Problem Set 1

Intro to `R`

[YOUR NAME]

Due Date: 2025-06-06 by 11:59PM

Getting Set Up

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...`. Accept defaults and save this file as `[LAST NAME]_ps1.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps1.Rmd` file. Then change the `author: [Your Name]` to your name.

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in.

This problem set is worth 22 total points, plus 2.5 extra credit points. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments. To submit, compile the completed problem set and upload the PDF file to Drobox on Friday by midnight. If you use AI for help, choose to save your output as a PDF and submit this with the problem set as well. Also note that I will not respond to Campuswire messages after 2PM ET on Friday, so don't wait until the last minute to get started!

Good luck!

If you collaborated with a colleague and/or used AI for any help on this problem set, document here. Write the names of your classmates and/or upload a PDF of your AI prompt and output with your problem set:

Part 1: All about college

[10 points possible; 0.5 extra credit points]

Question 0 [0 points]

Require `tidyverse` and load the `sc_debt.Rds` data by assigning it to an object named `df`.

```r
require() # Load tidyverse
```

```
## Loading required package:
```

```r
df <- read_rds() # Load the dataset directly from github
```

```
## Error in read_rds(): could not find function "read_rds"
```

Question 1 [1 point]

Which school has the lowest admission rate (`adm_rate`) and which state is it in (`stabbr`)?

```r
df %>%
  arrange() %>% # Arrange by the admission rate
  select() # Select the school name, the admission rate, and the state
```

```
## Error in df %>% arrange() %>% select(): could not find function "%>%"
```

> Write answer here

Question 2 [1 point]

Which are the top 10 schools by average SAT score (`sat_avg`)?

```r
df %>%
  arrange() %>% # arrange by SAT scores in descending order
  select() %>% # Select the school name and SAT score
  print() # Print the first 12 rows (hint: there is a tie)
```

```
## Error in df %>% arrange() %>% select() %>% print(): could not find function "%>%"
```

> Write answer here

Question 3 [1 point]

Create a new variable called `adm_rate_pct` which is the admissions rate multiplied by 100 to convert from a 0-to-1 decimal to a 0-to-100 percentage point.

```r
df <- df %>% # Use the object assignment operator to overwrite the df object
  mutate() # Create the new variable adm_rate_pct
```

```
## Error in df %>% mutate(): could not find function "%>%"
```

Question 4 [1 point]

Calculate the average SAT score and median earnings of recent graduates by state.

```r
df %>%
  group_by() %>% # Calculate state-by-state with group_by()
  summarise(sat_avg = , # Summarise the average SAT
            earn_avg = ) # Summarise the average earnings
```

```
## Error in df %>% group_by() %>% summarise(sat_avg = , earn_avg = ): could not find function "%>%"
```

Extra Credit [0.5 points]

Plot the average SAT score (x-axis) against the median earnings of recent graduates (y-axis) by school, and add the line of best fit. What relationship do you observe? Why do you think this relationship exists?

```
# INSERT CODE HERE
```

> Write answer here

Question 5 [3 points]

Research Question: Do students who graduate from smaller schools (i.e., schools with smaller student bodies) make more money in their future careers? Before looking at the data, write out what you think the answer is, and explain why you think so.

> Write a few sentences here.

Based on this research question, what is the outcome / dependent / Y variable and what is the explanatory / independent / X variable? Create the scatterplot of the data based on this answer, along with a line of best fit. Is your answer to the research question supported?

```
df %>%
  ggplot(aes(x = , # Put the explanatory variable on the x-axis
             y = )) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth() + # Add line of best fit
  labs(title = '', # give the plot meaningful labels to help the viewer understand it
       x = '',
       y = '')
```

```
## Error in df %>% ggplot(aes(x = , y = )): could not find function "%>%"
```

> Write a few sentences here.

Question 6 [2 points]

Does this relationship change by whether the school is a research university? Using the filter() function, create two versions of the plot, one for research universities and the other for non-research universities.

```
df %>%
  filter() %>% # Filter to non-research universities
  ggplot(aes(x = , # Put the explanatory variable on the x-axis
             y = )) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth() + # Add line of best fit
  labs(title = '', # give the plot meaningful labels to help the viewer understand it
       subtitle = '',
       x = '',
       y = '')
```

```
## Error in df %>% filter() %>% ggplot(aes(x = , y = )): could not find function "%>%"
```

```
df %>%
  filter() %>% # Filter to research universities
  ggplot(aes(x = , # Put the explanatory variable on the x-axis
             y = )) +  # Put the outcome variable on the y-axis
  geom_point() + # Create a scatterplot
  geom_smooth() + # Add line of best fit
  labs(title = '', # give the plot meaningful labels to help the viewer understand it
       subtitle = '',
       x = '',
       y = '')
```

```
## Error in df %>% filter() %>% ggplot(aes(x = , y = )): could not find function "%>%"
```

Question 7 [1 point]

Instead of creating two separate plots, color the points by whether the school is a research university. To do this, you first need to modify the research_u variable to be categorical (it is currently stored as numeric). To do this, use the mutate command with `ifelse()` to create a new variable called `research_u_cat` which is either "Research" if `research_u` is equal to 1, and "Non-Research" otherwise.

```
df <- df %>%
  mutate(research_u_cat = ifelse()) # Create a labeled version of the research_u variable
```

```
## Error in df %>% mutate(research_u_cat = ifelse()): could not find function "%>%"
```

```
df %>%
  ggplot(aes(x = , # Put the explanatory variable on the x-axis
             y = , # Put the outcome variable on the y-axis
             color = )) + # Color the points by the new variable you created above
  geom_point() + # Create a scatterplot
  geom_smooth() + # Add line of best fit
  labs(title = '', # give the plot meaningful labels to help the viewer understand it
       x = '',
       color = '',
       y = '')
```

```
## Error in df %>% ggplot(aes(x = , y = , color = )): could not find function "%>%"
```

Part 2: Learning about the 2020 elections from Michigan exit polling

[6 points; +1 extra credit points available]

For part 2 of this problem set, we will be using the `MI2020_ExitPoll.Rds` file from the course github page.

Question 8 [1 point]

Require an additional package called `labelled` (remember to `install.packages("labelled")` if you don't have it yet) and load the `MI2020_ExitPoll.Rds` data to an object called `MI_raw`. (Tip: use the `read_rds()` function with the link to the raw data.)

```
require()
```

```
## Loading required package:
```

```
MI_raw <- read_rds('')
```

```
## Error in read_rds(""): could not find function "read_rds"
```

What is the unit of analysis in this dataset? How many variables does it have? How many observations?

> Write answer here

Question 9 [1 point]

This has too much information that we don't care about. Create a new object called `MI_clean` that contains only the following variables:

- AGE10
- SEX
- PARTYID
- EDUC18
- PRSMI20
- QLT20
- LGBT
- BRNAGAIN
- LATINOS
- QRACEAI
- WEIGHT

and then list which of these variables contain missing data recorded as `NA`. How many respondents were not asked certain questions?

```
MI_clean <- MI_raw %>%
  select() # Select the requested variables
```

```
## Error in MI_raw %>% select(): could not find function "%>%"
```

```
summary() # Identify which have missing data recorded as NA
```

```
## Error in summary.default(): argument "object" is missing, with no default
```

> Write answer here

Question 10 [1 point]

Are there **unit non-response** *data in the PRSMI20 variable? If so, how are they recorded? What about the PARTYID variable? How many people refused to answer both of these questions?*

```
MI_clean %>%
  count() # Tip: use count() function to look at your variables.
```

```
## Error in MI_clean %>% count(): could not find function "%>%"
```

> Write answer here.

Question 11 [1 points]

Let's create a new variable called `preschoice` that converts PRSMI20 to a character. To do this, install the `labelled` package if you haven't already, then use the `to_character()` function from the `labelled` package. Now `count()` the number of respondents who reported voting for each candidate. How many respondents voted for candidate Trump in 2020? How many respondents refused to tell us who they voted for?

```
MI_clean <- MI_clean %>%
  mutate(preschoice = ) # Convert to character
```

```
## Error in MI_clean %>% mutate(preschoice = ): could not find function "%>%"
```

```
MI_clean %>%
  count()
```

```
## Error in MI_clean %>% count(): could not find function "%>%"
```

> Write answer here

Question 12 [2 points]

What proportion of women supported Trump?

```
# Women Trump supporters
MI_clean %>%
  drop_na() %>% # Drop any missing values for preschoice
  filter() %>% # Filter to only women
  count() %>% # Count the number of women who supported each candidate
  mutate(share = ) # Calculate the proportion of women who supported Trump
```

```
## Error in MI_clean %>% drop_na() %>% filter() %>% count() %>% mutate(share = ): could not find functi
```

```
# Alternative approach
MI_clean %>%
  drop_na() %>% # Drop any missing values for preschoice
  mutate(trumpSupp = ifelse()) %>% # Create "dummy" variable for whether the person voted for Trump or
  group_by() %>% # Group by gender
  summarise(share = mean(trumpSupp)) # Calculate proportion who supported Trump
```

```
## Error in MI_clean %>% drop_na() %>% mutate(trumpSupp = ifelse()) %>% group_by() %>% : could not find
```

> Write answer here.

Extra Credit [1 point]

Among women, which age group sees the highest support for Trump? To answer, you will need to calculate the proportion of women who supported Trump by age-group to determine which age-group had the highest Trump support among women. You will need to clean the AGE10 variable before completing this problem, just like we did with the PRSMI20 variable. Call the new variable "Age". HINT: to make your life easier (and not write a 10-level nested ifelse() function), try asking ChatGPT for help with this prompt: "I have a labelled variable in R that I want to convert to text. How can I do this?"

```
# Insert code here.
```

Write answer here

Part 3: NBA Jam, "Boom-shakalaka!"

[6 points; +1 extra credit point available]

Question 13 [1 point]

Plot the distribution of field goals attempted by all NBA players in the 2018-2019 season. Explain why you chose the visualization that you did. Then add a vertical line indicating the mean and median number of points in the data. Color the median line blue and the mean line red. Why is the median lower than the mean?

```
nba %>%
  ggplot() + # Put the fga variable on the x-axis of a ggplot.
   geom_...() + # Choose the appropriate geom function to visualize.
  labs() + # Add labels
      geom_vline() + # Median vertical line (blue)
      geom_vline() # Mean vertical line (red)
```

```
## Error in nba %>% ggplot(): could not find function "%>%"
```

Write answer here.

Question 14 [1 point]

Now examine the `country` *variable. Visualize this variable using the appropriate* `geom_...`*, and justify your reason for choosing it. Tweak the plot to put the country labels on the y-axis, ordered by frequency. Which country are most NBA players from? What is weird about your answer, and what might explain it?*

```
nba %>%
  count() %>% # count the number of players by country
  ggplot() + # place the country on the y-axis, reordered by the number of players. Put the number of p
  geom_...() + # Choose the best geom
  labs() # Add labels
```

```
## Error in nba %>% count() %>% ggplot(): could not find function "%>%"
```

Write answer here

Question 15 [3 points]

Let's pretend we are consulting for an NBA organization. The owner and GM tell us they are interested in the relationship between the player's age (`agePlayer`) and the amount of points they score (`pts`). Please answer the following research question and provide a theory supporting your answer: "Do older NBA players score more points than younger players?"

> Write answer here

Based on your answer above, what is the outcome / dependent / Y variable and what is the explanatory / independent / X variable? Why?

> Write answer here

Create a univariate visualization of both the X and Y variables. Choose the best `geom_...()` based on the variable type, and make sure to label your plots!

```r
# X variable
nba %>%
  ggplot() + # Put the X variable on the x-axis
  geom_...() +  # Choose the best geom given the variable type (make sure to look at it if you aren't s
  labs()     # Add labels
```

```
## Error in nba %>% ggplot(): could not find function "%>%"
```

```r
# Y variable
nba %>%
  ggplot(...) + # Put the Y variable on the x-axis
  geom_...() +  # Choose the best geom given the variable type (make sure to look at it if you aren't s
  labs(...)     # Add labels
```

```
## Error in nba %>% ggplot(...): could not find function "%>%"
```

Question 16 [1 point]

Now analyze the data by creating a multivariate visualization that shows the relationship between age and points. Add a STRAIGHT line of best fit with `geom_smooth()`.

```r
nba %>%
  ggplot() + # Put the X variable on the x-axis, and the Y variable on the y-axis
  geom_...() +  # Choose the best geom given both variable types
  geom_smooth() + # Add a STRAIGHT line of best fit
  labs()     # Add labels
```

```
## Error in nba %>% ggplot(): could not find function "%>%"
```

Based on your analysis, does the data support or reject your hypothesis from Question 3?

> Write answer here

Extra Credit [1 point]

Let's look for evidence of a "curvelinear" relationship between player age and points scored. To do so, first calculate the average points scored by age. Then plot this relationship using a multivariate visualization. Add a line of best fit with `geom_smooth()` *but DON'T use* `method = "lm"`*. What do you conclude? Why?*

```
# INSERT CODE HERE
```

Write answer here