Homework #4

Carson Crenshaw (cgc8gdt)

Insert packages

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.3      v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(gcookbook)
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
# Download Hmisc package
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

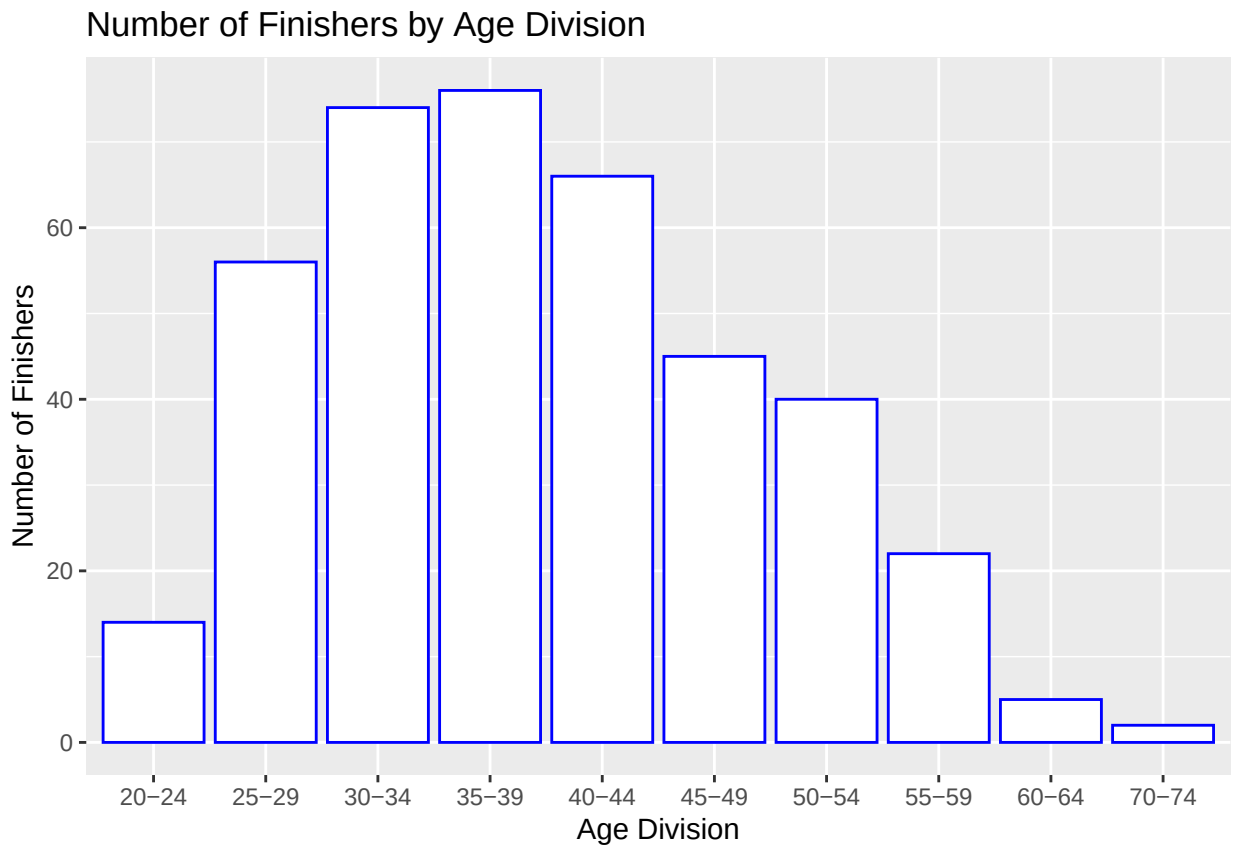
```
##
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

Problem 1

Part a

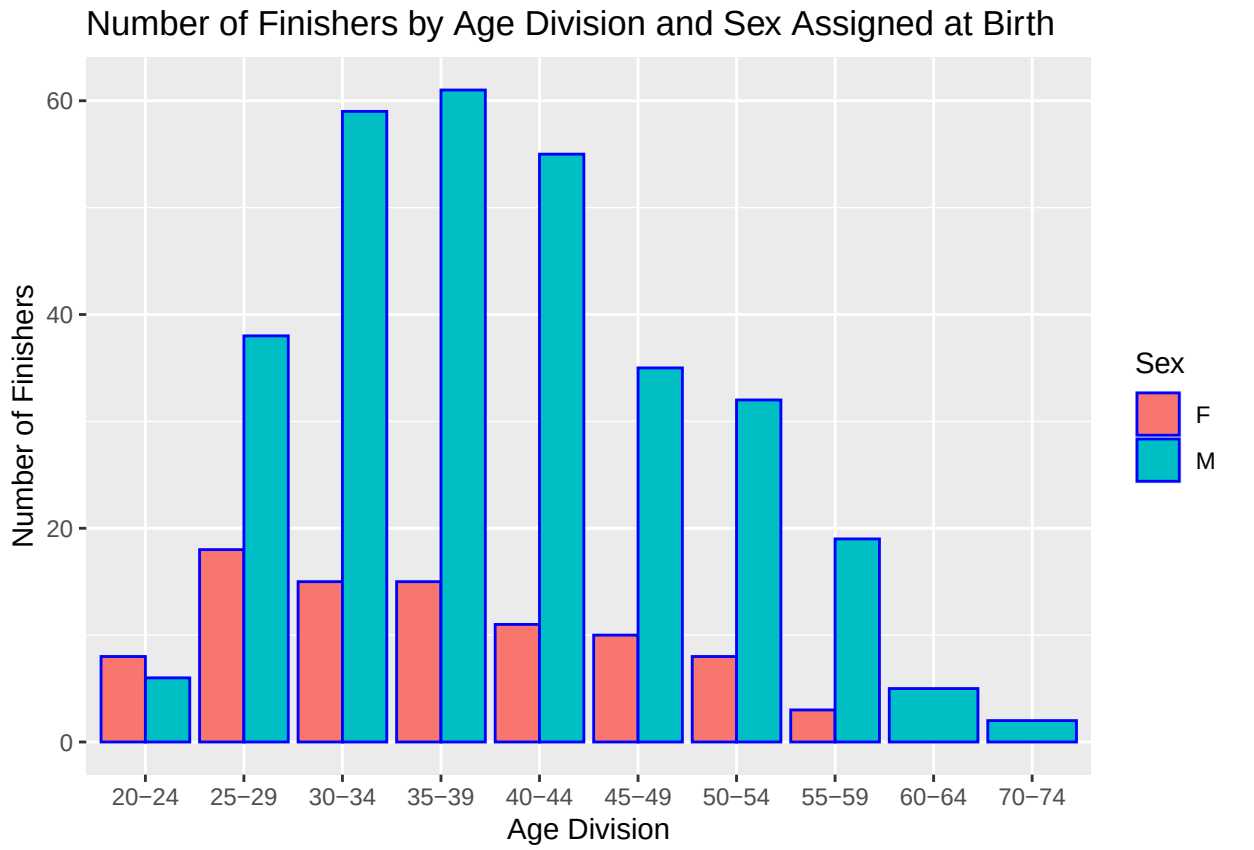
```
# Load the dataset; assign the dataframe to object called 'nym2019'
nym2019 <- read.table("nym2019.txt", header=TRUE)

ggplot(nym2019, aes(x=DivAge)) + geom_bar(fill="white", color="blue") +
  labs(title="Number of Finishers by Age Division",
                     x="Age Division", y="Number of Finishers")
```



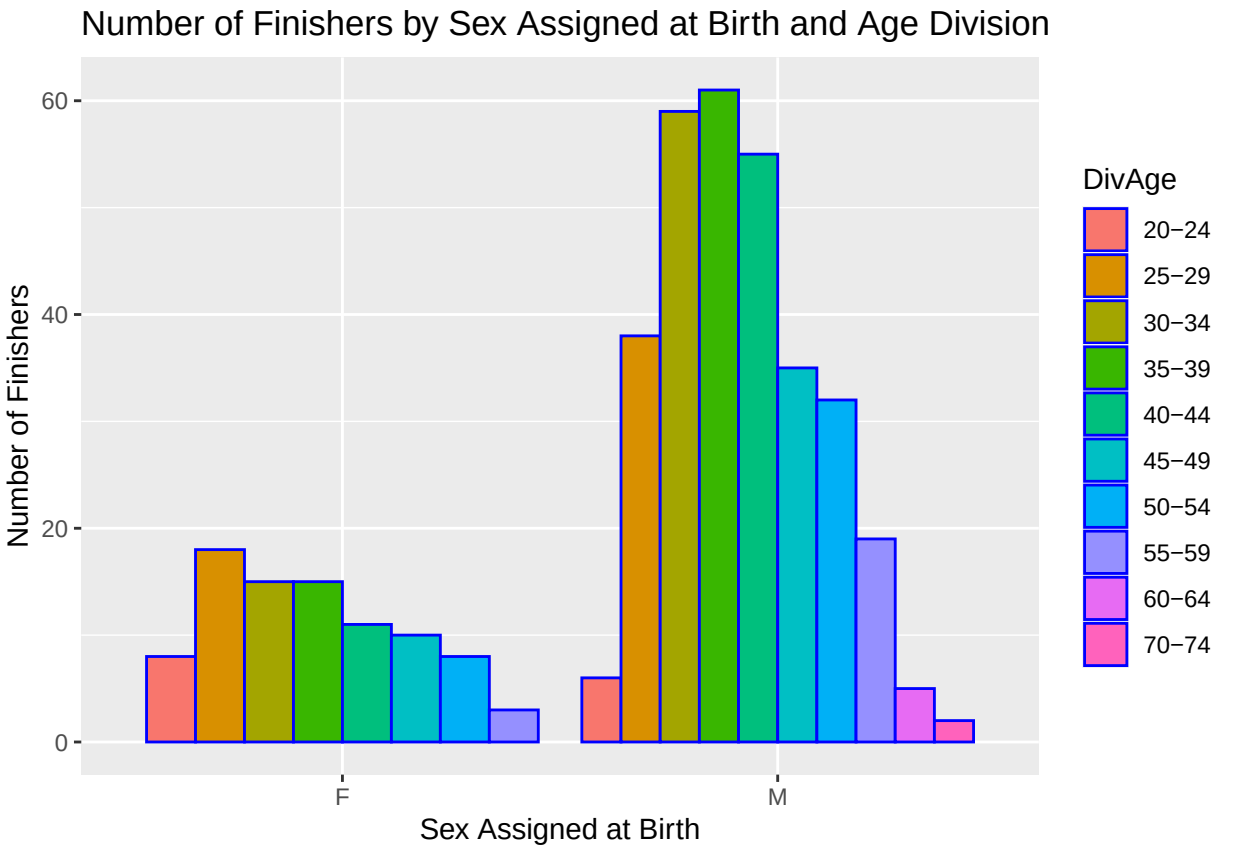
Part b

```
ggplot(nym2019, aes(x=DivAge, fill=Sex)) + geom_bar(color="blue",
                                                    position = 'dodge') +
  labs(title="Number of Finishers by Age Division and Sex Assigned at Birth",
                      x="Age Division", y="Number of Finishers")
```



Part c

```
ggplot(nym2019, aes(x=Sex, fill=DivAge)) + geom_bar(color="blue",
                                                    position = 'dodge') +
  labs(title="Number of Finishers by Sex Assigned at Birth and Age Division",
                       x="Sex Assigned at Birth", y="Number of Finishers")
```



Part d

The most basic conclusion derived from the plots created in parts(b) and (c) is that it is more common to see male than female marathon finishers. There seems to be a strong relationship between being a man and finishing a marathon. In other words, male marathon finishers vastly outnumber female marathon finishers. There is a considerable difference across all age groups between men and women except for the youngest (20-24) division. There is also a total lack of female finishers in the 60+ age groups.

When considering the number of finishers within the age divisions, however, one can conclude that marathon runners are often individuals who have considerable running experience. For both sexes, the distributions of finishers are right-skewed. The average age of marathon finishers are therefore older than one might have expected. Although the average female

marathon finisher (early-30s) tends to be slightly younger than the male counterpart (late-30s), the existence of older finishers skews the data.

The comparison between bar charts help illustrate the differences between groups. The previous plots ultimately show that there are considerable differences in spread, center, and shape between male and female marathon finishers. The female marathon finishers have a smaller spread, a younger center, and a more compact shape.

Part e

Despite the graph now showing the proportion of finishers of each sex assigned at birth across the age divisions relative to the overall number of finishers of each sex, most of my conclusions remain the same. The shape, spread, and center between the two sexes with respect to age divisions are still different, but they are considerably closer with respect to each other than the previous two graphs in part (b) and (c).

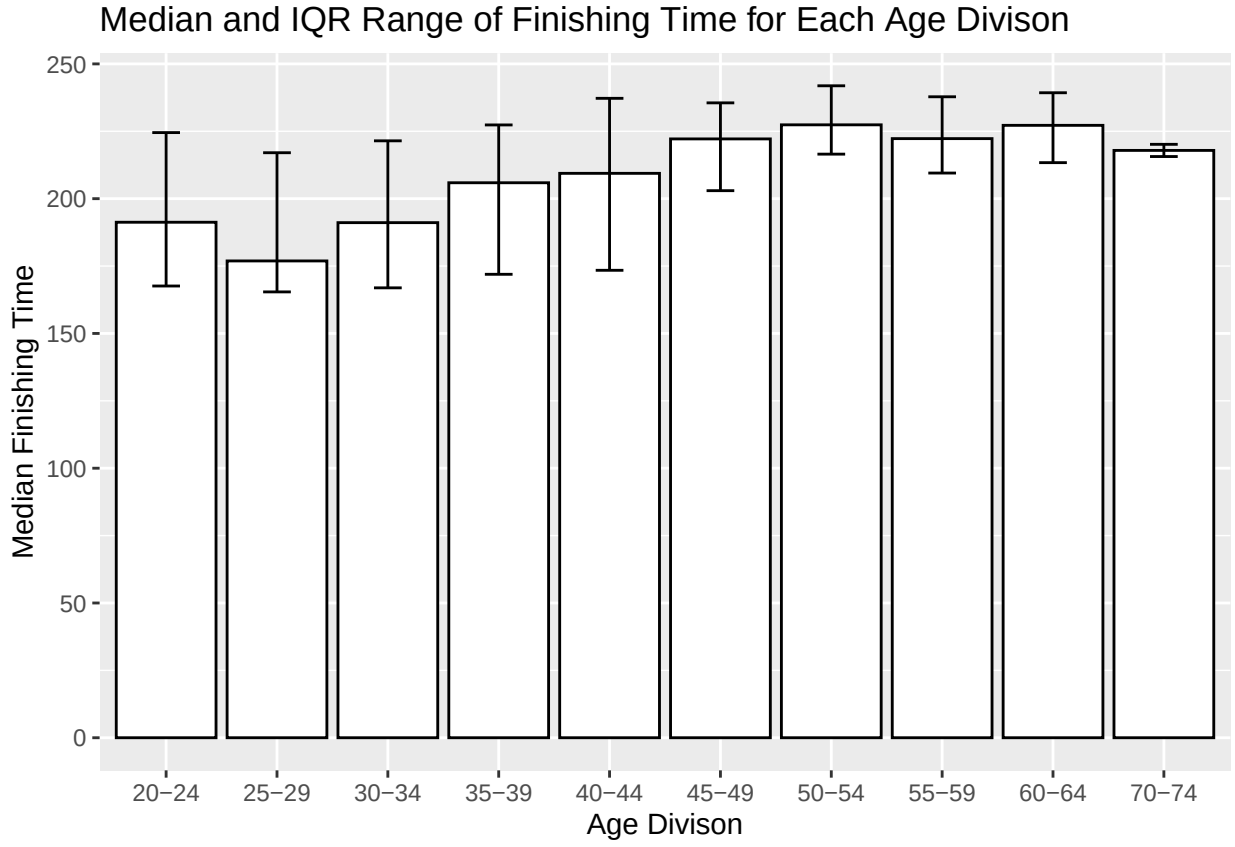
One can still conclude that the average female marathon finisher is younger than the average male finisher, as the existence of male finishers over the age of 60+ brings their average higher. Female finishers still tend to have a slightly smaller spread, but the shapes of the graphs are much more similar than before (the raw counts diminishes the shape of the female graph).

These new graphs, however, make it harder to conclude that there is a strong relationship between finishing a marathon and being male.

Part f

```
# Create a new dataframe with the median and IQR values
IQRnym2019 <- group_by(nym2019, DivAge) %>%
  summarise(min = min(Time),
            q1 = quantile(Time, 0.25),
            median = median(Time),
            q3 = quantile(Time, 0.75,
            max = max(Time)))

ggplot(IQRnym2019, aes(x=DivAge, y=median)) + geom_bar(stat = "identity",
                                                       color='black', fill='white') +
  geom_errorbar(aes(x=DivAge, ymin=q1, ymax=q3), width=0.25) +
  labs(title="Median and IQR Range of Finishing Time for Each Age Divison",
                     x="Age Divison", y="Median Finishing Time")
```



Part g

From the plot created in part (f), one can conclude that there is a positive relationship between age and median finishing time: As age increases, so does the median finishing time

of the marathon. This relationship mostly holds true throughout the graph, but can be seen to flatten out as the marathon runner gets above 50.

As an additional conclusion, there seems to be an inverse relationship between the variability in finishing times and age of the marathon participant: As the age of a participant increases, the IQR spread of finishing time decreases. The plot above illustrates that although the younger marathon finishers often have a faster median finishing time, those same young finishers have a wider IQR range. The larger IQR values indicate that the central portion of the data for younger finishers is more spread out than the older participants. Although a definite answer cannot be found within this initial observation, a possible explanation for the wider IQR ranges could be because of the larger number of participants in the younger age divisions.

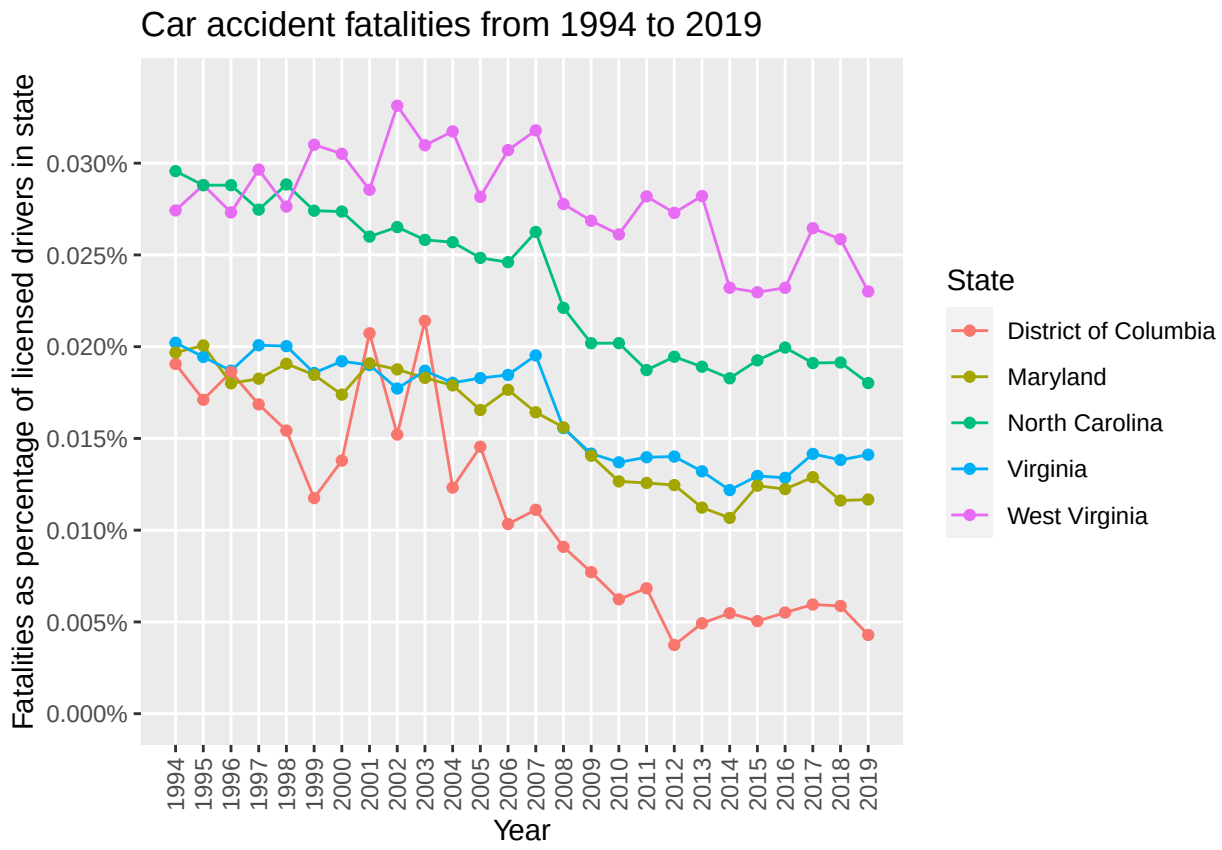
Problem 2

Original Graph

```
fatalities <- read.csv('fatalities.csv')
Graphic2 <- ggplot(fatalities, aes(x=Year, y=(Fatalities/Licensed.Drivers)/1000,
                                   color=State)) +
  geom_line() +
  geom_point()
```

Modified Graph

```
Graphic2 + labs(title="Car accident fatalities from 1994 to 2019") +
  scale_x_continuous(name="Year", breaks=c(1994:2019)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
  scale_y_continuous(name="Fatalities as percentage of licensed drivers in state",
                     limits =c(0, 0.00034),
                     breaks=seq(0.0000,0.00030,0.00005),
                     labels = scales::percent) +
  theme(panel.grid.minor.x=element_blank()) +
  theme(panel.grid.minor.y=element_blank())
```



Problem 3

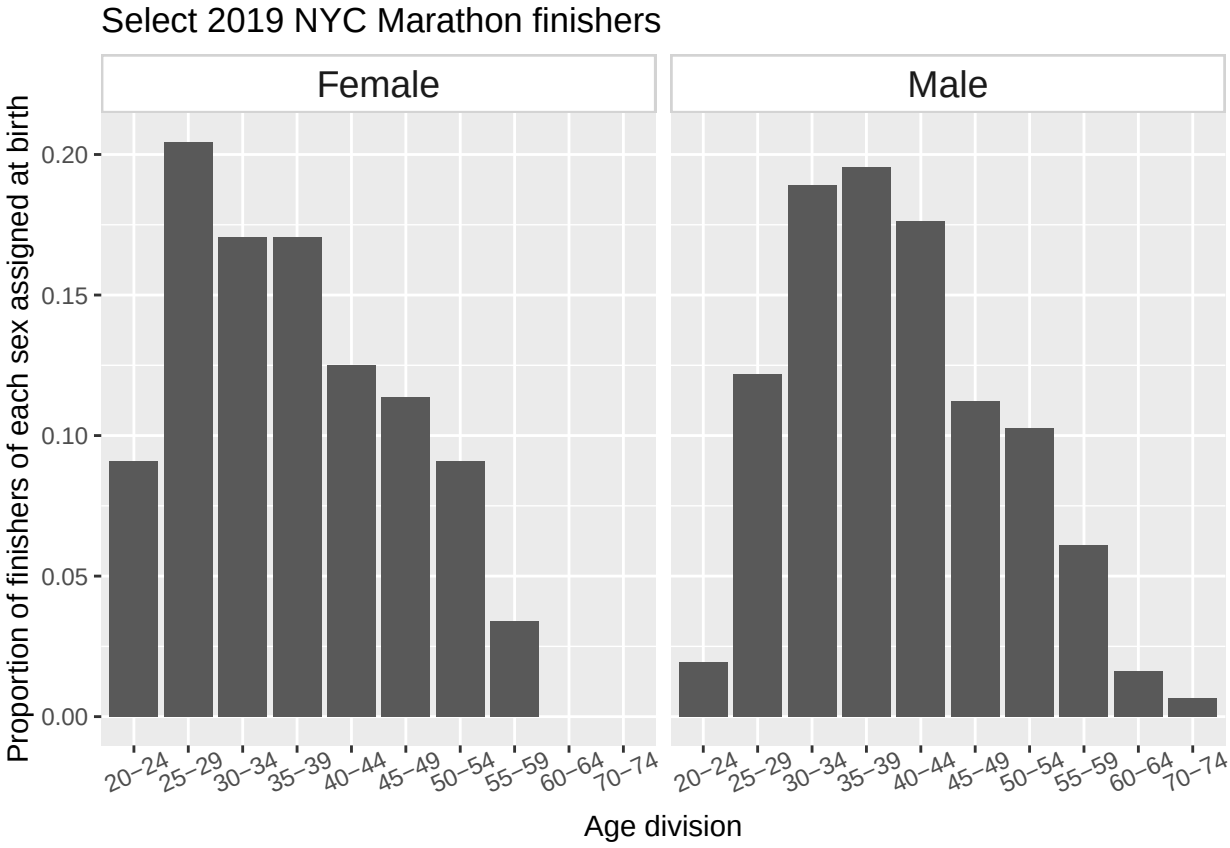
Original Graph

```
Graphic3 <- ggplot(nym2019, aes(x=DivAge)) + geom_bar(aes(y=..prop.., group = Sex))+
  facet_grid(.~Sex) +
  labs(y="prop")
```

Modified Graph

```
NewSex <- c("F" = "Female", "M" = "Male")

Graphic3 + facet_grid(.~Sex, labeller = labeller(Sex = NewSex)) +
  theme(strip.text = element_text(size = rel(1.25)),
    strip.background = element_rect(fill = "white", colour = "lightgray",
                                    size = 0.75)) +
  theme(axis.text.x = element_text(angle = 25, vjust = 0.85)) +
  labs(title="Select 2019 NYC Marathon finishers", x="Age division",
       y="Proportion of finishers of each sex assigned at birth")
```



Problem 4

Original Graph

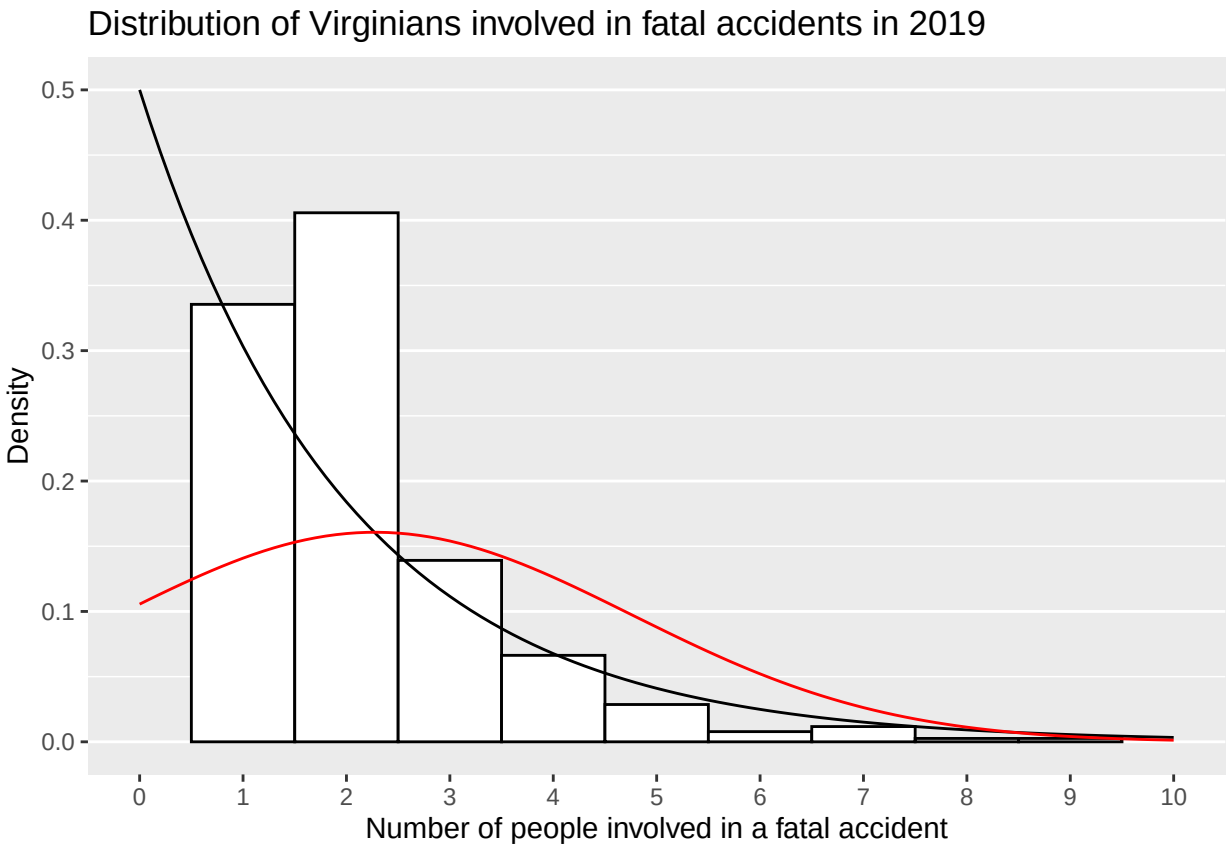
```
# Load the dataset; assign the dataframe to object called 'fatal_accidents'
fatal_accidents <- read.csv('fatal accidents.csv')
fatal_accidents <- group_by(fatal_accidents, Case.number)%>%
  mutate(People.count=sum(People.count.IN, People.count.OUT))
onlyVA <- filter(fatal_accidents, State == "Virginia")

xbar <- mean(onlyVA$People.count)
s <- sd(onlyVA$People.count)

Graphic4 <- ggplot(onlyVA, aes(x = People.count)) +
  geom_histogram(aes(y = ..density..),
                 colour = 1, fill = "white", binwidth=1) +
  stat_function(fun=dchisq, args=list(df=2)) +
  stat_function(fun=dnorm, args=list(mean=xbar, sd=s), color="red")
```

Modified Graph

```
Graphic4 + labs(title="Distribution of Virginians involved in fatal accidents in 2019",
       y="Density", x='Number of people involved in a fatal accident') +
  theme(panel.grid.major.x=element_blank(), panel.grid.minor.x=element_blank()) +
  scale_x_continuous(breaks = c(0:10), limits = c(0,10))
```



Problem 5

The density histogram above is a graph which represents the distribution of values in a dataset. The orientation of the bins in the plot above illustrate that the distribution has a small spread and outliers which skew the distribution to the right. Although the average number of people involved in a fatal accident is still low (approximately 2.5 people), the plot above indicates that the median value would be lower. One can also conclude that the most common number of people involved in a fatal accident was 2 in 2019, as this category has the highest percentage of observations in the dataset (40%).

In addition to the histogram itself, the chi-square and normal distribution lines allow one to draw further conclusions about the distribution of the number of people involved in fatal accidents. Even when considering the true mean and standard deviation from the dataset, the distribution of people in fatal accidents in Virginia (2019) does not follow the normal curve that is expected. This supports the previous conclusion that the true median value will be lower than the average number of people involved in fatal accidents. This is important because it over-estimates the amount of people usually affected by fatal accidents.

It should also be noted that the standard deviation of the plot is high. Seeing that the normal distribution line is considerably flat, a high standard deviation allows one to conclude that the distribution of people is relatively spread out. While the average number of people involved in a fatal accident may be low, it is quite dangerous to consider that there is a large amount of variation in this set of values.