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 fu

nction in the outliers package in R.

# Solution:

# Full code used and results underneath:

setwd("C:/Users/………………..”)

uscrime <- read.table("uscrime.txt", stringsAsFactors = FALSE, header = T)

head(uscrime)

summary(uscrime$Crime)

library(outliers)

#plot in a variety of ways for visual confirmation of predicted outliers

plot(uscrime$Crime)

plot(uscrime[,16],type ="b")

hist(uscrime[,16],type ="b")

qqnorm(uscrime$Crime)

boxplot(uscrime$Crime)

#grubbs test for outliers

grubbs.test(uscrime$Crime, type = 10)

#remove first outlier found and plot again before testing dataset again

crime1<- uscrime[-26,16]

plot(crime1)

boxplot(crime1)

#grubbs test #2 and remove second outlier found

grubbs.test(crime1, type = 10)

crime2<- crime1[-4]

plot(crime2)

#grubbs test #3

grubbs.test(crime2, type = 10)

hist(crime2)

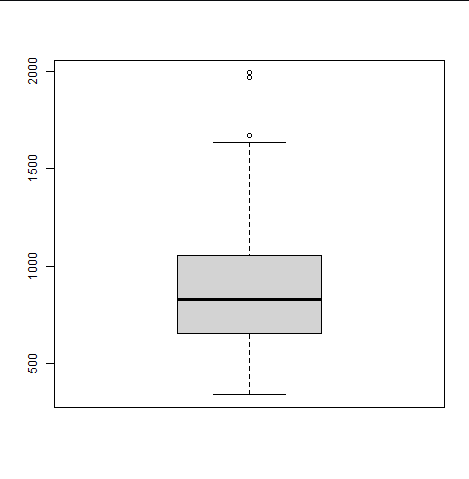
boxplot(crime2)

# Data set summary:

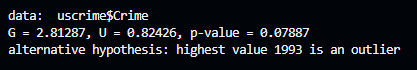
# 

# Initial data plots:

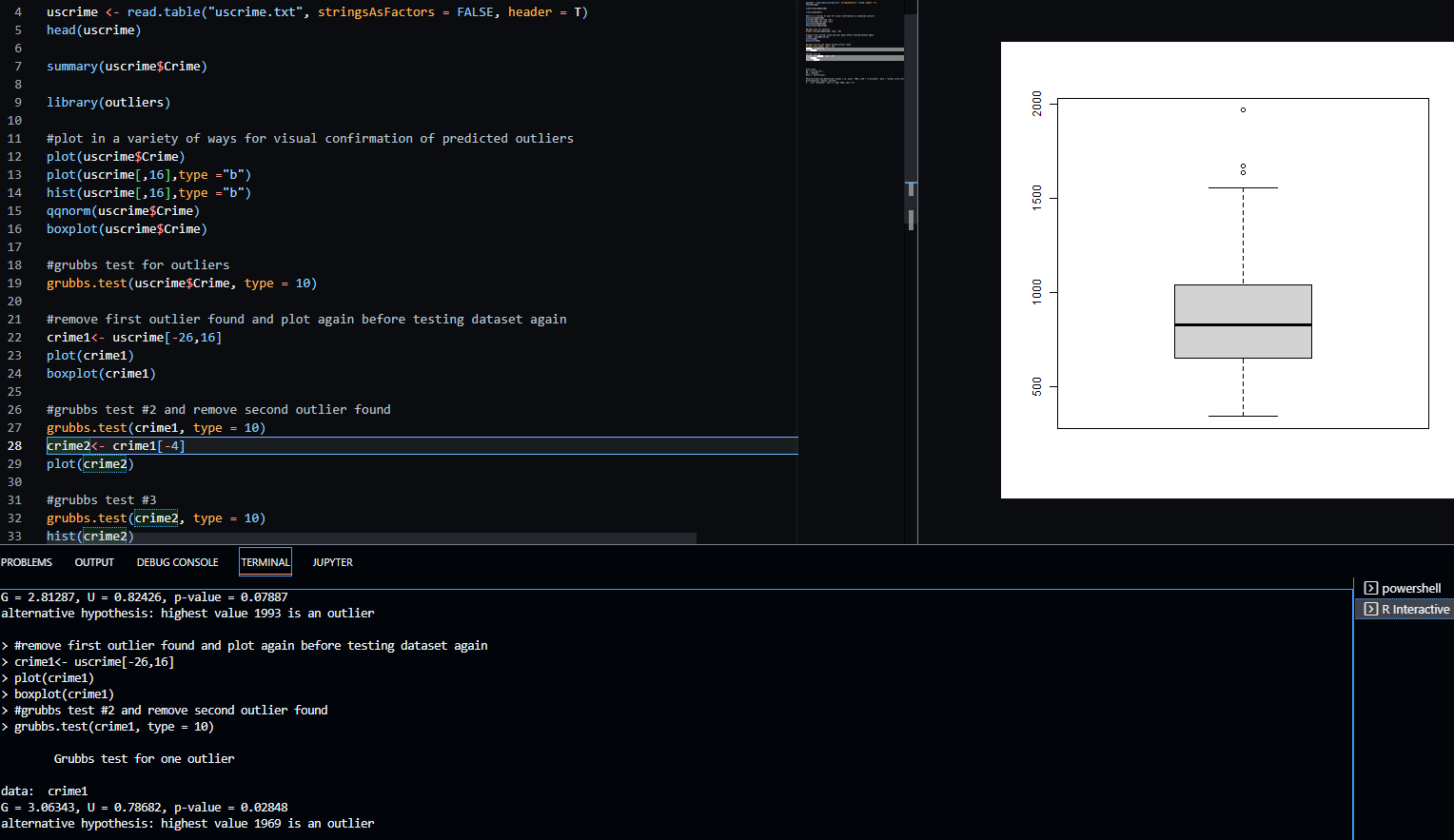
# 



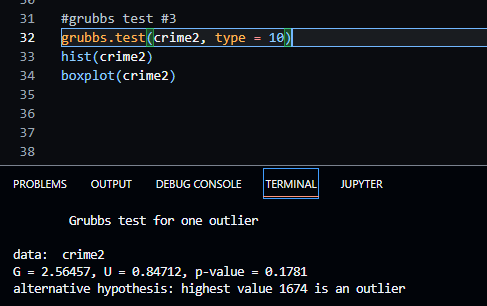
# First Grubbs test showing there’s over a 7% chance that we’d encounter an outlier so far from the others by chance alone, if all data were really sampled from a simple Gaussian normal distribution. Under 5% would normally be what we look for but based on the plots this will be considered an outlier and eliminated.



# Elimination of row 26 and grubbs test #2. Row 4 is now also an outlier with a p-value = 0.0284 after the removal of the first outliers and that is also removed.



# Grubbs test #3 for additional outliers were ran and though the code determines that value 1674 is an outlier, the p-value = 0.178 so this data point is not eliminated.



# Final data plots after all outliers are removed.

