Spring 2019 MCB 432

Dimensionality Reduction in Large Datasets

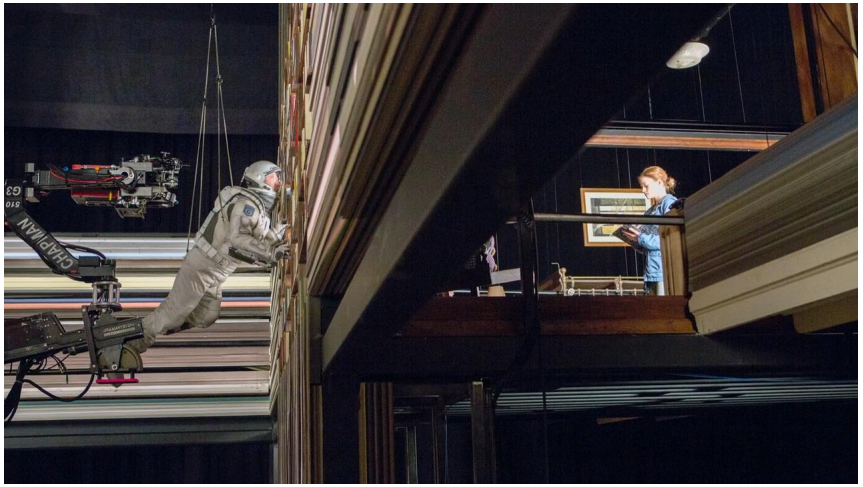
Feb 12, 2019

Presented by – Shriyaa Mittal

Lesson plan for today

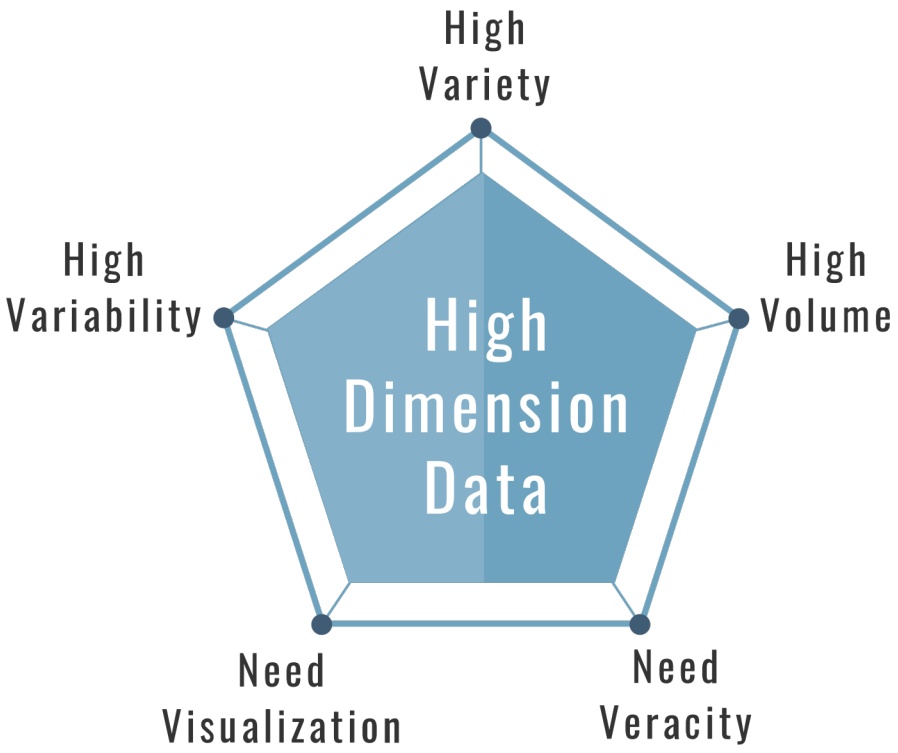
1. Identify the need for dimensionality reduction
2. List some common dimensionality reduction methods
3. Discuss Principal Component Analysis (PCA)
4. Demonstrate the math behind PCA
5. Use R to perform PCA on a dataset about eating habits in UK
6. Use our PCA model to predict “classes” for unknown data
7. Use R to perform principal coordinates analysis (PCoA) to recreate a map of the United States
8. Understand the limitation of dimensionality reduction

The Curse of Dimensionality



The Curse of Dimensionality

Refers to the problems that arise in analyzing large datasets.



The Curse of Dimensionality

The goal for dimensionality reduction is to find a low-dimensional representation of the data that retains as much information as possible.

- Space required to store the data is reduced
- Less dimensions lead to less computation time
- Some algorithms do not perform well when we have a large dimensions
- Remove redundant data
- It helps in visualizing data

The Curse of Dimensionality Reduction Methods!!

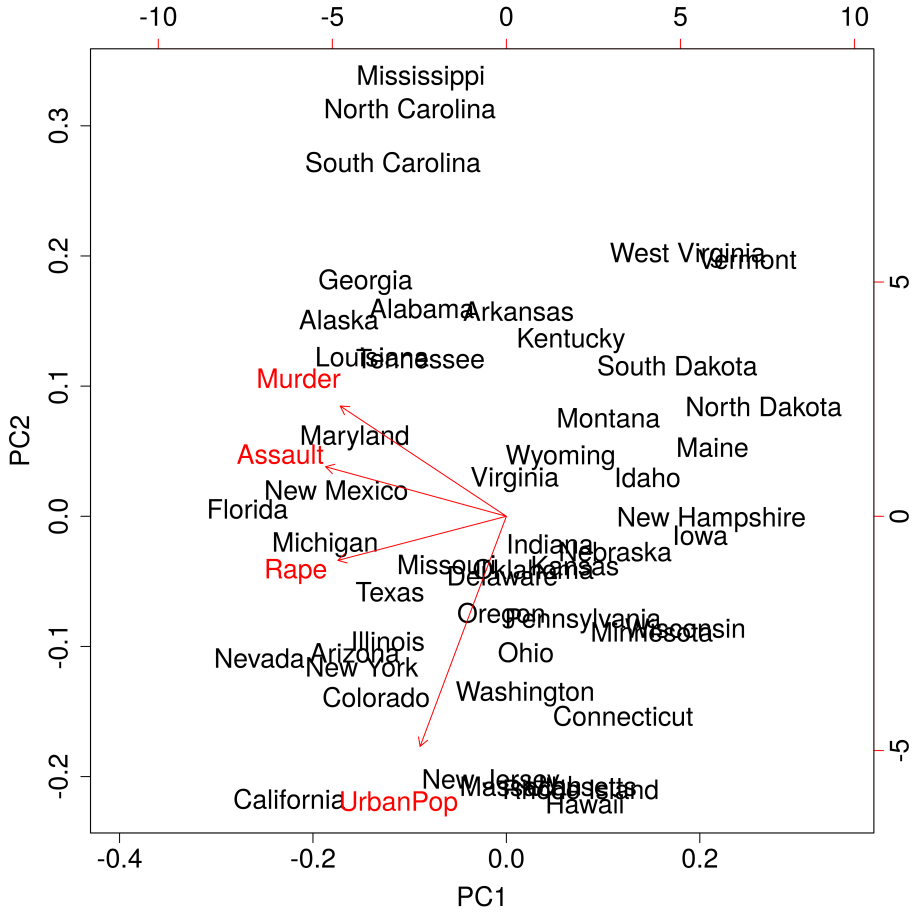
Decision Trees
Random forest
High correlation
Backward feature estimation
Factor analysis
Principal component analysis
Kernel PCA
Graph-based kernel PCA
Principal coordinate analysis
Linear discriminant analysis
Generalized discriminant analysis
Time-lagged independent component analysis
Autoencoders
Non-negative matrix factorization

Which one to use? Depends on your end goal.

Good resource: https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/

Remember this? We have done PCA before.

Principal component analysis (PCA) on USArrests dataset in R



How did we interpret this plot?

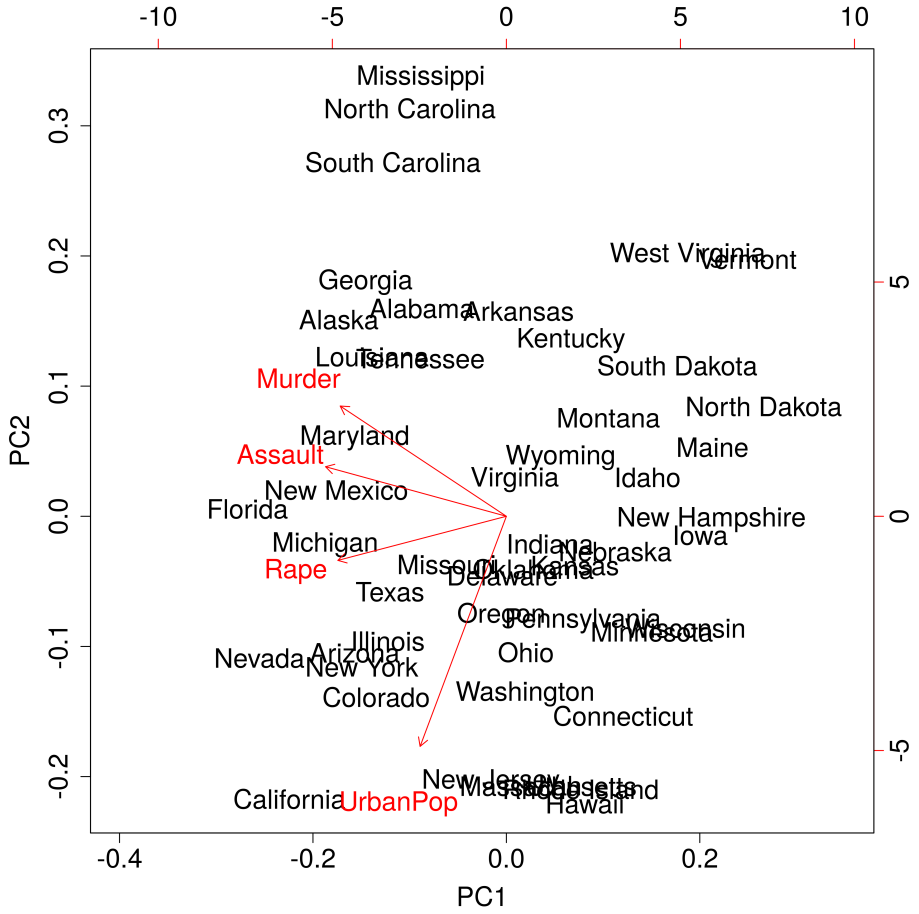
Red arrows represent variables and arrow direction represents the direction which explains the most variation.

The closeness of the red arrows variables indicates that they are correlated.

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

Remember this? We have done PCA before.

Principal component analysis (PCA) on USArrests dataset in R



How did we interpret this plot?

Red arrows represent variables and arrow direction represents the direction which explains the most variation.

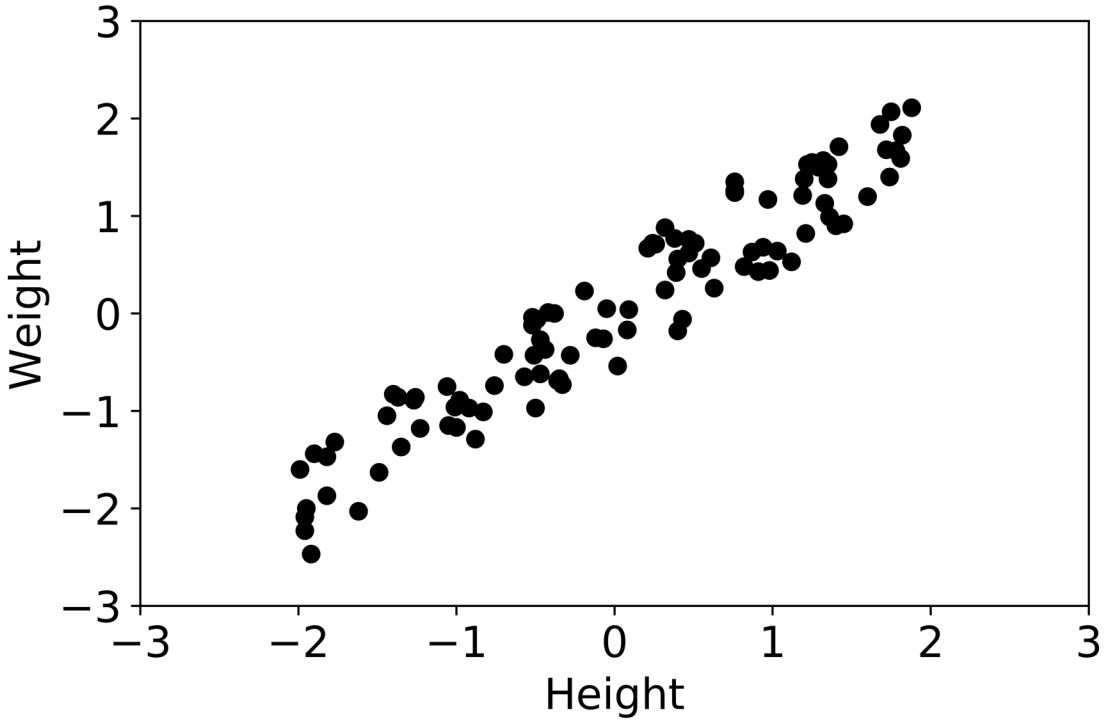
The closeness of the red arrows variables indicates that they are correlated.

Did we reduce dimensions here?

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

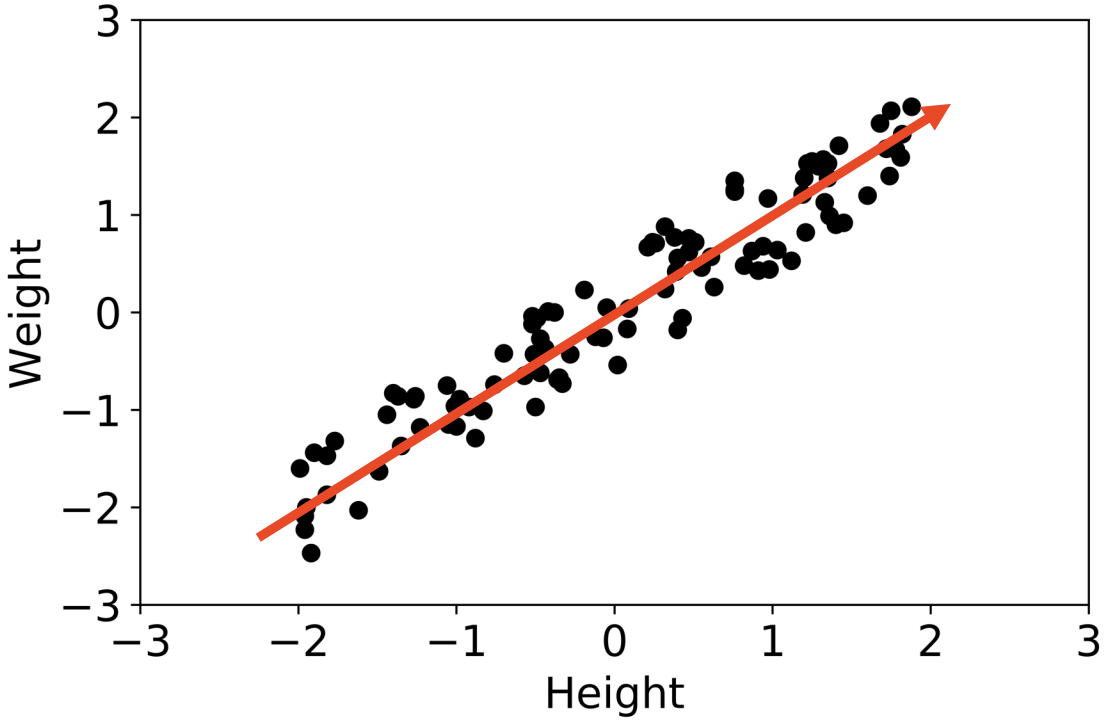
What is PCA?

Consider a data set of heights and weights of people



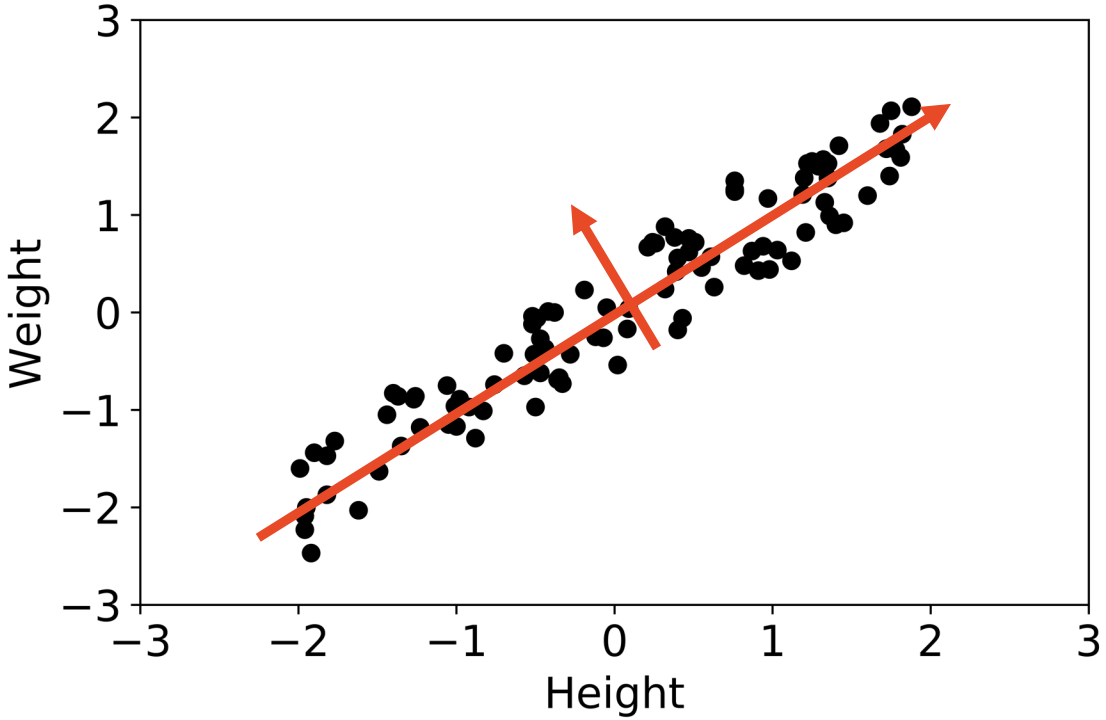
What is PCA?

Consider a data set of heights and weights of people



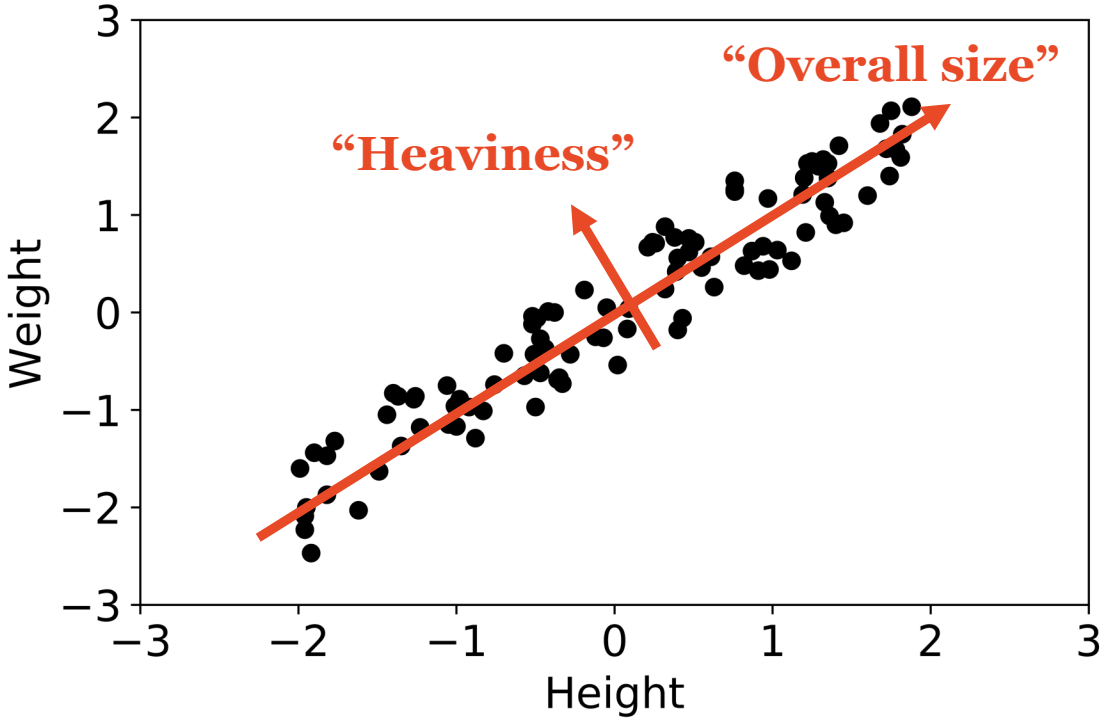
What is PCA?

Consider a data set of heights and weights of people



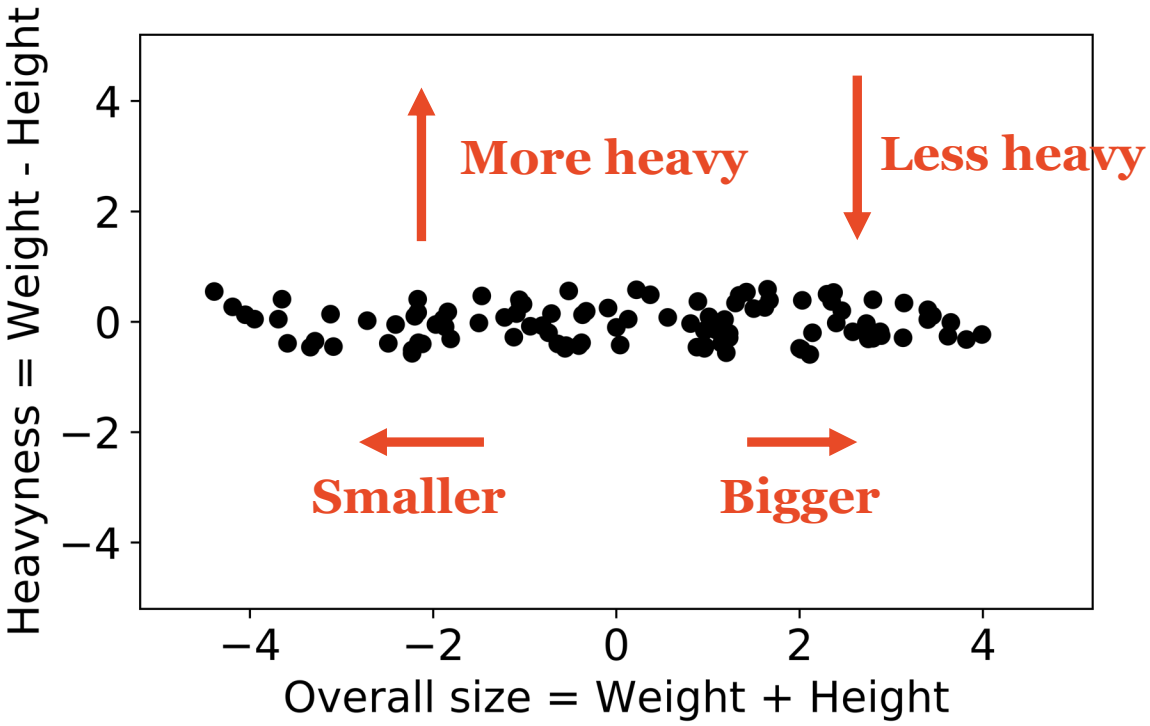
What is PCA?

Consider a data set of heights and weights of people



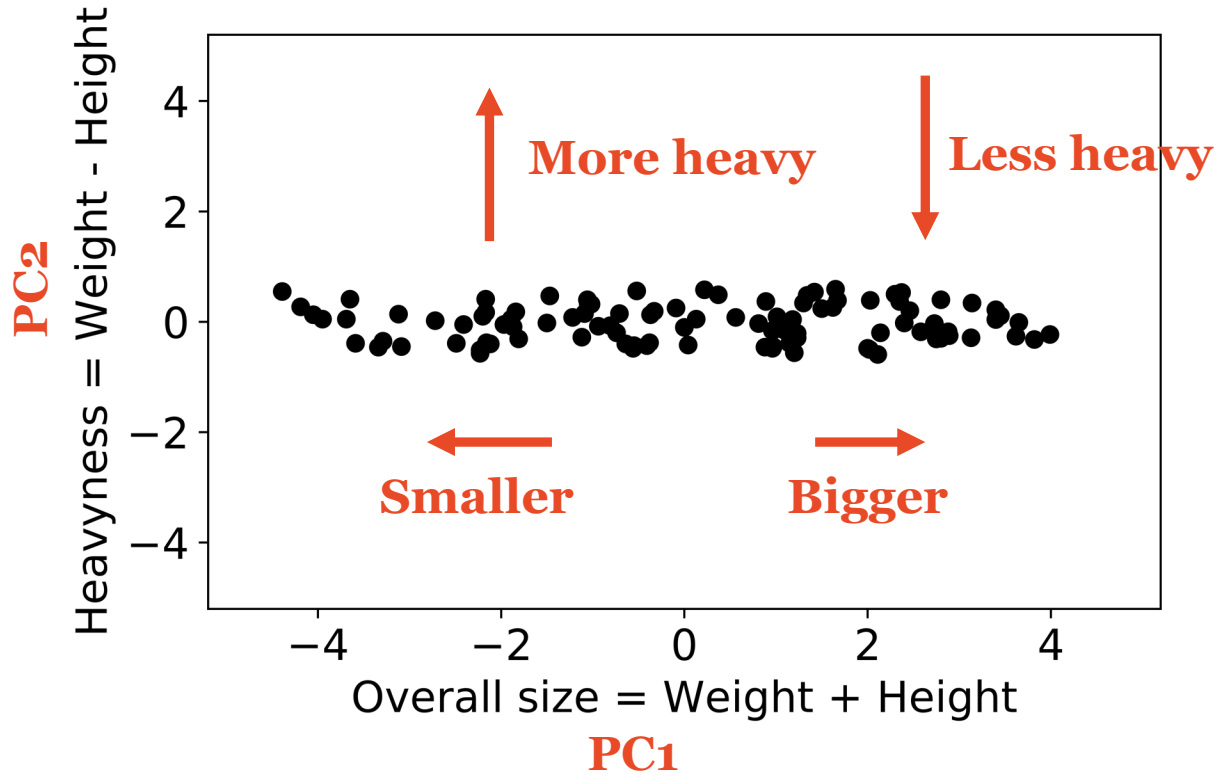
What is PCA?

Consider a data set of heights and weights of people, PCA on this data set reframes data in terms of overall size (small or big) and heaviness (less or more heavy)



What is PCA?

Consider a data set of heights and weights of people, PCA on this data set reframes data in terms of overall size (small or big) and heaviness (less or more heavy)



If we're going to only see the data along one dimension, it might be better to make that dimension the principal component (PC) with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data.

The math behind PCA

Height or Weight

Variance of one variable:

Mean of the variable

$$\mathrm{Var}(X) = \frac{1}{n}\sum_j (\bar{x} - x_j)^2 = \sigma_X^2$$

Covariance of two variables:

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_j (\bar{x} - x_j)(\bar{y} - y_j) = \sigma_{XY}^2$$

Height Weight

The math behind PCA

Covariance matrix of n variables $X_1 \ldots X_n$:

$$\mathbf{C} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix}$$

For 2 variables, say Height (X1) and Weight (X2), the C matrix will be 2 X 2

The math behind PCA

PCA diagonalizes the covariance matrix **C**:

diagonal matrix

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^T$$

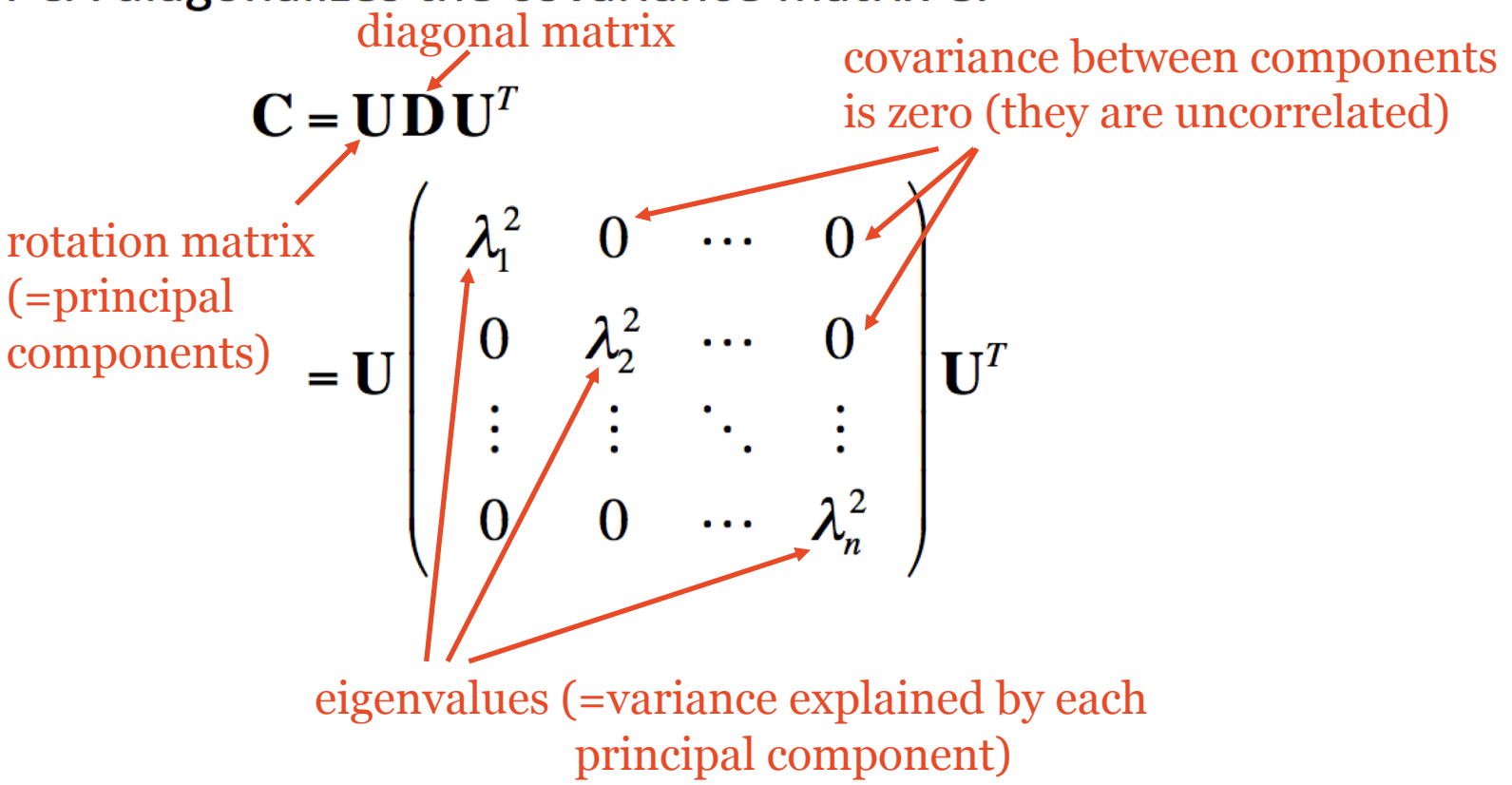
rotation matrix (=principal components)

$$= \mathbf{U} \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^2 \end{pmatrix} \mathbf{U}^T$$

For 2 variables, say Height (X1) and Weight (X2), the C matrix will be 2 X 2

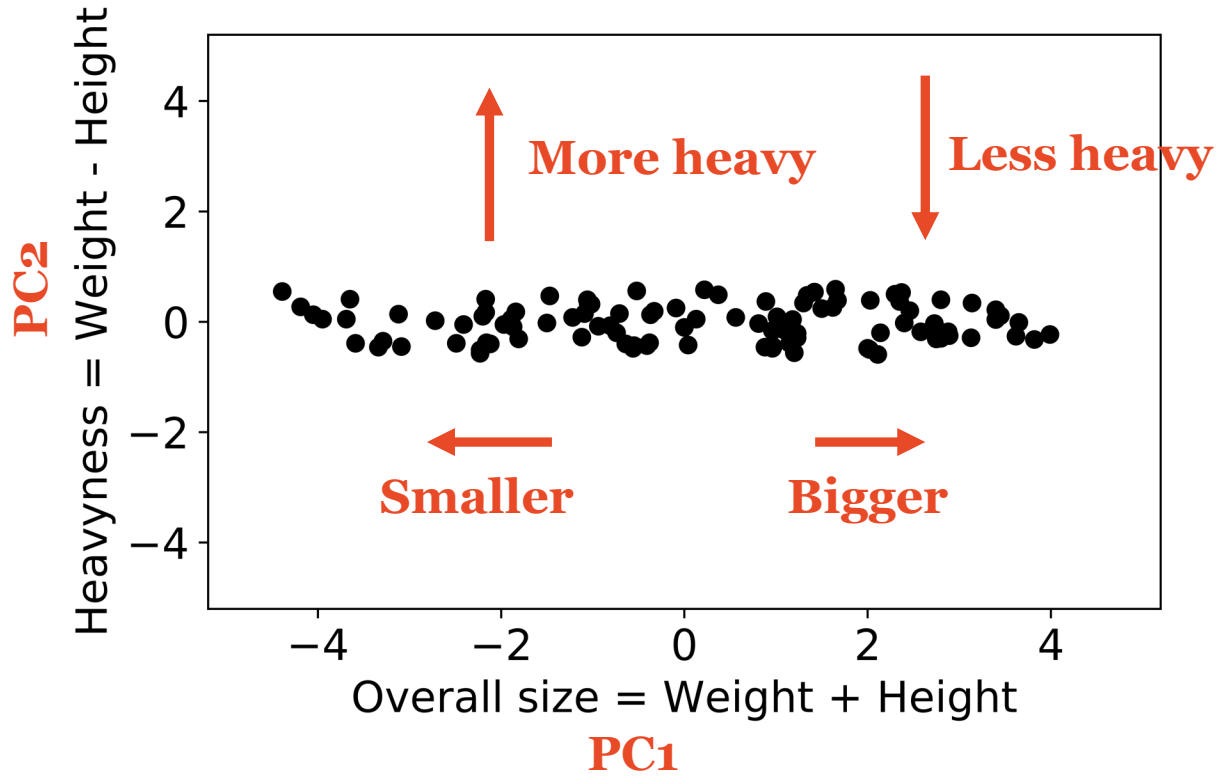
The math behind PCA

PCA diagonalizes the covariance matrix **C**:



What is PCA?

In our earlier example, overall size and heaviness are uncorrelated



Lets summarize the math in simple language

PCA calculates **principal components** (PC) for the data such that:

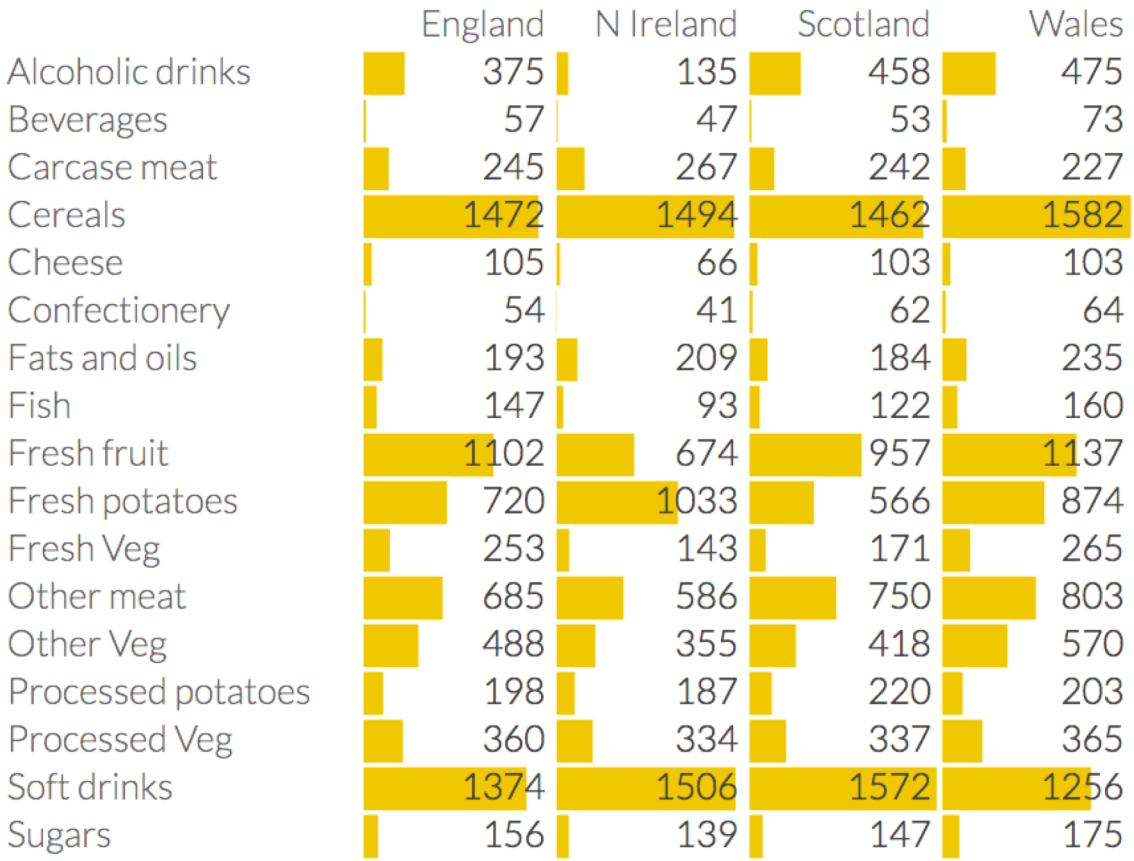
1. Each PC maximizes the variance in the underlying data
2. Each PC is uncorrelated with other PCs

The PCs calculated by PCA are indicative of "important features" in the data.

Mathematically, PCs are combinations of our initial variables (or features).
For example, in our 2D data with Height and Weight, the PCs were weight + height (overall size) and height – weight (heavyness).

A more complex dataset: Eating in the UK (17D problem)

Average consumption of 17 types of food in grams per person per week for every country in the UK.

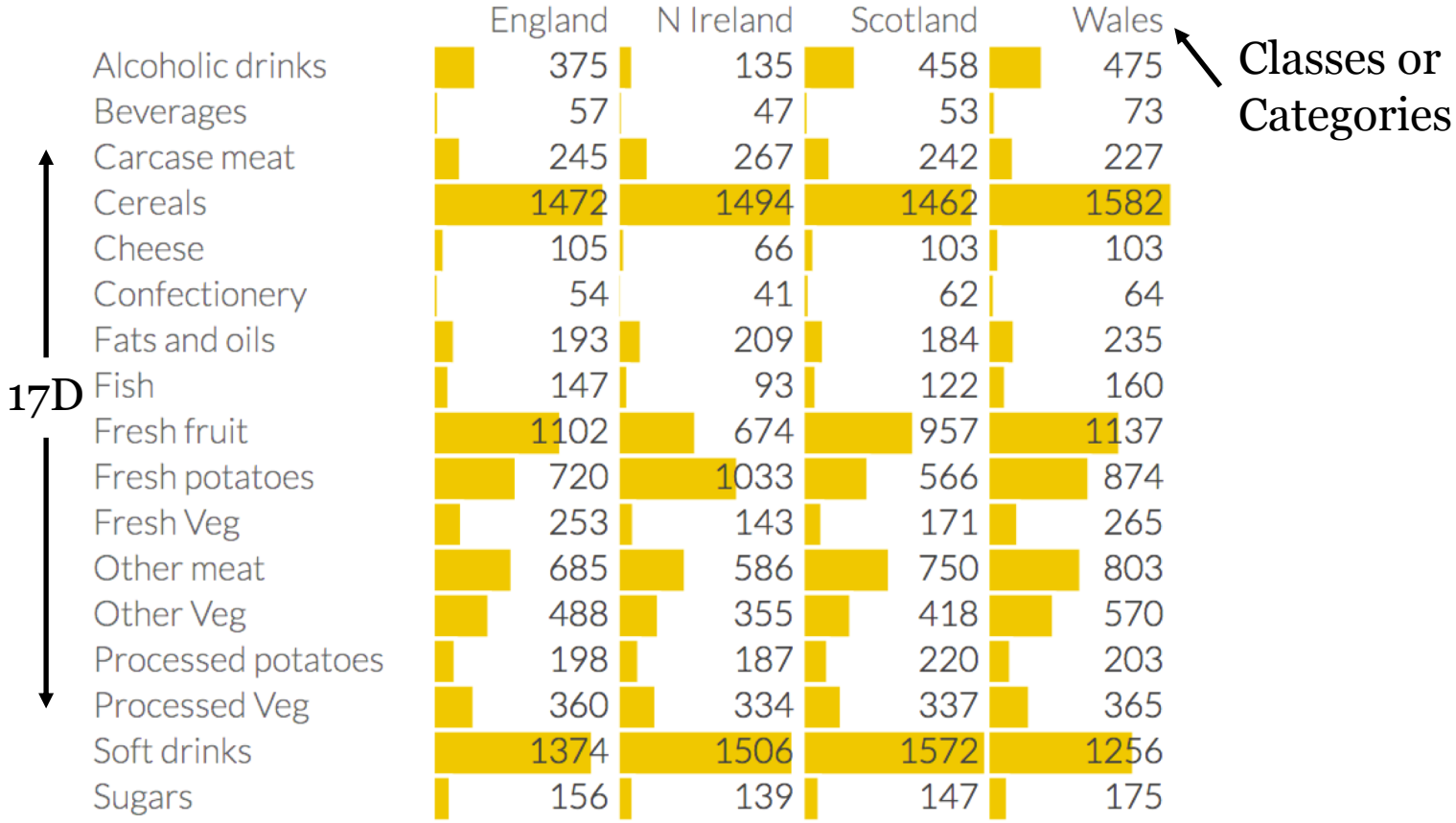


	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

http://setosa.io/ev/principal-component-analysis/

A more complex dataset: Eating in the UK (17D problem)

Average consumption of 17 types of food in grams per person per week for every country in the UK.



What do you think PC1 will tell us? Which of the 17 features look most informative?

http://setosa.io/ev/principal-component-analysis/

PCA on “Eating in the UK” dataset in R

```
## Load dataset
> UK <- read.table(“./Eating_in_the_UK.txt”)
> head(UK)
```

```
          Alcoholic_drinks Beverages Carcase_meat Cereals Cheese Confectionery
England                375        57          245    1472    105            54
Ireland                135        45          267    1494     66            41
Scotland               458        53          242    1462    103            62
Wales                  475        73          227    1582    103            64
         Fats_and_Oils Fish Fresh_fruit Fresh_potatoes Fresh_veg Other_meat
England            193  147        1102            720       253        685
Ireland            209   93         674           1033       143        586
Scotland           184  122         957            566       171        750
Wales              235  160        1137            874       265        803
         Other_veg Processed_potatoes Processed_veg Soft_drinks Sugars
England        488                198           360        1374    156
Ireland        355                187           334        1506    139
Scotland       418                220           337        1572    147
Wales          570                203           365        1256    175
```


https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on "Eating in the UK" dataset in R

```
## Use R function prcomp to perform PCA on the dataset
## Store PCA in variable name UK.pca
> UK.pca <- prcomp(UK)

## Mean for each of the 17 features
> UK.pca$center
```

```
  Alcoholic_drinks          Beverages       Carcase_meat            Cereals
            360.75              57.00             245.25            1502.50
           Cheese      Confectionery      Fats_and_Oils               Fish
            94.25              55.25             205.25             130.50
      Fresh_fruit     Fresh_potatoes         Fresh_veg        Other_meat
            967.50             798.25             208.00             706.00
       Other_veg Processed_potatoes      Processed_veg       Soft_drinks
            457.75             202.00             349.00            1427.00
           Sugars
            154.25
```

Based on these mean values, can we make a better guess for the features PC1 will pick to be most important?
HINT: Look at the data and identify values which vary a lot from their mean.

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

```
## contribution of features to PCs
> UK.pca

Standard deviations (1, .., p=4):
[1] 3.241774e+02 2.127422e+02 7.387441e+01 2.609126e-14

Rotation (n x k) = (17 x 4):
                          PC1          PC2         PC3          PC4
Alcoholic_drinks   -0.463926412 -0.113597925 -0.49853570 -0.485318885
Beverages          -0.029215997  0.029689993 -0.04076224 -0.451451957
Carcase_meat        0.047924691 -0.013910062  0.06366807  0.248787502
Cereals            -0.047711786  0.212596967 -0.35886552  0.016276590
Cheese             -0.056949645 -0.016019816  0.02395099 -0.069076834
Confectionery      -0.029647720 -0.005953961 -0.05231919  0.008651543
Fats_and_Oils      -0.005198777  0.095389616 -0.12523086  0.110378690
Fish               -0.084410250  0.050746596  0.03907227 -0.009925226
Fresh_fruit        -0.632594491  0.177673648  0.40019958 -0.169900442
Fresh_potatoes      0.401330805  0.715081633 -0.20677677 -0.285804501
Fresh_veg          -0.151843305  0.144887702  0.21383604  0.097872349
Other_meat         -0.258898986  0.015296905 -0.55383358  0.527457674
Other_veg          -0.243585069  0.225427668 -0.05331574  0.175077066
Processed_potatoes -0.026882233 -0.042855596 -0.07364515 -0.016493355
Processed_veg      -0.036487206  0.045449072  0.05289459  0.053373496
Soft_drinks         0.232251879 -0.555112651 -0.16942322 -0.203782858
Sugars             -0.037620227  0.043018117 -0.03605668  0.087903232
```

See anything different about Alcoholic drinks, Fresh fruit, Fresh potatoes. Look at the data once more.

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

```
## rotation matrix of the PCA
> UK.pca$x
```

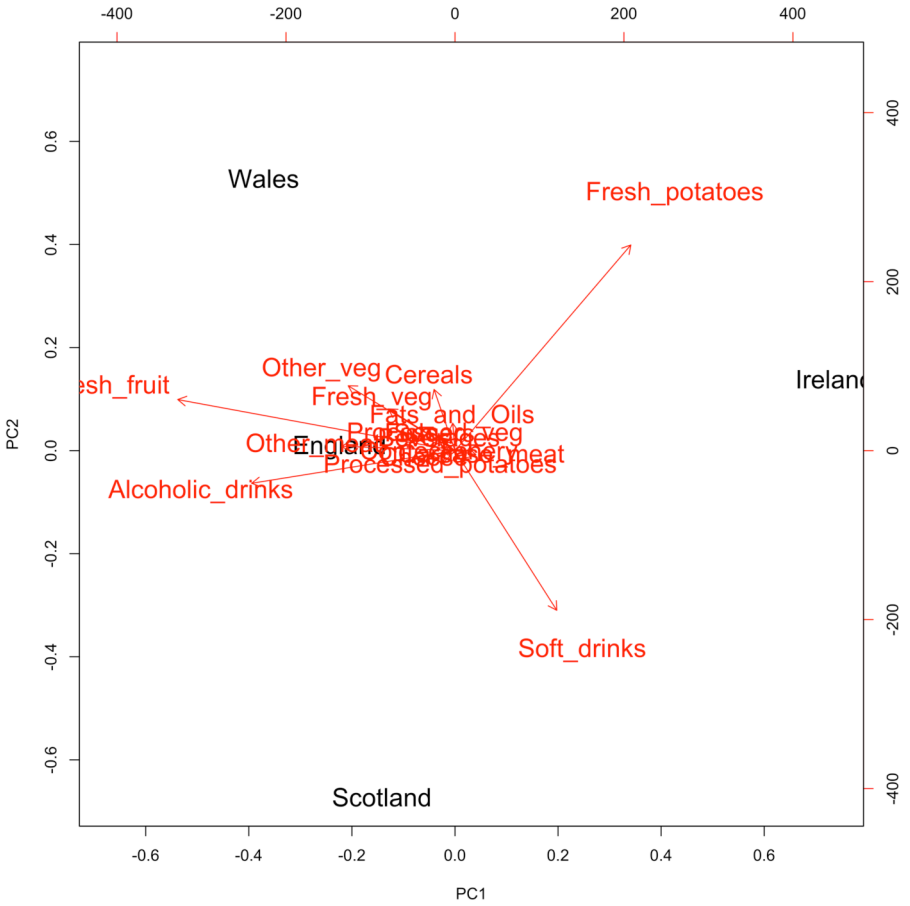
	PC1	PC2	PC3	PC4
England	-144.99358	2.532006	105.767195	-2.883804e-14
Ireland	477.42419	58.924299	-4.881029	1.481038e-13
Scotland	-91.84833	-286.081684	-44.411763	-5.812018e-13
Wales	-240.58227	224.625380	-56.474402	4.638512e-13

How many PCs did the prcomp() function create?

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

```
## visualize PC1 and PC2
> biplot(UK.pca, cex=1.6,
family='sans’)
# or
> biplot(UK.pca, cex=1.6)
```



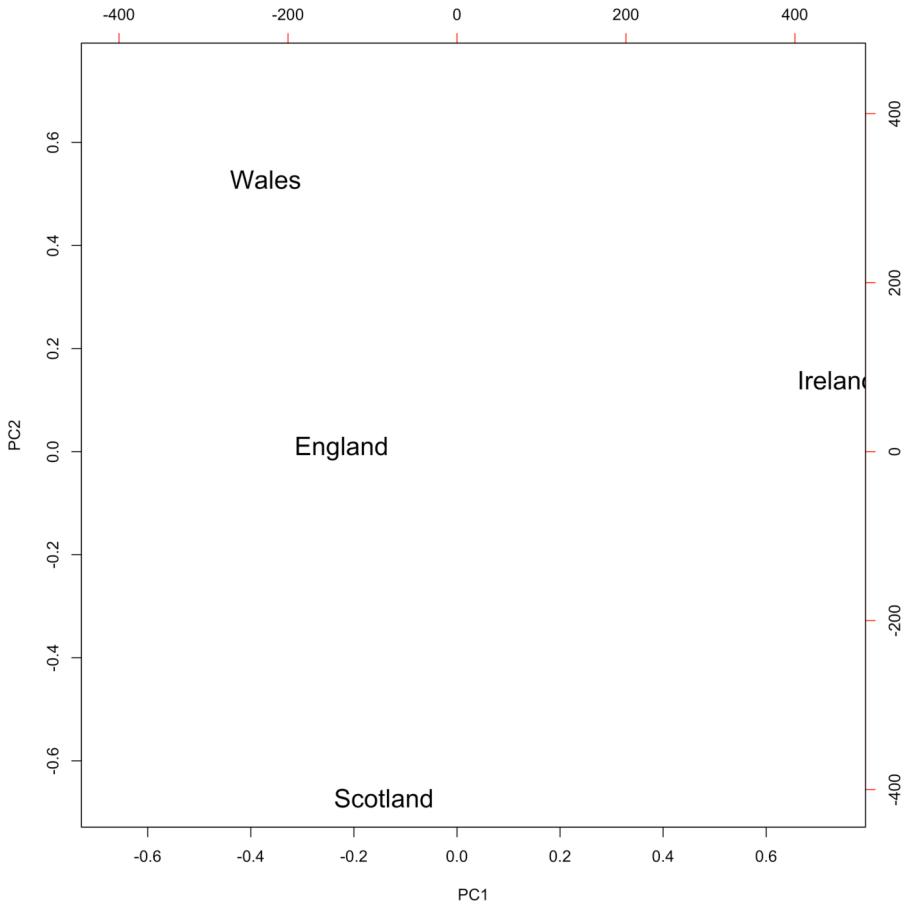
Which country is the outlier along PC1?

Which variables contribute the most to reach this conclusion? Hint: What variables contribute the most to PC1?

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

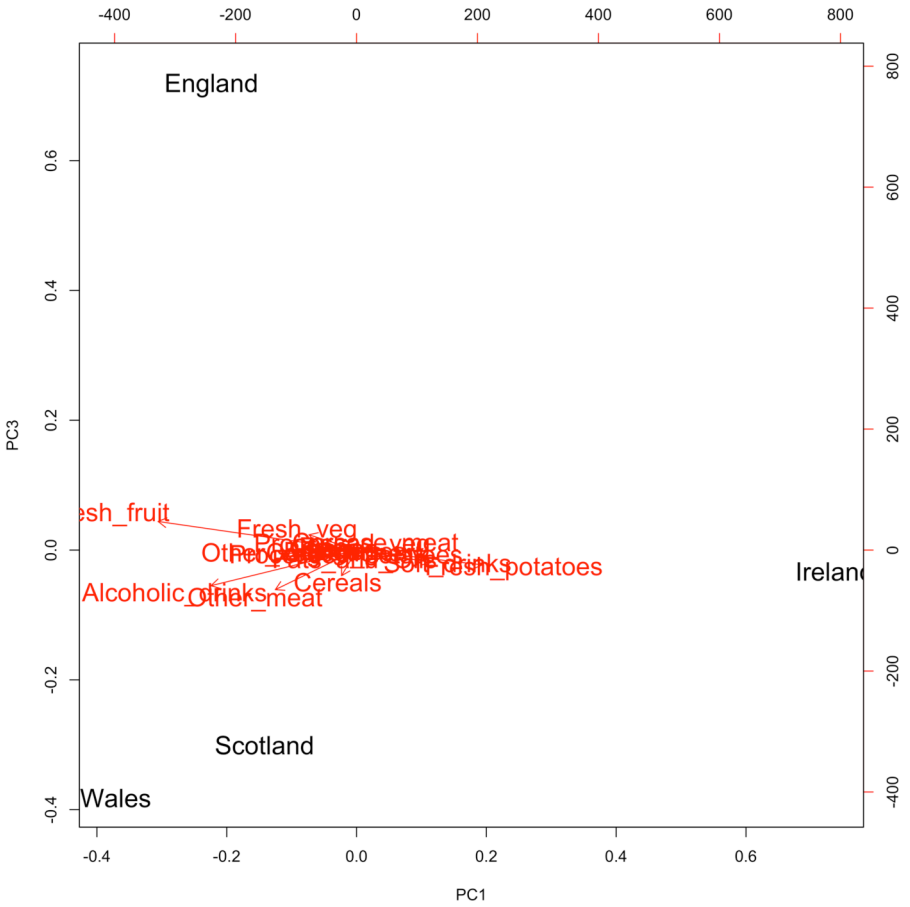
```
## visualize PC1 and PC2
> biplot(UK.pca, var.axes=F,
ylabs=NULL, cex=1.6,
family='sans’)
# or
> biplot(UK.pca, var.axes=F,
ylabs=NULL)
```



https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

```
## visualize PC1 and PC3
> biplot(UK.pca, cex=1.6,
c(1,3), family='sans’)
# or
> biplot(UK.pca,
c(1,3),cex=1.6)
```



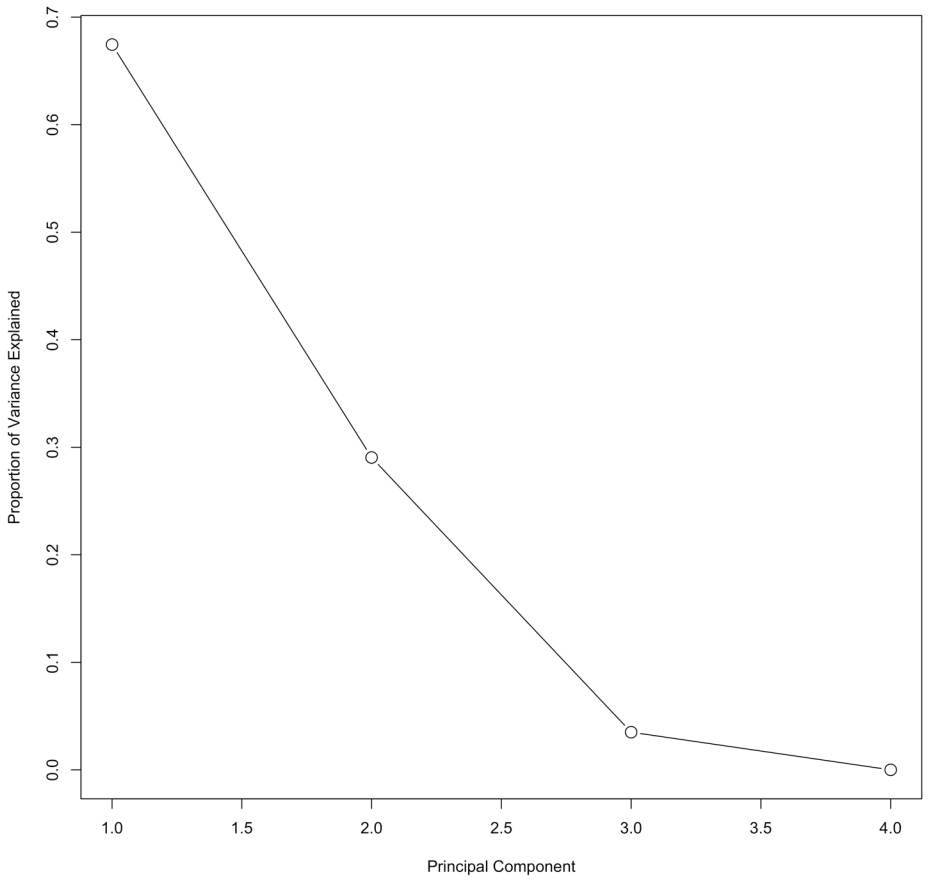
Similarly, visualize PC1 and PC4, PC2 and PC3, PC2 and PC4, PC3 and PC4.

What information to PC3 and PC4 provide?

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

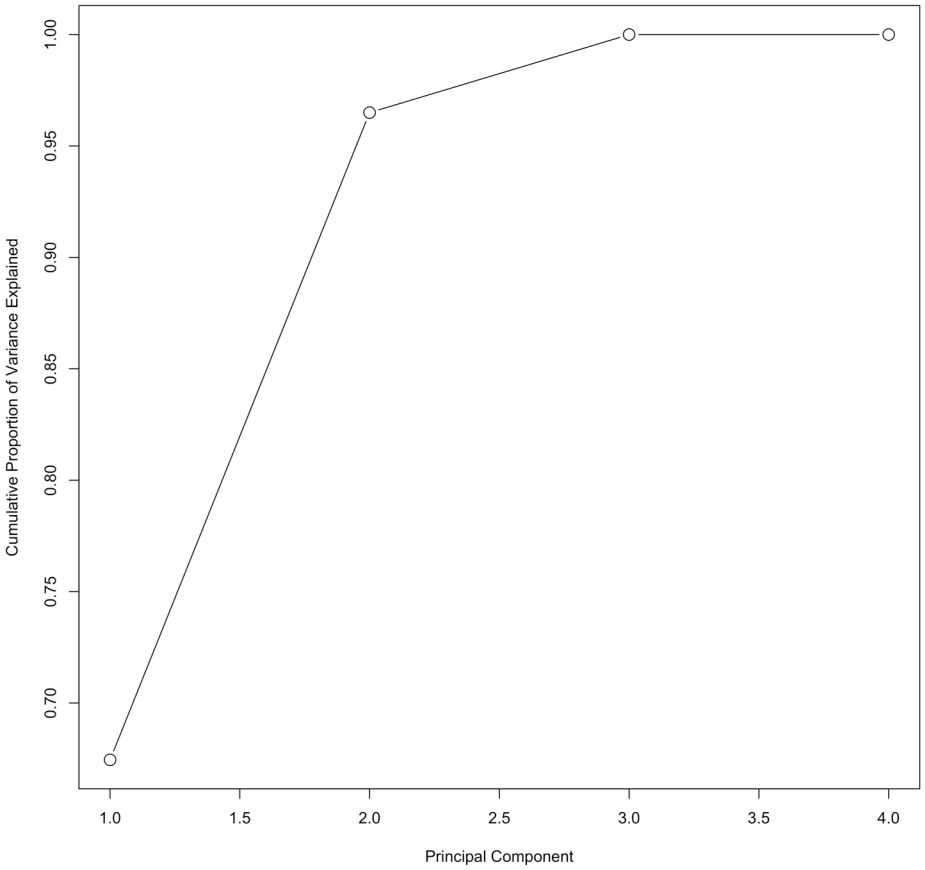
```
## proportion of variance in
the PCs
> std_dev <- UK.pca$sdev
> pr_var <- std_dev^2
> prop_varex <-
pr_var/sum(pr_var)
> plot(prop_varex,
xlab="Principal Component",
ylab="Proportion of Variance
Explained", type="b”,
family=‘sans’)
```



https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

```
## cumulative variation
> std_dev <- UK.pca$sdev
> pr_var <- std_dev^2
> prop_varex <-
pr_var/sum(pr_var)
> plot(cumsum(prop_varex),
xlab="Principal Component",
ylab="Cumulative Proportion of
Variance Explained", type="b”,
family=‘sans’)
```



https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on “Eating in the UK” dataset in R

```
## PCA with only 2 PCs
> UK.pca <- prcomp(UK,rank=2)
> UK.pca
```

```
Standard deviations (1, .., p=4):
[1] 3.241774e+02 2.127422e+02 7.387441e+01 2.609126e-14

Rotation (n x k) = (17 x 2):
                           PC1          PC2
Alcoholic_drinks   -0.463926412 -0.113597925
Beverages          -0.029215997  0.029689993
Carcase_meat        0.047924691 -0.013910062
Cereals            -0.047711786  0.212596967
Cheese             -0.056949645 -0.016019816
Confectionery      -0.029647720 -0.005953961
Fats_and_Oils      -0.005198777  0.095389616
Fish               -0.084410250  0.050746596
Fresh_fruit        -0.632594491  0.177673648
Fresh_potatoes      0.401330805  0.715081633
Fresh_veg          -0.151843305  0.144887702
Other_meat         -0.258898986  0.015296905
Other_veg          -0.243585069  0.225427668
Processed_potatoes -0.026882233 -0.042855596
Processed_veg      -0.036487206  0.045449072
Soft_drinks         0.232251879 -0.555112651
Sugars             -0.037620227  0.043018117
```


https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on "Eating in the UK" dataset in R

```
## PCA with 5 PCs
> UK.pca <- prcomp(UK,rank=5)
> UK.pca
```

We got only 4 PCs, why?

```
Standard deviations (1, .., p=4):
[1] 3.241774e+02 2.127422e+02 7.387441e+01 2.609126e-14

Rotation (n x k) = (17 x 4):
                          PC1          PC2         PC3          PC4
Alcoholic_drinks   -0.463926412 -0.113597925 -0.49853570 -0.485318885
Beverages          -0.029215997  0.029689993 -0.04076224 -0.451451957
Carcase_meat        0.047924691 -0.013910062  0.06366807  0.248787502
Cereals            -0.047711786  0.212596967 -0.35886552  0.016276590
Cheese             -0.056949645 -0.016019816  0.02395099 -0.069076834
Confectionery      -0.029647720 -0.005953961 -0.05231919  0.008651543
Fats_and_Oils      -0.005198777  0.095389616 -0.12523086  0.110378690
Fish               -0.084410250  0.050746596  0.03907227 -0.009925226
Fresh_fruit        -0.632594491  0.177673648  0.40019958 -0.169900442
Fresh_potatoes      0.401330805  0.715081633 -0.20677677 -0.285804501
Fresh_veg          -0.151843305  0.144887702  0.21383604  0.097872349
Other_meat         -0.258898986  0.015296905 -0.55383358  0.527457674
Other_veg          -0.243585069  0.225427668 -0.05331574  0.175077066
Processed_potatoes -0.026882233 -0.042855596 -0.07364515 -0.016493355
Processed_veg      -0.036487206  0.045449072  0.05289459  0.053373496
Soft_drinks         0.232251879 -0.555112651 -0.16942322 -0.203782858
Sugars             -0.037620227  0.043018117 -0.03605668  0.087903232
```


https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

PCA on "Eating in the UK" dataset in R

```
## PCA with 5 PCs
> h <- read.table("household_data.txt")
> h
           Alcoholic_drinks Beverages Carcase_meat Cereals Cheese Confectionery
Household1              375        57          245    1472    105            52
Household2              476        75          229    1584    105            66
           Fats_and_Oils Fish Fresh_fruit Fresh_potatoes Fresh_veg Other_meat
Household1           193  147        1102            720       253        685
Household2           237  162        1139            875       264        806
           Other_veg Processed_potatoes Processed_veg Soft_drinks Sugars
Household1       485                198           360        1374    156
Household2       572                205           367        1258    177

> predict(UK.pca,h)

                PC1        PC2       PC3        PC4
Household1 -144.2035   1.867631 106.03178 -0.5425343
Household2 -243.1700 225.620146 -59.71472  0.2730122
> UK.pca$x

                PC1         PC2        PC3           PC4
England  -144.99358    2.532006 105.767195 -2.883804e-14
Ireland   477.42419   58.924299  -4.881029  1.481038e-13
Scotland  -91.84833 -286.081684 -44.411763 -5.812018e-13
Wales    -240.58227  224.625380 -56.474402  4.638512e-13
```

Do PC1 and PC2 of Household1 seem close to that for England. Hence, our PCA model assigned Household 1 to England.

What about Household2?

Can you see this visually too?

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

Using Principal Coordinate Analysis (PCoA) for distance matrix

```
## Load US cities dataset
> cities <- as.matrix(UScitiesD)
> cities
```

	Atlanta	Chicago	Denver	Houston	LosAngeles	Miami	NewYork
Atlanta	0	587	1212	701	1936	604	748
Chicago	587	0	920	940	1745	1188	713
Denver	1212	920	0	879	831	1726	1631
Houston	701	940	879	0	1374	968	1420
LosAngeles	1936	1745	831	1374	0	2339	2451
Miami	604	1188	1726	968	2339	0	1092
NewYork	748	713	1631	1420	2451	1092	0
SanFrancisco	2139	1858	949	1645	347	2594	2571
Seattle	2182	1737	1021	1891	959	2734	2408
Washington.DC	543	597	1494	1220	2300	923	205

	SanFrancisco	Seattle	Washington.DC
Atlanta	2139	2182	543
Chicago	1858	1737	597
Denver	949	1021	1494
Houston	1645	1891	1220
LosAngeles	347	959	2300
Miami	2594	2734	923
NewYork	2571	2408	205
SanFrancisco	0	678	2442
Seattle	678	0	2329
Washington.DC	2442	2329	0

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

Using Principal Coordinate Analysis (PCoA) for distance matrix

```
## Load US cities dataset
## k is the number of principal coordinates
> city.location <- cmdscale(cities, k=2)
> city.location
```

```
                 [,1]      [,2]
Atlanta       -718.7594  142.99427
Chicago       -382.0558 -340.83962
Denver         481.6023  -25.28504
Houston       -161.4663  572.76991
LosAngeles    1203.7380  390.10029
Miami        -1133.5271  581.90731
NewYork      -1072.2357 -519.02423
SanFrancisco  1420.6033  112.58920
Seattle       1341.7225 -579.73928
Washington.DC -979.6220 -335.47281
```

Try the following:

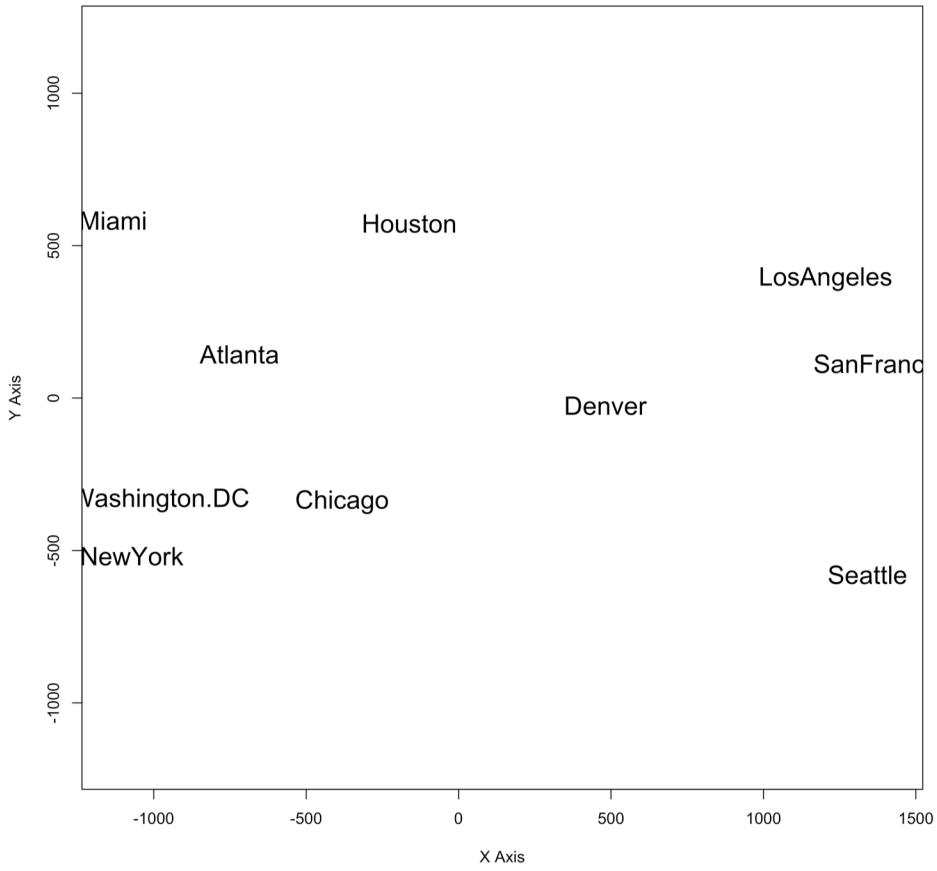
?cmdscale

Press ‘q’ to escape. You can do this for any R function.

https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

Using Principal Coordinate Analysis (PCoA) for distance matrix

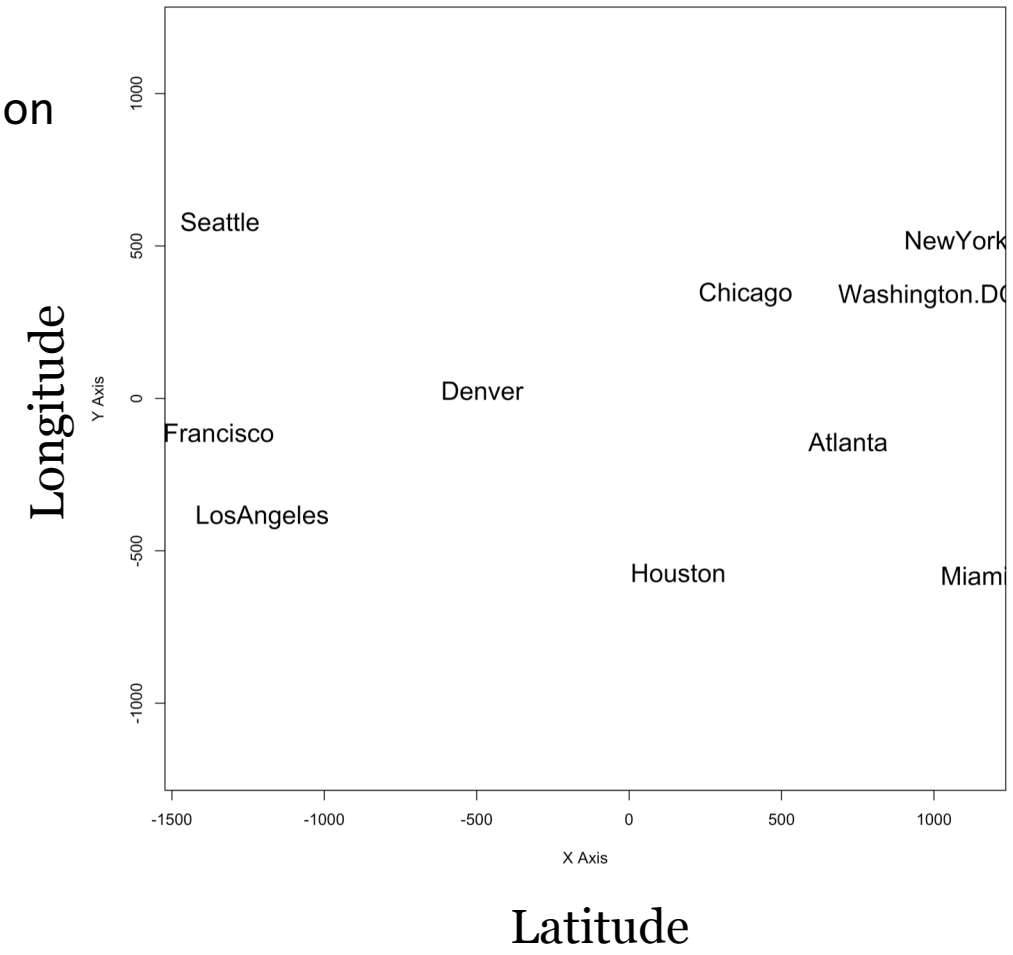
```
## Plot
> x <- city.location[,1]
> y <- city.location[,2]
> plot(x, y, type="n", asp=1,
xlab="", ylab="")
> text(x,y,labels(UScitiesD))
```



https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

Using Principal Coordinate Analysis (PCoA) for distance matrix

```
## Plot
> city.location <- -city.location
> x <- city.location[,1]
> y <- city.location[,2]
> plot(x, y, type="n", asp=1,
xlab="", ylab="")
> text(x,y,labels(UScitiesD))
```



https://github.com/shriyaamittal/MCB432/blob/master/dimensionality_reduction/R_commands

Some limitations of dimensionality reduction

- It may lead to some amount of data loss.
- PCs are linear combinations of features, sometimes this may not be desirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.
- We may not know how many principal components to keep - in practice, some thumb rules are applied.