Group 4 - Lab 2 - Visualization

Kevin Neville, Oscar Pettersson, Gustav Sternelöv, Vuong Tran

September 12, 2016

Assignment 1

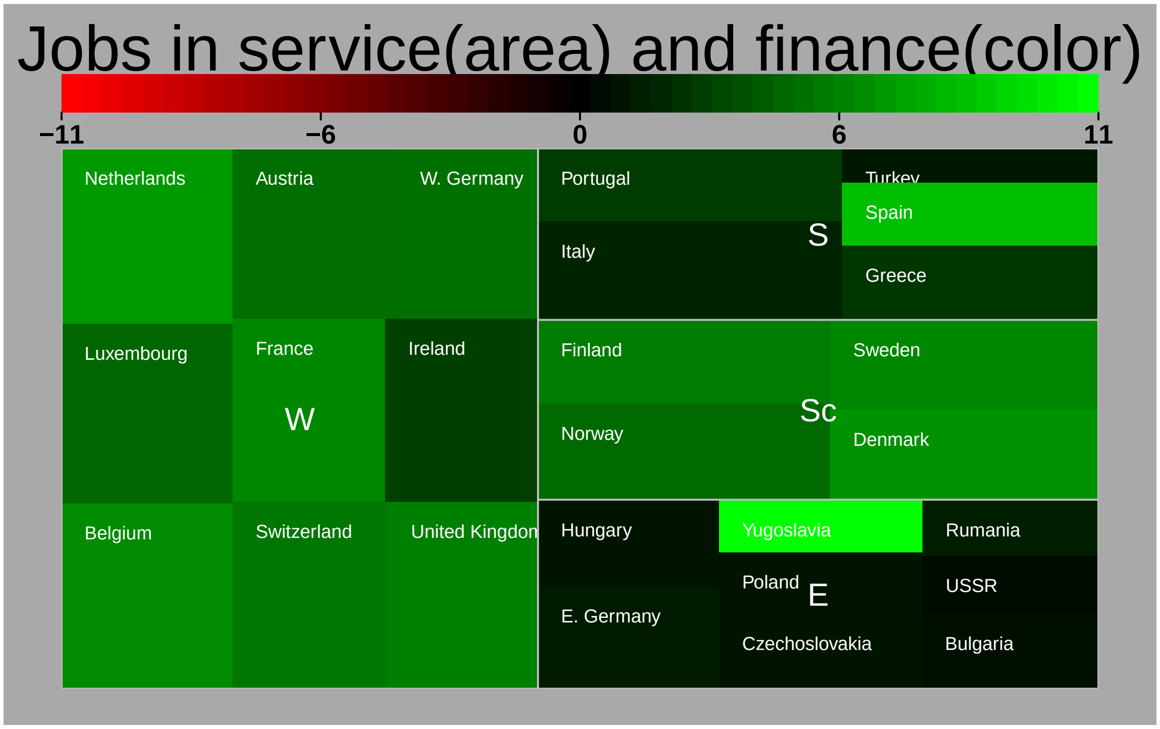
1)

The column *group* is added to the data set. The first six observations in the data set are shown below.

```
##       Country  Agr Min  Man  PS  Con   SI Fin  SPS  TC group
## 1     Belgium  3.3 0.9 27.6 0.9  8.2 19.1 6.2 26.6 7.2     W
## 2     Denmark  9.2 0.1 21.8 0.6  8.3 14.6 6.5 32.2 7.1    Sc
## 3      France 10.8 0.8 27.5 0.9  8.9 16.8 6.0 22.6 5.7     W
## 4 W. Germany  6.7 1.3 35.8 0.9  7.3 14.4 5.0 22.3 6.1     W
## 5     Ireland 23.2 1.0 20.7 1.3  7.5 16.8 2.8 20.8 6.1     W
## 6       Italy 15.9 0.6 27.6 0.5 10.0 18.1 1.6 20.1 5.7     S
```

2)

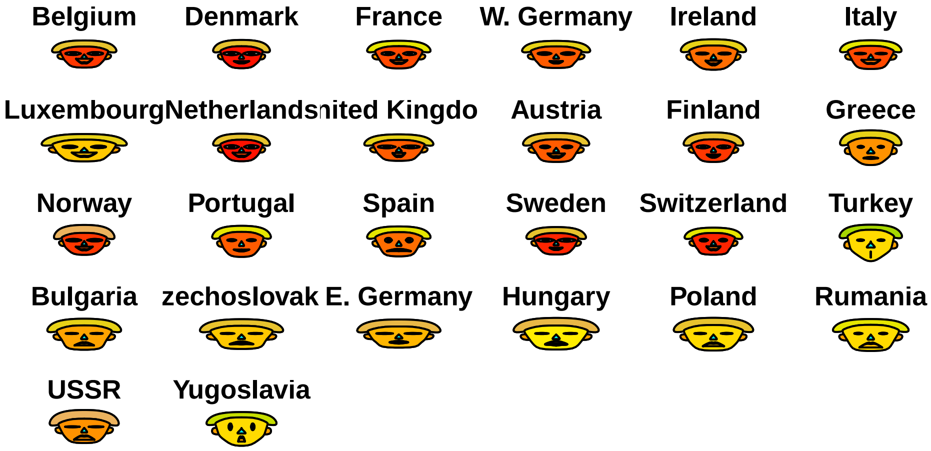
A tree map of the data is plotted where the rectangles are sorted after group, size is given by percentage employed in service areas and color shows percentage employed in finance areas.



By looking at the tree map there seem to be three different outliers, Yugoslavia, Spain and Turkey. Yugoslavia has a unusally big percentage employed in the finance sector compared to the other eastern countries. Turkey has a low percentage employed in the service sector for being an country in the southern region. The countries in the western and Scandinavian are more homogeneous groups than the other two.

3)

Chernoff faces are produced and all the quantitative variables are used. Since the use of Chernoff faces demands 15 variables and the data set just contains 9 variables, 6 extra variables are added to the data set. For all observations is the value for these extra variables set to 1. Hence, the new variables will have no effect on the results since the parts of the face they control will be the same for all observations.



```
## effect of variables:
##  modified item       Var
##  "height of face   " "Agr"
##  "width of face    " "Min"
##  "structure of face" "Man"
##  "height of mouth  " "PS"
##  "width of mouth   " "Con"
##  "smiling          " "SI"
##  "height of eyes   " "Fin"
##  "width of eyes    " "SPS"
##  "height of hair   " "TC"
##  "width of hair   "  "V12"
##  "style of hair   "  "V13"
##  "height of nose  "  "V14"
##  "width of nose   "  "V15"
##  "width of ear    "  "V16"
##  "height of ear   "  "V17"
```

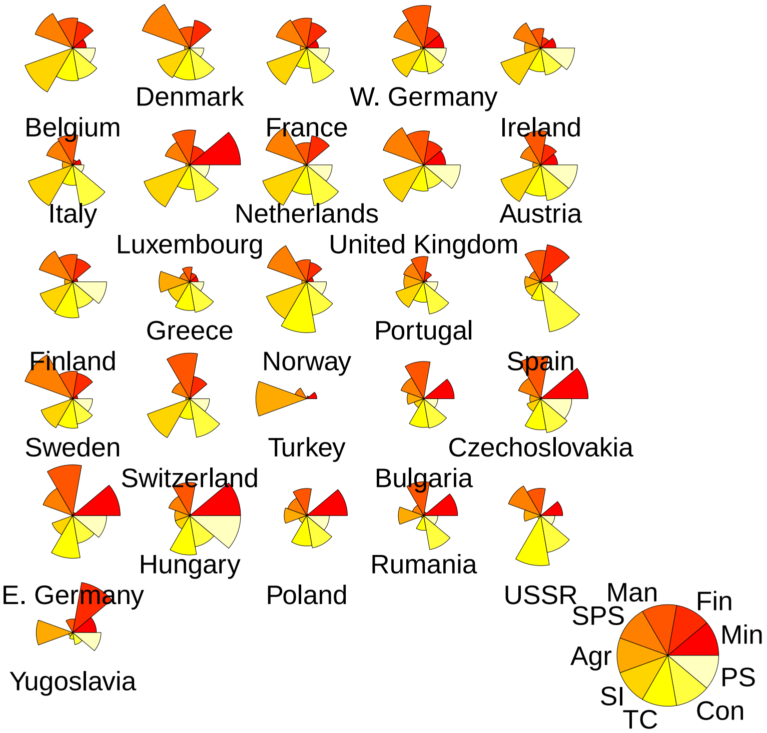
We think that two clusters can be found when analysing the chernoff faces. The first cluster consists of the western and Scandinavian countries and the second of the eastern countries. The majority of the western and Scandinavian countries have red faces, big eyes and in general quite small faces. Chernoff faces for the eastern countries are yellow, has small eyes and bigger faces than the others. Again, Yugoslavia and Turkey appears to be outliers as they have faces with unusual combinations of shapes and colours.

4)

The variables are reordered by using single link hierarchical clustering and the columns in the data set are now ordered as follows.

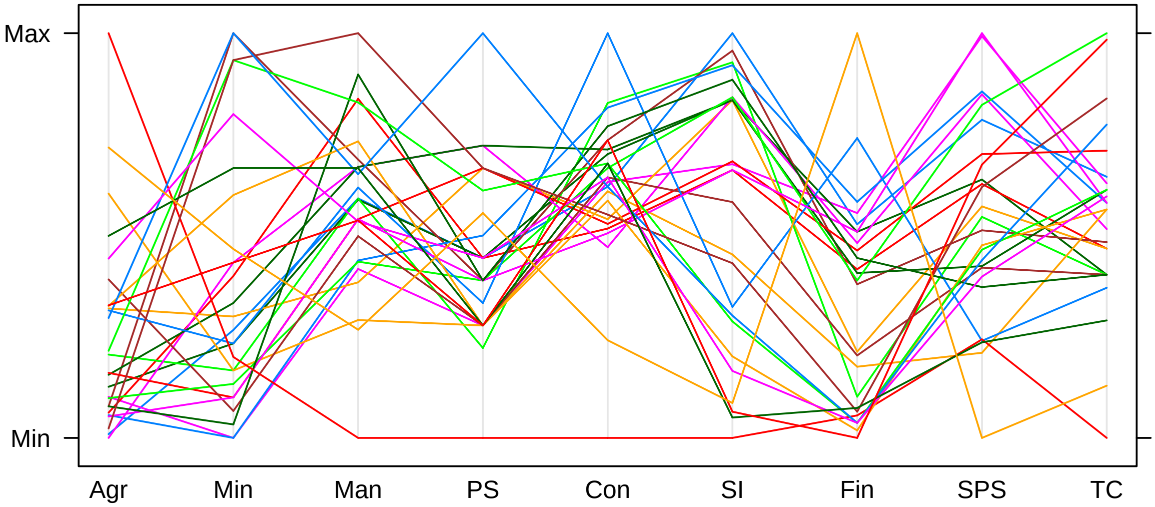
```
##   Min Fin  Man  SPS Agr   SI  TC Con  PS
## 1 0.9 6.2 27.6 26.6 3.3 19.1 7.2 8.2 0.9
```

Then, segment charts are produced by using the reordered data and all the quantitative variables in it.



The same pattern is given by the segment charts as by the chernoff faces. However, we think that the segment charts are easier to interpret. All variables are weighted equally, which perhaps not is the case with chernoff faces since the colour of the face or the size of the face may have stronger influence on the analysis than the ears.

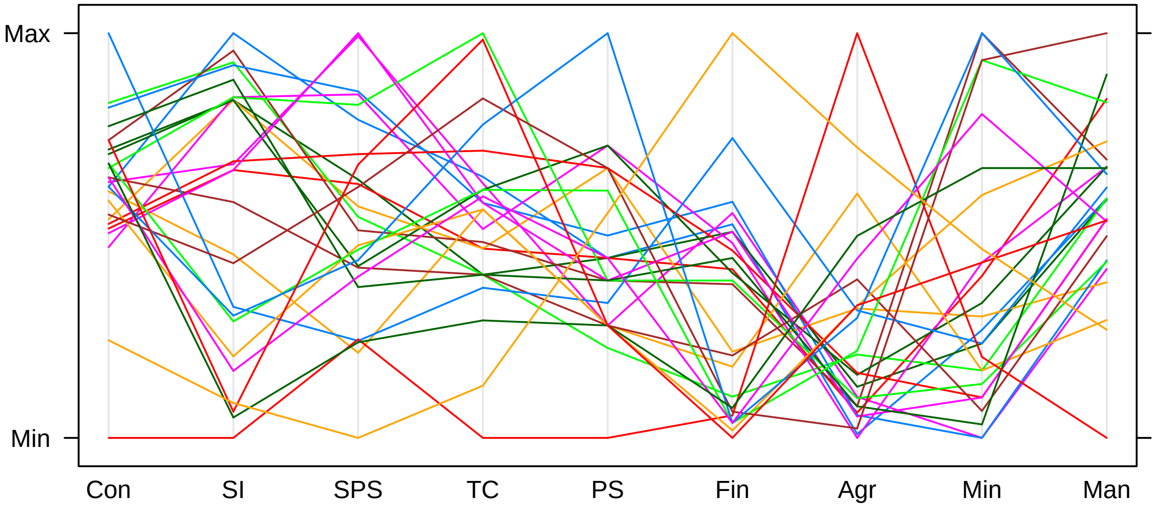
5)



We find it hard to interpret this plot. The potential clusters are hard to distinguish. One outlier can though easily be seen, the red line in the lower part of the graph which has very low values for the variables *Man*, *PS*, *Con* and *SI*. No clear correlation between any pair of the variables can be seen.

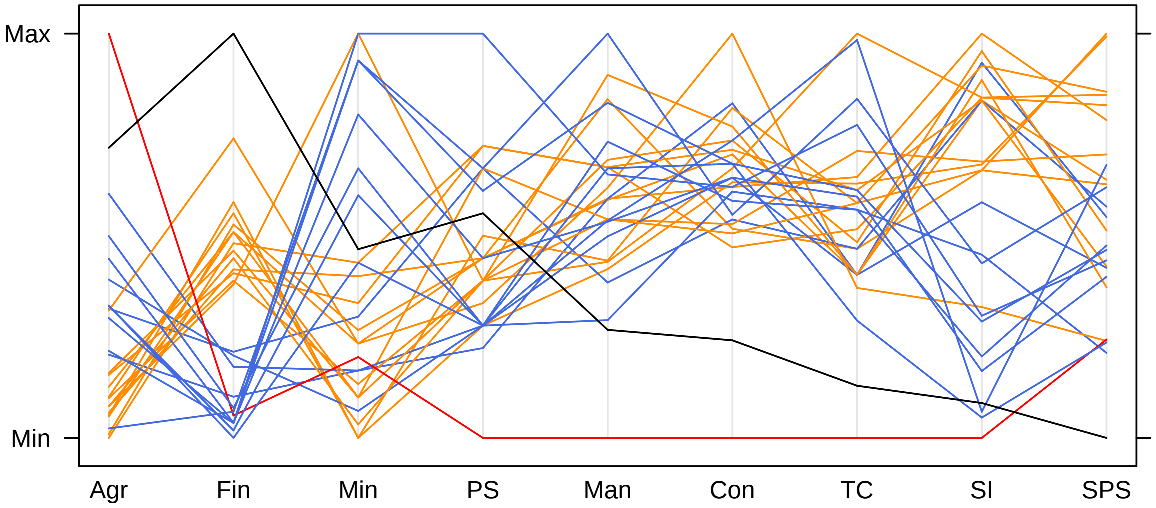
6)

The instructions given in the lab are performed and a parallel plot with the permuted columns is created.



Compared to the graph in *5)* it has become a little bit easier to find at least one cluster. For the observations with high values of *Fin* the values of *Agr* is low, so these observations are probably in the same cluster.

7)



The two clusters found are mainly given by the variable *Fin*, but also by the variables *SI* and *Arg*. Two outliers are found where the one coloured in red has very low values for the variables *PS*, *Man*, *Con*, *TC* and *SI*. The second outlier has a very high value for *Fin* and very low for *SPS*, *TC* and *Con*.

The first cluster, coloured in blue, consists of the countries which are included in the table below. In this group are all the eastern countries but Yugoslavia and all southern countries but Spain and Turkey. There is also one of the counries in the western region, Ireland.

```
##           Country group
## 5          Ireland     W
## 6            Italy     S
## 12          Greece     S
## 14        Portugal     S
## 19        Bulgaria     E
## 20 Czechoslovakia     E
## 21      E. Germany     E
## 22         Hungary     E
## 23          Poland     E
## 24         Rumania     E
## 25            USSR     E
```

The second cluster, coloured in orange, are presented in the next table. In this cluster are all of the Scandinavian countries and all the western except Ireland. Spain is the only of the southern countries in this cluster.

```
##           Country group
## 1         Belgium     W
## 2         Denmark    Sc
## 3          France     W
## 4      W. Germany     W
## 7      Luxembourg     W
## 8     Netherlands     W
## 9  United Kingdom     W
## 10        Austria     W
## 11        Finland    Sc
## 13         Norway    Sc
## 15          Spain     S
## 16         Sweden    Sc
## 17    Switzerland     W
```

The table with the two outliers is shown below. This table contains the last southern country, Turkey, and the only eastern country that not was in the first cluster, Yugoslavia.

```
##        Country group
## 18     Turkey     S
## 26 Yugoslavia     E
```

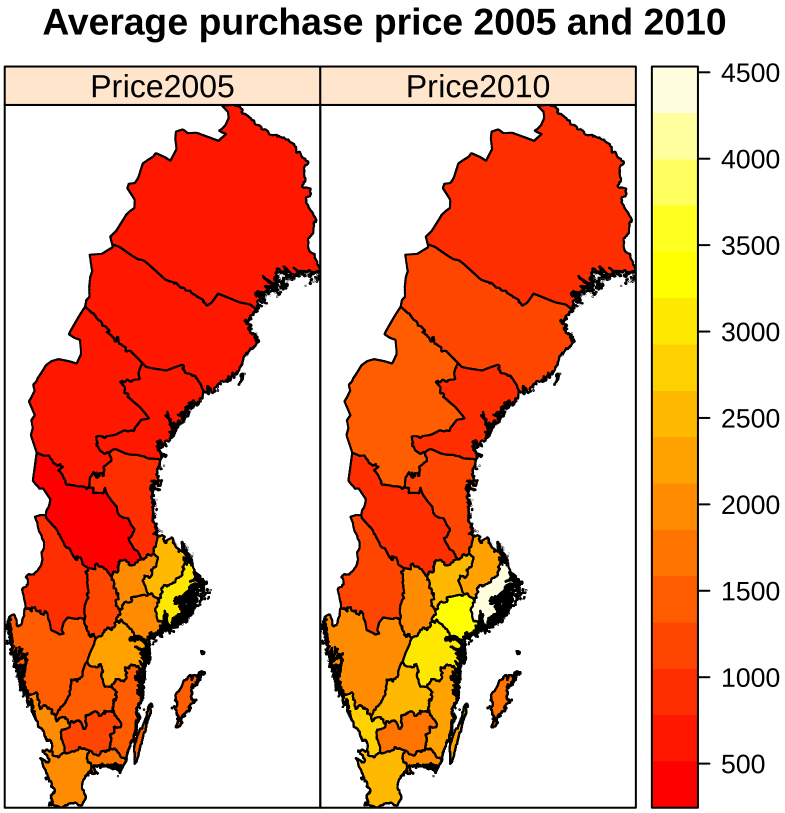
The clusters are clearly related to the group variable. In the first clusters are all the eastern countries except Yugoslavia and in the second cluster are all the Scandinavian and western countries, except Ireland. The only group which not seem to be well related to the clusters is the south. Three of the south countries are in the first cluster, one in the second and one is an outlier.
An interpretation on the analysis of the employment data is that countries in the same region tends to be similar to each other.

8)

The order of the variables made the parallel plots easier to analyse. Hence, the parallel plot in *5)* was the hardest to analyse since it was not ordered nor coloured. Of the two others, the parallel plot in *7)* was the easiest to analyse. The order was better and the colouring of the clusters helped a lot.
The clusters obtained by the parallel plots were very similar to those obtained by the chernoff faces and segment charts. In both cases there were on cluster with the majority of the western and Scandinavian countries and one with the eastern countries.

Assignment 2

1-3)



Regarding the average prices in 2005 for agricultural real estates, the northern part of Sweden is cheaper. It is most expensive in Stockholm and the nearby regions. In general, the average prices of agricultural real estates are lower in the south half of Sweden. For the average prices in 2010, the same interpretation can be given. It is more expensive on the south half than the north, and Stockholm is the most expensive region. A comparsion of the average prices in 2005 and 2010 gives that the average price has increased in all of the regions. In Stockholm, the average price has increased with approximately one million.

Group contribution

Each member of the group has contributed with their own graphs, written conclusions and participated in discussions about the lab.

R code

```
## ---- echo=FALSE---------------------------------------------------------
## 1)
jobs <- read.table(file = "C:\\Users\\Gustav\\Documents\\Visualization\\Lab2\\jobs.txt",
    header = TRUE, sep = "\t")
jobs$group <- 0
jobs[c(1, 3, 4, 5, 7, 8, 9, 10, 17), 11] <- "W"
jobs[c(2, 11, 13, 16), 11] <- "Sc"
jobs[c(6, 12, 14, 15, 18), 11] <- "S"
jobs[c(19, 20, 21, 22, 23, 24, 25, 26), 11] <- "E"
head(jobs)

## ---- echo=FALSE, fig.height=3.75, fig.width=6, warning=FALSE,
## message=FALSE, fig.align='center'---- 2)
library(portfolio)
map.market(id = jobs$Country, area = jobs$SI, group = jobs$group, color = jobs$Fin,
    main = "Jobs in service(area) and finance(color)", lab = c(T, T))


## ---- echo=FALSE, message=FALSE, warning=FALSE---------------------------
## 3)
library(aplpack)
jobs[, 12:17] <- 1
par(mar = c(1.1, 4.1, 4.1, 2.1))
faces(as.matrix(jobs[, c(2:10, 12:17)]), labels = jobs$Country, print.info = TRUE)

## ---- echo=FALSE, warning=FALSE, message=FALSE---------------------------
library(gclus)
library(graphics)
jobs <- jobs[, 1:11]
d <- dist(t(scale(jobs[, 2:10])))
jobs_order <- order.single(d)
jobs2 <- jobs[, jobs_order + 1]
head(jobs2, 1)

## ---- echo=FALSE---------------------------------------------------------
palette(heat.colors(9))
stars(jobs2, labels = as.character(jobs$Country), draw.segments = TRUE, key.labels = names(jobs2),
    key.loc = c(14, 1.85))


## ---- echo=FALSE, fig.height=3.5, fig.width=7, fig.align='center'--------
## 5)
parallelplot(jobs[, 2:10], horizontal.axis = FALSE)
```

```
## ---- echo=FALSE, fig.height=3.5, fig.width=7, fig.align='center',
## warning=FALSE---- 6)
set.seed(123445)
## a)
job_cor <- 1 - cor(jobs[, 2:10], )
## b)
library(TSP)
res <- solve_TSP(TSP(job_cor))
jobs2 <- jobs[, as.integer(res) + 1]
## c)
parallelplot(jobs2, horizontal.axis = FALSE)

## ---- echo=FALSE, message=FALSE, warning=FALSE, fig.height=3.5,
## fig.width=7, fig.align='center'----
library(seriation)
res2 <- seriate(as.dist(job_cor), method = "HC")
jobs_order2 <- get_order(res2)
jobs3 <- jobs[, jobs_order2 + 1]

color = 1 + (jobs3$Fin - min(jobs3$Fin) > 0.3 * (max(jobs3$Fin) - min(jobs3$Fin)))
color[18] <- 3
color[26] <- 4
color[color == 1] <- "royalblue"
color[color == 2] <- "darkorange"
color[color == 3] <- "red"
color[color == 4] <- "black"
parallelplot(jobs3, horizontal.axis = FALSE, col = color)

## ---- echo=FALSE---------------------------------------------------------
group = 1 + (jobs3$Fin - min(jobs3$Fin) > 0.3 * (max(jobs3$Fin) - min(jobs3$Fin)))
group[18] <- 3
group[26] <- 4
jobs$group2 <- group
subset(jobs, jobs$group2 == 1)[, c(1, 11)]

## ---- echo=FALSE---------------------------------------------------------
subset(jobs, jobs$group2 == 2)[, c(1, 11)]

## ---- echo=FALSE---------------------------------------------------------
subset(jobs, jobs$group2 == 3 | jobs$group2 == 4)[, c(1, 11)]

## ---- echo=FALSE, message=FALSE, warning=FALSE---------------------------
houseP <- read.csv("C:\\Users\\Gustav\\Documents\\Visualization\\Lab2\\prices_0510.csv",
    sep = ";")

map1 <- readRDS("C:\\Users\\Gustav\\Downloads\\SWE_adm1.rds")

temp <- map1@data
new <- merge(temp, houseP, by.x = "NAME_1", by.y = "County", all.y = F, all.x = T,
    sort = FALSE)

map1@data$Price2005 <- new$X2005
map1@data$Price2010 <- new$X2010
```

```r
spplot(map1, zcol = c("Price2005", "Price2010"), main = "Average purchase price 2005 and 2010",
    col.regions = heat.colors(16))


## ----code=readLines(knitr::purl('C:\\Users\\Gustav\\Documents\\Visualization\\Lab2\\Group4_lab2.Rmd',d
## = 1)), eval = FALSE, tidy=TRUE----
```