Miterm Project Presentation

Zihuan Qiao Team Member: CJ

2016/10/24

1. Introdunction to Dataset

a. Data Background

Gas and oil industries play significant roles in a nation's economy. And development of gas and oil industries depend a lot on locations. In this project, by looking at the County-level annual gross withdrawals of oil and gas in US, we try to explore some useful information about the distribution of oil and gas withdrawals in different counties and states through years from 2000 to 2011. County-level data from oil and/or natural gas producing States—for onshore production in the lower 48 States only—are compiled on a State-by-State basis.

b. Data Source

Data used in this project is aquired from ERS which stands for Economic Research Service. Data used in this project can be downloaded from website:http://www.ers.usda.gov/data-products/county-level-oil-and-gas-production-in-the-us.aspx.

Most States have production statistics available by county, field, or well, and these data were compiled by ERS at the county level to create a database of county-level production, annually for 2000 through 2011. The dataset is also maintained by ERS. Up till now, the County-level data has been updated to year 2011. Currently, an ERS update to this data product is not planned.

2. Data Wrangling

Noticing raw data is not clean enough for further exploration because several column headers are values, not variable names. So we use commands gather, mutate, filter and select from tidyr and dplyr packages to do the data wrangling. Finaly, tidy data is saved as oilTidyData.txt for later use.

Related code and result are as follows:

```r
library(tidyr)
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
```

```r
#load data
data.raw <- read.csv("oilgascounty.csv")

#convert data type to data frame
data0 <- data.frame(data.raw)
head(data0)
```

```
##   FIPS geoid Stabr    County_Name Rural_Urban_Continuum_Code_2013
## 1 1001  1001    AL Autauga County                               2
## 2 1003  1003    AL Baldwin County                               3
## 3 1005  1005    AL Barbour County                               6
## 4 1007  1007    AL    Bibb County                               1
## 5 1009  1009    AL  Blount County                               1
## 6 1011  1011    AL Bullock County                               6
##   Urban_Influence_2013 Metro_Nonmetro_2013 Metro_Micro_Noncore_2013
## 1                    2                   1                        2
## 2                    2                   1                        2
## 3                    6                   0                        0
## 4                    1                   1                        2
## 5                    1                   1                        2
## 6                    6                   0                        0
##   oil2000 oil2001 oil2002 oil2003 oil2004 oil2005 oil2006 oil2007 oil2008
## 1       0       0       0       0       0       0       0       0       0
## 2  138072  134666  138011  127985  130763  118043  103992  112303   97623
## 3       0       0       0       0       0       0       0       0       0
## 4       0       0       0       0       0       0       0       0       0
## 5       0       0       0       0       0       0       0       0       0
## 6       0       0       0       0       0       0       0       0       0
##   oil2009 oil2010 oil2011  gas2000  gas2001   gas2002   gas2003  gas2004
## 1       0       0       0        0        0         0         0        0
## 2   84982  101955   94638 72543902 98699994 107142655 101510068 90146850
## 3       0       0       0        0        0         0         0        0
## 4       0       0       0        0        0         0         0        0
## 5       0       0       0        0        0         0         0        0
## 6       0       0       0        0        0         0         0        0
##    gas2005  gas2006  gas2007  gas2008  gas2009  gas2010  gas2011
## 1        0        0        0        0        0        0        0
## 2 84536875 83951640 82876786 78547145 68525628 63069025 51041072
## 3        0        0        0        0        0        0        0
## 4     8301    98853   480015   684143   551719   453132   400504
## 5        0        0        0    20516    61054     3594    21496
## 6        0        0        0        0        0        0        0
##   oil_change_group gas_change_group oil_gas_change_group
## 1       Status Quo       Status Quo           Status Quo
## 2       Status Quo        H_Decline            H_Decline
## 3       Status Quo       Status Quo           Status Quo
## 4       Status Quo       Status Quo           Status Quo
## 5       Status Quo       Status Quo           Status Quo
## 6       Status Quo       Status Quo           Status Quo
```

```r
#gather oil data from 2000 to 2011
data.oil0 <- gather(data0, year, oilwithdraw, oil2000:oil2011)
data.oil <- mutate(data.oil0, year=(gsub("oil","",year)))
```

```r
#gather gas data from 2000 to 2011 on the biasis of data.oil
data.gas0 <- gather(data.oil, year2, gaswithdraw, gas2000:gas2011)
data.gas <- mutate(data.gas0, year2=(gsub("gas","",year2)))

#delete verbose rows where year != year2
data1 <- filter(data.gas, year == year2)

#delete verbose column year2
data <- select(data1, -year2)

#adjust columns order
Tidydata <- data[, c(1:8, 12:14, 9:11)]
head(Tidydata)
```

```
##   FIPS geoid Stabr    County_Name Rural_Urban_Continuum_Code_2013
## 1 1001  1001    AL Autauga County                               2
## 2 1003  1003    AL Baldwin County                               3
## 3 1005  1005    AL Barbour County                               6
## 4 1007  1007    AL    Bibb County                               1
## 5 1009  1009    AL  Blount County                               1
## 6 1011  1011    AL Bullock County                               6
##   Urban_Influence_2013 Metro_Nonmetro_2013 Metro_Micro_Noncore_2013 year
## 1                    2                   1                        2 2000
## 2                    2                   1                        2 2000
## 3                    6                   0                        0 2000
## 4                    1                   1                        2 2000
## 5                    1                   1                        2 2000
## 6                    6                   0                        0 2000
##   oilwithdraw gaswithdraw oil_change_group gas_change_group
## 1           0           0       Status Quo       Status Quo
## 2      138072    72543902       Status Quo        H_Decline
## 3           0           0       Status Quo       Status Quo
## 4           0           0       Status Quo       Status Quo
## 5           0           0       Status Quo       Status Quo
## 6           0           0       Status Quo       Status Quo
##   oil_gas_change_group
## 1           Status Quo
## 2            H_Decline
## 3           Status Quo
## 4           Status Quo
## 5           Status Quo
## 6           Status Quo
```

```r
write.table(Tidydata, "oilTidyData.txt")
```

3. Data Summarization

a. Variable Descriptions

```
dfnew <- read.table("oilTidyData.txt")
dfnew <- data.frame(dfnew)
dim(dfnew)
```

```
## [1] 37308    14
```

There are 14 variables in the tidy data. Their names and descriptions are as follows:

Variable Name	Description and Variable Labels
FIPS	Five-digit Federal Information Processing Standard (FIPS) code (num
geoid	FIPS code with leading zero (string)
Stabr	State abbreviation (string)
County Name	County name (string)
Rural Urban Continuum Code2013	Rural-urban Continuum Code, 2013 (see code descriptions)
Urban Influence 2013	Urban Influence Code, 2013 (see code descriptions)
Metro Nonmetro2013	Metro-nonmetro 2013 (0=nonmetro, 1=metro)
Metro Micro Noncore2013	Metro Micro Noncore indicator 2013 (0=nonmetro noncore, 1=nonmetro micropolita
year	year of data
oilwithdraw	Annual gross withdrawals (barrels) of crude oil, for the year specified in the v
gaswithdraw	Annual gross withdrawals (1,000 cubic feet) of natural gas, for the year specified in
oil change group	Categorical variable based upon change in the dollar value of oil produ
gas change group	Categorical variable based upon change in the dollar value of natural gas p
oil gas change group	Categorical variable based on the change in the dollar value of the sum of oil and na

b. Tidy Data Summarization

```
#basic data summary: mean, max, min, etc.
dfnew <- read.table("oilTidyData.txt")
str(dfnew)
```

```
## 'data.frame':    37308 obs. of  14 variables:
##  $ FIPS                          : int  1001 1003 1005 1007 1009 1011 1013 1015 1017 1019 ...
##  $ geoid                         : int  1001 1003 1005 1007 1009 1011 1013 1015 1017 1019 ...
##  $ Stabr                         : Factor w/ 49 levels "AL","AR","AZ",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ County_Name                   : Factor w/ 1842 levels "Abbeville County",..: 80 87 98 147 162 222
##  $ Rural_Urban_Continuum_Code_2013: int  2 3 6 1 1 6 6 3 6 6 ...
##  $ Urban_Influence_2013          : int  2 2 6 1 1 6 6 2 5 6 ...
##  $ Metro_Nonmetro_2013           : int  1 1 0 1 1 0 0 1 0 0 ...
##  $ Metro_Micro_Noncore_2013      : int  2 2 0 2 2 0 0 2 1 0 ...
##  $ year                          : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ oilwithdraw                   : int  0 138072 0 0 0 0 0 0 0 0 ...
##  $ gaswithdraw                   : int  0 72543902 0 0 0 0 0 0 0 0 ...
##  $ oil_change_group              : Factor w/ 3 levels "H_Decline","H_Growth",..: 3 3 3 3 3 3 3 3 3
##  $ gas_change_group              : Factor w/ 3 levels "H_Decline","H_Growth",..: 3 1 3 3 3 3 3 3 3
##  $ oil_gas_change_group          : Factor w/ 3 levels "H_Decline","H_Growth",..: 3 1 3 3 3 3 3 3 3
```

```
summary(dfnew)
```

```
##       FIPS           geoid            Stabr                  County_Name   
##  Min.   : 1001   Min.   : 1001   TX     : 3048   Washington County:  360  
##  1st Qu.:19045   1st Qu.:19045   GA     : 1908   Jefferson County :  300  
##  Median :29213   Median :29213   VA     : 1608   Franklin County  :  288  
##  Mean   :30679   Mean   :30679   KY     : 1440   Jackson County   :  276  
##  3rd Qu.:46009   3rd Qu.:46009   MO     : 1380   Lincoln County   :  276  
##  Max.   :56045   Max.   :56045   KS     : 1260   Madison County   :  228  
##                                  (Other):26664   (Other)          :35580  
##  Rural_Urban_Continuum_Code_2013 Urban_Influence_2013 Metro_Nonmetro_2013
##  Min.   :1.000                   Min.   : 1.000       Min.   :0.0000     
##  1st Qu.:2.000                   1st Qu.: 2.000       1st Qu.:0.0000     
##  Median :6.000                   Median : 5.000       Median :0.0000     
##  Mean   :4.986                   Mean   : 5.224       Mean   :0.3734     
##  3rd Qu.:7.000                   3rd Qu.: 8.000       3rd Qu.:1.0000     
##  Max.   :9.000                   Max.   :12.000       Max.   :1.0000     
##                                                                          
##  Metro_Micro_Noncore_2013      year       oilwithdraw       
##  Min.   :0.0000           Min.   :2000   Min.   :        0  
##  1st Qu.:0.0000           1st Qu.:2003   1st Qu.:        0  
##  Median :1.0000           Median :2006   Median :        0  
##  Mean   :0.9518           Mean   :2006   Mean   :   368432  
##  3rd Qu.:2.0000           3rd Qu.:2008   3rd Qu.:     9980  
##  Max.   :2.0000           Max.   :2011   Max.   :208781424  
##                                                             
##   gaswithdraw          oil_change_group   gas_change_group
##  Min.   :0.000e+00   H_Decline : 1464   H_Decline : 1860  
##  1st Qu.:0.000e+00   H_Growth  : 1284   H_Growth  : 2088  
##  Median :0.000e+00   Status Quo:34560   Status Quo:33360  
##  Mean   :5.809e+06                                        
##  3rd Qu.:5.102e+04                                        
##  Max.   :1.198e+09                                        
##                                                           
##  oil_gas_change_group
##  H_Decline : 2544    
##  H_Growth  : 2616    
##  Status Quo:32148    
##                      
##                      
##                      
## 
```

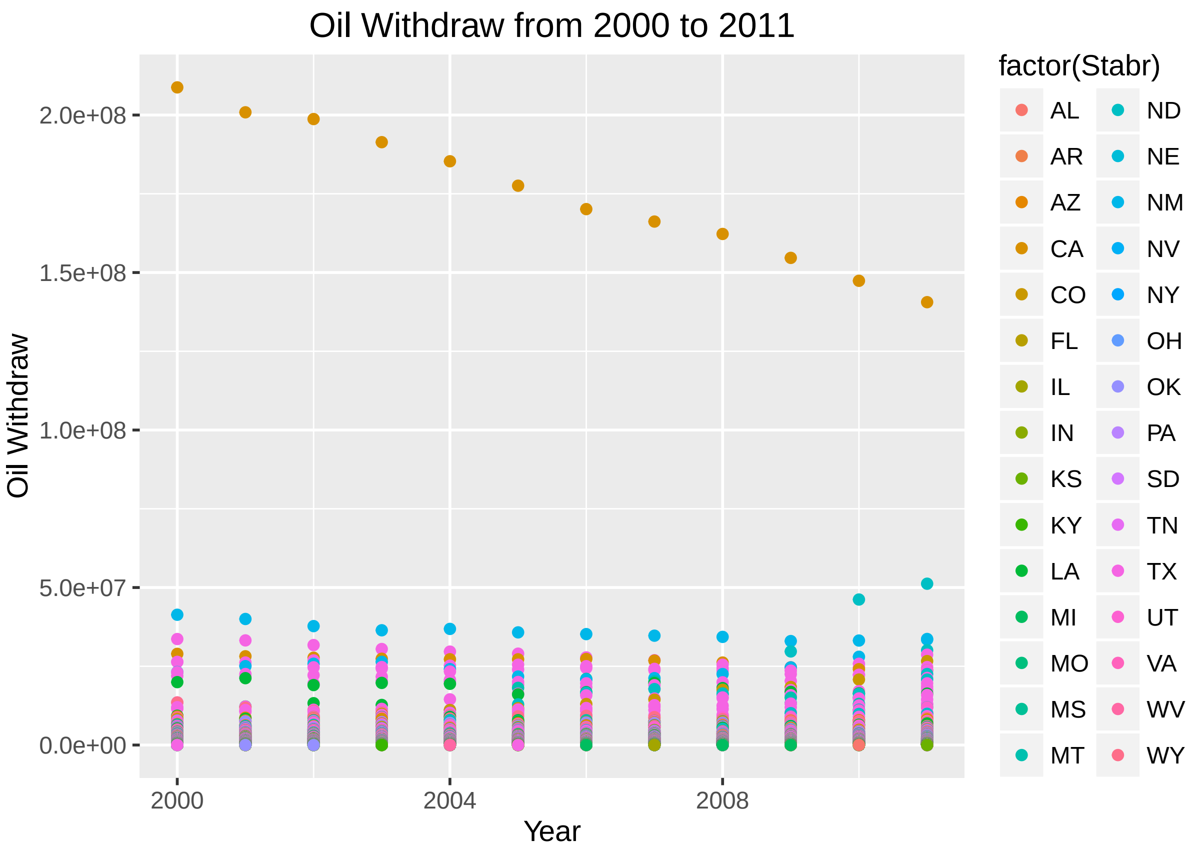
Tidy data has 14 variables with 37308 observations. Basic summarization including min, 1st Qu, median, mean, 3rd Qu, max for each variable is shown above.

4. Data Exploration

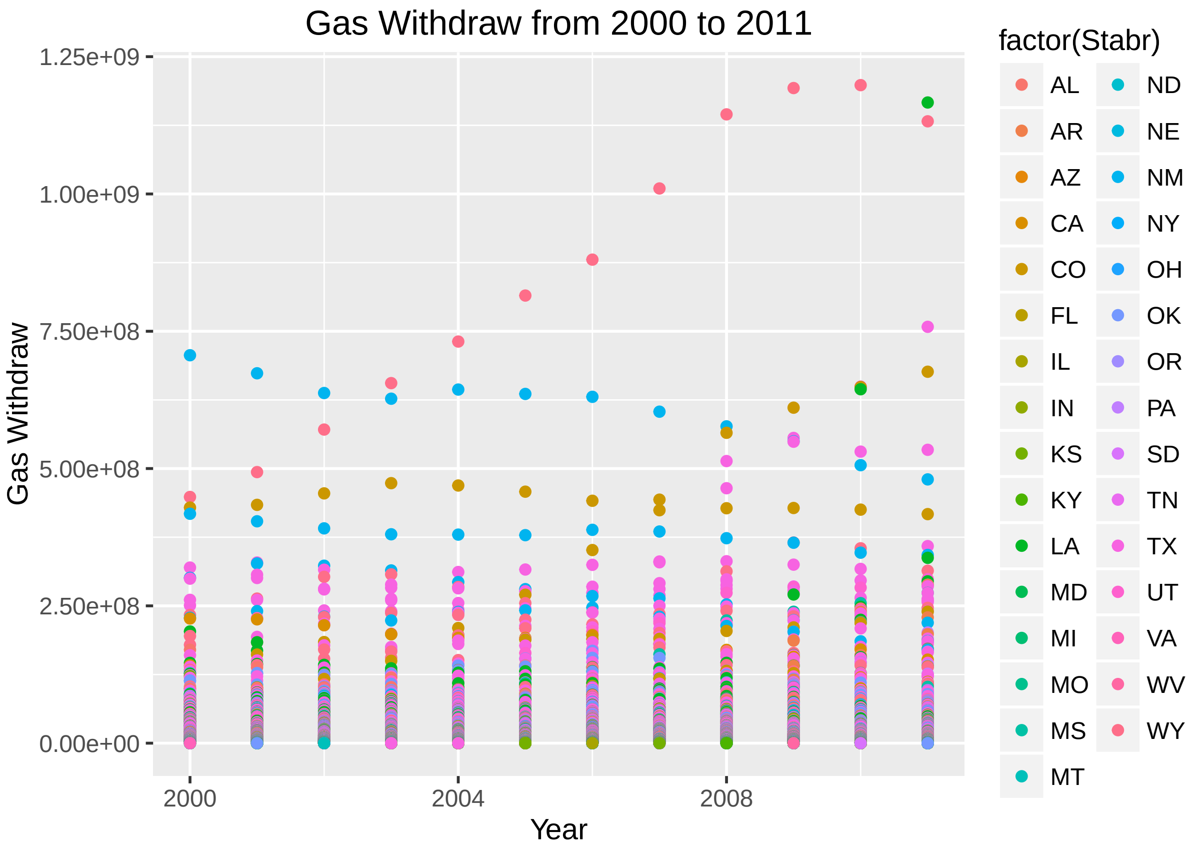
a. Year Based Exploration

(i) National Oil and Gas gross Withdrawals from 2000 to 2011

```
library(ggplot2)
library(dplyr)


#read tidy data
dfnew <- read.table("oilTidyData.txt")
dfnew <- data.frame(dfnew)

#select top 10000 oil/ gas withdraw county data
dfnew.oil <- arrange(dfnew, desc(oilwithdraw))
dfnew.oil <- dfnew.oil[1:10000,]
dfnew.gas <- arrange(dfnew, desc(gaswithdraw))
dfnew.gas <- dfnew.gas[1:10000,]

#draw point plot and line to indicate oil/ gas withdraws from year to year in terms of state
gg <- ggplot(dfnew.oil, aes(x = year, y = oilwithdraw, colour = factor(Stabr)))
gg + geom_point() + labs(title = "Oil Withdraw from 2000 to 2011", x = "Year", y = "Oil Withdraw")
```



```
qq <- ggplot(dfnew.gas, aes(x = year, y = gaswithdraw, colour = factor(Stabr)))
qq + geom_point() + labs(title = "Gas Withdraw from 2000 to 2011", x = "Year", y = "Gas Withdraw")
```



Point plot shows the oil annual gross withdraw of each state from 2000 to 2011. From the plot, we can see the level of oil withdrawals from 2000 to 2011 of each state. Also we can see the trend of each state from year to year. For example, CA has the highest level of oil withdrawals from 2000 to 2011 And its level of oil withdrawals is in a continuous decrease from 2000 to 2011.

Similarly, Point plot shows the gas annual gross withdraw of each state from 2000 to 2011. From the plot, we can see the level of oil withdrawals from 2000 to 2011 of each state. Also we can see the trend of each state from year to year. But situation with gas annual gross eithdrawals is much more complicated than the sitution with oil annual gross eithdrawals. NE has the highes level of gas annual gross eithdrawals in the first three years but it continues to drop through the years while AL keeps increasing its gas annual gross eithdrawals and surpass NE in 2003. But later in 2011, it is surpassed by LA.

b. Location Based Exploration

(i) National Oil and Gas Gross Withdrawals of Each State

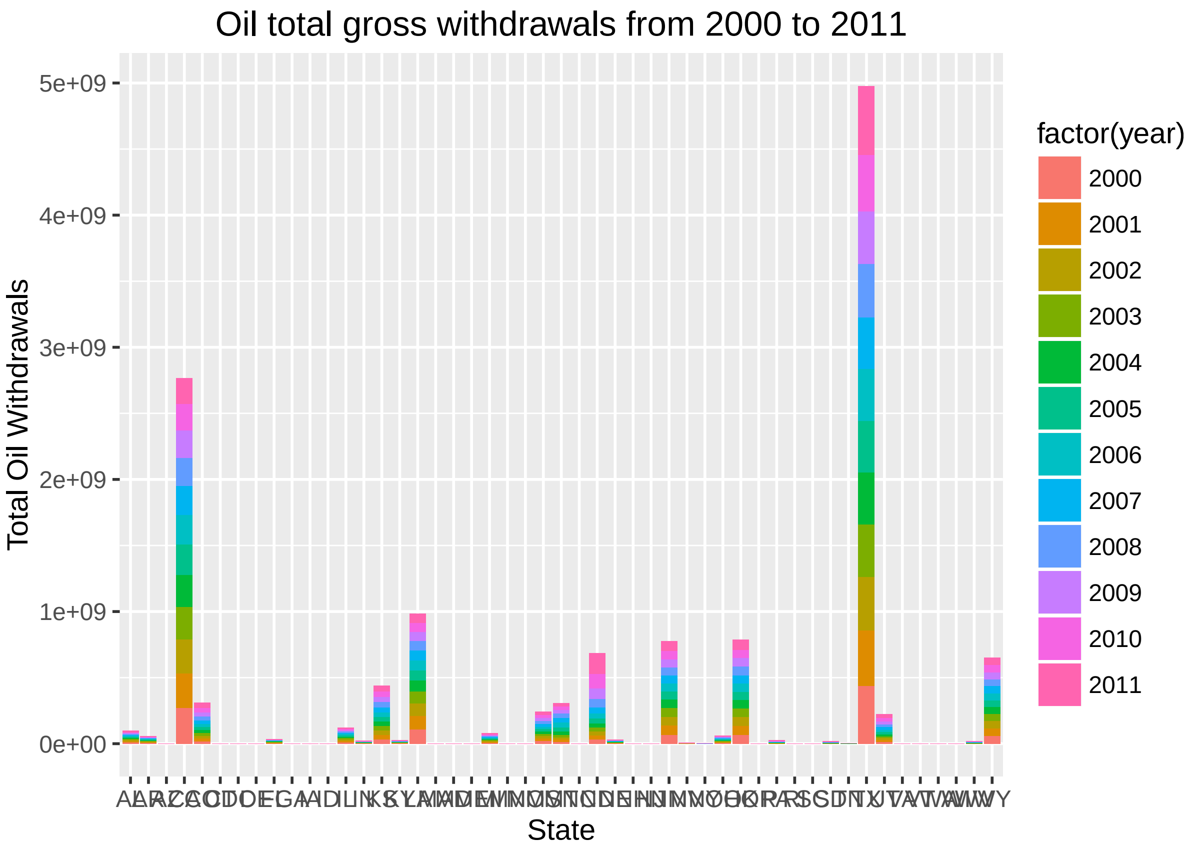
```
#Oil and Gas total gross withdrawals distribution on the state level

library(ggplot2)
library(magrittr)
```

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

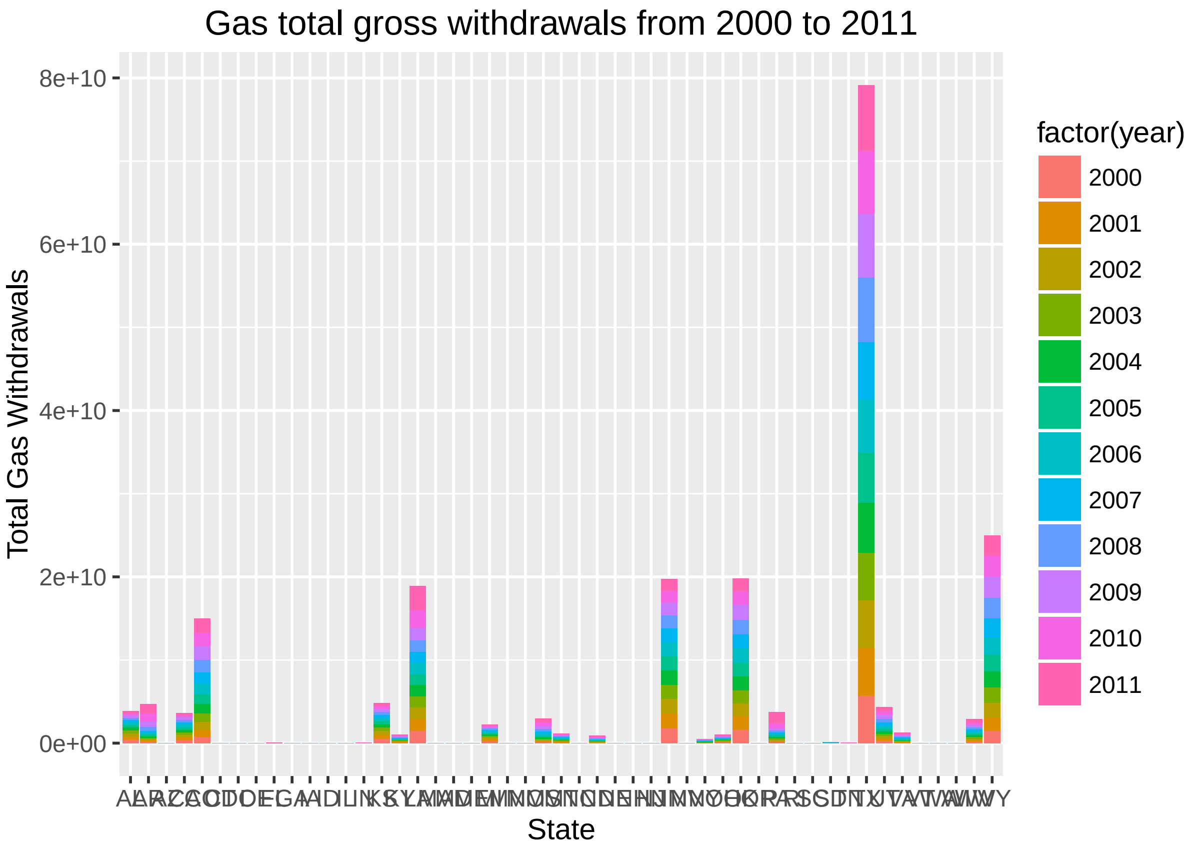
```
## The following object is masked from 'package:tidyr':
```

```
##
##     extract
```

```
#read tidy data and select subset
dfnew <- read.table("oilTidyData.txt")
attach(dfnew)
dfnew.state <- dfnew %>% select(Stabr, year : gaswithdraw)


#Draw bar graghics: Oil total gross withdrawals distribution on the state level
gg <- ggplot(data = dfnew.state, aes(x = Stabr, y = oilwithdraw, fill = factor(year)))
gg + geom_bar(stat = "identity") +
  labs(title = "Oil total gross withdrawals from 2000 to 2011", x = "State", y = "Total Oil Withdrawals
```

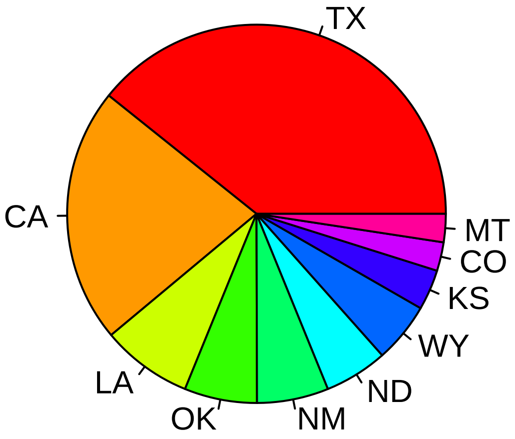


```
#Draw bar graghics: Gas total gross withdrawals distribution on the state level
gg <- ggplot(data = dfnew.state, aes(x = Stabr, y = gaswithdraw, fill = factor(year)))
gg + geom_bar(stat = "identity") +
  labs(title = "Gas total gross withdrawals from 2000 to 2011", x = "State", y = "Total Gas Withdrawals
```



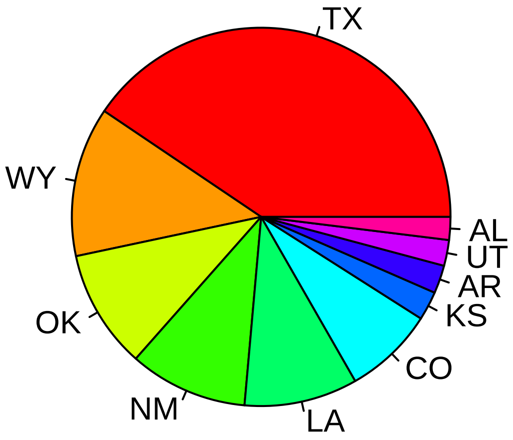
```
# Present top 10 biggest oil/ gas production states in piechart
library(ggplot2)


#read tidy data
dfnew <- read.table("oilTidyData.txt")
dfnew <- data.frame(dfnew)

# Present top 10 biggest oil production states in piechart
newstatesoil <- dfnew %>% group_by(Stabr) %>% summarize(sum_oil = sum(as.numeric(oilwithdraw)))
# We find that the pie chart is too dense. Let's list top 10 states, and its relative pie chart
newstatedata <- newstatesoil[order(-newstatesoil$sum_oil),]
# Now, here comes the top 10 states in oilwithdraw
Cleanstatedata <- newstatedata[1:10,]
pie(Cleanstatedata$sum_oil, labels = Cleanstatedata$Stabr, col = rainbow(length(Cleanstatedata$Stabr)),
```

Pie chart of oil withdraws in top 10 states



```r
# Present top 10 biggest gas production states in piechart
newstatesgas <- dfnew %>% group_by(Stabr) %>% summarize(sum_gas = sum(as.numeric(gaswithdraw)))
newstatedata1 <- newstatesgas[order(-newstatesgas$sum_gas),]
# Now, here comes the top 10 states in gaswithdraw
Cleanstatedata1 <- newstatedata1[1:10,]
pie(Cleanstatedata1$sum_gas, labels = Cleanstatedata1$Stabr, col = rainbow(length(Cleanstatedata1$Stabr)
```

Pie chart of gas withdraws in top 10 states



Oil and gas total gross withdrawals from 2000 to 2011 are shown by bar graph. Each bar consists of annual gross withdrawals from 2000 to 2011 of each state, each color represent a specific year.

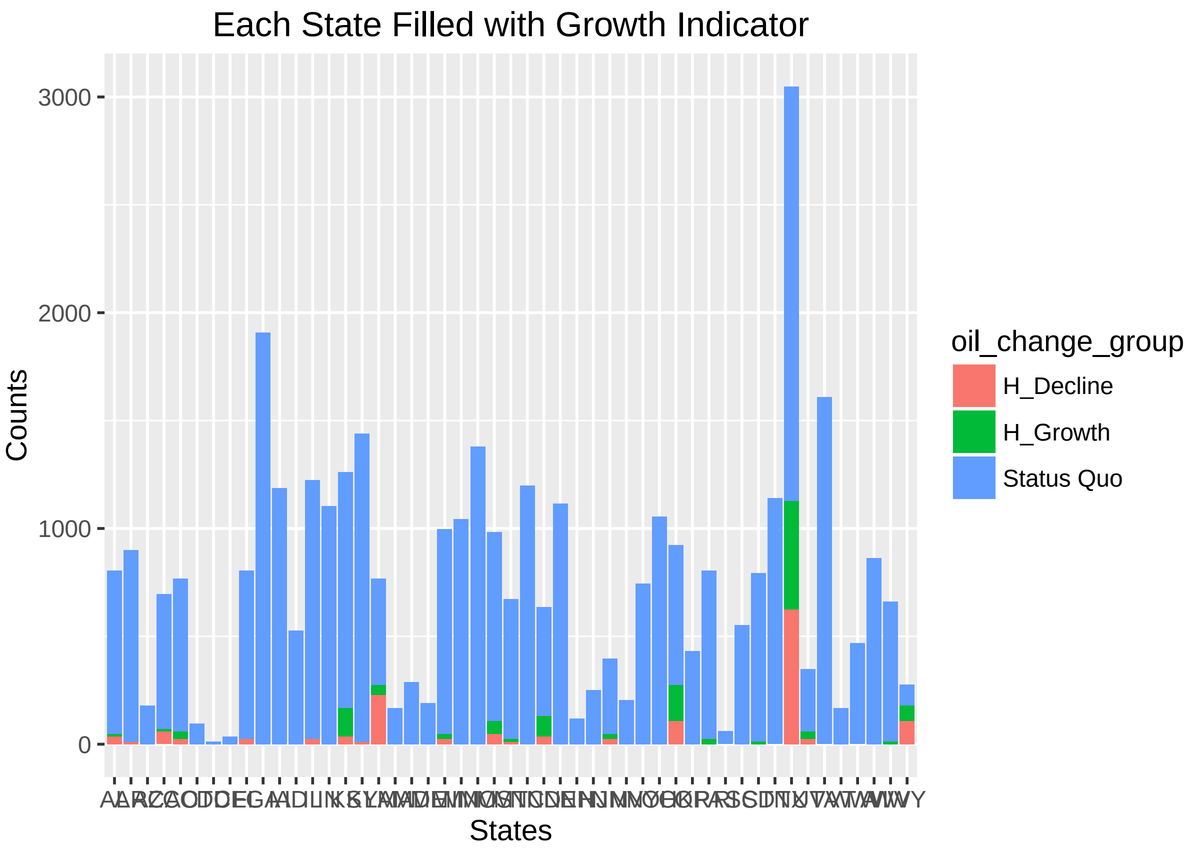
We can see from the , TX has the highest level of oil total gross withdrawals from 2000 to 2011 and CA follows. Moreover, states tend to have higher level of oil total gross withdrawals from 2000 to 2011 in 2011.

Similar expanation can be applied to gas total gross withdrawals from 2000 to 2011. TX has the highest level of oil total gross withdrawals from 2000 to 2011 and WY, LA, MN, OK follows whose gaps are not that big.

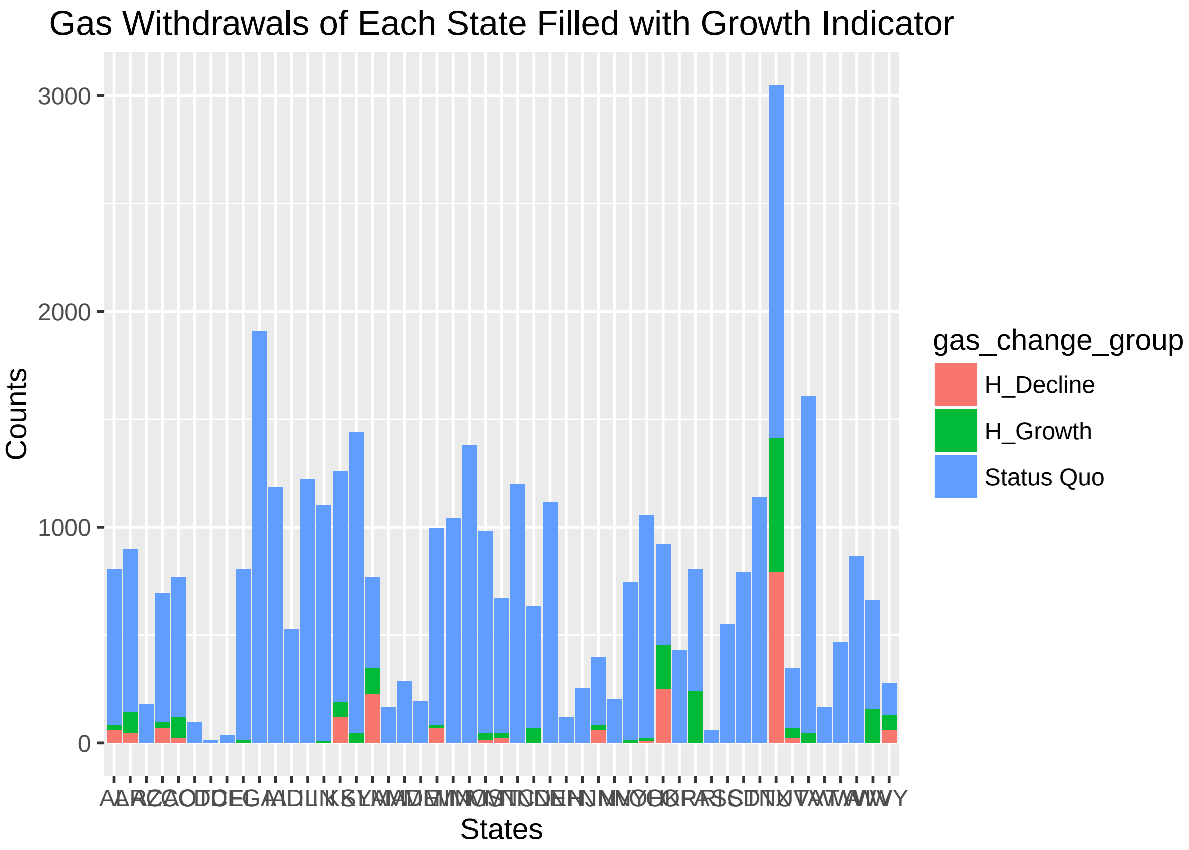
However, because there are 48 states shown in the bar gragh, the x lable is little bit too close which add difficulty in identificaiton, two corresponding pie charts are drawn choosing data from oil/ gas gross withdrawals top 10 states. The same conclusion as above can be drawn easier by looking at the pie chart which is clearer.

(ii) Oil and gas withdrawals change from 2000 to 2011 on the county level

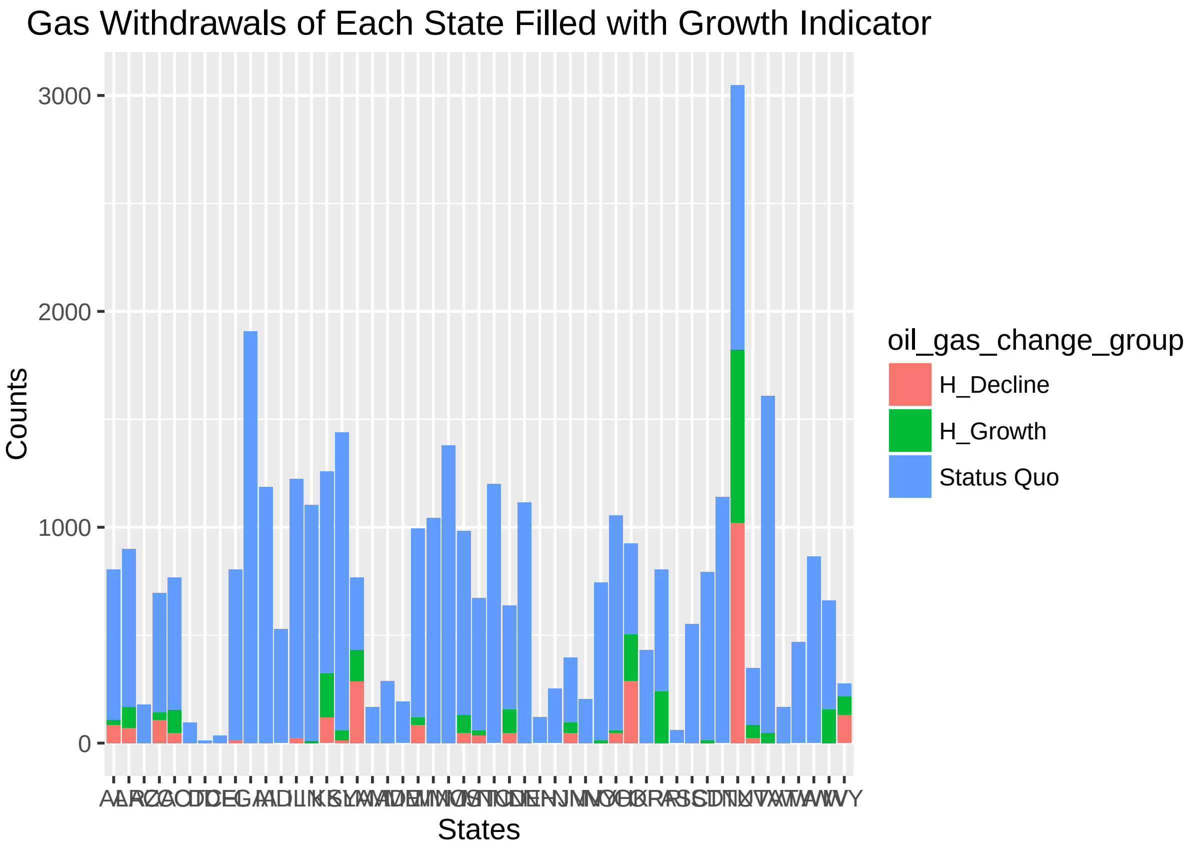
```r
# Each State filled with growth indicator
library(ggplot2)


#read tidy data
dfnew <- read.table("oilTidyData.txt")
dfnew <- data.frame(dfnew)

#draw  bar graphics
state_oil_growth<- ggplot(dfnew, aes(x=Stabr, fill=oil_change_group))+
  geom_bar() + labs(title = "Each State Filled with Growth Indicator", x = "States", y = "Counts")
state_oil_growth
```



```r
state_gas_growth<- ggplot(dfnew, aes(x=Stabr, fill=gas_change_group))+
  geom_bar() + labs(title = "Gas Withdrawals of Each State Filled with Growth Indicator", x = "States",
state_gas_growth
```



```
state_oil_gas_growth<- ggplot(dfnew, aes(x=Stabr, fill=oil_gas_change_group))+
  geom_bar() + labs(title = "Gas Withdrawals of Each State Filled with Growth Indicator", x = "States",
state_oil_gas_growth
```

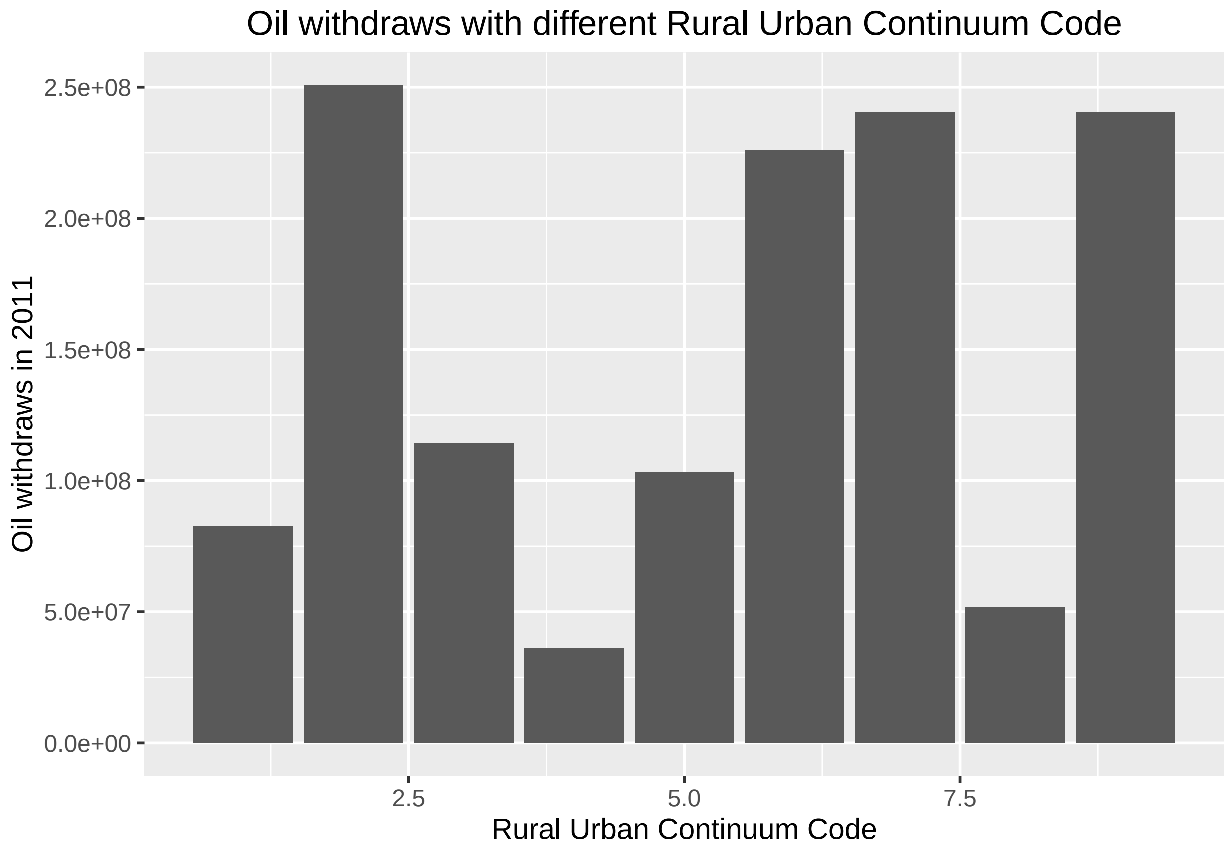


Oil gas change group is a categorical variable based upon change in the dollar value of oil production. There are three level indicating different change range where H Growth indicates grows more than 20 million, H Decline indicates grows less than 20 million, and Status Quo indicates the between situation.

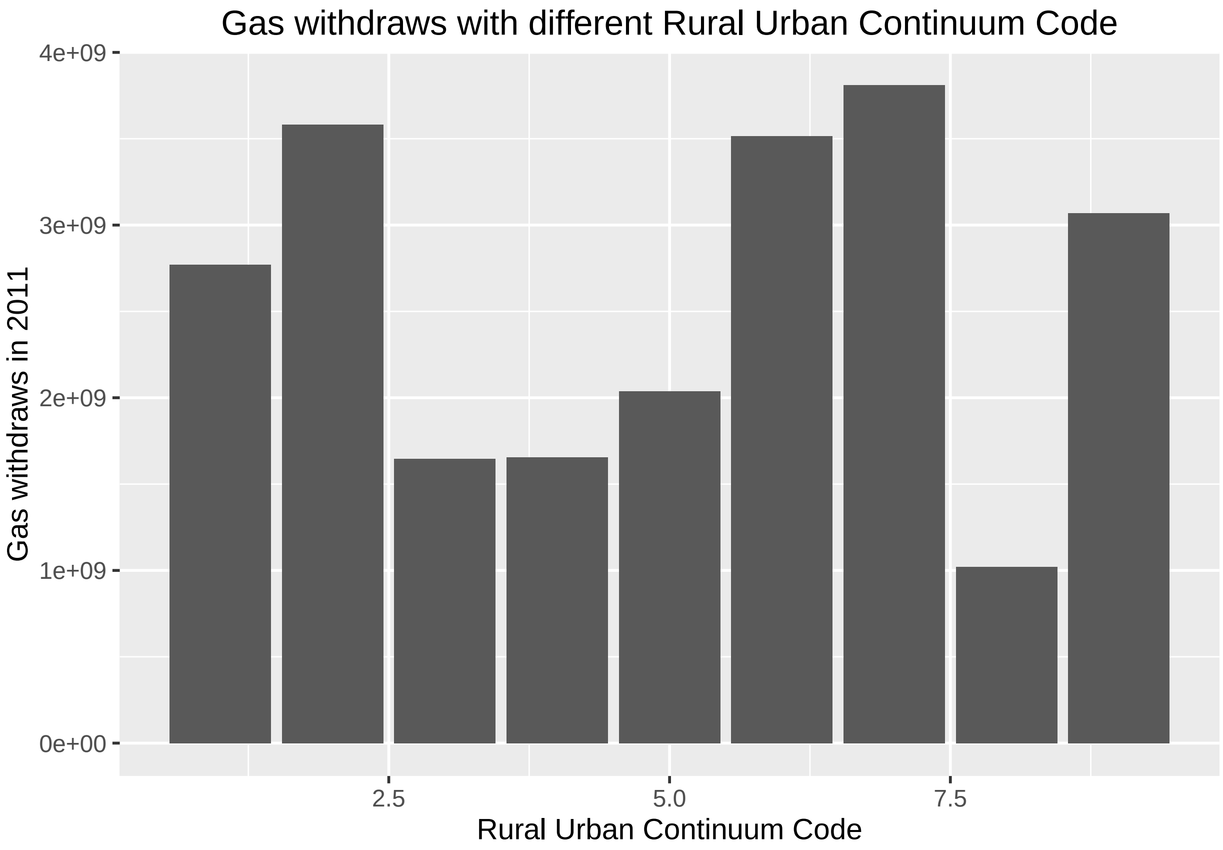
The graghs drawn indicates the proportion of counties in each state with different growth rate of oil, gas or oil and gas annual gross withdrawals from 2000 to 2011. We can see from both graghs that Status Quo takes the biggest proportion in almost every state.

(iii) Oil and gas withdrawals analysis with Rural Urban Continuum Code

```r
#Oil and Gas total gross withdrawals distribution on the Rural Urban Continuum Code
library(ggplot2)


#read tidy data and select oil and gas Rural Urban Continuum Code data in year 2011
dfnew <- read.table("oilTidyData.txt")
dfnew <- data.frame(dfnew)
dfnew.RUC <- select(dfnew, County_Name, Rural_Urban_Continuum_Code_2013, year:gaswithdraw)
dfnew.RUC2011 <- filter(dfnew.RUC, year == 2011)

#bar graphics indicating oil and gas withdraws corresponding to different Rural Urban Continuum Code
gg <- ggplot(data = dfnew.RUC2011, aes(x = Rural_Urban_Continuum_Code_2013, y = oilwithdraw))
gg + geom_bar(stat = "identity") + labs(title = "Oil withdraws with different Rural Urban Continuum Cod
```



```
gg <- ggplot(data = dfnew.RUC2011, aes(x = Rural_Urban_Continuum_Code_2013, y = gaswithdraw))
gg + geom_bar(stat = "identity") + labs(title = "Gas withdraws with different Rural Urban Continuum Cod
```



ERS Rural-Urban Continuum Codes distinguish metropolitan (metro) counties by the population size of their metro area, and nonmetropolitan (nonmetro) counties by degree of urbanization and adjacency to metro areas. The Office of Management and Budget's 2013 metro and nonmetro categories have been subdivided into three metro and six nonmetro groupings, resulting in a nine-part county classification. The codes provide researchers working with county data a more detailed residential classification, beyond a simple metro-nonmetro dichotomy, for the analysis of trends related to degree of rurality and metro proximity.

The values of code and their meanings are listed as follows:

Code	Description
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

These graghs are drawn with data in 2011. Counties in metro areas of 250,000 to 1 million population have the biggest oil withdrawals while Urban population of 20,000 or more, adjacent to a metro area have the smallest.

Counties in Urban population of 2,500 to 19,999, not adjacent to a metro area while Completely rural or less than 2,500 urban population, adjacent to a metro area.

5. Contribution

In this project, we have two main contributions. Firstly, we provide a way of cleaning the raw data to get county-level oil and gas anual gross withdrawals tidy data for further exploration. Secondly, we explore the tidy data by showing a series vivid graghs and charts from two aspects, time and location. These work can be taken as solid foundation for future work.

6. Future Work

Future work can be done in a quantitative way by modeling in Statistics on the basis of descriptive statistics analysis provided in this project.

Some ideas for future work include:
a) What is the relationship between county type(metro or nonmetro) and oil/ gas gross annual withdrawals?
b) What is the relationship between county county population size and oil/ gas gross annual withdrawals?
c) Oil/ gas gross annual withdrawals prediction using data from past years