

# Assignments

This page will contain all the assignments you submit for the class.

## Instructions for all assignments

I want you to submit your assignment as a PDF, so I can keep a record of what the code looked like that day. I also want you to include your answers on your personal GitHub website. This will be good practice for editing your website and it will help you produce something you can keep after the class is over.

1. Download the Assignment1.Rmd file from Canvas. You can use this as a template for writing your answers. It's the same as what you can see on my website in the Assignments tab. Once we're done with this I'll edit the text on the website to include the solutions.
2. On RStudio, open a new R script in RStudio (File > New File > R Script). This is where you can test out your R code. You'll write your R commands and draw plots here.
3. Once you have finalized your code, copy and paste your results into this template (Assignment 1.Rmd). For example, if you produced a plot as the solution to one of the problems, you can copy and paste the R code in R markdown by using the ```{r} ``` command. Answer the questions in full sentences and Save.
4. Produce a PDF file with your answers. To do this, knit to PDF (use Knit button at the top of RStudio), locate the PDF file in your docs folder (it's in the same folder as the Rproj), and submit that on on Canvas in Assignment 1.
5. Build Website, go to GitHub desktop, commit and push. Now your solutions should be on your website as well.

## Assignment 1

### Collaborators:

### Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

```
#install.packages("datasets")  
library(datasets)
```

```
USArrests
```

##	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
## 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

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package `datasets`, so there's no need to load data from your computer. Why is it useful to rename the dataset?

It is useful to renamed USArrests to dat.us because it is easier to write and it is good practice to rewrite data for yourself so you can create your own data which can be replicated by another person if they use the original data.

```
dat.us <- USArrests  
  
head(dat.us)
```

```
##           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
```

## Problem 2

Use this command to make the state names into a new variable called State.

```
dat$state <- tolower(rownames(USArrests))
```

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset USArrests.

```
names(dat.us)
```

```
## [1] "Murder"  "Assault" "UrbanPop" "Rape"
```

**Answer:** The four variables are Murder, Assault, UrbanPop, Rape.

## Problem 3

What type of variable (from the DVB chapter) is Murder?

**Answer:** Murder is a quantitative variable.

What R Type of variable is it?

**Answer:** Murder is numeric.

## Problem 4

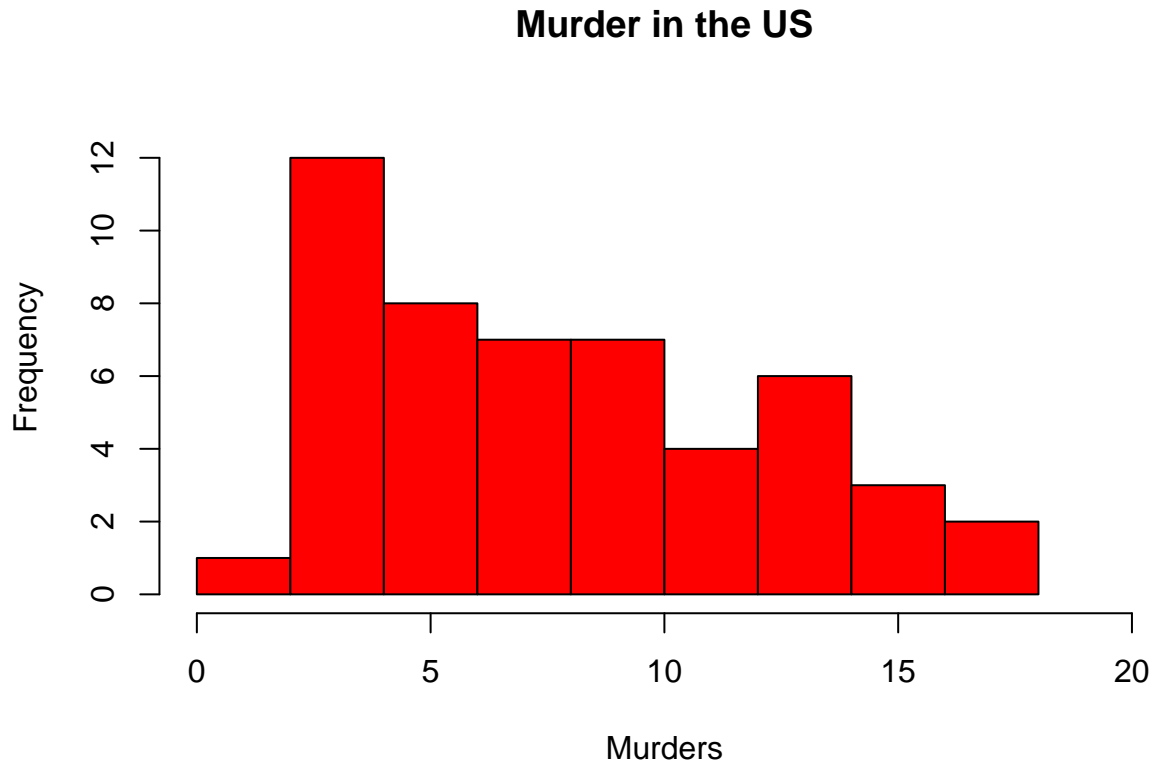
What information is contained in this dataset, in general? What do the numbers mean?

**Answer:** The dataset contains the data of murder, assault, rape and urbanpop from all 50 US states. The numbers represent the frequency of arrests for one of the four variables in a state during the time frame that the data was collected.

### Problem 5

Draw a histogram of `Murder` with proper labels and title.

```
hist(dat.us$Murder, xlab= "Murders", ylab="Frequency", main= "Murder in the US", xlim=(c(0, 20)), ylim=
```



### Problem 6

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

```
summary(dat.us$Murder)
```

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

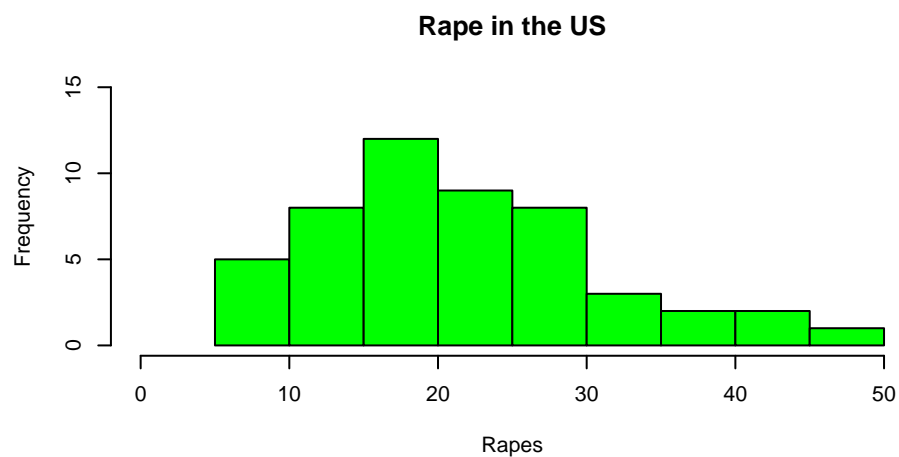
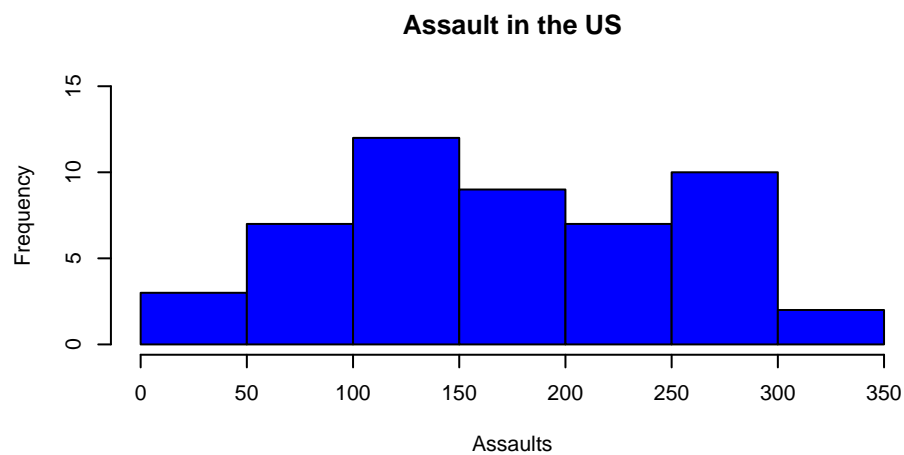
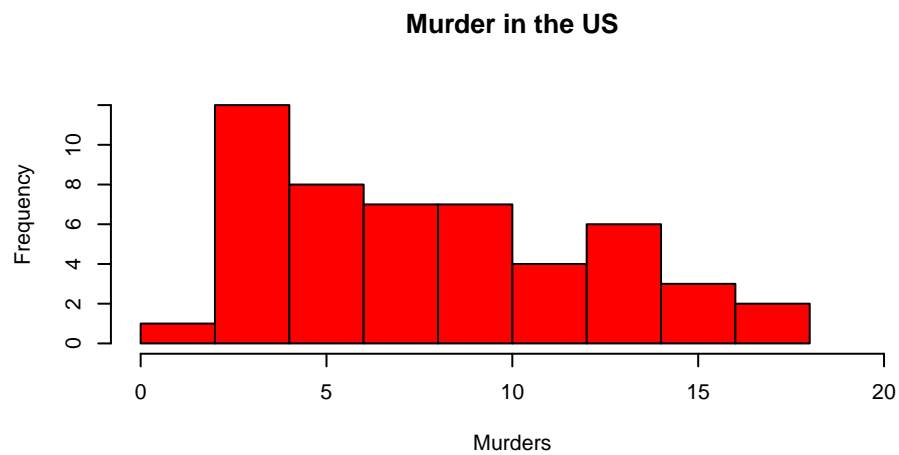
**Answer:** The mean for `Murder` is 7.788 and the median is 7.250. The mean is the amount that each subject would have if all of the values were added together and evenly distributed. If all 50 states had the same frequency of arrests for murders then it would be 7.788. The median is the middle value where exactly 50% of the values fall either above or below it. In the US, 50% of states have an arrest for murder frequency above 7.250 and the other 50% is below that. The median is highly robust because it is not greatly affected by outliers. The mean is the most common measure of central tendency but it is not robust because it will change based on the skewness of the distribution. A quartile indicates an interval that contains 25% or a quarter of the data. The first quartile for “Murder” is 4.075 which means that 25% of the “Murder” data falls

below 4.075 and the 3rd quartile is 11.250 which means that 25% of US states have a frequency of arrests for murder that is higher than 11.250. R gives you the 1st and 3rd quartile because those values are useful in determining the interquartile range (IQR). The IQR is the central half which means that 50% of the data falls within the 1st and 3rd quartile. In a box plot, values 1.5 IQRs above or below the tails are considered outliers.

### Problem 7

Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`. Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

```
par(mfrow=c(3,1))  
  
hist(dat.us$Murder, xlab= "Murders", ylab="Frequency", main= "Murder in the US", xlim=(c(0, 20)), ylim=  
  
hist(dat.us$Assault, xlab= "Assaults", ylab="Frequency", main= "Assault in the US", xlim=(c(0, 350)), y=  
  
hist(dat.us$Rape, xlab= "Rapes", ylab="Frequency", main= "Rape in the US", xlim=(c(0, 50)), ylim=(c(0, )
```



What does the command `par` do, in your own words (you can look this up by asking R `?par`)?

**Answer:** Command `par` is used to set parameters. The `mfrow` input allows you to create an array to plot multiple graphs on one window. The command `par(mfrow=c(3,1))` allows three graphs to be plotted in three rows.

What can you learn from plotting the histograms together?

Answer: When the histograms are plotted together it is easier to compare the skewness and spread of each plot. You can see where each histogram has its peaks and outliers.

## Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
install.packages("maps")
install.packages("ggplot2")

library('maps')
library('ggplot2')

ggplot(dat.us, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```

What does this code do? Explain what each line is doing.

Answer: The lines `library('maps')` and `library('ggplot2')` are pulling from the packages that were installed. The line `ggplot(dat.us, aes(map_id=state, fill=Murder))` is creating a ggplot with the `USArrests` dataset. The plot is set with an aesthetic of a map with the US states. Each state is filled in with its respective murder arrest data. The line `geom_map(map=map_data("state"))` contains the map coordinates for each US state. The last line `'expand_limits(x=map_data("state")$long, y = map_data("state")$lat)'` ensures that the limits of the plot include a single value for all plots. The x and y axis of this plot contains the value of "state" from the map data. The x axis is longitude and the y axis is latitude. Together this code creates a map of the US with each state filled in with its value for murder arrests. The darker blue indicates that the murder arrest frequency is 5 and below and the light blue indicates that it is 15 and above.

## Assignment 2

(Coming soon)