Big Data Analytics

Lab 3a

1 Description

Your task for today is to practice the concepts of correlation and regression on different datasets - both manually and using the R functions.

2 Tasks

2.1 Task 1

Explore `Boston` dataset presented in the lecture. After you load `datasets` package, you can find it in the variable `Boston`. Observe the relationships between different variables. Try this:

- scatter plot of two variables - function `plot(...)`,
- you can add a regression line after you draw a plot - function `abline(lm(y x))`,
- correlation of different variables - function `cor(...)`.

2.2 Task 2

You have the following data set, manually calculate:

- correlation coefficient,
- regression function.

Compare your results with R results. You can get the regression line formula with the `lm(...)` function.

x	4	7	5	6	1	5	9	10	10	3
y	33	37	34	32	32	38	43	37	40	33

2.3 Task 4

The following data set gives the average heights and weights for American women aged 30-39 (source: The World Almanac and Book of Facts, 1975).

x	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
y	115	117	120	123	126	129	132	135	139	142	146	150	154	159	164

What is the estimated regression line? Compute Coefficient of Determination. What else can you tell about this data set?

2.4 Task 5

Explore the data set `mtcars`. Perform the analysis of basic statistics for selected variables - these should be numerical variables. Analyse the relationship between the variables.