Assignment 6

Bhakti Sangoi

February 19, 2018

Reading Libraries

```
library(ggthemes)
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.1     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts ---------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
library(readxl)
```

GitHub

- https://github.com/sangoibhakti/NEU-DA5020.git

Reading files

```
education_data <- read_csv("FipsEducationsDA5020.csv")
```

```
## Parsed with column specification:
## cols(
##   fips = col_integer(),
##   year = col_integer(),
##   percent_measure = col_character(),
##   percent = col_double(),
##   county_state = col_character(),
##   rural_urban_cont_code = col_character(),
##   description = col_character()
## )
```

```
unemployment_data <- read_csv("FipsUnemploymentDA5020.csv")
```

```
## Parsed with column specification:
## cols(
##   fips = col_integer(),
##   year = col_integer(),
##   percent_unemployed = col_double()
## )
```

Q1.

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.

```
spread_unemployment<-spread(unemployment_data, key = year, value = percent_unemployed)
education<-education_data %>% separate(county_state, into = c("state", "county"))
```

```
## Warning: Too many values at 62884 locations: 21, 22, 23, 24, 25, 26, 27,
## 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, ...
```

```
spread_Edu1<-spread(education, key = year, value = percent)
```

Q2.

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.

```
#creating Education Tibble
#Education Tibble contains Fips code, Year, Percent and Rurual urban code
Edu_tibble<-spread_Edu1%>%
                select (fips,`1970`:`2015`,percent_measure,rural_urban_cont_code)%>%
                as_tibble()

#Creating Fips Tibble
#Fips Tibble is associated with Fips code, State and County
fips_tibble<-spread_Edu1%>%
                select(fips,state,county)%>%
                distinct()
seperate_fips <-separate(fips_tibble, fips,into = c("state_code", "county_code"), sep=2,    remove=FALS
              as_tibble()
```

```r
#Creating Rural Urban Code Tibble
#Rural Urban Code Tibble contains Rural Urban Code and its description
rural_urban_code_tibble<-spread_Edu1%>%
                select(rural_urban_cont_code,description)%>%
                filter(!rural_urban_cont_code=="NULL")%>%
                distinct()%>%
                as_tibble()
```

Q3.

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?

Answer:

- Edu_tibble - Education Tibble
- fips_tibble - Fips Tibble
- rural_urban_code_tibble - Rural Urban Code Tibble
- The fips column in Education Tibble is a Foreign Key. It refers to specific rows from the fibs tibble. Fips and Year combined is the composite key for Education Tibble.
- Fips column in Fibs Tibble is the primary Key.
- rural_urban_cont_code is the primary key in Rural Urban Code Tibble

Q4.

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?

```r
spread_Edu1 %>%
  select(county, percent_measure,`1970`,`2015`) %>%
#Required years i.e 1970 and 2015
  filter(percent_measure =="percent_less than_hs_diploma",county == "Nantucket")
```

```
## # A tibble: 1 x 4
##   county    percent_measure              `1970` `2015`
##   <chr>     <chr>                         <dbl>  <dbl>
## 1 Nantucket percent_less than_hs_diploma   33.7   5.20
```

```r
#Filtering on country and percent measure
```

4.1

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

```
spread_Edu1 %>%
        select(state, county, percent_measure,`2015`,rural_urban_cont_code)%>%
        filter(percent_measure =="percent_less than_hs_diploma", state == "AL",!is.na(rural_urban_cont_c
  #Filter out Rural Urban Code with NA's
  #Filter on state AL and percent measure not receiving a high school diploma
        group_by(state)%>%
        summarise(mean(`2015`))
```

```
## # A tibble: 1 x 2
##   state `mean(\`2015\`)`
##   <chr>            <dbl>
## 1 AL                19.8
```

```
  #Calculating average percentage
```

4.2

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

```
spread_Edu1 %>%
  select(state, percent_measure,rural_urban_cont_code,`2015`) %>%
  filter(percent_measure =="percent_four_plus_years_college", state == "MA", !is.na(rural_urban_cont_coc
  #Filter out Rural Urban Code with NA's
  #Filter on state MA and percent measure college graduates
  group_by(state)%>%
  summarise(mean(`2015`))
```

```
## # A tibble: 1 x 2
##   state `mean(\`2015\`)`
##   <chr>            <dbl>
## 1 MA                38.5
```

```
  #Calculating average percentage
```

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.

```
spread_Edu1 %>%
  select(state,county,rural_urban_cont_code,percent_measure,`1970`,`1980`,`1990`,`2000`,`2015`)%>%
  filter(percent_measure =="percent_less than_hs_diploma",state == "AL", !is.na(rural_urban_cont_code))
  #Filter out Rural Urban Code with NA's
  #Filter on state AL and percent measure not attaining a high school diploma
  group_by(county)%>%
  summarise(mean(`1970`),mean(`1980`),mean(`1990`),mean(`2000`),mean(`2015`))
```

```
## # A tibble: 68 x 6
##    county   `mean(\`1970\`)` `mean(\`1980\`)` `mean(\`1990\`)`
##    <chr>                <dbl>            <dbl>            <dbl>
##  1 Alabama               58.7             43.5             33.1
##  2 Autauga               54.8             40.6             30.0
##  3 Baldwin               59.4             39.7             26.8
##  4 Barbour               68.8             55.1             44.4
##  5 Bibb                  73.1             59.5             48.2
##  6 Blount                70.5             53.9             39.5
##  7 Bullock               72.9             59.0             51.0
##  8 Butler                70.5             55.0             47.2
##  9 Calhoun               58.6             43.3             32.6
## 10 Chambers              67.4             54.6             45.7
## # ... with 58 more rows, and 2 more variables: `mean(\`2000\`)` <dbl>,
## #   `mean(\`2015\`)` <dbl>
```

```
  #Calculating average for each of the calendar years
```

4.4

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

```
spread_Edu1 %>%
  group_by(rural_urban_cont_code)%>%
  #Grouping with rural urban code
  summarise(common=n())%>%
  top_n(1)
```

```
## Selecting by common
```

```
## # A tibble: 1 x 2
##   rural_urban_cont_code common
##   <chr>                  <int>
## 1 6                       2372
```

```
  #selecting and displaying only the top rural urban code
```

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?

```
spread_Edu1 %>%
  select(county,state,rural_urban_cont_code) %>%
  filter(rural_urban_cont_code=="NULL")%>%
  #Country not having rural_urban_county_code
  select(county,state)%>%
  #displaying only 2 fields
  group_by(county,state)%>%
  #grouping by County and State
  distinct()%>%
  #Returning on duplicate records
  arrange(state)
```

```
## # A tibble: 51 x 2
## # Groups: county, state [51]
##    county      state
##    <chr>       <chr>
##  1 Alaska      AK
##  2 Alabama     AL
##  3 Arkansas    AR
##  4 Arizona     AZ
##  5 California  CA
##  6 Colorado    CO
##  7 Connecticut CT
##  8 District    DC
##  9 Delaware    DE
## 10 Florida     FL
## # ... with 41 more rows
```

```
  #Arranging in alphabetical order by State
```

This result set represents the Counties that are not having rural urban code.

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?

```
education%>%
  select(state, county,percent_measure,year) %>%
  filter(state == "MS", percent_measure =="percent_four_plus_years_college", year=="2010")%>%
  select(county,year,percent_measure)%>%
  #Filter on state MS and percent measure of college graduates
  group_by(county)%>%
  arrange(county)
```

```
## # A tibble: 0 x 3
## # Groups: county [0]
## # ... with 3 variables: county <chr>, year <int>, percent_measure <chr>
```

```
#There are no counties with the above filter because year 2010 is not present in the dataset.

#Since Year 2010 is not in the dataset, calculating it for 2000 and 2015
spread_Edu1%>%
  select(state, county,percent_measure,`2000`,`2015`) %>%
  filter(state == "MS", percent_measure =="percent_four_plus_years_college")%>%
  #Filter on state MS and percent measure of college graduates
  group_by(county)%>%
  arrange(county)
```

```
## # A tibble: 83 x 5
## # Groups: county [82]
##    state county    percent_measure                 `2000` `2015`
##    <chr> <chr>     <chr>                            <dbl>  <dbl>
##  1 MS    Adams     percent_four_plus_years_college  17.5    17.8
##  2 MS    Alcorn    percent_four_plus_years_college  11.7    16.2
##  3 MS    Amite     percent_four_plus_years_college   9.40   12.1
##  4 MS    Attala    percent_four_plus_years_college  11.6    14.9
##  5 MS    Benton    percent_four_plus_years_college   7.80   10.6
##  6 MS    Bolivar   percent_four_plus_years_college  18.8    21.1
##  7 MS    Calhoun   percent_four_plus_years_college  10.2    11.2
##  8 MS    Carroll   percent_four_plus_years_college  10.9    13.6
##  9 MS    Chickasaw percent_four_plus_years_college   9.50   10.7
## 10 MS    Choctaw   percent_four_plus_years_college  11.2    13.5
## # ... with 73 more rows
```

The result represents that the Year 2010 is not present in the 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.

```
education_unemployment <- left_join(spread_Edu1, spread_unemployment, by = "fips" ) #Using Left Join, j

#Calculating average unemployment rate for the year 2015
avg <- mean(education_unemployment$`2015.y`,na.rm=TRUE)
avg <-as.double(avg)

part2<- education_unemployment %>%
   select(fips, state, county,`2015.y`) %>% distinct()

#Displaying unemployment rate of COunty,State which is more than average unemployment rate for the year
education_unemployment %>%
  select(fips,state,county,`2015.y`)%>%
  filter(`2015.y` >= avg) %>%
  inner_join(part2, by = "fips")%>%
  select(state.x,county.x,`2015.y.x`)%>%
  arrange(desc(`2015.y.x`))%>%
  distinct()
```

```
## # A tibble: 1,405 x 3
##    state.x county.x  `2015.y.x`
##    <chr>   <chr>          <dbl>
##  1 CA      Imperial        24.0
##  2 AK      Kusilvak        23.2
##  3 AZ      Yuma            21.8
##  4 AK      Yukon           18.0
##  5 NM      Luna            17.6
##  6 MS      Issaquena       16.9
##  7 AK      Northwest       15.5
##  8 CA      Colusa          15.3
##  9 AK      Hoonah          15.0
## 10 MS      Jefferson       14.9
## # ... with 1,395 more rows
```

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.

```
education_unemployment %>%
  select(fips, state,county,percent_measure,`2015.x`,`2015.y`) %>%
  filter((percent_measure =="percent_four_plus_years_college"),`2015.y`>`2015.x`) %>%
  #2015.x column represent percentage of college graduates from education dataset
  #2015.y column represent percentage of unemployed citizen
  #Filtering on unemployed citizens is greater than college graduates
  select(state,county) %>%
  arrange(state)
```

```
## # A tibble: 51 x 2
##    state county
##    <chr> <chr>
##  1 AK    Bethel
##  2 AK    Kusilvak
##  3 AK    Northwest
##  4 AK    Yukon
##  5 AL    Conecuh
##  6 AL    Greene
##  7 AL    Wilcox
##  8 AZ    Apache
##  9 AZ    Yuma
## 10 CA    Colusa
## # ... with 41 more rows
```

```
  #Arranging alphabetically by state
```

4.9

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

```
Edu_tibble %>% inner_join(fips_tibble, by = 'fips') %>%
  #joining to tibbles by fips
          select(county, state, `2015`, percent_measure) %>%
          filter(percent_measure =="percent_four_plus_years_college")%>%
          #filtering on college graduates
          arrange(desc(`2015`)) %>%
          head(1)
```

```
## # A tibble: 1 x 4
##   county state `2015` percent_measure
##   <chr>  <chr>  <dbl> <chr>
## 1 Falls  VA      78.8 percent_four_plus_years_college
```

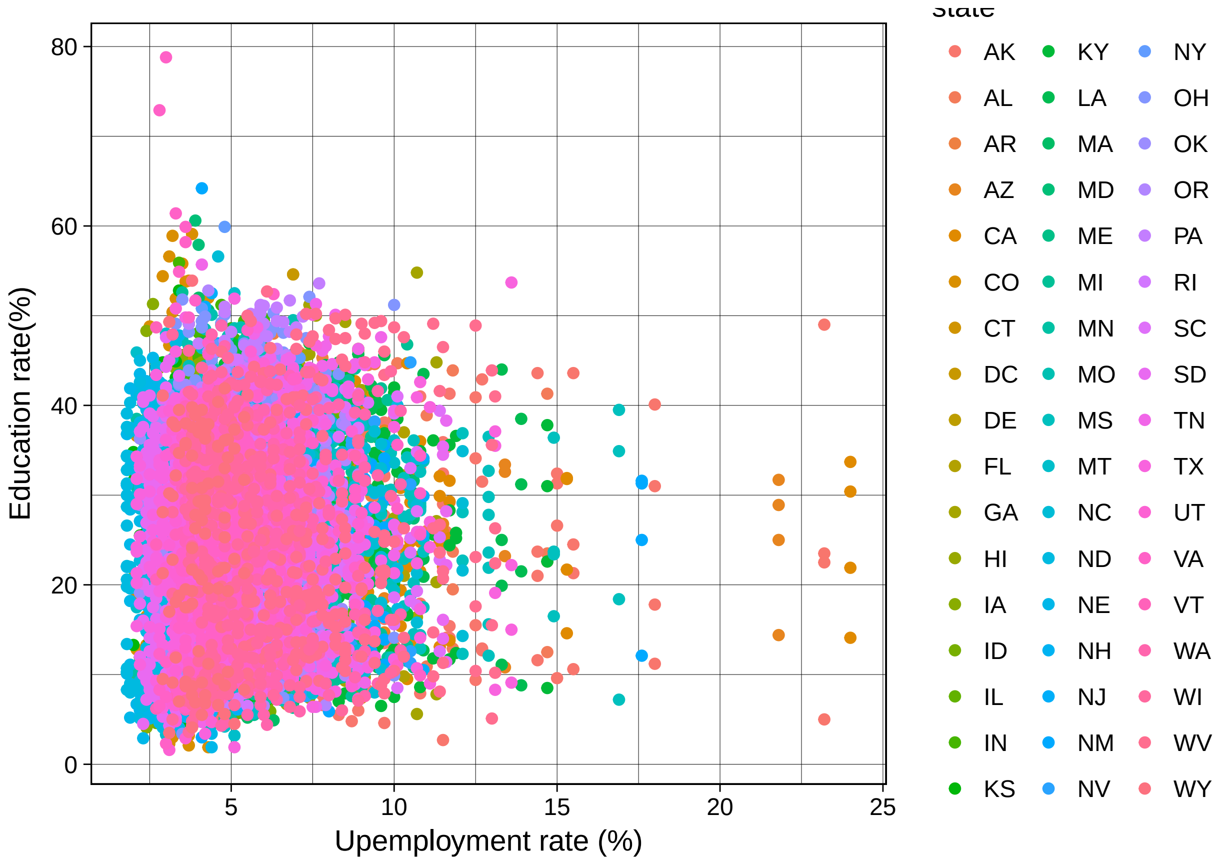
```
          #arranging in descending order and then displaying only the first row to find highest county,
```

Q5

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.

```
education_unemployment%>%
  group_by(state)%>%
  ggplot()+
  geom_point(mapping = aes(x = education_unemployment$`2015.x`, y = education_unemployment$`2015.y`,col
  theme_linedraw() +
  labs(
          x = "Education rate(%)",
          y = "Upemployment rate (%)"
      )+
  coord_flip()
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```

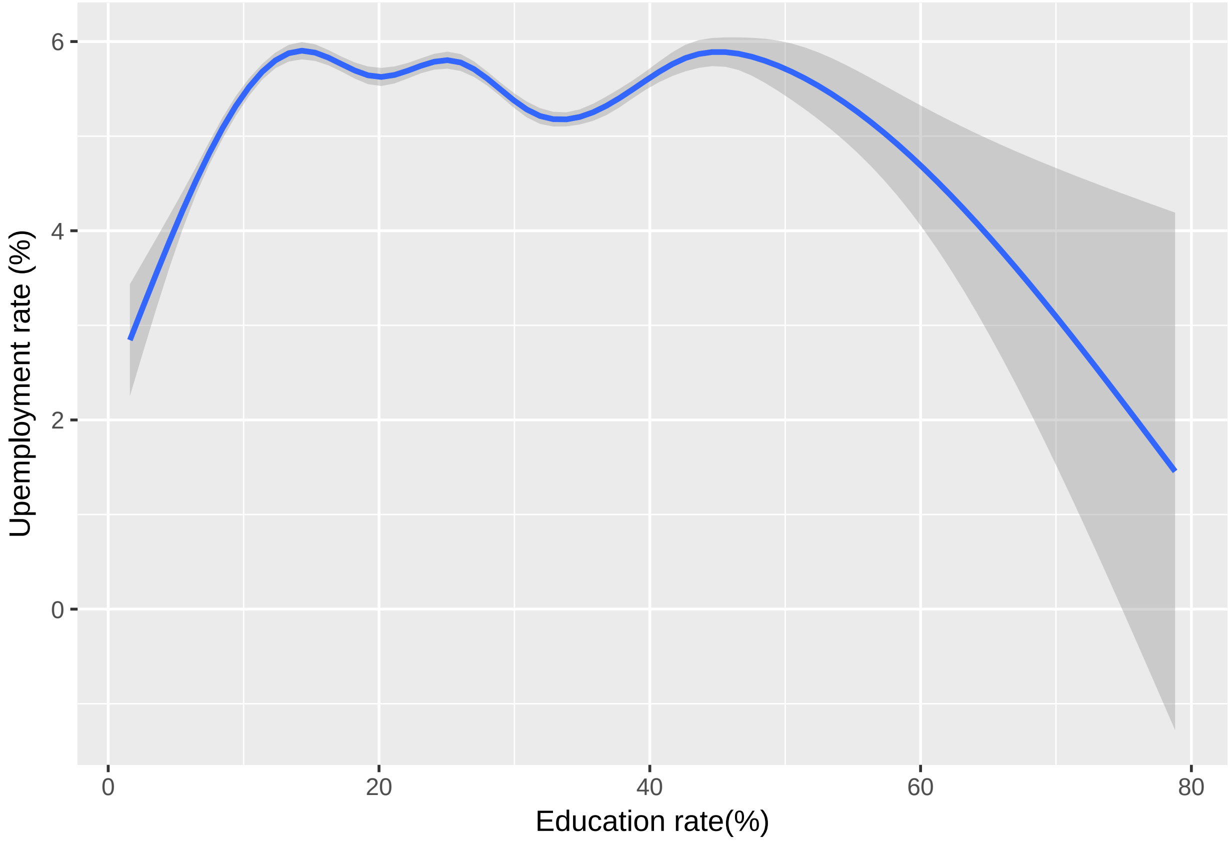


```
#the unemployment percent is highest when the education is at the lowest.


ggplot(data = education_unemployment)+
  geom_smooth(mapping = aes(x = education_unemployment$`2015.x`, y= education_unemployment$`2015.y`))+
  labs(
          x = "Education rate(%)",
          y = "Upemployment rate (%)"
      )
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 4 rows containing non-finite values (stat_smooth).
```



```
#As Education rate decreases, unemployment increases
```