# Lab 7: Data Pre-processing and Mining

Question Sheet

Dr Mercedes Torres Torres

## Introduction

This question sheet presents a series of exercises designed to make you use R to pre-process and mine a raw dataset. We will be using the *Team* dataset for all exercises in this lab.

In this lab session, you will learn to use R to:

- Check for missing or repeated values
- Deal with missing or repeated values
- Calculate and deal with outliers
- Create transformed attributes
- Reduce the dimensions of a dataset using Principal Component Analysis

## 1 Data Pre-processing

1. Some players are repeated. Delete rows with players who have the same name and surname.
   a. How many unique players are there?
   b. How many female players are there?
2. Analyse the number of missing values for each attribute and instance. Is there any attribute that should be deleted? Are there any instances that should be deleted? Why?
3. Some ages are missing. Find those empty values and replace them with a suitable value.
   a. Which value have you used? Why?
4. Some heights, weights and speeds are missing as well. Find those and replace them with a suitable value in each case. Take into consideration the players' gender.
   a. Which value have you used? Why?
5. Set all missing values in the "Selected" column to *U* for *unknown*.
6. Some players have no playing position. Choose a sensible strategy and deal with them.
   a. What strategy have you chosen to use? Why?

7.Using the equation given in class, identify any outliers in the following attributes: height and speed. If any are found, deal with them in a suitable manner.

a. If there are any outliers, which strategy have you chosen to deal with them? Why?

b. Give an example of another strategy that could have been applied and explain the effect it would ha

8. Include one more attribute in the database which normalizes the speed attribute between 0 and 1.

9. Include one more attribute in the database, Body Mass Index (BMI), to give you more information about the player's physical condition.

10. Analyse your data by creating a summary table overall and a summary table for each home team. Include centrality and dispersion measures. Use the pre-processed data to answer the following questions:

    a. What is the team with most payers?

    b. What is the overall mean salary?

    c. What is the overall median speed?

## 2 Data Mining

1. Now that your data is clean and transformed, answer the following:

    a. How many different teams with more men than women does Narnia have?

    b. What is the mean age and salary of male players in Dragon Island?

    c. What is the median height of female forward players in Bim?

    d. What is the home team of the fastest player?

    e. Show a histogram of the frequency of the positions from the fastest 40 players.

    f. What is the gender of the goalkeeper with fewer goals against them?

    g. Generate a pie chart with the percentages of selected players from each region. Which region has more pre-selected players?

    h. What is the team which spends most on salaries? And the one which spends less?

    i. Which is the team with most forwards? And the team with less defenders?

    j. Which team has the biggest difference in salaries (i.e. the difference between the most paid player and the least paid player is the largest).

    k. How many players were initially not selected?

    l. Which team has the best forwards (i.e. their average scoring is the highest)?

2. Change the "Selected" attribute to $N$ (for *No*) for those players who are under 16 and over 40.

3. Change the "Selected" attribute to $N$ for those players whose BMI is under 18 and over 24.

4. Change the "Selected" attribute to $N$ for those players whose salary is over 1.5 times the overall median.

5. Change the "Selected" attribute to $N$ for all forwards whose speed is less than the mean of the goalkeepers speed.

6. Change the "Selected" attribute to $N$ (for *No*) for all players whose years of experience are over 8.

7. Change the "Selected" attribute to $N$ for all forwards who have scored less than 20 goals and all midfielders who have scored with less than 15 goals.

8. Change the "Selected" attribute to $N$ (for *No*) for all goalkeepers who have more than 15 goals scored against them.

9. Change the “Selected” attribute to $Y$ (for *yes*) for those forwards whose speed is over the 3rd quartile speed for all forwards.
10. Change the “Selected” attribute to $Y$ for the top 3 goalkeepers (those who have the least amount of goals scored against them).
11. Change the “Selected” attribute to $Y$ for those midfielders and defenders who are between 18 and 29 years.
12. Change the “Selected” attribute to $Y$ for those defenders who have scored 6 or more goals.
13. Change the “Selected” attribute to $Y$ for those midfielders who have scored 9 or more goals.
14. How many players are left as Selected?
15. Select all appropriate measurements from the players, including (but not limited to) height, weight and BMI. Apply PCA to them and study:

    a. How many attributes would you need to obtain a variance of over 60%?
    b. How many attributes would you need for a variance of over 90%?
    c. Which linear combination of attribute would give you PC1 and PC2?
    d. Are there any attributes that are not helpful?