Starting Analysis & Visualization of Spatial Data with R

Aaron Brown

Sunday, May 17, 2015

Contents

1 Introduction

This introduction was taken as part of the Spatial.ly: R Spatial Tips site. The site was built to give R users an introduction into spatial analysis using R.

Attribution: Based on A Short Introduction to R by Richard Harris (www.social-statistics.org).

This document is based on the workbooks discussed here.

1.1 Packages Required

This analysis uses several packages in the analysis.

1. **ProjectTemplate** library to manage the files and structure of the project.
2. **maptools** and **rgdal** as required on this page.
3. **ggplot2** is used with the second workbook to demonstrate how to use *ggplot2* with spatial data.

This project is loaded by executing the following codechunk.

```
library(knitr)
opts_knit$set(root.dir = '..')
```

```
library(ProjectTemplate)
load.project()
```

```
## Loading project configuration
## Autoloading helper functions
## Autoloading packages
##  Loading package: maptools
## Loading required package: maptools
## Loading required package: sp
## Checking rgeos availability: FALSE
##      Note: when rgeos is not available, polygon geometry     computations in maptools depend on gpcli
##      which has a restricted licence. It is disabled by default;
##      to enable gpclib, type gpclibPermit()
##  Loading package: rgdal
## Loading required package: rgdal
## rgdal: version: 0.9-1, (SVN revision 518)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 1.11.0, released 2014/04/16
## Path to GDAL shared files: C:/Users/abrow/Documents/R/projects/personal/Spatial.ly Proj/packrat/lib/i
## GDAL does not use iconv for recoding strings.
## Loaded PROJ.4 runtime: Rel. 4.8.0, 6 March 2012, [PJ_VERSION: 480]
## Path to PROJ.4 shared files: C:/Users/abrow/Documents/R/projects/personal/Spatial.ly Proj/packrat/lib
##  Loading package: downloader
## Loading required package: downloader
##  Loading package: ggplot2
## Loading required package: ggplot2
##  Loading package: RgoogleMaps
## Loading required package: RgoogleMaps
##  Loading package: png
## Loading required package: png
##  Loading package: sp
##  Loading package: spdep
## Loading required package: spdep
## Loading required package: Matrix
## Autoloading cache
## Autoloading data
## Munging data
```

1.2 Files downloaded

There are two workbooks associated with this analysis and can be found here.

The following codechunk download the worksheet and raw data associated with the Richard Harris workbook and the unnamed second workbook.

```
data.dir <- 'data/02'
if(!file.exists(paste0(data.dir, 'Intro_R_Data.zip'))){
      url <- 'http://spatialanalysis.co.uk/wp-content/uploads/2013/04/Intro_R_Data.zip'
      download(url, dest = paste0(data.dir, '/Intro_R_Data.zip'), mode = 'wb')
}
```

```
## Warning in download.file(url, ...): downloaded length 20571 != reported
## length 20571
```

```
unzip(paste0(data.dir, '/Intro_R_Data.zip'), overwrite = TRUE, exdir = data.dir)
```

```
if(!file.exists(paste0(data.dir, '/intro_to_R1.pdf'))){
      url <- 'http://www.social-statistics.org/wp-content/uploads/2012/12/intro_to_R1.pdf'
      download(url, dest = paste0(data.dir, '/intro_to_R1.pdf'), mode = 'wb')
}

if(!file.exists(paste0(data.dir, '/R-ggplot2-data.zip'))){
      url <- 'http://spatialanalysis.co.uk/wp-content/uploads/2013/04/R-ggplot2-data.zip'
      download(url, dest = paste0(data.dir, '/R-ggplot2-data.zip'), mode = 'wb')
}
unzip(paste0(data.dir, '/R-ggplot2-data.zip'), overwrite = TRUE, exdir = data.dir)

if(!file.exists(paste0(data.dir, '/james_cheshire_ggplot_intro_blog.pdf'))){
      url <- 'http://spatialanalysis.co.uk/wp-content/uploads/2013/04/james_cheshire_ggplot_intro_blog.p
      download(url, dest = paste0(data.dir, '/james_cheshire_ggplot_intro_blog.pdf'), mode = 'wb')
}
```

2 Loading and Previewing Schools Data

```
schools.dat <- read.csv(paste0(data.dir, '/schools.csv'))
```

Reviewing the data

```
head(schools.dat); tail(schools.dat)
```

```
##   attainment   fsm   esl   SEN white blk.car blk.afr indian pakistani
## 1       27.0 0.634 0.586 0.034 0.230   0.012   0.169  0.009     0.019
## 2       29.0 0.398 0.384 0.000 0.343   0.095   0.105  0.006     0.006
## 3       26.4 0.680 0.952 0.040 0.025   0.016   0.257  0.006     0.007
## 4       27.3 0.409 0.294 0.010 0.234   0.062   0.150  0.038     0.062
## 5       28.3 0.418 0.262 0.046 0.513   0.000   0.081  0.000     0.000
## 6       30.0 0.349 0.353 0.018 0.269   0.045   0.149  0.008     0.000
##   bangladeshi chinese coe rc vol.con other.faith selective Easting
## 1       0.219   0.021   0  0       0           0         0  529733
## 2       0.078   0.005   0  0       0           0         0  529020
## 3       0.517   0.000   0  0       0           0         0  527842
## 4       0.086   0.000   0  0       0           0         0  525182
## 5       0.029   0.000   0  0       0           0         0  530130
## 6       0.143   0.020   0  0       0           0         0  528663
##   Northing       Long      Lat
## 1   186496 -0.1298076 51.56243
## 2   187023 -0.1398946 51.56733
## 3   182833 -0.1584024 51.52994
## 4   185772 -0.1956843 51.55695
## 5   186087 -0.1242348 51.55866
## 6   184769 -0.1458664 51.54715
```

```
##     attainment   fsm   esl   SEN white blk.car blk.afr indian pakistani
## 362       26.9 0.316 0.472 0.000 0.236   0.167   0.132  0.010     0.105
## 363       25.9 0.357 0.428 0.029 0.109   0.010   0.069  0.092     0.401
## 364       25.6 0.417 0.636 0.000 0.133   0.069   0.161  0.103     0.071
```

```
## 365       28.1 0.218 0.380 0.026 0.110   0.190   0.231  0.037     0.000
## 366       28.1 0.145 0.161 0.046 0.548   0.068   0.029  0.039     0.030
## 367       28.6 0.156 0.235 0.028 0.577   0.058   0.007  0.000     0.016
##     bangladeshi chinese coe rc vol.con other.faith selective Easting
## 362       0.020   0.000   0  0       0           0         0  538157
## 363       0.033   0.006   0  0       0           0         0  537827
## 364       0.041   0.000   0  0       0           0         0  538518
## 365       0.000   0.000   0  1       0           0         0  540033
## 366       0.009   0.000   0  0       0           0         0  538443
## 367       0.006   0.009   0  0       0           0         0  538384
##     Northing         Long      Lat
## 362   191575 -0.006359028 51.60607
## 363   187422 -0.012750450 51.56883
## 364   186885 -0.002998281 51.56384
## 365   190931  0.020456165 51.59982
## 366   193814 -0.001348053 51.62612
## 367   193125 -0.002471794 51.61994
```

Getting a summary of each column

```
summary(schools.dat)
```

```
##    attainment         fsm              esl              SEN
##  Min.   :24.10   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:26.80   1st Qu.:0.1365   1st Qu.:0.1530   1st Qu.:0.0080
##  Median :27.80   Median :0.2430   Median :0.3220   Median :0.0180
##  Mean   :27.86   Mean   :0.2689   Mean   :0.3493   Mean   :0.0213
##  3rd Qu.:28.60   3rd Qu.:0.3900   3rd Qu.:0.5105   3rd Qu.:0.0300
##  Max.   :33.10   Max.   :0.7990   Max.   :0.9920   Max.   :0.0980
##      white           blk.car           blk.afr           indian
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.1385   1st Qu.:0.01500   1st Qu.:0.0465   1st Qu.:0.00550
##  Median :0.3120   Median :0.04800   Median :0.1010   Median :0.01800
##  Mean   :0.3632   Mean   :0.07178   Mean   :0.1228   Mean   :0.05054
##  3rd Qu.:0.5780   3rd Qu.:0.11050   3rd Qu.:0.1780   3rd Qu.:0.05200
##  Max.   :0.9280   Max.   :0.48500   Max.   :0.6250   Max.   :0.82900
##    pakistani        bangladeshi         chinese             coe
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00100   1st Qu.:0.00300   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.01200   Median :0.01100   Median :0.00500   Median :0.00000
##  Mean   :0.03822   Mean   :0.05331   Mean   :0.00855   Mean   :0.06812
##  3rd Qu.:0.04300   3rd Qu.:0.03350   3rd Qu.:0.01000   3rd Qu.:0.00000
##  Max.   :0.40100   Max.   :0.97400   Max.   :0.11700   Max.   :1.00000
##        rc            vol.con         other.faith        selective
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.1717   Mean   :0.01362   Mean   :0.01907   Mean   :0.05177
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##     Easting          Northing           Long               Lat
##  Min.   :504816   Min.   :159523   Min.   :-0.48962   Min.   :51.32
##  1st Qu.:524172   1st Qu.:173996   1st Qu.:-0.21100   1st Qu.:51.45
```

```
##  Median :531958   Median :181950   Median :-0.10080   Median :51.52
##  Mean   :531832   Mean   :180479   Mean   :-0.10177   Mean   :51.51
##  3rd Qu.:539940   3rd Qu.:187424   3rd Qu.: 0.01364   3rd Qu.:51.57
##  Max.   :556625   Max.   :199227   Max.   : 0.25854   Max.   :51.68
```

The names of each column are

```
names(schools.dat)
```

```
##  [1] "attainment"  "fsm"         "esl"         "SEN"         "white"
##  [6] "blk.car"     "blk.afr"     "indian"      "pakistani"   "bangladeshi"
## [11] "chinese"     "coe"         "rc"          "vol.con"     "other.faith"
## [16] "selective"   "Easting"     "Northing"    "Long"        "Lat"
```

Checking the structure of *schools.dat*

```
ncol(schools.dat)
```

```
## [1] 20
```

```
nrow(schools.dat)
```

```
## [1] 367
```

```
complete.cases(schools.dat)
```

```
##   [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [85] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
##  [99] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [113] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [127] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [141] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [155] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [169] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [183] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [197] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [211] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [225] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [239] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [253] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [267] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [281] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [295] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [309] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [323] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [337] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [351] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [365] TRUE TRUE TRUE
```

2.1 Initial Data Visualizations

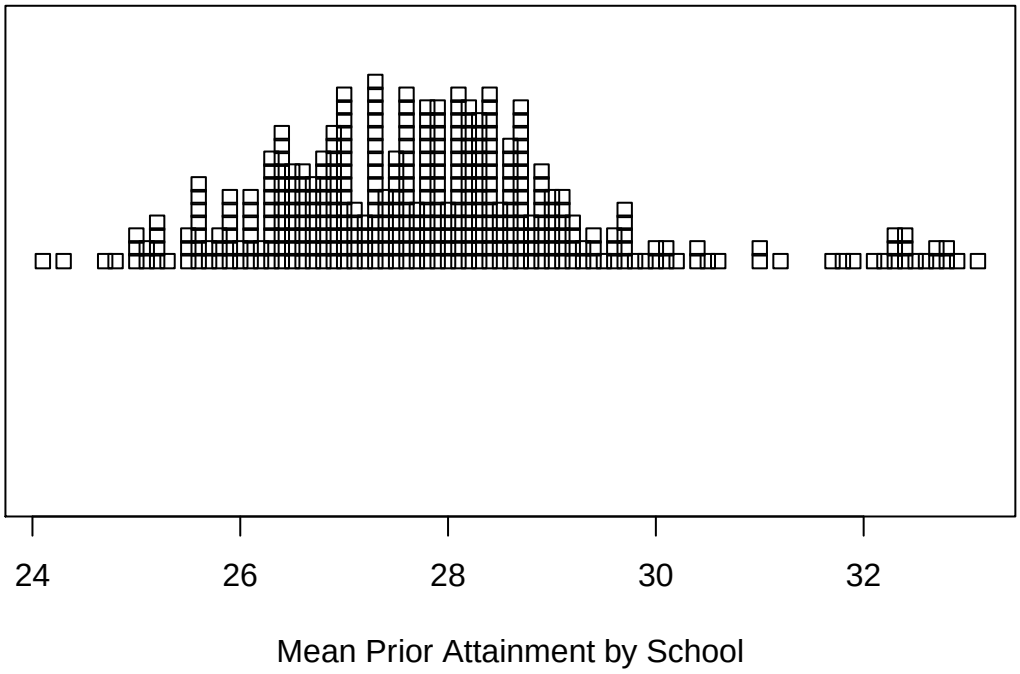
The file schools.csv contains information about the location and some attributes of schools in Greater London (in 2008). The locations are given as a grid reference (Easting, Northing). The information is not real but is realistic. It should not, however, be used to make inferences about real schools in London.

Of particular interest is the average attainment on leaving primary school of pupils entering their first year of secondary school. Do some schools in London attract higher attaining pupils more than others? The variable attainment contains this information. A stripchart and then a histogram will show that (not surprisingly) there is variation in the average prior attainment by school.

Here the histogram is scaled so the total area sums to one.

To this we can add a rug plot. . . also a density curve, a Normal curve for comparison and a legend.

```
attach(schools.dat)
stripchart(attainment, method="stack", xlab="Mean Prior Attainment by School")
```

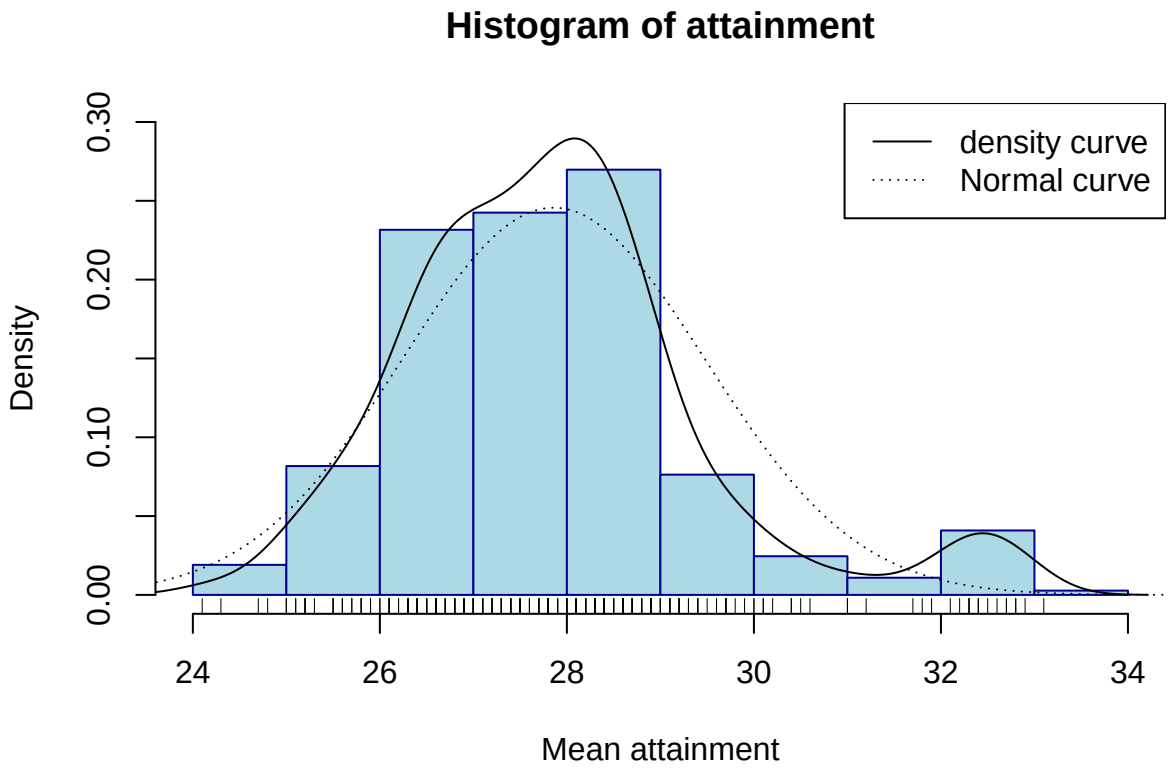



```
hist(attainment, col="light blue", border="dark blue", freq=F, ylim=c(0,0.30),xlab="Mean attainment")

rug(attainment)

lines(density(sort(attainment)))
xx <- seq(from=23, to=35, by=0.1)
yy <- dnorm(xx, mean(attainment), sd(attainment))
lines(xx, yy, lty="dotted")
rm(xx, yy)
```

```
legend("topright", legend=c("density curve","Normal curve"),
       lty=c("solid","dotted"))
```



If would be interesting to know if attainment varies by school type. A simple way to consider this is to produce a box plot. The data contain a series of dummy variables for each of a series of school types (Voluntary Aided Church of England: coe = 1; Voluntary Aided Roman Catholic: rc = 1; Voluntary controlled faith school: vol.con = 1; another type of faith school: other.faith = 1; a selective school (with an entrance exam): selective = 1). We will combine these into a single, categorical variable then produce the box plot showing the distribution of average attainment by school type.

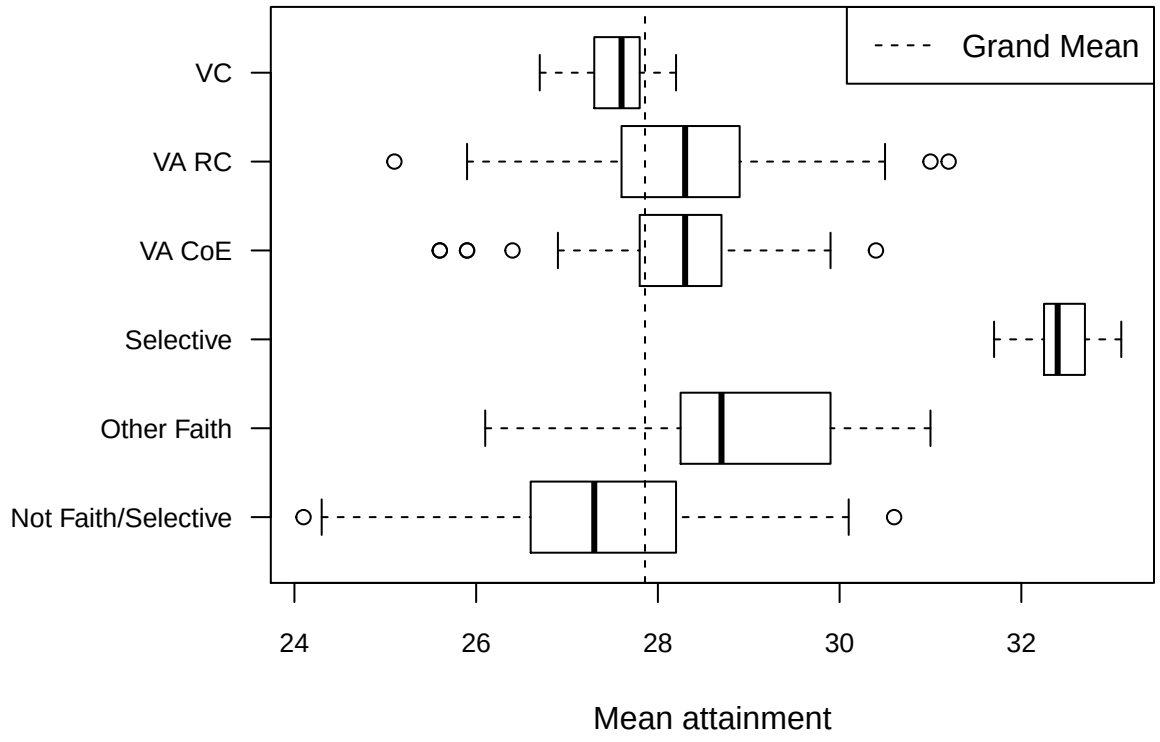
First the categorical variable:

```
school.type <- rep("Not Faith/Selective", times=nrow(schools.dat))
school.type[coe==1] <- "VA CoE"
school.type[rc==1] <- "VA RC"
school.type[vol.con==1] <- "VC"
school.type[other.faith==1] <- "Other Faith"
school.type[selective==1] <- "Selective"
school.type <- factor(school.type)
levels(school.type)
```

```
## [1] "Not Faith/Selective" "Other Faith"         "Selective"
## [4] "VA CoE"              "VA RC"               "VC"
```

Now the box plots:

```r
par(mai=c(1,1.4,0.5,0.5))      # Changes the graphic margins
boxplot(attainment ~ school.type, horizontal=T, xlab="Mean attainment", las=1,
        cex.axis=0.8)     # Includes options to draw the boxes and labels horizontally
abline(v=mean(attainment), lty="dashed")    # Adds the mean value to the plot
legend("topright", legend="Grand Mean", lty="dashed")
```



Not surprisingly, the selective schools recruit the pupils with highest average prior attainment.

2.2 Simple Statistics

It appears that there are differences in the levels of prior attainment of pupils in different school types. We can test whether the variation is significant using an analysis of variance.

```r
summary(aov(attainment ~ school.type))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## school.type   5  479.8   95.95   71.42 <2e-16 ***
## Residuals   361  485.0    1.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is, at a greater than 99.9% confidence ($F = 71.42$, $p < 0.001$). We might also be interested in comparing those schools with the highest and lowest proportions of Free School Meal eligible pupils to see if they are recruiting pupils with equal or differing mean prior attainment.

```
# Finds the attainment scores for schools with the highest proportions of FSM pupils
attainment.high.fsm.schools <- attainment[fsm > quantile(fsm, probs=0.75)]
# Finds the attainment scores for schools with the lowest proportions of FSM pupils
attainment.low.fsm.schools <- attainment[fsm < quantile(fsm, probs=0.25)]
t.test(attainment.high.fsm.schools, attainment.low.fsm.schools)
```

```
##
##  Welch Two Sample t-test
##
## data:  attainment.high.fsm.schools and attainment.low.fsm.schools
## t = -15.0431, df = 154.164, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.437206 -2.639240
## sample estimates:
## mean of x mean of y
##  26.58352  29.62174
```

It comes as little surprise to learn that those schools with the greatest proportions of FSM eligible pupils are also those recruiting lower attaining pupils on average (mean attainment 26.6 Vs 29.6, t = -15.0, $p < 0.001$). Exploring this further, the Pearson correlation between the mean prior attainment of pupils entering each school and the proportion of them that are FSM eligible is -0.689, and significant ($p < 0.001$):

```
round(cor(fsm, attainment),3)
```

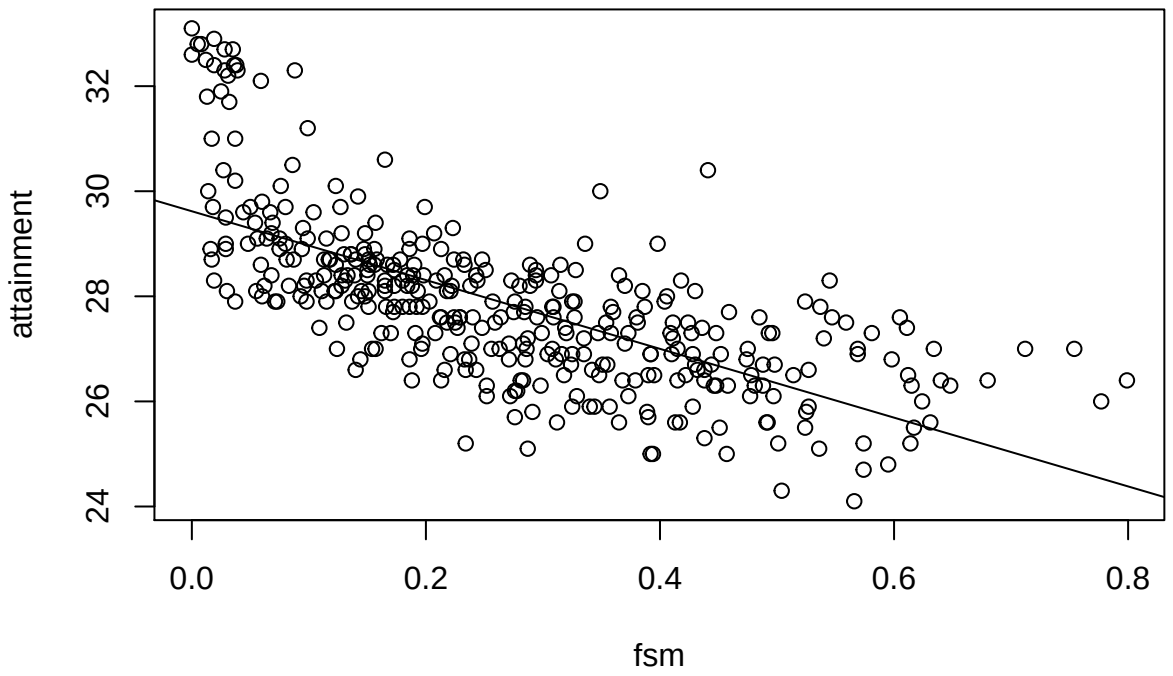
```
## [1] -0.689
```

```
cor.test(fsm, attainment)
```

```
##
##  Pearson's product-moment correlation
##
## data:  fsm and attainment
## t = -18.1731, df = 365, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7394165 -0.6313939
## sample estimates:
##       cor
## -0.6892159
```

Of course, the use of the Pearson correlation assumes that the relationship is linear, so let's check:

```
plot(attainment ~ fsm)
# Adds a line of best fit (a regression line)
abline(lm(attainment ~ fsm))
```



There is some suggestion the relationship might be curvilinear. However, we will ignore that here. Finally, some regression models. The first seeks to explain the mean prior attainment scores for the

schools in London by the proportion of their intake who are free school meal eligible. (The result is the regression line shown on the scatterplot above). The second adds a variable giving the proportion of the intake of a white ethnic group. The third adds a dummy variable indicating whether the school is selective or not.

```
model1 <- lm(attainment ~ fsm, data=schools.dat)
summary(model1)
```

```
##
## Call:
## lm(formula = attainment ~ fsm, data = schools.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8871 -0.7413 -0.1186  0.5487  3.6681
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.6190     0.1148  258.12   <2e-16 ***
## fsm          -6.5469     0.3603  -18.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.178 on 365 degrees of freedom
## Multiple R-squared:  0.475,  Adjusted R-squared:  0.4736
## F-statistic: 330.3 on 1 and 365 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(attainment ~ fsm + white, data=schools.dat)
summary(model2)
```

```
##
## Call:
## lm(formula = attainment ~ fsm + white, data = schools.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9442 -0.7295 -0.1335  0.5111  3.7837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.1250     0.1979  152.21  < 2e-16 ***
## fsm          -7.2502     0.4214  -17.20  < 2e-16 ***
## white        -0.8722     0.2796   -3.12  0.00196 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.164 on 364 degrees of freedom
## Multiple R-squared:  0.4887, Adjusted R-squared:  0.4859
## F-statistic: 173.9 on 2 and 364 DF,  p-value: < 2.2e-16
```

```
model3 <- update(model2, . ~ . + selective)
summary(model3)
```

```
##
## Call:
## lm(formula = attainment ~ fsm + white + selective, data = schools.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6262 -0.5620  0.0537  0.5607  3.6215
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  29.1706     0.1689 172.712   <2e-16 ***
## fsm          -5.2381     0.3591 -14.586   <2e-16 ***
## white        -0.2299     0.2249  -1.022    0.307
## selective     3.4768     0.2338  14.872   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9189 on 363 degrees of freedom
## Multiple R-squared:  0.6823, Adjusted R-squared:  0.6796
## F-statistic: 259.8 on 3 and 363 DF,  p-value: < 2.2e-16
```

```
lm(formula = attainment ~ fsm + white + selective, data = schools.dat)
```

```
##
## Call:
## lm(formula = attainment ~ fsm + white + selective, data = schools.dat)
##
## Coefficients:
## (Intercept)          fsm        white    selective
##     29.1706      -5.2381      -0.2299       3.4768
```

```
model4 <- update(model3, . ~ . - white)
anova(model4, model3)
```

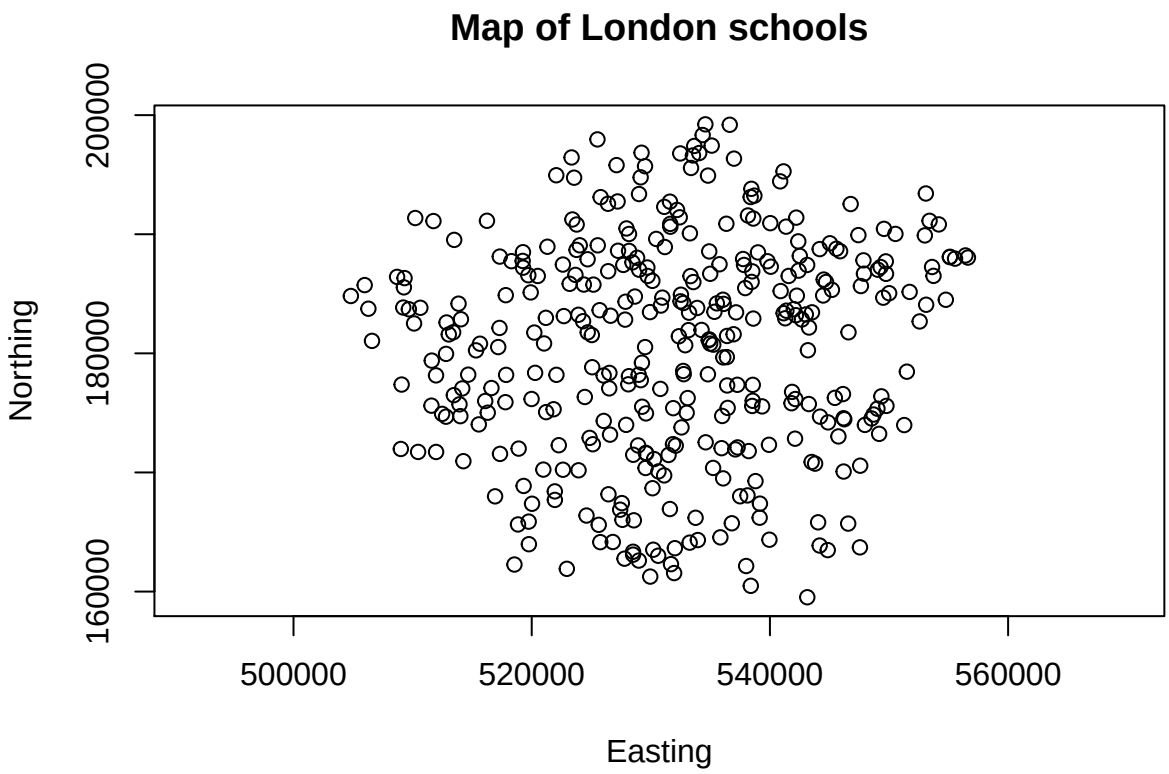
```
## Analysis of Variance Table
##
## Model 1: attainment ~ fsm + selective
## Model 2: attainment ~ fsm + white + selective
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    364 307.42
## 2    363 306.54  1   0.88222 1.0447 0.3074
```

The residual error, measured by the residual sum of squares (RSS), is not very different for the two models, and that difference, 0.882, is not significant ($F = 1.045$, $p = 0.307$).

3 Mapping Spatial Data

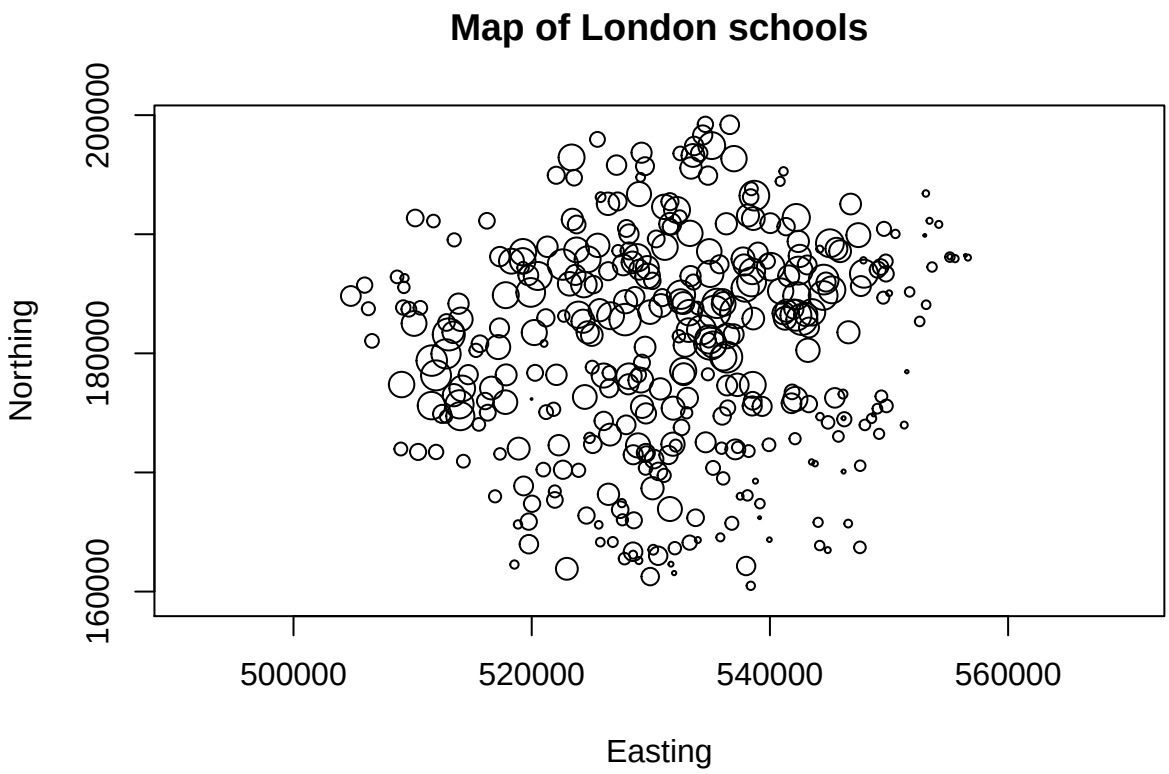
The schools data contain geographical coordinates and are therefore geographical data. Consequently they can be mapped. The simplest way for point data is to use a 2-dimensional plot, making sure the aspect ratio is fixed correctly.

```
plot(Easting, Northing, asp=1, main="Map of London schools")
```



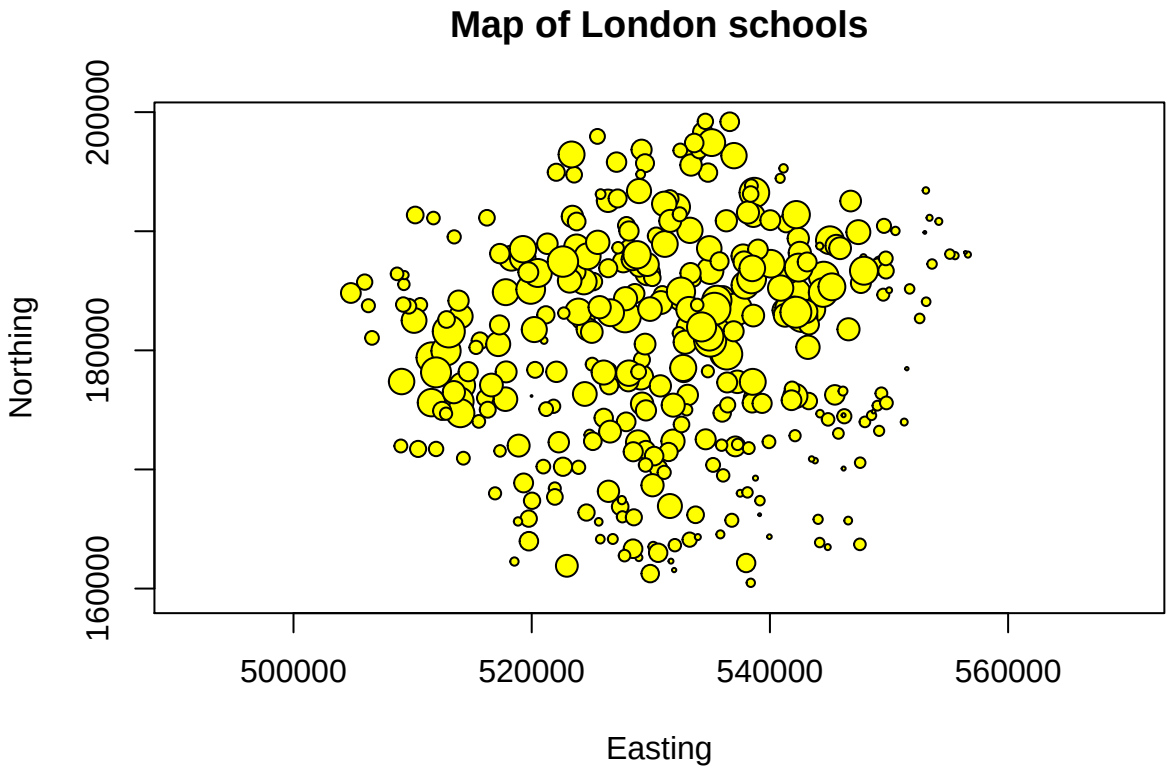
Amongst the attribute data for the schools, the variable esl gives the proportion of pupils who speak English as an additional language. It would be interesting for the size of the symbol on the map to be proportional to it.

```
plot(Easting, Northing, asp=1, main="Map of London schools", cex=sqrt(esl*5))
```



It might also be nice to add a little colour to the map. We might, for example, change the default plotting 'character' to a filled circle with a yellow background.

```
plot(Easting, Northing, asp=1, main="Map of London schools", cex=sqrt(esl*5), pch=21, bg="yellow")
```



A more interesting option would be to have the circles filled with a colour gradient that is related to a second variable in the data – the proportion of pupils eligible for free school meals for example.

To achieve this, we can begin by creating a simple colour palette:

```
palette <- c("yellow","orange","red","purple")
```

We now cut the free school meals eligibility variable into quartiles (four classes, each containing approximately the same number of observations).

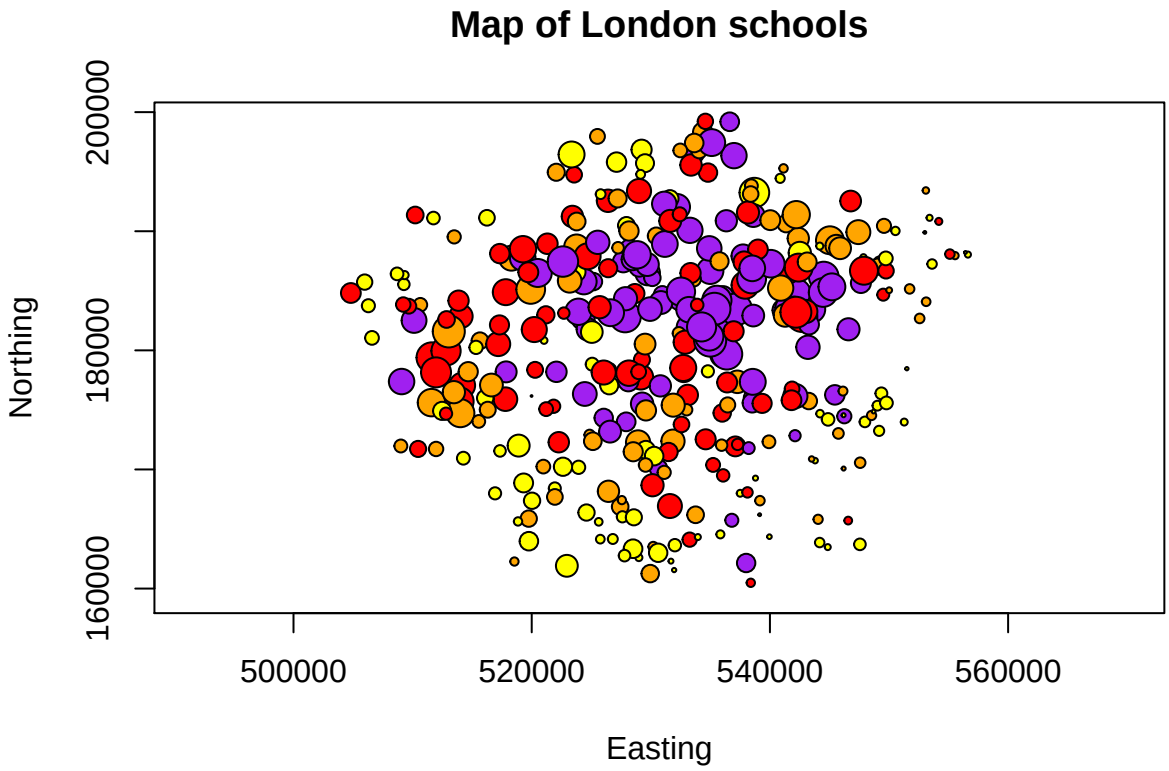
```
map.class <-  cut(fsm, quantile(fsm), labels=FALSE, include.lowest=TRUE)
```

What has happened is that the fsm variable has been split into four groups with the value 1 given to the first quarter of the data (schools with the lowest proportions of eligible pupils), the value 2 given to the next quarter, then 3, and finally the value 4 for schools with the highest proportions of FSM eligible pupils.

There are, then, now four map classes and the same number of colours in the palette. Schools in map class 1 (and with the lowest proportion of fsm-eligible pupils) will be coloured yellow, the next class will be orange, and so forth.

Bringing it all together,

```
plot(Easting, Northing, asp=1, main="Map of London schools",
     cex=sqrt(esl*5), pch=21, bg=palette[map.class])
```



It would be good to add a legend, and perhaps a scale bar and North arrow. Nevertheless, as a first map in R this isn't too bad!

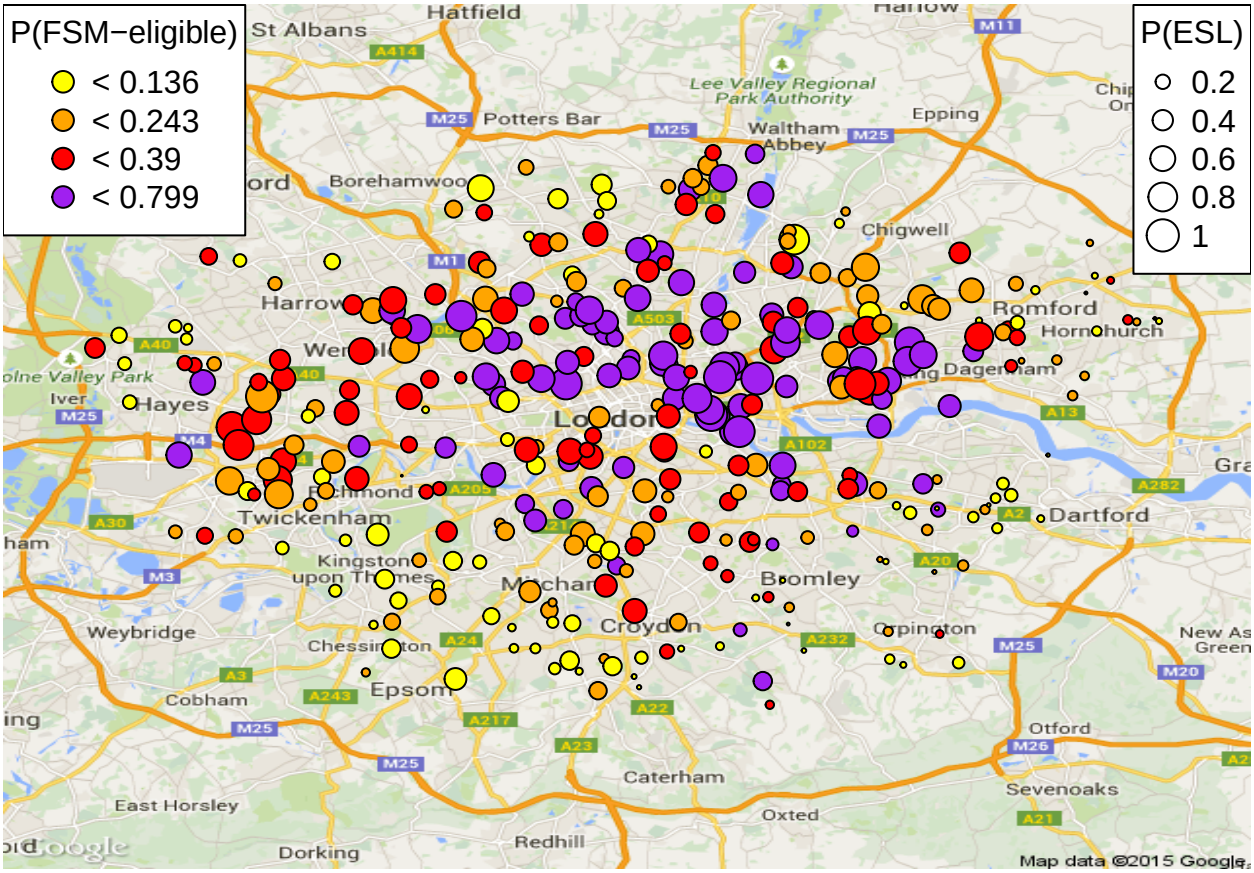
Why don't we be a bit more ambitious and overlay the map on a Google Maps tile, adding a legend as we do so? This requires us to load an additional library for R and to have an active Internet connection.

Assuming that the data frame, schools.dat, remains in the workspace and attached (it will be if you have followed the instructions above), and that the colour palette created above has not been deleted, then the map shown below is created with the following code:

```
MyMap <- MapBackground(lat=Lat, lon=Long)
```

```
## [1] "http://maps.google.com/maps/api/staticmap?center=51.4962565,-0.1155408&zoom=10&size=640x640&map
## center, zoom:  51.49626 -0.1155408 10
```

```
PlotOnStaticMap(MyMap, Lat, Long, cex=sqrt(esl*5), pch=21, bg=palette[map.class])
legend("topleft", legend=paste("<",tapply(fsm, map.class, max)),
       pch=21, pt.bg=palette, pt.cex=1.5, bg="white", title="P(FSM-eligible)")
legVals <- seq(from=0.2,to=1,by=0.2)
legend("topright", legend=round(legVals,3), pch=21, pt.bg="white",
       pt.cex=sqrt(legVals*5), bg="white", title="P(ESL)")
```



Remember that the data are simulated. The points shown on the map are not the true locations of schools in London.

3.1 Some simple geographical analysis

Remember the regression models from earlier? It would be interesting to test the assumption that the residuals exhibit independence by looking for spatial dependencies. To do this we will consider to what degree the residual value for any one school correlates with the mean residual value for its six nearest other schools (the choice of six is completely arbitrary).

First, we will take a copy of the schools data and convert that into an explicitly spatial object in R:

```
detach(schools.dat)
schools.xy <- schools.dat
attach(schools.xy)
coordinates(schools.xy) <- c("Easting", "Northing")
# Converts into a spatial object
class(schools.xy)
```

```
## [1] "SpatialPointsDataFrame"
## attr(,"package")
## [1] "sp"
```

```
detach(schools.xy)
# proj4string(schools.xy) <- CRS("+proj=tmerc datum=OSGB36")
```

```
proj4string(schools.xy) <- CRS("+proj=tmerc +lat_0=49 +lon_0=-2 +k=0.9996012717 +x_0=400000
                               +y_0=-100000 +ellps=airy
                               +towgs84=446.448,-125.157,542.060,0.1502,0.2470,0.8421,-20.4894
                               +units=m +no_defs")

# Sets the Coordinate Referencing System
```

Second, we find the six nearest neighbours for each school.

```
nearest.six <- knearneigh(schools.xy, k=6, RANN=F)
# RANN = F to override the use of the RANN package that may not be installed
```

We can learn from this that the six nearest schools to the first school in the data (row 1) are schools 5, 38, 2, 40, 223 and 6:

```
nearest.six$nn[1,]
```

```
## [1]   5  38   2  40 223   6
```

The neighbours object, nearest.six, is an object of class knn:

```
class(nearest.six)
```

```
## [1] "knn"
```

It is next converted into the more generic class of neighbours.

```
neighbours <- knn2nb(nearest.six)
class(neighbours)
```

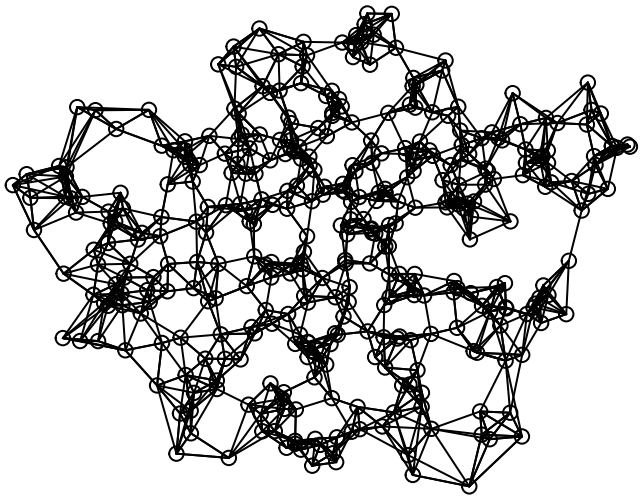
```
## [1] "nb"
```

```
summary(neighbours)
```

```
## Neighbour list object:
## Number of regions: 367
## Number of nonzero links: 2202
## Percentage nonzero weights: 1.634877
## Average number of links: 6
## Non-symmetric neighbours list
## Link number distribution:
##
##   6
## 367
## 367 least connected regions:
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 3
## 367 most connected regions:
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 3
```

The connections between each point and its neighbours can then be plotted. It may take a few minutes.

```
plot(neighbours, coordinates(schools.xy))
```



Having identified the six nearest neighbours to each school we could give each equal weight in a spatial weights matrix or, alternatively, decrease the weight with distance away (so the first nearest neighbour gets most weight and the sixth nearest the least). Creating a matrix with equal weight given to all neighbours is straightforward.

```
spatial.weights <- nb2listw(neighbours)
```

(The other possibility will not be considered further here but is achieved by creating then supplying a list of general weights to the function)

We now have all the information required to test whether there are spatial dependencies in the residuals. The answer is yes (Moran's I = 0.218, p < 0.001, indicating positive spatial autocorrelation).

```
lm.morantest(model4, spatial.weights)
```

```
##
##  Global Moran's I for regression residuals
##
## data:
## model: lm(formula = attainment ~ fsm + selective, data =
## schools.dat)
## weights: spatial.weights
##
```

```
## Moran I statistic standard deviate = 7.9152, p-value = 1.235e-15
## alternative hypothesis: greater
## sample estimates:
## Observed Moran's I        Expectation           Variance
##       0.2181914682      -0.0038585704       0.0007870118
```