**Lab\_2 Practical exercises**

1. Using the builtin state.x77 and USArrests datasets:  
• Convert to data table format - use the command option to include the rowname as a column.

* Find union and intersection of the column names of the two data tables.
* Merge the two data tables into a new one called USdata.dt; count the number of rows and columns of the resulting data table (answer: 50, 15) and the number of missing values in the data table (answer: 0).
* Check that the table is ordered by State, then use the setorder() function to reorder by other columns. Then return to order by State.
* Do a scatter plot of the two “Murder” variables and compute their correlation up to 3 significant digits (answer: 0.934)
* Add two new variables to USdata.dt: MeanMurder and MaxMurder to store the state-wise average and maximum of the two “Murder” variables (hint: look at ?max), then remove the two “Murder” variables from the data table.

2. Using the builtin airquality dataframe:

* + Count the number of observed values per column and report the percentage of missing values per column.
  + Make a copy of the dataframe converting to data table.
  + Impute the missing values to the mean. Only for columns with imputed values plot side by side

histograms of raw (unimputed) and imputed columns.

* + Write function impute.to.monthly.mean(x, month) (where month is a vector of the same length of x) that imputes missing values according to the mean value for each month, and repeat the imputation using this function.
  + Report maximum absolute difference (maxi |xi −yi|) and mean absolute difference (between imputation to the mean and imputation to the monthly mean (answer: Ozone: 18.51, 3.55,

Solar.R: 14.07, 0.4).

* + For Ozone only, compare graphically the distributions of the unimputed data, the data imputed to the mean, and the data imputed to the monthly mean and justify the differences you see.

3. Using the diab01.txt dataset (available on Learn):

* Fit a linear regression model for Y adjusted for age, sex and total cholesterol (TC). Create a vector called results.table which stores 4 numbers: regression coefficient and *p*-value for total cholesterol, R2 and adjusted R2 of the model (answer: 0.0925, 0.6999, 0.0319, -7e-04).
* Build all other possible linear regression models for Y using age, sex and one other predictor at a time. For each predictor, append a row of results to results.table in the same format as before. At the end, add row names to results.table to correspond to the predictors used. Identify the predictor that produces the best performing model (answer: BMI).
* Starting from the set of covariates used in the model with best performance determined above, write a loop to fit all possible linear regression models that use one additional predictor. Produce a table of R2 and adjusted R2 of all models you fitted in the loop. Report the adjusted R2 of the model of best fit (answer: 0.252).

4. Using the birthweight.txt dataset (available on Learn):

* + Explore the dataset and prepare it for the analysis (do not impute it).
  + Summarize the distribution of birth weight for babies born to women who smoked during pregnancy and for babies born to women who didn’t. Report the percentage of babies born weighing under 2.5kg in the two strata of smoking status (answer: 8.26%, 3.37%).
  + Fit a linear regression model to establish if there is an association between birth weight and smoking and how much birth weight changes according to smoking.
  + By how much is the average length of gestation different for first born children? Is the difference statistically significant? Is the mother’s pre-pregnancy weight associated with length of gestation?

• Is birth weight associated with the mother’s pre-pregnancy weight? Is the association independent of height of the mother?

• Use the R6 reference class from the lab and modify the P-value default to a threshold of 0.001. Add a new method to the class named qqplot.residuals that returns a qqplot of the residuals from the model already stored in the object. Use qqnorm and qqline and ensure that the function returns the plot as an object that may be reused in a separate document.