CSA1673 - Data Warehousing and Data mining.

W-2 Assignment-02.

Name: T.Sumanth
Reg No: 192011452

Date: 02-02-2023

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 a list of all electronics prices for commonly sold items (rounded to the nearest dollar) The numbers have been 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, 19, 18, 19, 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 partioning 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) Partioning using equal frequency

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

$\frac{40}{3} = 13$ observations.

The bin boundaries for equal frequency partioning 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 $= (1+1+5+5+5+8+8+10+10+10)/13$

$= 6$

Bin 2: mean $= \frac{(10+10+10+12+14+14+14+15+15+15+15+15+15)}{13}$

$= 15.$

Bin 3: mean

$$= \frac{15+15+15+18+18+18+18+18+20+20+20+20+20+20+20+20+21+21+21+21+25+25+25+25+25+28+28+30+30+30}{13}$$

$= 24$

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 ~~other~~ either equal ~~to~~ frequency or data smoothing, 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 to 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:

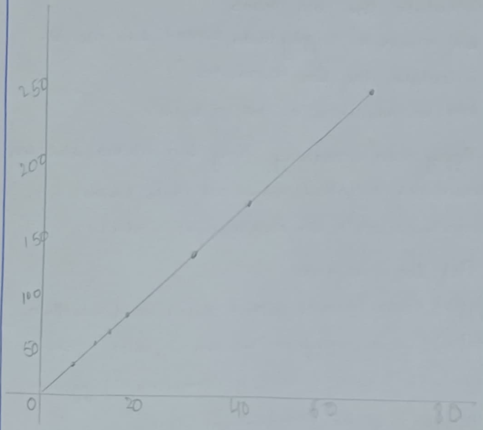
```
Load the ggplot2 library
library (ggplot2)
Create a vector of the prices data
<-c(1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14,
15, 15, 15, 15, 15, 15, 18, 19, 18, 19, 18, 19, 18, 18, 20, 20, 20, 20, 20,
20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30,
30) .
Partition the data using equal ~~binned data~~
binned <- data (data, breaks, =3, tables = 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(binwidth =1 , color = "black",
fill = "white") + labs (x = "price", y = "Frequency") + ggtitle ("
Histogram of the smoothed all electronics prices ")
Show the plot
plot (ggplot (data.frame (smoothed_data), aes(smoothed_
data)) ).
```

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. 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).



250
200
150
100
50
0
20
40
60
80