Dev10 Data Capstone
Group 4: AlgoRhythm

Raymond Sepulveda, Pemba Sherpa, Scott Partacz, Lindsey Oh

Introduction

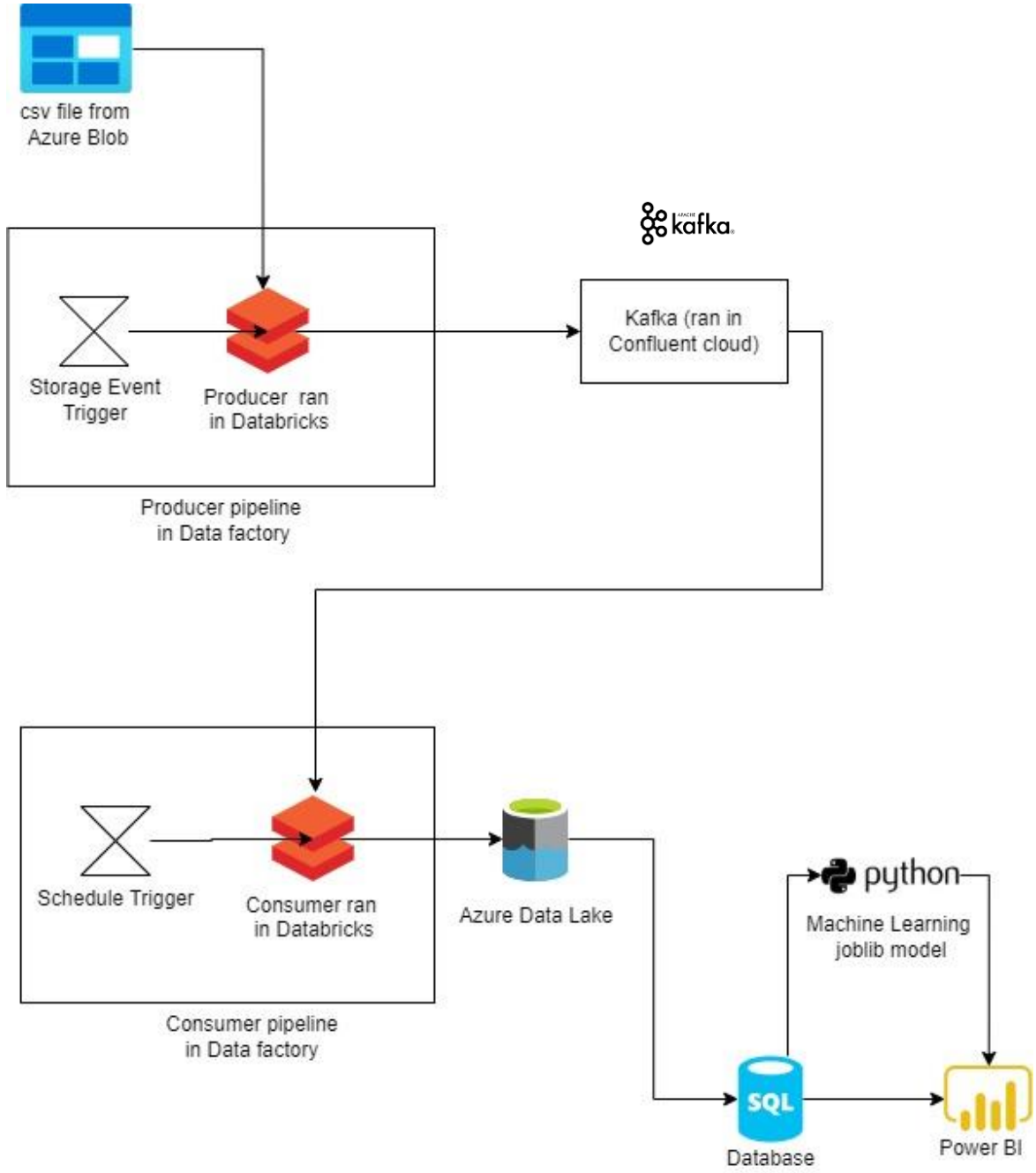
- Project Domain: Health Insurance
- Data Sources
  - CPS ASEC 2021
  - SAHIE API
- Project Purpose:
  - Cluster data
  - Predict based on clusters and demographics

Initial Questions

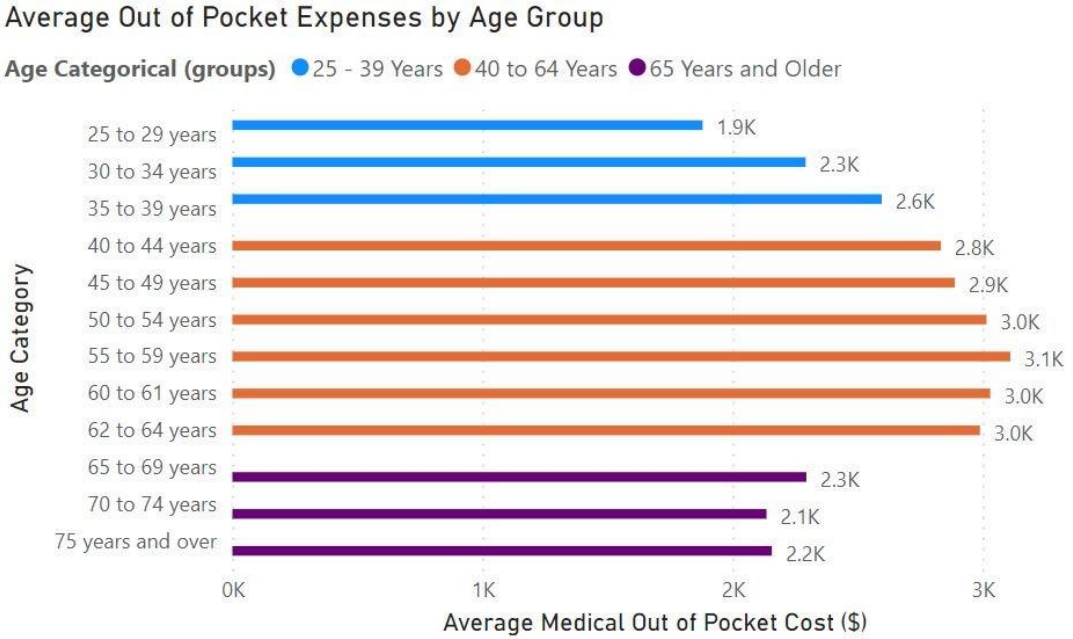
- What percent of each state in the U.S. has health insurance?
- How do disability and health status affect out-of-pocket expenses for the insured?
- How do socioeconomic factors affect the value of out-of-pocket expenses for the insured?
    - How does age affect out-of-pocket expenses?
    - How does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

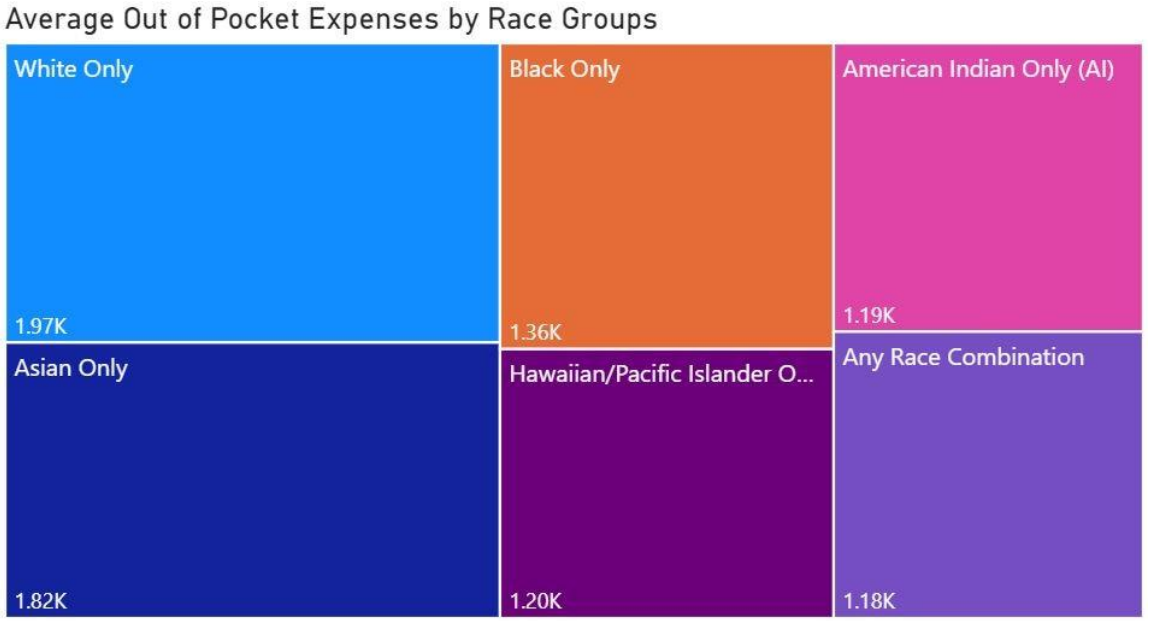
Hypothesis: Demographics will affect out-of-pocket expenses.

Data Platform

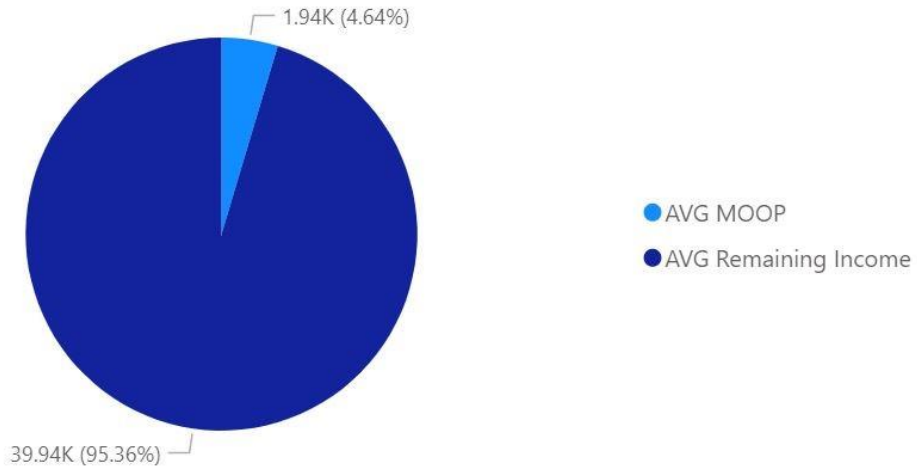


Research Process: Visualizations

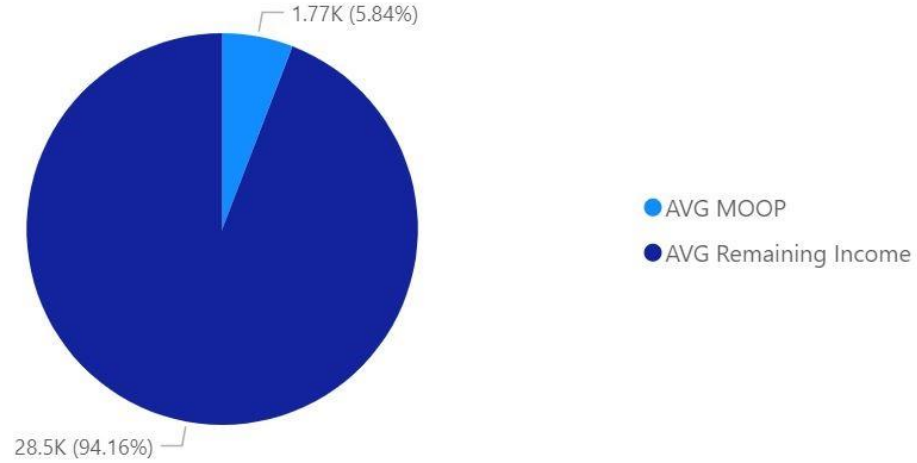




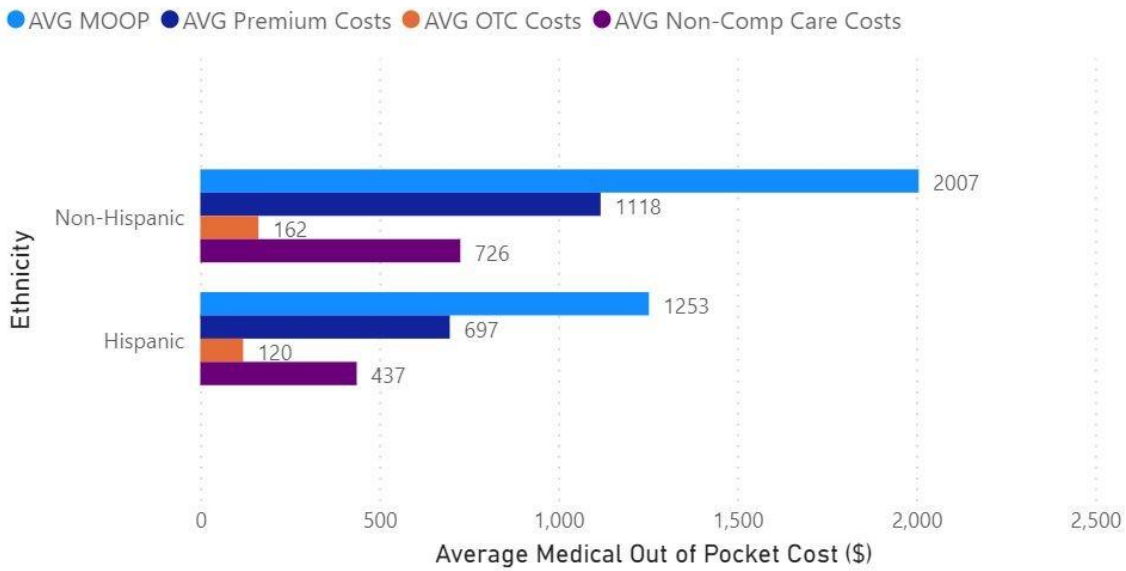
Male Out of Pocket Costs in Proportion to Adjusted Gross Income



Female Out of Pocket Costs in Proportion to Adjusted Gross Income



Average Out of Pocket Expense Breakdown by Ethnicity

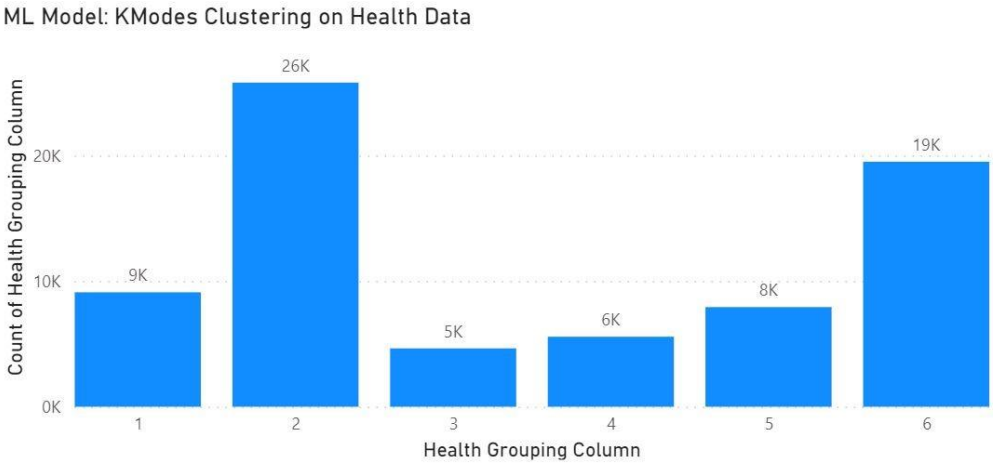


Machine Learning Exploration

- Goal – accurately predict how much an individual would spend on out-of-pocket health expenses
- Clustering model – group people based on their health-related quantifiers (DBSCAN, K-means, **K-modes**)
- Regression model – predict the total of out-of-pocket expenses (Linear regression, Gradient Boosting Regressor, and Random Forest)
- Classification model – predict which bin some would fall into. (**Random Forest**)

Kmodes

- Unsupervised Learning
- Used to group people based on health-related quantifiers



Random Forest

- Predict which bin individuals will fall into.
- Accuracy Score Overall: 76.1%

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**

Recommendation and Conclusion

- No significant difference in OOP costs in proportion to AGI across different demographic groups.
- These results account for the laws that exists to prevent discrimination based on various demographics
- Given more time and access to better quality data, our team would expand our data analysis over more than one year.

Questions?