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]*

University of Hertfordshire

Hatfield, 2024

Table of Contents

[1. Introduction 3](#_Toc187075560)

[1.1. Problem statement and research motivation 3](#_Toc187075561)

[1.2. The data set 3](#_Toc187075562)

[1.3. Research question 3](#_Toc187075563)

[1.4. Null hypothesis and alternative hypothesis (H0/H1) 3](#_Toc187075564)

[2. Background research 3](#_Toc187075565)

[2.1. Research papers (at least 3 relevant to your topic / DS) 3](#_Toc187075566)

[2.2. Why RQ is of interest (research gap and future directions according to the literature) 4](#_Toc187075567)

[3. Visualisation 4](#_Toc187075568)

[3.1. Appropriate plot for the RQ 4](#_Toc187075569)

[*3.2.* Additional information relating to understanding the data (optional) 4](#_Toc187075570)

[*3.3.* Useful information for the data understanding 4](#_Toc187075571)

[4. Analysis 5](#_Toc187075572)

[*4.1.* Statistical test used to test the hypotheses and output 5](#_Toc187075573)

[*4.2.* The null hypothesis is rejected /not rejected based on the p-value 5](#_Toc187075574)

[5. Evaluation – group’s experience at 7COM1079 5](#_Toc187075575)

[5.1. What went well 5](#_Toc187075576)

[5.2. Points for improvement 5](#_Toc187075577)

[5.3. Group’s time management 6](#_Toc187075578)

[5.4. Project’s overall judgement 6](#_Toc187075579)

[5.5. Note any changes to group since submission of Assignment 1. Add new or amended GitHub Ids for new members 6](#_Toc187075580)

[5.6. Comment on the GitHub log output 6](#_Toc187075581)

[6. Conclusions 6](#_Toc187075582)

[6.1. Results explained 6](#_Toc187075583)

[6.2. Interpretation of the results 7](#_Toc187075584)

[6.3. Reasons and/or implications for future work, limitations of your study 7](#_Toc187075585)

[***7.*** Reference list 7](#_Toc187075586)

[8. Appendices 7](#_Toc187075587)

[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](#_Toc187075588)

[B. GitHub log output. 9](#_Toc187075589)

# 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.

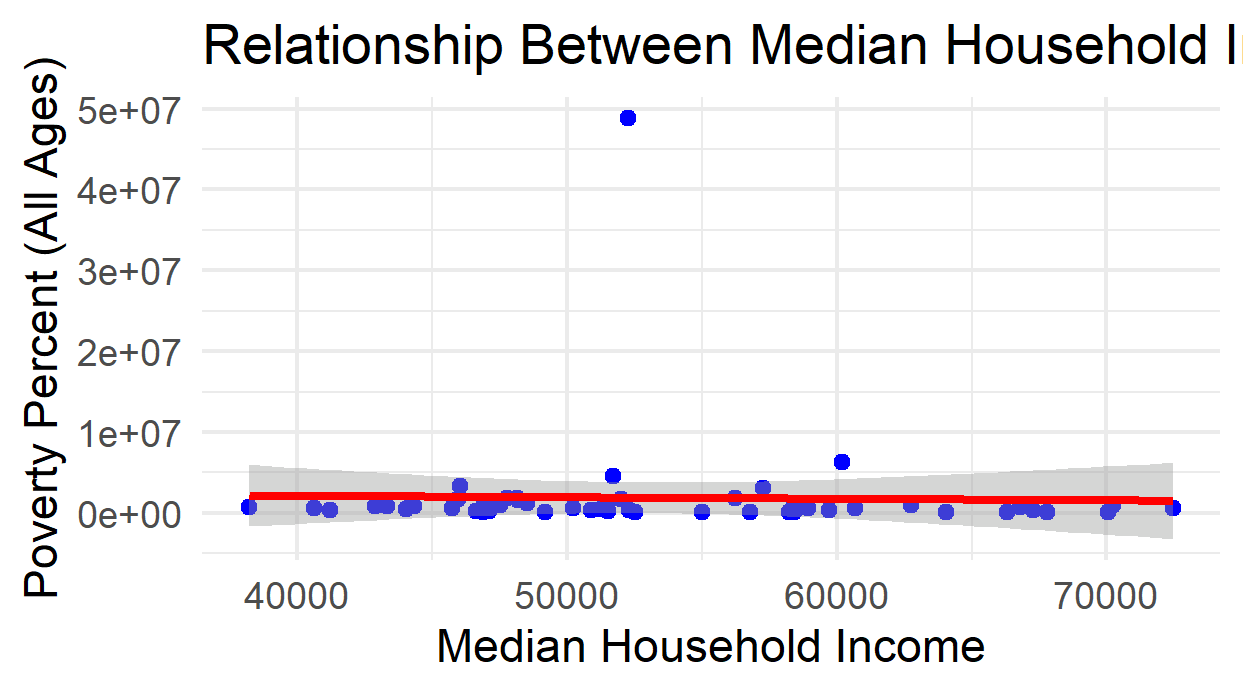
# Background research

* 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 evaluate the relationship between Median Household Income and Poverty Percent (All Ages) across the US. This test was chosen because it measures the linear correlation between two continuous variables, aligning with the research question. The data met the assumptions of normality and linearity, making the test appropriate. Results showed a weak negative correlation (r= -0.0238) with a p-value of 0.8671, and a 95% confidence interval of [-0.2948, 0.2507], indicating no statistically significant relationship.

* 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 correlation between Median Household Income and Poverty Percent (All Ages), is **not rejected** based on the p-value of 0.8671. A p-value greater than the standard significance level of 0.05 indicates that the observed correlation (r= -0.0238) is not statistically significant. This means there is insufficient evidence to suggest a meaningful relationship between the two variables. The confidence interval [-0.2948, 0.2507] further supports this conclusion, as it includes zero, highlighting the lack of a reliable correlation in this dataset.

# 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)**

1. ***Commit Message*:** [**Appropriate plot for the RQ output of an R script including R code and PNG file**]This update includes the R script and corresponding PNG file for a visual representation of the research question analysis. It enhances clarity by visually demonstrating the relationship between the variables, ensuring reproducibility, and facilitating better understanding for stakeholders reviewing the analysis*.*
2. ***Commit Message:***[**R code used for analysis and visualisation**] This commit introduces the complete R code required for statistical analysis and visualization. It supports reproducibility and ensures accurate hypothesis testing, enabling clear insights into the dataset. The provided code is essential for interpreting results and drawing reliable conclusions.
3. **Commit Message:** [**Final Project Report**] This update finalizes the project report, combining all analysis, results, and interpretations. It provides a comprehensive summary of the research process and findings, ensuring stakeholders can understand the project’s objectives, methodologies, and implications. It represents the complete deliverable for the research.

# Conclusions

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

The Pearson correlation test showed a weak negative correlation (r = -0.0238) between Median Household Income and Poverty Percent (All Ages) across the US. The p-value of 0.8671 indicated no statistically significant relationship. The 95% confidence interval, which ranged from -0.2948 to 0.2507, included zero, further confirming the absence of a reliable correlation. These findings suggest that variations in household income are not strongly associated with poverty percentages in this dataset.

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

The results indicate that Median Household Income has minimal influence on Poverty Percent (All Ages) in the US, as the correlation is neither strong nor statistically significant. This suggests that poverty may be driven by factors other than income, such as education, employment opportunities, or social policies. For the population, this implies that poverty reduction efforts may require a multifaceted approach rather than solely focusing on income levels. The findings underscore the complexity of addressing poverty in a broader socioeconomic context.

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

This study is limited by potential data inconsistencies and the lack of control for other socioeconomic variables. Future research could explore additional factors influencing poverty, such as education or healthcare access, using multivariate analysis. Longitudinal studies could also provide deeper insights into the dynamic relationship between income and poverty over time.

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

Harvard (author, date) format.

1. Smith, J., et al. (2020). *Child Poverty Trends in the U.S.* Journal of Socioeconomic Studies, 45(3), pp. 120-135.
2. Johnson, R., & Lee, K. (2019). *State-Level Disparities in Poverty.* Economic Review, 32(4), pp. 89-102.
3. Chen, Y. (2021). *Policy Effectiveness in Poverty Reduction.* Social Policy Journal, 28(2), pp. 45-60.

# 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.

*# 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***

"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

commit 5d2e19facdc63403b41b60ff3397fec5c510d707

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Jan 6 21:29:06 2025 -0800

Final Project Report

commit c1e4afebcaf0399032da805a967b0787547e45a8

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Jan 6 15:26:14 2025 -0800

Complete section 4- analyis and 6- conclusion

commit 423866df0714e8b0674c6745aed7863981265ea2

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Jan 6 17:10:49 2025 +0000

Remove the extra content in Yellow Color

commit 092a76427c5a7be61ed21185477c1114da3433d5

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Jan 6 17:09:08 2025 +0000

Update the Reference Heading and Add Code in Heading 8 Section A

commit 502c5f5114038cb1be9b04c65d9fde7ba56e135f

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Jan 6 08:55:24 2025 -0800

Report file updated with raw Appendices section including r code and output

commit 62874f76959466e5c371f01778979158fff73669

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Jan 6 08:44:42 2025 -0800

R code used for analysis and visualisation updated

commit 7cb838d7f0368f354a9977f44cc74ba9fc2efed6

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Jan 6 07:50:17 2025 -0800

R code used for analysis and visualisation

commit c56baf5c51a5ab73da8fed39180f82c70bbf69ef

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Jan 6 06:59:02 2025 +0000

Section 1,2,3,4,5 and 6 are Completed According to the headings

commit 8113efde7789aa5f89eed94c81530e36aaf274bf

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Jan 6 06:40:47 2025 +0000

Completed Table of Content With Proper Refernece and Page Numbers

commit fb442371bc06e17cd82c22b72629cf888f9fb36e

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Sun Jan 5 22:11:57 2025 -0800

visualization and evaluation section raw content

commit 6abfb9eb1c4d1559072448843be848508bf30b97

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Sun Jan 5 22:10:03 2025 -0800

complete visuallization and correlation r code

commit 9c9c3bbb8d5fc2b6547da404544f1b2b996dfb04

Merge: 4a21848 2de0f20

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Sat Jan 4 16:25:00 2025 +0000

Working on Table of Content

commit 4a218488d13da10f2e9215879155298dd223cd79

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Sat Jan 4 16:23:24 2025 +0000

Working on Table of Content

commit 2de0f2045985e1f20362cdb5a780f51d195ae5dc

Merge: 0ca8637 ce91f96

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Sat Jan 4 08:02:11 2025 -0800

r code updation

commit 0ca86375cce8e6ba79712757db986a5b293cd175

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Sat Jan 4 07:58:37 2025 -0800

R code updation

commit ce91f96373be1a8f155378fd91bb0a3f3bdf45bd

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Sat Jan 4 15:19:17 2025 +0000

Check the document and update the changes

commit e3ae73d16cf4524ad0d4141f91b590a7a26a160c

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Tue Dec 31 07:45:05 2024 -0800

updated report 3.2 and 3.3 & please align according to formate.

commit d8b8c7544e8f8b9ddcf651e8a91e001aedf27c58

Merge: 2b3e5e1 568384c

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Dec 30 16:35:55 2024 +0000

Updated Report File Section 3.1 Visualization

commit 2b3e5e1fa96bbac1a4c40503b4a54d87a6277fc0

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Dec 30 16:32:42 2024 +0000

Updated Report File Section 3.1 Visualization

commit 568384cdf1769b0b1ce6b5b88d5debc2eaad3d3d

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 08:25:02 2024 -0800

Appropriate plot for the RQ output of an R script including r code and png file

commit 5d554a2e0467d49791617a9828699c22a5703f4d

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 08:17:44 2024 -0800

My First

commit fd214794552c417745d2f8401f7857fa5db8d90b

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Dec 30 15:47:59 2024 +0000

Update Section Introduction

commit 4f376507f57371c4c112cc07abaf32fbcf8e9664

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 07:37:23 2024 -0800

remove extra files

commit a6e85afb1abea61056620d4de9bd4836370d10ce

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 07:36:38 2024 -0800

load dataset and check dataset columns and rows in RStudio using rcode

commit ebb0af035167eb9efba085318830ad62d39ac5d8

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 07:27:16 2024 -0800

loadin dataset

commit 82b168c54920f36c22266b94bfbb8c0a066c0a10

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 07:24:43 2024 -0800

deleting extra files

commit 778dd8b4871e546778042cba57b13645bfc3f623

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 07:22:38 2024 -0800

checking checking

commit 4d393f1b8caace79cbe89015e4f85144446be041

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 07:07:12 2024 -0800

Accessing Project in RStudio

commit 118658bf511d931e9cf7c1c9aa88712ca6bea623

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 06:50:14 2024 -0800

My First

commit 9a0cb749c12e148e65f1cc0d4a8b2e81bef8d066

Author: Mirza Taimur Zafar <mz24abd@herts.ac.uk>

Date: Mon Dec 30 04:53:06 2024 +0000

accessing project folder

commit 65422949f0abfb45a7c62b5621cff60b3f921380

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 04:46:51 2024 +0000

a78 txt file changings

commit 47296b9f5c27e6b70f5cd9c1fd4bf0bbba1afbd7

Author: Muhammad Usman <mu24aan@herts.ac.uk>

Date: Mon Dec 30 04:42:38 2024 +0000

a78 project creation