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

Final report title: (*the topic of your research.)*

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 *add page numbers here*

1. Introduction
   1. Problem statement and research motivation
   2. The data set
   3. Research question
   4. Null hypothesis and alternative hypothesis (H0/H1)

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS)
   2. Why RQ is of interest (research gap and future directions according to the literature)

1. Visualisation
   1. Appropriate plot for the RQ *output of an R script (NOT a screenshot)*
   2. Additional information relating to understanding the data (optional)
   3. Useful information for the data understanding

1. Analysis
   1. Statistical test used to test the hypotheses and output
   2. The null hypothesis is rejected /not rejected based on the p-value

1. Evaluation – group’s experience at 7COM1079
   1. What went well
   2. Points for improvement
   3. Group’s time management
   4. Project’s overall judgement
   5. Comment on GitHub log output

1. Conclusions
   1. Results explained.
   2. Interpretation of the results
   3. Reasons and/or implications for future work, limitations of your stud

1. Reference list

Harvard (author, date) format.

1. Appendices
2. R code used for analysis and visualisation.
3. GitHub log output.

**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**

* 1. Appropriate plot for the RQ(**50 words)**

A graph of a graph of income

Description automatically generated with medium confidence

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

* 1. **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.

* 1. **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**

* 1. **What Went Well**

Our team has achieved significant milestones in collaborative research projects, particularly in leveraging R programming to address data visualization challenges. Despite the potential challenges posed by differing laboratory schedules, effective communication and coordination were seamlessly achieved through meticulous planning and the strategic use of collaborative tools.

One of the critical tools that underpinned our workflow was Trello, which served as the backbone of our planning and task management processes. From clearly defining project milestones to assigning responsibilities and establishing deadlines, Trello enabled us to maintain organization and focus. By delineating individual tasks, we ensured that each team member had clear ownership of their responsibilities while collectively contributing to the overarching goals of the project.

GitHub was equally instrumental in the success of our project, providing a robust platform for the development, testing, and refinement of our codebase. Through diligent repository management, we avoided code conflicts during merges, ensuring that our repository remained clean, functional, and well-documented. Every contribution was carefully reviewed and integrated with feedback from team members, exemplifying our collaborative synergy even when working remotely.

The project was a testament to the collective efforts and diverse skill sets of the team. Each member brought unique expertise that not only enhanced the quality of the code but also contributed to a technically superior and visually compelling output. Challenges encountered throughout the project were met with a proactive and solution-oriented mindset, fostering an environment of growth and learning.

In summary, this project highlights our ability to balance technical complexity with effective teamwork. It underscores the importance of meticulous planning, clear communication, and a shared commitment to excellence, all of which were pivotal to our success.

**5.2 Points for Improvement**

While the project overall ended up as a good success story, certain points surely provided an avenue for improvement:

**Time Conflicts:** Aligning meetings among team members to non-existent lab hours would remain one of the greatest challenges. However, most of this was bridged by the virtual discussions held within Google Meet. That flexibility with using virtual meetings made possible scenarios where team members could connect from different locations and discuss critical aspects of the project. It might still be quite little with time overlaps, making some final decisions slow. In future projects, we may augment this system even further to help bring everyone up to speed by recording the Google Meet sessions for those not present, or by centralizing shared documentation on meeting notes.

Slack, which had been primarily used for tutor communications, could create a whole new layer for communication and support - something that would allow us to get inputs or feedback quickly, clear up doubts, or share updates with tutors. A future vision would expand Slack even more to include asynchronous teams discussion or even some shared channels to add another layer of communication with Google Meet, as it would also suffice for some settings.

**Initial Learning Curve:** R programming had one of the steepest learning curves, and the visualization libraries posed the most challenge to some of the team members who have been introduced to those tools. It caused a slowdown, even though, like magic, the group eventually adapted through personal efforts and support from the colleagues. The tutors on Slack helped to address many technical queries, but it could have been more proactive in addressing the learning curve with dedicated workshops or collaborative coding sessions scheduled earlier into the timeline of the project. Curated resources such as an already gathered collection of tutorials and documentation shared in team meetings or on Slack would form another effective basis for building confidence among the members.

**Commit Documentation:** GitHub was a very helpful tool to work with in getting the project repository well organized yet commit documentation had some inconsistencies leading to some confuse in it at times. As a team, we were able to take full advantage of GitHub for version control. However, some commit messages actually did not contain sufficient detail about the reasons for making those changes or the context of those edits that makes code review painful. Clear and comprehensive guidelines focused on writing commit messages would better clarify and trace changes.

**5.3 Group’s Time Management**

Our team exercised an excellent manage of time by following the project schedule and milestone achievement dates. The Trello boards which were effective for distributing tasks, clarifying roles, and setting deliverable deadlines were systematic so that all remained on the same page, notwithstanding the differences in the lab schedule. Updates and task watching on the Trello platform gave everyone accountability and shared transparency as we organized ourselves toward our targeted goals.

Some delays were caused by unknown bugs and integration troubles most especially during code contributions merging of bits. Such delays were fairly resolved without them majoring on the overall timeline using collaborative problem solving. Timely communication and decision-making were further strengthened by the use of Google Meet. We could certainly use a buffer period in future projects to enhance flexibility and fortify the team even better in managing its time.

**5.4 Project’s Overall Judgement**

This project successfully achieved all its objectives, delivering valuable research outputs characterized by striking visualizations. The R program we developed is not only highly functional but also user-friendly, effectively and precisely addressing the identified visualization challenges. The visual outputs provide meaningful interpretations of the data while exemplifying the team’s technical expertise and commitment to excellence.

The success of this project was largely driven by teamwork and the strategic use of collaboration tools. Trello provided a structured framework for planning and task management, while GitHub facilitated seamless version control and efficient integration of contributions. These tools, complemented by proactive communication through Google Meet, ensured consistent coordination and productivity throughout the project’s lifecycle. As a result, the final output is cohesive, technically robust, and a testament to both the team’s collaborative efforts and the efficiency of the workflow.

**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).**