Homework 1

Anahi Rodriguez

due 8/31/2022

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 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 appropriate. This will also help you identify and locate any errors more easily.

Two Technical Notes Before the Assignment:

In order to knit the document to a pdf document, you need to have some supporting software in place. You have two choices: installing a LaTeX editor on your computer, or using an R package that replaces the LaTeX install. If you choose to use the R version (recommended if you don't already have access to LaTeX), you'll want to copy and paste the following line of code (remove the # at the beginning of the line so that R knows to run it) into your Console (the lower left quadrant). You can then either leave the following line of code without making any changes or delete this line of code from your RMarkdown document. Once you have done this once, you should not need to do so again.

You'll also want to make sure that you have ggplot (the visualization package) available for you to use. We'll load the library later in the document. If you haven't already installed the package on your computer, you'll need to download the package. To do so, run the following line of code in your Console after removing the # at the beginning of the line. You only need to do this on your computer once, and it will be available each time you open RStudio. You can then either leave the following line of code without making any changes or delete the line of code from your RMarkdown document.

Warning: These two lines of code should only be run from the Console. To do this, copy and paste the code to the Console (lower left quadrant of RStudio) without the hashtag and space at the beginning of the code. You will not need to run these lines of code again, as you'll have downloaded the packages to your RStudio. You can leave the two lines of code as they currently exist in your RMarkdown document, or you can delete these lines of code. If you leave the code active in your document (remove the hashtag and space but keep the code), you will receive an error message when you try to knit the document.

Exercise 1: Initial Setup [5 points]

The first five points for this assignment come from setting up the document and your environment correctly. You will earn these points by knitting your document and creating a pdf.

I recommend that you knit the document to a pdf now, before beginning the main coding exercises. This way, you ensure that your environment is set up correctly and know that the document was properly formatted

before you begin editing the document. To knit the document, click the ball of yarn with a knitting needle on it or the word "Knit" beside it.

Exercise 2: Formatting & Submitting [5 points]

The next 5 points of this assignment will be earned for completing the following tasks:

- Including your name in the header of the document
- Assigning pages correctly on Gradescope during submission

Please also assign page 1 with your name for this exercise.

Exercise 3: Calculations [15 points]

R is a powerful calculator. Let's perform some basic operations with R in this exercise.

part a

Translate the following mathematical statement into R code, and calculate the result:

$(7+5)^{9-3} \times \frac{55}{6}$

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

((7 + 5)^6) * (55/6)
```

```
## [1] 27371520
```

part b

In addition to using R as a calculator, we can use built-in functions of R to simplify our calculations. Below, use the mathematical constant $\pi \approx 3.14159$, represented by `pi` in R, to calculate the value of $\pi^{2.3}$. Make sure this value is printed.

```
# Use this code chunk to answer the question, by replacing this line or
# adding a new line below it.
pi^(2.3)
```

```
## [1] 13.91377
```

part c

Create a vector that consists of the integers from 3 to 9, including both 3 and 9. Assign the vector to the variable **z**.

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

z <- (3:9)
z
```

```
## [1] 3 4 5 6 7 8 9
```

part d

Using the vector **z** that you created in part c, generate a new vector **y** with values of $\pi^z - \pi^{2.3}$. Print **y**.

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

y <- (pi^z) - pi^(2.3)
y
```

```
## [1]    17.09251    83.49532   292.10592   947.47543  3006.37946  9474.61725
## [7] 29795.18557
```

part e

We can chain functions and operations, allowing us to use multiple operations and functions in one line. In one line of code, square the values contained in each entry of **y**, then sum all of those values, and finally take the square root of the resulting sum. Print the resulting value.

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

print(sqrt(sum(y^2)))
```

```
## [1] 31425.31
```

Exercise 4: Dataset Basics [10 points]

While it can be very helpful that R is powerful in an abstract situation, we often want to answer questions using data. The remaining exercises will allow us to apply some of the built-in statistical functions to a dataset. We'll look at a road casualties dataset from Great Britain in the 1960s to 1980s. You can read more about the dataset and variables using the Help feature to the right or by typing `?Seatbelts` into the R Console below.

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

The dataset is now stored in R as `sb`.

Hint: the Day 1 Demo file on Canvas will be helpful for Exercises 4-7.

part a

Using R code, determine how many columns (variables) and rows (months) are contained in the dataset. Write the solution below the code, replacing the blank lines with the appropriate numbers.

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

dim(sb)
```

```
## [1] 192   8
```

Answer:

This dataset has 192 rows & 8 columns.

part b

Print the first 6 rows of the dataset.

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

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
```

part c

Looking at the first 6 rows of the dataset that you've printed above, what do you notice? What questions do you have about this dataset?

Answer:

Looking more into the dataset, I see that this data only covers one year here the law was in place, making it likely to see 0s in the 'law' column. ***

Exercise 5: Numerical Summaries [15 points]

Suppose that the Department of Transportation is interested in comparing the number of drivers killed or seriously injured (`drivers`) with the number of front-seat passengers killed or seriously injured (`front`).

part a

First, generate numerical summaries for the `drivers` variable. Be sure to calculate the mean, the five number summary, and the standard deviation (try searching on Google or looking ahead in the notes for this function if you haven't seen it in class yet).

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

summary(sb$drivers)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1057    1462    1631    1670    1851    2654
```

```
sd(sb$drivers)
```

```
## [1] 289.611
```

part b

For the `drivers` variable: Which is larger, the mean or the median? Think back to what this might tell you about the shape of the distribution. What do you anticipate about the shape of the distribution from the mean, the median, and the five number summary?

Note: We won't grade your prediction based on correctness; we're hoping that you'll think about what the data means here.

Answer: The mean is smaller than the mean, positively skewing the distribution. This tells us that more than half of the observations have a 'driver' value that is at least as large as its average value.

part c

Now, calculate the same numerical summaries from **part a** for the `front` variable.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   426.0   715.5   828.5   837.2   950.8  1299.0
```

```
sd(sb$front)
```

```
## [1] 175.099
```

part d

Compare the numerical summaries for the drivers from **part a** with the numerical summaries for the front seat passengers in **part c**. What do you notice? What real world implications might this have?

Answer: the variance is larger in the 'drivers' variable and the 'front' variable has a more normal distribution

---

Exercise 6: Visualizing and Interpreting One Variable [10 points]

We generated numerical summaries for the number of deaths and serious injuries of drivers and of front seat passengers in the last problem. Now, let's visualize what the distributions for these variables look like.

To do this, we'll load the `ggplot2` package, which allows you to create graphics like we did in class.

```
library(ggplot2)
```

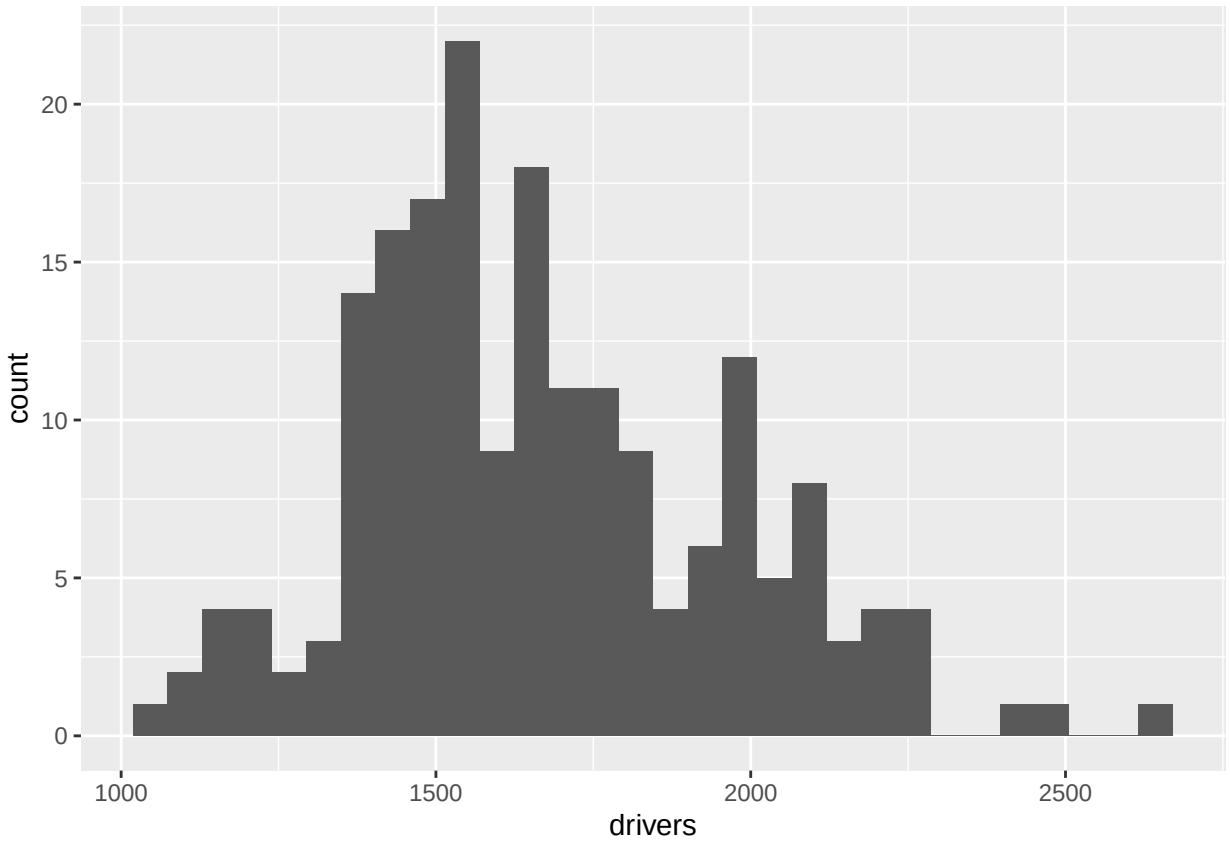
part a

First, generate a histogram of the `drivers` variable.

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

ggplot(data = sb, aes(x = drivers)) +
  geom_histogram()
```

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



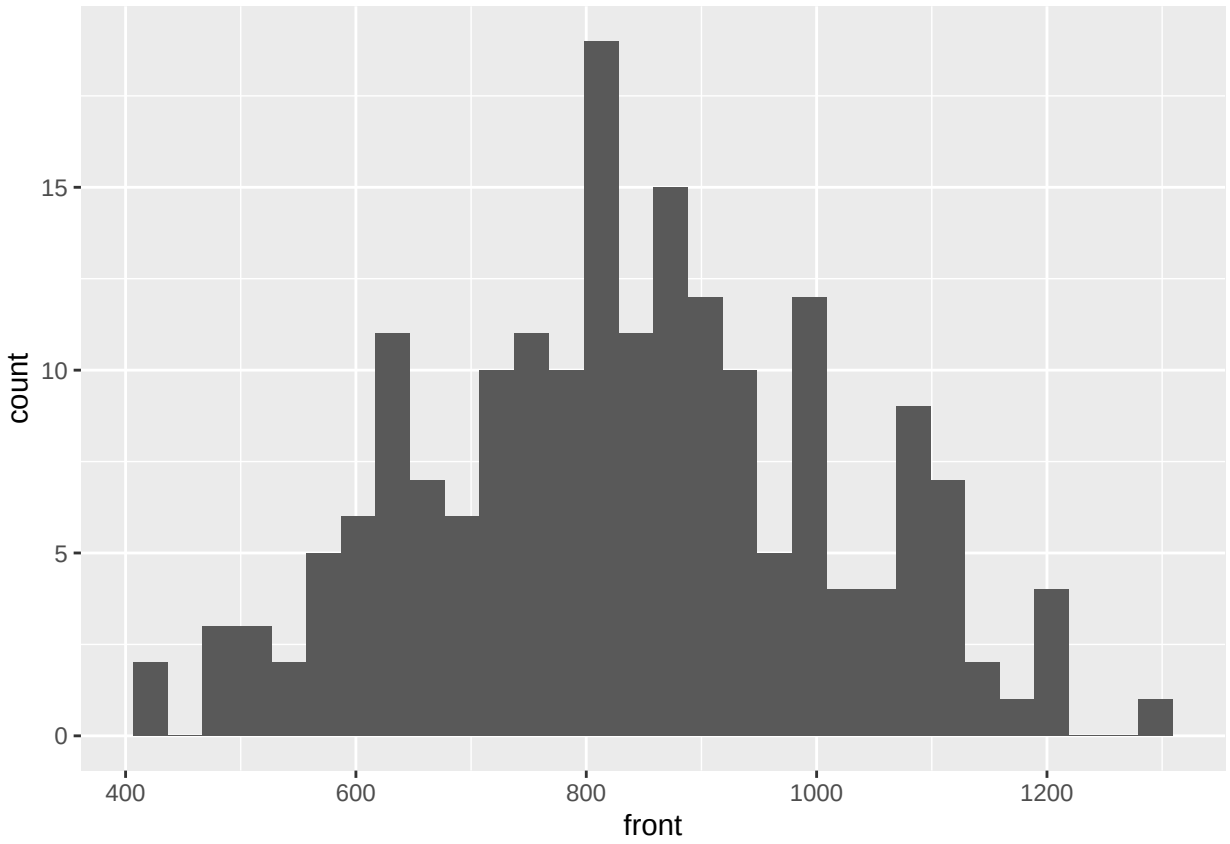
part b

Now, generate a histogram of the `front` variable.

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

ggplot(data = sb, aes(x = front)) +
  geom_histogram()
```

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



part c

How would you describe the number of deaths and serious injuries of drivers and of front seat passengers to an intern in the Department of Transportation? What do you notice from these two graphs?

Answer: This is a very serious issue for both the driver and front seat passengers, however, it is causing more deaths among drivers. The peak of these histograms has a larger frequency as well as a higher death toll among the driver death data than the front seat passenger death data accounts for.

Exercise 7: Scatterplots of Two or More Variables [20 points]

Finally, we'll create a scatterplot of the relationship between deaths and serious injuries of drivers compared to those of front seat passengers.

To do this, we'll load the `ggplot2` package, which allows you to create graphics like we did in class. You should delete the hashtag and space at the beginning of this line of code to load the package. Note that this line of code will not run if you have not previously installed the `ggplot2` package as described at the beginning of this Homework1.Rmd file.

```
library(ggplot2)
```

part a

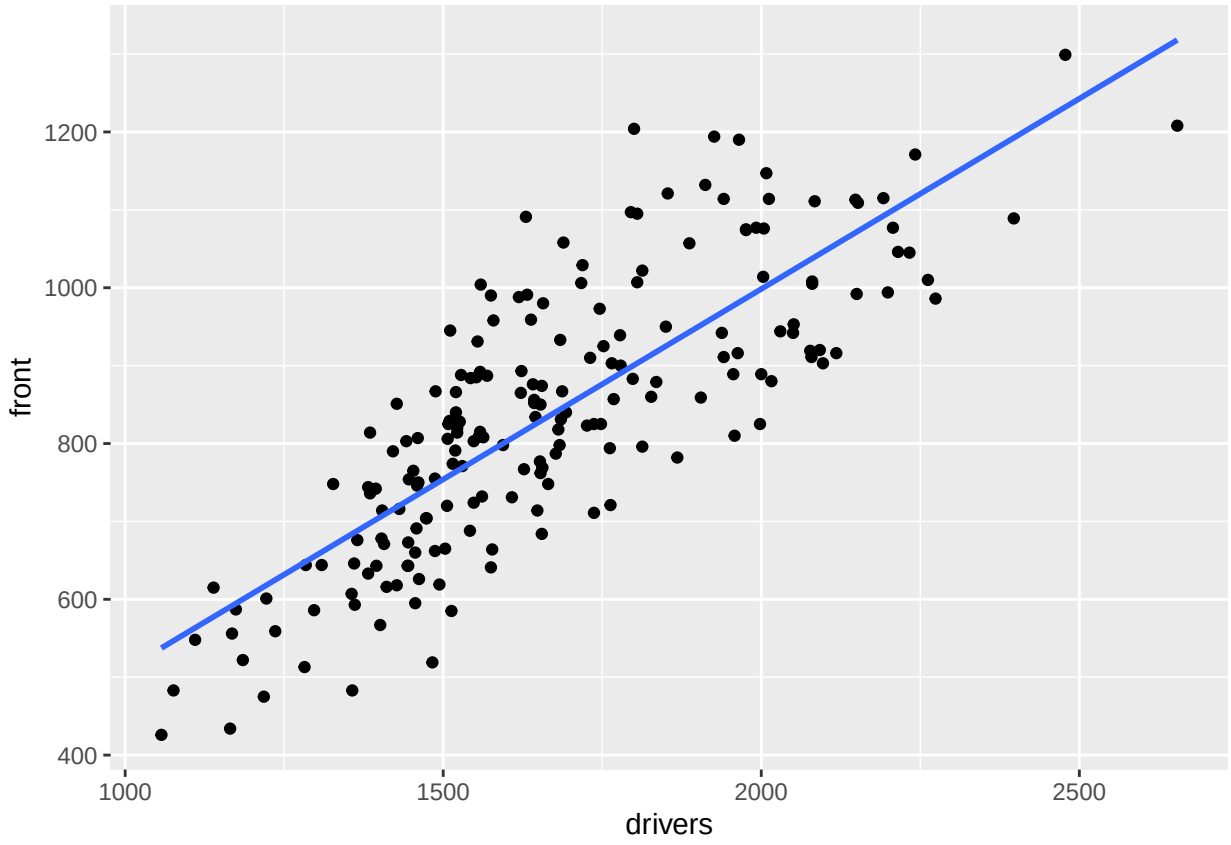
Modify the dataset name and variable names from the following code found in the Day 1 Demo R file to create a scatterplot of the Seatbelts data. Place the deaths and serious injuries from drivers on the x axis and the deaths and serious injuries from front seat passengers on the y axis.

ggplot(data = coasters, mapping = aes(y = Speed, x = Height)) + geom_point()

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

ggplot(data = sb, aes(x = drivers, y = front)) +
  geom_point() +
  geom_smooth(method = 'lm', se = F)
```

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



part b

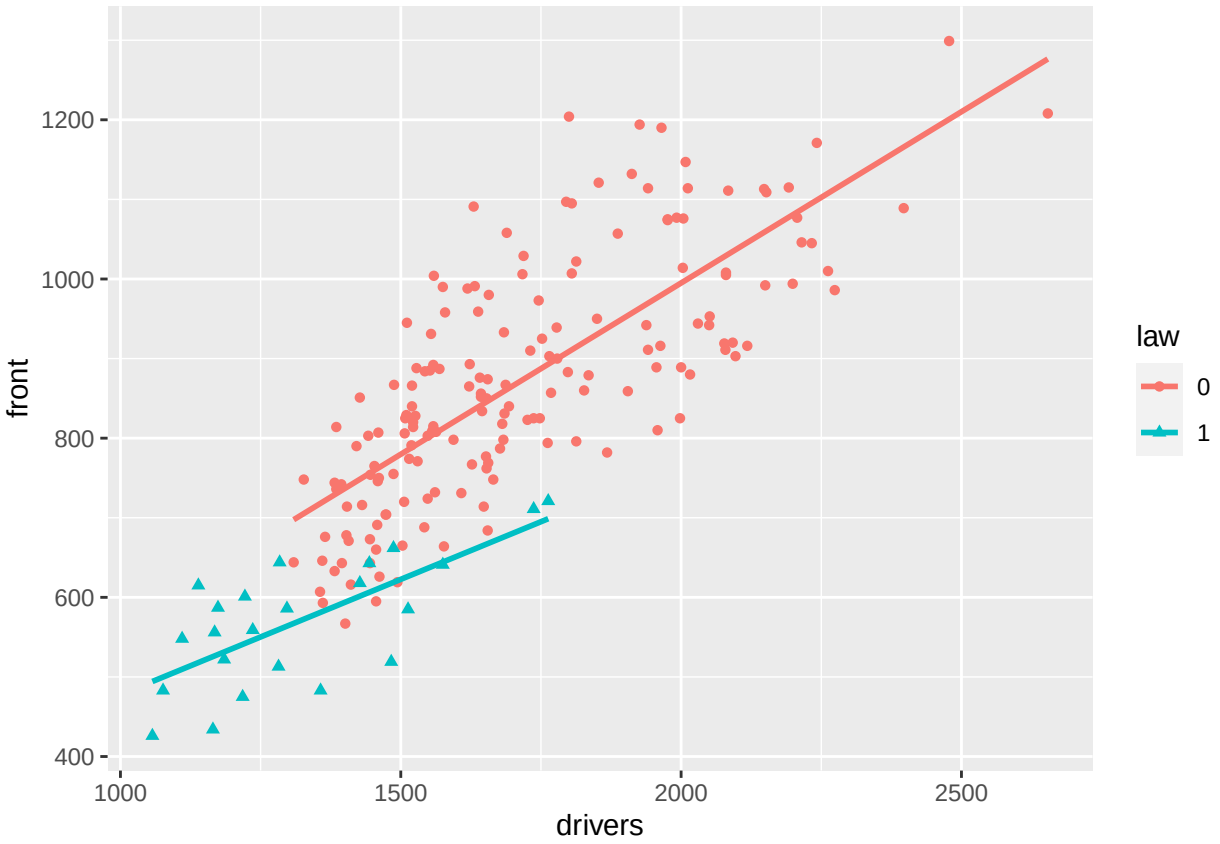
Now, add in the variable `law`, which represents whether a law requiring the use of seatbelts was in place for the given month or not. Incorporate the `law` variable in the shape and color of the points in the scatterplot. Add lines of best fit to this graph, as well. Below, you will find sample code from the Day 1 Demo R file.

ggplot(data = coasters, mapping = aes(y = Speed, x = Height, shape = Track, color = Track)) + geom_point() + geom_smooth(method = 'lm', se = F)

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

ggplot(data = sb, aes(x = drivers, y = front, shape = law, color = law)) +
  geom_point() +
  geom_smooth(method = 'lm', se = F)
```

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



part c

What do you notice from the scatterplots generated in **parts a & b** above?

Answer: Death tolls tend to be lower after the law has been passed among both driver and front seat deaths. Months with the highest death tolls are concentrated within cases without the law.

part d

What would be the next step that you would want to take if you were to continue looking at this dataset? You may choose to answer some or all of the following prompts for this question.

- If you were to continue analyzing the drivers and/or front variables, what would you take as the next step?
- What questions would you want to explore using any or all other variables in the dataset?
- What other visualizations might you pursue?
- If you could have any additional variable or information added to the dataset, what would you ask for? How would you use that information?
- Thinking about the data more deeply, do you have any concerns about what might be missing from the data? What types of information might not have been recorded? Are there any ambiguous variables or situations in the data?

There is not one correct answer to this question; you can be creative in your approach. No need to perform any suggested analyses for this problem.

Answer: I would ask for actual years and months assocaited with the observation as well as maybe the makeup of sex between these variables. There could also be more information stored on the amount of people in the car during these accidents.

Exercise 8: Debugging [20 points]

This coding exercise includes opportunities for debugging (fixing) common errors in RMarkdown. For this exercise, you will be provided with chunks of code that will prevent you from knitting the document, have some error in them, or are not formatted correctly. Because these chunks have errors, they also have an exclamation mark and a space (!) added at the beginning of the lines to allow you to knit the document initially. Please remove the exclamation marks and spaces at the beginning of the lines, fix the errors, and then knit the resulting chunk. Additionally, explain the error that you corrected.

For this exercise, you do not need to change any `R` code or functions; focus on the formatting of the chunks, successfully knitting the document, and having appropriate formatting for the document.

Hint: Work through one part at a time to help isolate and correct any errors. Knit the document now to ensure there are no other errors present in the document prior to starting this exercise.

part a

First, remove the exclamation mark and space that are at the beginning of the next three lines of code. Then try knitting the document. You should see an error. Make adjustment(s) to the next three lines of code until the document successfully knits. Then note what you changed or what error was initially present.

```
sum(1:10)
```

```
## [1] 55
```

Answer: The error was: needed to close the rchunk.

part b

```
 sum(11:20)
```

```
## [1] 155
```

Answer: The error was: two chunks had same name

part c

```
(1:10)^2
```

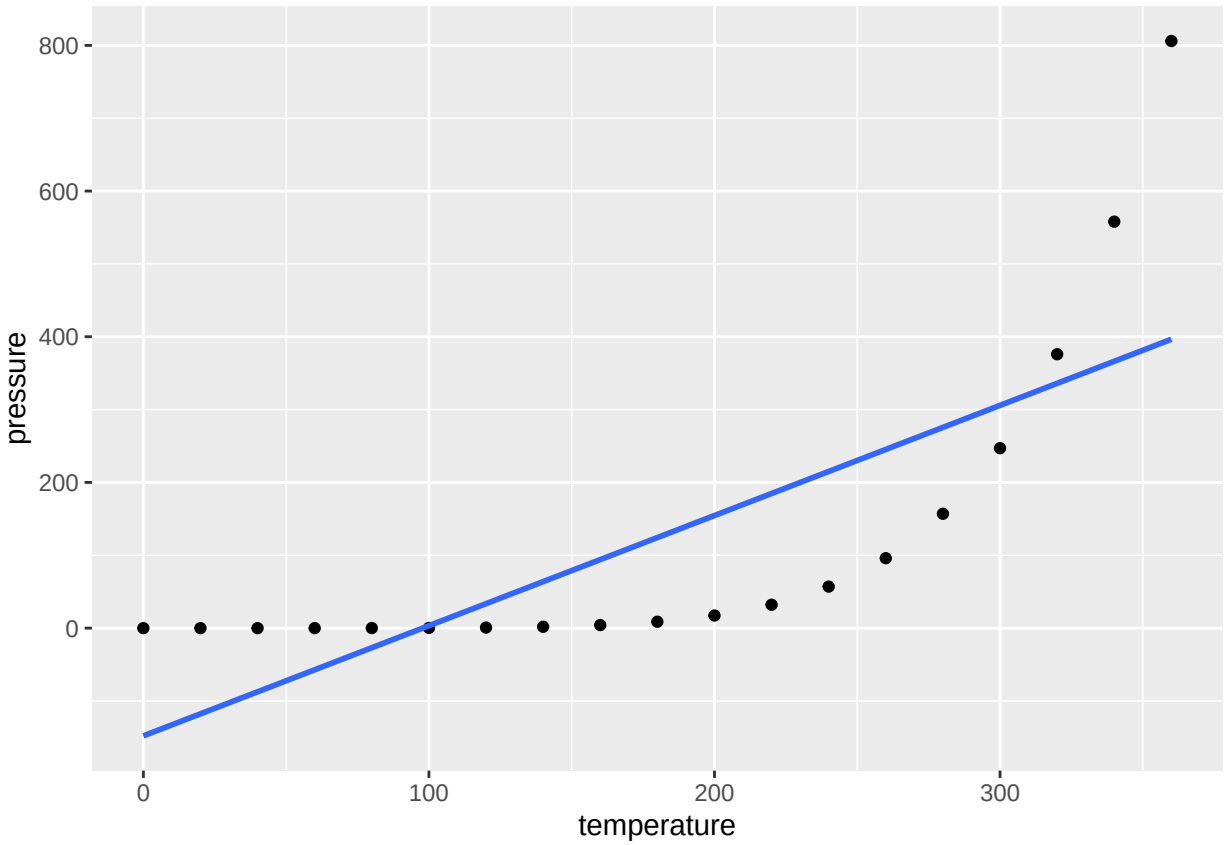
```
##  [1]   1   4   9  16  25  36  49  64  81 100
```

Answer: The error was: cannot have a chunk within another chunk

part d

```
library(ggplot2)
ggplot(data = pressure, mapping = aes(x = temperature, y = pressure)) + geom_point() + geom_smooth(meth
```

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



Hint: This part results in a formatting error. It does not generate an error that prevents the document from successfully knitting.

Answer: The error was: dont want a linear model for this method