

We R Under Way: A Data Science Portfolio

Kaydee S. Barker

Contents

1	Introduction	7
2	Interactive Graphing: Discharge of the Poudre River	9
2.1	Background on the Poudre River	9
2.2	Interactive Discharge Chart	10
3	Looking at Effects of Fire on Vegetation	11
3.1	Introduction	11
3.2	What is the correlation between NDVI and NDMI?	11
3.3	What month is the greenest month on average?	13
3.4	What month is the snowiest on average?	15
4	Fire Effects on Fish Populations	17
4.1	Pre versus post fire fish length and mass	17
4.2	Linear regression of fish mass vs. length for before and after the fire	20
5	Extracting and Visualizing Meteorological Data	27
5.1	1. Extract the meteorological data URLs. Here we want you to use the <code>rvest</code> package to get the URLs for the <code>SASP forcing</code> and <code>SBSP_forcing</code> meteorological datasets.	27
5.2	2. Download the meteorological data. Use the <code>download_file</code> and <code>str_split_fixed</code> commands to download the data and save it in your data folder. You can use a for loop or a map function.	28
5.3	3. Write a custom function to read in the data and append a site column to the data.	29

5.4	4. Use the <code>map</code> function to read in both meteorological files. Display a summary of your tibble.	29
5.5	5. Make a line plot of mean temp by year by site (using the <code>air temp [K]</code> variable). Is there anything suspicious in the plot? Adjust your filtering if needed.	31
5.6	6. Write a function that makes line plots of monthly average temperature at each site for a given year. Use a for loop to make these plots for 2005 to 2010.	31
5.7	Bonus: Make a plot of average daily precipitation by day of year (averaged across all available years)	36
6	Spatial Analysis in R	37
6.1	Loading in data	37
6.2	Part one	40
6.3	Part two	43
7	Linear Regressions, Quadratic Fits, Residuals, and Spatial	49
7.1	Weather Data Analysis	49
7.2	Extract Winneshiek County corn yields, fit a linear time trend, make a plot. Is there a significant time trend?	51
7.3	Fit a quadratic time trend (i.e., $\text{year} + \text{year}^2$) and make a plot. Is there evidence for slowing yield growth?	52
7.4	Time Series: Let's analyze the relationship between temperature and yields for the Winneshiek County time series. Use data on yield and summer avg Tmax. Is adding year or Tmax^2 to your model helpful? Make a plot and interpret the results.	54
7.5	Cross-Section: Analyze the relationship between temperature and yield across all counties in 2018. Is there a relationship? Interpret the results.	58
7.6	Panel: One way to leverage multiple time series is to group all data into what is called a "panel" regression. Convert the county ID code ("countyfp" or "county_ansi") into factor using <code>as.factor</code> , then include this variable in a regression using all counties' yield and summer temperature data. How does the significance of your temperature coefficients (Tmax , Tmax^2) change? Make a plot comparing actual and fitted yields and interpret the results of your model.	60

- 7.7 Soybeans: Download NASS data on soybean yields and explore either a time series relationship for a given county, the cross-sectional relationship for a given year, or a panel across all counties and years. 64
- 7.8 Bonus: Find a package to make a county map of Iowa displaying some sort of information about yields or weather. Interpret your map. 70

Writing and code by Kaydee Barker, assignments by Dr. Ross and Dr. Mueller (SOCR 580A7), Dr. Lefsky (ESS 330) of Colorado State University. Data cited within chapters.

Chapter 1

Introduction

“There are two kinds of data scientists: 1) Those who can extrapolate from incomplete data.”

I began my foray into R in the spring of 2020, first teaching myself some basic syntax and then using it for statistical analysis on my research projects as an undergraduate researcher at Colorado State University (CSU). With the help of my research mentors and many amazing people of the internet, I was able to fumble my way forward and learn a number of techniques to analyze and visualize data in R. I have since been building on my R and data science skills, including with the help of two key courses at CSU: “Quantitative Reasoning for Ecosystem Science” (ESS 330) and “Introduction to Environmental Data Science” (SOCR 580A7). Since I can’t yet publish data from my research projects, this portfolio is constructed of public data examples, primarily from my coursework in those two courses. Its purpose is a) to serve as a reference for myself and others learning to use R for environmental analyses, and b) to demonstrate my current R knowledge to advisors and colleagues.

Chapter 2

Interactive Graphing: Discharge of the Poudre River

“Someone asked me to name two structures that hold water. I was like, ‘well... damn!’ ”

This assignment used a unique package of R Markdown (dygraphs) in order to create an interactive chart.

Data and assignment provided by Dr. Matthew Ross and Dr. Nathan Mueller of Colorado State University.

2.1 Background on the Poudre River

Cache La Poudre River is an important watershed that supports **agriculture, industry, recreation, and residential needs** on the Front Range of Colorado. It also provides for cottonwood forest, shrub, and grassland ecosystems that support wildlife from the mountains down to the prairies. The unique **bio-diversity** and **history** of the Cache La Poudre watershed are valued widely; 45 miles along the Poudre are encompassed in a National Heritage Area. The history of Cache La Poudre is linked to the *history of the West*, because its banks supported the first major irrigation-based agricultural settlement of its kind in 1870, which would soon spread through the Arid West.

2.2 Interactive Discharge Chart

```
q <- readNWISdv(siteNumbers = '06752260',
                parameterCd = '00060',
                startDate = '2017-01-01',
                endDate = '2022-01-01') %>%
  rename(q = 'X_00060_00003')

q_xts <- xts(q$q, order.by = q$Date)

dygraph(q_xts) %>%
  dyAxis("y", label = "Discharge (cfs)") %>%
  dyOptions(drawPoints = TRUE, pointSize = 2)
```

PhantomJS not found. You can install it with `webshot::install_phantomjs()`. If it is

Discharge of the Poudre River in cubic feet per second from January 2017 to December 2021.

Chapter 3

Looking at Effects of Fire on Vegetation

“What happens when a wildfire tells you a joke? You get burned!”

This assignment demonstrates the benefit of visualizing data to see potential correlations.

Data and assignment provided by Dr. Matthew Ross and Dr. Nathan Mueller of Colorado State University.

3.1 Introduction

The Hayman Fire, started by arson in summer of 2002, was the largest wildfire in Colorado history until the 2020 wildfire season. It burned a large area of over 138 thousand acres between the Kenosha Mountains and Pikes Peak, affecting wildlife and causing water quality concerns for the Front Range populations through damage to watersheds that contribute to the South Platte River.

3.2 What is the correlation between NDVI and NDMI?

The Normalized Difference Vegetation Index (NDVI) is positively correlated with the Normalized Difference Moisture Index (NDMI). In everyday terms, NDVI indicates plant health as shown by how well leaves reflect near infrared and red light, while NDMI represents plant water content and is calculated from near infrared and short-wave infrared reflectance values (Agricolus, 2018).

These values can also tell us about how much vegetative cover there is at a given site, with the lowest NDVI (<0.1) and NDMI (<-0.8) values indicating bare soil.

Not surprisingly, the plot below shows that canopy cover is greatly decreased for the burned site compared to the unburned site.

```
#ggplot of wide set in summer
full_wide %>%
  filter(month %in% c(6,7,8,9,10)) %>%
  filter(year >= 2002) %>%
  ggplot(., aes(x=ndmi,y=ndvi, color=treatment)) +
  geom_point(shape=1) +
  xlab("NDMI") + ylab("NDVI") +
  ggtitle("Burned vs. Unburned Vegetation") +
  theme_few(base_size = 16) +
  scale_color_brewer(palette = "Set2") +
  theme(panel.grid.major=element_blank(), panel.grid.minor=element_blank(), legend.pos=
```

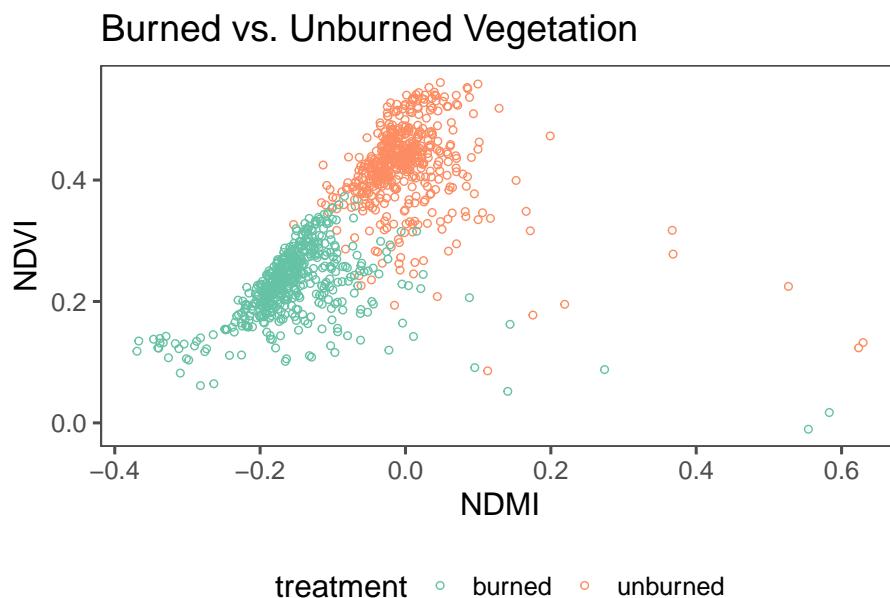


Figure 3.1: NDVI and NDMI values from 2002 to 2019 in Colorado sites that were burned (teal) or left unburned (orange) during the Hayman Fire.

As may be expected, vegetative growth (NDVI) is positively associated with the previous winter's snowfall, as shown in the plot below.

```

#ggplot winter NDSI to summer NDVI
ggplot(ndvi_ndsi, aes(x = mean_NDVI, y = mean_NDSI)) +
  geom_point(fill = "blue",
             shape = 21,
             size = 2) +
  geom_smooth(method = "lm",
             se = TRUE,
             lty = 1,
             color = "black",
             fill = "lightgrey",
             size = 1) +
  xlab("Mean NDSI") + ylab("Mean NDVI") +
  ggtitle("Winter NDSI vs. Summer NDVI") +
  theme_few(base_size = 16) +
  scale_y_continuous(breaks = pretty(c(-0.4,0.5), n = 4)) +
  scale_x_continuous(breaks = pretty(c(0.2,0.5), n = 6)) +
  theme(panel.grid.major=element_blank(), panel.grid.minor=element_blank(), legend.position="bottom")

## `geom_smooth()` using formula 'y ~ x'

```

3.3 What month is the greenest month on average?

If we plot monthly means of NDVI, we can see that the greenest month in Colorado is August.

```

#ggplot of monthly means
monthly_sum %>%
  filter(data == "ndvi") %>%
  mutate_at(vars(month), funs(factor)) %>%
  ggplot(., aes(x=month, y=value_mean, fill=month)) +
  geom_bar(stat = "identity", width = 0.7, position = "dodge") +
  geom_errorbar(aes(ymin=value_mean-value_std.error, ymax=value_mean+value_std.error),
               colour = "black", width = 0.7, position = "dodge") +
  scale_x_discrete(labels=c("5"="May", "6"="June", "7"="Jul.", "8"="Aug.", "9"="Sept.")) +
  xlab("Month") + ylab("NDVI") +
  ggtitle("Average NDVI per Month") +
  theme_few() +
  scale_fill_brewer(palette = "Greens") +
  theme(panel.grid.major=element_blank(),
        panel.grid.minor=element_blank(), legend.position="none")

```

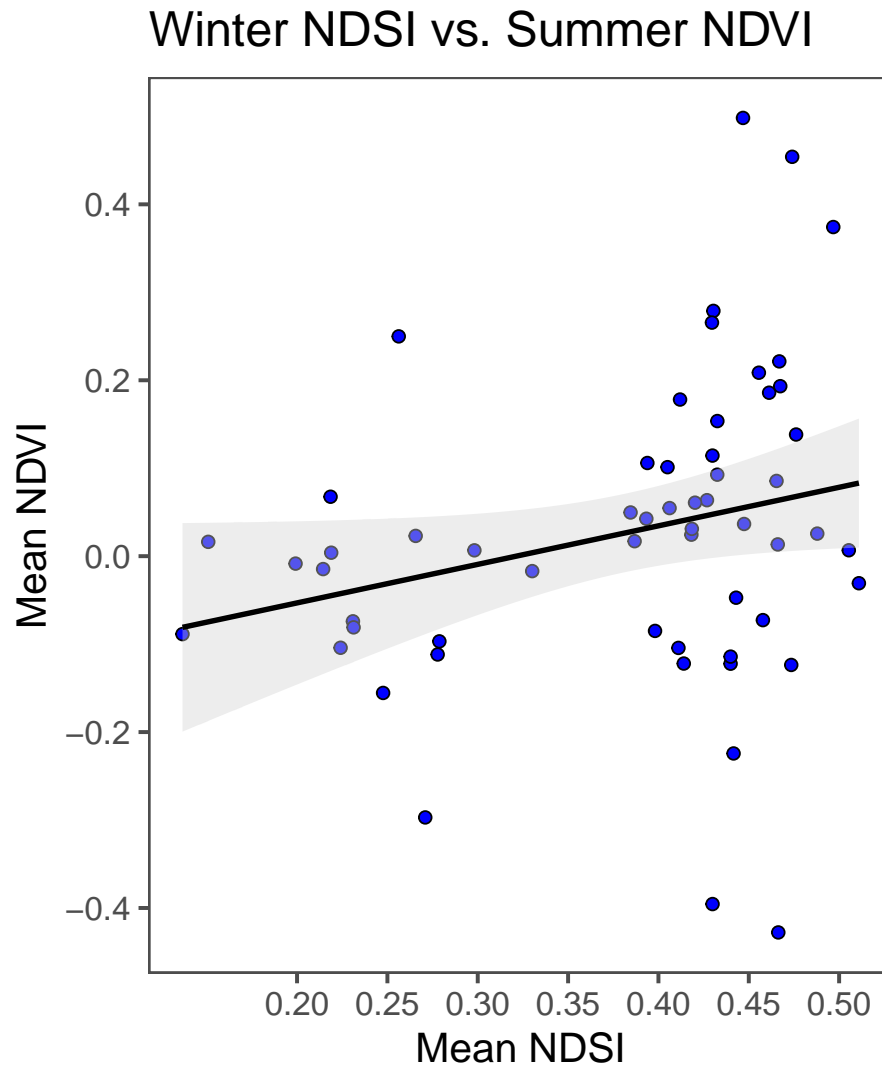


Figure 3.2: Linear models for mean summer NDVI and mean winter NDSI for pre- and post-burn and burned and unburned sites.

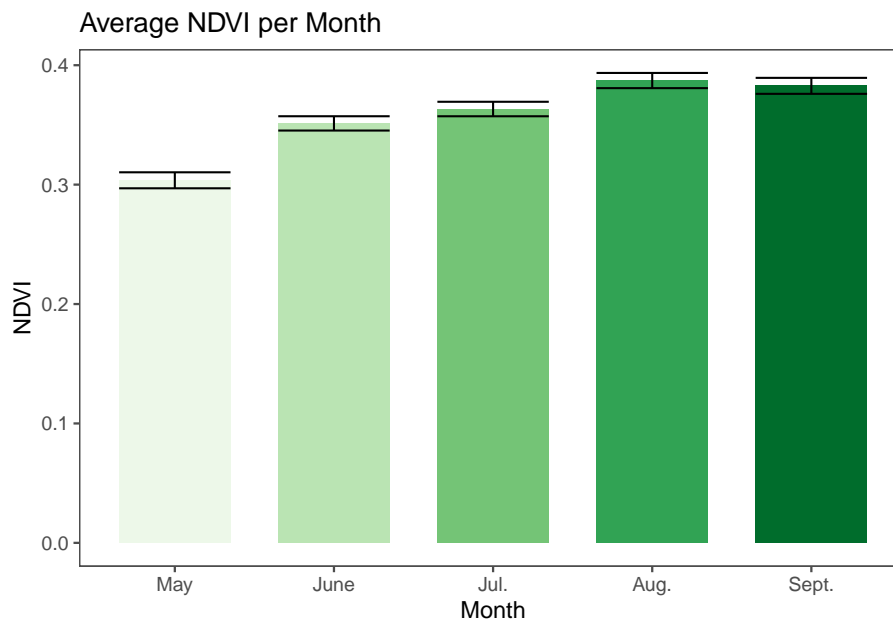


Figure 3.3: Mean NDVI and standard error per summer month across sites from 1984 to 2019.

3.4 What month is the snowiest on average?

If we plot the NDSI means for the winter months, we can see that the highest snowfall is January.

```
# Change ordering manually and make month into factor
monthly_win$month <- factor(monthly_win$month,
                             levels = c("11", "12", "1", "2", "3"))

monthly_win %>%
  filter(data == "ndsi") %>%
  ggplot(., aes(x=month, y=value_mean, fill=month)) +
  geom_bar(stat = "identity", width = 0.7, position = "dodge") +
  geom_errorbar(aes(ymin=value_mean-value_std.error, ymax=value_mean+value_std.error),
                colour = "black", width = 0.7, position = "dodge") +
  scale_x_discrete(labels=c("11"="Nov.", "12"="Dec", "1"="Jan.", "2"="Feb.",
                             "3"="Mar.")) +
  xlab("Month") + ylab("NDSI") +
  ggtitle("Average NDSI per Month") +
  theme_few() +
  scale_fill_brewer(palette = "Purples") +
```

```
theme(panel.grid.major=element_blank(), panel.grid.minor=element_blank(),  
      legend.position="none")
```

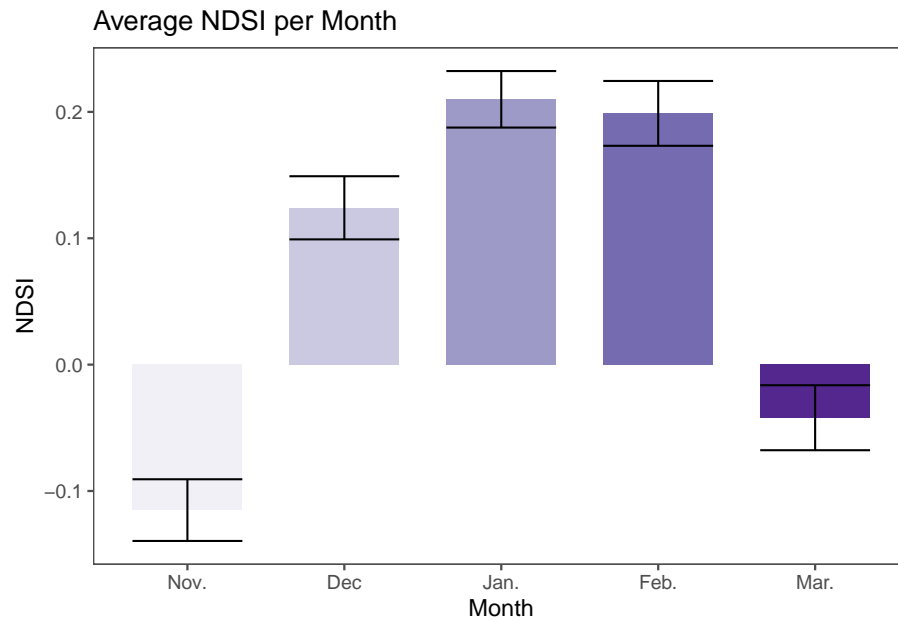


Figure 3.4: Mean NDSI and standard error per winter month across sites from 1984 to 2019.

Chapter 4

Fire Effects on Fish Populations

Wildfires don't only impact vegetation, but a wide variety of abiotic and biotic elements of the ecosystem. In this assignment, I looked at how fish in the Cache La Poudre Watershed were impacted by the High Park Fire in 2012.

Data and assignment provided by Dr. Michael Lefsky of Colorado State University.

4.1 Pre versus post fire fish length and mass

```
#summarize fishdata_R by time (another way to do this without subsets)
summary(fishdata_4R[fishdata_4R$time=="pre-fire",])
```

```
##      time      capture_id    length_cm    mass_g
## Length:100    Min.   :  1.00    Min.   :  5.00    Min.   :  66
## Class :character 1st Qu.: 25.75    1st Qu.:15.00    1st Qu.:132
## Mode  :character Median : 50.50    Median :18.00    Median :151
##              Mean  : 50.50    Mean   :19.16    Mean   :154
##              3rd Qu.: 75.25    3rd Qu.:23.00    3rd Qu.:182
##              Max.   :100.00    Max.    :32.00    Max.    :252
```

```
summary(fishdata_4R[fishdata_4R$time=="post-fire",])
```

```
##      time      capture_id    length_cm    mass_g
## Length:97    Min.   :  1    Min.   :  5.00    Min.   : 45.0
```

```
## Class :character 1st Qu.:25 1st Qu.:15.00 1st Qu.: 89.0
## Mode :character Median :49 Median :20.00 Median :113.0
## Mean :49 Mean :19.76 Mean :107.9
## 3rd Qu.:73 3rd Qu.:25.00 3rd Qu.:126.0
## Max. :97 Max. :38.00 Max. :157.0
```

```
# create function to run statistics
```

```
lab_stats <- function(x) c(sd(x),sd(x)^2,sd(x)/sqrt(length(x))) #calculate standard de
```

```
#Pre-fire statistics
```

```
lab_stats(fishdata_4R[fishdata_4R$time=="pre-fire",]$length_cm) #fish length
```

```
## [1] 6.2145479 38.6206061 0.6214548
```

```
lab_stats(fishdata_4R[fishdata_4R$time=="pre-fire",]$mass_g) #fish mass
```

```
## [1] 36.277409 1316.050404 3.627741
```

```
#Post-fire statistics
```

```
lab_stats(fishdata_4R[fishdata_4R$time=="post-fire",]$length_cm) #fish length
```

```
## [1] 7.0574624 49.8077749 0.7165767
```

```
lab_stats(fishdata_4R[fishdata_4R$time=="post-fire",]$mass_g) #fish mass
```

```
## [1] 26.894853 723.333119 2.730759
```

```
# 1-way ANOVA on pre- vs. post-fire mass and length
```

```
summary(aov(fishdata_4R$length_cm~fishdata_4R$time)) #ANOVA for fish length pre vs. po
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## fishdata_4R$time  1      18    17.90   0.406  0.525
## Residuals      195     8605    44.13
```

```
summary(aov(fishdata_4R$mass_g~fishdata_4R$time)) #ANOVA for fish mass pre vs. post fi
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## fishdata_4R$time  1 104798  104798  102.3 <2e-16 ***
## Residuals      195 199729    1024
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Make a 2 x 2 matrix of histograms for pre- and post-fire mass and length
par(mfrow=c(2,2)) #tell R how I want figures arranged
```

```
#Pre-fire histograms
```

```
hist(fishdata_4R[fishdata_4R$time == "pre-fire",]$length_cm,main="Pre-fire length (cm)",xlab="Length (cm)",ylab="Frequency")
```

```
hist(fishdata_4R[fishdata_4R$time == "pre-fire",]$mass_g,main="Pre-fire mass (g)",xlab="Mass (g)",ylab="Frequency")
```

```
#Post-fire histograms
```

```
hist(fishdata_4R[fishdata_4R$time == "post-fire",]$length_cm,main="Post-fire length (cm)",xlab="Length (cm)",ylab="Frequency")
```

```
hist(fishdata_4R[fishdata_4R$time == "post-fire",]$mass_g,main="Post-fire mass (g)",xlab="Mass (g)",ylab="Frequency")
```

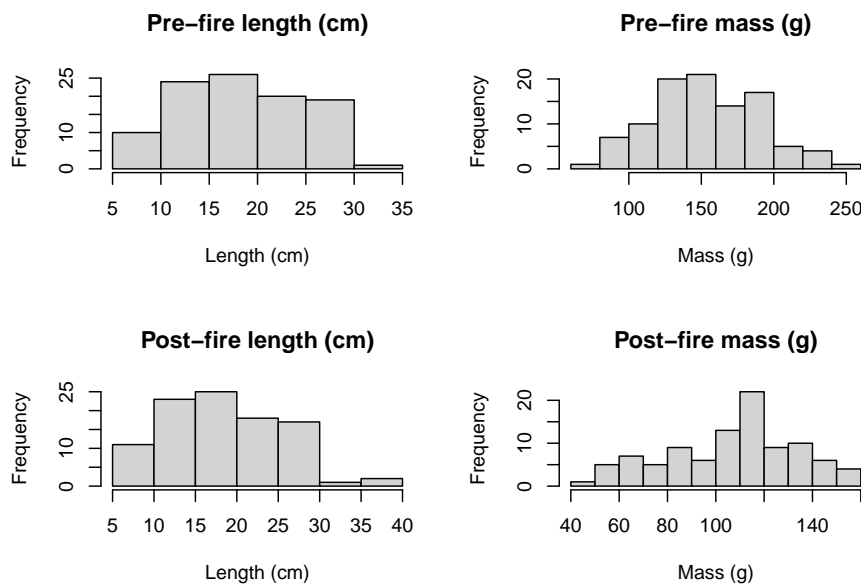


Figure 4.1: Histograms showing frequency of various lengths in centimeters and masses in grams of fish in Cache La Poudre Watershed in 2012 before the High Park Fire (Pre-fire) and in 2013 after the High Park Fire (Post-fire).

```
# Make a 2 x 2 matrix of histograms for pre- and post-fire mass and length
par(mfrow=c(2,2)) #tell R how I want figures arranged
```

```
# Make two boxplots side by side
```

```
par(mfrow=c(1,2)) #tell R I want two plots
```

```
boxplot(fishdata_4R$length_cm~fishdata_4R$time, main="Length (cm)",ylab = "Frequency",xlab="Time")
```

```
boxplot(fishdata_4R$mass_g~fishdata_4R$time, main="Mass (g)",ylab = "Frequency",xlab="Time") #length
```

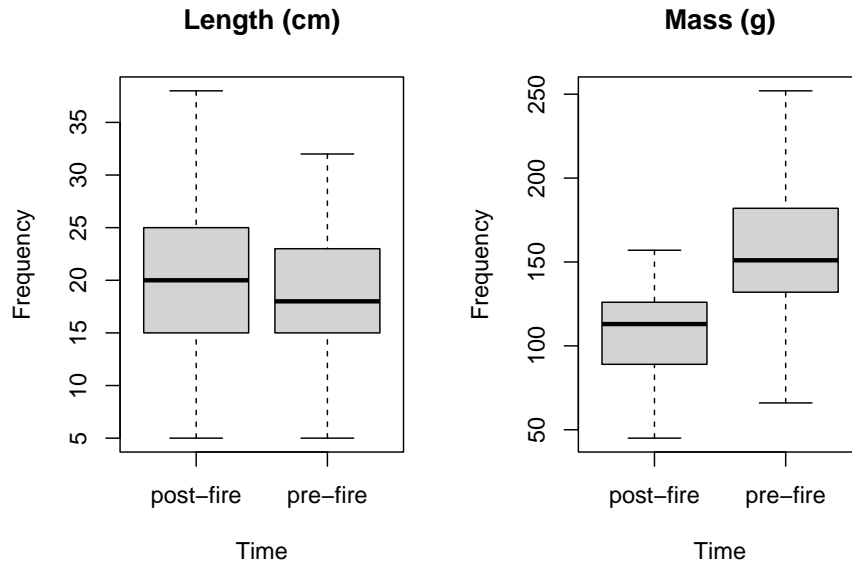


Figure 4.2: Boxplots for fish length in centimeters and mass in grams pre and post fire.

```
# Reset setting for plots
par(mfrow=c(1,1)) #return to single plot
```

4.2 Linear regression of fish mass vs. length for before and after the fire

```
# Pre-fire
# Scatterplot of length and mass where length is the independent variable and mass is
plot(mass_g ~ length_cm, data=fishdata_4R[fishdata_4R$time=="pre-fire",], xlab="Length",
title("Pre-fire Fish Mass vs. Length"))

# Linear regression on mass vs.length
lm_pre <- lm(mass_g ~ length_cm, data=fishdata_4R[fishdata_4R$time=="pre-fire",])
abline(lm_pre) #Adds the trendline to the regression scatterplot
```

4.2. LINEAR REGRESSION OF FISH MASS VS. LENGTH FOR BEFORE AND AFTER THE FIRE21

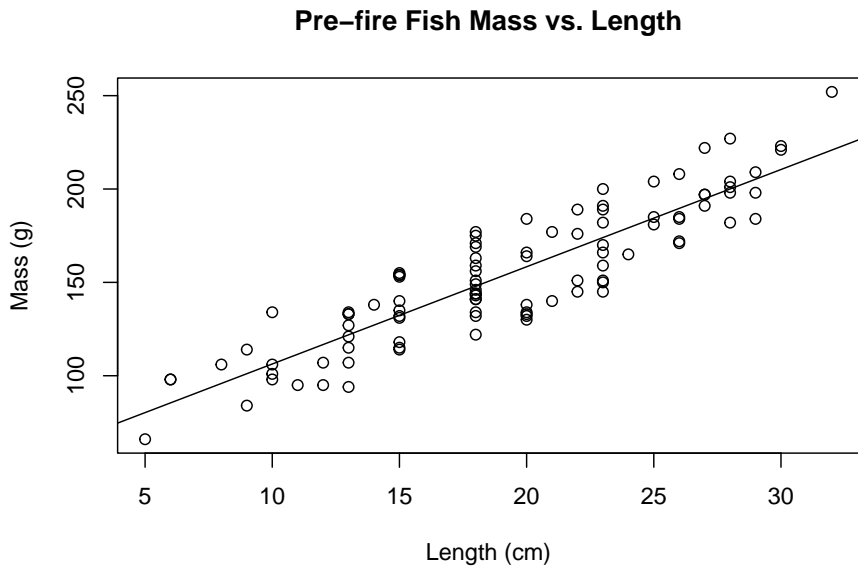


Figure 4.3: Scatterplot and linear regression line of fish length in centimeters versus fish mass in grams in Cache La Poudre in 2012 before the High Park Fire.

```
summary(aov(lm_pre)) #shows the results of the pre-fire linear regression ANOVA
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## length_cm    1 103690   103690     382 <2e-16 ***
## Residuals    98  26599      271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_pre) #shows equation of the line, multiple R-squared value
```

```
##
## Call:
## lm(formula = mass_g ~ length_cm, data = fishdata_4R[fishdata_4R$time ==
## "pre-fire", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.987 -14.472  -0.307   12.543   31.144
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.2113     5.3641   10.11  <2e-16 ***
## length_cm      5.2077     0.2664   19.55  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.47 on 98 degrees of freedom
## Multiple R-squared:  0.7958, Adjusted R-squared:  0.7938
## F-statistic:   382 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
# Post-fire
```

```
# Scatterplot of length and mass where length is the independent variable and mass is
plot(mass_g ~ length_cm, data=fishdata_4R[fishdata_4R$time=="post-fire",], xlab="Length",
title("Post-fire Fish Mass vs. Length"))
```

```
# Linear regression on mass vs.length
```

```
lm_post <- lm(mass_g ~ length_cm, data=fishdata_4R[fishdata_4R$time=="post-fire",])
abline(lm_post) #Adds the trendline to the regression scatterplot
```

```
summary(aov(lm_post)) #shows the results of the pre-fire linear regression ANOVA
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
```

4.2. LINEAR REGRESSION OF FISH MASS VS. LENGTH FOR BEFORE AND AFTER THE FIRE23

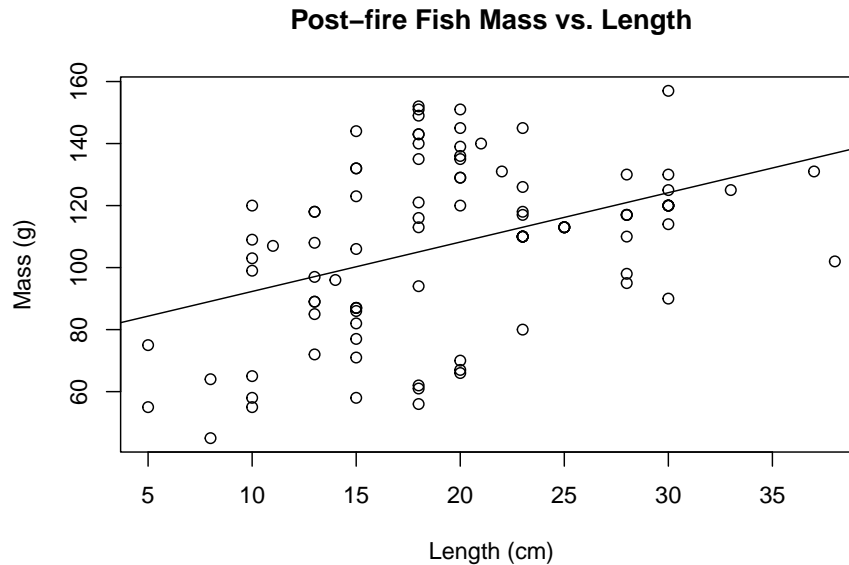


Figure 4.4: Scatterplot and linear regression line of fish length in centimeters versus fish mass in grams in Cache La Poudre in 2013 after the High Park Fire.

```
## length_cm    1  12126   12126    20.1 2.05e-05 ***
## Residuals   95  57313    603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm_post) #shows equation of the line, multiple R-squared value
```

```
##
## Call:
## lm(formula = mass_g ~ length_cm, data = fishdata_4R[fishdata_4R$time ==
## "post-fire", ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -49.048 -13.271  -3.011   19.582   46.952
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.3830     7.4498  10.253 < 2e-16 ***
## length_cm     1.5925     0.3552   4.483 2.05e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.56 on 95 degrees of freedom
## Multiple R-squared:  0.1746, Adjusted R-squared:  0.1659
## F-statistic: 20.1 on 1 and 95 DF,  p-value: 2.054e-05
```

```
#Pre- and Post-Fire on same graph
```

```
# First plot the pre-fire linear regression
```

```
# ylim sets the range of the y-axis; pch="+" makes points appear as plus signs; col="b"
plot(mass_g ~length_cm,data=fishdata_4R[fishdata_4R$time == "pre-fire",],xlab="Length",
title("Pre-Fire (+) and Post-Fire (o) Mass vs. Length"))
```

```
# Run linear regression of pre-fire mass and length to obtain the trend line.
```

```
lm_pre=lm(mass_g ~ length_cm,data=fishdata_4R[fishdata_4R$time == "pre-fire",])
abline(lm_pre,col="blue") #adds a trendline to the plot and makes the line blue
```

```
# Overlay the post-fire linear regression onto the plot of the pre-fire linear regression
```

```
# Plots post-fire data as o's and colors them red
```

```
points(mass_g ~length_cm,data=fishdata_4R[fishdata_4R$time == "post-fire",],xlab="Length",
```

```
# Run linear regression of post-fire mass and length to obtain the trend line.
```

```
lm_post=lm(mass_g ~ length_cm,data=fishdata_4R[fishdata_4R$time == "post-fire",])
abline(lm_post,col="red") #adds a trendline to the post-fire linear regression and m
```


4.2. LINEAR REGRESSION OF FISH MASS VS. LENGTH FOR BEFORE AND AFTER THE FIRE25

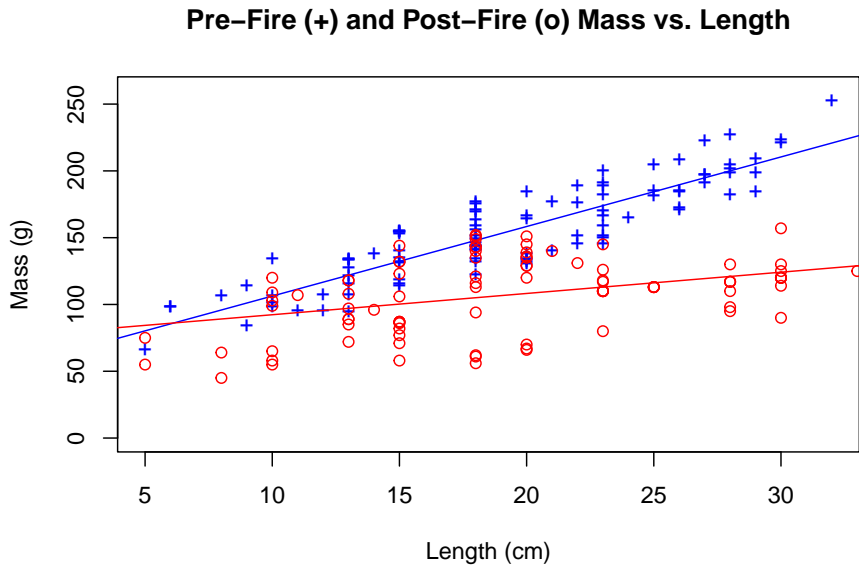


Figure 4.5: Scatterplot and linear regression line of fish length in centimeters versus fish mass in grams in Cache La Poudre in 2012 before the High Park Fire (blue, +) and in 2013 after the High Park Fire (red, o).

Chapter 5

Extracting and Visualizing Meteorological Data

“What do you call dangerous precipitation? A rain of terror.”

For this assignment, we used custom functions to read in and look at average meteorological data scraped from a public data archive.

Data is from Snowstudies.org. Assignment by Dr. Matthew Ross and Dr. Nathan Mueller of Colorado State University.

- 5.1 1. Extract the meteorological data URLs.**
Here we want you to use the `rvest` package to get the URLs for the `SASP forcing` and `SBSP_forcing` meteorological datasets.

```
# Read HTML page
snowarchive <- read_html("https://snowstudies.org/archived-data/")

# Read link with specific pattern
links <- snowarchive %>%
  html_nodes('a') %>% #look for links
  .[grepl('forcing',.)] %>% #filter to only links with "forcing" term
  html_attr('href') #tell it these are urls

links # view
```

```
## [1] "https://snowstudies.org/wp-content/uploads/2022/02/SBB_SASP_Forcing_Data.txt"
## [2] "https://snowstudies.org/wp-content/uploads/2022/02/SBB_SBSP_Forcing_Data.txt"
```

5.2 2. Download the meteorological data. Use the `download_file` and `str_split_fixed` commands to download the data and save it in your data folder. You can use a for loop or a map function.

```
# Grab only the name of the file by splitting out on forward slashes
splits <- str_split_fixed(links, '/', 8)
```

```
# Keep only the 8th column
files <- splits[, 8]
```

```
files
```

```
## [1] "SBB_SASP_Forcing_Data.txt" "SBB_SBSP_Forcing_Data.txt"
```

```
# Generate a file list for where the data goes
file_names <- paste0('Data_sci_bookdown/data/snow/', files)
```

```
# For loop that downloads each - i for every instance, length function tells how many
for(i in 1:length(file_names)){
  download.file(links[i], destfile=file_names[i])
}
```

```
# Download via map function
# map2(links, file_names, download.file)
```

```
# Map version of the for loop (downloading files)
downloaded <- file.exists(file_names)
evaluate <- !all(downloaded) # sees if files are downloaded (T/F)
if(evaluate == T){
  map2(links[1:2], file_names[1:2], download.file)
}else{print('data downloaded')}
```

```
## [1] "data downloaded"
```

5.3. 3. WRITE A CUSTOM FUNCTION TO READ IN THE DATA AND APPEND A SITE COLUMN TO THE DATA

5.3 3. Write a custom function to read in the data and append a site column to the data.

```
# Traditional read in

SASP <- read.csv("Data_sci_bookdown/data/snow/SBB_SASP_Forcing_Data.csv") %>%
  select(1,2,3,7,10)

colnames(SASP) <- c("year", "month", "day", "precip", "temp")

SBSP <- read.csv("Data_sci_bookdown/data/snow/SBB_SBSP_Forcing_Data.csv") %>%
  select(1,2,3,7,10)

colnames(SBSP) <- c("year", "month", "day", "precip", "temp")

# Combine csvs
alldata <- rbind(SASP, SBSP)

# Read in via new function

# Grab headers from metadata pdf
library(pdftools)

## Using poppler version 20.12.1

headers <- pdf_text('https://snowstudies.org/wp-content/uploads/2022/02/Serially-Complete-Metadata.pdf')
readr::read_lines(.) %>%
  trimws(.) %>%
  str_split_fixed(., '\\.', 2) %>%
  .[,2] %>%
  .[1:26] %>%
  str_trim(side = "left")
```

5.4 4. Use the map function to read in both meteorological files. Display a summary of your tibble.

```
# Pull site name out of the file name and read in the .txt files
read_data <- function(file){
```

```

name = str_split_fixed(file, '_', 2)[, 2] %>%
  gsub('_Forcing_Data.txt', '', .)
df <- read_fwf(file) %>%
  select(year=1, month=2, day=3, hour=4, precip=7, air_temp=10) %>% #choose and name
  mutate(site = name) #add column
}

alldata2 <- map_dfr(file_names, read_data)

```

```

## Rows: 69168 Columns: 19
## -- Column specification -----
##
## chr (2): X12, X14
## dbl (17): X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X13, X15, X16, X17, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 69168 Columns: 19
## -- Column specification -----
##
## chr (2): X12, X14
## dbl (17): X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X13, X15, X16, X17, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```

```
summary(alldata2)
```

```

##      year      month      day      hour
## Min.   :2003   Min.   : 1.000   Min.   : 1.00   Min.   : 0.00
## 1st Qu.:2005   1st Qu.: 3.000   1st Qu.: 8.00   1st Qu.: 5.75
## Median :2007   Median : 6.000   Median :16.00   Median :11.50
## Mean   :2007   Mean   : 6.472   Mean   :15.76   Mean   :11.50
## 3rd Qu.:2009   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:17.25
## Max.   :2011   Max.   :12.000   Max.   :31.00   Max.   :23.00
##      precip      air_temp      site
## Min.   :0.000e+00   Min.   :242.1   Length:138336
## 1st Qu.:0.000e+00   1st Qu.:265.8   Class :character
## Median :0.000e+00   Median :272.6   Mode  :character
## Mean   :3.838e-05   Mean   :272.6
## 3rd Qu.:0.000e+00   3rd Qu.:279.7
## Max.   :6.111e-03   Max.   :295.8

```

5.5. 5. MAKE A LINE PLOT OF MEAN TEMP BY YEAR BY SITE (USING THE AIR TEMP [K] VARIABLE). IS

5.5 5. Make a line plot of mean temp by year by site (using the air temp [K] variable). Is there anything suspicious in the plot? Adjust your filtering if needed.

```
temp_yearly <- alldata2 %>%
group_by(year, site) %>%
summarise(mean_temp = mean(`air_temp`, na.rm=T))
```

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups`
## argument.
```

```
ggplot(temp_yearly, aes(x=year, y=mean_temp, color=site)) +
  geom_point() + geom_line() +
  xlab("Year") + ylab("Mean Temperature (Degrees Kelvin)") +
  ggthemes::theme_few() +
  scale_color_brewer(palette = "Set2") +
  scale_x_continuous(breaks = pretty(c(2003,2012), n = 6)) +
  theme(legend.position="bottom")
```

5.6 6. Write a function that makes line plots of monthly average temperature at each site for a given year. Use a for loop to make these plots for 2005 to 2010.

```
temp_monthly <- alldata2 %>%
  group_by(year, month, site) %>%
  summarize(mean_temp = mean(`air_temp`, na.rm=T))
```

```
## `summarise()` has grouped output by 'year', 'month'. You can override using the
## `.groups` argument.
```

```
par(mfrow=c(5,1))

plot_monthly <- function(year.no) {
  plot <- temp_monthly %>%
```

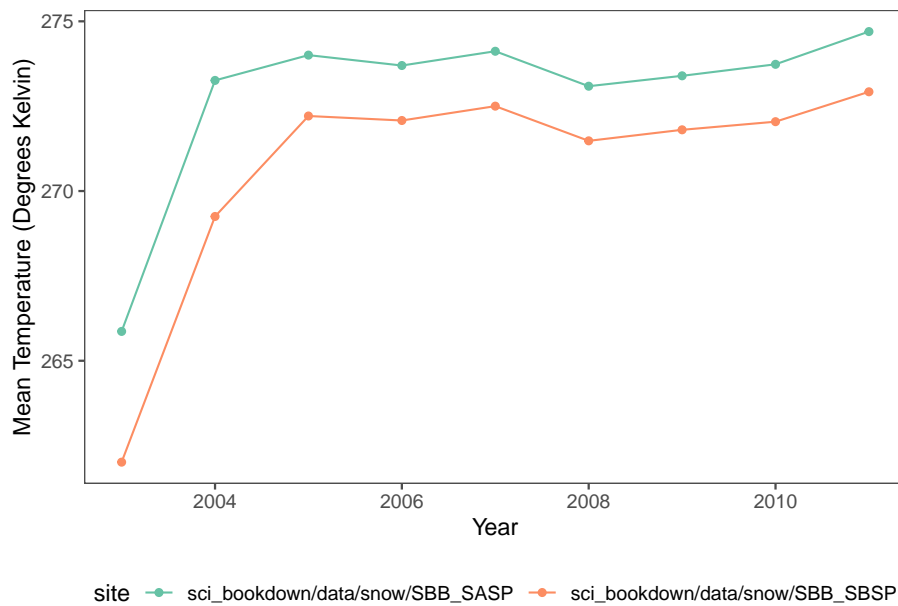


Figure 5.1: Mean temperature of the SASP (teal) and SBSP (orange) sites from 2003 to 2012, in degrees Kelvin.

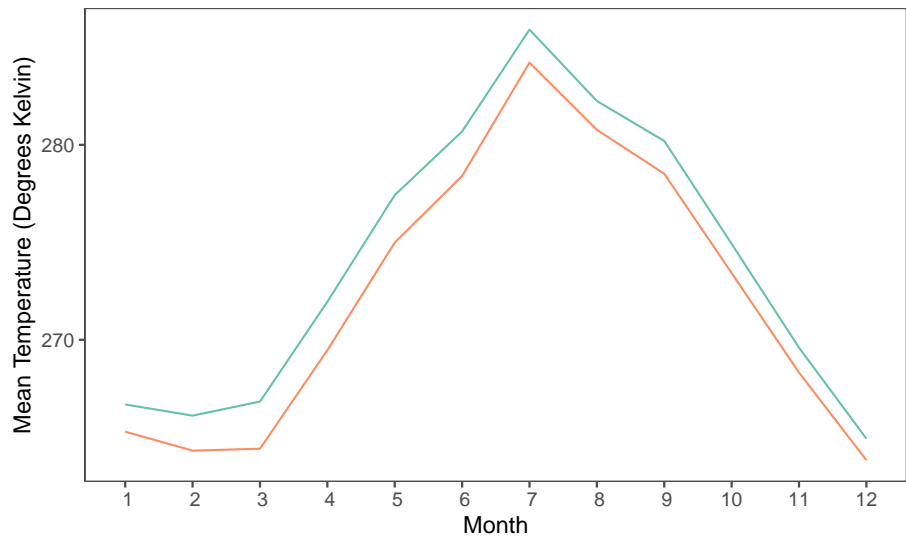
```

filter(year == year.no) %>%
ggplot(aes(x=month, y=mean_temp, color=site)) +
  geom_line() +
  xlab("Month") + ylab("Mean Temperature (Degrees Kelvin)") +
  ggthemes::theme_few() +
  scale_color_brewer(palette = "Set2") +
  scale_x_discrete(limits = c(1,2,3,4,5,6,7,8,9,10,11,12)) +
  scale_y_continuous(breaks = pretty(c(255,290), n = 4)) +
  theme(legend.position="bottom")
print(plot)
}

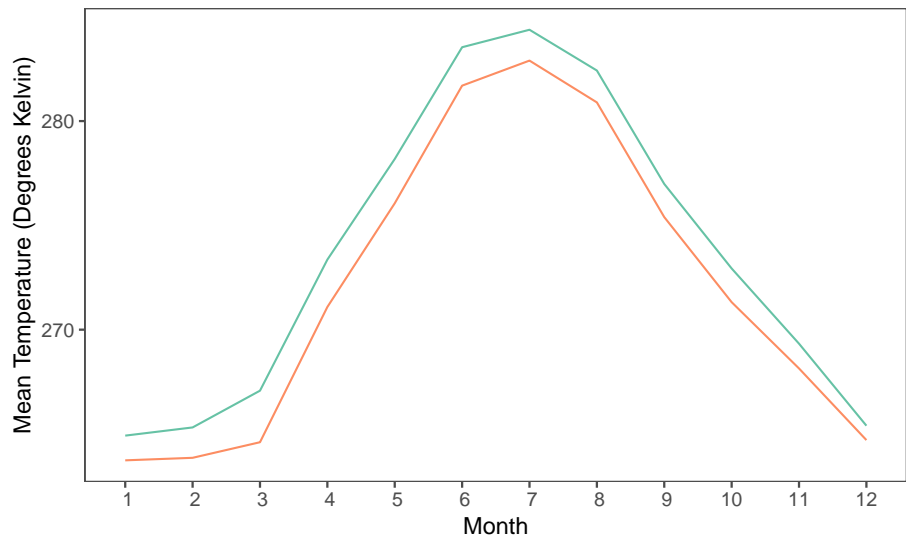
for(i in 2005:2010){
  plot_monthly(i)
}

```

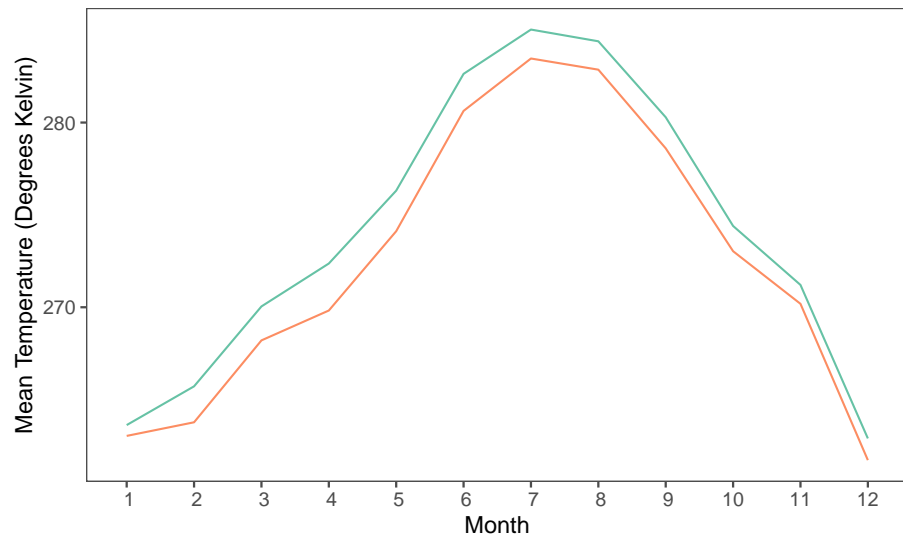

5.6. 6. WRITE A FUNCTION THAT MAKES LINE PLOTS OF MONTHLY AVERAGE TEMPERATURE AT EA



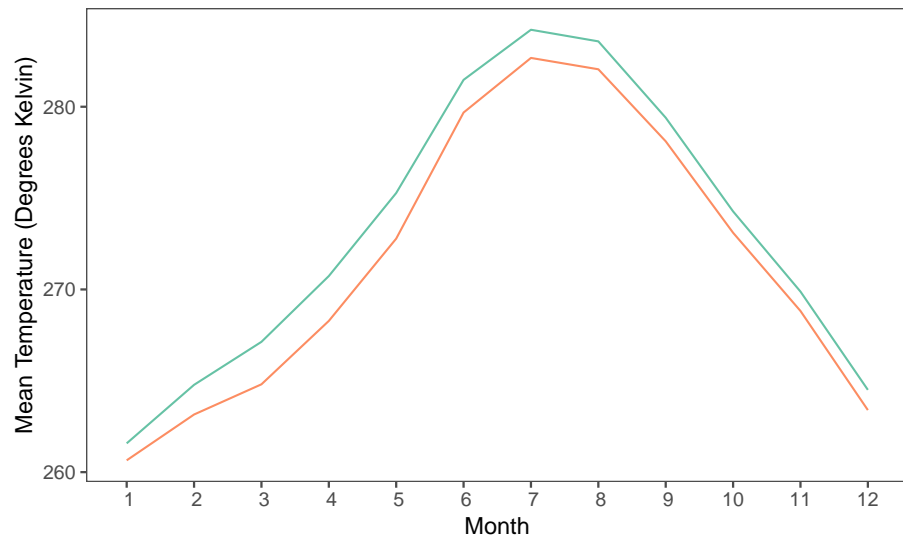
site sci_bookdown/data/snow/SBB_SASP sci_bookdown/data/snow/SBB_SBSP



site sci_bookdown/data/snow/SBB_SASP sci_bookdown/data/snow/SBB_SBSP

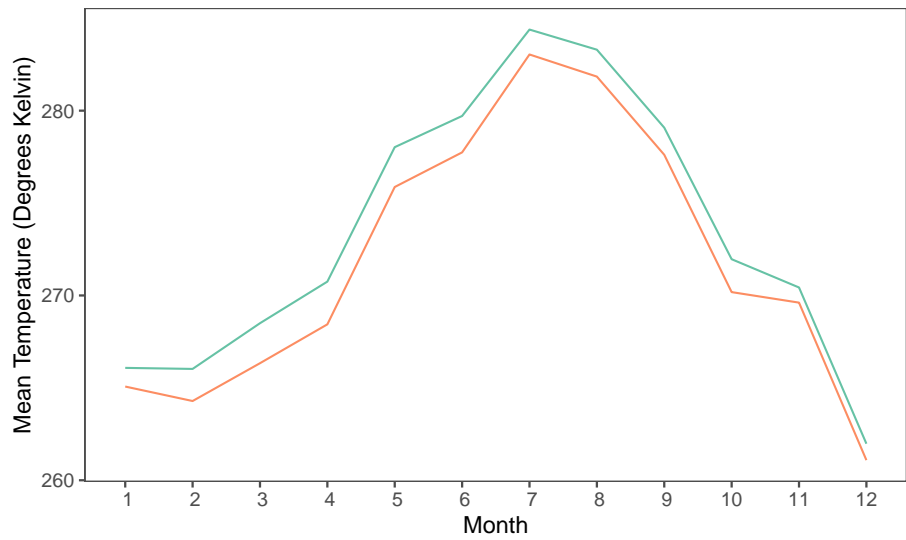


site — sci_bookdown/data/snow/SBB_SASP — sci_bookdown/data/snow/SBB_SBSP

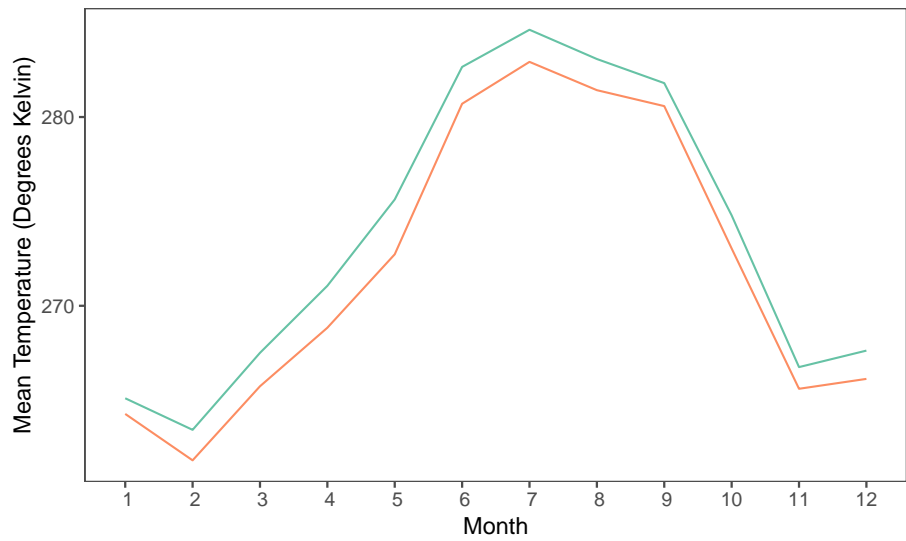


site — sci_bookdown/data/snow/SBB_SASP — sci_bookdown/data/snow/SBB_SBSP

5.6. 6. WRITE A FUNCTION THAT MAKES LINE PLOTS OF MONTHLY AVERAGE TEMPERATURE AT EA



site sci_bookdown/data/snow/SBB_SASP sci_bookdown/data/snow/SBB_SBSP



site sci_bookdown/data/snow/SBB_SASP sci_bookdown/data/snow/SBB_SBSP

5.7 Bonus: Make a plot of average daily precipitation by day of year (averaged across all available years)

```
precip_daily <- alldata2 %>%
  mutate(date = make_date(year, month, day),
         day_no = yday(date)) %>%
  group_by(day_no) %>%
  summarize(mean_precip = mean(`precip`*86400, na.rm=T))

ggplot(precip_daily, aes(x=day_no, y=mean_precip)) +
  geom_line() +
  xlab("Day of Year") + ylab("Mean Precipitation (mm/day)") +
  ggthemes::theme_few() +
  scale_color_brewer(palette = "Set2") +
  scale_y_continuous(breaks = pretty(c(0,14), n = 7)) +
  scale_x_continuous(breaks = pretty(c(1,365), n = 8))
```

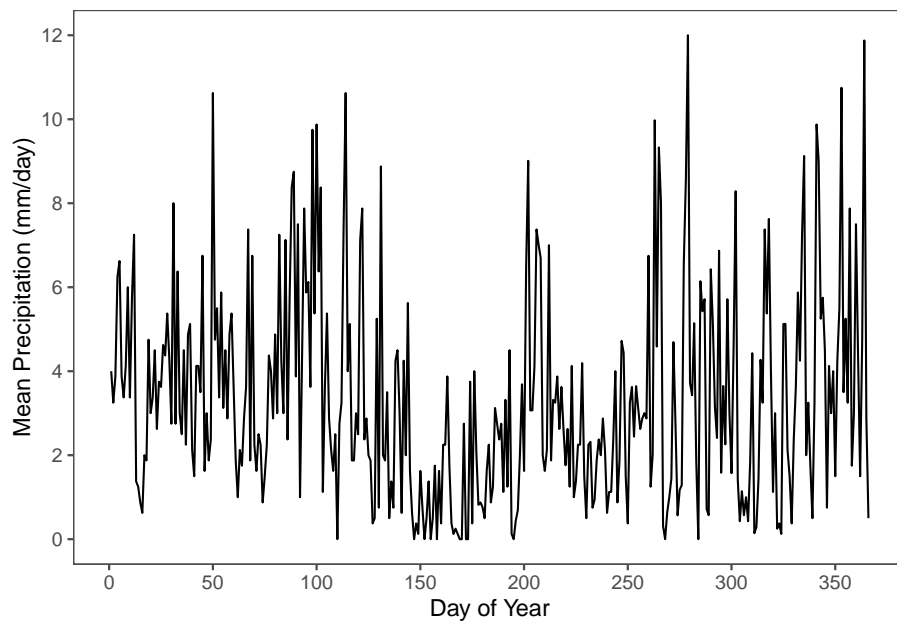


Figure 5.2: Mean daily precipitation by day of year, averaged from 2003 to 2012.

Chapter 6

Spatial Analysis in R

“Why are latitude and longitude so smart? Because they have so many degrees!”

In this assignment, I learned to use R for spatial analyses.

Data is from the LAGOS dataset. Assignment by Dr. Matthew Ross and Dr. Nathan Mueller of Colorado State University.

6.1 Loading in data

6.1.1 First download and then specifically grab the locus (or site lat longs)

```
# #Lagos download script  
#LAGOSNE::lagosne_get(dest_folder = LAGOSNE::lagos_path(), overwrite = TRUE)
```

```
#Load in lagos  
lagos <- lagosne_load()
```

```
## Warning in (function (version = NULL, fpath = NA) : LAGOSNE version unspecified,  
## loading version: 1.087.3
```

```
#Grab the lake centroid info  
lake_centers <- lagos$locus
```

```
# Make an sf object
spatial_lakes <- st_as_sf(lake_centers, coords=c('nhd_long', 'nhd_lat'),
                          crs=4326)

#Grab the water quality data
nutr <- lagos$epi_nutr

#Look at column names
#names(nutr)
```

6.1.2 Convert to spatial data

```
#Look at the column names
#names(lake_centers)

#Look at the structure
#str(lake_centers)

#View the full dataset
#View(lake_centers %>% slice(1:100))

spatial_lakes <- st_as_sf(x = lake_centers, coords = c("nhd_long", "nhd_lat"), crs = 4326)
  st_transform(2163)

#mapview(spatial_lakes)

#Subset for plotting
subset_spatial <- spatial_lakes %>%
  slice(1:100)

subset_baser <- spatial_lakes[1:100,]

#Dynamic mapviewer
#mapview(subset_spatial)
```

6.1.3 Subset to only Minnesota

```
states <- us_states()

#Plot all the states to check if they loaded
#mapview(states)
```

```

minnesota <- states %>%
  filter(name == 'Minnesota') %>%
  st_transform(2163)
#mapview(minnesota)

#Subset lakes based on spatial position
minnesota_lakes <- spatial_lakes[minnesota,]

#Plotting the first 1000 lakes
minnesota_lakes %>%
  arrange(-lake_area_ha) %>%
  slice(1:1000)

## Simple feature collection with 1000 features and 16 fields
## Geometry type: POINT
## Dimension: XY
## Bounding box: xmin: 254441 ymin: -154522.4 xmax: 755222.3 ymax: 464949.4
## Projected CRS: NAD27 / US National Atlas Equal Area
## First 10 features:
##   lagoslakeid  nhdid      gnis_name lake_area_ha lake_perim_meters
## 1      15162 123319728 Lake of the Woods 123779.817      401005.02
## 2      34986 105567868 Lower Red Lake 66650.332      115825.47
## 3      2498 120019294 Mille Lacs Lake 51867.225      151701.94
## 4      39213 105567402 Upper Red Lake 48288.325      99828.05
## 5       996 120018981 Leech Lake 41824.352      344259.98
## 6       583 120019513 Lake Winnibigoshish 22566.124      86722.10
## 7        73 120019354 Rainy Lake 18522.551      660313.32
## 8      2554 105954753 Vermilion Lake 15736.590      509617.01
## 9      2161 120019371 Kabetogama Lake 9037.249      288750.31
## 10     3119 166868528 Cass Lake 8375.173      85326.14
##   nhd_fcode nhd_ftype iws_zoneid hu4_zoneid hu6_zoneid hu8_zoneid hu12_zoneid
## 1      39004      390 IWS_37547 HU4_26 HU6_36 HU8_468 HU12_13912
## 2      39004      390 IWS_34899 HU4_54 HU6_74 HU8_327 HU12_14600
## 3      39004      390 IWS_22933 HU4_25 HU6_73 HU8_344 HU12_10875
## 4      39004      390 IWS_33471 HU4_54 HU6_74 HU8_327 HU12_14204
## 5      39004      390 IWS_23572 HU4_25 HU6_35 HU8_332 HU12_14479
## 6      39004      390 IWS_22455 HU4_25 HU6_35 HU8_331 HU12_14543
## 7      39004      390 IWS_37542 HU4_26 HU6_36 HU8_473 HU12_13942
## 8      39004      390 IWS_36424 HU4_26 HU6_36 HU8_131 HU12_14405
## 9      39004      390 IWS_36301 HU4_26 HU6_36 HU8_130 HU12_14395
## 10     39004      390 IWS_21080 HU4_25 HU6_35 HU8_331 HU12_13957
##   edu_zoneid county_zoneid state_zoneid elevation_m geometry
## 1      EDU_56 County_435 State_14 323.5090 POINT (366706.2 464949.4)
## 2      EDU_16 County_455 State_14 358.1656 POINT (371974.2 341706.5)
## 3      EDU_43 County_484 State_14 381.7920 POINT (489582.1 157109.5)

```

```
## 4      EDU_16      County_455      State_14      358.3096 POINT (389013.3 360819.5)
## 5      EDU_42      County_424      State_14      395.2420 POINT (422409.7 255724.9)
## 6      EDU_42      County_424      State_14      396.1560 POINT (437872.1 286675)
## 7      EDU_55      County_446      State_14      338.0670 POINT (515833.6 420274.2)
## 8      EDU_3       County_446      State_14      414.1680 POINT (566966.7 347059.1)
## 9      EDU_55      County_446      State_14      339.2530 POINT (519199.2 408290.2)
## 10     EDU_42      County_424      State_14      396.7710 POINT (410563.2 281005.2)
```

```
#mapview(., zcol = 'lake_area_ha')
```

6.2 Part one

6.2.1 Show a map outline of Iowa and Illinois (similar to Minnesota map upstream)

```
Istates <- states %>%
  filter(name == 'Iowa' | name == 'Illinois') %>%
  st_transform(2163)
mapview(Istates, canvas = TRUE)
```

6.2.2 Subset LAGOS data to these sites, how many sites are in Illinois and Iowa combined? How does this compare to Minnesota?

```
Istates_lakes <- spatial_lakes[Istates,]
nrow(Istates_lakes)
```

```
## [1] 16466
```

```
Istates_count <- length(Istates_lakes$lagoslakeid)
nrow(minnesota_lakes)
```

```
## [1] 29038
```



```
Minn_count <- length(minnesota_lakes$lagoslakeid)
```

Iowa and Illinois have 16466 lakes combined, much less than the number of lakes that Minnesota alone has, 29038.

6.2.3 What is the distribution of lake size in Iowa vs. Minnesota?

- Here I want to see a histogram plot with lake size on x-axis and frequency on y axis (check out `geom_histogram`)

```
iowa <- states %>%
  filter(name == 'Iowa') %>%
  st_transform(2163)

iowa_lakes <- spatial_lakes[iowa,]

combined <- rbind(iowa_lakes, minnesota_lakes)

ggplot(combined, aes(x= lake_area_ha)) +
  ggthemes::theme_few() + theme(legend.position="bottom") +
  xlab("Lake Area (ha)") + ylab("Count") +
  scale_x_continuous(trans = "log10", labels = scales::comma) +
  geom_histogram(data = minnesota_lakes, color = "red", alpha = 0.2) +
  geom_histogram(data = iowa_lakes, color = "blue", alpha = 0.2) +
  scale_fill_manual(values=c("blue", "red"), "State")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

6.2.4 Make an interactive plot of lakes in Iowa and Illinois and color them by lake area in hectares

```
Istates_map = Istates_lakes %>%
  arrange(-lake_area_ha) %>%
  slice(1:1000)

mapview(Istates_map, zcol = 'lake_area_ha', canvas = TRUE)
```

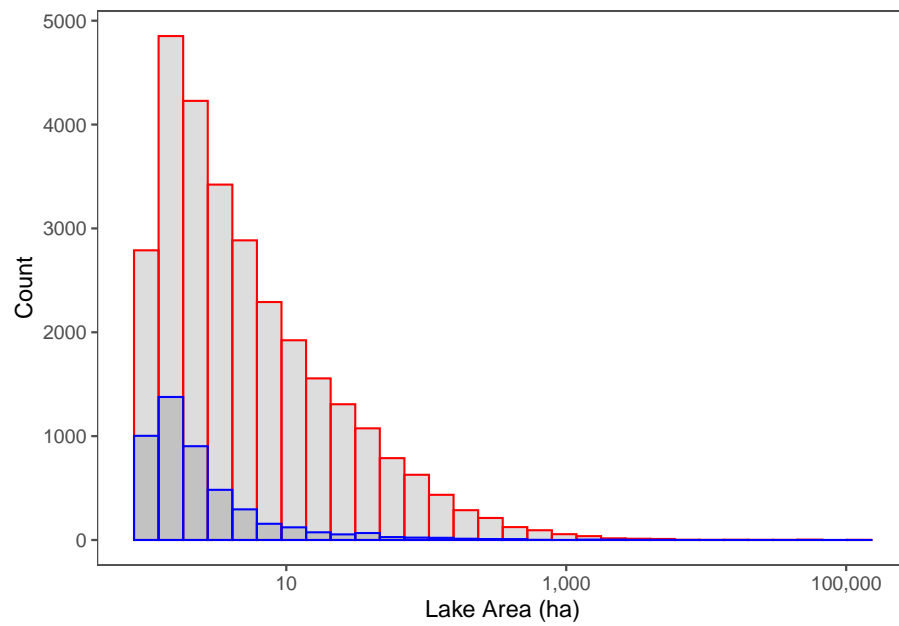


Figure 6.1: The number of lakes with a given area, in hectares, in Minnesota (red) and Iowa (blue).

6.2.5 What other data sources might we use to understand how reservoirs and natural lakes vary in size in these three states?

We might use the US Geological Survey (USGS) National Water Informational System (NWIS) and its National Water Dashboard as a data source, and look at gage height (indicating lake depth) as another parameter for lake size variation. The USGS National Hydrography Dataset (NHD) is another data source that would, similarly to Lagos, give us a surface area metric for lakes in the various states.

6.3 Part two

6.3.1 Subsets

6.3.1.1 Columns nutr to only keep key info that we want

```
clarity_only <- nutr %>%
  dplyr::select(lagoslakeid,sampleddate,chla,doc,secchi) %>%
  mutate(sampledate = as.character(sampledate) %>% ymd(.))
```

6.3.1.2 Keep sites with at least 200 observations

```
#Look at the number of rows of dataset
nrow(clarity_only)

chla_secchi <- clarity_only %>%
  filter(!is.na(chla),
         !is.na(secchi))

# How many observations did we lose?
# nrow(clarity_only) - nrow(chla_secchi)

# Keep only the lakes with at least 200 observations of secchi and chla
chla_secchi_200 <- chla_secchi %>%
  group_by(lagoslakeid) %>%
  mutate(count = n()) %>%
  filter(count > 200)
```

6.3.1.3 Join water quality data to spatial data

```
spatial_200 <- inner_join(spatial_lakes, chla_secchi_200 %>%
  distinct(lagoslakeid, .keep_all=T),
  by='lagoslakeid')
```

6.3.2 Mean Chlorophyll A map

```
### Take the mean chl_a and secchi by lake

means_200 <- chla_secchi_200 %>%
  # Take summary by lake id
  group_by(lagoslakeid) %>%
  # take mean chl_a per lake id
  summarize(mean_chl = mean(chla, na.rm=T),
             mean_secchi=mean(secchi, na.rm=T)) %>%
  #Get rid of NAs
  filter(!is.na(mean_chl),
         !is.na(mean_secchi)) %>%
  # Take the log base 10 of the mean_chl
  mutate(log10_mean_chl = log10(mean_chl))

#Join datasets
mean_spatial <- inner_join(spatial_lakes, means_200,
  by='lagoslakeid')

#Make a map
mapview(mean_spatial, zcol='log10_mean_chl', layer.name = "Mean Chlorophyll A Content")
```

6.3.3 What is the correlation between Secchi Disk Depth and Chlorophyll a for sites with at least 200 observations?

```
ggplot(means_200) +
  geom_point(aes(mean_secchi, mean_chl)) +
  ggthemes::theme_few() +
  xlab("Mean Secchi Disk Depth") + ylab("Mean Chlorophyll Content")
```

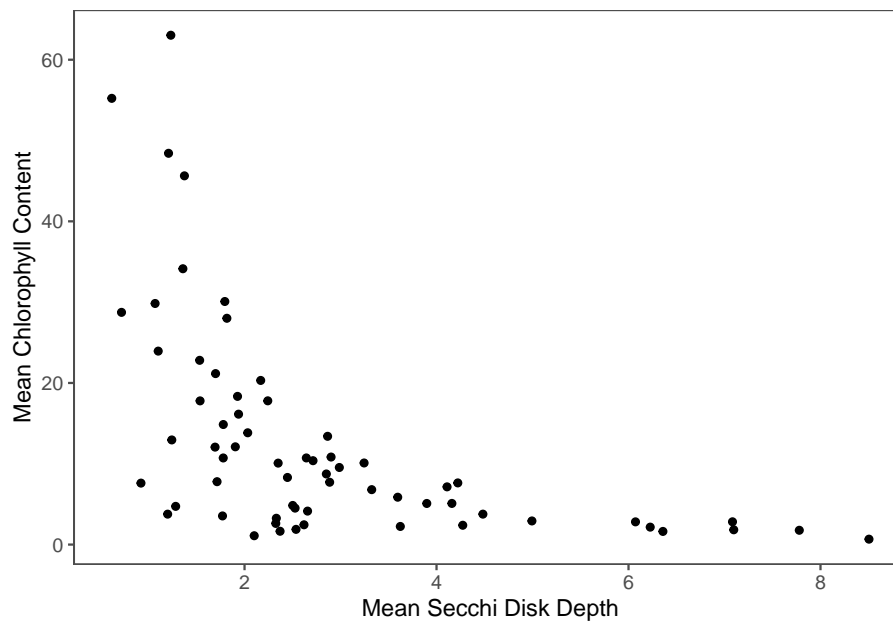


Figure 6.2: Chlorophyll content has a negative correlation with Secchi disk depth at sites with at least 200 observations.

6.3.3.1 Why might this be the case?

Secchi disks measure water clarity; the deeper the disk, the clearer the water (1). Chlorophyll content in lakes is generally a reliable marker of algae content, so that high chlorophyll values indicate high algal biomass and corresponding low water clarity (2). Additionally, chlorophyll may be used as a proxy for water quality, since high algal biomass is associated with high nutrient pollution in the process of eutrophication (2). High pollution may further decrease water clarity, so that the relationship between chlorophyll and Secchi disk depth may be expected.

1. “The Secchi Dip-in - What Is a Secchi Disk?” North American Lake Management Society (NALMS), <https://www.nalms.org/secchidipin/monitoring-methods/the-secchi-disk/what-is-a-secchi-disk/>.
2. Filazzola, A., Mahdian, O., Shuvo, A. et al. A database of chlorophyll and water chemistry in freshwater lakes. *Sci Data* 7, 310 (2020). <https://doi-org.ezproxy2.library.colostate.edu/10.1038/s41597-020-00648-2>

6.3.4 What states have the most data?

6.3.4.1 Make a lagos spatial dataset that has the total number of counts per site.

```
site_counts <- chla_secchi %>%
  group_by(lagoslakeid) %>%
  mutate(count = n())

lake_counts <- inner_join(site_counts, lake_centers, by= "lagoslakeid")%>%
  dplyr::select(lagoslakeid,nhd_long,nhd_lat, count, secchi, chla)

spatial_counts <- st_as_sf(lake_counts,coords=c("nhd_long","nhd_lat"),
                           crs=4326)
```

6.3.4.2 Join this point dataset to the us_boundaries data.

```
states <- us_states()

states_counts <- st_join(spatial_counts, states)
```

6.3.4.3 Group by state and sum all the observations in that state and arrange that data from most to least total observations per state.

```
sum_statecount <- states_counts %>%
  group_by(state_name) %>%
  summarize(sum = sum(count)) %>%
  arrange(desc(sum))

sumtable <- tibble(sum_statecount)

view(sumtable)

#ggplot(data = sumtable, aes(x=state_name, y=sum, fill=state_name)) +
#  geom_bar(stat = "identity", width = 0.3, position = "dodge") +
#  ggthemes::theme_few() +
#  xlab("State") + ylab(expression(paste("# of Observations")))
```

Minnesota has the most observations. Vermont, has the next most observations, but less than half of Minnesota's observations. South Dakota has the least number of observations in the dataset.

6.3.5 Is there a spatial pattern in Secchi disk depth for lakes with at least 200 observations?

```
mapview(mean_spatial, zcol='mean_secchi', layer.name = "Mean Secchi Disk Depth")
```


Chapter 7

Linear Regressions, Quadratic Fits, Residuals, and Spatial

“How many data scientists does it take to change a light bulb? That depends. It is really a matter of power.”

This assignment combined several methods to look at relationships between crop yields and weather data over time.

Data from USDA National Agricultural Statistical Service (NASS). Assignment by Dr. Matthew Ross and Dr. Nathan Mueller of Colorado State University.

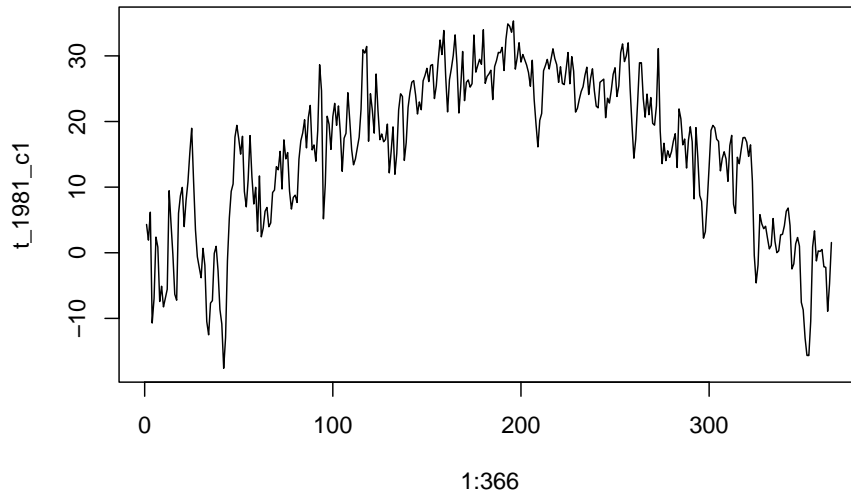
7.1 Weather Data Analysis

7.1.1 Load the PRISM daily maximum temperatures

```
# daily max temperature  
# dimensions: counties x days x years  
prism <- readMat("Data_sci_bookdown/data/prismiowa.mat")  
  
# look at county #1  
t_1981_c1 <- prism$tmaxdaily.iowa[1,,1] #first county, all days, first year  
t_1981_c1[366] #check for leap year (366 days)
```

```
## [1] NaN
```

```
plot(1:366, t_1981_c1, type = "l") #base r plot
```



```
# assign dimension names to tmax matrix
dimnames(prism$tmaxdaily.iowa) <- list(prism$COUNTYFP, 1:366, prism$years) #add dimensions

# converted 3d matrix into a data frame
tmaxdf <- as.data.frame.table(prism$tmaxdaily.iowa)

# relabel the columns
colnames(tmaxdf) <- c("countyfp", "doy", "year", "tmax") #name columns
tmaxdf <- tibble(tmaxdf) #tidyverse table
```

7.1.2 Download NASS corn yield data

```
# set our API key with NASS
nassqs_auth(key = "B9113AF8-85C4-3CEE-8D93-6E885D49E24F") #Here put in API code from U

# parameters to query on
params <- list(commodity_desc = "CORN", util_practice_desc = "GRAIN", prodn_practice_desc = "CORN")

# download
cornyieldsall <- nassqs_yields(params)
```

7.2. EXTRACT WINNESHIEK COUNTY CORN YIELDS, FIT A LINEAR TIME TREND, MAKE A PLOT. IS TH

```
## |  
  
cornyieldsall$county_ansi <- as.numeric(cornyieldsall$county_ansi)  
cornyieldsall$yield <- as.numeric(cornyieldsall$Value)  
  
# clean and filter this dataset  
cornyields <- select(cornyieldsall, county_ansi, county_name, yield, year) %>%  
  filter(!is.na(county_ansi) & !is.na(yield))  
cornyields <- tibble(cornyields)
```

7.2 Extract Winneshiek County corn yields, fit a linear time trend, make a plot. Is there a significant time trend?

```
winneccorn <- cornyields %>%  
  filter(county_ansi == "191")  
  
cornlm <- lm(yield ~ year, data = winneccorn)  
summary(cornlm) #P=1.77e-13 R^2= 0.755  
  
##  
## Call:  
## lm(formula = yield ~ year, data = winneccorn)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -51.163  -1.841   2.363   9.437  24.376   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept) -4763.290    448.286  -10.63 4.46e-13 ***  
## year         2.457       0.224   10.96 1.77e-13 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 16.97 on 39 degrees of freedom  
## Multiple R-squared:  0.7551, Adjusted R-squared:  0.7488   
## F-statistic: 120.2 on 1 and 39 DF,  p-value: 1.767e-13
```

```
ggplot(winneccorn, mapping = aes(x = year, y = yield)) +
  geom_point() +
  theme_bw() +
  labs(x = "Year", y = "Corn Yield") +
  geom_smooth(method = lm, se=TRUE, color="#78917E", fill="#C5DDB3")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

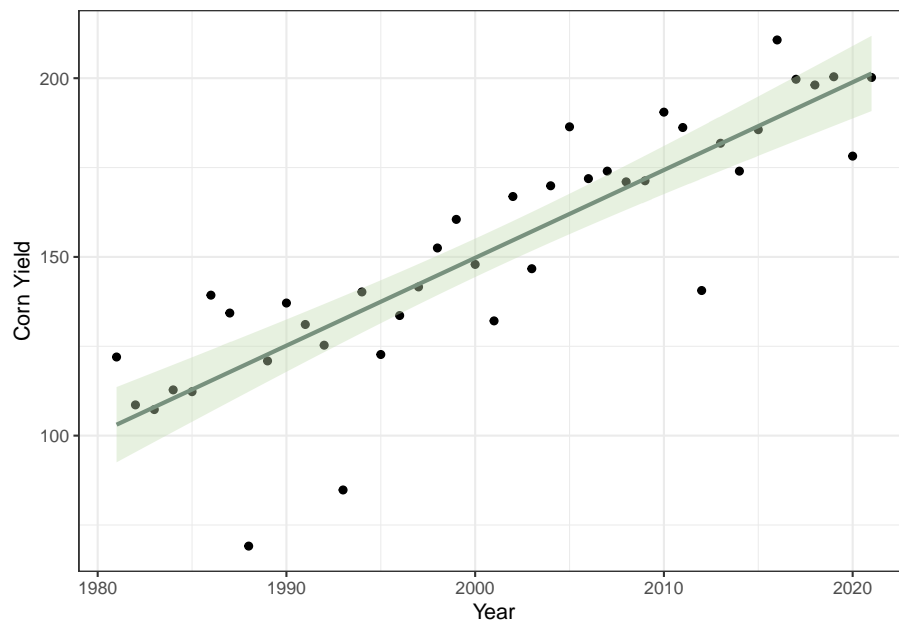


Figure 7.1: Linear regression of corn yields over time (years) in Winneshieck County, Iowa.

There is a significant positive correlation between corn yields and years in Winneshieck County, with an R-squared value of 0.755 and a P-value of 1.77e-13.

7.3 Fit a quadratic time trend (i.e., $\text{year} + \text{year}^2$) and make a plot. Is there evidence for slowing yield growth?

7.3. FIT A QUADRATIC TIME TREND (I.E., $\text{YEAR} + \text{YEAR}^2$) AND MAKE A PLOT. IS THERE EVIDENCE FOR

```
winnecorn$yearsq <- winnecorn$year^2 #square explanatory variables for quadratic

lm_cornquad <- lm(yield ~ year + yearsq, winnecorn)
summary(lm_cornquad)
```

```
##
## Call:
## lm(formula = yield ~ year + yearsq, data = winnecorn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -51.384  -3.115   1.388   9.743  25.324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.583e+04  8.580e+04   0.301   0.765
## year        -2.812e+01  8.576e+01  -0.328   0.745
## yearsq         7.641e-03  2.143e-02   0.357   0.723
##
## Residual standard error: 17.17 on 38 degrees of freedom
## Multiple R-squared:  0.7559, Adjusted R-squared:  0.7431
## F-statistic: 58.84 on 2 and 38 DF,  p-value: 2.311e-12
```

```
winnecorn$y_fitted <- lm_cornquad$fitted.values

#with the fitted values, create a non-linear trend
ggplot(winnecorn) +
  geom_point(mapping = aes(x = year, y = yield)) +
  geom_line(mapping = aes(x = year, y = y_fitted)) +
  theme_bw() +
  labs(x = "Year", y = "Corn Yield")
```

When we fit a quadratic line to the data, we find that it follows very closely to a linear regression, suggesting a fairly linear relationship between corn yields and years in Winneshiek County. There is no evidence of slowing yield growth in the model.

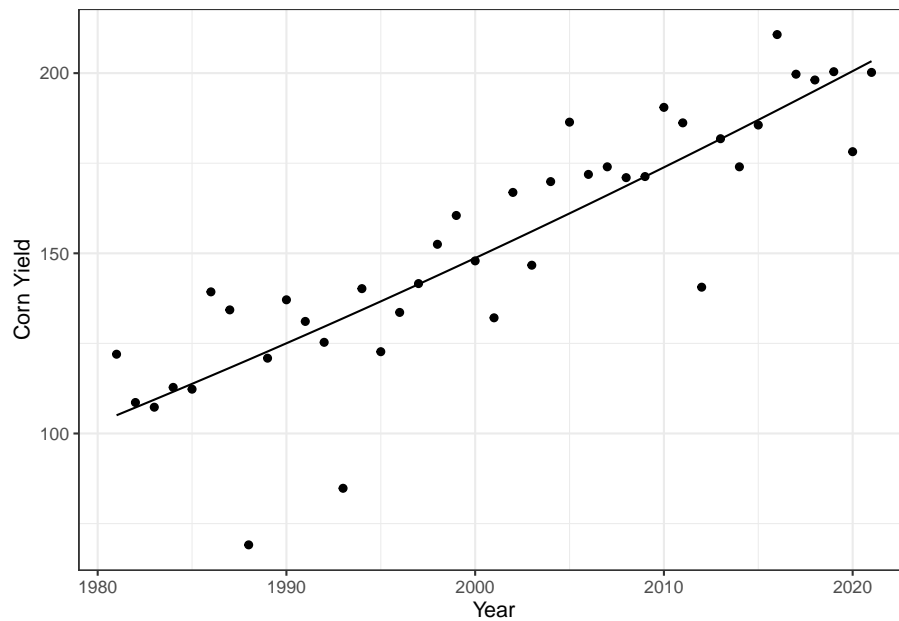


Figure 7.2: Quadratic fit of corn yields over time (years) in Winneshiek County, Iowa.

7.4 Time Series: Let's analyze the relationship between temperature and yields for the Winneshiek County time series. Use data on yield and summer avg Tmax. Is adding year or T_{\max}^2 to your model helpful? Make a plot and interpret the results.

```
# Winneshiek County summer temp maxes
tmaxdf$doy <- as.numeric(tmaxdf$doy)
tmaxdf$year <- as.numeric(as.character(tmaxdf$year))
tmaxdf$tmax <- as.numeric(tmaxdf$tmax)

winnesummer <- tmaxdf %>%
  filter(countyfp==191 & doym >= 152 & doym <= 243) %>% #day 152= June 1, 243= Aug 31
  group_by(year) %>%
  summarize(meantmax = mean(tmax))
```

7.4. TIME SERIES: LET'S ANALYZE THE RELATIONSHIP BETWEEN TEMPERATURE AND YIELDS FOR T

```
lm_summertmax <- lm(meantmax ~ year, winnesummer)
summary(lm_summertmax) #not sig
```

```
##
## Call:
## lm(formula = meantmax ~ year, data = winnesummer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.5189 -0.7867 -0.0341  0.6859  3.7415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.57670    36.44848   1.141   0.262
## year         -0.00747     0.01823  -0.410   0.684
##
## Residual standard error: 1.232 on 36 degrees of freedom
## Multiple R-squared:  0.004644,    Adjusted R-squared:  -0.02301
## F-statistic: 0.168 on 1 and 36 DF,  p-value: 0.6844
```

```
winnesummer$yearsq <- winnesummer$year^2 #square explanatory variables for quadratic
winnesummer$tmassq <- winnesummer$meantmax^2
```

```
lm_summerquad <- lm(meantmax ~ year + yearsq, winnesummer)
summary(lm_summerquad)
```

```
##
## Call:
## lm(formula = meantmax ~ year + yearsq, data = winnesummer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4617 -0.8812 -0.0530  0.7204  3.7308
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.618e+03  7.519e+03   0.481   0.633
## year        -3.585e+00  7.521e+00  -0.477   0.637
## yearsq       8.946e-04  1.881e-03   0.476   0.637
##
## Residual standard error: 1.246 on 35 degrees of freedom
## Multiple R-squared:  0.01104,    Adjusted R-squared:  -0.04547
## F-statistic: 0.1953 on 2 and 35 DF,  p-value: 0.8235
```

```

winnesummer$t_fitted <- lm_summerquad$fitted.values

# Join yield and temp data
winne <- inner_join(winnecorn, winnesummer)

## Joining, by = c("year", "yearsq")

lmwinne <- lm(yield ~ yearsq + tmaxsq, data = winne)
summary(lmwinne)

##
## Call:
## lm(formula = yield ~ yearsq + tmaxsq, data = winne)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.353  -7.496   2.089   9.806  27.874
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.314e+03  2.557e+02  -9.047 1.09e-10 ***
## yearsq       6.274e-04  6.295e-05   9.968 9.22e-12 ***
## tmaxsq      -6.445e-02  4.245e-02  -1.518  0.138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.97 on 35 degrees of freedom
## Multiple R-squared:  0.7492, Adjusted R-squared:  0.7349
## F-statistic: 52.28 on 2 and 35 DF,  p-value: 3.074e-11

winne$allfit <- lmwinne$fitted.values

ggplot(winne) +
  geom_point(mapping = aes(x = year, y = yield)) +
  geom_line(mapping = aes(x = year, y = allfit, color="red")) +
  geom_line(mapping = aes(x = year, y = y_fitted, color="blue")) +
  theme_bw() +
  scale_colour_manual(name = "Model",
    values = c("red"="red", "blue"="blue"), labels = c("Fit with Max Temp and Year"
  labs(x = "Year", y = "Corn Yield")

```

Adding maximum temperature trends to the model shows a similar trend, but peaks and dips in the fitted line highlight some of the outlying yield values and

7.4. TIME SERIES: LET'S ANALYZE THE RELATIONSHIP BETWEEN TEMPERATURE AND YIELDS FOR T

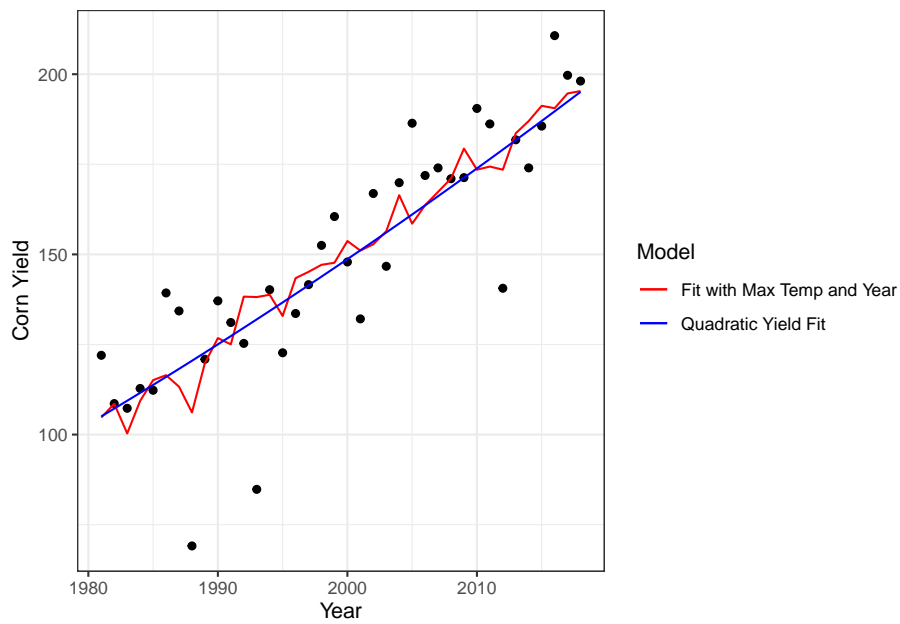


Figure 7.3: Comparative quadratic fit of corn yields over time (blue) and fitted line with maximum summer temperatures as well (red) in Winneshieck County, Iowa.

suggest an underlying relationship between maximum temperatures and yields. However, the relationship between squared maximum temperature and yield has a P-value of 0.14, compared with the year squared P-value of $9.22\text{e-}12$, so it is clearly not the important driver of trends. This model has an R-squared value of 0.749, around the same as (slightly lower than) the simple linear regression model with only yield vs. year. Thus, adding temperature doesn't significantly add to our understanding of yield trends in Winneshieck County.

7.5 Cross-Section: Analyze the relationship between temperature and yield across all counties in 2018. Is there a relationship? Interpret the results.

```
corn2018 <- cornyields %>%
  filter(year == "2018") %>%
  mutate_at(vars(county_ansi), funs(factor))

tmax2018 <- tmaxdf %>%
  filter(year == "2018") %>%
  filter(doy >= 152 & doy <= 243) %>%
  group_by(countyfp) %>%
  rename("county_ansi" = "countyfp") %>%
  summarize(meantmax = mean(tmax))

yieldtemp_2018 <- inner_join(corn2018, tmax2018, by="county_ansi") %>%
  mutate(tmaxsq = (meantmax^2))

yt_lm <- lm(yield ~ meantmax + tmaxsq, data = yieldtemp_2018)
summary(yt_lm)
```

```
##
## Call:
## lm(formula = yield ~ meantmax + tmaxsq, data = yieldtemp_2018)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
##	-44.221	-15.399	5.007	14.541	30.879

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-5501.602	1860.830	-2.957	0.00397 **

7.5. CROSS-SECTION: ANALYZE THE RELATIONSHIP BETWEEN TEMPERATURE AND YIELD ACROSS A

```
## meantmax      406.789    131.493    3.094  0.00263 **
## tmaxsq       -7.256      2.321   -3.126  0.00239 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.75 on 90 degrees of freedom
## Multiple R-squared:  0.1317, Adjusted R-squared:  0.1124
## F-statistic: 6.827 on 2 and 90 DF,  p-value: 0.001736
```

```
yieldtemp_2018$ytfit <- yt_lm$fitted.values

ggplot(yieldtemp_2018) +
  geom_point(mapping = aes(x = meantmax, y = yield)) +
  geom_line(mapping = aes(x = meantmax, y = ytfit, color="red")) +
  theme_bw() + theme(legend.position="none") +
  labs(x = "Mean Max Temperature (C)", y = "Corn Yield")
```

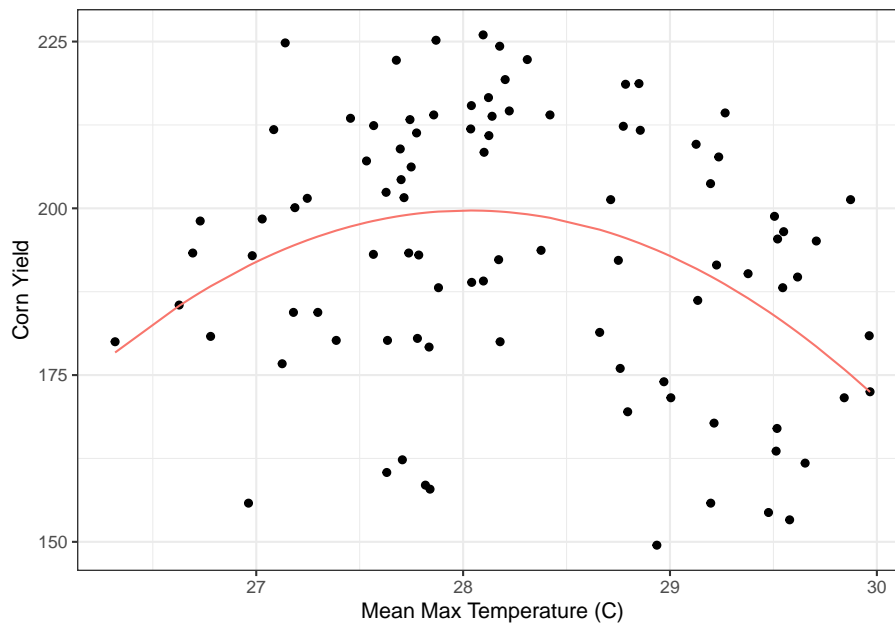


Figure 7.4: Quadratic fit of corn yields versus maximum summer temperatures (Degrees C) across Iowa.

There is a clear relationship with maximum temperatures and corn yields demonstrated in Figure 4. As we might expect, there appears to be a “sweet spot” in regard to temperature, with corn crops performing best at moderate temperatures and yields falling off at both low and high temperature years.

Lower mean maximum temperatures may indicate even lower temperatures that can shock crops, and high means are likely to cause high evaporation and withering. $P < 0.003$ for the relationship between temperature and corn yield across Iowa.

7.6 Panel: One way to leverage multiple time series is to group all data into what is called a “panel” regression. Convert the county ID code (“countyfp” or “county_ansi”) into factor using `as.factor`, then include this variable in a regression using all counties’ yield and summer temperature data. How does the significance of your temperature coefficients (T_{\max} , T_{\max}^2) change? Make a plot comparing actual and fitted yields and interpret the results of your model.

```
corn_all <- cornyields %>%
  mutate_at(vars(county_ansi), funs(factor))

tmax_all <- tmaxdf %>%
  filter(doy >= 152 & doy <= 243) %>%
  group_by(countyfp, year) %>%
  rename("county_ansi" = "countyfp") %>%
  summarize(meantmax = mean(tmax))
```

```
## `summarise()` has grouped output by 'county_ansi'. You can override using the
## `.groups` argument.
```

```
yieldtemp_all <- inner_join(corn_all, tmax_all) %>%
  mutate(tmaxsq = (meantmax^2)) %>%
  mutate(yearsq = year^2)
```

```
## Joining, by = c("county_ansi", "year")
```

7.6. PANEL: ONE WAY TO LEVERAGE MULTIPLE TIME SERIES IS TO GROUP ALL DATA INTO WHAT IS C

```

ytc_lm <- lm(yield ~ year + meantmax + tmaxsq + county_ansi, data = yieldtemp_all)
summary(ytc_lm)

```

```

##
## Call:
## lm(formula = yield ~ year + meantmax + tmaxsq + county_ansi,
##     data = yieldtemp_all)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -81.645  -9.720   1.924  13.232  40.409
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.826e+03  9.804e+01 -59.431  < 2e-16 ***
## year          2.203e+00  2.836e-02  77.664  < 2e-16 ***
## meantmax      1.182e+02  6.108e+00  19.352  < 2e-16 ***
## tmaxsq       -2.225e+00  1.085e-01 -20.503  < 2e-16 ***
## county_ansi3  -4.527e+00  4.321e+00  -1.048  0.294839
## county_ansi5   2.716e+00  4.343e+00   0.625  0.531743
## county_ansi7  -1.828e+01  4.350e+00  -4.203  2.70e-05 ***
## county_ansi9   5.068e+00  4.323e+00   1.172  0.241144
## county_ansi11  7.186e+00  4.325e+00   1.661  0.096732 .
## county_ansi13  7.289e+00  4.329e+00   1.684  0.092303 .
## county_ansi15  1.498e+01  4.323e+00   3.466  0.000534 ***
## county_ansi17  1.133e+01  4.332e+00   2.615  0.008966 **
## county_ansi19  7.651e+00  4.334e+00   1.765  0.077577 .
## county_ansi21  8.640e+00  4.328e+00   1.996  0.045974 *
## county_ansi23  9.089e+00  4.327e+00   2.100  0.035779 *
## county_ansi25  1.039e+01  4.326e+00   2.401  0.016400 *
## county_ansi27  9.666e+00  4.323e+00   2.236  0.025421 *
## county_ansi29  6.145e+00  4.321e+00   1.422  0.155092
## county_ansi31  1.579e+01  4.324e+00   3.651  0.000264 ***
## county_ansi33  4.582e+00  4.338e+00   1.056  0.290980
## county_ansi35  1.390e+01  4.325e+00   3.213  0.001325 **
## county_ansi37  2.169e+00  4.341e+00   0.500  0.617274
## county_ansi39 -2.404e+01  4.350e+00 -5.527  3.48e-08 ***
## county_ansi41  6.611e+00  4.329e+00   1.527  0.126809
## county_ansi43  8.864e+00  4.337e+00   2.044  0.041033 *
## county_ansi45  1.055e+01  4.325e+00   2.439  0.014756 *
## county_ansi47  6.528e+00  4.324e+00   1.510  0.131221
## county_ansi49  1.081e+01  4.321e+00   2.502  0.012386 *
## county_ansi51 -1.457e+01  4.352e+00 -3.349  0.000820 ***
## county_ansi53 -1.603e+01  4.350e+00 -3.686  0.000232 ***
## county_ansi55  9.423e+00  4.338e+00   2.172  0.029916 *

```

```

## county_ansi57 1.050e+01 4.321e+00 2.429 0.015186 *
## county_ansi59 2.906e+00 4.336e+00 0.670 0.502836
## county_ansi61 9.795e+00 4.340e+00 2.257 0.024059 *
## county_ansi63 7.232e+00 4.340e+00 1.666 0.095754 .
## county_ansi65 7.319e+00 4.341e+00 1.686 0.091905 .
## county_ansi67 4.791e+00 4.334e+00 1.106 0.269008
## county_ansi69 1.131e+01 4.330e+00 2.612 0.009035 **
## county_ansi71 1.358e+01 4.330e+00 3.136 0.001726 **
## county_ansi73 1.462e+01 4.321e+00 3.382 0.000727 ***
## county_ansi75 1.151e+01 4.328e+00 2.659 0.007863 **
## county_ansi77 3.379e+00 4.321e+00 0.782 0.434297
## county_ansi79 1.315e+01 4.324e+00 3.042 0.002370 **
## county_ansi81 8.706e+00 4.340e+00 2.006 0.044917 *
## county_ansi83 1.395e+01 4.326e+00 3.225 0.001271 **
## county_ansi85 6.891e+00 4.321e+00 1.595 0.110834
## county_ansi87 5.280e+00 4.321e+00 1.222 0.221864
## county_ansi89 9.433e-01 4.364e+00 0.216 0.828875
## county_ansi91 9.881e+00 4.334e+00 2.280 0.022661 *
## county_ansi93 1.186e+01 4.325e+00 2.743 0.006124 **
## county_ansi95 7.214e+00 4.322e+00 1.669 0.095161 .
## county_ansi97 -1.386e+00 4.330e+00 -0.320 0.748823
## county_ansi99 1.440e+01 4.322e+00 3.332 0.000871 ***
## county_ansi101 5.352e-01 4.325e+00 0.124 0.901510
## county_ansi103 4.380e+00 4.322e+00 1.013 0.310971
## county_ansi105 7.730e+00 4.328e+00 1.786 0.074158 .
## county_ansi107 2.203e+00 4.321e+00 0.510 0.610224
## county_ansi109 1.222e+01 4.335e+00 2.819 0.004839 **
## county_ansi111 1.779e+00 4.324e+00 0.411 0.680740
## county_ansi113 6.415e+00 4.326e+00 1.483 0.138218
## county_ansi115 7.330e+00 4.322e+00 1.696 0.089966 .
## county_ansi117 -2.168e+01 4.381e+00 -4.949 7.81e-07 ***
## county_ansi119 9.328e+00 4.325e+00 2.157 0.031063 *
## county_ansi121 -2.587e+00 4.321e+00 -0.599 0.549390
## county_ansi123 8.152e+00 4.321e+00 1.887 0.059302 .
## county_ansi125 1.919e+00 4.321e+00 0.444 0.656948
## county_ansi127 1.418e+01 4.326e+00 3.278 0.001055 **
## county_ansi129 1.023e+01 4.385e+00 2.332 0.019741 *
## county_ansi131 7.285e+00 4.352e+00 1.674 0.094242 .
## county_ansi133 7.987e-01 4.321e+00 0.185 0.853378
## county_ansi135 -1.585e+01 4.350e+00 -3.643 0.000273 ***
## county_ansi137 5.885e+00 4.322e+00 1.362 0.173381
## county_ansi139 8.283e+00 4.321e+00 1.917 0.055337 .
## county_ansi141 1.423e+01 4.328e+00 3.288 0.001018 **
## county_ansi143 8.743e+00 4.337e+00 2.016 0.043890 *
## county_ansi145 -3.674e-01 4.322e+00 -0.085 0.932261
## county_ansi147 7.261e+00 4.330e+00 1.677 0.093601 .

```

7.6. PANEL: ONE WAY TO LEVERAGE MULTIPLE TIME SERIES IS TO GROUP ALL DATA INTO WHAT IS C

```
## county_ansi149 7.352e+00 4.322e+00 1.701 0.089007 .
## county_ansi151 1.150e+01 4.326e+00 2.659 0.007880 **
## county_ansi153 1.403e+01 4.321e+00 3.247 0.001178 **
## county_ansi155 1.127e+01 4.350e+00 2.590 0.009627 **
## county_ansi157 1.055e+01 4.322e+00 2.441 0.014702 *
## county_ansi159 -2.070e+01 4.321e+00 -4.792 1.72e-06 ***
## county_ansi161 9.390e+00 4.326e+00 2.170 0.030050 *
## county_ansi163 1.628e+01 4.323e+00 3.765 0.000169 ***
## county_ansi165 7.673e+00 4.323e+00 1.775 0.075966 .
## county_ansi167 1.558e+01 4.323e+00 3.603 0.000318 ***
## county_ansi169 1.122e+01 4.325e+00 2.593 0.009543 **
## county_ansi171 9.740e+00 4.325e+00 2.252 0.024387 *
## county_ansi173 -1.404e+01 4.350e+00 -3.228 0.001256 **
## county_ansi175 -1.155e+01 4.350e+00 -2.655 0.007967 **
## county_ansi177 -5.278e+00 4.329e+00 -1.219 0.222881
## county_ansi179 -3.220e+00 4.351e+00 -0.740 0.459267
## county_ansi181 -2.159e+00 4.321e+00 -0.500 0.617309
## county_ansi183 1.042e+01 4.321e+00 2.410 0.015981 *
## county_ansi185 -2.189e+01 4.350e+00 -5.033 5.07e-07 ***
## county_ansi187 1.421e+01 4.326e+00 3.285 0.001029 **
## county_ansi189 8.236e+00 4.344e+00 1.896 0.058035 .
## county_ansi191 4.567e+00 4.350e+00 1.050 0.293826
## county_ansi193 2.799e+00 4.321e+00 0.648 0.517252
## county_ansi195 6.123e+00 4.356e+00 1.406 0.159892
## county_ansi197 1.156e+01 4.329e+00 2.669 0.007634 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.83 on 3646 degrees of freedom
## Multiple R-squared:  0.7207, Adjusted R-squared:  0.7129
## F-statistic: 93.13 on 101 and 3646 DF,  p-value: < 2.2e-16

yieldtemp_all$fittedyield <- ytc_lm$fitted.values

ggplot(yieldtemp_all, mapping = aes(x = fittedyield, y = yield)) +
  geom_point() +
  geom_smooth(method="lm") +
  theme_bw() + theme(legend.position="none") +
  labs(x = "Fitted Yield", y = "Actual Yield")

## `geom_smooth()` using formula 'y ~ x'

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

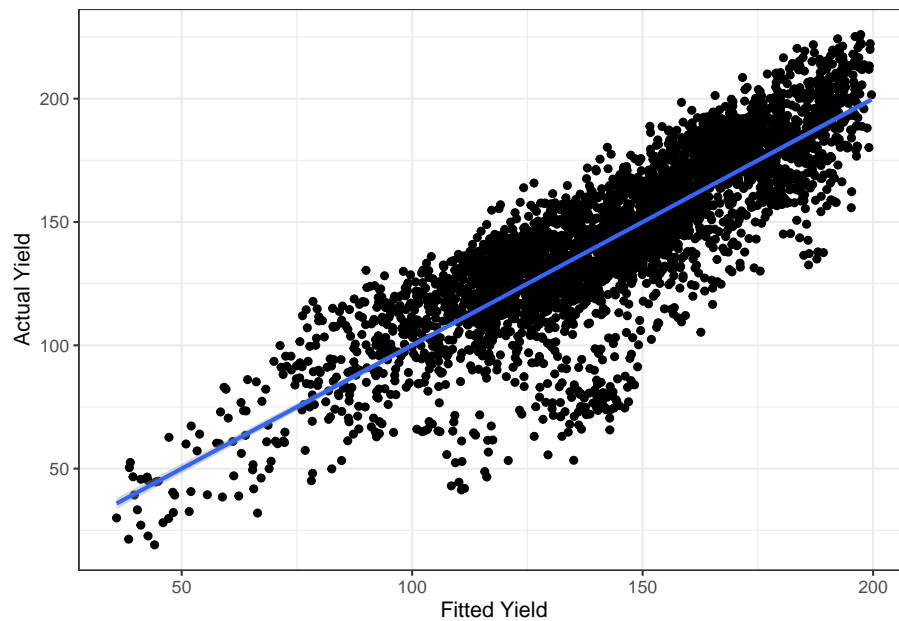


Figure 7.5: Fitted model yield values versus actual yield values for all counties of Iowa over all available years, from 1981 to 2018.

As a panel regression of all counties over all years, the statistical significance of year, mean maximum temperature, and squared maximum temperature as predictors of yield becomes stronger ($P < 2e-16$ for each). The R squared value for the model is 0.721, indicating a pretty good fit, as is evident in Figure 5. However, the residuals for the model are pretty wide (Figures 5, 6) and the data may not be very normally distributed (Figure 6).

7.7 Soybeans: Download NASS data on soybean yields and explore either a time series relationship for a given county, the cross-sectional relationship for a given year, or a panel across all counties and years.

```
# parameters to query on
params2 <- list(commodity_desc = "SOYBEANS", prodn_practice_desc = "ALL PRODUCTION PRA")
```


7.7. SOYBEANS: DOWNLOAD NASS DATA ON SOYBEAN YIELDS AND EXPLORE EITHER A TIME SERIES

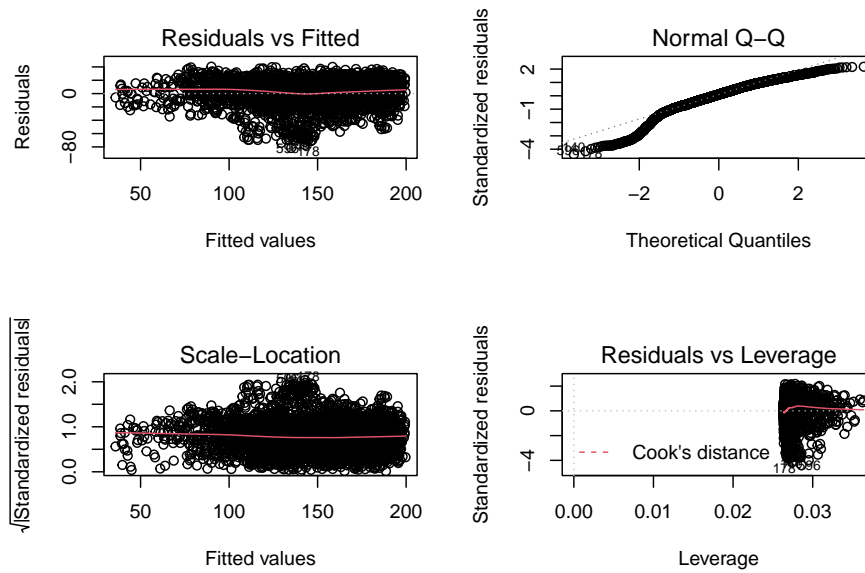


Figure 7.6: Residuals (top left), Normal Q-Q (top right), Scale-Location (bottom left), and Cook's Distance (bottom right) plots for the panel regression model.

```

# download
soyyieldsall <- nassqs_yields(params)

## |

soyyieldsall$county_ansi <- as.numeric(soyyieldsall$county_ansi)
soyyieldsall$yield <- as.numeric(soyyieldsall$Value)

# clean and filter this dataset
soy <- select(soyyieldsall, county_ansi, county_name, yield, year) %>%
  filter(!is.na(county_ansi) & !is.na(yield))
soy <- tibble(soy)

soy_panel <- soy %>%
  mutate_at(vars(county_ansi), funs(factor)) %>%
  mutate(yearsq = year^2)

soypanel_lm <- lm(yield ~ year + yearsq + county_ansi, data = soy_panel)
summary(soypanel_lm)

##
## Call:
## lm(formula = yield ~ year + yearsq + county_ansi, data = soy_panel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.865   -9.428    3.328   14.357   54.326
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.646e+04  1.090e+04   4.263 2.06e-05 ***
## year        -4.859e+01  1.089e+01  -4.461 8.38e-06 ***
## yearsq       1.272e-02  2.722e-03   4.672 3.09e-06 ***
## county_ansi3 -3.974e+00  4.809e+00  -0.826 0.408591
## county_ansi5  1.128e+01  4.749e+00   2.376 0.017550 *
## county_ansi7 -1.878e+01  4.778e+00  -3.931 8.61e-05 ***
## county_ansi9  9.877e+00  4.749e+00   2.080 0.037621 *
## county_ansi11 1.206e+01  4.749e+00   2.539 0.011162 *
## county_ansi13 1.391e+01  4.749e+00   2.930 0.003413 **
## county_ansi15 1.785e+01  4.749e+00   3.759 0.000173 ***
## county_ansi17 1.899e+01  4.749e+00   3.998 6.50e-05 ***
## county_ansi19 1.609e+01  4.749e+00   3.389 0.000709 ***
## county_ansi21 1.519e+01  4.749e+00   3.198 0.001396 **
## county_ansi23 1.612e+01  4.749e+00   3.394 0.000696 ***

```

7.7. SOYBEANS: DOWNLOAD NASS DATA ON SOYBEAN YIELDS AND EXPLORE EITHER A TIME SERIES

```
## county_ansi25 1.603e+01 4.749e+00 3.375 0.000744 ***
## county_ansi27 1.435e+01 4.749e+00 3.023 0.002523 **
## county_ansi29 7.386e+00 4.749e+00 1.555 0.119953
## county_ansi31 1.942e+01 4.749e+00 4.088 4.43e-05 ***
## county_ansi33 1.250e+01 4.749e+00 2.632 0.008515 **
## county_ansi35 1.961e+01 4.778e+00 4.105 4.13e-05 ***
## county_ansi37 1.127e+01 4.749e+00 2.373 0.017672 *
## county_ansi39 -2.380e+01 4.841e+00 -4.916 9.19e-07 ***
## county_ansi41 1.302e+01 4.749e+00 2.742 0.006141 **
## county_ansi43 1.659e+01 4.749e+00 3.493 0.000483 ***
## county_ansi45 1.556e+01 4.749e+00 3.276 0.001061 **
## county_ansi47 1.300e+01 4.749e+00 2.737 0.006228 **
## county_ansi49 1.131e+01 4.749e+00 2.382 0.017259 *
## county_ansi51 -1.838e+01 4.778e+00 -3.847 0.000121 ***
## county_ansi53 -1.681e+01 4.841e+00 -3.472 0.000521 ***
## county_ansi55 1.775e+01 4.749e+00 3.738 0.000188 ***
## county_ansi57 9.205e+00 4.778e+00 1.926 0.054127 .
## county_ansi59 9.501e+00 4.749e+00 2.001 0.045507 *
## county_ansi61 1.838e+01 4.749e+00 3.871 0.000110 ***
## county_ansi63 1.503e+01 4.778e+00 3.145 0.001676 **
## county_ansi65 1.555e+01 4.749e+00 3.275 0.001065 **
## county_ansi67 1.283e+01 4.749e+00 2.702 0.006920 **
## county_ansi69 1.810e+01 4.749e+00 3.810 0.000141 ***
## county_ansi71 5.259e+00 4.778e+00 1.100 0.271183
## county_ansi73 1.644e+01 4.749e+00 3.461 0.000545 ***
## county_ansi75 1.853e+01 4.749e+00 3.901 9.75e-05 ***
## county_ansi77 5.347e+00 4.749e+00 1.126 0.260249
## county_ansi79 1.771e+01 4.778e+00 3.706 0.000214 ***
## county_ansi81 1.716e+01 4.749e+00 3.614 0.000306 ***
## county_ansi83 1.836e+01 4.749e+00 3.866 0.000112 ***
## county_ansi85 6.372e+00 4.809e+00 1.325 0.185252
## county_ansi87 2.204e+00 4.749e+00 0.464 0.642690
## county_ansi89 9.845e+00 4.749e+00 2.073 0.038239 *
## county_ansi91 1.758e+01 4.749e+00 3.703 0.000216 ***
## county_ansi93 1.808e+01 4.778e+00 3.784 0.000156 ***
## county_ansi95 9.245e+00 4.778e+00 1.935 0.053090 .
## county_ansi97 5.382e+00 4.749e+00 1.133 0.257218
## county_ansi99 1.823e+01 4.778e+00 3.815 0.000138 ***
## county_ansi101 -6.065e+00 4.749e+00 -1.277 0.201671
## county_ansi103 6.598e+00 4.778e+00 1.381 0.167447
## county_ansi105 1.373e+01 4.778e+00 2.874 0.004069 **
## county_ansi107 1.030e+00 4.749e+00 0.217 0.828258
## county_ansi109 1.968e+01 4.749e+00 4.145 3.47e-05 ***
## county_ansi111 -4.248e+00 4.749e+00 -0.894 0.371157
## county_ansi113 1.244e+01 4.778e+00 2.604 0.009251 **
## county_ansi115 4.608e+00 4.749e+00 0.970 0.331929
```

```

## county_ansi117 -2.150e+01 4.841e+00 -4.441 9.20e-06 ***
## county_ansi119 1.472e+01 4.749e+00 3.100 0.001952 **
## county_ansi121 -1.277e+00 4.749e+00 -0.269 0.788032
## county_ansi123 8.308e+00 4.749e+00 1.749 0.080294 .
## county_ansi125 1.995e+00 4.778e+00 0.418 0.676297
## county_ansi127 2.112e+01 4.778e+00 4.419 1.02e-05 ***
## county_ansi129 5.192e+00 4.809e+00 1.080 0.280353
## county_ansi131 1.591e+01 4.749e+00 3.349 0.000818 ***
## county_ansi133 1.767e-01 4.749e+00 0.037 0.970326
## county_ansi135 -1.612e+01 4.778e+00 -3.373 0.000751 ***
## county_ansi137 4.108e+00 4.749e+00 0.865 0.387051
## county_ansi139 8.369e+00 4.749e+00 1.762 0.078100 .
## county_ansi141 2.129e+01 4.749e+00 4.483 7.57e-06 ***
## county_ansi143 1.615e+01 4.749e+00 3.400 0.000680 ***
## county_ansi145 -2.245e+00 4.749e+00 -0.473 0.636402
## county_ansi147 1.413e+01 4.749e+00 2.975 0.002950 **
## county_ansi149 1.129e+01 4.778e+00 2.362 0.018212 *
## county_ansi151 1.801e+01 4.778e+00 3.770 0.000166 ***
## county_ansi153 1.401e+01 4.749e+00 2.949 0.003206 **
## county_ansi155 1.124e+01 4.778e+00 2.352 0.018734 *
## county_ansi157 1.181e+01 4.749e+00 2.487 0.012908 *
## county_ansi159 -2.035e+01 4.778e+00 -4.258 2.11e-05 ***
## county_ansi161 1.602e+01 4.749e+00 3.373 0.000751 ***
## county_ansi163 1.991e+01 4.749e+00 4.192 2.83e-05 ***
## county_ansi165 1.242e+01 4.778e+00 2.599 0.009393 **
## county_ansi167 2.019e+01 4.749e+00 4.252 2.17e-05 ***
## county_ansi169 1.555e+01 4.749e+00 3.275 0.001067 **
## county_ansi171 1.451e+01 4.749e+00 3.055 0.002266 **
## county_ansi173 -1.499e+01 4.778e+00 -3.137 0.001720 **
## county_ansi175 -1.023e+01 4.778e+00 -2.142 0.032273 *
## county_ansi177 -1.466e+01 4.749e+00 -3.086 0.002041 **
## county_ansi179 -5.652e+00 4.778e+00 -1.183 0.236907
## county_ansi181 -2.277e+00 4.809e+00 -0.473 0.635891
## county_ansi183 8.595e+00 4.778e+00 1.799 0.072137 .
## county_ansi185 -2.086e+01 4.778e+00 -4.365 1.30e-05 ***
## county_ansi187 2.017e+01 4.749e+00 4.247 2.21e-05 ***
## county_ansi189 1.687e+01 4.749e+00 3.552 0.000387 ***
## county_ansi191 1.306e+01 4.749e+00 2.750 0.005980 **
## county_ansi193 6.472e+00 4.749e+00 1.363 0.173049
## county_ansi195 1.503e+01 4.749e+00 3.164 0.001566 **
## county_ansi197 1.807e+01 4.749e+00 3.805 0.000144 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.37 on 3914 degrees of freedom
## Multiple R-squared:  0.6598, Adjusted R-squared:  0.6511

```

7.7. SOYBEANS: DOWNLOAD NASS DATA ON SOYBEAN YIELDS AND EXPLORE EITHER A TIME SERIES

```
## F-statistic: 75.91 on 100 and 3914 DF,  p-value: < 2.2e-16
```

```
soy_panel$fittedyield <- soypanel_lm$fitted.values

ggplot(soy_panel, mapping = aes(x = year, y = yield)) +
  geom_point() +
  geom_line(mapping = aes(x = year, y = fittedyield, color="red")) +
  geom_smooth(method="lm") +
  theme_bw() + theme(legend.position="none") +
  labs(x = "Year", y = "Soy Yield")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

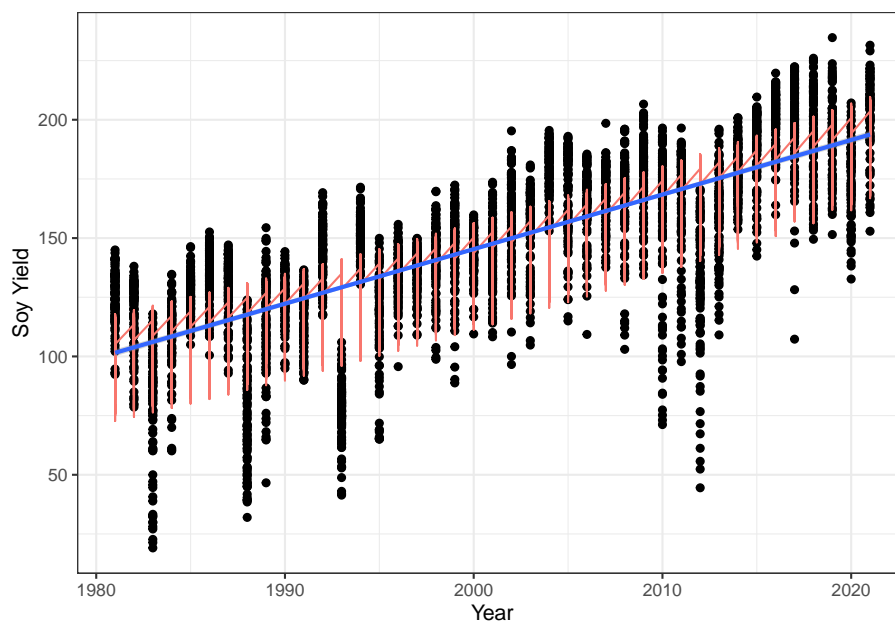


Figure 7.7: Soy yields over time (years) across all counties of Iowa, with a panel fit (orange), and linear fit (blue). Panel regression R squared is 0.660, with $p < 8.5e-06$ for year and years squared versus yield.

Like with corn yields, soy yields in Iowa follow an upward trend over time, though with wide residuals (Figure 7).

7.8 Bonus: Find a package to make a county map of Iowa displaying some sort of information about yields or weather. Interpret your map.

```
library(sf) #Spatial package that can read and create shapefiles
library(mapview) #Interactive maps
library(USAboundaries) #USA states and counties

counties <- us_counties

Iowa_ct <- counties(states = 'iowa')
#str(Iowa_ct)
#mapview(Iowa_ct)

summer2018 <- tmaxdf %>%
  filter(doy >= 152 & doy <= 243, year==2018) %>% #day 152= June 1, 243= Aug 31
  group_by(countyfp, year) %>%
  summarize(meantmax = mean(tmax))
```

`summarise()` has grouped output by 'countyfp'. You can override using the
`.groups` argument.

```
Itemp_2018 <- merge(Iowa_ct, summer2018)

mapview(Itemp_2018, zcol = 'meantmax', col.regions=brewer.pal(9, "OrRd"), layer.name = 'meantmax')
```