Data Wrangling Project

Tom Gocken

Wednesday, January 20, 2016

### Step 1: Load datasets

* Data for counties of interest was queried from <http://wonder.cdc.gov/EnvironmentalData.html>
* Each environmental data type was downloaded in its own tab-delimited file and a dataset was created for each file.

airtemp <- read.delim("data/Air Temperature.txt")  
precip <- read.delim("data/Precipitation.txt")  
sunlight <- read.delim("data/Sunlight.txt")  
surfacetemp <- read.delim("data/Surface Temperature.txt")  
particulate <- read.delim("data/Particulate Matter.txt")

### Step 2: Manage NA's

* **dplyr** package loaded for wrangling functions.

library(dplyr)

* Missing numeric values from certain columns in original files were populated with the string "Missing".
* "Missing" strings were converted to NA using **type.convert** function.

airtemp <- mutate(airtemp, heat\_index =   
 type.convert(as.character(Avg.Daily.Max.Heat.Index..F.),   
 na.strings = "Missing")  
 )   
surfacetemp <- mutate(surfacetemp, day\_surface\_temp =   
 type.convert(as.character(Avg.Day.Land.Surface.Temperature..F.),   
 na.strings = "Missing"),  
 night\_surface\_temp =  
 type.convert(as.character(Avg.Night.Land.Surface.Temperature..F.),   
 na.strings = "Missing")  
 )

Note: The same result could have been accomplished using **gsub** function:

airtemp <- mutate(airtemp, heat\_index = as.numeric(gsub("Missing", NA,   
as.character((airtemp$Avg.Daily.Max.Heat.Index..F.)))))

### Step 3: Join data into a single tidy dataset

* Rows from individual datasets represent unique county/date observations. To ensure that all rows from every dataset were included in the combined dataset, full joins were used.
* Except for environmental data columns, variables were named consistently in downloaded files. As a result, natural joins on all variables with common names were possible without the need for a "by" argument.

joindat <- full\_join(airtemp, precip)

## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Month", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Year.Code")

joindat <- full\_join(joindat, sunlight)

## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Month", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Year.Code")

joindat <- full\_join(joindat, surfacetemp)

## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Month", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Year.Code")

joindat <- full\_join(joindat, particulate)

## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Month", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Year.Code")

Date variable was created by concatenating year, month, and day columns and coverting to date class.

joindat <- mutate(joindat, date = as.Date(paste(joindat$Year.Code,   
 joindat$Month.Code,   
 joindat$Day.of.Month.Code,   
 sep="-")))

Select statement used to assign concise variable names in common format to columms of interest.

envdat <- select(joindat,  
 county = County,  
 year = Year,  
 day\_of\_yr = Day.of.Year,  
 date,  
 max\_air\_temp = Avg.Daily.Max.Air.Temperature..F.,  
 min\_air\_temp = Avg.Daily.Min.Air.Temperature..F.,  
 heat\_index,  
 precip = Avg.Daily.Precipitation..mm.,  
 sunlight = Avg.Daily.Sunlight..KJ.m².,  
 day\_surface\_temp,  
 night\_surface\_temp,  
 particulate\_matter = Avg.Fine.Particulate.Matter..µg.m³.  
 )

Growing degree units (GDUs), also known as growing degree days, were calculated by taking the average of the daily maximum and minimum temperatures compared to a base temperature, T(base), as follows:

GDU = ((T(max) + T(min)) / 2) - T(base)

where T(max) is equal to the maximum daily temperature but not greater than a defined upper limit and T(min) is equal to the maximum daily temperature but not less than the base temperature. The upper limit and base in this project were set to 50°F and 86°F (10°C and 50°C), respectively, typical values for corn.

Accumulated GDUs (AGDUs) were calculated using the **cumsum** function grouped by county and year and ordered by date. AGDUs provide a standard measure of accumulated heat during a growing season. The maturity of a plant variety is often expressed in AGDUs after planting, rather than days, since days to maturity vary by location and season.

References:  
<http://en.wikipedia.org/wiki/Growing_degree-day>  
<http://agron-www.agron.iastate.edu/Courses/agron212/Calculations/GDD.htm>

envdat <- mutate(envdat, gdu = ifelse(max\_air\_temp < 50, 0,  
 (((ifelse(max\_air\_temp > 86, 86, max\_air\_temp)   
 + ifelse(min\_air\_temp < 50, 50, min\_air\_temp)) / 2) - 50)))  
envdat <- transform(envdat, agdu = ave(gdu, paste(county, year),   
 FUN = cumsum))

### Step 4: Summarize and view data

summary(envdat)

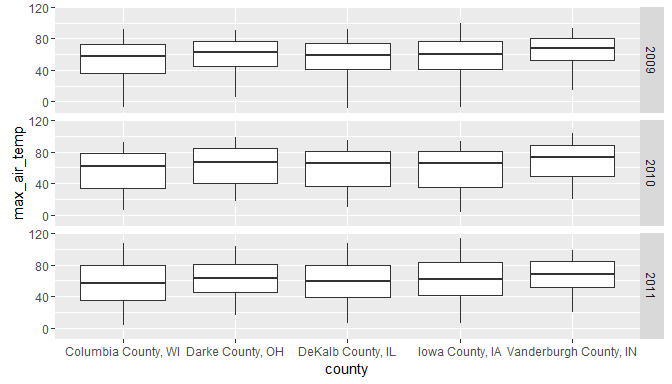
## county year day\_of\_yr   
## Columbia County, WI :7305 Min. :1992 Min. : 1.0   
## Darke County, OH :7305 1st Qu.:1996 1st Qu.: 92.0   
## DeKalb County, IL :7305 Median :2001 Median :183.0   
## Iowa County, IA :7305 Mean :2001 Mean :183.1   
## Vanderburgh County, IN:7305 3rd Qu.:2006 3rd Qu.:274.0   
## Max. :2011 Max. :366.0   
##   
## date max\_air\_temp min\_air\_temp heat\_index   
## Min. :1992-01-01 Min. :-16.81 Min. :-34.48 Min. : 78.40   
## 1st Qu.:1996-12-31 1st Qu.: 41.12 1st Qu.: 28.58 1st Qu.: 84.10   
## Median :2001-12-31 Median : 62.17 Median : 43.37 Median : 88.35   
## Mean :2001-12-31 Mean : 59.32 Mean : 42.48 Mean : 90.22   
## 3rd Qu.:2006-12-31 3rd Qu.: 78.13 3rd Qu.: 58.40 3rd Qu.: 94.80   
## Max. :2011-12-31 Max. :114.08 Max. : 84.07 Max. :128.49   
## NA's :28356   
## precip sunlight day\_surface\_temp night\_surface\_temp  
## Min. : 0.000 Min. : 1450 Min. : -8.23 Min. :-26.27   
## 1st Qu.: 0.000 1st Qu.: 8334 1st Qu.: 47.97 1st Qu.: 24.80   
## Median : 0.100 Median :14337 Median : 71.75 Median : 42.29   
## Mean : 2.754 Mean :14768 Mean : 63.88 Mean : 39.20   
## 3rd Qu.: 1.900 3rd Qu.:20725 3rd Qu.: 81.42 3rd Qu.: 55.77   
## Max. :133.900 Max. :30876 Max. :108.53 Max. : 77.10   
## NA's :28644 NA's :28042   
## particulate\_matter gdu agdu   
## Min. : 0.00 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 8.30 1st Qu.: 0.000 1st Qu.: 83.41   
## Median :12.03 Median : 6.185 Median :1275.78   
## Mean :13.14 Mean : 9.547 Mean :1577.38   
## 3rd Qu.:16.56 3rd Qu.:18.200 3rd Qu.:2916.39   
## Max. :55.30 Max. :35.035 Max. :4957.62   
## NA's :20090

Box plot and line graphs created using **ggplot2**.

library(ggplot2)

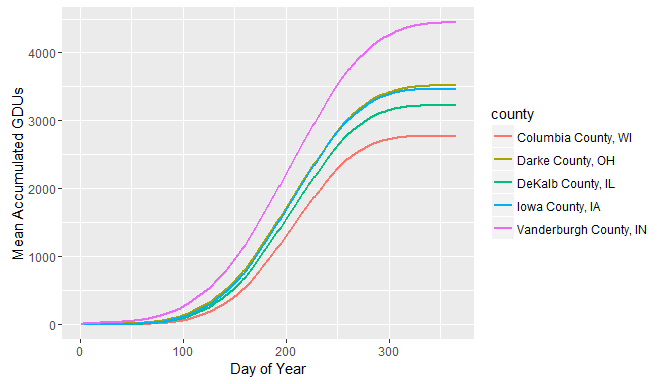
Differences in max air temp by year and county (2009-2011):

envdat\_3yr <- filter(envdat, year %in% c(2009, 2010, 2011))  
envdat\_3yr$year <- as.factor(envdat\_3yr$year)  
qplot(county, max\_air\_temp, data = envdat\_3yr, geom = "boxplot",   
 facets = year ~ .)



Differences in accumulated GDUs by county, across years:

county\_means <- envdat %>%  
 filter(day\_of\_yr != 366) %>% # exclude extra leap year day  
 group\_by(county, day\_of\_yr) %>%  
 summarize(agdu\_mean = mean(agdu))  
qplot(day\_of\_yr, agdu\_mean, data = county\_means, geom = "line", color = county,   
 xlab = "Day of Year", ylab = "Mean Accumulated GDUs") + geom\_line(size = 1.0)



Differences in accumulated GDUs by 5-year means, across counties:

envdat <- mutate(envdat, yr\_group = ifelse(year < 1997, "1992-1996",  
 ifelse(year < 2002, "1997-2001",  
 ifelse(year < 2007, "2002-2006",  
 "2007-2011"))))  
yr\_means <- envdat %>%  
 filter(day\_of\_yr != 366) %>% # exclude extra leap year day  
 group\_by(yr\_group, day\_of\_yr) %>%  
 summarize(agdu\_mean = mean(agdu))  
qplot(day\_of\_yr, agdu\_mean, data = yr\_means, geom = "line",   
 color = yr\_group, xlab = "Day of Year",   
 ylab = "Mean Accumulated GDUs") + geom\_line(size = 1.0)

