CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Showcasing evidence of climate change - shinyApp

AUTHOR: Korede Adegboye

SUPERVISOR: Dave Campbell

DATE: May 7, 2020

Table of Contents

1 Data Documentation

1.1 Temperature Data

Canada, Department of Environment and Climate Change temperature data *[1]* is listed as *Second Generation of Homogenized Temperatures for Canada*. There are minimum, maximum and mean temperatures, each containing data for 338 locations in the form of .txt files. Of concern, is the file format. The first two lines include important station characteristics. Specifically, those characteristics are listed to be station number, city and province. Further, the features of the file are given. To summarize the features, monthly temperatures are given for the respective years. The months include January to December. Annual and seasonal temperatures are also included. Years may range from 1880 to 2017. In other words, sone locations have data as early as 1880, and some have been updated to December 2017. Therefore, there is still wide a range of years, across all locations.

It is noted that, some monthly temperature values are missing. Leaving possible reason for concern in performing analysis on the data. These missing values, otherwise known as default values, are indicated to be "-9999.9". In addition, non-integer values are included. Namely, data flags. Missing data has been labelled "M", and estimated data labelled by "E". These are indicated after the monthly value. Finally, an "a" flag represents an adjusted value. Clearly, it will need to be decided, whether the flags are of importance to our analysis. If, no flag is given then this suggests that the original data is used.

As for readability, the data is comma-delimited and convenient to follow. Structured, in a table like setting. The unit for temperature is given by "°C", remaining consistent with Canadas standard unit. Thus, no conversions will need to be made. In addition, the minimum, maximum and mean temperatures are given by average daily temperatures throughout the month.

To conclude, the data files are complete and other sources of data do not need to be used.

2 Objective

The objective of this project is to illustrate evidence of climate change for a specified region. To start, cleaning the data is of much significance. The use of Jupyter Notebook will give insight into the data that is being dealt with. Mainly, one can see the decisions made and the reason for those decisions. In all, making use of the R coding language, and its available packages.

Next, it is intended to perform statistical methods. Given the temperature data, multilinear regression will be the statistical method used. Before performing the regression, an input data frame consisting of temperature, year, city and province will need to be created. Furthermore, retaining an output data frame showcasing statistical results. Briefly, noting the results and checking for any unusual results.

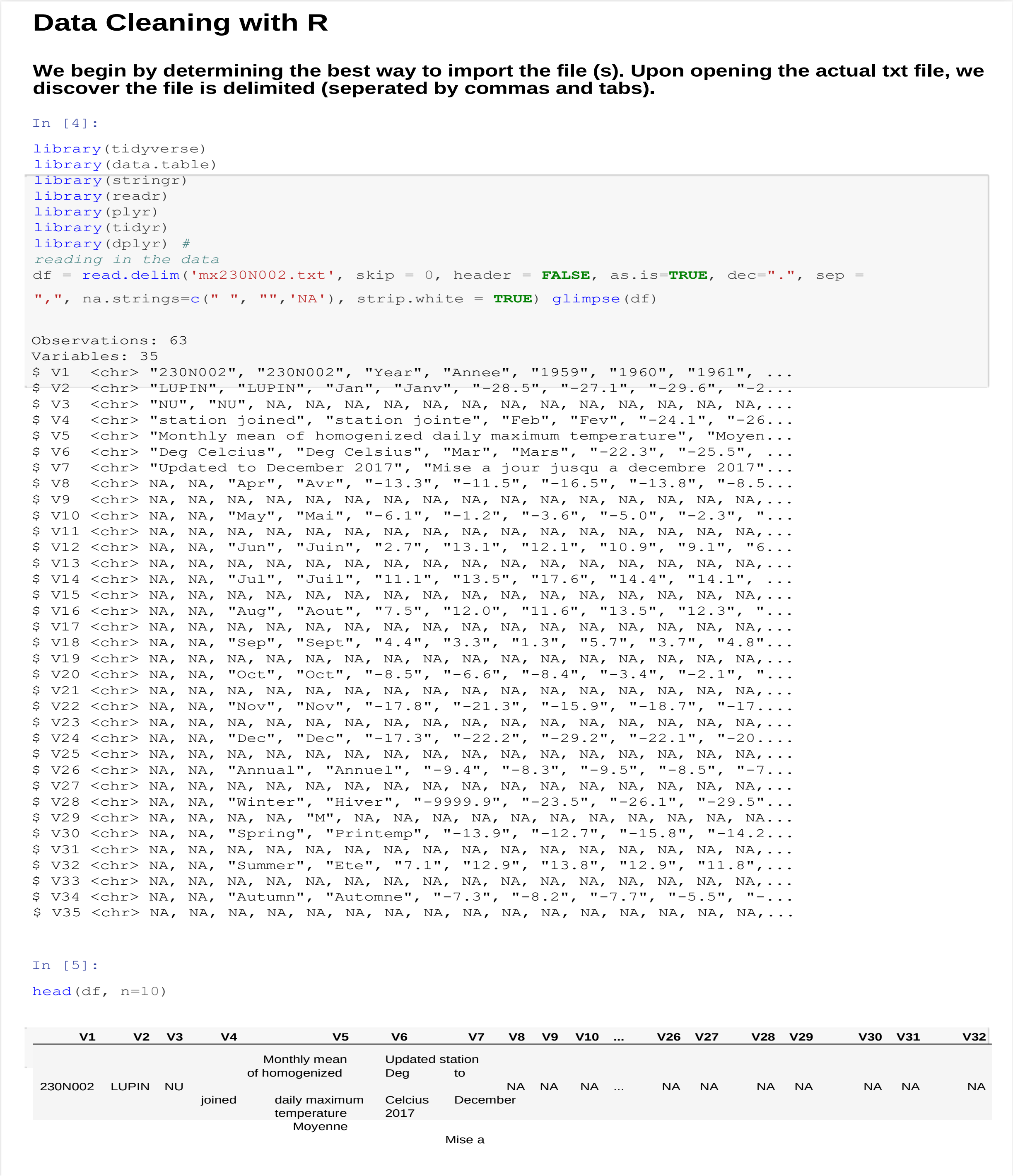
Following, we consider the development of a shiny application. A family-friendly interface allowing users to understand the evidence of climate change for a specified region. Regions, being city, province or nation. Software engineering demonstrates function reusability, interactions between server, UI and shiny modules. Hence, developing a back end that is easy to follow, and accounts for complex development of the application.

Climate change concerns vary for people across Canada. Whether it is of business, political or personal interest. Moving forward, alongside region there are other variables to consider. These variables include, minimum, mean, maximum temperatures, month, year to start, plot type and statistical result. Therefore, these results will be displayed by plots. Showing the user, what is going on and the current variables (features) of the plot is of much significance.

Now that exploratory data analysis can be conducted, further insights can be deduced. Finding that some results are like those found by Environment Canada or other credible data analysts. Further, the application intends to focus on displaying evidence of climate change. Guiding users to a trustworthy comprehension of the concerns they want answered.

3 Data Cleaning

The following demonstration was done with Jupyter Notebook. It uses a R language package found in Anaconda Navigator. The data cleaning notebook is available at *https://github.com/korede97/shinyClimate/blob/master/data_cleaning.ipynb.*



Data Cleaning with R

We begin by determining the best way to import the file (s). Upon opening the actual txt file, we discover the file is delimited (seperated by commas and tabs).

In [4]:

```
library(tidyverse)
library(data.table)
library(stringr)
library(readr)
library(plyr)
library(tidyr)
library(dplyr) #
reading in the data
df = read.delim('mx230N002.txt', skip = 0, header = FALSE, as.is=TRUE, dec=".", sep =
",", na.strings=c(" ", "",'NA'), strip.white = TRUE) glimpse(df)
```

```
Observations: 63
Variables: 35
$ V1  <chr> "230N002", "230N002", "Year", "Annee", "1959", "1960", "1961", ...
$ V2  <chr> "LUPIN", "LUPIN", "Jan", "Janv", "-28.5", "-27.1", "-29.6", "-2...
$ V3  <chr> "NU", "NU", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V4  <chr> "station joined", "station jointe", "Feb", "Fev", "-24.1", "-26...
$ V5  <chr> "Monthly mean of homogenized daily maximum temperature", "Moyen...
$ V6  <chr> "Deg Celcius", "Deg Celsius", "Mar", "Mars", "-22.3", "-25.5", ...
$ V7  <chr> "Updated to December 2017", "Mise a jour jusqu a decembre 2017"...
$ V8  <chr> NA, NA, "Apr", "Avr", "-13.3", "-11.5", "-16.5", "-13.8", "-8.5...
$ V9  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V10 <chr> NA, NA, "May", "Mai", "-6.1", "-1.2", "-3.6", "-5.0", "-2.3", "...
$ V11 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V12 <chr> NA, NA, "Jun", "Juin", "2.7", "13.1", "12.1", "10.9", "9.1", "6...
$ V13 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V14 <chr> NA, NA, "Jul", "Juil", "11.1", "13.5", "17.6", "14.4", "14.1", ...
$ V15 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V16 <chr> NA, NA, "Aug", "Aout", "7.5", "12.0", "11.6", "13.5", "12.3", "...
$ V17 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V18 <chr> NA, NA, "Sep", "Sept", "4.4", "3.3", "1.3", "5.7", "3.7", "4.8"...
$ V19 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V20 <chr> NA, NA, "Oct", "Oct", "-8.5", "-6.6", "-8.4", "-3.4", "-2.1", "...
$ V21 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V22 <chr> NA, NA, "Nov", "Nov", "-17.8", "-21.3", "-15.9", "-18.7", "-17....
$ V23 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V24 <chr> NA, NA, "Dec", "Dec", "-17.3", "-22.2", "-29.2", "-22.1", "-20....
$ V25 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V26 <chr> NA, NA, "Annual", "Annuel", "-9.4", "-8.3", "-9.5", "-8.5", "-7...
$ V27 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V28 <chr> NA, NA, "Winter", "Hiver", "-9999.9", "-23.5", "-26.1", "-29.5"...
$ V29 <chr> NA, NA, NA, NA, "M", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
$ V30 <chr> NA, NA, "Spring", "Printemp", "-13.9", "-12.7", "-15.8", "-14.2...
$ V31 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V32 <chr> NA, NA, "Summer", "Ete", "7.1", "12.9", "13.8", "12.9", "11.8",...
$ V33 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
$ V34 <chr> NA, NA, "Autumn", "Automne", "-7.3", "-8.2", "-7.7", "-5.5", "-...
$ V35 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

In [5]:

```
head(df, n=10)
```

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V26	V27	V28	V29	V30	V31	V32
230N002	LUPIN	NU	station joined	Monthly mean of homogenized daily maximum temperature	Deg Celcius	Updated to December 2017	NA	NA	NA	...	NA	NA	NA	NA	NA	NA	NA
				Moyenne		Mise a											

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	...	V26	V27	V28	V29	V30	V31	V32
230N002	LUPIN	NU	station jointe	mensuelle des temperatures quotidiennes maximales homogeneisees	Deg Celsius	Mise a jour jusqu a decembre 2017	NA	NA	NA	...	NA	NA	NA	NA	NA	NA	NA
Year	Jan	NA	Feb	NA	Mar	NA	Apr	NA	May	...	Annual	NA	Winter	NA	Spring	NA	Summer
Annee	Janv	NA	Fev	NA	Mars	NA	Avr	NA	Mai	...	Annuel	NA	Hiver	NA	Printemp	NA	Ete
1959	-28.5	NA	-24.1	NA	-22.3	NA	-13.3	NA	-6.1	...	-9.4	NA	-9999.9	M	-13.9	NA	7.1
1960	-27.1	NA	-26.2	NA	-25.5	NA	-11.5	NA	-1.2	...	-8.3	NA	-23.5	NA	-12.7	NA	12.9
1961	-29.6	NA	-26.4	NA	-27.3	NA	-16.5	NA	-3.6	...	-9.5	NA	-26.1	NA	-15.8	NA	13.8
1962	-26.5	NA	-32.8	NA	-23.8	NA	-13.8	NA	-5.0	...	-8.5	NA	-29.5	NA	-14.2	NA	12.9
1963	-28.8	NA	-26.0	NA	-26.5	NA	-8.5	NA	-2.3	...	-7.7	NA	-25.6	NA	-12.4	NA	11.8
1964	-31.1	NA	-26.9	NA	-28.3	NA	-15.8	NA	-0.9	...	-8.8	NA	-26.1	NA	-15.0	NA	12.4

The tidyverse package will become evidently useful as we proceed

- glimpse (helps) provides a summary of the data
- There are 63 observations (rows) and 35 variables (columns)
- all data is stored as characters
- After looking at the first few rows 10 rows of our current frame. We see,
- There is some information we want to remove. Specifically, the first 2 rows of the data frame
- Almost every other column is filled with NAs
- The 3rd row will later be used as our column names, and V1 as our row names
- Our data mainly deals with positive, and negative values.
- Unusual data includes: default: -9999.9 and the letters 'E' and 'M' used to mark if the data was estimated or missing

Constructing desired Dataframe

Earlier it has been noticed that the majority NAs appear in a patterned fashion. Since, its clear I went ahead and removed those columns from our data frame. Note also, letters, 'M' and 'E" will be removed.

Extract the header and combine it back to the data.

It is important to check for whitespaces when it comes to extracting names.

In [7]:

```
df = read.delim('mx230N002.txt', skip = 0, header = FALSE, as.is=TRUE, dec=".", sep = ",",
na.strings=c(" ", "",'NA'), strip.white = TRUE)

seq(from = 3, to = 35, by = 2)

df <- select(df, -seq(from = 3, to = 35, by = 2))

data <- slice(df, 5:n())
# reset df to show this step...
(hdr <- slice(df, 3))
# check for whitespaces in the column names

# unlist((hdr))
# (c  = select(hdr, contains(" ")))
# (st  = select(hdr, starts_with(" ")))
# (ew = select(hdr, ends_with(" ")))

is.na(hdr)

head ((df <- rename(data, hdr)), n=5)
```

```
3  5  7  9  11  13  15  17  19  21  23  25  27  29  31  33  35
```

V1	V2	V4	V6	V8	V10	V12	V14	V16	V18	V20	V22	V24	V26	V28	V30	V32	V34

V1	V2	V4	V6	V8	V10	V12	V14	V16	V18	V20	V22	V24	V26	V28	V30	V32	V34
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Winter	Spring	Summer	Autumn

V1	V2	V4	V6	V8	V10	V12	V14	V16	V18	V20	V22	V24	V26	V28	V30	V:
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALS

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Winter	Spring	Summer	Autumn
1959	-28.5	-24.1	-22.3	-13.3	-6.1	2.7	11.1	7.5	4.4	-8.5	-17.8	-17.3	-9.4	-9999.9	-13.9	7.1	-7.3
1960	-27.1	-26.2	-25.5	-11.5	-1.2	13.1	13.5	12.0	3.3	-6.6	-21.3	-22.2	-8.3	-23.5	-12.7	12.9	-8.2
1961	-29.6	-26.4	-27.3	-16.5	-3.6	12.1	17.6	11.6	1.3	-8.4	-15.9	-29.2	-9.5	-26.1	-15.8	13.8	-7.7
1962	-26.5	-32.8	-23.8	-13.8	-5.0	10.9	14.4	13.5	5.7	-3.4	-18.7	-22.1	-8.5	-29.5	-14.2	12.9	-5.5
1963	-28.8	-26.0	-26.5	-8.5	-2.3	9.1	14.1	12.3	3.7	-2.1	-17.3	-20.3	-7.7	-25.6	-12.4	11.8	-5.2

Standard default Values...

In this case we have -9999.9

In some situations, it has been suggested to replace the NAs or default values with the mean of the column. Chosen to go against this, since regression regression smoothly 'fill' in these blanks.

In [9]:

```
# head(df, n=10)
df <- data.frame(lapply(df, function(x){
      gsub("-9999.9", "NA", x)
    }))
df
```

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Winter	Spring	Summer	Autumn
1959	-28.5	-24.1	-22.3	-13.3	-6.1	2.7	11.1	7.5	4.4	-8.5	-17.8	-17.3	-9.4	NA	-13.9	7.1	-7.3
1960	-27.1	-26.2	-25.5	-11.5	-1.2	13.1	13.5	12.0	3.3	-6.6	-21.3	-22.2	-8.3	-23.5	-12.7	12.9	-8.2
1961	-29.6	-26.4	-27.3	-16.5	-3.6	12.1	17.6	11.6	1.3	-8.4	-15.9	-29.2	-9.5	-26.1	-15.8	13.8	-7.7
1962	-26.5	-32.8	-23.8	-13.8	-5.0	10.9	14.4	13.5	5.7	-3.4	-18.7	-22.1	-8.5	-29.5	-14.2	12.9	-5.5
1963	-28.8	-26.0	-26.5	-8.5	-2.3	9.1	14.1	12.3	3.7	-2.1	-17.3	-20.3	-7.7	-25.6	-12.4	11.8	-5.2
1964	-31.1	-26.9	-28.3	-15.8	-0.9	6.5	16.4	14.3	4.8	-4.0	-14.5	-25.7	-8.8	-26.1	-15.0	12.4	-4.6
1965	-27.6	-35.7	-19.1	-10.2	-0.7	6.5	15.5	13.2	1.5	-5.0	-16.0	-24.0	-8.5	-29.7	-10.0	11.7	-6.5
1966	-35.9	-27.1	-21.0	-14.7	-0.2	13.5	16.4	14.8	7.1	-8.9	-22.2	-21.6	-8.3	-29.0	-12.0	14.9	-8.0
1967	-28.0	-31.0	-23.9	-12.8	-2.8	5.8	13.7	12.6	5.9	-2.9	-13.7	-20.8	-8.2	-26.9	-13.2	10.7	-3.6
1968	-30.1	-23.4	-17.0	-14.0	-2.1	8.9	11.7	11.2	5.3	-2.7	-13.8	-24.2	-7.5	-24.8	-11.0	10.6	-3.7
1969	-30.5	-20.8	-20.8	-9.0	-3.6	4.1	15.3	13.0	4.6	-1.8	-17.1	-16.4	-6.9	-25.2	-11.1	10.8	-4.8
1970	-27.1	-28.0	-21.0	-10.3	-2.8	10.9	16.4	12.7	4.3	-3.8	-18.0	-30.2	-8.1	-23.8	-11.4	13.3	-5.8
1971	-28.6	-25.4	-20.8	-11.0	1.3	11.6	14.8	13.1	5.8	-2.9	-18.9	-25.9	-7.2	-28.1	-10.2	13.2	-5.3
1972	-31.5	-31.7	-19.7	-15.5	-1.6	8.7	14.1	13.0	0.3	-7.8	-17.2	-25.8	-9.6	-29.7	-12.3	11.9	-8.2
1973	-24.3	-26.6	-21.7	-11.4	5.2	15.6	18.2	14.1	9.2	-0.7	-16.9	-23.7	-5.3	-25.6	-9.3	16.0	-2.8
1974	-28.6	-27.7	-25.5	-13.3	0.0	12.2	15.9	11.5	2.1	-10.8	-14.2	-22.5	-8.4	-26.7	-12.9	13.2	-7.6
1975	-33.2	-25.8	-23.4	-7.2	1.1	12.9	15.6	15.6	7.4	-4.9	-17.2	-25.5	-7.1	-27.2	-9.8	14.7	-4.9
1976	-27.4	-27.4	-24.2	-4.7	2.1	9.7	17.3	12.7	6.4	-4.0	-14.6	-26.8	-6.7	-26.8	-8.9	13.2	-4.1
1977	-23.9	-21.1	-21.6	-9.9	0.3	11.2	14.1	10.0	7.4	-3.3	-16.4	-26.9	-6.7	-23.9	-10.4	11.8	-4.1
1978	-26.2	-20.7	-22.9	-13.3	-4.2	4.9	9.3	10.2	4.2	-10.8	-18.3	-25.1	-9.4	-24.6	-13.5	8.1	-8.3
1979	-23.5	-36.3	-26.7	-14.0	-2.6	5.6	15.4	9.9	4.1	-5.8	-11.7	-21.0	-8.9	-28.3	-14.4	10.3	-4.5
1980	-27.9	-20.2	-22.0	-9.5	2.4	9.3	13.6	14.3	3.0	-4.0	-18.3	-28.4	-7.3	-23.0	-9.7	12.4	-6.4
1981	-16.4	-23.4	-16.4	-14.4	0.6	9.4	14.0	15.0	3.7	-3.9	-14.5	-20.7	-5.6	-22.7	-10.1	12.8	-4.9
1982	-32.9	-24.7	-23.7	-14.7	-1.8	9.3	16.3	11.3	3.7	-4.3	-22.3	-26.5	-9.2	-26.1	-13.4	12.3	-7.6
1983	-27.5	-29.8	-24.7	-12.3	-9.0	9.9	15.0	13.5	5.1	-6.3	-10.7	-23.5	-8.4	-27.9	-15.3	12.8	-4.0

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Winter	Spring	Summer	Autumn
1984	-28.6	-22.0	-31.6	-8.3	0.5	12.5	17.2	12.8	2.4	-7.0	-18.3	-29.1	-7.5	-24.7	-9.8	14.2	-7.6
1985	-24.0	-31.6	-22.5	-15.6	1.5	11.1	11.7	10.1	3.7	-7.2	-19.5	-20.5	-8.6	-28.2	-12.2	11.0	-7.7
1986	-27.7	-22.9	-23.5	-14.3	-2.0	7.1	15.4	10.3	4.5	-6.9	-20.6	-20.3	-8.4	-23.7	-13.3	10.9	-7.7
1987	-20.1	-21.1	-22.0	-10.1	-2.7	9.9	14.8	8.5	7.1	-5.3	-17.0	-16.0	-6.2	-20.5	-11.6	11.1	-5.1
1988	-28.2	-28.2	-18.8	-11.3	-6.2	11.4	16.3	14.5	6.2	-5.8	-21.4	-23.9	-8.0	-24.1	-12.1	14.1	-7.0
1989	-29.9	-18.3	-25.7	-10.9	-4.1	13.1	17.9	17.6	4.4	-4.8	-22.8	-24.8	-7.4	-24.0	-13.6	16.2	-7.7
1990	-28.9	-31.0	-17.1	-10.4	-3.5	8.3	15.5	9.4	3.8	-7.2	-21.8	-27.5	-9.2	-28.2	-10.3	11.1	-8.4
1991	-28.3	-26.5	-24.1	-9.7	2.0	12.3	16.5	14.1	2.0	-8.3	-19.7	-24.0	-7.8	-27.4	-10.6	14.3	-8.7
1992	-25.3	-25.5	-19.1	-13.1	-2.3	8.5	16.4	12.2	1.1	-5.8	-14.9	-23.0	-7.6	-24.9	-11.5	12.4	-6.5
1993	-21.7	-24.8	-17.4	-12.9	-2.6	8.4	14.0	13.0	3.4	-6.2	-17.8	-25.7	-7.5	-23.2	-11.0	11.8	-6.9
1994	-30.8	-30.5	-17.8	-13.2	1.9	13.1	19.6	15.0	5.2	-3.2	-15.3	-20.0	-6.3	-29.0	-9.7	15.9	-4.4
1995	-21.6	-24.9	-21.3	-11.0	-3.8	12.3	13.2	12.4	3.4	-6.6	-17.4	-26.4	-7.6	-22.2	-12.0	12.6	-6.9
1996	-28.4	-23.2	-22.0	-12.0	-2.5	15.1	19.2	10.5	7.5	-7.1	-16.9	-23.7	-7.0	-26.0	-12.2	14.9	-5.5
1997	-26.5	-23.6	-23.9	-11.5	-3.1	9.9	18.5	13.8	7.2	-8.7	-11.9	-19.5	-6.6	-24.6	-12.8	14.1	-4.5
1998	-30.5	-21.7	-18.2	-6.2	3.5	12.7	18.2	14.9	6.0	-1.8	-10.1	-19.1	-4.4	-23.9	-7.0	15.3	-2.0
1999	-26.0	-20.0	-15.7	-9.5	-1.4	11.3	12.7	12.3	4.6	-6.6	-13.6	-21.5	-6.1	-21.7	-8.9	12.1	-5.2
2000	-23.0	-21.2	-18.3	-12.2	-0.5	12.3	20.6	12.5	3.6	-5.8	-14.7	-26.1	-6.1	-21.9	-10.3	15.1	-5.6
2001	-21.5	-23.9	-19.9	-11.1	-1.7	7.2	16.6	12.3	9.7	-7.6	-17.2	-19.2	-6.4	-23.8	-10.9	12.0	-5.0
2002	-24.9	-26.9	-22.0	-15.7	-4.2	12.5	16.7	11.3	5.1	-6.3	-13.8	-16.3	-7.0	-23.7	-14.0	13.5	-5.0
2003	-23.6	-28.5	-22.4	-9.1	-1.1	9.9	18.5	14.7	6.3	-1.5	-14.9	-21.3	-6.1	-22.8	-10.9	14.4	-3.4
2004	-29.7	-24.9	-25.1	-14.9	-7.4	9.1	17.0	9.7	3.0	-7.8	-18.3	-26.3	-9.6	-25.3	-15.8	11.9	-7.7
2005	-25.7	-26.2	-19.8	-8.0	-4.2	8.5	14.0	12.5	2.8	-4.2	-14.4	-17.4	-6.8	-26.1	-10.7	11.7	-5.3
2006	-22.7	-19.4	-15.0	-8.5	3.2	15.3	15.5	15.6	8.1	-3.8	-18.4	-17.3	-4.0	-19.8	-6.8	15.5	-4.7
2007	NA	-25.3	-24.3	-9.7	-2.2	9.9	18.6	11.1	1.8	-5.1	-18.9	-24.0	NA	NA	-12.1	13.2	-7.4
2008	-24.5	NA	NA	-10.9	-0.2	10.8	17.5	11.9	2.4	-3.3	-14.6	NA	NA	NA	NA	13.4	-5.2
2009	NA	NA	NA	NA	NA	NA	NA	NA	7.9	-6.7	-13.9	-21.9	NA	NA	NA	NA	-4.2
2010	-24.2	NA	NA	-4.6	-3.5	11.4	17.3	14.4	5.1	-3.3	-12.7	-21.2	NA	NA	NA	14.4	-3.6
2011	NA	NA	NA	-15.1	1.2	10.6	19.1	14.8	6.7	-2.8	-16.6	-20.9	NA	NA	NA	14.8	-4.2
2012	-23.8	-17.1	NA	-10.4	2.5	15.1	19.2	15.3	10.3	-4.2	-17.9	-26.4	NA	-20.6	NA	16.5	-3.9
2013	-27.6	NA	NA	-12.9	-1.3	16.9	14.4	16.1	6.5	-1.6	-15.8	-26.1	NA	NA	NA	15.8	-3.6
2014	NA	NA	NA	-11.4	1.5	15.0	19.6	12.4	3.4	-4.6	-16.9	-22.2	NA	NA	NA	15.7	-6.0
2015	-23.8	NA	-18.8	-12.6	2.6	13.0	15.4	15.2	6.7	-7.2	-14.0	-21.0	NA	NA	-9.6	14.5	-4.8
2016	NA	NA	NA	-13.5	3.6	13.4	17.5	15.6	6.3	-5.5	-11.5	NA	NA	NA	NA	15.5	-3.6
2017	NA	NA	NA	NA	2.6	12.6	17.8	18.6	9.4	NA	-18.1	-19.9	NA	NA	NA	16.3	NA

Saving files under new names and new and directories

Extracting the station number, city and province for file name use.

In [6]:

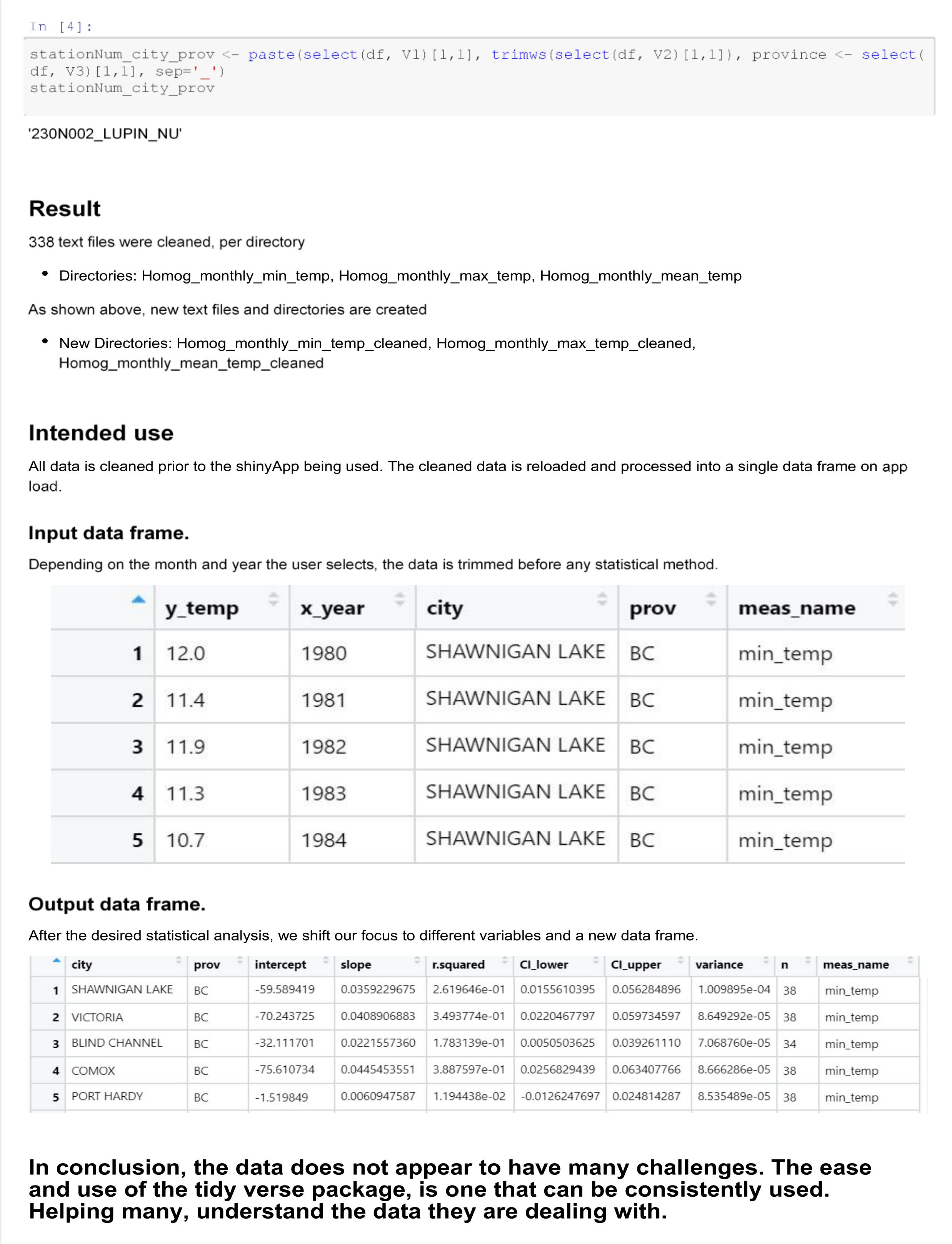
```
df = read.delim('mx230N002.txt', skip = 0, header = FALSE, as.is=TRUE, dec=".", sep = ",",
na.strings=c(" ", "",'NA'), strip.white = TRUE)

(station_number <- select(df, V1)[1,1])
(city <- select(df, V2)[1,1])
(province <- select(df, V3)[1,1])
```

```
'230N002'

'LUPIN'

'NU'
```



In [4]:

```
stationNum_city_prov <- paste(select(df, V1)[1,1], trimws(select(df, V2)[1,1]), province <- select(
df, V3)[1,1], sep='_')
stationNum_city_prov
```

'230N002_LUPIN_NU'

Result

338 text files were cleaned, per directory

- Directories: Homog_monthly_min_temp, Homog_monthly_max_temp, Homog_monthly_mean_temp

As shown above, new text files and directories are created

- New Directories: Homog_monthly_min_temp_cleaned, Homog_monthly_max_temp_cleaned, Homog_monthly_mean_temp_cleaned

Intended use

All data is cleaned prior to the shinyApp being used. The cleaned data is reloaded and processed into a single data frame on app load.

Input data frame.

Depending on the month and year the user selects, the data is trimmed before any statistical method.

	y_temp	x_year	city	prov	meas_name
1	12.0	1980	SHAWNIGAN LAKE	BC	min_temp
2	11.4	1981	SHAWNIGAN LAKE	BC	min_temp
3	11.9	1982	SHAWNIGAN LAKE	BC	min_temp
4	11.3	1983	SHAWNIGAN LAKE	BC	min_temp
5	10.7	1984	SHAWNIGAN LAKE	BC	min_temp

Output data frame.

After the desired statistical analysis, we shift our focus to different variables and a new data frame.

	city	prov	intercept	slope	r.squared	CI_lower	CI_upper	variance	n	meas_name
1	SHAWNIGAN LAKE	BC	-59.589419	0.0359229675	2.619646e-01	0.0155610395	0.056284896	1.009895e-04	38	min_temp
2	VICTORIA	BC	-70.243725	0.0408906883	3.493774e-01	0.0220467797	0.059734597	8.649292e-05	38	min_temp
3	BLIND CHANNEL	BC	-32.111701	0.0221557360	1.783139e-01	0.0050503625	0.039261110	7.068760e-05	34	min_temp
4	COMOX	BC	-75.610734	0.0445453551	3.887597e-01	0.0256829439	0.063407766	8.666286e-05	38	min_temp
5	PORT HARDY	BC	-1.519849	0.0060947587	1.194438e-02	-0.0126247697	0.024814287	8.535489e-05	38	min_temp

In conclusion, the data does not appear to have many challenges. The ease and use of the tidy verse package, is one that can be consistently used. Helping many, understand the data they are dealing with.

Figure 3.1: Notebook demonstrating the data cleaning process

4 Software Engineering

4.1 Description

The name of the application is *shinyClimate*. Built with R studio, and its shinyapp feature. It consists of a server, UI and 4 shiny modules. The app was designed to provide a family friendly illustration of Canadian climate change. Further, giving users a trustworthy understanding of the strength of climate change in each region (city, province, Canada). The website application, shinyClimate was written with R and source code is available at *https://github.com/korede97/shinyClimate.*

4.2 UML (Unified Modeling Language) Diagrams

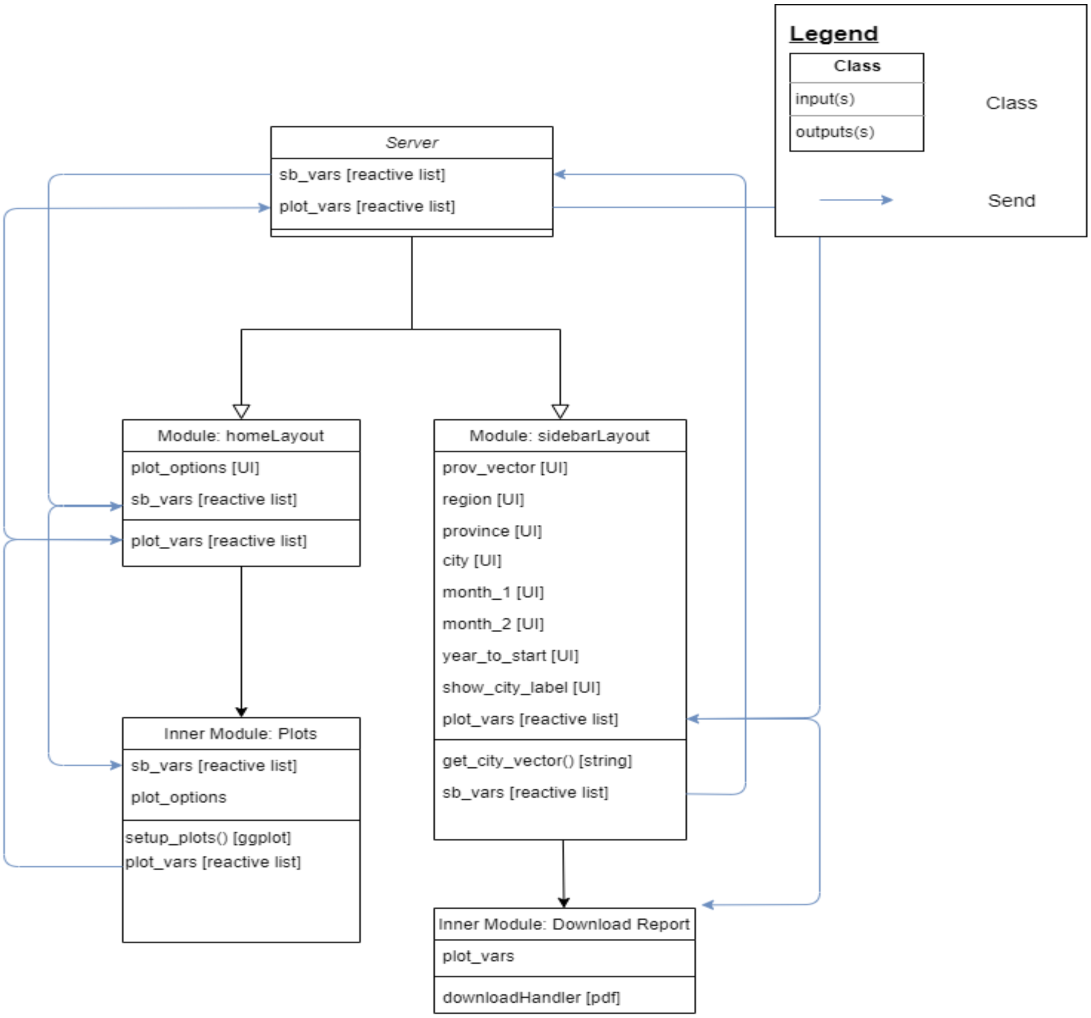


Figure 4.2.1: Class Diagram for *shinyClimate*

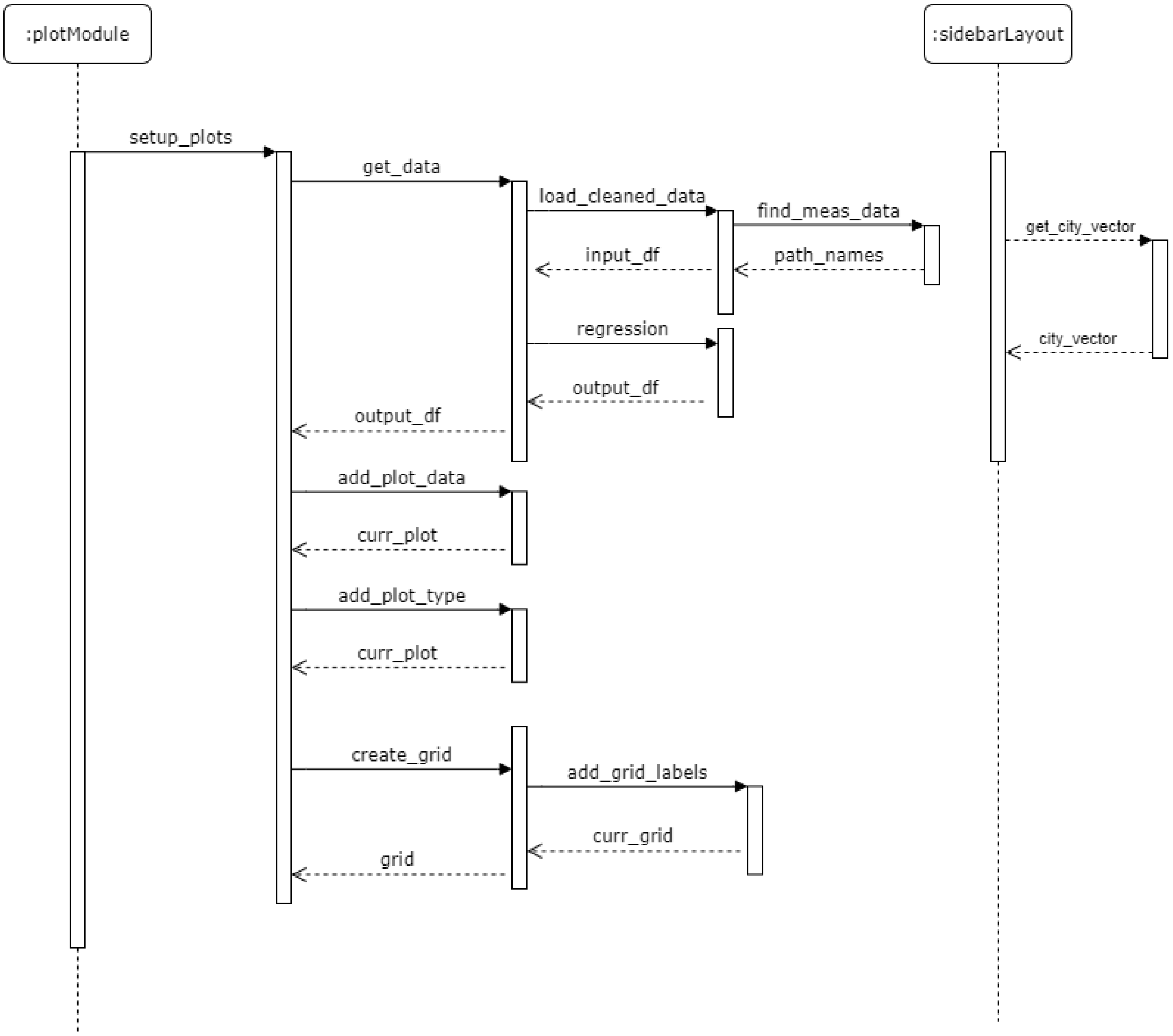


Figure 4.2.2: Sequence Diagram for *shinyClimate*

5 Data Analysis

The statistical method chosen is Multilinear Regression. In mathematical terms, the equation is given by,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Where, $\boldsymbol{y}$ is a response vector formed by temperature values. $\boldsymbol{X},$ the design matrix formed by 1s and the years. $\boldsymbol{\beta}$ denotes a vector of regression coefficients, and $\boldsymbol{\varepsilon}{\sim}N(\boldsymbol{0}, \boldsymbol{\sigma^2})$ is a vector of normally distributed error terms. Consequently, the population variance, $\boldsymbol{\sigma^2}$ is not available. Hence, it must be estimated. Note, the coefficients are the intercept and slope.

Given the temperature data, we have $\boldsymbol{y}$ and $\boldsymbol{X}$. Now it is necessary to estimate the coefficients. In other words,

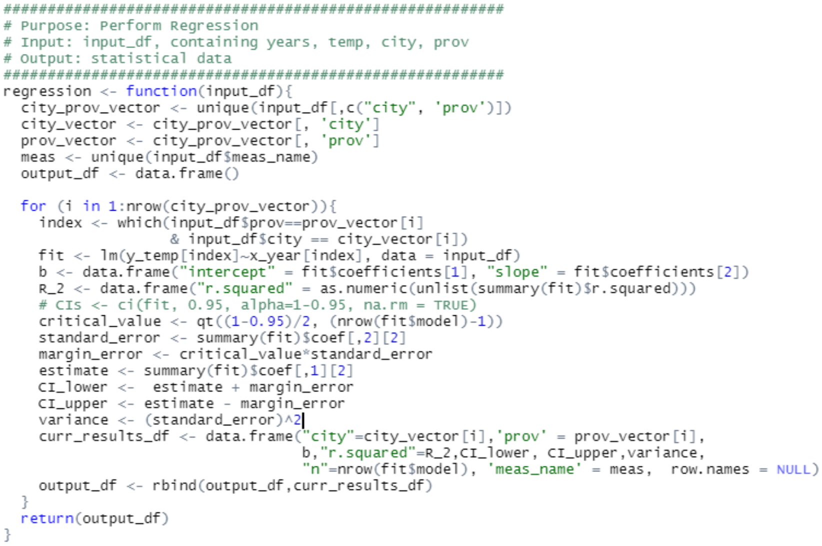
$$\widehat{\boldsymbol{\beta}} \; = \; (\boldsymbol{X^T X})^{-1} \, \boldsymbol{X^T} \, \boldsymbol{y}$$

Furthermore, the multilinear regression line equation is estimated by

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

, where $\widehat{\boldsymbol{y}}$ is a vector of fitted values.

Programmatically, R has a built-in function named *lm.* Conveniently fitting linear models. Given the data, $\boldsymbol{y}$ and $\boldsymbol{X}$ are grouped by respective cities. Further $\widehat{\boldsymbol{\beta}}$, R^2, $\hat{\sigma}$ can be extracted from the returned data frame. Other statistical values such as confidence interval, and $\hat{\sigma}^2$ can be calculated.



```
#########################################################
# Purpose: Perform Regression
# Input: input_df, containing years, temp, city, prov
# Output: statistical data
#########################################################
regression <- function(input_df){
  city_prov_vector <- unique(input_df[,c("city", 'prov')])
  city_vector <- city_prov_vector[, 'city']
  prov_vector <- city_prov_vector[, 'prov']
  meas <- unique(input_df$meas_name)
  output_df <- data.frame()

  for (i in 1:nrow(city_prov_vector)){
    index <- which(input_df$prov==prov_vector[i]
                   & input_df$city == city_vector[i])
    fit <- lm(y_temp[index]~x_year[index], data = input_df)
    b <- data.frame("intercept" = fit$coefficients[1], "slope" = fit$coefficients[2])
    R_2 <- data.frame("r.squared" = as.numeric(unlist(summary(fit)$r.squared)))
    # CIs <- ci(fit, 0.95, alpha=1-0.95, na.rm = TRUE)
    critical_value <- qt((1-0.95)/2, (nrow(fit$model)-1))
    standard_error <- summary(fit)$coef[,2][2]
    margin_error <- critical_value*standard_error
    estimate <- summary(fit)$coef[,1][2]
    CI_lower <-  estimate + margin_error
    CI_upper <- estimate - margin_error
    variance <- (standard_error)^2
    curr_results_df <- data.frame("city"=city_vector[i],'prov' = prov_vector[i],
                                  b,"r.squared"=R_2,CI_lower, CI_upper,variance,
                                  "n"=nrow(fit$model), 'meas_name' = meas,  row.names = NULL)
    output_df <- rbind(output_df,curr_results_df)
  }
  return(output_df)
}
```

Figure 5.1: Regression function from shinyClimate

Throughout, extreme values are the focus. With an intention of providing the trustworthy evidence of climate change. The default months, January and July resemble months after the solstice. In addition, plots of minimum and maximum temperatures for a given month, are displayed together.

To further analyze a specific region, begin with the slopes to evaluate the steepness of the temperature trend. Proceed to, confidence intervals considering the variation of the slopes found within a region. Finally observe R^2, the coefficient of determination, to assess the goodness of the model. In other words, the amount of the variation in slopes that is explained by the temperatures over the years.

When, analyzing the confidence intervals our goal is to determine whether the slope is significant or not. Formally, we must conduct hypothesis tests. Define the null hypothesis, H_0 as confidence interval contains zero and alternative hypothesis, H_A as confidence interval does not contain zero. If we reject H_0, the slope is significant. If we do not reject H_0, the slope is not significant. Note that, if the confidence interval contains zero, we do not know whether the slope is increasing or decreasing. Hence, we cannot make a strong conclusion about the slope. Otherwise, if the confidence interval does not contain zero it is exclusively increasing or decreasing.

The following examples were performed in R Studio using R and software is available from the site *http://ka97.shinyapps.io/shinyClimate.*

5.1 City – Example

[City – Flin Flon]

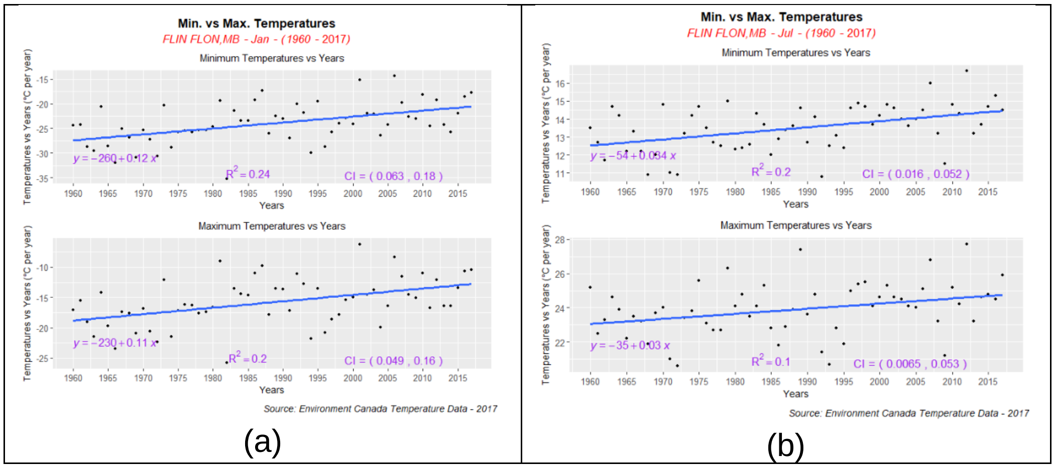
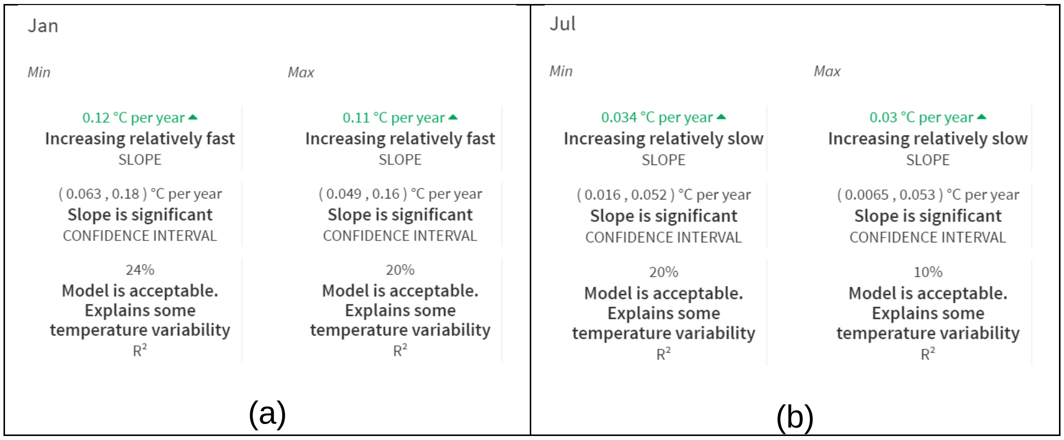


Figure 4.1.1: Regression line plot – Minimum vs Maximum temperatures for Toronto. Plot includes the regression line equation ($\boldsymbol{y = b_0 + b_1 x}$), coefficient of determination ($\boldsymbol{R^2}$), and Confidence Intervals. (a) the month of January and (b) the month of July



Jan

Min

0.12 °C per year ▲
Increasing relatively fast
SLOPE

(0.063 , 0.18) °C per year
Slope is significant
CONFIDENCE INTERVAL

24%
Model is acceptable. Explains some temperature variability
R²

Max

0.11 °C per year ▲
Increasing relatively fast
SLOPE

(0.049 , 0.16) °C per year
Slope is significant
CONFIDENCE INTERVAL

20%
Model is acceptable. Explains some temperature variability
R²

(a)

Jul

Min

0.034 °C per year ▲
Increasing relatively slow
SLOPE

(0.016 , 0.052) °C per year
Slope is significant
CONFIDENCE INTERVAL

20%
Model is acceptable. Explains some temperature variability
R²

Max

0.03 °C per year ▲
Increasing relatively slow
SLOPE

(0.0065 , 0.053) °C per year
Slope is significant
CONFIDENCE INTERVAL

10%
Model is acceptable. Explains some temperature variability
R²

(b)

Figure 4.1.2: Distribution shapes for plots in *Figure 4.1.1*

The minimum and maximum temperatures for January indicate that temperatures are increasing relatively fast. Taking a deeper look at the winter season months is worthwhile. Similarly, for the month of July, minimum and maximum temperature increase, but at a relatively slow rate. Hence, January temperatures are increasing faster than that of July.

The 95% confidence intervals suggest that the slopes are significant. Therefore, there is strong evidence of temperature increase in January and July. Moreover, there is slight evidence of temperature increase for each month. Note, January and February resemble the solstice months. Thus, we are using these extremes to imply whether there is evidence of temperature increase throughout the year.

Finally, R^2 for both months generally implies that the model is acceptable. Hence, it explains some temperature variability. The value of R^2 remains acceptable, since we know there are other factors affecting temperature. In addition, there is slight evidence of climate change since the model explains some variability.

5.2 Province – Example

[Province – Manitoba]

Histograms show the approximate distribution of values. It can be further analyzed by its shape, spread and outliers. Shapes could possibly be defined by, left-skewed, right-skewed, bell-shaped, uniform or bimodal.

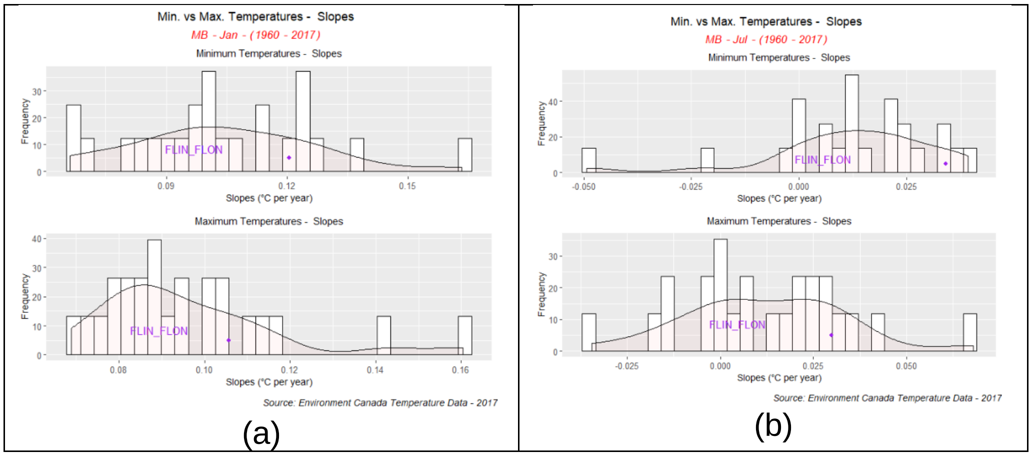
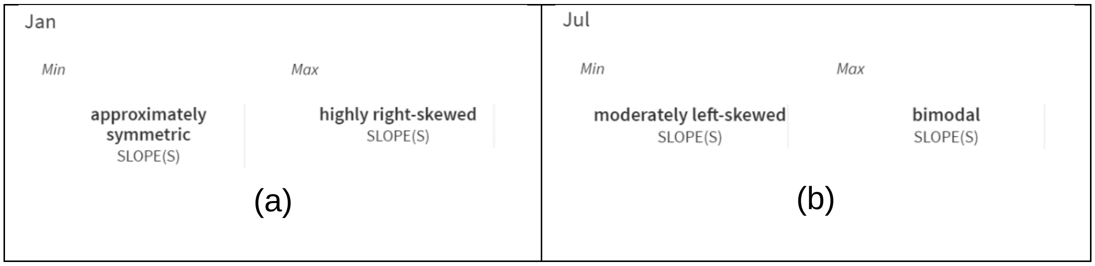


Figure 4.2.1: Histogram of slopes – Minimum vs Maximum temperatures for Ontario. (a) the month of January and (b) the month of July



Jan

Min

approximately symmetric
SLOPE(S)

Max

highly right-skewed
SLOPE(S)

(a)

Jul

Min

moderately left-skewed
SLOPE(S)

Max

bimodal
SLOPE(S)

(b)

Figure 4.2.2: Distribution shapes for plots in *Figure 4.2.1*

To start, the minimum temperatures for January indicate that the distribution is approximately symmetric. It has a large variation, and some outliers. Thus, the central tendency of slopes is about 0.10 °C per year. The maximum temperature for January indicates a highly right-skewed distribution, with a central tendency around 0.09 °C per year. Thus, most of Manitoba is increasing relatively fast for the month of January. In contrast, the minimum temperatures for July, propose a moderately left-skewed distribution. With a spread ranging from negative to positive slopes and outliers. The bulk of the values seem to be around 0.001°C per year, a slowly increasing rate. Maximum temperatures for July indicate a bimodal distribution. More specifically, two distributions can be found. Further exploration into grouping the cities for the month of July may be necessary to make a stronger conclusion. Clearly, January is increasing faster than July. Also, note for all distributions the Flin Flon label suggests that the city's slope is higher than majority of the cities in the province.

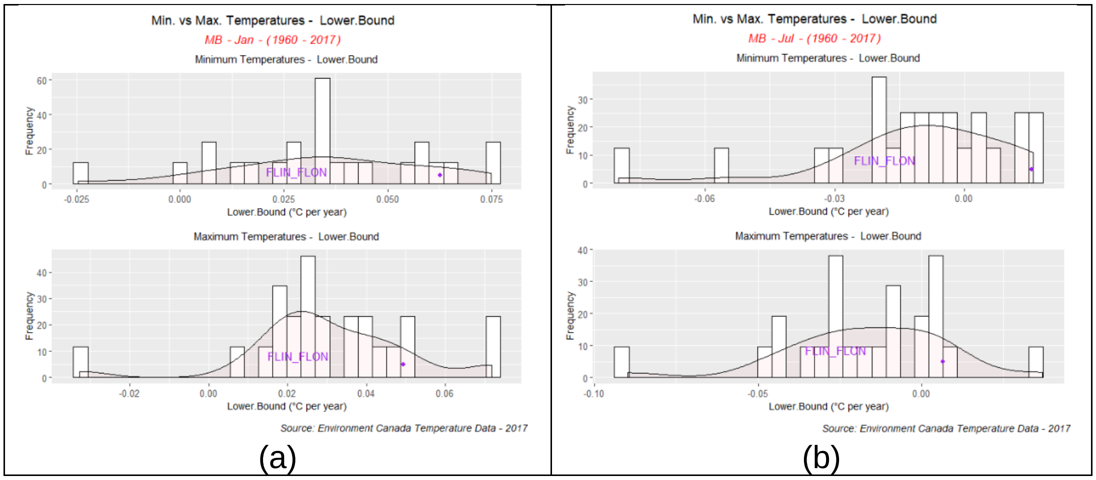


Figure 4.2.3: Histogram of lower confidence bound of slopes – Minimum vs Maximum temperatures for Ontario. (a) the month of January and (b) the month of July

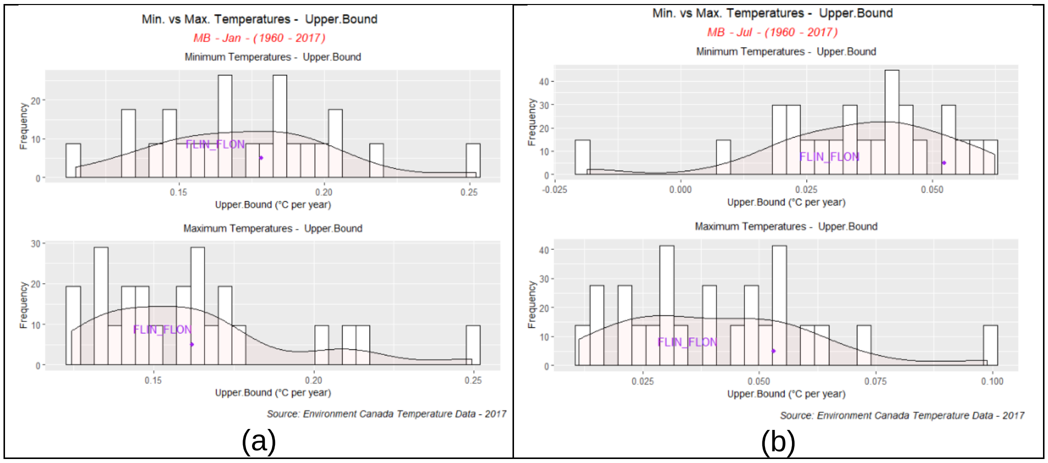
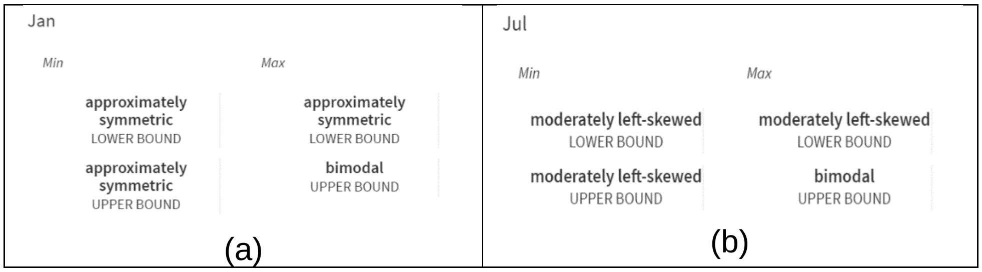


Figure 4.2.4: Histogram of upper confidence bound of slopes for Ontario – Minimum vs Maximum temperatures for Ontario. (a) the month of January and (b) the month of July



Jan

Min	Max
approximately symmetric LOWER BOUND	approximately symmetric LOWER BOUND
approximately symmetric UPPER BOUND	bimodal UPPER BOUND

(a)

Jul

Min	Max
moderately left-skewed LOWER BOUND	moderately left-skewed LOWER BOUND
moderately left-skewed UPPER BOUND	bimodal UPPER BOUND

(b)

Figure 4.2.5: Distribution shapes for plots in *Figure 4.2.3* and *Figure 4.2.4*

Given the 95% lower and upper bounds for January. We have that the slope is significant. Hence, Manitoba is experiencing an increase in temperature for the month of January. Next off, minimum and maximum confidence bounds for July suggest that the slope is not significant or slightly significant. More specifically, the lower bounds indicate that trend for the province is towards a negative slope. Whereas the upper bounds show that the trend is positive.

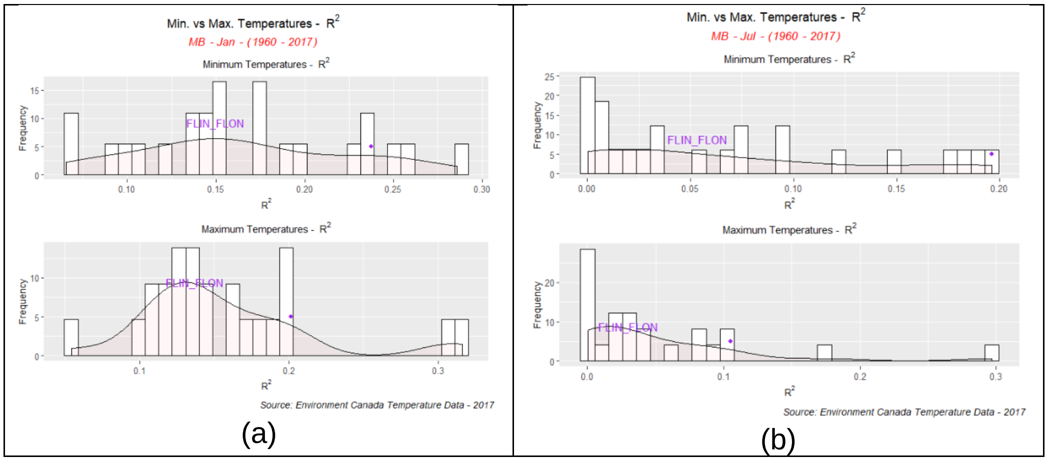
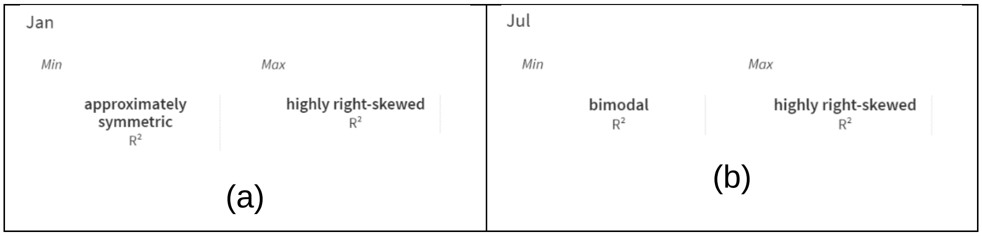


Figure 4.2.6: Histogram of $\boldsymbol{R^2}$ for slopes – Minimum vs Maximum temperatures for Ontario. (a) the month of January and (b) the month of July



Jan

Min approximately symmetric R^2

Max highly right-skewed R^2

(a)

Jul

Min bimodal R^2

Max highly right-skewed R^2

(b)

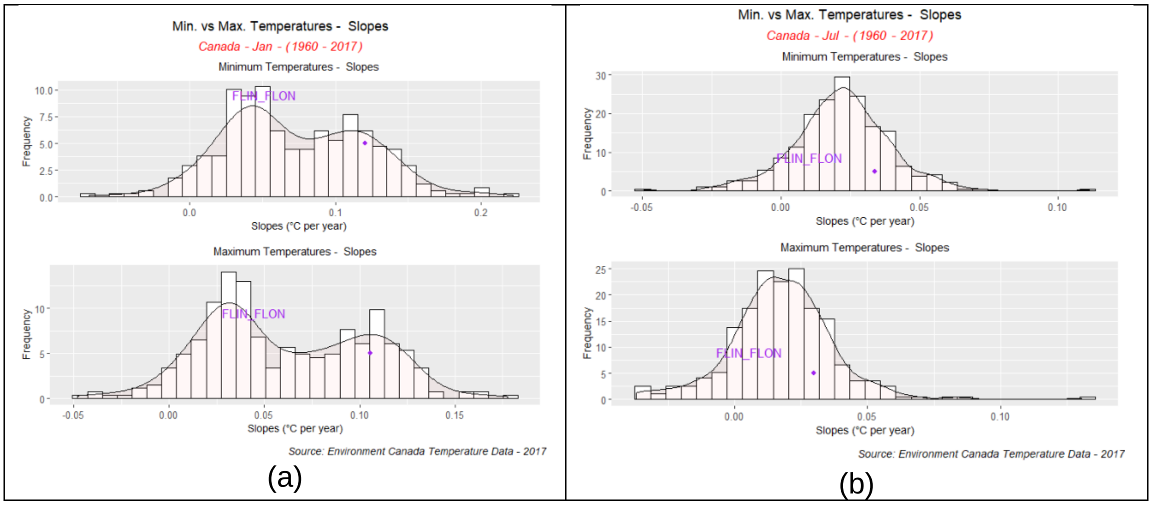
Figure 4.2.7: Distribution shapes for plots in *Figure 4.2.3*

January indicates that the goodness of the model is acceptable, whereas for July it is weak. Hence, variation of values and the bulk of values for January suggest some of the increase in temperature is explained by the increase in years. As for July, approximately none of the increase in temperature is explained by increase in years. Therefore, adding variables to the regression model may be necessary. Further, discovering which factors have the most effect on climate change.

In conclusion, Manitoba is experiencing significant evidence of temperature increase for January and the model is acceptable. In contrast, July has slight evidence of temperature increase. More specifically, we should further investigate by adding more factors, finding some grouping for the province or only analyze the cities within.

5.3 Nation – Example

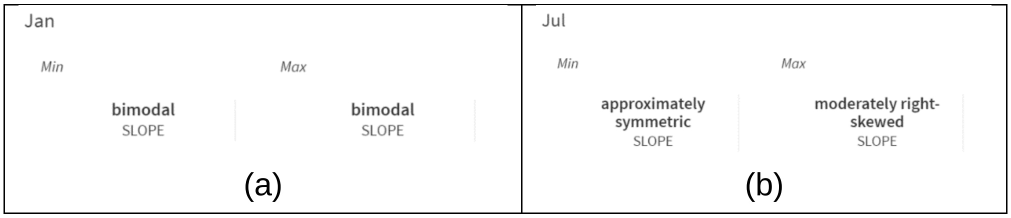
[Nation – Canada]



(a)

(b)

Figure 4.3.1: Histogram of slopes – Minimum vs Maximum temperatures for Canada. (a) the month of January and (b) the month of July



Jan

Min

bimodal
SLOPE

Max

bimodal
SLOPE

(a)

Jul

Min

approximately symmetric
SLOPE

Max

moderately right-skewed
SLOPE

(b)

Figure 4.3.2: Distribution shapes for plots in *Figure 4.3.1*

The slopes for minimum and maximum temperature for January propose a bimodal distribution. Thus, summary of data is proportional. Where, one proportion is best characterized by the range of smaller slopes and the other proportion by the range of larger slopes. More importantly, the distributions indicate that Canada is undergoing an increase in temperature for January. Additionally, temperatures are increasing relatively slow and increasing relatively fast. Regarding July, we see that the slopes are summarized differently than that of January. Minimum and maximum temperatures suggest that the central tendency is near 0.025 °C per year. Thus, slopes are increasing relatively slow.

Boxplots allow us to analyze the skewness of data illustrated by the box in respect to the center line. Outliers given by the dots. And variation given by the whiskers. Most importantly, we can conveniently compare the distribution of data across all provinces.

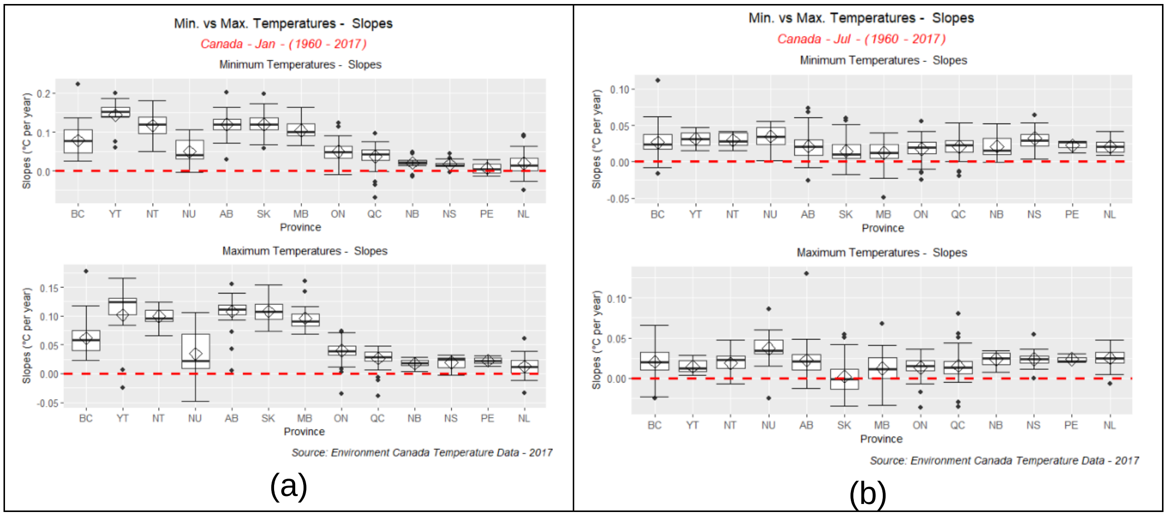


Figure 4.3.3: Boxplot of slopes – Minimum vs Maximum temperatures for Canada. (a) the month of January and (b) the month of July

First off, for the month of January, the center lines for each province is above zero. The variation of slopes for each province varies. Northwest Territories, Nunavut and Prince Edward Island have no outliers. Suggesting that, the exploratory analysis into those provinces would be significant. Yukon, Northwest Territories, Alberta, Saskatchewan and Manitoba consistently have the highest medians. Thus, these provinces must have more similarities. Adding factors to discover what the similarities are, is worthwhile. Similarly, for Ontario, Quebec, New Brunswick, Nova Scotia, Prince Edward Island and Newfoundland. Note, Nunavut seems to be characterized by its own climatic region. Secondly, for the month of July the median lines are approximately the same. Additionally, we are seeing that there are less outliers. Clearly, the slopes for January are increasing faster than that of July.

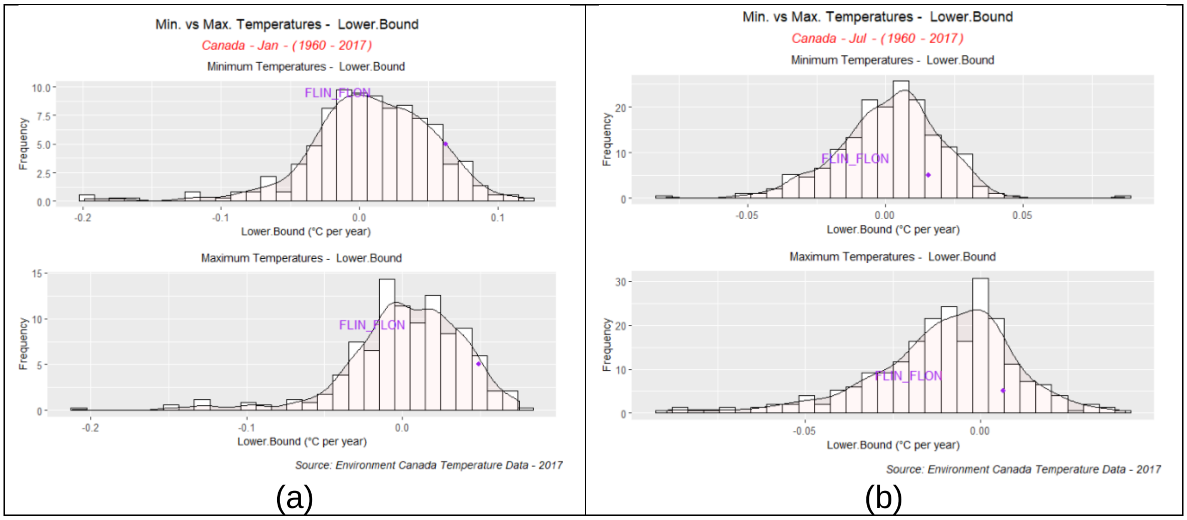


Figure 4.3.5: Histogram of lower confidence bound for slopes – Minimum vs Maximum temperatures for Canada. (a) the month of January and (b) the month of July

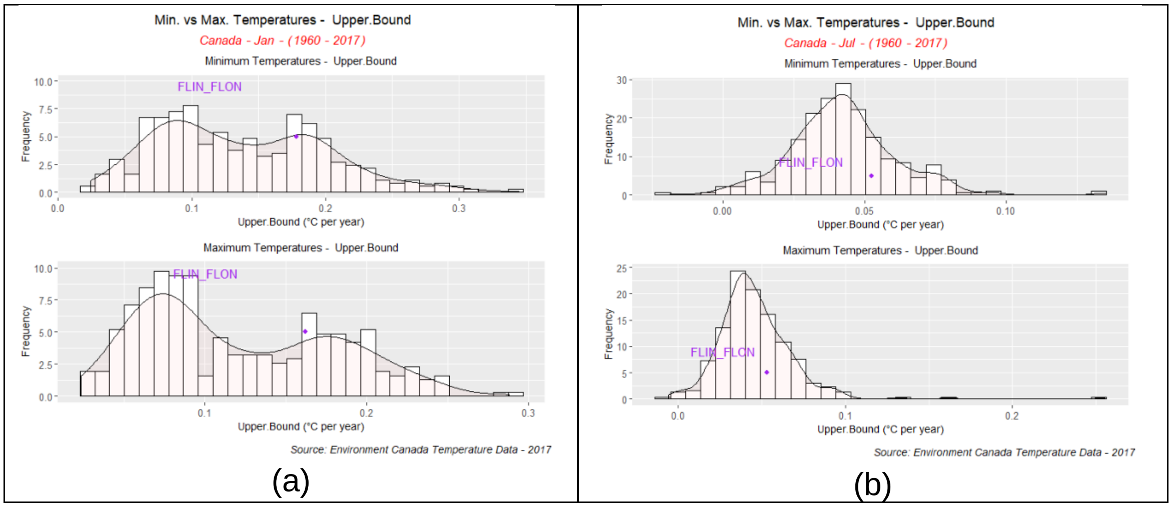
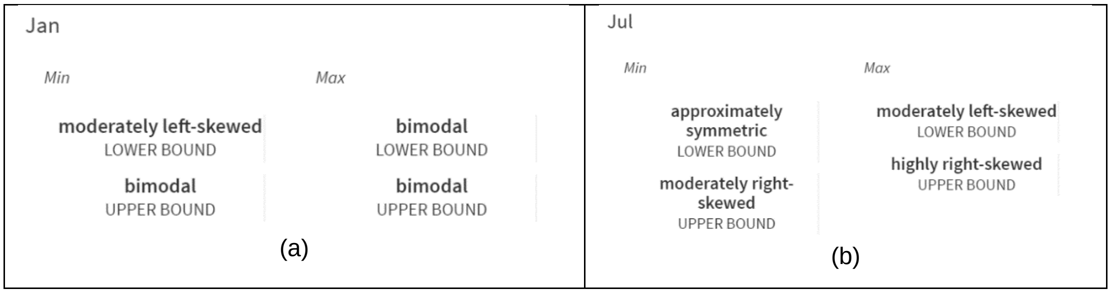


Figure 4.3.6: Histogram of upper confidence bound for slopes – Minimum vs Maximum temperatures for Canada. (a) the month of January and (b) the month of July



Jan

Min

moderately left-skewed
LOWER BOUND

bimodal
UPPER BOUND

Max

bimodal
LOWER BOUND

bimodal
UPPER BOUND

(a)

Jul

Min

approximately symmetric
LOWER BOUND

moderately right-skewed
UPPER BOUND

Max

moderately left-skewed
LOWER BOUND

highly right-skewed
UPPER BOUND

(b)

Figure 4.3.7: Distribution shapes for plots in *Figure 4.3.5* and *Figure 4.3.6*

Here, we are most interested in the testing the significance of the trends. For January, we cannot clearly deduce the significance of the trend. We previously assumed that the distribution of slopes can be further characterized. But do not know what that characterization may be. Note, further analysis is needed for this characterization. Presently, the proportion with a range of larger trends is significant, whereas the one with a range of smaller trends region trend is slightly significant. The slopes for January are moderately significant, regardless of proportions. In comparison to July, the trends are not significant, say slightly significant.

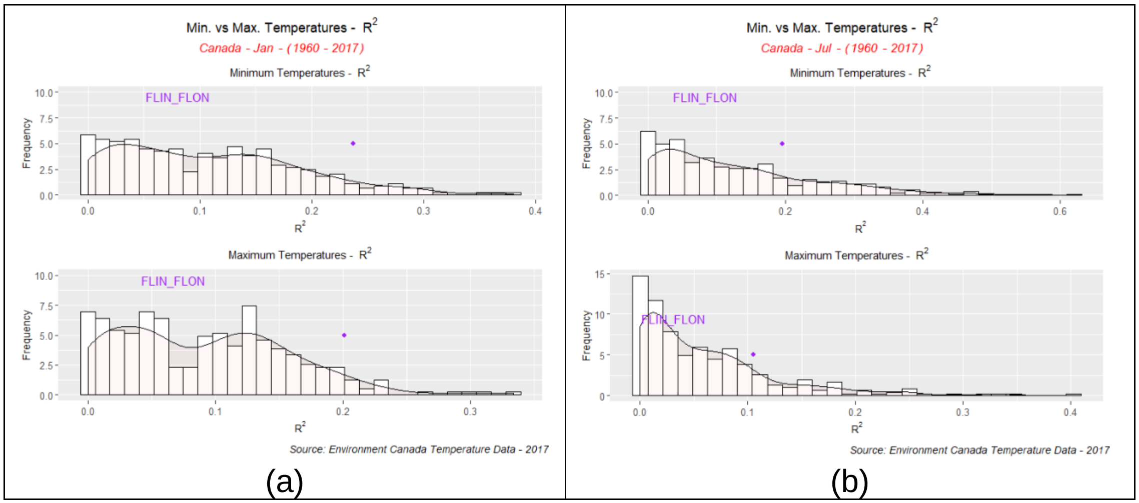


Figure 4.3.8: Histogram of $\boldsymbol{R^2}$ for slopes – Minimum vs Maximum temperatures for Canada. (a) the month of January and (b) the month of July

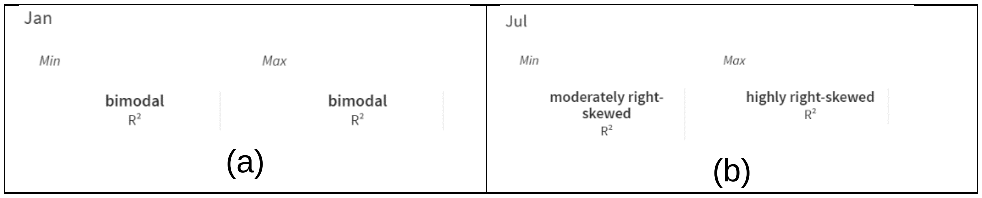


Figure 4.3.9: Distribution shapes for plots in *Figure 4.3.8*

The coefficient of determination for January, similarly, suggests a bimodal distribution exists. The models given the bimodal distribution of slopes are somewhat acceptable. Furthermore, better models can be constructed to make good decisions. Specifically, since the slopes and R^2 give bimodal distribution, there is reason to believe some sort of characterization can be made to better analyze the data for January. In contrast, July indicates the model is also somewhat acceptable. More specifically, the distributions begin from a low coefficient of determination but there is right skewness.

In conclusion, for January, significant evidence of temperature increase exists. This suggests, the winter season is experiencing increasing temperatures. For July there is slight evidence of temperature increase throughout Canada.

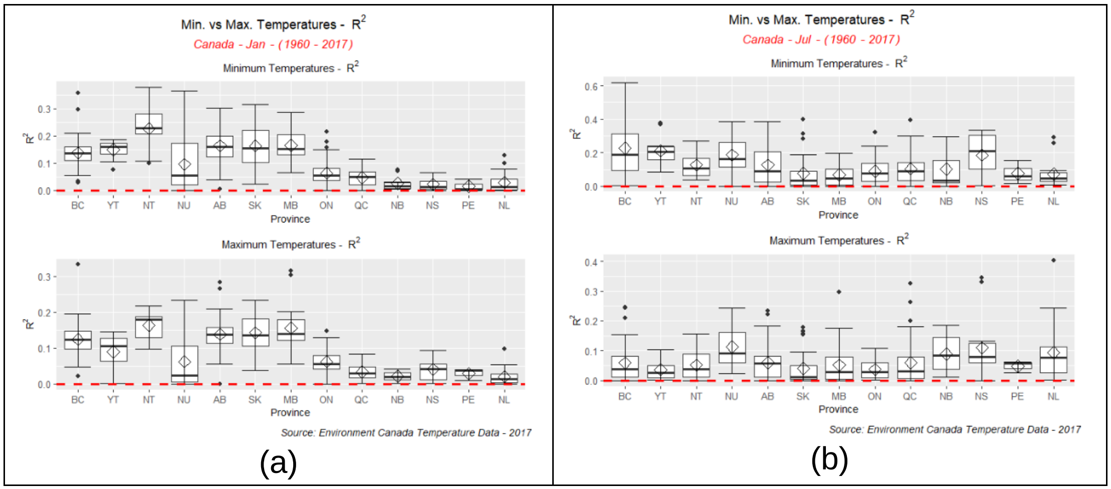


Figure 4.3.10: Boxplot of $\boldsymbol{R^2}$ for slopes – Minimum vs Maximum temperatures for Canada. (a) the month of January and (b) the month of July

It is clear, that R^2 across all provinces suggests increase in years somewhat explains an increase in temperature. Furthermore, the models that have temperatures increasing relatively fast for January seem to have a larger R^2. Thus, we conclude that the provinces that do not explain much of the variability in temperatures need another factor that better defines the variability.

5.4 Summary

Temperature slopes vary within Canada, provinces and cities. Majority of Canada is in fact, experiencing increasing temperatures. Yukon and the Northwest Territories are northern regions, which temperatures are increasing at a faster rate than the southern regions**.** December through February warm faster than the other months. Minimum temperatures can be decreasing (or increasing) while maximum temperatures increasing (or decreasing).

Cities with extreme warming, within a province; Yukon Territory (Dawson, Pelly Ranch), Northwest Territories (Fort Good Hope, Fort Smith), Nunavut (Pelly Bay, Ennadai Lake), British Columbia (Creston, Glacier, Kelowna), Manitoba (Flin Flon, Norway House), Saskatchewan (Waskesiu Lake, Loon Lake, Yellow Grass), Alberta (Coronation, Entrance), Quebec (Bagotville, Kuujjuarapik), Ontario (Beatrice, Cornwall, Dryden, Ottawa), Newfoundland and Labrador (Cartwright, Nain), Prince Edward Island (Charlottetown, Monticello, Summerside), Nova Scotia (Collegeville, Greenwood, Yarmouth), New Brunswick (Moncton, Woodstock)

6 Appendix A: Supplementary data

Adegboye, K. (2020, February). korede97/shinyClimate. Retrieved May 2020, from *https://github.com/korede97/shinyClimate*

Adegboye, K. (2020, February). korede97/shinyClimate. Retrieved May 2020, from *https://github.com/korede97/shinyClimate/blob/master/data_cleaning.ipynb*

Adegboye, K. (2020, April). Showcasing evidence of climate change. Retrieved May 7, 2020, from http://ka97.shinyapps.io/shinyClimate

7 References

Climate Change Canada. (2017, August 9). Government of Canada. Retrieved from *https://www.canada.ca/en/environment-climate-change/services/climate-change/science-research-data/climate-trends-variability/adjusted-homogenized-canadian-data/surface-air-temperature-access.html*