UWDS2-Wk1-Weather-jms206

Jim Stearns, NetID=jms206

13 January 2015

Weather Data

Assignment Overview

"The atmospheric science department at UW has been collecting weather data from a station set up on their rooftop (in Seattle) since 1998. The station records time, temperature, humidity, and similar measurements once every minute. Let's explore this dataset in order to become familiar with the data, sanity check to verify data quality, and finally look for some simple relationship and trends. Use the same data set we used in class (weather_data_2000_2014.csv). If you need to download the dataset again, a download link is available on the Catalyst site."

Dataset Acquisition and Preparation

The course "Catalyst" web page contains a link for the 2000-2014 data on a dropbox location

The dataset is about 60MB, so be patient, and download only if it isn't already available on this system.

```
# Adjust for your environment.
homeDir <- "/Users/jimstearns/GoogleDrive/Learning/Courses/UWPCE-DataScience/UWDS-Repo-JimStearns206

stopifnot(file.exists(homeDir))
setwd(homeDir)

# Sorry, but this URL seems transitory, and it's a dropbox that doesn't support programmatic downloa
# ... But for now (Jan 2015), it works, manually
#urlToWeatherData <- "https://www.dropbox.com/s/5wot0sbhcsy6x99/weather_data_2000_2014.csv.zip?dl=1"
#if (!file.exists(pathToWeatherData)) {
#    download.file(urlToWeatherData, pathToWeatherData, method="curl")
#}
```

```
# Save some time by unzipping externally. R takes forever.
# read.csv can unzip, but it took forever, approaching a hang.
# So unzipped outside read.csv, and use here.
pathToWeatherData <- "weather_data_2000_2014.csv"
stopifnot(file.exists(pathToWeatherData))
if (!exists("wd_df")) {
    wd_df <- read.csv(pathToWeatherData, header=TRUE)
}
```

```
library(data.table)
wd <- data.table(wd_df)

#Header:
#Year,Month,Day,Time,RHum (%),Temp (F),Wind Direct,Speed (knot),Gust (knot),Rain (inch),Radiation (W
wd_names=c('Year', 'Month', 'Day', 'Time', 'RHumPct', 'TempF', 'WindDir', 'WindSpeed', 'WindGust','R
names(wd) <- wd_names
```

```
## Warning in `names<-.data.table`(`*tmp*`, value = c("Year", "Month", "Day",
## : The names(x)<-value syntax copies the whole table. This is due to <- in
## R itself. Please change to setnames(x,old,new) which does not copy and is
## faster. See help('setnames'). You can safely ignore this warning if it is
## inconvenient to change right now. Setting options(warn=2) turns this
## warning into an error, so you can then use traceback() to find and change
## your names<- calls.
```

Data Exploration and Quality Issue(s)

Assignment

a. "Using summary statistics, histograms, boxplots, or other means identify and describe at least one data quality issue in the dataset."

Implementation

```
str(wd)
```

```
## Classes 'data.table' and 'data.frame':   6992901 obs. of  12 variables:
##  $ Year            : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ Month           : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ Day             : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Time        : Factor w/ 86400 levels "00:00:00","00:00:01",..: 25 85 145 205 265 325 385 445 505 565
##  $ RHumPct         : int  42 41 42 42 42 42 43 42 42 43 ...
##  $ TempF           : int  61 61 61 61 60 60 60 60 60 59 ...
##  $ WindDir         : int  321 320 298 322 330 328 320 325 345 359 ...
##  $ WindSpeed       : int  4 7 9 8 11 10 5 9 11 5 ...
##  $ WindGust        : int  6 9 11 10 15 13 7 12 13 9 ...
##  $ RainInches      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ RadiationWattsM2: num  371 376 344 227 202 ...
##  $ PresMbar        : num  NA NA NA NA NA NA NA NA NA NA ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
head(wd)
```

```
##    Year Month Day     Time RHumPct TempF WindDir WindSpeed WindGust
## 1: 2000     4   1 00:00:24      42    61     321         4        6
## 2: 2000     4   1 00:01:24      41    61     320         7        9
## 3: 2000     4   1 00:02:24      42    61     298         9       11
## 4: 2000     4   1 00:03:24      42    61     322         8       10
## 5: 2000     4   1 00:04:24      42    60     330        11       15
## 6: 2000     4   1 00:05:24      42    60     328        10       13
##    RainInches RadiationWattsM2 PresMbar
## 1:          0           370.69       NA
## 2:          0           375.70       NA
## 3:          0           344.48       NA
## 4:          0           226.86       NA
## 5:          0           201.93       NA
## 6:          0           242.24       NA
```

```
summary(wd)
```

```
##       Year          Month            Day              Time
##  Min.   :2000   Min.   : 1.00   Min.   : 1.00   15:00:49:     99
##  1st Qu.:2003   1st Qu.: 4.00   1st Qu.: 8.00   15:30:49:     99
##  Median :2007   Median : 7.00   Median :16.00   15:35:49:     99
##  Mean   :2007   Mean   : 6.56   Mean   :15.72   15:40:49:     99
##  3rd Qu.:2010   3rd Qu.: 9.00   3rd Qu.:23.00   15:45:49:     99
##  Max.   :2014   Max.   :12.00   Max.   :31.00   15:50:49:     99
##                                                 (Other) :6992307
##     RHumPct              TempF              WindDir
##  Min.   :-99999.00   Min.   :-99999.00   Min.   :     0.0
##  1st Qu.:    59.00   1st Qu.:    45.00   1st Qu.:    96.0
##  Median :    72.00   Median :    52.00   Median :   182.0
##  Mean   :    67.87   Mean   :    52.75   Mean   :   177.8
##  3rd Qu.:    81.00   3rd Qu.:    60.00   3rd Qu.:   244.0
##  Max.   :   512.00   Max.   :  9596.00   Max.   :943290.0
##
##    WindSpeed           WindGust        RainInches
##  Min.   :   0.000   Min.   :     0   Min.   :   0.0000
##  1st Qu.:   2.000   1st Qu.:     3   1st Qu.:   0.0000
##  Median :   4.000   Median :     5   Median :   0.0000
##  Mean   :   4.631   Mean   :     6   Mean   :   0.0004
##  3rd Qu.:   6.000   3rd Qu.:     8   3rd Qu.:   0.0000
##  Max.   :9526.000   Max.   :952290   Max.   :1833.3000
##
##  RadiationWattsM2      PresMbar
##  Min.   :    0.00   Min.   : 980
##  1st Qu.:    0.00   1st Qu.:1014
##  Median :    2.66   Median :1018
##  Mean   :  129.17   Mean   :1018
##  3rd Qu.:  165.90   3rd Qu.:1023
##  Max.   :26177.00   Max.   :1043
##                     NA's   :5853991
```

Summary() is sufficient to identify eight fields with one or more data quality issues:

1. Relative Humidity, Percent (RHumPct): Values below 0 and greater than 100.
2. Temperature, Farenheit (TempF): A minimum value of -99999 and maximum value of 9596.
3. Wind direction (WindDir): Values above 360.
4. Wind Speed in Knots (WindSpeed): a maximum value of 9526.
5. Wind Gust in Knots (WindGust): a maximum value of 952290.
6. Rain in Inches (RainInches): even for Seattle, a maximum value of 1833 inches has never occurred.
7. Radiation in watts/meterˆ2 (RadiationWattsM2): a maximum value of 26177.
8. Pressure in millibars: over 5M Not-Availables (NA's)

How many outliers in a dataset with 6992901 observations:

1. 3574 = Relative Humidity, Percent (RHumPct): Values below 0 and greater than 100:
2. 874 = Temperature, Farenheit (TempF): Temperature < -40 or > 130.
3. 1 = Wind direction (WindDir): Values above 360.
4. 2 = Wind Speed in Knots (WindSpeed): values above 200.
5. 1 = Wind Gust in Knots (WindGust): values above 300.
6. 2 = Rain in Inches (RainInches) > 10: even for Seattle, a maximum value of 1833 inches has never occurred in a 10 minute period.
7. 5430 = Radiation in watts/meterˆ2 (RadiationWattsM2): a value greater than 1000.
8. 5853991 = Pressure in millibars: entries with value of NA.

Filtering Questionable Data

Assignment

b. "Filter the data to remove the questionable data you identified in part a. How much of the data is affected? Hint: some of the functions "length", "nrow", "ncol", "is.na", and "which" may be helpful."

Implementation

For the purpose of looking for and analyzing monthly trends, I put Pressure to the side: most of that column's entries are NA - over 5M out of less than 7M.

```
wdF <- wd[, PresMbar:=NULL]
```

Remove every row with an out-of-bounds value in any of the remaining columns.

```
nRowBeforeFiltering <- nrow(wdF)
wdF <- wdF[!(wdF$RHumPct < 0 | wdF$RHumPct > 100),]
nrow(wdF)
```

```
## [1] 6989327
```

```
wdF <- wdF[!(wdF$TempF < -40 | wdF$TempF > 130),]
nrow(wdF)
```

```
## [1] 6988825
```

```
wdF <- wdF[!(wdF$WindDir > 360),]
nrow(wdF)
```

```
## [1] 6988825
```

```
wdF <- wdF[!(wdF$WindSpeed > 200),]
nrow(wdF)
```

```
## [1] 6988824
```

```
wdF <- wdF[!(wdF$WindGust > 300),]
nrow(wdF)
```

```
## [1] 6988824
```

```
wdF <- wdF[!(wdF$RainInches > 10),]
nrow(wdF)
```

```
## [1] 6988823
```

```
wdF <- wdF[!(wdF$RadiationWattsM2 > 1000),]
nrow(wdF)
```

```
## [1] 6983394
```

```
nRowAfterFiltering <- nrow(wdF)
```

9507 out of 6992901 unfiltered rows were removed.

Monthly Trend

Assignment

c. "Look for and describe a monthly trend in the data."

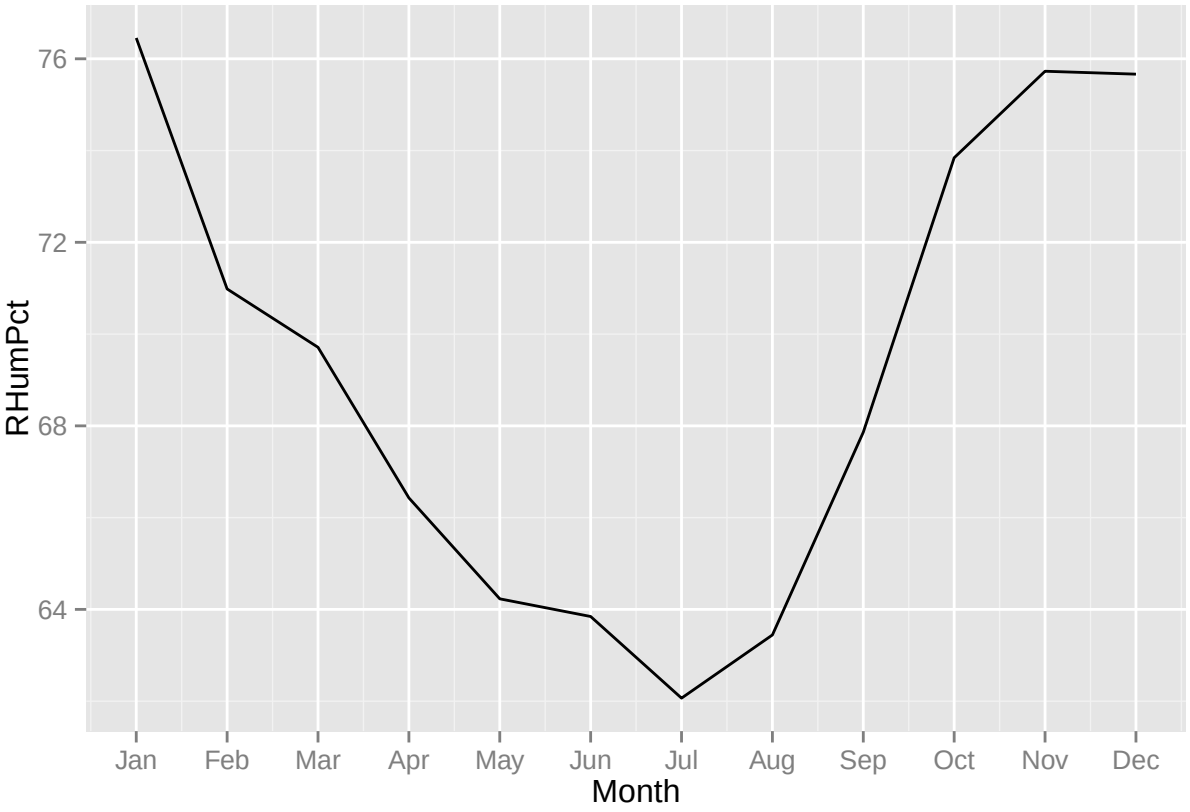
Assumption: we're looking for trends over months in every year, not over years. I.e. aggregate each month for all years.

Implementation

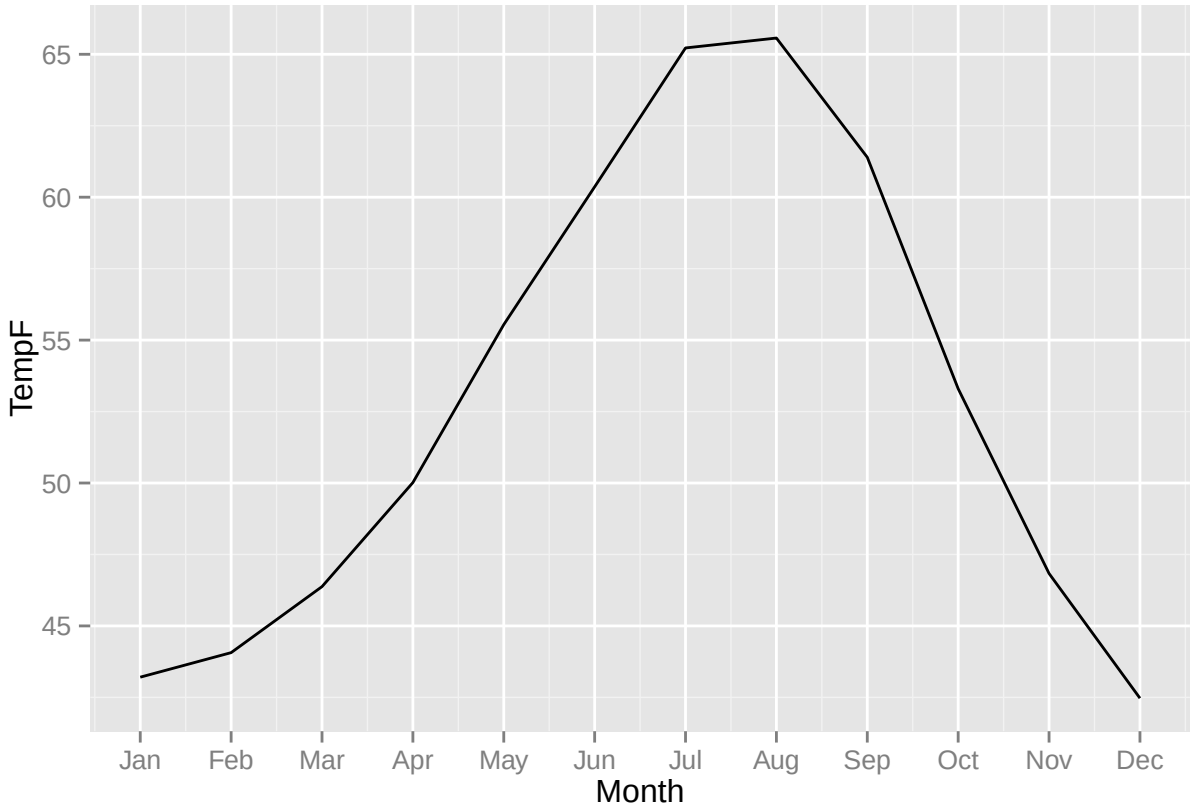
```
# Skip (don't average) columns: Year, Day, Time
columnsToAggregate=c(wdF$RHumPct, wdF$TempF, wdF$WindDir, wdF$windGust, wdF$RainInches)
wdSumByMon <- aggregate(cbind(RHumPct, TempF, WindDir, WindSpeed, WindGust, RainInches) ~ Month, dat
```

Plots of each of the six metrics of interest, by month:

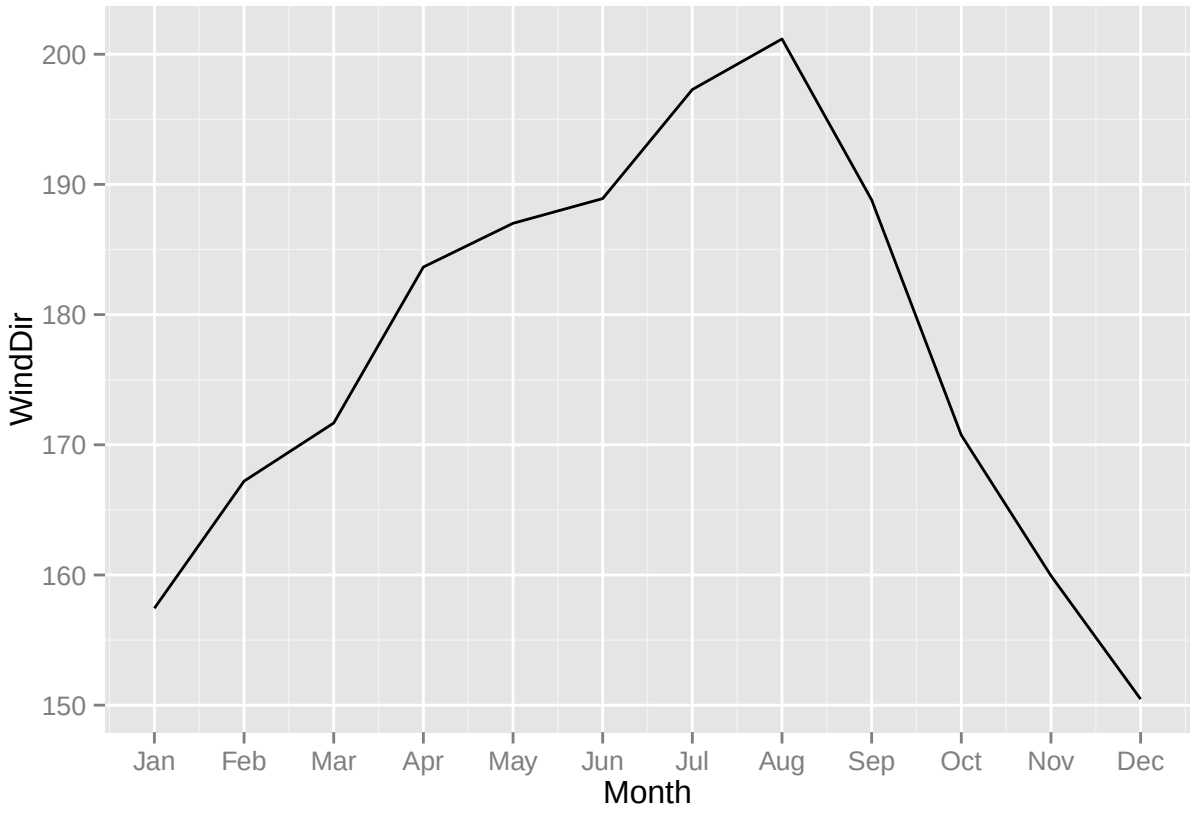
```
library("ggplot2")
xAxisLabels <- c("Jan","Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")
ggplot(wdSumByMon, aes(x=Month)) + scale_x_continuous(breaks=seq(1,12), labels=xAxisLabels) +
    geom_line(aes(y=RHumPct))
```



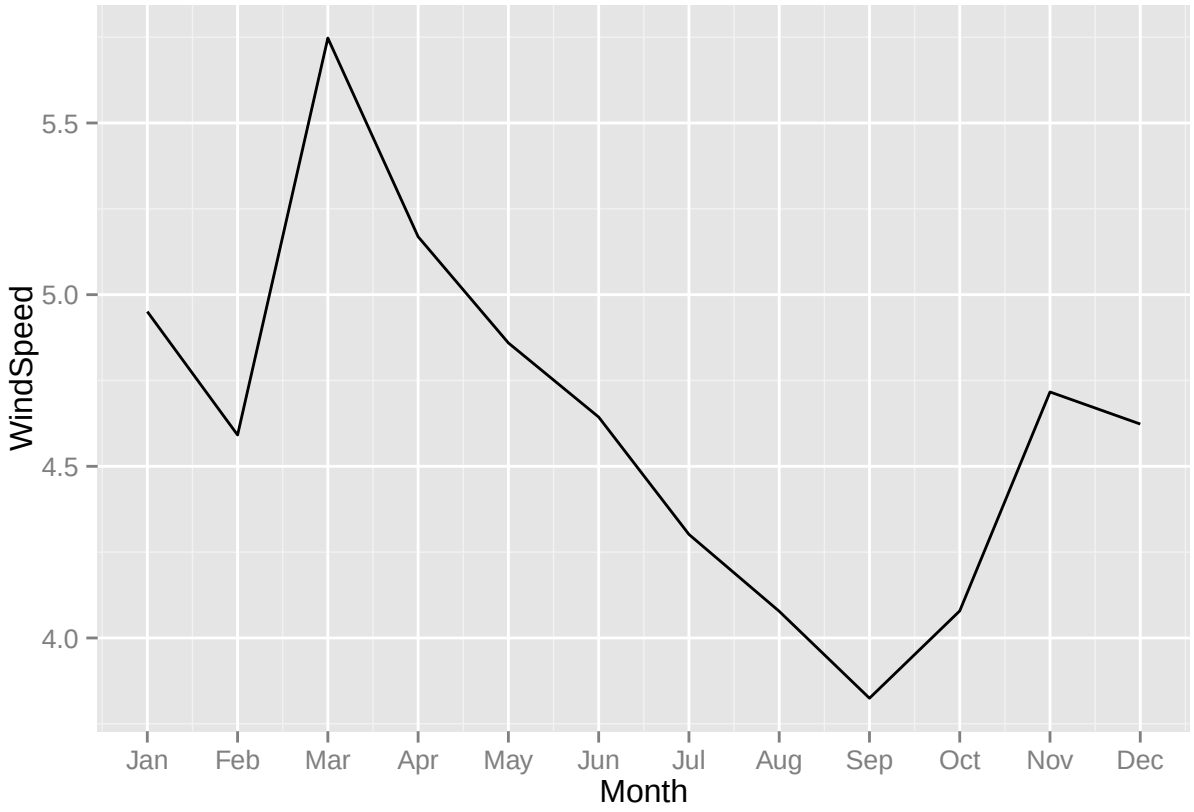

```
ggplot(wdSumByMon, aes(x=Month)) + scale_x_continuous(breaks=seq(1,12), labels=xAxisLabels) +
    geom_line(aes(y=TempF))
```



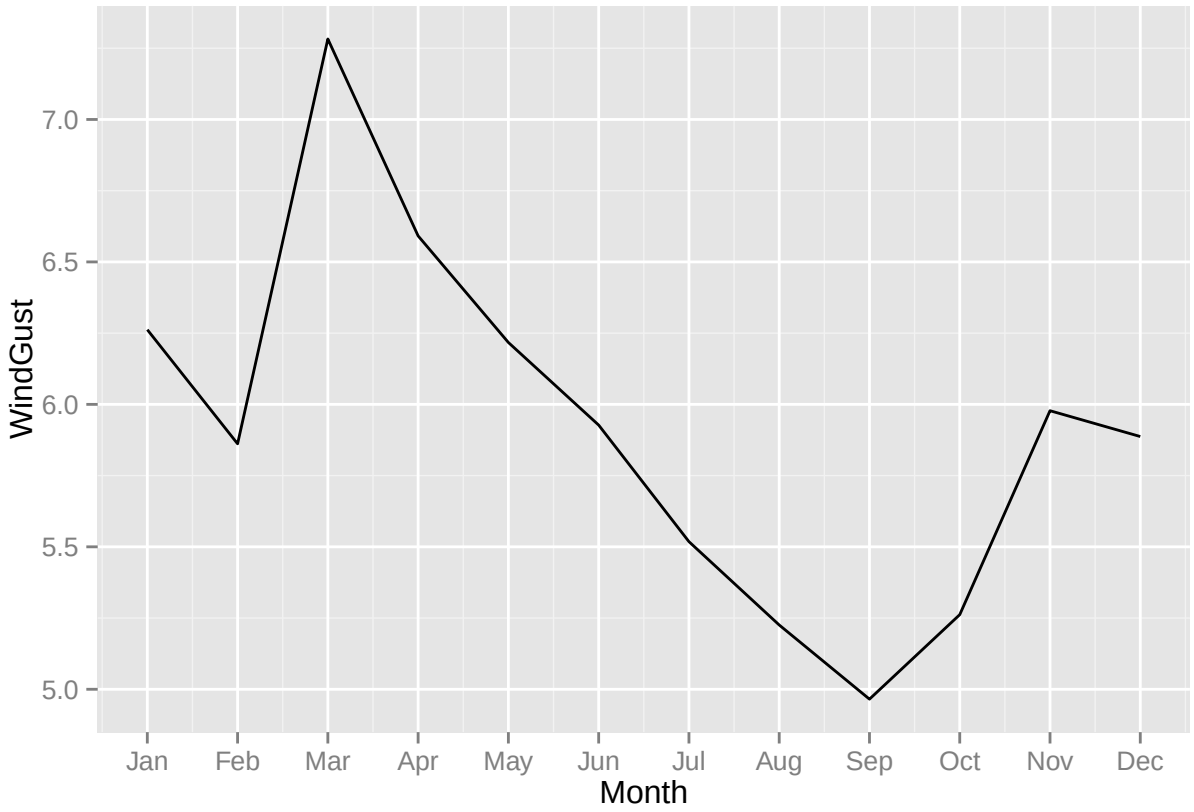

```
ggplot(wdSumByMon, aes(x=Month)) + scale_x_continuous(breaks=seq(1,12), labels=xAxisLabels) +
    geom_line(aes(y=WindDir))
```



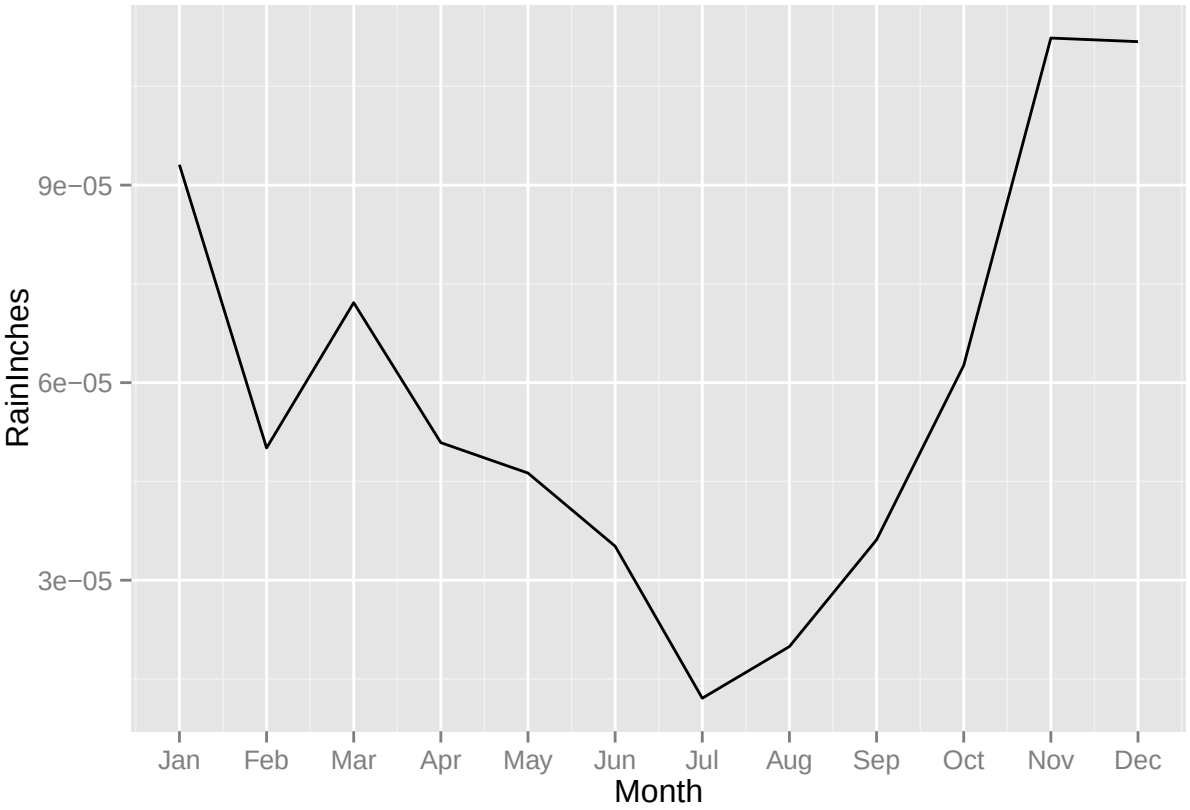

```
ggplot(wdSumByMon, aes(x=Month)) + scale_x_continuous(breaks=seq(1,12), labels=xAxisLabels) +
    geom_line(aes(y=WindSpeed))
```




```
ggplot(wdSumByMon, aes(x=Month)) + scale_x_continuous(breaks=seq(1,12), labels=xAxisLabels) +
    geom_line(aes(y=WindGust))
```




```
ggplot(wdSumByMon, aes(x=Month)) + scale_x_continuous(breaks=seq(1,12), labels=xAxisLabels) +
    geom_line(aes(y=RainInches))
```



Monthly trends observed in the above plot:

- Obvious (for Seattle):
  - Warmest in August, coolest in December/January
  - Driest in the summer, wettest in November/December
  - Least humid in summer, most humid in winter
  - Lightest winds, both steady and gust, in September. Strongest in March.
- Less Obvious:
  - The average wind direction, all seasons is mostly from the south - between 150 and 200 degrees.
  - The wind direction tends to the SSW in the summer and to the SSE in the winter.