Group C Project 1 JSON vignette: *The one with Jason and temperatures*

Jesse DeLaRosa and Sophia Melenikiotis

6/12/2019

Contents

JSON Files: What they are, what they do, and what we did with one

Introduction to JSONs

JSON files, as described by fileinfo are files "that stores simple data structures and objects in JavaScript Object Notation (JSON) format." They are used often in storing data from web applications and servers, sometimes moving data between the two. Naturally, we found our data set from such a source, though we'll go further into detail on that matter in a later section.

JSON and R

R has a few great packages for reading JSON data sets into R. We started by seeing if `tidyverse` had options for this, and while they did not, they did recommend using `jsonlite` as an easy package to read in JSON files. Another option is `rjson`, and both `rjson` and `jsonlite` provide access to a great function for reading in JSON datasets into R: `fromJSON()`, which lets you read in a JSON and convert it into a data frame. We chose to use `fromJSON()` from `jsonlite` since that package got `tidyverse`'s recommendation.

An example of JSON data: Raleigh's surface temperature

Ever increasing concerns of the state of the Earth's climate has led many a meterologist to collect data on temperatures of as many regions of the world as possible. Berkeley Earth is an organization of researchers who have collected data pertaining to the Earth's temperature across over a billion data points dating back to the 18th century, all freely available to the public. The Chapel Hill Open Data website pulled a piece of this massive dataset and created a JSON file with information pertaining to Raleigh's average surface temperature from over 3000 days between 1743 and 2013.

Reading the data

First, we'll demonstrate how easy it is to pull data from a JSON file into R using the `fromJSON` function and then show what the data frame looks like before describing what the variables represent.

```
library(tidyverse)
library(jsonlite)

RaleighJson <- fromJSON("C:/Users/Jesse DeLaRosa/Desktop/Project/Grad School/ST 558 Data Science for Sta
head(RaleighJson)
```

```
##      city       country geopoint.lat geopoint.lon averagetemperature
## 1 Raleigh United States     36.28333    -79.93333              7.157
## 2 Raleigh United States     36.28333    -79.93333              2.801
## 3 Raleigh United States     36.28333    -79.93333              9.227
## 4 Raleigh United States     36.28333    -79.93333              9.835
## 5 Raleigh United States     36.28333    -79.93333             19.754
## 6 Raleigh United States     36.28333    -79.93333             25.242
##           dt averagetemperatureuncertainty
## 1 1829-12-01                         5.711
## 2 1971-01-01                         0.162
## 3 1789-11-01                         1.845
## 4 1923-03-01                         0.319
## 5 1762-09-01                         4.989
## 6 1821-08-01                         1.289
```

The `city`, `country`, `latitude`, and `longitude` values correspond to where the data is coming from. Since the good people of Chapel Hill only pulled the Raleigh data, however, all of these values are the same across the 3,239 records we're working with. `Average temperature` is the average temperature of the day, `dt` is the date the data was recorded, and `Average temperature uncertainty` corresponds to the measured accuracy of the measurement.

We'll convert this data frame to a tibble before we continue, to make the data easier to work with `tidyverse`.

```
## # A tibble: 3,239 x 6
##    city  country geopoint$lat  $lon averagetemperat~ dt    averagetemperat~
##    <chr> <chr>          <dbl> <dbl>            <dbl> <chr>            <dbl>
##  1 Rale~ United~         36.3 -79.9             7.16 1829~            5.71
##  2 Rale~ United~         36.3 -79.9             2.80 1971~            0.162
##  3 Rale~ United~         36.3 -79.9             9.23 1789~            1.84
##  4 Rale~ United~         36.3 -79.9             9.84 1923~            0.319
##  5 Rale~ United~         36.3 -79.9            19.8  1762~            4.99
##  6 Rale~ United~         36.3 -79.9            25.2  1821~            1.29
##  7 Rale~ United~         36.3 -79.9             6.79 1953~            0.219
##  8 Rale~ United~         36.3 -79.9            26.2  1773~            1.78
##  9 Rale~ United~         36.3 -79.9            23.5  1877~            1.51
## 10 Rale~ United~         36.3 -79.9            22.1  1934~            0.256
## # ... with 3,229 more rows
```

```
## [1] "Raleigh"
```

While the dataset that Chapel Hill compiled this data from is much, much larger than the 3000 observations we'll work with here, the JSON file they created from Berkeley Earth was specifically for Raleigh. So we'll focus on variations in temperature across different points in time and seasons. First, howevever, several new variables will need to be created in order to execute such analyses.

Creating new variables

With growing concerns about temperature changes over time, having a data set that spans Raleigh's history back to when it was under British rule should give us a good idea of how such changes look. We'll examine these in terms of season and across years. First we'll create a variable specifically for each observation's year and month, then use those for each observation's season and year range.

```
## # A tibble: 3,239 x 4
##     Year Month Season YearRange
##    <int> <int> <chr>  <chr>
##  1  1829    12 Winter 1793-1842
##  2  1971     1 Winter 1943-1992
##  3  1789    11 Fall   1743-1792
##  4  1923     3 Spring 1893-1942
##  5  1762     9 Fall   1743-1792
##  6  1821     8 Summer 1793-1842
##  7  1953     1 Winter 1943-1992
##  8  1773     7 Summer 1743-1792
##  9  1877     6 Summer 1843-1892
## 10  1934     9 Fall   1893-1942
## # ... with 3,229 more rows
```

Numeric summaries

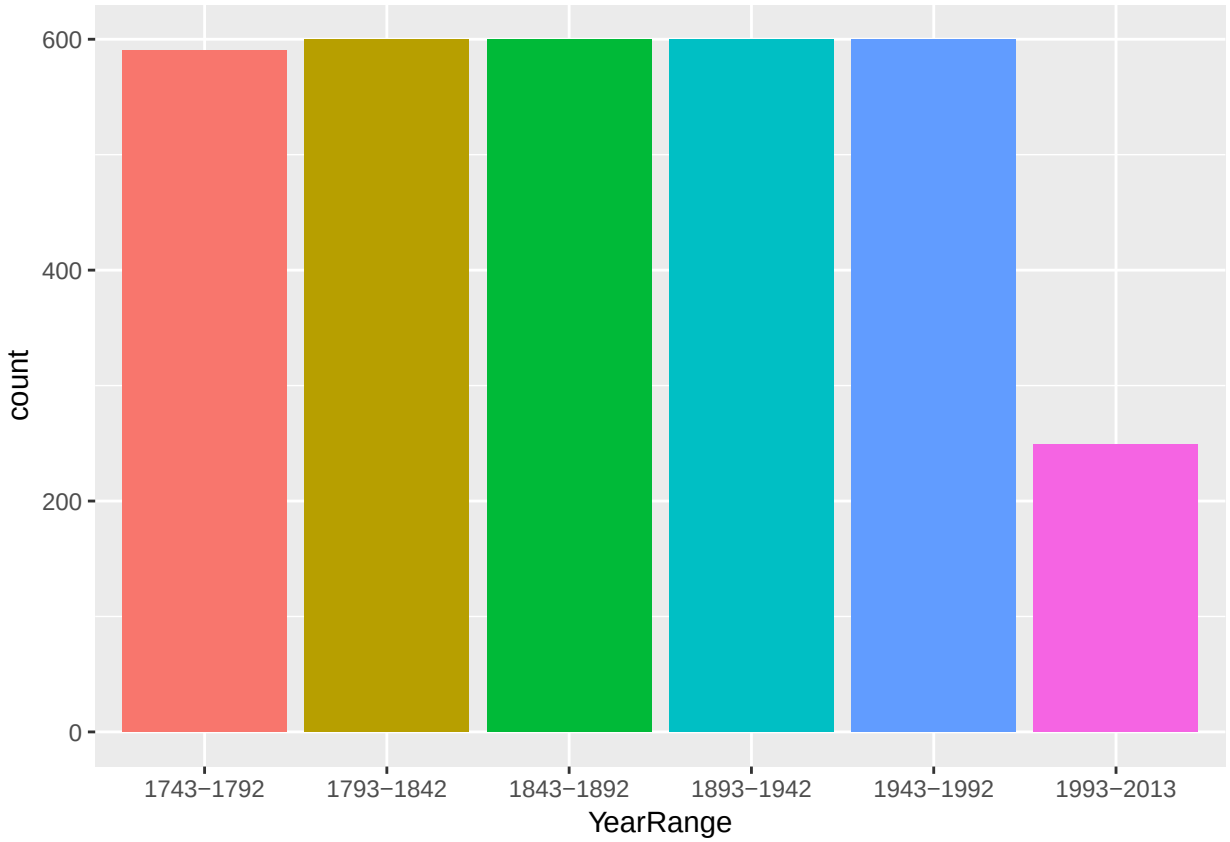
Year Range and Seasons Distribution

To start with, let's look at how spread out the observations are across `Year ranges` and `Season`, to see how distributed the data collected has been.

Year Range Distribution

Table 1: Distribution of Observations across Year Ranges

Year Range	# of Observations
1743-1792	590
1793-1842	600
1843-1892	600
1893-1942	600
1943-1992	600
1993-2013	249

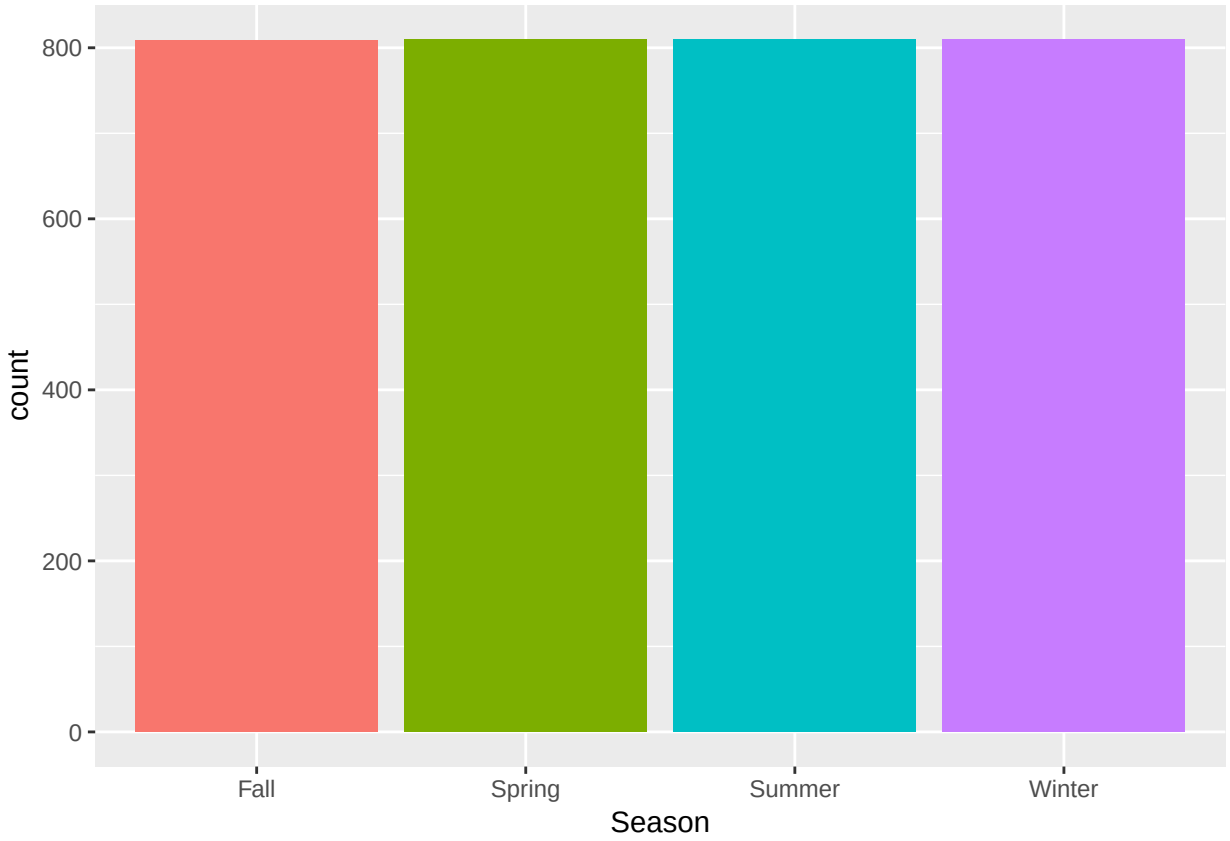


It appears we have a fairly consistent distribution across all the year ranges, except for notably the latest range, likely due to the smaller scope of its range.

Season Distribution

Table 2: Distribution of Observations across Seasons

Season	# of Observations
Fall	809
Spring	810
Summer	810
Winter	810



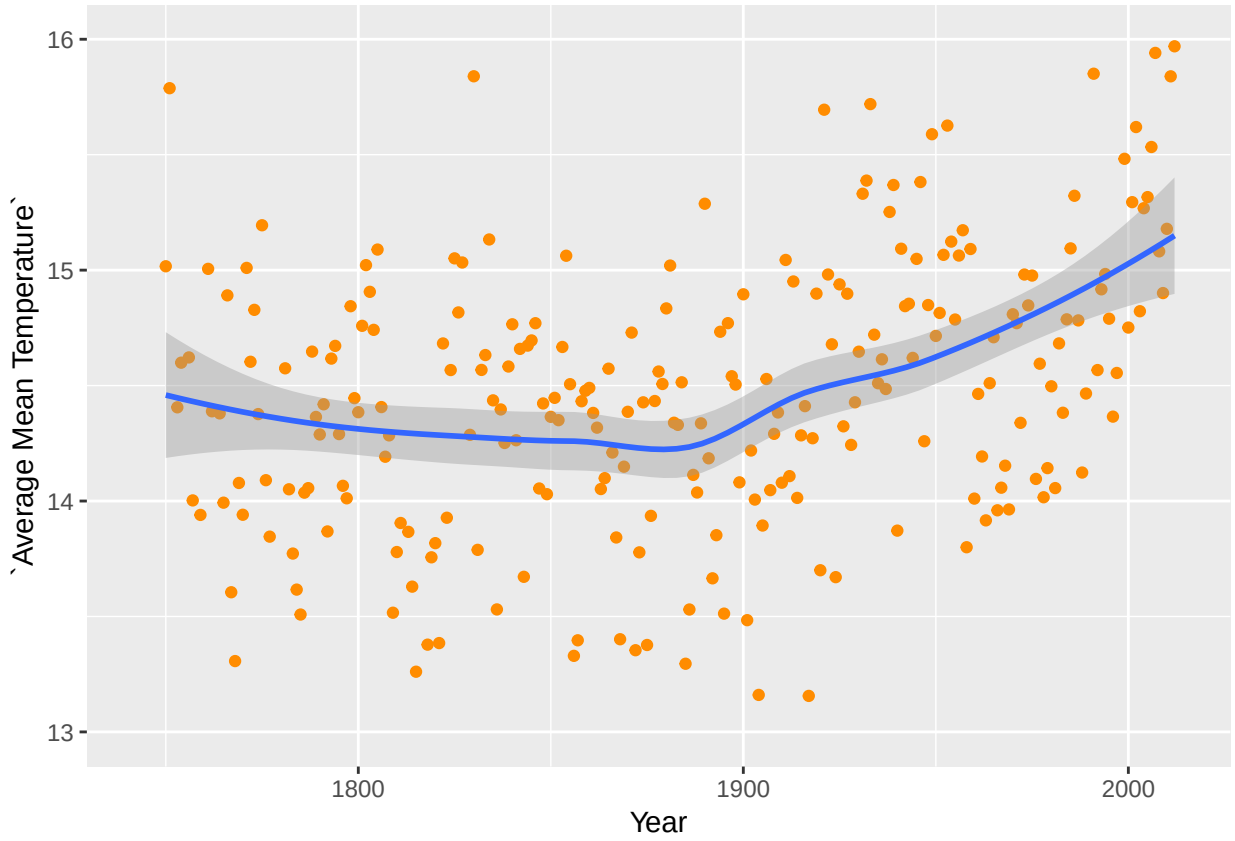
Fortunately, the seasons are uniformally distributed. In this case, it means that while we created the variables of `Year Range` and `Season`, the uniform or near uniform distributions of both mean further evaluations aren't much skewed across conditions.

Visualizations

Average Temperature

Let's start visualizing our data by asking the simple question: does it appear that average temperatures in Raleigh have risen since 1743?

Average Temperature across years

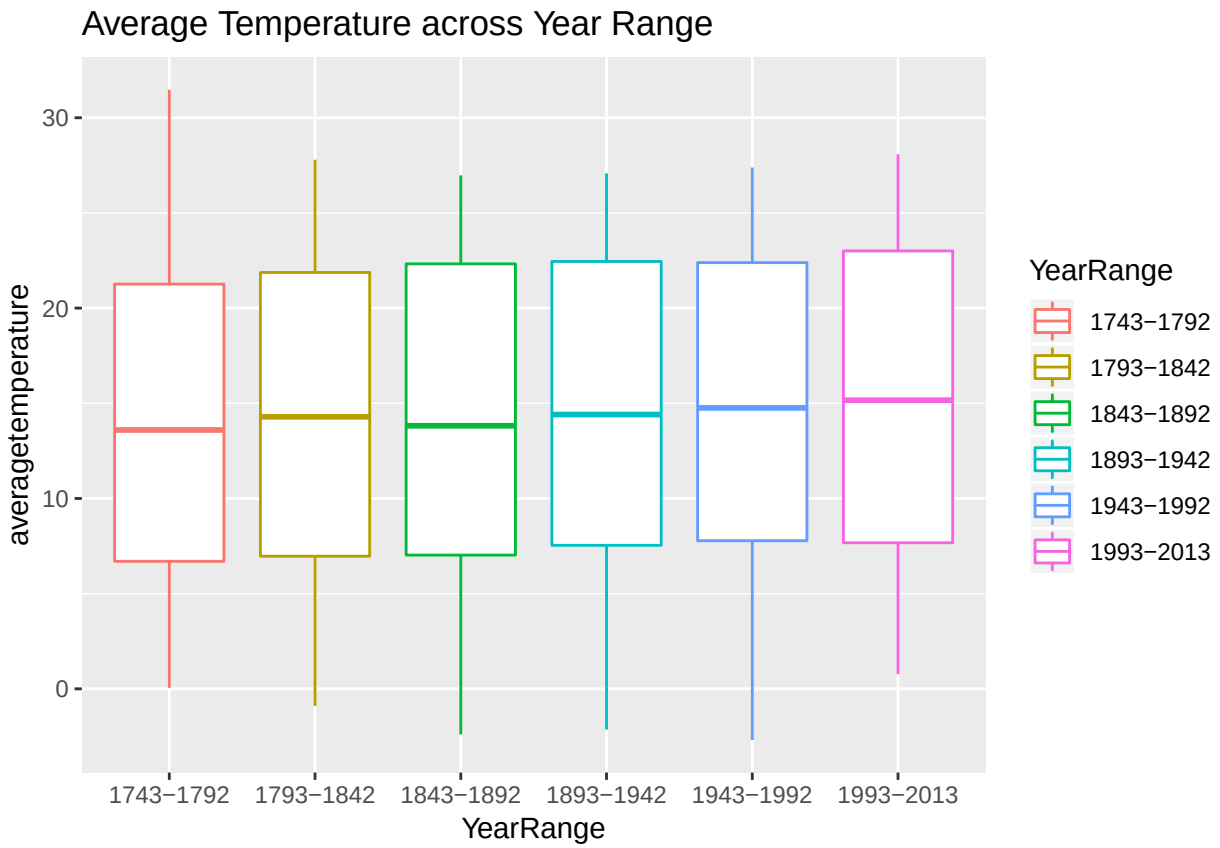


Short answer: yes!

While there appears to be an upward trend over time, there's a great apparent variance across the time series. For visualization sake, let's see what happens if we average the averages across time periods and go from there.

Average Temperature across year ranges

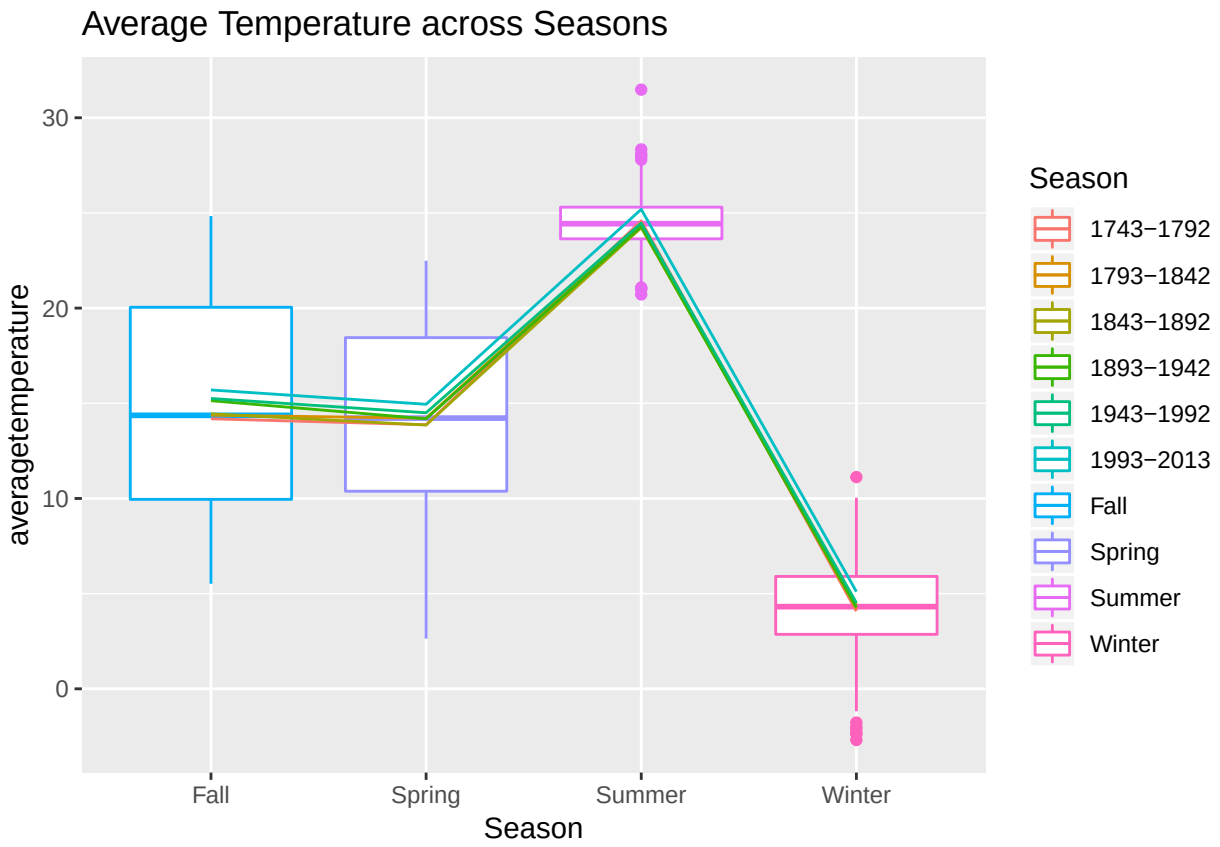
Here we'll look at the average temperatures across year ranges. This could get graphic...



This graph supports the same pattern that the smoothing line in the previous graph indicated, indicating an upward trend.

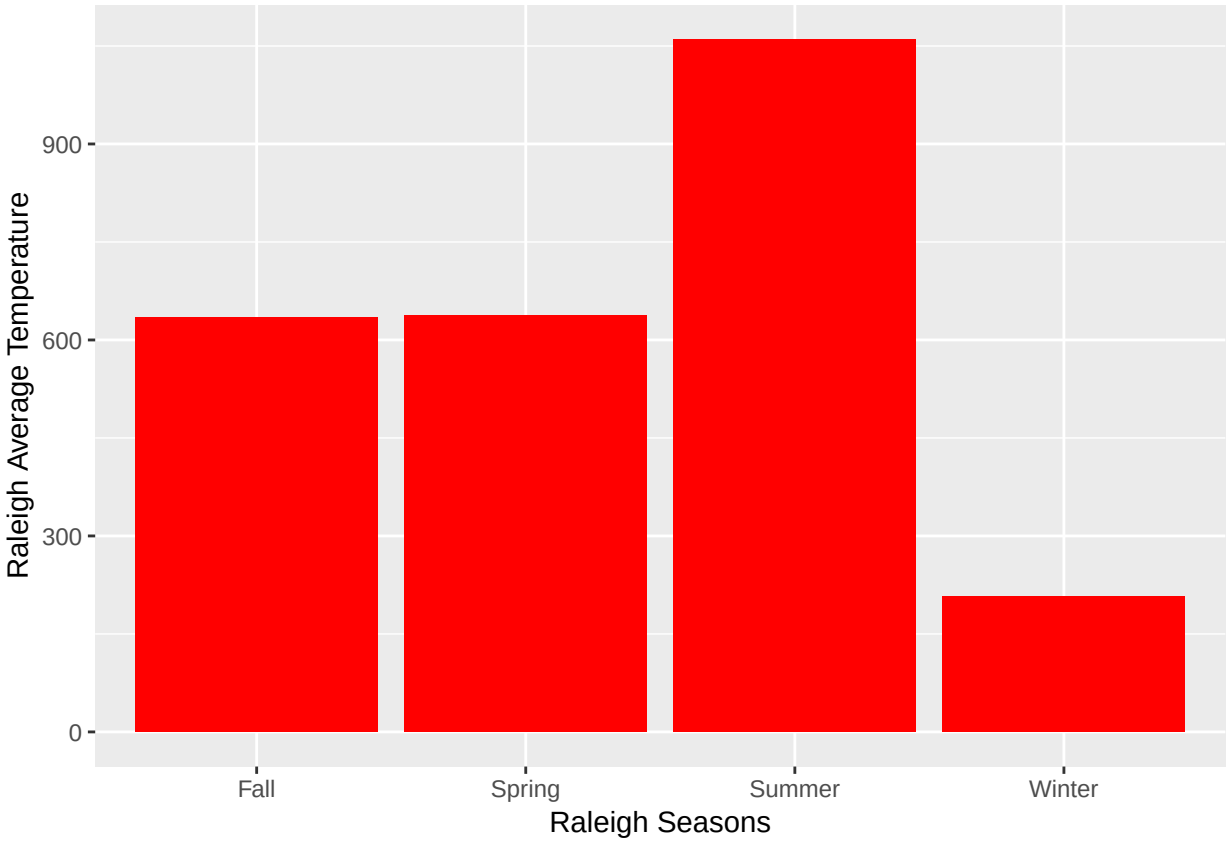
Average Temperature across seasons

Here we'll look at average temperatures across the seasons in Raleigh.



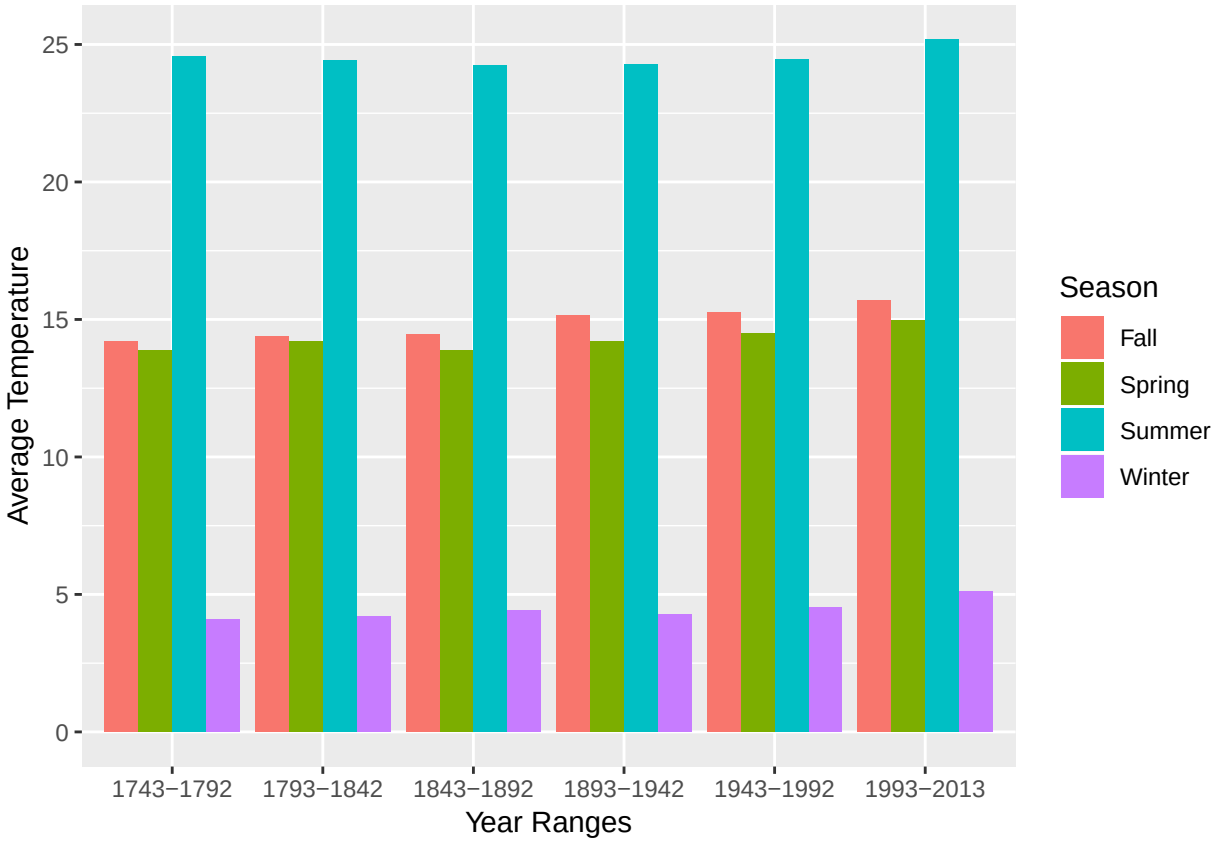
To my surprise as a Raleigh local, the winters are historically colder than the summers! Sometimes its hard to tell the difference...

The chart below shows the averages for the different seasons. If we filter down to the specific seasons for 2000 and beyond, we can see what year had the highest and lowest average temperature in said season.

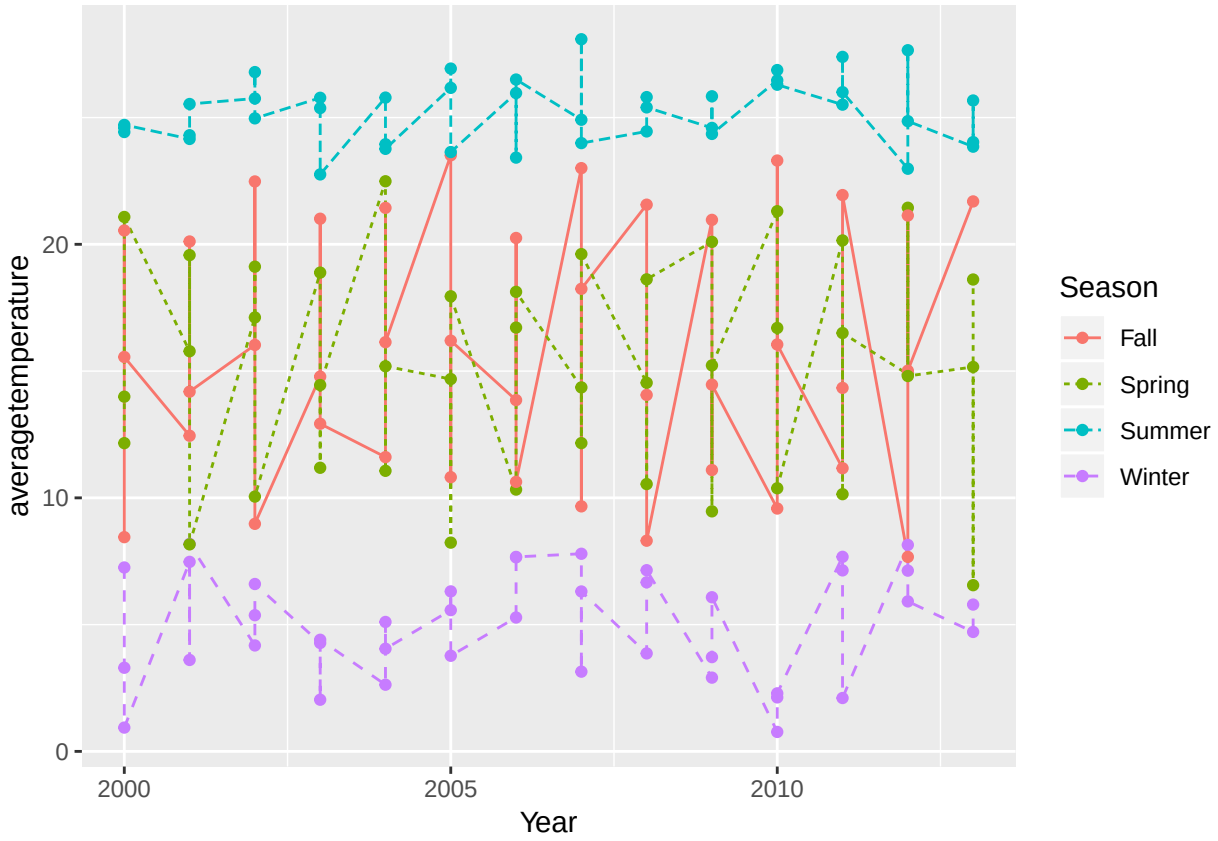


Average Temperature across seasons and year ranges

Here we'll group the average temperatures across year ranges and seasons and look at them side by side.

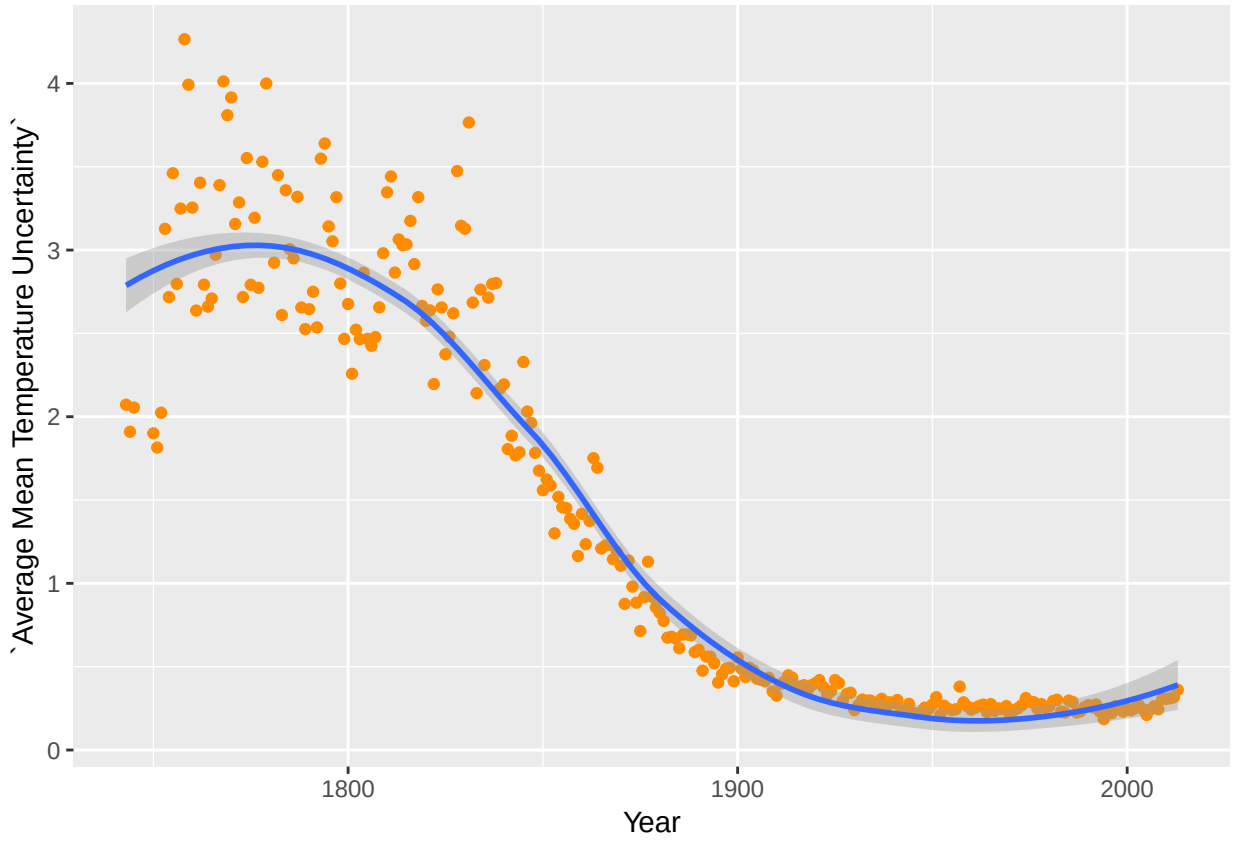


The following chart shows a line chart starting in 2000 and ending in 2013 for each season. We see the biggest spike at 28.089 degrees in 2007 and biggest drop in 2010. We see that like the bar chart above that Spring and Fall fall relatively close together.



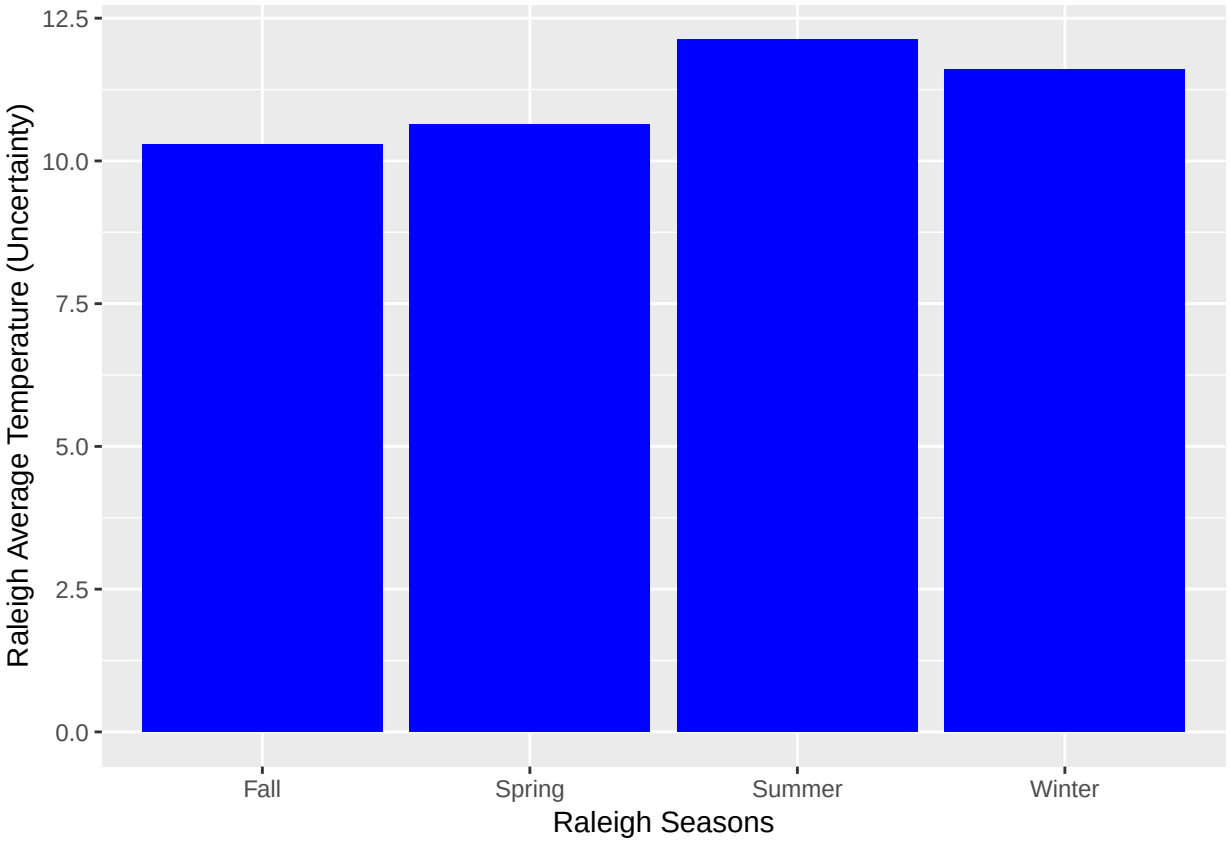
Trend of Average Temperature Uncertainty

As history presses onward, the uncertainty associated with the recorded values of average temperature should go down, no? Here we'll look at the average temperature uncertainty over the years.

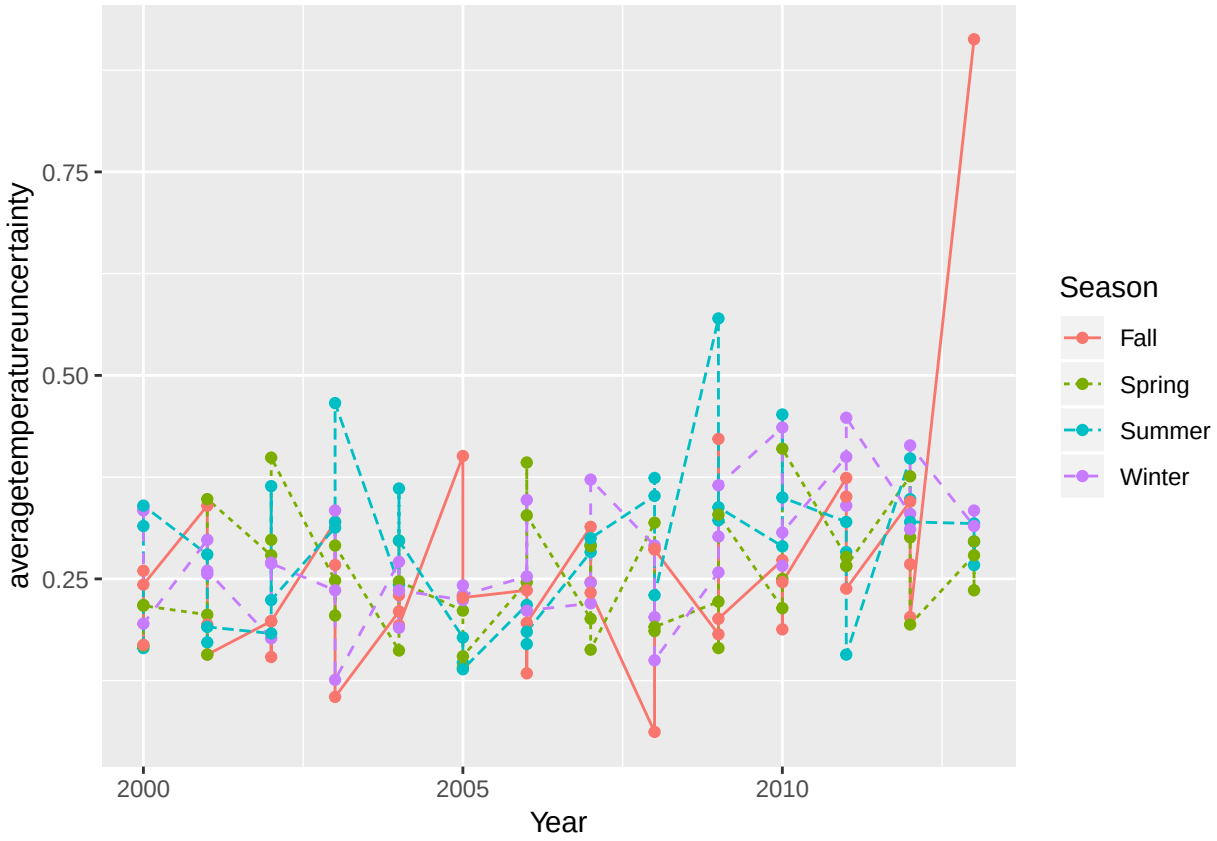


We see here how the relationship between Year and Uncertainty is quadratic! Here's to the apparent rising uncertainty measurements of our more recent years changing back soon...

The graph below shows the season values for the average temperatures under uncertainty. We see how summer is only slightly higher compared to the above bar chart.

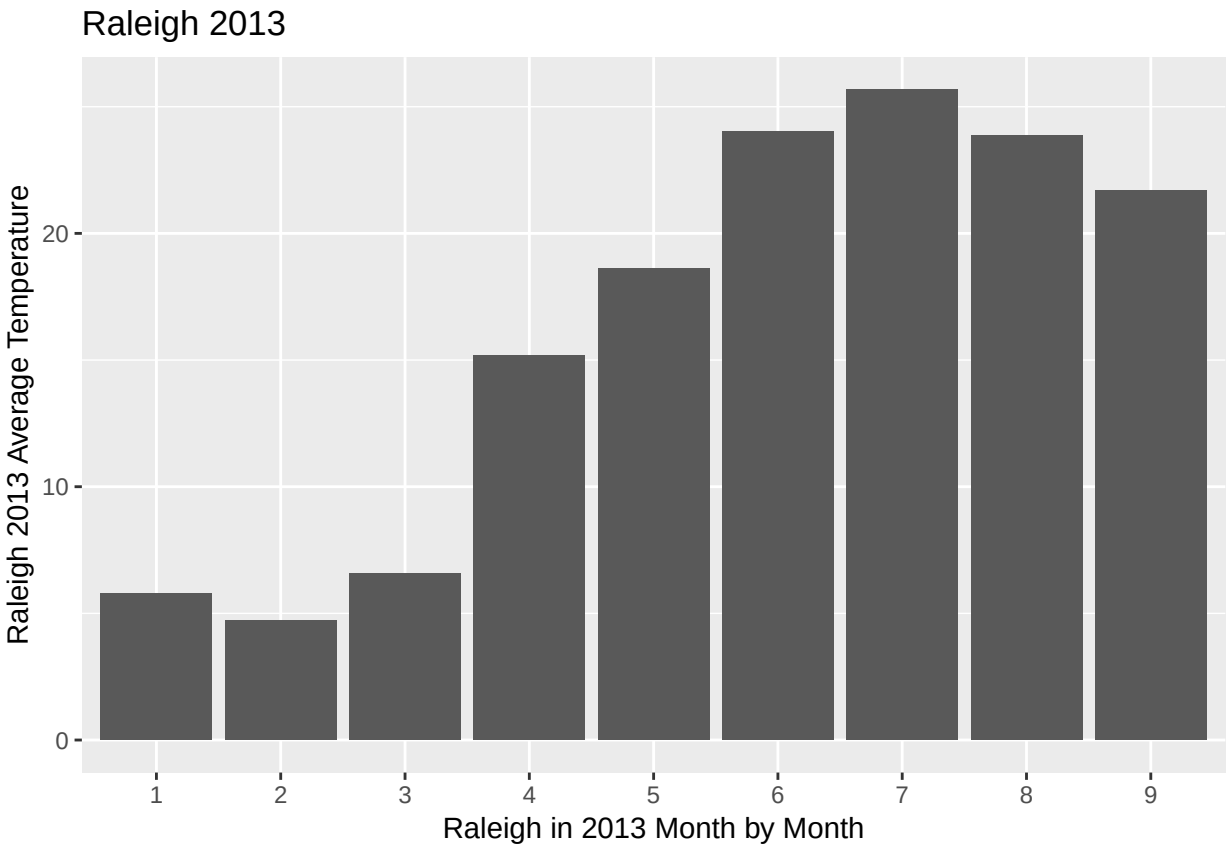


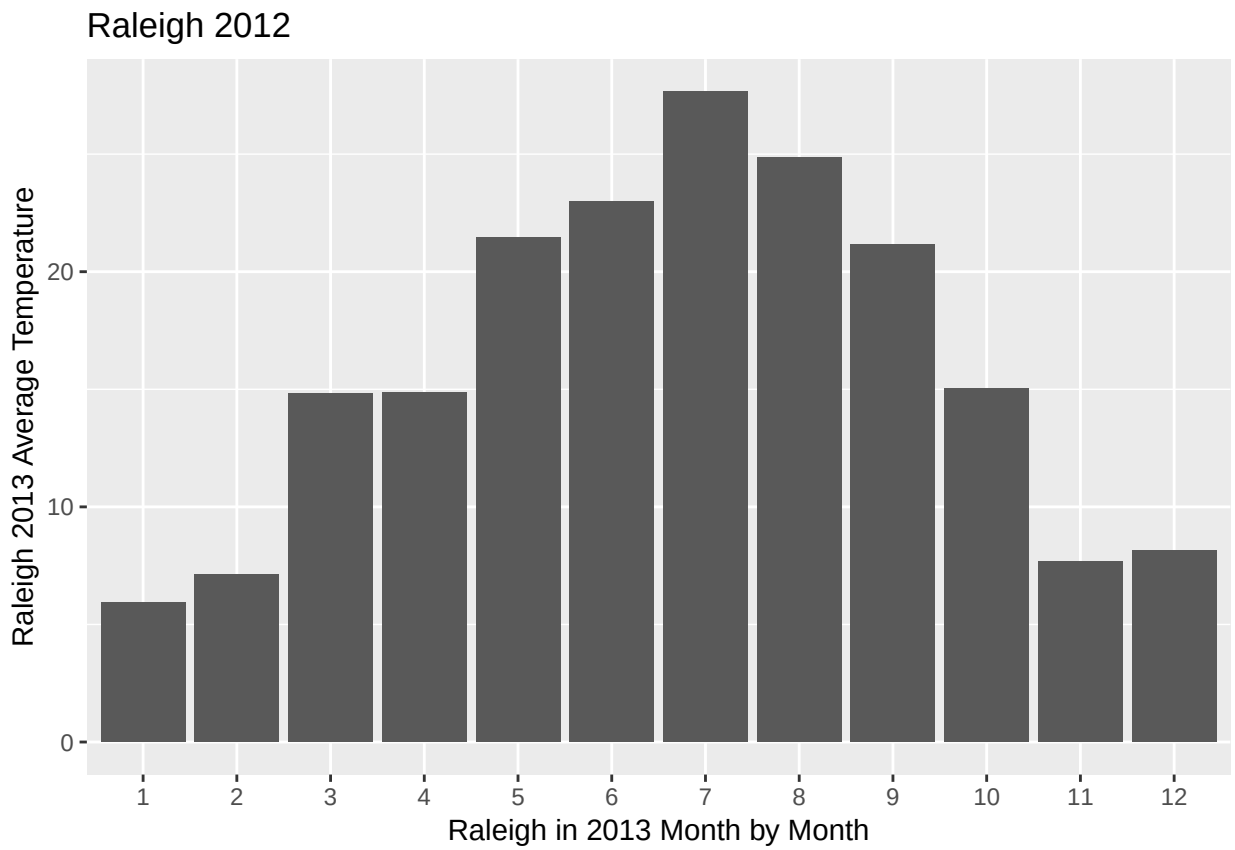
We see from the line graph below how the temperatures are relatively close together in the 2000s, but we see a big spike in temperature in the Fall. This could be a potential outlier that we could consider to remove if we wanted to do a prediction model.



Average Temperature 2012 and 2013

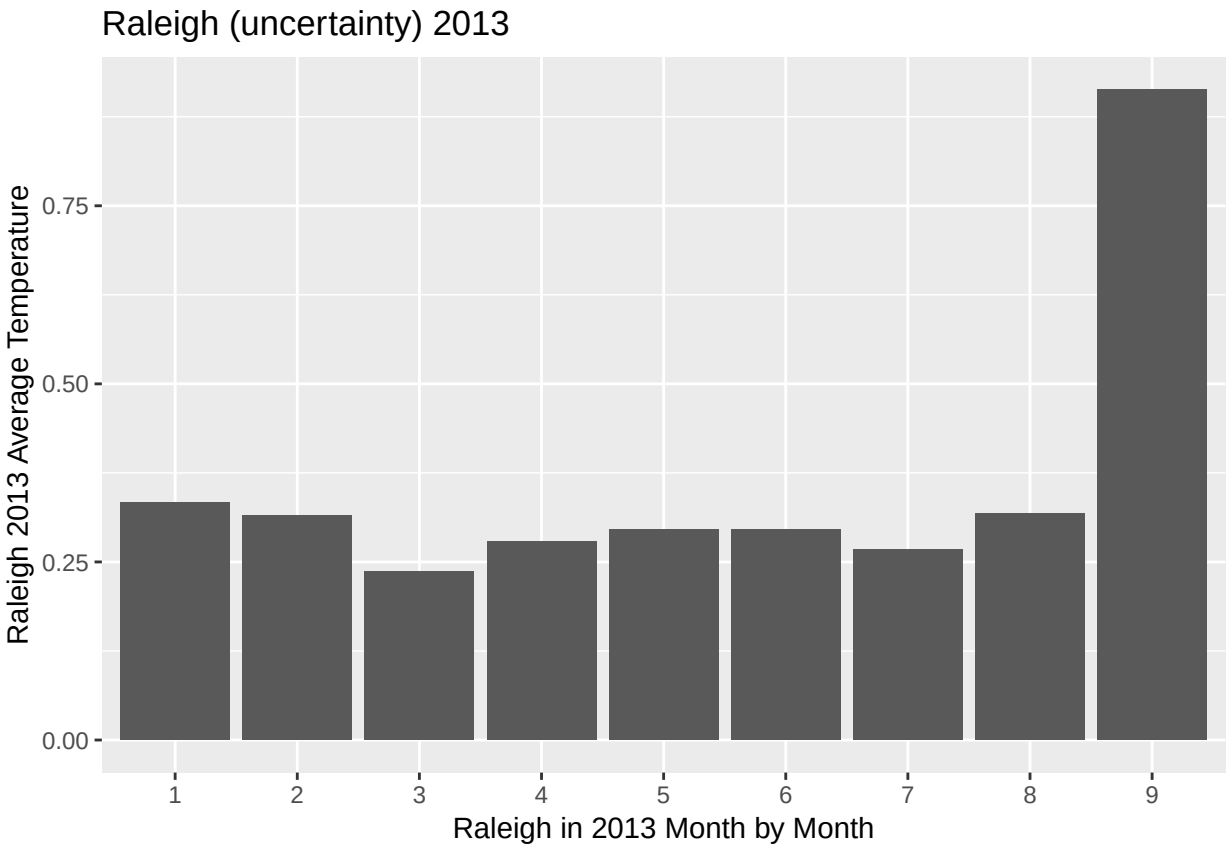
This looks at the year distribution of 2013 month by month. Notice how it shows a left skewed distribution with our hottest month in July. Also, if we look at the graph below it we see the month to month comparison of 2012. We see that July is still the hottest month, but it shows to be more normally distributed.

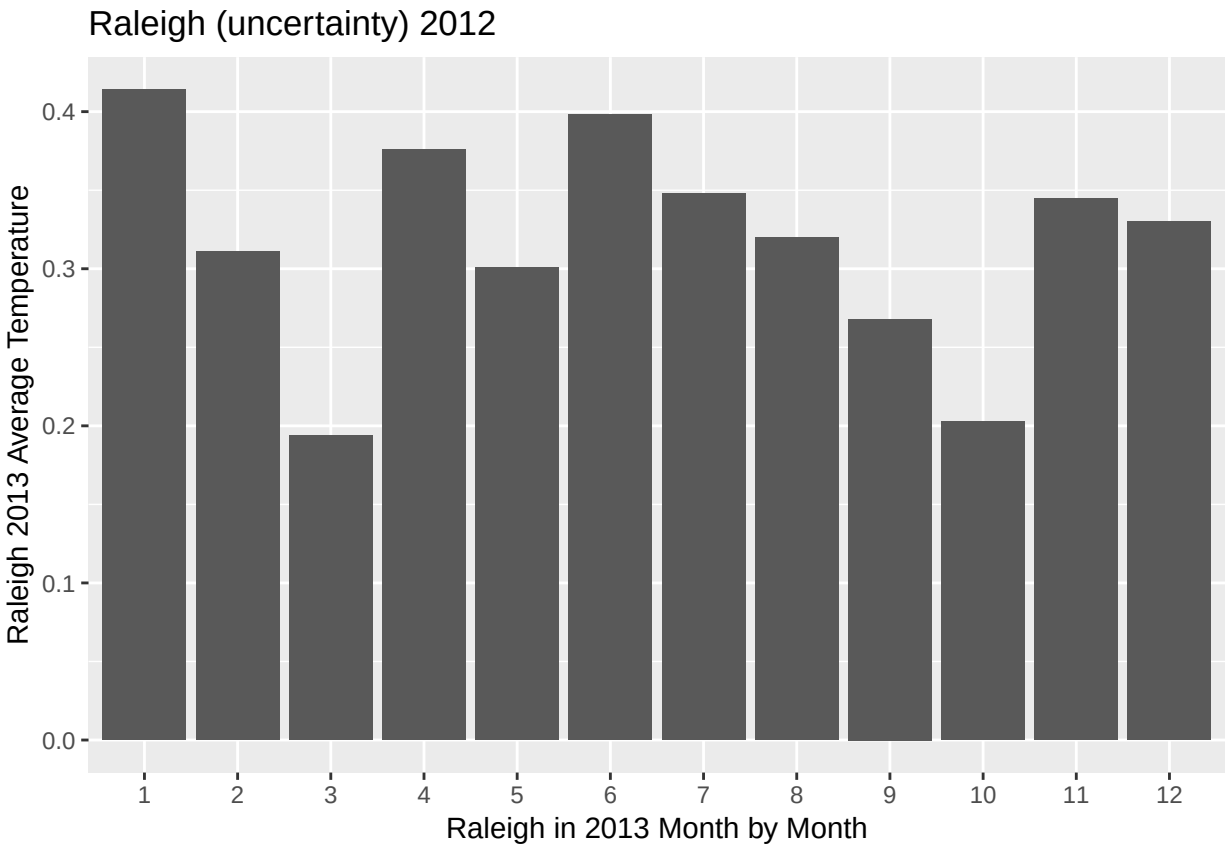

Raleigh 2013
Raleigh 2013 Average Temperature
0
10
20
1
2
3
4
5
6
7
8
9
Raleigh in 2013 Month by Month



Average Temperature (uncertainty) 2012 and 2013

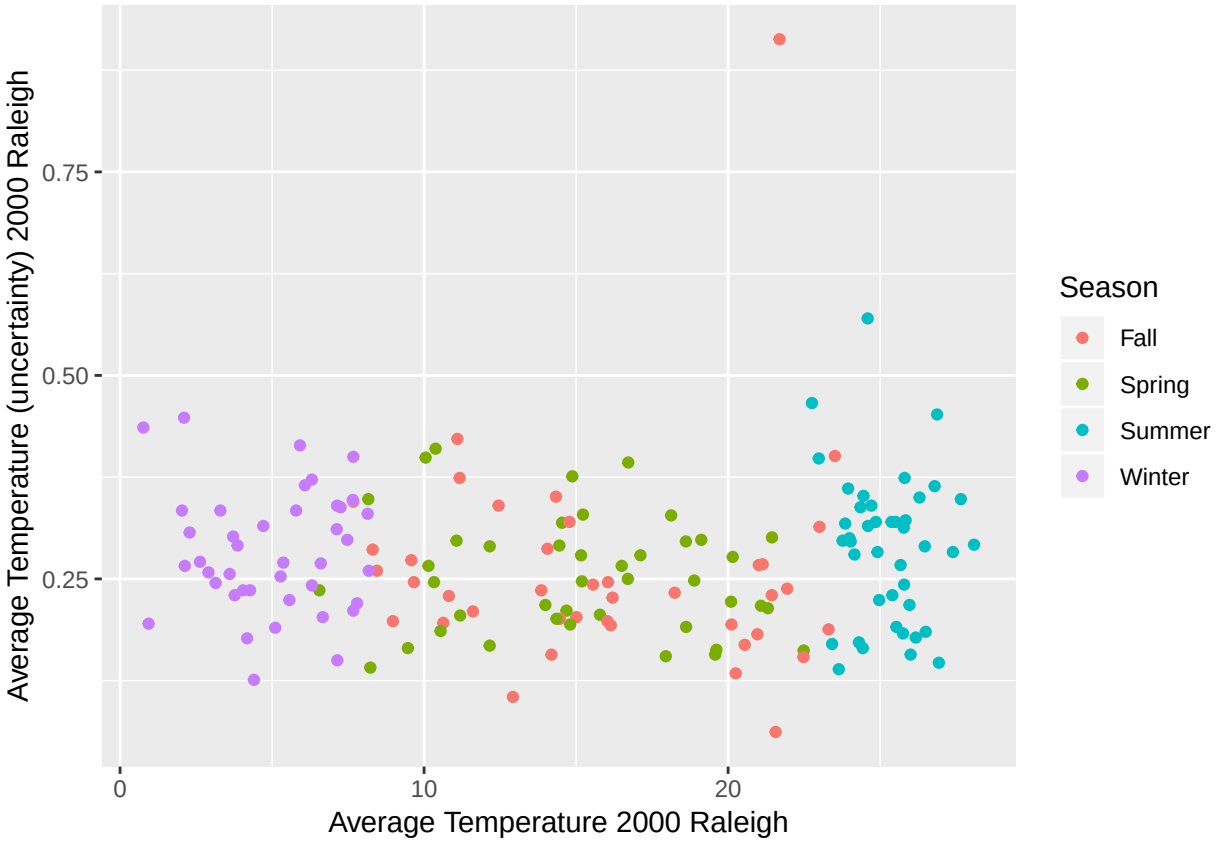
Looking at 2013, the average temperature shows to be pretty steady but we clearly see an outlier in September. Looking at 2012, the temperature shows to be high in January and doesn't show a big increase until June and then drops again until November.


Raleigh (uncertainty) 2013
Raleigh 2013 Average Temperature
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
Raleigh in 2013 Month by Month



Average Temperature and Uncertainty

Below shows a scatterplot of average temperature versus average temperature (uncertainty). We see how most points show at the bottom with the exception of the potential outlier in the Fall like mentioned above.



Conclusion

We see how JSON are efficient files that are pretty easy to handle with R to access data sets saved online. We looked at one such data set to examine average temperatures and measurement uncertainty across Raleigh's history and in more recent years. In conclusion, R's your friend when it comes to handling JSON files, Raleigh's temperatures seem to be on the rise, which is certainly evident to us locals when we sweat on Christmas Eve.