DA5020 - Week 10 SQLite and comparing dplyr to SQL

2018-03-25

This week you are responsible for chapters 10, 11, 12 in the "Data Collection, Integration and Analysis" textbook. Review each chapter separately and work through all examples in the text BEFORE starting the assignment. You will use the schema you developed in homework 6 to store data in SQLite.

This week's assignment you use the relational schema you designed in week 6 and store data into the SQLite relational database system. Load the Unemployment and Educational data files into R studio. One file contains yearly unemployment rates from 1970 to 2015, for counties in the US. The other file contains aggregated data percentages on the highest level of education achieved for each census member. The levels of education are: "less than a high school diploma", "high school diploma awarded", "attended some college", "college graduate and beyond". The census tracks the information at the county level and uses a fips number to represent a specific county within a U.S. state. The fips number is a 5 digit number where the first two digits of the fips number represents a U.S. state, while the last three digits represent a specific county within that state.

Questions

1. Revisit the census schema you created for homework 6. After installing SQLite, implement the tables for your database design in SQLite and load the data into the correct tables using either SQL INSERT statements or CSV loads. Make sure the database design is normalized (at least 3NF) and has minimal redundancy. Make sure your SQLite tables have primary keys as well as foreign keys for relationships. (20 points)

```r
library(RSQLite)
library(tidyverse)

education <-  read_csv("FipsEducationsDA5020.csv") %>%
  spread(key = percent_measure, value = percent ) %>%
  separate(county_state, into = c("state","county"))
```

```
## Warning: Too many values at 15721 locations: 6, 7, 8, 9, 10, 11, 12, 13,
## 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, ...
```

```r
unemployment <- read_csv("FipsUnemploymentDA5020.csv")
colnames(unemployment) <- c("fipsnumber","year1","percent_unemployed")

fips<-unique(cbind.data.frame((education$fips),(education$state),(education$county)))
colnames(fips) <- c("fipsnumber","state","county")

rural_urban_code <- unique(cbind.data.frame(as.character(education$rural_urban_cont_code),as.character(
colnames(rural_urban_code) <- c("rurl_urbn_code","description")
rural_urban_code <- as_tibble(rural_urban_code)

database <- dbConnect(SQLite(), dbname="Fips.sqlite")
dbWriteTable(conn = database, name = 'education', value = education,
             row.names=F, header=T, overwrite=T)
dbWriteTable(conn = database, name = 'fips', value = fips,
             row.names=F, header=T, overwrite=T)
dbWriteTable(conn = database, name = "rural_urban", value = rural_urban_code,
             row.names=F, header=T, overwrite=T)
```

```
dbWriteTable(conn = database, name = "unemployment", value = unemployment,
             row.names=F, header=T, overwrite=T)
dbListTables(database)
```

```
## [1] "education"         "education1"        "education_statics"
## [4] "fips"              "rural_urban"       "unemployment"
```

Normalization

```
education1 <- unique(cbind.data.frame(as.character(education$fips), as.character(education$year),
                                             as.character(education$rural_urban_cont_code), as.characte
                                             as.character(education$percent_has_some_college), as.chara

colnames(education1) <- c("fips","year", "ru_code","percent_four_plus_years_college",
                                 "percent_has_some_college","percent_hs_diploma","percent_less_than_hs_

dbWriteTable(conn = database, name = "education1", value = education1, row.names = F, overwrite=T)

head(dbReadTable(database,"education1"))
```

```
##   fips year ru_code percent_four_plus_years_college
## 1 1000 1970    NULL                             7.8
## 2 1000 1980    NULL                            12.2
## 3 1000 1990    NULL                            15.7
## 4 1000 2000    NULL                              19
## 5 1000 2015    NULL                            23.5
## 6 1001 1970       2                             6.4
##   percent_has_some_college percent_hs_diploma percent_less_than_hs_diploma
## 1                      7.5               25.9                         58.7
## 2                     12.5               31.8                         43.5
## 3                     21.7               29.4                         33.1
## 4                     25.9               30.4                         24.7
## 5                     29.7                 31                         15.7
## 6                      7.7               31.1                         54.8
```

```
head(dbGetQuery(database,"select fips,year,state,county,ru_code,description,percent_four_plus_years_coll
                      from education1
                      INNER JOIN fips on education1.fips=fips.fipsnumber
                      INNER JOIN rural_urban on education1.ru_code=rural_urban.rurl_urbn_code"),10)
```

```
##    fips year state  county ru_code
## 1  1000 1970    AL Alabama    NULL
## 2  1000 1980    AL Alabama    NULL
## 3  1000 1990    AL Alabama    NULL
## 4  1000 2000    AL Alabama    NULL
## 5  1000 2015    AL Alabama    NULL
## 6  1001 1970    AL Autauga       2
## 7  1001 1980    AL Autauga       2
## 8  1001 1990    AL Autauga       2
## 9  1001 2000    AL Autauga       2
## 10 1001 2015    AL Autauga       2
##                                                    description
## 1                                                         NULL
## 2                                                         NULL
## 3                                                         NULL
## 4                                                         NULL
## 5                                                        NULL
## 6  Counties in metro areas of 250,000 to 1 million population
## 7  Counties in metro areas of 250,000 to 1 million population
## 8  Counties in metro areas of 250,000 to 1 million population
## 9  Counties in metro areas of 250,000 to 1 million population
## 10 Counties in metro areas of 250,000 to 1 million population
##    percent_four_plus_years_college percent_has_some_college
## 1                              7.8                      7.5
## 2                             12.2                     12.5
## 3                             15.7                     21.7
## 4                               19                     25.9
## 5                             23.5                     29.7
## 6                              6.4                      7.7
## 7                             12.1                     12.1
## 8                             14.5                     23.5
## 9                               18                     26.9
## 10                            23.2                     30.4
##    percent_hs_diploma percent_less_than_hs_diploma
## 1                25.9                         58.7
## 2                31.8                         43.5
## 3                29.4                         33.1
## 4                30.4                         24.7
## 5                  31                         15.7
## 6                31.1                         54.8
## 7                35.2                         40.6
## 8                  32                           30
## 9                33.8                         21.3
## 10               33.5                         12.8
```

2. Write SQL expressions to answer the following queries: (40 points)

- 2.0 In the year 1970, what is the population percent that did not earn a high school diploma for the Nantucket county in Massachusetts? What about the year 2015?

```
dbGetQuery(database, "select year,state,county,[percent_less than_hs_diploma]
           from education
           where (year = 1970 and county = 'Nantucket')")
```

```
##   year state    county percent_less than_hs_diploma
## 1 1970    MA Nantucket                         33.7
```

```
dbGetQuery(database, "select year,state,county,[percent_less than_hs_diploma]
           from education
           where (year = 2015 and county = 'Nantucket')")
```

```
##   year state    county percent_less than_hs_diploma
## 1 2015    MA Nantucket                          5.2
```

33.7% is the percent of the population not attaining a high school diploma for the Nantucket county in Massachusetts in 1970. 5.2% is the percent of the population not attaining a high school diploma for the Nantucket county in Massachusetts in 2015.

- 2.1 What is the average population percentage that did not earn a high school diploma for the counties in Alabama for the year 2015?

```
head(dbGetQuery(database, "select year,state,county,AVG([percent_less than_hs_diploma])
           from education
           WHERE (year = 2015 and state = 'AL')
           GROUP BY county"))
```

```
##   year state  county AVG([percent_less than_hs_diploma])
## 1 2015    AL Alabama                                15.7
## 2 2015    AL Autauga                                12.8
## 3 2015    AL Baldwin                                10.5
## 4 2015    AL Barbour                                26.7
## 5 2015    AL    Bibb                                19.3
## 6 2015    AL  Blount                                21.5
```

- 2.2 What is the average percentage of college graduates for the counties in the state of Massachusetts for the year 2015?

```
head(dbGetQuery(database,"select year,state,county,AVG(percent_four_plus_years_college)
           from education
           where (year = 2015 and state = 'MA')
           GROUP BY county"))
```

```
##   year state     county AVG(percent_four_plus_years_college)
## 1 2015    MA Barnstable                                 40.1
## 2 2015    MA  Berkshire                                 31.6
## 3 2015    MA    Bristol                                 25.9
## 4 2015    MA      Dukes                                 40.3
## 5 2015    MA      Essex                                 37.5
## 6 2015    MA   Franklin                                 35.2
```

- 2.3 Determine the average percentage of the population that did not earn a high school diploma for the counties in Alabama for each year within the dataset. The result should return the calendar year and the average percentage drop out rate for that year.

```
dbGetQuery(database, "select year,AVG([percent_less than_hs_diploma]) as Avg_drop_out
           from education
           where state = 'AL'
           GROUP BY year")
```

```
##   year Avg_drop_out
## 1 1970     65.15882
## 2 1980     50.62059
## 3 1990     40.10000
## 4 2000     30.26471
## 5 2015     19.75882
```

- 2.4 What is the most common rural_urban code for the U.S. counties?

```
dbGetQuery(database,"select rural_urban_cont_code, count(*) as rur_urb_count
           from education
           GROUP BY rural_urban_cont_code
           ORDER BY count(*) DESC ")
```

```
##    rural_urban_cont_code rur_urb_count
## 1                      6          2961
## 2                      7          2165
## 3                      1          2153
## 4                      9          2091
## 5                      2          1890
## 6                      3          1779
## 7                      8          1097
## 8                      4          1070
## 9                      5           460
## 10                  NULL           255
```

6 is the most common rural_urban code for US counties

- 2.5 Which counties have not been coded with a rural urban code? Return a result that contains two fields: County, State for the counties that has not been assigned a rural urban code. Do not return duplicate values in the result. Order the result alphabetically by state.

```
head(dbGetQuery(database, "select state, county
           from education
           where rural_urban_cont_code = 'NULL'
           GROUP BY state, county"))
```

```
##   state     county
## 1    AK     Alaska
## 2    AL    Alabama
## 3    AR   Arkansas
## 4    AZ    Arizona
## 5    CA California
## 6    CO   Colorado
```

- 2.6 What is the minimal percentage of college graduates for the counties in the state of Mississippi for the year 2010? There is no data available for the year 2010 in this dataset

```
dbGetQuery(database, "select year,state,county,MIN(percent_four_plus_years_college)
           from education
           where (year = 2010 and state = 'MS')")
```

```
##   year state county MIN(percent_four_plus_years_college)
## 1   NA  <NA>   <NA>                                   NA
```

- 2.7 Which state contains the most number of counties that have not been provided a rural urban code?

```
head(dbGetQuery(database, "select state, county, count(*) as Total
           from education
           where rural_urban_cont_code = 'NULL'
           GROUP BY state, county"))
```

```
##   state     county Total
## 1    AK     Alaska     5
## 2    AL    Alabama     5
## 3    AR   Arkansas     5
## 4    AZ    Arizona     5
## 5    CA California     5
## 6    CO   Colorado     5
```

- 2.8 In the year 2015, which fip counties, U.S. states contain a higher percentage of unemployed citizens than the percentage of college graduates? List the county name and the state name. Order the result alphabetically by state.

```
head(dbGetQuery(database, "select state,county,percent_four_plus_years_college, percent_unemployed
           from education
           inner join unemployment on unemployment.fipsnumber = education.fips
           where (year = 2015 and(percent_unemployed > percent_four_plus_years_college))
           Group by year ,state, county"))
```

```
##   state    county percent_four_plus_years_college percent_unemployed
## 1    AK    Bethel                             11.6                16.1
## 2    AK  Kusilvak                              5.0                23.8
## 3    AK      Lake                             12.9                16.3
## 4    AK Northwest                             10.6                17.0
## 5    AK     Yukon                             11.2                18.9
## 6    AL   Barbour                             12.5                14.3
```

- 2.9 Return the county, U.S. state and year that contains the highest percentage of college graduates in this dataset?

```
dbGetQuery(database,"select year,county,state, MAX(percent_four_plus_years_college) as Max_College_gradu
           from education")
```

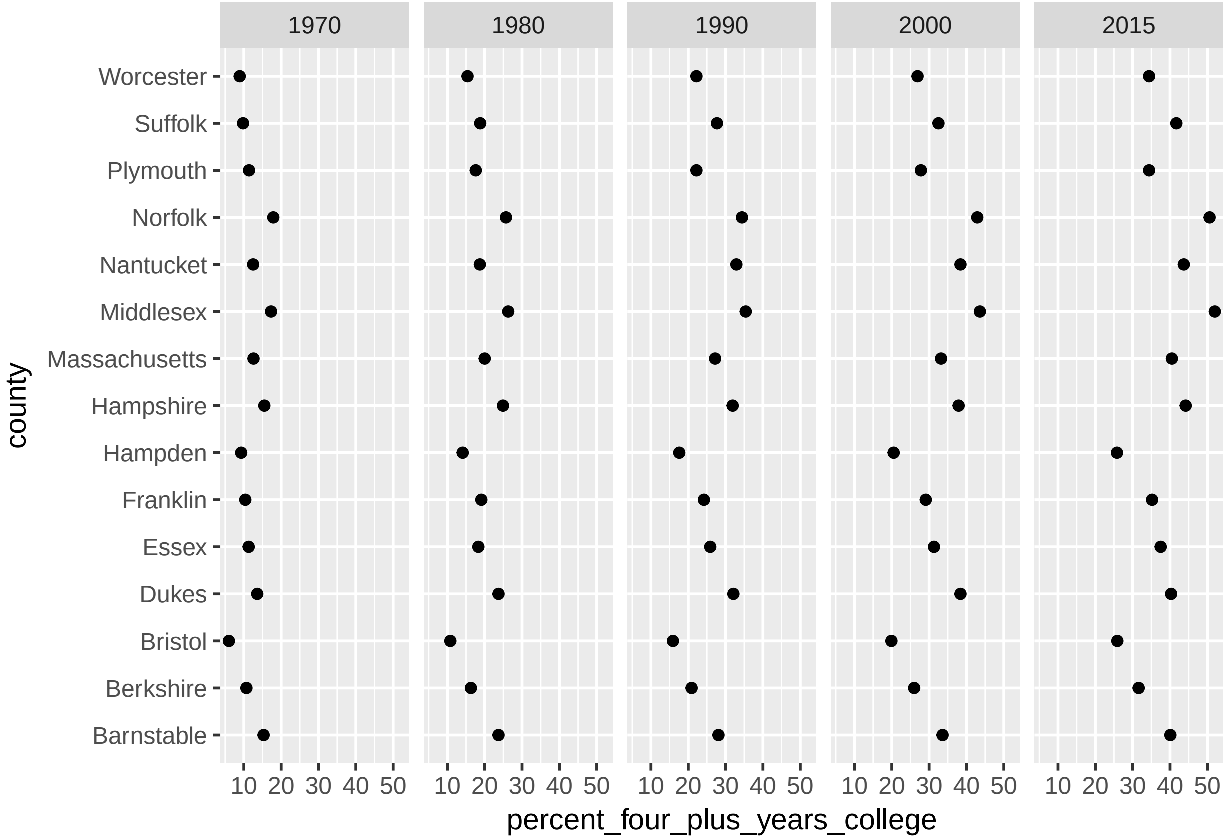
```
##   year county state Max_College_graduates
## 1 2015  Falls    VA                  78.8
```

3. Compare your SQL SELECT statements to your dplyr statements written to answer the same questions. Do you have a preference between the two methods? State your reasons for your preference. (10 points)

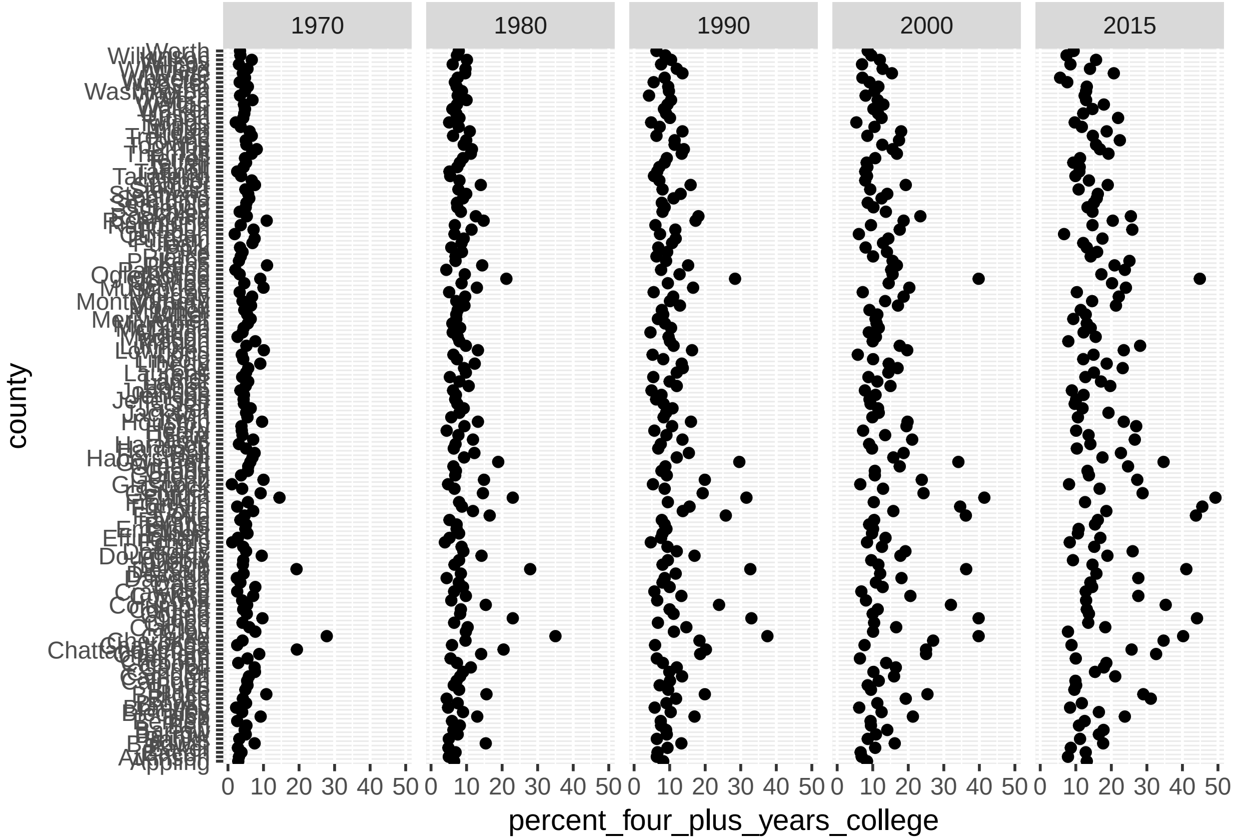
I prefer SQL over dplyr statements as in SQL, 3N is used which prevents redundancy in the data. WHERE is very easy to use There seem to be more options for selecting the parameters in the query in SQLite as compared to dplyr

4. Write a R function named get_state_county_education_data_dplyr(edf, state), it accepts a data frame containing education data and a state's abbreviation for arguments and produces a chart that shows the change in education across time for each county in that state. Use dplyr to extract the data. Write a few R statements that call the function with different state values. (5 points)

```
get_state_county_education_data_dplyr <- function(edf, State) {

  df <-   edf %>%
    filter(state == State) %>%
    select(year,county,state,percent_four_plus_years_college, percent_has_some_college,
           percent_hs_diploma,`percent_less than_hs_diploma`)

  ggplot(data = df)+
    geom_point(mapping = aes(x = percent_four_plus_years_college, y = county))+
    facet_wrap(~ year, nrow =  1)

}

get_state_county_education_data_dplyr(edf =  education, State = 'MA')
```

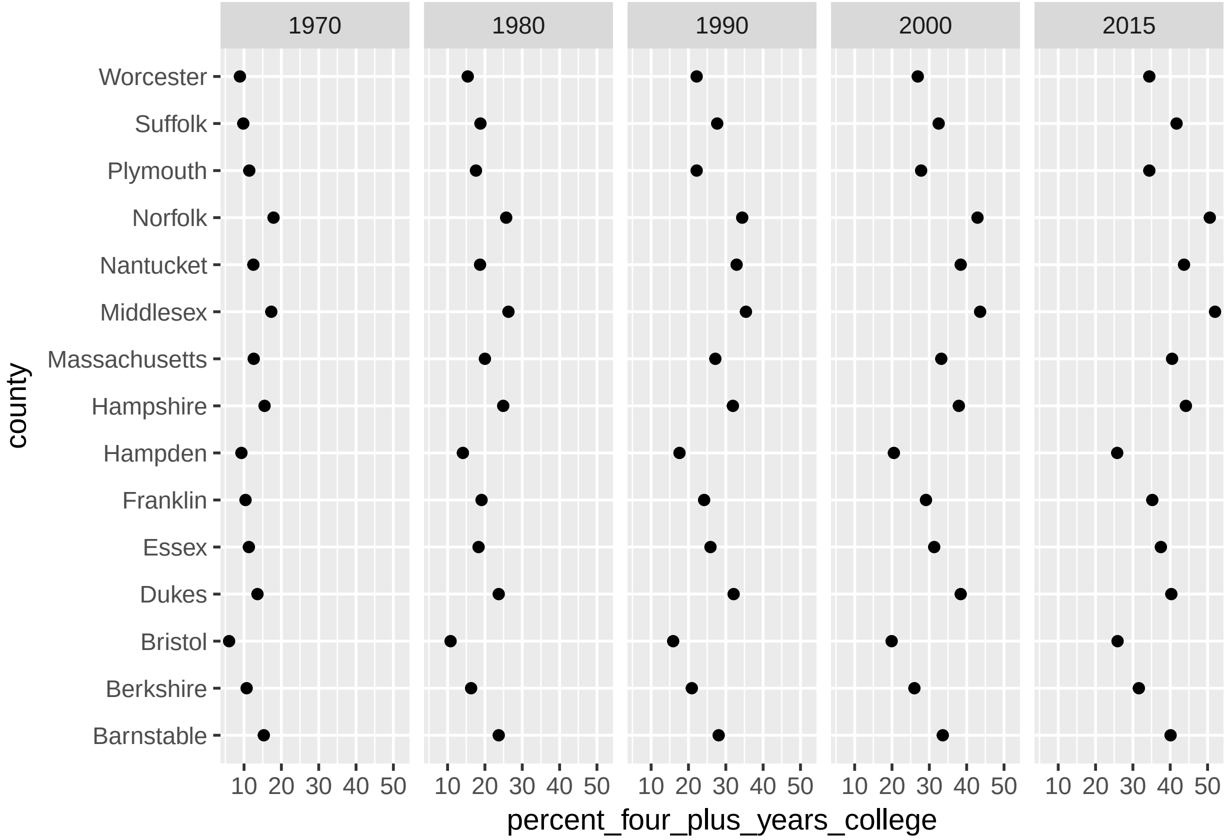


```
get_state_county_education_data_dplyr(edf =  education, State = 'GA')
```

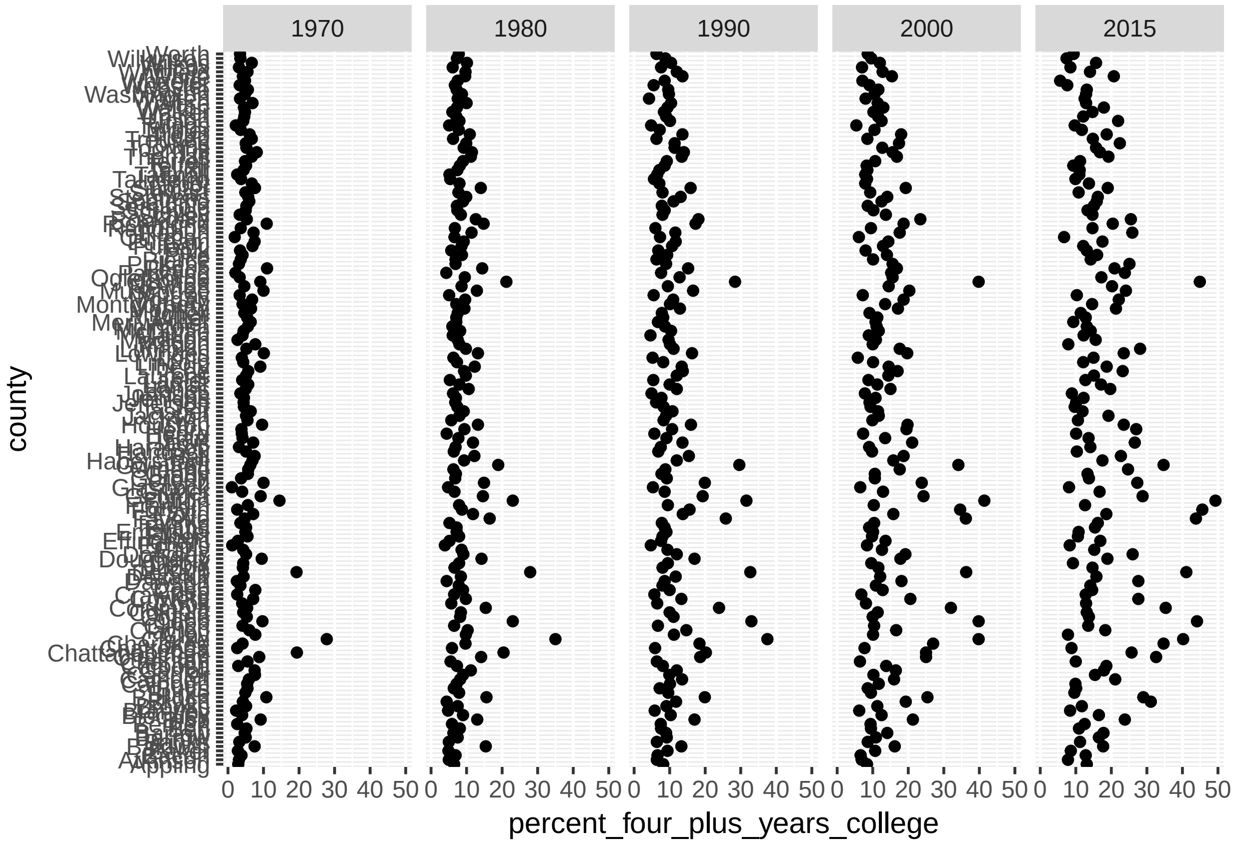


5. Write a R function named get_state_county_education_data_sql(edSQL, state), it accepts a SQL database connection containing education data and a state's abbreviation for arguments and produces a chart that shows the change in education across time for each county in that state. Use SQL SELECT to extract the data from the database. Write a few R statements that call the function with different state values. (10 points)

```
get_state_county_education_data_sql <- function(edf, State) {

  d <- "select year,county,percent_four_plus_years_college from %s where state = '%s' "
  d <- sprintf(d,edf,State)
  d <-  dbGetQuery(database,d)

  ggplot(data = d)+
    geom_point(mapping = aes(x = percent_four_plus_years_college, y = county))+
    facet_wrap(~year, nrow = 1)

}
get_state_county_education_data_sql(edf =  'education', State = 'MA')
```

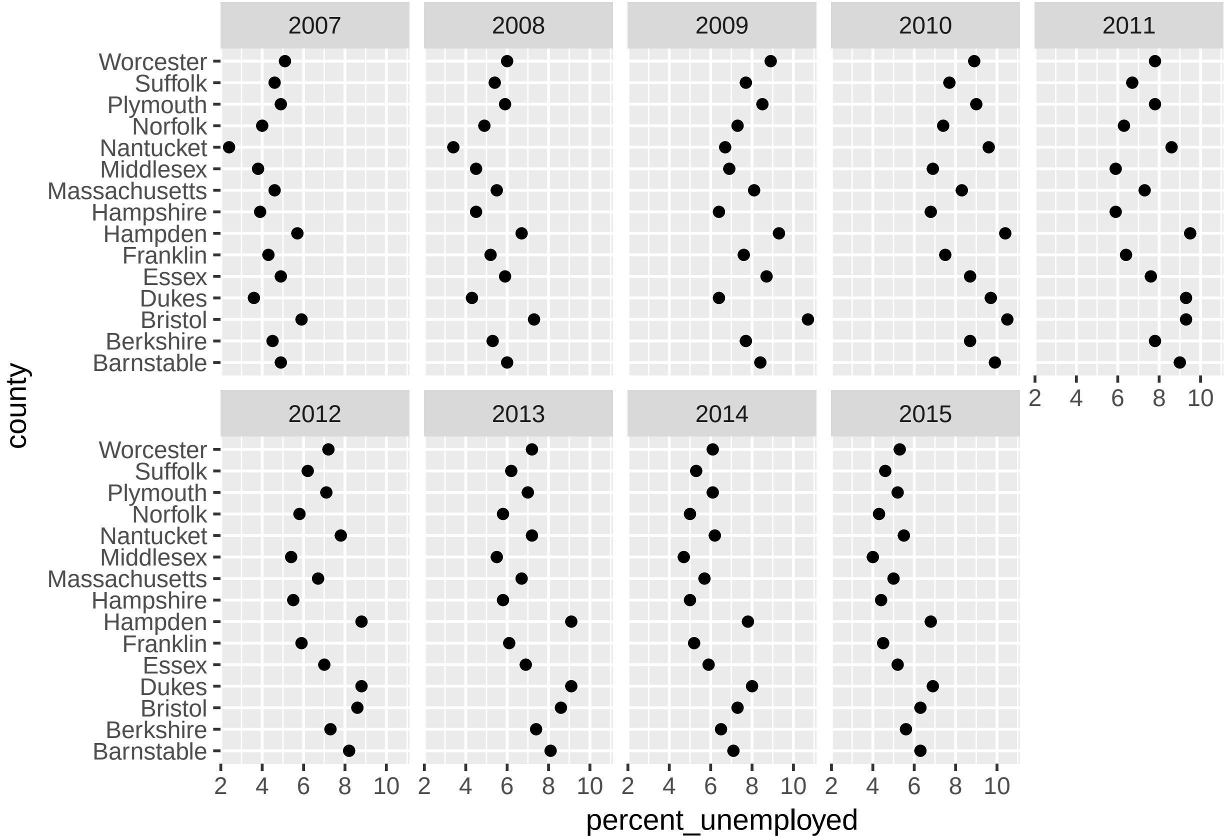


```
get_state_county_education_data_sql(edf =  'education', State = 'GA')
```

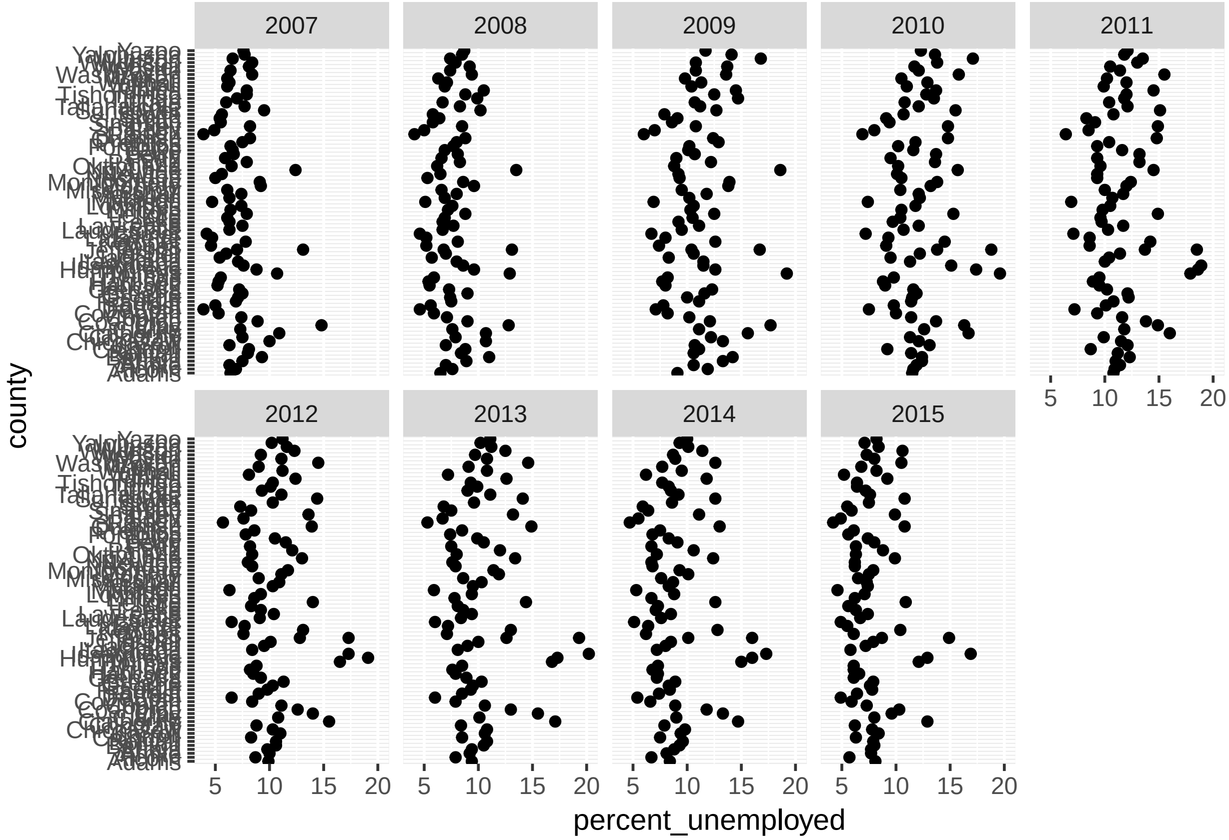


6. Write a R function named get_state_county_unemployment_data_dplyr(udf, state), it accepts a data frame containing unemployment data and state's abbreviation and produces a chart that shows the change in unemployment across time for each county in that state. Use dplyr to extract the data. Write a few R statements that call the function with different state values. (5 points)

```
fips_un <- inner_join(fips,unemployment)
get_state_county_unemployment_data_dplyr <- function(udf, State) {

  df1 <-   fips_un %>%
    filter(state == State) %>%
    select(year1,state,county,percent_unemployed)

  ggplot(data = df1)+
    geom_point(mapping = aes(x = percent_unemployed, y = county))+
    facet_wrap(~year1, nrow =  2)

}

get_state_county_unemployment_data_dplyr(udf = fips_un, State = 'MA' )
```

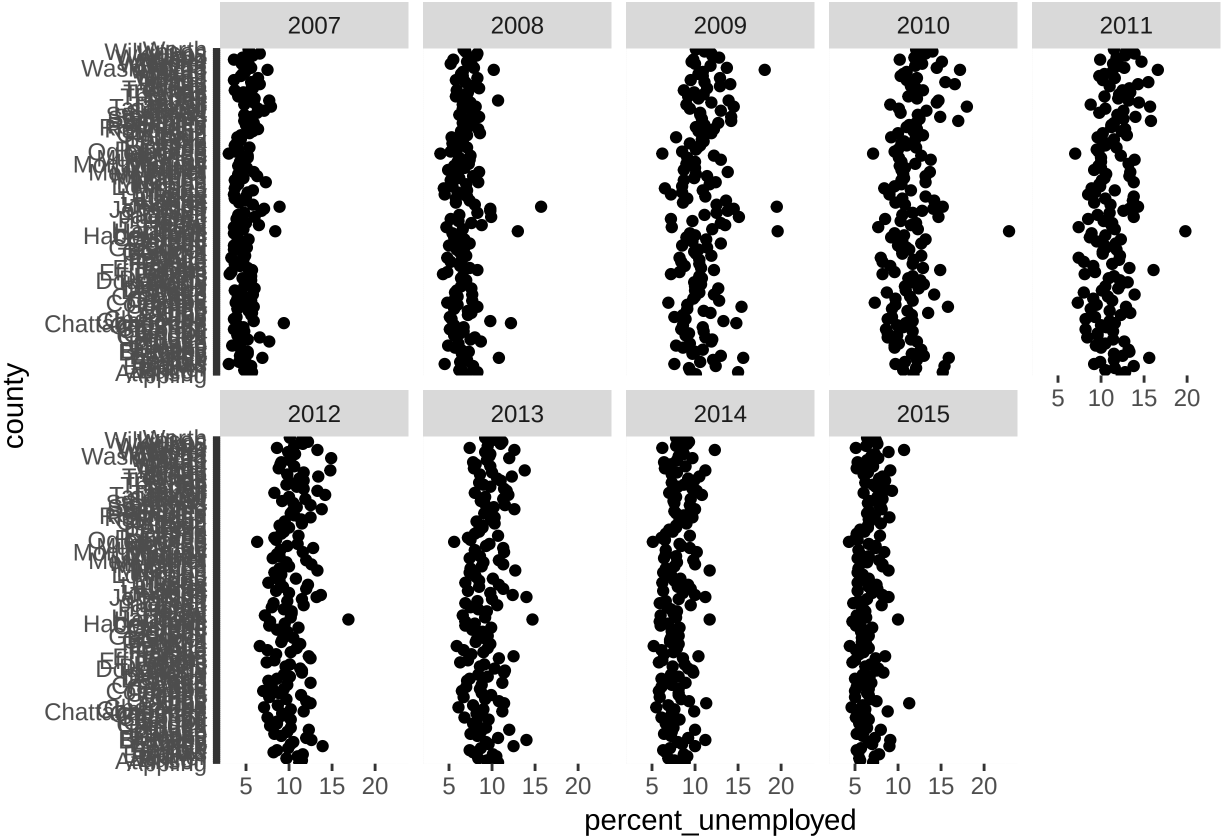


```r
get_state_county_unemployment_data_dplyr(udf = fips_un, State = 'MS' )
```



7. Write a R function named get_state_county_unemployment_data_sql(udfSQL, state), it accepts a SQL database oject containing unemployment data and state's abbreviation and produces a chart that shows the change in education across time for each county in that state. Use SQL SELECT to extract the data. Write a few R statements that call the function with different state values. (10 points)

```
get_state_county_unemployment_data_sql <- function(udf, State) {

  df2 <- "select year1,state,county,percent_unemployed from %s a inner join fips f
  on a.fipsnumber = f.fipsnumber where state = '%s' "

  df2 <- sprintf(df2, udf, State)

  df2 <- dbGetQuery(database, df2)

  ggplot(data = df2)+
    geom_point(mapping = aes(x = percent_unemployed, y = county))+
    facet_wrap(~year1, nrow =  2)

}

get_state_county_unemployment_data_sql(udf = 'unemployment', State = 'GA')
```



Submission

You need to submit an .Rmd extension file as well as the generated pdf file. Be sure to state all the assumptions and give explanations as comments in the .Rmd file wherever needed to help us assess your submission. Please name the submission file LAST_FirstInitial_1.Rmd for example for John Smith's 1st assignment, the file should be named Smith_J_1.Rmd.