ISYE 6501 Intro Analytics Modeling – HW3

Question 5.1 Using crime data from the file `uscrime.txt` 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.

The first step is to read in the dataset and do some simple data summary for the last column. From the code, I get to know this dataset contains **47 data points**. The last column has minimal value **324**, maximum value **1,993**, and average value **905.1**. Use the `Boxplot` function in R to plot the data. As we can see, it identified three outliers with the maximum values (Figure 1).

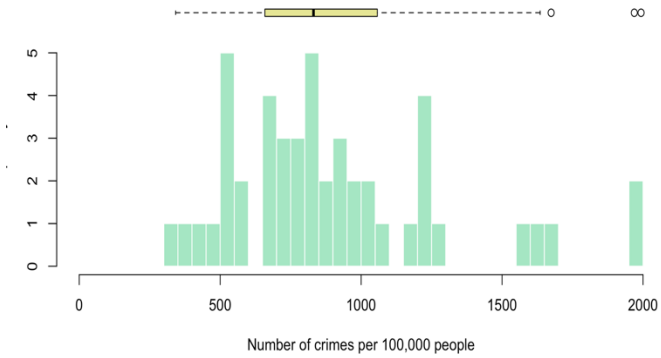

Fig.1

Then I used function `grubbs.test` in the `outliers` package, the outcome shows the maximum value 1993 is the outlier with p-value 0.079, which is not significant. The same with the G value by comparing to the G Critical Value(http://www.sediment.uni-goettingen.de/staff/dunkl/software/pep-grubbs.pdf).

But from the definition of Grubbs' test, its hypothesis is defined as **Ha: There is exactly one outlier in the data set.** But obviously. from the histogram, there is more than one outlier in the dataset. When I change the type to 20 to detect if the dataset contains two outliers on the same tail. I got the error because of the sample size: `Error in qgrubbs(q, n, type, rev = TRUE) : n must be in range 3-30`. By using `chisq.out.test`, I got P-value 0.00491, which is significant at 0.005 level. I assume we get the p-value 0.079, because there is another data point very close to the maximum value.

To test if there is another outlier besides the maximum value, I used `grubbs.test` on the dataset after removing the highest value 1,993. This time I got a P-value 0.028 which is significant at 0.05 level, that the value 1,969 is an outlier. By keep trying this, we get p-value above 0.1 for the three values 1,674, 1,635, and 1,555. And finally, we got p-value = 1 after removing all the values mentioned above.

Overall, I think `grubbs.test` is not very reliable for the dataset with multiple outliers and close to each other. I think boxplot or histogram is very straightforward to visualize outliers, which can work as good references for other automated tools.

```
Grubbs test for one outlier
data:  df$Crime
G = 2.81290, U = 0.82426, p-value = 0.07887
alternative hypothesis: highest value 1993 is an outlier

crime2<-rm.outlier(df$Crime)
> grubbs.test(crime2)
Grubbs test for one outlier
data:  crime2
G = 3.06340, U = 0.78682, p-value = 0.02848
alternative hypothesis: highest value 1969 is an outlier
```

Question 6.1 Describe a situation or problem from your job, everyday life, current events, etc., for which a Change Detection model would be appropriate. Applying the CUSUM technique, how would you choose the critical value and the threshold?

Every month, I will be received data of last month's consumer eligibility, I think the CUSUM technique can be used there to automate the data quality control and monitored the change. That would be very easy to check the decrements and increments of member amount, also easier to find out any data issue.

In this case, member enrollment is very time sensitive, I hope the model to detect change early so that the company can launch some campaign to solve the issue in time. For example, in the first month there are 1,000,000 eligible members, and decreasing at the rate 1,000 people/month. I will set the C to 0, so the model can be sensitive. At month 6th we will lose 5,000 people which is 0.5% of the initial member amount, I would like to detect the change before that, by calculation the S_5 =5000, so I will set the T value of 5,000.

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. You can get the data that you need from the file `temps.txt` or online. You can use R if you'd like, but it's straightforward enough that an Excel spreadsheet can easily do the job too.

I used the T=250 C=4 for the CUSUM approach, green and red color indicates if the St score greater than T value or not (Figure 2). As we can see, using this combination, the summer ends from 11-Sep (2012) to 22-Oct (2013).

DAY	St_1996	St_1997	St_1998	St_1999	St_2000	St_2001	St_2002	St_2003	St_2004	St_2005	St_2006	St_2007	St_2008	St_2009	St_2010	St_2011	St_2012	St_2013	St_2014	St_2015
7-Sep	169.419	41.9948	94.848	11.329	166.883	47.5461	46.7492	26.167	97.7312	35.2093	109.192	46.3968	102.229	87.458	44.8245	67.0421	233.252	14.7537	0	78.6575
8-Sep	169.477	40.4948	93.9623	10.1433	181.654	47.9032	48.9206	30.9813	110.617	36.895	115.521	49.6968	99.7435	86.9008	42.9245	85.5993	235.623	9.02518	3.55714	83.8432
9-Sep	167.561	42.9596	102.948	22.7771	191.288	48.1568	47.1178	34.7418	112.476	37.5711	121.76	51.3728	99.2646	89.3093	42.0372	96.0077	242.891	4.36321	7.06419	85.9981
10-Sep	172.575	49.3346	112.795	25.3743	198.816	45.4484	41.3956	41.4085	114.309	40.21	125.941	53.0256	99.7785	95.6287	39.19	103.313	249.071	0.7521	8.5503	90.0954
11-Sep	175.548	53.6497	119.549	27.9359	204.268	44.7498	35.7518	46.9975	116.118	42.8127	131.05	54.6558	104.231	106.793	34.4091	108.546	256.154	0	6.07084	93.1502
12-Sep	176.507	53.9605	123.252	30.463	207.673	47.0201	35.1166	51.5246	120.861	43.4073	138.064	59.2233	104.676	111.888	36.5983	112.722	265.113	0	1.65193	104.056
13-Sep	187.32	53.2805	121.972	33.943	210.046	49.2601	49.29	53.0312	127.514	42.0207	156.824	64.7167	106.103	117.901	41.7183	114.868	273.953	0	3.21193	117.776
14-Sep	197.005	52.6094	122.682	36.3903	213.375	49.4969	60.3163	56.4918	134.08	40.6522	165.468	74.0851	105.537	121.862	41.8368	114.026	280.703	3.47368	12.6461	129.341
15-Sep	199.654	50.9601	123.383	36.8319	222.583	58.6138	58.3682	59.9073	146.483	36.3405	172.027	82.3448	113.862	128.732	41.9536	119.117	283.417	4.92823	19.9838	137.796
16-Sep	206.218	49.3319	123.089	45.1652	236.609	67.6138	59.4067	64.2663	156.752	35.0457	177.514	96.4217	121.093	132.552	44.0434	137.963	287.084	1.42823	21.3043	147.129
17-Sep	212.699	45.7496	123.785	53.3931	252.432	72.5505	62.4067	68.5701	161.954	35.7419	180.957	108.346	139.093	145.21	40.1826	149.66	293.666	7.84595	22.6081	152.395
18-Sep	223.049	44.1871	129.41	60.5306	268.057	77.4255	70.3067	72.8201	168.079	34.4544	184.357	119.133	144.03	150.798	38.3451	163.185	304.116	15.171	24.8831	157.595
19-Sep	232.284	39.6809	136.941	70.5429	272.625	81.2527	79.0968	74.0547	175.116	31.2075	193.641	126.825	149.895	162.242	37.518	176.543	312.462	16.4796	28.1177	158.78
20-Sep	241.406	31.2785	142.404	91.299	274.173	83.0575	81.8529	72.3108	187.007	28.0001	208.739	135.41	157.663	166.633	32.7497	184.799	322.682	15.7967	32.3006	157.975
21-Sep	251.406	28.905	147.802	100.938	285.583	82.8648	86.5517	75.5276	196.778	24.8315	221.678	146.856	169.289	172.946	29.0268	189.004	326.85	27.9653	32.4813	169.036
22-Sep	258.322	45.3335	147.206	117.378	300.809	81.6862	98.1112	85.6229	201.492	25.6529	227.547	150.261	171.885	174.243	28.3125	199.087	328.993	31.0963	37.6004	180.952
23-Sep	262.193	51.6865	150.571	130.66	307.95	79.5333	104.594	89.6699	202.198	25.4764	229.394	156.59	176.426	176.514	28.5948	206.087	337.04	34.1904	47.6004	187.788
24-Sep	266.018	55.9888	156.862	140.823	312.043	96.1844	119.896	94.6583	204.872	24.3136	233.197	156.916	184.868	176.782	31.839	213.994	347.959	48.1206	56.4957	201.462
25-Sep	266.834	76.0578	162.092	147.904	318.066	115.61	138.976	97.6123	208.505	26.1297	245.852	159.214	189.258	175.069	30.1034	215.879	354.798	65.8448	66.2773	221.899
26-Sep	270.607	91.9441	165.285	153.915	337.862	128.882	151.908	100.533	214.073	35.8343	255.397	164.453	195.577	175.353	45.1943	219.72	357.605	72.4925	78.9137	238.149
27-Sep	279.281	113.585	165.476	163.814	354.469	139.04	164.695	103.42	226.5	36.5309	263.847	168.644	204.79	180.578	55.1718	225.495	358.402	78.0768	87.453	254.216
28-Sep	291.815	131.029	172.587	171.625	366.935	146.118	173.384	115.176	232.855	38.2087	270.225	172.789	205.001	182.778	69.9829	236.151	361.169	85.5768	99.853	266.15
29-Sep	307.177	139.381	184.565	182.317	381.243	160.041	187.911	133.725	239.141	39.868	287.411	180.843	205.21	195.833	81.6643	241.744	365.883	93.9834	115.084	275.974
30-Sep	330.286	138.739	196.413	198.828	393.417	173.812	196.346	147.128	243.38	44.4767	299.466	190.789	205.416	206.767	96.186	253.212	382.416	101.309	117.29	277.778
1-Oct	351.168	148.987	197.252	213.183	403.481	183.479	201.722	159.396	245.595	48.0466	303.476	199.639	217.492	215.606	107.584	277.416	395.803	103.611	117.495	293.412
2-Oct	365.891	161.104	205.996	225.406	411.46	188.096	205.063	175.492	247.787	49.5998	309.423	206.415	229.439	225.34	119.85	298.395	412.016	105.887	118.686	313.827
3-Oct	368.586	171.114	215.638	228.596	416.386	191.675	208.368	193.397	251.934	53.1156	314.318	218.068	235.323	233.982	141.881	312.226	426.079	108.14	126.791	334.027
4-Oct	385.107	178.052	220.232	244.617	422.251	196.206	213.618	200.231	255.049	54.6156	317.183	226.63	238.177	244.514	164.673	320.966	432.079	107.4	147.677	350.059
5-Oct	405.416	181.949	224.778	258.493	427.065	201.68	213.865	206.004	259.121	60.0589	318.038	234.115	241.002	266.812	184.261	326.646	438.017	107.658	162.409	362.955
6-Oct	427.498	184.816	238.187	274.207	440.739	216.007	215.1	214.688	268.1	74.3548	328.783	240.533	244.787	279.976	200.679	334.248	442.905	107.913	170.062	372.751
7-Oct	453.316	187.654	242.55	287.783	461.204	232.169	222.262	225.263	276.989	88.5063	343.379	245.897	248.535	285.087	209.012	343.753	462.592	118.065	173.678	377.499
8-Oct	461.056	190.464	259.74	301.223	492.354	237.279	238.262	235.733	287.769	101.526	357.829	251.207	262.145	289.157	216.272	354.153	487.032	130.095	173.298	381.209
9-Oct	476.64	195.227	273.79	315.52	523.195	255.209	259.054	246.1	300.423	117.388	365.205	258.444	272.65	288.236	220.49	370.391	504.299	138.045	172.922	385.872
10-Oct	490.091	197.962	286.711	329.677	544.823	266.033	275.681	256.364	312.952	126.162	371.519	265.611	281.071	298.216	223.676	390.43	518.427	141.957	172.549	400.392
11-Oct	506.382	200.671	294.556	342.706	559.308	274.771	284.225	273.462	319.418	129.9	377.772	286.572	288.421	305.129	226.832	413.246	530.436	146.821	172.18	412.79
12-Oct	522.517	206.325	302.325	358.581	571.673	280.453	286.744	278.51	329.783	141.525	392.877	302.38	296.69	322.869	235.899	427.919	538.369	149.657	172.805	422.098
13-Oct	534.536	210.934	310.02	380.248	581.94	286.082	294.192	283.51	348.964	150.068	416.753	316.056	304.881	333.508	244.88	441.462	550.188	151.476	173.424	426.364
14-Oct	540.498	227.387	320.615	390.814	592.11	294.629	312.465	288.462	368.955	157.54	433.47	325.641	310.022	350.979	260.71	451.905	559.914	158.231	183.943	433.562
15-Oct	544.423	248.639	327.147	403.262	600.203	303.096	341.465	302.285	389.749	163.951	449.041	335.136	314.125	373.241	271.439	459.279	569.55	173.839	200.308	437.721
16-Oct	549.303	275.639	334.61	411.633	605.25	324.365	361.28	313.007	401.435	173.275	475.365	346.525	316.208	395.296	284.05	462.621	582.068	183.357	215.53	445.805
17-Oct	552.156	293.474	342.996	416.954	610.25	347.42	382.895	327.594	409.049	183.504	489.557	351.865	332.144	427.058	292.582	464.942	593.481	194.77	220.704	460.75
18-Oct	570.838	313.129	350.314	431.145	615.204	366.302	400.349	340.067	418.577	187.695	497.675	362.11	349.917	454.567	303.019	472.197	605.781	205.088	229.795	479.522
19-Oct	592.325	327.651	353.602	450.163	627.051	378.077	414.674	347.472	432.973	189.866	506.712	373.254	369.511	475.882	309.397	492.269	618.961	225.223	241.777	500.108
20-Oct	608.664	341.053	363.799	474.957	638.792	385.782	426.888	351.838	444.268	192.018	522.604	385.29	387.94	490.07	320.674	520.09	634.006	238.241	253.652	513.572
21-Oct	613.956	354.336	375.887	495.568	648.447	389.454	440.976	350.219	453.48	197.124	538.357	394.246	400.259	501.158	328.878	543.701	643.962	249.162	261.457	521.962
22-Oct	617.219	375.433	397.782	507.077	654.052	392.104	461.88	354.561	462.612	215.072	552.979	409.07	416.434	509.176	338.992	562.148	650.857	262.96	276.133	527.304
23-Oct	632.349	395.354	419.486	534.346	663.574	393.738	487.558	358.865	474.638	229.889	584.327	414.844	438.417	519.106	351	578.453	658.683	282.586	288.698	530.617
24-Oct	643.384	416.096	432.081	559.398	673.014	397.342	504.093	368.09	483.586	258.458	612.431	442.379	468.141	531.925	357.949	591.642	665.45	299.069	299.172	538.858
25-Oct	654.324	423.771	441.594	579.278	679.398	406.863	528.417	377.235	487.5	288.766	641.285	467.695	484.722	548.601	369.795	601.745	671.168	325.325	306.582	551.986
26-Oct	663.189	435.347	447.06	594.033	688.703	425.228	544.604	391.26	490.39	311.876	662.955	485.856	498.188	565.134	372.617	610.77	676.837	346.401	306.989	568.969
27-Oct	671.978	460.709	452.48	602.713	694.955	456.328	559.663	409.134	491.273	332.809	680.476	504.856	522.448	586.487	384.339	617.737	692.375	359.368	307.393	596.717
28-Oct	674.745	487.842	457.855	613.305	699.172	483.203	568.646	433.801	495.123	354.559	700.826	520.723	555.432	596.754	397.948	629.604	721.667	377.184	314.734	602.417
29-Oct	676.497	505.825	464.177	624.801	708.313	501.922	577.555	445.371	501.916	374.145	712.082	544.392	579.217	602.969	417.394	656.248	750.716	383.945	325.982	616.004
30-Oct	678.235	521.678	466.481	633.227	715.395	511.562	593.334	449.904	505.678	388.613	722.254	562.909	596.856	618.059	433.706	680.691	779.528	387.674	342.097	629.479
31-Oct	680.95	543.353	471.741	641.585	721.427	522.115	616.919	456.384	505.443	401.97	737.303	577.307	612.368	630.051	445.917	700.967	799.178	395.341	363.04	650.78

Fig.2

2. Use a CUSUM approach to make a judgment of whether Atlanta's summer climate has gotten warmer in that time (and if so, when).

Using the defined summer days from the question1, I created two measurements to judge if the summer climate has gotten warmer. The first is the number of summer days, using c=0 and t=20, starting from 2013, it indicates the summer is getting longer. In terms of avg summer temperature, in 2011 it shows increment but after 2014 the CUSUM model does not see it keep increasing anymore (Figure 3).

Year	Summer Day	St_Days
1996	97	0
1997	107	5
1998	99	3
1999	96	0
2000	78	0
2001	100	3.833333333
2002	100	7.119047619
2003	99	9.119047619
2004	95	7.341269841
2005	79	0
2006	87	0
2007	99	4.333333333
2008	99	8.333333333
2009	96	9.261904762
2010	105	18.52857143
2011	91	14.09107143
2012	72	0
2013	113	17.88888889
2014	229	144.7309942
2015	221	257.6309942

Year	Avg_Summer_Temp	St_avg_temp
1996	86.30927835	0
1997	84.25233645	0
1998	86.36363636	0.72188598
1999	87.02083333	1.75619819
2000	89.20779221	4.33321505
2001	84.11	2.23256894
2002	87	2.90915798
2003	83.57575758	0.50496127
2004	84.14736842	0
2005	86.69620253	0.82788201
2006	87.65517241	2.45229282
2007	88.36363636	4.59076135
2008	85.74747475	4.14981388
2009	84.53125	2.61101111
2010	86.55619048	3.06473964
2011	89.59340659	6.33750025
2012	90.18055556	9.97035631
2013	84.03834808	7.60041331
2014	83.94308943	5.26495911
2015	83.30081301	2.43611503

Fig.3

```
# ISYE 6501 Intro Analytics Modeling - HW3
# IP uscrime.txt

# Loading and examining data
df<-read.delim("uscrime.txt", header = TRUE, sep = "\t")
#### Display Head Lines ####
head(df,2)
#### Show summary, number of row of last column##
summary(df$Crime)
nrow(df)
#Question 5.1  Using crime data from the file uscrime.txt test
#to see whether there are any outliers in the last column (number of crimes p
er 100,000 people).
#Use the grubbs.test function in the outliers package in R.

#### Boxplot and Histgram ####

boxplot(df$Crime)
# Layout to split the screen
layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,8))
# Draw the boxplot and the histogram
par(mar=c(0, 3.1, 1.1, 2.1))
boxplot(df$Crime , horizontal=TRUE, xaxt="n",ylim=c(0,2000),col=rgb(0.8,0.8,0
,0.5) , frame=F)
```

```r
par(mar=c(4, 3.1, 1.1, 2.1))
hist(df$Crime,breaks=25,col=rgb(0.2,0.8,0.5,0.5) , border=F , main="" ,xlab="
Number of crimes per 100,000 people",xlim=c(0,2000))

#### Library outliers####
library(outliers)

grubbs.test(df$Crime)

crime2<-rm.outlier(df$Crime)
grubbs.test(crime2)

crime3<-rm.outlier(crime2)
grubbs.test(crime3)

crime4<-rm.outlier(crime3)
grubbs.test(crime4)

crime5<-rm.outlier(crime4)
grubbs.test(crime5)

crime6<-rm.outlier(crime5)
grubbs.test(crime6)
```