insurance sales?

Finally, we will have a project report and presentation ready to present to an audience with a diverse background.