**Question 5.1**

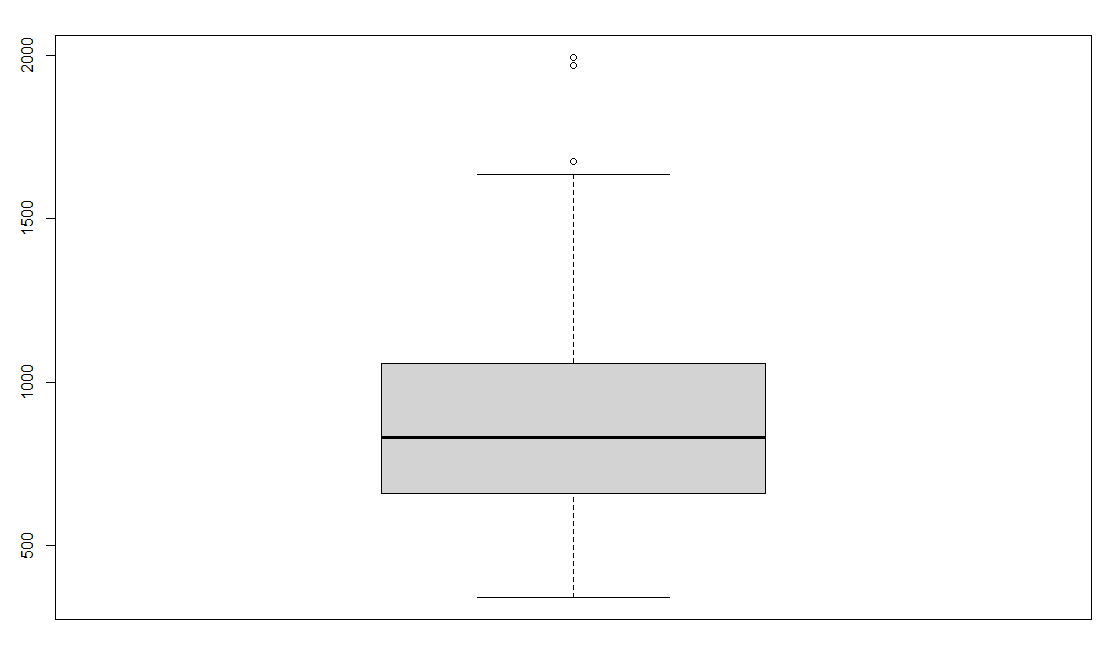
Using crime data from the file uscrime.txt (<http://www.statsci.org/data/general/uscrime.txt>, description at <http://www.statsci.org/data/general/uscrime.html>), test to see whether there are any outliers in the last column (number of crimes per 100,000 people). Use the grubbs.test function in the outliers package in R.

In conclusion, there are not outliers in the last column when we conduct test if both min and max value are outliers. And there are no outliers opposite tails. However, if we choose to test two side with both min and max value, we have two outliers which are: 1993, 1969.

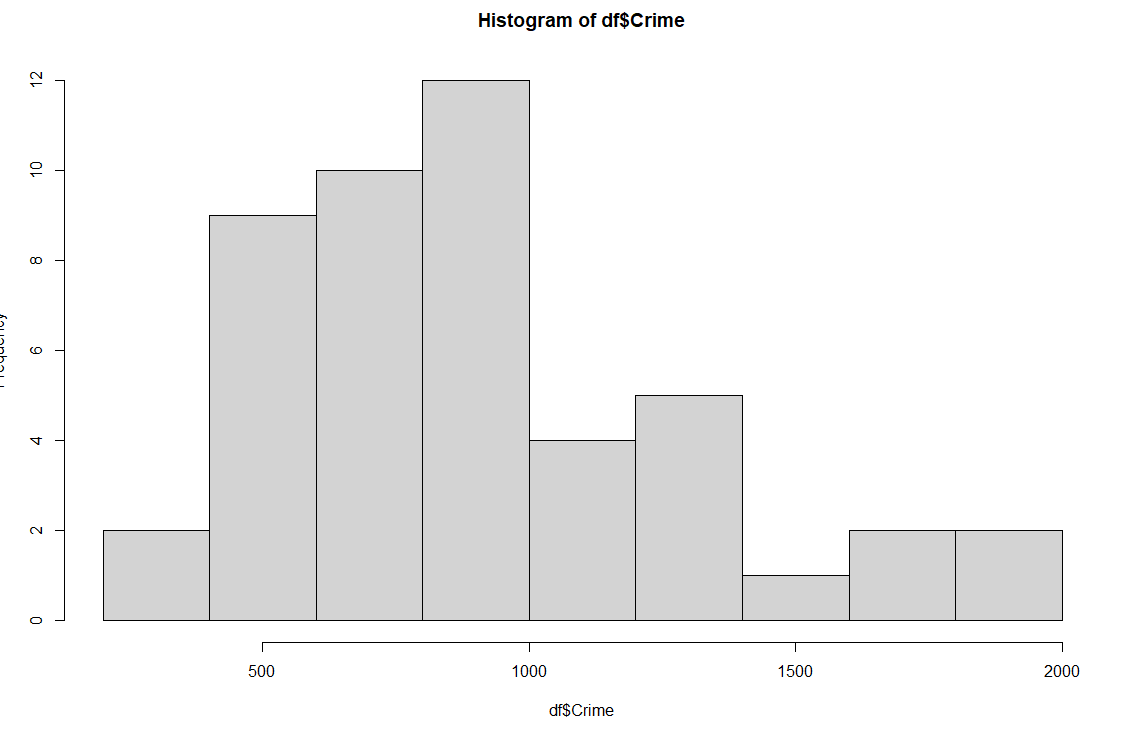
# Basic Explanation and codes are listed below:

First, we input the data, and use visualization to identify the outliers. However, this method is not always right.

# 



# From the boxplot, we can see few points are outranged. Majority of the data are close together, and a few outliers are clearly separated. It's reasonable to think that the point circled in top line close enough to the majority of the data that we can't be certain that it's an outlier. When this happens, we can use a statistical test called Grubbs' test to make sure.



# I plot a histogram in here for several reasons:

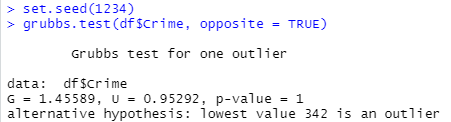
First, Grubbs' test works by assuming that the data are normally distributed and it is therefore important to first ensure that this assumption is plausible for the data.

# Second, A histogram provides a common way to check the normality assumption visually.

From the hist plot above, It is roughly a normal distribution. Hence we can apply grubbs test.

# 图形用户界面, 文本, 应用程序 描述已自动生成

We need to make our hypothesis first. H0: there is no outliers in the data. H1: there is outliers in the data. The results show g statistic is 2.81287, p is 0.07887. Because p is larger than significance value 0.05(in this case). We can not reject null hypothesis and conclude that highest value 1993 is an outlier.



# We set parameter opposite is true to look at other tail of the distribution. With same h1 and h0 hypothesis, the p =1 is larger than the significance value 0.05. Hence, we cannot reject null hypothesis and conclude that lowest value 342 is an outlier.

文本

描述已自动生成

# Next, we set type = 11 to check if both highest value and lowest value are outliers with same hypothesis test. The p = 1 is larger than the significance value, hence we cannot reject null hypothesis and conclude that 342 and 1993 are an outliers. Then we used which.max to find the index, and filtered it out, rerun the test.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

# With same hypothesis tests and significance value, but this time when we exclude the index with max values, the p = 0.02848 of new grubbs test is lower than 0.05 significance value, which we can reject null hypothesis and conclude that highest value 1969 is an outliers.

For the convenience of check all the data in last column of dataset, we write a loop here.

# 表格 中度可信度描述已自动生成

No outliers! Results are same for one side test, for opposite = TRUE test(other tail), for min-max test.

# 

When we test two side with min-max, we can found out that 1993 and 1969 are the outliers in the dataset.

# In conclusion, when we tried one tail test, the significance value is 0.05, we can not find any outliers. However, when we tried two side test, the significance value is 0.05/2, which accompanied by lower p value, we get to find two outliers.