Usage of Sources of Energy

Code

Data summary

Introduction:

Today's generation is not new to the terminology: "Save The Environment". This was not even a thing about two or three decades ago. But today everyone is talking about it. So what exactly happened in the span of two decades. Under the idea of "DEVELOPMENT", we focused only on the luxuries that we could achieve for ourselves and completely ignored the environment. The development included increased use of non renewable sources of energy which increased the damage being caused. Our aim is to find the usage of sources of energy over the 5 year period which can help us determine if the citizens did really understand the emerging environment crisis and shift towards the renewable sources of energy or not. The analysis has been done on the basis of State wise and Division wise categories.

I have used the housing data from the PUMS micro data (https://data.census.gov/mdat/#/search?ds=ACSPUMS5Y2018 (https://data.census.gov/mdat/#/search?ds=ACSPUMS5Y2018)) for the analysis as it includes the columns like Electricity Usage, Fuel Cost and a few more which are the sources which use energy in various forms to generate the output.

Importing the Data and Renaming:

As the data has a lot of columns, it is advisable to import the columns that are needed for the analysis of the data. If not done so, the processing time of the data may increase drastically. We create a vector of the required fields and use 'fread' function to import the csv files and discard un-required fields on the go.

	Record_Type <chr>	Division <int>	State <int>	Adj_Housing_Factor <int>	Adj_Income_Factor <int>	Property_Value <int>	Fan
1	H	6	1	1054015	1061971	*NA*	
2	H	6	1	1054015	1061971	25000	
3	H	6	1	1054015	1061971	80000	

	Record_Type <chr>	Division <int>	State <int>	Adj_Housing_Factor <int>	Adj_Income_Factor <int>	Property_Value <int>	Fan
4	H	6	1	1054015	1061971	*NA*	
5	H	6	1	1054015	1061971	*NA*	
6	H	6	1	1054015	1061971	18000	

6 rows | 1-8 of 16 columns

As far as the removal of null values is concerned, we will do on the go while using it for the analysis.

Methodology and Findings:

Generating the weighted variables:

We start by generating the weighted variables for the analysis. We multiply the column with the weight value and divide it by 100000 to keep the data in the easily understandable format. The data which is collected on monthly basis has been multiplied by 12 months. All the values in the dataset are based on Annual Basis. The dataframes which require weighted values can be weighted as follows:

Hide

```
#Fuel Weight
data_all["Wt_Fuel_Cost"] = (data_all["Fuel_Cost"] * data_all["Adj_Housing_Factor"]) / 1000000

#Electricity Weight
data_all["Wt_Electricity_Cost"] = (data_all["Electricity_Cost"] * data_all["Adj_Housing_Factor"]
* 12) / 1000000

#Gas Weight
data_all["Wt_Gas_Cost"] = (data_all["Gas_Cost"] * data_all["Adj_Housing_Factor"] * 12) / 1000000

#Water Weight
data_all["Wt_Water_Cost"] = data_all["Water_Cost"] * data_all ["Adj_Housing_Factor"] / 1000000

#Insurance Weight
data_all["Wt_Insurance"]=data_all["Insurance"]*data_all["Adj_Housing_Factor"]/1000000

#Weighted Annual Family Income
data_all["Wt_Family_Income"]=data_all["Family_Income"]*data_all["Adj_Housing_Factor"]*12/1000000

#Weighted Annual Rent
data_all["Wt_Rent"]=data_all["Rent"]*data_all["Adj_Housing_Factor"]*12/1000000

summary(data_all)
```

```
Record_Type           Division         State       Adj_Housing_Factor
Length:7487361     Min.   :1.000   Min.   : 1.00   Min.   :1000000
Class :character   1st Qu.:3.000   1st Qu.:12.00   1st Qu.:1021505
Mode  :character   Median :5.000   Median :27.00   Median :1034680
                   Mean   :5.124   Mean   :27.83   Mean   :1029136
                   3rd Qu.:7.000   3rd Qu.:42.00   3rd Qu.:1036463
                   Max.   :9.000   Max.   :56.00   Max.   :1054015

Adj_Income_Factor Property_Value    Family_Income     Electricity_Cost
Min.   :1011189   Min.   :    100   Min.   : -21500   Min.   :  1.0
1st Qu.:1029257   1st Qu.: 100000   1st Qu.:  39000   1st Qu.: 70.0
Median :1035988   Median : 180000   Median :  70000   Median :120.0
Mean   :1036534   Mean   : 276505   Mean   :  94503   Mean   :138.5
3rd Qu.:1045195   3rd Qu.: 320000   3rd Qu.: 116030   3rd Qu.:180.0
Max.   :1061971   Max.   :6308000   Max.   :3164000   Max.   :660.0
                  NA's   :3139781   NA's   :3402406   NA's   :1355137
  Fuel_Cost       House_Heating_Fuel   Water_Cost      Lot_Size_in_Acres
Min.   :   1.0    Min.   :1.0        Min.   :   1.0    Min.   :1.0
1st Qu.:   2.0    1st Qu.:1.0        1st Qu.:   2.0    1st Qu.:1.0
Median :   2.0    Median :2.0        Median : 220.0    Median :1.0
Mean   : 115.6    Mean   :2.2        Mean   : 411.5    Mean   :1.3
3rd Qu.:   2.0    3rd Qu.:3.0        3rd Qu.: 660.0    3rd Qu.:1.0
Max.   :7800.0    Max.   :9.0        Max.   :4600.0    Max.   :3.0
NA's   :1355137   NA's   :1355137    NA's   :1355137   NA's   :2170040
   Gas_Cost        Property_Tax       Insurance            Rent
Min.   :  1.0     Min.   : 1        Min.   :   0      Min.   :   4
1st Qu.:  3.0     1st Qu.:19        1st Qu.: 450      1st Qu.: 520
Median : 20.0     Median :31        Median : 800      Median : 790
Mean   : 46.3     Mean   :34        Mean   : 988      Mean   : 921
3rd Qu.: 60.0     3rd Qu.:50        3rd Qu.:1200      3rd Qu.:1200
Max.   :640.0     Max.   :68        Max.   :9400      Max.   :4000
NA's   :1355137   NA's   :3202733   NA's   :3202733   NA's   :5660370
Wt_Electricity_Cost  Wt_Fuel_Cost      Wt_Gas_Cost      Wt_Water_Cost
Min.   :  12.0      Min.   :   1.0    Min.   :  12.0    Min.   :   1.0
1st Qu.: 870.6      1st Qu.:   2.0    1st Qu.:  36.8    1st Qu.:   2.1
Median :1471.0      Median :   2.1    Median : 240.0    Median : 207.3
Mean   :1710.1      Mean   :  66.3    Mean   : 572.5    Mean   : 390.0
3rd Qu.:2276.7      3rd Qu.:   2.1    3rd Qu.: 746.3    3rd Qu.: 641.5
Max.   :8194.7      Max.   :2145.2    Max.   :7946.3    Max.   :2145.2
NA's   :1355137     NA's   :1452266   NA's   :1355137   NA's   :1436290
 Wt_Insurance     Wt_Family_Income     Wt_Rent
Min.   :   0      Min.   :-25296    Min.   :   48
1st Qu.: 414      1st Qu.:     0    1st Qu.: 6219
Median : 776      Median :     0    Median : 9328
Mean   : 797      Mean   :  5031    Mean   :10222
3rd Qu.:1200      3rd Qu.:  9561    3rd Qu.:13484
Max.   :2145      Max.   : 25742    Max.   :25742
NA's   :3546011   NA's   :7432588   NA's   :5750860
```

I have labelled the factors so as to understand provide a better understanding of the analysis which is essentially based on the states and regions of United States of America.

Hide

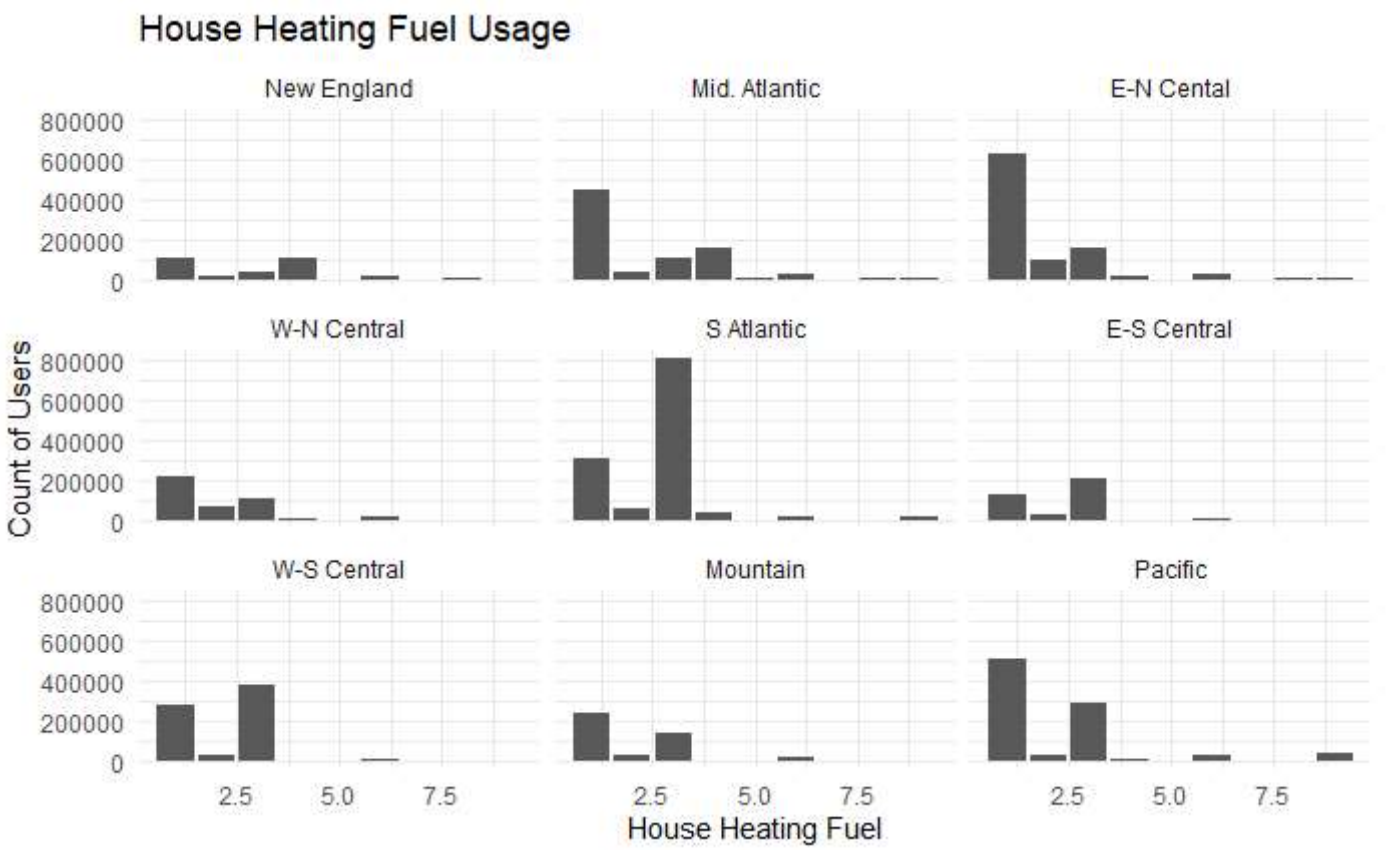
```
data_all$Division <- factor(data_all$Division,
                            levels =c(1,2,3,4,5,6,7,8,9),
                            labels = c("New England","Mid. Atlantic",
                                       "E-N Cental","W-N Central",
                                       "S Atlantic","E-S Central",
                                       "W-S Central",
                                       "Mountain","Pacific"))

data_all$State <- factor(data_all$State,
                         levels =c(1,2,4,5,6,8,9,
                                   10,11,12,13,15,16,17,18,
                                   19,20,21,22,23,24,25,26,27,
                                   28,29,30,31,32,33,34,35,36,
                                   37,38,39,40,41,42,44,45,
                                   46,47,48,49,50,51,53,54,
                                   55,56,72),
                         labels =c("AL","AK","AZ","AR","CA","CO","CT","DE",
"DC","FL","GA","HI","ID","IL","IN","IA",                                  "KS","KY","LA","ME","M
D","MA","MI","MN",                                  "MS","MO","MT","NE","NV","NH","NJ","NM",
"NY","NC","ND","OH","OK","OR","PA","RI",                                  "SC","SD","TN","TX","U
T","VT","VA","WA",                                  "WV","WI","WY","PR"))
```

Analysis

House Heating Fuel:

We are now interested in finding the method used for House heating. From the below graphs, we can easily understand that across country most people use Utility Gas for the house heating except for the South Atlantic. The people of South Atlantic region depend on electricity for the house heating. The maximum count of usages for the Utlilty Gas is approximately 600000 whereas for the electricity it is 800000. From the State-wise distribution, we can see that the use of other sources of energy is quite negligible. Hence, we are quite sure that the energy sources used by the most of the people across the United States is harmful to the nature.

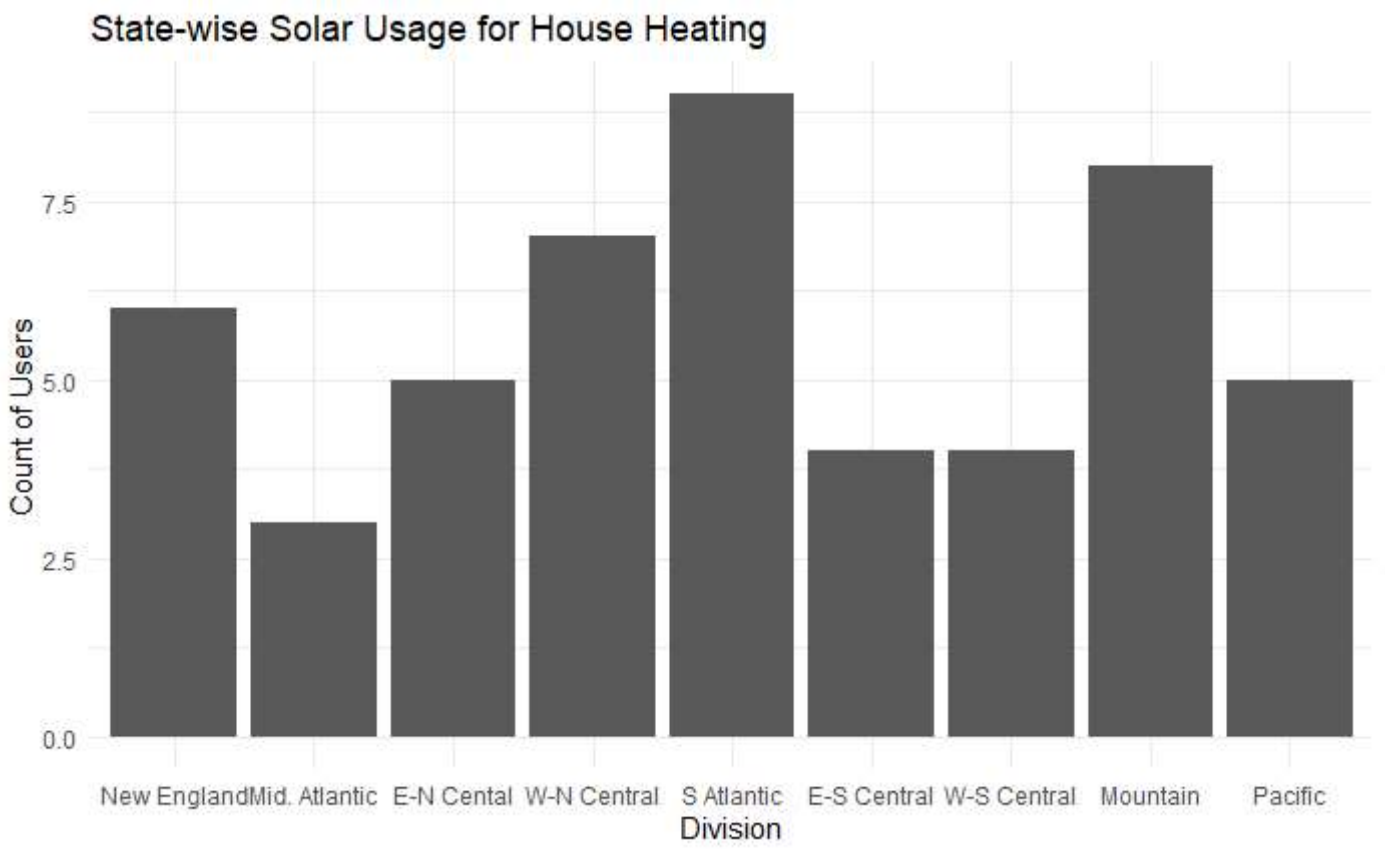


Now we shift our focus from the other sources of the fuel to the renewable sources of energy like Solar Energy. The total number of users for solar enegry are 7171 in the whole of United States. This number is extremely small when compared to the other sources of energy mentioned in the above paragraphs. However, California State and Mountain Region of the USA have the maximum count of solar users.

State	Division	Number_of_Users
<fctr>	<fctr>	<int>
AL	E-S Central	7
AK	Pacific	2
AZ	Mountain	610
AR	W-S Central	9
CA	Pacific	2769
CO	Mountain	278

6 rows

The large difference in the number of users is clearly visibile in both the graphs. In the above graph, the number of solar users are not visible however upon searching for specific users of solar energy, we come to know the real scenario.



Fuel Cost

The fuel cost is highest in the Middle Atlantic Region. The cities with the highest fuel cost are in the New York and Pennsylvania. The tax rates in New York are quite high.

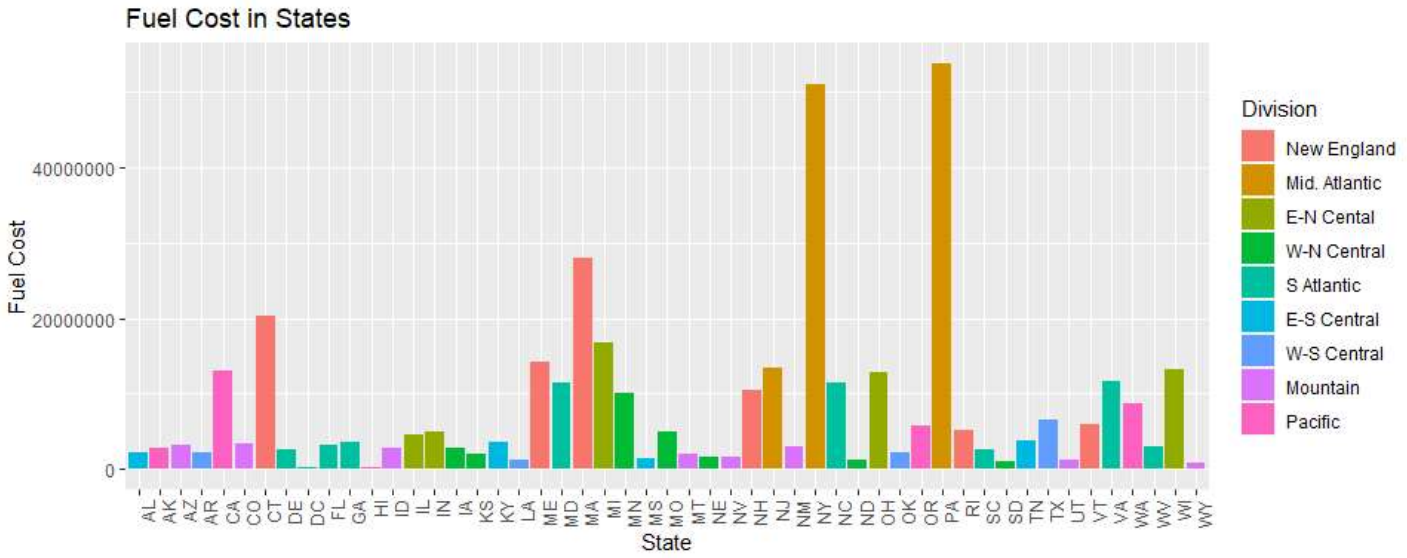
Hide

```
fuel <- data_all %>%
  select(Wt_Fuel_Cost,State,Division,Fuel_Cost) %>%
  na.omit()


fuel$Fuel_Cost <- factor(fuel$Fuel_Cost,
                   levels =c(1,2),
                   labels = c("Inc. in Rent","Fuel Not Used"))

fig(10,4)

fuel %>%
  ggplot()+geom_col(aes(x=State,y=Wt_Fuel_Cost,fill=Division))+
  theme(axis.text.x = element_text(angle = 90, hjust=1))+ggtitle("Fuel Cost in States")+ylab("Fu
el Cost")
```



Gas Cost:

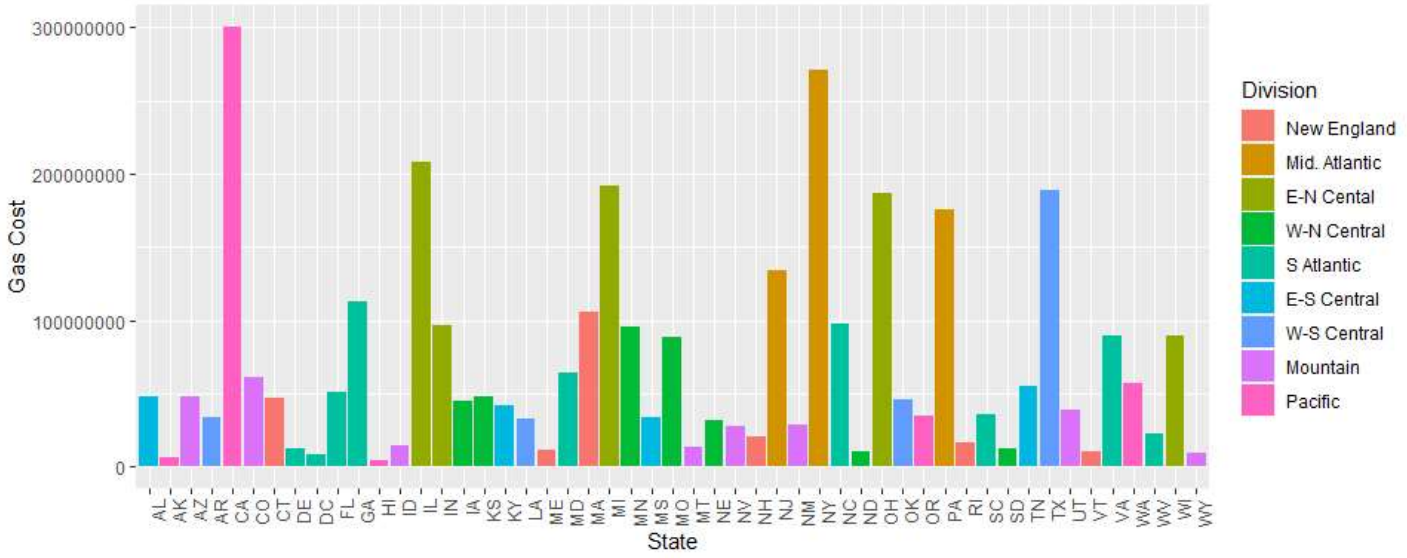
The cost of Gas is highest in the state of California followed by New York. The high taxes and strict environmental laws make the gas price comparatively high.

Hide

```
gas <- data_all %>%
  select(Wt_Gas_Cost,State,Gas_Cost,Division) %>%
  na.omit()

gas$Gas_Cost <- factor(gas$Gas_Cost,
                   levels =c(1,2,3),
                   labels = c("Inc. in Rent","Incl. in Electricity",
                              "Gas Not Used"))

ggplot(gas)+geom_col(aes(x=State,y=Wt_Gas_Cost,fill=Division))+theme(axis.text.x = element_text
(angle = 90, hjust=1))+ylab("Gas Cost")
```



The water cost is maximum in the Texas state due to draught like conditions which aroused between a the 5 year period when the data was collected. Due to high consumption of water in California, the government planned to levy high water tax on drinking water to stop people from misusing the available water resources.

Water Cost

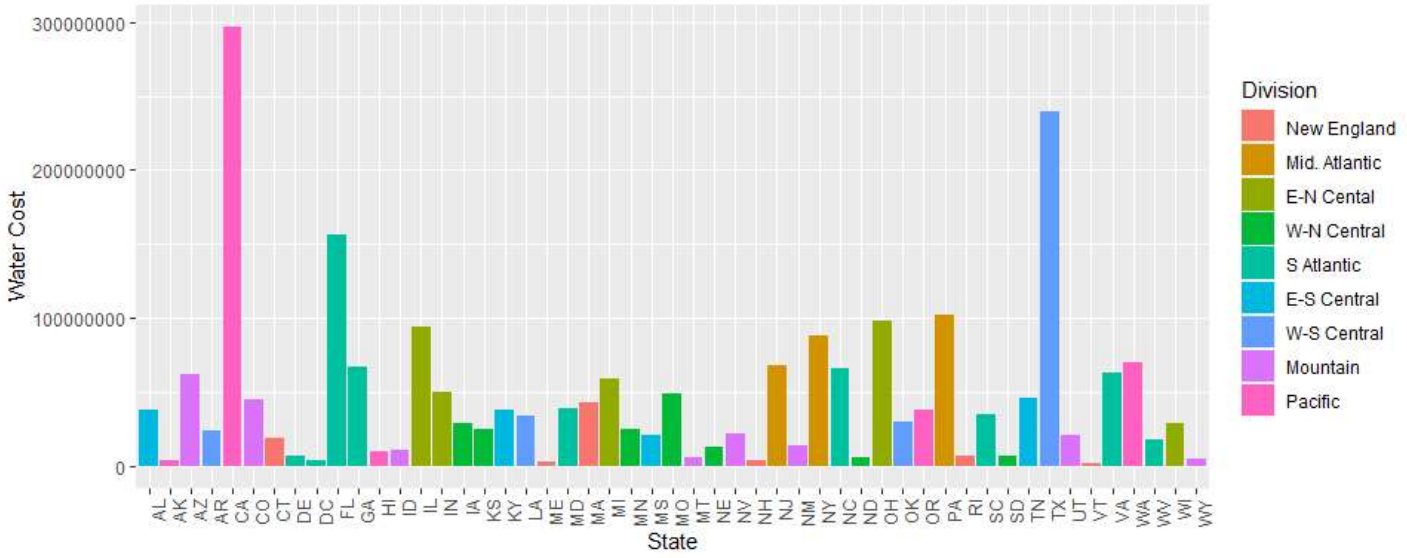
Hide

```
#WATER

water <- data_all %>%
  select(State,Wt_Water_Cost,Water_Cost,Division)%>% na.omit()

water$Water_Cost <- factor(water$Water_Cost,
                     levels =c(1,2),
                     labels = c("Inc. in Rent","Not Used"))

ggplot(water)+
  geom_col(aes(x=State,y=Wt_Water_Cost,fill=Division))+
  theme(axis.text.x = element_text(angle = 90, hjust=1))+ylab("Water Cost")
```



Suspected Solution

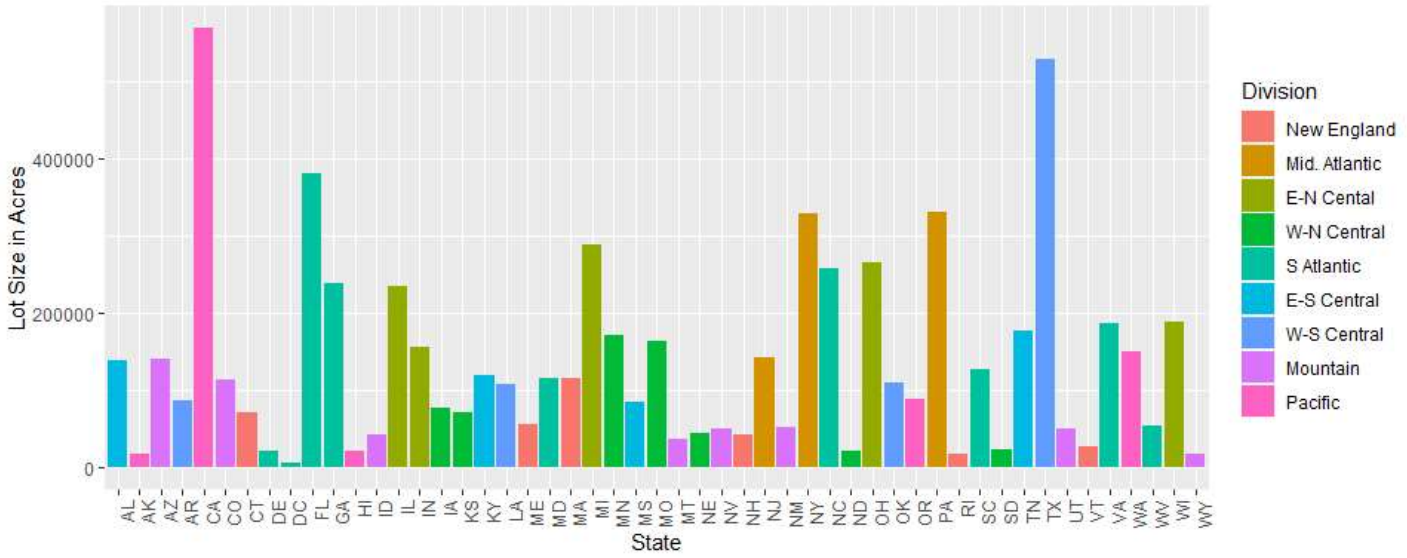
To solve all the crisis, we need land to shift to non renewable sources of energy. So I plotted the land availability in Acres and found that although the prices levied on the necessities are high, the only states with maximum land area available are California, Texas and Florida. These barren lands can be used for solar energy projects.

Hide

```
#ACRES

acres <- data_all %>% select(State, Division, Lot_Size_in_Acres) %>% na.omit()

ggplot(acres, aes(State, Lot_Size_in_Acres,fill=Division)) + geom_col()+theme(axis.text.x = elem
ent_text(angle = 90, hjust=1))+ylab("Lot Size in Acres")
```



###Linear Regression Model

Linear regression models are used to show or predict the relationship between two variables or factors. We are using multiple variables to specify predictor for the Property Tax.

We use this method to establish that a correlation exists between variables. But correlation is not the same as causation. Even a line in a simple linear regression that fits the data points well may not say something definitive about a cause-and-effect relationship.

####Linear Regression Model to determine relationship between Property Tax and Other Charges

Hide

```
data_for_model <- data_all %>%
  select(Wt_Insurance,Property_Value,
         Property_Tax,State,House_Heating_Fuel,Division,
         Wt_Water_Cost,Wt_Gas_Cost,
         Wt_Fuel_Cost,Wt_Electricity_Cost)

model_1 <- lm(Property_Tax~Wt_Insurance+Property_Value+Division+
              Wt_Water_Cost+Wt_Gas_Cost+Wt_Electricity_Cost+Wt_Fuel_Cost
              ,data = data_for_model)

summary(model_1)
```

```
Call:
lm(formula = Property_Tax ~ Wt_Insurance + Property_Value + Division +
    Wt_Water_Cost + Wt_Gas_Cost + Wt_Electricity_Cost + Wt_Fuel_Cost,
    data = data_for_model)

Residuals:
     Min       1Q   Median       3Q      Max
-194.401   -9.803   -0.762    9.873   59.648

Coefficients:
                            Estimate     Std. Error t value
(Intercept)            31.5059546981   0.0434159044  725.68
Wt_Insurance            0.0087301825   0.0000147275  592.78
Property_Value          0.0000236696   0.0000000289  818.94
DivisionMid. Atlantic  -2.5037353018   0.0439244727  -57.00
DivisionE-N Cental     -9.9470208928   0.0430627380 -230.99
DivisionW-N Central   -17.2702085639   0.0476662972 -362.31
DivisionS Atlantic    -18.8408740926   0.0427971266 -440.24
DivisionE-S Central   -27.3461073951   0.0493737393 -553.86
DivisionW-S Central   -18.6255469540   0.0460142951 -404.78
DivisionMountain      -21.3239266719   0.0479975182 -444.27
DivisionPacific       -12.4599424385   0.0447492310 -278.44
Wt_Water_Cost           0.0028707511   0.0000170572  168.30
Wt_Gas_Cost             0.0003612438   0.0000085507   42.25
Wt_Electricity_Cost     0.0005326696   0.0000069033   77.16
Wt_Fuel_Cost           -0.0018616855   0.0000258834  -71.93
                              Pr(>|t|)
(Intercept)           <0.0000000000000002 ***
Wt_Insurance          <0.0000000000000002 ***
Property_Value        <0.0000000000000002 ***
DivisionMid. Atlantic <0.0000000000000002 ***
DivisionE-N Cental    <0.0000000000000002 ***
DivisionW-N Central   <0.0000000000000002 ***
DivisionS Atlantic    <0.0000000000000002 ***
DivisionE-S Central   <0.0000000000000002 ***
DivisionW-S Central   <0.0000000000000002 ***
DivisionMountain      <0.0000000000000002 ***
DivisionPacific       <0.0000000000000002 ***
Wt_Water_Cost         <0.0000000000000002 ***
Wt_Gas_Cost           <0.0000000000000002 ***
Wt_Electricity_Cost   <0.0000000000000002 ***
Wt_Fuel_Cost          <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.92 on 3810098 degrees of freedom
  (3677248 observations deleted due to missingness)
Multiple R-squared:  0.4087,    Adjusted R-squared:  0.4086
F-statistic: 1.881e+05 on 14 and 3810098 DF,  p-value: < 0.00000000000000022
```

Residuals: The section summarizes the residuals, the error between the prediction of the model and the actual results. Smaller residuals are better. Many a times this value is dependent on the number of variables and how they are taken. Since this dataset has values which are non continuous on time series basis, it is hard to obtain a smaller number of Residuals.

Coefficients: For each variable and the intercept, a weight is produced and that weight has other attributes like the standard error, a t-test value and significance. The average error is ~4% and the significance for the variables is almost 95%.

Estimate: This is the weight given to the variable. In the simple regression case (one variable plus the intercept), The estimate values are quite small and negligible which states that for every increase in property value, the effect will be minimal on the dependent variables.

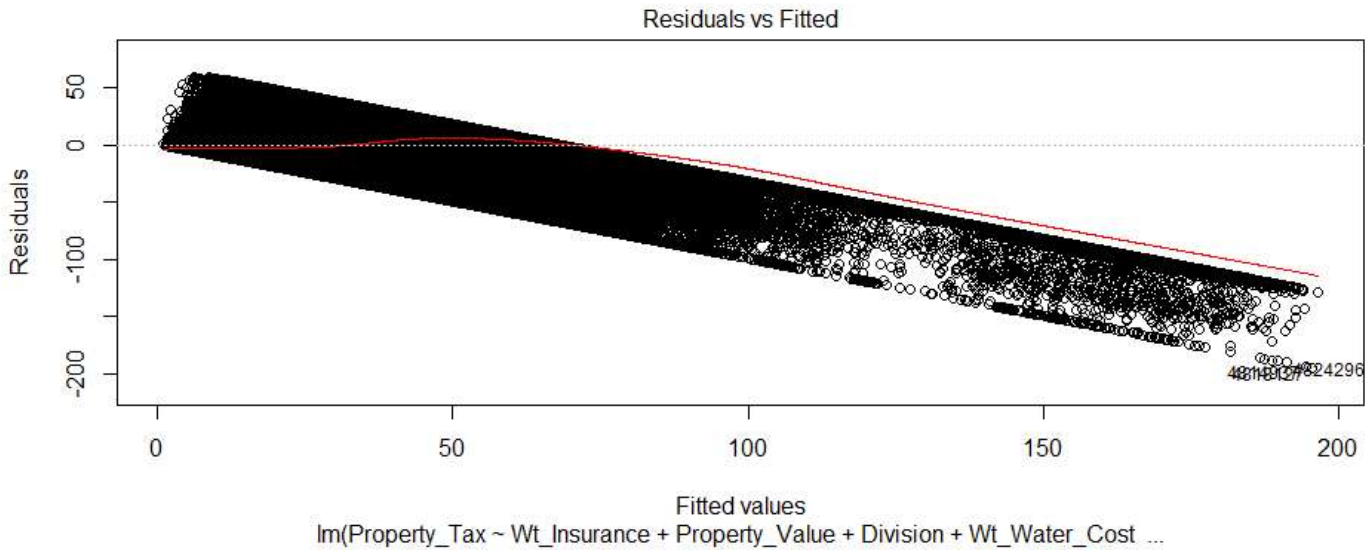
Std. Error: This tells you how precisely was the estimate measured. It's really only useful for calculating the t-value. The maximum standard error is 4.3% which is not a very large value. So the chances of error in the predicter variables should be minimum.

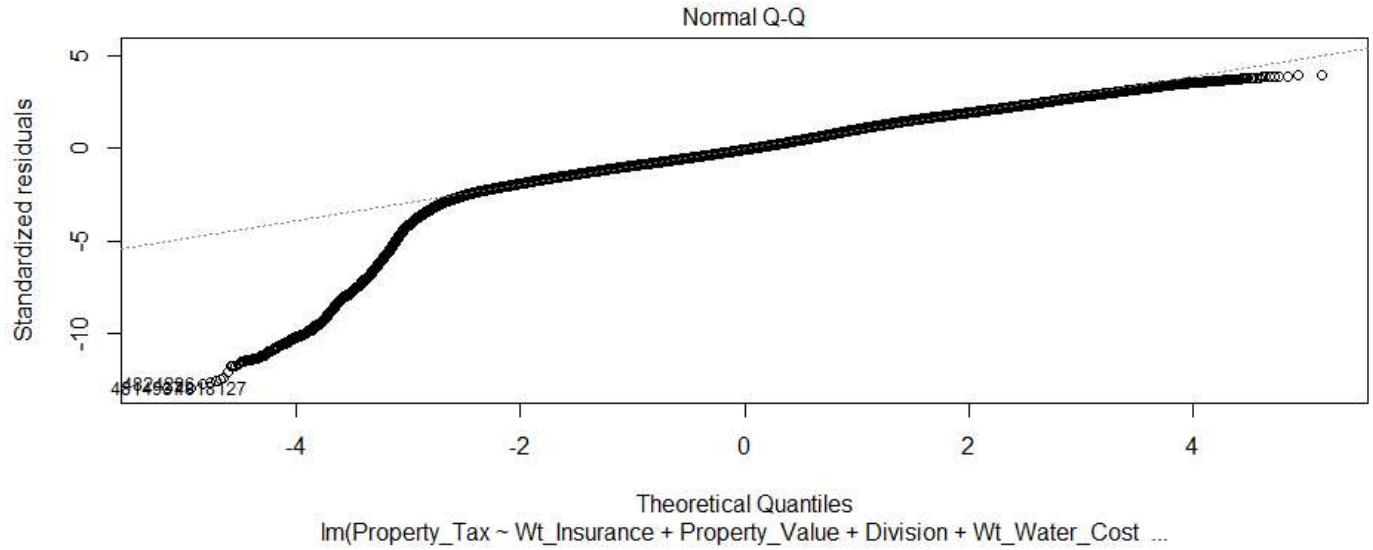
t-value and Pr(>[t]): The t-value is calculated by taking the coefficient divided by the Std. Error. It is then used to test whether or not the coefficient is significantly different from zero. If it isn't significant, then the coefficient really isn't adding anything to the model but the three stars at the end of the coefficients show that all of them are affecting the model analysis with over 95% effect.Those stars are also called the Pr(>|t|) or significance level.

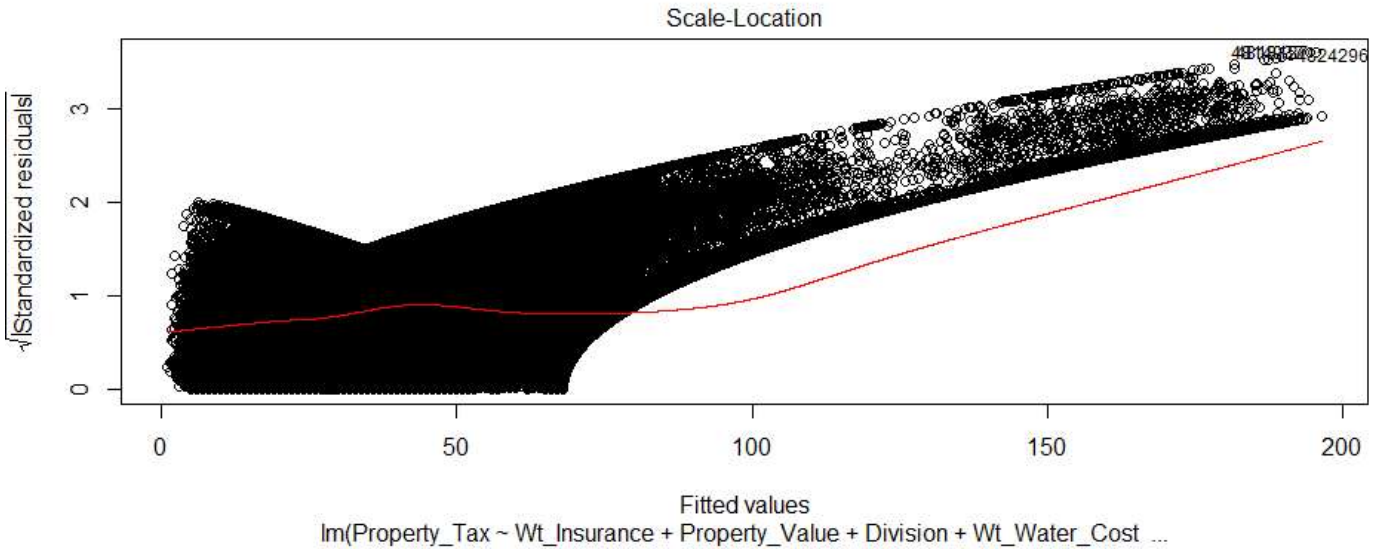
Performance Measures: Three sets of measurements are provided: Residual Standard Error: This is the standard deviation of the residuals. Smaller is better. Hence, we have the Residual Error as ~14%.

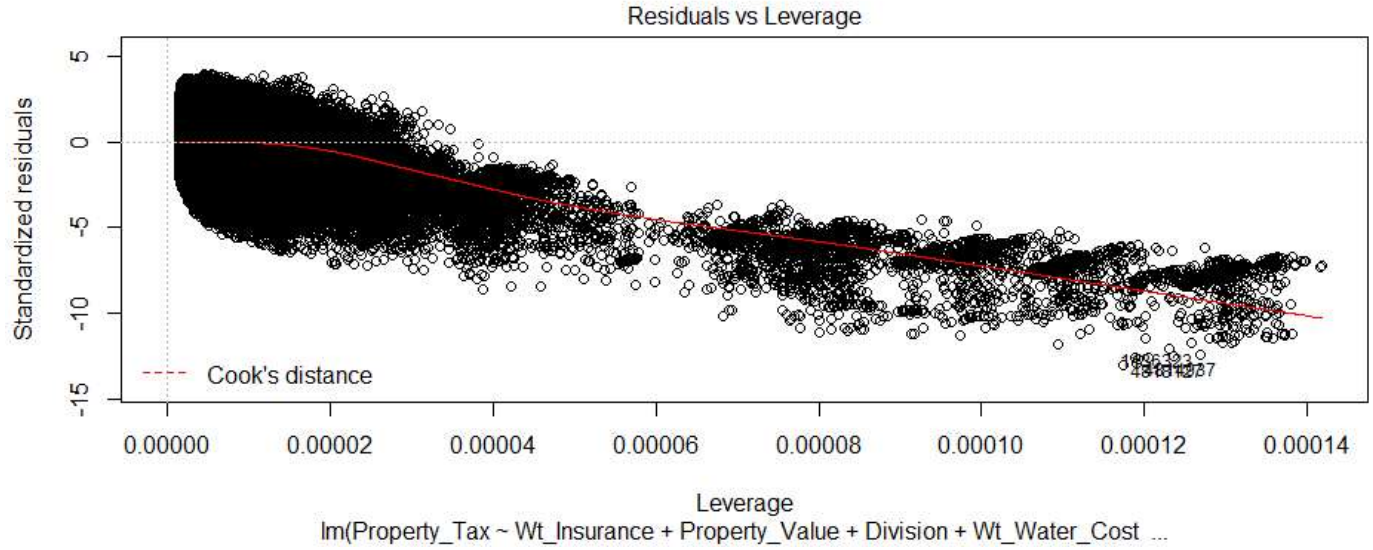
Multiple / Adjusted R-Square: R-squared shows the amount of variance explained by the model. Adjusted R-Square takes into account the number of variables. The R square value is also supposed to be the accuracy of the model which is ~40% in our case. But it is always not correct to reject a model on the basis of the R squared Values.

F-Statistic: The F-test checks if at least one variable's weight is significantly different than zero. This is a global test to help asses a model. The p-value is much less than 0.05 which shows that there is a relationship between the predictors and dependents.









Residual vs Fitted Graph: The points initially are not scattered hence the randomness in variance is negligible but the model soon starts to deviate away from the linearity. Hence the variables are linear at the initial stages but they soon tend to become non linear.

Normal Q-Q Plot: The residuals follow dotted line perfectly hence they are normally distributed. However, in order for the p-values to be believable, the residuals from the regression must look approximately normally distributed.

Scale Location Plot: This plot shows how residuals are spread among the predictors. Due to a large number of variables it is tough to determine the variance. However, on a scarce view, it can be figured that the variance is quite high among the variables. Towards the end values, the variance is quite high and does not relate to the expected or predicted values.

Residuals vs Leverage : According to the Cook's Distance, there are not many outliers in the plot. However, the model does not follow the trend at towards the end which makes it tough to conclude if the model is accurate enough or not.

Model Training and Prediction:

Hide

```
sample_data <- sample(seq_len(nrow(data_all)),size = floor(0.75 * nrow(data_all)))
train_data <- data_all[sample_data,]
test_data <- data_all[-sample_data,]

model_train <- lm(Property_Tax~Wt_Insurance+Property_Value+Division+
              Wt_Water_Cost+Wt_Gas_Cost+Wt_Electricity_Cost+Wt_Fuel_Cost
              ,data = train_data)

summary(model_train)
```

```
Call:
lm(formula = Property_Tax ~ Wt_Insurance + Property_Value + Division +
    Wt_Water_Cost + Wt_Gas_Cost + Wt_Electricity_Cost + Wt_Fuel_Cost,
    data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-194.681   -9.800   -0.764    9.867   59.148

Coefficients:
                            Estimate      Std. Error t value
(Intercept)            31.48390653708   0.05009364907  628.50
Wt_Insurance            0.00873129064   0.00001700235  513.53
Property_Value          0.00002372237   0.00000003336  711.20
DivisionMid. Atlantic  -2.49831736269   0.05068479449  -49.29
DivisionE-N Cental     -9.92674000516   0.04968680254 -199.79
DivisionW-N Central   -17.21952809948   0.05501457742 -313.00
DivisionS Atlantic    -18.82278940612   0.04938204930 -381.17
DivisionE-S Central   -27.32052196867   0.05697766027 -479.50
DivisionW-S Central   -18.61735826985   0.05310603928 -350.57
DivisionMountain      -21.31798255738   0.05539477929 -384.84
DivisionPacific       -12.46323236931   0.05163761297 -241.36
Wt_Water_Cost           0.00286466458   0.00001969086  145.48
Wt_Gas_Cost             0.00036091482   0.00000986020   36.60
Wt_Electricity_Cost     0.00053034195   0.00000796992   66.54
Wt_Fuel_Cost           -0.00187630194   0.00002987647  -62.80
                              Pr(>|t|)
(Intercept)           <0.0000000000000002 ***
Wt_Insurance          <0.0000000000000002 ***
Property_Value        <0.0000000000000002 ***
DivisionMid. Atlantic <0.0000000000000002 ***
DivisionE-N Cental    <0.0000000000000002 ***
DivisionW-N Central   <0.0000000000000002 ***
DivisionS Atlantic    <0.0000000000000002 ***
DivisionE-S Central   <0.0000000000000002 ***
DivisionW-S Central   <0.0000000000000002 ***
DivisionMountain      <0.0000000000000002 ***
DivisionPacific       <0.0000000000000002 ***
Wt_Water_Cost         <0.0000000000000002 ***
Wt_Gas_Cost           <0.0000000000000002 ***
Wt_Electricity_Cost   <0.0000000000000002 ***
Wt_Fuel_Cost          <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.92 on 2858262 degrees of freedom
  (2757243 observations deleted due to missingness)
Multiple R-squared:  0.4089,    Adjusted R-squared:  0.4089
F-statistic: 1.412e+05 on 14 and 2858262 DF,  p-value: < 0.00000000000000022
```

The model that we trained for the prediction of the data with 75% of trained data generated the same values as the linear regression model above. Hence, there is not a large difference due to training of the model.

Hide

```
data_predict <- predict(model_train,newdata = test_data)
predicted_values <- data.frame(cbind(test_data$Property_Tax,data_predict))
colnames(predicted_values)=c("Actual Values","Predicted Values")

head(predicted_values)
```

	Actual Values <dbl>	Predicted Values <dbl>
5	*NA*	*NA*
6	3	7.245384
7	26	*NA*
8	5	24.460531
15	2	13.729899
18	*NA*	*NA*

6 rows

There is a difference of 20-30 values between the predicted and the actual values. Hence, I think that the model is not an appropriate fit to predict something upon.

Conclusion

The United States of America ranks second in the list of countries dependent on the renewable technology. However, as we saw in the PUMS data 2013-2015, not a lot has changed when comparing renewable sources of energy to the other sources of energy. The solution to implement the sources of energy requires a lot of land and money.

I personally felt that the statistical part of the project is quite weak as I could not perform multiple tests. Moreover, I was more focused in making the data more visually better by trying to plot it on the map or animating it. However, the model does work out fine and generates satisfactory results but when keeping the practical scenario in picture, I think that this has a lot of future work. Looking at my systems hardware, I felt that this is the best I could do after multiple system failures.