Project Proposal
Data Science Capstone

Title

Generational Differences in Student Loan Debt and the Implications on Homeownership Trends

Research Question

How do differences in student loan debt between older and current generations affect homeownership trends?

Motivation

Understanding the difference in student loan debt composition across different generations holds personal significance to us as members of the current generation facing the burden of student loans. It provides insight into the economic and social dynamics that shape our society and addresses a pressing societal issue affecting millions of individuals. It speaks to the experiences of students navigating the complexities of higher education financing and the challenges they face in achieving future financial stability, including homeownership. By examining these dynamics, our research aims to empower individuals with knowledge to make informed financial decisions and advocate for systemic reforms.

This study also reflects on the evolving societal norms and expectations surrounding higher education and homeownership. It highlights the shift towards higher education attendance and the resulting increase in student loan debt, which affects traditional milestones such as home ownership. By exploring the differing financial landscapes faced by various generations, this research fosters a deeper understanding of the unique challenges and opportunities each generation encounters.

Insights from our study can inform the design of educational programs, policy interventions, and lending practices aimed at mitigating the adverse effects of student loan debt on individuals and society at large. Ultimately, our research contributes to building a more inclusive and resilient society by addressing systemic barriers to economic opportunity and wealth accumulation.

In summary, the proposed research holds significant merit, with the potential to drive positive societal change. A better understanding of generational differences in student loan debt and its implications on homeownership rates can lead to more informed policies, enhanced educational programs, and a more equitable and prosperous future.

Data

For this research project, data will be acquired from several reputable and publicly accessible sources that provide detailed information on student loan debt, demographic characteristics, and

homeownership trends. We'll be compiling data about student loan debt and homeownership rates by age. While the following datasets have already been created and cleaned, we plan to join these datasets by age so that our final dataset is unique to this project.

The Federal Reserve compiled these statistics on family debt holdings from 1989-2022. This dataset includes information regarding age, income percentile, and debt broken out by education loans, mortgage loans, auto loans, and other debts by survey year.
Family holdings of debt, by selected characteristics of families and type of debt, 1989–2022 surveys:
https://www.federalreserve.gov/econres/files/scf2022_tables_public_nominal_historical.xlsx

This data portfolio from the Federal Student Aid office features datasets that include year, age range, outstanding debts, and number of borrowers.
Direct Loan Portfolio by Borrower Age:
https://studentaid.gov/data-center/student/portfolio

Lastly, this dataset from Census.gov lists the number of householders in thousands based on age and year.
Household Estimates for the United States, by Age of Householder: 1982 - Present:
https://www.census.gov/housing/hvs/data/histtab12.xlsx

Methodology Requirements

To incorporate statistical analysis and inference into the project, we will calculate the mean, median, and standard deviation of student loan debt by age group to understand central tendencies and variability. We will summarize the distribution of debt types (student loans and mortgages) across different age groups by year. Regression analysis will be performed to model the relationship between age, student loan debt, and homeownership rates, specifically to understand the impact of student loan debt on the probability of homeownership. Confidence intervals for the mean student loan debt and homeownership rates will be calculated to quantify the precision of our estimates, reporting 95% confidence intervals to provide a range within which the true population parameters are expected to lie. Hypothesis tests, such as t-tests, will determine whether observed differences in debt levels between generations are statistically significant, using p-values to assess the likelihood that the observed differences occurred by chance.

We will utilize several types of data visualizations to demonstrate our conclusions effectively. A line graph will depict general trends in student loans over time, while another graph will compare mortgage loans across different age groups. A bar chart will be used to compare the average student loan debt across different generations, with each bar representing a different generation (Baby Boomers, Gen X, Millennials, Gen Z) and error bars indicating confidence intervals or standard deviations. Additionally, a scatter plot will examine the relationship

between student loan debt and homeownership rates, where each point represents a different demographic group or geographic region.

The data engineering component will primarily involve the acquisition and organization of data. This includes conducting exploratory data analysis (EDA) in SQL and joining datasets to ensure a coherent and analyzable data structure.

Machine learning techniques will be employed to enhance the analysis. Time series forecasting models will predict future trends in student loan debt and homeownership rates, helping policymakers and stakeholders anticipate potential challenges and develop proactive strategies. Clustering algorithms, such as k-means clustering, will group individuals based on their debt profiles and homeownership status to identify distinct patterns and trends. Predictive modeling will forecast future trends in student loan debt and homeownership rates and predict the probability of homeownership based on student loan debt and other demographic factors.

Ethical concerns will be addressed by ensuring the privacy of individuals whose data is being analyzed by anonymizing student loan debt and homeownership details. Transparency about potential biases and limitations in the data and analysis will be maintained. Findings will be presented with clear explanations, avoiding sensationalism, and statistical methods will be used to provide context and avoid misleading conclusions.