Assignment 1

S. Natalia Alvarado Pachon & Noriko Magara

October 2nd, 2015

Using RStudio and Markdown

Description

This document is the first attempt to use **Markdown** and **R**. We 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

R Studio has built-in Data sets that users can use to generate outputs and practice their skills.

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

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

For this exercise, we chose the data set ***USArrests*** which contains the number of arrests per 100,000 inhabitants per type of crime (assault, murder or rape) and the percentage of the urban population in 1973. These data are indicated per State.

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

	Murder	Assault	UrbanPop	Rape
Missouri	9.0	178	70	28.2
Montana	6.0	109	53	16.4
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

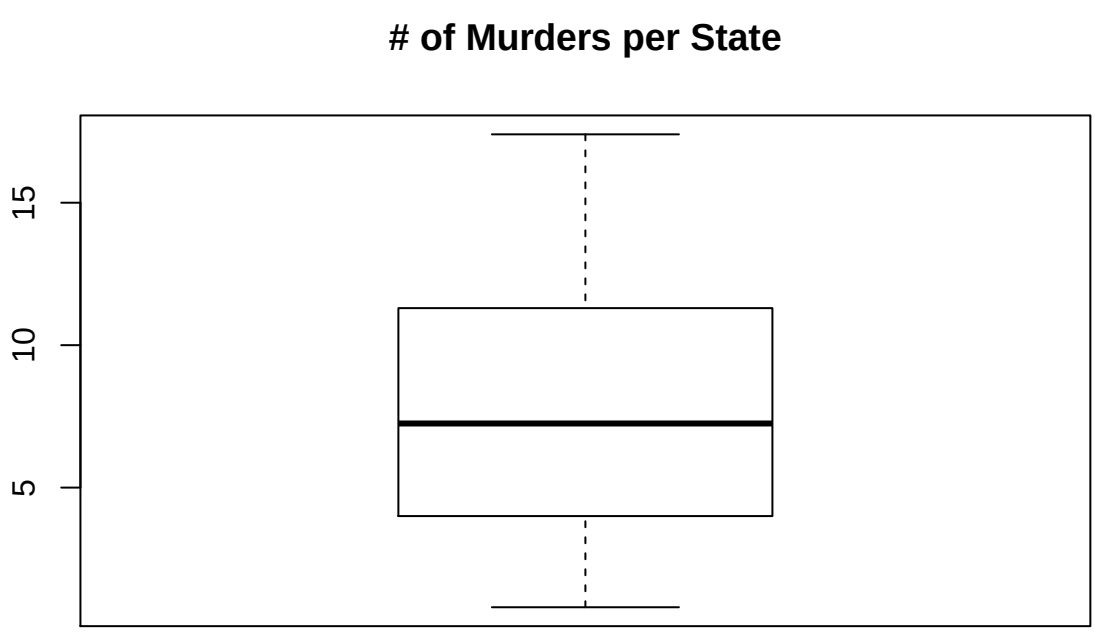
Given that the typology of the arrest can affect dramatically the number of incidents and the violence of the crime, we chose to focus only on Arrests related to Murders. This type of crime is better defined and the outcome can only be given in absolute terms, while variables such as assault can suffer from problems such as lack of report.

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

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

This summary shows the mean number of reported murders per 100,000 on this given year was of 7,2. However, strong differences can be found within the country, **North Dakota** accounted for 0,8 murders while **Georgia accounted** for 17.4.

This indicates that the distribution of the data may not be as concentrated around the mean or that these rates are outliers. To have a better understanding of the distribution, a boxplot can explain the distribution according to the different quartiles:



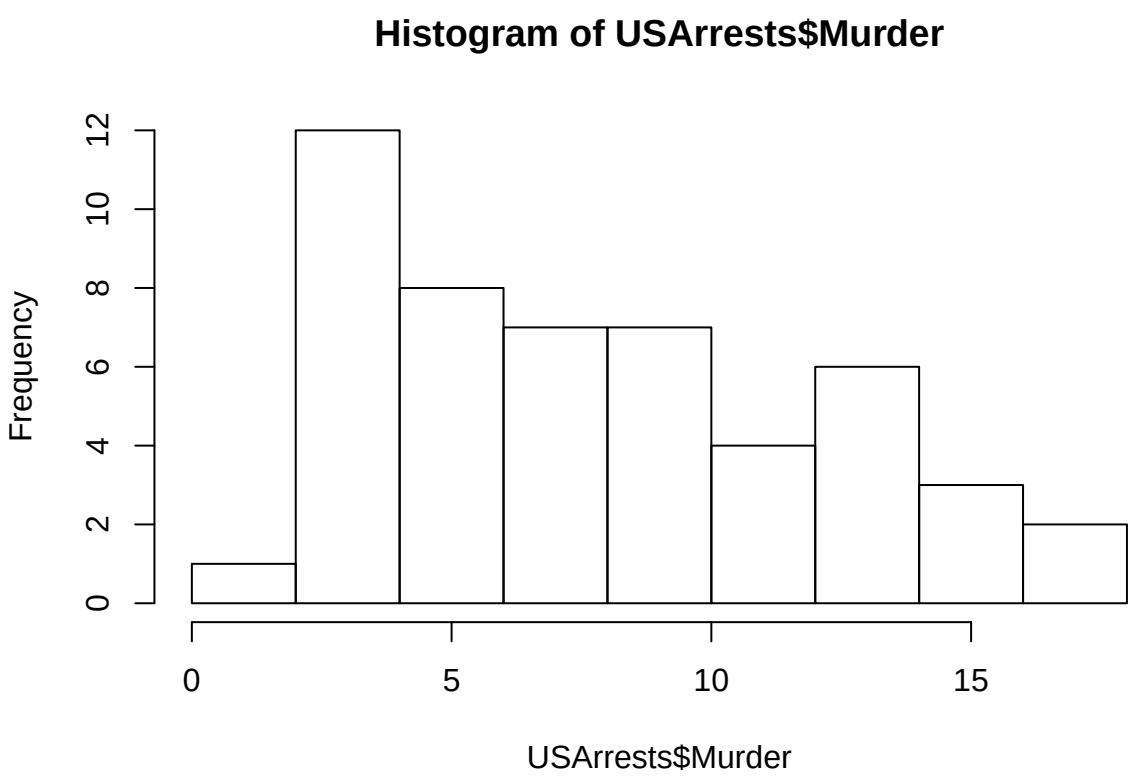
This box plot seems to suggest most observations are in the second quartile.

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:



This histogram shows a distribution skewed to the left. However, no information can be inferred from this data set but the number of crimes committed in this given year. To add other variables could help identify patterns or causalities.

Analysis on viorent crimes

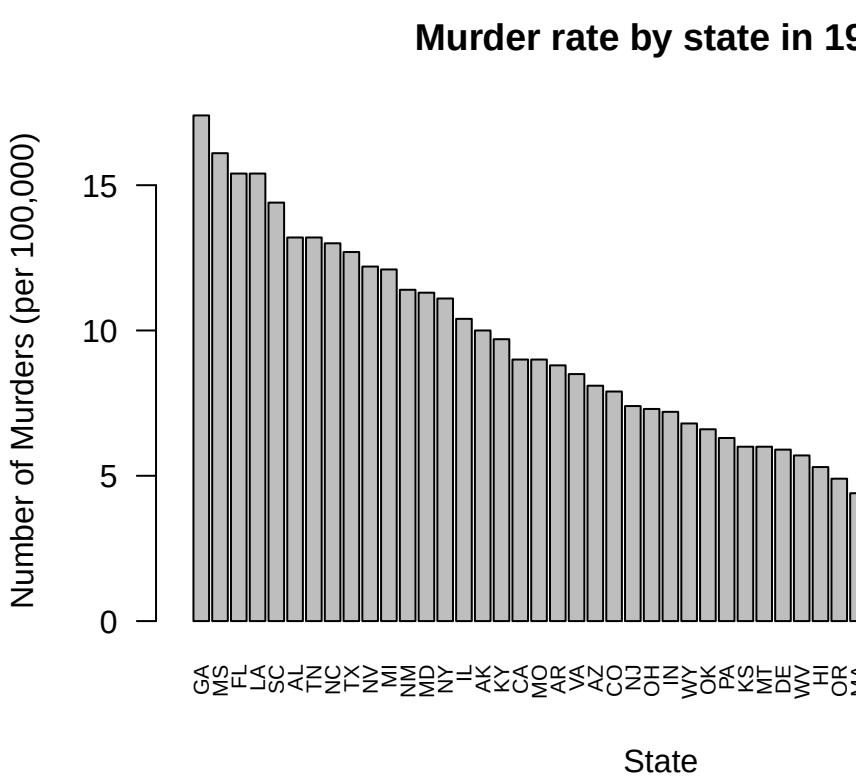
Combining some more variables together

Another built-in data set ***state*** contains information related to the 50 states from 1976. Here, ***state.abb*** (state name abbreviation) and ***state.x77*** (basic demographics) where combined to the original ***USArrests*** data frame and created new one called ***UScombined***, which now has abbreviations and some demographics such as population, income, etc.

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

	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
Florida	FL	15.4	335	80	31.9	8277	4815	1.3	70.66	10.7

	ABB	Murder	Assault	UrbanPop	Rape	Population	Income	Illiteracy	Life.Exp	Murder
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

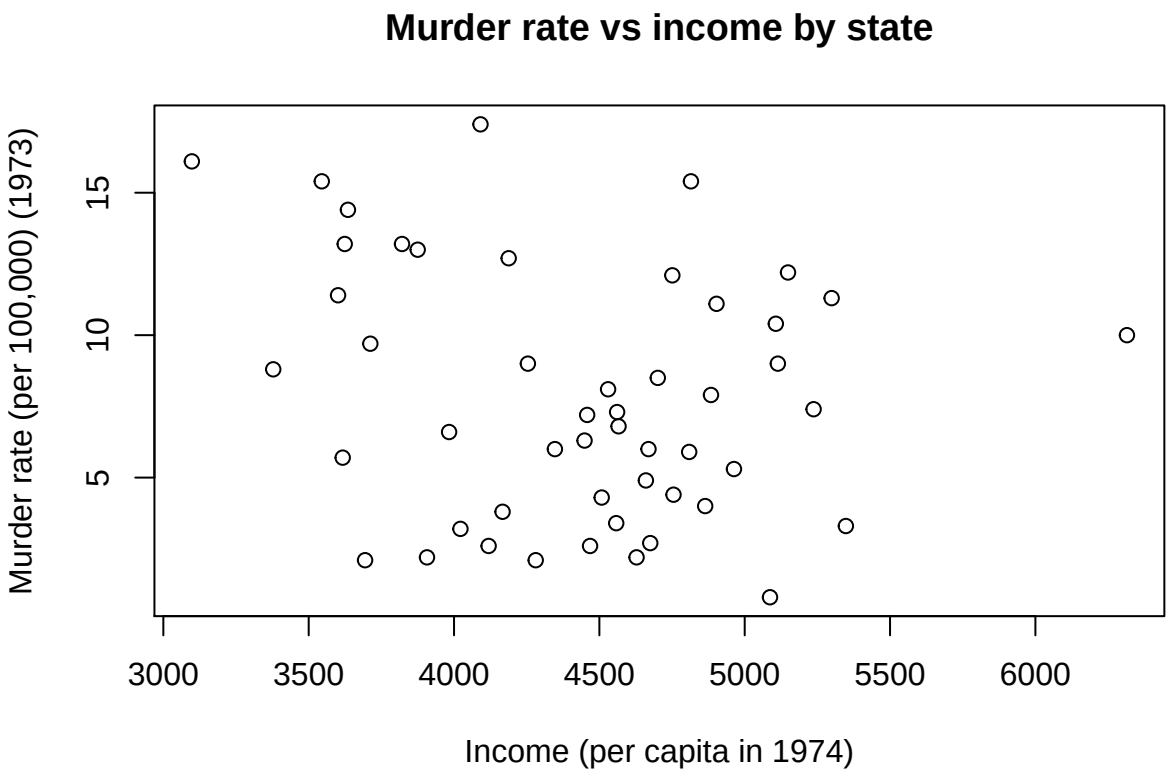


Now, the bar chart can be labeled by abbreviations.

Muder rate vs income

For this exercise, the relationship between murder rate and other variable will be explored. The link between poverty and criminality has been one of the main topics for research in the academia with many social scientists supporting the claim that there is, at least, a high correlation between both. In this case the variable Murders and Income will be jointly scrutinized.

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



It is very difficult to tell whether this relation has any correlation due to the way the observations are scattered. However, a line can be fitted through the points, trying to facilitate the way the relationship is observed.

This fitted line allows to see a negative relationship between income and murder rates; the more income, the less murders are reported in a State.

Correlation test Next, in order to formally test the correlation, a test can be made.

```
##
##  Pearson's product-moment correlation
##
## data:  UScombined$Income and UScombined$Murder
## t = -1.5268, df = 48, p-value = 0.1334
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.46565891  0.06716608
## sample estimates:
##       cor
## -0.215205
```

The result indicates that there is a statistically insignificant negative correlation of -0.215205.

Linear regression Finally, a simple bi-variate linear regression is estimated, to see the possible impact of this relation.

```
##
## Call:
```

```
## lm(formula = UScombined$Murder ~ UScombined$Income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8196 -3.2865 -0.5779  4.1364  9.0860
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       14.5544782  4.4734048   3.254  0.00209 **
## UScombined$Income -0.0015254  0.0009991  -1.527  0.13338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.298 on 48 degrees of freedom
## Multiple R-squared:  0.04631,    Adjusted R-squared:  0.02644
## F-statistic: 2.331 on 1 and 48 DF,  p-value: 0.1334
```

The result shows that each additional unit of income per capita is associated on average with a -0.0015254 increase (which is equivalent to a 0.0015254 decrease) in murder rate per 100,000. However, this estimates are not statistically significant, as shown by theyr P-values.

From the theoretical perspective, this model has several problems.

- Endogeneity: the model assumes that income affects murder rate, but income can also be correlated with variables not accounted for and captures in the error term.
- Lack of control variables: definitely, there are other variables that could help explains murder rates. By excluding them, the estimate are probably biased. A recommendation for this is to follow theory.