Midterm Project for Data Science in R

Quan Zhou

10/23/2016

County-level oil and gas production

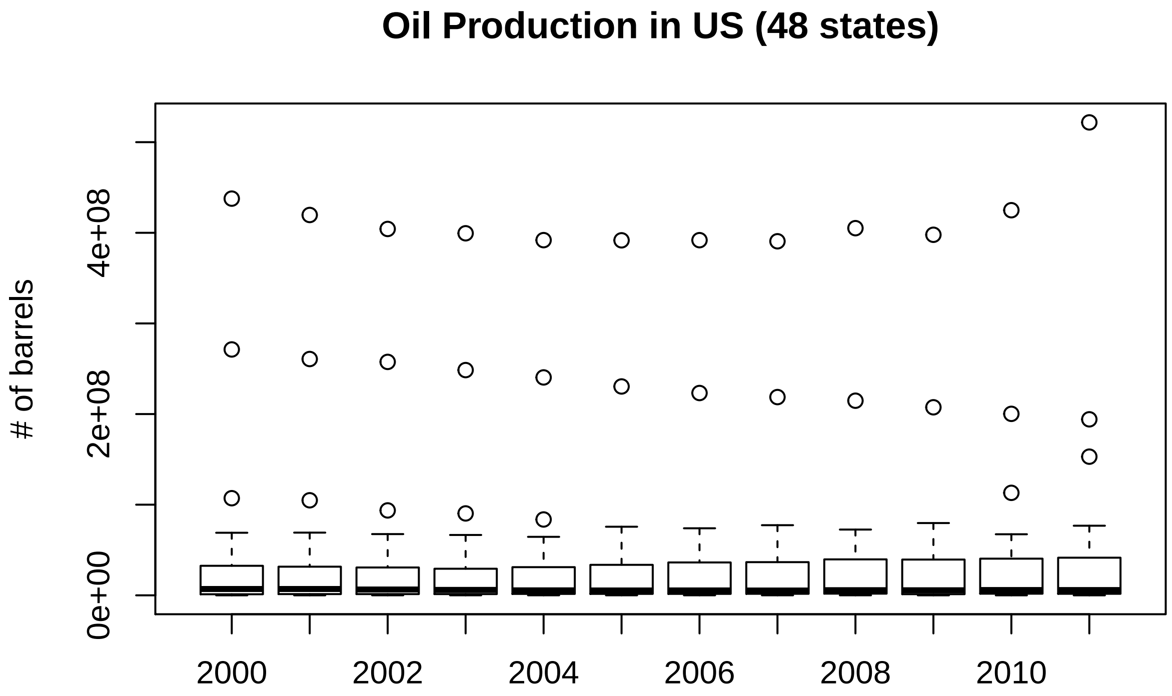
```
# load in the data via the Import Dataset feature in the workspace
df <- read.csv(file = "~/Documents/HW/MA615/MidtermProject/oilgascounty.csv")
set.seed(100)
# remove columns that are not useful
df <- df[,-c(1,2,5,6,7,8,33,34,35)]
# divide dataframe into two dataframes: one called df_oil and the other df_gas
df_oil <- df[c(1,2,3:14)]
df_gas <- df[c(1,2,15:26)]
# convert integer values to numeric
df_oil[, 3:14] <- sapply(df_oil[, 3:14], as.numeric)
df_gas[, 3:14] <- sapply(df_gas[, 3:14], as.numeric)

# split the dataframe by states and store them as a list
# list_oil_by_state <- split(df_oil, df_oil$Stabr)
# list_gas_by_state <- split(df_gas, df_gas$Stabr)

# use aggregate to compute the sum for each state
oilsum<-aggregate(df_oil[, 3:14], list(State=df_oil$Stabr), sum)
gassum<-aggregate(df_gas[, 3:14], list(State=df_gas$Stabr), sum)

# Go through each row and determine if any state produced nothing over 12 years
oilsum<-oilsum[apply(oilsum, 1, function(o) ! ( any(as.numeric(o[2:13])==0))),]
gassum<-gassum[apply(gassum, 1, function(g) ! ( any(as.numeric(g[2:13])==0))),]

# We are done clean and tidy the oil and gas data
saveRDS(oilsum, file="oil.rda")
saveRDS(gassum, file="gas.rda")
```

Plotting oil production

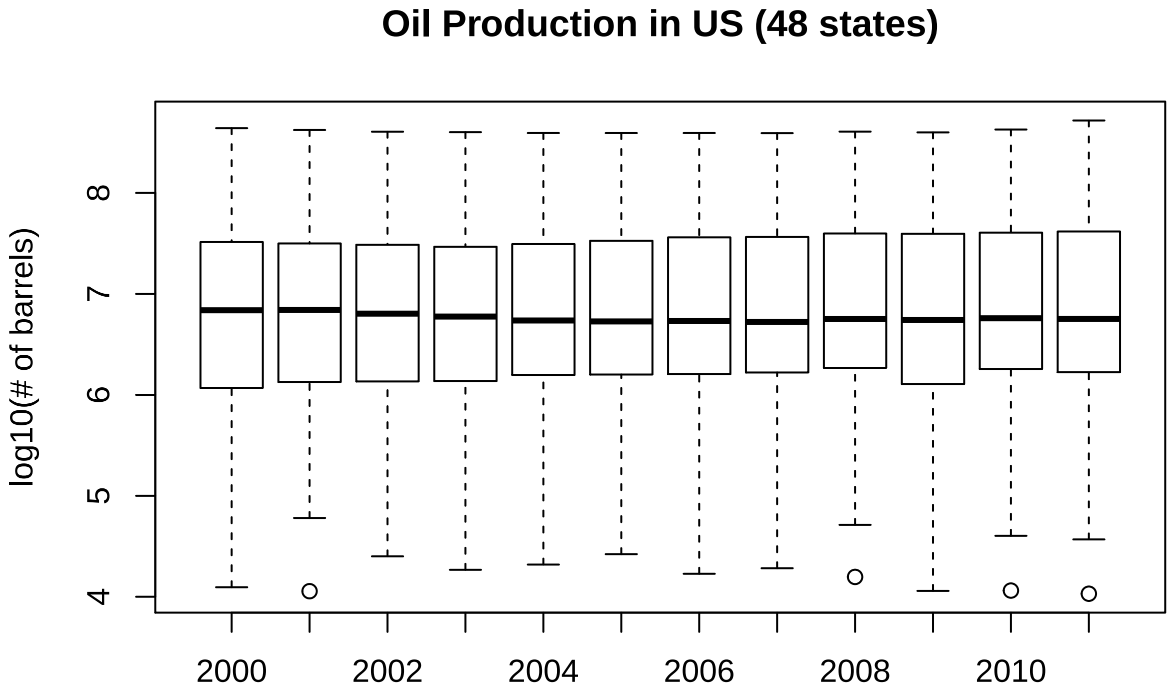
```
# output statistics of oil and gas production
summary(oilsum[2:13])
```

```
##     oil2000             oil2001             oil2002
##  Min.   :    12418   Min.   :    11344   Min.   :    25110
##  1st Qu.:  1224338   1st Qu.:  1376536   1st Qu.:  1428459
##  Median :  6971772   Median :  6970699   Median :  6423510
##  Mean   : 40408306   Mean   : 39151238   Mean   : 37732866
##  3rd Qu.: 29510202   3rd Qu.: 28774348   3rd Qu.: 28182201
##  Max.   :437700231   Max.   :419634532   Max.   :404223421
##     oil2003             oil2004             oil2005
##  Min.   :    18489   Min.   :    20816   Min.   :    26417
##  1st Qu.:  1437801   1st Qu.:  1611746   1st Qu.:  1589299
##  Median :  5982368   Median :  5469592   Median :  5344706
##  Mean   : 36976349   Mean   : 36311026   Mean   : 35946045
##  3rd Qu.: 27405895   3rd Qu.: 29493290   3rd Qu.: 33411226
##  Max.   :399461473   Max.   :391896994   Max.   :391691263
##     oil2006             oil2007             oil2008
##  Min.   :    16881   Min.   :    19155   Min.   :    15712
##  1st Qu.:  1628042   1st Qu.:  1677861   1st Qu.:  1883152
##  Median :  5392808   Median :  5302684   Median :  5659828
##  Mean   : 36160064   Mean   : 36378418   Mean   : 37576201
##  3rd Qu.: 36137736   3rd Qu.: 36169382   3rd Qu.: 37648659
##  Max.   :391870785   Max.   :390621796   Max.   :405114648
##     oil2009             oil2010             oil2011
##  Min.   :    11430   Min.   :    11508   Min.   :    10712
##  1st Qu.:  1374993   1st Qu.:  1809066   1st Qu.:  1734578
##  Median :  5550575   Median :  5822812   Median :  5796684
##  Mean   : 37390160   Mean   : 39294441   Mean   : 44857266
##  3rd Qu.: 37155044   3rd Qu.: 38538030   3rd Qu.: 40928700
##  Max.   :397818942   Max.   :424899287   Max.   :521790261
```

```
# Normal boxplot
boxplot(oilsum[2:13], names = c("2000", "2001", "2002", "2003", "2004", "2005",
    "2006", "2007", "2008", "2009", "2010", "2011"), main = "Oil Production in US (48 states)",
    ylab = "# of barrels")
```



```
# since the top producing states, TX, ND, and AL have much larger values, we
# use log
log_oilsum <- log10(oilsum[2:13])
boxplot(log_oilsum[1:12], names = c("2000", "2001", "2002", "2003", "2004",
    "2005", "2006", "2007", "2008", "2009", "2010", "2011"), main = "Oil Production in US (48 states)",
    ylab = "log10(# of barrels)")
# a better visualization for log plot using ggplot/plotly
library(ggplot2)
```



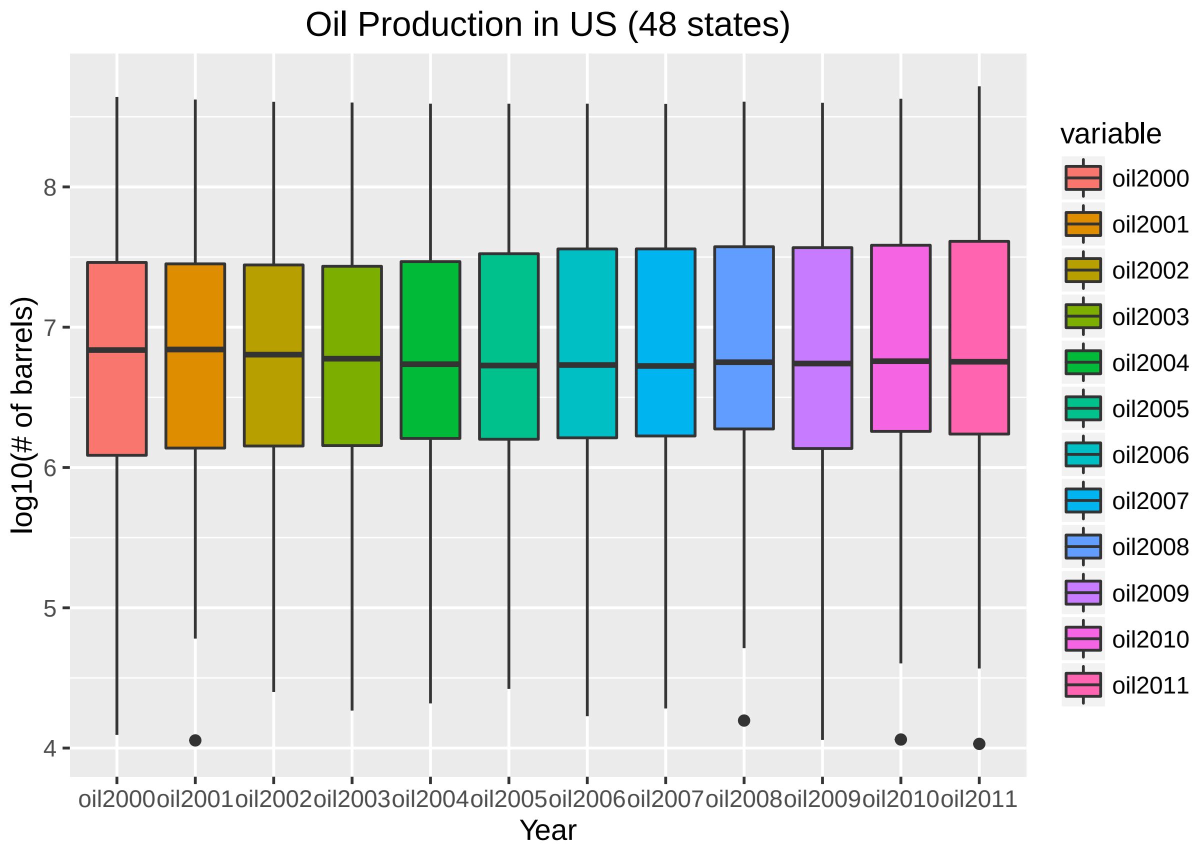
```
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
# need reshape to further simply the two-way data
library("reshape2", lib.loc = "/Library/Frameworks/R.framework/Versions/3.3/Resources/library")
oilsum2 <- melt(log_oilsum)
```

```
## No id variables; using all as measure variables
```

```r
# convert year into vector

ggplot(oilsum2, aes(x = variable, y = value)) + geom_boxplot(aes(fill = variable)) +
    xlab("Year") + ylab("log10(# of barrels)") + ggtitle("Oil Production in US (48 states)")
```



##

Plotting Gas production

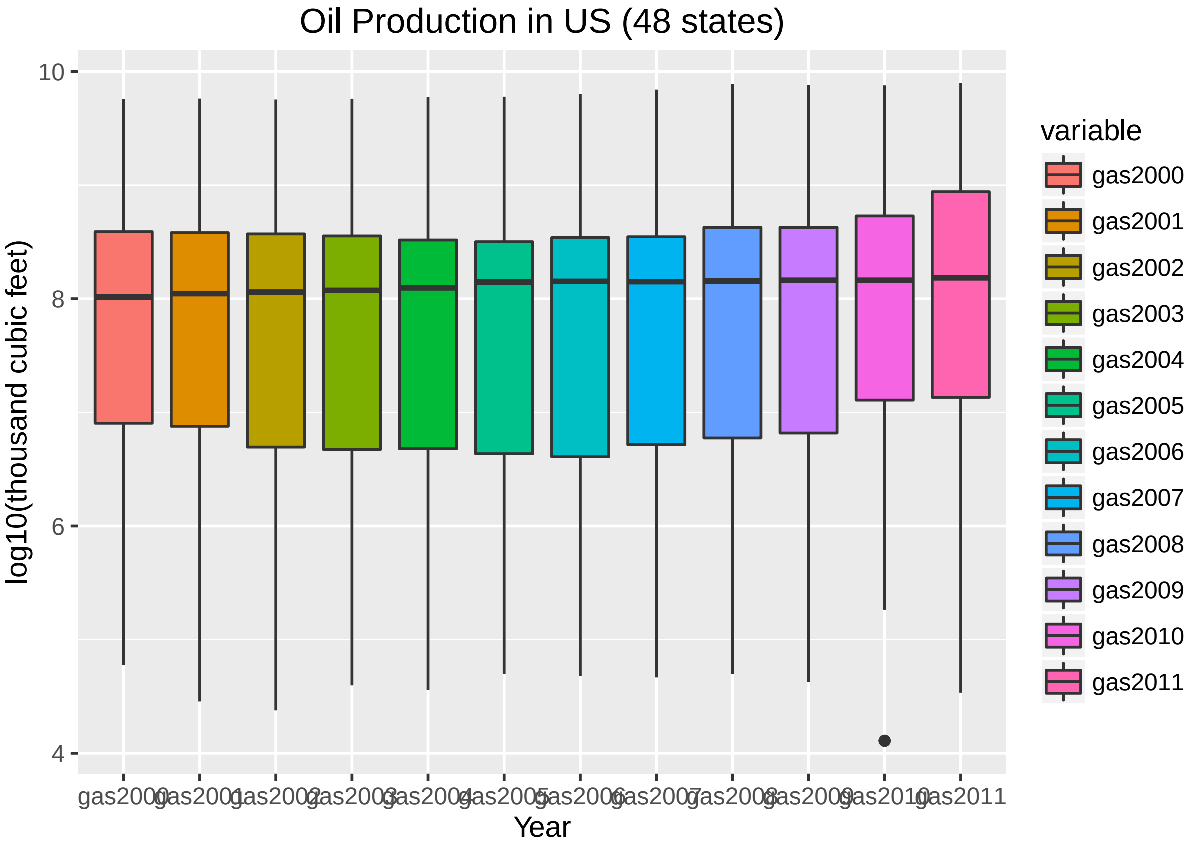
```r
# output statistics of gas production
summary(gassum[2:13])
```

```
##     gas2000             gas2001             gas2002
##  Min.   :5.946e+04   Min.   :2.855e+04   Min.   :2.382e+04
##  1st Qu.:8.152e+06   1st Qu.:7.781e+06   1st Qu.:5.589e+06
##  Median :1.042e+08   Median :1.137e+08   Median :1.184e+08
##  Mean   :5.241e+08   Mean   :5.330e+08   Mean   :5.291e+08
##  3rd Qu.:3.899e+08   3rd Qu.:3.819e+08   3rd Qu.:3.730e+08
##  Max.   :5.713e+09   Max.   :5.780e+09   Max.   :5.677e+09
##     gas2003             gas2004             gas2005
##  Min.   :3.955e+04   Min.   :3.582e+04   Min.   :4.959e+04
##  1st Qu.:5.523e+06   1st Qu.:5.602e+06   1st Qu.:5.185e+06
##  Median :1.243e+08   Median :1.285e+08   Median :1.454e+08
##  Mean   :5.341e+08   Mean   :5.499e+08   Mean   :5.532e+08
##  3rd Qu.:3.582e+08   3rd Qu.:3.297e+08   3rd Qu.:3.176e+08
##  Max.   :5.770e+09   Max.   :5.998e+09   Max.   :6.009e+09
##     gas2006             gas2007             gas2008
##  Min.   :4.757e+04   Min.   :4.646e+04   Min.   :4.954e+04
##  1st Qu.:4.868e+06   1st Qu.:5.926e+06   1st Qu.:6.550e+06
##  Median :1.458e+08   Median :1.435e+08   Median :1.445e+08
##  Mean   :5.794e+08   Mean   :6.075e+08   Mean   :6.630e+08
##  3rd Qu.:3.456e+08   3rd Qu.:3.526e+08   3rd Qu.:4.264e+08
##  Max.   :6.350e+09   Max.   :6.938e+09   Max.   :7.778e+09
##     gas2009             gas2010             gas2011
##  Min.   :4.255e+04   Min.   :1.287e+04   Min.   :3.411e+04
##  1st Qu.:7.174e+06   1st Qu.:1.285e+07   1st Qu.:1.371e+07
##  Median :1.456e+08   Median :1.455e+08   Median :1.531e+08
##  Mean   :6.765e+08   Mean   :7.040e+08   Mean   :7.700e+08
##  3rd Qu.:4.270e+08   3rd Qu.:5.393e+08   3rd Qu.:9.274e+08
##  Max.   :7.654e+09   Max.   :7.559e+09   Max.   :7.906e+09
```

```
# plot using ggplot/plotly
gassum2 <- melt(log10(gassum[2:13]))
```

```
## No id variables; using all as measure variables
```

```
# convert year into vector
ggplot(gassum2, aes(x = variable, y = value)) + geom_boxplot(aes(fill = variable)) +
    xlab("Year") + ylab("log10(thousand cubic feet)") + ggtitle("Oil Production in US (48 states)")
```


Oil Production in US (48 states)
log10(thousand cubic feet)
gas2000
gas2001
gas2002
gas2003
gas2004
gas2005
gas2006
gas2007
gas2008
gas2009
gas2010
gas2011
Year
variable
gas2000
gas2001
gas2002
gas2003
gas2004
gas2005
gas2006
gas2007
gas2008
gas2009
gas2010
gas2011