Correlation: Variance, Covariance, and Plotting

Kim Wiggins

Park names, park sizes and visitors printed below:

```
parks
```

```
##                   size visitors
## Arcadia           47.4     2.05
## Bryce Canyon      35.8     1.02
## Cuyahoga Valley   32.9     2.53
## Everglades      1508.5     1.23
## Grand Canyon    1217.4     4.40
## Grand Tenton     310.0     2.46
## Great Smoky      521.8     9.19
## Hot Springs        5.6     1.34
## Olympic          922.7     3.14
## Mount Rainier    235.6     1.17
## Rocky Mountain   265.8     2.80
## Shenandoah       199.0     1.09
## Yellostone      2219.8     2.84
## Yosemite         761.3     3.30
## Zion             146.6     2.59
```

Use the `cov()` and `cor()` commands to view the sample covariance matrix and the sample correlation matrices.

```
cov(parks)
```

```
##                 size   visitors
## size     424177.2998 228.416524
## visitors    228.4165   4.132295
```

Covariance matrix, Parks data

```
cor(parks)
```

```
##               size  visitors
## size     1.0000000 0.1725274
## visitors 0.1725274 1.0000000
```

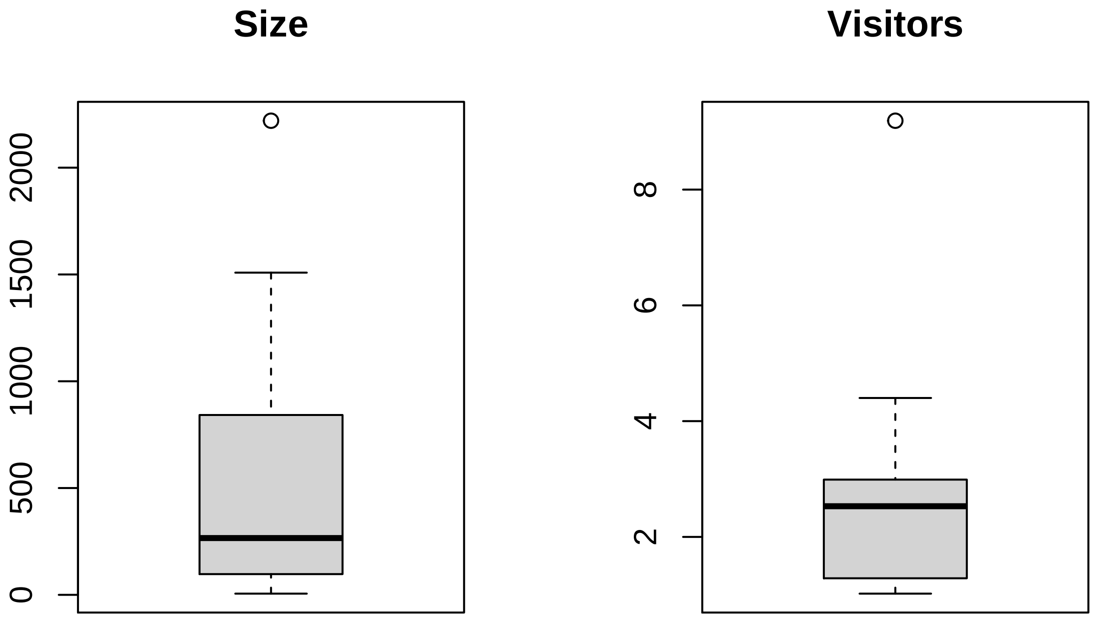
Correlation Matrix, Parks Data

The correlation matrix displays the Pearson correlation coefficient between two variables to demonstrate how strongly those variables are related to each other. As the scaled form of covariance, correlation coefficients are standardized (non-scalar) and dimensionless (unit-free), and can thus be interpreted easily on a scale from -1 to +1. Covariance values vary from negative infinity to positive infinity. While covariance indicates the direction of the relationship only, correlation indicates direction and strength, resulting in a proportion metric measuring how much on average these variables differ with regard to one another.

Because the covariance matrix isn't scaled, we can only interpret direction. The size of the park is measured on a different scale than the visitor count, so interpretation regarding strength of the direction is not appropriate using the covariance matrix. In this case, the covariance matrix tells us that the relationship is positive, but the correlation coefficient of 0.173 being so close to 0 indicates that the linear relationship is not strong. The variables may have some other strong relationship that is not linear, and we must take care to not interpret a 0 (or values close to 0) as indicating lack of any relationship - only lack of a strong linear one.

We can also draw a matrix of plots containing histograms and boxplots of park sizes and number of visitors, making sure to give appropriate titles and axes labels for the histograms and clearly interpret the graph/s.

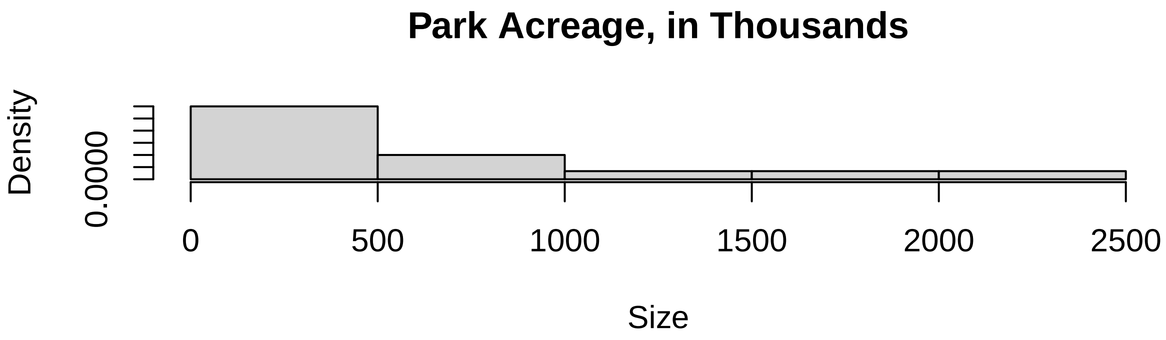
```r
layout(matrix(1:2, nc=2))
# 1 by 2 rows
boxplot(parks$size, main = "Size")
boxplot(parks$visitors, main = "Visitors")
```

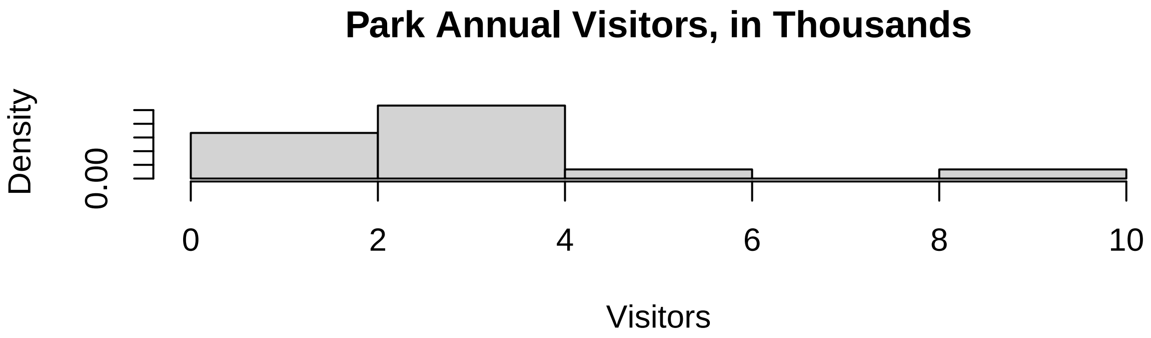


Examination of the boxplots indicate two outliers: one for each variable. Visual inspection of the elements reveal the potential outliers to be Yellowstone for Size, and Great Smoky for visitors.
Be cautious to consider these *potential* outliers - they could be anomalies or belong to a bimodal distribution and actually be quite typical of the data's behavior.

Histograms

```r
par(mfrow=c(2,1))
hist(parks$size, main = "Park Acreage, in Thousands", xlab = "Size", prob= T)
hist(parks$visitors, main = "Park Annual Visitors, in Thousands", xlab = "Visitors", prob = T)
```

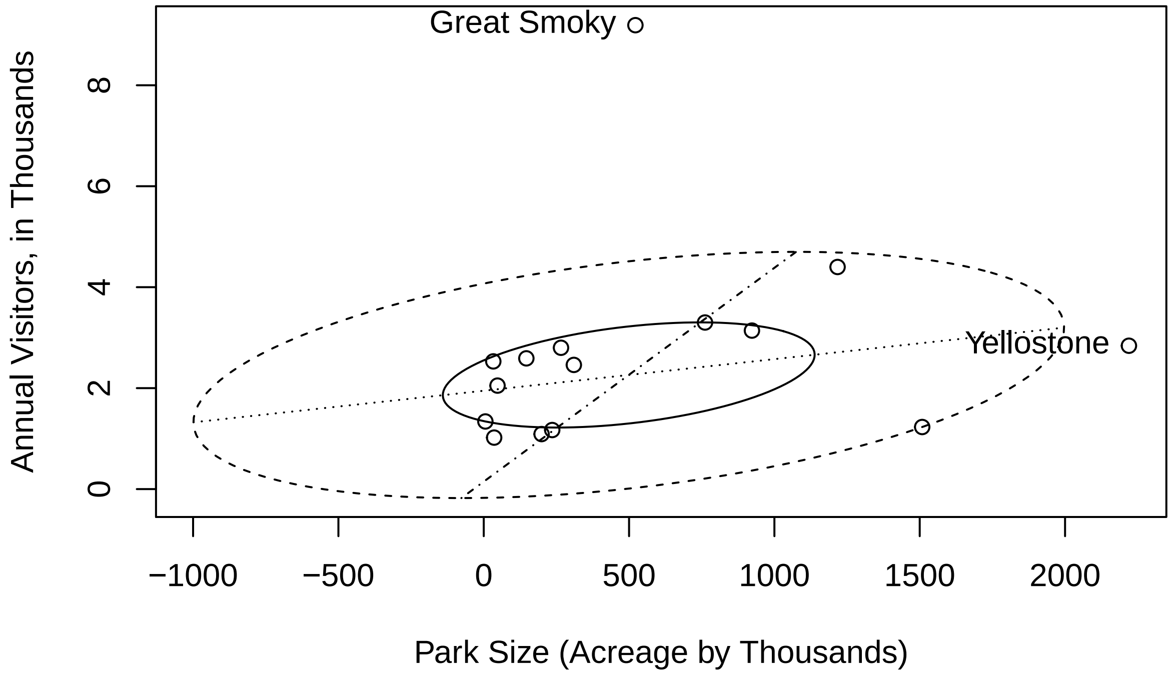




Histograms indicate that most values for Size fall within the 0-1000 range. Visitor ranges indicate the majority of observations fall between 1 and 4, with a potential outlier in the 8-10 range.

Explore for possible outliers. Use appropriate graphs to draw conclusions. Calculate correlation coefficient between park size and number of park visitors before and after removing possible outliers. Comment on your findings.

```
outpark = match(parkname <- c("Great Smoky", "Yellostone"), rownames(parks))
bvbox(parks, mtitle = "Bivariate Boxplot, Parks", xlab = size.lab, ylab = visitors.lab)
text(parks$size[outpark], parks$visitors[outpark], labels = parkname, pos = c(2,2,4,2,2))
```



The outliers are labeled, and in this case, will be removed from future analysis. In future cases, that is likely not the best practice, but for the purposes of comparing before and after correlations here, we will remove them.

After exluding the outliers, the correlation coefficient increases from 0.173 to 0.399.

```
with(parks, cor(size, visitors))
```

```
## [1] 0.1725274
```

```
with(parks, cor(size[-outpark], visitors[-outpark]))
```

```
## [1] 0.398539
```