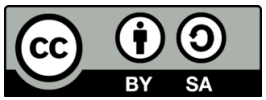
THE UNIVERSITY *of* EDINBURGH

Data Science in Medicine

Lecture 3: Visualising Data

Dr Areti Manataki

Usher Institute
The University of Edinburgh



Usher Institute
Population Health Sciences & Informatics

In the previous lecture

- Why is statistical analysis useful?
  - It allows us to describe existing data and to discover patterns in the data
  - It allows us to make inferences about a population based on a sample from that population.
- Data may be qualitative (categorical or ordinal scale) or quantitative (interval or ratio scale).
- Summary statistics involve measures of central tendency (e.g. mean, median, mode) and measures of dispersion (e.g. range, variance, standard deviation).

In the previous lecture

- For example, for this set of course grades {69, 70, 86, 42, 54, 79, 69} we have:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{69+70+86+42+54+79+69}{7} = \frac{469}{7} = 67$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

$$= \frac{(69-67)^2 + (70-67)^2 + (86-67)^2 + (42-67)^2 + (54-67)^2 + (79-67)^2 + (69-67)^2}{7}$$

$$= 188$$

In this lecture

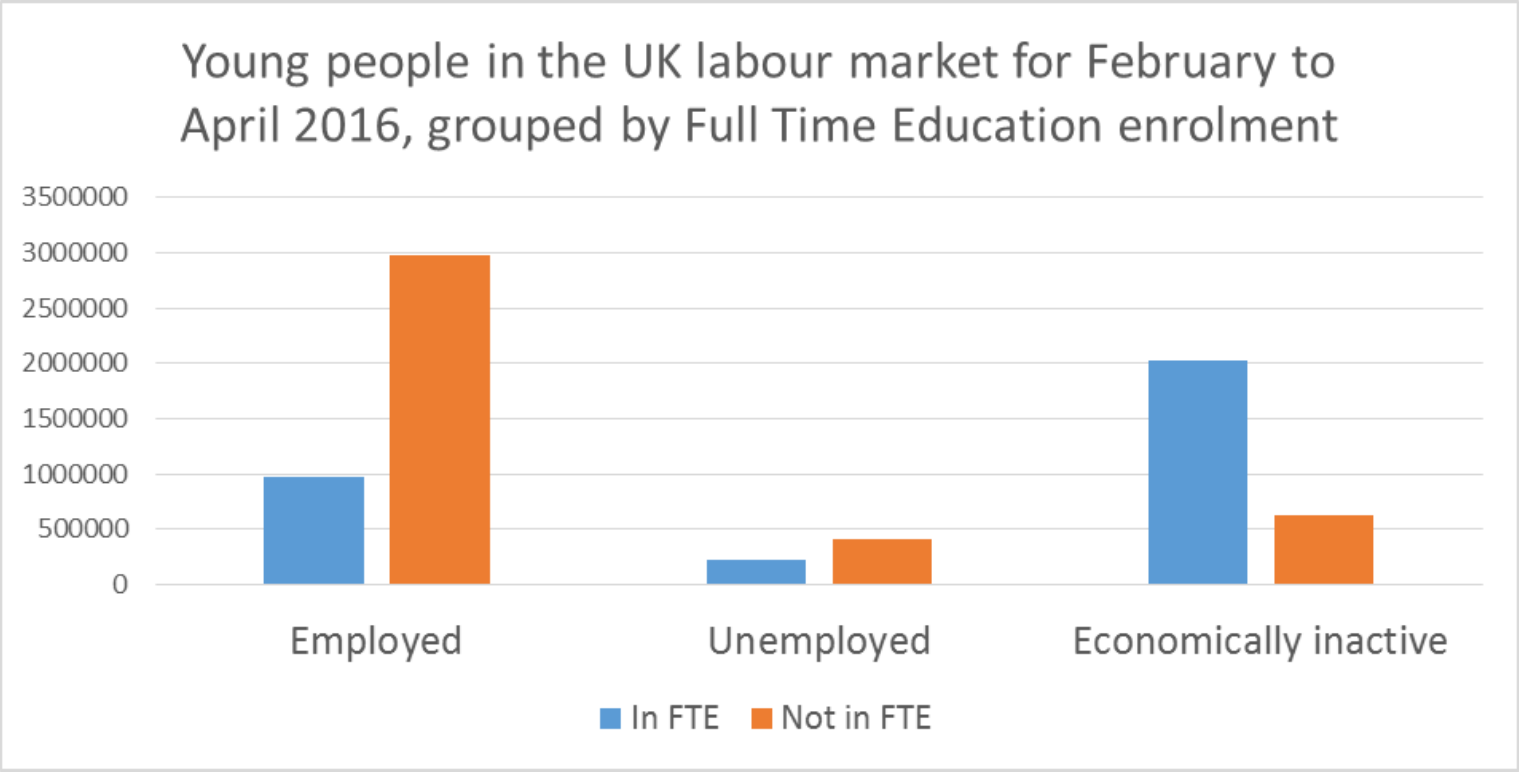
- Visualising data
  - Qualitative data: bar charts, pie charts
  - Quantitative data: histograms, box plots
  - Bivariate: scatter plots, line graphs
- Introduction to R
  - R essentials
  - Calculating summary statistics
  - Drawing simple graphs

Bar charts

- Bar charts are appropriate for data on a categorical or ordinal scale.
- They display the number of times each category occurs in the data.
- Several variations: grouped bar charts, stacked bar charts, etc.

Bar chart example

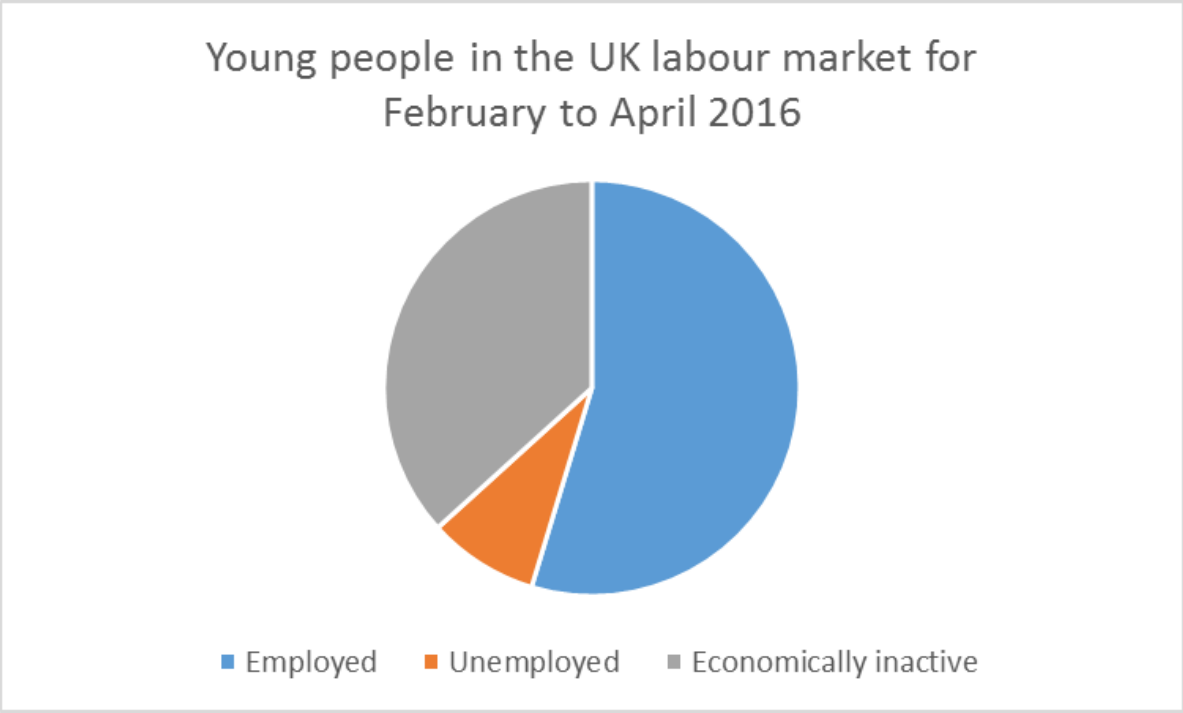
Grouped bar chart example



Pie charts

- Pie charts are also appropriate for data on a categorical or ordinal scale.
- They illustrate the relative proportion of data in each category.
- They capture part-whole relationships: the different slices should add up to a meaningful whole.
- They are most useful when there are only a few categories and the differences among these categories are fairly large.

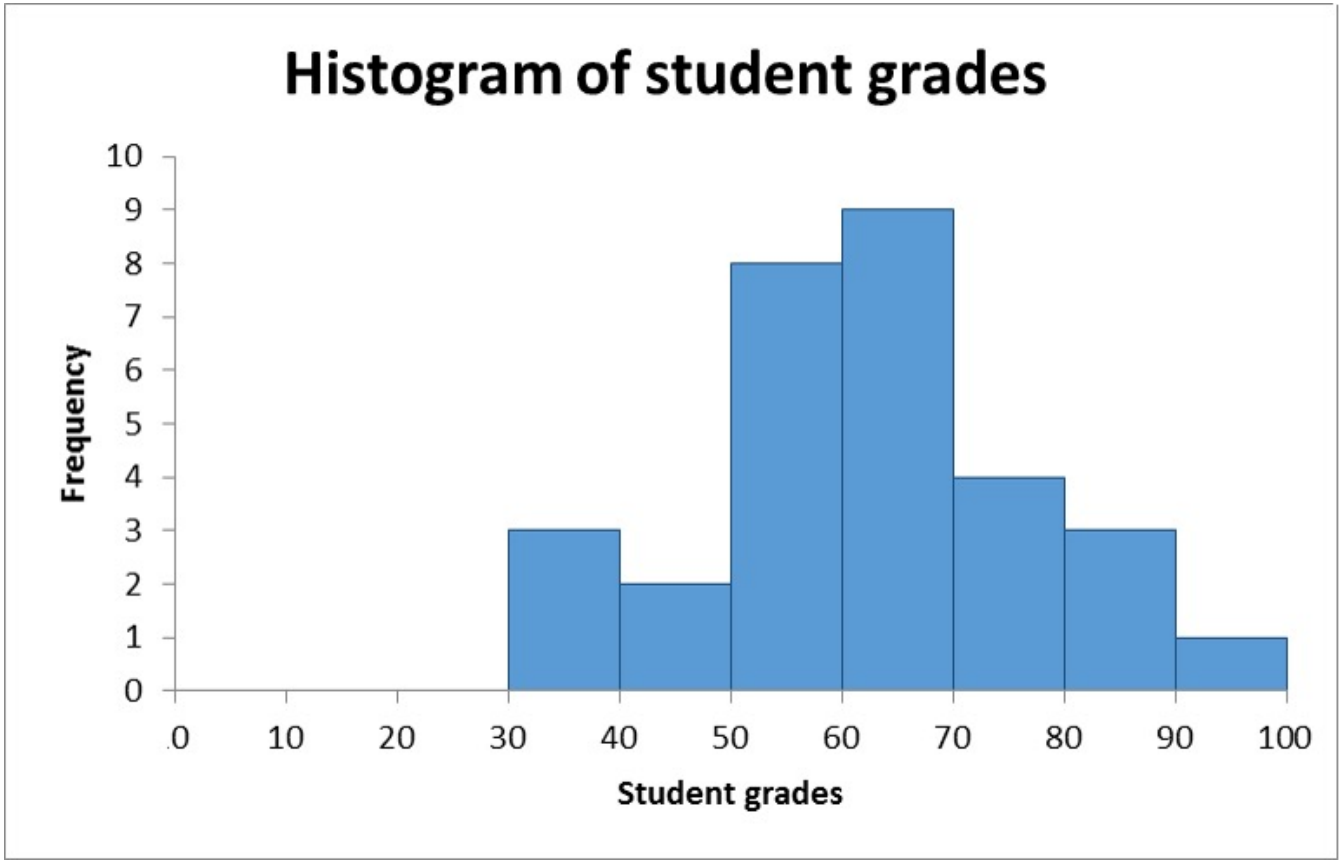
Pie chart example



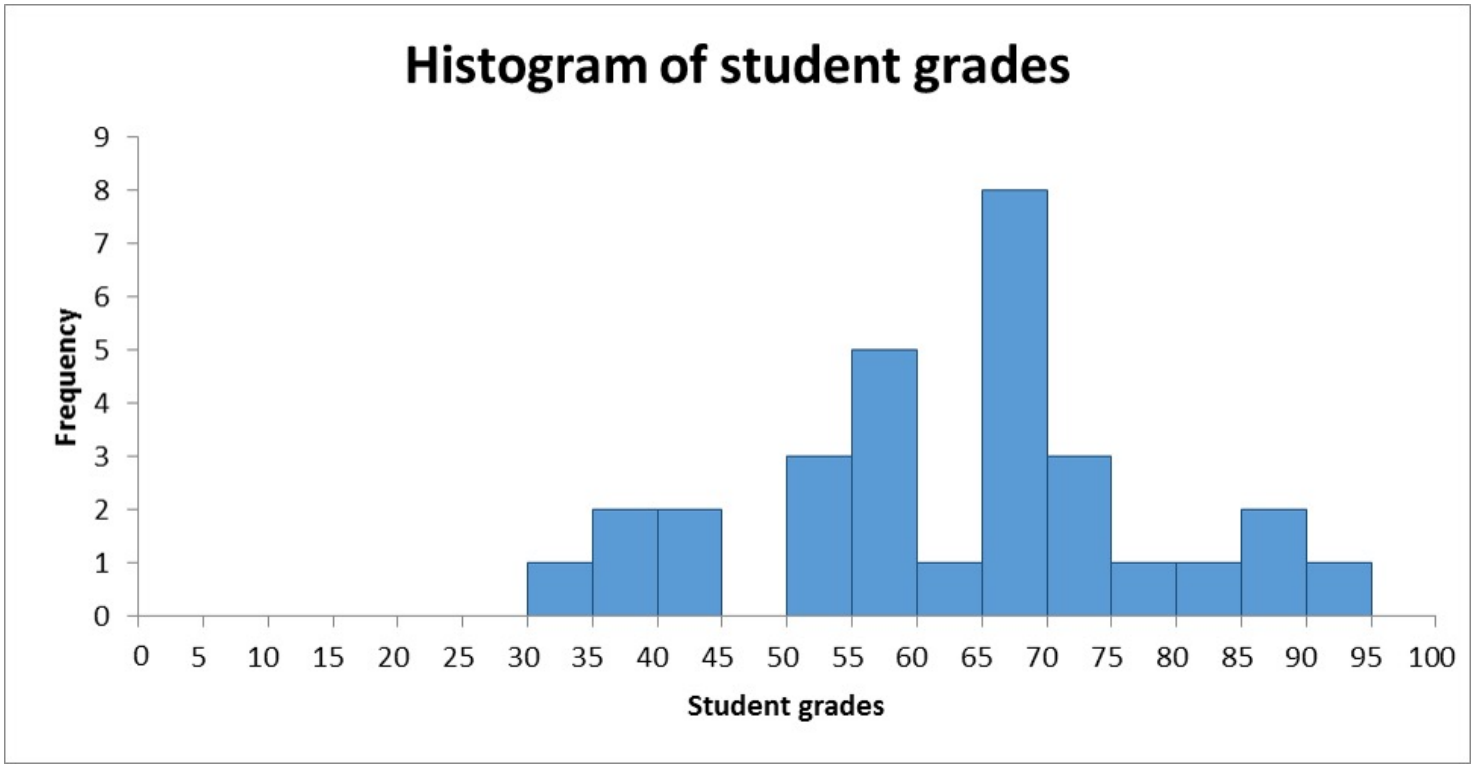
Histograms

- Histograms are appropriate for data on an interval or ratio scale.
- The data are collected into bands, and the histogram shows the frequency with which values occur in the data.
- The choice of the number and width of bars can drastically affect the appearance of a histogram.

Histogram example

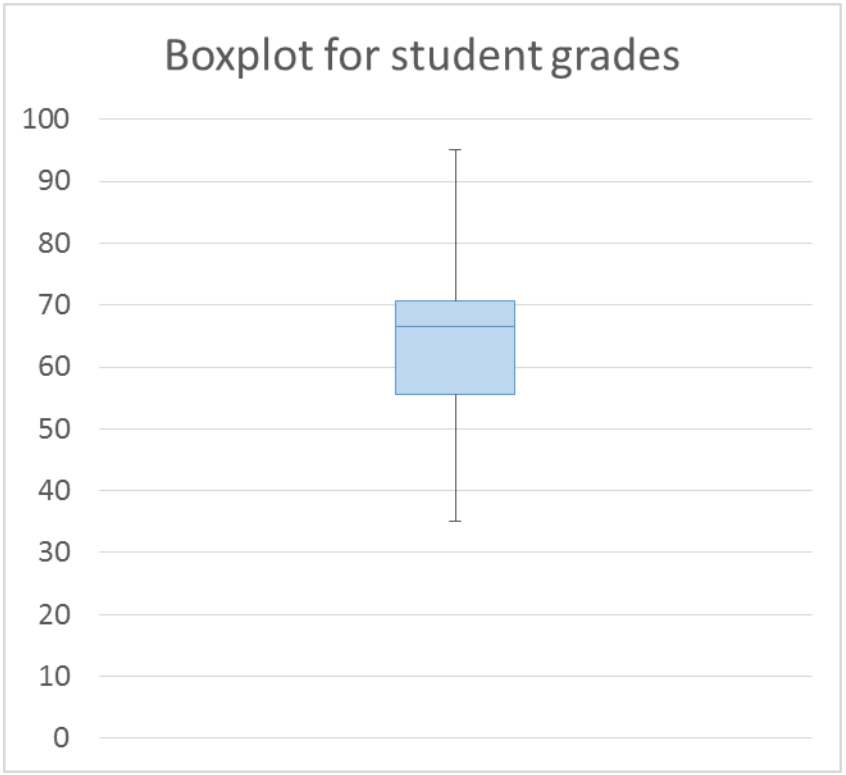


Another histogram example



Boxplots

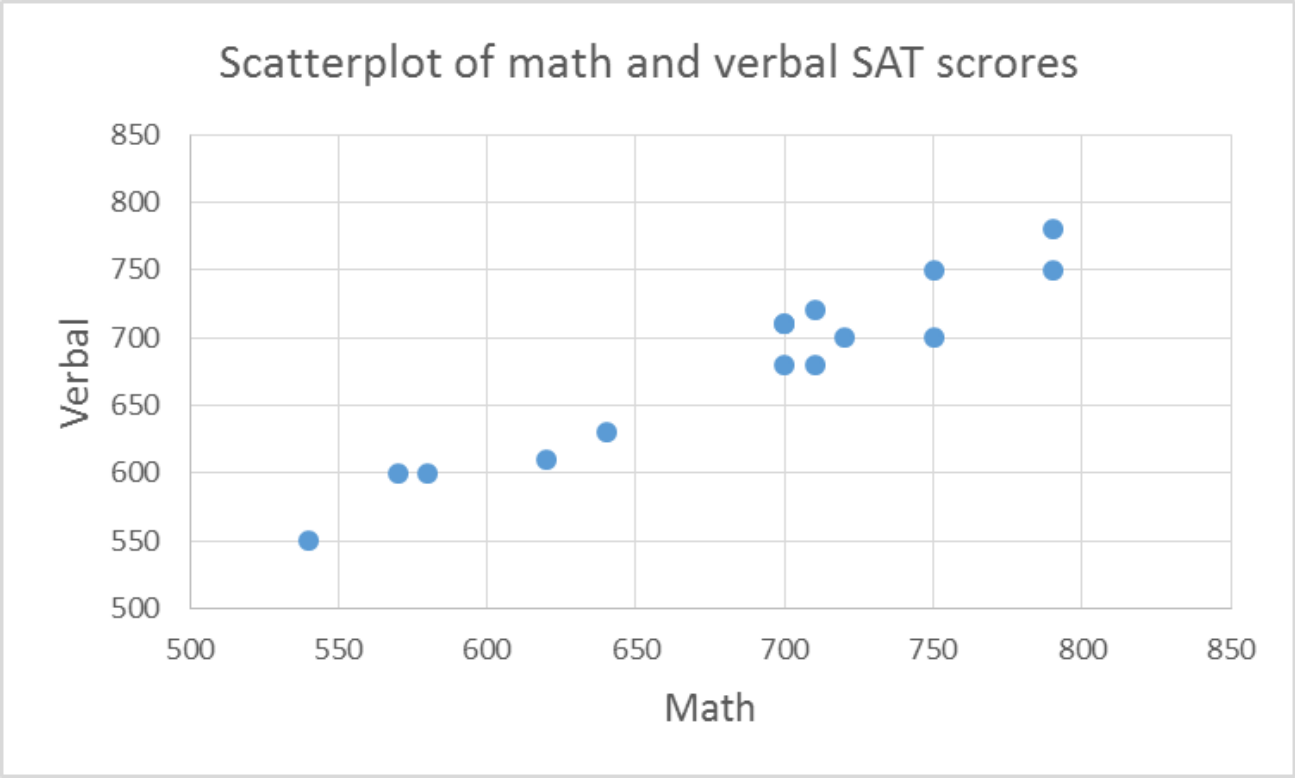
- Boxplots are also appropriate for data on an interval or ratio scale.
- They depict the median, the interquartile range, and the range of the data.



Scatterplots

- Scatterplots are widely used for visualising the relationship between two variables.
- They define each point in a dataset by two values (from the two variables of interest) and plot each point on a pair of axes.

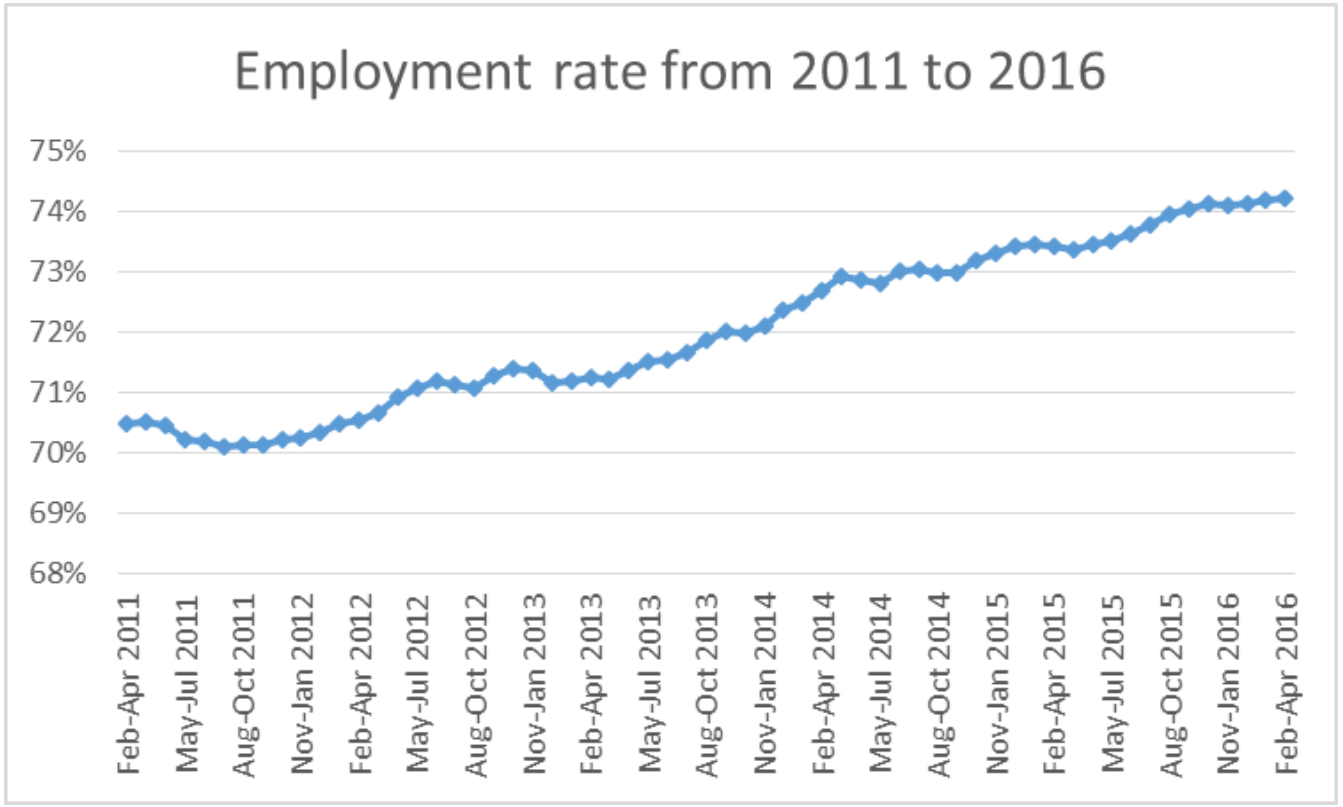
Scatterplot example



Line graphs

- Line graphs are also used to display the relationship between two variables, usually between time on the x-axis and some other variable on the y-axis.
- They show a trend over time.

Line graph example



Introduction to R

- R is a programming language and software environment for statistical computing and graphics.
- It is increasingly used in both academia and industry.
- It is user-developed and freely available.
- It can perform most of the tasks available in statistical software.

Why use R

- Advanced statistical analysis
- Flexible
- Free
- Runs on almost any standard operating system
- Open source project
- Frequent releases and active development
- Extensive documentation, tutorials, webpages

Basic commands in R

- Simple calculations

```
> 5 + 3
```

- Variable assignment

```
> age <- 28
```

- Using existing functions for calculations

```
> sqrt(225)
```

- Creating a vector

```
> weekly_sales <- c(200, 120, 130, 125, 220)
```

- Getting information out of a vector

```
> weekly_sales[3]
> weekly_sales[weekly_sales > 180]
```

Basic commands in R

- Installing packages

```
> install.packages("somepackage")
```

- Loading packages

```
> library("somepackage")
```

- Setting working directory

```
> setwd("~/myDirectory")
```

- Reading a csv file

```
> data <- read.csv("somedata.csv")
```

Getting a feel for the data

- Get the structure of the object

```
> str(data)
```

- Get the head of the object

```
> head(data)
```

- Get the names of the object

```
> names(data)
```

- Get the entire object

```
> data
```

Calculating summary statistics

- Calculate the mean of a variable

```
> mean(data$someVariable)
```

- Calculate the median of a variable

```
> median(data$someVariable)
```

- Estimate the population variance of a variable

```
> var(data$someVariable)
```

- Estimate the population standard deviation of a variable

```
> sd(data$someVariable)
```

Calculating summary statistics

- Get an overall summary of an object or a variable

```
> summary(data)
> summary(data$someVariable)
```

- Calculate descriptive statistics separately for each group

```
> by(data, data$someVariable, summary)
```

Visualising data

- Plot a variable (the type of graph generated depends on the type of the variable)

```
> plot(data$someVariable)
```

- Draw a histogram for some variable

```
> hist(data$someVariable)
```

- Draw a boxplot for some variable

```
> boxplot(data$someVariable)
```

- Draw a bar chart for some variable

```
> freq <- table(data$someVariable)
> barplot(freq)
```

Visualising data

- Draw a pie chart for some variable

```
> freq <- table(data$someVariable)
> pie(freq)
```

- Draw a scatterplot for two variables

```
> plot(data$variable1, data$variable2)
```

Conclusions

- What kind of visualisations are possible depends on the kind of data.
  - Qualitative data: bar charts and pie charts
  - Quantitative data: histograms and box plots
  - Bivariate data: scatter plots and line graphs
- R is a programming language for statistical analysis of data.
  - We covered some basics in R
  - More in the labs!