# My First RMarkdown Document

*Andre Guimaraes Duarte*

*July 15, 2016*

## R Markdown

This is my first R Markdown document. Iam required to submit all **MSAN 593** homework in RMkardown. I am going to import a large dataset from the Housing Affordability Data System (HADS) from Data,gov using the `read.csv` function. The Housing Affordability Data System (HADS) is a set of gousing-unit level datasets that measures the affordability of housing units and the housing cost burdens of households, relative to area median incomes, poverty level incomes, and Fair Market Rents.

```
read.csv("~/Desktop/hadsData.txt")
```

This fails for a few reasons, namely, I read in the file and stored it to no where. So I wasted my time waiting for R to read in the file, and then when it finally did, it printed some rows to the Console window, and then the following message was printed to the console `[ reached getOption("max.print") -- omitted 64434 rows ]` and voila, the data disappeared faster than it loaded. Now I know better.

```
hadsData <- read.csv("~/Desktop/hadsData.txt")
```

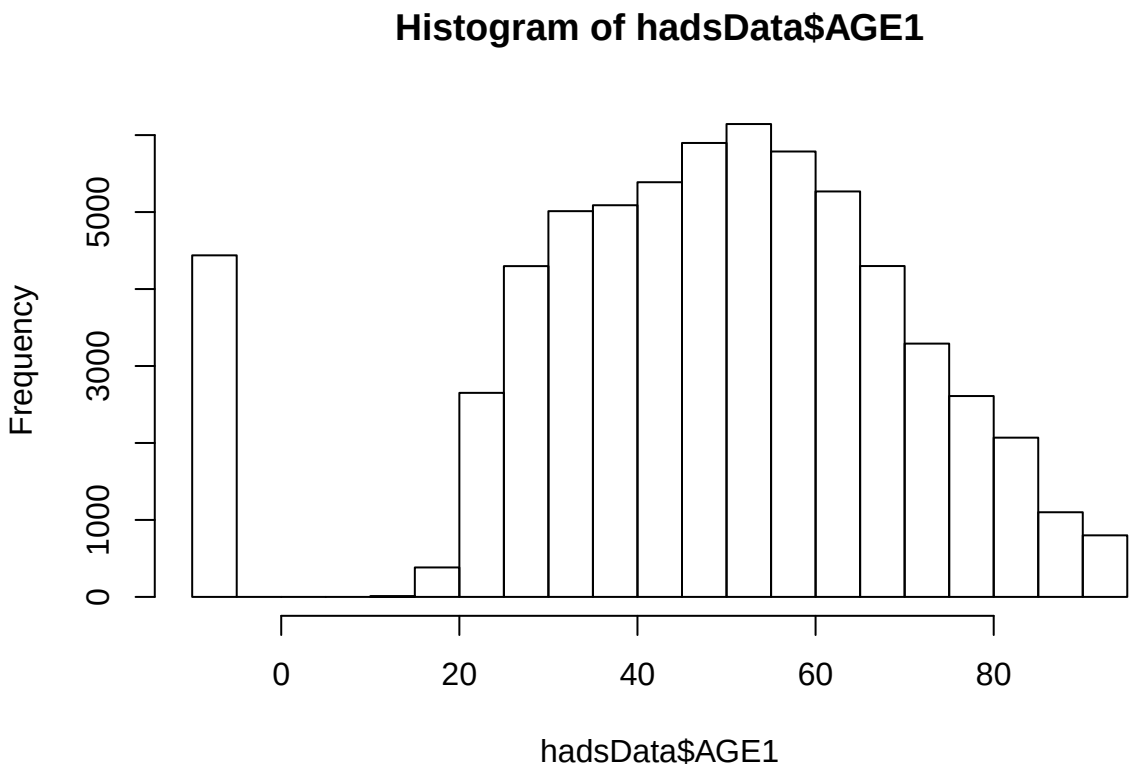
The HADS dataset has 64535 rows and 99 columns. It would be imprudent to run `str()` on all 99 variables in the dataset, so I will just show the first ten.

```
## 'data.frame':    64535 obs. of  10 variables:
##  $ CONTROL: Factor w/ 64535 levels "'100003130103'",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ AGE1   : int  82 50 53 67 26 56 50 26 60 26 ...
##  $ METRO3 : Factor w/ 5 levels "'1'","'2'","'3'",..: 3 5 5 5 1 2 1 4 5 4 ...
##  $ REGION : Factor w/ 4 levels "'1'","'2'","'3'",..: 1 3 3 3 3 3 3 4 4 2 ...
##  $ LMED   : int  73738 55846 55846 55846 60991 62066 60991 52322 50296 63221 ...
##  $ FMR    : int  956 1100 1100 949 737 657 988 773 1125 552 ...
##  $ L30    : int  15738 17165 13750 13750 14801 13170 16646 13489 13115 13338 ...
##  $ L50    : int  26213 28604 22897 22897 24628 21924 27713 22471 21859 22199 ...
##  $ L80    : int  40322 45744 36614 36614 39421 35073 44340 35929 34939 35501 ...
##  $ IPOV   : int  11067 24218 15470 13964 15492 12005 18050 15992 15452 12005 ...
```

I am particularly interested in further exploring the `AGE1` and `REGION` variables of the data set. I will now create two subsections, one for each variable.

### AGE1

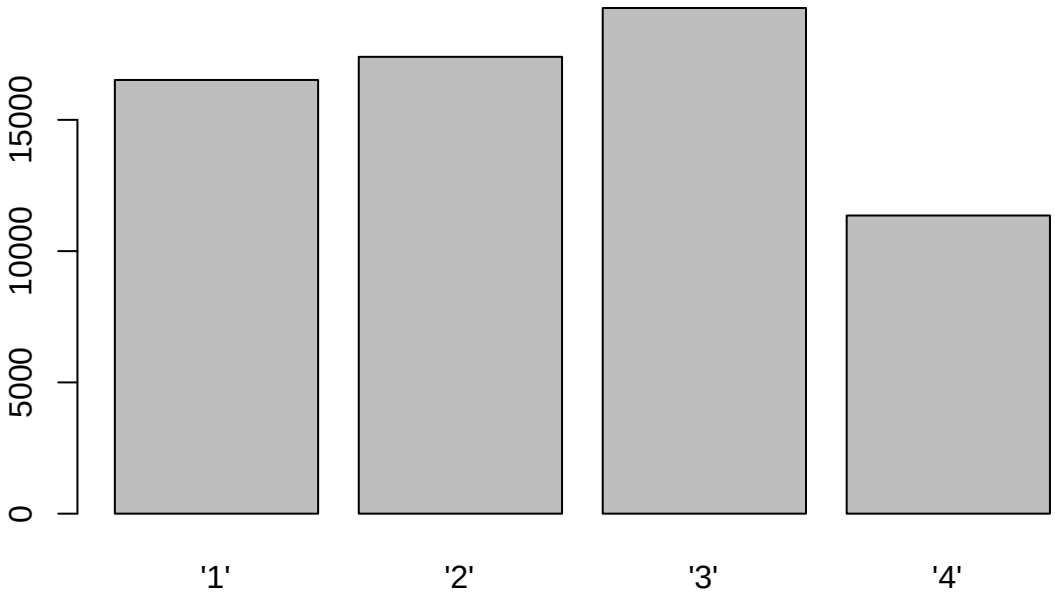
`AGE1` is defined by the HADS data dictionary as the age of the head of household. The mean age of household is 47.97 and the min and max are -9 and 93 respectively. Clearly there is something funky going on in the data if the minimum age is -9. Anyhow, I will generate a histogram of the ages:




This is not a very pretty histogram, but it gets the idea across.

### `REGION`

`REGION` is defined by the HADS data dictionary as the census region. I observe that even though `type` of `REGION` is `integer`, the `class` is `factor`, so I will generate a bar graph to evaluate the frequencies of occurence of each region:




So that we can get exact numbers, I can also generate a contingency table:

```
##
```

```
##   '1'   '2'   '3'   '4'
## 16519 17400 19260 11356
```

## Conclusion

This brings me to the end of my first RMarkdown document.