Exam 2

Devon Rudd

## Introduction

This analysis is a personal project to determine a good time to hike around Colfax California. Weather data is sourced from [NOAA](https://www.ncdc.noaa.gov/ghcn-daily-description) using rnoaa package and air quality data is sourced public sources with the ropenaq package.

load('airquality.RDa')  
load('weather.RDa')  
  
## functions  
temp\_convert<- function(x){ (x/10)\*(9/5) + 32}  
## recommeded packages  
library(tidyr)  
library(dplyr)  
library(ggplot2)  
library(gridExtra)

### 1. Explore maximum temperature

* Use temperature function to create columns for max (tmax), min (tmin) and observed (tobs) temperature in Fahrenheit.

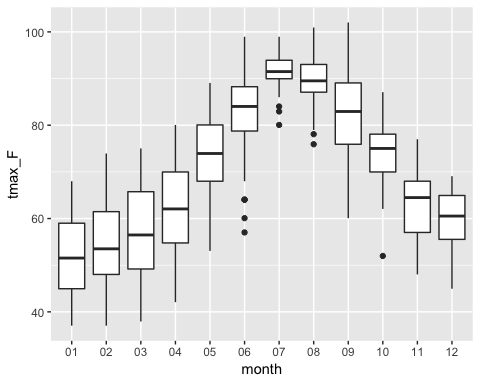
df <- weather\_Colfax  
df$tmax\_F <- sapply(df$tmax, temp\_convert)  
df$tmin\_F <- sapply(df$tmin, temp\_convert)  
df$tobs\_F <- sapply(df$tobs, temp\_convert)

* Create a factor() variable for months()

df$month <- format(df$date, "%m")  
months <- factor(df$month)

* Create a boxpot to view the maximum temperature as a function of the month.

ggplot(df, aes(x=month, y=tmax\_F))+ geom\_boxplot(data = df)



* Based on the box plot,
  + What months have a smaller temperature variation? **Based on the boxplot, it appears that July and August have the smallest teperature variation.**
  + If ideal hiking weather has a max temperature between 60 and 80, how many months are NOT ideal for hiking? Describe the exclusion criteria use. **The best months two go hiking would be in May and October as the max box value for them does not get above 80 degrees and lowest does not get below 60 degrees.**

### 2. Explore air quality

* Tidy up the air quality data by filtering for the particles parameter (pm25) and summarizing the data by day with the max of pm25.

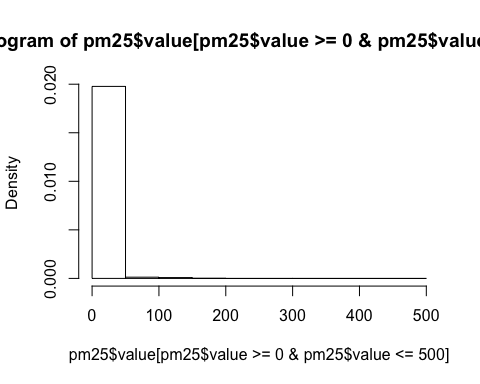
df\_aq <- airquality\_Colfax  
  
pm25 <- df\_aq %>%  
 filter(parameter == "pm25") %>%  
 filter(value >= 0)

* Enrich that data with a factor variable for air quality category based on

|  |  |
| --- | --- |
| AQI Category | Particles (PM2.5) |
| Good | 0 - 50 |
| Moderate | 51 - 100 |
| Unhealthy for Sensitive Groups | 101 - 150 |
| Unhealthy | 151 - 200 |
| Very Unhealthy | 201 - 300 |
| Hazardous | 301 - 500 |

* Create a histogram for the max air quality.

histogram <- hist(pm25$value[pm25$value >= 0 & pm25$value <= 500], breaks=c(0, 50, 100, 150, 200, 300, 500))

 \* Create bar chart for AQI category.

bar <- ggplot(pm25, aes(value)) + geom\_bar()

* Use grid.arrange(plot1, plot2, ncol=2) to plot side by side

# Did not work  
# grid.arrange(histogram, bar, ncol=2)

* Compare the histogram and bar chart:
  + What is does each graph best communicate? **The histogram best communicates the values that fall within the AQI definitions (“Good”, “Unhealthy”, etc.) and the bar chart better communicates the exact values that were measured. The bar chart would be better for scientific studies while the histogram would be better for the medical field researching exposure to the airquality levels.**
  + How could you combine into one graph? **You could use the rug display to show where the concentration lies within the histgram underneath the histogram.**

### 3. Trend weather and air quality by date

* Merge or join the summarized air quality data with the weather data

# date\_conv <- function(x) {  
# f <- as.Date(as.POSIXct(x, tz="PST"))  
# return(f)  
# }  
  
names(weather\_Colfax)[names(weather\_Colfax) == "date"] <- "dateLocal"  
airquality\_Colfax$dateLocal <- as.Date(airquality\_Colfax$dateLocal)  
# sapply(airquality\_Colfax$dateLocal, date\_conv)  
# merged <- inner\_join(pm25, weather\_Colfax, by="dateLocal")

* Plot the max daily tempurature and air quality by day. Plot either side by side or layered on the same graph.
* Evaluate the graph (s)
  + Any problems with the data?
  + What are one or two items this graph is not representing or miscommunicating to inform ideal hiking weather?
* Challenge: Create a final visualization by adding multiple aesthetics, reference lines or annotation to indicate ideal dates for hiking.