Date 21/01/2023

# Datawave house and Data mining

CSA1675

Assignment-1

M. Venkata Hari
192011164
CSE

## Read:-

### Creating a file:-

using file.create() function a new file can be created from console or translates if already exist.

syntax:

```
file.create(" ")
```

Example:-

```
# create a file
# the file created can be seen
# in your working directory.
```

### Reading a file:-

using read table() function R, files can be read and output is shown as data frame.

Syntax:-

```
read.table(file)
```

Example:-

```
# Reading text file
new.iris <- read.table(file = "GFG.txt")
# print
print(new.iris)
```

## writing into a csv file:

R can create csv file form existing, data-frame the write.csv() function is used to create the csv file. this file gets created in the working directory.

```
# create a data frame
data <- read.csv("input.csv")
retval <- subset(data, as.Data(start-date) > as.Date("2004-10-01"))
# write filtered data into a new file.
write.csv(retval, "output.csv")
newdata <- read.csv("output.csv")
print(newdata)
```

## Data Manipulation

round(x,n):-

```
# round the values of x to n decimal places
```

ceiling(x):-

```
# vector x of smallest integer
```

floor(x):-

```
# vector x of largest integer
```

as.integer:-

```
# Truncates real x to integer
```

(compare to round(x,0)).

### Statistics

min() → lowest value from given data

mean() → Average value

median() → middle value Q1, Q2, Q3

Sum():- → Total

var() → # produces the variance covariance matrix

sd() → # Standard deviation

## Transformation:-

five num:-

```
# Turkey five numbers min, lowerhinge, minge, median, upper hinge max.
```

Table() →

```
# Frequency counts of entries ideally the entries are factors (although it works with integers (or) even reals).
```

scale(data, scale = T)

→ # centres around the mean and scales by the sd

### Input and display:-

read table (filename, header (True) → #read files with tables in first row

```
# read a table (or) space. delimited file.
read.table(filename, header = True, sep = ".")
# read csv files
x = c(1:10) → # create a data vector with elements 1-10
```

rect = c(x,y) → # combine them into vector or length 2n.

mat = cbind(x,y) → # combine them into an x2 matrix

(2) Suppose that the data for the analysis includes the attribute age. the age values for the data tuples are 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

*) the first quartile ($Q_1$) is the 25th percentile and the third quartile ($Q_3$) is the 75th percentile in a data set.

To find $Q_1$ and $Q_3$, we first need to order the data set and find the medians. for odd number of elements in the dataset.

Median = $(N+1)/2^{th}$ element of the sorted dataset, where N is the number of elements in the dataset

for even number of elements in the dataset Median = ($N/2^{th}$ element + ($N/2+1$)th element)/2 of sorted dataset where N is the number of elements in the dataset.

Here we have 26 elements in the dataset, so the median is the average of the 18th and 14th elements which are 19 & 20 respectively

Therefore $Q_2 = (19+20)/2$

$= 19.5$

Now that we have $Q_2$, we can find $Q_1$ and $Q_3$ by finding the median of the lower and upper halves of the data set respectively

for the lower half of the dataset, we have the following values

13, 15, 16, 16, 19

the median of this set is 16, so $Q_1 = 16$

for the upper half of the data set, we have following values.

20, 20, 21, 22, 22, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

the median of this set is 35, so

$Q_3 = 35$

therefore the first quartile ($Q_1$) = 16 and the third quartile ($Q_3$) = 35