R Project

Jacob Rozran

2/28/2017

PROBLEM 1

Read in the data.

```
library(readxl)

we_dat <- read_excel("Philadelphia Temperatures - ALL.xlsx")

summary(we_dat)
```

```
##      Month              Day              Year           High
##  Min.   : 1.000   Min.   : 1.00   Min.   :1872   Min.   :-999.00
##  1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:1904   1st Qu.:  47.00
##  Median : 7.000   Median :16.00   Median :1936   Median :  64.00
##  Mean   : 6.514   Mean   :15.76   Mean   :1936   Mean   :  44.47
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1967   3rd Qu.:  79.00
##  Max.   :12.000   Max.   :31.00   Max.   :1999   Max.   : 106.00
##       Low              Precip              Snow
##  Min.   :-999.0   Min.   :-999.000   Min.   :-999.0
##  1st Qu.:  33.0   1st Qu.:   0.000   1st Qu.:   0.0
##  Median :  46.0   Median :   0.000   Median :   0.0
##  Mean   :  27.9   Mean   :   7.087   Mean   :-100.9
##  3rd Qu.:  61.0   3rd Qu.:   4.000   3rd Qu.:   0.0
##  Max.   :  82.0   Max.   : 663.000   Max.   : 276.0
```

PROBLEM 2

Data manipulation.

PART A

Change values of -999 to NA, as they were not collected. From the summary above, we can see that High, Low, Precip, and Snow have values of -999; the other variables/columns do not.

```
we_dat$High[we_dat$High == -999] <- NA
we_dat$Low[we_dat$Low == -999] <- NA
we_dat$Precip[we_dat$Precip == -999] <- NA
we_dat$Snow[we_dat$Snow == -999] <- NA

summary(we_dat)
```

```
##      Month             Day             Year           High
##  Min.   : 1.000   Min.   : 1.00   Min.   :1872   Min.   :  5.00
##  1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:1904   1st Qu.: 48.00
##  Median : 7.000   Median :16.00   Median :1936   Median : 65.00
##  Mean   : 6.514   Mean   :15.76   Mean   :1936   Mean   : 63.22
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1967   3rd Qu.: 80.00
##  Max.   :12.000   Max.   :31.00   Max.   :1999   Max.   :106.00
##                                                  NA's   :827
##       Low              Precip            Snow
##  Min.   :-11.00   Min.   : -1.00   Min.   : -1.000
##  1st Qu.: 33.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median : 46.00   Median :  0.00   Median :  0.000
##  Mean   : 46.35   Mean   : 11.14   Mean   :  0.556
##  3rd Qu.: 61.00   3rd Qu.:  4.00   3rd Qu.:  0.000
##  Max.   : 82.00   Max.   :663.00   Max.   :276.000
##  NA's   :827      NA's   :188      NA's   :4754
```

PART B

As the only value less than 0 is -1, indicating a trace amount of snow or precipitation, I am changing it to a 1. This indicates a tenth inch of snow and hundredth inch of precipitation. That seems like a “trace” to me.

```
we_dat$Precip[we_dat$Precip < 0] <- 1
we_dat$Snow[we_dat$Snow < 0] <- 1

summary(we_dat)
```

```
##      Month             Day             Year           High
##  Min.   : 1.000   Min.   : 1.00   Min.   :1872   Min.   :  5.00
##  1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:1904   1st Qu.: 48.00
##  Median : 7.000   Median :16.00   Median :1936   Median : 65.00
##  Mean   : 6.514   Mean   :15.76   Mean   :1936   Mean   : 63.22
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1967   3rd Qu.: 80.00
##  Max.   :12.000   Max.   :31.00   Max.   :1999   Max.   :106.00
##                                                  NA's   :827
##       Low              Precip           Snow
##  Min.   :-11.00   Min.   :  0.0   Min.   :  0.000
##  1st Qu.: 33.00   1st Qu.:  0.0   1st Qu.:  0.000
##  Median : 46.00   Median :  0.0   Median :  0.000
##  Mean   : 46.35   Mean   : 11.4   Mean   :  0.647
##  3rd Qu.: 61.00   3rd Qu.:  4.0   3rd Qu.:  0.000
##  Max.   : 82.00   Max.   :663.0   Max.   :276.000
##  NA's   :827      NA's   :188     NA's   :4754
```

PART C

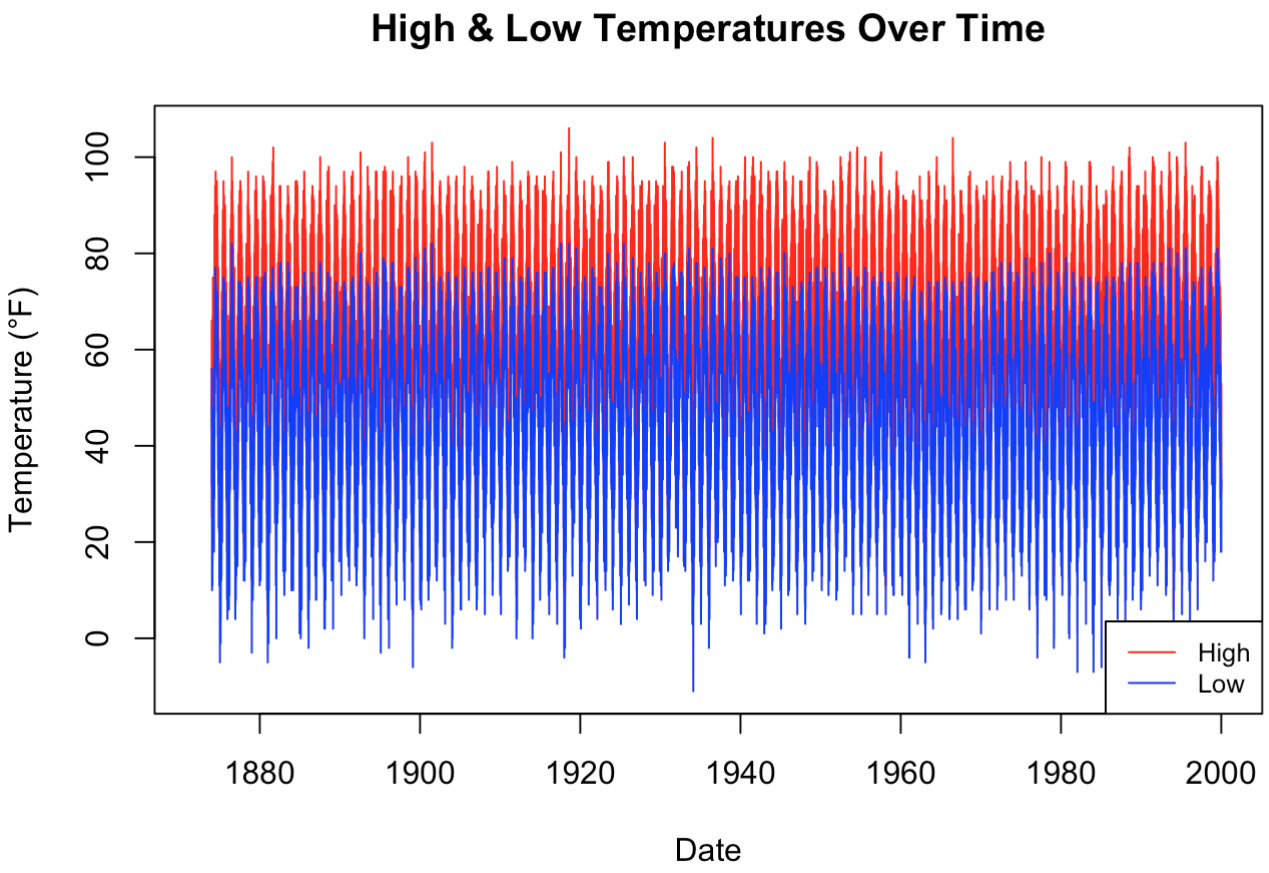
```
we_dat$Precip_inches <- we_dat$Precip / 100
we_dat$Snow_inches <- we_dat$Snow / 10

summary(we_dat)
```

```
##      Month              Day             Year           High      
##  Min.   : 1.000   Min.   : 1.00   Min.   :1872   Min.   :  5.00  
##  1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:1904   1st Qu.: 48.00  
##  Median : 7.000   Median :16.00   Median :1936   Median : 65.00  
##  Mean   : 6.514   Mean   :15.76   Mean   :1936   Mean   : 63.22  
##  3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1967   3rd Qu.: 80.00  
##  Max.   :12.000   Max.   :31.00   Max.   :1999   Max.   :106.00  
##                                                  NA's   :827     
##       Low             Precip           Snow          Precip_inches    
##  Min.   :-11.00   Min.   :  0.0   Min.   :  0.000   Min.   :0.000  
##  1st Qu.: 33.00   1st Qu.:  0.0   1st Qu.:  0.000   1st Qu.:0.000  
##  Median : 46.00   Median :  0.0   Median :  0.000   Median :0.000  
##  Mean   : 46.35   Mean   : 11.4   Mean   :  0.647   Mean   :0.114  
##  3rd Qu.: 61.00   3rd Qu.:  4.0   3rd Qu.:  0.000   3rd Qu.:0.040  
##  Max.   : 82.00   Max.   :663.0   Max.   :276.000   Max.   :6.630  
##  NA's   :827      NA's   :188     NA's   :4754      NA's   :188    
##   Snow_inches     
##  Min.   : 0.000  
##  1st Qu.: 0.000  
##  Median : 0.000  
##  Mean   : 0.065  
##  3rd Qu.: 0.000  
##  Max.   :27.600  
##  NA's   :4754
```

PROBLEM 3

Plot the daily highs and lows on the same plot.

```
we_dat$Date <- as.Date(paste(we_dat$Year, "-", we_dat$Month, "-", we_dat$Day,
                             sep = ""))

plot(x = we_dat$Date, y = we_dat$High, type = "l", col = "red",
     ylim = c(min(we_dat$Low, na.rm = TRUE), max(we_dat$High, na.rm = TRUE)),
     main = "High & Low Temperatures Over Time", xlab = "Date",
     ylab = "Temperature (°F)")
points(x = we_dat$Date, y = we_dat$Low, type = "l", col = "blue")
legend("bottomright", lty = c(1, 1), col = c("red", "blue"),
       legend = c("High", "Low"), cex = 0.8)
```



There is not much you can tell in terms of temperature trend from this graph. There seems to be a few periods of higher low temperatures in the 1930's. There also seems to be a recent trend of higher low temperatures in the 1980's and 1990's. High temperatures seem more or less constant over time.

PROBLEM 4

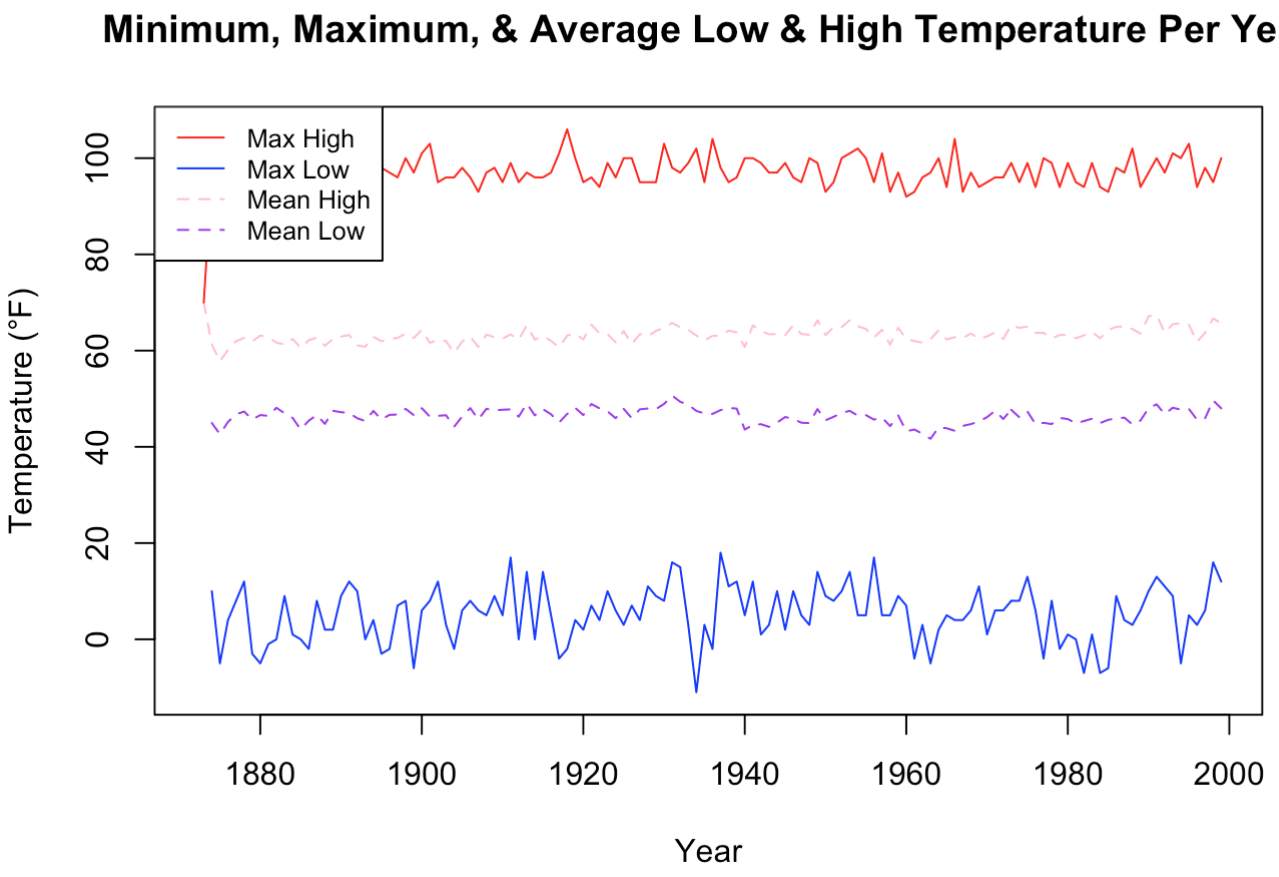
Plot the yearly minimum, maximum, and average low and high temperature per year.

```
library(plyr)

we_dat_year <- ddply(we_dat, .(Year), summarize, min_low = min(Low, na.rm = TRUE),
                     mean_low = mean(Low, na.rm = TRUE),
                     max_high = max(High, na.rm = TRUE),
                     mean_high = mean(High, na.rm = TRUE))
```

```
## Warning in max(High, na.rm = TRUE): no non-missing arguments to max;
## returning -Inf
```

```
## Warning in min(Low, na.rm = TRUE): no non-missing arguments to min;
## returning Inf
```

```
we_dat_year$min_low[is.nan(we_dat_year$min_low) | is.infinite(we_dat_year$min_low)] <- N
A
we_dat_year$mean_low[is.nan(we_dat_year$mean_low) | is.infinite(we_dat_year$mean_low)] <
- NA
we_dat_year$mean_high[is.nan(we_dat_year$mean_high) |
is.infinite(we_dat_year$mean_high)] <- NA
we_dat_year$max_high[is.nan(we_dat_year$max_high) | is.infinite(we_dat_year$max_high)] <
- NA

plot(x = we_dat_year$Year, y = we_dat_year$min_low, type = "l", col = "blue",
     ylim = c(min(we_dat_year$min_low, na.rm = TRUE), max(we_dat_year$max_high, na.rm =
TRUE)),
     main = "Minimum, Maximum, & Average Low & High Temperature Per Year",
     ylab = "Temperature (°F)", xlab = "Year")
points(x = we_dat_year$Year, y = we_dat_year$mean_low, type = "l", col = "purple", lty =
 2)
points(x = we_dat_year$Year, y = we_dat_year$mean_high, type = "l", col = "pink", lty =
2)
points(x = we_dat_year$Year, y = we_dat_year$max_high, type = "l", col = "red", lty = 1)
legend("topleft", lty = c(1, 1, 2, 2), col = c("red", "blue", "pink", "purple"),
       legend = c("Max High", "Max Low", "Mean High", "Mean Low"), cex = 0.8)
```



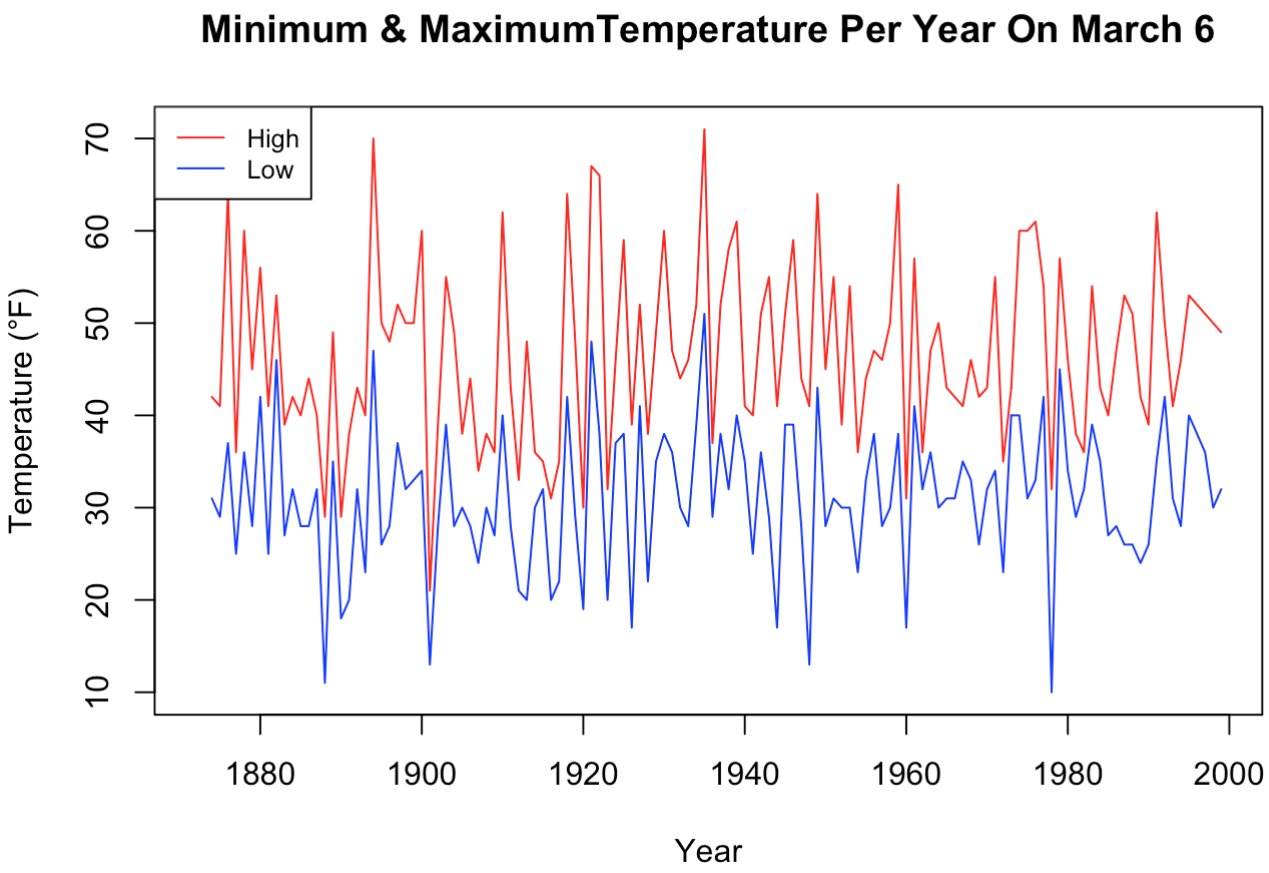
PROBLEM 5

Analysis on the data from March 6th.

PART A

Plot the min and max temp per year on the same plot.

```
we_dat_0306 <- we_dat[we_dat$Month == 3 & we_dat$Day == 6, ]

plot(x = we_dat_0306$Year, y = we_dat_0306$Low, type = "l", col = "blue",
     ylim = c(min(we_dat_0306$Low, na.rm = TRUE), max(we_dat_0306$High, na.rm = TRUE)),
     main = "Minimum & MaximumTemperature Per Year On March 6",
     ylab = "Temperature (°F)", xlab = "Year")
points(x = we_dat_0306$Year, y = we_dat_0306$High, type = "l", col = "red")
legend("topleft", lty = c(1, 1), col = c("red", "blue"), legend = c("High", "Low"), cex
= 0.8)
```

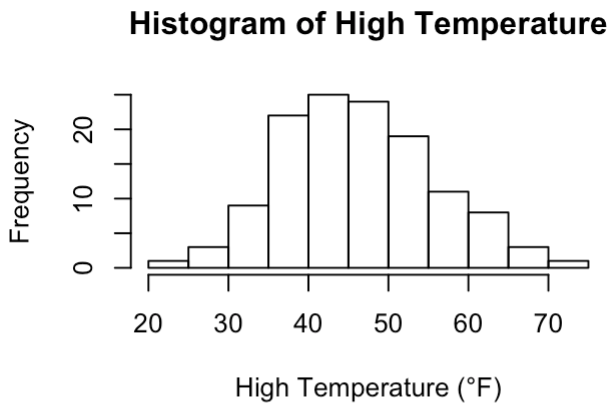


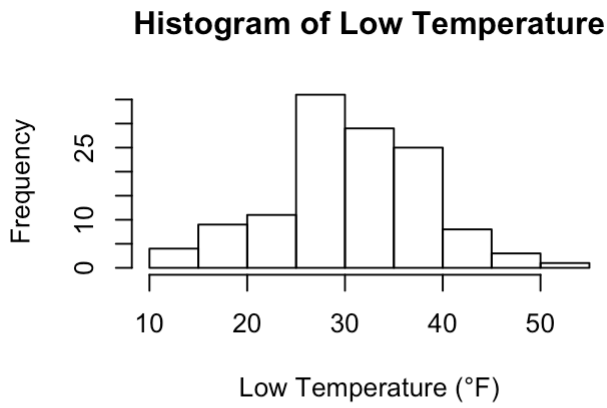
The high and low temperature on March 6th is much more volitile than that of the high and low temperature for the entire year. It is possible to get freezing, winter temperatures or moderate, spring temperatures during this time of the year. Over the course of the year, however, high and low temperatures tend to be similar from year to year.

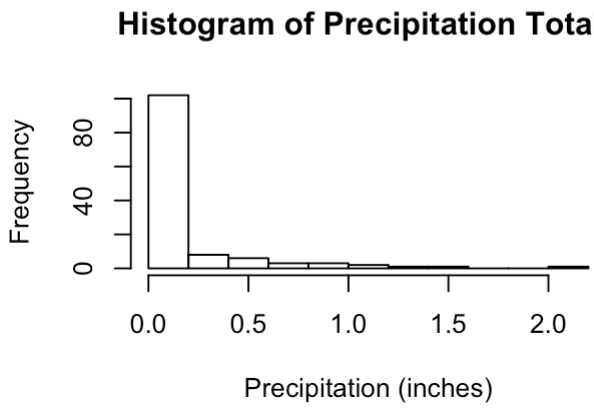
PART B

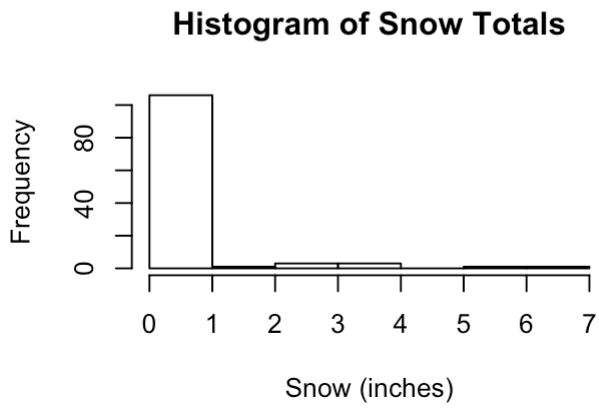
Create histograms of High and Low temperature and Precipitation and Snow totals.

```
par(mfrow = c(2, 2))

hist(we_dat_0306$High, main = "Histogram of High Temperature", xlab = "High Temperature
 (°F)")
hist(we_dat_0306$Low, main = "Histogram of Low Temperature", xlab = "Low Temperature (°
F)")
hist(we_dat_0306$Precip_inches, main = "Histogram of Precipitation Totals", xlab = "Prec
ipitation (inches)")
hist(we_dat_0306$Snow_inches, main = "Histogram of Snow Totals", xlab = "Snow (inches)")
```









The high temperature is almost normally distributed - probably positively skewed, though. The higher-than-the-mean temperatures happen more frequently the lower side. The mean is in the 45-50°F range. There is a range of around 55°F.

The low temperature has a very interesting distribution - there are some lower temperature days, but at around 25°F, there is a large step up. The average is around 30°F and there is less of a range of temperatures than the highs: about 45°F.

Most days have no precipitation or snow. Rain (preciptiation) happens more often than snow, though. When it does snow, it is up to 7 inches. When it rains, it is up to 2 inches.

PART C

Create a 90% Confidence Interval for mean using a t-distribution.

90% CI: $\bar{x} \pm t_{\alpha/2,N-1}(s / \text{sqrt}(N))$

```
upper_high <- (mean(we_dat_0306$High, na.rm = TRUE) +
                       qt(0.05,
                          df = length(we_dat_0306$High[!is.na(we_dat_0306$High)]) - 1,
                          lower.tail = FALSE) *
                       (sd(we_dat_0306$High, na.rm = TRUE) /

sqrt(length(we_dat_0306$High[!is.na(we_dat_0306$High)])))
lower_high <- (mean(we_dat_0306$High, na.rm = TRUE) -
                       qt(0.05,
                          df = length(we_dat_0306$High[!is.na(we_dat_0306$High)]) - 1,
                          lower.tail = FALSE) *
                       (sd(we_dat_0306$High, na.rm = TRUE) /

sqrt(length(we_dat_0306$High[!is.na(we_dat_0306$High)])))

upper_low <- (mean(we_dat_0306$Low, na.rm = TRUE) +
                       qt(0.05,
                          df = length(we_dat_0306$Low[!is.na(we_dat_0306$Low)]) - 1,
                          lower.tail = FALSE) *
                       (sd(we_dat_0306$Low, na.rm = TRUE) /
                                sqrt(length(we_dat_0306$Low[!is.na(we_dat_0306$Low)]))))
lower_low <- (mean(we_dat_0306$Low, na.rm = TRUE) -
                       qt(0.05,
                          df = length(we_dat_0306$Low[!is.na(we_dat_0306$Low)]) - 1,
                          lower.tail = FALSE) *
                       (sd(we_dat_0306$Low, na.rm = TRUE) /
                                sqrt(length(we_dat_0306$Low[!is.na(we_dat_0306$Low)]))))

upper_precip <- (mean(we_dat_0306$Precip_inches, na.rm = TRUE) +
                       qt(0.05,
                          df = length(we_dat_0306$Precip_inches[!is.na(we_dat_0306$Preci
p_inches)]) - 1,
                          lower.tail = FALSE) *
                       (sd(we_dat_0306$Precip_inches, na.rm = TRUE) /

sqrt(length(we_dat_0306$Precip_inches[!is.na(we_dat_0306$Precip_inches)])))
lower_precip <- (mean(we_dat_0306$Precip_inches, na.rm = TRUE) -
                       qt(0.05,
                          df = length(we_dat_0306$Precip_inches[!is.na(we_dat_0306$Preci
p_inches)]) - 1,
                          lower.tail = FALSE) *
                       (sd(we_dat_0306$Precip_inches, na.rm = TRUE) /

sqrt(length(we_dat_0306$Precip_inches[!is.na(we_dat_0306$Precip_inches)])))

upper_snow <- (mean(we_dat_0306$Snow_inches, na.rm = TRUE) +
                       qt(0.05,
                          df = length(we_dat_0306$Snow_inches[!is.na(we_dat_0306$Snow_in
ches)]) - 1,
                          lower.tail = FALSE) *
                       (sd(we_dat_0306$Snow_inches, na.rm = TRUE) /
                                sqrt(length(we_dat_0306$Snow_inches[!is.na(we_dat_0306$S
now_inches)]))))
```

```
lower_snow <- (mean(we_dat_0306$Snow_inches, na.rm = TRUE) -
                      qt(0.05,
                         df = length(we_dat_0306$Snow_inches[!is.na(we_dat_0306$Snow_in
ches)]) - 1,
                         lower.tail = FALSE) *
                      (sd(we_dat_0306$Snow_inches, na.rm = TRUE) /
                              sqrt(length(we_dat_0306$Snow_inches[!is.na(we_dat_0306$S
now_inches)]))))
```

90% CI for High Temp (°F) mean: (45.3209, 48.2029)
90% CI for Low Temp (°F) mean: (29.9831, 32.2708)
90% CI for Snow Total (inches) mean: (0.1566, 0.4886)
90% CI for Precip Total (inches) mean: (0.1024, 0.1984)

PART D

Create a 90% for median using bootstrap sampling.

```
high_samples <- replicate(10000, sample(we_dat_0306$High[!is.na(we_dat_0306$High)],
                                        replace = TRUE))
high_means <- apply(high_samples, 2, median)

low_samples <- replicate(10000, sample(we_dat_0306$Low[!is.na(we_dat_0306$Low)],
                                        replace = TRUE))
low_means <- apply(low_samples, 2, median)

precip_samples <- replicate(10000, sample(we_dat_0306$Precip_inches[!is.na(we_dat_0306$P
recip_inches)],
                                        replace = TRUE))
precip_means <- apply(precip_samples, 2, median)

snow_samples <- replicate(10000, sample(we_dat_0306$Snow_inches[!is.na(we_dat_0306$Snow_
inches)],
                                        replace = TRUE))
snow_means <- apply(snow_samples, 2, median)
```

90% CI for High Temp (°F) median: (44, 48)
90% CI for Low Temp (°F) median: (30, 32)
90% CI for Snow Total (inches) median: (0, 0)
90% CI for Precip Total (inches) median: (0, 0.01)

PART E

The median is best for the highly skewed variables: snow and precipitation totals. On any given day, you'd expect the snow and precipitation totals to be closer to the median (~0), than the mean.

The mean and the median are very similar for the temperature variables as the data is not *that* skewed.