Hi Caity,

This is a sufficient demonstration of your data restructuring skills. I see you are using the pivot and join functions appropriately. I also see that you check your wrangling regularly, which is good practice! Well done! Here are some issues that you may want to consider

* I missed if you checked cases that couldn’t be joined. That’s always good to check!
* Can you label 2014-2018 data as 2010? You do this for the education data and that is quite a stretch. I couldn’t find a good justification for this decision either.
* The figures for Q2 couldn’t fully answer the questions “how do education and income on the county level relate” and “does this relationship differ between rural and urban counties and over time?”. Your figures spoke to the question “how does personal income in counties develop over time and between rural-urban areas” only.

Grade: B+

EDUC 423A/SOC 302A: Assignment 3

Caity McGinley

2/28/2021

# Honor Code Statement

We strongly encourage students to form study groups and students may discuss and work on assignments in groups. We expect that each student understands their own submission. As such, students must write their submissions independently and clearly disclose the names of all other students who were part of their study group. Additionally, lifting code or solutions directly from the internet (e.g., Google, GitHub, Stack Overflow) is a violation of the [Stanford Honor Code](https://communitystandards.stanford.edu/policies-and-guidance/honor-code). We take academic honesty and Honor Code violations extremely seriously and expect the same of students. If you have questions about what may or may not constitute an Honor Code violation, please reach out the teaching team.

**Email:** [caitym26@stanford.eud](mailto:caitym26@stanford.eud)

**SUID:** 06211809 **Study Group:** Tangaloa Kaufusi

I acknowledge and agree to abide by the Honor Code.

**Signed:** Caity McGinley

# Setup and cleaning the data

# Include all code required to load packages, import, clean, pivot and join the data here.

library(knitr)

## Warning: package 'knitr' was built under R version 4.0.3

library (dplyr) No need to attach dplyr as it is part of the tidyverse library ☺

## Warning: package 'dplyr' was built under R version 4.0.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

library(tidyverse)

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

## v tibble 3.0.3 v purrr 0.3.4  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## Warning: package 'readr' was built under R version 4.0.4

## Warning: package 'stringr' was built under R version 4.0.3

## Warning: package 'forcats' was built under R version 4.0.3

## -- Conflicts ------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x ggplot2::%+%() masks psych::%+%()  
## x ggplot2::alpha() masks psych::alpha()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(tidytext)

## Warning: package 'tidytext' was built under R version 4.0.3

library(textdata)

## Warning: package 'textdata' was built under R version 4.0.3

library(janitor)

## Warning: package 'janitor' was built under R version 4.0.3

##   
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test

library(stringr)  
library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.0.3

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

library(grid)  
library(stringi)

library(janitor)  
library(readr)

setwd("C:/Users/cmcgi/Downloads/Soc 302A\_Lab1")  
  
#loading data  
Education <- read\_csv("Education.csv") #The Education table only includes the 2003 and 2013 versions of the Rural-Urban Continuum Code, which brings us to the next collection of data tables.

## Warning: Missing column names filled in: 'X2' [2], 'X3' [3], 'X4' [4], 'X5' [5],  
## 'X6' [6], 'X7' [7], 'X8' [8], 'X9' [9], 'X10' [10], 'X11' [11], 'X12' [12],  
## 'X13' [13], 'X14' [14], 'X15' [15], 'X16' [16], 'X17' [17], 'X18' [18],  
## 'X19' [19], 'X20' [20], 'X21' [21], 'X22' [22], 'X23' [23], 'X24' [24],  
## 'X25' [25], 'X26' [26], 'X27' [27], 'X28' [28], 'X29' [29], 'X30' [30],  
## 'X31' [31], 'X32' [32], 'X33' [33], 'X34' [34], 'X35' [35], 'X36' [36],  
## 'X37' [37], 'X38' [38], 'X39' [39], 'X40' [40], 'X41' [41], 'X42' [42],  
## 'X43' [43], 'X44' [44], 'X45' [45], 'X46' [46], 'X47' [47]

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## .default = col\_character()  
## )  
## i Use `spec()` for the full column specifications.

rucpr2003 <- read\_csv("pr2003.csv") #in spanish, possibly can change variable names

## Warning: Missing column names filled in: 'X7' [7], 'X8' [8], 'X9' [9],  
## 'X10' [10], 'X11' [11]

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## `FIPS Code` = col\_double(),  
## State = col\_character(),  
## `Municipio Name` = col\_character(),  
## `Population 2003` = col\_number(),  
## `Rural-urban Continuum Code, 2003` = col\_double(),  
## `Description of the 2003 Code` = col\_character(),  
## X7 = col\_logical(),  
## X8 = col\_logical(),  
## X9 = col\_logical(),  
## X10 = col\_logical(),  
## X11 = col\_logical()  
## )

ruc1974 <- read\_csv("ruralurbancodes1974.csv")

## Warning: Missing column names filled in: 'X5' [5], 'X6' [6], 'X7' [7]

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## `FIPS Code` = col\_character(),  
## State = col\_character(),  
## `County Name` = col\_character(),  
## `1974 Rural-urban Continuum Code` = col\_double(),  
## X5 = col\_logical(),  
## X6 = col\_logical(),  
## X7 = col\_logical()  
## )

ruc2003 <- read\_csv("ruralurbancodes2003.csv") #looks like cd8393 but has three more columns, rename description, will need to pivot

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## `FIPS Code` = col\_character(),  
## State = col\_character(),  
## `County Name` = col\_character(),  
## `1993 Rural-urban Continuum Code` = col\_double(),  
## `2003 Rural-urban Continuum Code` = col\_double(),  
## `2000 Population` = col\_number(),  
## `Percent of workers in nonmetro counties commuting to central counties of adjacent metro areas` = col\_double(),  
## `Description for 2003 codes` = col\_character()  
## )

ruc2013 <- read\_csv("ruralurbancodes2013.csv") #get rid \_ in "count\_name" and 2013\_pop, and rename the continuum code variable

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## FIPS = col\_character(),  
## State = col\_character(),  
## County\_Name = col\_character(),  
## Population\_2010 = col\_number(),  
## RUCC\_2013 = col\_double(),  
## Description = col\_character()  
## )

ruc8393<- read\_csv("cd8393.csv") #change to lowercase

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## FIPS = col\_character(),  
## State = col\_character(),  
## `County Name` = col\_character(),  
## `1983 Rural-urban Continuum Code` = col\_double(),  
## `1993 Rural-urban Continuum Code` = col\_double()  
## )

ruc93 <- read\_csv("code93.csv") #Don't need this one

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## `FIPS Code` = col\_character(),  
## State = col\_character(),  
## `County Name` = col\_character(),  
## `Rural-urban Continuuum Code, 1993` = col\_double()  
## )

CAIN <- read\_csv("CAINC1\_\_ALL\_AREAS\_1969\_2019.csv")

##   
## -- Column specification ---------------------------------------------------------------------------------------------------  
## cols(  
## .default = col\_character(),  
## Region = col\_double(),  
## LineCode = col\_double()  
## )  
## i Use `spec()` for the full column specifications.

## Warning: 4 parsing failures.  
## row col expected actual file  
## 9595 -- 59 columns 1 columns 'CAINC1\_\_ALL\_AREAS\_1969\_2019.csv'  
## 9596 -- 59 columns 1 columns 'CAINC1\_\_ALL\_AREAS\_1969\_2019.csv'  
## 9597 -- 59 columns 1 columns 'CAINC1\_\_ALL\_AREAS\_1969\_2019.csv'  
## 9598 -- 59 columns 1 columns 'CAINC1\_\_ALL\_AREAS\_1969\_2019.csv'

#Plans  
#1 get rid of useless rows/columns  
#2 put new first row onto variable names   
#3 Clean up other data sets  
# 4 Then join be like variables   
  
#### Cleaning ####  
#janitor   
ruc8393 <- ruc8393 %>%   
clean\_names() %>%   
select (- county\_name, - state)  
ruc1974 <- ruc1974 %>%   
clean\_names() %>%   
select (- county\_name, - state, - x5, - x6, - x7)  
ruc2003 <- ruc2003 %>%   
clean\_names() %>%   
select (- county\_name, - state, - x2000\_population, -description\_for\_2003\_codes)  
ruc2013 <- ruc2013 %>%   
clean\_names() %>%   
select (- county\_name, - state, - description, -population\_2010)  
rucpr2003 <- rucpr2003 %>%   
clean\_names() %>%   
select (- x7, - x8, - x9, - x10, - x11, -state, -population\_2003, -municipio\_name)   
CAIN <- CAIN %>%   
clean\_names()   
  
CAIN <- CAIN %>%   
rename(fips\_code = geo\_fips, county\_name = geo\_name) %>%   
select (-industry\_classification, -table\_name, -county\_name, -unit, -region, -line\_code)  
  
#Rename area\_name to county\_name  
Education <- Education %>%  
 row\_to\_names(row\_number = 4) %>%   
clean\_names()%>% rename(county\_name = area\_name) %>%   
select(-county\_name)  
  
  
#Renaming fips  
ruc8393 <- ruc8393 %>%   
rename(fips\_code = fips)  
  
ruc2013 <- ruc2013 %>%   
rename(fips\_code = fips, x2013\_rural\_urban\_continuum\_code = rucc\_2013)  
  
#RUCC Code Renaming  
rucpr2003 <- rucpr2003 %>%   
rename(x2003\_rural\_urban\_continuum\_code = rural\_urban\_continuum\_code\_2003) %>%   
select (- description\_of\_the\_2003\_code)  
  
  
#order joining to get years in order   
  
#merging RUC Code Data   
data1 <- left\_join(ruc1974, ruc8393, by = c("fips\_code"))   
  
data1 <- left\_join(data1, ruc2003, by = c("fips\_code"))  
  
data1 <- left\_join(data1, ruc2013, by = c("fips\_code")) %>%   
select (- x1993\_rural\_urban\_continuum\_code.y) %>%   
rename (x1993\_rural\_urban\_continuum\_code = x1993\_rural\_urban\_continuum\_code.x)

How do you deal with 1993 popping up in several tables? Which one will you use?   
  
data1 <- data1 %>%   
lapply(as.numeric) %>%   
as.data.frame()

## Warning in lapply(., as.numeric): NAs introduced by coercion

data1 <- left\_join(data1, rucpr2003, by = c("fips\_code", "x2003\_rural\_urban\_continuum\_code"), na.rm = TRUE)  
  
data1 <- data1 %>%  
 pivot\_longer(names\_to=c('year', 'Rural\_Urban\_Code'), values\_to='value', names\_sep='\_', cols= starts\_with("x"))

Try this instead

pivot\_longer(cols = -c(fips, description),

names\_to = "year",

values\_to = "rural\_urban\_continuum")

## Warning: Expected 2 pieces. Additional pieces discarded in 5 rows [1, 2, 3, 4,  
## 5].

data1 <- data1 %>%   
select (- Rural\_Urban\_Code) %>%   
rename(Rural\_Urban\_code = value)  
  
data1 <- data1 %>%  
 mutate(year = str\_replace\_all(year, "x", ""))  
   
#checking  
class(data1$year)

## [1] "character"

data1$year <- as.numeric(data1$year, na.rm = TRUE)  
class(data1$year)

## [1] "numeric"

class(data1$fips\_code)

## [1] "numeric"

glimpse(data1)

## Rows: 15,805  
## Columns: 4  
## $ fips\_code <dbl> ...  
## $ percent\_of\_workers\_in\_nonmetro\_counties\_commuting\_to\_central\_counties\_of\_adjacent\_metro\_areas <dbl> ...  
## $ year <dbl> ...  
## $ Rural\_Urban\_code <dbl> ...

rucc <- data1 %>%   
select(-percent\_of\_workers\_in\_nonmetro\_counties\_commuting\_to\_central\_counties\_of\_adjacent\_metro\_areas)  
  
#Education data set   
  
Education$fips\_code <- as.numeric(Education$fips\_code, na.rm = TRUE)  
  
Education <- Education %>%   
select(fips\_code, starts\_with("percent"))  
  
  
glimpse(Education)

#What group am I interested in? %percent of adults with percent\_of\_adults\_with\_a\_bachelors\_degree\_or\_higher  
  
Education <- Education %>%   
rename (percent\_of\_adults\_with\_a\_bachelors\_degree\_or\_higher\_1970 = percent\_of\_adults\_completing\_four\_years\_of\_college\_or\_higher\_1970, percent\_of\_adults\_with\_a\_bachelors\_degree\_or\_higher\_1980 = percent\_of\_adults\_completing\_four\_years\_of\_college\_or\_higher\_1980)  
  
#note done yet  
Education <- Education %>%  
 pivot\_longer(names\_to= 'education', values\_to='value', cols= starts\_with("percent"))

Education <- Education %>% filter(str\_detect(education, "bachelors"))  
  
   
 Education <- Education %>%  
 mutate(year = str\_remove\_all(education, pattern ="[[:alpha:]]"))   
   
unique(Education$year)

## [1] "\_\_\_\_\_\_\_\_\_1970" "\_\_\_\_\_\_\_\_\_1980" "\_\_\_\_\_\_\_\_\_1990" "\_\_\_\_\_\_\_\_\_2000"   
## [5] "\_\_\_\_\_\_\_\_\_2014\_18"

table(Education$year)

##   
## \_\_\_\_\_\_\_\_\_1970 \_\_\_\_\_\_\_\_\_1980 \_\_\_\_\_\_\_\_\_1990 \_\_\_\_\_\_\_\_\_2000   
## 3283 3283 3283 3283   
## \_\_\_\_\_\_\_\_\_2014\_18   
## 3283

Education <- Education %>%  
 mutate(year = case\_when(  
 year == "\_\_\_\_\_\_\_\_\_1970" ~ 1974,  
 year == "\_\_\_\_\_\_\_\_\_1980" ~ 1983,   
 year == "\_\_\_\_\_\_\_\_\_1990" ~ 1993,  
 year == "\_\_\_\_\_\_\_\_\_2000" ~ 2003,   
 year == "\_\_\_\_\_\_\_\_\_2014\_18" ~ 2013))  
  
  
yeah this did not really work as intended

Alternatives

year = str\_extract(variable,

"\\d{4,}...|\\d{4,}")

#CAIN data  
  
CAIN <- pivot\_wider(  
 CAIN,  
 id\_cols = NULL,  
 names\_from = description,  
 names\_sep = "\_",  
 values\_from = x1969:x2019)

I see what you are doing here, but you don’t really need the values of description to become variables. You will only need one variable, so it is easier to filter for one. Example

filter(description == "Per capita personal income (dollars) 2/")

CAIN <- CAIN %>% pivot\_longer(names\_to=c('year', 'measure'), values\_to='value', names\_sep='\_', cols= starts\_with("x"))

Because now you get the variable names back into variable values again! That is, the variable measure = the old variable description  
  
CAIN <- CAIN %>%  
 mutate(year = str\_replace\_all(year, "x", ""))  
   
class(CAIN$year)

## [1] "character"

CAIN$year <- as.numeric(CAIN$year)  
class(CAIN$year)

## [1] "numeric"

CAIN$fips\_code <- as.numeric(CAIN$fips\_code)

## Warning: NAs introduced by coercion you want to check this, because this is your key variable for joining

CAIN$value <- as.numeric(CAIN$value)

## Warning: NAs introduced by coercion

class(CAIN$measure)

## [1] "character"

CAIN <- na.omit(CAIN)   
  
CAIN2 <- pivot\_wider(  
 CAIN,  
 id\_cols = NULL,  
 names\_from = measure,  
 names\_prefix = "",  
 names\_sep = "\_",  
 names\_glue = NULL,  
 names\_sort = FALSE,  
 names\_repair = "check\_unique",  
 values\_from = value)  
   
   
CAIN2 <- CAIN2 %>%   
clean\_names()  
  
# Last merge  
complete <- left\_join(rucc, CAIN2, by='fips\_code', 'year') %>%  
 left\_join(., Education, by='fips\_code')   
complete <- complete %>%   
select (-year.x , - year.y)

#Last cleaning of names  
complete <- complete %>%   
clean\_names()  
  
lapply(complete,class)

## $fips\_code  
## [1] "numeric"  
##   
## $rural\_urban\_code  
## [1] "numeric"  
##   
## $personal\_income\_thousands\_of\_dollars  
## [1] "numeric"  
##   
## $population\_persons\_1  
## [1] "numeric"  
##   
## $per\_capita\_personal\_income\_dollars\_2  
## [1] "numeric"  
##   
## $education  
## [1] "character"  
##   
## $value  
## [1] "character"  
##   
## $year  
## [1] "numeric"

complete$value <- as.numeric(complete$value)  
  
lapply(complete,class)

## $fips\_code  
## [1] "numeric"  
##   
## $rural\_urban\_code  
## [1] "numeric"  
##   
## $personal\_income\_thousands\_of\_dollars  
## [1] "numeric"  
##   
## $population\_persons\_1  
## [1] "numeric"  
##   
## $per\_capita\_personal\_income\_dollars\_2  
## [1] "numeric"  
##   
## $education  
## [1] "character"  
##   
## $value  
## [1] "numeric"  
##   
## $year  
## [1] "numeric"

#DONEEEEE! Yay!

YAYY ☺

# How does the relationship between education and income varies between urban and rural areas over time?

**Q1: How is educational inequality between areas with different urbanization levels changing over time?**

# Include all code required to generate your visualization here.  
  
# Cleaning

complete <- na.omit(complete)   
  
# Getting rid of scientific notation.

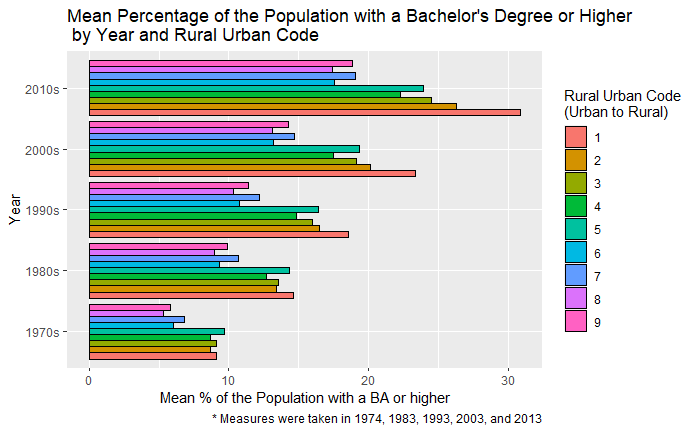
options(scipen = 999)  
  
# Checking  
lapply(complete,class)

## $fips\_code  
## [1] "numeric"  
##   
## $rural\_urban\_code  
## [1] "numeric"  
##   
## $personal\_income\_thousands\_of\_dollars  
## [1] "numeric"  
##   
## $population\_persons\_1  
## [1] "numeric"  
##   
## $per\_capita\_personal\_income\_dollars\_2  
## [1] "numeric"  
##   
## $education  
## [1] "character"  
##   
## $value  
## [1] "numeric"  
##   
## $year  
## [1] "numeric"

complete <- filter(complete, rural\_urban\_code %in% 1:9)  
urban <- filter(complete, rural\_urban\_code %in% 1:3)  
rural <- filter(complete, rural\_urban\_code %in% 4:9)  
  
urban <- urban %>%   
 clean\_names()  
  
rural <-rural %>%   
 clean\_names()  
  
complete$rural\_urban\_code <- as.character(complete$rural\_urban\_code)  
urban$rural\_urban\_code <- as.character(urban$rural\_urban\_code)  
rural$rural\_urban\_code <- as.character(rural$rural\_urban\_code)  
  
  
completedata3 <- complete %>%   
 group\_by(rural\_urban\_code, year) %>%   
 summarise(mean = mean(value, na.rm = TRUE))

## `summarise()` regrouping output by 'rural\_urban\_code' (override with `.groups` argument)

ggplot(completedata3, aes(x = reorder(year, mean), y = mean, fill = rural\_urban\_code)) +  
 geom\_col(position = "dodge", color = "black") +   
 scale\_x\_discrete(labels = c("1970s", "1980s", "1990s",   
 "2000s", "2010s")) +  
 coord\_flip() +  
 labs(title = "Mean Percentage of the Population with a Bachelor's Degree or Higher \n by Year and Rural Urban Code",  
 caption = "\* Measures were taken in 1974, 1983, 1993, 2003, and 2013",   
 y = "Mean % of the Population with a BA or higher",  
 x = "Year",  
 fill = "Rural Urban Code \n(Urban to Rural)")



Good going on this plot. I LOVE the notation about the UR code, but I don’t know if anyone who didn’t know the assignment would know that “measures taking from” refers to the UR code. If I didn’t know better, I’d think that the % bachelor degree holders data was taken from 174, 1983, 1993, 2003 and 2013 (and that is not the case)

Placing the year on the y axis is a unconventional and I am not sure if it is a good idea to name the year 2010’s as such, because the education data comes from 2014-2018.

ruraldata <- rural %>%   
 group\_by(rural\_urban\_code, year) %>%   
 summarise(mean = mean(per\_capita\_personal\_income\_dollars\_2, na.rm = TRUE))

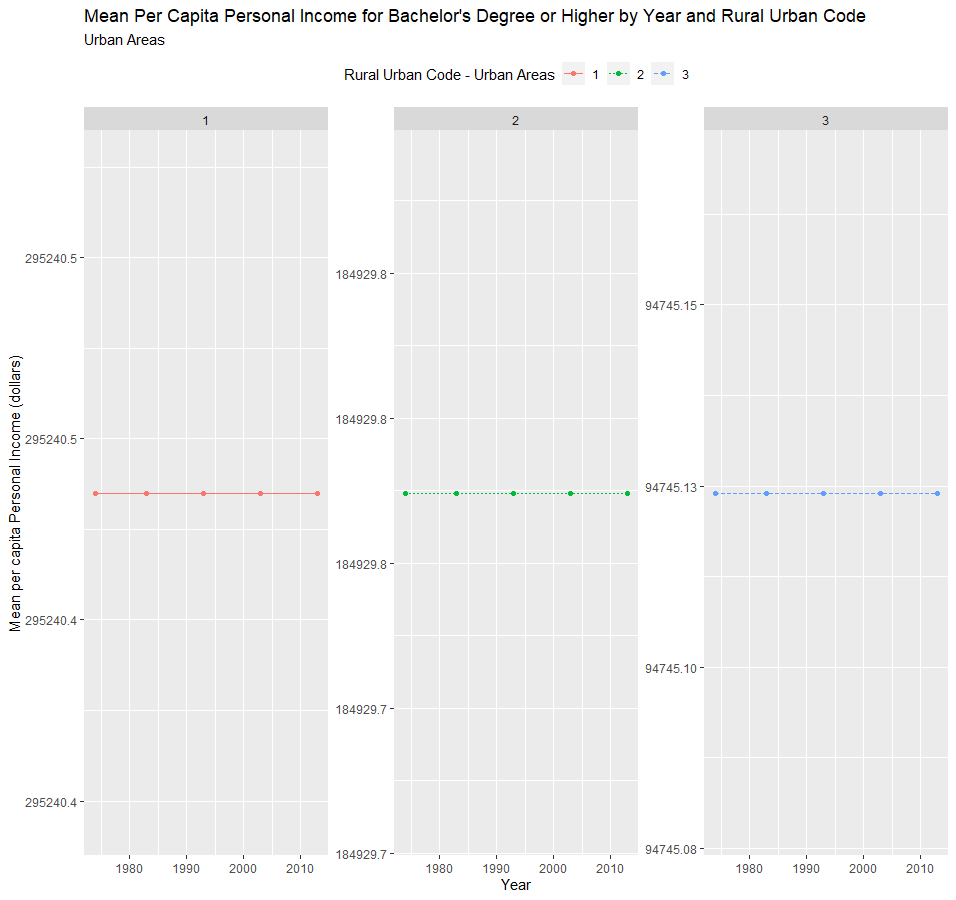
## `summarise()` regrouping output by 'rural\_urban\_code' (override with `.groups` argument)

urbandata <- urban %>%   
 group\_by(rural\_urban\_code, year) %>%   
 summarise(mean = mean(population\_persons\_1, na.rm = TRUE))

Why are you taking the mean population?

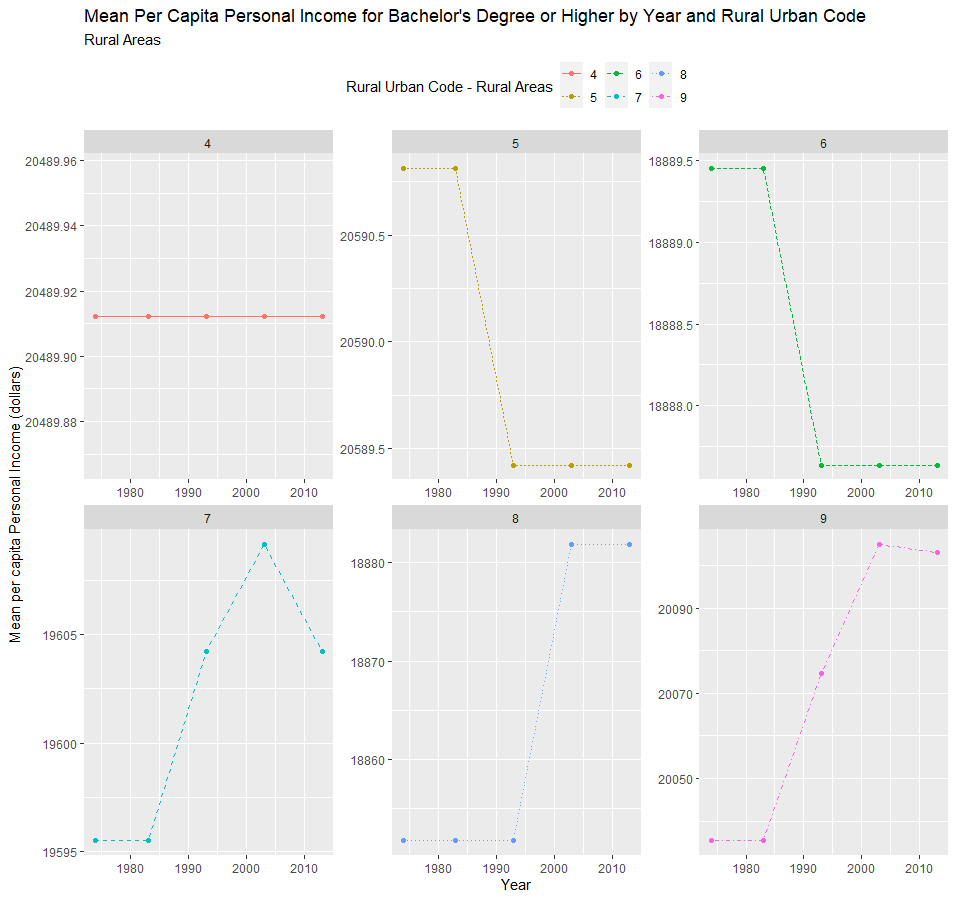
## `summarise()` regrouping output by 'rural\_urban\_code' (override with `.groups` argument)

ggplot(urbandata, aes(x=year, y=mean, group=rural\_urban\_code)) +  
 geom\_line(aes(linetype=rural\_urban\_code, color=rural\_urban\_code))+  
 geom\_point(aes(color=rural\_urban\_code))+  
 theme(legend.position="top") +  
 facet\_wrap(~rural\_urban\_code, scales = "free") +  
 labs(title = "Mean Per Capita Personal Income for Bachelor's Degree or Higher by Year and Rural Urban Code",  
 subtitle = "Urban Areas",  
 caption = "\* Measures were taken in 1974, 1983, 1993, 2003, and 2013",   
 y = "Mean per capita Personal Income (dollars)",  
 x = "Year",  
 color = "Rural Urban Code - Urban Areas",   
 linetype = "Rural Urban Code - Urban Areas")



Hmm I am not sure what this plot is showing. From your code it looks like you are plotting not capita but population size. But I think you didn’t mean to do that as your next plot is income. Other than that, I don’t think it is right that the population stays so stable over time, do you? Note that we did not ask you to assess income inequality over time.

ggplot(ruraldata, aes(x=year, y=mean, group=rural\_urban\_code)) +  
 geom\_line(aes(linetype=rural\_urban\_code, color=rural\_urban\_code))+  
 geom\_point(aes(color=rural\_urban\_code))+  
 facet\_wrap(~rural\_urban\_code, scales = "free") +  
 theme(legend.position="top") +  
 labs(title = "Mean Per Capita Personal Income for Bachelor's Degree or Higher by Year and Rural Urban Code",  
 subtitle = "Rural Areas",  
 caption = "\* Measures were taken in 1974, 1983, 1993, 2003, and 2013",   
 y = "Mean per capita Personal Income (dollars)",  
 x = "Year",  
 color = "Rural Urban Code - Rural Areas",   
 linetype = "Rural Urban Code - Rural Areas")

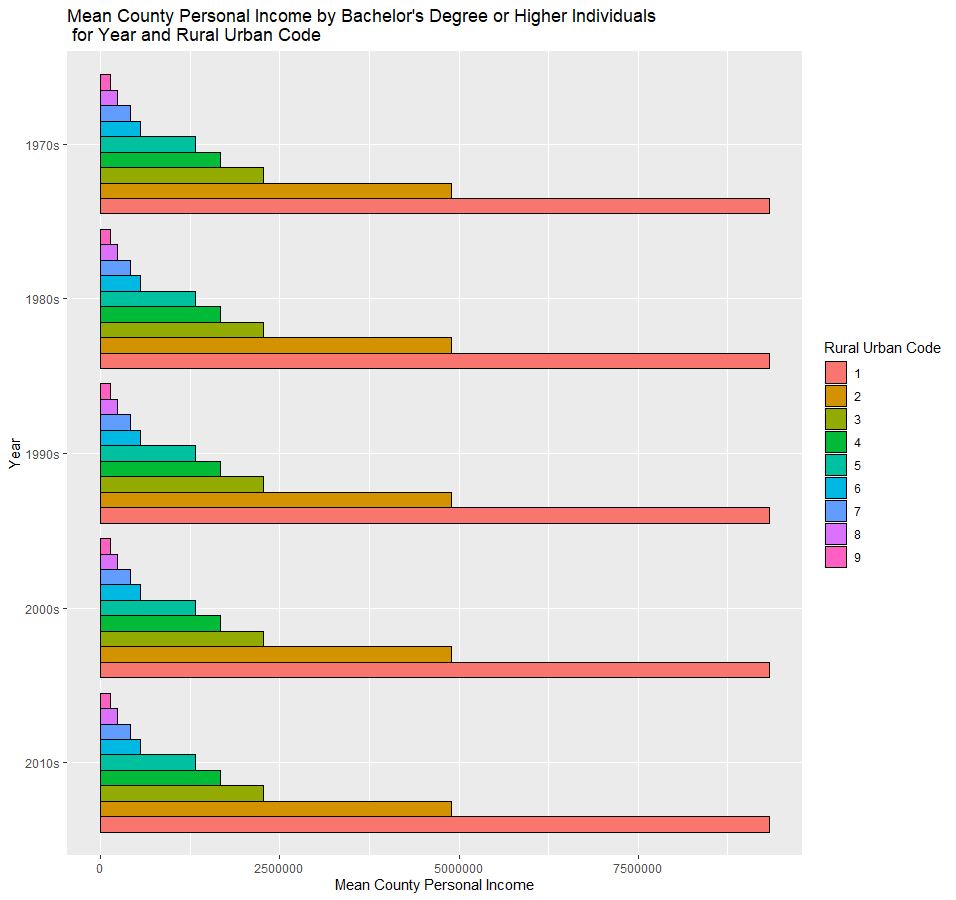


**Q2: How is county educational attainment level related to county personal income vary between areas with different urbanization levels and does this relationship change over time?**

# Include all code required to generate your visualization here.  
  
  
completedata2 <- complete %>%   
 group\_by(rural\_urban\_code, year) %>%   
 summarise(mean = mean(personal\_income\_thousands\_of\_dollars, na.rm = TRUE))

## `summarise()` regrouping output by 'rural\_urban\_code' (override with `.groups` argument)

ggplot(completedata2, aes(x = reorder(year, mean), y = mean, fill = rural\_urban\_code)) +  
 geom\_col(position = "dodge", color = "black") +   
 scale\_x\_discrete(labels = c("2010s", "2000s", "1990s",   
 "1980s", "1970s")) +  
 coord\_flip() +  
 labs(title = "Mean County Personal Income by Bachelor's Degree or Higher Individuals \n for Year and Rural Urban Code",  
 y = "Mean Personal Income",  
 x = "Year",  
 fill = "Rural Urban Code")



This plot shows mean personal income in a county over time by rural urban code but the question was about the relationship between income and education. I can’t assess that from this plot because I can’t see variation in % college degree holders between counties. Also, is personal income not sensitive to county size? Why not pick income per capita?

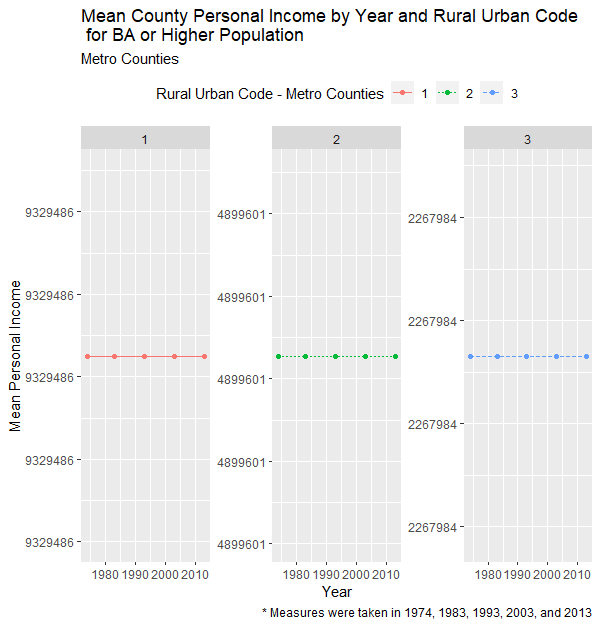
ruraldata2 <- rural %>%   
 group\_by(rural\_urban\_code, year) %>%   
 summarise(mean = mean(personal\_income\_thousands\_of\_dollars, na.rm = TRUE))

## `summarise()` regrouping output by 'rural\_urban\_code' (override with `.groups` argument)

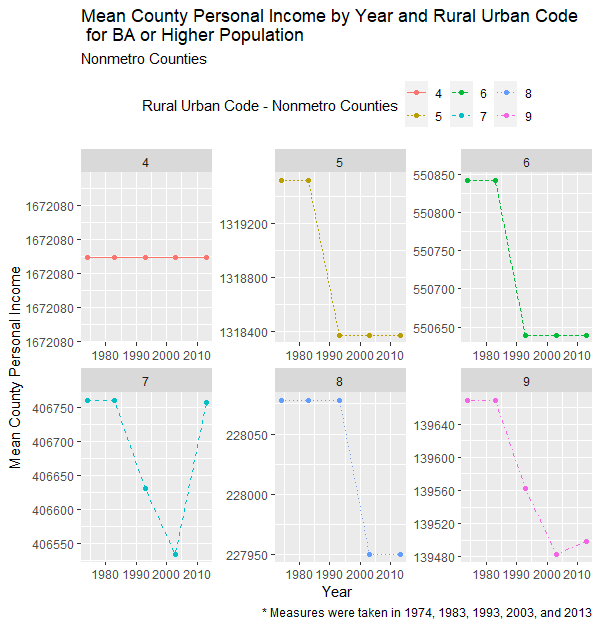
urbandata2 <- urban %>%   
 group\_by(rural\_urban\_code, year) %>%   
 summarise(mean = mean(personal\_income\_thousands\_of\_dollars, na.rm = TRUE))

## `summarise()` regrouping output by 'rural\_urban\_code' (override with `.groups` argument)

ggplot(urbandata2, aes(x=year, y=mean, group=rural\_urban\_code)) +  
 geom\_line(aes(linetype=rural\_urban\_code, color=rural\_urban\_code))+  
 geom\_point(aes(color=rural\_urban\_code))+  
 theme(legend.position="top") +  
 facet\_wrap(~rural\_urban\_code, scales = "free") +  
 labs(title = "Mean County Personal Income by Year and Rural Urban Code \n for BA or Higher Population",  
 subtitle = "Metro Counties",  
 caption = "\* Measures were taken in 1974, 1983, 1993, 2003, and 2013",   
 y = "Mean Personal Income",  
 x = "Year",  
 color = "Rural Urban Code - Metro Counties",   
 linetype = "Rural Urban Code - Metro Counties")



ggplot(ruraldata2, aes(x=year, y=mean, group=rural\_urban\_code)) +  
 geom\_line(aes(linetype=rural\_urban\_code, color=rural\_urban\_code))+  
 geom\_point(aes(color=rural\_urban\_code))+  
 facet\_wrap(~rural\_urban\_code, scales = "free") +  
 theme(legend.position="top") +  
 labs(title = "Mean County Personal Income by Year and Rural Urban Code \n for BA or Higher Population",  
 subtitle = "Nonmetro Counties",  
 caption = "\* Measures were taken in 1974, 1983, 1993, 2003, and 2013",   
 y = "Mean County Personal Income",  
 x = "Year",  
 color = "Rural Urban Code - Nonmetro Counties",   
 linetype = "Rural Urban Code - Nonmetro Counties")



I’m not sure how these visualizations help answering the RQ

**Discussion**

*Referencing your visualizations from above, answer the three research questions.*

**Q1: How is educational inequality between areas with different urbanization levels changing over time?**

In my plot visulaizing the mean percentage of individual’s who hold a BA or higher in the population, I can see that no matter the urban code, the mean percentage of the population with a BA or high is growing! Some urban codes, such as 8 and 9, which represents “Completely rural or less than 2,500 urban population, adjacent to a metro area” and “Completely rural or less than 2,500 urban population, non-adjacent to a metro area” are more stagnant from the 80s and 90s than more urban areas, such as codes 1 or 2 (the codes go from metro to more rural). However, overall, the data shows that more people are attending college, completing their BA and pursuing/completing more graduate work. There seems to be a particular boom in the 2000s and 2010s as well, indicating more individuals attended college after 2000.yes. but did you also notice that the gap between rural and urban is increasing over time? So urban areas have increasingly more and more college graduates, but this increase of college graduates isn’t as as steep for rural areas. Education inequality between rural and urban areas is rising.

For example



**Q2: How is county educational attainment level related to county personal income vary between areas with different urbanization levels and does this relationship change over time?**

My line plots showing mean county personal income by bachelor’s degree well you show mean personal income by year. Not by % bachelor degree holders of higher show that, for metro counties (codes 1-3) the average mean county income stays relatively stagnant over time. However, there is a difference based on rural urban codes for metro counties. 1, or “counties in metro areas of 1 million population or more” have the highest mean county personal income for BA or higher population, followed by 2, or “Counties in metro areas of 250,000 to 1 million population”, and lastly, 3, or “Counties in metro areas of fewer than 250,000 population”. This makes sense as more people in a county can help contribute to a higher county personal income. Yes, that is why it personal income per capita is a more useful variable

Again, as seen in the metro counties, the codes with the larger populations (4-5-6) make much more on average for the county personal income than the more rural counties (7-8-9). This is to be expected as there are less oppurtunities for higher paying work and the smaller population. However, in nonmetro counties, the stagnant trend in average county person income is not replicated. Interestingly, code 4, which represents “Urban population of 20,000 or more, adjacent to a metro area” aka the suburbs, conforms to the stagnant tend seen in metro counties. However, as the rural urban codes become increasing more rural (nonmetro), we can see flunctuations in the mean county person income. Generally, for RUC 5 and 6, the average county person income level dropped by ~$100 as has remained that way from the 1990s to the 2010s. The more rural areas, 7, 8, and 9, saw drops too. 8 had a similar trend to 5 and 6, but instead of dipping in the 90s, it dipped in the 2000s and stayed there for the 2010s. 7 and 9 show gradual dips in the average county personal income in the 1990s and 2000s, but 7 recovers to pre 1990 averages by 2010. While 9 does make a recovery, the 2010 levels don’t reach pre-1990 averages. Regardless, the averages for all codes only vary by ~$50-100.

**Q3: What difficulties do you encounter in comparing personal income between areas over time?**

The personal incomes are difficult to compare because we only have one measure per decade. I had to summarize an entire decade based off of one yearly measure, which doesn’t show yearly flunctuation. Ok and why is that important? In addition, I had to take averages because there were just too many values to accurately portray the personal income. Additionally, I chose not to look at fips\_code, which would have given me individual counties by state. I don’t fully understand why you didn’t just join on fips I grouped the counties by their rural urban codes because it would be more summaritive and give insight into how population and city structures influence county personal income level. Another issue is that counties are often redrawn over time, meaning that there fips\_code might change. Yes that is correct. So some counties will be impossible to join on their fips code. But not using the fips\_code, I can get a better general picture of metro counties vs nonmetro counties as well as individual variation between those types of counties as they become increasingly urban or increasingly rural. Well you can’t plot on the county level anymore, so you lose a lot of variation