DWDM LAB EXERCISES – 23 JULY 2019

Scax: NOTATION FOR SCALAR VARIABLES

Vecx: NOTATION FOR VECTOR VARIABLES

ls()

c()

rep()

table()

matrix()

dataframe()

seq()

rm()

data()

read.csv()

read.table()

comments: #

ctrl + L : to clear the console

summary()

sd()

quartile()

boxplot()

plot()

getwd()

setwd()

view()

str() : internal structure

library()

length()

dim() : vector dimensions

ls()

names(): data frame headers / cols/ attribs

hist()

xtabs()

by()

mean()

median()

array()

print()

assignment operator <- or =

Arithmetic operators

Relational operators

Logical operators

R statements: if, if…else, switch, repeat, while, for, break, next

functionname <- function(arg list)

load()

save()

DWDM LAB EXERCISES – 30 JULY 2019

STATISTICAL FUNCTIONS

1) Mean
2) Median
3) Min
4) Max
5) Var
6) Summary
7) Hist
8) Table
9) Boxplot
10) Plot
11) Xtabs()

Data related commands

1. read.csv()
2. read.xls() //requires some additional pacages to be installed to get it work.
3. library()
4. data()
5. view()
6. str()
7. require()
8. dim()
9. length()
10. names()
11. head()

12. rownames()

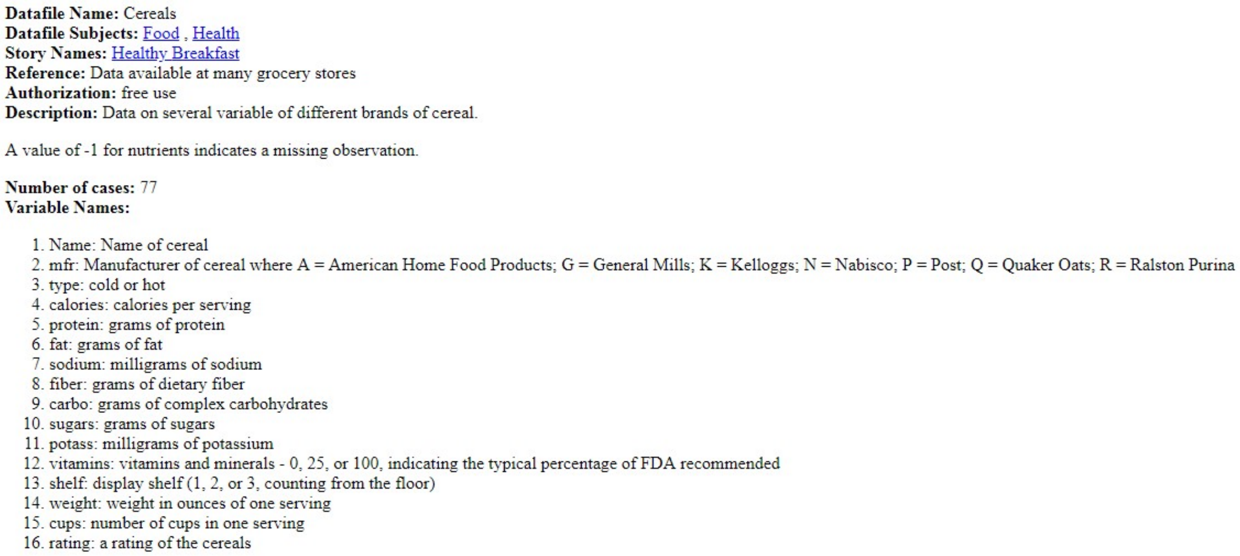
DWDM LAB EXERCISES – 30 JULY 2019

1. Tabulate number of cylinders in the mtcars dataset table.mtcars$cylinder
2. Find the five number summary of milespergallon
3. Draw a histogram for hp
4. Box plot the miles per gallon
5. Find the avg weight of all cars
6. Find the car with the minimum displacement
7. Find the car with the maximum qsec
8. Find the median horse power and find the car with the highest fuel efficiency
9. Find the car with the lowest fuel efficiency
10. Find the car with the best hp
11. Tabulate mpg for different number of gears
12. Do side by side boxplot of mpg for cars with am (automatic transmission) Vs cars with Manual transmission

DWDM LAB-06AUG-2019

\\172.16.68.115

Cereals1 data set: First convert the .xls into ,.csv and import into R.



Datafile Name: Cereals
Datafile Subjects: Food , Health
Story Names: Healthy Breakfast
Reference: Data available at many grocery stores
Authorization: free use
Description: Data on several variable of different brands of cereal.

A value of -1 for nutrients indicates a missing observation.

Number of cases: 77
Variable Names:

1. Name: Name of cereal
2. mfr: Manufacturer of cereal where A = American Home Food Products; G = General Mills; K = Kelloggs; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina
3. type: cold or hot
4. calories: calories per serving
5. protein: grams of protein
6. fat: grams of fat
7. sodium: milligrams of sodium
8. fiber: grams of dietary fiber
9. carbo: grams of complex carbohydrates
10. sugars: grams of sugars
11. potass: milligrams of potassium
12. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended
13. shelf: display shelf (1, 2, or 3, counting from the floor)
14. weight: weight in ounces of one serving
15. cups: number of cups in one serving
16. rating: a rating of the cereals

1) Tabulate the following attributes: mfr, and type of cereals

2) Display the 5 number summary for all nutritional attributes starting from calories to vitamins

Note: A value of -1 in nutrients indicates a missing observation.

3) For missing values find and replace with arithmetic mean of the attributes.

4) Find and replace outliers with median.

5) Compare the 5 number summary before and after preprocessing.

6) Draw side-by-side box plots of Calories of Hot Vs cold cereals.

7) Are the attributes calories and consumer rating correlated?

8) Are the attributes mfr and consumer rating correlated?

9) Which is the best Vs worst cereal in terms of user rating?

10) Which is the best Vs worst cereal in terms of calories?

11) Rate the top 5 cereals in terms of user rating?

12) Identify the cereal with the highest sodium.

13) Identify the cereal with the lowest carbohydrate.

14) Are the variable shelf and sugar correlated?

15) Identify the manufacturer of the cereal with the highest sugar content.

cor: The built-in function of R.

factor(): Converting character data type inot nominal

chisq.test():

You will be doing preprocessing.

Replace missing values with mean

Find outliers using 1.5 * IQR

USING WEKA – 23AUG2019

1. Preprocessing: Use numeric to nominal filter
2. For question 2 change property "car" = 2 and set index to cook books.

1. FOR THE BANK.xls dataset perform preprocessing and find atleast 5 interesting rules with pep on the RHS of the rule.

2. Use the supermarket.arff dataset and perform the following tasks:

   (i) Split the dataset into 2 datasets: one containing items and another containing departments.
   (ii) For the item dataset find the most frequent itemsets ranked by lift.
   (iii) Find the top five selling items in the dataset
   (iv) For the topmost selling item, find association rules with the item on the RHS of the rule.
   (v) Find the top 5 association rules for the department dataset ranked as per the lift.
   (vi) Find top 5 association rules with total in the RHS using the item dataset.

1. Using the Weather Nominal dataset and J48 algorithm, write down the classifier rules where the class label is play. Compare the accuracy of the classifier with the confusion matrix for two testing strategies: cross-validation and 66% split.

2. Diabetes.arff: Using this dataset perform required preprocessing and generate the top 5 association rules using the apriori algorithm with minimum lift of 1.

3. Using the same dataset, create classifiers using Naïve Bayesian and the Bayesian Belief Network. Compare the classifier accuracy using a Confusion Matrix.
1. Use the bank dataset. Perform preprocessing and find top 5 classifier rules with pep as the class label on the RHS.

2. Use the diabetes dataset, perform preprocessing, and find all relevant classifier rules for class equals tested positive using C4.5. Visualize the classifier tree generated. Rank the predictor variables as per their relevance to the class label. Compare the accuracy of C4.5 against KNN for k = 1 and 3.

```
g1 <- subset(mtcars, gear == 3, select = c(mpg, gear))
g2 <- subset(mtcars, gear ==4, select = c(mpg, gear))
g3 <- subset(mtcars, gear == 5, select = c(mpg, gear))

gears1 <- c(g1, g2, g3)
```

```
avg1 = mean(g1$mpg)
avg2 = mean(g2$mpg)
avg3 = mean(g3$mpg)
```

or

```
gears2 <- c(mean(g1$mpg), mean(g2$mpg), mean(g3$mpg))

avgvecx1 <- c(avg1, avg2, avg3)

gears1 <- c(3, 4, 5)

xtabs(~avgvecx1 + gears1)
```

```
                  gears1
avgvecx1           3 4 5
```

```
16.1066666666667 1 0 0
21.38            0 0 1
24.5333333333333 0 1 0
```

15 Marks

7 marks from introduction and preprocessing:

4 marks numericals +

8 marks from association rule mining and data warehousing

4 Marks numericals

Assignment -02: 26-aug-2019 to 31 aug 2019

Topics: Association rule mining

Monday 26-Aug-2019