Obtaining population data

Chris Shaw

28 May 2016

Population data

This file downloads detailed US population data by year state and age group from:

http://seer.cancer.gov/popdata/download.html

This is a large dataset which is then processed to obtain relevant summarised datasets:

- population of each state by year

The raw data is obtained from the url if it doesn't already exist:

```
filename<-"us.1969_2014.19ages.adjusted.txt.gz"
data_url <-"http://seer.cancer.gov/popdata/yr1969_2014.19ages/us.1969_2014.19ages.adjusted.txt.gz"

if (!file.exists(filename)){
        download.file(data_url, filename)
}

# Load data and rename data column to something more friendly
#
raw_data<-read.table(filename, stringsAsFactors = F)
names(raw_data)[1]<-"raw"
head(raw_data)
```

```
##                           raw
## 1 1969AL01001991910000000159
## 2 1969AL01001991910100000657
## 3 1969AL01001991910200001137
## 4 1969AL01001991910300000956
## 5 1969AL01001991910400000721
## 6 1969AL01001991910500000424
```

The adjusted dataset is used which takes into account the population shifts in certain states following hurricane Katrina.

From time to time the data sets are updated - check and change the data_url above to get the latest data

It can be seen that the raw data is compacted into just one column. The next section parses and summarises this column.

Data format

The format of this data is described here:

http://seer.cancer.gov/popdata/popdic.html

The key elements needed for this module are:

field	start column	end column	Data type
Year	1	4	numeric
State postal abbreviation	5	6	

|character | |Population |19 |26 |numeric |

Field	Start Col	End Col	Data Type
Year	1	4	numeric
State	5	6	char
Population	19	26	numeric

The following code extracts the relevant columns and caches the results.

```r
raw_data$year<-as.numeric(substr(raw_data$raw, 1,4))
raw_data$state<-as.character(substr(raw_data$raw, 5,6))
raw_data$population<-as.numeric(substr(raw_data$raw, 19,26))
```

Create a new dataset

In this case we just want population per state per year:

```r
summary <- raw_data %>% group_by(year, state) %>% summarise(population=sum(population))

# write out as a csv
write.table(summary, "states.csv", sep=",", row.names = F)
```