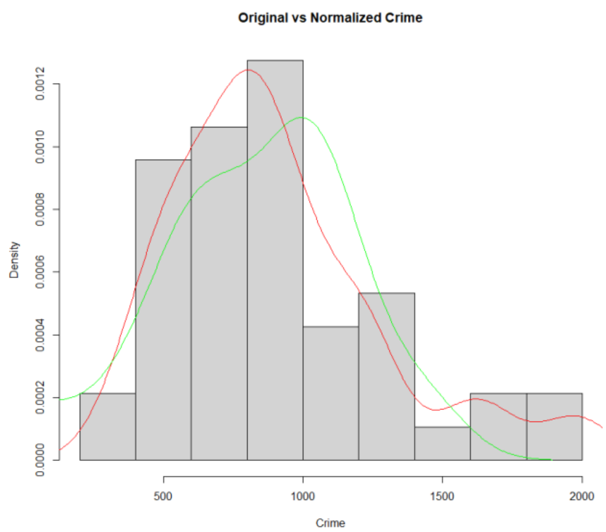
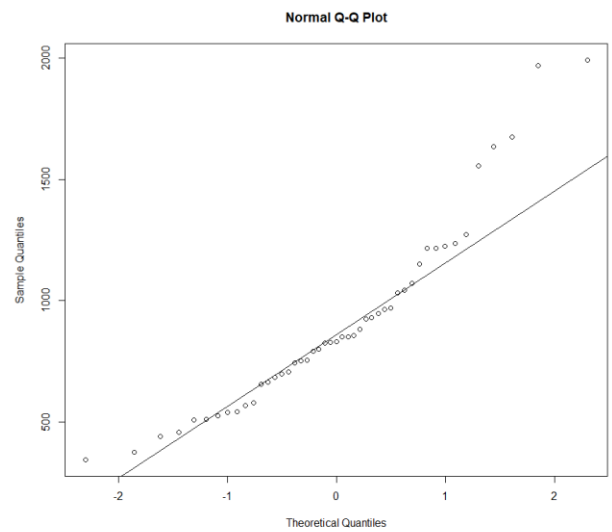
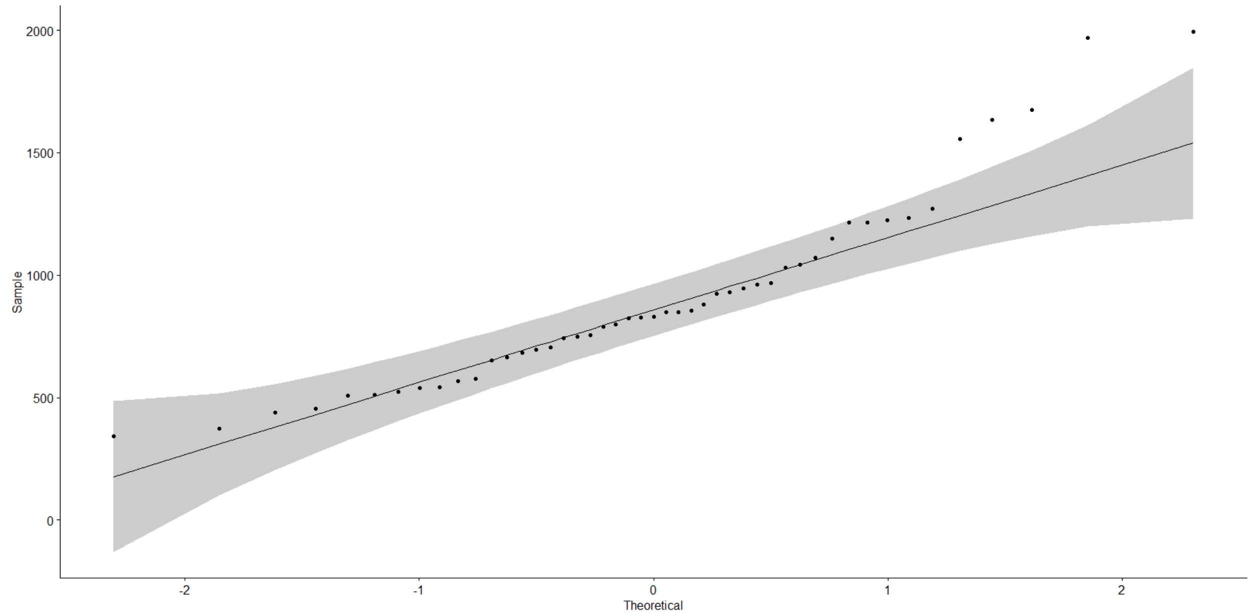
Question 5.1: Using crime data, test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

First things first, we want to evaluate the quality of our data. As shown below we plotted a histogram and a qqplot with a perfect line to see how far the data deviates from the theoretical assumption (line) to constitute it to be normal. Meanwhile, the graph on the left shows the histogram with the kernel density of the histogram in red. Also, the green line shows the "normal" distribution of the data, we normalized the data points and it allowed us to plot it in the same graph.

Moreover, the plot below (2nd one) is about the same as the qqplot but it better demonstrates a 95% confidence interval of where a normal data ought to be. As we can see there are several points outside that boundary justifying that this crime data is not normal.







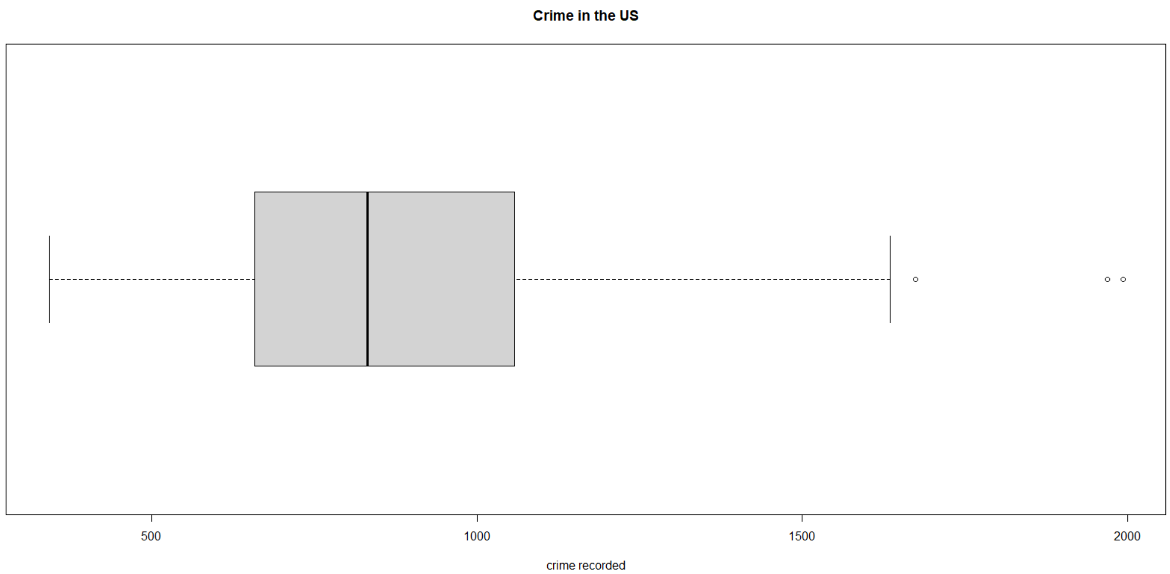

```
> # Method 3: Formal Statistical Tests
> shapiro.test(df_crime)

        Shapiro-Wilk normality test

data:  df_crime
W = 0.91273, p-value = 0.001882

> ad.test(df_crime)

        Anderson-Darling normality test

data:  df_crime
A = 1.1725, p-value = 0.004139
```

The third screenshot is a formal statistical test for the Shapiro-Wilks and Anderson- Darling test for normality. The better of the two is Shapiro-Wilks but I feel like it does not hurt to double check with a graphical and statistical hypothesis approach. In both tests, the hypothesis is as follows (for both):

$$H_0: Data\ comes\ from\ a\ normal\ distribution$$

$$H_A: Data\ does\ NOT\ comes\ from\ a\ normal\ distribution$$

Since the p-value is less than $\alpha = 0.1, .05, and\ 0.01$ confidence level, we have sufficient evidence to reject the null hypothesis via Shapiro-wilks tests and Anderson-Darling test for normality. There for we can conclude that the data does not come from a normal distribution.




```
> grubbs.test(df_crime)

        Grubbs test for one outlier

data:  df_crime
G = 2.81287, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier
```

```
> rosnerTest(df_crime)
$distribution
[1] "Normal"

$statistic
     R.1      R.2      R.3 
2.812874 3.063425 2.564571 

$sample.size
[1] 47

$parameters
k 
3 

$alpha
[1] 0.05

$crit.value
lambda.1 lambda.2 lambda.3 
3.103243 3.094456 3.085425 

$n.outliers
[1] 0

$alternative
[1] "Up to 3 observations are not\n                                 from the same Distribution."

$method
[1] "Rosner's Test for Outliers"

$data
 [1]  791 1635  578 1969 1234  682  963 1555  856  705 1674  849  511  664  798  946  539  929  750 1225  742  439 1216
[24]  968  523 1993  342 1216 1043  696  373  754 1072  923  653 1272  831  566  826 1151  880  542  823 1030  455  508
[47]  849

$data.name
[1] "df_crime"

$bad.obs
[1] 0

$all.stats
  i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 905.0851 386.7627  1993      26 2.812874   3.103243   FALSE
2 1 881.4348 355.0161  1969       4 3.063425   3.094456   FALSE
3 2 857.2667 318.4678  1674      11 2.564571   3.085425   FALSE
```

```
> BPS = boxplot.stats(df_crime);BPS
$stats
[1]  342.0  658.5  831.0 1057.5 1635.0

$n
[1] 47

$conf
[1] 739.0438 922.9562

$out
[1] 1969 1674 1993
```

```
> interqrange
[1] 399
> LB = BPS$stats[2] - 1.5*interqrange; LB
[1] 60
> UB = BPS$stats[4] + 1.5*interqrange; UB
[1] 1656
```

Now for outlier detection, plotted the traditional box plot and it displayed the min, 25th quartile, median, 75th quartile, and maximum. As you can see the potential outliers are the three dots beyond the max value of 1635 of 47 data points.

Moreover, I calculated the IQR to calculate the lower and upper bound as shown:

$$Q_1 - 1.5(IQR) = Lower\ Bound$$

$$Q_3 + 1.5(IQR) = Upper\ Bound$$

What this tells us is that anything below the lower bound is a lower outlier and anything above the upper bound is an outlier. Thankfully, the function has provided 3 data points to where it is out one of these boundaries and those are values: 1674,1969, and 1993

Funny enough, I conducted two statistical tests one required by the assignment (Grubbs test) and an additional one (Rosner test) to verify with the first one. The statistical hypothesis test is as follows:

For Grubbs:

$$H_0: Highest\ value\ is\ NOT\ an\ outlier$$

$$H_A: Highest\ value\ is\ an\ outlier$$

For Rosner:

$$H_0: There\ are\ NO\ outliers$$

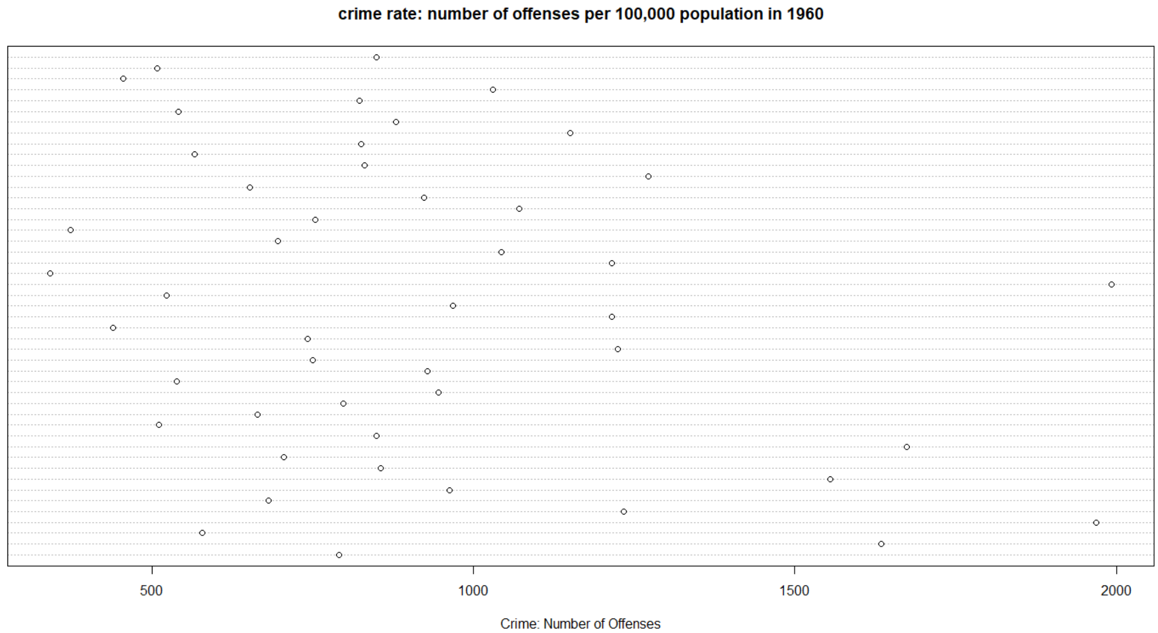
$$H_A: There\ exists\ at\ least\ one\ outlier$$

At $\alpha = .01\ and\ .05$ confidence level, we fail to reject the null hypothesis because 0.07 is greater than 0.05 boundary. So, we do not have sufficient evidence to reject the null hypothesis where the highest value (1993) is NOT an outlier.

However, at $\alpha = .1$, it is statistically significant thus we would have sufficient evidence to reject he null hypothesis and in favor of the alternative hypothesis that 1993 is an outlier.

Similarly, the Rosner test provides a similar output as the Grubbs test. Unlike the Grubbs test, the Rosner tests for multiple outliers, where student T test is to Grubbs, F test is to Rosner. Thankfully, it provides an indicator of whether the value is an outlier and like the Grubbs test all three data points mention is not an outlier.

Overall, the tests could be doing one of two things. The first is that they require a higher threshold than what is provided by the dataset to be classified as an outlier. Second, it could be what Dr. Sokol mention in the module in clustering where it could a group type of outlier where the algorithm picks it up but does not classify it as an extreme outlier since it is near a few data points with similar characteristics.



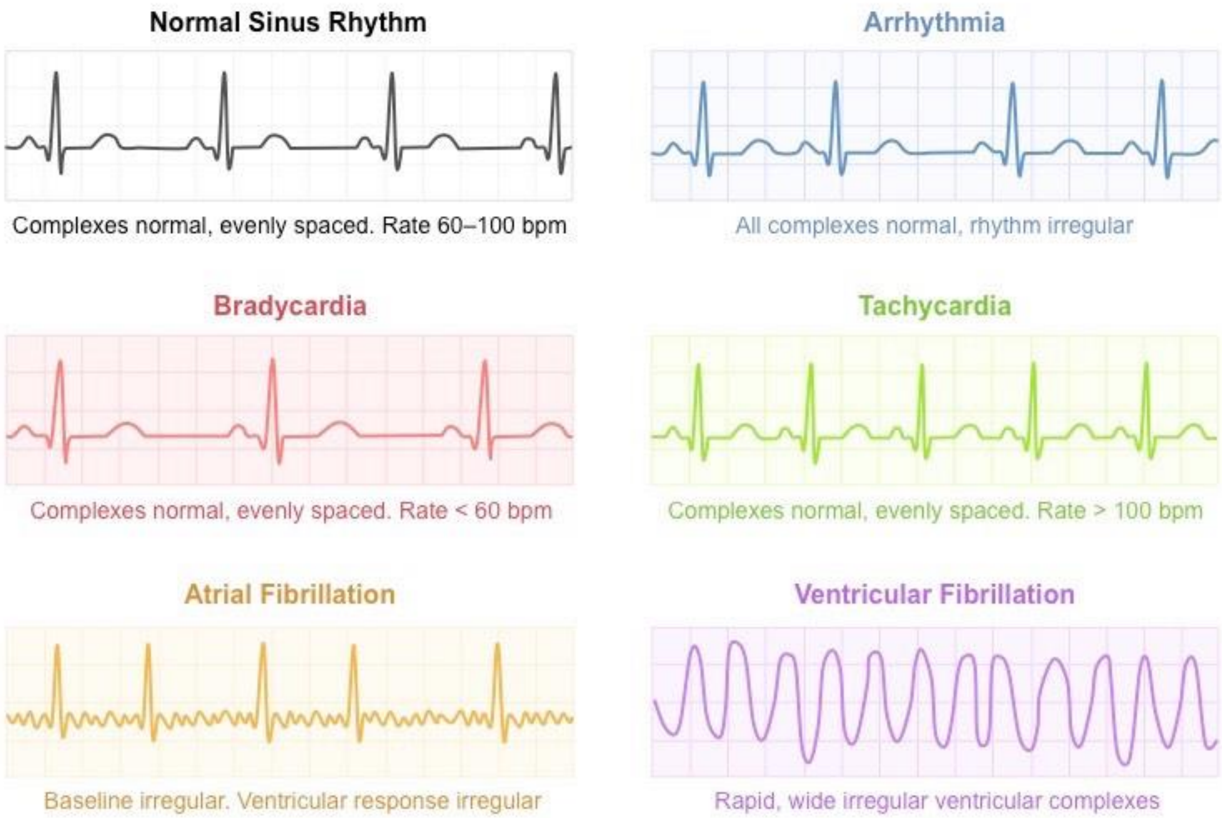
This dot plot shows the distribution for the 47 data points across its range. As you can see around 2000 there are some points there probably not making it as true extreme outlier.

It is up to the user to see whether to classify the three points as an outlier the numerical approach using IQR for upper and lower boundary tells us yes, these points are outliers along with the graphical approach of the box plot. However, the statistical tests (Grubbs and Rosner) tell us otherwise. Depending on what the objective is one could classify these outliers to avoid any potential issues that an outlier may cause or proceed forward with your analysis or remove them or use imputation if you are not satisfied with it.

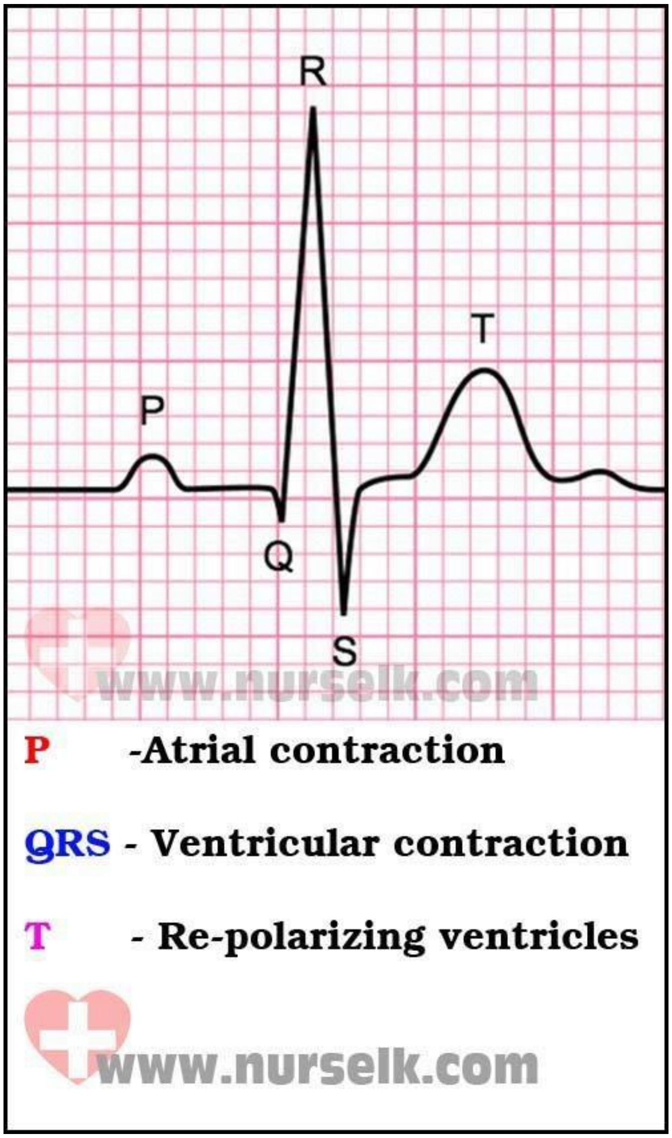
Question 6.1: Describe a change or situation where change detection would be appropriate. Apply CUSUM, how would you choose a critical value and threshold?

A fun explanation or example where CUSUM could be used is detecting irregularities in the heartbeat. More formally, an electrocardiography (ECG) is a test used in the medical field where the medical professionals record a patient’s heart electrical activity. To do this, the medical professional attaches electrodes on the skin of the chest, and the machine will evaluate the electrical activity.

ECG types of rhythms with explanation (Cornell, B)



This graph above with the captions show the different outcomes when you measure with an ECG. The ideal condition for humans is the Normal Sinus Rhythm as explained with the caption below.



The interpretation is as follows:

P is when your atrial one of the heart chambers contracts to get blood in

QRS: is when your ventricles contract to push blood away from the heart

T: think of it as a recharge period where your heart waits until the next flow of blood.

I asked my girlfriend who is an ICU nurse, she said that there are 2 indicators where we should be aware of heart rate and peaks. The peaks the medical professionals count by hand or what the computer evaluates the heartbeat.

In the heartbeat case, we can implement a CUSUM approach because it gives us definitive boundaries to assets the patients by. She said that the there are 2 heart points beats you should be concerned about 130 and 160. So, these two are our lower threshold and upper threshold range (T). Ideally, we want no abnormal detection from a patient, but it happens in the real world.

Regarding the C value, or sensitivity parameter, we want to elect a parameter where above 160 the outcome S_t is above threshold T because at this point the patient is in extreme danger and needs medical attention. Meanwhile, we do not want C to sensitive either where above 130 detects a change thus bring medical care but at 130 the medical profession gives the patient drugs to bring their heart rate down. A good value for C would be like 95% of the time makes upper S_t to trigger a detection change. This could help ease the burden on medical professionals.

Question 6.2.1: Using July through October daily-high-temperature data for Atlanta for 1996 through 2015, use a CUSUM approach to identify when unofficial summer ends (i.e., when the weather starts cooling off) each year. Data, temps.txt or online.

Using the 5-year moving average approach, we can split the data into 4 quarters, provided by office hours but this alone cannot isolate the dates where can approximate summer's end. We will have to use CUSUM with a low T value to indicates that it does not take much for the trigger detection to activate, i.e., summer conditions.

Our approximate ideal model is the Gregorian calendar where summer starts June 21st and ends September 21st; here we must mention that we have omitted variables because we start on July 1st instead of June 21st.

Let us use T as 67 and a C of 0

15-Sep	8	0	5	0	58	5	7	16	17	0	18	8	4	9	0	43	11	0	10	23
16-Sep	10	0	1	2	68	9	2	16	24	0	18	18	6	8	0	57	9	0	8	28
17-Sep	12	0	0	5	79	10	0	17	26	0	17	25	20	16	0	63	11	3	6	28
18-Sep	17	0	2	6	90	10	3	17	28	0	15	31	20	17	0	71	17	6	5	28
19-Sep	22	0	7	10	90	10	6	15	32	0	19	35	22	24	0	79	20	4	5	24
20-Sep	27	0	9	26	87	7	4	9	41	0	29	39	25	24	0	83	26	0	6	19
21-Sep	32	0	11	30	94	3	3	9	48	0	37	47	33	26	0	82	25	9	3	25
22-Sep	35	12	7	42	105	0	10	15	49	0	38	46	31	23	0	87	23	8	4	32
23-Sep	35	13	8	50	108	0	12	16	47	0	35	48	32	21	0	90	27	8	11	35
24-Sep	35	13	11	55	108	13	22	17	47	0	34	45	36	17	0	93	33	19	17	44
25-Sep	31	29	13	58	110	28	37	17	48	0	42	43	37	11	0	90	36	33	24	60
26-Sep	31	40	13	59	126	38	45	16	51	6	47	45	39	7	11	89	35	37	34	73
27-Sep	36	58	11	64	139	44	54	16	60	4	51	45	45	8	17	91	31	40	40	85
28-Sep	44	72	15	68	148	48	58	24	64	2	53	45	41	6	29	97	30	44	50	93
29-Sep	56	76	24	74	159	58	69	40	68	0	66	50	38	15	37	98	31	50	63	99
30-Sep	76	72	33	86	168	69	74	50	70	2	74	56	34	22	48	105	43	55	63	98
1-Oct	94	79	32	97	175	75	75	59	69	2	74	62	43	27	56	126	53	54	61	110
2-Oct	105	87	38	105	180	77	75	73	69	1	76	65	51	33	65	143	66	54	60	127
3-Oct	105	94	45	104	182	78	74	88	71	1	77	73	54	38	85	153	76	54	66	145
4-Oct	119	98	47	117	185	79	76	93	72	0	76	79	53	45	105	159	79	50	85	158
5-Oct	136	98	50	127	187	82	73	96	73	2	73	83	53	64	122	161	82	48	98	168
6-Oct	156	98	61	140	198	93	70	103	80	14	80	87	53	74	136	165	83	46	104	176
7-Oct	180	98	63	150	216	107	74	111	87	25	92	89	54	76	143	171	100	53	106	178
8-Oct	186	97	78	160	245	109	86	120	96	35	103	91	64	77	148	179	122	63	104	179
9-Oct	199	99	91	172	274	125	104	128	106	49	107	96	72	73	150	192	136	69	102	181
10-Oct	211	99	102	183	294	133	117	137	117	55	110	100	77	80	151	209	148	70	99	194
11-Oct	226	98	108	193	307	140	123	152	122	57	113	119	82	84	152	230	157	73	97	204
12-Oct	240	101	115	207	318	144	123	156	131	66	125	132	87	99	160	242	163	74	96	211
13-Oct	251	103	121	226	327	147	127	159	149	72	146	143	93	107	167	253	173	73	95	214
14-Oct	256	117	130	234	336	154	143	163	167	78	160	151	95	122	181	261	180	78	104	219
15-Oct	259	135	135	245	343	160	169	175	187	82	173	158	97	142	190	267	188	92	119	221
16-Oct	262	160	142	251	347	180	187	185	198	89	197	168	96	162	201	268	199	99	133	228
17-Oct	264	176	149	254	352	201	207	198	205	98	209	171	110	192	209	268	208	109	137	241
18-Oct	282	193	155	267	356	219	222	210	213	100	215	179	125	218	218	273	219	118	145	258
19-Oct	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

This tells us that warmer climate is are diminishing because a high T indicates that climate is cooling, and we have values that easily meets the condition. We would see more values between

75 and 85 happen more frequently. We would have to increase it to the maximum of the minimum which is 91 and give it 10 standard deviations for C (See next problem for how it is obtained).

24-Oct	47	74	26	74	33	0	80	0	23	27	78	18	57	73	0	52	6	27	6	5
25-Oct	47	71	25	83	29	0	94	0	19	46	96	31	63	81	1	51	2	43	3	6
26-Oct	46	72	21	87	28	8	100	3	12	59	107	38	65	87	0	49	0	53	0	12
27-Oct	45	86	16	86	24	28	104	11	2	69	114	47	79	98	1	45	5	56	0	29
28-Oct	37	103	11	86	18	45	103	25	0	81	124	52	101	98	4	47	23	64	0	25
29-Oct	29	111	7	87	17	53	101	27	0	90	125	65	115	94	14	63	42	60	1	28
30-Oct	21	116	0	86	14	53	107	21	0	94	125	74	122	99	20	77	61	54	7	31
31-Oct	14	128	0	84	10	53	121	18	0	98	130	78	128	101	22	88	70	52	18	43

This generate values that are cold since it took a lot to generate this condition. Again, a high T in this case will imply it is cold and the opposite is true for warm conditions.

To give an approximate time, it would be the beginning/mid-September to be the end of summer since more CUSUM values are popping up throughout the years.

Question 6.2.2: Use CUSUM method to make a judgement of whether ATL's summer has gotten warmer in that time (if so, then when)

For this problem, I kind of went overboard with the analysis using CUSUM. I originally thought of plotting many plots using values liken to a moving average based on two months, and monthly so I guess MA(2) and MA(4), respectfully. But for the simplicity of the audience, I decided to take an algebraic/calculus approach where we would partition the data into its segments and taking the maximum of the maximum, minimum of maximum, maximum of the minimum, and minimum of the minimum, as shown below

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Overall Max	Overall Min	Mean	SD
Max	99	95	95	99	101	93	97	91	95	94	98	104	95	95	97	99	105	92	95	97	105	91	98	7
Min	60	55	63	57	55	51	57	57	62	54	53	59	50	51	67	59	56	56	63	56	67	50	58.5	8.5
Bi-Month (Max)																								
Max (J/A)	99	93	95	99	101	93	97	91	95	94	98	104	95	95	97	99	105	92	95	97	105	91	98	7
Max (S/O)	91	95	91	96	87	90	95	88	86	91	88	91	90	87	96	96	92	91	92	92	96	86	91	5
Bi-Month (min																								
MIN (J/A)	79	72	79	73	75	81	77	73	75	76	80	79	78	75	83	80	80	66	76	74	83	66	74.5	8.5
MIN (S/O)	60	55	63	57	55	51	57	57	62	54	53	59	50	51	67	59	56	56	63	56	67	50	58.5	8.5
Monthly (Max)																								
July	99	93	95	99	100	91	95	90	95	94	96	95	95	95	97	97	105	91	93	94	105	90	97.5	7.5
August	93	91	93	98	101	93	97	91	91	94	98	104	94	95	96	99	99	92	95	97	104	91	97.5	6.5
September	91	95	91	96	87	90	95	88	86	91	88	91	90	87	96	96	92	91	92	92	96	86	91	5
October	84	82	86	84	82	81	87	84	82	85	86	83	83	85	86	85	83	85	86	82	87	81	84	3
Monthly (Min)																								
July	79	72	80	73	75	82	77	73	81	76	80	79	82	80	83	80	84	76	76	79	84	72	78	6
August	82	73	79	80	81	81	77	81	75	80	80	86	78	75	84	85	80	66	81	74	86	66	76	10
September	64	64	75	68	66	66	69	66	72	77	70	76	69	71	76	69	72	67	71	67	77	64	70.5	6.5
October	60	55	63	57	55	51	57	57	62	54	53	59	50	51	67	59	56	56	63	56	67	50	58.5	8.5
																				AVERAGE	87.7857143	73.857143	80.82143	6.964286
																				MAXIMUM	105	91	98	10
																				MINIMUM	67	50	58.5	3
																				STD DEV	14.5188396	16.218969	15.28273	1.834124

As you can see it is to grasp the condition of the hottest weathers by years, bi-monthly, monthly, then we take the row maximum, minimum, average, and standard deviations then aggregated all again for a final true value of the sample as shown in the highlighted yellow.

We can use this information for a best CUSUM guess for the threshold and critical values for C. For the threshold value let us use the averages of the averages which we take by taking the average of all 4 months by year as shown:

S_t (parameters)																					
mean (mu)	83.7154	81.6748	84.2602	83.3577	84.0325	81.5528	83.5854	81.4797	81.7642	83.3577	83.0488	85.3984	82.5122	80.9919	87.2114	85.2764	84.6504	81.6667	83.9431	83.3008	83.339
std dev (sigma or S)	8.54834	9.31902	6.40931	9.72333	9.51869	8.22452	9.42609	7.01795	6.66294	7.7334	9.79365	9.0334	8.73317	9.01319	7.44516	9.93116	9.25237	7.72654	6.59148	8.70927	8.4406

Threshold is 83.339 or 83 if we are rounding

And the C value is 8.4 or 8 if we are rounding

By Plugging in these values our the our created CUSUM function we get:

10-Oct	71	0	15	23	67	12	20	17	6	13	0	0	0	0	31	24	33	0	0
11-Oct	78	0	13	25	72	10	17	25	3	6	0	10	0	0	24	36	34	0	0
12-Oct	85	0	11	30	75	6	9	20	4	8	4	16	0	7	24	40	32	0	0
13-Oct	87	0	10	42	76	1	6	16	14	6	17	19	0	7	23	44	34	0	0
14-Oct	84	6	11	42	77	0	13	11	24	3	23	19	0	14	29	44	33	0	1
15-Oct	79	16	8	44	76	0	32	16	36	0	28	18	0	26	30	41	33	6	8
16-Oct	75	33	6	43	72	12	41	17	39	0	44	19	0	38	33	34	36	5	14
17-Oct	68	41	6	38	68	25	53	23	38	0	48	15	6	60	33	27	37	7	10
18-Oct	78	50	4	42	64	35	60	26	38	0	46	15	13	78	34	24	40	8	10
19-Oct	91	55	0	52	67	37	65	25	44	0	45	17	23	90	31	34	43	18	13
20-Oct	98	59	1	67	70	36	68	20	47	0	51	19	31	95	33	53	49	22	16
21-Oct	95	62	5	78	72	30	72	10	48	0	57	18	34	97	32	67	50	24	15
22-Oct	90	74	18	81	69	24	84	5	48	8	63	24	40	96	34	76	47	28	20
23-Oct	97	85	31	99	70	16	100	0	52	14	85	20	53	97	37	83	46	39	24
24-Oct	99	96	35	116	71	11	108	1	53	33	104	39	73	101	35	88	44	47	26
25-Oct	102	95	37	127	69	12	124	1	49	54	124	55	81	109	38	89	40	64	25
26-Oct	103	98	34	133	70	21	131	7	44	69	137	64	85	117	33	89	37	77	17
27-Oct	103	114	31	134	68	44	138	16	36	81	146	75	101	130	36	87	44	82	9
28-Oct	98	133	28	136	64	62	138	33	32	95	158	82	125	132	41	91	64	91	8
29-Oct	92	143	27	139	65	73	139	36	31	106	161	98	141	130	52	109	85	90	11
30-Oct	86	150	21	140	64	74	146	33	27	112	163	108	150	137	60	125	106	86	19
31-Oct	80	164	18	140	62	77	162	31	18	118	170	114	158	141	65	138	117	85	32

What this tells us is that in ATL, the summers are not as long when you have a high threshold. A high T indicates more cooling conditions in the state and seeing the average it makes sense that temperature decreases in the fall which is past September. Over the years, there is no indication that summer has increased like longer time, but I will mention that winter conditions is occurring more often through the span of the years.

References:

Cornell, B. (n.d.). *Electrocardiography*. BioNinja. Retrieved from https://ib.bioninja.com.au/standard-level/topic-6-human-physiology/62-the-blood-system/electrocardiography.html