Holsenbeck_S_6

Stephen Synchronicity

2017-10-15

```
this.dir <- dirname("Holsenbeck_S_6.Rmd")
setwd(this.dir)
dfEdu <- read_csv("FipsEducationsDA5020v2.csv")
dfUne <- read_csv("FipsUnemploymentDA5020.csv")
```

Homework Outline

Homework 6

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.

```
# Seperate state & county
dfEduStates <- dfEdu %>% filter(description == "NULL")
dfEdu <- dfEdu %>% filter(description != "NULL")
dfEdu <- dfEdu %>% separate(county_state, c("ST", "County"), extra = "merge")
```

```
library("forcats")
# Recode Factors in Education & Description
class(dfEdu$description)
```

```
## [1] "character"
```

```
unique(dfEdu$description)
```

```
## [1] "Counties in metro areas of 250,000 to 1 million population"

## [2] "Counties in metro areas of fewer than 250,000 population"

## [3] "Urban population of 2,500 to 19,999, adjacent to a metro area"
```

```
## [4] "Counties in metro areas of 1 million population or more"

## [5] "Completely rural or less than 2,500 urban population, not adjacent to
a metro area"
## [6] "Urban population of 2,500 to 19,999, not adjacent to a metro area"

## [7] "Completely rural or less than 2,500 urban population, adjacent to a m
etro area"
## [8] "Urban population of 20,000 or more, adjacent to a metro area"

## [9] "Urban population of 20,000 or more, not adjacent to a metro area"
```

```
levels(dfEdu$description)
```

```
## NULL
```

```
# Recode Description
dfEdu$description <- fct_recode(dfEdu$description, `metro, 1m+` = "Counties i
n metro areas of 1 million population or more",
    `metro, 250k-1m` = "Counties in metro areas of 250,000 to 1 million popul
ation",
    `metro, <250k` = "Counties in metro areas of fewer than 250,000 populatio
n",
    `urbanmetro, 20k+` = "Urban population of 20,000 or more, adjacent to a m
etro area",
    `urban, 20k+` = "Urban population of 20,000 or more, not adjacent to a me
tro area",
    `urbanmetro, 2.5-19.999k` = "Urban population of 2,500 to 19,999, adjacen
t to a metro area",
    `urban, 2.5-19.999k` = "Urban population of 2,500 to 19,999, not adjacent
 to a metro area",
    `ruralmetro, <2.5k` = "Completely rural or less than 2,500 urban populati
on, adjacent to a metro area",
    `rural, <2.5k` = "Completely rural or less than 2,500 urban population, n
ot adjacent to a metro area")
# order desc factor levels
levels(dfEdu$description) <- c("metro, 1m+", "metro, 250k-1m", "metro,
<250k", "urbanmetro, 20k+",
    "urban, 20k+", "urbanmetro, 2.5-19.999k", "urban, 2.5-19.999k", "ruralmet
ro, <2.5k",
    "rural, <2.5k")
```

```
class(dfEdu$percent_measure)
```

```
## [1] "character"
```

```
unique(dfEdu$percent_measure)
```

```
## [1] "percent_four_plus_years_college" "percent_has_some_college"
## [3] "percent_hs_diploma"              "percent_less than_hs_diploma"
```

```
dfEdu$percent_measure <- fct_recode(dfEdu$percent_measure, B = "percent_four_
plus_years_college",
    `<B` = "percent_has_some_college", HS = "percent_hs_diploma", `<HS` = "pe
rcent_less than_hs_diploma")
levels(dfEdu$percent_measure)
```

```
## [1] "B"   "<B"  "HS"  "<HS"
```

```
# Rename Cols
colnames(dfEdu)[3] <- "EduLvl"
colnames(dfEdu)[7] <- "RUCC"
colnames(dfUne)[3] <- "p.Une"
# Spread EduLvl
dfEdu <- dfEdu %>% spread(EduLvl, percent)
```

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.

```
Edu <- as.tibble(unique(dfEdu[, c(1:2, 7:10)]))
RUCC <- as.tibble(unique(dfEdu[, c(5:6)]))
fips <- as.tibble(unique(dfEdu[, c(1, 3:4)]))
```

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. The fips column in the education table is a foreign key that relates the education statistics to the fips primary key in the fips table that identifies the county associated with the fips number (the table is 3142 rows with the 50 States rows removed).
The primary key for the education tibble is the fips number and the year.
The RUCC tibble primary key is the RUCC number .

4.

(40 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?

A. In 1970: 33.7, in 2015: 5.2

```
fips %>% filter(County == "Nantucket")  #Find the fips
```

```
## # A tibble: 1 x 3
##    fips    ST    County
##   <int> <chr>     <chr>
## 1 25019    MA Nantucket
```

```
Edu %>% filter(fips == "25019" & year %in% c("1970", "2015"))  #Show the valu
es
```

```
## # A tibble: 2 x 6
##    fips  year     B  `<B`    HS `<HS`
##   <int> <int> <dbl> <dbl> <dbl> <dbl>
## 1 25019  1970  12.5  12.1  41.7  33.7
## 2 25019  2015  43.7  25.7  25.4   5.2
```

4.1

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

A. ~19.82

```
AL <- fips %>% filter(ST == "AL")
AL.edu <- left_join(AL, Edu, by = "fips")
(AL.edu.2015 <- AL.edu %>% filter(year == "2015") %>% group_by(year) %>% summ
arise(`m<HS` = mean(`<HS`
```

```
    na.rm = T)))
```

```
## # A tibble: 1 x 2
##    year  `m<HS`
##   <int>   <dbl>
## 1  2015 19.8194
```

4.2

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

A. ~9.32

```
MA <- fips %>% filter(ST == "MA")
MA.edu <- left_join(MA, Edu, by = "fips")
(MA.edu.2015 <- MA.edu %>% filter(year == "2015") %>% group_by(year) %>% summ
arise(`m<HS` = mean(`<HS`,
    na.rm = T)))
```

```
## # A tibble: 1 x 2
##    year   `m<HS`
##   <int>    <dbl>
## 1  2015 9.321429
```

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.

```
(AL.edu.yr <- AL.edu %>% group_by(year) %>% summarise(`m<HS` = mean(`<HS`, n
a.rm = T)))
```

```
## # A tibble: 5 x 2
##    year   `m<HS`
##   <int>    <dbl>
## 1  1970 65.25522
## 2  1980 50.72687
## 3  1990 40.20448
## 4  2000 30.34776
## 5  2015 19.81940
```

4.4

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

A. RUCC:6 Description: “urbanmetro, 2.5-19.999k”

```
dfEdu %>% group_by(RUCC) %>% count(RUCC)
```

```
## # A tibble: 9 x 2
## # Groups:   RUCC [9]
##    RUCC     n
##   <chr> <int>
## 1     1  2153
## 2     2  1890
## 3     3  1779
## 4     4  1070
## 5     5   460
## 6     6  2961
## 7     7  2165
## 8     8  1097
## 9     9  2091
```

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 has not been assigned a rural urban code. Do not return duplicate values in the result. Order the result alphabetically by state.

A. There actually aren’t any counties without RUCC, these are the state-wide summary statstics for all states and the District of Columbia that the census included but did not make apparent.

```
unique(dfEduStates %>% select(county_state) %>% separate(county_state,
c("ST", "County"),
    extra = "merge"))
```

```
## # A tibble: 51 x 2
##       ST                County
##    <chr>                 <chr>
##  1    AL               Alabama
##  2    AK                Alaska
##  3    AZ               Arizona
##  4    AR              Arkansas
##  5    CA            California
##  6    CO              Colorado
##  7    CT           Connecticut
```

```
##  8    DE             Delaware
##  9    DC District of Columbia
## 10    FL              Florida
## # ... with 41 more rows
```

4.6

What is the minimal percentage of college graduates for the counties in the state of Mississippi for the year 2010?

A)Data isn't available for the year 2010 in the education data set. Are we looking at the same data set? I'm going to assume for the purposes of answering the question that by 2010, 2000 is meant. In the year 2000, the minimal college graduates for MS was 7.1% in Issaquena County

```
MS.edu <- inner_join(Edu, fips %>% filter(ST == "MS"), by = "fips") %>% filte
r(year ==
    "2000") %>% arrange(B)
min(MS.edu$B)
```

```
## [1] 7.1
```

4.7

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

A. This question doesn't make sense given the data. The only places where the RUCC are not provided are for the entire State summary entries on a per year basis.

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.

A.

```
(Une.2015 <- inner_join(dfEdu %>% filter(year == 2015), dfUne %>% filter(year
 ==
    2015), by = "fips") %>% filter(p.Une > B) %>% select(ST, County, B, p.Un
e, -year.y) %>%
    arrange(ST))
```

```
## # A tibble: 51 x 4
##       ST    County     B p.Une
##    <chr>     <chr> <dbl> <dbl>
##  1    AK    Bethel  11.6  14.4
##  2    AK  Kusilvak   5.0  23.2
##  3    AK Northwest  10.6  15.5
```

```
##  4    AK     Yukon  11.2  18.0
##  5    AL   Conecuh   8.2   9.2
##  6    AL    Greene  10.9  11.0
##  7    AL    Wilcox  12.5  14.7
##  8    AZ    Apache  10.8  13.4
##  9    AZ      Yuma  14.4  21.8
## 10    CA    Colusa  14.6  15.3
## # ... with 41 more rows
```

4.9

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

A. I'm not sure if you mean the dataset from the previous question, or the overall dataset, so I'll do both. Colusa, CA in dataset from previous question, Falls Church,VA in 2015 in the overall dataset with 78.8% college grads.

```
Une.2015 %>% mutate(r.B = dense_rank(B)) %>% filter(r.B == max(r.B))
```

```
## # A tibble: 1 x 5
##      ST County     B p.Une   r.B
##   <chr>  <chr> <dbl> <dbl> <int>
## 1    CA Colusa  14.6  15.3    40
```

```
(Top.B <- inner_join(dfEdu, dfUne, by = c("fips", "year")) %>% select(ST, Cou
nty,
    B, p.Une, everything()) %>% mutate(r.B = dense_rank(B)) %>% filter(r.B ==
 max(r.B)))
```

```
## # A tibble: 1 x 12
##      ST County     B p.Une  fips  year  RUCC  description  `<B`    HS
##   <chr>  <chr> <dbl> <dbl> <int> <int> <chr>       <fctr> <dbl> <dbl>
## 1    VA  Falls  78.8     3 51610  2015     1 metro, <250k  11.4   7.5
## # ... with 2 more variables: `<HS` <dbl>, r.B <int>
```

5.

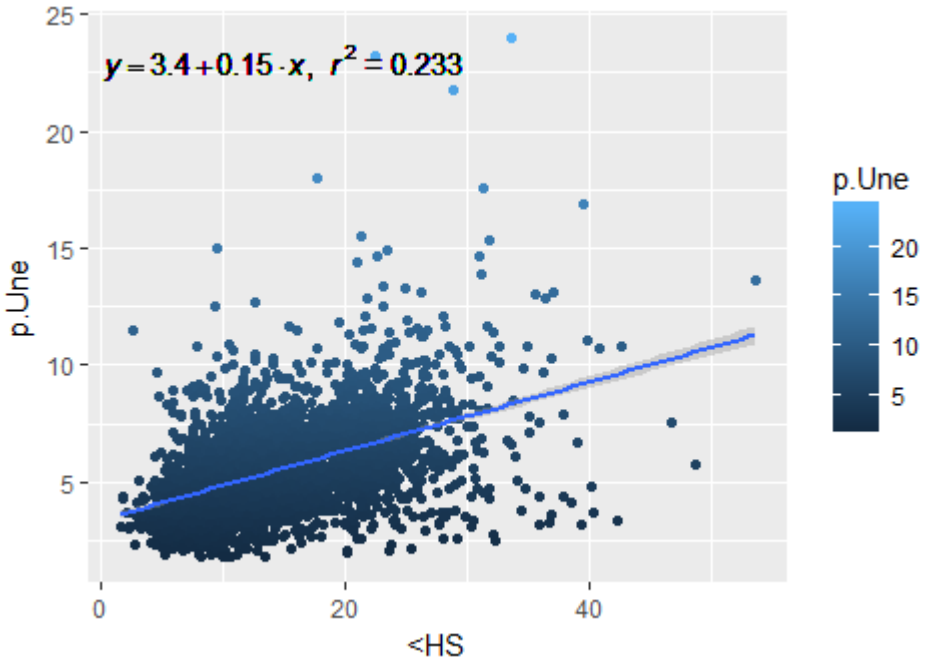
(20 points) *Open question*: explore the unemployment rate and the percent not attaining a high school diploma over the time period in common for the two datasets. What can you discover? Create a plot that supports your discovery.

A. It appears that there is correlation between percentages not attaining a HS diploma and the unemployment rate, but the relatively low R^2 R2 coeffecient indicates that the correlation is a weak one. The second graph shows that RUCC 6&8 might have a statistically significant higher

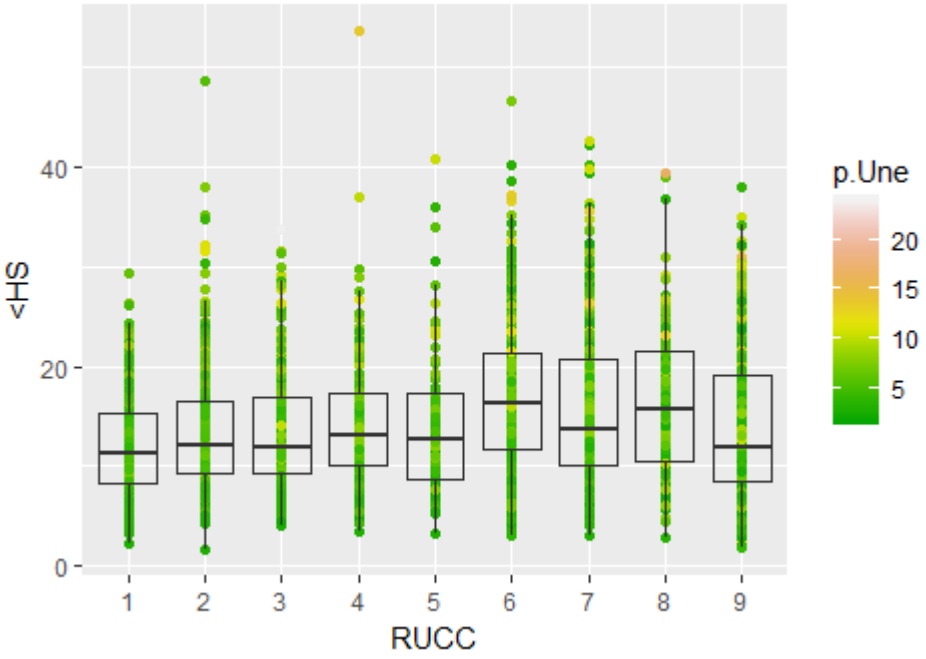
percentage of folks not finishing high school. The t-tests below indicate that this is indeed the case.

```
p.Edu <- inner_join(dfEdu, dfUne) %>% select(ST, County, p.Une, `<HS`, HS, `<
B`,
    B, everything())
```

```
# Function for linear reqression equation
lm_eqn = function(m) {

    l <- list(a = format(coef(m)[1], digits = 2), b = format(abs(coef(m)[2]),
 digits = 2),
        r2 = format(summary(m)$r.squared, digits = 3))

    if (coef(m)[2] >= 0) {
        eq <- substitute(italic(y) == a + b %.% italic(x) * "," ~
~italic(r)^2 ~
            "=" ~ r2, l)
    } else {
        eq <- substitute(italic(y) == a - b %.% italic(x) * "," ~
~italic(r)^2 ~
            "=" ~ r2, l)
    }

    as.character(as.expression(eq))
}

# Scatter plots with linear line of best fit and linear regression equation
ggplot(data = p.Edu, mapping = aes(x = `<HS`, y = p.Une)) + geom_point(mappin
g = aes(color = p.Une)) +
    geom_smooth(method = "lm") + geom_text(aes(x = 15, y = 23, label =
lm_eqn(lm(p.Une ~
    `<HS`, p.Edu))), parse = TRUE)
```




```
# Scatter plots by RUCC code with Boxplots
ggplot(data = p.Edu, mapping = aes(x = RUCC, y = `<HS`)) + geom_point(mapping
 = aes(color = p.Une)) +
    geom_boxplot(alpha = 0.01) + scale_colour_gradientn(colours = terrain.col
ors(10))
```




```
# T-tests for the two RUCC codes that appear to have higher averages
rucc8 <- p.Edu %>% filter(RUCC == "8")

rucc6 <- p.Edu %>% filter(RUCC == "6")
t.test(x = p.Edu$`<HS`, rucc8$`<HS`)
```

```
##
##  Welch Two Sample t-test
##
## data:  p.Edu$`<HS` and rucc8$`<HS`
## t = -3.2372, df = 248.57, p-value = 0.001371
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.488043 -0.605755
## sample estimates:
## mean of x mean of y
##  14.57128  16.11818
```

```
t.test(x = p.Edu$`<HS`, rucc6$`<HS`)
```

```
##
##  Welch Two Sample t-test
##
## data:  p.Edu$`<HS` and rucc6$`<HS`
## t = -8.3568, df = 836.29, p-value = 2.67e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.042477 -1.885109
## sample estimates:
## mean of x mean of y
##  14.57128  17.03508
```