APPLIED STATISTICAL ANALYSIS LAB

NAME: ADITI KULKARNI

ROLL NO: 55

YEAR: SY

DIVISION: E (E3)

SRN NO: 202201893

ASSIGNMENT 4

STATEMENT: To create two-way cross tabulations to explore the relationship between several variables and to use the Chart Builder to visualize the relationship Using the preexisting Census.csv data file.

THEORY:

data=read.csv(file.choose()) allows the user to interactively select a CSV file, which is then loaded into an R dataframe named data. It's a common way to import data into R.

str(data) provides a structured summary of the data dataframe, including its structure, data types, and a preview of the data. This helps in understanding the data's characteristics.

tab1=table(data$age, data$marital.status) creates a two-way contingency table tab1 to explore the relationship between the variables age and marital.status.

tab2=table(data$race, data$sex) creates another contingency table tab2 to explore the relationship between the variables race and sex.

margin.table(tab2, 1) calculates row totals for the tab2 contingency table, giving the total counts for each race category.

margin.table(tab2, 2) calculates column totals for the tab2 contingency table, giving the total counts for each sex category.

library(ggplot2) loads the ggplot2 package for creating data visualizations.

dt=data.frame(tab2) converts the tab2 contingency table into a data frame called dt.

colnames(dt)=c("Race","Sex","Freq") assigns meaningful column names to the dt data frame.

The subsequent lines of code use ggplot2 to create various visualizations such as bar charts and line plots to visually explore and represent the relationships between race and sex based on the data in the dt data frame.

SOURCE CODE: data=read.csv(file.choose())

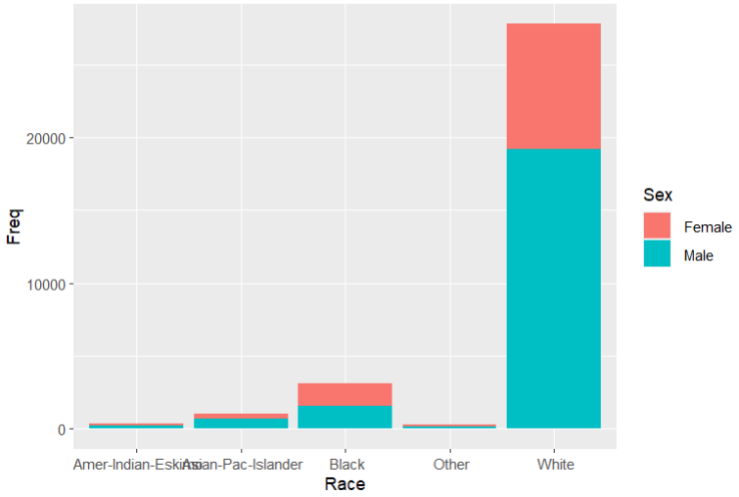
```
class(data)

View(data)

str(data)

#normal cross tables

tab1=table(data$age,data$marital.status) # comparing age with marital status

tab1

tab2=table(data$race,data$sex) # comparing race with sex

tab2

margin.table(tab2,1) # row totals

margin.table(tab2,2) # columns totals

prop.table(tab2) # proportions based on overall totals

prop.table(tab2,1) # proportions based on row totals

prop.table(tab2,2) #proportions based on column totals

# Bar Charts

library(ggplot2)

dt=data.frame(tab2)
```

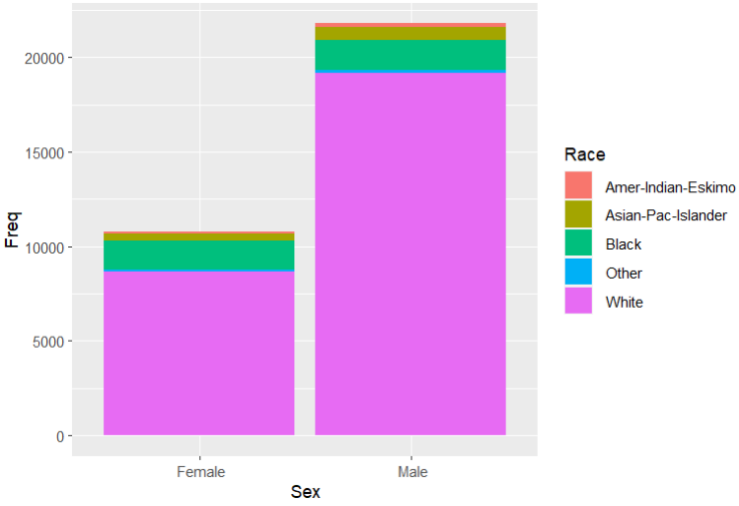
```
dt
colnames(dt)=c("Race","Sex","Freq")
dt
ggplot(dt,aes(x=Race, y=Freq , fill=Sex)) + geom_col()
ggplot(dt,aes(x=Sex, y=Freq , fill=Race)) + geom_col()
ggplot(dt, aes(x=Race, y=Freq, fill=Sex)) + geom_col(position="dodge")
# Line Plots
ggplot(dt, aes(x=Race, y=Freq, color=Sex, group=Sex)) + geom_line()
```

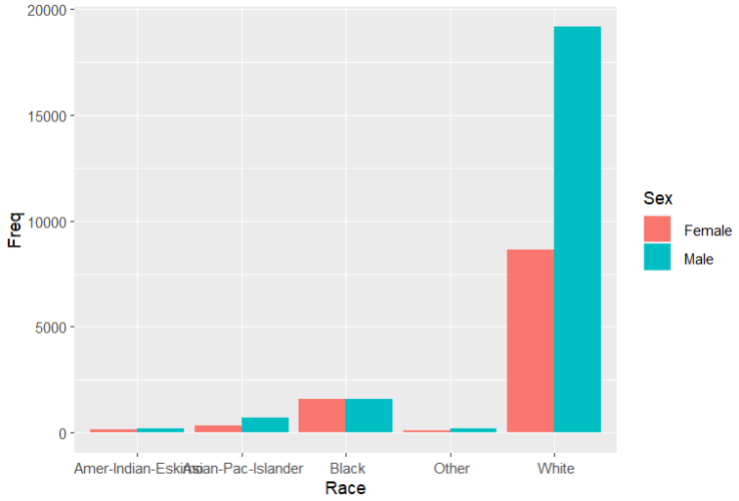
OUTPUT:

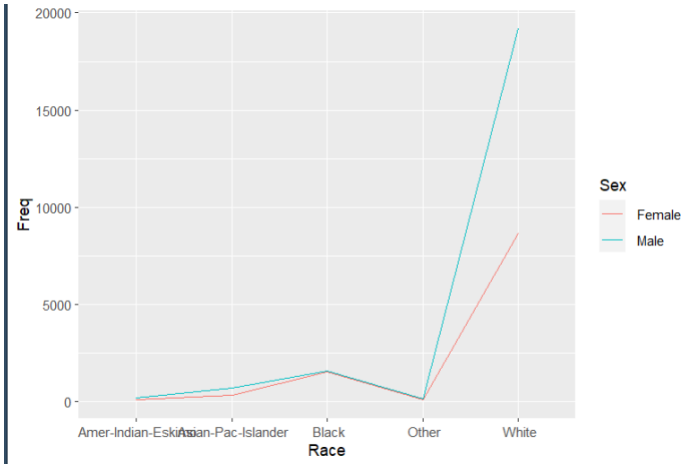
```
prop.table(tab2) # proportions based on overall totals

                         Female        Male
 Amer-Indian-Eskimo 0.003654679 0.005896625
 Asian-Pac-Islander 0.010626209 0.021283130
 Black              0.047756519 0.048186481
 Other              0.003347563 0.004975277
 White              0.265409539 0.588863978
prop.table(tab2,1) # proportions based on row totals

                        Female      Male
 Amer-Indian-Eskimo 0.3826367 0.6173633
 Asian-Pac-Islander 0.3330125 0.6669875
 Black              0.4977593 0.5022407
 Other              0.4022140 0.5977860
 White              0.3106845 0.6893155
prop.table(tab2,2) #proportions based on column totals

                         Female        Male
 Amer-Indian-Eskimo 0.011048185 0.008811381
 Asian-Pac-Islander 0.032123294 0.031803580
 Black              0.144369139 0.072005507
 Other              0.010119766 0.007434603
 White              0.802339616 0.879944929
```

```
> dt=data.frame(tab2)
> dt
                  Var1    Var2  Freq
1   Amer-Indian-Eskimo  Female   119
2   Asian-Pac-Islander  Female   346
3                Black  Female  1555
4                Other  Female   109
5                White  Female  8642
6   Amer-Indian-Eskimo    Male   192
7   Asian-Pac-Islander    Male   693
8                Black    Male  1569
9                Other    Male   162
10               White    Male 19174
> colnames(dt)=c("Race","Sex","Freq")
> dt
                  Race     Sex  Freq
1   Amer-Indian-Eskimo  Female   119
2   Asian-Pac-Islander  Female   346
3                Black  Female  1555
4                Other  Female   109
5                White  Female  8642
6   Amer-Indian-Eskimo    Male   192
7   Asian-Pac-Islander    Male   693
8                Black    Male  1569
9                Other    Male   162
10               White    Male 19174
```


20000
10000
0
Freq
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White
Race
Sex
Female
Male


20000
15000
10000
5000
0
Freq
Female
Male
Sex
Race
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White


20000
15000
10000
5000
0
Freq
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White
Race
Sex
Female
Male



CONCLUSION: This code showcases a typical data exploration and visualization workflow in R. It is designed to help users gain insights into the relationships between specific variables in the dataset, with a focus on age, marital status, race, and sex. The code enables users to create informative visualizations that facilitate a better understanding of the data's patterns and trends. Overall, it's a valuable tool for initial data analysis and visualization in R.