7COM1079-0901-2024 - Team Research and Development Project

Final report title: ***Is there a difference in the proportions of income across different levels of education among adults in the USA?***

Group ID: A125

Dataset number: DS122

Prepared by: Christima Mary Suresh Kumar (23084053),

Punnoose Panoor Punnoose (23103096),

Mirshad Rahman Naduvilakkalathil (),

Shidhiq Ali (),

Jackson simethy()

University of Hertfordshire

Hatfield, 2024

Table of Contents

[1. Introduction 3](#_Toc187076961)

[1.1 Problem statement and research motivation 3](#_Toc187076962)

[1.2 The data set 3](#_Toc187076963)

[1.3 Research question 3](#_Toc187076964)

[1.4 Null hypothesis and alternative hypothesis (H0/H1) 3](#_Toc187076965)

[2. Background research 4](#_Toc187076966)

[2.1 Research papers 4](#_Toc187076967)

[2.2 Why RQ is of interest 4](#_Toc187076968)

[3. Visualization 5](#_Toc187076969)

[3.1 Appropriate plot for the RQ(50 words) 5](#_Toc187076970)

[3.2 Additional information relating to understanding the data 5](#_Toc187076971)

[3.3 Useful information for the data understanding 5](#_Toc187076972)

[4. Analysis 6](#_Toc187076973)

[4.1 Statistical test used to test the hypotheses and output 6](#_Toc187076974)

[4.2 The null hypothesis is rejected /not rejected based on the p-value 6](#_Toc187076975)

[5. Evaluation 6](#_Toc187076976)

[5.1 What Went Well 6](#_Toc187076977)

[5.2 Points for Improvement 7](#_Toc187076978)

[5.3 Group’s Time Management 8](#_Toc187076979)

[5.4 Project’s Overall Judgement 8](#_Toc187076980)

[6.Conclusions 8](#_Toc187076981)

[6.1 Results explained 8](#_Toc187076982)

[6.2 Interpretation of the results 9](#_Toc187076983)

[6.3 Reasons and/or implications for future work, limitations of your study 9](#_Toc187076984)

[7.References 9](#_Toc187076985)

[8.Appendices 9](#_Toc187076986)

# **1**. Introduction

## 1.1 Problem statement and research motivation

Income inequality is a persistent global issue, with disparities often linked to educational attainment. While education is widely recognized as a key driver of earning potential, the specific impact of different education levels on income proportions remains underexplored. Previous studies have shown that education significantly affects income levels, highlighting the need for further investigation (Chakrabarty, 2018), This study investigates the relationship between education and income among adults in the USA to uncover patterns that can inform policies aimed at reducing inequality. Understanding how education levels influence income distribution can provide valuable insights for policymakers, educators, and organizations, enabling targeted interventions to promote economic mobility and address disparities in wealth distribution more effectively.

## 1.2 The data set

The dataset, titled *adult income1.csv*, contains 31,948 entries and 12 columns. It captures demographic and employment-related attributes, including age, workclass, education level, marital status, occupation, and income classification (≤50K or >50K). The independent variable is education.num, while the dependent variable is income (nominal). This dataset offers a comprehensive view of factors influencing income, making it ideal for analysing relationships between education levels and income proportions. Its richness enables robust statistical analysis to address the research question effectively.

## 1.3 Research question

**Is there a difference in the proportions of income across different levels of education among adults in the USA?**

To answer this, a chi-square test will be conducted to examine the relationship between education levels and income proportions, using statistical analysis to identify significant patterns and differences.

## 1.4 Null hypothesis and alternative hypothesis (H0/H1)

**Null Hypothesis (H₀):** There is no significant difference in the proportions of income across different levels of education among adults in the USA. In other words, education levels do not influence the likelihood of an individual's income falling into the ≤50K or >50K categories.

**Alternative Hypothesis (H₁):** There is a significant difference in the proportions of income across different levels of education among adults in the USA. This implies that education levels are a determining factor in income classification, with variations in income proportions observed between individuals with different educational attainment levels

# 2. Background research

## 2.1 Research papers

The relationship between demographic factors and income has been extensively studied using machine learning techniques. Chakrabarty and Biswas utilized the UCI Adult Dataset and applied Gradient Boosting to classify individuals into income brackets (>50K or ≤50K) (Navoneel Chakrabarty, 2018). They achieved a high prediction accuracy of 88.16%, highlighting education and employment features as critical determinants of income. This work emphasizes the value of advanced machine learning algorithms in uncovering patterns in socioeconomic data. Similarly, various machine learning algorithms, including decision trees, to analyse the Adult Census Income Dataset (Chet Lemon, 2018). His research underscored the importance of education and occupation as significant factors in predicting higher income levels, with a focus on improving data preprocessing techniques to enhance model performance. Lemon, Zelazo, and Mulakaluri also investigated the dataset using Naïve Bayes, Logistic Regression, and Decision Trees. They found Decision Trees to be the most effective, identifying education and hours worked per week as the most influential predictors of income. These studies collectively demonstrate the critical role of education in income prediction while highlighting gaps in evaluating proportional differences across educational tiers. This underscores the need for further research into the nuanced relationship between education and income.

## 2.2 Why RQ is of interest

Despite numerous studies utilizing the UCI Adult Dataset, there remains a gap in fully understanding the nuanced impact of education on income levels. Prior research confirms education as a critical factor in income prediction, but limited work has been done to evaluate the proportional differences across specific educational tiers. This research aims to address this gap by investigating the question: “*Is there a difference in the proportions of income across different levels of education among adults in the USA?*”. Addressing this question is vital for crafting policies aimed at reducing income inequality and enhancing economic mobility. Future directions include integrating longitudinal data to analyse trends over time and refining models to account for complex, non-linear relationships between education and income.

# 3. Visualization

## 3.1 Appropriate plot for the RQ(**50 words)**

A stacked bar plot is appropriate as it visualizes the difference in proportions of income categories (<=50K and >50K) across education levels. It allows direct comparison within each level, showing relative contributions of each category. The stacked structure highlights income distribution changes, aligning with the research question’s focus on proportion

A graph of a graph of income

Description automatically generated with medium confidence

## 3.2 Additional information relating to understanding the data

The plot visually highlights how higher education levels are associated with a larger proportion of individuals earning >50K. For example, individuals with a Doctorate or Master’s degree predominantly fall in the >50K income category, while lower education levels such as 10th, 11th, and Preschool have a majority earning <=50K.

## 3.3 Useful information for the data understanding

Key observations include a clear upward trend in income with increasing education level. Advanced degrees (e.g., Doctorate, Master’s) show the highest proportions of >50K incomes, whereas lower education levels have minimal representation in this category. Intermediate levels like bachelor's and associate degrees show a more balanced distribution of income.

# 4. Analysis

## 4.1 Statistical test used to test the hypotheses and output

To test the hypotheses, a **Chi-Squared Test of Independence** was conducted. This test is appropriate as both variables, **income** and **education level**, are categorical. The test evaluates whether there is a significant association between income categories (<=50K and >50K) and education levels.

The test produced the following output:

• **Chi-squared statistic**: 4352.3

• **Degrees of freedom**: 15

• **P-value**: < 2.2e-16

## 4.2 The null hypothesis is rejected /not rejected based on the p-value

Based on the p-value (< 2.2e-16), the null hypothesis is **rejected**. This indicates a statistically significant difference in proportions of income across different education levels. Thus, income distribution is not independent of education level.

# 5. Evaluation

## 5.1 What Went Well

Our teams have been excellent in solving challenges with data visualization done using R programming. Lab schedule conflicts presented a lot of planning involved and tools such as Trello and GitHub came in handy. Trello kept accountability and task management clear. GitHub was used to manage repositories for seamless collaboration. Each member's unique skills contributed to the quality of the project while the team tackled challenges with a growth mindset. It is the project that shows the right balance of technical complexity and teamwork-the success of planning and shared commitment.

## 5.2 Points for Improvement

Time conflicts hindered scheduling, despite virtual meetings on Google Meet. Recording sessions and centralizing meeting notes could improve efficiency. Expanding Slack beyond tutor communication to include team discussions and shared channels would streamline updates and feedback. R programming’s steep learning curve slowed progress; earlier workshops, curated resources, and collaborative coding sessions could ease adaptation. Lastly, inconsistent GitHub commit messages caused confusion; clearer guidelines for detailed commit documentation would improve code review and traceability.

## 5.3 Group’s Time Management

The phases of data collection, visualization, and testing broke the project into smaller bits, which made it possible for the team to manage time optimally. Trello and GitHub took care of efficient task-tracking and collaborative efforts and combined with regular updates and flexible meetings to ensure that progress continues. Although there were some minor last-minute bug fixes, all the milestones were met in time through persistent hard work and good teamwork.

## 5.4 Project’s Overall Judgement

The project reached its milestones by realizing effective visualisation of results from a well-functioning user-friendly R program that addressed the problem efficiently. Trello and GitHub have also made a significant contribution to task management and collaborative undertaking between members, while proactive communication via Google Meet was instrumental in effective coordination. It is clear from the output: technical excellence, collaborative teamwork, and positive commitment to the quality process right through.

# 6.Conclusions

## 6.1 Results explained

Analysis of the dataset revealed significant differences in income proportions between educational levels. Higher income groups (>50K) demonstrated a substantial association with higher educational accomplishment, whilst lower education levels tended to fall into the ≤50K category. Those with a doctorate or master's degree, for instance, are stacked toward the higher wage brackets. The chi-square test provided statistical support for these, confirming the study's hypothesis regarding the relationship between education and income distribution.

## 6.2 Interpretation of the results

These Results highlight how important education is in shaping adult income distribution in the United States. According to the findings, strategies that increase access to higher education may help to reduce income inequality. Targeted interventions are crucial to closing the income gap for groups with few educational options. In a larger sense, the study highlights the important connection between economic mobility and educational achievement, highlighting education as a tool for lowering the income gap.

## 6.3 Reasons and/or implications for future work, limitations of your study

The study's accuracy in evaluating the complex financial effects of schooling is limited by its dependence on categorical income data. Future studies might include other factors like experience and employment sector. Adding continuous information would further enhance our understanding of the long-term economic impacts of schooling by revealing trends in income over time.

# 7.References

Chakrabarty, S. B. (2018). A Statistical Approach to Adult Census Income. Greater Noida: IEEE.

Chet Lemon, C. Z. (2018). *Predicting if income exceeds $50,000 per year based on 1994 US Census Data with Simple Classification Techniques.*

Garn SM, B. S. (1977). Level of education, level of income, and level of fatness in adults. (pp. 721-725). ScienceDirect.

Navoneel Chakrabarty, S. B. (2018). *A Statistical Approach to Adult Census Income Level Prediction.*

Paul C. Glick, H. P. (1956). Educational Level and Potential Income. *American Sociological Review*, 307-312.

# 8.Appendices

1. *# Load library*

*library(ggplot2)*

*# Load the data*

*data <- read.csv("adult income1.csv")*

*# Filter relevant columns*

*filtered\_data <- data[, c("education", "income")]*

*# Summarize data to calculate proportions*

*proportions <- as.data.frame(prop.table(table(filtered\_data$education, filtered\_data$income), margin = 1))*

*colnames(proportions) <- c("Education", "Income", "Proportion")*

*# Create a stacked bar plot*

*ggplot(proportions, aes(x = Education, y = Proportion, fill = Income)) +*

*geom\_bar(stat = "identity", position = "stack") +*

*theme\_minimal() +*

*labs(title = "Proportion of Income by Education Level",*

*x = "Education Level",*

*y = "Proportion",*

*fill = "Income") +*

*theme(axis.text.x = element\_text(angle = 45, hjust = 1))*

*# Perform a chi-squared test*

*chi\_test <- chisq.test(table(filtered\_data$education, filtered\_data$income))*

*print(chi\_test).*