AGGREGATION

SYRACUSE UNIVERSITY
School of Information Studies

DATA TRANSFORMATION

Aggregation

Attribute transformation

Dimensionality reduction and feature selection (covered in future weeks)

AGGREGATION

Combining two or more rows or columns into a single row or column

Purpose:

Data reduction

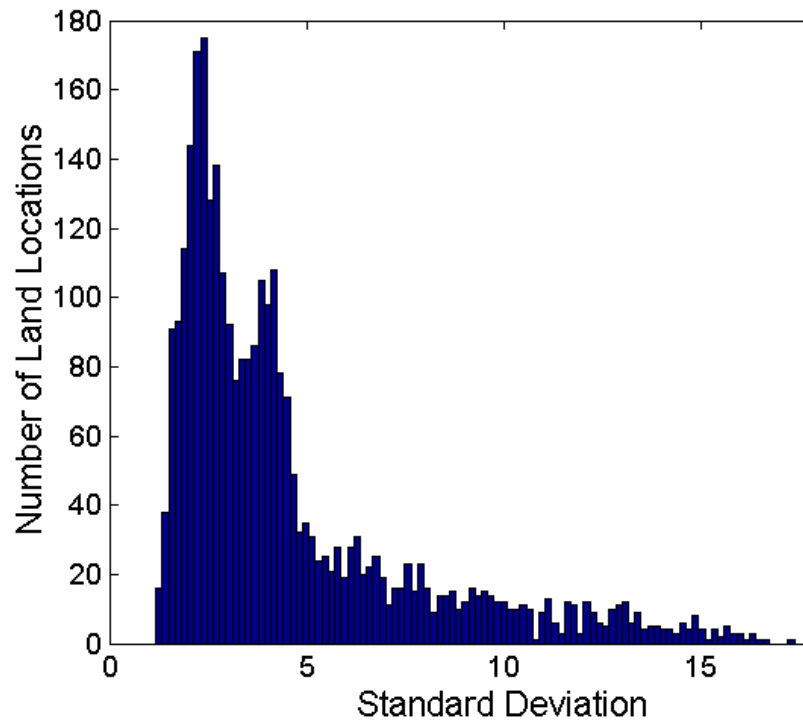All test scores merged into one total score

Change of scale

Cities aggregated into regions, states, countries, etc.
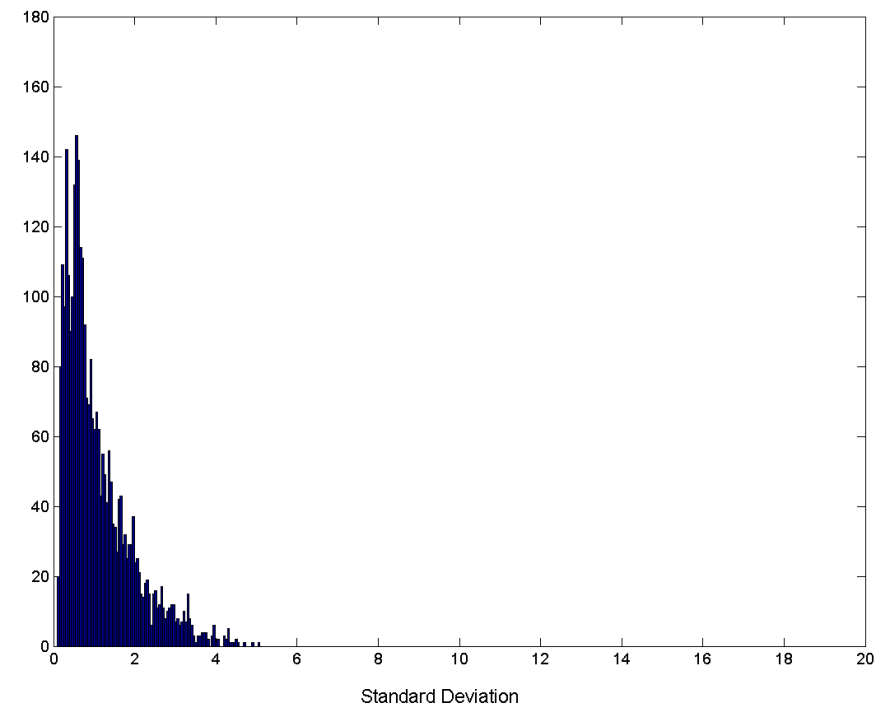
More "stable" data

Aggregated data tends to have less noise

AGGREGATION

Variation of Precipitation in Australia

Standard Deviation of Average
Monthly Precipitation

Standard Deviation of
Average Yearly Precipitation

AGGREGATE ROWS

```
> sales <- read.csv("/Users/byu/Desktop/data/sales.csv")
> salesByRegion <- aggregate(cbind(Mon,Tue,Wed,Thu,Fri,Sat,Sun
),by=list(Group.region=Region),FUN=sum)
> salesByRegion
  Group.region Mon Tue Wed Thu Fri Sat Sun
1            A 150 190 186 188 170 101 102
2            B 267 342 306 304 341 327 305
3            C 298 366 381 400 407 402 495
4            D 390 395 385 365 376 342 375
```

AGGREGATE COLUMNS

```
> InWeekend <- rowSums(sales[,c("Sat","Sun")])
> salesNew <- data.frame(sales,InWeekend)
> salesNew
  Region Store Mon Tue Wed Thu Fri Sat Sun InWeekend
1      A    S1 100 125 119 110 116  59  57       116
2      A    S2  50  65  67  78  54  42  45        87
3      B    S3  78  89  81  92  97  82  85       167
4      B    S4  90 120 105  97 107 119 120       239
5      B    S5  99 133 120 115 137 126 100       226
6      C    S6 130 190 211 200 187 187 195       382
7      C    S7 168 176 170 200 220 215 300       515
8      D    S8 200 210 190 195 187 170 175       345
9      D    S9 190 185 195 170 189 172 200       372
```

AGGREGATE ROWS AND COLUMNS

```
> salesInWeekend <-aggregate(InWeekend, by=list(Region), FUN=m
ean)
> salesInWeekend
  Group.1        x
1       A 101.5000
2       B 210.6667
3       C 448.5000
4       D 358.5000
```

SYRACUSE UNIVERSITY
School of Information Studies