7COM1079-0901-2025 - Team Research and Development Project

Final report title: Analysis of Gender-Based Income Disparities in the United States using the 1994 Adult Census.

Group ID: A 74

Dataset number: DS096 – adult income1.csv

Prepared by: Hariprasad Murugesan - 24135993
Phani Muppalla - 24113322
Hariharaan Murugesan Salaja - 24160923
Prabhaharan Radhakrishnan - 24138313
Rajath Krishna Naik – 24157852

University of Hertfordshire
Hatfield, 2025

Table of Contents

# 1. Introduction

## 1.1 Problem statement and research motivation

Income inequality between men and women continues to be a prevailing issue in all most every country but for our research we choose United States. There are many research studies indicating that gender tends to affect earnings and subsequently access to higher-paying occupations. According to (Platt, et al., 2016) wage gap in gender causes mood disorders and depression. Men earn more than women even in traditionally female occupations, for example male registered nurses have historically out-earned female registered nurses. Considering these factors, we focus on testing these proportional differences to gain factual knowledge into gender-based income patterns using actual census data.

## 1.2 The data set

The dataset used in this research is the Adult Income Census dataset, which contains 31,947 rows and 12 attributes, including age, education, occupation, race, sex, and category of income. It originates from the 1994 U.S. Census Bureau database and is one of the most extensively used datasets when studying income classification, demographic disparities, and social patterns in the United States.

## 1.3 Research question**.** State your RQ

**RQ**: **Is there a significant difference in the proportions of individuals earning less than 50K and more than 50K between males and females in the United States in 1994?**

This question aims to determine whether income distribution in United States varies by gender when comparing the two standard income categories present in the adult income dataset during the year 1994.

## 1.4 Null hypothesis and alternative hypothesis (H0/H1)

H0 (Null Hypothesis):

The null hypothesis ($H_0$) is that gender (male vs. female) and income category (<50K vs. >50K) are not related in the dataset. This means that any differences observed in the distribution of income categories between males and females occur purely by chance and cannot be attributed to gender.

H1(Alternative Hypothesis):

The alternative hypothesis ($H_1$) is that gender (male vs. female) and income category (<50K vs. >50K) are related. According to this hypothesis, gender influences the likelihood of belonging to either income category. Earlier studies using U.S. Census data have shown clear income differences between men and women, which is why it makes sense to test this idea (Blau, 2017).

# 2 Background research

## 2.1 Research papers

1)" The Gender Wage Gap: Extent, Trends, and Explanations." (Blau, 2017)

The following article compares how men and women differ based on different wage levels, such as high, middle, and low earners. They focus mainly on the shares of males and females in each part of the wage distribution to show why women still earn less compared to men.

2)" Unequal depression for equal work? How the wage gap explains gendered disparities in mood disorders" (Platt, et al., 2016).

In this journal, the authors group workers by income and compare the proportions between men and women in the lower and higher pay categories. Then, they relate these proportional differences in earnings to find differences in depression and anxiety rates between genders.

3) "Gender wage gap and its associated factors: An examination of traditional gender ideology, education, and occupation" (Langdon & Klomegah, 2013).

In this paper, they did statistical analysis using Chi-square and logistic regression then found that gender is a powerful predictor of earrings. The study shows that women were three times more likely than men to be in a low-income group (under $50,000). This gap persists even after controlling for education and occupation showing it's not just about difference in job choices.

## 2.2 Why RQ is of interest?

This research question is important because gender-based income inequality remains one of the most concerning social and economic issues in the United States. Though earlier studies documented the gender wage gap, most focused on continuous income differences rather than proportional differences between defined income groups. While there are clear categories (<50K and >50K) in the adult dataset, the literature rarely investigates whether these proportions

differ significantly by gender. This addresses a gap that shows more clearly how income is distributed across males and females.
Fairness research can be informed by this insight, policy discussions can be directed by it, and it points to structural disparities found in real census data.

# 3 Visualisation

## 3.1 Appropriate graphs for the RQ output of an R script

Fig 1. United States: Proportion of Gender within Income Brackets:

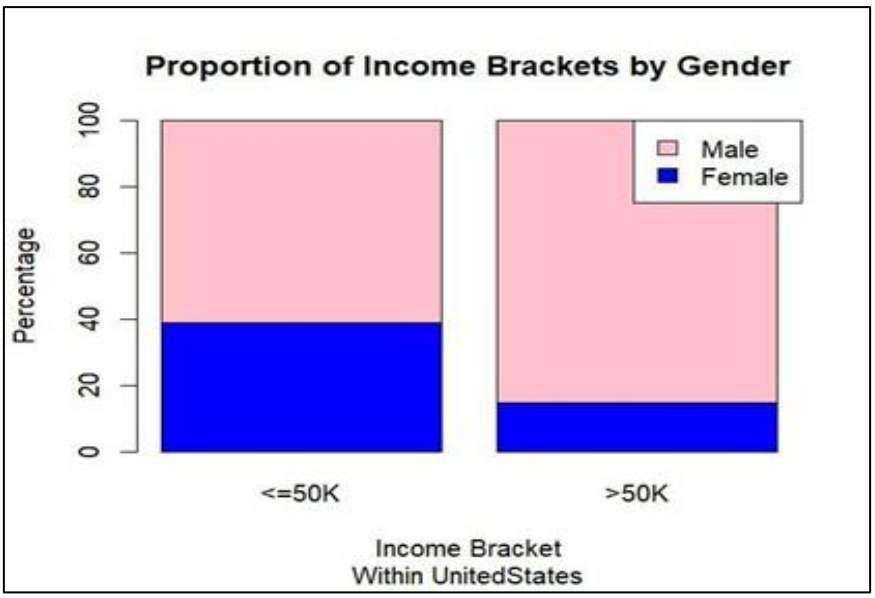
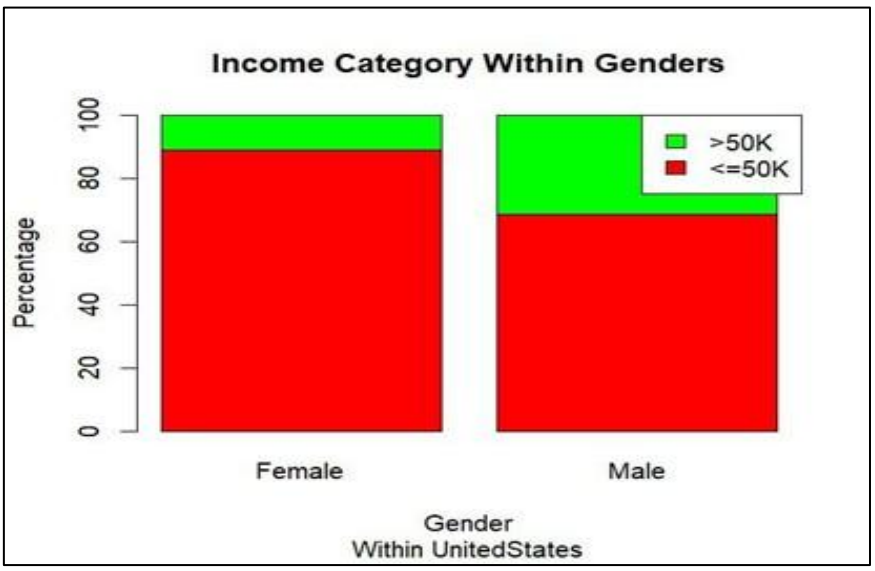



Fig 2. United States: Income category Distribution within Genders:



Contingency table:

United states: Sex vs Income Bracket Contingency table

| | **<= 50K** | **>50K** |
|---|---|---|
| **Female** | **8594** | **1068** |
| **Male** | **13371** | **6087** |

Sample size: n = 29,120
Female: 9,662 total (93.3% earn <=50K)
Male: 19,458 total (68.7% earn <=50K)

For comparison of proportions analysis, we will use a chi-square test to check for the likelihood that there is a relationship between the two nominal variables. This non-parametric test makes no assumptions about the shape of the data (which is nominal, not interval) so we do not include a histogram for this test.

### 3.2 Additional information relating to understanding the data

The bar plot clearly shows that males have a larger green section, meaning more of them earn over 50K. The contingency table supports this with actual numbers, confirming that the pattern seen in the chart is backed by data and suitable for the chi-square test.

### 3.3 Useful information for the data understanding

The plot clearly shows that most women fall into the ≤50K income group, while men appear more often in the >50K range. Although both genders mostly earn below 50K, men are still more likely to be high earners. This trend suggests gender may influence income levels.

## 4 Analysis:

### 4.1 Statistical test used to test the hypotheses and output

We applied Pearson's Chi-square test of independence to the United States contingency table because both variables male, female and income bracket which is 1(≤50K), 2(>50K) are nominal. Our research question clearly asks whether the income proportions differ by gender, so a proportions-based test is essential. The Adult Income dataset supplies large cell counts, satisfying the expected frequency assumption and justifying a non-parametric test that does not impose distributional constraints on categorical data.

### 4.2 The null hypothesis is rejected /not rejected based on the p-value

The test with United-States produced $x^2 = 1424.4$ with $df = 1$ and $p < 2.2 \times 10^{-16}$, far below alpha = 0.05. Therefore, we rejected the null hypothesis of equal income proportions across genders. Practically only about 11 % of females fall in the >50K bracket, compared with roughly 31 % of males, indicating a strong gender income gap. Because the sample size is approximately about 32000 records, the result is statistically robust and most likely it is not a sampling fluke. This effect also has real world meaning - males dominate high income positions, while females are concentrated in the ≤50K category, implying the need to explore other causes.

## 5 Evaluation – group's experience at 7COM1079

### 5.1 What went well

Our group coordinated well together, and everyone communicated clearly during the project. We split the tasks for each member fairly, and each member contributed in a useful way. Choosing the United States as our only country simplified our workflow and let us deepen the narrative. The feedback we received during the tutorial sessions greatly helped us refine our approach early, making the rest of the work much simpler.

### 5.2 Points for improvement

Initially, we ended up enthusiastically working with many countries from the dataset and later decided to remove them from the scope. At times, we also found it difficult to sync up due to our different schedules, which caused some tasks to be delayed beyond the planned time. To address this, we started connecting via Google Meet three times a week to keep track of progress. This improved our coordination and ensured timely delivery.

### 5.3 Group's time management

Overall, our time management was good since we completed the important tasks on time. However, in some cases, making sense of the results and finalising slides took longer, costing us roughly 2–3 working days more. Setting up weekly goals and using a shared checklist helped us stay more organised and finish tasks on time.

### 5.4 Project's overall judgement

Our United-States only project was executed well by us. The cleaned data, chi-square test, bar charts, and statistics all obviously point toward the same gender pay gap. Even though we initially had some issues, our final execution was strong, and we learned and sharpened many new skills.

### 5.5 Note any changes to group since original allocation if applicable. Add new or amended GitHub Ids for new members

No changes occurred after the original allocation; all members and GitHub handles remain as initially registered.

### 5.6 Comment on the GitHub log output

Appendix B Section 8 highlights the log snippet. Key entries include:

- One team member handled coding and statistical analysis.
- Another focused on the visualization and analysis.
- Other remaining members equally prepared the report.
- All collaborated, supported each other, documented details in Git, and contributed equally throughout the project.

## 6 Conclusions

### 6.1 Results explained

The United States contingency table shows starkly different distributions: 93% of women versus 69% of men earn ≤50K, while 31% of men but only 11% of women surpass >50K. Pearson's Chi-square statistic of 1424.4 (df = 1, p < 2.2×10^-16) confirms the gap is not due to sampling noise. Because the contingency counts, visuals, and statistics all come from the same cleaned DS096 data, the results consistently convey the same conclusion.

### 6.2 Interpretation of the results

Rejecting the null hypothesis confirms that there is a link between gender and income. Specifically, men are three times more likely to earn over $50k, demonstrating a gender pay disparity. Furthermore, visual evidence from bar charts and chi square test results validates this significant disparity.

### 6.3 Reasons and/or implications for future work, limitations of your study

We only used one segment of the data and a basic two variables to test, while we tried to use other variables like education, marital status and occupation too. Next time we would include with those factors, check trends over time, and maybe gather qualitative notes to see why the gap persists.

## 7 References:


1. Blau, F. D. K. L. M., 2017. The gender wage gap: Extent, trends, and explanations. *Journal of economic literature,* 55(3), pp. 789-865.
Available at:
https://www.aeaweb.org/articles?id=10.1257%2Fjel.20160995&trk=article-ssr-frontend-pulse_x-social-details_comments-action_comment-text [Accessed: 09/12/25]

2. Langdon, D. L. & Klomegah, R., 2013. Gender wage gap and its associated factors: An examination of traditional gender ideology, education, and occupation. *International Review of Modern Sociology,* pp. 173-200.
Available at: https://www.jstor.org/stable/43496468?seq=1 [Accessed: 09/12/25]

3. Platt, J., Prins, S., Bates, L. & Keyes, K., 2016. Unequal depression for equal work? How the wage gap explains gendered disparities in mood disorders. *Social Science & Medicine,* Volume 149, pp. 1-8.
Available at: https://www.sciencedirect.com/science/article/pii/S0277953615302616 [Accessed: 09/12/25]

# 8 Appendices

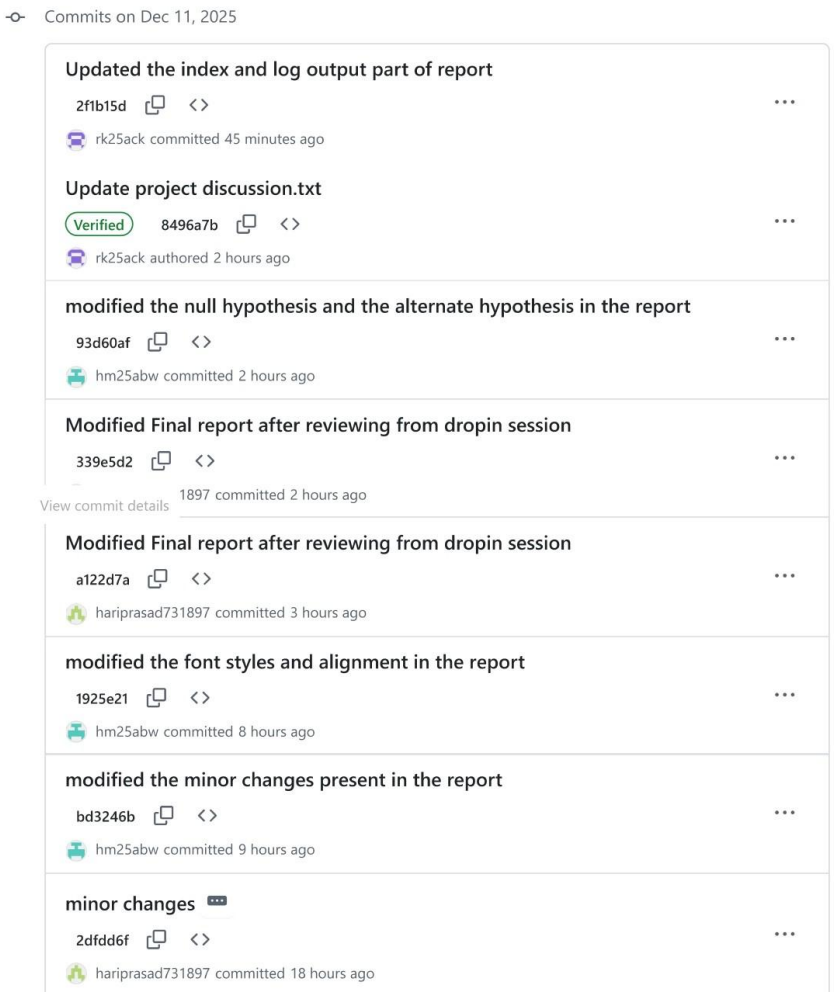
## 8.1 R code used for analysis and visualisation:

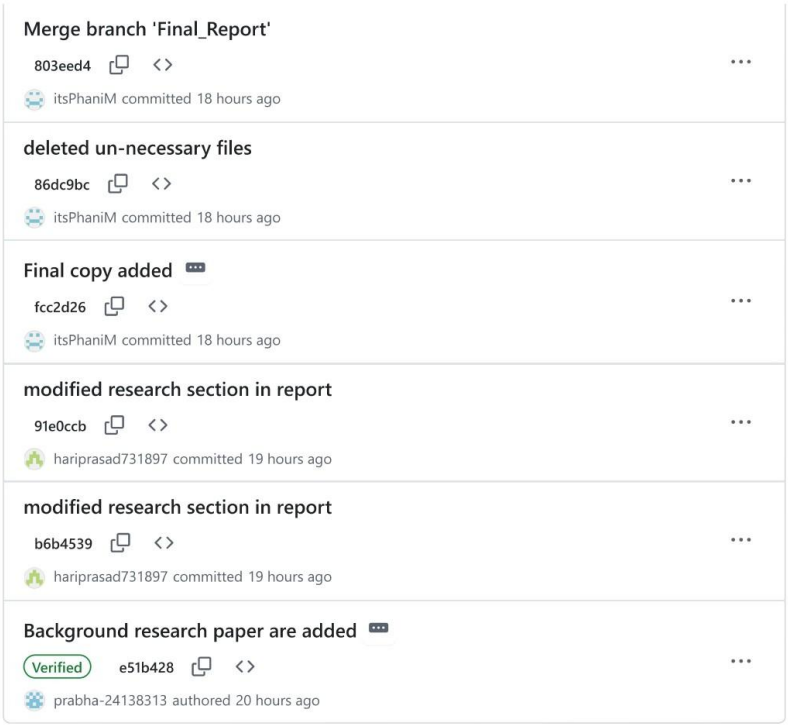
```
# Import data
library(readr)
df <- read_csv("adult income1.csv")

# Clean and filter dataset to only include 'Male' and 'Female'
df2 <- subset(df, sex == "Male" | sex == "Female")
df2 <- subset(df, native.country == "United-States" )

# Create a new binary column for income bracket
df2$income_bracket <- ifelse(df2$income == "<=50K", "<=50K", ">50K")

# Build a contingency table: Gender vs Income bracket
 usa<- table(df2$sex[df2$native.country=="United-States"],
df2$income_bracket[df2$native.country=="United-States"] )


# Chi-squared test for difference in proportions
chisq.test(usa)

# Convert counts to percentages by column: proportion inside each income
group
percentages <- prop.table(usa, margin=2) * 100

# Plot: Stacked bar chart of gender proportions inside each income group
barplot(percentages, col = c("orange", "black"), xlab = "Income Bracket",
ylab = "Percentage",
      main = "Proportion of Income Brackets by Gender", ylim = c(0, 100),
      legend.text = rownames(usa), args.legend = list(x = "topright"),sub =
"Within UnitedStates")

# Optional: Also view proportions within each gender (transpose for barplot)
tpercentages <- prop.table(t(usa), margin=2) * 100
barplot(tpercentages, col = c("red", "green"), xlab = "Gender", ylab =
"Percentage",
      main = "Income Category Within Genders", ylim = c(0, 100),
      legend.text = colnames(usa), args.legend = list(x = "topright"),sub =
"Within UnitedStates")
```

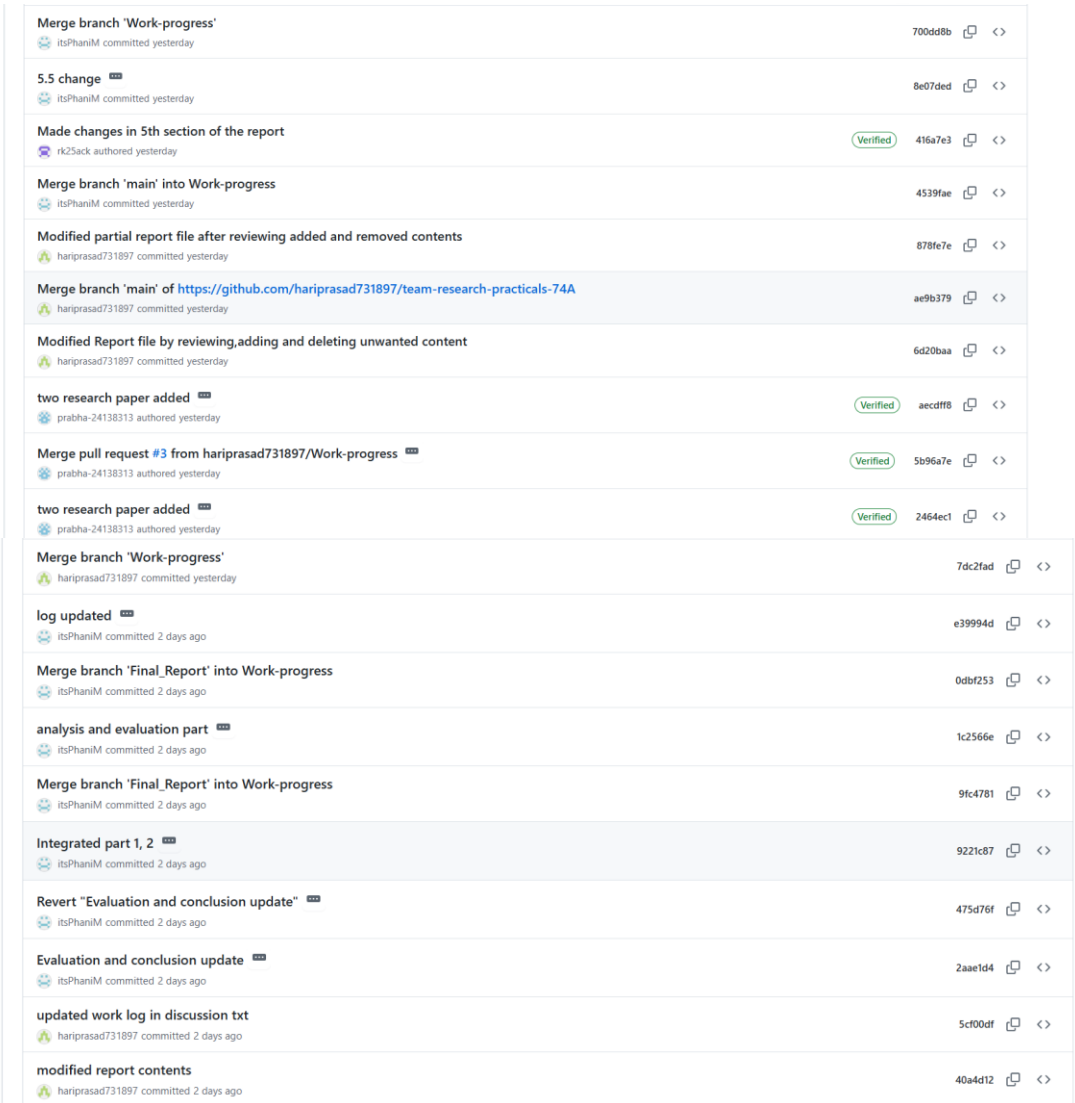
## 8.2 GitHub log output.

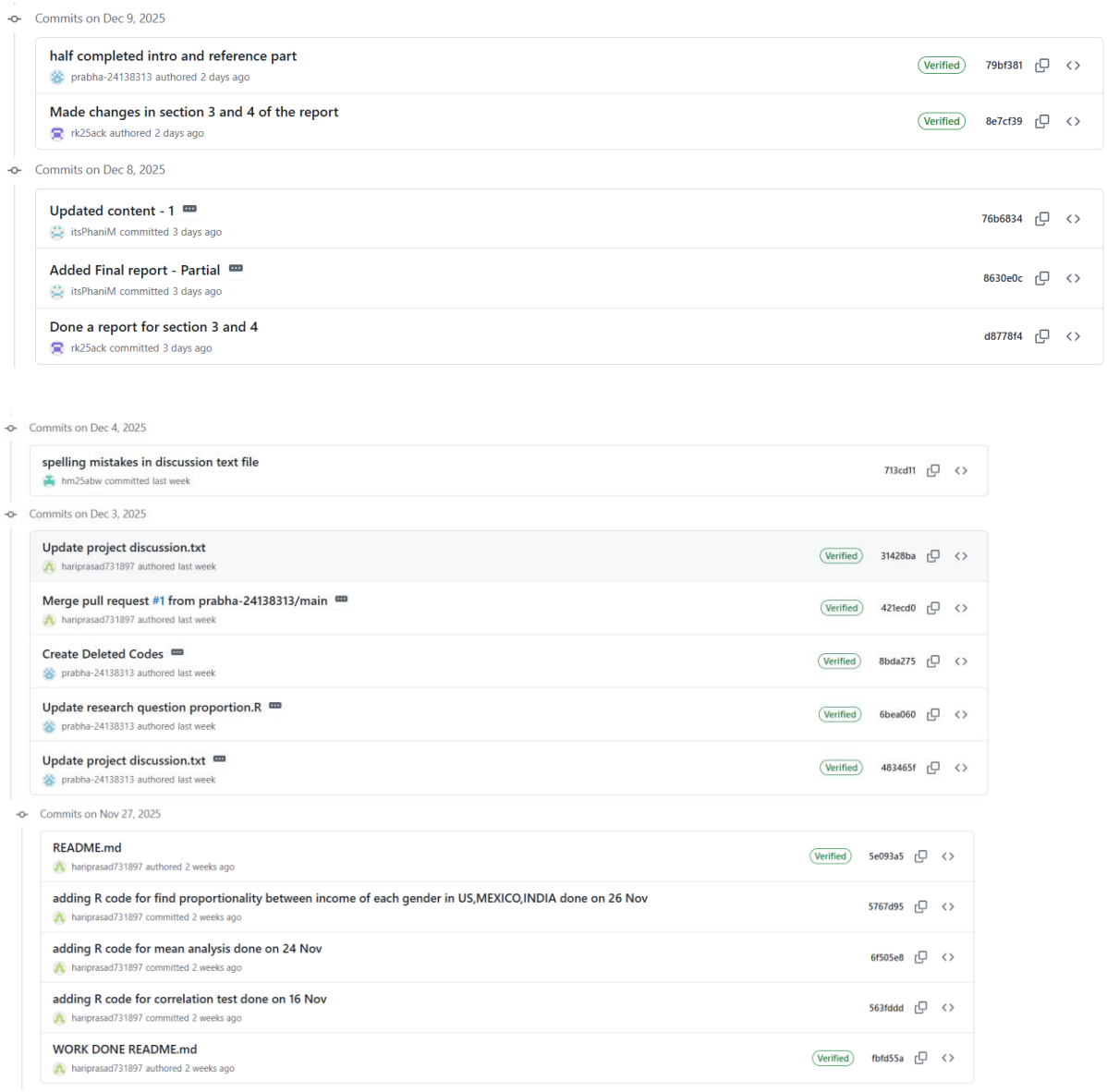
| Commits | User | UserMail |
|---|---|---|
| 20 | hariprasad731897 | hm25abl@herts.ac.uk |
| 4 | hm25abw | hm25abw@herts.ac.uk |
| 18 | itsPhani | pm25aao@herts.ac.uk |
| 9 | prabha-24138313 | pr25aar@herts.ac.uk |
| 7 | rk25ack | rk25ack@herts.ac.uk |

## Screenshots of GIT Commit Logs:




Merge branch 'Final_Report'
803eed4
itsPhaniM committed 18 hours ago
deleted un-necessary files
86dc9bc
itsPhaniM committed 18 hours ago
Final copy added
fcc2d26
itsPhaniM committed 18 hours ago
modified research section in report
91e0ccb
hariprasad731897 committed 19 hours ago
modified research section in report
b6b4539
hariprasad731897 committed 19 hours ago
Background research paper are added
Verified
e51b428
prabha-24138313 authored 20 hours ago


Merge branch 'Work-progress'
itsPhaniM committed yesterday
700dd8b
5.5 change
itsPhaniM committed yesterday
8e07ded
Made changes in 5th section of the report
rk25ack authored yesterday
Verified
416a7e3
Merge branch 'main' into Work-progress
itsPhaniM committed yesterday
4539fae
Modified partial report file after reviewing added and removed contents
hariprasad731897 committed yesterday
878fe7e
Merge branch 'main' of https://github.com/hariprasad731897/team-research-practicals-74A
hariprasad731897 committed yesterday
ae9b379
Modified Report file by reviewing,adding and deleting unwanted content
hariprasad731897 committed yesterday
6d20baa
two research paper added
prabha-24138313 authored yesterday
Verified
aecdff8
Merge pull request #3 from hariprasad731897/Work-progress
prabha-24138313 authored yesterday
Verified
5b96a7e
two research paper added
prabha-24138313 authored yesterday
Verified
2464ec1
Merge branch 'Work-progress'
hariprasad731897 committed yesterday
7dc2fad
log updated
itsPhaniM committed 2 days ago
e39994d
Merge branch 'Final_Report' into Work-progress
itsPhaniM committed 2 days ago
0dbf253
analysis and evaluation part
itsPhaniM committed 2 days ago
1c2566e
Merge branch 'Final_Report' into Work-progress
itsPhaniM committed 2 days ago
9fc4781
Integrated part 1, 2
itsPhaniM committed 2 days ago
9221c87
Revert "Evaluation and conclusion update"
itsPhaniM committed 2 days ago
475d76f
Evaluation and conclusion update
itsPhaniM committed 2 days ago
2aae1d4
updated work log in discussion txt
hariprasad731897 committed 2 days ago
5cf00df
modified report contents
hariprasad731897 committed 2 days ago
40a4d12


Commits on Dec 9, 2025
half completed intro and reference part
prabha-24138313 authored 2 days ago
Verified
79bf381
Made changes in section 3 and 4 of the report
rk25ack authored 2 days ago
Verified
8e7cf39
Commits on Dec 8, 2025
Updated content - 1
itsPhaniM committed 3 days ago
76b6834
Added Final report - Partial
itsPhaniM committed 3 days ago
8630e0c
Done a report for section 3 and 4
rk25ack committed 3 days ago
d8778f4
Commits on Dec 4, 2025
spelling mistakes in discussion text file
hm25abw committed last week
713cd11
Commits on Dec 3, 2025
Update project discussion.txt
hariprasad731897 authored last week
Verified
31428ba
Merge pull request #1 from prabha-24138313/main
hariprasad731897 authored last week
Verified
421ecd0
Create Deleted Codes
prabha-24138313 authored last week
Verified
8bda275
Update research question proportion.R
prabha-24138313 authored last week
Verified
6bea060
Update project discussion.txt
prabha-24138313 authored last week
Verified
483465f
Commits on Nov 27, 2025
README.md
hariprasad731897 authored 2 weeks ago
Verified
5e093a5
adding R code for find proportionality between income of each gender in US,MEXICO,INDIA done on 26 Nov
hariprasad731897 committed 2 weeks ago
5767d95
adding R code for mean analysis done on 24 Nov
hariprasad731897 committed 2 weeks ago
6f505e8
adding R code for correlation test done on 16 Nov
hariprasad731897 committed 2 weeks ago
563fddd
WORK DONE README.md
hariprasad731897 authored 2 weeks ago
Verified
fbfd55a