Assignment 2

Name: Danyl Fernandes
Roll : 72
ID : 2020012004
Sub: DWM

1) Smoothing by equal frequency bins
∴ 12 / 3 = 4

Bin 1 : 6, 9, 12, 13
Bin 2 : 15, 25, 50, 70
Bin 3 : 72, 92, 204, 232

i) Using bin means :
- Calculate mean for each bin & replace values with mean

$$\mu 1 = \frac{6+9+12+13}{4} = 40/4 = 10$$

$$\mu 2 = \frac{15+25+50+70}{4} = 160/4 = 40$$

$$\mu 3 = \frac{72+92+204+232}{4} = 600/4 = 150$$

∴ Bin 1 = 10, 10, 10, 10
Bin 2 = 40, 40, 40
Bin 3 = 150, 150, 150, 150

ii) Using bin boundaries

- Calaulate min & max values & replace values with closest boundaly

Bin 1 = min = 6
mx = 13

Bin 2 = min = 15
max = 70

Bin 3 = min = 72
max = 232

∴ Bin 1 = 6, 6, 13, 13
Bin 2 = 15, 15, 70, 70
Bin 3 = 72, 72, 232; 232

2) i) Use min-max noxmalization to transfom the value 45 for age onto the range [0,1]

$$v' = \frac{v - v_{minA}}{max_a - min_a}(new-max_a - new-min_a) + new-min_a$$

$min_a = 13$, $max_a = 72$
$v = 45$, $new\_max_a = 1$
$new - min_a = 0$

$$v' = \frac{45-13}{72-13}(1-0)+0$$

$$\boxed{v' = 0.542372}$$

ii) Z score normalization

$$v' = \frac{v - \mu_A}{\sigma_A}$$

$$\mu_A = \frac{525}{15} = 35$$

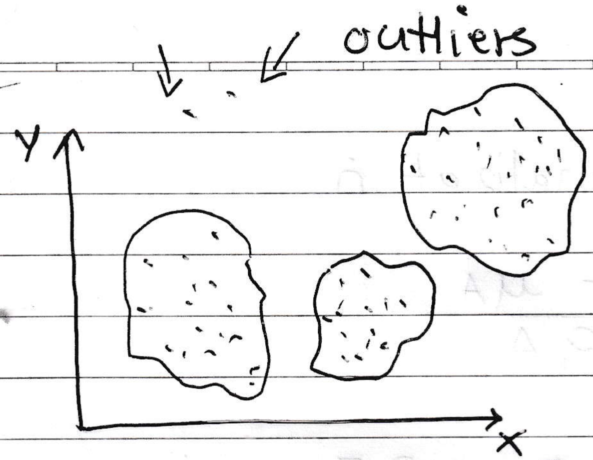
$v = 45$, $\sigma_A = 20.64$

$$v' = \frac{45-35}{20.64} = 0.484496$$

$$\boxed{v' = 0.484496}$$

3) i) Method of Clustering

- The extreme values that drastrically deviate from the dataset observations are called outliers.
- Analysing these helps in identifying anomalous observations
- In Clustering, we group similar values in clusters called "clusters".
- These methods may also be used to detect unusual activities or fraudaulent transactions



ii) Regression

- Regnession is a data mining technique used to predict a range of numeric values given in a dataset
- Regression is used also for smoothing

Typos

a) Linear : Linear regression finds the best line to fit two variables or attributes so that one variable can be used to predict the other

Denoted as $y = \alpha + \beta_x$

b) Multiple linear : Used when more than two variables are in use, the data is to be fit to multidimensional surface.

$$\beta = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

Where $\bar{x}$ & $\bar{y}$ are mean of variables x & y respectively