Project 1: Exploring Weather Trends

## Outline

* Purpose
* Tools and Methods
* Data
* Analysis
* Results

### Purpose

Describing the similarities and differences between global temperature trends, and temperature trends in the closest big city to where I live, San Jose and my hometowns, Los Angeles and Taipei.

### Tools and Methods

I wrote SQL queries for extracting the city and global data from database, and saved them as csv files. The report was generated by RMarkdown. I used the libraries readr, dplyr, ggplot and caret to manipulate the data, and create line charts

### Data

I extracted the data from the database, and picked my current location, San Jose, and my hometowns, Los Angeles and Taipei.

* city list includes the variables, city and country  
  SELECT \*  
  FROM city\_list  
  ORDER BY country, city
* city data includes the variables, city, country, year, avg\_temp  
  SELECT \*  
  FROM city\_data  
  WHERE city IN (‘San Jose’, ‘Los Angeles’,‘San Francisco’) AND country = ‘United States’
* city data

SELECT \* FROM city\_data WHERE city = ‘Taipei’

* global data includes the variables, year and avg\_temp  
  SELECT \*  
  FROM global\_data

### Analysis

1. I imported the csv files into R.

Note: I only kept the period of time from 1849 to 2013 for each dataset because it is easy and fair for comparison.

1. I created a function for calculating moving averages by using global dataset. I calculated 3 year, 5 year, 7 year and 10 year moving averages and generated a line chart for all of them. In the line charts, the lines of 3 year and 5 year look relatively rough compared to 7 year and 10 year. However, the line of 10 year is too smooth and takes out some of detail. Thus, I found that the line of 7 year moving average is smooth and does not lack too much of information of the data.

Note: During the calculation of moving average, 1st to 6th year in the dataset were missing because 7 year moving average is calculated from 7th year. To avoid showing the line of NA group in line charts, I removed them from the datasets.

1. I generated the following line charts.

* 7 year moving average in Global
* 7 year moving average in San Jose
* 7 year moving average in Global versus San Jose
* 7 year moving average in Los Angeles
* 7 year moving average in Taipei
* 7 year moving average in the world, San Jose, Los Angeles and Taipei

### Results

1. Compared the 7 year moving average temperature in my current city, San Jose with in the world from 1849 to 2013

* The temperature trends in the world and San Jose has gradually become hotter and hotter, especially from 1975 to 2013.
* The temperature in San Jose has not been consistent going up. Before 1975, the temperature has been dramatically up and down. The temperature in the world has been consistent going up, especially from 1975 to 2013.
* The average temperature in San Jose is hotter than in the world, around 6 degrees difference.
* In the world, the year is highly and positively correlated with average temperature than in San Jose.
* Using linear regression for estimating the average temperature in San Jose based on the global temperature is not accurate. The estimated temperatures are very off from the actual temperatures, might choose different models in the future.

1. Compared the 7 year moving average temperature among the world, San Jose, Los Angeles and Taipei from 1849 to 2013.

* The temperature in Taipei has been the hottest over the time compared to the world, San Jose and Los Angeles.
* The temperatures in San Jose and in Los Angeles have the similar trends over the years even though the temperature in Los Angeles has been always higher than in San Jose over the years.
* The temperatures in San Jose, Los Angeles, and Taipei are all higher than in the world.
* The world has highest correlation between year and average temperature compared to San Jose, Los Angeles and Taipei.

## 1) Reading datasets  
library(readr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

# a. Global   
global <- read\_csv("global.csv")

## Parsed with column specification:  
## cols(  
## year = col\_integer(),  
## avg\_temp = col\_double()  
## )

# Make the time period the same like San Jose and Los Angeles  
global2 <- global %>%  
 filter(year >= 1849 & year <= 2013)  
  
# b. My current Location: San Jose  
cities <- read\_csv("cities.csv")

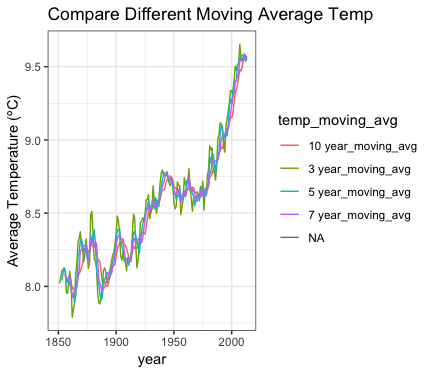
## Parsed with column specification:  
## cols(  
## year = col\_integer(),  
## city = col\_character(),  
## country = col\_character(),  
## avg\_temp = col\_double()  
## )

sj <- cities %>%  
 filter(city == 'San Jose') %>%  
 select(year,avg\_temp)  
  
# c. The city I stayed before the Bay: Los Angeles  
la <- cities %>%  
 filter(city == 'Los Angeles') %>%  
 select(year,avg\_temp)  
  
# d. My hometown: Taipei  
tpi <- read\_csv("tpi.csv") %>%  
 filter(year >= 1849 & year <= 2013) %>%  
 select(year,avg\_temp)

## Parsed with column specification:  
## cols(  
## year = col\_integer(),  
## city = col\_character(),  
## country = col\_character(),  
## avg\_temp = col\_double()  
## )

## 2) Create a function for calculating moving average  
move <- function(data,loc,y){  
 mov\_year = data.frame()  
 locat = data.frame()  
 year = data.frame()  
 movavg = data.frame()  
 for(i in 1:length(data$avg\_temp)){  
 if (i < y){  
 x <- NA  
 z <- NA  
 }  
 else{  
 x <- data %>%  
 summarize(movmean = mean(avg\_temp[(i-(y-1)):i]))  
 z <- paste(y,'year\_moving\_avg')  
 }  
 movavg = rbind(movavg,data.frame(as.numeric(x)))  
 year = rbind(year,z)  
 locat = rbind(locat,loc)  
 }  
 names(locat) <- 'location'  
 names(movavg) <- 'mov\_avg'  
 names(year) = 'temp\_moving\_avg'  
 mov\_year <- cbind(data,movavg,year,locat)  
 return(mov\_year)  
}  
  
globalall <- rbind(move(global2,'Global',3),move(global2,'Global',5),move(global2,'Global',7),move(global2,'Global',10))  
  
library(ggplot2)  
ggplot(data=globalall, aes(x=year, y=mov\_avg, color=temp\_moving\_avg))+geom\_line()+  
 ylab("Average Temperature (ºC)")**+**theme\_bw()+ggtitle("Compare Different Moving Average Temp")

## Warning: Removed 21 rows containing missing values (geom\_path).



# Remove the NAs  
global7 <- move(global2,'Global',7) %>%  
 filter(year >= 1855)  
head(global7)

## year avg\_temp mov\_avg temp\_moving\_avg location  
## 1 1855 8.11 8.074286 7 year\_moving\_avg Global  
## 2 1856 8.00 8.077143 7 year\_moving\_avg Global  
## 3 1857 7.76 8.057143 7 year\_moving\_avg Global  
## 4 1858 8.10 8.045714 7 year\_moving\_avg Global  
## 5 1859 8.25 8.067143 7 year\_moving\_avg Global  
## 6 1860 7.96 8.055714 7 year\_moving\_avg Global

sj7 <- move(sj,'San Jose',7) %>%  
 filter(year >= 1855)  
head(sj7)

## year avg\_temp mov\_avg temp\_moving\_avg location  
## 1 1855 14.20 14.10000 7 year\_moving\_avg San Jose  
## 2 1856 14.10 14.09714 7 year\_moving\_avg San Jose  
## 3 1857 14.78 14.23714 7 year\_moving\_avg San Jose  
## 4 1858 14.19 14.20857 7 year\_moving\_avg San Jose  
## 5 1859 13.71 14.19429 7 year\_moving\_avg San Jose  
## 6 1860 13.81 14.11000 7 year\_moving\_avg San Jose

glo\_sj <- rbind(global7,sj7)  
  
la7 <- move(la,'Los Angeles',7) %>%  
 filter(year >= 1855)  
head(la7)

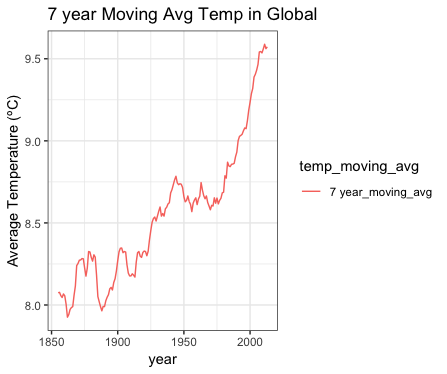
## year avg\_temp mov\_avg temp\_moving\_avg location  
## 1 1855 15.94 15.72571 7 year\_moving\_avg Los Angeles  
## 2 1856 15.52 15.69857 7 year\_moving\_avg Los Angeles  
## 3 1857 16.19 15.82857 7 year\_moving\_avg Los Angeles  
## 4 1858 15.67 15.84857 7 year\_moving\_avg Los Angeles  
## 5 1859 15.29 15.80286 7 year\_moving\_avg Los Angeles  
## 6 1860 15.41 15.68000 7 year\_moving\_avg Los Angeles

tpi7 <- move(tpi,'Taipei',7) %>%   
 filter(year >= 1855)  
head(tpi7)

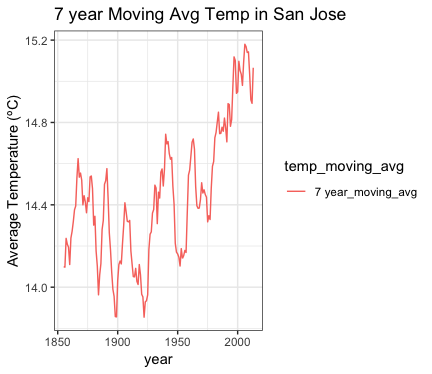
## year avg\_temp mov\_avg temp\_moving\_avg location  
## 1 1855 21.89 21.94714 7 year\_moving\_avg Taipei  
## 2 1856 21.54 21.92714 7 year\_moving\_avg Taipei  
## 3 1857 21.94 21.93571 7 year\_moving\_avg Taipei  
## 4 1858 21.67 21.91000 7 year\_moving\_avg Taipei  
## 5 1859 21.87 21.90714 7 year\_moving\_avg Taipei  
## 6 1860 21.70 21.83286 7 year\_moving\_avg Taipei

glob\_city <- rbind(global7,sj7,la7,tpi7)

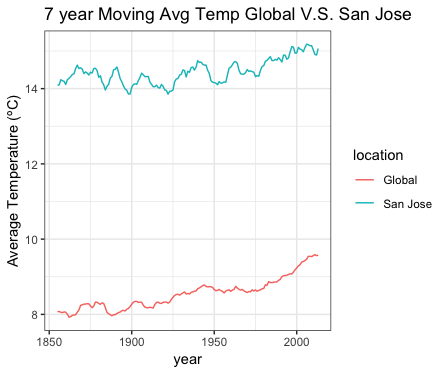
## 3) Data Visualization  
library(ggplot2)  
  
ggplot(data=global7, aes(x=year, y=mov\_avg, color=temp\_moving\_avg))+geom\_line()+  
 ylab("Average Temperature (ºC)")+theme\_bw()+ggtitle("7 year Moving Avg Temp in Global")



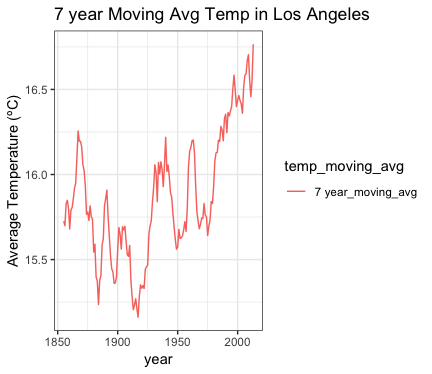
ggplot(data=sj7, aes(x=year, y=mov\_avg, color=temp\_moving\_avg))+geom\_line()+  
 ylab("Average Temperature (ºC)")+theme\_bw()+ggtitle("7 year Moving Avg Temp in San Jose")



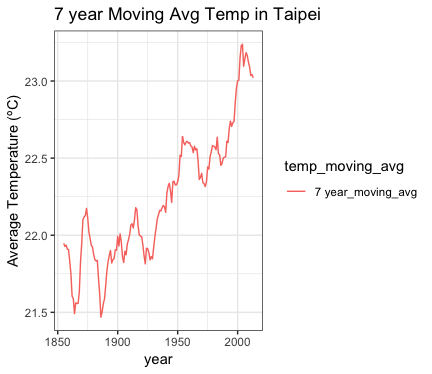
ggplot(data=glo\_sj, aes(x=year, y=mov\_avg, color=location))+geom\_line()+  
 ylab("Average Temperature (ºC)")+theme\_bw()+ggtitle("7 year Moving Avg Tem Global V.S. San Jose")



ggplot(data=la7, aes(x=year, y=mov\_avg, color=temp\_moving\_avg))+geom\_line()+  
 ylab("Average Temperature (ºC)")+theme\_bw()+ggtitle("7 year Moving Avg Temp in Los Angeles")



ggplot(data=tpi7, aes(x=year, y=mov\_avg, color=temp\_moving\_avg))+geom\_line()+  
 ylab("Average Temperature (ºC)")+theme\_bw()+ggtitle("7 year Moving Avg Temp in Taipei")



ggplot(data=glob\_city, aes(x=year, y=mov\_avg, color=location))+geom\_line()+  
 ylab("Average Temperature (ºC)")+theme\_bw()+ggtitle("7 year Moving Avg Temp")

