Reece’s Portfolio Project

Welcome to my assessment of changes in population, and education in different regions within the US. The reason I have chosen to look into this topic is to understand the underlying causes of city growth.

I would like to understand if there is a direct correlation between education of citizens and how that effects population growth. The information I have chose to work with is from the Economic Research Services.

The data that is compiled is based on averaged amounts of data for each state and county within those states. This data is being pulled from three separate files which all maintain different types of data all relating to the same regions.

I acknowledge that there is limitations to this data when it comes to auxiliary factors like national events or nature disaster. I will be overcoming this issue by finding patterns and using a model which best represents the general trend of the data then cross referencing the anomalies with potential events that may have had an influence in that region. Another limitation is the scope of which this data contains. Due to inconsistencies we will only be making estimations between 2013 and 2017 to mitigate error due to policy chance for how the US government collected the data.

The first thing we will need to do is collect our data and import it to view the raw data we will be manipulating. In addition, we will install all the needed libraries in advance which will make it easier to analyze the findings.

```
suppressMessages(library(tidyverse))
suppressMessages(library(dplyr))
suppressMessages((library(tidyr)))
```

```
##  [1] "forcats"   "stringr"   "dplyr"     "purrr"     "readr"    
##  [6] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"    
## [11] "graphics"  "grDevices" "utils"     "datasets"  "methods"  
## [16] "base"
```

```
suppressMessages(library(naniar))
```

```
## Warning: package 'naniar' was built under R version 3.5.3
```

```
unTidyPopulationDF <- read.csv(
  "https://raw.githubusercontent.com/introdsci/DataScience-rtresendez/master/datasets/Population
Estimates.csv")

unTidyEducationDF <- read.csv(
  "https://raw.githubusercontent.com/introdsci/DataScience-rtresendez/master/datasets/Education.
csv")

unTidyUnemploymentDF <- read.csv(
  "https://raw.githubusercontent.com/introdsci/DataScience-rtresendez/master/datasets/Unemployme
nt.csv")
```

As to spare you from looking at te headache dataframes we have made I will explain what has currently been done. We have now stored the data into Dataframes. unfortunately, due to a conversation error when making the CVS file, the column names did not properly migrate. This means we will be tidying up this data to make sure everything is looking nice.

```
colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "Educational.attainment.for.adults.age.25.and.older.for.the.U.S...States..and.counties..1970.2017"] <- "FIPS_Code"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X"] <- "State"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X.1"] <- "County"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X.4"] <- "RUCC"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X.5"] <- "UCC"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X.38"] <- "LessHSD"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X.39"] <- "HSDOnly"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X.40"] <- "AADeg"

colnames(unTidyEducationDF)[colnames(unTidyEducationDF) == "X.41"] <- "BD"
```

The code above is renaming all the old columns from the untidy dataset and renaming them so it will be easier to understand what we are dealing with. Here is a quick legend that explains each of those variables

- FIPC_Code – A unique code designed to acknowledge the differences in city based on location and population by designating one of 9 numbers from Metropolitan cities to rural isolated regions.
- State – US states or territories
- County – Subregion of all US territories
- RUCC – Rural-Urban continuum code
- UCC – Urban continuum code
- LessHSD – Peoples with less than a high school diploma
- HSDOnly – Peoples with only a high school diploma
- AADeg – Peoples with some college or an AA degree
- BD – Peoples with a bachelors degree or higher

Now we will go the same process of the other two untidy datasets.

```
colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "Population.estimates.for.the.U.S...States..and.counties..2010.18..see.the.second.tab.in.this.workbook.for.variable.name.descriptions."] <- "FIPS"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X"] <- "State"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.1"] <- "County"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.3"] <- "URCC"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.5"] <- "UCC"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.12"] <- "PopEst13"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.13"] <- "PopEst14"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.14"] <- "PopEst15"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.15"] <- "PopEst16"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.16"] <- "PopEst17"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.30"] <- "Births13"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.31"] <- "Births14"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.32"] <- "Births15"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.33"] <- "Births16"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.34"] <- "Births17"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.39"] <- "Deaths13"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.40"] <- "Deaths14"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.41"] <- "Deaths15"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.42"] <- "Deaths16"

colnames(unTidyPopulationDF)[colnames(unTidyPopulationDF) == "X.43"] <- "Deaths17"
```

The above code has a few of the same variables which are easily recognized. Below is a legend for all new variables in this dataframe

- PopEst13-17 – These are the estimations of the current population between 2013 and 2017
- Births13-167 – These are the estimations for births between 2013 and 2017
- Deaths13-17 – These are the estimations of deaths between 2013 and 2017

```
colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "Unemployment.and.median.household.income.for.the.U.S...States..and.counties..2007.18"] <- "FIPS"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X"] <- "State"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.1"] <- "County"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.2"] <- "URCC"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.3"] <- "UCC"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.4"] <- "Metro_Status"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.29"] <- "Labor_Force13"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.30"] <- "Employed13"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.31"] <- "Unemployed13"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.32"] <- "UnempRt13"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.33"] <- "Labor_Force14"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.34"] <- "Employed14"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.35"] <- "Unemployed14"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.36"] <- "UnempRt14"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.37"] <- "Labor_Force15"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.38"] <- "Employed15"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.39"] <- "Unemployed15"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.40"] <- "UnempRt15"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.41"] <- "Labor_Force16"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.42"] <- "Employed16"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.43"] <- "Unemployed16"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.44"] <- "UnempRt16"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.45"] <- "Labor_Force17"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.46"] <- "Employed17"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.47"] <- "Unemployed17"

colnames(unTidyUnemploymentDF)[colnames(unTidyUnemploymentDF) == "X.48"] <- "UnempRt17"
```

Once again there are a few new variables and a few old ones. Below is the final legend for this data

- Metro_Status – Binary representation of whether a region is metro or not
- Labor_Force13-17 – Active amount of people working between 2013-2017
- Employed13-17 – Amount of people employed between 2013-2017
- Unemployed – Amount of people unemployed between 2013-2017
- unEmpRt13-17 – Unemployment rates between 2013-12017 per region

Now that we have renamed all the variables we will be using for now it is time to produce some tibbles which we will use for all data models in the future.

We will be using a total of 3 tibbles: Education, Population, and Employment.

Below is code that is going to populate our tibbles using the renamed columns from the untidy datasets

Please note that below the tibble is an expression that simply removes the specified rows from the tibble. The reason this is in here is to make sure we do not have empty rows of data that does not mean anything.

```
Education <- tibble(FIPS = unTidyEducationDF$FIPS_Code,
                    State = unTidyEducationDF$State,
                    County = unTidyEducationDF$County,
                    RUCC = unTidyEducationDF$RUCC,
                    UCC = unTidyEducationDF$UCC,
                    LessHSD = unTidyEducationDF$LessHSD,
                    HSDOnly = unTidyEducationDF$HSDOnly,
                    AADeg = unTidyEducationDF$AADeg,
                    BD = unTidyEducationDF$BD)

Education <- Education[-c(1:4),]

Population <- tibble(FIPS = unTidyPopulationDF$FIPS,
                     State = unTidyPopulationDF$State,
                     County = unTidyPopulationDF$County,
                     RUCC = unTidyPopulationDF$URCC,
                     UCC = unTidyPopulationDF$UCC,
                     PopEst13 = unTidyPopulationDF$PopEst13,
                     PopEst14 = unTidyPopulationDF$PopEst14,
                     PopEst15 = unTidyPopulationDF$PopEst15,
                     PopEst16 = unTidyPopulationDF$PopEst16,
                     PopEst17 = unTidyPopulationDF$PopEst17,
                     Births13 = unTidyPopulationDF$Births13,
                     Births14 = unTidyPopulationDF$Births14,
                     Births15 = unTidyPopulationDF$Births15,
                     Births16 = unTidyPopulationDF$Births16,
                     Births17 = unTidyPopulationDF$Births17,
                     Deaths13 = unTidyPopulationDF$Deaths13,
                     Deaths14 = unTidyPopulationDF$Deaths14,
                     Deaths15 = unTidyPopulationDF$Deaths15,
                     Deaths16 = unTidyPopulationDF$Deaths16,
                     Deaths17 = unTidyPopulationDF$Deaths17)

Population <- Population[-c(1:2),]

Employment <- tibble(FIPS = unTidyUnemploymentDF$FIPS,
                     State = unTidyUnemploymentDF$State,
                     County = unTidyUnemploymentDF$County,
                     RUCC = unTidyUnemploymentDF$URCC,
                     UCC = unTidyUnemploymentDF$UCC,
                     Metro_Status = unTidyUnemploymentDF$Metro_Status,
                     Labor_Force13 = unTidyUnemploymentDF$Labor_Force13,
                     Labor_Force14 = unTidyUnemploymentDF$Labor_Force14,
                     Labor_Force15 = unTidyUnemploymentDF$Labor_Force15,
                     Labor_Force16 = unTidyUnemploymentDF$Labor_Force16,
                     Labor_Force17 = unTidyUnemploymentDF$Labor_Force17,
                     Employed13 = unTidyUnemploymentDF$Employed13,
                     Employed14 = unTidyUnemploymentDF$Employed14,
                     Employed15 = unTidyUnemploymentDF$Employed15,
                     Employed16 = unTidyUnemploymentDF$Employed16,
                     Employed17 = unTidyUnemploymentDF$Employed17,
                     Unemployed13 = unTidyUnemploymentDF$Unemployed13,
                     Unemployed14 = unTidyUnemploymentDF$Unemployed14,
                     Unemployed15 = unTidyUnemploymentDF$Unemployed15,
                     Unemployed16 = unTidyUnemploymentDF$Unemployed16,
                     Unemployed17 = unTidyUnemploymentDF$Unemployed17,
                     UmpRt13 = unTidyUnemploymentDF$UnempRt13,
                     UmpRt14 = unTidyUnemploymentDF$UnempRt14,
                     UmpRt15 = unTidyUnemploymentDF$UnempRt15,
                     UmpRt16 = unTidyUnemploymentDF$UnempRt16,
                     UmpRt17 = unTidyUnemploymentDF$UnempRt17)

Employment <- Employment[-c(1:7),]
```

Now that we have some nicer looking tibbles we can finish up the tidying process to ensure this data is ready to be manipulated. We will be filling in the blank spaces with 0 as they are categorical variables. Since 0 is an undefined category in this dataset we are going to define it for all broader regions like the whole US or whole states.

There is one exception to this change and that is the status of whether the area is metro or not, since the US and states don’t have a given value we will be replacing that value with NA.

```
Population$RUCC <- sub("^$", "0", Population$RUCC)
Population$UCC <- sub("^$", "0", Population$UCC)

Employment$RUCC <- sub("^$", "0", Employment$RUCC)
Employment$UCC <- sub("^$", "0", Employment$UCC)
Employment$Metro_Status <- sub("^$", "NA", Employment$Metro_Status)

Education$RUCC <- sub("^$", "0", Education$RUCC)
Education$UCC <- sub("^$", "0", Education$UCC)
```

This has concluded our data Discovery and Preparation Stage. Below is a summary of what modifications have been made to create this now workable dataset.

We have created 3 Tibbles from data provided by Economic Research Services.

The tibble for Population contains 20 variables. There are 3 unique categorical variables that define the dataset. Those are State, FIPS, and County. All other variables are continuous with respect to the given region. They contain information about current population and growth and decay from natural causes.

The tibble for Employment contains 26 variables. Similar to the Population tibble they also have the same 3 unique categorical variables. The remaining variables for continuous with respect to their region. They contain information for employment status for citizens.

The tibble for Education contains 9 variables. With the same 3 categorical variables they also have continuous data with respect to their region. They contain information on current education status within a given region.

My hope is by the end of this project to answer the following questions:

Is there a correlation between the percent of a population that is educated, and the growth of that Region?

Does the Employment Rate of a region cause changes in population fluctuation, and does that effect the population of educated peoples in the area.

Is there a reasonable trend that can be seen when comparing the labor force with peoples who have a higher level of education.