Homework #3

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   1.2.1      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
```

Problem 1

Part a

```
# Load the dataset; assign the dataframe to object called 'fatal_accidents'
fatal_accidents <- read.csv('fatal accidents.csv')
# Mutate is the most efficient tidyverse function which can add a new column
fatal_accidents <- group_by(fatal_accidents, Case.number)%>%
  mutate(People.count=sum(People.count.IN, People.count.OUT))
head(data.frame(fatal_accidents))
```

```
##                 State Case.number Vehicle.count People.count.IN
## 1 District of Columbia      110001             1               1
## 2 District of Columbia      110002             1               1
## 3 District of Columbia      110003             1               1
## 4 District of Columbia      110004             2               2
## 5 District of Columbia      110005             1               2
## 6 District of Columbia      110006             4               7
##   People.count.OUT Day Month Year Day.of.week Hour Minute People.count
## 1                1  11     2 2019           2   23     34            2
## 2                1  20     2 2019           4   18     25            2
## 3                1   5     3 2019           3   21      1            2
## 4                0  13     5 2019           2    5     19            2
## 5                0   4     8 2019           1    4      7            2
## 6                0   5     4 2019           6    2     45            7
```

Part b

```
group_by(fatal_accidents, State)%>%
  summarize(AvgNumofVehicles = mean(Vehicle.count),
            AvgNumofPeople = mean(People.count))
```

```
## # A tibble: 5 x 3
##   State                AvgNumofVehicles AvgNumofPeople
##   <chr>                           <dbl>          <dbl>
## 1 District of Columbia             1.55           2.95
## 2 Maryland                         1.64           2.59
## 3 North Carolina                   1.54           2.34
## 4 Virginia                         1.51           2.28
## 5 West Virginia                    1.50           2.38
```

Part c

```
group_by(fatal_accidents, State)%>%
  summarize(MinNumofVehicles = min(Vehicle.count),
            AvgNumofVehicles = mean(Vehicle.count),
            MaxNumofVehicles = max(Vehicle.count))
```

```
## # A tibble: 5 x 4
##   State                MinNumofVehicles AvgNumofVehicles MaxNumofVehicles
##   <chr>                           <int>            <dbl>             <int>
## 1 District of Columbia                1             1.55                 4
## 2 Maryland                            1             1.64                12
## 3 North Carolina                      1             1.54                 7
## 4 Virginia                            1             1.51                 8
## 5 West Virginia                       1             1.50                 5
```

Part d

From the tables created in part (b) and (c), one can observe that most car accidents taking place in 2019 across all 5 states involved only one or two cars and no more than 3 people on average. While we might assume that massive pile-ups and car accidents are more common when it comes to traffic fatalities because of news coverage or fear, these are illustrated above to be more rare. Smaller accidents which result in fatalities are far more common.

Part e

```
fatal_accidents %>%
  filter(State == "Virginia") %>%
  group_by(Month) %>%
  summarise(NumofAccidents = n())
```

```
## # A tibble: 12 x 2
##    Month NumofAccidents
##    <int>          <int>
##  1     1             63
##  2     2             55
##  3     3             57
##  4     4             60
##  5     5             66
##  6     6             62
##  7     7             55
```

```
##  8     8             69
##  9     9             79
## 10    10             78
## 11    11             70
## 12    12             60
```

Part f

```
fatal_accidents %>%
  filter(State == "Virginia", Month == 6 | Month == 7 | Month == 8) %>%
  group_by(Day.of.week) %>%
  summarise(MedianNumofVehicles = median(Vehicle.count),
            AvgNumofVehicles = mean(Vehicle.count))
```

```
## # A tibble: 7 x 3
##   Day.of.week MedianNumofVehicles AvgNumofVehicles
##         <int>               <dbl>            <dbl>
## 1           1                   1             1.54
## 2           2                   1             1.62
## 3           3                   2             1.61
## 4           4                   1             1.39
## 5           5                   1             1.54
## 6           6                   1             1.29
## 7           7                   1             1.78
```

Part g

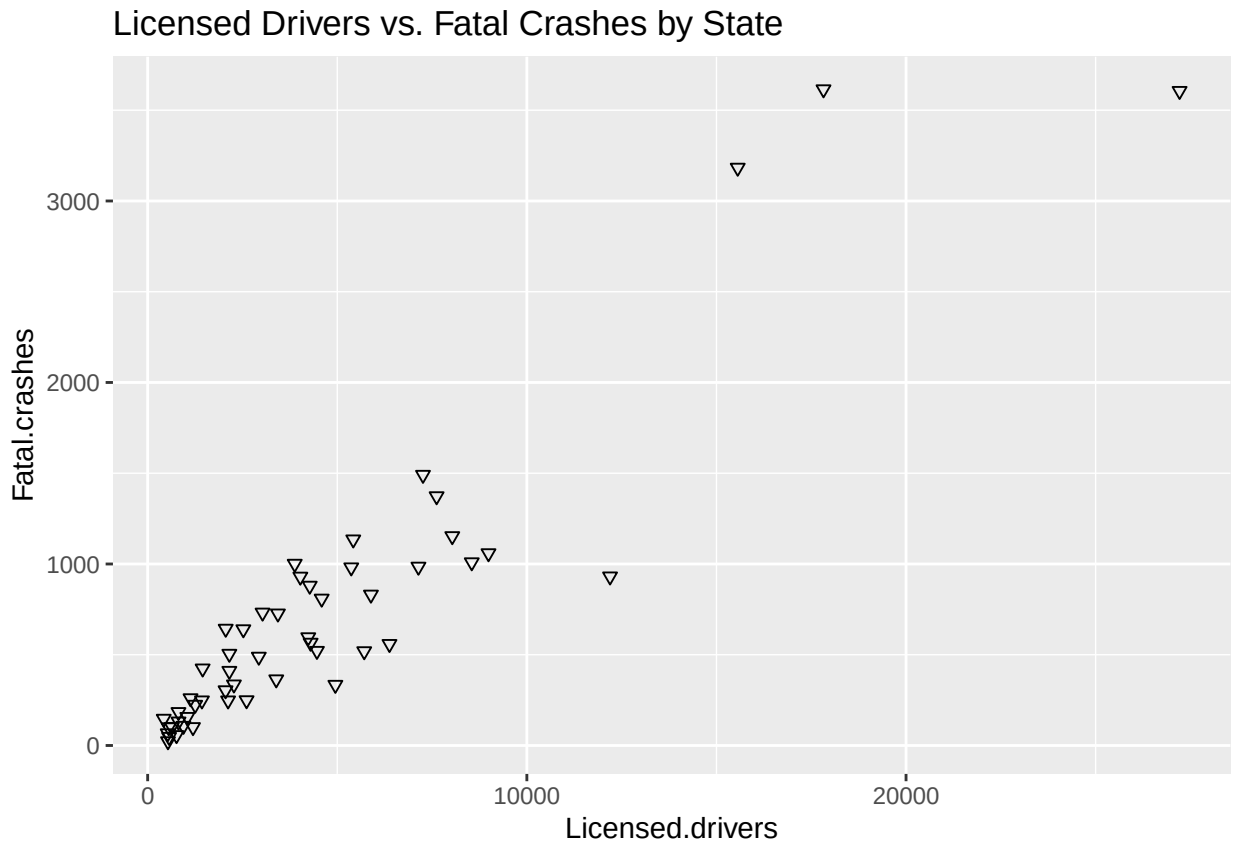
From the tables created in part (e) and (f), one can observe that the end of the year often sees the most fatal car accidents (late summer, early fall) in Virginia. The least amount of fatal accidents commonly occur at the beginning of the year. Similar to the averages across all states, Virginia follows the trend of accidents involving no more than 2 cars.

Problem 2

```
# Load the dataset; assign the dataframe to object called 'state_crashes'
state_crashes <- read.csv('state crashes.csv')
```

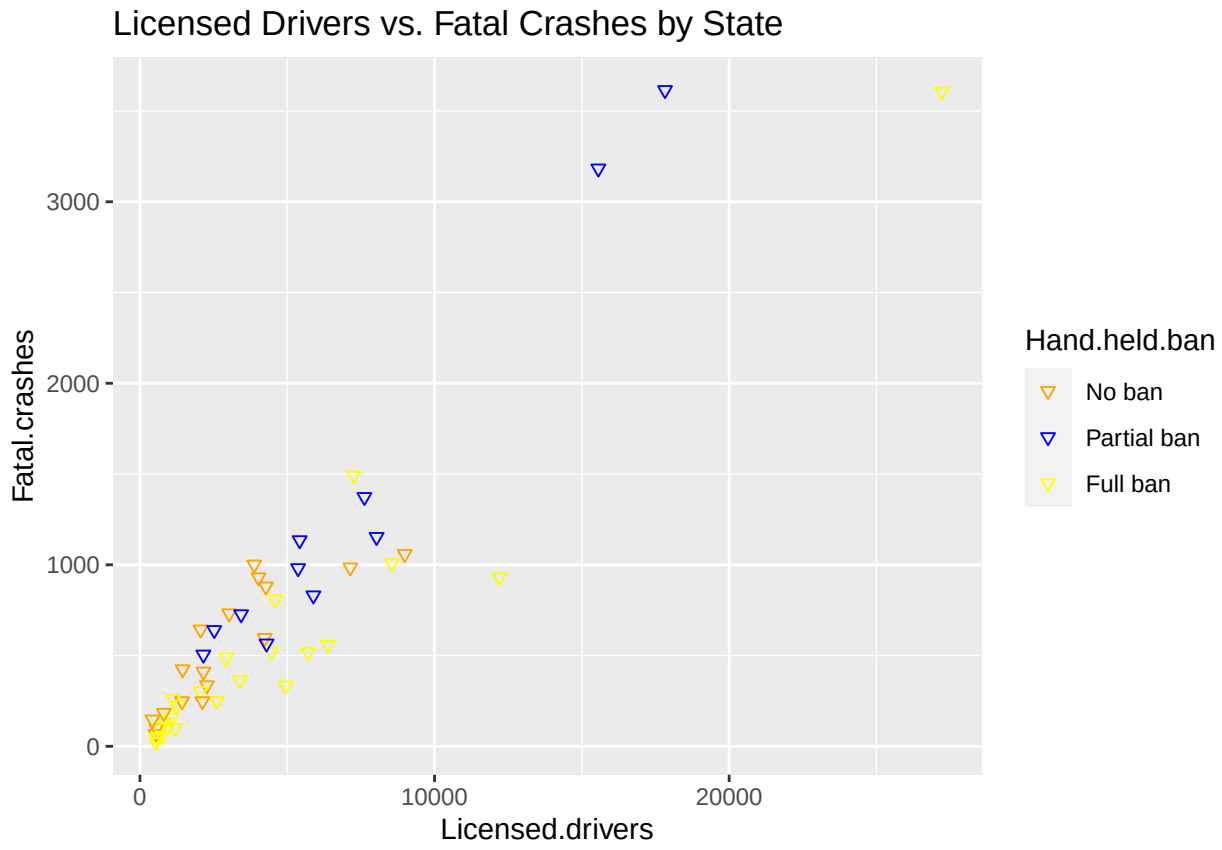
Part a

```
ggplot(state_crashes, aes(x=Licensed.drivers, y=Fatal.crashes)) +
  geom_point(shape = 6) +
  labs(title = "Licensed Drivers vs. Fatal Crashes by State")
```



Part b

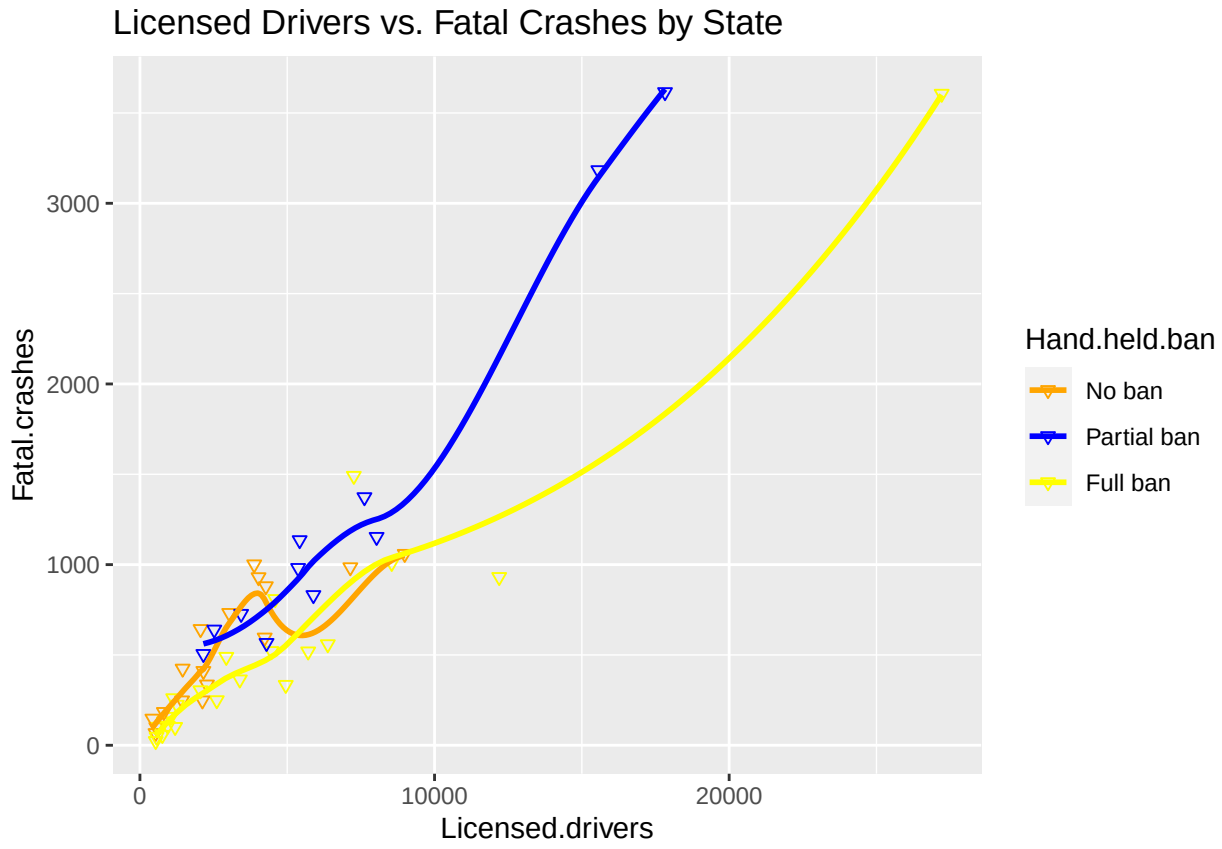
```
state_crashes$Hand.held.ban <- factor(state_crashes$Hand.held.ban,
                                      labels=c("No ban","Partial ban", "Full ban"))

ggplot(state_crashes, aes(x=Licensed.drivers, y=Fatal.crashes, color=Hand.held.ban)) +
  geom_point(shape = 6) +
  scale_color_manual(values = c("orange","blue", "yellow")) +
  labs(title = "Licensed Drivers vs. Fatal Crashes by State")
```



Part c

```
ggplot(state_crashes, aes(x=Licensed.drivers, y=Fatal.crashes, color=Hand.held.ban)) +
  geom_point(shape = 6) +
  scale_color_manual(values = c("orange","blue", "yellow")) +
  geom_smooth(se=F) +
  labs(title = "Licensed Drivers vs. Fatal Crashes by State")
```

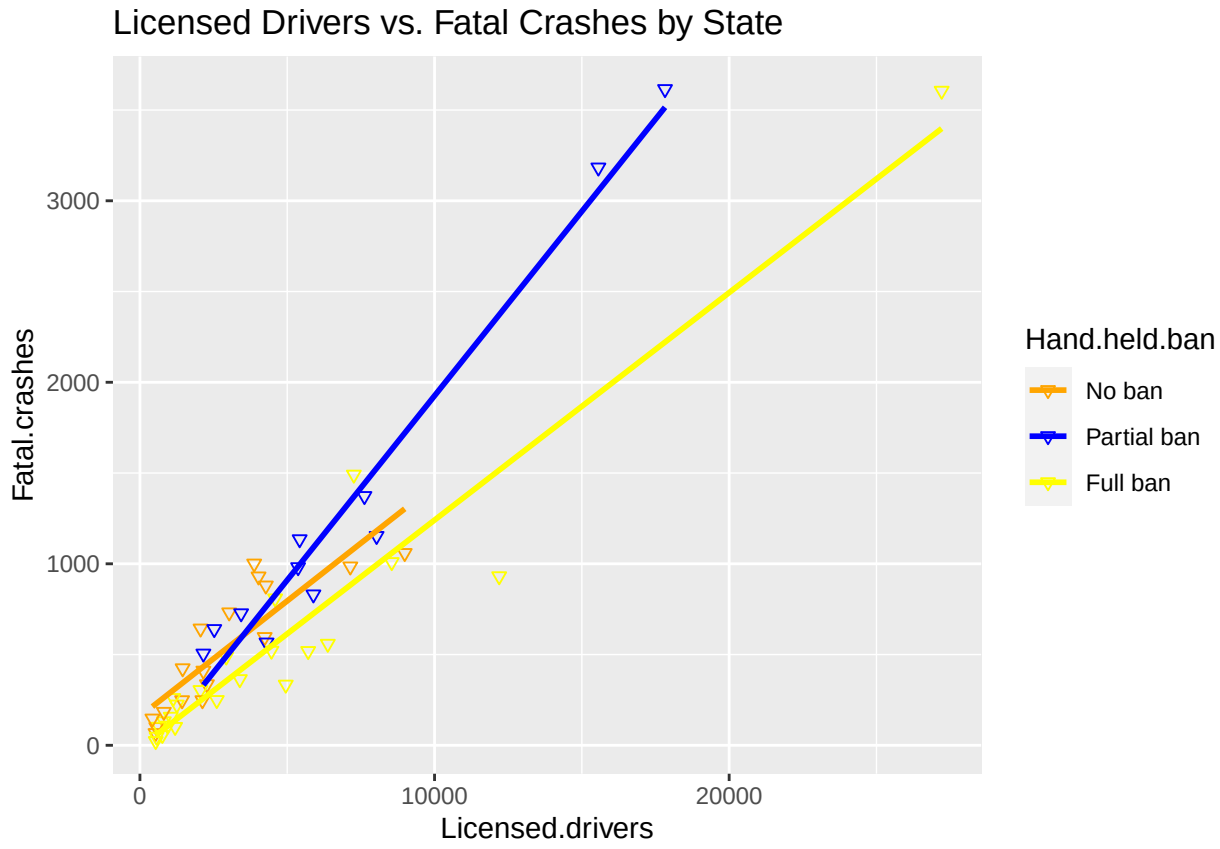
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Part d

```
ggplot(state_crashes, aes(x=Licensed.drivers, y=Fatal.crashes, color=Hand.held.ban)) +
  geom_point(shape = 6) +
  scale_color_manual(values = c("orange","blue", "yellow")) +
  geom_smooth(method=lm, se=F) +
  labs(title = "Licensed Drivers vs. Fatal Crashes by State")
```

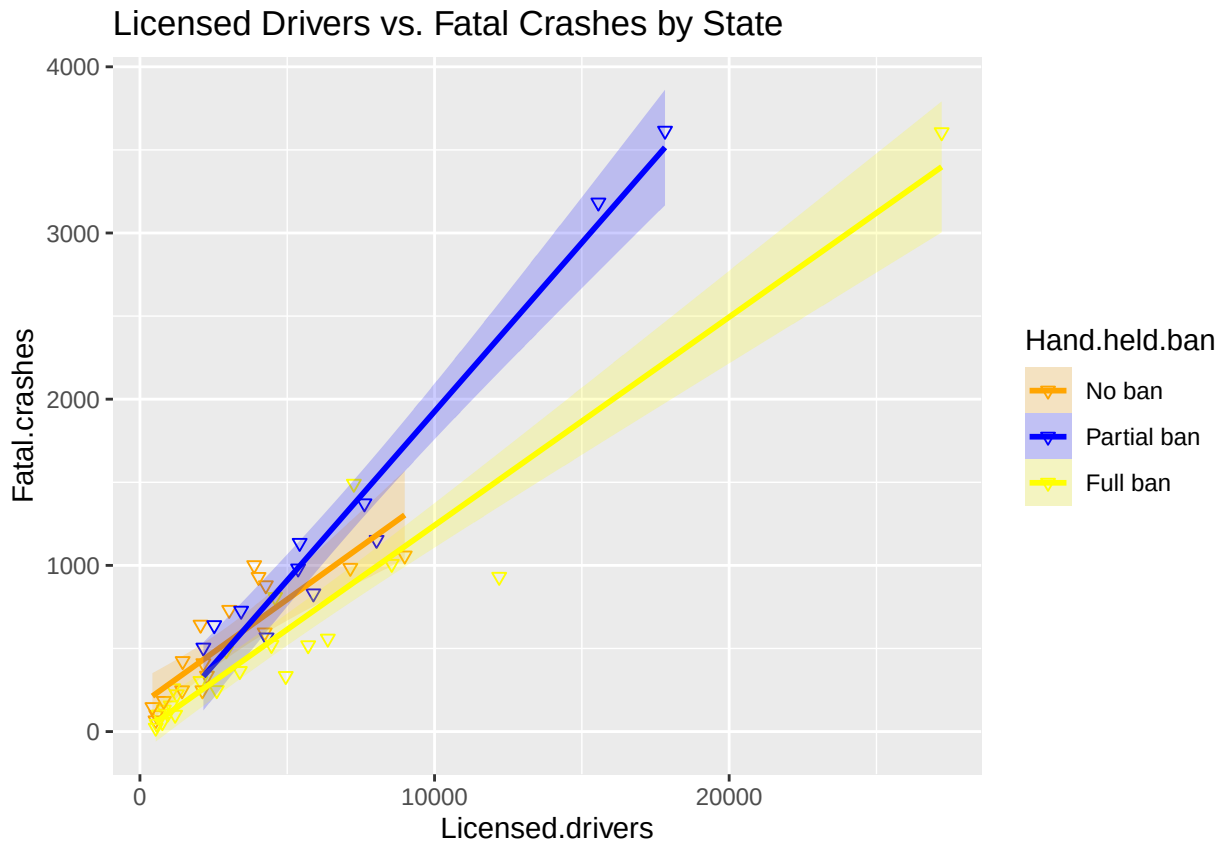
```
## `geom_smooth()` using formula = 'y ~ x'
```



Part e

```
ggplot(state_crashes, aes(x=Licensed.drivers, y=Fatal.crashes, color=Hand.held.ban)) +
  geom_point(shape = 6) +
  geom_smooth(method=lm, aes(fill=Hand.held.ban), alpha=0.2) +
  scale_color_manual(values = c("orange","blue", "yellow")) +
  scale_fill_manual(values = c("orange","blue", "yellow")) +
  labs(title = "Licensed Drivers vs. Fatal Crashes by State")
```

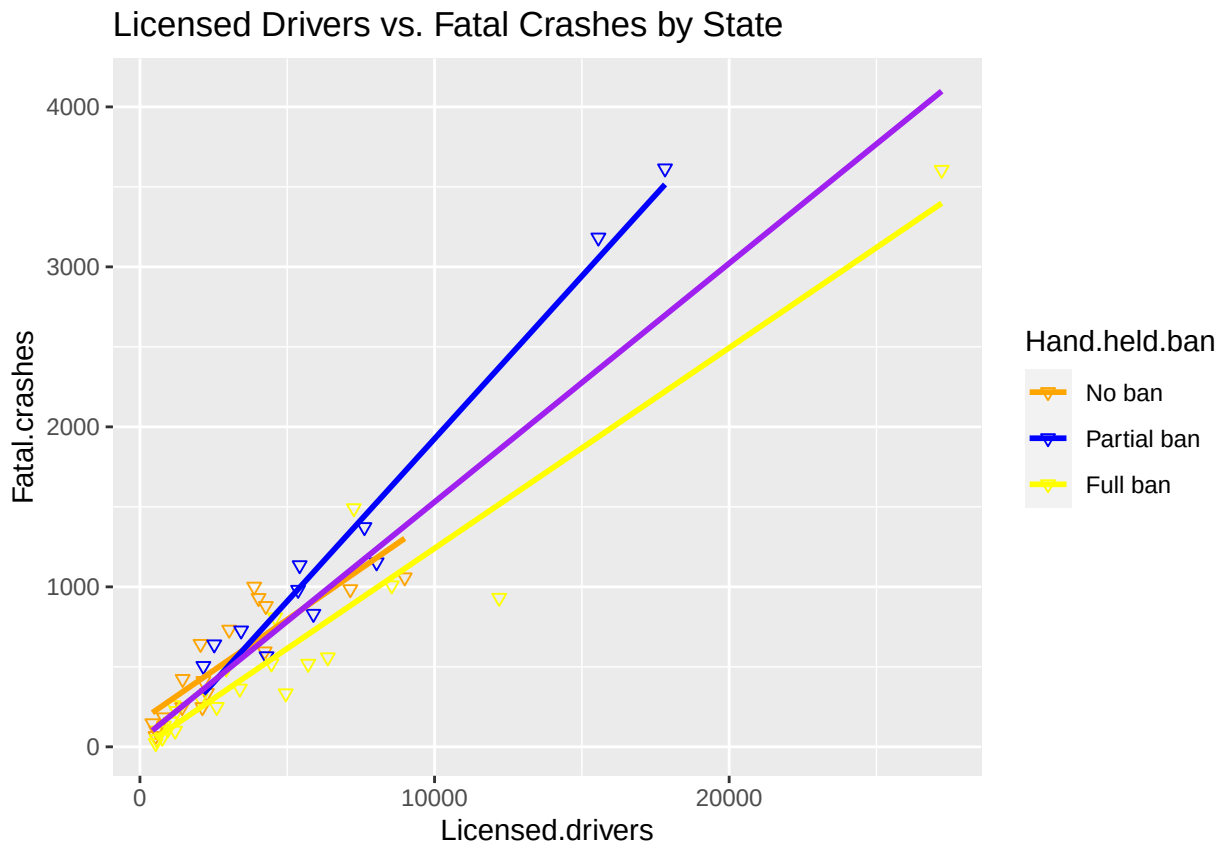
```
## `geom_smooth()` using formula = 'y ~ x'
```



Part f

```
#The purple regression line describes all states
ggplot(state_crashes, aes(x=Licensed.drivers, y=Fatal.crashes, color=Hand.held.ban)) +
  geom_point(shape = 6) +
  scale_color_manual(values = c("orange","blue", "yellow")) +
  geom_smooth(method=lm, se=F) +
  geom_smooth(method=lm, se=F, color='purple') +
  labs(title = "Licensed Drivers vs. Fatal Crashes by State")
```

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



Part g

When considering the information presented in the previous graphs, the most basic conclusion holds that more licensed drivers per state has a positive relationship with fatal crashes. While one cannot say that more licensed drivers causes more fatal crashes, there is a clear trend between the two variables. Additionally, states with a full ban on hand-held devices

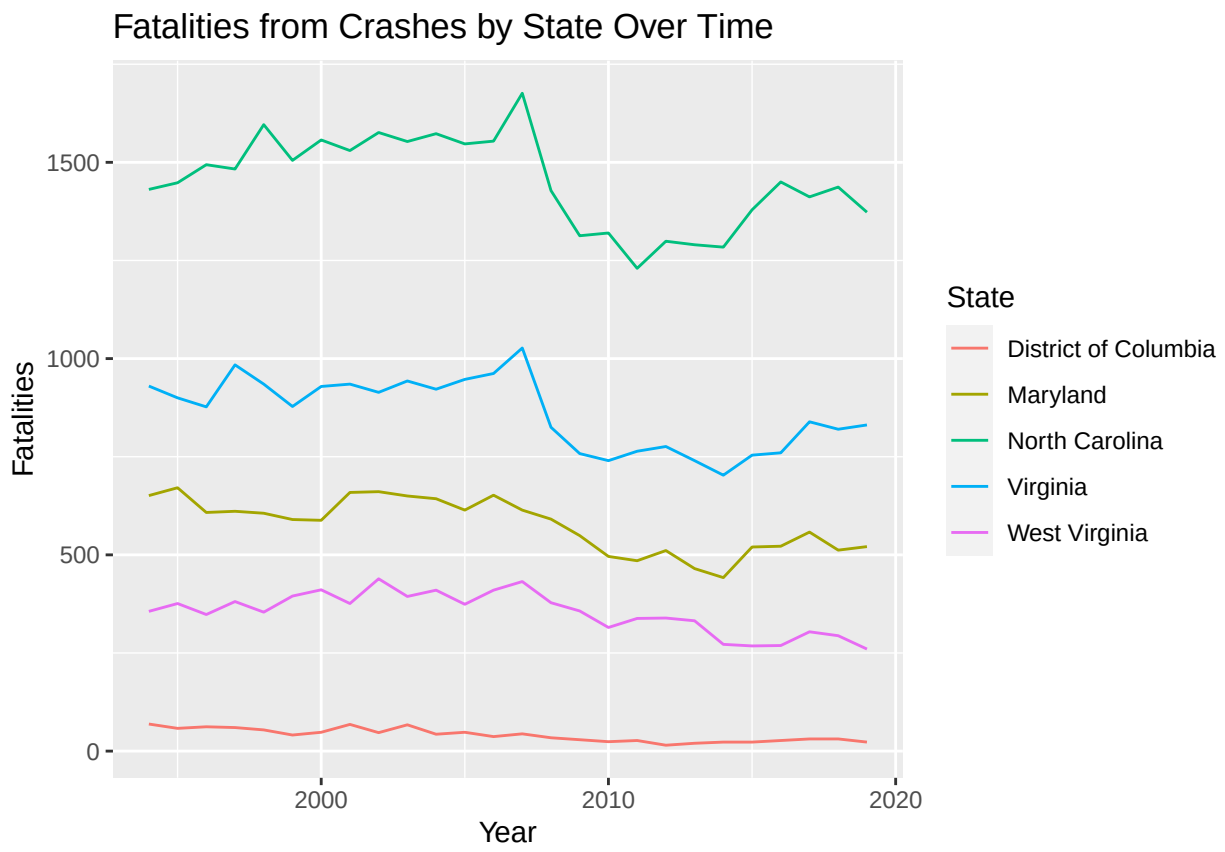
see a lower rate of fatal crashes regardless of the amount of licensed drivers (when compared to no and partial bans, as well as the average regression line between all states). Again, causation cannot be derived from these graphs, but clear relationships can be denoted.

Problem 3

```
# Load the dataset; assign the dataframe to object called 'fatalities'
fatalities <- read.csv('fatalities.csv')
```

Part a

```
ggplot(fatalities, aes(x=Year, y=Fatalities, color=State)) + geom_line() +
  labs(title = "Fatalities from Crashes by State Over Time")
```

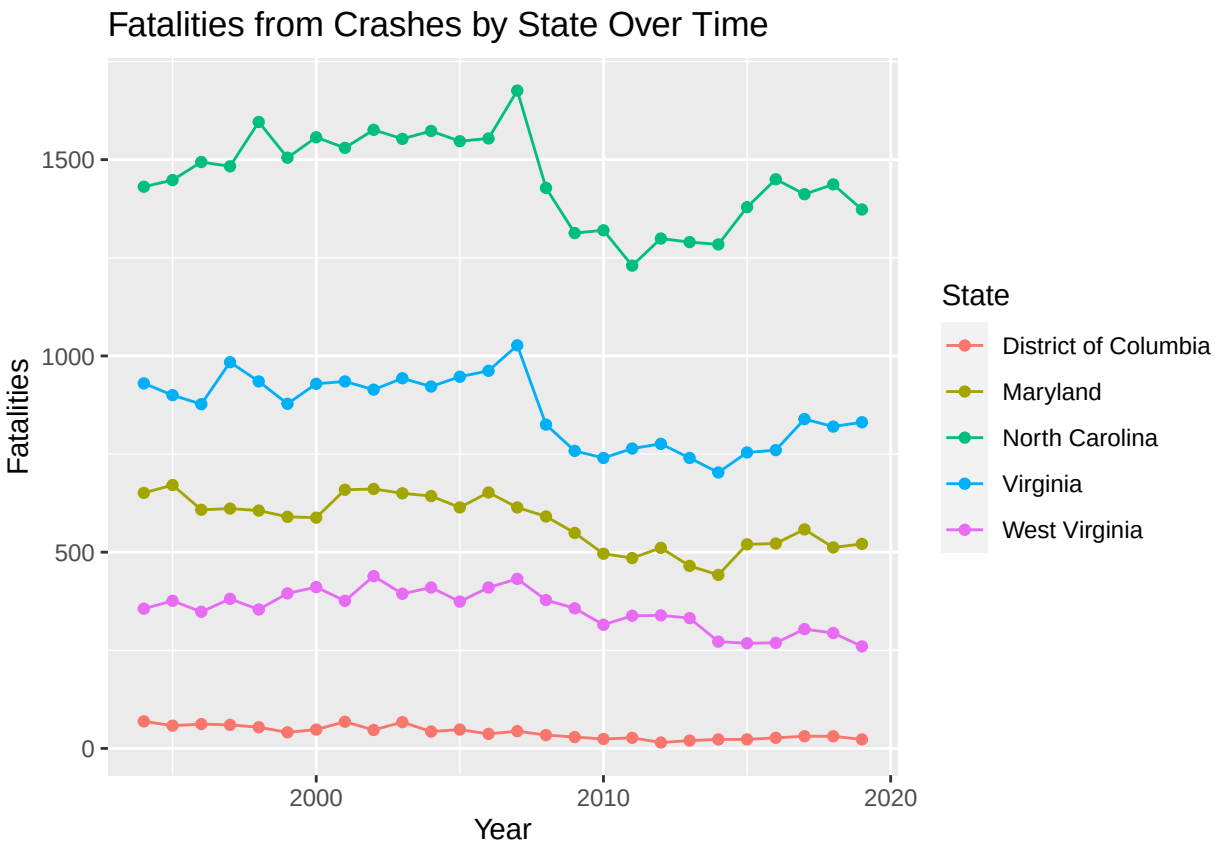


Part b

From the plot created in part (a), one can conclude that the states seem to have a somewhat similar trend over time. While the number of total fatalities per year vary widely by state, all followed a steady/constant rate until a moderate drop in fatalities occurred around 2008 which led to a downturn in car fatalities. Although the overall trend might be similar, there is a little variation in the recent number of fatalities: North Carolina, Virginia, and Maryland all seem to have faced a slight increase in fatalities while West Virginia and Washington D.C. have continued to decrease.

Part c

```
ggplot(fatalities, aes(x=Year, y=Fatalities, color=State)) +
  geom_line() +
  geom_point() +
  labs(title = "Fatalities from Crashes by State Over Time")
```

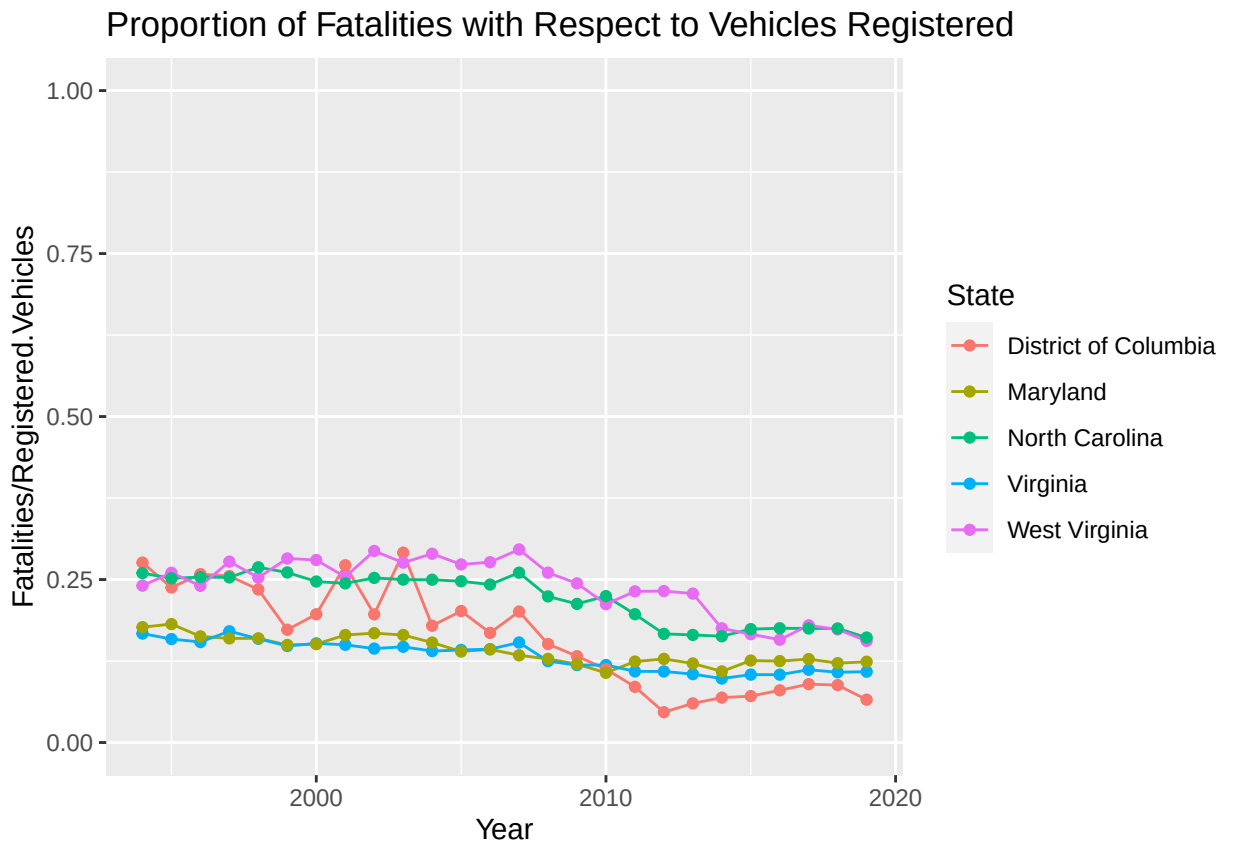


Part d

One advantage to using the plot created in part (c) instead of the plot created in part (a) is that specific points of downturn or increase can be pinpointed to a specific date/data point. In graph a, the conclusions made based on time are assumed because exact points are not visualized.

Part e

```
ggplot(fatalities, aes(x=Year, y=Fatalities/Registered.Vehicles, color=State)) +
  geom_line() + geom_point() +
  ylim(0, 1) +
  labs(title = "Proportion of Fatalities with Respect to Vehicles Registered")
```

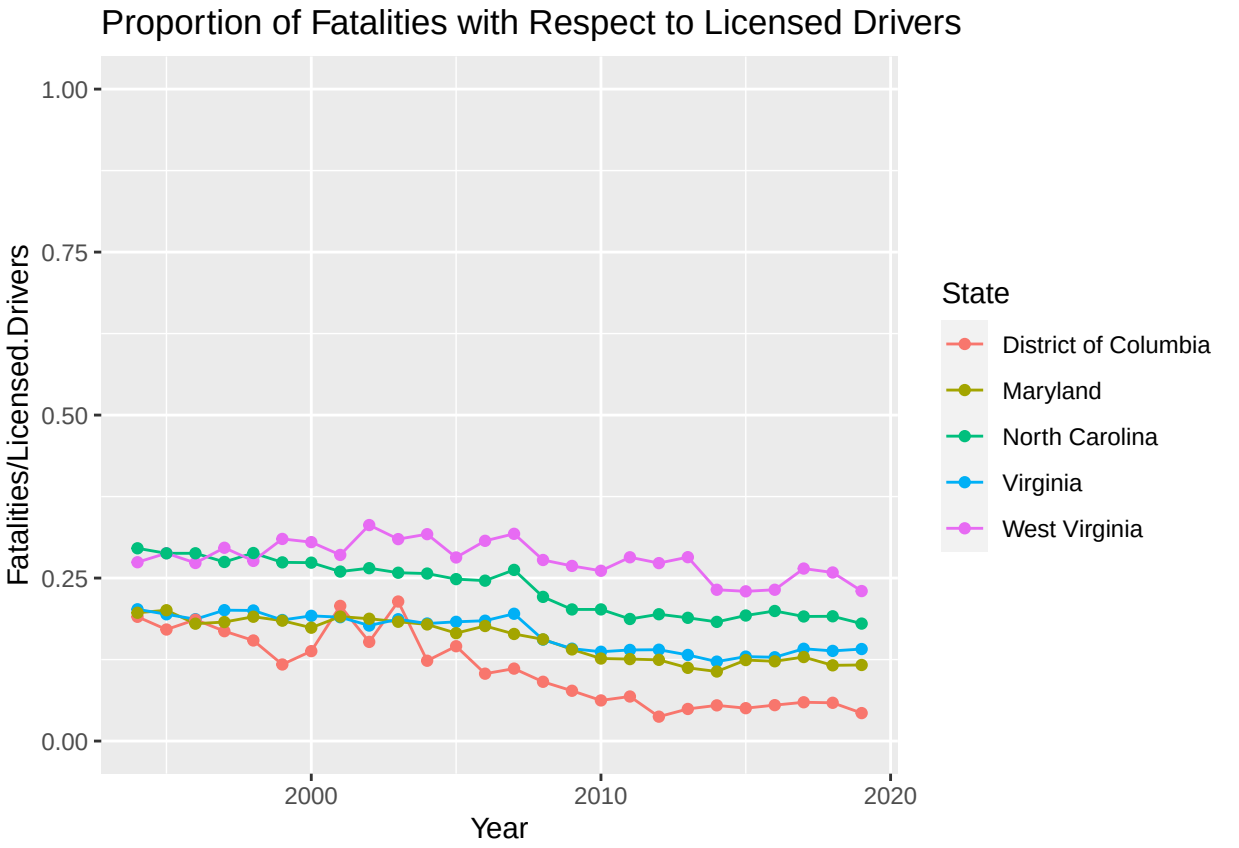


Part f

The proportion graph above provides a new understanding of the fatalities data. I had previously concluded that the trends between certain states are relatively similar but hold some variation in recent years. The plot in part (e), however, shows that when considering the amount of registered vehicles, the trend lines in fatalities all follow the same steady decline. The amount of fatalities is also not as great as originally assumed.

Part g

```
ggplot(fatalities, aes(x=Year, y=Fatalities/Licensed.Drivers, color=State)) +
  geom_line() +
  geom_point() +
  ylim(0, 1) +
  labs(title = "Proportion of Fatalities with Respect to Licensed Drivers")
```



Part h

As I had concluded in part (f), the new proportion graph does change some of my conclusions drawn in part (b). Not only are the proportions more similar than the raw counts of fatalities offered by the graph in part (b), but the data reflects a comparable trend in the decrease of fatalities across states. While the raw fatalities graph offers some valuable information, the two proportion graphs in part (f) and part (h) offer a more realistic understanding of the fatal car crash phenomenon.