# Energy, Price and GDP evaluation report for years 2010-2014

Research Question:

The question is - which are the top ranked states in the US which have high GDP and have high consumption of non-polluting renewable energy resources and also have lower average energy prices.

Relevance: The purpose is to show that states which rely on non polluting renewable energy resources can also have highGDP and lower energy prices.This analysis explores whether there is an economic case to be made for adopting non polluting renewable energy resources.

Setting my working data

```
setwd("C:/Users/Veeru/Documents/R_script_assignment2")
getwd()
[1] "C:/Users/Veeru/Documents/R_script_assignment2"
```

Uploading necessary packages

```
library(tidyverse)
library(ggplot2)
library(gridExtra)#for table creation
library(ggpubr)
```

Since the first line is header, setting header is equal to TRUE.

```
df <- read.table("EnergyCensus&EconomicDataUS2010-2014.csv", header = TRUE,
sep = ",")
```

Attributes give me all variable names, so that i can refer it in case of requirement

```
attributes(df)
$`names`
  [1] "StateCodes"            "State"
  [3] "Region"                "Division"
  [5] "Coast"                 "Great.Lakes"
  [7] "TotalC2010"            "TotalC2011"
```

```
 [9] "TotalC2012"            "TotalC2013"
[11] "TotalC2014"            "TotalP2010"
[13] "TotalP2011"            "TotalP2012"
[15] "TotalP2013"            "TotalP2014"
[17] "TotalE2010"            "TotalE2011"
[19] "TotalE2012"            "TotalE2013"
[21] "TotalE2014"            "TotalPrice2010"
[23] "TotalPrice2011"        "TotalPrice2012"
[25] "TotalPrice2013"        "TotalPrice2014"
[27] "TotalC10.11"           "TotalC11.12"
[29] "TotalC12.13"           "TotalC13.14"
[31] "TotalP10.11"           "TotalP11.12"
[33] "TotalP12.13"           "TotalP13.14"
[35] "TotalE10.11"           "TotalE11.12"
[37] "TotalE12.13"           "TotalE13.14"
[39] "TotalPrice10.11"       "TotalPrice11.12"
[41] "TotalPrice12.13"       "TotalPrice13.14"
[43] "BiomassC2010"          "BiomassC2011"
[45] "BiomassC2012"          "BiomassC2013"
[47] "BiomassC2014"          "CoalC2010"
[49] "CoalC2011"             "CoalC2012"
[51] "CoalC2013"             "CoalC2014"
[53] "CoalP2010"             "CoalP2011"
[55] "CoalP2012"             "CoalP2013"
[57] "CoalP2014"             "CoalE2010"
[59] "CoalE2011"             "CoalE2012"
[61] "CoalE2013"             "CoalE2014"
[63] "CoalPrice2010"         "CoalPrice2011"
[65] "CoalPrice2012"         "CoalPrice2013"
[67] "CoalPrice2014"         "ElecC2010"
[69] "ElecC2011"             "ElecC2012"
[71] "ElecC2013"             "ElecC2014"
[73] "ElecE2010"             "ElecE2011"
[75] "ElecE2012"             "ElecE2013"
```

```
 [77] "ElecE2014"              "ElecPrice2010"
 [79] "ElecPrice2011"          "ElecPrice2012"
 [81] "ElecPrice2013"          "ElecPrice2014"
 [83] "FossFuelC2010"          "FossFuelC2011"
 [85] "FossFuelC2012"          "FossFuelC2013"
 [87] "FossFuelC2014"          "GeoC2010"
 [89] "GeoC2011"               "GeoC2012"
 [91] "GeoC2013"               "GeoC2014"
 [93] "GeoP2010"               "GeoP2011"
 [95] "GeoP2012"               "GeoP2013"
 [97] "GeoP2014"               "HydroC2010"
 [99] "HydroC2011"             "HydroC2012"
[101] "HydroC2013"             "HydroC2014"
[103] "HydroP2010"             "HydroP2011"
[105] "HydroP2012"             "HydroP2013"
[107] "HydroP2014"             "NatGasC2010"
[109] "NatGasC2011"            "NatGasC2012"
[111] "NatGasC2013"            "NatGasC2014"
[113] "NatGasE2010"            "NatGasE2011"
[115] "NatGasE2012"            "NatGasE2013"
[117] "NatGasE2014"            "NatGasPrice2010"
[119] "NatGasPrice2011"        "NatGasPrice2012"
[121] "NatGasPrice2013"        "NatGasPrice2014"
[123] "LPGC2010"               "LPGC2011"
[125] "LPGC2012"               "LPGC2013"
[127] "LPGC2014"               "LPGE2010"
[129] "LPGE2011"               "LPGE2012"
[131] "LPGE2013"               "LPGE2014"
[133] "LPGPrice2010"           "LPGPrice2011"
[135] "LPGPrice2012"           "LPGPrice2013"
[137] "LPGPrice2014"           "GDP2010Q1"
[139] "GDP2010Q2"              "GDP2010Q3"
[141] "GDP2010Q4"              "GDP2010"
[143] "GDP2011Q1"              "GDP2011Q2"
```

```
[145] "GDP2011Q3"             "GDP2011Q4"
[147] "GDP2011"               "GDP2012Q1"
[149] "GDP2012Q2"             "GDP2012Q3"
[151] "GDP2012Q4"             "GDP2012"
[153] "GDP2013Q1"             "GDP2013Q2"
[155] "GDP2013Q3"             "GDP2013Q4"
[157] "GDP2013"               "GDP2014Q1"
[159] "GDP2014Q2"             "GDP2014Q3"
[161] "GDP2014Q4"             "GDP2014"
[163] "CENSUS2010POP"         "POPESTIMATE2010"
[165] "POPESTIMATE2011"       "POPESTIMATE2012"
[167] "POPESTIMATE2013"       "POPESTIMATE2014"
[169] "RBIRTH2011"            "RBIRTH2012"
[171] "RBIRTH2013"            "RBIRTH2014"
[173] "RDEATH2011"            "RDEATH2012"
[175] "RDEATH2013"            "RDEATH2014"
[177] "RNATURALINC2011"       "RNATURALINC2012"
[179] "RNATURALINC2013"       "RNATURALINC2014"
[181] "RINTERNATIONALMIG2011" "RINTERNATIONALMIG2012"
[183] "RINTERNATIONALMIG2013" "RINTERNATIONALMIG2014"
[185] "RDOMESTICMIG2011"      "RDOMESTICMIG2012"
[187] "RDOMESTICMIG2013"      "RDOMESTICMIG2014"
[189] "RNETMIG2011"           "RNETMIG2012"
[191] "RNETMIG2013"           "RNETMIG2014"

$class
[1] "data.frame"

$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
[24] 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
[47] 47 48 49 50 51 52
```

Glimpse/str gives the type of variables. The first two variables are categorical and rest all are continuous. The df has 52 observations and 192 variables. The subset of data frame i am choosing for my question is the average price variable, the GDP variables, the geothermal consumption and hydro power consumption variables for answering my question.

```
str(df)
```

```
'data.frame':   52 obs. of  192 variables:
 $ StateCodes          : Factor w/ 52 levels "AK","AL","AR",..: 2 1 4 3 5 6 7 9 10 11 ...
 $ State               : Factor w/ 52 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 10 11 ...
 $ Region              : int  3 4 4 3 4 4 1 3 3 3 ...
 $ Division            : int  6 9 8 7 9 8 1 5 5 5 ...
 $ Coast               : int  1 1 0 0 1 0 1 1 1 1 ...
 $ Great.Lakes         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TotalC2010          : int  1931522 653221 1383531 1120632 7760629 1513547 764970 250212 4282673 3100144 ...
```

checking for any NA’s in the df and we have 8 of them.

```
table(is.na(df))
```

```

FALSE  TRUE 
 9976     8 
```

```
na <- df%>%select(everything()) %>% summarise_all(funs(sum(is.na(.))))
na
```

| StateCodes | State | Region | Division | Coast | Great.Lakes |
|---:|---:|---:|---:|---:|---:|
| <int> | <int> | <int> | <int> | <int> | <int> |
| 0 | 0 | 1 | 1 | 1 | 1 |

1 row | 1-6 of 192 columns

The following code shows the row number and the position where NA’s are found.[rows,columns]

This means that 52nd row is incomplete, this row has the data for US as whole which i do not need since my focus is on states.

```
df[!complete.cases(df),]
```

| | StateCodes | State | Region | Division | Coast |
|---|---|---|---|---|---|
| | <fctr> | <fctr> | <int> | <int> | <int> |
| 52 | US | United States | *NA* | *NA* | *NA* |

1 row | 1-6 of 192 columns

Removing the row with NAs

The rows are reduced from 52 to 51, which is the actual number of states.

```
new_df <- na.omit(df)
nrow(new_df)
[1] 51
```

Creating data frame “GDP” which has GDP values for each state from 2010 to 2014 adding new variable “mean_gdp” which has mean calculated from the 5 years of GDP values for each state

```
gdp <- new_df %>% select(2,142,147,152,157,162) %>%
mutate(mean_gdp=(GDP2010+GDP2011+GDP2012+GDP2013+GDP2014)/5)
names(gdp)
[1] "State"    "GDP2010"  "GDP2011"  "GDP2012"  "GDP2013"  "GDP2014"
[7] "mean_gdp"
```

Since the plot consisting for all years GDP values is messy and doesnt help in answering the question, focusing only on the mean GDP values for creating plot

Arranging according to the states which have higher GDP values, California Ranks first with GDP $2130822.40 in million USD

```
gdpmean <- gdp %>% select(1,7) %>% arrange(desc(mean_gdp))
gdpmean
```

| State | mean_gdp |
|---|---:|
| <fctr> | <dbl> |
| California | 2130822.40 |
| Texas | 1434499.30 |
| New York | 1292133.65 |
| Florida | 772514.65 |
| Illinois | 701511.90 |
| Pennsylvania | 627958.05 |
| Ohio | 541790.95 |
| New Jersey | 516889.80 |
| North Carolina | 442848.80 |
| Virginia | 442823.00 |

1-10 of 51 rows

creating a subset of Average Price named “total_price”. This data frame has variables having Average Price from 2010 to 2014 for each state. The average price is in USD per million BTU. New Variable mean_price is created which has the mean values of all five years. Note: BTU - British Thermal Unit

```
total_price <- new_df %>% select(2,22:26) %>%

mutate(mean_price=(TotalPrice2010+TotalPrice2011+TotalPrice2012+TotalPrice201
3+

TotalPrice2014)/5)

names(total_price)

[1] "State"          "TotalPrice2010" "TotalPrice2011" "TotalPrice2012"
```

```
[5] "TotalPrice2013" "TotalPrice2014" "mean_price"
```

The price to be arranged from smaller to greater, since the part of my question is to find the states that pay least price for their energy. Louisiana stands to be on top, the average price for energy is $15.93. Hawaii pays highest that is $37

```
pricemean <- total_price %>% select(1,7) %>% arrange((mean_price))
pricemean
```

| **State** | **mean_price** |
|---|---:|
| <fctr> | <dbl> |
| Louisiana | 15.928 |
| North Dakota | 17.286 |
| Indiana | 17.296 |
| Iowa | 17.626 |
| Wyoming | 17.760 |
| Illinois | 18.766 |
| Nebraska | 18.900 |
| Arkansas | 18.992 |
| Texas | 19.028 |
| Alabama | 19.108 |

1-10 of 51 rows

Creating a subset of states consuming geo-thermal energy.New column called “geo_mean” is created, this column has the average value from the year 2010-2014. The states are then arranged according to the one consuming high geo-thermal energy.

```
geo_con <- new_df %>% select(2,88:92) %>%
mutate(geo_mean=(GeoC2010+GeoC2011+GeoC2012+GeoC2013+GeoC2014)/5)

names(geo_con)
```

```
[1] "State"    "GeoC2010" "GeoC2011" "GeoC2012" "GeoC2013" "GeoC2014"

[7] "geo_mean"
```

```
gmean <- geo_con %>% select(1,7) %>% arrange(desc(geo_mean))

gmean
```

| State | geo_mean |
|---|---:|
| <fctr> | <dbl> |
| California | 121424.8 |
| Nevada | 24491.4 |
| Florida | 9896.6 |
| Michigan | 5113.0 |
| Indiana | 4569.6 |
| Utah | 4217.8 |
| Ohio | 3383.6 |
| Kentucky | 2676.0 |
| Texas | 2447.6 |

| State | geo_mean |
|---|---:|
| <fctr> | <dbl> |
| Hawaii | 2338.0 |

1-10 of 51 rows

Creating a subset of states consuming hydro-power.New column called “hydro_mean” is created, this column has the average value from the year 2010-2014. The states are then arranged according to the one consuming higher hydro power.

```
hydro_con <- new_df %>% select(2,98:102) %>%
mutate(hydro_mean=(HydroC2010+HydroC2011+HydroC2012+HydroC2013+HydroC2014)/5)

names(hydro_con)
```

```
[1] "State"      "HydroC2010" "HydroC2011" "HydroC2012" "HydroC2013"

[6] "HydroC2014" "hydro_mean"
```

```
hmean <- hydro_con %>% select(1,7) %>% arrange(desc(hydro_mean))

hmean
```

| State | hydro_mean |
|---|---:|
| <fctr> | <dbl> |
| Washington | 782207.6 |
| Oregon | 347050.8 |
| California | 275776.0 |
| New York | 248294.0 |
| Montana | 104552.8 |
| Idaho | 98023.0 |

| State | hydro_mean |
|---|---:|
| <fctr> | <dbl> |
| Alabama | 91017.4 |
| Tennessee | 90947.4 |
| Arizona | 66455.2 |
| South Dakota | 52655.8 |

1-10 of 51 rows

The lollipop plot. This plot represents the top twenty states among the 51 states of USA who have higher GDP value.

```
gdp_plot <- ggplot(gdpmean[1:20,], aes(x=mean_gdp,
y=reorder(State,mean_gdp))) +
geom_point(size=5,color="#66CC33",alpha=0.7)+labs(x = "Average GDP(in million
USD)", y="States", title="Top 20 States with highest GDP from 2010-
2014")+theme_bw() + theme(axis.title.y=element_blank(),panel.border =
element_blank(), panel.grid.major = element_blank(),panel.grid.minor =
element_blank(), axis.line = element_blank(), axis.ticks = element_blank()) +
geom_segment(aes(y=State, x=0, yend=State,
xend=mean_gdp),color="#66CC33",alpha=0.3 )
```

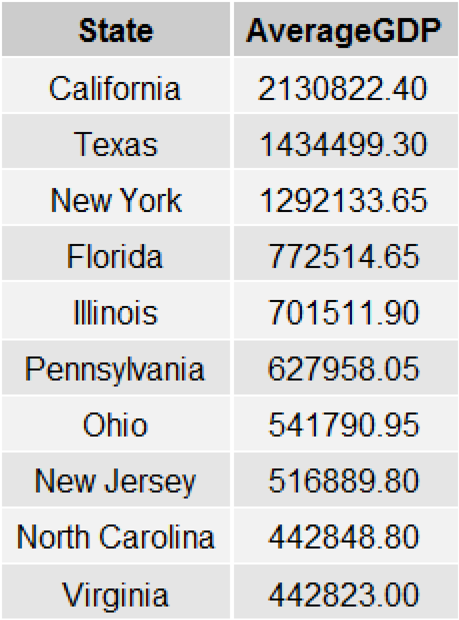
Data for table creation

```
gdp_table <- cbind(State=c("California","Texas","New York",
"Florida","Illinois","Pennsylvania","Ohio","New Jersey","North
Carolina","Virginia"),

AverageGDP=c("2130822.40","1434499.30","1292133.65","772514.65","701511.90","
627958.05","541790.95","516889.80","442848.80","442823.00"))

gdp_table
```

```
      State            AverageGDP

 [1,] "California"     "2130822.40"

 [2,] "Texas"          "1434499.30"

 [3,] "New York"       "1292133.65"

 [4,] "Florida"        "772514.65"
 [5,] "Illinois"       "701511.90"
 [6,] "Pennsylvania"   "627958.05"
 [7,] "Ohio"           "541790.95"
 [8,] "New Jersey"     "516889.80"
 [9,] "North Carolina" "442848.80"
[10,] "Virginia"       "442823.00"
```

The table gives the information of top ten states which are having higher Average GDP in five years, 2010 to 2014

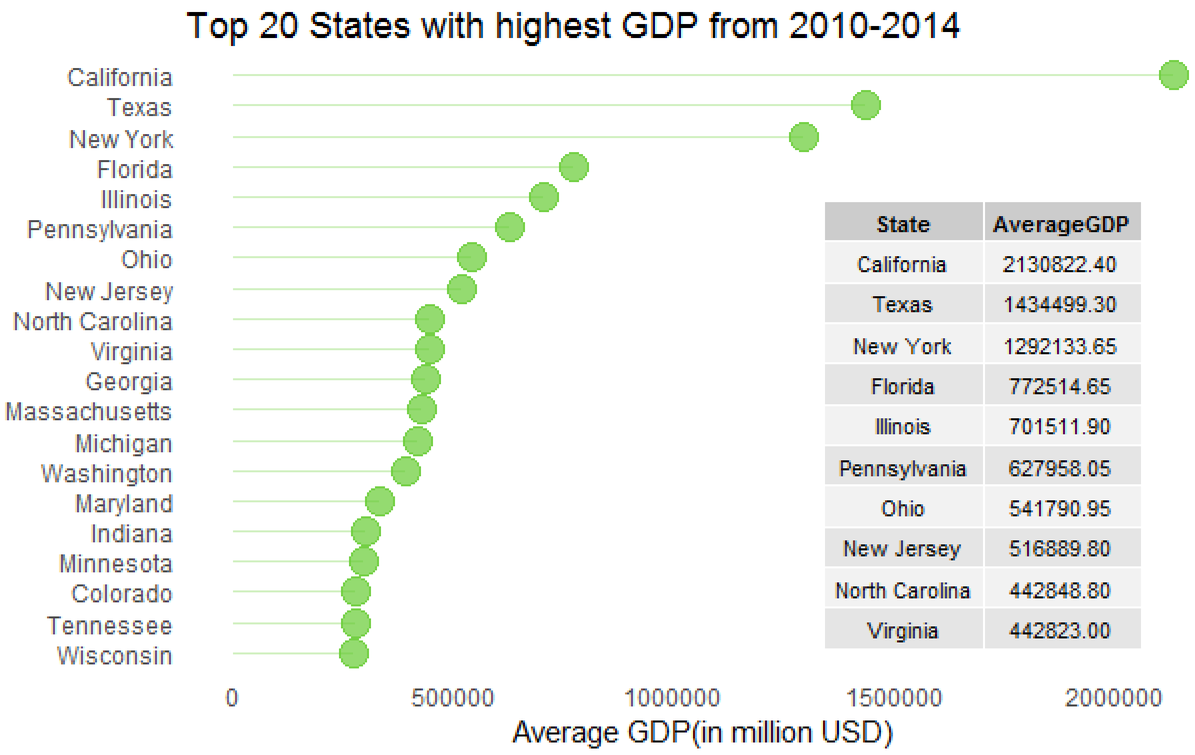
```
qplot(1:10, 1:10, geom = "blank") + theme_bw() + theme_void() +
annotation_custom(grob = tableGrob(gdp_table))
```



| State | AverageGDP |
|---|---|
| California | 2130822.40 |
| Texas | 1434499.30 |
| New York | 1292133.65 |
| Florida | 772514.65 |
| Illinois | 701511.90 |
| Pennsylvania | 627958.05 |
| Ohio | 541790.95 |
| New Jersey | 516889.80 |
| North Carolina | 442848.80 |
| Virginia | 442823.00 |

To adjust the proportions of width to height (padding) and the size of the text and cells (base_size).

```
gdp_plot+annotation_custom(tableGrob(gdp_table, theme =
ttheme_default(base_size = 8, padding = unit(c(3,3),"mm"))), xmin=3000000,
xmax=400000, ymin=2, ymax=15)
```



| State | AverageGDP |
|---|---|
| California | 2130822.40 |
| Texas | 1434499.30 |
| New York | 1292133.65 |
| Florida | 772514.65 |
| Illinois | 701511.90 |
| Pennsylvania | 627958.05 |
| Ohio | 541790.95 |
| New Jersey | 516889.80 |
| North Carolina | 442848.80 |
| Virginia | 442823.00 |

```
#ggsave(file="gdp.png",height=6, dpi=300) used for saving the plot
```

The Lollipop plot for Lowest Average price.

Limitation 1:

I found geom_segment difficult to plot when using with xlim(). The points originally where plotting at far left, so i used xlim() so that i could use space to plot table. But geom_segment() does not plot well when using xlim(). I had to specially use scale_x_continuous() function to expand and set limits manually for geom_segment() to work.

```
price_plot <- ggplot(pricemean[1:20,], aes(x=mean_price,
y=reorder(State,mean_price))) +
geom_point(size=5,color="#FFCC00",alpha=0.7)+labs(x = "Average Price(in
USD/million BTU)", y="States", title="Top 20 states with low energy prices
from 2010-2014")+theme_bw() + theme(panel.border = element_blank(),
panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
axis.line = element_blank(), axis.ticks =
element_blank(),axis.title.y=element_blank()) +geom_segment(aes(y=State,
x=13, yend=State, xend=mean_price),color="#FFCC00",alpha=0.3
)+scale_x_continuous(expand = c(13, 30))+scale_x_continuous(limits = c(13,
30))
```

```
Scale for 'x' is already present. Adding another scale for 'x', which

will replace the existing scale.
```

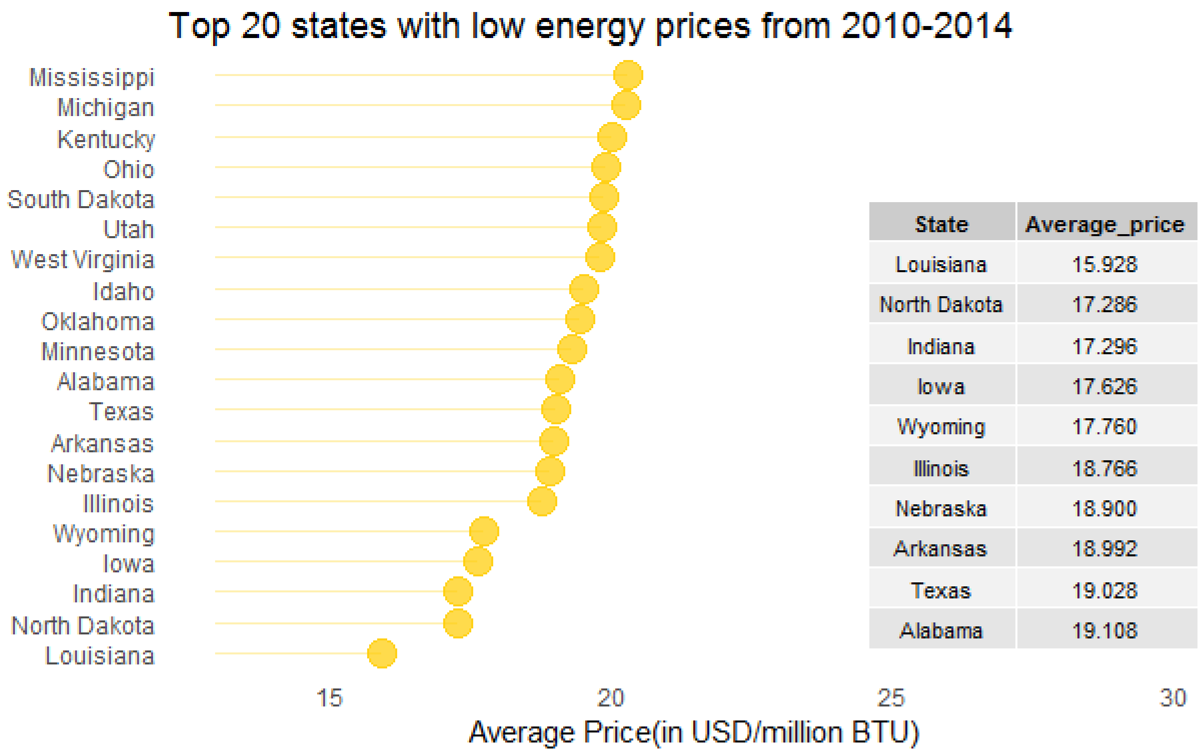
The Table data for Average Price table

```
price_table <- cbind(State=c("Louisiana","North Dakota","Indiana",
"Iowa","Wyoming","Illinois","Nebraska","Arkansas","Texas","Alabama"),

Average_price=c("15.928","17.286","17.296","17.626","17.760","18.766","18.900
","18.992","19.028","19.108"))

price_table
```

```
      State          Average_price
 [1,] "Louisiana"    "15.928"
 [2,] "North Dakota" "17.286"
 [3,] "Indiana"      "17.296"
 [4,] "Iowa"         "17.626"
 [5,] "Wyoming"      "17.760"
 [6,] "Illinois"     "18.766"
 [7,] "Nebraska"     "18.900"
 [8,] "Arkansas"     "18.992"
 [9,] "Texas"        "19.028"
[10,] "Alabama"      "19.108"
```

To adjust the proportions of width to height (padding) and the size of the text and cells (base_size).

```
price_plot+annotation_custom(tableGrob(price_table, theme =
ttheme_default(base_size = 8, padding = unit(c(3,3),"mm"))), xmin=25,
xmax=30, ymin=2, ymax=15)
```



| State | Average_price |
|---|---|
| Louisiana | 15.928 |
| North Dakota | 17.286 |
| Indiana | 17.296 |
| Iowa | 17.626 |
| Wyoming | 17.760 |
| Illinois | 18.766 |
| Nebraska | 18.900 |
| Arkansas | 18.992 |
| Texas | 19.028 |
| Alabama | 19.108 |

```
#ggsave(file="price.png",height=6, dpi=300)
```

Top ten states with high geo thermal energy consumption in billion BTU is combined with top ten states with high hydro power consumption in billion BTU(British thermal units) and is then saved in data frame “green_energy”

```
geo<-gmean[1:10,]
hydro <- hmean[1:10,]
geo <- mutate(geo, EnergyResource="Geo-Thermal")
hydro <- mutate(hydro, EnergyResource="Hydro-Power")
names(geo)[2] <- "mean"
names(hydro)[2] <- "mean"
green_energy <- rbind(geo,hydro)
green_energy
```

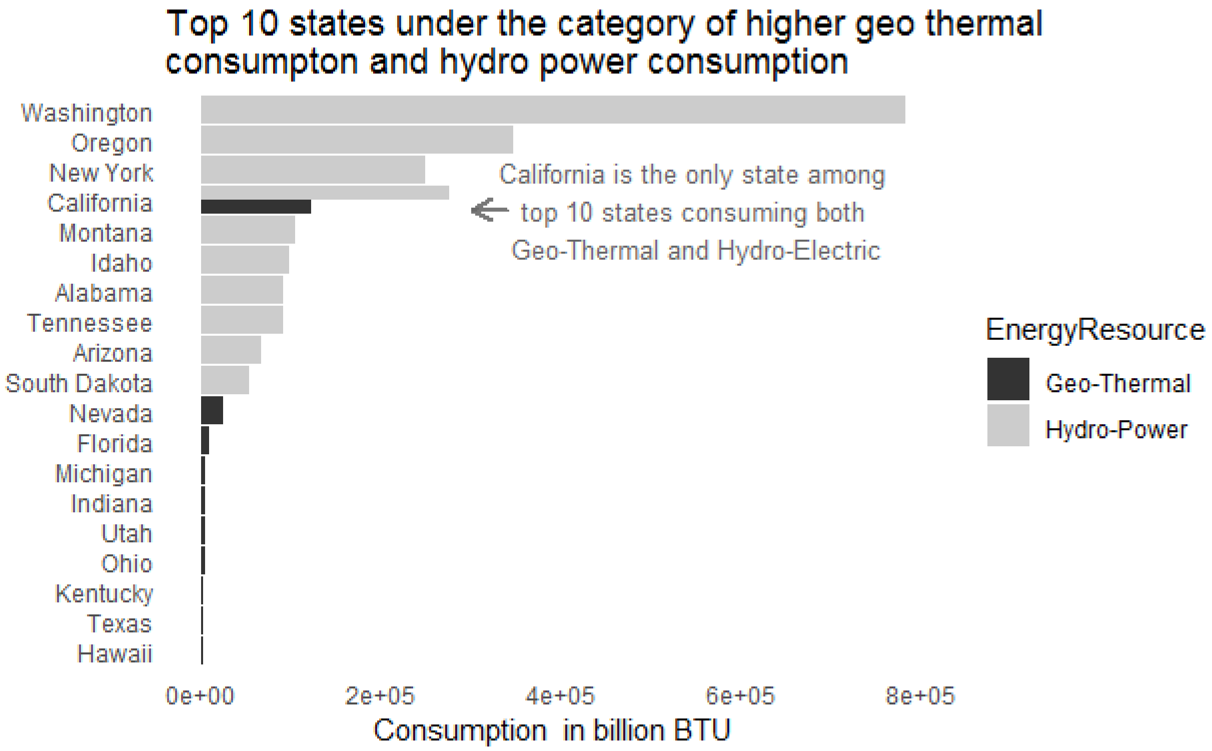
| State | mean | EnergyResource |
|---|---:|---|
| <fctr> | <dbl> | <chr> |
| California | 121424.8 | Geo-Thermal |
| Nevada | 24491.4 | Geo-Thermal |
| Florida | 9896.6 | Geo-Thermal |
| Michigan | 5113.0 | Geo-Thermal |
| Indiana | 4569.6 | Geo-Thermal |
| Utah | 4217.8 | Geo-Thermal |
| Ohio | 3383.6 | Geo-Thermal |
| Kentucky | 2676.0 | Geo-Thermal |
| Texas | 2447.6 | Geo-Thermal |
| Hawaii | 2338.0 | Geo-Thermal |

1-10 of 20 rows

The data frame green_energy is then used to plot grouped bar graph showing top 10 states consuming Geothermal energy and top 10 states consuming Hydro power.

```
ggplot(data=green_energy,
aes(x=reorder(State,mean),y=mean,fill=EnergyResource)) +
geom_col(position="dodge",stat="identity") + coord_flip() +ggtitle("Top 10
states under the category of higher geo thermal \nconsumpton and hydro power
consumption")+scale_fill_grey()+labs(y = "Consumption  in billion
BTU")+theme_bw() + theme(panel.border = element_blank(), panel.grid.major =
element_blank(), panel.grid.minor = element_blank(), axis.line =
element_blank(), axis.ticks =
element_blank(),axis.title.y=element_blank())+geom_segment(aes(x = 15.7, y =
340000, xend = 15.7, yend = 300000),size=1,lineend =
"round",alpha=0.1,color="#666666", linejoin = "round",arrow = arrow(length =
unit(0.3, "cm")))+ annotate("text",color="#666666", x = 15.7, y =
550000,size=3.5,label = "California is the only state among \ntop 10 states
consuming both \nGeo-Thermal and Hydro-Electric")
```

```
Ignoring unknown parameters: stat
```



Top 10 states under the category of higher geo thermal consumpton and hydro power consumption

California is the only state among top 10 states consuming both Geo-Thermal and Hydro-Electric

Combining dataframe ‘pricemean’ which has the states arranged with respect to lower average price on Energy and ‘gmean which has the states sorted according to higher geo thermal consumption and ’hmean’which has states sorted as per higher hydro power consumption and ’gdpmean’ which has states ordered as per high GDP value. All these dataframe combined to data frame ‘total’.

The total data frame is then filtered to align the states which have greater hydro and geo thermal consumption, Less average price in energy and greater GDP

```
total1<- left_join(pricemean,gmean, by ='State')

total2 <- left_join(total1,hmean, by = 'State')

total <- left_join(total2,gdpmean, by = 'State')

total
```

| State | mean_price | geo_mean | hydro_mean |
|---|---:|---:|---:|
| <fctr> | <dbl> | <dbl> | <dbl> |
| Louisiana | 15.928 | 1825.0 | 9552.6 |
| North Dakota | 17.286 | 968.2 | 22062.2 |
| Indiana | 17.296 | 4569.6 | 3949.0 |

| State | mean_price | geo_mean | hydro_mean |
|---|---:|---:|---:|
| <fctr> | <dbl> | <dbl> | <dbl> |
| Iowa | 17.626 | 1274.4 | 8207.0 |
| Wyoming | 17.760 | 651.4 | 9084.8 |
| Illinois | 18.766 | 1999.2 | 1197.2 |
| Nebraska | 18.900 | 1196.4 | 12445.2 |
| Arkansas | 18.992 | 789.4 | 27157.8 |
| Texas | 19.028 | 2447.6 | 6318.6 |
| Alabama | 19.108 | 139.2 | 91017.4 |

1-10 of 51 rows | 1-4 of 5 columns

The Highest Average price is of Hawaii which is $36.978/million BTU of energy.

Connecticut has lowest Geo Thermal consumption which is 20.4 billions BTU.

Disctrict of Columbia, Delaware and Missisippi had no hydro electric consumption.

Vermount having least GDP of $28077.90 in million USD

Using filter, sorting the data frame such that Average price is sorted in the order less than hawaii, geo thermal consumption gretaer than connecticut's, Hydro power consumption greater than 0 and GDP greater than Vermount's

The intention behind this to get a dataframe which gets the states aligned to having higher geo thermal and hydro power consumption, having higher GDP and also having less average price of Energy.

```
filter(total, mean_price<=36.978, mean_gdp>=28077.90, geo_mean>=20.4,
hydro_mean>=0.0)
```

| State | mean_price | geo_mean | hydro_mean |
|---|---:|---:|---:|
| <fctr> | <dbl> | <dbl> | <dbl> |
| Louisiana | 15.928 | 1825.0 | 9552.6 |
| North Dakota | 17.286 | 968.2 | 22062.2 |
| Indiana | 17.296 | 4569.6 | 3949.0 |
| Iowa | 17.626 | 1274.4 | 8207.0 |
| Wyoming | 17.760 | 651.4 | 9084.8 |
| Illinois | 18.766 | 1999.2 | 1197.2 |
| Nebraska | 18.900 | 1196.4 | 12445.2 |

| State | mean_price | geo_mean | hydro_mean |
|---|---:|---:|---:|
| <fctr> | <dbl> | <dbl> | <dbl> |
| Arkansas | 18.992 | 789.4 | 27157.8 |
| Texas | 19.028 | 2447.6 | 6318.6 |
| Alabama | 19.108 | 139.2 | 91017.4 |

1-10 of 51 rows | 1-4 of 5 columns

### Limitation2:

Unfortunately, Filter function only aligns the states with respect to first variable and other three as tie breakers. For instance Louisiana has less Average price, but it has less geo thermal/hydro power consumption and less GDP and still tops the list and Texas has higher GDP but less hydro consumption than louisiana and greater geo thermal consumption than louisiana and still in ninth place.

I want the states to be sorted which are majority of top ranks satisfying the conditions to answer my question.

Using arrange function to check if my question can be answered and it produces the same result as above.

```
arrange(total, mean_price, desc(geo_mean),desc(hydro_mean),desc(mean_gdp))
```

| State | mean_price | geo_mean | hydro_mean |
|---|---:|---:|---:|
| <fctr> | <dbl> | <dbl> | <dbl> |
| Louisiana | 15.928 | 1825.0 | 9552.6 |
| North Dakota | 17.286 | 968.2 | 22062.2 |
| Indiana | 17.296 | 4569.6 | 3949.0 |
| Iowa | 17.626 | 1274.4 | 8207.0 |
| Wyoming | 17.760 | 651.4 | 9084.8 |
| Illinois | 18.766 | 1999.2 | 1197.2 |
| Nebraska | 18.900 | 1196.4 | 12445.2 |
| Arkansas | 18.992 | 789.4 | 27157.8 |
| Texas | 19.028 | 2447.6 | 6318.6 |
| Alabama | 19.108 | 139.2 | 91017.4 |

1-10 of 51 rows | 1-4 of 5 columns

Unsuccessful by above attempts, I will rank states in each four Data frame of high GDP,geo thermal/hydro power consumption and least average price. At the end i will add all the ranks and arrange the state which has least sum value. This list is the answer for my question.

Ranking the states as per the higher GDP values

```
gdp_rank <- gdpmean %>% select(1) %>% mutate(gdprank = 1:n())
gdp_rank
```

| State<br><fctr> | gdprank<br><int> |
|---|---:|
| California | 1 |
| Texas | 2 |
| New York | 3 |
| Florida | 4 |
| Illinois | 5 |
| Pennsylvania | 6 |
| Ohio | 7 |
| New Jersey | 8 |
| North Carolina | 9 |
| Virginia | 10 |

1-10 of 51 rows

Ranking the states having less Average price on their Energy

```
price_rank <- pricemean %>% select(1) %>% mutate(pricerank = 1:n())
price_rank
```

| State<br><fctr> | pricerank<br><int> |
|---|---:|
| Louisiana | 1 |
| North Dakota | 2 |
| Indiana | 3 |
| Iowa | 4 |

| State | pricerank |
|:---|---:|
| <fctr> | <int> |
| Wyoming | 5 |
| Illinois | 6 |
| Nebraska | 7 |
| Arkansas | 8 |
| Texas | 9 |
| Alabama | 10 |

1-10 of 51 rows

Ranking the states consuming more Geo thermal energy

```
geo_rank <- gmean %>% select(1) %>% mutate(georank = 1:n())
geo_rank
```

| State | georank |
|:---|---:|
| <fctr> | <int> |
| California | 1 |
| Nevada | 2 |
| Florida | 3 |
| Michigan | 4 |
| Indiana | 5 |
| Utah | 6 |
| Ohio | 7 |
| Kentucky | 8 |
| Texas | 9 |
| Hawaii | 10 |

1-10 of 51 rows

Ranking the state consuming more hydro power

```
hydro_rank <- hmean %>% select(1) %>% mutate(hydrorank = 1:n())
hydro_rank
```

| State | hydrorank |
|---|---:|
| <fctr> | <int> |
| Washington | 1 |
| Oregon | 2 |
| California | 3 |
| New York | 4 |
| Montana | 5 |
| Idaho | 6 |
| Alabama | 7 |
| Tennessee | 8 |
| Arizona | 9 |
| South Dakota | 10 |

1-10 of 51 rows

joining all the above dataframe and adding their rank to find the states which answer the question

```
rank1<- left_join(price_rank,geo_rank, by ='State')
rank2 <- left_join(rank1,hydro_rank, by = 'State')
rank3 <- left_join(rank2,gdp_rank, by = 'State')
rank4 <- mutate(rank3, states_rank=(pricerank+georank+hydrorank+gdprank))
ranking <- arrange(rank4, states_rank)
ranking
```

| State | pricerank | georank | hydrorank | gdprank |
|---|---:|---:|---:|---:|
| <fctr> | <int> | <int> | <int> | <int> |
| California | 39 | 1 | 3 | 1 |
| Texas | 9 | 9 | 37 | 2 |
| Washington | 25 | 21 | 1 | 14 |
| Michigan | 19 | 4 | 27 | 13 |
| Pennsylvania | 31 | 11 | 16 | 6 |
| Indiana | 3 | 5 | 41 | 16 |
| Kentucky | 18 | 8 | 14 | 27 |

| State | pricerank | georank | hydrorank | gdprank |
|---|---:|---:|---:|---:|
| <fctr> | <int> | <int> | <int> | <int> |
| Illinois | 6 | 13 | 44 | 5 |
| New York | 42 | 20 | 4 | 3 |
| Ohio | 17 | 7 | 39 | 7 |

1-10 of 51 rows | 1-5 of 6 columns

The Answer to my question

```
state_rank <- ranking %>% select(1) %>% mutate(rank = 1:n())
state_rank
```

| State | rank |
|---|---:|
| <fctr> | <int> |
| California | 1 |
| Texas | 2 |
| Washington | 3 |
| Michigan | 4 |
| Pennsylvania | 5 |
| Indiana | 6 |
| Kentucky | 7 |
| Illinois | 8 |
| New York | 9 |
| Ohio | 10 |

1-10 of 51 rows

Using map_data() function to get state Data frame, the state names are mentioned under region variable. I will have to change the ‘State’ variable name to region so that i can merge my dataframe ‘state_rank’ with that of ggplot’s ‘state’ dataframe.

```
state0 <- map_data("state")
```

```
Attaching package: <U+393C><U+3E31>maps<U+393C><U+3E32>
```

```
The following object is masked from
<U+393C><U+3E31>package:purrr<U+393C><U+3E32>:


    map
```

```
str(state0)
'data.frame':   15537 obs. of  6 variables:
 $ long     : num  -87.5 -87.5 -87.5 -87.5 -87.6 ...
 $ lat      : num  30.4 30.4 30.4 30.3 30.3 ...
 $ group    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ order    : int  1 2 3 4 5 6 7 8 9 10 ...
 $ region   : chr  "alabama" "alabama" "alabama" "alabama" ...
 $ subregion: chr  NA NA NA NA ...
```

Changing the ‘state_rank’ data frame’s variable name of ‘States’ to ‘region’.Also, the state names in ggplots ‘state’ data frame is all lower case and ‘state_rank’ data frame’s state names are in caps. Therefore converting them to lower case for merging purpose.Additionally, ‘state_rank’ data frame has state variable of factor class, coverting it to character class as it is in ggplot’s ‘state’ data frame for ‘region’ variable.

```
names(state_rank)[1]<-"region"
state_rank$region <- as.character(state_rank$region)
state_rank$region <- tolower(state_rank$region)
str(state_rank)
'data.frame':   51 obs. of  2 variables:
 $ region: chr  "california" "texas" "washington" "michigan" ...
 $ rank  : int  1 2 3 4 5 6 7 8 9 10 ...
```

merging the data frame ‘state_rank’ and ‘state0’

```
map <- merge(state0,state_rank)
map
```

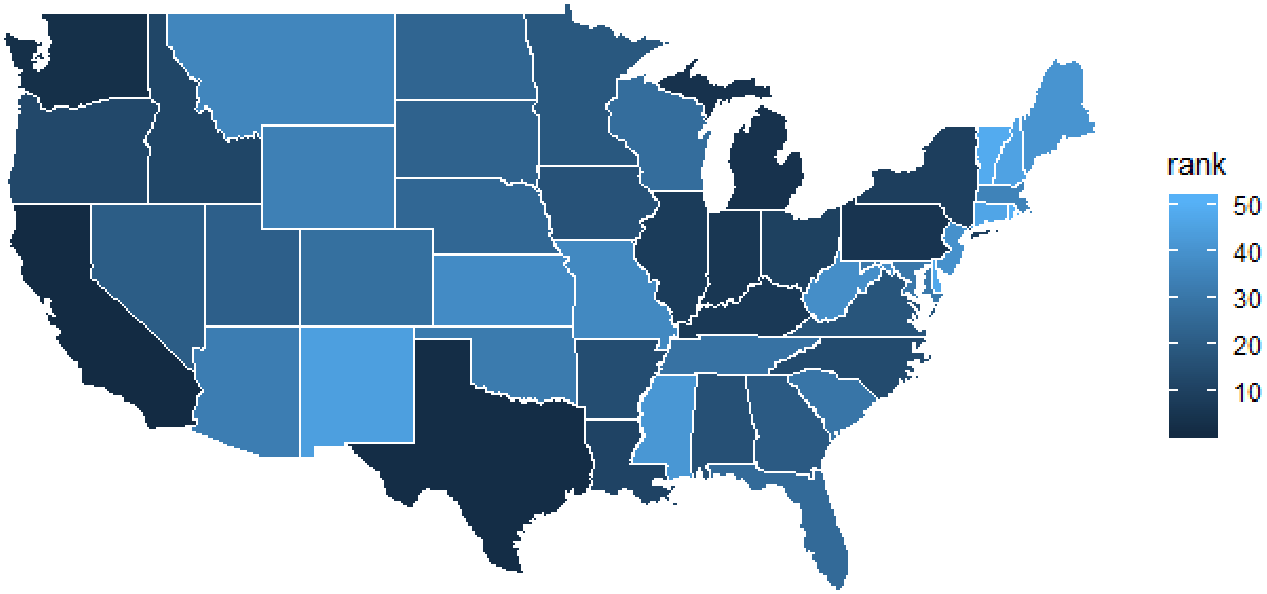
| region | long | lat | group | order | subregion | rank |
|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <int> | <chr> | <int> |

| region | long | lat | group | order | subregion | rank |
|---|---|---|---|---|---|---|
| <chr> | <dbl> | <dbl> | <dbl> | <int> | <chr> | <int> |
| alabama | -87.46201 | 30.38968 | 1 | 1 | *NA* | 16 |
| alabama | -87.48493 | 30.37249 | 1 | 2 | *NA* | 16 |
| alabama | -87.52503 | 30.37249 | 1 | 3 | *NA* | 16 |
| alabama | -87.53076 | 30.33239 | 1 | 4 | *NA* | 16 |
| alabama | -87.57087 | 30.32665 | 1 | 5 | *NA* | 16 |
| alabama | -87.58806 | 30.32665 | 1 | 6 | *NA* | 16 |
| alabama | -87.59379 | 30.30947 | 1 | 7 | *NA* | 16 |
| alabama | -87.59379 | 30.28655 | 1 | 8 | *NA* | 16 |
| alabama | -87.67400 | 30.27509 | 1 | 9 | *NA* | 16 |
| alabama | -87.81152 | 30.25790 | 1 | 10 | *NA* | 16 |

1-10 of 15,537 rows

Plotting the polygon map, with darker shade assigned to higher rank and lighter to lower.(theme_map() did not work and to use theme_void() instead)

```
statemap <- ggplot(map, aes(x=long, y= lat, fill=rank,
group=group))+geom_polygon(col="white") + coord_map()+theme_void()

statemap
```



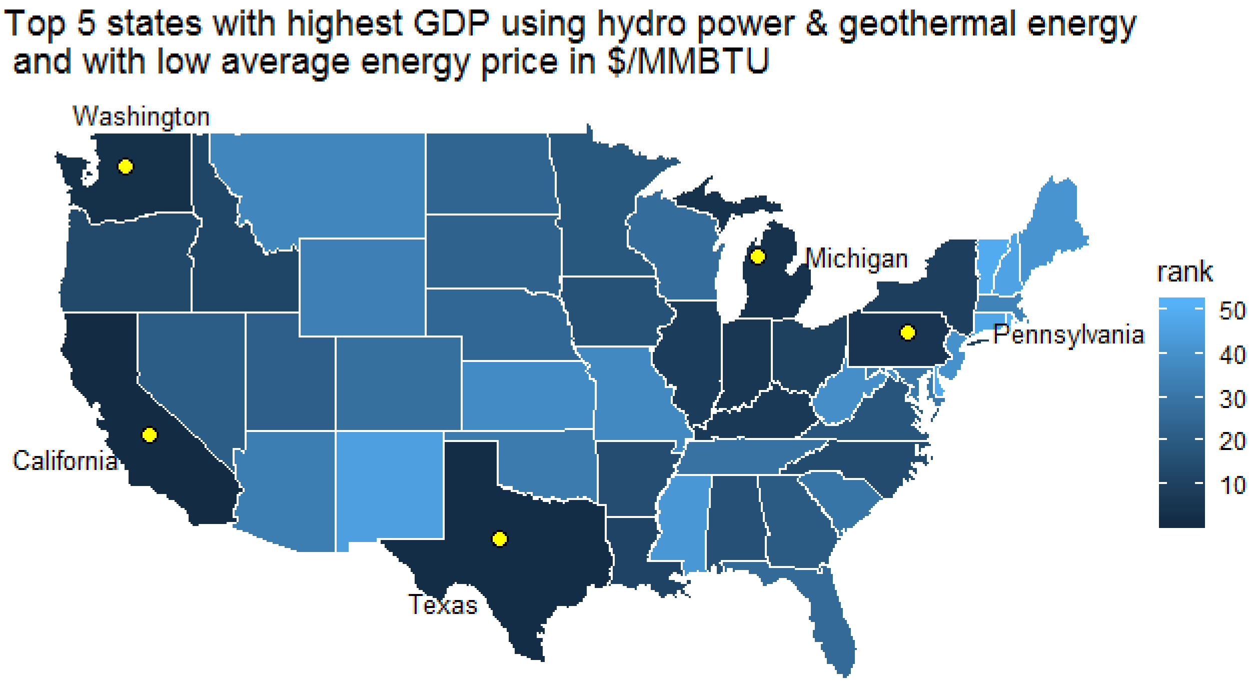
Creating dataframe ‘labs’ for latitude and longitude points for the top five states.

```
labs <- data.frame( long = c(-119.4179,-99.9018,-120.7401,-85.6024,-
77.1945),lat = c(36.7783,31.9686,47.7511,44.3148,41.2033))
```

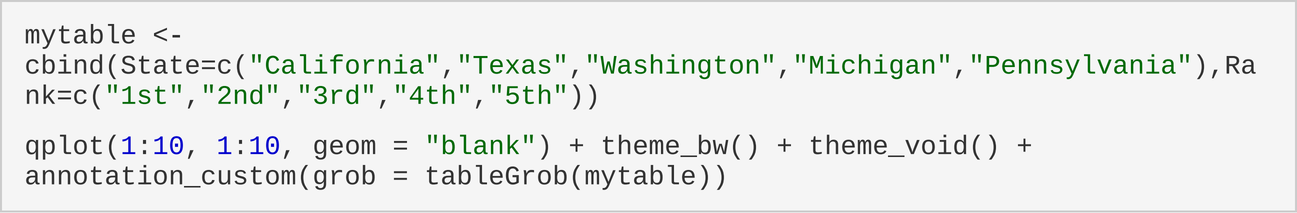
plotting the points on the map which represents the top five countries which aligns with the question’s requirements.

```
plotmap <- statemap + geom_point(data = labs, aes(x = long, y = lat,
group=NULL), shape = 21, color = "black",fill = "yellow", size = 3) +
annotate(geom="text", x=-124,
y=35.7783,label="California",color="black",size=3.5) + annotate(geom="text",
x=-102.9018, y=29, label="Texas",color="black",size=3.5) +
annotate(geom="text", x=-119.7401, y=49.7511,
label="Washington",color="black",size=3.5) + annotate(geom="text", x=-80,
y=44.3148, label="Michigan",color="black",size=3.5) + annotate(geom="text",
x=-68.1945, y=41.2033, label="Pennsylvania",color="black",size=3.5)
```

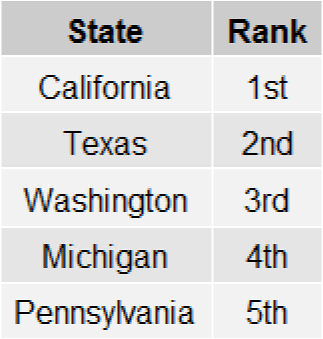
```
plotmap +ggtitle("Top 5 states with highest GDP using hydro power &
geothermal energy \n and with low average energy price in $/MMBTU")
```


Top 5 states with highest GDP using hydro power & geothermal energy
and with low average energy price in $/MMBTU
Washington
Michigan
Pennsylvania
California
Texas
rank
50
40
30
20
10

Unable to plot the table on the map, since it is only allowed for cartesians co-ordinates. ggarrange() can be used to include the map and table in one plot, but the map looked stretched.



```
mytable <-
cbind(State=c("California","Texas","Washington","Michigan","Pennsylvania"),Ra
nk=c("1st","2nd","3rd","4th","5th"))

qplot(1:10, 1:10, geom = "blank") + theme_bw() + theme_void() +
annotation_custom(grob = tableGrob(mytable))
```



| State | Rank |
|---|---|
| California | 1st |
| Texas | 2nd |
| Washington | 3rd |
| Michigan | 4th |
| Pennsylvania | 5th |

Appendix A:

**data source:** https://www.kaggle.com/lislejoem/us_energy_census_gdp_10-14

The GDP lollipop plot gdp_plot: a)ggplot(gdpmean[1:20,], aes(x=mean_gdp, y=reorder(State,mean_gdp)))I am choosing only 20 states value which are sorted in gdpmean data frame with higher gdp values coming first. Choosing all 50 states does not produce clear plot, however clear plots can be achieved by using ggsave() function allowing the plot to be saved with height and width we choose, this does give clear plots since the states are spread out but those plot do not fit on power point presentation. ‘Y’ axis is assigned to state name, by using reorder option so that in plot the states are aligned in the same order as in my data frame. b) geom_point(size=5,color=“#66CC33”,alpha=0.7)+labs(x = “Average GDP(in million USD)”, y=“States”, title=“Top 20 States with highest GDP from 2010-2014”)+theme_bw() geom_point() is plot the dots, the size and color is added. All the labels are specified under labs() function. theme_bw() replaces the exsisting plot background to white. c)theme(axis.title.y=element_blank(),panel.border = element_blank(), panel.grid.major = element_blank(),panel.grid.minor = element_blank(), axis.line = element_blank(), axis.ticks =

element_blank()) Inside theme() function ‘y’ title is removed since i found it unnecessary, border, grids, tick are removed too. d)gdp_table: For table insertion in the lollipop plot, new data frame gdp_table is created for top ten states having higher GDP. cbind() is column bind for variables to align one beside each other. e)qplot: quick plot to plot table. I have plotted with empty background so that the table can be placed over the lollipop plot. Average Price lollipop plot The lollipop plot for least average price for top 20 states

geom_segment(aes(y=State, x=13, yend=State, xend=mean_price),color=“#FFCC00”,alpha=0.3 )+scale_x_continuous(expand = c(13, 30))+scale_x_continuous(limits = c(13, 30)) I found geom_segment difficult to plot when using with xlim(). The points originally where plotting at far left, so i used xlim() so that i could use space to plot table. But geom_segment() does not plot well when using xlim(). I had to specially use scale_x_continuous() function to expand and set limits manually for geom_segment() to work. Rest of the plot attributes are changed same as for GDP lollipop plot.

Geo thermal and hydro power bar plot: The idea behind this plot is to merge the plot of top 10 high geothermal energy states with that of higher hydro power consuming states and show how many top states consume both hydro and geo thermal. As we see in the plot there are 19 states in total. Only california comes under top ten rank for both types of energy consumption.
In the code geom_col() is used instead of geom_bar(), since i wanted the bars to represent the values and not the counts. The co-ordinates are flipped for bars to be horizontal.

Mapping states which combine all the four condition that answer the question. All the data frames with mean calculated are joined together using left_join() function. The function seems to joined it arranging the data on the basis first variable mentioned as argument . Since this does not answer my question used filter() function to arrange the states satisfying the four condition. Further tried arrange() function , got the same data frame as above.
I had to rank the states as per individual condition and them sum their ranks, the states with least amount of sum tops the list and answers my question. The same data frame is then used to plot it in map.

To plot map, two data frame is required one the ranking data frame which i fond and other ggplots ‘state’ data frame. Both are merged , and plotted on map using ggplot(). The gradient is done as per the ranking and additional attributes are added. Points of top 5 states are plotted on the map using geom_point() feeding it with the latitude and longitude points to plot on map.