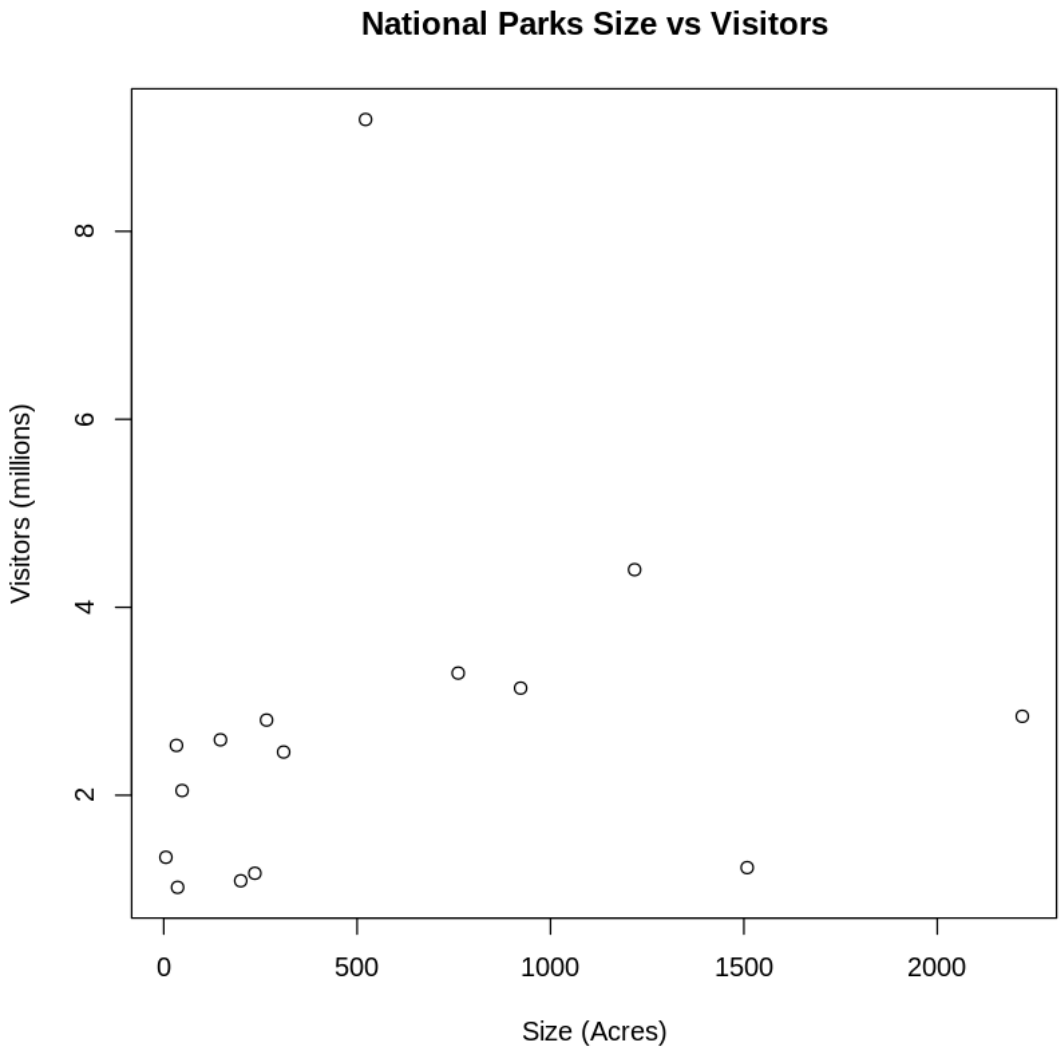
Question 3

January 28, 2021

[1]:
```
library(readxl)
park <- read_excel("NationalParks.xlsx")
```

(a) Bivariate scatter plot

[2]:
```
plot(park[,2:3], main ="National Parks Size vs Visitors")
```



(b) Compute the sample correlation matrix

[3]:
```
#data matrix
X <-matrix(c(park$`Size (Acres)`, park$`Visitors (millions)`),␣
↪nrow=15,ncol=2,byrow=F)
```

[4]:
```
#sample correlation matrix
sample_cor <- cor(X)
sample_cor
```

A matrix: 2 × 2 of type dbl

1.0000000	0.1725274
0.1725274	1.0000000

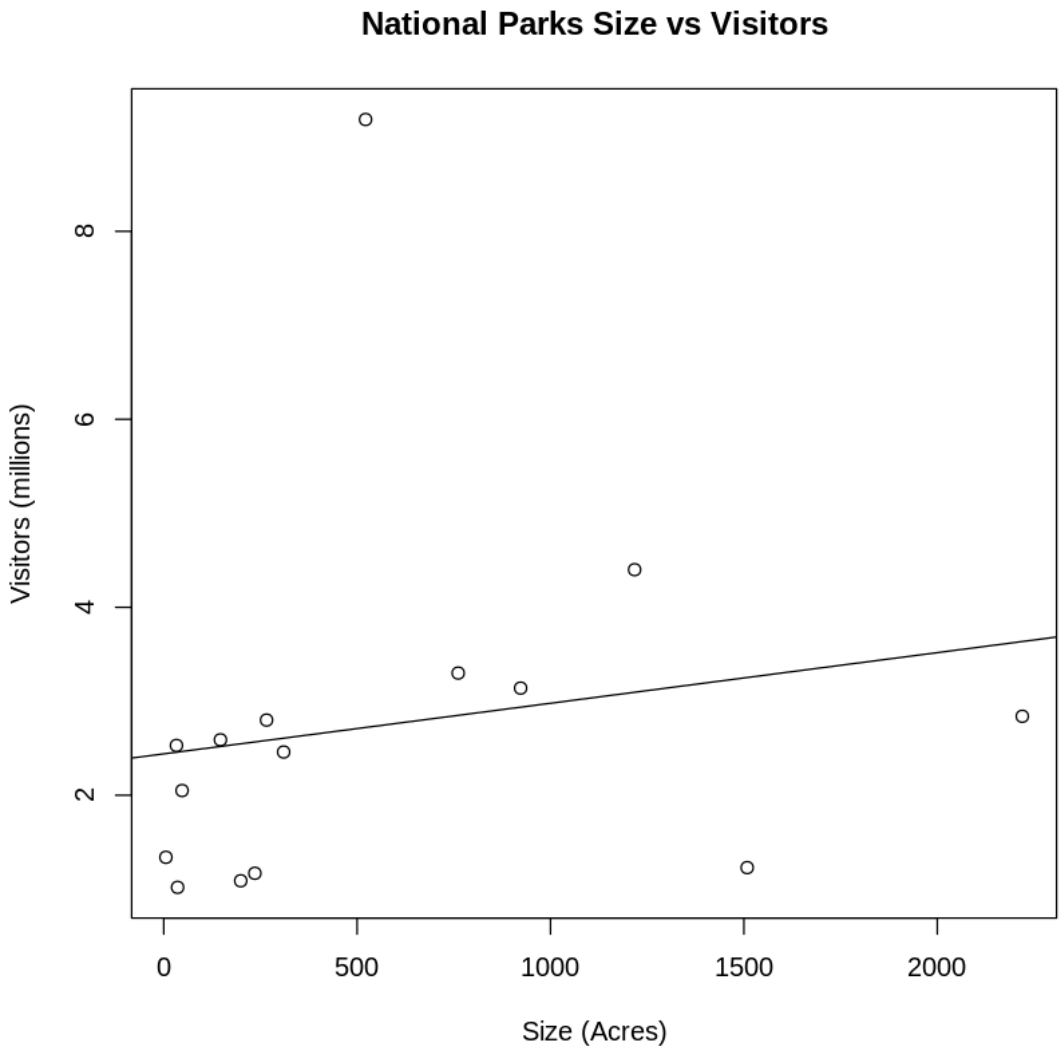
The sample correlation matrix is:

$$R = \begin{bmatrix} 1.0000000 & 0.1725274 \\ 0.1725274 & 1.0000000 \end{bmatrix} \tag{1}$$

(c) Identify the park that is unusual. Give a reason.

As we can see, from the plot below the park with visitors more than 9 millions is the outlier, which is Great Smoky National Park. It is far from the rest of the data points, and it is the farthest from the regression line.

[5]:
```
#plot the data to check outlier
plot(park[,2:3], main ="National Parks Size vs Visitors")
#regression line
abline(lm(park$`Visitors (millions)` ~ park$`Size (Acres)`))
```



(d)

[6]:
```
#create a new data matrix by removing Great Smoky National Park from the old␣
↪data matrix (row 7)
new_X <- X[-7,]
#new sample correlation matrix
new_sample_cor <- cor(new_X)
new_sample_cor
```

A matrix: 2 × 2 of type dbl

1.0000000	0.3907829
0.3907829	1.0000000

The new sample correlation matrix is:

$$R = \begin{bmatrix} 1.0000000 & 0.3907829 \\ 0.3907829 & 1.0000000 \end{bmatrix} \tag{2}$$

By removing the unusual data point, the sample correlation between park size and visitor numbers become more than doubled in magnitude, from 0.17253 to 0.39078. Showing a greater positive linear relationship between the two variables.