# Module 4 - Assignment 1

## Tate, Levi

### Data Transformation

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.4 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.0

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

state\_income <- read\_csv("state\_income.csv")

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## id = col\_double(),  
## State\_Code = col\_double(),  
## State\_Name = col\_character(),  
## State\_ab = col\_character(),  
## County = col\_character(),  
## City = col\_character(),  
## Place = col\_character(),  
## Type = col\_character(),  
## Primary = col\_character(),  
## Zip\_Code = col\_double(),  
## Area\_Code = col\_double(),  
## ALand = col\_double(),  
## AWater = col\_double(),  
## Lat = col\_double(),  
## Lon = col\_double(),  
## Mean = col\_double(),  
## Median = col\_double(),  
## Stdev = col\_double()  
## )

## Warning: 1 parsing failure.  
## row col expected actual file  
## 27548 Area\_Code a double M 'state\_income.csv'

#### State Incomes

I will be creating a subset of the data from state\_income. Variables used:  
1.) **State\_Name**: The state name reported by the U.S. Census Bureau for the specified geographic location.  
2.) **State\_ab**: The abbreviated state name reported by the U.S. Census Bureau for the specified geographic location.  
3.) **County**: The county name reported by the U.S. Census Bureau for the specified geographic location.  
4.) **City**: The city name reported by the U.S. Census Bureau for the specified geographic location.  
5.) **Type**:The place Type reported by the U.S. Census Bureau for the specified geographic location  
6.) **ALand**:The Square area of land at the geographic or track location.  
7.) **Mean**: The mean household income of the specified geographic location.  
8.) **Median**: The median household income of the specified geographic location.  
9.) **Stdev**: The standard deviation of the household income for the specified geographic location.

state\_income2 <- select(state\_income, State\_Name, State\_ab, County, City, Type,  
 ALand, Mean, Median, Stdev)  
select(state\_income2, State\_ab, everything())

## # A tibble: 32,526 x 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 Chickasaw City 1.09e7 38773 30506 33101  
## 2 AL Alabama Barbour Coun… Louisville City 2.61e7 37725 19528 43789  
## 3 AL Alabama Shelby County Columbiana City 4.48e7 54606 31930 57348  
## 4 AL Alabama Mobile County Satsuma City 3.69e7 63919 52814 47707  
## 5 AL Alabama Mobile County Dauphin I… Town 1.62e7 77948 67225 54270  
## 6 AL Alabama Cullman Coun… Cullman Town 8.91e6 50715 42643 35886  
## 7 AL Alabama Escambia Cou… East Brew… 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 Co… Sylacauga CDP 4.52e7 54807 41712 51359  
## # … with 32,516 more rows

head(state\_income2, n = 10)

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

state\_income2 <- rename(state\_income2, SquareArea = ALand, IncomeMean = Mean,   
 IncomeMedian = Median, IncomeStdev = Stdev)  
head(state\_income2, n = 10)

## # A tibble: 10 x 9  
## State\_Name State\_ab County City Type SquareArea IncomeMean IncomeMedian  
## <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl>  
## 1 Alabama AL Mobil… Chic… City 10894952 38773 30506  
## 2 Alabama AL Barbo… Loui… City 26070325 37725 19528  
## 3 Alabama AL Shelb… Colu… City 44835274 54606 31930  
## 4 Alabama AL Mobil… Sats… City 36878729 63919 52814  
## 5 Alabama AL Mobil… Daup… Town 16204185 77948 67225  
## 6 Alabama AL Cullm… Cull… Town 8913021 50715 42643  
## 7 Alabama AL Escam… East… City 8826252 33737 23610  
## 8 Alabama AL Elmor… Coos… Town 10222339 46319 40242  
## 9 Alabama AL Morga… Eva Town 10544874 57994 39591  
## 10 Alabama AL Talla… Syla… CDP 45178321 54807 41712  
## # … with 1 more variable: IncomeStdev <dbl>

NC\_income <- filter(state\_income2, State\_ab == "NC")  
head(NC\_income, n = 10)

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

#### NC Incomes

I will be using the NC\_income dataset to create summaries of the incomes within North Carolina including summaries by county, city and type.

arrange(NC\_income, County) %>%  
head(NC\_income, n = 10)

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

summary1 <- group\_by(NC\_income, County)  
summary1 <- summarise (summary1, mean = mean(IncomeMean))

## `summarise()` ungrouping output (override with `.groups` argument)

summary2 <- NC\_income %>%   
 group\_by(City) %>%  
 summarise(mean = mean(IncomeMean))

## `summarise()` ungrouping output (override with `.groups` argument)

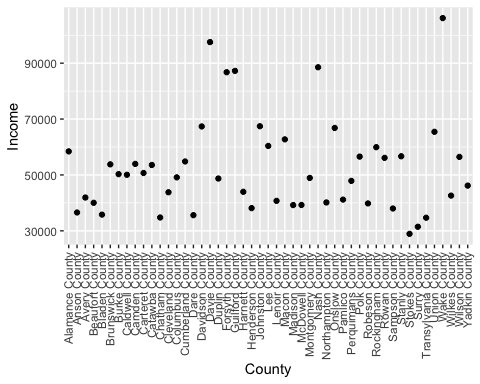
summary1 shows the mean household income by County. summary2 shows the mean household income by City.

summary3 <- NC\_income %>%  
 group\_by(Type) %>%  
 summarise(mean = mean(IncomeMean))

## `summarise()` ungrouping output (override with `.groups` argument)

#### Income Visualization

ggplot(data = summary1, mapping = aes(x = County, y = mean)) +  
geom\_point() +  
labs(y = "Income") +  
theme(axis.text.x = element\_text(angle = 90,vjust =0.5, hjust=1))

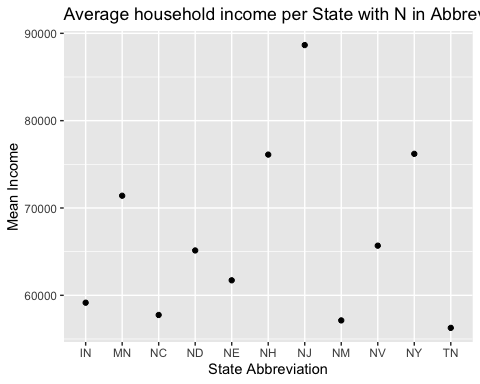


Wake County has the largest average income per household: ~$107,000. Stokes County has the lowest average income per household: ~29,000. Surry County has the second lowest average income per household: ~$32,000.

AvgStateIncome <- state\_income2 %>%  
 group\_by(State\_ab) %>%  
 summarise(mean = mean(IncomeMean)) %>%  
 filter(grepl("N", State\_ab))

## `summarise()` ungrouping output (override with `.groups` argument)

ggplot(data = AvgStateIncome, mapping = aes(x = State\_ab, y = mean)) +  
 geom\_point() +  
 labs(title = "Average household income per State with N in Abbreviation",  
 x = "State Abbreviation",  
 y = "Mean Income")



New Jersey has the largest average income per household: ~$87,000. Tennessee has the lowest average income per household: ~$53,000.