Homework 3

Max Ryoo (hr2ee)

library(ggplot2) library(car) library(gcookbook) library(MASS) library(Hmisc)

Problem 1

A

```
setwd("/Users/maxryoo/Documents/Fall 2018/STAT3080/HW3")
nym2002 <- read.table("nym2002.txt", header=TRUE)
print(nym2002[1:5,])
```

```
##   place gender age home     time
## 1  3592   Male  52  GBR 217.4833
## 2 13853 Female  40   NY 272.5500
## 3 12256   Male  31  FRA 265.2833
## 4 10457 Female  33   MI 256.1500
## 5  9686   Male  33   NY 252.2500
```

I set my working directory as above which contains the file to be read in. I used the read.table to read in the file and stored it into a dataframe called nym2002. I proceeded to print out the first 5 rows of the dataframe.

B

```
num_finishers <- nrow(nym2002)
print(num_finishers)
```

```
## [1] 68
```

I used the nrow function to determine the number of finishers in the dataframe. I then stored it in num_finishers and printed the result.

C

```
youngest <- min(nym2002$age)
oldest <- max(nym2002$age)
print(youngest)
```

```
## [1] 25
```

```
print(oldest)
```

```
## [1] 71
```

I used the min and max function on the age column of the nym2002 dataframe. I then stored it into youngest and oldest respectivly, which I printed to be 25 and 71 years of age.

D

```
fastest <- nym2002[which(nym2002$time == min(nym2002$time)),]$age
slowest <- nym2002[which(nym2002$time == max(nym2002$time)),]$age
print(c("slowest: ", slowest ," fastest: ", fastest))
```

```
## [1] "slowest: "  "58"         " fastest: " "40"
```

I used the which function to find which entires had the min and max times. I then looked at that index and took only the age column. The slowest runner was 58 years old and the fastest was 40 years old.

E

```
us_finisher <- nym2002[which(nchar(as.character(nym2002$home)) ==2),]
num_us <- nrow(us_finisher)
print(num_us)
```

```
## [1] 38
```

I first took the home column and made it into a vector. I then used the nchar function to see the length of the characters since countries have 3 instead of 2. I then picked the entries that are states or us territories. I then took the number of rows and printed the result of 38.

F

```
oldest <- nym2002[which(nym2002$time == max(nym2002$time)),]$age
best_time_older <- min(nym2002[which(nym2002$age > oldest), ]$time)
old_position <- nym2002[which(nym2002$time == best_time_older), ]$place
print(old_position)
```

```
## [1] 15229
```

I first found the age of the slowest runner. I then found the fastest time from the people who were older than her (best_time_older). I searched for that time in the original nym2002 and subsetted the place, which I printed and saw was 15229.

G

```
position_fastest <- nym2002[which(nym2002$time == min(nym2002$time)),]$place
```

```
print(position_fastest)
```

```
## [1] 200
```

I found which entry has the minimum time. I then subsetted the place for the position, which I printed and found to be 200.

H

```
countries <- nym2002[which(nchar(as.character(nym2002$home)) == 3),]
unqiue_countries<- c(as.character(unique(countries$home)), "US")
print(length(unqiue_countries))
```

```
## [1] 13
```

I first got all entries that have countries as a home in the dataset (countries are three characters). I then made it into a vector using as.character function for the countries and added "US" for the states. I then printed the unique countries.

Problem 2

A

```
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```
library(gcookbook)
```

```
##
## Attaching package: 'gcookbook'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     countries
```

```
library(MASS)
library(Hmisc)
```

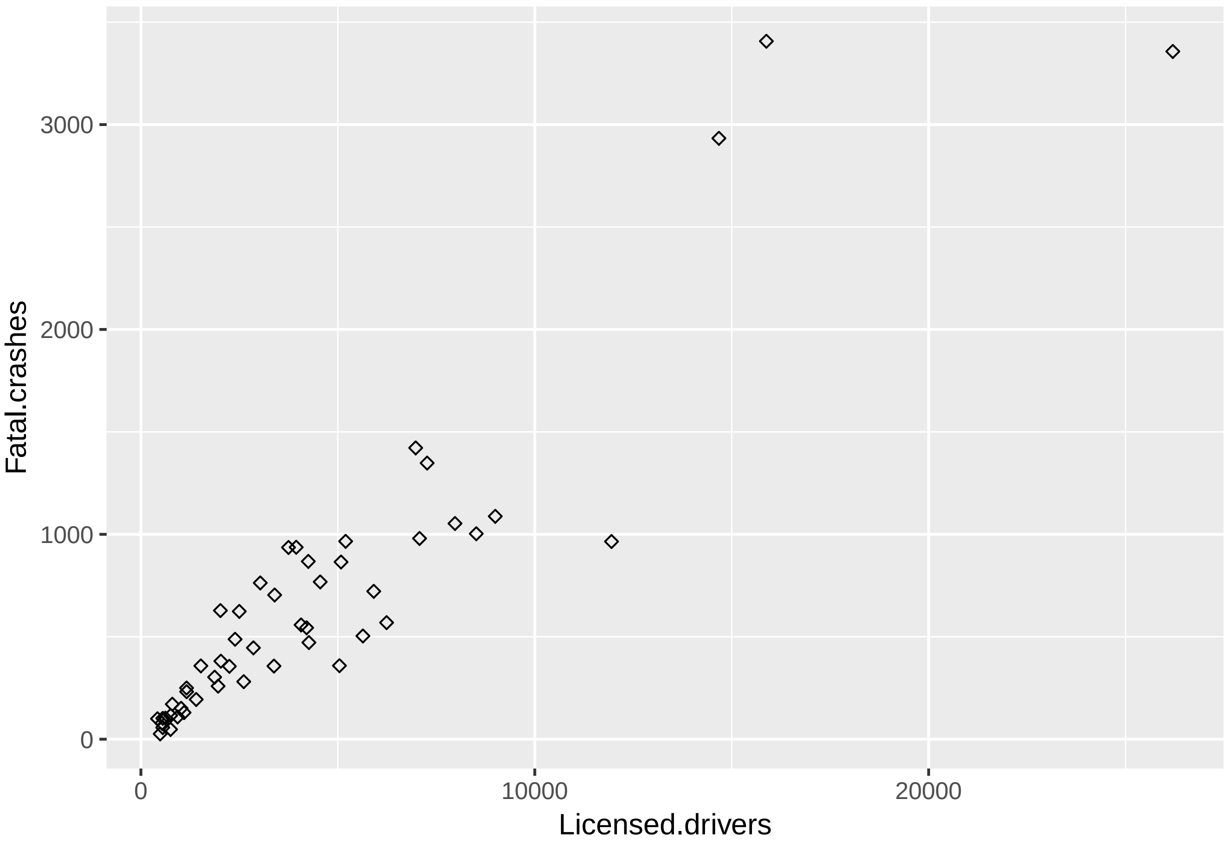
```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

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

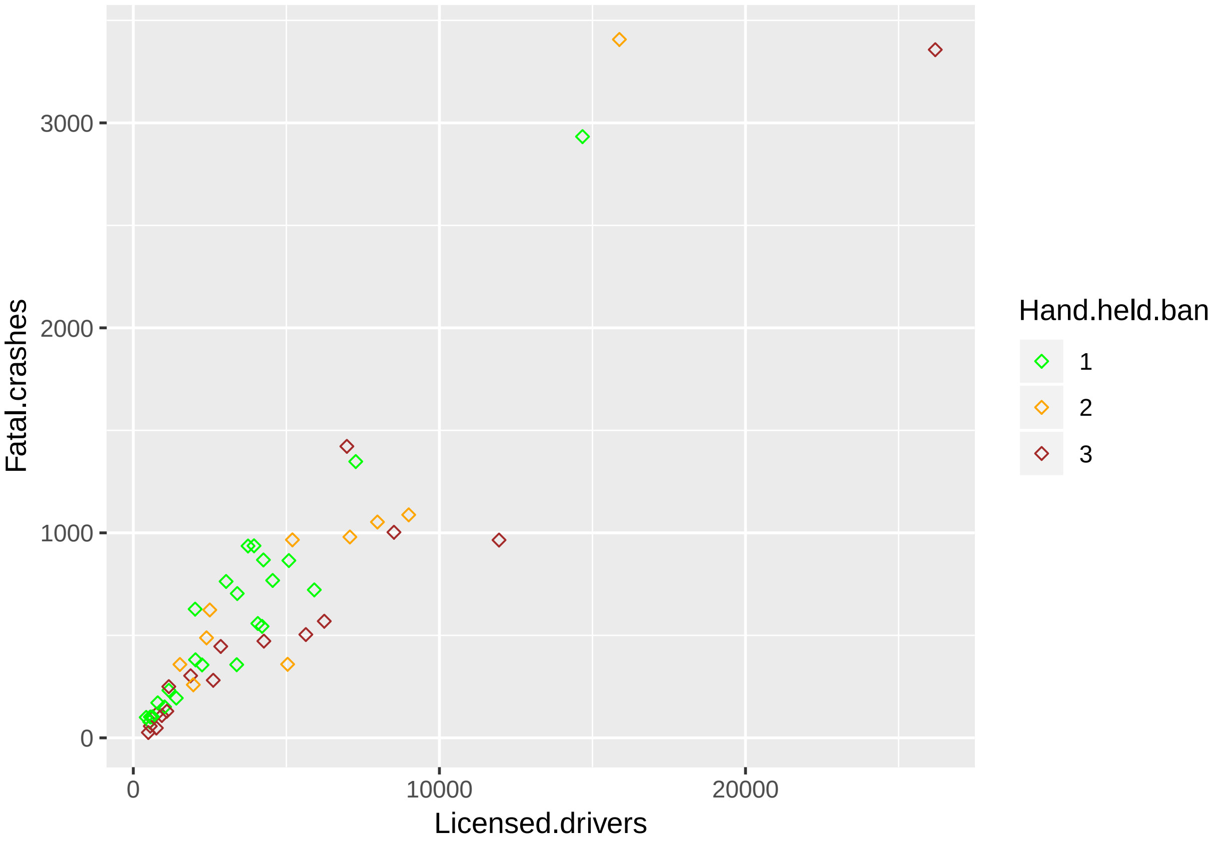
```
crash <- read.csv("state crashes.csv")
crash_plot <- ggplot(crash, aes(x=Licensed.drivers,
y=Fatal.crashes)) + geom_point(shape=23)
print(crash_plot)
```



I read in the csv and made a scatter plot through ggplot2. I made the shape a diamond with no fill in.

B

```
crash_plot <- ggplot(crash, aes(x=Licensed.drivers, y=Fatal.crashes,
colour=as.character(Hand.held.ban)))
crash_handheld <- crash_plot + geom_point(
shape=23) + scale_colour_manual("Hand.held.ban",
values=c("1"="green","2"="orange", "3"="brown"))
print(crash_handheld)
```

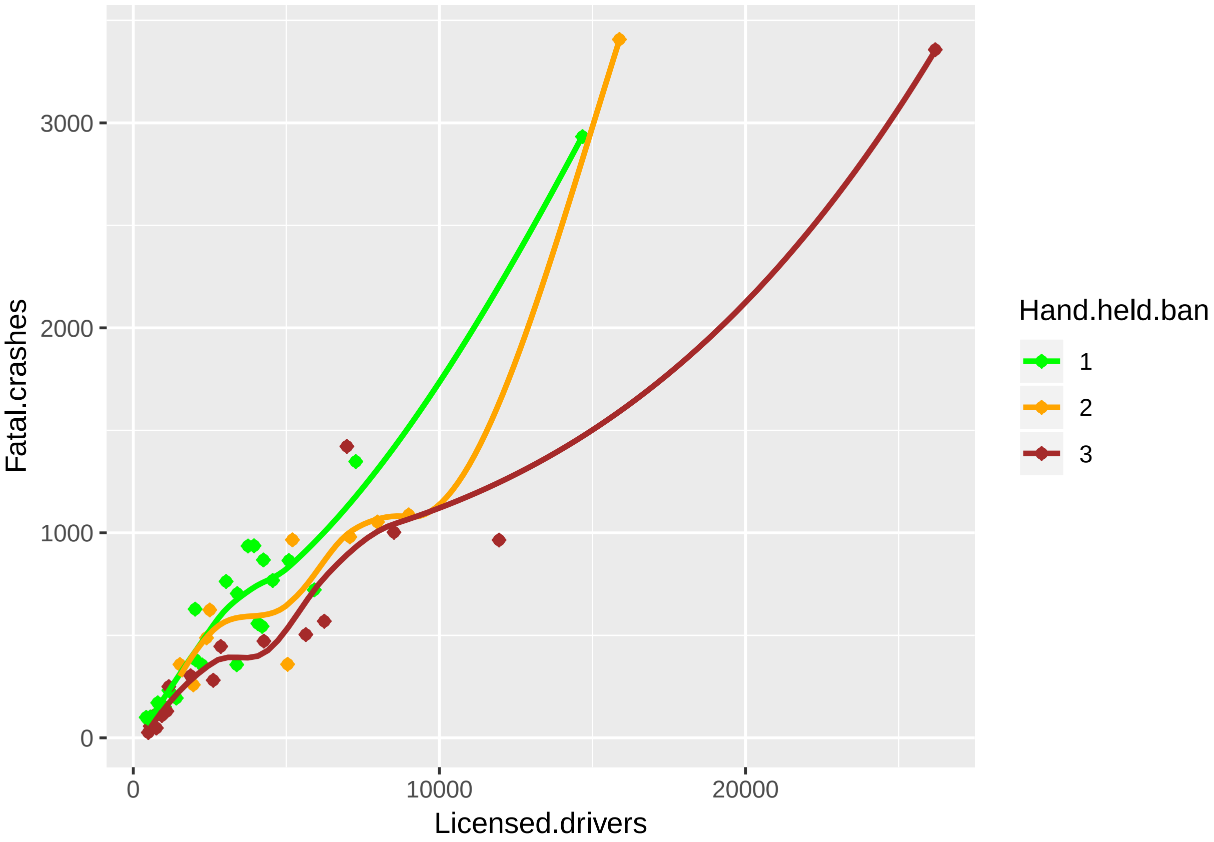


I made the crash_plot display by colours and changed the settings in the sacle_colour_manual function. I had to change the Hand.held.ban column into characters to show correctly for the scale_colour_manual function.

C

```
smooth_crash <- crash_handheld + geom_point(
) + geom_smooth(se=F)
print(smooth_crash)
```

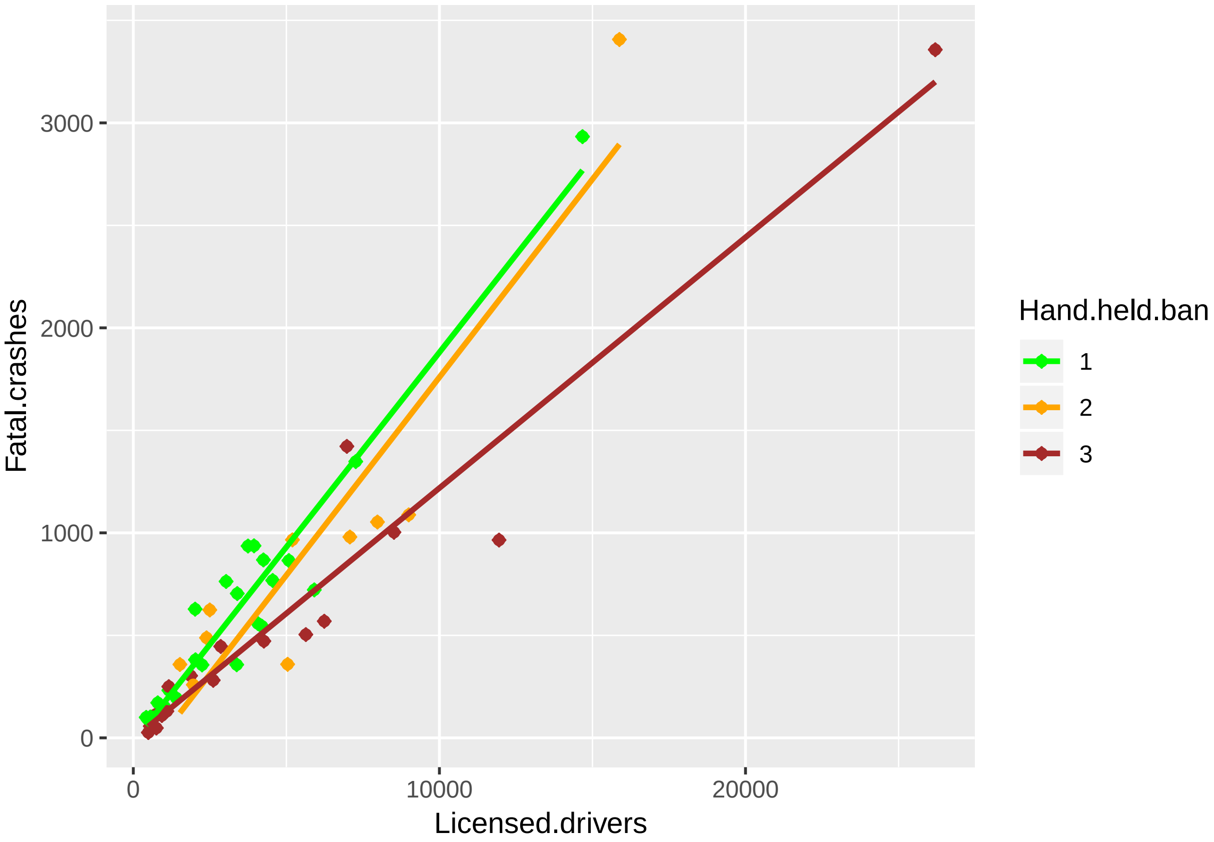
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



I added a smooth line to crash_handheld plot with the function geom_smooth. I added se=F to get rid of the confidence interval.

D

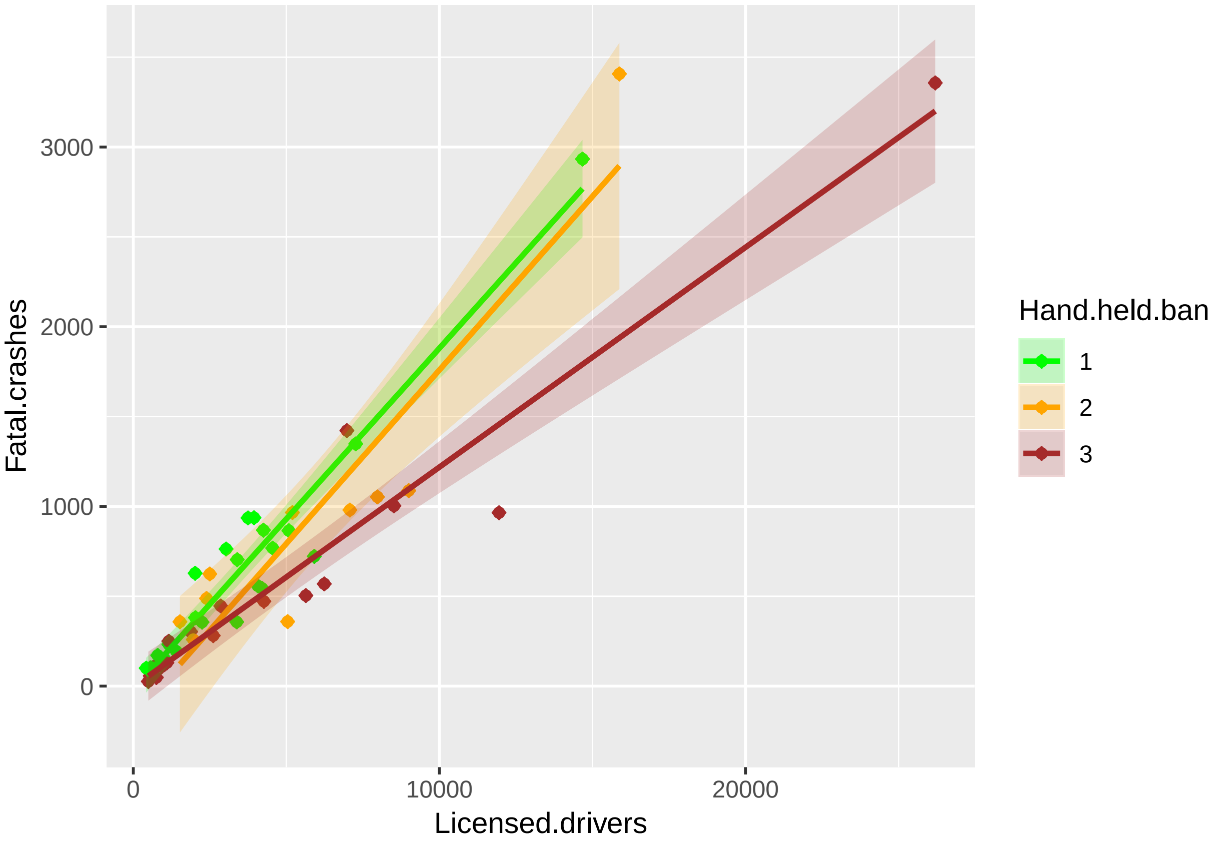
```
smooth_crash_reg <- crash_handheld+ geom_point() + geom_smooth(method=lm, se=F)
print(smooth_crash_reg)
```



I first loaded the crash_handheld plot. I then added the layer of geom_smooth. I added se=F to get rid of the confidence interval. I added method=lm for the regression linear model. I then printed the result.

E

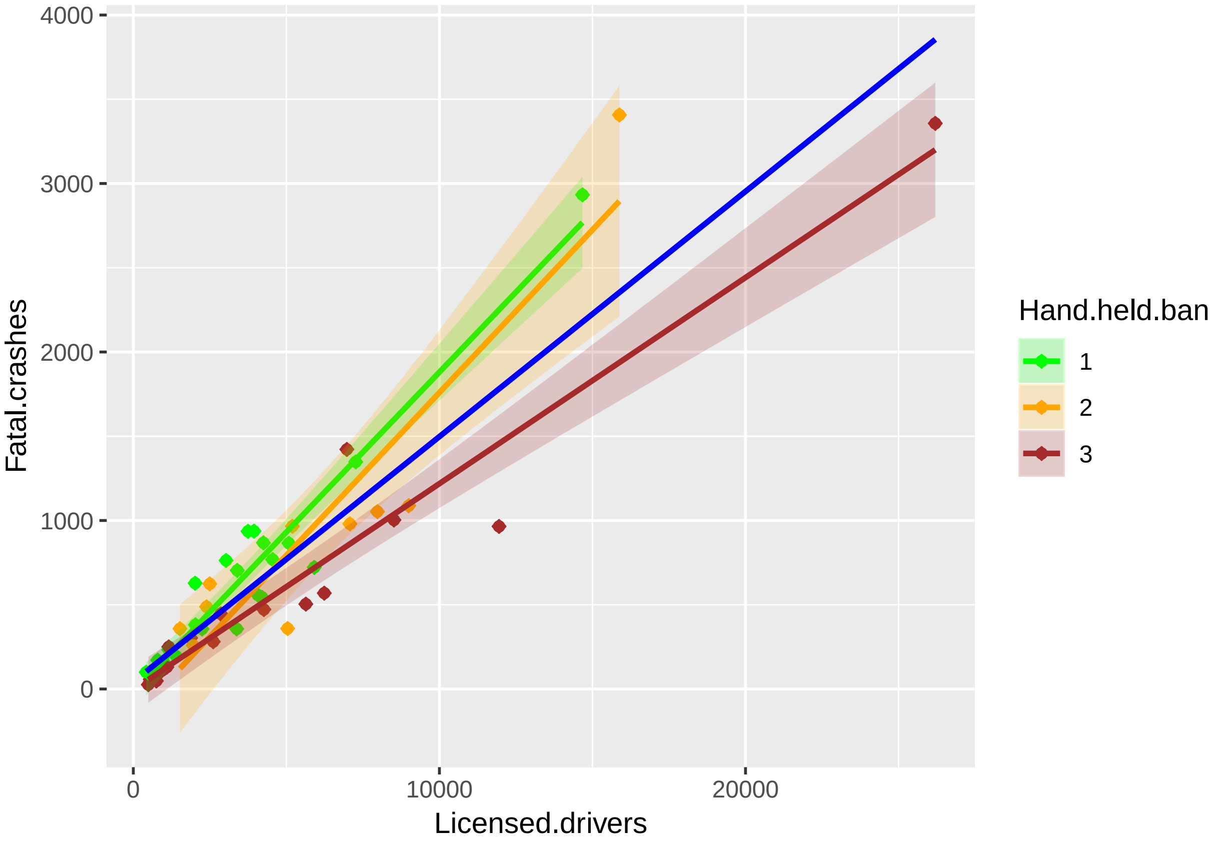
```
smooth_crash1 <- crash_handheld + geom_point(
) + geom_smooth(aes(fill=as.character(Hand.held.ban)
),method=lm,alpha=0.2) + scale_fill_manual("Hand.held.ban",
values=c("1"="green","2"="orange", "3"="brown"))
print(smooth_crash1)
```



I first loaded the previous graphs. I added alpha=0.2 for the transparency of the confidence interval. I then set the fill of those confidence intervals with scale_fill_manual with the same parameters as the scale_colour_manual previously used. I also added aes parameter of fills to add another fill color to the original graph.

F

```
all_lm <- smooth_crash1 + geom_smooth(method = 'lm',
aes(colour = NA), colour = 'blue', se=F)
print(all_lm)
```



I first loaded the smooth_crash1 plot from the previous problem. I proceeded to get the linaer model where there was no colour differentiation of Hand.held.ban. I got rid of the confidence as well. I then stored that in all_lm and printed the plot.

G

You can conclude that the more licensed drivers there are the more fata crashes there were. The relationship seemed to have a strong linear relationship. However, this is no link to causation. Also we can conclude that if a handheld device was given then there occured more fatal crashses.

References

1. <https://stats.stackexchange.com/questions/5253/ how-do-i-get-the-number-of-rows-of-a-data-frame-in-r>
2. <https://stackoverflow.com/questions/40003028 /extracting-unique-values-from-data-frame-using-r>
3. <http://sape.inf.usi.ch/quick-reference/ggplot2/shape
4. <https://stackoverflow.com/questions/ 38788357/change-bar-plot-colour-in-geom-bar-with-ggplot2-in-r>

5. <https://stackoverflow.com/questions/16562859/ ggplot2-colour-geom-point-by-factor-but-geom-smooth-based-on-all-data>