Bing-Je_Wu_HW7

Bing-Je Wu

5/18/2019

Step 1: Load the Data

1. Read the data – using the gdata package we have previously used.

```
library("gdata")
income_df <- read.xls("MedianZIP_2_2.xlsx")
```

2. Clean up the dataframe

a. Remove any info at the front of the file that's not needed

```
income_df<- income_df[-1,]
```

b. Update the column names (zip, median, mean, population)

```
colnames(income_df) <- c("zip", "median", "mean", "population")
```

use str_pad() to fill the 4 digits zipcode into 5 digit zipcode

```
library(stringr)
income_df$zip <- str_pad(string=income_df$zip, width = 5,side = "left", pad = "0")
```

convert data type and format for future anaylsis

```
income_df$zip <- as.character(income_df$zip)
income_df$median <- gsub("\\,","",income_df$median)
income_df$median <- as.numeric(income_df$median)
income_df$mean <- gsub("\\,","",income_df$mean)
income_df$mean <- as.numeric(income_df$mean)
income_df$population <- gsub("\\,","",income_df$population)
income_df$population <- as.numeric(income_df$population)
```

3. Load the 'zipcode' package

4. Merge the zip code information from the two data frames (merge into one dataframe)

```
names(income_df)
```

```
## [1] "zip"        "median"     "mean"       "population"
```

```
names(zipcode)
```

```
## [1] "zip"       "city"      "state"     "latitude"  "longitude"
```

```
income_merge <- merge(x=income_df,y=zipcode, by.x ="zip", by.y ="zip")
```

5. Remove Hawaii and Alaska (just focus on the ‘lower 48’ states)

```
income_merge<-income_merge[income_merge$state != "AK" & income_merge$state != "HI",]
# remove DC from the dataset
income_merge<-income_merge[which(income_merge$state !="DC"),]
```

Step 2: Show the income & population per state

1. Create a simpler dataframe, with just the average median income and the the population for each state.

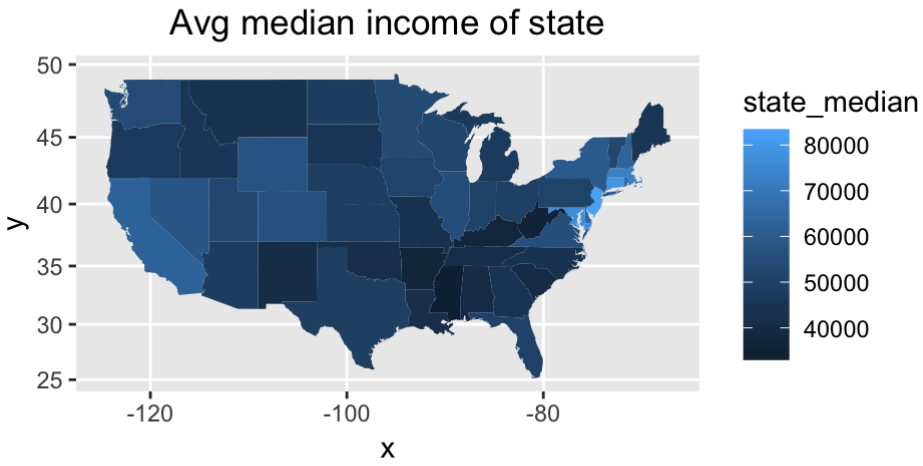
```
state_median<- tapply(income_merge$median, INDEX = income_merge$state, FUN=mean)
state_pop<- tapply(income_merge$population, INDEX = income_merge$state, FUN=sum)
simple_df <- data.frame(state_median, state_pop)
```

2. Add the state abbreviations and the state names as new columns (make sure the state names are all lower case)

```
simple_df$state <- rownames(simple_df)
simple_df$stateName <-state.name[match(simple_df$state,state.abb)]
simple_df$stateName <- tolower(simple_df$stateName)
```

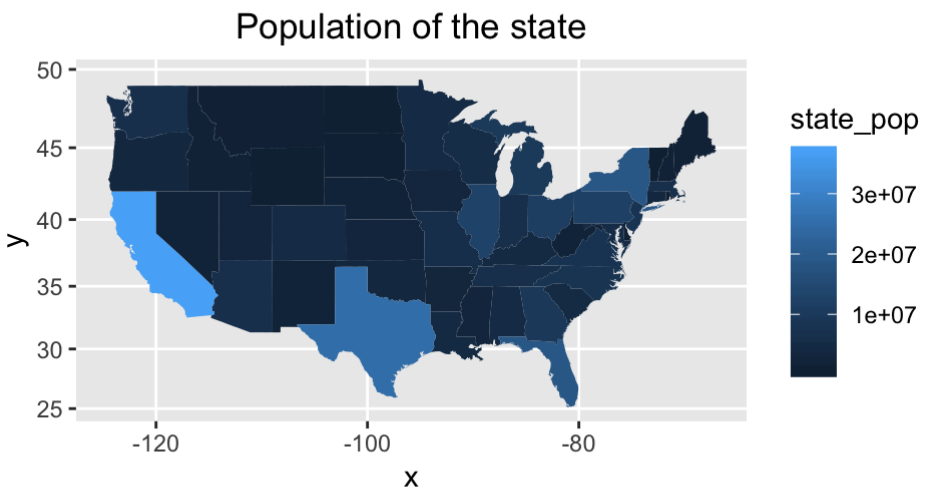
3. Show the U.S. map, representing the color with the average median income of that state

```
library(ggplot2)
library(maps)
library(mapproj)
us <- map_data("state")
ggplot(simple_df, aes(map_id=stateName)) +
  geom_map(map=us, aes(fill=state_median)) +
  expand_limits(x=us$long, y=us$lat) + coord_map() +
  ggtitle("Avg median income of state") +
  theme(plot.title = element_text(hjust = 0.5))
```



4. Create a second map with color representing the population of the state

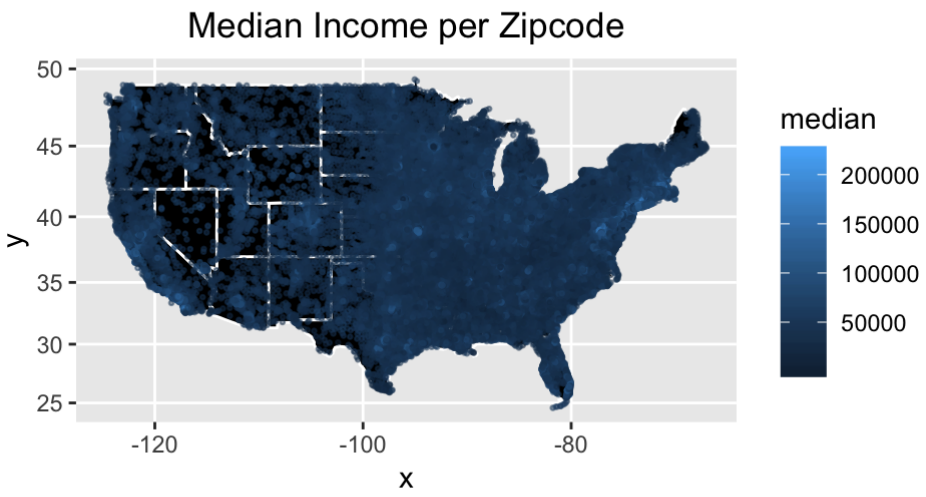
```
library(ggplot2)
library(maps)
library(mapproj)
us <- map_data("state")
ggplot(simple_df, aes(map_id=stateName)) +
  geom_map(map=us, aes(fill=state_pop)) +
  expand_limits(x=us$long, y=us$lat) + coord_map() +
  ggtitle("Population of the state") +
  theme(plot.title = element_text(hjust = 0.5))
```



Step 3: Show the income per zip code

1. Have draw each zip code on the map, where the color of the 'dot' is based on the median income. To make the map look appealing, have the background of the map be black.

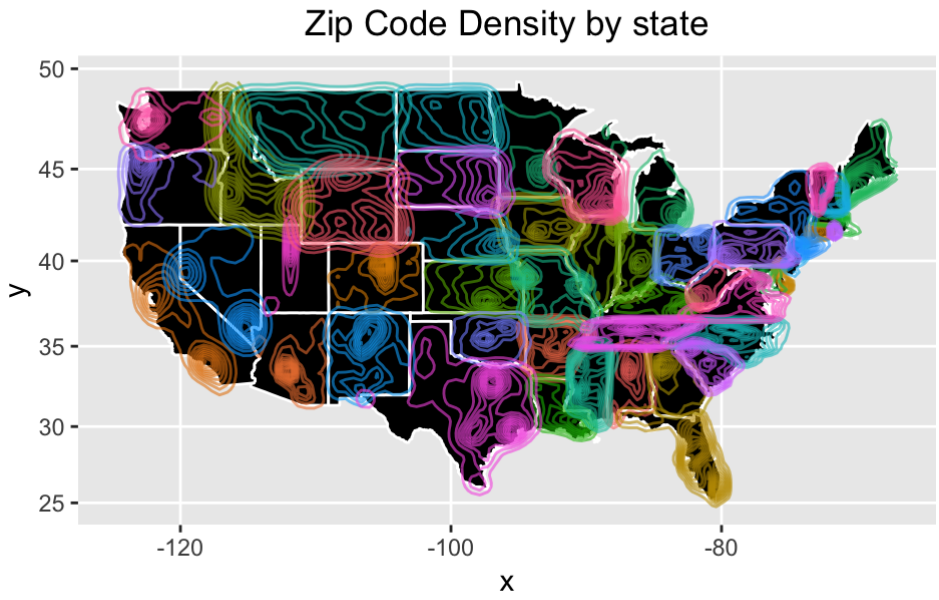
```
library(openintro)
# format state name into income_merge dataset
income_merge$stateName <- abbr2state(income_merge$state)
income_merge$stateName <- tolower(income_merge$stateName)
# create map
ggplot(income_merge, aes(map_id=stateName)) +
  geom_map(map=us, aes(),fill="black", color="white") +
  expand_limits(x=us$long, y=us$lat) +
  coord_map() +
  geom_point(aes(x=longitude, y= latitude, color=median), alpha=0.5, size=0.5) +
  ggtitle("Median Income per Zipcode") +
  theme(plot.title = element_text(hjust = 0.5))
```



Step 4: Show Zip Code Density

1. Now generate a different map, one where we can easily see where there are lots of zip codes, and where there are few (using the 'stat_density2d' function).

```
ggplot(income_merge, aes(map_id=stateName)) +
  geom_map(map=us, fill="black", color="white") +
  expand_limits(x=us$long, y=us$lat) +
  coord_map() +
  stat_density2d(aes(x=longitude, y= latitude,color=income_merge$state), alpha=0.6) +
  ggtitle("Zip Code Density by state") +
  theme(plot.title=element_text(hjust = 0.5), legend.position = "none")
```

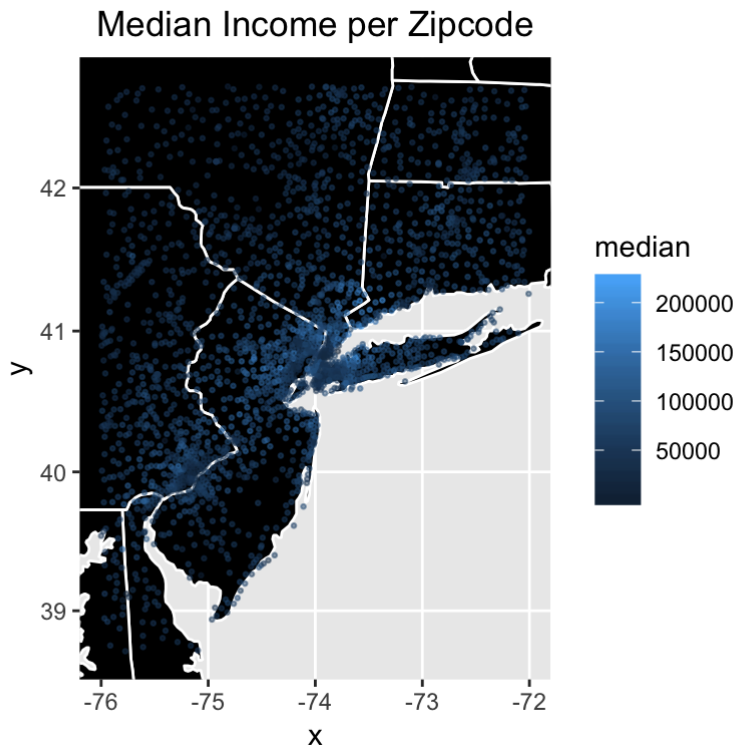


Step 5: Zoom in to the region around NYC

1. Repeat steps 3 & 4, but have the image / map be of the northeast U.S. (centered around New York).

Zoom in the region arounf NYC for step 3

```
ggplot(income_merge, aes(map_id=stateName)) +
  geom_map(map=us, aes(),fill="black", color="white") +
  expand_limits(x=us$long, y=us$lat) +
  coord_map() +
  geom_point(aes(x=longitude, y= latitude, color=median), alpha=0.5, size=0.5) +
  ggtitle("Median Income per Zipcode") +
  theme(plot.title = element_text(hjust = 0.5)) +
  xlim(-76.0,-72.0) + ylim(38.7,42.7)
```



Zoom in the region arounf NYC for step 4

```
ggplot(income_merge, aes(map_id=stateName)) +
  geom_map(map=us, fill="black", color="white") +
  expand_limits(x=us$long, y=us$lat) +
  coord_map() + stat_density2d(aes(x=longitude, y= latitude), alpha=0.6) +
  ggtitle("Zip Code Density") +
  theme(plot.title=element_text(hjust = 0.5)) +
  xlim(-76.0,-72.0) + ylim(38.7,42.7)
```