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).*

1. commit 3c0ff30717d6c6b1e5141ae92b51c34d3e329fc9

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Tue Jan 7 14:26:16 2025 +0000

Edits in evaluation and final revision

commit 0223600559a76670e0a275d80e4092124a2815ea

Author: js23agd <js23agd@herts.ac.uk>

Date: Tue Jan 7 13:03:58 2025 +0000

Evaluation word count reduced

commit 884be89ab1ec66b5bb4108f5b94c8b90057bbc80

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 17:32:53 2025 +0000

adding table of contents and formatting the final report doc

commit 58e4e5f392274d7834a265f3c6a57fdf4ae2e2e1

Merge: e85ac48 a569266

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Mon Jan 6 22:33:33 2025 +0530

Merge branch 'main' of https://github.com/cs24aba/A125-Team-Research-and-Development-Project

commit e85ac4891b4273fe1e4bafe7b223e841df41ec9a

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Mon Jan 6 22:33:17 2025 +0530

compiling the final report document

commit a5692660d4785cc30756e1bc6e751b7e57ad2b91

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 16:09:29 2025 +0000

fixes and rewrites in conclusion

commit 56a6e9ef25d5d3b455e4ecf080441a56b60966ff

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 15:58:43 2025 +0000

added analysis section

commit 887869a8eda4f0c8d8f22d8801751a64ca5bfdc8

Merge: afcd4e8 12d0696

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 15:42:59 2025 +0000

Merge branch 'main' of https://github.com/cs24aba/A125-Team-Research-and-Development-Project

commit afcd4e8ae4e322f9e3aef322e5f762788d264946

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 15:42:41 2025 +0000

fixed barplot image in readme

commit 12d06960f6c5b85bda6a44ac20e9d0ab95be6073

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Mon Jan 6 21:08:45 2025 +0530

final report file added

commit 7af4340301acd694af4b0a83871747d7d9d1ffe3

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 15:36:09 2025 +0000

slight changes in the report

commit 7161c53773acf78195643a427fce13c5a35e18b2

Merge: 9f2c821 74f1697

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 15:34:29 2025 +0000

Merge branch 'main' of https://github.com/cs24aba/A125-Team-Research-and-Development-Project

commit 74f1697995020fde2ee8988425ec0340bb50120d

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Mon Jan 6 21:00:56 2025 +0530

background research updated

commit 9f2c82104c763b5263c43bc5e588bfe840177653

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Mon Jan 6 15:30:08 2025 +0000

making edits on evaluation

commit c034838127b51a002bcec592d1ca17f568996a68

Merge: 645e5a0 d5095c6

Author: mn23adn <mn23adn@herts.ac.uk>

Date: Sun Jan 5 19:17:41 2025 +0000

uploaded background research and RQ interest section for final report

commit d5095c6f6da80bb748cf9d3dfae083aa5b9e75fc

Author: SIDHIQ ALI <sidhiqsidhiqali@gmail.com>

Date: Sun Jan 5 16:50:06 2025 +0000

Add files via upload

Adding the conclusion for the report

commit 1d827352cbb0914742779ad78bdf635dfdda243e

Author: js23agd <js23agd@herts.ac.uk>

Date: Sun Jan 5 16:27:11 2025 +0000

Evaluation section added

commit f46be2409424416cb973e1e83025015ef1ed97f2

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Fri Jan 3 18:48:33 2025 +0530

introduction added to the report

commit a18bd5210727e7d0bd7b734b92191b4ea24ccadf

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Thu Jan 2 16:23:46 2025 +0000

Added section 3 (visualization) of the report

commit 26f1088896e5432796504d127c9100382f673492

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Tue Dec 24 13:50:12 2024 +0000

fixing git tracking issue

commit 62037baa9f8eb1a4f935834ae8a9ac75924ccbe1

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Tue Dec 24 13:40:27 2024 +0000

added a list of literature resource relevant to the rq

commit 2f410f67c3faec4303ab0f0cc29400648af62550

Author: cs24aba <cs24aba@herts.ac.uk>

Date: Tue Dec 24 13:15:24 2024 +0000

reverting the R script and plot back to the previous one without percentage labels

commit e24b7770a816ee91772c84ea69553785b37a5968

Merge: 1ba495e ec16241

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Fri Dec 13 19:06:23 2024 +0530

Merge branch 'main' of https://github.com/cs24aba/A125-Team-Research-and-Development-Project

commit 1ba495e6db26a315d692343ba5f6ddb5c7ea1ac8

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Fri Dec 13 19:05:53 2024 +0530

report document added

commit ec162419f261ef78f7739a166db104e5868b4f02

Merge: 5e992dd df12c6f

Author: SIDHIQ ALI <zs24act@herts.ac.uk>

Date: Sat Dec 7 16:04:25 2024 +0000

Merge pull request #2 from cs24aba/zs24act

updated the stacked bar plot with the labeled image

commit df12c6f15a014859248d9649dd6a6712ac85b31d

Author: zs24act <zs24act@herts.ac.uk>

Date: Sat Dec 7 16:03:29 2024 +0000

updated the stacked bar plot with the labeled image

commit 5e992ddb6210f225b124f552a53a0a7a6bd4e02e

Merge: 5f53747 7534aae

Author: SIDHIQ ALI <zs24act@herts.ac.uk>

Date: Sat Dec 7 15:58:13 2024 +0000

Merge pull request #1 from cs24aba/zs24act

Added percentage labels to stack bar plot to improve the readability

commit 7534aaeecf21e2f018e41b59bdd22aa6323d82bb

Author: zs24act <zs24act@herts.ac.uk>

Date: Sat Dec 7 15:56:52 2024 +0000

Added percentage labels to stack bar plot to improve the readability

commit 645e5a03d83f01c7db14abbef752332539da4b3d

Author: mn23adn <mn23adn@herts.ac.uk>

Date: Mon Dec 2 15:42:13 2024 +0000

updated readme file with visualization method used in the analysis.

commit 5f53747638acecb43ffdd35ac1999c94437d5955

Author: mn23adn <mn23adn@herts.ac.uk>

Date: Mon Dec 2 15:42:13 2024 +0000

updated readme file with visualization method used in the analysis.

commit 5383d21fac9639792c2b4cef905d38df0cd1ab4d

Author: mn23adn <mn23adn@herts.ac.uk>

Date: Mon Dec 2 15:30:01 2024 +0000

added visualization presentation

commit d0148061b46d1533992ade411267259ff6ecc4fb

Author: mn23adn <mn23adn@herts.ac.uk>

Date: Mon Dec 2 15:30:01 2024 +0000

added visualization presentation

commit 55afc6e27e14d056b53005a06ee90e99297a11c2

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Wed Nov 27 20:57:46 2024 +0530

old rq folder updated

commit 06276f75bac7e94c908b3ce9c871683354886d93

Author: pp24abe <pp24abe@herts.ac.uk>

Date: Wed Nov 27 20:57:46 2024 +0530

old rq folder updated

commit 4584941af99fd04a426e8df35e4fe0bcae0ae486

Author: rayan <ppp.csb1923@saintgits.org>

Date: Wed Nov 27 20:39:19 2024 +0530

added old rq and diagrams in oldrq folder

commit ad9c9fe3a75006a2b5d56eb1528e5c683ecce35c

Author: rayan <ppp.csb1923@saintgits.org>

Date: Wed Nov 27 20:39:19 2024 +0530

added old rq and diagrams in oldrq folder

commit 4b42d798f7cae3c095e057f270ba722b39e7574c

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 20:25:52 2024 +0000

added images and updated readme with detailed project info

commit 22e201e71e307f065a1447a3b555dd374f99dfa9

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 20:25:52 2024 +0000

added images and updated readme with detailed project info

commit 678c6959340ad869c27494e737b061fe4924d6bf

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 19:05:27 2024 +0000

added .DS\_store to gitignore

commit e77fd1d4c6e765755f2f2e3fb39156cdea03f251

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 19:05:27 2024 +0000

added .DS\_store to gitignore

commit ca0a7ed88a21c559507f543b9b1650ffacf150bc

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 18:57:02 2024 +0000

stacked bar plot to visualize the proportions of income across different education levels

commit 9311ee90b2b0a28f32167dab1ae5289e5f186655

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 18:57:02 2024 +0000

stacked bar plot to visualize the proportions of income across different education levels

commit 2d755da641943b806b56d57e2eacd44e92dec54b

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 17:55:42 2024 +0000

added finalized research question presentation

commit 83e4794825abb8b881b2a48966ca91561e187307

Author: christima <cs24aba@herts.ac.uk>

Date: Thu Nov 21 17:55:42 2024 +0000

added finalized research question presentation

commit 72dcf59ae70eab9356c84d876ccc853c948a3b48

Author: js23agd <js23agd@herts.ac.uk>

Date: Tue Nov 5 04:07:15 2024 +0000

RQ presentation slides

commit fa544ac339f664da7016d5d3c84e137e061c28c9

Author: js23agd <js23agd@herts.ac.uk>

Date: Tue Nov 5 04:07:15 2024 +0000

RQ presentation slides

commit f09c5a0c7c3e857370b6dde9ba344a029a9f2645

Author: christima <cs24aba@herts.ac.uk>

Date: Sat Oct 26 23:00:22 2024 +0100

adding allocated dataset