Social Security Administration: OASDI Beneficiaries by State and ZIP Code, 2015

CUNY MSDA - DATA607 - Project 2_c

Completed by: Duubar Villalobos Jimenez mydvtech@gmail.com

March 12, 2017

Figure 1:

The goal of this assignment is to give you practice in preparing different datasets for downstream analysis work.

Your task is to:

(1) Choose any **three** of the **"wide" datasets** identified in the Week 5 Discussion items. (You may use your own dataset; please don't use my Sample Post dataset, since that was used in your Week 5 assignment!)

For each of the three chosen datasets:

- Create a **.CSV** file (or optionally, a **MySQL** database!) that includes all of the information included in the dataset. You're encouraged to use a "wide" structure similar to how the information appears in the discussion item, so that you can practice tidying and transformations as described below.
- Read the information from your **.CSV** file into **R**, and use **tidyr** and **dplyr** as needed to tidy and transform your data. [Most of your grade will be based on this step!]
- Perform the analysis requested in the discussion item.
- Your code should be in an R Markdown file, posted to rpubs.com, and should include narrative descriptions of your data cleanup work, analysis, and conclusions.

(2) Please include in your homework submission, for each of the three chosen datasets:

The **URL** to the **.Rmd** file in your **GitHub** repository, and The URL for your **rpubs.com** web page.

PROCEDURE

Library definitions

```
library(knitr)
library(gdata)
library(stringr)
library(tidyr)
library(dplyr)
library(zipcode)
library(ggplot2)
```

OASDI NY State Only.

Dataset url location:

url: https://www.ssa.gov/policy/docs/statcomps/

I will be exploring the OASDI Beneficiaries by State and ZIP Code, 2015.

This annual publication focuses on the Social Security beneficiary population at the ZIP Code level. It presents basic program data on the number and type of beneficiaries and the amount of benefits paid in each state, Social Security Administration field office, and ZIP Code. It also shows the number of beneficiaries aged 65 or older.

This annual publication focuses on the Social Security beneficiary population-people receiving Old-Age, Survivors, and Disability Insurance (OASDI) benefits-at the ZIP Code level. It presents basic program data on the number and type of beneficiaries and the amount of benefits paid in each state, Social Security Administration field office, and ZIP Code. It also shows the number of men and women aged 65 or older receiving benefits. The data include only persons whose benefits are currently payable. Those whose benefits were withheld are excluded.

Cherice Jefferies in the Office of Statistical Analysis and Support programmed and compiled the data for this report. Staff of the Office of Information Resources edited the report and prepared it for web publication.

This is a complete Dataset from the federal government website managed by the Social Security Administration.

Last Updated:

This is a complete set of all data for 2015.

Date: October 2016.

Data Provided by:

Office of Retirement and Disability Policy. Office of Research, Evaluation, and Statistics.

Dataset Owner:

Social Security Administration (SSA) and the Government of the United States of America.

Dictionary

This dataset does not seem to have a dictionary. The download link is for a **.xlsx** file containing the desired data. However there's a **.pdf** file describing the data.

Filename: oasdi_zip15.xlsx

For simplicity reasons, I will read the raw data directly from the source.

URL and Raw data name and location definitions:

```
url <- "https://www.ssa.gov/policy/docs/statcomps/oasdi_zip/2015/"
xlsxfile <- "oasdi_zip15.xlsx"
rm(xlsxfile)
```

I tried loading the data from the original location as raw as possible from the **.XLSX** file but found several problems trying to read it. I ended up unmerging the cells by opening the excel file for the State of **New York** sheet and saving as a **.csv** file.

For reproducibility purposes I have uploaded the untoched **.csv** file onto my **GitHub** repository.

```
url <- "https://raw.githubusercontent.com/dvillalobos/MSDA/master/607/Projects/Project2/"
csvfile <- "oasdi_zip15.csv"
```

Function to download .csv file, and extract information from it

```
downloadCSV <- function(myurl, mycsvfile){
  myurl <- paste(myurl,mycsvfile, sep="")
  my.data <- read.csv(myurl, header=FALSE, stringsAsFactors =FALSE )
  head(my.data)
  return(my.data)
}
```

Imported file structure display

```
my.data <- downloadCSV(myurl= url, mycsvfile= csvfile)
```

```
kable(head(my.data))
```

V1
New York
Number of beneficiaries with benefits in current-payment status and total monthly benefits, by field office and ZIP Code, D
Field office and ZIP Code
(thousands of dollars) Number of OASDI beneficiaries aged 65 or older
All areas a

In summary, this data needs to be cleaned up.

Data transformation

Now that I have the data frame I will transform it in order to create some possible outcomes from the given information; for this, I will start by excluding small portion of it.

Excluding Information:

Excluding top and bottom unwanted Rows:

This procedure will exclude the unwanted information contained in the first six rows, then I will exclude the information contained in at the bottom of the file 1995 to 2002 becoming from 1990 to the end in the new data frame.

```
my.new.data <- my.data[-c(1:6), ]
my.new.data <-my.new.data[-c(1989:3003), ]
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
7	Albany			54,020	37,555	8,015	3,175	1,575	3,700	71,518	54,461	4,301	39,255
8		12007		50	40	10	0	0	0	74	58	0	40
9		12009		1,520	1,145	175	75	65	60	2,095	1,669	105	1,175
10		12023		490	340	60	35	20	35	630	470	46	350
11		12024		40	25	10	5	0	0	58	41	7	30
12		12033		1,790	1,390	185	105	45	65	2,458	1,996	142	1,415

Exclude unwanted V3 column:

```
str(my.new.data)
```

```
## 'data.frame':    1988 obs. of  13 variables:
##  $ V1 : chr  "Albany" "" "" "" ...
##  $ V2 : chr  "" "12007" "12009" "12023" ...
##  $ V3 : chr  "" "" "" "" ...
##  $ V4 : chr  "54,020" "50" "1,520" "490" ...
##  $ V5 : chr  "37,555" "40" "1,145" "340" ...
##  $ V6 : chr  "8,015" "10" "175" "60" ...
##  $ V7 : chr  "3,175" "0" "75" "35" ...
##  $ V8 : chr  "1,575" "0" "65" "20" ...
##  $ V9 : chr  "3,700" "0" "60" "35" ...
##  $ V10: chr  "71,518" "74" "2,095" "630" ...
##  $ V11: chr  "54,461" "58" "1,669" "470" ...
##  $ V12: chr  "4,301" "0" "105" "46" ...
##  $ V13: chr  "39,255" "40" "1,175" "350" ...
```

```
my.new.data <- my.new.data %>% subset(select=-c(V3))
str(my.new.data)
```

```
## 'data.frame':    1988 obs. of  12 variables:
##  $ V1 : chr  "Albany" "" "" "" ...
##  $ V2 : chr  "" "12007" "12009" "12023" ...
##  $ V4 : chr  "54,020" "50" "1,520" "490" ...
##  $ V5 : chr  "37,555" "40" "1,145" "340" ...
##  $ V6 : chr  "8,015" "10" "175" "60" ...
##  $ V7 : chr  "3,175" "0" "75" "35" ...
##  $ V8 : chr  "1,575" "0" "65" "20" ...
##  $ V9 : chr  "3,700" "0" "60" "35" ...
##  $ V10: chr  "71,518" "74" "2,095" "630" ...
##  $ V11: chr  "54,461" "58" "1,669" "470" ...
##  $ V12: chr  "4,301" "0" "105" "46" ...
##  $ V13: chr  "39,255" "40" "1,175" "350" ...
```

	V1	V2	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13
7	Albany		54,020	37,555	8,015	3,175	1,575	3,700	71,518	54,461	4,301	39,255
8		12007	50	40	10	0	0	0	74	58	0	40
9		12009	1,520	1,145	175	75	65	60	2,095	1,669	105	1,175
10		12023	490	340	60	35	20	35	630	470	46	350
11		12024	40	25	10	5	0	0	58	41	7	30
12		12033	1,790	1,390	185	105	45	65	2,458	1,996	142	1,415

Renaming Columns

```r
names(my.new.data) <- c("County","Zipcode","n Total", "n Retired", "n Disabled", "n Widow & Parents", "r
```

	County	Zipcode	n Total	n Retired	n Disabled	n Widow & Parents	n Spouses	n Children	$ All Beneficia
7	Albany		54,020	37,555	8,015	3,175	1,575	3,700	71,518
8		12007	50	40	10	0	0	0	74
9		12009	1,520	1,145	175	75	65	60	2,095
10		12023	490	340	60	35	20	35	630
11		12024	40	25	10	5	0	0	58
12		12033	1,790	1,390	185	105	45	65	2,458

Need to split data into 2 data frames

a) Zip Code Table which is going to include:

- Data for Numbers
- Data for Monthly Benefits

b) County Table which is going to include:

- Data for Numbers
- Data for Monthly Benefits

Separate results:

First I will separate County Summary data from zip code data.

```r
# Creating a County Data Frame
my.new.data$County <- str_replace_all(my.new.data$County," ","")
my.county.data <- my.new.data %>% subset(County != "")
rownames(my.county.data) <- NULL
```

County Summary Table.

County	Zipcode	n Total	n Retired	n Disabled	n Widow & Parents	n Spouses	n Chil
Albany		54,020	37,555	8,015	3,175	1,575	3,700
Babylon		36,020	23,460	5,905	2,530	1,380	2,745
Batavia		43,525	29,635	6,980	2,750	1,395	2,765
Binghamton		68,620	45,655	11,070	4,305	2,325	5,265

County	Zipcode	n Total	n Retired	n Disabled	n Widow & Parents	n Spouses	n Chil
Bronx,East		47,275	29,245	8,655	3,260	2,075	4,040
Bronx,HuntsPoint		7,525	3,895	1,810	515	345	960
Bronx,LaconiaAvenue		44,445	29,245	7,825	2,295	1,420	3,660
Bronx,North		30,275	17,235	7,005	1,810	1,315	2,910
Bronx,South		36,405	19,000	8,840	2,495	1,765	4,305
Bronx,WestFarms		15,550	7,930	4,055	1,010	665	1,890
Brooklyn,BedfordHeights		36,995	24,440	6,500	2,040	1,055	2,960
Brooklyn,BoroHall		69,915	46,875	8,960	5,020	4,145	4,915
Brooklyn,Bushwick		29,115	16,940	5,785	2,160	1,650	2,580
Brooklyn,Canarsie		28,570	18,040	5,175	1,645	950	2,760
Brooklyn,CypressHills		54,775	34,110	9,660	3,460	2,880	4,665
Brooklyn,Flatbush		68,195	48,450	8,695	3,575	3,040	4,435
Brooklyn,NewUtrecht		71,155	49,090	9,255	4,680	4,550	3,580
Buffalo		108,520	69,925	18,335	8,005	3,670	8,585
Corning		26,640	17,415	4,600	1,770	925	1,930
Dunkirk		13,310	8,665	2,370	925	455	895
Elmira		28,885	18,585	5,375	1,875	910	2,140
Flushing		77,595	57,980	6,345	5,085	5,150	3,035
Freeport		121,675	86,210	14,520	8,145	5,260	7,540
Geneva		63,290	43,880	9,240	3,475	1,790	4,905
Gloversville		20,940	13,705	3,640	1,280	550	1,765
Hudson		27,125	18,610	4,090	1,690	815	1,920
Ithaca		15,985	11,490	2,050	875	595	975
Jamaica		77,395	55,295	9,960	4,055	3,000	5,085
Jamestown		20,360	13,580	3,330	1,340	645	1,465
LongIslandCity		50,430	35,720	5,940	3,450	3,000	2,320
Melville		77,860	54,895	8,735	5,020	3,845	5,365
Mineola		129,535	95,300	11,205	8,880	7,290	6,860
Monticello		17,775	11,395	3,085	1,070	550	1,675
NewRochelle		46,695	34,215	4,790	2,865	2,220	2,605
NewYorkCity,Downtown		37,075	28,185	3,510	1,605	1,360	2,415
NewYorkCity,EastHarlem		21,930	14,500	3,270	1,600	1,060	1,500
NewYorkCity,EastVillage		25,900	18,110	3,435	1,700	1,460	1,195
NewYorkCity,Midtown		83,785	66,955	5,485	4,445	3,890	3,010
NewYorkCity,Uptown		53,220	35,600	9,335	2,940	1,815	3,530
NewYorkCity,WashingtonHeights		43,770	30,030	6,820	2,430	2,085	2,405
Newburgh		65,100	42,410	10,510	4,060	2,225	5,895
NiagaraFalls		52,475	33,495	9,435	3,910	1,935	3,700
Ogdensburg		25,220	15,275	4,705	2,050	1,180	2,010
Olean		30,505	19,810	5,260	2,040	1,080	2,315
Oneonta		26,535	18,300	3,720	1,675	945	1,895
Oswego		30,345	18,785	5,850	2,060	1,125	2,525
Patchogue		139,600	92,085	21,455	9,265	5,605	11,190
Peekskill		40,585	30,710	3,550	2,385	1,755	2,185
Plattsburgh		37,785	23,570	7,145	2,655	1,380	3,035
Poughkeepsie		118,085	80,280	17,455	7,195	4,130	9,025
Queensbury		45,700	30,190	7,520	3,110	1,585	3,295
RegoPark		73,955	53,815	8,200	4,500	4,450	2,990
RidgeRoad		94,935	63,455	14,050	7,450	3,475	6,505
Riverhead		41,565	31,290	3,835	2,465	1,685	2,290
Rochester,Downtown		87,335	56,480	15,905	4,795	2,870	7,285
Rochester,Greece		70,080	50,400	9,185	4,120	2,335	4,040

County	Zipcode	n Total	n Retired	n Disabled	n Widow & Parents	n Spouses	n Chil
RockawayPark		18,660	11,350	3,665	1,135	680	1,830
Schenectady		82,075	57,055	11,330	5,165	2,825	5,700
StatenIsland		63,730	41,000	10,060	4,455	3,125	5,090
StatenIsland,HylanBlvd		25,050	15,500	4,650	1,615	1,020	2,265
Syracuse		125,495	84,985	19,815	7,855	3,920	8,920
Troy		42,255	28,415	7,070	2,590	1,115	3,065
Utica		80,270	52,280	13,535	5,070	2,445	6,940
Watertown		27,705	17,585	4,615	2,085	1,190	2,230
WestNyack		55,535	39,935	5,535	3,025	2,595	4,445
WhitePlains		39,775	30,350	3,195	2,205	1,950	2,075
Yonkers		39,910	28,285	5,085	2,495	1,740	2,305

Zip Code Table by employing **anti_join()** from **dplyr** function.

```
my.zipcode.data <- anti_join(my.new.data, my.county.data, by="County")
my.zipcode.data <- my.zipcode.data %>% subset(select=-c(County))
```

Zip Code table.

Zipcode	n Total	n Retired	n Disabled	n Widow & Parents	n Spouses	n Children	$ All Beneficiaries	$ Retire
12007	50	40	10	0	0	0	74	58
12009	1,520	1,145	175	75	65	60	2,095	1,669
12023	490	340	60	35	20	35	630	470
12024	40	25	10	5	0	0	58	41
12033	1,790	1,390	185	105	45	65	2,458	1,996
12041	115	90	10	10	0	5	161	126

Data Exploration

From the above table we can explore a few things as follows:

Geographical distribution:

Distribution by Region:

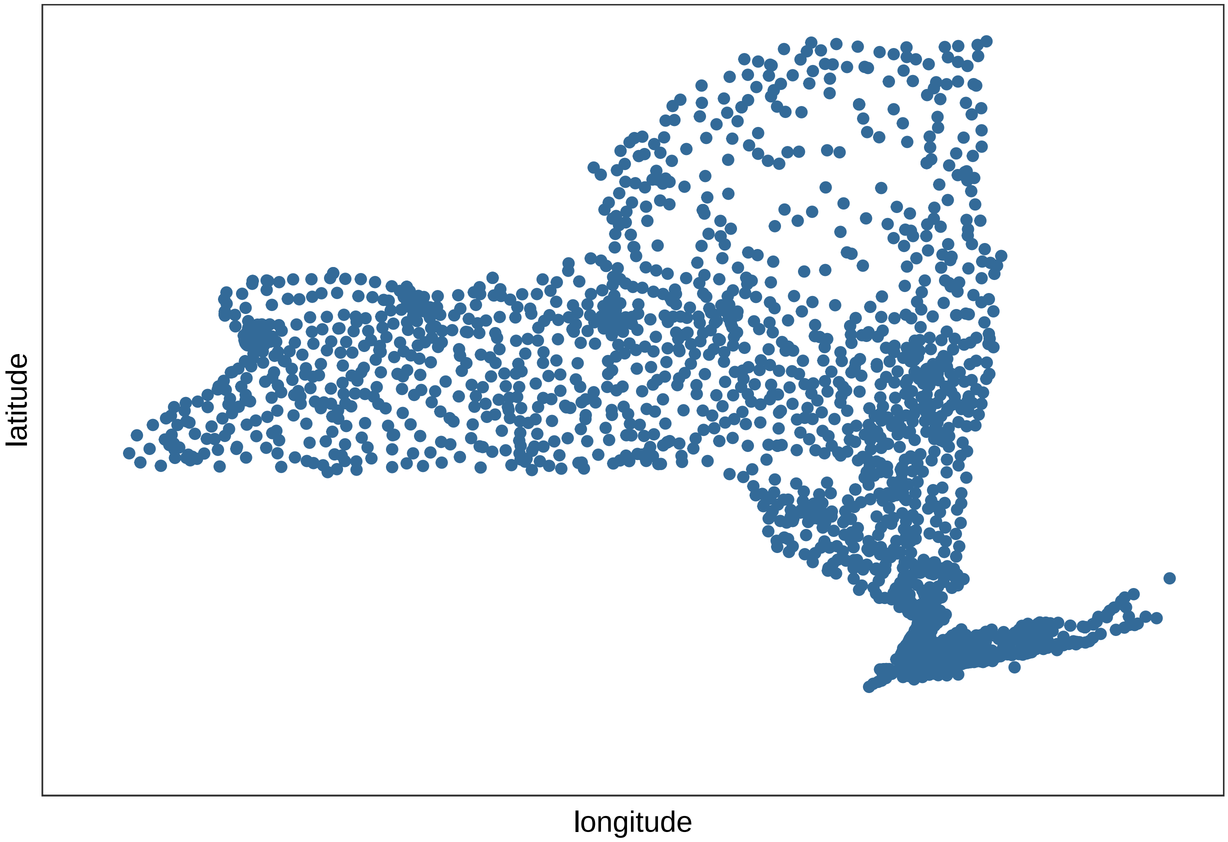
Distribution of OASDI Beneficiaries during 2015 by the zipcode (Region).

```
# Merge Zipcodes with the zipcode library
USzipCodes <- my.zipcode.data
USzipCodes$Zipcode <- clean.zipcodes(USzipCodes$Zipcode)
data(zipcode)
USzipCodes <- merge(USzipCodes, zipcode, by.x='Zipcode', by.y='zip')

# Creating ggplot of matches ZipCodes
g <- ggplot(data=USzipCodes) + geom_point(aes(x=longitude, y=latitude, colour=1))

# simplify display and limit to the "lower 48"
g <- g + theme_bw() + scale_x_continuous(limits = c(-80,-72), breaks = NULL)
g <- g + scale_y_continuous(limits = c(40,45), breaks = NULL)
```

```r
g <- g + theme(legend.position="none")
g
```



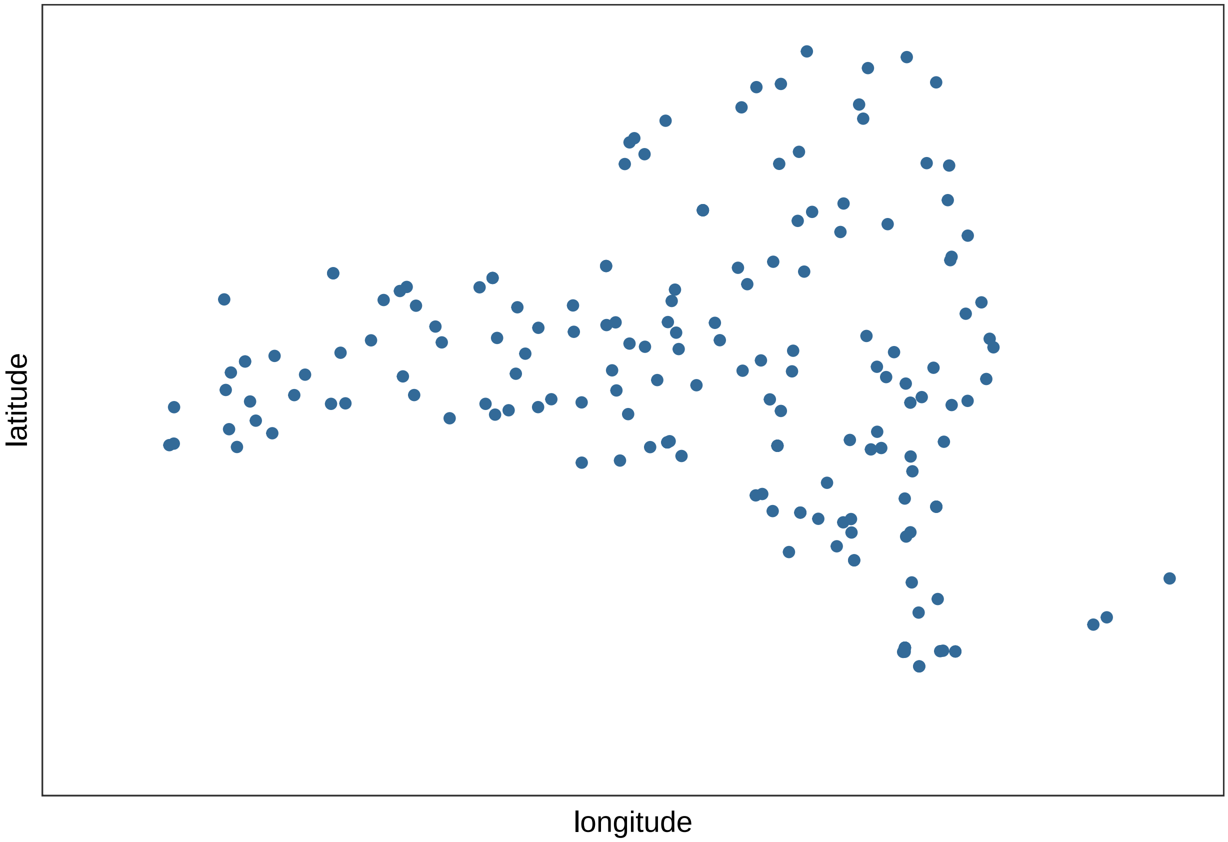
Zip codes in which children have not received benefits.

Zipcode	city	state	latitude	longitude
06390	Fishers Island	NY	41.26194	-72.00708
10020	New York	NY	40.75867	-73.98024
10101	New York	NY	40.78075	-73.97718
10107	New York	NY	40.76643	-73.98273
10123	New York	NY	40.75149	-73.99054
10129	New York	NY	40.78075	-73.97718

Geographical distribution.

```r
# Creating ggplot of matches ZipCodes
g <- ggplot(data=USzipCodesChild) + geom_point(aes(x=longitude, y=latitude, colour=3))

# simplify display and limit to the "lower 48"
g <- g + theme_bw() + scale_x_continuous(limits = c(-80,-72), breaks = NULL)
g <- g + scale_y_continuous(limits = c(40,45), breaks = NULL)
g <- g + theme(legend.position="none")
g
```



Conclusions

This is an interesting analysis and I believe it can play a great role in local discoveries related to OASDI Beneficiaries since it covers immediate surrounding areas.

For example, from the **Region** distribution we can visualize how the distribution is over the respective zipcodes having a better perspective on how respective populations are distributed.

Final conclusion:

Since there's more data available, it will be interesting to perform more comparisons in regards of the years and states and any other "correlation" that we could find related to other entities as well.