It is known that R provides a wide range of functions for data preprocessing, calculation, manipulation, and graphical displays. In particular, it includes a well-developed, simple programming language that users can extend by adding new functions on their own. Many such extensions of the R language in the form of packages are easily downloadable from the Comprehensive R Archive Network ([CRAN](https://cran.r-project.org/)).

The purpose of this exercise is to help you be familiar with the R language, which will be used extensively in this course. As an illustration, we consider student dataset (studentdata) available in the [LearnBayes](https://cran.r-project.org/web/packages/LearnBayes/index.html) R package. This data was collected based on answers to a sheet of questions given to a large number of students in introductory statistics classes. This data contains 657 observations on the following 11 variables below:

|  |
| --- |
| **Student** student number  **Height** height in inches  **Gender** gender  **Shoes** number of pairs of shoes owned  **Number** number chosen between 1 and 10  **Dvds** name of movie dvds owned  **ToSleep** time the person went to sleep the previous night (hours past midnight)  **WakeUp** time the person woke up the next morning  **Haircut** cost of last haircut including tip  **Job** number of hours working on a job per week  **Drink** usual drink at suppertime among milk, water, and pop |

Install LearnBayes package in R/Rstudio and then access studentdata

|  |
| --- |
| #Install the LearnBayes package  #Keep in mind that R is case-sensitive  install.packages('LearnBayes')  #You just need to install once and then you can directly use  #so long as you access the LearnBayes package  library(LearnBayes)  #Access studentdata from the LearnBayes package  data(studentdata)  attach(studentdata)  #show part of data  head(studentdata) |

After accessing the studentdata, we can now use R to answer the following questions:

1. The variable **Dvds** in the student dataset contains the number of movie DVDs owned by students in the class.
2. Construct a histogram of this variable using the **hist** command in R.
3. Summarize this variable using the **summary** command in R.
4. Use the **table** command in R to construct a frequency table of the individual values of Dvds that were observed. If one constructs a barplot of these tabled values using the command

|  |
| --- |
| barplot(table(Dvds),col='red') |

one will see that particular response values are very popular. Is there any explanation for these popular values for the number of DVDs owned?

1. The variable **Height** contains the height (in inches) of each student in the class.
2. Construct parallel boxplots of the heights using the **Gender** variable. Hint: boxplot(Height~Gender)
3. If one assigns the boxplot output to a variable

|  |
| --- |
| output=boxplot(Height~Gender) |

then **output** is a list that contains statistics used in constructing the boxplots. Print **output** to see the statistics that are stored.

1. On average, how much taller are male students than female students? 3
2. The variables **ToSleep** and **WakeUp** contain, respectively, the time to bed and wake-up time for each student the previous evening. (The data are recorded as hours past midnight, so a value of −2 indicates 10 p.m.)
3. Construct a scatterplot of ToSleep and WakeUp.
4. Find a least-squares fit to these data using the **lm** command and then place the least-squares fit on the scatterplot using the **abline** command.

|  |
| --- |
| plot(ToSleep, WakeUp)  fit = lm(WakeUp~ToSleep)  summary(fit)  abline(fit, col='blue', lwd=2) |