EDA - Exam 2

Aaron Niskin

December 06, 2016

This is a 90 minute exam. You are to write an R Markdown document that provides scripts for carrying out the following tasks. It is the work, not the answers, which will be evaluated.

You may use the electronic resources at your disposal, but please do your own work.

Do as much as you can. Mail your pdf to mcdonald@ncf.edu at the end of the 90 minute period.

Navigate to the following page:

https://www.kaggle.com/lislejoem/us_energy_census_gdp_10-14

Read the annotation. Download the dataset and read it into memory (this may require you to register at kaggle - it's easy). Alternatively, grab the data from the canvas page under "files," then "Data." Initial exploration

1. **Vet the data for completeness. Briefly discuss your conclusions.**

   Our data has 52, 1 rows, and 51 complete rows.

```
tmp <- complete_cases(data)
tmp <- tmp[tmp$PercentComplete < 100, ]
```

```
printDf(tmp, title = "Fields with less than 100 percent complete cases")
```

FieldName	CompleteCases	PercentComplete
Region	51	98.08
Division	51	98.08
Coast	51	98.08
Great.Lakes	51	98.08
RDOMESTICMIG2011	51	98.08
RDOMESTICMIG2012	51	98.08
RDOMESTICMIG2013	51	98.08
RDOMESTICMIG2014	51	98.08

Table 1: Fields with less than 100 percent complete cases

   So for a data set with 52 rows, it's doing pretty well on completeness. Furthermore, that one "incomplete case" is for the entire US and those fields don't apply. So it's not really an 'incomplete' case.

2. **Filter out any observations which do not correpsond to US states.**

   This works because every state has complete cases.

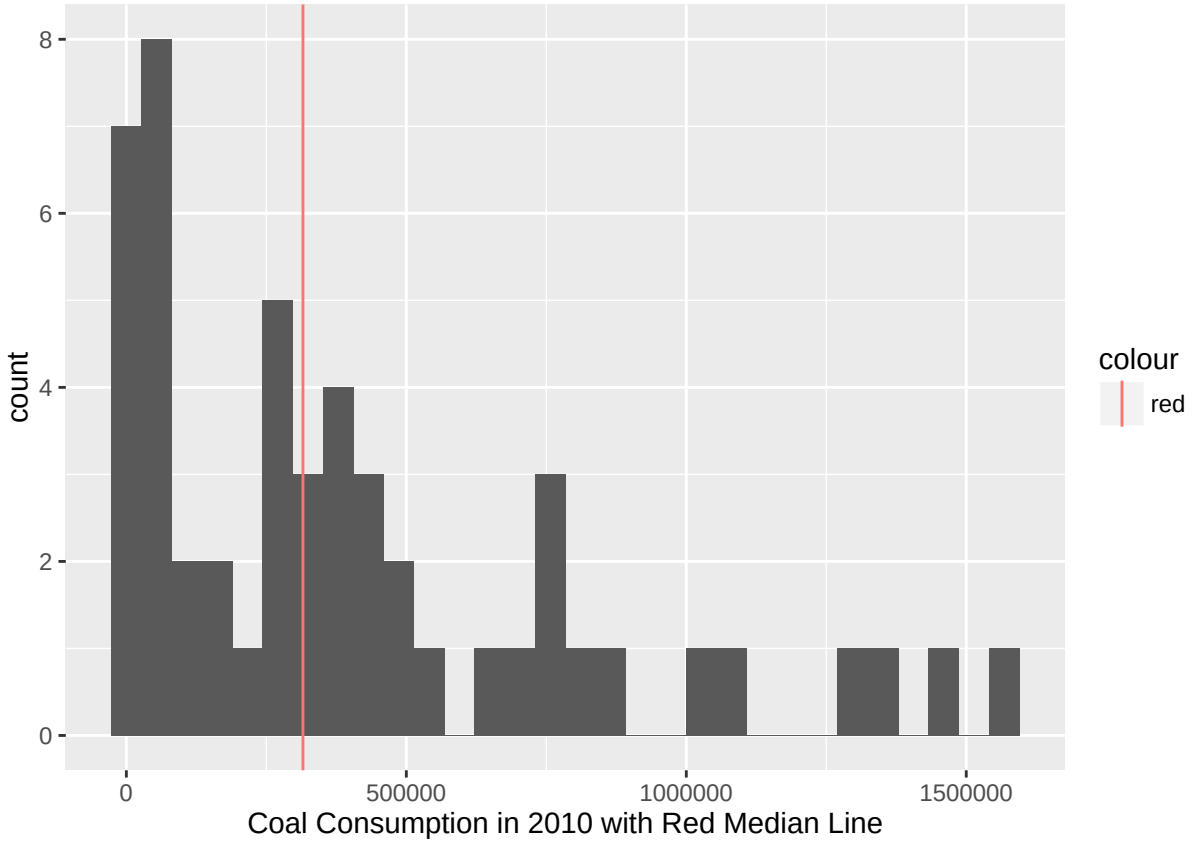
```
data_states <- data[complete.cases(data), ]
```

3. **Subset the energy variables to produce a data frame which contains information concerning coal and electricity data (as well as all variables not contained in the energy variables).**

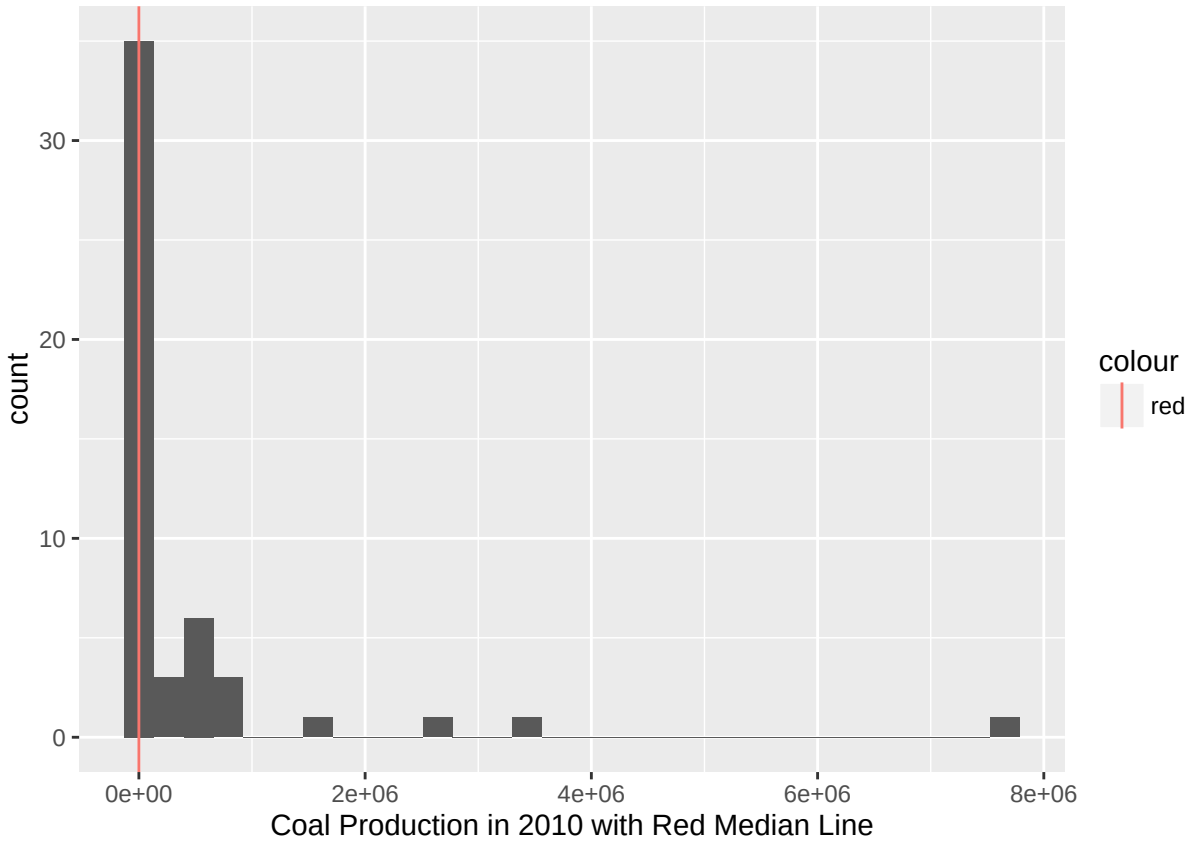
```
data_states_coal_elect <- data_states[, c(1:5, 47:81, 122:length(names(data_states)))]
```

4. **Using ggplot2, create histograms and boxplots for total coal consumption and total coal production for the year 2010. Identify any production outliers above median.**

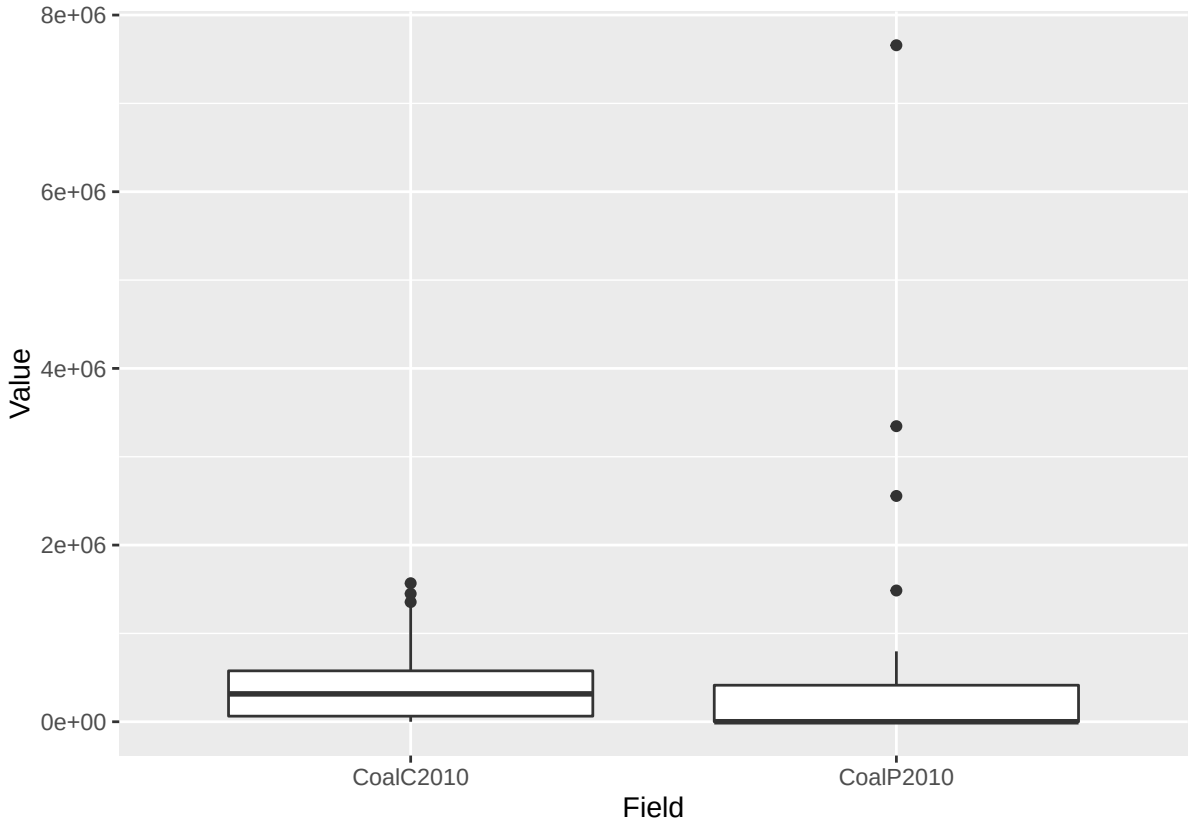
```
plt <- ggplot(data = data_states_coal_elect)
plt + geom_histogram(aes(x = CoalC2010)) + labs(x = "Coal Consumption in 2010 with Red Median Line"
    geom_vline(aes(xintercept = median(CoalC2010), color = "red"))
```



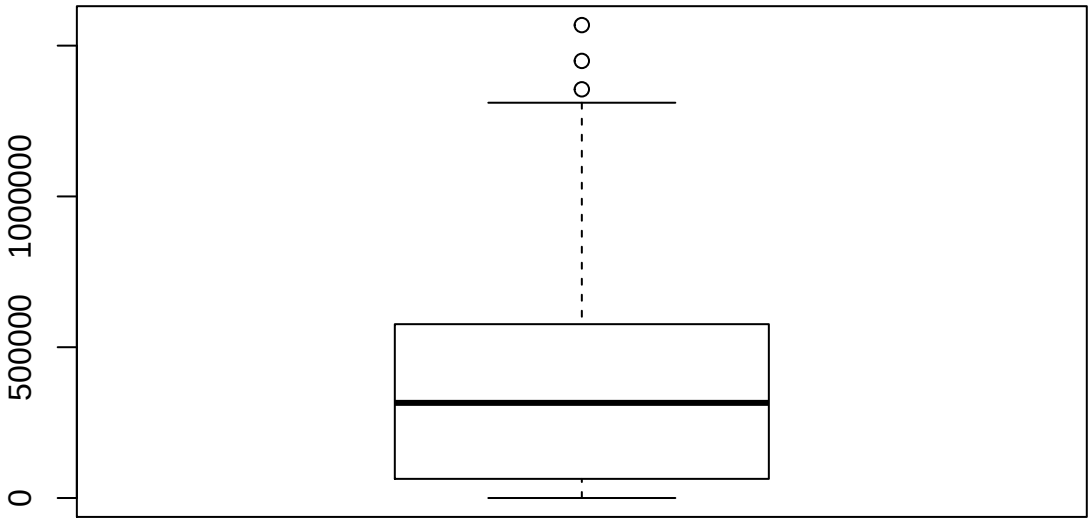

```
plt + geom_histogram(aes(x = CoalP2010)) + labs(x = "Coal Production in 2010 with Red Median Line")
    geom_vline(aes(xintercept = median(CoalP2010), color = "red"))
```



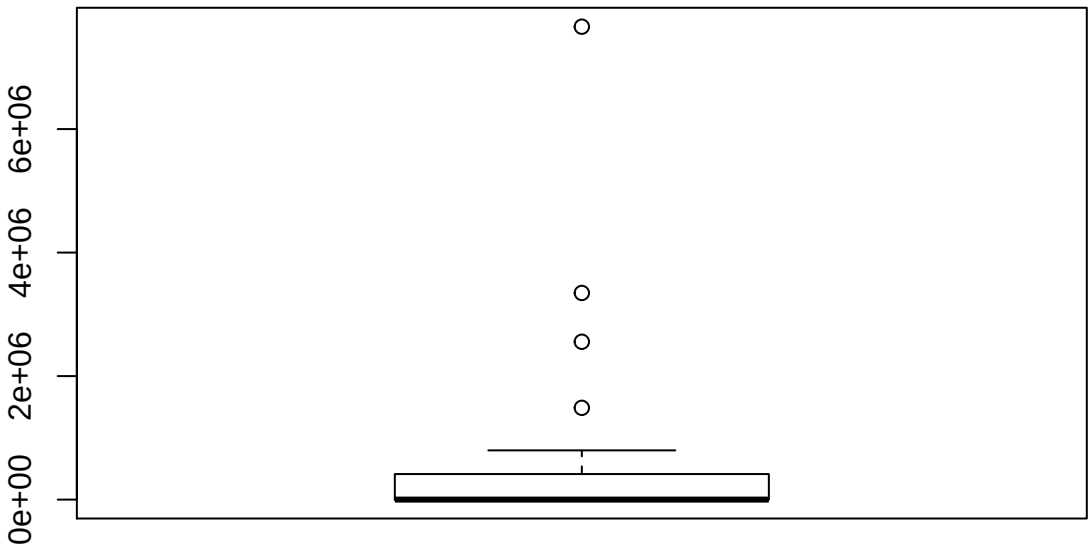

```
tmp <- data_states_coal_elect %>% gather("Field", "Value", c(6, 11))
ggplot(data = tmp) + geom_boxplot(aes(x = Field, y = Value))
```




```
rm(tmp)
boxplot(data_states_coal_elect$CoalC2010)
```




```
boxplot(data_states_coal_elect$CoalP2010)
```



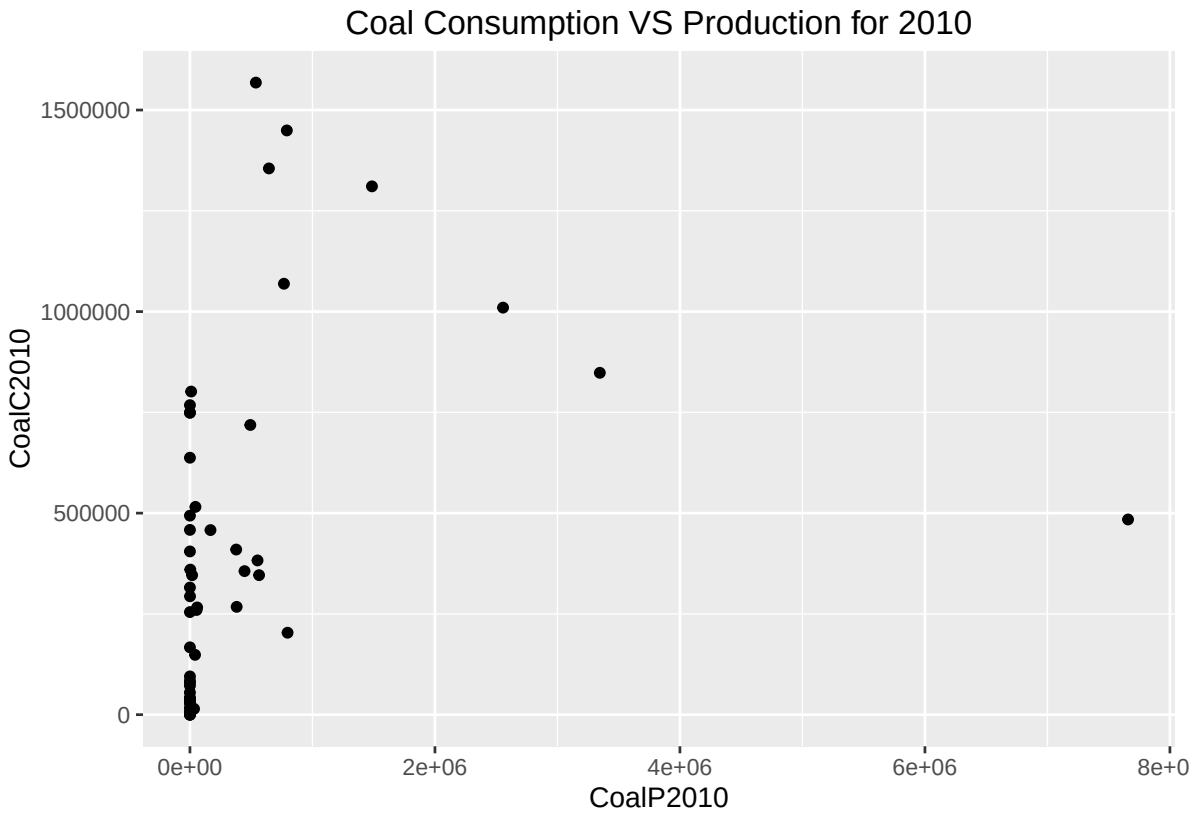

```
prod_med = median(data_states_coal_elect$CoalP2010)
data_states_coal_elect[data_states_coal_elect$CoalP2010 > prod_med, "StateCodes"]
```

```
##  [1] AK AL AR AZ CO IL IN KS KY LA MD MO MS MT ND NM OH OK PA TN TX UT VA
## [24] WV WY
## 52 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY
```

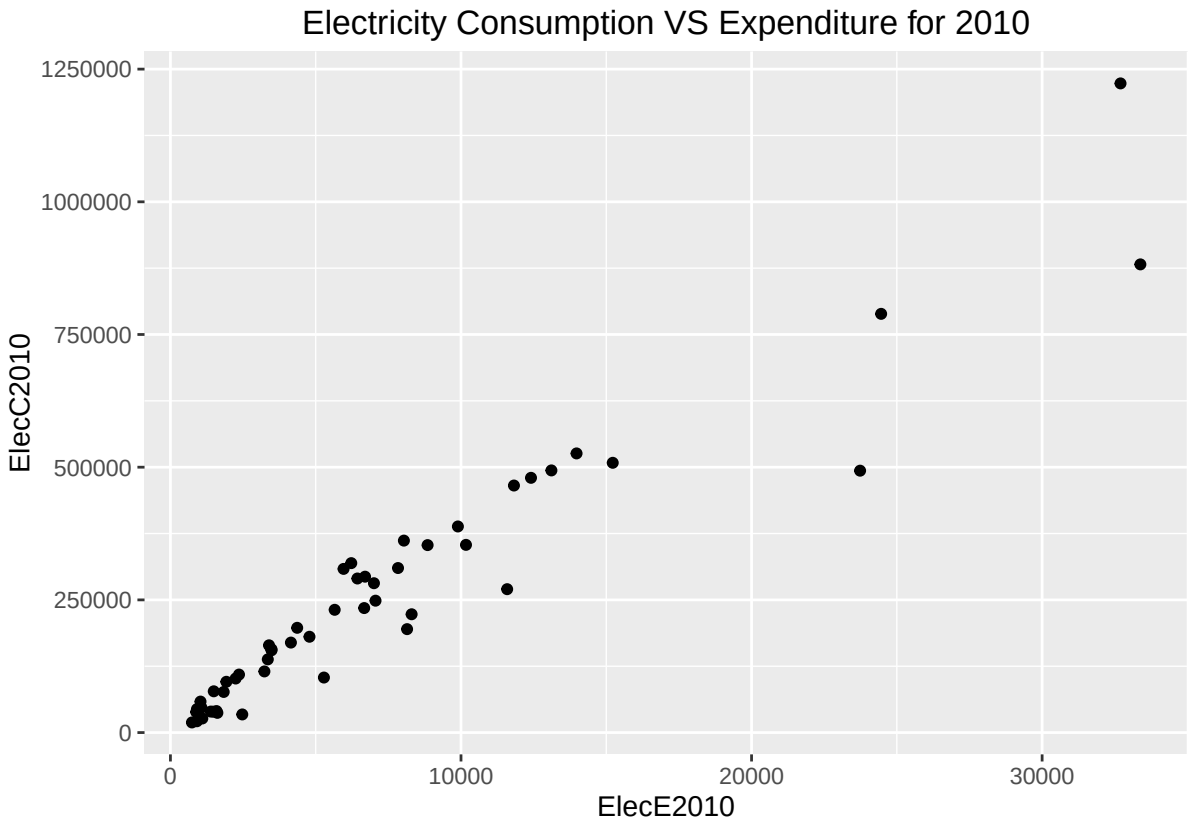
This is because the median is zero. So any coal producing state would be an outlier accourding to this definition.

5. **Using ggplot2, create a scatterplots for total coal consumption/production and total electricity con- sumption/production for the year 2010. Discuss correlation.**

```
ggplot(data = data_states_coal_elect) + geom_point(aes(x = CoalP2010, y = CoalC2010)) +
    labs(title = "Coal Consumption VS Production for 2010")
```

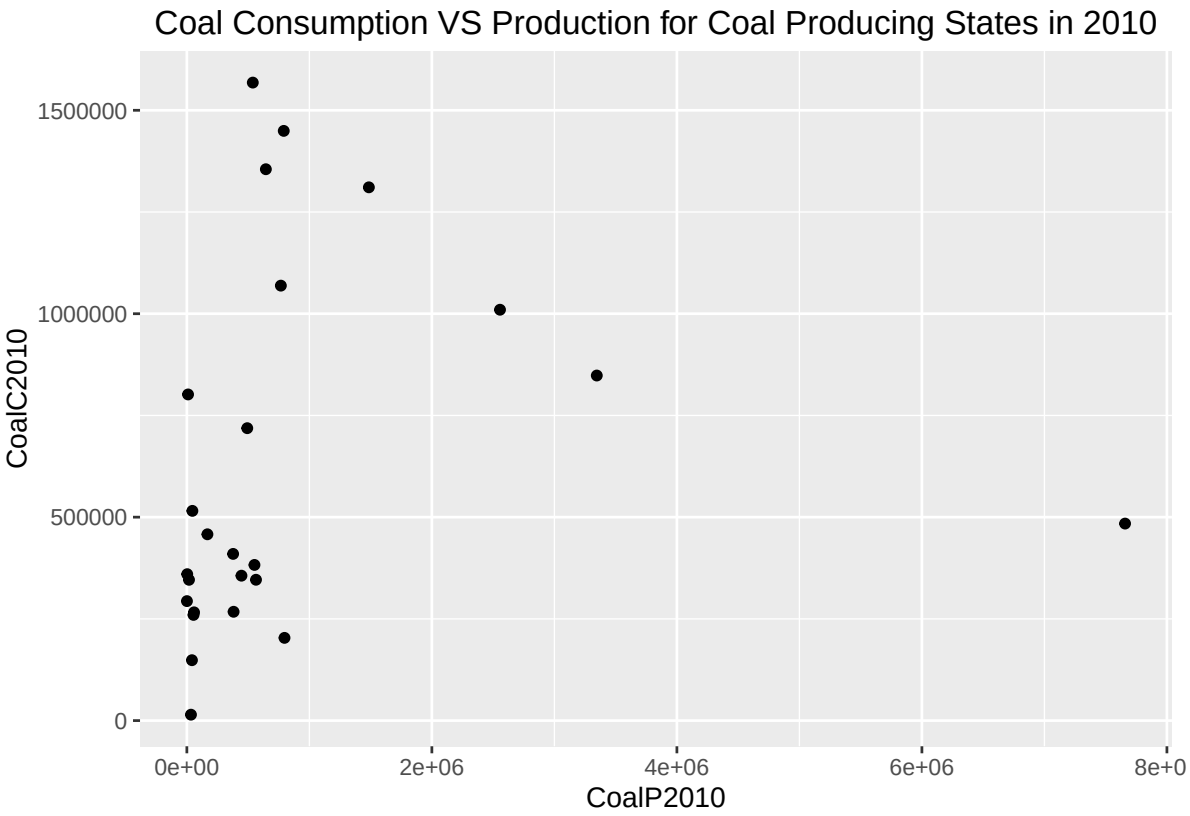



```
ggplot(data = data_states_coal_elect) + geom_point(aes(x = ElecE2010, y = ElecC2010)) +
    labs(title = "Electricity Consumption VS Expenditure for 2010")
```



The electricity consumption is highly correlated with expenditure, but coal consumption is not very highly correlated with production.

```
ggplot(data = data_states_coal_elect[data_states_coal_elect$CoalP2010 > 0, ]) +
    geom_point(aes(x = CoalP2010, y = CoalC2010)) + labs(title = "Coal Consumption VS Production fo
```



And still, nothing really too convincing there.

Aggregation

6. **Aggregate total coal consumption and total coal production by region. Report the corresponding IQRs.**

```
tmp <- getNumericCols(data_states_coal_elect) %>% group_by(by = Region) %>%
    summarise_all(funs(sum))
tmp_quants <- quantile(tmp$CoalC2010)
tmp_quants[4] - tmp_quants[2]
```

```
##     75%
## 5935172
```

The above was just for CoalC2010, but you can `lapply` for the rest.

7. **Aggregate total coal consumption and total coal production by division. Report the corresponding medians.**

```
tmp <- getNumericCols(data_states_coal_elect) %>% group_by(by = Division) %>%
    summarise_all(funs(sum))
tmp <- tmp %>% summarise_all(funs(median))
```

```
printDf(tmp[, 6:dim(tmp)[2]])
```

CoalC2010	CoalC2011	CoalC2012	CoalC2013	CoalC2014	CoalP2010	CoalP2011	CoalP2012	CoalP2013	C
2392434	2250328	1962251	2061844	2033410	1485775	1511491	1390644	1379262	

Table 2:

Mutation and filtering

8. **For the year 2010, normalize total coal consumption and total coal production by population and add this as another variable. Report the range of the new variable**

```
data_states_coal_elect$CoalC2010_pop_norm <- data_states_coal_elect$CoalC2010/data_states_coal_elec
max(data_states_coal_elect$CoalC2010_pop_norm) - min(data_states_coal_elect$CoalC2010_pop_norm)
```

```
## [1] 0.8580351
```

9. **Filter the new variable to retain the observations corresponding to the top ten values. Report the state abbreviations of the 2010 populations of the corresponding states.**

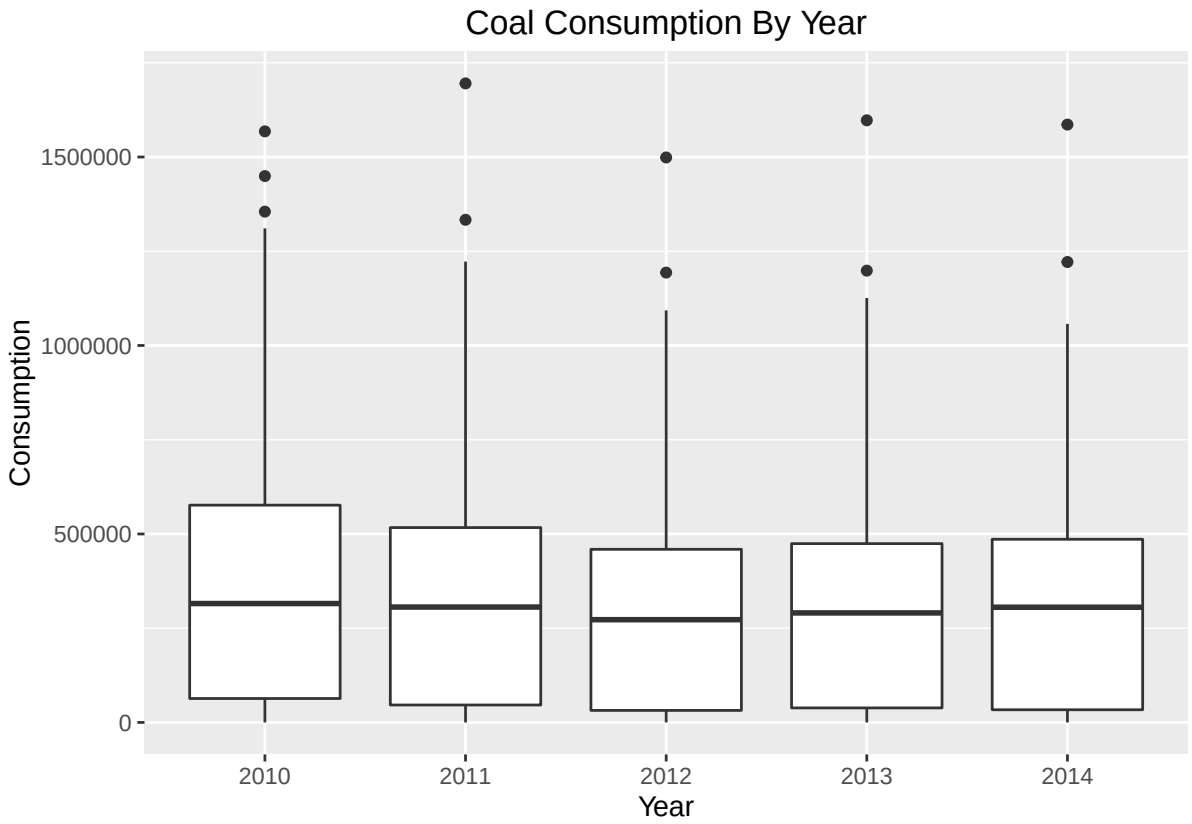
```
data_states_coal_elect <- data_states_coal_elect %>% arrange(desc(CoalC2010_pop_norm))
data_states_coal_elect[1:10, "StateCodes"]
```

```
##  [1] WY ND WV KY IN MT IA AZ AL NE
## 52 Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA ... WY
```

Plotting multiple variables

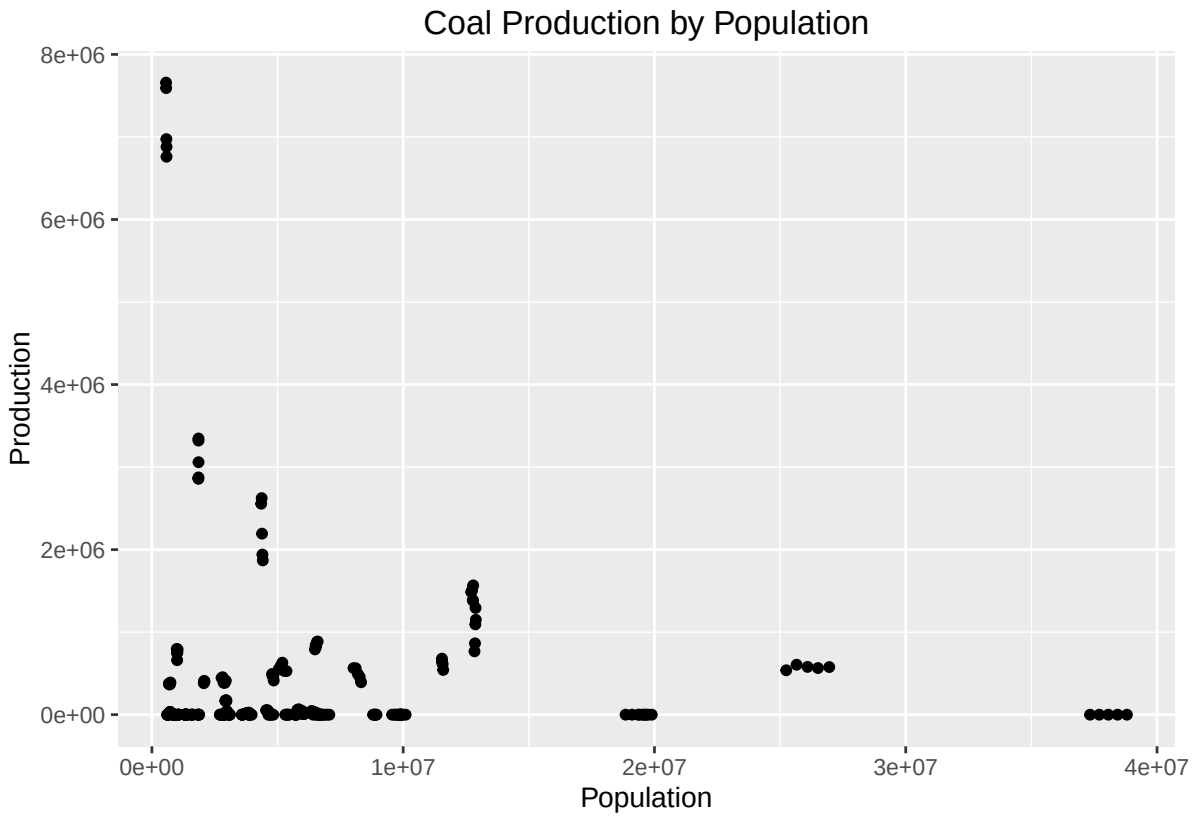
10. **Using ggplot2, facet on year and generate boxplots for total coal consumption (you might consider reshaping the data).**

```
tmp <- data_states_coal_elect %>% gather("Year", "Consumption", c(6:10))
tmp$Year <- gsub("CoalC", "", tmp$Year)
ggplot(data = tmp) + geom_boxplot(aes(x = Year, y = Consumption)) + labs(title = "Coal Consumption
```

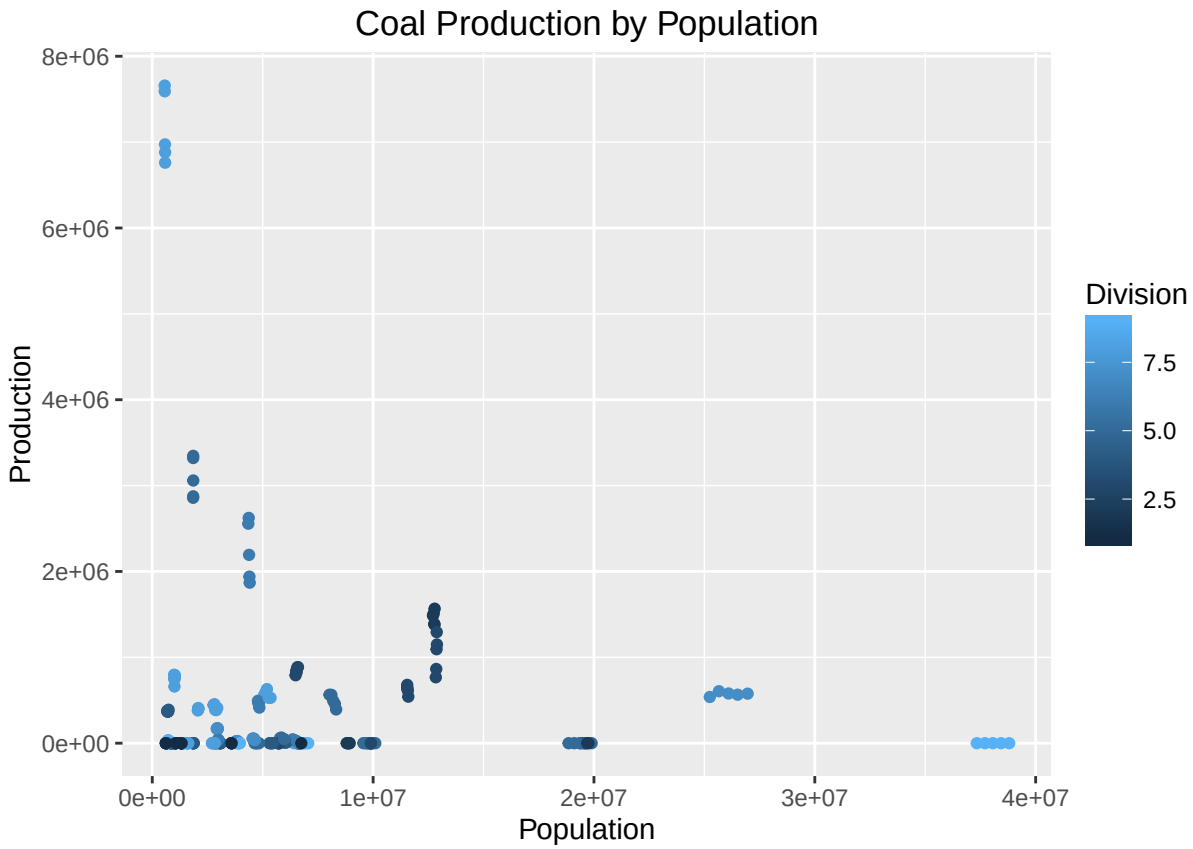


11. **Generate a scatterplot of coal production and population for the year 2010 and color by region. Generate a second plot colored by division.**

```
tmp2 <- data_states %>% gather("Year", "Population", c(163:167))
tmp2$Year <- gsub("POPESTIMATE", "", tmp2$Year)
tmp <- inner_join(tmp, tmp2, by = c("Year", "Region", "StateCodes", "Division"))
tmp2 <- data_states_coal_elect %>% gather("Year", "Production", c(11:15))
tmp2$Year <- gsub("CoalP", "", tmp2$Year)
tmp <- inner_join(tmp, tmp2, by = c("Year", "Region", "StateCodes", "Division"))
ggplot(data = tmp) + geom_point(aes(y = Production, x = Population)) + labs(title = "Coal Productio
```

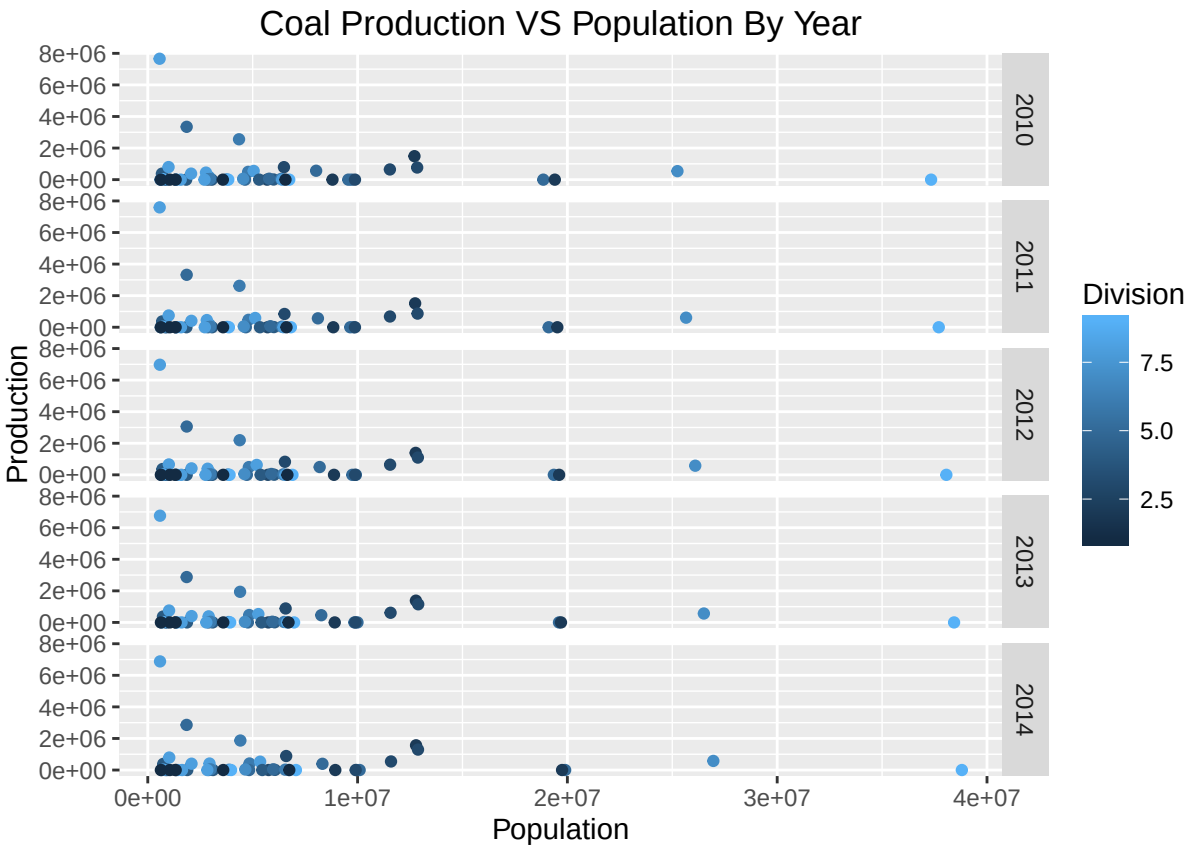



```
ggplot(data = tmp) + geom_point(aes(y = Production, x = Population, color = Division)) +
    labs(title = "Coal Production by Population")
```



12. **Generate a scatterplot of coal production and population, colored by region, faceted on year.**

```
ggplot(data = tmp) + geom_point(aes(y = Production, x = Population, color = Division)) +
    labs(title = "Coal Production VS Population By Year") + facet_grid(Year ~
    .)
```


Coal Production VS Population By Year
Production
Population
Division
7.5
5.0
2.5
2010
2011
2012
2013
2014
8e+06
6e+06
4e+06
2e+06
0e+00
0e+00
1e+07
2e+07
3e+07
4e+07