CST/Math/ASCS Data Mining	Fall 2021
Tu, Th: 2:00pm 3:15pm	**HW 1 UG**

Part I.

This part of the homework must be done by hand + calculators! – no standard functions in R, Mathematica, Matllab or SAS allowed.

You should use either one of the following options to find characteristic polynomial and its roots and eigenvectors of covariance matrix (e.g. http://www.wolframalpha.com/):

- a TI calculator (has to be TI-89 or higher to have required functions),
- R as a calculator (i.e. you need to show results of each intermediate step listed)
- one of the online calculators

You must show the results after each step.

Given data fame X that consists of 6 data objects, each with 2 attributes:

	X1	X2
1	19	12
2	22	6
3	6	9
4	3	15
5	2	13
6	20	5

1. [1pt] Convert data frame into centered matrix
2. [1pt] Compute covariance matrix C_X (with estimated sample mean version)
3. [1pt] Compute characteristic polynomial of C_X and eigenvalues of C_X
4. [1pt] find principal components/rotation matrix P (such that $X = PY$)
5. [1pt] how much variance (%) is explained by new first principal component $\boldsymbol{p}_1$?
6. [1pt] Compute PCA transformation (rotation) of X (without centering) to obtain $Y = P^{-1}X$.

Note:
It is a good exercise to do this in R. If you are using R as calculator then you must provide the script that you have written.
You may need elementary matrix operations:
Multiply matrices A and B: $A \,\%*\%\, B$
Transpose matrix A: $t(A)$
To create list: $v <- c(1,2,3,4)$
To coerce list to matrix: $as.matrix(v)$ makes it 1 row matrix

To compute characteristic polynomial in R use package pracma, $charpoly$ function. For eigenvectors you can use $eigen$ function (in base – no packages needed).

Part 2

7. Imagine that you have run PCA on data gathered from one of the questionnaires gathered by the car manufacturer in which 10'000 people gave their age, gender, country of residence and the car model they purchased.

i) [1pt] You find that all N eigenvectors (N=space dimensionality of the dataset) cover the same % of the data variance. What is N here? What is this % of variance?

ii)[1pt] How would you interpret the results in i? What if 1 eigenvector cover 99% of the data variance.