HW7

Diego Valdes

February 25, 2019

```
rm(list=ls()) # clear work space
#dev.off(dev.list()["RStudioGD"]) # clear plots

suppressWarnings(require(ggplot2))
```

```
## Loading required package: ggplot2
```

```
suppressWarnings(require(maps))
```

```
## Loading required package: maps
```

```
suppressWarnings(require(mapproj))
```

```
## Loading required package: mapproj
```

```
suppressWarnings(require(openxlsx))
```

```
## Loading required package: openxlsx
```

```
suppressWarnings(require(zipcode))
```

```
## Loading required package: zipcode
```

```
suppressWarnings(require(openintro))
```

```
## Loading required package: openintro
## Please visit openintro.org for free statistics materials
##
## Attaching package: 'openintro'
## The following object is masked from 'package:ggplot2':
##
##     diamonds
## The following objects are masked from 'package:datasets':
##
##     cars, trees
```

```
getwd()
```

```
## [1] "C:/Users/dvjr2/Google Drive/Documents/Syracuse/IST_687/HW"
```

```
setwd("C:/Users/dvjr2/Google Drive/Documents/Syracuse/IST_687/HW/")

fileName = "MedianZIP_2_2.xlsx" # manually removed first row and converted to csv

data = read.xlsx(fileName, colNames = FALSE) # read the file
data = data[-1:-2 , ] # remove unwanted bs

colnames(data) = c("zip", "median", "mean", "population") # col names

data$zip = clean.zipcodes(data$zip) # clean the zip codes
```

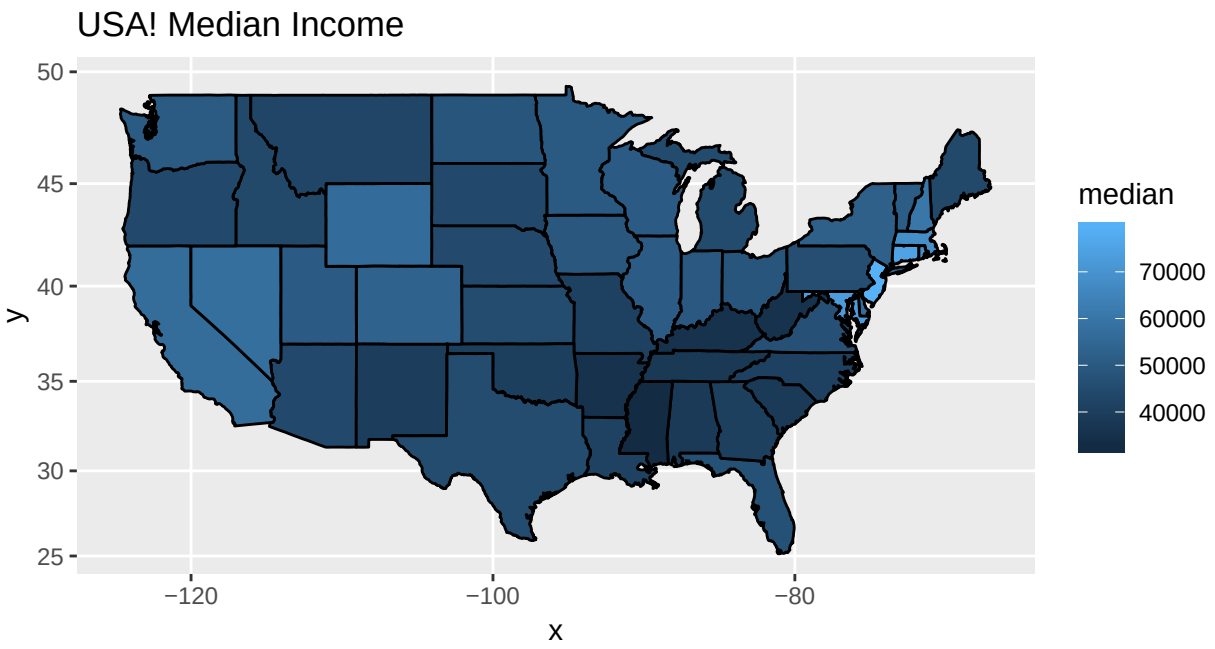
```r
data("zipcode") # load zipcode data

newData = zipcode[zipcode$state != 'HI' & zipcode$state != 'AK', ] # remove HI and AK data

stateData = merge(newData, data, by = "zip") # merge the two df into one all mighty and powerful super
str(stateData) # take a look
```

```
## 'data.frame':    32321 obs. of  8 variables:
##  $ zip       : chr  "01001" "01002" "01003" "01005" ...
##  $ city      : chr  "Agawam" "Amherst" "Amherst" "Barre" ...
##  $ state     : chr  "MA" "MA" "MA" "MA" ...
##  $ latitude  : num  42.1 42.4 42.4 42.4 42.3 ...
##  $ longitude : num  -72.6 -72.5 -72.6 -72.1 -72.4 ...
##  $ median    : chr  "56662.573499999999" "49853.417699999998" "28462" "75423" ...
##  $ mean      : chr  "66687.750899999999" "75062.634300000005" "35121" "82442" ...
##  $ population: chr  "16445" "28069" "8491" "4798" ...
```

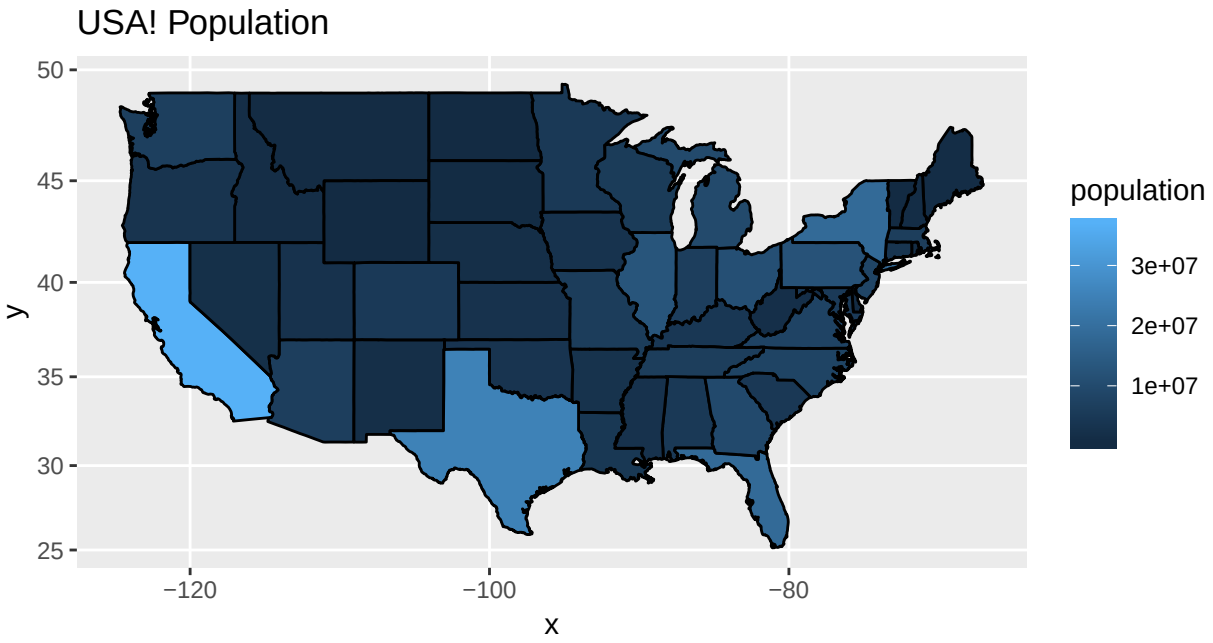
```r
# lets do some generic cleaning
stateData$stateName = tolower(abbr2state(stateData$state))
stateData$state = tolower(stateData$state)
stateData$median = as.numeric(stateData$median)
stateData$mean = as.numeric(stateData$mean)
```

```
## Warning: NAs introduced by coercion
```

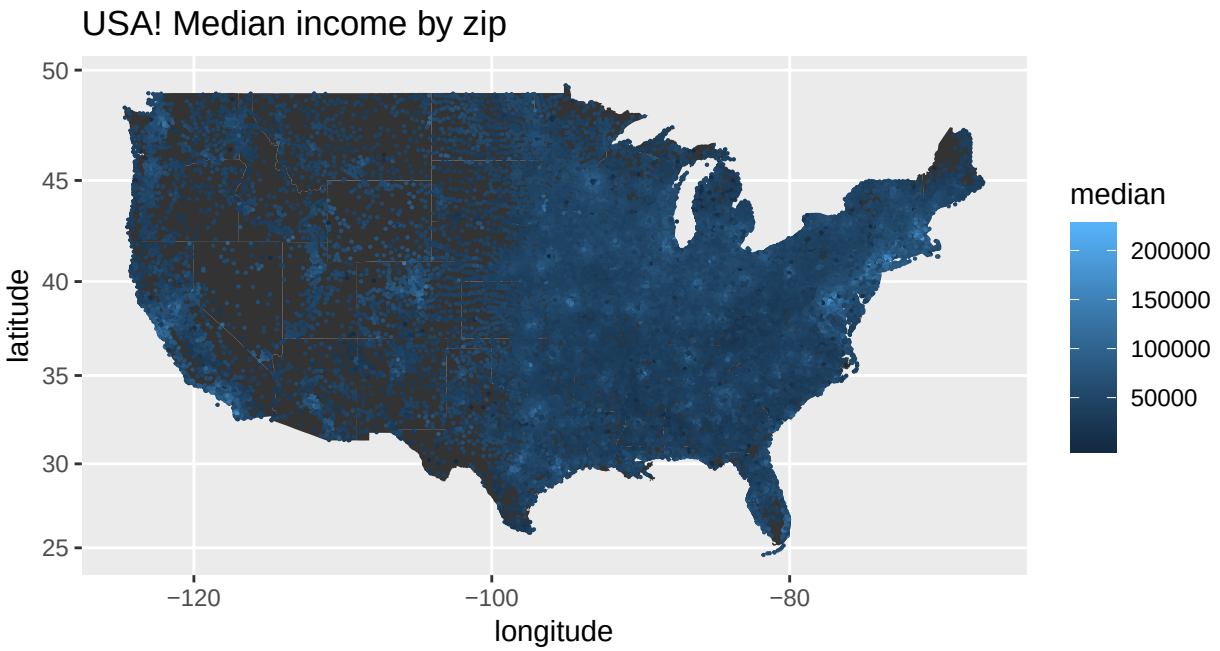
```r
stateData$population = as.numeric(stateData$population)

stateIncome = stateData[ , c(2, 3, 6, 7, 8, 9)] # state income data

us = map_data("state") # get map data

#themap = ggplot(stateIncome, aes(map_id = stateName)) +
#  geom_map(map = us, fill = "white",  color = "black") +
#  expand_limits(x = us$long, y = us$lat) +
#  coord_map() + ggtitle("USA!")

# mutate the data for plotting
dataMedian = as.data.frame(tapply(stateIncome$median, stateIncome$stateName, median))
dataPop = as.data.frame(tapply(stateIncome$population, stateIncome$stateName, sum))
dataMedian$states = rownames(dataMedian)
dataPop$states = rownames(dataPop)

colnames(dataMedian) = c("median", "states")
colnames(dataPop) = c("population", "states")

# median income
ggplot(dataMedian, aes(map_id = states)) +
  geom_map(map = us, aes(fill = median), color = "black") +
  expand_limits(x = us$long, y = us$lat) +
  coord_map() + ggtitle("USA! Median Income")
```



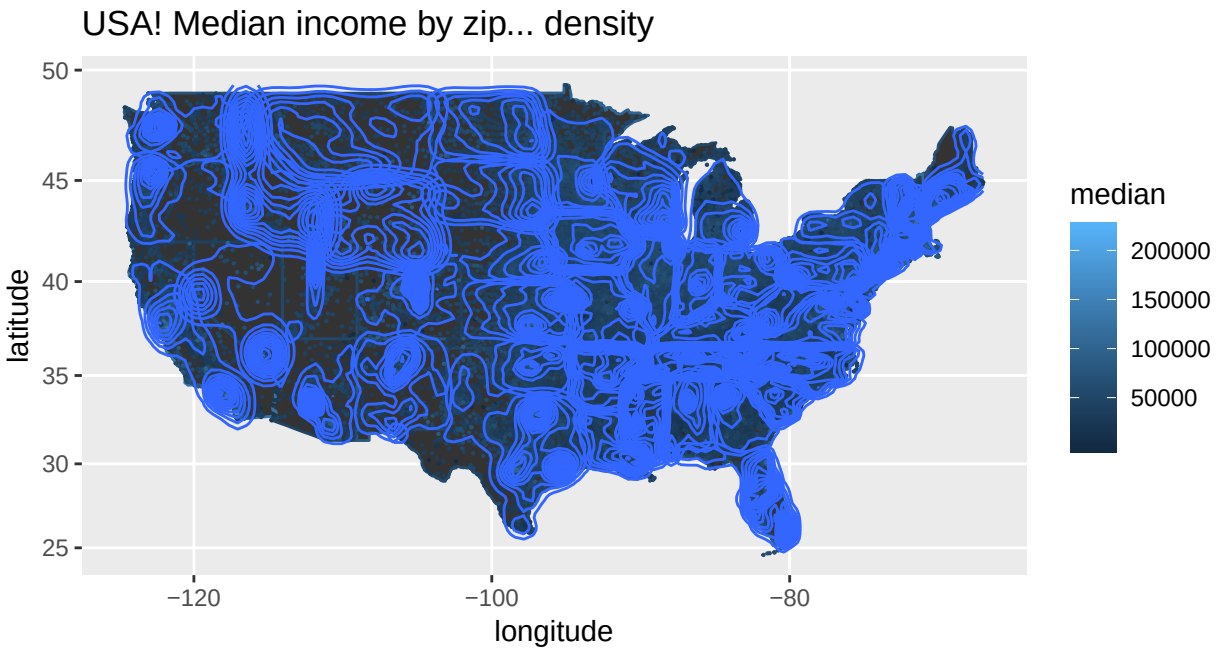

```
# population
ggplot(dataPop, aes(map_id = states)) +
  geom_map(map = us, aes(fill = population), color = "black") +
  expand_limits(x = us$long, y = us$lat) +
  coord_map() + ggtitle("USA! Population")
```



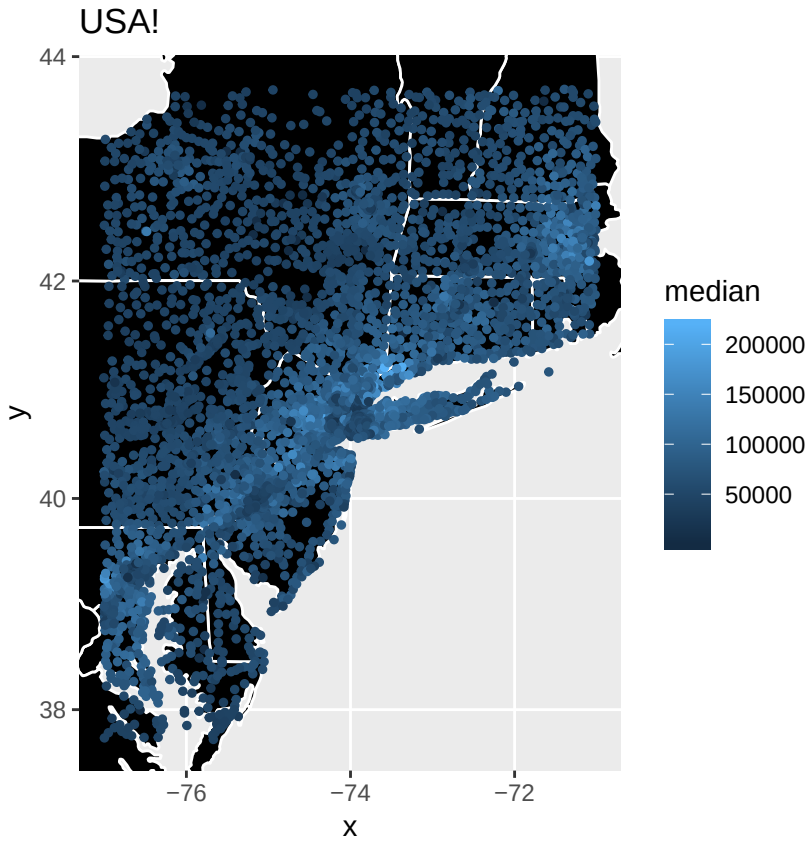

```
zipMedian = stateData[ , c(1,4,5,6, 9)] # filter to plot them zips

# themap + geom_point(data = zipMedian, aes(x = longitude, y = latitude, color = median))

# median by zip
ggplot(zipMedian, aes(map_id = stateName)) +
  geom_map(map = us) +
  coord_map() + ggtitle("USA! Median income by zip") +
  geom_point(aes(x = longitude, y = latitude ,color = median), size = .1)
```



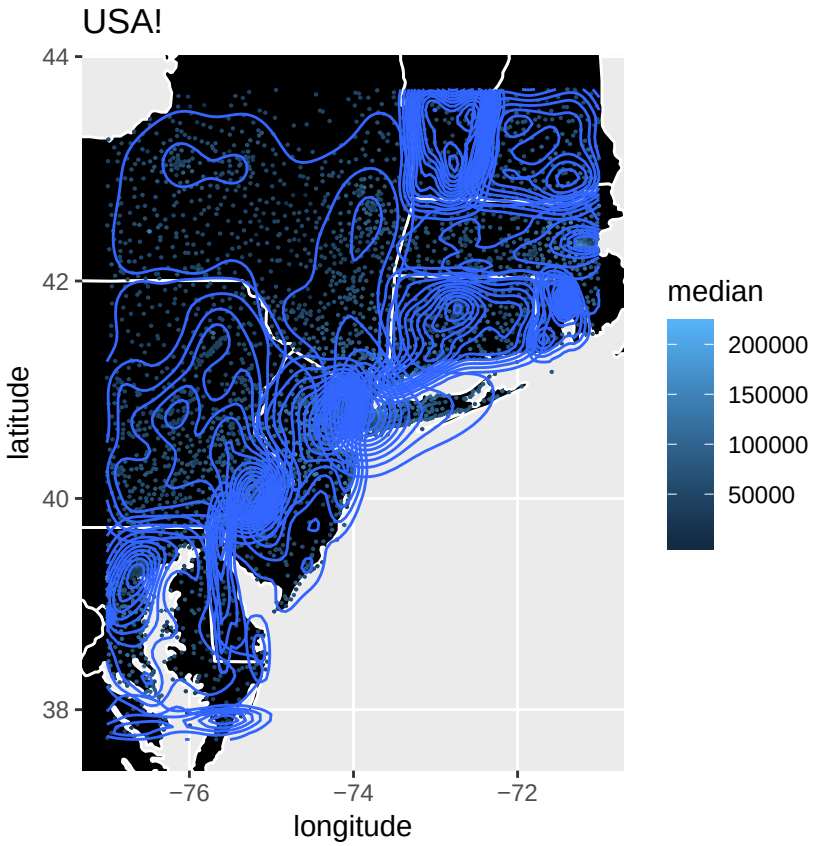

```
# median by zip
ggplot(zipMedian, aes(map_id = stateName, x = longitude, y = latitude, color = median)) +
  geom_map(map = us) +
  coord_map() + ggtitle("USA! Median income by zip... density") +
  geom_point(size = .1) +
  stat_density_2d()
```




```
# center NY
lon = -74.00594
lat = 40.71278
zoom = 3

# map limts
xlimit = c(lon - zoom, lon + zoom)
ylimit = c(lat -zoom, lat + zoom)

# filter data only for NY region
nyData = zipMedian
nyData = nyData[nyData$longitude < xlimit[2], ]
nyData = nyData[nyData$longitude > xlimit[1], ]
nyData = nyData[nyData$latitude > ylimit[1], ]
nyData = nyData[nyData$latitude < ylimit[2], ]

# NY data by zip
ggplot(nyData, aes(map_id = stateName)) +
  geom_map(map = us, fill = "black",  color = "white") +
  coord_map() + ggtitle("USA!") + expand_limits(x = xlimit, y = ylimit) +
  geom_point(aes(x = longitude, y = latitude ,color = median), size = 1)
```




```
# NY data by zip... now with density
ggplot(nyData, aes(map_id = stateName, x = longitude, y = latitude ,color = median)) +
  geom_map(map = us, fill = "black",  color = "white") +
  coord_map() + ggtitle("USA!") + expand_limits(x = xlimit, y = ylimit) +
  geom_point(size = .1) +
  stat_density_2d()
```


USA!
44
42
40
38
latitude
−76
−74
−72
longitude
median
200000
150000
100000
50000