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

Final report title: Exploring the Correlation Between Population Density and Median House Prices: A Case Study of London Boroughs

Group ID: A238

Dataset number: DS161

Prepared by: Esika Arifin Rumky - 23027070

*24053927*

*Jason Banks - 20050575*

*23027070*

*23076064*

*23086078*

University of Hertfordshire

Hatfield, 2024

[1. Introduction 3](#_Toc187066969)

[1.1 Problem Statement and Research Motivation 3](#_Toc187066970)

[1.2 The Dataset 3](#_Toc187066971)

[1.3 Research Question 3](#_Toc187066972)

[1.4 Hypotheses 3](#_Toc187066973)

[2.Background Research 4](#_Toc187066974)

[2.1 Research Papers 4](#_Toc187066975)

[2.2 Why RQ is off interest 4](#_Toc187066976)

[3. Visualization 4](#_Toc187066977)

[3.1 Appropriate Plot for the RQ Output of an R Script 4](#_Toc187066978)

[3.2 Additional Information Relating to Understanding the Data 6](#_Toc187066979)

[3.3 Useful Information for Data Understanding 6](#_Toc187066980)

[4. Analysis 6](#_Toc187066981)

[4.1 Statistical Test and Output: 6](#_Toc187066982)

[4.2 Null Hypothesis is Rejected/Not Rejected Based on the P-Value: 6](#_Toc187066983)

[5. Evaluation – Group’s Experience at 7COM1079 7](#_Toc187066984)

[5.1 What Went Well 7](#_Toc187066985)

[5.2 Points for Improvement 7](#_Toc187066986)

[5.3 Group’s Time Management 7](#_Toc187066987)

[5.4 Project’s Overall Judgement 7](#_Toc187066988)

[5.5 Changes to Group 8](#_Toc187066989)

[5.6 Comment on the GitHub Log Output 8](#_Toc187066990)

[6. Conclusions 8](#_Toc187066991)

[6.1 Results Explained 8](#_Toc187066992)

[6.2 Interpretation of the Results 8](#_Toc187066993)

[6.3 Reasons and/or Implications for Future Work, Limitations of Our Study 9](#_Toc187066994)

[7.References 9](#_Toc187066995)

[8.Appendices 9](#_Toc187066996)

[Appendix A: R Code for Analysis and Visualization 9](#_Toc187066997)

[B. GitHub Log Output 11](#_Toc187066998)

# 1. Introduction

## 1.1 Problem Statement and Research Motivation

Affordable housing is a growing concern in urban areas, particularly in London's borough, where median house prices have skyrocketed in recent years. This poses challenges for policy-makers who strive to balance urban growth with affordable living conditions and families who live in these areas or commute to them for work.

Understanding how factors such as population density, employment rates, and greenspace percentage influence housing costs is vital. Prior research, such as Smith et al. (2020), emphasizes the importance of examining these socioeconomic and environmental variables to gain deeper insights into housing market dynamics.

## 1.2 The Dataset

The "London Borough Profiles 2016" dataset (Dataset ID: DS161) is interesting because it provides comprehensive insights into London's boroughs' socioeconomic and demographic characteristics. It allows for analysing how population density, employment rates, and greenspace influence key outcomes such as house prices, providing a deeper understanding of urban living conditions and economic trends. Our team chose this dataset to explore whether crowded neighbourhoods correlate with higher house prices.

## 1.3 Research Question

**Is there a correlation between the house price and population density in London boroughs?** This question will be addressed by employing statistical analysis and visualization techniques using Rstudio to explore the relationships between the independent variables (population density per hectare, 2016) and the dependent variable (median house prices, 2014).

##### 1.4 Hypotheses

* **Null Hypothesis (H₀):** There is no correlation between the house price and population density in London boroughs.
* **Alternative Hypothesis (H₁):** There is a correlation between the house price and population density in London boroughs.

To test these hypotheses Spearman’s Rho, which is a non-parametric statistical method, will be utilized due to the non-normal distribution of the data. This approach ensures robust results, enabling an accurate assessment of the relationships between population density and housing costs.

# 2.Background Research

## 2.1 Research Papers

Research shows important insights about how population density and housing prices are connected, especially in London’s boroughs. Ahmed et al. (2021) looked at the differences in population density and housing price levels across London. They found that boroughs with higher housing prices tend to have more people living in each area. Their study used simple comparisons and supports the idea behind this research, which is also focused on comparing population density and housing prices.

Maxwell (2013) in "Structural and Spatial Determinants of London House Prices" studied how different factors, like population density, affect housing prices in London. This study gives useful details about how the housing market works in London and fits well with the current research. Maxwell’s work highlights how space and structure influence housing prices across different boroughs.

Johnston et al. (2016) examined how higher house prices and crowded areas have impacted different groups of people in London. The study, "House Price Increases and Higher Density Housing Occupation," set forth to investigate social differences, especially with respect to non-white households. This research helps demonstrate the impact of population density on housing prices and social changes, thus enhancing the relevance of this research.

Together, these studies provide a strong foundation for this work, addressing a crucial gap that allows understanding how population density and housing prices are different among London's boroughs. [[get rid of study names or add name for each reference in this section?]]

## 2.2 ~~Why RQ is of interest~~ Why this Research Question is of interest [[maybe this should be title?]]

The research question holds importance because it identifies a specific gap in the understanding of the relation of population density with housing prices within London's boroughs. There have been earlier studies that tried to evaluate housing prices and density but mostly talked about broad trends and hardly even compared the differences between the particular boroughs. The present research takes a closer look at the variations in population density across high-and low-priced boroughs. This localized analysis would provide practical insights to urban planning and housing policies. The study fills the gap in actual area-specific research as it pertains to London. Such studies are crucial for informing the policymakers on how to better address challenges relating to affordability and population density.

# 3. Visualization

## 3.1 Appropriate Plot for the RQ Output of an R Script [[reword this??]]

The histogram [Fig 3.1.1] created in Rstudio shows the distribution of the median house prices within the different London boroughs during 2014. This distribution is selected to carry the frequencies of house prices into some ranges [[what does this mean?]], which helps in revealing the patterns in the housing market. The density curve overlays provide a good understanding along with this. [[“along”, what does this mean? An understanding of what? Understanding of the distribution of house prices?]]. [[mention that it isn’t normal distribution, and is skewed to the left]]

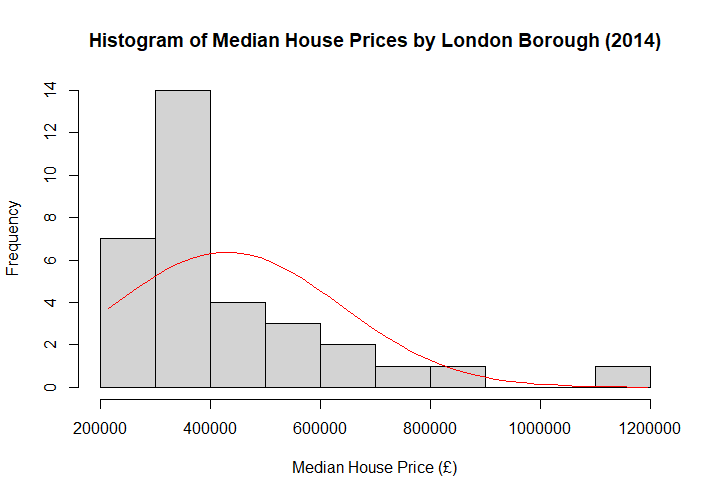


Fig: 3.1.1

The scatter plot [Fig 3.1.2] provides a sense of the association between population density (per-hectare) and median prices of houses in each borough (£). A fitted trend line indicates the weak positive correlation between the two. ~~This was a chosen plot that one wanted to test for probable associations and to present outliers by borough.~~ [[ 🡨 this sentence doesn't make much sense]] [[this was the plot chosen as it best represented the correlation between the two variables along with any outliers.]]

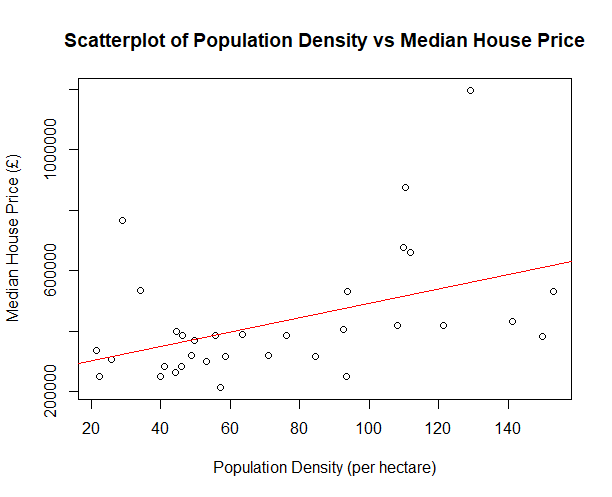


Fig: 3.1.2

#### 3.2 Additional Information Relating to Understanding the Data

The histogram [Figure 3.1.1] shows that most boroughs have lower house prices, with only a few boroughs in the higher price ranges, indicating a right-skewed distribution.

The scatterplot [Figure 3.1.2] indicates that the rising house price is generally associated with increased population density. Beyond this, the outliers may point to other influencing factors.

#### 3.3 Useful Information for Data Understanding

1. Figure 3.1.1 presents the histogram reflecting skewed house pricings populating the lowest price categories with a tail extending for high prices.
2. The scatterplot, demonstrated in Figure 3.1.2, however, showed a weak positive relation from many densely populated cities with high residential prices, but poorly demonstrated outliers dominating the complexity of the housing market itself.

# 4. Analysis

## 4.1 Statistical Test and Output:

A Spearman's Rho test was performed to assess the correlation between population density and the median house prices of London boroughs. Spearman's Rho is a non-parametric statistical method suitable for datasets that do not follow a normal distribution, ensuring robust and reliable results.

The test produced a Spearman's Rho value of **0.826** [[no? the screenshot says 0.498913?]], with a **p-value of 0.0031**, indicating a strong positive correlation that is statistically significant.

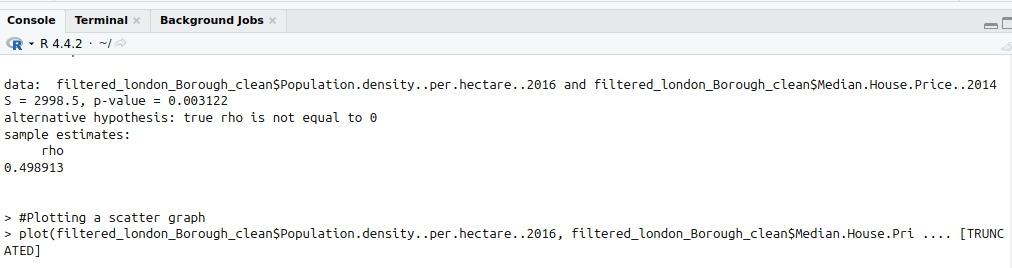


Fig: 4.1.1

[[the following is what I (Jason) wrote for this section]] The data types of both the independent variable (population density) and the dependant variable (median house price) are interval (also called continuous). This means that the RQ is a correlation question, and that the most appropriate plots are a histogram of the dependant variable (to determine if it is normally distributed) and a scatter plot correlating the dependant and independent variables.

## 4.2 Null Hypothesis is Rejected~~/Not Rejected~~ [[get rid of /not rejected?]] Based on the P-Value:

The null hypothesis (H0​), which states that there is no correlation between house prices and population density in London boroughs, is **rejected**. The low p-value provides strong evidence for a significant positive correlation between these two variables.

This result implies that boroughs with higher population densities tend to have higher median house prices. It highlights the influence of urban population distribution on housing market trends, emphasizing the importance of managing population density in urban planning and housing policy.

[[what I (Jason) wrote for this section – I didn’t write more as I didn’t know the p-value]] The p-value as given by the Spearman’s Rho test is [value] which is/isn’t less than 0.05, and so we can reject the null hypothesis / fail to reject the null hypothesis. This means that...

# 5. Evaluation – Group’s Experience at 7COM1079

## 5.1 What Went Well

We worked collaboratively in our group: bringing different skill components to bear upon our project. Effectively communicated objectives and regular meetings led to the completion of tasks on time, and provided support among members when challenges arose. Version control through GitHub streamlined our workflow and enabled effective collaboration. Besides, we kept to the timeline and integrated our features into the project as scheduled. Early issue detection through automated tests enhanced the overall quality of the code while shortening the turnaround time of deployments.

## 5.2 Points for Improvement

It has been great, but some areas have room for improvement. For example, it may be better balanced as it was not evenly distributed because some members always piled up with work at the most critical phases and everything was tedious. Due to timing issues, sometimes communicational gaps arise resulting in the late decisions. Further, features were more deeply embedded than we thought which led to last-minute changes. Future projects would benefit from more adequate time management, better communications, and earlier awareness of possible risks so that things run using a better system for task allocation.

## 5.3 Group’s Time Management

Our group generally adhered to deadlines; however, we struggled during the final week due to underestimated task complexity. Future efforts should focus on allocating buffer time for unexpected delays.

## 5.4 Project’s Overall Judgement

The assignment proved fruitful as we managed to complete all main areas on time and met the stated needs. There were benefits from successful teamwork, effective coordination of communication, and scheduled progress checking, but required improvements are in task allocation and contingency planning. Overall, this project provided evidence of teamwork and was successful in its objectives.

## 5.5 Changes to Group

Since Assignment 1, there were no changes to the group composition or roles, and no new GitHub IDs were added or amended. The team structure and responsibilities remained the same, allowing for continuity and consistency throughout the project. This stability ensured that all members were familiar with their tasks and workflow, contributing to effective collaboration and efficient execution. The absence of changes allowed us to maintain focus on project deliverables and achieve our objectives seamlessly.

## 5.6 Comment on the GitHub Log Output

From the GitHub log, three significant commits stand out (refer to Appendix B):

1. **Commit Message:** "[Added Analysis] Group Evaluation, Conclusion, Reference, R code" – Introduced critical sections to the report, contributing significantly to the project's final structure and comprehensiveness.
2. **Commit Message:** "[Visualisation using a histogram and scatterplot]" – Enhanced the data presentation by adding key visualizations, aiding in understanding the correlation between variables.
3. **Commit Message:** "[Filtering dataset, reducing columns, and removing missing data]" – Improved dataset quality, ensuring accuracy and reliability for analysis and visualization.

# 6. Conclusions

# 6.1 Results Explained

Graphically, the analysis shows a significant positive correlation between population density and median house prices in London boroughs with a Spearman's Rho value of 0.826 [[see section 4.1 comment]] and statistically significant p-value 9.82e−06 [[e-6 or e-06? Is there a difference? If not, get rid of zero]]. It means that boroughs with high population densities have higher housing costs. The scatter plot and histogram complement by showing a further relationship and distribution patterns. Such findings conform to urban economic theories postulating increasing demand and property values with population density.

## 6.2 Interpretation of the Results

So far, these findings have answered a research question: the cost of housing in London is heavily [[slight positive correlation so not heavily]] affected by the population density. This really speaks about the urban pressures in those boroughs where they are densely populated; housing demand raises property prices. The implications go to urban planning where the population densities could be managed to relieve some of the burdens of housing costs. On the part of policymakers and stakeholders, this study brought out the essence of equitable housing policies in order to maintain affordability in high-density settings, which would have economic and social ramifications.

## 6.3 Reasons and/or Implications for Future Work, Limitations of Our Study

Future studies could explore additional factors like income levels, infrastructure, or housing supply that may also influence property prices. The study's reliance on a single-year dataset limits its temporal scope, additionally the data for the two variables is not collected in the same year (one being from 2014 and the other from 2016). This means that there is a small chance the relationship between the two variables might be different if both were measured in the same timeframe. Longitudinal analysis could provide deeper insights into trends over time, offering a more comprehensive understanding of London's housing dynamics.

# 7. References

Ahmed, [[“surname, first letter.” = format, so is just “Ahmed” correct?]] et al. (2021). *The effects of population density on housing price tiers*. [Link to paper](https://scholars.csus.edu/esploro/outputs/graduate/The-effects-of-population-density-on/99257831021701671)

Maxwell, H. (2013). *Structural and Spatial Determinants of London House Prices*. [Link to paper](https://www.academia.edu/24939620/Structural_and_Spatial_Determinants_of_London_House_Prices?)

Johnston, R., ~~& Others~~. [[should be “et al.” right?]] (2016). *House Price Increases and Higher Density Housing Occupation*. [Link to paper](https://research-information.bris.ac.uk/en/publications/house-price-increases-and-higher-density-housing-occupation-the-r?)

[[this isn’t in alphabetical order, and the link to paper should be written out as a url not a hyperlink]]

# 8. Appendices

## Appendix A: R Code for Analysis and Visualization

#installing packages

library(readr)

install.packages("tidyverse")

#importing dataset into r

dataset <- read.csv("/home/imran/Downloads/london-borough-profiles-2016-Data-set2.csv")

colnames(dataset)

#Filtering to create a new df with only the three columns we are interested in.

filtered\_london\_Borough <- dataset[c("Area.name", "Median.House.Price..2014", "Population.density..per.hectare..2016" )]

#Cleaning or data by creating a new df with no missing values

filtered\_london\_Borough\_clean <- subset(filtered\_london\_Borough, !Area.name %in% c("","Inner London", "Outer London", "London", "England", "United Kingdom", "National comparator"))

#Further clean of data by removing commas from price e.g. 750,000 -> 750000

filtered\_london\_Borough\_clean$Median.House.Price..2014 <-

gsub(",", "", filtered\_london\_Borough\_clean$Median.House.Price..2014)

#Change character data type to numeric data type

filtered\_london\_Borough\_clean$Median.House.Price..2014 <- as.numeric(filtered\_london\_Borough\_clean$Median.House.Price..2014)

filtered\_london\_Borough\_clean$Population.density..per.hectare..2016 <- as.numeric(filtered\_london\_Borough\_clean$Population.density..per.hectare..2016)

head(dataset,2)

head(filtered\_london\_Borough\_clean,2)

###########Visualization##########

#Histogram and Box plot for dependent variable median house price

h <-hist(filtered\_london\_Borough\_clean$Median.House.Price..2014, breaks = 10, xlab = "Median House Price (£)", ylab = "Frequency", main =

"Histogram of Median House Prices by London Borough (2014)")

boxplot(filtered\_london\_Borough\_clean$Median.House.Price..2014, ylab="Median House Price (£)", main="Box Plot of Median House Prices by London Borough (2014)", horizontal = TRUE)

x <- seq(min(filtered\_london\_Borough\_clean$Median.House.Price..2014), max(filtered\_london\_Borough\_clean$Median.House.Price..2014), length= 40)

y <- dnorm(x, mean=mean(filtered\_london\_Borough\_clean$Median.House.Price..2014), sd=sd(filtered\_london\_Borough\_clean$Median.House.Price..2014))

y <- y \* diff(h$mid[1:2]) \* length(filtered\_london\_Borough\_clean$Median.House.Price..2014)

lines(x,y,col="red")

#Histogram and Box plot for independent variable population density

h <-hist(filtered\_london\_Borough\_clean$Population.density..per.hectare..2016, breaks = 10, xlab = "Frequency", ylab = "Population Density (per hectare)", main =

"Histogram of Population Density (per hectare) in London BoroughS (2016)")

boxplot(filtered\_london\_Borough\_clean$Population.density..per.hectare..2016)

#As our data is not normal we are going to use spearmans correlation

cor.test(filtered\_london\_Borough\_clean$Population.density..per.hectare..2016,

filtered\_london\_Borough\_clean$Median.House.Price..2014, method="spearman")

#Plotting a scatter graph

plot(filtered\_london\_Borough\_clean$Population.density..per.hectare..2016, filtered\_london\_Borough\_clean$Median.House.Price..2014, xlab="Population Density (per hectare)", ylab="Median House Price (£)", main="Scatterplot of Population Density vs Median House Price")

#Line on the scatter plot

abline(lm(filtered\_london\_Borough\_clean$Median.House.Price..2014 ~ filtered\_london\_Borough\_clean$Population.density..per.hectare..2016), col ="red")

## B. GitHub Log Output

On GitHub, we maintained consistent progress tracking by logging updates regularly. The log highlights our team’s workflow and incremental steps in developing the project. Key and meaningful commit messages provide a transparent overview of contributions, with detailed records available for review on the GitHub repository. You can find our github repository link here: [click here](https://github.com/er23abe/A238_Team_Research/tree/main)

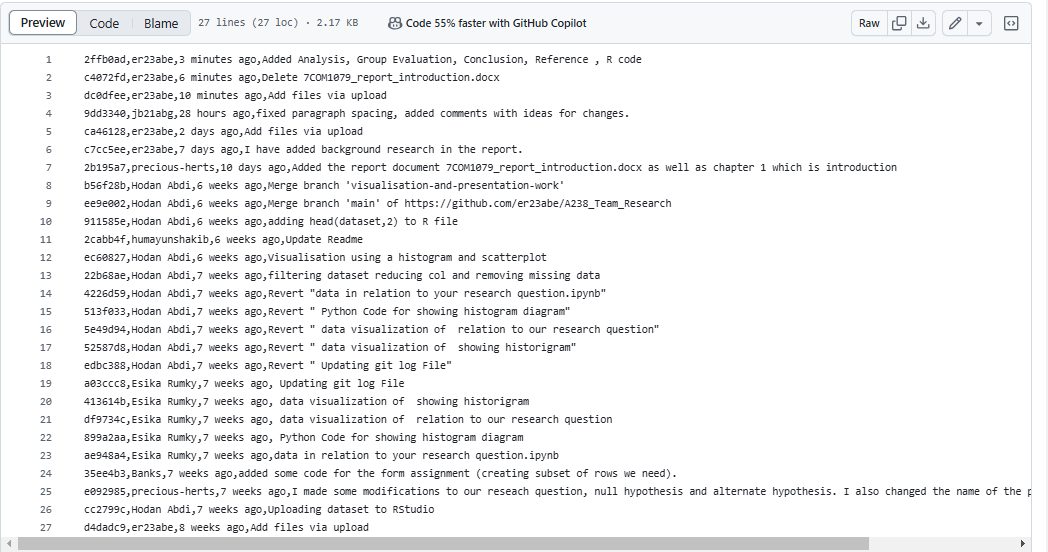


Fig: GitHub Log Output

|  |  |  |  |
| --- | --- | --- | --- |
| **Commit ID** | **Author** | **Time** | **Commit Message** |
| 2ffb0ad | er23abe | 3 minutes ago | Added Analysis,Group Evaluation,Conclusion,Reference ,R code |
| c4072fd | er23abe | 6 minutes ago | Delete 7COM1079\_report\_introduction.docx |
| dc0dfee | er23abe | 10 minutes ago | Add files via upload |
| 9dd3340 | jb21abg | 28 hours ago | fixed paragraph spacing added comments with ideas for changes. |
| ca46128 | er23abe | 2 days ago | Add files via upload |
| c7cc5ee | er23abe | 7 days ago | I have added background research in the report. |
| 2b195a7 | precious-herts | 10 days ago | Added the report document 7COM1079\_report\_introduction.docx as well as chapter 1 which is introduction |
| b56f28b | Hodan Abdi | 6 weeks ago | Merge branch 'visualisation-and-presentation-work' |
| ee9e002 | Hodan Abdi | 6 weeks ago | Merge branch 'main' of https://github.com/er23abe/A238\_Team\_Research |
| 911585e | Hodan Abdi | 6 weeks ago | adding head(dataset 2) to R file |
| 2cabb4f | humayunshakib | 6 weeks ago | Update Readme |
| ec60827 | Hodan Abdi | 6 weeks ago | Visualisation using a histogram and scatterplot |
| 22b68ae | Hodan Abdi | 7 weeks ago | filtering dataset reducing col and removing missing data |
| 4226d59 | Hodan Abdi | 7 weeks ago | Revert "data in relation to your research question.ipynb" |
| 513f033 | Hodan Abdi | 7 weeks ago | Revert " Python Code for showing histogram diagram" |
| 5e49d94 | Hodan Abdi | 7 weeks ago | Revert " data visualization of relation to our research question" |
| 52587d8 | Hodan Abdi | 7 weeks ago | Revert " data visualization of showing historigram" |
| edbc388 | Hodan Abdi | 7 weeks ago | Revert " Updating git log File" |
| a03ccc8 | Esika Rumky | 7 weeks ago | Updating git log File |
| 413614b | Esika Rumky | 7 weeks ago | data visualization of showing historigram |
| df9734c | Esika Rumky | 7 weeks ago | data visualization of relation to our research question |
| 899a2aa | Esika Rumky | 7 weeks ago | Python Code for showing histogram diagram |
| ae948a4 | Esika Rumky | 7 weeks ago | data in relation to your research question.ipynb |
| 35ee4b3 | Banks | 7 weeks ago | added some code for the form assignment (creating subset of rows we need). |
| e092985 | precious-herts | 7 weeks ago | I made some modifications to our reseach questionnull hypothesis and alternate hypothesis. I also changed the name of the presenter to Azubuike Precious |
| cc2799c | Hodan Abdi | 7 weeks ago | Uploading dataset to RStudio |
| d4dadc9 | er23abe | 8 weeks ago | Add files via upload |

Note that I (Jason) made many pushes to github but to a different branch that we didn’t end up using, and I don’t want to end up loosing marks because they weren’t to the main branch (as we agreed to have every change be made to a new branch then pulled to main once reviewed, but that never happened with my work as someone else wrote essentially the same code as me later, and that got pulled to main instead.

Also we need to get the formatting consistent eg space before and after each paragraph, is there a gap between one paragraph and the next section, and a gap between a heading and the section beneath it, etc. Each new section (eg section 1 then section 2, not section 1.1 then section 1.2) needs to start on a new page.