Module 4 Assignment 1: Data Transformation

Tabitha Hagen

2022-07-28

# Module 4 - Assignment 1

## Hagen, Tabitha

### Data Transformation

# R Setup

knitr::opts\_chunk$set(echo = TRUE)  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.7 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

State\_Income <- read\_csv("state\_income.csv") # copy original file

## Warning: One or more parsing issues, see `problems()` for details

## Rows: 32526 Columns: 18

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (7): State\_Name, State\_ab, County, City, Place, Type, Primary  
## dbl (11): id, State\_Code, Zip\_Code, Area\_Code, ALand, AWater, Lat, Lon, Mean...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Part 1: Exploring the data and using select (), rename () and filter ()

#### 1A Creating and Ordering a Subset of the data

#### State Incomes  
  
glimpse (State\_Income) #Look at the copied file

## Rows: 32,526  
## Columns: 18  
## $ id <dbl> 1011000, 1011010, 1011020, 1011030, 1011040, 1011050, 10110…  
## $ State\_Code <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…  
## $ State\_Name <chr> "Alabama", "Alabama", "Alabama", "Alabama", "Alabama", "Ala…  
## $ State\_ab <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL",…  
## $ County <chr> "Mobile County", "Barbour County", "Shelby County", "Mobile…  
## $ City <chr> "Chickasaw", "Louisville", "Columbiana", "Satsuma", "Dauphi…  
## $ Place <chr> "Chickasaw city", "Clio city", "Columbiana city", "Creola c…  
## $ Type <chr> "City", "City", "City", "City", "Town", "Town", "City", "To…  
## $ Primary <chr> "place", "place", "place", "place", "place", "place", "plac…  
## $ Zip\_Code <dbl> 36611, 36048, 35051, 36572, 36528, 35057, 36426, 36020, 356…  
## $ Area\_Code <dbl> 251, 334, 205, 251, 251, 256, 251, 334, 256, 256, 256, 256,…  
## $ ALand <dbl> 10894952, 26070325, 44835274, 36878729, 16204185, 8913021, …  
## $ AWater <dbl> 909156, 23254, 261034, 2374530, 413605152, 26837, 91015, 17…  
## $ Lat <dbl> 30.77145, 31.70852, 33.19145, 30.87434, 30.25091, 34.04541,…  
## $ Lon <dbl> -88.07970, -85.61104, -86.61562, -88.00944, -88.17127, -86.…  
## $ Mean <dbl> 38773, 37725, 54606, 63919, 77948, 50715, 33737, 46319, 579…  
## $ Median <dbl> 30506, 19528, 31930, 52814, 67225, 42643, 23610, 40242, 395…  
## $ Stdev <dbl> 33101, 43789, 57348, 47707, 54270, 35886, 28256, 38941, 472…

state\_income2 <- select(State\_Income, State\_Name, State\_ab, County, City, Type, ALand, Mean, Median, Stdev) #create a new dataset   
  
state\_income2 <- select(state\_income2, State\_ab,everything()) # rearrange the column order within the dataset  
  
head(state\_income2, n=10) # Look at the 1st 10 rows in the table

## # A tibble: 10 × 9  
## State\_ab State\_Name County City Type ALand Mean Median Stdev  
## <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 AL Alabama Mobile County Chickas… City 1.09e7 38773 30506 33101  
## 2 AL Alabama Barbour County Louisvi… City 2.61e7 37725 19528 43789  
## 3 AL Alabama Shelby County Columbi… City 4.48e7 54606 31930 57348  
## 4 AL Alabama Mobile County Satsuma City 3.69e7 63919 52814 47707  
## 5 AL Alabama Mobile County Dauphin… Town 1.62e7 77948 67225 54270  
## 6 AL Alabama Cullman County Cullman Town 8.91e6 50715 42643 35886  
## 7 AL Alabama Escambia County East Br… City 8.83e6 33737 23610 28256  
## 8 AL Alabama Elmore County Coosada Town 1.02e7 46319 40242 38941  
## 9 AL Alabama Morgan County Eva Town 1.05e7 57994 39591 47235  
## 10 AL Alabama Talladega County Sylacau… CDP 4.52e7 54807 41712 51359

#### 1B Renaming Columns/Variablesin the Subset

state\_income2 <- rename(state\_income2, SquareArea = ALand, IncomeMean = Mean, IncomeMedian = Median, IncomeStDev = Stdev) # making names more user\_friendly  
  
head(state\_income2, n=10) # Look at the 1st 10 rows in the table

## # A tibble: 10 × 9  
## State\_ab State\_Name County City Type SquareArea IncomeMean IncomeMedian  
## <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 AL Alabama Mobile Co… Chic… City 10894952 38773 30506  
## 2 AL Alabama Barbour C… Loui… City 26070325 37725 19528  
## 3 AL Alabama Shelby Co… Colu… City 44835274 54606 31930  
## 4 AL Alabama Mobile Co… Sats… City 36878729 63919 52814  
## 5 AL Alabama Mobile Co… Daup… Town 16204185 77948 67225  
## 6 AL Alabama Cullman C… Cull… Town 8913021 50715 42643  
## 7 AL Alabama Escambia … East… City 8826252 33737 23610  
## 8 AL Alabama Elmore Co… Coos… Town 10222339 46319 40242  
## 9 AL Alabama Morgan Co… Eva Town 10544874 57994 39591  
## 10 AL Alabama Talladega… Syla… CDP 45178321 54807 41712  
## # … with 1 more variable: IncomeStDev <dbl>

#### 1C Create another Subset with only NC values

NC\_income <- filter(state\_income2,State\_ab == "NC" | State\_Name == "North Carolina") # filter by NC info  
  
head(NC\_income, n=10) # Look at the 1st 10 rows in the table

## # A tibble: 10 × 9  
## State\_ab State\_Name County City Type SquareArea IncomeMean IncomeMedian  
## <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 NC North Carolina Alama… Elon CDP 3515396 89973 300000  
## 2 NC North Carolina Johns… Wend… Town 23956770 67438 300000  
## 3 NC North Carolina Samps… Sted… Town 1353212 43538 25196  
## 4 NC North Carolina Hende… Hend… CDP 2625120 38120 31430  
## 5 NC North Carolina Beauf… Pine… Town 4121722 30468 17951  
## 6 NC North Carolina Davie… Clem… Town 5903422 97561 80720  
## 7 NC North Carolina Blade… Blad… Town 5737410 38588 20838  
## 8 NC North Carolina Samps… Clin… CDP 8562785 34778 23603  
## 9 NC North Carolina Lee C… Broa… Town 3350431 60384 52298  
## 10 NC North Carolina Guilf… Burl… City 75533002 54337 300000  
## # … with 1 more variable: IncomeStDev <dbl>

# Part 2: NC Incomes – arrange () and summarize ()

#### NC Incomes

First we created a copy of the original data set “state\_income” which had 32,526 observations or rows with 18 variables. Then we created a subset with only 9 of those columns or variables of information. We even renamed some of those columns to a more user-friendly name. Lastly, we took a subset of that table which contains only NC information. Now, that NC\_income dataset will be used to create summaries of the incomes within North Carolina utilizing the variables or columns named County, City, and Type.

#### Part 2A: NC Summaries by County and City (using pipes)

NC\_income <- arrange(NC\_income, County) # arrange the 915 observation table by County - only 49 counties  
head(NC\_income, n=10) # Look at the 1st 10 rows in the table

## # A tibble: 10 × 9  
## State\_ab State\_Name County City Type SquareArea IncomeMean IncomeMedian  
## <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 NC North Carolina Alama… Elon CDP 3515396 89973 300000  
## 2 NC North Carolina Alama… Meba… City 23213152 67397 55632  
## 3 NC North Carolina Alama… Hend… Track 12734435 57073 41022  
## 4 NC North Carolina Alama… Ahos… Track 199246026 54071 42038  
## 5 NC North Carolina Alama… Red … Track 93319263 30673 20786  
## 6 NC North Carolina Alama… Stat… Track 10829691 40174 27569  
## 7 NC North Carolina Alama… Supp… Track 29875162 45625 32324  
## 8 NC North Carolina Alama… Stat… Track 37718022 55177 48504  
## 9 NC North Carolina Alama… Moor… Track 13853696 106274 83085  
## 10 NC North Carolina Alama… Moor… Track 7037037 93463 79991  
## # … with 1 more variable: IncomeStDev <dbl>

summary1 <- group\_by(NC\_income, County) # creating a new dataset explicitly by County  
summary1 <- summarise (summary1, mean = mean(IncomeMean)) # adding Mean Income to the dataset  
head(summary1, n=10) # Look at the 1st 10 rows in the table

## # A tibble: 10 × 2  
## County mean  
## <chr> <dbl>  
## 1 Alamance County 58430.  
## 2 Anson County 36559.  
## 3 Avery County 41915   
## 4 Beaufort County 40029.  
## 5 Bladen County 35796   
## 6 Brunswick County 53794.  
## 7 Burke County 50283.  
## 8 Caldwell County 50050   
## 9 Camden County 53950   
## 10 Carteret County 50700

summary2 <- NC\_income %>% # Specifying the dataFRAME that will be used for all commands after this pipe  
 group\_by(City) %>% # Grouping info by City, which lists 335 cities  
 summarise (mean = mean(IncomeMean)) # adding Mean Income to the dataset  
head(summary2, n=10) # Look at the 1st 10 rows in the table

## # A tibble: 10 × 2  
## City mean  
## <chr> <dbl>  
## 1 Aberdeen 71839   
## 2 Advance 96650   
## 3 Ahoskie 51360   
## 4 Albemarle 61028.  
## 5 Andrews 39213   
## 6 Angier 58414.  
## 7 Apex 113786.  
## 8 Arapahoe 58334   
## 9 Arden 85641   
## 10 Ash 50636

In Summary 1, above, the dataset tables lists 49 Counties and their respective Income Means. Most of the original 915 observations from NC were based in Alamance County. NC has 100 counties but only 49 are represented by the original NC\_income dataset.

In Summary 2, above, the data is sorted by Cities. There are 335 cities listed. Again, they are still mostly in Alamance County, to the east of the economic Triad region and northwest of the Research Triangle economic region. However, there is more information to analyze in this dataset compared to Summary 1.

In the final Summary 3 dataset, 5 Types of Community are listed as descibed by the U.S. Census Bureau for the specified geographic location. Therefore, there are only 5 rows or observations in this table.

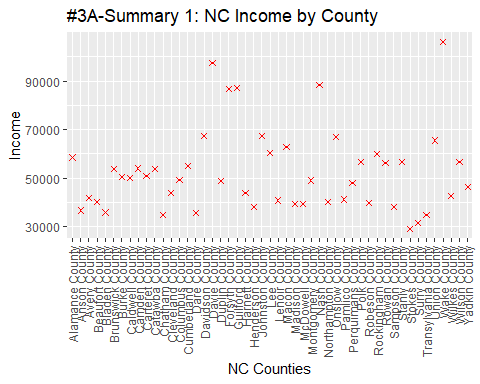
#### Part 2B: NC Summaries by Type

summary3 <- NC\_income %>% # Specifying the dataFRAME that will be used for all commands after this pipe  
 group\_by(Type) %>% # Grouping info by Type, which lists 5 Types of Community  
 summarise (mean = mean(IncomeMean)) # adding Mean Income to the dataset  
head(summary3, n=10) # Look at the 1st 10 rows in the tabl

## # A tibble: 5 × 2  
## Type mean  
## <chr> <dbl>  
## 1 CDP 45853.  
## 2 City 55884.  
## 3 Town 53116.  
## 4 Track 58381.  
## 5 Village 0

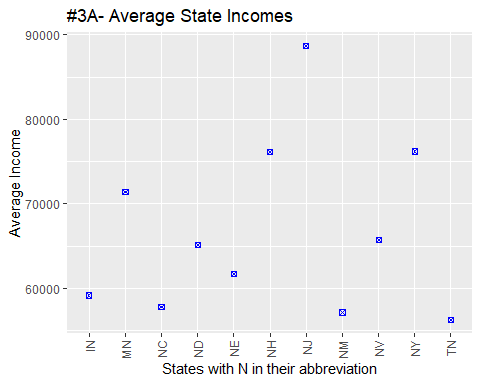
# Part 3 – Visualizing Results

#### Income Visualization of Summary 1 by County  
ggplot(summary1, aes(x=County, y=mean)) + # create a scatterplot graph for Summary 1  
 geom\_point(shape=4, color="red") + # plot the data points  
 geom\_smooth(method="lm", formula= "y~x" ) + # Adds a fitted line to the graph  
 theme(axis.text.x = element\_text(angle = 90,vjust =0.5, hjust=1)) + # make x axis labels vertical  
 labs(title = "#3A-Summary 1: NC Income by County", # Plot Title of Scatterplot  
 x = "NC Counties", # x axis label  
 y = "Income" # y axis label  
 )



You can see that Wake County, which has the state Capital, has the highest Income, according to Summary 1. The lowest income appears to be in Stokes County, followed closely by Surry County. Both of these counties are set in the Appalachian Mountains country land bordering on the Virgina State line. In contrast, Wake County is very urban and hosts a variety of government and higher education jobs.

#### Income Visualization of different states  
AvgStateIncome <- state\_income2 %>% # Specifying the dataFRAME that will be used for all commands after this pipe  
 group\_by(State\_ab) %>% # Grouping info by State\_ab  
 summarise (mean = mean(IncomeMean)) %>% # adding Mean Income to the dataset  
 filter (grepl("N", State\_ab)) # Include only the observations where there an N in the state\_ab  
  
ggplot(AvgStateIncome, aes(x=State\_ab, y=mean)) + # create a scatterplot graph   
 geom\_point(shape=7, color="blue") + # plot the data points  
 theme(axis.text.x = element\_text(angle = 90,vjust =0.5, hjust=1)) + # make x axis labels vertical  
 labs(title = "#3A- Average State Incomes", # Plot Title of Scatterplot  
 x = "States with N in their abbreviation", # x axis label  
 y = "Average Income" # y axis label  
 )



New Jersey showed a higher average income than the other ten states that were listed in this dataset. Their population is significantly more dense and urban than other states like Minnesota. I am unsure of when this data was collected because 2022 NC average Income surely would not be near the bottom of this list. With our exploding tech presence (Google, Amazon, Broadband, and Apple), our Average Incomes are swelling in the current market. Even through COVID lockdowns, our unemployment rates were fairly low because technology enabled employees to work at home. Therefore, their incomes stayed remarkable high compared to other states.