Homework 2

Anahi Rodriguez

Due 9/7/2022

Homework Instructions

Each exercise will be worth 10 points, unless otherwise noted.

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 multiple choice questions, please bold your selected answer.

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: Adding Variables to a Dataset [10 points]

As a first step in analyzing the `sb` dataset, 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 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.

```
# Use this code chunk to answer the question, by replacing this line or
# adding a new line below it.
sb$DriversInjured =  sb$drivers - sb$DriversKilled
```

part b

The `sb` dataset contains the number of rear-seat passengers killed or seriously injured in the `rear` variable. While there are often many seats available in the rear of a car, the driver and the front-seat passenger each refer to one specific location in the car. Therefore, create an `AllFront` variable that records the number of deaths or serious injuries for drivers and front-seat passengers, or in other words, all individuals sitting at the front of the car.

```
# Use this code chunk to answer the question, by replacing this line or
# adding a new line below it.
sb$AllFront = sb$front + sb$drivers
```

part c

Confirm that your two variables have 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.

```
# Use this code chunk to answer the question, by replacing this line or
# adding a new line below it.
names(sb)
```

```
##  [1] "DriversKilled"  "drivers"        "front"          "rear"
##  [5] "kms"            "PetrolPrice"    "VanKilled"      "law"
##  [9] "DriversInjured" "AllFront"
```

Answer: I used the names variable around 'sb' to see all of the variables included to make sure the two newest variables created actaully exist in the dataframe

Exercise 2: 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 have an affect on those in the front of the car, both drivers and passengers?

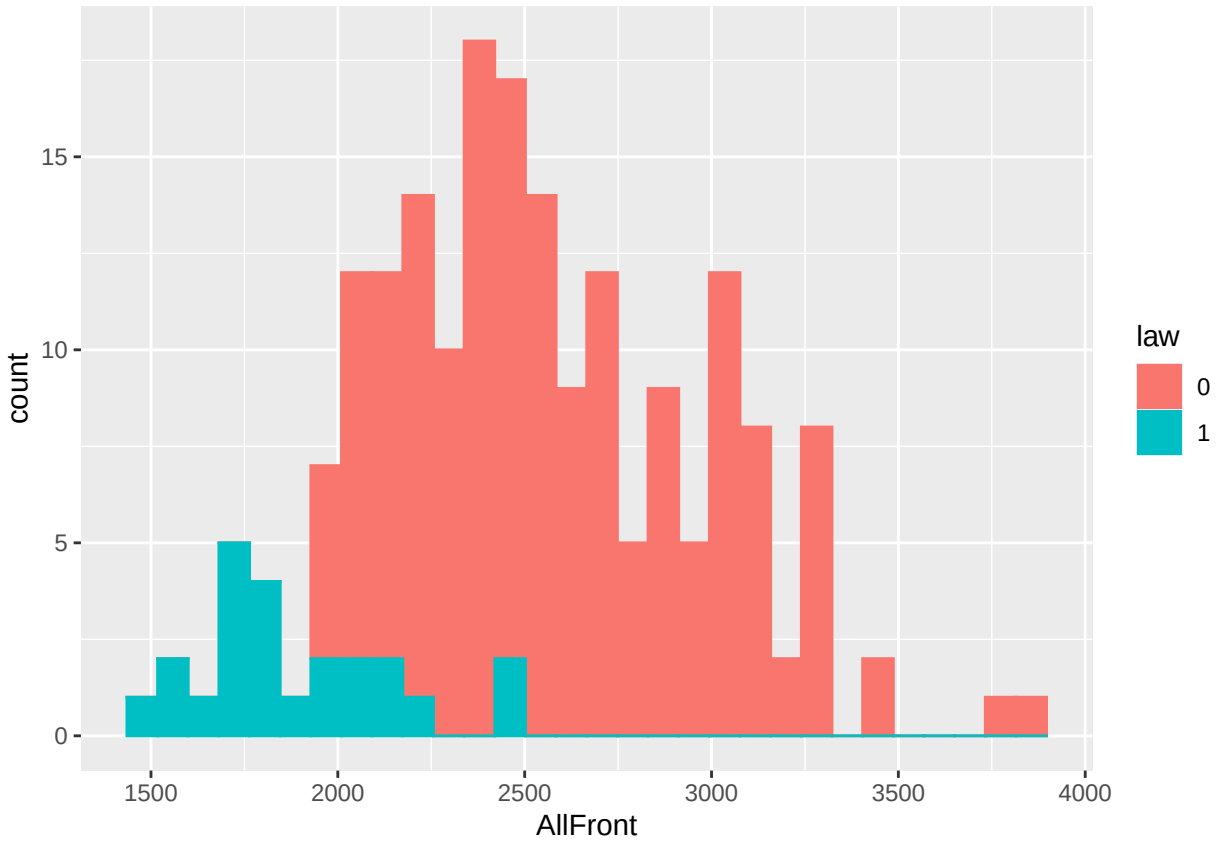
part a

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

```
# Use this code chunk to answer the question, by replacing this line or
# adding a new line below it.

ggplot(sb, aes(x = AllFront, color = law, fill = law)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



part b

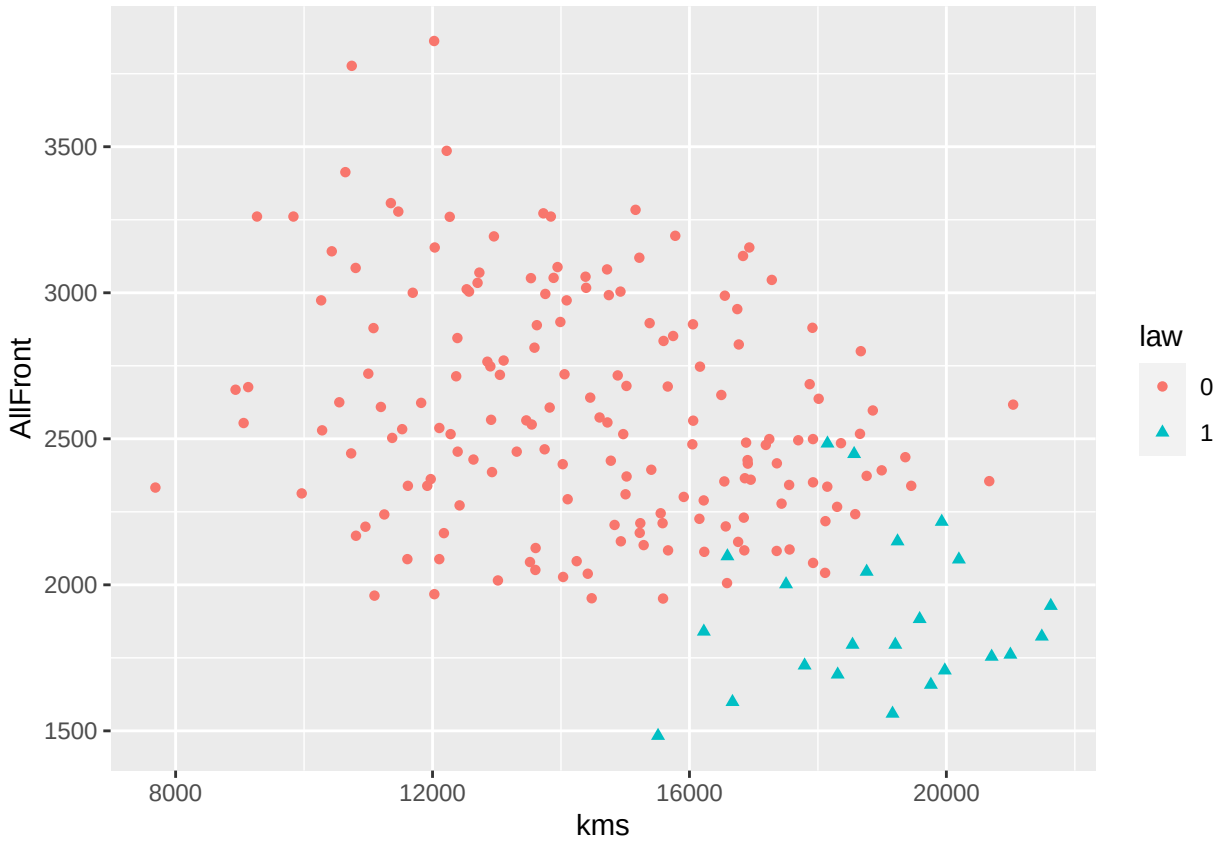
What do you notice from this graph?

Answer: Death or serious injuries are much less common when the seatbelt law is in place than when there is no law in place. When the law is in place, there still exists injuries and deaths but in much smaller numbers every month.

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 drivers and front-seat passengers (`AllFront`). 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.

```
# Use this code chunk to answer the question, by replacing this line or
# adding a new line below it.
ggplot(sb, aes(x = kms, y = AllFront, color = law, shape = law)) +
  geom_point()
```



part d

What do you notice from this graph? Discuss how your understanding has changed from **part b**, if at all.

Answer: It is much more likely for an individual to be injured or killed when driving shorter distance when then law was not in place and similarly, the deaths or injuries that occur when the law is in place, happen when individuals are driving larger distances.

Exercise 3: Logical Statements [20 points]

part a

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

```
## Use this code chunk to answer the question.
dim(sb)
```

```
## [1] 192  10
```

```
law_months = subset(sb, sb$law == 1)
dim(law_months)
```

```
## [1] 23 10
```

Answer: The dataset spans over 192 months and 23 of those include months where the law was in place

---

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?

```
## Use this code chunk to answer the question.

deadly = subset(sb, sb$DriversKilled > 100)
dim(deadly)/dim(sb)
```

```
## [1] 0.8177083 1.0000000
```

```
157/192
```

```
## [1] 0.8177083
```

Answer: 81.77% of months in this dataset would classify as a deadly month for drivers

---

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?

```
## Use this code chunk to answer the question.
deadly_van = subset(sb, sb$VanKilled > 10)
dim(deadly_van)/dim(sb)
```

```
## [1] 0.359375 1.000000
```

```
69/192
```

```
## [1] 0.359375
```

Answer: 35.93% of the months in this dataset would classify as a deadly month for van drivers.

---

part d

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

```
## Use this code chunk to answer the question.

deadly_either_or = subset(sb, sb$DriversKilled > 100 | sb$VanKilled > 10)
dim(deadly_either_or)/dim(sb)
```

```
## [1] 0.8333333 1.0000000
```

```
160/192
```

```
## [1] 0.8333333
```

```
deadly_both = subset(sb, sb$DriversKilled > 100 & sb$VanKilled > 10)
dim(deadly_both)/dim(sb)
```

```
## [1] 0.34375 1.00000
```

```
66/192
```

```
## [1] 0.34375
```

Answer 83.33% of this dataset was deadly for either drivers or van drivers or both. 34.37% of this dataset was deadly for both drivers and van drivers.

---

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**.

```
## Use this code chunk to answer the question.
priya_death = subset(sb, sb$DriversKilled > 90)
dim(priya_death)-dim(deadly)
```

```
## [1] 20  0
```

```
177 - 157
```

```
## [1] 20
```

Answer 20 More months would be considered deadly using Priya's cutoff compared to our original definition

---

Exercise 4: 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.

```
## Use this code chunk to answer the questions.
withlaw = subset(sb, sb$law == 1)
withoutlaw = subset(sb, sb$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.

```
## Use this code chunk to answer the question.
dim(sb)[1] == (dim(withlaw)[1] + dim(withoutlaw)[1])
```

```
## [1] TRUE
```

Answer/support:

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?

```
## Use this code chunk to answer the question
summary(withlaw$DriversInjured)
```

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

```
summary(withoutlaw$DriversInjured)
```

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

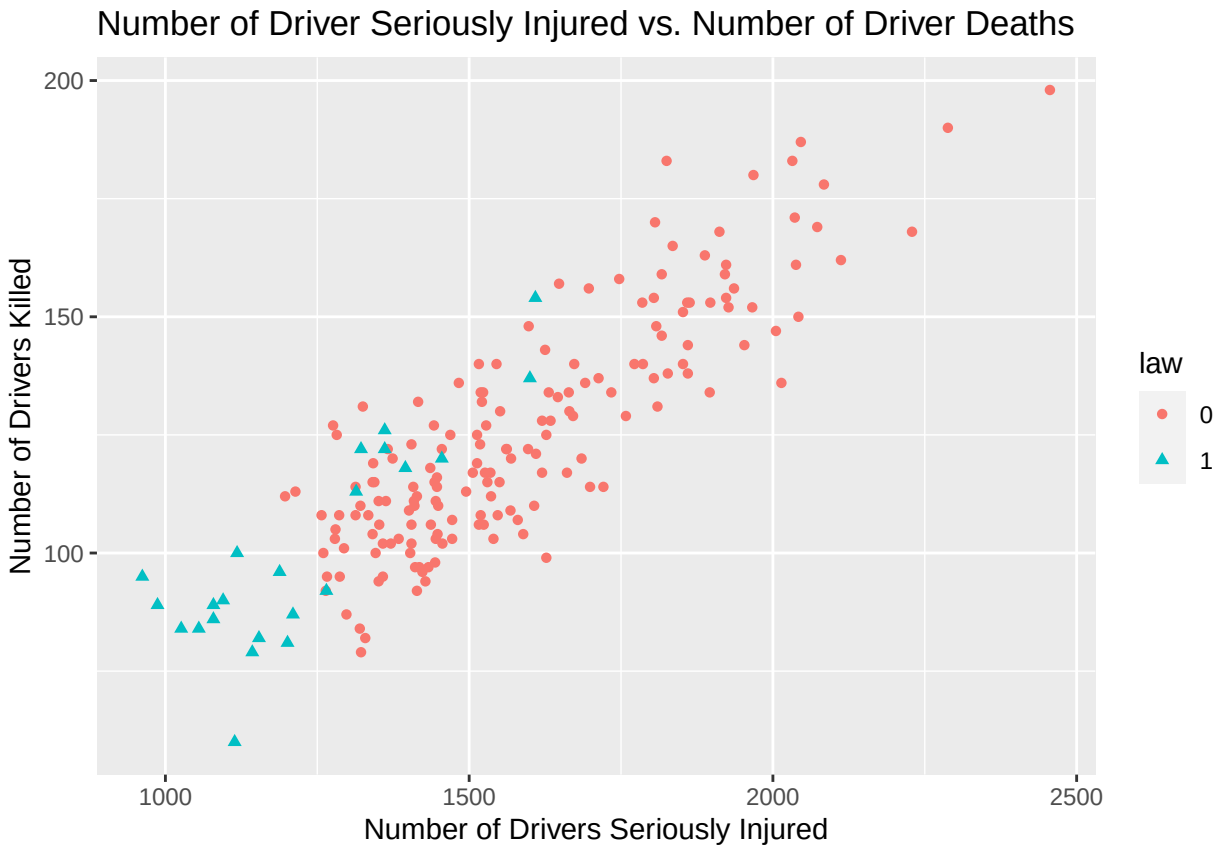
Answer: The smallest number of drivers injured when the law is not in place is almost 200 people greater than the smallest number of drivers injured when the law is in place. Similarly, the average number of drivers injured is over 300 people larger when the law is not in place compared to when the law is in place. I do think that their exists a difference in amount of drivers killed based on compulsory seat belt usage.

Exercise 5: 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.

```
## Use this code chunk to answer the question.
ggplot(sb, aes(x = DriversInjured, y = DriversKilled, color = law, shape = law)) +
  geom_point() +
  labs(title = "Number of Driver Seriously Injured vs. Number of Driver Deaths",
       x = "Number of Drivers Seriously Injured",
       y = "Number of Drivers Killed")
```



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: I would explain that the threat of serious injury is higher than the threat of being killed but that both are higher when the law was not in place. With the law in place, the number of driver deaths and serious injuries decreases as most of the blue triangles are congegrated in the bottom left of the scater plot.

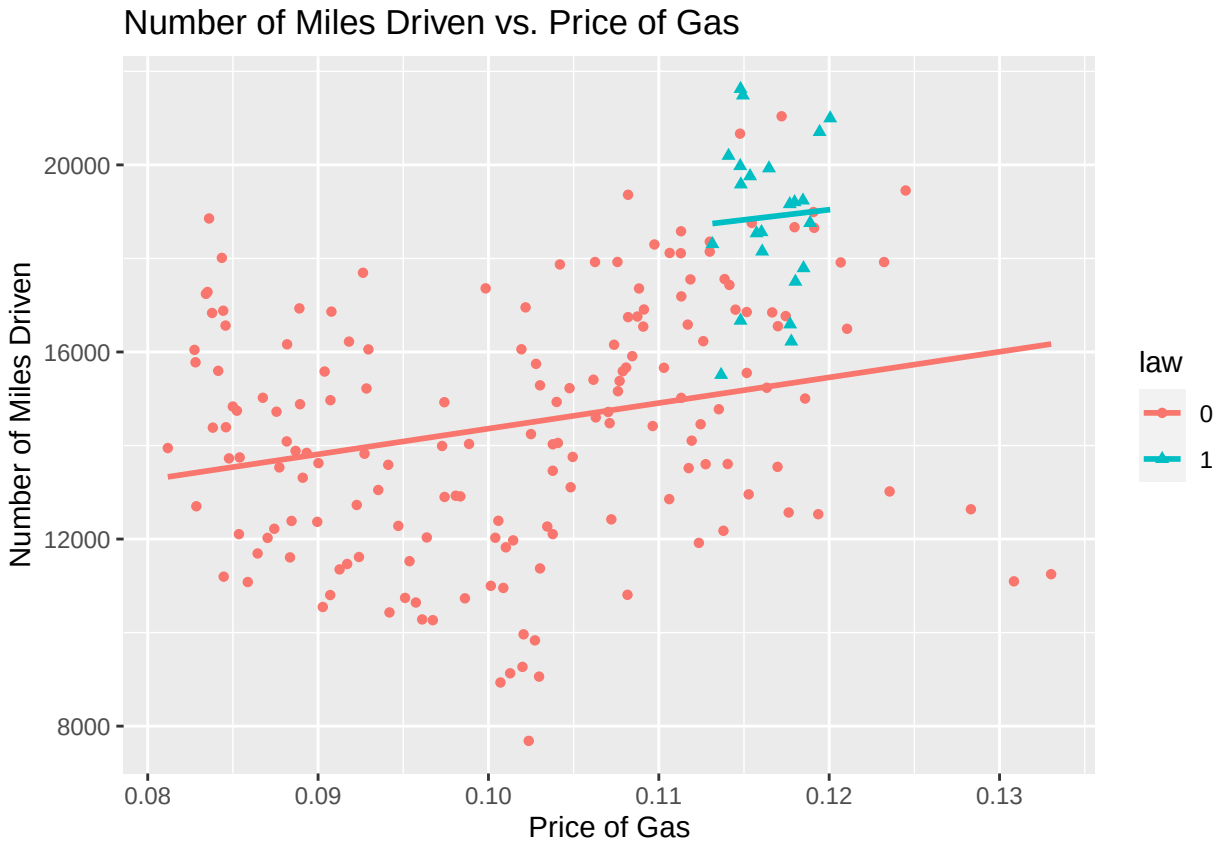
Exercise 6: Line of Best Fit [10 points]

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.

```
## Use this code chunk to answer the question.
ggplot(sb, aes(x = PetrolPrice, y = kms, color = law, shape = law)) +
  geom_point() +
  geom_smooth(method = 'lm', se = F, formula = 'y~x') +
  labs(title = "Number of Miles Driven vs. Price of Gas",
```

```
       x = "Price of Gas",
       y = "Number of Miles Driven")
```



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

Answer: I would say that gas prices affect driving decisions in that people tend to drive slightly further distances when the price of gas is more expensive. We can see this relationship by the positive slopes in the lines.

Exercise 7: 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 analyses, not a recreation of something you have done in Homework 1 or Homework 2.

First, write your goal for your exploration below.

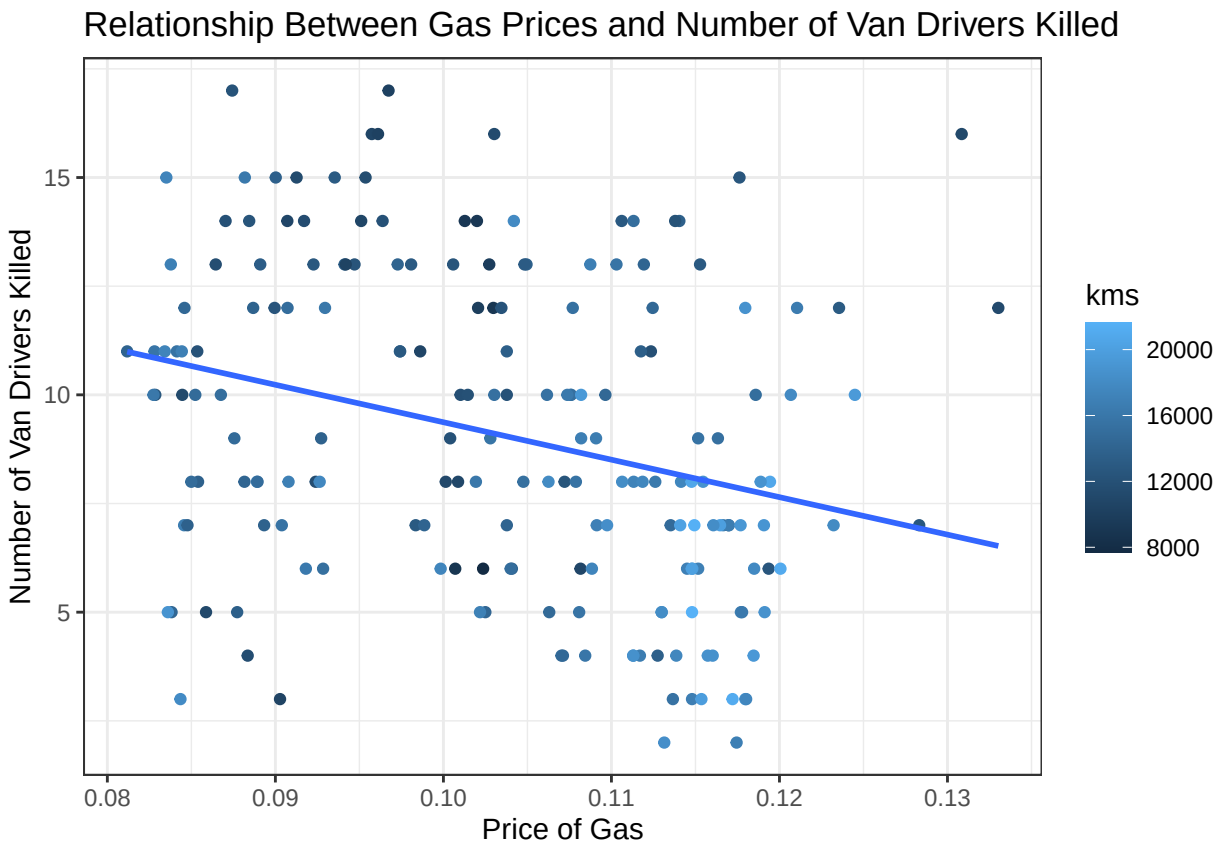
Goal: How the price of gas affects the number of Van driver deaths

Then, complete the analyses.

```
## Use this code chunk to perform your data analysis.

ggplot(sb, aes(x = PetrolPrice, y = VanKilled, color = kms)) +
  geom_point() +
```

```
  geom_smooth(method  = 'lm', se = F, formula = 'y~x' ) +
  labs(title = "Relationship Between Gas Prices and Number of Van Drivers Killed",
       x = "Price of Gas",
       y = "Number of Van Drivers Killed") +
  theme_bw()
```




```
low_gas = subset(sb, sb$PetrolPrice < mean(sb$PetrolPrice))
high_gas = subset(sb, sb$PetrolPrice > mean(sb$PetrolPrice))

summary(low_gas$VanKilled)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    8.00   11.00   10.42   13.00   17.00
```

```
summary(high_gas$VanKilled)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   5.000   7.000   7.853  10.000  16.000
```

```
cor(sb$VanKilled,sb$PetrolPrice)
```

```
## [1] -0.2885584
```

```
cor(sb$VanKilled, sb$kms)
```

```
## [1] -0.4980356
```

```
cor(low_gas$VanKilled, low_gas$kms)
```

```
## [1] -0.3230645
```

```
cor(high_gas$VanKilled, high_gas$kms)
```

```
## [1] -0.4552476
```

Finally, write up your findings:

Answer: I am interested in investigating the relationship between the price of gas and the number of van driver deaths. To do this, I am plotting these against each other in a scatter plot and adding a line of best fit. Additionally, I have added miles driven by color coding the dots according to total distance driven with longer distances colored a lighter blue and shorter distances a darker blue.

It seems that as the price of gas increases, less van drivers deaths occur. In fact, the average number of deaths among observations where gas is lower than the average price is at 10.42 compared to 7.85 for observations with gas prices greater than the average gas price among all observations. Additionally, there exists an overall negative correlation of -0.498 between the number of van drivers killed and the number of miles driven. This does not change much for cases where the price of gas is greater than the average as it only slighlty increase to -0.455. However, among cases where the price of gas was lower than average, the correlation between number of van drivers killed and total miles driven increases to -0.323.

Exercise 8: Formatting [5 points]

The last 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 this page for this exercise (Exercise 8)
- all code is printed and readable for each question
- generated a pdf file