Assignment 1

Import Data into R

```
library(readxl)
library(ggplot2)

#reading data and setting col names
population <- read_excel("population.xls",
                         skip = 9, col_names = FALSE, na = '.')

colNm = c("Code", "County Municipality", "Population", "Population growth",
          "Live Births", "Deaths", "Population surplus", "In_mig_tot", "In_mig_from_sc",
          "In_mig_from_ros", "In_mig_from_ab", "Out_mig_tot", "Out_mig_from_sc",
          "Out_mig_from_ros", "Out_mig_from_ab", "Net_mig_tot", "Net_mig_from_sc",
          "Net_mig_from_ros", "Net_mig_from_ab", "Adjustments")
colnames(population) = colNm

#splitting counties and cities
population$keep = population$Code <= 25
counties = population[population$keep,c("Code", "County Municipality", "Population")]
cities = population[!population$keep,c("Code", "County Municipality", "Population")]
```

Creating a function to select random city

```
#function to randomly select a city
selectRandCity = function(data){
  total_pop = sum(data$Population)
  data$prob = data$Population/total_pop
  data$cumPop = cumsum(data$prob)
  randNum = runif(1, 0, 1)
  selected_ind = which.max((data$cumPop > randNum )*1)
  return(selected_ind)
}
```

In this part we are creating a function that selects and returns a randomly selected row index. The random selection is based on the population of the cities. Cities with lager population have higher probability of getting selected.

I have converted the populations column into probebilities by dividing it by the total population. The cumPop column is the cumulative sum of the prob column, which is probabilities of the cities. After this a random number is generated between 0 and 1 and the city having the cumulative probability in that range is selected.

Using the function created in the previous step

```
#setting seed to get reproducable results
set.seed(123456)
#selecting 20 random cities
selectedCities = cities[1,]
for(i in 1:20){
  ind = selectRandCity(cities)
```

```
  selectedCities[i,] = cities[ind,]
  cities = cities[-ind,]
}
```

Selected Cities

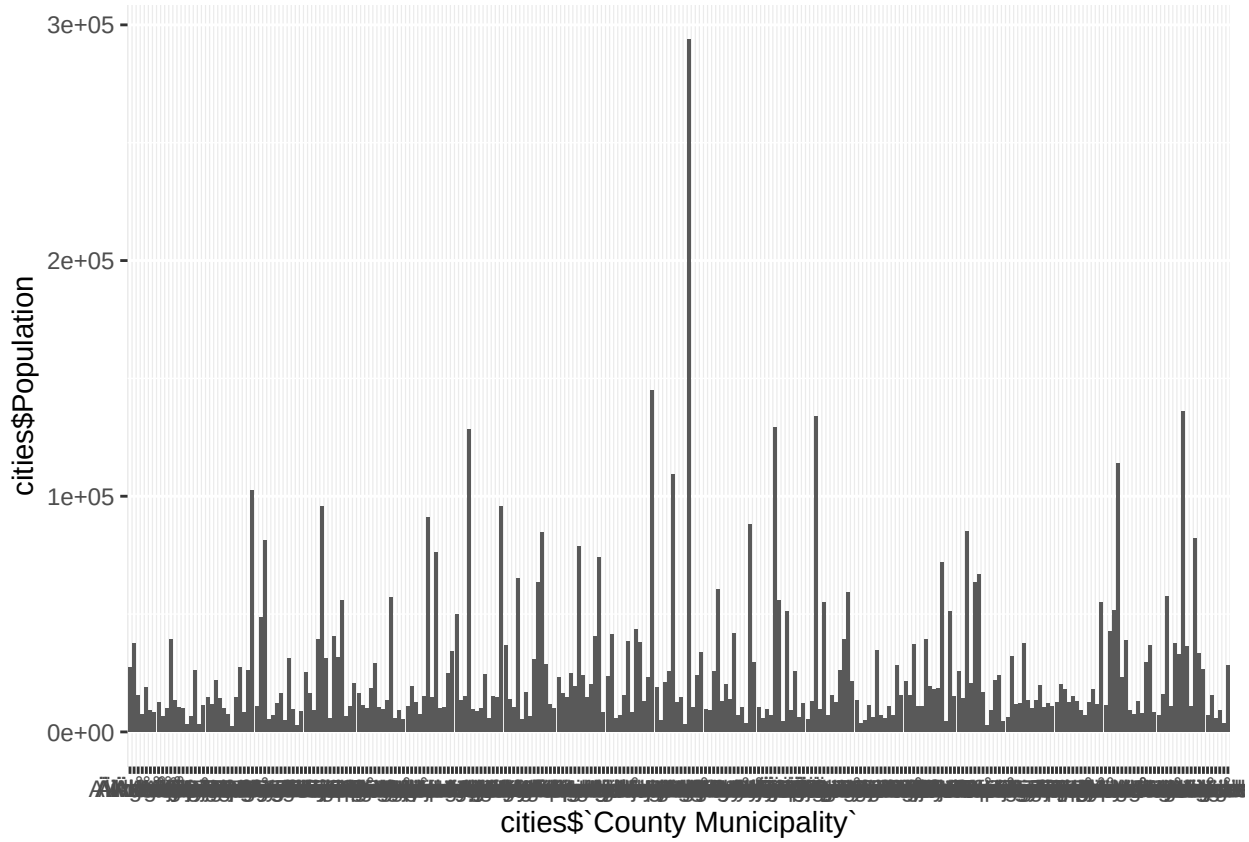
```
#print selected cities
print(selectedCities)
```

```
# A tibble: 20 x 3
    Code `County Municipality` Population
 * <dbl> <chr>                       <dbl>
 1  1883 Karlskoga                   29742
 2  1491 Ulricehamn                  22753
 3   880 Kalmar                      62388
 4   680 Jönköping                  126331
 5   781 Ljungby                     27410
 6   160 Täby                        63014
 7  1287 Trelleborg                  41891
 8   180 Stockholm                  829417
 9  2580 Luleå                       73950
10   380 Uppsala                    194751
11  1982 Fagersta                    12249
12  1480 Göteborg                   507330
13  2281 Sundsvall                   95533
14  2180 Gävle                       94352
15  2581 Piteå                       40860
16  2121 Ovanåker                    11530
17  2061 Smedjebacken                10758
18   483 Katrineholm                 32303
19   881 Nybro                       19576
20  1861 Hallsberg                   15235
```

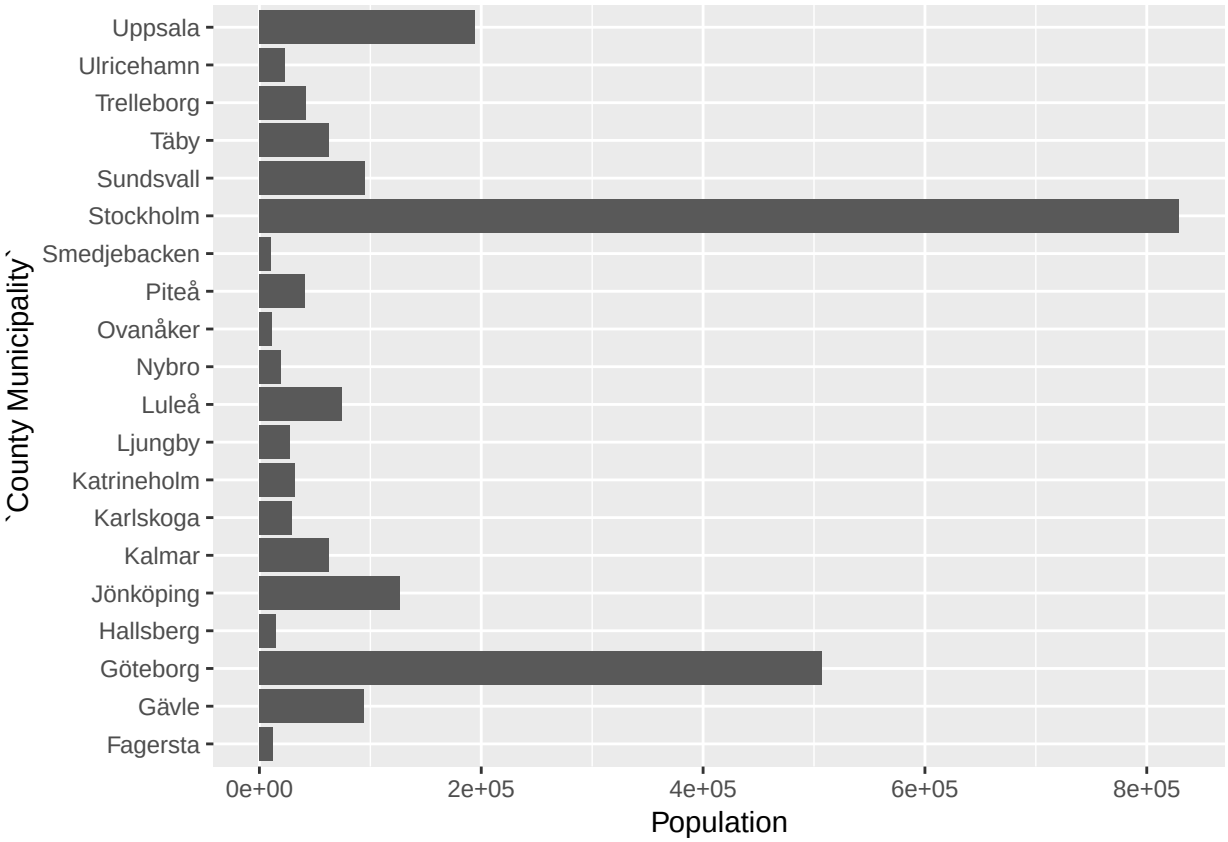
The cities with the largest population are selected in this random process. Stockholm and Goteborg which are one of the largest populated cities in Sweden get selected almost every time as they have a very large probability of getting selected.

Plot showing population of cities selected

```
ggplot(cities, aes(cities$`County Municipality` , cities$Population)) +
  geom_histogram(stat = "identity")
```




```
ggplot(selectedCities, aes(`County Municipality` , Population)) +
  geom_histogram(stat = "identity") + coord_flip()
```



Some of the cities with the largest population, like Stockholm, Gothemborg, Upsella, were selected as they were given higher priority than others. Stockholm has the largest population, so it has the highest probability of getting selected. This turns out to be true as it gets selected almost every time we run a simulation with different seed.

The majority of the cities do not have too large a population, so most of the 20 random cities selected is made up from these. They have low probability of gettin selected but there are too many of such cities, so these cities are the ones that fill up the majority of the 20 random picks.