**Question 5.1**

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

Here’s one possible solution. Please note that a good solution doesn’t have to try all of the possibilities in the code; they’re shown to help you learn, but they’re not necessary.

The file solution 5.1.R contains R code and some explanation for the following approach.

First, because the Grubbs test assumes normality, we start by running a normality test that you’ll probably remember from basic statistics: the Shapiro-Wilk test. The test actually suggests that the data is not normally distributed (p=0.001882) – but looking at the Q-Q plot below, it seems that the reason for the non-normality is the tails, which might imply that the test is affected by potential outliers. The middle of the distribution looks normal, so we’ll go ahead with the Grubbs test.

# 

Figure 1. Q-Q plot of the Crime column.

Note here that this is really a judgment call. On the one hand, it could be that the Shapiro-Wilk test is identifying that the tails, especially on the upper end, are really not normally-distributed, enough so that the extreme values aren’t really outliers, they’re just part of the distribution. On the other hand, it could be that the distribution really is close enough to normal, and the reason it fails the Shapiro-Wilk test is that there’s outlying data. The Grubbs test’s validity depends on which of these is closer to true.

In this case, let’s go on with the Grubbs test. At worst, it’ll either show that there aren’t outliers, or it’ll identify potential outliers – then we would (if this was more than a homework assignment) investigate those data points more carefully to see what’s going on, to determine whether they seem like a real part of the distribution or whether they’re real outliers.

It turns out that the lowest-crime city is unlikely to be an outlier (p-value so close to 1 that it just comes up as 1).

On the other hand, the highest-crime city might be an outlier (p=0.079), and if we remove it, the second-highest-crime city also appears to be an outlier (p=0.028). The box-and-whisker plot below shows the outliers more clearly.

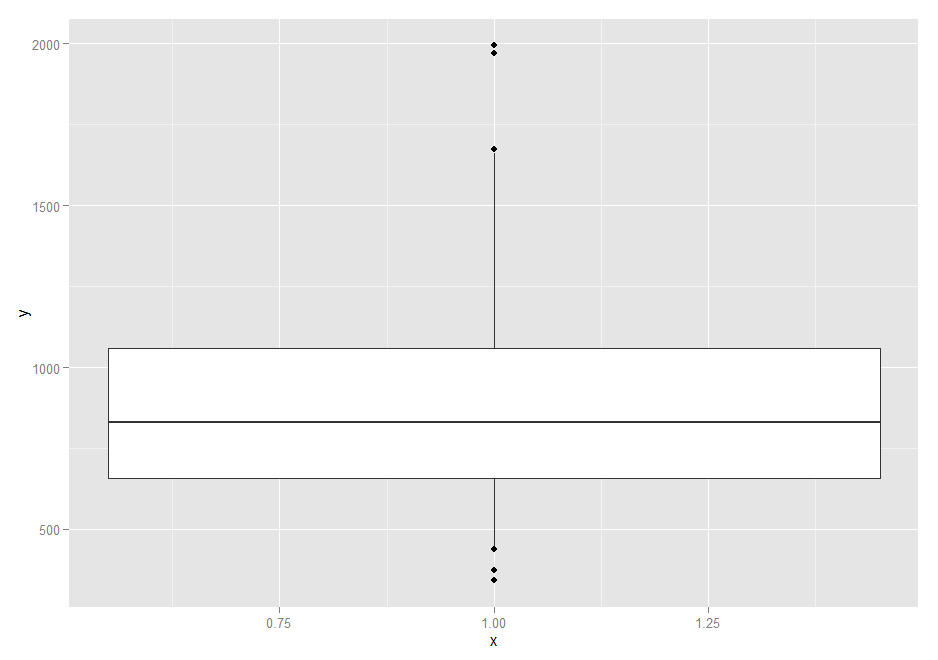


Figure 3. Box-and-whisker plot of the Crime column.

# Note that some people tried to determine whether there was an outlier on the high end and on the low end simultaneously, using the “type=11” parameter in the grubbs.test() R function. The problem with this approach is that the answer it returns is “no” – because they’re not *both* outliers (as we saw, the lowest-crime city isn’t an outlier). That result hides the fact that the highest-crime city probably *is* an outlier (and in fact, so is the second-highest). So using the “type=10” parameter is generally a better approach; it tests one side, and adding the “opposite=TRUE” parameter tests the other side. See solution 5.1.R for details.