MONASH University

MONASH BUSINESS SCHOOL

Distance

High Dimensional Data Analysis

Anastasios Panagiotelis

Lecture 3

Why distance?

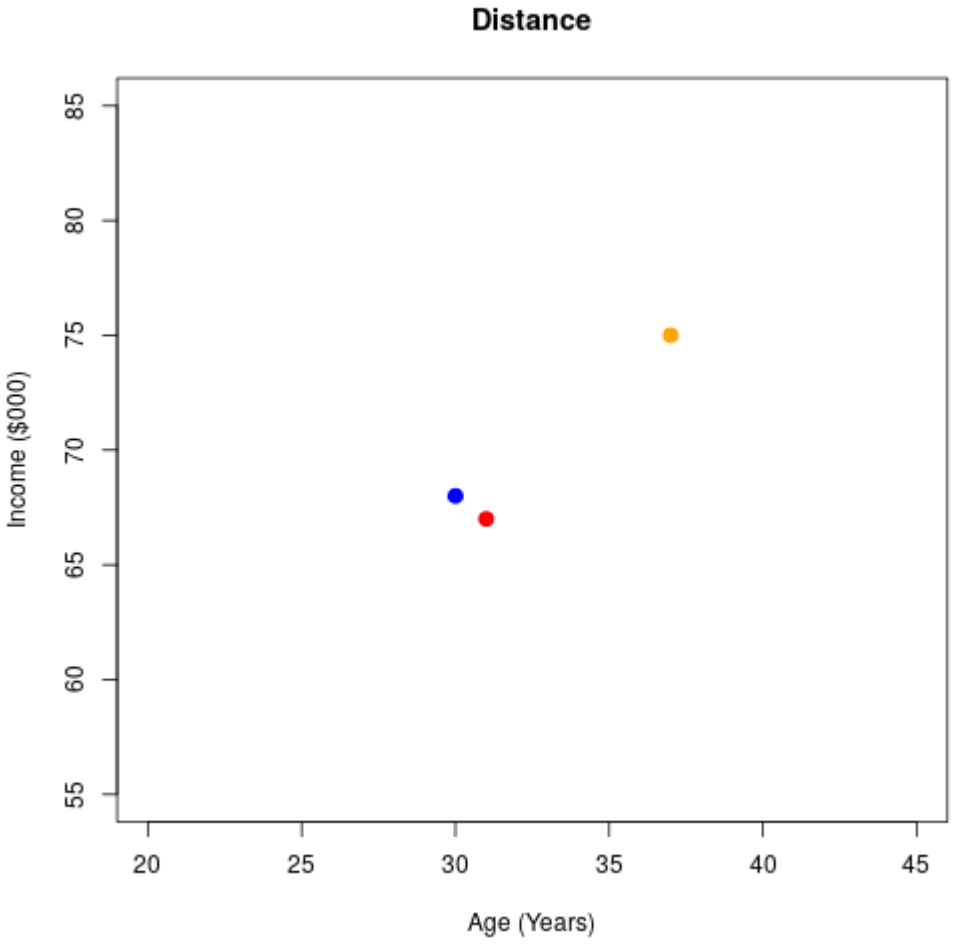
Why distance?

- Many problems that involve thinking about how *similar* or dissimilar two observations are. For example:
  - May use the same marketing strategy for *similar* demographic groups.
  - May lend money to applicants who are *similar* to those who pay debts back.
- Arguably the most important concept in data analysis is *distance*

Simple example

- Consider 3 individuals:
  - Mr Orange: 37 years of age earns $75k a year
  - Mr Red: 31 years of age earns $67k a year
  - Mr Blue: 30 years of age earns $68k a year
- Which two are the most similar?

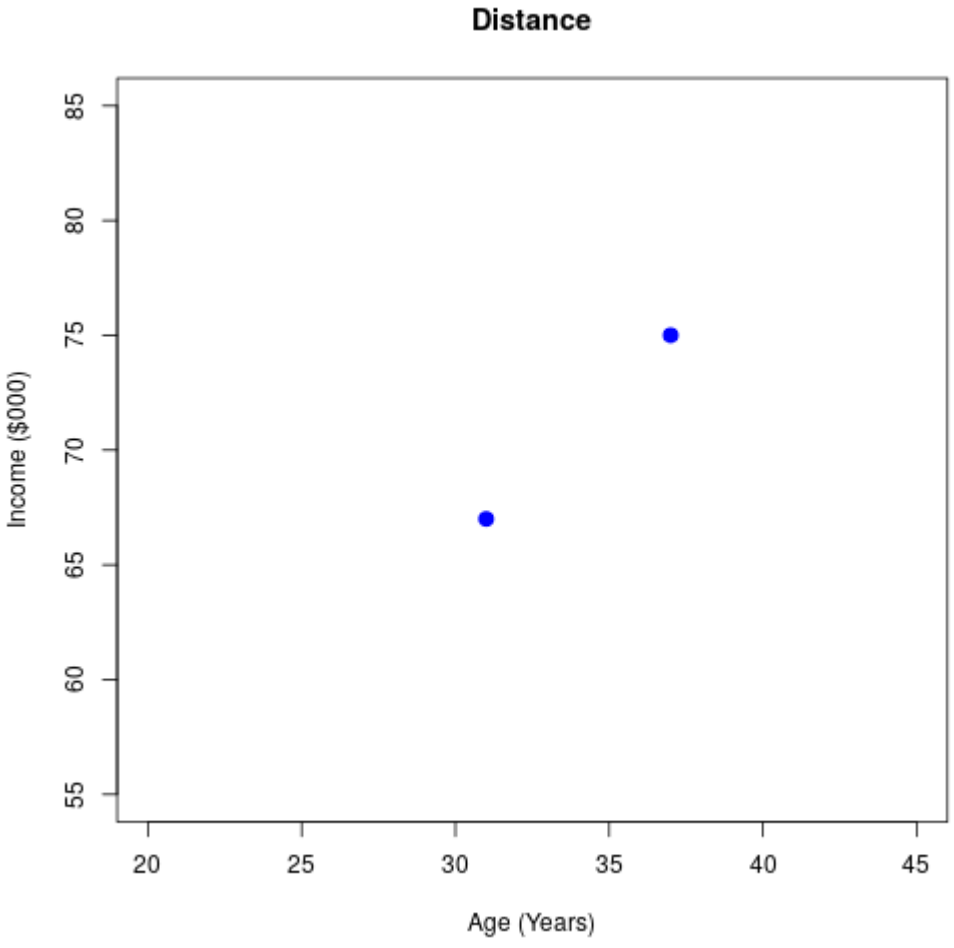
On a scatterplot



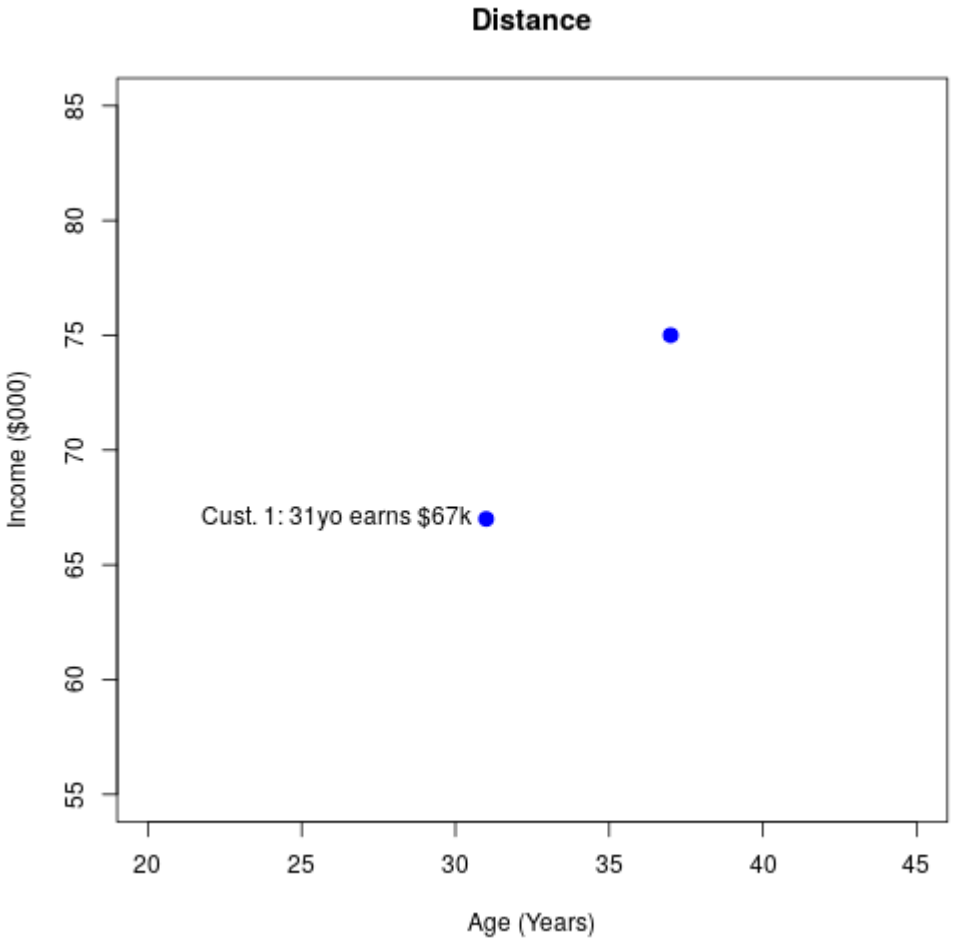
Distance as a number

- It is easy to think about three individuals but what if there are thousands of individuals?
  - In this case it will be useful to attach some number to the distance between pairs of individuals
  - We will do it with a simple application of Pythagoras' theorem.

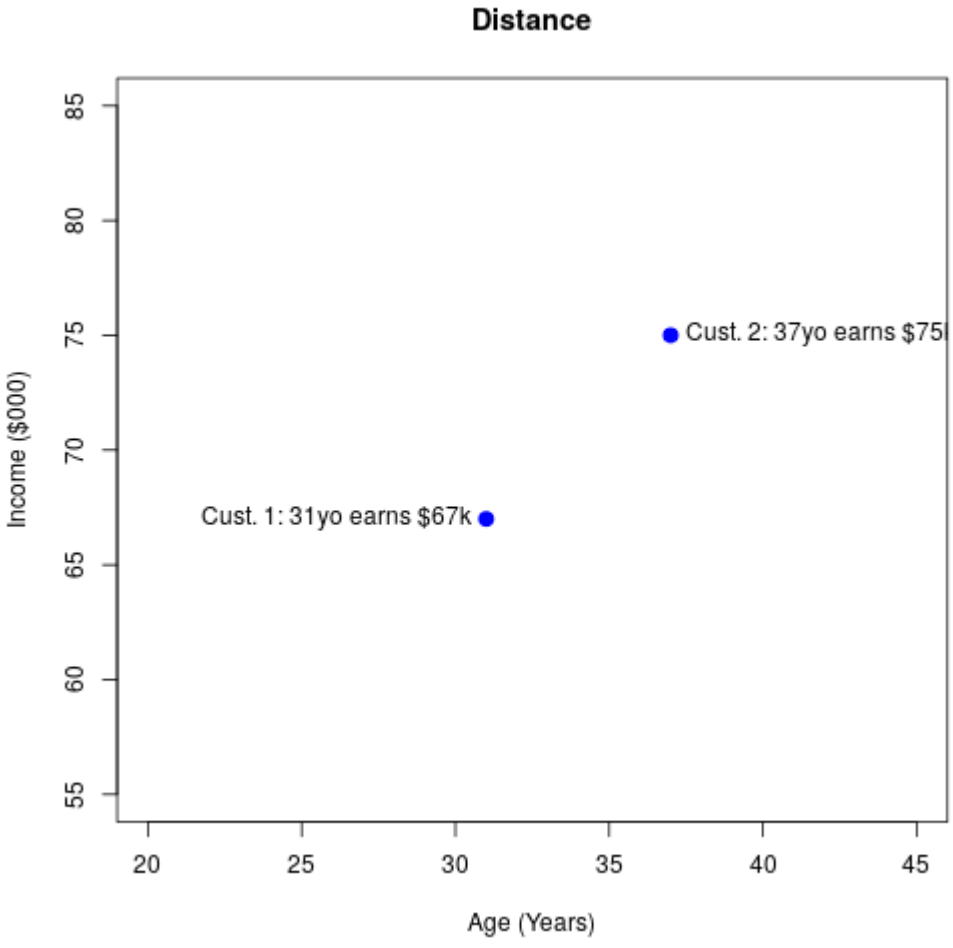
Finding the Distance



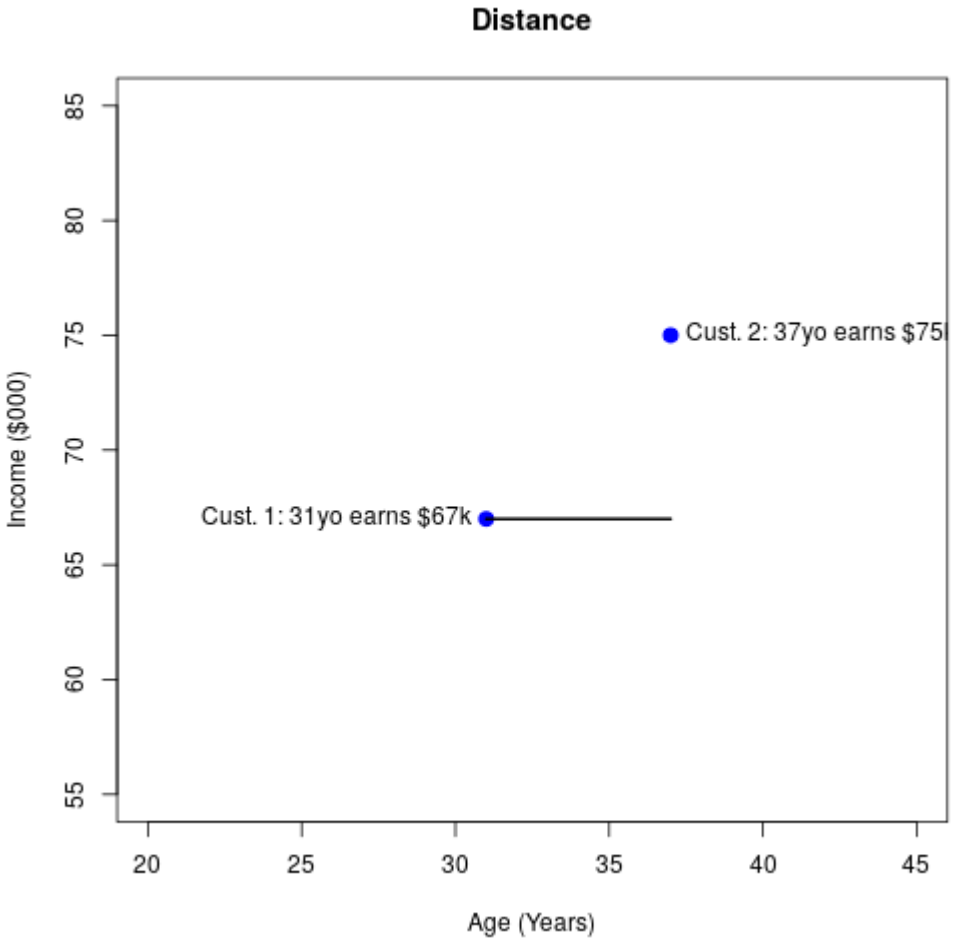
Finding the Distance



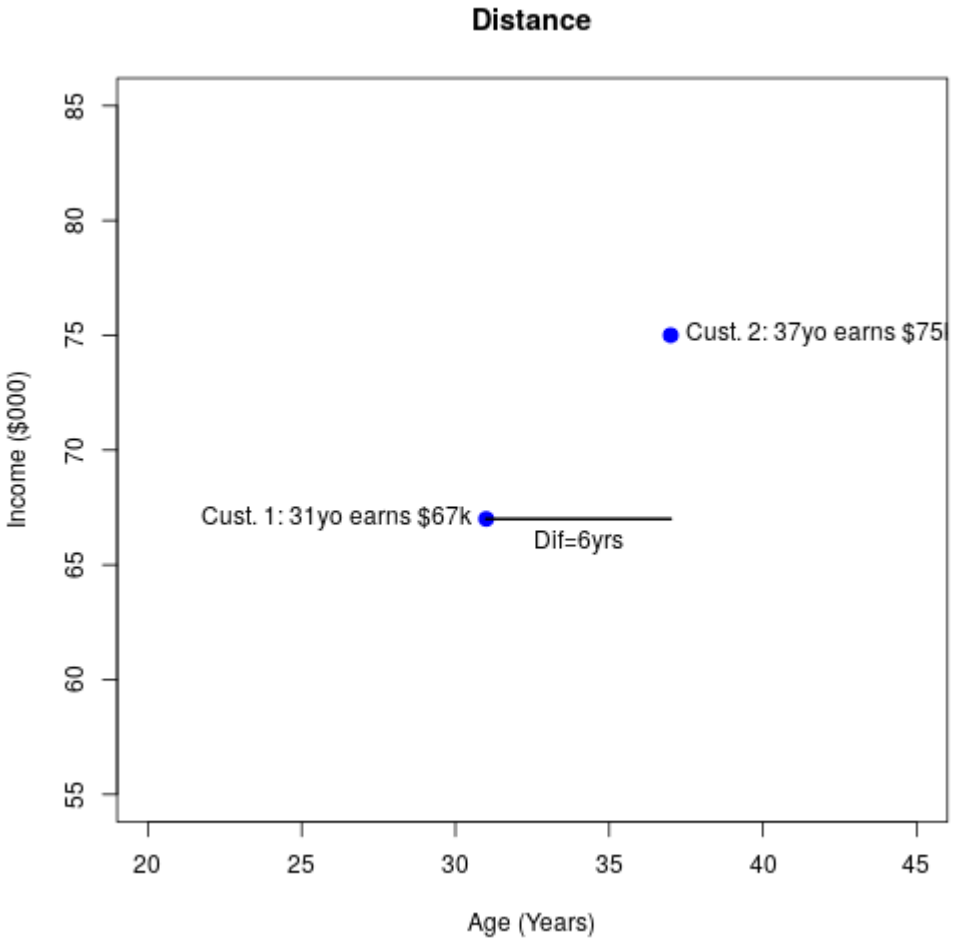
Finding the Distance



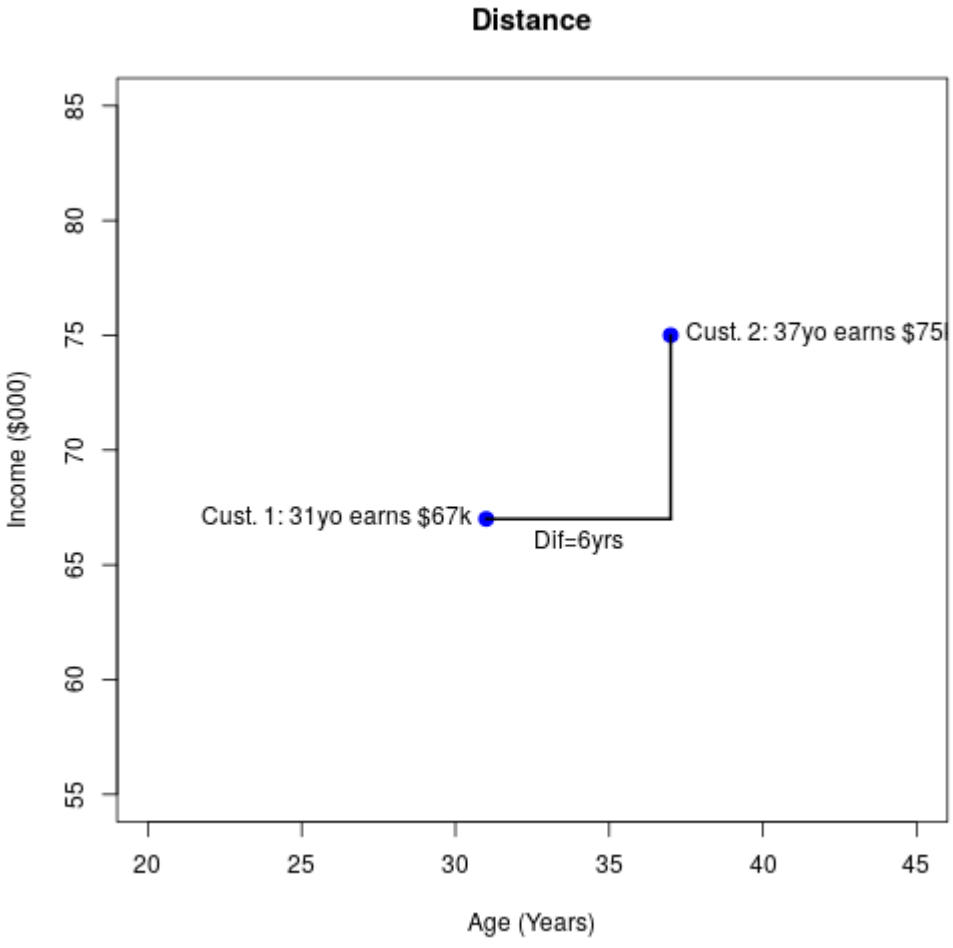
Finding the Distance



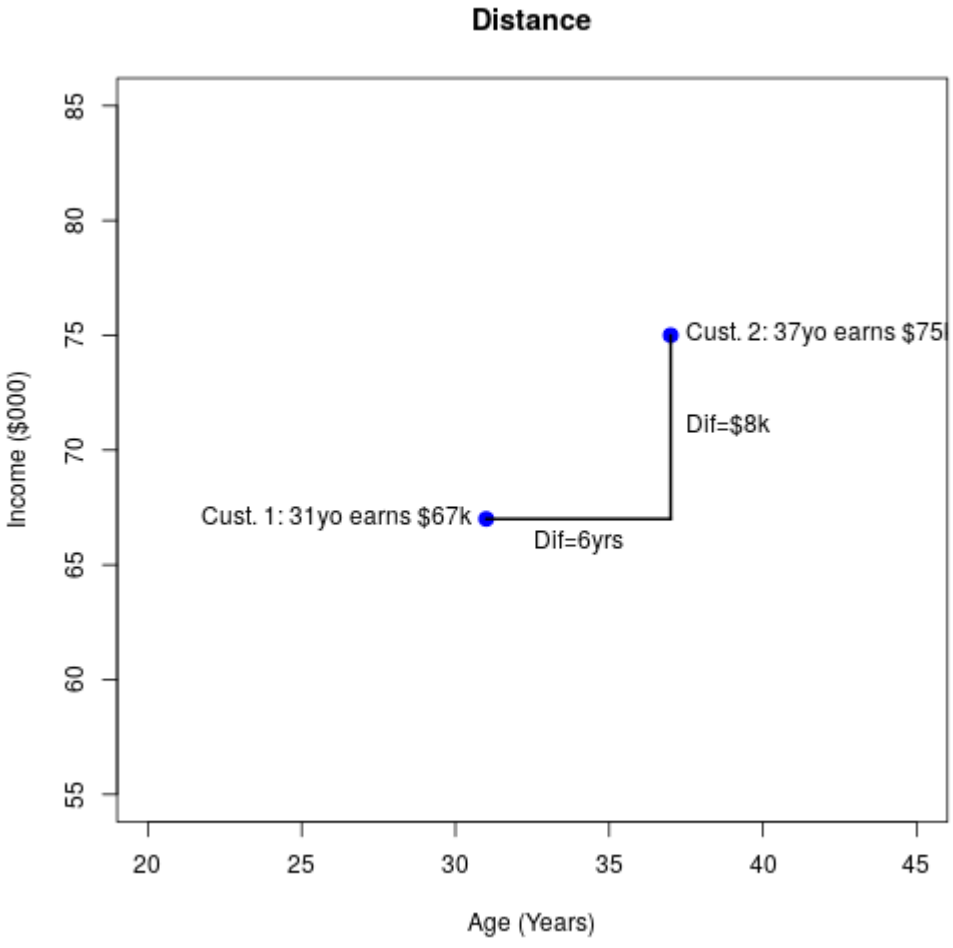
Finding the Distance



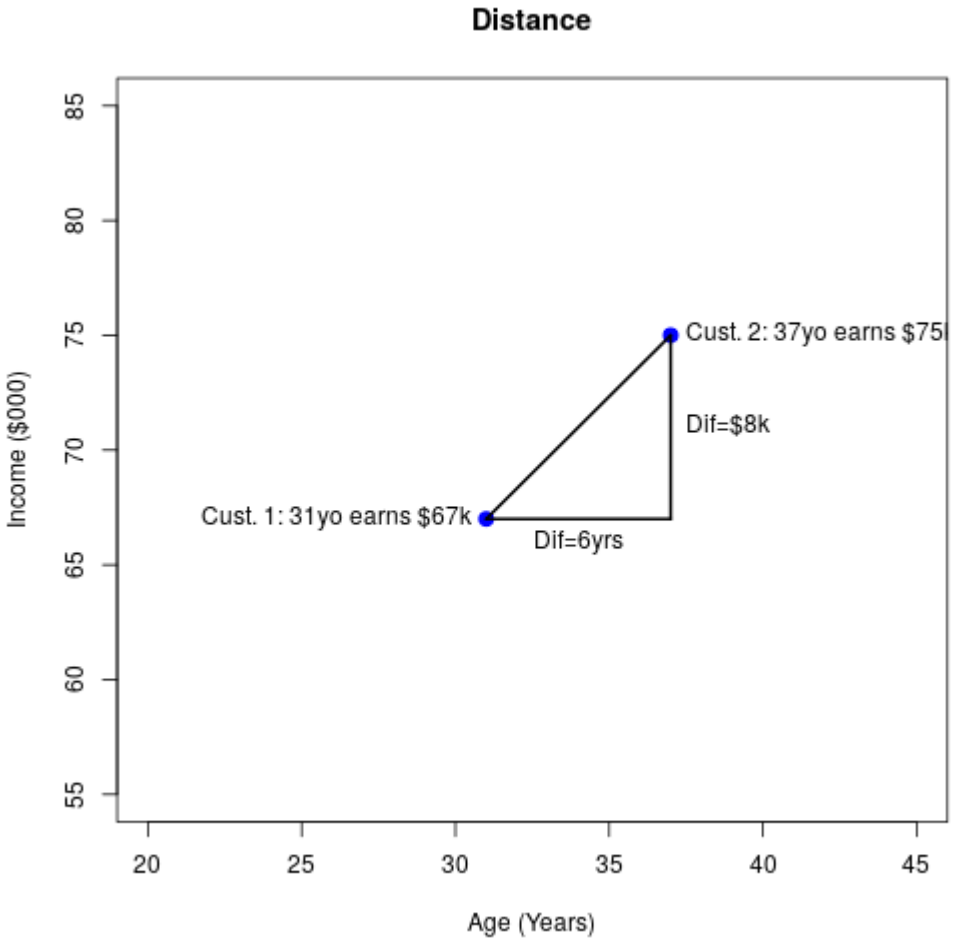
Finding the Distance



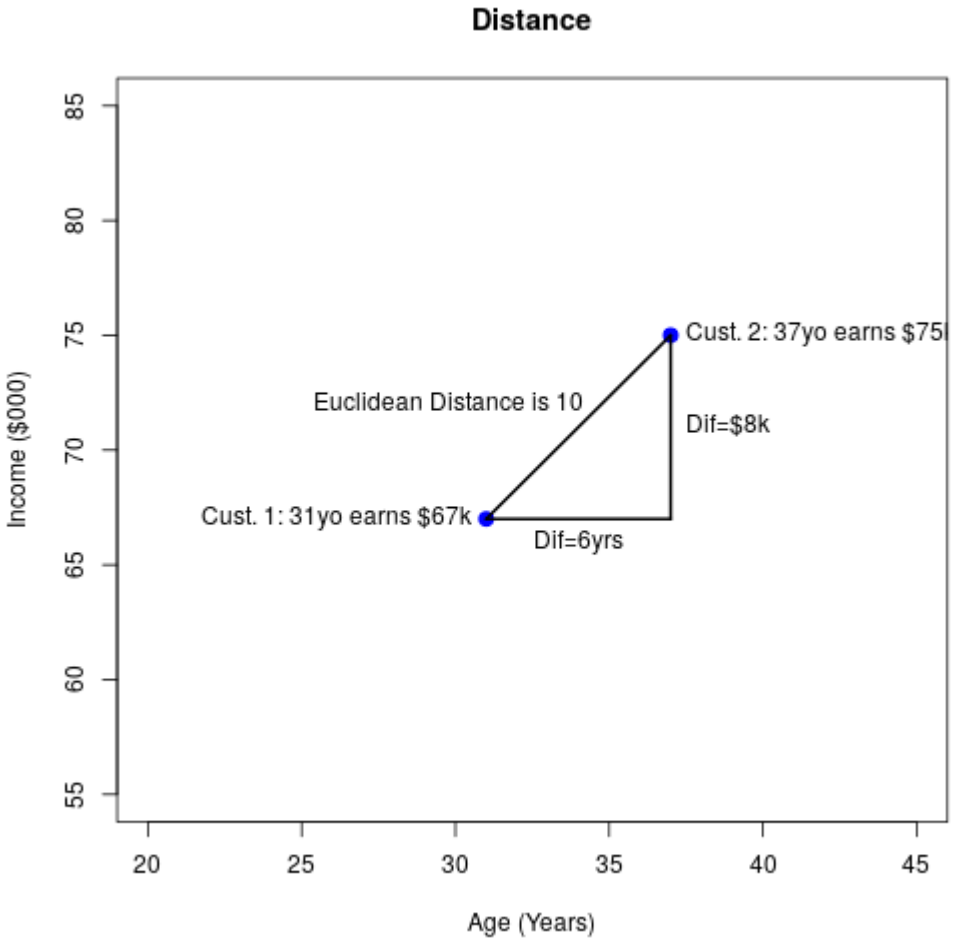
Finding the Distance



Finding the Distance



Finding the Distance



Euclidean distance

- In general there are more than two variables.
- Is there a way to apply our intuition in 2 dimensions to higher dimensions?
  - Pythagoras' theorem can be *generalised* to higher dimensions.
  - This results in a concept of distance called *Euclidean distance*.

Euclidean distance

We measure p variables for two observations: x_j is the measurement of variable j for observation $\mathbf{x}$, y_j is the measurement of variable j for observation $\mathbf{y}$. *Euclidean* distance between $\mathbf{x}$ and $\mathbf{y}$ is:

$$D\left(\mathbf{x},\mathbf{y}\right)=\sqrt{\sum_{j=1}^{p}\left(x_j-y_j\right)^2}$$

Vectors

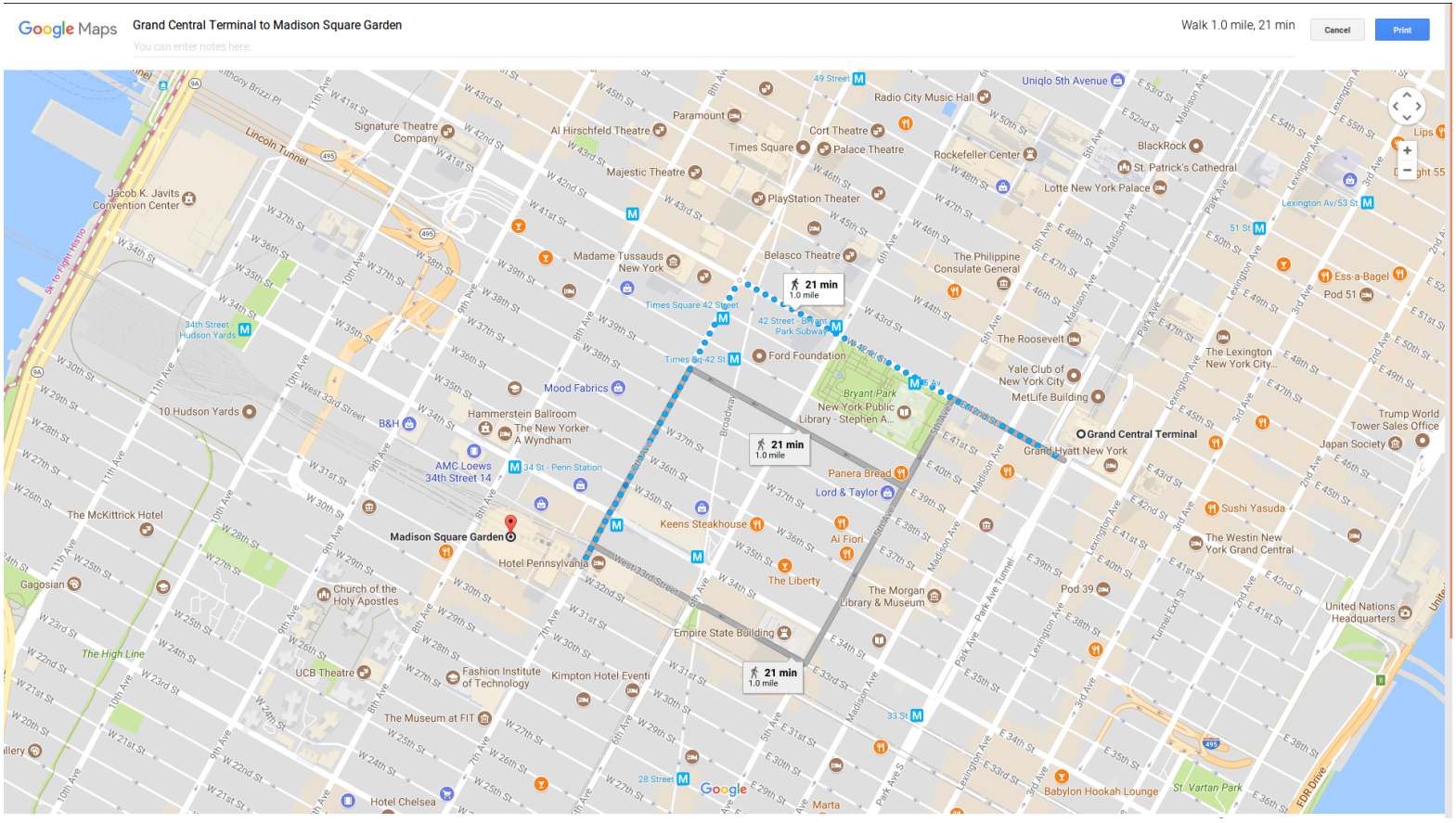
- Notice that $\mathbf{x}$ and $\mathbf{y}$ are examples of **vectors**.
- For example $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ where x_1 is age and x_2 is income.
- We can think of a data point as
  - A vector of attributes or measurements
  - A point in space
- These are the same thing.

Other kinds of distance

- We will nearly always use Euclidean Distance in this unit, however there are other ways of understanding distance
- One example is the *Manhattan Distance* also known as block distance.

$$D(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} |x_j - y_j|$$

Manhattan Distance



Distance and Standardising data

- We must be careful about the units of measurement.
- Euclidean (and Manhattan) distance change for variables measured in *different units*.
- For this reason, it is common to calculate distance after the *standardising* data.
- If the variables are all measured in the same units, then this standardisation is unecessary.
- Some distances are not sensitive to units of measurement (e.g. Mahalanobis Distance)

Distance in R

- R has its own special object for distances known as a `dist` object
- It can be obtained using the `dist()` function
- We are going to find Euclidean distances between the beers in the beers dataset. Use:
  - Only beers with price greater than $4.50
  - Only numeric variables.
  - Standardised data
  - Use the function `dist` to get the distances.

Load packages and data

```
library(dplyr)
Beer<-readRDS('Beer.rds')
```

Find Distances

```
Beer%>%filter(price>4.5)%>% #Only expensi
  select_if(is.numeric)%>% #Only numeric
  scale%>%
  dist->d
```

1	2	3	4	5
0.0000	3.4298	3.8333	4.1632	4.1950
3.4298	0.0000	2.3009	2.8076	1.6260
3.8333	2.3009	0.0000	1.1482	3.2339
4.1632	2.8076	1.1482	0.0000	3.3188
4.1950	1.6260	3.2339	3.3188	0.0000

Labels

- Only numeric variables were used to compute distances.
- The names of the beers are not attached to the `dist` object.
- This can be achived by assigning the beer names to `attributes(d)$Labels`
- Here `d` is the `dist` object.

Use Beer Names

```
Beer%>%filter(price>4.5)%>% #Only expensi
  pull(beer)-> #Get beer names
  attributes(d)$Labels #"Attach" them to
```

	Anchor Steam	Becks	Heineken	Kirin	St Pauli Girl
Anchor Steam	0.0000	3.4298	3.8333	4.1632	4.1950
Becks	3.4298	0.0000	2.3009	2.8076	1.6260
Heineken	3.8333	2.3009	0.0000	1.1482	3.2339
Kirin	4.1632	2.8076	1.1482	0.0000	3.3188
St Pauli Girl	4.1950	1.6260	3.2339	3.3188	0.0000

Your Turn

- Compute the distance without standardising the data.
- Compute the Manhattan distance for standardised data.
- Compute the Manhattan distance for unstandardised data.

Non-Metric

Non-metric Data

- Can we define distance when the variables are non metric?
- The answer is yes!
- We will discuss two approaches:
  - Jaccard Similarity/ Distance
  - Dummy Variables

First a motivation

- Many people use music streaming services like Spotify.
- One of the attractions of these services is they they recommend artists based on the favourite artists of other users who have similar taste in music.
- The data in this case is in the form of a list of favourite artists.

Distance in musical taste

- Suppose there are three customers with the following favourite artists
  - Customer A: Maroon 5, Ariana Grande, Ed Sheeran, Cardi B
  - Customer B: Maroon 5, Ed Sheeran, BTS
  - Customer C: Cardi B, Drake, Future
- How do we measure which customers have similar taste and which have different taste?

Jaccard Similarity and Distance

- Jaccard similarity gives us a measure of how close two *sets* are, in this case the set of each customers favourite musician. The formula is

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

- Where $|A \cap B|$ is the number of elements in both set A and set B and $|A \cup B|$ is the number of elements in either set A or set B.

Jaccard Similarity

- In our example
  - $A \cap B = \{\text{Maroon 5, Ed Sheeran}\}$
  - $|A \cap B| = 2$
  - $A \cup B = \{\text{Maroon 5, Ariana Grande, Ed Sheeran, Cardi B, BTS}\}$
  - $|A \cup B| = 5$
- The Jaccard similarity will be $J = 2/5 = 0.4$.

  The Jaccard *distance* is

  $d_J = 1 - J = 1 - 0.4 = 0.6$

Using dummy variables

- Alternatively the same data can be coded using dummy variables:
  - $X_j = 1$ if artist j is a favourite of customer x
  - $X_j = 0$ otherwise
- The usual distance measures such as Euclidean or Manhattan distance can then be used.

Conclusions

- That concludes the topic on distance.
- This is relevant to the following topics
  - Cluster Analysis
  - Multidimensional Scaling (MDS)
- Now one final exercise

Distances between tweets

- Find someone on Twitter or a similar social media site
  - Find the first two tweets
  - Think of a way to compute a Jaccard distance between their tweets
- Hint: Think of the words used in the tweet as a *set*