Business Statistics 41000

Lecture 2

Mladen Kolar

Outline of today's topics

Linearly related variables

- Linear regression
- Linear functions and linear combinations
- Mean and variance of a linear combination
- Portfolios

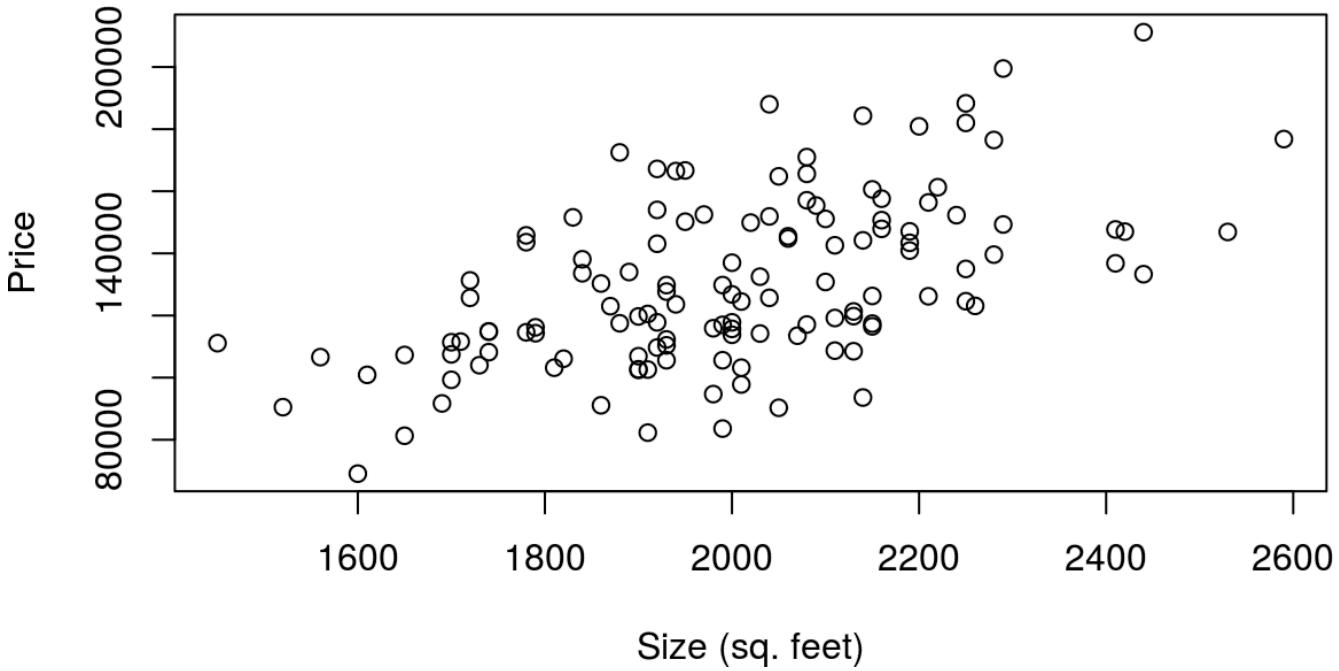
Clustering

- Distance and similarity

Linear regression

This is data on 128 homes. It includes their sales price (in dollars) and interior size (in square feet).

```
homep_df = read.csv("housesp1.csv")
plot(homep_df$size, homep_df$price, xlab="Size (sq. feet)", ylab="Price")
```



Linear regression

Clearly, the data are correlated.

```
cor(homep_df)
```

```
##          size   price
## size  1.00000 0.55298
## price 0.55298 1.00000
```

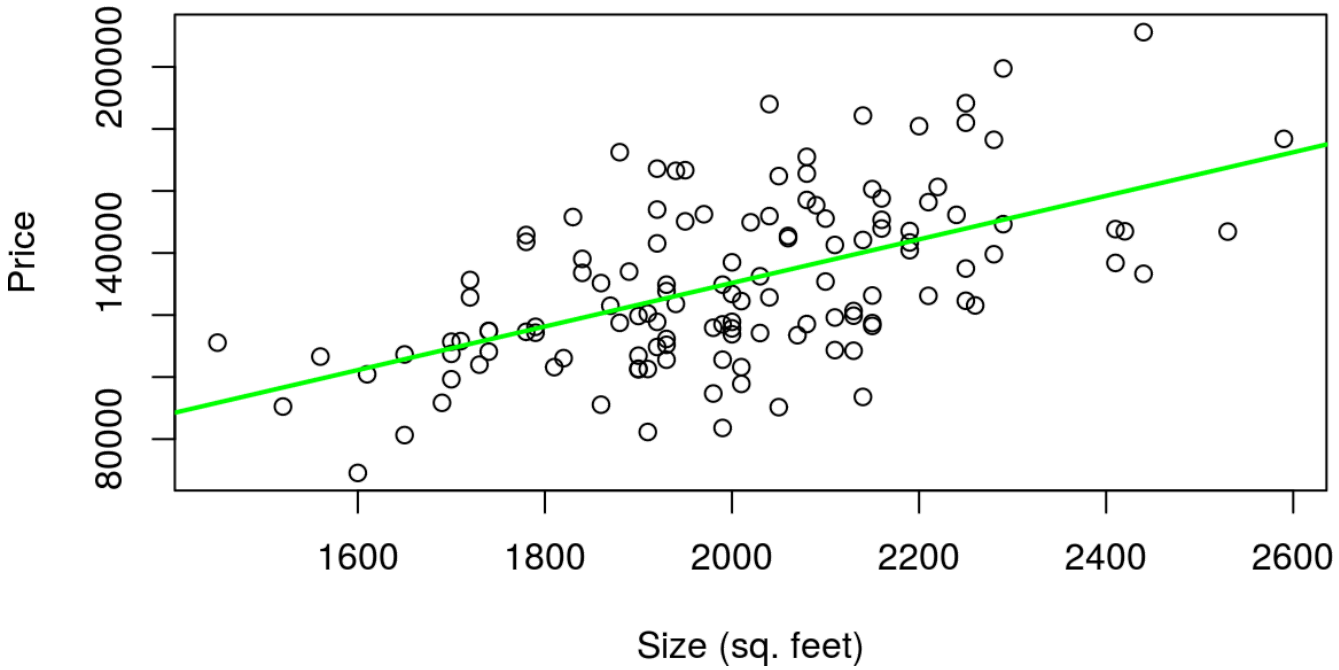
It looks like we could fit a line through the data.

But, which line? And, what is the equation of that line?

Why would we want to do this?

Here is the data with the regression line drawn through it.

```
plot(homep_df$size, homep_df$price, xlab="Size (sq. feet)", ylab="Price")
reg = lm(price ~ size, homep_df)
abline(reg, lw=2, col="green")
```



Linear regression fits a line through data.

Let y be the house prices and x be the size of the house.

(Univariate) linear regression helps us find the **linear relationship** or **linear function** between two variables

$$y = c_0 + c_1 \cdot x.$$

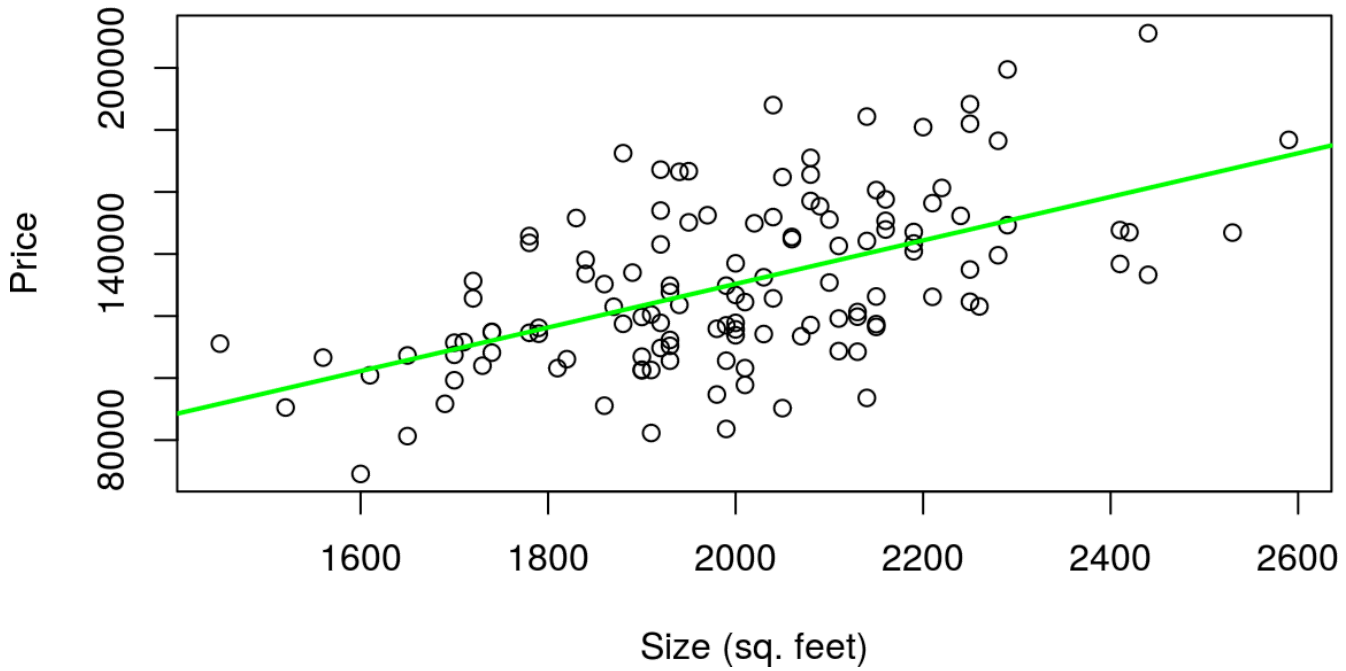
The variable c_0 is the **intercept**.

The variable c_1 is a number called the **slope**.

When we "run a regression," we obtain values for the intercept and slope coefficients given our data.
Note: the linear relationship holds approximately.

```
summary(reg)
```

```
##
## Call:
## lm(formula = price ~ size, data = homep_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -46593 -16644  -1610  15124  54829
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10091.13   18966.10   -0.53      0.6
## size            70.23       9.43    7.45  1.3e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22500 on 126 degrees of freedom
## Multiple R-squared:  0.306,  Adjusted R-squared:   0.3
## F-statistic: 55.5 on 1 and 126 DF,  p-value: 1.3e-11
```



The green line has formula

$$y = -10091.13 + 70.23 \cdot x$$

What is the interpretation?

Linear regression formulas

$$\text{slope} = \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x} \cdot r_{xy}$$

$$\text{intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

The formulas for the slope and intercept just use the sample mean, sample covariance, and sample variance.

We will study this in more detail later in the class.

The slope formula takes the covariance and "standardizes" it so that its units are (units of y)/(units of x).

The intercept formula will make the regression line pass through the point $(\bar{x}, \bar{y})$.

Regression and prediction

You have a house on the market with size 2200 sq. ft.

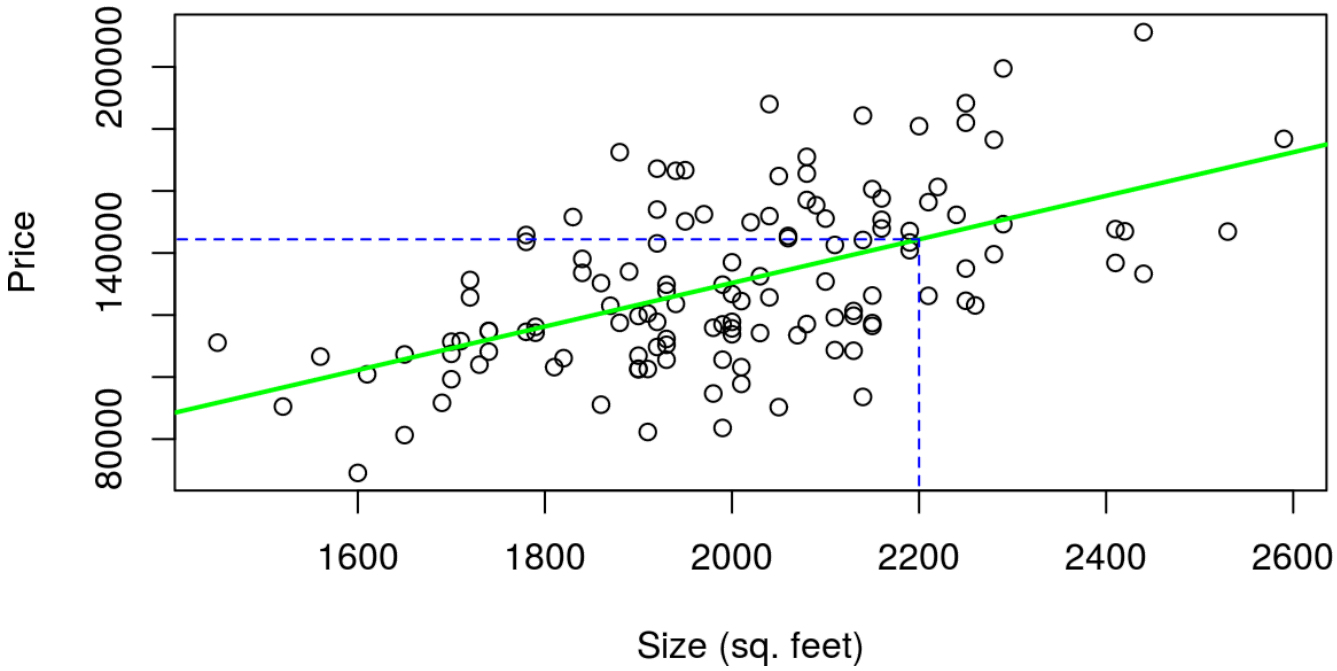
Can we predict at what price the house will sell?

We could use the sample mean or median of prices but this doesn't take size into account.

Regression allows us to use information on size to form our prediction.

We plug the size (2200 sq. ft) into our equation.

Predicted price: $-10091.13 + 70.23 \cdot 2200 = \$144,407$



Additional comments on regression

Because we are using more information (in this case the information on the size of the home), the predictions we make are (hopefully!) better in some sense.

Importantly, regression is based on the same concepts (sample means, sample covariance and variances) that we learned in so far in the class.

It's simply an alternative way to use our information.

There is nothing mysterious about it!

Sample mean and variance of a linear function

Suppose y is a linear function of x.

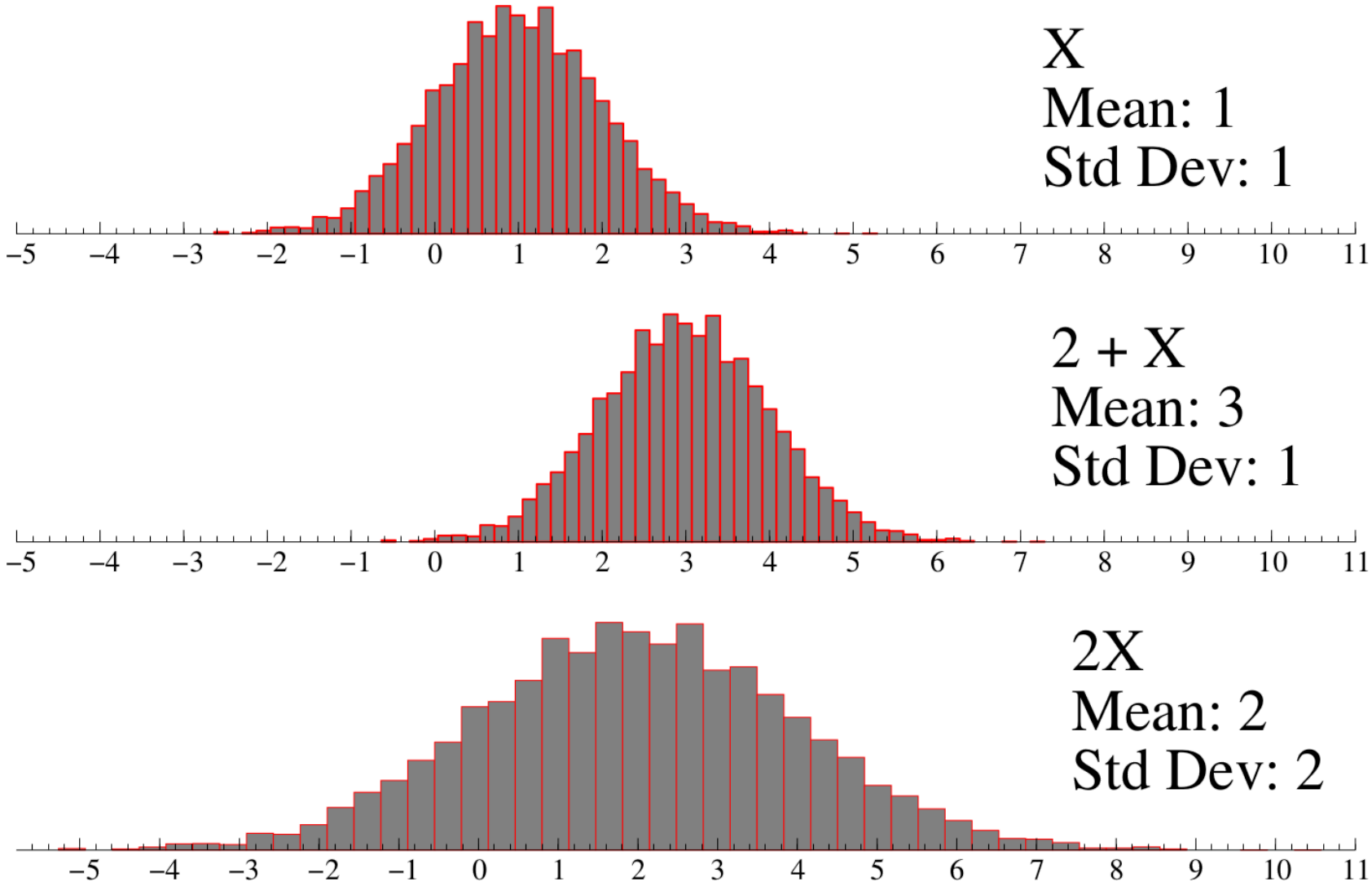
$$y = c_0 + c_1 x$$

How are the sample mean and variance (std. dev.) of y related to those of x?

In other words, given $\overline{x}$ and s_x^2, what separate affects do multiplying by c_1 and adding c_0 have?

Seeing the effects graphically

Histogram of 1000 data points



Sample mean and variance of a linear function

Suppose

$$y = c_0 + c_1 x$$

Then

$$\begin{aligned} \bar{y} &= c_0 + c_1 \bar{x} \\ s_y^2 &= c_1^2 s_x^2 \\ s_y &= |c_1| \, s_x \end{aligned}$$

Example

Suppose x has sample mean 100 and sample standard deviation 10.

What are the sample mean, sample variance, and sample standard deviation of y when

1. $y = 2 \cdot x$
2. $y = 5 + x$
3. $y = 5 - 2 \cdot x$

NOTE: Answers are on the next slide.

Example

Suppose x has sample mean 100 and sample standard deviation 10.

What are the sample mean, sample variance, and sample standard deviation of y when

1. $y = 2 \cdot x$ $\quad \bar{y} = 200, \;\; s_y^2 = 400, \;\; s_y = 20$
2. $y = 5 + x$ $\quad \bar{y} = 105, \;\; s_y^2 = 100, \;\; s_y = 10$
3. $y = 5 - 2 \cdot x$ $\quad \bar{y} = -195, \;\; s_y^2 = 400, \;\; s_y = 20$

Linear combinations

We may want a variable y to be a function of more than one variable. Assume we have k different variables $x_1, \ldots, x_k$.

A variable y is a **linear combination** if it is related to several other variables $x_1, \ldots, x_k$ by the formula

$$y = c_0 + c_1 x_1 + c_2 x_2 + \ldots + c_k x_k$$

c_0 is the intercept
$c_1, \ldots, c_k$ are coefficients

Portfolios

What is a portfolio?

Why do investors create portfolios?

We would like to understand average return of a portfolio and its risk.

How is the risk of a portfolio related to financial assets that it contains?

Portfolios

Suppose you have \$100 to invest.

Let x_1 be the return on asset 1. If $x_1 = 0.1$ and you put all your money into asset 1, then you will have $\$100 \cdot (1 + 0.1) = \$110$ at the end of the period.

Let x_2 be the return on asset 2. If $x_2 = 0.15$ and you put all your money into asset 2, then you will have $\$100 \cdot (1 + 0.15) = \$115$ at the end of the period.

What happens if you put $\frac{1}{2}$ your money in asset 1 and $\frac{1}{2}$ your money in asset 2?

Portfolios

At the end of the period you will have

$$0.5 * 100 * (1 + 0.1) + 0.5 * 100 * (1 + 0.15) = 100 * [1 + (.5 * .1) + (.5 * .15)]$$
$$55 \quad + \quad 57.5 = \$112.50$$

In other words, if we put $\frac{1}{2}$ of our money in asset 1 and $\frac{1}{2}$ in asset 2 the return on the portfolio is

$$R_p = \frac{1}{2}x_1 + \frac{1}{2}x_2$$

The return on the portfolio is a **linear combination** of the returns on the individual assets.

In general, suppose you have $\$M$ dollars to invest in two assets with returns x_1 and x_2.

Let w_1 be the fraction of your wealth that you choose to put in x_1.

Assume that our portfolio weights sum to one, $w_1 + w_2 = 1$.

$$\begin{aligned} w_1 M(1 + x_1) + w_2 M(1 + x_2) &= M[w_1 + w_2 + (w_1 x_1) + (w_2 x_2)] \\ &= M[1 + w_1 x_1 + w_2 x_2] \end{aligned}$$

The return on our portfolio is

$$R_p = w_1 x_1 + w_2 x_2$$

Portfolios with m assets

Suppose we have m possible assets.

Let x_i denote the return on the ith asset.

Let w_i denote the percentage of wealth invested in the ith asset.

Then, the return on the portfolio is:

$$R_p = \sum_{i=1}^{m} w_i x_i$$

The return on the portfolio is a linear combination of individual asset returns, where the coefficients are equal to the fraction of wealth invested.

Example: Country returns

Consider building a portfolio using monthly country returns data and the two variables are Hong Kong and USA.

We place $\frac{1}{2}$ of our wealth in each asset.

```
countryReturn_df = read.csv("CountryMonthlyReturns.csv")
port_df = countryReturn_df[,c("usa", "honkong")]
port_df$port1 = 0.5*port_df$honkong + 0.5*port_df$usa
head(port_df)
```

```
##     usa honkong  port1
## 1  0.04    0.02  0.030
## 2 -0.03    0.06  0.015
## 3  0.01    0.02  0.015
## 4  0.01   -0.03 -0.010
## 5  0.05    0.08  0.065
## 6  0.00    0.02  0.010
```

The sample means of Hong Kong, USA, and the portfolio.

```
mean(port_df$honkong)
```

```
## [1] 0.021028
```

```
mean(port_df$usa)
```

```
## [1] 0.013458
```

```
mean(port_df$port1)
```

```
## [1] 0.017243
```

Note that the sample mean of the portfolio is

$$\begin{aligned}\overline{R}_p &= w_1 \cdot \overline{\texttt{honkong}} + w_2 \cdot \overline{\texttt{usa}} \\ &= 0.5 \cdot \overline{\texttt{honkong}} + 0.5 \cdot \overline{\texttt{usa}} \\ &= 0.5 \cdot (0.021028) + 0.5 \cdot (0.013458) = 0.017243\end{aligned}$$

The sample variance and standard deviation of Hong Kong, USA, and the portfolio.

```
var(port_df$honkong)
```

```
## [1] 0.005215
```

```
var(port_df$usa)
```

```
## [1] 0.0011077
```

```
var(port_df$port1)
```

```
## [1] 0.0020959
```

```
sd(port_df$honkong)
```

```
## [1] 0.072215
```

```
sd(port_df$usa)
```

```
## [1] 0.033283
```

```
sd(port_df$port1)
```

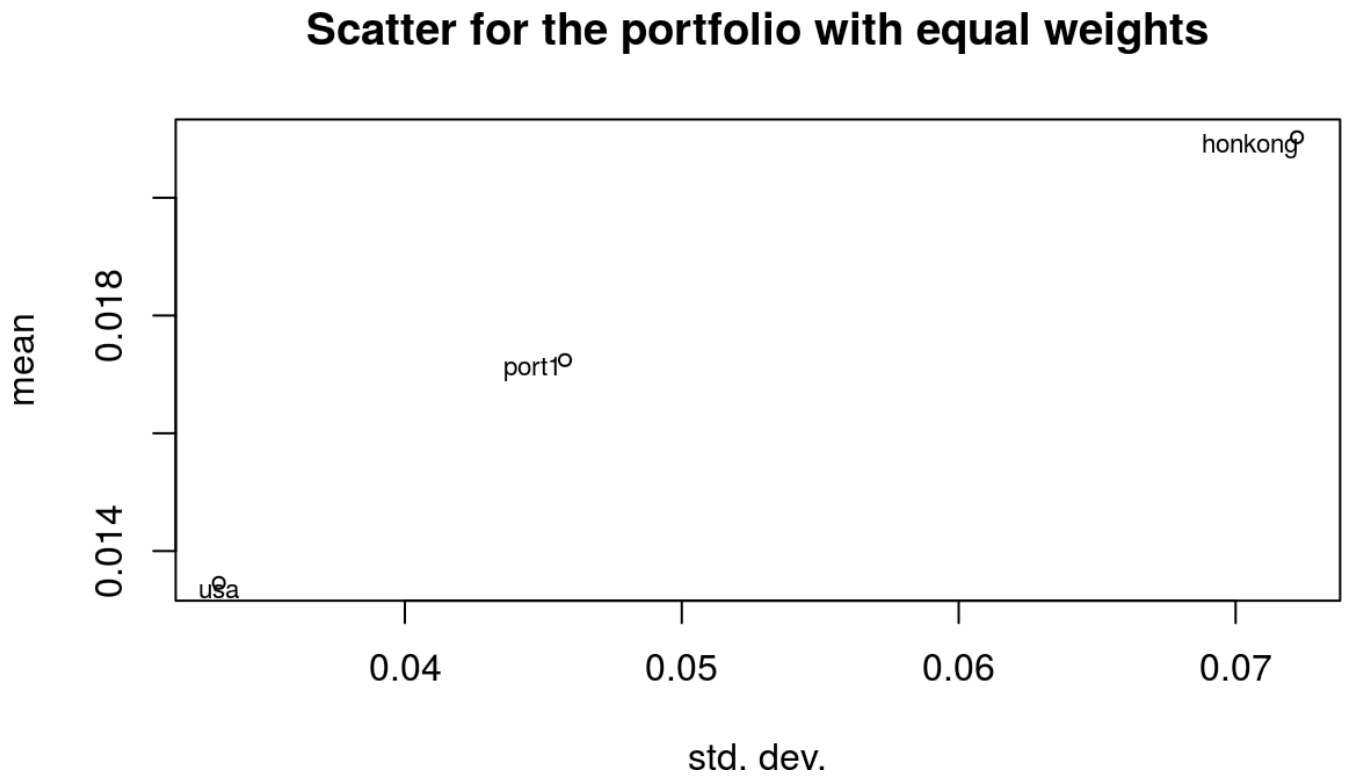
```
## [1] 0.045781
```

Note that the sample standard deviation of the portfolio, $s_{\texttt{port}}$ is **smaller** than half the sum of $s_{\texttt{honkong}}$ and $s_{\texttt{usa}}$.

$$s_{\texttt{port}} < 0.5 \cdot s_{\texttt{honkong}} + 0.5 \cdot s_{\texttt{usa}}$$

$$0.046 < 0.5 \cdot (0.072) + 0.5 \cdot (0.033) = 0.053$$

Diversification



The sample mean of the portfolio is half-way between the sample mean of Hong Kong and USA.

The sample standard dev is less than half-way between $s_{\texttt{usa}}$ and $s_{\texttt{honkong}}$.

Mean and variance of a linear combination

Suppose

$$y = c_0 + c_1 x_1 + c_2 x_2$$

Then

$$\bar{y} = c_0 + c_1 \bar{x}_1 + c_2 \bar{x}_2$$

$$s_y^2 = c_1^2 s_{x_1}^2 + c_2^2 s_{x_2}^2 + 2 c_1 c_2 s_{x_1 x_2}$$

Notice that when we have two variables we must take their covariance $s_{x_1 x_2}$ into account.

Reminder about correlations and covariances

Remember, we defined the sample correlation of two variables and as the sample covariance divided by the standard deviations

$$r_{x_1x_2} = \frac{s_{x_1x_2}}{s_{x_1}\, s_{x_2}}.$$

So, if we know the sample correlation and the sample standard deviations we can determine the sample covariance

$$s_{x_1x_2} = r_{x_1x_2}\, s_{x_1}\, s_{x_2}$$

If we know the sample standard deviations, then all we need to know is either the sample correlation **or** the sample covariance.

Example: Country returns

The sample covariance matrix is

```
cov(port_df)
```

```
##                usa   honkong     port1
## usa     0.0011077 0.0010304 0.0010691
## honkong 0.0010304 0.0052150 0.0031227
## port1   0.0010691 0.0031227 0.0020959
```

Using the formula

$$
\begin{aligned}
s^2_{\texttt{port}} &= w_1^2 s^2_{\texttt{honkong}} + w_2^2 s^2_{\texttt{usa}} + 2 w_1 w_2 s_{\texttt{honkong,usa}} \\
&= 0.5^2(0.0052) + 0.5^2(0.0011) + 2(0.5^2)(0.001) \\
&= 0.0021
\end{aligned}
$$

$$
s_{\texttt{port}} = \sqrt{0.0021} = 0.0458
$$

Understanding the contribution of covariance

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

Explore sample standard deviation of portfolio returns when

- assets are uncorrelated
- assets have strong positive correlation
- assets have strong negative correlation

Mean and variance of a linear combination

Three right hand side variables x_1, x_2, and x_3.

Suppose

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3$$

Then

$$\bar{y} = c_0 + c_1 \bar{x}_1 + c_2 \bar{x}_2 + c_3 \bar{x}_3$$

$$s_y^2 = c_1^2 s_{x_1}^2 + c_2^2 s_{x_2}^2 + c_3^2 s_{x_3}^2$$

$$+ 2 \left[c_1 c_2 s_{x_1 x_2} + c_1 c_3 s_{x_1 x_3} + c_2 c_3 s_{x_2 x_3} \right]$$

Notice that the variance formula now has 3 covariance terms.

Example: mutual funds

Vanguard Windsor (VWNDX), Pimco bond fund (PTTRX), DWS Global (LBF)
Portfolio: $\texttt{port} = 0.1\ \texttt{VWNDX} + 0.7\ \texttt{PTTRX} + 0.2\ \texttt{LBF}$

```
mfr_df = read.csv("mutualFundReturn.csv")
port_df = mfr_df[,c("VWNDX", "PTTRX", "LBF")]
port_df$port = 0.1*port_df$VWNDX + 0.7*port_df$PTTRX + 0.2*port_df$LBF
cov(port_df)
```

```
##            VWNDX      PTTRX        LBF       port
## VWNDX 2.6305e-03 7.2351e-05 0.00193761 0.00070121
## PTTRX 7.2351e-05 1.3748e-04 0.00024021 0.00015152
## LBF   1.9376e-03 2.4021e-04 0.00534785 0.00143148
## port  7.0121e-04 1.5152e-04 0.00143148 0.00046248
```

$$
\begin{aligned}
& s^2_{\texttt{port}} \\
&= w_1^2 \cdot s^2_{\texttt{VWNDX}} + w_2^2 \cdot s^2_{\texttt{PTTRX}} + w_3^2 \cdot s^2_{\texttt{LBF}} \\
&\quad + 2 \cdot w_1 \cdot w_2 \cdot s_{\texttt{VWNDX,PTTRX}} + 2 \cdot w_1 \cdot w_3 \cdot s_{\texttt{VWNDX,LBF}} + 2 \cdot w_2 \cdot w_3 \cdot s_{\texttt{PTTRX,LBF}} \\
&= 0.1^2 \cdot s^2_{\texttt{VWNDX}} + 0.7^2 \cdot s^2_{\texttt{PTTRX}} + 0.2^2 \cdot s^2_{\texttt{LBF}} \\
&\quad + 2 \cdot 0.1 \cdot 0.7 \cdot s_{\texttt{VWNDX,PTTRX}} + 2 \cdot 0.1 \cdot 0.2 \cdot s_{\texttt{VWNDX,LBF}} + 2 \cdot 0.7 \cdot 0.2 \cdot s_{\texttt{PTTRX,LBF}} \\
&= 4.6248e-04
\end{aligned}
$$

Further remarks on linear combinations

For linear combinations greater than 3 right-hand side variables (say k variables), the mean and variance formulas can be generalized.

The mean formula is: $\overline{y} = c_0 + \sum_{j=1}^{k} c_j \overline{x}_j$

The variance formulas take into account all pairwise combinations of the covariances.

I will not ask you to calculate by hand any formulas with more than 3 right-hand side variables.

If you take the portfolios class from the finance group, you will learn about building portfolios by taking linear combinations of assets. The goal is to choose the weights to build good portfolios that are on or close to the "efficient frontier."

Clustering

Similarity

Many statistical tools rely on similarity between observations.

If two observations are **similar** in some ways, they often share other characteristics as well.

We have already seen an example of predicting the price of a house based on its size.

Clustering

The main idea behind clustering is to find groups of observations that are similar to each other,

- but not so similar to other observations

Exploratory tool: we can identify new business opportunities - way to find interesting patterns in data

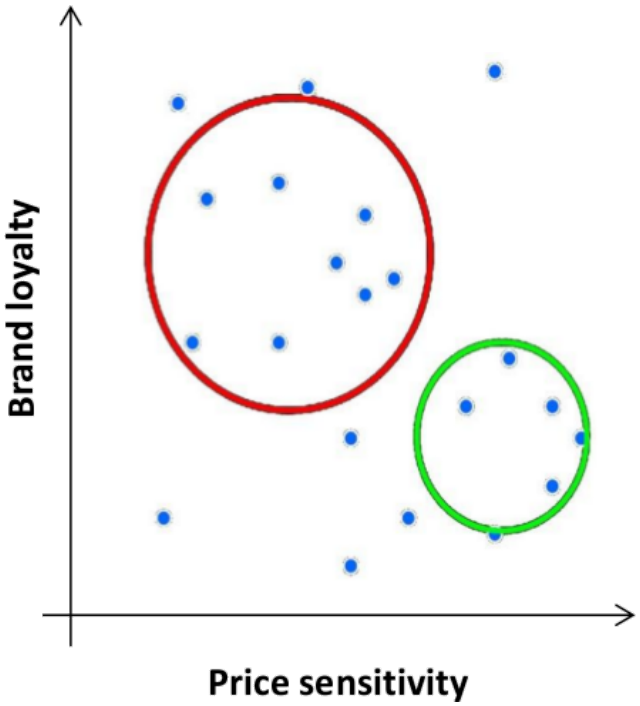
Can be used to perform data reduction.
Treat everyone in a group the same way.

Example: Market Segmentation

Cluster analysis is used in marketing to identify groups of *similar* customers in a market, that is, perform market segmentation.

Once we have groups, we can characterize them based on purchasing patterns. **Here, we often need domain expertize**.

How many groups do we need?
Clustering is an exploratory technique. Try different number of clusters and see what makes sense.



Example: shopping attitudes

Questionaire

- V1: Shopping is fun
- V2: Shopping is bad for your budget
- V3: I combine shopping with eating out
- V4: I try to get the best buys while shopping
- V5: I don’t care about shopping
- V6: You can save a lot of money by comparing prices

Responses: 1 = strongly disagree, 7 = strongly agree

Are there segments of shoppers with similar attitudes?
What are they?

```
(shopping_df = read.csv("shopping.csv", row.names=1))
```

```
##    V1 V2 V3 V4 V5 V6
## 1   6  4  7  3  2  3
## 2   2  3  1  4  5  4
## 3   7  2  6  4  1  3
## 4   4  6  4  5  3  6
## 5   1  3  2  2  6  4
## 6   6  4  6  3  3  4
## 7   5  3  6  3  3  4
## 8   7  3  7  4  1  4
## 9   2  4  3  3  6  3
## 10  3  5  3  6  4  6
## 11  1  3  2  3  5  3
## 12  5  4  5  4  2  4
## 13  2  2  1  5  4  4
## 14  4  6  4  6  4  7
## 15  6  5  4  2  1  4
## 16  3  5  4  6  4  7
## 17  4  4  7  2  2  5
## 18  3  7  2  6  4  3
## 19  4  6  3  7  2  7
## 20  2  3  2  4  7  2
```

```
plot(x=shopping_df[,1], y=shopping_df[,2], main="Shopping attitudes",
     xlab="Shopping is fun", ylab="Shopping is bad for your budget")
```

Cluster shoppers into three groups.

```
# setting the seed will allow us to reproduce results
# otherwise the result will change every time
set.seed(1)
(grpShopper = kmeans(shopping_df, centers=3, nstart=100))
```

```
## K-means clustering with 3 clusters of sizes 6, 8, 6
##
## Cluster means:
##       V1     V2     V3    V4    V5     V6
## 1 3.5000 5.8333 3.3333 6.000 3.500 6.0000
## 2 5.7500 3.6250 6.0000 3.125 1.875 3.8750
## 3 1.6667 3.0000 1.8333 3.500 5.500 3.3333
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  2  3  2  1  3  2  2  2  3  1  3  2  3  1  2  1  2  1  1  3
##
## Within cluster sum of squares by cluster:
## [1] 25.167 34.000 20.500
##  (between_SS / total_SS =  75.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

`grp$cluster` holds cluster assignments for each observation.

```
plot(x=shopping_df[,1], y=shopping_df[,2], main="Three clusters",
     col=grpShopper$cluster+1,
     xlab="Shopping is fun", ylab="Shopping is bad for your budget")
points(x=grpShopper$centers[,1], y=grpShopper$centers[,2], col=c(2,3,4), pch=3)
```

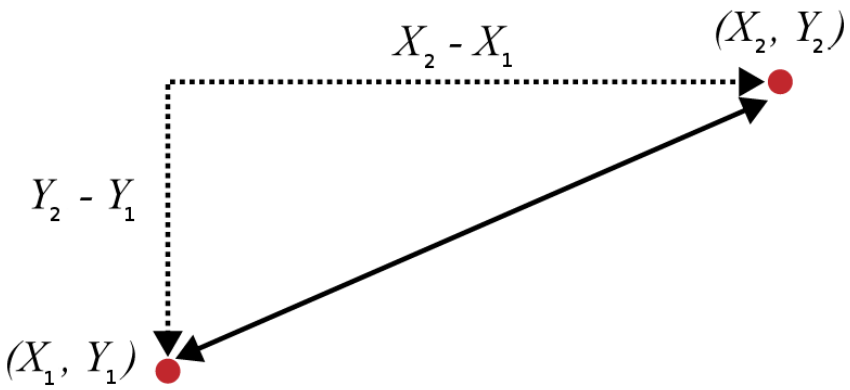
A closer look at centers

	1	2	3
V1: Shopping is fun	3.50	5.75	1.67
V2: Shopping is bad for your budget	5.83	3.63	3.00
V3: I combine shopping with eating out	3.33	6.00	1.83
V4: I try to get the best buys while shopping	6.00	3.13	3.50
V5: I don’t care about shopping	3.50	1.88	5.50
V6: You can save a lot of money by comparing prices	6.00	3.88	3.33

1 = strongly disagree, 7 = strongly agree

How to measure similarity?

Distance as a measure of similarity



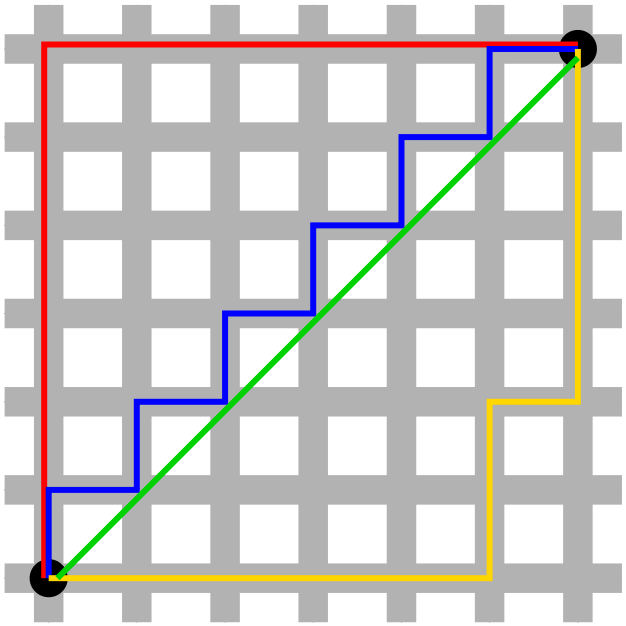
$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

In general:

$$\sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \ldots + (Z_1 - Z_2)^2}$$

This is called **Euclidean distance**.

Another distance – Manhattan



In general:

$$|A_1 - A_2| + |B_1 - B_2| + \ldots + |Z_1 - Z_2|$$

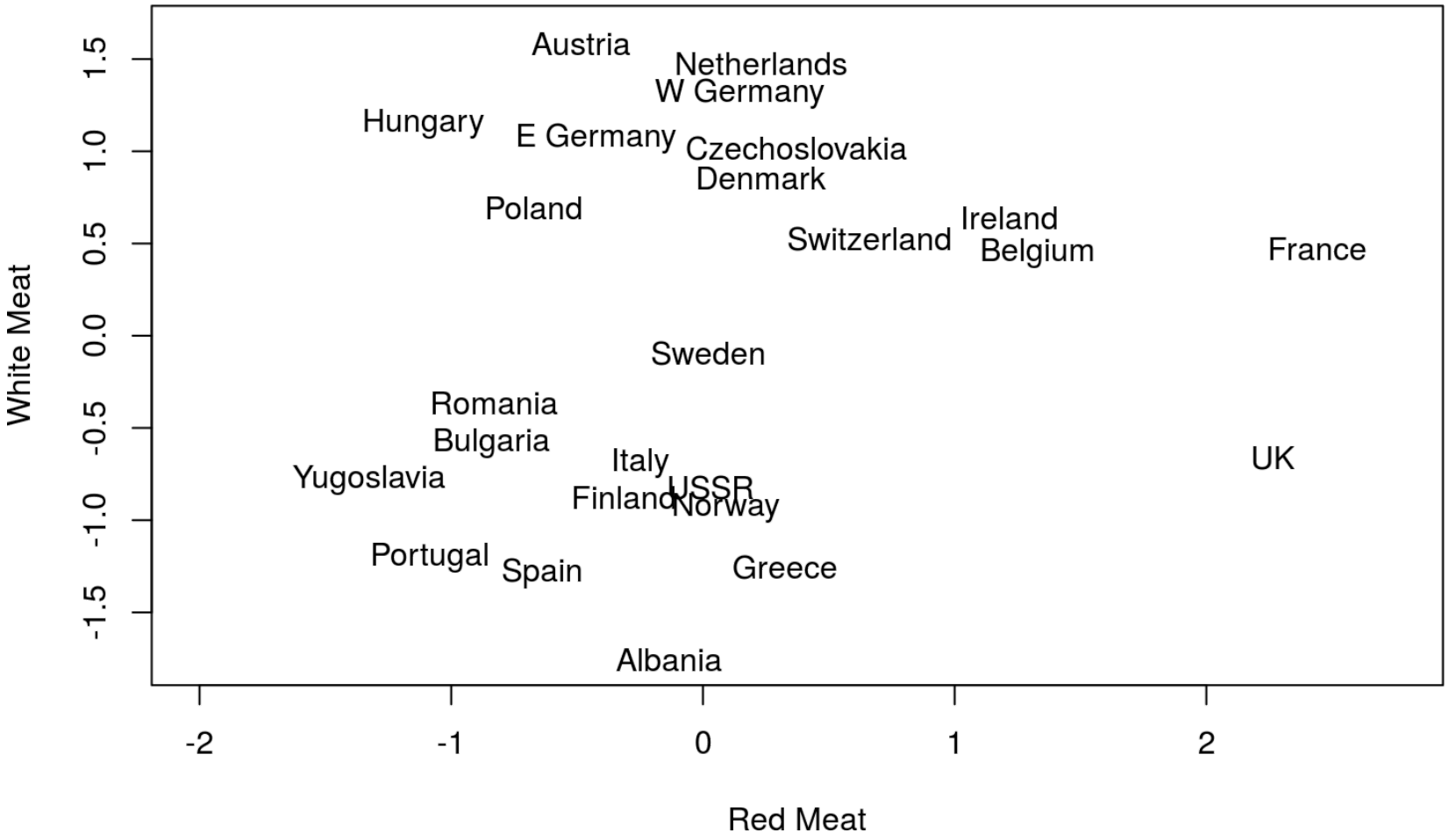
Example: Protein Consumption in Europe

Protein consumption by country, in grams per person per day

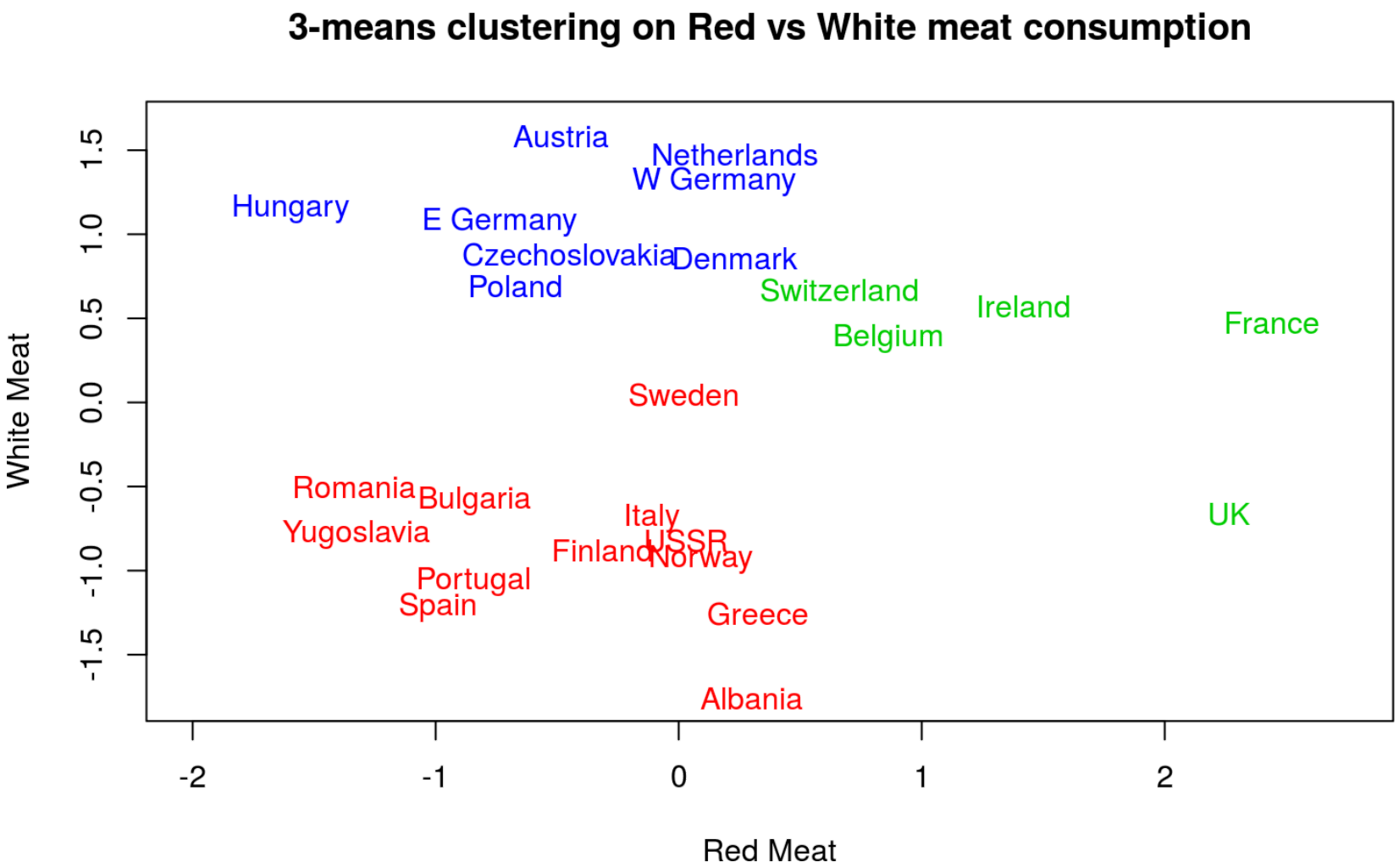
```
food_df = read.csv("protein.csv", row.names=1) # 1st column is country name
head(food_df)
```

```
##                RedMeat WhiteMeat Eggs Milk Fish Cereals Starch Nuts Fr.Veg
## Albania           10.1       1.4  0.5  8.9  0.2    42.3    0.6  5.5    1.7
## Austria            8.9      14.0  4.3 19.9  2.1    28.0    3.6  1.3    4.3
## Belgium           13.5       9.3  4.1 17.5  4.5    26.6    5.7  2.1    4.0
## Bulgaria           7.8       6.0  1.6  8.3  1.2    56.7    1.1  3.7    4.2
## Czechoslovakia     9.7      11.4  2.8 12.5  2.0    34.3    5.0  1.1    4.0
## Denmark           10.6      10.8  3.7 25.0  9.9    21.9    4.8  0.7    2.4
```

```
# scale the data so that every column has sample mean equal to zero
# and variance equal to 1
food_scaled = scale(food_df)
plot(food_scaled[,"RedMeat"], food_scaled[,"WhiteMeat"], xlim=c(-2,2.75),
     type="n", xlab="Red Meat", ylab="White Meat")
pointLabel(food_scaled[,"RedMeat"], food_scaled[,"WhiteMeat"],
           labels=rownames(food_scaled))
```



```r
set.seed(1)
## first, consider just Red and White meat clusters
grpMeat <- kmeans(food_scaled[,c("WhiteMeat","RedMeat")], centers=3, nstart=100)
plot(food_scaled[,"RedMeat"], food_scaled[,"WhiteMeat"], xlim=c(-2,2.75),
     type="n", xlab="Red Meat", ylab="White Meat",
     main="3-means clustering on Red vs White meat consumption")
pointLabel(food_scaled[,"RedMeat"], food_scaled[,"WhiteMeat"],
           labels=rownames(food_scaled), col=grpMeat$cluster+1)
```



Consumption is in units of standard deviation from the mean.

Cluster 1

```
rownames(food_df)[grpMeat$cluster==1]
```

```
##  [1] "Albania"    "Bulgaria"   "Finland"    "Greece"     "Italy"
##  [6] "Norway"     "Portugal"   "Romania"    "Spain"      "Sweden"
## [11] "USSR"       "Yugoslavia"
```

Cluster 2

```
rownames(food_df)[grpMeat$cluster==2]
```

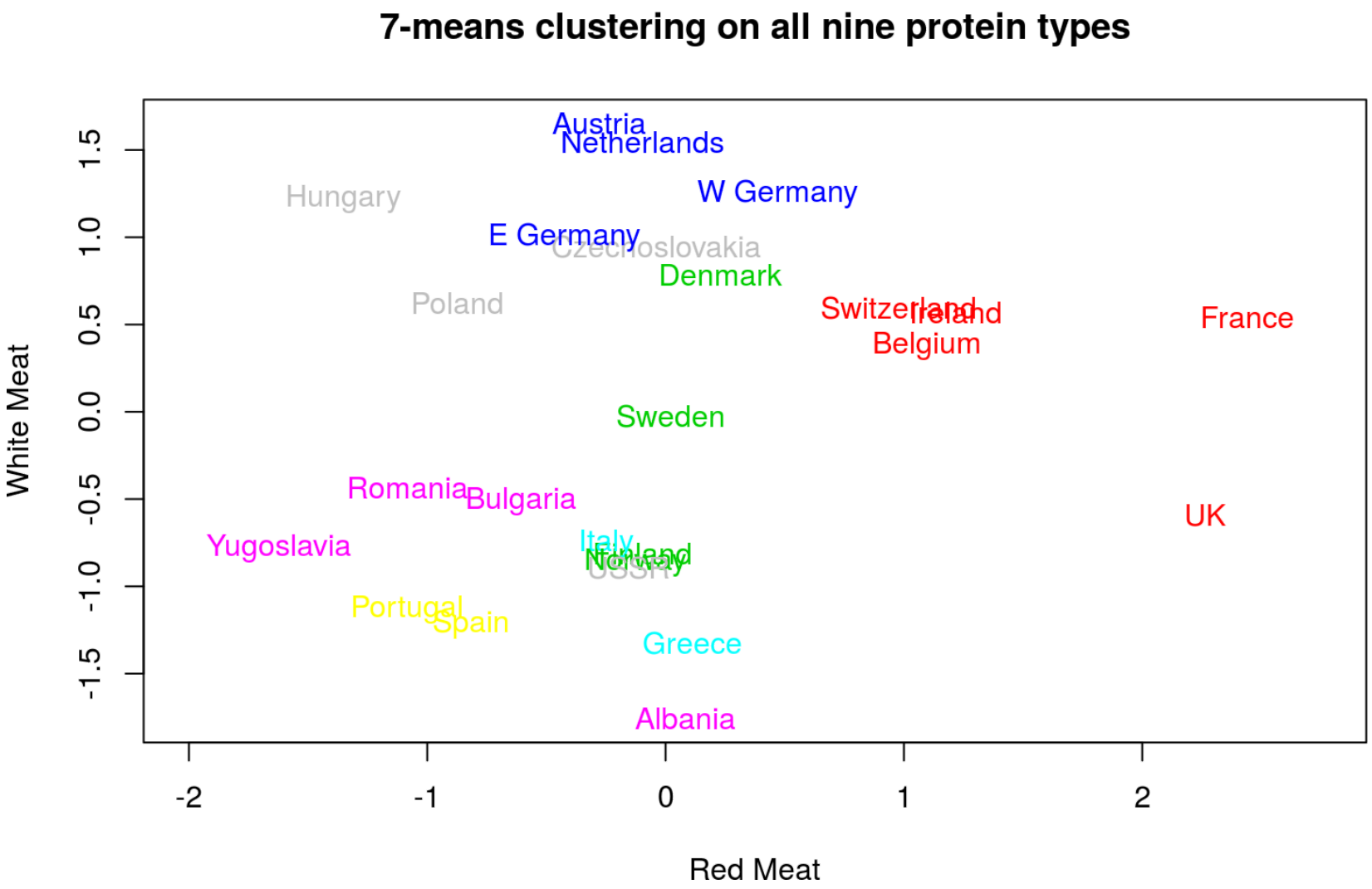
```
## [1] "Belgium"     "France"      "Ireland"     "Switzerland" "UK"
```

Cluster 3

```
rownames(food_df)[grpMeat$cluster==3]
```

```
## [1] "Austria"        "Czechoslovakia" "Denmark"        "E Germany"
## [5] "Hungary"        "Netherlands"    "Poland"         "W Germany"
```

```
grpProtein <- kmeans(food_scaled, centers=7, nstart=100)
plot(food_scaled[,"RedMeat"], food_scaled[,"WhiteMeat"], xlim=c(-2,2.75),
     type="n", xlab="Red Meat", ylab="White Meat",
     main="7-means clustering on all nine protein types")
text(food_scaled[,"RedMeat"], food_scaled[,"WhiteMeat"],
     labels=rownames(food_scaled), col=grpProtein$cluster+1)
```



Plotting the red vs white plane, but clustering on all variables.

```
rownames(food_df)[grpProtein$cluster==1]
```

```
## [1] "Belgium"     "France"      "Ireland"     "Switzerland" "UK"
```

```
rownames(food_df)[grpProtein$cluster==2]
```

```
## [1] "Denmark" "Finland" "Norway"  "Sweden"
```

```
rownames(food_df)[grpProtein$cluster==3]
```

```
## [1] "Austria"     "E Germany"   "Netherlands" "W Germany"
```

```
rownames(food_df)[grpProtein$cluster==4]
```

```
## [1] "Greece" "Italy"
```

```
rownames(food_df)[grpProtein$cluster==5]
```

```
## [1] "Albania"    "Bulgaria"   "Romania"    "Yugoslavia"
```

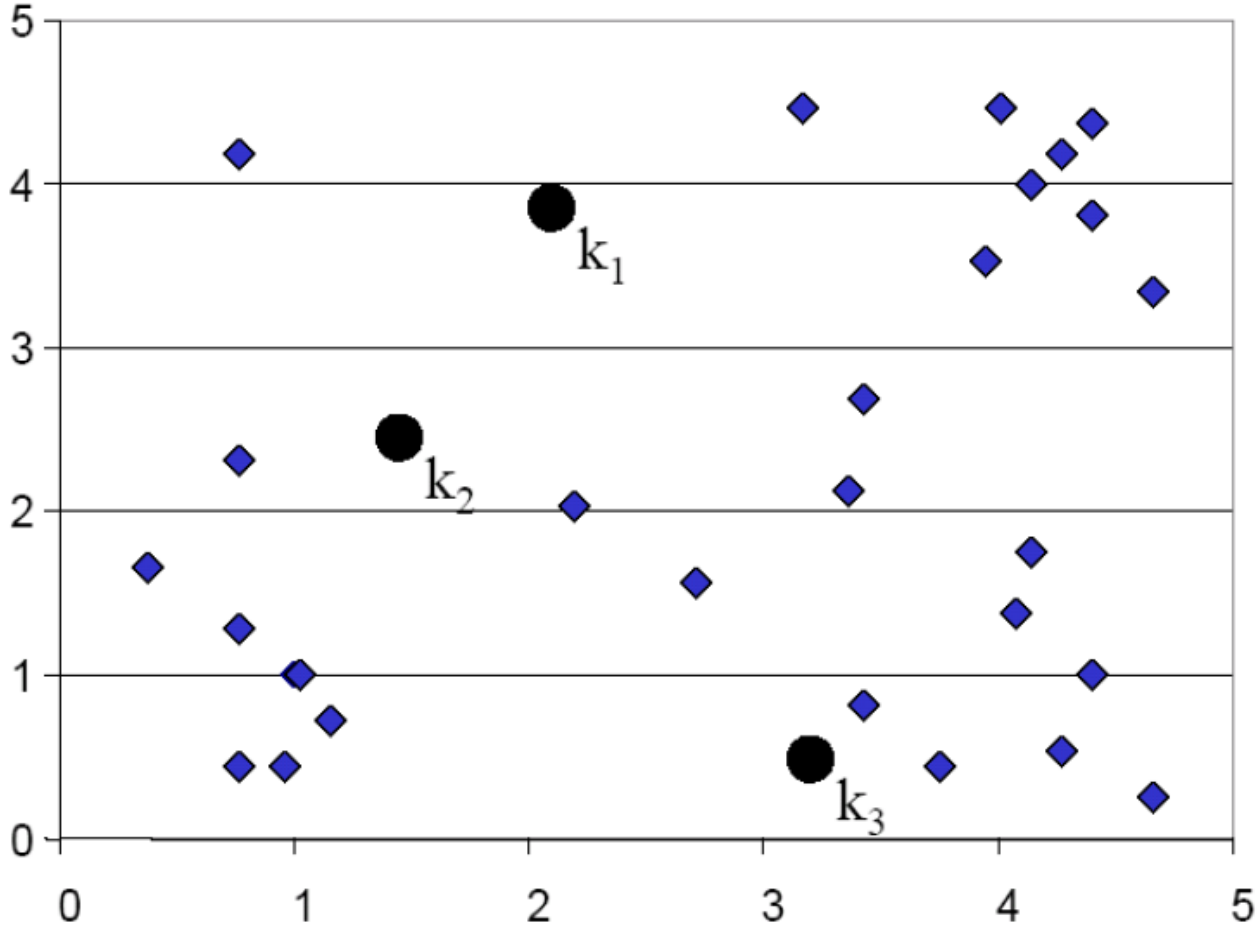
```
rownames(food_df)[grpProtein$cluster==6]
```

```
## [1] "Portugal" "Spain"
```

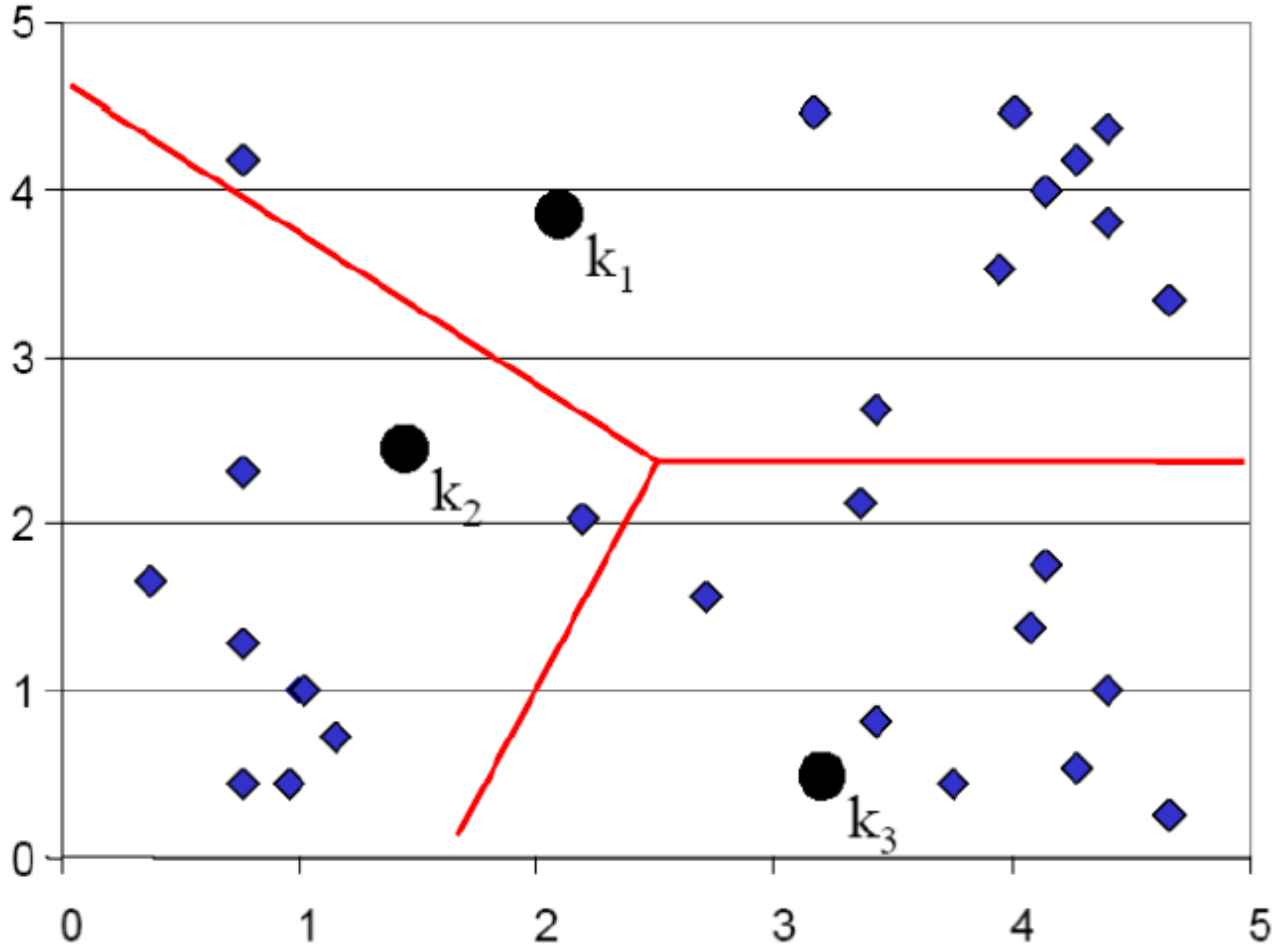
```
rownames(food_df)[grpProtein$cluster==7]
```

```
## [1] "Czechoslovakia" "Hungary"        "Poland"         "USSR"
```

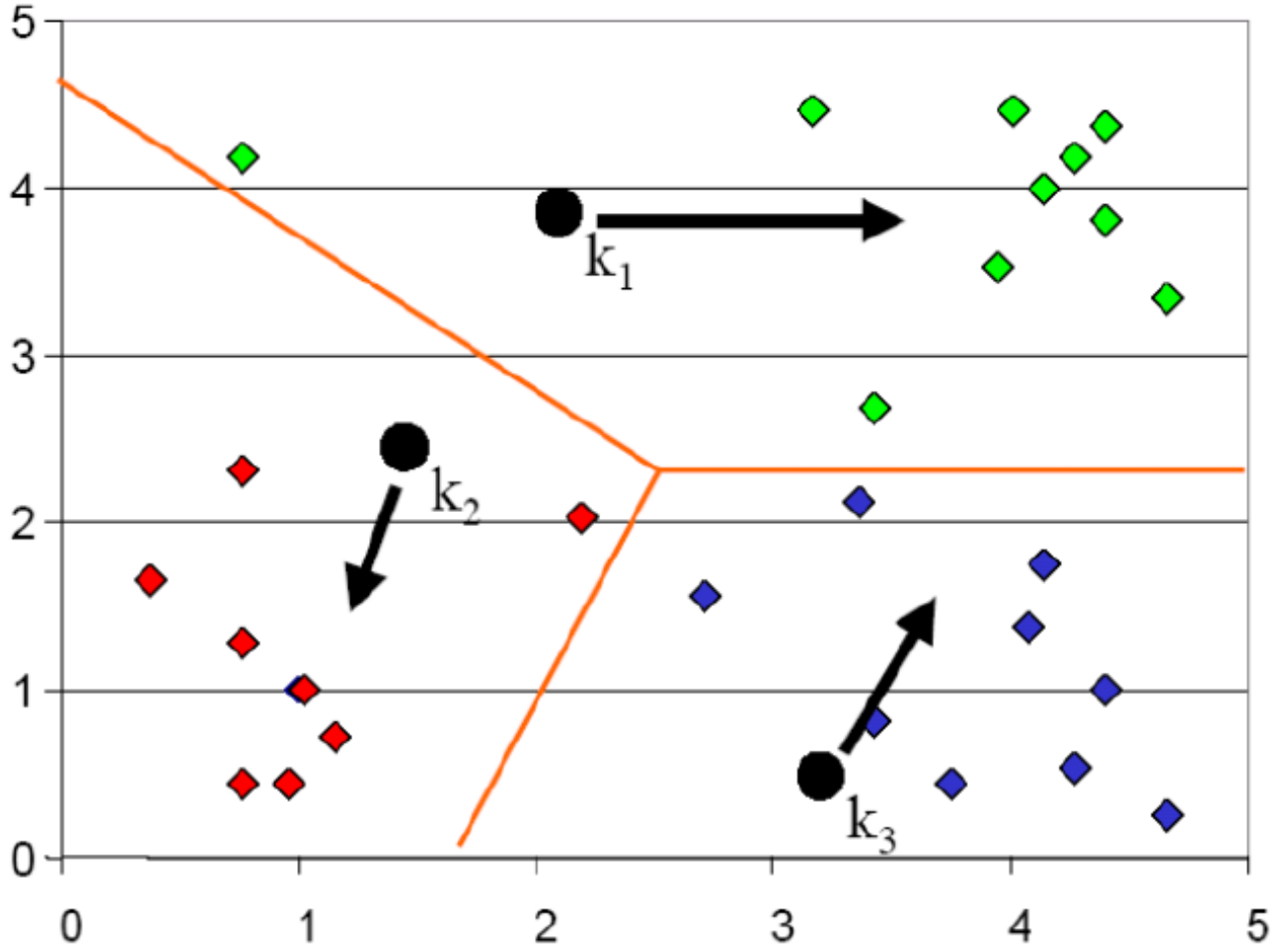
How does clustering actually work?



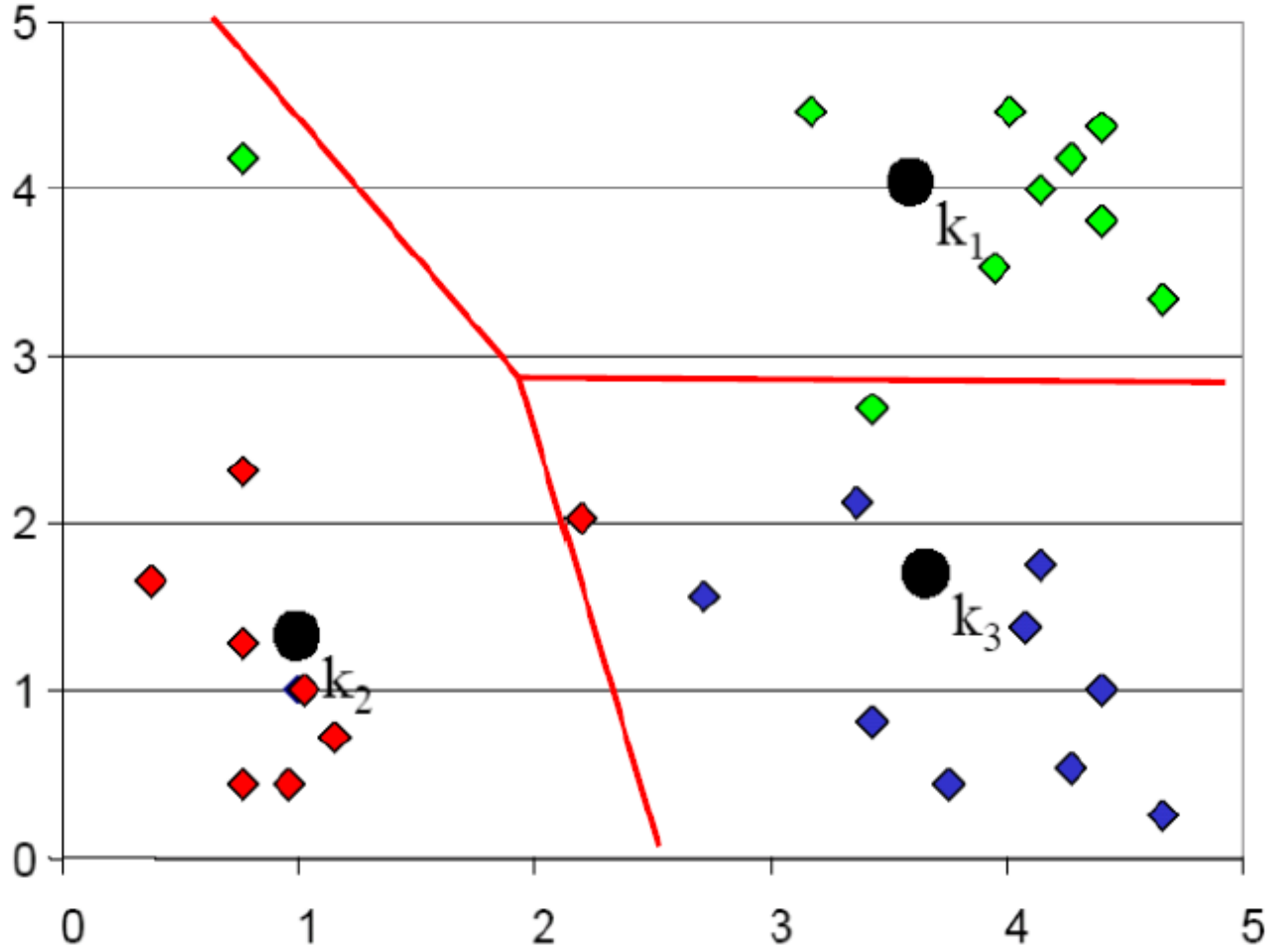
How does clustering actually work?



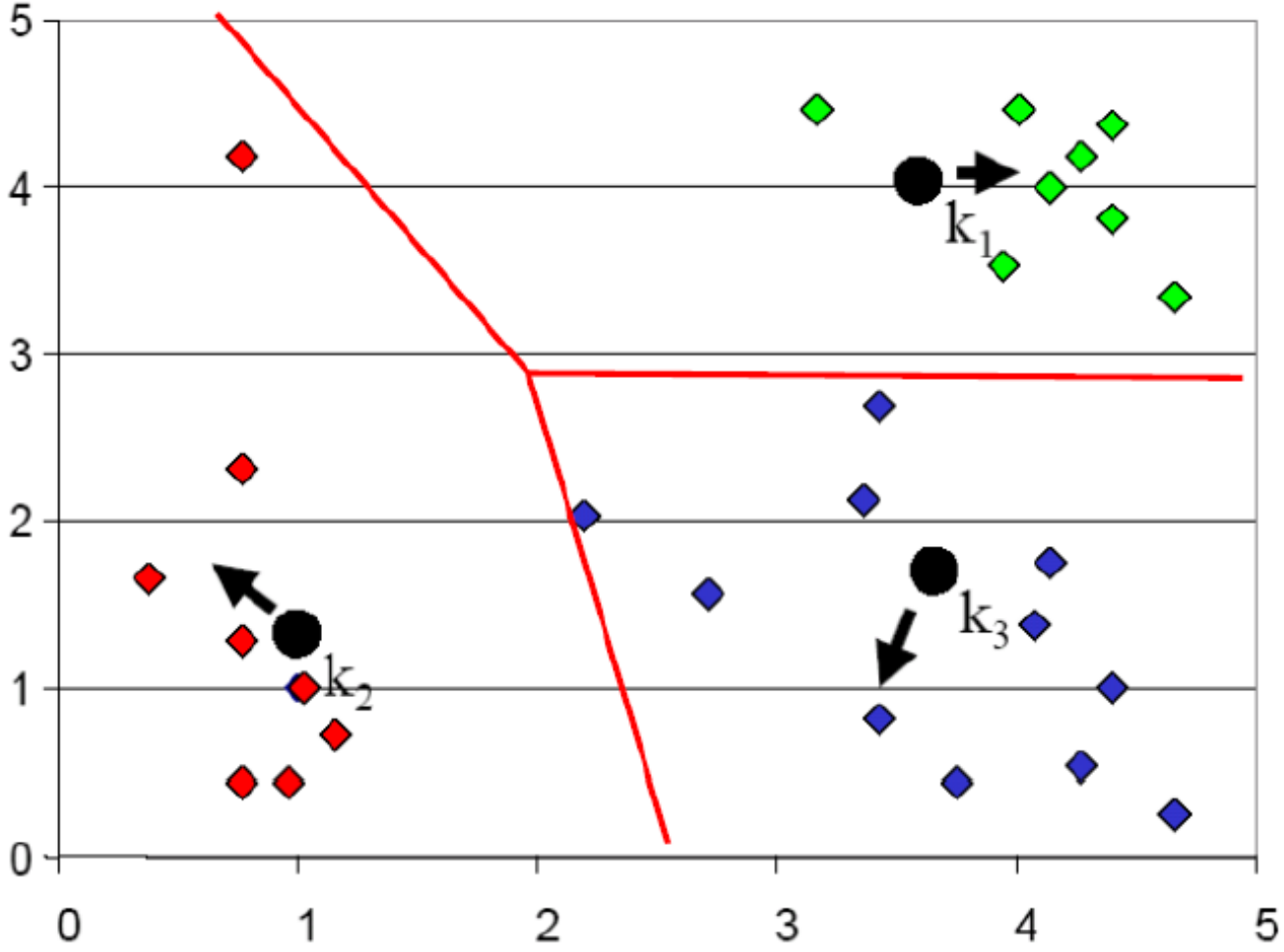
How does clustering actually work?



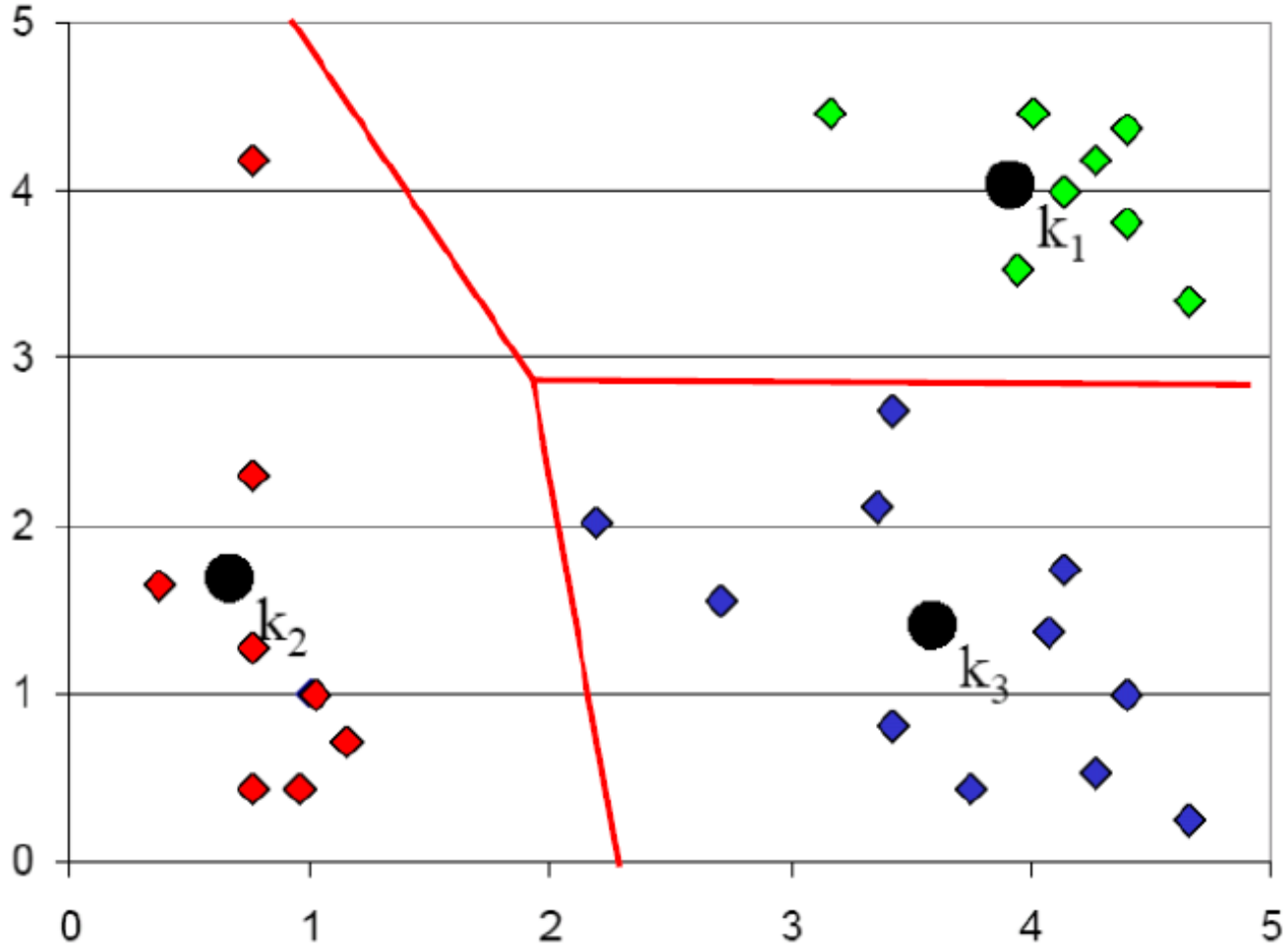
How does clustering actually work?



How does clustering actually work?



How does clustering actually work?



Discussion of clustering

We do not always look to solve a particular business problem. Sometimes we just want to explore data and find new opportunities.

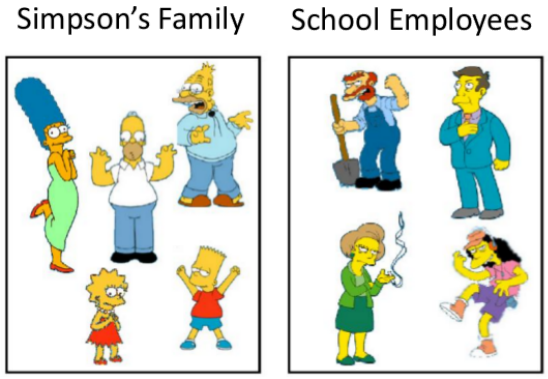
Cluster Analysis allows us to simplify across respondents.

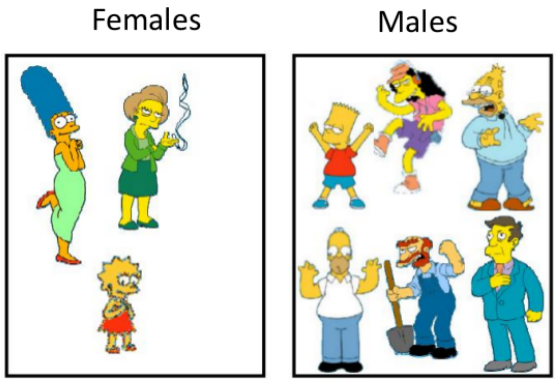
When used effectively, it can guide our business strategy.

We often need business expertise to understand whether output of clustering is meaningful.

We can get different results by changing similarity measure.

- Clustering is subjective





Similarities in business

Advertisers want to serve online ads to consumers that are similar to their current good customers.

A cable company wants to identify customers similar to those that have previously cancelled the service and incentivize them to keep paying for the subscription.

Amazon and Netflix use similarity to provide recommendations.

- "People who like X also like Y"
- "Customers with your browsing history have also looked at …"

Medical doctors provide diagnosis based on previous similar cases.

Lawyers argue cases by citing legal precedence.

Machine learning and statistical systems help doctors and lawyers with such case-based reasoning.

Example: Single malt Scotch whiskey

Dataset consists of 109 Scotch whiskeys (**scotch.csv**)

Tasting notes:

- **color** (14 values): yellow, very pale, pale, pale gold, gold, old gold, full gold, amber, etc.
- **nose** (12 values): aromatic, peaty, sweet, light, fresh, dry, grassy, etc.
- **body** (8 values): soft, medium, full, round, smooth, light, firm, oily
- **palate** (15 values): full, dry, sherry, big, fruity, grassy, smoky, salty, etc.
- **finish** (19 values): full, dry, warm, light, smooth, clean, fruity, grassy, etc

Total of 68 characteristics. We denote each by a 0/1 variable.

Whether Scotch has a particular characteristic or not.

Example: Single malt Scotch whiskey

We tried Ardbeg and really liked it.

Can we find other Scotches similar to it?

Tasting notes for Ardbeg:

- *Color*: sherry
- *Nose*: peat, dry, sea
- *Body*: medium, full, light, firm
- *Palate*: sweet
- *Finish*: salt

Scotch	Dist	Description
Ardbeg	0.00	sherry; peat, dry, sea; medium, full, light, firm; sweet; salt
Bowmore	0.43	old gold; peat, dry, grass, sea; medium, light; sweet; salt, quick
Inchgower	0.61	gold; aroma, peat, sweet, dry, sea, rich; medium, smooth; sweet; salt; salt, long
Pulteney	0.61	amber; fresh, dry, sea; light, firm; smooth, sweet, spice, salt; warm, salt, long
Bunnahabhain	0.62	gold; fresh, sea; medium, light, firm; clean, fruit, sweet; full
Glenury Royal	0.63	bronze; peat, dry; medium; light, firm; dry, smoke, sweet; dry, fruit, smoke, sweet very long

More examples

How to find fraudulent accounts on Facebook?

How to determine credit card limit for a new customer?

How to determine monthly premium for car insurance?

How to identify groups of similar users?