Homework 5

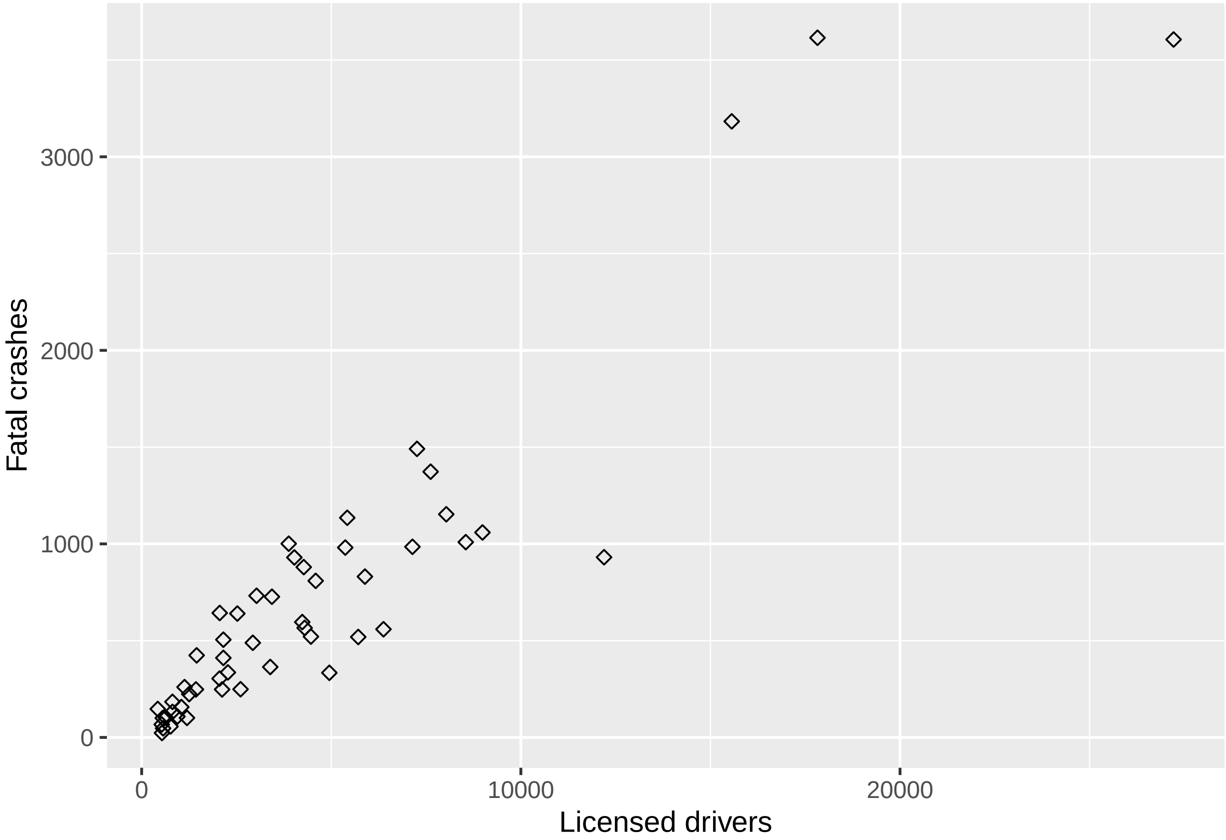
Akhil Havaldar (ash2sfp)

Question 1

a)

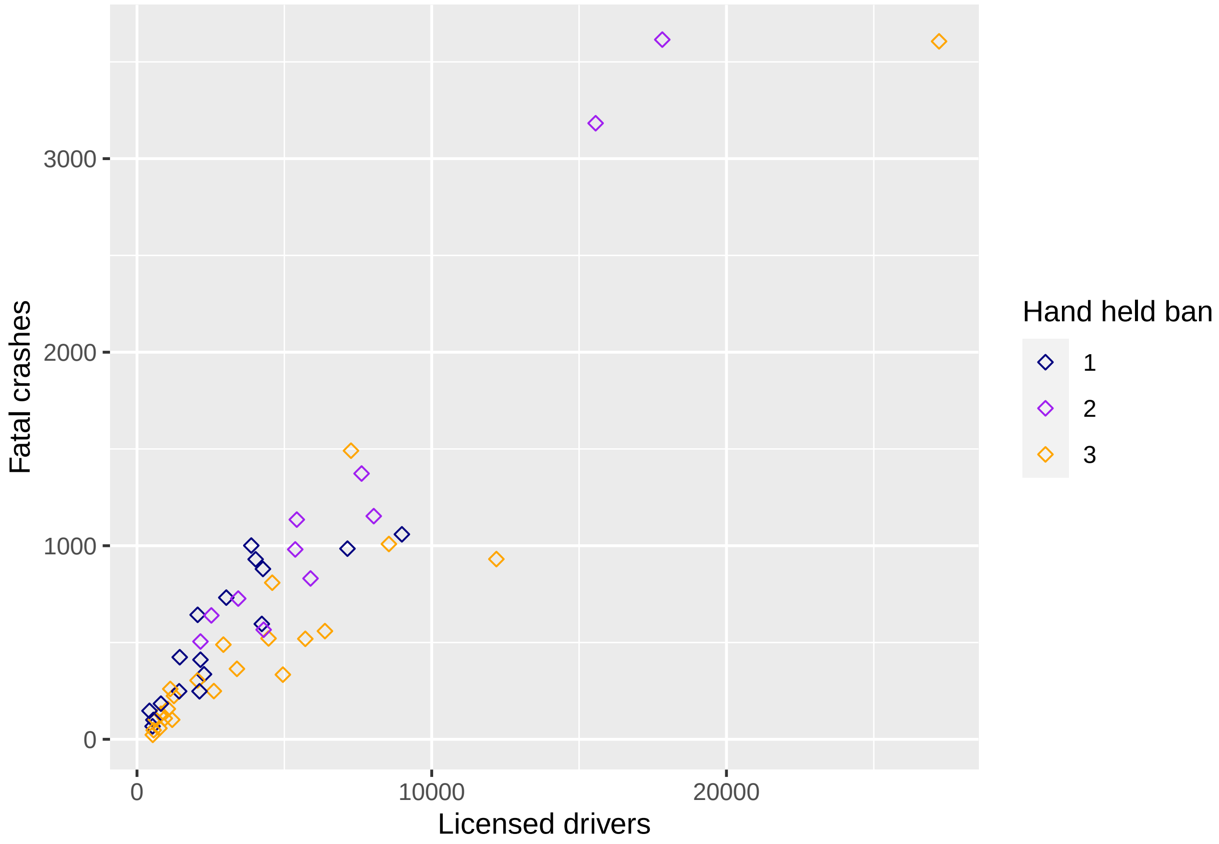
```
library(readr)
library(ggplot2)
library(tidyverse)

dat <- read_csv("state crashes.csv")
ggplot(dat, aes(x=`Licensed drivers`, y=`Fatal crashes`)) + geom_point(shape=5)
```



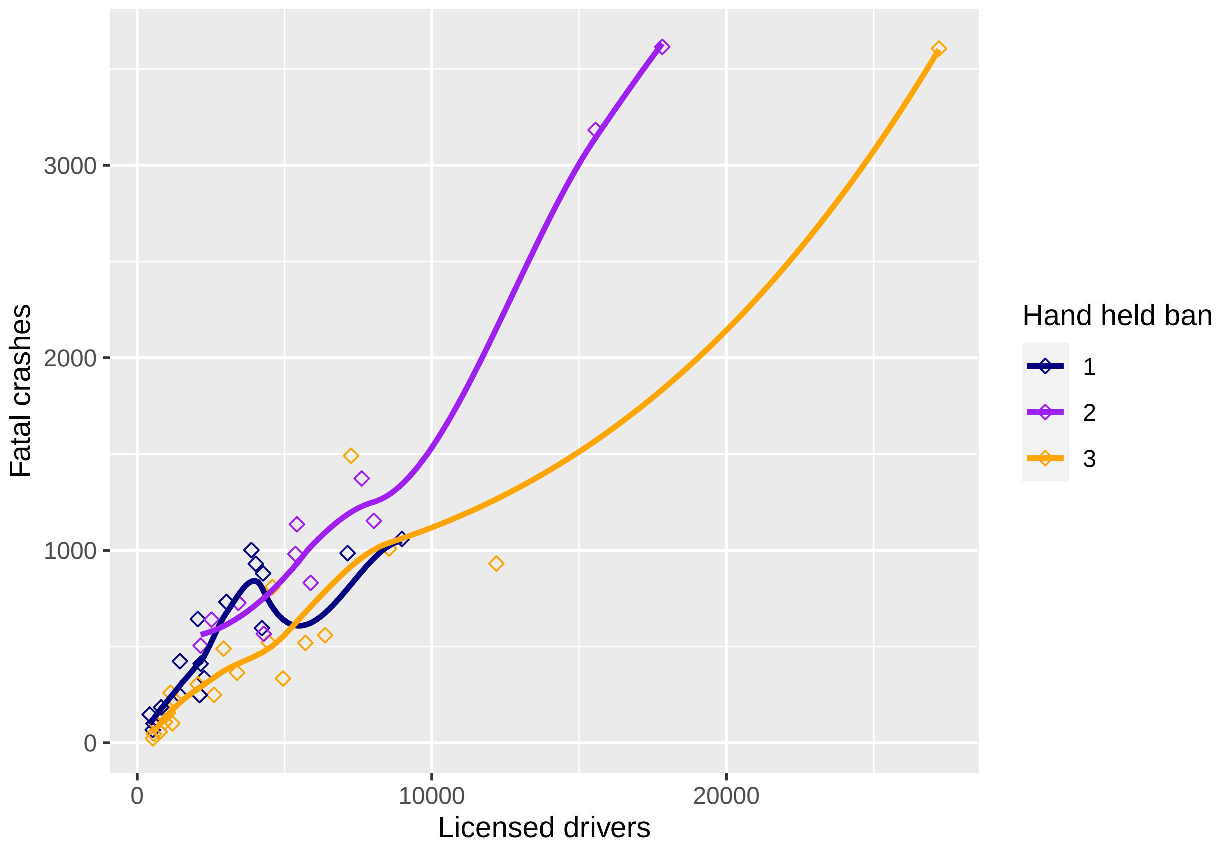
b)

```
dat$`Hand held ban` <- as.factor(dat$`Hand held ban`)
ggplot(dat, aes(x=`Licensed drivers`, y=`Fatal crashes`,
                color=`Hand held ban`)) + geom_point(shape=5)+
  scale_color_manual(values=c(`1`="navyblue",`2`="purple", `3`="orange"))
```



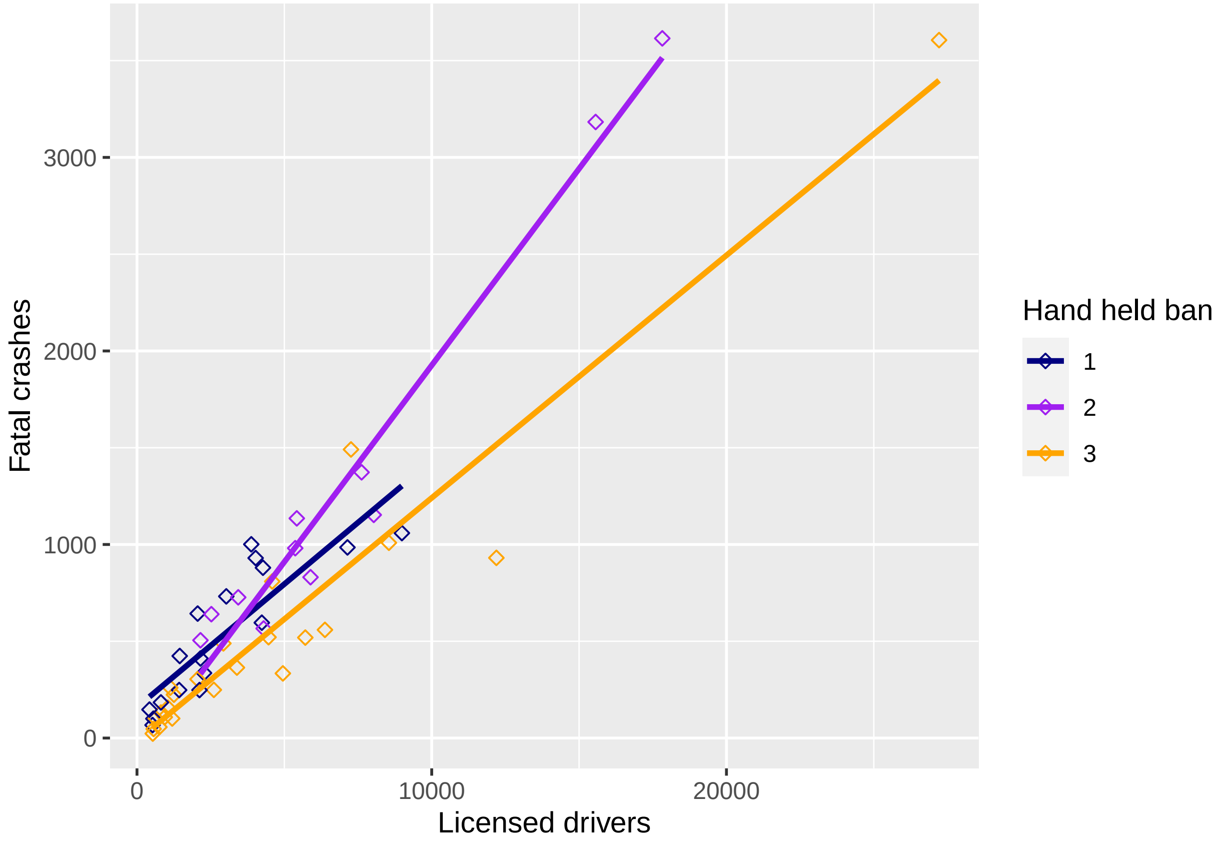
c)

```
ggplot(dat, aes(x=`Licensed drivers`,
                y=`Fatal crashes`, color=`Hand held ban`)) +
  geom_point(shape=5)+
  scale_color_manual(values=c(`1`="navyblue",`2`="purple", `3`="orange")) +
  geom_smooth(se=F)
```



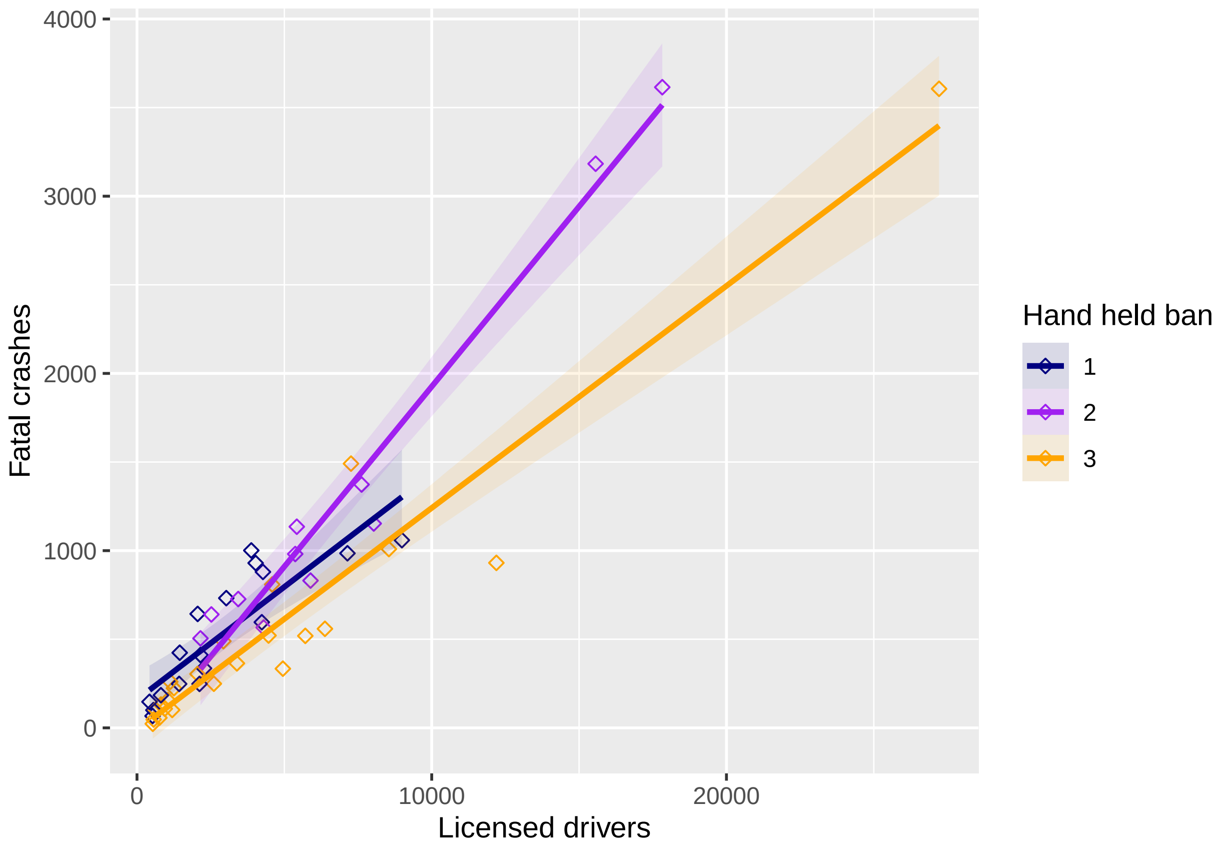
d)

```
ggplot(dat, aes(x=`Licensed drivers`, y=`Fatal crashes`,
                color=`Hand held ban`)) +
  geom_point(shape=5)+
  scale_color_manual(values=c(`1`="navyblue",`2`="purple", `3`="orange")) +
  geom_smooth(method=lm, se=F)
```



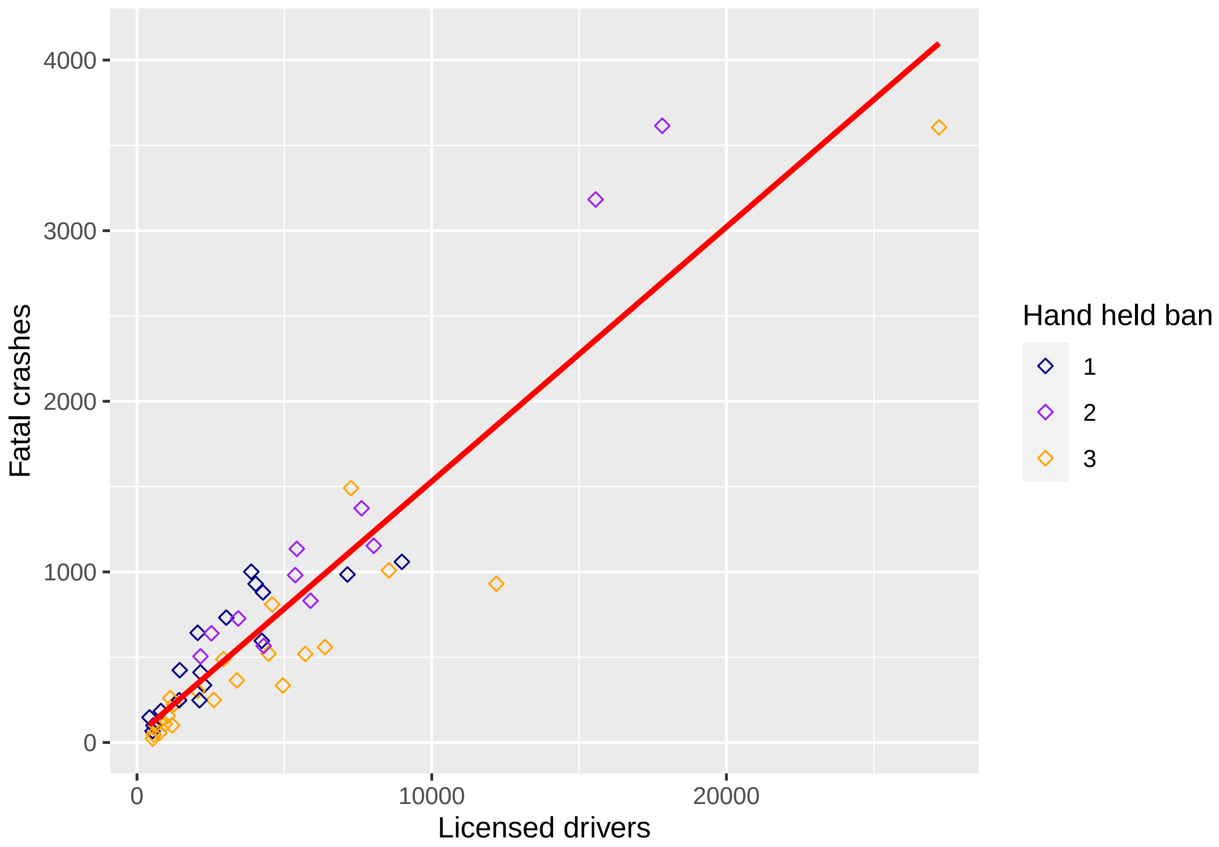
e)

```
ggplot(dat, aes(x=`Licensed drivers`, y=`Fatal crashes`,
                color=`Hand held ban`, fill=`Hand held ban`)) +
  geom_point(shape=5)+
  scale_color_manual(values=c(`1`="navyblue",`2`="purple", `3`="orange")) +
  geom_smooth(method=lm,
              alpha=0.1)+scale_fill_manual(values=c("navyblue","purple","orange"))
```



f)

```
ggplot(dat, aes(x=`Licensed drivers`,
                y=`Fatal crashes`, color=`Hand held ban`)) +
  geom_point(shape=5)+
  scale_color_manual(values=c(`1`="navy",`2`="purple", `3`="orange")) +
  geom_smooth(method=lm, se=F, color="red")
```



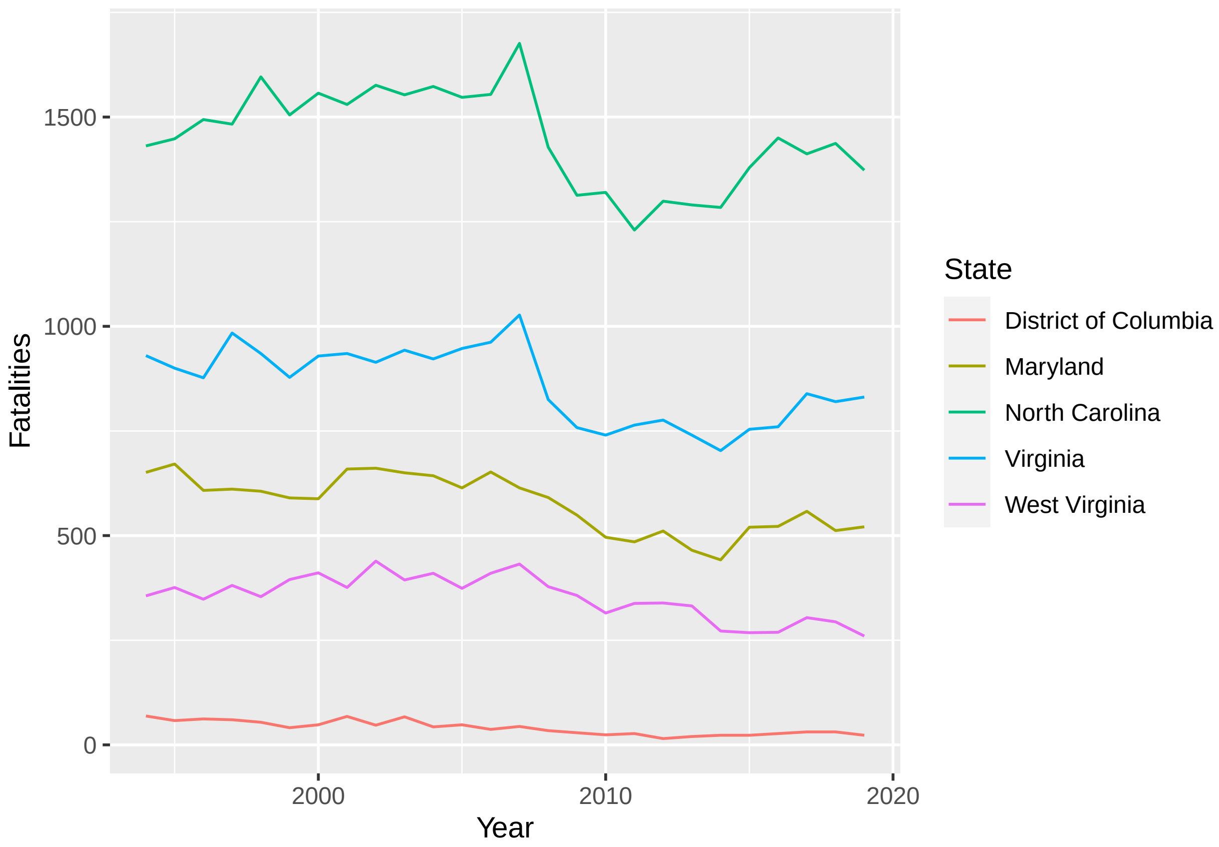
g)

- From the graph, we can see a correlation between licensed drivers and number of fatal crashes. This makes sense since if you have more drivers, crashes will generally occur more often. From the color of the points, we can see that the states that have banned hand held devices while driving have a lower number of fatal crashes (the orange points are more concentrated in the lower left region).

Question 2

a)

```
fat <- read_csv("fatalities.csv")
ggplot(fat, aes(x=Year, y=Fatalities, color = State))+
  geom_line()
```

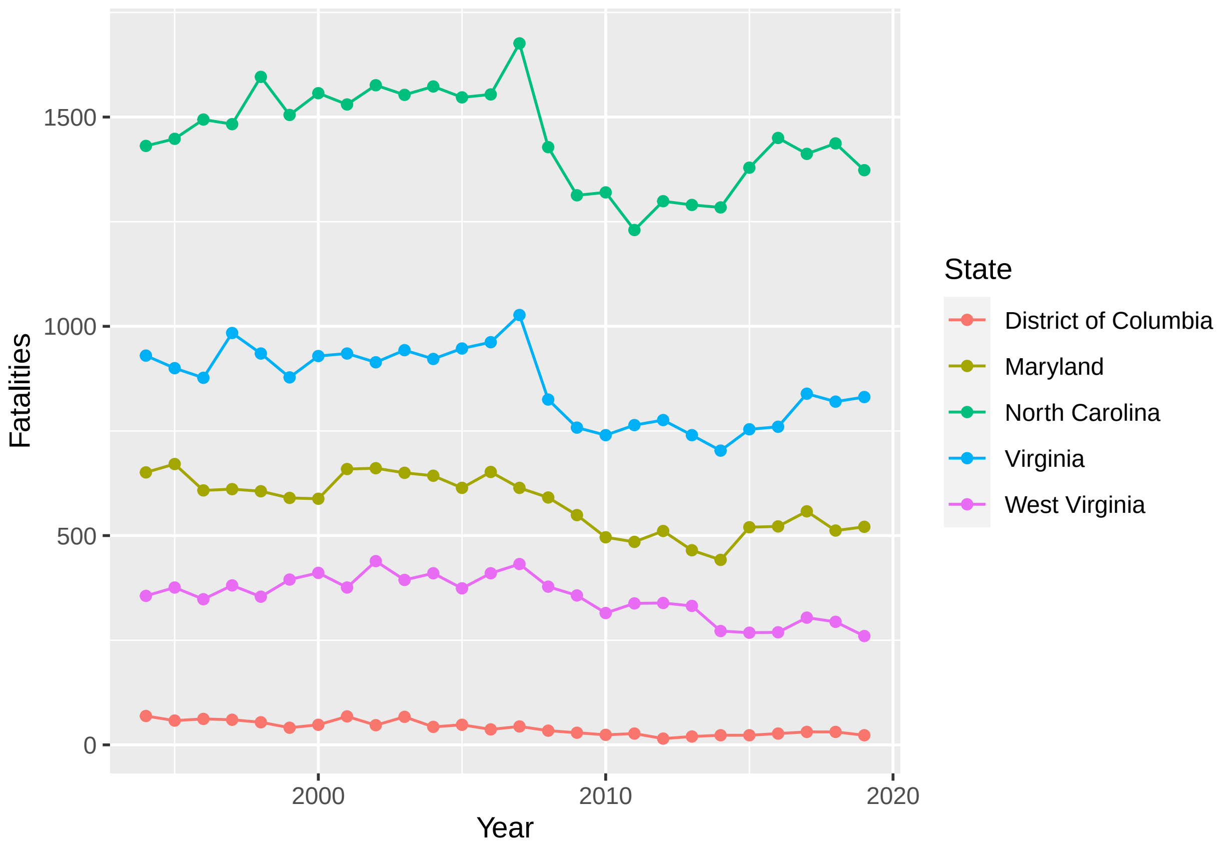


b)

- From the graph in (a), we can see the states exhibit similar trends over time, but all have much different values. For example, both Virginia and North Carolina have a drop off in fatalities around 2007-08, but Virginia has around 1000 fatalities while North Carolina has around 1500 fatalities.

c)

```
ggplot(fat, aes(x=Year, y=Fatalities, color = State)) +
  geom_line()+geom_point()
```

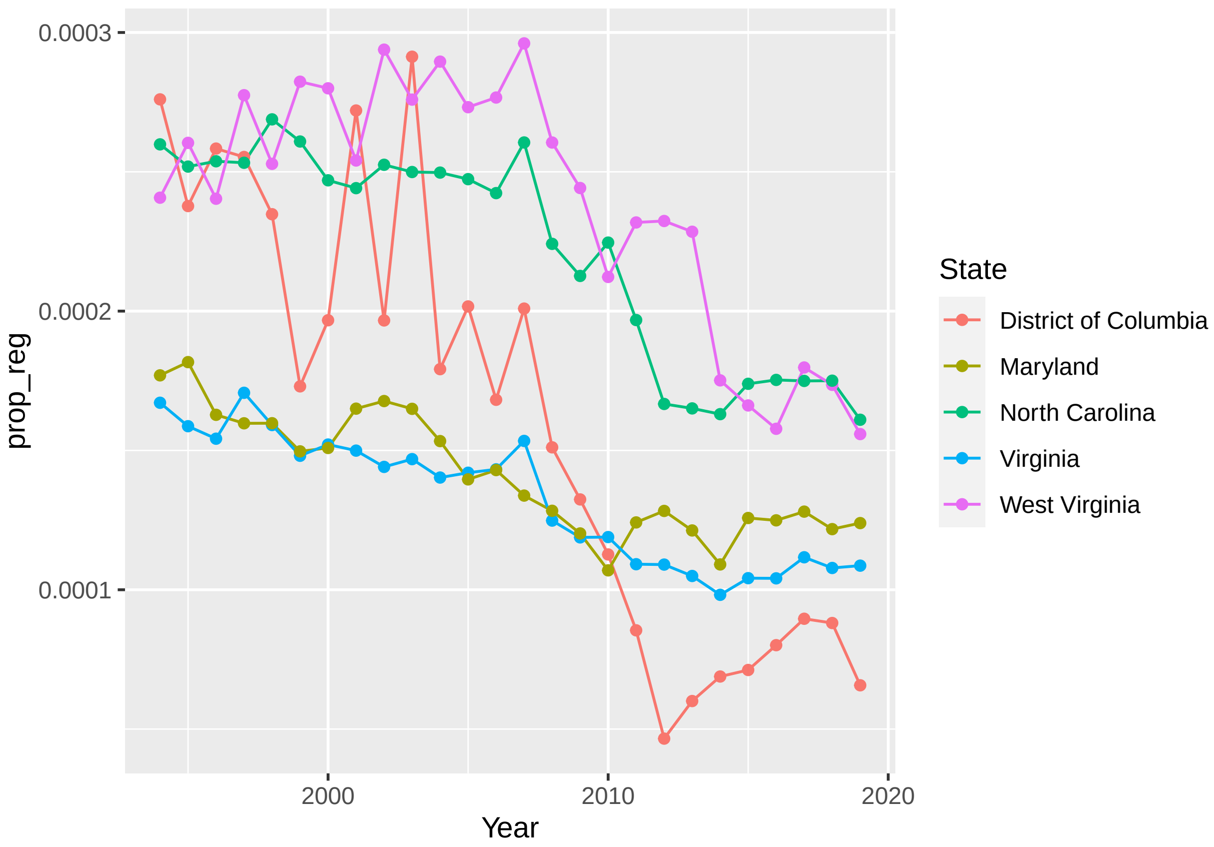


d)

- The graph in (c) allows us to see the exact years of all the points in the data, whereas in (a) we cannot see that as clearly.

e)

```
library(scales)
fat$prop_reg <- (fat$Fatalities/fat$`Registered Vehicles`)
ggplot(fat, aes(x=Year, y=prop_reg, color = State)) +
  geom_line()+geom_point()+scale_y_continuous(labels =
                                              label_number(scale=
                                                           0.001))
```

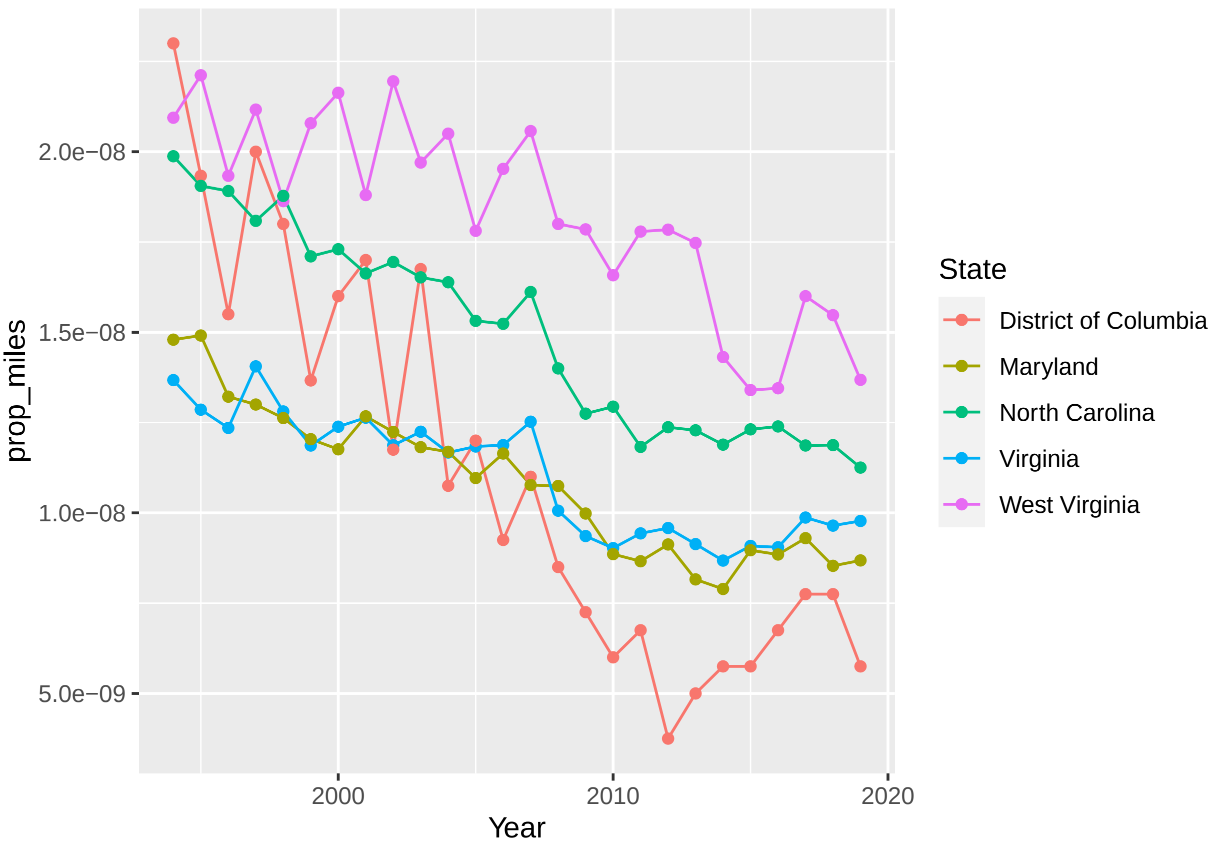


f)

- Yes. Here we can see that the states do not exhibit similar trends. In fact, there is a lot more variation in fatalities per total vehicles registered by each state than just fatalities by state. For example, now it seems like West Virginia is the most dangerous driving state for fatalities instead of North Carolina or Virginia. With the change in scale, the graph becomes easier to interpret in terms of fatality values. Now, we can read the graph as the number of fatalities per registeres vehicle over the years.

g)

```
fat$prop_miles <- (fat$Fatalities/(fat$`Vehicle Miles`*1000000000))
ggplot(fat, aes(x=Year, y=prop_miles, color = State)) +
  geom_line()+geom_point()
```

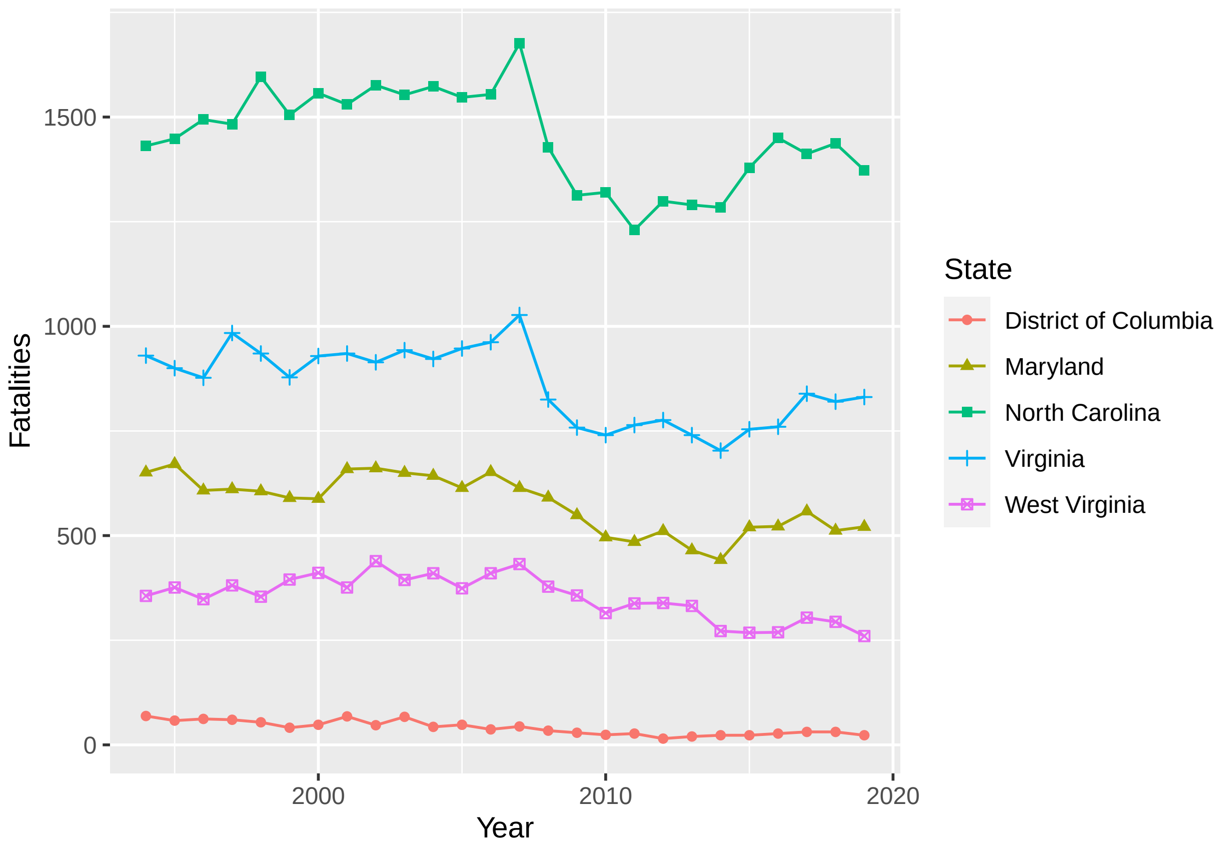


h)

- Between the plots in (e) and (g) the trends seem relatively similar. The biggest difference in the 2 graphs is how D.C. appears. In (e) there seems to be a fluctuation in the number of cars registered between 2000-05 which could be the reason why the proportion of fatalities to number of miles driven also falls off during these years. Less cars registered leads to a decreased number of miles driven. The interpretation of the graph after scaling is now the amount of fatalities per mile driven, which can give us a better sense of the number of fatalities that occur when someone drives a short distance.

i)

```
ggplot(fat, aes(x=Year, y=Fatalities, color = State)) +
  geom_line()+geom_point(aes(shape=State))
```



j)

- Varying point type relays pretty much the same information as varying by color, but this makes it even more clear the lines are varying by a different variable (state).