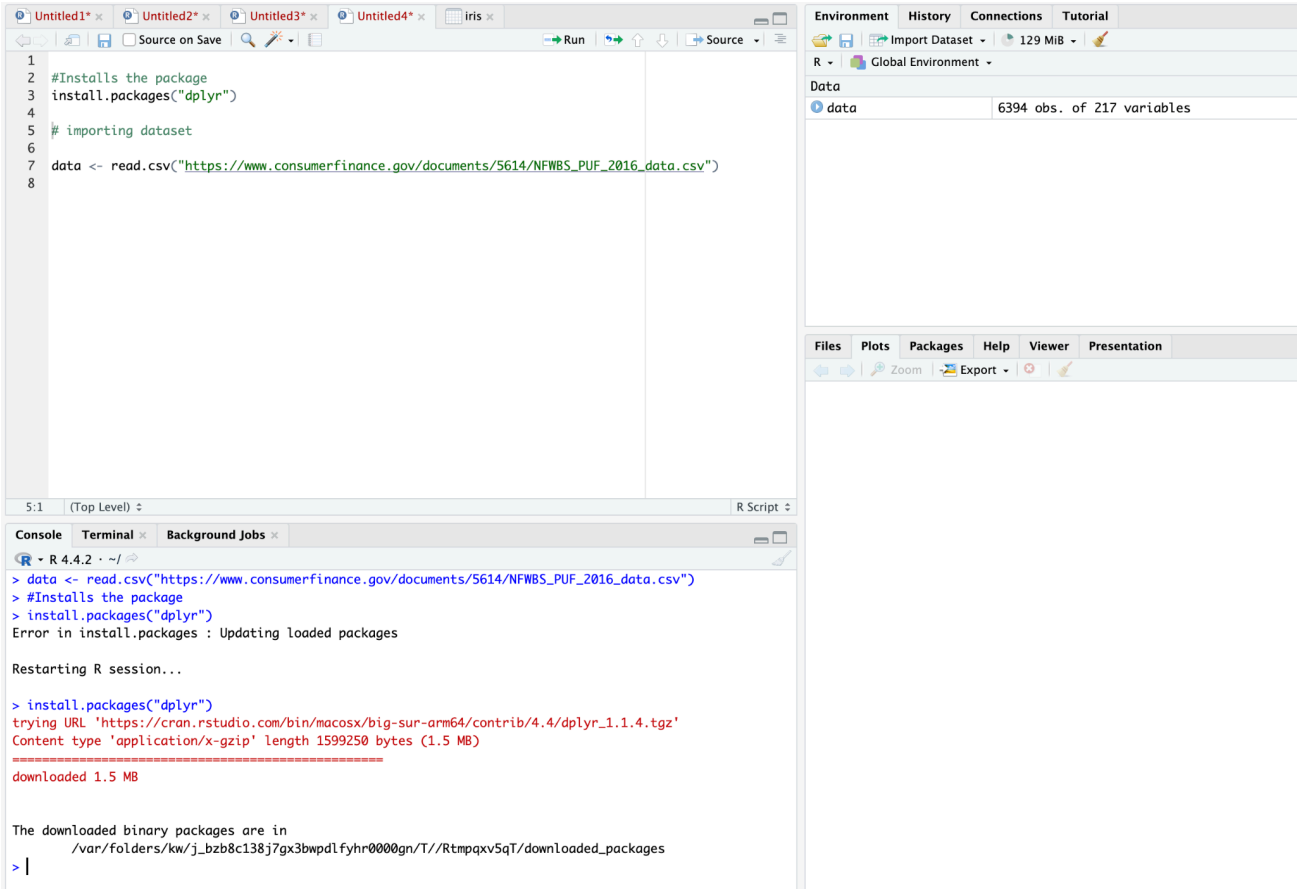
Analyzing Survey Data in R

Importing a .csv file directly from the web

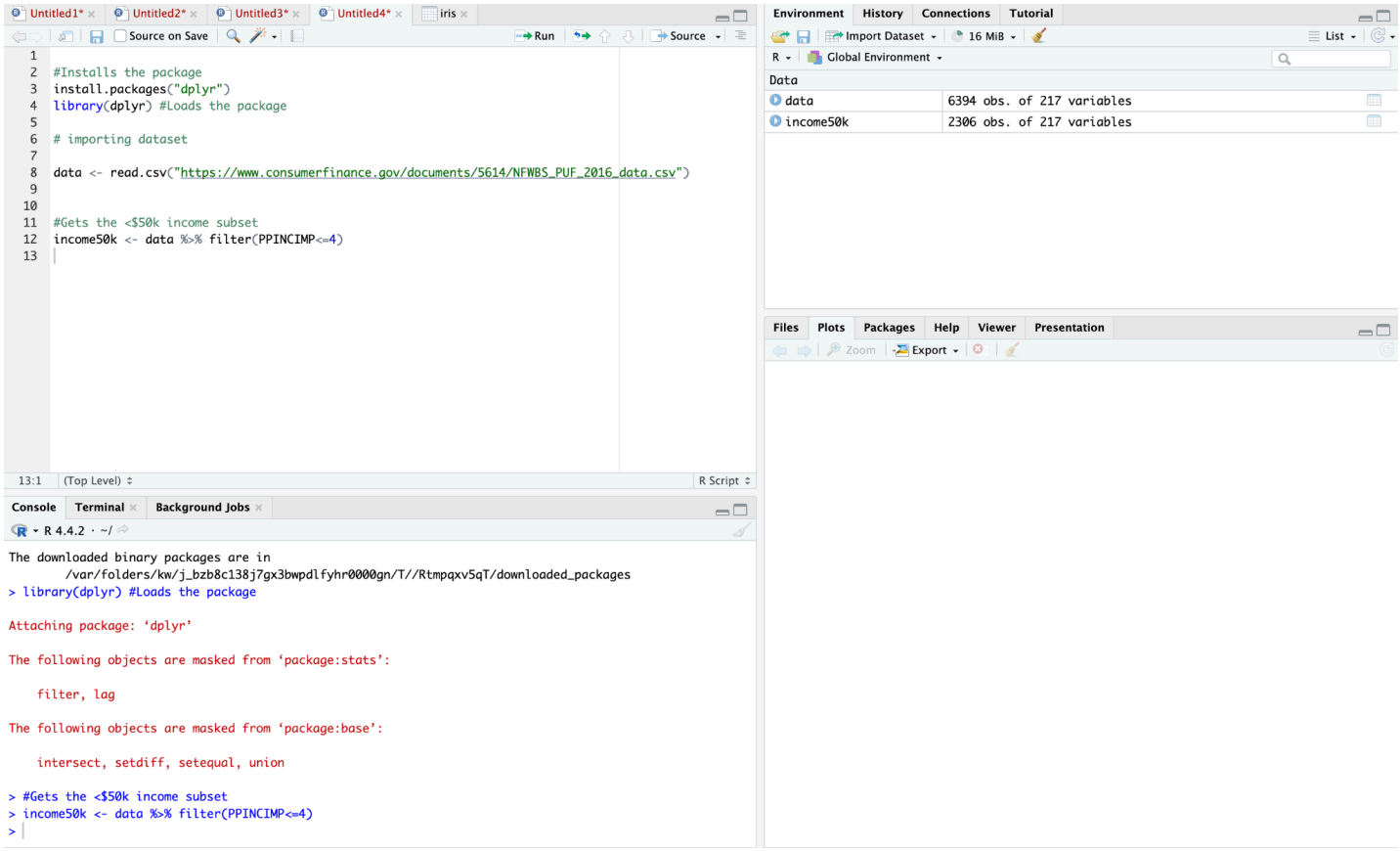
importing dataset

data <-
read.csv("https://www.consumerfinance.gov/documents/5614/NFWBS_PUF_2016_data.csv")



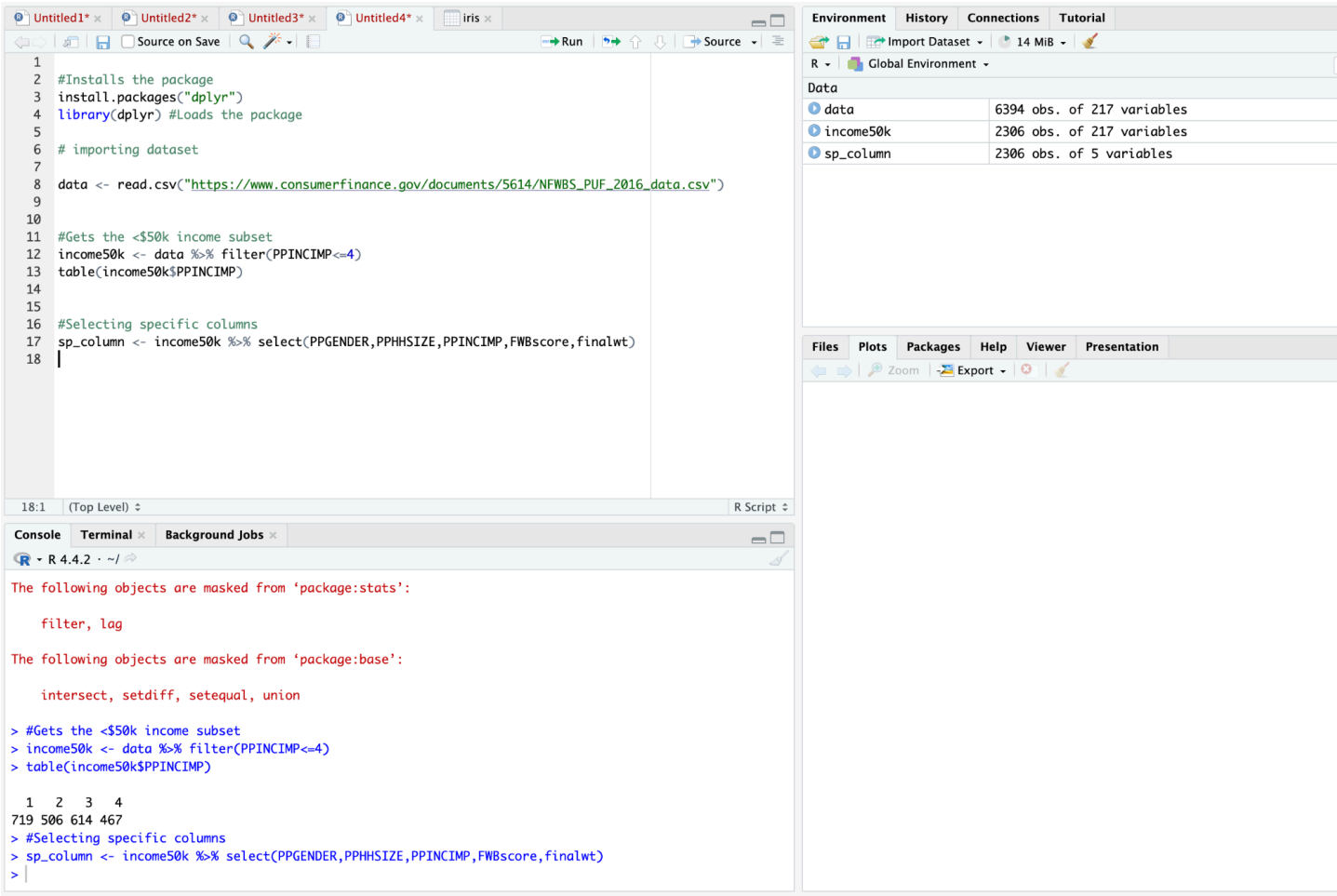
Creating a subset

#Gets the <$50k income subset
income50k <- data %>% filter(PPINCIMP<=4)



Selecting specific columns

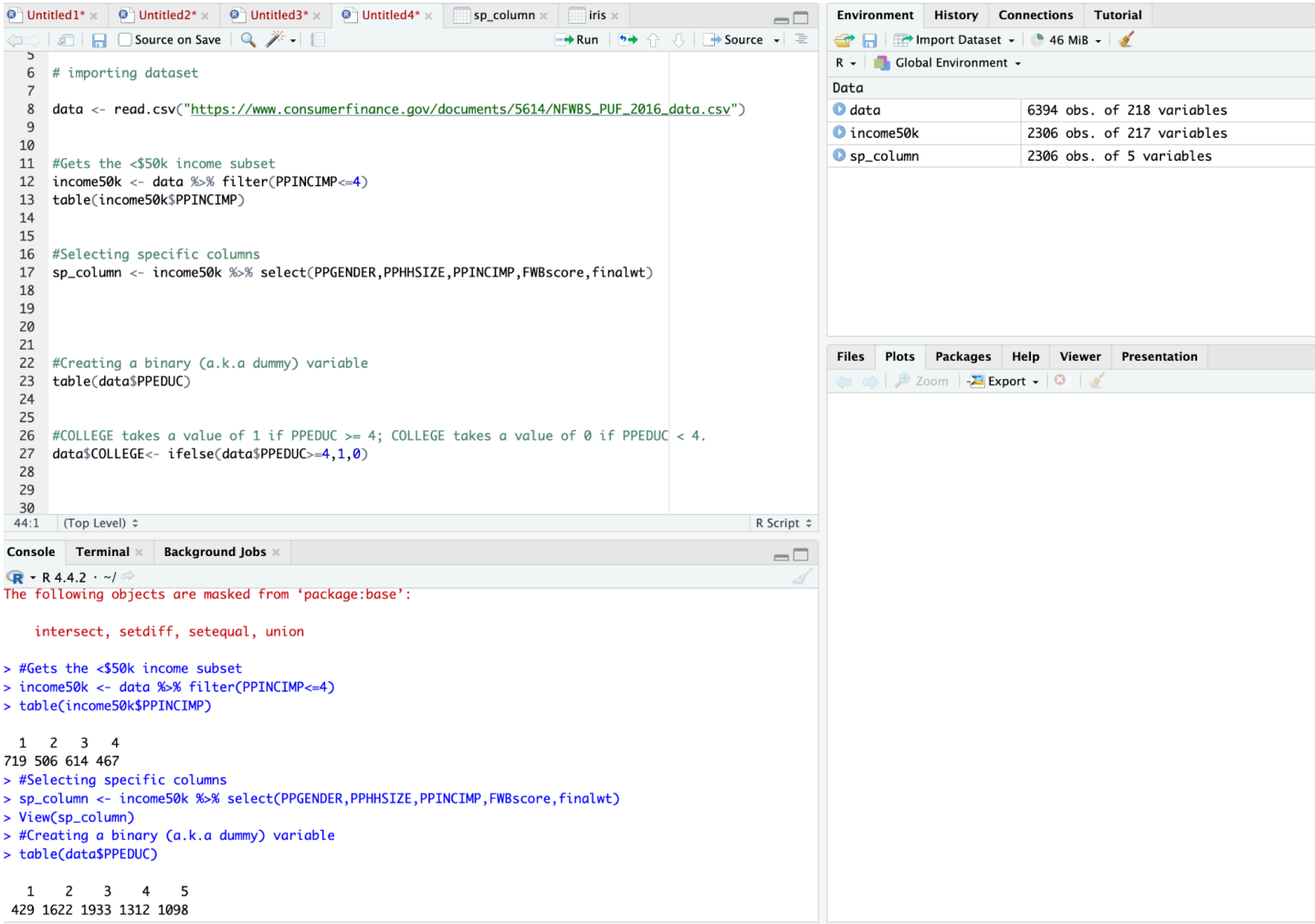
```
sp_column <- income50k %>% select(PPGENDER,PPHHSIZE,PPINCIMP,FWBscore,finalwt)
```



Creating a binary (a.k.a dummy) variable

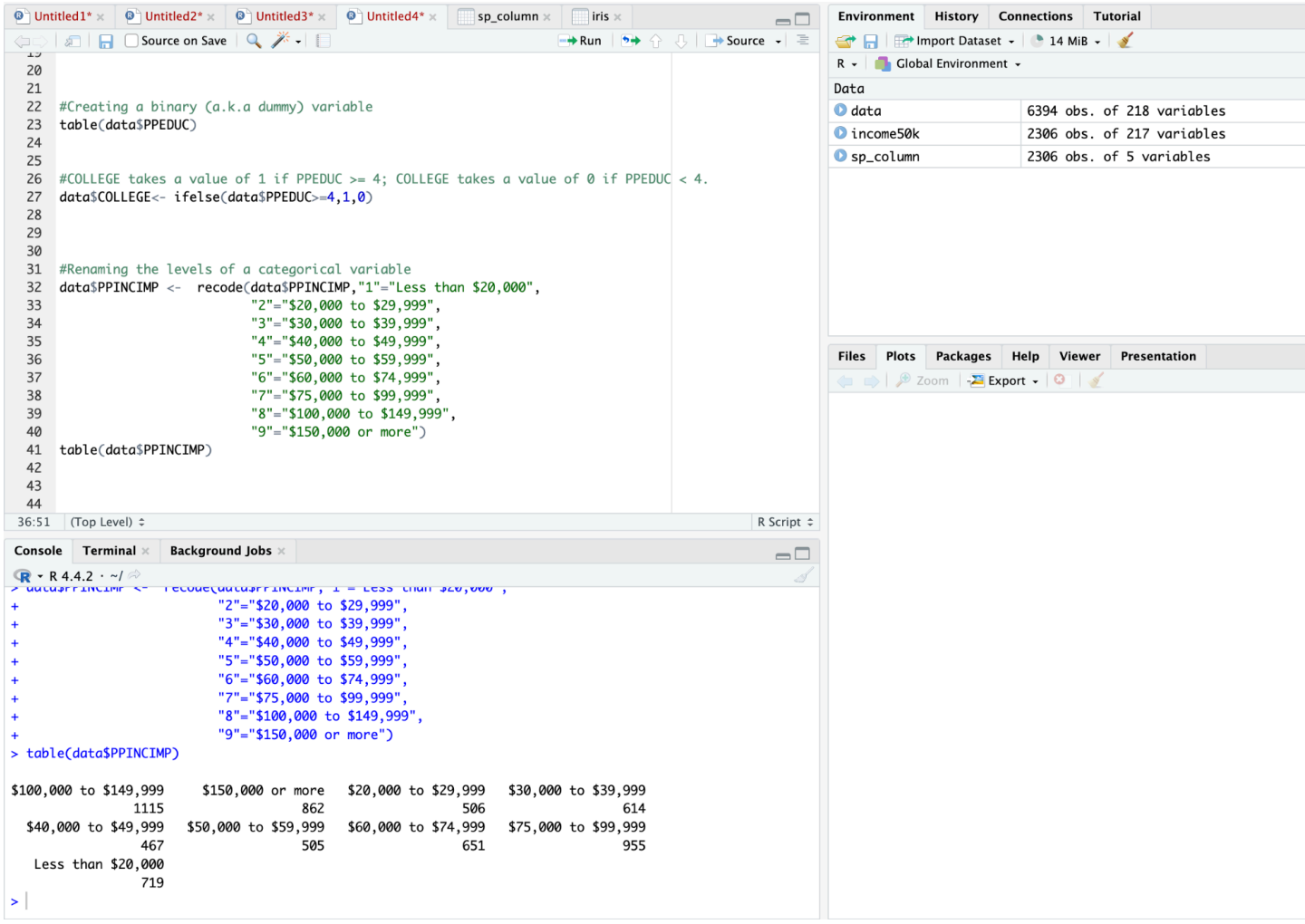
table(data$PPEDUC)

#COLLEGE takes a value of 1 if PPEDUC >= 4; COLLEGE takes a value of 0 if PPEDUC < 4.
data$COLLEGE<- ifelse(data$PPEDUC>=4,1,0)



Renaming the levels of a categorical variable

```
#Renaming the levels of a categorical variable
data$PPINCIMP <-  recode(data$PPINCIMP,"1"="Less than $20,000",
                         "2"="$20,000 to $29,999",
                         "3"="$30,000 to $39,999",
                         "4"="$40,000 to $49,999",
                         "5"="$50,000 to $59,999",
                         "6"="$60,000 to $74,999",
                         "7"="$75,000 to $99,999",
                         "8"="$100,000 to $149,999",
                         "9"="$150,000 or more")
table(data$PPINCIMP)
```



Creating a new categorical variable

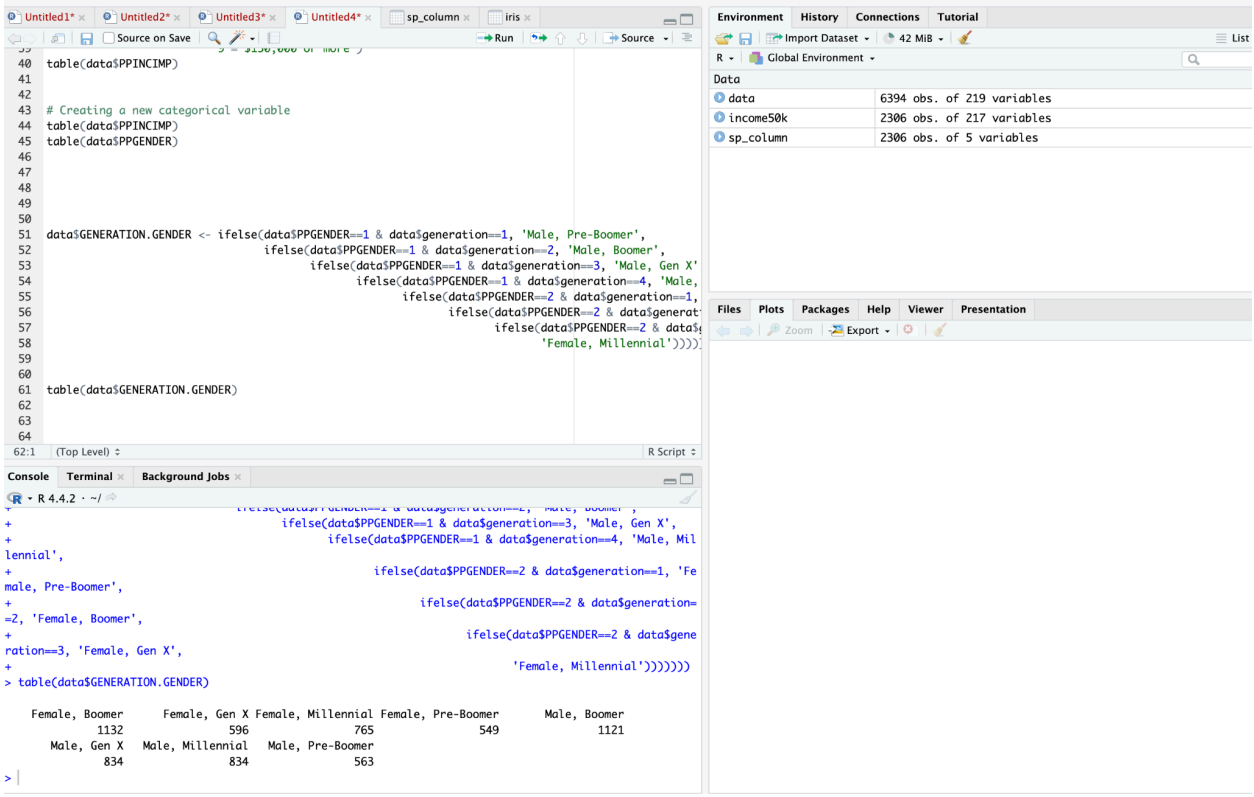
```
data$GENERATION.GENDER <- ifelse(data$PPGENDER==1 & data$generation==1, 'Male,
Pre-Boomer',
                          ifelse(data$PPGENDER==1 & data$generation==2, 'Male, Boomer',
                              ifelse(data$PPGENDER==1 & data$generation==3, 'Male, Gen X',
                                  ifelse(data$PPGENDER==1 & data$generation==4, 'Male,
Millennial',
                                         ifelse(data$PPGENDER==2 & data$generation==1, 'Female,
Pre-Boomer',
                                                ifelse(data$PPGENDER==2 & data$generation==2,
'Female, Boomer',
                                                       ifelse(data$PPGENDER==2 & data$generation==3,
'Female, Gen X',
                                                              'Female, Millennial')))))))


#I have data$GENERATION.GENDER at the start of the code because I made a new variable
called GENERATION.GENDER."
```

We write ifelse() conditions seven times since the intersection of the two variables (generation and PPGENDER) has eight categories.

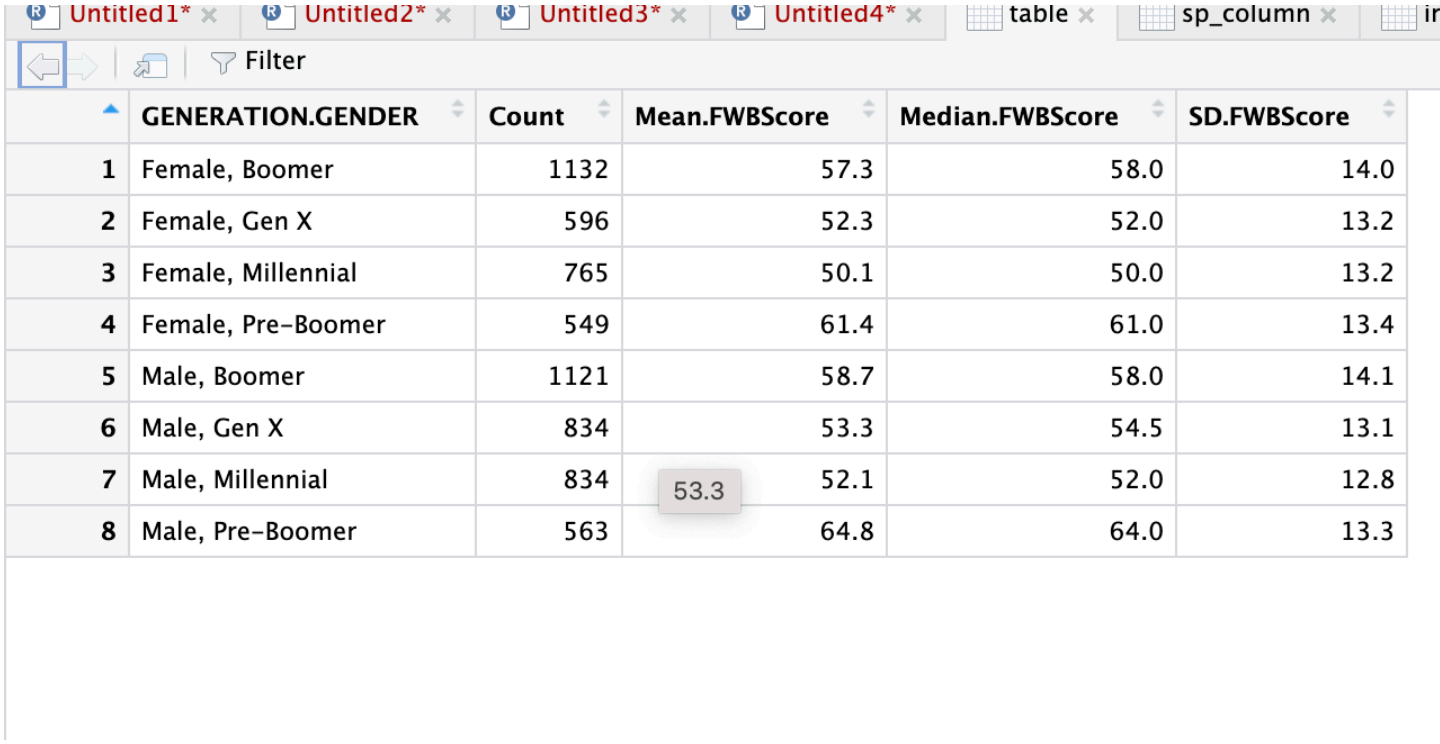
The last category does not require a condition because a response will be placed in the final group ('Female, Millennial') if the seven conditions, which pertain to seven categories, do not match.

```
table(data$GENERATION.GENDER)
```


Untitled1*
Untitled2*
Untitled3*
Untitled4*
sp_column
iris
Source on Save
Run
Source
table(data$PPINCIMP)
# Creating a new categorical variable
table(data$PPINCIMP)
table(data$PPGENDER)
data$GENERATION.GENDER <- ifelse(data$PPGENDER==1 & data$generation==1, 'Male, Pre-Boomer',
ifelse(data$PPGENDER==1 & data$generation==2, 'Male, Boomer',
ifelse(data$PPGENDER==1 & data$generation==3, 'Male, Gen X'
ifelse(data$PPGENDER==1 & data$generation==4, 'Male,
ifelse(data$PPGENDER==2 & data$generation==1,
ifelse(data$PPGENDER==2 & data$generat
ifelse(data$PPGENDER==2 & data$
'Female, Millennial'))))
table(data$GENERATION.GENDER)
62:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.4.2 · ~/
Female, Boomer Female, Gen X Female, Millennial Female, Pre-Boomer Male, Boomer
1132 596 765 549 1121
Male, Gen X Male, Millennial Male, Pre-Boomer
834 834 563
Environment
History
Connections
Tutorial
Import Dataset
42 MiB
List
R
Global Environment
Data
data 6394 obs. of 219 variables
income50k 2306 obs. of 217 variables
sp_column 2306 obs. of 5 variables
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export

Creating a summary statistics table

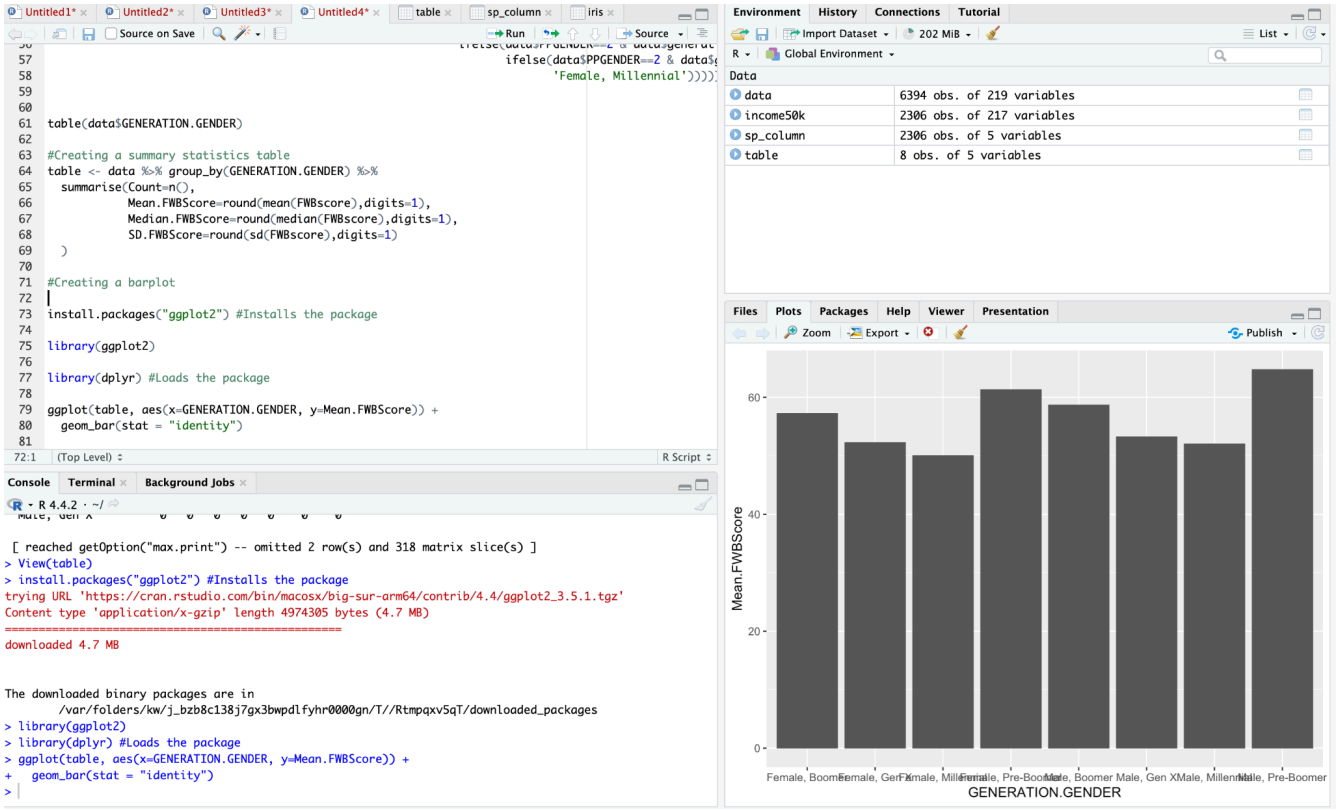
```
table <- data %>% group_by(GENERATION.GENDER) %>%
  summarise(Count=n(),
            Mean.FWBScore=round(mean(FWBscore),digits=1),
            Median.FWBScore=round(median(FWBscore),digits=1),
            SD.FWBScore=round(sd(FWBscore),digits=1)
  )
```



	GENERATION.GENDER	Count	Mean.FWBScore	Median.FWBScore	SD.FWBScore
1	Female, Boomer	1132	57.3	58.0	14.0
2	Female, Gen X	596	52.3	52.0	13.2
3	Female, Millennial	765	50.1	50.0	13.2
4	Female, Pre-Boomer	549	61.4	61.0	13.4
5	Male, Boomer	1121	58.7	58.0	14.1
6	Male, Gen X	834	53.3	54.5	13.1
7	Male, Millennial	834	53.3	52.0	12.8
8	Male, Pre-Boomer	563	64.8	64.0	13.3

Creating a barplot

install.packages("ggplot2") #Installs the package
library(ggplot2)
library(dplyr) #Loads the package
ggplot(table, aes(x=GENERATION.GENDER, y=Mean.FWBScore)) +
  geom_bar(stat = "identity")



ggplot(table, aes(x=GENERATION.GENDER, y=Mean.FWBScore)) +
  geom_bar(stat = "identity")+
  coord_flip()+
  theme_light()+
  labs(y="Average Financial Well-Being Score", x=" ")

```
  summarise(Count=n(),
            Mean.FWBScore=round(mean(FWBscore),digits=1),
            Median.FWBScore=round(median(FWBscore),digits=1),
            SD.FWBScore=round(sd(FWBscore),digits=1)
  )

#Creating a barplot

install.packages("ggplot2") #Installs the package

library(ggplot2)

library(dplyr) #Loads the package

ggplot(table, aes(x=GENERATION.GENDER, y=Mean.FWBScore)) +
  geom_bar(stat = "identity")



ggplot(table, aes(x=GENERATION.GENDER, y=Mean.FWBScore)) +
  geom_bar(stat = "identity")+
  coord_flip()+
  theme_light()+
  labs(y="Average Financial Well-Being Score", x=" ")
```

```
Content type 'application/x-gzip' length 4974305 bytes (4.7 MB)
==================================================
downloaded 4.7 MB


The downloaded binary packages are in
        /var/folders/kw/j_bzb8c138j7gx3bwpdlfyhr0000gn/T//Rtmpqxv5qT/downloaded_packages
> library(ggplot2)
> library(dplyr) #Loads the package
> ggplot(table, aes(x=GENERATION.GENDER, y=Mean.FWBScore)) +
+   geom_bar(stat = "identity")
> ggplot(table, aes(x=GENERATION.GENDER, y=Mean.FWBScore)) +
+   geom_bar(stat = "identity")+
+   coord_flip()+
+   theme_light()+
+   labs(y="Average Financial Well-Being Score", x=" ")
>
```

Data	
data	6394 obs. of 219 variables
income50k	2306 obs. of 217 variables
sp_column	2306 obs. of 5 variables
table	8 obs. of 5 variables