**CIA - 4**

**(Semester - II)**

**PROGRAMME NAME: CMS/EMS                                       MARKS: 40**

**COURSE NAME: R Programming                                      Date:10.3.2020**

**COURSECODE:STAT232                                                                                                     INSTRUCTIONS**

·**Make sure your mobile phone is switched off and placed in the designated   places.**

·**Malpractices will be viewed very seriously**.

· **Read the questions carefully and choose any one question for answering within**

**first half an hour from the commencement of examination.**

· **Once the question is chosen it cannot be further changed at any cost.**

· **Solve the question and submit the report in Google class room. Further take a**

**hardcopy of the report and submit to the concerned teacher.**

· **The report should contain proper title, introduction, analysis and conclusion.**

· **Usage of ggplot2, dplyr, lattice packages are compulsory for all the questions.**

**Answer any one question:**

**Q1**:**LungCapDataset**

**The variables of LungCap Dataset are described below**

* ***LungCap*** It’s the lung capacity of the person
* ***Age*** It’s how old is the person
* ***Height*** It’s how tall is the person
* ***Smoke*** If the person smokes or doesn’t smoke
* ***Gender*** If are male or female
* ***Caesarean*** If they’re born by caesarean

1.Import the LungCapData into R and  find the summary of the dataset

2.Find the strength of the relationship between Age and Height and perform a scatterplot

3. Perform all the graphical visualizations as below

     a. Add a title to the plot and Label x-axis and y-axis

     b.Rotate the values on the y-axis and also change the limits of x-axis and y-axis

     c.Change the size of the plotting characters and change the color of the characters to red.

4.Draw the regression line to the plot predicting height using age and change the color of the line to blue and the width of the line

5.Identify gender on the plot for the age male and female

**Q2:UScerealDataset**

The UScereal data set has been collected from the 1993-ASA Statistical Graphics Exposition, and is taken from the mandatory F&DA food label. The recorded variables are:

* mfr-Manufacturer, represented by its first initial: G=General Mills, K=Kelloggs, N=Nabisco, P=Post, Q=Quaker Oats, R=Ralston Purina.
* calories - number of calories in one portion.
* protein - grams of protein in one portion.
* fat - grams of fat in one portion.
* sodium - milligrams of sodium in one portion.
* fibre - grams of dietary fibre in one portion.
* carbo -grams of complex carbohydrates in one portion.
* sugars -grams of sugars in one portion.
* shelf-display shelf (1, 2, or 3, counting from the floor).
* potassium - grams of potassium.
* vitamins-vitamins and minerals (none, enriched, or 100%).

Import the given "UScereal" dataset to investigate the following

1. Understand and prepare dataset for analysis. Calculate summary statistics of a dataset and find maximum protein value of each Manufacturer

2. Using the *ggplot2* package, generate any three different plots using appropriate variables. Use appropriate plot to find the outlier on *calories* for each *shelf*. Explain your findings on generated plots.

3. Find the strength of the relationship among variables, plot the relationship and explain your understanding.

4.   Create a simple linear regression model and plot it. Explain your understanding.

**Q3**:**Vocab Dataset**

The observations in the “Vocab” dataset are the respondents to U.S. General Social Surveys, 1972-2016. The variables in the dataset are:

year: year of the survey

sex: sex of the respondent, Male or Female

education: number of years of education

vocabulary: vocabulary test scores

Import the “Vocab” dataset in R and answer the following:

1. Obtain the basic statistical summary of each variable of the given data set and explain it.
2. Construct and explain any two suitable bivariate plots with important graphical parameters using the variables sex, education and vocabulary.(Use ggplot2 package)
3. Obtain the degree of linear relationship between education and vocabulary test score, and represent it with the scatter diagram. Explain your findings.
4. Set the equation of linear relationship between education and vocabulary test score and answer the following:

i. Show the obtained line of regression in the above scatter diagram with red color.

ii.Find the average vocabulary test score if years of education are 15.

iii.What is the average vocabulary test score if years of education are zero.

**Q4: ChickWeight data set**

The ChickWeight dataset is a data frame which represents the progression of weight of several chicks. The little chicklings are each given a specific diet. There are four types of diet and the farmer wants to know which one fattens the chicks the fastest.

1. Obtain the basic statistical summary of all numerical variables of the given data set and explain it. Are there any missing observations?

2. Use ggplot and map  time to x and weight  to y. Add color=diet and draw conclusions.

3. How are weight and time correlated? What about diet and weight? Explain your findings.

4.Form a linear equation of linear relationship between time and weight  and answer the following questions:

a)   Obtain scatter diagram and display line of regression in  blue color .

b) What will be  weight of the chicken at time=11

c)   What should be the time for weight=185?

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* ALL THE BEST\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*