Homework 3

Charles Ancel

9/13/2023

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 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.

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

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

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

Exercise 2: 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.

```r
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 `dots` object in R for those observations in the `datasaurus_dozen` dataset that take the value `dots` for the variable `dataset`
- Create a `circle` object in R for those observations in the `datasaurus_dozen` dataset that take the value `circle` for the variable `dataset`
- Create a `wide_lines` object in R for those observations in the `datasaurus_dozen` dataset that take the value `wide_lines` for the variable `dataset`.

```r
dino <- datasaurus_dozen[datasaurus_dozen$dataset == "dino", ]
dots <- datasaurus_dozen[datasaurus_dozen$dataset == "dots", ]
circle <- datasaurus_dozen[datasaurus_dozen$dataset == "circle", ]
wide_lines <- datasaurus_dozen[datasaurus_dozen$dataset == "wide_lines", ]
```

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

```r
dino_stats <- list(
  rows_columns = dim(dino),
  mean_x = mean(dino$x),
  mean_y = mean(dino$y)
)
```

```r
dots_stats <- list(
  rows_columns = dim(dots),
  mean_x = mean(dots$x),
  mean_y = mean(dots$y)
)

circle_stats <- list(
  rows_columns = dim(circle),
  mean_x = mean(circle$x),
  mean_y = mean(circle$y)
)

wide_lines_stats <- list(
  rows_columns = dim(wide_lines),
  mean_x = mean(wide_lines$x),
  mean_y = mean(wide_lines$y)
)

list(dino = dino_stats, dots = dots_stats, circle = circle_stats, wide_lines = wide_lines_stats)
```

```
## $dino
## $dino$rows_columns
## [1] 142   3
##
## $dino$mean_x
## [1] 54.26327
##
## $dino$mean_y
## [1] 47.83225
##
##
## $dots
## $dots$rows_columns
## [1] 142   3
##
## $dots$mean_x
## [1] 54.2603
##
## $dots$mean_y
## [1] 47.83983
##
##
## $circle
## $circle$rows_columns
## [1] 142   3
##
## $circle$mean_x
## [1] 54.26732
##
## $circle$mean_y
## [1] 47.83772
##
##
```

```
## $wide_lines
## $wide_lines$rows_columns
## [1] 142   3
##
## $wide_lines$mean_x
## [1] 54.26692
##
## $wide_lines$mean_y
## [1] 47.8316
```

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

```
# For the dino dataset
dino_cor <- cor(dino$x, dino$y)
dino_lm <- lm(y ~ x, data = dino)
dino_coeffs <- coef(dino_lm)

# For the dots dataset
dots_cor <- cor(dots$x, dots$y)
dots_lm <- lm(y ~ x, data = dots)
dots_coeffs <- coef(dots_lm)

# For the circle dataset
circle_cor <- cor(circle$x, circle$y)
circle_lm <- lm(y ~ x, data = circle)
circle_coeffs <- coef(circle_lm)

# For the wide_lines dataset
wide_lines_cor <- cor(wide_lines$x, wide_lines$y)
wide_lines_lm <- lm(y ~ x, data = wide_lines)
wide_lines_coeffs <- coef(wide_lines_lm)

# Print out the results
list(
  dino = list(correlation = dino_cor, coefficients = dino_coeffs),
  dots = list(correlation = dots_cor, coefficients = dots_coeffs),
  circle = list(correlation = circle_cor, coefficients = circle_coeffs),
  wide_lines = list(correlation = wide_lines_cor, coefficients = wide_lines_coeffs)
)
```

```
## $dino
## $dino$correlation
## [1] -0.06447185
##
## $dino$coefficients
## (Intercept)           x
##  53.4529784  -0.1035825
##
##
## $dots
## $dots$correlation
## [1] -0.06034144
##
## $dots$coefficients
## (Intercept)           x
##  53.0983419  -0.0969127
##
##
## $circle
## $circle$correlation
## [1] -0.06834336
##
## $circle$coefficients
## (Intercept)           x
##  53.7970450  -0.1098143
##
##
## $wide_lines
## $wide_lines$correlation
## [1] -0.06657523
##
## $wide_lines$coefficients
## (Intercept)           x
##  53.6349489  -0.1069408
```

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. We are not looking for one correct answer for the second question.

Answer: From **parts b & c**:

1. **Similarity:** The datasets may display different visual patterns but have similar statistical properties (means, correlations, and coefficients).

2. **Deceptive Summary Statistics:** Relying solely on summary statistics can be misleading, underscoring the importance of data visualization.

Speculation on underlying causes:

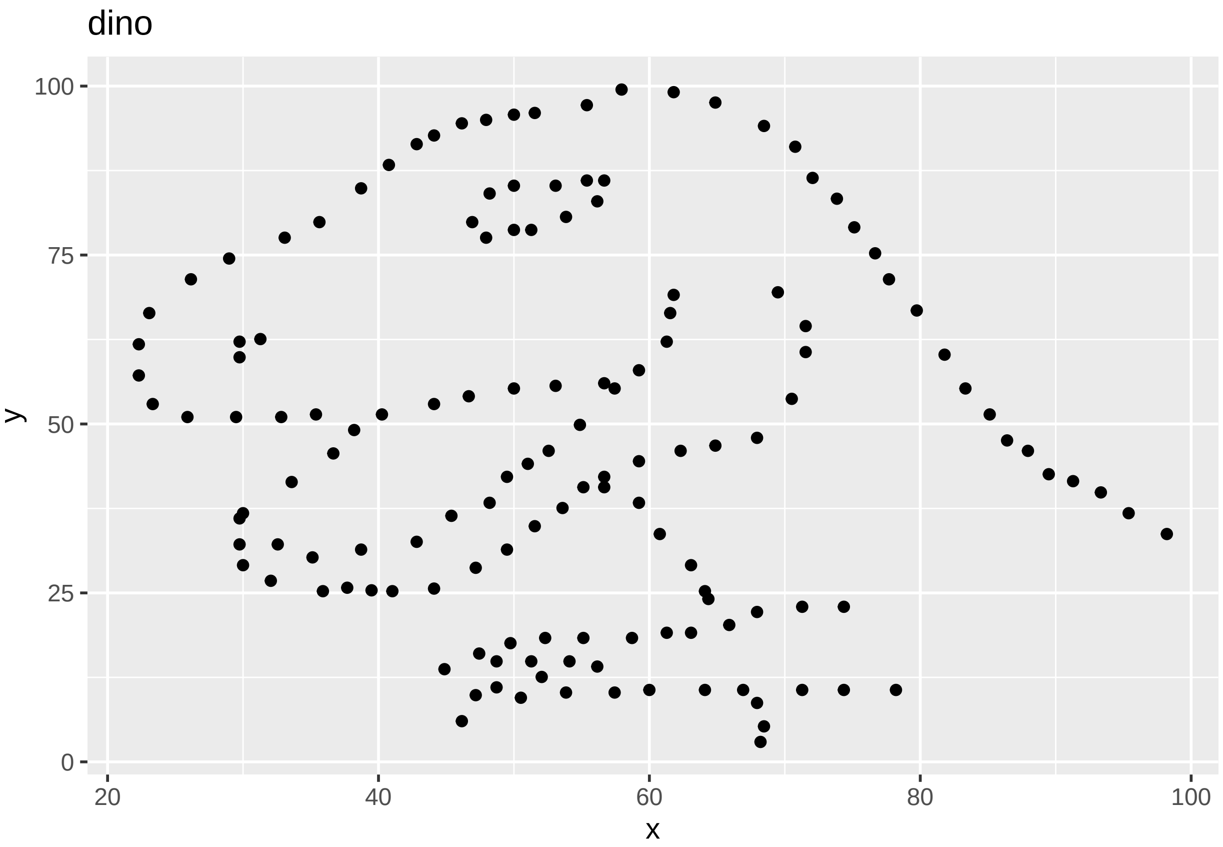
- **Intentional Design:** The datasets in the `datasaurus_dozen` were crafted to emphasize the value of visual data exploration, showing that datasets can have similar statistics but differ visually.

- **Lesson:** This emphasizes that while statistics provide insights, they can't replace visual data exploration. Always visualize your data to understand its nuances.

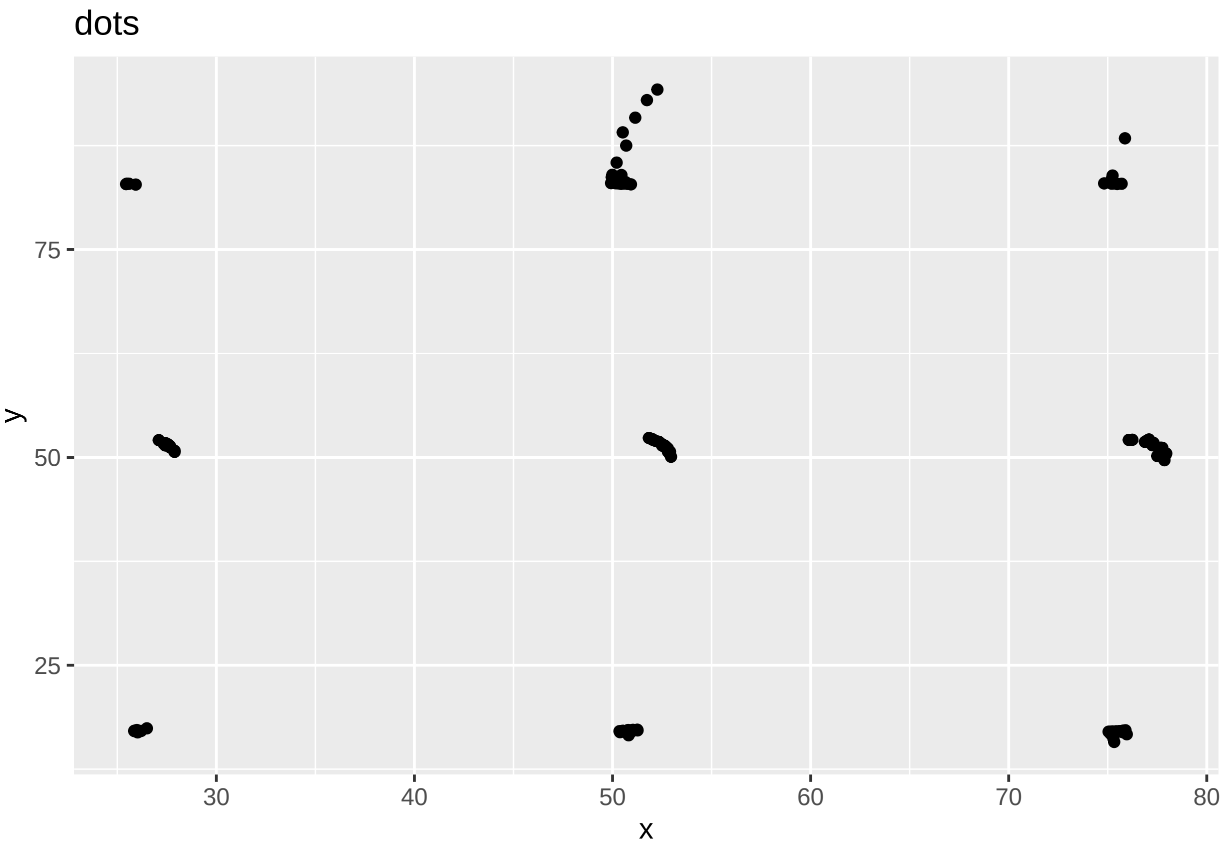
part e

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.

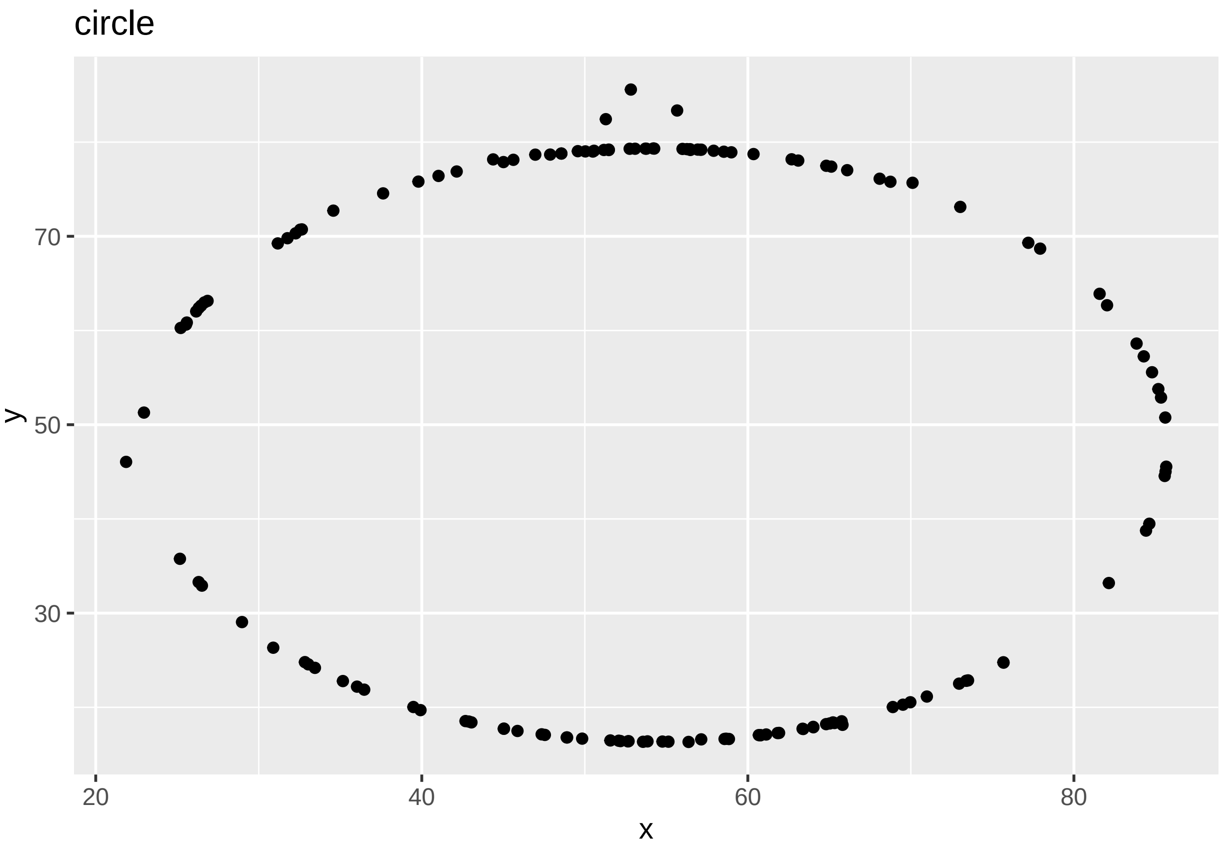
```r
# Dino plot
ggplot(dino, aes(x, y)) +
  geom_point() +
  labs(title = "dino", x = "x", y = "y")
```



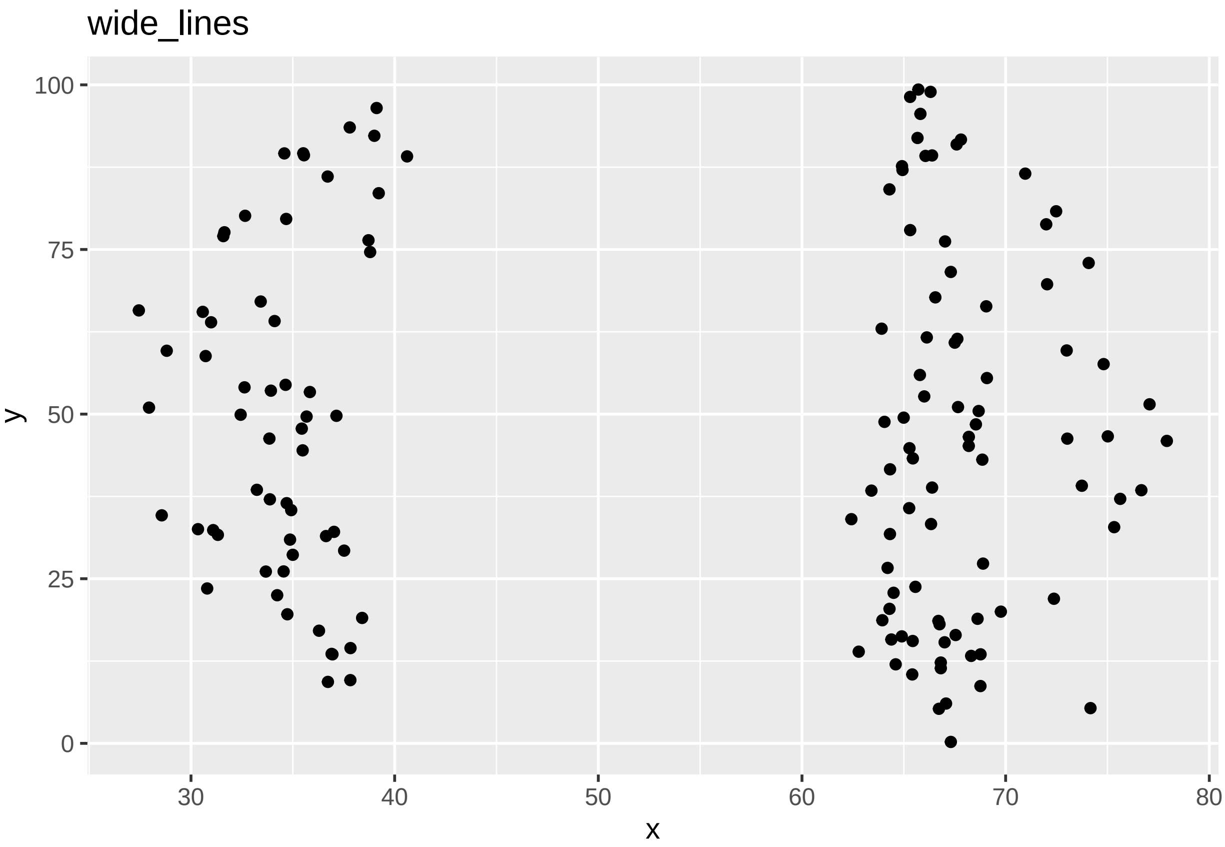
```
# Dots plot
ggplot(dots, aes(x, y)) +
  geom_point() +
  labs(title = "dots", x = "x", y = "y")
```



```
# Circle plot
ggplot(circle, aes(x, y)) +
  geom_point() +
  labs(title = "circle", x = "x", y = "y")
```



```r
# Wide_lines plot
ggplot(wide_lines, aes(x, y)) +
  geom_point() +
  labs(title = "wide_lines", x = "x", y = "y")
```



part f

What did you observe from the graphs? Why is visualizing data a crucial first step before fitting a model?

Answer:

Observations from the Graphs:

- **Dino:** As the name suggests, the data points form a shape resembling a dinosaur.
- **Dots:** This dataset displays a scattered pattern, much like random dots.
- **Circle:** The data points form a distinct circular shape.
- **Wide Lines:** The data appears to be distributed along two parallel wide lines.

Importance of Visualizing Data:

1. **Reveal Data Structure:** Visuals make apparent the actual structure or pattern of the data, which might not be deduced from statistics alone.
2. **Informative Insights:** Visual patterns, like shapes, can offer context or insights about the nature or source of the data.
3. **Avoid Assumptions:** Relying solely on summary statistics can lead to incorrect assumptions about data distribution or behavior.
4. **Guide Analysis:** Recognizing data patterns can guide subsequent analysis or model choice.

In essence, while statistics provide a numerical summary, visuals offer a tangible understanding, emphasizing the need for both in data analysis.

Exercise 3: Variable Types [10 points]

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

part a

The most popular movie (as determined by the box office sales for the weekend).

Answer: Categorical (nominal)

part b

Film ratings assigned by the Motion Picture Association.

Answer: Categorical (ordinal) - Ratings have a specific order but the differences between them aren't uniform (e.g., the difference between PG and PG-13 is not necessarily the same as between PG-13 and R).

part c

Production time for a film.

Answer: Quantitative (continuous) - Production time can take any value within a range and can be measured with great precision, e.g., 2.5 months.

part d

Number of employees on the payroll for a film.

Answer: Quantitative (discrete) - The number of employees is a countable quantity.

part e

How many Oscar nominations a film receives.

Answer: Quantitative (discrete) - The number of nominations is also a countable quantity.

Exercise 4: 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 (eldest vs. youngest):** Explanatory variable (or independent variable)

- **Adult Height:** Response variable (or dependent variable)
- **Sex:** Confounding variable (a variable that might be an extraneous influence on the study, potentially affecting the relationship between the explanatory and response variables)

Type of Study:

This study is **observational**. Yanis is not manipulating any variables but rather observing and recording data from participants as they are.

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:** Explanatory variable (or independent variable)
- **Meal Time (speed with which you eat a meal):** Response variable (or dependent variable)
- **Talk Time during Meal:** Confounding variable (it's believed to be related to both the geographic region and meal time)

Type of Study:

This study is **observational**. Fernando is collecting data without manipulating any variables.

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:

- **Number of Dining Companions:** Explanatory variable (or independent variable) - This is being manipulated by Brenda in the study.
- **Ordering Choices:** Response variable (or dependent variable)
- **Meal Time:** Response variable (or dependent variable) - It's another outcome that Brenda is studying.

Type of Study:

This study is **experimental**. Brenda is actively manipulating a variable (number of dining companions) to observe its effects on the outcomes (ordering choices and meal time).

Exercise 5: Birthweight, Descriptive Summaries [13 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 an `R` function)? How many variables (use an `R` function)? Where and when was this data collected (not from an `R` function)? Provide these details in a sentence after your code block.

```
library(MASS)
# Getting the number of observations
num_observations <- nrow(birthwt)

# Getting the number of variables
num_variables <- ncol(birthwt)

num_observations
```

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

```
num_variables
```

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

Answer: The birthwt dataset contains 189 observations and 10 variables. The data was collected at Baystate Medical Center, Springfield, Mass during 1986.

part b

We'll be using this dataset to predict the birthweight of babies 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?

Hint: You might want to read about the data more or examine the variables in the data to help answer this question.

Note: You are not being asked to determine if causality is at play here. Instead, you are giving two reasons - one that indicates causality could be reasonable and one that questions whether causality is reasonable.

Answer:

Support for Causality: - Biological Foundations: Some variables, like mother's age, weight, and smoking status, have well-understood biological implications on birthweight, indicating a possible causal link.

Challenges to Causality: - Potential Confounders: Not all influencing factors might be captured in the dataset, which can confound perceived relationships. **- Nature of the Data:** The dataset is observational. Without interventions or controls, determining clear causality becomes complex.

In summary, while there are compelling reasons to consider causal relationships, the dataset's structure and potential omissions warrant a cautious approach to causal conclusions.

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?

```r
# Assuming you have loaded the MASS package and its dataset
library(MASS)

# Selecting the relevant columns
selected_data <- birthwt[, c("bwt", "age", "lwt")]

# Creating the correlation matrix
cor_matrix <- cor(selected_data)

cor_matrix
```

```
##            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 (`lwt`) has the highest correlation with the baby's birthweight (`bwt`) compared to the mother's age.

part d

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:

Additional Variables: 1. **Prenatal Care:** Frequency of medical check-ups during pregnancy. 2. **Maternal Diet:** Nutritional intake during pregnancy. 3. **Environmental Exposures:** Exposure to toxins or pollutants. 4. **Genetic Factors:** Family health history or genetic conditions. 5. **Stress Levels:** Mother's mental well-being during pregnancy. 6. **Substance Use:** Consumption of alcohol, drugs, or medications.

Concerns About Data Use: 1. **Over-simplification:** Limited variables might not capture the full picture. 2. **Ethical Concerns:** Risk of stigmatizing certain demographic groups. 3. **Potential Bias:** Without data collection context, interpretations might be skewed.

Additional Information Desired: 1. **Collection Method:** How participants were chosen. 2. **Handling Confounders:** Efforts to control external influencing factors. 3. **Data Timeline:** Time frame and external events during data collection.

In essence, while the data offers insights, understanding its context, potential gaps, and implications is crucial.

Exercise 6: Interpreting a Linear Model for Birthweight [27 points]

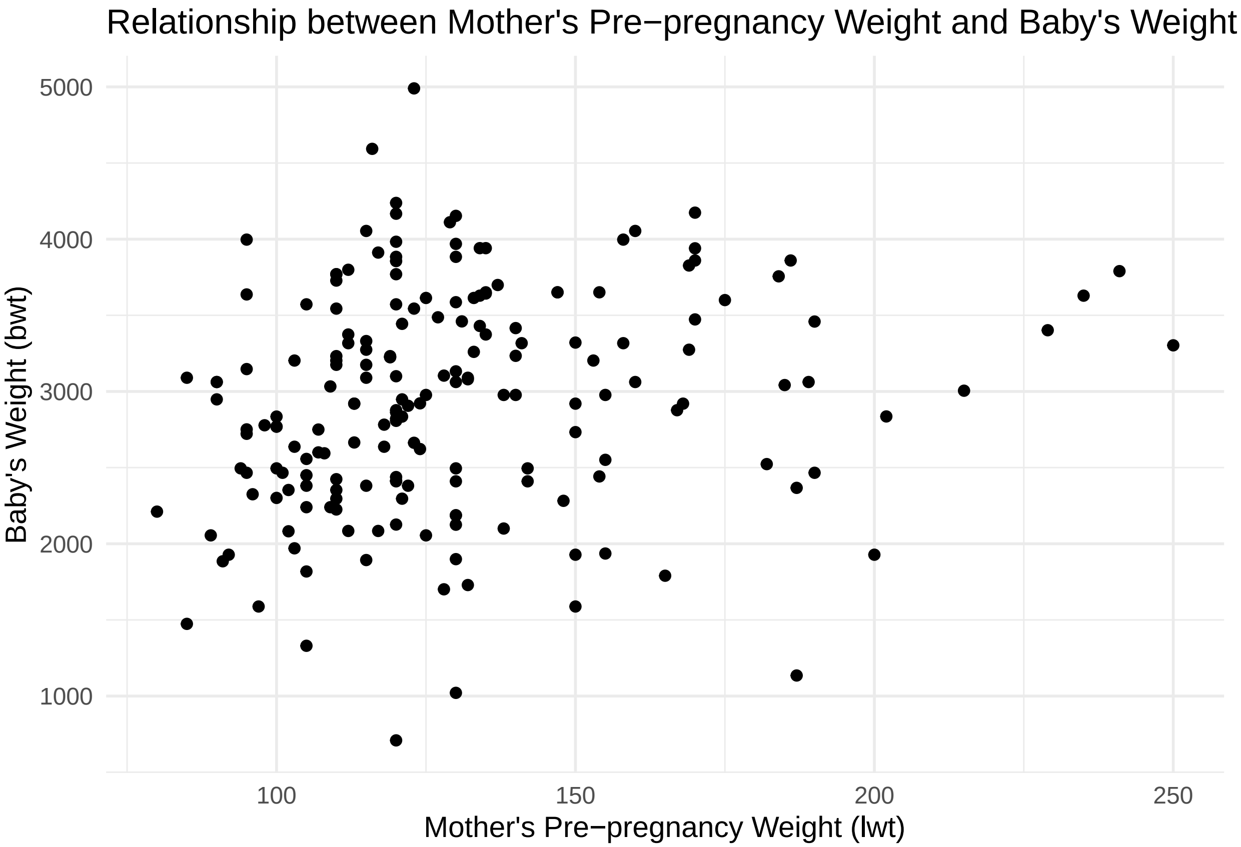
We'll continue analyzing the `birthwt` dataset that we started looking at in the last Exercise. For this question, we'll focus on the variables `bwt` and `lwt`.

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.

```r
library(ggplot2)

# Scatter plot for lwt vs bwt
ggplot(birthwt, aes(x = lwt, y = bwt)) +
  geom_point() +
  labs(title = "Relationship between Mother's Pre-pregnancy Weight and Baby's Weight",
       x = "Mother's Pre-pregnancy Weight (lwt)",
       y = "Baby's Weight (bwt)") +
  theme_minimal()
```



Answer: The scatter plot shows a mild positive trend between mother's pre-pregnancy weight (`lwt`) and baby's weight (`bwt`). A correlation of 0.186(from Exercise 5c) indicates a weak positive relationship. Most data points fall between 100-150 pounds for `lwt` and 2000-4000 grams for `bwt`. This suggests babies of mothers with higher pre-pregnancy weights tend to have slightly higher birth weights, but the relationship is not very strong. Correlation doesn't imply causation, and other factors could influence birthweight.

part b

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

```r
# Fitting the linear model
model <- lm(bwt ~ lwt, data = birthwt)

# Displaying the model summary to retrieve coefficients
summary(model)
```

```
##
```

```
## 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:

Based on the provided output for the linear model:

For the relationship between `bwt` (baby's weight) and `lwt` (mother's pre-pregnancy weight):

- The intercept (β_0) is 2369.624. This means that the predicted baby's weight when the mother's pre-pregnancy weight is zero (though not a practical scenario) is approximately 2369.624 grams.
- The slope (β_1) is 4.429. This indicates that for every additional pound increase in the mother's pre-pregnancy weight, the baby's weight is predicted to increase by approximately 4.429 grams.

The linear model predicting baby's weight (`bwt`) from mother's pre-pregnancy weight (`lwt`) can be written as:

$$\text{bwt} = 2369.624 + 4.429 \times \text{lwt}$$

part c

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

Answer:

Intercept (2369.624): This represents the predicted baby's weight in grams when the mother's pre-pregnancy weight is zero. While this is a theoretical value, it essentially sets a baseline for the linear model.

Slope (4.429): For every one-pound increase in the mother's pre-pregnancy weight (`lwt`), the baby's birthweight (`bwt`) is expected to increase by approximately 4.429 grams, holding all else constant.

part d

Is the intercept meaningful? Briefly explain.

Answer: The intercept, in this context, is not practically meaningful. It suggests the predicted weight of a baby when the mother's pre-pregnancy weight is zero pounds, which is not a realistic scenario. While the intercept provides a mathematical foundation for the linear model, it doesn't offer any real-world interpretation in this specific situation.

part e

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.

```
# Given coefficients and lwt value
intercept <- 2369.624
slope <- 4.429
lwt_value <- 147

# Calculate predicted bwt
predicted_bwt <- intercept + slope * lwt_value

# Calculate the residual
observed_bwt <- 3000
residual <- observed_bwt - predicted_bwt

predicted_bwt
```

```
## [1] 3020.687
```

```
residual
```

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

Answer:

To calculate the estimated mean baby's birthweight for a mother with a pre-pregnancy weight of 147 pounds, you can use the linear model equation:

$$\text{bwt} = 2369.624 + 4.429 \times \text{lwt}$$

Substitute lwt with 147 to get the estimated birthweight.

For the residual:

Residual = Observed value - Predicted value

In this case: Observed value = 3000 g (given) Predicted value = The result from the above calculation

1. **Predicted Birthweight:**

$$\text{bwt} = 2369.624 + (4.429 \times 147) = 3020.687$$

2. **Residual:**

$$\text{Residual} = 3000 - \text{Predicted bwt} = -20.687$$

part f

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?

```
observed_bwt = 2743
residual_value = -40
# Calculate lwt using the relationship
lwt_original <- (observed_bwt - intercept - residual_value) / slope
lwt_original
```

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

Answer:

To find the mother's weight given the baby's weight and the residual, we can use the relationship:

$$\text{Residual} = \text{Observed bwt} - \text{Predicted bwt}$$

From the linear model:

$$\text{Predicted bwt} = \text{intercept} + \text{slope} \times \text{lwt}$$

Given:

$$\text{Observed bwt} = 2743 \text{ g}$$

$$\text{Residual} = -40 \text{ g}$$

Substitute in the observed bwt and residual to find the predicted bwt, and then rearrange the equation to solve for lwt:

$$\text{lwt} = \frac{\text{Observed bwt} - \text{intercept} - \text{Residual}}{\text{slope}}$$

Now, we can calculate lwt using the given values and the coefficients from the linear model.

$$\text{lwt} = \frac{2743 - 2369.624 - (-40)}{4.429} = 93.33394$$