Data Transformation and Discretization

Hazim Fitri

2024-12-18

Contents

Data Transformation and Discretization

Normalization

Min-Max Normalization

```
dataAP3 <- read.csv('dataAP3.csv', header = T)
head(dataAP3)
```

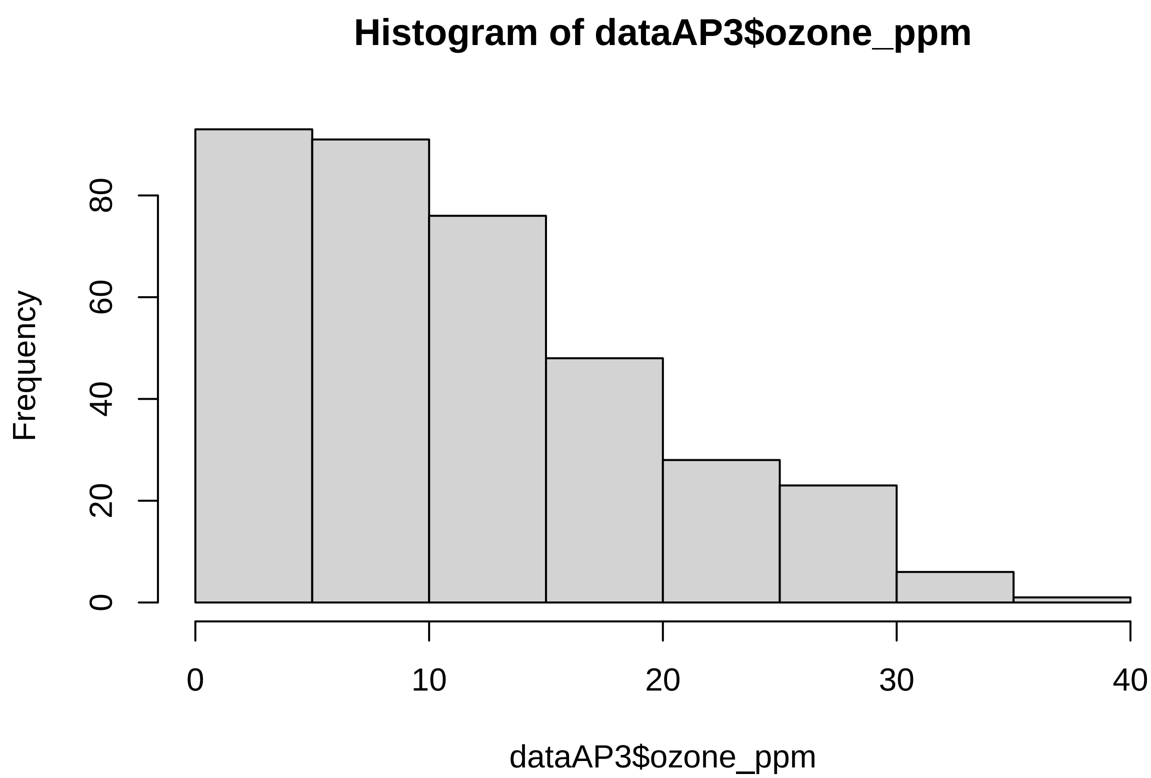
```
##   X Month Day_of_month Day_of_week ozone_ppm pressure_height.hPA Wind_speed.mph
## 1 1     1            1           4      3.01                5480              8
## 2 2     1            2           5      3.20                5660              6
## 3 3     1            3           6      2.70                5710              4
## 4 4     1            4           7      5.18                5700              3
## 5 5     1            5           1      5.34                5760              3
## 6 6     1            6           2      5.77                5720              4
##   Temperature_Celcius Inversion_base_height.IBH Pressure_gradient.Psi.ft
## 1                  30                      5000                      -15
## 2                  38                      1601                      -14
## 3                  40                      2693                      -25
## 4                  45                       590                      -24
## 5                  54                      1450                       25
## 6                  35                      1568                       15
##   Inversion_temperature.ivC Visibility_pAerosol
## 1                     30.56                 200
## 2                     46.94                 300
## 3                     47.66                 250
## 4                     55.04                 100
## 5                     57.02                  60
## 6                     53.78                  60
```

```
dataAP3 = dataAP3[,-c(1)]
head(dataAP3)
```

```
##   Month Day_of_month Day_of_week ozone_ppm pressure_height.hPA Wind_speed.mph
## 1     1            1           4      3.01                5480              8
## 2     1            2           5      3.20                5660              6
## 3     1            3           6      2.70                5710              4
## 4     1            4           7      5.18                5700              3
## 5     1            5           1      5.34                5760              3
## 6     1            6           2      5.77                5720              4
##   Temperature_Celcius Inversion_base_height.IBH Pressure_gradient.Psi.ft
## 1                  30                      5000                      -15
## 2                  38                      1601                      -14
## 3                  40                      2693                      -25
## 4                  45                       590                      -24
## 5                  54                      1450                       25
## 6                  35                      1568                       15
##   Inversion_temperature.ivC Visibility_pAerosol
## 1                     30.56                 200
## 2                     46.94                 300
```

```
## 3                  47.66                250
## 4                  55.04                100
## 5                  57.02                 60
## 6                  53.78                 60
```

```
hist(dataAP3$ozone_ppm)
```

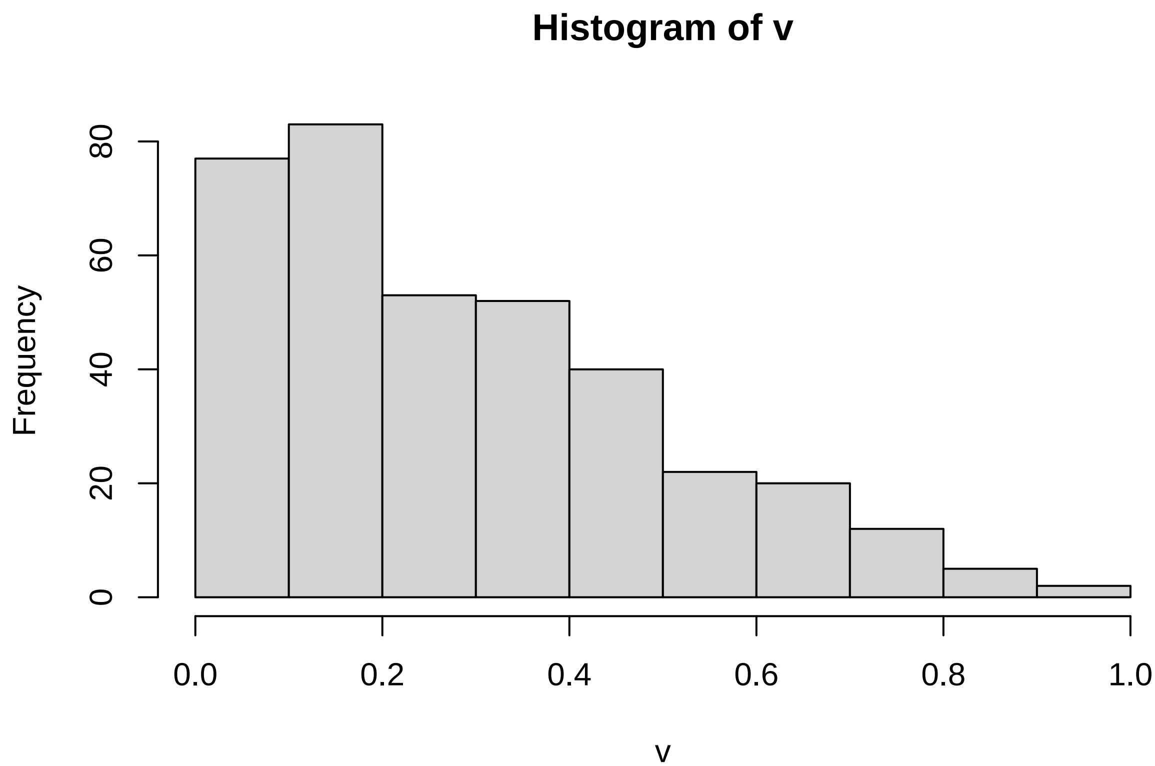


$$V = \frac{\left[X - \min(X)\right] \times \left[baru\_\max(X) - baru\_\min(X)\right]}{\max(X) - \min(X)} + baru\_\min(X)$$

```
min_ozone = min(dataAP3$ozone_ppm)
max_ozone = max(dataAP3$ozone_ppm)
v = ((dataAP3$ozone_ppm - min_ozone) * (1 - 0)) / (max_ozone-min_ozone)
head(v)
```

```
## [1] 0.06146001 0.06655931 0.05314010 0.11969941 0.12399356 0.13553408
```

```
hist(v)
```



Z-score Normalization

```
mean_hpa = mean(dataAP3$pressure_height.hPA)
sd_hpa = sd(dataAP3$pressure_height.hPA)
z_score_hpa = (dataAP3$pressure_height.hPA - mean_hpa) / sd_hpa
head(z_score_hpa)
```

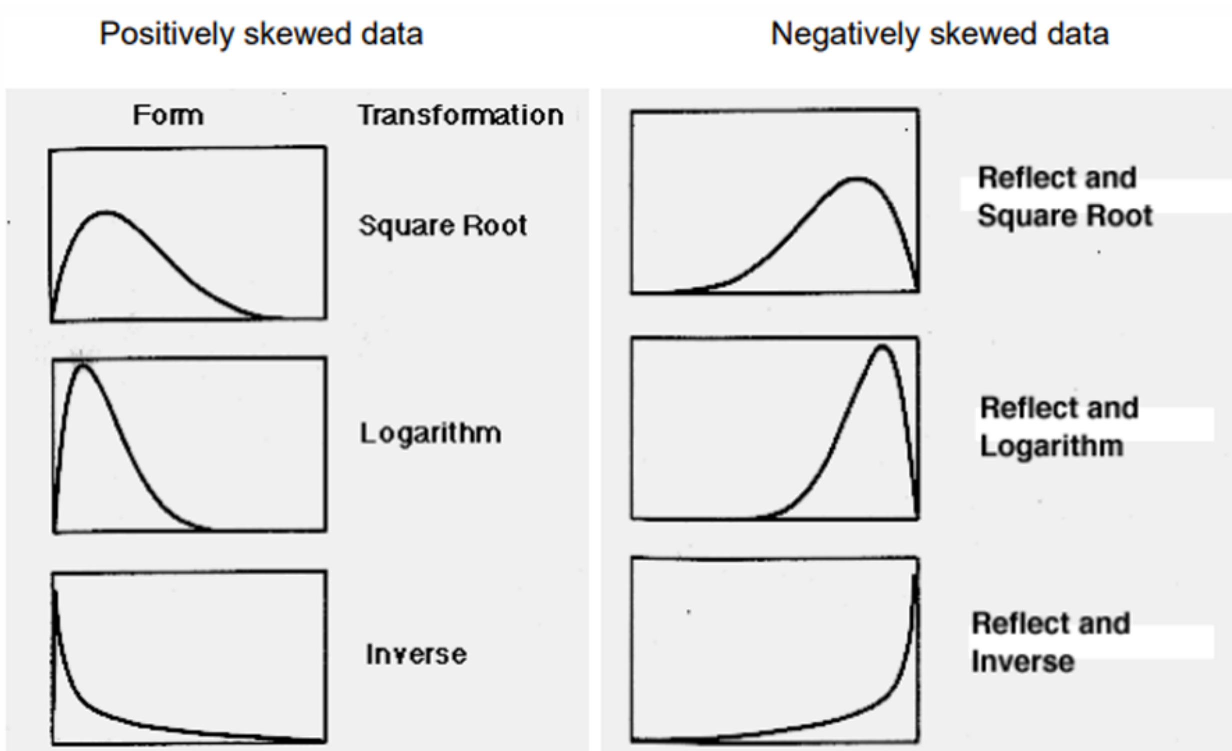
```
## [1] -2.58185122 -0.87803114 -0.40474779 -0.49940446  0.06853557 -0.31009112
```

Decimal Scaling

```
pHnew = dataAP3$pressure_height.hPA/1000
head(pHnew)
```

```
## [1] 5.48 5.66 5.71 5.70 5.76 5.72
```

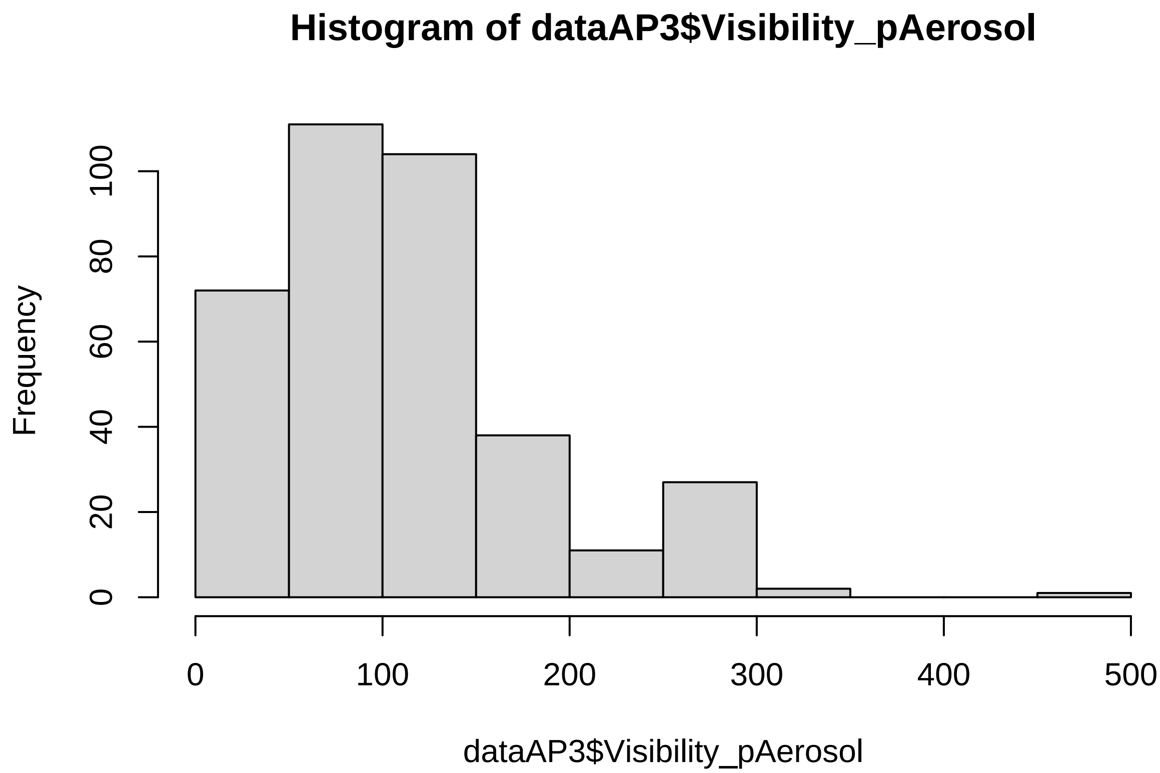
Normaling Data Distribution



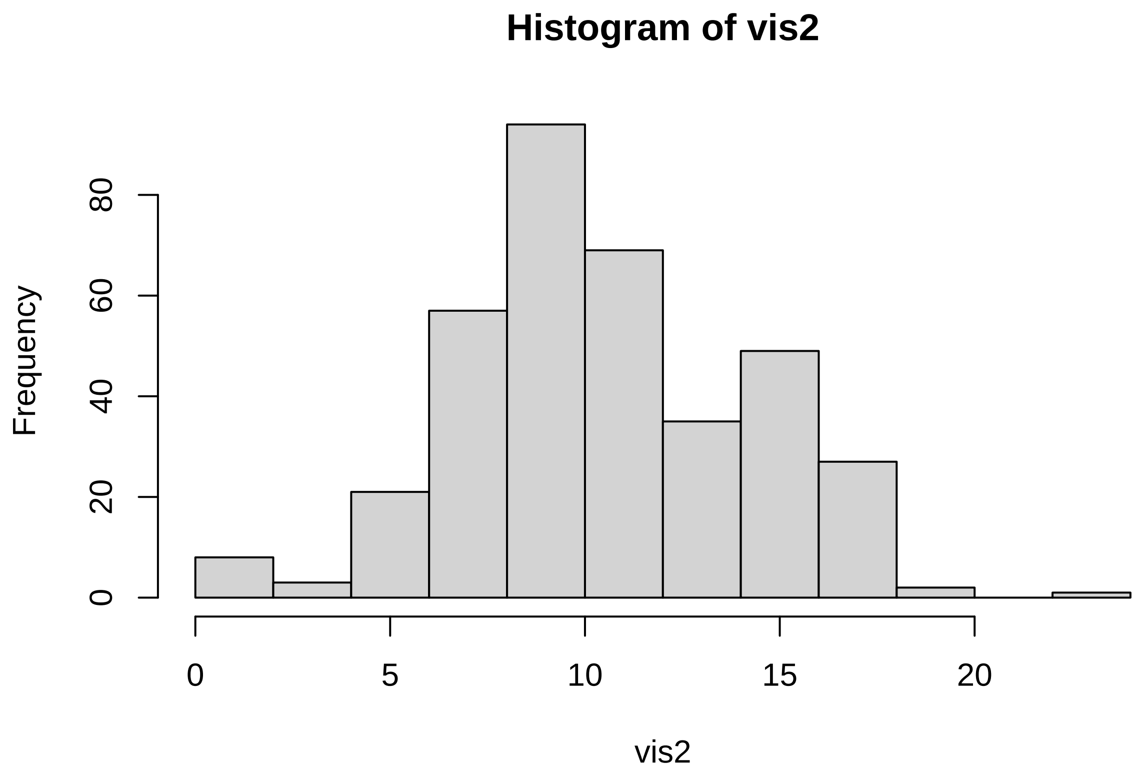

```r
dataAP3$Visibility_pAerosol
```

```
##   [1] 200 300 250 100  60  60 100 250 120 120 120 150  40 200 250 200 200 150
##  [19]  10 140 250 200 150 140  50   0  70 150 150 120  40 120   6  30 100 200
##  [37]  60 350 250 350 300 300 300 200 100 250 200 200  40   2 300 300 300 300
##  [55] 300 150 150  80  40  40  80 300 200 500 140 140 140 100 140 200 120 300
##  [73] 300 150   2  50  70  17 140 140 300 200 250  80  60 100 150 150 200 100
##  [91] 300 120 100 200 200 200 300 300 250 120 140 200 140  80 300 100 300 200
## [109] 120 100 120  60 120 100 100  27  40 140 150 100 100 120 150 100 120  80
## [127] 120 140 120  70  80  70  40  20  17  40  50  50  70  80 120 120 100 120
## [145] 120 200 120  40  70 100 120 100 120  70  80 100 100 120 120 120 150 140
## [163] 140 140 140  60  30  17  80  60 100 120 150 120 140 140 120 120  80 140
## [181] 140 150 120 120 140 100  50  40 100  80 100  60  50  70  80  80  80  90
## [199] 120 120 100  60  40  50  40  70  80  80  80  80  80 100 120 150 200 150
## [217] 150 150 150 100 100 100  30  80  70  60 150 200 200 200 250 300  70 300
## [235] 150 300  30 100 100  17  20   4  70  30  70  60  40  50  70 140 100 120
## [253] 100  70 150  50  70  40  70 120 140 140 100  50  70  40  40 100 120 120
## [271] 140 120  70 150 200 200 200  70  40  50  17  80 250 200   2  20   7  30
## [289]  50  70  17  80  50  60  60  80  50  50  40  40 300 200 150 100 100  60
## [307] 150 150 200 300 120  30 100  50  20 200 120 300 200  70 140 150 200   4
## [325]  40  30  30   2   0  30  60 150 100 250 150 200 200 200  80  60 300 200
## [343] 300  50  40  70 150 150  70 200 120 150 150  60  70 150 300 100  70  40
## [361] 140 200  70  40 100  70
```

```r
hist(dataAP3$Visibility_pAerosol)
```




```
vis2 = sqrt(dataAP3$Visibility_pAerosol)
hist(vis2)
```



Assessing Normality

Histogram & Boxplot

Normal Quantile Plot (Q-Q Plot)

Goodnes-of-fit test

Kolmogorov-Smirnov

Shapiro-Wilk

Anderson-Darling

Discretization

Unsupervised Learning

This method need the knowledge of the industry and can be made manually for example like the financial class (B40, M40, T20)

```r
library(infotheo)
data("USArrests")
attach(USArrests)
USArrests
```

2 category

```
##                Murder Assault UrbanPop Rape
## Alabama          13.2     236       58 21.2
## Alaska           10.0     263       48 44.5
## Arizona           8.1     294       80 31.0
## Arkansas          8.8     190       50 19.5
## California        9.0     276       91 40.6
## Colorado          7.9     204       78 38.7
## Connecticut       3.3     110       77 11.1
## Delaware          5.9     238       72 15.8
## Florida          15.4     335       80 31.9
## Georgia          17.4     211       60 25.8
## Hawaii            5.3      46       83 20.2
## Idaho             2.6     120       54 14.2
## Illinois         10.4     249       83 24.0
## Indiana           7.2     113       65 21.0
## Iowa              2.2      56       57 11.3
## Kansas            6.0     115       66 18.0
## Kentucky          9.7     109       52 16.3
## Louisiana        15.4     249       66 22.2
## Maine             2.1      83       51  7.8
## Maryland         11.3     300       67 27.8
## Massachusetts     4.4     149       85 16.3
## Michigan         12.1     255       74 35.1
## Minnesota         2.7      72       66 14.9
## Mississippi      16.1     259       44 17.1
## Missouri          9.0     178       70 28.2
## Montana           6.0     109       53 16.4
## Nebraska          4.3     102       62 16.5
## Nevada           12.2     252       81 46.0
## New Hampshire     2.1      57       56  9.5
## New Jersey        7.4     159       89 18.8
## New Mexico       11.4     285       70 32.1
## New York         11.1     254       86 26.1
## North Carolina   13.0     337       45 16.1
## North Dakota      0.8      45       44  7.3
## Ohio              7.3     120       75 21.4
## Oklahoma          6.6     151       68 20.0
## Oregon            4.9     159       67 29.3
## Pennsylvania      6.3     106       72 14.9
## Rhode Island      3.4     174       87  8.3
## South Carolina   14.4     279       48 22.5
## South Dakota      3.8      86       45 12.8
## Tennessee        13.2     188       59 26.9
## Texas            12.7     201       80 25.5
## Utah              3.2     120       80 22.9
```

```
## Vermont           2.2      48       32 11.2
## Virginia          8.5     156       63 20.7
## Washington        4.0     145       73 26.2
## West Virginia     5.7      81       39  9.3
## Wisconsin         2.6      53       66 10.8
## Wyoming           6.8     161       60 15.6
```

```
cutoff = 10 # Need domain explanation
status_m = ifelse(Murder<10,'Low Risk','High Risk')
head(status_m)
```

```
## [1] "High Risk" "High Risk" "Low Risk"  "Low Risk"  "Low Risk"  "Low Risk"
```

```
library(car)
```

More than 2 category

```
## Warning: package 'car' was built under R version 4.4.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.4.2
```

```
status_den = Recode(UrbanPop, "0:50 = 'Low Density';
                               51:70 = 'Moderate Density';
                               else = 'High Density'")
head(status_den)
```

```
## [1] "Moderate Density" "Low Density"      "High Density"     "Low Density"
## [5] "High Density"     "High Density"
```

```
assault_status = discretize(Assault, 'equalwidth', 4)
unique(assault_status)
```

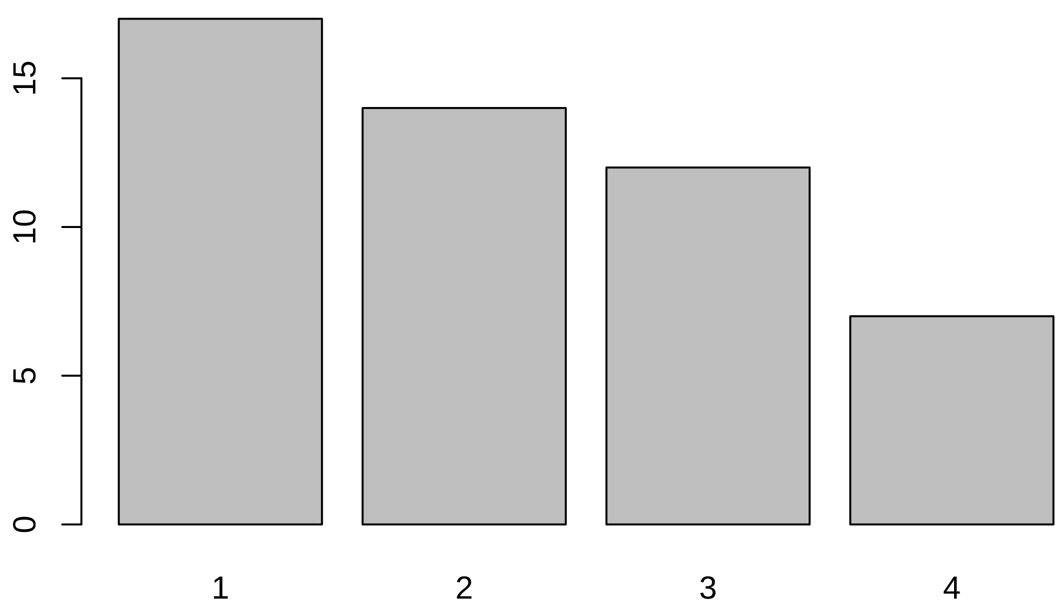
Equal-width

```
##   X
## 1 3
## 3 4
## 4 2
## 7 1
```

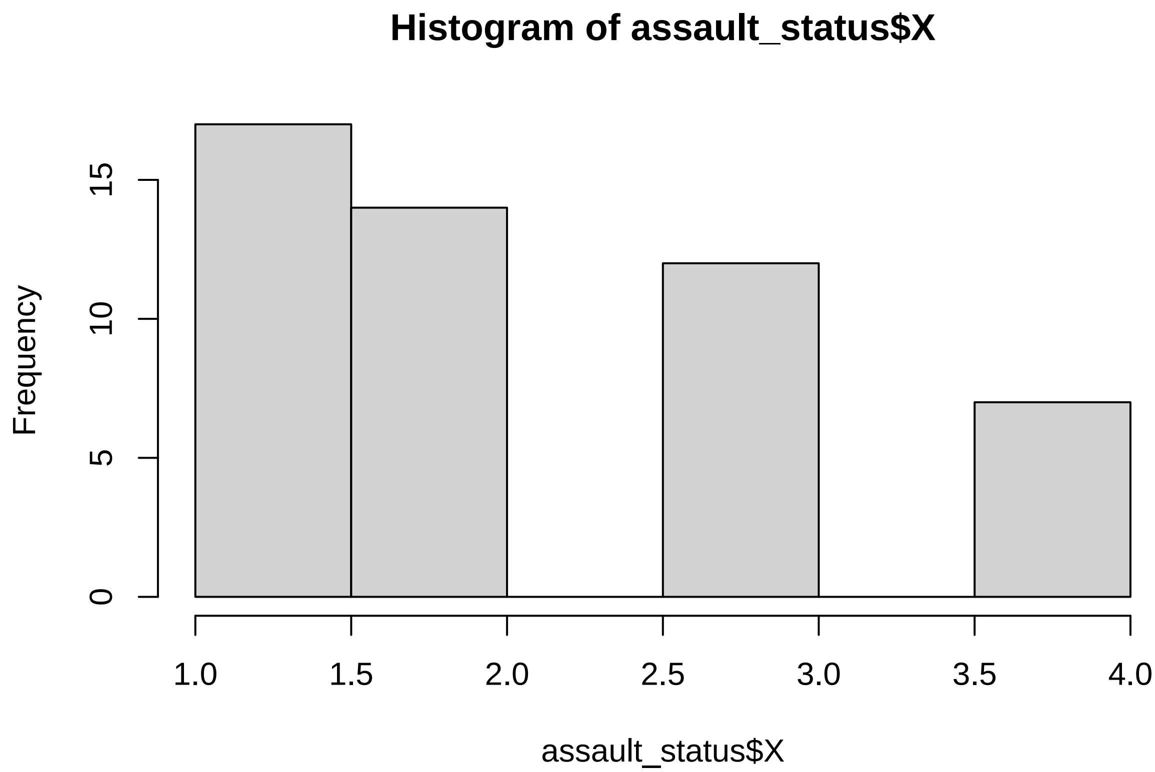
```
head(assault_status)
```

```
##   X
## 1 3
## 2 3
## 3 4
## 4 2
## 5 4
## 6 3
```

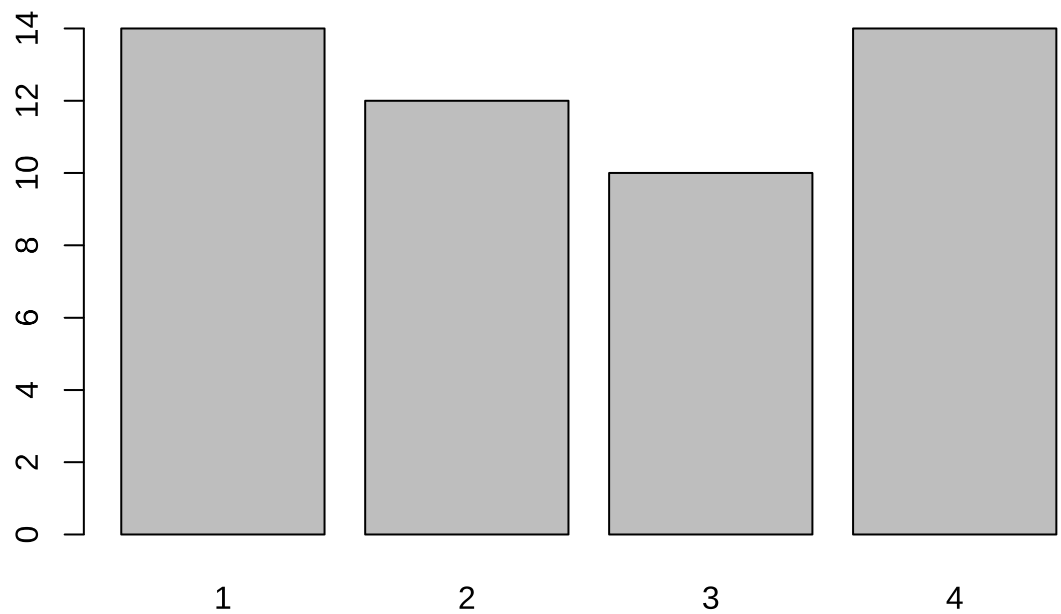
```
barplot(table(assault_status$X))
```




```
hist(assault_status$X)
```




```
barplot(table(discretize(Assault, 'equalfreq', 4)$X))
```



Equal Frequency

Supervised Learning

```
library(discretization)
data(iris)
iris2 = chi2(iris, alp=0.05)$Disc.data
head(iris2)
```

Chi2

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1            1           3            1           1  setosa
## 2            1           2            1           1  setosa
## 3            1           2            1           1  setosa
## 4            1           2            1           1  setosa
## 5            1           3            1           1  setosa
## 6            1           3            1           1  setosa
```

Attribute formation

Smoothing