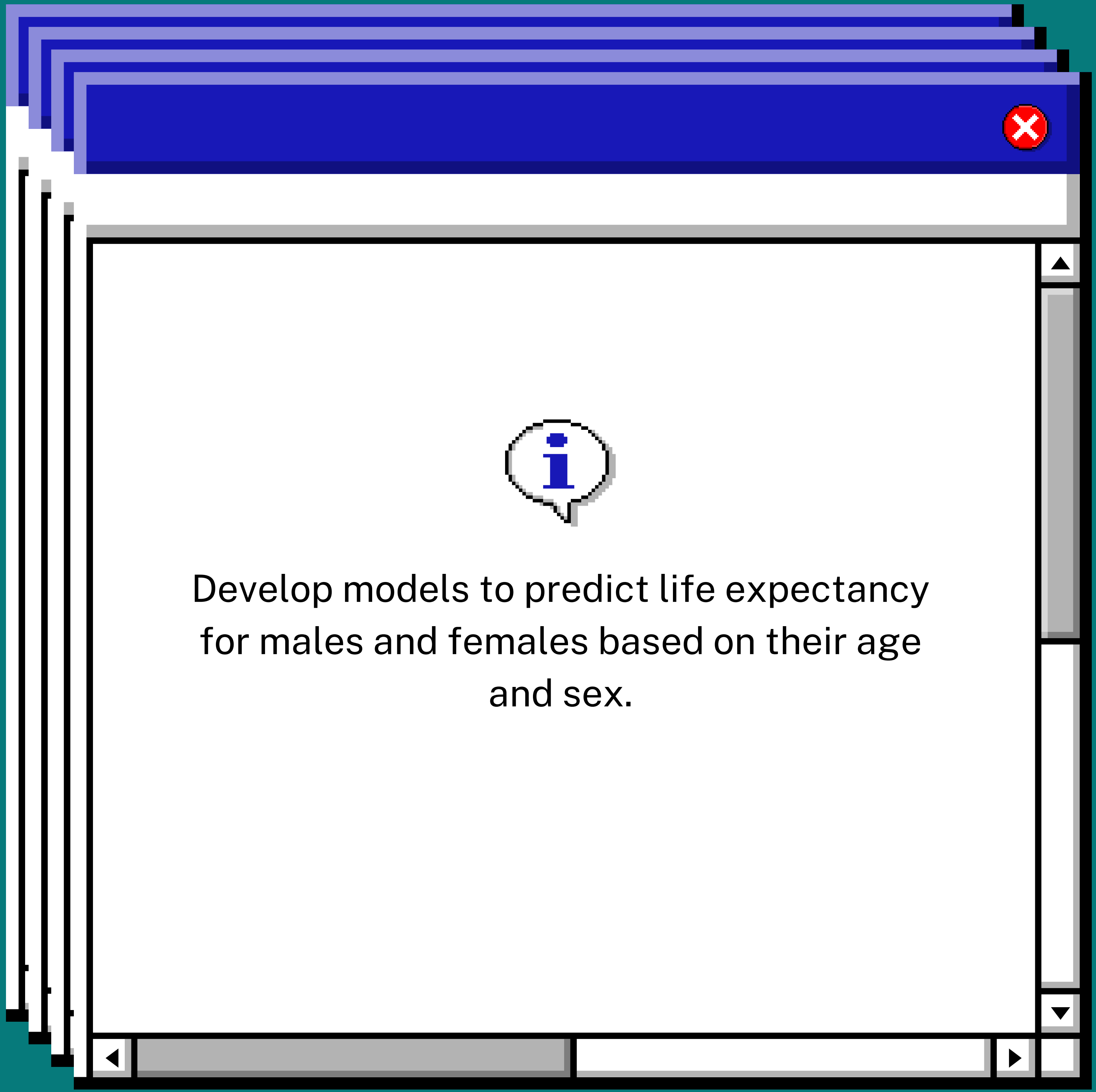
Life expectancy based on age and sex

A hopefully not too boring presentation.

11:11PM

Problem Statement



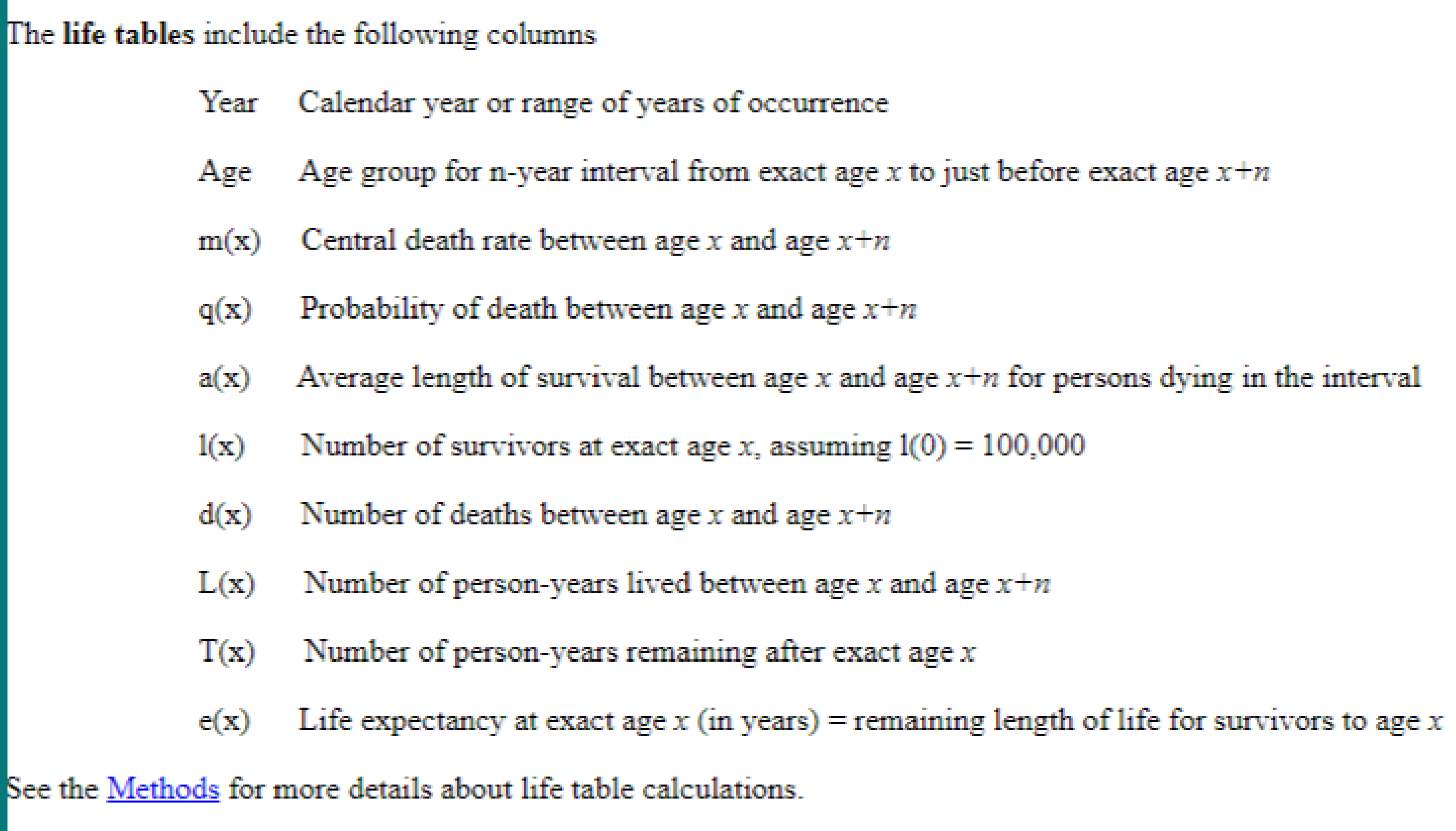
Develop models to predict life expectancy for males and females based on their age and sex.

Back

Data cleaning and Wrangling

Back

Dataset



The **life tables** include the following columns

Year Calendar year or range of years of occurrence

Age Age group for n-year interval from exact age *x* to just before exact age *x+n*

m(x) Central death rate between age *x* and age *x+n*

q(x) Probability of death between age *x* and age *x+n*

a(x) Average length of survival between age *x* and age *x+n* for persons dying in the interval

l(x) Number of survivors at exact age *x*, assuming l(0) = 100,000

d(x) Number of deaths between age *x* and age *x+n*

L(x) Number of person-years lived between age *x* and age *x+n*

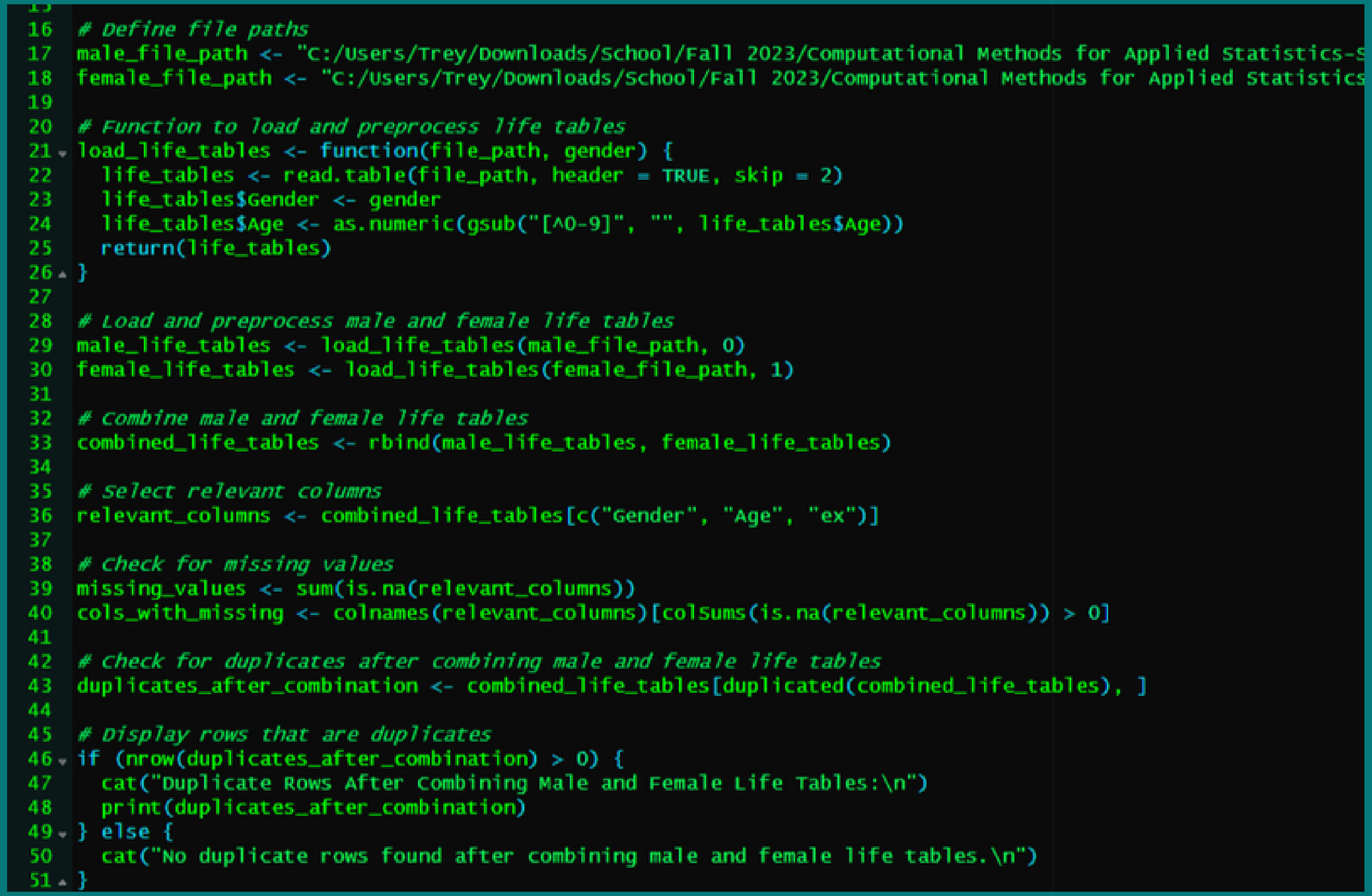
T(x) Number of person-years remaining after exact age *x*

e(x) Life expectancy at exact age *x* (in years) = remaining length of life for survivors to age *x*

See the Methods for more details about life table calculations.

Utilized life tables

I used the life tables found at the Human Mortality Database. I then limited my data to the relevant columns. Being Age, Sex, and e(x)-Life Expectancy.



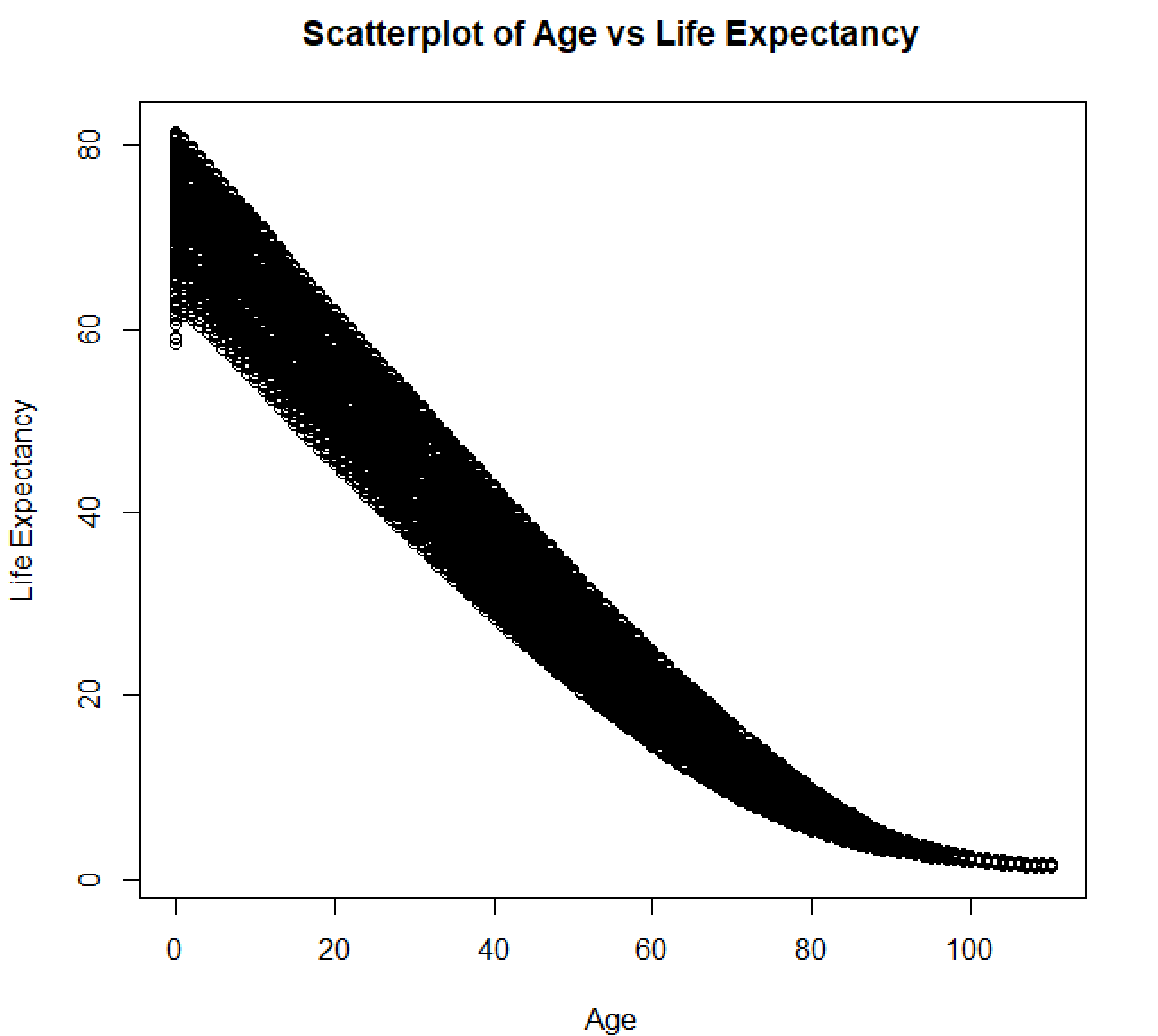
```r
# Define file paths
male_file_path <- "C:/Users/Trey/Downloads/School/Fall 2023/Computational Methods for Applied Statistics-S
female_file_path <- "C:/Users/Trey/Downloads/School/Fall 2023/Computational Methods for Applied Statistics

# Function to load and preprocess life tables
load_life_tables <- function(file_path, gender) {
  life_tables <- read.table(file_path, header = TRUE, skip = 2)
  life_tables$Gender <- gender
  life_tables$Age <- as.numeric(gsub("[^0-9]", "", life_tables$Age))
  return(life_tables)
}

# Load and preprocess male and female life tables
male_life_tables <- load_life_tables(male_file_path, 0)
female_life_tables <- load_life_tables(female_file_path, 1)

# Combine male and female life tables
combined_life_tables <- rbind(male_life_tables, female_life_tables)

# Select relevant columns
relevant_columns <- combined_life_tables[c("Gender", "Age", "ex")]

# Check for missing values
missing_values <- sum(is.na(relevant_columns))
cols_with_missing <- colnames(relevant_columns)[colSums(is.na(relevant_columns)) > 0]

# Check for duplicates after combining male and female life tables
duplicates_after_combination <- combined_life_tables[duplicated(combined_life_tables), ]

# Display rows that are duplicates
if (nrow(duplicates_after_combination) > 0) {
  cat("Duplicate Rows After Combining Male and Female Life Tables:\n")
  print(duplicates_after_combination)
} else {
  cat("No duplicate rows found after combining male and female life tables.\n")
}
```

Code used

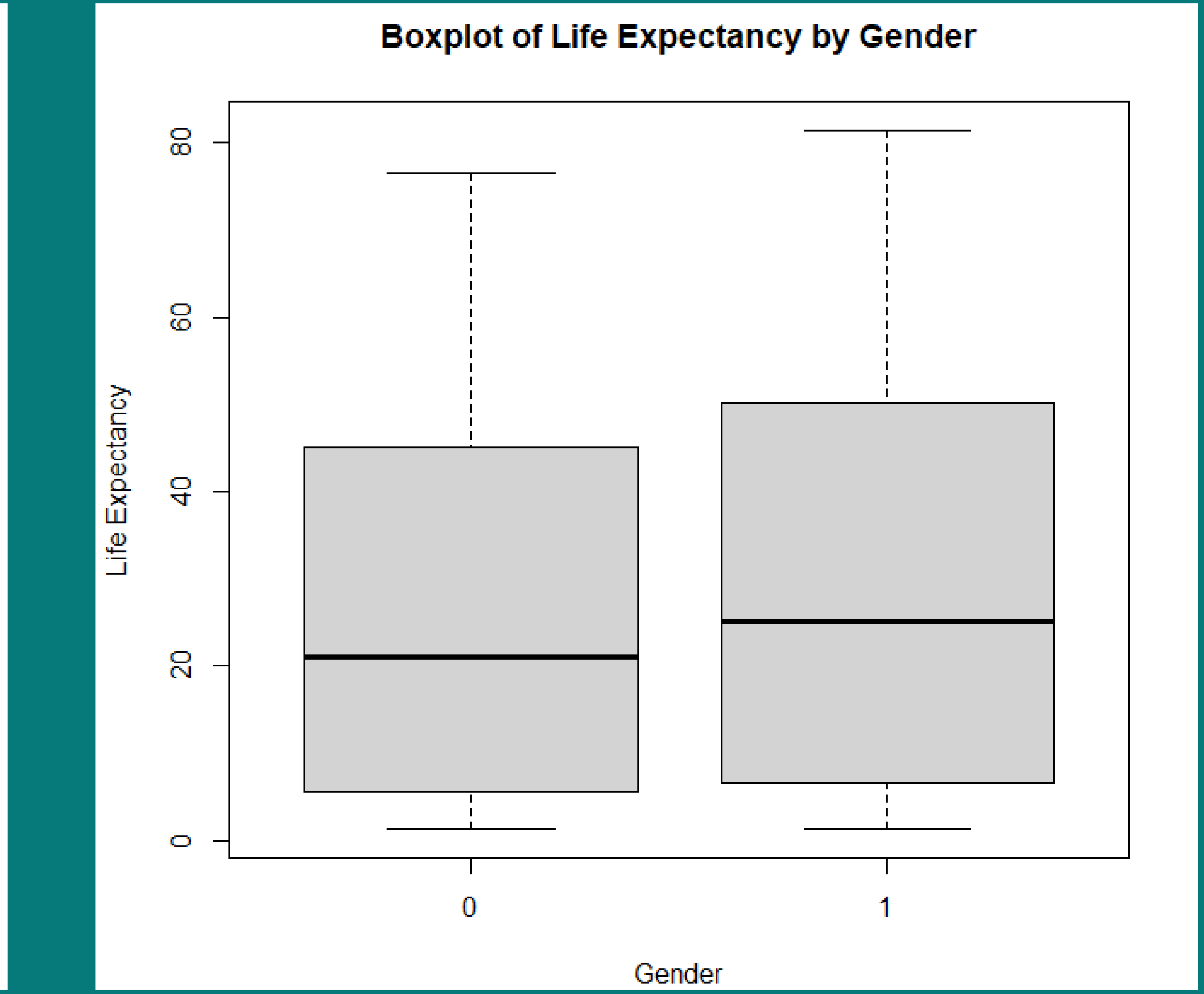
I had to combine the life tables of USA females and males then I found no missing

Back

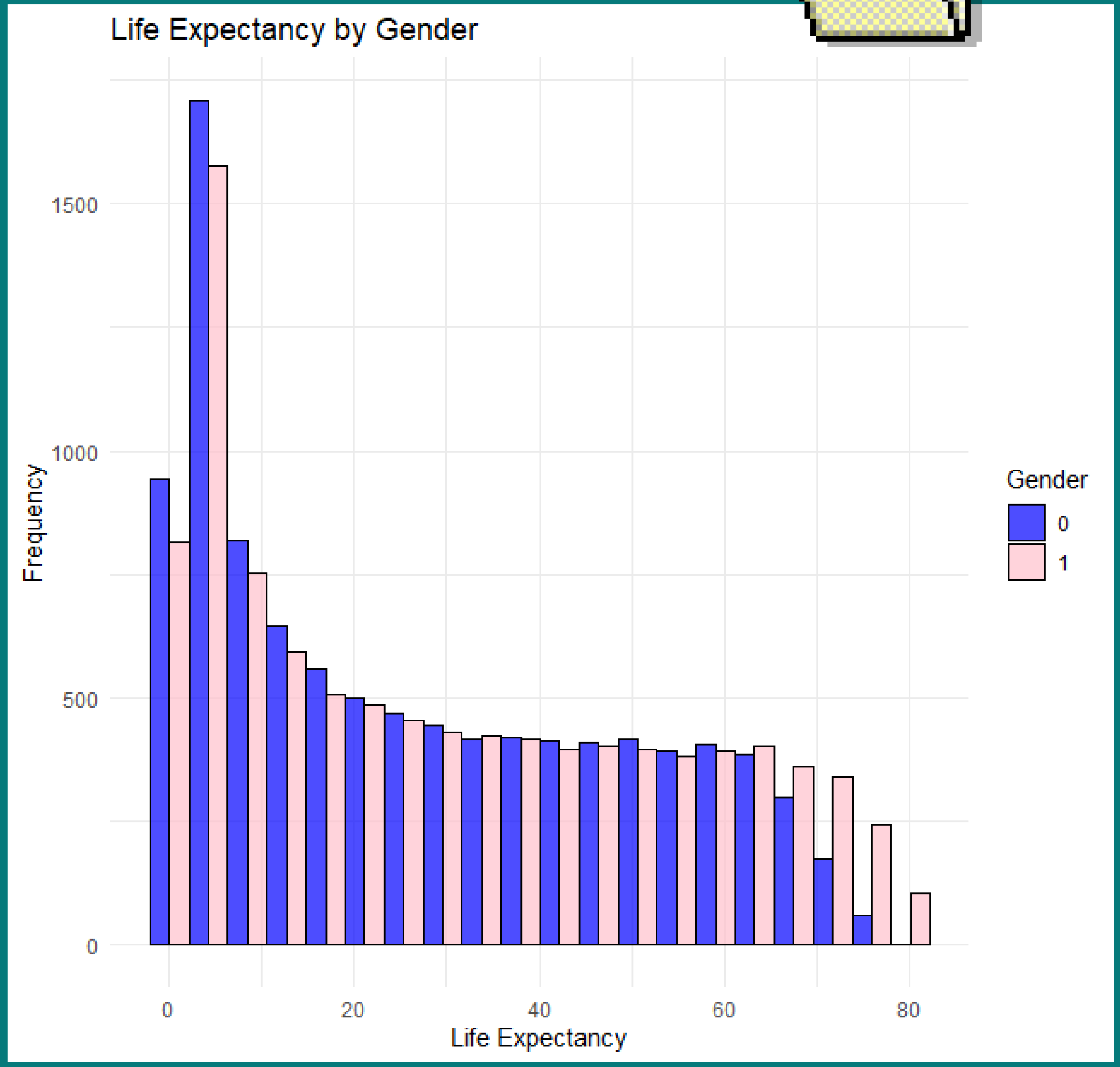
Exploratory Data Analysis


Scatterplot of Age vs Life Expectancy

Incase you did not know, as your Age increases…your life expectancy decreases.


Boxplot of Life Expectancy by Gender

0=male 1=female
Women live slightly longer than Men on average.


Life Expectancy by Gender

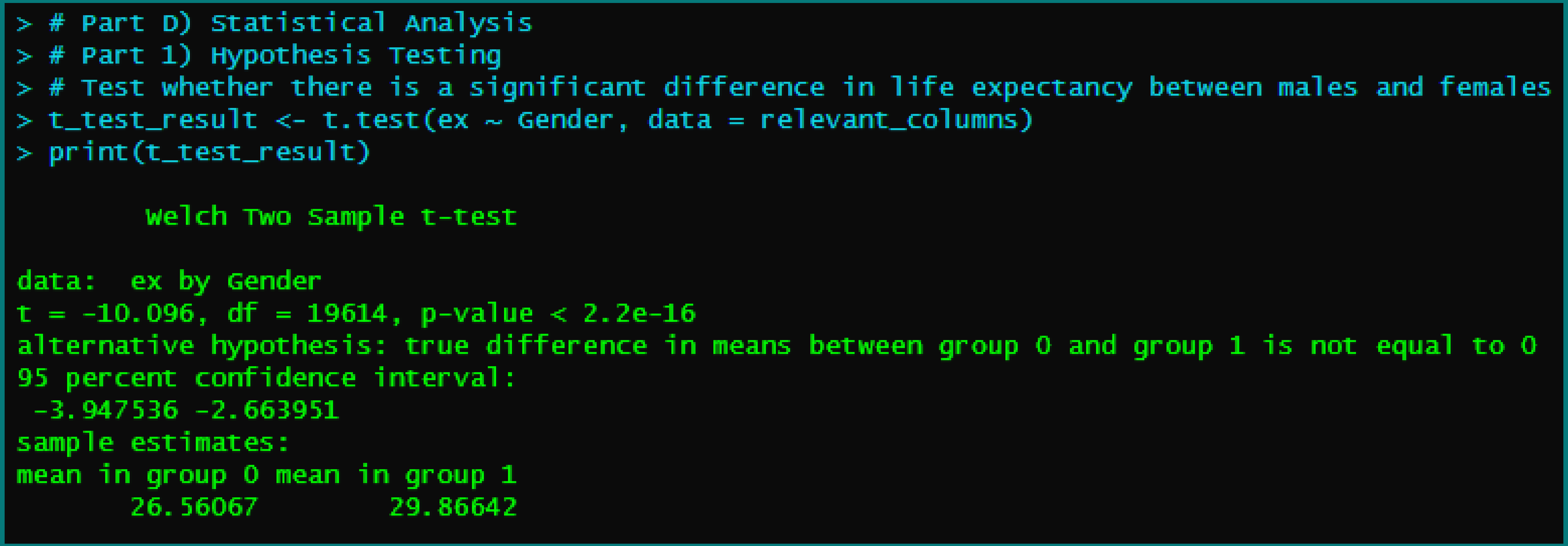
Here is a histogram of life expectancy frequencies since 1933 displaying a similar concept as the boxplots.

Back

Statistical Analysis

Back

Statistical Analysis



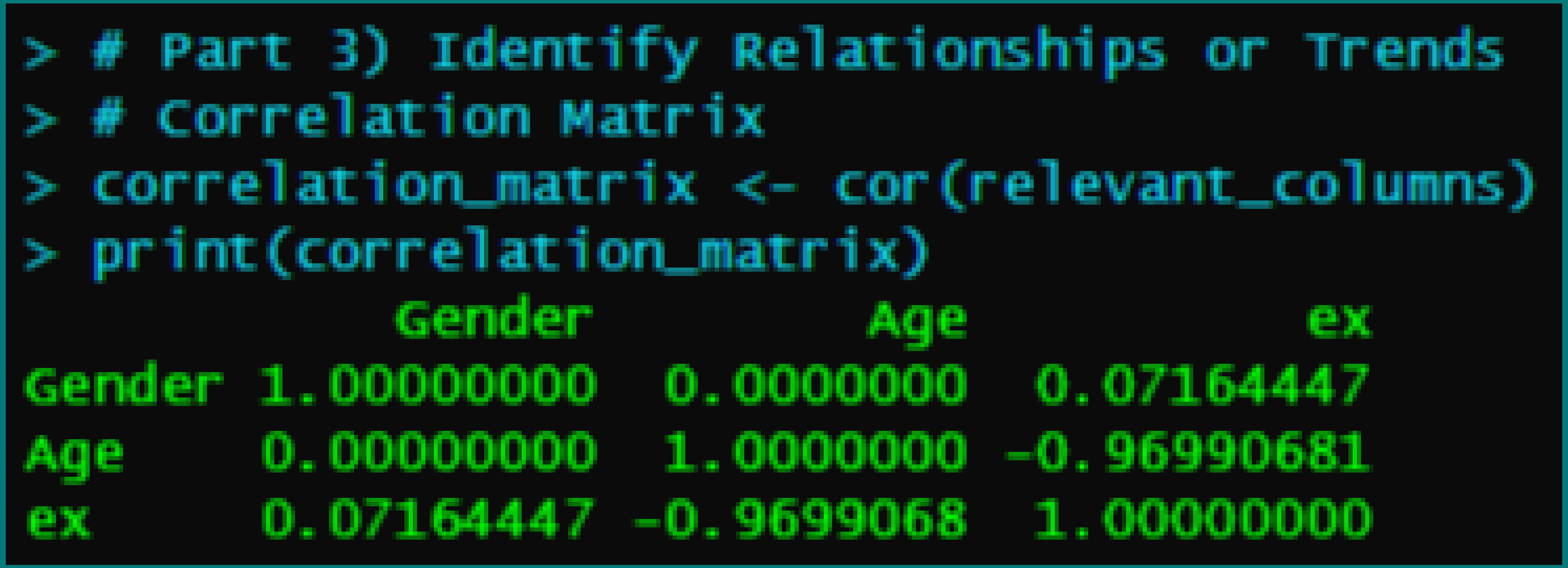
```
> # Part D) Statistical Analysis
> # Part 1) Hypothesis Testing
> # Test whether there is a significant difference in life expectancy between males and females
> t_test_result <- t.test(ex ~ Gender, data = relevant_columns)
> print(t_test_result)

	Welch Two Sample t-test

data:  ex by Gender
t = -10.096, df = 19614, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -3.947536 -2.663951
sample estimates:
mean in group 0 mean in group 1 
       26.56067        29.86642 
```

Hypothesis test

Preformed a hypothesis test to test if there was a significant difference in the means. Based on this test, there is a significant difference in life expectancy between males and females. The mean life expectancy for females is estimated to be higher than that for males, and this difference is statistically significant.

Back

Statistical Analysis



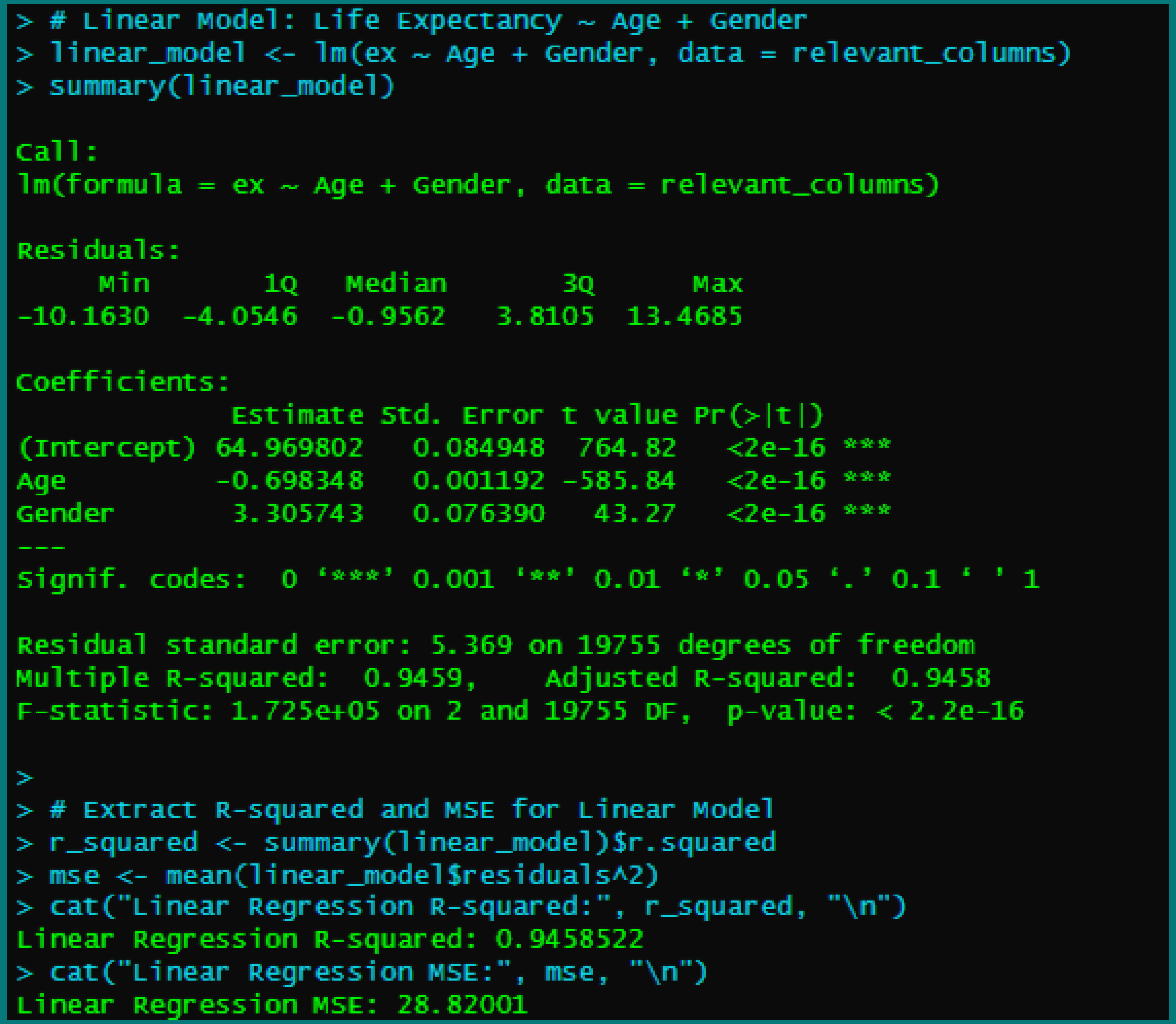
```
> # Part 3) Identify Relationships or Trends
> # Correlation Matrix
> correlation_matrix <- cor(relevant_columns)
> print(correlation_matrix)
           Gender        Age          ex
Gender 1.00000000  0.0000000  0.07164447
Age    0.00000000  1.0000000 -0.96990681
ex     0.07164447 -0.9699068  1.00000000
```

Correlation Analysis

The correlation matrix suggests a strong negative linear relationship between age and life expectancy, which is consistent with the general understanding that older individuals tend to have lower life expectancies. Then because Gender is a binary in the dataset, its correlations are meaningless.

Back

Statistical Analysis



```
> # Linear Model: Life Expectancy ~ Age + Gender
> linear_model <- lm(ex ~ Age + Gender, data = relevant_columns)
> summary(linear_model)

Call:
lm(formula = ex ~ Age + Gender, data = relevant_columns)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1630  -4.0546  -0.9562   3.8105  13.4685 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 64.969802   0.084948  764.82   <2e-16 ***
Age         -0.698348   0.001192 -585.84   <2e-16 ***
Gender       3.305743   0.076390   43.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.369 on 19755 degrees of freedom
Multiple R-squared:  0.9459,	Adjusted R-squared:  0.9458 
F-statistic: 1.725e+05 on 2 and 19755 DF,  p-value: < 2.2e-16

> 
> # Extract R-squared and MSE for Linear Model
> r_squared <- summary(linear_model)$r.squared
> mse <- mean(linear_model$residuals^2)
> cat("Linear Regression R-squared:", r_squared, "\n")
Linear Regression R-squared: 0.9458522 
> cat("Linear Regression MSE:", mse, "\n")
Linear Regression MSE: 28.82001 
```

Regression analysis

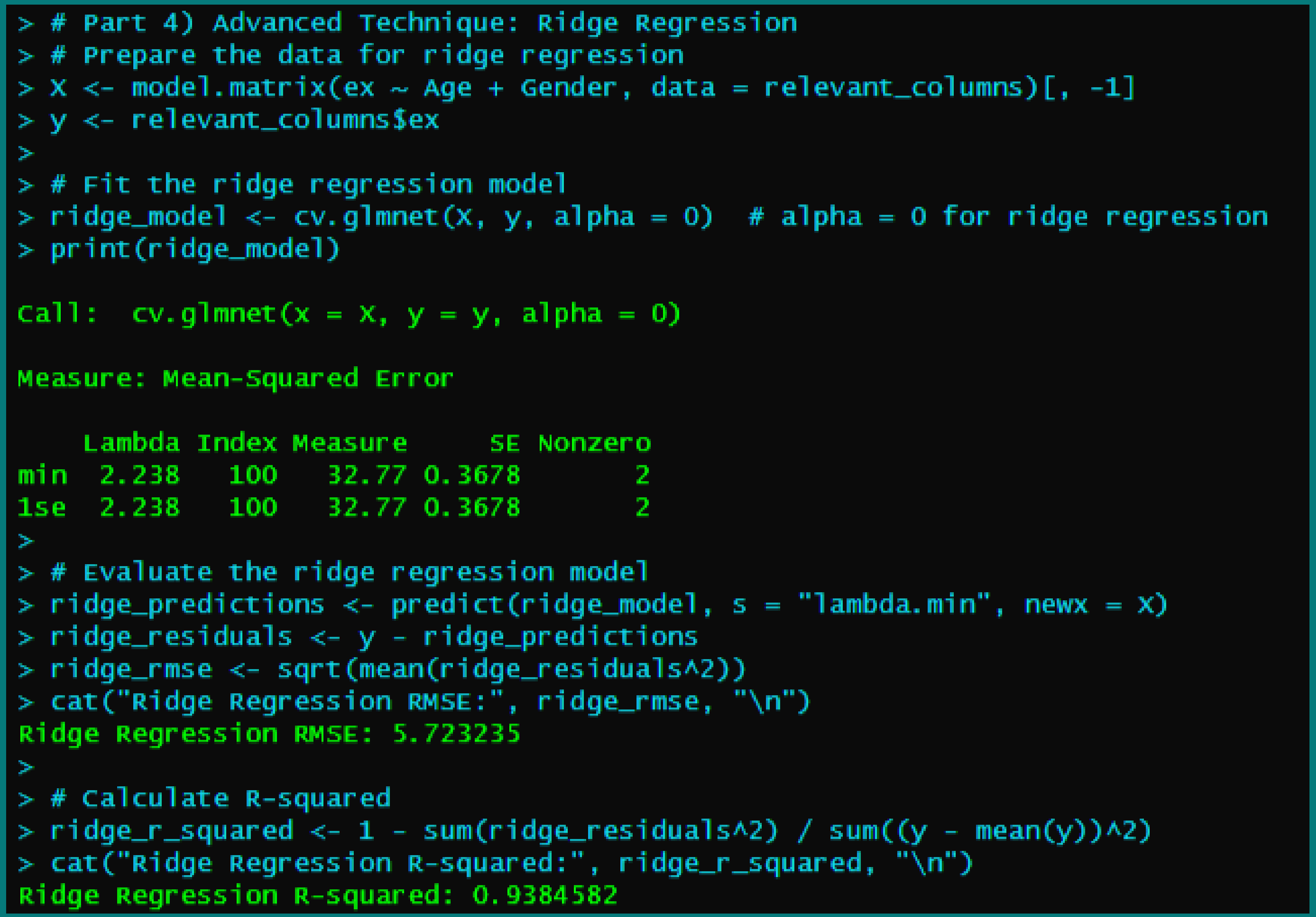
For each year increase in age, life expectancy is estimated to decrease by approximately 0.698 units, holding gender constant.

This also suggests that females, on average, have a higher life expectancy than males by approximately 3.31 years, holding age constant.

Approximately 94.59% of the variability in life expectancy is explained by age and gender.

Back

Statistical Analysis-advanced techniques



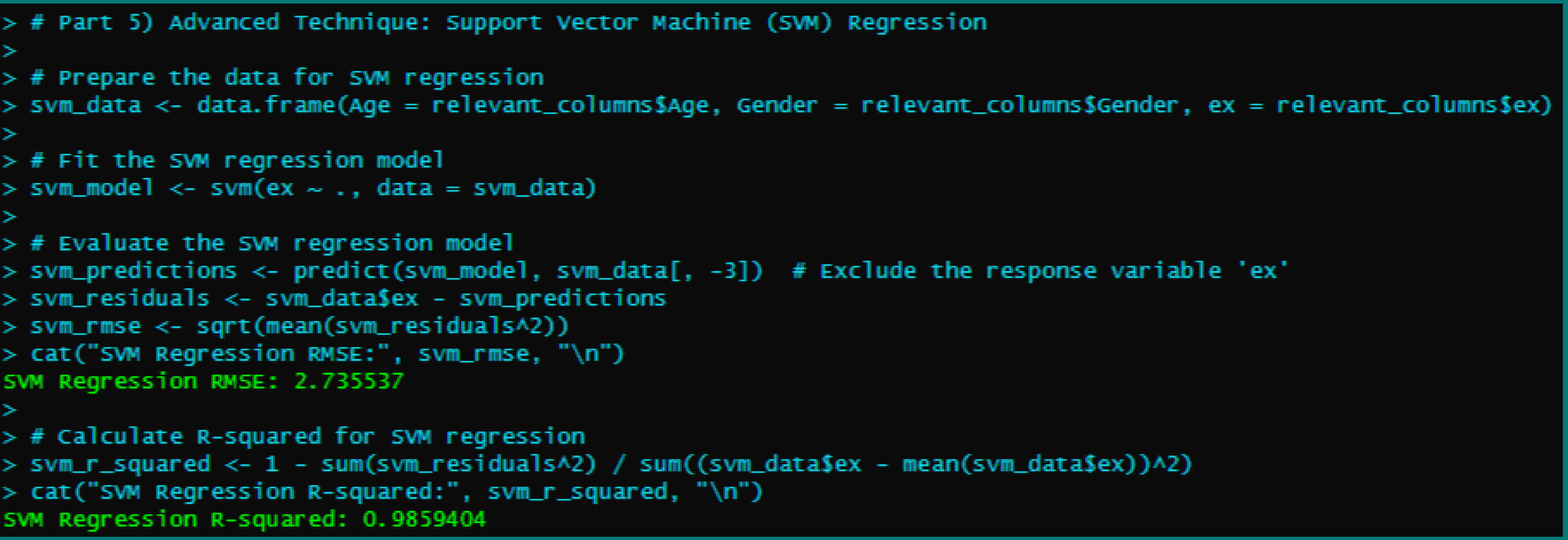
```
> # Part 4) Advanced Technique: Ridge Regression
> # Prepare the data for ridge regression
> X <- model.matrix(ex ~ Age + Gender, data = relevant_columns)[, -1]
> y <- relevant_columns$ex
>
> # Fit the ridge regression model
> ridge_model <- cv.glmnet(X, y, alpha = 0)  # alpha = 0 for ridge regression
> print(ridge_model)

Call:  cv.glmnet(x = X, y = y, alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure     SE Nonzero
min  2.238   100   32.77 0.3678       2
1se  2.238   100   32.77 0.3678       2
>
> # Evaluate the ridge regression model
> ridge_predictions <- predict(ridge_model, s = "lambda.min", newx = X)
> ridge_residuals <- y - ridge_predictions
> ridge_rmse <- sqrt(mean(ridge_residuals^2))
> cat("Ridge Regression RMSE:", ridge_rmse, "\n")
Ridge Regression RMSE: 5.723235
>
> # Calculate R-squared
> ridge_r_squared <- 1 - sum(ridge_residuals^2) / sum((y - mean(y))^2)
> cat("Ridge Regression R-squared:", ridge_r_squared, "\n")
Ridge Regression R-squared: 0.9384582
```

Ridge Regression analysis

- The ridge regression model with the chosen lambda performs well, as indicated by the low RMSE and high R-squared.
- The selected lambda suggests a balance between model complexity and goodness of fit. A larger lambda would result in a more parsimonious model, but it might sacrifice some goodness of fit.
- The non-zero coefficients indicate which variables are contributing to the prediction. Since you have two nonzero coefficients, the ridge regression model is effectively selecting the most important variables for prediction.
- The R-squared value suggests that the ridge regression model explains a substantial portion of the variability in life expectancy.

Back

Statistical Analysis-advanced techniques



```
> # Part 5) Advanced Technique: Support Vector Machine (SVM) Regression
>
> # Prepare the data for SVM regression
> svm_data <- data.frame(Age = relevant_columns$Age, Gender = relevant_columns$Gender, ex = relevant_columns$ex)
>
> # Fit the SVM regression model
> svm_model <- svm(ex ~ ., data = svm_data)
>
> # Evaluate the SVM regression model
> svm_predictions <- predict(svm_model, svm_data[, -3])  # Exclude the response variable 'ex'
> svm_residuals <- svm_data$ex - svm_predictions
> svm_rmse <- sqrt(mean(svm_residuals^2))
> cat("SVM Regression RMSE:", svm_rmse, "\n")
SVM Regression RMSE: 2.735537
>
> # Calculate R-squared for SVM regression
> svm_r_squared <- 1 - sum(svm_residuals^2) / sum((svm_data$ex - mean(svm_data$ex))^2)
> cat("SVM Regression R-squared:", svm_r_squared, "\n")
SVM Regression R-squared: 0.9859404
```

Support Vector Machine

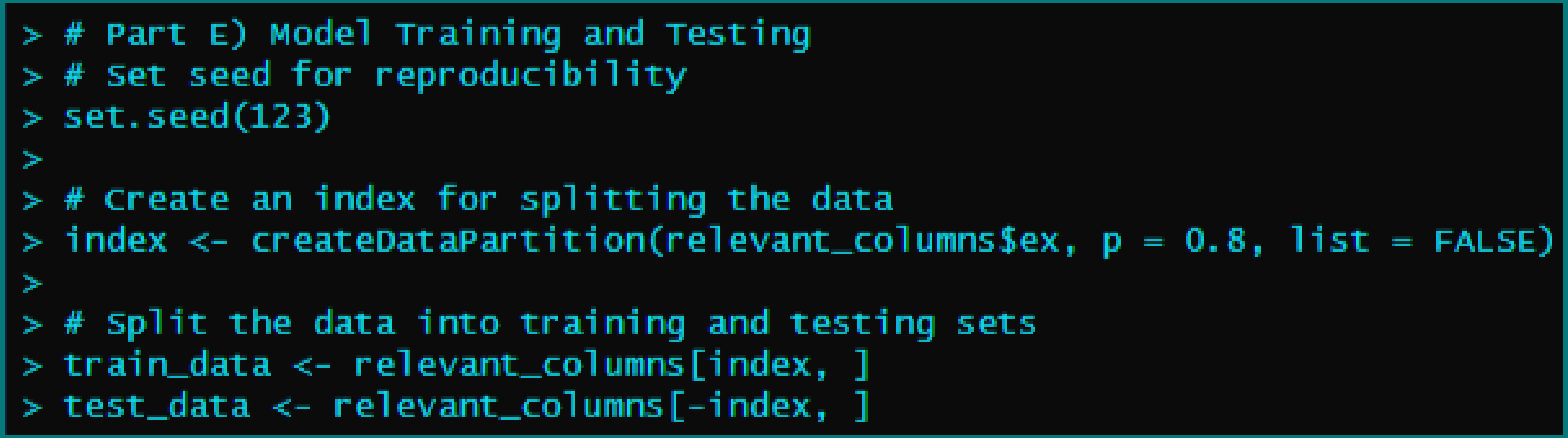
The SVM regression model, as evaluated based on the provided metrics, appears to perform well. The combination of a low RMSE and a high R-squared suggests that the model provides accurate predictions and explains a significant portion of the variability in the response variable. This makes the SVM regression model a promising and effective tool for predicting the 'ex' variable based on the given features.

Back

Predictive Modeling

Back

Hold-out validation

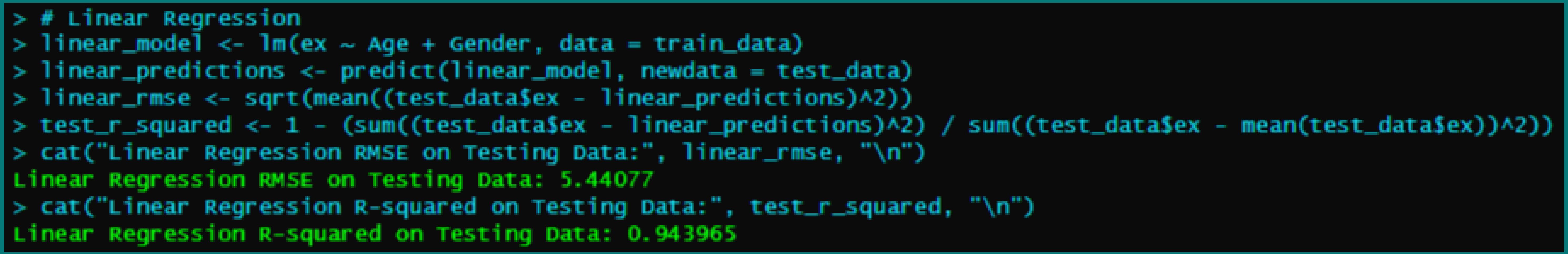


```
> # Part E) Model Training and Testing
> # Set seed for reproducibility
> set.seed(123)
>
> # Create an index for splitting the data
> index <- createDataPartition(relevant_columns$ex, p = 0.8, list = FALSE)
>
> # Split the data into training and testing sets
> train_data <- relevant_columns[index, ]
> test_data <- relevant_columns[-index, ]
```

80% of the dataset goes to the training set and 20% to the test set

Back

Modeling and Prediction



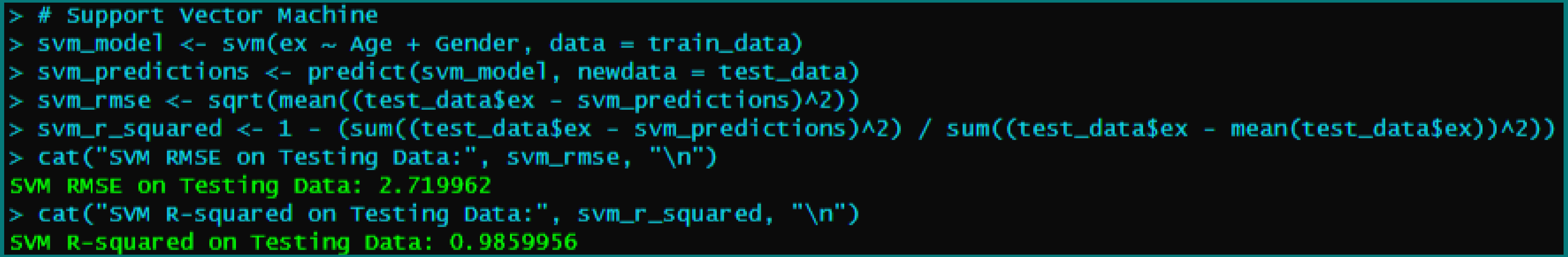
```
> # Linear Regression
> linear_model <- lm(ex ~ Age + Gender, data = train_data)
> linear_predictions <- predict(linear_model, newdata = test_data)
> linear_rmse <- sqrt(mean((test_data$ex - linear_predictions)^2))
> test_r_squared <- 1 - (sum((test_data$ex - linear_predictions)^2) / sum((test_data$ex - mean(test_data$ex))^2))
> cat("Linear Regression RMSE on Testing Data:", linear_rmse, "\n")
Linear Regression RMSE on Testing Data: 5.44077
> cat("Linear Regression R-squared on Testing Data:", test_r_squared, "\n")
Linear Regression R-squared on Testing Data: 0.943965
```

Linear Regression

The Linear Regression model, as evaluated on the testing data, demonstrates good predictive performance for life expectancy. The moderate RMSE and high R-squared indicate that the model generalizes well to unseen data, providing accurate and reliable predictions. This makes the Linear Regression model a suitable choice for predicting life expectancy based on the given features

Back

Modeling and Prediction



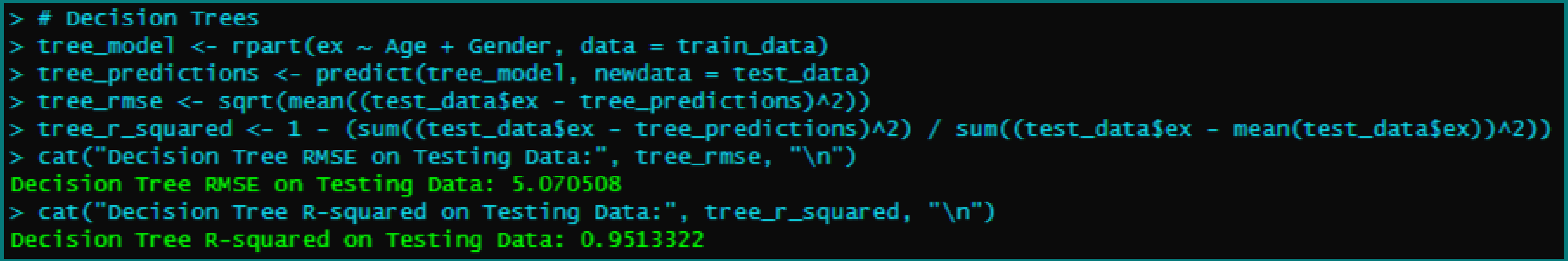
```
> # Support Vector Machine
> svm_model <- svm(ex ~ Age + Gender, data = train_data)
> svm_predictions <- predict(svm_model, newdata = test_data)
> svm_rmse <- sqrt(mean((test_data$ex - svm_predictions)^2))
> svm_r_squared <- 1 - (sum((test_data$ex - svm_predictions)^2) / sum((test_data$ex - mean(test_data$ex))^2))
> cat("SVM RMSE on Testing Data:", svm_rmse, "\n")
SVM RMSE on Testing Data: 2.719962
> cat("SVM R-squared on Testing Data:", svm_r_squared, "\n")
SVM R-squared on Testing Data: 0.9859956
```

Support Vector Machine

The Support Vector Machine model, as evaluated on the testing data, demonstrates strong predictive performance for life expectancy. The low RMSE and high R-squared indicate that the model generalizes well to unseen data, providing accurate and reliable predictions. This makes the SVM model a suitable choice for predicting life expectancy

Back

Modeling and Prediction



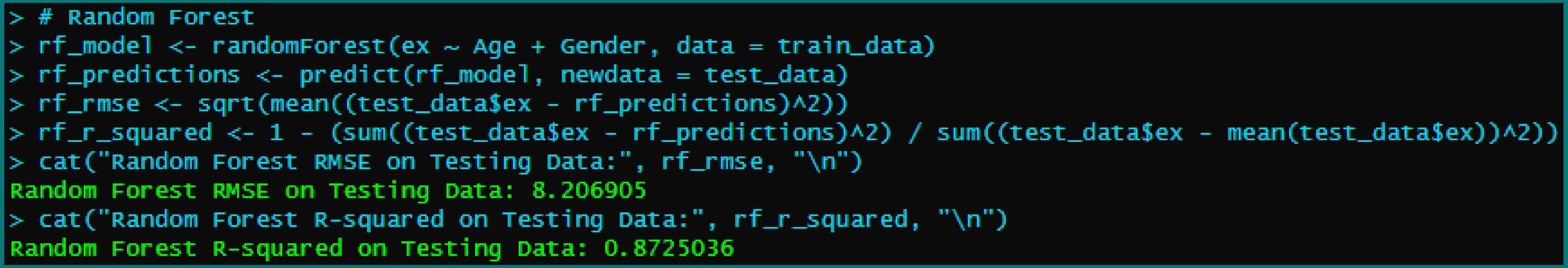
```
> # Decision Trees
> tree_model <- rpart(ex ~ Age + Gender, data = train_data)
> tree_predictions <- predict(tree_model, newdata = test_data)
> tree_rmse <- sqrt(mean((test_data$ex - tree_predictions)^2))
> tree_r_squared <- 1 - (sum((test_data$ex - tree_predictions)^2) / sum((test_data$ex - mean(test_data$ex))^2))
> cat("Decision Tree RMSE on Testing Data:", tree_rmse, "\n")
Decision Tree RMSE on Testing Data: 5.070508
> cat("Decision Tree R-squared on Testing Data:", tree_r_squared, "\n")
Decision Tree R-squared on Testing Data: 0.9513322
```

Decision Tree

The Decision Tree model, as evaluated on the testing data, demonstrates good predictive performance for life expectancy. The moderate RMSE and high R-squared indicate that the model generalizes well to unseen data, providing accurate and reliable predictions. Decision Trees are known for their ability to capture complex relationships in the data, and this result suggests that the model effectively leverages these capabilities.

Back

Modeling and Prediction



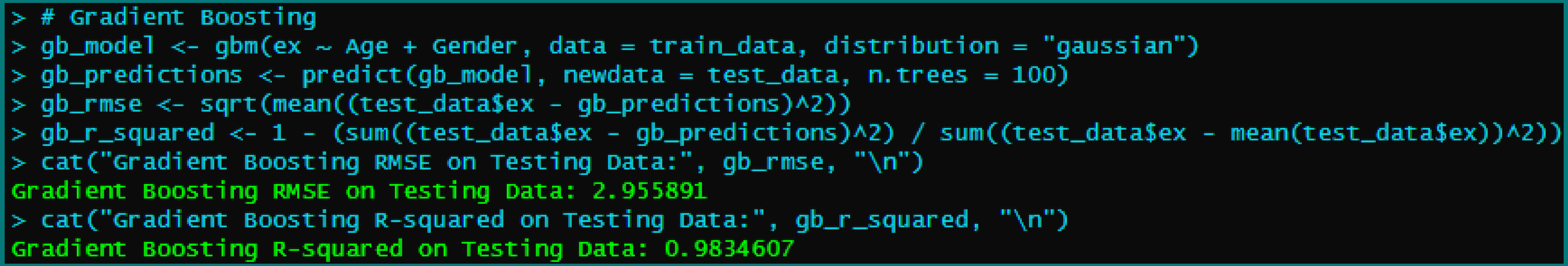
```
> # Random Forest
> rf_model <- randomForest(ex ~ Age + Gender, data = train_data)
> rf_predictions <- predict(rf_model, newdata = test_data)
> rf_rmse <- sqrt(mean((test_data$ex - rf_predictions)^2))
> rf_r_squared <- 1 - (sum((test_data$ex - rf_predictions)^2) / sum((test_data$ex - mean(test_data$ex))^2))
> cat("Random Forest RMSE on Testing Data:", rf_rmse, "\n")
Random Forest RMSE on Testing Data: 8.206905
> cat("Random Forest R-squared on Testing Data:", rf_r_squared, "\n")
Random Forest R-squared on Testing Data: 0.8725036
```

Random Forest

The Random Forest model, as evaluated on the testing data, demonstrates good predictive performance for life expectancy. While the RMSE is relatively higher, the R-squared value indicates that the model still explains a substantial portion of the variability in the response variable. Random Forest models are known for their robustness and ability to handle complex relationships in the data. This result suggests that the Random Forest leverages these characteristics to provide accurate predictions on the testing data.

Back

Modeling and Prediction



```
> # Gradient Boosting
> gb_model <- gbm(ex ~ Age + Gender, data = train_data, distribution = "gaussian")
> gb_predictions <- predict(gb_model, newdata = test_data, n.trees = 100)
> gb_rmse <- sqrt(mean((test_data$ex - gb_predictions)^2))
> gb_r_squared <- 1 - (sum((test_data$ex - gb_predictions)^2) / sum((test_data$ex - mean(test_data$ex))^2))
> cat("Gradient Boosting RMSE on Testing Data:", gb_rmse, "\n")
Gradient Boosting RMSE on Testing Data: 2.955891
> cat("Gradient Boosting R-squared on Testing Data:", gb_r_squared, "\n")
Gradient Boosting R-squared on Testing Data: 0.9834607
```

Gradient Boosting

The Gradient Boosting model, as evaluated on the testing data, demonstrates excellent predictive performance for life expectancy. The low RMSE and high R-squared value indicate that the model provides accurate predictions and effectively captures the patterns in the data. Gradient Boosting is known for its ability to handle complex relationships and nonlinearities, and this result suggests that the model leverages these capabilities to provide highly accurate predictions.

Back

Conclusion

Back

Conclusion

Most all models preformed well, with only Random Forests preforming slightly worse than the others.

1. Limitations:
   - The analysis is based on a specific set of features (age and gender), and there may be other factors influencing life expectancy that are not considered.
   - The models assume a linear relationship between features and life expectancy, which may not capture complex non-linear patterns.
2. Improvements and Future Research:
   - Feature Expansion: Explore additional features that could impact life expectancy, such as socioeconomic factors, healthcare access, and lifestyle variables.
   - Model Complexity: Experiment with more complex models or ensemble methods to capture intricate relationships within the data.
   - Data Quality: Ensure data quality by addressing any missing values, outliers, or inconsistencies that may impact model performance.
3. Ethical Considerations:
   - Consider ethical implications related to the use of demographic data, ensuring fairness and avoiding biases in predictive models.
4. External Validation:
   - Validate the models on external datasets to assess generalizability and robustness.

Back

Thank you!

Insert a parting or call-to-action message here.

Resource Page

HMD. Human Mortality Database. Human Mortality Database: USA lifetables, males and females. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France). Available at www.mortality.org (data downloaded on 11/1/2023).