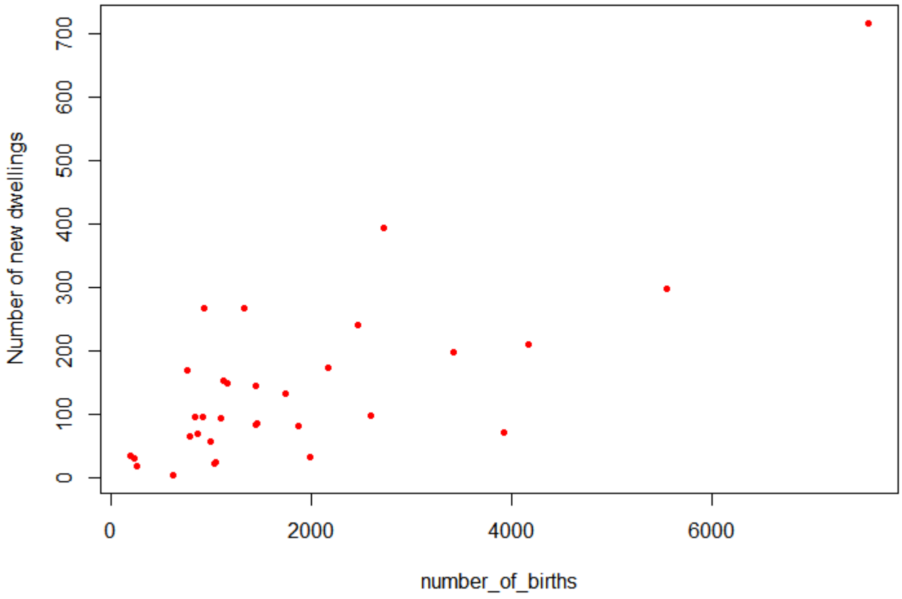
At this stage of our project, we are trying to enter the datasets, collected from statistics.gov.scot into the environment of R with the aid of RStudio.

To begin with, our previously selected datasets, both have as main dimension, the year (or time). Our main conception was to find out if there was a significant difference between the years, based on a specific geographic area (a council area, or country level).

However, the main difficulty we have encountered it was related with the structure of the .csv file, exported from the statistics.gov.scot. We had to do several transformations, as all the values were added in the same row. In the end of this paper you could find some of the commands that we used in our approach. But above all this, the most important fact is that plot() & pairs() were not the appropriate functions to export a time series plot.

For this reason, we had to follow a different approach and focus on our second dataset (number of births & number of new completed dwellings), in order to depict the state in a specific time period for all council areas. In this way, we had to also make the assumption that the results from total dwellings completions will be from Q4-2010 and not from the whole year (because of the dataset structure).

The results for 2010 are shown below



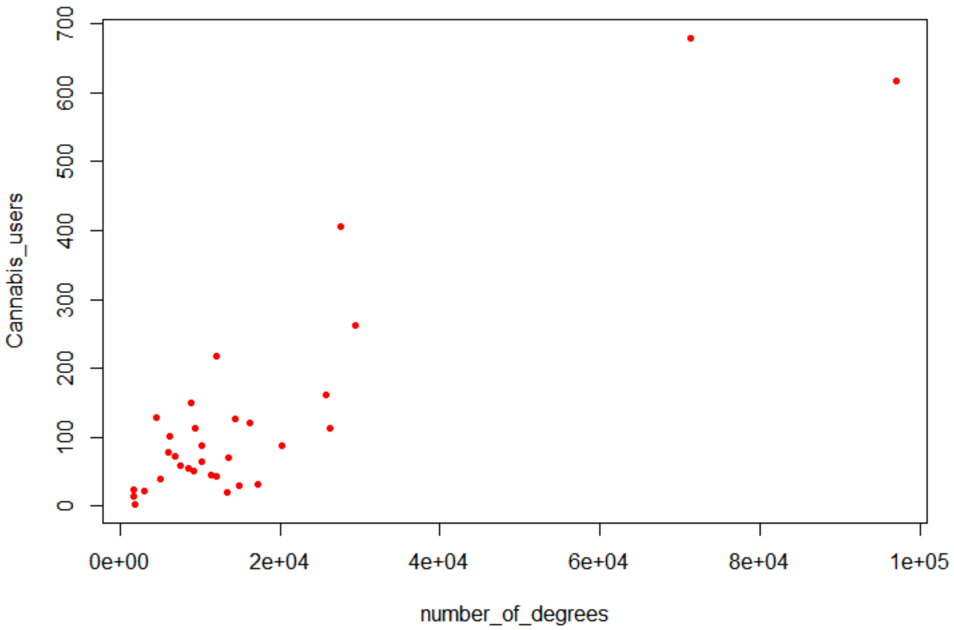
```
plot(birts_dwellings$Births, birts_dwellings$New_dwellings, col="red" ,
xlab="number_of_births" , ylab="Number of new dwellings", pch=20)
```

From the results we can assume the following:

Most of the observations are close in the start of the axes, where very a few are far from them. This makes sense as in Scotland some council areas are more “crowded” than the average. For example Glasgow City has 7565 births & 716 dwelling completions. Also it seems that these two variables are correlated, but still, we have to continue to further statistical analysis.

Also, in order to find an example of year to year match (and not quartile to year) we have tried a scenario of the total number of degree holders and registered cannabis users in Scotland.

The results for 2004 are shown below



Unfortunately, for an unknown reason, the axis of “number of degrees” is written in a scientific mode. (1e+05=100.000)

From the illustrated observations we can assume the same things in accordance with the previous example.

Findings

We still need to spend time to determine our scenario and its goal.

Focusing on a specific year, we can only depict the current situation and we cannot predict future tendencies.

Another approach is to depict two or three specific years and then compare-contrast them.

We will probably leave aside the “time series” scenarios.

Further suggestions for study

- Find a way to aggregate Q1,Q2,Q3,Q4 of 2010 for the total dwellings completions
- Find an appropriate package to demonstrate time series

Commands used for transforming matrices

```
Delete Rows
e <- e[-c(1,3,4,5,6), ,drop=F]

Delete Columns
e < -e [,-c(1,3,4,5,6), drop=F]

Rename Columns
names(your_matrix)[1] <- "Degree_Count"

Change the variable names
colnames(e)[colnames(e) == '1'] <- 'V1'

Convert matrices to data frames
e <- as.data.frame(e)

Combine two matrices into one
sc2 <- cbind(e, r)

Package for extended import tools in R
library(readr)
```

For the elective course of Information Systems Development

Kokovidis Symeon

April 2017