Data Wrangling Project

Tom Gocken

Wednesday, January 20, 2016

Step 1: Load CDC environmental datasets

- Data for counties of interest was queried from http://wonder.cdc.gov/EnvironmentalData.html (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.

```
setwd("C:/Projects/springboard-wrangling")
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 = "Missin
g"))

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: Load and reshape monthly ERSST data measuring El Nino / La Nina effects

- Historical El Nino/ La Nina data source: http://www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml (http://www.cpc.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml)

```
el_nino <- read.csv("../data/el_nino.csv")
el_nino <- rename(el_nino, year = Year, "1" = DJF, "2" = JFM, "3" = FMA, "4" =
MAM, "5" = AMJ, "6" = MJJ,
                  "7" = JJA, "8" = JAS, "9" = ASO, "10" = SON, "11" = OND, "1
2" = NDJ)
```

- Gather monthly data into single column using **tidyr** package

```
el_nino2 <- tidyr::gather(el_nino, "month", "ersst", 2:13)
el_nino2 <- mutate(el_nino2, month = as.integer(el_nino2$month))
```

Step 4: Join data into a single tidy dataset

```
joindat <- left_join(airtemp, precip)
```

```
## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Mont
h", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Y
ear.Code")
```

```
joindat <- left_join(joindat, sunlight)
```

```
## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Mont
h", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Y
ear.Code")
```

```
joindat <- left_join(joindat, surfacetemp)
```

```
## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Mont
h", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Y
ear.Code")
```

```
joindat <- left_join(joindat, particulate)
```

```
## Joining by: c("Notes", "County", "County.Code", "Year", "Year.Code", "Mont
h", "Month.Code", "Day.of.Month", "Day.of.Month.Code", "Day.of.Year", "Day.of.Y
ear.Code")
```

```
joindat <- left_join(joindat, el_nino2, by = c("Year" = "year", "Month.Code" =
"month"))
```

Date variable 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,
                 month = Month.Code,
                 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.,
                 ersst
                 )
```

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://en.wikipedia.org/wiki/Growing_degree-day)
http://agron-www.agron.iastate.edu/Courses/agron212/Calculations/GDD.htm (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 5: Summarize and view data

```
summary(envdat)
```

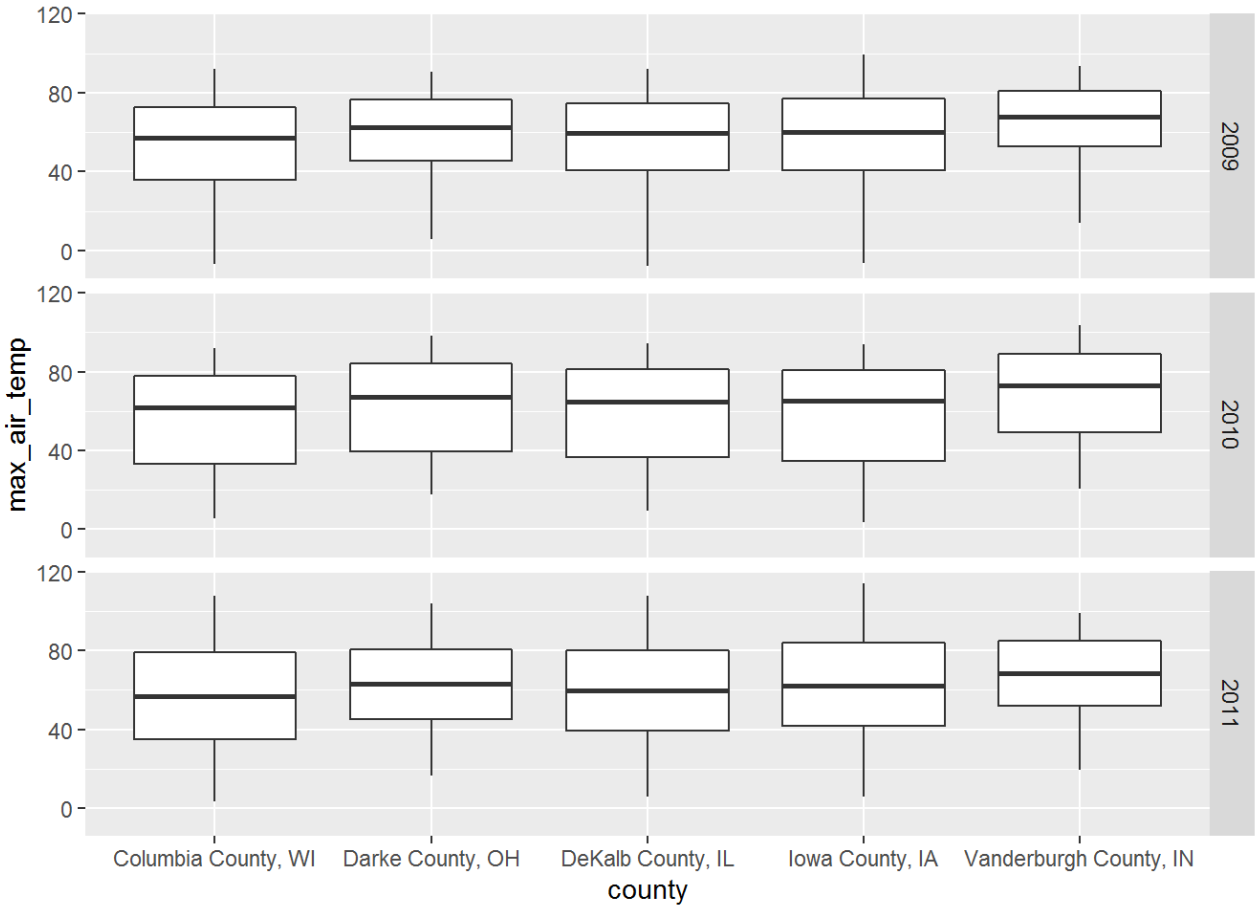
```
##                     county          year          month
##  Columbia County, WI   :7305   Min.   :1992   Min.   : 1.000
##  Darke County, OH      :7305   1st Qu.:1996   1st Qu.: 4.000
##  DeKalb County, IL     :7305   Median :2001   Median : 7.000
##  Iowa County, IA       :7305   Mean   :2001   Mean   : 6.523
##  Vanderburgh County, IN:7305   3rd Qu.:2006   3rd Qu.:10.000
##                                Max.   :2011   Max.   :12.000
##
##    day_of_yr          date             max_air_temp     min_air_temp
##  Min.   :  1.0   Min.   :1992-01-01   Min.   :-16.81   Min.   :-34.48
##  1st Qu.: 92.0   1st Qu.:1996-12-31   1st Qu.: 41.12   1st Qu.: 28.58
##  Median :183.0   Median :2001-12-31   Median : 62.17   Median : 43.37
##  Mean   :183.1   Mean   :2001-12-31   Mean   : 59.32   Mean   : 42.48
##  3rd Qu.:274.0   3rd Qu.:2006-12-31   3rd Qu.: 78.13   3rd Qu.: 58.40
##  Max.   :366.0   Max.   :2011-12-31   Max.   :114.08   Max.   : 84.07
##
##    heat_index         precip           sunlight     day_surface_temp
##  Min.   : 78.40   Min.   :  0.000   Min.   : 1450   Min.   : -8.23
##  1st Qu.: 84.10   1st Qu.:  0.000   1st Qu.: 8334   1st Qu.: 47.97
##  Median : 88.35   Median :  0.100   Median :14337   Median : 71.75
##  Mean   : 90.22   Mean   :  2.754   Mean   :14768   Mean   : 63.88
##  3rd Qu.: 94.80   3rd Qu.:  1.900   3rd Qu.:20725   3rd Qu.: 81.42
##  Max.   :128.49   Max.   :133.900   Max.   :30876   Max.   :108.53
##  NA's   :28356                                      NA's   :28644
##  night_surface_temp particulate_matter     ersst               gdu
##  Min.   :-26.27     Min.   : 0.00      Min.   :-1.60000   Min.   : 0.000
##  1st Qu.: 24.80     1st Qu.: 8.30      1st Qu.:-0.70000   1st Qu.: 0.000
##  Median : 42.29     Median :12.03      Median : 0.00000   Median : 6.185
##  Mean   : 39.20     Mean   :13.14      Mean   :-0.02597   Mean   : 9.547
##  3rd Qu.: 55.77     3rd Qu.:16.56      3rd Qu.: 0.50000   3rd Qu.:18.200
##  Max.   : 77.10     Max.   :55.30      Max.   : 2.30000   Max.   :35.035
##  NA's   :28042      NA's   :20090
##       agdu
##  Min.   :   0.00
##  1st Qu.:  83.41
##  Median :1275.78
##  Mean   :1577.38
##  3rd Qu.:2916.39
##  Max.   :4957.62
##
```

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

```
library(ggplot2)
```

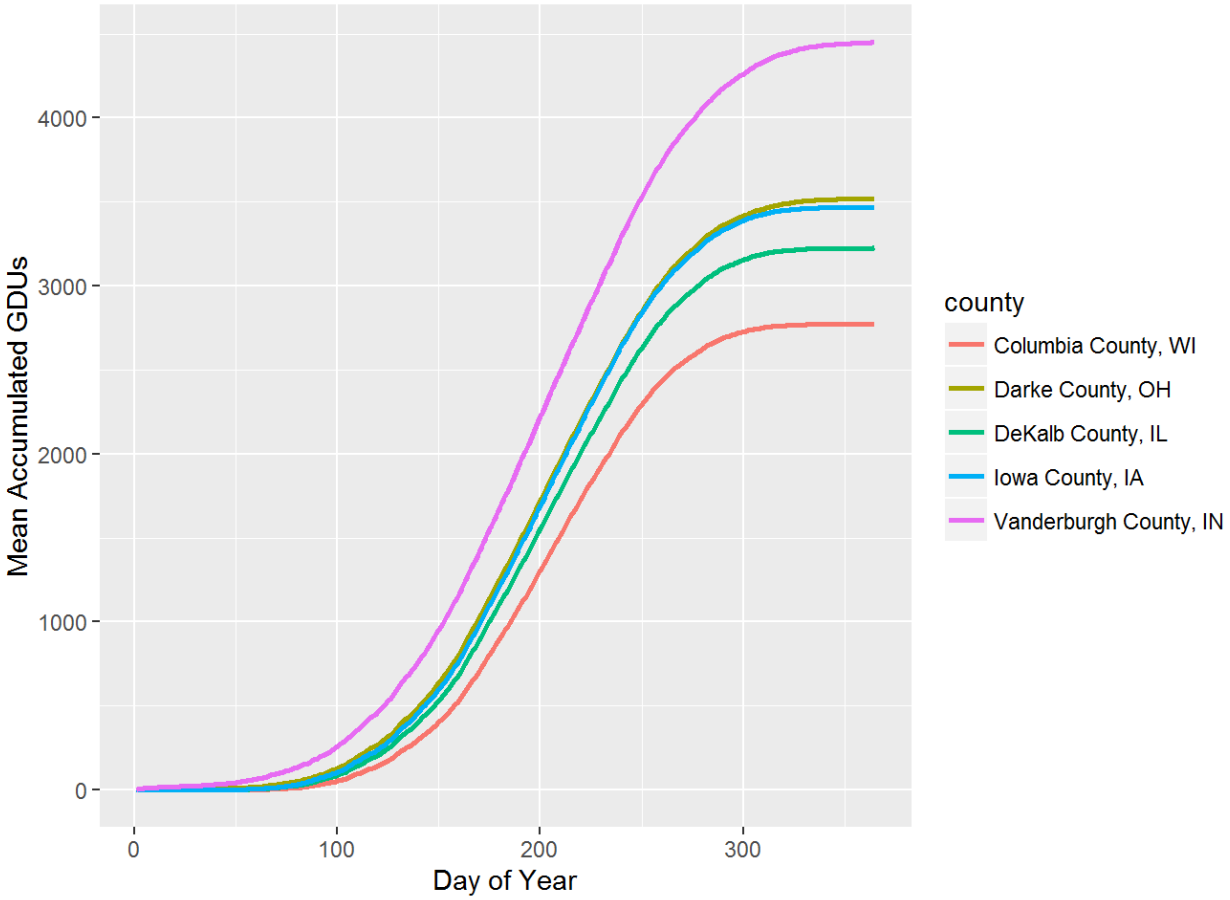
Differences in max air temp by year and county (2009-2011):

```
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 ~ .)
```



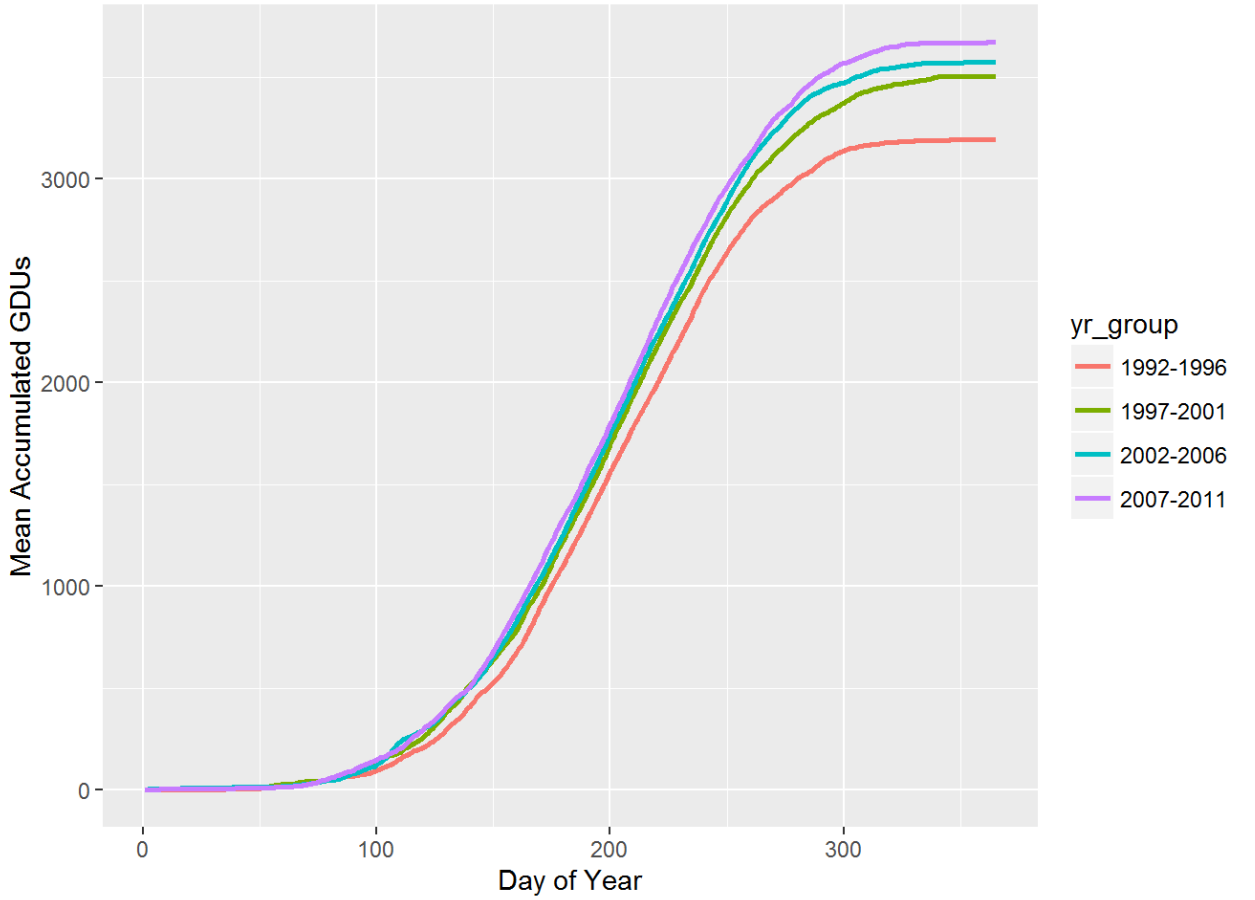
Differences in accumulated GDUs by county, across years:

```
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 = count
y,
      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)
```




```
# Save results for future use
write.table(envdat, "../data/envdat.txt", sep = "\t")
```