In this session, we will

- look at how to subset data;
- learn how to implement simple linear algebra routines;
- explore our data using simple plotting and summary statistic techniques.

Subsetting data

Download the data file “Music.csv” from https://www.scss.tcd.ie/~arwhite/Teaching/STU33011.html and read it into R, giving it the name music. (See previous worksheet for details).

music has been read in as a data.frame object. (This is distinct from a matrix object. We will discuss this more in later labs.) To select elements of music we enter a command of the form music[i,j], where i denotes the row number(s) and j denotes the column number(s).

For example, music[1, 1] denotes the element in the first row and first column of music.

```
music[1, 1]  
```

## [1] "Dancing Queen"

music[c(1, 2, 3), 2] denotes the elements in the second column of the first three rows of music:

```
music[c(1, 2, 3), 2]  
```

## [1] "Abba" "Abba" "Abba"

music[1,] denotes the first row of music:

```
music[1,]  
```

## X Artist Type LVar LAve LMax LFEner LFreq
## 1 Dancing Queen Abba Rock 17600756 -90.00687 29921 105.921 59.57379

music[, 2] denotes the second column of music:

```
music[, 2]  
```

## [1] "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Abba"  "Vivaldi"  "Vivaldi"  "Vivaldi"  "Vivaldi"  "Vivaldi"  "Vivaldi"  "Mozart"  "Mozart"  "Mozart"  "Mozart"  "Mozart"  "Mozart"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  "Eels"  

## 1

Exercise

- Obtain the entries in the music data frame corresponding to rows 1, 3 and 4 and columns 2 through 5 and assign this subset of music as M2.

- Obtain the entries in the music data frame corresponding to columns 4 through 8 (i.e., the numeric variables) and assign this subset of music as music_num.

Plots and summary statistics

We can visualise data using the plot function. Say we wish to plot a quadratic curve. What does each line of the below code do?

```r
x <- sample(-1000:1000, size = 200, replace = FALSE)
x <- sort(x)
y <- x^2 + 3*x + 1
plot(x, y, type = "l", col = "red")
```

To make simple statistical summaries of numeric variables, use mean, sd, var, cov, and cor. For example to take the mean of variable 4 of the music dataset, just use music[, 4]. Unfortunately, in this case, it’s hard to get a good sense of what this means when the variable terms are not clearly defined. To get covariance and correlation estimates use:
cov(music[, 4:8])

<table>
<thead>
<tr>
<th></th>
<th>LVar</th>
<th>LAve</th>
<th>LMax</th>
<th>LFEnner</th>
<th>LFreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVar</td>
<td>6.983700e+14</td>
<td>1.393563e+08</td>
<td>1.487901e+11</td>
<td>1.008562e+08</td>
<td>-1.218687e+09</td>
</tr>
<tr>
<td>LAve</td>
<td>1.393563e+08</td>
<td>2.299347e+03</td>
<td>-1.484280e+04</td>
<td>4.635271e+00</td>
<td>1.538224e+03</td>
</tr>
<tr>
<td>LMax</td>
<td>1.487901e+11</td>
<td>-1.484280e+04</td>
<td>7.676086e+07</td>
<td>1.987469e+04</td>
<td>-4.122962e+05</td>
</tr>
<tr>
<td>LFEnner</td>
<td>1.008562e+08</td>
<td>4.635271e+00</td>
<td>1.987469e+04</td>
<td>3.005762e+01</td>
<td>-2.599052e+02</td>
</tr>
<tr>
<td>LFreq</td>
<td>-1.218687e+09</td>
<td>1.538224e+03</td>
<td>-4.122962e+05</td>
<td>-2.599052e+02</td>
<td>3.122046e+04</td>
</tr>
</tbody>
</table>

cor(music[, 4:8])

<table>
<thead>
<tr>
<th></th>
<th>LVar</th>
<th>LAve</th>
<th>LMax</th>
<th>LFEnner</th>
<th>LFreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>LVar</td>
<td>1.00000000</td>
<td>0.11168505</td>
<td>0.6426304</td>
<td>0.69611717</td>
<td>-0.2609936</td>
</tr>
<tr>
<td>LAve</td>
<td>0.1116850</td>
<td>1.00000000</td>
<td>-0.0358804</td>
<td>0.01790643</td>
<td>0.1843789</td>
</tr>
<tr>
<td>LMax</td>
<td>0.6426304</td>
<td>-0.0358804</td>
<td>1.0000000</td>
<td>0.41376440</td>
<td>-0.2663298</td>
</tr>
<tr>
<td>LFEnner</td>
<td>0.6961172</td>
<td>0.01790643</td>
<td>0.4137644</td>
<td>1.00000000</td>
<td>-0.2682983</td>
</tr>
<tr>
<td>LFreq</td>
<td>-0.2609936</td>
<td>0.18437887</td>
<td>-0.2663298</td>
<td>-0.26829833</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>

Exercise

- Visualise variables 7 and 8 in the music dataset. How does the behaviour of the data correspond with the variables's correlation?

Matrix Eigendecomposition

Eigendecomposition of matrices in R is simple when using the inbuilt eigen function. Make sure to check the help file for this command.

Exercise

- Create a matrix \( A \) such that:

\[
A = \begin{pmatrix}
1 & 2 \\
2 & 5
\end{pmatrix}
\]

Now enter the following command:

\[
\text{res} \leftarrow \text{eigen}(A)
\]

The above assigns the result of the eigendecomposition of \( A \) to the name \( \text{res} \). Typing \( \text{res} \) into the console then calls the result up. It would have been fine to simply enter \( \text{eigen}(A) \) to see the result of the function, but assigning it with a name means we can use it again in the future without having to repeat the earlier command.

\[
\text{res}
\]

```
## eigen() decomposition
## $values
## [1] 5.8284271 0.1715729
##
## $vectors
##
## [,1]        [,2]
## [1,] 0.3826834 -0.9238795
## [2,] 0.9238795  0.3826834
```

The numbers following the term $values are the eigenvalues of \( A \), given in decreasing order. The columns of the matrix returned after the term $vectors are the corresponding orthonormal eigenvectors. That is to say the output informs us that the matrix \( A \) has one eigenvalue of 5.8284271 that corresponds to
orthonormal eigenvector (0.3826834, 0.9238795), and another smaller eigenvalue of 0.1715729 that corresponds
to orthonormal eigenvector (-0.9238795, 0.3826834).

To request the eigenvalues and eigenvectors of the matrix $A$ directly, now enter the following:

res$values

## [1] 5.8284271 0.1715729

res$vectors

## [,1]        [,2]
## [1,] 0.3826834 -0.9238795
## [2,] 0.9238795  0.3826834

Exercise

- Write an R code that directly returns only the first eigenvalue of $A$.
- Write an R code that directly returns only the second eigenvector of $A$.

Matrix Multiplication

We can check to see if eigen returns correct eigenvalues and eigenvectors. Recall that $Av = \lambda v$ for any
eigenvector $v$ and its corresponding eigenvalue $\lambda$. In R, the command %*% is used to multiply two matrices
together. (What happens if * is used instead?) Use %*% to view the help file.

To check the result for the leading eigenvalue and eigenvector, enter the following:

$A \%*\% res$vectors[,1]

## [,1]
## [1,] 2.230442
## [2,] 5.384765

Does this result agree with the following?

res$value[1] * res$vectors[,1]

Exercise

- Check that the second eigenvalue and eigenvector are also correct.
- The cranial length and cranial breadth of 35 female frogs have expected value $\mu = (23, 24)^T$ and
covariance $\Sigma = \begin{pmatrix} 17.7 & 20.3 \\ 20.3 & 24.4 \end{pmatrix}$. What are the eigenvalues and eigenvectors of $\Sigma$?
- Can you think of any other ways to double check whether or not the eigenvalues are appropriate? Hint: in R we transpose a matrix or vector with the t function, and make a diagonal matrix using the diag function.