MA4202-MA7202 Introduction to Functional Data Analysis 2024-25

Mini Project

Table of Contents

Introduction

The study seeks to relate various trends and statistical differences in cancer incidence across different age cohorts and periods. The dataset is the breast cancer incidence rates of Australia for the ages from 1921 to 2001. The "fda package" in R is used to analyse temporal trends and age-specific variation in cancer rates.

Research Question

- What is the spread of the cancer rates by age group over time?
- Are there big differences between cancer rates among the age groups over time?
- What has been the overall trend of breast cancer rates for all age groups combined over time?
- How much variation in cancer rates within each age group exists over time?
- Are there any special events that drive an increase in cancer rates over certain periods?

Clarification Of The Choice Of Method

The choice of methods in this analysis follows from the desire to study both temporal trends and temporal differences in incidence rates in the population of women with breast cancer. The "fda package" is used to facilitate the handling and analysis of functional data, which is well-suited for analysing trends over time. It is used in standard deviation calculations to quantify each age group's variability with time, since the cancer rate is always constant within each cohort (Clift *et al.*, 2023).

Motivations For Selecting Such Methods

The "fda package" is taken as the time series dataset in this study's functional data analysis. The preferred method to see if the differences between age groups over time are significant, as well as whether any of the observed patterns are statistically significant, is "ANOVA". Within age groups, further interpretation of the data is done by quantification of variability using standard deviation analysis (Teles *et al.*, 2021).

Full Method Description

The trend and variance in breast cancer rate are investigated by making use of interactions of the "fda", "ANOVA", and data visualisation techniques. The data is plotted by the cancer rates by age group over time to reveal any tendencies, visualising the data for Research Question 1 (Zhou *et al.*, 2022). Statistical differences of cancer rates across age groups are calculated by analysing with

"ANOVA" to test Research Question 2. The dataset is split into two periods before and after 1950, and the mean cancer rates per period are compared for Research Question 5.

Justification Of Results

```
List of 5
 $ x    : num [1:9] 47 52 57 62 67 72 77 82 87
 $ y    : num [1:9, 1:81] 33.5 59.1 49.8 55.8 56 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:9] "47" "52" "57" "62" ...
  .. ..$ : chr [1:81] "1921" "1922" "1923" "1924" ...
 $ time : Time-Series [1:81] from 1921 to 2001: 1921 1922 1923 1924 1925 ...
 $ xname: chr "Age"
 $ yname: chr "Cancer rate"
 - attr(*, "class")= chr [1:2] "fts" "fds"
```

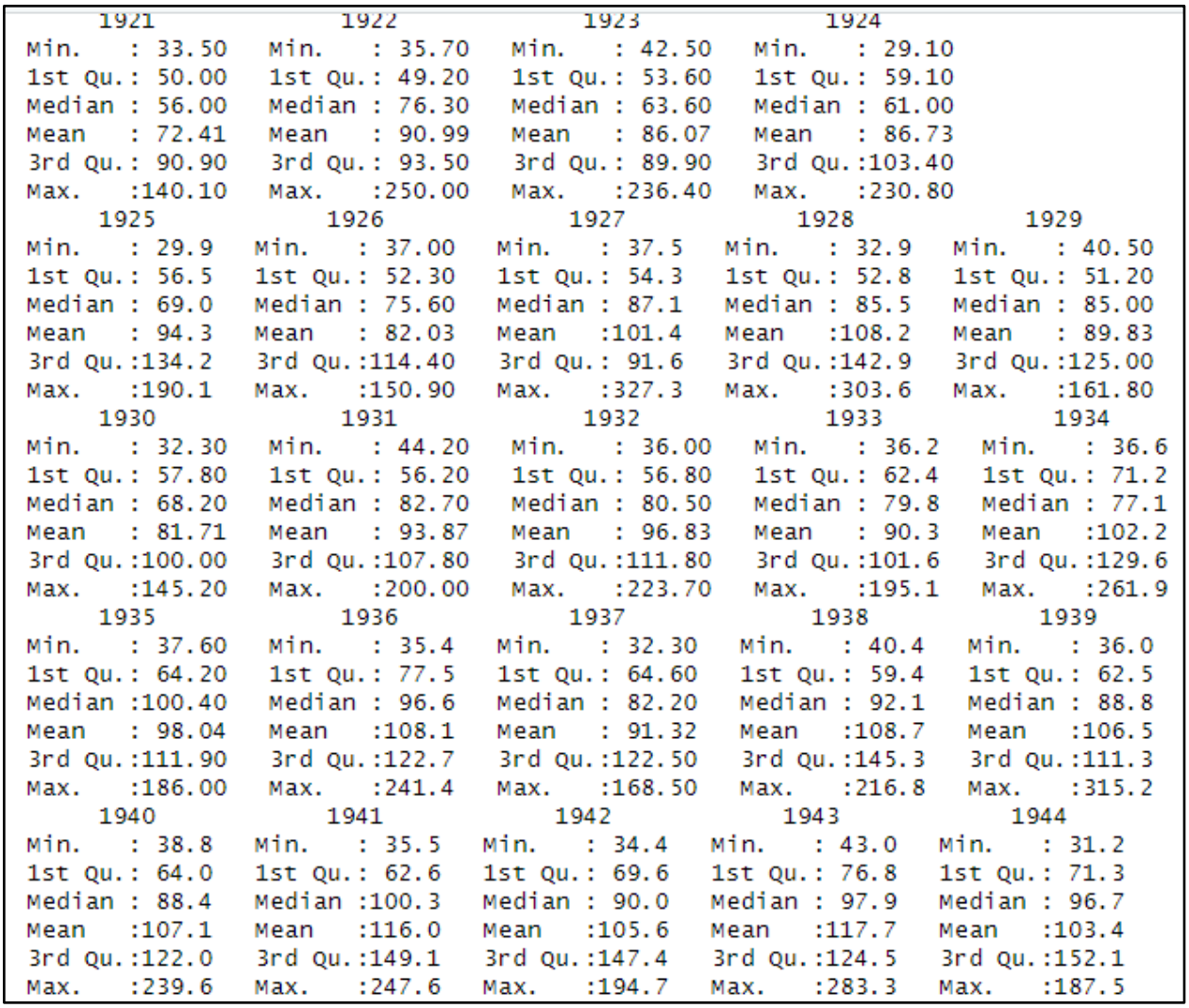
Figure 1: Structure of the dataset

The structure of the dataset is shown in the above figure, and the variables and data types of the dataset are displayed. It is a description of the dataset organization based on the rates of breast cancer across the age groups.

```
> # Checking for missing values in the entire dataset
> sum(is.na(Cancerrate))
[1] 0
```

Figure 2: Checking for missing values

The above figure illustrates that the missing values with null appearance are indicated by the check in the dataset. It ensures that the data that is used in the analysis is complete and reliable.



1921	1922	1923	1924	
Min. : 33.50	Min. : 35.70	Min. : 42.50	Min. : 29.10	
1st Qu.: 50.00	1st Qu.: 49.20	1st Qu.: 53.60	1st Qu.: 59.10	
Median : 56.00	Median : 76.30	Median : 63.60	Median : 61.00	
Mean : 72.41	Mean : 90.99	Mean : 86.07	Mean : 86.73	
3rd Qu.: 90.90	3rd Qu.: 93.50	3rd Qu.: 89.90	3rd Qu.:103.40	
Max. :140.10	Max. :250.00	Max. :236.40	Max. :230.80	
1925	1926	1927	1928	1929
Min. : 29.9	Min. : 37.00	Min. : 37.5	Min. : 32.9	Min. : 40.50
1st Qu.: 56.5	1st Qu.: 52.30	1st Qu.: 54.3	1st Qu.: 52.8	1st Qu.: 51.20
Median : 69.0	Median : 75.60	Median : 87.1	Median : 85.5	Median : 85.00
Mean : 94.3	Mean : 82.03	Mean :101.4	Mean :108.2	Mean : 89.83
3rd Qu.:134.2	3rd Qu.:114.40	3rd Qu.: 91.6	3rd Qu.:142.9	3rd Qu.:125.00
Max. :190.1	Max. :150.90	Max. :327.3	Max. :303.6	Max. :161.80
1930	1931	1932	1933	1934
Min. : 32.30	Min. : 44.20	Min. : 36.00	Min. : 36.2	Min. : 36.6
1st Qu.: 57.80	1st Qu.: 56.20	1st Qu.: 56.80	1st Qu.: 62.4	1st Qu.: 71.2
Median : 68.20	Median : 82.70	Median : 80.50	Median : 79.8	Median : 77.1
Mean : 81.71	Mean : 93.87	Mean : 96.83	Mean : 90.3	Mean :102.2
3rd Qu.:100.00	3rd Qu.:107.80	3rd Qu.:111.80	3rd Qu.:101.6	3rd Qu.:129.6
Max. :145.20	Max. :200.00	Max. :223.70	Max. :195.1	Max. :261.9
1935	1936	1937	1938	1939
Min. : 37.60	Min. : 35.4	Min. : 32.30	Min. : 40.4	Min. : 36.0
1st Qu.: 64.20	1st Qu.: 77.5	1st Qu.: 64.60	1st Qu.: 59.4	1st Qu.: 62.5
Median :100.40	Median : 96.6	Median : 82.20	Median : 92.1	Median : 88.8
Mean : 98.04	Mean :108.1	Mean : 91.32	Mean :108.7	Mean :106.5
3rd Qu.:111.90	3rd Qu.:122.7	3rd Qu.:122.50	3rd Qu.:145.3	3rd Qu.:111.3
Max. :186.00	Max. :241.4	Max. :168.50	Max. :216.8	Max. :315.2
1940	1941	1942	1943	1944
Min. : 38.8	Min. : 35.5	Min. : 34.4	Min. : 43.0	Min. : 31.2
1st Qu.: 64.0	1st Qu.: 62.6	1st Qu.: 69.6	1st Qu.: 76.8	1st Qu.: 71.3
Median : 88.4	Median :100.3	Median : 90.0	Median : 97.9	Median : 96.7
Mean :107.1	Mean :116.0	Mean :105.6	Mean :117.7	Mean :103.4
3rd Qu.:122.0	3rd Qu.:149.1	3rd Qu.:147.4	3rd Qu.:124.5	3rd Qu.:152.1
Max. :239.6	Max. :247.6	Max. :194.7	Max. :283.3	Max. :187.5

Figure 3: Summary statistics

The above figure shows that summary statistics consist of the cancer rates, measures such as mean, median, and standard deviation. It explains the spread of breast cancer rates across age groups.

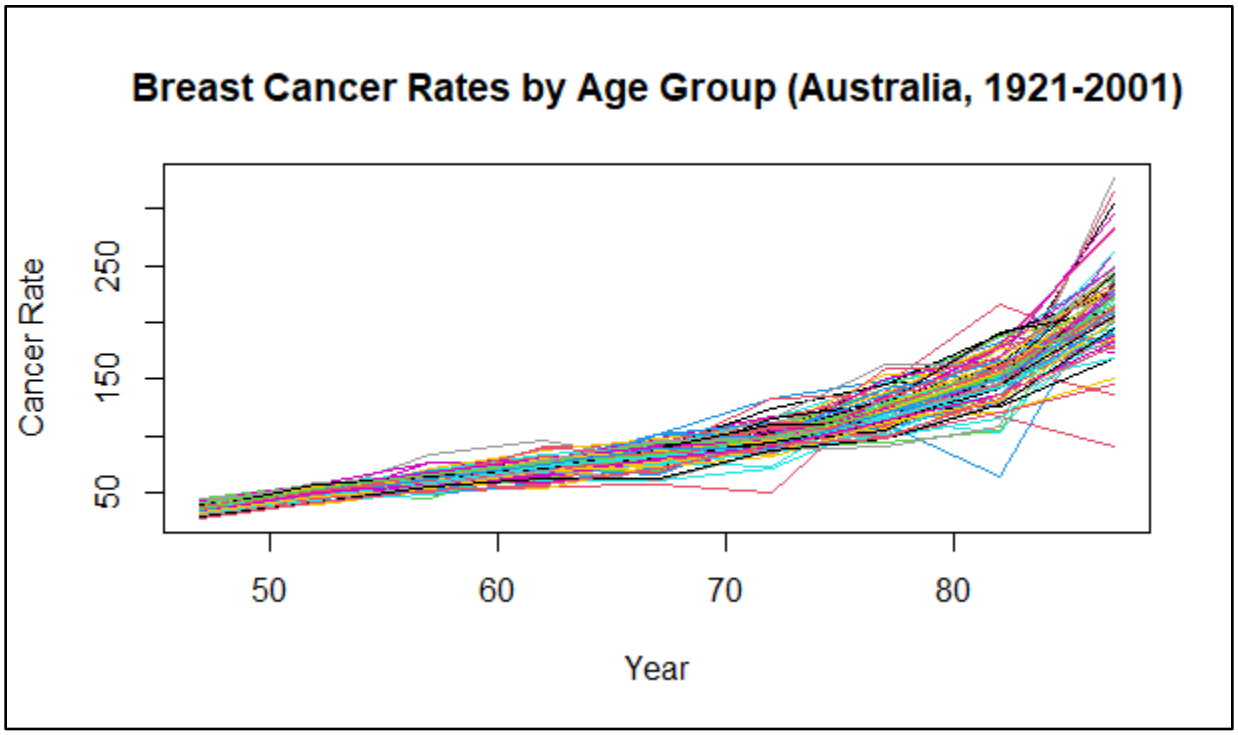


Figure 4: Breast Cancer Rates by Age Group (Australia, 1921-2001)

The above figure shows the trend of breast cancer rates over the various age groups through time. It visualizes those variations in rate and depicts the difference between different age groups.

```
             Df  Sum Sq Mean Sq  F value   Pr(>F)
AgeGroup      8 2099037  262380 1021.544  < 2e-16 ***
Year         80   49226     615    2.396 2.32e-09 ***
Residuals   640  164381     257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: ANOVA result

The above figure shows the result of ANOVA that compares the rates of cancer between age groups in a statistical way. It determines whether or not there are substantial differences in rates for the series of years studied.

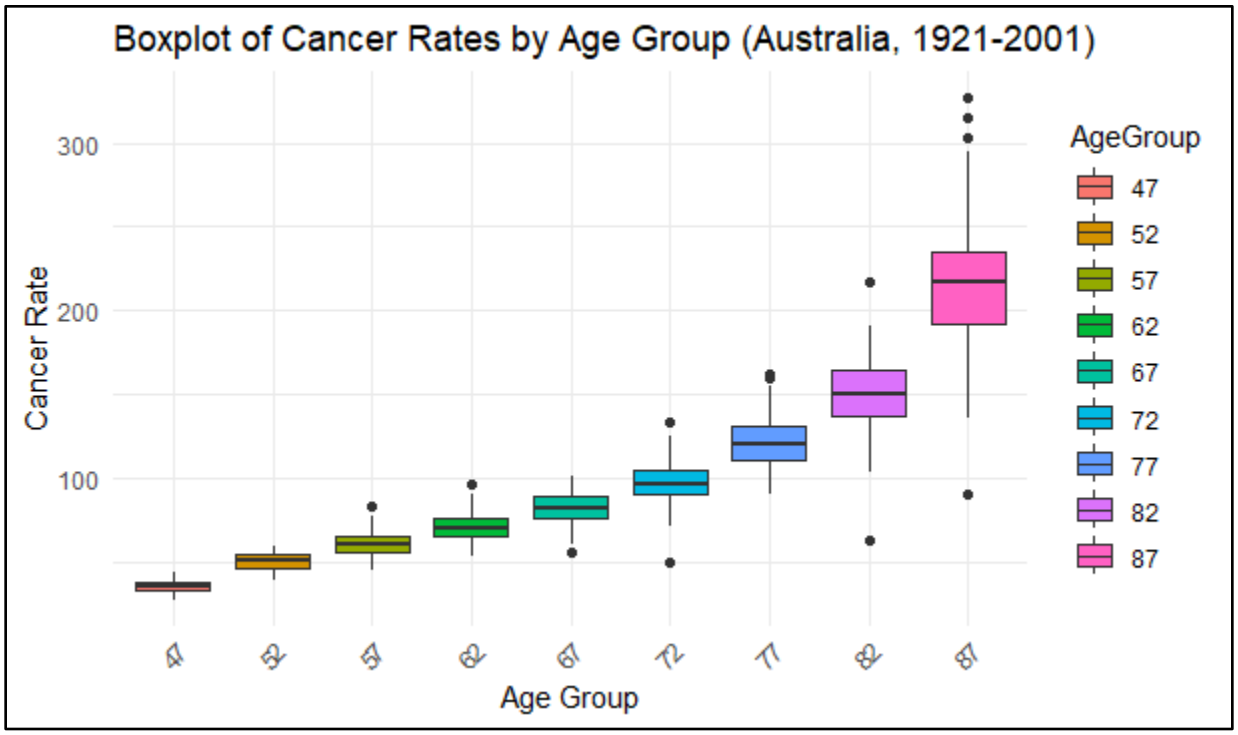


Figure 6: Boxplot of Cancer Rates by Age Group (Australia, 1921-2001)

A boxplot of cancer rates among different age groups is given in the above figure. It visually summarizes the distribution of the cancer rates and picks out outliers or major trends in the data.

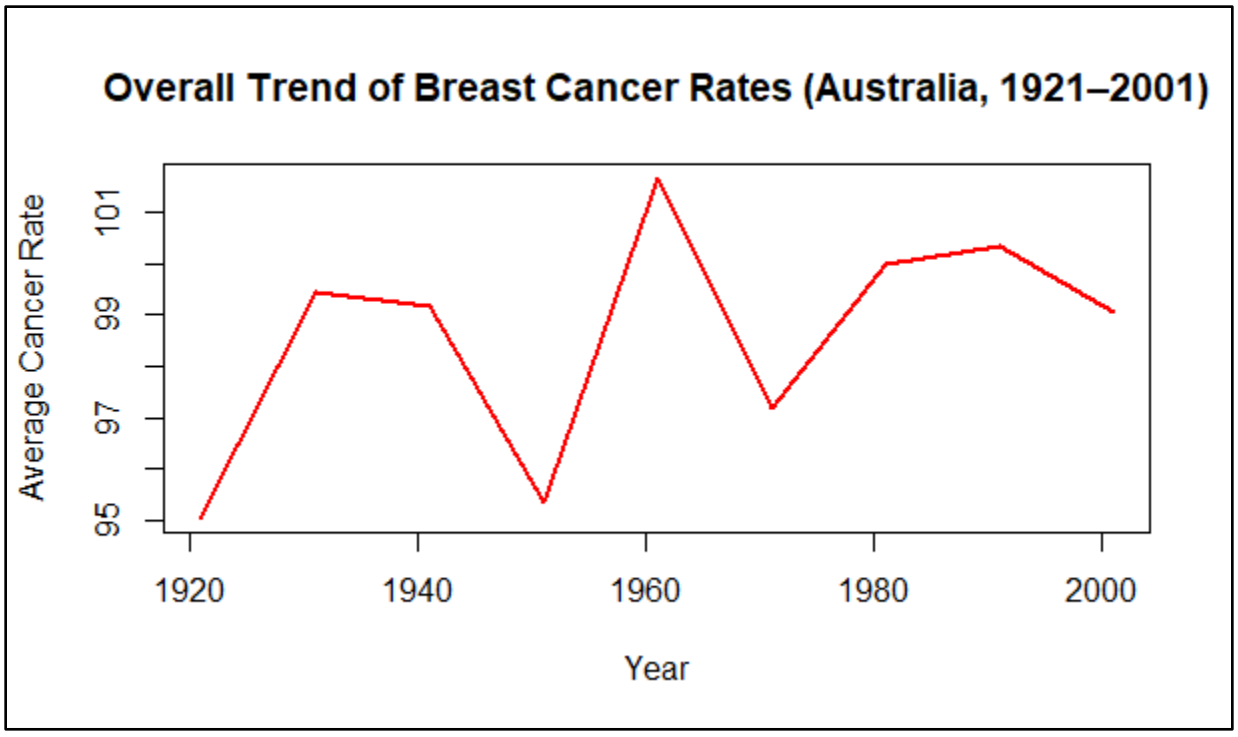


Figure 7: Overall Trend of Breast Cancer Rates (Australia, 1921-2001)

The above figure shows the global trend of the rate of incidence of breast cancer. It displays the combined cancer rate by all age groups and the general pattern, as well as the fluctuations in the decades.

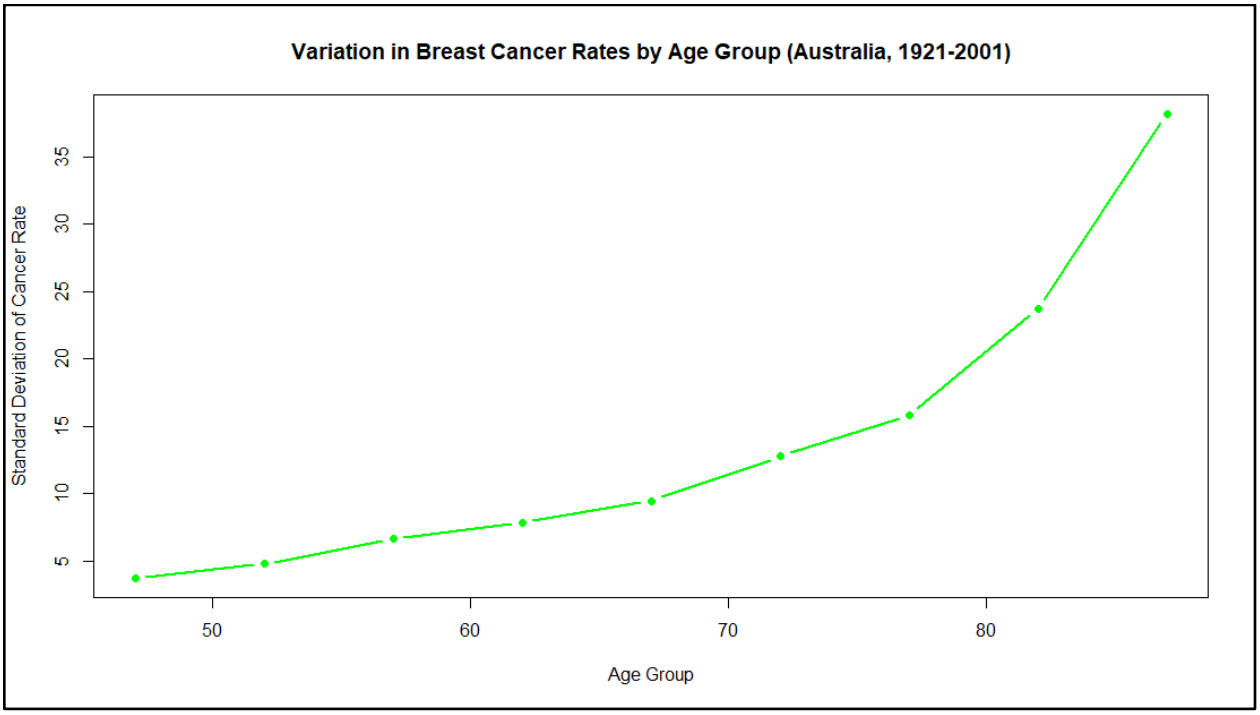


Figure 8: Variation in Breast Cancer Rates by Age Group (Australia, 1921-2001)

The above figure shows that cancer rates are displayed by age group. The rate standard deviations show the degree of variability across each age group over time.

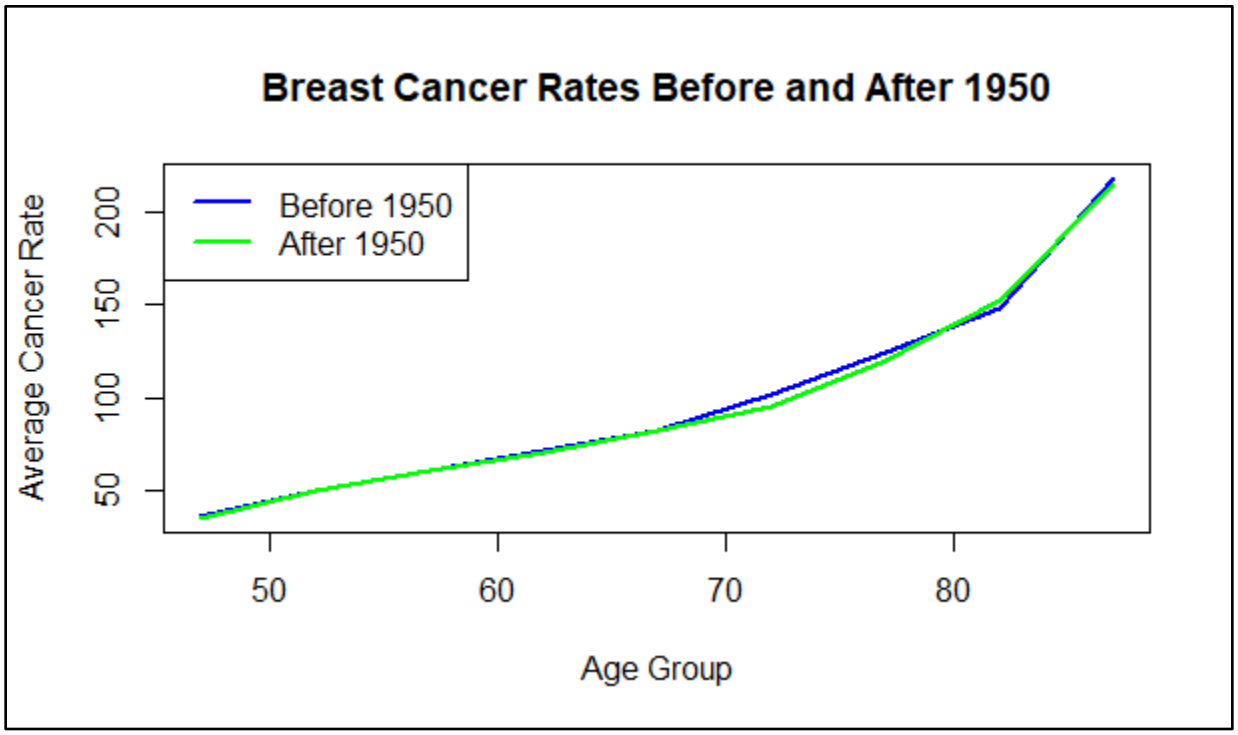


Figure 9: Breast Cancer Rates Before and After 1950

The above figure shows the average cancer rates in each age group before and after 1950. It finds trends and changes in the rates in the two separate periods within the dataset.

R Code Used To Obtain Results Through The Chosen Methods

```
# Installing the fda package
install.packages("fda")

# Loading the fda package
library(fda)

# Loading the dataset
data(Cancerrate)

# Viewing the structure
str(Cancerrate)


# Viewing dataset in spreadsheet-style view
View(Cancerrate)

# Viewing the first 10 rows of the dataset
head(Cancerrate, 10)
```

```
# Checking for missing values in the entire dataset
sum(is.na(Cancerrate))

# Summary statistics
summary(Cancerrate$y)

# Research Question 1. What is the spread of the cancer rates by age group over time?

# Plotting the cancer rates for each age group to see the trends over time
plot(Cancerrate, main = "Breast Cancer Rates by Age Group (Australia, 1921-2001)",
    xlab = "Year", ylab = "Cancer Rate", col = 2:10, lty = 1)

# Research Question 2. Are there big differences between cancer rates among the age groups
over time?

cancer_matrix <- Cancerrate$y

cancer_matrix_transposed <- t(cancer_matrix)

yearVec <- Cancerrate$argvals
age_groups <- seq(47, 87, by = 5)

cancer_df <- as.data.frame(cancer_matrix_transposed)
colnames(cancer_df) <- paste0("Age_", age_groups)
cancer_df$Year <- yearVec

install.packages("reshape2")
library(reshape2)

str(cancer_df)


# Adding the Year column
cancer_df$Year <- seq(1921, 2001, by = 10)


# Add the year column (as above)
cancer_df$Year <- seq(1921, 2001, by = 10)

# Melt the data from wide to long format
cancer_data_long <- melt(cancer_df, id.vars = "Year",
                         variable.name = "AgeGroup",
                         value.name = "CancerRate")
```

```
# Performing ANOVA to check differences between age groups for each year
anova_results <- aov(CancerRate ~ AgeGroup + Year, data = cancer_data_long)

# Summary of the ANOVA results
summary(anova_results)


# Loading the heatmap library
library(ggplot2)

# Creating the boxplot
ggplot(cancer_data_long, aes(x = AgeGroup, y = CancerRate, fill = AgeGroup)) +
  geom_boxplot() +
  labs(title = "Boxplot of Cancer Rates by Age Group (Australia, 1921-2001)",
       x = "Age Group", y = "Cancer Rate") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Research Question 3. What has been the overall trend of breast cancer rates for all age
groups combined over time?

# Computing the mean across all age groups for each year
mean_cancer_rate <- apply(cancer_matrix_transposed, 1, mean)

# Plotting the overall trend of cancer rates over time
unique_years <- unique(cancer_data_long$Year)
mean_cancer_rate <- tapply(cancer_data_long$CancerRate, cancer_data_long$Year, mean)
length(mean_cancer_rate)
length(unique_years)
plot(unique_years, mean_cancer_rate, type = "l", col = "red", lwd = 2,
     main = "Overall Trend of Breast Cancer Rates (Australia, 1921–2001)",
     xlab = "Year", ylab = "Average Cancer Rate")


# Research Question 4. How much variation in cancer rates within each age group exists over
time?

# Calculating the standard deviation of cancer rates for each age group across all years
std_dev_cancer_rate <- apply(cancer_matrix, 1, sd)

# Ensuring that the length of std_dev_cancer_rate matches the number of age groups
age_groups <- seq(47, 87, by = 5)

# Checking the lengths to ensure they match
length(age_groups)
length(std_dev_cancer_rate)
```

```
# Plotting the standard deviation for each age group
plot(age_groups, std_dev_cancer_rate, type = "b", col = "green", pch = 16, lwd = 2,
    main = "Variation in Breast Cancer Rates by Age Group (Australia, 1921-2001)",
    xlab = "Age Group", ylab = "Standard Deviation of Cancer Rate")

# Research Question 5. Are there any special events that drive an increase in cancer rates over certain periods?

# Splitting the data into two periods: before 1950 and after 1950
before_1950 <- yearVec < 1950
after_1950 <- yearVec >= 1950

# Checking if the length of yearVec matches the number of columns in the matrix
length(yearVec)
dim(cancer_matrix_transposed)

# Defining the years before and after 1950
before_1950 <- yearVec < 1950
after_1950 <- yearVec >= 1950

# Ensuring that the logical vectors have the correct length
sum(before_1950)
sum(after_1950)

# Calculating the mean cancer rate for each age group before 1950
mean_before_1950 <- apply(cancer_matrix_transposed[before_1950, ], 2, mean)

# Calculating the mean cancer rate for each age group after 1950
mean_after_1950 <- apply(cancer_matrix_transposed[after_1950, ], 2, mean)

# Viewing the calculated means
mean_before_1950
mean_after_1950


# Checking the lengths of the vectors
length(mean_before_1950)
length(mean_after_1950)


mean_before_1950
mean_after_1950

# Checking for any missing values
sum(is.na(mean_before_1950))
```

```
sum(is.na(mean_after_1950))

# Age groups
age_groups <- c(47, 52, 57, 62, 67, 72, 77, 82, 87)

# Checking the lengths
length(age_groups)
length(mean_before_1950)
length(mean_after_1950)

summary(mean_before_1950)
summary(mean_after_1950)

sum(is.finite(mean_before_1950))
sum(is.finite(mean_after_1950))

# Ensure 'Year' exists in the long-form data
head(cancer_data_long)

# Convert Year to numeric if needed
cancer_data_long$Year <- as.numeric(as.character(cancer_data_long$Year))

# Get means for each AgeGroup before and after 1950
mean_before_1950 <- tapply(cancer_data_long$CancerRate[cancer_data_long$Year < 1950],
                           cancer_data_long$AgeGroup[cancer_data_long$Year < 1950],
                           mean, na.rm = TRUE)

mean_after_1950 <- tapply(cancer_data_long$CancerRate[cancer_data_long$Year >= 1950],
                          cancer_data_long$AgeGroup[cancer_data_long$Year >= 1950],
                          mean, na.rm = TRUE)



# Plotting the data
common_age_groups <- intersect(names(mean_before_1950), names(mean_after_1950))

plot(as.numeric(gsub("Age_", "", common_age_groups)),
     mean_before_1950[common_age_groups],
     type = "l", col = "blue", lwd = 2,
     xlim = c(47, 87),
     ylim = range(c(mean_before_1950[common_age_groups],
mean_after_1950[common_age_groups])),
     main = "Breast Cancer Rates Before and After 1950",
     xlab = "Age Group", ylab = "Average Cancer Rate")

lines(as.numeric(gsub("Age_", "", common_age_groups)),
```

```
    mean_after_1950[common_age_groups], col = "red", lwd = 2)

legend("topleft", legend = c("Before 1950", "After 1950"), col = c("blue", "red"), lwd = 2)
```

Conclusion

Significant trends in breast cancer rates have been found across different age groups in Australia from 1921 to 2001. The research shows that cancer rates are both over time and amongst age groups. The study shows interesting patterns of change in breast cancer rates and overall has some important implications for continuing research and targeted health interventions.

References

Clift, A.K., Dodwell, D., Lord, S., Petrou, S., Brady, M., Collins, G.S. and Hippisley-Cox, J., 2023. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. *bmj*, *381*.

Teles, R.H.G., Yano, R.S., Villarinho, N.J., Yamagata, A.S., Jaeger, R.G., Meybohm, P., Burek, M. and Freitas, V.M., 2021. Advances in breast cancer management and extracellular vesicle research, a bibliometric analysis. *Current Oncology*, *28*(6), pp.4504-4520.

Zhou, Y., Jia, N., Ding, M. and Yuan, K., 2022. Effects of exercise on inflammatory factors and IGF system in breast cancer survivors: a meta-analysis. *BMC Women's Health*, *22*(1), p.507.