DATAWARE HOUSE AND DATA MINING

ASSIGNMENT-1

VARALA SIRISHA
192011492
CSA1675

Read:

Creating a file:

Using file.Creater, function, a new file can be created from console or trancates 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 dataframe

Syntax:

Read.table(file)

Ex:-

```
# Reading txt 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 dataframe
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 integers >x.
floor(x) # vector x of largest integer <x.
as.integer # truncates real x to integers
             (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

```
fivenum() → # Tukey five numbers min,
            Lowerhinge, median, upper hinge, max,
Table() → # frequency counts of entires,
          ideally the entires are factors
          (although it work with integers
          or even reals)
Scale(data, scale = T) ↓
  # centers around the mean
  and scales by sd).
```

Input and display

```
read table(file name, header=
                    True) →
# read files with tables in first
  row
# read a tab 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
Vect = c(x,y) → # Combine them
               into vector (or) length 2n.
mat = (bind(x,y) → # Combine them
       into an x 2 matrix.
```

2) The age values for the data types are

18, 15, 16, 16, 19, 20, 20, 21, 22, 25, 25, 25, 30, 33, 33, 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 data set, 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 $Q_2 = \frac{(19+20)}{2}$

$= 19.5$

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

For the lower half of the datasets, 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 dataset, 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, 30

$Q_3 = 35$

∴ The first Quartile (Q_1) = 16 and the third Quartile (Q_3) = 35

$Q_1 = 16$; $Q_3 = 35$