Department of Computer Science and Engineering

Ramaiah Institute of Technology

Bangalore-560054

Data Analytics Laboratory – CSL717

Question Bank

1. **i) Store all dept (9 depts) students names,USN,Dept names, 5 subject grades and SGPA in csv file using ms excel.**

a. Extract each dept students names separately.
b. Extract S grade scores in all subjects in each dept seperately.
c. Extract students who have scored at least S grades in any 2 subjects
d. Extract students who have scored above 9 SGPA in each dept

1. **Store all dept faculty names with designation and salary details.**

a. extract each dept faculty details separately
b. extract Professors of each dept seperately
c. extract people who earn more than 1.5 lakh in each dept where their designation is prof, associate or assistant
d.Find out the cost of professors in each dept .(sum up their salary to get cost of them)
e. Find the cost of each dept faculty
f. Find out the average cost of faculty in each dept.
g.Which dept has highest average cost of faculty
i. Which dept has lowest cost of faculty

---

2. **Store all dept (9 depts) students names,USN,Dept names, 5 subject grades and SGPA in csv file using ms excel.**

i. Store students marks numerically, transform into grades and store in new dataframe
ii. Check whether students grades are identical or not in each subject
iii. Extract students' marks in each subject separately. If the student has scored greater than 80 map it as "good", if it is between (80 and 60) map it as "moderate", if it is between (40 and 60) map it as "need improvement", else map it as "poor".
iv. Consider dataset given in (1.a) , map the S,A, grades as " GOOD"; map ,B,C grades as "average", D,E grades as "below average"; 'F' grade as "poor".
v. Transform dept names to numerical data.
vi. Using factor() and mapvalues() convert dept names to numerical data.
vii. Create table from student data with USN and names only.
viii. Display the typeof each column.
ix. Write separate functions to perform all the above functions separately and call them in R script.
x. Write a function to perform statistical analysis of students data.

xi. Use sapply to perform 1.c, 1d

---

3. Store all dept faculty names with designation and salary details.

a. Store faculty salary numerically, transform into factor.
(eg. 50000 to 75000 as 1, 75000 to 100000 as 2 , and so on)
b. Check whether faculty paper publication count and number of training program attended are same or not.
c. Extract paper published count separately. If the count is grater than 15 map it as "Excellent",
if it is (10-15)map it as "good", if it is (5 to 10) map it as " moderate"), if it is (1-5) map it as "need to improve", else "poor, start your research".
d. Consider the solution of 1c. Map "Excellent", "good" as " Good performers".
Map " Moderate" and "need to improve" as " Ok, Keep it up". Else " map to " You may be fired!"
e. Transform designations to numerical data
f. Using factor() and mapvalues() convert designations to numerical data.
g. create table for faculty names and designations.
h.Write separate functions to perform all the above functions separately and call them in R script.
i. Write a function to perform statistical analysis of faculty data to identify faculty performance of depts..
j. Use sapply to perform all the above functions.

---

4. Use student data set

i. Plot , in each dept, how many students have scored above 9 SGPA

ii. Create subset of students , who have scored S grade in any subject and failed in any subject.

iii.Find out average SGPA of each dept students.

iv.Find out average score of each subject for each dept.

v. Extract 10 toppers of each dept.

vi. Sort students details of each dept separately.

vii.Search for a particular student name in the data set, and retrieve his/her details.

Use faculty data set

i. Plot, in each dept, how many faculty are earning more than 1 lakh.
ii. create subset of faculty, who have published more than 10 papers and their designation is Associate professor.
iii. Find out average papers published by each dept, designation wise.
iv. Find out Average training programs attended by faculty each dept wise.

v. Extract top 3 performers among faculty dept wise(more papers published and more training programmes attended)
vi. Sort faculty details , dept wise separately.
vii. Search for faculty name in the data set and retrive his/her details.

---

5. Use Faculty data set

a. Change the column names of faculty data set.

b. Use map values() , as.factor() and transform ()

- to change the designation column to have numerical values. 1- Prof, 2-Asso.Prof, 3-Asst.Prof;
- to change gender column 1-Male, 2-Female

c. Using with() and tapply() , calculate the mean training programs attended and no. of papers published in each department. Format it for markdown.

d. Using with() and aggregate() , calculate the mean training programs attended and no. of papers published in each department. Format it for markdown.

e. Check whether the mean value of no. of papers published in depts. are influenced by training programs attended and designation. Do regression analysis using aggregate().

f. Create the table output for designation and papers published. Use with() and table().

g. Find the odds of lower no. of paper published with respect to designation and no.of training programs attended using the output of question (f).

h. Is the designation affects the training programs attended? Check it with the data. Prove it.

i. Find the correlation of papers published and training programs attended using with(), cor().

j. Find the correlation of training programs attended and designation using with(), cor().

k. Using by() combine the operations of above questions (i) and (j) using function. And do the correlation analysis using cor() with in the function.

l. Plot average training programs attended against designations(only 3 designations) of the institution using plot()

m. Plot average papers published against designations of the institution using plot().

n. Change the x axis , color and y-axis labels respectively. Add legends.

o. Plot the above graphs in (l) and (m) using with() and plot().

p. Draw scatter plot for above questions. Draw box plot for above question. Draw bar plots for above questions. Draw single variable plots for above question.

q. plot the prof, asst.prof and asso.prof average performance in different colors using rep(), colorpalatte functions. Represent each desgination average performance by different symbols.

---

6. Use student data set

a. Make some string entries in student marks. Make some numerical entries in names

b. Convert the marks to numerical data and show.

c. Display the mark as factors.

d. Display the marks as characters

e. Display the type of marks columns

f. Using gsub remove character data in marks column. Using gsub remove numerical data in name column

g. Clean the name and marks column and put it in a new student data set variable using transform function

h. Use table to get summary of student data

i. Use sapply to perform the cleaning of data mentioned above

j. Use lapply to perform the cleaning of data mentioned above.

k. Display the summary using summary()

L. Include gender details for students.

m. Add Mr or Ms. For each student, using paste command . Display all students details.

n. Define user defined functions to perform the above operations.

o. Use while loop, for loop to access students marks and find the grades. Put this with in a function

p. Use apply, sapply, lapply and tapply to perform the above operation over all the columns of students data set.

q. Use with() function to apply the above operation over students dataset.

r. Use any() function to apply the above operation over students dataset.

---

7. Use Student data set

a. Change the column names of Student data set.

b. Use map values() , as.factor() and transform ()

- To change the Grade column(S,A,B,etc) to have numerical values. 1- S, 2-A, 3-B,etc.;
- To change gender column 1-Male, 2-Female

c. Using with() and tapply() , calculate the mean of marks in each subject scored by students in each department and mean of CGPA of students in each dept with respect to gender. Format it for markdown.

d. Using with() and aggregate() , calculate the mean of marks in each subject scored by students in each department and the gender of students in each dept. Format it for markdown.

e. Check whether the mean value of each subject marks in depts influenced by the gender of students in each dept or not. Do regression analysis using aggregate().

f. Create the table output for mean scores in subjects in each dept gender wise. Use with() and table().

g. Find the odds of lower no. of mean marks in subjects with respect to gender of students in the depts using the output of question (f).

h. Is the Gender of students affects the CGPA of students? Check it with the data. Prove it.

i. Find the correlation of Gender of students and Marks and CGPA scored using with(), cor().

k. Using by() combine the operations of above questions (i) and (j) using function. And do the correlation analysis using cor() with in the function.

l. Plot average marks in subjects , CGPA scored against depts of the institution using plot()

m. Plot average marks in subjects , CGPA scored against gender in each dept of the institution using plot().

n. Change the x axis , color and y-axis labels respectively. Add legends.

o. Plot the above graphs in (l) and (m) using with() and plot().

p. Draw scatter plot for above questions. Draw box plot for above question. Draw bar plots for above questions. Draw single variable plots for above question.

q. plot the female students and male students performance in different colors using rep(), colorpalatte functions. Represent each gender's average performance(dept wise) by different symbols.

---

8. **Use Student data set**

a. Change the column names of stud data set.

b. Use with() and plot(), plot graph against dept and average CGPA semesterwise.

c. Use qplot(), to plot the same mentioned in (b)

d. Use qplot() with attributes color, shape, x and y labels.

e. Display the dimension of stud data set.

f. using ggplot() and geom_point() , plot the same mentioned in (b)

g. in ggplot() , change the point size.

h. change color of points in ggplot().

i. define your own color palatte and use it in ggplot()

j. Using facet_wrap(), plot for (f), with respect to dept

k. Using facet_grid(), perform the (j)

l. Draw bar graph for (b) using ggplot and geom_bar().

m. Add legends in graph.

n. Draw smooth curves using stat_smooth() and by changing size for (b).

o. Add title to ggplot() of your graph and draw regression lines and geom points for (b)

p. Draw ggplot() for dept wise students' average CGPA.

q. Draw grom Polygon for (p)

---

9. Use Faculty data set

a. Change the column names of faculty data set.

b. Use with() and plot(), plot graph against dept and average Papers published.

c. Use qplot(), to plot the same mentioned in (b)

d. Use qplot() with attributes color, shape, x and y labels.

e. Display the dimension of faculty data set.

f. using ggplot() and geom_point() , plot the same mentioned in (b)

g. in ggplot() , change the point size.

h. change color of points in ggplot().

i. define your own color palatte and use it in ggplot()

j. Using facet_wrap(), plot for (f), with respect to desgination.(Prof, Asso prof, Asst Prof)

k. Using facet_grid(), perform the (j)

l. Draw bar graph for (b) using ggplot and geom_bar().

m. Add legends in graph.

n. Draw smooth curves using stat_smooth() and by changing size for (b).

o. Add title to ggplot() of your graph and draw regression lines and geom points for (b)

p. Draw ggplot() for dept wise average training programs attened by Asst prof.

q. Draw grom Polygon for (p)

---

10. Use Student data set

i. Compare the no. of students scored above 9 CGPA by departments(eg. CSE,ECE,CV,ISE,MEch) and gender wise(F,M). Create boxplot using qplot ,showing how these counts varies between departments according to gender.
ii. Assess whether the difference in (a) is statistically significant using aggregate().
iii. Assess whether the difference in (a) is statistically significant using two sample t.test(). Assess access the t-test information element by element.
iv. Calculate difference in means of CGPA between different depts(eg. CSE,ECE,CV,ISE,MEch)
v. Using with() and t.test() perform (c).
vi. Generate 1000 simulation data for student dataset using user defined function for controlled group (500 data)and treatment group(500 data).
vii. Draw box-plots, density plots and do t-test for (f)
viii. Iterate the operation in (f) for 50 times and do t-test each of the time. Capture P-values of t-test for all 50 times and plot using q-plot() and ggplot() and compare.
ix. Perform non-parametric test using wilcox.test() to do (b).
x. Do (i) using with().
xi. Check the class of your result in (i).
xii. Check the non-normality in (b) using qqnorm(), qqline() and rnorm().
xiii. Use your simulate function to simulate data and do (l).
xiv. Use table function to create summary for (b). Perform fisher.test(). Check the attributes.Perform chi-square test() over the summary data.
xv. Plot the proportions of (b) or (n) using melt() , ggplot(), geom_bar() and geom_errorbar().