Date 03/02/2023. Data warehouse and Data mining CSA1675

M. Venkata Hari 192011464 CSE

Assignment - 2

① Imagine that you have selected data from the all electronics data warehouse for analysis. The data set will be huge. The following data are sorted: 1,1,5,5,5,5,5,8, 8,10,10,10,10,12,14,14,14,15,15,15, 15,15,15,18,18,18,18,18,18,18, 20,20,20,20,20,20,20,21,21,21, 21,25,25,25,25,25,28,28, 30,30,30. (i) partition the dataset using an equal-frequency partitioning method with bin equal to 3 (ii) apply data smoothing using bin means and bin boundary (iii) plot histogram for the above frequency division

(i) partitioning using equal frequency:

We divide the data set into 3 equal frequency bins, each containing the same number of observations. To calculate the bin boundaries, we count the number of observations in the data set and divide that by the number of bins. In this case 3. Each bin will contain 40/3 = 13 observations.

The bin boundaries for equal frequency partitioning method are:

Bin 1 :- 1 - 12

Bin 2 :- 12 - 21

Bin 3 :- 21 - 30

ii) Data smoothing using bin means and bin boundaries:

For data smoothing, we calculate the mean of the each bin and use that as the representative value for all observations in that bin.

Bin 1: Mean (1+1+5+5+5+5+5+8+8+10+10+10+10)/13 = 6

Bin 2: Mean = (10+10+10+12+14+14+14+15+15+15+15+15+15)/13 = 15

Bin: Mean = (15+15+15+18+18+18+18+18+20+20+20+20+20+20+20+20+21+21+21+21+25+25+25+25+25+28+28+30+30+30)/13 = 24

The bin boundaries for smoothed data using bin means are:

Bin 1 :- 6 - 12

Bin 2 :- 12 - 21

Bin 3 :- 21 - 30

iii) Plotting histogram:

Using bin boundaries obtained from either equal frequency (or) data smooth we can plot a histogram by creating bars of the same width that span the bin boundaries and the height of each bar is proportional and the frequency of observations in that bin. The x-axis represents the price of the item and the y-axis represents the frequency of observations.

R-Program:-

Load the ggplot2 library

```
library(ggplot2)
```

Create a vector of the prices

```
data <- c(1,1,5,5,5,5,5, 8,8,10,10,10,10,12,14,14,14,15,15,15,15,15,15,18,18,18,18,
18,18,18,18,20,20,20,20,20,20,20,20,21,21,21,21,25,25,25,25,25,28,28,30,30,30)
```

Partition the data using equal frequency partitioning with bin = 3

```
binned_data <- cut(data, breaks = 3, lables = c("1-19", "20-39", "40+"), right = false)
```

calculate the bin means

```
bin_means <- tapply(data, binned_data, mean)
```

calculate the bin boundaries

```
bin_boundaries <- c(-inf, 19, 39, inf)
```

Apply data smoothing using bin means and bin boundaries

```
smoothed_data <- cut(data, breaks = bin_boundaries, labels = bin_means, right = false)
```

Plot the histogram

```
ggplot(data.frame(smoothed_data), aes(smoothed
```

-data)) + geom - histogram (bin width = 1, color = "black", fill = "white") + labs (x = "price", y = "frequency") + ggtitle ("Histogram of smoothed all electronics prices")

show the plot

plot (ggplot (data. frame (smoothed-data), aes (smoothed-data)))

(a) the following table would be plotted as (x,y) points, with the first column being the x values as number of mobile phones sold and the second column being the y values as money. So to use the scatter plot for how many mobile phones sold.

X →	4	1	5	7	10	2	50	25	90	36
Y →	12	5	13	19	31	7	153	72	275	110

the scatter plot for the given table can be plotted as follows.

(4,12), (1,5), (5,13), (7,19) (10,31), (2,7), (50,153), (25, 72), (90,275), (36,110).