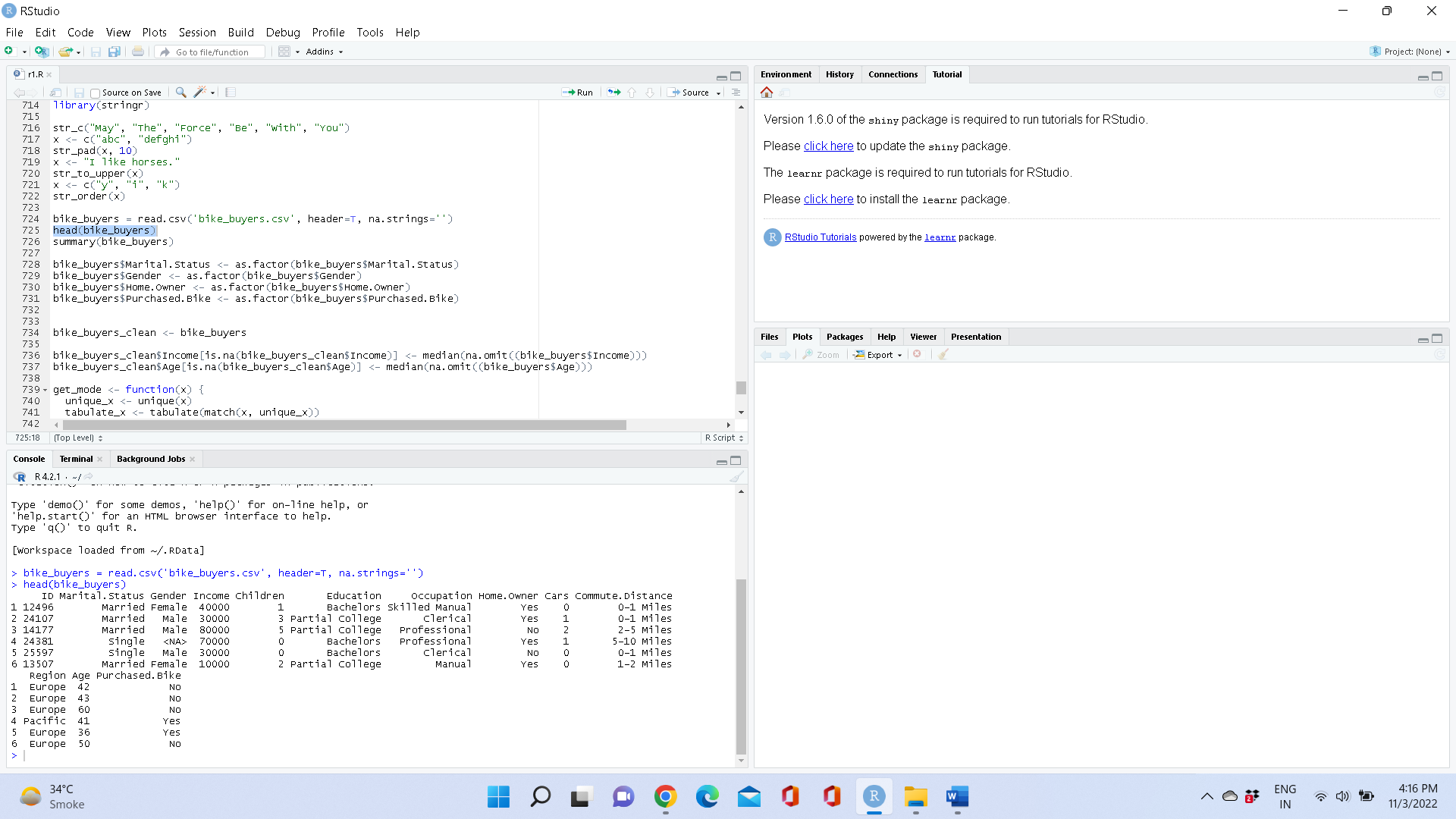
I have selected bike\_buyers dataset for this project.

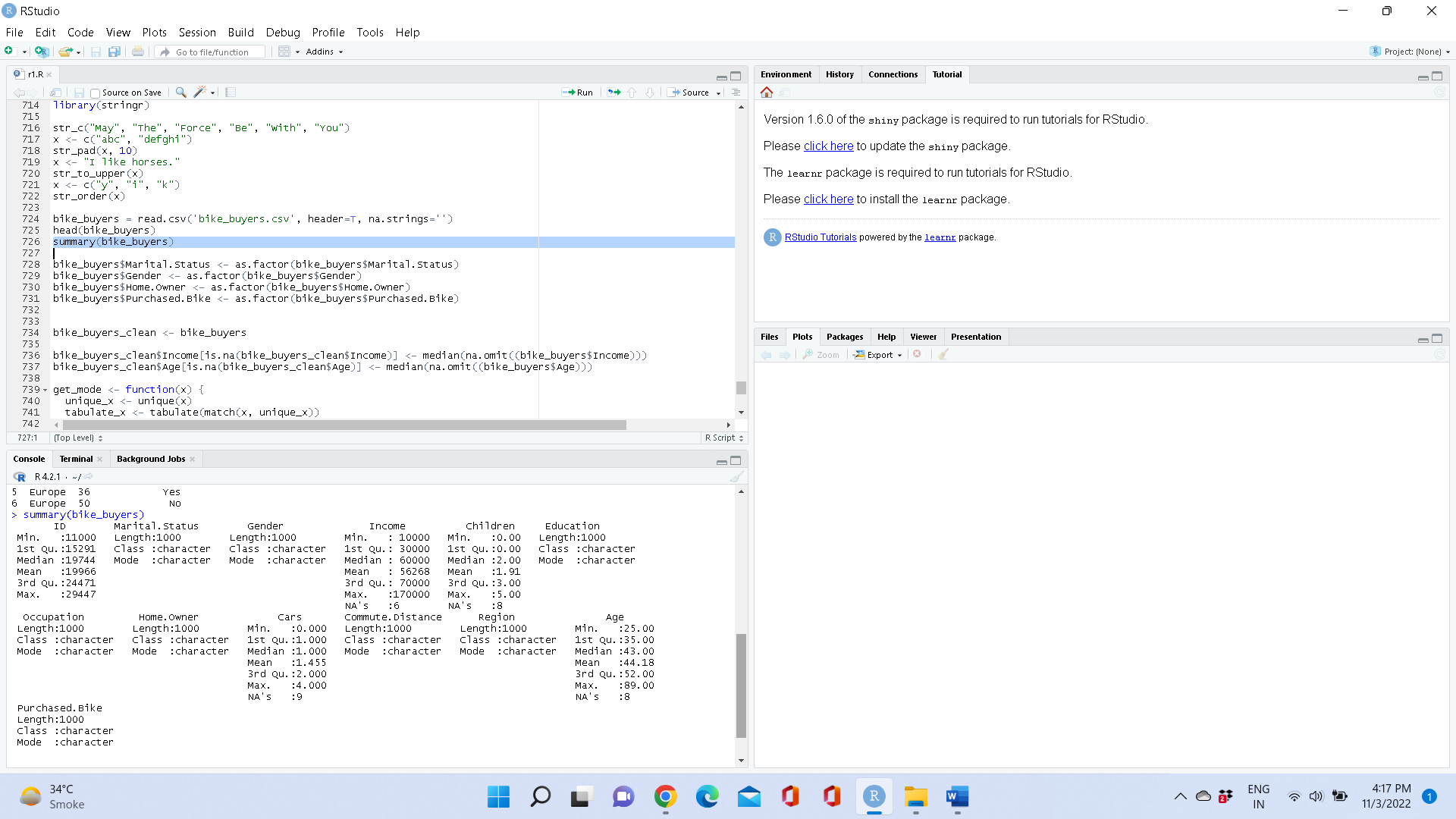
Reading and high-level analysis:

*bike\_buyers = read.csv('bike\_buyers.csv', header=T, na.strings='')*

*head(bike\_buyers)*



*summary(bike\_buyers)*



Assign Factors to String Values:

*bike\_buyers$Marital.Status <- as.factor(bike\_buyers$Marital.Status)*

*bike\_buyers$Gender <- as.factor(bike\_buyers$Gender)*

*bike\_buyers$Home.Owner <- as.factor(bike\_buyers$Home.Owner)*

*bike\_buyers$Purchased.Bike <- as.factor(bike\_buyers$Purchased.Bike)*

Dealing with NA values:

*bike\_buyers\_clean <- bike\_buyers*

We will replace NA values in Age and Income feature with Median.

Replace NA values in Marital.Status, Gender, Children, Home.Owner with Mode.

Replace NA values in Cars from Mean

# replace with median

*bike\_buyers\_clean$Income[is.na(bike\_buyers\_clean$Income)] <-median(na.omit((bike\_buyers$Income)))*

*bike\_buyers\_clean$Age[is.na(bike\_buyers\_clean$Age)] <- median(na.omit((bike\_buyers$Age)))*

*get\_mode <- function(x) {*

*unique\_x <- unique(x)*

*tabulate\_x <- tabulate(match(x, unique\_x))*

*unique\_x[tabulate\_x == max(tabulate\_x)]*

*}*

*# replace with mode*

*bike\_buyers\_clean$Marital.Status[is.na(bike\_buyers\_clean$Marital.Status)] <- get\_mode(bike\_buyers$Marital.Status)*

*bike\_buyers\_clean$Gender[is.na(bike\_buyers\_clean$Gender)] <- get\_mode(bike\_buyers$Gender)*

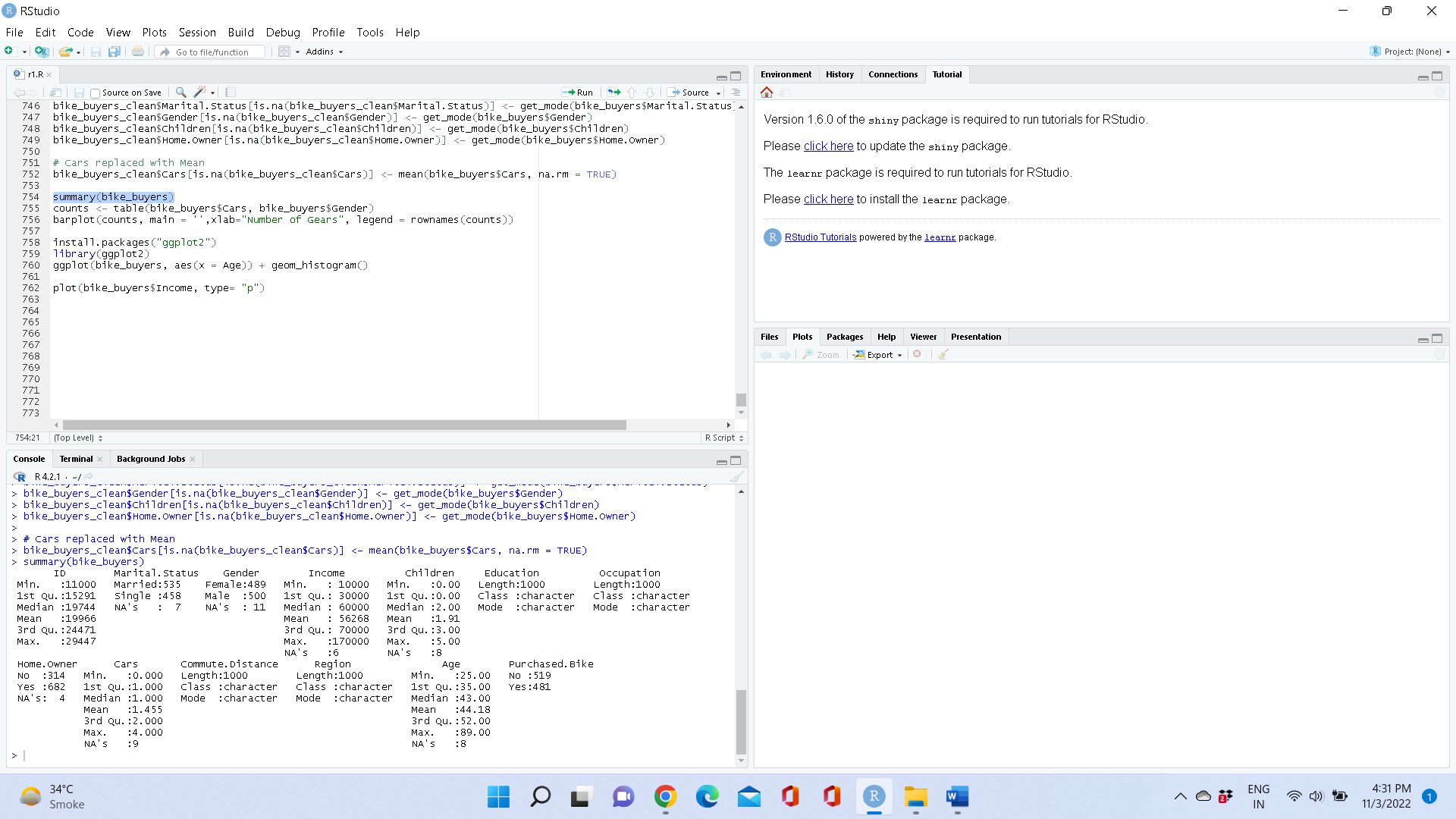
*bike\_buyers\_clean$Children[is.na(bike\_buyers\_clean$Children)] <- get\_mode(bike\_buyers$Children)*

*bike\_buyers\_clean$Home.Owner[is.na(bike\_buyers\_clean$Home.Owner)] <- get\_mode(bike\_buyers$Home.Owner)*

*# Cars replaced with Mean*

*bike\_buyers\_clean$Cars[is.na(bike\_buyers\_clean$Cars)] <- mean(bike\_buyers$Cars, na.rm = TRUE)*

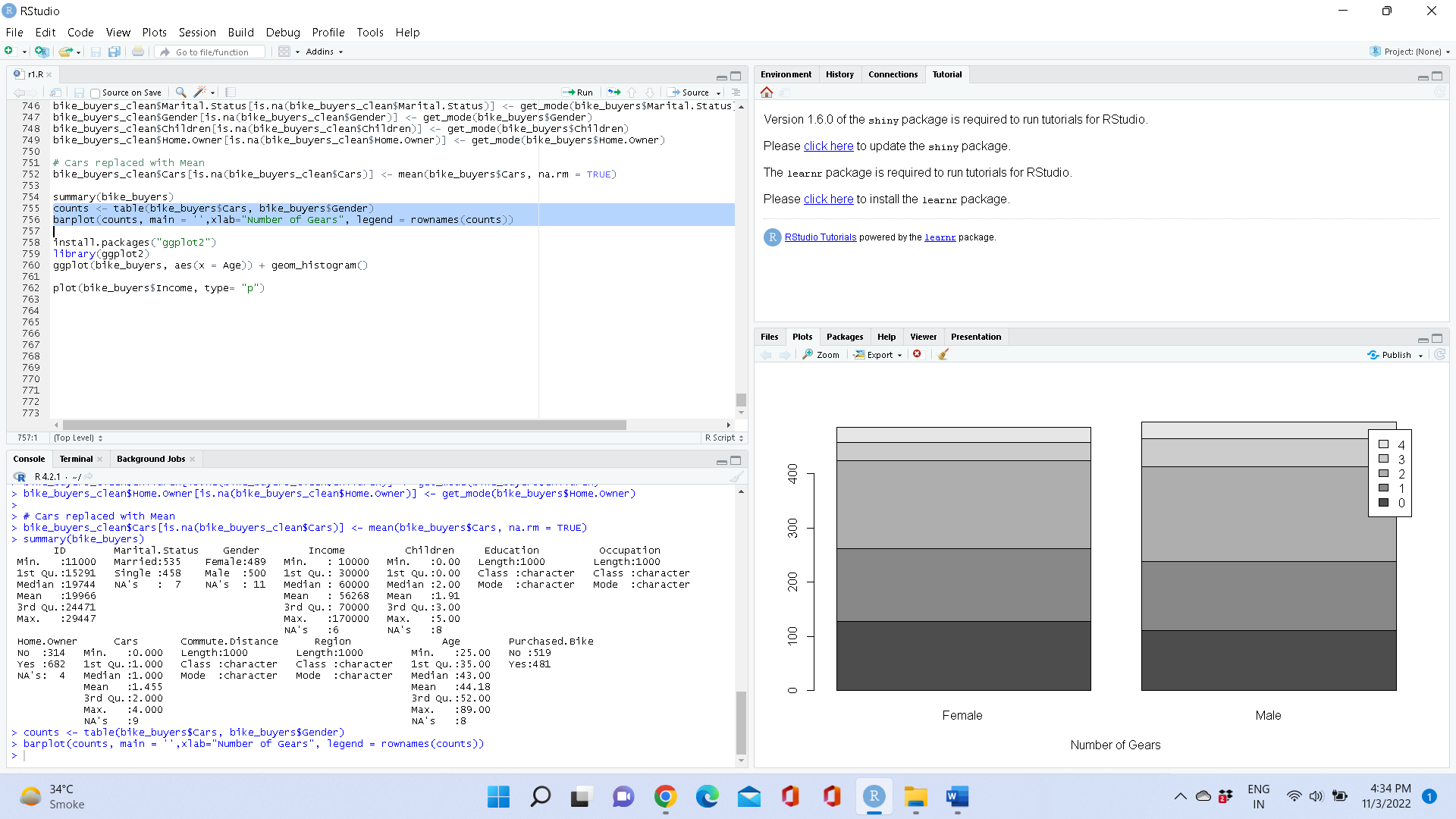
*summary(bike\_buyers)*



Visualizations:

*counts <- table(bike\_buyers$Cars, bike\_buyers$Gender)*

*barplot(counts, main = '',xlab="Number of Gears", legend = rownames(counts))*

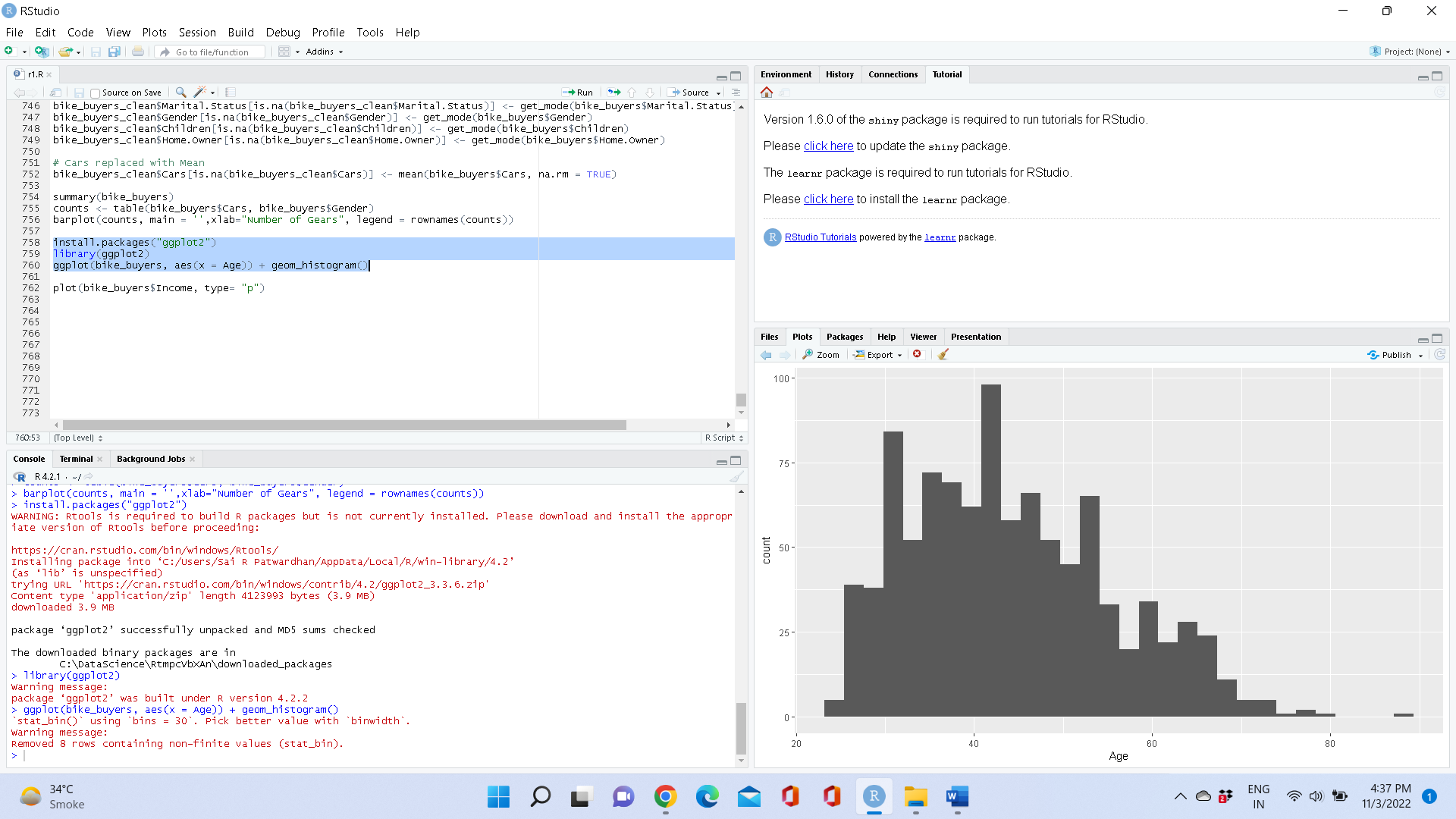
*\*

**Conclusion: There is no much difference in number of Male and Female bike buyers.**

*install.packages("ggplot2")*

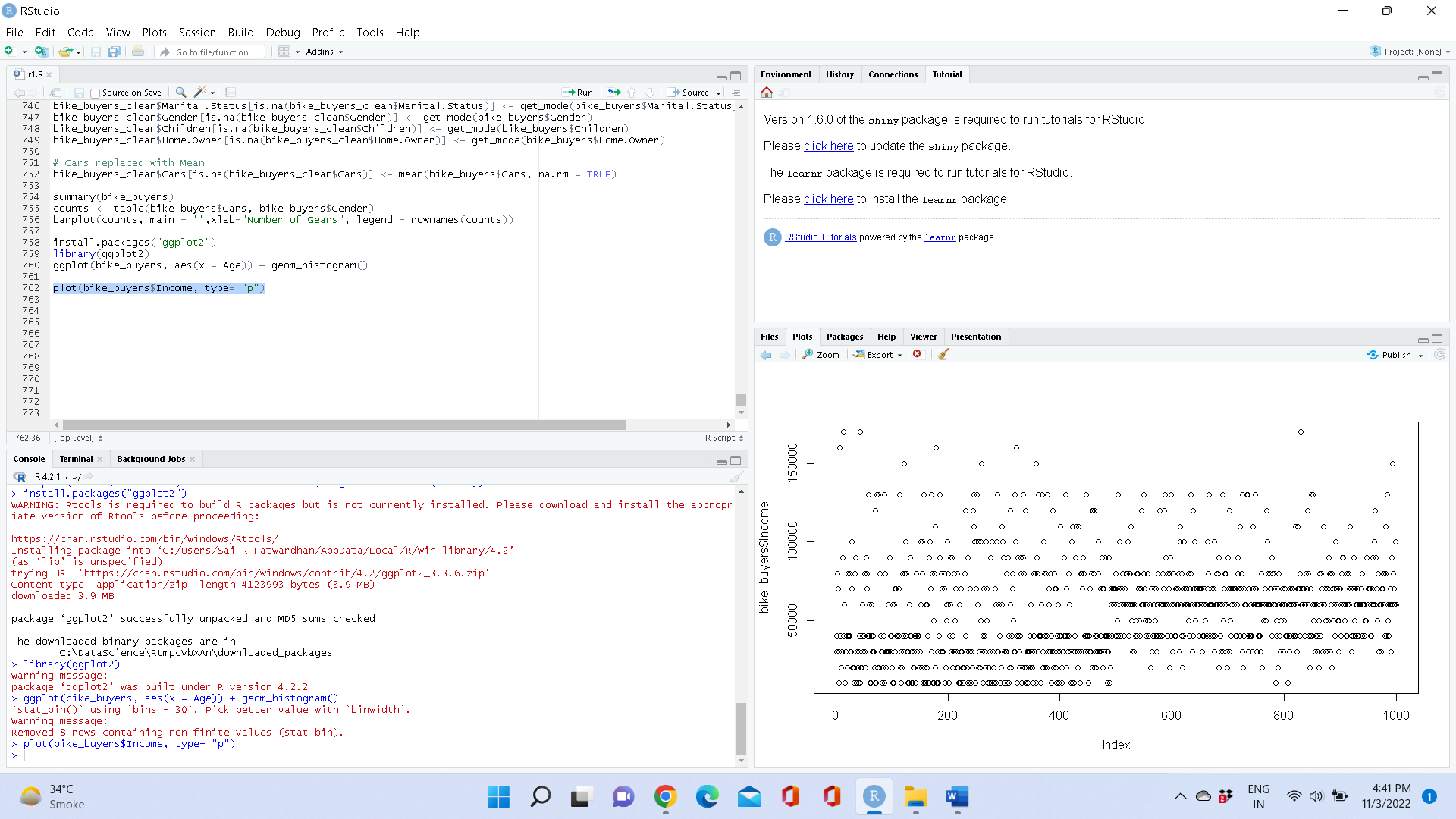
*library(ggplot2)*

*ggplot(bike\_buyers, aes(x = Age)) + geom\_histogram()*



**Conclusion: Most buyers are between 30-50 years of age.**

*plot(bike\_buyers$Income, type= "p")*



**Conclusion : People with higher income don’t purchase bike.**