Lome Report

Allison Zembrodt and Rachel Driehaus

2/15/2018

Previous Report

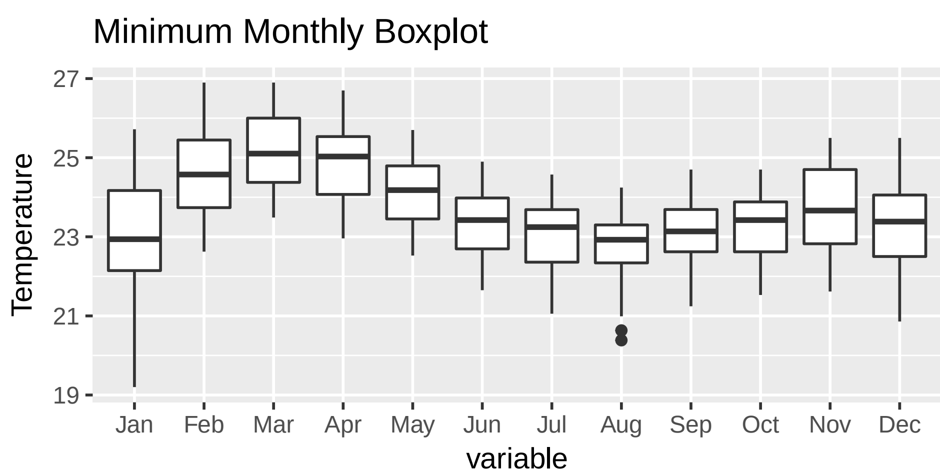
In the previous report, the group stated that both the maximum and the minimum yearly temperatures were best modeled by a simple linear regression. With the new monthly data set, we would like to test and see if we can find a better fitting model.

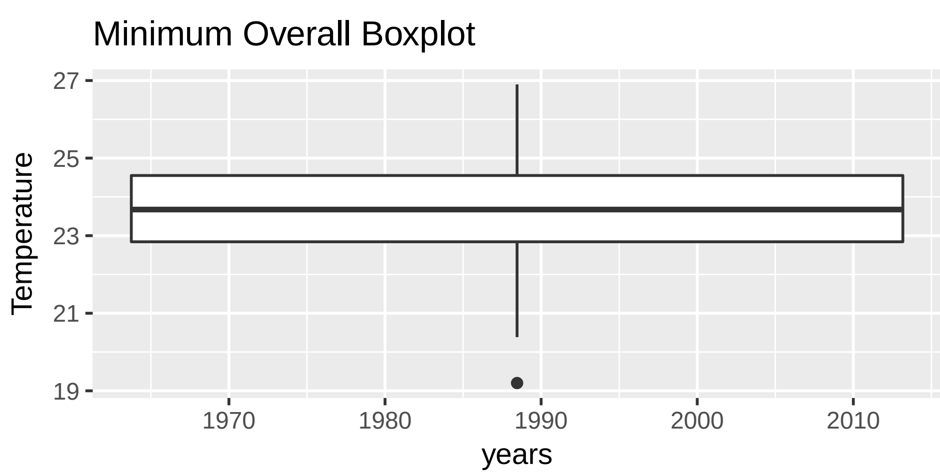
Background Information

Lomé is a coastal city. Since it is on the coast, the climate is more temperate than the other inland cities. It is also surrounded by a lagoon. Recently, the city of Lomé has grown due to an addition of an oil refinery. This could have a large impact on the future climate.

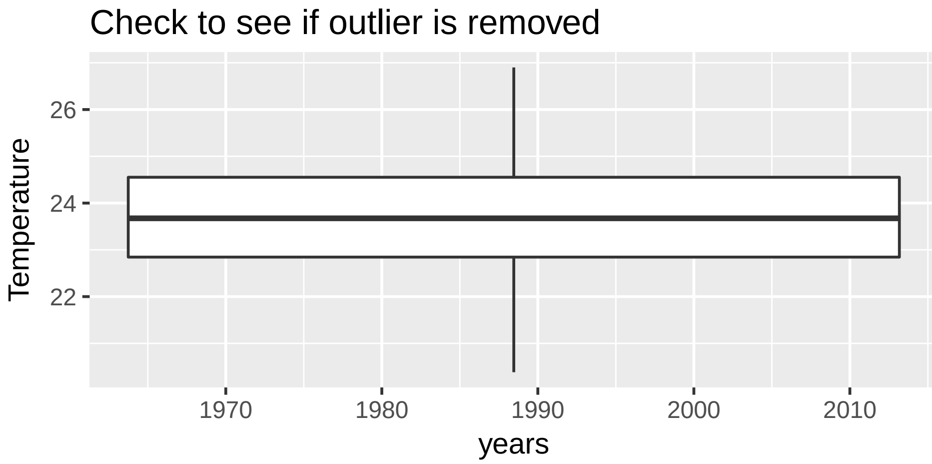
Minimum

Outliers





Based on the boxplots, we removed one value from January 1983 with a temperature of 19.2 degrees. This was the only outlier of all the data together. We chose not to remove the 2 points in August because they were not significant outliers in terms of the whole data. The reason these two points are showing up as outliers in the monthly box plots is because the IQR is so much smaller than the other months. The 2 points in August are not outliers in terms of all the data. The last boxplot is after the removal of the outliers.



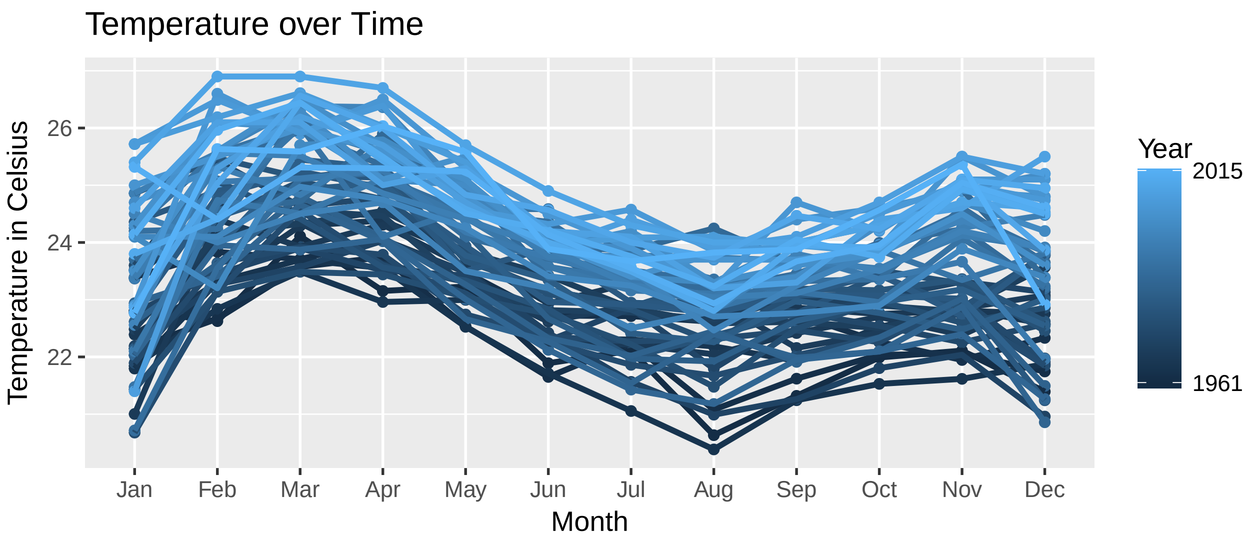
Modeling

Since the other group had already checked the linear and quadratic models, we wanted to start by adding sines and cosines to our model.

Before we dove in, we wanted to see if there was any periodicity in the years. To do this we graphed all of the points vs. month and ran a line through the years. We added color to the graph to point out which lines were the later years and which were more recent. This graph clearly shows how the temperatures are increasing over time.

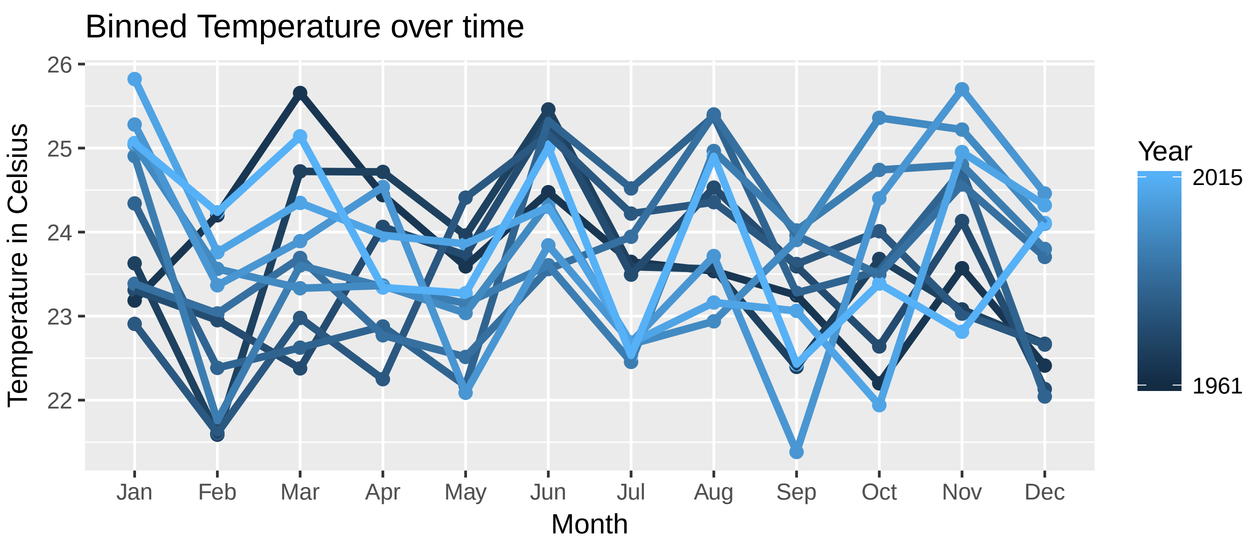
Interesting Graph

In this graph we plotted the temperatures by month. We then added lines to correspond to each of the years. The lines have a color gradient that allows you to see the warming of temperatures very clearly. The light colors correspond to the most recent years, while the darker colors are further back. As you can see the lighter color lines are falling at the top of the graph.



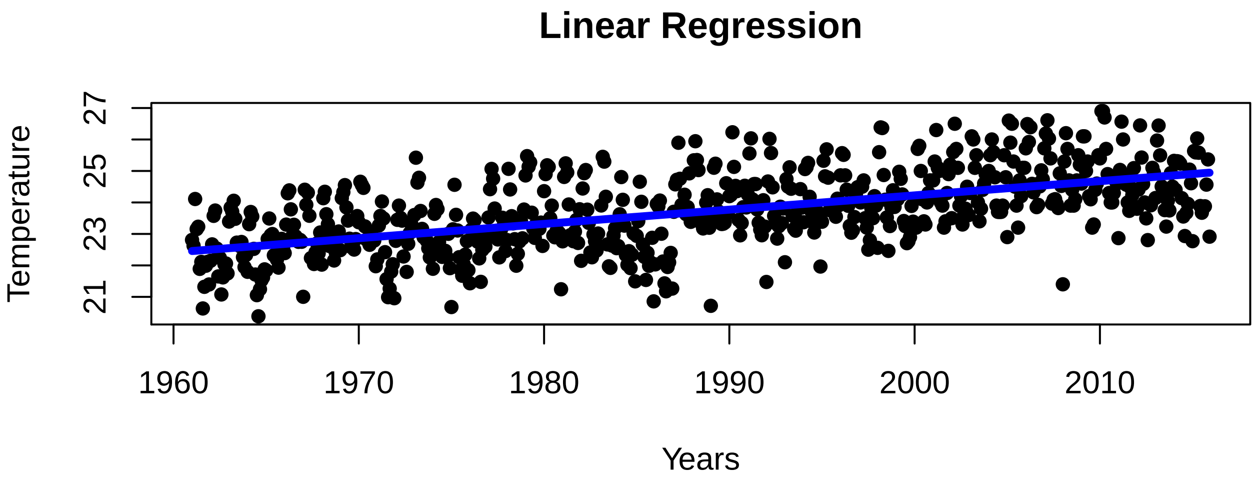
Binning previous graph to make it clearer

While the previous graph was very cool, we wanted to bin the data to see if it made it any clearer. We did the averages of 5 years for each month. This way we just have 11 lines instead of 55. Again the colors correspond to the years.



Linear Regression

With these graphs and the periodicity in mind, we began working on our models. To begin, we re-ran the linear regression to see what it looked like now that we had the monthly data.



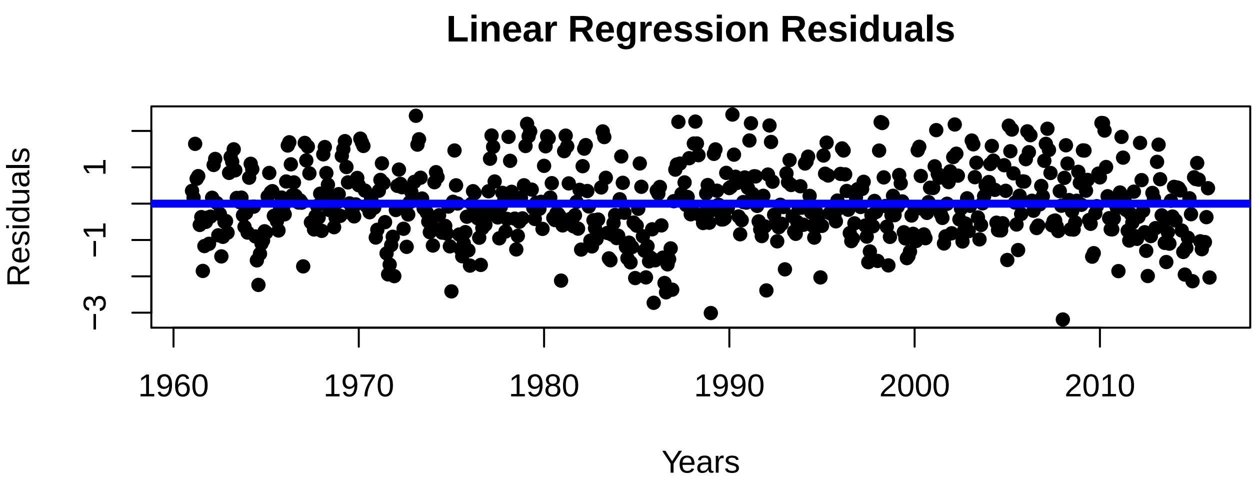
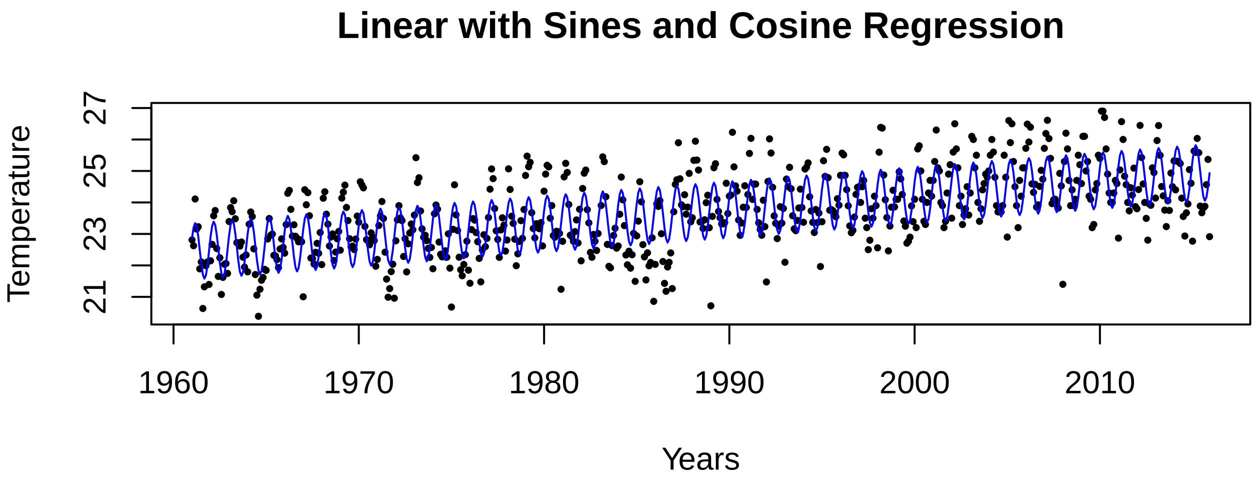
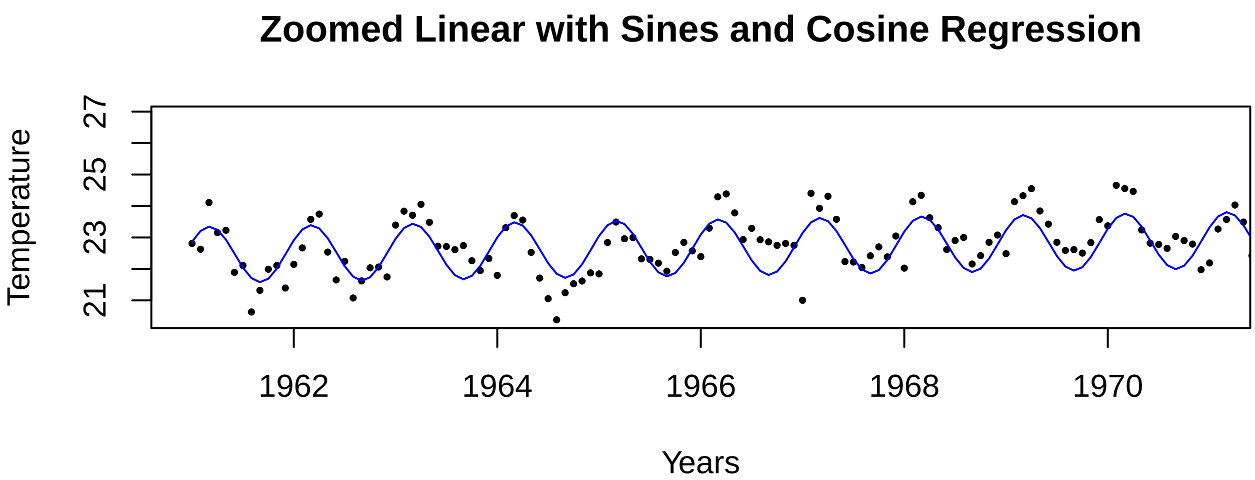


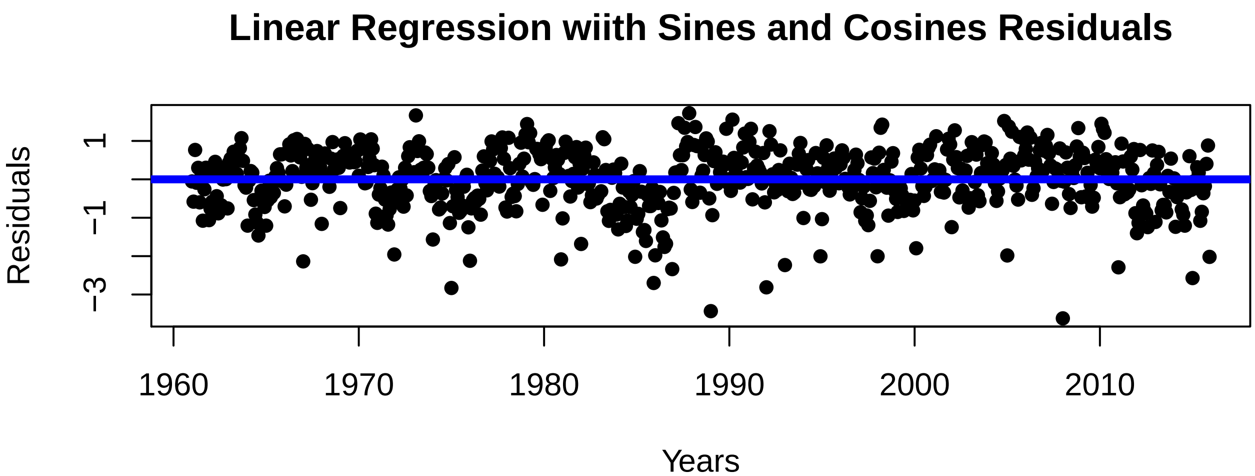
Table 1: Minimum Temperature Linear Regression Analysis

	Dependent variable:
	value
lDate	0.045*** (0.002)
Constant	−66.406*** (4.857)
Observations	659
R^2	0.344
Adjusted R^2	0.343
Residual Std. Error	0.996 (df = 657)
F Statistic	344.224*** (df = 1; 657)
Note:	*p<0.1; **p<0.05; ***p<0.01

The results of the linear regression, now knowing the monthly averages, is bleak. With an Adjusted R squared of .343, we feel there must be a better option. Since we have already noticed a yearly period, adding a sine and cosine term would be appropriate.







The adjusted R-squared for this equation is .5937. We then looked at the quadratic model with sines and cosines and found that the quadratic term was not significant. The linear regession with the sines and cosines is our best model to fit the data. It may not be the most accurate, but it is the best we have.

Table 2: Minimum Temperature Linear Regression Analysis with Sines and Cosines

	Dependent variable:
	value
lDate	0.045*** (0.002)
Constant	−66.406*** (4.857)
Observations	659
R^2	0.344
Adjusted R^2	0.343
Residual Std. Error	0.996 (df = 657)
F Statistic	344.224*** (df = 1; 657)
Note:	*p<0.1; **p<0.05; ***p<0.01

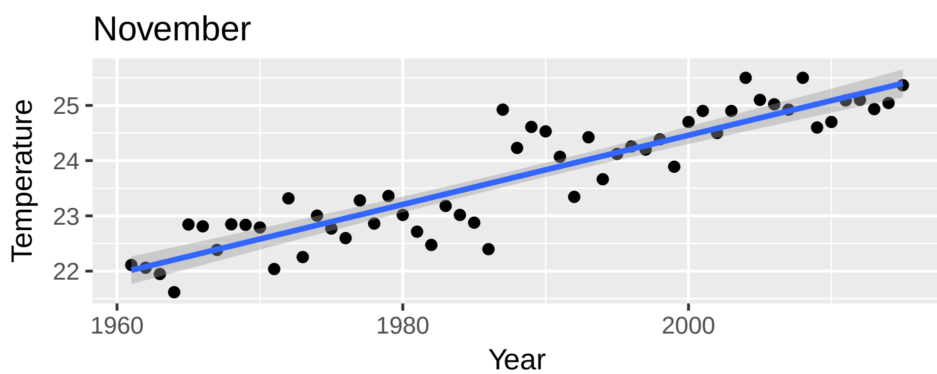


Table 3: November Minimum Temperature Linear Regression Analysis

	Dependent variable:
	Nov
Year	0.063*** (0.004)
Constant	−100.867*** (7.944)
Observations	55
R^2	0.823
Adjusted R^2	0.819
Residual Std. Error	0.470 (df = 53)
F Statistic	245.937*** (df = 1; 53)
Note:	*p<0.1; **p<0.05; ***p<0.01

Lastly for the minimums, we looked at all of the monthly data individually to see if some of the months were accurrately predicted by linear or quadratic regression. The most accurate model was a linear regression for November. This produced an R^2 of ~.82.

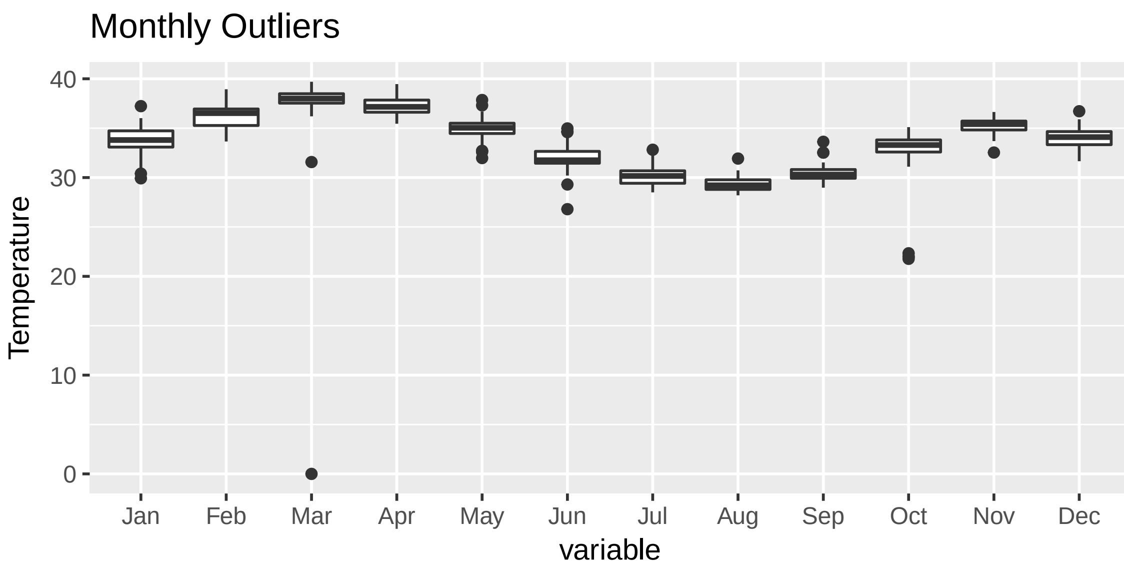
Minimum Conclusion

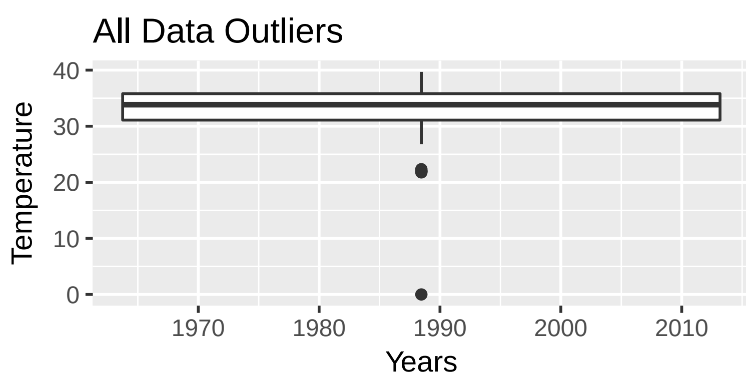
In regards to the minimum monthly temperatures, we believe our best model is a linear regression with sines and cosines terms. While the R^2 is not the best, the trends seem to match up and it produces better results than simple linear or quadratic.

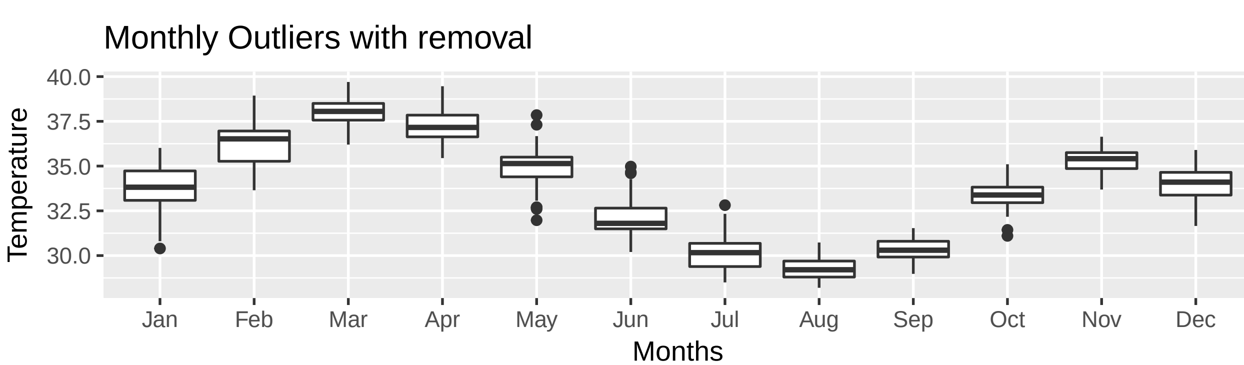
Maximum Data

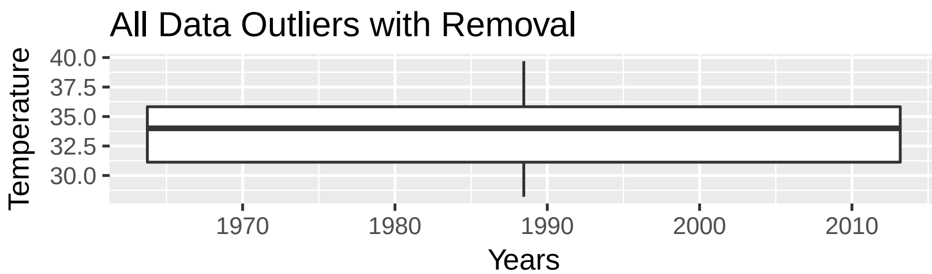
At first glance, there are some very weird things happening in the maximum data. 1972 and 1973, and 1993 and 1994 have the exact same data points for every row. What should be done with this? Should we remove both rows? Maybe only one is inaccurate? For the content of this report, We removed the years 1973 and 1994. The data seemed in line with the rest so, We doubt that both of the years were truly wrong.

To begin modeling, we started by identifying outliers. We ended up removing about 16 points. We could have done more, but we were worried about losing important information, that doesn't seem out of the realm of possibility. We did remove a 0 in march of 1983, other points in 8 different months.





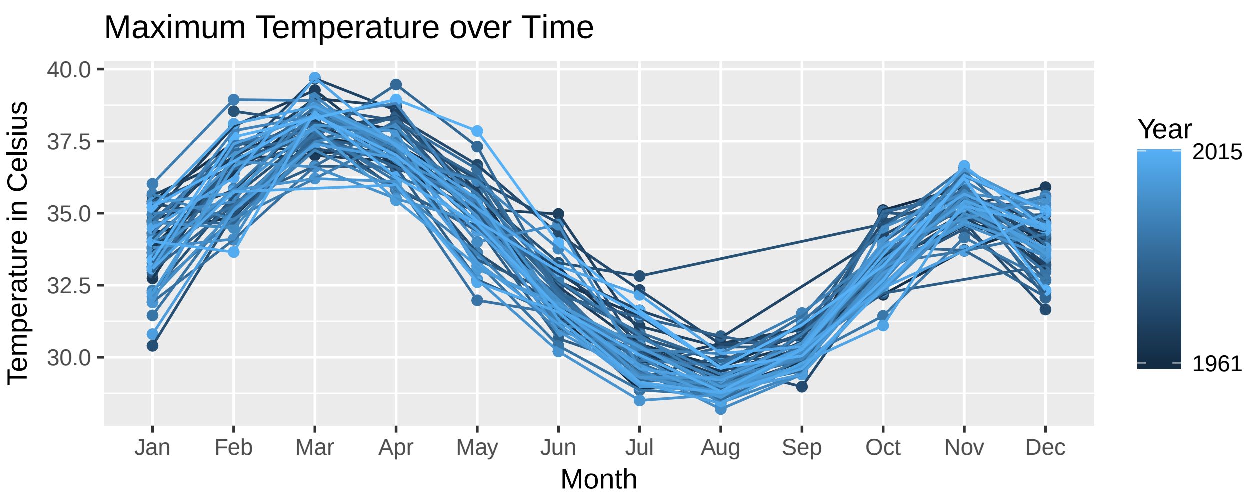




We chose to leave the 2 Octobers outliers because they are less than a degree away from points that are not considered outliers. The point in July is less than .3 from the others. We felt like all of these were within the possibility of real temperatures.

Interesting Graph

Since we had such good results doing this graph for the minimum temperatures, we decided to try it with the maximum temperatures. As you can tell, it is very different from the previous graph. The colors overlap, there is no clear distinction in recent years from older years. You may also notice two lines that cut across the middle from about July to October, this is because we removed a lot of outliers in that year. The next analysis might remove that year all together.



Regression

To start our modeling, we wanted to try running linear regression on the monthly data. The results were even worse than the minimum. The Adjusted R-squared is .00153. Hopefully, adding sines and cosines in our next approach will improve the model.

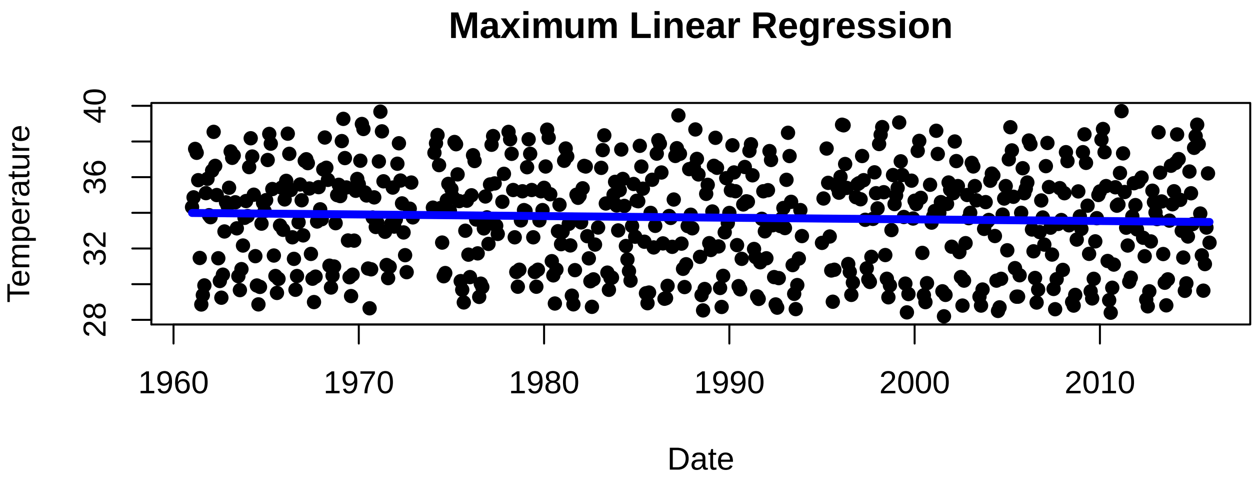
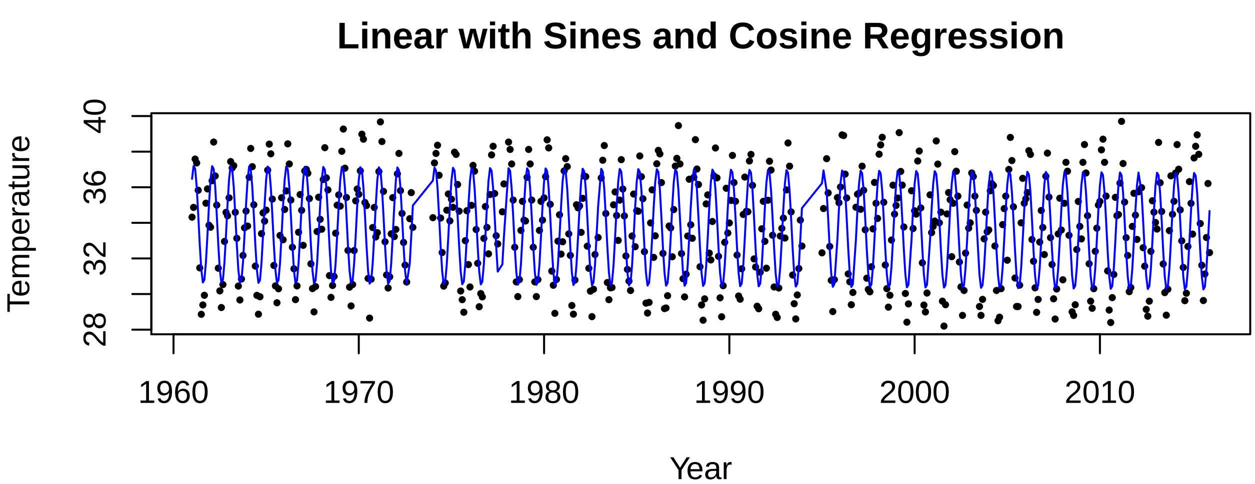


Table 4: Maximum Temperature Linear Regression Analysis

	Dependent variable:
	value
Date	−0.009 (0.007)
Constant	52.033*** (14.220)
Observations	621
R^2	0.003
Adjusted R^2	0.001
Residual Std. Error	2.868 (df = 619)
F Statistic	1.655 (df = 1; 619)
Note:	*p<0.1; **p<0.05; ***p<0.01

Sines and Cosines



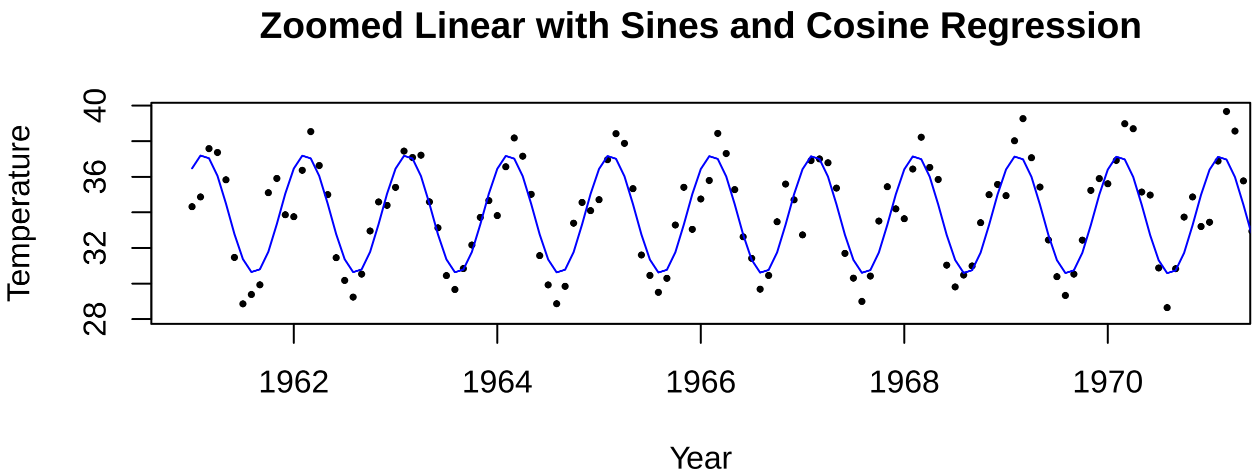


Table 5: Maximum Temperature Linear Regression with Sines and Cosines

	Dependent variable:
	value
Date)	2.131*** (0.094)
Date)	2.545*** (0.093)
Date	−0.007* (0.004)
Constant	48.064*** (8.165)
Observations	621
R^2	0.672
Adjusted R^2	0.671
Residual Std. Error	1.646 (df = 617)
F Statistic	422.078*** (df = 3; 617)
Note:	*p<0.1; **p<0.05; ***p<0.01

It is very interesting to note that the x term in our model is very slightly negative. The term is significant, but the coefficent is only -.008. However, the adjusted R^2 on this model is .671. This is an even better fit than we got for our minimum temperatures. We also tested a cubic, but the cubic term was not significant. We would choose this as our model for the maximum temperatures.

CONCLUSIONS

In conclusion, the best model for both the minimum temperatures and the maximum temperatures is the linear regression with sine and cosine terms. The "interesting graph" we created for the minimum temperatures data really shows the rise in the climate change for the minimum temperatures. The maximum temperatures, however, does not show this same trend

What information might be useful moving forward?

It would be useful to know what is happening in the maximums' data. The years being repeated is clearly a mistake, but is the real data stored somewhere else? We would not have even noticed the repition if we didn't have to individually pluck out the outliers. Moving forward, we would like to know if the outliers that were excluded were actually the correct data. We would also like to know what instruments were used, how regularly was the data collected, and how consistent was the person/people collecting the data. We would also like to learn more about the opening of the oil refinery and the effect that could have on their climate.