Homework 3

Anahi Rodriguez

Due 9/14/2022

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

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

We'll use new packages for this homework assignment. You'll need to install the datasauRus package in your Console (just once). You'll want to keep this line commented out when knitting your document. The MASS package should come pre-installed, but you can always confirm by re-installing the package below. The `install.packages` functions should not be run in your RMarkdown document. You may choose to either leave them commented out (retain the hashtag at the beginning of the line) or to delete the starter code chunk.

```
# install.packages('datasauRus')
# install.packages('MASS')
```

```
library(ggplot2)
library(datasauRus)
library(MASS)
```

Exercise 1: Datasaurus [30 points]

For this question, we'll use the data contained in the `datasaurus_dozen` dataset within the `datasauRus` package. Make sure that you have installed and loaded the `datasauRus` package before you begin working on this exercise.

The `datasaurus_dozen` contains three variables:

- dataset, with 13 options
- x, and

- y.

It may help to look at the first few rows of the `datasaurus_dozen` dataset.

```
head(datasaurus_dozen)
```

```
## # A tibble: 6 x 3
##   dataset     x     y
##   <chr>   <dbl> <dbl>
## 1 dino     55.4  97.2
## 2 dino     51.5  96.0
## 3 dino     46.2  94.5
## 4 dino     42.8  91.4
## 5 dino     40.8  88.3
## 6 dino     38.7  84.9
```

part a

Let's begin by creating the following four datasets:

- Create a `dino` object in R for those observations in the `datasaurus_dozen` dataset that take the value `dino` for the variable `dataset`.
- Create a `high_lines` object in R for those observations in the `datasaurus_dozen` dataset that take the value `high_lines` for the variable `dataset`
- Create a `star` object in R for those observations in the `datasaurus_dozen` dataset that take the value `star` for the variable `dataset`
- Create a `x_shape` object in R for those observations in the `datasaurus_dozen` dataset that take the value `x_shape` for the variable `dataset`.

```
# Use this code chunk to answer this question.
dino = subset(datasaurus_dozen, datasaurus_dozen$dataset == 'dino')
high_lines = subset(datasaurus_dozen, datasaurus_dozen$dataset == 'high_lines')
star = subset(datasaurus_dozen, datasaurus_dozen$dataset == 'star')
x_shape = subset(datasaurus_dozen, datasaurus_dozen$dataset == 'x_shape')
```

part b

For each of the four R objects you created in **part a**, report the following statistics:

- number of rows & columns
- mean of x
- mean of y

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

```
## [1] 142   3
```

```
dim(high_lines)
```

```
## [1] 142   3
```

```
dim(star)
```

```
## [1] 142   3
```

```
dim(x_shape)
```

```
## [1] 142   3
```

```
mean(dino$x)
```

```
## [1] 54.26327
```

```
mean(high_lines$x)
```

```
## [1] 54.26881
```

```
mean(star$x)
```

```
## [1] 54.26734
```

```
mean(x_shape$x)
```

```
## [1] 54.26015
```

```
mean(dino$y)
```

```
## [1] 47.83225
```

```
mean(high_lines$y)
```

```
## [1] 47.83545
```

```
mean(star$y)
```

```
## [1] 47.83955
```

```
mean(x_shape$y)
```

```
## [1] 47.83972
```

part c

For each of these four R objects, report the following:

- the correlation of x and y
- the coefficients for the linear model predicting y from x

```
# Use this code chunk to answer this question.
cor(dino$x, dino$y)
```

```
## [1] -0.06447185
```

```
cor(high_lines$x, high_lines$y)
```

```
## [1] -0.06850422
```

```
cor(star$x, star$y)
```

```
## [1] -0.0629611
```

```
cor(x_shape$x, x_shape$y)
```

```
## [1] -0.06558334
```

```
dino_lm = lm(y ~ x, data = dino)
high_lines_lm = lm(y ~ x, data = high_lines)
star_lm = lm(y ~ x, data = star)
x_shape_lm = lm(y ~ x, data = x_shape)

coef(dino_lm)
```

```
## (Intercept)           x
##  53.4529784  -0.1035825
```

```
coef(high_lines_lm)
```

```
## (Intercept)           x
##  53.8087932  -0.1100695
```

```
coef(star_lm)
```

```
## (Intercept)           x
##   53.326679   -0.101113
```

```
coef(x_shape_lm)
```

```
## (Intercept)           x
##  53.5542263  -0.1053169
```

part d

What do you notice from your results in **parts b & c**? What might be the underlying cause of your results from each of these datasets?

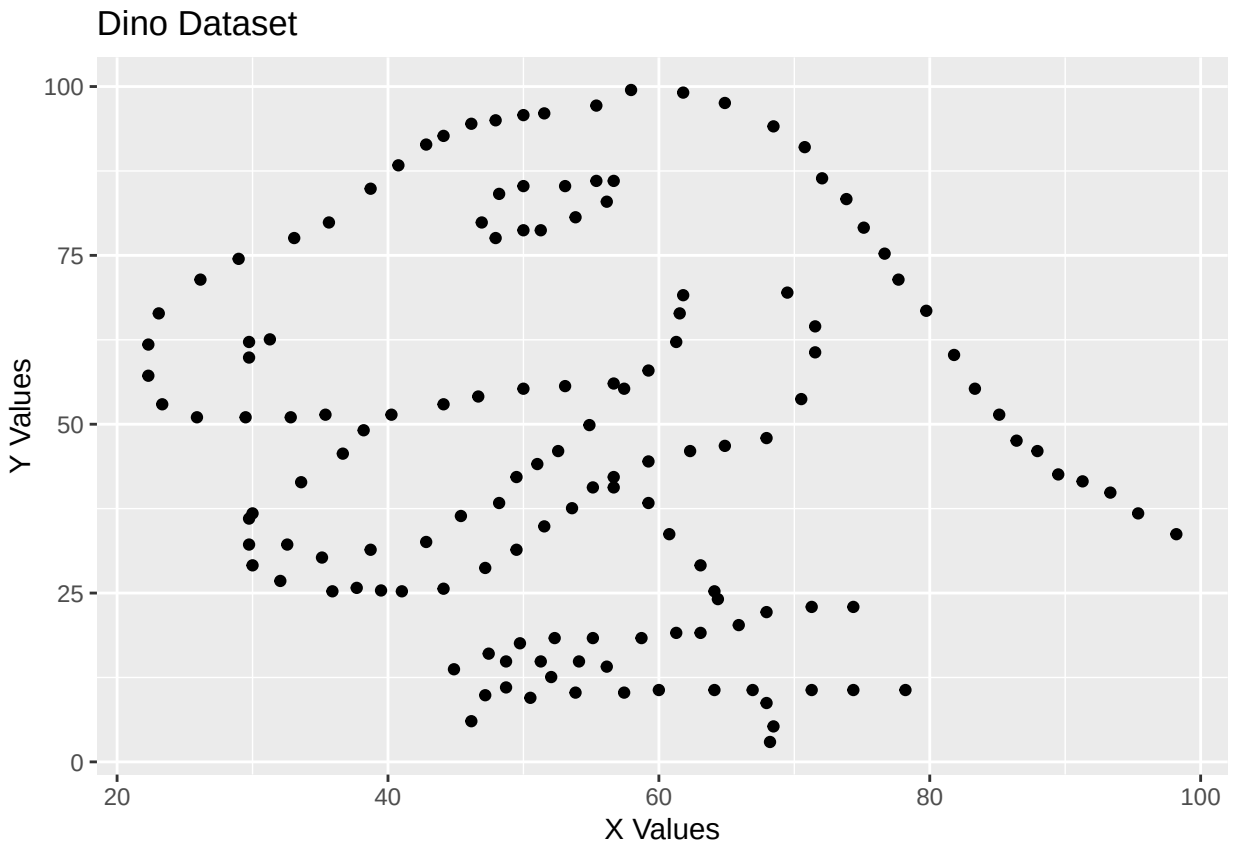
Note: there is a correct answer for the first question of d. The second question is asking you to speculate as to what might be occurring without a correct answer.

Answer: The average x value within these 4 datasets is very close to the intercept from the models when x is used to predict y. I believe that this means that the values of the x-coordinate and the y-coordinate are very close values for each point.

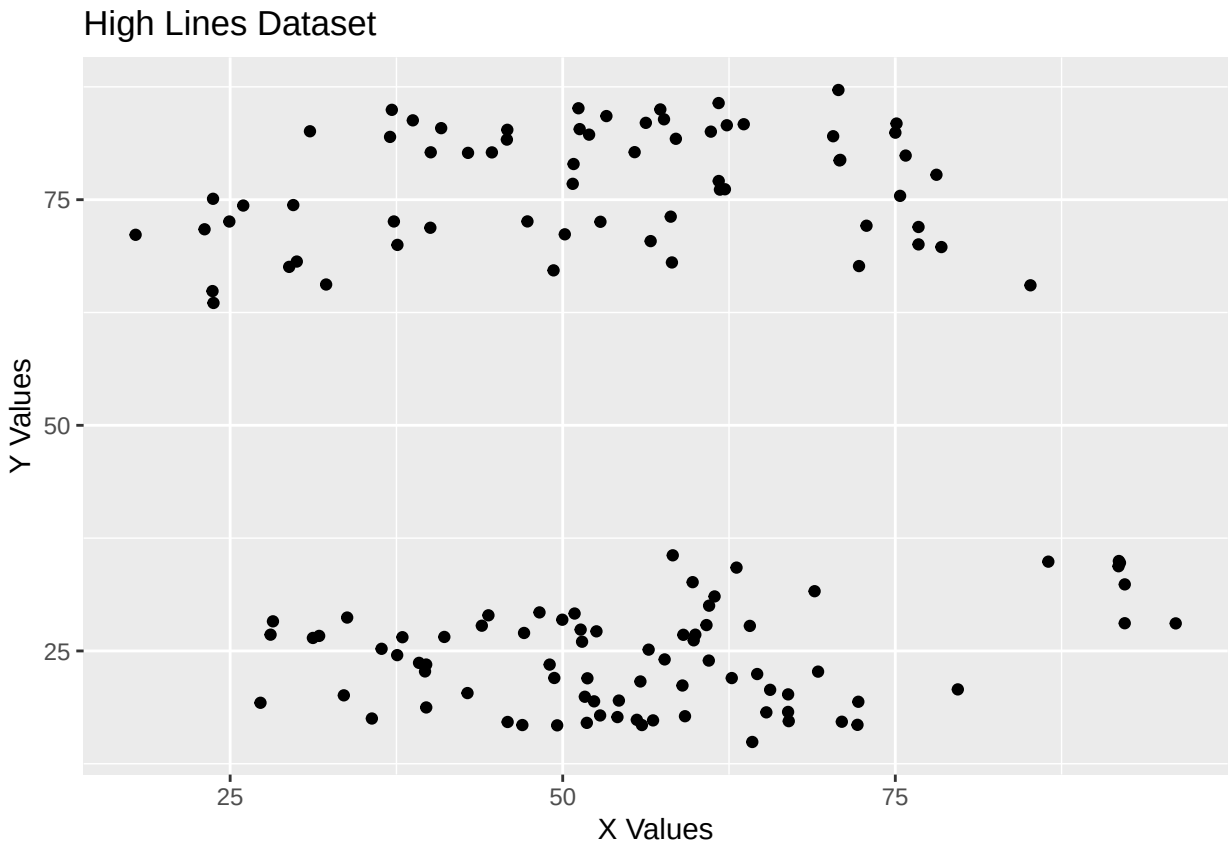
part e

Graph each of the datasets, using the x and y variables for their respective axes. Include the dataset name in the title of each graph. Axes labels of x and y are sufficient for this problem.

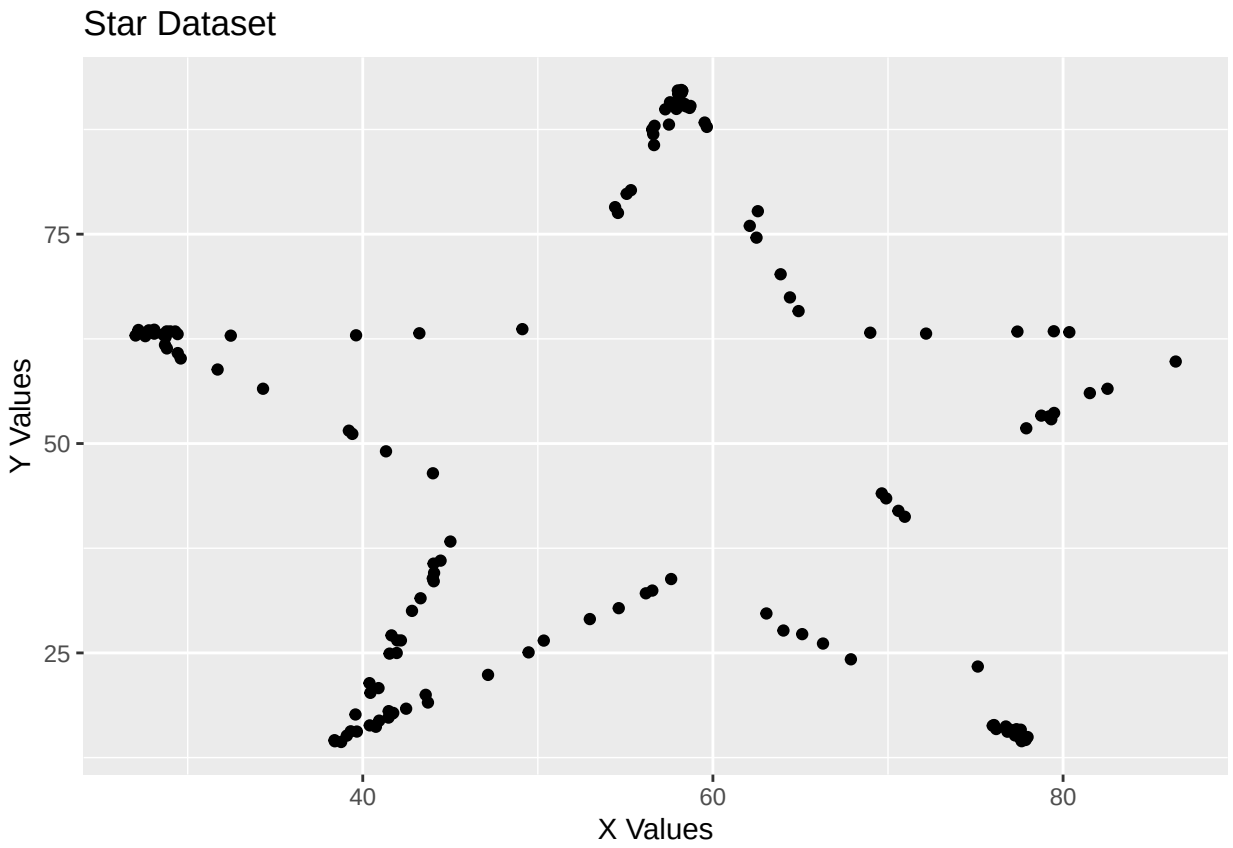
```
# Use this code chunk to answer this question.
ggplot(data = dino, aes(x = x, y = y)) +
  geom_point() +
  labs(x = 'X Values', y = 'Y Values', title = 'Dino Dataset')
```



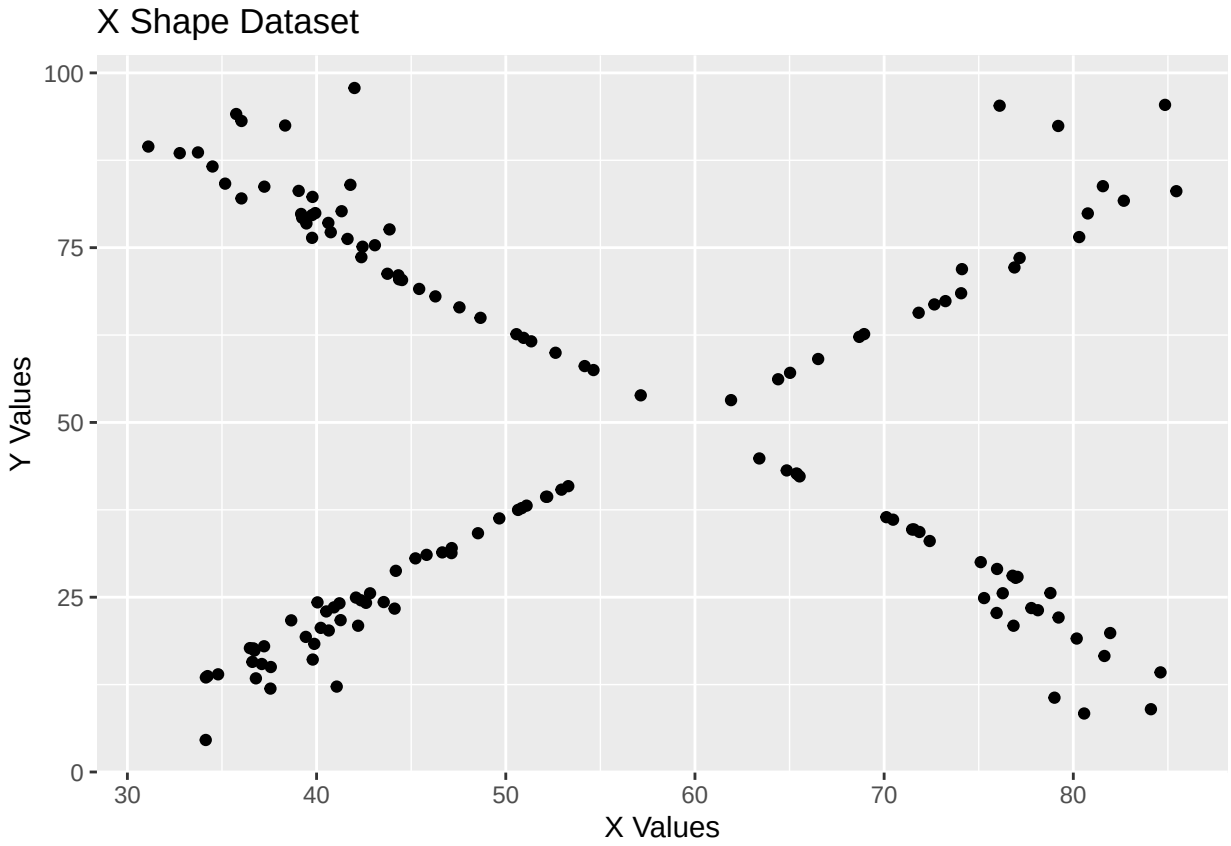

```
ggplot(data = high_lines, aes(x = x, y = y)) +
  geom_point() +
  labs(x = 'X Values', y = 'Y Values', title = 'High Lines Dataset')
```




```
ggplot(data = star, aes(x = x, y = y)) +
  geom_point() +
  labs(x = 'X Values', y = 'Y Values', title = 'Star Dataset')
```




```
ggplot(data = x_shape, aes(x = x, y = y)) +
  geom_point() +
  labs(x = 'X Values', y = 'Y Values', title = 'X Shape Dataset')
```



part f

What did you observe from the graphs? What is a takeaway point from this exercise?

Answer: I observed that these graphs draw the shape in the variable 'dataset'. Filtering data can give us a better idea of what makes up the observations in a variable.

Exercise 2: Variable Types [10 points]

Indicate the variable type for each of the variables described. Be sure to be specific, including the general and specific labels (e.g. quantitative discrete).

part a

Popularity of a roller coaster, as measured by the number of riders on September 1, 2022.

Answer: quantitative discrete

part b

Theme park where a roller coaster is located.

Answer: nominal categorical

part c

Time (in minutes) a roller coaster was operating on September 1, 2022.

Answer: quantitative continuous

part d

Target audience for the roller coaster, based on age grouping.

Answer: ordinal categorical

part e

How many breakdown events resulting in a pause from operation that a roller coaster experienced between September 1, 2020 and August 31, 2021.

Answer: quantitative discrete

Exercise 3: Variable Roles & Study Types [15 points]

For each of the following proposed studies, indicate the **variable roles** for each variable described. Is the study described **experimental or observational**?

part a

Yanis suspects that the **eldest child** in a family grows to be the shortest adult, and that the **youngest child** grows to be tallest. Berza reminds Yanis that **adult height** is also affected by **sex**, so Yanis decides to record that as well for all participants. Yanis recruits adult participants with at least one sibling for this study.

Variables:

- Birth order (eldest vs. youngest)
- Adult Height
- Sex

Answer: Birth order is an explanatory variable, adult height is the response variable, and sex is a cofounder variable. The study is experimental.

part b

Do your friends approach mealtime the same way that you do? Some students report when they come to college that they eat faster than their friends do. One student, Alex, speculates that the **speed with which you eat a meal** is determined by where you are from **geographically**. Jennifer reminds him that additional factors, like how much you **talk while you eat**, are also related both to your geographic region and to how long it takes to eat a meal. Fernando finds this theory interesting, and so decides to gather data on these variables from a campus dining hall.

Variables:

- Meal Time
- Geographic Region
- Talk Time during Meal

Answer: Geographic region is an explanatory variable, Meal time is the response variable, and talk time during a meal is a cofounder variable. This study is observational.

part c

Inspired by Fernando's study in the dining hall, Brenda designs her own study. Brenda wants to know if **ordering choices** and **eating time** depend on **how many people you are seated with**. Brenda designs a study where entrants to the dining hall are randomly assigned to eat at a table by themselves, with 1 friend, with 2 friends, or with 3 friends. The food ordered and the time spent eating are both recorded.

Variables:

- Ordering Choices
- Meal Time
- Number of Dining Companions

Answer: Ordering choices is a cofounder, eating time is the response variable, and number of dining companions is the explanatory variable. This study is experimental.

Exercise 4: Birthweight, Descriptive Summaries [15 points]

In this exercise, we'll work with the `birthwt` dataset contained within the `MASS` package. Read through the documentation using the Help command below. If you would like to prevent new browser windows from reopening every time you knit the document, you may opt to comment this line of code out by adding a hashtag at the beginning of the line.

```
#?birthwt
```

part a

How many observations are in this dataset (use `R` function)? How many variables (use `R` function)? Where and when was this data collected (not from an `R` function)? Provide these details in a sentence after your code block.

```
# Use this code chunk for your answer.
dim(birthwt)
```

```
## [1] 189  10
```

Answer: There are 189 observations and 10 variables. The observations from this data were collected at Baystate Medical Center, Springfield, Mass in 1986.

part b

We'll be using this dataset to predict baby's birthweight using the other variables in the dataset. In this part, we'll think about reasons that causality might be plausible for this specific scenario and reasons causality might not be supported based on the underlying behavior of the variables of interest. Without performing any numerical analyses, what reason(s) support a causal relationship between the other variables, excluding the indicator of a low birth weight, and the baby's birthweight? What reason(s) undermine any determination of causality or suggest that causality might not be a reasonable explanation for these variables?

Answer: We might think that smoking may be associated with the number of previous premature labours, mothers weight in pounds at last menstrual cycle, history of hypertension, presence of uterine irritability, and number of physician visits during the first trimester as smoking generally negatively affects one's health. Additionally, age can also have some casual relationship with many of the same variables as smoking does because similar to smoking, age will also have an adverse effect on your health and risks associated with your health. However, on the other hand, since so many of these variables are related to each other in different directions and magnitudes, it may be difficult to seperate which variables actually have a casual relationship with one another.

part c

Create a correlation matrix of the baby's birthweight, the mother's age, and the mother's weight. Which of the possible explanatory variables has the highest correlation with the baby's birthweight?

```
# Use this code chunk for your answer.
cor(birthwt[,c(10, 2, 3)])
```

```
##            bwt        age       lwt
## bwt 1.00000000 0.09031781 0.1857333
## age 0.09031781 1.00000000 0.1800732
## lwt 0.18573328 0.18007315 1.0000000
```

Answer: The mother's weight in pounds at her last menstrual cycle has the highest correlation with baby's birth weight.

part d

A doctor would like to analyze the birth weight data with an aim of generating results that could be applied to her current patients. Is this an appropriate use of the dataset? Explain.

Answer: I do think that this is an appropriate use of the dataset as long as most, if not all, the variables are included in a regression as different effects of other variables can be controlled for. However, correlations alone would not be an appropriate use as there exists many cofounding variables that effect these relationships and can give a misleading conclusion.

part e

Thinking critically about this data, are there additional variables that could be added? Do you have concerns about how this data might be used? Any additional information you'd like to know about the data?

Answer: I think that there could be an extension of the race variable to include more than simply 3 categories. This data could also include income as the health of an individual and their family is heavily determined by their wealth. Additionally, height of both the mother and the father could also give a better picture of the situation. And laslty, any other variables pertaining to the health of the mother would only make the conclusions stronger as its common knowledge that the mother's health will affect the child somehow. I am concenred that this data could be misleading if certain variables such as history of hypertension or other health defects/issues are used because this is a variable that is also affected by the variable smoking and without these cofounders included, results can be under/over biased in some direction.

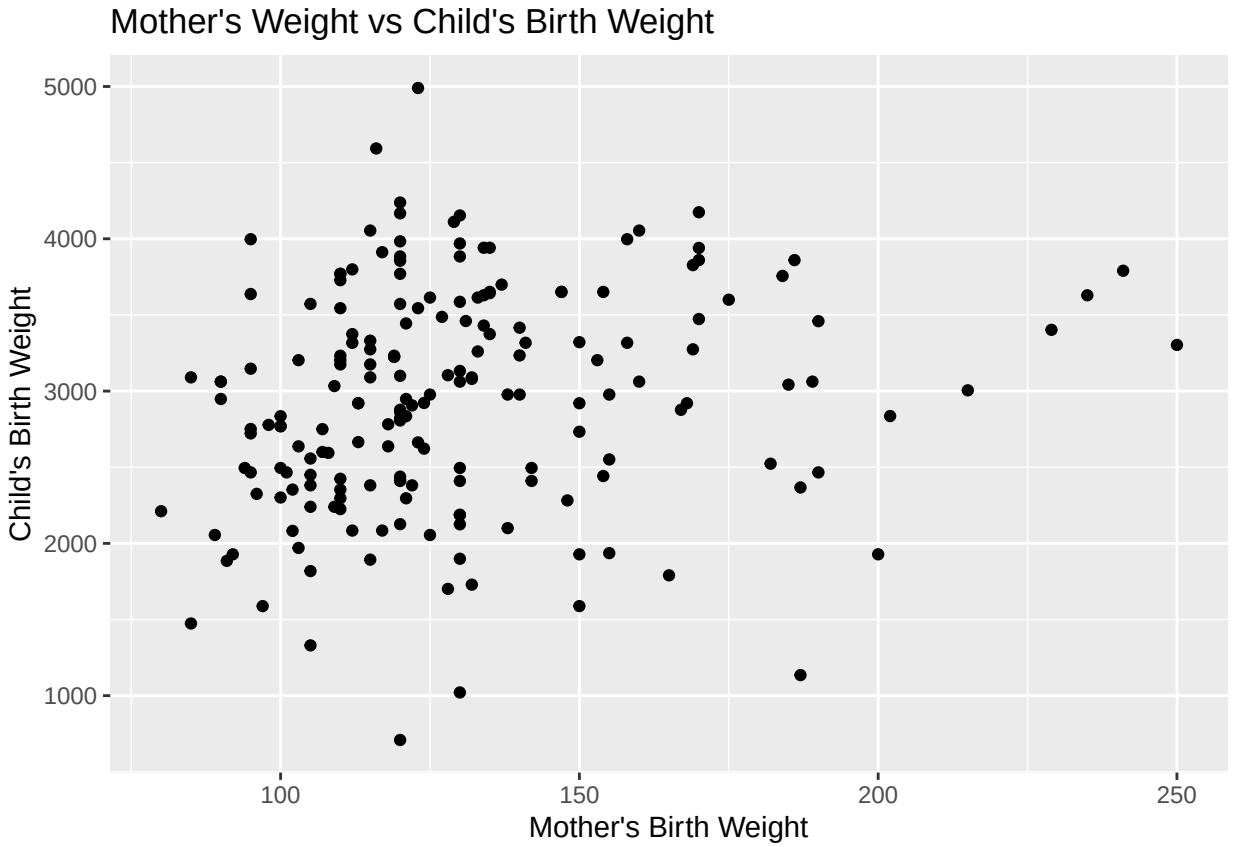
Exercise 5: Interpreting a Linear Model for Birthweight [25 points]

We'll continue analyzing the `birthwt` dataset that we started looking at in the last Exercise. For this question, we'll focus on the variables `bwt` and `lwt`.

part a

Visualize the relationship between the mother's pre-pregnancy weight and the baby's weight. Make sure to also provide appropriate titles and axes labels for your graph. Then, interpret this relationship.

```
# Use this code chunk for your answer.
ggplot(data = birthwt, aes(x = lwt, y = bwt)) +
  geom_point() +
  labs(title = "Mother's Weight vs Child's Birth Weight",
       x = "Mother's Birth Weight",
       y = "Child's Birth Weight")
```



Answer: There exists a weak but positive relationship between the mother's weight and the child's birth weight. As the mother's weight increases, so does the child's birth weight increase on average slightly. This relationship may be linear but it seems that perhaps a different type of relationship may fit better. Additionally, there do exists a few outliers that come from a greater mother's weight and maybe one or two from the child's birth weight.

part b

Fit a linear model that predicts the baby's weight from the mother's pre-pregnancy weight. Write that model out below.

```
# Use this code chunk for your answer.
lm = lm(bwt ~ lwt, data = birthwt)
summary(lm)
```

```
##
## Call:
## lm(formula = bwt ~ lwt, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2192.12  -497.97    -3.84   508.32  2075.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2369.624    228.493  10.371   <2e-16 ***
## lwt            4.429      1.713   2.585   0.0105 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 718.4 on 187 degrees of freedom
## Multiple R-squared:  0.0345, Adjusted R-squared:  0.02933
## F-statistic: 6.681 on 1 and 187 DF,  p-value: 0.0105
```

Answer: estimated child birth weight = 2369.624 + (Mother's weight * 4.429)

part c

Interpret each of the fitted coefficients (intercept and slope) for this model.

Answer: On average, when the mother's weight is 0 pounds, we expect the average birth weight of their child to be 2369.624 grams. For each additional pound heavier the mother is, expect on average for the birth weight of the baby to increase by 4.429 grams.

part d

Calculate the estimated mean baby's birthweight for a mother with a pre-pregnancy weight of 147 pounds. What is the residual for a mother with a pre-pregnancy weight of 147 and a baby's birthweight of 3000 g.

```
# Use this code chunk (if needed) for your answer.
3000 - (2369.624 + (4.429 * 147))
```

```
## [1] -20.687
```

Answer: the residual is -20.687 grams

part e

One of the mother's weights was accidentally removed from the dataset. However, we know the corresponding baby's weight (2743 g) and the residual (-40 g). What was the original mother's weight?

```
# You may use this code chunk for your calculation, or you may type your calculation below.
observed = 2743
# -40 = 2743 - x
predicted = (-2743 - 40)/-1
predicted
```

```
## [1] 2783
```

```
# predicted = 2369.624 + (4.429 * x)
(2783 - 2369.624)/4.429
```

```
## [1] 93.33394
```

```
# check answer
2369.624 + (93.33394* 4.429 )
```

```
## [1] 2783
```

```
observed - (2783)
```

```
## [1] -40
```

Answer: The mothers weight was originally 93.33394 pounds

Exercise 6: 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 on the document
- properly assigned pages to exercises on Gradescope
- select **page 1 (with your name)** and this page for this exercise (Exercise 6)
- all code is printed and readable for each question
- generated a pdf file