Tasks for LuciLinx R Workshop

Part I

Section 4. VARIABLES AND BASIC OPERATIONS

4a. Compare two numbers: e^{π} and π^{e}.
pi, exp(), "^", ">"

4b. Create a vector of exponents of 2: $2^0, 2^1, 2^2, \dots 2^{10}$
i:j, "^"

4c. Output the results of 4b as a vector of strings with template: "2^***i*** = ***x***".
print, sprint

4d. Output the results 4b, showing only even exponents.
print, "%%"

Section 5. DATA IMPORT AND EXPORT

5a. Dataset from http://edu.sablab.net/data/txt/shop.txt contains records about customers, collected by a women's apparel store. Check its structure. View its summary.
read.table, fix, str, summary, head

5b. For the "shop" table, save into a new text file only the records for customers, who paid using Visa card.
write.table

Section 7. DATA VISUALIZATION

7a. Use dataset from http://edu.sablab.net/data/txt/mice.txt. Build distributions for male and female body weights in one plot.
plot, density

7b. Draw boxplots, showing variability of bleeding time for mice of different strains.
boxplot

Part II

Section 8. DESCRIPTIVE STATISTICS

8a. Calculate number of mice with bleeding time bigger than 2 minutes
sum

8b. For dataset Mice replace starting weight of any mice by 1000 (assume, there is a mistype). Calculate mean, median, standard deviation and median absolute deviation (MAD) of this weight. Compare the results with original measures.
mean, median, sd, mad

Section 9. PCA AND CLUSTERING

9a. Acute lymphoblastic leukemia (ALL), is a form of leukemia, or cancer of the white blood cells characterized by excess lymphoblasts. File at http://edu.sablab.net/data/txt/all_data.txt contains the results of a full-trancript profiling for ALL patients and healthy donors using Affymetrix microarrays. The data were downloaded from ArrayExpress repository and normalized. The expression values in the table are in $\log_2$ scale. Perform and visualize PCA for the patients.

Hint: transform data before PCA using t() function.
prcomp, t

Section 10. RANDOM NUMBERS

10a. Test central limit theorem. Build distributions of sum of *n* uniform random variables, where n is 1, 2, 3, 6. Compare the latest (n=6) with a normal distribution.
runif, rnorm, qqplot

Section 11. STATISTICAL TESTS

11a. For Mice data (http://edu.sablab.net/data/txt/mice.txt) compare two weight parameters: *ending weight* and *weight changes* for 2 strains "C58/J" and "CAST/EiJ". Test hypothesis about means and variances these parameters. Try non-parametric Wilcoxon method for as well.
t.test, var.test, wilcox.test

Section 12. ANOVA AND LINEAR REGRESSION

Data presented in the ***leukemia.txt*** (http://edu.sablab.net/data/txt /leukemia.txt) were collected for two groups of patients, who died of acute myelogenous leukemia (AML). Patients were classified into the two groups according to the presence or absence of a morphologic characteristic of white cells. Patients termed AG positive were identified by the presence of Auer rods and/or significant granulature of the leukemic cells in the bone marrow at diagnosis. For AG-negative patients, these factors were absent. Leukemia is a cancer characterized by an overproliferation of white blood cells; the higher the white blood count (WBC), the more severe the disease.

8a. Separately for each morphologic group, AG-positive and AG-negative, draw a scatter diagram to show a possible association between the survival time and the log WBC (take the log yourself) and check if a linear model is justified. If so, estimate a survival time for a patient with WBC = 20000.
lm, log, summary

8b. Create an additional factor – low or high WBC level (e.g. – higher or lower than median) and perform 2-factor ANOVA analysis of the survival time.
aov, as.factor, cbind, summary