IS602 Project - Data Cleaning and Pre-Analysis

Max Wagner

May 3, 2016

Introduction

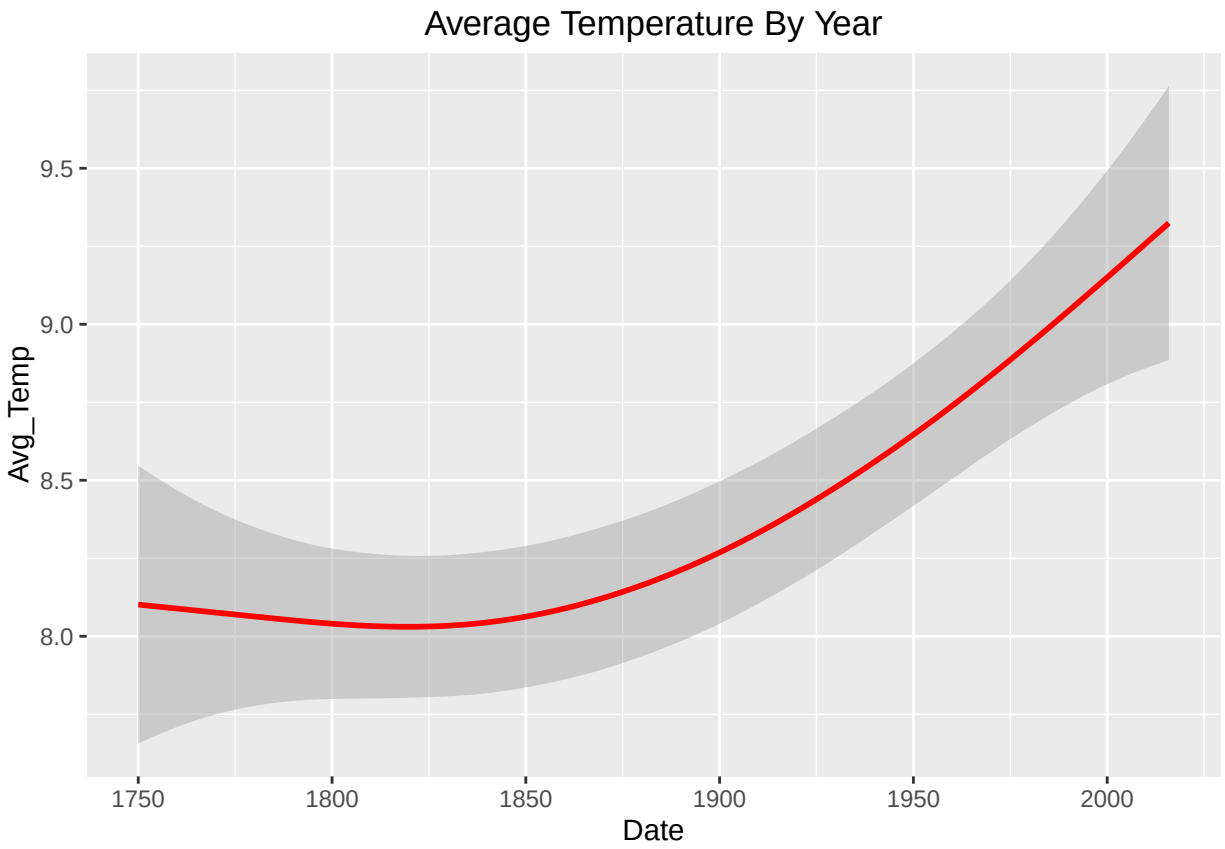
Berkley Earth has gathered climate and weather data from 1750 until present day and has compiled it into multiple csv files. I would like to look at temperature over time, which regions have changed the most, and if specific types of regions are more prone to temperature change.

I chose R Markdown as my main resource as it is significantly more presentable and easy to put into a report form than my other options.

Analysis

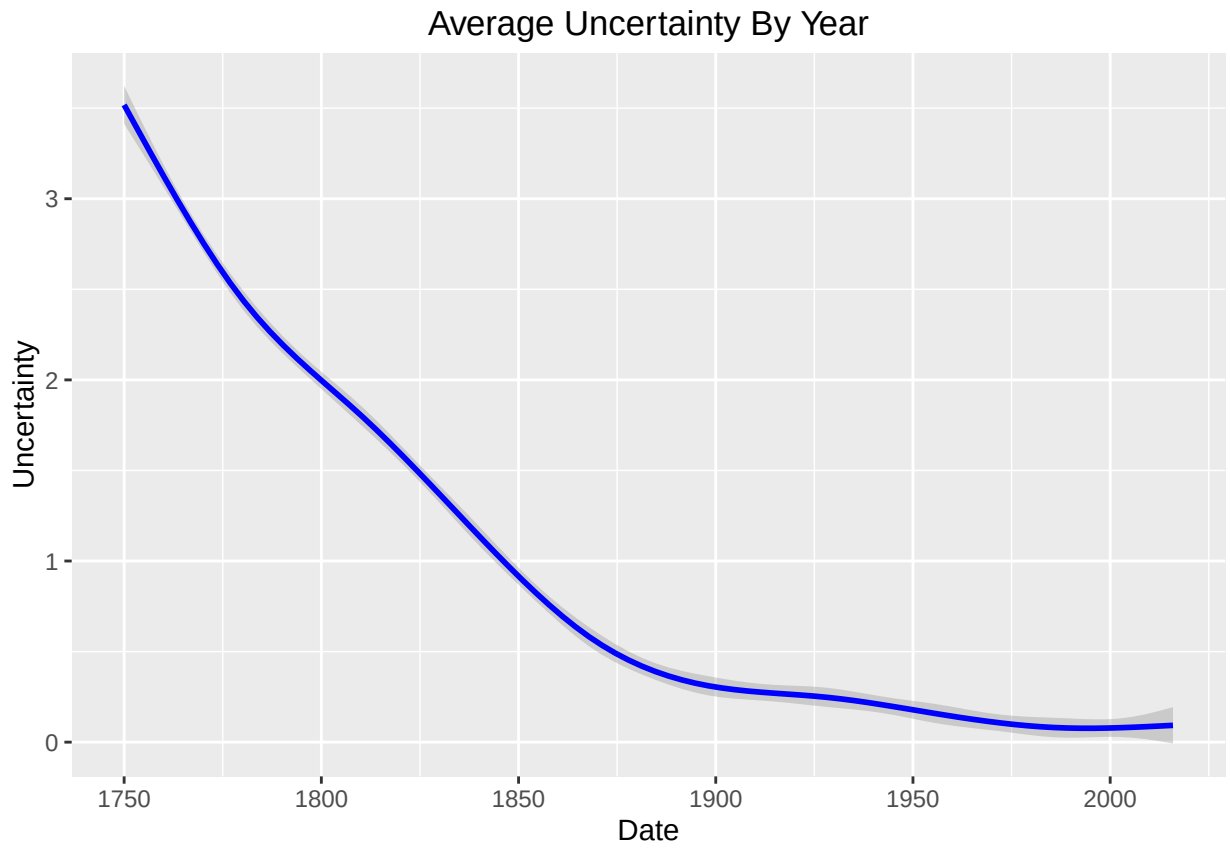
This small first section cleans up some of the global temperature values and then plots the average temperature by year. 'ggplot2' will automatically remove NA values in this case. Due to the amount of points, I removed the points from the graph and left only the smoothed fit line. From the graph, it is apparent that global temperature has indeed risen by roughly 1.5 degree from 1750 until present day. This means that the project is not pointless and that there should be an end result of some kind.

```
library(ggplot2)
library(plyr)
library(dplyr)

global <- read.csv("GlobalTemperatures.csv")
global$dt <- as.Date(global$dt, "%Y-%m-%d") #convert dates
globaltemps <- global[1:3] #drop other columns for purpose of initial graphs
colnames(globaltemps) <- c("Date", "Avg_Temp", "Uncertainty")

ggplot(data = globaltemps, aes(x = Date, y = Avg_Temp)) +
  geom_smooth(color = "red") +
  labs(title = "Average Temperature By Year")
```



Another interesting piece to look at before moving on to more individualized analysis is the change in uncertainty over the recording years. As expected the uncertainty is significantly higher in the 1700s than in present day. The change is from roughly 3.5% to 0.1%.

```
ggplot(data = globaltemps, aes(x = Date, y = Uncertainty)) +
  geom_smooth(color = "blue") +
  labs(title = "Average Uncertainty By Year")
```

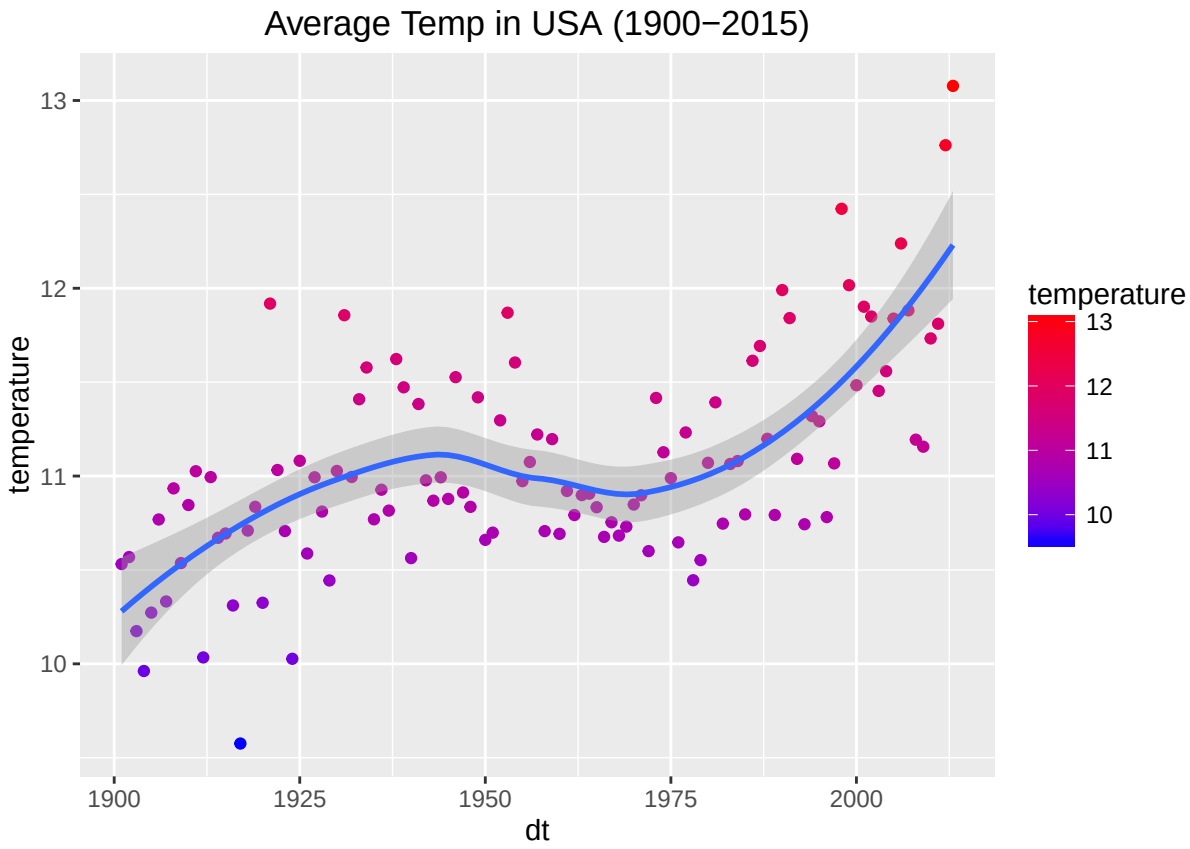


Going on from here, It may be best to exclude the earliest years of the data, as there are numerous blank values, and the uncertainty is high in comparison with the later records. 1900 marks the change over in the data set to a more modern method of collection and record keeping. For the sake of accuracy, I will only include samples from 1900 onward for the remainder of the project. I can clean up one of the csv's to fit more of what I need.

```
majors <- read.csv("GlobalLandTemperaturesByState.csv")
majors <- na.omit(majors) #omit na's
majors$dt <- as.Date(majors$dt, "%Y-%m-%d") #convert dates
majors$dt <- as.numeric(format(majors$dt, "%Y")) #only keep the year, the rest isn't important
majors_1900 <- filter(majors, dt > 1900) #1900+ only
majors_usa <- filter(majors, Country == "United States" & dt > 1900) #usa only and 1900+
majors_usa$State <- revalue(majors_usa$State, c("Georgia (State)" = "Georgia")) # fix the bad georgia n
majors_nyc <- filter(majors, State == "New York" & dt > 1900) #nyc only and 1900+
```

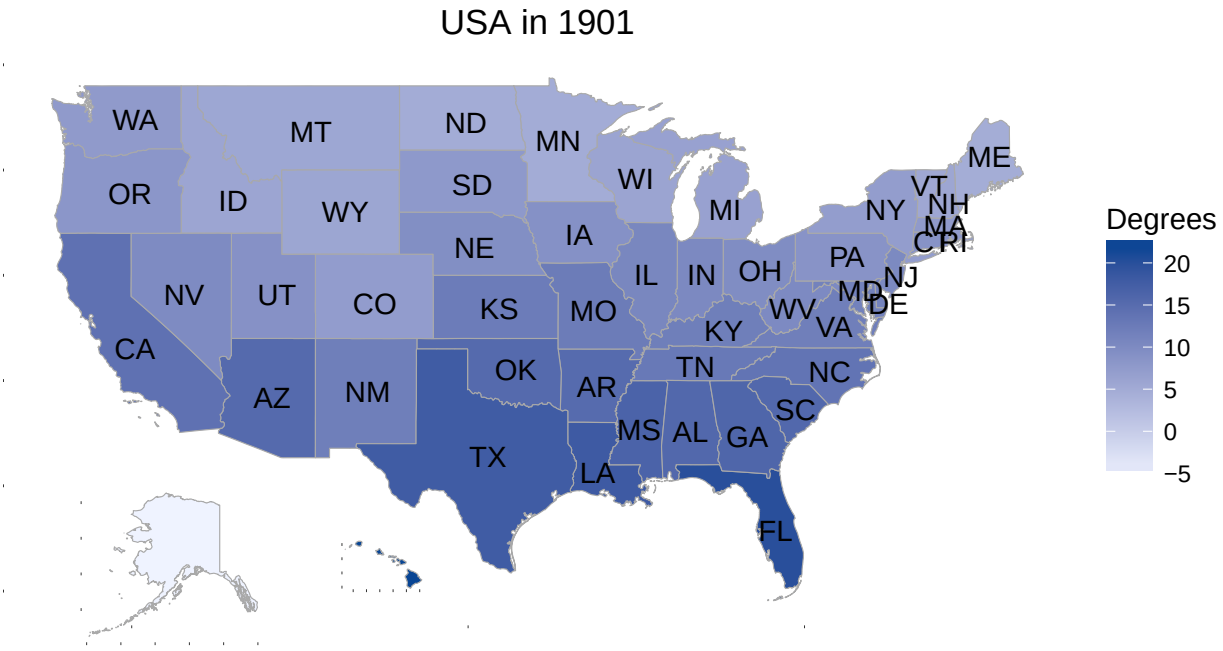
The next section deals with purely cities in the United States. To begin it will just be a simple average of temperatures over time. Again, we can see a definite increase in average temperatures, however there is a dip in the middle that was not present in the overall data from the global readings.

```
bystateavg <- ddply(majors_usa, ~dt, summarise, temperature = mean(AverageTemperature))
ggplot(data = bystateavg, aes(x = dt, y = temperature, color = temperature)) +
  geom_point() + geom_smooth() + labs(title = "Average Temp in USA (1900-2015)", xlab = "Year", ylab =
  scale_color_gradient(low = "blue", high = "red")
```

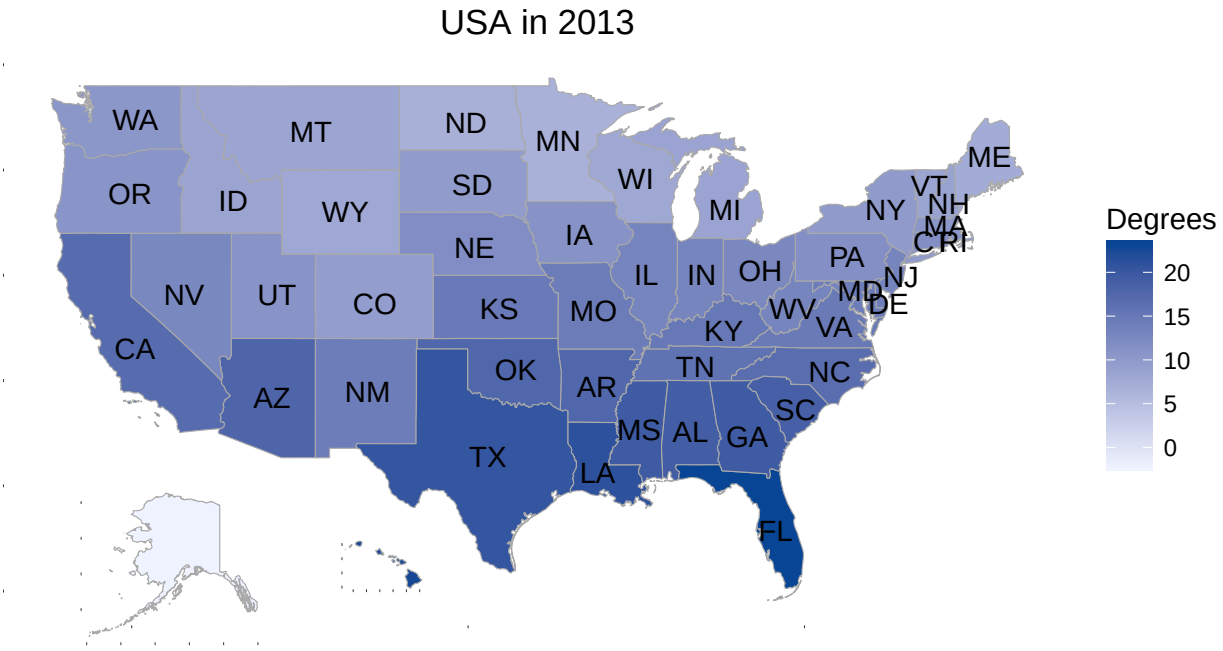


Now we can plot the temperatures onto a map.

```
library(choroplethr)
library(choroplethrMaps)

usa_map <- majors_usa[c(1:2, 4)] # select some columns
colnames(usa_map) <- c("year", "value", "region") # rename to fit the map package guidelines
usa_map$region<-tolower(usa_map$region) #region has to be lowercase

usa_map1901 <- filter(usa_map, year==1901)
usa_map1901 <- ddply(usa_map1901, ~region, summarise, value = mean(value))

usa_map2013 <- filter(usa_map, year==2013)
usa_map2013 <- ddply(usa_map2013, ~region, summarise, value = mean(value))

state_choropleth(usa_map1901, title = "USA in 1901", legend = "Degrees", num_colors = 1)
```

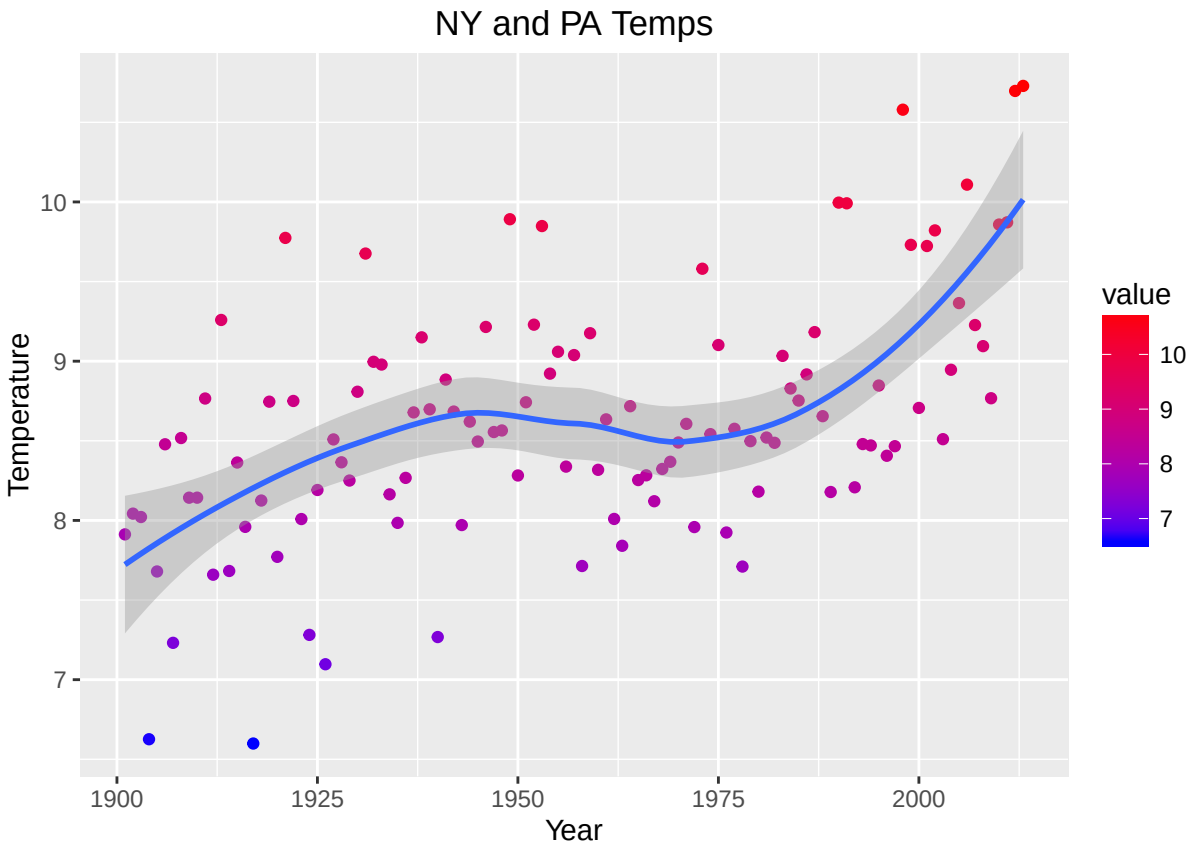



```
state_choropleth(usa_map2013, title = "USA in 2013", legend = "Degrees", num_colors = 1)
```



The difference when looking at the entire country seems noticable, but barely, given that the scale is quite large. Let's try just looking at a few states in the Northeast instead, specifically NY and PA. This shows a similar trend as the overall country, but with a smaller scale.

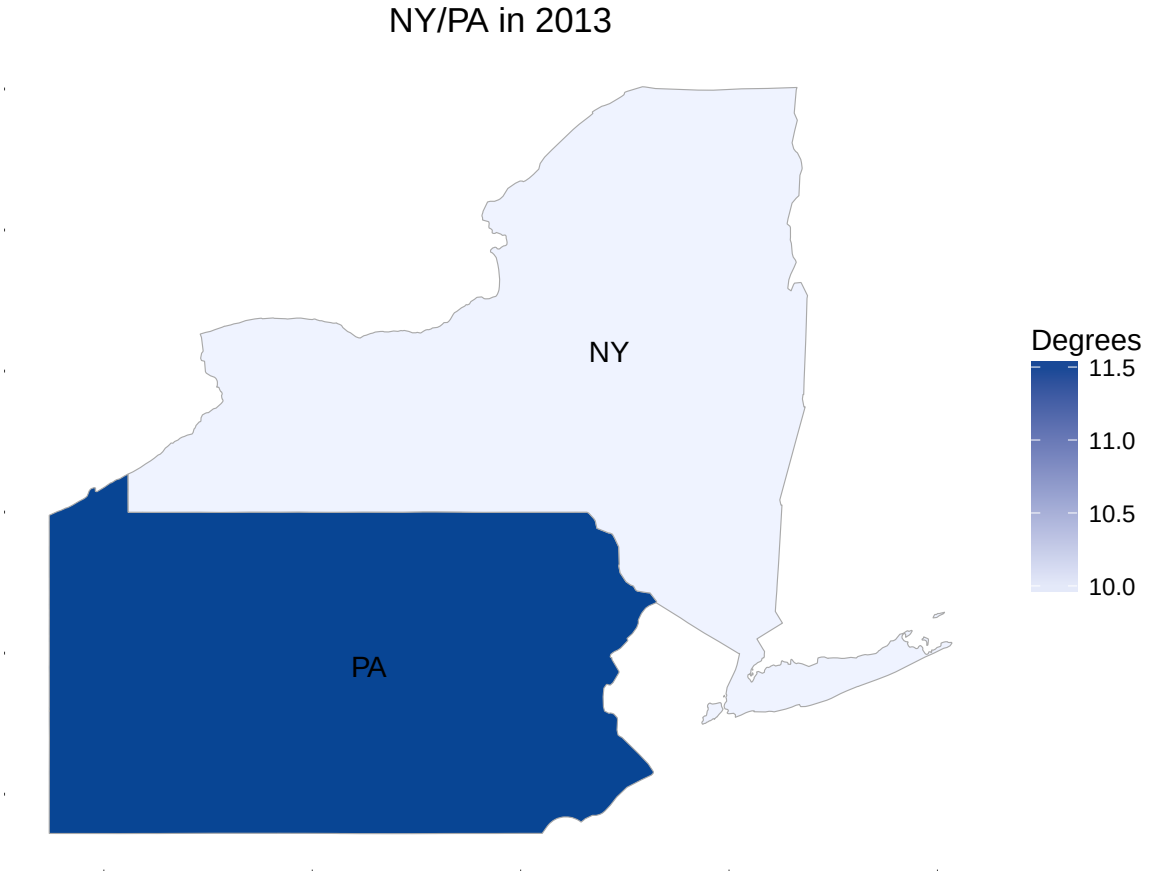
```
nypamap <- filter(usa_map, region == "new york" | region == "pennsylvania")
nypamap <- ddply(nypamap, ~year, summarise, value = mean(value))

qplot(data = nypamap, x = year, y = value, geom = c("point", "smooth"), main = "NY and PA Temps", xlab =
  aes(color = value) +
  scale_color_gradient(low = "blue", high = "red")
```



Now we can look at a map of just these two states. In this case the colors remain the same, however the scale changes enormously from (7.2:8.4) to (10.0:11.5).

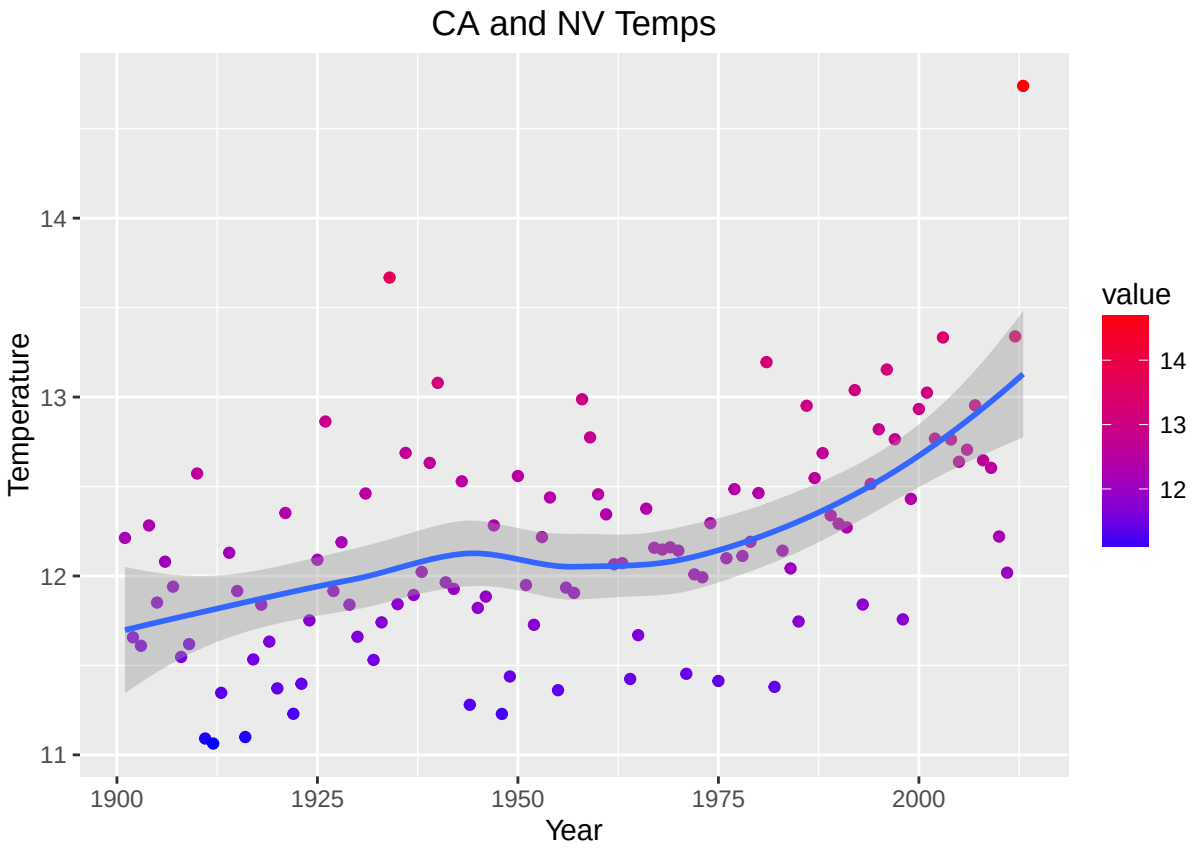
```
state_choropleth(usa_map1901, title = "NY/PA in 1901", legend = "Degrees", num_colors = 1, zoom = c("ne
```


```
state_choropleth(usa_map2013, title = "NY/PA in 2013", legend = "Degrees", num_colors = 1, zoom = c("ne
```



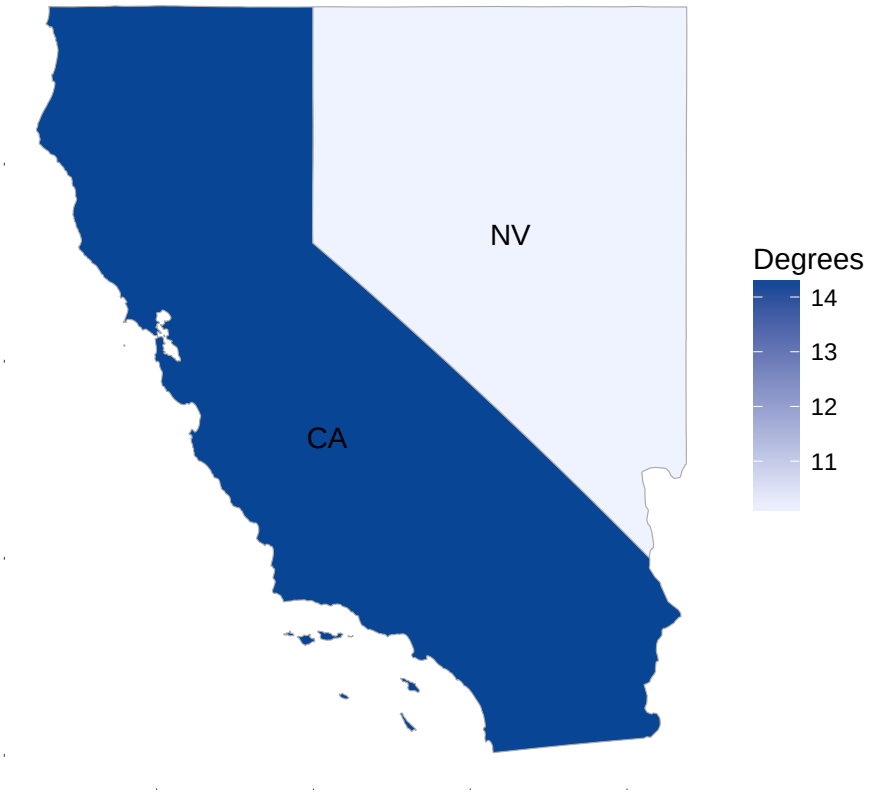
Let's add a warmer region to compare the changes. California and Nevada will get the same treatment as New York and Pennsylvania. The first gradiant map shows a much less steep curve in comparison to the NY/PA map, and a much smaller temperature range.

```
camap <- filter(usa_map, region == "california" | region == "nevada")
camap <- ddply(camap, ~year, summarise, value = mean(value))

qplot(data = camap, x = year, y = value, geom = c("point", "smooth"), main = "CA and NV Temps", xlab =
  aes(color = value) +
  scale_color_gradient(low = "blue", high = "red")
```

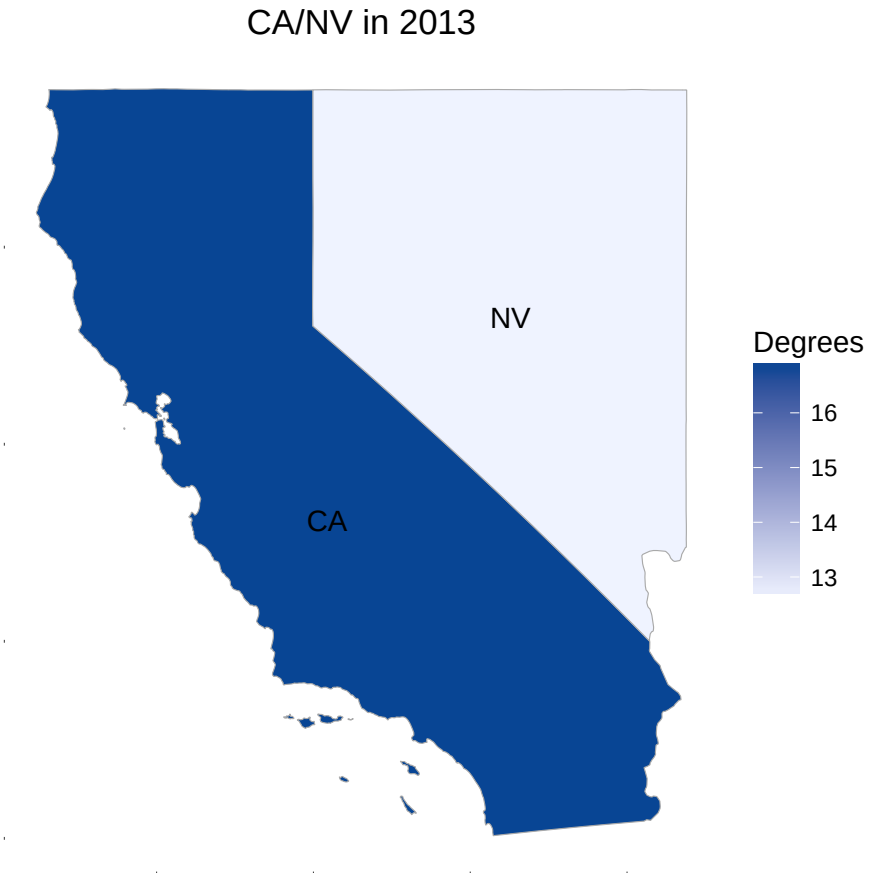


And now to zoom in on them in the map. Interestingly enough we still see a fairly notable change even for a warmer region of the country with the range changing from 11:14 to 13:16. A difference here is that the range contains itself, meaning that the change was not as large for the warmer region of CA/NV when compared to the colder region of NY/PA.

```
state_choropleth(usa_map1901, title = "CA/NV in 1901", legend = "Degrees", num_colors = 1, zoom = c("ca
```




```
state_choropleth(usa_map2013, title = "CA/NV in 2013", legend = "Degrees", num_colors = 1, zoom = c("ca
```



Conclusions

While I strayed away from looking at particular seasons in lieu of looking at specific states and regions, I believe that the results were more reliable this way as it averages an entire year instead of only a few months. This negates having a particularly cold winter or hot summer. The use of the `choropleth` package was incredibly useful to visualize the changes instead of looking purely at numbers. I had planned to focus on NYC near the end of the project, but realized that the only data from the New York State was from NYC, so it made sense to compare to other states instead of just a particular city.

From the results of the brief study, it seems beyond reasonable doubt that the average temperature is rising. How much it has risen depends on the region, but globally it has risen by a measurable amount.

Sources