# DA5020 - Week 6 Assignment Tidy and Relational Data Operations

*2018-02-19*

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.

```
library(tidyverse)
Ed <- read_csv("FipsEducationsDA5020.csv")
Un <- read_csv("FipsUnemploymentDA5020.csv")

Ed1 <- spread(Ed, key = percent_measure, value = percent)
Ed2 <- separate(Ed1, 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, ...
```

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.

```
# Rename the 10th column so that it can be selected for making the tibble
names(Ed2)[10] <- "percent_less_hs_diploma"
education <- select(Ed2, fips, year, percent_four_plus_years_college, percent_has_some_college, percent_
education <- unique(education)
education <- as_tibble(education)

fips <- select(Ed2, fips, County, State)
fips <- unique(fips)
fips <- as_tibble(fips)

rural_urban_code <- select(Ed2, fips, rural_urban_cont_code, description)
```

```
rural_urban_code <- rural_urban_code[!duplicated(rural_urban_code$description), ]
rural_urban_code <- rural_urban_code[-1,] #Remove the first observation to get rid of NULL value
rural_urban_code <- as_tibble(rural_urban_code)
```

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?

Ans) The fips column in the education table is a foreign key for the education tibble. The primary key for the education tibble is the year and fips. The primary key of the rural_urban code tibble is rural_urban_count_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?

```
N1970 <- Ed2 %>% filter(year == "1970", County == "Nantucket")
N1970
```

```
## # A tibble: 1 x 10
##    fips  year State County    rural~ descript~ percen~ perce~ perc~ perce~
##   <int> <int> <chr> <chr>     <chr>  <chr>       <dbl>  <dbl> <dbl>  <dbl>
## 1 25019  1970 MA    Nantucket 7      Urban po~    12.5   12.1  41.7   33.7
```

```
# 33.7%
N2015 <- Ed2 %>% filter(year == "2015", County == "Nantucket")
N2015
```

```
## # A tibble: 1 x 10
##    fips  year State County    rural~ descript~ percen~ perce~ perc~ perce~
##   <int> <int> <chr> <chr>     <chr>  <chr>       <dbl>  <dbl> <dbl>  <dbl>
## 1 25019  2015 MA    Nantucket 7      Urban po~    43.7   25.7  25.4   5.20
```

```
# 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.

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

```
NHS <- Ed2 %>% filter(year == "2015", State == "AL") %>% select(year, State, County, percent_less_hs_di
# 19.75%
```

19.75% is the average percentage not receiving a high school diploma for the counties in Alabama for the year 2015

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

```
MCG <- Ed2 %>% filter(year == "2015", State == "MA") %>% select(year, State, County, percent_four_plus_
head(MCG)
```

```
## # A tibble: 6 x 5
##    year State County        percent_four_plus_years_college   Avg
##   <int> <chr> <chr>                                   <dbl> <dbl>
## 1  2015 MA    Massachusetts                            40.5  38.5
## 2  2015 MA    Barnstable                               40.1  38.5
## 3  2015 MA    Berkshire                                31.6  38.5
```

```
## 4  2015 MA    Bristol                                   25.9  38.5
## 5  2015 MA    Dukes                                     40.3  38.5
## 6  2015 MA    Essex                                     37.5  38.5
```

```
# 38.53%
```

38.52% is the average percentage of college graduates for the counties in the state of Massachusetts for the year 2015

- 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.

```
AHS <- Ed2 %>% filter(State == "AL") %>% select(year, State, County, percent_less_hs_diploma)  %>% grou
head(AHS)
```

```
## # A tibble: 5 x 2
##    year   Avg
##   <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? 6 is the most common rural_urban code for US counties

```
Mcruc <- Ed2 %>% group_by(rural_urban_cont_code) %>% count(rural_urban_cont_code)
# 6
head(Mcruc)
```

```
## # A tibble: 6 x 2
## # Groups:   rural_urban_cont_code [6]
##   rural_urban_cont_code     n
##   <chr>                 <int>
## 1 1                      2153
## 2 2                      1890
## 3 3                      1779
## 4 4                      1070
## 5 5                       460
## 6 6                      2961
```

- 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?

```
NEd2 <- Ed2 %>% filter(rural_urban_cont_code=="NULL")
A <- NEd2 %>% select(County, State)
A <- unique(A)
# This results represents the cases where the name of the states is the same as the county.
head(A)
```

```
## # A tibble: 6 x 2
##   County     State
##   <chr>      <chr>
## 1 Alabama    AL
## 2 Alaska     AK
```

```
## 3 Arizona    AZ
## 4 Arkansas   AR
## 5 California CA
## 6 Colorado   CO
```

- 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?

```
MICG <- Ed2 %>% filter(State == "MS", year == 2010)
# There is no data available for the year 2010 in this dataset
```

- 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.

```
AbUn <- inner_join(Ed2 %>% filter(year == 2015),Un %>% filter(year == 2015), by="fips") %>% filter(perc
head(AbUn)
```

```
## # A tibble: 6 x 3
##   County    State percent_unemployed
##   <chr>     <chr>              <dbl>
## 1 Imperial  CA                  24.0
## 2 Kusilvak  AK                  23.2
## 3 Yuma      AZ                  21.8
## 4 Yukon     AK                  18.0
## 5 Luna      NM                  17.6
## 6 Issaquena MS                  16.9
```

- 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.

```
Un2015 <- inner_join(Ed2 %>% filter(year==2015),Un %>% filter(year==2015), by="fips")
Un2015 <- Un2015 %>% filter(percent_unemployed > percent_four_plus_years_college) %>% select(State, Cou
head(Un2015)
```

```
## # A tibble: 6 x 4
##   State County    percent_unemployed percent_four_plus_years_college
##   <chr> <chr>                  <dbl>                           <dbl>
## 1 AK    Bethel                 14.4                            11.6
## 2 AK    Kusilvak               23.2                             5.00
## 3 AK    Northwest              15.5                            10.6
## 4 AK    Yukon                  18.0                            11.2
## 5 AL    Conecuh                 9.20                            8.20
## 6 AL    Greene                 11.0                            10.9
```

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

```
HPCG <- Ed2 %>% select(County, State, year, percent_four_plus_years_college)
summarise(HPCG, Highest = max(percent_four_plus_years_college))
```

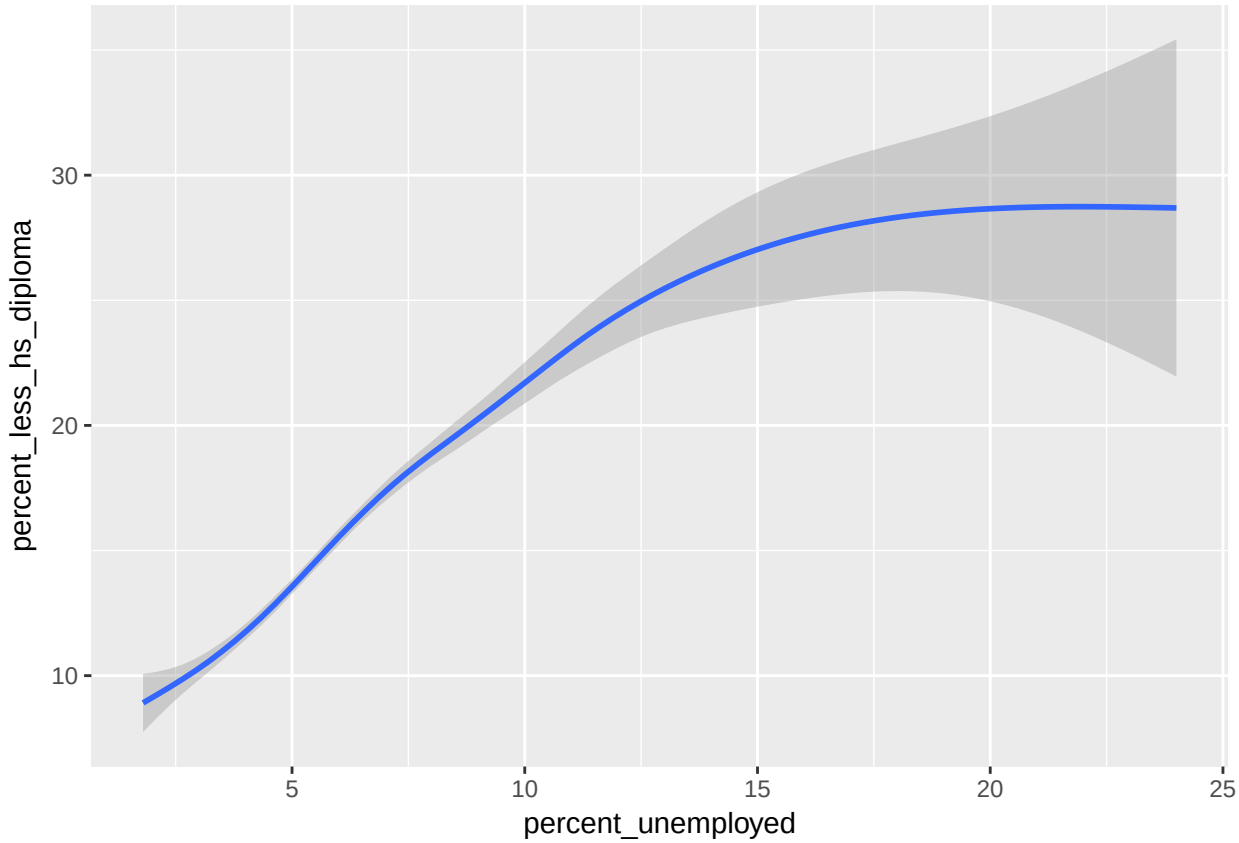
```
## # A tibble: 1 x 1
##   Highest
##     <dbl>
## 1    78.8
```

```
H <- HPCG %>% filter(percent_four_plus_years_college == 78.8)
H
```

```
## # A tibble: 1 x 4
##   County State  year percent_four_plus_years_college
##   <chr>  <chr> <int>                           <dbl>
## 1 Falls  VA     2015                            78.8
```

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.

```
UnHs <- inner_join(Ed2, Un, by = c("fips","year"))
ggplot(data = UnHs, mapping = aes(x = percent_unemployed, y = percent_less_hs_diploma))+
  geom_smooth(mapping = aes(colour = percent_less_hs_diploma))
```




```
# We can see that there is a direct correlation between the the unemployment rate and the percent not
# attaining a high school diploma.
# Higher the percent not having a high school diploma, higher the rate of unemployment.
```

## 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.