STA 100
Homework 2
Due: Monday, Oct 16^{th}

Book Homework

These prompts correspond to "Book Homework" portion of the homework on Canvas. You turn in the answers to these questions online.

You must show proper probability notation for each problem such as $P(A)$ or $\hat{P}(A)$. For notation such as $P(A \cap B)$, writing out words such as $P(A$ and $B)$ is suffice, especially if you are typing up solutions.

1. Consider the following contingency (frequency) table, in which two species of mice were tested for a specific parasite:

	Notation:	I	I^c	
Notation		Infected	Not Infected	Total
S_1	Species 1	38	16	54
S_2	Species 2	20	35	55
		58	51	109

   (a) Estimate the probability that a randomly selected mouse was species 1. $\hat{P}(S_1) = \frac{54}{109}$

   (b) Estimate the probability that a randomly selected mouse was infected. $\hat{P}(I) = \frac{58}{109}$

   (c) Estimate the probability that a randomly selected mouse was both infected and species 1. $\hat{P}(I \cap S_1) = \frac{38}{109}$

   (d) Estimate the probability that a randomly selected mouse was not infected and species 2. $\hat{P}(I^c \cap S_2) = \frac{35}{109}$

2. Continue with the data from Problem 1.

   (a) If a mouse was species 1, what is the estimated probability they were infected? $\hat{P}(I | S_1) = \frac{P(I \cap S_1)}{P(S_1)} = \frac{38}{54}$

   (b) If a mouse was species 2, what is the estimated probability they were infected? $\hat{P}(I | S_2) = \frac{P(I \cap S_2)}{P(S_2)} = \frac{20}{55}$

   (c) What is the estimated probability that an infected mouse was species 1? $\hat{P}(S_1 | I) = \frac{P(S_1 \cap I)}{P(I)} = \frac{38}{58}$

   (d) What is the estimated probability that an infected mouse was species 2? $\hat{P}(S_2 | I) = \frac{P(S_2 \cap I)}{P(I)} = \frac{20}{58}$

   (e) Are the events that a mouse is species 1 and a mouse was infected independent? $\hat{P}(S_1 \cap I) = \hat{P}(S_1) \cdot \hat{P}(I) \longrightarrow \left(\frac{54}{109}\right)\left(\frac{58}{109}\right) \neq \frac{38}{109}$ not independent

3. For a particular disease, the probability of the disease is 0.04. If someone has the disease, the probability they test positive is 0.95. If they do not have the disease, the probability they test negative is 0.99.

   (a) Estimate the probability someone both tests positive and has the disease.

   (b) Estimate the probability that someone tests positive.

   (c) Estimate the probability that if someone tested positive, they have the disease.

   (d) Estimate the probability that if someone tests negative, they do not have the disease.

4. Answer the following questions with TRUE or FALSE. Explain your answers with a sentence or two, or you may find it helpful to draw a Venn diagram to demonstrate your answer.

   (a) The intersection of two events A and B can be larger than the union of the same two events A and B.

   (b) The probability of a single event A must be smaller than or equal to the union of two events A and B.

   (c) The condition probability of A given B must be smaller than the intersection of the same two events A and B

   (d) If two events are independent, that means that $Pr(A \cup B) = Pr(A \cap B)$

For a particular disease, the probability of the disease is 0.04. If someone has the disease, the probability they test positive is 0.95. If they do not have the disease, the probability they test negative is 0.99.

$P(D+) = 0.04$

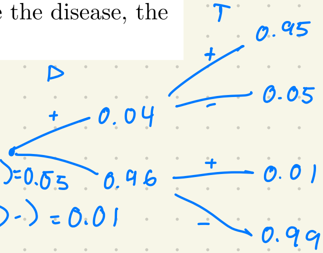
$P(D-) = 0.96$

$P(T+ \mid D+) = 0.95$

$P(T- \mid D-) = 0.99$

$P(T- \mid D+) = 0.05$

$P(T+ \mid D-) = 0.01$



(a) Estimate the probability someone both tests positive and has the disease.

$$P(T+ \cap D+) = P(T+ \mid D+) \cdot P(D+) = 0.95 \cdot 0.04 = 0.038$$

(b) Estimate the probability that someone tests positive.

$$P(T+) = P(T+ \cap D+) + P(T+ \cap D-) = P(T+ \mid D+) \cdot P(D+) + P(T+ \mid D-) \cdot P(D-)$$
$$= (0.95 \cdot 0.04) + (0.01)(0.96)$$
$$= 0.0476$$

(c) Estimate the probability that if someone tested positive, they have the disease.

$$P(D+ \mid T+) = \frac{P(D+ \cap T+)}{P(T+)} = \frac{P(T+ \mid D+) \cdot P(D+)}{0.0476} = \frac{(0.95)(0.04)}{0.0476}$$
$$= 0.798$$

(d) Estimate the probability that if someone tests negative, they do not have the disease.

$$P(D- \mid T-) = \frac{P(T- \mid D-)\, P(D-)}{P(T- \mid D-) \cdot P(D-) + P(T- \mid D+) \cdot P(D+)} = \frac{(0.99)(0.96)}{(0.99)(0.96) + (0.05)(0.04)}$$

$$= 0.998$$

4. Answer the following questions with TRUE or FALSE. Explain your answers with a sentence or two, or you may find it helpful to draw a Venn diagram to demonstrate your answer.

F (a) The intersection of two events A and B can be larger than the union of the same two events A and B.

T (b) The probability of a single event A must be smaller than or equal to the union of two events A and B.

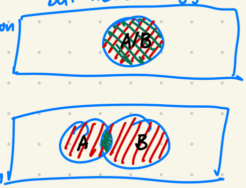
F (c) The condition probability of A given B must be smaller than the intersection of the same two events A and B

F (d) If two events are independent, that means that $Pr(A \cup B) = Pr(A \cap B)$

(a) False Best it can be is equal but never bigger.

intersect = union

intersect < union



= Union

= Intersect

(b) True b/c of

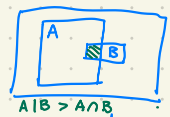
$$P(A \cup B) = \underset{P(A)}{P(A) + P(B)} - P(A \cap B)$$

never negative

always $[0, P(B)/P(A)]$

(c)

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$



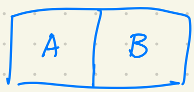
False because $P(B) \leq 1$ and that means denominator can decrease making probability of $P(A|B)$ increase. And possibly over take $P(A \cap B)$ which has 1 as it's denominator.

(d) False

$P(A|B) = P(A)$

$P(B|A) = P(B)$

$P(A \cap B) = P(A) \cdot P(B)$



$P(A \cup B) = 1$

$P(A \cap B) = \emptyset$

R Homework

These prompts correspond to "R Portion" of the homeworks on Canvas. You use R to find the answers to the following questions, and submit your answers online.

I. You will be working with the dataset `colors.csv`, which has the following columns:

Column 1: `Eye`: The eye color of the subject

Column 2: `Sex`: The hair color of the subject

Column 3: `GPA`: The college GPA of the subject

(a) Plot a barplot of the eye color of the subjects. What color is the least common?

(b) Plot a side-by-side barplot of the subjects, using sex and eye color. Comparing men and women, who has a higher probability of brown eyes? Be sure to choose the grouping of the bars that make it easier to read and interpret the result.

(c) Plot a histogram of GPA. What is the most common interval of GPA?

(d) Plot a side-by-side boxplot of GPA by eye color. Which eye color has the highest minimum?

(e) Refer to the previous side-by-side boxplot of GPA by eye color. Which eye color has the highest 25^{th} percentile?

(f) Refer to the previous side-by-side boxplot of GPA by eye color. Which eye color has the most outliers?

sta_100_hw_2

Tiffany Chan

2023-10-10

R Markdown

```
getwd()
```

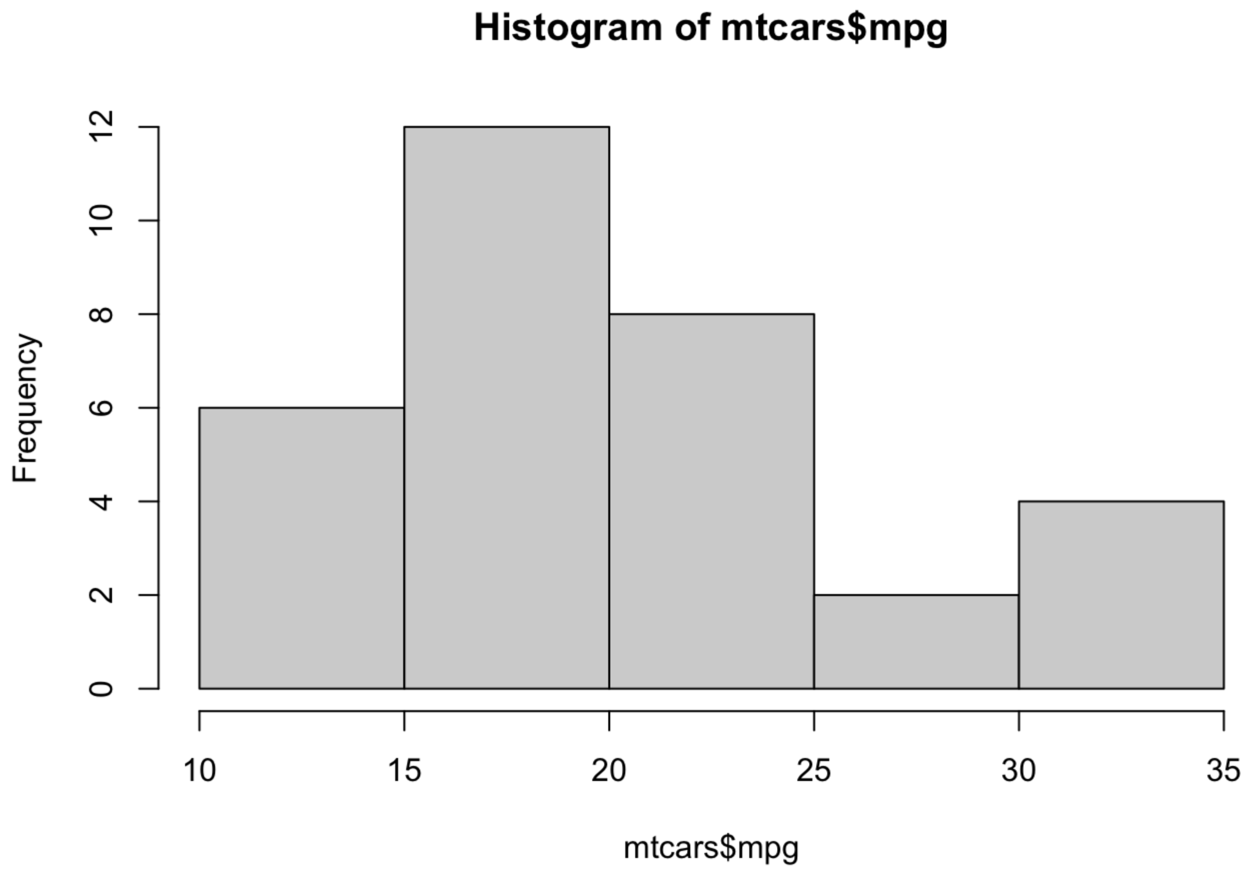
```
## [1] "/Users/tiffanysmacbookpro/Desktop/sta100F2023_tchan"
```

```
gender = rep(c("female","male"), c(1835, 2691))
admitted = rep(c("yes", "no", "yes", "no"), c(557, 1278, 1198, 1493))
dept = rep(c("A", "B", "C", "D", "E", "F", "A", "B", "C", "D", "E", "F"),
           c(89, 17, 202, 131, 94, 24, 19, 8, 391, 244, 299, 317))
dept2 = rep(c("A", "B", "C", "D", "E", "F", "A", "B", "C", "D", "E", "F"),
            c(512, 353, 120, 138, 53, 22, 313, 207, 205, 279, 138, 351))
department = c(dept, dept2)
ucb = data.frame(gender, admitted,department)
rm(gender, admitted, dept, dept2, department)
```

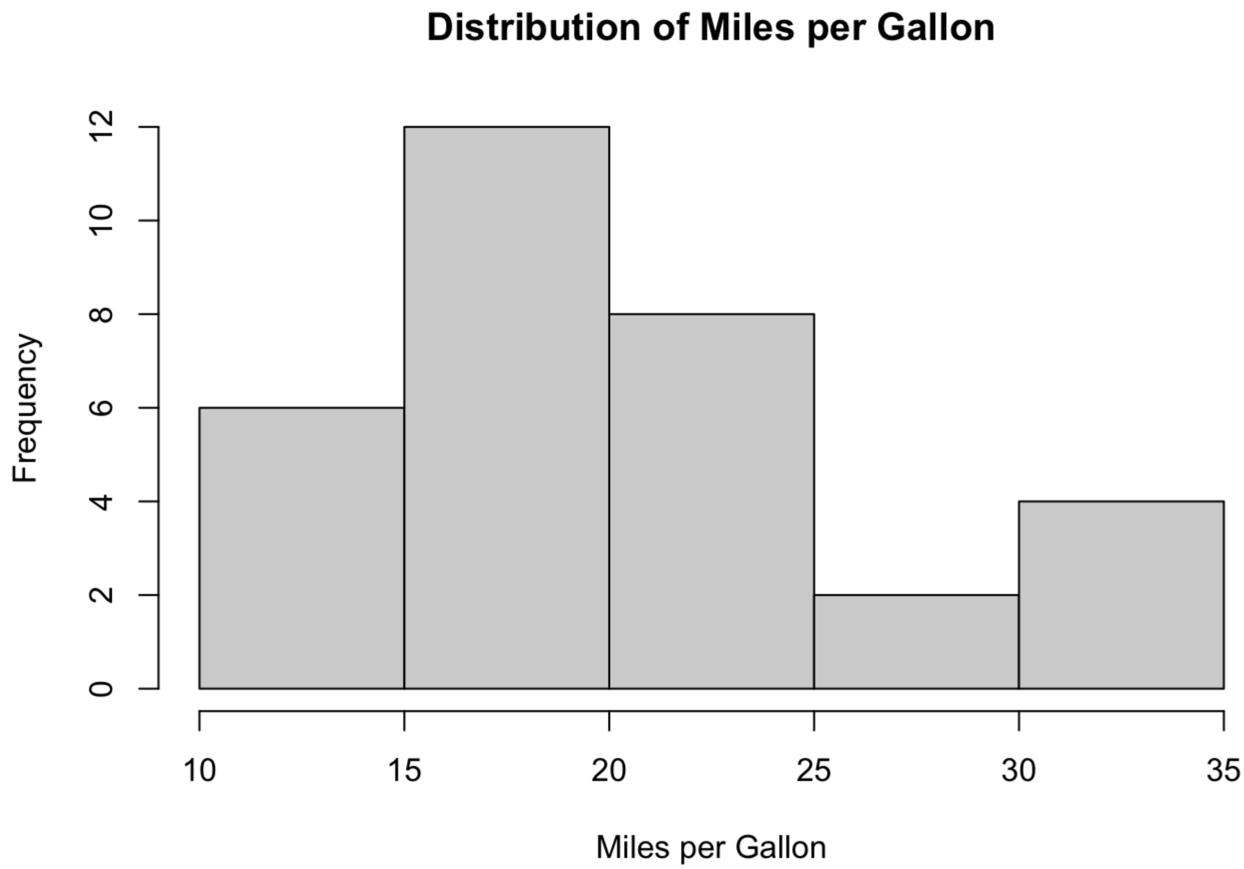
```
head(ucb)
```

```
##   gender admitted department
## 1 female      yes          A
## 2 female      yes          A
## 3 female      yes          A
## 4 female      yes          A
## 5 female      yes          A
## 6 female      yes          A
```

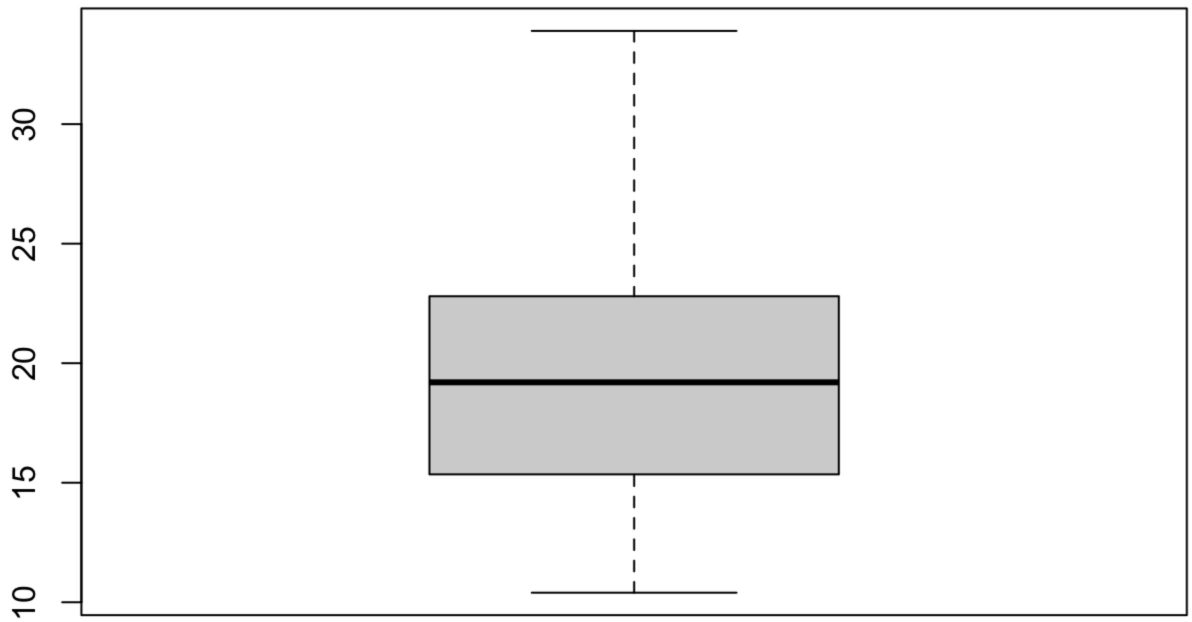
```
hist(mtcars$mpg)
```



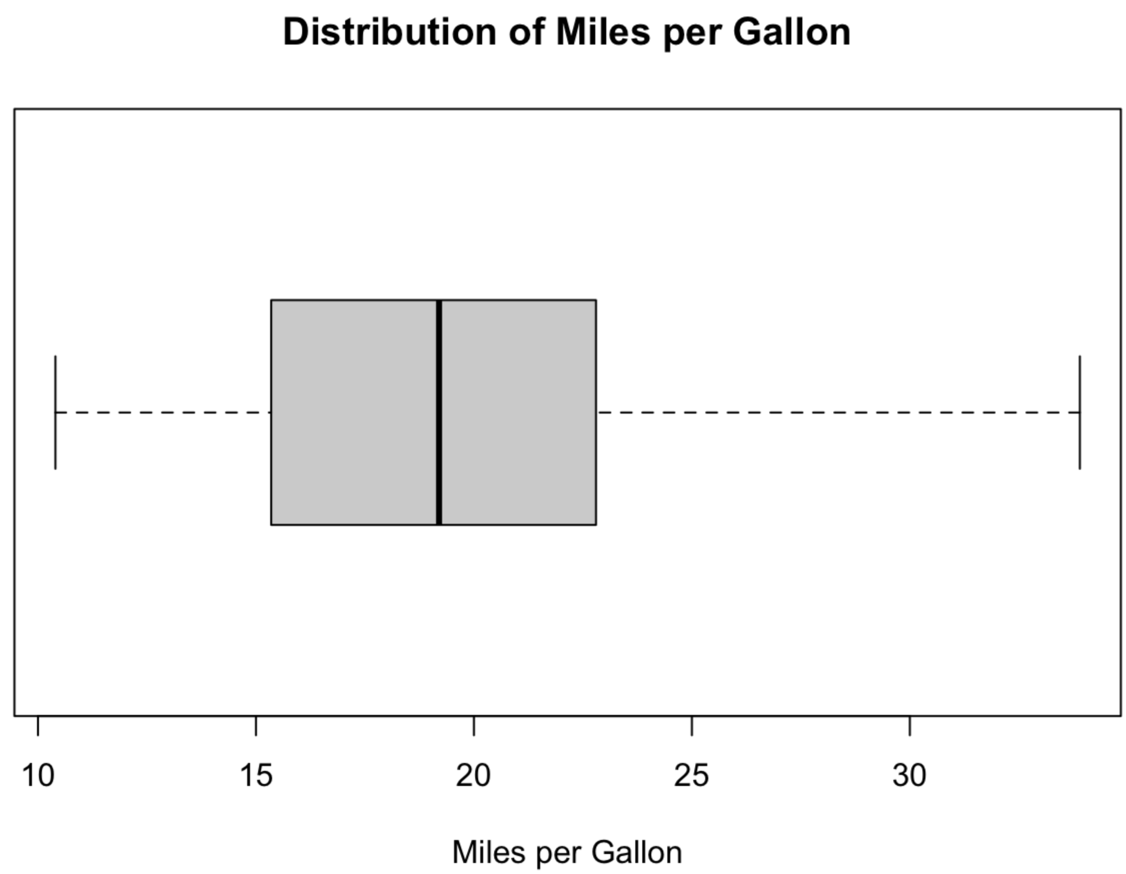

```
hist(mtcars$mpg, xlab = "Miles per Gallon", main = "Distribution of Miles per Gallon" )
```




```
boxplot(mtcars$mpg)
```




```
boxplot(mtcars$mpg, xlab="Miles per Gallon", main="Distribution of Miles per Gallon", ho
rizontal = TRUE)
```



You will be working with the dataset colors.csv, which has the following columns: Column 1: Eye: The eye color of the subject Column 2: Sex: The hair color of the subject Column 3: GPA: The college GPA of the subject

```
colors<-read.csv("/Users/tiffanysmacbookpro/Desktop/sta100F2023_tchan/data/colors.csv")
```

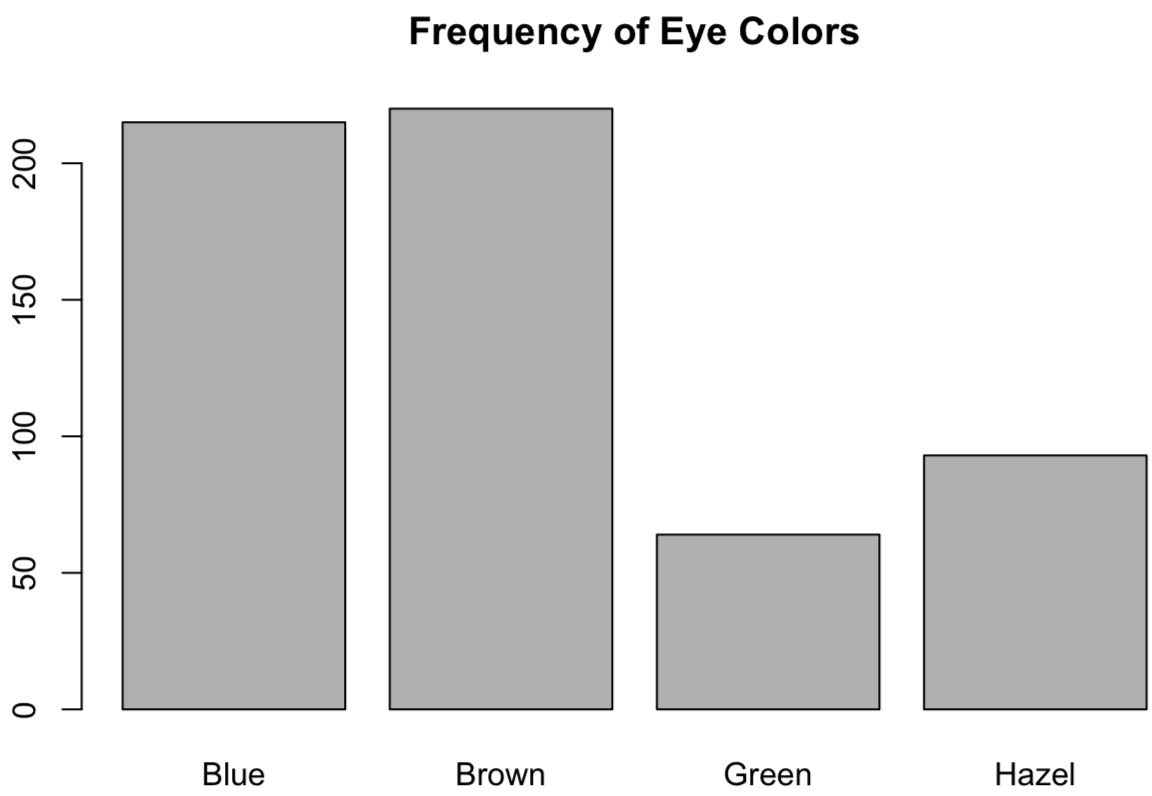
```
names(colors)
```

```
## [1] "Eye" "Sex" "GPA"
```

```
#(a) Plot a barplot of the eye color of the subjects. What
#color is the least common?
```

```
eye.table=table(colors$Eye)
```

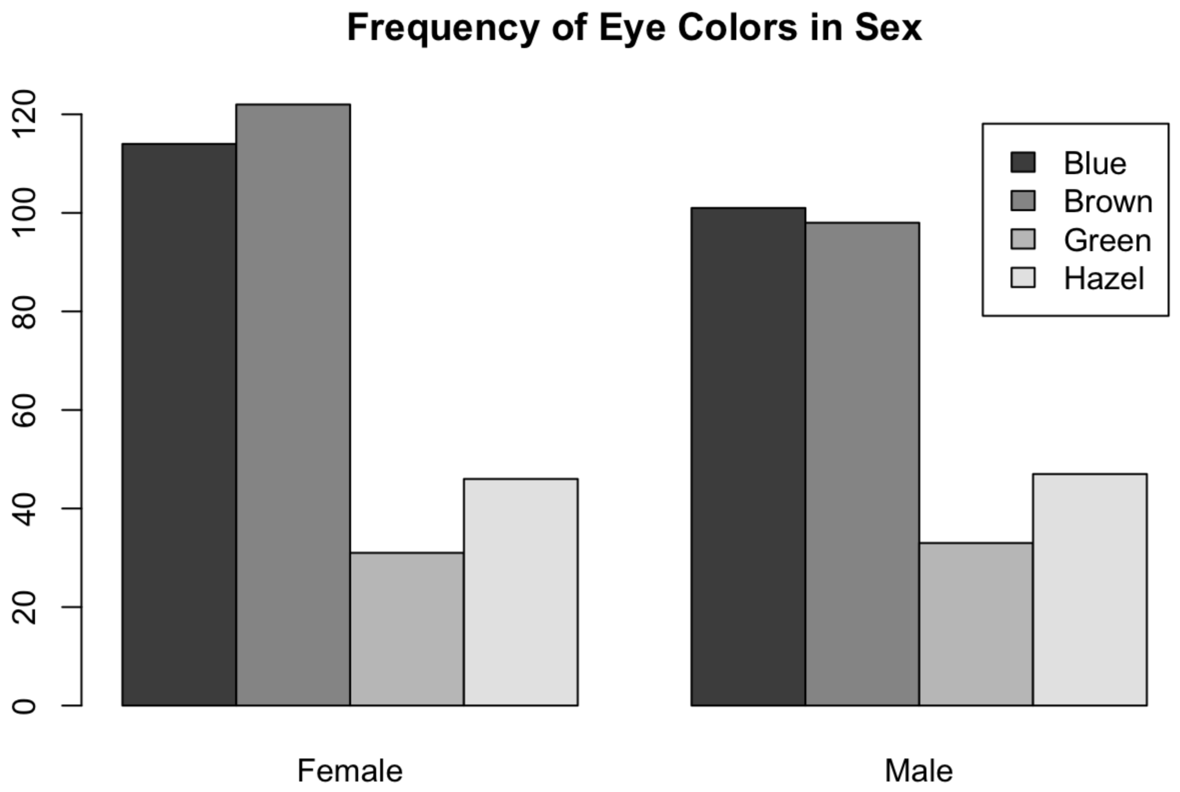
```
barplot(eye.table, main="Frequency of Eye Colors")
```




```
#Least common eye color is green eyes.
```

```
#(b) Plot a side-by-side barplot of the subjects, using sex
#and eye color. Comparing men and women, who
#has a higher probability of brown eyes? Be sure to
#choose the grouping of the bars that make it easier
#to read and interpret the result.
```

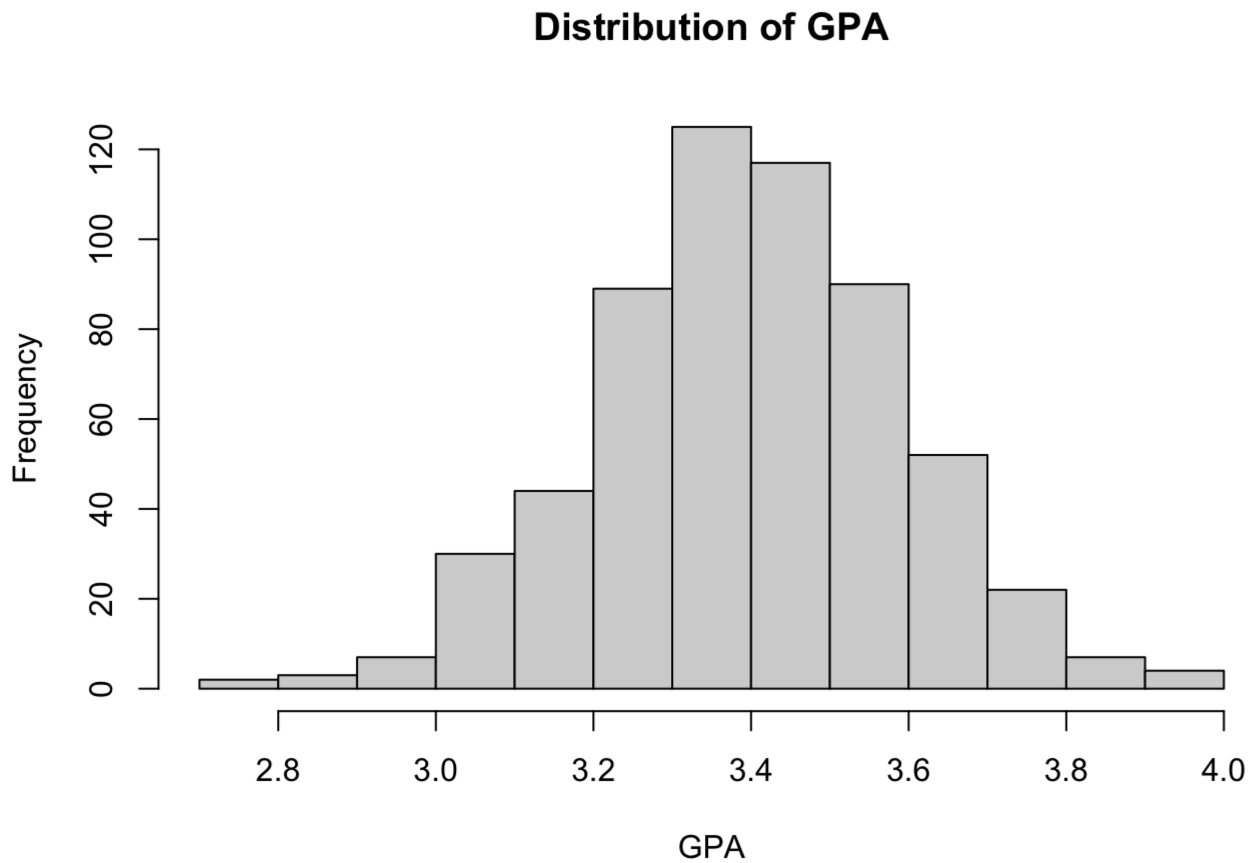
```
eye.sex.table=table(colors$Eye,colors$Sex)
barplot(eye.sex.table, main="Frequency of Eye Colors in Sex", beside = TRUE, legend=row.
names(eye.sex.table))
```




```
#Females had a higher probability of having brown eyes
```

```
#(c) Plot a histogram of GPA. What is the most common
#interval of GPA?
```

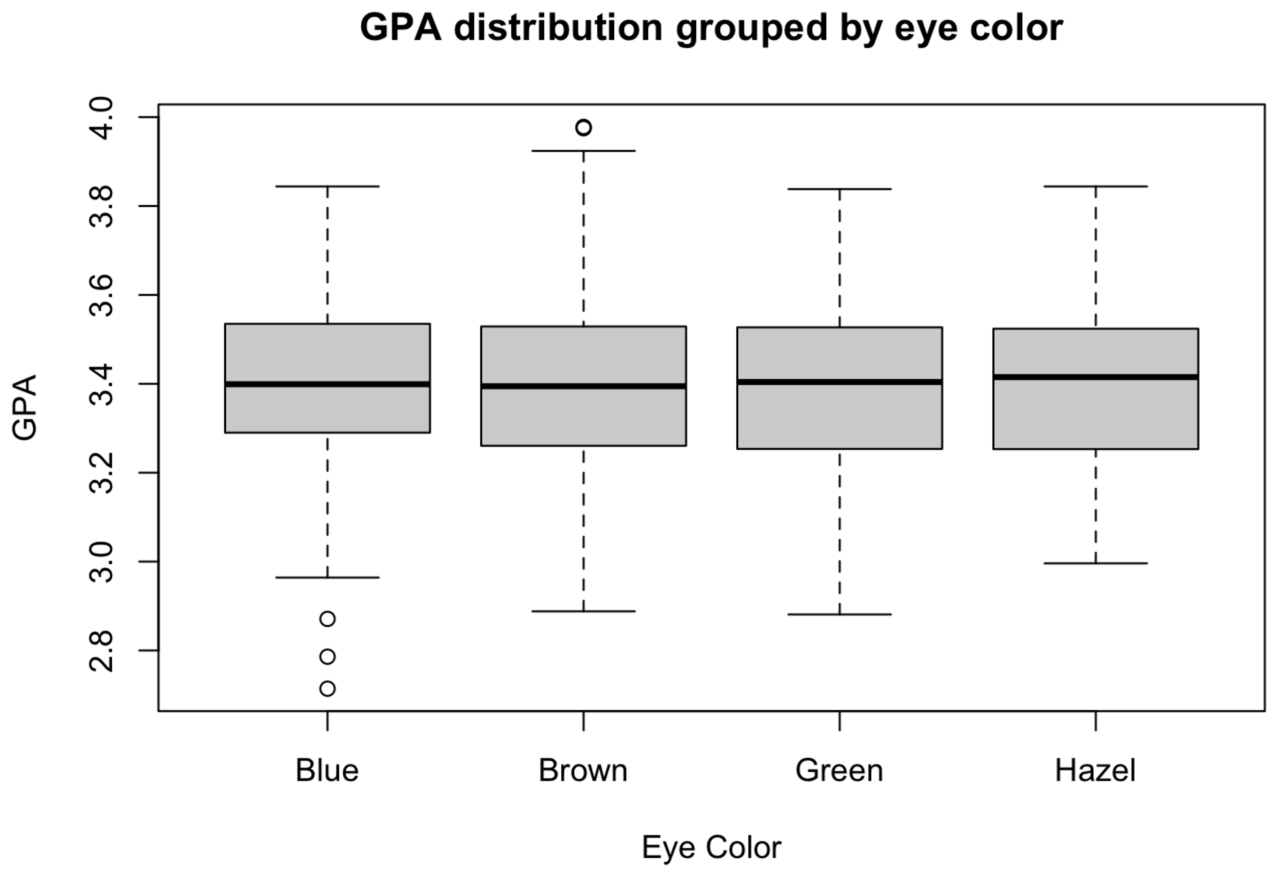
```
hist(colors$GPA, main="Distribution of GPA", xlab = "GPA")
```




```
#Most common interval ranged from 3.3 to 3.4
```

```
#(d) Plot a side-by-side boxplot of GPA by eye color.
#Which eye color has the highest minimum?
```

```
boxplot(GPA ~ Eye, data=colors, xlab = "Eye Color", main = "GPA distribution grouped by
eye color")
```




```
#Hazel eye color had the highest minimum.
```

```
#(e) Refer to the previous side-by-side boxplot of GPA
#by eye color. Which eye color has the highest 25th
#percentile?
```

```
#Blue eyes had the highest 25th percentile.
```

```
#(f) Refer to the previous side-by-side boxplot of GPA
#by eye color. Which eye color has the most outliers?
```

```
#Blue eyes had the most outliers.
```