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)))))
}
```

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

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

Clean Data

```
# 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 ZIP wise total data line by keeping category level values
agi_ny <- agi_ny[agi_ny$X!="",]

# Remove 0 Tot_Returns from data set
agi_ny <- agi_ny[agi_ny$Tot_Returns!="**",]
```

Process Data

Here we will do the following:

- Calculate ZIP wise total number of returns

- Calculate percentage of category population within the ZIP
- Adjust AGI based upon population representation for each category within the ZIP
- Aggregate this adjusted AGI on ZIP

```
# Add Zip wise total as a separate column and adjust total returns based upon population %
agi_ny$ZipTotReturns <- ave(nmbr(agi_ny$Tot_Returns), agi_ny$ZIP, FUN=sum)
agi_ny$PopulationPC <- round(nmbr(agi_ny$Tot_Returns)/agi_ny$ZipTotReturns,4)

# Adjust AGI based upon their weight
agi_ny$Adj_AGI <- round(nmbr(agi_ny$Tot_AGI)*agi_ny$PopulationPC,2)

# Now calculate ZIP level weighted average AGI
suppressMessages(library(sqldf))
sql <- "select ZIP, ZipTotReturns as TotReturns, sum(adj_agi) as Adj_AGI"
sql <- paste(sql, ", round(sum(adj_agi)/ZipTotReturns,2) as Avg_AGI")
sql <- paste(sql, "from agi_ny")
sql <- paste(sql, "group by ZIP, ZipTotReturns")
agi_ny <- suppressMessages(sqldf(sql))

# Display data
head(agi_ny)
```

```
##     ZIP TotReturns Adj_AGI Avg_AGI
## 1 10001      13300  258283   19.42
## 2 10002      43460  288646    6.64
## 3 10003      29360 1039011   35.39
## 4 10004       2420  264348  109.23
## 5 10005       5580 1408852  252.48
## 6 10006       2310   78374   33.93
```

```
tail(agi_ny)
```

```
##        ZIP TotReturns Adj_AGI Avg_AGI
## 1540 14897        390    2856    7.32
## 1541 14898        590    4445    7.53
## 1542 14901       5570   37661    6.76
## 1543 14903       3700   28350    7.66
## 1544 14904       7310   53577    7.33
## 1545 14905       4350   36641    8.42
```

Now we will take the natural log for the Avg_AGI column and use that as the Index for Income. We will use this index to normalize the data.

```
agi_ny$Indx <- round(log(agi_ny$Avg_AGI), 2)
summary(agi_ny)
```

```
##       ZIP          TotReturns       Adj_AGI          Avg_AGI
##  Min.   :10001   Min.   :   90   Min.   :   1141   Min.   :  6.3
##  1st Qu.:11749   1st Qu.:  680   1st Qu.:   5721   1st Qu.:  7.7
##  Median :12808   Median : 2020   Median :  18128   Median :  8.3
##  Mean   :12723   Mean   : 5914   Mean   :  82842   Mean   : 13.0
```

```
##  3rd Qu.:13786   3rd Qu.: 7470   3rd Qu.:  83503   3rd Qu.: 10.4
##  Max.   :14905   Max.   :50490   Max.   :3116244   Max.   :322.2
##       Indx
##  Min.   :1.83
##  1st Qu.:2.04
##  Median :2.11
##  Mean   :2.30
##  3rd Qu.:2.34
##  Max.   :5.78
```

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

```
##     ZIP TotReturns Adj_AGI Avg_AGI Indx
## 1 10001      13300  258283   19.42 2.97
## 2 10002      43460  288646    6.64 1.89
## 3 10003      29360 1039011   35.39 3.57
## 4 10004       2420  264348  109.23 4.69
## 5 10005       5580 1408852  252.48 5.53
## 6 10006       2310   78374   33.93 3.52
```

```
tail(agi_ny)
```

```
##        ZIP TotReturns Adj_AGI Avg_AGI Indx
## 1540 14897        390    2856    7.32 1.99
## 1541 14898        590    4445    7.53 2.02
## 1542 14901       5570   37661    6.76 1.91
## 1543 14903       3700   28350    7.66 2.04
## 1544 14904       7310   53577    7.33 1.99
## 1545 14905       4350   36641    8.42 2.13
```

Writing AGI Index data to CSV file

```
outfl <- "agi.csv"
write.csv(agi_ny, outfl, row.names=FALSE)
```