Practical Data Science
(Multivariate Stats)

Write R Scripts or use R to perform any mathematical operations while solving the following problems.

Problem 1: Employees who have similar profiles

You are given a sample data set of employee details. Assume that two employees are similar if the Euclidian distance between any two employees is very less. You can represent each employee as a vector/point in 3-dimensional space.

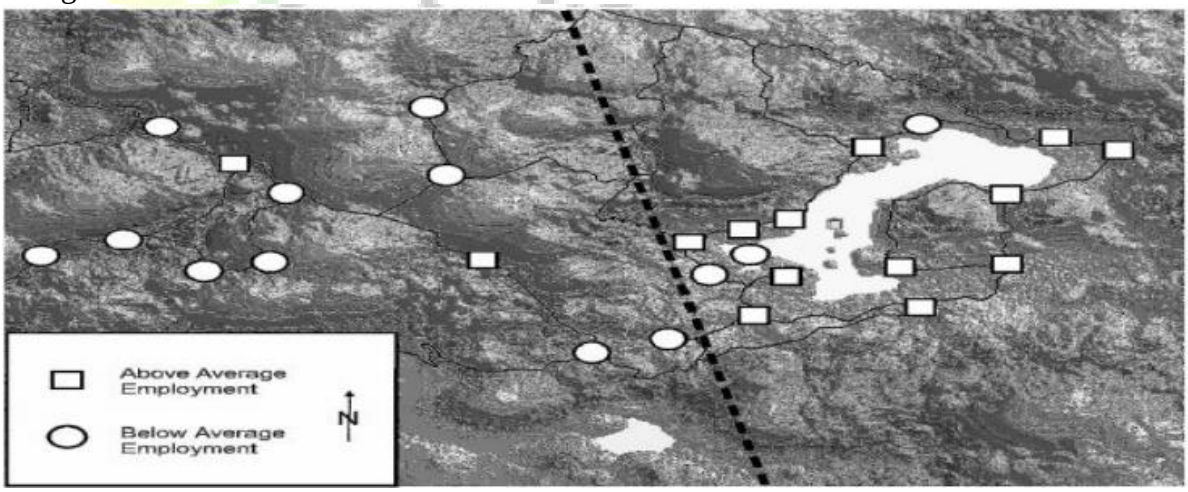
Empid	Salary	Age	Experience
1	25000	24	4
2	40000	27	5
3	55000	32	7
4	27000	25	5
5	53000	30	5
6	26000	35	10

Find the employees who have similar profiles,

a. without normalizing the data
b. with 0-1 normalization
c. with Z-score normalization

Problem 2: Living place vs Education

The map below shows employment levels for two groups of villages in India: those near the lake and those in the mountains (the dashed line separates these two groups). You have to determine whether there is a difference in employment between these two groups. In other words, you have to find whether there is significant relationship between the living area and education variables.



Practical Data Science
(Multivariate Stats)

Problem 3: Fair die or not

Alice has a die which has faces numbered from 1 to 6. He observed following frequencies out of 120 trials.

Die value:	1	2	3	4	5	6
Frequency:	17	20	29	20	18	16

He thinks that the die may be biased. What do you think?

Problem 4: Weekday vs Class absentees relation

The table below shows the number of students absent on particular days in a week.

Day:	M	Tu	W	Th	F
Num of Absentees:	125	88	85	94	108

Find the expected frequencies if it is assumed that number of absentees is independent of the day of week. Do the observed and expected frequencies are significant? What are the chi-square statistic and p-values?

Problem 5: Transport Survey

In the survey of transport, electors from three different areas of a large city were asked whether they prefer money to be spent on improving road transportation or not. The replies are shown in the following table:

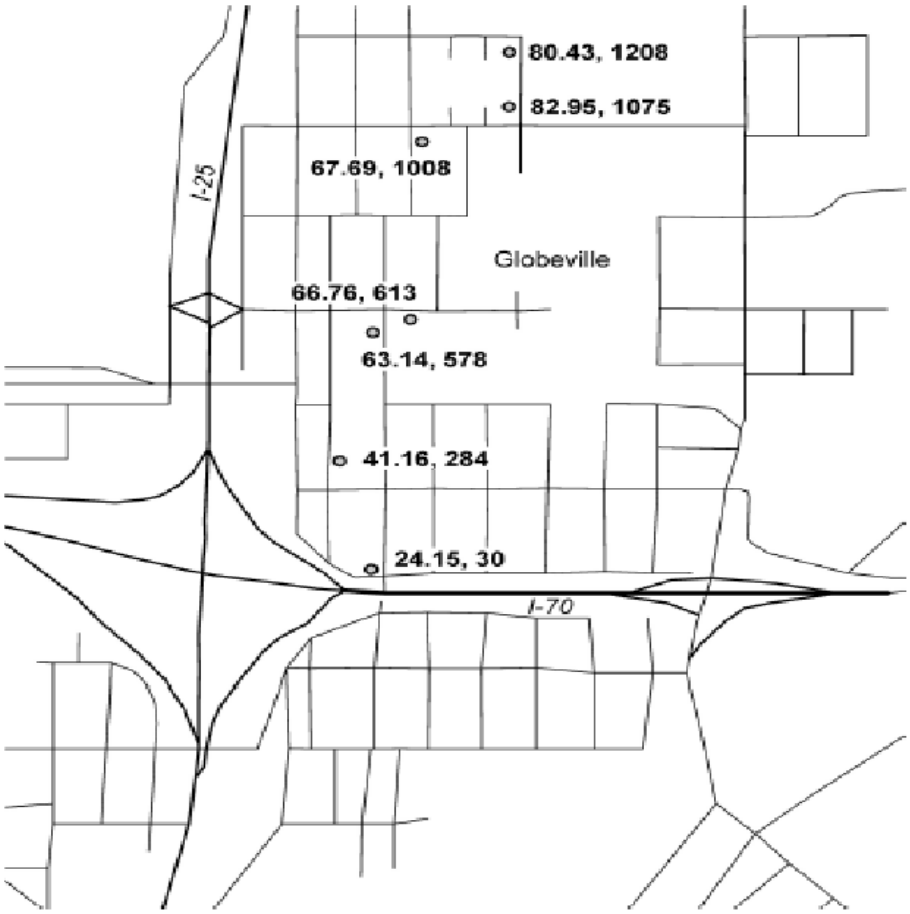
	A	B	C
Road Transport Preferred:	78	46	24
Non-road Transport Preferred:	22	34	36

Is the area independent of transportation preferred?

Problem 6: Near Highway Property valuable or not

The map below depicts the value per square foot for 7 houses in the Globeville neighborhood of Denver, Colorado. The elevated portion of I-70(National highway) is shown as a heavy black line on the map. Each house is labeled with the value per square foot and the distance from I-70 in meters. Is there correlation between property value and distance variables?

Practical Data Science
(Multivariate Stats)



Problem 7: Covariance, Standardized Covariance & Correlation matrices

The dataset Olympics.dat contains the result of the decathlon at the olympic games in Atlanta at 1996 and is available at algorithmica repository. Answer the following questions:

a) Find the covariance matrix for given dataset.
b) Apply z-score transformation to covariance matrix and find the resultant matrix.
c) Find the correlation matrix for given dataset.
d) Convert the covariance matrix to correlation matrix using cov2cor() function.