Exercise set 1: Data manipulation in R, Ch2

Deadline **Febraury 6**

Submit your answers individually **in learn**

It is enough you provide the R-code, with comments if necessary.

1. Use the R code below to generate *vec1*

   ```
   vec1<-c(0,2,3,0,2,11,0,7,NA)
   ```

   (a) Use indexing to remove the NA.

   (b) Make a logical vector indicating the elements equal to zero

   (c) Use the logical vector to pick out the zero values and store them in a vector called 'zeros'.

   (d) Check how many zeros you have in *vec1* by taking the length of the vector *zeros* (use function *length()*).

2. In this exercise you will take data from a table and store the information in a data frame that you export it to a plain-text file which is easy to work with in R and any other statistical package. The example is very similar to the example given in the course book, pp. 45-47.

   Link: Pages in the book

   (a) Construct a data frame from the table below, including the three variables W (average wage/h), YEAR (including the years for each observation) and Gender (including characters for Women/Men). The number of rows should be as many as the number of values in the table (18).

   (b) Export the data frame as a comma separated (.csv) file.

	2003	2004	2005	2006	2007	2008	2009	2010	2011
Average Wage/h									
Men	120	122	124	130	136	140	143	150	155
Women	109	112	115	121	128	132	135	140	148

3. For this exercise you need to download the data-file *Freedman.xlsx* from the following link: **Link: Freedman.xlsx**

   (a) Import the data to R with either tha package *XLConnect* or *openxlsx*.

   (b) Use summary() on the imported data.

   (c) Sometimes numerical variables are read as character but it is possible to use *as.numeric()* on these variables to define them as numeric. All variables except for the variable *City* should be numeric, make sure they are.

   (d) In preferable way, control the accuracy of the imported version to the original excel file (Debugging). Nothing advanced is needed, just pick out some values in Excel and control that you have the values in the same position in the R data frame. If you like, you can compute some row or column sums in Excel and compare them to the row and column sums you obtain in R.

4. For this exercise we will use the *Prestige* data frame in the package *car*.

   (a) Install the package, if you haven't done this already, and load the data frame.

   (b) Read the help file for the data to learn about the variables (?Prestige)

   (c) Select a subset of the data for occupations with more than 50% women and call the subset 'sub_Prestige'.

   (d) Use the subset and compute the average *prestige* score.

   (e) Now compute the average *prestige* score for occupations with less than 50% women

   (f) For this question use the complete *Prestige* data again. Make a *for*-loop to compute the average prestige score for the three different types of occupations. Automatically store the three means in a vector. The loop should be general, i.e. even if the types of occupations were 100000 one should be able to use your loop.