Problem Set 1

Name:

Due: Tues, Sept 22

Rubric

Each question is graded on effort, not correctness. Responses should demonstrate engagement with concepts, code, and examples from the previous class. Vague, generic, copy-pasted responses, or responses that do not attempt to fully address the question will receive reduced credit. You may work with your classmates on this assignment, but you should each submit your own work. Please submit your response as a .pdf file to canvas by the due date posted on canvas. **No late work will be accepted.**

However, each question will be assigned a hypothetical point value indicating the accuracy of your response, which I hope you find to be useful feedback as you prepare for exams in this course.

Points	Criteria
5	The solution is correct and well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect.
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification.
0	Either the solution is missing entirely, or the author makes no nontrivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information).

You may need to install some packages before you are able to render the document. **Please pay attention to formatting throughout! You should look at your rendered .pdf file and ensure that the formatting looks good before submitting.**

> **Note**
>
> Trying running `help("install.packages")` to learn more about package installation.

Practice with R and RStudio

1. Run the following code so that I know what version of R you are running. You will need the `tidyverse` package installed to run the following code chunk.

```
sessionInfo()
```

```
R version 4.6.0 (2026-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26200)

Matrix products: default
  LAPACK version 3.12.1

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

time zone: America/New_York
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] lubridate_1.9.5 forcats_1.0.1   stringr_1.6.0   dplyr_1.2.1
 [5] purrr_1.2.2     readr_2.2.0     tidyr_1.3.2     tibble_3.3.1
 [9] ggplot2_4.0.3   tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] gtable_0.3.6       jsonlite_2.0.0     compiler_4.6.0     tidyselect_1.2.1
 [5] scales_1.4.0       yaml_2.3.12        fastmap_1.2.0      R6_2.6.1
 [9] generics_0.1.4     knitr_1.51         pillar_1.11.1      RColorBrewer_1.1-3
```

```
[13] tzdb_0.5.0         rlang_1.2.0        stringi_1.8.7      xfun_0.59
[17] S7_0.2.2           otel_0.2.0         timechange_0.4.0   cli_3.6.6
[21] withr_3.0.3        magrittr_2.0.5     digest_0.6.39      grid_4.6.0
[25] rstudioapi_0.19.0  hms_1.1.4          lifecycle_1.0.5    vctrs_0.7.3
[29] evaluate_1.0.5     glue_1.8.1         farver_2.1.2       rmarkdown_2.31
[33] tools_4.6.0        pkgconfig_2.0.3    htmltools_0.5.9
```

A key learning objective of this course is to develop the ability to communicate statistics, both verbally and in writing. To do so, you must be able to properly format a written report. RStudio makes formatting reports fairly easy, but it will likely take some practice. You will not learn everything that R is capable of in this activity. Instead, this activity will help you get your feet wet with R, and you will learn more about R by exposure throughout the semester. R is an **extremely** marketable skill with employers, so I invite you to further explore R on your own or with me during office hours this semester. For all the following questions, use the Markdown basics link for help. Or, ask ChatGPT how to accomplish these tasks in Markdown.

2. Type your name in boldface.

3. Type your major (or intended major) in italics - it is okay to say you do not know.

4. Make a list of a few of your interests.

   - list 1

5. Include a hyperlink to something interesting or funny.

   google

6. Include a picture that you find interesting.

`R` is an object based coding language, meaning we first create objects, then apply functions to those objects to conduct analyses. Above, we created an object called `toads` that contains data on toad breeding in the Tibetan Plateau.

7. Type `toads` into the console and hit return. What type of object is toads?

```
toads <- read_csv("data-JEB.csv")
str(toads)
class(toads)
```

8. There are many types of objects in R. We will work extensively with data frames (`tibbles`) and vectors. Vectors may be created by **concatenating** a comma-separated list of values, or by **extracting** columns from a data set.

```
# build a vector from scratch
buildavector <- c(42, 7, 69, 8.3)
str(buildavector)
```

```
 num [1:4] 42 7 69 8.3
```

```
class(buildavector)
```

```
[1] "numeric"
```

```
# vectors can also be made of character data
rickroll <- c("never", "gonnna", "give", "you", "up", "never", "gonna")
str(rickroll)
```

```
 chr [1:7] "never" "gonnna" "give" "you" "up" "never" "gonna"
```

```
class(rickroll)
```

```
[1] "character"
```

```
# extract a vector from a data frame
alt <- toads$altitude
str(alt)
```

```
 num [1:431] 3462 3462 3462 3462 3462 ...
```

```
class(alt)
```

```
[1] "numeric"
```

9. Once we have created an object, we can use functions to manipulate the object and create useful summaries.

```
# compute the mean of alt: type ?mean in the console

# compute the standard deviation from alt: type ?sd in the console

# subtract the mean from alt

# divide the difference between alt and mean(alt) by the standard deviation of alt
```

Practice with statistics

10. The following code imports the `barre_weather` data and prints the first six rows. This data set, provided by the state of Vermont, contains daily measurements from the weather station at the Knapp State Airport west of Barre, VT.

> **!** Important
>
> In order to run the code below that imports the weather data, you must ensure that the data set is saved in the same folder as this .qmd file, and that your current working directory is set to the source file location. This can easily be done in RStudio by selecting the following options: **Session → Set working directory → To source file location**.
> Additionally, you will need to install the `tidyverse` package by running `install.packages("tidyverse")` in the console.

```
library(tidyverse)
barre_weather <- readr::read_csv("barre_weather.csv")
head(barre_weather)
```

```
# A tibble: 6 x 11
  Date       `Percipitation in.` `Snow Depth in.` `Snowfall in.` `Max Temp F`
  <chr>                    <dbl>            <dbl>          <dbl>        <dbl>
1 06/01/1948                0                   0              0         80.1
2 06/02/1948                0                   0              0         79.0
3 06/03/1948                0                   0              0         82.0
4 06/04/1948                0                   0              0         84.0
5 06/05/1948                0.15                0              0         62.1
6 06/06/1948                0                   0              0         69.1
# i 6 more variables: `Min Temp F` <dbl>, `Ave Wind Spd mph` <dbl>, FOG <lgl>,
#   Sleet <lgl>, `Smoke/Haze` <lgl>, Thunder <lgl>
```

a) What are the observational units for this data set?

b) Determine whether each of the following variables from the `barre_weather` data set are numeric or categorical. If a variable is categorical, determine whether it is nominal or ordinal.

- Percipitation in.: type here
- Snow Depth in.:
- Snowfall in.:
- Max Temp F:
- Ave Wind Spd mph:
- FOG:
- Sleet:
- Smoke/Haze:

c) Describe how one could convert the maximum temperature variable to an ordinal, categorical variable?

d) Suppose researchers were interested in determining whether the average precipitation differed between days with fog and days without fog. What would the response variable be? What about the explanatory variable?

11. Researchers collected data to examine the relationship between air pollutants and preterm births in Southern California. During the study air pollution levels were measured by air quality monitoring stations. Specifically, levels of carbon monoxide were recorded in parts per million, nitrogen dioxide and ozone in parts per hundred million, and coarse particulate matter (PM_{10}) in $\mu g/m^3$. Length of gestation data were collected on 143,196 births between the years 1989 and 1993, and air pollution exposure during gestation was calculated for each birth. The analysis suggested that increased ambient PM_{10} and, to a lesser degree, CO concentrations may be associated with the occurrence of preterm births.

a) Is this an observational study or an experiment? Explain your reasoning.

b) What are the cases in this study and how many are there?

c) What is the main research question of the study?

d) What are the variables in this study? Identify each variable as numerical or categorical.

e) Are we able to reasonably conclude that exposure to increased ambient PM_{10} caused increased preterm births? Why or why not?

12. Researchers investigating the effects of gamification (application of game-design elements and game principles in non-game contexts) on learning statistics randomly assigned 365 college students in a statistics course to one of four groups; one of these groups had no reading exercises and no gamification, one group had reading but no gamification, one group had gamification but no reading, and a final group had gamification and reading. Students in all groups also attended lectures. The study found that gamification had a positive impact on student learning compared to traditional teaching methods involving reading exercises.

a) Is this an observational study or an experiment? Explain your reasoning.

b) What are the cases in this study and how many are there?

c) What is the response variable in the study?

d) What is the explanatory variable in the study and what type of variable is it?

e) Are we able to reasonably conclude that gamification caused an increase in student learning? Why or why not?

13. Select a comparison from the spurious correlation website, download the the figure and include it in this document. Comment on what might explain the observed relationship - it is okay to be outlandish!