Atakpame Final

Allison Zembrodt

4/29/2018

Preprocessing of Our Data

Before I can run any analysis, I need to filter out the outliers of our data and add in the Ensos and SSTs. In order to merge the files, I need to make sure that the Decimal Years from the FinalTogo data and the EnsoSST data are the same. To do this, I rounded both files to 3 decimal places. I also removed the extra column in the csv so it is not present in our data set. Finally, I subtracted 1961 from the years to make them start at 0 and go to 55.

```
FinalTogo$DecYear = round(FinalTogo$DecYear, digits =3)
EnsoSST <- EnsoSST %>% select(-X1)
#Merge the files together based on Decimal year
FinalTogo <- merge(FinalTogo, EnsoSST)
FinalTogo$DecYear <- FinalTogo$DecYear -1961
```

Split the Data

I used the tidyverse and dplyr to split our data into 3 files. Min removes just the minimum outliers, Max removes just the maximum outliers, and Rainfalldata removes all outliers and null values. Now, I can rebuild the models from the modeling team.

```
Min <- FinalTogo %>% filter(MinOutlier=="No")
Max <- FinalTogo %>% filter(MaxOutlier == "No")
Rainfalldata <- FinalTogo %>% filter(MaxOutlier== 'No') %>% filter(MinOutlier== 'No')
Rainfalldata= Rainfalldata[complete.cases(Rainfalldata$Rainfall),]
```

Add Predicted Column for all of the datasets

Using the linear model built in predict function, I predicted the values for all of our data points. I then column-bound those predicted values so that there is a column for the predicted values in all of the data files. This makes it much easier to look at the predicted vs. actual data. It also makes graphing the resiudals much easier.

```
Min = cbind(
  Min,
  minPredict= predict(minlinear, newdata = Min)
)
Max= cbind(
  Max,
  maxPredict= predict(maxlinear, newdata = Max)
)
Rainfalldata = cbind(
  Rainfalldata,
  rainPredict = predict(rainfallLinear, newdata = Rainfalldata)
)
```

Filter for City

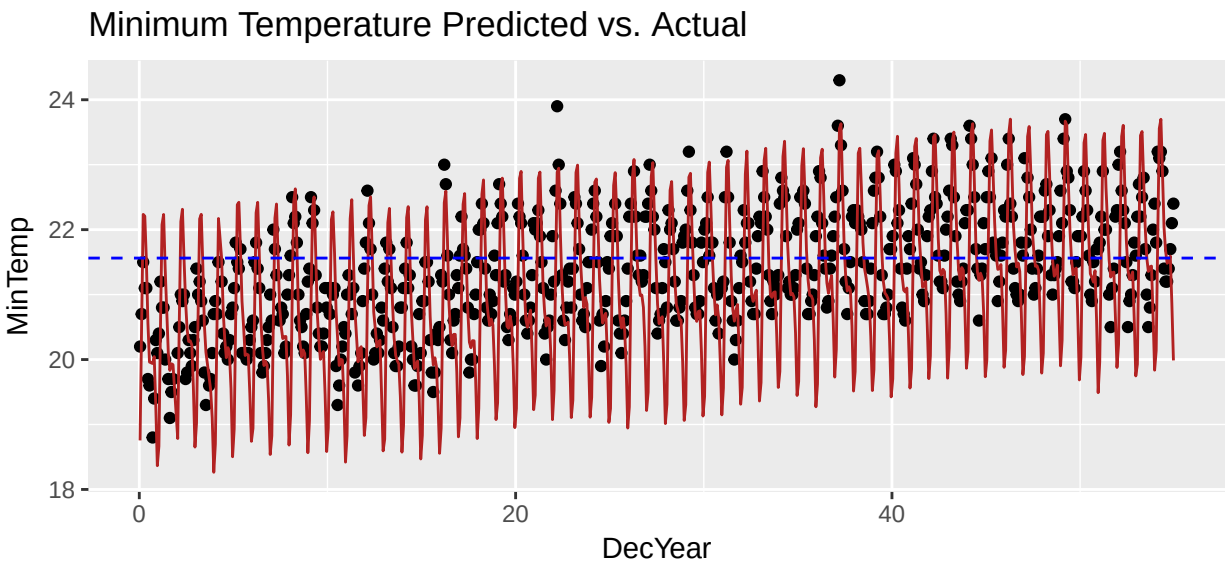
Now that our data is completely preprocessed, I can look at my city, Atakpame. Again I used the function 'filter' to subset the data for my specific city. I filtered all three big files, Min, Max, and Rainfalldata, and remamed the files AMin, AMax and ARain.

```
AMin = Min %>% filter(location =='Atakpame')
AMax = Max %>% filter(location =="Atakpame")
ARain = Rainfalldata %>% filter(location == "Atakpame")
```

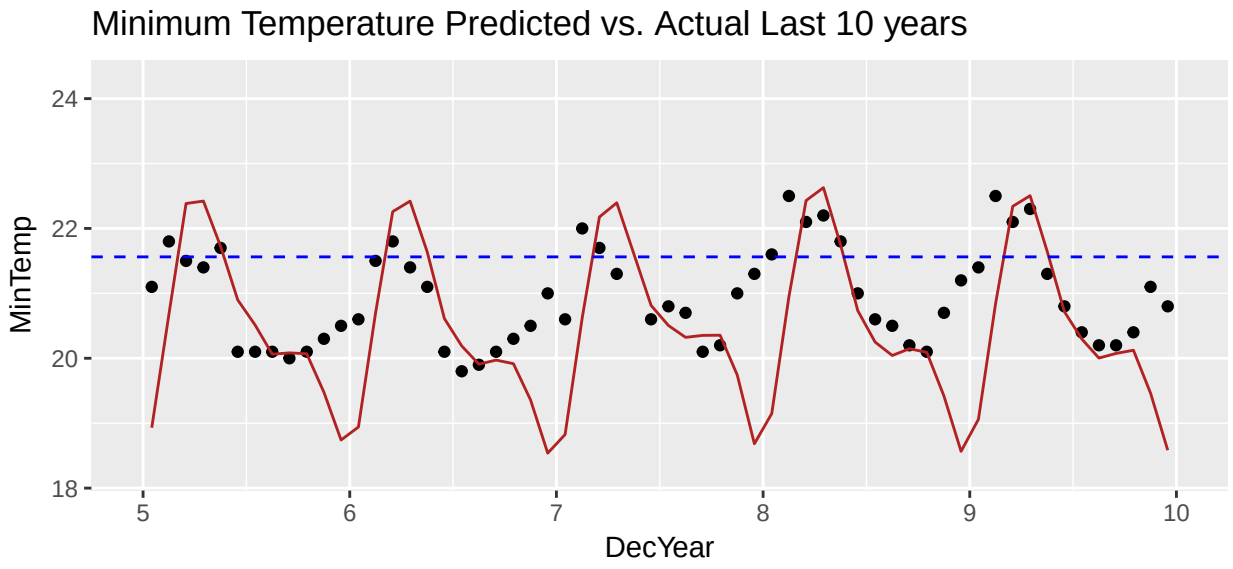
Minimum Analysis

To begin my analysis, I wanted to look at how the predicted vs. actual values compare for my city. The black points are the actual values and the red line is the predicted values. If you zoom in on the model to look at a 10 year period, I noticed that the periods are out of sync. A specific example is if you look closely around the years 6,7,8, and 9, our model is predicting low values, but the actual values are increasing.

```
ggplot(AMin) + geom_point(aes(x = DecYear,y = MinTemp ))+
  geom_line( aes(x= DecYear, y = minPredict), color = 'Firebrick') +
  geom_hline(yintercept=21.5622, linetype="dashed", color = "blue") +
  labs(title = "Minimum Temperature Predicted vs. Actual ")
```




```
ggplot(AMin) + geom_point(aes(x = DecYear,y = MinTemp ))+
  geom_line( aes(x= DecYear, y = minPredict), color = 'Firebrick')+
  geom_hline(yintercept=21.5622, linetype="dashed", color = "blue")+
  labs(title = "Minimum Temperature Predicted vs. Actual Last 10 years")+xlim(5, 10)
```

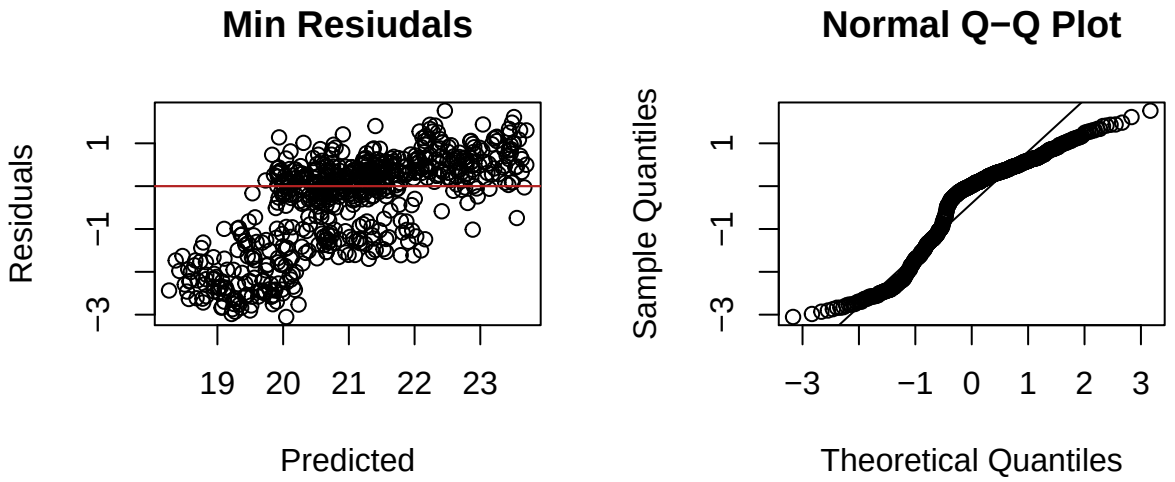


Residuals

There is definetly a pattern in the residuals. As we discussed in class, the model appears to fit the cities on the northern end of the country much better than the southern end. I think this model is completely out of sync for our city and is causing the mean to be a better predictor than the model we created. Our cities' residuals definetly have a pattern and the QQ plot looks like an S. There are sever departures from normality in our residuals and qq-plot.

```
par(mfrow = c(1,2))
plot(AMin$minPredict, AMin$minPredict - AMin$MinTemp, xlab = "Predicted", ylab = "Residuals")
abline(h=0, col = "Firebrick", lwd= 1.1)
title("Min Resiudals")

minp = as.vector(AMin$minPredict)
mina = as.vector(AMin$MinTemp)
qqnorm(minp-mina)
qqline(minp-mina)
```



Min "Our Squared" Function

This equation allows us to have a negative R-squared if the mean preforms better than our model. When I run this function on our Atakpame data, I get -.5714. This is horrible. It means that our model is better predicted with the mean. Since our city preforms this badly, I wanted to see if our city was abnormal, or if other cities preform this badly as well.

```
OurSquaredMin = function(df){
  ss.e =0
  ss.tot = 0
  for (i in 1:nrow(df)){
    ss.e = ss.e + (df$MinTemp[i]-df$minPredict[i])^2
    ss.tot= ss.tot +(df$MinTemp[i]-mean(Min$MinTemp))^2
  }
  1-(ss.e/ss.tot)
}
```

```
OurSquaredMin(AMin)
```

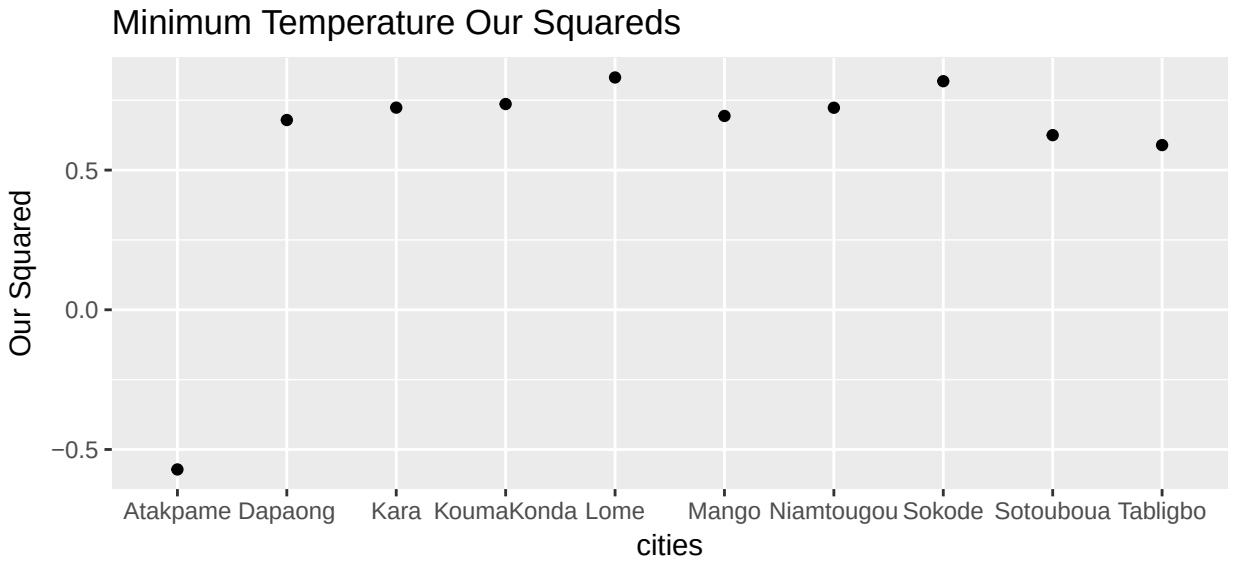
```
## [1] -0.5714952
```

I built a for loop to run this function on all of the cities' data. As you can see from the graph, Atakpame preforms much worse than all the other cities. All other cities have an Our Squared of at least .5. Our city is definitely outside the norm.

```
#Our Squareds for Minimum
Rmin= c()
cities = c("Dapaong", "Mango", "Niamtougou", "Kara", "Sokode", "Sotouboua", "KoumaKonda", "Atakpame", "
for (i in 1:10){
  x= Min %>%filter(location == cities[i])
  Rmin <- c(Rmin,OurSquaredMin(x))
}
MinOur = data.frame(cities, Rmin)
MinOur %>% filter(cities =="Atakpame")
```

```
##     cities       Rmin
## 1 Atakpame -0.5714952
```

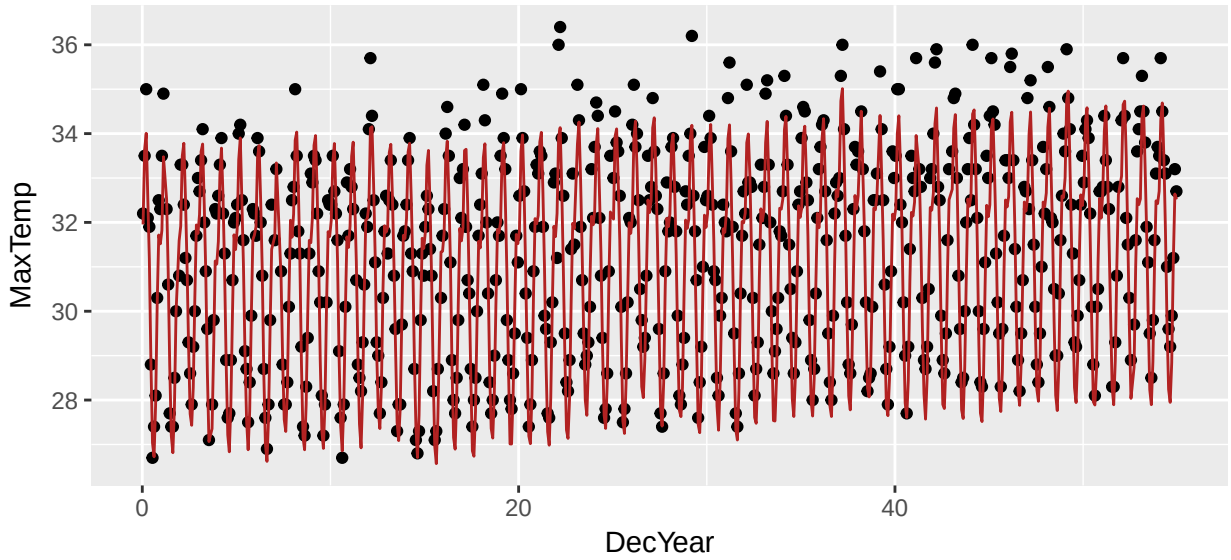
```
ggplot(MinOur, aes(x=cities, y = Rmin))+geom_point()+ labs(y = "Our Squared", title = "Minimum Temperatu
```



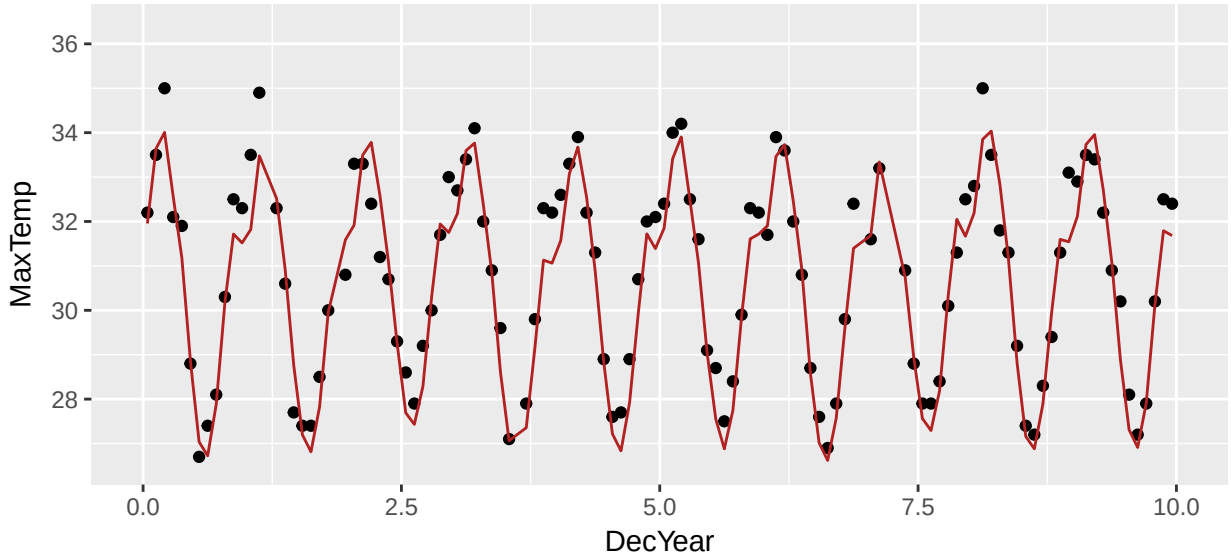
Maximum Analysis

Now, that we have looked at the minimum model and how it does on our city data, let's look at the maximum model. Again, the black points are the actual values and the red line is the predicted values. If you zoom in on the model to look at a 10 year period, I noticed that the model is fitting really well. The general trend of the temperatures is well modeled with this equation. There are a few extreme outliers, but for the most part everything lines up.

```
ggplot(AMax) + geom_point(aes(x = DecYear,y = MaxTemp ))+
  geom_line( aes(x= DecYear, y = maxPredict), color = 'Firebrick')
```




```
ggplot(AMax) + geom_point(aes(x = DecYear,y = MaxTemp ))+
  geom_line( aes(x= DecYear, y = maxPredict), color = 'Firebrick')+xlim(0, 10)
```

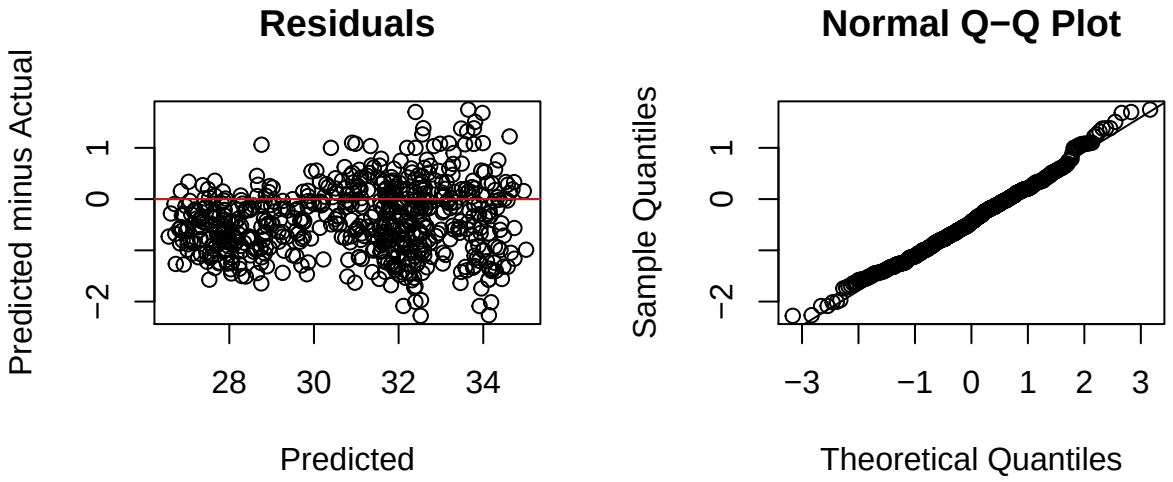


Maximum Residuals

For the maximum model, my cities residuals look ok. I also wanted to look at the QQ plot as another test for normality. This QQ plot is pretty much what we would expect to see for a normal distribution.

```r
#Maximum Temperature
par(mfrow = c(1,2))
plot(AMax$maxPredict, AMax$maxPredict - AMax$MaxTemp, xlab = "Predicted", ylab = "Predicted minus Actual
abline(h=0, col = "Firebrick", lwd= 1.1)
title("Residuals")
maxp = as.vector(AMax$maxPredict)
maxa = as.vector(AMax$MaxTemp)

qqnorm(maxp-maxa)
qqline(maxp-maxa)
```



Max Our Squared

This function was built to calculate the "Our Squared" of our maximum model for our city. Atakpame does really well with the maximum model. The "Our Squared" is almost .9. Again, I wanted to check and see if our city did as well as the others. Overall, I think that the maximum looks great! Our R-squared is one of the highest, and I would say that the maximum temperature model fits well.

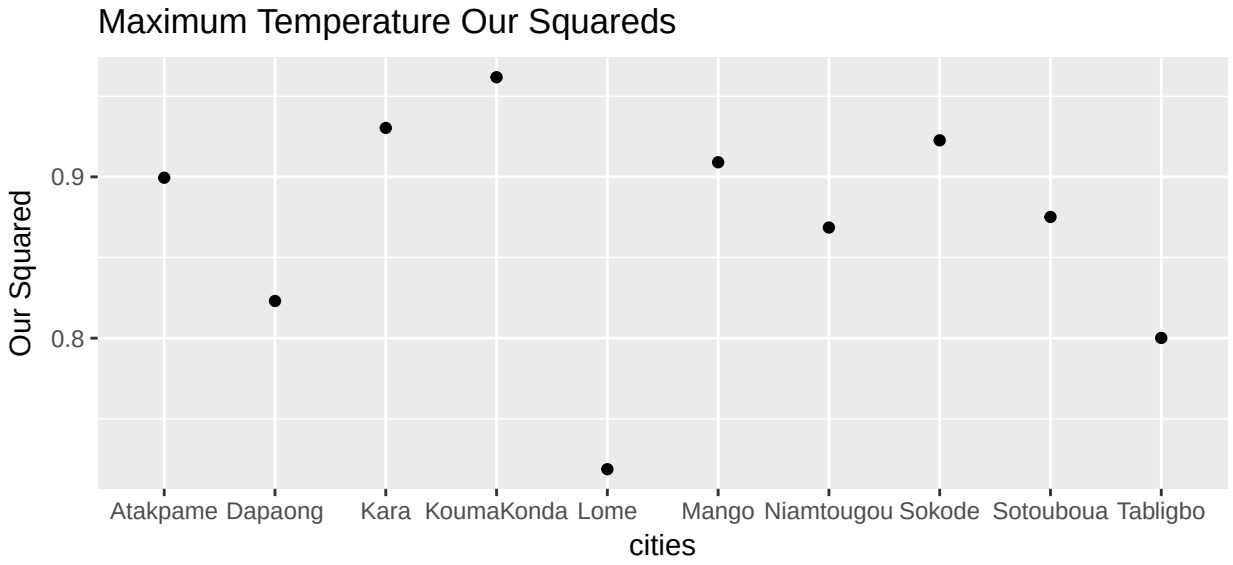
```r
OurSquaredMax = function(df){
  ss.e =0
  ss.tot = 0
  for (i in 1:nrow(df)){
    ss.e = ss.e + (df$MaxTemp[i]-df$maxPredict[i])^2
    ss.tot= ss.tot +(df$MaxTemp[i]-mean(Max$MaxTemp))^2
  }
  1-(ss.e/ss.tot)
}
```

```r
#Maximum
Rmax= c()
cities = c("Dapaong", "Mango", "Niamtougou", "Kara", "Sokode", "Sotouboua", "KoumaKonda", "Atakpame", "
for (i in 1:10){
  x= Max %>%filter(location == cities[i])
  Rmax <- c(Rmax,OurSquaredMax(x))
}

MaxOur = data.frame(cities, Rmax)
MaxOur %>% filter(cities =="Atakpame")
```

```
##     cities      Rmax
## 1 Atakpame 0.8994645
```

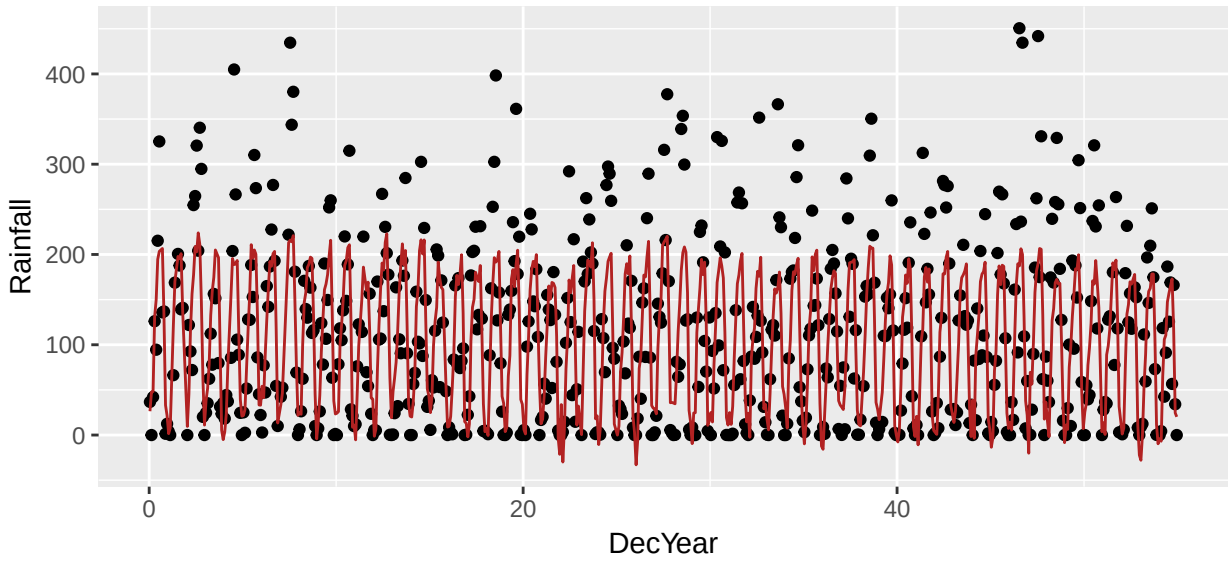
```
ggplot(MaxOur, aes(x=cities, y = Rmax))+geom_point()+ labs(y = "Our Squared", title = "Maximum Temperatu
```



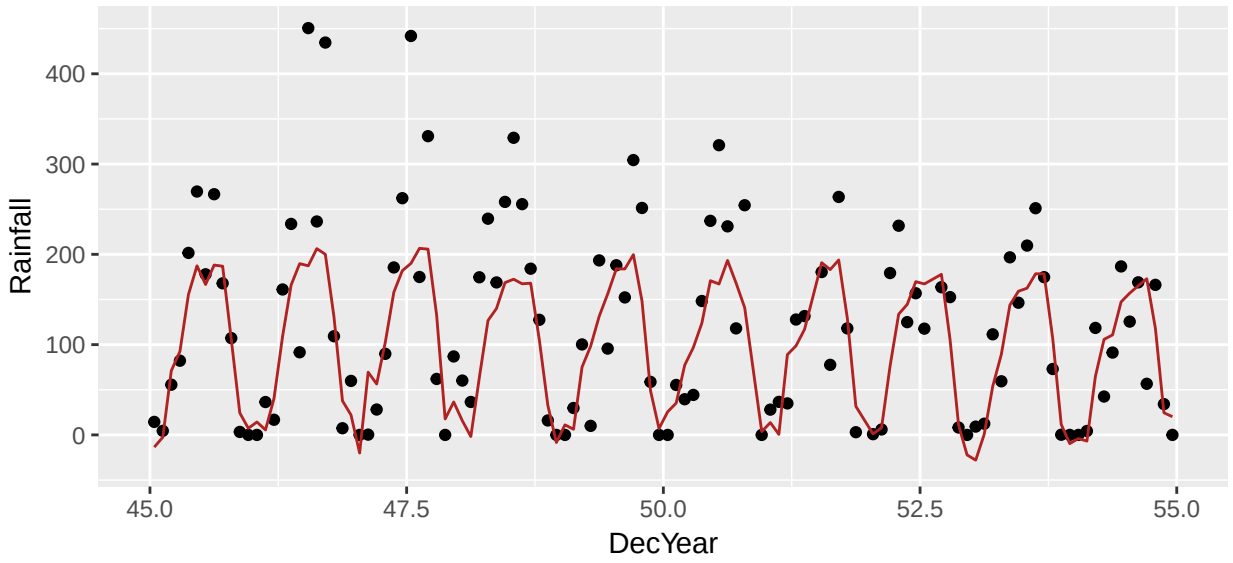
Rainfall Analysis

Finally, I want to take a look at the rainfall model and see how well it predicts for our city. As you can see from the first graph, many of the high points are NOT caputred by the model. When you take a closer look at the last 15 years, it is easier to look at the fit of the model.

```
ggplot(ARain)+geom_point(aes(x= DecYear, y = Rainfall))+geom_line(aes(x = DecYear, y = rainPredict), col
```




```
ggplot(ARain)+geom_point(aes(x= DecYear, y = Rainfall))+geom_line(aes(x = DecYear, y = rainPredict), col
```



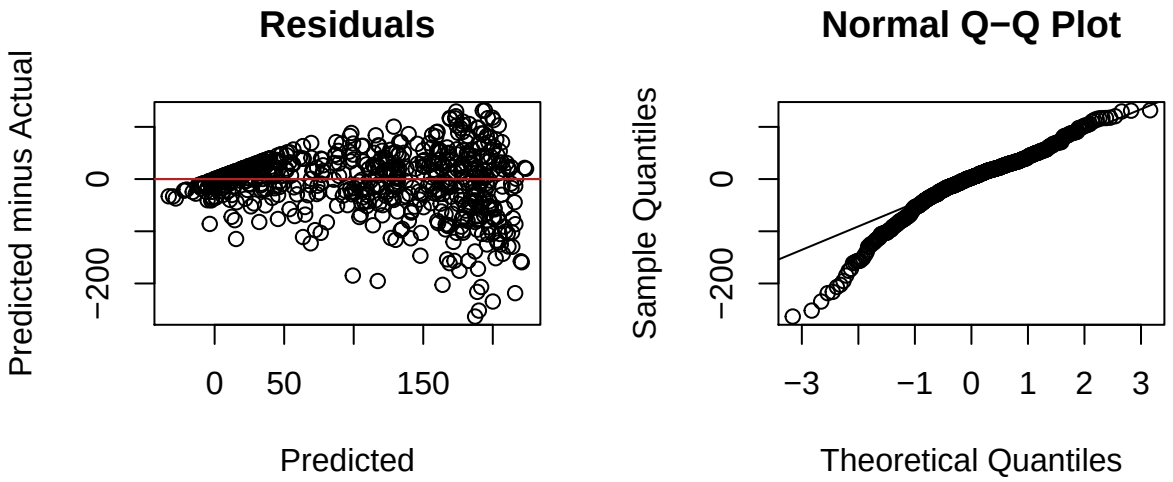
Residuals

Now, I want to look at the residuals of the rainfall model for my city. They are very clearly not normal. I can tell just by glancing at the resiudals plot that they are not normally distributed. There are so many zeros in this data that I expected our data to be abnormal.

```
#Rainfall
par(mfrow = c(1,2))
plot(ARain$rainPredict, ARain$rainPredict - ARain$Rainfall, xlab = "Predicted", ylab = "Predicted minus

abline(h=0, col = "Firebrick", lwd= 1.1)
title("Residuals")

rainp = as.vector(ARain$rainPredict)
raina = as.vector(ARain$Rainfall)
qqnorm(rainp-raina)
qqline(rainp-raina)
```



##Rainfall Our Squared Finally, I want to look at the "Our Squared" for our city for the rainfall function and see how it compares to other cities. Our city gets about .65 which is pretty average for all the cities. Overall, I think that the rainfall model does about as well as can be expected for Atakpame.

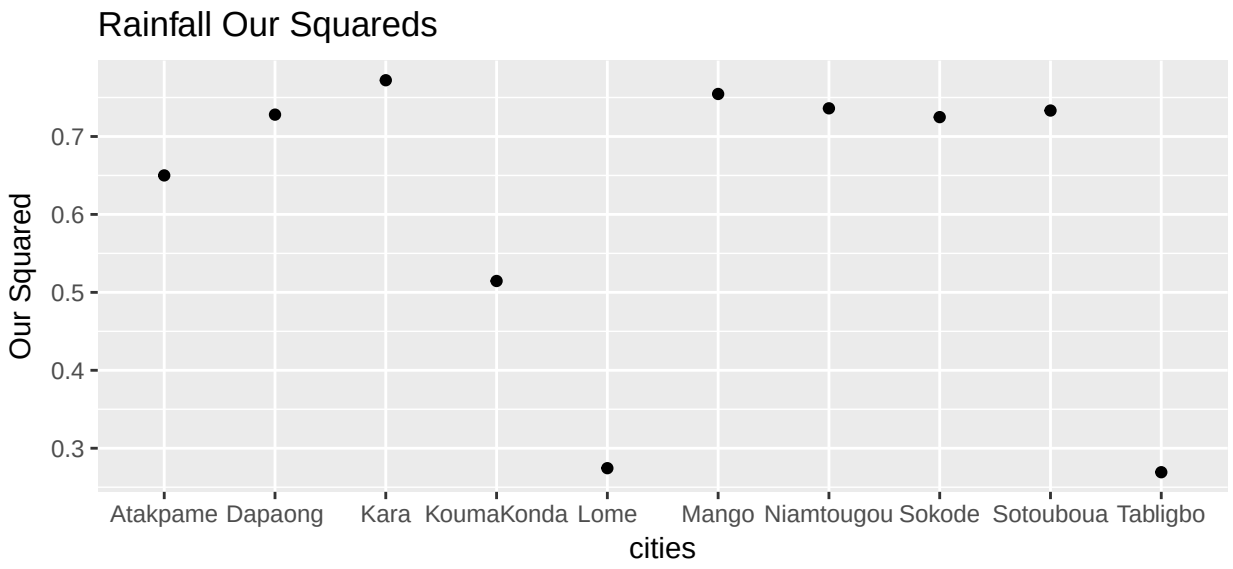
```r
OurSquaredRainfall = function(df){
  ss.e =0
  ss.tot =0
  for(i in 1:nrow(df)){
    ss.e = ss.e +(df$Rainfall[i]- df$rainPredict[i])^2
    ss.tot = ss.tot + (df$Rainfall[i]-mean(Rainfalldata$Rainfall))^2
  }
  1-(ss.e/ss.tot)
}
```

```r
#Rainfall
Rrain= c()
cities = c("Dapaong", "Mango", "Niamtougou", "Kara", "Sokode", "Sotouboua", "KoumaKonda", "Atakpame", "
for (i in 1:10){
  x= Rainfalldata %>%filter(location == cities[i])
  OurSquaredRainfall(x)
  Rrain <- c(Rrain,OurSquaredRainfall(x))
}

RainOur = data.frame(cities, Rrain)
RainOur %>% filter(cities =="Atakpame")
```

```
##     cities     Rrain
## 1 Atakpame 0.6500426
```

```r
ggplot(RainOur, aes(x=cities, y = Rrain))+geom_point()+ labs(y = "Our Squared", title = "Rainfall Our S
```



Final Words on the Models

The minimum temperature model did not fit the city of Atakpame very well. In fact, just using the overall mean minimum temperature would have been a better predictor of the minimum temperatures in Atakpame. When I closely examined the actual Atakpame minimum temperatures verses the predicted values from the model, I noticed that the periods were out of sync. This is why the residuals and qq-plot had severe departures from normality. Overall, the minimum temperature model did not work well for our city, and the mean minimum temperature would have been a better fit. The rainfall model did fine with predicting

the periods of our city, but it did not do well in capturing the extremes. The "Our Squared" for Atakpame rainfall was .65. Finally, the maximum model, however, worked very well for our city. The "Our Squared" values was at 0.9. This means that 90% of the variability in our data was explained by the model. Overall, the minimum model worked very poorly for Atakpame, the rainfall model was mediocre, but the maximum model fit great.