Data Science Problems

## 3. Data Wrangling

Data set of people is based names on the [1990 Census](https://catalog.data.gov/dataset/names-from-census-1990) and state population by age estimated from Census [factfinder](https://factfinder.census.gov).

load('sample\_people.Rda')

### 3.1 Update the data frame (sample\_people)

Separate the person’s name into a first and last name column.

* Try with strsplit() and lapply() functions to select the appropriate element.
* Now try with the tidyr package using the seperate() function.
* Show results with the head() function

### 3.2 Create a summary table by state.

Recommend using the dplyr package with group\_by() and summarize() functions.

* What is the sample average and median population by state? Give both the frequency and relative frequency of the stats using inline code on a rmarkdown table:

|  |  |  |
| --- | --- | --- |
| Statistic | Average | Median |
| Frequency | *STAT1* | *STAT2* |
| Relative Frequency | *STAT3* | *STAT4* |

* Order the summary table from largest to smallest population.
* [Add a html table](https://rmarkdown.rstudio.com/lesson-7.html) of the 5 larges states using kable()
* Add a plot of all the state populations using plot()
* Describe the sample population data based on the table and the plot including the probability of an individual in the sample being from California.

### 3.3 Add the R default data state.name & state.region

* Create reference data frame with state.name & state.region
* Add the region to the sample\_people data frame with the merge() function.
* Sample one person, what is his/her percentile based on their age?
* For that person, what is his/her percentile based on their age for his/her region?
* For that person, what is his/her percentile based on their age for his/her state?
* Using the mutate(), create columns for each persons age percentile for US, their region and their state. Validate calculation based on the person sampled.