DATA SET TYPES

SYRACUSE UNIVERSITY
School of Information Studies

STUDY GUIDE: KEY CONCEPTS

Make sure you understand the following key concepts by the end of Week 2:

- Data set types
  - Records, transactions, images, sequences, audios
- Variable types
  - Nominal or categorical, ordinal, numeric (interval and ratio)
- Data quality issues
  - Outliers, missing values, duplicate data
- Data summary and visualization
- Data transformation

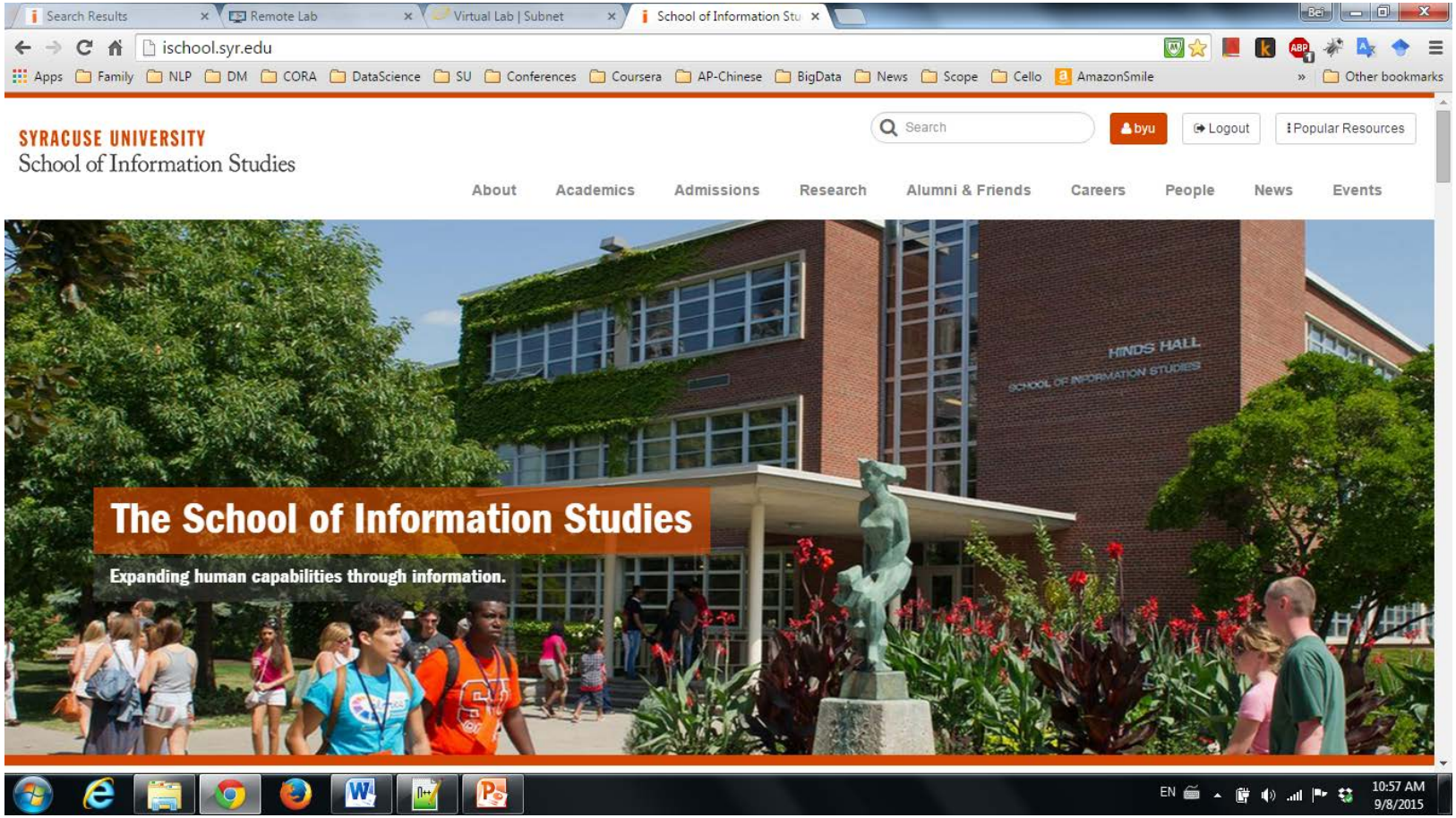
DATA SET TYPES

Record data: Data in the tabular format

- Each row is a data example.
- Each column is an attribute.
- Most common type of data set.

NAME	HIGHEST DEGREE	AGE	BLOOD TYPE
Jane	Middle School	25	A
John	High School	30	B
Amy	College	34	O
Larry	Grad School	31	AB

NONRECORD DATA



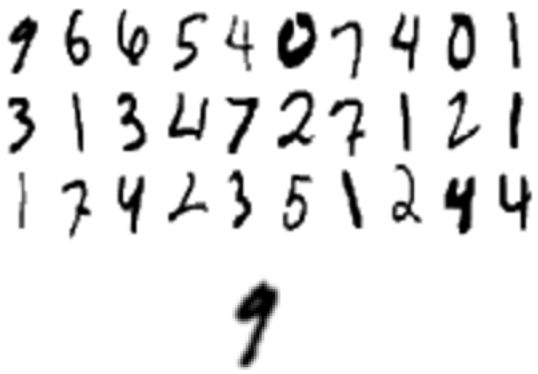
NONRECORD DATA: TEXT DOCUMENTS

Some data sets are not born as record data but can be converted to record format.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

IMAGE DATA

https://www.kaggle.com/c/digit-recognizer



	A	B	C	D	E
1	label	pixel0	pixel1	pixel2	pixel3
2	4	0	0	0	0
3	5	0	0	0	0
4	0	0	0	0	0
5	2	0	0	0	0
6	1	0	0	0	0
7	4	0	0	0	0
8	9	0	0	0	0
9	6	0	0	0	0
10	8	0	0	0	0

Each image is 28*28 pixels = 784 total.
Each pixel has a single pixel value [0, 255] associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker.

SEQUENCE DATA

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

PLAGIARISM DETECTION

Edit distance: The minimum number of steps needed to transform one sequence to the other

E.g., to transform “ABCD” to “ABCE,” one step is needed to transform “D” to “E.”

The algorithms used for comparing genomic sequences were used to detect plagiarism (e.g., turnitin.com) by replacing the nucleotides A, T, C, and G with words in text documents.

TRANSACTION DATA

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

TRANSACTION DATA

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Converted to record data

TID	Bread?	Coke?	Milk?	Diaper?	Beer?
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	1	0

SPARSE MATRIX

Most values in the matrix are “0”

Too many columns

Too few with nonzero values

TID	Bread?	Coke?	Milk?	Diaper?	Beer?
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	1	0

STORAGE OF SPARSE MATRIX

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Requires less space

TID	Bread?	Coke?	Milk?	Diaper?	Beer?
1	1	1	1	0	0
2	1	0	0	0	1
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	1	0

Requires more space

NETWORK DATA

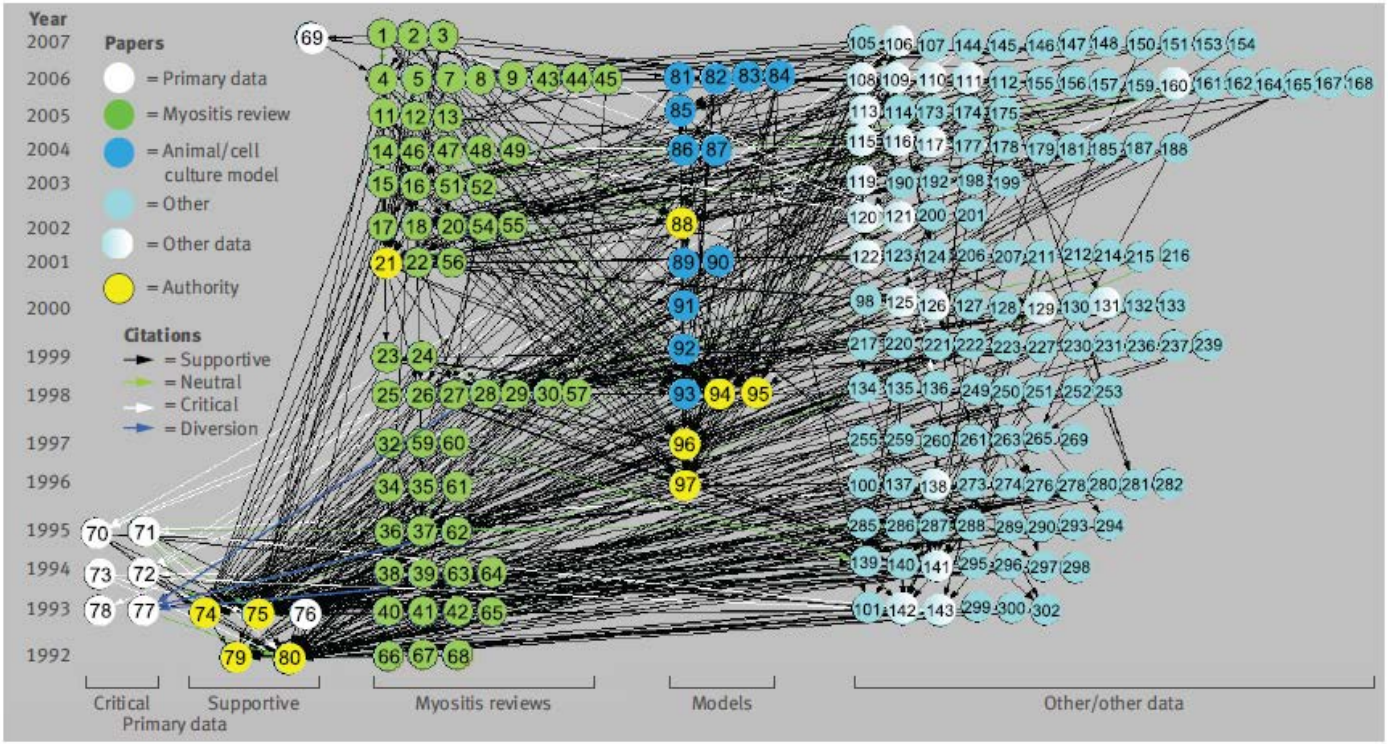


Fig 1 | Claim specific citation network. Citations regarding claim that β amyloid precursor protein mRNA or protein, or β amyloid protein, is abnormally present in inclusion body myositis muscle. The network is organised according to paper category and year of publication. Authority status (yellow) was defined computationally by network theory. Many citations flow to supportive primary data but not critical data. Papers are represented as nodes (n=218) and citations as directed edges (supportive n=636, neutral n=18, critical n=21, diversion n=3). Twenty four papers contain statements pertaining to claim but do not make or receive citations about it (not shown).

Steven G. Greenberg (2009). How citation distortions create unfound authority: analysis of citation network. BMJ 2009; 339: b2680 http://dx.doi.org/10.1136/bmj.b2680

REVIEW OF DATA SET TYPES

Record data

Nonrecord data

- Text data
- Image data
- Sequence data
- Transaction data
- Network data

ATTRIBUTE TYPES

SYRACUSE UNIVERSITY
School of Information Studies

HOW TO PREPARE DATA FOR ANALYSIS

Understand the meaning of data.

Assess the quality of data.

Transform data for analysis if necessary.

TEXTBOOK EXAMPLE 2.1

An illustration of data-related issues

Understand the definition of each attribute and the meaning of its values (e.g., fields 5, 4).

Understand the potential relationship between each attribute and your analysis goal (e.g., fields 1, 2, and 3 and target field 5).

Understand potential data quality problems, such as missing data, noise, and outliers (e.g., field 4).

UNDERSTAND THE MEANING OF VARIABLES

It might not be a trivial task, as this video shows.

ATTRIBUTE TYPES

SYRACUSE UNIVERSITY
School of Information Studies

ATTRIBUTE

An attribute is a property or characteristic of an object.

The value of an attribute can be different for different data examples.

DATA EXAMPLE: PERSON	ATTRIBUTE 1: HIGHEST DEGREE	ATTRIBUTE 2: AGE	ATTRIBUTE 3: BLOOD TYPE
Jane	Middle School	25	A
John	High School	30	B
Amy	College	34	O
Larry	Grad School	31	AB

ATTRIBUTE TYPES

Textbook Table 2.2

Four main types of attributes:

- Nominal
- Ordinal
- Interval
- Ratio

NOMINAL ATTRIBUTES

Definition: The values of a nominal attribute are just "names."

NAME	HIGHEST DEGREE	AGE	BLOOD TYPE
Jane	Middle School	25	A
John	High School	30	B
Amy	College	34	O
Larry	Grad School	31	AB

ORDINAL ATTRIBUTE

An ordinal attribute describes data objects in a qualitative, ordered way.

NAME	HIGHEST DEGREE	AGE	BLOOD TYPE
Jane	Middle School	25	A
John	High School	30	B
Amy	College	34	O
Larry	Grad School	31	AB

NUMERIC

Numbers that describe a measurable quantity

"How many"

"How much"

NAME	HIGHEST DEGREE	AGE	BLOOD TYPE
Jane	Middle School	25	A
John	High School	30	B
Amy	College	34	O
Larry	Grad School	31	AB

SIGNIFICANT DIGITS

Use a scale to measure people's weight.

Assume the scale is accurate to 0.1 lb.

If a person weighs three times and gets 150.0, 150.4, and 150.6, is the average weight 150.3333?

No, because the scale is accurate only to 0.1 lbs.

http://www.chem.sc.edu/faculty/morgan/resources/sigfigs/

SIGNIFICANT DIGITS IN R

```
> a = c(150.0,150.4,150.6)
> mean(a)
[1] 150.3333
> signif(mean(x), digits = 4)
[1] 150.3
```

INTERVAL VS. RATIO

Is there an arbitrary zero?

Yes: Ratio

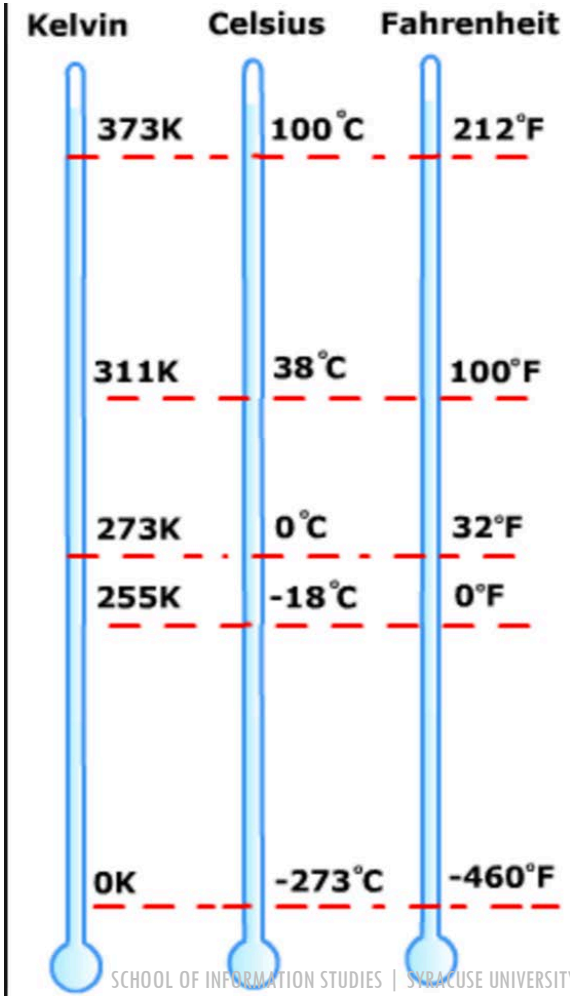
E.g., Kelvin scale of temperature, in which “0” means the lowest possible temperature. At this temperature, all atoms stop moving.

No: Interval

E.g., Celsius or Fahrenheit scale of temperature, in which “0” is **not** the lowest temperature.

RATIO VS. INTERVAL

A temperature of 200 degrees Kelvin is twice as warm as 100 degrees Kelvin.



A temperature of 200 degrees Fahrenheit is **not** twice as warm as 100 degrees Fahrenheit.

CONVERT ATTRIBUTE TYPE IN R

SYRACUSE UNIVERSITY
School of Information Studies

CONVERT DATA TYPE IN R

When reading data into tools like R, the tool might not interpret the data types correctly.

Examine data definitions in R:

```
> titanic <- read.csv("/Users/byu/Desktop/Data/titanic-train.csv", na.string = c(""))
```

```
> str(titanic)
```

OUTPUT

'data.frame': 891 obs. of 11 variables:

$ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...

$ Survived : int 0 1 1 1 0 0 0 0 1 1 ...

$ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...

$ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...

$ Age : num 22 38 26 35 35 NA 54 2 27 14 ...

$ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...

$ Parch : int 0 0 0 0 0 0 0 1 2 0 ...

$ Ticket : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 345 133 ...

$ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

$ Cabin : Factor w/ 147 levels "A10","A14","A16",..: NA 82 NA 56 NA NA 130 NA ...

$ Embarked : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...

PROBLEM WITH WRONG DATA TYPE

When R misinterpreted nominal variable “PassengerId” as numeric, it would calculate the mean and variance of passenger IDs, which does not make sense.

```
> summary(titanic)
  PassengerId    Survived Pclass      Sex
 Min.   :  1.0   0:549    1:216   female:314
 1st Qu.:223.5   1:342    2:184   male   :577
 Median :446.0            3:491
 Mean   :446.0
 3rd Qu.:668.5
 Max.   :891.0
```

CONVERT DATA TYPE IN R

R treats nominal variables as “factors”:

```
> titanic$Survived=factor(titanic$Survived)
> str(titanic)
```

Output:

```
...
 $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
...
```

CONVERT DATA TYPE IN R

R treats ordinal variables as “ordered factors”:

```
> titanic$Pclass=ordered(titanic$Pclass)
> str(titanic)
```

Output:

```
...
 $ Pclass     : Ord.factor w/ 3 levels "1"<"2"<"3": 3 1 3 1 3 3 1 3 3 2 ...
...
```

ORDERED FACTOR IN R

Month defined as a list:

```
> mons=c("Jan", "Jan", "Feb", "Feb", "Mar", "Apr", "May", "Jun"
, "Jul", "Aug", "Sep", "Oct", "Oct", "Nov", "Dec", "Dec")
> table(mons)
mons
Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
  1   1   2   2   2   1   1   1   1   1   2   1
```

Month defined as an ordered factor:

```
> mons_factor=factor(mons, levels=c("Jan", "Feb", "Mar", "Apr",
"May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"), ordere
d=TRUE)
> table(mons_factor)
mons_factor
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
  2   2   1   1   1   1   1   1   1   2   1   2
```

DATA QUALITY ISSUES

SYRACUSE UNIVERSITY
School of Information Studies

DATA QUALITY ISSUES

Noise

Outliers

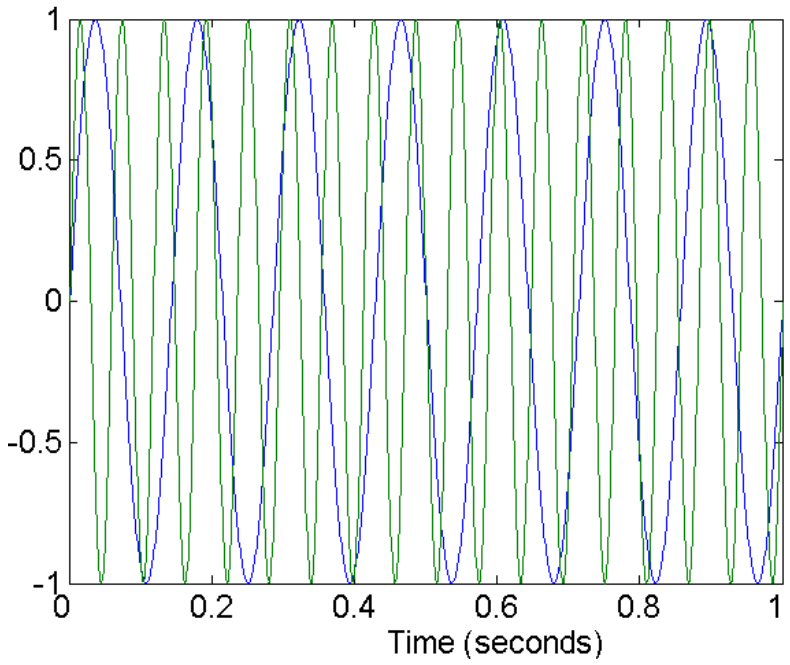
Missing values

Duplicate data

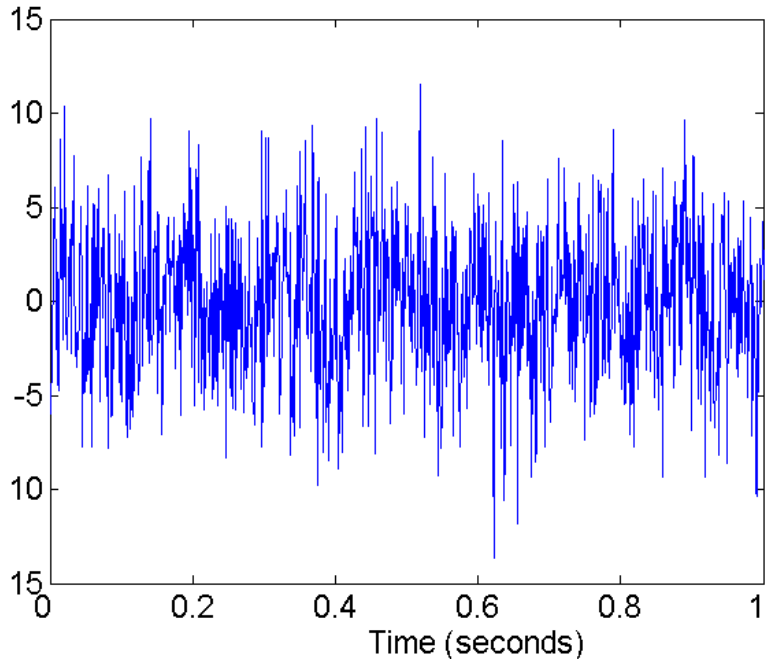
NOISE

Noise refers to modification of original values.

Examples: Distortion of a person's voice when talking on a poor-quality phone and "snow" on television screen



Two Sine Waves

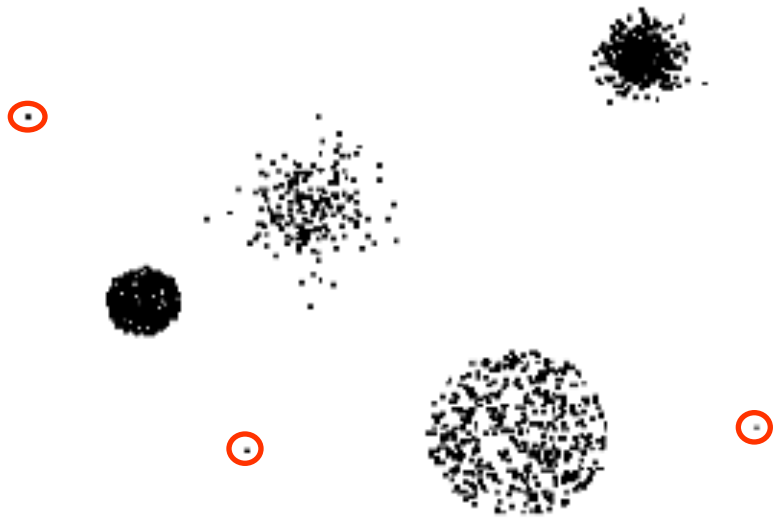


Two Sine Waves + Noise

OUTLIERS

Outliers are data objects with characteristics that are considerably different from most of the other data objects in the data set.

E.g., 250 would be an outlier for variable "people's age."



OUTLIERS SHOULD BE DETECTED AND ANALYZED CAREFULLY

Each year, satellites measure the ozone level over Antarctica.

In the early 1980s, however, scientists were so astounded in detecting a dramatic seasonal drop in ozone levels over Antarctica by a flyover that they spent two years rechecking their satellite data.

They discovered that satellites had dutifully been recording the ozone collapse, but the computers had not raised an alert because they were programmed to reject such extreme data as anomalies.

http://www.stat.wmich.edu/s160/book/node8.html

MISSING VALUES

Why are values missing?

- Information is not collected.
  (E.g., people decline to give their age and weight.)
- Attributes may not be applicable to all cases.
  (E.g., annual income is not applicable to most children.)

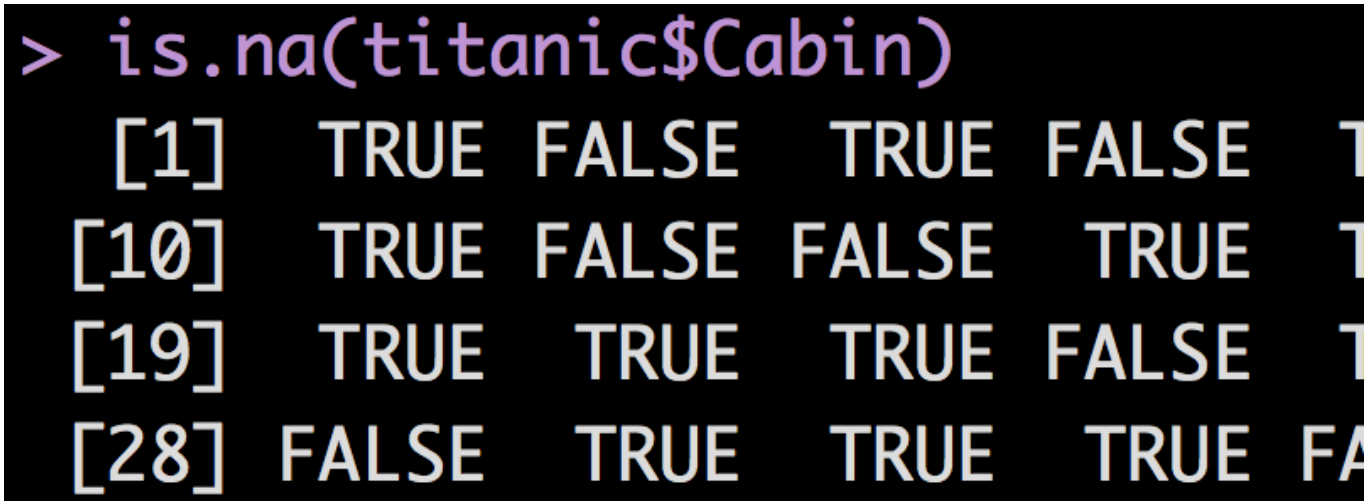
Handling missing values:

- Eliminate data objects.
- Ignore the missing value during analysis.
- Estimate missing values and replace them.

CHECK MISSING VALUES IN R

>is.na(titanic)

>is.na(titanic$Cabin)



FIND COMPLETE RECORDS

```
> titanic[complete.cases(titanic),]
    PassengerId Survived Pclass    Sex   Age SibSp Parch
2             2        1      1 female 38.00     1     0
4             4        1      1 female 35.00     1     0
7             7        0      1   male 54.00     0     0
11           11        1      3 female  4.00     1     1
12           12        1      1 female 58.00     0     0
22           22        1      2   male 34.00     0     0
```

```
> nrow(titanic[!complete.cases(titanic),])
[1] 708
```

```
> nrow(titanic[complete.cases(titanic),])
[1] 183
```

COUNT NUMBER OF MISSING VALUES

```
> length(which(is.na(titanic$Age)))
[1] 177
```

Is “age” still a useful variable for predicting survivors?

ESTIMATE AND REPLACE MISSING VALUES

```
> titanic$Age[is.na(titanic$Age)] <- mean(titanic$Age, na.rm =
TRUE)
> length(which(is.na(titanic$Age)))
[1] 0
```

REMOVE RECORDS WITH MISSING VALUES

```
> titanic_new <- titanic[complete.cases(titanic),]
> nrow(titanic_new)
[1] 202
> titanic_new2 <- na.omit(titanic)
> nrow(titanic_new2)
[1] 202
```

Isn't the number of complete cases 183?

It was, but remember the missing values in "age" have been replaced by average age.

REMOVE VARIABLES WITH MISSING VALUES

```
> myVars=c("Pclass", "Sex", "Age", "SibSp", "Fare", "Survived")
> titanic_new3 <- titanic[myVars]
> str(titanic_new3)
'data.frame':   891 obs. of  6 variables:
 $ Pclass  : Ord.factor w/ 3 levels "1"<"2"<"3": 3 1 3 1 3 3 1
3 3 2 ...
 $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2
2 1 1 ...
 $ Age     : num  22 38 26 35 35 ...
 $ SibSp   : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Survived: Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ..
```

DUPLICATE DATA

Data set may include data objects that are duplicates or almost duplicates of one another.

- Major issue when merging data from heterogenous sources

Examples:

- Same person with multiple e-mail addresses

Data cleaning:

- Process of dealing with duplicate data issues

AN EXAMPLE OF DUPLICATE DATA

An Amazon Mechanical Turk worker set up two accounts and finished a task twice in order to get double payment.

Two identical records were sent to the data collector.

How to identify them?

- Check IP address.
- Compare similarity between records.

CHECK AND REMOVE DUPLICATED RECORDS

```
>  nrow(titanic[duplicated(titanic),])
[1] 0
```

```
> titanic_new4 <- titanic[!duplicated(titanic),]
> nrow(titanic_new4)
[1] 891
```

REVIEW DATA QUALITY ISSUES

Noise

Outliers

Missing values

Duplicate data

SUMMARY STATISTICS

SYRACUSE UNIVERSITY
School of Information Studies

SUMMARY STATISTICS

Common summary statistics:

- Central tendency of data
- Data spread

Different statistical measures for different variable types:

- Numeric
- Nominal

NUMERIC VARIABLES

Central tendency
- Mean
- Median
- Mode

Data spread
- Standard deviation
- Variance
- Min., max., quartiles

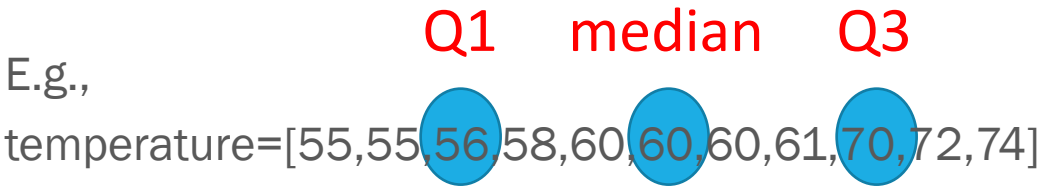
DATA SPREAD

Range: Max. – Min.

Variance/standard deviation

Quartile: Q1 (25%), Q2 (50%), Q3 (75%), Q4 (100%)

E.g.,



Interquartile range (IQR): Q3 – Q1 = 70 – 56 = 14

SUMMARIZE NUMERIC VARIABLE IN R

Summarize the whole data set:

```
summary(titanic)
```

Summarize central tendency:

```
mean(titanic$Age)
median(titanic$Age)
freq=table(titanic$Age)
table(titanic$Age)[which.max(table(titanic$Age))]  # mode
```

SUMMARIZE NUMERIC VARIABLE IN R

Summarize data spread:

```
var(titanic$Age) # variance
sd(titanic$Age) # standard deviation
max(titanic$Age)
min(titanic$Age)
range <- max(titanic$Age) – min(titanic$Age)
qt <- quantile(titanic$Age, na.rm=TRUE) # quartile, remove missing values
IQR=qt[['75%']]-qt[['25%']] # Interquartile range
```

NOMINAL VARIABLES

Central tendency: Mode

Data spread: Distribution

```
> table(titanic$Sex)

female   male
   314    577
> table(titanic$Sex)[which.max(table(titanic$Sex))]
male
 577
```

ORDINAL VARIABLES

Can we treat ordinal variables as nominal or numeric?

CUSTOMER ID	RATING
1	*
2	*****
3	**
4	***
5	****
6	**
…	…
Average	?

VISUALIZATION

SYRACUSE UNIVERSITY
School of Information Studies

DATA VISUALIZATION

Complementary approach for data analysis in a visual way

Basic tools for numeric variables

- Histogram: Show distribution of one variable.
- Box plot: Use five key values to show distribution.
- Scatter plot: Plot relationship between two variables.

Basic tools for nominal variables

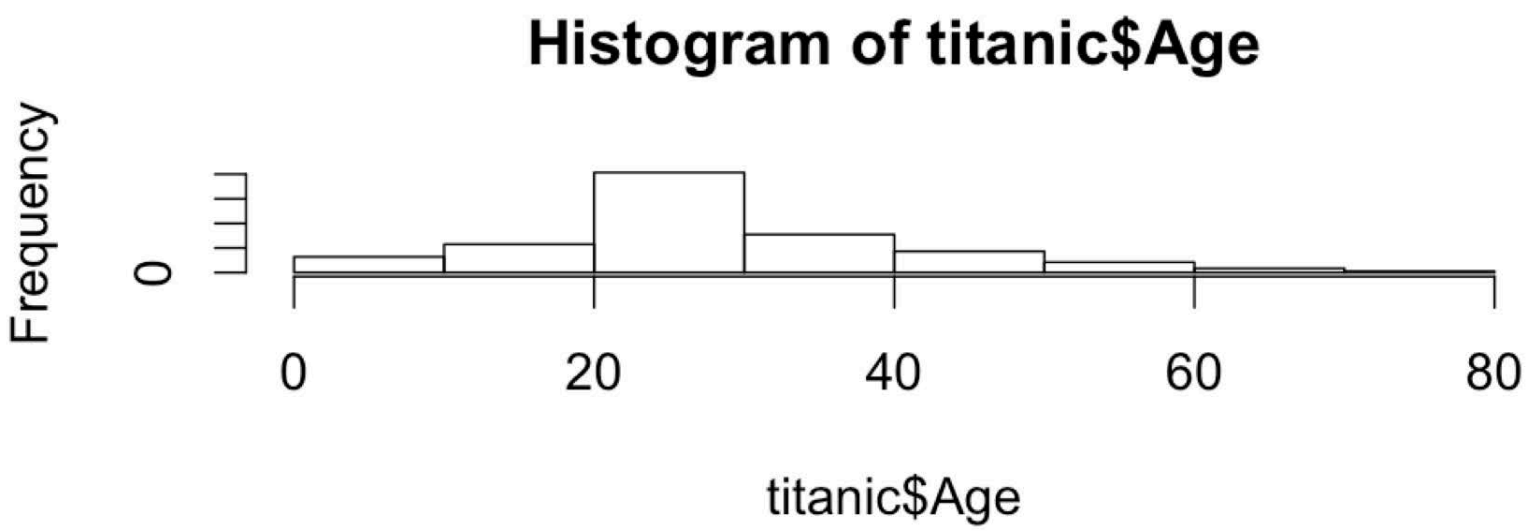
- Pie chart or bar chart to show frequencies
- Cross-tab to show relationship between two variables

HISTOGRAM

```
> hist(titanic$Age)
>
```

Files | Plots | Packages | Help | Viewer

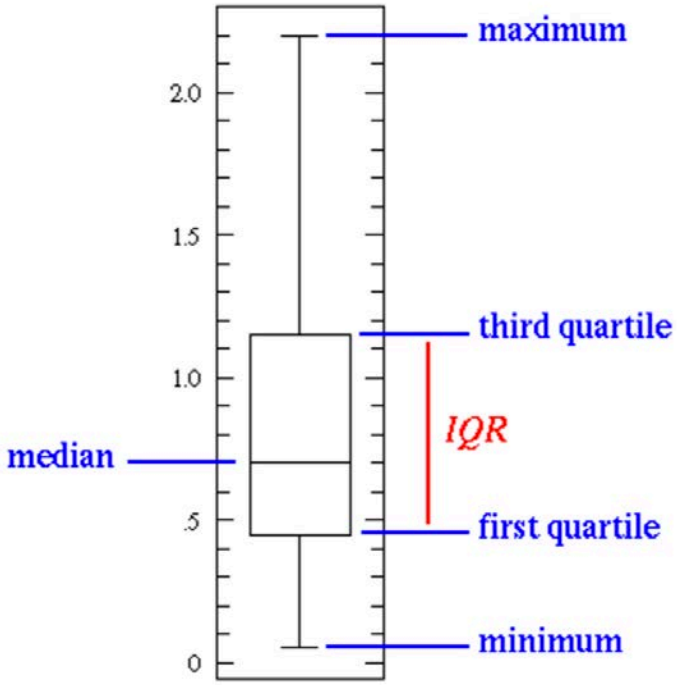
Zoom | Export | Pu



BOX PLOT

Five values of box plot
- Bottom of box: Q1
- Top of box: Q3
- Band near the middle of box: Median
- Upper whisker = min.(max, Q3 + 1.5IQR)
- Lower whisker = max.(min, Q1 – 1.5IQR)

Use visualization for outlier detection
- http://www.stat.wmich.edu/s160/book/node8.html
- http://www.r-bloggers.com/about-boxplot/

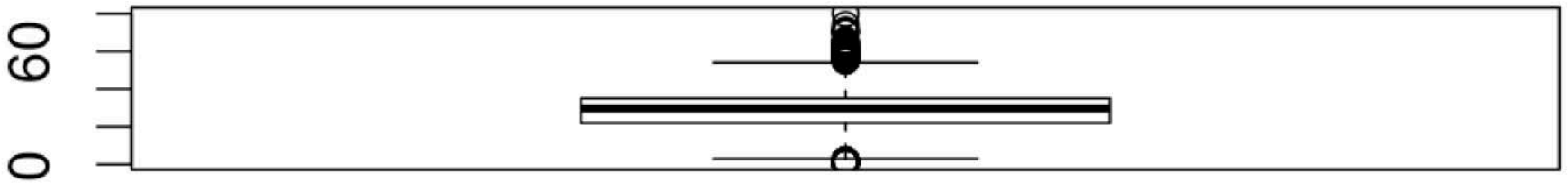


BOX PLOT

```
> boxplot(titanic$Age)
> |
```

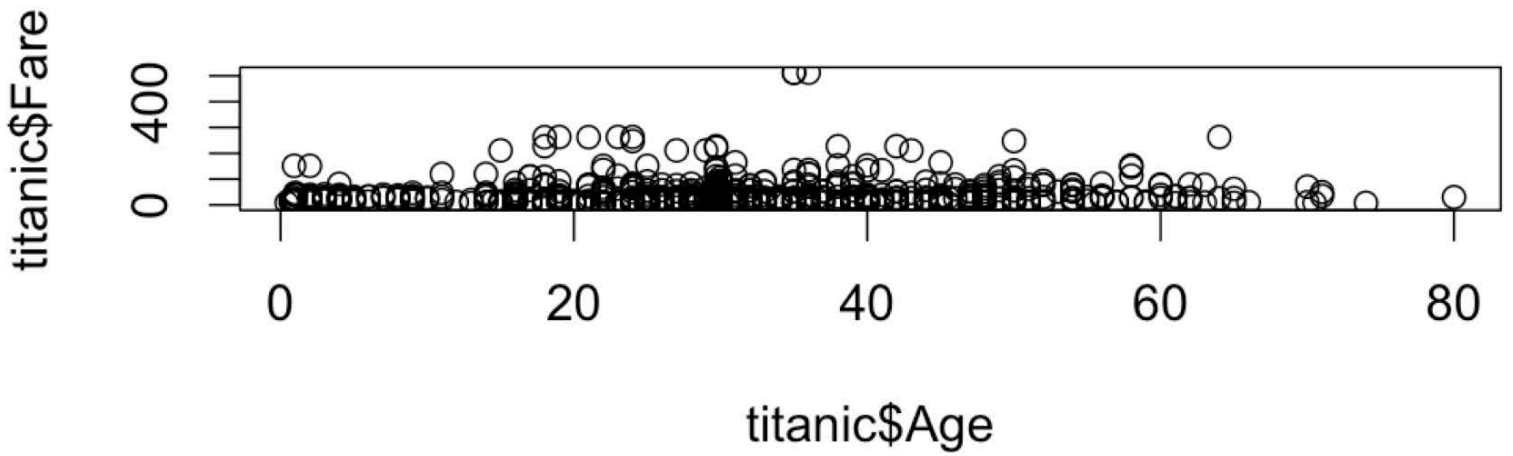
Files | Plots | Packages | Help | Viewer

Zoom | Export | Pu



SCATTER-PLOT TWO NUMERIC VARIABLES

```
> plot(titanic$Age, titanic$Fare)
```



CROSS-TAB TWO NOMINAL VARIABLES

```
> table(titanic$Sex, titanic$Survived)

            0   1
  female   81 233
  male    468 109
```

RELATIONSHIP BETWEEN A NOMINAL AND A NUMERIC VARIABLE

```
> male=titanic[titanic$Sex=='male',]
> mean(male$Fare)
[1] 25.52389
> female=titanic[titanic$Sex=='female',]
> mean(female$Fare)
[1] 44.47982
```

CORRELATION VS. CAUSATION

Did ice cream cause drowning?
https://www.youtube.com/watch?v=8B271L3NtAw

"Cargo Cult Science" (Richard Feynman)
http://calteches.library.caltech.edu/51/2/CargoCult.htm
Search "a man named Young" in the article, and read that story.

AGGREGATION

SYRACUSE UNIVERSITY
School of Information Studies

DATA TRANSFORMATION

Aggregation

Attribute transformation

Dimensionality reduction and feature selection (covered in future weeks)

AGGREGATION

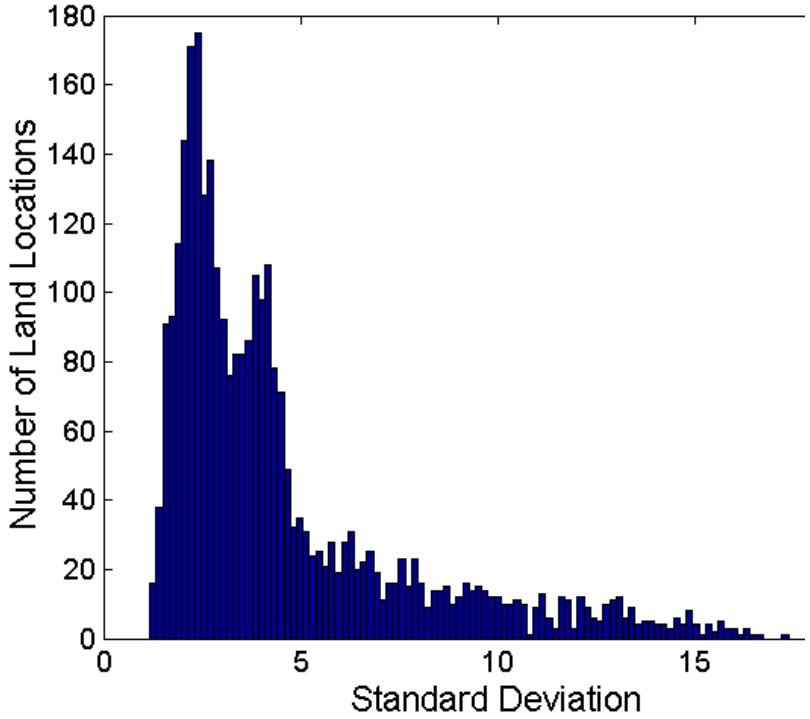
Combining two or more rows or columns into a single row or column

Purpose:

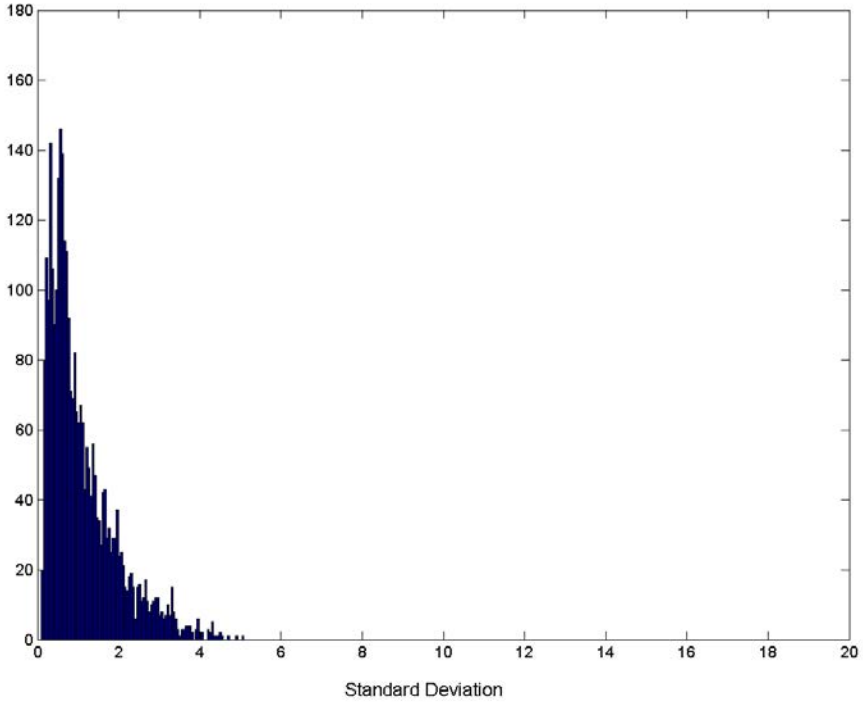
- Data reduction
  - All test scores merged into one total score
- Change of scale
  - Cities aggregated into regions, states, countries, etc.
- More “stable” data
  - Aggregated data tends to have less noise

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
```

TRANSFORMATION

SYRACUSE UNIVERSITY
School of Information Studies

ATTRIBUTE TRANSFORMATION

Sometimes, the original values of an attribute need to be transformed for purpose of analysis.

Some common transformations:

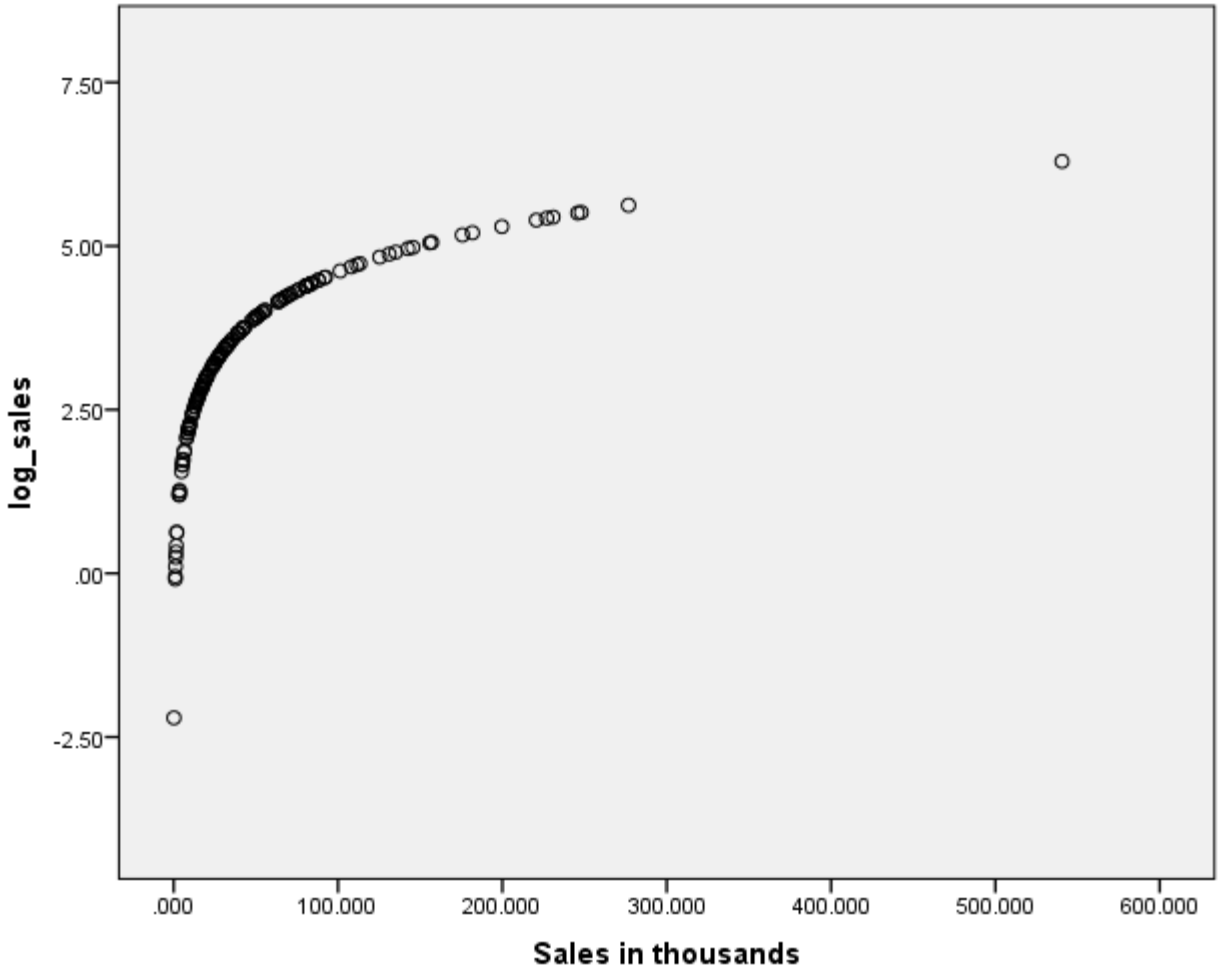
- Discretization
- Log transformation
- Normalization
  - Z-score
  - Min_max

DISCRETIZATION (BINNING)

Discretization is a process to transform a continuous attribute to a discrete one.

```
> age <- cut(titanic$Age, breaks = c(0,10,20,30,40,50,60,Inf),la
bels=c("child","teens","twenties","thirties","fourties","fifties
","old"))
> age
  [1] twenties thirties twenties thirties thirties <NA>
  [7] fifties  child    twenties teens    child    fifties
 [13] teens    thirties teens    fifties  child    <NA>
 [19] thirties <NA>     thirties thirties teens    twenties
```

LOG TRANSFORMATION

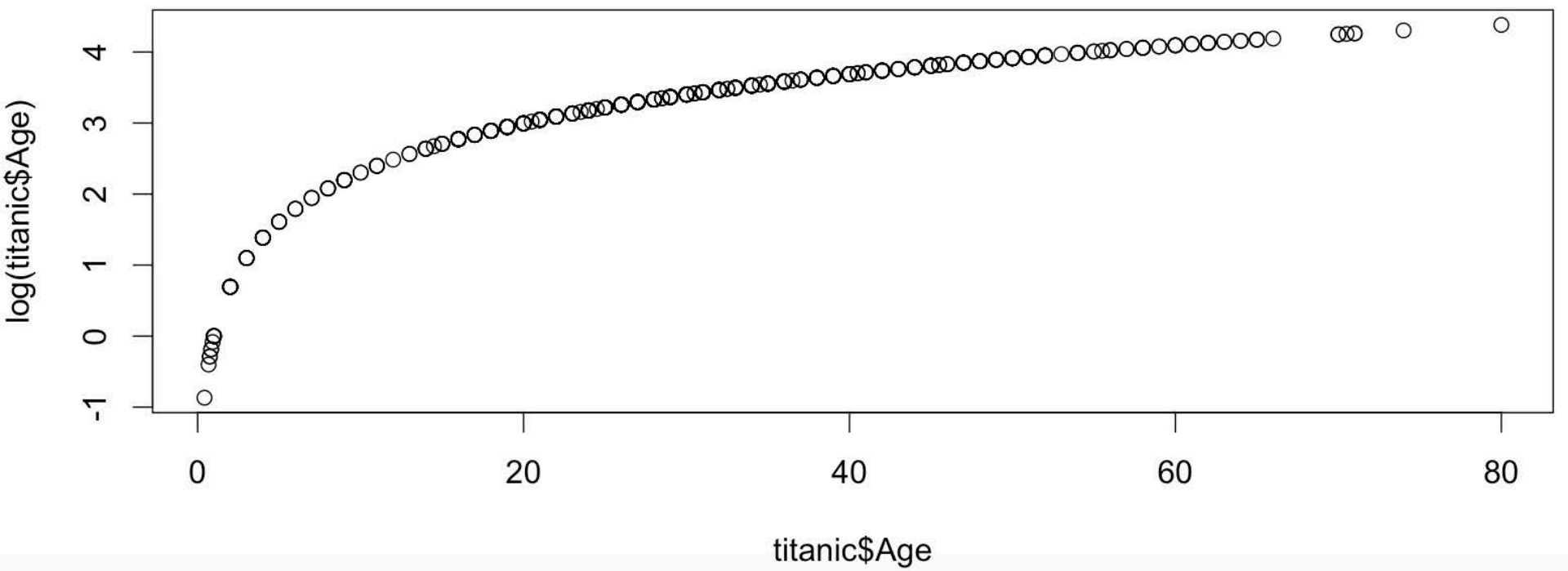


Log transformation leaves the analysis more robust with outliers.

```
> plot(titanic$Age, log(titanic$Age))
> 
```

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publi



Z-SCORE TRANSFORMATION

A data analysis problem: What Facebook messages sent from restaurants are popular among fans? Popularity is measured by the number of comments received.

Two restaurants: McDonald's and Lemon Grass

- McDonald's has millions of fans on Facebook, while Lemon Grass has thousands.
- A message from McDonald's received 1,000 comments.
- A message from Lemon Grass also received 1,000 comments.
- Which message is more popular, or are they equally popular?

http://research.microsoft.com/en-us/um/beijing/events/ia2011/n7.pdf

Z-SCORE TRANSFORMATION

The message from Lemon Grass seems more popular, but right now the face values look the same: 1,000.

How to demonstrate the real difference in popularity?

Z-SCORE TRANSFORMATION

Assume:

McDonald's

- Average number of comments: *u* = 2,000
- Standard deviation: *sd* = 500

Lemon Grass

- Average number of comments: *u* = 200
- Standard deviation: *sd* = 50

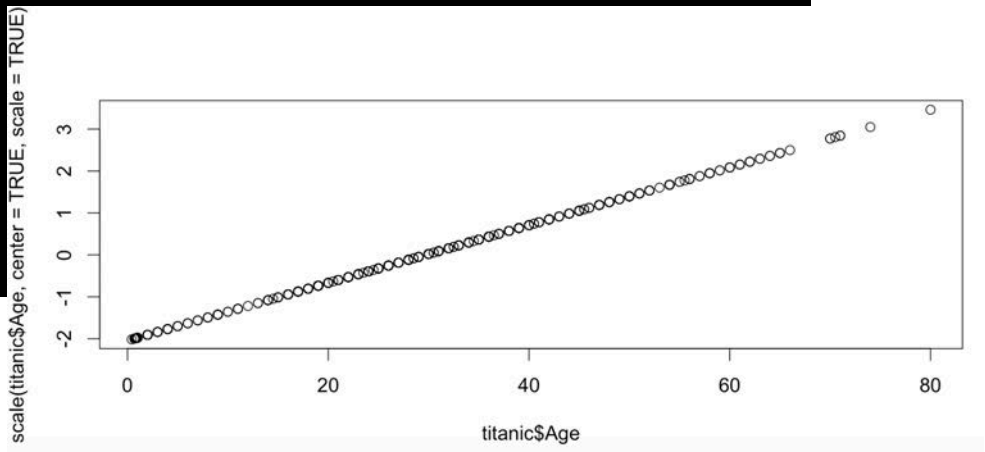
$Z(x) = (x - u)/sd$

Z-SCORE TRANSFORMATION

Facebook Messages	# Comments	Z-Score	
McDonald's msg 1	1,000	-2	
McDonald's msg 2	500	-3	
…			
Lemon Grass msg 1	1,000	16	
Lemon Grass msg 2	500	6	
…			

Z-SCORE IN R

```
> scale(titanic$Age, center = TRUE, scale = TRUE)
              [,1]
  [1,] -0.53000510
  [2,]  0.57143041
  [3,] -0.25464622
  [4,]  0.36491125
  [5,]  0.36491125
  [6,]          NA
```



MIN_MAX TRANSFORMATION

Assume:

McDonald's

- Minimum number of comments: *min* = 50
- Maximum number of comments: *max* = 10,000

Lemon Grass

- Minimum number of comments: *min* = 10
- Maximum number of comments: *max* = 2,000

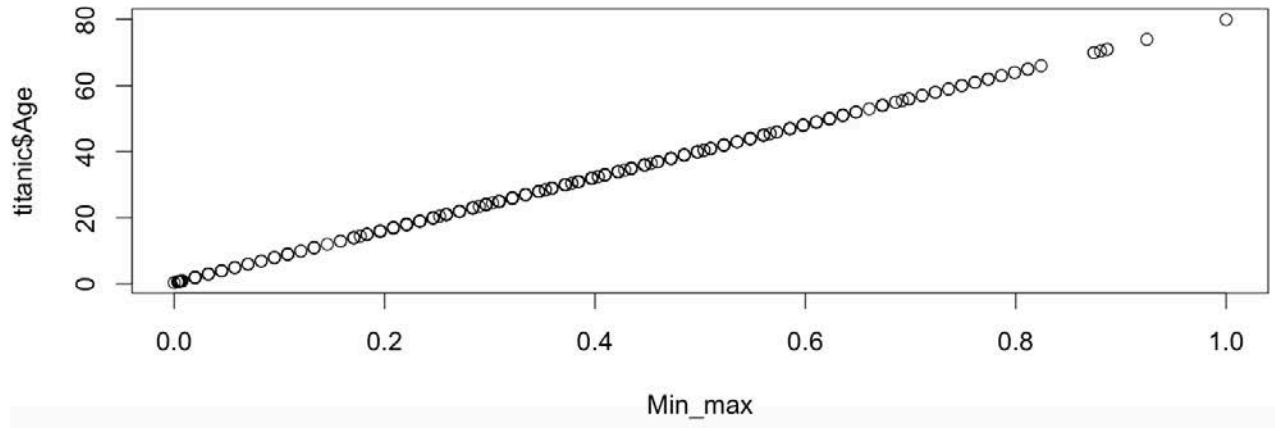
Min_max(x) = (x-min)/(max-min)

MIN_MAX TRANSFORMATION

Facebook Messages	# Comments	Z-Score	Min_Max
McDonald's msg 1	1,000	-2	.10
McDonald's msg 2	500	-3	.05
…			
Lemon Grass msg 1	1,000	16	.50
Lemon Grass msg 2	500	6	.25
…			

MIN_MAX IN R

```
> Min_max <- (titanic$Age-min(titanic$Age,na.rm=TRUE))/(max(tita
nic$Age,na.rm=TRUE)-min(titanic$Age,na.rm=TRUE))
> Min_max
  [1] 0.271173662 0.472229203 0.321437547 0.434531289
  [5] 0.434531289           NA 0.673284745 0.019854235
  [9] 0.334003518 0.170645891 0.044986177 0.723548630
 [13] 0.246041719 0.484795175 0.170645891 0.685850716
```



MANY MORE TRANSFORMATIONS

TFIDF (Textbook exercise 16 on page 92)

16. Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i}, \tag{2.1}$$

where df_i is the number of documents in which the i^{th} term appears and is known as the **document frequency** of the term. This transformation is known as the **inverse document frequency** transformation.

(a) What is the effect of this transformation if a term occurs in one document? In every document?

Terms that occur in every document have 0 weight, while those that occur in one document have maximum weight, i.e., $\log m$.

(b) What might be the purpose of this transformation?

This normalization reflects the observation that terms that occur in every document do not have any power to distinguish one document from another, while those that are relatively rare do.

A REVIEW OF DATA TRANSFORMATION

Aggregation

Discretization

Log transformation

Z-score transformation

Min_max transformation

SAMPLING

SYRACUSE UNIVERSITY
School of Information Studies

SAMPLING

Why sampling?

Sampling when obtaining or analyzing the entire set of data of interest is too expensive or time consuming,

And the sample is representative, meaning, it has approximately the same property (of interest) as the original set of data.

Therefore, the analysis results on the sample data may be reliably generalized to the entire data set.

SAMPLING METHODS

A sampling task: To sample 300 college students in Syracuse University to study their social media use patterns

Convenience sampling:

Sample iSchool students to represent SU students.

Random sampling:

Randomly sample students around campus.

Stratified sampling:

Sample equal number of students from each school.

Systematic sampling:

E.g., sort students' SUID numbers in increasing order; pick the 1st, 11th, 21st, 31st, … , students until 300 students are sampled.

SAMPLING WITH AND WITHOUT REPLACEMENT

Sampling without replacement:

One item would occur in the sample at most once.

Sampling with replacement:

One item may occur in the sample multiple times.

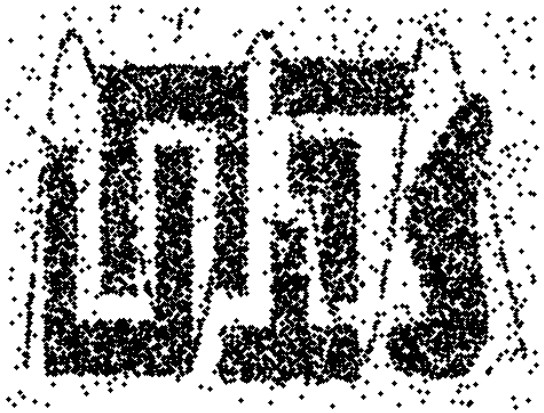
RANDOM SAMPLING

```
> sample <- titanic[sample(1:nrow(titanic), 100, replace=FALSE),
]
> table(titanic$Survived)

  0   1
549 342
> table(sample$Survived)

 0  1
62 38
```

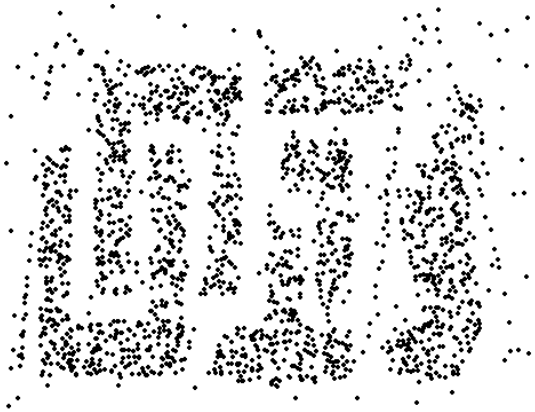
SYSTEMATIC SAMPLING

```
> #sample lines #1, #11, #21, ...
> ss=titanic[seq(1, nrow(titanic),10),]
> nrow(ss)
[1] 90
```

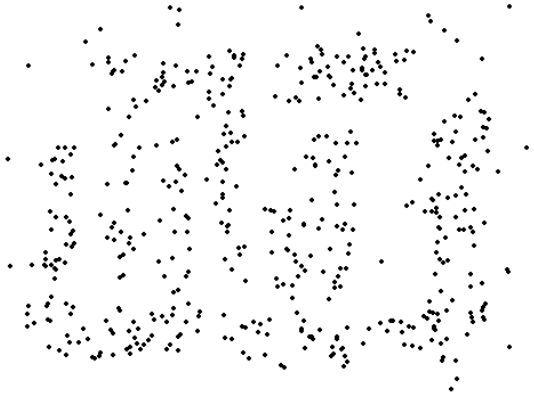
SAMPLE SIZE



8,000 points



2,000 Points



500 Points

A REVIEW OF SAMPLING METHODS

Convenience sampling

Random sampling

Stratified sampling

Systematic sampling

Sample representativeness