**Analysis of Income Dynamics Using GSS Data**

Jessica Wang

PSY 8712

**Analysis of Income Dynamics Using GSS Data**

The purpose of this project is to examine key socio-economic factors influencing income dynamics and perceptions of income inequality within the General Social Survey (GSS). The analyses aim to discern patterns and predict outcomes based on demographic and socio-economic variables, employing statistical methods and machine learning techniques.

**Hypotheses and Research Questions**

RQ1: How well can machine learning models predict an individual's income based on age, education level, marital status, and number of children?

RQ2: What is the correlation between age and income across different genders?

RQ3: Do educational levels significantly influence perceptions of income inequality?

**Method**

**Open Science Materials**

Binder is an open-source online service that allows users to create custom computing environments that can be shared and used by many remote users. Anyone interested can view the static content of the research but also interact with the live Jupyter Notebooks used for the project. This approach enhances transparency and replicability by allowing others to execute the code and view the results in real-time. Binder environment for this project can be accessed via the following link: <https://mybinder.org/v2/gh/wan00967/psy8712-final.git/HEAD?urlpath=rstudio>.

GitHub is a platform for version control and collaboration. It allows multiple people to work together on projects from anywhere. For this project, GitHub has been used to host all the code, datasets, and documentation involved in the study. This ensures that all materials are accessible to the public, promoting an open and collaborative approach to science.

To access the study materials on GitHub:

1. Visit the GitHub repository at <https://github.com/wan00967/psy8712-final.git>.
2. The main page of the repository (the README file) includes an overview of the project, instructions on how to use the files, and links to the actual datasets and scripts used in the analyses.

**Participants**

The dataset consists of responses from participants in the 2022 General Social Survey. The unit of analysis primarily includes adult individuals spanning various age groups and socio-economic backgrounds across the United States.

**Measures**

Age: Participant's age.

Income: Total family income.

Educational Level: Highest degree obtained.

Marital Status: Current marital status.

Number of Children: Number of children.

**Procedure**

Participants provided responses through structured interviews, capturing diverse aspects of social life and economic conditions.

**Analyses**

**Descriptive Statistics and Static Visualizations**

**Table 1**

*Descriptive statistics*

|  |  |  |
| --- | --- | --- |
| **Variable** | **Mean** | **SD** |
| Age | 43.592 | 14.012 |
| Income | 11.482 | 1.697 |

**Figure 1**

**A graph of a graph showing the age and income

Description automatically generated**

**Interactive Visualization**

An interactive dashboard developed in Shiny allows users to explore how income correlates with age across different demographic variables, enhancing understanding of demographic impacts on income.

**Data Cleaning**

The data from the 2022 General Social Survey was cleaned, focusing on keeping variables of interest like work status, age, education, marital status, number of children, perceptions of wealth equality, happiness, and income, discarding any records with missing values. Variables were appropriately converted—categorical variables to factors with explicit levels, and continuous variables like age and income to numeric types—to facilitate later analyses.

**Analysis**

RQ1: Machine learning models varied in predictive accuracy, generally very low, with a random forest model showing the highest accuracy in cross-validation.

Table 2

*Machine learning results*

|  |  |  |
| --- | --- | --- |
| Algo | cv\_rsq | ho\_rsq |
| lm | 0.16 | 0.10 |
| glmnet | 0.17 | 0.10 |
| ranger | 0.67 | 0.10 |
| xgbTree | 0.70 | 0.05 |

RQ2: Correlation analysis showed a very weak positive correlation for males (r = 0.0316) and a negligible negative correlation for females (r = -0.0112).

**Table 3**

*Correlations between age and income by sex*

|  |  |
| --- | --- |
| Sex | Correlation |
| Male | 0.03 |
| Female | -0.01 |

RQ3: ANOVA testing indicated no significant differences in perceptions of income inequality across educational levels (p = 0.448).

**Figure 2**

A graph of a bar graph

Description automatically generated with medium confidence

**Reflection**

During this class, I learned the importance of precise data cleaning and preparation, which are essential for ensuring the accuracy of subsequent statistical analyses. Moving forward, I will implement these procedures in my research. I also learned the utility of interactive visualizations, such as Shiny apps, in making data more accessible and interpretable. It could be useful to incorporate these tools in the future to improve data presentation to improve understanding and engagement. I also recognized the benefits of reproducible research practices, including the use of R Markdown and GitHub. I think the most valuable thing that I learned was how to structure the process of working with data. This makes everything much more organized and understandable for me and the people I could be working with.