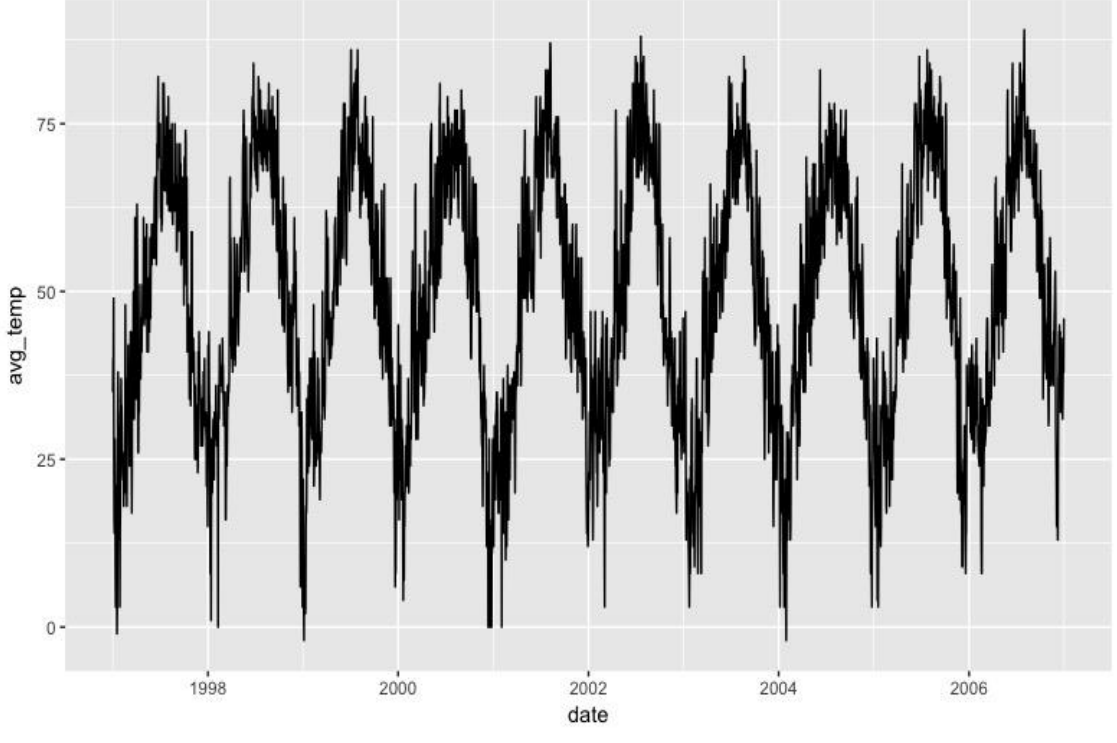
1. Load the data using the read_csv() function. Look at the resulting tibble using the head() function. How many columns are there? What are they and what are their types? Use the nrow() function to find the number of rows.

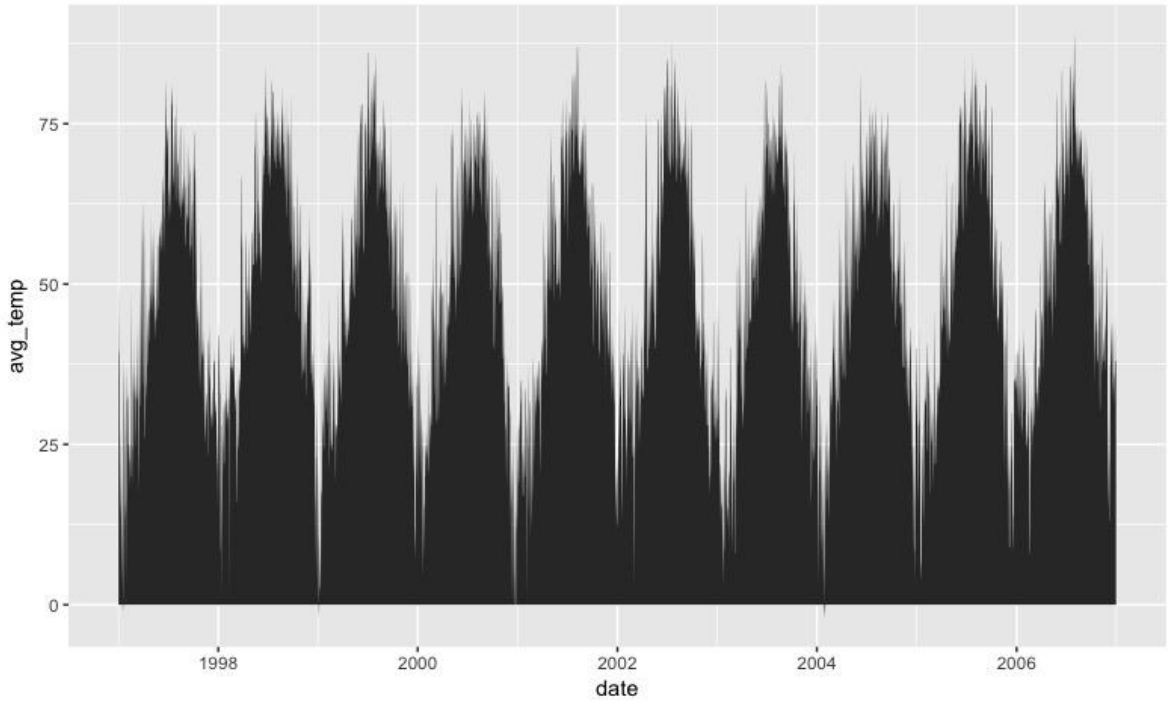
- A tibble: 6 x 5. I am using head(mkw) command.
- There are 5 columns. I am using ncol(mkw) command.
- There are date and its data type <date>, avg_temp and its data type <dbl>, snowfall and its data type <dbl>, rainfall and its data type <dbl> and avg_wind_speed and its data type <dbl>.
- There are 3640 rows. I am using nrow(mkw) command.

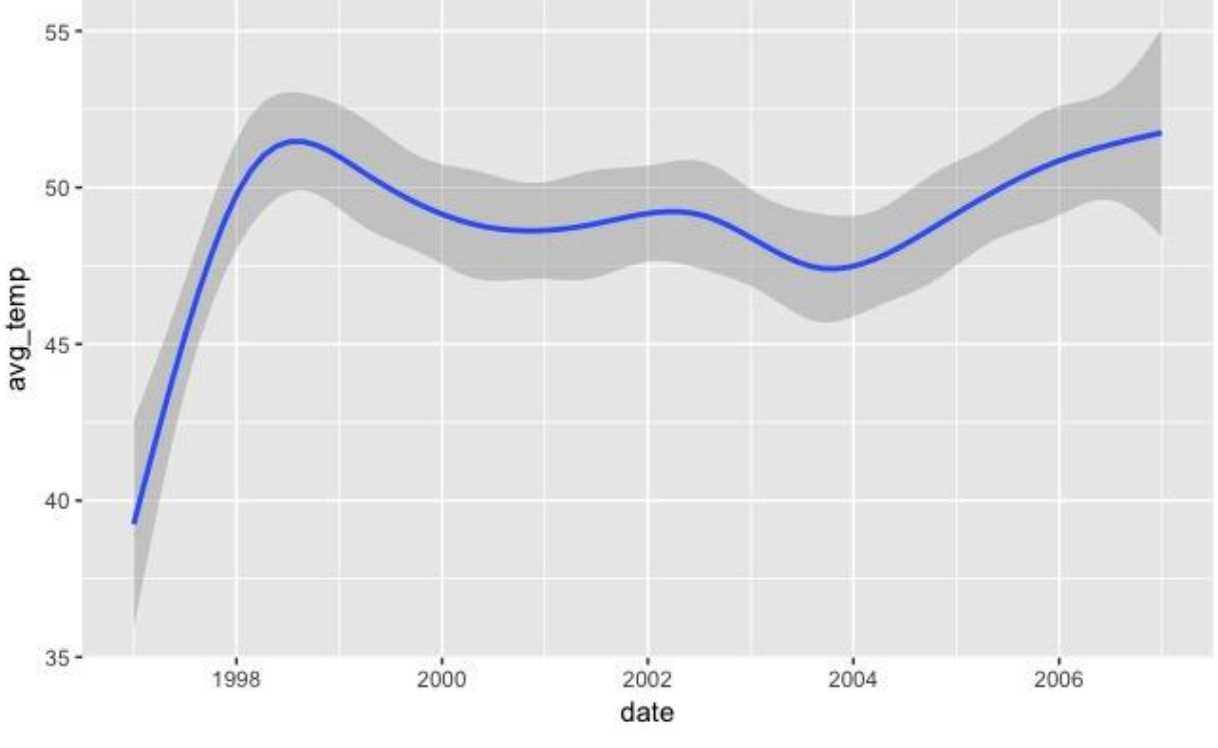
```
> head(mw)
# A tibble: 6 x 5
  date       avg_temp snowfall rainfall avg_wind_speed
  <date>        <dbl>    <dbl>    <dbl>          <dbl>
1 1997-01-01     35.0        0   0.0200           7.10
2 1997-01-02     40.0        0   0                5.50
3 1997-01-03     39.0        0   0                6.00
4 1997-01-04     49.0        0   0.420            0.600
5 1997-01-05     27.0        0   0               18.6
6 1997-01-06     16.0        0   0               13.9
```

2. Plot the avg_temp over time. What two columns do you need? What type of plot is appropriate? Do you see a repeating pattern?

- For this plot I need Date and avg_temp columns.
- I think geom_line, geom_smpoth and area plot is appropriate but geom_line is more appropriate. Because in line plot we can easily see repeating pattern.
- I use ggplot(data =mkw) + geom_line(mapping = aes(x = date, y = avg_temp)) for line. And I use ggplot(data = mkw) + geom_smooth(mapping = aes(x = date, y = avg_temp)) for smooth.
- Also, I use ggplot(data = mkw) + geom_area(mapping = aes(x = date, y = avg_temp)) for area.
- Yes, there are repeating pattern in my output.

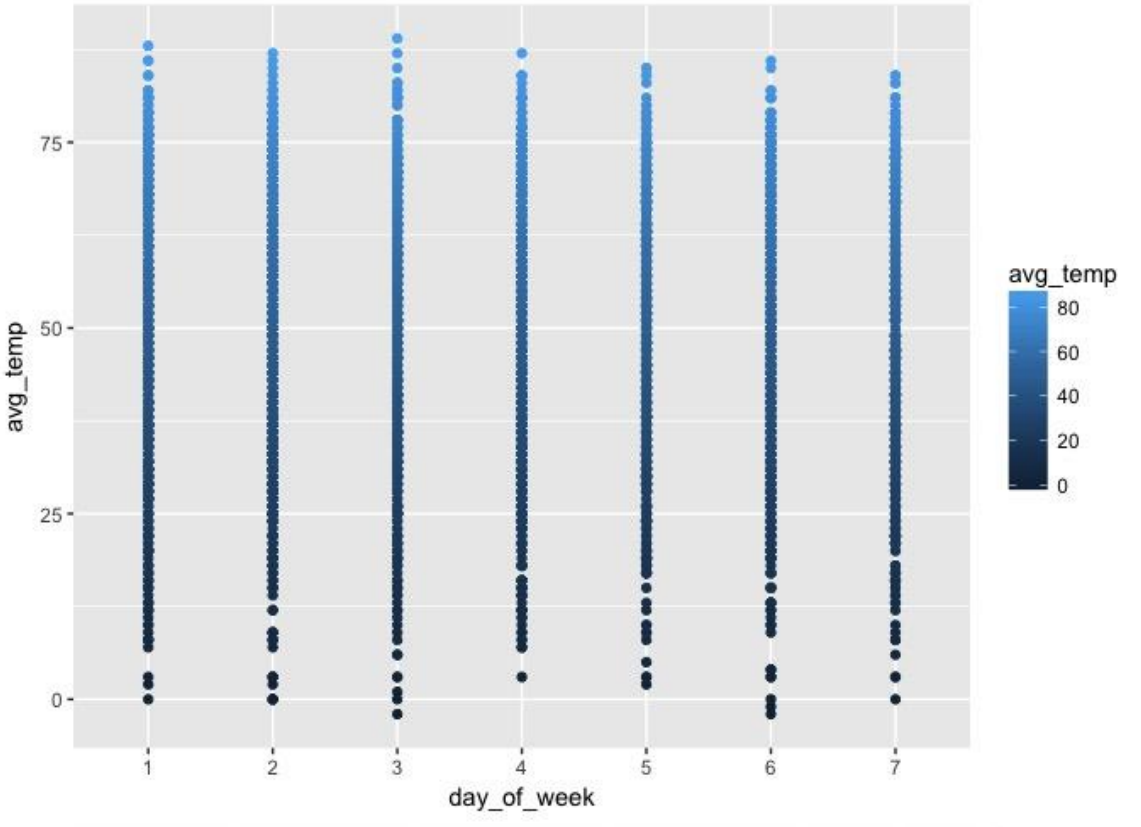


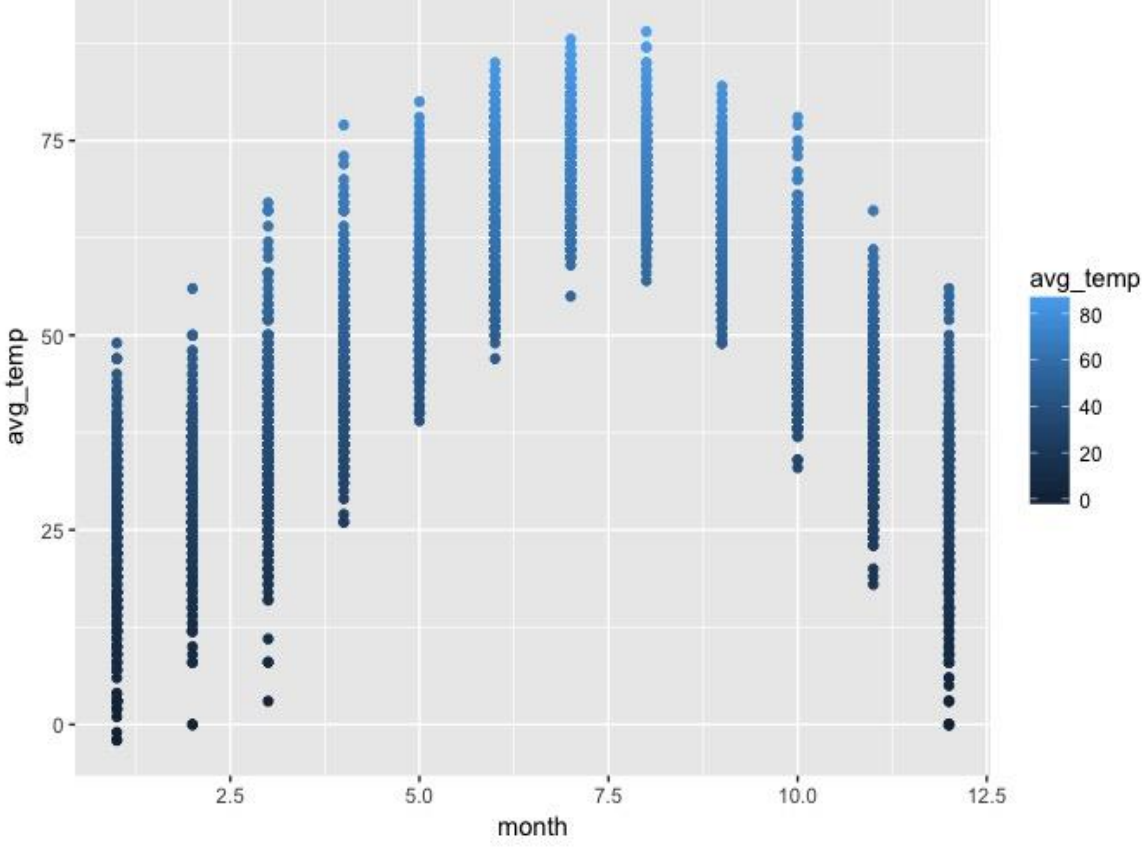

75
50
25
0
avg_temp
1998
2000
2002
2004
2006
date

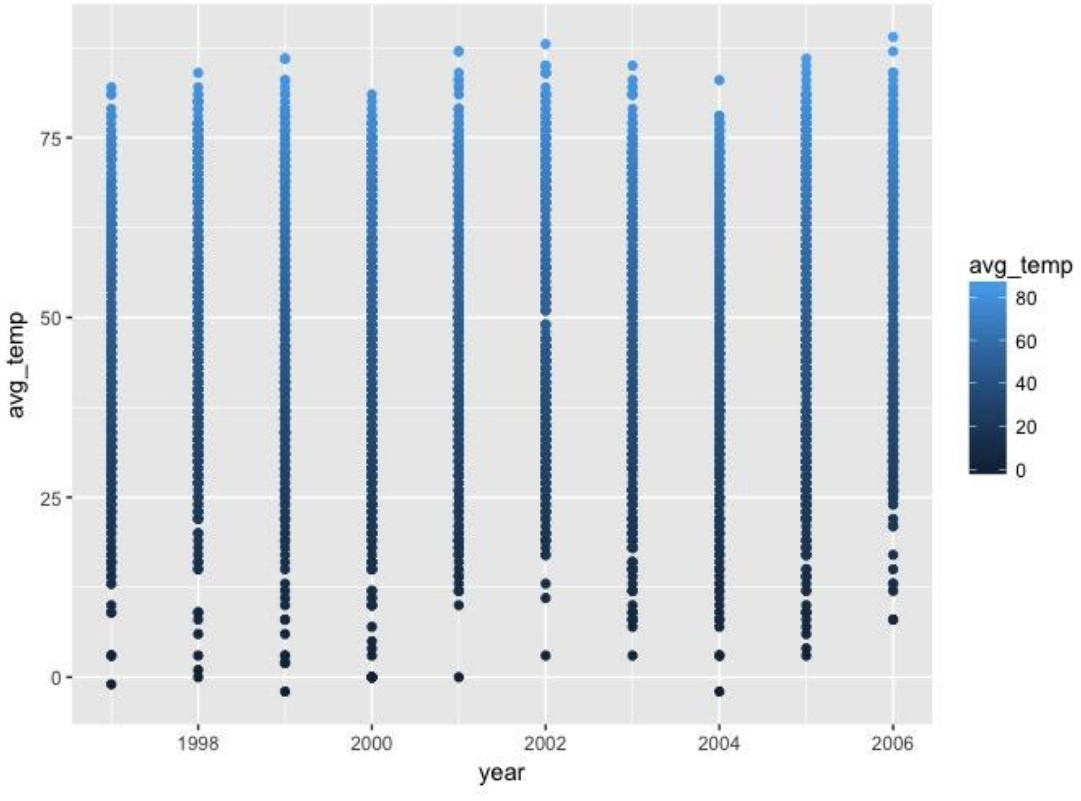

55
50
45
40
35
avg_temp
1998
2000
2002
2004
2006
date

3. Let's identify the timescales (e.g., year, month, day of the week) which best predict the variation in the temperature data. Extract the year, month, and day of the week into new columns. (Hint: Review how we manipulated date/times in the slides and reference the lubridate manual on D2L.) Save the output of head() on the resulting tibble. What types should they be? Generate plots for temperature vs the year, month, and day of the week. What type of plots are appropriate? Which timescales are most predictive of the variations?
   - Save the output of head() on the resulting tibble.
     - library(lubridate)
     - mutate(mkw, day_of_week = as.factor(wday(mkw$date)))
     - mkw <- mutate(mkw, day_of_week = as.factor(wday(mkw$date)))
     - mutate(mkw, year = year(mkw$date))
     - mkw <- mutate(mkw, year = year(mkw$date))
     - mutate(mkw, month = month(mkw$date))
     - mkw <- mutate(mkw, month = month(mkw$date))
     - head(mkw)
   - Here the data type for year, month, and day of week.

day_of_week	year	month
<fct>	<dbl>	<dbl>
4	1997	1.00
5	1997	1.00
6	1997	1.00
7	1997	1.00
1	1997	1.00
2	1997	1.00

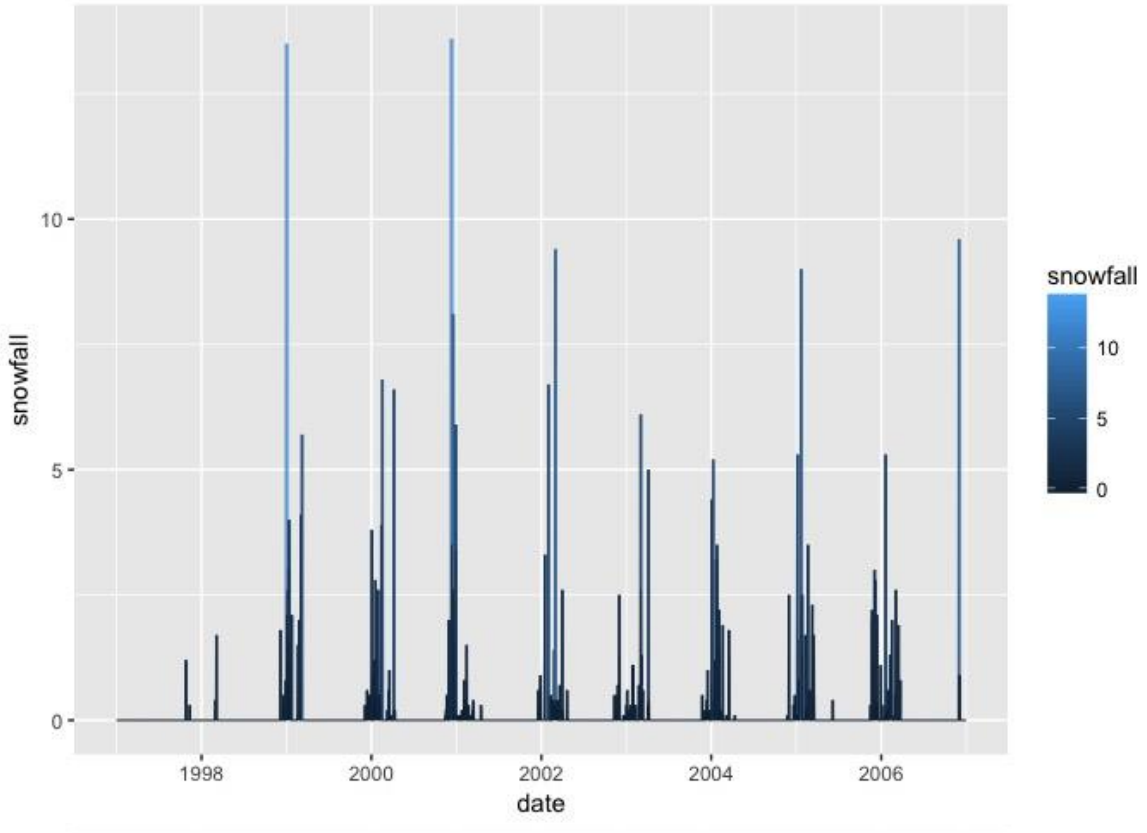

avg_temp
75
50
25
0
1
2
3
4
5
6
7
day_of_week
avg_temp
80
60
40
20
0

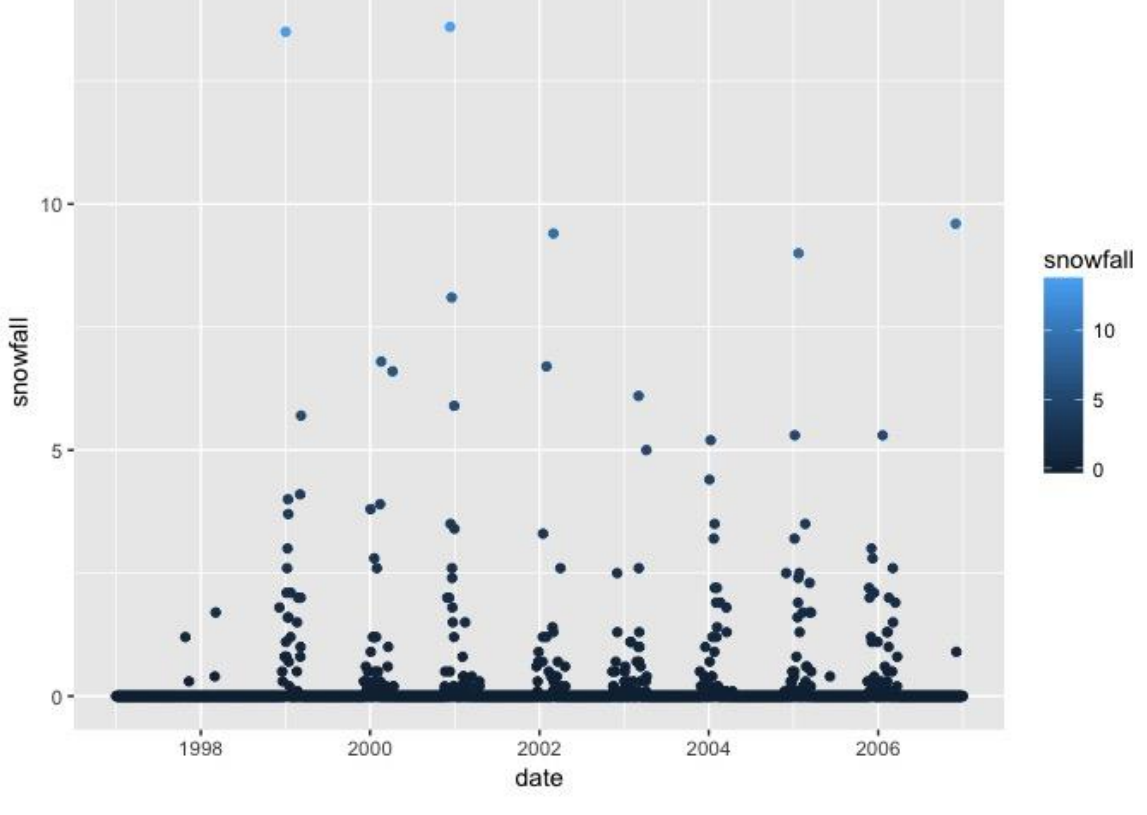

avg_temp
75
50
25
0
2.5
5.0
7.5
10.0
12.5
month
avg_temp
80
60
40
20
0



4. Plot the snowfall over time. What two columns do you need? Try plotting the data with both the geom_line() and geom_point() plot types – which one do you think is better suited?

- I am going to use Date and snowfall columns.
- I think Point Plot is better suited. You can easily differentiate the lower amount of snowfall with the point plot then the line.
- ggplot(data = mkw) + geom_line(mapping = aes(x = date, y = snowfall, colour = snowfall))
- ggplot(data = mkw) + geom_point(mapping = aes(x = date, y = snowfall, colour = snowfall))

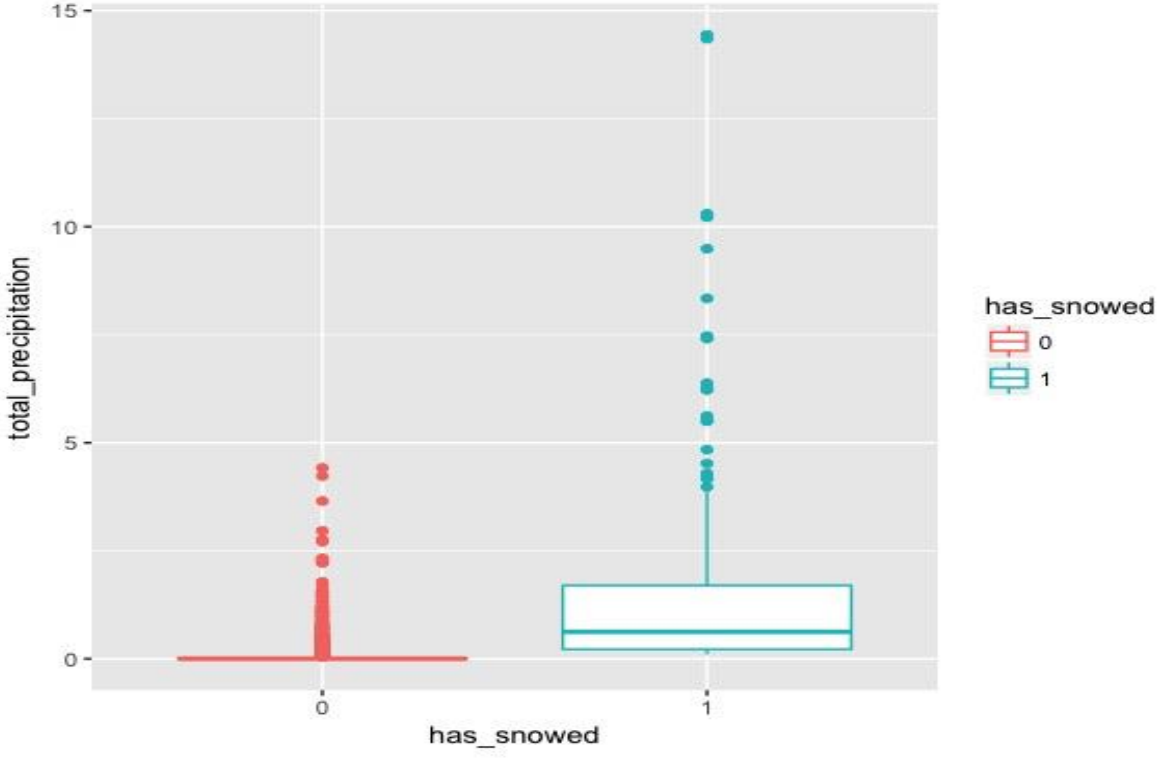

snowfall
10
5
0
1998
2000
2002
2004
2006
date
snowfall
10
5
0

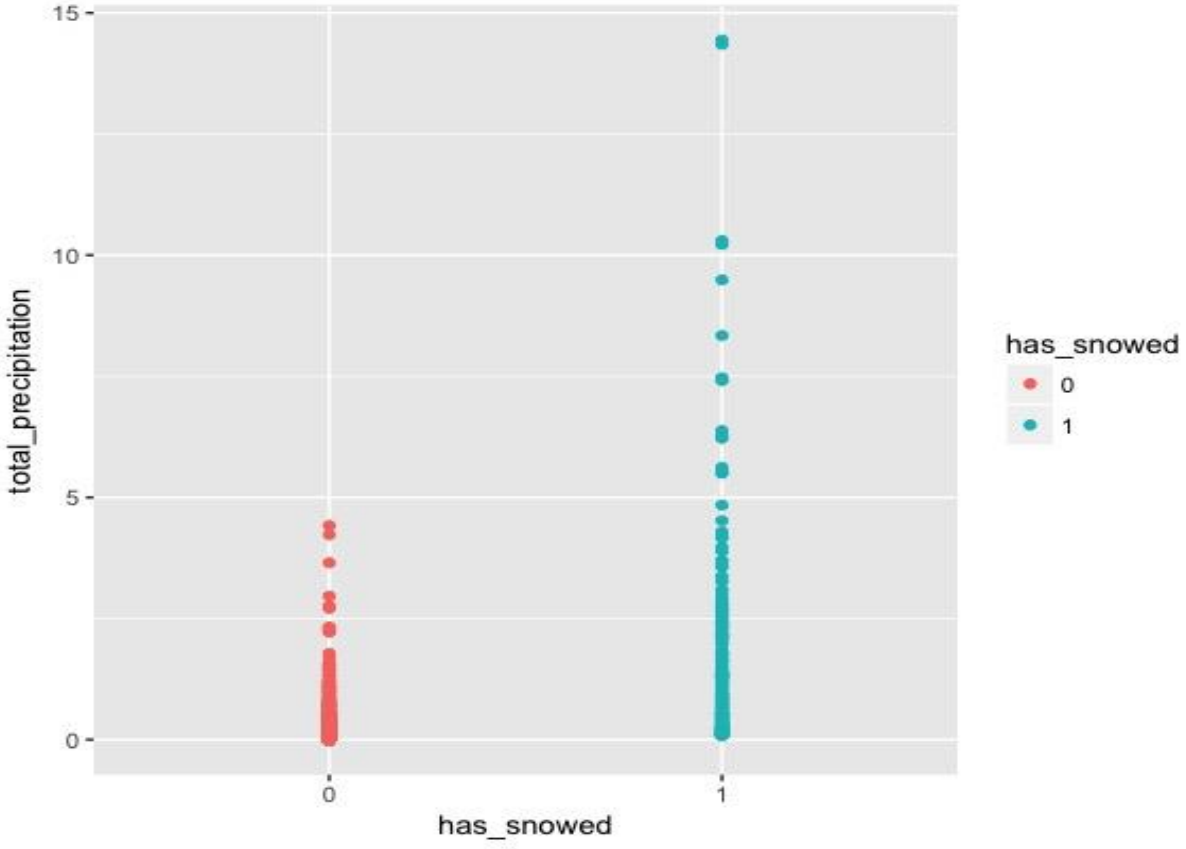

snowfall
10
5
0
1998
2000
2002
2004
2006
date
snowfall
10
5
0

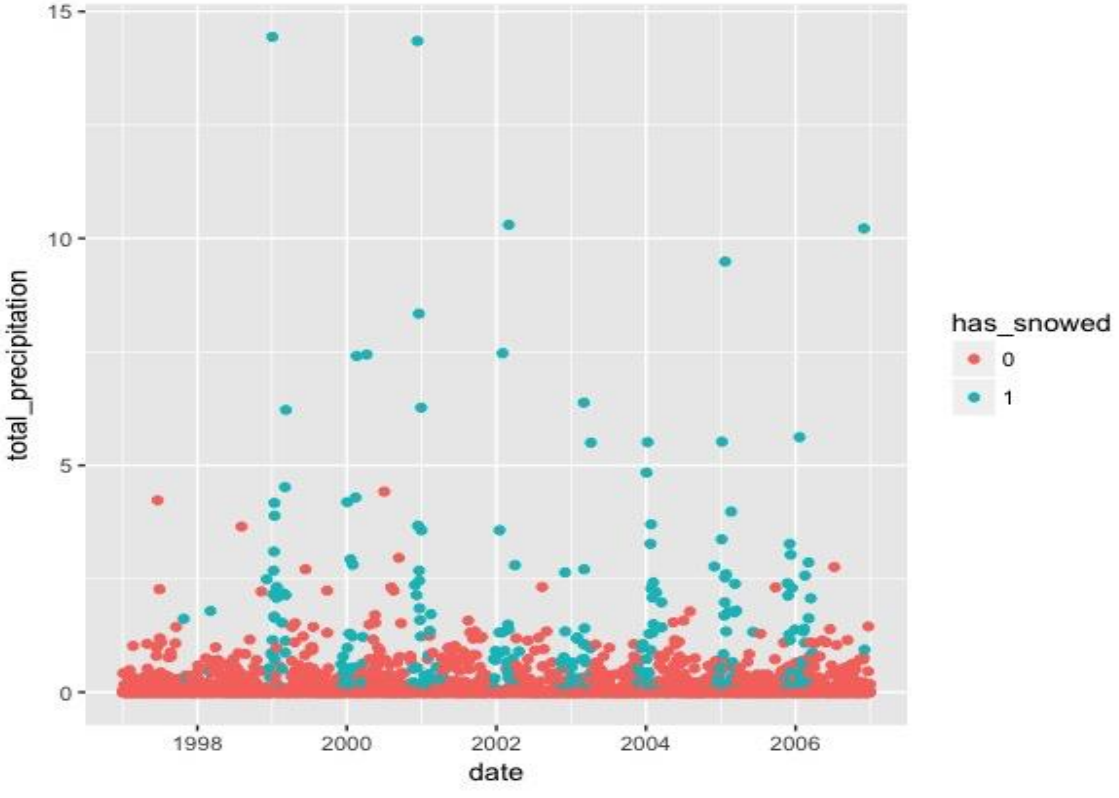
5. Snow only occurs at lower temperatures. Create a new column called total_precipitation that contains the sums of the snowfall and rainfall columns. Then, create another column called has_snowed which is a factor or logical (Boolean) type indicating whether it snowed that day. (Hint: Think about how we used ifelse() to replace NAs with the string "none".) Save the output of head() on the resulting tibble. Lastly, plot total_precipitation vs has_snowed and color it by has_snowed. What type of plot is most appropriate?

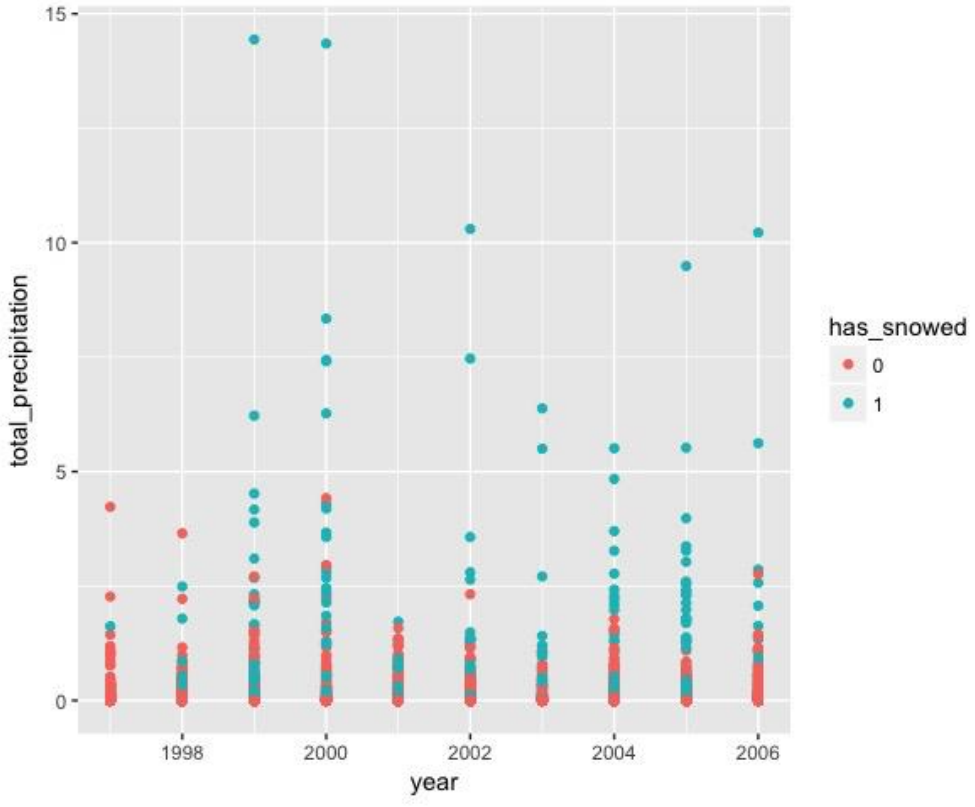
    - Here I create a two new columns called total_precipitation and has_snowed.

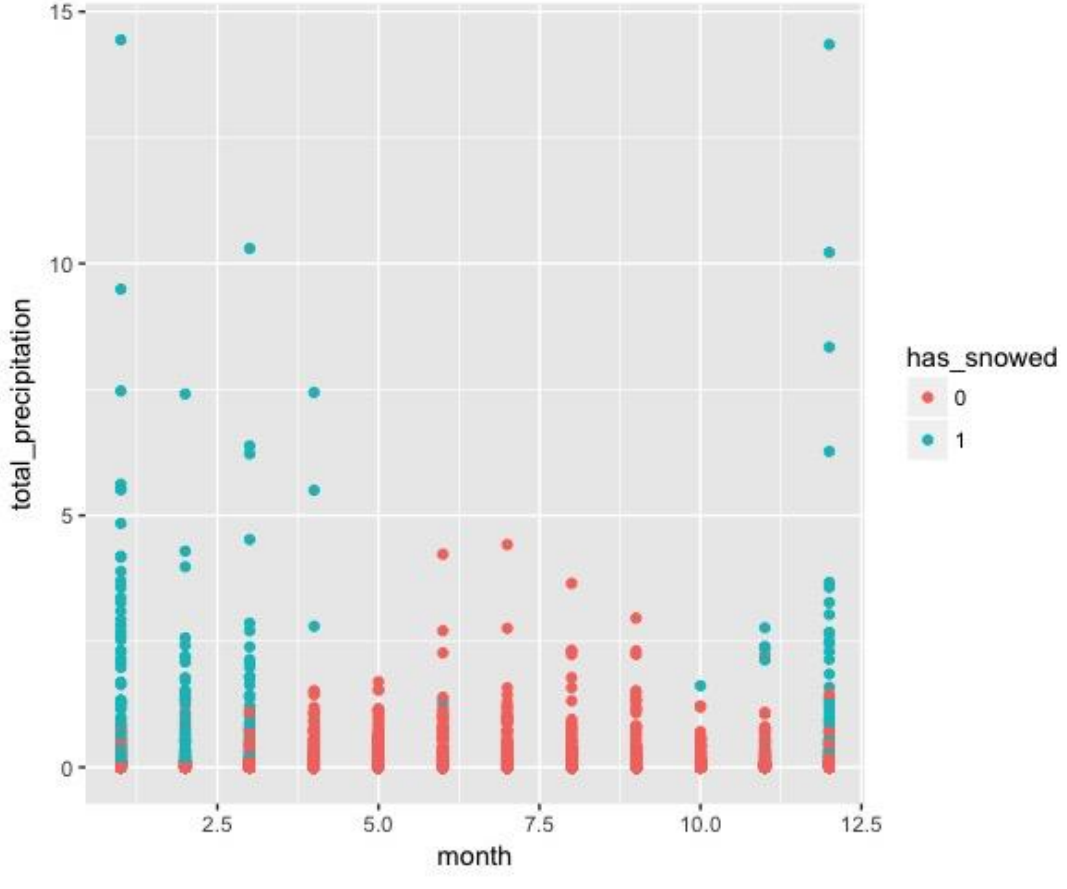
total_precipitat…	has_snowed
<dbl>	<fct>
0.0200	0
0	0
0	0
0.420	0
0	0
0	0

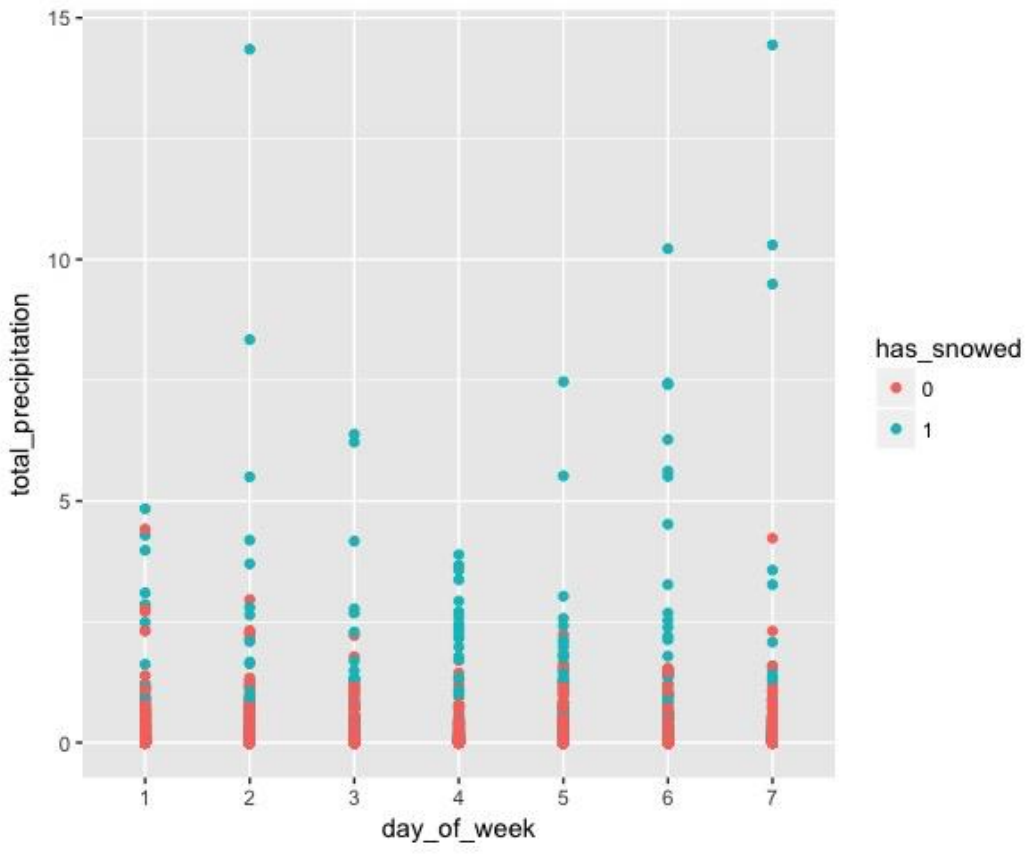



15
10
5
0
total_precipitation
0
1
has_snowed
has_snowed
0
1


15
10
5
0
total_precipitation
1998
2000
2002
2004
2006
date
has_snowed
0
1

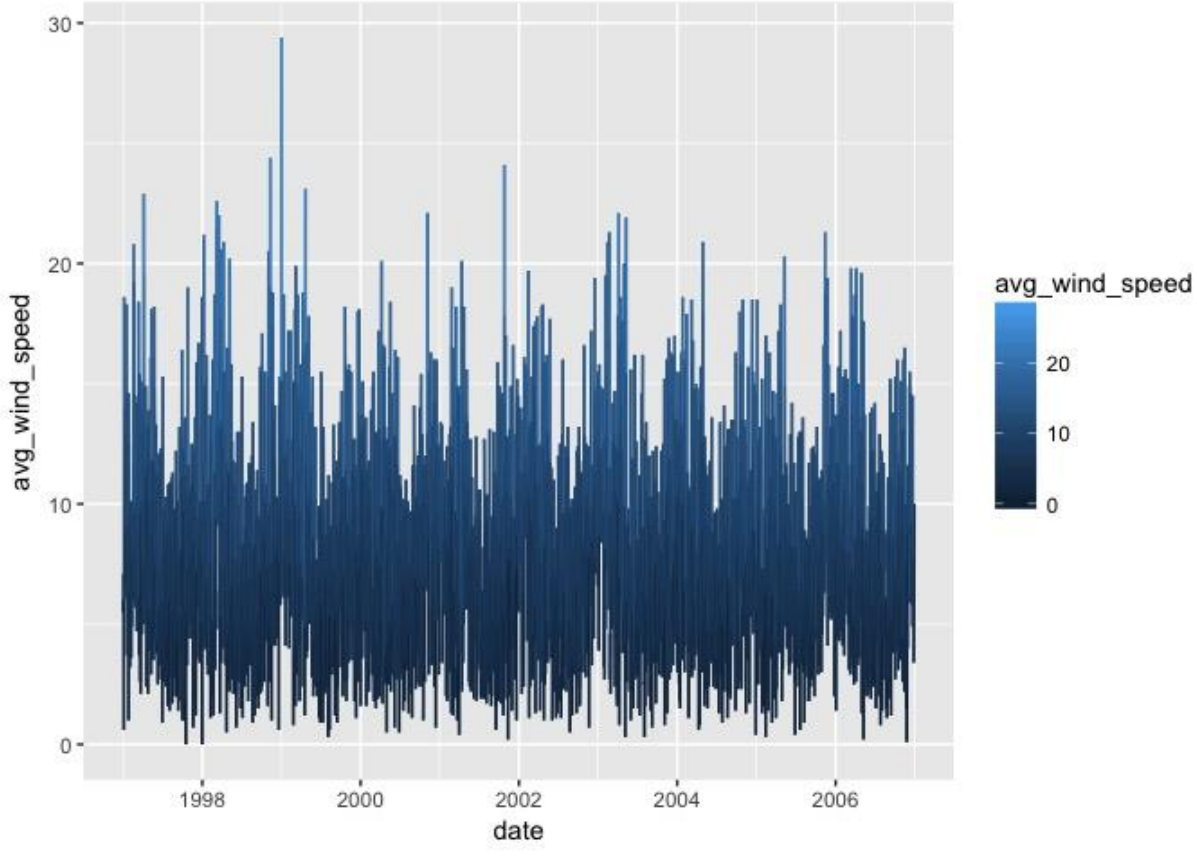

15
10
5
0
total_precipitation
1998
2000
2002
2004
2006
year
has_snowed
0
1

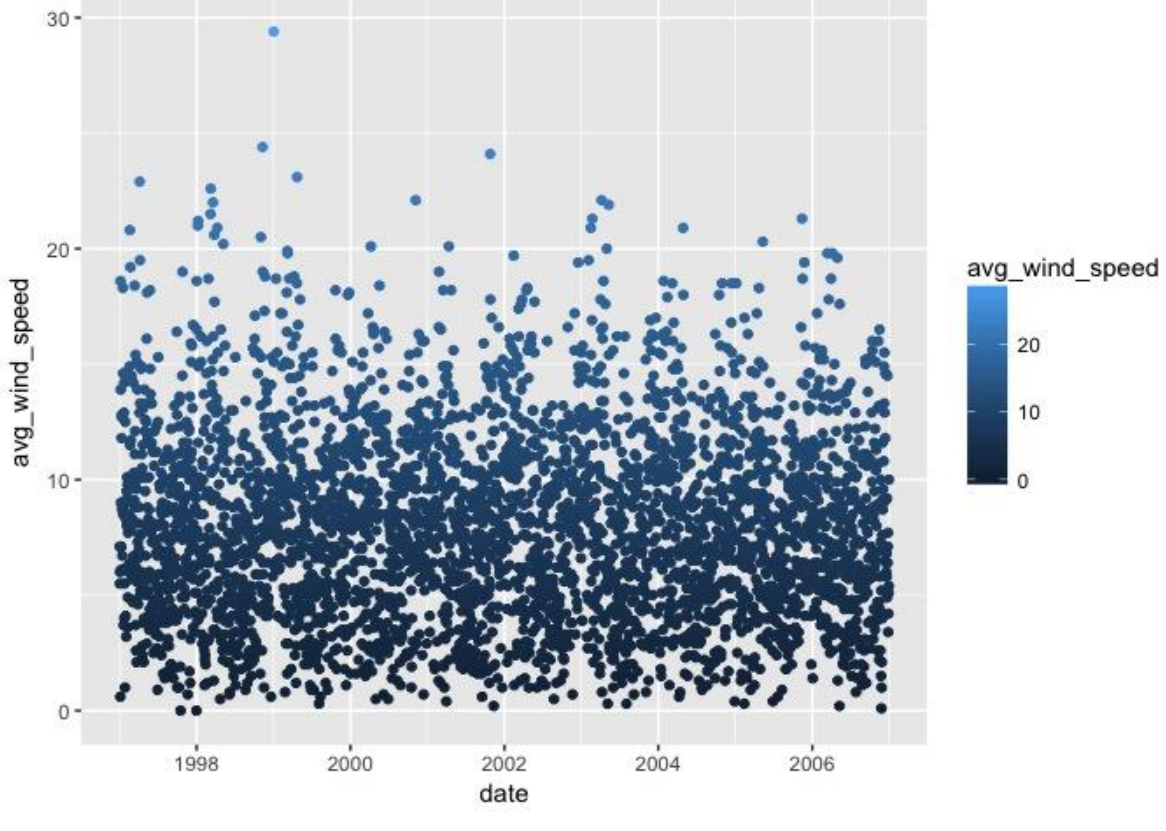

15
10
5
0
total_precipitation
2.5
5.0
7.5
10.0
12.5
month
has_snowed
0
1



- Geom_point plot is appropriate for this data because we can see all data clearly.

6. Plot the avg_wind_speed over time. Do you see a pattern?
   - ggplot(data = mkw) + geom_line(mapping = aes(x = date, y = avg_wind_speed, col = avg_wind_speed))
   - ggplot(data = mkw) + geom_point(mapping = aes(x = date, y = avg_wind_speed, col = avg_wind_speed))
   - ggplot(data = mkw) + geom_area(mapping = aes(x = date, y = avg_wind_speed, col = "red"))
   - Yes, I can see some pattern. I can easily differentiate avg_wind_speed vs date with different plot.


30
20
10
0
avg_wind_speed
1998
2000
2002
2004
2006
date
avg_wind_speed
20
10
0


30
20
10
0
avg_wind_speed
1998
2000
2002
2004
2006
date
avg_wind_speed
20
10
0

7. Plot the snowfall vs the avg_wind_speed. Do you see a pattern?

- ggplot(data = mkw) + geom_line(mapping = aes(y = snowfall, x = avg_wind_speed, col = avg_wind_speed))
- ggplot(data = mkw) + geom_point(mapping = aes(y = snowfall, x = avg_wind_speed, col = avg_wind_))
- ggplot(data = mkw) + geom_area(mapping = aes(y = snowfall, x = avg_wind_speed, col = "red"))
- Point plot we can see avg_wind_speed vs snowfall data clearly.