A 50 Year Analysis of Temperature and Precipitation Data at San Francisco International Airport

Author: Hilary Hickingbotham

Introduction

Mark Twain famously said, "The coldest winter I ever spent was a summer in San Francisco." I can attest to this statement, as I grew up just south of San Francisco. For this project, I wanted to look at 50 years of weather data (1970-2020) from San Francisco International Airport, and expand my skills using R and look for trends in the temperature and precipitation data to answer the following questions:

- When is the hottest time of the year, and does it often rain during that time?
- When is the coldest time of year, and does it often rain during that time?
- What are the precipitation levels throughout the year?
- What conclusions can be made by looking at the temperature and precipitation data?

Gathered Data

The data was gathered using the NOAA database, looking specifically at the weather station located at San Francisco International Airport (Station USW00023234). Data pulled was for Max and Min Temperature and Measured Precipitation from 1Jan1970-31Dec2020. In total, there were 18,628 entries in the data set.

Data Cleaning and Organization

Before any analysis could be performed, a few data cleaning and organizational tasks were completed in preparation. These tasks included:

- Remove NAs in the data if any.
- Fixing the Date in R from a “character” to “date” variable for time analysis.
- Added a column that would assign TRUE/FALSE if there was any measured precipitation on that day.

```
#Identifying NA values and removing them
(cols_withNA <- apply(sfoweather50, 2, function(x) sum(is.na(x))))
sfoweather50_2 <- sfoweather50[complete.cases(sfoweather50),]
View(sfoweather50_2)

#Fixing the date variable
sfoweather50$date_fixed <- as.Date(sfoweather50$DATE, format = "%Y-%m-%d")

#adding the True/False indicator for precipitation
sfoweather50$RAIN[sfoweather50$PRCP > 0] <- TRUE
sfoweather50$RAIN[sfoweather50$PRCP == 0] <- FALSE
```

Data Summary

A call to R will give a nice summary of the data as seen below.

```
Summary(sfoweather50)
```

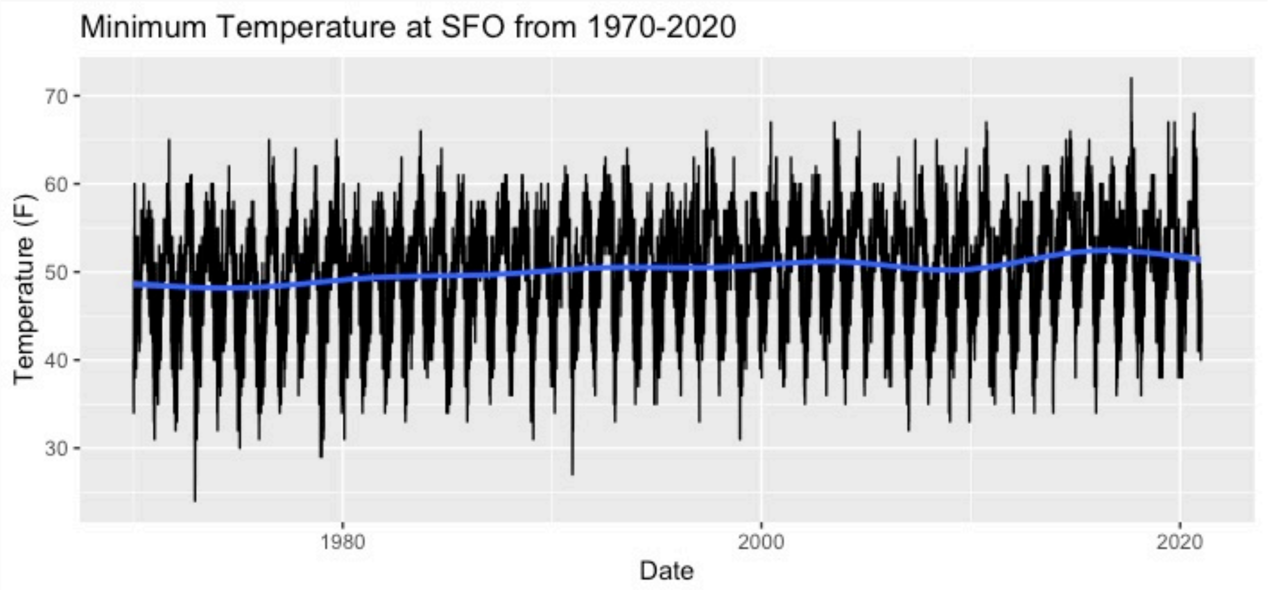
```
STATION               NAME                DATE                 PRCP        
 Length:18608       Length:18608       Length:18608       Min.   :0.00000  
 Class :character   Class :character   Class :character   1st Qu.:0.00000  
 Mode  :character   Mode  :character   Mode  :character   Median :0.00000  
                                                          Mean   :0.05331  
                                                          3rd Qu.:0.00000  
                                                          Max.   :5.59000  
      TMAX             TMIN          RAIN          date_fixed        
 Min.   : 37.00   Min.   :24.00   FALSE:15356   Min.   :1970-01-01  
 1st Qu.: 60.00   1st Qu.:47.00   TRUE : 3252   1st Qu.:1982-09-26  
 Median : 65.00   Median :51.00                 Median :1995-06-22  
 Mean   : 65.84   Mean   :50.23                 Mean   :1995-06-23  
 3rd Qu.: 71.00   3rd Qu.:54.00                 3rd Qu.:2008-03-17  
 Max.   :105.00   Max.   :72.00                 Max.   :2020-12-31  
```

Compared to other places in the United States, it only precipitated 17.45% of days from 1970-2020 in San Francisco. The median TMAX was 65° and the median TMIN was 51°.

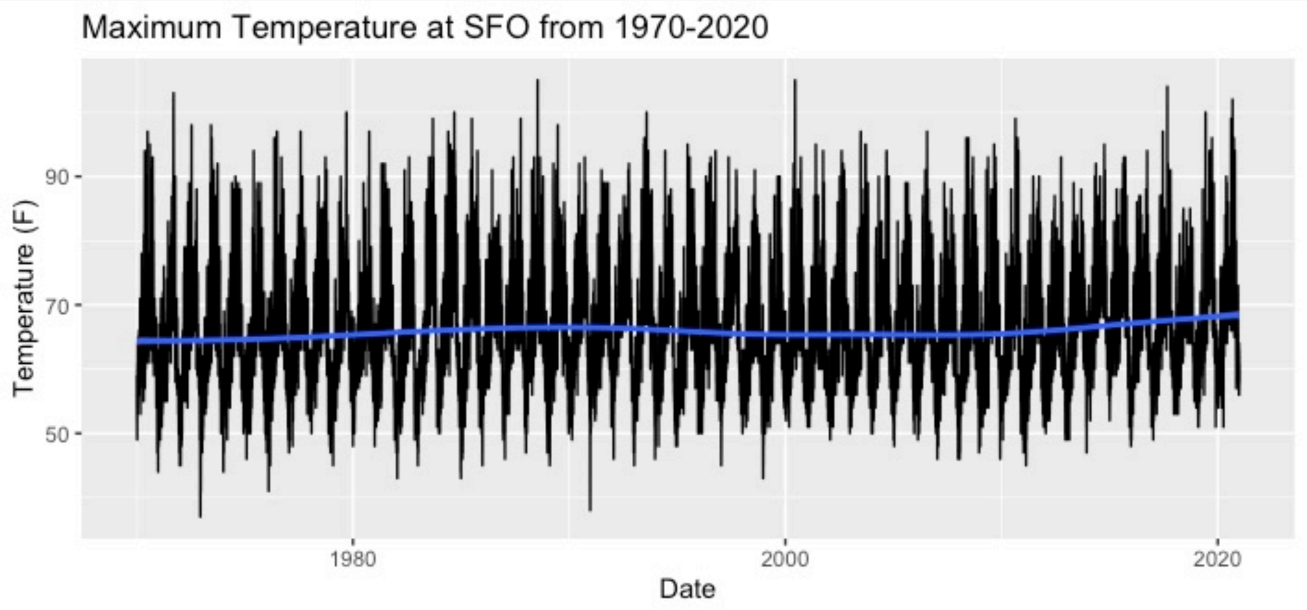
Basic Graphs

A few basic graphs were generated, using the ggplot2 package. These graphs include the geom_smooth() function that will generate a linear model predictor trend line.

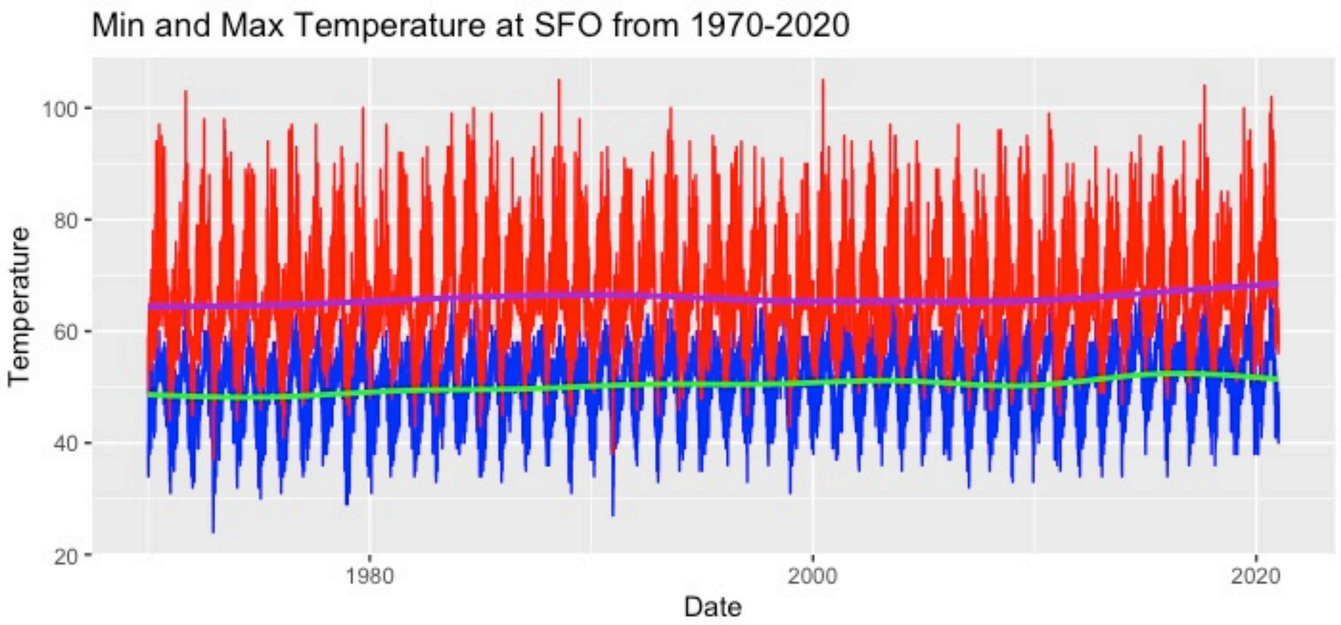
```
ggplot(sfoweather50_2, aes(x= date_fixed, y = TMIN)) + geom_line() +
  geom_smooth() +
  xlab('Date') +
  ylab('Temperature (F)') +
  ggtitle('Minimum Temperature at SFO from 1970-2020')
```



```
ggplot(sfoweather50_2, aes(x= date_fixed, y = TMAX)) + geom_line() +
  geom_smooth() +
  xlab('Date') +
  ylab('Temperature (F)') +
  ggtitle('Maximum Temperature at SFO from 1970-2020')
```



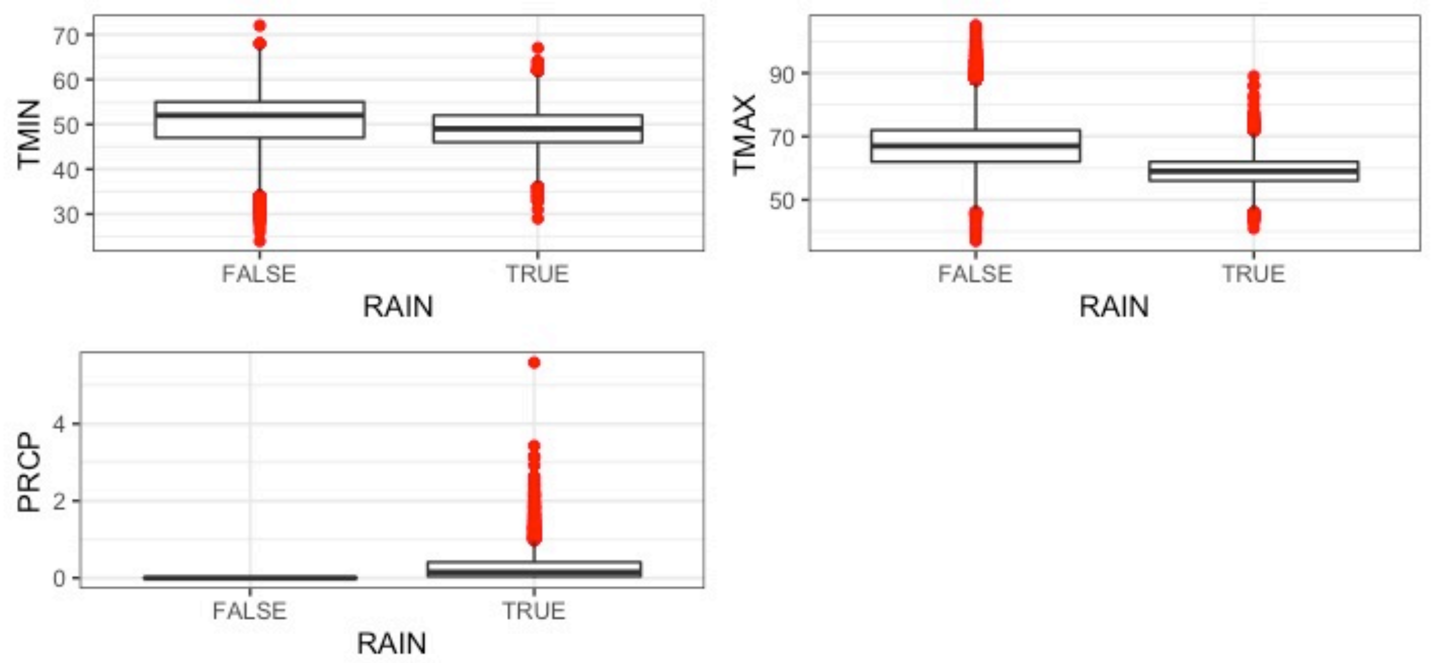
```
#plotted together
ggplot() +
  geom_line(data = sfoweather50_2, aes(x= date_fixed, y = TMIN), color = "blue") +
  geom_line(data= sfoweather50_2, aes(x= date_fixed, y = TMAX), color = "red") +
  geom_smooth(data = sfoweather50_2, aes(x= date_fixed, y = TMIN), color = "green") +
  geom_smooth(data = sfoweather50_2, aes(x= date_fixed, y = TMAX), color = "purple") +
  xlab('Date') +
  ylab('Temperature') +
  ggtitle('Min and Max Temperature at SFO from 1970-2020')
```



Data Visualization

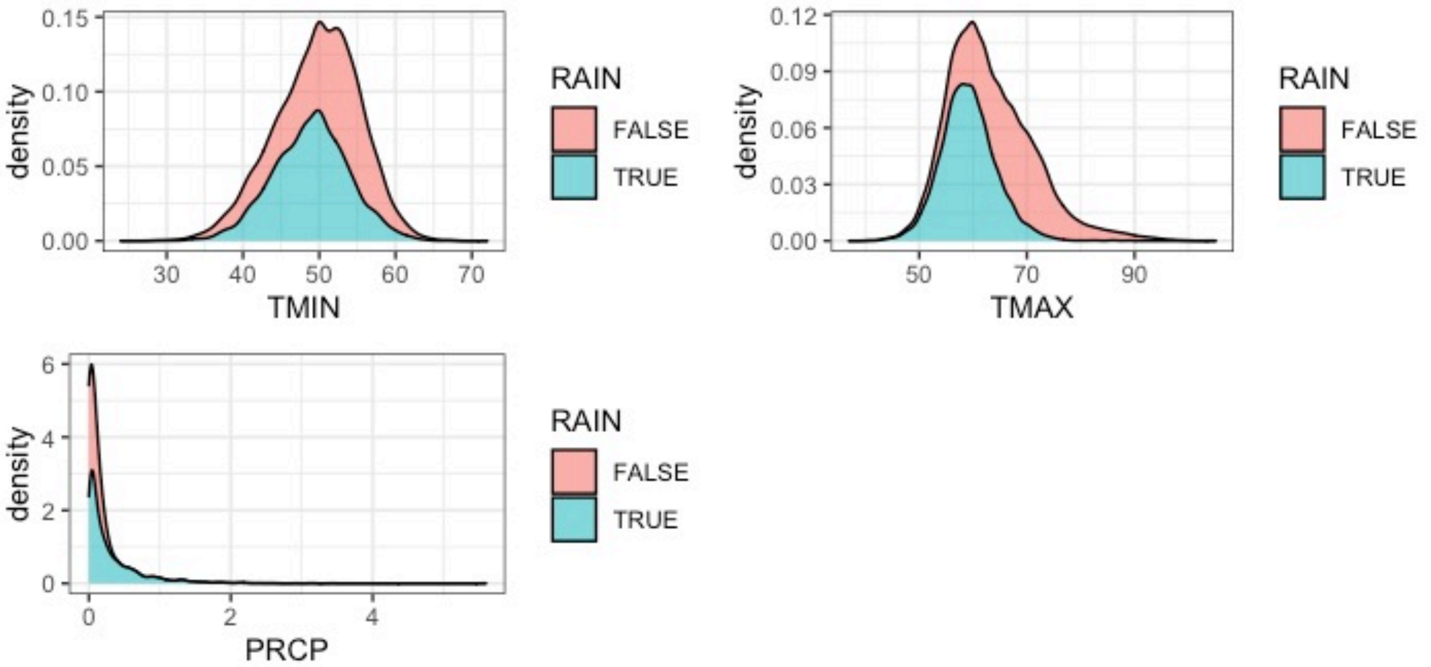
I wanted to be able to visualize the data if a few different ways prior to any analysis.

```
sfoweather50_2$RAIN <- as.factor(sfoweather50_2$RAIN)
a1 <- sfoweather50_2 %>% ggplot(aes(RAIN , TMIN)) + geom_boxplot(outlier.color = "red") +
theme_bw()
a2 <- sfoweather50_2 %>% ggplot(aes(RAIN , TMAX)) + geom_boxplot(outlier.color = "red") +
theme_bw()
a3 <- sfoweather50_2 %>% ggplot(aes(RAIN , PRCP)) + geom_boxplot(outlier.color = "red") +
theme_bw()
grid.arrange(a1, a2, a3, ncol = 2, nrow = 2)
```



This box plot is looking at the three variables, Minimum Temperature (TMIN), Maximum Temperature (TMAX), and Precipitation (PRCP) against the TRUE/FALSE variable as to whether there was any precipitation. Looking at these graphs, some conclusions that could be made are that the TMIN temperature for precipitation is slightly lower than no precipitation and the TMAX temperature for precipitation is also lower than no precipitation. The third box plot, is exactly as we expect, with precipitation FALSE being entirely at 0, and precipitation having a distribution if TRUE.

```
a4 <- sfoweather50_2 %>% ggplot(aes(TMIN , fill = RAIN)) + geom_density(position =
"stack" , alpha = 0.6) + theme_bw()
a5 <- sfoweather50_2 %>% ggplot(aes(TMAX, fill = RAIN)) + geom_density(position = "stack"
, alpha = 0.6) + theme_bw()
a6 <- sfoweather50_2 %>% ggplot(aes(PRCP, fill = RAIN)) + geom_density(position = "stack"
, alpha = 0.6) + theme_bw()
grid.arrange(a4, a5, a6, ncol = 2, nrow = 2)
```



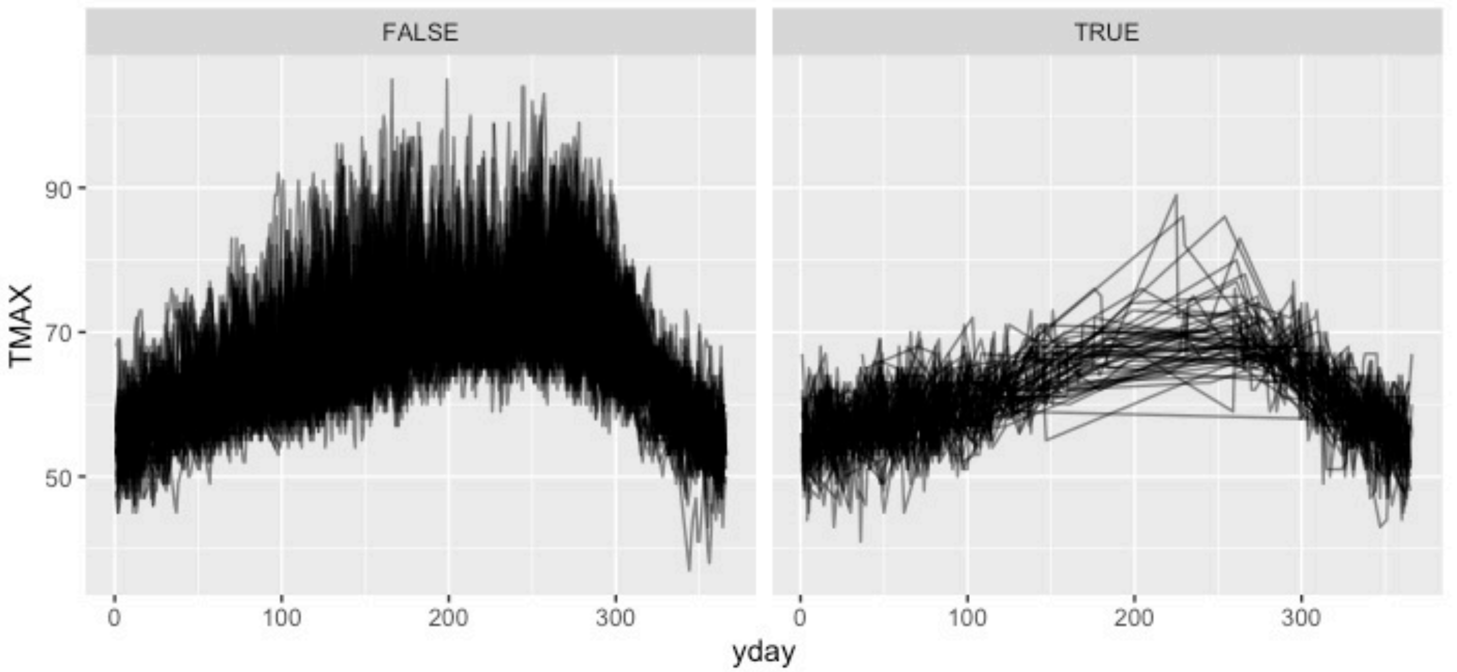
These graphs are looking at the distributions of the TMIN, TMAX and PRCP with and overlay of the precipitation TRUE/FALSE variable. From the graphs, TMIN looks relatively symmetric, TMAX looks right skewed, and PRCP looks similar to an exponential distribution with a large lamda λ.

Data Analysis

Next, I added a column for years, yday, and month in order to analyze the temperatures and precipitation by month.

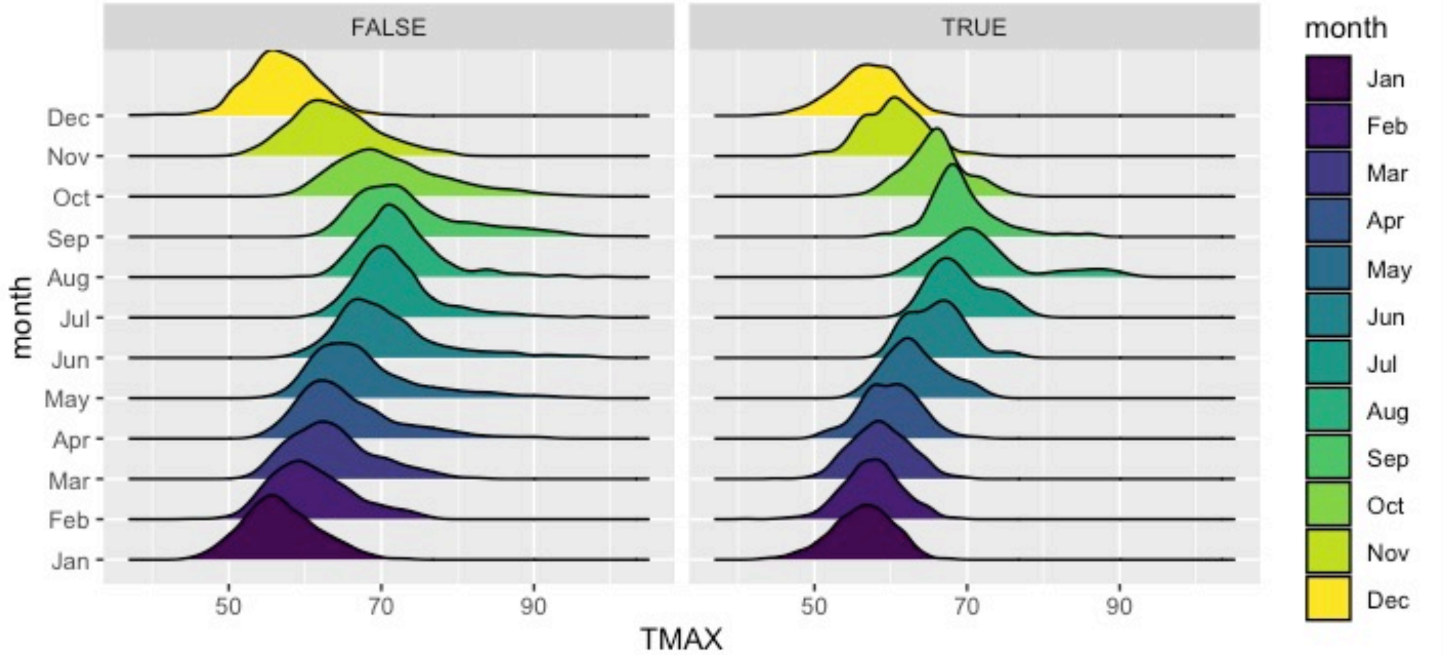
```
# Add columns for year, yday and month
seat_dailysfo50 <- sfoweather50_2 %>%
  mutate(
    year = year(date_fixed),
    yday = yday(date_fixed),
    month = month(date_fixed, label = TRUE))

#Maximum Temperatures
ggplot(seat_dailysfo50, aes(x = yday, y = TMAX)) +
  geom_line(aes(group = year), alpha = 0.5) + facet_wrap(~RAIN)
```



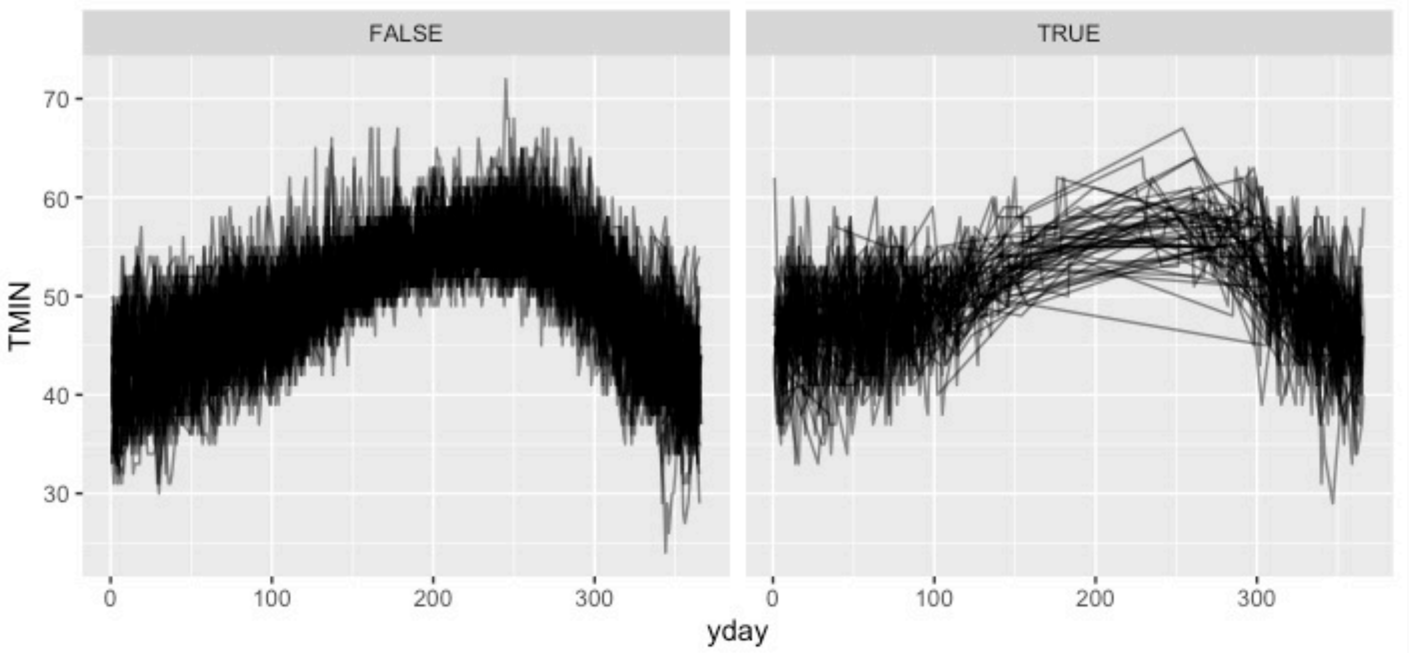
From this graph, we see that TMAX increases from Day 1 (Jan 1) through approximately day 250 (Sept 7) for both precipitation types. Because only 17.45% of days from 1970-2020 had precipitation TRUE, it's unsurprising that the graph looks jumbled during the summer months. This may be an indication that it's less likely to rain in the summer months in San Francisco, which I will explore later.

```
ggplot(seat_dailysfo50, aes(x = TMAX, y = month, height = ..density.., fill = month)) +
  geom_density_ridges(stat = "density") + facet_wrap(~RAIN)
```



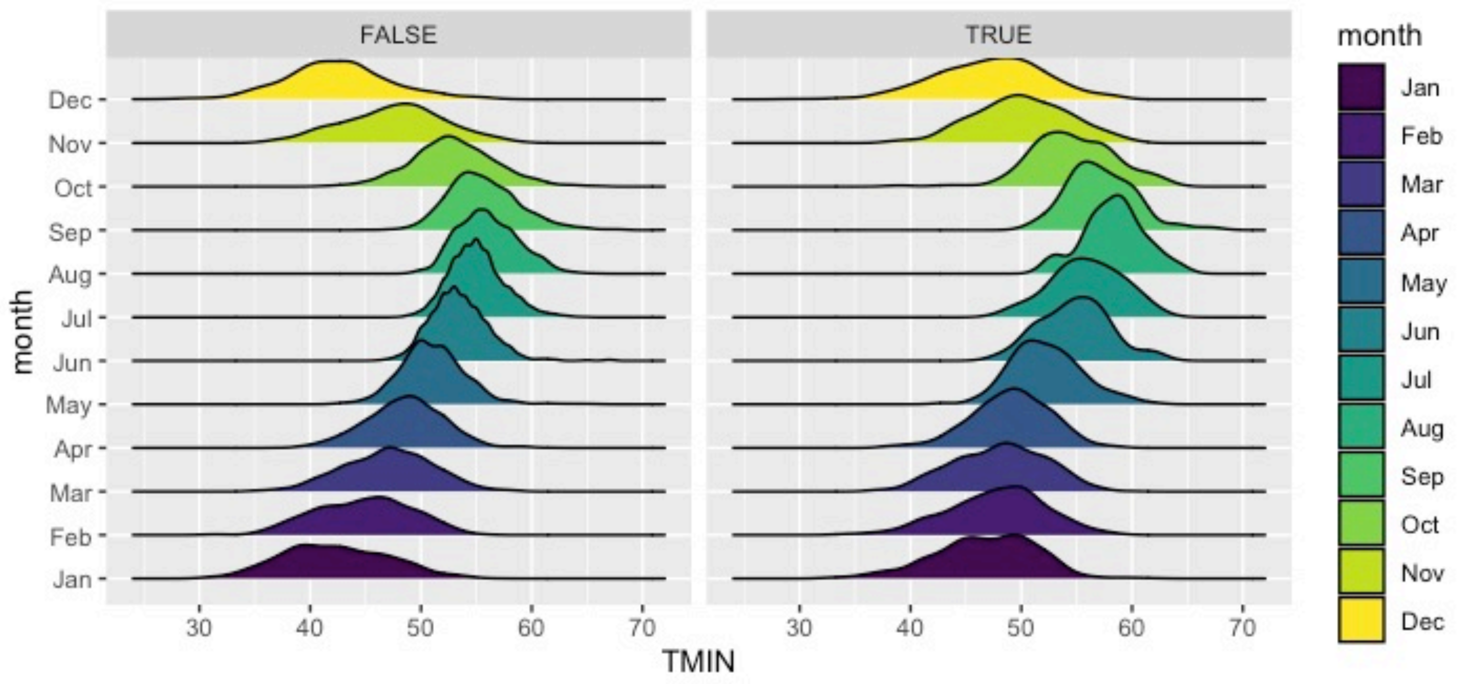
From this plot, we see that whether the precipitation is TRUE or FALSE, hottest temperatures are found during July, August, and September. We can see a pattern of temperature rise from January until its peak in August before falling from September through December.

```
ggplot(seat_dailysfo50, aes(x = yday, y = TMIN)) +
  geom_line(aes(group = year), alpha = 0.5) + facet_wrap(~RAIN)
```



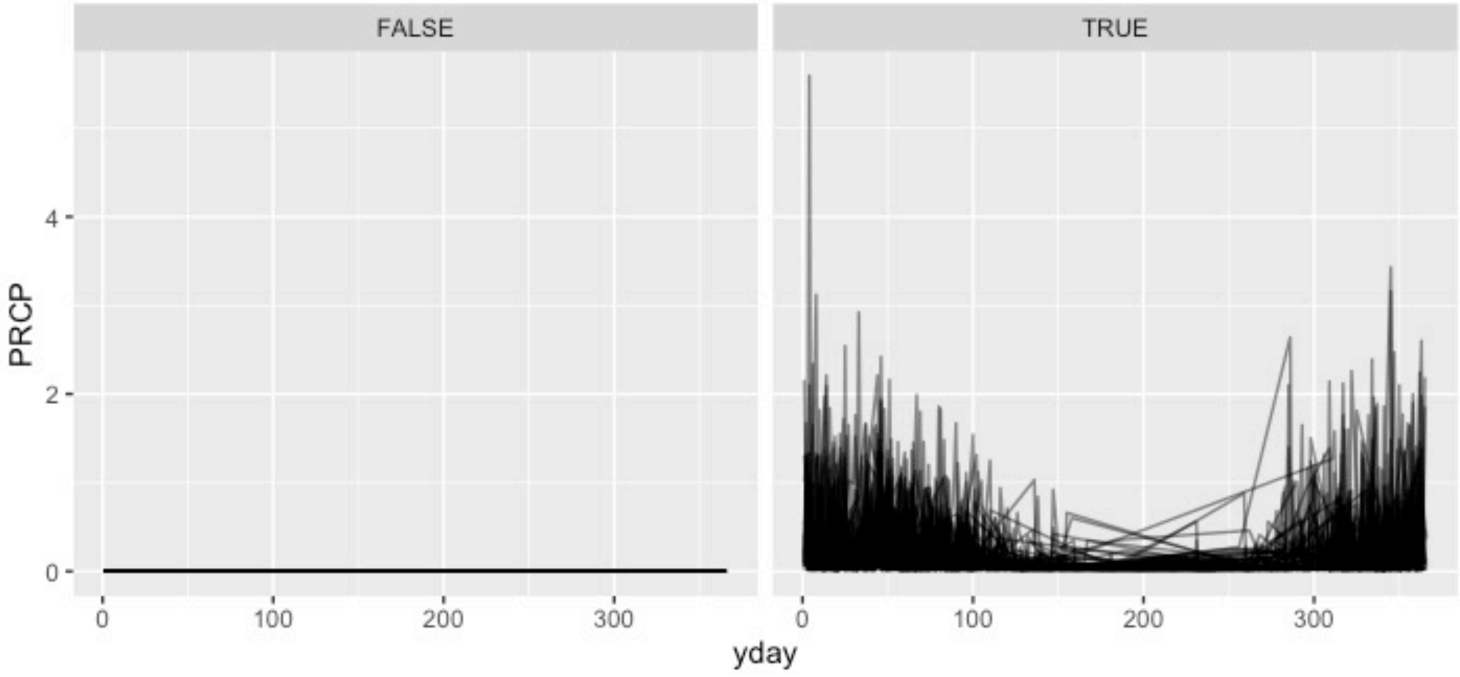
Similar to TMAX, we see that TMIN increases from Day 1 (Jan 1) through approximately day 250 (Sept 7) for both precipitation types. Even more pronounced is the jumbling of the graph for precipitation TRUE during the summer months. This may be an indication that it’s less likely to rain in the summer months in San Francisco, which I will explore later.

```
ggplot(seat_dailysfo50, aes(x = TMIN, y = month, height = ..density.. , fill = month)) +
  geom_density_ridges(stat = "density")  + facet_wrap(~RAIN)
```



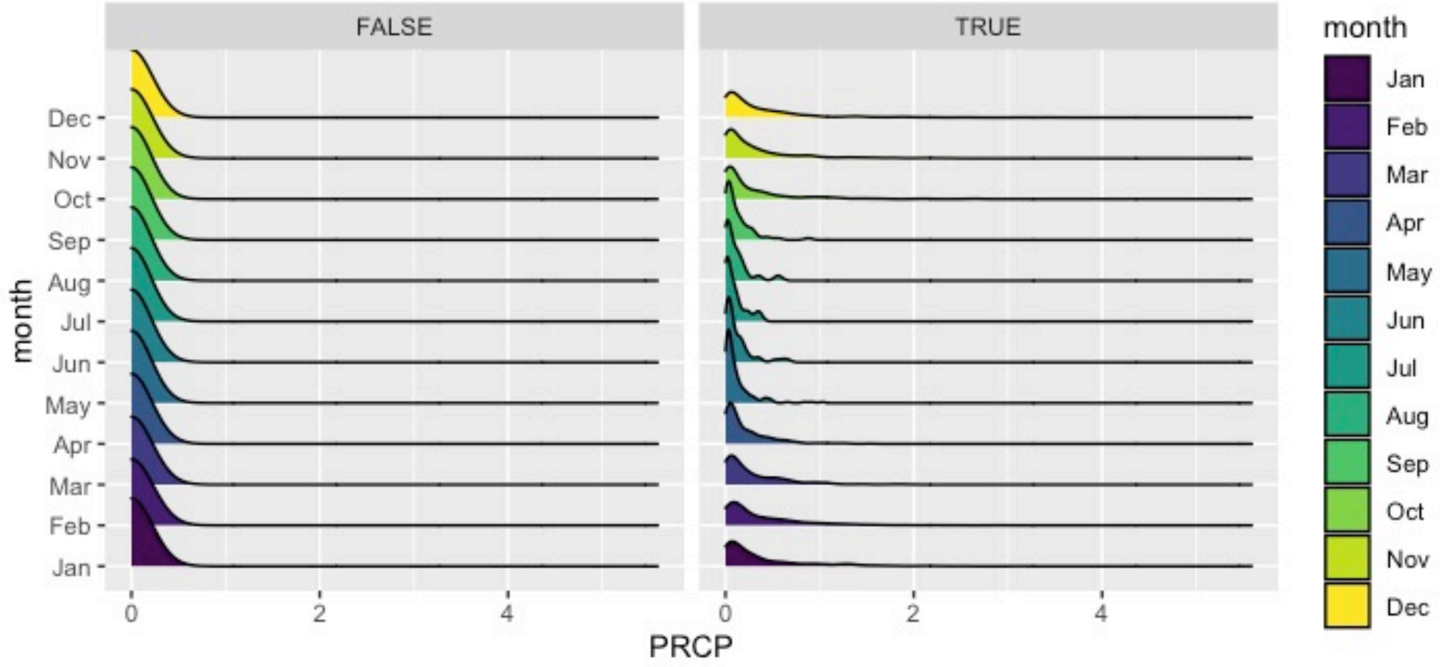
From this plot, we see that whether the precipitation is TRUE or FALSE, coldest temperatures are found during December, January, and February. We can see a pattern of temperature falling from September until its coldest point in December before rising from February through September.

```
ggplot(seat_dailysfo50, aes(x = yday, y = PRCP)) +
  geom_line(aes(group = year), alpha = 0.5) + facet_wrap(~RAIN)
```



The graph on the left is exactly to be expected, that when precipitation is FALSE, the measured precipitation will be 0. When precipitation is TRUE, the values have a wide "V" shape form, from high levels in 0 days then gradually goes down until the 200th day. The precipitation values rise until it peaks again around the 350th day.

```
ggplot(seat_dailysfo50, aes(x = PRCP, y = month, height = ..density.., fill = month)) +
  geom_density_ridges(stat = "density") + facet_wrap(~RAIN)
```



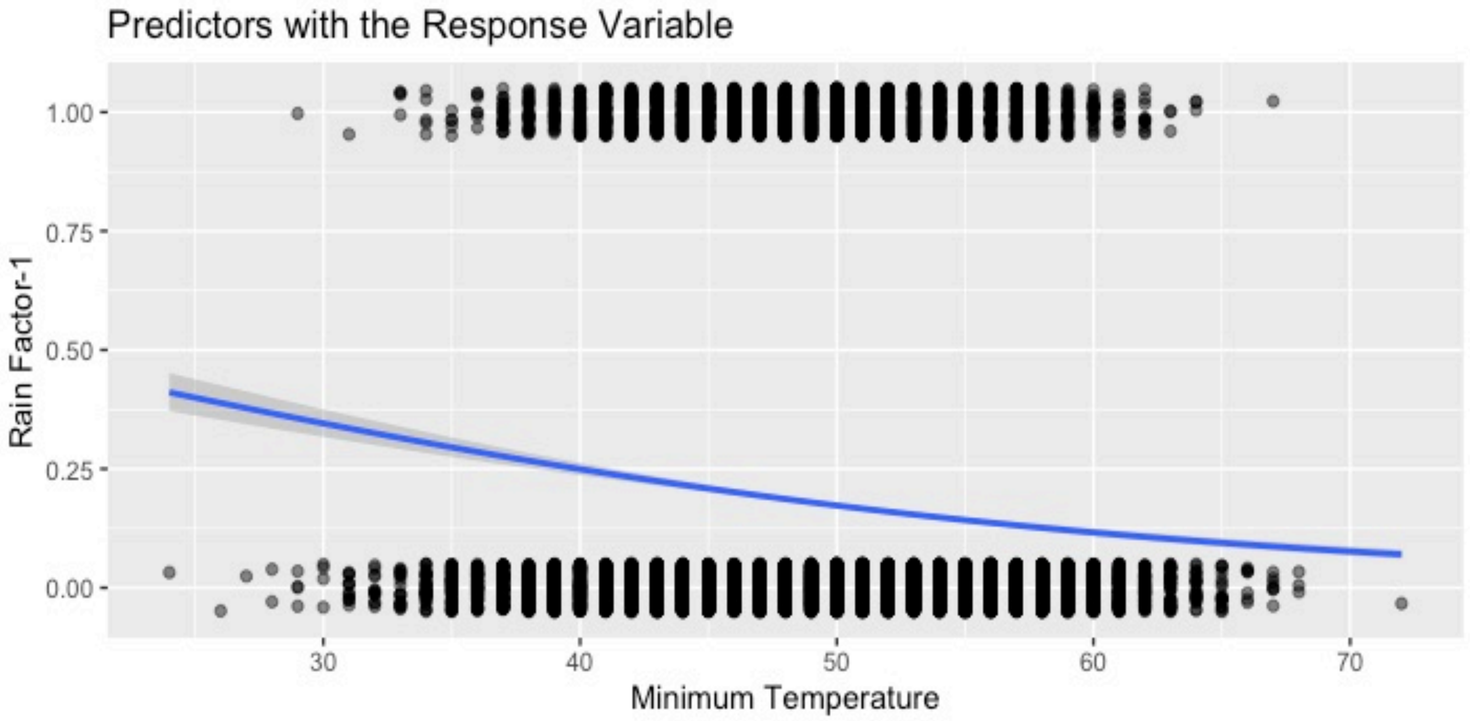
From this plot, we see that if precipitation is FALSE, there are identical distributions around 0, which is to be expected. This may be a default from R to show some sort of analysis. When precipitation is TRUE, daily precipitation values are very small in the summer season, and increase to larger precipitation values in the winter and spring seasons.

Precipitation Predictions

Next, I wanted to see if I could predict whether it would precipitate given a min and max temperature based on the responses.

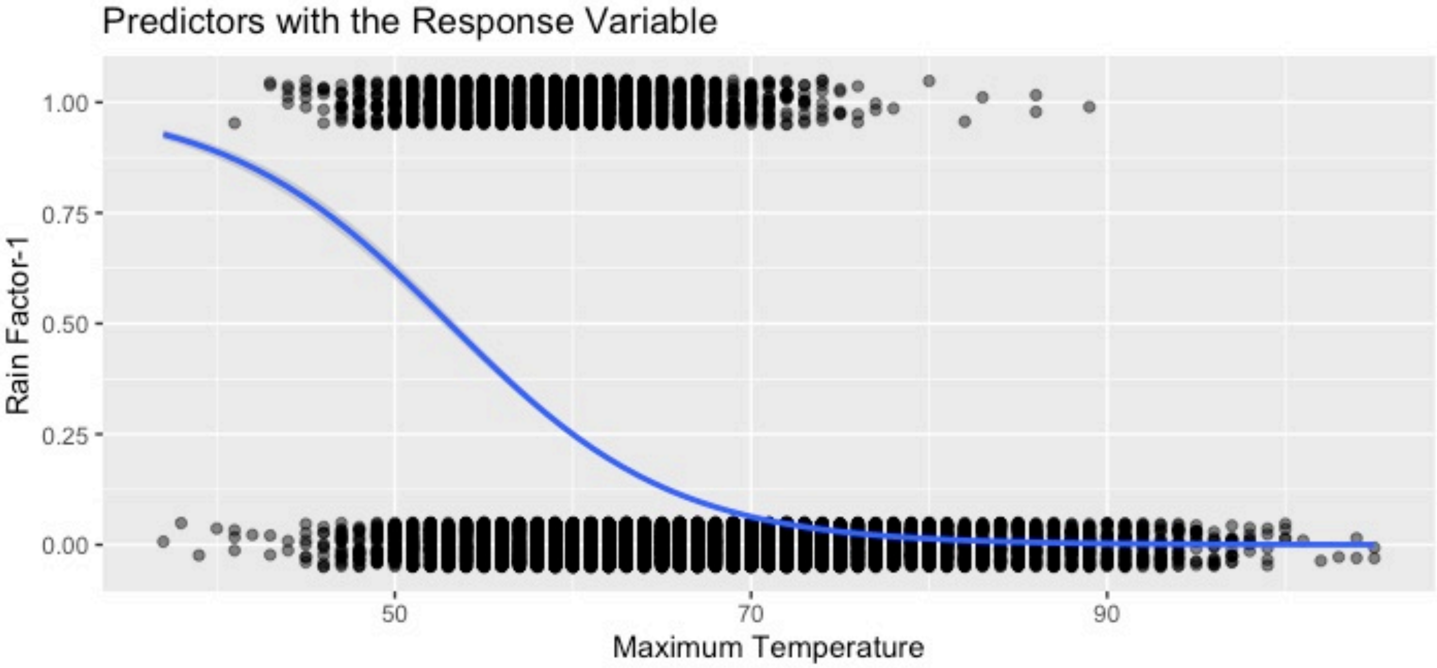
```
#plotting predictors with the response variable
binomial_smooth <- function(...) {
  geom_smooth(method = "glm", method.args = list(family = "binomial"), ...) }

sfoweather50_2 %>% ggplot(aes(x = TMIN, as.numeric(RAIN) - 1)) +
  geom_jitter(width = 0, height = 0.05, alpha = 0.5) + binomial_smooth() +
  xlab('Minimum Temperature') +
  ylab('Rain Factor-1') +
  ggtitle('Predictors with the Response Variable')
```



Looking at predictive curve line, as the minimum temperature increases, our curve line goes down implying that it won't likely precipitate. Around 30°, there is approximately a 30% of rain.

```
sfoweather50_2 %>% ggplot(aes(x = TMAX, as.numeric(RAIN) - 1)) +
  geom_jitter(width = 0, height = 0.05, alpha = 0.5) + binomial_smooth() +
  xlab('Maximum Temperature') +
  ylab('Rain Factor-1') +
  ggtitle('Predictors with the Response Variable')
```



Looking at the predictive model for TMAX, for colder temperatures below 60°, it is more likely to precipitate. As TMAX increases, our curve drops, and above 70° there it is very unlikely to precipitate.

Conclusions

At the beginning of this report, there were a few questions I wanted to attempt to answer from the data analysis. Revisiting these questions:

When is the hottest time of the year, and does it often rain during that time?

  - From the data, the hottest times of year in San Francisco are from July to September, and it often doesn't rain during that time.

- When is the coldest time of year, and does it often rain during that time?
  - The coldest months of the year in San Francisco are from December to February and it is most likely to rain during this time.
- What are the precipitation levels throughout the year?
  - It doesn't rain often in San Francisco, and when it does, it is often not much accumulation. It is most likely to precipitate during the winter and spring seasons. I can attest to this anecdotally, because growing up, I could walk or ride my bike to school all year round without too much fear of getting caught in the rain. It also wasn't very cold during the winters, as evident from the TMIN and TMAX distributions during the winter, so I didn't own a winter coat until I went to college in the Northeast.
- What conclusions can be made by looking at the temperature and precipitation data?
  - Only 17.45% of days between 1970-2020 had measured precipitation from 1970-2020, so San Francisco has great weather for someone who is looking for a place where it doesn't precipitate much. While it may precipitate less in a desert, an advantage of San Francisco is its relatively mild winter and summers. Lastly, Mark Twain was right, don't expect hot summers in San Francisco, yet a light jacket may be all you need for all seasons.