KID: [K2283306@kingston.ac.uk](mailto:K2283306@kingston.ac.uk)

CI7330 – Data Analytics and Visualisation

Assignment 2: 19th January 2023

**Task 1: Descriptive Statistics**

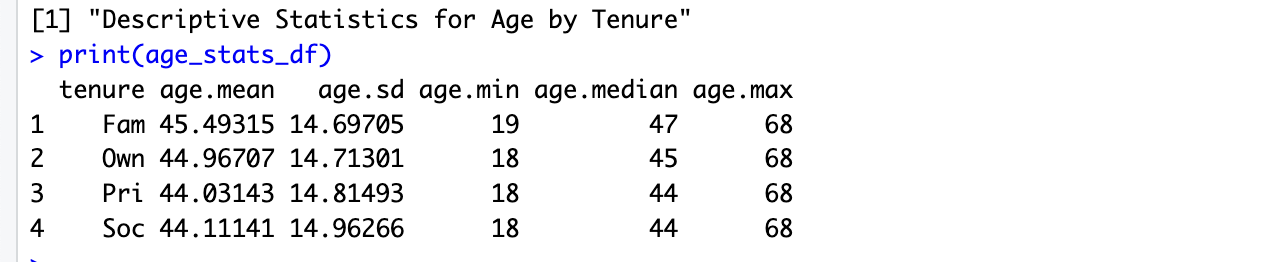


Table 1: Descriptive Statistics for age

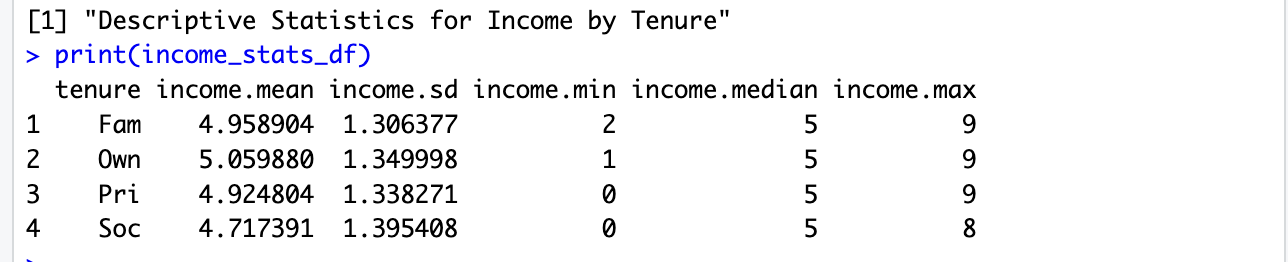


Table 2: Descriptive Statistics for income

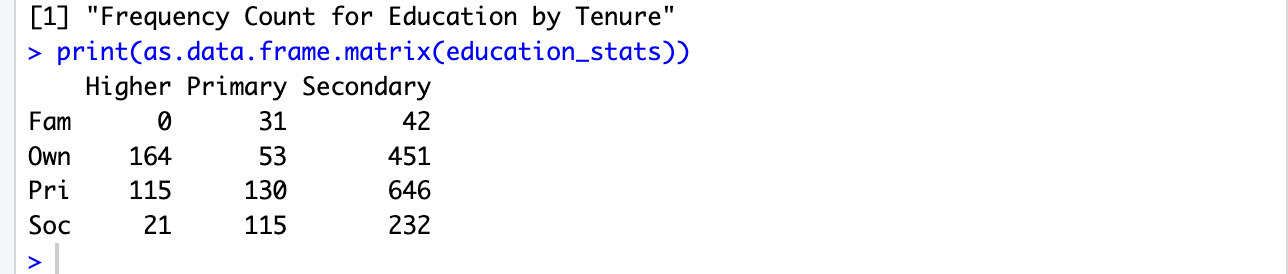
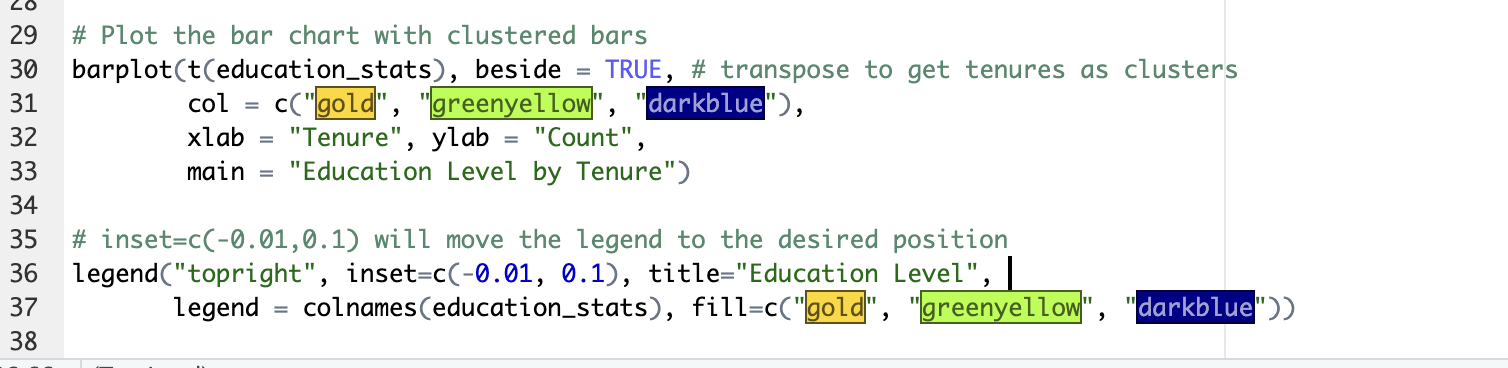


Table 3: Descriptive Statistics for education

The data reveals a uniform age profile across various housing tenures, predominantly centred around the mid-forties. Homeowners stand out with marginally higher incomes and a significant representation in higher education, hinting at a link between property ownership and socio-economic advantages. Conversely, the broad income spectrum within private tenants suggests diverse financial backgrounds. The absence of higher education attainment among family tenants underscores educational inequities linked to housing status. Overall, the analysis indicates a complex interplay between housing tenure, economic conditions, and educational levels.

**Task 2: Clustered Bar Plot**



Code Snippet 1: Clustered Bar Plot

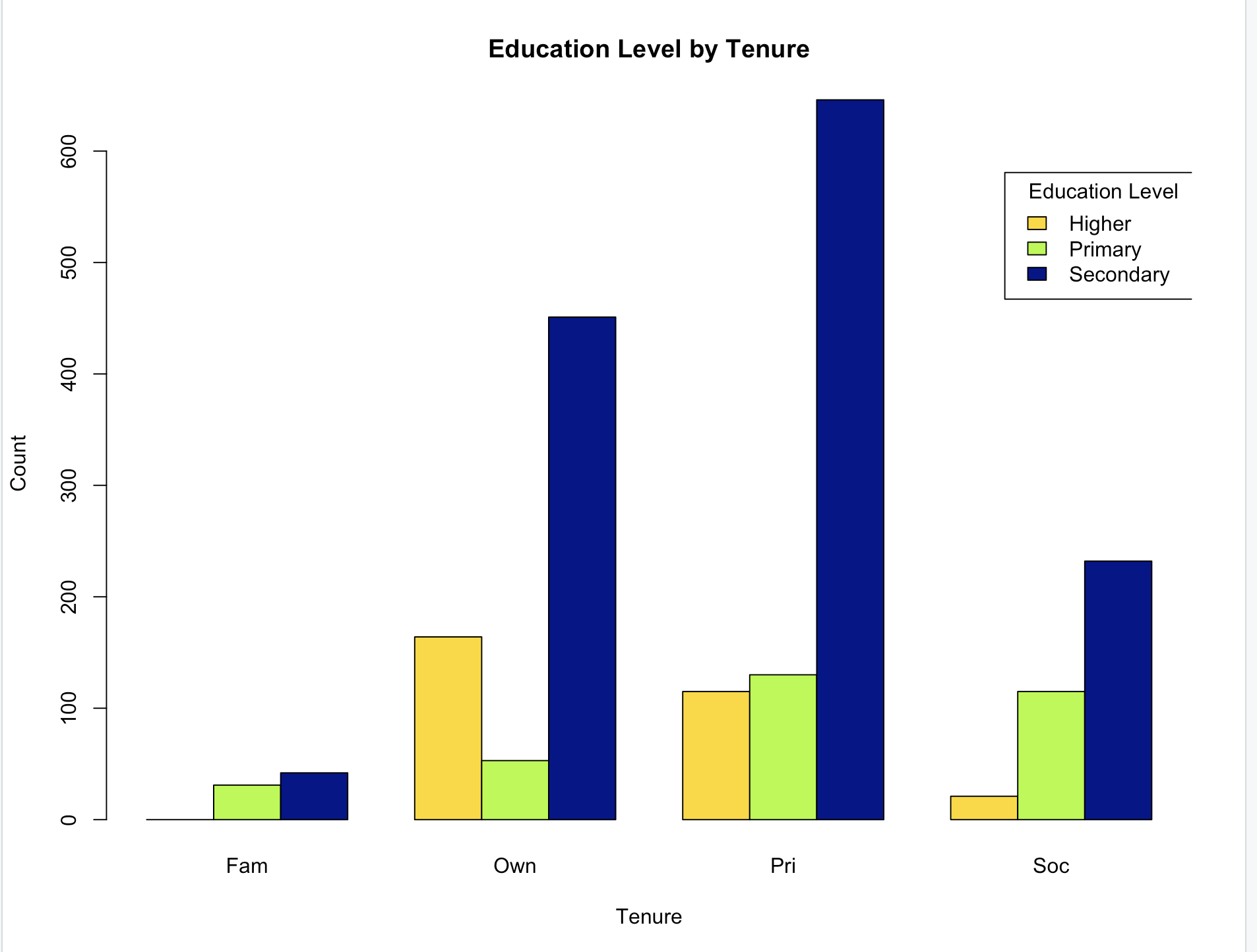
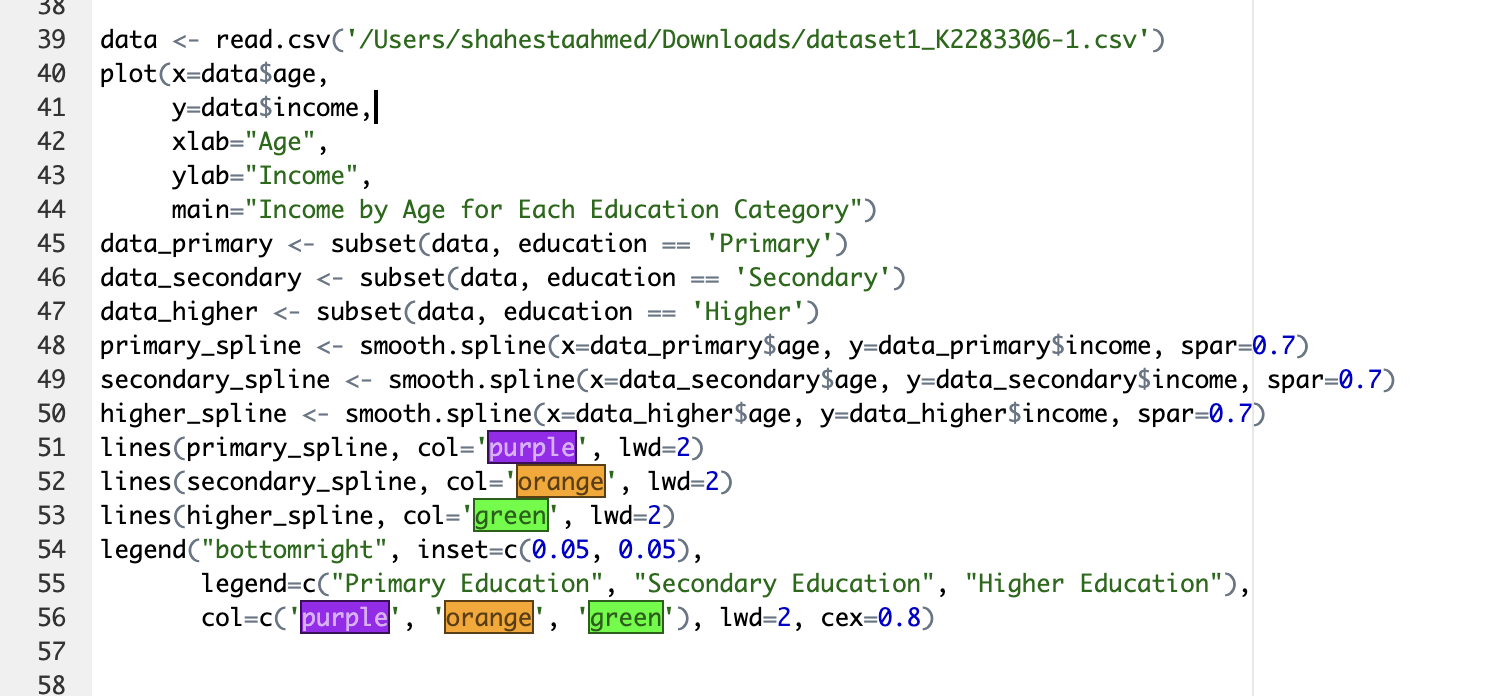


Figure 1: Clustered bar plot for tenure across education levels

The chart displays the relationship between housing tenure and education levels, revealing that individuals with secondary education are the most prevalent across all housing types. Homeowners and private renters show a higher proportion of secondary education, while those with primary education are least represented. This suggests targeted policy engagement may be necessary for those with lower education levels in different housing situations.

**Task 3: Spline Lines for level of education**



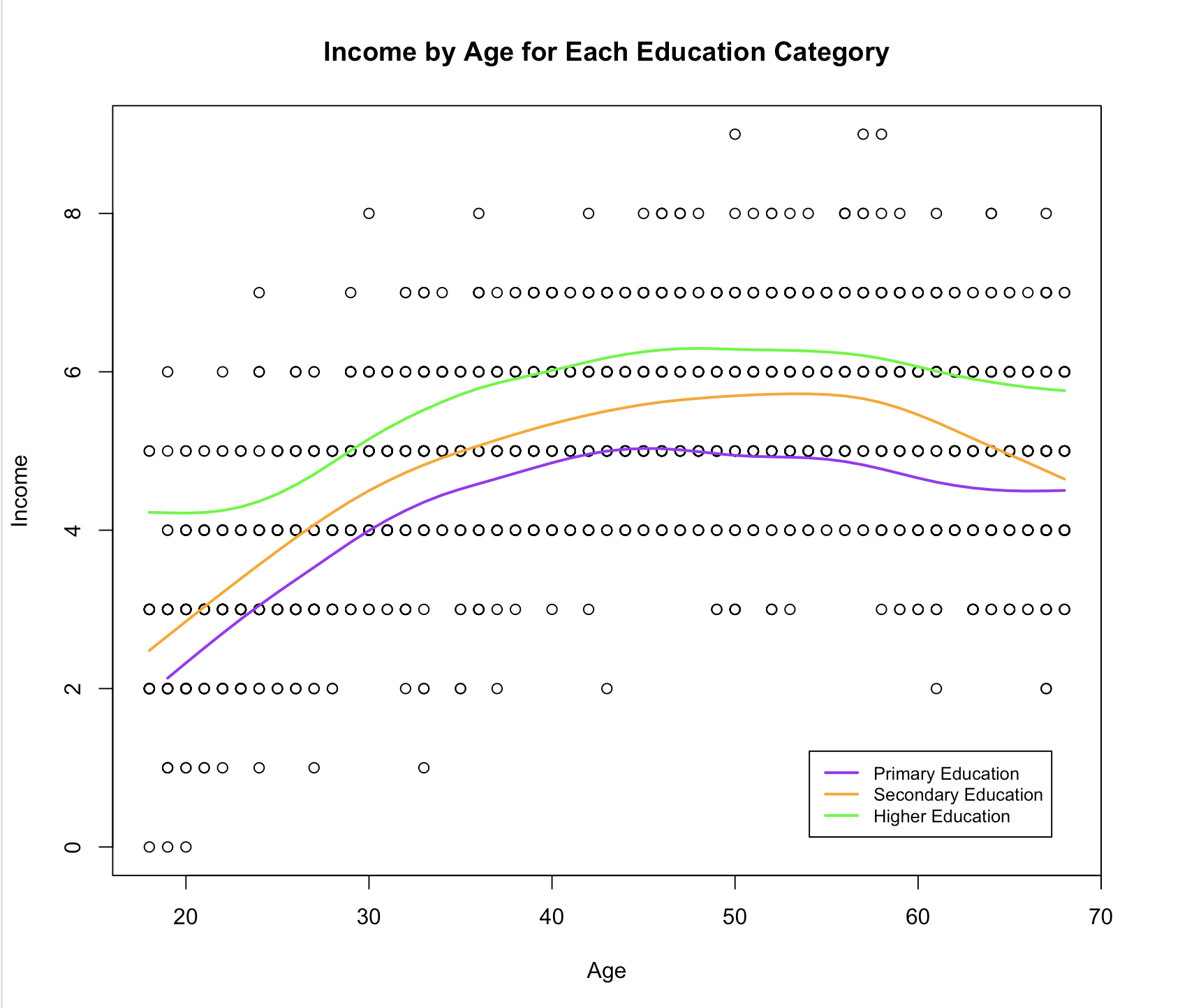
Code Snippet 2: Spline Graphs  
  


Figure 2: Spline lines for education levels with income and age as scatter points

The scatterplot shows that as age increases, income tends to stabilize regardless of education level. Individuals with higher education generally have higher incomes across all ages, followed by those with secondary education, while those with primary education typically have the lowest incomes. The convergence of income levels in later years suggests that the advantage of higher education on income may diminish with age.

**Task 4: Marginal effect (Predicted Probability)**

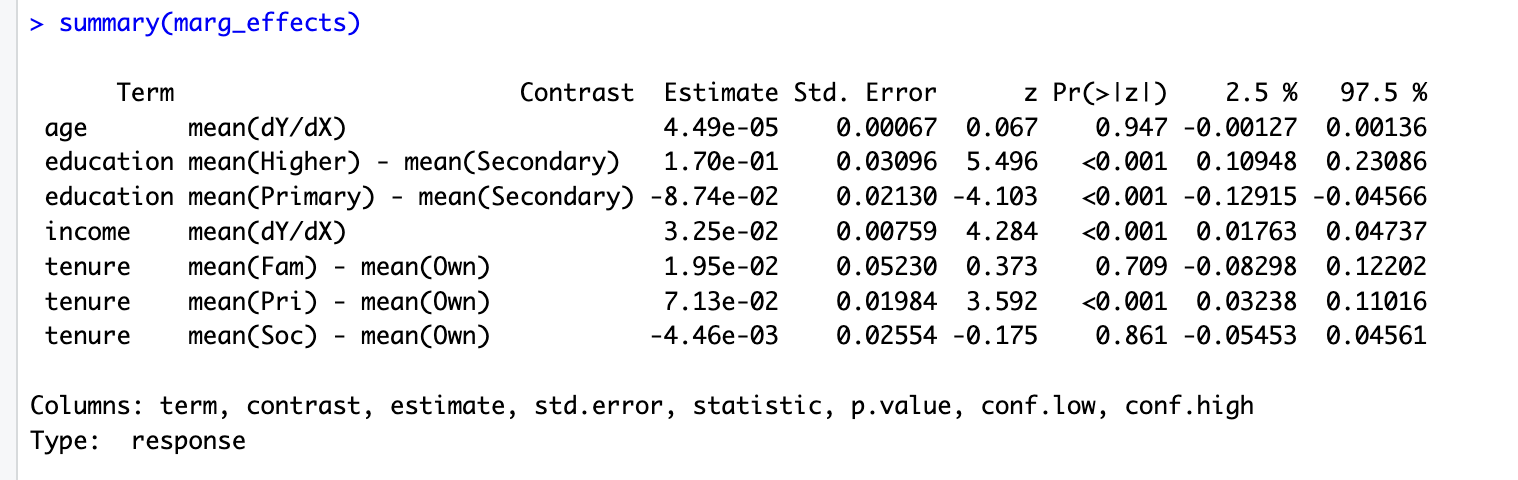


Table 4: Marginal effects summary

The study found that private renters and those with higher education are more likely to support the party, while age has little effect on voting preferences. Higher income also correlates with increased support. Targeting educated, higher-earning renters could expand the party's base.

**Task 5: Formula predicting the log-odds of intention to vote**

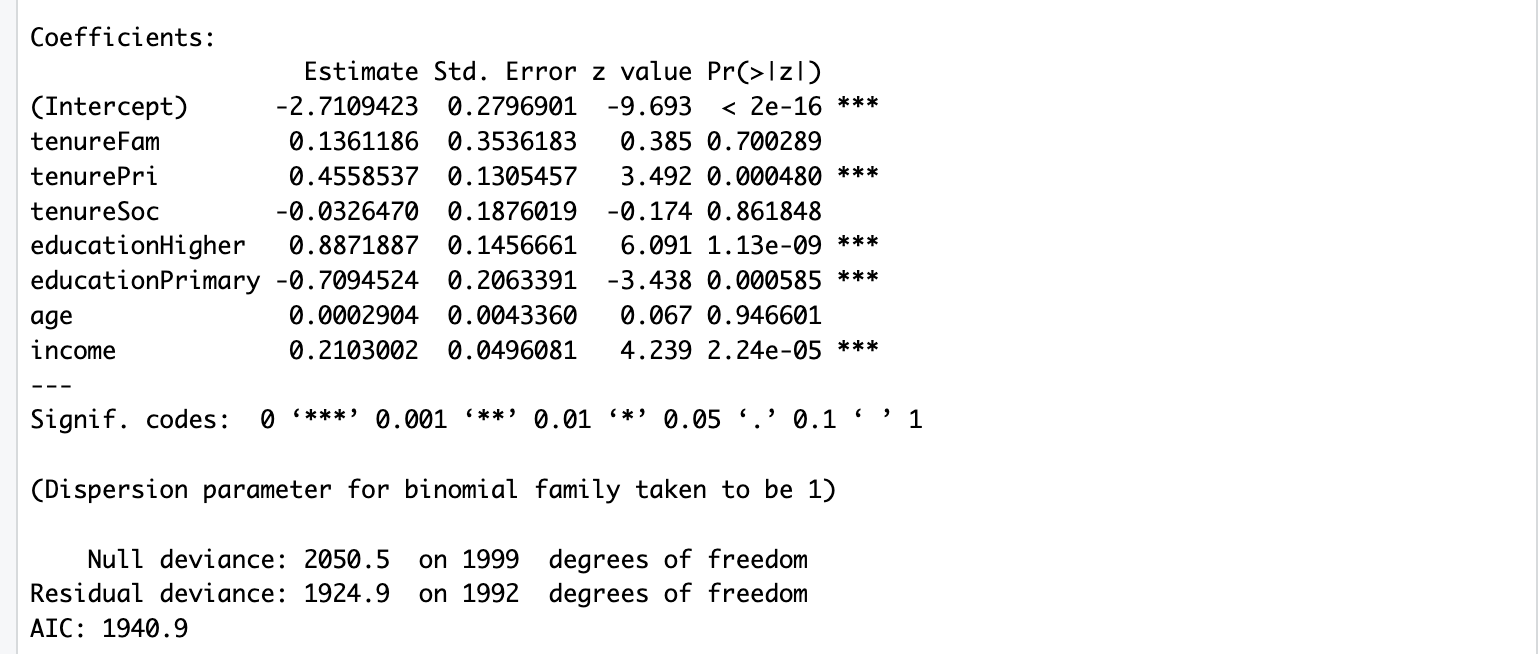


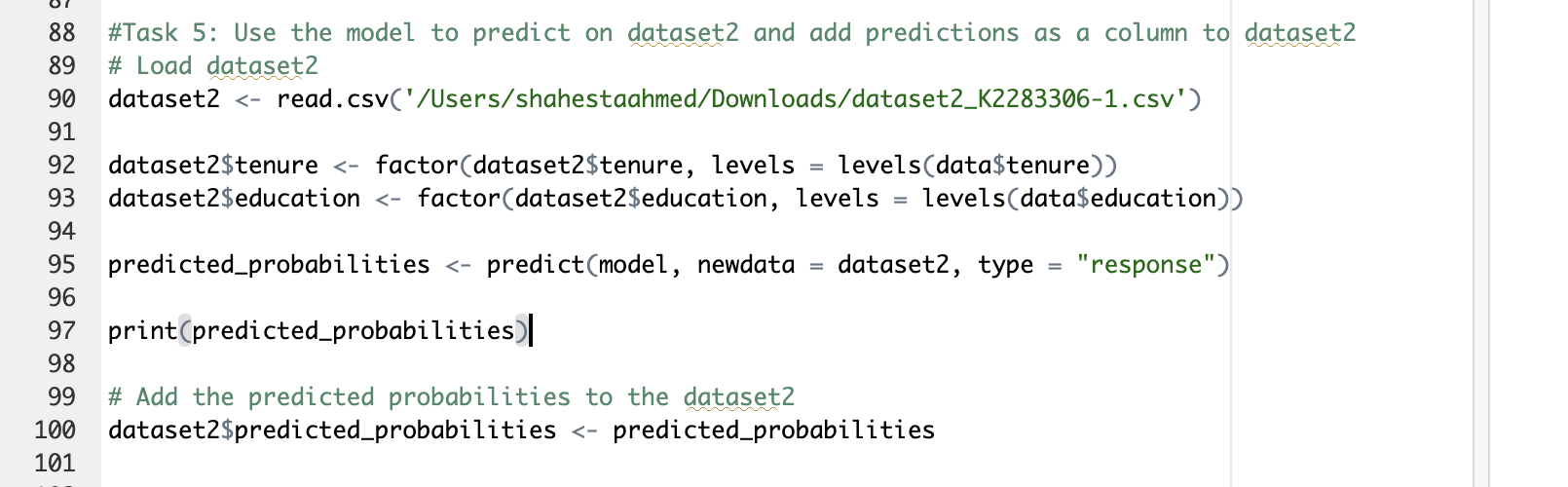
Table 5: Model coefficients

The coefficients are derived from above.

Formula:

*log-odds(Vote=1) = -2.7109 + 0.1361 × TenureFam + 0.4559 × TenurePri - 0.0326 × TenureSoc + 0.8872 × EducationHigher - 0.7095 × EducationPrimary + 0.0003 × Age + 0.2103 × Income*

Using the predict() function to get predicted probabilities for Happytown



Code Snippet 3: Model for prediction

**R Code for Analysis:**

#Task 1: Descriptive Statistics

# Load the data

data <- read.csv('/Users/shahestaahmed/Downloads/dataset1\_K2283306-1.csv')

# Descriptive statistics for 'age' grouped by 'tenure'

age\_stats <- aggregate(age ~ tenure, data, function(x) c(mean = mean(x), sd = sd(x), min = min(x), median = median(x), max = max(x)))

age\_stats\_df <- do.call(data.frame, age\_stats)

print("Descriptive Statistics for Age by Tenure")

print(age\_stats\_df)

# Descriptive statistics for 'income' grouped by 'tenure'

income\_stats <- aggregate(income ~ tenure, data, function(x) c(mean = mean(x), sd = sd(x), min = min(x), median = median(x), max = max(x)))

income\_stats\_df <- do.call(data.frame, income\_stats)

print("Descriptive Statistics for Income by Tenure")

print(income\_stats\_df)

# Frequency count for 'education' grouped by 'tenure'

education\_stats <- table(data$tenure, data$education)

print("Frequency Count for Education by Tenure")

print(as.data.frame.matrix(education\_stats))

#Task 2: Clustered Bar Plot

data$tenure <- factor(data$tenure)

data$education <- factor(data$education, levels = c("Higher", "Primary", "Secondary"))

# Create the cross-tabulation table with tenure as rows and education as columns

education\_stats <- table(data$tenure, data$education)

# Plot the bar chart with clustered bars

barplot(t(education\_stats), beside = TRUE, # transpose to get tenures as clusters

col = c("gold", "greenyellow", "darkblue"),

xlab = "Tenure", ylab = "Count",

main = "Education Level by Tenure")

# inset=c(-0.01,0.1) will move the legend to the desired position

legend("topright", inset=c(-0.01, 0.1), title="Education Level",

legend = colnames(education\_stats), fill=c("gold", "greenyellow", "darkblue"))

#Task 3: Creating spline charts for education levels

data <- read.csv('/Users/shahestaahmed/Downloads/dataset1\_K2283306-1.csv')

plot(x=data$age,

y=data$income,

xlab="Age",

ylab="Income",

main="Income by Age for Each Education Category")

data\_primary <- subset(data, education == 'Primary')

data\_secondary <- subset(data, education == 'Secondary')

data\_higher <- subset(data, education == 'Higher')

primary\_spline <- smooth.spline(x=data\_primary$age, y=data\_primary$income, spar=0.7)

secondary\_spline <- smooth.spline(x=data\_secondary$age, y=data\_secondary$income, spar=0.7)

higher\_spline <- smooth.spline(x=data\_higher$age, y=data\_higher$income, spar=0.7)

lines(primary\_spline, col='purple', lwd=2)

lines(secondary\_spline, col='orange', lwd=2)

lines(higher\_spline, col='green', lwd=2)

legend("bottomright", inset=c(0.05, 0.05),

legend=c("Primary Education", "Secondary Education", "Higher Education"),

col=c('purple', 'orange', 'green'), lwd=2, cex=0.8)

#Task 4: Calculating coefficients and confidence intervals

# Load marginaleffects library

install.packages("marginaleffects")

library(marginaleffects)

data <- read.csv('/Users/shahestaahmed/Downloads/dataset1\_K2283306-1.csv')

data$tenure <- relevel(factor(data$tenure), ref = "Own")

data$education <- relevel(factor(data$education), ref = "Secondary")

# Fit the logistic regression model

model <- glm(vote ~ tenure + education + age + income, data = data, family = binomial())

# Display the coefficients and confidence intervals

summary(model)

confint(model)

# Calculate and display the marginal effects

marg\_effects <- marginaleffects(model)

summary(marg\_effects)

confint(marg\_effects)

#Task 5: Use the model to predict on dataset2 and add predictions as a column to dataset2

# Load dataset2

dataset2 <- read.csv('/Users/shahestaahmed/Downloads/dataset2\_K2283306-1.csv')

dataset2$tenure <- factor(dataset2$tenure, levels = levels(data$tenure))

dataset2$education <- factor(dataset2$education, levels = levels(data$education))

predicted\_probabilities <- predict(model, newdata = dataset2, type = "response")

print(predicted\_probabilities)

# Add the predicted probabilities to the dataset2

dataset2$predicted\_probabilities <- predicted\_probabilities