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

Final report title: (Analysis of Patient Data for Epidemiology Study*)*

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 4](#_Toc186807045)

[1.1 Problem statement and research motivation 4](#_Toc186807046)

[1.2 The data set 4](#_Toc186807047)

[1.3 Research Question 4](#_Toc186807048)

[1.4 Null hypothesis and alternative hypothesis (H0/H1) 4](#_Toc186807049)

[2. Background research 6](#_Toc186807050)

[2.1 Research papers (at least 3 relevant to your topic / DS) 6](#_Toc186807051)

[2.2 Why RQ is of interest (research gap and future directions according to the literature) 6](#_Toc186807052)

[3. Visualisation 7](#_Toc186807053)

[3.1 Appropriate plot for the RQ output of an R script 7](#_Toc186807054)

[3.2 Additional information relating to understanding the data (optional) 8](#_Toc186807055)

[3.3 Useful information for the data understanding 8](#_Toc186807056)

[4. Analysis 9](#_Toc186807057)

[4.1 Statistical test used to test the hypotheses and output 9](#_Toc186807058)

[4.2 The null hypothesis is rejected /not rejected based on the p-value 9](#_Toc186807059)

[5. Evaluation 10](#_Toc186807060)

[5.1 What Went Well 10](#_Toc186807061)

[5.2 Points for improvement 10](#_Toc186807062)

[5.3 Group’s time management 10](#_Toc186807063)

[5.4 Project’s Overall Judgment 10](#_Toc186807064)

[6. Conclusions 11](#_Toc186807065)

[6.1 Results explained 11](#_Toc186807066)

[6.2 Interpretation of the results 11](#_Toc186807067)

[6.3 Reasons and/or implications for future work, limitations of your study 11](#_Toc186807068)

[7. Reference 12](#_Toc186807069)

[8. Appendices 13](#_Toc186807070)

[R code is used for analysis and visualization analysis. R code with the appropriate statistics to test the hypotheses. 13](#_Toc186807071)

List of Figures

[Figure 1: Histogram of Patient's Age with Bell Curve 7](#_Toc186807072)

[Figure 2: Histogram of Sex with Density Plot 7](#_Toc186807073)

[Figure 3: Scatterplot of Patient sex vs Patient Age 8](#_Toc186807074)

[Figure 4: Results 9](#_Toc186807075)

# 1. Introduction

## 1.1 Problem statement and research motivation

The growth of chronic and infectious diseases demands a clear understanding of how demographics such as age and sex play into the course of disease progression and health outcomes. Much of epidemiology research lacks effective analysis of such variables, so further research required for improvement of these healthcare practices. Understanding patient data can uncover patterns in that information for customized treatment planning and enhanced public health measures. As represented by Strulik and Grossmann, (2024), demographic-specific insights are important in shaping policy that is effective for healthcare. This study aims to fulfill the need for comprehensive analysis, thereby bridging the knowledge gap, and enhancing epidemiological precision using data-driven approaches.

## 1.2 The data set

The dataset contains 5,165 patient records with information on demographics: age, sex, country, and city, as well as clinical factors such as symptoms, confirmed and released dates, and types of infection cases. Most of the fields are well-documented, but there are missing values for some data, like symptom onset dates and deceased dates. The dataset covers different locations and has important variables to analyse demographic effects on health outcomes. This structured dataset offers a rich foundation for exploring epidemiological patterns.

## 1.3 Research Question

The work integrates the research question, involving the analysis of certain demographic variables such as age and sex and the impact these have on disease advancement and health outcomes in epidemiological research. In response, the kind of statistical analysis involving data mining and exploration methodology such as graphing and mapping tools will be used to explore the relative relationship between demographics and patient health as well as treatment outcome.

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

The null hypothesis, H0, states that there is no significant association between demographic factors, such as age and sex, and patient health outcomes. In other words, the null hypothesis holds that these factors do not influence disease progression. On the other hand, the alternative hypothesis, H1, assumes that there are significant associations, indicating that demographic characteristics do affect patient health and treatment outcomes. This research uses statistical tools, including regression analysis and hypothesis testing, to confirm or invalidate the null hypothesis by analysing the dataset. The findings will add to a deeper understanding of what role demographic factors play in shaping epidemiological patterns and healthcare strategies.

# 2. Background research

## 2.1 Research papers (at least 3 relevant to your topic / DS)

Analysis of the three research papers points out the utilization of datasets to drive substantial research advancements. The first study essentially deals with “epidemiological data”, which draws on various datasets such as public health records, disease surveillance reports, and population statistics. They are important for the establishment of trends and risk factors that have to be considered in improving strategies for disease prevention. The second paper talks about “healthcare data analytics, using datasets from EHRs”, patient histories, and medical imaging (Alkhatib, Khoei and Ghapanchi, 2016). Such datasets would be quite vital in enhancing predictive models for patient outcomes, also for optimizing healthcare delivery while minimizing inefficiencies in the operation.

The third research paper examines the “epidemiological methodologies” implemented using datasets such as national registries and longitudinal studies. It makes large-scale analysis possible concerning patterns of diseases and mortality trends, whose implications are critical for informing public health policies (Gokhale et al., 2020). Together, the papers demonstrate how important datasets are in advancing evidence-based research and decision-making. They stress the fact that innovation in healthcare and epidemiology is promoted through well-structured, dependable datasets to achieve better societal outcomes.

## 2.2 Why RQ is of interest (research gap and future directions according to the literature)

The research question poses a great interest because the discovery of gaps in the current knowledge and the possibility of charting the course of future research are important factors. The present research already tends to focus on the issues of combining different datasets into a unified one, dealing with data quality, and handling such problems as data scarcity in medical and pandemic investigations. These gaps suggest opportunities for the development of new methodologies as well as the refining of existing frameworks (Shadbolt et al., 2022). Moreover, the chosen areas can be further investigated in future studies and applied to creating more reliable analytical models, establishing effective cross-disciplinary collaboration to resolve public health issues most efficiently. Also by using innovative technologies such as artificial intelligence and machine learning to fill the gaps between the existing studies.

# 3. Visualisation

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

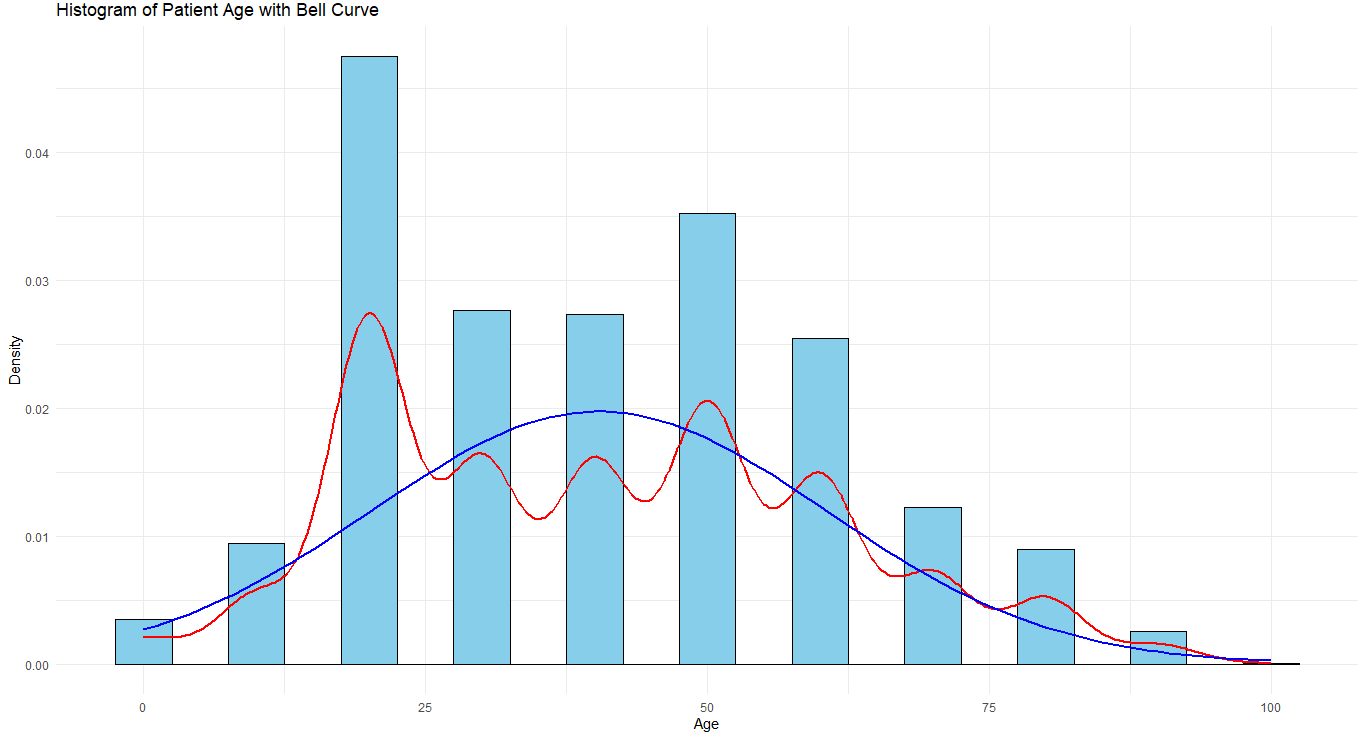


Figure 1: Histogram of Patient's Age with Bell Curve

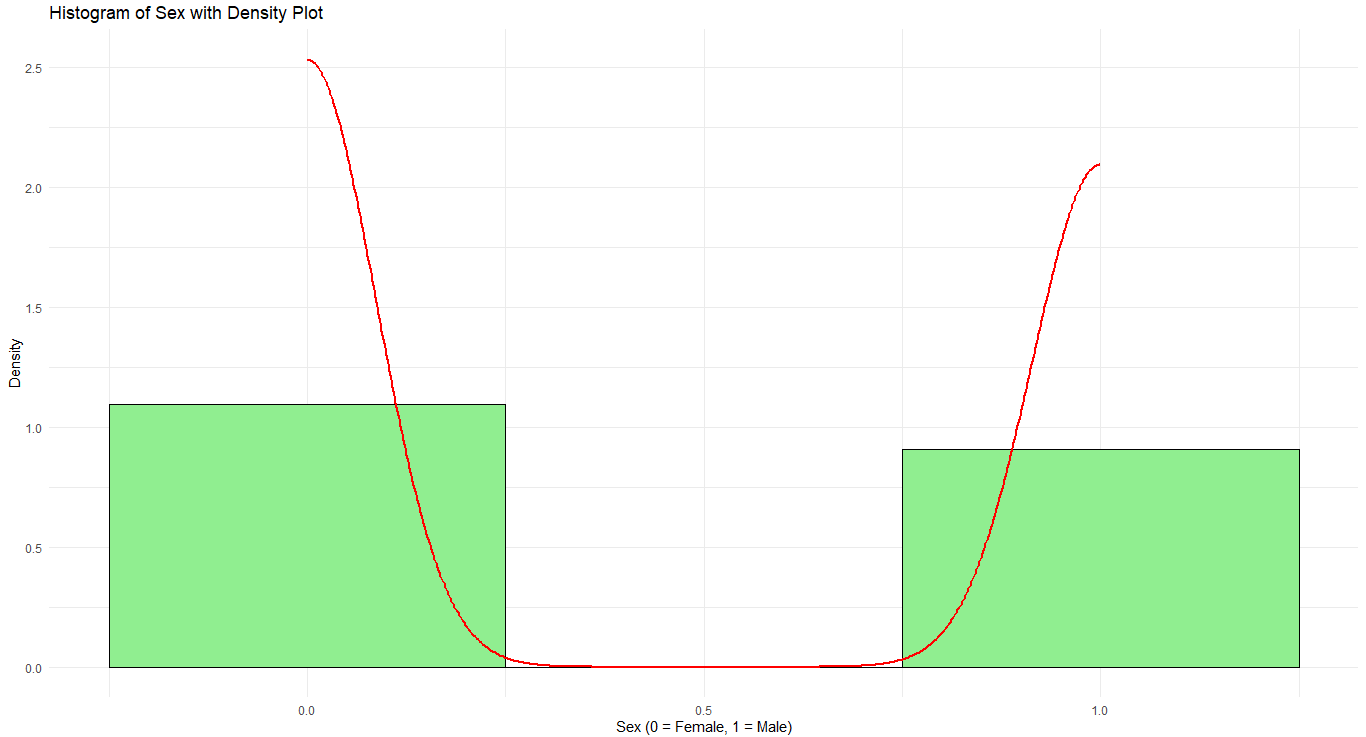


Figure 2: Histogram of Sex with Density Plot

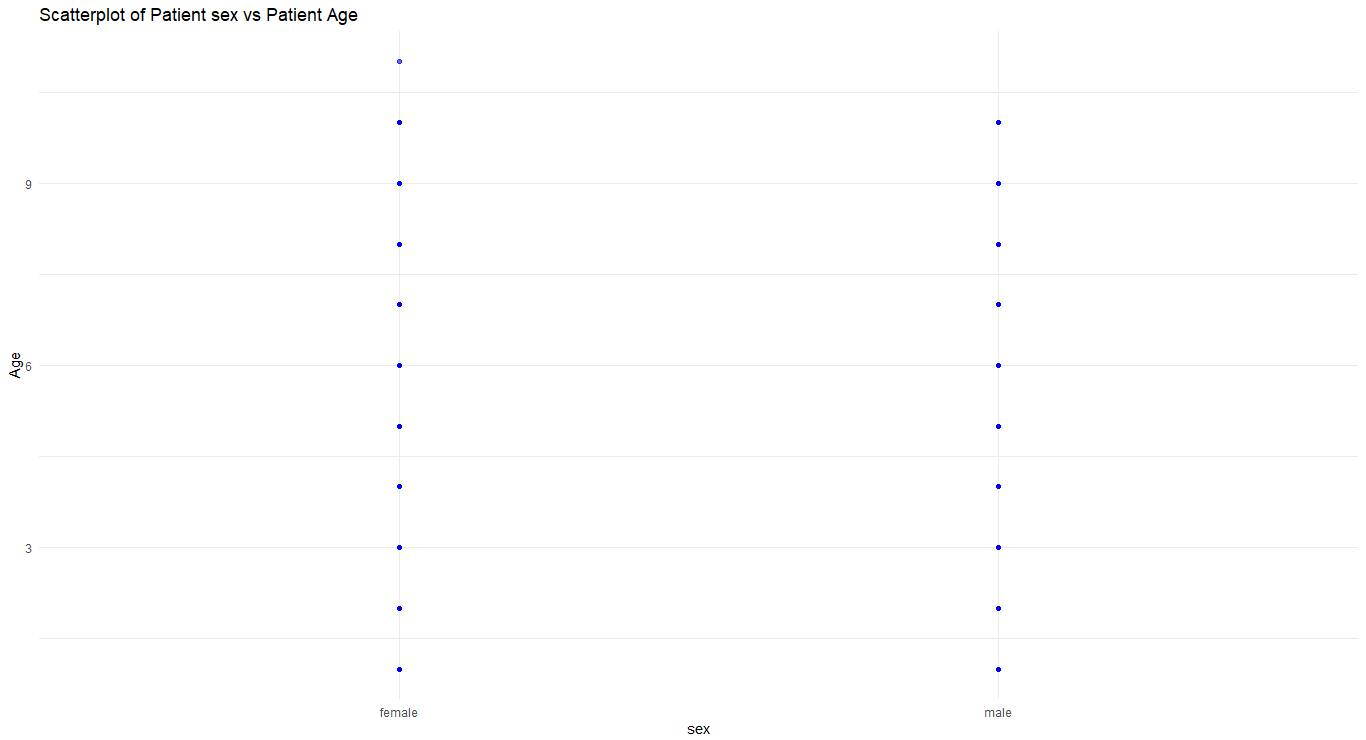


Figure 3: Scatterplot of Patient sex vs Patient Age

This image presents an effective scatterplot of the patient sex vs patient age. Both male and female patients have an age ranging from 0 to 9. Also, there is an effective distribution of ages for male and female. This visualization showcases the even distribution of patient ages across both sexes. Gender and age can be easily compared with the use of the scatterplot presented in the given image. All of these plots meet the necessities of the research question.

## 3.2 Additional information relating to understanding the data (optional)

The histogram depicts the general age distribution, which shows frequency patterns. The density plot illustrates the gender-based age distribution, which clarifies differences or overlaps (Shekapure, Dharmale and Shekapure, 2022). The scatter plot examines the relationship between age and gender, which identifies trends or anomalies. All these visualizations together provide a comprehensive understanding of the data and answer the research question effectively.

## 3.3 Useful information for the data understanding

These plots reveal critical information a histogram shows a biased age distribution that peaks in some age ranges and a density plot shows how men and women cluster at different ages the scatter plot allows the identification of associations between variables, as well as outliers in data, providing additional insight.

# 4. Analysis

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

A correlation was run to test the hypothesis between sex (numerically coded) and age. A correlation test is appropriate because the research question involves examining the relationship between two continuous variables: binary numeric values for age and sex. The test statistic of t = -7.3415, df = 3783 (p-value = 2.572e-13) reveals a statistically significant, although weak, negative correlation. Since sex is a numeric variable used in this test, it is a proper test to assess the association.

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

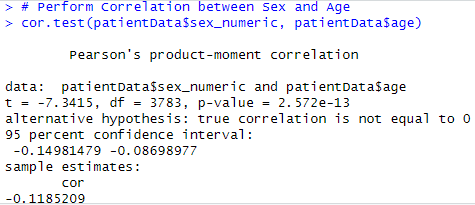


Figure 4: Results

In the null hypothesis, there is no significant association between sex and age. Since the p-value of the above is much smaller than common significance level 0.05 then there is enough reason to reject the null hypothesis. This means that sex and age are related in the dataset statistically significant. The correlation coefficient of -0.12, however, suggests that this is weak relationship. While relationship is significant, but practically insignificant, suggesting that other factors may influence the outcomes more than age and sex alone.

# 5. Evaluation

## 5.1 What Went Well

Throughout the project, strong group collaboration was demonstrated. Every member of the group participated equally and due to this all the tasks are managed on time. The data was cleaned and analysed well and the handling of missing values and consistency of the dataset was clearly communicated amongst the team. The visualizations were popular and members had deep discussions about the results. The group was able to complete tasks and deliverables within the project date and timeline and appropriate results were achieved.

## 5.2 Points for improvement

The team could improve in one area and which is communication in the beginning stages. Task allocation was unclear to some team members, either because the confusion could have been addressed earlier. All the members could also have delved deeper into exploring more in depth the various relationships of demographic factors to health outcomes using advanced statistical methods: for example, regression analysis. Group could have spent more time on exploratory data analysis (EDA) than statistically testing results, to achieve a more holistic view of our findings.

## 5.3 Group’s time management

In terms of time management group was mostly efficient, however, all the members did spend longer than planned on the cleaning and initial exploratory analysis for data. But the group was able to pull it back from the later stages, by managing assignments wisely. The project was completed in timeline and deadlines were matched.

## 5.4 Project’s Overall Judgment

The project was overall successful. The group worked well together, turned out a compelling analysis from the dataset, and made a well-constructed analysis. Things were held up a little in the early part, but all the members stayed focused and reached project goals. The results were very clear, very relevant and very professional.

# 6. Conclusions

## 6.1 Results explained

Correlation test shows a weak negative relationship (as shown by correlation coefficient -0.12) between sex (numerically coded) and age for the dataset. This indicates a slight inverse relationship: One variable tends to increase and the other tends to decrease as one variable increases. It is found that the relationship is very highly significant (>0.001), and so unlikely to be due to chance. The correlation was small, suggesting that sex and age are only weakly correlated, given the dataset.

## 6.2 Interpretation of the results

Results with respect to the research question imply the dataset is weakly dependent on age and sex. Also different is the sex distribution which is slightly greater among younger patients than in older patients. At the same time, the weak correlation means that age and sex do not have a strong correlation, or sex does not account for a huge amount of variation in age. These findings in a broader epidemiologic context suggest that other demographic or health factors may explain more of the variation in health outcomes than is accounted for by age and sex.

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

Other factors like socioeconomic status or underlying health conditions should be even more influential on health outcomes, and so there are more opportunities for future work. They include the poor correlation between the sex and the age, indicating the probable resonance of the relationship is not as strong as was predicted. If the dataset is larger and more diverse, this could provide more conclusive results.

# 7. Reference

‌Alkhatib, M., Khoei, A.T. - and Ghapanchi, A. (2016). Analysis of Research in Healthcare Data Analytics. *arXiv (Cornell University)*. [online] doi:https://doi.org/10.48550/arxiv.1606.01354.

Gokhale, K.M., Chandan, J.S., Toulis, K., Gkoutos, G., Tino, P. and Nirantharakumar, K. (2020). Data extraction for epidemiological research (DExtER): a novel tool for automated clinical epidemiology studies. *European Journal of Epidemiology*, [online] 36(2), pp.165–178. doi:https://doi.org/10.1007/s10654-020-00677-6.

‌Shadbolt, N., Brett, A., Chen, M., Marion, G., McKendrick, I.J., Griffiths, J.P. -, Pellis, L., Reeve, R. and Swallow, B. (2022). The challenges of data in future pandemics. *Epidemics*, [online] 40, pp.100612–100612. doi:https://doi.org/10.1016/j.epidem.2022.100612.

Shekapure, S., Dharmale, G. and Shekapure, N. (2022). Clinical Data Analysis of Patient and Recommendation. *NeuroQuantology*, [online] 20(6), pp.6148–6155. Available at: https://www.researchgate.net/publication/362347728\_Clinical\_Data\_Analysis\_of\_Patient\_and\_Recommendation [Accessed 3 Jan. 2025].

‌Strulik, H. and Grossmann, V. (2024). The economics of aging with infectious and chronic diseases. *Economics & Human Biology*, [online] 52, p.101319. doi:https://doi.org/10.1016/j.ehb.2023.101319.

‌

# 8. Appendices

## R code is used for analysis and visualization analysis. R code with the appropriate statistics to test the hypotheses.

# Load required libraries

library(tidyverse)   # For data manipulation and visualization

library(ggplot2)     # For visualizations

# Load the dataset

patientData <- read\_csv("PatientInfo.csv", show\_col\_types = FALSE)

# Clean column names to make them syntactically valid

patientData <- patientData %>% rename\_with(~make.names(.), everything())

# Print the head of the dataset

head(patientData)

# Print the columns of the dataset

colnames(patientData)

# Check missng values in dataset

colSums(is.na(patientData))

# Impute 'contact\_number' with the median

patientData <- patientData %>%

  mutate(contact\_number = ifelse(is.na(contact\_number),

                                 median(contact\_number, na.rm = TRUE),

                                 contact\_number))

# Impute 'sex' and 'infection\_case' with the mode

calculateMode <- function(x) {

  ux <- na.omit(unique(x))

  ux[which.max(tabulate(match(x, ux)))]

}

patientData <- patientData %>%

  mutate(sex = ifelse(is.na(sex), calculateMode(sex), sex),

         infection\_case = ifelse(is.na(infection\_case), calculateMode(infection\_case), infection\_case))

# Remove missing values in 'state'

patientData <- patientData %>%

  filter(!is.na(state))

# Remove irrelevant or highly missing columns

patientData <- patientData %>%

  select(-c(symptom\_onset\_date, deceased\_date))

# Handle missing 'city' with "Unknown"

patientData <- patientData %>%

  mutate(city = ifelse(is.na(city), "Unknown", city))

# Replace remaining missing values in dates with placeholders

patientData <- patientData %>%

  mutate(across(ends\_with("\_date"), ~ ifelse(is.na(.), as.Date("1900-01-01"), as.Date(.))))

# Replace missing values with in infected\_by with "Unknown"

patientData <- patientData %>%

  mutate(infected\_by = ifelse(is.na(infected\_by), "Unknown", infected\_by))

# Verify Missing Values After Cleaning

colSums(is.na(patientData))

# Print summary of the Cleaned Dataset

summary(patientData)

# Print structure of the Cleaned Dataset

str(patientData)

# Remove non-finite values from age column and convert age to numeric

patientData <- patientData %>%

  filter(!is.na(age) & is.finite(as.numeric(gsub("s", "", age)))) %>%

  mutate(age = as.numeric(gsub("s", "", age)))

# Convert sex to numeric

patientData <- patientData %>% mutate(sex\_numeric = ifelse(sex == "male", 1, 0))

# Plot the Histogram of Age with Bell Curve

ggplot(patientData, aes(x = age)) +

  geom\_histogram(aes(y = after\_stat(density)), binwidth = 5, fill = "skyblue", color = "black") +

  geom\_density(color = "red", linewidth = 1) +

  stat\_function(fun = dnorm,

                args = list(mean = mean(patientData$age, na.rm = TRUE),

                            sd = sd(patientData$age, na.rm = TRUE)),

                color = "blue", linewidth = 1) +

  labs(title = "Histogram of Patient Age with Bell Curve",

       x = "Age", y = "Density") +

  theme\_minimal()

# Plot the Histogram of Sex with Density Plot

ggplot(patientData, aes(x = sex\_numeric)) +

  geom\_histogram(aes(y = ..density..), binwidth = 0.5, fill = "lightgreen", color = "black") +

  geom\_density(color = "red", linewidth = 1) +

  labs(title = "Histogram of Sex with Density Plot",

       x = "Sex (0 = Female, 1 = Male)", y = "Density") +

  theme\_minimal()

# Plot Scatterplot of sex vs Age with Trend Line

ggplot(patientData, aes(x = sex, y = as.numeric(as.factor(age)))) +

  geom\_point(color = "blue", alpha = 0.6) +

  geom\_smooth(method = "lm", color = "red", se = FALSE) +

  labs(title = "Scatterplot of Patient sex vs Patient Age",

       x = "sex", y = "Age") +

  theme\_minimal()

# Perform Correlation between Sex and Age

cor.test(patientData$sex\_numeric, patientData$age)