midterm project

Lu Lin

October 25, 2016

data introduction

I get the data from United States Department of Agriculture Economic Rearch Service(USDA). This dataset is aboout the County-level oil and gas production in the US.In the dataset, it also gives us the demograohic information about the coounty.

variable information

-FIPS

-geoid

-Stabr

-County_Name

-Rural_Urban_Continuum_Code_2013

-Urban_Influence_2013

-Metro_Nonmetro_2013 Metro-nonmetro 2013

-Metro_Micro_Noncore_2013

-oil2000, oil2001, . . . , oil2011

-gas2000, gas2001, . . . , gas2011

-oil_change_group

-gas_change_group

-oil_gas_change_group

Note: for more detailed information about the variables, please see the appendix.

data cleanning

The purpose of data cleaning is to make the data tidy, which is convenient for the future analysis. According to the pricinples for the tidy data, which is defined by Hadley Wickham. I make some modification to the original dataset.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape2)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:reshape2':
##
##     smiths
```

```
library(ggplot2)
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
```

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

```
#import the data from a csv file
og <- read.csv("oil.csv",header=TRUE,stringsAsFactors = FALSE)
colnames(og)
```

```
##  [1] "FIPS"                           "geoid"
##  [3] "Stabr"                          "County_Name"
##  [5] "Rural_Urban_Continuum_Code_2013" "Urban_Influence_2013"
##  [7] "Metro_Nonmetro_2013"            "Metro_Micro_Noncore_2013"
##  [9] "oil2000"                        "oil2001"
## [11] "oil2002"                        "oil2003"
## [13] "oil2004"                        "oil2005"
## [15] "oil2006"                        "oil2007"
## [17] "oil2008"                        "oil2009"
## [19] "oil2010"                        "oil2011"
## [21] "gas2000"                        "gas2001"
## [23] "gas2002"                        "gas2003"
## [25] "gas2004"                        "gas2005"
## [27] "gas2006"                        "gas2007"
## [29] "gas2008"                        "gas2009"
## [31] "gas2010"                        "gas2011"
## [33] "oil_change_group"               "gas_change_group"
## [35] "oil_gas_change_group"
```

This code is for import the dataset which is contained in a csv file.

```r
#make the demographic information into one table
demo <- og[c(1,5,6,7,8)]
colnames(demo)
```

```
## [1] "FIPS"                          "Rural_Urban_Continuum_Code_2013"
## [3] "Urban_Influence_2013"          "Metro_Nonmetro_2013"
## [5] "Metro_Micro_Noncore_2013"
```

```r
#make the oil and gas production information into one table
ogproduction <-select(og, -(Rural_Urban_Continuum_Code_2013:Metro_Micro_Noncore_2013))
colnames(ogproduction)
```

```
##  [1] "FIPS"                "geoid"               "Stabr"
##  [4] "County_Name"         "oil2000"             "oil2001"
##  [7] "oil2002"             "oil2003"             "oil2004"
## [10] "oil2005"             "oil2006"             "oil2007"
## [13] "oil2008"             "oil2009"             "oil2010"
## [16] "oil2011"             "gas2000"             "gas2001"
## [19] "gas2002"             "gas2003"             "gas2004"
## [22] "gas2005"             "gas2006"             "gas2007"
## [25] "gas2008"             "gas2009"             "gas2010"
## [28] "gas2011"             "oil_change_group"    "gas_change_group"
## [31] "oil_gas_change_group"
```

The data contains two parts of the county information. One is demographic information in 2013 and another one is the information for oil and gas amount from 2000 to 2011.Therefore, it is better to seperate them into two tables.

```r
#melt the oil and gas columns.
#one column for year, one column for oil amount, one column for gas amount
oil1<- gather(ogproduction,"year","oil", oil2000:oil2011)
og1<- gather(oil1,"year","gas",gas2000:gas2011)
colnames(og1)
```

```
##  [1] "FIPS"                "geoid"               "Stabr"
##  [4] "County_Name"         "oil_change_group"    "gas_change_group"
##  [7] "oil_gas_change_group" "year"               "oil"
## [10] "year"                "gas"
```

In the column variables, the name of the variables are oil2000,oil2001,. . . ,oil2011. These are not appropriate column names, Because they contain two informations. One is the year information and another one is the oil amount information.Thus, we should have two seperated columns. One column is for the yearand one column for the oil production information. This is also applied to the columns which contain the information for gas.

```r
#delete a repeated year column
og1[10] <- NULL
colnames(og1)
```

```
##  [1] "FIPS"                "geoid"               "Stabr"
##  [4] "County_Name"         "oil_change_group"    "gas_change_group"
##  [7] "oil_gas_change_group" "year"               "oil"
## [10] "gas"
```

After the modification, there are two same year column, so we delete one of them.

```
head(og1$year)
```

```
## [1] "oil2000" "oil2000" "oil2000" "oil2000" "oil2000" "oil2000"
```

```
#delete the oil words from the values
og1$year[og1$year == "oil2000"] <- "2000"
og1$year[og1$year == "oil2001"] <- "2001"
og1$year[og1$year == "oil2002"] <- "2002"
og1$year[og1$year == "oil2003"] <- "2003"
og1$year[og1$year == "oil2004"] <- "2004"
og1$year[og1$year == "oil2005"] <- "2005"
og1$year[og1$year == "oil2006"] <- "2006"
og1$year[og1$year == "oil2007"] <- "2007"
og1$year[og1$year == "oil2008"] <- "2008"
og1$year[og1$year == "oil2009"] <- "2009"
og1$year[og1$year == "oil2010"] <- "2010"
og1$year[og1$year == "oil2011"] <- "2011"
head(og1$year)
```

```
## [1] "2000" "2000" "2000" "2000" "2000" "2000"
```

For the year column, we only want the number instead of the oil and number, so we use the number to subsititute the original values.

```
#delete one column in the demo table, which contains the repeated informatino from another table
demo$Metro_Nonmetro_2013 <- NULL
```

For the demographic data table, the Metro_Nonmetro_2013 column contains the same information as the Metro_Micro_Noncore_2013 column,so we delete the former one.

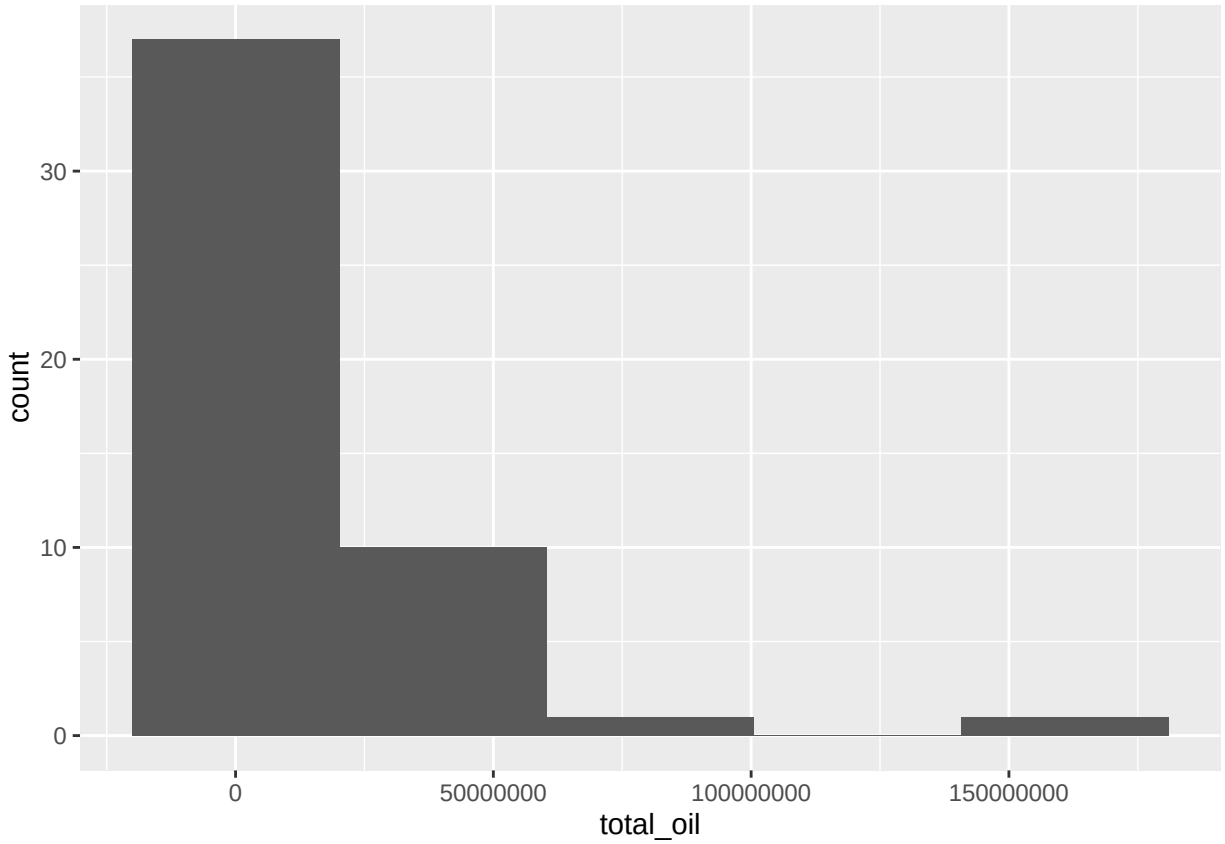
data summerizing

After data cleaning, we can do data summerizing to dig out more information.

```
#make a table of the total oil amount per state
tempog <- data.frame( og1$FIPS, og1$Stabr, og1$oil)
colnames(tempog) <- c("FIPS", "State", "Oil")
oil_by_state <- summarise(group_by(tempog, State), sum(as.numeric(Oil)))
colnames(oil_by_state)[2] <- "total_oil"
```

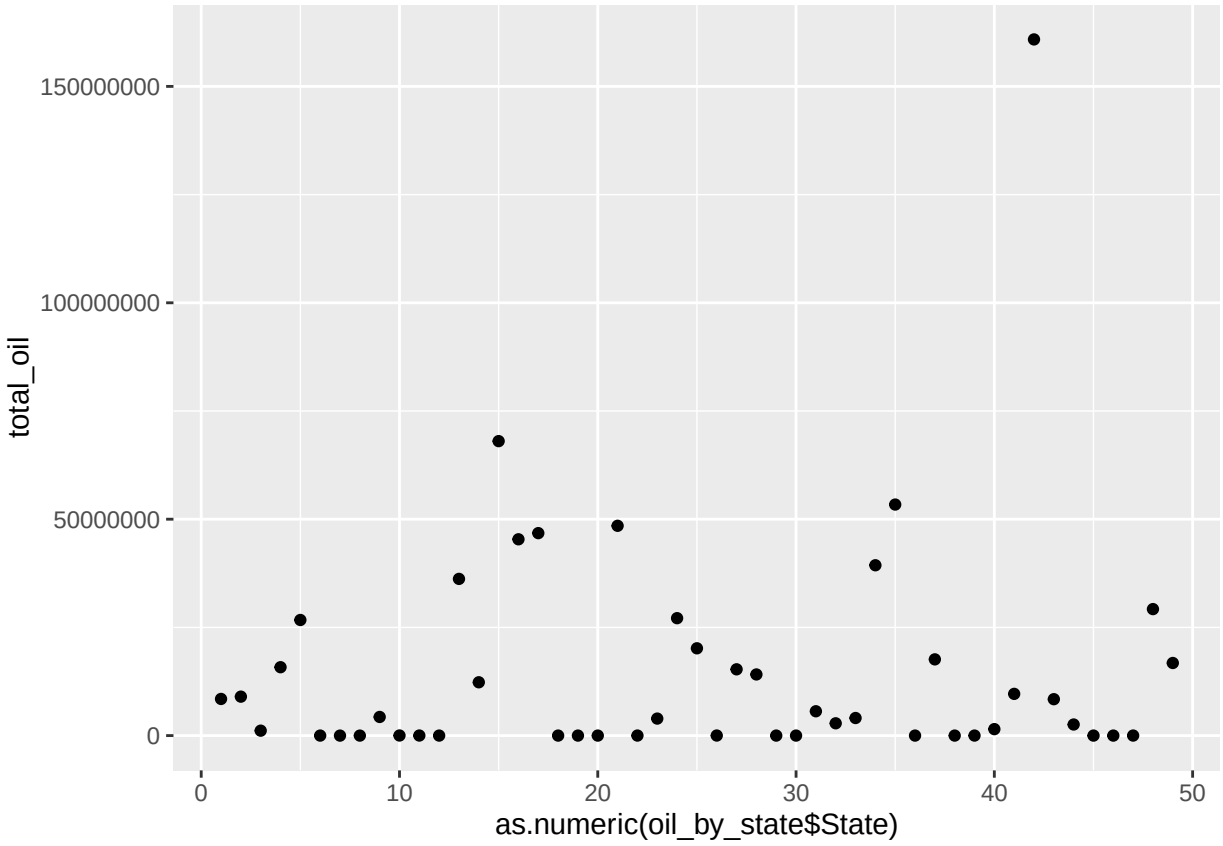
I want to know the total amount of oil from 2000 to 2011 for different state. Therefore, we make a table which contains the state names and the total oil amount.

```
#make a scatterplot for the total oil amount of different states
options(scipen=5)
ggplot(oil_by_state,aes(total_oil))+geom_histogram(bins = 5)
```



Based on the oil_by_state table, we build a scatterplot, which we can visually see the total amount of oil on the graph.

```
#count the number of state for different oil amount
ggplot(oil_by_state,aes(x=as.numeric(oil_by_state$State),y=total_oil))+geom_point()
```



Based on the table, I want to know the number of states for different oil amount.

Appendix

-FIPS
Five digit Federal Information Processing Standard (FIPS) code (numeric)

-geoid
FIPS code with leading zero (string)

-Stabr
State abbreviation (string)

-County_Name
County name (string)

-Rural_Urban_Continuum_Code_2013

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

-Urban_Influence_2013

descriptions: 1 In large metro area of 1+ million residents 2 In small metro area of less than 1 million residents Nonmetropolitan counties 3 Micropolitan area adjacent to large metro area 4 Noncore adjacent to

large metro area 5 Micropolitan area adjacent to small metro area 6 Noncore adjacent to small metro area and contains a town of at least 2,500 residents 7 Noncore adjacent to small metro area and does not contain a town of at least 2,500 residents 8 Micropolitan area not adjacent to a metro area 9 Noncore adjacent to micro area and contains a town of at least 2,500 residents 10 Noncore adjacent to micro area and does not contain a town of at least 2,500 residents 11 Noncore not adjacent to metro or micro area and contains a town of at least 2,500 residents 12 Noncore not adjacent to metro or micro area and does not contain a town of at least 2,500 residents

-Metro_Nonmetro_2013 Metro-nonmetro 2013 0 nonmetro 1 metro

-Metro_Micro_Noncore_2013
0 nonmetro noncore 1 nonmetro micropolitan 2 metropolitan

-oil2000, oil2001, . . . , oil2011 Annual gross withdrawals (barrels) of crude oil, for the year specified in the variable name

-gas2000, gas2001, . . . , gas2011 Annual gross withdrawals (thousand cubic feet) of natural gas, for the year specified in the variable name

-oil_change_group
Categorical variable based upon change in the dollar value of oil production, 2000-11. Values are H_Growth (>=$20 million),H Decline (<=-$20 million), Status Quo (change between +/- $20 million)

-gas_change_group
Categorical variable based upon change in the dollar value of natural gas production, 2000-11. Values are H_Growth (>=$20 million), H_Decline (<=-$20 million), Status Quo (change between +/- $20 million)

-oil_gas_change_group
Categorical variable based on the change in the dollar value of the sum of oil and natural gas production, 2000-11. Values are H_Growth (>=$20 million),H_Decline (<=-$20 million), Status Quo (change between +/- $20 million)