DA5020 - Week 6 Assignment Tidy and Relational Data Operations

2019-10-15

This week's assignment is about tidying up the structure of data collected by the US census. 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. (20 points) Download the unemployment and education data files from blackboard and save the files to your working directory folder. Load both the unemployment data and the education data into R. Review the education data. Identify where variable names are actually values for a specific variable. Identify when multiple rows are data for the same entity. Identify when specific columns contain more than one atomic value. Tidy up the education data using spread, gather and separate.

```
a <- read.csv("FipsEducationsDA5020.csv")
b <- read.csv("FipsUnemploymentDA5020.csv")
#install.packages("stringr")
library(stringr)
#install.packages("tidyr")
library(tidyr)


#every measurement for a year and fips is reapeted 4 times which is not good, so we use spread function
a.new <- a%>%
  spread(key=percent_measure,value=percent)

#Seperating the state and counties
a.sep <- a.new %>%
  separate(county_state, into = c("state","county"))
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 15721 rows
## [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24,
## 25, ...].
```

2. (15 points) Break apart the education data into three distinct tibbles. One tibble named education contains the education data, another tibble named fips, contains the fips number definition, and the third tibble named rural_urban_code contains the textual description of the 9 different urban to rural data descriptions. These three tibbles must be linked together to represent the relationships between the tibbles. For example, the fips table will contain 3,192 rows, where each row represents the definition of a fips number (County, State). Each row in the education table will contain the educational attainment of a spcific county. It also will contain a fips number since this data is specific to a county within a state.

```
library(dplyr)
#install.packages("tibble")
library(tibble)

#fips table
d <- select(a.sep, fips,county,state)
fips <- as_data_frame(d)%>%
  group_by(fips,county,state)  %>%
  summarize()
```

```
## Warning: `as_data_frame()` is deprecated, use `as_tibble()` (but mind the new semantics).
## This warning is displayed once per session.
```

```
#rural_urban_code table
d1 <- select(a.sep, rural_urban_cont_code,description)
rural_urban_code <- as_data_frame(d1)%>%
  group_by(rural_urban_cont_code,description)  %>%
  summarize()


#Making the education tibble
d2 <- select(a.sep,fips,percent_four_plus_years_college,percent_has_some_college,percent_hs_diploma,`pe

education <- as_data_frame(d2)%>%
  group_by(fips) %>%
  summarise_each(funs(mean),AVG.P.four_plus_years_college=percent_four_plus_years_college,AVG.P.percent_
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once per session.
```

3. (5 points) Answer the following questions about your tibbles: The fips column in the education table - is it a foreign or a primary key for the education tibble? What is the primary key for your education tibble? The rural_urban code tibble should only contain 9 rows. What is its primary key?

```
#A primary key uniquely identifies an observation in the table. For example, fips$fips is a primary in

#in my education table, there are 5 variables: fips + 4 education degrees. fips are foreign key since i

#rural_urban_cont_code is a primary key since it uniquely identifies an observation in rural_urban_code
```

4. (50 points) Write expressions to answer the following queries:

- 4.0 In the year 1970, what is the percent of the population not attaining a high school diploma for the Nantucket county in Massachusetts? What about the year 2015?

```
# Percent not attaining a high school diploma in MA and Nantucket county in 1970 and 2015
#Filter works on the rows
#select works on the columns (variables)
#group_by gathers all the same parameters in column and make them ready for other analysis by summarize

filter(a.sep, state=="MA",county=="Nantucket",year=="1970") %>%
  select(`percent_less than_hs_diploma`) %>%
head() ##33.7%
```

```
##   percent_less than_hs_diploma
## 1                         33.7
```

```
filter(a.sep, state=="MA",county=="Nantucket",year=="2015") %>%
  select(`percent_less than_hs_diploma`) %>%
head() #5.2%
```

```
##   percent_less than_hs_diploma
## 1                          5.2
```

- 4.1 What is the average percentage not receiving a high school diploma for the counties in Alabama for the year 2015?

```
s<- filter (a.sep, state=="AL",year== "2015") %>%
  select(`percent_less than_hs_diploma`)

head(mean(s$`percent_less than_hs_diploma`))
```

```
## [1] 19.75882
```

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

```
x<- filter (a.sep, state=="MA",year== "2015") %>%
  select(percent_four_plus_years_college)

head(mean(x$percent_four_plus_years_college))
```

```
## [1] 38.52667
```

- 4.3 Determine the average percentage of population not attaining 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 not attaining a high school diploma for that year.

```
filter (a.sep, state=="AL") %>%
  select(year,`percent_less than_hs_diploma`) %>%
  group_by(year) %>%
  summarise(avg.not.hs.diploma=mean(`percent_less than_hs_diploma`)) %>%
  head()
```

```
## # A tibble: 5 x 2
##    year avg.not.hs.diploma
##   <int>              <dbl>
## 1  1970               65.2
## 2  1980               50.6
## 3  1990               40.1
## 4  2000               30.3
## 5  2015               19.8
```

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

```
#we count them and then sort them as a descending order.
#code 6 is the most common one.
a.sep %>%
count(rural_urban_cont_code) %>%
arrange(desc(n))
```

```
## # A tibble: 10 x 2
##    rural_urban_cont_code     n
##    <fct>                 <int>
##  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
```

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

```
#whenever the name of county is exactly the name of state, rural urban code is NULL. for 5 years it has
q <- a.sep %>%
filter (rural_urban_cont_code=="NULL")%>%
select(state,county,rural_urban_cont_code) %>%
group_by(state,county,rural_urban_cont_code) %>%
  summarise()
q <- q[order(q$state),]#making in alphabetical order
```

- 4.6 What is the minimal percentage of college graduates for the counties in the state of Mississippi for the year 2010? What does the result represent?

```
#There is no data for year 2010, I calculate it for 2015
a.sep %>%
filter (state=="MS",year== "2015") %>%
  select(county,percent_four_plus_years_college) %>%
  arrange(desc(percent_four_plus_years_college)) %>%
  tail()
```

```
##          county percent_four_plus_years_college
## 78    Chickasaw                             10.7
## 79       Benton                             10.6
## 80 Tallahatchie                              8.5
## 81        Perry                              8.3
## 82       Greene                              8.2
## 83    Issaquena                              7.2
```

```
a.sep %>%
filter (state=="MS",year== "2015") %>%
  select(county,percent_four_plus_years_college) %>%
  summarise(min(percent_four_plus_years_college))
```

```
##   min(percent_four_plus_years_college)
## 1                                  7.2
```

```
#the mimimum percentage belongs to Issaquena which is 7.2 %
```

- 4.7 In the year 2015, which fip counties, are above the average unemployment rate? Provide the county name, U.S. state name and the unemployment rate in the result. Sort in descending order by unemployment rate.

```
v <- b %>%
  filter(year=="2015")
mean(v$percent_unemployed) #average is 5.528102
```

```
## [1] 5.528102
```

```
z <- inner_join(v,fips, by="fips")
desc.2015 <- z%>%
filter(percent_unemployed>5.528102) %>%
  arrange(desc(percent_unemployed)) %>%
  select(state,county,percent_unemployed)
```

- 4.8 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.

```
n <- filter(a.sep,year=="2015")
m <- filter(b,year=="2015") %>%
  select(fips,percent_unemployed)
l<- merge(n,m,by="fips")
k <- l %>%
  filter(percent_unemployed>percent_four_plus_years_college) %>%
  select(state,county,percent_unemployed,percent_four_plus_years_college)

k <- k[order(k$state),]#making in alphabetical order
```

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

```
a.sep %>%
  select(county,year,state,percent_four_plus_years_college) %>%
  arrange(desc(percent_four_plus_years_college)) %>%
  head()
```

```
##       county year state percent_four_plus_years_college
## 1      Falls 2015    VA                            78.8
## 2  Arlington 2015    VA                            72.9
## 3        Los 2015    NM                            64.2
## 4      Falls 2000    VA                            63.7
## 5 Alexandria 2015    VA                            61.4
## 6     Howard 2015    MD                            60.6
```

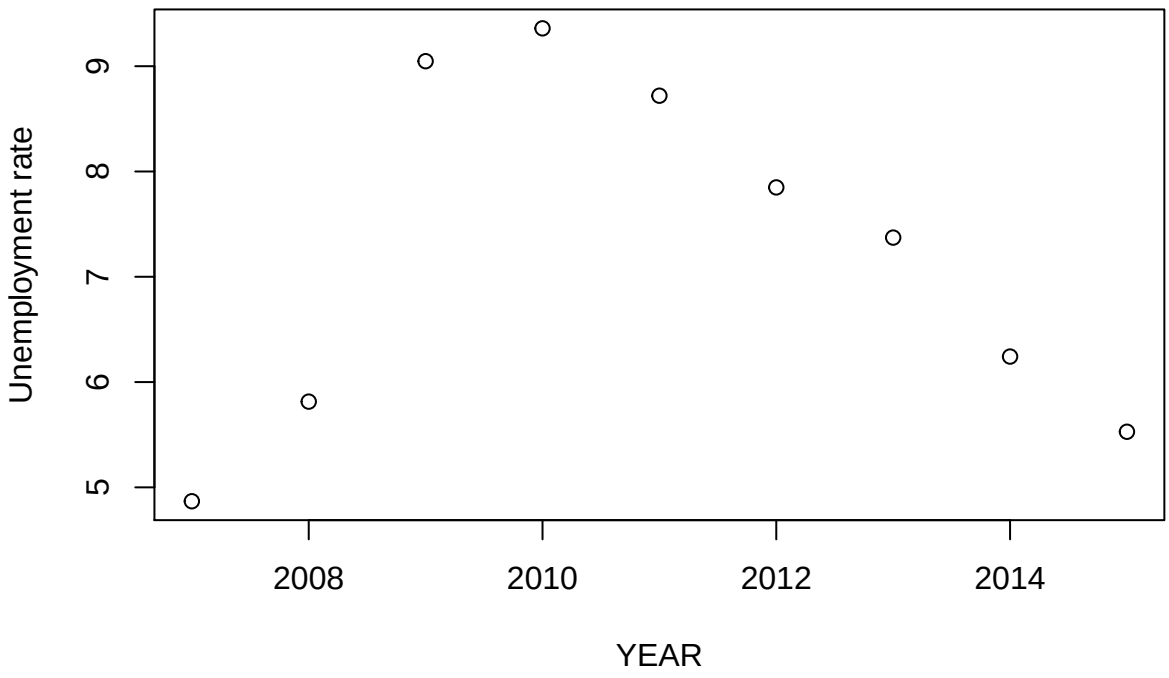
```
# The highest percentage goes to county "Falls" and state "VA" in 2015
```

5. (10 points) *Open question*: explore the unemployment rate and the percent not attaining a high school diploma over the time period in common for the two datasets. What can you discover? Create a plot that supports your discovery.

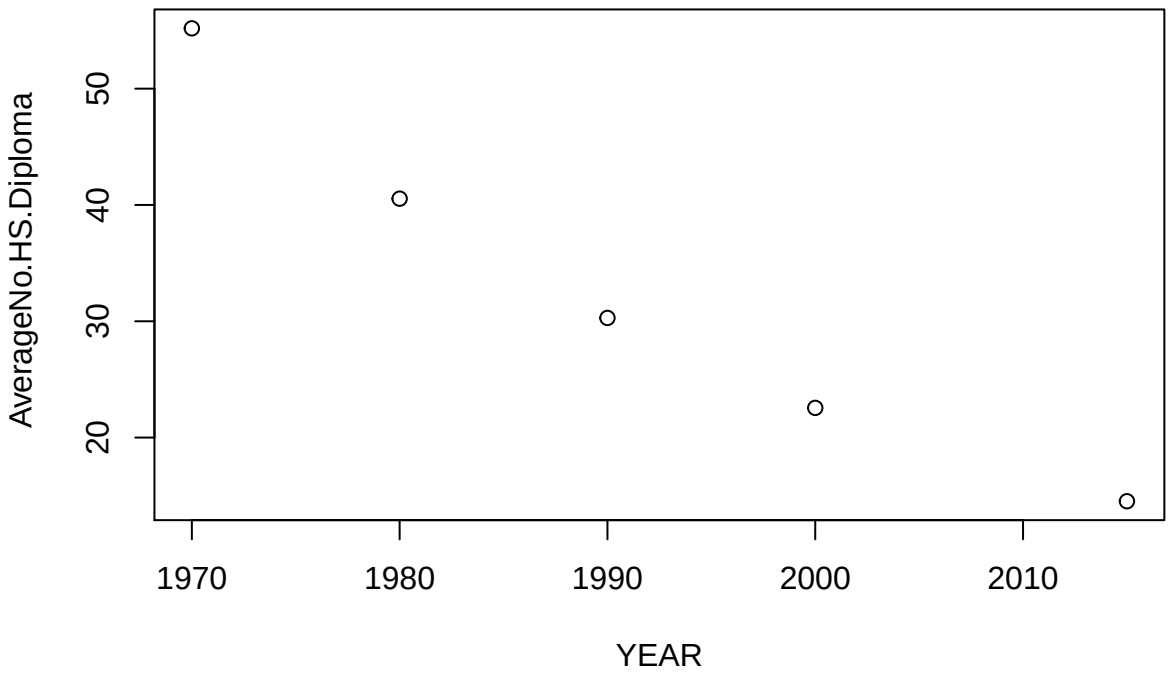
```
t <- b %>%
  group_by(year) %>%
  summarise(AVG.UNEMPLOYMENT=mean(percent_unemployed))

r <- a.sep %>%
  group_by(year) %>%
  summarise(AVG.NO.HS.Diploma=mean(`percent_less than_hs_diploma`))

plot(x=t$year,y=t$AVG.UNEMPLOYMENT,xlab="YEAR",ylab="Unemployment rate")
```




```
# we see that it is increasing at the begining and then drops between 2007 and 2015
plot(x=r$year,y=r$AVG.NO.HS.Diploma,xlab="YEAR",ylab=" AverageNo.HS.Diploma")
```




```
# It has a descending order as time goes on showing that number of educated poeople with high school di
```

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.