Midterm Project

Wayne Wang

October 11, 2016

Abstract

This report demonstrates the process of cleaning the USDA County level Oil and Natural Gas Production data in the U.S. during 2000 to 2011.

After obtaining a tidy data set, we explore the data and generates:

1. Trend plots for oil and natural gas production at the national level.
2. Trend plots by metropolitan status.
3. Summary table of value changes regarding oil and natural gas production.

Data cleaning

Import raw data

```r
wd <- "C:/Users/Wayne/Desktop/MA615_DataScienceInR/Assignments/MidtermProject/"
oilgascounty <- fread(paste0(wd,"oilgascounty.csv"), data.table=FALSE)
```

Examine raw data

```r
varnames <- names(oilgascounty)
varnames
```

```
##  [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"
```

> **Looks like the raw data has spread columns of oil and gas by year which require some cleaning prior to data analysis.**

Our goal is to reshape the oilgascounty data into a tidy data that contains:

1. identification variables
2. changes in dollar production during 2000 - 2011 for oil, gas and oil + gas
3. year
4. oil
5. gas

```r
# Separate identification columns, oil, gas into three data set
id <- oilgascounty[varnames[c(1:8,33:35)]] %>% data.table() %>% setkey(FIPS)
oil <- select(oilgascounty, +FIPS, contains("oil2"))
gas <- select(oilgascounty, +FIPS, contains("gas2"))

# Reshape oil, remove "oil" string from year column, prep for merging
oil_tidy <- gather(oil, "label", "oil", 2:ncol(oil)) %>%
            separate(label,into = c("delete","year"), sep="l", remove=TRUE) %>%
            select(-delete) %>%
            data.table() %>%
            setkey(FIPS,year)

# Reshape gas, remove "gas" string from year column, prep for merging
gas_tidy <- gather(gas, "label", "gas", 2:ncol(gas)) %>%
            separate(label,into = c("delete","year"), sep="s", remove=TRUE) %>%
            select(-delete) %>%
            data.table() %>%
            setkey(FIPS,year)

# First merge oil_tidy, gas_tidy by FIPS and year
# then merge with id by FIPS
oilgascounty_tidy <- id[gas_tidy[oil_tidy] %>% setkey(FIPS)]

# Convert oil and gas into numeric variable, using gsub() to remove comma
# from both variables first
oilgascounty_tidy$oil <- gsub(",", "", oilgascounty_tidy$oil) %>% as.numeric()
oilgascounty_tidy$gas <-  gsub(",", "", oilgascounty_tidy$gas) %>% as.numeric()
```

Examine tidy data

```r
names(oilgascounty_tidy)
```

```
##  [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] "oil_change_group"                "gas_change_group"
## [11] "oil_gas_change_group"            "year"
## [13] "gas"                             "oil"
```

All columns are in tidy format. Note that we did not process Rural_urban_* variables since these variables were measured once in 2013 and are likely to serve as cross sectional parameters, that is assuming that these parameters remains the same over the entire study period.

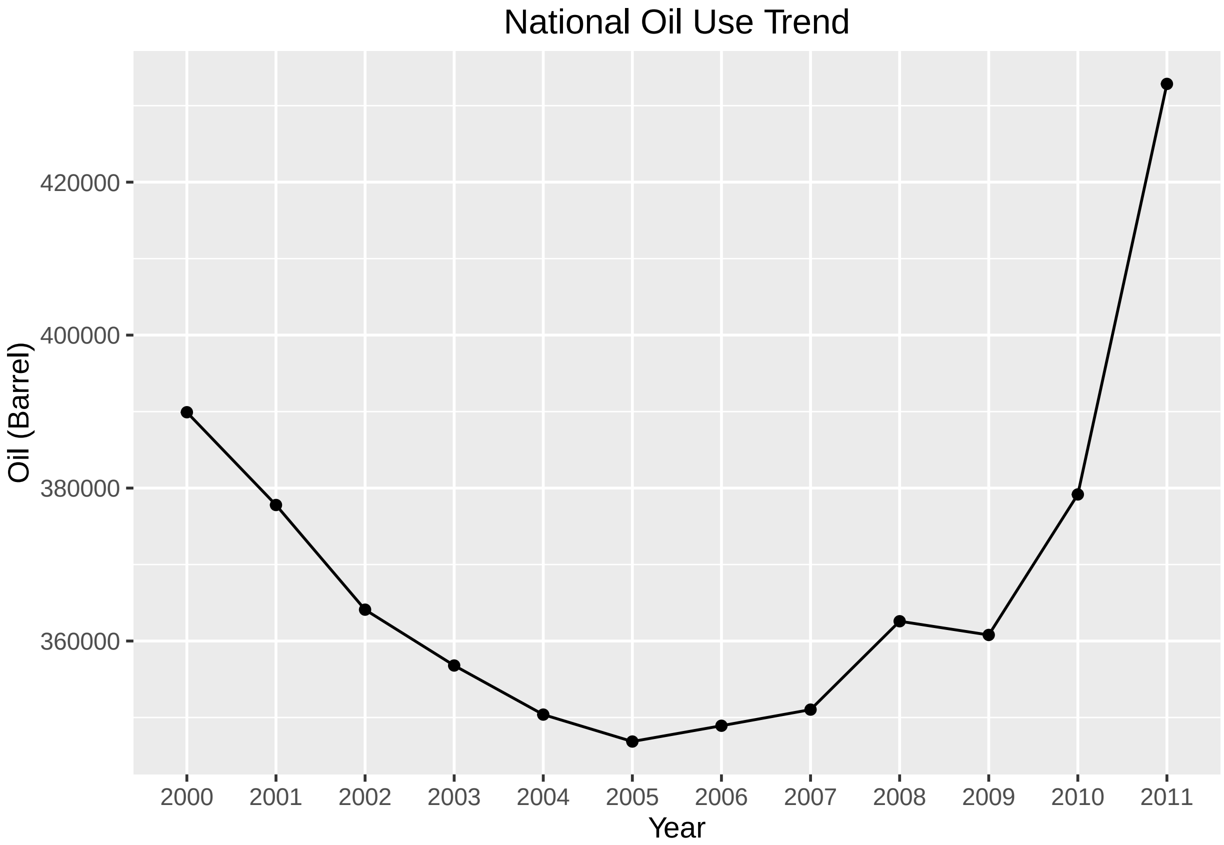
Exploring the tidy data

National Level Summary

```
# National fuel use summary
National_summary <- select(oilgascounty_tidy, c(year,oil,gas)) %>%
                    group_by(year) %>%
                    summarise_each(funs(mean))
```

1. Oil trend

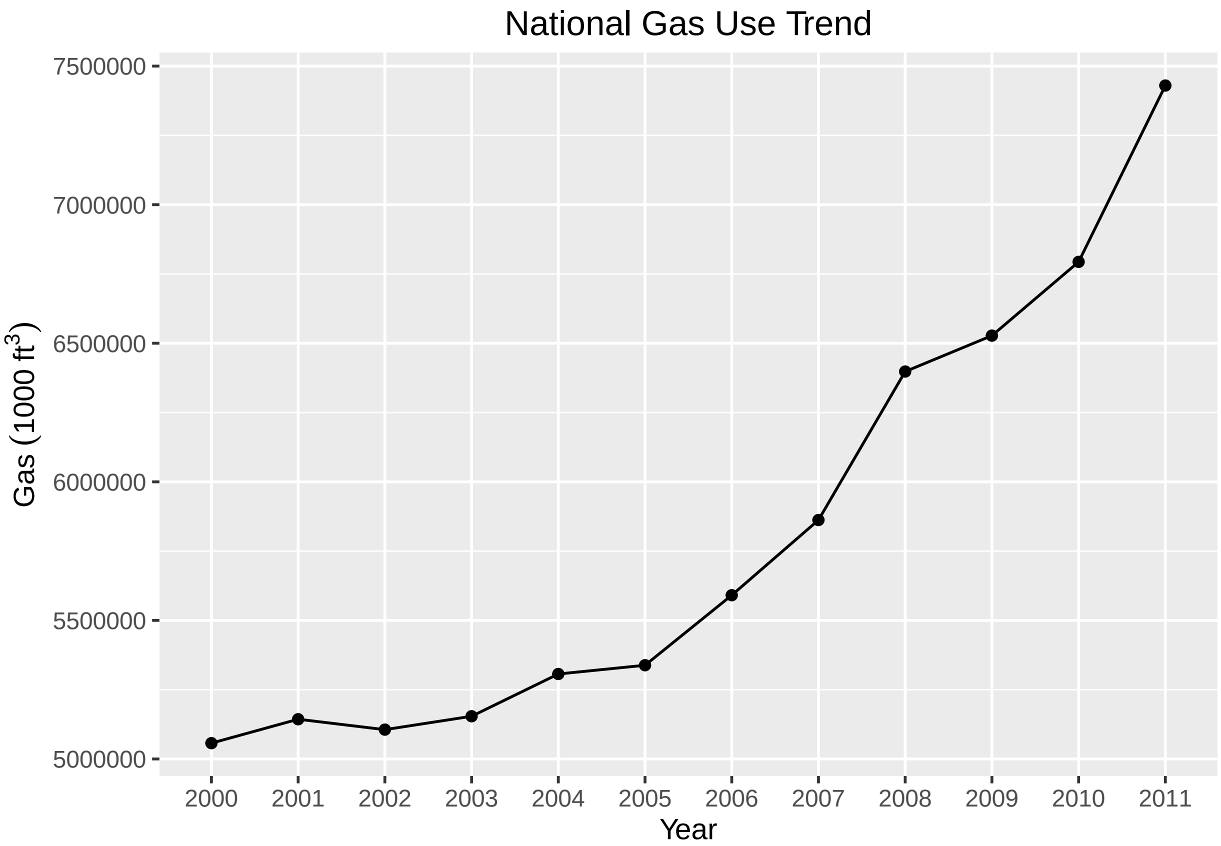
```
# National oil use trend
oil.plot <- ggplot(National_summary, aes(x=year, y=oil, group=1)) +
                   ggtitle("National Oil Use Trend") +
                   ylab("Oil (Barrel)") + xlab("Year") +
                   geom_point() + geom_line()
plot(oil.plot)
```



Oil demand first shows a decrease during 2000~2005 but after 2005 oil demand grows rapidly and exceeded the previous highest demand in 2011.

2. Gas trend

```
gas.plot <- ggplot(National_summary, aes(x=year, y=gas, group=1)) +
                   ggtitle("National Gas Use Trend") +
                   ylab(expression("Gas"~(1000~ft^{3}))) + xlab("Year") +
                   geom_point() + geom_line()
plot(gas.plot)
```



Gas usage generally increases over time, but the increase rate accelerated after 2005.

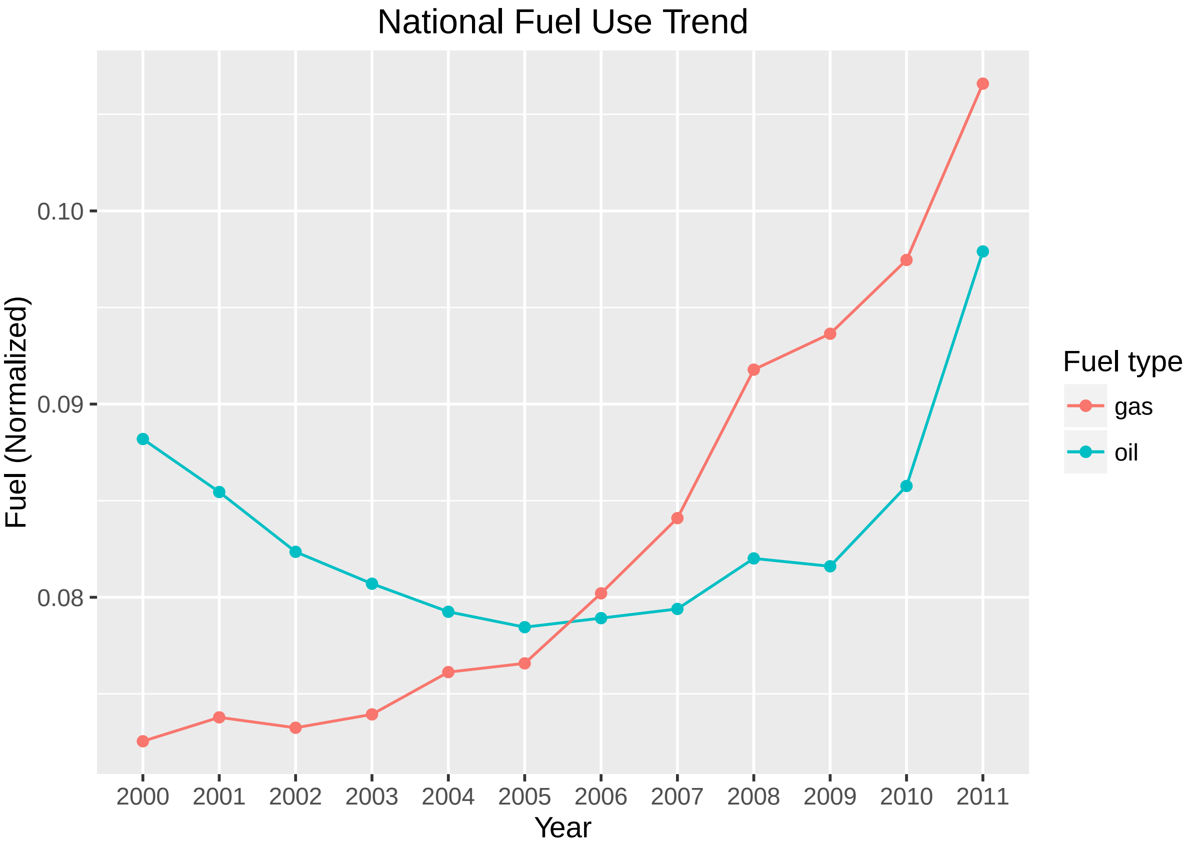
Below we compare the relative consumption of oil and gas together.

Since oil and gas are measured using different units, we need to normalize their values in order to make them comparable.

```
#Normalize gas and oil in order to compare trend over time
National_summary$oil_normal <- National_summary$oil / sum(National_summary$oil)
National_summary$gas_normal <- National_summary$gas / sum(National_summary$gas)
```

```
oil_gas <- ggplot(National_summary,aes(x=year,y=oil_normal,group=1,colour="oil")) +
                  ggtitle("National Fuel Use Trend") +  ylab("Fuel (Normalized)") +
                  xlab("Year") +labs(colour="Fuel type") + geom_point() +
                  geom_line() +
                  geom_point(aes(x=year,y=gas_normal, group=1,colour="gas")) +
                  geom_line(aes(x=year,y=gas_normal, group=1,colour="gas"))

plot(oil_gas)
```



It seems that after 2005, both oil and gas demand increases but gas usage posed a more rapid growth than oil usage.

Now let's examine trends by metropolitan status

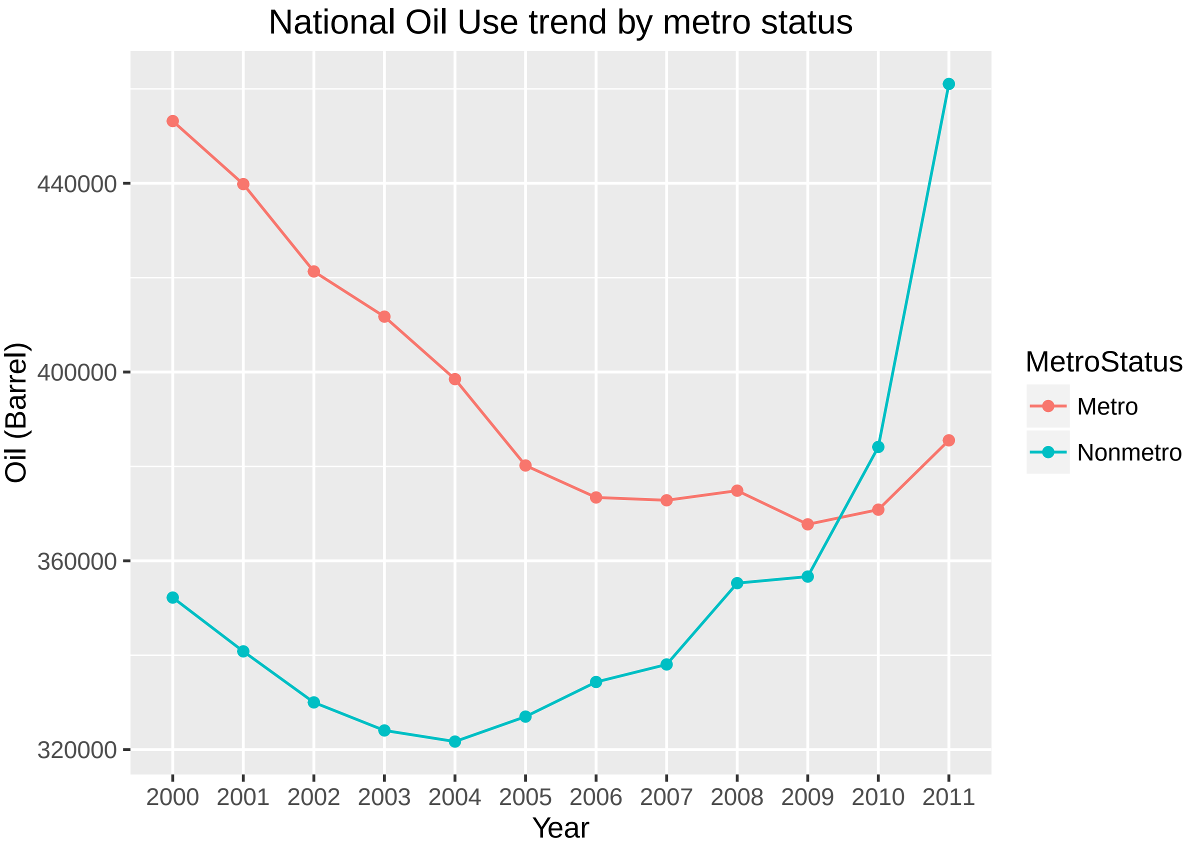
```
National_summary_bymetro <- select(oilgascounty_tidy,c(year,Metro_Nonmetro_2013,oil,gas)) %>%
          group_by(year,Metro_Nonmetro_2013) %>%
          summarise_each(funs(mean))

National_summary_bymetro$MetroStatus <- ifelse(National_summary_bymetro$Metro_Nonmetro_2013==1,'Metro',
```

3. Oil trend by metropolitan status

```
# National oil use trend by metro status
oil.plot2 <- ggplot(National_summary_bymetro,
        aes(x=year, y=oil, group= MetroStatus, colour=MetroStatus)) +
        ggtitle("National Oil Use trend by metro status") +
        ylab("Oil (Barrel)") + xlab("Year") +
        geom_point() +
        geom_line()

plot(oil.plot2)
```

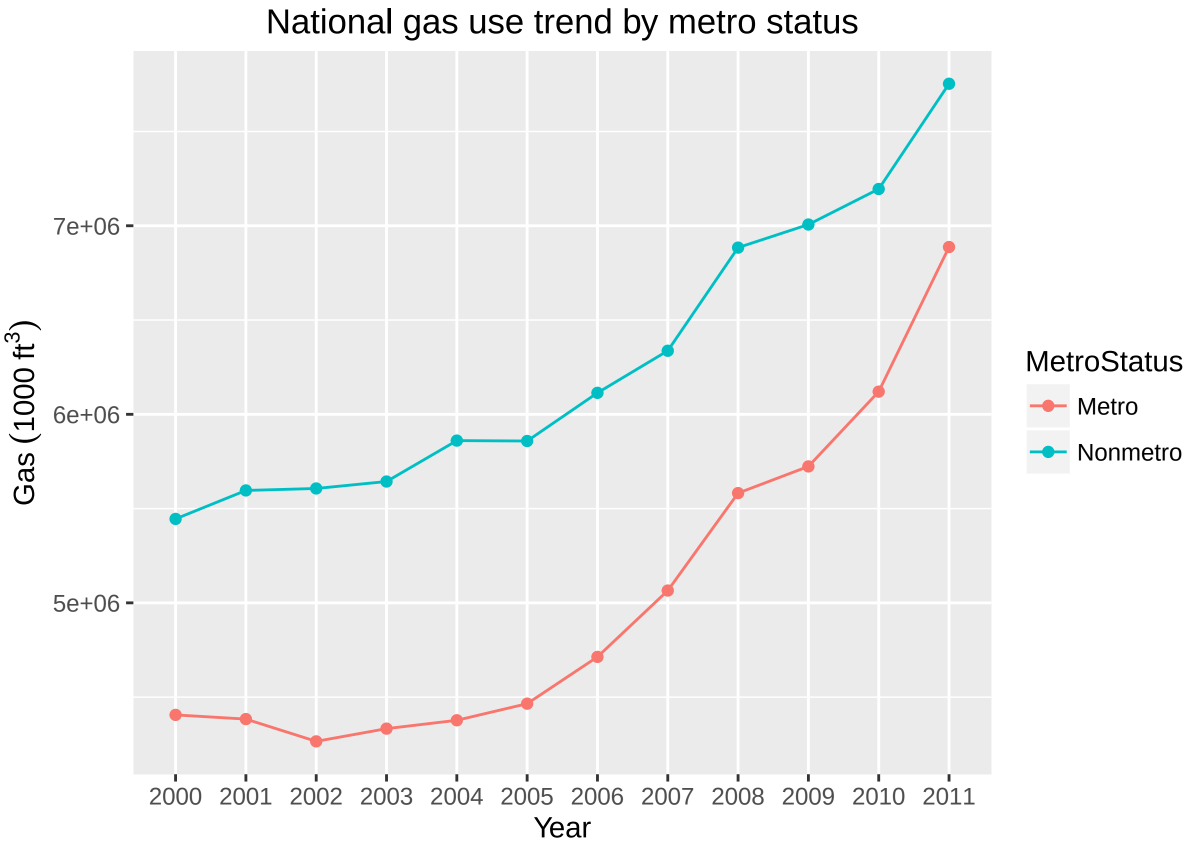


Oil production generally decreased over time in metropolitan counties, while in non metropolitan counties oil production rapidly increased after 2004.

4. Gas trend by metropolitan status

```
# National gas use trend by metro status
gas.plot2 <- ggplot(National_summary_bymetro,
                aes(x=year, y=gas, group= MetroStatus, colour=MetroStatus)) +
                ggtitle("National gas use trend by metro status") +
                ylab(expression("Gas"~(1000~ft^{3}))) + xlab("Year") +
                geom_point() + geom_line()

plot(gas.plot2)
```



> **Both metro and nonmetro counties posed a increase trend in natural gas production over the study period.**

Finally, let's look at the change in dollar value of oil and gas production during the study period.

From data documentation we learned that:

1. High growth (H_Growth): changes in dollar value >= 20 million
2. High decline (H_Decline): changes in dollar value <= -20 million
3. Status Quo (Status Quo): changes between +/- 20 million

For this part of the exploration, we can use the original raw data to examine the proportion of high growth, high decline and status quo for oil and gas of the entire nation.

```
# Oil
total <- as.numeric(nrow(oilgascounty))
oil_g <- round(length(which(oilgascounty$oil_change_group %in% 'H_Growth')) / total * 100, 2)
```

```
oil_d <- round(length(which(oilgascounty$oil_change_group %in% 'H_Decline')) / total *100, 2)
oil_sq <- round(length(which(oilgascounty$oil_change_group %in% 'Status Quo')) / total *100, 2)

# gas
gas_g <- round(length(which(oilgascounty$gas_change_group %in% 'H_Growth')) / total * 100, 2)
gas_d <- round(length(which(oilgascounty$gas_change_group %in% 'H_Decline')) / total *100, 2)
gas_sq <- round(length(which(oilgascounty$gas_change_group %in% 'Status Quo')) / total *100, 2)
```

This table shows the percent of counties of the US of oil/ gas useage growth, decline and status quo

Change (%)	Oil	Gas
Growth	3.44	5.6
Decline	3.92	4.99
Satus Quo	92.63	89.42