Bio4158 - Devoir 3

Michael Pham (300129636), Yacine Marouf (300112014), Tristan Lachance (300059877)

07 October, 2022

Here are the needed packages

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.7     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(performance)
library(see)
library(patchwork)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(pwr)
library(broom)
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(Rmisc)
```

```
## Loading required package: lattice
```

Loading data into R

```
setwd("/Users/apple/Desktop/BIO_4158 /BIO 4158 lab/Data sets BIO 4158")
getwd()
```

```
## [1] "/Users/apple/Desktop/BIO_4158 /BIO 4158 lab/Data sets BIO 4158"
```

```
climate <- read_csv("climate.csv")
```

```
## Rows: 348 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): loc
## dbl (3): reg, year, ave
##
## 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.
```

```
head(climate)
```

```
## # A tibble: 6 x 4
##     reg loc    year   ave
##   <dbl> <chr> <dbl> <dbl>
## 1     1 A      1950  NA
## 2     1 A      1951 -18.6
## 3     1 A      1952 -17.6
## 4     1 A      1953 -16.8
## 5     1 A      1954 -17.8
## 6     1 A      1955 -18.2
```

Answer the following questions. Total marks = 30.

1. What is the 'biological' (scientific) hypothesis? (1 point)

Climate change is observed to be warming. There is a variation in changes in temperature across different geographic location, and this pheonomena is causes by the variation in human activities across region

2. What does this hypothesis predict in the context of her study? (2 points)

This hypothesis predicts that warming is more prevalent in the Northern area compared to the Southern Area

3. What is the associated null hypothesis? (2 points)

H_o: North Area and South Are have the same increase in temparature

4. To test this hypothesis, the student will need to quantify the extent of any climate change separately for each location. She's interested not only in the point estimate of the rate of climate change for each location (which she needs for testing her biological hypothesis above), but also in its significance in each location (i.e. she wants to separately test whether there is evidence that climate has changed in any way at each site). Let's start by conducting a regression analysis to test whether there is evidence of climate change in **Yarmouth**

a. What is the statistical null hypothesis relevant to the analysis ofthe Yarmouth data? (2 points)

The rate of change in temaperature is 0

b) What is the statistical model in the verbal form ?

Annual_Average_temperature = Year+ Error

c) Provide a relevant plot of the Yarthmouth data

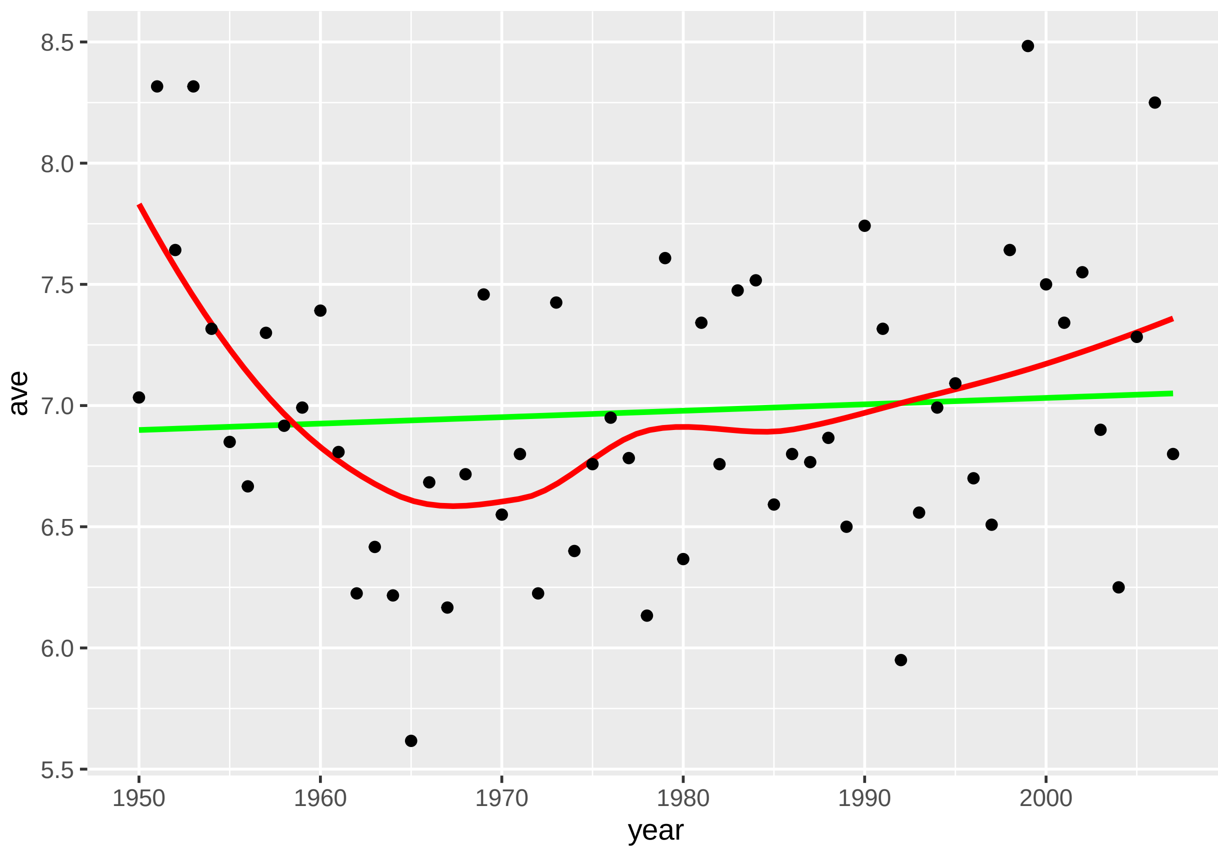
In this case, we would use the scatter plot

```r
    Yarmouth <- climate %>% filter(loc=="Y") %>% arrange(year)

head(Yarmouth)
```

```
## # A tibble: 6 x 4
##     reg loc    year   ave
##   <dbl> <chr> <dbl> <dbl>
## 1     0 Y      1950  7.03
## 2     0 Y      1951  8.32
## 3     0 Y      1952  7.64
## 4     0 Y      1953  8.32
## 5     0 Y      1954  7.32
## 6     0 Y      1955  6.85
```

```r
mygraph <- ggplot(
  data = Yarmouth[!is.na(Yarmouth$ave), ], # source of data
  aes(x = year, y = ave)
)
# plot data points, regression, loess trace
mygraph <- mygraph +
  stat_smooth(method = lm, se = FALSE, color = "green") + # add linear regression, but no SE shading
  stat_smooth(color = "red", se = FALSE) + # add loess
  geom_point() # add data points

mygraph
```

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

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



d) Fit the model with and shows the R output

```
RegModel.1 <- lm(ave ~ year, data = Yarmouth)
summary(RegModel.1)
```

```
##
## Call:
## lm(formula = ave ~ year, data = Yarmouth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3221 -0.4016 -0.1265  0.4505  1.4545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.732995   9.466686   0.183    0.855
## year        0.002649   0.004785   0.554    0.582
##
## Residual standard error: 0.61 on 56 degrees of freedom
## Multiple R-squared:  0.005445,   Adjusted R-squared:  -0.01231
## F-statistic: 0.3066 on 1 and 56 DF,  p-value: 0.582
```

e) State the statistical assumptions for the analysis of Yarmouth data and provide some evidence that you examined them

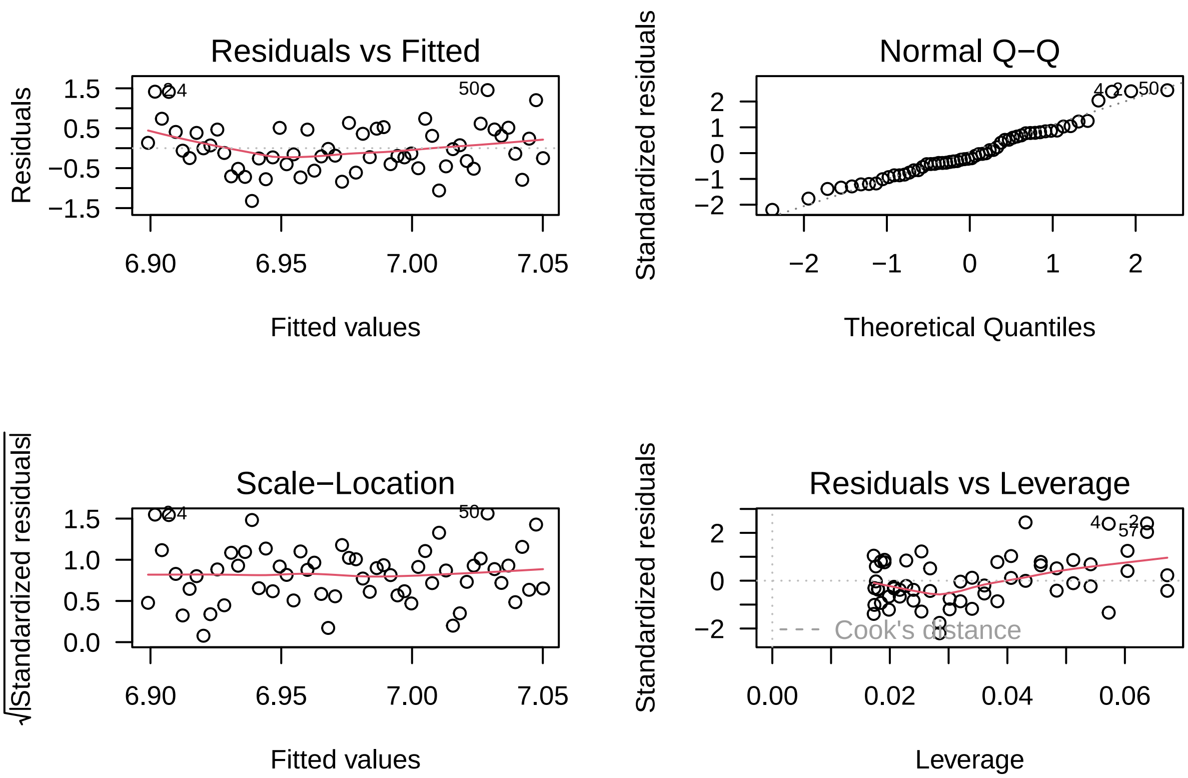
There are several assumptions that we have to make sure:

- No errors in X measurements
- Relationship between X and Y is linear
- Independence of residuals (no serial aurocorrelation)
- Residuals are normally distributed
- Homoscedasticity of residuals (even spread of residuals on X-axis)

We will provide the graphs for visualization purpose only. We would not use them to intepret the valdity of our assumption

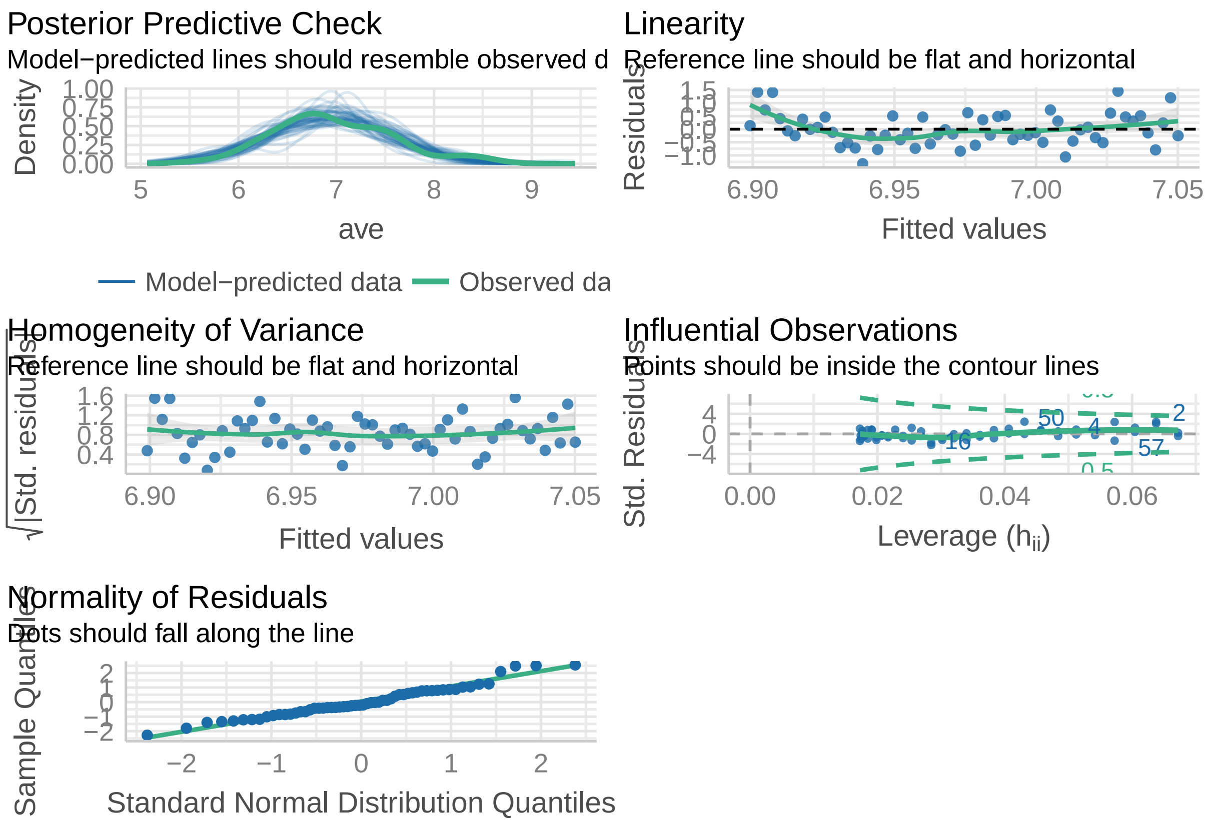
residuals plot

```
par(mfrow = c(2, 2), las = 1)
plot(RegModel.1)
```



Linearity plots

```
check_model(RegModel.1)
```



From here, we will start checking the validity of the assumptions using different types of test

We use the Breusch-Pagan test examines whether the variability of the residuals is constant with respect to increasing fitted values

```
bptest(RegModel.1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  RegModel.1
## BP = 0.23686, df = 1, p-value = 0.6265
```

From the output, we can see that the **p-value is 0.625 >0.05**, thus we fail to reject the null hypothesis. **Meaning, the variance is constant**

Next, we will use the the Durbin-Watson test to detect serial autocorrelation in the residuals. Under the assumption of no autocorrelation, the D statistic is 2.

```
dwtest(RegModel.1)
```

```
##
##  Durbin-Watson test
##
## data:  RegModel.1
## DW = 1.4933, p-value = 0.01735
## alternative hypothesis: true autocorrelation is greater than 0
```

From the output, we the **p-value is 0.01735 <0.05**, so we reject the null hypothesis. This means that **there is a autocorrelation** in the residuals

Thirdly, we will use the the RESET test is a test of the assumption of linearity. If the linearity assumption is met, the RESET statistic will be close to 1.

```
resettest(RegModel.1)
```

```
##
##  RESET test
##
## data:  RegModel.1
## RESET = 7.689, df1 = 2, df2 = 54, p-value = 0.001153
```

From the output, we can see that the **RESET point = 7.689**, which is way bigger than 1 ,and the **p-value is 0.001153 <0.05**. Thus, we can reject the Null Hypothesis and conclude that there is no linear relationship.

Finally, we will use the Shapiro-Wilk normality test on the residual to confirm that the deviation from normality of the residuals is not large

```
shapiro.test(residuals(RegModel.1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(RegModel.1)
## W = 0.97111, p-value = 0.1807
```

From the output, we can see that the **p-value is 0.1807 > 0.05**. Thus, the values **do not deviate too much from Normal Distribution**

Overall, we can see that the data violate 2 assumptions: linearity and correlation bewteen residuals. Otherwise, the data pass all other assumptions mentioned above

4f) Whta is the statistical conclusion (inference - do you aceept the null ?)

We will show the output of the model

```
summary(RegModel.1)
```

```
##
## Call:
## lm(formula = ave ~ year, data = Yarmouth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3221 -0.4016 -0.1265  0.4505  1.4545
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.732995   9.466686   0.183    0.855
## year        0.002649   0.004785   0.554    0.582
##
## Residual standard error: 0.61 on 56 degrees of freedom
## Multiple R-squared:  0.005445,   Adjusted R-squared:  -0.01231
## F-statistic: 0.3066 on 1 and 56 DF,  p-value: 0.582
```

From the output, we can see that the **p-value is 0.582**. Thus, we fail to reject the null Hypothesis. This means that the slope is 0 ($\hat{\beta}$)=0). We can also interpret this as the independent variable **year** have no statistically significant relationship with the dependent variable **average temperature**

4g) Provide a "biological" conclusion concerning climate change in Yarmouth (i.e: qualified statemement about climate change at this site including an estimate of raw effect sie, its precision and whether a trend of this magnitude matters).

Calculation of effective size:

$$d = \frac{b}{sb\sqrt{n-k-1}} = \frac{0.002649}{0.004785\sqrt{58-1-1}} = 0.0739$$

So, **effect size is 0.00739**

Precision would be based on the standard error. We can also calculate the 95% confidence interval to show the precision.

My biological conclusion is that: The change in temperature in Yarmouth is not a serious issue.

For the precision, I can calculate the confidence interval of the slope

```
confint(RegModel.1)
```

```
##                   2.5 %      97.5 %
## (Intercept) -17.231054546 20.69704551
## year         -0.006935455  0.01223399
```

We can see that. Given the estimation of the slope, the confidence interval is significantly large. Thus, this analysis has a **low precision**.

The **magnitude of the trend** is small statistically. However, according to IPCC, only 6^oC increased would cause a dire consequence. Thus, it may still be worthwhile to put the little increasing tred into consideration.

4h) Provide a statement about your confidence in the statistical conclusion (for example with respect to the violation of certain assumptions, to potential biases created by missing data, to suboptimal experimental design, and/or to power issues).

The data that we are analyzing comes from the city Yarmouth.

```
head(Yarmouth,6)
```

```
## # A tibble: 6 x 4
##     reg loc    year   ave
##   <dbl> <chr> <dbl> <dbl>
## 1     0 Y      1950  7.03
## 2     0 Y      1951  8.32
## 3     0 Y      1952  7.64
## 4     0 Y      1953  8.32
## 5     0 Y      1954  7.32
## 6     0 Y      1955  6.85
```

- First, I want to address whether or not the data the violate assumptions for linear regression model. From several tests being done in part e, we can see that the data violate 2 assumptions (there no linear relationship between independenta variable, and there is an autocorrelation between the residuals). Thus, I am **not confident** that the data would give a good analysis.

- Secondly, I want to address bias due to missing data (NA values). For this data set Yarmouth, that is not an issue because we do not have NA values in the data set.
- Thirdly, I want to discuss the experimental design of the data. Since the data is collected from the goverment website, there is not enough sufficient information for me to have a conclusion on how the data should be collected.
- Fourth, I want to calculate the power of the hypotheis test on the estimation of climate change rate in Yarth mouth

We have: $n = 58, d = 0.0739, \alpha = 0.05$

```
pwr.t.test(n = 58,
           d = 0.0739,
           sig.level = 0.05,
           type = "one.sample")
```

```
##
##      One-sample t test power calculation
##
##               n = 58
##               d = 0.0739
##       sig.level = 0.05
##           power = 0.08575476
##     alternative = two.sided
```

From the output, we can see that the **power= 0.0857**, which is very low. From this, we know that there is a high chance we will have type II error in our analysis.

Overall, From addressing quality of data and power, I conclude that we **should not** make any conclusion from this analysis

5. Yarmouth is only one of the six locations for which she needs to quantify the extent of any climate change. Plot the temporal trend for temperature separately for each station (2 points), and estimate the rate of change for each station (2 points).

First, we will create multiple subsets, each associates to one city.

```
# Split climate in multiple dataframe, each associates with one city
cities <- climate %>%
  group_split(loc)

# Let's examine the structure of variable cities
str(cities)
```

```
## list<tibble[,4]> [1:6]
## $ : tibble [58 x 4] (S3: tbl_df/tbl/data.frame)
##  ..$ reg : num [1:58] 1 1 1 1 1 1 1 1 1 1 ...
##  ..$ loc : chr [1:58] "A" "A" "A" "A" ...
##  ..$ year: num [1:58] 1950 1951 1952 1953 1954 ...
##  ..$ ave : num [1:58] NA -18.6 -17.6 -16.9 -17.9 ...
## $ : tibble [58 x 4] (S3: tbl_df/tbl/data.frame)
##  ..$ reg : num [1:58] 1 1 1 1 1 1 1 1 1 1 ...
```

```
##   ..$ loc : chr [1:58] "M" "M" "M" "M" ...
##   ..$ year: num [1:58] 1950 1951 1952 1953 1954 ...
##   ..$ ave : num [1:58] -18.5 -17 -16.9 -17.4 -17.7 ...
## $ : tibble [58 x 4] (S3: tbl_df/tbl/data.frame)
##   ..$ reg : num [1:58] 1 1 1 1 1 1 1 1 1 1 ...
##   ..$ loc : chr [1:58] "S" "S" "S" "S" ...
##   ..$ year: num [1:58] 1950 1951 1952 1953 1954 ...
##   ..$ ave : num [1:58] NA NA NA NA NA ...
## $ : tibble [58 x 4] (S3: tbl_df/tbl/data.frame)
##   ..$ reg : num [1:58] 0 0 0 0 0 0 0 0 0 0 ...
##   ..$ loc : chr [1:58] "V" "V" "V" "V" ...
##   ..$ year: num [1:58] 1950 1951 1952 1953 1954 ...
##   ..$ ave : num [1:58] 9.1 9.74 9.94 10.6 9.69 ...
## $ : tibble [58 x 4] (S3: tbl_df/tbl/data.frame)
##   ..$ reg : num [1:58] 0 0 0 0 0 0 0 0 0 0 ...
##   ..$ loc : chr [1:58] "W" "W" "W" "W" ...
##   ..$ year: num [1:58] 1950 1951 1952 1953 1954 ...
##   ..$ ave : num [1:58] 8.62 8.94 10.09 10.38 9.69 ...
## $ : tibble [58 x 4] (S3: tbl_df/tbl/data.frame)
##   ..$ reg : num [1:58] 0 0 0 0 0 0 0 0 0 0 ...
##   ..$ loc : chr [1:58] "Y" "Y" "Y" "Y" ...
##   ..$ year: num [1:58] 1950 1951 1952 1953 1954 ...
##   ..$ ave : num [1:58] 7.03 8.32 7.64 8.32 7.32 ...
## @ ptype: tibble [0 x 4] (S3: tbl_df/tbl/data.frame)
##   ..$ reg : num(0)
##   ..$ loc : chr(0)
##   ..$ year: num(0)
##   ..$ ave : num(0)
```

```
# name each elment of "cities" with the correct name
Alert <- cities[[1]]
Mould_Bay <- cities[[2]]
Sachs_Habor <- cities[[3]]
Victoria <- cities[[4]]
Windsor <- cities[[5]]
Yarmouth <- cities[[6]]
```

Second, we will draw multiple temporal-trend plots. One for each city. To do this, I will write a function that input **city** and output the **scatter plot**. This function would be based heavily on the lab manual provided by professor Julien Martin

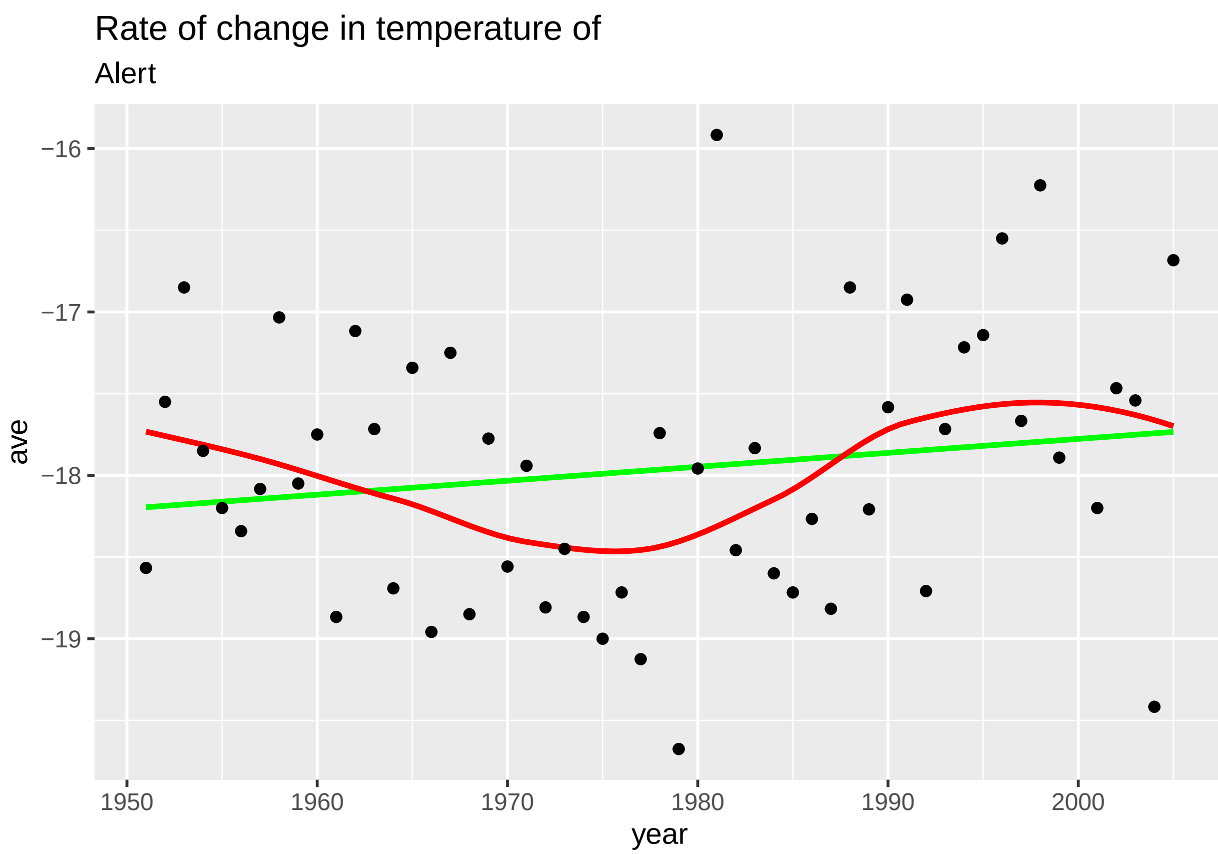
```
plotting_temporal_trend <- function (city){

  #city: Yarmouth, Windsord,...as a data frame

mygraph <- ggplot(
  data = city[!is.na(city$ave), ], # source of data
  aes(x = year, y = ave)
)

# plot data points, regression, loess trace
mygraph <- mygraph +
  stat_smooth(method = lm, se = FALSE, color = "green") + # add linear regression, but no SE shading
  stat_smooth(color = "red", se = FALSE) + # add loess
```

```
  geom_point()+
  ggtitle( "Rate of change in temperature of", deparse(substitute(city)))         # add data points

mygraph
}
```

Next, I will draw all 6 temporal trend graphs for all 6 cities

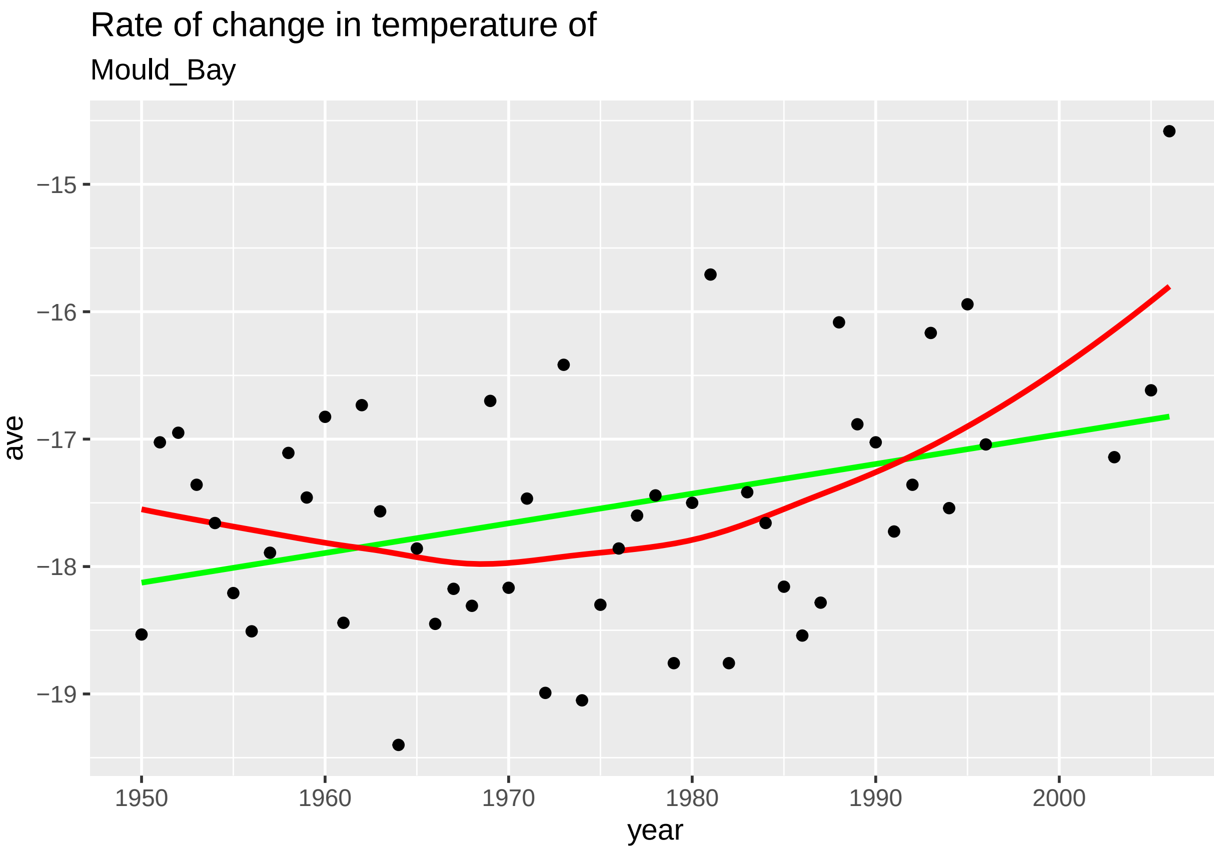
```
plotting_temporal_trend(Alert)
```

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

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



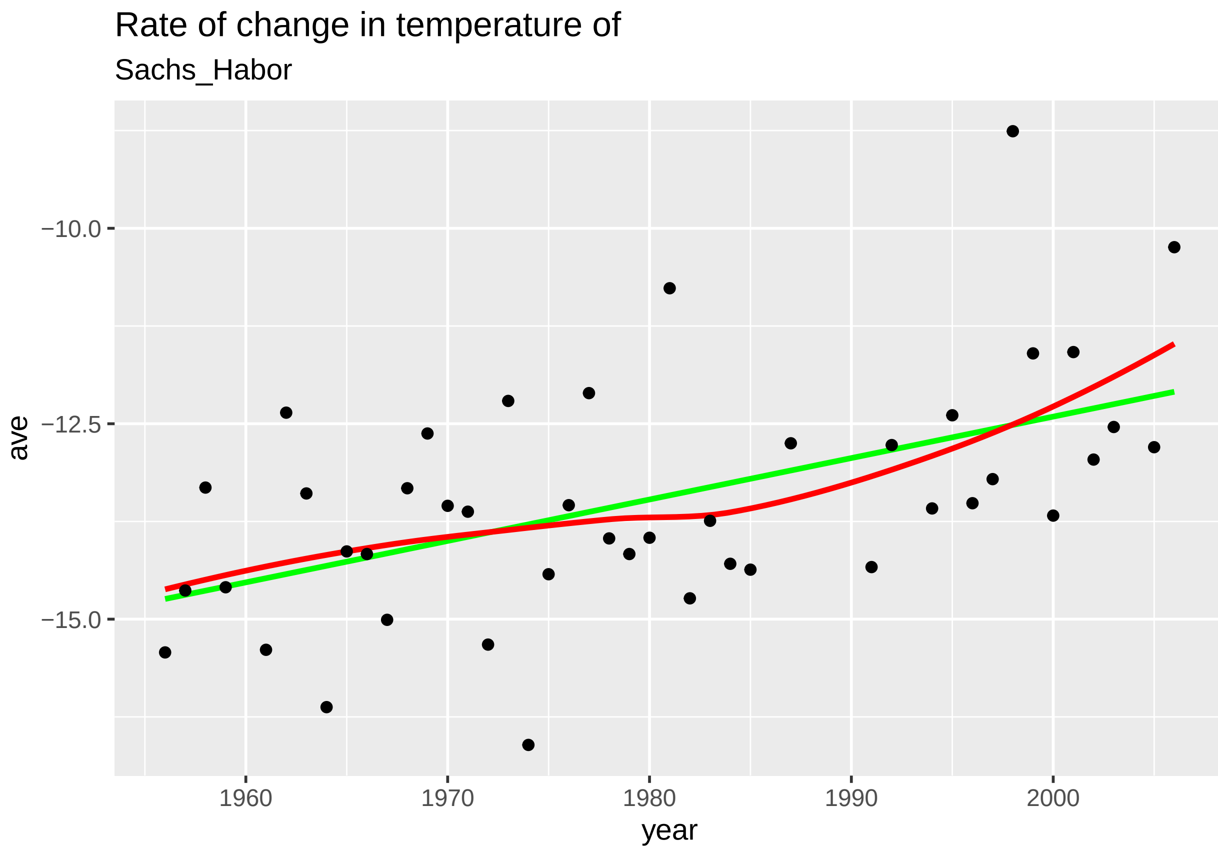

```
plotting_temporal_trend(Mould_Bay)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



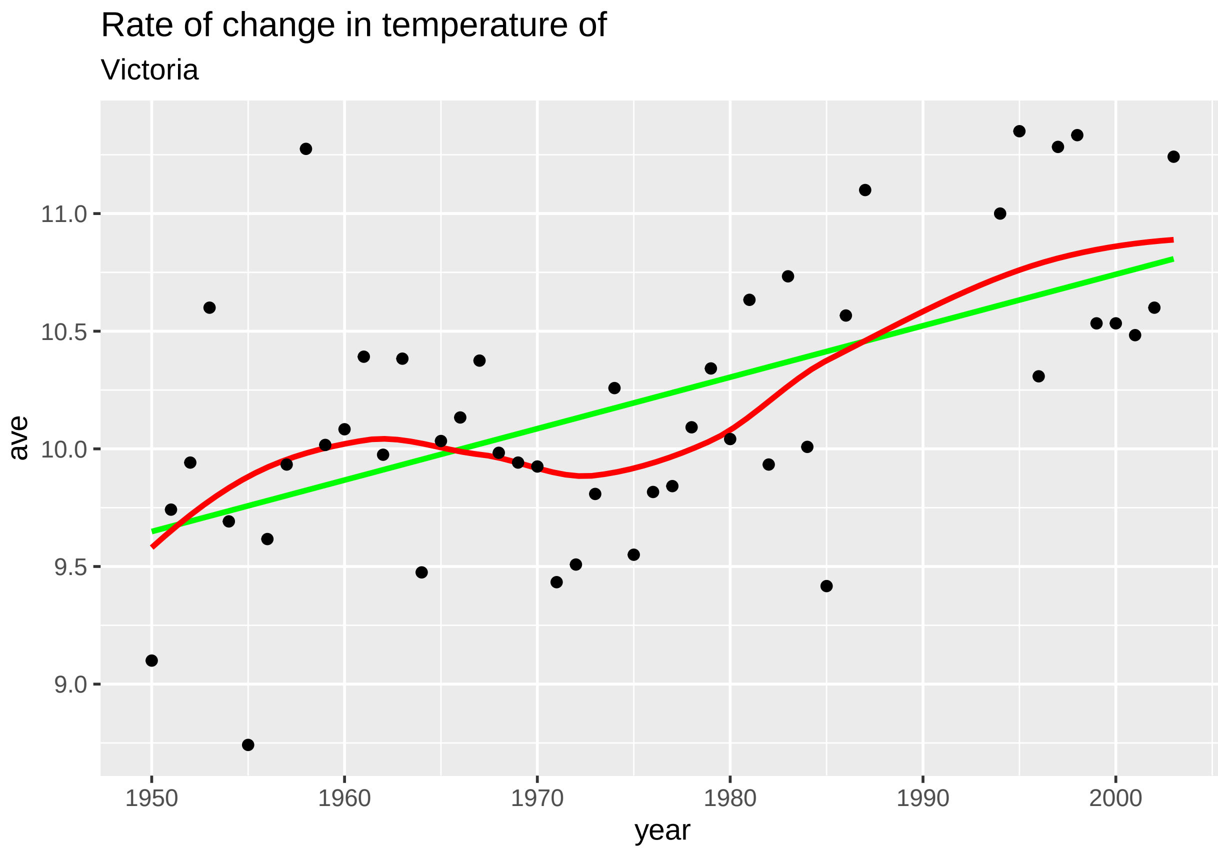

```
plotting_temporal_trend(Sachs_Habor)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



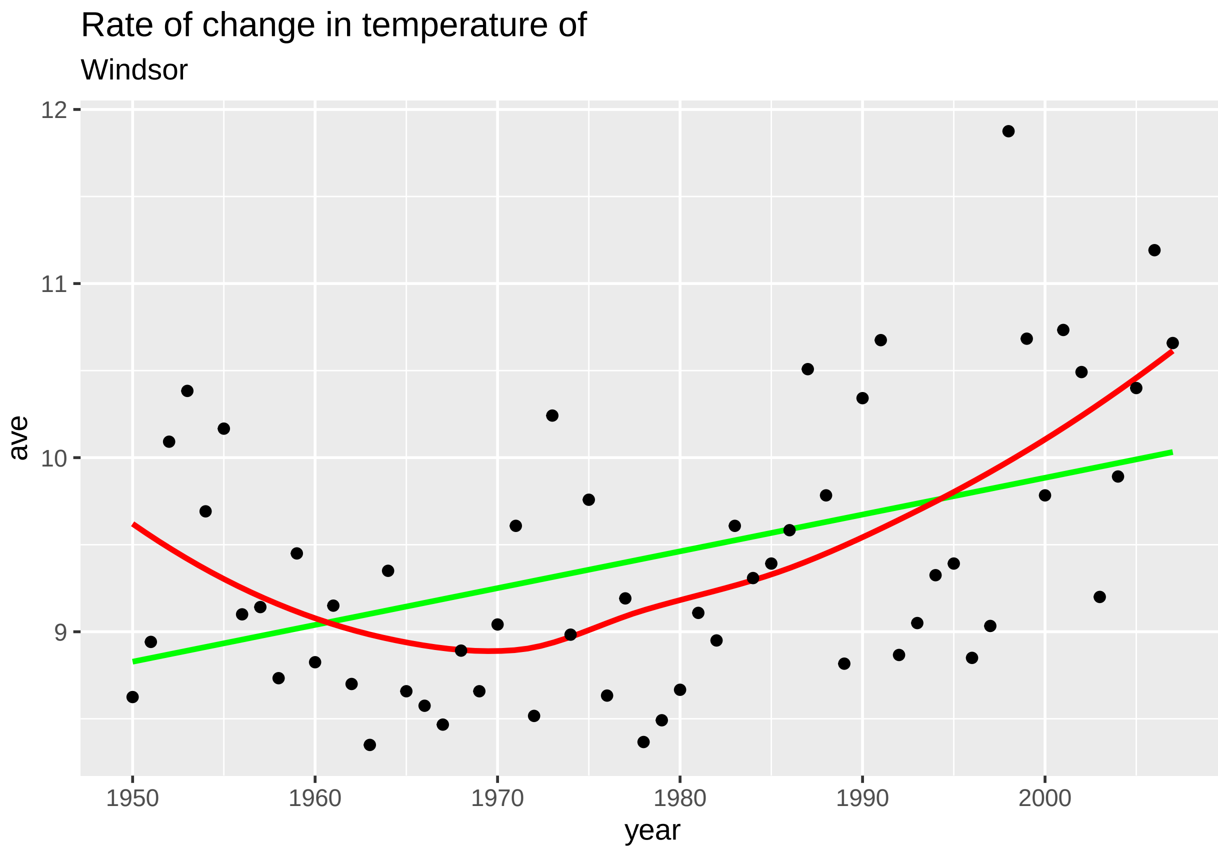

```
plotting_temporal_trend(Victoria)
```

```
## 'geom_smooth()' using formula 'y ~ x'
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```



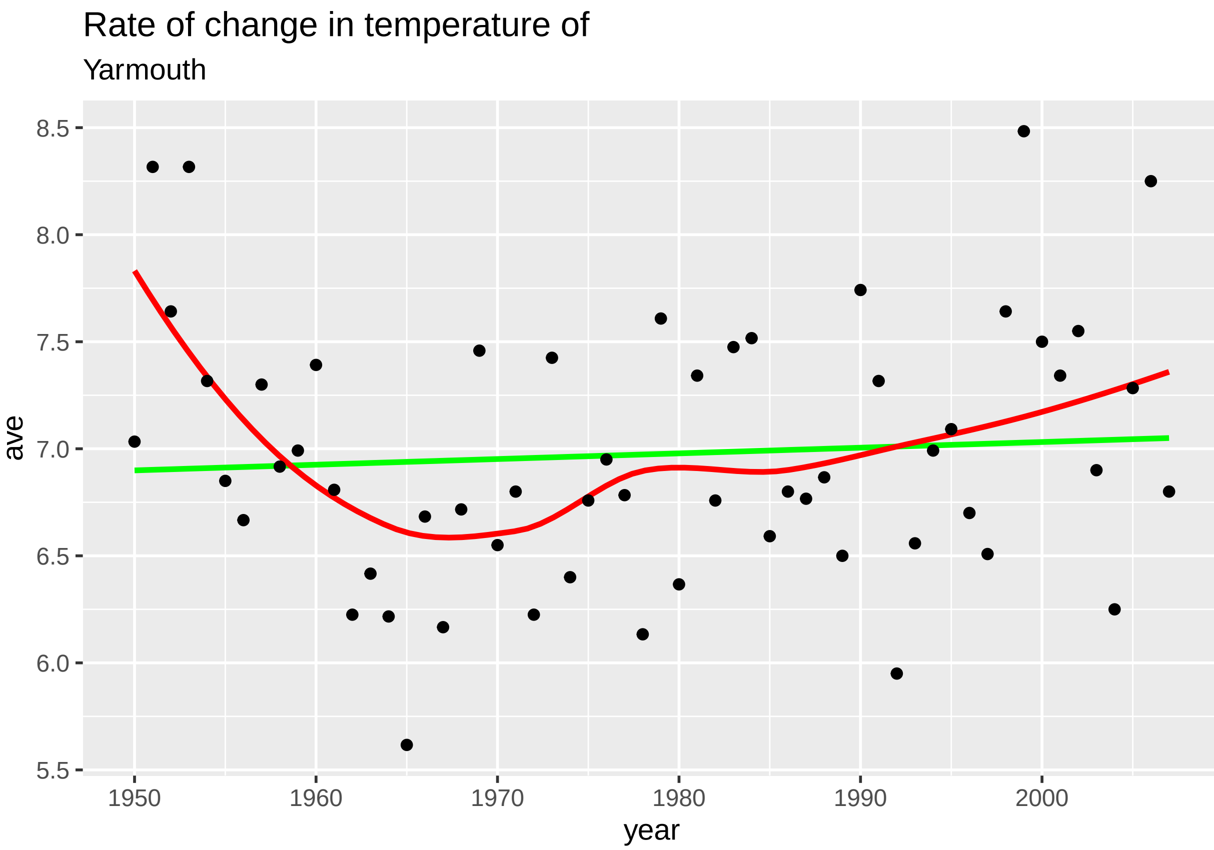

```
plotting_temporal_trend(Windsor)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```




```
plotting_temporal_trend(Yarmouth)
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



After the temporal trends, we want to estimate the rate of change (or the slope) in temperature for each city.

I will summarize all linear regression result in one table, so it is easy to compare

```
fitted_models = climate %>%
                group_by(loc) %>%
                do(fit_ave = lm(ave ~year, data = .)) %>%
                ungroup %>%
                mutate(yearCoef = map(fit_ave, tidy))%>%
                unnest(yearCoef)

fitted_models
```

```
## # A tibble: 12 x 7
##    loc   fit_ave term        estimate std.error statistic   p.value
##    <chr> <list>  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
##  1 A     <lm>    (Intercept)  -34.8     14.0        -2.49 0.0160
##  2 A     <lm>    year           0.00854  0.00708     1.21 0.233
##  3 M     <lm>    (Intercept)  -63.5     16.3        -3.90 0.000301
##  4 M     <lm>    year           0.0233   0.00825     2.82 0.00695
##  5 S     <lm>    (Intercept) -118.      25.7        -4.60 0.0000386
##  6 S     <lm>    year           0.0530   0.0130      4.08 0.000199
##  7 V     <lm>    (Intercept)  -33.0      8.90       -3.71 0.000563
##  8 V     <lm>    year           0.0219   0.00451     4.85 0.0000145
##  9 W     <lm>    (Intercept)  -32.4     11.1        -2.92 0.00502
## 10 W     <lm>    year           0.0211   0.00560     3.77 0.000392
```

```
## 11 Y     <lm>    (Intercept)    1.73      9.47        0.183 0.855
## 12 Y     <lm>    year           0.00265   0.00478     0.554 0.582
```

From the table, we can see the *slope* of each city is the **estimation**. I will display the result in a simpler table down below

```
rate_of_change_table <-fitted_models %>%
                       filter(term == "year")%>%
                       select(loc, estimate) %>%
                       transmute(city= loc,  rate_of_change=estimate)

rate_of_change_table
```

```
## # A tibble: 6 x 2
##   city  rate_of_change
##   <chr>          <dbl>
## 1 A            0.00854
## 2 M            0.0233
## 3 S            0.0530
## 4 V            0.0219
## 5 W            0.0211
## 6 Y            0.00265
```

6. Compute the mean warming rate in the North and in the South by averaging the rates calculated separately for each station (1 point). Calculate a 95% CI for each of these means based on the 3 replicates stations. (2 points)

From queston 5, we have this data:

```
rate_of_change_table <- rate_of_change_table %>% mutate(region= c("North","North","North","South","Sout
                     select(region, city, rate_of_change)

rate_of_change_table
```

```
## # A tibble: 6 x 3
##   region city  rate_of_change
##   <chr>  <chr>          <dbl>
## 1 North  A            0.00854
## 2 North  M            0.0233
## 3 North  S            0.0530
## 4 South  V            0.0219
## 5 South  W            0.0211
## 6 South  Y            0.00265
```

We then calculate the mean warming rate and standard deviation of warming rate for the North and the South

```
rate_of_change_table %>% group_by(region) %>%
  summarise_at(vars(rate_of_change),list(region_mean = mean, region_sd= sd))
```

```
## # A tibble: 2 x 3
##   region region_mean region_sd
##   <chr>        <dbl>     <dbl>
## 1 North       0.0283    0.0227
## 2 South       0.0152    0.0109
```

To calculate the 95% Confidence interval for each mean, we would use the t-test, with the degree of freedom of n-1. We the sample size n=3, so n-1 = 2. We will use `t.test()` to solve the confidence interval

So we have:

$$n = 3, \alpha = 0,05$$

The formula to calculate confidence interval is:

$$\bar{u} \pm t_{\alpha/2,n-1} * se(\bar{u})$$

Here is the confidence interval for the mean warming rate in the **North**

```
north_result <- as.vector((rate_of_change_table %>% filter(region =="North"))$rate_of_change)

t.test(north_result)
```

```
##
##  One Sample t-test
##
## data:  north_result
## t = 2.1621, df = 2, p-value = 0.1631
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.02799273  0.08454226
## sample estimates:
##  mean of x
## 0.02827476
```

It is [**-0.02799273 0.08454226**]

Here is the confidence interval for the mean warming rate in the **South**

```
south_result <- as.vector((rate_of_change_table %>% filter(region =="South"))$rate_of_change)

t.test(south_result)
```

```
##
##  One Sample t-test
##
## data:  south_result
## t = 2.4203, df = 2, p-value = 0.1366
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.01183302  0.04226223
## sample estimates:
##  mean of x
## 0.01521461
```

It is: [**-0.01183302 , 0.04226223**]