QUEEN'S UNIVERSITY BELFAST
SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Data Analysis & Visualisation

CSC3062

BEng (CS & SE), MEng (CS & SE), BIT & CIT

Dr Reza Rafiee

Semester 1 2019

Quick review on some technical terms

- Dataset (data set, cohort)
- Variables (or features) & feature space
- Observation (sample)
- Variation
- Dimension
- Pattern

Principal component analysis (PCA)

- A **linear** dimensionality-reduction technique
  - **Transforming** variables (or features) of a large dataset (i.e., multivariate data) into a smaller one that still contains most of the information in the large dataset

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

Principal component analysis (PCA)

Reducing data by **projecting** (geometrically) into a lower dimensions which called principal components (PCs)

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

PCA example

Let's look at
a question & analysis for better
understanding of PCA

PCA example (17 dimensions)

Eating in the UK

Assume we have a dataset including 17 features/dimensions (Table 1). This table shows the average consumption of 17 types of food in grams per person per week for every country in the UK.

The table shows some interesting variations across different food types, but overall differences aren't so notable. Let's see if PCA can eliminate dimensions to emphasize how countries differ.

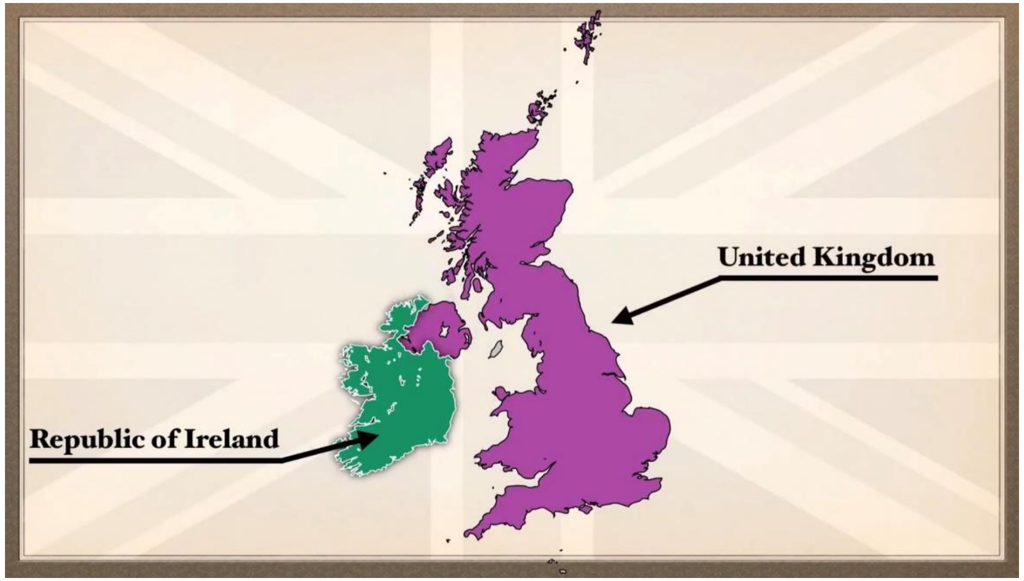
http://www.sdss.jhu.edu/~szalay/class/2016-oldold/SignalProcPCA.pdf

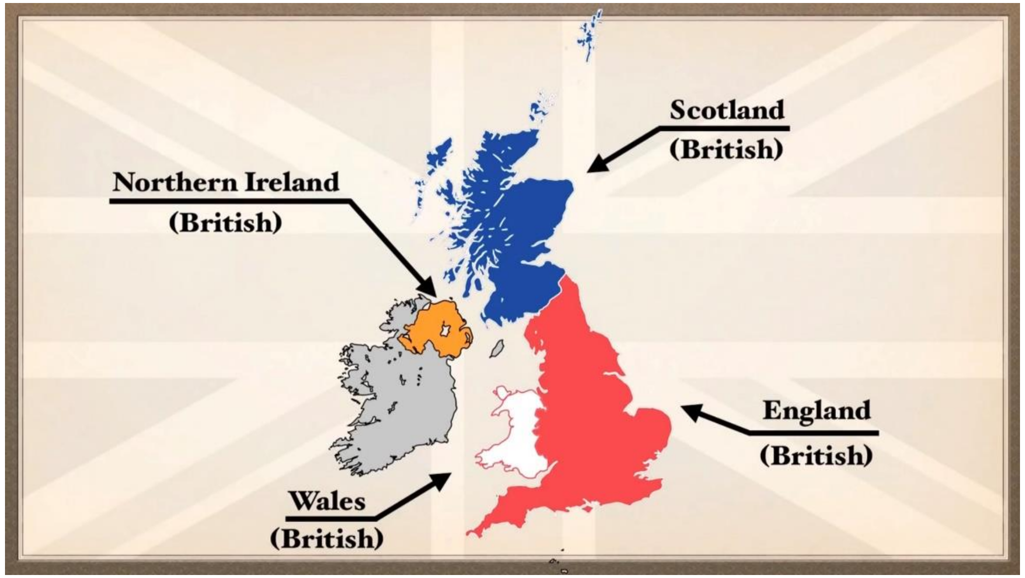
	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	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

Table 1: UK food consumption in 1997 (g/person/week). Source: DEFRA website

PCA example (17 dimensions)

Eating in the UK





PCA example (17 dimensions)

Eating in the UK

Assume we have a dataset including 17 features/dimensions (Table 1). This table shows the *average consumption* of 17 types of food in grams per person per week for every country in the UK.

The table shows some interesting **variations across different food types**, but overall differences aren't so notable. Let's see **if PCA can eliminate dimensions to emphasize how countries differ**.

	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

Table 1: UK food consumption in 1997 (g/person/week). Source: DEFRA website

PCA example (17 dimensions)

Eating in the UK – Question?

Can PCA reduce the dimension of this dataset (i.e., eliminate dimensions) to highlight how countries differ?

PCA analysis using *prcomp()* package

	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

Input_dataset

?

	England	N Ireland	Scotland	Wales
PC1	-144.993	477.3916	-91.8693	-240.529
PC2	2.532999	58.90186	-286.082	224.6469
PC3	105.7689	-4.8779	-44.4155	-56.4756

Reduced dataset

Summarises of features

PCA analysis using *prcomp()* package

	England	N Ireland	Scotland	Wales
PC1	-144.993	477.3916	-91.8693	-240.529
PC2	2.532999	58.90186	-286.082	224.6469
PC3	105.7689	-4.8779	-44.4155	-56.4756

Reduced dataset

Question: how many new variables (PCs) will be acceptable when using this transformation (or projection)?

PCA analysis using *prcomp()* package

Input_dataset

	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

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)
 # scale =T is appropriate for high-dimensional data
```

PCA analysis using *prcomp()* package

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)
 # scale =T is appropriate for high-dimensional data
summary(PCA_Model_prcomp)
# Importance of components:
#                               PC1      PC2      PC3       PC4
# Standard deviation     324.1502 212.7478 73.87622 3.828e-14
# Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
# Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00
```

PCA analysis using *prcomp()* package

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)
 # scale =T is appropriate for high-dimensional data
summary(PCA_Model_prcomp)
# Importance of components:
#                            PC1      PC2      PC3       PC4
# Standard deviation     324.1502 212.7478 73.87622 3.828e-14
# Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
# Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00

PCA_Model_prcomp$x   # Showing the principle components
#                PC1         PC2        PC3           PC4
# England   -144.99315    2.532999 105.768945 -3.765391e-14
# N Ireland  477.39164   58.901862  -4.877895  1.667659e-13
# Scotland   -91.86934 -286.081786 -44.415495 -8.860586e-13
# Wales     -240.52915  224.646925 -56.475555  7.770000e-13
```

PCA analysis using *prcomp()* package

```
PCA_Model_prcomp <- prcomp(t(Input_dataset), center = T, scale=F)
 # scale =T is appropriate for high-dimensional data
summary(PCA_Model_prcomp)
# Importance of components:
#                            PC1      PC2      PC3       PC4
# Standard deviation     324.1502 212.7478 73.87622 3.828e-14
# Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
# Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00
```

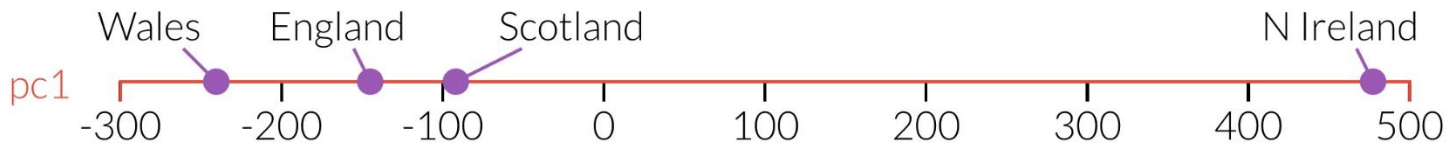
```
PCA_Model_prcomp$x   # Showing the principle components
#                  PC1         PC2        PC3           PC4
# England   -144.99315    2.532999 105.768945 -3.765391e-14
# N Ireland  477.39164   58.901862  -4.877895  1.667659e-13
# Scotland   -91.86934 -286.081786 -44.415495 -8.860586e-13
# Wales     -240.52915  224.646925 -56.475555  7.770000e-13
```

PCA analysis using *prcomp()* package

Eating in the UK

Here is the plot of the data along the first principal component (PCA).

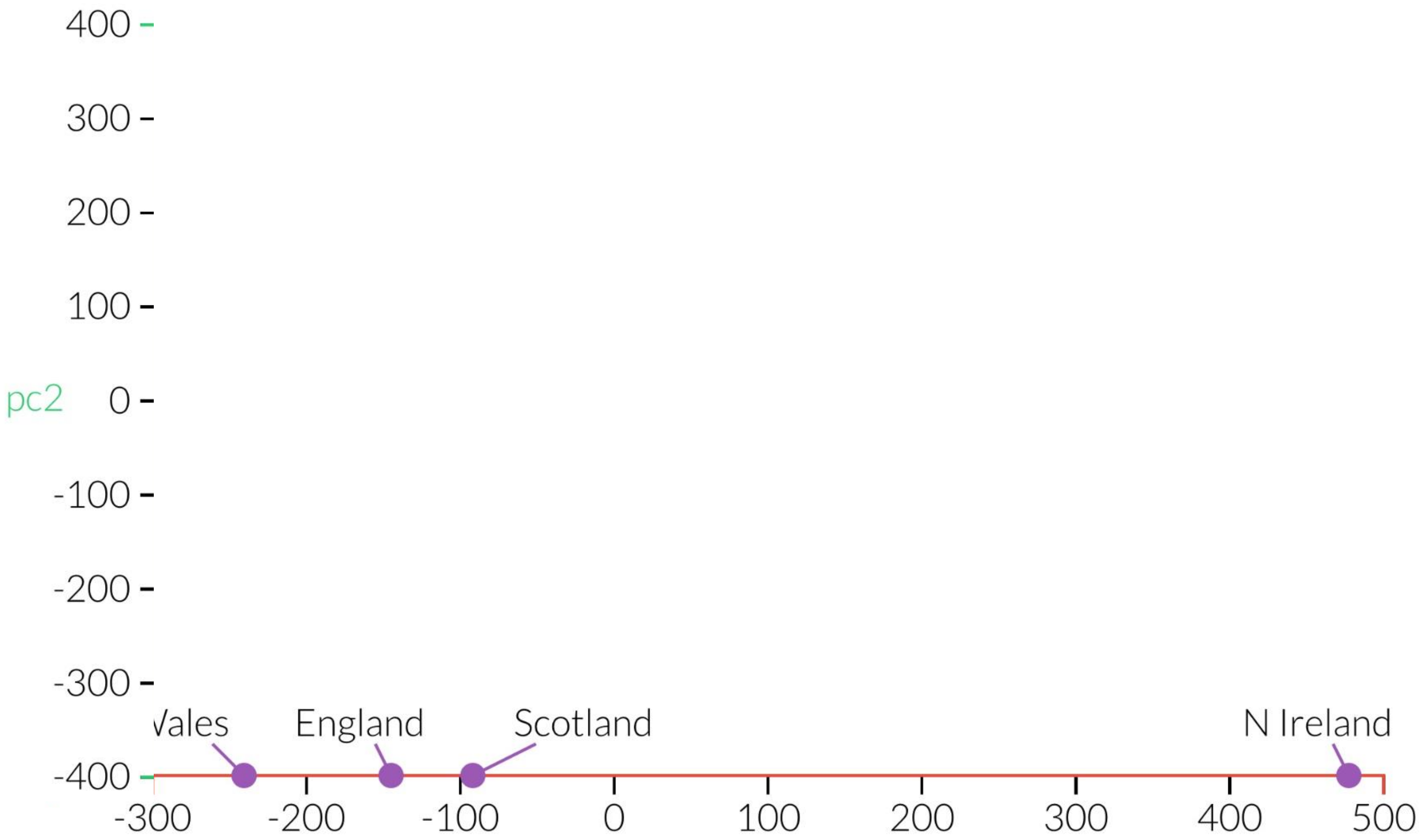
	PC1
# England	**-144.99315**
# N Ireland	**477.39164**
# Scotland	**-91.86934**
# Wales	**-240.52915**



PCA analysis using *prcomp()* package

Eating in the UK

Adding the second principal components (PCA).

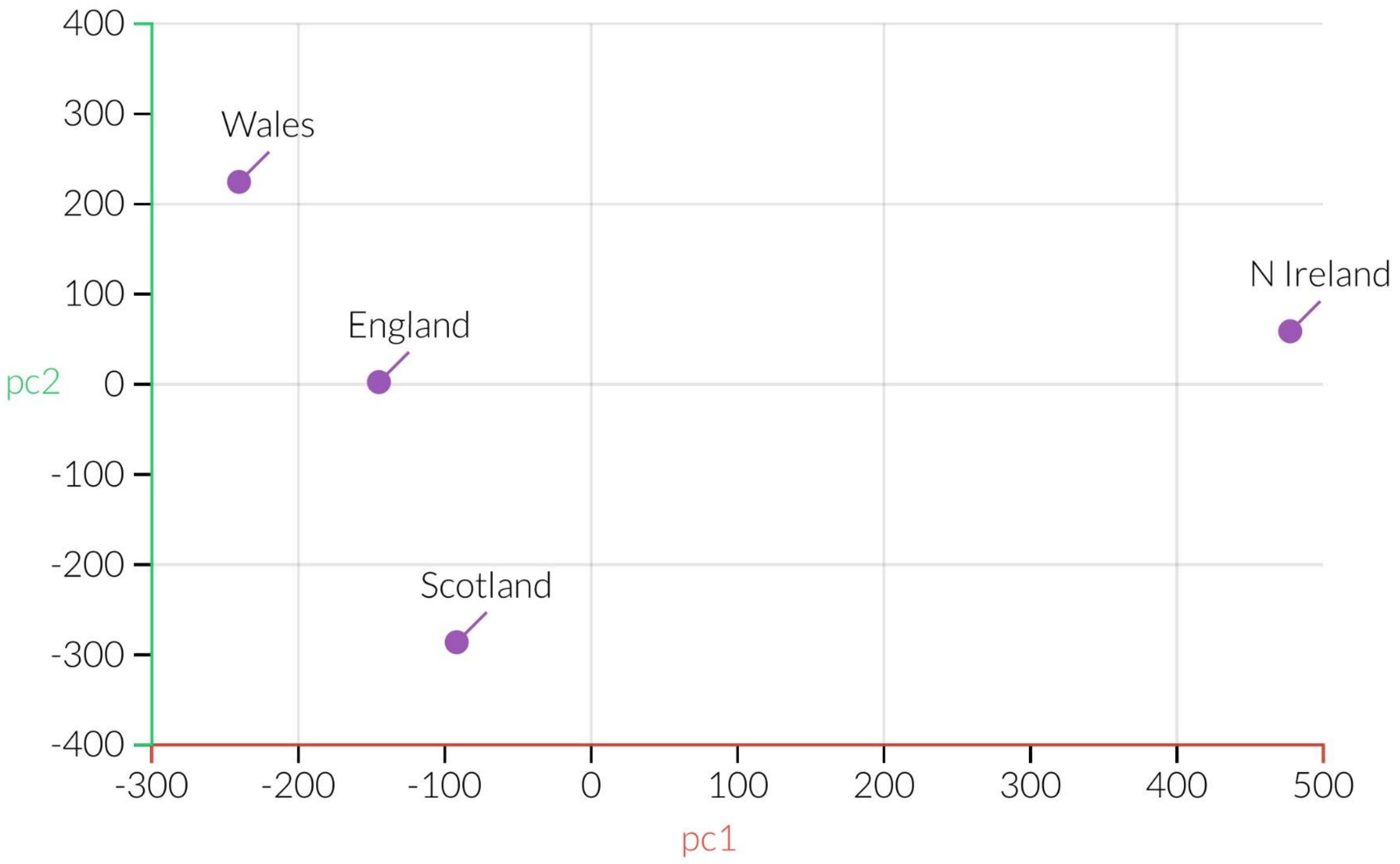


PCA analysis using *prcomp()* package

Eating in the UK

The first and second principal components (PCA).

#	PC1	PC2
# England	-144.99315	2.532999
# N Ireland	477.39164	58.901862
# Scotland	-91.86934	-286.081786
# Wales	-240.52915	224.646925



Interpret the PCs

Eating in the UK

Assume we have a dataset including 17 features/dimensions (Table 1). This table shows the average consumption of 17 types of food in grams per person per week for every country in the UK.

The Northern Irish eat way more grams of <u>fresh potatoes</u> and way fewer of <u>fresh fruits</u>, <u>cheese</u>, <u>fish</u> and <u>alcoholic drinks</u>.

	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

Table 1: UK food consumption in 1997 (g/person/week). Source: DEFRA website

Principal component analysis (PCA)

- A **linear** dimensionality-reduction technique
  - **Transforming** variables (or features) of a large dataset (i.e., multivariate data) into a smaller one that still contains most of the information in the large dataset
- By using PCA, we **reduce the number of features** of a dataset, while preserving as much information as possible.
- Reducing data by **projecting** (geometrically) into a lower dimensions which called principal components (PCs)

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

Directions with most variance

- Principal components are the underlying structure in the data.

- Principal components are **the directions** where there is the **most variance**, or the directions where the data is most spread out.

How to find the direction where there is most variance?

Directions with most **variance**

Standard deviation (SD) is a measure of how spread out numbers are

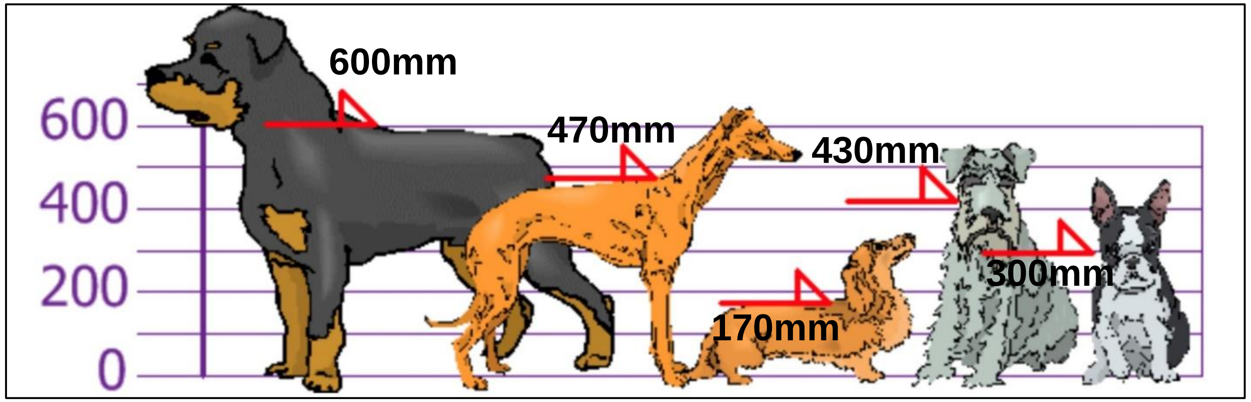
SD is the square root of the variance

(Sigma: $\boldsymbol{\sigma}$) $\boldsymbol{\sigma}^2$

We aim to measure the average, SD and variance of the heights of following dogs

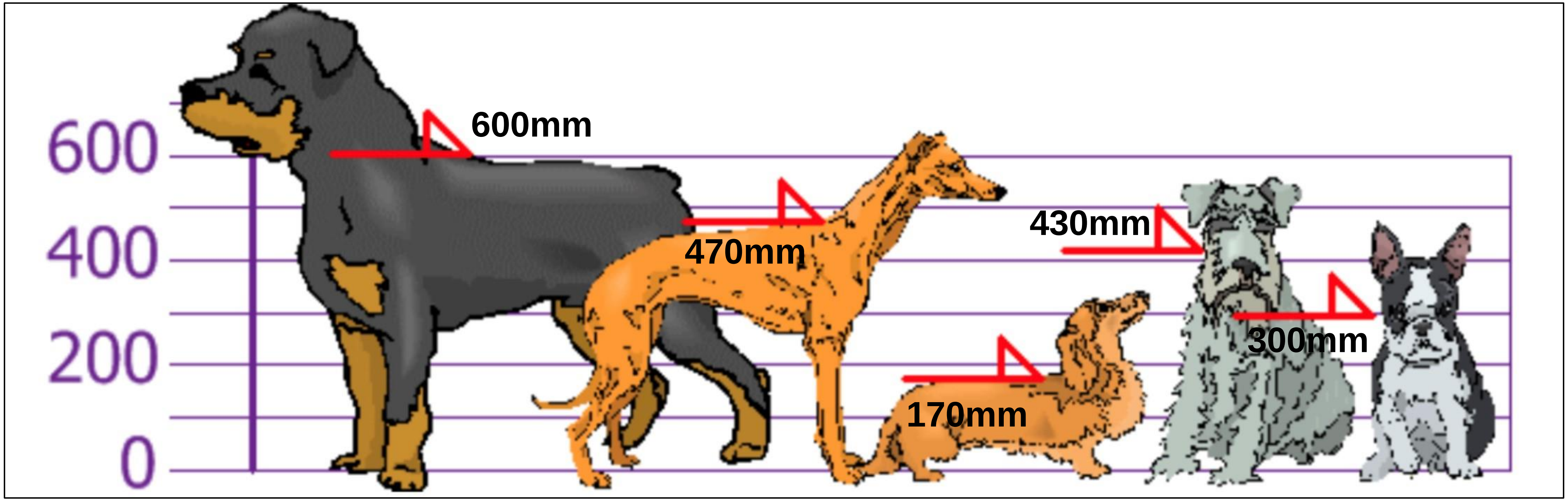
https://studylib.net/doc/6625962/standard-deviation-and-variance-explanation

What is variance?



Directions with most **variance**

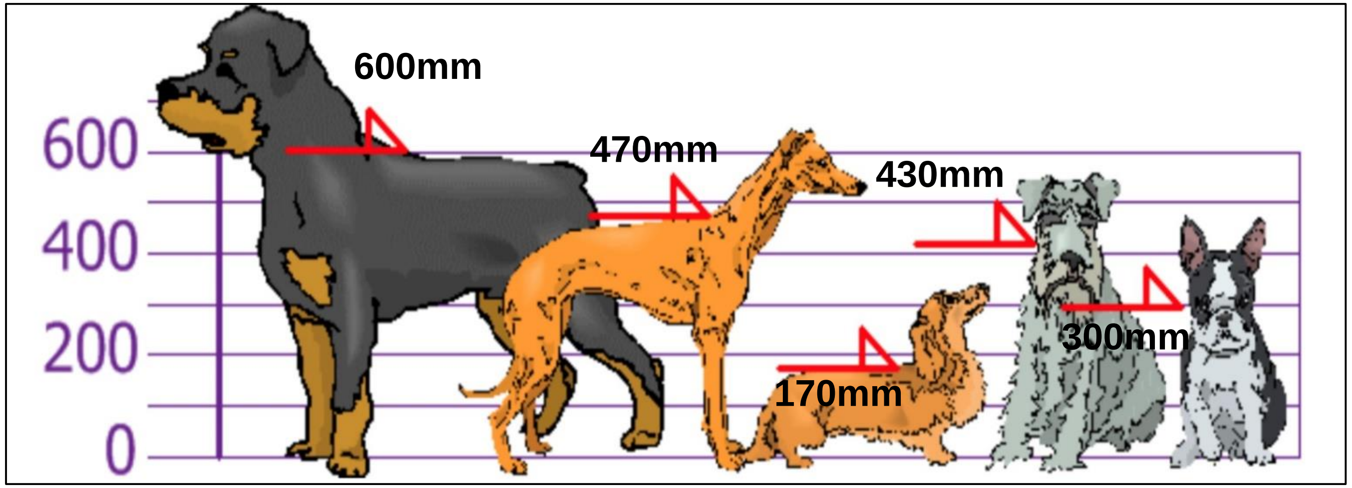
We aim to measure the average, SD and variance of the **heights** of following **dogs**



Sample (observation)
Variable or feature (i.e, random variable)
Statistical measurement

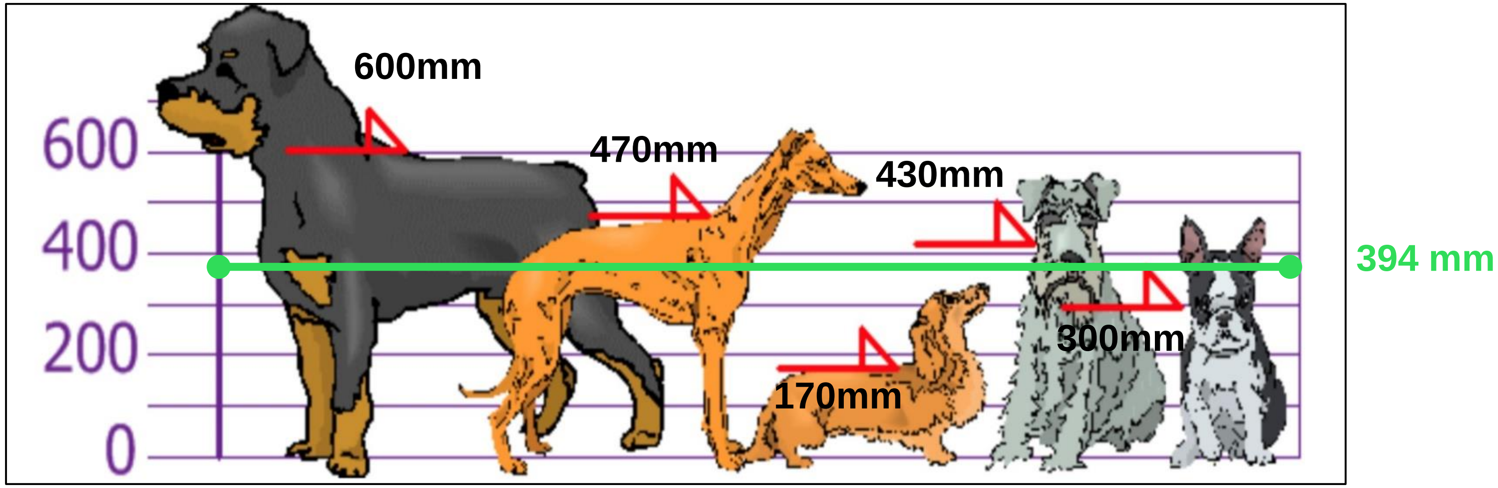
https://studylib.net/doc/6625962/standard-deviation-and-variance-explanation

Directions with most **variance**



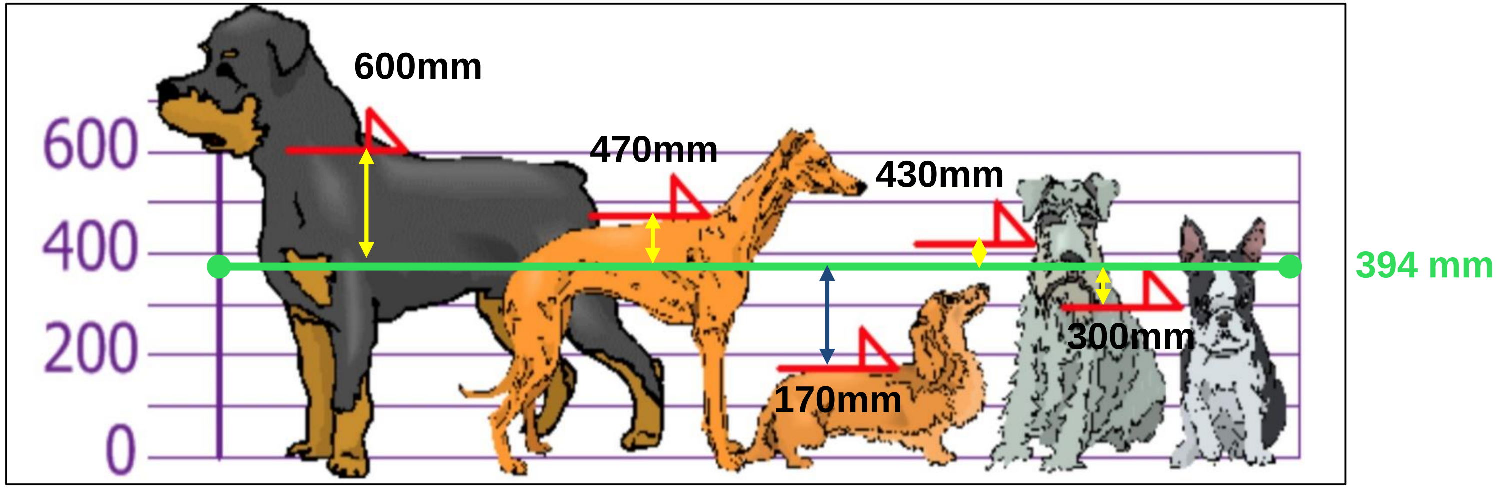
$$Mean = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

Directions with most **variance**

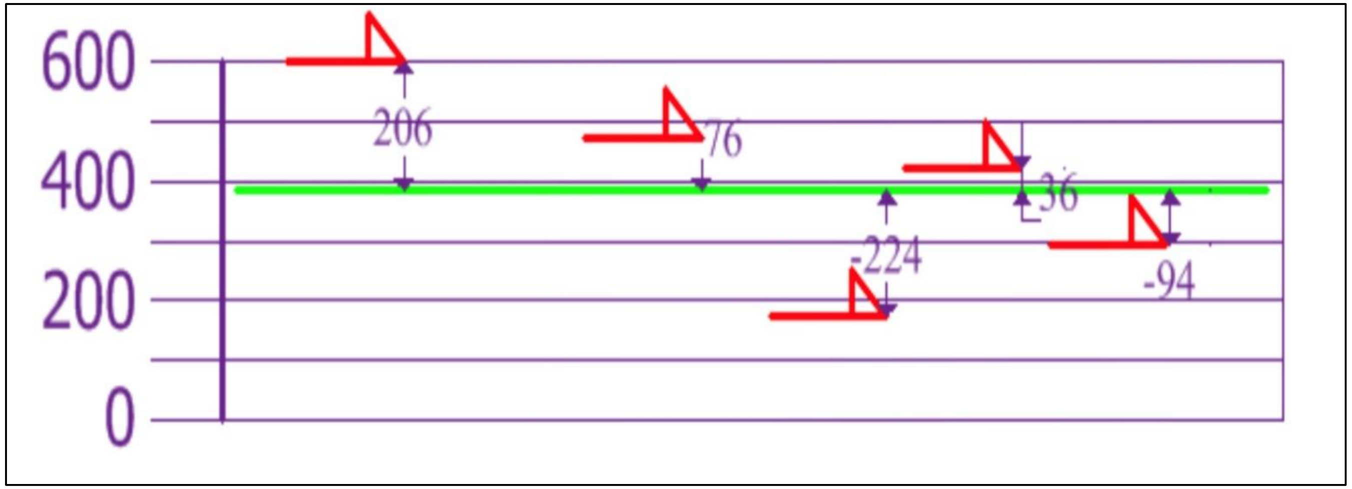


$$Mean = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

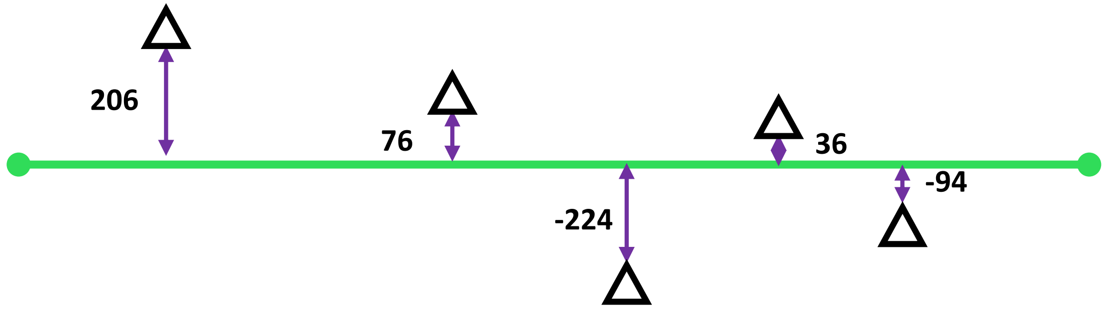
Directions with most **variance**



Now, we calculate each dogs **difference from the Mean**



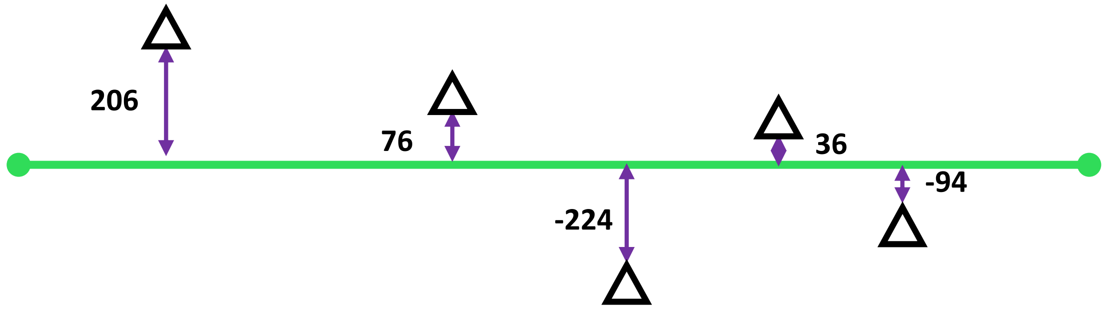
Directions with most **variance**



$$\mu = Mean = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} = \frac{108{,}520}{5} = 21{,}704$$

Directions with most **variance**



$$\mu = Mean = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} = \frac{108{,}520}{5} = 21{,}704$$

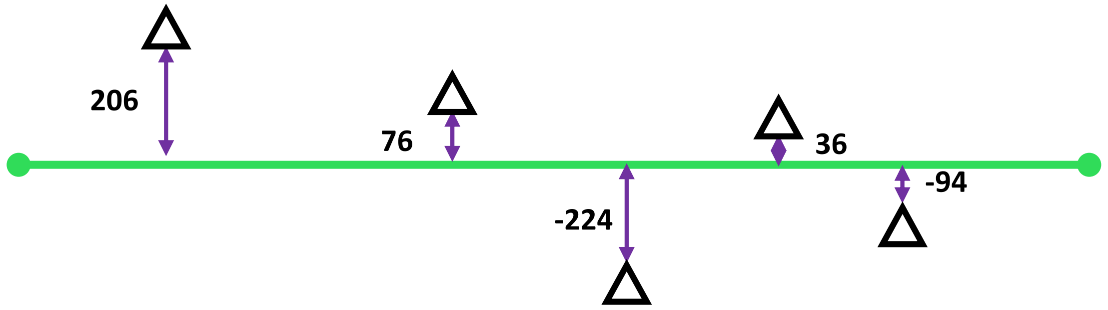
$$\sigma = \sqrt{21{,}704} = 147.32 \approx 147\ mm$$

Directions with most **variance**

$\mu = 394$

$\sigma^2 = 21{,}704$

$\sigma \approx 147\ mm$

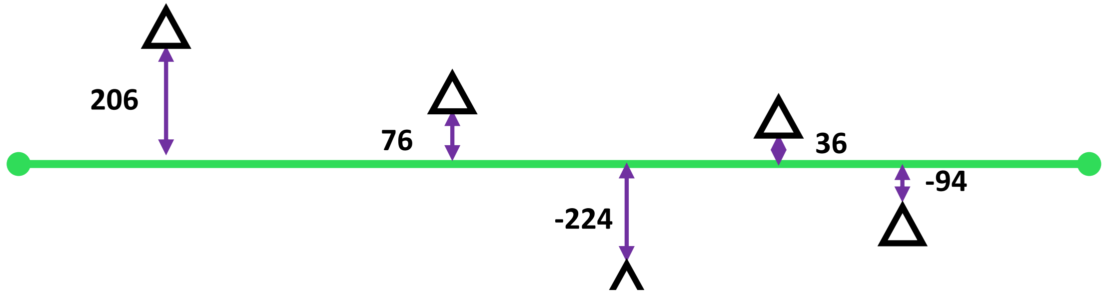


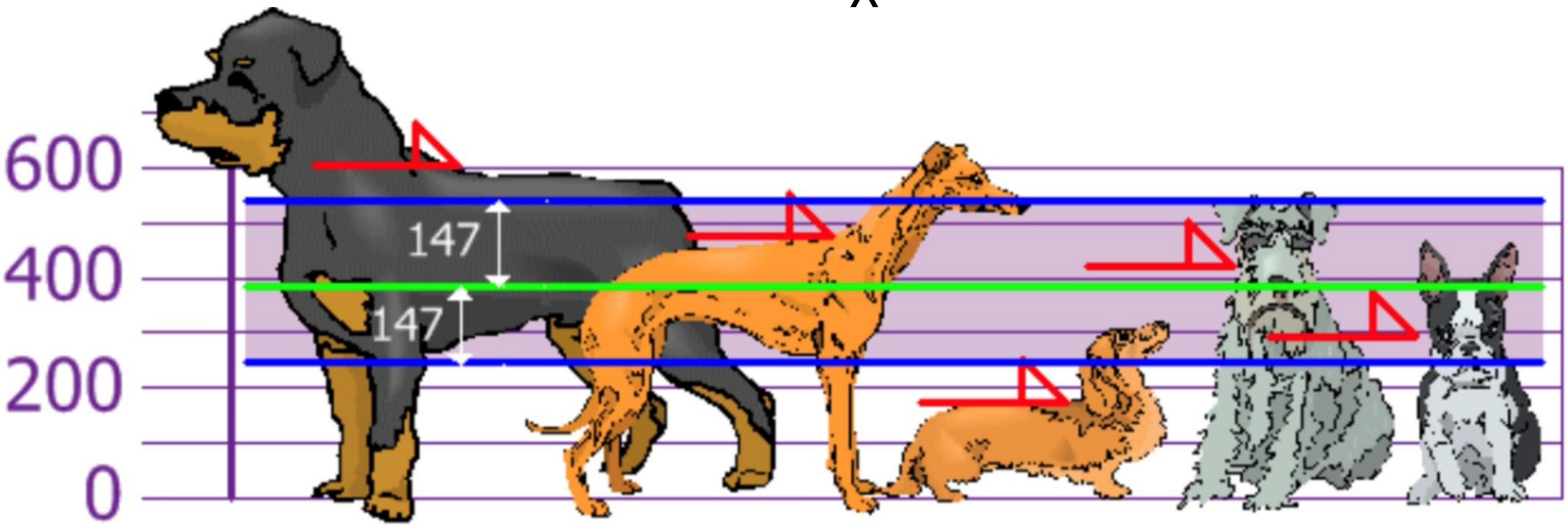
Directions with most **variance**

$\mu = 394$

$\sigma^2 = 21{,}704$

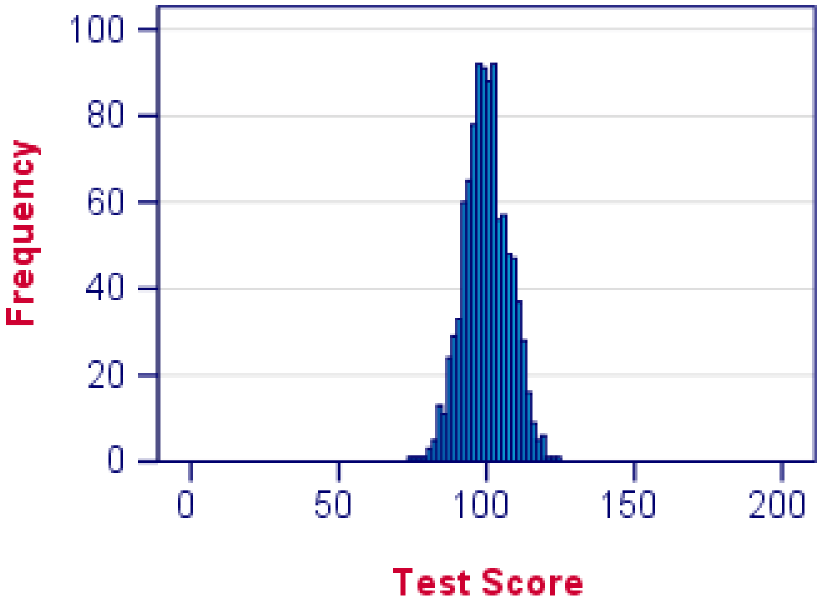
$\sigma \approx 147\ mm$

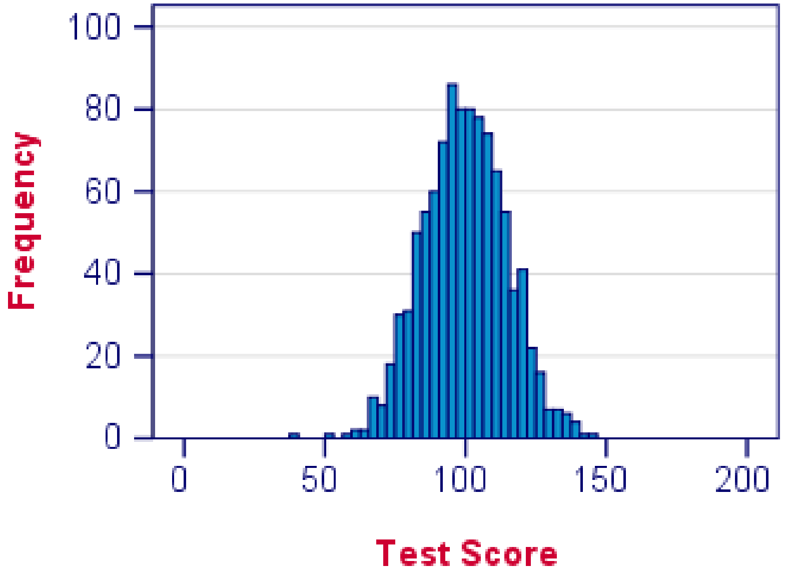


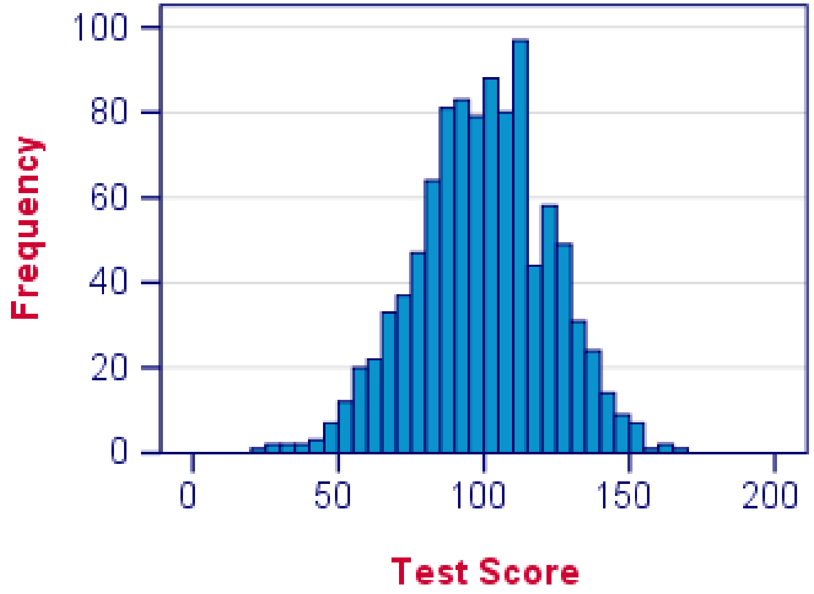


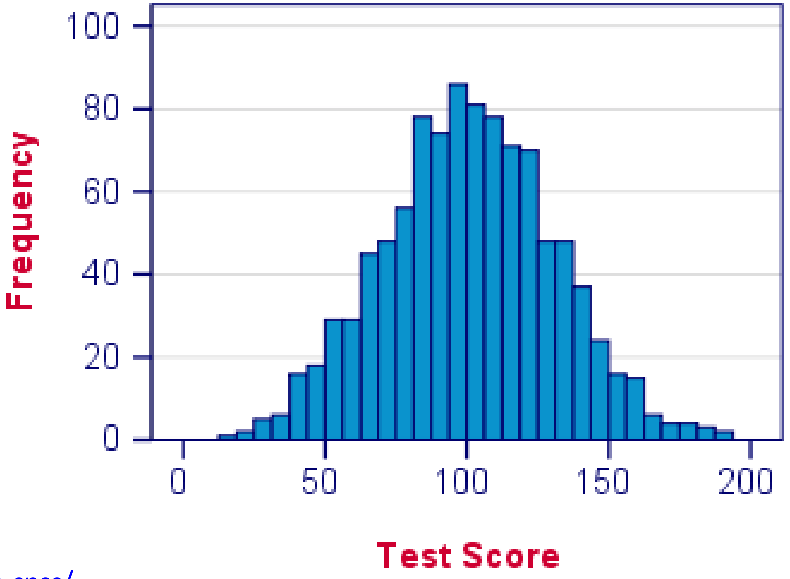
So, using the Standard Deviation we have a **"standard"** way of knowing what is **normal**, and what is **extra large** or **extra small**

Directions with most **variance**



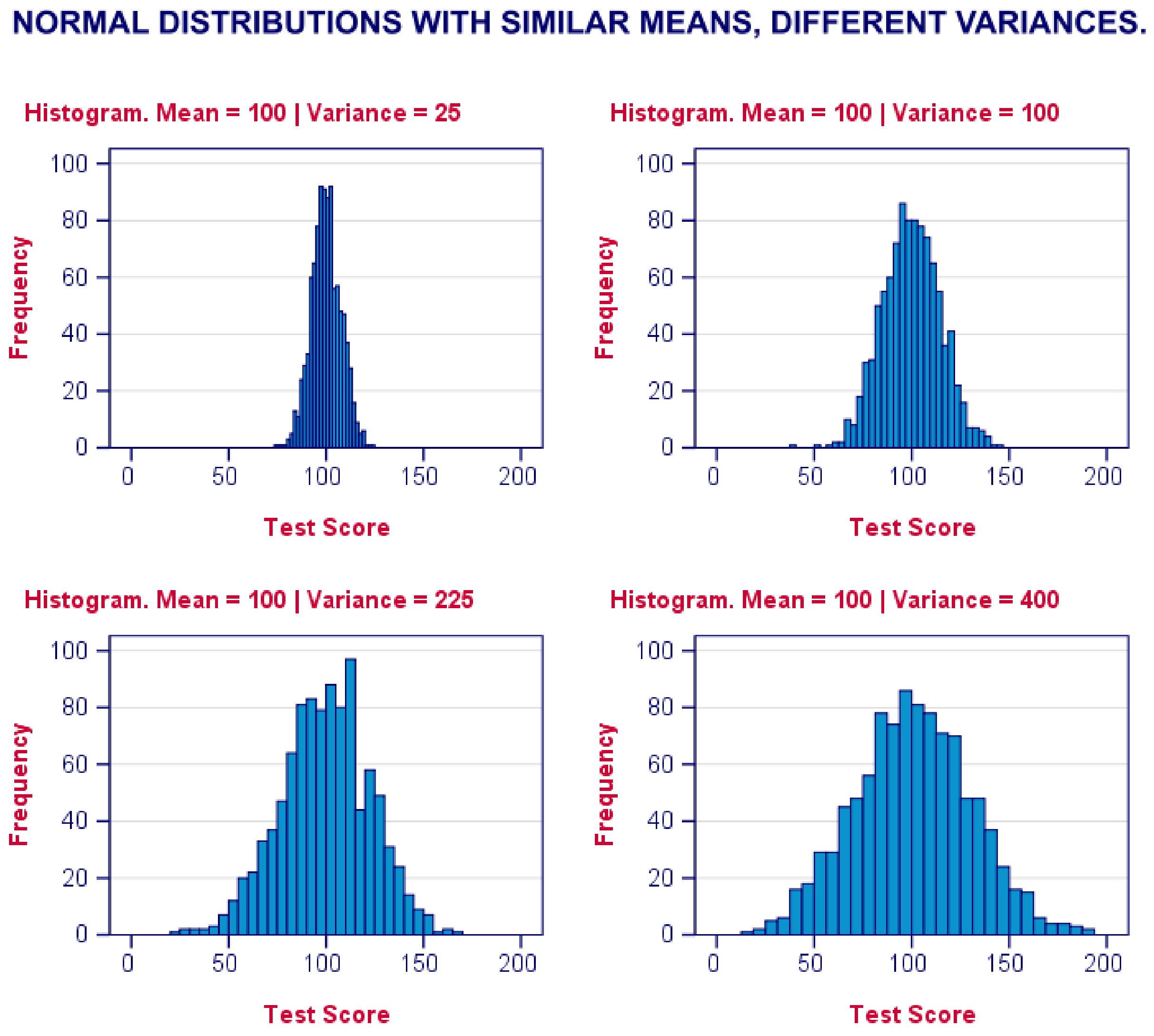






https://www.spss-tutorials.com/levenes-test-in-spss/

Directions with most variance

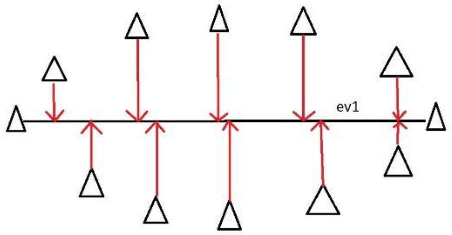


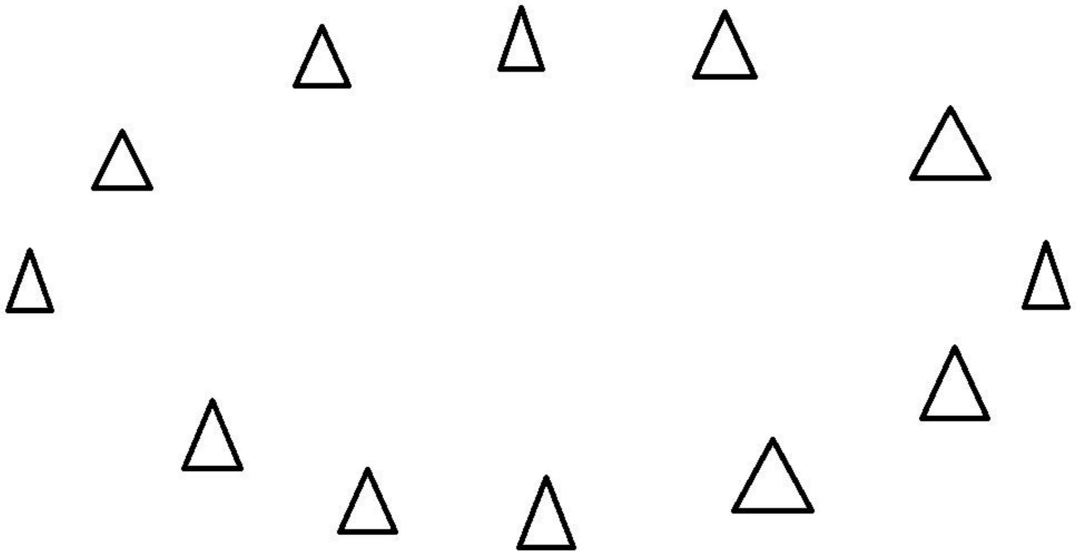
https://www.spss-tutorials.com/levenes-test-in-spss/

Directions with most variance

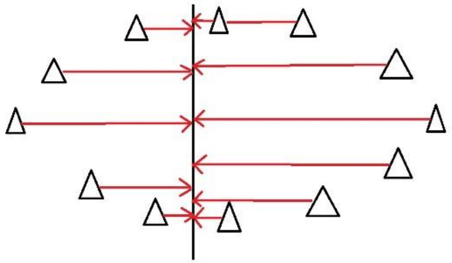
Assume that the triangles are data points.

?





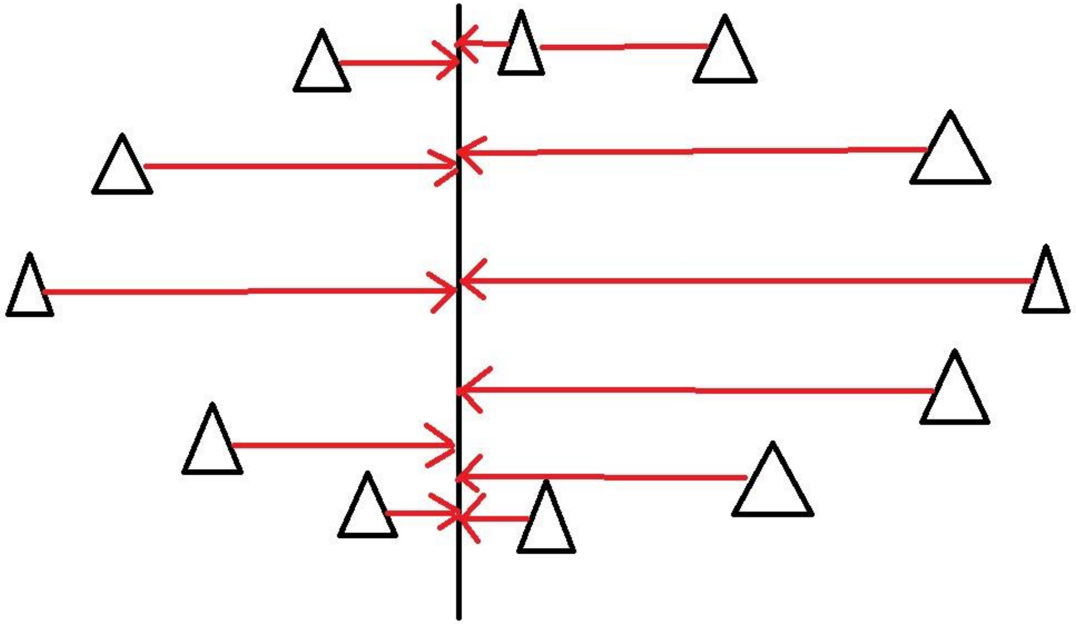
?



Find the straight line where the data is most spread out when projected onto it.

Directions with most variance

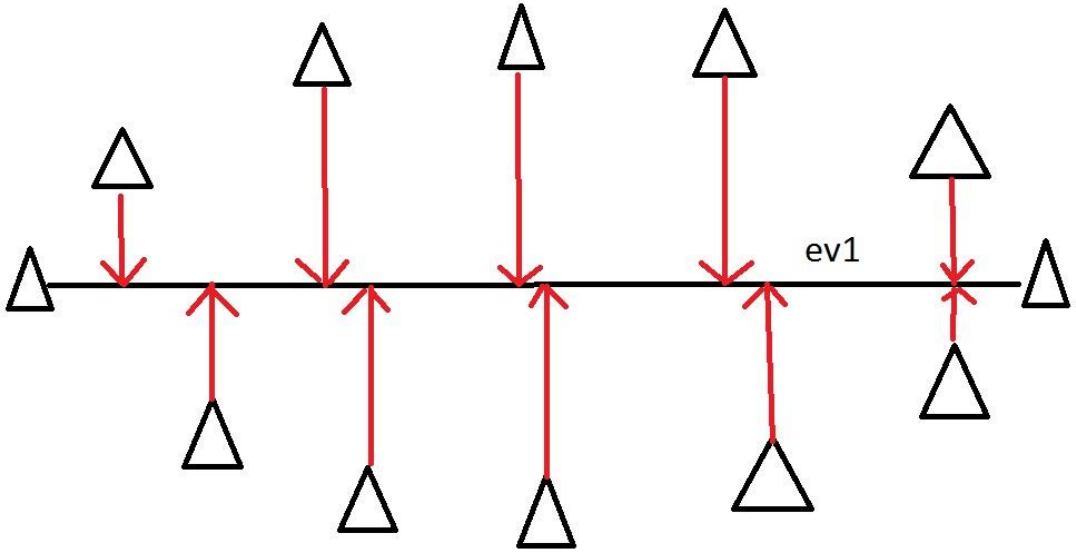
A vertical straight line with the points projected on to it will look like this.



The data is not very spread out here, therefore it does not have a large variance. It is probably not the principal component.

Directions with most variance

Now consider a horizontal line with lines from data points projected on this:

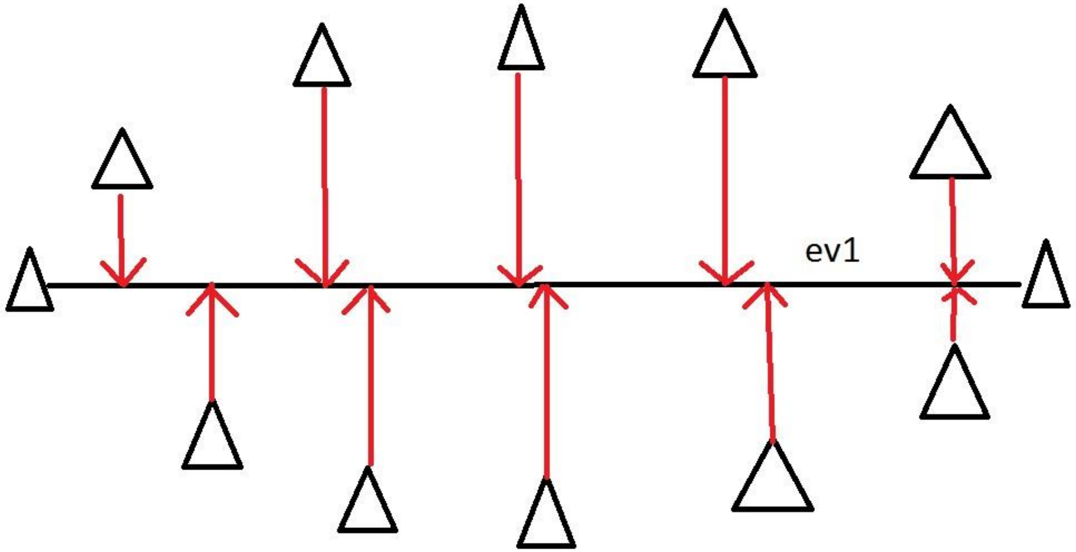


On this line the data is way more spread out and it has a large variance.

The horizontal line is therefore the principal component in this example.

PC | eigenvector and eigenvalue

Now consider a horizontal line with lines from data points projected on this:

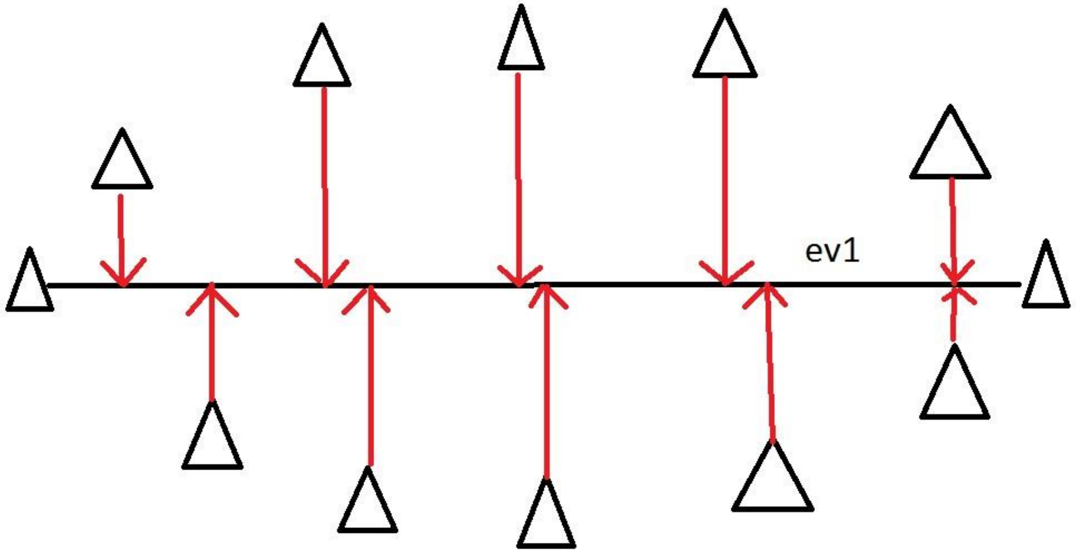


On this line the data is way more spread out and it has a large variance.

The horizontal line is therefore the principal component in this example.

PC | eigenvector and eigenvalue

The horizontal line is therefore the principal component in this example.



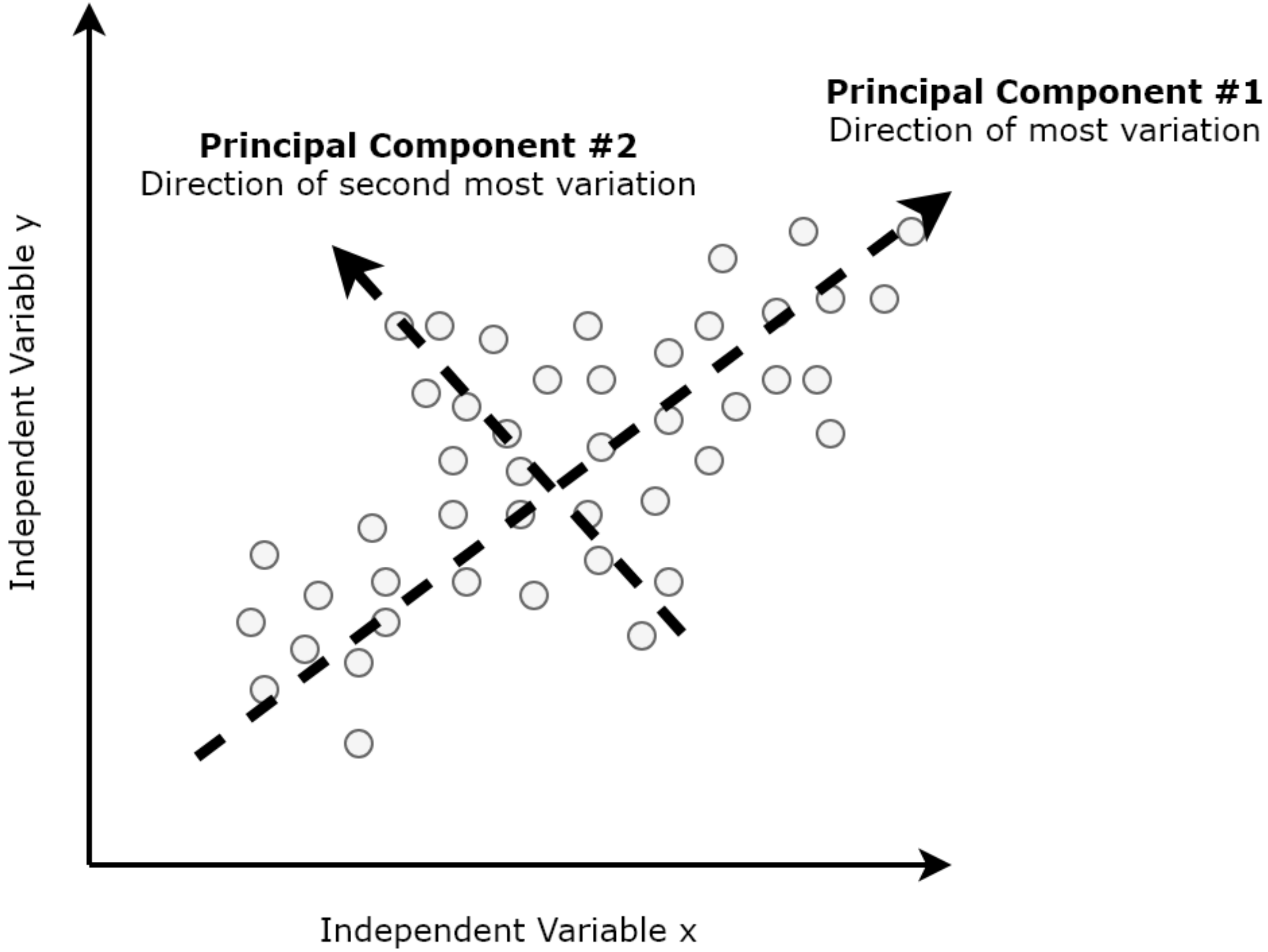
The direction of this line is called **eigenvector**.

An **eigenvalue** is a number telling us how spread out the data is on the line.

QUEEN'S UNIVERSITY BELFAST | SCHOOL OF ELECTRONICS, ELECTRICAL ENGNIEERING AND COMPUTER SCIENCE

Eigenvector and eigenvalue

The Principal component directions are directions in the feature space along which the original data are high variable.



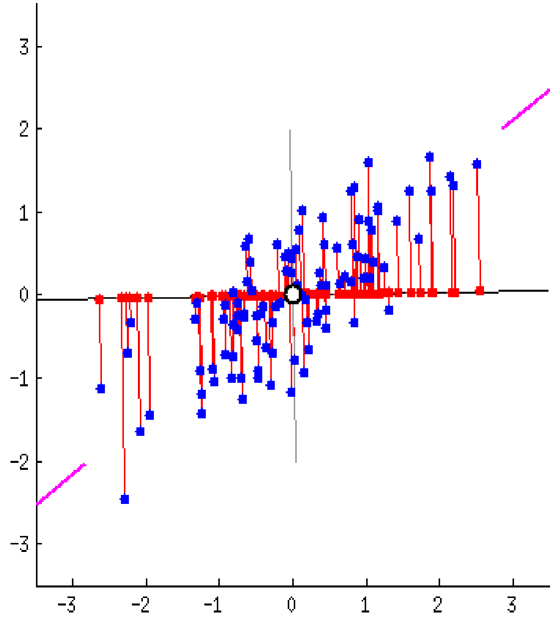
An **eigenvector** is a direction of the line

An **eigenvalue** is a number telling us how spread out the data is on the line

The eigenvector with the highest eigenvalue is therefore the principal component

PCA

- PCA allows us to **summarise** and to **visualise** the information in a data set containing individuals/observations/samples described by multiple inter-correlated quantitative variables.
- Each variable could be considered as a different dimension. If you have more than 3 variables in your data sets, it could be very difficult to visualize a multi-dimensional hyperspace.



PCA

PCA is used **to extract** the important information from a **multivariate** data table and to express this information as a set of few **new variables** called **PCs** (**principal components)**. These **new variables** correspond to **a linear combination of the originals**. The number of principal components is less than or equal to the number of **original variables**.

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/

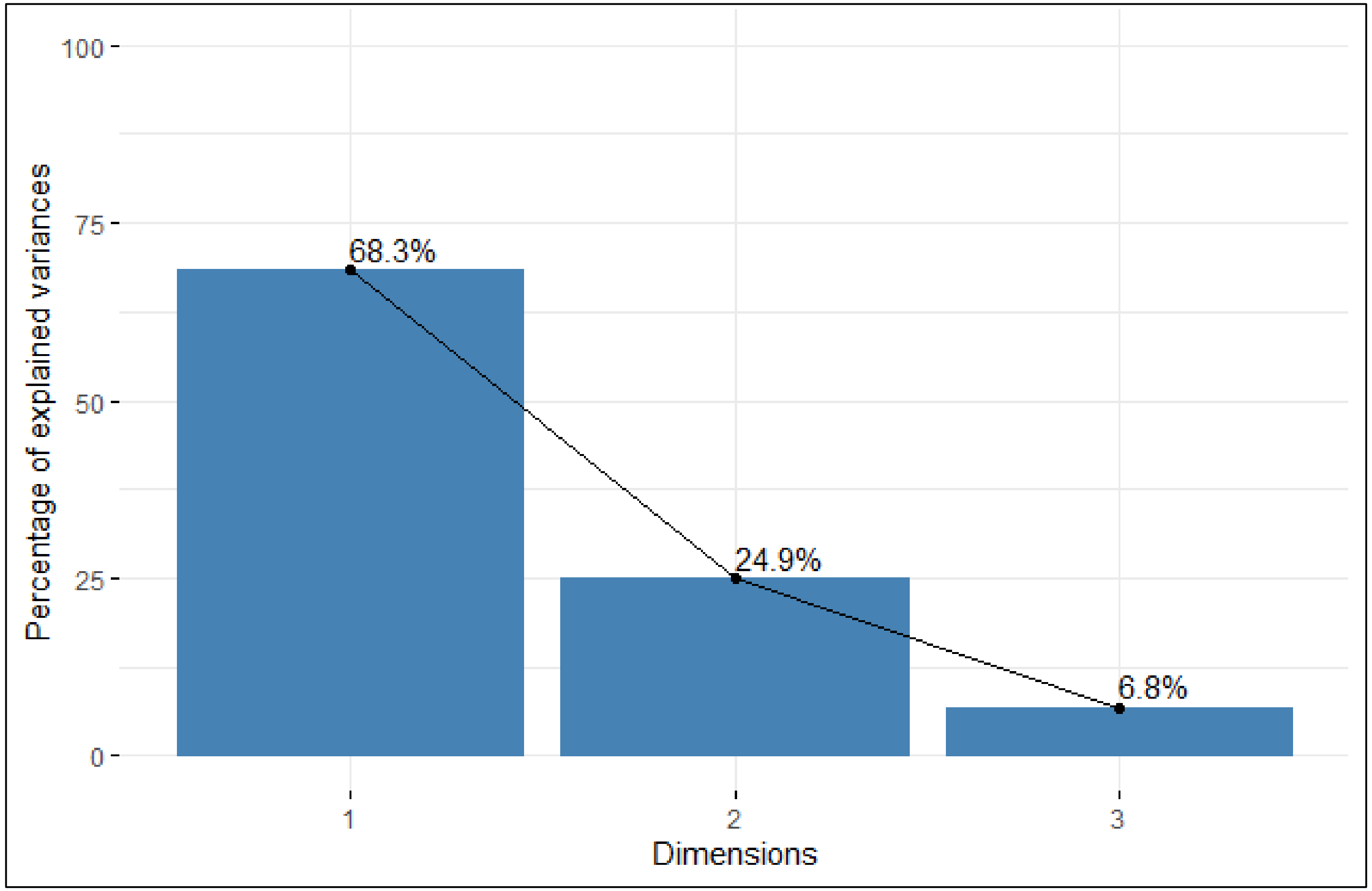
PCA

- A technique which is used to emphasise **variation** and **reveal strong patterns** in a dataset. It is often used to make data easy to **explore** and **visualise**.
- It's an **unsupervised learning method** and is similar to clustering.
- It could be considered as a **compression method**.
- Each **feature** could be considered as a different **dimension**. If you have more than 3 features in your dataset, it could be very difficult to visualise!
- Trade of between accuracy and **simplicity**

PCA analysis using FactoMineR() package

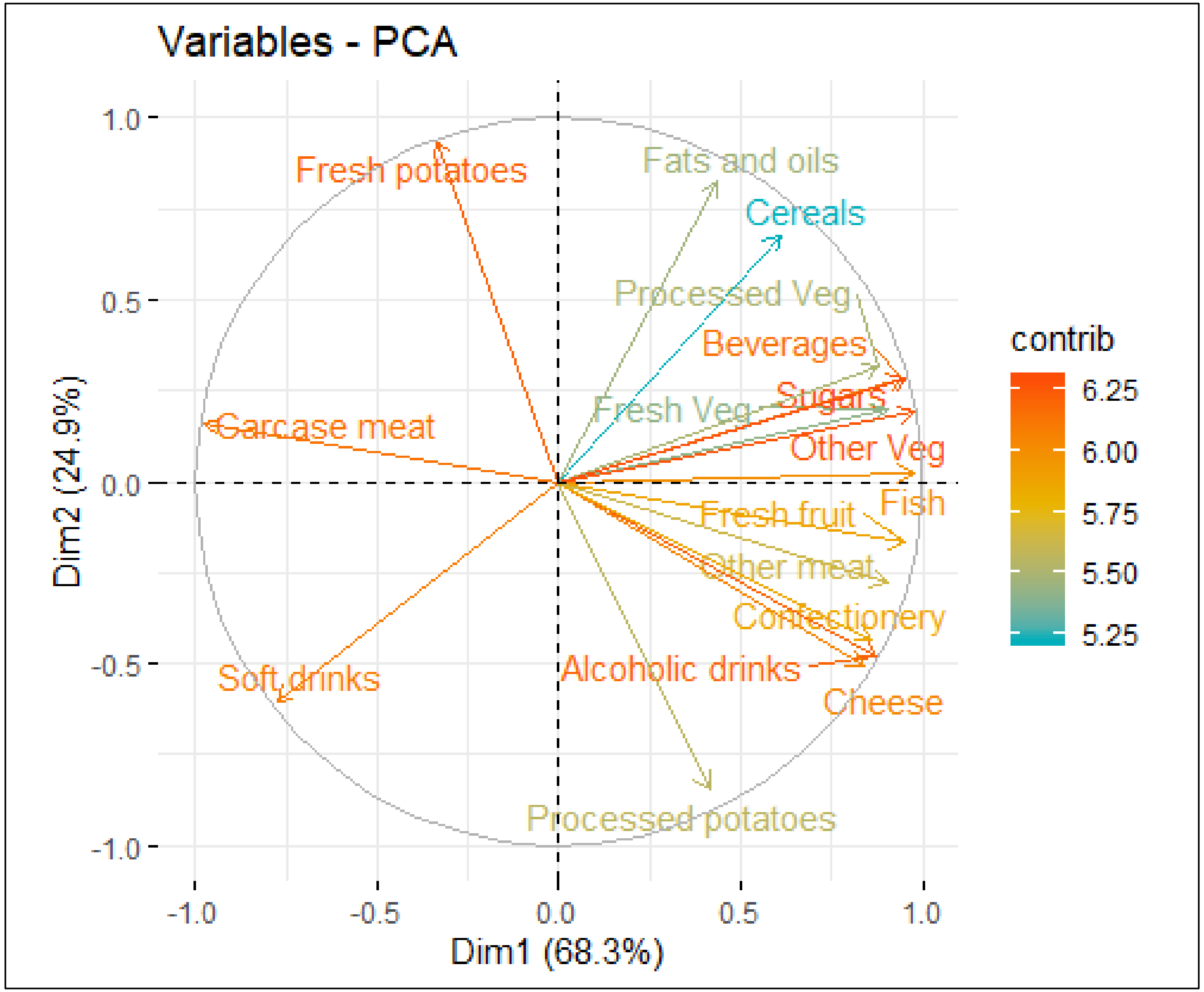
```
PCA_Model_Input_Dataset <- PCA(Input_dataset, scale.unit = TRUE, ncp = 17, graph = TRUE)
# Rows are individuals and columns are numeric variables, ncp: number of dimensions
print(PCA_Model_Input_Dataset)
```

```
** Results for the Principal Component Analysis (PCA)
** The analysis was performed on 4 individuals, described by 17 variables
* The results are available in the following objects:
    name description
1 "$eig" "eigenvalues"
2 "$var" "results for the variables“
3 "$var$coord" "coord. for the variables“
4 "$var$cor" "correlations variables - dimensions“
5 "$var$cos2" "cos2 for the variables“
6 "$var$contrib" "contributions of the variables“
7 "$ind" "results for the individuals“
8 "$ind$coord" "coord. for the individuals"
9 "$ind$cos2" "cos2 for the individuals“
10 "$ind$contrib" "contributions of the individuals“
11 "$call" "summary statistics“
12 "$call$centre" "mean of the variables“
13 "$call$ecart.type" "standard error of the variables"
14 "$call$row.w" "weights for the individuals"
15 "$call$col.w" "weights for the variables"
```




```
fviz_eig(PCA_Model_Input_Dataset, addlabels = TRUE, ncp = 3, ylim = c(0, 100))
```

The Northern Irish eat way more grams of fresh potatoes.



Any Questions?