Homework 2
CS699 A1, Spring 2025

Due: 2/10

- You must show all calculations and important intermediate steps/results. Otherwise, you will lose points even if your answers are correct.
- If you use R, you must submit the R code file.

Problem 1 (5 points). Consider the following contingency table that summarizes values of two categorical variables:

	Red	Green	Blue
Low	12	6	21
Medium	23	15	22
High	10	13	19

Using the chi-square test method that we discussed in the class, determine whether there is a correlation between the two variables. Use significance level 5%.

You should not use any software tool for this problem, except for only calculation purposes. You must calculate expected values and the test statistic yourself.

Problem 2 (5 points). Use *hw2_p2.csv* file for this problem. The dataset has 4 variables and 100 tuples. You may use any tool for this problem, including *R* and *Excel*.

(1). Show the correlation matrix (with all four variables).
(2). Which two variables have the strongest correlation?

Problem 3 (10 points). Use *hw2_p3.csv* file for this problem. Use R for this problem.

(1). Standardize all variables using the z-score method.
(2). Apply PCA to the standardized dataset and show the screenshot of the summary of the result.
(3). How many principal components do you need if you want to keep more than 80% of total variance? How many principal components you need if you want to keep more than 90% of total variance?
(4). Show the first six tuples of the transformed dataset (which has principal components as new variables).

Submission: You need to submit the following two files:

- *hw2.doc* or *hw2.pdf,* which includes answers to all problems.
- *hw2.R,* which was written for problem 3.

Include all files in a single archive file and name it *LastName_FirstName_HW2.EXT*. Here, "*EXT*" is an appropriate file extension (e.g., *zip* or *rar*) and submit it to Blackboard.