

Creating Scatter plots in R

Ralph Mansson

Introduction

Scatter plots

A scatter plot is a graph used to investigate the relationship between two variables in a data set. The x and y axes are used for the values of the two variables and a symbol on the graph represents the combination for each pair of values in the data set.

To illustrate creating a scatter plot we will use a simple data set for the population of the UK between 1992 and 2009. This data is saved in a data frame `uk.df` using the following command:

```
uk.df = data.frame(Year = 1992:2009,  
Population = c(57770, 57933, 58096, 58258,  
58418, 58577, 58743, 58925, 59131, 59363,  
59618, 59894, 60186, 60489, 60804, 61129,  
61461, 61796))
```

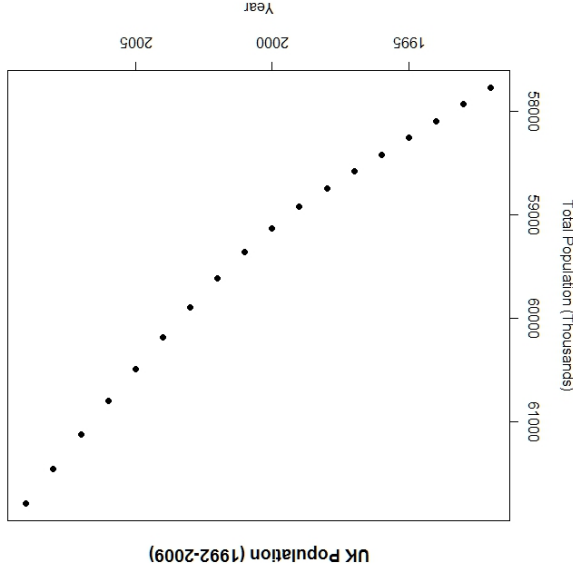
The data has been recorded in thousands to save space on the graphs.

Base Graphics

The general purpose `plot` function, which is part of the **base** graphics system, is used to create a scatter plot for the UK population data. The first two arguments to the function are the x and y variables respectively. The following code will create a scatter plot:

```
plot(uk.df$Year, uk.df$Population,  
xlab = "Year", ylab = "Total Population  
(Thousands)", main = "UK Population  
(1992-2009)", pch = 16)
```

The labels for the x and y axes are specified via the `xlab` and `ylab` arguments to the plot function and the `main` argument specifies the title for the plot.



The graph itself is plain and functional.

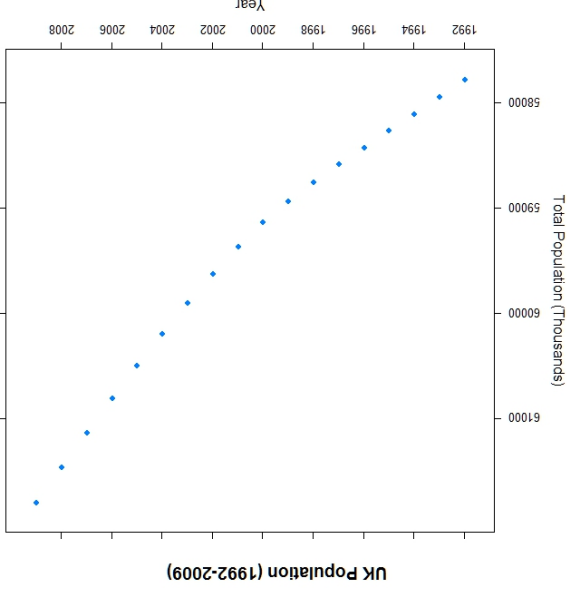
Lattice Graphics

The `lattice` graphics package provides a function `xyplot` to create scatter plots and is very similar to the `base` graphics approach. The first argument to the function is a formula describing the relationship to be plotted on the graph, with the y variable preceding the x variable. The data frame is specified with the `data` argument to simplify the expression in the formula. The code used is as follows:

```
xyplot(Population ~ Year, data = uk.df,
       xlab = "Year", ylab = "Total Population
(Thousands)", main = "UK Population
(1992-2009)", scales = list(x = list(at =
seq(1992, 2009, 2))))
```

The axis labels and the overall title are specified in the same way as the `base` graphics system. Some fine tuning of the labels on the x axis is undertaken with

the `scales` argument to indicate that we want every second year to be included on the label starting in 1992 and running until 2009. The `lattice` graph is shown here:



Overall the graph produced by the `lattice` package is similar to the `base` graphics with some improvements to the layout via the labels.

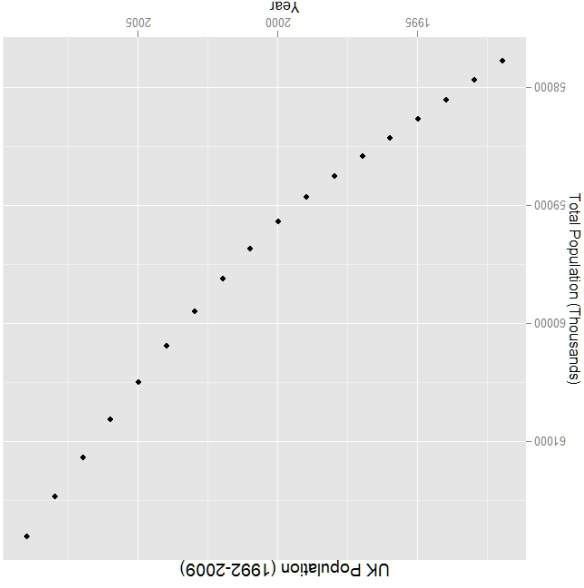
ggplot2 Graphics

The `ggplot2` function is used to create graphs with the `ggplot2` package. The first argument is the data frame with the data to be plotted and the `aes` argument specifies the aesthetics associated with the graph. In the case below the `Year` variable appears on the x axis and the `Population` variable on the y axis.

```
ggplot(uk.df, aes(Year, Population)) +
  geom_point() + xlab("Year") + ylab("Total
Population (Thousands)") + opts(title = "UK
Population (1992-2009)")
```

The `geom_point` specifies the type of graph to create (a scatter plot) and the labels for the graph are created by adding them to the graph with the `xlab`, `ylab` and

`opts` functions. The graph is shown below:



This graph is not greatly different to the scatter plot created using the `base` and `lattice` packages. The default theme in the `ggplot2` package has a gray background with white grid lines that allows easy visual recognition of graphs created using this package.

GMR-2010-003: Creating Scatter plots in R
©2010 GM-RAM Limited

This leaflet is part of a series covering Statistical Analysis using the R Statistical Software.

<http://www.gm-ram.com>
<http://www.wekaleamstudios.co.uk>