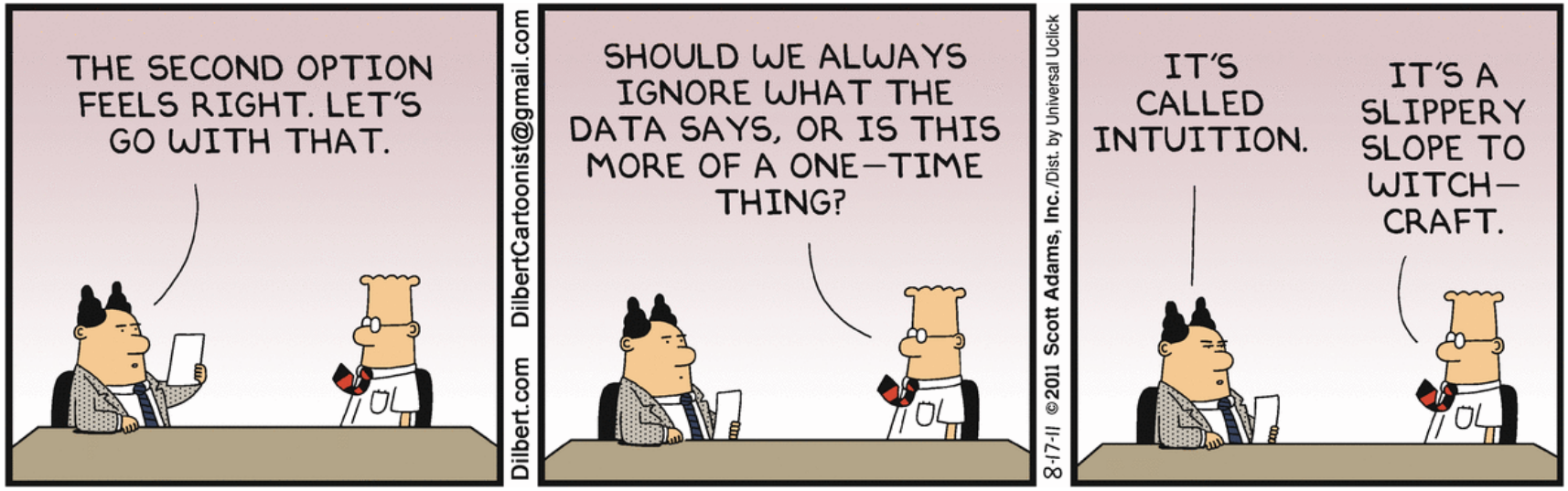
Business Statistics 41000

Lecture 1

Mladen Kolar

Why are you here?



Why are you here?

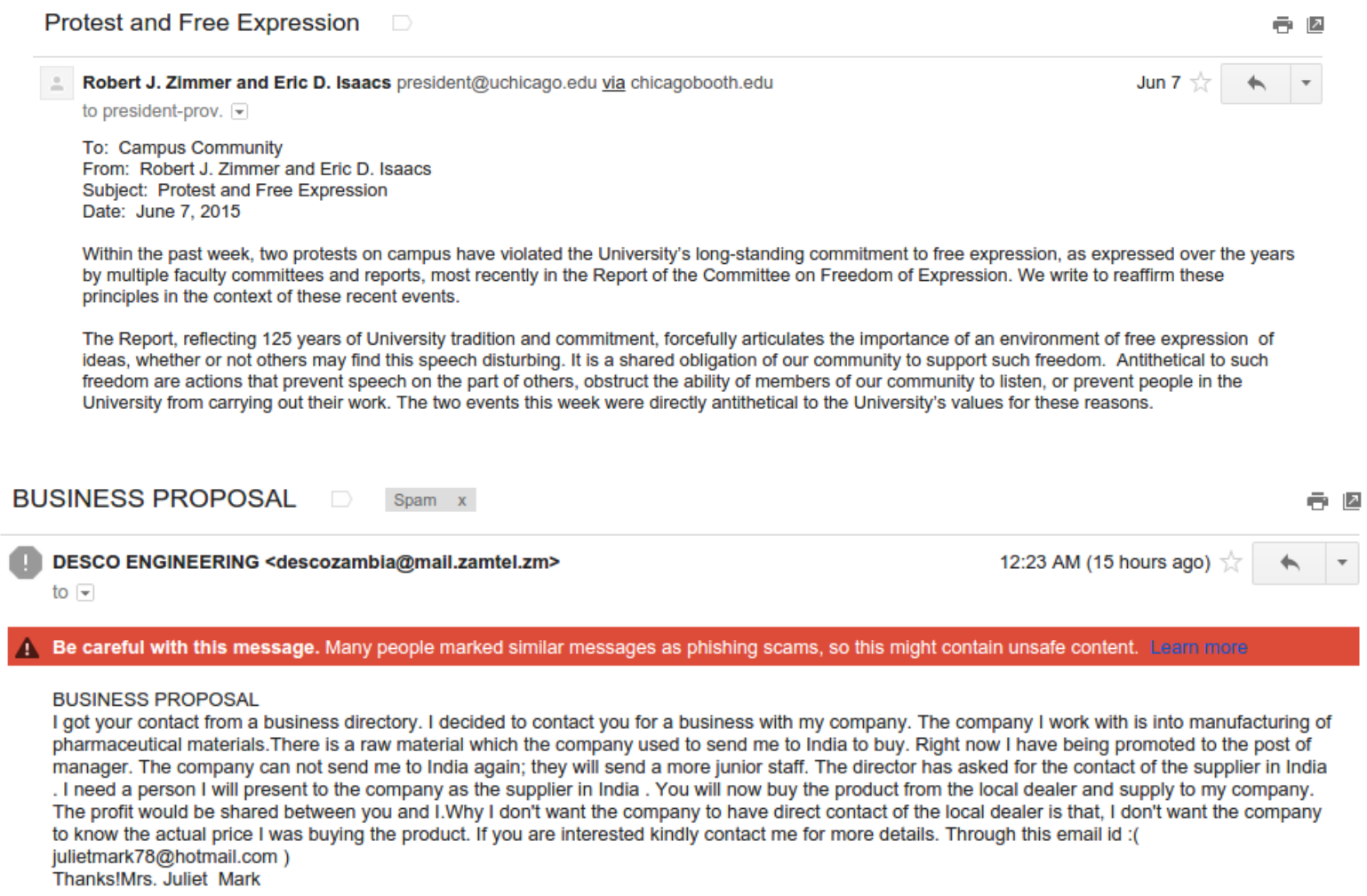
Because

- you are passionate about the subject
- you want to learn about fundations for many transformative technologies of the 21st century
- you eventually want to earn big $$$
- …

No matter the reason, we hope you will get something useful out of the class.

Remember that the more effort you put into the class, the more you will get out.

Spam Filtering



Protest and Free Expression

Robert J. Zimmer and Eric D. Isaacs president@uchicago.edu via chicagobooth.edu Jun 7

to president-prov.

To: Campus Community
From: Robert J. Zimmer and Eric D. Isaacs
Subject: Protest and Free Expression
Date: June 7, 2015

Within the past week, two protests on campus have violated the University's long-standing commitment to free expression, as expressed over the years by multiple faculty committees and reports, most recently in the Report of the Committee on Freedom of Expression. We write to reaffirm these principles in the context of these recent events.

The Report, reflecting 125 years of University tradition and commitment, forcefully articulates the importance of an environment of free expression of ideas, whether or not others may find this speech disturbing. It is a shared obligation of our community to support such freedom. Antithetical to such freedom are actions that prevent speech on the part of others, obstruct the ability of members of our community to listen, or prevent people in the University from carrying out their work. The two events this week were directly antithetical to the University's values for these reasons.

BUSINESS PROPOSAL Spam x

DESCO ENGINEERING <descozambia@mail.zamtel.zm> 12:23 AM (15 hours ago)

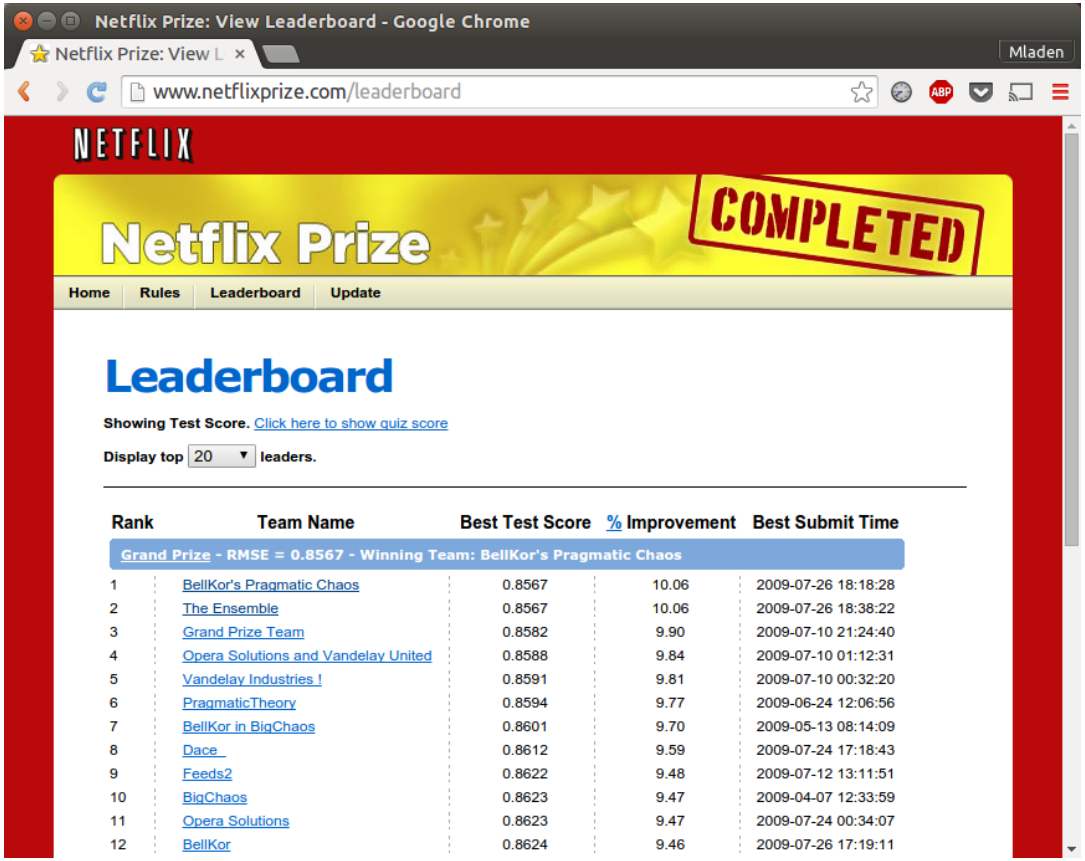
to

Be careful with this message. Many people marked similar messages as phishing scams, so this might contain unsafe content. Learn more

BUSINESS PROPOSAL
I got your contact from a business directory. I decided to contact you for a business with my company. The company I work with is into manufacturing of pharmaceutical materials.There is a raw material which the company used to send me to India to buy. Right now I have being promoted to the post of manager. The company can not send me to India again; they will send a more junior staff. The director has asked for the contact of the supplier in India . I need a person I will present to the company as the supplier in India . You will now buy the product from the local dealer and supply to my company. The profit would be shared between you and I.Why I don't want the company to have direct contact of the local dealer is that, I don't want the company to know the actual price I was buying the product. If you are interested kindly contact me for more details. Through this email id :(julietmark78@hotmail.com)
Thanks!Mrs. Juliet Mark

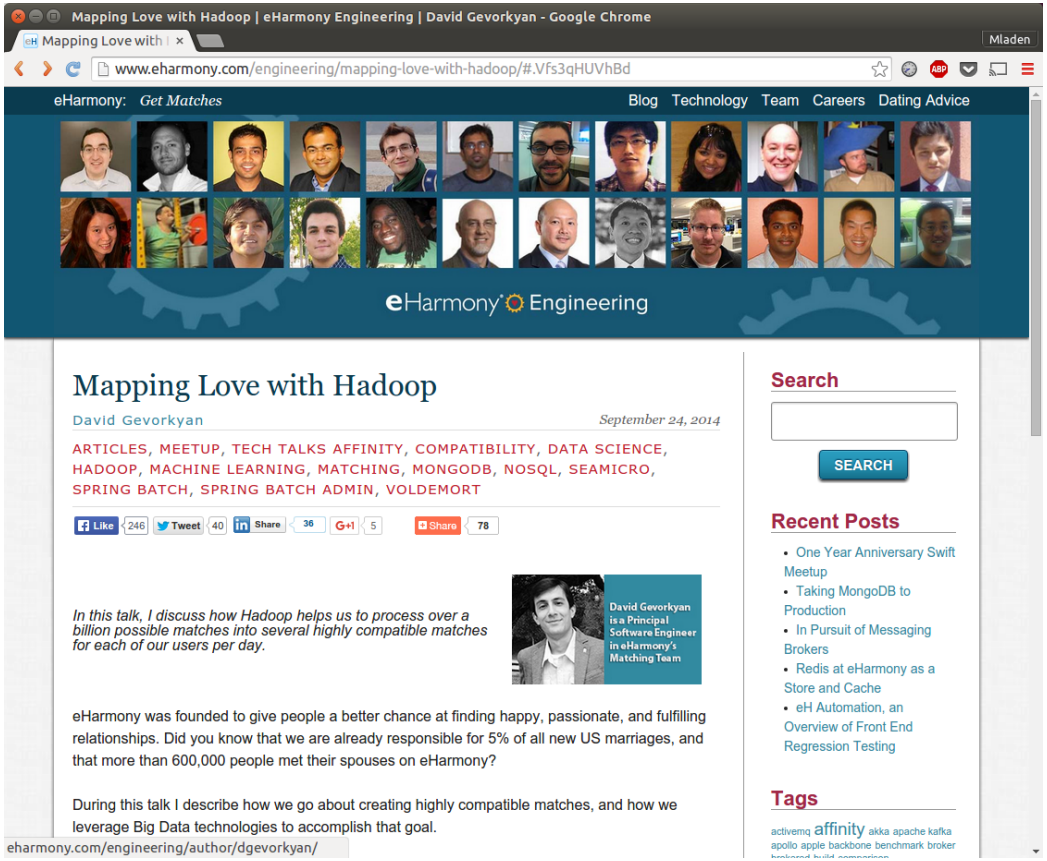
Spam or Ham

Personal recommendation



$1M prize!

Personal recommendation



Fall in love with statistics

Transforming Healthcare

Heritage Health Prize

What is statistics?

Data from real world

- lots of data
- we need to analyze for insights

Models capture uncertainty

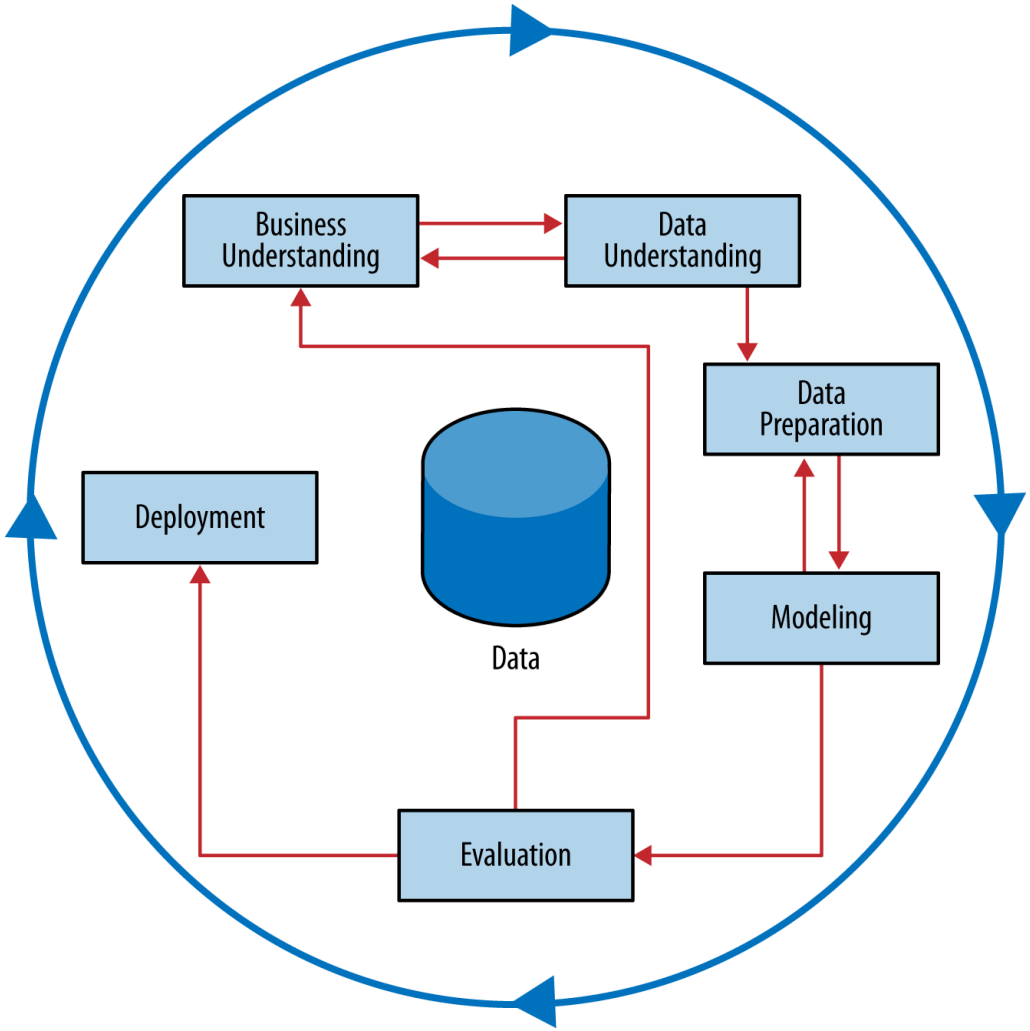
- explain the real world
- predict outcomes
- test competing ideas

http://www.economist.com/node/15579717

Statistics is the link between the models and the data.

- For example, which model is the "best" model?

Crisp DM diagram



Course structure

We begin with a discussion of summarizing real world data (means, variances, covariances, …).

We next introduce models. Since the real world is uncertain our models will involve uncertainty/probability.

We then combine the data with the models to discuss estimation and testing. We begin with very simple models and move to more complex models.

Outline of today's topics

Types of data

Descriptive Statistics

- Tables
- Histogram
- The Time Series Plot
- The Mean and Median
- The Sample Variance and Standard Deviation
- The Empirical Rule
- Covariance and Correlation

Statistical methodology

Common steps in statistical analysis

1. Formulate the problem
2. Collect data
3. Visualize the data (*NOTE: We will start here*)
4. Perform the necessary statistical calculations
5. Interpret the results

Here is some data (our sample):

```
##   age sex soc edu Reg inc colas restrntEve juice cigs antique news
## 1  67   2   3   1   3  12     1          0     1    0       1    0
## 2  51   2   3   8   3  10     1          1     0    1       1    0
## 3  63   2   3   1   2  13     1          1     0    1       1    0
## 4  45   2   4   3   1  18     1          1     1    0       1    0
## 5  52   2   5   2   8  13     0          1     1    0       1    0
## 6  58   2   1   3   9  17     0          0     1    1       1    1
```

The data is from a large survey carried out by a marketing research company in Britain (see `marketing.csv`).

Each row corresponds to a household.

Each column describes a different characteristic of that household.

The characteristics are called **variables**.

The rows are called **observations**.

Types of data

Variables are the fundamental objects in statistics. They come in different types

- **Numeric** versus **categorical** data.
- You can separate these by asking whether the data is qualitative versus quantitative.
- If it is categorical data, it may be **ordered categorical**. In other words, the order of the categories is meaningful.
- If it is numeric data, it may be **discrete** or **continuous**.

Numeric data

In the marketing data set, the variable "age" is simply the age (in years) of the responder.

This is what we call a **numeric** variable.

- The variable has units.
- Taking an average is meaningful.

For example, using the British marketing data, we can compute the average age of respondents or we can compute the difference in ages between two respondents.

Numeric data

Numeric data can also be split up into two types:

- discrete
- continuous

Examples:

- The amount of snow in Chicago in February is usually thought of as continuous.
- We may think of the variable age as continuous even though age is measured in years and we could list the possible values.
- Number of children is thought of as a discrete variable.

Categorical data

A variable like "Reg" is a **categorical variable**. It describes in which geographic region of Britain the household is located. Each number is really just a code for a region.

1	"Scotland"
2	"North West"
3	"North"
4	"Yorkshire & Humberside"
5	"East Midlands"
6	"East Anglia"
7	"South East"
8	"Greater London"
9	"South West"
10	"Wales"
11	"West Midlands"

Instead of using numbers we could have used text strings in the data file.

Ordered categorical data

The variable "soc" stands for socioeconomic status and there are six categories.

soc	Meaning
1	Higher managerial, administrative and professional
2	Intermediate managerial, administrative and professional
3	Supervisory, clerical and junior managerial, administrative and professional
4	Skilled manual workers
5	Semi-skilled and unskilled manual workers
6	State pensioners, casual and lowest grade workers, unemployed with state benefits only

Ordered categorical data

The variable "soc" is categorical because it takes on codes 1 to 6. Each code represents a different socioeconomic status.

It is ordered categorical data because 1 indicates the highest social grade, while 6 indicates the lowest social grade

Variable "soc" is ordered like "age" but does not have units.

It does not make sense to average "soc" or compute differences in "soc"" whereas it does with a numeric variable like "age".

Types of data: more examples

Let's continue looking at the marketing data set and some of the definitions of the data:

"age": measured in years
"sex": 1 = male; 2 = female
"edu": education, terminal age of education

edu	Terminal age of education
1	14 or under
2	15
3	16
4	17
5	18
6	19
7	20
8	21 to 23
9	24 and older

Types of data: more examples

"Reg": region of Britain
"inc": total family income before tax measured in British pounds

```
>      inc       Income before tax       Inc       Income before tax
>
>       1         1,999 or less          11        11,000 - 11,999
>       2         2,000 - 2,999          12        12,000 - 14,999
>       3         3,000 - 3,999          13        15,000 - 19,999
>       4         4,000 - 4,999          14        20,000 - 24,999
>       5         5,000 - 5,999          15        25,000 - 29,999
>       6         6,000 - 6,999          16        30,000 - 34,999
>       7         7,000 - 7,999          17        35,000 - 39,999
>       8         8,000 - 8,999          18        40,000 - 49,999
>       9         9,000 - 9,999          19        50,000 or over
>       10       10,000 - 10,999         20        Not stated
```

Both "edu" and "inc" could have been numeric but are broken into ranges and are therefore ordered categorical.

Types of data: more examples

"cola": 1/0 — do (do not) buy cola
"restE": 1/0 — do (do not) dine out at restaurants in the evening
"juice": 1/0 — do (do not) buy juice
"cigs": 1/0 — do (do not) buy cigarettes

Dummy variables
- categorical variables that take two values, 0 or 1.

Very commonly used to indicate that something:

- did happen — 1, or
- did not happen — 0

Types of data: more examples

```
##   age sex soc cigs antique news enders friends simpsons football
## 1  67   2   3    0       1    0      0       0        0        0
## 2  51   2   3    1       1    0      1       1        0        0
## 3  63   2   3    1       1    0      1       0        0        0
## 4  45   2   4    0       1    0      0       0        0        0
## 5  52   2   5    0       1    0      1       1        0        0
## 6  58   2   1    1       1    1      0       0        0        0
```

"enders": 1/0 — do (do not) watch East Ender
"friend": 1/0 — do (do not) watch Friends
"antiq": 1/0 — do (do not) watch Antiques Roadshow
"news": 1/0 — do (do not) watch BBC news
"simp": 1/0 — do (do not) watch Simpsons
"foot": 1/0 — do (do not) watch football

British marketing data

Our data has information on

- Demographics: age and income
- Product category usage
- Media exposure (tv shows)

What is the point? Why collect this data?

- We want to see how product usage relates to demographics.
- Who drinks cola?
- When should a firm like Pepsi or Coca-cola advertise?

Summary: types of data

Data comes in one of three types

- numeric
- categorical
- ordered categorical

Numeric variables are either discrete or continuous.

Categorical variables often take on the value 0 or 1 to indicate that something did or did not happen.

It is meaningful to perform mathematical operations for some types of data but not others.

Exploring data

Visualization

We want to get a sense of what our data look like.
We start by investigating each variables on its own.
We will consider several types of viewpoints:

1. Tables and bargraphs
2. Histograms
3. Time series plots

Later, we will investigate how variables relate to one another.

People who watch Friends tend to drink cola. Smokers tend to get cancer.

Loading data in R

A basic component of working with data is knowing your working directory

- The two main commands are `getwd()` and `setwd()` (see section 2.4)

```
setwd("/home/mkolar/Downloads")
```

```
marketing_df = read.csv("marketing.csv")
dim(marketing_df)
```

```
## [1] 1000   16
```

Tables

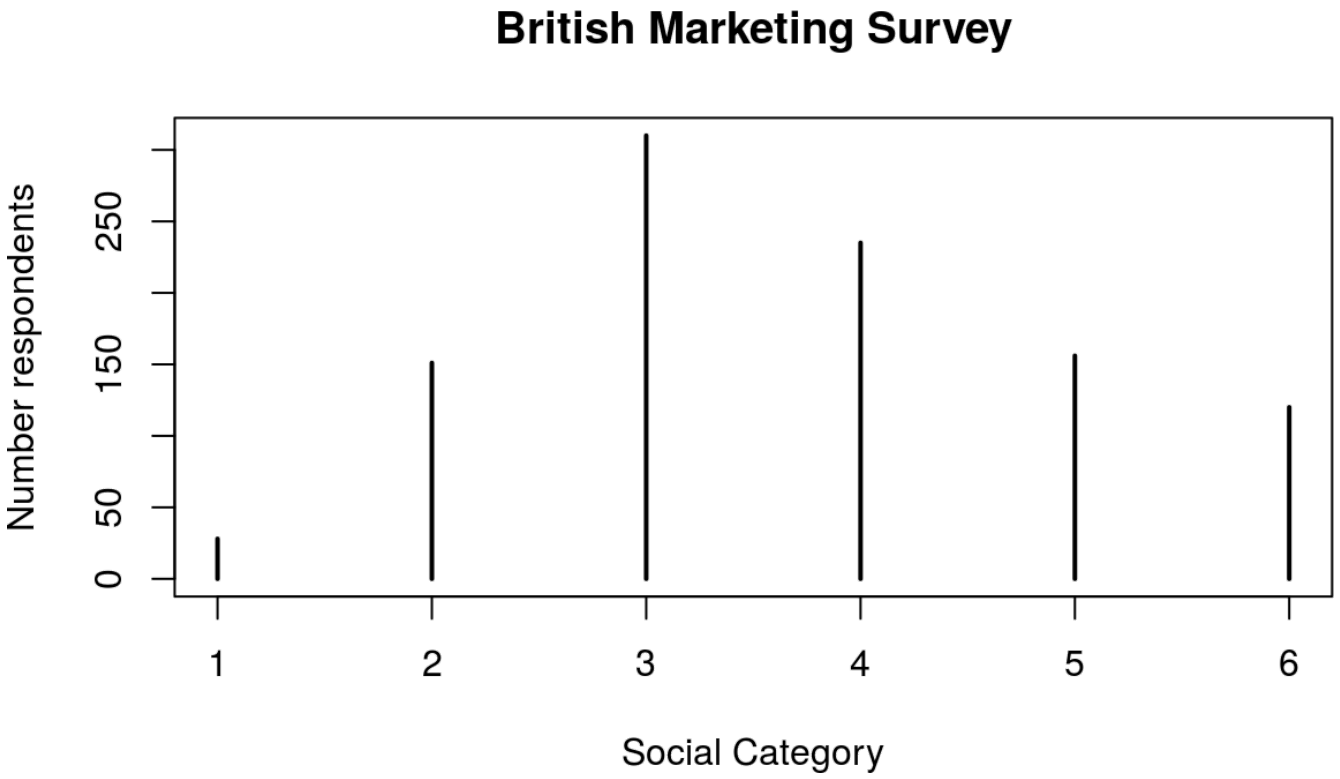
Consider the variable "soc" from the marketing data set. It had six categories. To view a categorical variable, we can create the following table:

```
soc_table = table(marketing_df$soc)
soc_table
```

```
##
##   1   2   3   4   5   6
##  28 151 310 235 156 120
```

It summarizes the number of households in each category.
(In our case, n = 1000 households.)

```
plot(soc_table, xlab='Social Category', ylab='Number respondents',
                main='British Marketing Survey')
```

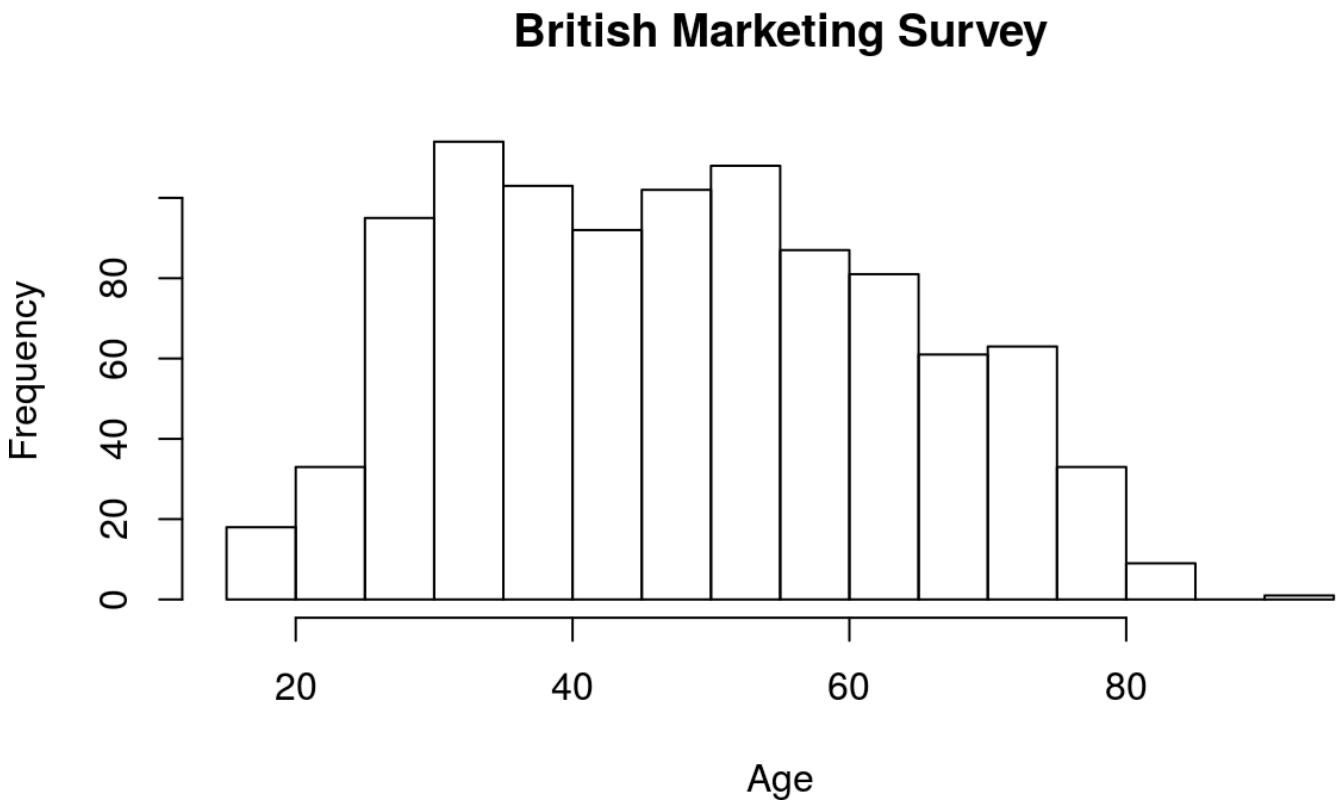


The height of each bar equals the number of observations in that category.

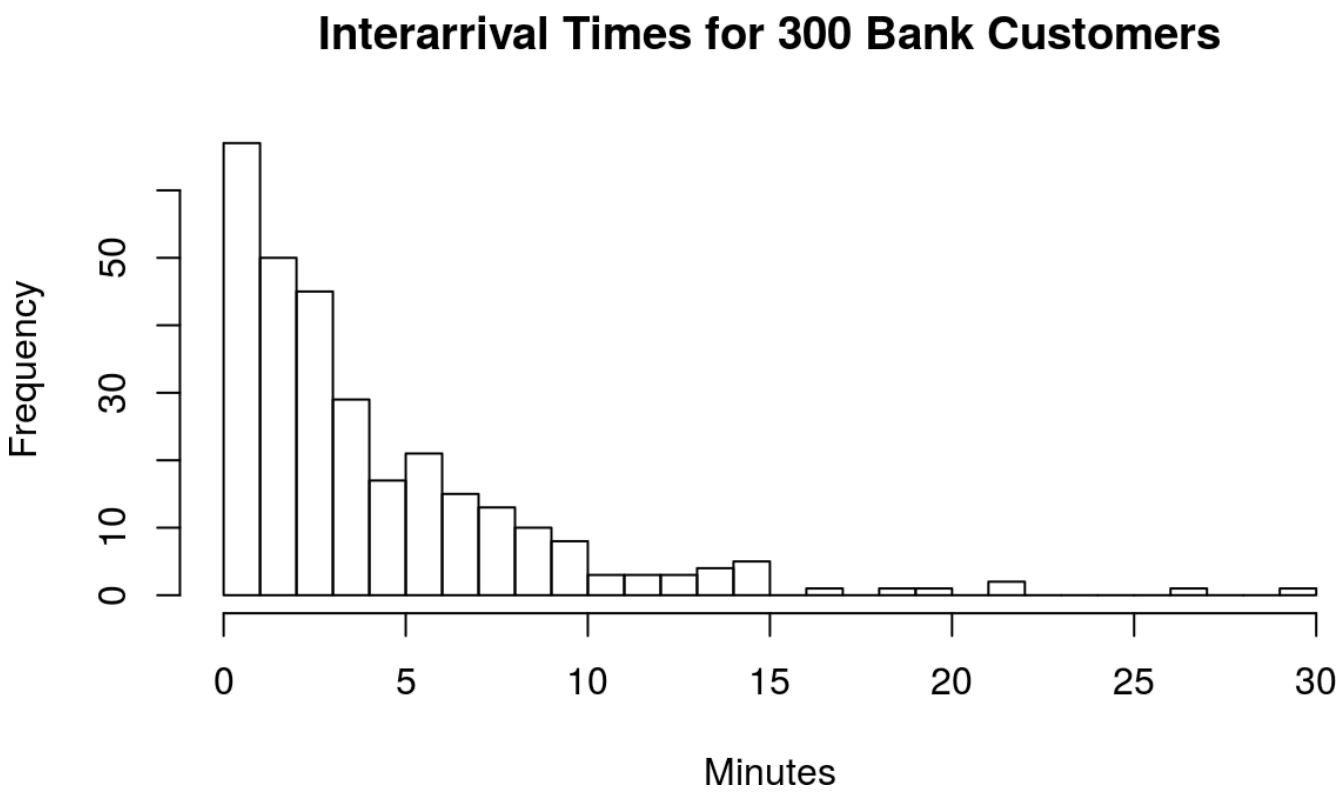
Histograms

In a **histogram**, we take a numeric variable and break it into *bins* and then plot the number in each bin.

```
hist(marketing_df$age, xlab='Age', main='British Marketing Survey')
```



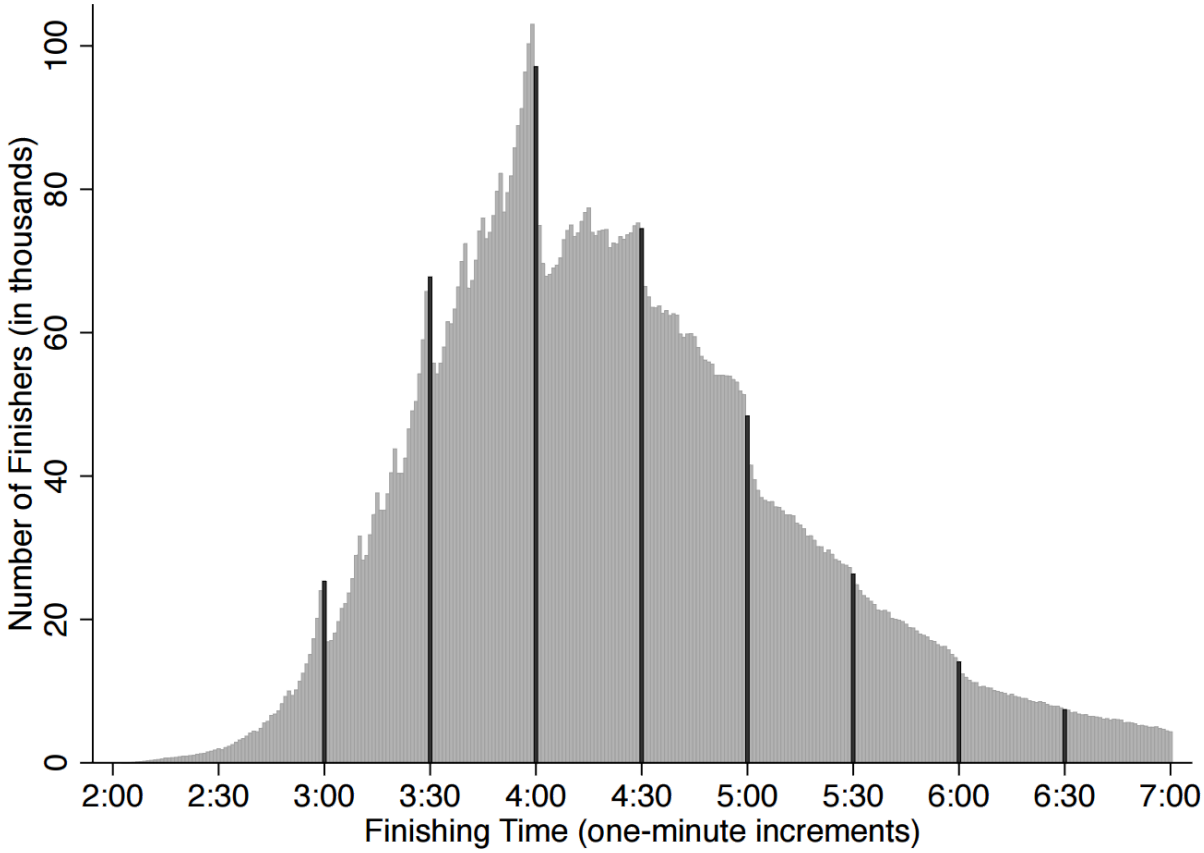
```
bank_df = read.csv("bank.csv")
hist(bank_df[,1], breaks = 25,
     xlab='Minutes', main='Interarrival Times for 300 Bank Customers')
```



These data have a "heavy right tail" which is called **right-skewed**.

Example: Marathon finishing times

Sample size: $n = 9,789,093$. The dark bars highlight the density in the minute bin just prior to each 30 minute threshold.



Histograms

What is the effect of the number of bins on a histogram?

Demo: https://mlakolar.shinyapps.io/simpleHistogram/

Histograms: Summary

A histogram gives us a smooth picture of the data.

The height of each bar tells us how many observations are in the corresponding interval.

The number of bins determines the degree of resolution.

- A small number of bins, that is, each bin has a large width, gives us a very smooth (flat) histogram with no detail.
- A large number of bins, that is, each bin has a small width, gives us detailed histogram, maybe even too many details.

There is no special way of determining the bin width.
I use my eye.

Plotting time series data

Time series data

The marketing survey data is what economists call *cross-sectional* data. At a particular point in time, the households represent a **sample** of all British households.

In cross-sectional data, order does not matter.
Later in the course, we will formalize this in terms of independence.

With **time series** data, each observation corresponds to a point in time. Order matters.

Time series data

Consider the Standard & Poors 500 Index

```
gspc_df = read.csv("GSPC.csv")
head(gspc_df, n = 4)
```

```
##       Date   Open   High    Low  Close     Volume Adj.Close
## 1 1/3/2002 1154.7 1165.3 1154.0 1165.3 1398900000    1165.3
## 2 1/4/2002 1165.3 1176.6 1163.4 1172.5 1513000000    1172.5
## 3 1/7/2002 1172.5 1177.0 1163.6 1164.9 1308300000    1164.9
## 4 1/8/2002 1164.9 1167.6 1157.5 1160.7 1258800000    1160.7
```

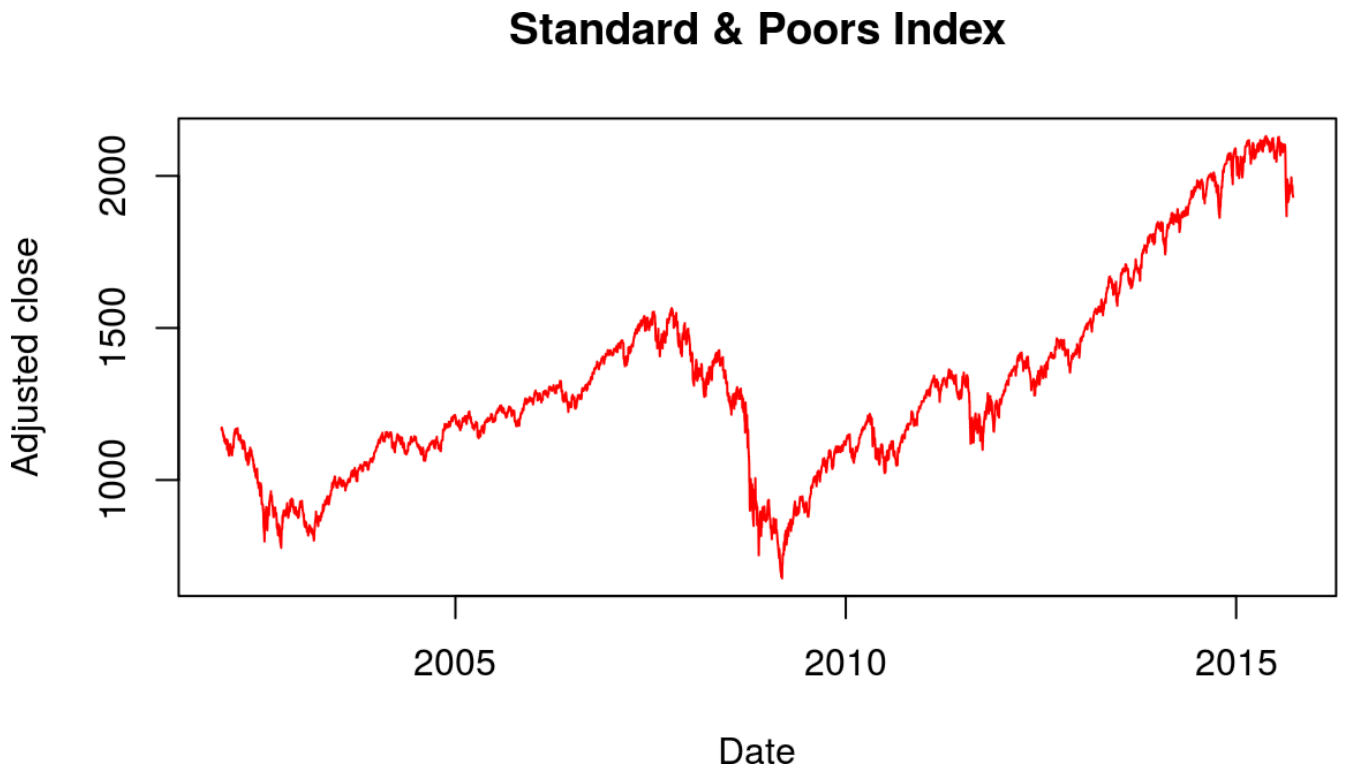
Clearly, order matters since 3-Jan-02 comes before 4-Jan-02.

Specifically, what we observe on 4-Jan-02 depends on what happened the day before.

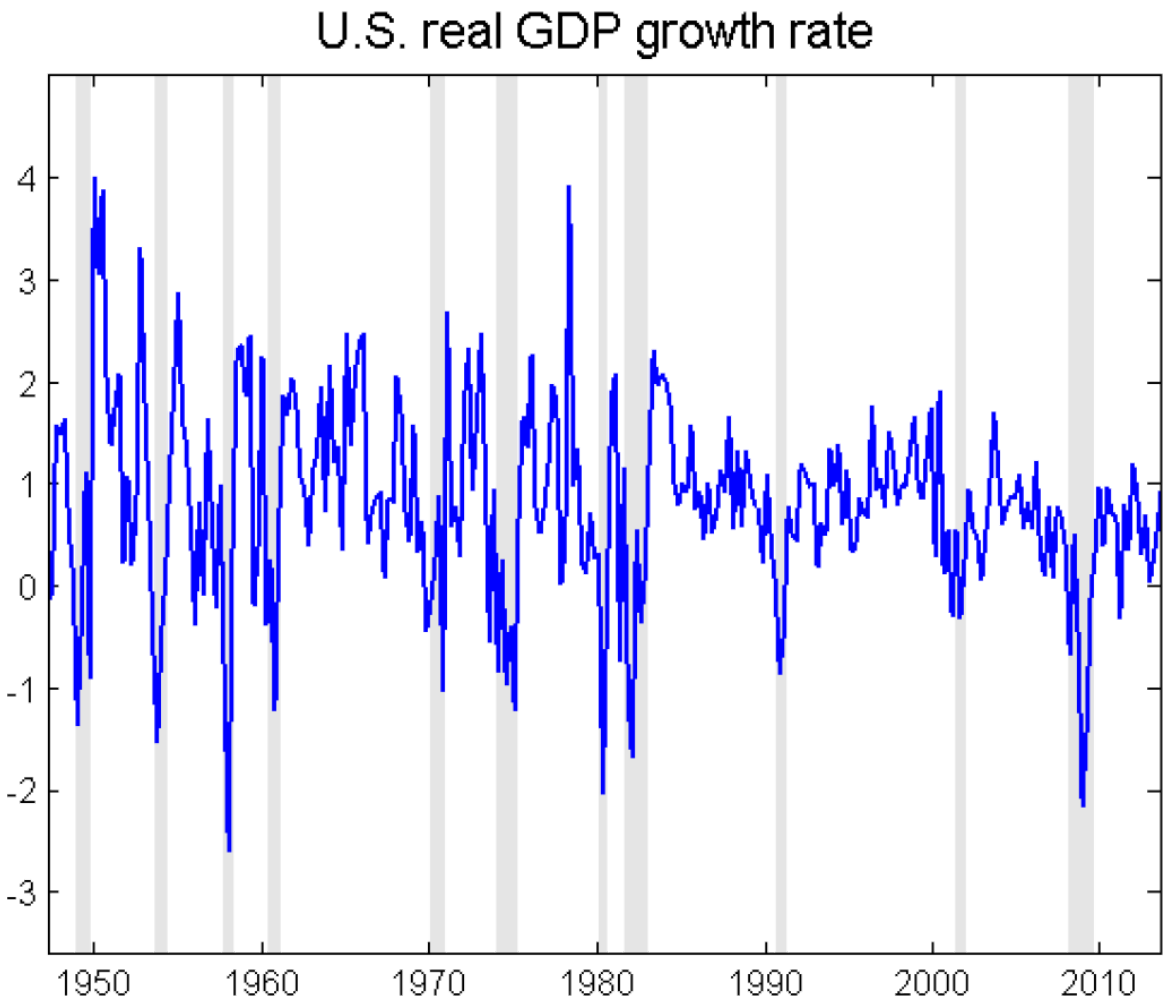
In a time series plot, we plot the data against time.

Time series plot of the adjusted close

```
plot(x=as.Date(gspc_df$Date, format = "%m/%d/%Y"),
     y=gspc_df$Adj.Close, type="l", col="red",
     xlab="Date", ylab="Adjusted close", main = "Standard & Poors Index")
```



Time Series Plot Quarterly real GDP growth rate for the United States from Q2 1947 to Q4 2015.



Recessions are labeled in gray.

Plotting simple returns

The return on an asset is the percentage increase in wealth invested in the asset over a given time period.

If you invest B at the beginning of the time period you get

$$E = B + r \cdot B = (1 + r) \cdot B$$

at the end of the time period, where r is the return.

Clearly, given E and B we can calculate the return r:

$$r = \frac{E - B}{B}$$

The Canadian Returns Data

Here are 107 monthly returns on a broadbased portfolio of Canadian assets (more on portfolios later).

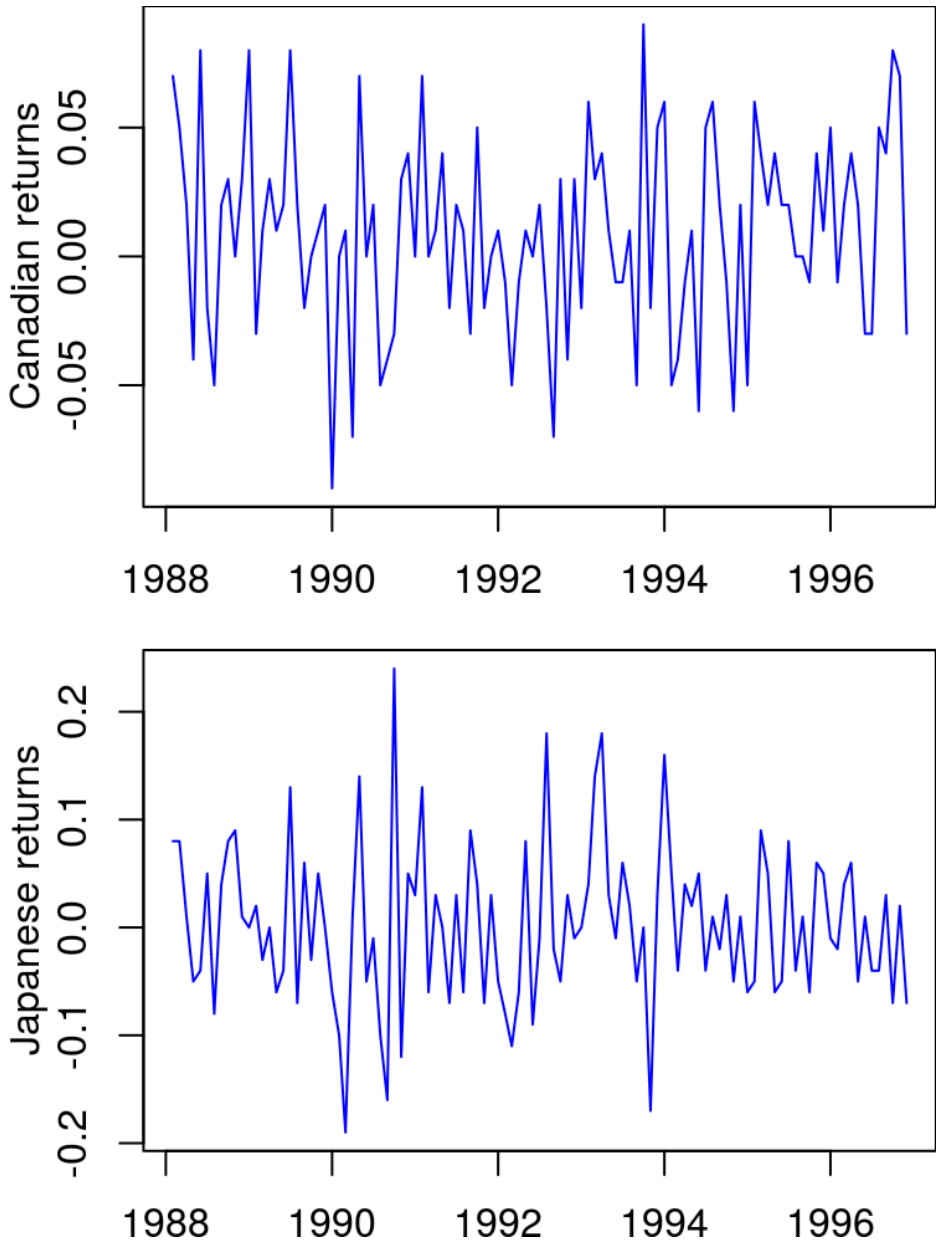
```
countryReturn_df = read.csv("CountryMonthlyReturns.csv")
countryReturn_df$canada
```

```
##   [1]  0.07  0.05  0.02 -0.04  0.08 -0.02 -0.05  0.02  0.03  0.00  0.03
##  [12]  0.08 -0.03  0.01  0.03  0.01  0.02  0.08  0.02 -0.02  0.00  0.01
##  [23]  0.02 -0.09  0.00  0.01 -0.07  0.07  0.00  0.02 -0.05 -0.04 -0.03
##  [34]  0.03  0.04  0.00  0.07  0.00  0.01  0.04 -0.02  0.02  0.01 -0.03
##  [45]  0.05 -0.02  0.00  0.01 -0.01 -0.05 -0.01  0.01  0.00  0.02 -0.02
##  [56] -0.07  0.03 -0.04  0.03 -0.02  0.06  0.03  0.04  0.01 -0.01 -0.01
##  [67]  0.01 -0.05  0.09 -0.02  0.05  0.06 -0.05 -0.04 -0.01  0.01 -0.06
##  [78]  0.05  0.06  0.02 -0.01 -0.06  0.02 -0.05  0.06  0.04  0.02  0.04
##  [89]  0.02  0.02  0.00  0.00 -0.01  0.04  0.01  0.05 -0.01  0.02  0.04
## [100]  0.02 -0.03 -0.03  0.05  0.04  0.08  0.07 -0.03
```

Each number corresponds to a month. They are given in time order (go across rows first). Our first observation is 0.07.

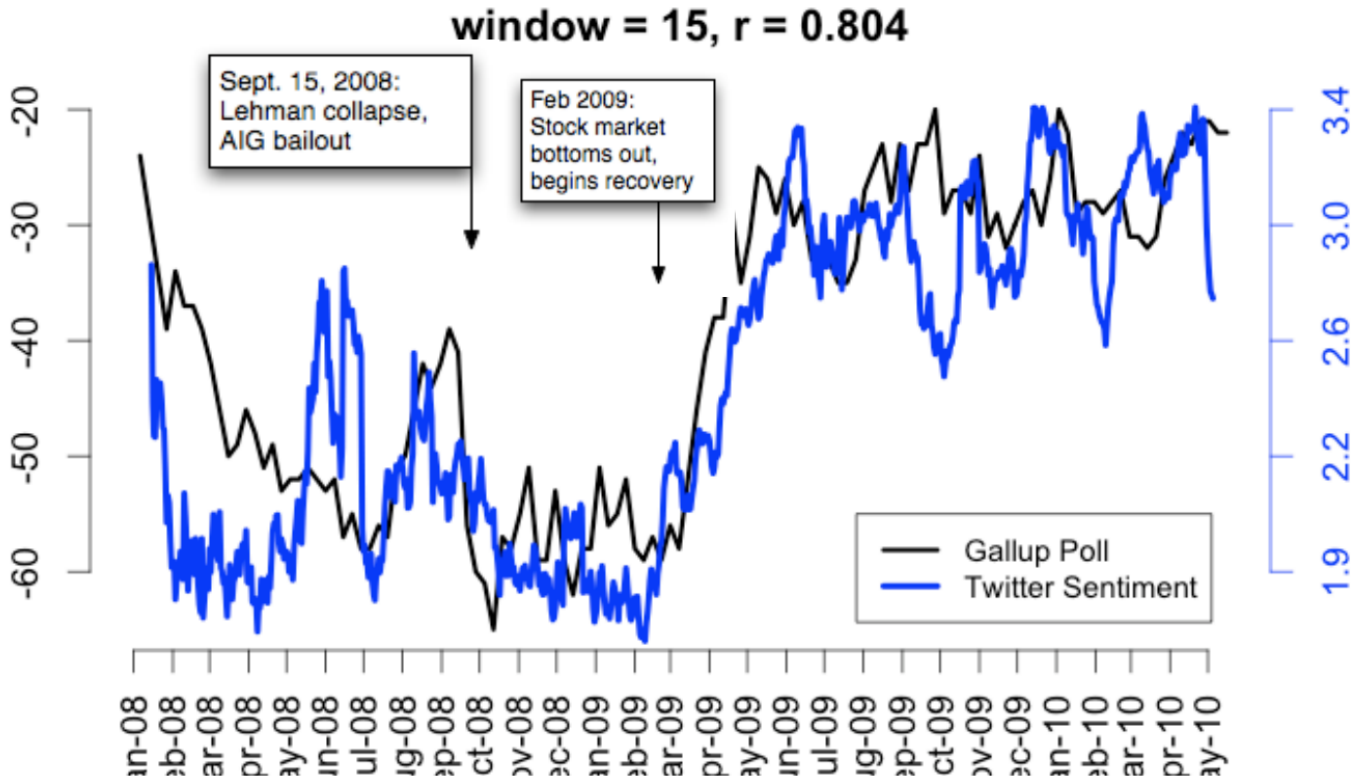
In the first month, the return was 0.07, while in the 11th, it was 0.03.

Monthly returns on a portfolio of Canadian and Japanese stocks.
How do these returns compare?



Plotting time series plots together

Twitter sentiment versus Gallup Poll of Consumer Confidence



Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010.

From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSMP2010

Other time series plots examples

Wikipedia Page Views Statistics: Plots Wikipedia article traffic statistics.

Google trends: This shows you the ups-and-downs of the public's interest in a particular topic using how often we search for it.

Google correlate: Reverse engineers the problem!

Summary statistics

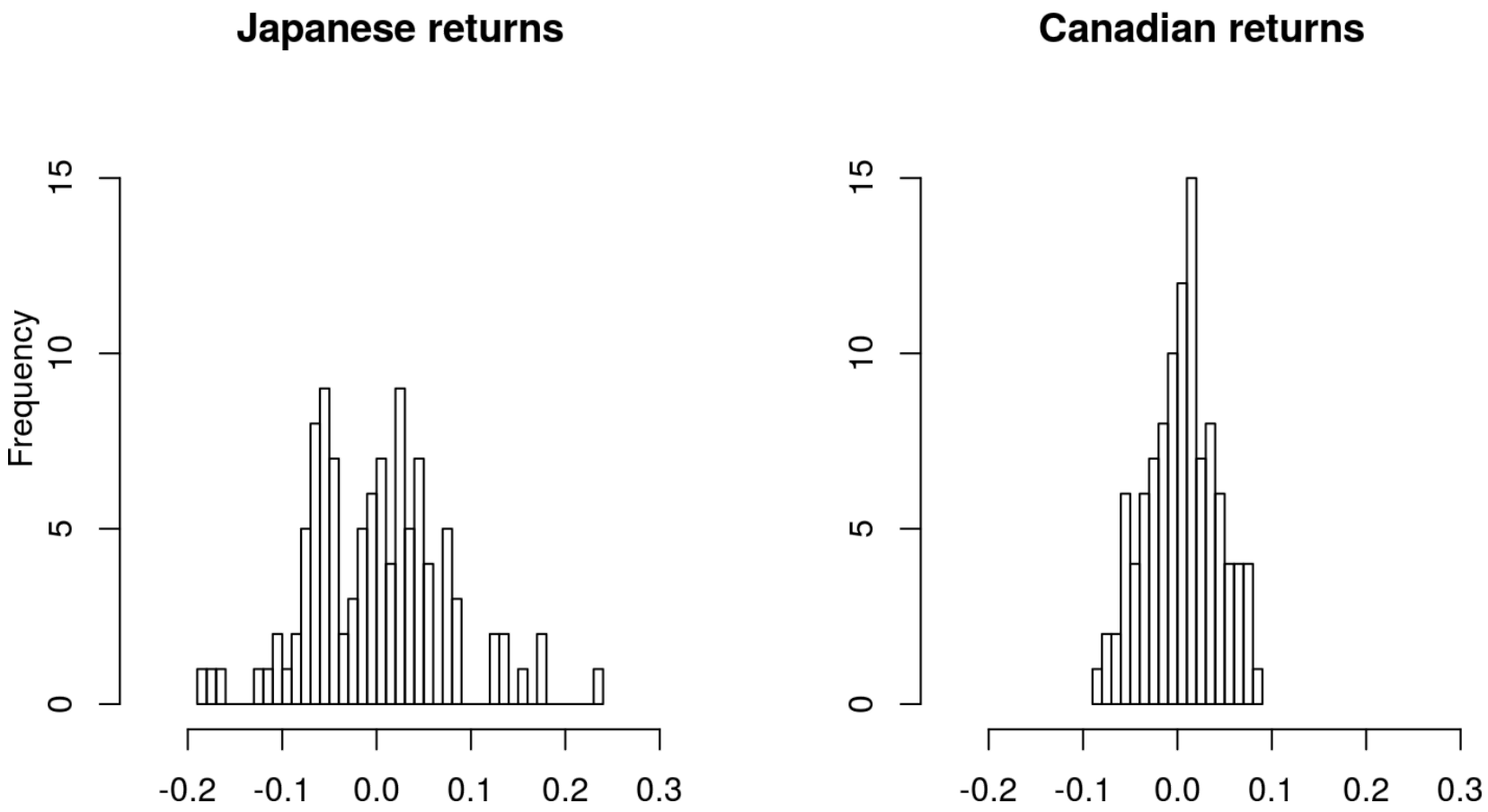
Summarizing a Single Numeric Variable

We have looked at graphs. Suppose we are now interested in having numerical summaries of the data rather than graphical representations.

Two important features of any numeric variable are:

1. What is a typical or average value?
2. How spread out or ‘variable’ are the values?

Monthly returns on Canadian portfolio and Japanese portfolio.
The x-axis has the same scale!



They seem to be centered roughly at the same place but Japan has more spread.
How can we summarize this?

Measures of central tendency

The following descriptive statistics capture central tendency of data

- **sample mean**
- **sample median**

Let us compare the means of the Canadian and Japanese returns.

```
mean(countryReturn_df$canada)
```

```
## [1] 0.0090654
```

```
mean(countryReturn_df$japan)
```

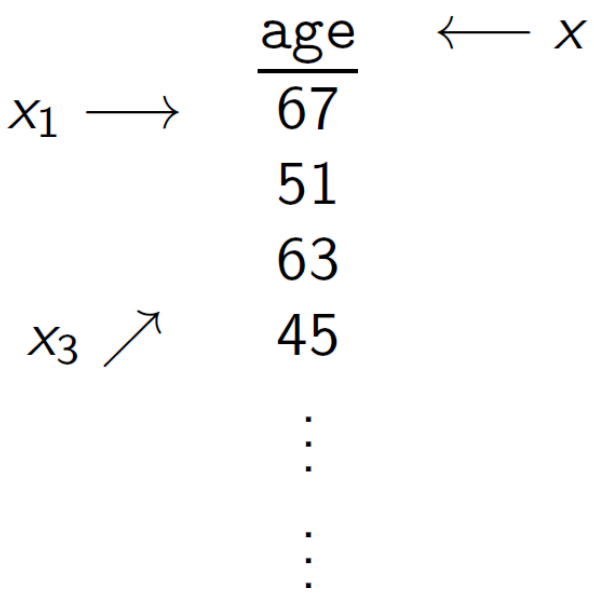
```
## [1] 0.0023364
```

This is a big difference.

It was hard to see this difference in the histogram because the difference is small compared to the variation.

Sample mean

Our data consists of a bunch of numbers.
It is common to denote a general set of numbers as: $x_1, x_2, x_3, \ldots, x_n$.
The number of numbers n is often called the sample size.



There is nothing special about using the letter x we could just as easily call them y or z.

Remember that sometimes the order of the observations has a meaning.
For example, in a time series such as return data.

In other cases such as the British survey data, order does not matter.

Sample mean

The sample mean is the average of the numbers "x":

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The symbol $\bar{x}$ is called **x bar** and is commonly used to denote the mean of the numbers.

Simple Example

```
x = marketing_df$age[1:5]
x
```

```
## [1] 67 51 63 45 52
```

```
x[2]
```

```
## [1] 51
```

```
x[5]
```

```
## [1] 52
```

```
n = length(x)
n
```

```
## [1] 5
```

```
mean(x)
```

```
## [1] 55.6
```

$$
\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{5}\sum_{i=1}^{5} x_i \\
&= \frac{1}{5}(67 + 51 + 63 + 45 + 52) \\
&= 56.6
\end{aligned}
$$

Sample median

The median is another measure of central tendency.

After ordering the values from smallest to largest, the **sample median** is the middle value of the data.

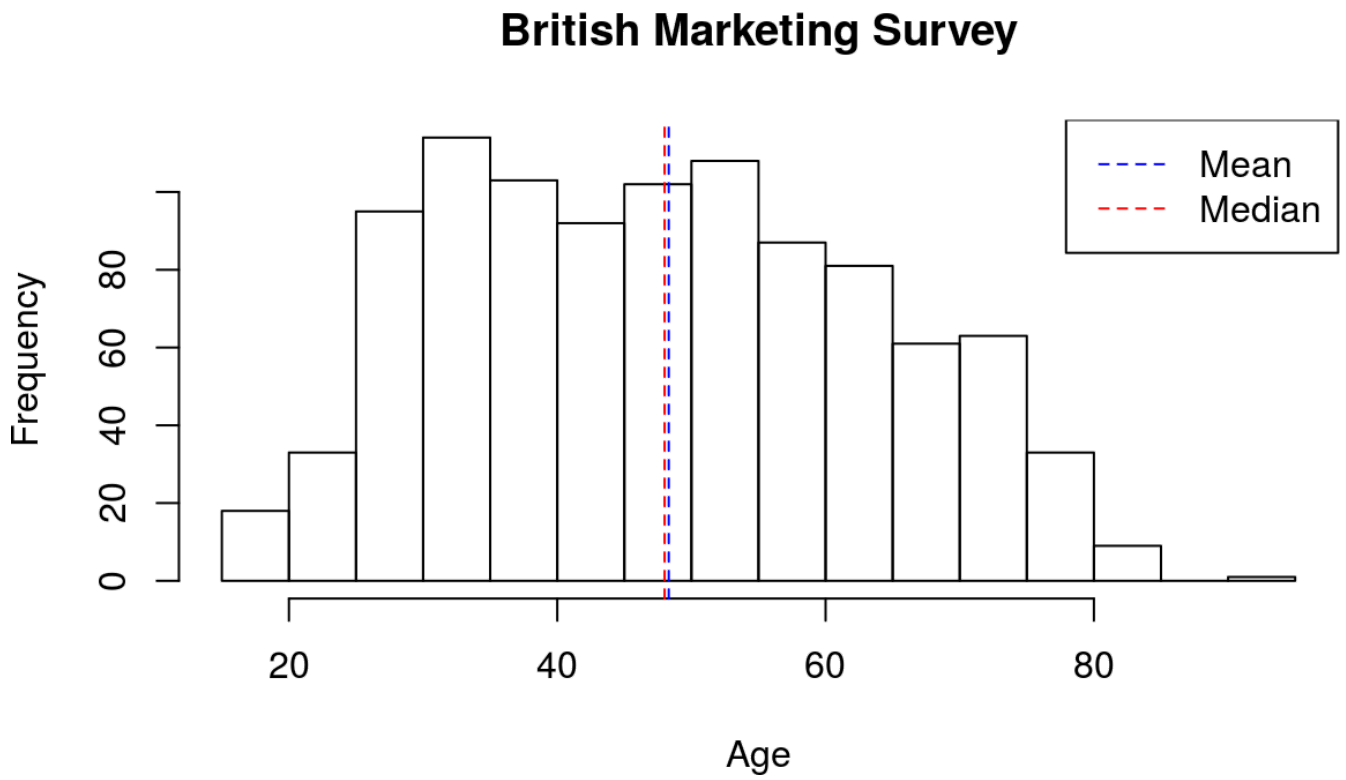
If n is even, then there is no single median value. It is common to define the sample median as the average of the two middle values.

Example 1
Data: 1 4 7 8 10
Median: 7

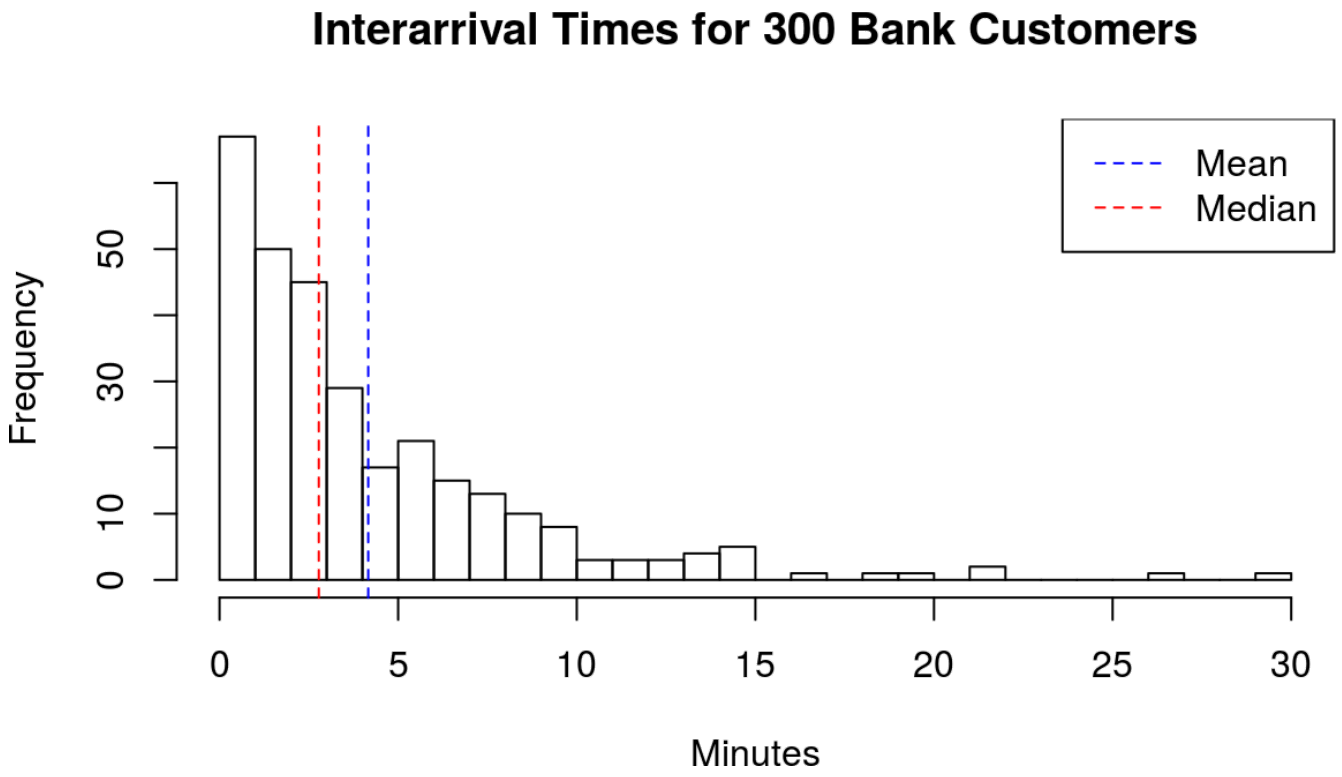
Example 2
Data: 1 4 5 8
Median: 4.5

Mean vs Median

```
hist(marketing_df$age, xlab='Age', main='British Marketing Survey')
abline(v = mean(marketing_df$age), col="blue", lty=2)
abline(v = median(marketing_df$age), col="red", lty=2)
legend(x = "topright", c("Mean", "Median"), col = c("blue", "red"), lty = c(2, 2))
```



```
bank_df = read.csv("bank.csv")
hist(bank_df[,1], breaks = 25,
     xlab='Minutes', main='Interarrival Times for 300 Bank Customers')
abline(v = mean(bank_df[,1]), col="blue", lty=2)
abline(v = median(bank_df[,1]), col="red", lty=2)
legend(x = "topright", c("Mean", "Median"), col = c("blue", "red"), lty = c(2, 2))
```



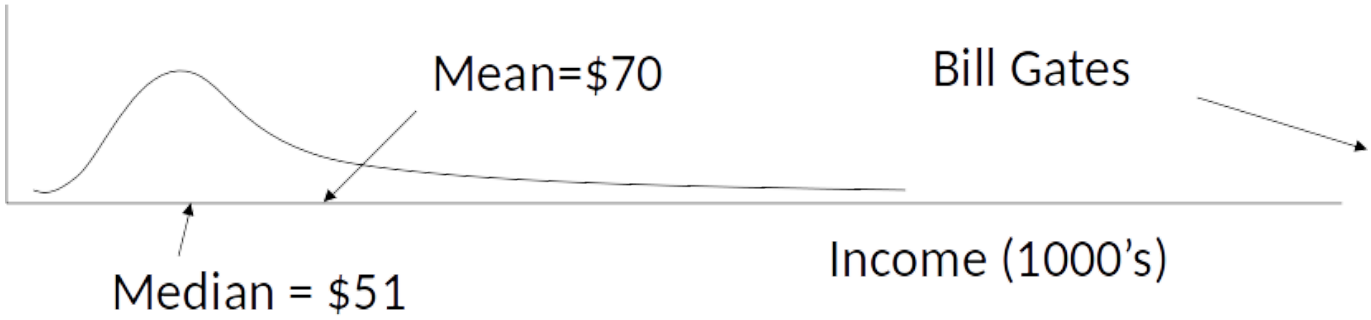
When the mean is bigger than the median, the data are right-skewed. Large values are “pulling” the mean to the right.

Mean vs Median

Although both the mean and median are good measures of the "center" of the distribution, the median is less sensitive to outliers.

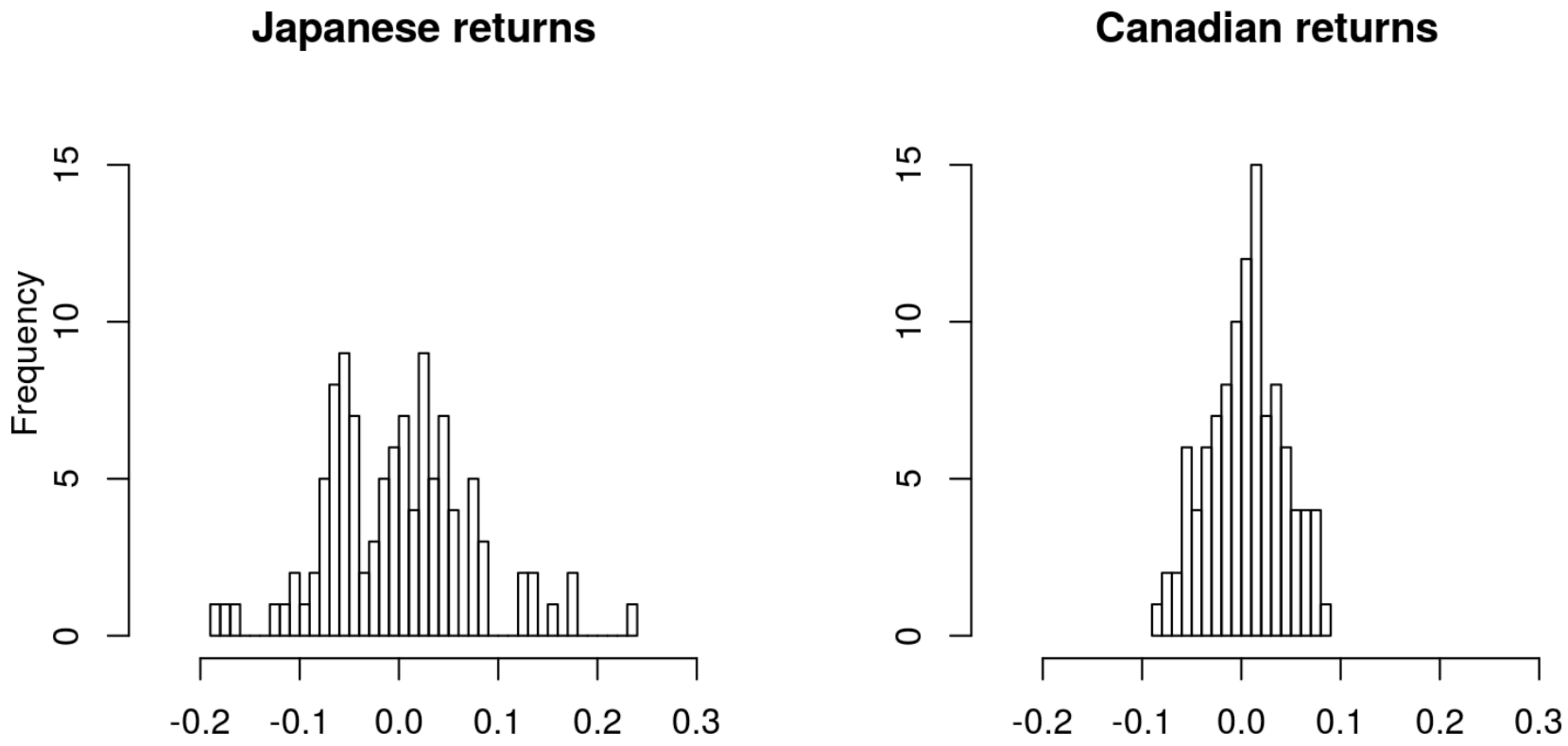
The median is often nice to report when there are extreme values in the data.

Example: Think about US household income



The Sample Variance and Standard Deviation

The mean and median provide information about the center of the distribution, but they provide no information about the dispersion or variability of the data.

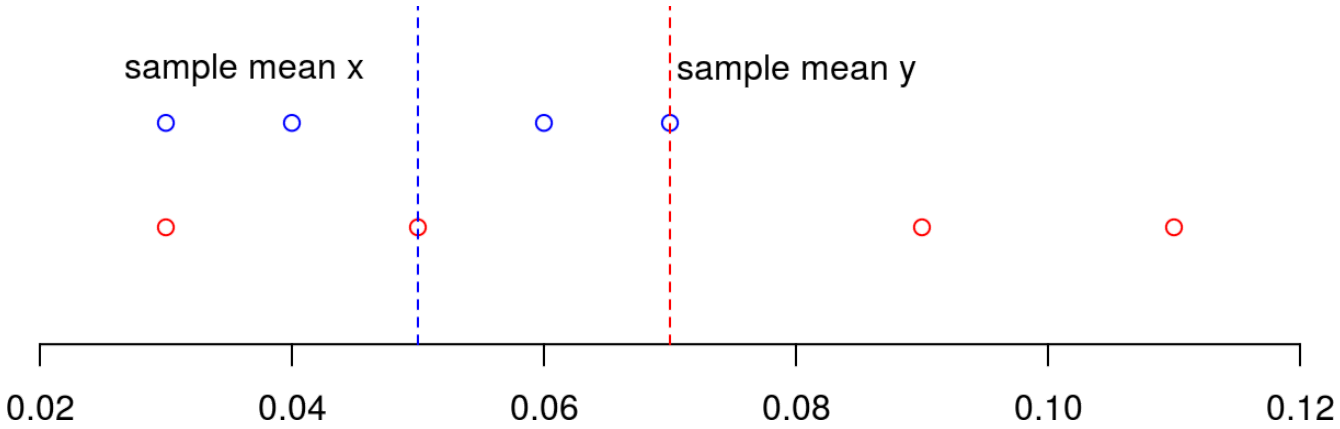


Simple Example

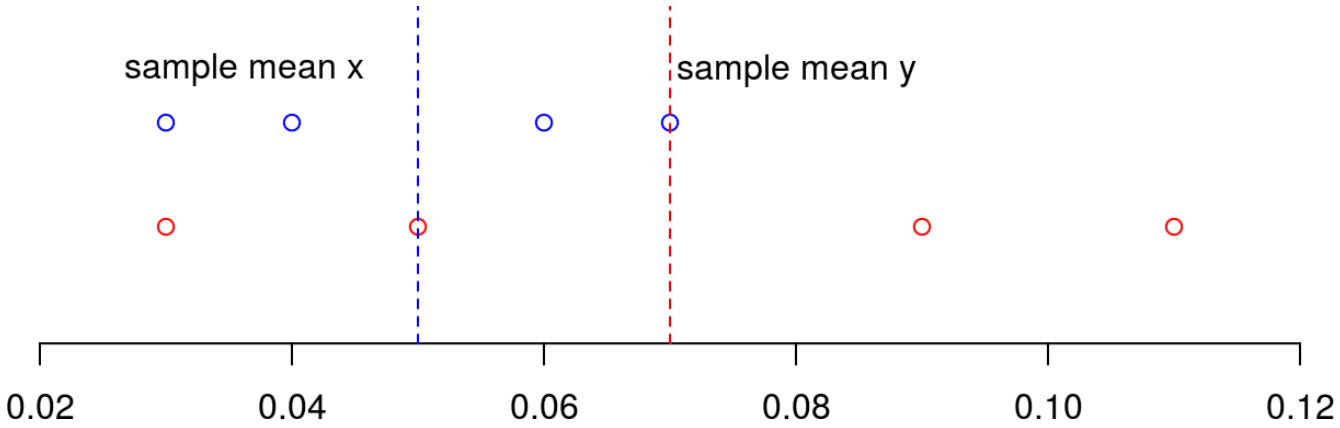
Suppose we have yearly returns data for two assets.

```
toydf = data.frame(x=c(0.07,0.06,0.04,0.03), y=c(0.11,0.05,0.09,0.03))
toydf
```

```
##      x    y
## 1 0.07 0.11
## 2 0.06 0.05
## 3 0.04 0.09
## 4 0.03 0.03
```



Simple Example




```
##        x     x-x̄     y     y-ȳ
## 1   0.07    0.02  0.11    0.04
## 2   0.06    0.01  0.05   -0.02
## 3   0.04   -0.01  0.09    0.02
## 4   0.03   -0.02  0.03   -0.04
```

The numbers in column 4 are bigger than numbers in column 2.

How do we summarize this?

We consider the average squared distance $\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$

The sample variance

The sample variance is the average squared distance from the mean.

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- For technical reasons, we divide by $n-1$ instaed of n. There is little practical difference.

What is the smallest value the variance can be?

What are the units of the variance?

The sample standard deviation

The sample standard deviation measures the variability in the same units our original data.

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Back to our simple example

```
var(toydf$x)

## [1] 0.00033333

sd(toydf$x)

## [1] 0.018257

var(toydf$y)

## [1] 0.0013333

sd(toydf$y)

## [1] 0.036515
```

```
##        x     x-x̄      y     y-ȳ
## 1   0.07    0.02   0.11    0.04
## 2   0.06    0.01   0.05   -0.02
## 3   0.04   -0.01   0.09    0.02
## 4   0.03   -0.02   0.03   -0.04
```

The sample standard deviation of y is greater than x.

This captures numerically the fact that y has "more variation" about its mean than x.

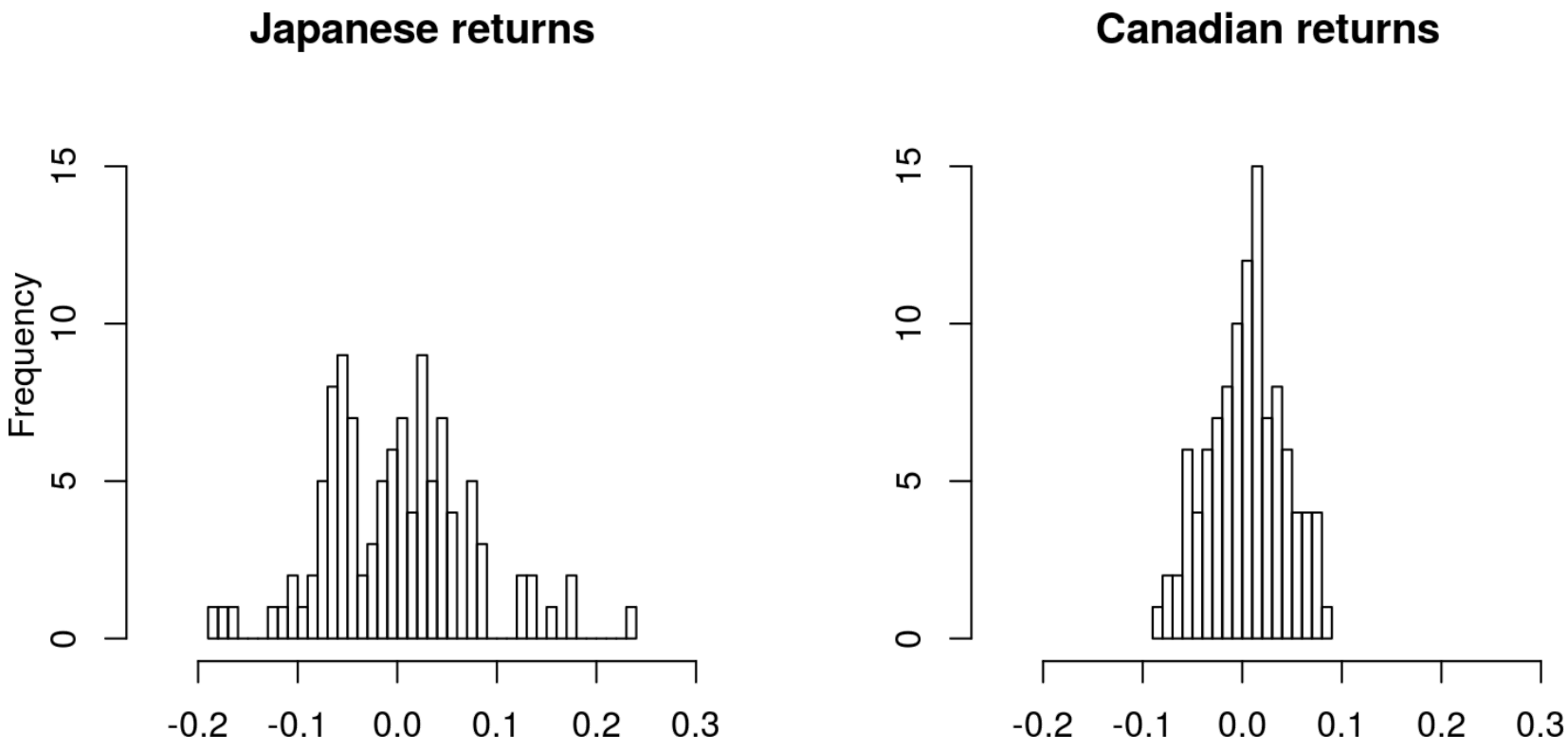
We can do the calculation by hand as well:

$$
\begin{aligned}
s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2 = \frac{1}{3}\Big((0.02)^2+(0.01)^2+(-0.01)^2+(-0.02)^2\Big) \\
&= 0.000333 \\
s_x &= \sqrt{0.000333} = 0.01826
\end{aligned}
$$

The sample standard deviation captures how the Japanese returns are more spread out than the Candian returns.

```
sapply(countryReturn_df[c("japan","canada")], sd)
```

```
##    japan   canada
## 0.073684 0.038327
```



Empirical rule

We have a way to compute the sample mean and sample standard deviation.

How do these numbers relate back to data?

A good **empirical rule** of thumb goes as follows

- Approximately 68% of the data is in the interval

$$(\overline{x} - s_x, \overline{x} + s_x) = \overline{x} \pm s_x$$

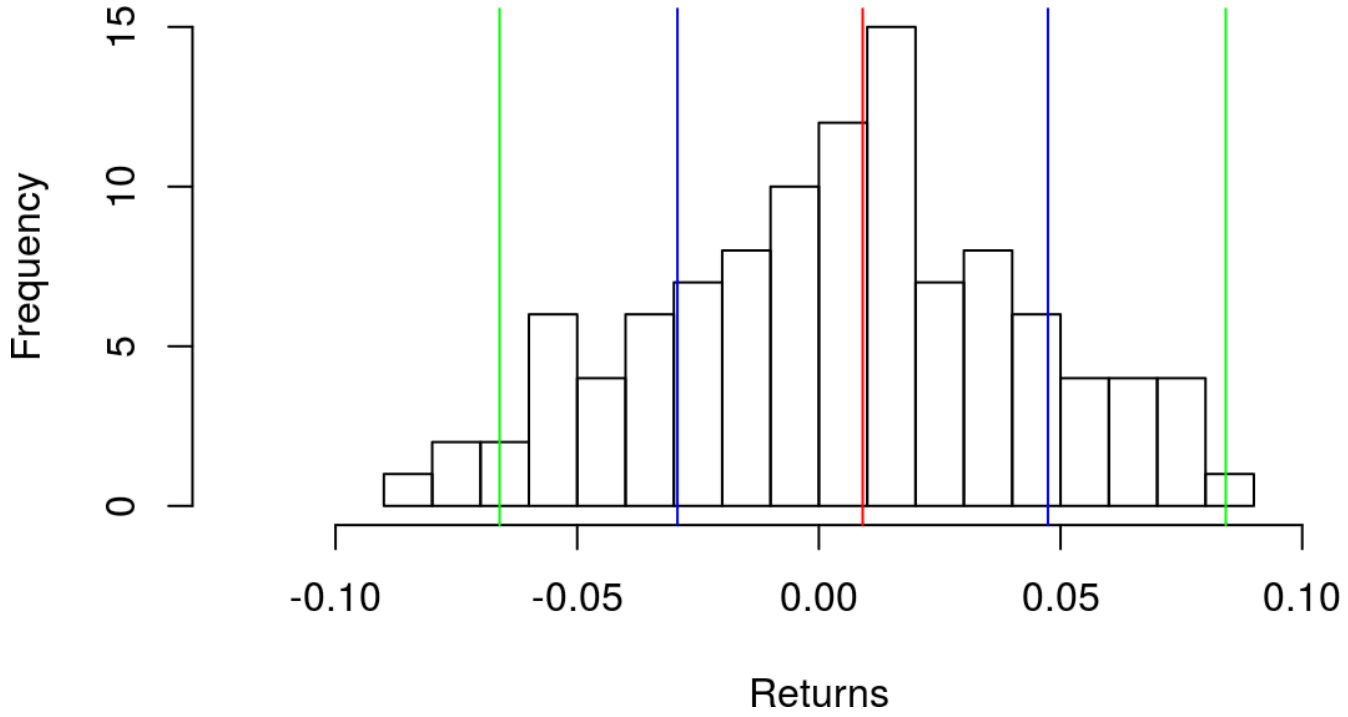
- Approximately 95% of the data is in the interval

$$(\overline{x} - 2 \cdot s_x, \overline{x} + 2 \cdot s_x) = \overline{x} \pm 2 \cdot s_x$$

IMPORTANT: This is true only for "mound-shaped" data.

Empirical rule

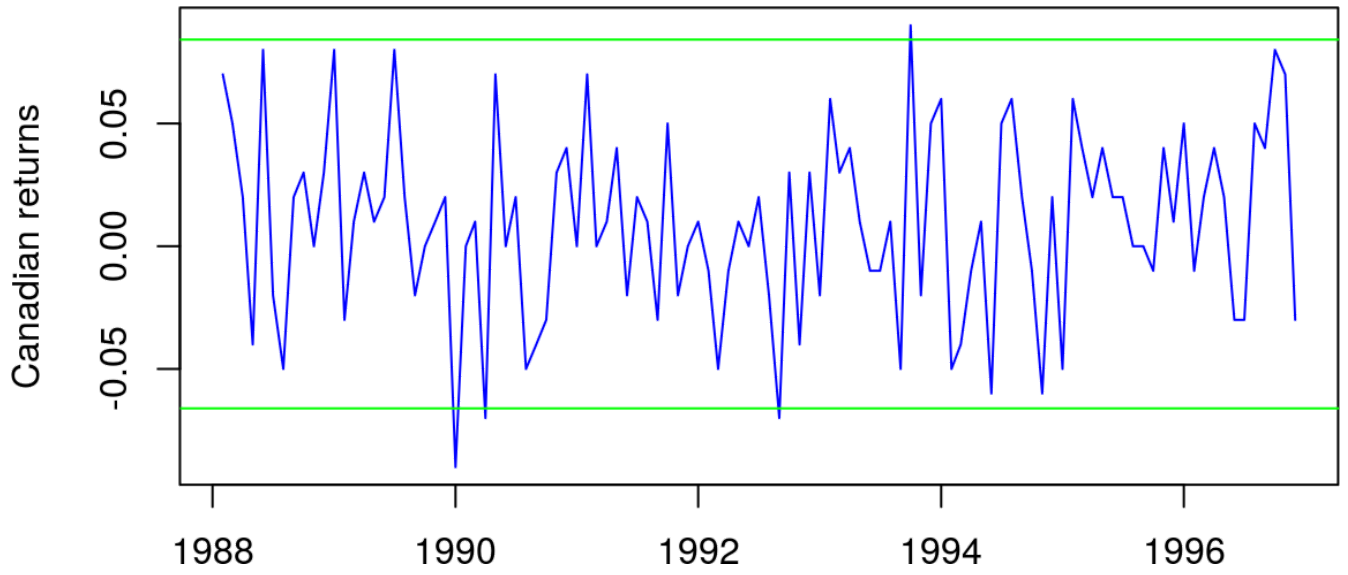
We can see this visually for the Canadian monthly return data.



The empirical rule says that roughly 95% of the data lies inside the green lines and 68% of the data lies inside the blue lines.

Empirical rule

A better visual may be the time series plot.



With $n = 107$, the empirical rule suggests that there should be 5 points lying outside the green lines. There are 4. Not bad!

Why shouly you care?

Eyeball standard deviation

Can be used to predict the future observation

Identify anomalies

Example: Comparing Mutual Funds

Let's compare mutual funds by looking at their means and standard deviations. We have monthly data on 12 different assets from July 1996 to Dec. 2013. (see mutualFundReturn.cvs)

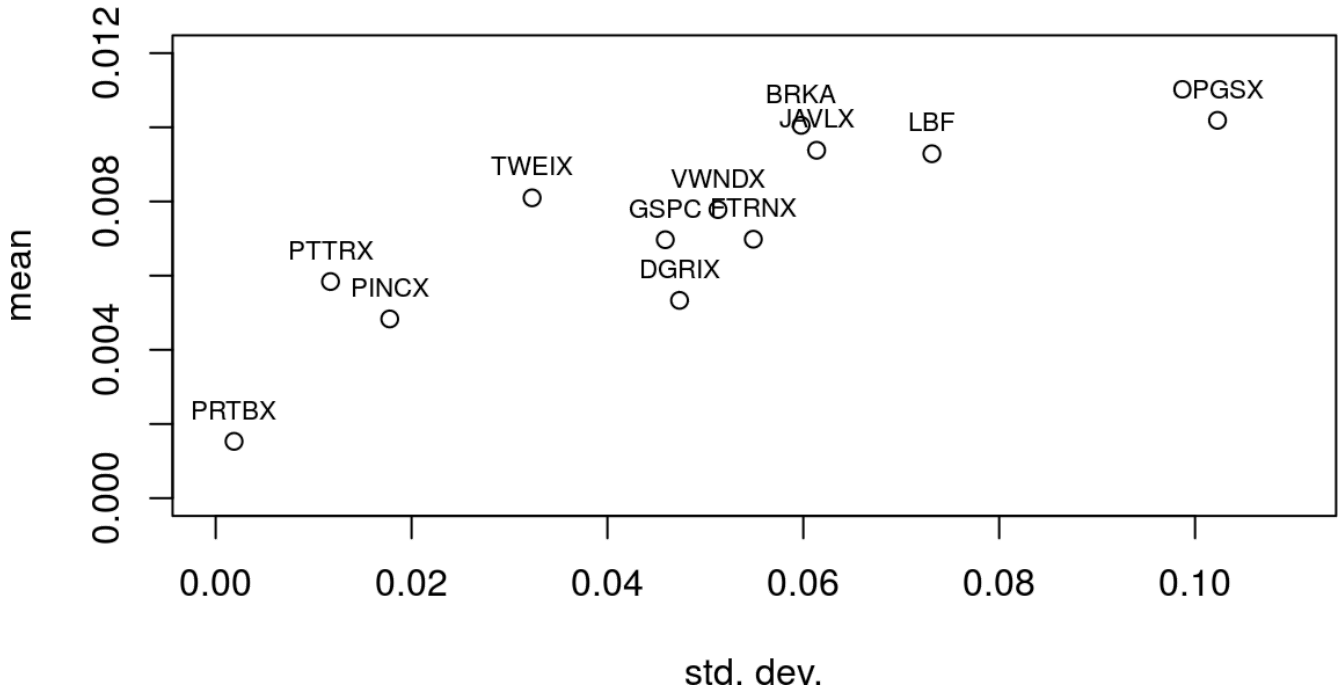
Ticker Symbol	Fund name
TWEIX	American Century Equity Income (large value)
BRKA	Berkshire Hathaway Holding Company A shares
DGRIX	Dreyfus Growth & Income (large growth)
LBF	DWS Global High Income Fund (bonds)
FTRNX	Fidelity Trend Fund (large growth)
JAVLX	Janus Twenty (large growth)
OPGSX	Oppenheimer Gold & Special Minerals
PRTBX	Permanent Portfolio Treasury Bill (ultrashort bonds)
PTTRX	Pimco Funds Total Return (intermediate bonds)
PINCX	Putnam Income
GSPC	S&P 500 index
VWNDX	Vanguard Windsor (large value)

```r
# drop first column as it represents date
mfr_df = read.csv("mutualFundReturn.csv")[,-1]
mean.mfr = sapply(mfr_df, mean)
sd.mfr = sapply(mfr_df, sd)
cbind(mean.mfr, sd.mfr)
```

```
##        mean.mfr    sd.mfr
## TWEIX 0.0080977 0.0323099
## BRKA  0.0100519 0.0597827
## DGRIX 0.0053327 0.0473610
## LBF   0.0092868 0.0731290
## FTRNX 0.0069820 0.0548984
## JAVLX 0.0093803 0.0613612
## OPGSX 0.0101873 0.1022936
## PRTBX 0.0015320 0.0018773
## PTTRX 0.0058405 0.0117253
## PINCX 0.0048325 0.0177709
## GSPC  0.0069683 0.0459095
## VWNDX 0.0077809 0.0512879
```

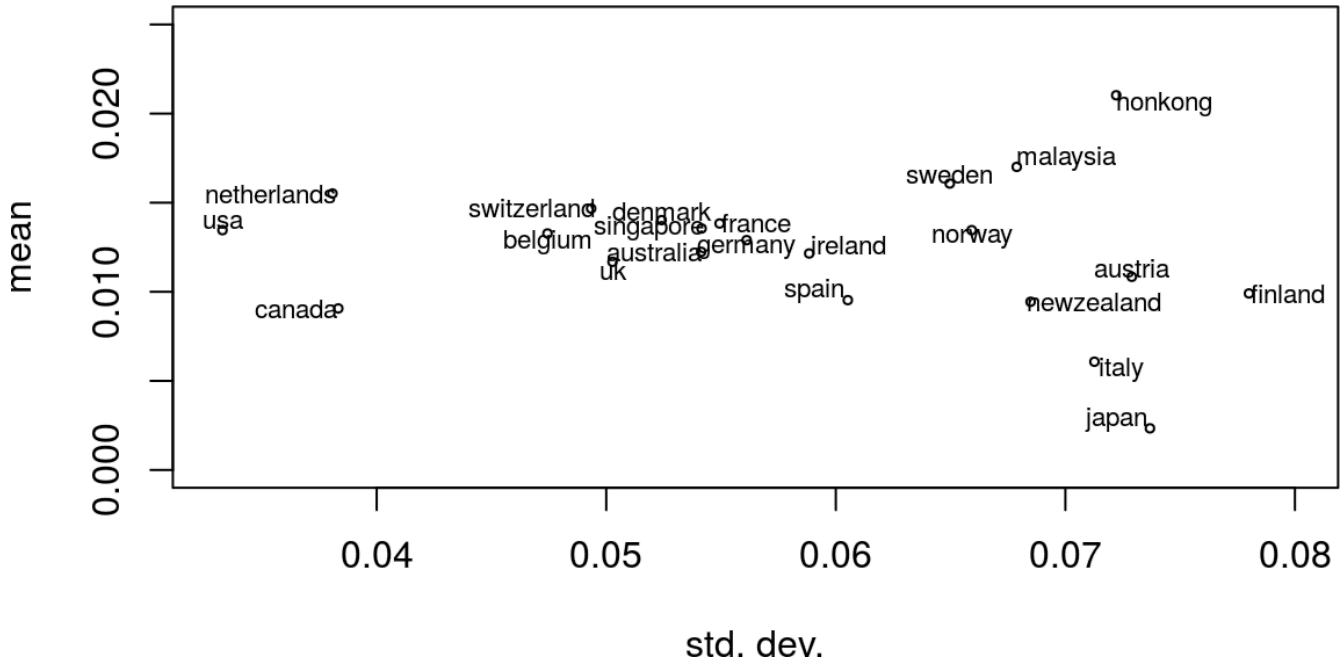
The largest return is the Oppenheimer Gold fund (OPGSX) but it has the highest variability. Unsurprisingly, the asset with the lowest variability is the ultra-short bond fund (PRTBX).

```
plot(sd.mfr, mean.mfr, type="p", xlim=c(0, 0.11), ylim=c(0, 0.012),
     main="", xlab = "std. dev.", ylab = "mean")
text(sd.mfr, mean.mfr, labels=names(mfr_df), cex= 0.7, pos=3)
```



If you're a fund manager, where do you want to be on this plot?

Example: Comparing Country Returns



Other summary statistics

Remember, we are still only considering numeric variables.

```
summary(countryReturn_df[,2:5])
```

```
##    australia          austria           belgium            canada        
##  Min.   :-0.1500   Min.   :-0.2300   Min.   :-0.1000   Min.   :-0.09000  
##  1st Qu.:-0.0200   1st Qu.:-0.0300   1st Qu.:-0.0100   1st Qu.:-0.02000  
##  Median : 0.0100   Median : 0.0100   Median : 0.0100   Median : 0.01000  
##  Mean   : 0.0122   Mean   : 0.0108   Mean   : 0.0133   Mean   : 0.00907  
##  3rd Qu.: 0.0500   3rd Qu.: 0.0500   3rd Qu.: 0.0300   3rd Qu.: 0.03500  
##  Max.   : 0.1700   Max.   : 0.2500   Max.   : 0.2500   Max.   : 0.09000
```

```
quantile(countryReturn_df$canada)
```

```
##     0%    25%    50%    75%   100% 
## -0.090 -0.020  0.010  0.035  0.090
```

```
range(countryReturn_df$canada)
```

```
## [1] -0.09  0.09
```

```
IQR(countryReturn_df$canada)
```

```
## [1] 0.055
```

Percentiles

The pth percentile is a number such that p% of the observations are less than it.

Examples:

- The 25th quantile is the number such that 25% of the values are less than it and 75% are larger.
- The median is the 50th percentile because 50% of the data is smaller and 50% is larger.

IMPORTANT: Percentiles are also known as **quantiles**.

```
quantile(countryReturn_df$canada, 0.3)
```

```
##   30%
## -0.01
```

Quartiles and IQR

The first, second, and third **quartiles** are the 25th, 50th, and 75th percentiles.

Q_1 — first quartile (splits off the lowest 25% of data from the highest 75%)
Q_2 — second quartile (median, cuts a dataset in half)
Q_3 — third quartile (splits off the highest 25% of data from the lowest 75%)

```
quantile(countryReturn_df$canada, probs=c(0.25, 0.5, 0.75))
```

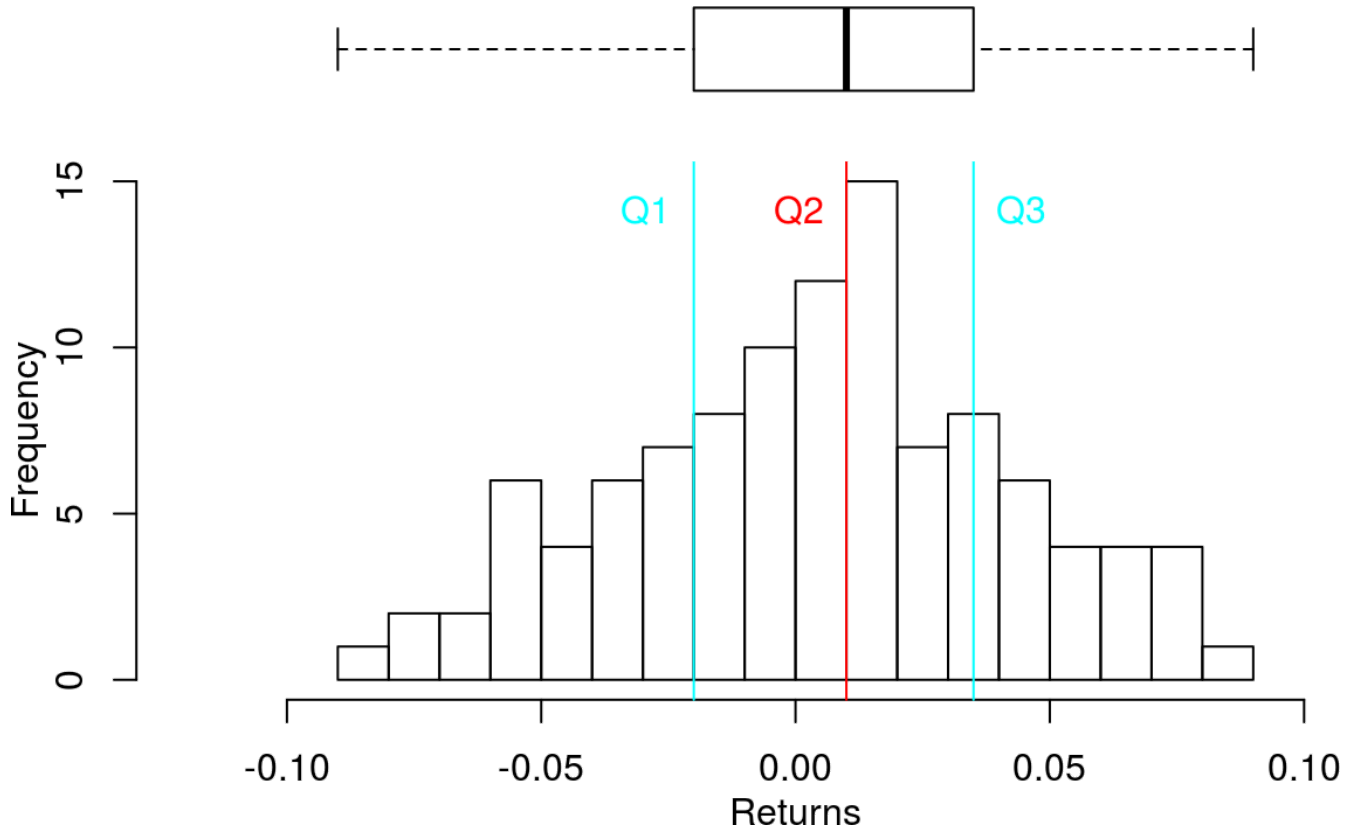
```
##    25%    50%    75%
## -0.020  0.010  0.035
```

The interquartile range (IQR) is the difference between the first and third quartiles.

$$IQR = Q_3 - Q_1$$

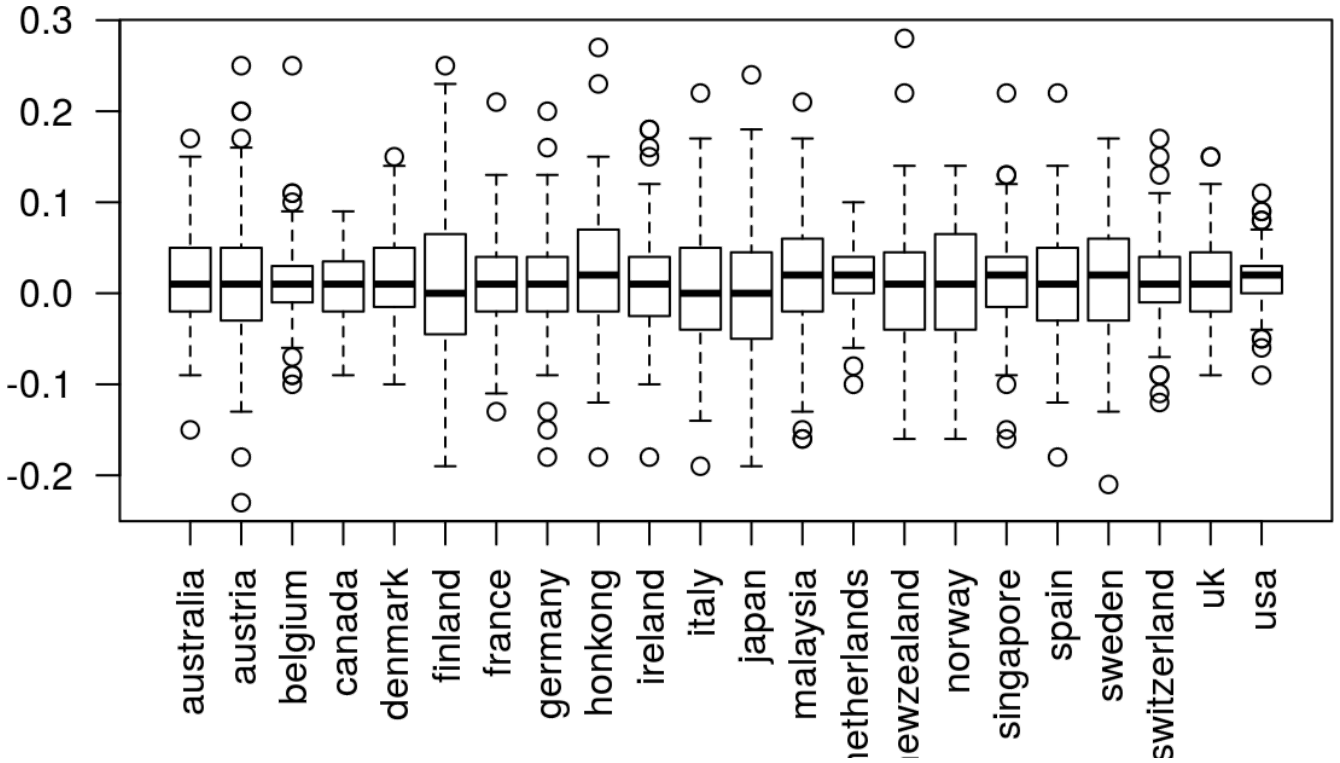
```
IQR(countryReturn_df$canada)
```

```
## [1] 0.055
```


Histogram of Canadian returns
Q1
Q2
Q3
Frequency
0
5
10
15
-0.10
-0.05
0.00
0.05
0.10
Returns

Boxplot of returns per country

```
boxplot(x = as.list(countryReturn_df[,-1]), las=2)
```



The length of whiskers is $1.5 \cdot IQR$.

Circles denote *"outliers"*.

Outliers

Boxplot denotes an observation as an outlier if it is

- smaller than $Q_1 - 1.5 \cdot IQR$
- larger than $Q_3 + 1.5 \cdot IQR$

You should investigate outliers to see if they are interesting observations, measurement errors, or something else.

Important: Do not blindly throw away data.

Relationships between two variables

Covariance and Correlation

The mean and standard deviation help us summarize a bunch of numbers which are measurements of just one thing.

A fundamental and totally different question is how one thing relates to another.

We used a scatterplot to look at two things: the mean and standard deviation of different assets.

In this section of the notes we look at scatterplots and how correlation can be used to summarize them.

Scatter Plot

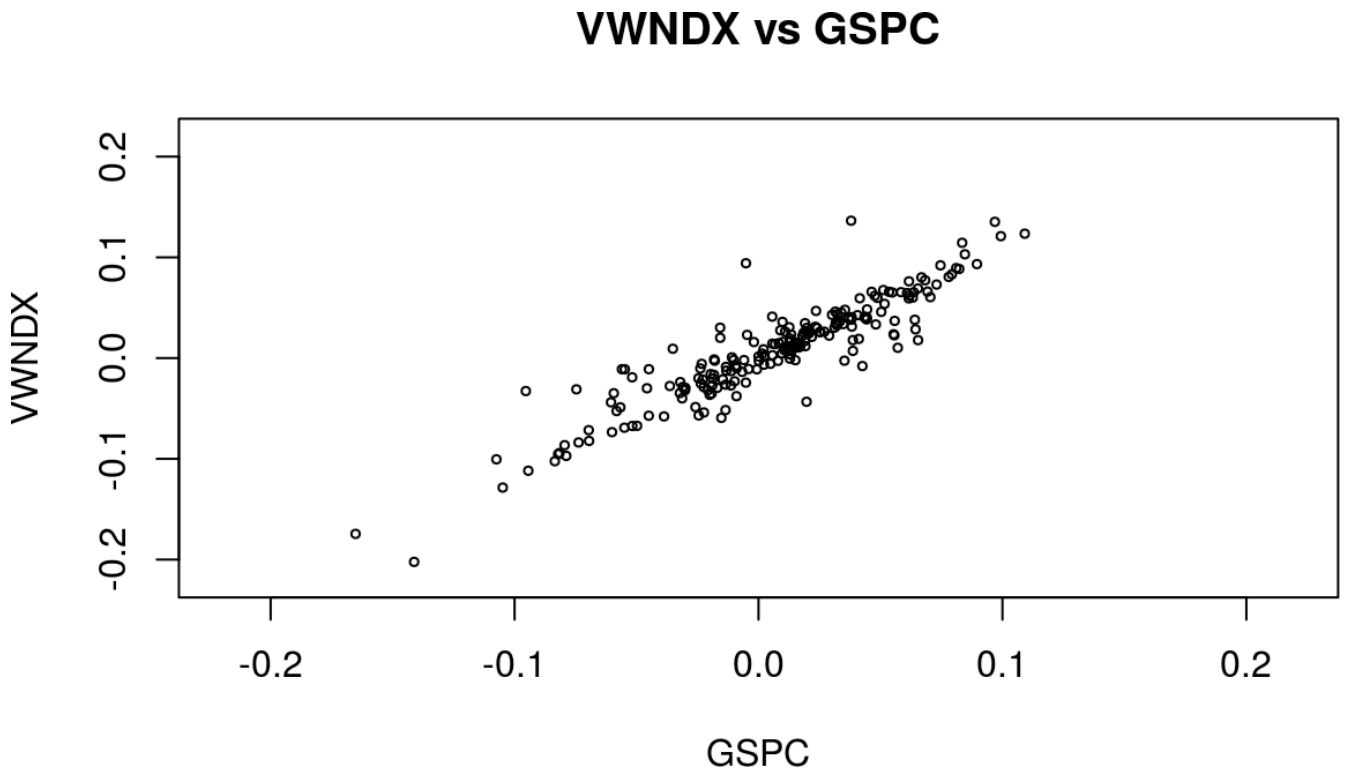
Scatter plot is used to visualize relationship between two numeric variables.

```
head(mfr_df[,c("VWNDX", "GSPC")])
```

```
##       VWNDX     GSPC
## 1 -0.010996 -0.04494
## 2  0.026815  0.01926
## 3  0.036943  0.05582
## 4  0.043612  0.03233
## 5  0.072984  0.07300
## 6 -0.005689 -0.02329
```

Are returns on a mutual fund related to the market?

Scatter Plot



Each point corresponds to a monthly return.

Scatter Plot

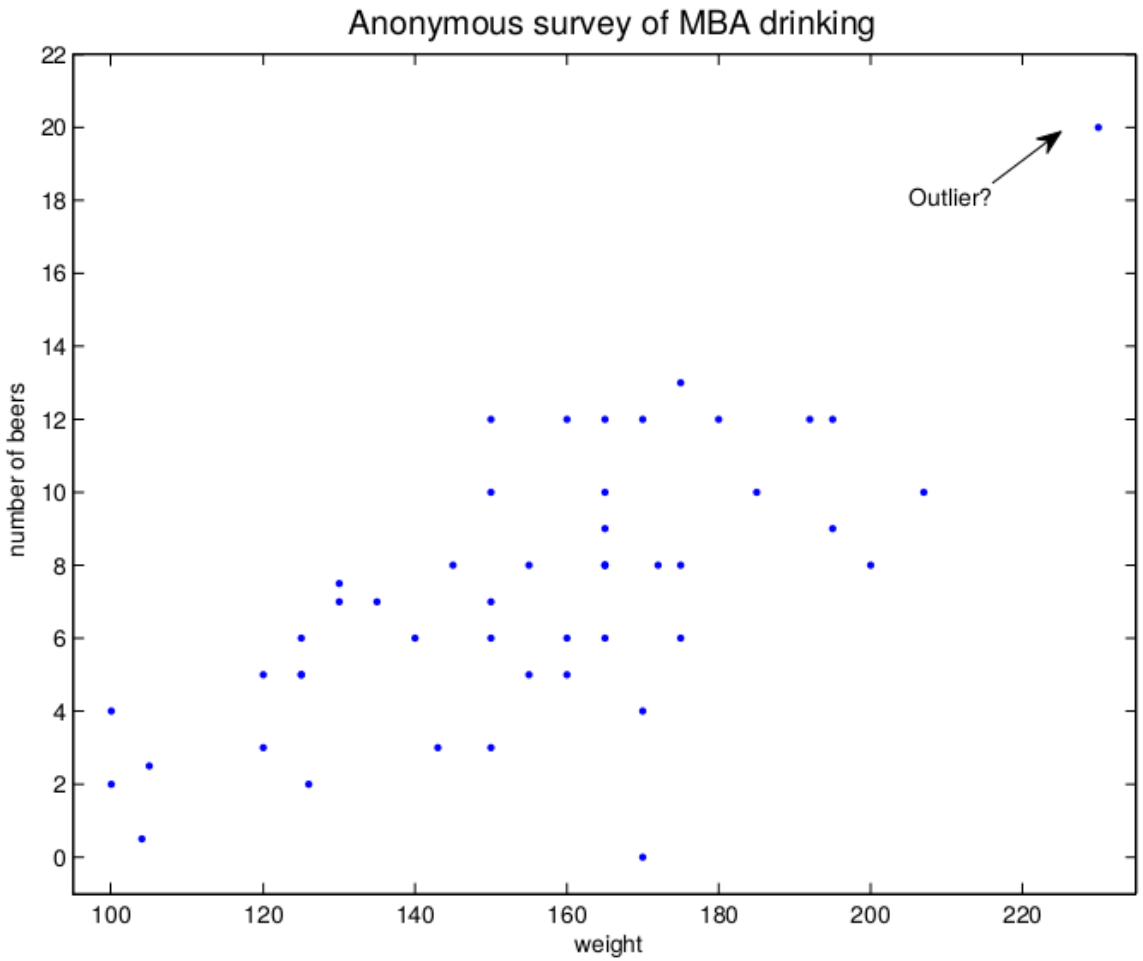
```
beer_df = read.csv("BeerConsumptionMBA.csv")
head(beer_df)
```

```
##   nbeer weight height age gender
## 1    12    192     72  26      0
## 2    12    160     66  27      0
## 3     5    155     65  25      0
## 4     5    120     66  28      0
## 5     7    150     67  28      0
## 6    13    175     71  31      0
```

Each row corresponds to an individual and their drinking "ability."

Each person has recorded the number of beers they can drink and their weight.

Scatter Plot



Covariance and correlation

The plot enables us to see the relationship between x and y. In both examples it does look like there is a relationship.

Even more, the relationship looks linear in that it looks like we could draw a line through the plot to capture the pattern.

Covariance and correlation summarize how strong a linear relationship there is between two variables.

Covariance and correlation

The **sample covariance** between data x and y is defined as

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}) (y_i - \overline{y})$$

- What are the units of the covariance?

The **sample correlation** between data x and y is defined as

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

The sample correlation is simply the sample covariance divided by the standard deviation of x and y, respectively.

- What are the units of the correlation?

Properties of the Sample Correlation

The sample correlation always lies between -1 and 1, i.e. $-1 \leq r_{xy} \leq 1$.

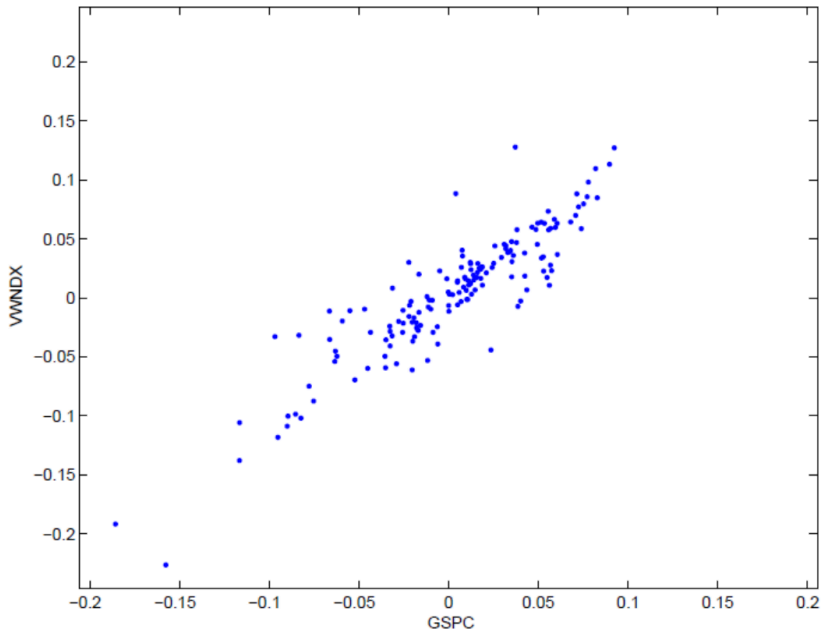
The closer r_{xy} is to 1 the stronger the linear relationship is with a *positive slope.* When one variable increases the other tends to increase.

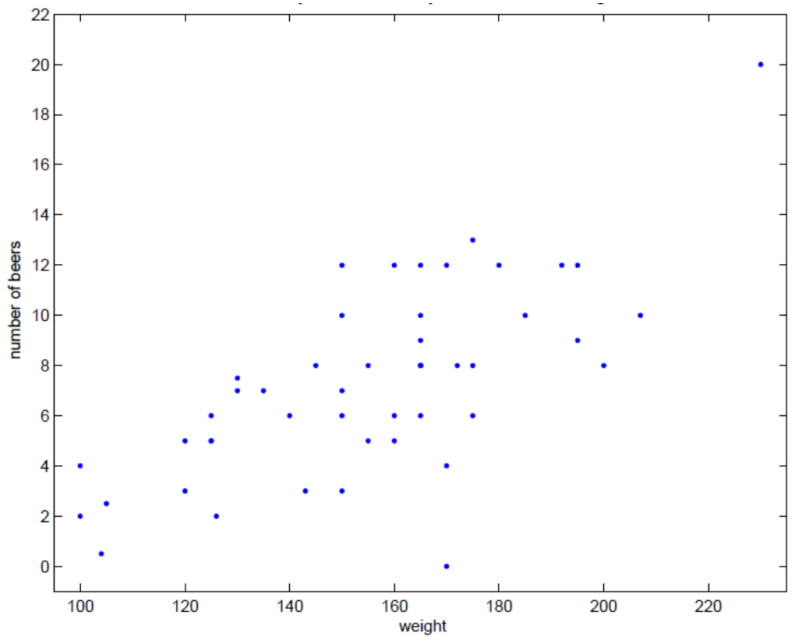
The closer r_{xy} is to -1 the stronger the linear relationship is with a *negative slope.* When one variable increases the other tends to decrease.

A correlation of 1 is a straight line with positive slope.

Correlation

Compare the mutual fund data and the beer data.

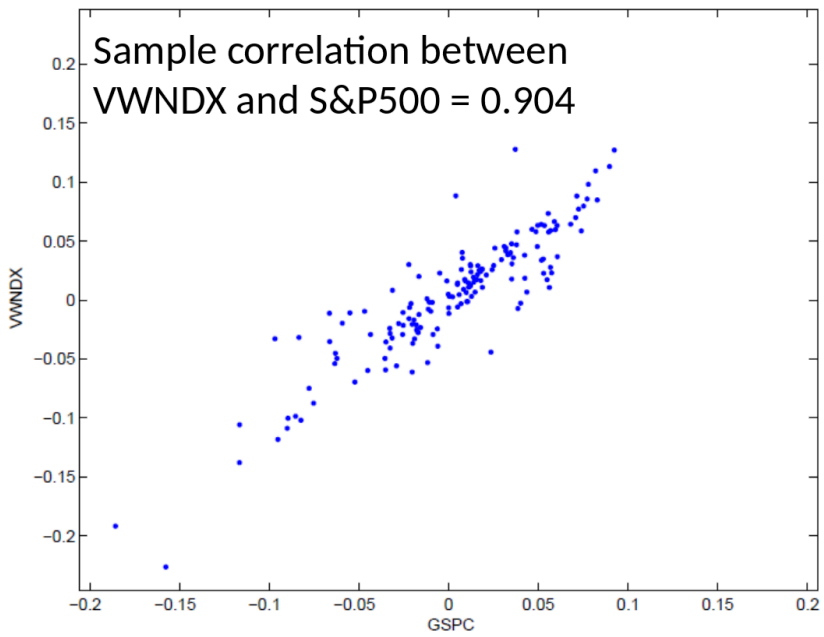


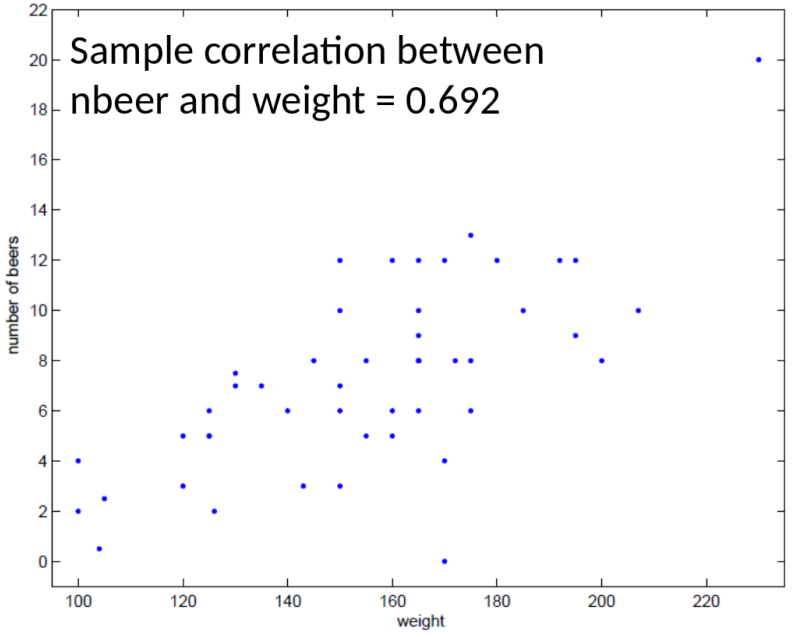


Which appears to be more correlated?

Correlation

Compare the mutual fund data and the beer data.





Which appears to be more correlated?

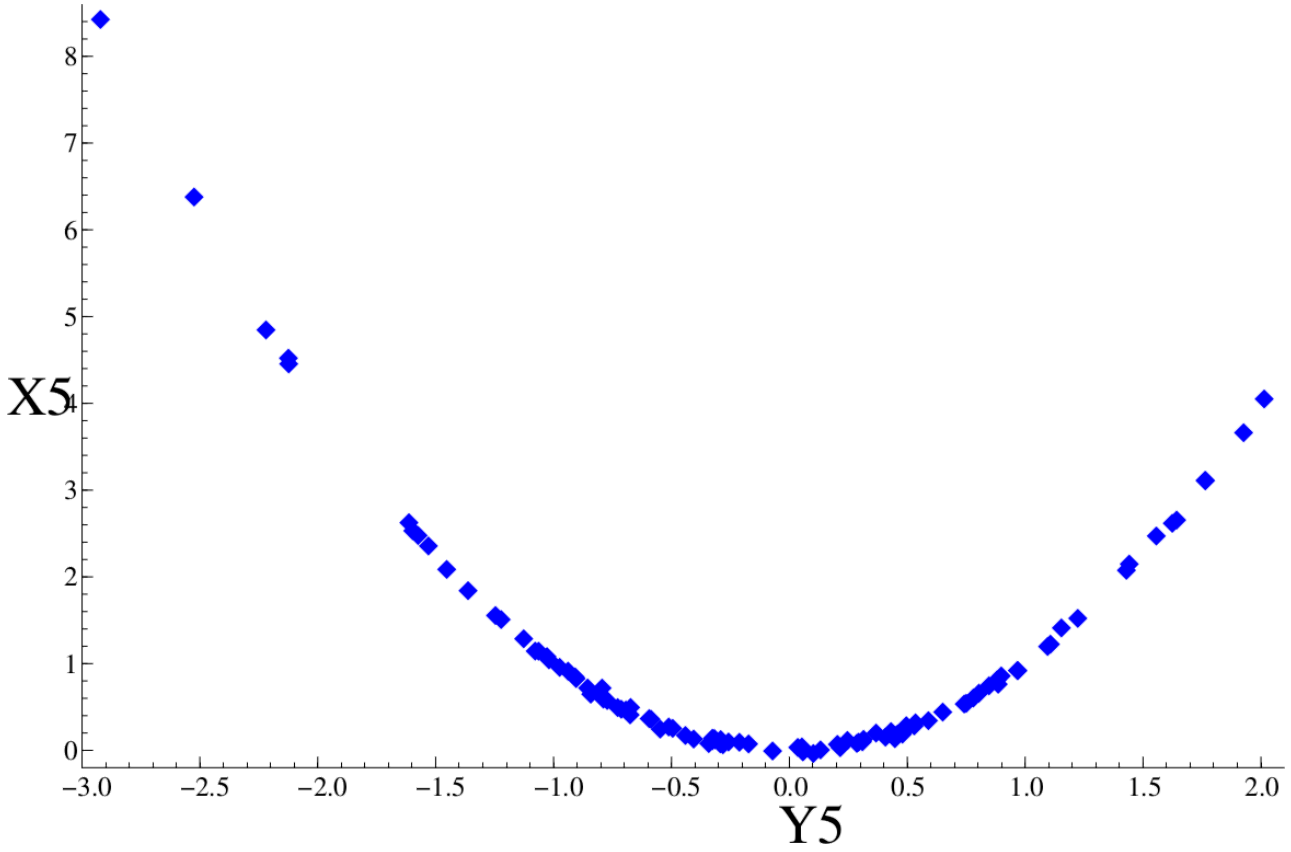
The larger correlation between VWNDX and S&P500 indicates that the linear relationship is STRONGER.

Correlation: further examples

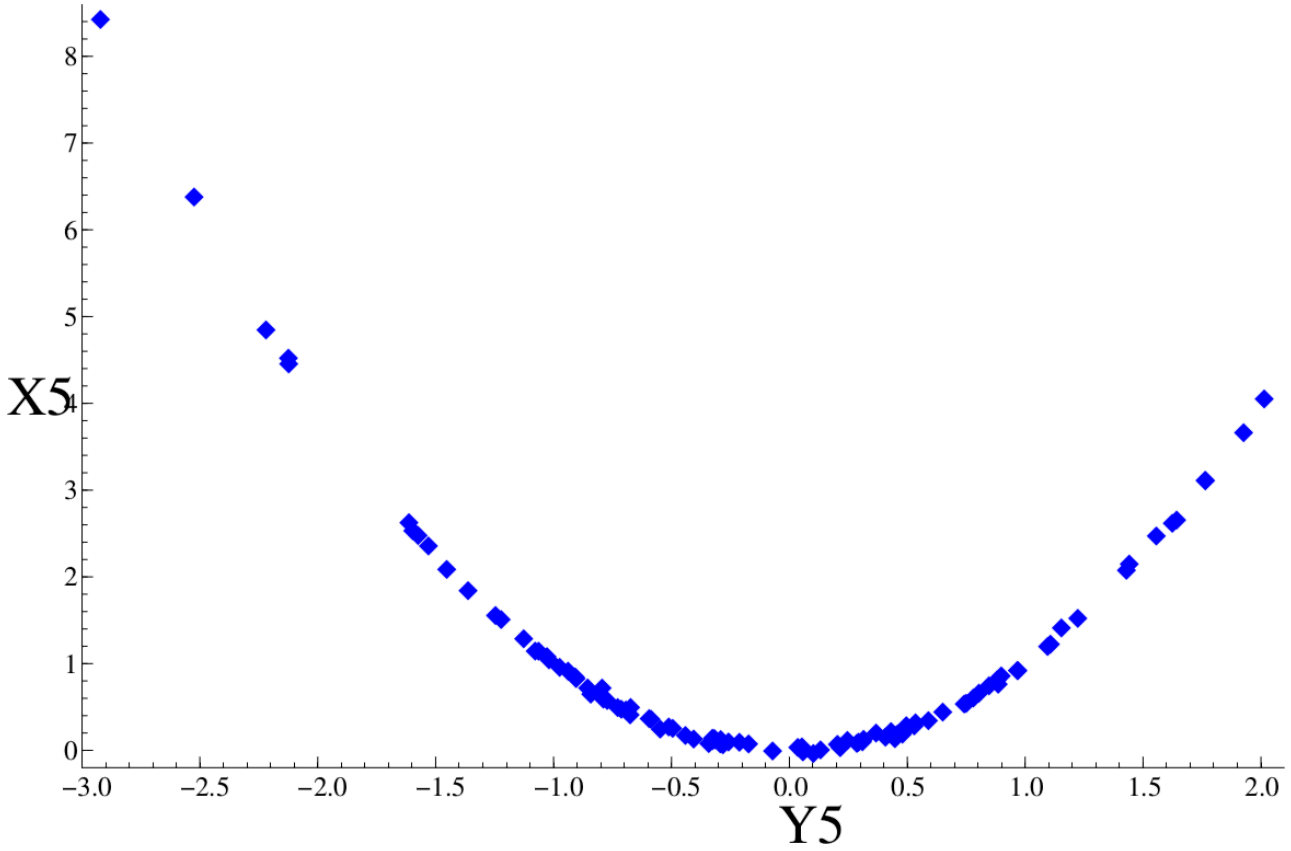
Illustrate correlation between two variables:
https://mlakolar.shinyapps.io/bivariateCorrelation/

Guess the correlation between two variables:
https://mlakolar.shinyapps.io/correlationGame/

Correlation: Be cautious



Correlation: Be cautious

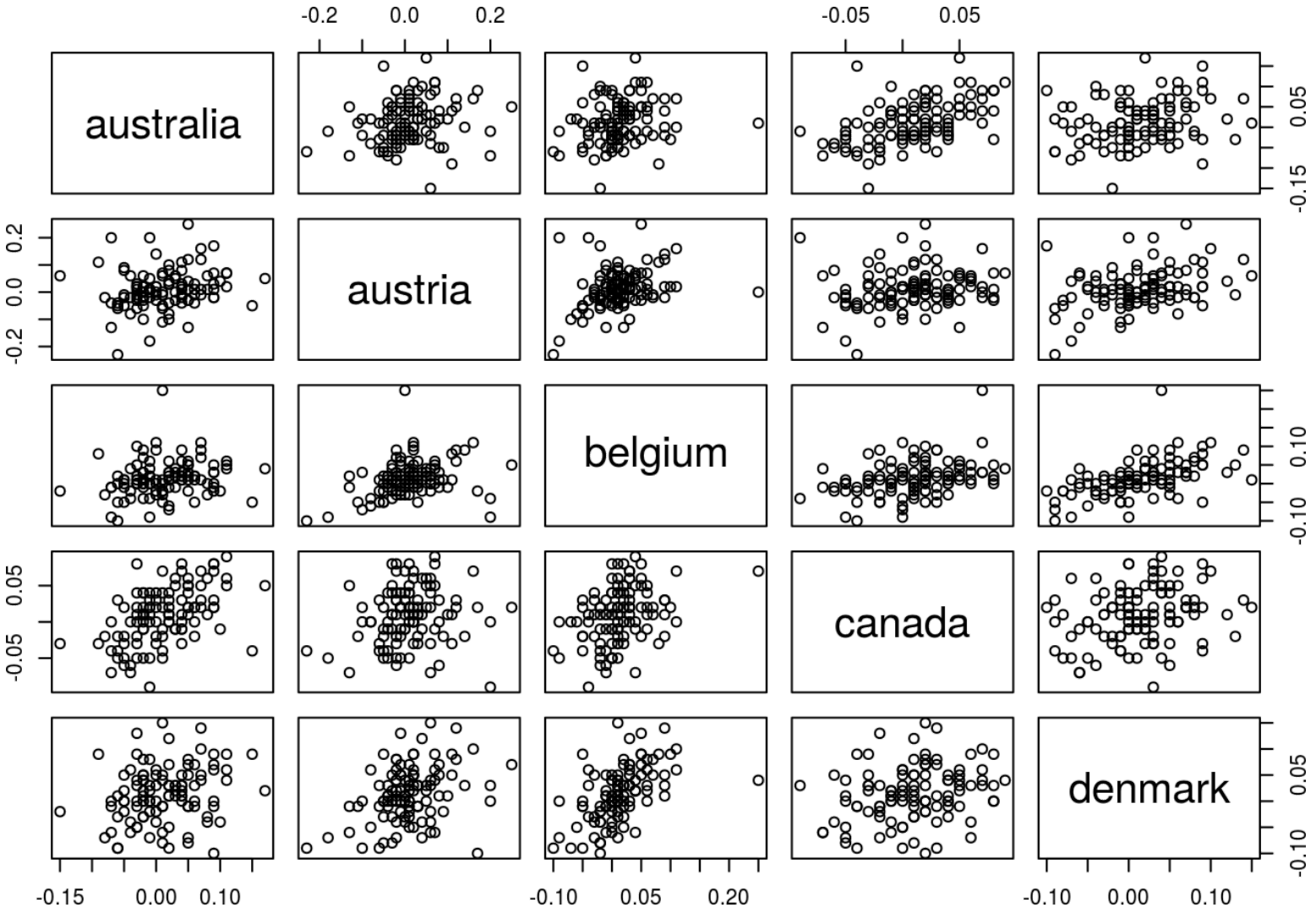


IMPORTANT: Correlation only measures a **LINEAR** relationship.

Clearly, the variables x_5 and y_5 are highly (nonlinearly) related.
Correlation between x_5 and y_5 is -0.26

Many scatter plots

```
pairs(countryReturn_df[,2:6])
```



The Sample Correlation Matrix

The **sample correlation matrix** is a table of all sample correlations between each pair of variables. It can be used to summarize a large number of plots.

```
cor(countryReturn_df[,2:6])
```

```
##           australia austria belgium  canada denmark
## australia   1.00000 0.18838 0.18923 0.50662 0.22987
## austria     0.18838 1.00000 0.31020 0.15393 0.34334
## belgium     0.18923 0.31020 1.00000 0.35712 0.53421
## canada      0.50662 0.15393 0.35712 1.00000 0.25127
## denmark     0.22987 0.34334 0.53421 0.25127 1.00000
```

Why are all the diagonal entries equal to one? Why are the upper and lower off-diagonals the same?

Using the correlation and covariance formulas

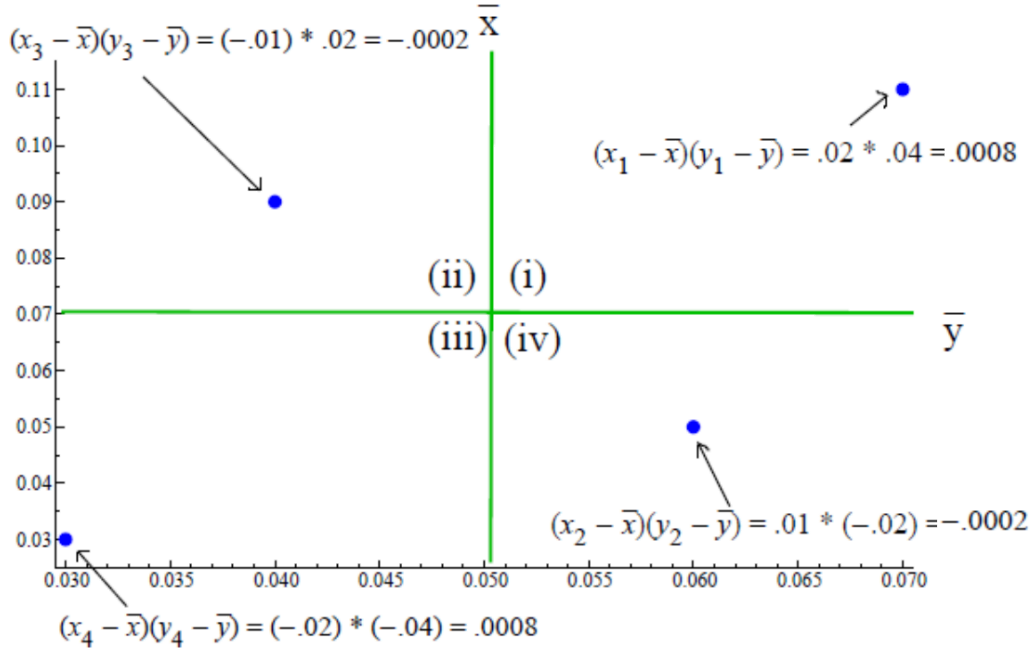
Why are the formulas for cov and corr capturing the relationship?

We will revisit our earlier example.

	x	x-x̄	y	y-ȳ
1	0.07	0.02	0.11	0.04
2	0.06	0.01	0.05	-0.02
3	0.04	-0.01	0.09	0.02
4	0.03	-0.02	0.03	-0.04

$$
\begin{aligned}
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right) \\
&= \frac{1}{3}(0.02*0.04+0.01*-0.02-0.01*0.02-0.02*-0.04) \\
&= \frac{1}{3}(0.0008-0.0002-0.0002+0.0008) \\
&= \frac{1}{3}(0.0012) = 0.0004
\end{aligned}
$$

Each of the four points contributes to the covariance.
Notice what determines the sign (-, +) and magnitude of each contribution.
Let's see which point does what.



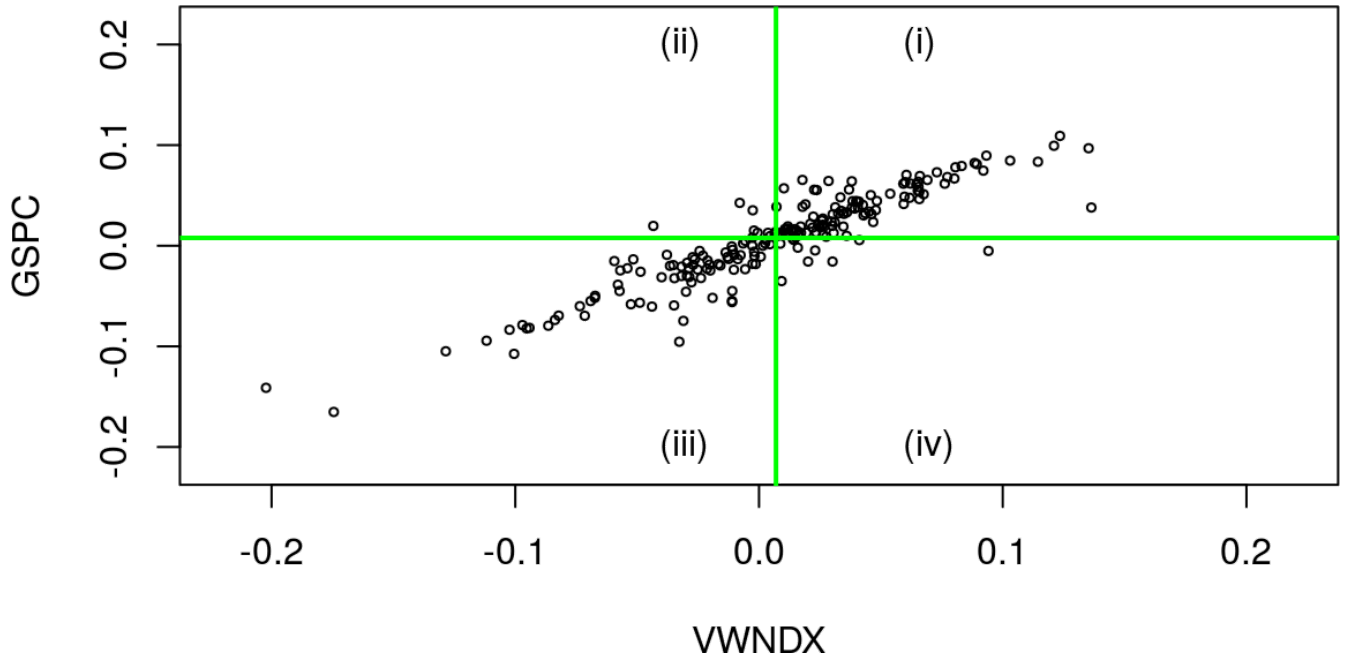
Points in quadrant (iii) have both x and y less than their means so they make a positive contribution to the covariance.

Points in quadrant (i) have both x and y larger than their means so they make a positive contribution to the covariance.

In (ii) and (iv) one of x and y is less than its mean and the other is greater so we get a negative contribution.

The sign (-, +) of the covariance tells us in which quadrants our data lies "on average."

There are lots of data points in quadrants (i) and (iii) which make positive contributions.



There are only a few data points in quadrants (ii) and (iv) which make negative contributions.

How Changes in Units Affect the Covariance

Suppose we have data on 4 individuals' education and income:

```
##      yearsSchool       Income
## 1             13       34,500
## 2             21      120,200
## 3             16       54,950
## 4             18       98,100
```

The sample covariance is 128650.

What are the units of measurement?

How Changes in Units Affect the Covariance

What if we measure income in thousands of dollars instead of dollars?

```
##      yearsSchool       Income
## 1             13         34.5
## 2             21        120.2
## 3             16        54.95
## 4             18         98.1
```

The sample covariance is 128.65!!

What are the units of measurement?

Key Points to Remember about Covariance

A positive covariance implies that when a variable is above (below) its mean the other variable tends to be above (below) its mean.

A negative covariance implies that when one variable is above (below) its mean the other variable tends to be below (above) its mean.

The units of the covariance are (typically) not meaningful.

The magnitude of the covariance is not easy to interpret.

Focus on the sign of the covariance. It tells us which quadrant we should expect to see our data relative to the mean.

How Changes in Units Affect the Correlation

Consider the same data on individuals' education and income as above:

```
##      yearsSchool       Income
## 1             13       34,500
## 2             21      120,200
## 3             16       54,950
## 4             18       98,100
```

The sample correlation is 0.975.

What are the units of measurement?

How Changes in Units Affect the Correlation

Again, what if we measure income in thousands of dollars?

```
##      yearsSchool       Income
## 1             13         34.5
## 2             21        120.2
## 3             16        54.95
## 4             18         98.1
```

The sample correlation remains 0.975!!

What are the units of measurement?

Key Points to Remember about Correlation

The correlation always has the same sign as the covariance because we are simply dividing by standard deviations which are always positive.

The correlation can be more informative than the covariance because it is easier to interpret as a measure of strength.

Correlation is unit-less and always lies between -1 and 1.

Interpretation: close to 1 means a strong positive relationship.

Interpretation: close to -1 means a strong negative relationship.