DATA Project

ARMY

2021-10-14

Contents

Title: Weather Change for Calgary vs. Edmonton

Data Collection:

- The data set (**Eighty years of Candian climate data**) is obtained from *Kaggle*. This dataset has been compiled from public sources, mainly from *Government of Canada*. The whole data set consists of daily temperatures and precipitation from 13 Canadian centers. Precipitation is either rain or snow (likely snow in winter months).
- Because few of Canada's weather stations have been operating continuously, so the final data set was patched together.
- For our interest we only focus on Alberta Weather change mainly Calgary and Edmonton. As part of effort to tidy the data set, we removed records that are NA in either city.
- In order to get a more tangible conclusion, we are mainly looking at temperature or precipitation over the last 30 years.
- To make our analysis more comparable, daily records are grouped into 4 different seasons, i.e.: "winter" "spring" "summer" "autumn".

Purpose/Motivation:

- There are several reasons we want to investigate the weather change in Calgary vs. Edmonton
  - On the high level, we are curious to see if our local environment is trending the same way as the global warming over the past 30 years.
  - On the practical level, we want to investigate the overall weather pattern locally, so that we can provide educated advice like travel tips for different seasons in Alberta.
  - Also because Calgary and Edmonton are geographically closed to each other, they are theoretically quite correlated in terms of weather. As such, we want to explore the strength of this association so that data in one location can be used to to predict the other if somehow data is hard to achieve for one particular location.

Statistical Analysis:

- A list of visualizaiton used in this report:
  - Scatter plot to show the the temperature/precipitation change overtime
  - Barplot to compare the mean value of cities
  - Pairwise plot to show association between cities
  - Time series line graph to show change overtime
  - Histogram/Density plot to show underlying distribution
  - qqplot to check normality assumption
- A list of Statistical Methods:
  - 95% confidence interval for population mean and proportion p
  - Bootstrap distribution and confidence interval
  - Testing hypothesis
    - Estimation of one population mean
    - Estimation of two population difference (mean_calgary - mean_edmonton)
    - Testing of equal variance
  - Simple Linear regression

Conclusion:

NOTE: Maybe we should try to think of one plot that can summarize everything?

- Is Alberta's local weather trending the same way as the global warming over the past 30 years?
- What are Alberta's local weather patterns like? What advice would be suggested to travelers coming to Alberta for each season?
- Do Calgary and Edmonton have similar weather?

Reference:

- https://weatherspark.com/y/2349/Average-Weather-in-Calgary-Canada-Year-Round
- https://www.kaggle.com/aturner374/eighty-years-of-canadian-climate-data

Data Visualizaiton

Data Collection and Transformation

- We will first need to manually add another column of seasons based on: spring runs from March 1 to May 31; summer runs from June 1 to August 31; fall (autumn) runs from September 1 to November 30; and. winter runs from December 1 to February 28 (February 29 in a leap year).
- In order to get a more tangible conclusion, we are mainly looking at temperature or precipitation over the last 30 years
- The dataset was presented in two formats:
  - wide format where all the CITY sit in the heading
  - long format where all the CITY sit in one column, with DATE and SEASON acting the combinatorial ID

```
#first read in the Canadian_climate_history data file
setwd("/Users/yongpengfu/OneDrive - University of Calgary/Master in Data Science and
↪  Analytics/Courses/DATA 602 L03 - (Fall 2021) - Statistical Data
↪  Analysis/Assignment/Project/DATA602_Project")
Canadian_climate_history <- as_tibble(read.csv("Canadian_climate_history.csv",
↪  header = T))
#convert LOCAL_DATE to Year-Month-Date
Canadian_climate_history$LOCAL_DATE <- as.Date(Canadian_climate_history$LOCAL_DATE,
↪  format = "%d-%b-%Y")

#So I think we can group different months manually according to different seasons.
↪  And this becomes a new factor column
Canadian_climate_history %<>% mutate(season =
↪  time2season(Canadian_climate_history$LOCAL_DATE, out.fmt = "seasons")) %>%
↪  mutate(season = replace(season, season == "autumm", "autumn"))

#This is a tidy dataframe into long format for ggplot later on.
Canadian_climate_history_long <- Canadian_climate_history %>%  gather(City,Measure,
↪  -LOCAL_DATE, -season) %>% separate(City, into = c("Mean", "Variable", "CITY"),
↪  sep = "_") %>% unite("VARIABLE", Mean:Variable, remove = F) %>%
↪  dplyr::select(-c(Mean, Variable))

#In order to get a more tangible conclusion, we are mainly looking at temperature or
↪  precipitation over the last 30 years.
Canadian_climate_history_last30 <- Canadian_climate_history %>% filter(LOCAL_DATE >=
↪  "1991-01-01")
Canadian_climate_history_long_last_30<- Canadian_climate_history_long %>%
↪  filter(LOCAL_DATE >= "1991-01-01")

#show the first few records for both dataset
Canadian_climate_history_last30 %>% head(3)
```

```
## # A tibble: 3 x 28
##   LOCAL_DATE MEAN_TEMPERATUR~ TOTAL_PRECIPITA~ MEAN_TEMPERATUR~ TOTAL_PRECIPITA~
##   <date>                <dbl>            <dbl>            <dbl>            <dbl>
## 1 1991-01-01            -21.7                0            -25.7                0
## 2 1991-01-02            -18.7                0            -23.5                0
## 3 1991-01-03            -19.4                0            -20.2                0
## # ... with 23 more variables: MEAN_TEMPERATURE_HALIFAX <dbl>,
## #   TOTAL_PRECIPITATION_HALIFAX <dbl>, MEAN_TEMPERATURE_MONCTON <dbl>,
## #   TOTAL_PRECIPITATION_MONCTON <dbl>, MEAN_TEMPERATURE_MONTREAL <dbl>,
## #   TOTAL_PRECIPITATION_MONTREAL <dbl>, MEAN_TEMPERATURE_OTTAWA <dbl>,
## #   TOTAL_PRECIPITATION_OTTAWA <dbl>, MEAN_TEMPERATURE_QUEBEC <dbl>,
## #   TOTAL_PRECIPITATION_QUEBEC <dbl>, MEAN_TEMPERATURE_SASKATOON <dbl>,
## #   TOTAL_PRECIPITATION_SASKATOON <dbl>, MEAN_TEMPERATURE_STJOHNS <dbl>,
## #   TOTAL_PRECIPITATION_STJOHNS <dbl>, MEAN_TEMPERATURE_TORONTO <dbl>,
## #   TOTAL_PRECIPITATION_TORONTO <dbl>, MEAN_TEMPERATURE_VANCOUVER <dbl>,
## #   TOTAL_PRECIPITATION_VANCOUVER <dbl>, MEAN_TEMPERATURE_WHITEHORSE <dbl>,
## #   TOTAL_PRECIPITATION_WHITEHORSE <dbl>, MEAN_TEMPERATURE_WINNIPEG <dbl>,
## #   TOTAL_PRECIPITATION_WINNIPEG <dbl>, season <chr>
```

```
Canadian_climate_history_long_last_30 %>% head(3)
```

```
## # A tibble: 3 x 5
##   LOCAL_DATE season VARIABLE         CITY    Measure
##   <date>     <chr>  <chr>            <chr>     <dbl>
## 1 1991-01-01 winter MEAN_TEMPERATURE CALGARY   -21.7
## 2 1991-01-02 winter MEAN_TEMPERATURE CALGARY   -18.7
## 3 1991-01-03 winter MEAN_TEMPERATURE CALGARY   -19.4
```

- For our interest we only focus on Alberta Weather change mainly Calgary and Edmonton. As part of effort to tidy the data set, we removed records that are NA in either city.

```
#How many cities recorded and retrieve their names
all_citys <- sapply(strsplit(names(Canadian_climate_history[-1]), "_"), "[", 3)%>%
↪  discard(is.na)
unique(all_citys) %>% print
```

```
##  [1] "CALGARY"    "EDMONTON"   "HALIFAX"    "MONCTON"    "MONTREAL"
##  [6] "OTTAWA"     "QUEBEC"     "SASKATOON"  "STJOHNS"    "TORONTO"
## [11] "VANCOUVER"  "WHITEHORSE" "WINNIPEG"
```

```
#choose the "CALGARY", "EDMONTON" for further analysis
selected_city <- seq(2,length(names(Canadian_climate_history))-1)[all_citys %in%
↪  c("CALGARY", "EDMONTON")]
Canadian_climate_history_last30_Cal_Edm <-
↪  Canadian_climate_history_last30[,c(1,selected_city,
↪  length(names(Canadian_climate_history)))]
Canadian_climate_history_long_last_30_Cal_Edm <-
↪  Canadian_climate_history_long_last_30%>% filter(CITY == "CALGARY" | CITY ==
↪  "EDMONTON")
#show the first few records for both dataset
Canadian_climate_history_last30_Cal_Edm %>% head(3)
```

```
## # A tibble: 3 x 6
##   LOCAL_DATE MEAN_TEMPERATUR~ TOTAL_PRECIPITA~ MEAN_TEMPERATUR~ TOTAL_PRECIPITA~
##   <date>                <dbl>            <dbl>            <dbl>            <dbl>
## 1 1991-01-01            -21.7                0            -25.7                0
## 2 1991-01-02            -18.7                0            -23.5                0
## 3 1991-01-03            -19.4                0            -20.2                0
## # ... with 1 more variable: season <chr>
```

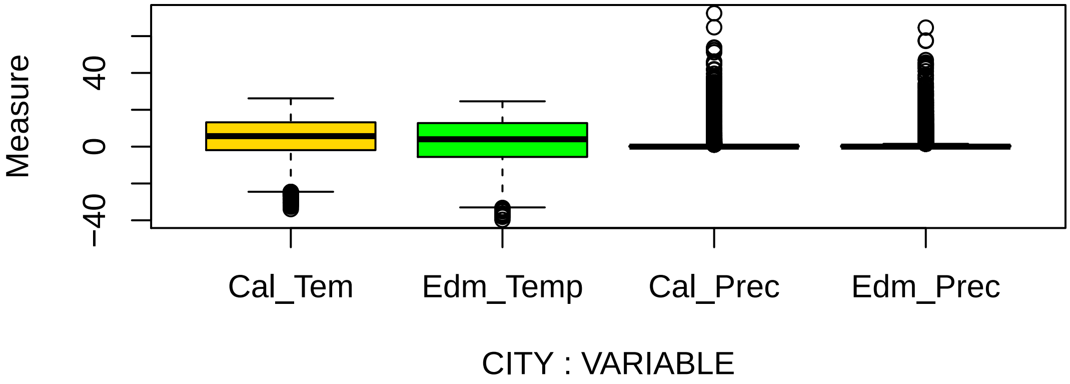
```
Canadian_climate_history_long_last_30_Cal_Edm %>% head(3)
```

```
## # A tibble: 3 x 5
##   LOCAL_DATE season VARIABLE         CITY    Measure
##   <date>     <chr>  <chr>            <chr>     <dbl>
## 1 1991-01-01 winter MEAN_TEMPERATURE CALGARY   -21.7
## 2 1991-01-02 winter MEAN_TEMPERATURE CALGARY   -18.7
## 3 1991-01-03 winter MEAN_TEMPERATURE CALGARY   -19.4
```

Boxplot for Calgary and Edmonton

- We want to look at the spread of Temperature and Precipitation over 2 cities.
- This is to provide some indication of the data symmetry and skewness.

```
#first lets take a look at the boxplot for Calgary Temperature and Percipitation
boxplot(Measure~CITY*VARIABLE,data = Canadian_climate_history_long_last_30_Cal_Edm,
↪  col=c("gold", "green", "red", "blue"),
        names = c("Cal_Tem", "Edm_Temp", "Cal_Prec", "Edm_Prec"))
```

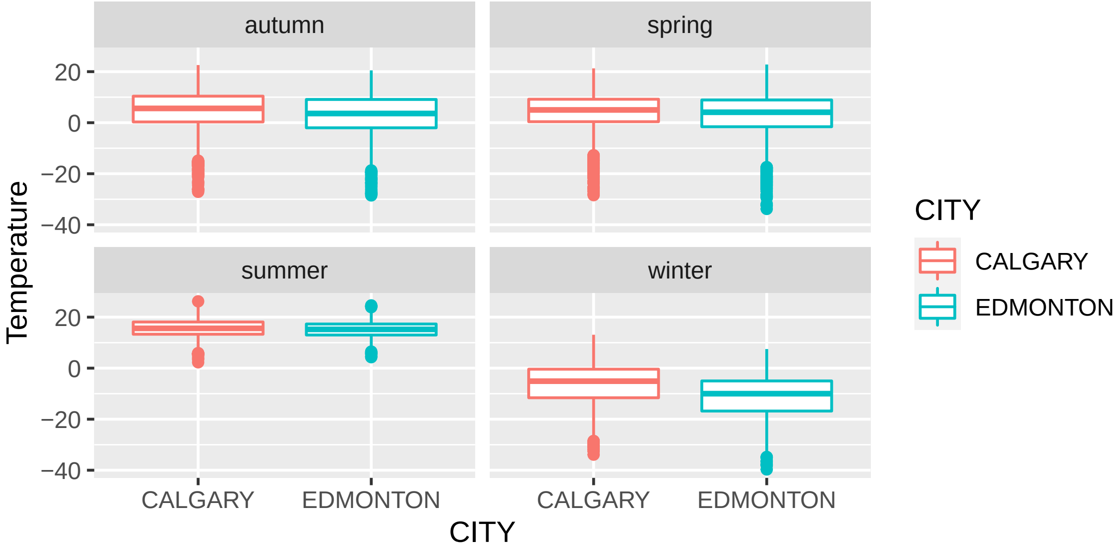


- From this boxplot, we can see Temperature for both Calgary and Edmonton are quite clustered together from 0 to 20 degree, with Edmonton mean temperature in general a little lower than Calgary. There are outliers below lower fence, indicating there are some extreme cold temperature. This is not a big surprise in winter for Alberta. However, it does urge us to separate them further into different seasons. - Precipitation on the other hand has more "outliers" above the upper fence, indicating precipitation is highly skewed. We need to deal with this with extra cautions when it comes to parametric testing.

- Lets take a look at the Temperature by Seasons boxplots.
  - Based on the following plot, it does align with previous finding that Calgary in general has higher temperature than Edmonton, especially during "autumn" and "winter".

```
ggplot(Canadian_climate_history_long_last_30_Cal_Edm %>% filter(VARIABLE ==
↪  "MEAN_TEMPERATURE")) + geom_boxplot(aes(x= CITY, y = Measure, color = CITY)) +
↪  facet_wrap(~season) + ylab("Temperature")
```

```
## Warning: Removed 283 rows containing non-finite values (stat_boxplot).
```



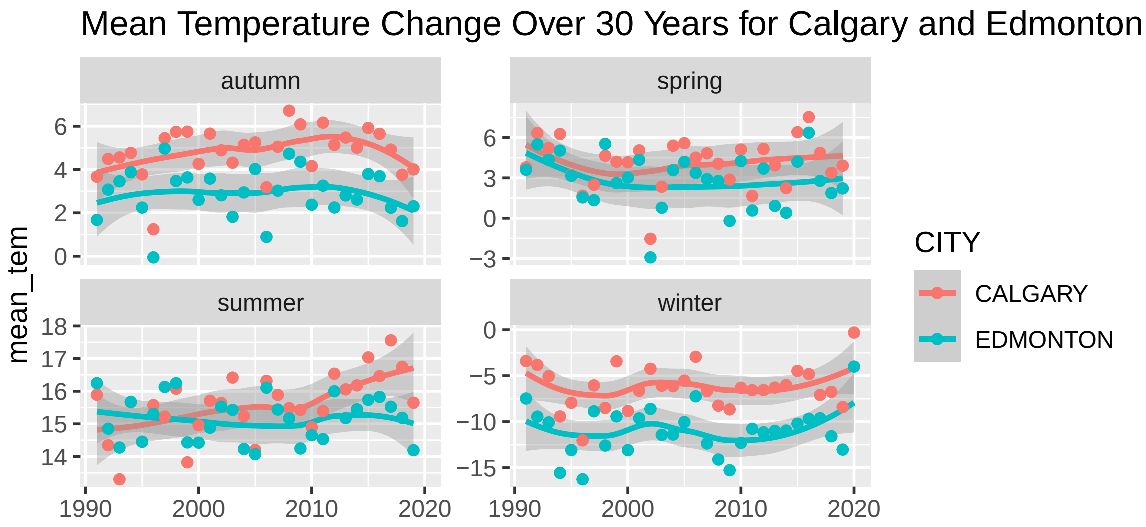
Scatter plot for Temperature change over time

- We will only focus on Temperature change over the last 30 years for both cities, hoping to reveal some agreed trending as global warming.
- We do not analyze Precipitation here is simply we think total precipitation may not be affected much by global warming.

```
#you need to convert the categorical year to numeric year, otherwise you will see
↪  the following error:
# The "Each group consists of only one observation" error message happens because
↪  your x aesthetic is a factor. ggplot takes that to mean that your independent
↪  variable is categorical, which doesn't make sense in conjunction with geom_line.
ggplot(Canadian_climate_history_long_last_30_Cal_Edm %>% filter(VARIABLE ==
↪  "MEAN_TEMPERATURE") %>% mutate(year =as.numeric(format(LOCAL_DATE, "%Y"))) %>%
↪  group_by(CITY,season,year) %>% summarise(mean_tem = mean(Measure, na.rm = T)),
↪  aes(x = year, y = mean_tem, color = CITY)) +
  geom_smooth()+ geom_point()+
  xlab("")+
  facet_wrap(~season, scales = "free_y") + labs(title="Mean Temperature Change Over
↪  30 Years for Calgary and Edmonton")
```

```
## `summarise()` has grouped output by 'CITY', 'season'. You can override using the `.groups` argumen
```

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

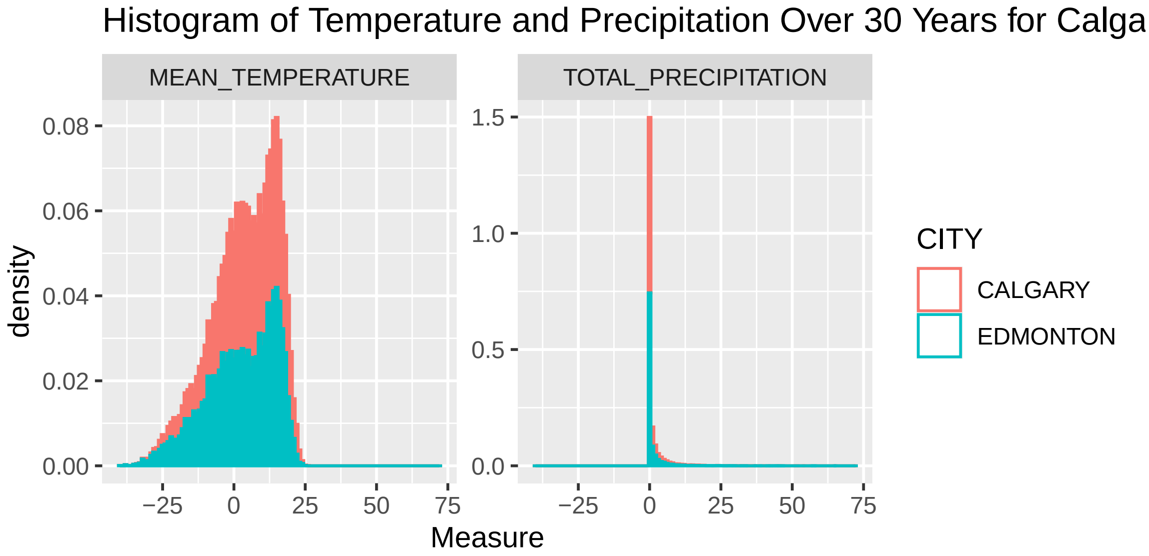


From the above scatter plot, we can already find some trending for temperature change over time. Particularly for Calgary in summer, there is an upward mean-year temperature going in the past 30 years. This is concerning of course, but we need to perform some formal Linear Regression analysis to confirm this conclusion, as shown in the section of Statistical Analysis.

#Statistical Analysis ## Distribution of Temperature and Precipitation - For Calgary and Edmonton - We are looking at the histogram of Temperature and Precipitation for Calgary and Edmonton, so that underlying data distribution can be revealed.

```
#Get the dataset for Calgary and Edmonton only for the last 30 years
Canadian_climate_history_long_Cal_Edm_last_30_noNa <- Canadian_climate_history_long
↪  %>% filter(CITY == "CALGARY" | CITY == "EDMONTON", LOCAL_DATE >= "1991-01-01",
↪  !is.na(Measure))

#Plot out the Histogram of Temperature and Precipitation Over 30 Years for Calgary
↪  and Edmonton
ggplot(Canadian_climate_history_long_Cal_Edm_last_30_noNa, aes(x = Measure, color =
↪  CITY)) +
  geom_histogram(fill = "white", binwidth = 1, aes(y=..density..))+
  facet_wrap(~VARIABLE, scales = "free_y") + labs(title="Histogram of Temperature
↪  and Precipitation Over 30 Years for Calgary and Edmonton")
```



The above plot clearly shows that Temperature follows a normal distribution, while Precipitation does not. This is a the foundation how we are gonna do the following parametric analysis.

Construct a 90% confidence interval for the mean temperature/precipitation by seasons for Calgary.

- The following chunk of code is for temperature CI in Calgary

```
#For Calculating the 90% confidence interval for the mean temperature by season for
↪  Calgary and Edmonton.
calgary_last30_tem <- Canadian_climate_history_long_Cal_Edm_last_30_noNa %>%
↪  filter(VARIABLE == "MEAN_TEMPERATURE", CITY == "CALGARY")
calgary_last30_tem$season <- factor(calgary_last30_tem$season, levels = c("winter",
↪  "spring", "autumn","summer" ))
edmonton_last30_tem <- Canadian_climate_history_long_Cal_Edm_last_30_noNa %>%
↪  filter(VARIABLE == "MEAN_TEMPERATURE", CITY == "EDMONTON")
edmonton_last30_tem$season <- factor(edmonton_last30_tem$season, levels =
↪  c("winter", "spring", "autumn","summer" ))
```

```
#For Calculating the 90% confidence interval for the mean precipitation by season
↪  for Calgary and Edmonton.
calgary_last30_perc <- Canadian_climate_history_long_Cal_Edm_last_30_noNa %>%
↪  filter(VARIABLE == "TOTAL_PRECIPITATION", CITY == "CALGARY")
calgary_last30_perc$season <- factor(calgary_last30_perc$season, levels =
↪  c("winter", "spring", "autumn","summer" ))
edmonton_last30_perc <- Canadian_climate_history_long_Cal_Edm_last_30_noNa %>%
↪  filter(VARIABLE == "TOTAL_PRECIPITATION", CITY == "EDMONTON")
edmonton_last30_perc$season <- factor(edmonton_last30_perc$season, levels =
↪  c("winter", "spring", "autumn","summer" ))

#Write a function for one-sample CI
function_CI <- function(x_bar, za2, sd, n){
  x_bar + za2 * sd/sqrt(n)
}

#We have the following parameters
c_Tem_bar =  tapply(calgary_last30_tem$Measure, calgary_last30_tem$season, mean)
alpha = 0.1
za2 = qnorm(1-alpha/2)
c_Tem_sd = tapply(calgary_last30_tem$Measure, calgary_last30_tem$season, sd)
c_Tem_n = tapply(calgary_last30_tem$Measure, calgary_last30_tem$season, length)
c_Tem_CI = list() #declare a place holder

#Collect the lower and upper interval for 5% level for both Temperature and
↪  precipitation
for (i in 1:4){
  c_Tem_CI[[i]] <- c(function_CI(c_Tem_bar[i], -za2,
↪  c_Tem_sd[i],c_Tem_n[i]),function_CI(c_Tem_bar[i], za2, c_Tem_sd[i], c_Tem_n[i]))
}
```

From the previous calculation, we are 90% confident that the true mean Temperature of Calgary for different seasons are as follows:

```
## # A tibble: 4 x 3
##   Seasons_Temperature Lower_CI Upper_CI
##   <chr>                  <dbl>    <dbl>
## 1 winter                 -6.73    -6.19
## 2 spring                  3.86     4.35
## 3 autumn                  4.58     5.09
## 4 summer                 15.5     15.7
```

We can see Temperature have narrow confidence interval for all different seasons, indicating there is small variability for them.

- Because the Precipitation does not follow normal distribution, we will use bootstrap to calculate the confidence interval.

```
#Calculate the bootstrap Confidence Interval for Mean Precipitation from Calgary.
set.seed(2021)
```

```r
Canadian_climate_history_long_Cal_last_30_noNa <- Canadian_climate_history_long %>%
↪  filter(CITY == "CALGARY" , LOCAL_DATE >= "1991-01-01", !is.na(Measure))

# for spring
calg_precip_spring <- Canadian_climate_history_long_Cal_last_30_noNa %>%
↪  filter(VARIABLE == "TOTAL_PRECIPITATION")  %>% filter(season == "spring")

# for winter
calg_precip_winter <- Canadian_climate_history_long_Cal_last_30_noNa %>%
↪  filter(VARIABLE == "TOTAL_PRECIPITATION")  %>% filter(season == "winter")

# for autumn
calg_precip_autumn <- Canadian_climate_history_long_Cal_last_30_noNa %>%
↪  filter(VARIABLE == "TOTAL_PRECIPITATION")  %>% filter(season == "autumn")

# for summer
calg_precip_summer <- Canadian_climate_history_long_Cal_last_30_noNa %>%
↪  filter(VARIABLE == "TOTAL_PRECIPITATION")  %>% filter(season == "summer")

# define the function that will be used in the bootstrap function
calg_precip_stat = function(d,i){
  d2 = d[i,]
  return(mean(d2$Measure))
}

# call the bootstrap function for spring
calg_precip_spring = boot(data = calg_precip_spring, statistic = calg_precip_stat, R
↪  = 2000)

#call the boot.ci function to find the CI for spring
boot.ci(boot.out = calg_precip_spring, conf = 0.95, type = c("perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = calg_precip_spring, conf = 0.95, type = c("perc"))
##
## Intervals :
## Level     Percentile
## 95%   ( 1.023,  1.289 )
## Calculations and Intervals on Original Scale
```

```r
# for summer
calg_precip_summer = boot(data = calg_precip_summer, statistic = calg_precip_stat, R
↪  = 2000)

#call the boot.ci function to find the CI for summer
boot.ci(boot.out = calg_precip_summer, conf = 0.95, type = c("perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
```

```
##
## CALL :
## boot.ci(boot.out = calg_precip_summer, conf = 0.95, type = c("perc"))
##
## Intervals :
## Level     Percentile
## 95%   ( 2.249,  2.703 )
## Calculations and Intervals on Original Scale
```

```
# for winter
calg_precip_winter = boot(data = calg_precip_winter, statistic = calg_precip_stat, R
↪  = 2000)

#call the boot.ci function to find the CI for winter
boot.ci(boot.out = calg_precip_winter, conf = 0.95, type = c("perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = calg_precip_winter, conf = 0.95, type = c("perc"))
##
## Intervals :
## Level     Percentile
## 95%   ( 0.3649,  0.4558 )
## Calculations and Intervals on Original Scale
```

```
# for autumn
calg_precip_autumn = boot(data = calg_precip_autumn, statistic = calg_precip_stat, R
↪  = 2000)

#call the boot.ci function to find the CI for autumn
boot.ci(boot.out = calg_precip_autumn, conf = 0.95, type = c("perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = calg_precip_autumn, conf = 0.95, type = c("perc"))
##
## Intervals :
## Level     Percentile
## 95%   ( 0.7030,  0.9082 )
## Calculations and Intervals on Original Scale
```

From the previous calculations, we are 95% confident that the true mean precipitation of Calgary for different seasons are as follows:

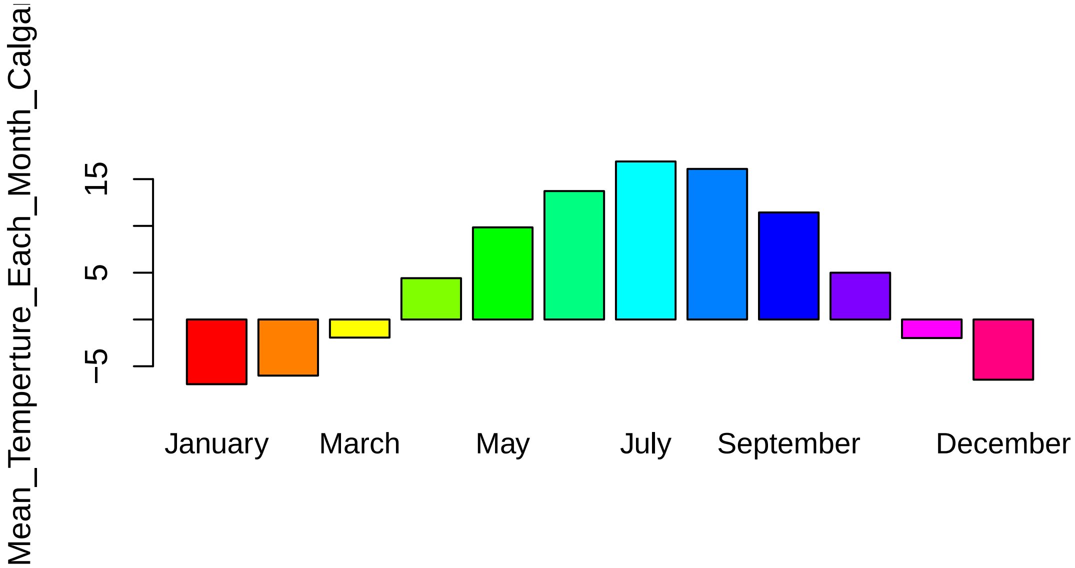
Season	Lower CI	Upper CI
Winter	0.3649	0.4558
Spring	1.023	1.289
Autumn	0.7030	0.9082
Summer	2.249	2.703

What month is likely to be the hottest on average in Calgary? What month is likely to be the coldest on average in Calgary?

- It was reported that the hottest and coldest month are July and January for Calgary as follows:
  - The hottest month of the year is July, with an average of 17°C.
  - The coldest month of the year is January, with an average -6°C.

Barplot to indicate which month is the warmest and coldest in Calgary

```
#What month is likely to be the hottest on average? What month is likely to be the
↪  coldest on average?
calgary_last30_tem_mean_month <- calgary_last30_tem %>% mutate(Month =
↪  format(calgary_last30_tem$LOCAL_DATE, "%B")) %>% group_by(Month) %>%
↪  summarise(Mean_Tem_Month = mean(Measure))
unique_month <- calgary_last30_tem %>% mutate(Month =
↪  format(calgary_last30_tem$LOCAL_DATE, "%B")) %>% select(Month) %>% unique %>%
↪  mutate(Month_ID = 1:12) %>% column_to_rownames(var = "Month")
#add the Month_ID to calgary_last30_tem_mean_month
calgary_last30_tem_mean_month %<>% mutate(Month_ID =
↪  unique_month[calgary_last30_tem_mean_month$Month,]) %>% arrange(Month_ID)
#plot out the mean temperature for each month from Calgary
barplot(calgary_last30_tem_mean_month$Mean_Tem_Month, names.arg =
↪  calgary_last30_tem_mean_month$Month, col = rainbow(12), cex.names=0.9, ylab =
↪  "Mean_Temperture_Each_Month_Calgary")
```



So it is very obvious that **July** is likely to be the hottest month of the year on average, and **January** is likely to be the coldest moth of the year on average.

One sample test for populaiton mean using z test for July warm temperature in Calgary

1. It was reported that the hottest month of the year is July in Calgary, with an average 17°C. I suspect that the average is higher than that, so I selected the past 30 years of July Temperature for each day. I want to test: H0: u = 17 vs. Ha: u > 17 at level $\alpha = 0.05$ This is a large-sample right-tail normal test. Because n>30, we can use the large-sample z-test to approximate the t-test.

```
calgary_last30_tem_July <- calgary_last30_tem %>% mutate(Month =
↪  format(calgary_last30_tem$LOCAL_DATE, "%B")) %>% filter(Month == "July")
calgary_last30_tem_July_measure <- calgary_last30_tem_July$Measure
#Lets create the function to calculate z
z_statistic <- function(x_bar, u, sd, n){
  (x_bar - u)/(sd/sqrt(n))
}
#We have the following parameter:
x_bar = mean(calgary_last30_tem_July_measure)
u0 = 17
sd = sd(calgary_last30_tem_July_measure)
n = length(calgary_last30_tem_July_measure)
alpha = 0.05
#obtain the z-test statistic value
z_cal <- z_statistic(x_bar, u0, sd, n)
#Obtain the P-value for the calculated z value
P_value <- 1- pnorm(z_cal)
P_value
```

```
## [1] 0.8680107
```

Because P value is 0.868 which is way larger than alpha = 0.05, so we fail to reject H0 at 5% level. As such we conclude that we are 95% confident that there is no evidence to show the true mean temperature for July is higher than 17°C.

One sample test for populaiton mean using z test for January cold temperature in Calgary

1. It was reported that the coldest month of the year is January in Calgary, with an average -6°C. From my experience living in Calgary, I suspect that the average is not -6°C, so I selected the past 30 years of January Temperature for each day. I want to test: H0: u = -6 vs. Ha: u $\neq$ 17 at level $\alpha = 0.05$ This is a large-sample two-tail normal test. Because n>30, we can use the large-sample z-test to approximate the t-test.

```
calgary_last30_tem_January <- calgary_last30_tem %>% mutate(Month =
↪  format(calgary_last30_tem$LOCAL_DATE, "%B")) %>% filter(Month == "January")
calgary_last30_tem_January_measure <- calgary_last30_tem_January$Measure
#Lets create the function to calculate z
z_statistic <- function(x_bar, u, sd, n){
  (x_bar - u)/(sd/sqrt(n))
}
#We have the following parameter:
x_bar = mean(calgary_last30_tem_January_measure)
u0 = -6
sd = sd(calgary_last30_tem_January_measure)
n = length(calgary_last30_tem_January_measure)
alpha = 0.05/2
z_value = qnorm(alpha)
#obtain the z-test statistic value
z_cal <- z_statistic(x_bar, u0, sd, n)
```

Because the calculated z = -3.0386155, which is outside -1.959964 and -1.959964, so we reject H0 at 5% level. As such we conclude that we are 95% confident that the average temperature for January is not -6°C. *NEED TO EDIT THIS CONCLUSION*

Construct a 90% confidence interval for the proportion of days below 0°C for Calgary vs. the proportion of days below 0°C for Edmonton.

This is to answer do we have more days with warm temperature (>0) in Calgary compared to Edmonton. We want to test H0: p_hat_cal - p_hat_edm = 0 vs. p_hat_cal - p_hat_edm > 0

```
#I am using the last 30 years Temperature data in Calgary to do the test.
calgary_last30_tem_measure <- calgary_last30_tem$Measure
edmonton_last30_tem_measure <- edmonton_last30_tem$Measure
x_cal_warm <- sum(calgary_last30_tem_measure>0)
n_cal_warm <- length(calgary_last30_tem_measure)
x_edm_warm <- sum(edmonton_last30_tem_measure>0)
n_edm_warm <- length(edmonton_last30_tem_measure)
p_hat_cal <- x_cal_warm/n_cal_warm
p_hat_edm <- x_edm_warm/n_edm_warm
alpha = 0.05
za2 <- qnorm(1-alpha)
lower_CI <- (p_hat_cal - p_hat_edm) - za2*sqrt(p_hat_cal*(1-p_hat_cal)/n_cal_warm +
↪   p_hat_edm*(1-p_hat_edm)/n_edm_warm)
upper_CI <- (p_hat_cal - p_hat_edm) + za2*sqrt(p_hat_cal*(1-p_hat_cal)/n_cal_warm +
↪   p_hat_edm*(1-p_hat_edm)/n_edm_warm)

cat("At 5% level, the confidence interval is between (", lower_CI, ",", upper_CI,
↪   "). Because 0 is not contained inside actually smaller than this interval, we
↪   reject H0. We can conclude that we are 95% confident that Calgary has more warm
↪   days compared to Edmonton." )
```

```
## At 5% level, the confidence interval is between ( 0.07575521 , 0.09737276 ). Because 0 is not cont
```

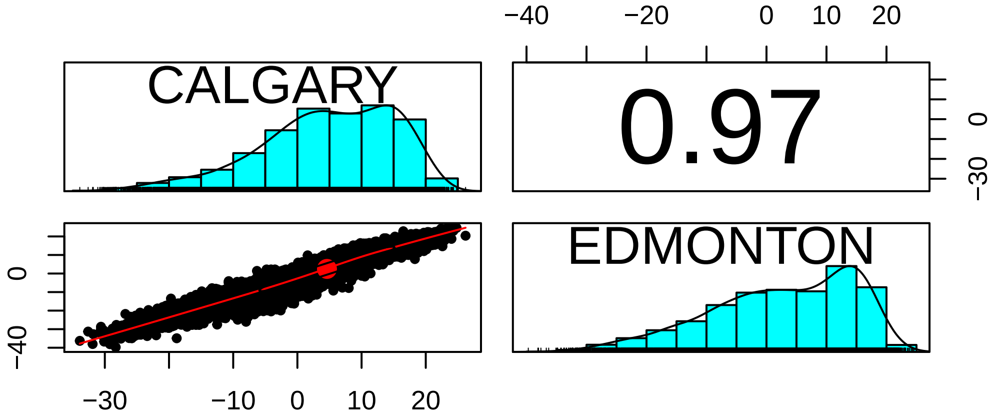
If there is a correlation for the Temperature between Calgary and Edmonton?

- Because Calgary and Edmonton are geographically closed to each other, they are theoretically quite correlated in terms of weather. As such, we want to explore the strength of this association so that data in one location can be used to to predict the other if somehow data is hard to achieve for one particular location.

```
#Make a scatter plot and add the fitted regression line to the plot
Canadian_climate_history_wide_Cal_Edm_last_30_noNa_tem <-
↪   Canadian_climate_history_long_Cal_Edm_last_30_noNa %>% filter(VARIABLE ==
↪   "MEAN_TEMPERATURE") %>% dcast(LOCAL_DATE+season+VARIABLE~CITY)
```

```
## Using Measure as value column: use value.var to override.
```

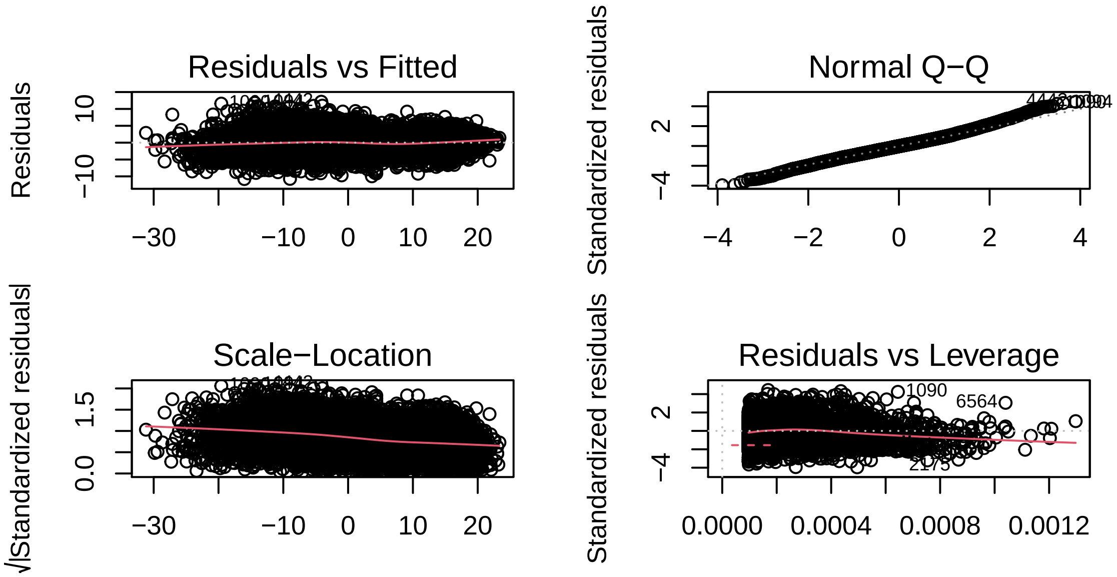
```
pairs.panels(Canadian_climate_history_wide_Cal_Edm_last_30_noNa_tem %>%
↪   dplyr::select(CALGARY, EDMONTON))
```



```
#Assess statistically the adequacy of the fit of the linear model using appropriate
↪  statistics and model diagnostics.
#Use EDMONTON to predict CALGARY
cal_edm_tem_fit <- lm(CALGARY~EDMONTON, data =
↪  Canadian_climate_history_wide_Cal_Edm_last_30_noNa_tem)
summary(cal_edm_tem_fit)
```

```
##
## Call:
## lm(formula = CALGARY ~ EDMONTON, data = Canadian_climate_history_wide_Cal_Edm_last_30_noNa_tem)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7909  -1.7707  -0.0567   1.6705  12.1331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.483489   0.027425   90.56   <2e-16 ***
## EDMONTON    0.848284   0.002238  379.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.728 on 10333 degrees of freedom
##   (233 observations deleted due to missingness)
## Multiple R-squared:  0.9329, Adjusted R-squared:  0.9329
## F-statistic: 1.437e+05 on 1 and 10333 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
par(mar = c(3,4,3,2.1))
plot(cal_edm_tem_fit)
```



From the linear regression, we can see that the overall F test has a Pvalue < 2e-16. The individual t tests (Pvalue for Intercept is <2e-16, Pvalue for slope is <2e-16) all suggest that the fitted model is highy statistically significant. The coefficient of determination is $R^2 = 0.933$.

From the model dignostics, we dont see things very abnormal. Both Residuals vs Fitted and Scale-Location show linearity and equal variance assumption met because the points are evenly distributed in the rectangular region. qqplot shows a near 45 degree line indicating a good normality assumption. There are only a few influential data points. Since we already get a good R2, I dont there is a need to re-test the model without these influential data points.

I highly recommend using model for predicting Calgary Temperature using Edmonton Temperature, because slope is highly statistically significant. And the R2 is above 0.9, which indicates a good linear regression. Besides, all the assumptions are met from the model dignostics.

Create the bootstrap distribution for mean_calgary_temp- mean_edmonton_temp.

- Create the bootstrap distribution for mean_calgary_temp- mean_edmonton_temp. Construct a 95% confidence interval. What can you infer from the result?
- I want to test if Calgary on leverage has different temperature than Edmonton all year round.

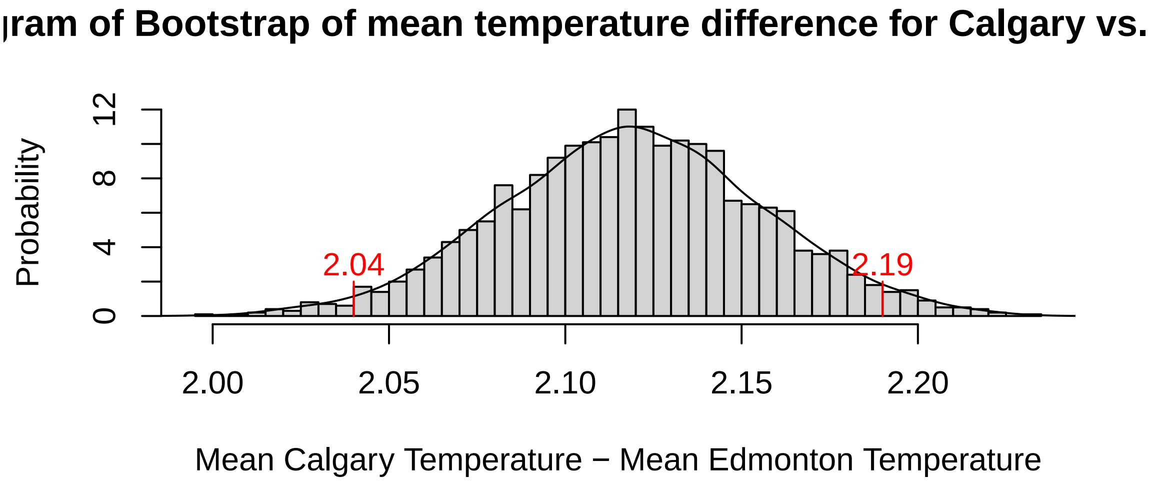
Let u1 and u2 to be the mean temperature for Calgary and Edmonton respectively. We are testing H0: u1 - u2 = 0 vs. u1 - u2 != 0. This is a two-sample two tail test. We reject H0 if 0 is contained within the confidence interval.

```
#We use boot to do the hypothesis test.
tem_diff <- function(d,i){
  d2 = d[i,]
  return(mean(d2$CALGARY, na.rm = T) - mean(d2$EDMONTON, na.rm = T))
}
set.seed(2021)
boot_tem_cal_edm <- boot(data =
↪   Canadian_climate_history_wide_Cal_Edm_last_30_noNa_tem, statistic = tem_diff, R
↪   = 2000)
boot.ci(boot_tem_cal_edm, conf = 0.95, type = "perc") #NOTE you cannot use bca
↪   interval because the sample size is too big for R=2000
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_tem_cal_edm, conf = 0.95, type = "perc")
##
## Intervals :
## Level     Percentile
## 95%   ( 2.045,  2.192 )
## Calculations and Intervals on Original Scale
```

At 95% confidence level, the percentile interval is (2.04, 2.19). Because u1-u2=0 is not contained inside this 95% CI , we reject H0. As such we conclude that we are 95% confident the Calgary on leverage has differernce temperature than Edmonton all year round.

```
hist(boot_tem_cal_edm$t, nclass = 40, probability = T, xlab = "Mean Calgary
↪  Temperature - Mean Edmonton Temperature",
     ylab = "Probability", main = "Histogram of Bootstrap of mean temperature
      ↪  difference for Calgary vs. Edmonton")
lines(density(boot_tem_cal_edm$t)) #add the density line
text(2.04, 3, col="red", 2.04) #I am ploting the perc interval
text(2.19, 3, col="red", 2.19)
lines(c(2.04,2.04), c(0,2), col="red")
lines(c(2.19,2.19), c(0,2), col="red")
```



Do Calgary and Edmonton see the same variance in temperature/precipitation?

This is to test the hypothesis of equal variances between Calgary and Edmonton for temperature/precipitation

– Null hypothesis ($\mathbf{H_0}$) = σ\{calgary, temp\}ˆ2 = σ\{edmonton, temp\}ˆ2

– Alternative hypothesis ($\mathbf{H_A}$) = σ\{calgary, temp\}ˆ2 $\neq$ σ\{edmonton, temp\}ˆ2

```
# city_stats <-  Canadian_climate_history_last30_long %>% mutate(year =
↪  format(LOCAL_DATE, "%Y")) %>% group_by(city, measures) %>%
↪  summarise(mean_city_val = mean(value, na.rm = T), sd_city_val = sd(value,
↪  na.rm), n_city_val = length(value))
```

```
#
# # for temperature
# calg_temp_stats = as_tibble(city_stats) %>% filter(city == "CALGARY") %>%
↪  filter(measures == "MEAN_TEMPERATURE" )
# edm_temp_stats = as_tibble(city_stats) %>% filter(city == "EDMONTON") %>%
↪  filter(measures == "MEAN_TEMPERATURE" )
#
# # define parameters for both cities
# calg_temp_mean = calg_temp_stats$mean_city_val
# calg_temp_s = calg_temp_stats$sd_city_val
# calg_temp_n = calg_temp_stats$n_city_val
#
# edm_temp_mean = edm_temp_stats$mean_city_val
# edm_temp_s = edm_temp_stats$sd_city_val
# edm_temp_n = edm_temp_stats$n_city_val
#
# # find F
# F_temp = calg_temp_s^2/edm_temp_s^2
#
# F_alpha = 0.05
#
# df_calg = calg_temp_n - 1
# df_edm = edm_temp_n - 1
#
# # find critical value range
# F_alpha_temp = qf(1-alpha/2, df_calg, df_edm)
# F_1minusalpha_temp = qf(alpha/2, df_calg, df_edm)
#
# print(F_temp)
# print(F_alpha_temp)
# print(F_1minusalpha_temp)
```

Since F (0.7540) is not within the critical value range (0.9498, 1.0529), we reject H0 at alpha = 0.05. We are 95% confident that the true σ{calgary, temp}^2 $\neq$ σ{edmonton, temp}^2.

For precipitation:

– Null hypothesis ($\mathbf{H_0}$) = σ{calgary, precipitation}^2 = σ{edmonton, precipitation}^2

– Alternative hypothesis ($\mathbf{H_A}$) = σ{calgary, precipitation}^2 $\neq$ σ{edmonton, precipitation}^2

```
# # for precipitation
#
# calg_precip_stats = as_tibble(city_stats) %>% filter(city == "CALGARY") %>%
↪  filter(measures == "MEAN_PRECIPITATION" )
# edm_precip_stats = as_tibble(city_stats) %>% filter(city == "EDMONTON") %>%
↪  filter(measures == "MEAN_PRECIPITATION" )
#
# # define parameters for both cities
# calg_precip_mean = calg_precip_stats$mean_city_val
# calg_precip_s = calg_precip_stats$sd_city_val
# calg_precip_n = calg_precip_stats$n_city_val
#
```

```
# edm_precip_mean = edm_precip_stats$mean_city_val
# edm_precip_s = edm_precip_stats$sd_city_val
# edm_precip_n = edm_precip_stats$n_city_val
#
# # find F
# F_precip= calg_precip_s^2/edm_precip_s^2
#
# alpha = 0.05
#
# df_calg_precip = calg_precip_n - 1
# df_edm_precip = edm_precip_n - 1
#
# # find critical value range
# F_alpha_precip = qf(1-alpha/2, df_calg_precip, df_edm_precip)
# F_1minusalpha_precip = qf(alpha/2, df_calg_precip, df_edm_precip)
#
# print(F_precip)
# print(F_alpha_precip)
# print(F_1minusalpha_precip)
```

Since F (1.1960) is not within the critical value range (0.9497, 1.053), we reject H0 at alpha = 0.05. We are 95% confident that the true σ{calgary, precipitation}ˆ2 $\neq$ σ{edmonton, precipitation}ˆ2.