# Assignment 1

*S. Natalia Alvarado Pachon & Noriko Magara*

*Octorber 2nd, 2015*

## Using RStudio and Markdown

### Description

This document is the first attempt to use **Markdown** and **R**. I will use the R Code constructed in the last session of the class as main input and will show the main codes used for the outputs.

### Using RStudio

RStudio has built-in Datasets that users can use to generate outputs and practice their skills.

To access the list of built-in lists in RStudio, the user must type: `data()` To select a dataset, the user can use the code: `data("swiss")`

To inspect the dataset, the user can use the code: `?swiss`

For this excersice, we chose the dataset ***USArrests***

List of ***Number of Crimes per State***

| | Murder | Assault | UrbanPop | Rape |
|---|---|---|---|---|
| Alabama | 13.2 | 236 | 58 | 21.2 |
| Alaska | 10.0 | 263 | 48 | 44.5 |
| Arizona | 8.1 | 294 | 80 | 31.0 |
| Arkansas | 8.8 | 190 | 50 | 19.5 |
| California | 9.0 | 276 | 91 | 40.6 |
| Colorado | 7.9 | 204 | 78 | 38.7 |
| Connecticut | 3.3 | 110 | 77 | 11.1 |
| Delaware | 5.9 | 238 | 72 | 15.8 |
| Florida | 15.4 | 335 | 80 | 31.9 |
| Georgia | 17.4 | 211 | 60 | 25.8 |
| Hawaii | 5.3 | 46 | 83 | 20.2 |
| Idaho | 2.6 | 120 | 54 | 14.2 |
| Illinois | 10.4 | 249 | 83 | 24.0 |
| Indiana | 7.2 | 113 | 65 | 21.0 |
| Iowa | 2.2 | 56 | 57 | 11.3 |
| Kansas | 6.0 | 115 | 66 | 18.0 |
| Kentucky | 9.7 | 109 | 52 | 16.3 |
| Louisiana | 15.4 | 249 | 66 | 22.2 |
| Maine | 2.1 | 83 | 51 | 7.8 |
| Maryland | 11.3 | 300 | 67 | 27.8 |
| Massachusetts | 4.4 | 149 | 85 | 16.3 |
| Michigan | 12.1 | 255 | 74 | 35.1 |
| Minnesota | 2.7 | 72 | 66 | 14.9 |
| Mississippi | 16.1 | 259 | 44 | 17.1 |
| Missouri | 9.0 | 178 | 70 | 28.2 |
| Montana | 6.0 | 109 | 53 | 16.4 |

| | Murder | Assault | UrbanPop | Rape |
|---|---|---|---|---|
| Nebraska | 4.3 | 102 | 62 | 16.5 |
| Nevada | 12.2 | 252 | 81 | 46.0 |
| New Hampshire | 2.1 | 57 | 56 | 9.5 |
| New Jersey | 7.4 | 159 | 89 | 18.8 |
| New Mexico | 11.4 | 285 | 70 | 32.1 |
| New York | 11.1 | 254 | 86 | 26.1 |
| North Carolina | 13.0 | 337 | 45 | 16.1 |
| North Dakota | 0.8 | 45 | 44 | 7.3 |
| Ohio | 7.3 | 120 | 75 | 21.4 |
| Oklahoma | 6.6 | 151 | 68 | 20.0 |
| Oregon | 4.9 | 159 | 67 | 29.3 |
| Pennsylvania | 6.3 | 106 | 72 | 14.9 |
| Rhode Island | 3.4 | 174 | 87 | 8.3 |
| South Carolina | 14.4 | 279 | 48 | 22.5 |
| South Dakota | 3.8 | 86 | 45 | 12.8 |
| Tennessee | 13.2 | 188 | 59 | 26.9 |
| Texas | 12.7 | 201 | 80 | 25.5 |
| Utah | 3.2 | 120 | 80 | 22.9 |
| Vermont | 2.2 | 48 | 32 | 11.2 |
| Virginia | 8.5 | 156 | 63 | 20.7 |
| Washington | 4.0 | 145 | 73 | 26.2 |
| West Virginia | 5.7 | 81 | 39 | 9.3 |
| Wisconsin | 2.6 | 53 | 66 | 10.8 |
| Wyoming | 6.8 | 161 | 60 | 15.6 |

## Arrests in the US when convictions are related to murders

Given that the typology of the arrest can affect dramatically the number of incidents and the violence of the crime, we chose to focus only on Arrests related to Murders.

```
data("USArrests")
```

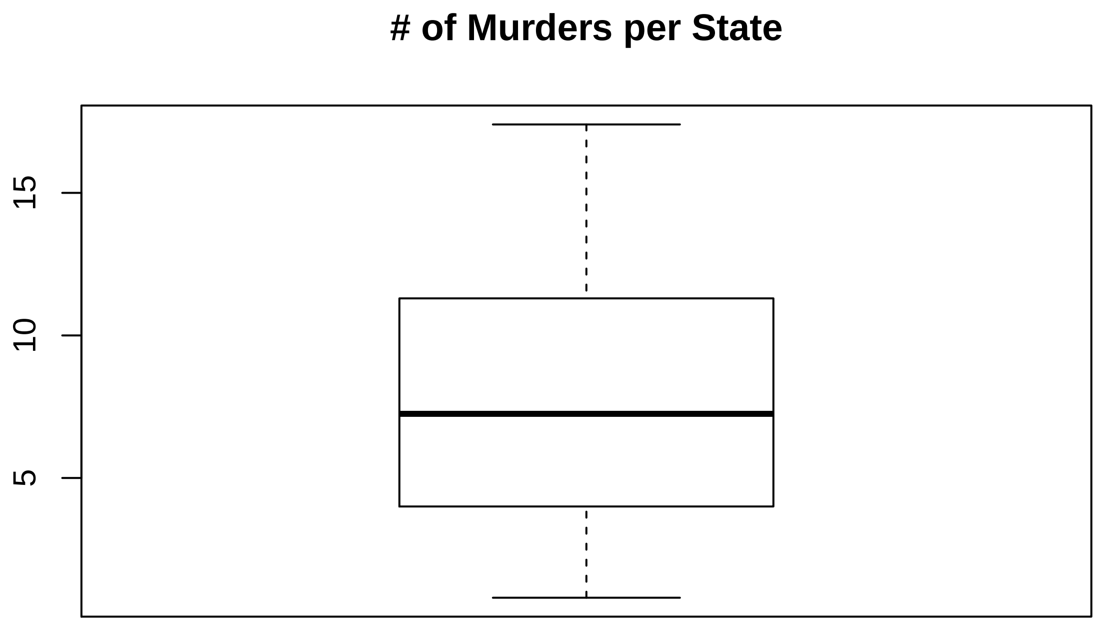
A summary of the information contained in the Dataset gives us a better detail of it

```
summary(USArrests$Murder)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

The boxplot for this Dataset:

```
boxplot(USArrests$Murder, main = '# of Murders per State')
```




summary(USArrests$Murder)boxplot(USArrests$Murder, main = ‘# of Murders per State’)

**Central Tendency and Variation**

The mean for the data in the US Arrest is ***7.788***

The median is ***7.25***

A histrogram shows more clearly the distribution of the data:

```r
hist(USArrests$Murder)
```


## Analysis on viorent crimes

### Combining some more variables together

Another built-in dataset ***state*** contains information related to the 50 states. Here, I combined ***state.abb*** (state name abbrebiation) and ***state.x77*** (basic demographics) to the original ***USArrests*** dataframe and created new one called ***UScombined***, which now has abbrebiations and some demographics such as population, income, etc.

The preparation of ***UScombined*** dataframe is coded in a separate code file to simplify the R Markdown file. The table below shows the whole resulting dataframe.

```r
source("USdataframe.R")
knitr::kable(UScombined)
```

| | ABB | Murder | Assault | UrbanPop | Rape | Population | Income | Illiteracy | Life.Exp | Murder |
|---|---|---|---|---|---|---|---|---|---|---|
| Alabama | AL | 13.2 | 236 | 58 | 21.2 | 3615 | 3624 | 2.1 | 69.05 | 15.1 |
| Alaska | AK | 10.0 | 263 | 48 | 44.5 | 365 | 6315 | 1.5 | 69.31 | 11.3 |
| Arizona | AZ | 8.1 | 294 | 80 | 31.0 | 2212 | 4530 | 1.8 | 70.55 | 7.8 |
| Arkansas | AR | 8.8 | 190 | 50 | 19.5 | 2110 | 3378 | 1.9 | 70.66 | 10.1 |
| California | CA | 9.0 | 276 | 91 | 40.6 | 21198 | 5114 | 1.1 | 71.71 | 10.3 |
| Colorado | CO | 7.9 | 204 | 78 | 38.7 | 2541 | 4884 | 0.7 | 72.06 | 6.8 |
| Connecticut | CT | 3.3 | 110 | 77 | 11.1 | 3100 | 5348 | 1.1 | 72.48 | 3.1 |
| Delaware | DE | 5.9 | 238 | 72 | 15.8 | 579 | 4809 | 0.9 | 70.06 | 6.2 |

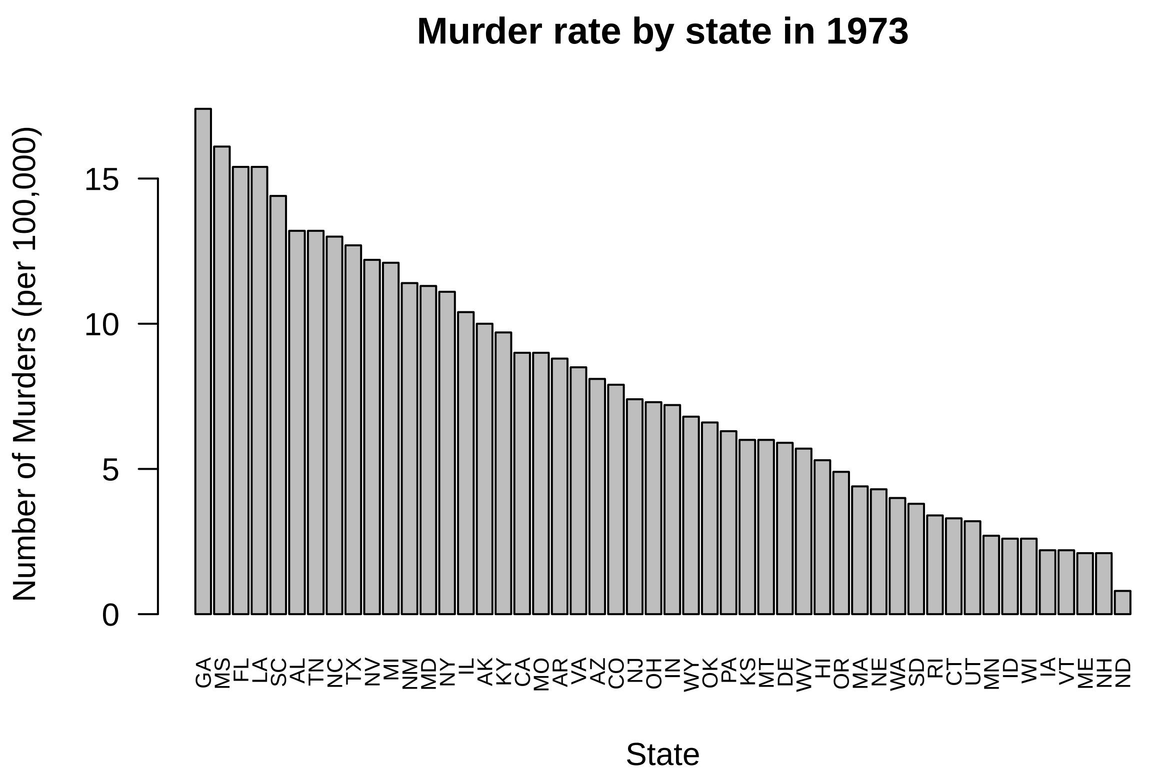
| | ABB | Murder | Assault | UrbanPop | Rape | Population | Income | Illiteracy | Life.Exp | Murder |
|---|---|---|---|---|---|---|---|---|---|---|
| Florida | FL | 15.4 | 335 | 80 | 31.9 | 8277 | 4815 | 1.3 | 70.66 | 10.7 |
| Georgia | GA | 17.4 | 211 | 60 | 25.8 | 4931 | 4091 | 2.0 | 68.54 | 13.9 |
| Hawaii | HI | 5.3 | 46 | 83 | 20.2 | 868 | 4963 | 1.9 | 73.60 | 6.2 |
| Idaho | ID | 2.6 | 120 | 54 | 14.2 | 813 | 4119 | 0.6 | 71.87 | 5.3 |
| Illinois | IL | 10.4 | 249 | 83 | 24.0 | 11197 | 5107 | 0.9 | 70.14 | 10.3 |
| Indiana | IN | 7.2 | 113 | 65 | 21.0 | 5313 | 4458 | 0.7 | 70.88 | 7.1 |
| Iowa | IA | 2.2 | 56 | 57 | 11.3 | 2861 | 4628 | 0.5 | 72.56 | 2.3 |
| Kansas | KS | 6.0 | 115 | 66 | 18.0 | 2280 | 4669 | 0.6 | 72.58 | 4.5 |
| Kentucky | KY | 9.7 | 109 | 52 | 16.3 | 3387 | 3712 | 1.6 | 70.10 | 10.6 |
| Louisiana | LA | 15.4 | 249 | 66 | 22.2 | 3806 | 3545 | 2.8 | 68.76 | 13.2 |
| Maine | ME | 2.1 | 83 | 51 | 7.8 | 1058 | 3694 | 0.7 | 70.39 | 2.7 |
| Maryland | MD | 11.3 | 300 | 67 | 27.8 | 4122 | 5299 | 0.9 | 70.22 | 8.5 |
| Massachusetts | MA | 4.4 | 149 | 85 | 16.3 | 5814 | 4755 | 1.1 | 71.83 | 3.3 |
| Michigan | MI | 12.1 | 255 | 74 | 35.1 | 9111 | 4751 | 0.9 | 70.63 | 11.1 |
| Minnesota | MN | 2.7 | 72 | 66 | 14.9 | 3921 | 4675 | 0.6 | 72.96 | 2.3 |
| Mississippi | MS | 16.1 | 259 | 44 | 17.1 | 2341 | 3098 | 2.4 | 68.09 | 12.5 |
| Missouri | MO | 9.0 | 178 | 70 | 28.2 | 4767 | 4254 | 0.8 | 70.69 | 9.3 |
| Montana | MT | 6.0 | 109 | 53 | 16.4 | 746 | 4347 | 0.6 | 70.56 | 5.0 |
| Nebraska | NE | 4.3 | 102 | 62 | 16.5 | 1544 | 4508 | 0.6 | 72.60 | 2.9 |
| Nevada | NV | 12.2 | 252 | 81 | 46.0 | 590 | 5149 | 0.5 | 69.03 | 11.5 |
| New Hampshire | NH | 2.1 | 57 | 56 | 9.5 | 812 | 4281 | 0.7 | 71.23 | 3.3 |
| New Jersey | NJ | 7.4 | 159 | 89 | 18.8 | 7333 | 5237 | 1.1 | 70.93 | 5.2 |
| New Mexico | NM | 11.4 | 285 | 70 | 32.1 | 1144 | 3601 | 2.2 | 70.32 | 9.7 |
| New York | NY | 11.1 | 254 | 86 | 26.1 | 18076 | 4903 | 1.4 | 70.55 | 10.9 |
| North Carolina | NC | 13.0 | 337 | 45 | 16.1 | 5441 | 3875 | 1.8 | 69.21 | 11.1 |
| North Dakota | ND | 0.8 | 45 | 44 | 7.3 | 637 | 5087 | 0.8 | 72.78 | 1.4 |
| Ohio | OH | 7.3 | 120 | 75 | 21.4 | 10735 | 4561 | 0.8 | 70.82 | 7.4 |
| Oklahoma | OK | 6.6 | 151 | 68 | 20.0 | 2715 | 3983 | 1.1 | 71.42 | 6.4 |
| Oregon | OR | 4.9 | 159 | 67 | 29.3 | 2284 | 4660 | 0.6 | 72.13 | 4.2 |
| Pennsylvania | PA | 6.3 | 106 | 72 | 14.9 | 11860 | 4449 | 1.0 | 70.43 | 6.1 |
| Rhode Island | RI | 3.4 | 174 | 87 | 8.3 | 931 | 4558 | 1.3 | 71.90 | 2.4 |
| South Carolina | SC | 14.4 | 279 | 48 | 22.5 | 2816 | 3635 | 2.3 | 67.96 | 11.6 |
| South Dakota | SD | 3.8 | 86 | 45 | 12.8 | 681 | 4167 | 0.5 | 72.08 | 1.7 |
| Tennessee | TN | 13.2 | 188 | 59 | 26.9 | 4173 | 3821 | 1.7 | 70.11 | 11.0 |
| Texas | TX | 12.7 | 201 | 80 | 25.5 | 12237 | 4188 | 2.2 | 70.90 | 12.2 |
| Utah | UT | 3.2 | 120 | 80 | 22.9 | 1203 | 4022 | 0.6 | 72.90 | 4.5 |
| Vermont | VT | 2.2 | 48 | 32 | 11.2 | 472 | 3907 | 0.6 | 71.64 | 5.5 |
| Virginia | VA | 8.5 | 156 | 63 | 20.7 | 4981 | 4701 | 1.4 | 70.08 | 9.5 |
| Washington | WA | 4.0 | 145 | 73 | 26.2 | 3559 | 4864 | 0.6 | 71.72 | 4.3 |
| West Virginia | WV | 5.7 | 81 | 39 | 9.3 | 1799 | 3617 | 1.4 | 69.48 | 6.7 |
| Wisconsin | WI | 2.6 | 53 | 66 | 10.8 | 4589 | 4468 | 0.7 | 72.48 | 3.0 |
| Wyoming | WY | 6.8 | 161 | 60 | 15.6 | 376 | 4566 | 0.6 | 70.29 | 6.9 |

Now, I can for example label the bar chart by abbrebiations.

```
# sort the dataframe
SortUScombined <- UScombined[order(-UScombined$Murder) , ]

barplot(SortUScombined$Murder, names.arg=SortUScombined$ABB,
        cex.names=0.7, las=2,
        main="Murder rate by state in 1973",
```

```
       xlab="State", ylab="Number of Murders (per 100,000)"
       )
```

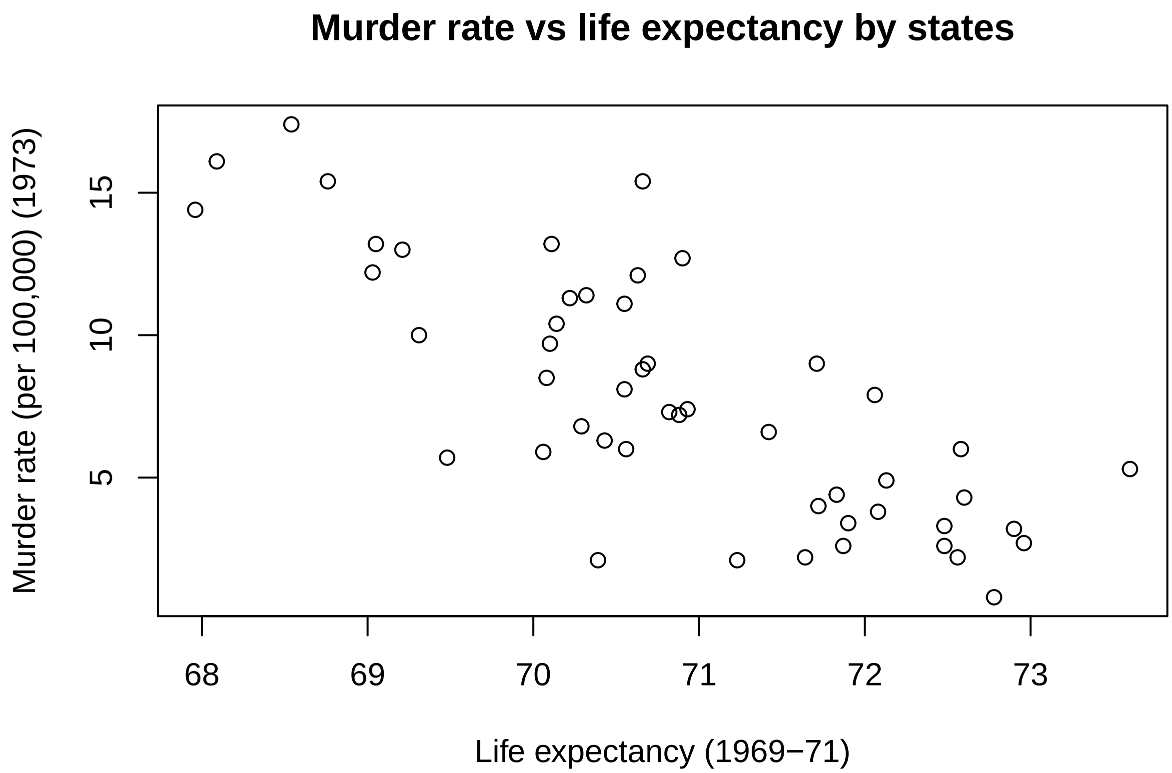



#### Muder rate vs life expectancy

I am interested in exploring the relationship between murder rate and other variable. Here I look into the relationship between murder rate and life expectancy.

**Scatter plot** Below is the scatter plot of the two variables.

```
plot(UScombined$Life.Exp, UScombined$Murder,
     main = 'Murder rate vs life expectancy by states',
     xlab = 'Life expectancy (1969-71)', ylab = 'Murder rate (per 100,000) (1973)'
     )
```




It seems like there is a negative correlation, meaning the longer people live, the lower the murder rate in the state. (In those states, probably people are mentally in good health, as well as physicall health.)

**Correlation test** Next, in order to formally test the correlation, I run a correlation test.

```
cor.test(UScombined$Life.Exp, UScombined$Murder)
```

```
##
##  Pearson's product-moment correlation
##
## data:  UScombined$Life.Exp and UScombined$Murder
## t = -8.5934, df = 48, p-value = 2.836e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8686217 -0.6385121
## sample estimates:
##        cor
## -0.7784985
```

The result indicates that there is a statistically significant negative correlation of -0.7784985.

**Linear regression** Finally, I estimate a simple bivariate linear regression.

```r
lmr <- lm(UScombined$Murder~UScombined$Life.Exp)
summary(lmr)
```

```
##
## Call:
## lm(formula = UScombined$Murder ~ UScombined$Life.Exp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.922 -1.582 -0.156  1.703  7.060
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         186.8206    20.8375   8.966 7.97e-12 ***
## UScombined$Life.Exp  -2.5259     0.2939  -8.593 2.84e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.762 on 48 degrees of freedom
## Multiple R-squared:  0.6061, Adjusted R-squared:  0.5979
## F-statistic: 73.85 on 1 and 48 DF,  p-value: 2.836e-11
```

The result shows that each additional year of life expectancy is on average associated with a -2.5259044 increase (which is equivalent to a 2.5259044 decrease) in murder rate per 100,000. However, this model has several problems.

- simultaneity: the model assumes that life expectancy affects murder rate, but murder rate may also affect life expectancy (the higher the murder rate is, the shorter people live (because more of them are killed!)). Sounds dangerous!
- Lack of control variables: definitely there are other variables that explains murder rates. By excluding them, the estimate may be biased.