Week6

Vinisha

October 13, 2017

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)

## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ---------------------------------------------
-

## filter(): dplyr, stats
## lag():    dplyr, stats

library(magrittr)

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract

library(ggplot2)
library(stringr)

#loading dataset into r
Education <- read.csv("FipsEducationsDA5020v2.csv")
Unemp <- read.csv("FipsUnemploymentDA5020(1).csv")

#Part 1 Identify where variable names are actually values for a specific
variable

Education_change <- spread(Education, key = percent_measure, value = percent)
```

```
# Combining County_state, rural_urban_count_code and description as it is
common for each set of fips
Education_change <- unite(Education_change, ruralurbancode_description,
rural_urban_cont_code, description, sep = "  _  ")


# Seprarating count_state to County and state
Education_change <- separate(Education_change, 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, ...

Education_change %>% head(10) %>%
  knitr::kable()
```

fips	year	state	county	ruralurbancode_description	percent_four_plus_years_college	percent_has_some_college	percent_hs_diploma	percent_lessthan_hs_diploma
1000	1970	AL	Alabama	NULL _ NULL	7.8	7.5	25.9	58.7
1000	1980	AL	Alabama	NULL _ NULL	12.2	12.5	31.8	43.5
1000	1990	AL	Alabama	NULL _ NULL	15.7	21.7	29.4	33.1
1000	2000	AL	Alabama	NULL _ NULL	19.0	25.9	30.4	24.7
1000	2015	AL	Alabama	NULL _ NULL	23.5	29.7	31.0	15.7
1001	1970	AL	Autauga	2 _ Counties in metro areas of 250,000 to 1 million population	6.4	7.7	31.1	54.8
1001	1980	AL	Autauga	2 _ Counties in metro areas of	12.1	12.1	35.2	40.6

				250,000 to 1 million population				
1001	1990	AL	Autauga	2 _ Counties in metro areas of 250,000 to 1 million population	14.5	23.5	32.0	30.0
1001	2000	AL	Autauga	2 _ Counties in metro areas of 250,000 to 1 million population	18.0	26.9	33.8	21.3
1001	2015	AL	Autauga	2 _ Counties in metro areas of 250,000 to 1 million population	23.2	30.4	33.5	12.8

```
#Part2
# Creating three tibble
# 1) Edu1: Education (fips, Year, Education percentage)
# 2) Edu2: Fips (Fips, State, County)
# 3) Edu3: RuralCode (Fips, RuralCode and Description)


Edu1 <- as.tibble(select(Education, fips, year, percent_measure, percent ))

Edu1 <- spread(Edu1, key = percent_measure, value = percent)

head(Edu1, 10) %>% knitr::kable()
```

fips	year	percent_four_plus_years_college	percent_has_some_college	percent_hs_diploma	percent_less_than_hs_diploma
1000	1970	7.8	7.5	25.9	58.7
1000	1980	12.2	12.5	31.8	43.5
1000	1990	15.7	21.7	29.4	33.1
100	200	19.0	25.9	30.4	24.7

0	0				
1000	2015	23.5	29.7	31.0	15.7
1001	1970	6.4	7.7	31.1	54.8
1001	1980	12.1	12.1	35.2	40.6
1001	1990	14.5	23.5	32.0	30.0
1001	2000	18.0	26.9	33.8	21.3
1001	2015	23.2	30.4	33.5	12.8

```
Edu2 <- as.tibble(unique(Education[ , c("fips", "county_state"
,"rural_urban_cont_code")]))

Edu2 <- separate(Edu2, county_state, into = c("state", "county"))

## Warning: Too many values at 3153 locations: 2, 3, 4, 5, 6, 7, 8, 9, 10,
11,
## 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, ...

head(Edu2, 10) %>% knitr::kable()
```

fips	state	county	rural_urban_cont_code
1000	AL	Alabama	NULL
1001	AL	Autauga	2
1003	AL	Baldwin	3
1005	AL	Barbour	6
1007	AL	Bibb	1
1009	AL	Blount	1
1011	AL	Bullock	6
1013	AL	Butler	6
1015	AL	Calhoun	3
1017	AL	Chambers	6

```
Edu3 <- as.tibble(unique(Education[ ,c("rural_urban_cont_code",
"description")]))



head(Edu3, 10) %>% knitr::kable()
```

rural_urban_cont_code	description
NULL	NULL

2	Counties in metro areas of 250,000 to 1 million population
3	Counties in metro areas of fewer than 250,000 population
6	Urban population of 2,500 to 19,999, adjacent to a metro area
1	Counties in metro areas of 1 million population or more
9	Completely rural or less than 2,500 urban population, not adjacent to a metro area
7	Urban population of 2,500 to 19,999, not adjacent to a metro area
8	Completely rural or less than 2,500 urban population, adjacent to a metro area
4	Urban population of 20,000 or more, adjacent to a metro area
5	Urban population of 20,000 or more, not adjacent to a metro area

```
#Part3

# Question1 )Fips column in the Edu1 tibble is one of the primary key in the
Education tibble as Fips and year together form the composite key and become
unique for the education table

# Question 2) The primary key for the education tibble is Composite key
formed by FIPS and YEAR column making a combination that represents each row
uniquely

# Question 3) The rural_urban code tibble contains 9 rows with the
rural_urban_cont_code as its primary key

#Part4

# 4.0

part4.0 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))

#getting the percent of people not attaning a high school diploma for state
Massachussetts, county Mantucket and for year 1970 and 2015

part4.0 <- part4.0 %>% select(fips, state, year, county, `percent_less
than_hs_diploma`)%>% filter(fips, year, state == "MA", county == "Nantucket",
year %in% c("1970", "2015"))

head(part4.0)

## # A tibble: 2 x 5
##    fips state  year    county `percent_less than_hs_diploma`
##   <int> <chr> <int>     <chr>                          <dbl>
## 1 25019    MA  1970 Nantucket                           33.7
## 2 25019    MA  2015 Nantucket                            5.2
```

```r
#4.1

#joining education tibbe and fips tibble
part4.1 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))

# getting averagr data for percent less than high school diploma for year
2015 and state Albama
part4.1 <- (part4.1 %>% select(year,state, `percent_less than_hs_diploma`)%>%
filter(year == "2015", state == "AL"))

part4.1 <- aggregate(x=part4.1$`percent_less than_hs_diploma`,
          by=list(part4.1$year, part4.1$state),
          FUN=mean)

head(part4.1)
```

```
##   Group.1 Group.2        x
## 1    2015      AL 19.75882
```

```r
#4.2

#joining education tibbe and fips tibble
part4.2 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))

# getting averagr data for percent of college graduates for year 2015 and
state Massachussetts
part4.2 <- (part4.2 %>% select(year,state, `percent_has_some_college`)%>%
filter(year == "2015", state == "MA"))

part4.2 <- aggregate(x=part4.2$percent_has_some_college,
          by=list(part4.2$year, part4.2$state),FUN=mean)

head(part4.2)
```

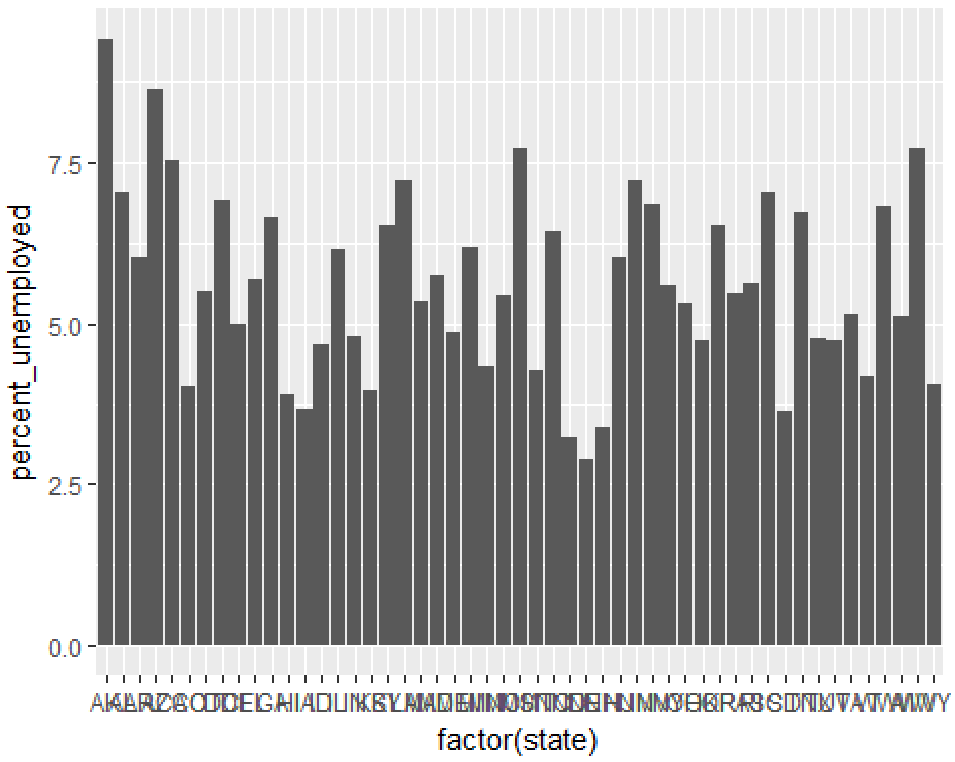
```
##   Group.1 Group.2        x
## 1    2015      MA 25.91333
```

```r
#4.3

#joining education tibbe and fips tibble
part4.3 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))


# getting averagr data for percent less than high school diploma for and
state Alabama
part4.3 <- (part4.3 %>% select(year,state, `percent_less than_hs_diploma`)%>%
filter(state == "AL"))

part4.3 <- aggregate(x=part4.3$`percent_less than_hs_diploma`,
          by=list(part4.3$year),
          FUN=mean)

head(part4.3)

##   Group.1        x
## 1    1970 65.15882
## 2    1980 50.62059
## 3    1990 40.10000
## 4    2000 30.26471
## 5    2015 19.75882

#4.4

# getting the most frequesntly occuring ruralurban code (maximum count)
Temp <- table(Edu2$rural_urban_cont_code)
freqname <- names(Temp)[Temp == max(Temp)]

head(freqname)

## [1] "6"

#4.5

#counties that have not been coded with a rural urban code
Part4.5 <- select(Edu2, state, county, rural_urban_cont_code)%>%
filter(rural_urban_cont_code == "NULL")

head(Part4.5, 10)

## # A tibble: 10 x 3
##    state      county rural_urban_cont_code
##    <chr>       <chr>                <fctr>
##   1    AL     Alabama                  NULL
##   2    AK      Alaska                  NULL
##   3    AZ     Arizona                  NULL
##   4    AR    Arkansas                  NULL
##   5    CA  California                  NULL
##   6    CO    Colorado                  NULL
##   7    CT Connecticut                  NULL
##   8    DE    Delaware                  NULL
##   9    DC    District                  NULL
## 10    FL     Florida                  NULL

#4.6

#joining education tibbe and fips tibble
Part4.6 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))

# getting averagr data for percent of college graduates for year 2010 and
state Mississippi
Part4.6 <- (Part4.6%>% select(year, state, percent_has_some_college) %>%
filter(state == "MS", year == "2010"))

head(Part4.6)

## # A tibble: 0 x 3
## # ... with 3 variables: year <int>, state <chr>,
## #   percent_has_some_college <dbl>

# returns zero rows as there is no record maintained for year 2010



#4.7

#getting state taht has maximum number of counties
Part4.7 <- (unique(select(Edu2, state, county)))

Temp <- table(Part4.6$state)
name <- as.tibble(names(Temp)[Temp == max(Temp)])

## Warning in max(Temp): no non-missing arguments to max; returning -Inf

# getting state and counties that do not have an urban code assigned to them
Part4.7a <- (select(Edu2, state, county, rural_urban_cont_code)) %>%
filter(rural_urban_cont_code == "NULL")

head(name)

## # A tibble: 0 x 0

head(Part4.7a)

## # A tibble: 6 x 3
##   state     county rural_urban_cont_code
##   <chr>      <chr>                <fctr>
## 1    AL    Alabama                  NULL
## 2    AK     Alaska                  NULL
## 3    AZ    Arizona                  NULL
## 4    AR   Arkansas                  NULL
## 5    CA California                  NULL
## 6    CO   Colorado                  NULL

#4.8
#joining education tibble and fips tibble
#joining unemployment table with previous tibble
Part4.8 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))
Part4.8 <- inner_join(Part4.8, Unemp, by = c("fips" = "fips", "year" =
"year"))


#which fip counties, U.S. states contain a higher percentage of unemployed
citizens than the percentage of college graduates, in year 2015
```

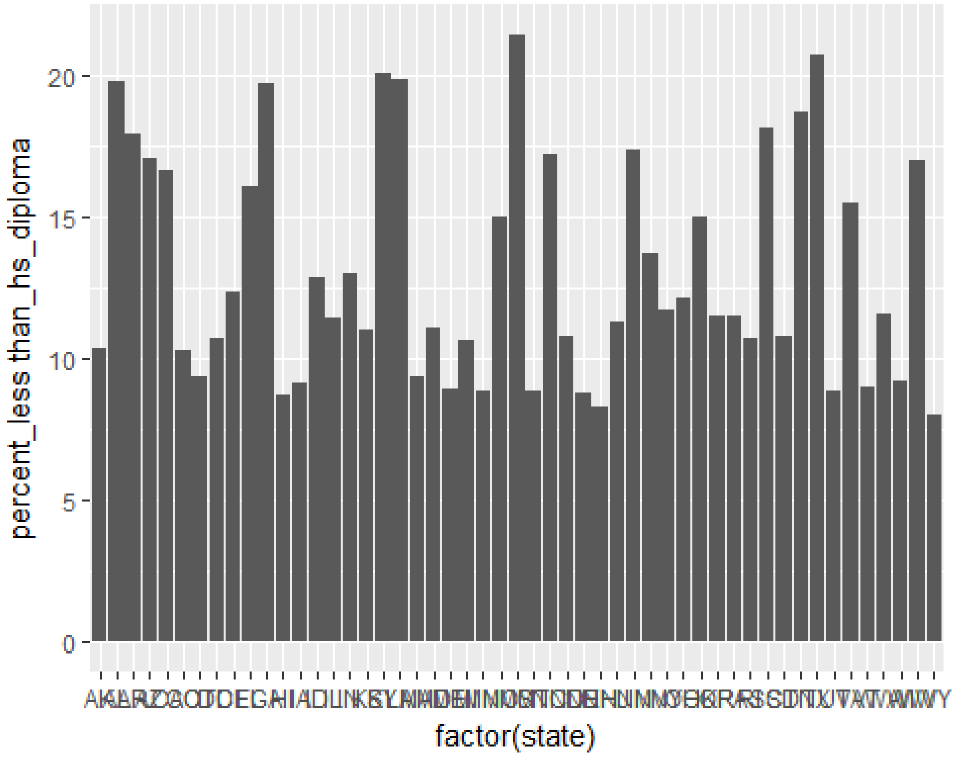
```r
Part4.8 <- Part4.8 %>% select(state, county, percent_has_some_college,
percent_unemployed ) %>% filter(percent_unemployed >
percent_has_some_college)

# none of the counties have percent_has_some_college < unemployed citizes



#4.9
#joining education tibbe and fips tibble
Part4.9 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))

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

Part4.9 <- Part4.9 %>% select(state, county, year, percent_has_some_college)
%>% filter(state == state, county == county, year == year,
percent_has_some_college == max(percent_has_some_college))

head(Part4.9)
```

```
## # A tibble: 1 x 4
##   state county  year percent_has_some_college
##   <chr>  <chr> <int>                    <dbl>
## 1    NE Banner  2015                     47.8
```

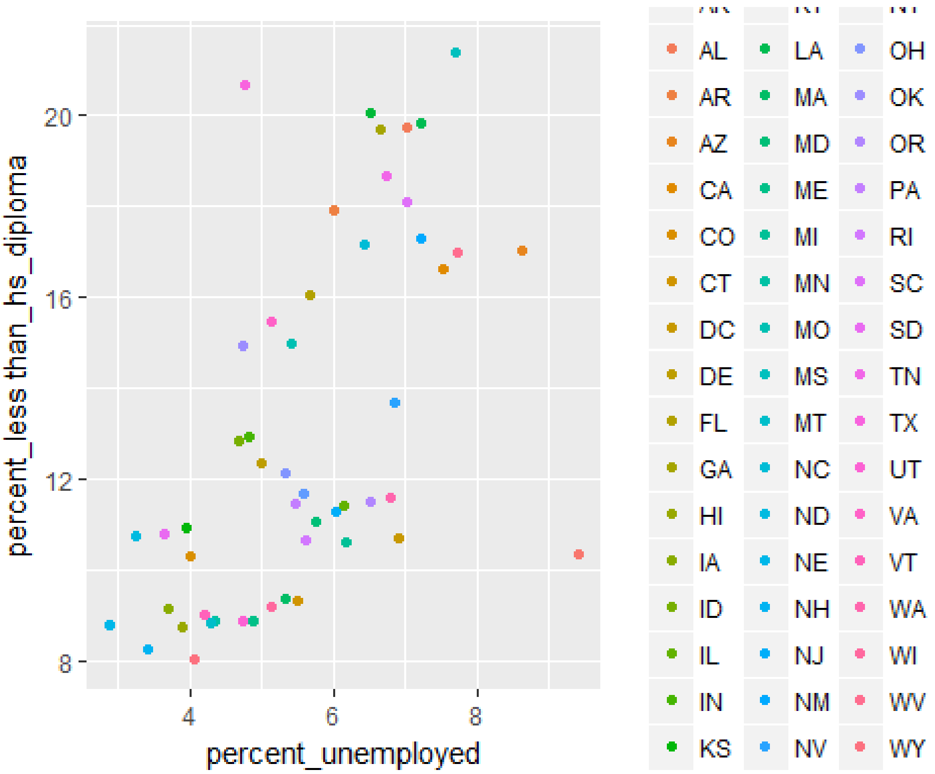
```r
#Part5

#joining education tibbe and fips tibble'
#joining unemployment table with previous tibble
Part5 <- inner_join(Edu1, Edu2, by = c("fips" = "fips"))
Part5 <- inner_join(Part5, Unemp, by = c("fips" = "fips", "year" = "year"))


Part5 <- Part5 %>% select(state, percent_unemployed, `percent_less
than_hs_diploma` )

# plotting mean percent_unemployed with respevt to each state
ggplot(Part5, aes(x=factor(state), y=percent_unemployed)) +
stat_summary(fun.y="mean", geom="bar")
```



```
# plotting mean percent_less_than _hs_diploma with respevt to each state
ggplot(Part5, aes(x=factor(state), y=`percent_less than_hs_diploma`)) +
stat_summary(fun.y="mean", geom="bar")
```



```
group <- group_by(Part5, state)

Part5.1 <- summarise(group, percent_unemployed = mean(percent_unemployed),
`percent_less than_hs_diploma` = mean(`percent_less than_hs_diploma`))

# scatter plot of mean unemployed percent v/s percent less than diploma
differentiating by state
ggplot(data = Part5.1)+ geom_point(mapping =
aes(percent_unemployed,`percent_less than_hs_diploma`, color = state ))
```


percent_less than_hs_diploma
percent_unemployed
AK
AL
AR
AZ
CA
CO
CT
DC
DE
FL
GA
HI
IA
ID
IL
IN
KS
KY
LA
MA
MD
ME
MI
MN
MO
MS
MT
NC
ND
NE
NH
NJ
NM
NV
NY
OH
OK
OR
PA
RI
SC
SD
TN
TX
UT
VA
VT
WA
WI
WV
WY