Statistics 202: Statistical Aspects of Data Mining

Professor David Mease

Tuesday, Thursday 9:00-10:15 AM Terman 156

Lecture 2 = Start chapter 2

Agenda:

1) Assign Chapters 1 and 2 Homework due 7/10
2) Lecture over first part of chapter 2

Homework Assignment:

Chapters 1 and 2 homework is due Tuesday 7/10

Either email to me (dmease@stanford.edu), bring it to class, or put it under my office door.

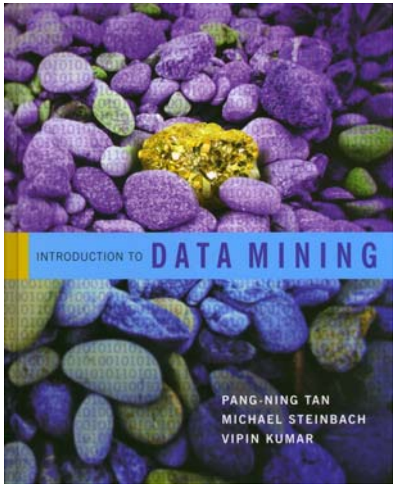
SCPD students may use email or use courier.

The assignment is posted at

http://www.stats202.com/homework.html

Introduction to Data Mining

by
Tan, Steinbach, Kumar



Chapter 2: Data

What is Data?

- An attribute is a property or characteristic of an object

- Examples: eye color of a person, temperature, etc.

- Attribute is also known as variable, field, characteristic, or feature

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- A collection of attributes describe an object

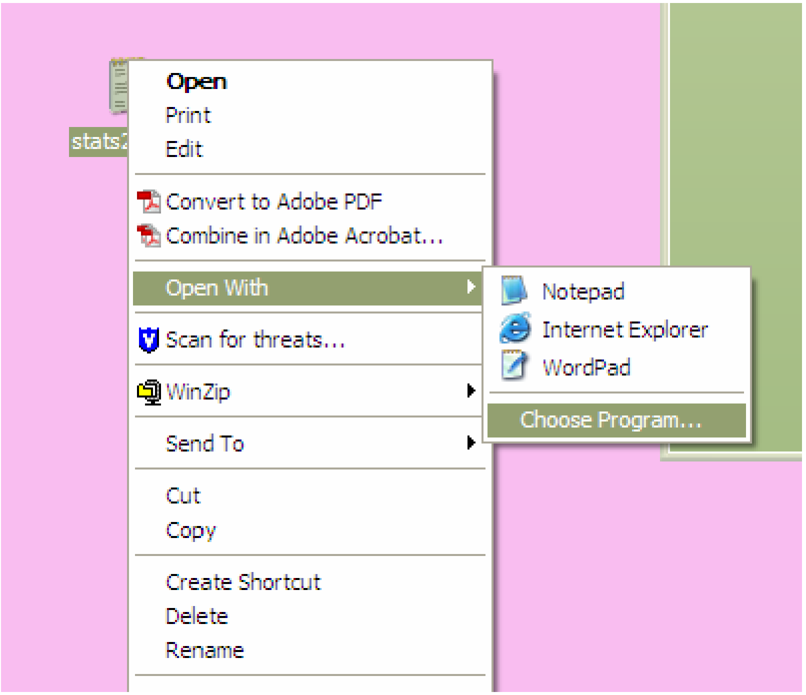
- Object is also known as record, point, case, sample, entity, instance, or observation

Reading Data into Excel

Download it from the web at

www.stats202.com/stats202log.txt

The right click on it and select “Open With” then “Choose Program”

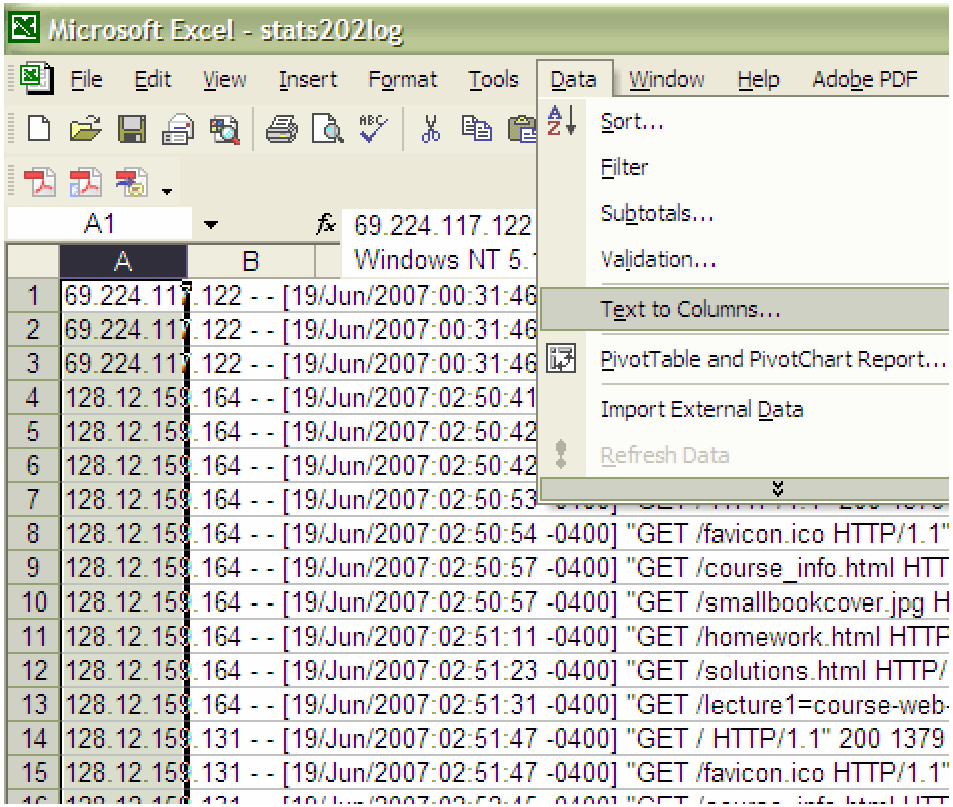


Reading Data into Excel

Choose Excel

Oops! Everything is in the first column!

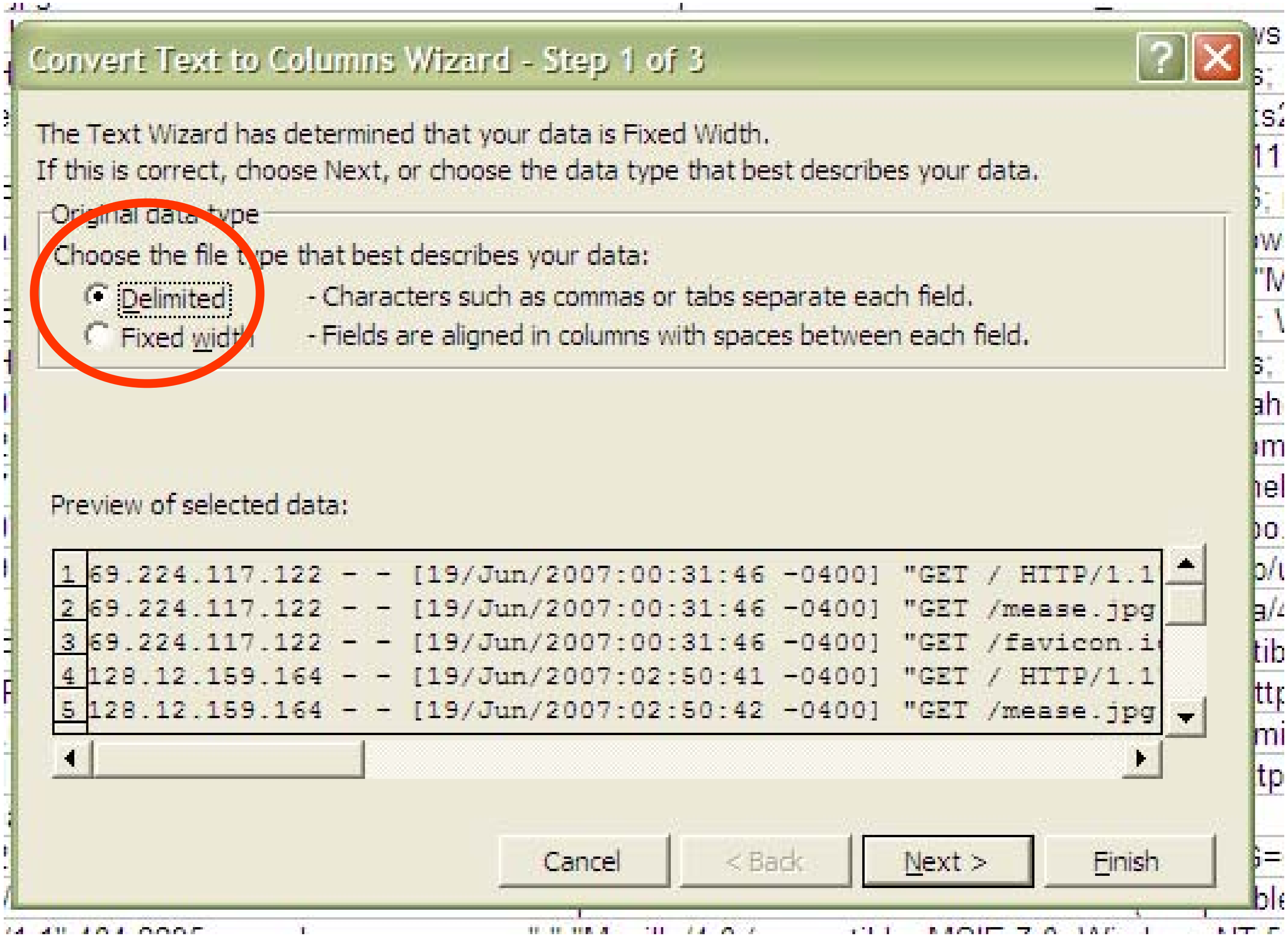
Solution: click on the first column then select “Data” then “Text to Columns”



```
   69.224.117.122
   Windows NT 5.
1  69.224.117.122 - - [19/Jun/2007:00:31:46
2  69.224.117.122 - - [19/Jun/2007:00:31:46
3  69.224.117.122 - - [19/Jun/2007:00:31:46
4  128.12.159.164 - - [19/Jun/2007:02:50:41
5  128.12.159.164 - - [19/Jun/2007:02:50:42
6  128.12.159.164 - - [19/Jun/2007:02:50:42
7  128.12.159.164 - - [19/Jun/2007:02:50:53
8  128.12.159.164 - - [19/Jun/2007:02:50:54 -0400] "GET /favicon.ico HTTP/1.1"
9  128.12.159.164 - - [19/Jun/2007:02:50:57 -0400] "GET /course_info.html HTT
10 128.12.159.164 - - [19/Jun/2007:02:50:57 -0400] "GET /smallbookcover.jpg H
11 128.12.159.164 - - [19/Jun/2007:02:51:11 -0400] "GET /homework.html HTTP
12 128.12.159.164 - - [19/Jun/2007:02:51:23 -0400] "GET /solutions.html HTTP/
13 128.12.159.164 - - [19/Jun/2007:02:51:31 -0400] "GET /lecture1=course-web-
14 128.12.159.131 - - [19/Jun/2007:02:51:47 -0400] "GET / HTTP/1.1" 200 1379
15 128.12.159.131 - - [19/Jun/2007:02:51:47 -0400] "GET /favicon.ico HTTP/1.1"
```

Reading Data into Excel

Choose “Delimited” and hit “Next”



Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Fixed Width.
If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

● Delimited - Characters such as commas or tabs separate each field.

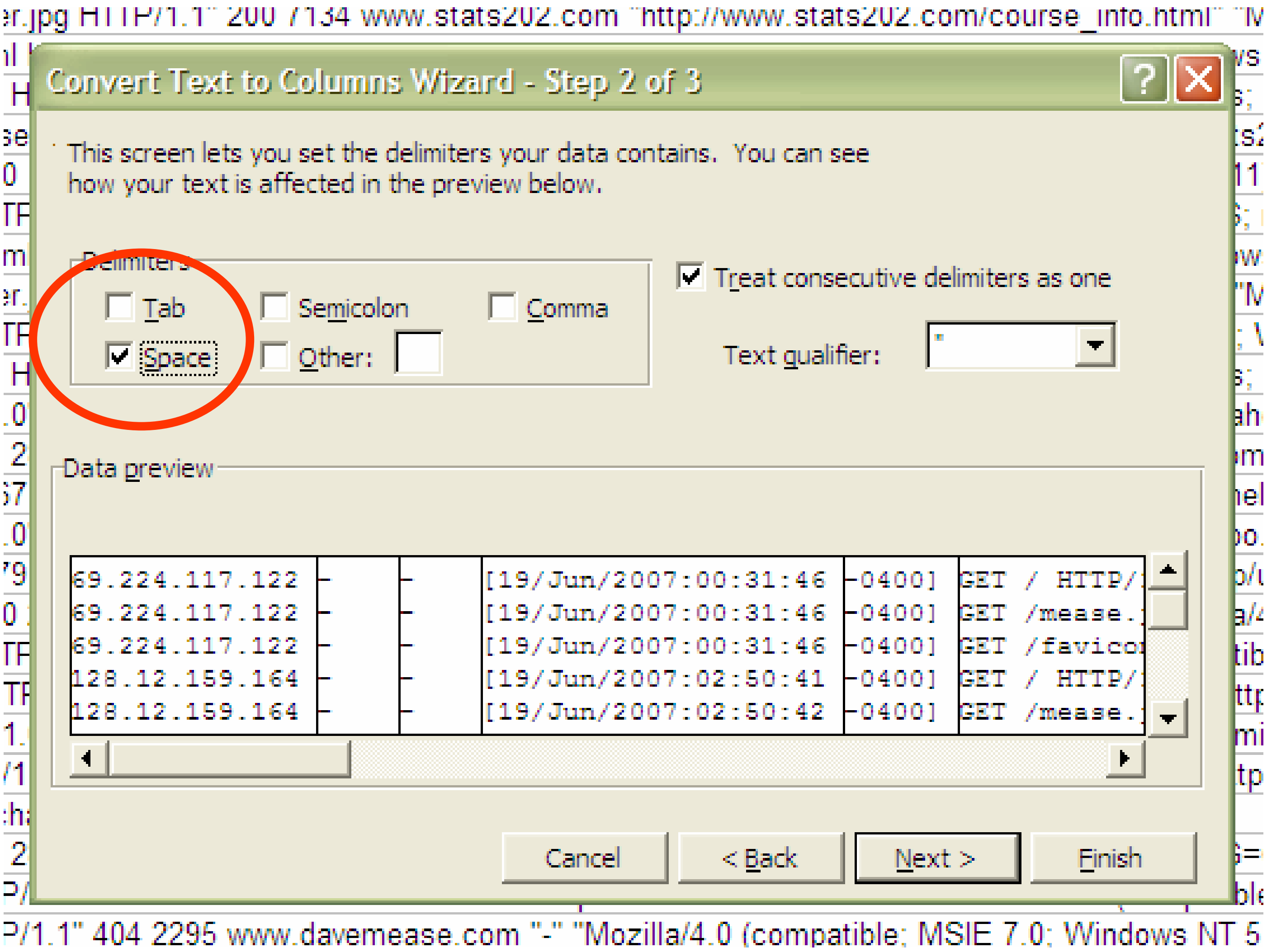
○ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
1 69.224.117.122 - - [19/Jun/2007:00:31:46 -0400] "GET / HTTP/1.1
2 69.224.117.122 - - [19/Jun/2007:00:31:46 -0400] "GET /mease.jpg
3 69.224.117.122 - - [19/Jun/2007:00:31:46 -0400] "GET /favicon.i
4 128.12.159.164 - - [19/Jun/2007:02:50:41 -0400] "GET / HTTP/1.1
5 128.12.159.164 - - [19/Jun/2007:02:50:42 -0400] "GET /mease.jpg
```

Reading Data into Excel

Check only “Space” then “Next” again



er.jpg HTTP/1.1" 200 7134 www.stats202.com "http://www.stats202.com/course_info.html" "M

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☐ Tab ☐ Semicolon ☐ Comma

☑ Space ☐ Other:

☑ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
69.224.117.122 - - [19/Jun/2007:00:31:46 -0400] GET / HTTP/
69.224.117.122 - - [19/Jun/2007:00:31:46 -0400] GET /mease.
69.224.117.122 - - [19/Jun/2007:00:31:46 -0400] GET /favico
128.12.159.164 - - [19/Jun/2007:02:50:41 -0400] GET / HTTP/
128.12.159.164 - - [19/Jun/2007:02:50:42 -0400] GET /mease.
```

P/1.1" 404 2295 www.davemease.com "-" "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5

Reading Data into Excel

Q: Why did this work? Why don’t all spaces cause column splits?

Reading Data into Excel

Q: Why did this work? Why don't all spaces cause column splits?

A: The file is *escaped* using quotes.

(read http://en.wikipedia.org/wiki/Delimiter for more information)

Reading Data into R

Download it from the web at

www.stats202.com/stats202log.txt

Set your working directory:

```
> setwd("C:/Documents and
Settings/Administrator/Desktop")
```

Read it in:

```
> data<-read.csv("stats202log.txt",
     sep=" ",header=F)
```

Reading Data into R

Look at the first 5 rows:

```
> data[1:5,]
```

```
              V1 V2 V3                    V4     V5                        V6  V7   V8                V9
1 69.224.117.122  -  - [19/Jun/2007:00:31:46 -0400]            GET / HTTP/1.1 200 2867 www.davemease.com
2 69.224.117.122  -  - [19/Jun/2007:00:31:46 -0400]   GET /mease.jpg HTTP/1.1 200 4583 www.davemease.com
3 69.224.117.122  -  - [19/Jun/2007:00:31:46 -0400] GET /favicon.ico HTTP/1.1 404 2295 www.davemease.com
4 128.12.159.164  -  - [19/Jun/2007:02:50:41 -0400]            GET / HTTP/1.1 200 2867 www.davemease.com
5 128.12.159.164  -  - [19/Jun/2007:02:50:42 -0400]   GET /mease.jpg HTTP/1.1 200 4583 www.davemease.com
                                                             V10                                                                                        V11 V12
1 http://search.msn.com/results.aspx?q=mease&first=21&FORM=PERE2                      Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322)   -
2                                      http://www.davemease.com/                      Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322)   -
3                                                              -                      Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322)   -
4                                                              - Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.4) Gecko/20070515 Firefox/2.0.0.4   -
5                                      http://www.davemease.com/ Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.8.1.4) Gecko/20070515 Firefox/2.0.0.4   -
```

Reading Data into R

Look at the first column:

```
> data[,1]
```

```
   [1] 69.224.117.122  69.224.117.122  69.224.117.122  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164
```

...

...

...

```
[1901] 65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11
[1911] 65.57.245.11    67.164.82.184   67.164.82.184   67.164.82.184   171.66.214.36   171.66.214.36   171.66.214.36   65.57.245.11    65.57.245.11    65.57.245.11
[1921] 65.57.245.11    65.57.245.11
73 Levels: 128.12.159.131 128.12.159.164 132.79.14.16 171.64.102.169 171.64.102.98 171.66.214.36 196.209.251.3 202.160.180.150 202.160.180.57 ... 89.100.163.185
```

Experimental Vs. Observational Data
(Important but not in book)

- **Experimental data describes data which was collected by someone who exercised strict control over all attributes.**

- **Observational data describes data which was collected with no such controls. Most all data used in data mining is observational data so be careful.**

- **Examples:**

-Diet Coke vs. Weight

-Carbon Dioxide in Atmosphere vs. Earth's Temperature

Types of Attributes:

Qualitative vs. Quantitative (P. 26)

- **Qualitative** (or **Categorical**) attributes represent distinct categories rather than numbers. Mathematical operations such as addition and subtraction do not make sense. Examples:

  eye color, letter grade, IP address, zip code

- **Quantitative** (or **Numeric**) are numbers and can be treated as such. Examples:

  weight, failures per hour, number of TVs, temperature

Types of Attributes (P. 25):

- All Qualitative (or Categorical) attributes are either Nominal or Ordinal.

Nominal = categories with no order
Ordinal = categories with a meaningful order

- All Quantitative (or Numeric) attributes are either Interval or Ratio.

Interval = no “true” zero, division makes no sense
Ratio = true zero exists, division makes sense

Types of Attributes:

- Some examples:
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit, GRE score
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $=$ $\neq$
  - Order: $<$ $>$
  - Addition: $+$ $-$
  - Multiplication: $*$ $/$

  - **Nominal** attribute: distinctness
  - **Ordinal** attribute: distinctness & order
  - **Interval** attribute: distinctness, order & addition
  - **Ratio** attribute: all 4 properties

Discrete vs. Continuous (P. 28)

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables
  - Note: binary attributes are a special case of discrete attributes which have only 2 values

- Continuous Attribute
  - Has real numbers as attribute values
  - Can compute as accurately as instruments allow
  - Examples: temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Discrete vs. Continuous (P. 28)

- **Qualitative (categorical) attributes are always discrete**

- **Quantitative (numeric) attributes can be either discrete or continuous**

In class exercise #3:

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

a) Number of telephones in your house
b) Size of French Fries (Medium or Large or X-Large)
c) Ownership of a cell phone
d) Number of local phone calls you made in a month
e) Length of longest phone call
f) Length of your foot
g) Price of your textbook
h) Zip code
i) Temperature in degrees Fahrenheit
j) Temperature in degrees Celsius
k) Temperature in kelvins

Types of Data in R

- R often distinguishes between qualitative (categorical) attributes and quantitative (numeric)

- In R,

qualitative (categorical) = “factor”

quantitative (numeric) = “numeric”

Types of Data in R

- For example, the IP address in the first column of stats202log.txt is a factor

```
> data[,1]

   [1] 69.224.117.122  69.224.117.122  69.224.117.122  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164  128.12.159.164

                                                        ...

                                                        ...
[1901] 65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11    65.57.245.11
[1911] 65.57.245.11    67.164.82.184   67.164.82.184   67.164.82.184   171.66.214.36   171.66.214.36   171.66.214.36   65.57.245.11    65.57.245.11    65.57.245.11
[1921] 65.57.245.11    65.57.245.11
73 Levels: 128.12.159.131 128.12.159.164 132.79.14.16 171.64.102.169 171.64.102.98 171.66.214.36 196.209.251.3 202.160.180.150 202.160.180.57 ... 89.100.163.185


> is.factor(data[,1])
[1] TRUE


> data[,1]+10
[1] NA NA NA NA NA NA NA NA ...
Warning message:
+ not meaningful for factors ...
```

Types of Data in R

- However, the 8th column looks like it should be numeric. Why is it not? How do we fix this?

```
> data[,8]

   [1] 2867    4583    2295    2867    4583    2295    1379    2294    4432    7134    2296    2297    3219968 1379    2294    4432    7134    2293    2297    2294

                                                          ...

[1901] 2294    4432    7134    2294    4432    7134    2294    2867    4583    2295    2294    4432    7134    2294    4432    7134    2294    2294    2294    2294
[1921] 2294    2294
Levels: - 1135151 122880 1379 1510 2290 2293 2294 2295 2296 2297 2309 238 241 246 248 250 2725487 280535 2867 3072 3219968 4432 4583 626 7134 7482


> is.factor(data[,8])
[1] TRUE
> is.numeric(data[,8])
[1] FALSE
```

Types of Data in R

- A: We should have told R that “-” means missing when we read it in.

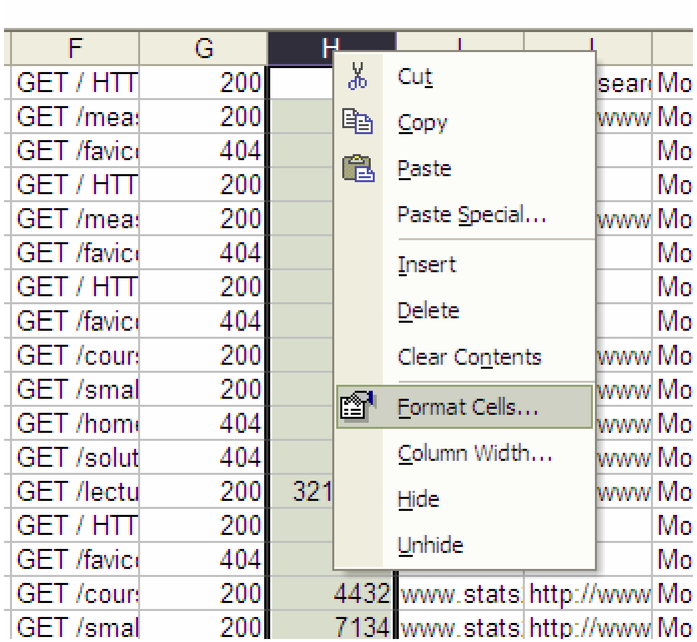
```
> data<-read.csv("stats202log.txt",
     sep=" ",header=F, na.strings = "-")

> is.factor(data[,8])
[1] FALSE
> is.numeric(data[,8])
[1] TRUE
```

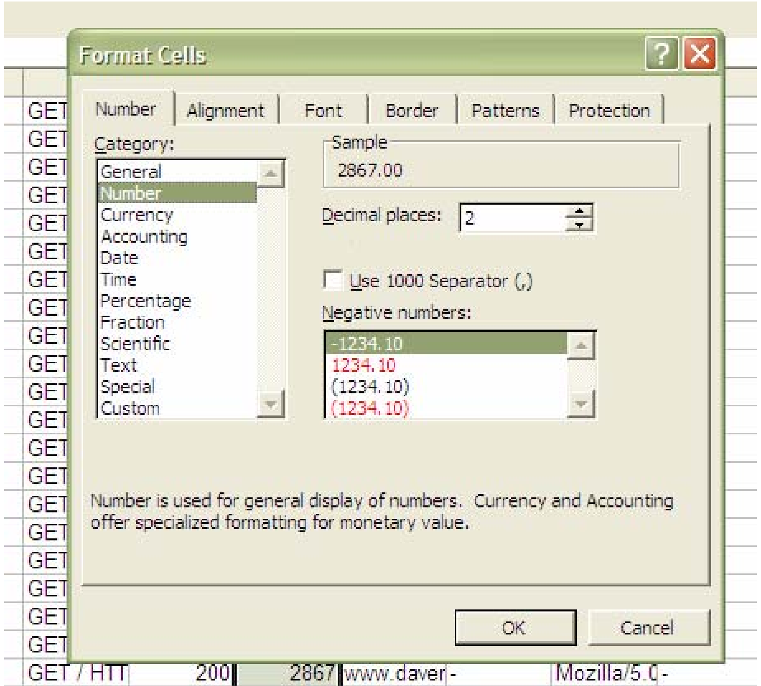
Types of Data in Excel

- Excel is not quite as picky and allows you to mix types more

- Also, you can change between a lot of different predefined formats in Excel by right clicking a column and then selecting “Format Cells”



F	G	H	I	J	
GET / HTT	200			sear	Mo
GET /mea:	200			www	Mo
GET /favic	404				Mo
GET / HTT	200				Mo
GET /mea:	200			www	Mo
GET /favic	404				Mo
GET / HTT	200				Mo
GET /favic	404				Mo
GET /cour:	200			www	Mo
GET /smal	200			www	Mo
GET /hom	404			www	Mo
GET /solut	404			www	Mo
GET /lectu	200	321		www	Mo
GET / HTT	200				Mo
GET /favic	404				Mo
GET /cour:	200	4432	www.stats	http://www	Mo
GET /smal	200	7134	www.stats	http://www	Mo



Category: Number

Sample: 2867.00

Decimal places: 2

□ Use 1000 Separator (,)

Negative numbers: -1234.10

Number is used for general display of numbers. Currency and Accounting offer specialized formatting for monetary value.

GET / HTT	200	2867	www.daver -	Mozilla/5.0 -