# Module 4 - Assignment 1

## Kiser, Stephen

### Data Transformation

library(tidyverse)

## -- Attaching packages ------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.5  
## v tidyr 1.0.2 v stringr 1.4.0  
## v readr 1.3.1 v 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")

## Parsed with 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

In this part of the assignment I will be taking a subset of data to which gives us smaller data to work with. I will be using 9 variables of the 18 variables in the dataset. We will be using the name of the state and abbreviation. Then we will be using the county and city name where the U.S. Census Bureau used for the data.Then we will use Aland which is the square ara of land at the location. Finally the last three variables we will be using are the mean household income, median, and standard deviation.

state\_income2 <- select(state\_income, State\_ab, State\_Name, County, City, Type, ALand, Mean, Median, Stdev)  
  
head(state\_income2,10)

## # A tibble: 10 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

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

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

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

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

#### NC Incomes

For this part of the assignment I will be using the North Carolina dataset I created in the previous part. We will be summarizing the incomes within the state of North carolina. We will be finding the incomes by county, city, and type. First, I willl need to arrange the data by ascending order in the County column.

NC\_income <- arrange(NC\_income, County)  
  
head(NC\_income, 10)

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

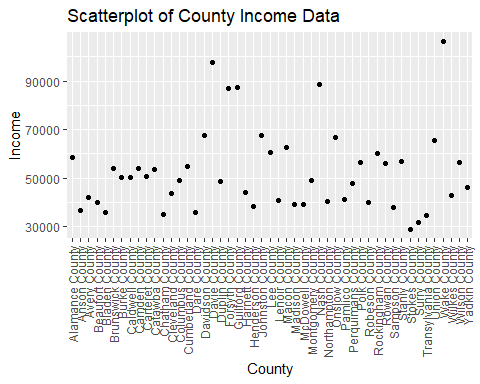
summary1 <- group\_by(NC\_income,County)  
summary1 <- summarise (summary1, mean= mean(IncomeMean))  
summary2 <- NC\_income %>%  
 group\_by(City) %>%  
 summarise (mean = mean(IncomeMean))

I created two new datasets called summary1 and summary2. Summary1 has the counties of North Carolina and has the mean income in that county. Summary2 is a dataset of cities in North Carolina and their average income. Both datasets are arranged in A-Z for both counties and cities.

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

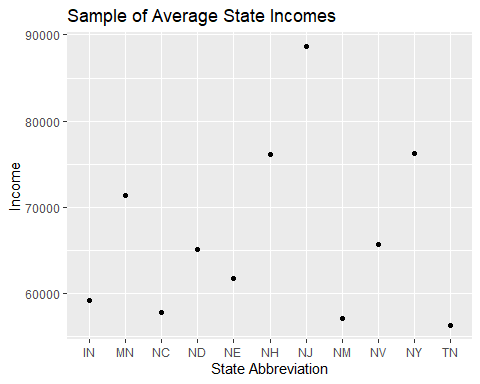
#### Income Visualization

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



After making the scatterplot from dataset summary1 I will now will find the highest income and lowest income county. The highest income county is Wake County. The lowest income county is Stokes County. The second lowest income county is Surry County.

AvgStateIncome <- state\_income2 %>%  
 group\_by(State\_ab) %>%  
 summarise(mean=mean(IncomeMean))  
  
AvgStateIncome <- filter(AvgStateIncome, grepl("N",State\_ab))  
  
ggplot(AvgStateIncome, aes(State\_ab, mean)) +  
 geom\_point() +  
 labs(title = "Sample of Average State Incomes", x= "State Abbreviation", y= "Income")



After creating this scatterplot of 11 states with an “N” in their abreviation I am able to see the average incomes. The largest average income was in New Jersey with 88657.64. The smallest average income was in Tennessee 56271.95. North Carolina had the third lowest average income of this sample.