# Capstone Project Executive Report

Lindsey Oh, Scott Partacz, Raymond Sepulveda, Pemba Sherpa

Dev10 Data Track, Genesis10

## Introduction

### Background

On the list of Countries with the Best Health Care Systems, the United States ranks thirtieth (Ireland, 2021). When it comes to Total Health Spending by Country, however, the United States ranks first (OECD, n.d.). The majority of an American's health expenses can be attributed to health insurance premiums and out-of-pocket (OOP) costs (Investopedia Team, 2021). Given the time constraints of this Capstone, our team narrowed the project focus down to United States health insurance data for the year 2021. Since the value of premiums does not vary within a calendar year, our team decided to examine OOP costs for the insured population. The purpose of this project is to categorize insured people into health status clusters based on their difficulty performing daily everyday tasks due to various conditions. These clusters, in combination with the age, race, income, and sex demographic data, are then compared to OOP costs to determine whether a relationship exists.

### Exploratory Questions

1. What percent of each state in the U.S. has health insurance?
2. How do disability and health status affect out-of-pocket expenses for the insured?
3. How do socioeconomic factors affect the value of out-of-pocket expenses for the insured?
    a. How does age affect out-of-pocket expenses?
    b. How does race/ethnicity affect your out-of-pocket expenses?
    c. How does income affect out-of-pocket expenses?
    d. Does gender play any role in out-of-pocket expenses?

## Research

### Data Sources

After extensive research, we decided on two sources of data, both from the United States Census Bureau. Our primary source is the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) data for 2021.

In addition to being the primary source of monthly labor force statistics, the CPS is used to collect data for a variety of other studies that keep the nation informed of the economic

and social well-being of its people. This is done by adding a set of supplemental questions to the monthly basic CPS questions. Supplemental inquiries vary month to month and cover a wide variety of topics such as child support, volunteerism, health insurance coverage, and school enrollment. Supplements are usually conducted annually or biannually, but the frequency and recurrence of a supplement depend completely on what best meets the needs of the supplement's sponsor. (U.S. Census Bureau, Nov 2021)

Our secondary source is the Small Area Health Insurance Estimates (SAHIE) application programming interface (API).

The Small Area Health Insurance Estimates (SAHIE) program was created to develop model-based estimates of health insurance coverage for counties and states. SAHIE is only source of single-year health insurance coverage estimates for all U.S. counties. This program is partially funded by the Centers for Disease Control and Prevention's (CDC) Division of Cancer Prevention and Control (DCPC). The SAHIE program models health insurance coverage by combining survey data with administrative records and other Census data sources. The model produces timely and accurate estimates of health insurance coverage. Additionally, the SAHIE program's model methodology and estimates have undergone internal U.S. Census Bureau review as well as external review. (U.S. Census Bureau, Oct 2021)

**Data Platform**

Before setting up the data platform, the original file was downloaded from the 2021 ASEC website. The Person Record comma-separated value (CSV) file was uploaded to Azure Data Lake.

The Data Platform is set up in Azure Data Factory. The first element in the Data Platform is the Producer, which is triggered to run whenever changes are made to the storage. The Producer Pipeline consists of this trigger and the Producer Data Brick. The Producer Data Brick loads the CSV file from Azure Data Lake as a PySpark data frame, drops unnecessary columns, and sends each row as a message in JSON format to Kafka. Kafka runs in Confluent, which "is a full-scale data streaming platform that enables you to easily access, store, and manage data as continuous, real-time streams" (Confluent, n.d.). The Consumer Pipeline consists of a time trigger and the Consumer Data Brick. The Consumer consumes messages from Kafka and stores them in a SQL database. Data from the SQL database is then accessed to create both a Power BI

Report as well as the machine learning models. Specific to Power BI, a data flow is set up to load the data in set time intervals. This concludes the data flow from its original source to its final destination in the form of visualizations.

**Exploratory Data Analysis**

The most recent SAHIE API data available is from the year 2019. Figure 1 was created based on this data with the purpose of providing an overview of health insurance coverage in the United States. At the time, Massachusetts had the greatest percentage of insured survey respondents (96.5%). Texas had the least percentage of insured survey respondents (79.3%). The Northeastern states had a greater percentage insured than Southern states. In addition, each state's population should be considered. Oklahoma had the second-to-least percentage of insured survey respondents (83.2%), which represents 540,944 Oklahoman respondents. By contrast, 5,114,881 Texan respondents were uninsured. On average, about 90% of all survey respondents had health insurance.

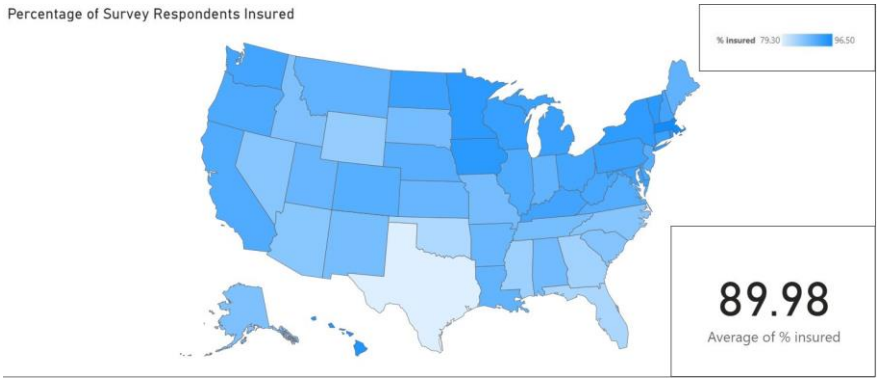


*Figure 1: Heat Map of US by Percentage of Survey Respondents Insured*

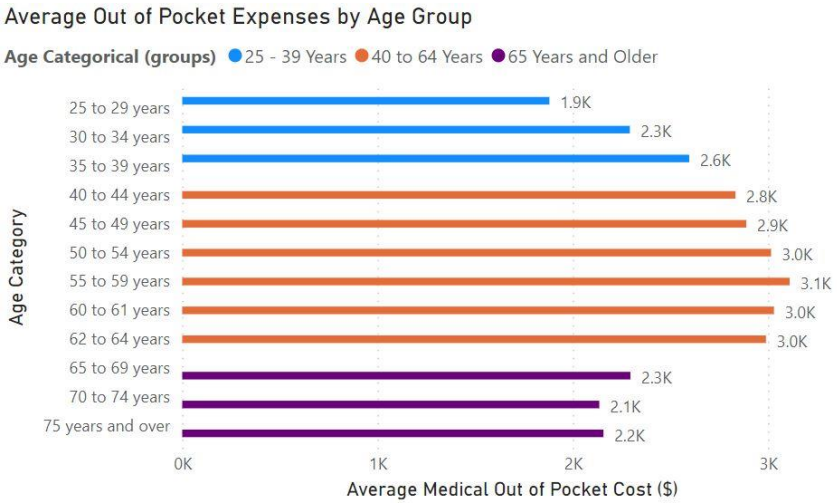


*Figure 2: Average Out of Pocket Expenses by Age Group*

Figures 2 through 5 were created based on the 2021 CPS ASEC data. These figures begin to explore Questions 3a through 3d. Figure 2 tells the story of how age affects average OOP costs. At age 25, a person may remain on their parents' health insurance, which could be why the 25-29 age group has the lowest OOP costs of all groups. The average OOP cost rises continually from 25-29 years to 55-59 years—a possible reflection of health issues accumulating with age. In the 60-61 and 62-64 age groups, the average OOP cost unexpectedly dips slightly before plummeting in the 65-69 age group. A person qualifies for Medicare once they turn 65, which may explain the plummet within this age group. People at age 65 switching from private to public coverage would also contribute to the sharp decrease in average OOP cost.

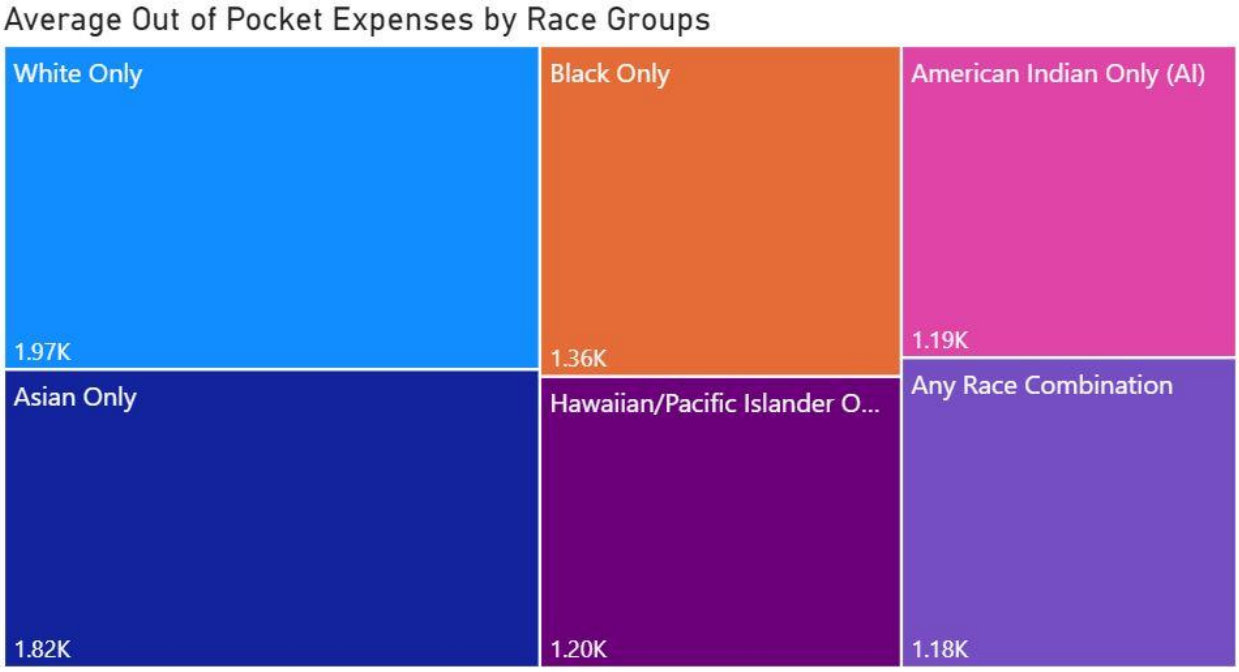


*Figure 3: Average Out of Pocket Expenses by Race Groups*

Average OOP costs by race groups can be seen in Figure 3. The White Only race group had the highest average of $1,970. The Any Race Combination race group had the lowest average of $1,180. Average OOP costs by ethnicity can be seen in Figure 4. Non-Hispanic White respondents had a significantly larger average OOP cost ($2,007) than Hispanic White respondents ($1,253). This same pattern holds true for the breakdown of costs: Non-Hispanic White respondents had larger average premium, over-the-counter, and non-comprehensive care costs than Hispanic White respondents.

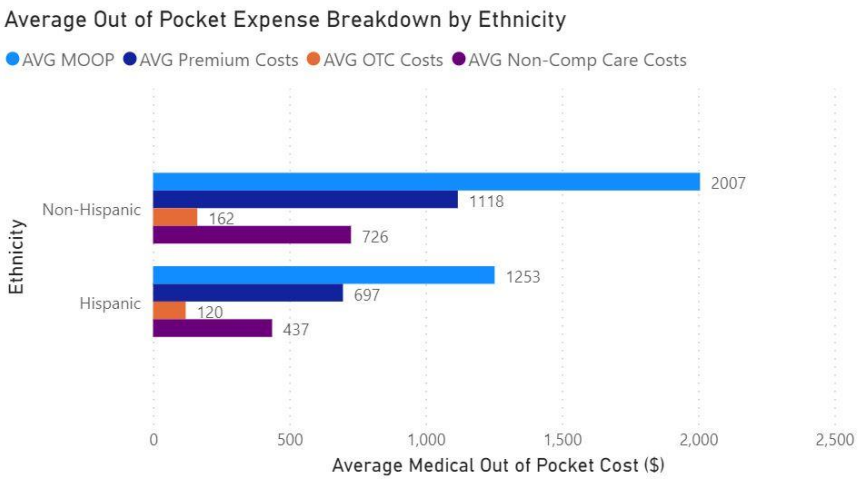


*Figure 4: Average Out of Pocket Expense Breakdown by Ethnicity*

Figure 5 compares the average OOP costs for females and males as a percent of the average adjusted gross income of the respective gender. While females ($1,770) pay less OOP costs than males ($1,940) on average, the expense is a greater percentage of the average female's income (5.8%) than that of the average male (4.6%).

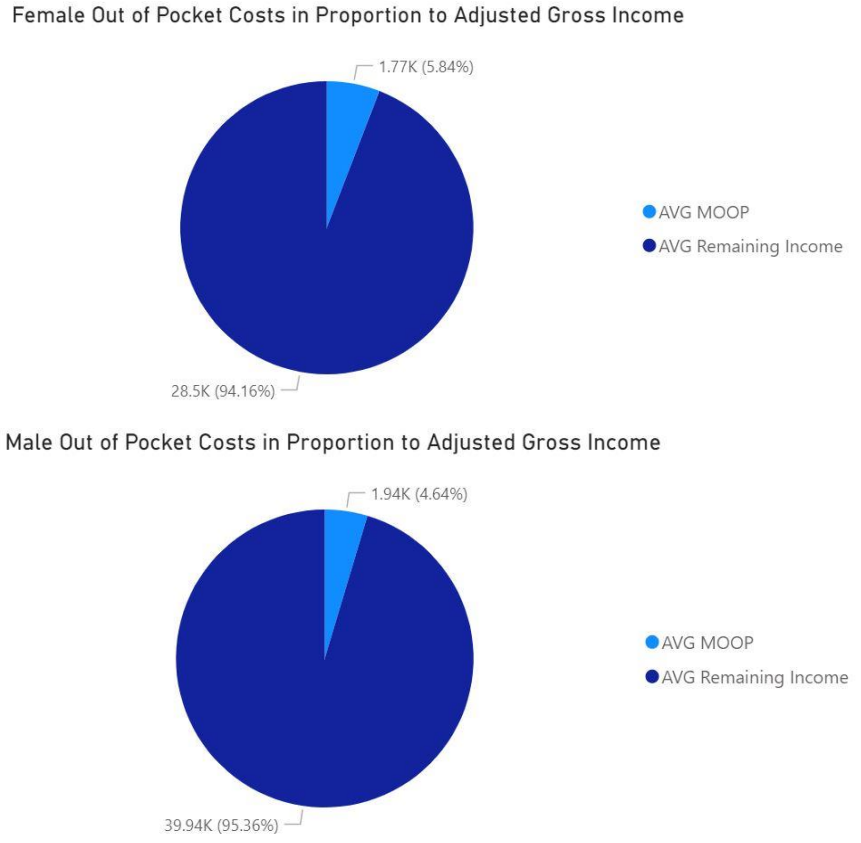


*Figure 5: Average Out of Pocket Costs in Proportion to Adjusted Gross Income by Gender*

**Machine Learning: Algorithm Selection**

After considering various clustering algorithms available to our team via Scikit-learn, we decided on K-Modes. Initially, we had considered using K-Means. However, the features of interest in our data are all categorical. K-Means is an appropriate algorithm for continuous/numerical data, not categorical data. Whereas K-Means uses means based on distances between data point, K-Modes uses modes based on matching categories between data points. The categorical features we used are:

- Does this person have difficulty dressing or bathing? (Yes/No)
- Is this person deaf or does this person have serious difficulty hearing? (Yes/No)
- Is this person blind or does this person have serious difficulty seeing even when wearing glasses? (Yes/No)
- Because of a physical, mental, or emotional condition, does this person have difficulty doing errands along such as visiting a doctor's office or shopping? (Yes/No)
- Does this person have serious difficulty walking or climbing stairs? (Yes/No)
- Because of a physical, mental, or emotional condition, does this person have serious difficulty concentrating, remembering, or making decisions? (Yes/No)
- Does this person have any of these disability conditions? (Yes/No)
- Age of person (in bins: 6 = 25 to 29 years, 7 = 30 to 34 years, 8 = 35 to 39 years, 9 = 40 to 44 years, 10 = 45 to 49 years, 11 = 50 to 54 years, 12 = 55 to 59 years, 13 = 60 to 61 years, 14 = 62 to 64 years, 15 = 65 to 69 years, 16 = 70 to 74 years, 17 = 75 years and over)
- Health status (1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor)

A scree plot was used to determine that we should set the number of health status clusters equal to 6. Dummy variables were created for the "Age of person" feature and "Health status" feature. The output was a column that denoted which health status cluster was assigned to each person by the K-modes algorithm.

This health status cluster column, combined with other demographic data, were used as the features for the second machine learning algorithm:

- Health Status Cluster
- Race (01 = White only, 02 = Black only, 03 = American Indian, Alaskan Native only (AI), 04 = Asian only, 05 = Hawaiian/Pacific Islander only (HP), 06 = White-Black, 07 =

White-AI, 08 = White-Asian, 09 = White-HP, 10 = Black-AI, 11 = Black-Asian, 12 = Black-HP, 13 = AI-Asian, 14 = AI-HP, 15 = Asian-HP, 16 = White-Black-AI, 17 = White-Black-Asian, 18 = White-Black-HP, 19 = White-AI-Asian, 20 = White-AI-HP, 21 = White-Asian-HP, 22 = Black-AI-Asian, 23 = White-Black-AI-Asian, 24 = White-AI-Asian-HP, 25 = Other 3 race comb., 26 = Other 4 or 5 race comb.)

- Did you ever serve on active duty in the U.S. Armed Forces (Yes/No)
- Are you Spanish, Hispanic, or Latino? (Yes/No)
- Sex (Male/Female)
- Public coverage last year (Yes/No)
- Any employment-based coverage last year (Yes/No)
- Any direct-purchase coverage last year (Yes/No)
- Any Marketplace coverage last year (Yes/No)
- Any non-Marketplace coverage last year (Yes/No)
- Medicaid coverage last year (Yes/No)
- Other means-tested coverage last year (Yes/No)
- Medicare coverage last year (Yes/No)
- Any TRICARE coverage last year (Yes/No)
- CHAMPVA coverage last year (Yes/No)
- VACARE coverage last year (Yes/No)

Dummy variables were created for the "Health Status Cluster" feature and "Race" feature. At first, we implemented various regression models: linear regression, logistic regression, gradient boosting (regression), random forests (regression). The results of these models were inconclusive—the correlation was negligible—so we modified our approach. Instead of predicting a dollar value for OOP costs, we decided to separate the OOP cost range into bins and then predict which bin a person would fall under. Random forests (classification) is the algorithm we selected to perform this task. Because the random forests (classification) algorithm predicts based on multiple decision trees instead of a single tree, it yields more accurate results.

## Machine Learning: Results

Figure 6 shows the bar graph that was created to visualize the clusters created using the K-Modes algorithm. A majority of the sample set falls under three clusters in our K-Modes

model, with 26.5% of the sample in Cluster 2, 19.4% of the sample in Cluster 6, and 9.2% of the sample in Cluster 1.

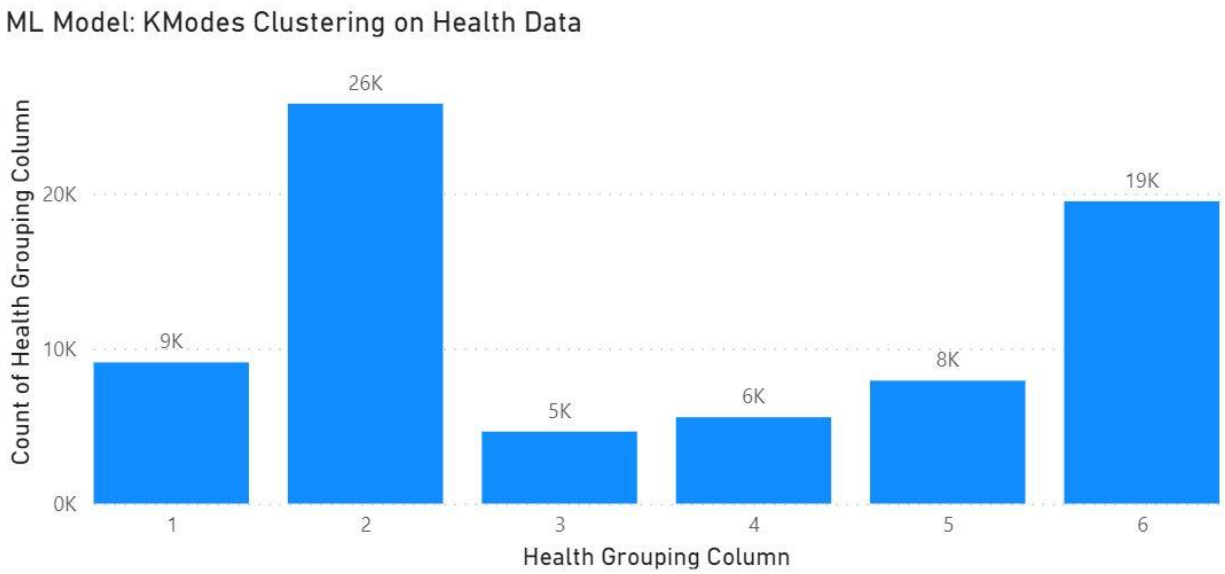


*Figure 6: K-Modes Clustering on Health Data*

A confusion matrix was created for the results of the random forest (classification) algorithm (Figure 7). Interestingly, the Classification model was somewhat accurate overall, with an accuracy rating of 76.1%. However, when it came to classifying Bins 1 and 4, the model was more accurate—with accuracy ratings of 99.3% and 81.3%, respectively.

**Bin 1**
$0-$1500
**Bin 2**
$1500-$3000
**Bin 3**
$3000-$4500
**Bin 4**
$4500+

ML Model: Out of Pocket Cost Random Forest Classification

| bins | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| 1 | 55492 | 289 | | 78 | **55859** |
| 2 | 7340 | 4956 | 14 | 3015 | **15325** |
| 3 | 2348 | 1638 | 212 | 5542 | **9740** |
| 4 | 2185 | 958 | 51 | 13925 | **17119** |
| **Total** | **67365** | **7841** | **277** | **22560** | **98043** |

*Figure 7: OOP Cost Random Forest (Classification) Confusion Matrix*

## Conclusion

Our hypothesis was that different demographic groups would have different OOP costs. However, the results of our project indicate otherwise. There is no significant difference in OOP costs across different demographic groups. In hindsight, these results account for the laws that exists to prevent discrimination based on various demographics.

**Recommendations and Next Steps**

We recommend that the insured population in the U.S., especially those who are purchasing health insurance for the first time, use our results to make informed decisions when selecting a coverage plan. The value of the plan premium is a significant indicator of how much the person will be paying for all OOP costs: premium, over-the-counter, and non-comprehensive care costs. We also recommend that health insurance companies, who have access to the most detailed breakdowns of data, expand upon our research with the purpose of adjusting policies to attract more clients.

Given more time and access to better quality data, our team would expand our data analysis over more than one year, since ASEC data is available from 1998. We would also search for data to perform a longitudinal analysis. This would provide insight into how health insurance premiums, medical OOP costs, and over-the-counter costs change over time. In addition, having more data with which to train the machine learning model may yield more accurate results than those of this project.

## References


Confluent. (n.d.). *What is Confluent Platform?* Confluent Documentation. Retrieved February 14, 2022, from https://docs.confluent.io/platform/current/platform.html

OECD. (n.d.). *Health spending*. OECD Data. Retrieved February 10, 2022, from https://data.oecd.org/healthres/health-spending.htm

Investopedia Team. (2021, June 29). *What Country Spends the Most on Healthcare?* Investopedia. Retrieved February 10, 2022, from https://www.investopedia.com /ask/answers/020915/what-country-spends-most-healthcare.asp

Ireland, S. (2021, April 27). *Revealed: Countries with the Best Health Care Systems, 2021*. CEOWORLD Magazine. Retrieved February 10, 2022, from https://ceoworld.biz / 2021/04/27/revealed-countries-with-the-best-health-care-systems-2021/

U.S. Census Bureau. (2021, October 8). *SAHIE API*. United States Census Bureau. Retrieved February 10, 2022, from https://www.census.gov/programs-surveys/sahie/data/api.html

U.S. Census Bureau. (2021, November 22). *About the Current Population Survey*. United States Census Bureau. Retrieved February 10, 2022, from https://www.census.gov/programs-surveys/cps/about.html