Cluster Analysis

Enrique J. De La Hoz D.

UTB - Data Science

Cluster Analysis

A form of exploratory data analysis ***(EDA)*** where ***observations*** are divided into meaningful groups that share common characteristics ***(features)***.

Clustering example

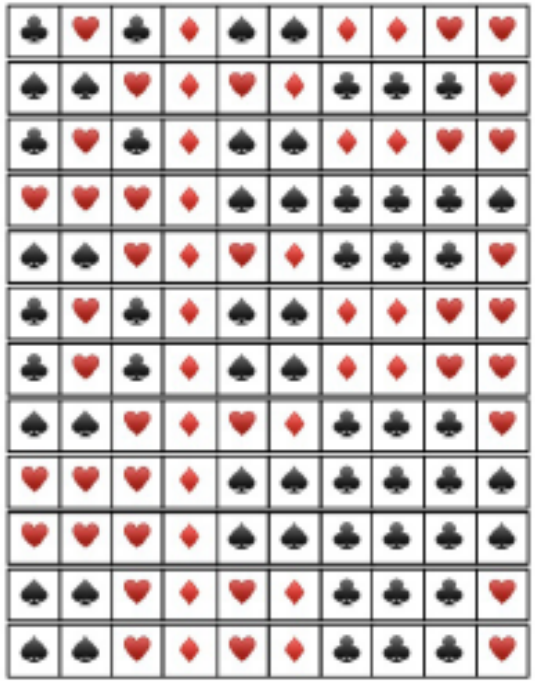


Figure 1:

Clustering example

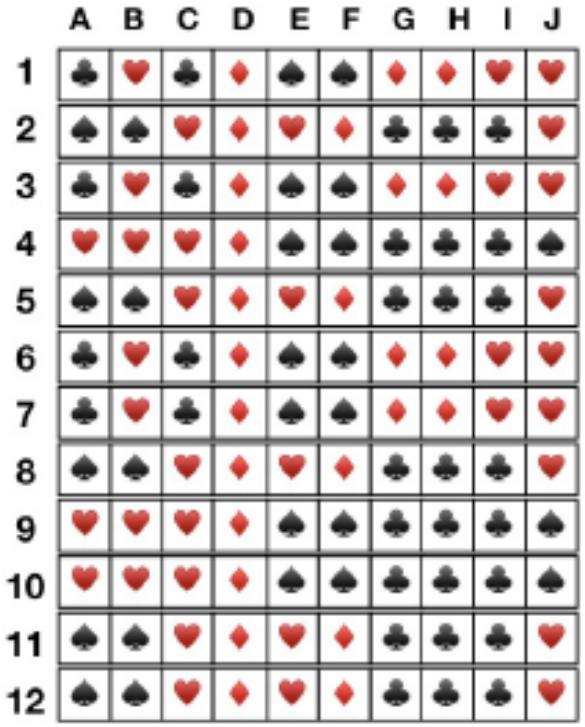


Figure 2:

Clustering example

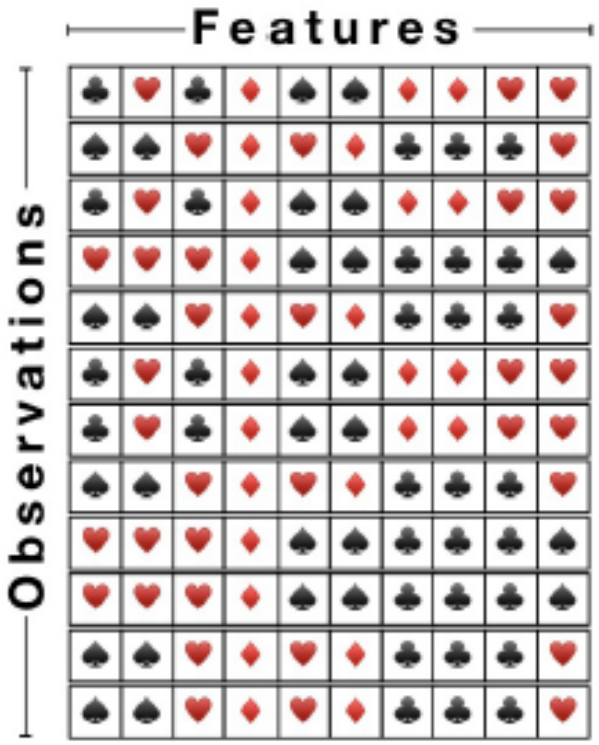


Figure 3:

Clustering example

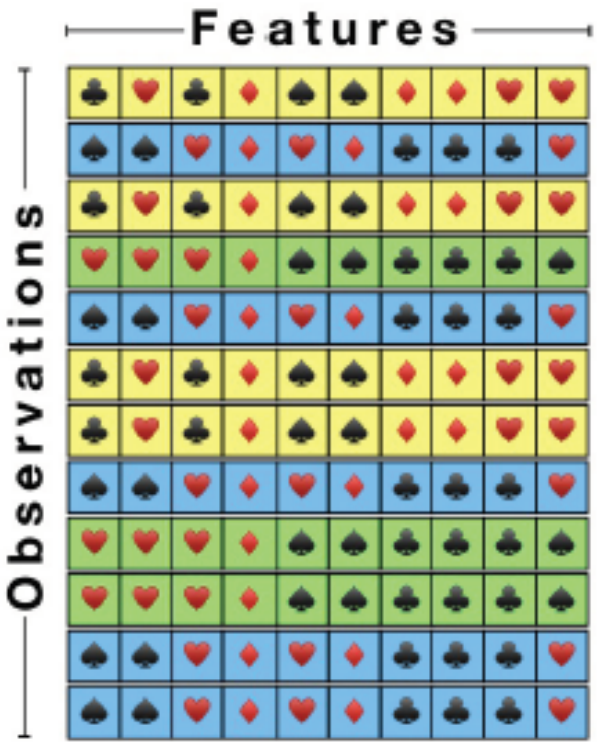


Figure 4:

Clustering example

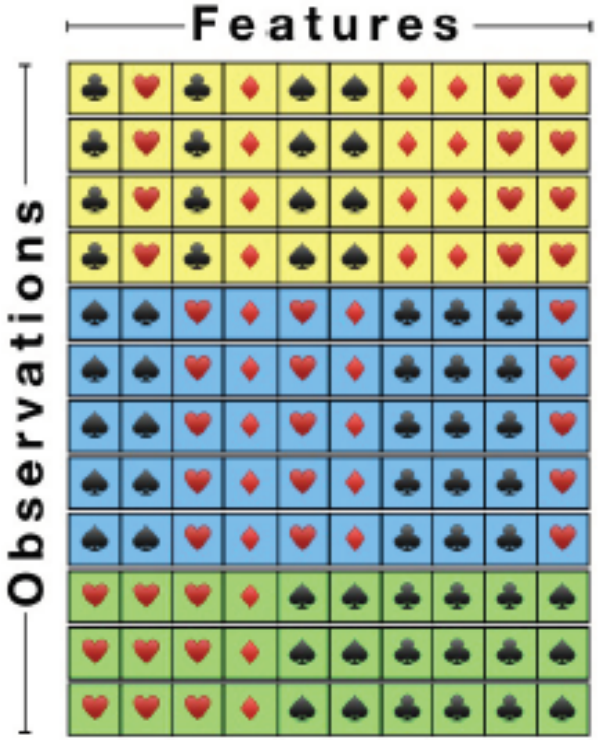


Figure 5:

The flow of cluster analysis

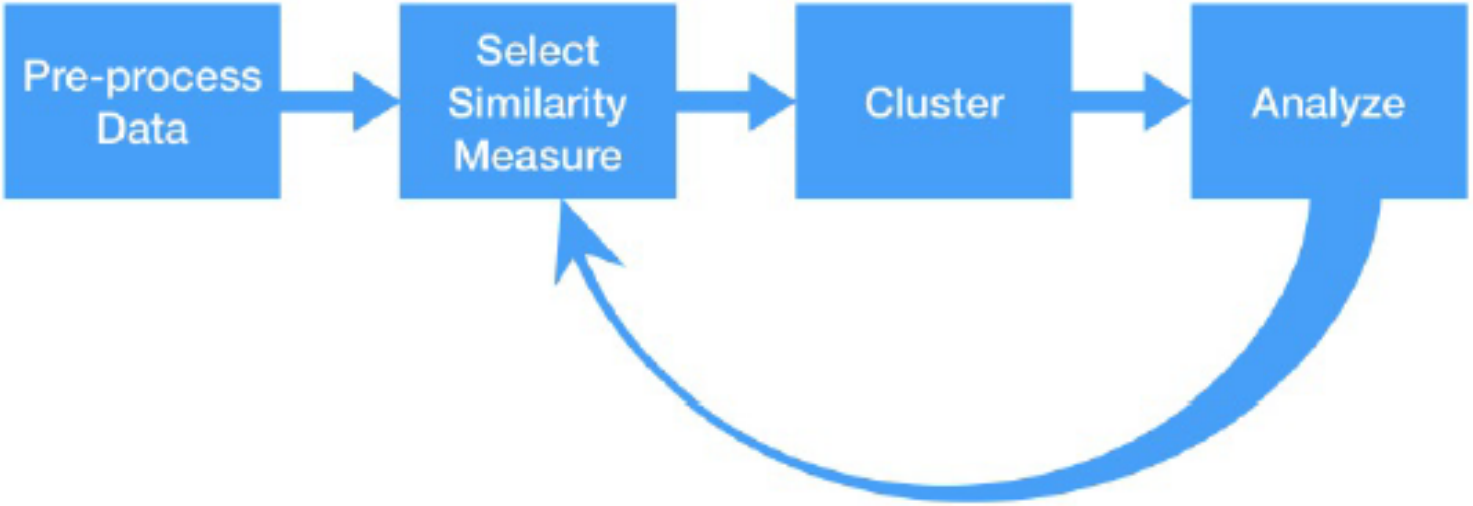


Figure 6:

Distance between two observations

- Distance vs Similarity
  - $ DISTANCE = 1 - SIMILARITY$

Distanace between two players

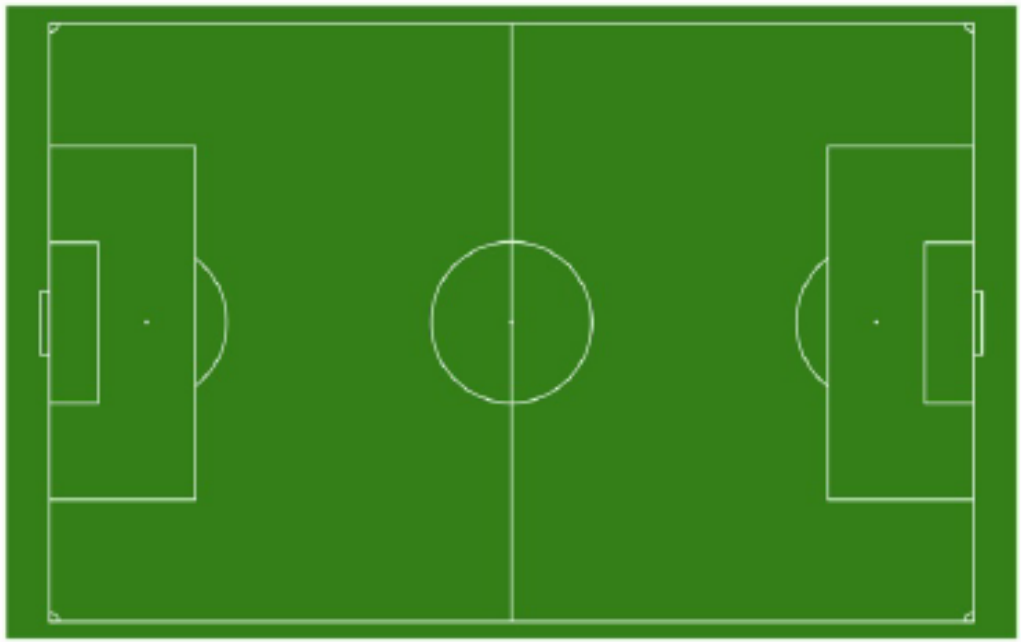


Figure 7:

Distanace between two players

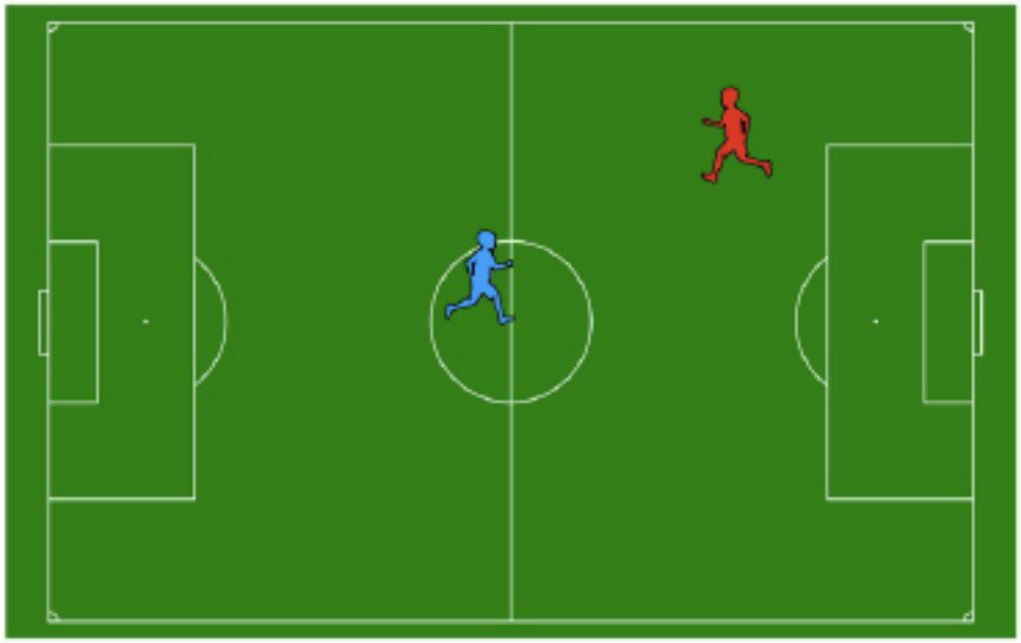


Figure 8:

Distanace between two players

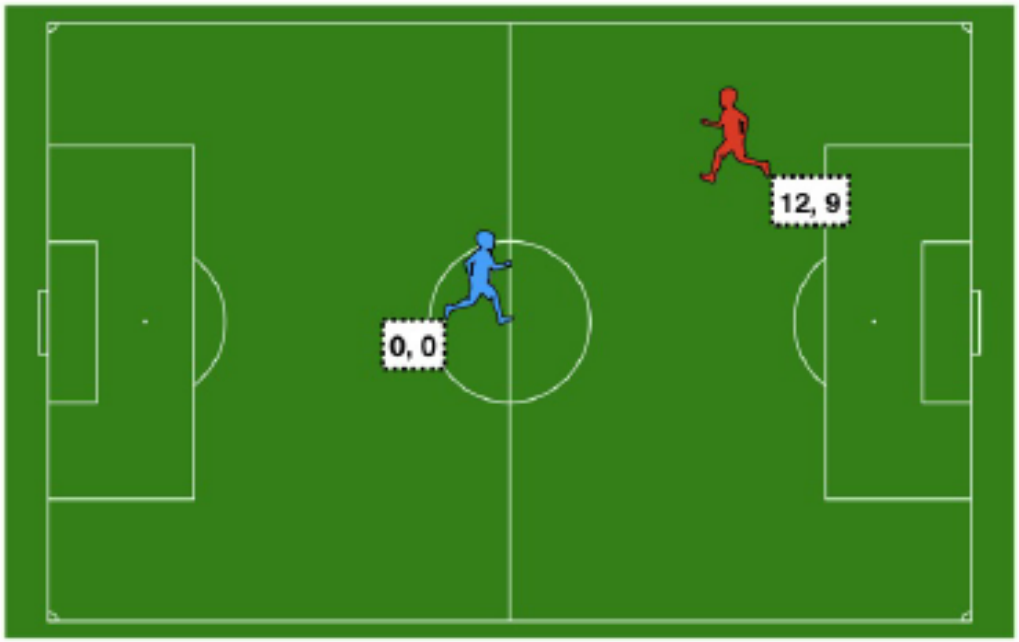


Figure 9:

Distanace between two players

	X	Y
Blue	0	0
Red	12	9

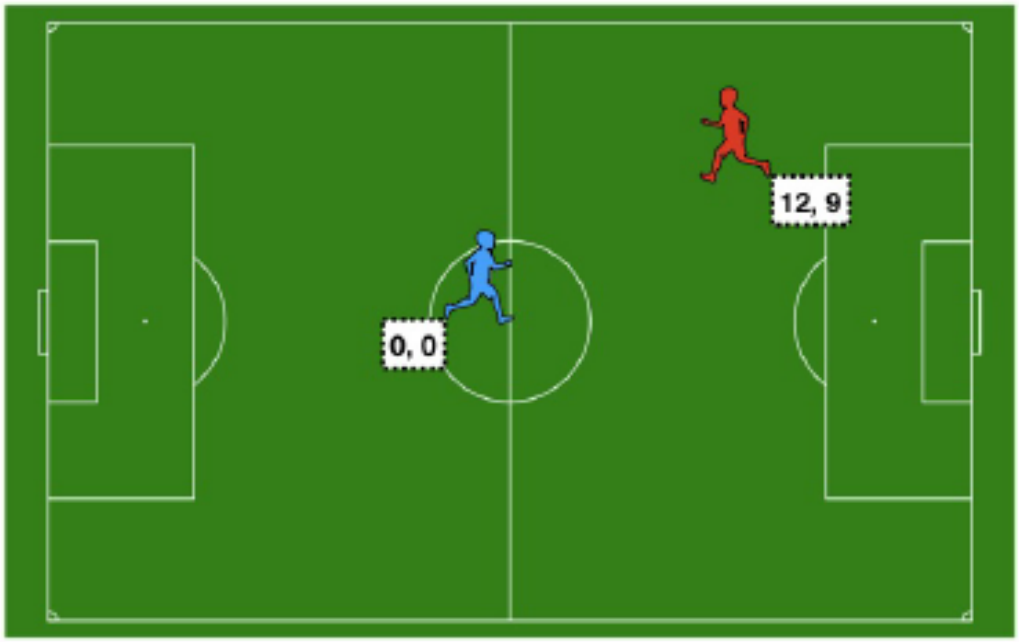


Figure 10:

Distanace between two players

	X	Y
Blue	0	0
Red	12	9

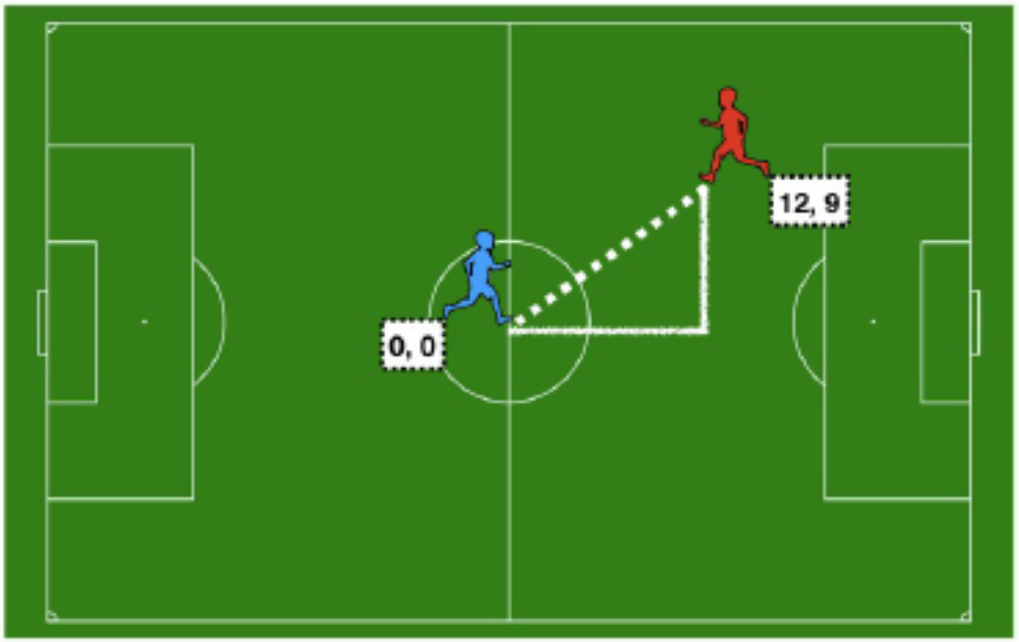


Figure 11:

Distanace between two players

	X	Y
Blue	0	0
Red	12	9

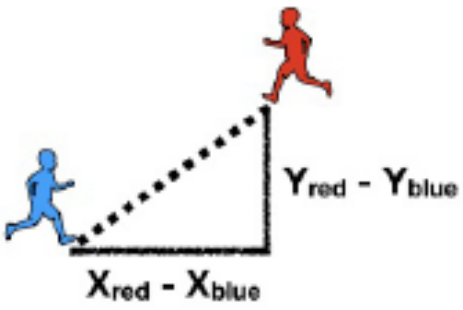


Figure 12:

Distanace between two players

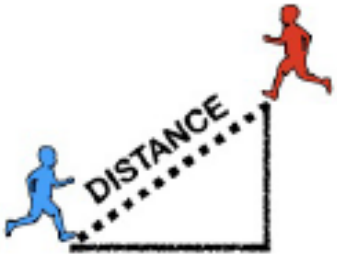
	X	Y
Blue	0	0
Red	12	9

$$= \sqrt{(X_{red} - X_{blue})^2 + (Y_{red} - Y_{blue})^2}$$

Figure 13:

Distanace between two players

	X	Y
Blue	0	0
Red	12	9

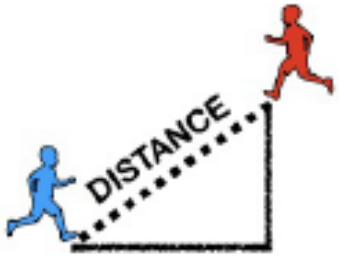


$$= \sqrt{(12 - 0)^2 + (9 - 0)^2}$$

Figure 14:

Distanace between two players

	X	Y
Blue	0	0
Red	12	9



= 15

Figure 15:

dist() Function

```r
two_players
```

```
##      X  Y
## BLUE 0  0
## RED  9 12
```

```r
dist(two_players, method = "euclidean")
```

```
##     BLUE
## RED   15
```

More than 2 Observations

```r
two_players_3
```

```
##        X  Y
## BLUE   0  0
## RED    9 12
## GREEN -2 19
```

```r
dist(two_players_3, method = "euclidean")
```

```
##           BLUE      RED
## RED   15.00000
## GREEN 19.10497 13.03840
```

Scaling the Features

Observation	Height (feet)	Weight (lbs)
1	6.0	200
2	6.0	202
3	8.0	200
...	...	...
...	...	...

Figure 16:

Scaling the Features

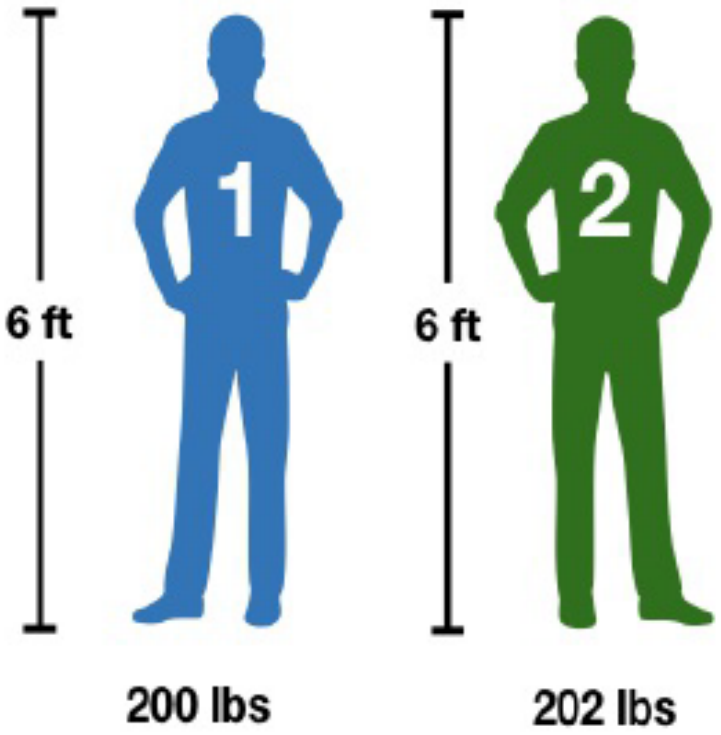


Figure 17:

Scaling the Features

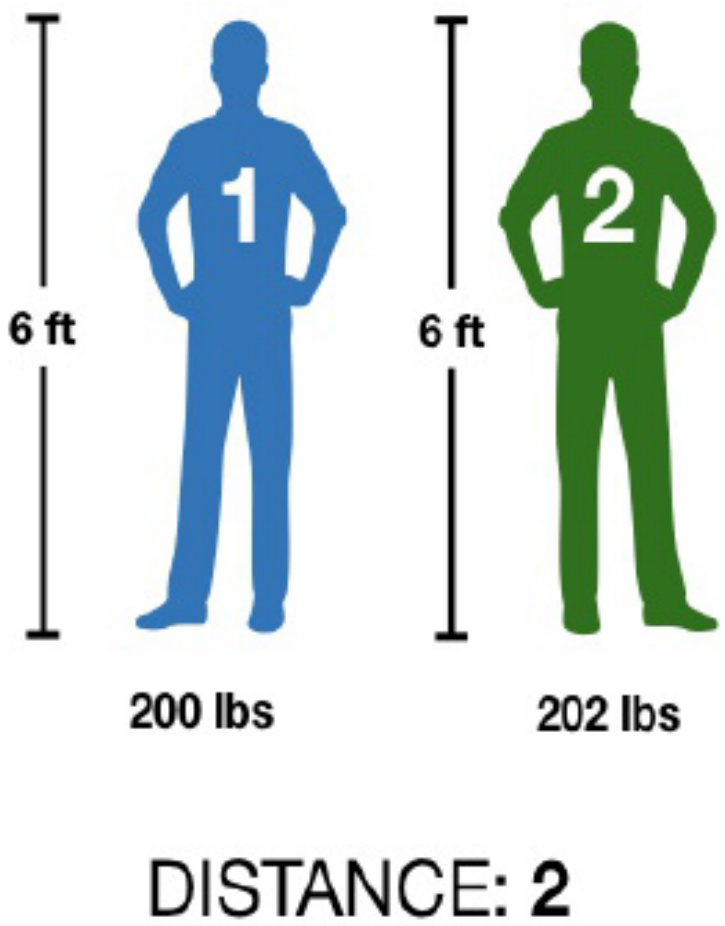


Figure 18:

Scaling the Features

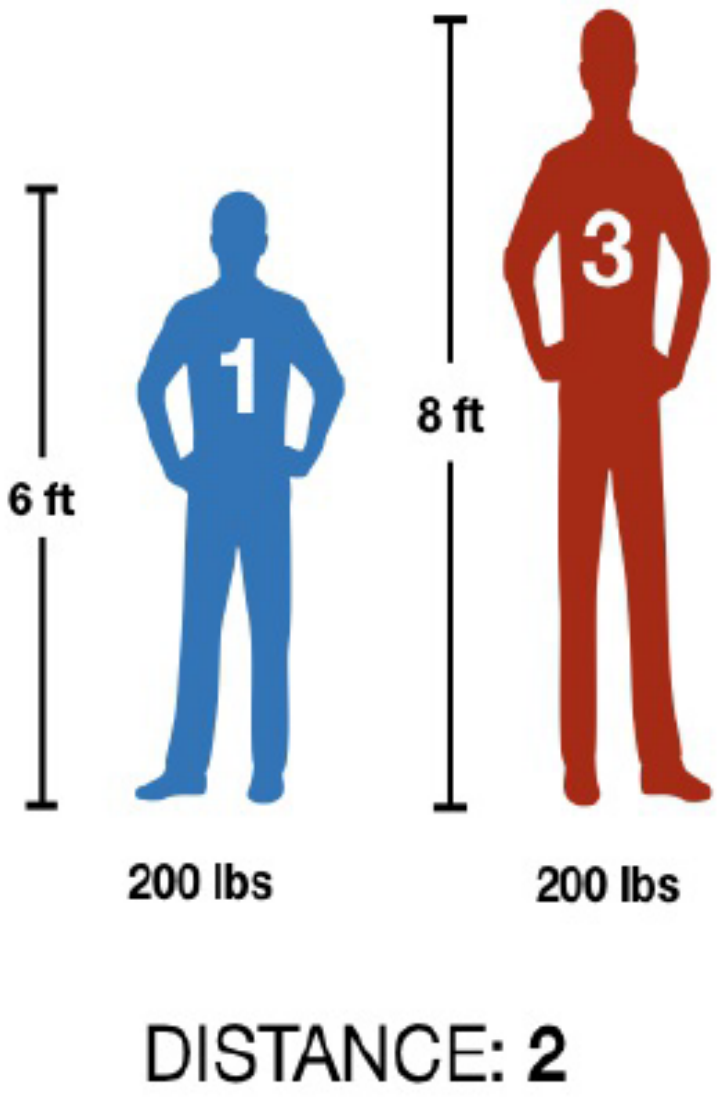


Figure 19:

Scaling the Features

Figure 20:

How to do it?

$$height_{scaled} = \frac{height - mean(height)}{sd(height)}$$

Scaling Distance

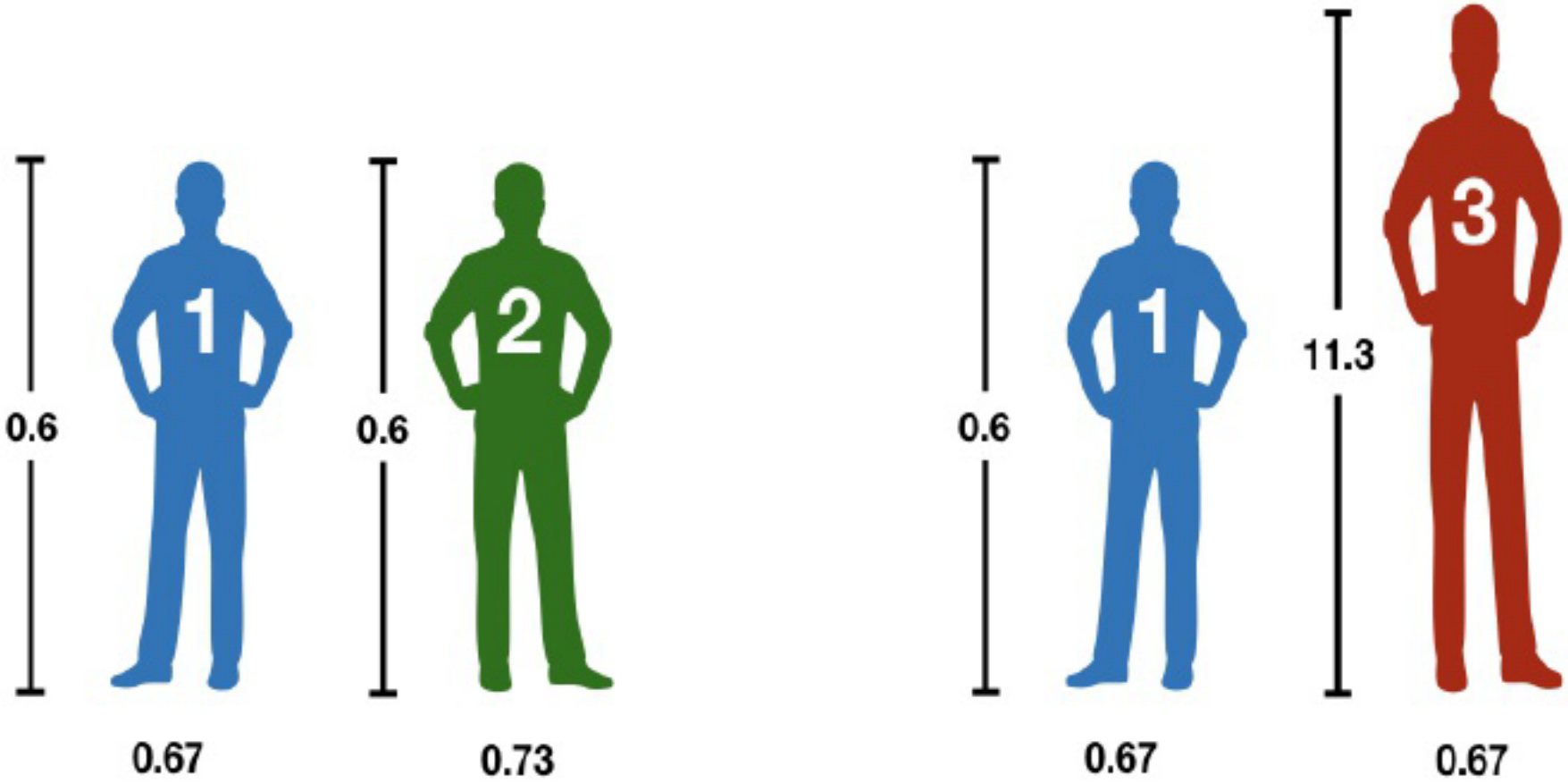


Figure 21:

Scaling Distance

Figure 22:

Using R

```
##   Height Weight
## 1      6    200
## 2      6    202
## 3      8    200
```

```
##          Height     Weight
## [1,] -0.5773503 -0.5773503
## [2,] -0.5773503  1.1547005
## [3,]  1.1547005 -0.5773503
## attr(,"scaled:center")
##     Height     Weight
##   6.666667 200.666667
## attr(,"scaled:scale")
##   Height   Weight
## 1.154701 1.154701
```

Distance for categorical Data

	wine	beer	whiskey	vodka
1	**TRUE**	**TRUE**	FALSE	FALSE
2	FALSE	**TRUE**	**TRUE**	**TRUE**
...	...	...	...	...

Figure 23:

Jaccard Index

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

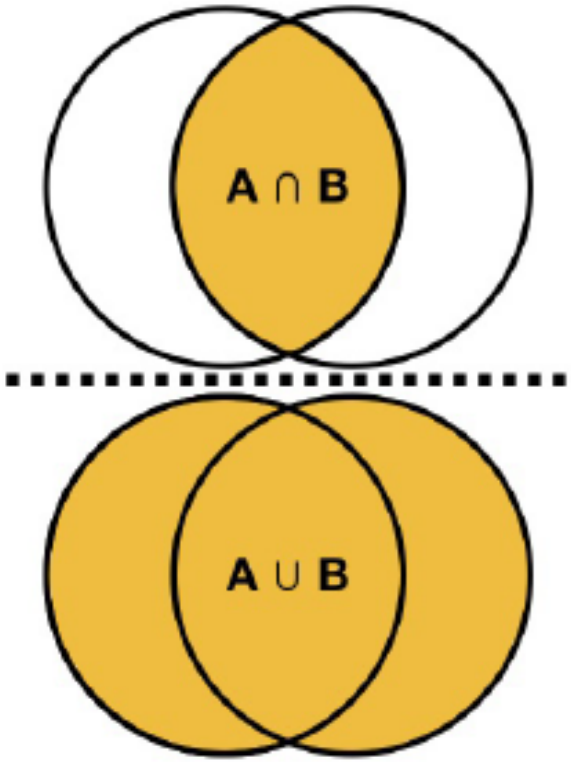


Figure 24:

Calculatind Jaccard Distance

	wine	beer	whiskey	vodka
1	**TRUE**	**TRUE**	FALSE	FALSE
2	FALSE	**TRUE**	**TRUE**	**TRUE**
...	...	...	...	...

Figure 25:

$$J(1,2) = \frac{1 \cap 2}{1 \cup 2} = \frac{1}{4} = 0.25$$

$$Distance(1,2) = 1 - J(1,2) = 0.75$$

Jaccard Distance in R

More than two categories

	color	sport
1	red	soccer
2	green	hockey
3	blue	hockey
4	blue	soccer
...	...	...

	colorblue	colorgreen	colorred	sporthockey	sportsocce
1	0	0	1	0	1
2	0	1	0	1	0
3	1	0	0	1	0
4	1	0	0	0	1
...	...	...	...	...	...

Figure 26:

Dummification in R

```
colorx
```

```
##   color  sport
## 1   red soccer
## 2 green hockey
## 3  blue hockey
## 4  blue soccer
```

```
library(dummies)
dum<- dummy.data.frame(colorx)
dum
```

```
##   colorblue colorgreen colorred sporthockey sportsoccer
## 1         0          0        1            0           1
## 2         0          1        0            1           0
## 3         1          0        0            1           0
## 4         1          0        0            0           1
```

Genera

```r
dist(dum, method = "binary")
```

```
##           1         2         3
## 2 1.0000000
## 3 1.0000000 0.6666667
## 4 0.6666667 1.0000000 0.6666667
```