AGI & Income Index

Partha Banerjee

Outcome of this work is to find Zip wise "Adjusted gross income (AGI)" and then calculate Income Index which will be used towards medicare data factoring for fraud detection. Source of the data: IRS Website.

Download data file **12zp33ny.xls** from IRS

- Open 12zp33ny.xls and remove header and footer descriptions since they will not be part of our intended data
- Save file as **12zp33ny.csv** as otherwise we need to add some programming complexity to read data directly from .xls file

```
# Function to return number for a factor
nmbr <- function(col) {
  return(suppressWarnings(as.numeric(gsub(",","", as.character(col)))))
}
```

```
fl <- "12zp33ny.csv"
```

```
# Read data file
agi_ny_raw <- read.csv(fl)

# Remove unnecessary columns by keeping ZIP, Number.of.returns,
# Size.of.adjusted.gross.income, and AGI.
# We need Size.of.adjusted.gross.income for filtering the data and will remove this
# column once filtering is over.
# Remove all NA values from data.
agi_ny <- agi_ny_raw[,c(1,2,3,10)]
agi_ny <- agi_ny[complete.cases(agi_ny),]
colnames(agi_ny) <- c('ZIP', 'X', 'Tot.Returns', 'Tot.AGI')

# AGI should be Total of AGI's / Total of Returns
agi_ny$AGI <- round(nmbr(agi_ny$Tot.AGI) / nmbr(agi_ny$Tot.Returns), 0)

# Remove ZIP codes 00000 and 99999 as they represent the total of all Zip values and
# nonresidential ZIP/Category code respectively and will not serve any purpose for us.
agi_ny <- agi_ny[ which(agi_ny$ZIP!=0 & agi_ny$ZIP!=99999),]

# Remove Size.of.adjusted.gross.income level data by keeping only ZIP level value.
agi_ny <- agi_ny[agi_ny$X=="",]
agi_ny <- agi_ny[,c(1,3,4,5)]
agi_ny <- agi_ny[order(agi_ny$ZIP),]

# Display data
head(agi_ny)
```

```
##      ZIP Tot.Returns   Tot.AGI AGI
## 10 10001      13,300 1,915,739 144
## 18 10002      43,460 1,944,420  45
## 26 10003      29,360 5,989,687 204
```

```
## 34 10004       2,420   879,963 364
## 42 10005       5,580 5,488,231 984
## 50 10006       2,310   407,119 176
```

```
tail(agi_ny)
```

```
##         ZIP Tot.Returns Tot.AGI AGI
## 12322 14897         390  16,375  42
## 12330 14898         590  23,042  39
## 12338 14901       5,570 193,855  35
## 12346 14903       3,700 212,137  57
## 12354 14904       7,310 255,507  35
## 12362 14905       4,350 279,642  64
```

Now we will find the median value for AGI and use that as standard to calculate the Index for Income. We will use this index to normalize the data.

```
m_agi <- median(agi_ny$AGI)
agi_ny$II <- round(agi_ny$AGI / m_agi, 2)
summary(agi_ny)
```

```
##       ZIP          Tot.Returns        Tot.AGI          AGI
##  Min.   :10001   120    :  13   5,094    :   2   Min.   : 21.0
##  1st Qu.:11749   160    :  12   73,153   :   2   1st Qu.: 42.0
##  Median :12808   270    :  11   1,002,305:   1   Median : 49.0
##  Mean   :12723   300    :  11   1,002,789:   1   Mean   : 68.9
##  3rd Qu.:13786   110    :  10   1,004,328:   1   3rd Qu.: 66.0
##  Max.   :14905   650    :  10   1,010,940:   1   Max.   :984.0
##                  (Other):1478   (Other)  :1537
##        II
##  Min.   : 0.43
##  1st Qu.: 0.86
##  Median : 1.00
##  Mean   : 1.41
##  3rd Qu.: 1.35
##  Max.   :20.08
##
```

```
# Display data
head(agi_ny)
```

```
##      ZIP Tot.Returns   Tot.AGI AGI    II
## 10 10001      13,300 1,915,739 144  2.94
## 18 10002      43,460 1,944,420  45  0.92
## 26 10003      29,360 5,989,687 204  4.16
## 34 10004       2,420   879,963 364  7.43
## 42 10005       5,580 5,488,231 984 20.08
## 50 10006       2,310   407,119 176  3.59
```

```
tail(agi_ny)
```

```
##         ZIP Tot.Returns Tot.AGI AGI   II
## 12322 14897         390  16,375  42 0.86
## 12330 14898         590  23,042  39 0.80
## 12338 14901       5,570 193,855  35 0.71
## 12346 14903       3,700 212,137  57 1.16
## 12354 14904       7,310 255,507  35 0.71
## 12362 14905       4,350 279,642  64 1.31
```