Homework 3

2025-01-27

Question 5.1

Using crime data from the file uscrime.txt (http://www.statsci.org/data/general/uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), 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.

Answer:

To begin the process of identifying the outliers, I loaded the data to become familiar with the dataset. By doing a preliminar overlook of the statistics with the summary function, I can see that the *maximum value* is **1993** while the *mean* is **905.1** which can give us an idea of the outliers and where they can be located in the quartiles within the dataset. It does seem like we might have values in that could be potential outliers.

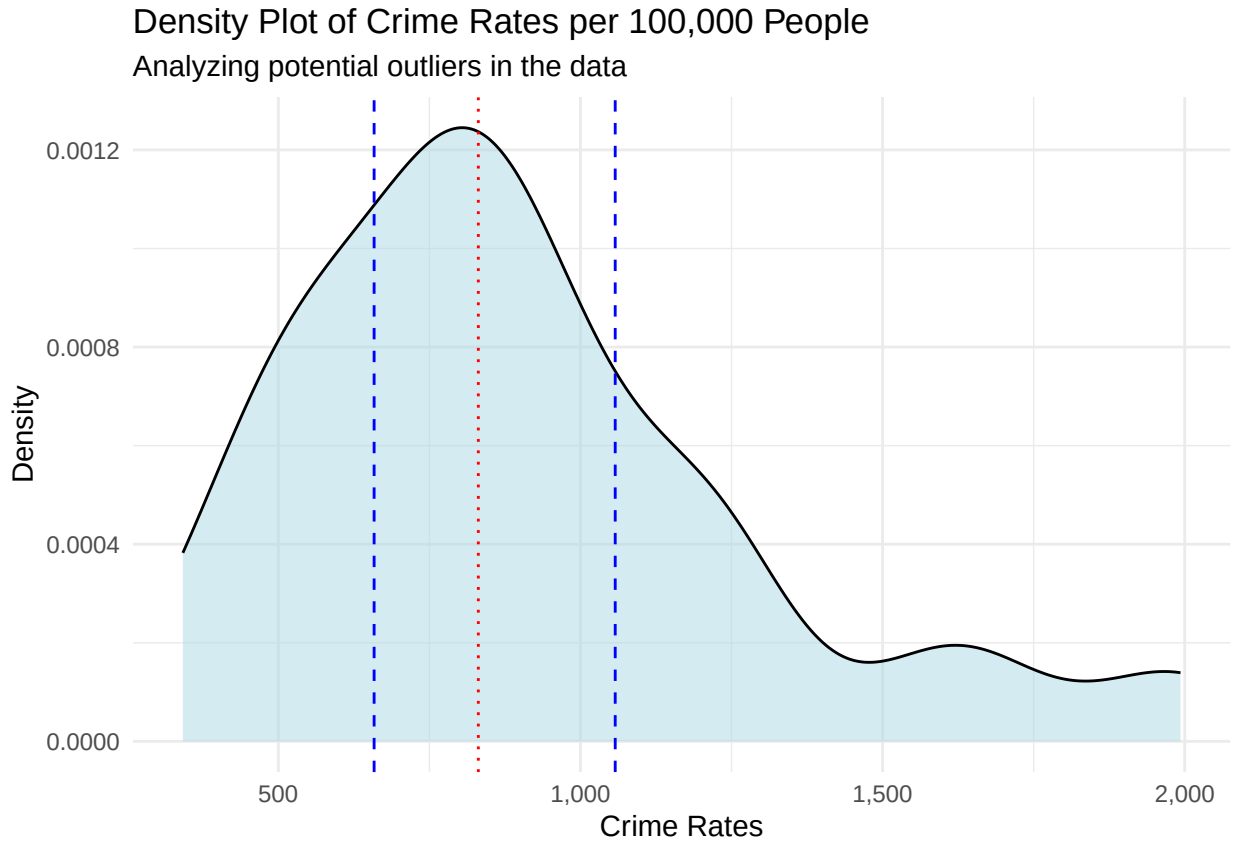
```
# Summary for crime rates
print(crime_summary <- summary(crime_rates))
```

```
##      Crime
##  Min.   : 342.0
##  1st Qu.: 658.5
##  Median : 831.0
##  Mean   : 905.1
##  3rd Qu.:1057.5
##  Max.   :1993.0
```

Next, I plotted the quartiles to further analyze possible outliers.

```
# Calculate skewness and kurtosis
crime_skewness <- skewness(crime_rates$Crime)
crime_kurtosis <- kurtosis(crime_rates$Crime)

# Detect outliers using the `outlier` function
extreme_outlier <- outlier(crime_rates$Crime)

# Plot density to visualize outliers
ggplot(crime_rates, aes(x = Crime)) +
  geom_density(fill = 'lightblue', alpha = 0.5) +
  theme_minimal() +
  scale_x_continuous(labels = scales::comma) +
  labs(
    title = 'Density Plot of Crime Rates per 100,000 People',
    subtitle = 'Analyzing potential outliers in the data',
    x = 'Crime Rates',
```

```
    y = 'Density'
  ) +
  geom_vline(
    xintercept = quantile(crime_rates$Crime, 0.25),
    color = 'blue',
    linetype = 'dashed'
  ) +
  geom_vline(
    xintercept = median(crime_rates$Crime),
    color = 'red',
    linetype = 'dotted'
  ) +
  geom_vline(
    xintercept = quantile(crime_rates$Crime, 0.75),
    color = 'blue',
    linetype = 'dashed'
  )
```




```
# Display initial results
cat('Summary of Crime Rates:\n', crime_summary, '\n')
```

```
## Summary of Crime Rates:
##  Min.   : 342.0   1st Qu.: 658.5   Median : 831.0   Mean   : 905.1   3rd Qu.:1057.5   Max.   :1993.0
```

```r
cat('Skewness: ', crime_skewness, '\n')
```

```
## Skewness:  1.08848
```

```r
cat('Kurtosis: ', crime_kurtosis, '\n')
```

```
## Kurtosis:  3.943658
```

```r
cat('Extreme Outlier Detected: ', extreme_outlier, '\n')
```

```
## Extreme Outlier Detected:  1993
```

As we can see in the density plot it shows a *right-skewed distribution*, a positive *skewness* of **1.08848** indicates that the distribution of the data is slightly right-skewed. This means there are a few larger values pulling the mean toward the right, but it's not an extreme skew. A *kurtosis* of* **3.94** suggests the data has heavier tails than a normal distribution, meaning there are more extreme outliers than what would be expected in a normal distribution.

Now, to answer the question I will use the grubss.test function in the outliers package in R.

```r
# Perform Grubbs' test for detecting outliers
grubbs_result <- grubbs.test(
  crime_rates$Crime,
  type = 10,
  two.sided = TRUE
)

# Extract and print relevant elements of the result
cat('Grubbs Test Result:\n')
```

```
## Grubbs Test Result:
```

```r
cat('Test Statistic: ', grubbs_result$statistic, '\n')
```

```
## Test Statistic:  2.812874 0.824255
```

```r
cat('P-value: ', grubbs_result$p.value, '\n')
```

```
## P-value:  0.1577497
```

```r
cat('Alternative Hypothesis: ', grubbs_result$alternative, '\n')
```

```
## Alternative Hypothesis:  highest value 1993 is an outlier
```

As a result of the Grubbs Test, we obtained a *p-value* of **0.1577497** which can be a insignificant against the *threshold* of **0.05**. Since **0.1577 > 0.05** you fail to reject the *null hypothesis*. There is insufficient evidence to conclude that the most extreme value in your dataset is an outlier. The value is not statistically significant as an outlier at the 5% level.

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?

Answer:

I am currently working a part-time job as a barista. One example where the CUSUM technique can be applied is to monitor the average coffee preparation time during peak hours by detecting small but sustained shifts in the process mean. Starting with a baseline average of 3 minutes, CUSUM calculates cumulative deviations from this reference value for each observed preparation time. By doing this, CUSUM can provide early warnings of inefficiencies or changes in the process, enabling the barista team to address issues like equipment maintenance or workflow adjustments proactively, ensuring consistent service quality.

Relevant predictors I would use include:

- Order Type: Whether it's a basic coffee (e.g., black coffee) or a more complex drink (e.g., latte, cappuccino).
- Time of Day: Morning rush, midday lull, evening.
- Staffing Levels: Number of baristas working during the measurement period.
- Experience Level: Average experience (in months/years) of the baristas on the shift.
- Equipment Used: Whether the espresso machine, grinder, or other tools were involved.

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, for example at http://www.iweathernet.com/atlanta-weather-records or https://www.wunderground.com/history/airport/KFTY/2015/7/1/CustomHistory.html . You can use R if you'd like, but it's straightforward enough that an Excel spreadsheet can easily do the job too.

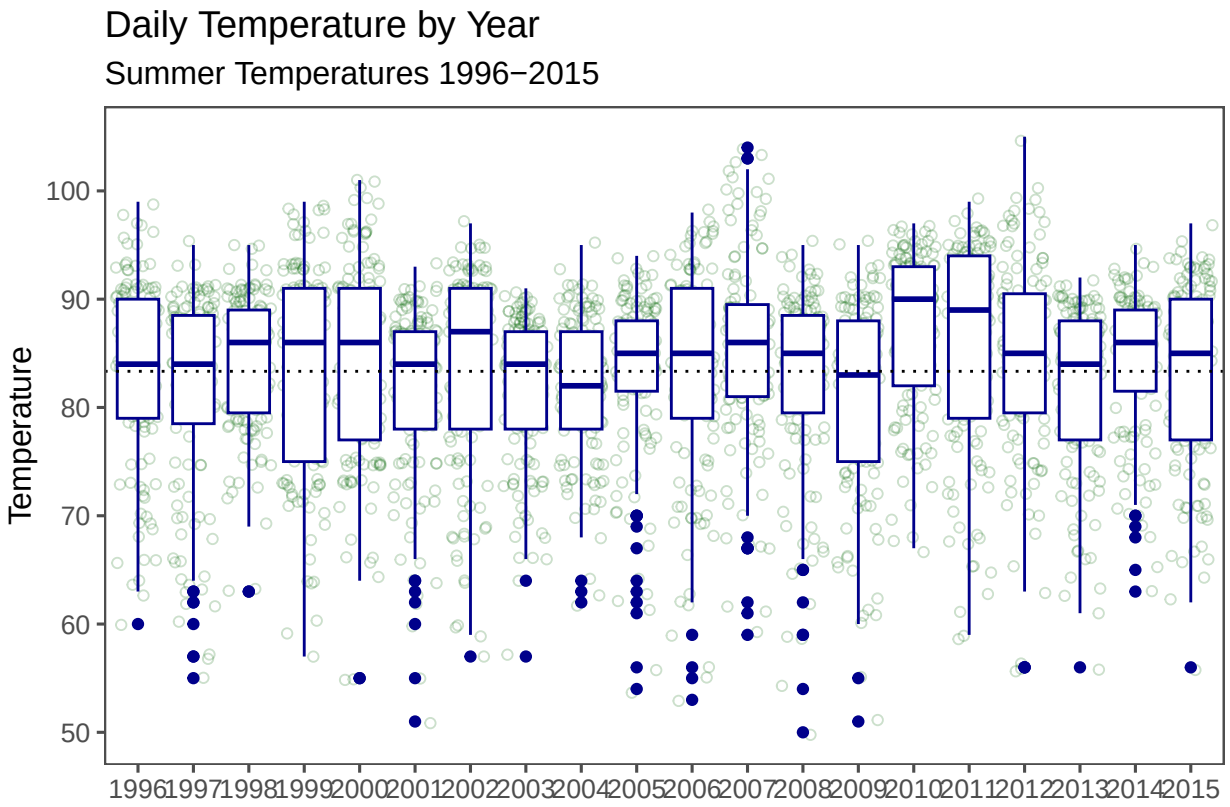
Answer 1:

```
# GTx_6501 Homework 3

# Workflow

# Best practice
rm(list = ls())

# Set up directory and packages
setwd('~/Desktop/GTX/Homework 3/')
# Load packages
pacman::p_load(tidyverse, kernlab, caret, kknn, outliers,modelr, ggthemes, corrplot,moments)

# Load the outliers library
library(readr)
```

```r
library(dplyr)
library(outliers)
library(ggplot2)
library(scales)
library(tidyr)
library(moments)
library(lubridate)

## Data manipulation
temps = read_delim('/Users/cn/Desktop/GTX/Homework 3/temps.txt', delim = '\t') %>%
  as_tibble() %>%
  gather(year, temp, -DAY) %>%
  mutate(year = as.factor(year),
         date = paste(DAY,year, sep = '-')) %>%
  mutate(date_val = dmy(date),
         color = ifelse(temp > mean(.$temp), 'Above', 'Below'),
         month = month(date_val),
         day = day(date_val)) %>%
  dplyr::select(date_val, DAY, year, temp,color, month, day)
```

```
## Rows: 123 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (1): DAY
## dbl (20): 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, ...
##
## 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.
```

```r
# Visualize trends
ggplot(data = temps, aes(x = year, y = temp)) +
  geom_jitter(pch = 21, alpha = .2, color = 'dark green') +
  geom_boxplot(color = 'dark blue') +
  theme_few() +
  theme(legend.position = 'none') +
  geom_hline(yintercept = mean(temps$temp), linetype = 'dotted') +
  xlab('') +
  ylab('Temperature') +
  labs(title = 'Daily Temperature by Year',
       subtitle = 'Summer Temperatures 1996-2015')
```

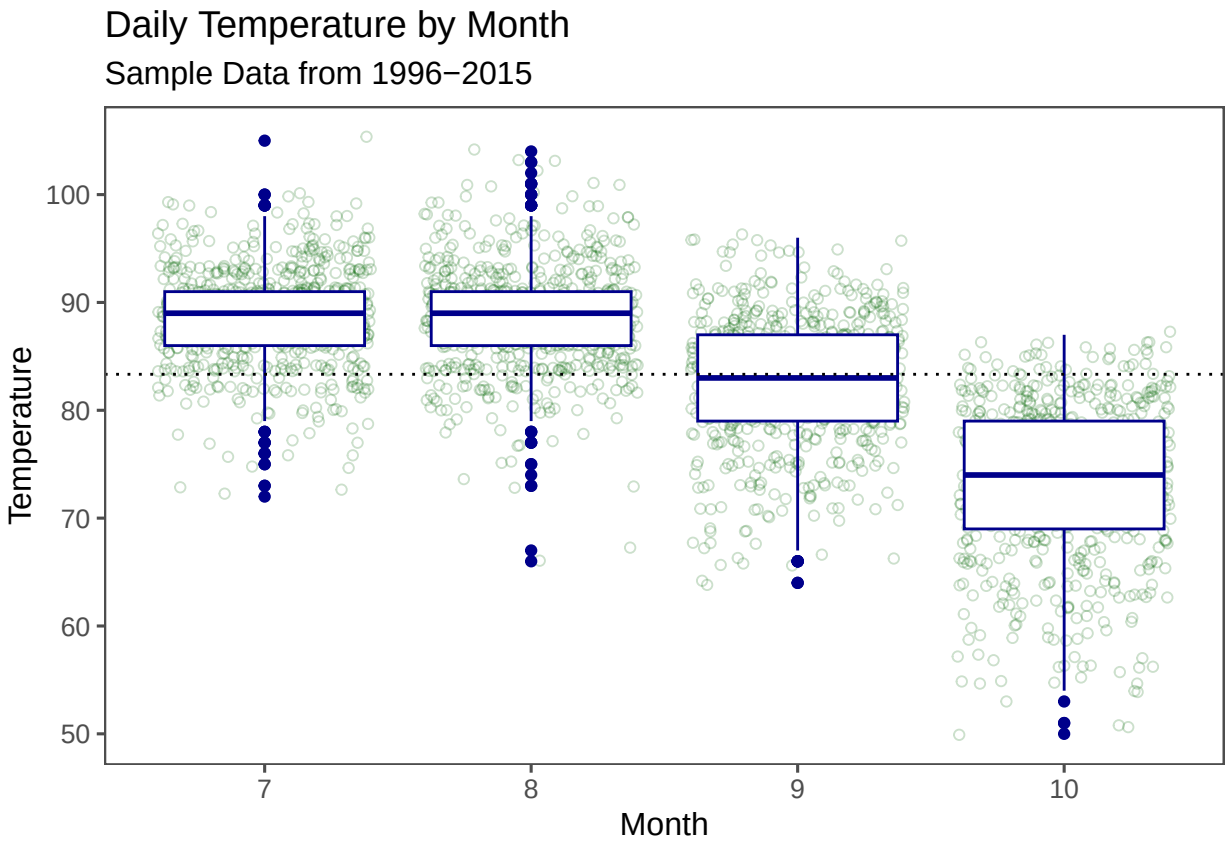


This graph illustrates the yearly distribution of daily high temperatures during the summer months (July to October) from 1996 to 2015. Across all years, median temperatures remain relatively stable, hovering around 85°F to 90°F, with a mean summer temperature of approximately 85°F (marked by the dotted line). This indicates consistent summer weather patterns over the 20-year period.

There are notable outliers, especially in September and October, where daily high temperatures occasionally drop below 70°F.

While the graph shows overall consistency during peak summer months (July and August), the presence of outliers in the later months aligns with the seasonal cooling trend observed in the first graph. Identifying when these outliers cluster below a specific threshold, such as 80°F, can help determine the point when summer ends for each year.

```
# Visualize trends
ggplot(data = temps, aes(x = as.factor(month), y = temp)) +
  geom_jitter(pch = 21, alpha = .2, color = 'dark green') +
  geom_boxplot(color = 'dark blue') +
  theme_few() +
  theme(legend.position = 'none') +
  geom_hline(yintercept = mean(temps$temp), linetype = 'dotted') +
  xlab('Month') +
  ylab('Temperature') +
  labs(title = 'Daily Temperature by Month',
       subtitle = 'Sample Data from 1996-2015')
```



The second graph shows the distribution of daily high temperatures from July through October, highlighting when the weather starts cooling off. In July and August, median temperatures are consistently high, around 90°F, with an interquartile range (IQR) between 85°F and 95°F, indicating stable and warm weather.

Outliers below 80°F are rare during these months, showing that temperatures seldom drop significantly during peak summer.

By October, the cooling trend becomes even more pronounced, with the median temperature dropping further to 75°F and the IQR shifting to 65°F to 80°F. This sharp decrease is accompanied by more frequent outliers below 70°F, signaling that the unofficial summer typically ends by mid to late September as cooler temperatures dominate in October.

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).

Answer 2:

```
# Load necessary library
library(qcc)
```

```
## Package 'qcc' version 2.7
```

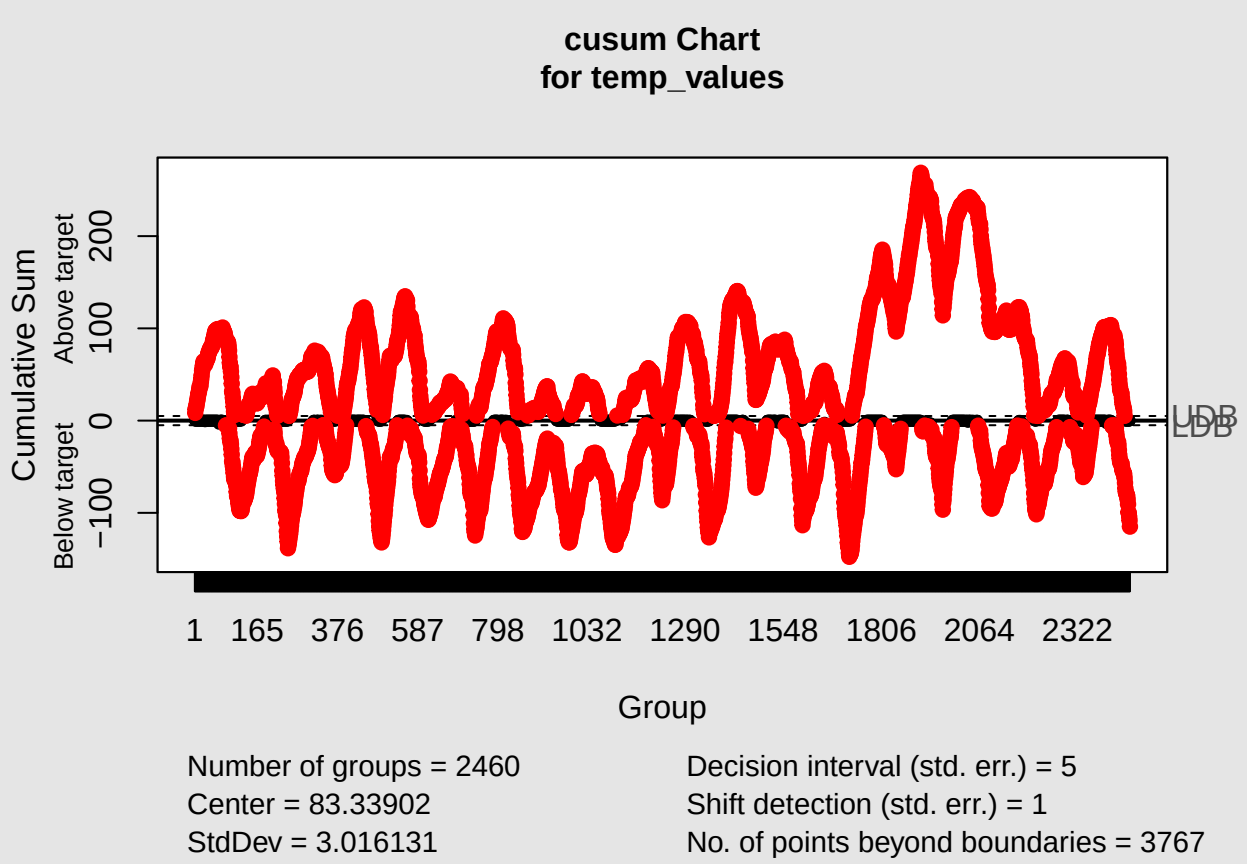
```
## Type 'citation("qcc")' for citing this R package in publications.
```

```r
# Load and manipulate data
temps <- read_delim('/Users/cn/Desktop/GTX/Homework 3/temps.txt', delim = '\t') %>%
  as_tibble() %>%
  gather(year, temp, -DAY) %>%
  mutate(year = as.factor(year),
         date = paste(DAY, year, sep = '-')) %>%
  mutate(date_val = dmy(date),
         color = ifelse(temp > mean(.$temp), 'Above', 'Below'),
         month = month(date_val),
         day = day(date_val)) %>%
  dplyr::select(date_val, DAY, year, temp, color, month, day)
```

```
## Rows: 123 Columns: 21
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr  (1): DAY
## dbl (20): 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, ...
##
## 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.
```

```r
# Extract temperature values for CUSUM analysis
temp_values <- temps$temp

# Perform CUSUM analysis
cusum_chart <- cusum(temp_values, decision.interval = 5, se.shift = 1)

# Plot the CUSUM chart
plot(cusum_chart, main = "CUSUM Chart for Temperature Data")
```



Yes, there is statistical evidence of warming, with the change becoming more apparent in the latter portion of the time series (around groups 1740-2156). The upward trend in the CUSUM plot during this period suggests a systematic increase in temperatures compared to the historical average.