eGRID Data Analysis

https://github.com/zhangxsaras/eGRID16

Xin Zhang

Abstract

Experimental overview. This section should be no longer than 250 words.

Contents

List of Tables

List of Figures

<Note: set up autoreferencing for figures and tables in your document>

1 Research Question and Rationale

2 Dataset Information

3 Exploratory Data Analysis and Wrangling

```
#Load dataset
GEN16 <- read.csv("./Data/Raw/egrid_GEN16.csv")
PLNT16 <- read.csv("./Data/Raw/egrid_PLNT16.csv")

#data wrangling
#This dataset has two names, one is the full name, and the first row is the abbreviati
#change column names to the abbr.
names(GEN16) <- lapply(GEN16[1, ], as.character)
names(PLNT16) <- lapply(PLNT16[1, ], as.character)

#filter out the first row and use abbr. as column names
GEN16 <- GEN16[2:26184,]
PLNT16 <- PLNT16[2:9710,]


#numeric data has comma, convert factor to numberic
class (GEN16$GENNTAN)
```

```
## [1] "factor"
```

```
GEN16$GENNTAN <-as.numeric(gsub(",", "", GEN16$GENNTAN))
PLNT16$PLNGENAN <-as.numeric(gsub(",", "", PLNT16$PLNGENAN))
PLNT16$PLCO2EQA <-as.numeric(gsub(",", "", PLNT16$PLCO2EQA))

#Year data as.Date
class (GEN16$GENYRONL)
```

```
## [1] "factor"
```

```
GEN16$GENYRONL<-as.Date(GEN16$GENYRONL,format = "%Y")

#filter data by the sequence of time
GEN16 = GEN16[order(GEN16[,'GENYRONL']),]

#GEN16 - sum the totoal generation by year
GEN16sel <- GEN16 %>%
  select(SEQGEN16, PSTATABB, GENNTAN, GENYRONL) %>%
  filter(!is.na(GENNTAN)) %>%
  filter(GENNTAN>0) %>%
  group_by(GENYRONL)%>%
  summarise(GENSUM = sum(GENNTAN))

#GEN16 - sum the totoal generation/CO2/plant numbers by state
PLNT16sel <- PLNT16 %>%
```

```r
  select(SEQPLT16, PSTATABB, PLPRMFL, PLNGENAN, PLCO2EQA) %>%
    filter(!is.na(PLNGENAN)&!is.na(PLCO2EQA)) %>%
  filter(PLNGENAN>0)%>%
  group_by(PSTATABB)%>%
  summarise(PLNTGEN = sum(PLNGENAN),
            ECO2 = sum(PLCO2EQA),
            Count=n())

#summary code for GEN16
colnames(GEN16sel)
```

```
## [1] "GENYRONL" "GENSUM"
```

```r
class(GEN16sel$GENSUM)
```

```
## [1] "numeric"
```

```r
class(GEN16sel$GENNTAN)
```

```
## Warning: Unknown or uninitialised column: 'GENNTAN'.
```

```
## [1] "NULL"
```

```r
summary(GEN16sel)
```

```
##     GENYRONL               GENSUM
##  Min.   :1891-04-15   Min.   :      876
##  1st Qu.:1925-10-14   1st Qu.:   794214
##  Median :1956-04-15   Median : 10767508
##  Mean   :1956-03-21   Mean   : 33200403
##  3rd Qu.:1986-10-14   3rd Qu.: 60662462
##  Max.   :2017-04-15   Max.   :213550203
```

```r
dim(GEN16sel)
```

```
## [1] 123   2
```

```r
head(GEN16sel)
```

```
## # A tibble: 6 x 2
##   GENYRONL   GENSUM
##   <date>      <dbl>
## 1 1891-04-15  24330
## 2 1893-04-15   1512
## 3 1896-04-15  21453
## 4 1898-04-15  25514
## 5 1899-04-15    876
## 6 1900-04-15  20899
```

```r
#summary code for PLNT16
colnames(PLNT16sel)
```

```
## [1] "PSTATABB" "PLNTGEN"  "ECO2"     "Count"
```

```r
class(PLNT16sel$PLNTGEN)
```

```
## [1] "numeric"
```

```r
class(PLNT16sel$ECO2)
```

```
## [1] "numeric"
```

```r
summary(PLNT16sel)
```

```
##     PSTATABB     PLNTGEN                ECO2                Count       
##  AK     : 1   Min.   :    76474   Min.   :    18470   Min.   :   2.0  
##  AL     : 1   1st Qu.: 34625868   1st Qu.: 11970062   1st Qu.:  61.0  
##  AR     : 1   Median : 60445059   Median : 31234830   Median :  96.0  
##  AZ     : 1   Mean   : 80006416   Mean   : 40119484   Mean   : 146.5  
##  CA     : 1   3rd Qu.:107747773   3rd Qu.: 54001623   3rd Qu.: 148.0  
##  CO     : 1   Max.   :453941341   Max.   :239363582   Max.   :1163.0  
##  (Other):45
```

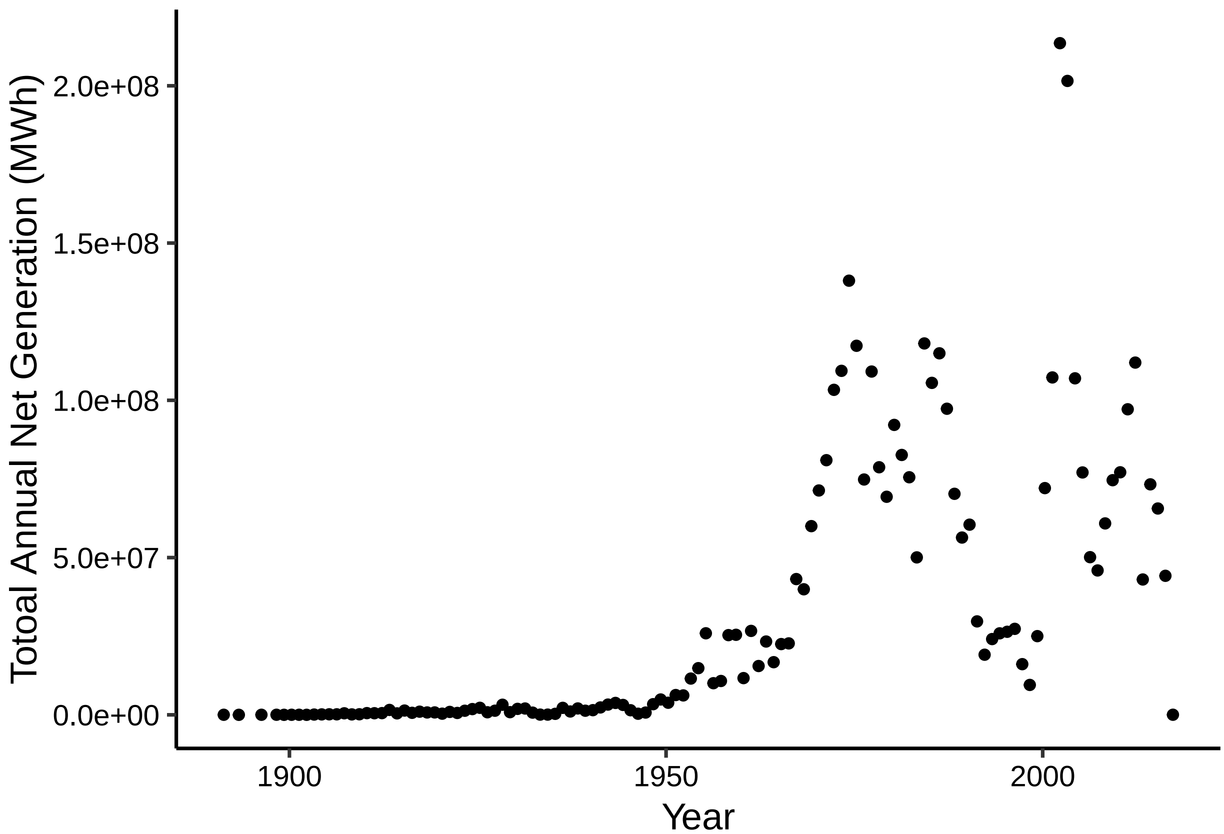
```r
dim(PLNT16sel)
```

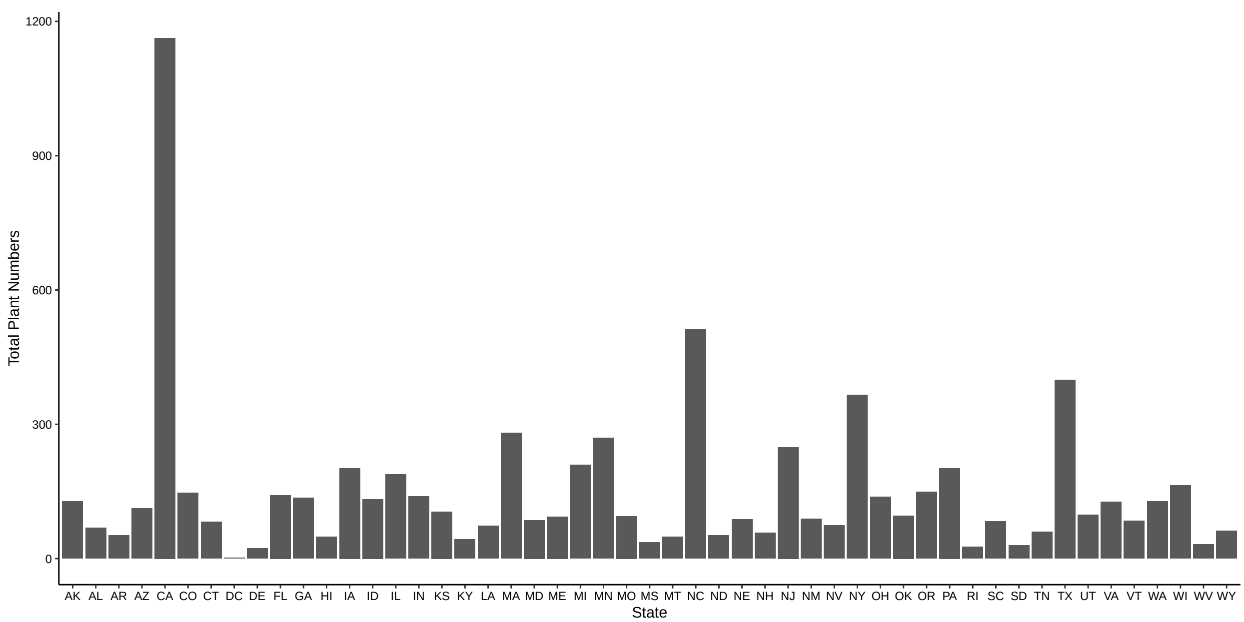
```
## [1] 51  4
```

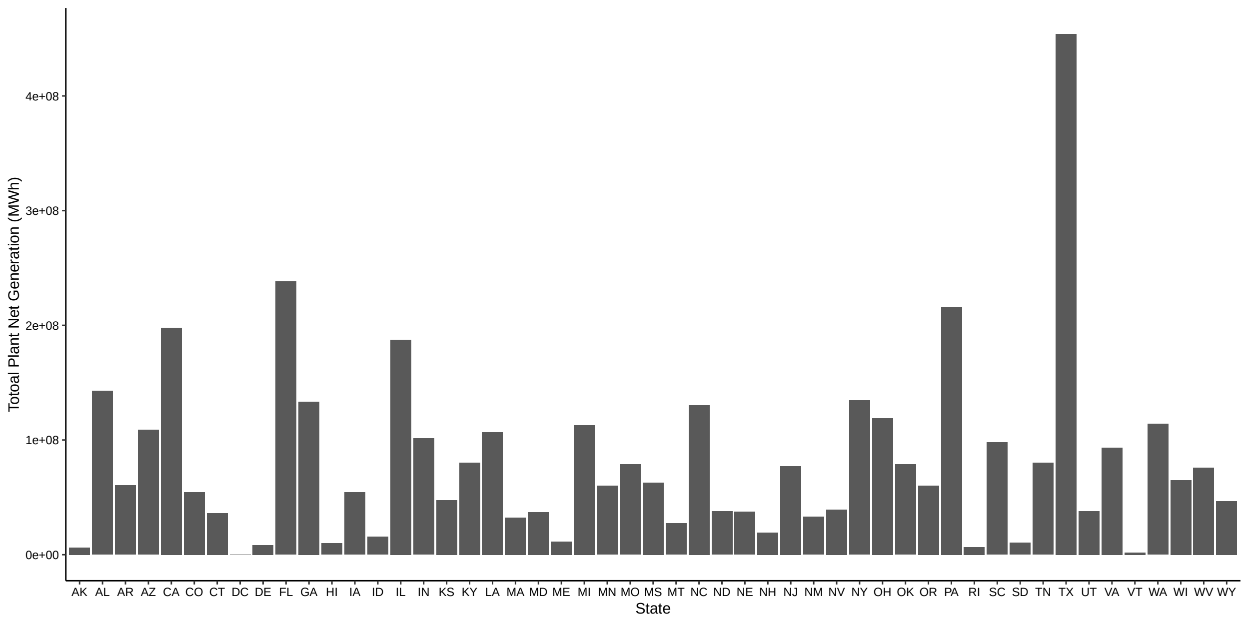
```r
head(PLNT16sel)
```

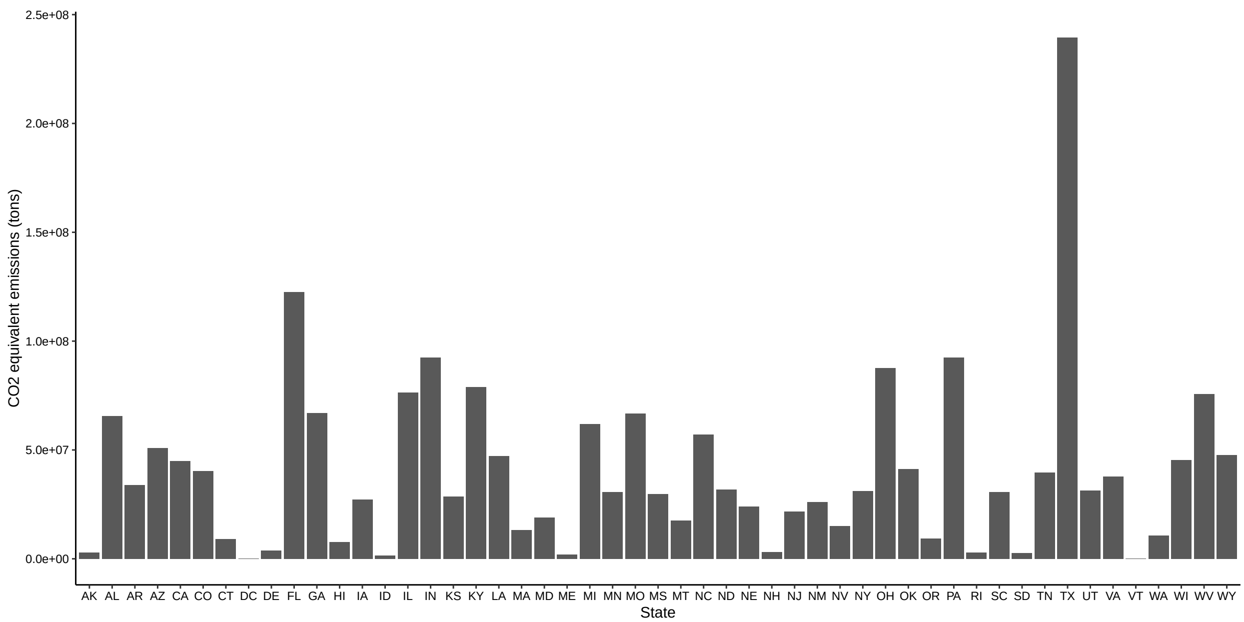
```
## # A tibble: 6 x 4
##   PSTATABB   PLNTGEN     ECO2 Count
##   <fct>        <dbl>    <dbl> <int>
## 1 AK         6339538  2944890   128
## 2 AL       142863565 65536193    69
## 3 AR        60445059 33927595    52
## 4 AZ       108734651 50946063   112
## 5 CA       197956373 44797489  1163
## 6 CO        54679959 40213412   147
```

```r
#save new datasets
#write.csv(GEN16sel, file = "./Data/Processed/GEN16sel_Processed.csv",row.names=FALSE)
#write.csv(PLNT16sel, file = "./Data/Processed/PLNT16sel_Processed.csv",row.names=FALS
#data prep for shiny app
#PLNT16orisel <- PLNT16 %>%
#  select(SEQPLT16, PSTATABB, PLPRMFL, PLNGENAN, PLCO2EQA) %>%
#    filter(!is.na(PLNGENAN)&!is.na(PLCO2EQA)) %>%
#  filter(PLNGENAN>0)
#write.csv(PLNT16orisel, file = "./Data/Processed/PLNT16ori_Processed.csv",row.names=F
```


Totoal Annual Net Generation (MWh)
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
1900
1950
2000
Year


Total Plant Numbers
0
300
600
900
1200
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
State


Totoal Plant Net Generation (MWh)
4e+08
3e+08
2e+08
1e+08
0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
State


CO2 equivalent emissions (tons)
2.5e+08
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0.0e+00
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
State

4 Analysis

```r
#Q1 Time series analysis on GEN16
# Use GLM to see if there is a significant time trend
GENTest.fixed <- gls(data = GEN16sel,
                     GENSUM ~ GENYRONL,
                      method = "REML")
summary(GENTest.fixed) # significnat trend t=10.23, p<0.005.
```

```
## Generalized least squares fit by REML
##   Model: GENSUM ~ GENYRONL
##   Data: GEN16sel
##        AIC      BIC    logLik
##   4562.782 4571.169 -2278.391
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 44750764 3129296.8 14.30058       0
## GENYRONL        2295     224.3 10.22992       0
##
##  Correlation:
##          (Intr)
## GENYRONL 0.361
##
## Standardized residuals:
##        Min          Q1         Med          Q3         Max
## -2.6061370 -0.5704619 -0.1403653  0.3843417  4.3789847
##
## Residual standard error: 32367813
## Degrees of freedom: 123 total; 121 residual
```

According to GLM, time has a significant effect on the total annual net generation of generators (t=10.23, $p<0.05$).

```r
# Run a Mann-Kendall test
mk.test(GEN16sel$GENSUM)
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM
## z = 11.311, n = 123, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 5.175000e+03 2.092503e+05 6.897241e-01
```

```r
# there is a trend over time according to this test (p<0.001), Z is positive, positive

# Test for change point
pettitt.test(GEN16sel$GENSUM) #changing point at 58 - Year 1952
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  GEN16sel$GENSUM
## U* = 3670, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                               58
```

```r
#GEN16sel[58,]

# Run separate Mann-Kendall for each change point
mk.test(GEN16sel$GENSUM[1:57])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[1:57]
## z = 6.8907, n = 57, p-value = 5.551e-12
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.002000e+03 2.110267e+04 6.278195e-01
```

```r
# there is a trend over time according to this test (p<0.001), Z is positive, positive
mk.test(GEN16sel$GENSUM[58:123])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[58:123]
## z = 2.8224, n = 66, p-value = 0.004767
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 5.110000e+02 3.265167e+04 2.382284e-01
```

```r
# there is a trend over time according to this test (p<0.05), Z is positive, positive

# Is there a second change point?
```

```r
pettitt.test(GEN16sel$GENSUM[58:123]) #there is! 17 - Year 1969
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  GEN16sel$GENSUM[58:123]
## U* = 681, p-value = 0.0001447
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              17
```

```r
#GEN16sel[75,]

# Run separate Mann-Kendall for each change point
mk.test(GEN16sel$GENSUM[58:74])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[58:74]
## z = 2.5951, n = 17, p-value = 0.009455
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##          S        varS         tau
##  64.0000000 589.3333333   0.4705882
```

```r
# there is a trend over time according to this test (p<0.05), Z is positive, positive
mk.test(GEN16sel$GENSUM[75:123])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[75:123]
## z = -2.0084, n = 49, p-value = 0.0446
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S           varS           tau
##  -234.0000000 13458.6666667    -0.1989796
```

```r
# there is a trend over time according to this test (p<0.05), Z is negative, negative

# Is there a third change point?
pettitt.test(GEN16sel$GENSUM[75:123]) #there is! 1987
```

```
##
##  Pettitt's test for single change-point detection
```

```
##
## data:  GEN16sel$GENSUM[75:123]
## U* = 322, p-value = 0.01123
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              19
```

```r
# GEN16sel[94,]
mk.test(GEN16sel$GENSUM[75:93])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[75:93]
## z = 0.69971, n = 19, p-value = 0.4841
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##          S       varS        tau
##  21.000000 817.000000   0.122807
```

```r
# no trend!
mk.test(GEN16sel$GENSUM[94:123])
```

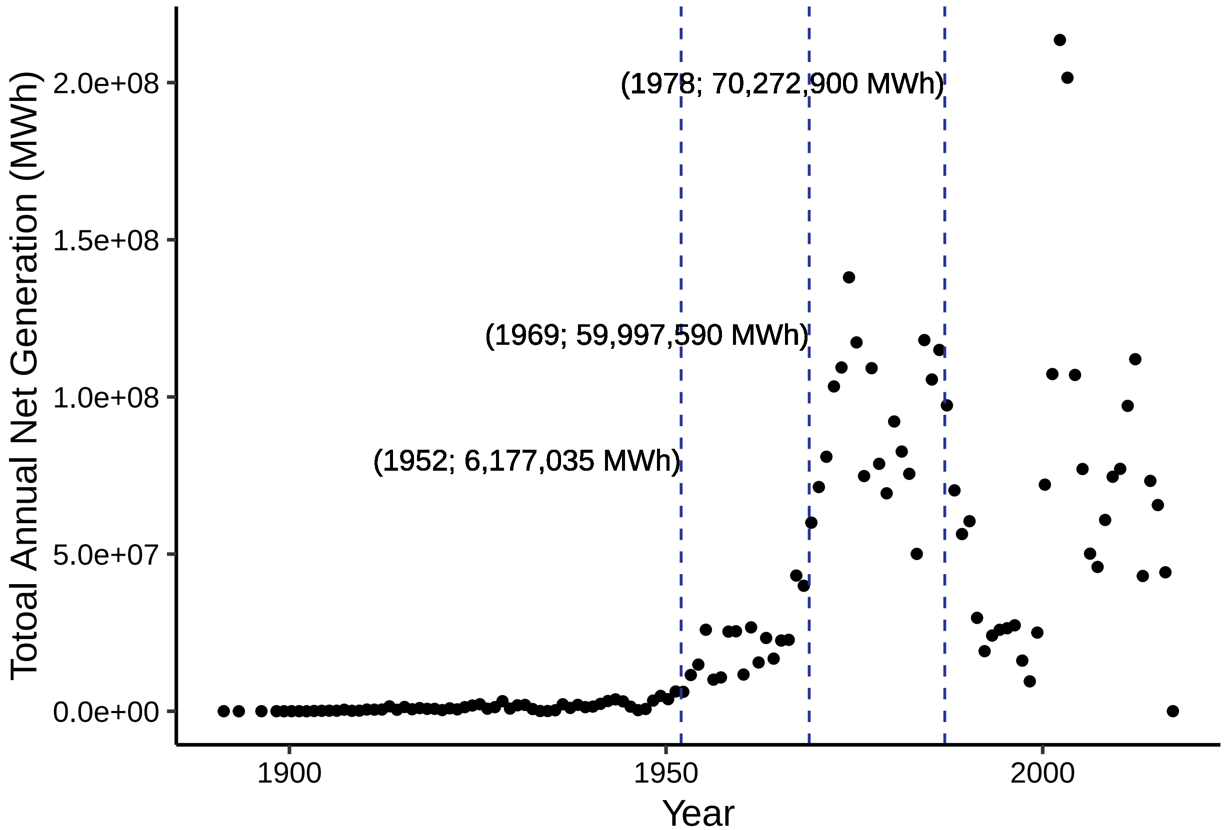
```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[94:123]
## z = 1.1775, n = 30, p-value = 0.239
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##           S        varS         tau
##   67.000000 3141.666667    0.154023
```

```r
# no trend!
# no trend before and after the changing point
```

According to the Mann-Kendall test, there is a trend over time for total annual net generation of generators ($p<0.001$), Z =11.311, which indicates a positive trend over time. Accroding to pettitt test. There are three changing points: Year 1952, Year 1969 and Year 1987 ($p<0.05$).

```r
ggplot(GEN16sel, aes(x = GENYRONL, y = GENSUM)) +
  geom_point() +
  labs(x= "Year", y="Totoal Annual Net Generation (MWh)")+
  geom_vline(xintercept = as.Date("1952-01-01"), color="#253494", lty = 2)+
  geom_vline(xintercept = as.Date("1969-01-01"), color="#253494", lty = 2)+
  geom_vline(xintercept = as.Date("1987-01-01"), color="#253494", lty = 2)+
```

```
geom_text(x = as.Date("1952-01-01"), y = 80000000, label = "(1952; 6,177,035 MWh)", hj
geom_text(x = as.Date("1969-01-01"), y = 120000000, label = "(1969; 59,997,590 MWh)",
geom_text(x = as.Date("1987-01-01"), y = 200000000, label = "(1978; 70,272,900 MWh)",
```



As is shown in the figure, before the first changing point 1952, the annual net generation of genrators in U.S. incrased very slowly each year and after 1952, the speed of generation change became faster. Annual net generation kept growing until 1969, this is the second change point. After 1969, there was a negative trend of annual net generation. The third changing point is Year 1987, and there was no clear pattern in 1969-1987 or after 1987.

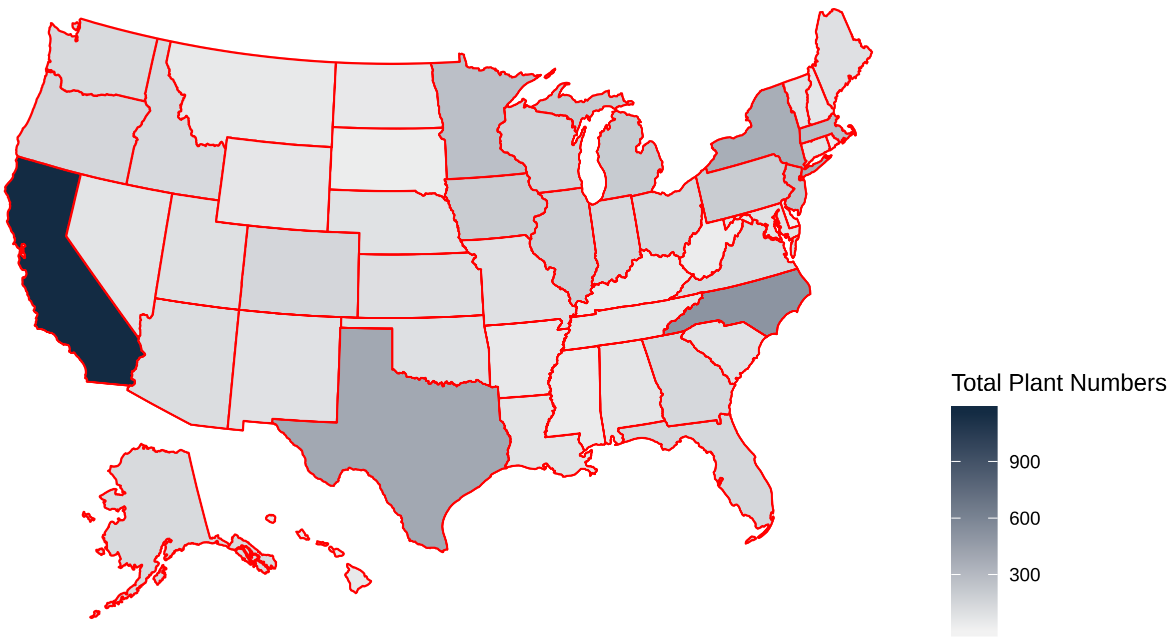
```
#Q2 spatial distribution analysis on PLNT16
#data wrangling
#add fips number to PLNT16sel data
library(usmap)
state_map <- us_map(regions = "states")
PLNT_State<- merge(state_map, PLNT16sel, by.x = "abbr", by.y = "PSTATABB")%>%
  select(abbr, fips, full, PLNTGEN, ECO2, Count)
PLNT_State = PLNT_State[!duplicated(PLNT_State$abbr),]

#join PLNT16 data to counties map data
states_sf<- st_read('./Data/RAW/States.shp')
```

```
## Reading layer `States' from data source `C:\Users\Xin Zhang\Desktop\eGRID16\Data\Raw\
## Simple feature collection with 52 features and 1 field
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -179.1743 ymin: 17.91377 xmax: 179.7739 ymax: 71.35256
## epsg (SRID):    4269
## proj4string:    +proj=longlat +datum=NAD83 +no_defs
```

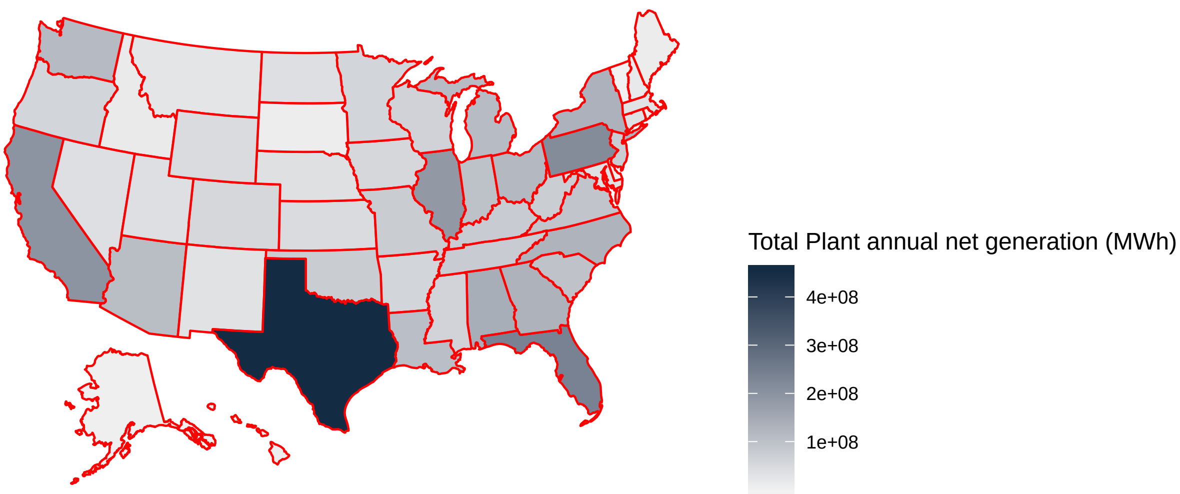
```r
st_crs(states_sf)
```

```
## Coordinate Reference System:
##   EPSG: 4269
##   proj4string: "+proj=longlat +datum=NAD83 +no_defs"
```

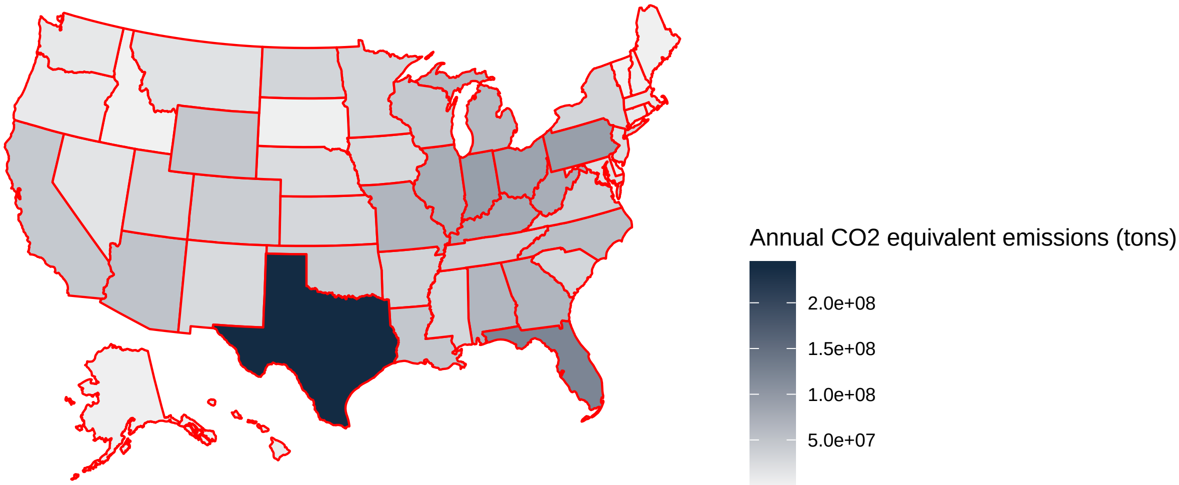
```r
PLNT_State_merge <- merge(states_sf, PLNT_State, by.x = "STATEFP", by.y = "fips")
#mapview(PLNT_State_merge)

#Plot plant numbers
#mapview(PLNT_State_merge['Count'], layer.name = "Total Plant Numbers")
plot_usmap(data = PLNT_State, values = "Count", lines = "red") +
    scale_fill_gradient("Total Plant Numbers",low='grey95',high='#132B43')+
  theme(legend.position = "right")
```



```r
#Plot PLNT Generation
#mapview(PLNT_State_merge['PLNTGEN'], layer.name = "Total Plant annual net generation
#ggplot() +
#  geom_sf(data = PLNT_State_merge, aes(fill=PLNTGEN))+
#  scale_fill_gradient("Total Plant annual net generation (MWh)",low='white',high='dar
#  theme(legend.position = "right")
plot_usmap(data = PLNT_State, values = "PLNTGEN", lines = "red") +
    scale_fill_gradient("Total Plant annual net generation (MWh)",low='grey95',high='#13
  theme(legend.position = "right")
```




```r
#Plot equivalent CO2
#mapview(PLNT_State_merge['ECO2'], layer.name = "Total Plant annual CO2 equivalent emi
plot_usmap(data = PLNT_State, values = "ECO2", lines = "red") +
    scale_fill_gradient("Annual CO2 equivalent emissions (tons)",low='grey95',high='#13
  theme(legend.position = "right")
```



Since our research scale is U.S., it also includes Alaska and this made it very difficult to map it using ggplot (AL will be far away from other sates and the whole map will be very small to show). Therefore, here I use the plot_usmap funciton instead.As the maps show, there are spatial distribution patterns. For total plant numbers, California ranks first but in general,there are more power plants in east cosat states than in west coast. Texas is another exception in the southern part that has relatively larger numbers of power plants. For plant annual net generation, in general, east coast states also have relatively higher net generation than west coast states, and Texas ranks first. Accordingly, Texas also has highest annual CO2 equivalent emissions, and east coat states have higher emissions than west coast states.

5 Summary and Conclusions

From the analysis above, we can see that time has a significant effect on the total annual net generation of generators. There is a trend over time for total annual net generation of generators that new power generators tend to have higher or net generation overtime.

There are also three changing points of the whole period: 1952, 1969, 1987. After 1952, there is a huge increase in power generation each year, this might be related to the technology thrive at the end of 1940s. The first hydraulic fracturing treatment was pumped on a gas well operated by Pan American Petroleum Corp in the Hugoton field in 1947 and it was the beginning of period of rapid electric industry growth (Oilscams.org, Access: 2019-4-14). However, at the end of 1960s, energy crisis caused the reduction of electricity generation and this energy crisis peaked at 1973 (Energy Crisis (1970s), Access: 2019-4-14). This accord with our second changing point here at 1969 and after 1969, there was a negative trend of annual net generation.

In the late 1970s, the government published more regulations on the electricity generation industry such as the National Energy Act in 1978 to exert more control on the electricity industry. In 1990s, the government published Congress passes Bush's Energy Policy Act (EPACT) to deregulate the electricity industry (Ballotpedia, Access: 2019-4-14). The back and force between strict and loose policies led to the fluctuations of the generator generations, and this might be a cause why there is a changing point of 1987 according to the pettitt test but there were no trend before and after this year according to the Mann-Kendall test.

From the maps, we can see that there was spatial distribution pattern for power plants generation in 2016. In general, eastern states had higher power plant numbers, electricity net generation, and equivalent CO2 emissions than western states. California ranked first for total power plants nubmers but its anual net generation and equivalent CO2 emission were both much less than Texas.The main reason for this was the type of power plants. In Texas, there are more traditonal power plants using coal and natural gas as fuel, but in California, there are more renewable energy power plants. Compared to traditional power plants (600 MW, 0.75), renewable power plants have lower nameplate capacity and capacity factor (1000MW, 0.1). Therefore, even though California had more power plants, it had lower electricity generation. Also, compared to traditional power plants, renwable power plants use clean energy and barly have any emissions, so California's equivalent CO2 emissions were also much lower than Texas.

To sum up, there is an installation time trend for generator annual net generation and there is a spatial distribution pattern of Power Plants in U.S..