Homework 1

Charles

2024-01-31

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 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. After you have done this once, you should not need to do so again.

```
# tinytex::install_tinytex()
```

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 with the hashtag commenting it out so that it doesn't run or delete the line of code from your RMarkdown document.

```
# install.packages('ggplot2')
```

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. You will also know that any later errors come from your code rather than a setting in 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 (line 3)
- 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:

$(14-5)^{4+7} \times \frac{8}{23}$

```
(14-5)^(4+7)* (8/23)
```

```
## [1] 10915151168
```

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 $\pi^{7.6}$. Make sure this value is printed.

```
pi^(7.6)
```

```
## [1] 6002.595
```

part c

Create a vector that consists of the integers from 4 to 11, including both 4 and 11. Assign the vector to the variable **z**.

```
z=seq(4,11,1)
```

part d

Using the vector **z** that you created in part c, generate a new vector **y** with values of $\pi^z - \pi^{7.6}$. Print **y**.

```
y = pi^z - pi^(7.6)
print(y)
```

```
## [1]  -5905.186  -5696.575  -5041.206  -2982.302   3485.936  23806.504  87645.453
## [8] 288201.423
```

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.

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

```
## [1] 302361.2
```

Exercise 4: Dataset Basics [10 points]

While it can be very helpful that R is powerful for calculations, 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 the 1960s to 1980s. Below, I adjust the data so that it's easier to work with.

```
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 and the Intro R tutorial on Canvas will be helpful for Exercises 4-7.

part a

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. What is the geographic location where the data was collected from?

Answer: The dataset Seatbelts is about car driver causalities and fatalities from Jan 1969 to Dec 1984

part b

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.

```
dim(sb)
```

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

```
nrow(sb)
```

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

```
ncol(sb)
```

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

Answer:

This dataset has 192 rows & 8 columns.

part c

Print the first 6 rows of the dataset.

```
head(sb, 6)
```

```
##   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 d

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:Drivers killed(col 1) is loosely correlated with the volume of drivers(col 2). I am interested if distance traveled(kms) or petrolprice will also be correlated with deaths.

Exercise 5: Numerical Summaries [15 points]

Suppose that the Department of Transportation is interested in comparing the number of passengers killed or seriously injured based on where they were sitting: in the front (`front`) or in the rear (`rear`).

part a

First, generate numerical summaries for the `front` variable. Be sure to calculate the mean, the five number summary, and the standard deviation.

```
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 b

For the `front` 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: For the front variable, the mean is larger than the median; this indicts that the data is positively skewed. This also means that when plotted, the graph would be right skewed slightly.

part c

Now, calculate the same numerical summaries from **part a** for the 'rear" variable.

```
summary(sb$rear)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   224.0   344.8   401.5   401.2   456.2   646.0
```

```r
sd(sb$rear)
```

```
## [1] 83.10221
```

part d

Compare the numerical summaries for the front-seat passengers from **part a** with the numerical summaries for the rear-seat passengers in **part c**. What do you notice? What real world implications might this have? What limitations are there for your conclusions?

Answer:Based on the available data, if you are in a car accident, the front passengers are twice as likely to be the ones injured. However, a portion of that is due to the fact that all cars have drivers in the front seat and thus every car in an accident has a chance to injure those in the front seat rather than a less than 100% chance that someone is in the rear seats.

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

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

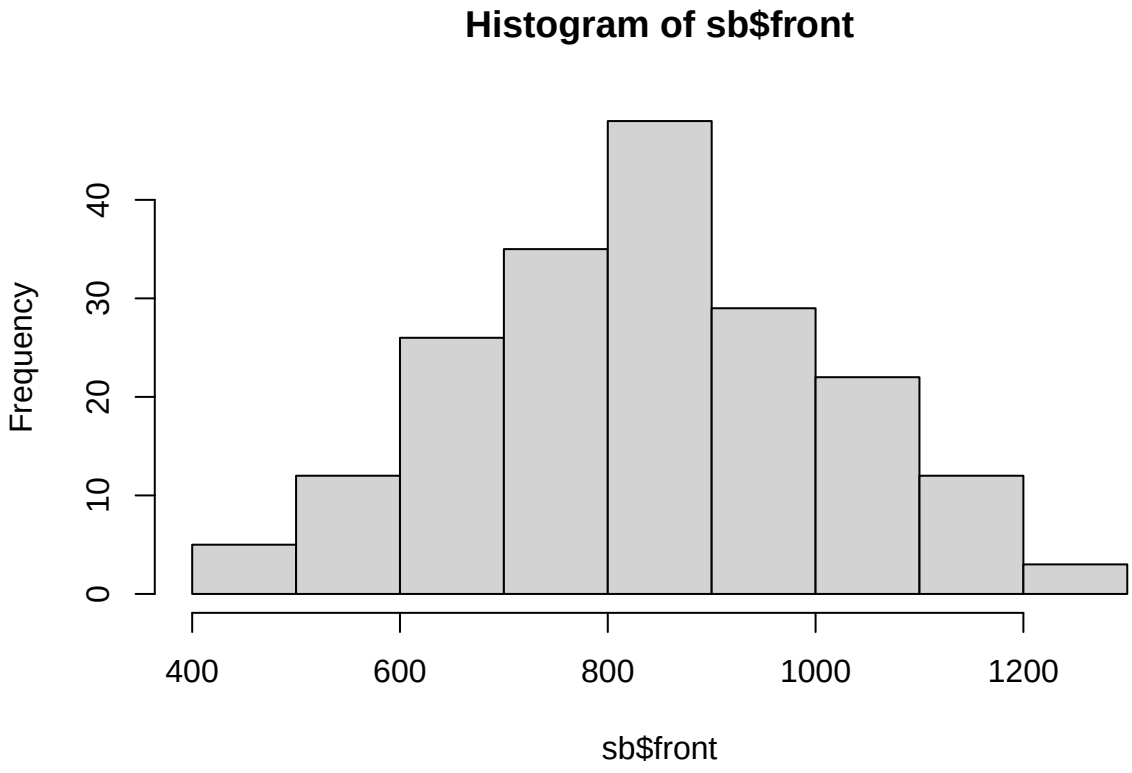
To do this, we'll load the `ggplot2` package, which allows you to create graphics like we did in class.

```r
library(ggplot2)
```

part a

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

```r
hist(sb$front)
```



part b

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

```
hist(sb$rear)
```

part c

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

Answer: From these two graphs, it is notable that both graphs share a very similar shape; They share an overall distribution and central tendency.

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

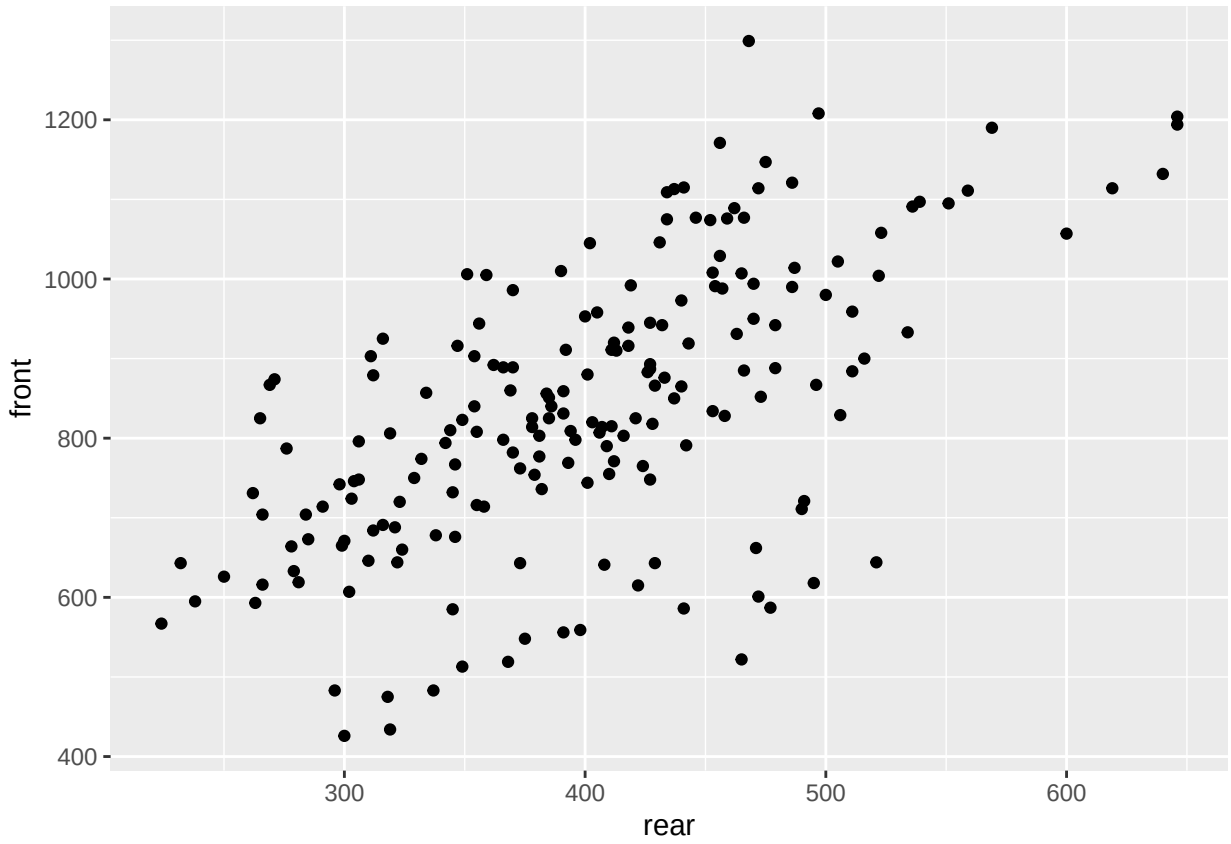
Finally, we'll create a scatterplot of the relationship between deaths and serious injuries of front- compared to those of rear-seat passengers.

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 front-seat passengers on the x axis and the deaths and serious injuries from rear-seat passengers on the y axis.

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

```
ggplot(data = sb, mapping = aes(y = front, x = rear)) +
geom_point()
```

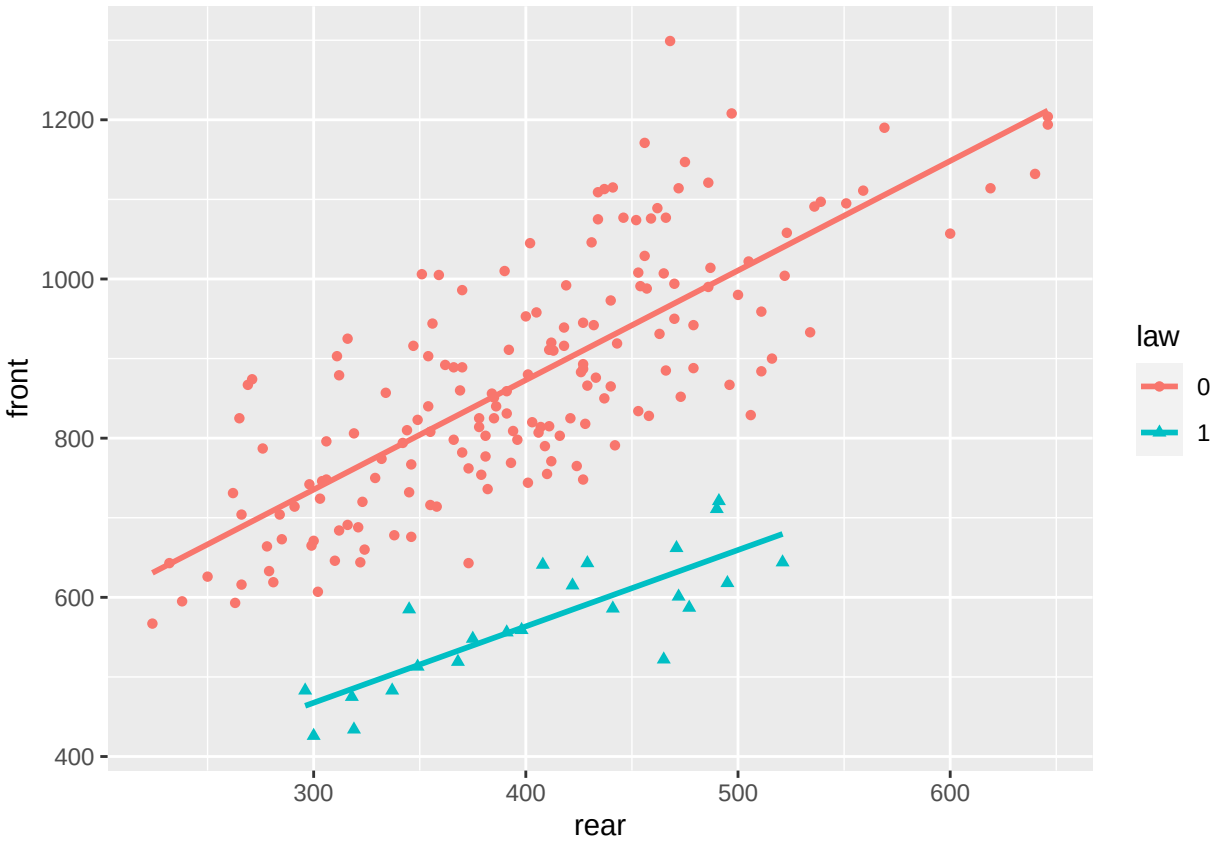


part b

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 = Drop, shape = Track, color = Track)) + geom_point() + geom_smooth(method = 'lm', se = F, formula = y ~ x)

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



part c

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

Answer:

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: If I were to continue analyzing the dataset `sb`, I would try to see how correlated the `front` and `rear` would be if you were to scale the `rear` by two or the `front` by two. Now if I was looking at the dataset holistically, I would look at the correlations between `driverskilled` and `kms`, and then potentially involve either the month or the `petrolprice` to see if either of those also played a part in drivers killed which would mean I could attempt to standardize to remove potential confounding variables. For visual insights, I'd use

scatter plots, correlation matrices, and time series plots. I'd appreciate additional data on vehicle types and weather conditions, as these could impact driver fatalities. My concerns include potential biases in data collection and clarity in definitions; for example, does drivers killed encompass immediate fatalities only, or later ones too? Understanding specifics behind variables like month is crucial to discern patterns accurately.

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: We were not in a chunk for coding.

part b

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

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

Answer: The error was: Needed to create and rename the chunk so it wasn't the same as the one above.

part c

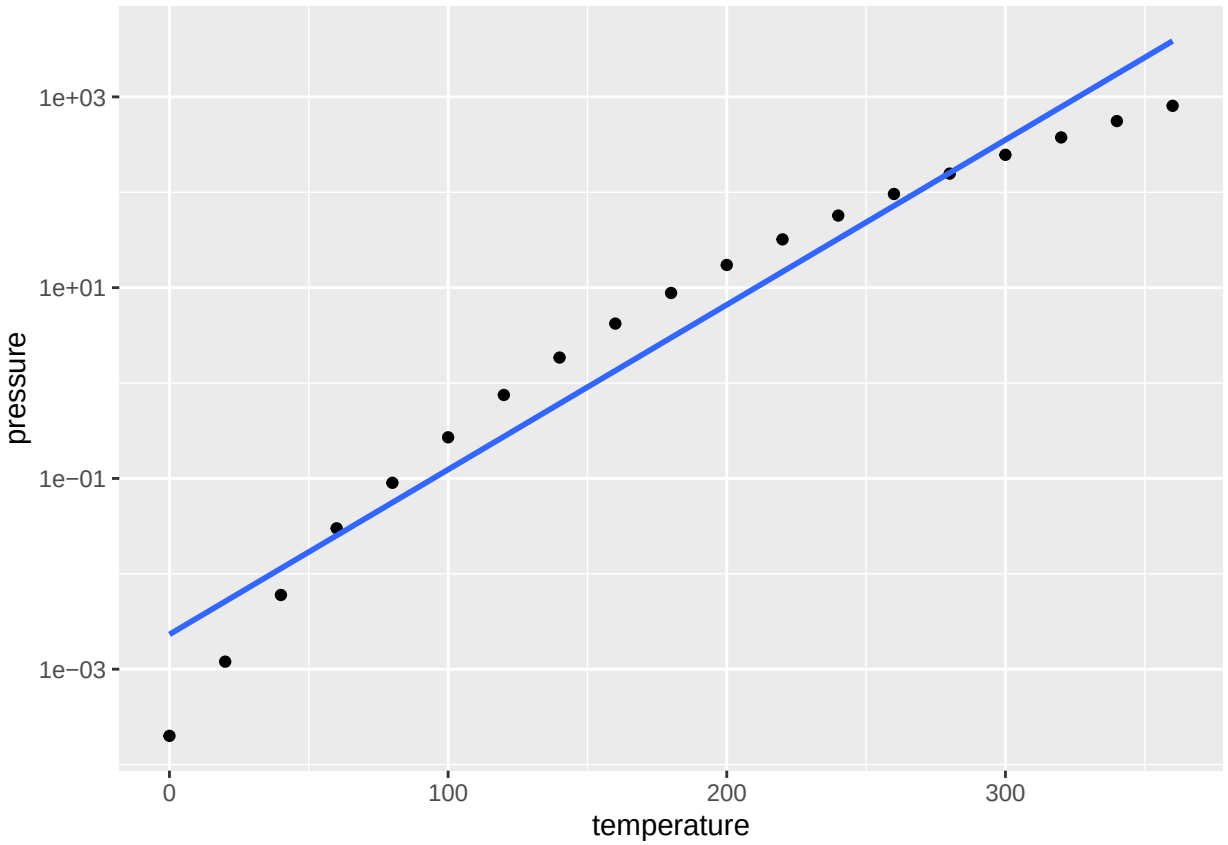
```
(1:10)^2
```

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

Answer: The error was: Misplaced/multiple chunk headers, incomplete chunk, and removed the before the ^ to allow it to be read in the chunk.

part d

```
ggplot(data = pressure, mapping = aes(x = temperature, y = pressure)) + geom_point() + scale_y_log10() +
```



Hint: This part results in a formatting issue. Be sure to look at the pdf version of the document. It does not generate an error that prevents the document from successfully knitting.

Answer: The issue was: Needed to include logarithmic scaling for the y axis so that it would line up better with what the graph is trying to convey.