Precipitation Distributions

Set-Up

First, we'll load the packages necesary for the analysis:

```
library(tidyverse)
library(plotly)
```

Read the data into a dataframe and check that the dataframe was constructed properly:

```
weather <- read.csv("weather_df.csv")
head(weather)
```

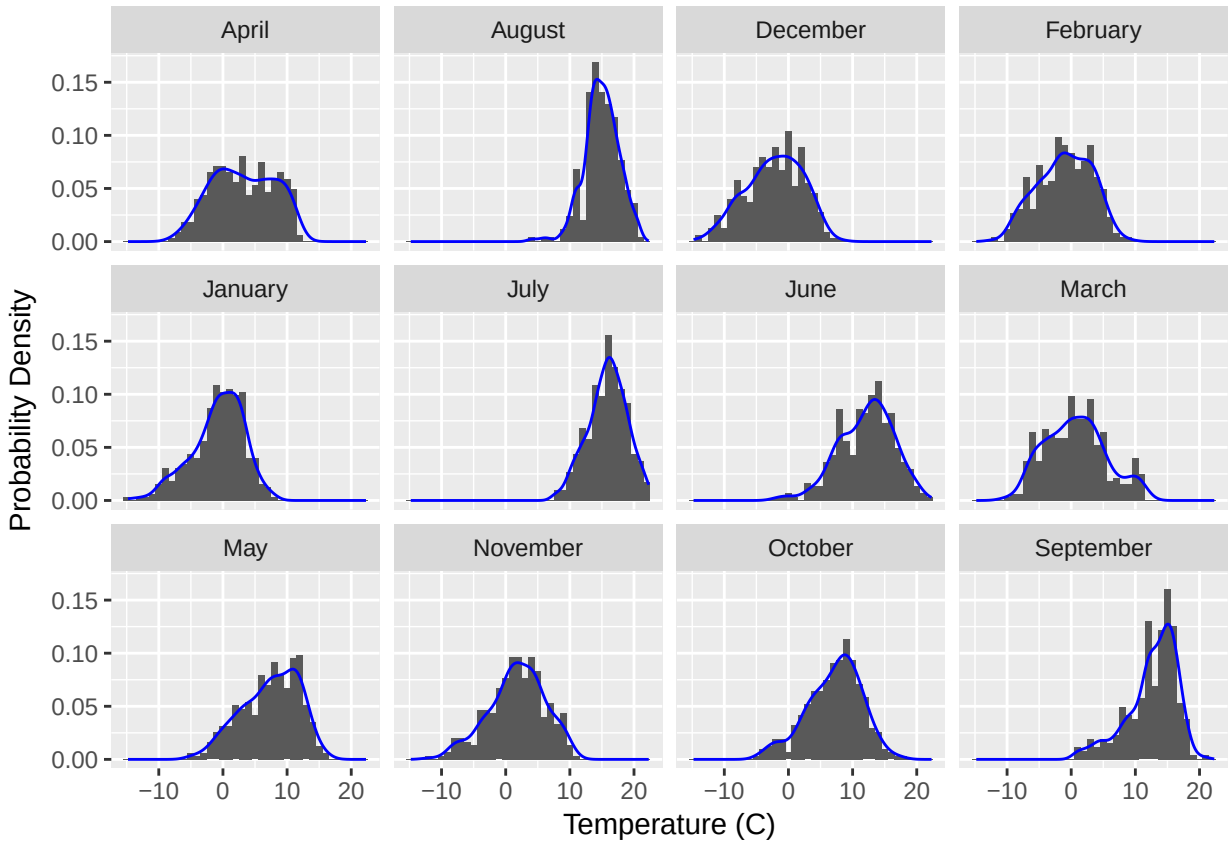
```
##      Date Year Day_of_Year Day_of_Run Solar_Rad Wind_Speed Wind_Direction
## 1 12/1/05 2005         335          1     0.097       2.99              2
## 2 12/2/05 2005         336          2     0.841       2.02            295
## 3 12/3/05 2005         337          3     2.794       2.22            126
## 4 12/4/05 2005         338          4     2.295       0.82             90
## 5 12/5/05 2005         339          5     2.322       0.46            148
## 6 12/6/05 2005         340          6     2.230       0.55            198
##   Wind_Gust Tavg Tmax  Tmin Havg Hmax Hmin Pressure Snow_Depth
## 1     19.89  1.9  3.0  -0.1   98   99   95      796      244.8
## 2     13.88 -4.7 -0.2  -7.3   90   98   79      799      250.8
## 3     18.19 -6.5 -3.5 -10.0   72   94   51      806      453.6
## 4     13.00 -5.9  0.6 -10.7   68   88   49      811      253.2
## 5      6.47 -3.1  3.7  -7.4   51   78   25      810      253.6
## 6      6.92 -2.1  4.9  -7.0   45   63   21      808      251.2
##   Total_Precip    Month
## 1        99.06 December
## 2         3.05 December
## 3         1.52 December
## 4         0.00 December
## 5         0.25 December
## 6         0.00 December
```

Single Variable Distributions

At this point we have the data properly formatted into a dataframe and we can start looking at how the variables are distributed. Understanding the range of values and the probability associated with each value provides a better understanding of the weather's behavoir at the weather station. For example, from viewing the monthly distributions in the previous analysis we have clear signals that March is the most predictably wet month of the calander, but that December can have higher precipitation totals with much lower predictability.

Let's start by looking at the temperature distribution separated by month:

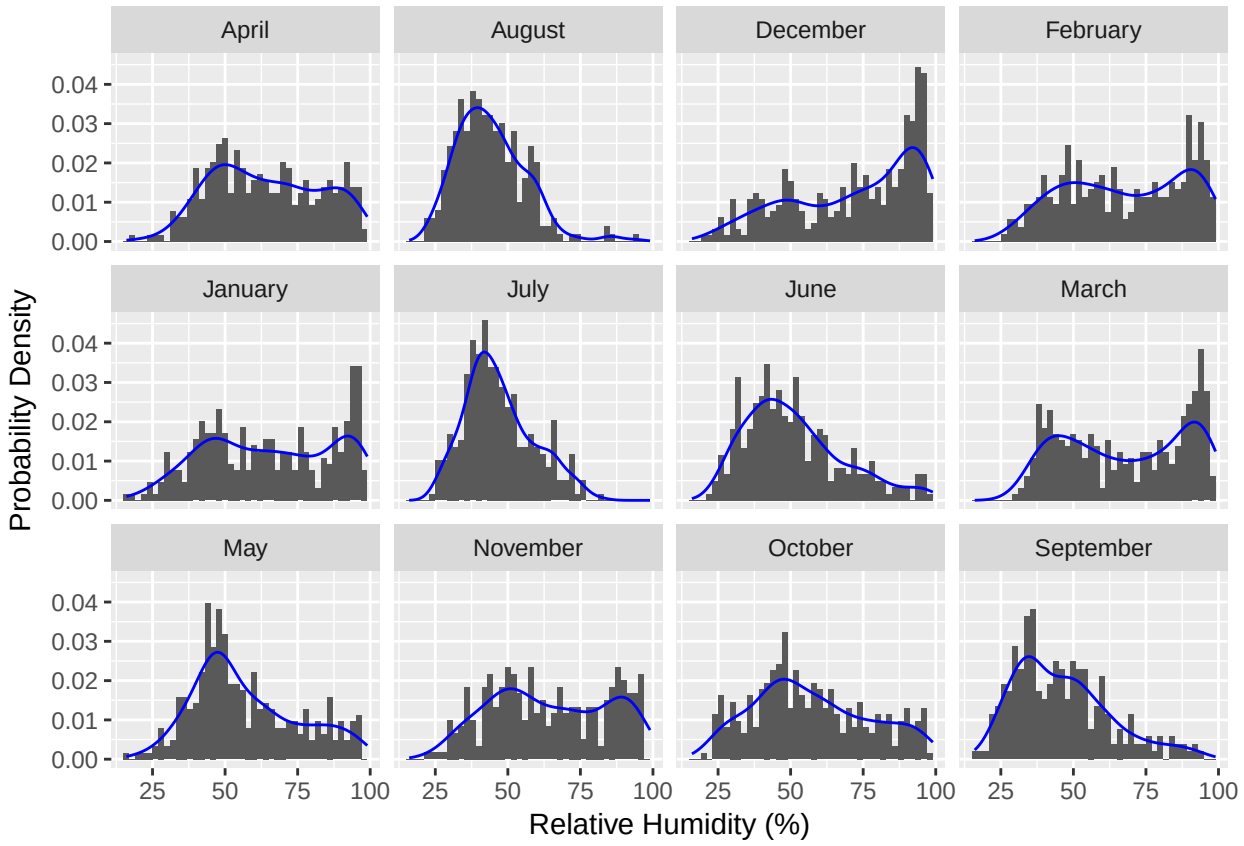
```
ggplot(data = weather) +
  geom_histogram(aes(x = Tavg, y = ..density..), binwidth = 1) +
  facet_wrap(~Month) +
  geom_line(aes(x = Tavg, y = ..density..), stat = 'density', color = 'blue') +
  xlab("Temperature (C)") +
  ylab("Probability Density")
```



The temperature distributions in each month show an approximately normal distribution centered around means that warm and cool depending on the month. April has an interesting distribution; it appears slightly biphasic. Potentially, these distributions indicate that in this transitional period it is likely to be either cold or warm with few days in between.

Lets take a look at how the humidity distribution changes by month:

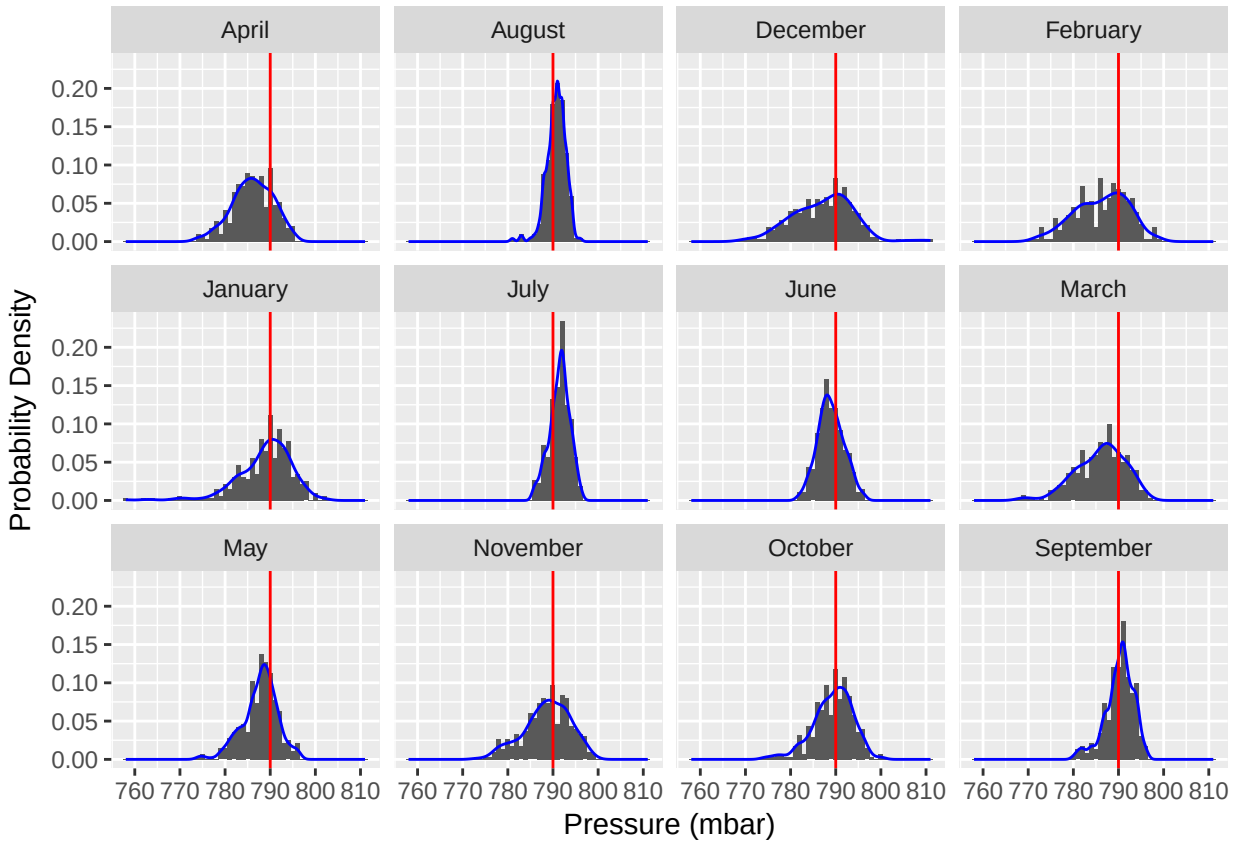
```
ggplot(data = weather) +
  geom_histogram(aes(x = Havg, y = ..density..), binwidth = 2) +
  facet_wrap(~Month) +
  geom_line(aes(x = Havg, y = ..density..), stat = 'density', color = 'blue') +
  xlab("Relative Humidity (%)") +
  ylab("Probability Density")
```



It is evident that the wetter months have a much greater range of humidity. As the weather becomes warmer and drier in the summer months, the variance decreases and the distribtion approaches a normal distribution.

Now lets see how the pressure distributions vary:

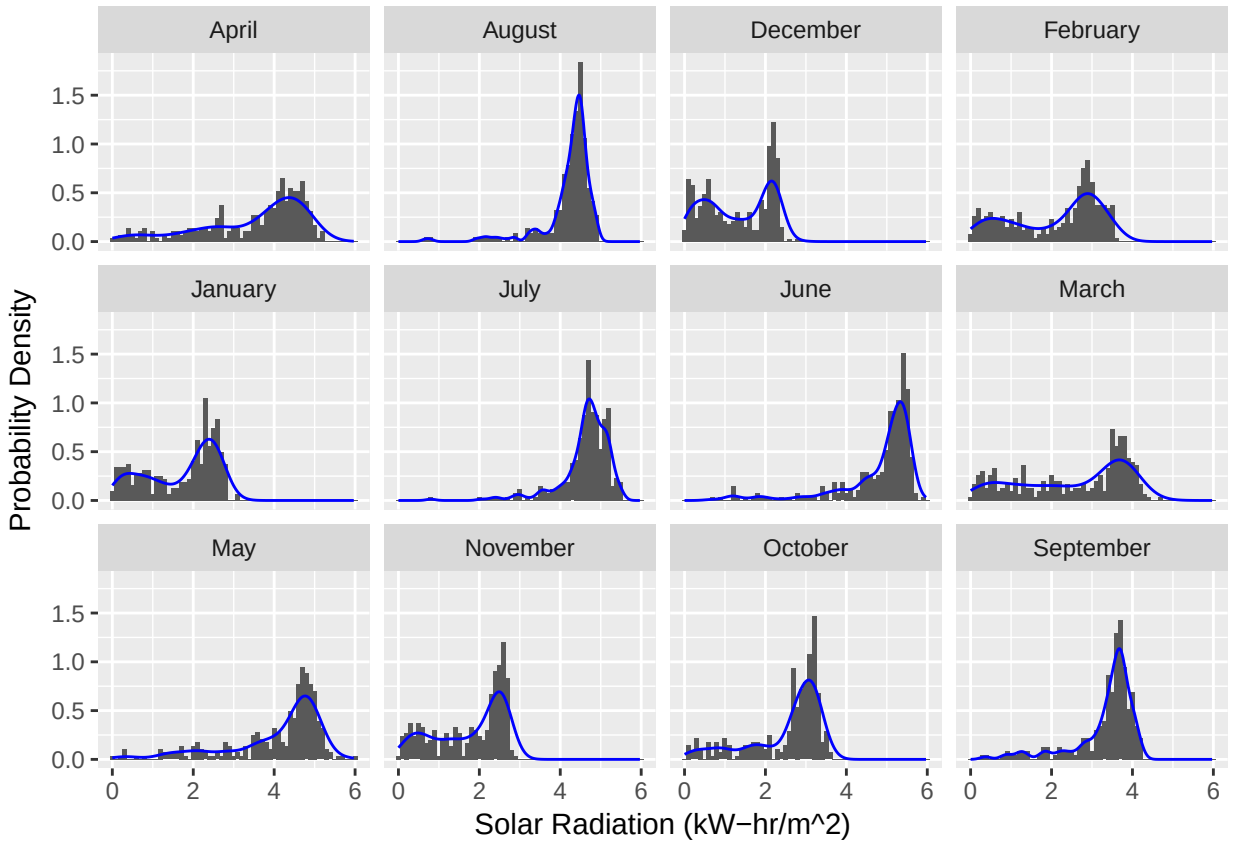
```
ggplot(data = weather) +
  geom_histogram(aes(x = Pressure, y = ..density..), binwidth = 1) +
  facet_wrap(~Month) +
  geom_line(aes(x = Pressure, y = ..density..), stat = 'density', color = 'blue') +
  geom_vline(aes(xintercept = 790), color = 'red') +
   xlab("Pressure (mbar)") +
  ylab("Probability Density")
```



The red vertical lines provide a reference at 790 mbar. These distributions show that, similar to humidity, the wetter months have much more variable pressure ranges. As summer sets in, the pressure distribution becomes much tighter and may increase slightly (especially in July, August, and September). Of course we would have to verify these claims with more rigorous hypothesis testing.

Solar Radiation:

```
ggplot(data = weather) +
  geom_histogram(aes(x = Solar_Rad, y = ..density..), binwidth = .1) +
  facet_wrap(~Month) +
  geom_line(aes(x = Solar_Rad, y = ..density..), stat = 'density', color = 'blue') +
   xlab("Solar Radiation (kW-hr/m^2)") +
  ylab("Probability Density")
```



Clearly, the solar radiation is much greater in summer than in winter, spring, and fall.

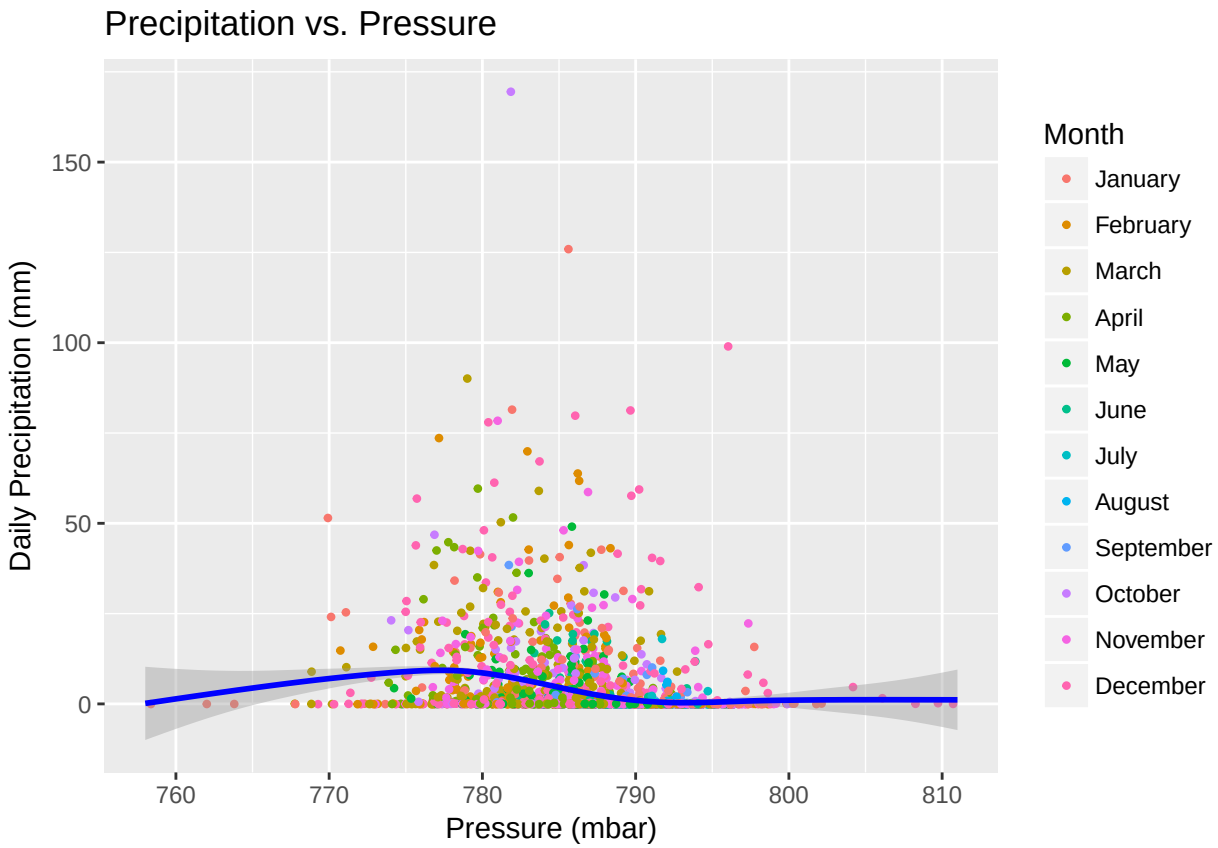
After viewing each of these different distributions we have a decent idea about how each of these factors vary throughout the year. Now, we can begin probing the relationships between each weather factor and how much precipitation actually falls. The goal with this single variable analysis is to observe how each weather factor affects the occurence of precipitation events.

Single Variable Relationships to Precipitation

Let's take a look at pressure and total daily precipitation:

```
# Reorder the months in the Months factor for more sensible plotting

weather$Month = factor(weather$Month, levels(weather$Month)[c(5, 4, 8, 1, 9,
                                                              7, 6, 2, 12, 11,
                                                              10, 3)])
```

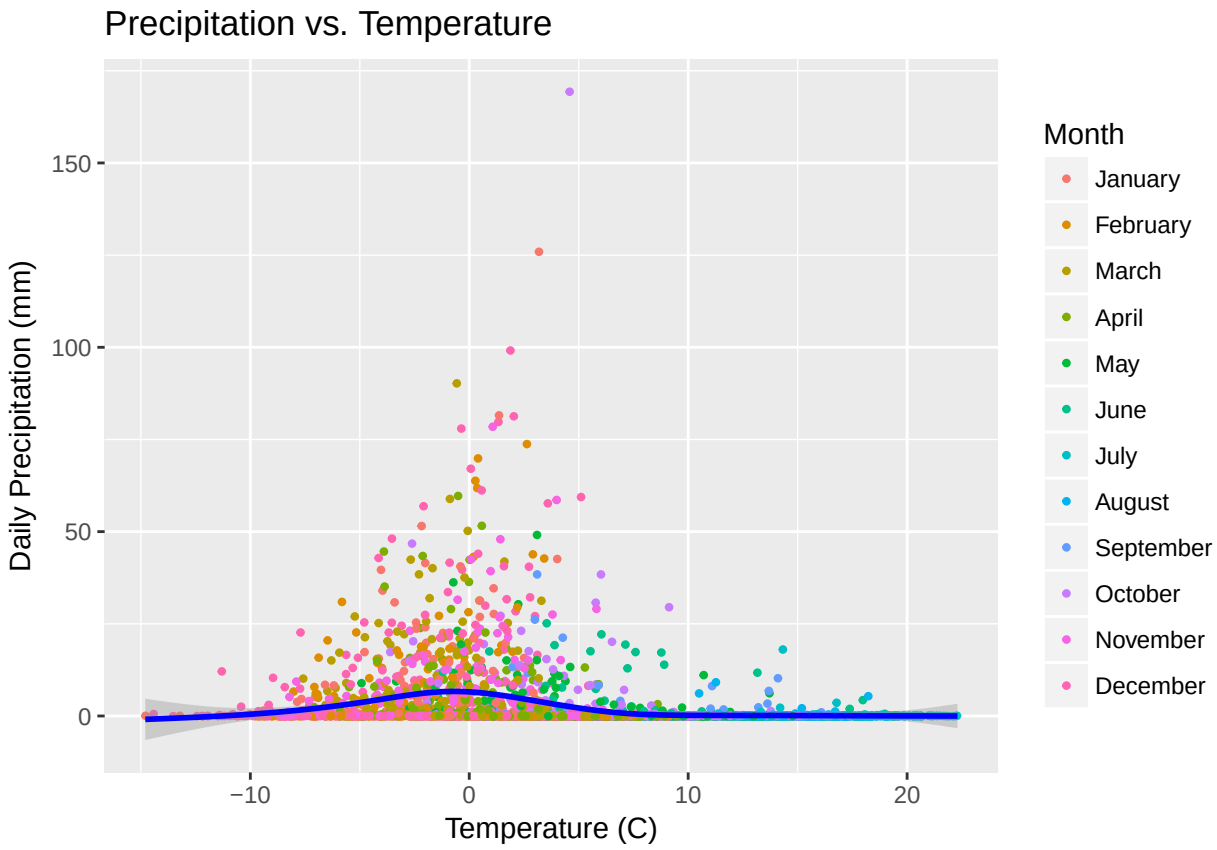
```
# Plot the pressure vs precipitation

ggplot(data = weather,mapping = aes(x = Pressure, y = Total_Precip)) +
  geom_jitter(mapping = aes(color = Month), size = .85) +
  geom_smooth(color = "blue") +
  ylab("Daily Precipitation (mm)") +
  xlab("Pressure (mbar)") +
  ggtitle("Precipitation vs. Pressure")
```



It appears that precipitaion events occur more often when pressure is in the lower portion of the range. This makes sense because unsettled weather is associated with low pressure weather systems.

Now let's take a look at precipitation vs temperature:

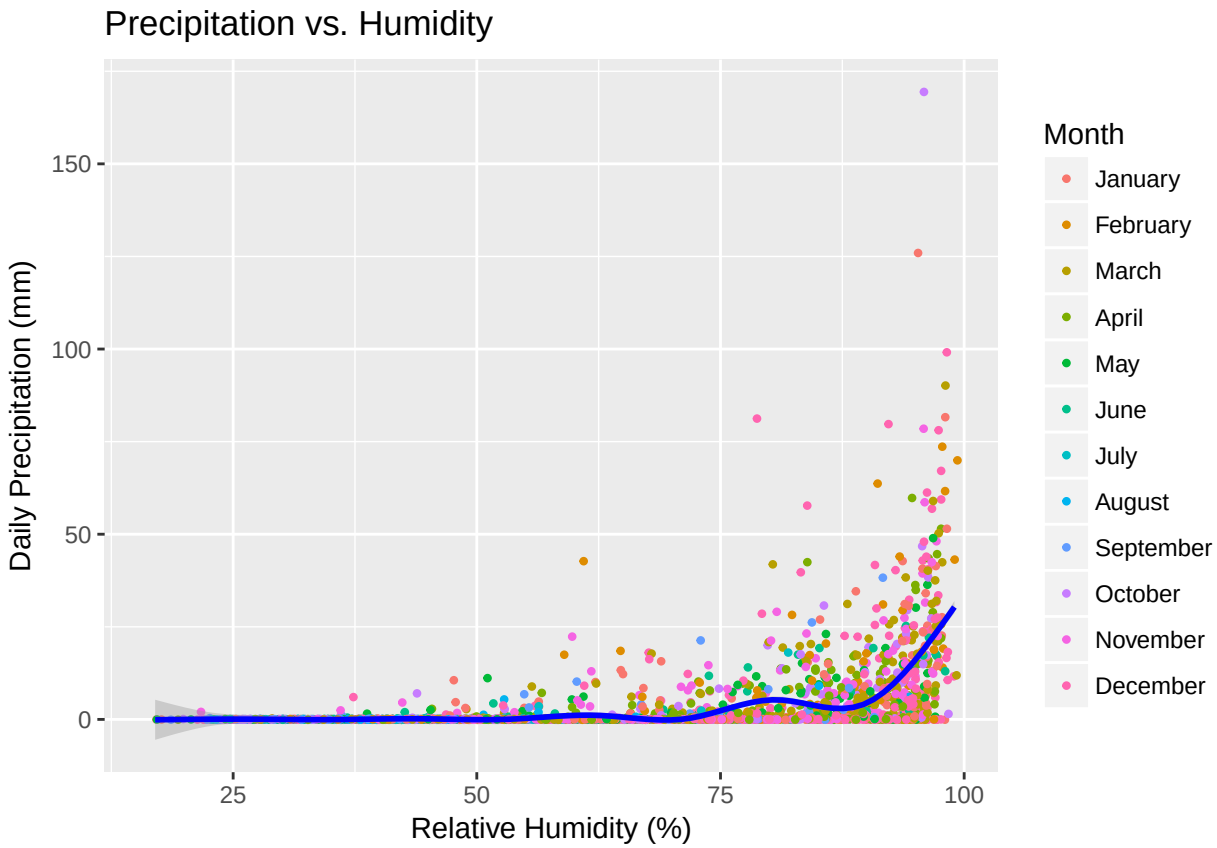
```
ggplot(data = weather,mapping = aes(x = Tavg, y = Total_Precip)) +
  geom_jitter(mapping = aes(color = Month), size = .85) +
  geom_smooth(color = "blue") +
  ylab("Daily Precipitation (mm)") +
  xlab("Temperature (C)") +
  ggtitle("Precipitation vs. Temperature")
```



Interestingly, most of the precipitation events seem to occur closer to 0 degrees celcius. Indeed, the largest precipitation events in the data set occur quite close to 0.

How about precipitation vs humidity?

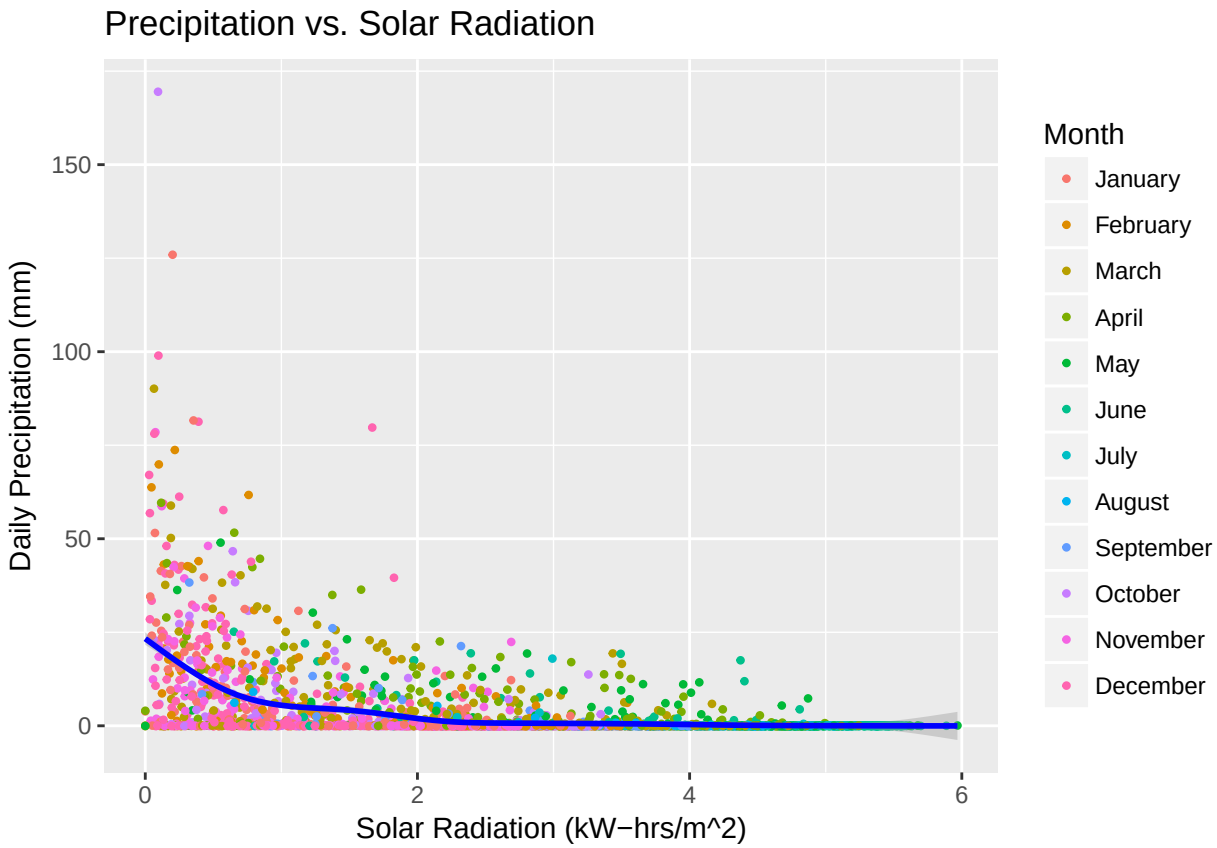
```
ggplot(data = weather,mapping = aes(x = Havg, y = Total_Precip)) +
  geom_jitter(mapping = aes(color = Month), size = .85) +
  geom_smooth(color = "blue") +
  ylab("Daily Precipitation (mm)") +
  xlab("Relative Humidity (%)") +
  ggtitle("Precipitation vs. Humidity")
```



This one is a bit obvious since humidity measures the water content in the air. Clearly, high humidity is necessary for precipitation to occur.

Precipitation vs Solar Radiation?

```
ggplot(data = weather,mapping = aes(x = Solar_Rad, y = Total_Precip)) +
  geom_jitter(mapping = aes(color = Month), size = .85) +
  geom_smooth(color = "blue") +
  ylab("Daily Precipitation (mm)") +
  xlab("Solar Radiation (kW-hrs/m^2)") +
  ggtitle("Precipitation vs. Solar Radiation")
```



Another obvious one. We would expect solar radiation to be low on days when preciptation occurs due to cloud cover.

This analysis provides a pretty clear picture of what factors combine to create precipitation events. The ideal day for large precipitation events is one in which the daily average temperature is close to 0 degrees celsius, the pressure is in the lower range (< 790 mbar), there is high humidity, and solar radiation is low. This simply describes a cold and stormy day! Now lets see if we can observe any interesting higher level patterns with a multivaraible approach.

Multivariable Relations to Precipitation

To start, lets see how temperature, pressure and precipitation interact:

```
plot_ly(weather, x = ~Tavg,
                 y = ~Pressure,
                 z = ~Total_Precip,
                 type = "scatter3d",
                 color = ~Havg,
                 mode = "markers",
                 marker = list(line = list(color = "black", width = .75)))
```

This plot clearly corroborates our conclusion in the single variable analysis that both lower pressure and temperature close to zero are favorable conditions for precipitation.

We'll let the rest of the distributions speak for themselves...

```
plot_ly(weather, x = ~Havg,
                 y = ~Pressure,
```

```
                  z = ~Total_Precip,
                  type = "scatter3d",
                  color = ~Month)
```

```
plot_ly(weather, x = ~Tavg,
                  y = ~Havg,
                  z = ~Total_Precip,
                  type = "scatter3d",
                  color = ~Month)
```

```
plot_ly(weather, x = ~Solar_Rad,
                  y = ~Pressure,
                  z = ~Total_Precip,
                  type = "scatter3d",
                  color = ~Month)
```

```
plot_ly(weather, x = ~Tavg,
                  y = ~Solar_Rad,
                  z = ~Total_Precip,
                  type = "scatter3d",
                  color = ~Month)
```

Conclusions

Anlysis of the distribtions of each factor have displayed what we might have been able to come up with intuitively; that precipitation events occur on cold and stormy days when there is consistent cloud cover and high atmospheric water content. Some more interesting results of the analysis are that the wetter months are often associated with larger variances in weather patterns, the largest precipitation events in the region occur when the temperature is close to 0 degrees celcius, and that the month of April shows a biphasic temperature distribution.