SOCIALLYDETERMINED

Data Imputation & Social Determinants of Health

Bay-esian Watch: Stephanie Balarezo, Carmin Berberich, Cara Dunnavant, and Megan Schaeb

Special thanks to Dr. Alan Lattimer, Mike Ackermann, and Quiyana Murphy for guiding and aiding us through this project

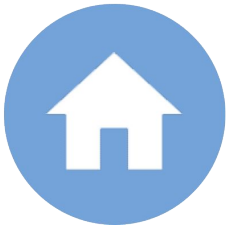
Patient sleeps in hospital bed. (2019). WVUToday. Retrieved October 16, 2022, from https://wvutoday.wvu.edu/stories/2019/11/14/too-much-light-may-darken-mood-of-hospital-patients-say-wvu-researchers.

Social Determinants of Health

Financial Strain

Food Insecurity

Transportation Barriers



Housing Instability

Health Literacy Challenges

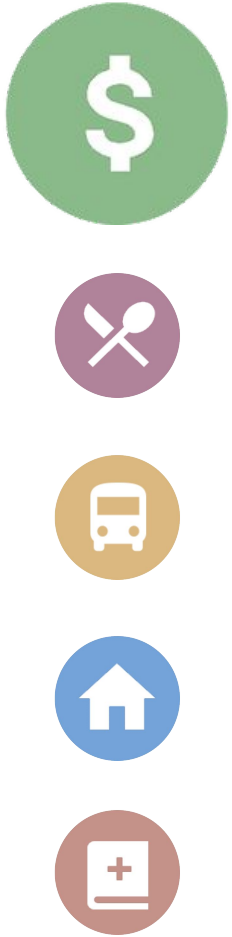
Problem Statement

Problem:

- Data used in risk score models is often missing or uncertain

Our Goal:

- Address missing data through imputation
- Address uncertain data through conditioning a background distribution
- Leverage data correlations, additional data sources, literature review



Literature Review

Addressing Health Equity and Social Determinants of Health Through Healthy People 2030 - Gómez, Cynthia

- Provided background on health equity and strategies to achieve it

Gómez, Cynthia A., et al. "Practice FullReport: Addressing Health Equity and Social Determinants of Health Through Healthy People 2030." *Journal of Public Health Management and Practice* 27.6 (2021): S249.4

Comparison of Performance of Data Imputation Methods for Numeric Dataset - Jadhav, Anil, et al.

- Evaluated 7 different numerical data imputation methods

Jadhav, Anil, et al. "Comparison of Performance of Data Imputation Methods for Numeric Dataset." *Applied Artificial Intelligence*, vol. 33, no. 10, 2019, pp. 913-933. *Taylor & Francis Online*, https://www.tandfonline.com/doi/full/10.1080/08839514.2019.1637138.

Bayesian network data imputation with application to survival tree analysis - Rancoita, Paola M.V., et al.

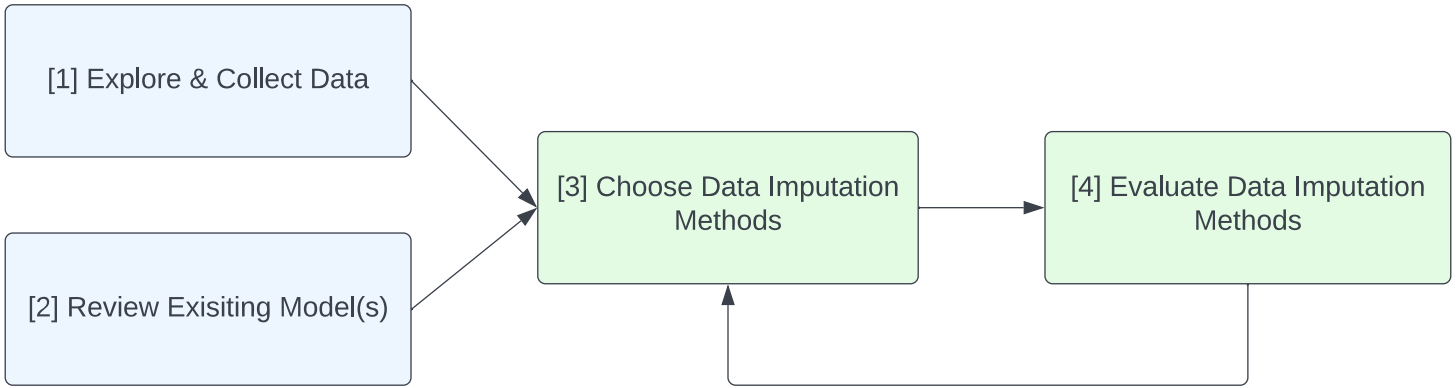
- Discussed Bayesian networks and model evaluation metrics

Rancoita, Paola M.V., et al. "Bayesian network data imputation with application to survival tree analysis." *Computational Statistics & Data Analysis*, vol. 93, no. January 2016, 2014, pp. 373-387. *ScienceDirect*, https://www.sciencedirect.com/science/article/pii/S0167947314003569. Accessed 15 September 2022.

Project Process for Socially Determined

Discover Relationships	Identify Distribution	Risk Score Produced
▷ Literature reviews ▷ Data Science tools/ Machine Learning to impute missing bins (ranges of values)	▷ Condition the distribution ▷ Use correlations to impute a likely value within the bin	▷ Obtain risk distributions ○ A risk score of 1 - 5

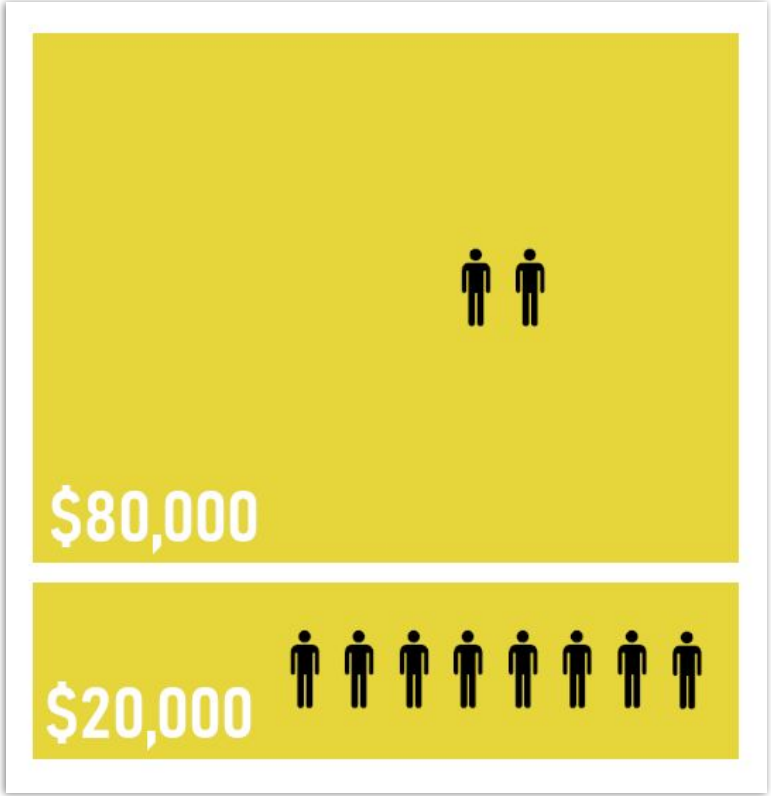
Project Components



Classification Strategies Chosen

- Imputation
  - **Decision Trees**
    - Works on a set of decisions derived from the data and its behavior
  - **Random Forest**
    - Consists of a large number of individual decision trees that operate as an ensemble
  - **K-Nearest Neighbor (kNN)**
    - Creates groups among individuals in the training data based on similar characteristics in the predictor data fields
- Evaluation
  - **Accuracy**
    - Calculated from the confusion matrix
    - Defined as the number of correct predictions divided by the total number of predictions (with +/- one bin margin of error)
  - **Chi-Square Test for Independence**
    - P-value **< 0.05** indicates a significant relationship between data fields

Distribution Strategies Chosen

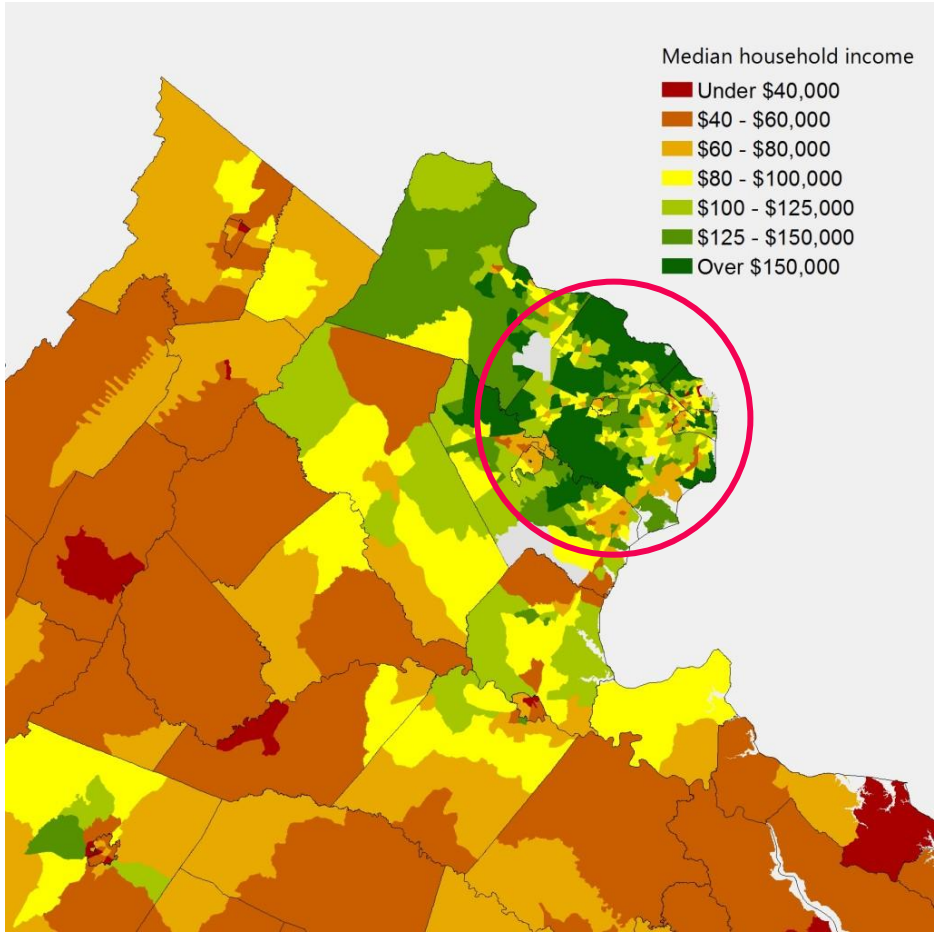


- Imputation
  - Condition background distribution on wealth score field
  - Identified as **Pareto Distribution**
    - Background distribution for U.S. net worth
- Evaluation
  - We were not able to complete an evaluation for the final imputation as we were not able to deduce a final distribution to impute from due to the lack of real-world wealth data available

Wikimedia Foundation. (n.d.). *Pareto principle*. Wikipedia. Retrieved December 9, 2022, from https://en.wikipedia.org/wiki/Pareto_principle

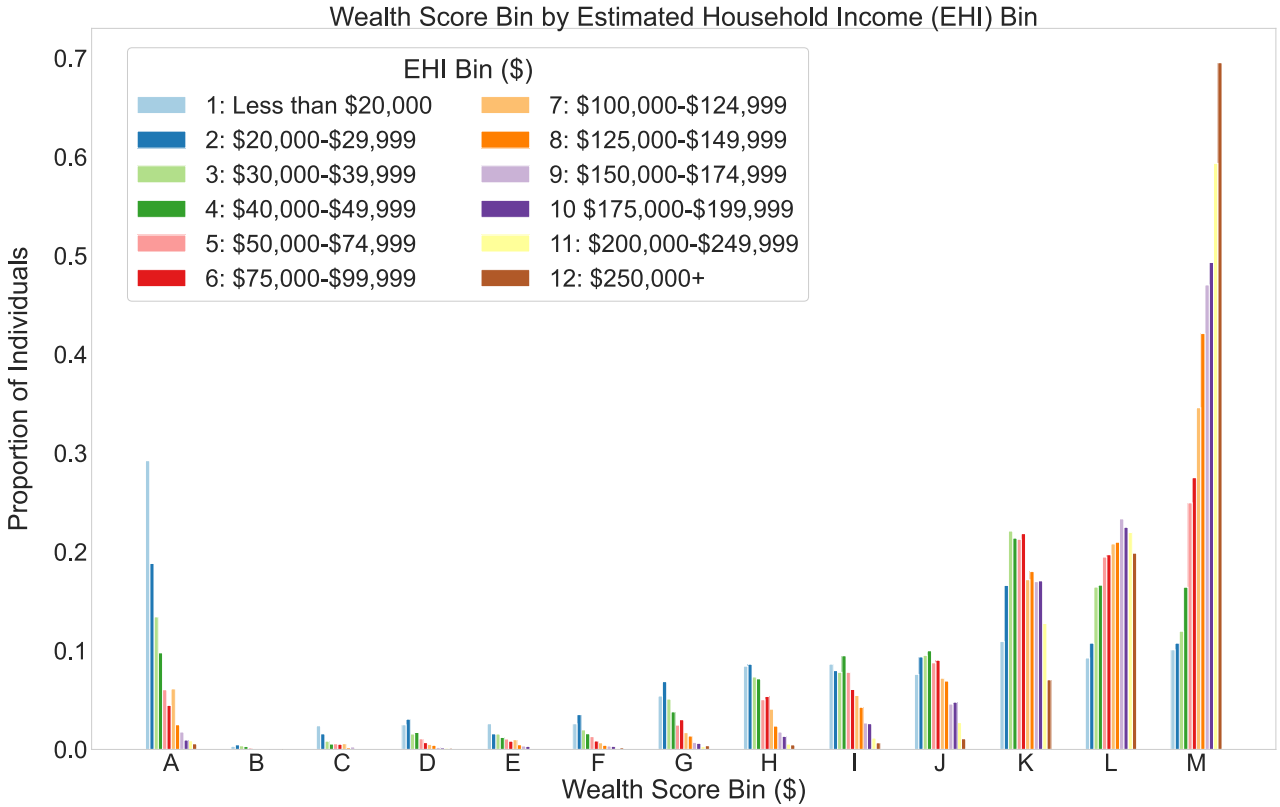
Limitations

- Lack of publicly available wealth data
  - Unable to deduce parameters for a final distribution
- Infutor data is biased
  - Only individuals from the Fairfax, VA, and DC area that made credit card transactions



Juday, L. (2015). StatChat. University of Virginia. Retrieved December 9, 2022, from https://statchatva.org/2015/11/23/northern-virginia/.

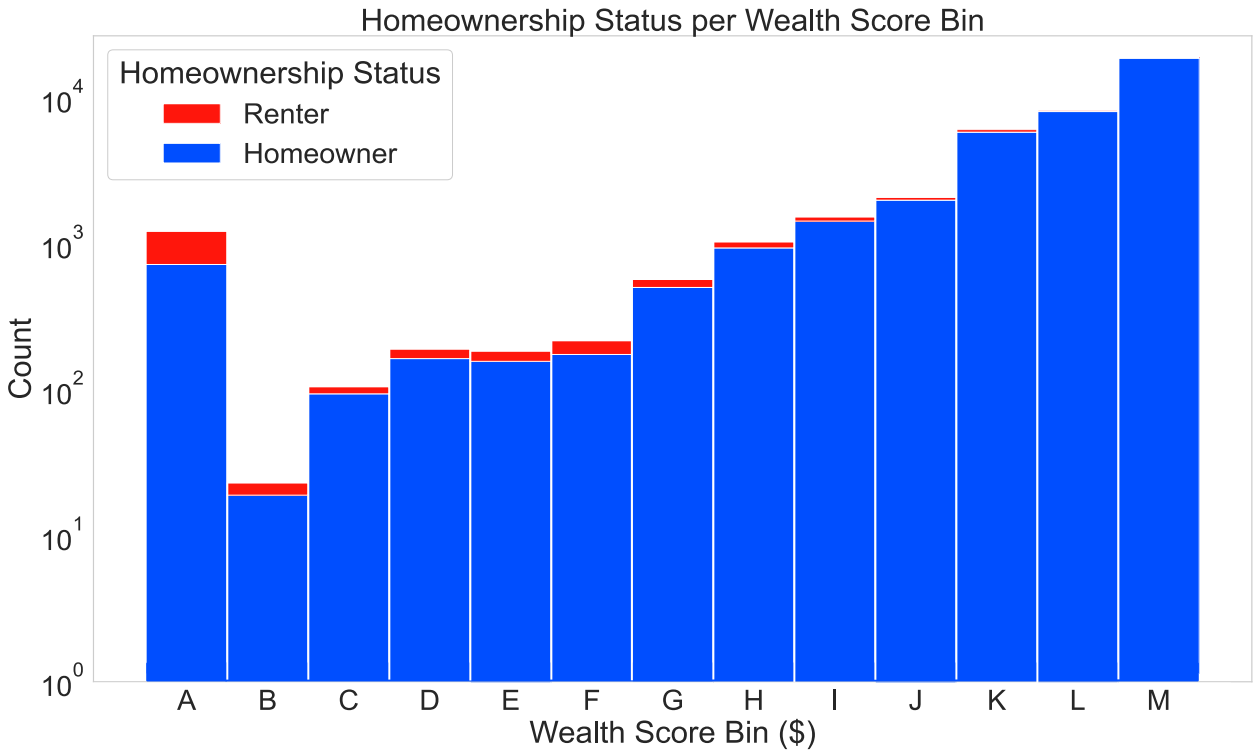
Results: Wealth & EHI



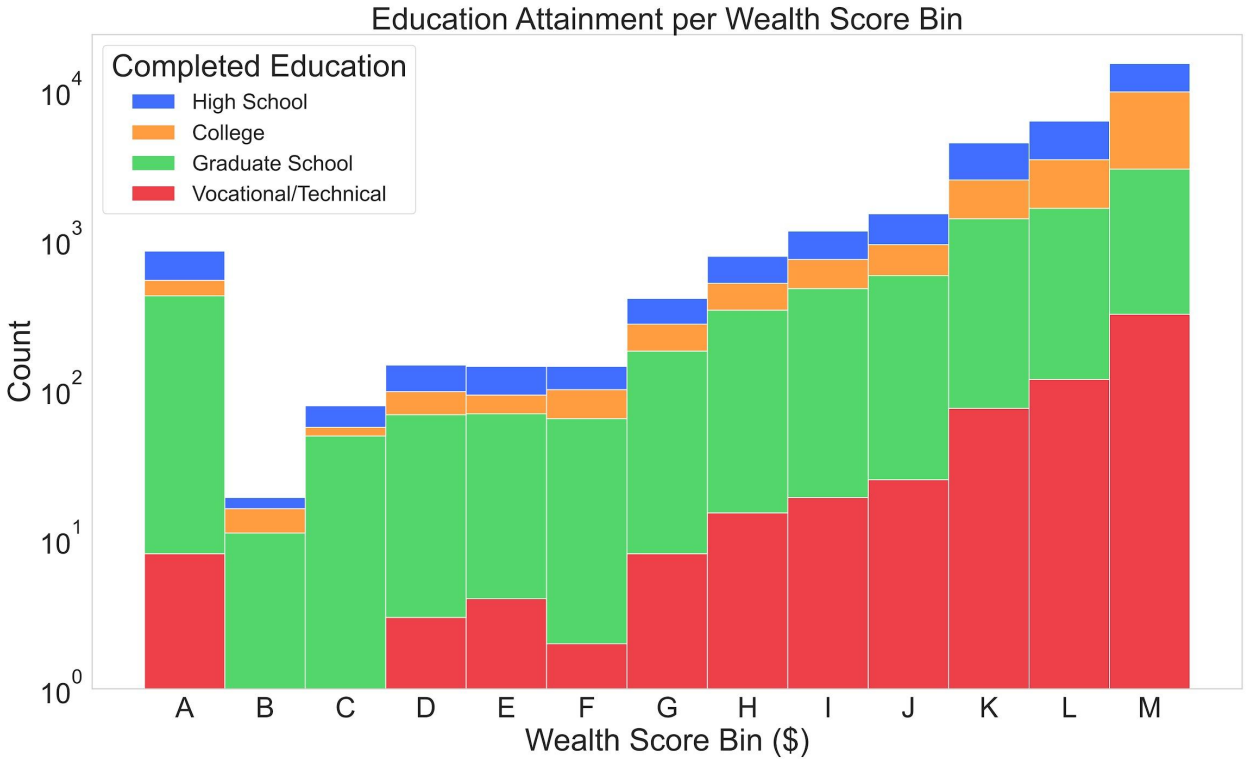
- **Direct relationship** between EHI and wealth
- Proportion of individuals with **highest EHI** is largest among those with highest wealth score
- Proportion of individuals with **lowest EHI** is largest among those with lowest wealth score
- Chi-square p-value **<0.001**

Results: Wealth & Homeownership

- As wealth increases, renter count decreases
- Chi-square p-value **<0.001**
  - Homeownership is significantly associated with wealth score



Results: Wealth & Education



- Lowest bin has the highest count of graduate school
  - Could be due to debt or working low paying jobs while in school
- Chi-square p-value **<0.001**
  - Education is significantly associated with wealth

Classification Method Results

Average Accuracy for Classification Models

Predictors Included		Decision Tree	Random Forest	kNN
1	EHI, Homeownership, and Education	0.718	0.720	0.721
2	EHI and Homeownership	0.707	0.707	0.710
3	EHI and Education	0.712	0.714	0.720

Results: Pareto Distribution for Wealth

$$p(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}$$

The Probability Density Function for the Pareto distribution

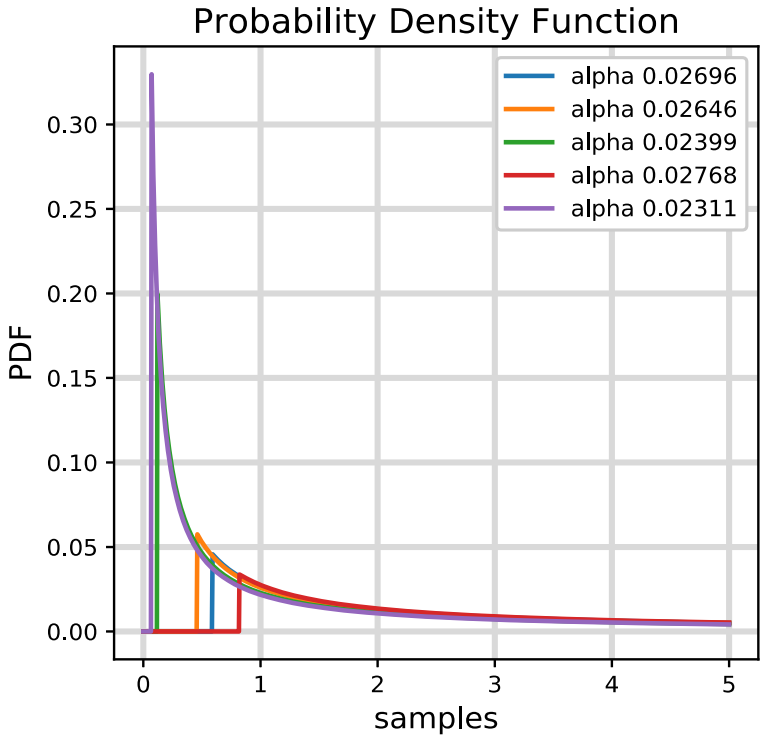
- x_0 represents scale
- α represents shape

$$\alpha = log_{(\frac{T}{x_0})}\sqrt{2}$$

Simulate α

- x_0 is randomized
- **T** represents an individual's median net worth

Results: Pareto Distribution for Wealth

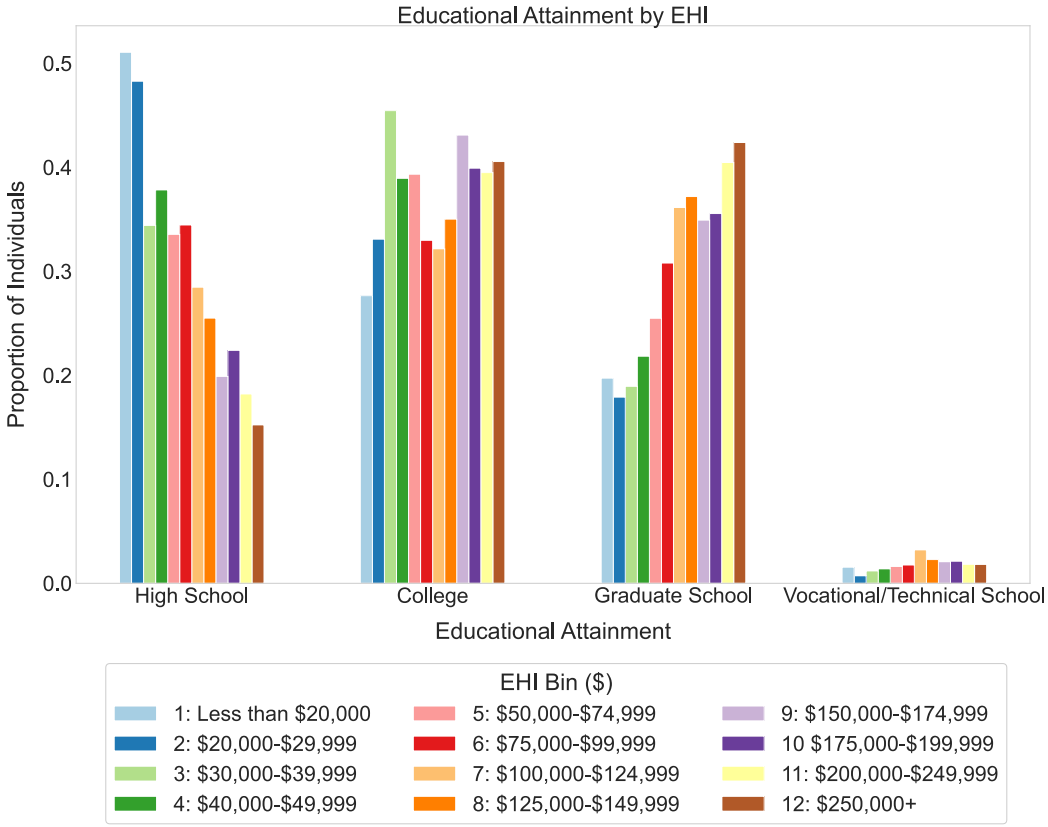


- Simulations of Pareto Distribution for an Individual
- Utilized Individual's Background Demographics

State	Median Net Worth
Alabama	88,910
Alaska	(B)
Arizona	149,300
Arkansas	78,100
California	247,500

Table: Sample Data for State Median Wealth from SIPP (*State-Level Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2020*); (B) denotes a missing median net worth due to lack of data

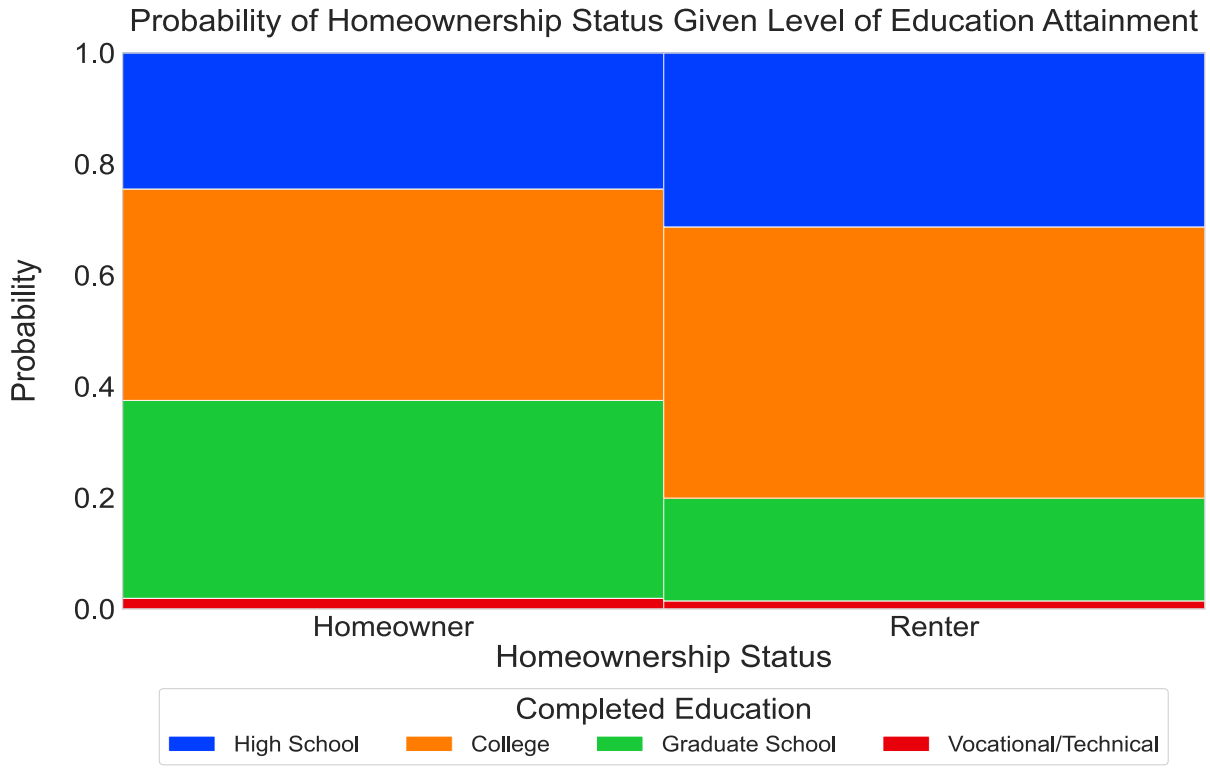
Results: Education & EHI



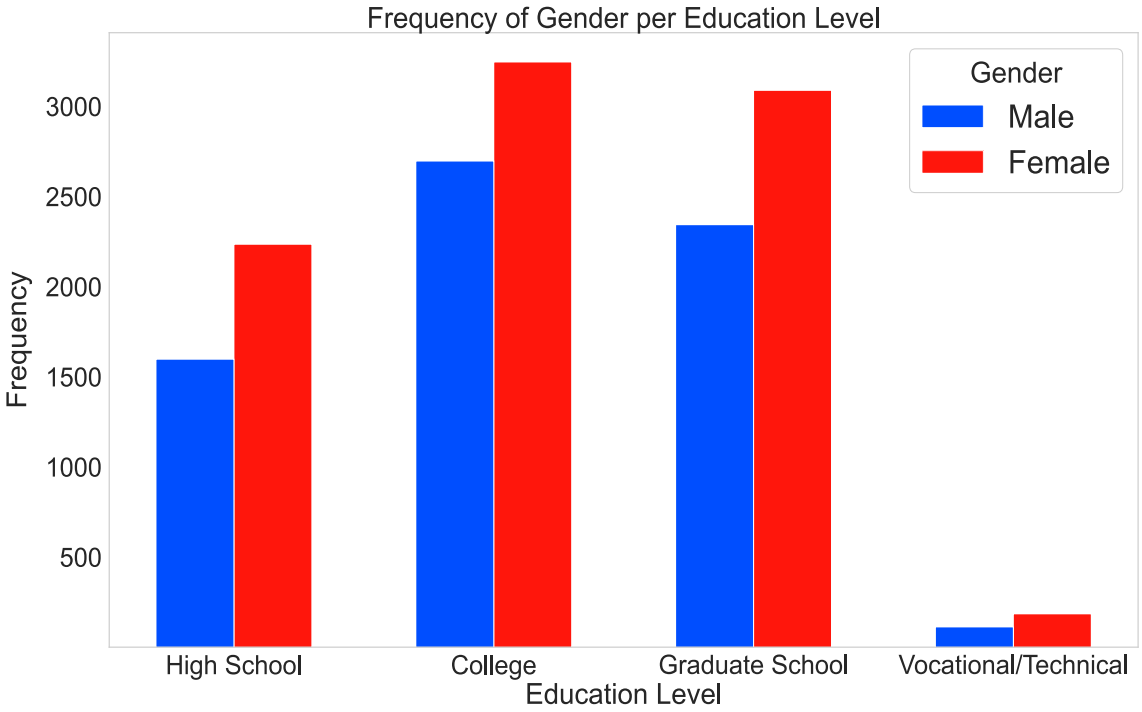
- **Direct relationship** between EHI and education
- Proportion of individuals with **highest EHI** is largest among those who attained a **graduate school** education
- Proportion of individuals with **lowest EHI** is largest among those who attained a **high school** education
- Chi-square p-value **<0.001**

Results: Education & Homeownership

- Among graduate students, they are **2x** as likely to own a house than to rent
- Approximately **82%** of all renters have education levels lower than grad school
  - Approximately **62%** of homeowners
- Chi-square p-value **<0.001**
  - Homeownership is significantly associated with education attainment



Results: Education & Gender



- In our dataset, women attain higher education than men across all 4 levels
- Chi-square p-value **<0.001**
  - Gender is significantly associated with education attainment

Interpretation of Results

- Wealth score can be predicted using EHI, homeowner status, and education achievement level
- Decision trees, random forest, and kNN are all viable ways of predicting one's wealth score bin
  - kNN has highest accuracy
- Pareto has been identified as background distribution for wealth
- Education can be predicted using EHI, homeowner status, and gender
- Improved risk score models can help inform Socially Determined's stakeholders on how to reduce the severity of adverse health outcomes

Conclusion

Main takeaways:

- We were able develop a process to address issues with
  - missing wealth data through imputation
  - uncertain wealth data by conditioning a background Pareto distribution
- Classification techniques were used to predict one's wealth score bin with over 70% accuracy (including margin of error)

Conclusion

Next steps:

- Repeat process with unbiased data to validate results
- Social scientist with access to wealth population data could determine Pareto distribution parameters for wealth model
- Run classification techniques on education and fit a distribution to it

We hope our findings can be used to improve Socially Determined's risk score models in an effort to help achieve health equity

Bibliography

- State-Level Wealth, Asset Ownership, & Debt of Households Detailed Tables: 2020 (2022, August 31). United States Census Bureau. Retrieved from https://www.census.gov/data/tables/2020/demo/wealth/state-wealth-asset-ownership.html

- Household Income Distribution Overall (2022). Northern Virginia Regional Commission. Retrieved from https://www.novaregiondashboard.com/household-income-distribution