Data Science Problems

## 4. Data Exploration

Congratulations! You have been hired as an analyst for a community development company that plans and develops 55+ communities for retirees. The first report you are assigned to work on is an exploratory analysis for a new location in the US.

## Recommended packages  
library(ggplot2)  
library(gridExtra)  
  
## weather station data  
load('weather\_us.Rda')  
load('stations\_us.Rda')  
load("city\_stations.Rda")  
  
## city data  
load('us\_cities.Rda')  
  
## functions  
temp\_convert<- function(x){ (x/10)\*(9/5) + 32}

### 4.1 Review the US city data

* Plot a box plot of us\_cities population (pop). What do you observe about the data?
* Retirees are interested in living close to a city that is not to large or to small. Exclude cities less than 1st quartile and above 3rd quartile. Name the new data frame us\_cities\_mid.
* Plot a bar chart of mid size cities by state. What 3 states have the most mid sized cities?

### 4.2 Review weather data

* Tidy up the data for analysis
  + Filter weather\_us for the first 5 columns (id, data, prcp, tmax, tmin)
  + Create a data frame of **unique** stations with the elevation, latitude and longitude using station\_data.
  + **Merge** or **Join** the station data and city station data into the weather\_us data frame.
* Convert units
  + tmax and tmin using the temp\_convert function
  + Filter extreme weather with min temp < -30 and max temp > 120
  + elevation from meters to feet. Multiply by 3.281
* Create 4 scatter plots for January (low temperature) and July (high temperature). Try geom\_point() with alpha = 0.1 due to amount of data.
  + January low temp vs latitude
  + January low temp vs elevation
  + July high temp vs latitude
  + July high temp vs elevation
* Combine plots into one graphic. Try grid.arrange(p1, p2, p3, p4, ncol=2, nrow=2).
  + Describe the relationship between temperature, latitude and elevation.
  + What happens after an elevation of about 1000 ft?

### 4.3 Recommendation list.

* Summarize weather data with grouping by city
  + Average min temp from January
  + Average max temp from July
  + Average elevation
  + Add each variable to the us\_cities\_mid from question one.
* Filter the data for good weather temperature.
  + Cold temperature > 32
  + Hot temperature < 100
  + Plot a bar chart of good weather mid sized cities by state. How does this compare to all mid sized cities in question 1.
* Assume the targeted demographic is interested in living in a higher elevation (> 1000 ft). What is the top recommended location?