Name: Your name here
Due: 2024/09/16

Homework 1

Be sure to submit **both** the .pdf and .qmd file to Canvas by Monday, September 16th at 11:59 pm. Additionally, when possible, your answers should be denoted with a callout box. For example, if the question is, "What is the meaning of life?"

> 42

1. [1 pt] Run the following code so that I know what version of R you have installed.

   ```
   sessionInfo()
   ```

   ```
   R version 4.3.0 (2023-04-21 ucrt)
   Platform: x86_64-w64-mingw32/x64 (64-bit)
   Running under: Windows 10 x64 (build 19045)

   Matrix products: default


   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

   loaded via a namespace (and not attached):
    [1] compiler_4.3.0  fastmap_1.1.1   cli_3.6.1       tools_4.3.0
    [5] htmltools_0.5.5 rstudioapi_0.14 yaml_2.3.7      rmarkdown_2.25
    [9] knitr_1.42      jsonlite_1.8.4  xfun_0.39       digest_0.6.31
   [13] rlang_1.1.1     evaluate_0.21
   ```

2. [10 pt] For this assignment, we will use a data set describing weather station measurements at the Knapp State Airport, west of Barre, VT. This data set is publicly available and provided by the

state of Vermont. The purpose of this question is to import and manicure the data, so that we have a clean time series for analysis later in the assignment.

a) [1 pt] Import the Barre weather station data and print the first six rows.

```
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>
```

b) [1 pt] Create a new data frame, called `barre_clean`, that includes a `year`, `month`, and `day` column.

```
barre_clean <- barre_weather |>
  mutate(
    date = mdy(Date),
    year = year(date),
    month = month(date),
    day = day(date)
  ) |>
  dplyr::select(-Date)
```

c) [2 pt] It is always advised to inspect the data before manipulating it. Create a table of the number of observed days by year. Filter to years with fewer than 365 days and greater than 366 days and print the resulting data frame. Do you notice anything strange?

```
barre_clean |>
  group_by(year) |>
  summarize(n_days = n()) |>
  filter(n_days <= 364 | n_days >= 367)
```

```
# A tibble: 5 x 2
   year n_days
  <dbl>  <int>
```

```
1  1948    214
2  1955    363
3  1996    351
4  1997    363
5  2015    171
```

> There are a handful of years ('55, '96, and '97) for which we observe fewer than 365 days. This is also true of 1948 and 2015, but that is because the survey period did not start on January 1st or end on December 31st.

d) [2 pt] Create a new data frame, called `barre_monthly`, that summarizes the total sum of snowfall in each month (you may ignore any oddities you found in the previous question, for now). The resulting data frame should have a `sum_snowfall` column, a `year` column, and a `month` column. Arrange the data frame by year and month. Print the first and last six rows of `barre_monthly`.

```r
barre_monthly <- barre_weather |>
  dplyr::select(
    Date,
    snowfall = `Snowfall in.`
  ) |>
  mutate(
    date = lubridate::parse_date_time(Date, "mdy"),
    year = lubridate::year(date),
    month = lubridate::month(date)
  ) |>
  group_by(year, month) |>
  summarize(sum_snowfall = sum(snowfall)) %>%
  ungroup %>%
  arrange(year, month)

head(barre_monthly)
```

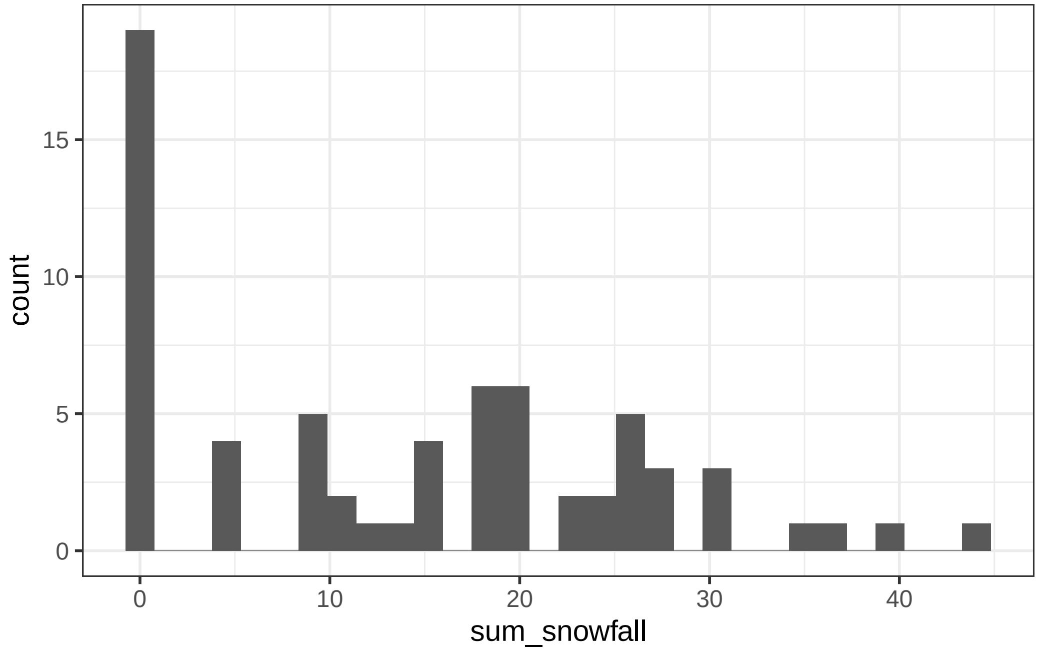
```
# A tibble: 6 x 3
   year month sum_snowfall
  <dbl> <dbl>        <dbl>
1  1948     6         0
2  1948     7         0
3  1948     8         0
4  1948     9         0
5  1948    10         2.01
6  1948    11         0.39
```

```r
tail(barre_monthly)
```

```
# A tibble: 6 x 3
   year month sum_snowfall
  <dbl> <dbl>        <dbl>
1  2015     1            0
2  2015     2            0
3  2015     3            0
4  2015     4            0
5  2015     5            0
6  2015     6            0
```

e) [1 pt] Create a histogram of the `sum_snowfall` in February. Does anything strike you as odd?

```
barre_monthly |>
  filter(month == 2) |>
  ggplot(aes(x = sum_snowfall)) +
  geom_histogram() +
  theme_bw()
```



Awful lot of 0's for sum total snowfall in February!

f) [1 pt] What years contain 0 sum total snowfall in February? Does anything strike you as odd?

```r
barre_monthly |>
  filter(month == 2, sum_snowfall == 0) |>
  dplyr::select(year) |> unlist() |> unname()
```

```
 [1] 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011
[16] 2012 2013 2014 2015
```

Strangely, these data suggest that every February between 1997 and 2015 saw 0 total inches of snowfall.

g) [2 pt] Create a `ts` object containing the sum total monthly snowfall between 1950 and 1996 (hopefully you are convinced that we should not include `1997:2015`), called `barre_ts`. What is the frequency of this time series?

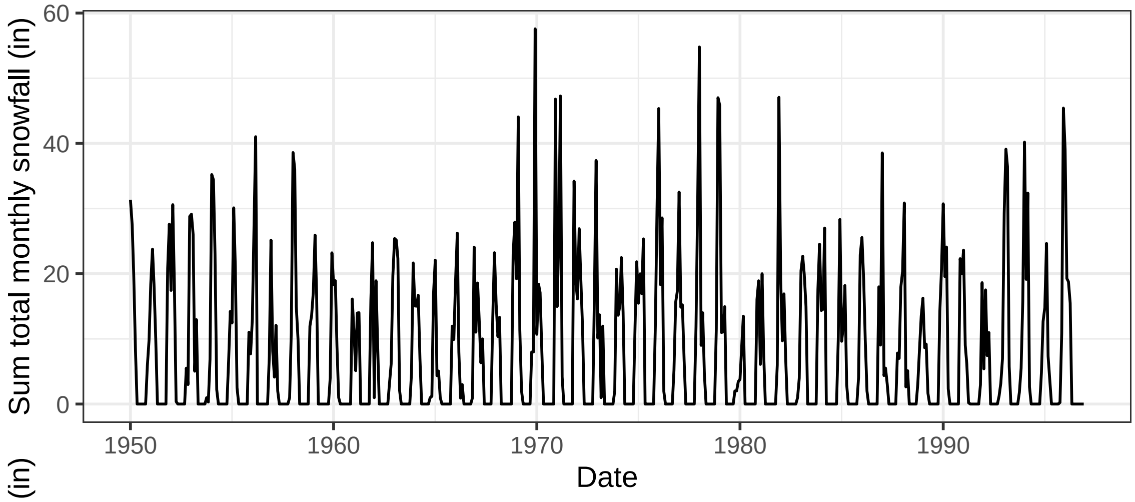
```r
barre_monthly_pre97 <- barre_monthly |> filter(year <= 1996 & year >= 1950)
barre_ts <- ts(
  data =  barre_monthly_pre97$sum_snowfall,
  start = c(1950, 1),
  end = c(1996, 12),
  freq = 12
)
```

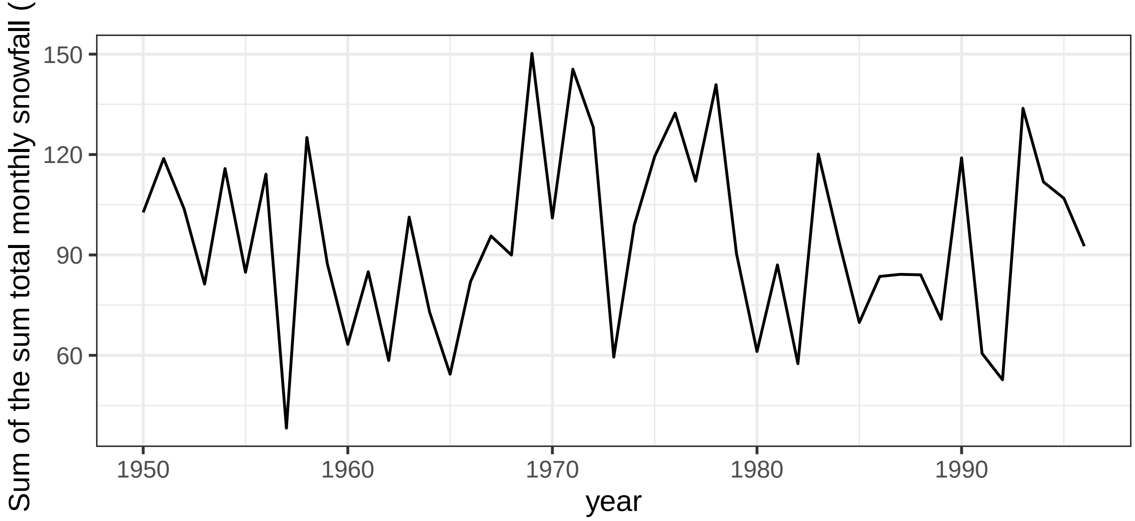
Since we have averaged across months, the frequency of this time series is 12.

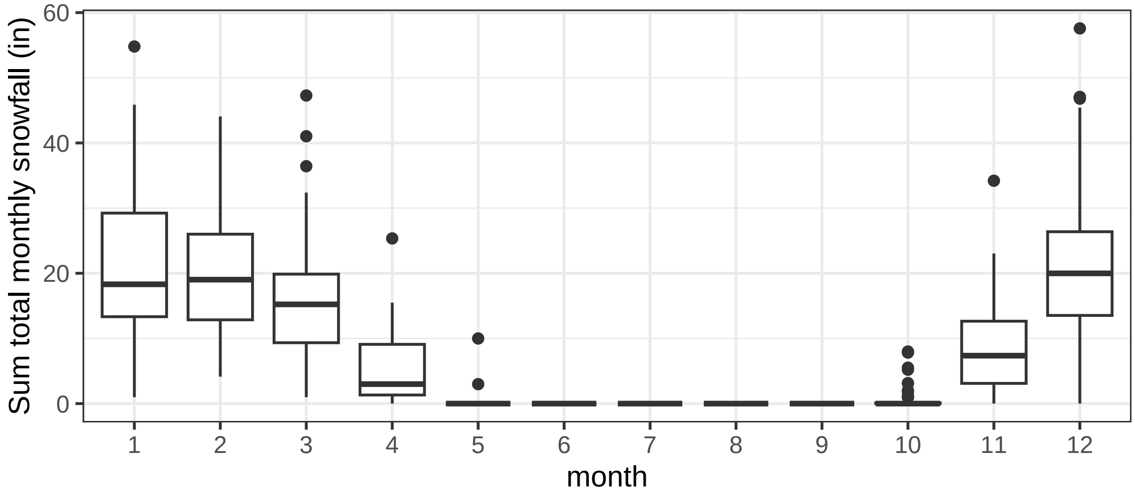
3. [5 pt] You should now have a clean `ts` object, called `barre_ts`. Using this object, describe the sum total monthly snowfall (in) at the Knapp State Airport weather station. Be sure to reference any apparent trends or seasonality in your response, including figures that support your statements.

The sum total snowfall (in) at the Knapp State Airport weather station shows no apparent trend between 1950 and 1996 (based on the middle figure below), but does display very strong seasonal effects (based on the bottom figure below). This is (obviously) not shocking, as we expect to see greater snowfall in the winter months than the summer months.

```r
p1 <- barre_monthly_pre97 |>
  mutate(date = ymd(paste0(year, "-", month, "-", 1))) |>
  ggplot(aes(x = date, y = sum_snowfall)) +
  geom_line() +
  theme_bw() +
  labs(x = "Date", y = "Sum total monthly snowfall (in)")

p2 <- barre_monthly_pre97 |>
  group_by(year) |>
  summarize(total_sum_snowfall = sum(sum_snowfall)) |>
  ggplot(aes(x = year, y = total_sum_snowfall)) +
  geom_line() +
  theme_bw() +
  labs(y = "Sum of the sum total monthly snowfall (in)")

p3 <- barre_monthly_pre97 |>
  mutate(month = factor(month)) |>
  ggplot() +
  geom_boxplot(aes(x = month, y = sum_snowfall)) +
  theme_bw() +
  labs(y = "Sum total monthly snowfall (in)")

gridExtra::grid.arrange(p1, p2, p3, nrow = 3)
```


Sum total monthly snowfall (in)
60
40
20
0
1950
1960
1970
1980
1990
Date


Sum of the sum total monthly snowfall (in)
150
120
90
60
1950
1960
1970
1980
1990
year


Sum total monthly snowfall (in)
60
40
20
0
1
2
3
4
5
6
7
8
9
10
11
12
month