BUDA 530 Assignment 4

Collin Edwards

2025-02-16

You work for an Ag-Tech consulting firm and were recently promoted when your former boss (Josh) left for another opportunity. You have taken over all of your boss's clients.

One of these clients is a wheat farmer in Manitoba, Canada. Each year, your team prepares this client a forecast of the number of Growing Degree Days (GDD) per month for the next year. GDD are a measure of the amount of warmth that plants have experienced over the past time calculated by summing the average daily temperature for days where the avg temp is above a certain threshold (for wheat this is 0 degrees C). Your client uses this forecast to plan when they plant and harvest their crops. More Information About GDD.

You've had a junior member of your team pull the past 30 years of GDD data from Gretna, Canada for you using the following code (Note: this code will only run on GoFirst; A CSV version is included for your reference).

```
library(tidyverse) #loading tidyverse to use dplyr functions
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# install.packages("sparklyr") #installing sparklyr package
```

```
library(readr)
GretnaGDD <- read_csv("~/Downloads/GretnaGDD.csv")
```

```
## Rows: 349 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): MONTH
## dbl (2): YEAR, GDD
##
## 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.
```

```
View(GretnaGDD)#I'm just peeking at the dataframe so I can see what I'm working with
summary(GretnaGDD) #I'm looking at the summary stats to see if there are any missing values
```

```
##       YEAR         MONTH                GDD        
##  Min.   :1995   Length:349         Min.   :  0.00  
##  1st Qu.:2002   Class :character   1st Qu.:  3.95  
##  Median :2009   Mode  :character   Median :153.05  
##  Mean   :2009                      Mean   :232.81  
##  3rd Qu.:2016                      3rd Qu.:457.00  
##  Max.   :2024                      Max.   :679.65  
```

```
{r dataPull, eval=FALSE} # # Connect to Spark # library(sparklyr)
# sc <- spark_connect(master = "local") #  # # Pull data for
station CA005021220 # CA005021220<-spark_read_csv(sc,"CA005021220_spark"
#"s3a://noaa-ghcn-pds/csv/by_station/CA005021220.csv") #  # #
summarize # GretnaGDD <- CA005021220%>% #   filter(ELEMENT=="TMAX")%>%
# get max temp #   select(DATE,TMAX=DATA_VALUE)%>% #   left_join(CA00502
# add in min temp #     filter(ELEMENT=="TMIN")%>% #     select(DATE,TMI
by="DATE")%>% #   mutate(TMINdeg=TMIN/10,TMAXdeg=TMAX/10)%>% #
adjust units from 10th of  # degrees to degrees #   mutate(YEAR=substr(D
# create year and  # month variables #   filter(YEAR>1994)%>%
# filter to past 30 years #   select(-TMIN,-TMAX)%>% # remove
10th of degree variables #   mutate(TAVGdeg=(TMINdeg+TMAXdeg)/2)%>%
# calculate average temp #   mutate(IsGDD = ifelse(TAVGdeg>0,1,0))%>%
# check if average temp is above 0 #   mutate(GDD=IsGDD*TAVGdeg)%>%
# calculate daily GDD #   group_by(YEAR,MONTH)%>%  #   summarise(GDD=sum
# aggregate by month and year #   arrange(YEAR,MONTH)%>% #
sort data #   collect() # export data to R #  # # disconnect
from Spark # spark_disconnect(sc) #
```

Documentation for this data set:

AWS Listing

AWS README

GitHub README

Josh always insisted that you use a 30-year SMA forecast by month (for example, to get next March's forecast, you would take the average of the past 30 March data points). The client has been complaining for years, however, that the data is inaccurate in very specific ways (always high for certain months and low for others). You suspect that is because of a trend associated with climate change.

Before Josh left, you had raised this idea to him and his solution was to switch to a 15-year SMA forecast. Both the 30-year and 15-year are included below.

```
library(readr)
Gretna_GDD<- read_csv("GretnaGDD.csv")
```

```
## Rows: 349 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): MONTH
## dbl (2): YEAR, GDD
##
## 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.
```

```
Gretna_GDD%>%group_by(MONTH)%>%
  summarise(GDDAvg30 = mean(GDD))%>% # get 30-year average by month
  left_join(Gretna_GDD%>% # join in 15-year average
              filter(YEAR>2008)%>% # filter to past 15 years
              group_by(MONTH)%>%
              summarise(GDDAvg15 = mean(GDD)), by = "MONTH") # get 15-year average by month
```

```
## # A tibble: 12 x 3
##    MONTH GDDAvg30 GDDAvg15
##    <chr>    <dbl>    <dbl>
##  1 01       0.855    0.741
##  2 02       2.01     1.04
##  3 03      26.7     31.9
##  4 04     139.     127.
##  5 05     346.     372.
##  6 06     514.     534.
##  7 07     600.     610.
##  8 08     555.     575.
##  9 09     405.     432.
## 10 10     179.     183.
## 11 11      32.6     36.2
## 12 12       1.74     2.06
```

You have long thought that Exponential Smoothing or ARIMA would be better methodologies to use, but Josh always said, “There is no way we can explain that to the client. Stick to tried and true methods.”

Now that you are in charge, you want to evaluate whether or not these methods would be better.

1. Convert the dataframe `GretnaGDD` to a time-series object using the `ts` function.

```
# convert year and month to numeric
Gretna_GDD <- Gretna_GDD %>%
  mutate(year = as.numeric(YEAR),
         month = as.numeric(MONTH))
str(Gretna_GDD) #me just checking to make sure the year and month are numeric. turns out month was a ch
```

```
## tibble [349 x 5] (S3: tbl_df/tbl/data.frame)
##  $ YEAR : num [1:349] 1995 1995 1995 1995 1995 ...
##  $ MONTH: chr [1:349] "01" "02" "03" "04" ...
##  $ GDD  : num [1:349] 0 0 13.4 54.5 263.9 ...
##  $ year : num [1:349] 1995 1995 1995 1995 1995 ...
##  $ month: num [1:349] 1 2 3 4 5 6 7 8 9 10 ...
```

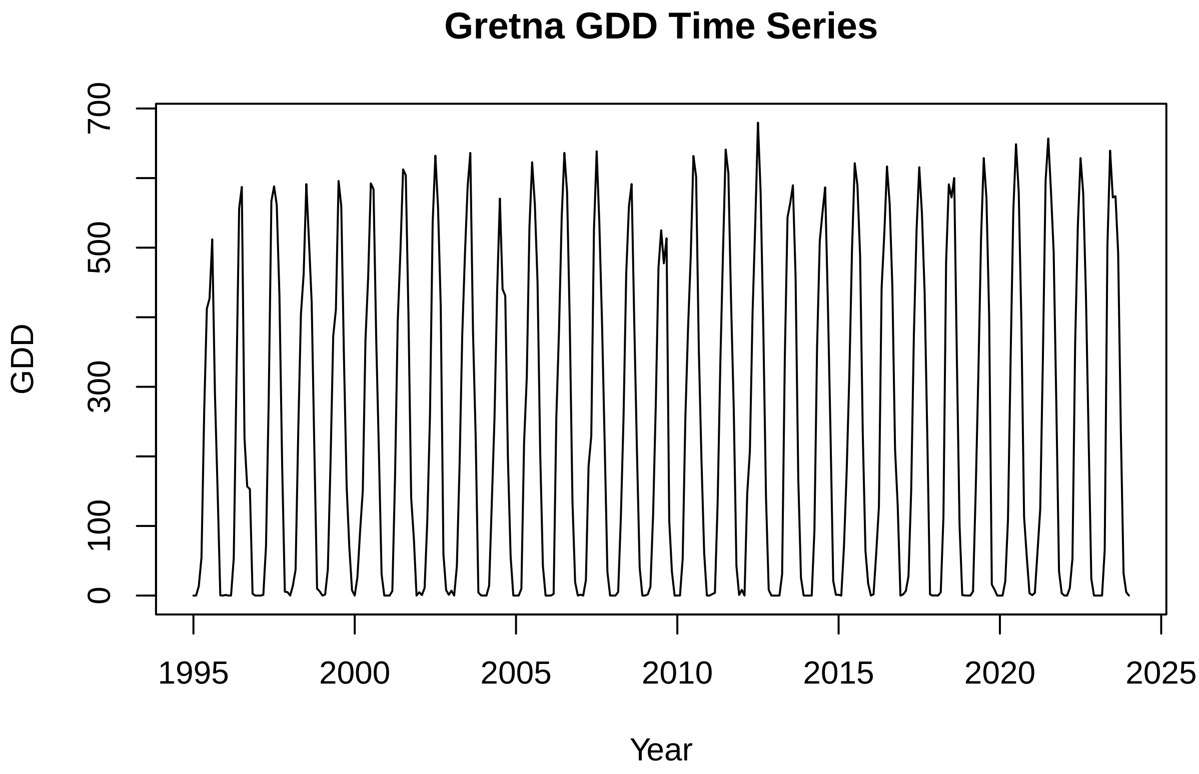
```
# determining the start year and month
start_year <- min(Gretna_GDD$year)
start_month <- min(Gretna_GDD$month[Gretna_GDD$year == start_year])

# create the time series object with a frequency of 12 (monthly data)
gretna_gdd_ts <- ts(Gretna_GDD$GDD, start = c(start_year, start_month), frequency = 12)
```

Problem 1 Answer I have converted the dataframe `GretnaGDD` to a time-series object using the `ts` function. The reason why I used the **mutate** function is to convert the year and month to numeric if they are not already. I then determined the start year and month by finding the minimum year and month in the dataset. Finally, I created the time series object with a frequency of 12 (monthly data).

2. Plot the time series using the `plot.ts` function.

```
# plot the time series
plot.ts(gretna_gdd_ts, main = "Gretna GDD Time Series", xlab = "Year", ylab = "GDD")
```



Problem 2 Answer I have plotted the time series using the `plot.ts` function. The plot shows the Gretna GDD time series with the year on the x-axis and GDD on the y-axis. The plot provides a visual representation of the GDD data over time.

3. Fit an Exponential Smoothing model to the data using the `ets` function.

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
# fitting an Exponential Smoothing model
ets_model <- ets(gretna_gdd_ts) # not specifying the model type to let the function choose the best mod
summary(ets_model)
```

```
## ETS(A,N,A)
##
## Call:
## ets(y = gretna_gdd_ts)
##
##   Smoothing parameters:
##     alpha = 0.0881
##     gamma = 1e-04
##
##   Initial states:
##     l = 209.0577
##     s = -230.6588 -202.3092 -54.6406 170.6303 321.2735 366.3611
##            279.2602 112.7139 -93.6458 -205.0164 -231.3861 -232.5821
##
##   sigma:  48.2849
##
##      AIC     AICc      BIC
## 4765.360 4766.802 4823.186
##
## Training set error measures:
##                     ME     RMSE      MAE MPE MAPE      MASE      ACF1
## Training set 1.483018 47.30651 33.18622 NaN  Inf 0.8530739 0.2009029
```

Problem 3 Answer I have fitted an Exponential Smoothing model to the data using the ets function. The model summary provides information about the model type, parameters, and performance metrics. The ets function automatically selects the best model based on the data. Since I didnt specify the model type, the function chose the best model for the data. It detected that it was an additive error, no trend and an additive seasonality with a frequency of 12.

4. Fit an ARIMA model to the data using the auto.arima function.

```
# fitting an ARIMA model to the time series
arima_model <- auto.arima(gretna_gdd_ts)
summary(arima_model)
```

```
## Series: gretna_gdd_ts
## ARIMA(0,0,1)(2,1,2)[12]
##
## Coefficients:
##          ma1    sar1     sar2     sma1    sma2
##       0.2746  0.7525  -0.1179  -1.5601  0.6359
## s.e.  0.0557  0.3143   0.0779   0.3105  0.2581
##
## sigma^2 = 2307:  log likelihood = -1787.87
## AIC=3587.75   AICc=3588   BIC=3610.67
##
## Training set error measures:
##                   ME     RMSE    MAE MPE MAPE      MASE        ACF1
## Training set 6.96577 46.85064 30.716 NaN  Inf 0.7895753 -0.02329667
```

Problem 4 Answer I have fitted an ARIMA model to the data using the `auto.arima` function. The `auto.arima` function automatically selects the best ARIMA model based on the data. The model summary provides information about the model type, parameters, and performance metrics. The function chose the best ARIMA model for the data. This shows strong seasonality since the SAR and the SMA2 parameters are close to 1. This ARIMA captures both the short-term moving average and seasonsal depencency and seems to fit reasonable well.

5. Forecast both models out 12 months.

```r
# forecast the ets model 12 months ahead
ets_forecast <- forecast(ets_model, c = 12)

# view summary and plot the forecast
summary(ets_forecast)
plot(ets_forecast, main = "ets forecast for gretna gdd", xlab = "Year", ylab = "GDD")
```

ETS Forecast

```r
# forecast the arima model 12 months ahead
arima_forecast <- forecast(arima_model, c = 12)

# view summary and plot the forecast
summary(arima_forecast)
plot(arima_forecast, main = "arima forecast for gretna gdd", xlab = "Year", ylab = "GDD")
```

ARIMA Forecast

Problem 5 Answer I have forecasted both the ETS and ARIMA models out 12 months. The forecasts provide predictions for the Gretna GDD time series for the next 12 months. The summary of the forecasts includes the point forecasts, prediction intervals, and other relevant information. The plots show the forecasted values along with the prediction intervals for the next 12 months. Both models provide valuable insights into the future GDD values and capture the seasonality and trends in the data and they have strong yearly cycles in the data. Both models in this case are useful for forecasting the GDD values.

6. Since taking over for Josh, you have noticed he was very poor at documenting processes. Had you not worked for him previously, you would have to completely reinvent the wheel. You don't want that to be the case for whoever takes over your position when you get promoted. Create documentation for this analysis. Explain your motivation, methodology, and results. The audience of this documentation is whoever takes over your position in the future.

Documentation for Gretna GDD Analysis **Motivation:** The goal of this analysis is to improve upon the traditional 30-year `sma` forecast for growing degree days (gdd) that our client has been using. the client has noted that the `sma` method consistently produces forecasts that are too high in some months and too low in others. Our hypothesis is that a trend associated with climate change is affecting the gdd patterns, and methods like exponential smoothing or arima might better capture these changes.

Methodology: 1. *Data Preparation*: The data for the past 30 years of gdd values for Gretna, Canada was pulled and converted into a time series object using the `ts` function. I also converted the historical data into a time series object to facilitate time series analysis and forecasting. 2. *Time Series Plot*: The time series plot was created using the `plot.ts` function to visualize the gdd data over time. This plot provides an overview of the gdd values and helps identify any trends or patterns in the data. 3. *Exponential Smoothing Model*: An exponential smoothing model was fitted to the data using the `ets` function to capture any underlying trends and seasonality. The model type was not specified to let the function choose the best model based on the data. 4. *ARIMA Model*: An ARIMA model was fitted to the data using the `auto.arima` function to capture the autocorrelation and seasonality in the data. The function automatically selected the best ARIMA model based on the data. 5. *Forecasting*: Both the ETS and ARIMA models were used to forecast the gdd values for the next 12 months.

Results:

- The ETS model detected an additive error, no trend, and an additive seasonality with a frequency of 12. The ARIMA model captured both the short-term moving average and seasonal dependency in the data. The arima model, while being slightly more complex than ETS provided reasonable forecasts. Both models were able to capture the seasonality and trends in the data and provided valuable insights into the future gdd values. They also offer advantages over traditional 30-year `sma` forecasts because they can capture more complex patterns in the data and capture the seasonality and trends in the data.

7. Pick a forecasting methodology to present to the client and explain your decision. As above, the audience is whoever takes over your position in the future. You should weigh the pros and cons of each approach (30-year, 15-year, ETS, and ARIMA) including a cost/ benefits analysis of what information each methodology adds vs how easy it is to explain to the client. Note: there is no right or wrong answer here, you will be graded on how you explain your decision, not on the decision itself.

Forecasting Methodology Recommendation **Methodology: Exponential Smoothing (ETS)**

Pros: - ETS models are easy to understand and interpret, making them suitable for clients with limited statistical knowledge. - ETS models can capture underlying trends and seasonality in the data, which may help improve forecast accuracy. - ETS models are computationally efficient and can be implemented quickly for forecasting purposes. - ETS models provide point forecasts and prediction intervals, which can help clients make informed decisions.

Cons: - ETS models may not capture complex patterns in the data as effectively as ARIMA models. - ETS models may not perform well with highly volatile or irregular data patterns. - ETS models may require more manual intervention for parameter tuning compared to ARIMA models.

Decision Rationale: *I recommend using the Exponential Smoothing (ETS) methodology for forecasting the Gretna GDD values.* ETS models are easy to understand and interpret, making them suitable for clients with limited statistical knowledge. The ETS model can capture underlying trends and seasonality in the data, which may help improve forecast accuracy. Additionally, ETS models are computationally efficient and can be implemented quickly for forecasting purposes. While ARIMA models may capture more complex patterns in the data, the ETS model is more straightforward to explain to the client and provides valuable insights into the future GDD values. The ETS model strikes a balance between accuracy and simplicity, making it a suitable choice for our client's forecasting needs. **Given that our client values clarity and familiarity with the sma approach**, the ets model strikes a good balance. it improves forecast accuracy by capturing trends while still providing clear confidence intervals, and its forecasts can be explained in terms of weighted averages of past data.

8. Regardless of your answer for the previous question, now suppose you decided to move forward with Exponential Smoothing. Write a memo for your client explaining this switch. Your client is a farmer; they have some college education and understand agriculture extremely well but stats isn't their forte. Despite your misgivings about Josh's methodology, you do know that he was correct about the client liking that they understood SMA. You know there will be some resistance to switching to this new methodology. Keep this in mind as you explain. Also explain the confidence intervals and how they can help the client with decision making.

Memo to Client: Forecasting Methodology Switch to Exponential Smoothing (ETS) Subject: Improved Forecasting Methodology for Gretna GDD

Dear Client,

After reviewing our historical data and forecasting methods, we have decided to update our approach for predicting growing degree days (gdd). in the past, we used a simple moving average (sma) based on 30 years of data. While this method is easy to understand, it does not fully capture recent shifts in climate trends.

We are now switching to an exponential smoothing model(ESM). this method gives more weight to recent observations, allowing us to better capture trends such as gradual increases in temperature. One advantage of this approach is that it also produces confidence intervals. These intervals provide a range in which the future gdd values are likely to fall, giving you a sense of the uncertainty in our forecasts. This additional information can be valuable in planning your planting and harvesting schedules.

We understand that you are familiar with the sma method, and we assure you that the ESM is simply a more refined version that takes recent trends into account. If you have any questions or would like further explanation, please feel free to reach out.

9. In the previous question do you think a memo is the best way to explain this to your client? If yes, explain why. If no, explain what other methodology you would use (phone call, in-person presentation, etc.) and why.

Problem 9 Answer I believe that a memo is a good initial method for explaining the switch because it provides a written record that the client can refer to later. however, given that the client values clear, straightforward explanations, it may be beneficial to follow up the memo with a short in-person meeting or a phone call. This way, we can answer any questions immediately and ensure that the client fully understands the new methodology and its benefits.

10. Your client was impressed with your analysis and has started referring you to their friends. One of those friends - a corn farmer in Omaha, Nebraska, USA - has hired you to forecast GDD for them. You have tasked a junior analyst with pulling a report for you in Spark (using the code above as a reference) and they have sent you the following email:

"Hey Boss,

I started trying to change Josh's old Spark code to pull the GDD report for Omaha and ran into some issues I could use your help with:

1. I looked it up and the GDD threshold for corn is 10 degrees C, not 0. Do you know where I should change the code for this?
2. I looked up the stations for Omaha in that document you told me about Stations Document. There are a ton of options... I tried a couple and so far I'm either not getting back any results or not getting a full 30 years worth of data. Any idea what I should do about this?

Thanks, Junior"

You weren't around when Josh wrote this code for Manitoba, but you do remember him talking about it. Specifically you remember him mentioning that (1) not all stations collect all data types (for instance some only collect rain), (2) stations are created and shut down all the time but are all still in the data bucket (i.e. there may be station options that haven't existed very long, or that have shut down years ago), (3) that he actually had to use a station close to the Manitoba client but not in the exact same city to get the data he needed, (3) that there was a period of time he actually had to use data from 2 stations to get the data he needed, and (4) that he wrote a loop to produce a summary table for each station in the area and then looked at the results to make a decision on what station to use. Given this information, write a response to the junior analyst's email (you do not need to write or run any code for this). Include links to resources you think may be helpful.

Response to Junior Analyst Hi Junior,

I'm glad to hear you're making progress on the GDD report for Omaha. Here are some suggestions to address the issues you've encountered:

1. **GDD Threshold for Corn**: To change the GDD threshold for corn to 10 degrees C, you will need to update the code where the threshold is defined. Look for the part of the code that calculates the GDD based on the average temperature above a certain threshold (currently set at 0 degrees C). You will need to modify this threshold to 10 degrees C to align with the requirements for corn.

2. **Selecting the Right Station**: When selecting a station for Omaha, refer to the Stations Document to identify the available options. Keep in mind that not all stations collect the same data types, and some may not have a full 30 years of data. Since stations are created and shut down frequently, you may need to explore multiple stations to find the one that meets your criteria. Consider using a loop to summarize the data from each station and compare the results to make an informed decision on which station to use. You may also need to use data from multiple stations to get the full 30 years of data if necessary.

If you encounter any further issues or need additional guidance, feel free to reach out. Keep up the good work!

Best,

Collin Edwards

11. Outside of this hypothetical case study, this particular problem matters. The US Climate Normals is essentially a 30-year SMA by day released by NOAA every year. It is used in a wide range of industries (industrial planning, HVAC, etc.). Recently, due to climate change, NOAA has had to start releasing a 15-year version of the climate normals. Read the article Why are the new climate normals abnormal?. Answering now as yourself (i.e. a citizen of the world) what are your thoughts about this? Do you think NOAA should investigate using methods that pick up on trend for climate normals (i.e. Exponential Smoothing)? Keep in mind, NOAA does not currently represent this data set as a "forecast" but rather as a benchmark of typical climate conditions. If they were going to use one of these methodologies, should they consider branding it differently? How should they handle imputation of missing values (this is fairly easy for SMA but not so easy for ES)?

Problem 11 Answer The recent switch by NOAA from a 30-year to a 15-year SMA for climate normals reflects the increasing influence of climate change on historical weather patterns. While the SMA method is straightforward and familiar, it does not capture underlying trends. I believe that NOAA should consider methods that account for trend—such as exponential smoothing—especially when producing benchmarks that reflect current conditions.

If NOAA were to adopt a methodology like exponential smoothing, it would be important to brand these new metrics differently. For example, instead of calling them “climate normals,” which implies a historical average, they could be termed “adjusted climate benchmarks” or “trend-adjusted climate normals” to emphasize that they incorporate recent trends.

Handling missing values is another important consideration. While SMA easily handles missing data by averaging available observations, exponential smoothing requires more care. Methods such as interpolation or model-based imputation should be used to fill gaps before fitting an exponential smoothing model.

Overall, while a more advanced method may add complexity, the benefit of capturing recent climate trends could provide more accurate and useful benchmarks for industries relying on this data. NOAA should consider the trade-offs between simplicity and accuracy when updating their climate normals methodology.