*Please delete all the cursive text before submission. It is here just for your reference*.

*Further: data set – DS, research question – RQ*

*The mark (****x words****) after each subchapter states the word count limit. This indicates the expected amount of information which you can exceed by 10% without losing the mark.*

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

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

Group ID:

Dataset number:

Prepared by: *[Name and ID of submitting student first],*

*[Name and ID of other group members]*

***Please make sure*** *the document spelled correctly (including image labels, section headings, and table of contents). Please use correct punctuation.*

*Make sure your report is grammatically correct.*

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.

*The list below outlines the chapter/subchapter numbers, names, word count limits, and explanations of what to write in each section.*

1. Introduction

* 1. Problem statement and research motivation **(100 words)**
* *What is the problem in the area we want to learn more about (motivation for study).*
* *Use at least one citation from related literature for top marks.*
  1. The data set **(75 words)**
  2. Research question **(50 words).** *Explain how you are going to answer your RQ.* **(50 words)**
  3. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**

1. Background research
   1. Research papers (at least 3 relevant to your topic / DS) **(200 words)**

* *Was the data set used for some research papers?* *Reference at least 3 relevant research papers to your topic / DS.*
  1. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

1. Visualisation
   1. Appropriate plot for the RQ *output of an R script (NOT a screenshot)* (**50 words)**

* *Explain the choice of the plot.*
* *Anything on the plot from R is not counted towards word count limit*
* *Make sure that the plot is from output of an R script (NOT a screenshot)*
* *Make sure that the plot has a caption or title, X and Y-axis labels, legend if appropriate and units.*
* *Make sure the title or caption and axis labels are informative.*
  1. Additional information relating to understanding the data (optional) (**50 words)**
* *Per plot: explain the purpose and insights.*
  1. Useful information for the data understanding (**50 words)**
* *Summarise key observations from the plot.*

1. Analysis
   1. Statistical test used to test the hypotheses and output (**75 words)**

* *Explain the choice of the test.*
* *Make sure the test is appropriate for the RQ and data.*
  1. The null hypothesis is rejected /not rejected based on the p-value (**100 words)** *(interpret the results)*

1. Evaluation – group’s experience at 7COM1079
   1. What went well **(75 words)**
   2. Points for improvement **(75 words)**
   3. Group’s time management (**50 words)**
   4. Project’s overall judgement (**50 words)**
   5. Note any changes to group since submission of Assignment 1. Add new or amended GitHub Ids for new members **(75 words, write only if applies to your group arrangements)**
   6. Comment on the GitHub log output **(50 words)**

*Please comment on the GitHub log output, and refer to it as being placed into**Appendix B.*

*From your Git log, select the three most significant commits during this project and include the following for each:*

1. ***Commit Message:*** *[Insert Commit Message] Brief explanation of the broader impact of the change*
2. ***Commit Message:*** *[Insert Commit Message] Brief explanation of the broader impact of the change*
3. ***Commit Message:*** *[Insert Commit Message] Brief explanation of the broader impact of the change*

1. Conclusions
   1. Results explained (**75 words)**
   2. Interpretation of the results (**75 words)**

* *Interpretation of what the results mean in terms of your RQ and the effect this may have on your population and the wider context of your topic.*
  1. Reasons and/or implications for future work, limitations of your study (**50 words)**

1. Reference list ***(not included in the work count)***

Harvard (author, date) format.

1. Appendices
2. R code used for analysis and visualisation ***(not included in the word count)***

Analysis.R code with the appropriate statistics to test the hypotheses.

* ***No word count****, but ensure the code is without redundant lines, well-commented and produces the correct output.*
* *Make sure it runs (look in Rscript.log for output from a statistical test)*
* *It should compute appropriate statistics to test the hypotheses*

1. 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 (ordinal), 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.