Homework 6

[redacted]

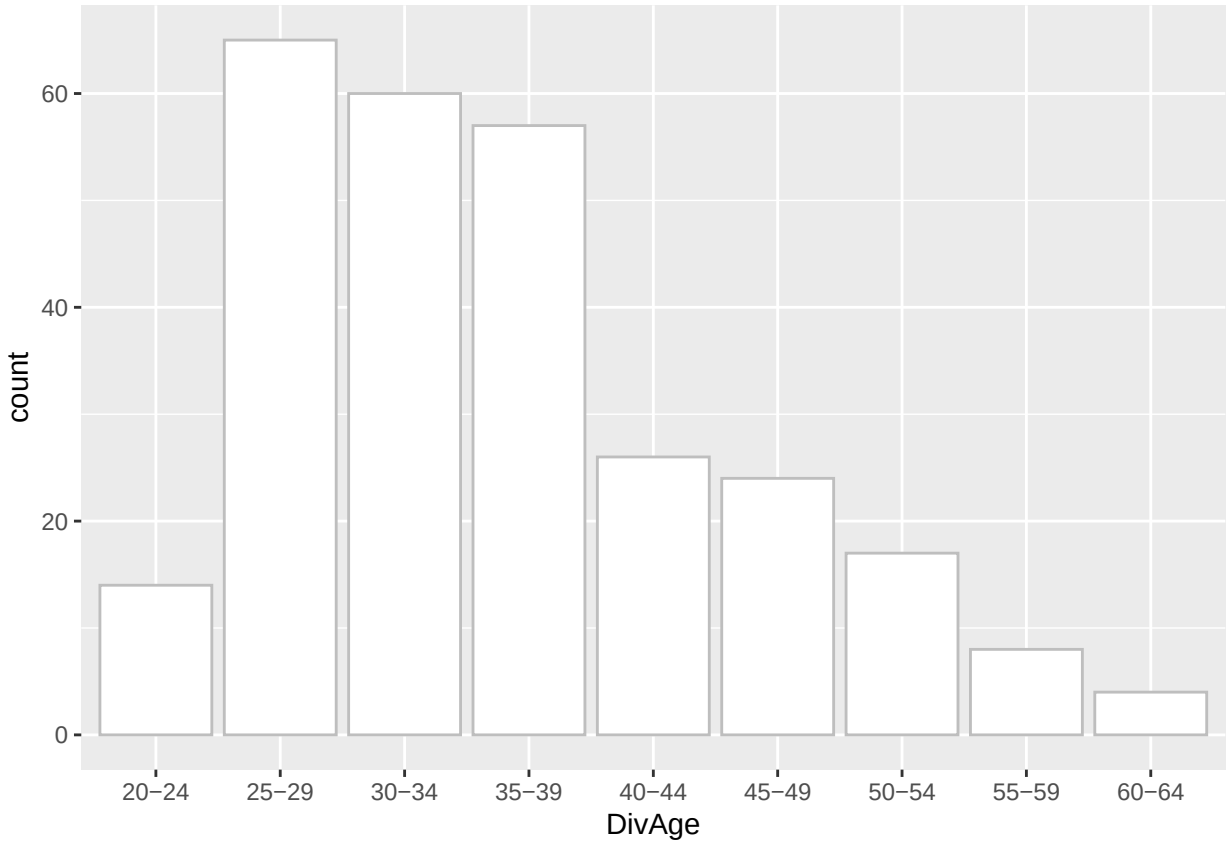
10/19/2022

Problem 1

```
library(ggplot2)
nym2021 <- read.table("/Users/[redacted]/Desktop/STAT 3080/nym2021.txt", header = TRUE)
```

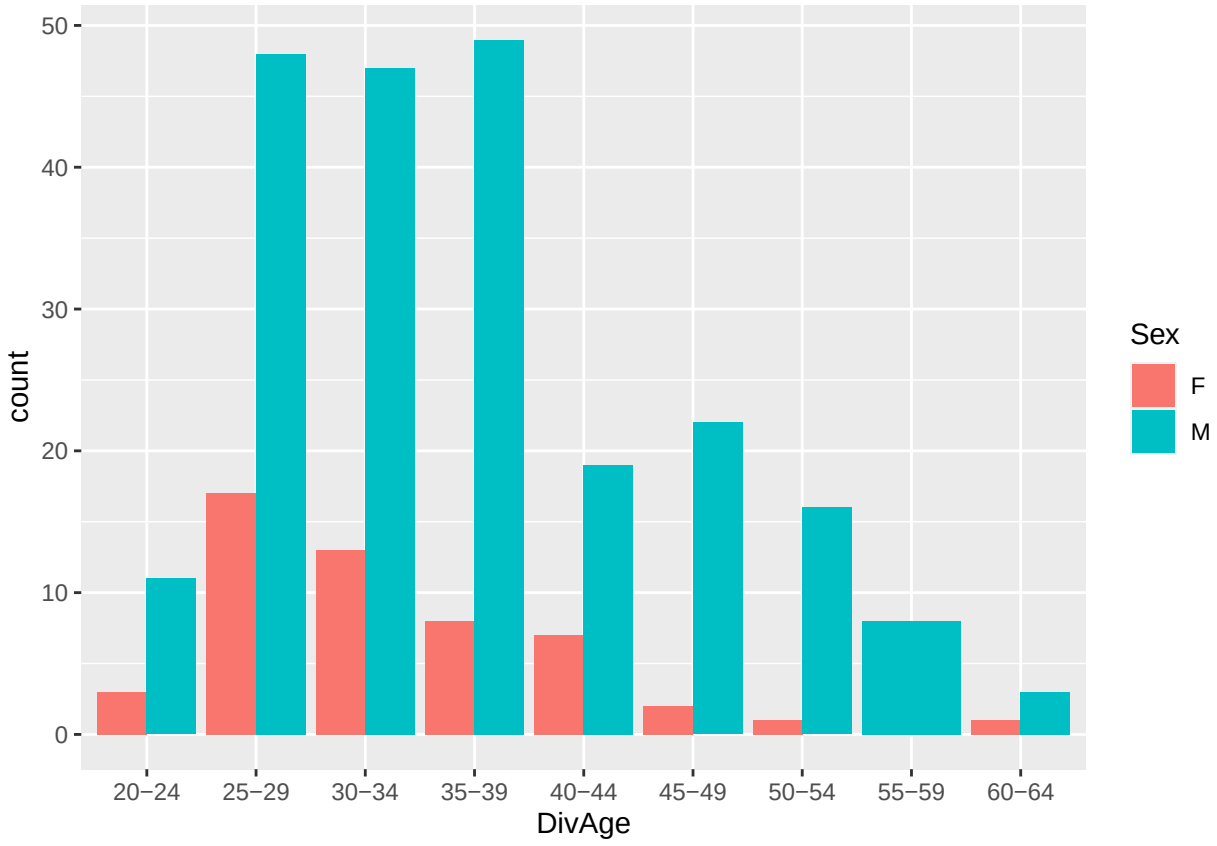
(a)

```
ggplot(nym2021, aes(x=DivAge)) + geom_bar(fill="white", color="gray")
```



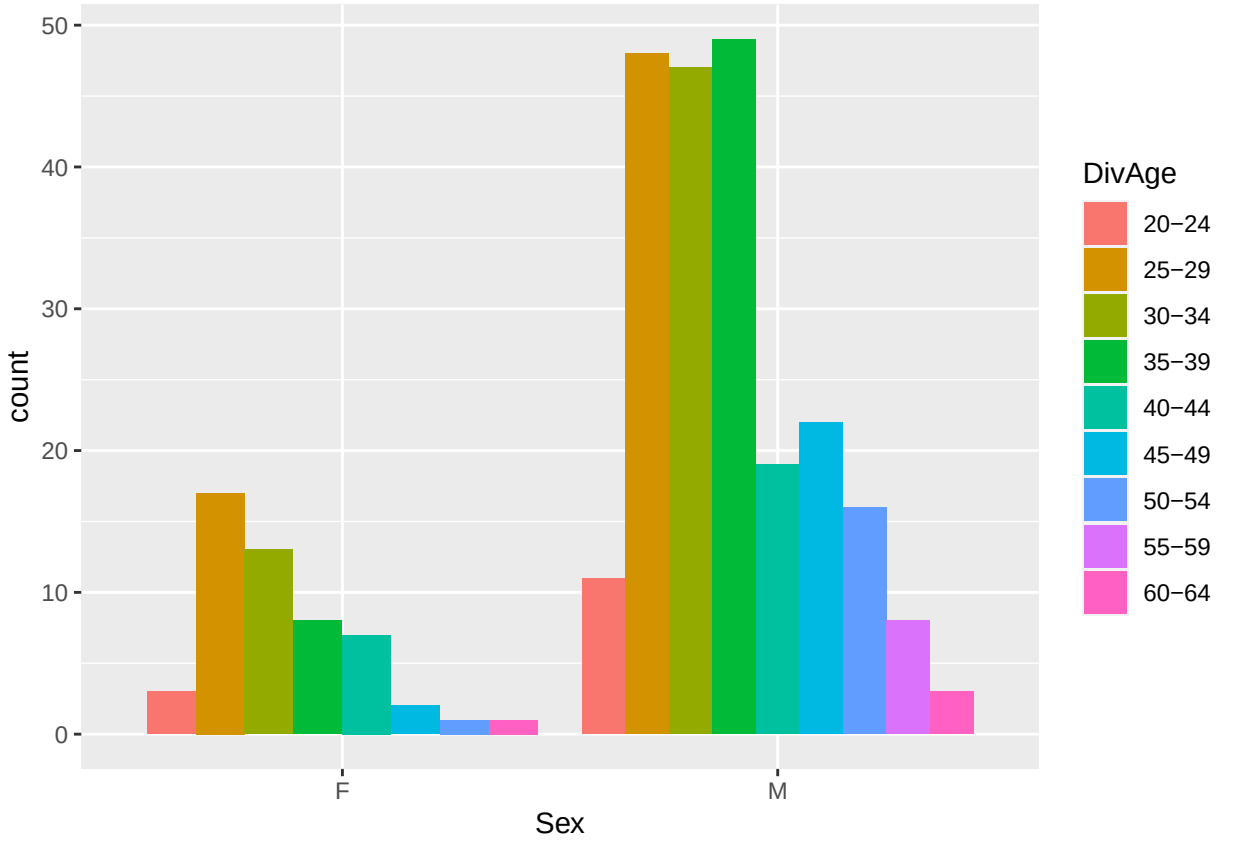
(b)

```
ggplot(nym2021, aes(x=DivAge, fill=Sex)) + geom_bar(position = "dodge")
```



(c)

```
ggplot(nym2021, aes(x=Sex, fill=DivAge)) + geom_bar(position = "dodge")
```



(d)

In the plot created in part B, it is apparent that in each age division, there are more men finishers than women. In the plot created in part C, this same conclusion can be drawn as the count of men finishers in each age division is substantially higher than that of women. Both graphs give us the same conclusions.

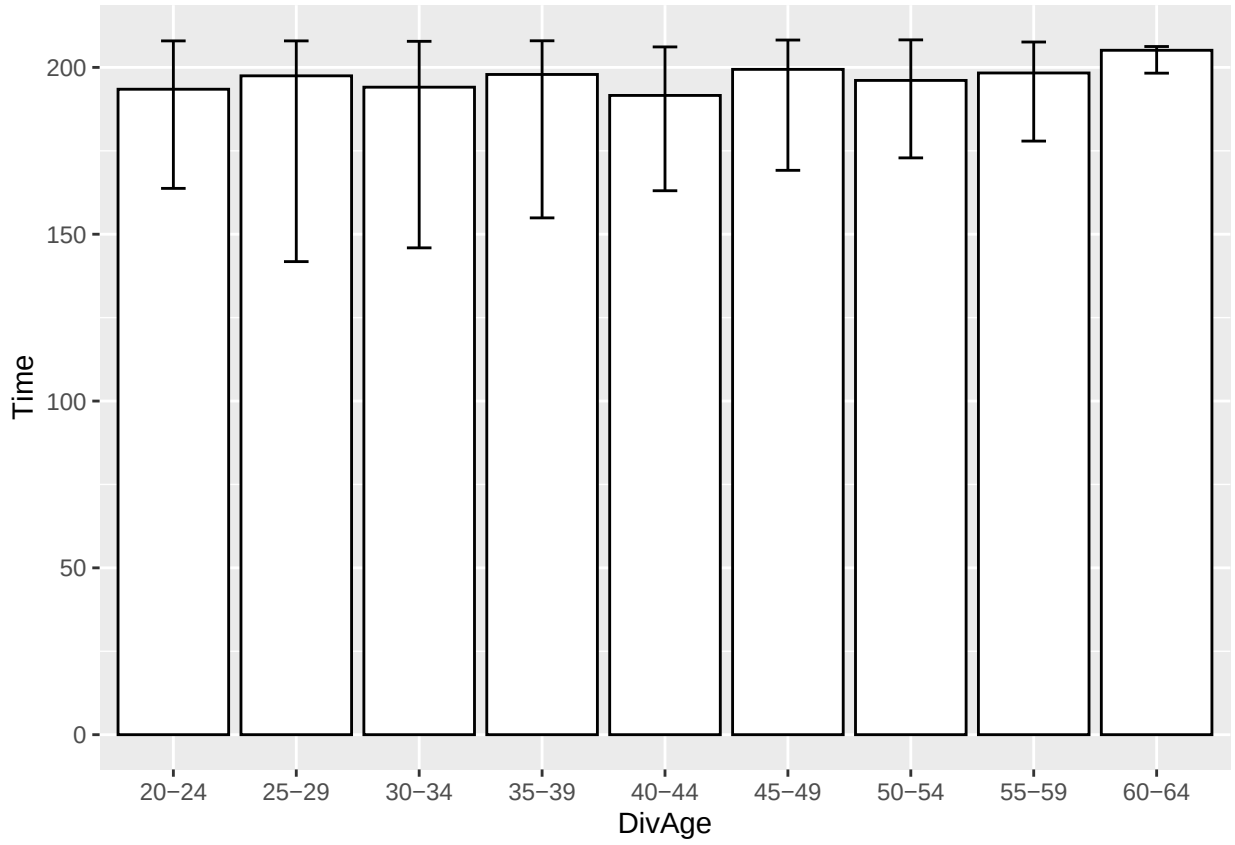
(e)

This plot will change my conclusions drawn in part D due to the fact that these proportions are only comparing values of the same sex with one another and not the total number of both men and women like the plots previously. This means that we can see that there is a larger proportion of women who compete in the 25-29 and 30-35 age group than proportion of men who compete in those age divisions.This is different than what we had previously concluded due to the fact that since there are more men competing than women, their propritons will be smaller as they have a bigger total(denominator).

(f)

```
plotF <- ggplot(nym2021, aes(x=DivAge, y=Time))
```

```
plotF + stat_summary(fun=median, geom="bar", fill="white", color="black") +
  stat_summary(fun.data=median_hilow, geom="errorbar", width=0.2)
```



(g)

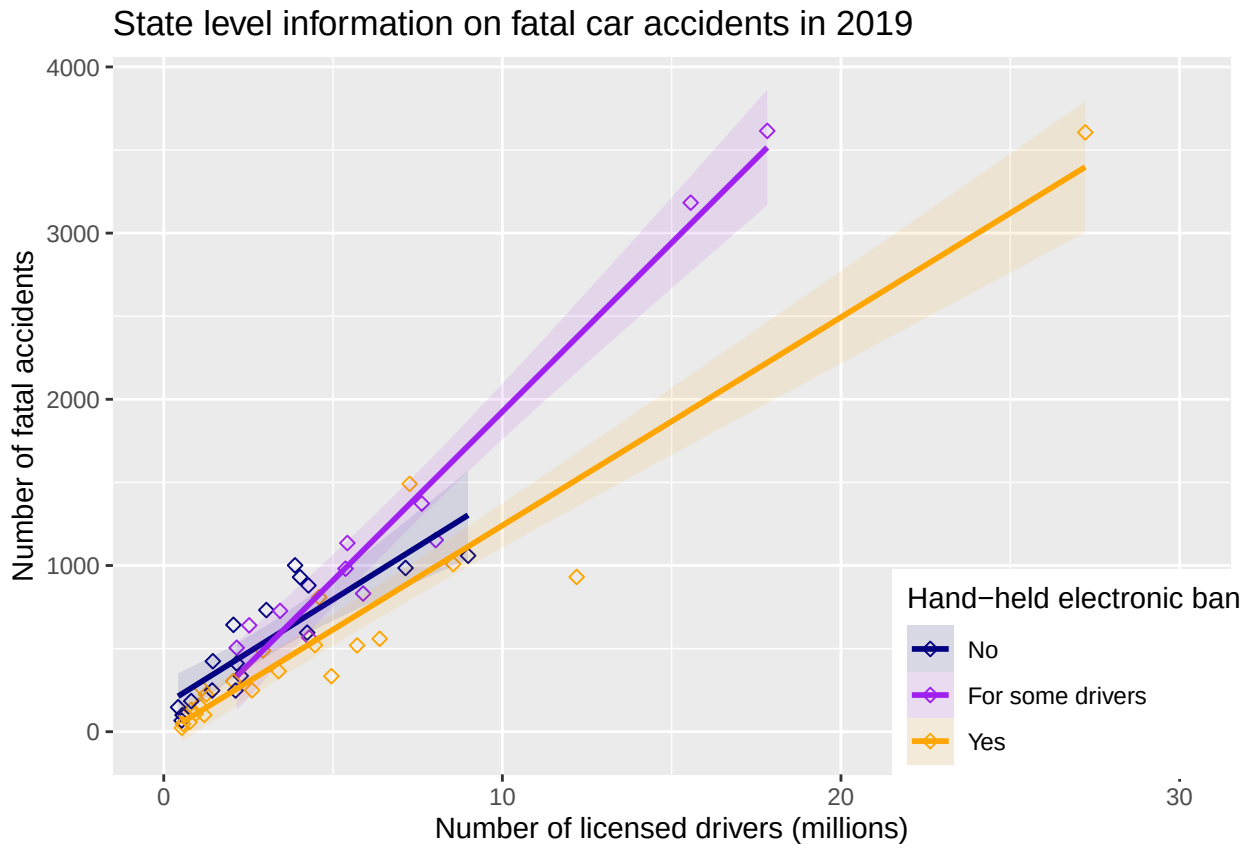
I can conclude that the median finishing times across all age divisions are relatively similar. The error bars give an idea of how precise our measurements are and it is apparent that there is more variation and less precision in the younger age groups.

Problem 2

```
crashes <-data.frame(read.csv("/Users/[redacted]/Desktop/STAT 3080/state crashes.csv",header=TRUE))

crashes$Hand.held.ban<- factor(crashes$Hand.held.ban)

Graphic2 <- ggplot(crashes, aes(x=(Licensed.drivers/1000), y=Fatal.crashes,
color=Hand.held.ban, fill=Hand.held.ban))

Graphic2 + geom_point(shape=5) +
  scale_color_manual(values=c("1"="navy blue","2"="purple", "3"="orange"),
                     labels=c("No","For some drivers","Yes"),
```

```
                     name="Hand-held electronic ban") +
  scale_fill_manual(values=c("1"="navy blue","2"="purple", "3"="orange"),
                    labels=c("No","For some drivers", "Yes"),
                    name="Hand-held electronic ban") +
  theme(legend.position = c(.86, .14)) +
  geom_smooth(method = lm, alpha=0.1) +
  labs(title="State level information on fatal car accidents in 2019",
       x="Number of licensed drivers (millions)", y="Number of fatal accidents") +
  coord_cartesian(xlim=c(0,30))
```

```
## `geom_smooth()` using formula 'y ~ x'
```

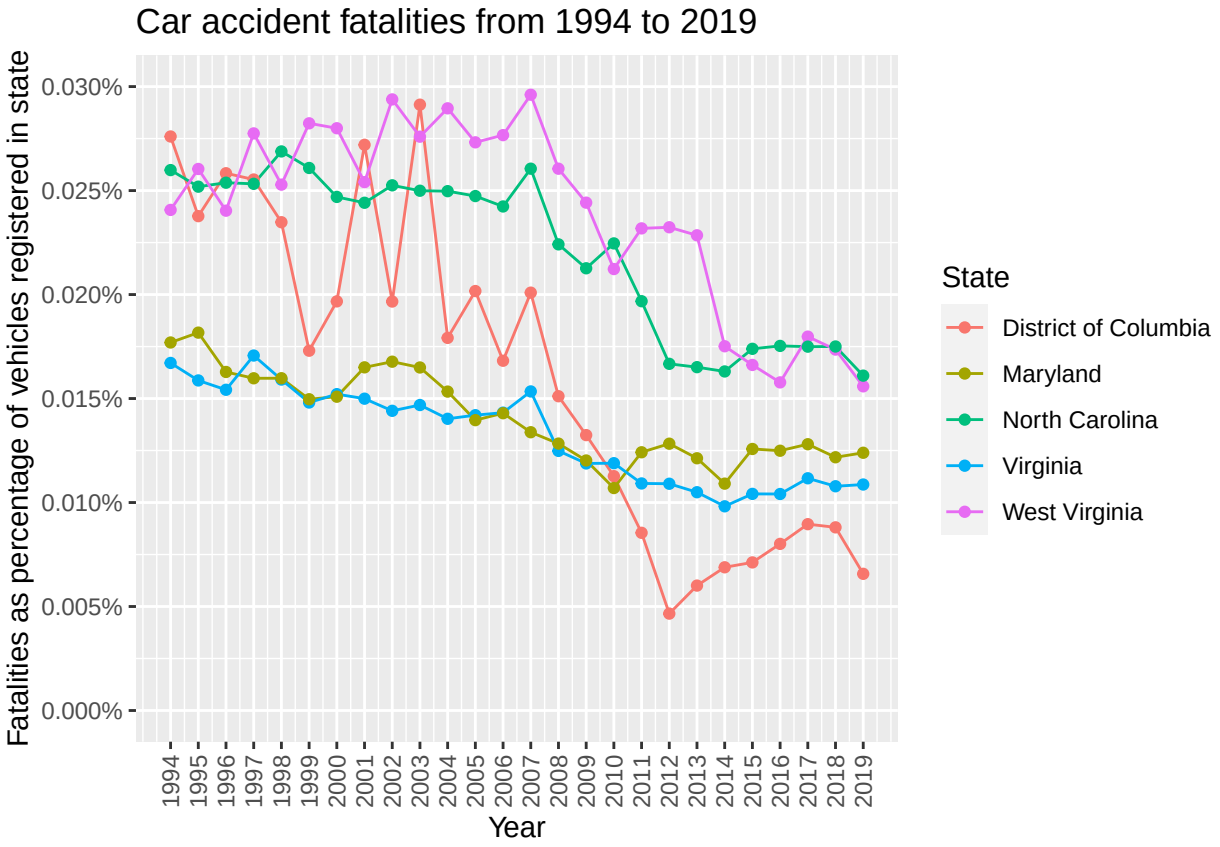


Problem 3

```
fatal <-data.frame(read.csv("/Users/[redacted]/Desktop/STAT 3080/fatalities.csv",header=TRUE))

Graphic3 <- ggplot(fatal, aes(x=Year, y=((Fatalities/Registered.Vehicles)/
1000), color=State))

Graphic3 + geom_line() + geom_point() +
  scale_x_continuous(breaks = seq(min(fatal$Year), max(fatal$Year), by = 1)) +
  theme( axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
```

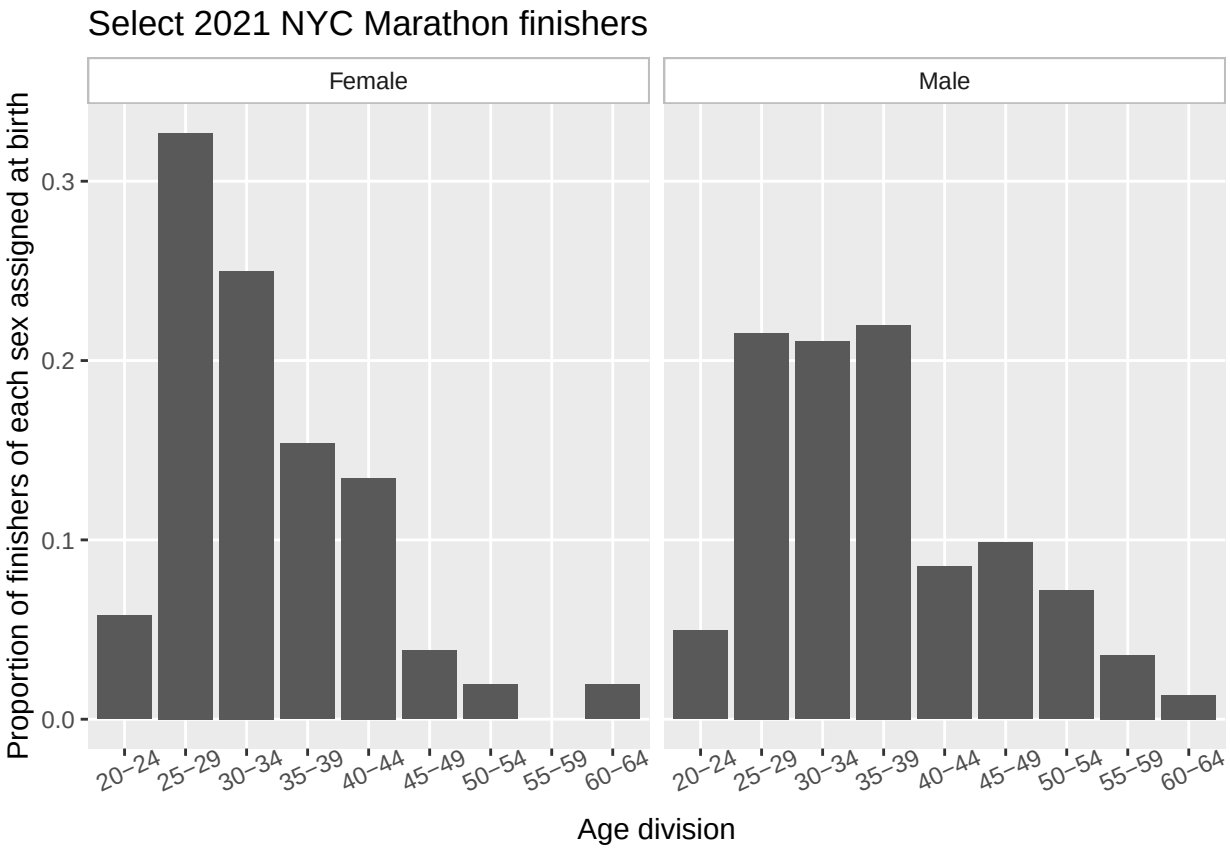
```
  labs(title="Car accident fatalities from 1994 to 2019",
       y= "Fatalities as percentage of vehicles registered in state") +
  scale_y_continuous(breaks = round(seq(min(fatal$Fatalities/fatal$Registered.Vehicles),
                                        max(fatal$Fatalities/fatal$Registered.Vehicles),by=0.5)),4) +
  scale_y_continuous(labels = scales::percent_format(accuracy = 0.001),
                     breaks = seq(0,3e-04,5e-05), limits = c(0,3e-04))
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```



Problem 4

```
Graphic4 <- ggplot(nym2021, aes(x=DivAge, y = ..prop.., group = Sex)) + geom_bar() +
  facet_grid(.~Sex, labeller=as_labeller(c("F" = "Female", "M" = "Male"))) +
  labs(title="Select 2021 NYC Marathon finishers", x="Age division",
       y="Proportion of finishers of each sex assigned at birth") +
  theme(strip.background = element_rect(fill = "white", color ="gray")) +
  theme(panel.grid.minor=element_blank(), axis.text.x = element_text(angle=25))
Graphic4
```

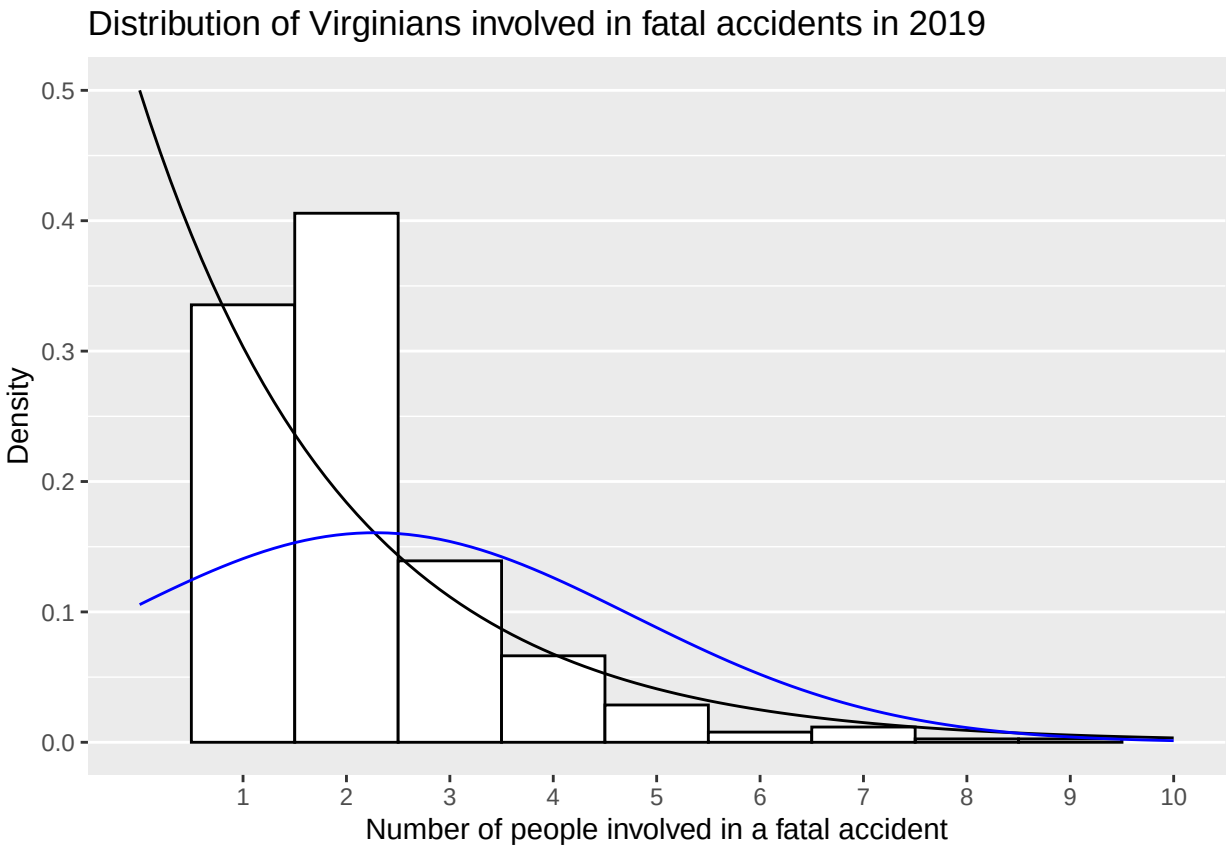


Problem 5

```
fatal1 <-data.frame(read.csv("/Users/[redacted]/Desktop/STAT 3080/fatal accidents.csv",header=TRUE)

Virginia <- fatal1[fatal1$State == "Virginia",]
Virginia$People.count <- Virginia$People.count.IN + Virginia$People.count.OUT

Graphic5 <-ggplot(Virginia, aes(x=People.count)) + geom_histogram(color ="black", fill="white",
          binwidth = 1, aes(y=..density..)) +
  stat_function(fun=dchisq, args = list(df=2)) + stat_function(fun=dnorm, color = "blue",
           args = list( mean = mean(Virginia$People.count),
           (sd = sd(Virginia$People.count))))

Graphic5 + scale_x_continuous(breaks=1:10, limits = c(0,10)) +
labs(title = "Distribution of Virginians involved in fatal accidents in 2019",
     x = "Number of people involved in a fatal accident", y = "Density")+
  theme(panel.grid.minor.x=element_blank(), panel.grid.major.x=element_blank())
```

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



Problem 6

I can conclude that the distribution of individuals involved in fatal car accidents is right skewed. Therefore, this means that there is a greater likelihood that fewer people will be involved in fatal car accidents.

References

1. StackOverlow.com
2. [redacted]