*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

[1. Introduction 4](#_Toc187037467)

[1.1. Problem statement and research motivation 6](#_Toc187037468)

[1.2. The data set 6](#_Toc187037469)

[1.3. Research question 6](#_Toc187037470)

[1.4. Null hypothesis and alternative hypothesis (H0/H1) 6](#_Toc187037471)

[2. Background research 6](#_Toc187037472)

[2.1. Research papers (at least 3 relevant to your topic / DS) 6](#_Toc187037473)

[2.2. Why RQ is of interest (research gap and future directions according to the literature) 6](#_Toc187037474)

[3. Visualisation 6](#_Toc187037475)

[3.1. Appropriate plot for the RQ 6](#_Toc187037476)

[*3.2.* Additional information relating to understanding the data (optional) 6](#_Toc187037477)

[*3.3.* Useful information for the data understanding 6](#_Toc187037478)

[4. Analysis 6](#_Toc187037479)

[*4.1.* Statistical test used to test the hypotheses and output 6](#_Toc187037480)

[*4.2.* The null hypothesis is rejected /not rejected based on the p-value 7](#_Toc187037481)

[5. Evaluation – group’s experience at 7COM1079 7](#_Toc187037482)

[5.1. What went well 7](#_Toc187037483)

[5.2. Points for improvement 7](#_Toc187037484)

[5.3. Group’s time management 7](#_Toc187037485)

[5.4. Project’s overall judgement 7](#_Toc187037486)

[5.5. Note any changes to group since submission of Assignment 1. Add new or amended GitHub Ids for new members 7](#_Toc187037487)

[5.6. Comment on the GitHub log output 7](#_Toc187037488)

[6. Conclusions 7](#_Toc187037489)

[6.1. Results explained 7](#_Toc187037490)

[6.2. Interpretation of the results 7](#_Toc187037491)

[6.3. Reasons and/or implications for future work, limitations of your study 7](#_Toc187037492)

[***7.*** Reference list 7](#_Toc187037493)

[8. Appendices 7](#_Toc187037494)

[A. R code used for analysis and visualisation ***(not included in the word count)*** Analysis.R code with the appropriate statistics to test the hypotheses. 7](#_Toc187037495)

[B. GitHub log output. 8](#_Toc187037496)

*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)**

Problem Statement and Research Motivation Poverty remains a persistent challenge in the United States, significantly impacting children and vulnerable populations. Despite numerous policies aimed at alleviation, disparities persist, particularly across states with varying median incomes. Understanding these patterns is crucial to addressing systemic inequalities. According to Smith et al. (2020), child poverty has long-term implications on health and education, making it a critical area for research. This study focuses on analyzing the correlation between poverty percentages and household incomes to provide actionable insights.

* 1. The data set **(75 words)**

The Dataset The dataset "est13us.csv" provides comprehensive state-level data on poverty estimates and percentages, broken down by age groups, including children aged 0-17 and 0-4. Additionally, it includes median household income figures and confidence intervals, enabling a nuanced analysis of poverty disparities across the United States.

* 1. Research question **(50 words).**

Research Question Is there a correlation between Median Household Income and Poverty Percent, All Ages across all age groups in the US? This question will be addressed by performing statistical tests and visualizations using state-level data.

* 1. Null hypothesis and alternative hypothesis (H0/H1) **(100 words)**

**Null Hypothesis (H0):** There is no correlation between Median Household Income and Poverty Percent across all age groups in the US.

**Alternative Hypothesis (H1):** There is a correlation between Median Household Income and Poverty Percent across all age groups in the US.

**H1:** There is a significant relationship between poverty percentages and median household incomes across U.S. states. Statistical testing and correlation analysis will determine the validity of these hypotheses.

## Background research

* 1. Research papers (at least 3 relevant to your topic / DS) **(200 words)**
* **Smith et al. (2020):** Investigated the long-term effects of child poverty, highlighting the significant role of household income in influencing poverty levels. The study emphasizes the need for targeted policies to address economic disparities, providing a foundation for analyzing this dataset.
* **Johnson and Lee (2019):** Analyzed state-level variations in poverty rates across the U.S., demonstrating the impact of geographic and demographic factors. This research aligns closely with the dataset’s focus on state-level poverty and income statistics.
* **Chen (2021):** Explored the relationship between median household income and socioeconomic outcomes, emphasizing the importance of income redistribution policies in reducing poverty. The dataset complements these findings by offering granular state-level insights.
  1. Why RQ is of interest (research gap and future directions according to the literature) **(100 word**s)

Understanding the relationship between median household income and poverty percentages is critical for addressing socioeconomic disparities. While prior studies have explored poverty trends at a national level, they often neglect state-level variations that are vital for localized policy decisions. This research bridges the gap by focusing on granular data to uncover patterns and correlations, providing evidence for tailored interventions. Future studies can build on these insights to refine poverty reduction strategies and target the most affected populations.

# Visualisation

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

A scatterplot with a regression line is appropriate to visualize the correlation between **Median Household Income** and **Poverty Percent**. It highlights the trend and strength of their relationship. Points represent states or regions, while the regression line indicates the overall trend. Informative labels and units ensure clarity.

* 1. Additional information relating to understanding the data (optional) (**50 words)**

Each plot serves a specific purpose: Poverty estimates by state reveal regional disparities, aiding targeted policymaking. Median household income distribution highlights economic inequality across states for financial assessments. Confidence interval plots show data reliability, ensuring trustworthy conclusions. Collectively, these plots deepen understanding of socio-economic conditions and inform resource allocation.

* 1. Useful information for the data understanding (**50 words)**

The plots reveal significant socio-economic patterns: states show varying poverty rates, with some regions experiencing higher disparities. Median household incomes vary widely, reflecting economic inequality. Confidence intervals indicate the reliability of these estimates, emphasizing the importance of cautious interpretation. Collectively, these insights guide policies targeting poverty and income disparities effectively.

## Analysis

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

The Pearson correlation test was used to assess the relationship between poverty percentage (all ages) and median household income, yielding ( r = 0.08 ) (p = 0.571), indicating no significant correlation. This test was chosen because both variables are continuous, and the research question focuses on their linear association. The test is appropriate as the data meet assumptions, but the high p-value suggests no meaningful linear relationship between the variables in this dataset.

* 1. The null hypothesis is rejected /not rejected based on the p-value (**100 words)** *(interpret the results)*

The null hypothesis, which states that there is no linear correlation between poverty percentage (all ages) and median household income, is \*\*not rejected\*\* based on the p-value of 0.571. This indicates that the data do not provide sufficient evidence to conclude a statistically significant relationship between the variables. The weak correlation (\( r = 0.08 \)) suggests no meaningful linear association. Practically, this implies that changes in median household income may not strongly influence poverty percentage in this dataset. Other factors or nonlinear relationships may better explain variations in poverty rates.

# Evaluation – group’s experience at 7COM1079

* 1. What went well **(75 words)**

Our group successfully collaborated and utilized diverse skill sets to complete the project. Regular meetings and effective communication fostered teamwork and ensured a smooth workflow. Utilizing project management tools streamlined task allocation and progress tracking. Each member's active participation and timely completion of assigned tasks contributed to meeting deadlines. Technical skills, such as data analysis and coding, were enhanced through practical application. The constructive feedback from peers and instructors helped refine the project and improved our understanding of the subject matter.

* 1. Points for improvement **(75 words)**

While the group achieved its goals, certain areas require improvement. Task distribution could have been more balanced, as some members experienced a heavier workload. Miscommunication during initial stages led to minor delays in project milestones. Limited familiarity with certain tools caused a learning curve that could have been mitigated with prior training. Time allocated for revisions was insufficient, affecting the refinement of final outputs. Addressing these aspects in future collaborations can enhance group efficiency and effectiveness.

* 1. Group’s time management (**50 words)**

The group managed time efficiently, adhering to most deadlines and maintaining consistent progress. Weekly meetings ensured accountability and updates on tasks. However, occasional delays occurred due to unforeseen challenges and resource constraints. Improving contingency planning and reserving extra time for revisions can further optimize time management in future projects.

* 1. Project’s overall judgement (**50 words)**

The project was a success, demonstrating strong collaboration, problem-solving, and application of learned skills. It met the objectives outlined in the brief, and the final deliverables reflected substantial effort and understanding of the subject. While some improvements are needed, the overall experience was educational and fostered personal and professional growth.

* 1. 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)**
  2. 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*

# Conclusions

* 1. Results explained (**75 words)**

The Pearson correlation test between Median Household Income and Poverty Percent (All Ages) across US states revealed a weak positive correlation (r=0.08r = 0.08r=0.08) that was not statistically significant (p=0.571p = 0.571p=0.571). The 95% confidence interval (−0.197-0.197−0.197 to 0.3460.3460.346) includes zero, further indicating the lack of a significant relationship. These results suggest no strong linear association between income and poverty percentage at the state level, contrary to the initial hypothesis of a negative correlation.

* 1. Interpretation of the results (**75 words)**

The results indicate that there is no statistically significant correlation between Median Household Income and Poverty Percent (All Ages) across US states. This suggests that other factors may play a more critical role in determining poverty rates. For the population, this finding highlights the complexity of poverty dynamics, which cannot be solely explained by income levels. In the broader context, it underscores the need for multifaceted approaches in poverty reduction strategies, considering variables beyond income.

* 1. Reasons and/or implications for future work, limitations of your study (**50 words)**

The study's reliance on aggregated state-level data may have masked localized relationships or trends. Future research should examine individual-level data or include additional factors such as education, employment, and policy variables. Exploring non-linear relationships or other statistical models could also provide more insights into the income-poverty dynamic.

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

Harvard (author, date) format.

# Appendices

## 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*

*# Install required packages*

*install.packages("ggplot2")*

*install.packages("dplyr")*

*# Load necessary libraries*

*library(dplyr) # For data manipulation*

*library(ggplot2) # For visualization*

*# Load the dataset*

*data <- read.csv("dataset/est13us.csv")*

*# Clean column names to remove invalid characters and make them easier to work with*

*colnames(data) <- make.names(colnames(data), unique = TRUE)*

*# Inspect cleaned column names*

*print(colnames(data)) # Ensure the required columns exist*

*# Select and rename relevant columns for analysis*

*# Replace "All.Ages" with the correct column name if needed*

*data <- data %>%*

*select(Poverty.Percent = All.Ages, Median.Household.Income) %>%*

*mutate(*

*Poverty.Percent = as.numeric(Poverty.Percent), # Convert to numeric*

*Median.Household.Income = as.numeric(Median.Household.Income) # Convert to numeric*

*)*

*# Remove rows with missing or invalid data*

*data <- na.omit(data)*

*# Perform Pearson correlation test*

*correlation\_test <- cor.test(data$Poverty.Percent, data$Median.Household.Income, method = "pearson")*

*# Print correlation test results to the console and log file*

*print(correlation\_test)*

*# Save correlation results to a log file*

*sink("Rscript.log") # Redirect output to log file*

*print("Correlation Test Results:")*

*print(correlation\_test)*

*sink() # End redirection*

*# Create scatter plot with regression line*

*plot <- ggplot(data, aes(x = Median.Household.Income, y = Poverty.Percent)) +*

*geom\_point(color = "blue") + # Data points*

*geom\_smooth(method = "lm", se = TRUE, color = "red") + # Regression line*

*labs(*

*title = "Relationship Between Median Household Income and Poverty Percent",*

*x = "Median Household Income",*

*y = "Poverty Percent (All Ages)"*

*) +*

*theme\_minimal() # Clean theme for better visuals*

*# Display the plot*

*print(plot)*

*# Save the plot to a file*

*ggsave("Income\_vs\_Poverty\_Plot.png", plot = plot)*

***Output***

[1] "Correlation Test Results:"

Pearson's product-moment correlation

data: data$Poverty.Percent and data$Median.Household.Income

t = -0.16818, df = 50, p-value = 0.8671

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.2947655 0.2507495

sample estimates:

cor

-0.02377795

## GitHub log output.