Practice Lab

Carey Kopeikin & Matthew Bardoe

10/31/2020

A New Kind of Assignment

Most of our assignments so far have been worksheets that you either work through on your own, or we work through together. These worksheets provide you with background on how to utilize `R` to investigate, manipulate, and display data. In our next assignment we will give you a much more open ended assignment. You will choose from between several datasets, and we ask you to create a document that shows your exploration of the data, and some of the interesting features of the data that you were able to find.

This leads to several questions:

- How do I do this?
- How will I be graded?

How do I do this

Follow the following steps to explore the data.

1. Understand the data. What are the variables? What do they mean what are the ranges that we see if it is a numeric variable. If it is a character variable would it be better as a factor, what are the levels if it is a factor. Can I create graphs to show the variation inside a variable? Can I regroup a variable? Can I create sub-dataframes of data and explore those?
2. How will I be graded? We will be looking for several things.

a. Understanding of the data. What are variables what do they mean and how many observations are in the dataset.

b. Good displays of the data. All displays have titles, labels where appropriate, are easy to read, and are colorful.

c. Good use of R-markdown. Do you have section titles? Do you use bold and other formatting well? Does my file look good as a pdf? https://www.dataquest.io/blog/r-markdown-guide-cheatsheet/#tve-jump-17333d75ada

d. Analysis of the displays of data. Each display of data comes with an explanation of what is shown in words. That explanation uses correct terminology and is accurate.

e. Creativity of the analysis. Did you find something cool, interesting, unexpected, remarkably clear. Can you create a display that shows it easily or beautifully.

How to Present Your Work

There should be three sections to your final lab:

1. Introduction: Details about the Data and questions you are going to investigate.
2. Observations and Analysis: Observations from the Data
3. Conclusion

Datasets

The four datasets are:

- **Aircraft-Wildlife Collisions**, `birds.csv` Information on this dataset is at: https://www.openintro.org/data/index.php?data=birds
- **Flights data**, `nycflights.csv` Information on this dataset is at: https://www.openintro.org/data/index.php?data=nycflights
- **Youth Risk Behavior Surveillance System** `yrbss.csv` Information on this dataset is at: https://www.openintro.org/data/index.php?data=yrbss
- **UK Smoking Data "`smoking.csv`"** Information on this dataset is at: https://www.openintro.org/data/index.php?data=smoking

Introduction

The data set tells us about the different demographics of smokers and non-smokers, such as gender, age, education qualification, and marital status. I want to examine first whether there is a direct relationship between smoking and singular variables, such as gender, marital status, and age. Then, I want to explore how several of these variables together affect smoking to see whether there is a trend in the type of people that are smokers or non-smokers.

observation and analysis

```
smoking.data <- read.csv("smoking.csv")
head(smoking.data)
```

```
##   gender age marital_status highest_qualification nationality ethnicity
## 1   Male  38       Divorced      No Qualification     British     White
## 2 Female  42         Single      No Qualification     British     White
## 3   Male  40        Married                Degree     English     White
## 4 Female  40        Married                Degree     English     White
## 5 Female  39        Married          GCSE/O Level     British     White
## 6 Female  37        Married          GCSE/O Level     British     White
##       gross_income    region smoke amt_weekends amt_weekdays    type
## 1   2,600 to 5,200 The North    No           NA           NA
## 2      Under 2,600 The North   Yes           12           12 Packets
## 3 28,600 to 36,400 The North    No           NA           NA
## 4 10,400 to 15,600 The North    No           NA           NA
## 5   2,600 to 5,200 The North    No           NA           NA
## 6 15,600 to 20,800 The North    No           NA           NA
```

After observing the data, we found that many of the categories would fit better if we make some changes.

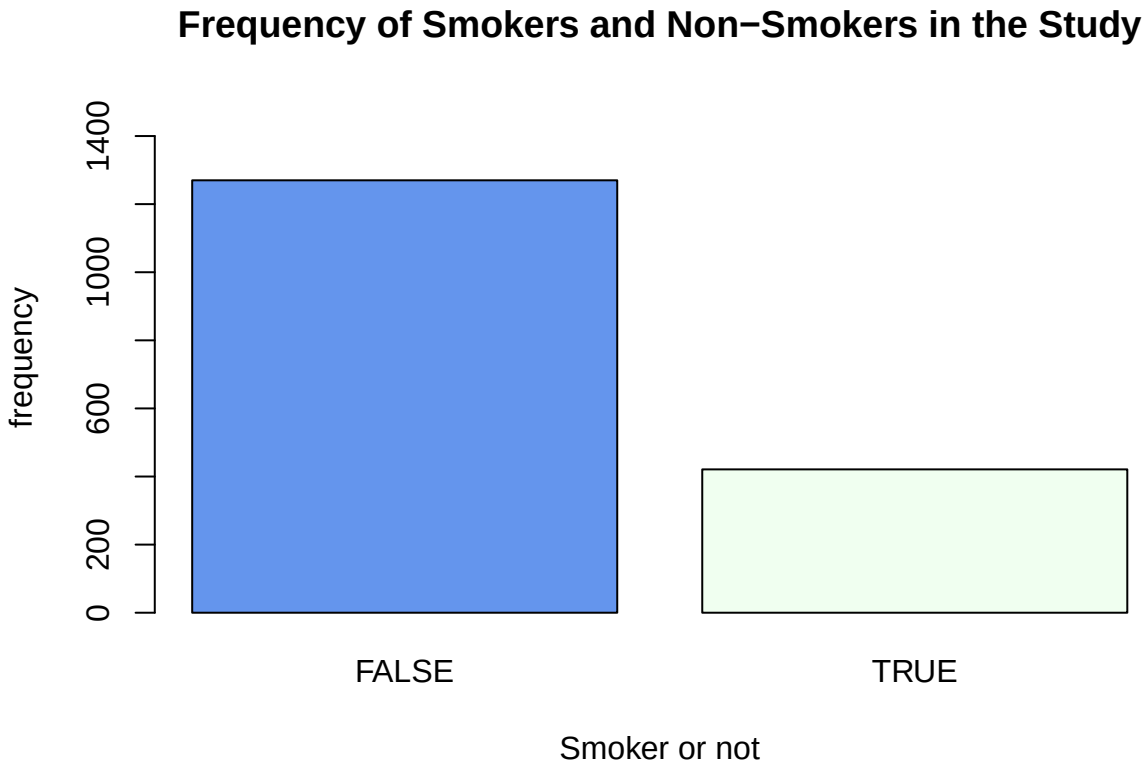
```
smoking.data$gender <- as.factor(smoking.data$gender)

smoking.data$smoke <- smoking.data$smoke == "Yes"

smoking.data$marital_status <- as.factor(smoking.data$marital_status)

smoking.data$highest_qualification <- as.factor(smoking.data$highest_qualification)

smoking.data$nationality <- as.factor(smoking.data$nationality)

smoking.data$ethnicity <- as.factor(smoking.data$ethnicity)

smoking.data$gross_income <- as.factor(smoking.data$gross_income)
```

```
head(smoking.data)
```

```
##   gender age marital_status highest_qualification nationality ethnicity
## 1   Male  38       Divorced      No Qualification     British     White
## 2 Female  42         Single      No Qualification     British     White
## 3   Male  40        Married                Degree     English     White
## 4 Female  40        Married                Degree     English     White
## 5 Female  39        Married          GCSE/O Level     British     White
## 6 Female  37        Married          GCSE/O Level     British     White
##       gross_income    region smoke amt_weekends amt_weekdays    type
## 1   2,600 to 5,200 The North FALSE           NA           NA
## 2      Under 2,600 The North  TRUE           12           12 Packets
## 3 28,600 to 36,400 The North FALSE           NA           NA
## 4 10,400 to 15,600 The North FALSE           NA           NA
## 5   2,600 to 5,200 The North FALSE           NA           NA
## 6 15,600 to 20,800 The North FALSE           NA           NA
```

First of all, we want to understand the an overview of how many people in this data set smoke and how many don't.

```
barplot( table( smoking.data$smoke),
         main = "Frequency of Smokers and Non-Smokers in the Study",
         xlab = "Smoker or not",
         ylab = "frequency",
         col = c("cornflowerblue", "honeydew") ,
         ylim=c(0,1500))
```

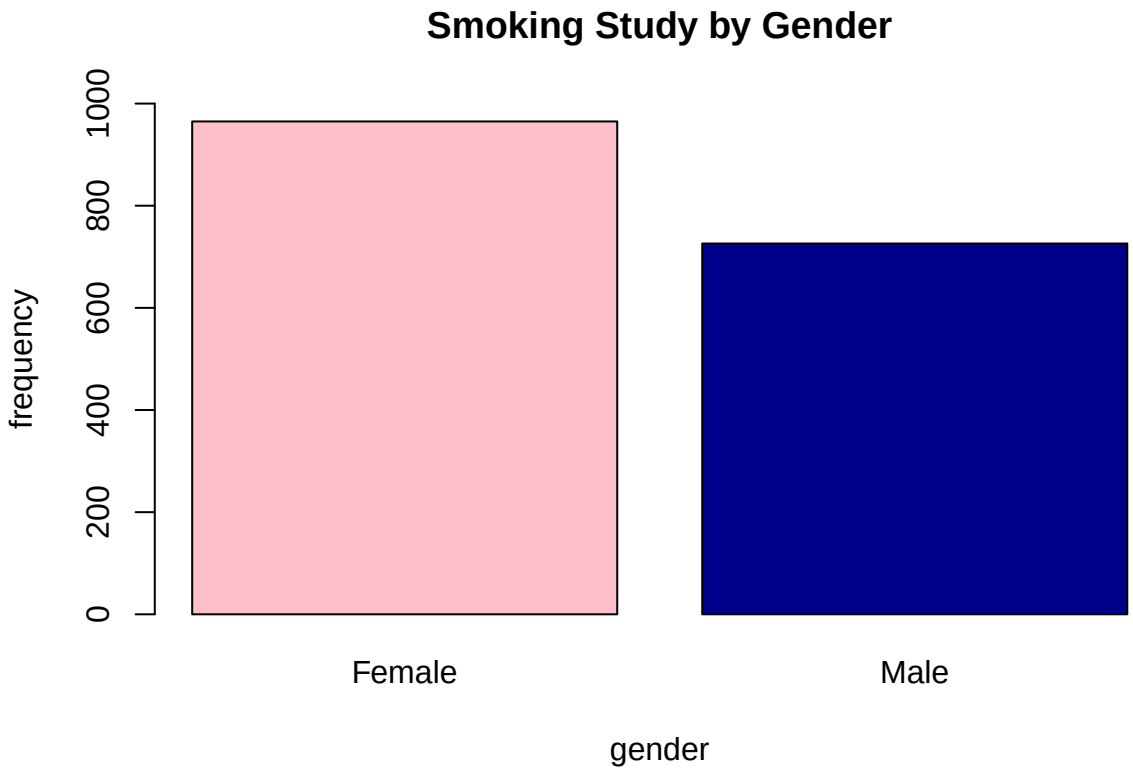


The bar graph shows that in this specific study, there are significantly more non-smokers than smokers.There are just under 1300 non-smokers while there is a little bit over 400 smokers, which means there are over 3x the amount of non-smokers. We can expect that this data reflects the demographics of smokers and non-smokers

across the UK population because, according to the data set website, the sample was random and varied and can be used for analyzing demographic characteristics of smokers.

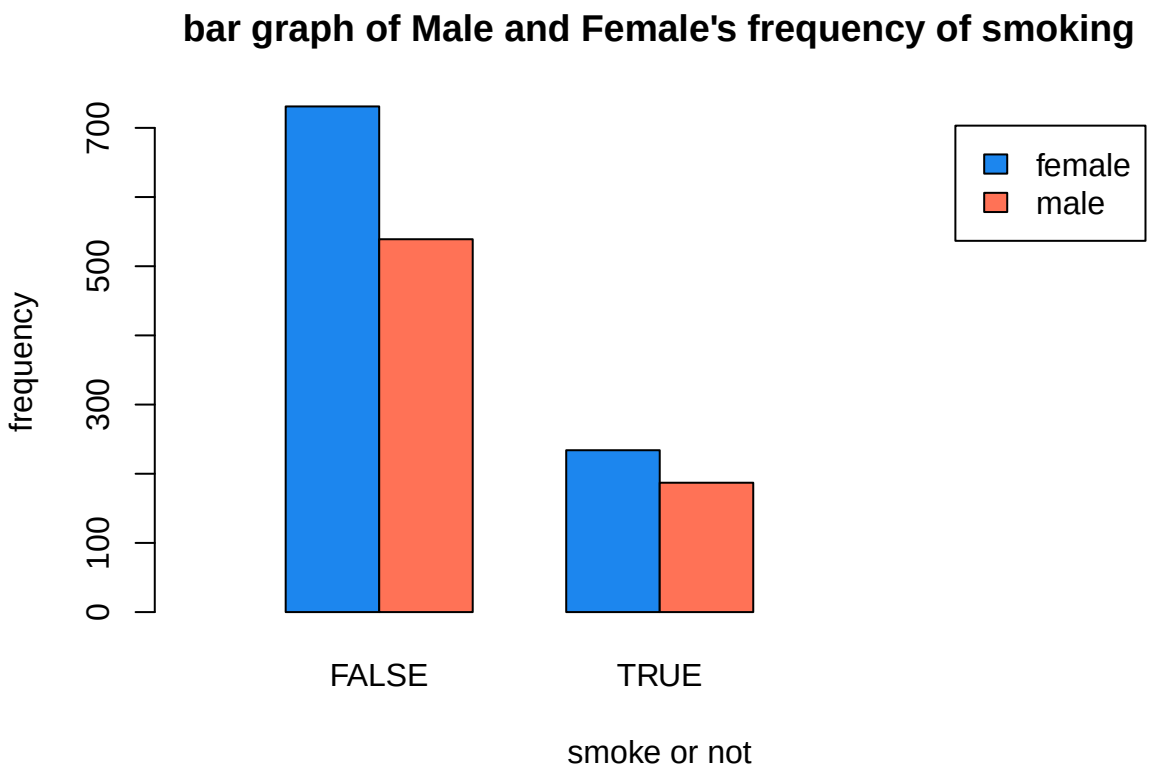
We wanted to understand the relationship between gender and smoking. Therefore, we made a barplot that compares the total data of the number of male and female that was investigated in the study and a barplot of non-smoker and smoker who are male and female.

```
barplot(table(smoking.data$gender),
        main = "Smoking Study by Gender",
        xlab = "gender",
        ylab = "frequency",
        col = c("pink", "blue4"
          ),
        ylim = c(0,1000))
```



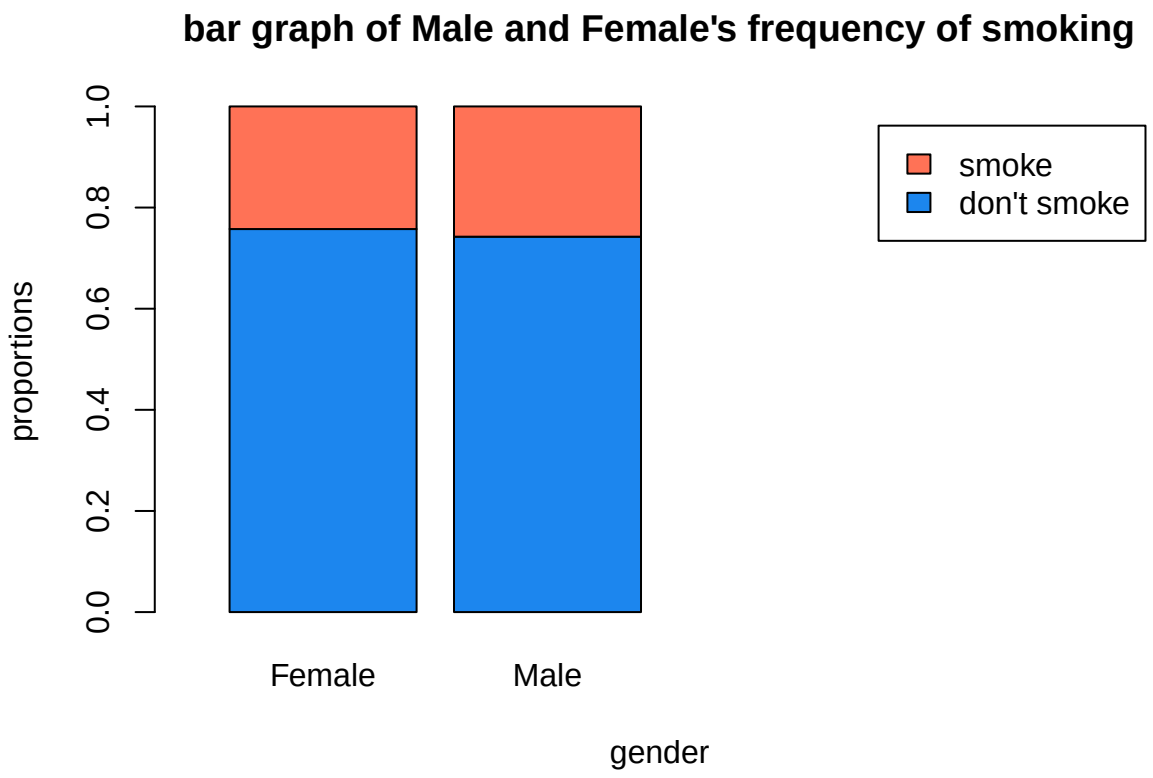
The study sampled significantly more females than males. While almost 1,000 females were surveyed, there were just under 750 males surveyed. This gives us a good estimate of the frequency of the gender of people participating in the survey. To go into more details of the smoking habits of both gender, we can create a side-by-side graph like below.

```
barplot(table(smoking.data$gender,smoking.data$smoke),
        beside = TRUE,
        main = "bar graph of Male and Female's frequency of smoking",
        xlab = "smoke or not",
        ylab = "frequency",
        col = c("dodgerblue2","coral1"),
        legend.text = c("female", "male"),
        xlim = c(0,10)
  )
```



From this graph, it is obvious that there is a higher frequency of total population who don't smoke than the total population of those who smoke. In both categories, there are more female than male, mainly because there are more data collected of those who are female. In order to understand whether smoking or not are related to gender, I will create a relative frequency stacked bar plot.

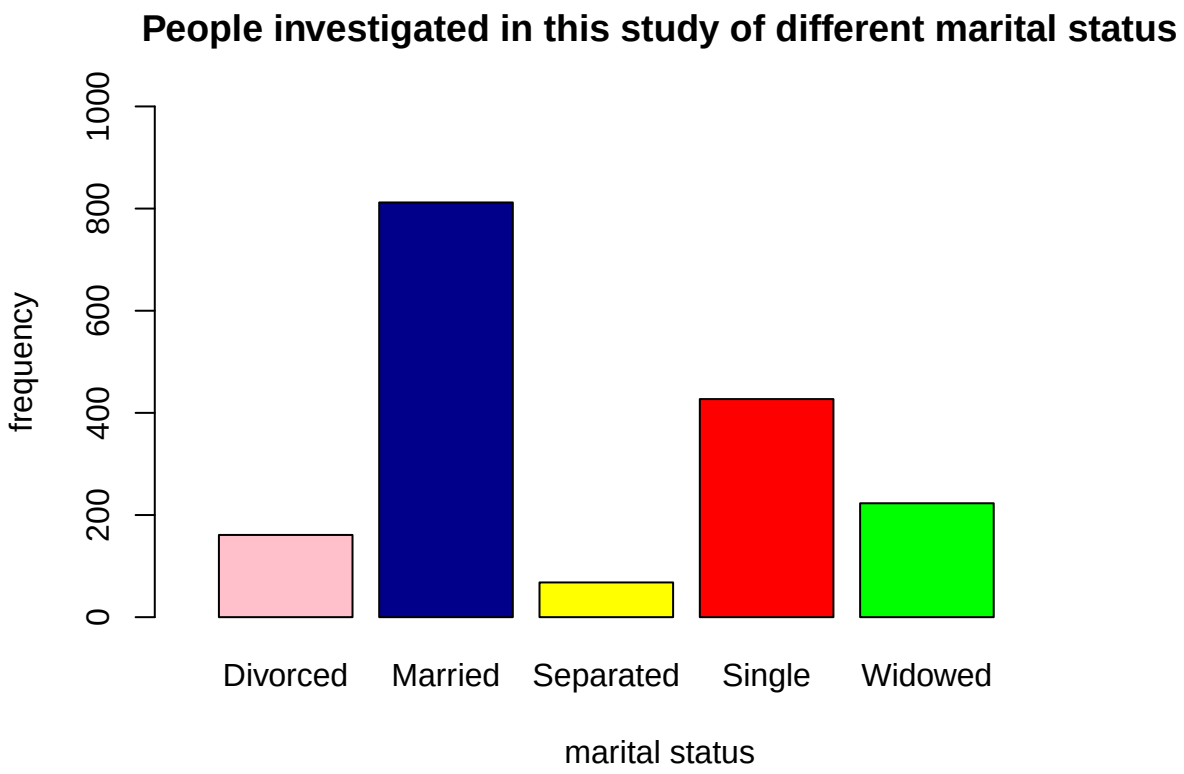
```
barplot(prop.table(table(smoking.data$smoke,smoking.data$gender),2),

        main = "bar graph of Male and Female's frequency of smoking",
        xlab = "gender",
        ylab = "proportions",
        col = c("dodgerblue2","coral1"),
        legend.text = c("don't smoke", "smoke"),
        xlim = c(0,5))
```



This graph presents us that the proportions of male and female that smoke is in fact approximately the same, with slightly more male smoking than female. The proportions give us a standardized way to compare the percentages of smokers and non-smokers by gender. From the bar graph, we can see that in both genders, the ratio of smoking to non smoking is about the same, with around 76% of both genders not smoking. The female percentage may be slightly higher. Even though in the previous graph, it appears that the proportion of female non-smokers to female smokers is bigger than that of male non-smokers to smokers, they are very close in percentage.

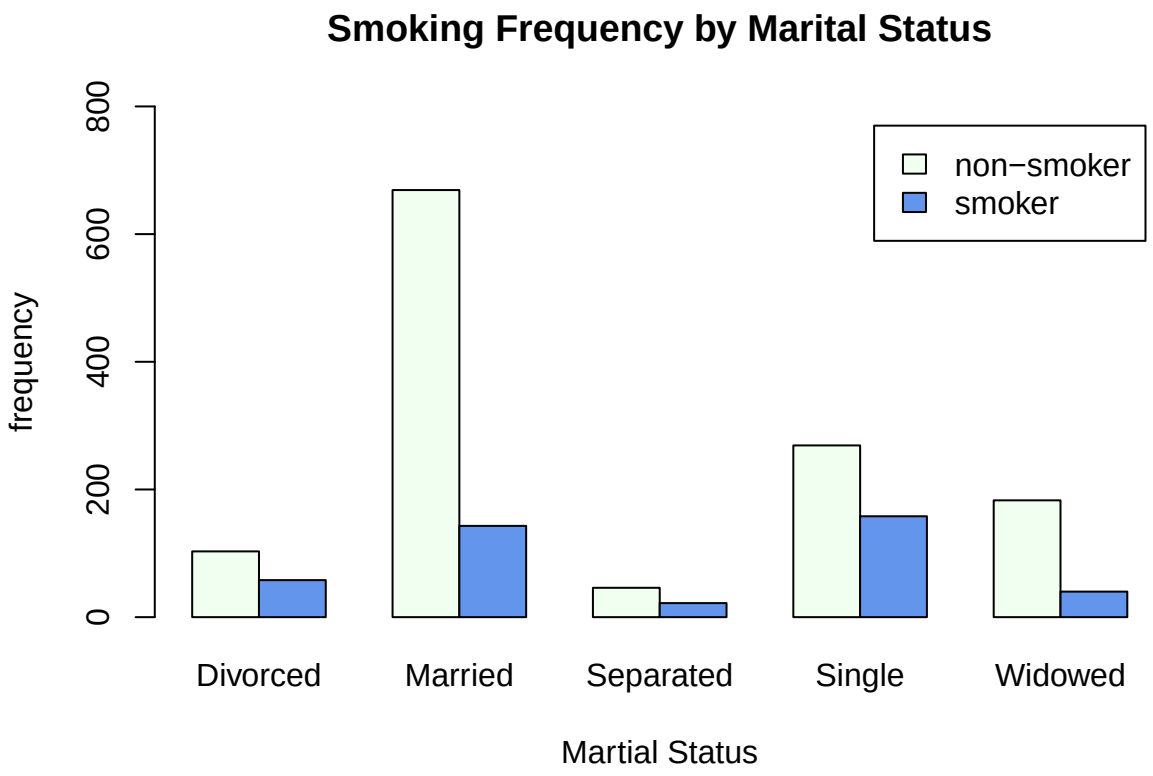
However, I also wonder if there is a relationship between smoking or not and marital status. Therefore, we made a graph first understanding the distributions of different marital status in the study, then we made a relative frequency stacked barplot so we can understand different marital status's frequency of smoking without bias of the different total number of people that has been investigated in the study.

```
barplot(table(smoking.data$marital_status),
        main = "People investigated in this study of different marital status",
        xlab = "marital status",
        ylab = "frequency",
        col = c("pink", "blue4", "yellow","red", "green"
          ),
        xlim = c(0,7),
        ylim = c(0,1000))
```



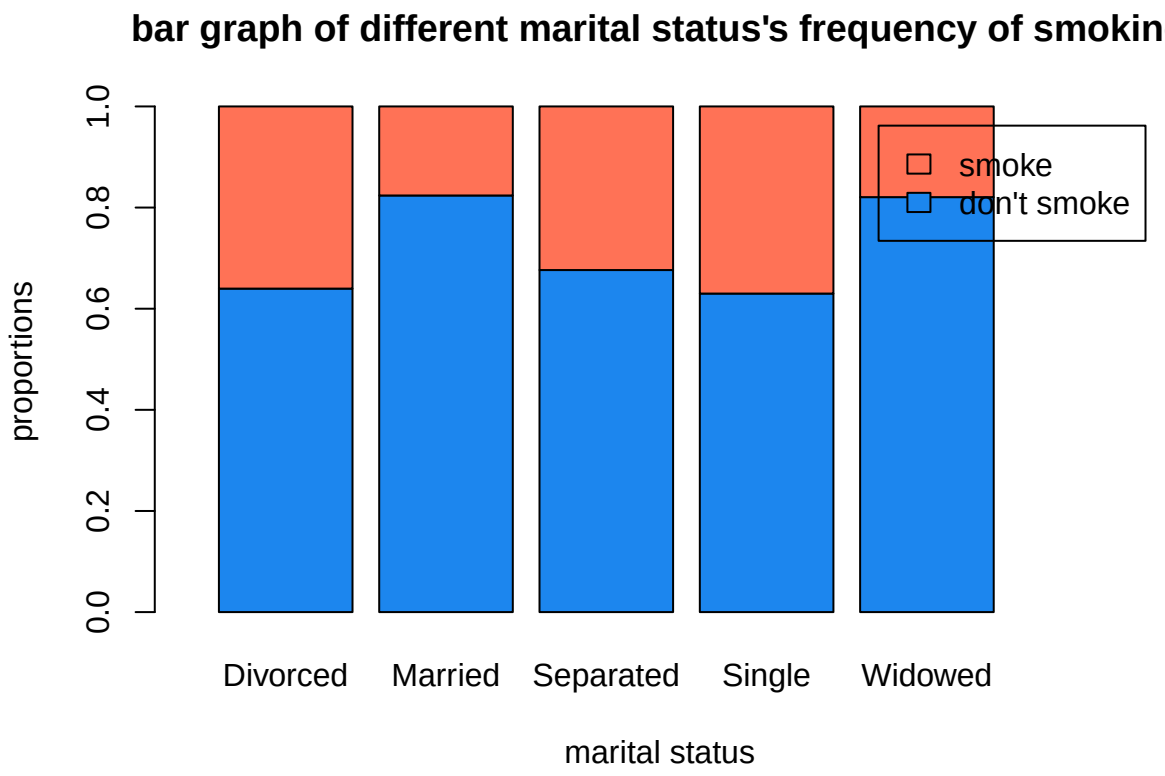
This graph shows that out of the total study, regardless of whether smoking or not smoking, there is a significant amount more of married people than any other category. Around 800 people in the study are married, with the second highest category of being single only a little over 400 people. The smallest category by far is separated, with only around 40 people.

```
barplot(table(smoking.data$smoke,smoking.data$marital_status),
        beside = TRUE,
        main = "Smoking Frequency by Marital Status",
        xlab = "Martial Status",
        ylab = "frequency",
        col = c("honeydew","cornflowerblue"),
        legend.text = c("non-smoker", "smoker"),
        ylim = c(0,800))
```



The bar graph shows us a side-by-side comparison of the frequency of smokers and non-smokers by marital status. It is hard to tell the proportions of smoking by marital status, making it hard to compare smoking percentages by marital status across the board. However, we can tell that in every single category, the frequency of non-smokers are higher than smokers, which means that there isn't necessarily one marital status that suggests that one may smoke. However, the proportions in the next graph may show a different story.

```
barplot(prop.table(table(smoking.data$smoke,smoking.data$marital_status),2),

        main = "bar graph of different marital status's frequency of smoking",
        xlab = "marital status",
        ylab = "proportions",
        col = c("dodgerblue2","coral1"),
        legend.text = c("don't smoke", "smoke"),
        xlim = c(0,7))
```



From this graph, we can observe that the proportion of those who smoke is the highest both in single and divorced group, and lowest in widowed and married groups, which we find very interesting. Around 80% of married and widowed people are not smokers, while almost 40% of single people smoke. This might suggest that single people are more likely to smoke.

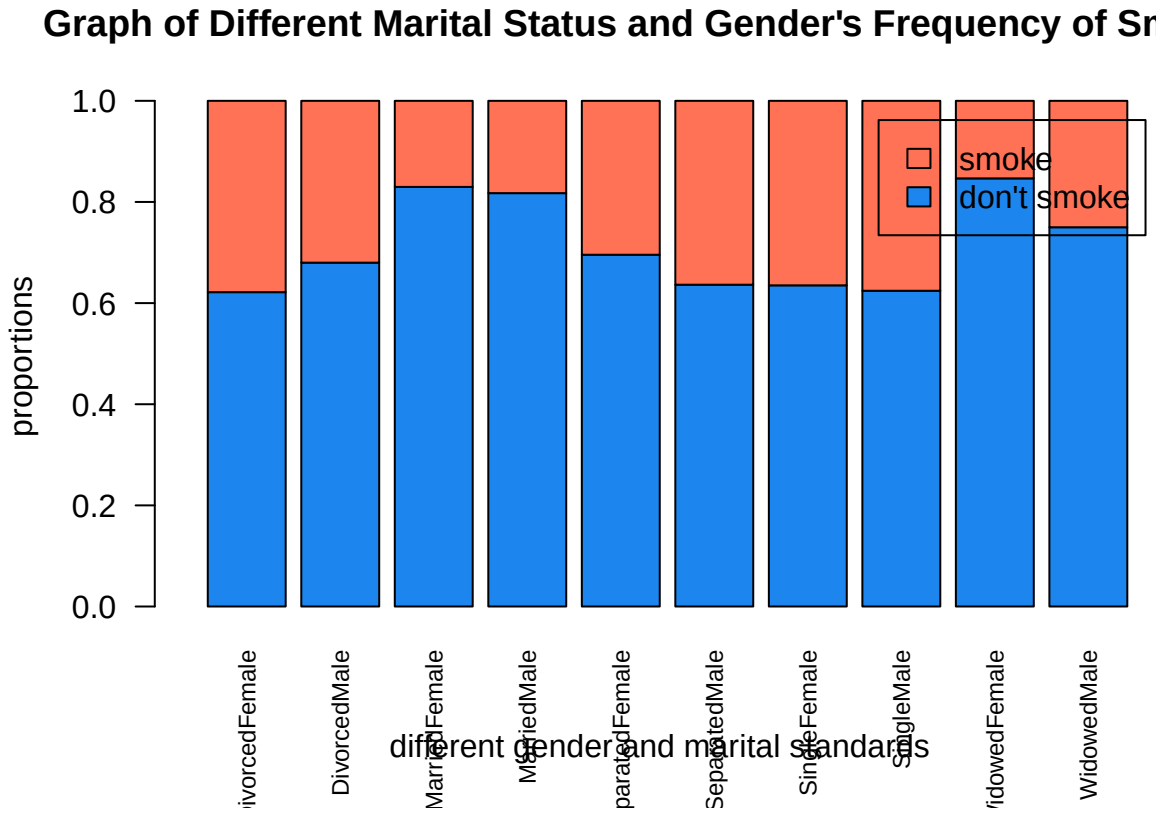
We than wondered whether it has anything to do with genders.

```
#creat new variable that combines the traits
smoking.data$marriedgender <- paste(smoking.data$marital_status, smoking.data$gender, sep="")

head(smoking.data)
```

```
##   gender age marital_status highest_qualification nationality ethnicity
## 1   Male  38       Divorced      No Qualification     British     White
## 2 Female  42         Single      No Qualification     British     White
## 3   Male  40        Married                Degree     English     White
## 4 Female  40        Married                Degree     English     White
## 5 Female  39        Married          GCSE/O Level     British     White
## 6 Female  37        Married          GCSE/O Level     British     White
##       gross_income    region smoke amt_weekends amt_weekdays    type
## 1   2,600 to 5,200 The North FALSE           NA           NA
## 2      Under 2,600 The North  TRUE           12           12 Packets
## 3 28,600 to 36,400 The North FALSE           NA           NA
## 4 10,400 to 15,600 The North FALSE           NA           NA
## 5   2,600 to 5,200 The North FALSE           NA           NA
## 6 15,600 to 20,800 The North FALSE           NA           NA
##   marriedgender
## 1  DivorcedMale
## 2  SingleFemale
## 3   MarriedMale
## 4 MarriedFemale
```

```
## 5 MarriedFemale
## 6 MarriedFemale
```

```
barplot(prop.table(table(smoking.data$smoke,smoking.data$marriedgender),2),

        main = "Graph of Different Marital Status and Gender's Frequency of Smoking",
        xlab = "different gender and marital standards",
        ylab = "proportions",
        col = c("dodgerblue2","coral1"),
        legend.text = c("don't smoke", "smoke"),
        xlim = c(0,12),
        las = 2,
        pch = 12,
        cex.names = 0.75)
```

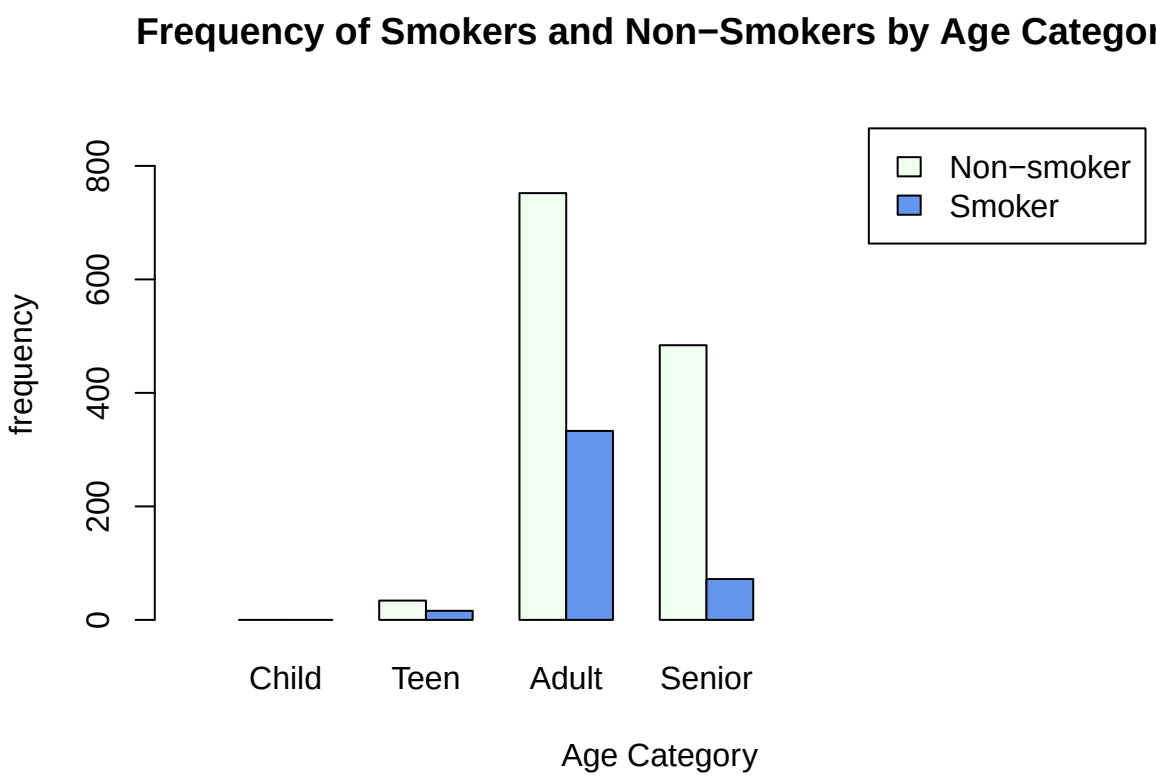


In the relative frequency stacked barplot above, we can observe that the groups with the greatest smoking proportion are divorced female, separated male, single female and single male, those with greatest smoking proportion include divorced female which is also pretty interesting. Married females and males along with female widows have the lowest smoking percentage, at around 82%. Divorced females, separated males, and single males and females have the highest smoking rate, all just above 60%. Analyzing smoking trends by two categories shows greater differences in the data, as we can see there are significant percentage differences between different categories.

We are also interested in the smoking age distribution in this data set. In order to analyze the data better, we first created a box plot of the age group.

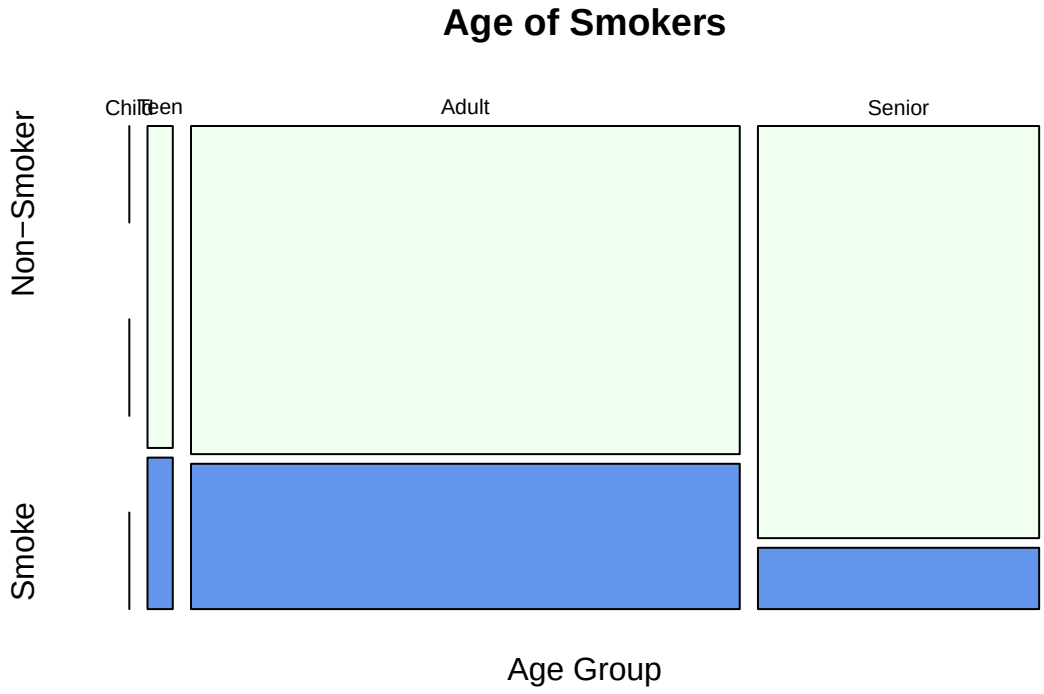
```
age.breaks <- c(0, 13, 19, 60, 120)

smoking.data$age.group <- cut(smoking.data$age,
                          breaks = age.breaks,
```

```
                          right = FALSE)


levels(smoking.data$age.group)=c("Child", "Teen", "Adult", "Senior")



barplot(table(smoking.data$smoke, smoking.data$age.group),
        beside = TRUE,
        main = "Frequency of Smokers and Non-Smokers by Age Category",
        xlab = "Age Category",
        ylab = "frequency",
        col = c("honeydew","cornflowerblue"),
        legend.text = c("Non-smoker" ,"Smoker"),
        xlim = c(0,20) ,
        ylim=c(0,900))
```



This comparative bar graph shows that in the study, there are way more adults surveyed than any other category, which makes sense, because children and teens aren't legally allowed to smoke, and seniors are probably wary of smoking because of their health. In every category, there are more non-smokers than smokers, but it is hard to tell the proportion difference from this graph. There are around 700 adult non-smokers and 300 adult smokers. These numbers are significantly higher than any other age group.

```
mosaicplot(smoking.data$age.group~smoking.data$smoke,
        main="Age of Smokers",
        col=c("honeydew","cornflowerblue"),
        xlab="Age Group",
        ylab="Smoke                        Non-Smoker",
        )
```

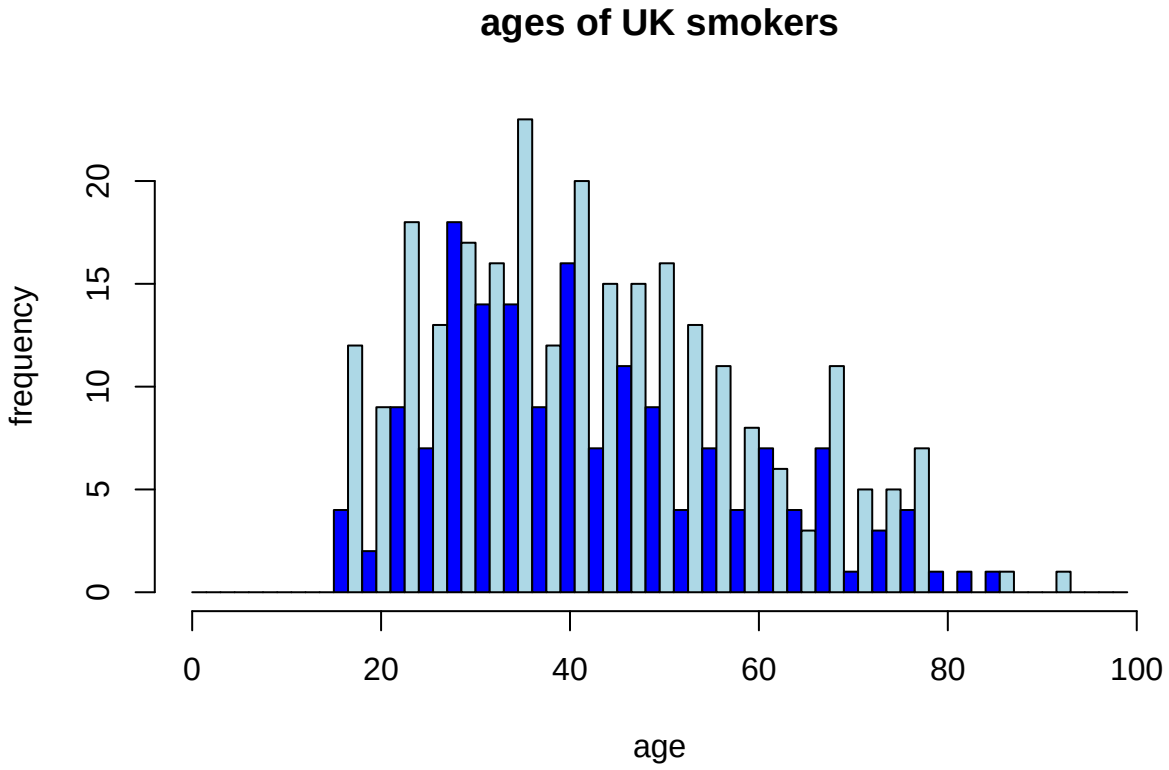


The mosiac plot gives us a good visual depiction of the proportion differences of smokers and non-smokers by age category. This graph tells us an interesting fact: the proportion of senior non-smokers is higher than adult and teens. The teen and adult smoking to non-smoking ratio is about the same.

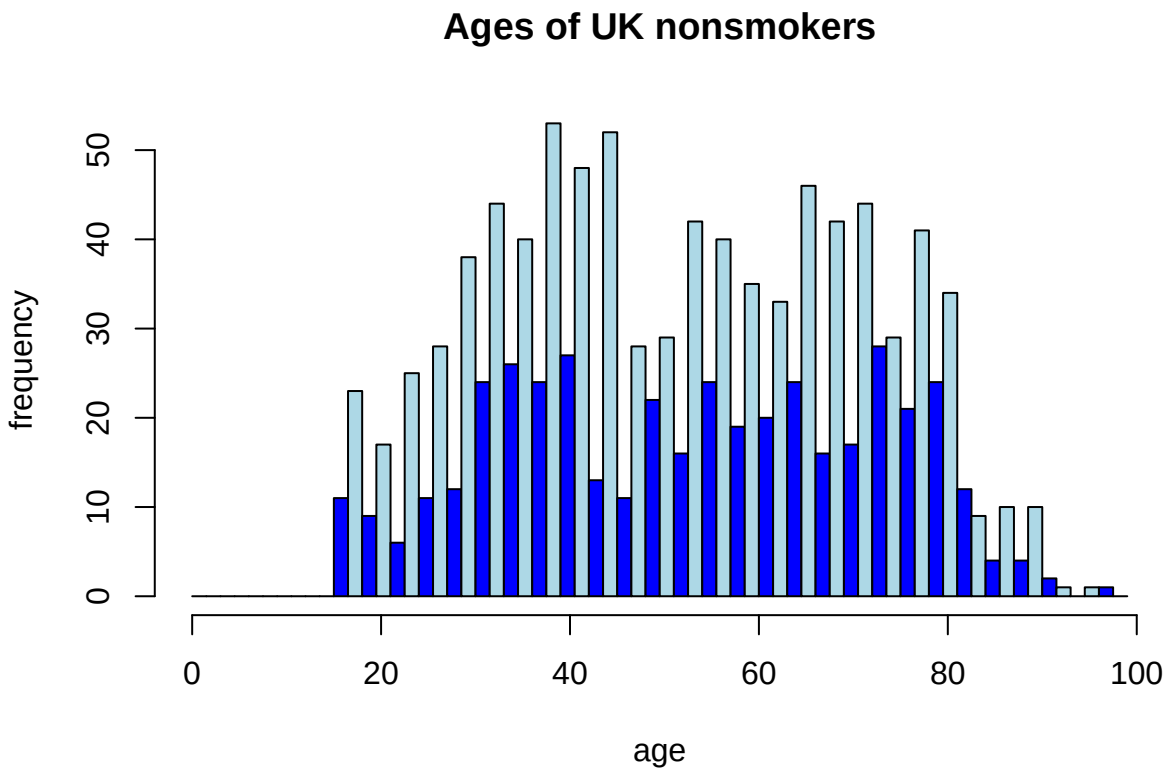
```
boxplot(smoking.data$age~smoking.data$smoke,
        main = "Average age of smokers and non-smokers",
        ylab = "Smoking age",
        xlab = "Smoker or not",
        cex = 1,
        pch = 20,
        col = "red",
        outcol = "blue")
```

This boxplot shows that the medium age of the non-smokers is slightly higher than the medium age of smokers.In addition, the maximum of age of non-smokers, around 97, is a lot larger than the maximum age of smokers, which is a outlier of around 92. We can possibly infer from the the boxplot that this is probably because people who smoke live shorter. For more details of the trends of the age for both smokers and non-smokers, we decided to use histograms to compare the data.

```
#make a smokers only data frame
smokingage.smokers <- smoking.data[ smoking.data$smoke == "TRUE", ]

hist(smokingage.smokers$age,
     main = "ages of UK smokers",
     xlab = "age",
     ylab = "frequency",
    col = c("blue", "lightblue"),
     breaks=seq(from=0, to=100, by=1.5))
```




```
#make a nonsmoker only data frame
smokingage.nonsmokers <- smoking.data[ smoking.data$smoke == "FALSE", ]

hist(smokingage.nonsmokers$age,
     main = "Ages of UK nonsmokers",
    xlab = "age",
    ylab = "frequency",
  col = c("blue", "lightblue"),
     breaks=seq(from=0, to=100, by=1.5))
```



From these two histograms, we can observe that the graph of UK smokers shows a trend that skews right, while the graph of UK nonsmokers shows a trend which skews left. This that probably there are more young people that have the habit of smoking, or that most smokers pass away relative younger than non-smokers.

Conclusion

The data set showed results that we pretty much expected in every category except for marital status. There are way more non-smokers than smokers, but the proportions of female smokers to non-smokers and male smokers to male non-smokers were about the same. We were not surprised that divorced and married people smoke the most, but when combined with gender, we were shocked by how divorced female, separated male, single female and single male smoke the most. The fact that the majority of people who smoked in this study were adults made sense, as well as the fact that among the seniors who participated in the study, a greater proportion than adults and teens did not smoke than smoked. It was also interesting that the age of U.K smokers skewed to the right while U.K non-smokers skewed to the left. It can be due to the ages of people who smoke are generally younger, or that most smokers live shorter lives than non-smokers, or simply that more young smokers were investigated in the study.