Homework 2

Charles Ancel

2023-09-07

Homework Instructions

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

```
library(ggplot2)
```

We will continue analyzing the Seatbelts dataset that we considered in Homework 1. Below, we perform the same preprocessing steps to prepare the dataset for analysis. Use the created `sb` R object for the exercises.

```
sb = as.data.frame(Seatbelts)
sb$law = as.factor(sb$law)
```

Exercise 1: Formatting [5 points]

The first five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name in the document header
- properly assigned pages to exercises on Gradescope
- select page 1 (with your name) and the page for this exercise
- all code is printed and readable for each question
- generated a pdf file

Exercise 2: Adding Variables to a Dataset [10 points]

As a first step in analyzing the `sb` dataset for this assignment, we'll add a few variables to the dataset. We'll analyze these variables throughout this assignment.

Note that for this assignment, because `sb` is already a copy of the original `Seatbelts` dataset from R, it is appropriate to adjust the `sb` dataset directly rather than creating a newly named dataframe, as discussed in lecture.

part a

The `sb` dataset contains the number of car drivers killed (`DriversKilled`) and the number of car drivers killed or seriously injured (`drivers`). Suppose that instead we are interested in the number of car drivers seriously injured.

Add a new variable, `DriversInjured`, to the `sb` dataset for the number of car drivers seriously injured but not killed. You can calculate this variable using the two variables mentioned above.

```
sb$DriversInjured = sb$drivers - sb$DriversKilled
```

part b

Confirm that your variable has been correctly added to the dataframe. There are many ways this can be accomplished, so select one method and explain how you know that it worked.

```
# Displaying the first few rows of the dataset to confirm the addition of our new column
head(sb)
```

```
##   DriversKilled drivers front rear   kms PetrolPrice VanKilled law
## 1           107    1687   867  269  9059   0.1029718        12   0
## 2            97    1508   825  265  7685   0.1023630         6   0
## 3           102    1507   806  319  9963   0.1020625        12   0
## 4            87    1385   814  407 10955   0.1008733         8   0
## 5           119    1632   991  454 11823   0.1010197        10   0
## 6           106    1511   945  427 12391   0.1005812        13   0
##   DriversInjured
## 1           1580
## 2           1411
## 3           1405
## 4           1298
## 5           1513
## 6           1405
```

I know that we have correctly added the `DriversInjured` column to the `sb` dataframe using `sb.head()`; we can see that `DriversInjured` is also calculated properly as well. **Answer:**

part c

Check that the values for this variable are reasonable. Again, there are many ways this can be accomplished. Select a method to confirm the values are reasonable; if they aren't explain why they are not and attempt to fix them. If they are, explain how they are.

```
# Check for any rows where DriversInjured has negative values
negative_values <- sum(sb$DriversInjured < 0)

# Displaying the summary statistics for DriversInjured
summary_stats <- summary(sb$DriversInjured)

negative_values
```

```
## [1] 0
```

```
summary_stats
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     962    1358    1516    1548    1715    2456
```

Answer:

`negative_values` returns 0, it indicates there are no negative values, which is good. The `summary_stats` should give an overview of the distribution, including the minimum, median, mean, maximum, and quartiles.

part d

The `sb` dataset contains the number of passengers killed or seriously injured in the `front` and in the `rear` variables. Create a new variable that records the total number of passengers that are killed or seriously injured, regardless of where the passengers are seated. Call this variable `AllPassenger`. Confirm that this variable was added correctly to the dataset.

```
# Adding the new variable
sb$AllPassenger = sb$front + sb$rear

# Confirming the variable was added correctly by displaying the first few rows
head(sb[, c("front", "rear", "AllPassenger")])
```

```
##   front rear AllPassenger
## 1   867  269         1136
## 2   825  265         1090
## 3   806  319         1125
## 4   814  407         1221
## 5   991  454         1445
## 6   945  427         1372
```

Exercise 3: Visualizing and Interpreting Two Variables [20 points]

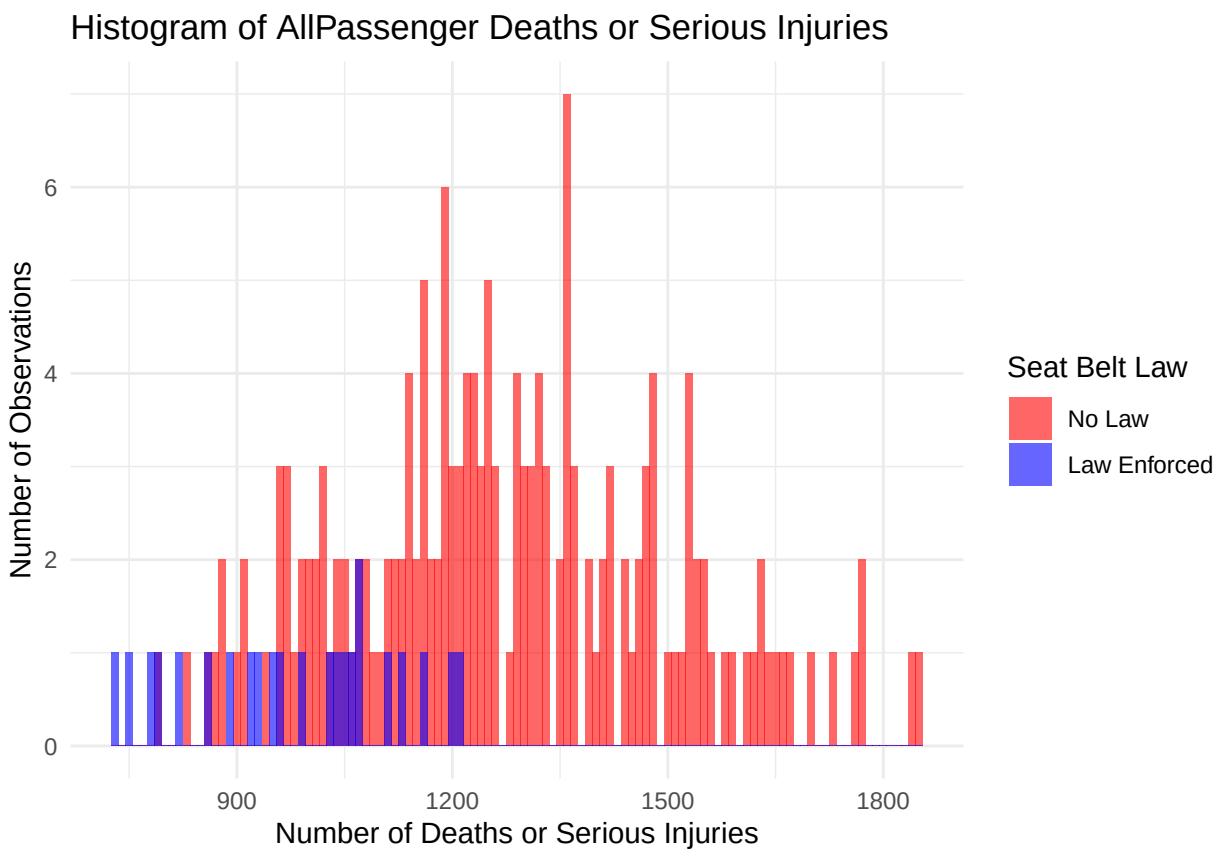
The Help file for the Seatbelts dataset indicates that the data were collected over a time when a compulsory seat belt law was introduced. We'll analyze differences related to this `law` variable throughout this homework assignment.

Previous studies have shown that seat belts helped reduce driver death and serious injury. Does this safety feature also have an affect on all passengers in the car?

part a

As a first look at understanding this relationship, create a histogram of the `AllPassenger` deaths or serious injuries. Be sure to include coloring that indicates whether the observation corresponds to a month with or a month without the compulsory seat belt `law`. Make sure that your histogram is clear and easy to read.

```
ggplot(data=sb, aes(x=AllPassenger, fill=factor(law))) +
  geom_histogram(binwidth=10, position="identity", alpha=0.6) +
  scale_fill_manual(values=c("red", "blue"), name="Seat Belt Law",
                    breaks=c(0, 1),
                    labels=c("No Law", "Law Enforced")) +
  labs(title="Histogram of AllPassenger Deaths or Serious Injuries",
       x="Number of Deaths or Serious Injuries",
       y="Number of Observations") +
  theme_minimal()
```



part b

What do you notice from this graph?

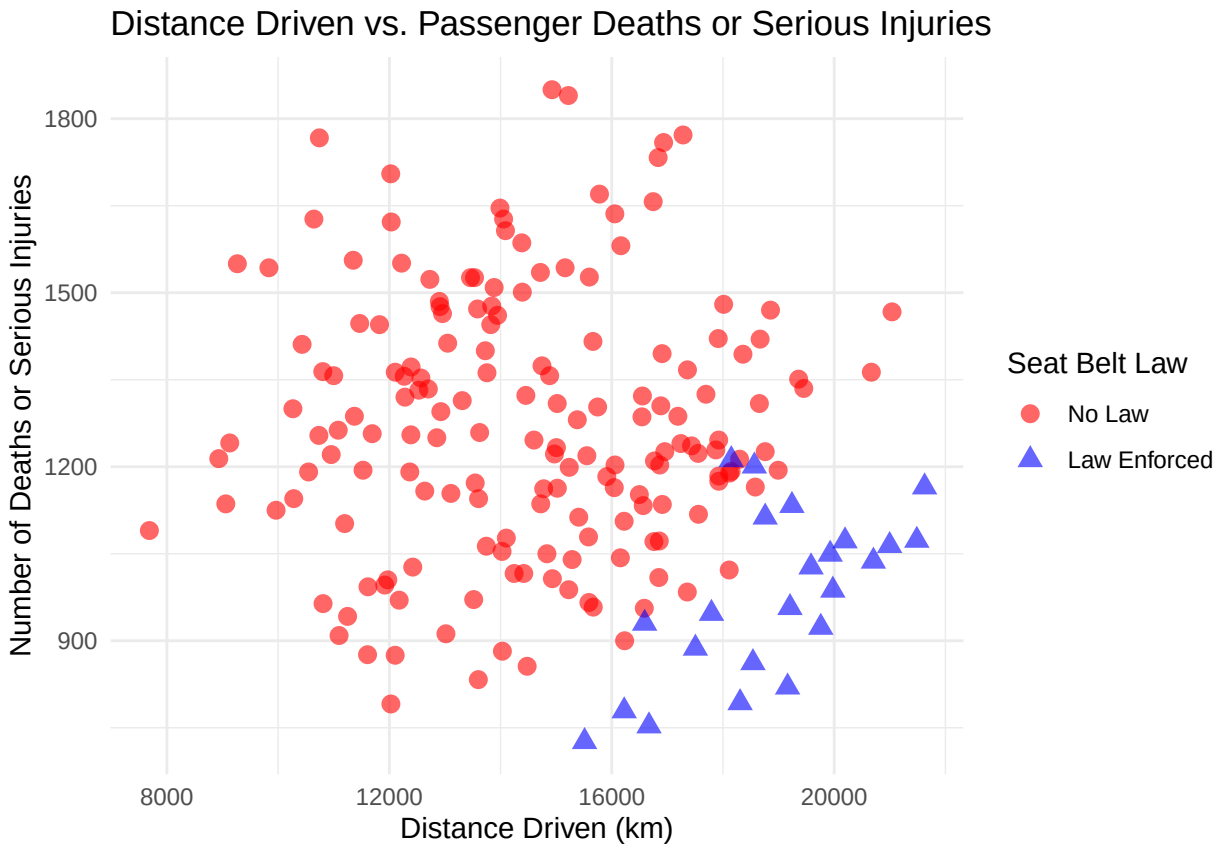
Answer: From the histogram, it's evident that the seatbelt law had a substantial positive effect, with most observations after the law's enforcement showing fewer than 1,200 deaths. This highlights the significant life-saving impact of the seatbelt legislation. However, while the visualization suggests a strong relationship, further statistical analyses would be essential to confirm the causal impact.

part c

One theory is that any differences in deaths or serious injuries could be related to the amount of driving that is done in any given month. To visualize this relationship, create a scatterplot of distance driven (`km`) vs. deaths or serious injuries of passengers (`AllPassenger`). Include the variable `law` in the color and shape of the points based on whether the law was active at the time. Make sure your scatterplot is clear and easy to read.

```
# Creating the scatterplot
ggplot(data=sb, aes(x=kms, y=AllPassenger, color=factor(law), shape=factor(law))) +
  geom_point(alpha=0.6, size=3) +
  scale_color_manual(values=c("red", "blue"), name="Seat Belt Law",
                     breaks=c(0, 1),
                     labels=c("No Law", "Law Enforced")) +
  scale_shape_manual(values=c(16, 17), name="Seat Belt Law",
                     breaks=c(0, 1),
                     labels=c("No Law", "Law Enforced")) +
  labs(title="Distance Driven vs. Passenger Deaths or Serious Injuries",
       x="Distance Driven (km)",
```

```
      y="Number of Deaths or Serious Injuries") +
  theme_minimal()
```



part d

What do you notice from this graph? Discuss any new or updated impressions about the data compared to **part b**, if any.

Answer: From the scatterplot, it's evident that the seatbelt law positively impacted passenger safety. Even during months with increased driving distances, passenger casualties remain lower when the law is in effect. Compared to part b, this graph adds clarity by showing that the reduction in deaths or injuries isn't merely due to reduced driving, further emphasizing the law's effectiveness.

Exercise 4: Line of Best Fit [10 points]

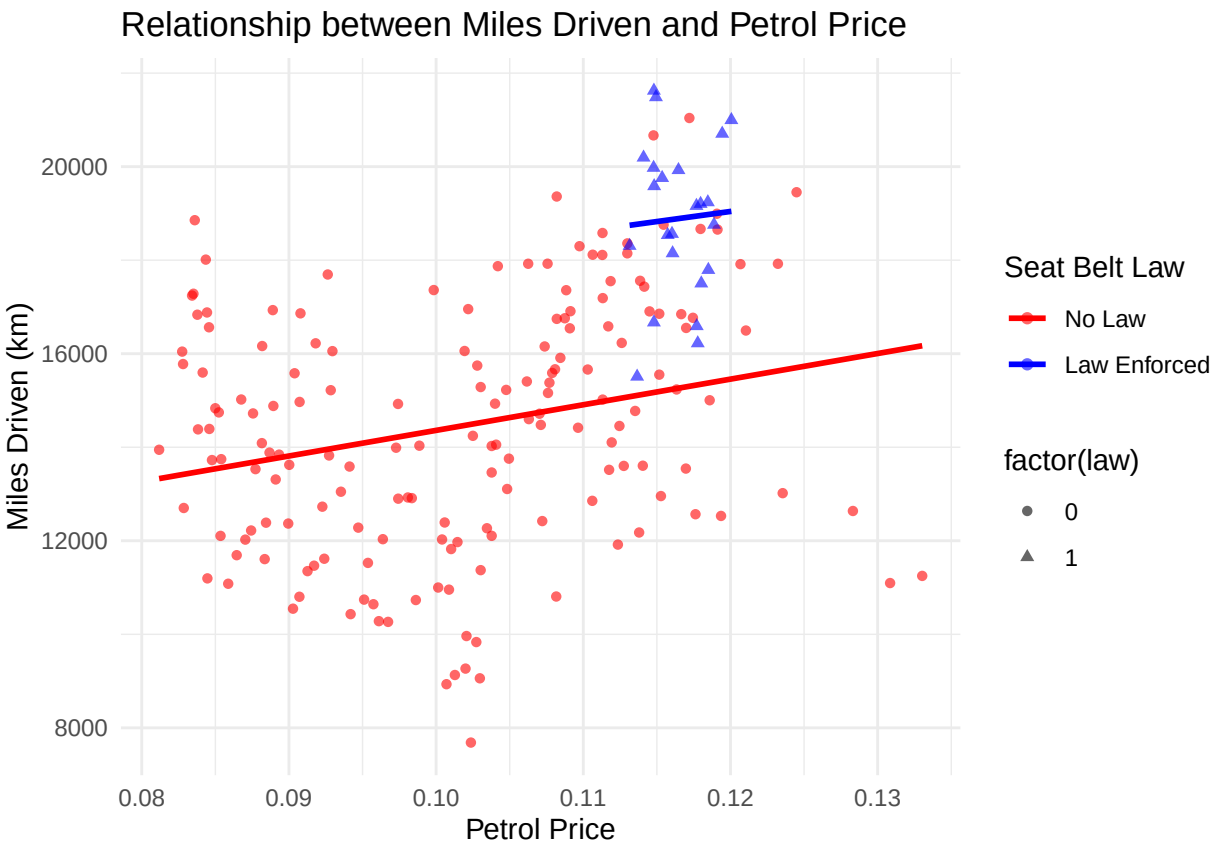
Do gas prices affect driving decisions? We are interested in predicting the number of miles driven (`km`) from the price of gas (`PetrolPrice`). Generate a scatterplot for these two variables. Make sure that your scatterplot meets the following characteristics:

- include clear axis labels & graph titles
- include two lines summarizing the relationship between `km` and `PetrolPrice`, one for months before the seatbelt law was passed and one for months after it was passed
- you may also include additional formatting, including colors and shapes, but these are not required.

```
# Scatterplot with lines of best fit
ggplot(data=sb, aes(x=PetrolPrice, y=kms, color=factor(law))) +
```

```
  geom_point(aes(shape=factor(law)), alpha=0.6) +
  geom_smooth(aes(group=factor(law)), method="lm", se=FALSE) +
  scale_color_manual(values=c("red", "blue"), name="Seat Belt Law",
                     breaks=c(0, 1),
                     labels=c("No Law", "Law Enforced")) +
  labs(title="Relationship between Miles Driven and Petrol Price",
       x="Petrol Price",
       y="Miles Driven (km)") +
  theme_minimal()
```

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



Would you say that gas prices affect driving decisions? Explain.

Answer: Given that both lines exhibit a relatively small positive slope, it suggests that there is only a slight increase in the distance traveled as gas prices rise. Therefore, while there is a minor positive correlation, it's evident that gas prices have a limited influence on driving decisions. In essence, other factors might play a more dominant role in influencing how much people drive than the price of gas alone. ***

Exercise 5: Logical Statements [20 points]

part a

How many months are included in this dataset? For how many months was the law in place in the data?

Use this code chunk to answer the question.

```
totalMonth = length(sb$law)
lawMonths = sum(sb$law == 1)

totalMonth
```

```
## [1] 192
```

```
lawMonths
```

```
## [1] 23
```

part b

Define a deadly month for drivers to be a month where more than 100 drivers were killed.

What proportion of months in this dataset would classify as a deadly month for drivers?

```
deadMonth = sum(sb$DriversKilled >=100)
(deadMonth/length(sb$DriversKilled)) * 100
```

```
## [1] 83.85417
```

part c

Define a deadly month for **van** drivers to be a month where more than 10 van drivers were killed.

What proportion of months in this dataset would classify as a deadly month for van drivers?

```
deadVan = sum(sb$VanKilled >= 10)
(deadVan/length(sb$VanKilled)) * 100
```

```
## [1] 45.3125
```

part d

For what proportion of months in this dataset was it deadly for at least one of drivers or van drivers? What proportion of months in this dataset was it deadly for both drivers and van drivers?

```
deadlyforOne <- sum(sb$DriversKilled >= 100 | sb$VanKilled >= 10)

deadlyforTwo <- sum(sb$DriversKilled >= 100 & sb$VanKilled >= 10)

propdeadlyForOne <- (deadlyforOne/length(sb$DriversKilled)) * 100

propdeadlyForTwo <- (deadlyforTwo/length(sb$DriversKilled)) * 100

propdeadlyForOne
```

```
## [1] 86.45833
```

```
propdeadlyForTwo
```

```
## [1] 42.70833
```

part e

Priya picked a different cutoff point for deciding a month was deadly. Priya decided that a month would be considered deadly for drivers if 90 or more drivers were killed. Determine how many additional months are

considered deadly for drivers based on Priya's cutoff compared to our original definition in **part b**.

```
priya = (sum(sb$DriversKilled >=90) - deadMonth)
priya
```

```
## [1] 17
```

Exercise 6: Subsetting Data [15 points]

We'll return to the idea of analyzing what the compulsory seat belt law accomplishes. To do this, it'll be helpful to subset the data into two individual datasets.

part a

Create a `withlaw` dataset containing the observations for the months where the law was active, and a `withoutlaw` dataset that contains the observations for the months before the law was active.

```
# Create a subset with observations when the law was active
withlaw <- subset(sb, law == 1)

# Create a subset with observations before the law was active
withoutlaw <- subset(sb, law == 0)
```

part b

Confirm that this separation worked. Do this using code along with reasoning based on interpreting your output. For this question, you should not print the entire dataset and check manually (by eye) that this separation worked.

```
# Check the number of observations
total_obs <- nrow(sb)
withlaw_obs <- nrow(withlaw)
withoutlaw_obs <- nrow(withoutlaw)

# Check for any law == 1 observations in withoutlaw dataset and vice versa
law_violation_without <- sum(withoutlaw$law == 1)
law_violation_with <- sum(withlaw$law == 0)

total_obs
```

```
## [1] 192
```

```
withlaw_obs
```

```
## [1] 23
```

```
withoutlaw_obs
```

```
## [1] 169
```

```
law_violation_without
```

```
## [1] 0
```

```
law_violation_with
```

```
## [1] 0
```

Answer/support: The sum of observations from the withlaw (23) and withoutlaw (169) datasets equals the total number of observations in the original sb dataset (192). This confirms the data was split correctly.

Additionally, both checks for law violations in the respective datasets resulted in a value of 0. This further confirms that the datasets were subsetted accurately based on the presence or absence of the law. Therefore, the separation of data worked as intended.

part c

Calculate summary statistics for the number of drivers seriously injured (`DriversInjured`) before the law was active and when the law was active. Describe what you see in these summary statistics (similarities & differences). If you were talking to a friend, would you suggest that there is a difference based on compulsory seat belt usage?

```
summary_withoutlaw <- summary(withoutlaw$DriversInjured)
summary_withlaw <- summary(withlaw$DriversInjured)

summary_withoutlaw
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1197    1408    1530    1592    1785    2456
```

```
summary_withlaw
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     962    1087    1188    1221    1342    1609
```

Answer:

Observations:

- The average serious injuries decreased from 1,592 before the law to 1,221 after its enactment.
- The maximum number of drivers seriously injured after the law (1,609) is even lower than the 1st quartile before the law (1,408).

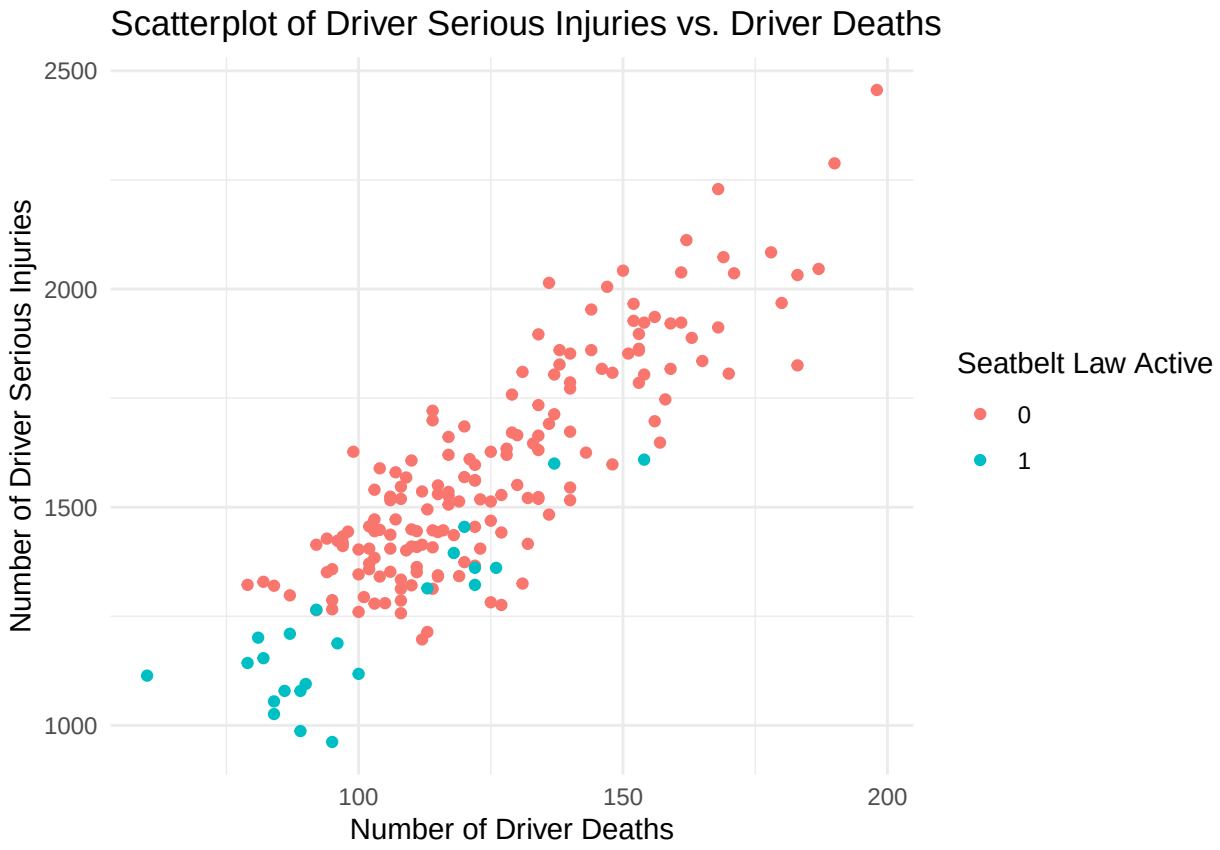
Conclusion: The compulsory seat belt law appears to have significantly reduced the number of drivers seriously injured. Based on this data, one would suggest that the seat belt law has made a positive impact on driver safety.

Exercise 7: Driver Recommendations [10 points]

part a

Create a scatterplot of the number of driver serious injuries vs. the number of driver deaths. Incorporate the law variable into this scatterplot, and make sure that the scatterplot has clear axes labels.

```
ggplot(data = sb, aes(x = DriversKilled, y = DriversInjured, color = as.factor(law))) +
  geom_point() +
  labs(
    x = "Number of Driver Deaths",
    y = "Number of Driver Serious Injuries",
    title = "Scatterplot of Driver Serious Injuries vs. Driver Deaths",
    color = "Seatbelt Law Active"
  ) +
  theme_minimal()
```



part b

Interpret this scatterplot, and explain the real world significance of this graph. For example, what might you tell a driver about how using a seatbelt might affect the risk of serious injury and the risk of death?

Answer:

Interpretation: The scatterplot reveals a positive correlation between driver deaths and serious injuries. Points representing months after the seatbelt law's enactment suggest a decrease in both deaths and injuries.

Real-World Significance: Seatbelt usage appears to reduce both death and serious injury risks in accidents.

Advice to a Driver: "Wearing a seatbelt not only reduces your risk of dying in an accident but also lessens severe injuries. It's a simple act with major benefits."

Exercise 8: Asking Questions & Exploring Data [10 points]

This exercise will be more open ended. Define and complete a new exploration of the seatbelts data in order to understand some aspect of the data. You may use graphs, numerical summaries, or some combination of the two in your exploration. You may return to your answer from Homework 1 Exercise 7 for additional inspiration. Be sure to write up your final findings in about 1-2 paragraphs. Make sure that your analysis is original, not a recreation of something you have done in Homework 1 or Homework 2.

First, write your goal for your exploration below.

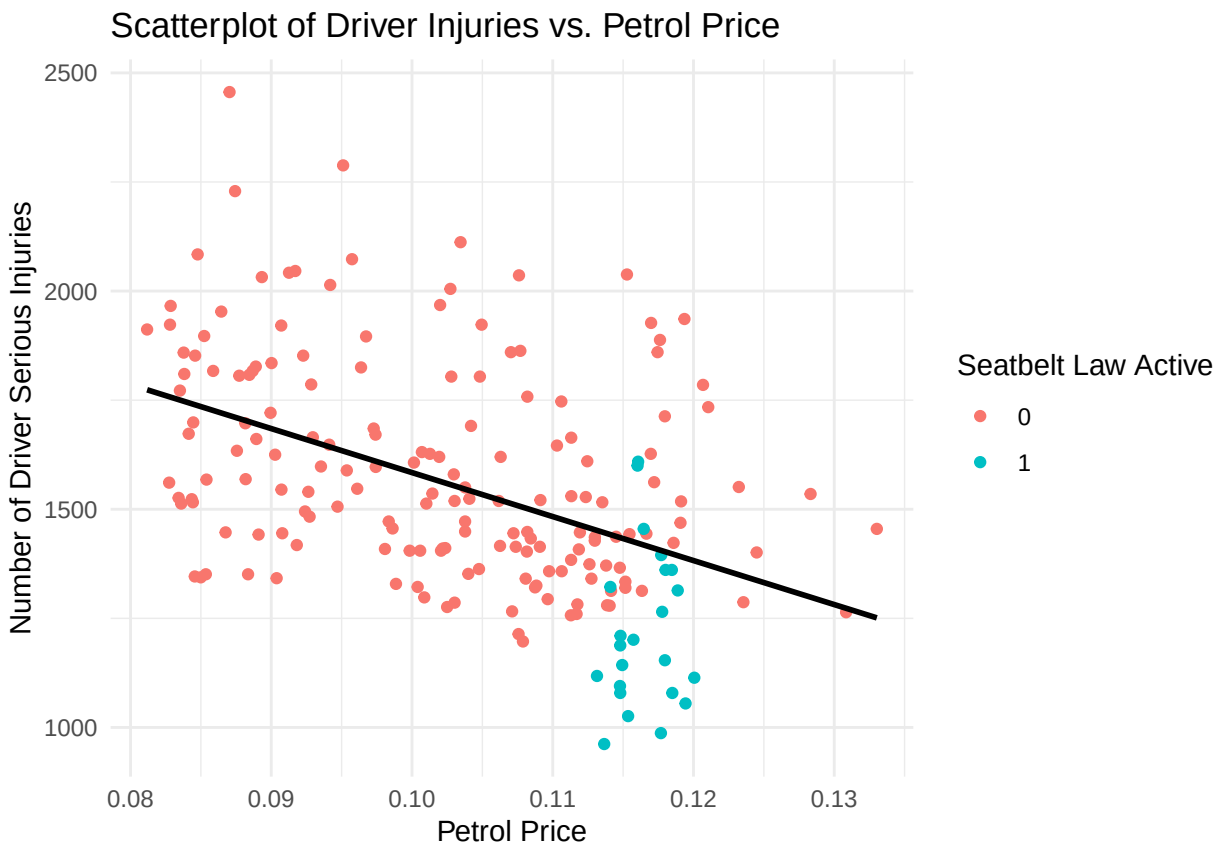
Goal:

To determine if petrol prices (`PetrolPrice`) influence the number of drivers seriously injured (`DriversInjured`).

Then, complete the analysis.

```
ggplot(data = sb, aes(x = PetrolPrice, y = DriversInjured)) +
  geom_point(aes(color = as.factor(law))) +
  geom_smooth(method = "lm", se = FALSE, color = "black") +
  labs(
    x = "Petrol Price",
    y = "Number of Driver Serious Injuries",
    title = "Scatterplot of Driver Injuries vs. Petrol Price",
    color = "Seatbelt Law Active"
  ) +
  theme_minimal()
```

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



Finally, write up your findings:

Answer: The scatterplot presents a relationship between petrol prices and the number of driver serious injuries. Initially, it may be assumed that as petrol prices rise, individuals might drive less, leading to fewer accidents and injuries. However, the actual trend, supported by the linear regression line, indicates only a subtle change in injuries with varying petrol prices. This suggests other factors, like safety measures (e.g., the seatbelt law) and vehicle safety improvements, could play more pivotal roles in determining injury rates. Additionally, any fluctuation in driving frequency due to petrol prices seems not to drastically impact the number of severe injuries.