G54DMA - Lab 5: Visualisation in R

Question Sheet

This question sheet presents a series of exercises designed to make you use R for data visualisation purposes. You will be applying the concepts you learned during the previous labs (data frames, reading data from files, etc.) to analyse and visualise data.

In this lab session, you will learn to:

1. **Create simple graphs to analyse your data using basic R.**
2. **Create sophisticated graphs to analyse data using the *ggplot2* library.**
3. **Choose appropriate visualisation methods for data analysis purposes.**

To begin, start your R editor (RStudio or your editor of choice). I recommend that you revise Lecture 4 and Lecture 5 and read the Instruction sheet carefully. Once you have finished reading them, start working on these exercises.

A. Basic Visualisation using R

In this section, we use the *Iris* dataset to create and analyse simple graphs.

1. Create a pie chart for the Species class in the *Iris* dataset. *Setosa* should be blue, *versicolor* should be yellow and *virginica* should be magenta. Save this chart as species_piechart.png.
2. Create a function called "plotPetalLength.R" that reads the *Iris* dataset and plots a histogram of petal lengths with:
   a. 30 bars in a colour of your choosing
   b. A sensible title
   c. Two lines of different colour showing the mean and the median
   d. A legend for the median and mean lines

3. Create a dotplot that groups samples into their species and shows how the petal lengths of all instances vary. Each group should have a different colour of your choosing.
    a. What can be inferred from this plot regarding the petal length of *setosa* plants in comparison to *versicolor* and *virginica* plants?
4. Create a scatterplot that plots sepal width versus petal length. Group points according to class.
    a. What can be inferred from the scatterplot?
5. Create a scatterplot matrix showing the relationship between all numerical variables in the *iris* dataset.
    a. Looking at the graphs, which attributes have a higher correlation?
6. Plot a 2-by-2 graph called "boxplots.png" in which each quadrant shows:
    a. A boxplot of all data and how they vary according to the length and width of the sepal and the petal.
    b. A boxplot of *setosa* plants and how they vary according to the length and width of the sepal and petal.
    c. A boxplot of *versicolor* plants and how they vary according to the length and width of the sepal and petal.
    d. A boxplot of *virginica* plants and how they vary according to the length and width of the sepal and petal.

    All plots should have sensible titles and both axes should be titled.

    Looking at the graph, answer the following questions:

    I. Which species has the largest sepal length?
    II. What does the "petal length" data of the whole data boxplot tell us about the distribution of iris?

B. Advanced Visualisation

In this section, we will be using the "*ggplot2*" package to create more complicated and more informative graphs in R. In terms of data, we will be using a dataset of footballers from different regions in Narnia. This dataset is called *team* and you can download it from Moodle.

Before attempting the following exercises, download *team* and familiarise yourself with the dataset and its attributes. Then, install and load "ggplot2" to your R session.

Remember that all graphs need to be titled and all axes labelled.

1. Create a histogram of that shows the age distribution from all players at intervals of 1 year.
    a. What is the most common age? How common is it?
2. Create a histogram that shows the salary distribution of players that are Defenders. Choose a sensible interval.
    a. How many 30-year-old players are there?
    b. What is the age of the youngest player in the team? How many of those players are there?
3. Repeat the previous histogram, but grouping the data according to Gender.
    a. How many women are aged 27?
    b. How many men are aged 33?
4. Create a histogram that shows the age distribution of Forward players grouped into teams. Choose a sensible age interval.
5. Create a scatterplot that shows the relationship between speed and height of players from Bim, Dragon Island and Calormen, grouped by Gender.
    a. Add a line that fits each group of the data. Consider the tendencies shown in this:
        i. How good are the fit lines?
        ii. Do faster male footballers tend to be taller or shorter?
6. Create a graph with separate scatterplots that show Weight vs Height of players grouped into their teams. Separate each scatterplot according to Gender. Title the graph and the axes sensibly. Figure 1 shows an example.

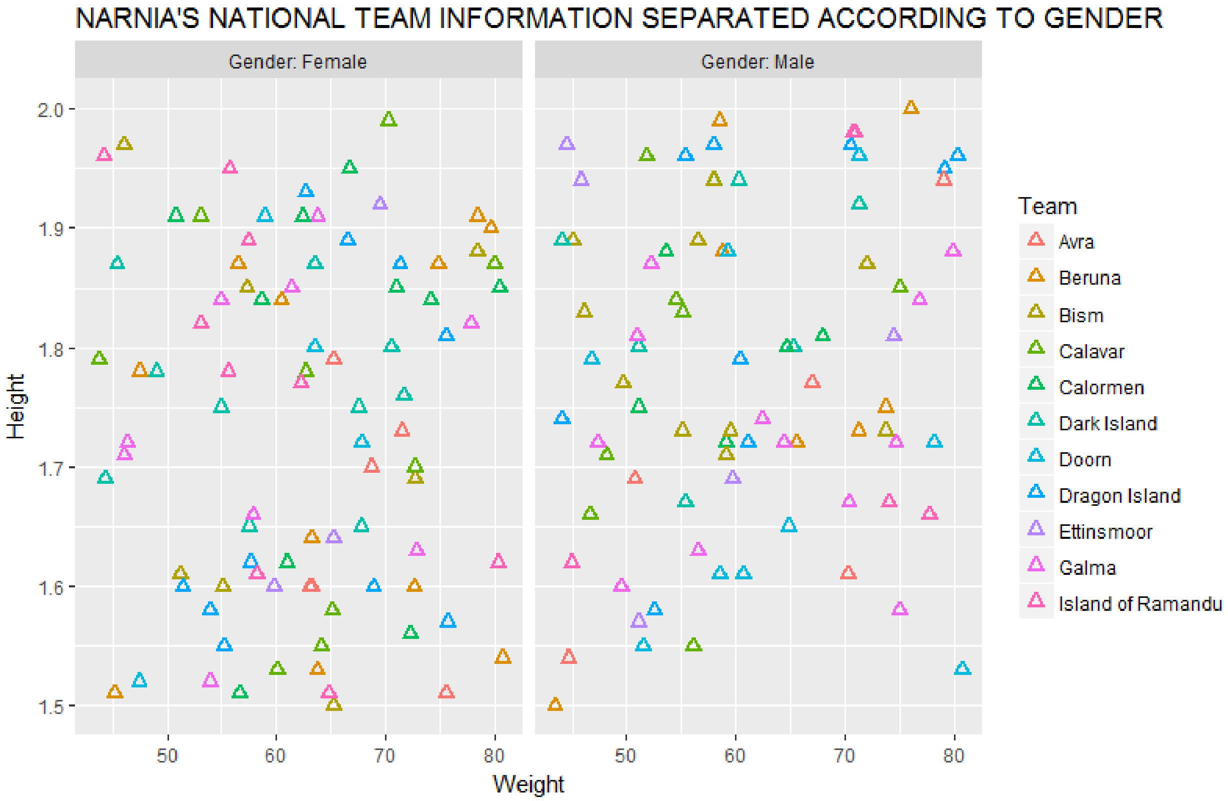

Figure 1. Narnia's National Team Information Separated According to gender.

7. Apply the functions and exercises done previously to replicate the graph shown in Figure 2. This graph shows:
    a. Members of all teams grouped by team using colour.
    b. Members of all teams grouped by gender using shape. Female players are shown with a diamond mark and male players are shown with a triangle mark.
    c. Those players who ply the Forward position have their speed shown as a label. (Hint: the ggrepel package will come in handy here).
    d. It is titled "NARNIA'S NATIONAL TEAM INFORMATION".
    e. The X axis is called "Weight of all players" and ranges between 40 and 85, with breaks at increments of 5. (Hint: the grid package will be helpful here).
    f. The Y axis is called "Height of all players" and ranges between 1.49 and 2.01, with breaks at intervals of 0.1.

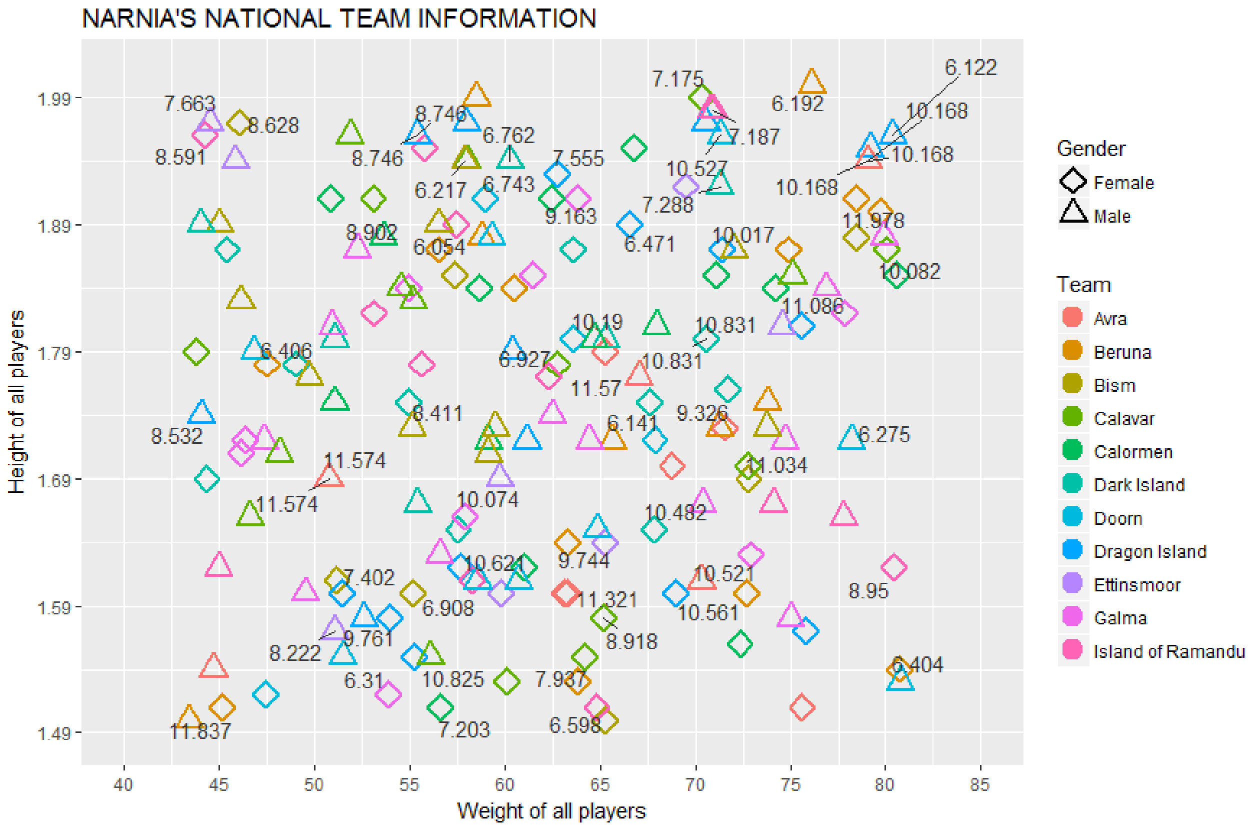

Figure 2. Narnia's National Team Information. All players from all teams, whether selected or not, are shown.

C. Choosing the right visualisation techniques

In this section, you will have to use the skills developed in the previous labs and use statistics and visualisation to provide a meaningful analysis of the *Reviews* dataset. You can download the *reviews.csv* dataset from Moodle.

The *Reviews* dataset contains over 84,500 instances of people rating different movies. The attributes included in the dataset are:

- userID: the ID of the reviewer. One reviewer can have a variable number of reviews.
- movieID: the movie reviewed by the reviewer.
- rating: how many stars the reviewer has given to the movie. The lowest score is 0.0 and the highest is 5.0.
- date: date and time in which the review was uploaded.

- movieTitle: the title of the movie reviewed.
- movieYear: the year the movie was released.
- movieGenre: the genre of the move. Although one movie can be of many different genres, for the purpose of this exercise, we will only consider the main genre of a movie.

You have been hired by *Calabaza Productions*, a production company interested in studying trends in reviews from movies released in the last hundred years. Your job is to provide a comprehensive set of statistics and visualisations that can help *Calabaza Productions* understand the dataset.

1. Analyse the data to familiarise yourself with the contents of the dataset. Calculate appropriate centrality and dispersion measurements for all attributes.

2. Use analytics to answer the following questions:
    a. How many reviewers have participated?
    b. How many movies have been reviewed?
    c. Which reviewer has the most reviews? And the least?
    d. Which genre has the most reviews? And the least?
    e. Which is the oldest movie reviewed? Who reviewed it (if more than one reviewer rated it, return all of them)?
    f. Which one(s) is (are) the highest-rated genres of the 1990s?
    g. What is the most reviewed movie? What is its average score?

3. Use visualisation to answer the following questions:
    a. What is the distribution of ratings for all movie genres?
    b. What is the distribution of Comedies according to release decade?
    c.
    d. How have the average rating of Comedies, Westerns, Dramas and Horror have changed since
    e. Do a monthly comparison of the number of Horror movies and Animations of reviews uploaded