Capstone (CYO) Project HarvardX PH125.9x

Gail McDaniel

December 31, 2020

Contents

1 Abstract

Scientific American published two articles, one in 2009 stating that the "entire world could get all of its energy—fuel as well as electricity—from wind, water and solar sources by 2030" (Jacobson, M.Z., & Delucchi, M.A., 2009), and another recommended the state of New York be powered with wind, water and sun, referred to "WWS" (Fischetti, M., 2013). This vision seeks to replace the use of fossil fuels (coal, oil and natural gas) with renewable energy sources worldwide by the year 2050. Renewable energy is energy from sources that are virtually inexhaustible and naturally replenishing, but finite in that the amount of energy is flow-limited, mechanically or electrically constrained, within a unit of time. The major types of renewable energy are:[1] biomass, hydropower, geothermal, wind and solar (Renewable Energy Explained*, 2018).

The purpose of this project is to determine if statistical analysis of published aggregate data available from the U.S. Energy Information Administration indicate that adequate electrical power could be generated with renewable sources. The published annual data is aggregated eliminating the possibility of analyzing whether renewable electrical energy generation could meet the daily, weekly and seasonal demands. Also, this project does not look at any of the financial costs for development nor land area needed to accommodate such sources of energy (Lyman, 2016).

1.1 Problem Statement

Does a statistical analysis of historical trend data indicate the feasibility for states to produce an adequate supply of electricity using renewable sources by 2050?

1.2 Tool Selection and Test Methodologies

In deciding which methods to use that work with continuous data, there are numerous statistical methodologies that could be used including KNN (K Nearest Neighbors) and SVM (Support Vector Machines). Decision Trees are a non-parametric supervised learning method that works with continuous and categorical data and can be used in both descriptive and predictive analytics. Logistic Regression and Naïve Bayes are non-descriptive methods that are both simple to implement. Logistic Regression is a probabilistic framework that provides confidence intervals and if additional training data is needed it can be quickly incorporated into the model.

- Decision Tree is easy to interpret and explain, can find nonlinear relationships and interactions within a dataset, is able to classify data, can work with incomplete data, and works with both continuous and categorical data.
- Naïve Bayes is incredibly simple to implement, works well with high dimension and is fast, works well with smaller training data, but assumes independence of data will perform poorly if not met.
- Logistic Regression is used to describe the relationship between one dependent binary variable and one or more nominal, ordinal, or ratio-level independent variable and works well when the dependent variable is binary [2]

For this analysis Linear Regression and Decision Tree will be used. Principal Component Analysis (PCA) will be used for dimension reduction. There are various R packages that can be utilized to aid in the PCA analysis and investigation

2 Data Exploration and Preparation

The U.S. Energy Information Administration (EIA) published data that is cleaned, complete, aggregated, at same scale, but may have been imputed (EIA, 2018). Two datasets are used in this study:

- Annual Sales by State, available from the U.S. Energy Information Administration, includes data consisting of Year, State, Industry Sector, Residential, Commercial, Industrial, Transportation, Other,

[1] https://www.nrc.gov/docs/ML0906/ML090680583.pdf
[2] https://towardsdatascience.com/statistical-testing-understanding-how-to-select-the-best-test-for-your-data-52141c305168

and Total. There are 8 independent variables (IVs) and 1 dependent variable (DV). This data is also a mix of qualitative and quantitative variables and contains 3580 observations.

- Annual Generation by State data, available from the U.S. Energy Information Administration, consists of Year, State, Type of Producer, Energy source, and amount of electricity generated, in Mega-watt hours. Four qualitative IVs and 1 quantitative DV are present in the data. Three of the Independent variables are categorical, and one continuous. The dependent variable is quantitative, the dataset contains 53,753 observations.

The datasets used in this study:

- Annual Retail sales of electricity to ultimate customers (EIA-861) - https://www.eia.gov/electricity/data/state/sales_annual.xlsx (Released: September 22, 2020)
- Net Generation by State by Type of Producer by Energy Source (EIA-906, EIA-920, and EIA-923)- https://www.eia.gov/electricity/data/state/annual_generation_state.xls (Released: September 22, 2020)

Note: *The files downloaded are Excel(xlsx) files they need to be converted to csv as well as a little manual cleaning. For details see Appendix A.*

2.1 Data Preparation

Data from the two datasets will be combined and transformed from narrow to wide for analysis and will provide 1530 observations with a mix of qualitative and quantitative data as shown below

The following variables are calculated for the study dataset:

- Renewable is a total of biomass, hydropower, geothermal, wind and solar. (Renewable Energy Explained*, 2018). Wood and Wood Derived Fuels are tracked separate by the EIA data and are included with the other Renewable sources for this study. (See Appendix B for Energy Sources Information)
- Fossil is a total of Coal, Natural Gas, Other Gasses, and Petroleum.
- The Ratio variable is calculated as a ratio of Renewable divided by Total.
- Level - if the ratio is .75 or greater the level is "High" otherwise the level will be "Low".

The initial process is to read, reshape, and enrich the data from both files and create a single dataset for the study. The initial structure of both files is narrow and long, for this analysis, the dataset also includes data collected from various sources but filtered to only include 'Total Electric Power Industry' or 'Total Power Industry'[3].

Loading Annual Generation Data

```
# Read the annual generation data
path = "annual_generation_state.csv"
db <- read.table(path, sep=",", header=TRUE)

#Select only observations of "Total Electric Power Industry"
db_filtered <- subset(db, TYPE.OF.PRODUCER == "Total Electric Power Industry")

# Lets remove the "US" rows since they're an aggregate of the states
db_filtered <- subset(db_filtered, STATE != "US-TOTAL" )
db_filtered <- subset(db_filtered, STATE != "US-Total" )

# let's drop columns "X", "X.1", "X.2", "X.3", "X.4", and "X.5"
db_filtered <- db_filtered[, -c(6:11)]
```

[3]https://www.eia.gov/electricity/monthly/

```r
# Rename some columns for ease of work
colnames(db_filtered)[colnames(db_filtered) == 'ï..YEAR'] <- 'Year'
colnames(db_filtered)[colnames(db_filtered) == 'STATE'] <- 'State'
colnames(db_filtered)[colnames(db_filtered) == 'TYPE.OF.PRODUCER'] <- 'TypeOfProducer'
colnames(db_filtered)[colnames(db_filtered) == 'ENERGY.SOURCE'] <- 'EnergySource'
colnames(db_filtered)[colnames(db_filtered) == 'GENERATION..Megawatthours.'] <- 'Generation'

# remove commas (,) from the generation column
db_filtered$Generation <- str_remove_all(db_filtered$Generation, "[,]")

# Let's make sure the 'Generation' column is numeric
db_filtered$Generation <-sapply(db_filtered$Generation,as.numeric)

# let's replase NA's with 0
db_filtered$Generation <- replace_na(db_filtered$Generation, 0)

nrow_narrow <- nrow(db_filtered)
ncol_narrow <- ncol(db_filtered)
```

The initial narrow shape of the data with multiple rows per state per data year, and after being reshaped the data is wide. Initial layout of the data – narrow and long (14360 observations by 5 variables)

After the records have been read-in to the data frame and filtered the records must be transformed. This can be accomplished using the spread() function in the "tidyr"[4] library, a package that provides a way to neatly group data for the r language.

```r
# reshape the data from narrow and Long to Wide and Short
wide_db <- spread(data = db_filtered,
              key = EnergySource,
              value = Generation)
```

Loading Annual Sales Data

```r
# Read the usage data
path = "sales_annual.csv"
sales = read.csv(path, header = TRUE)
names(sales)
```

```
## [1] "ï..Year"                  "State"
## [3] "Industry.Sector.Category" "Residential"
## [5] "Commercial"               "Industrial"
## [7] "Transportation"           "Other"
## [9] "Total"
```

```r
#Select only observations of "Total Electric Power Industry"
sales_filtered <- subset(sales, Industry.Sector.Category == "Total Electric Industry")


# let's drop columns "Residential","Commercial","Industrial","Transportation", and "Other"
sales_filtered <- sales_filtered[, -c(4:8)]


# Lets remove the "US" rows since they're an aggregate of the states
sales_filtered <- subset(sales_filtered, State != "US")
```

[4] https://tidyr.tidyverse.org/reference/spread.html

```
# Rename some columns for ease of work
colnames(sales_filtered)[colnames(sales_filtered) == 'ï..Year'] <- 'Year'
colnames(sales_filtered)[colnames(sales_filtered) == 'Industry.Sector.Category'] <- 'IndustrySector'


# remove commas (,) from the generation column
sales_filtered$Total <- str_remove_all(sales_filtered$Total, "[,]")

# Let's make sure the 'Total' column is numeric
sales_filtered <- sales_filtered %>% mutate_at('Total', as.numeric)

# let's replase NA's with 0
sales_filtered$Total <- replace_na(sales_filtered$Total, 0)
```

Herein, the data has been transferred to the new variable headings. There were various "Energy Source" attributes that were not identified for all states those are set to 0 in the dataset

Once the records in the dataframe have been pivoted, the next process is to add calculated columns. The Renewable column is the sum of Biomass, Geothermal, Hydroelectric, Solar, Wind and Wood. The Fossil column is the sum of Coal, Other Gases, Natural Gas and Petroleum. The Ratio column is calculated by dividing Renewable by Total. The Level column is calculated from the Ratio column if the ratio is .75 or greater then the level is "High" otherwise the level will be "Low". Date for the Usage column is from the Annual Sales dataset.

Final process in the data preparation is reorder the columns for ease of use before saving the dataset to a comma-separated-value (csv) file.

```
# reorder columns
wide_db <- wide_db[c("Year","State", "TypeOfProducer", "Biomass", "Geothermal", "ConvHydro",
  "Solar", "Wind", "Wood", "Coal",
  "OtherGases", "NatGas", "Petroleum", "Nuclear", "Other", "PSHydro",
  "Renewable", "Fossil", "Total", "Usage", "Ratio", "Level")]
```

3 Data Analysis

Exploratory Data Analysis plays a very important role in the Data Analysts Workflow, which includes univariate (1 – variable) and bivariate (2 – variable) analysis. This process includes analyzing data types, looking for missing values, identifying outliers and analyzing distributions of both numerical and categorical variables.

3.1 Univariate Analysis

One of the first steps in the analysis is to check continuous variables to see the distribution, as shown in Plot 1. To start this process, the summary() function provides basic descriptive statistics and frequencies that provide insight, such as central tendency, variation and shape of the data as shown below.

```
summary(wide_db)
```

```
##       Year         State           TypeOfProducer        Biomass
##  Min.   :1990   Length:1530        Length:1530        Min.   :   -149
##  1st Qu.:1997   Class :character   Class :character   1st Qu.:   8603
##  Median :2004   Mode  :character   Mode  :character   Median :  86954
##  Mean   :2004                                         Mean   : 369838
##  3rd Qu.:2012                                         3rd Qu.: 421312
##  Max.   :2019                                         Max.   :3631313
##    Geothermal         ConvHydro             Solar                Wind
```

```
##  Min.   :       0   Min.   :        0   Min.   :      -5   Min.   :       0
##  1st Qu.:       0   1st Qu.:   510798   1st Qu.:       0   1st Qu.:       0
##  Median :       0   Median :  1432290   Median :       0   Median :       0
##  Mean   :  296772   Mean   :  5523321   Mean   :  192757   Mean   : 1459561
##  3rd Qu.:       0   3rd Qu.:  3502178   3rd Qu.:      71   3rd Qu.:  502327
##  Max.   :14948113   Max.   :104170551   Max.   :28331513   Max.   :83620371
##       Wood              Coal             OtherGases          NatGas
##  Min.   :   -689   Min.   :        0   Min.   :      0   Min.   :     -348
##  1st Qu.:      0   1st Qu.:  3890619   1st Qu.:      0   1st Qu.:   764789
##  Median : 213070   Median : 25862853   Median :      0   Median :  4038746
##  Mean   : 743360   Mean   : 33318427   Mean   : 250522   Mean   : 16119373
##  3rd Qu.:1192304   3rd Qu.: 43832171   3rd Qu.: 106803   3rd Qu.: 15396471
##  Max.   :6218978   Max.   :157896535   Max.   :5999490   Max.   :255630021
##    Petroleum           Nuclear             Other            PSHydro
##  Min.   :  -17940   Min.   : -124722   Min.   : -48433   Min.   :-2240077
##  1st Qu.:   49876   1st Qu.:       0   1st Qu.:      0   1st Qu.:       0
##  Median :  214231   Median : 8445672   Median :   8696   Median :       0
##  Mean   : 1460328   Mean   :14542748   Mean   : 191767   Mean   : -109985
##  3rd Qu.:  950232   3rd Qu.:25100520   3rd Qu.: 211516   3rd Qu.:       0
##  Max.   :41521651   Max.   :98735488   Max.   :3710700   Max.   : 2506665
##    Renewable            Fossil              Total
##  Min.   :        0   Min.   :     2880   Min.   :    35499
##  1st Qu.:  1477363   1st Qu.: 16231479   1st Qu.: 30486570
##  Median :  3471671   Median : 35860292   Median : 53052890
##  Mean   :  8585609   Mean   : 51148649   Mean   : 74358789
##  3rd Qu.:  7779450   3rd Qu.: 72965855   3rd Qu.:102189452
##  Max.   :105410768   Max.   :362763329   Max.   :483201031
##      Usage               Ratio             Level
##  Min.   :  4253840   Min.   :0.00000   Length:1530
##  1st Qu.: 18721528   1st Qu.:0.03036   Class :character
##  Median : 49710432   Median :0.06107   Mode  :character
##  Mean   : 67466079   Mean   :0.15395
##  3rd Qu.: 90355092   3rd Qu.:0.16522
##  Max.   :429343404   Max.   :0.99916
```

The dataset contains 1530 observations with 22 variables. Two (2) variables are categorical and nineteen (19) are continuous.

3.1.1 Renewable Energy Sources

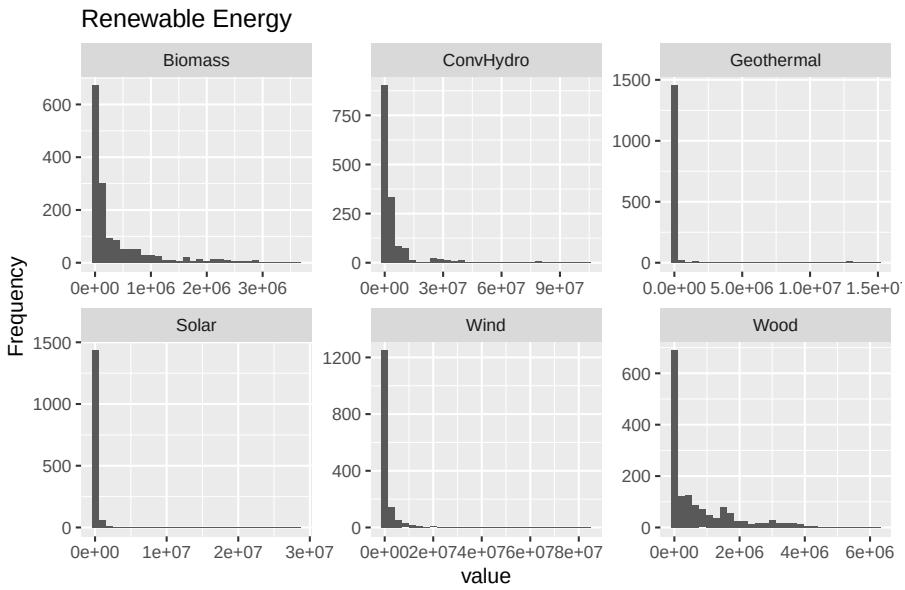
The following code filters the dataset by the major "Renewable" energy sources. From this filtered data a statistical summary is calculated, and various plots produced.

```
# Renewable Data
renewable_data <- wide_db[, c("Biomass", "ConvHydro", "Geothermal", "Solar", "Wind", "Wood")]

stargazer(renewable_data , type="text", summary=TRUE, rownames = TRUE,
      title="Renewable Sources", decimal.mark = ".", digits = 1,
        digits.extra = 1, digit.separator = "")
```

```
##
## Renewable Sources
## ========================================================================
## Statistic   N     Mean     St. Dev.  Min  Pctl(25) Pctl(75)     Max
## ------------------------------------------------------------------------
## Biomass    1530 369837.9   625792.8  -149  8603.2  421312.5   3631313
```

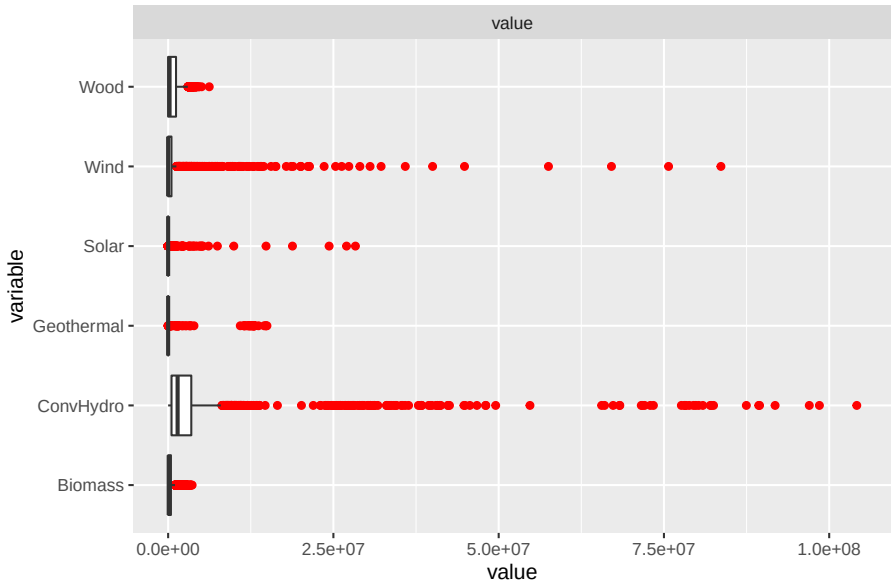
```
## ConvHydro   1530 5523321.0 13011833.0   0   510797.5 3502178.0 104170551
## Geothermal  1530  296772.4  1791770.0   0        0       0.0     14948113
## Solar       1530  192757.5  1448406.0  -5        0      70.8     28331513
## Wind        1530 1459561.0  5185336.0   0        0  502326.8     83620371
## Wood        1530  743359.7  1048492.0 -689      0     1192305.0   6218978
## ------------------------------------------------------------------------
```

Histogram of Renewable Sources



BoxPlot of Renewable Energy Sources

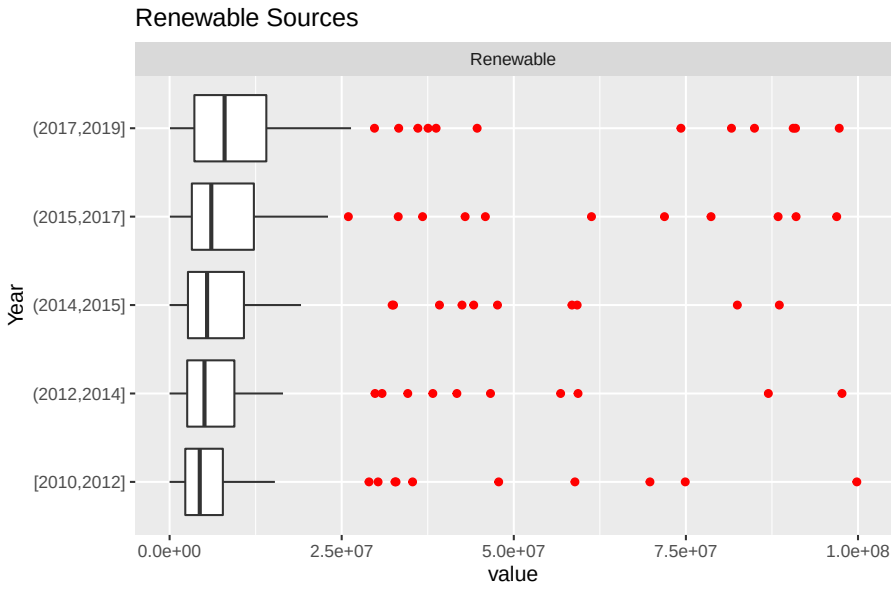
```
## No id variables; using all as measure variables
```



When looking at the box plot for all the years, 1990-2019, there is a considerable number of observations that are outliers. Since 'Renewable' were not really considered 25 years ago if the timeframe is shortened to the

last 10 years as in the following box plot there are considerably less observations that appear to be outliers.

10yr BoxPlot of Renewable Energy Sources



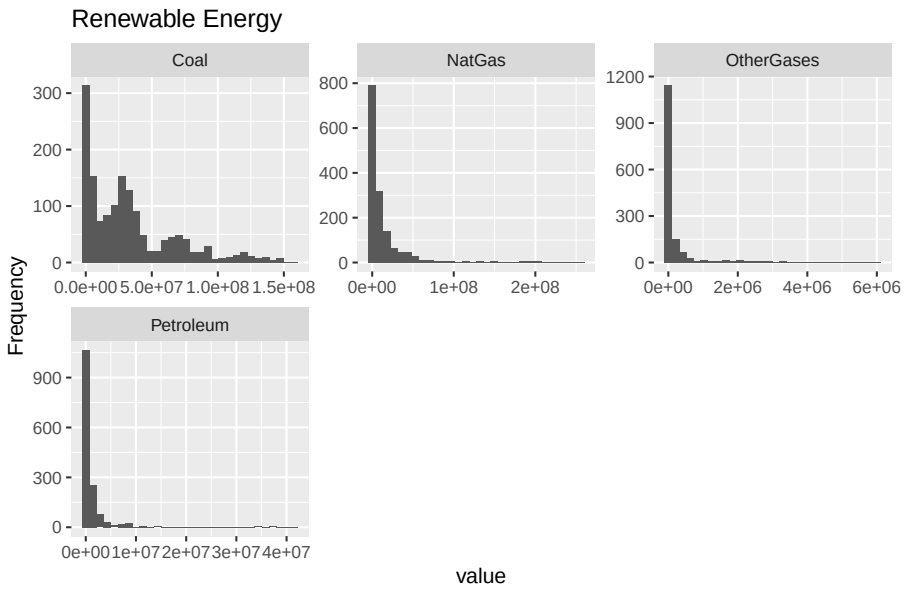
Fossil Fuel Sources

Renewable efforts vary from state to state and without additional filtering of the observations by State some observations still appear to be outliers. The following code filters the dataset by the major "Fossil Fuel" sources. From this filtered data, various plots and a statistical summary is calculated. For the box plot the data is pivoted for visual analysis.

```
##
## Fossil Fuel Sources
## ================================================================================
## Statistic    N      Mean      St. Dev.    Min   Pctl(25)  Pctl(75)     Max
## --------------------------------------------------------------------------------
## Coal        1530 33318427.0 33948229.0    0    3890619  43832171.0 157896535
## NatGas      1530 16119373.0 33011965.0  -348   764789.2  15396471  255630021
## OtherGases 1530   250522.0   676008.1     0       0       106803    5999490
## Petroleum   1530 1460328.0  4229965.0  -17940 49876.2    950232   41521651
## --------------------------------------------------------------------------------
```

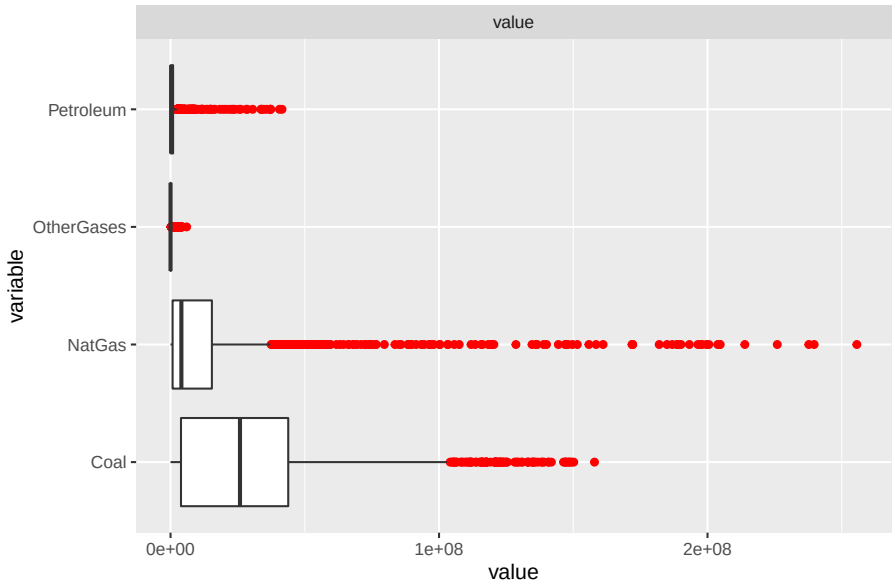
Other Gases has 0 for Pctl(25) and the mean is higher than the Pctl(75), this is a skewed distribution with a significant number of observations with value of 0. Reviewing at the fivenumber summary for OtherGases in the complete dataset summary the median is also 0,

Histogram of Fossil Fuel Sources



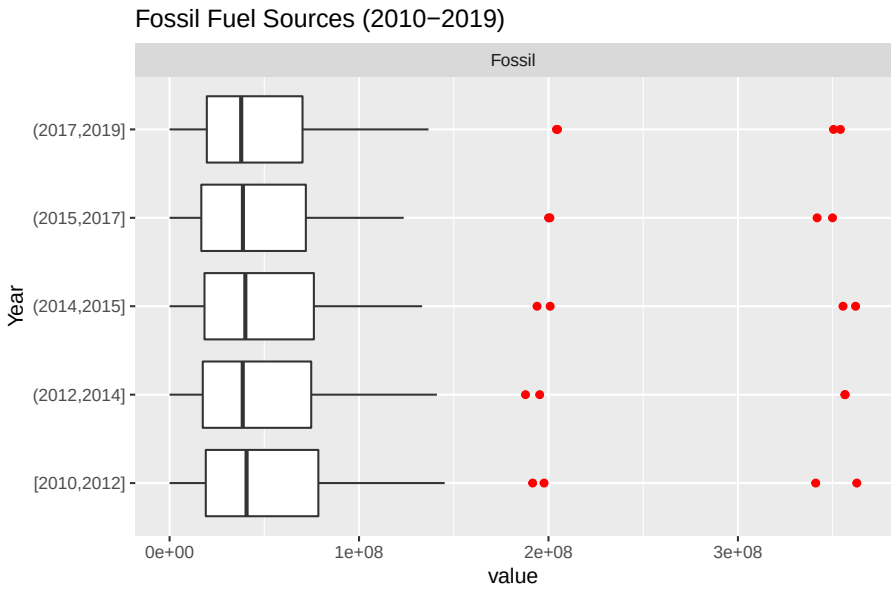
BoxPlot of Fossil Fuel Sources

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



When looking at the full dataset it appears there are a considerable number of outliers especially for Natural Gas. While filtering of the dataset by states may give a more accurate statistical description, in this analysis all states are combined and analyzed

10yr BoxPlot of Fossil Fuel Sources



Looking at the last 10 years there are considerably fewer observations that are outliers.

3.1.2 Other Energy Sources

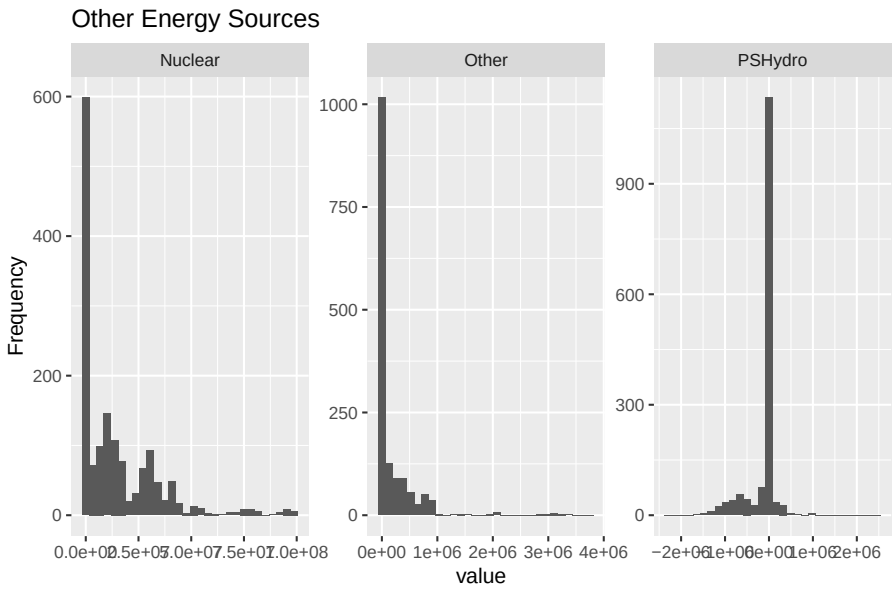
Other sources include pumped-storage hydroelectric, nuclear and other energy sources. PumpedStorage Hydroelectric (PSHydro) generates electricity for peak loads by using water previously pumped into an elevated reservoir during off-peak periods thus adding additional generation capacity ("Pumped-storage hydroelectric", n.d.). Nuclear is not a fossil fuel but a nonrenewable energy source taken from Uranium ore ("Nonrenewable Energy Explained" 2018). Most nuclear electric power plants use a process of splitting uranium atoms, this process is called fission. This process produces heat that turns water into steam, which is used to power turbine generators to produce electricity. Other includes batteries, chemicals, hydrogen, pitch, purchased steam, sulfur, miscellaneous technologies. Beginning in 2001, non-renewable waste (municipal solid waste from non-biogenic sources, and tire-derived fuels). ("Monthly Energy Review", 2019)

The following code filters the dataset by the major "Other" energy sources.

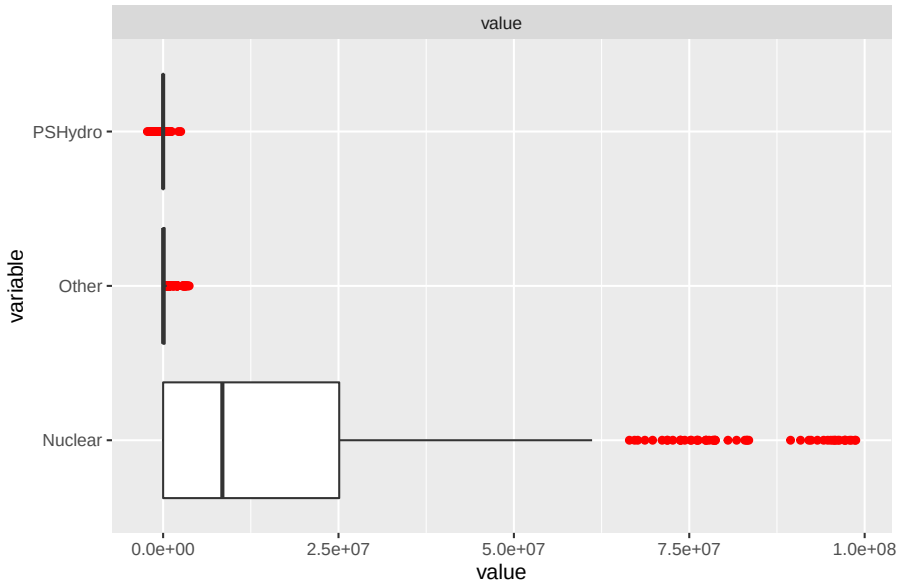
```
# Other Sources Data
other_data <- wide_db[, c("Nuclear", "Other", "PSHydro")]

stargazer(other_data , type="text", summary=TRUE, rownames = TRUE,
      title="Alternate/Other Sources", decimal.mark = ".", digits = 1,
        digits.extra = 1, digit.separator = "")
```

```
##
## Alternate/Other Sources
## ================================================================================
## Statistic  N      Mean      St. Dev.    Min    Pctl(25)  Pctl(75)     Max
## --------------------------------------------------------------------------------
## Nuclear   1530 14542748.0 19168969.0 -124722     0     25100520.0 98735488
## Other     1530  191767.2   443170.1   -48433     0      211515.8  3710700
## PSHydro   1530 -109984.7   336326.6  -2240077    0         0      2506665
## --------------------------------------------------------------------------------
```

Histogram of Other Energy Sources

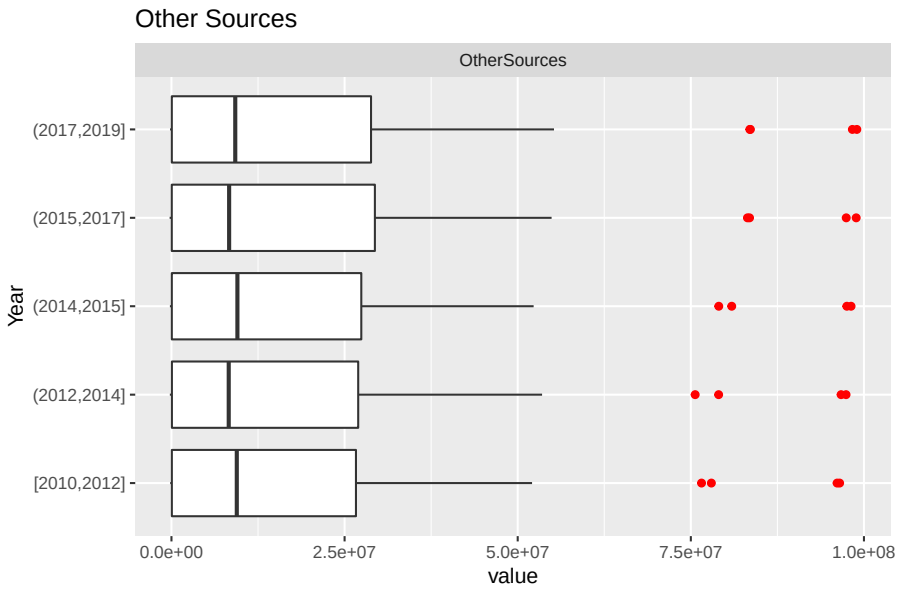



```
## No id variables; using all as measure variables
```



PSHydro generation totals are calculated by facility production minus energy used for pumping ("Monthly Energy Review", 2019).

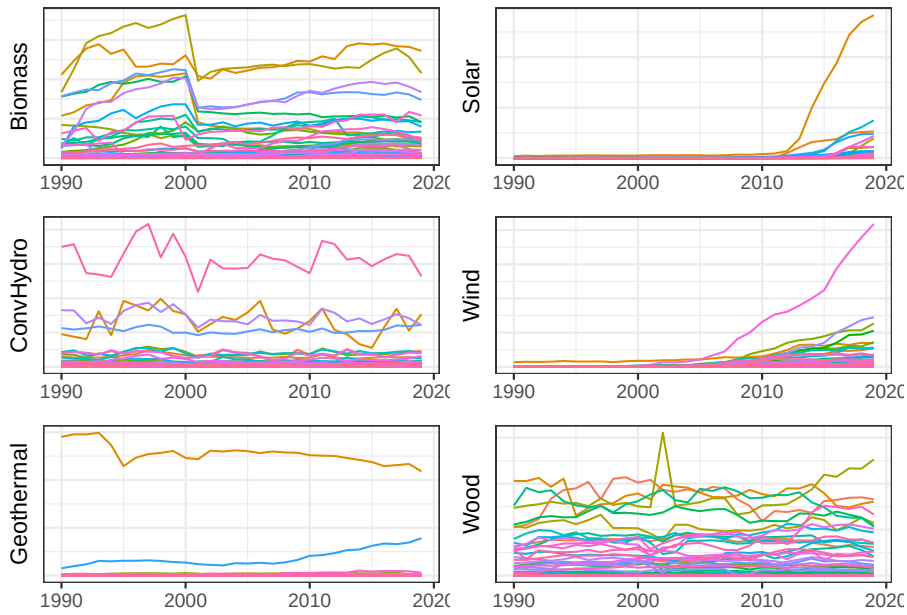
10Yr Boxplot of Other Energy Sources



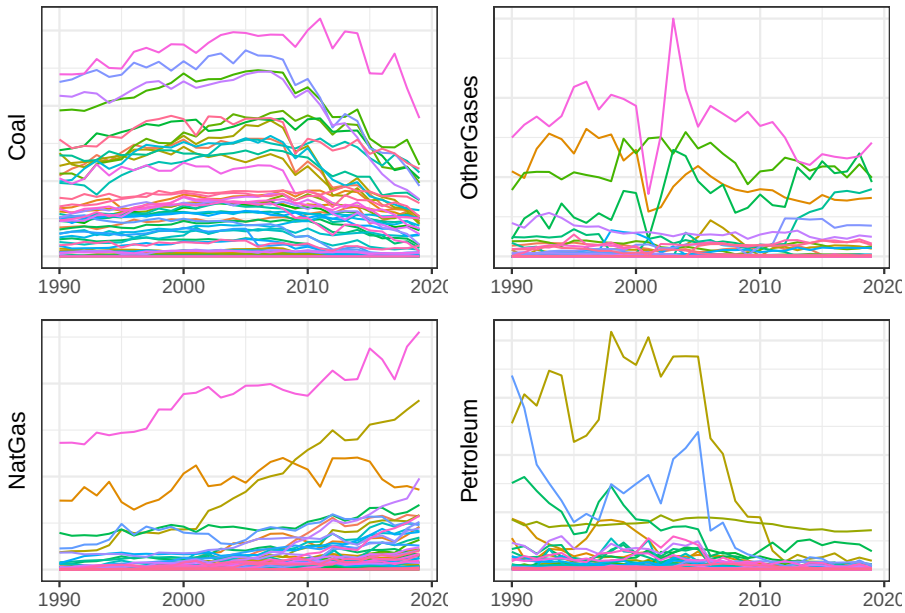
3.2 Bivariate Analysis

```
renewable_data <- wide_db[, c( "State", "Year", "Biomass", "ConvHydro", "Geothermal", "Solar", "Wind",
    group_by(State)
```

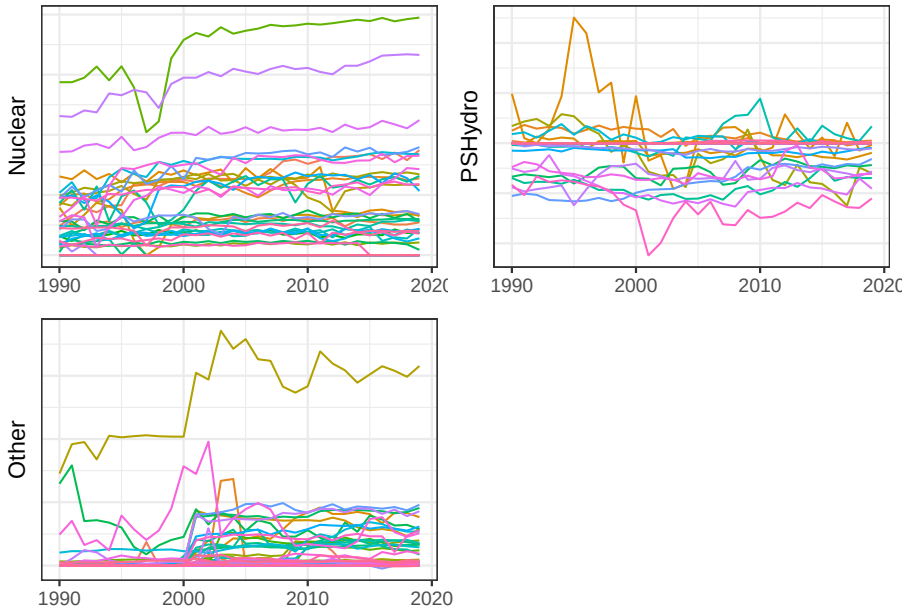
Renewable Sources



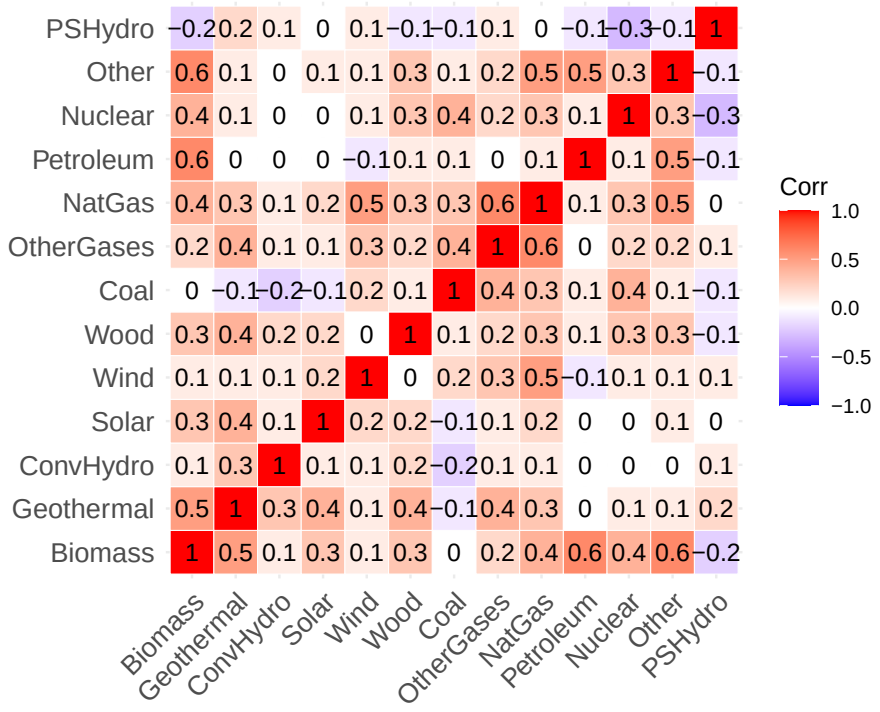
Fossil Fuel Sources



Other Energy Sources



Correlation Matrix of All Energy Sources



3.3 Multivariate Analysis

According to the Principal Component Analysis (PC4) the three most explanatory variables are Solar, Geothermal and Conventional Hydroelectric. Wood, Wind and Biomass are not significant in explaining the variance.

3.3.1 Principal Component Analysis(PCA)

PCA is a method used to reduce number of variables in your data by extracting important one from a large pool. It reduces the dimension of your data with the aim of retaining as much information as possible.[5]

```
## Standard deviations (1, .., p=13):
##  [1] 1.8752225 1.3705993 1.3028287 1.0915871 0.9782468 0.9199676 0.8248098
##  [8] 0.7751468 0.7440067 0.6994165 0.5295591 0.4065026 0.3784121
##
## Rotation (n x k) = (13 x 13):
##                    PC1         PC2         PC3         PC4          PC5
## Biomass     0.41372604 -0.16553323 -0.2974276  0.08327808 -0.098309987
## Geothermal  0.27580761  0.40709272 -0.2674088 -0.09110561  0.081821240
## ConvHydro   0.09808747  0.27407370 -0.3019930 -0.25702209  0.305048063
## Solar       0.19246291  0.30749758 -0.1780077 -0.05436718 -0.578028417
## Wind        0.18843312  0.26767444  0.3057114  0.13013091 -0.461789784
## Wood        0.29154573  0.03766735 -0.1588606 -0.37762452  0.304356326
## Coal        0.16851443 -0.17438058  0.5611183 -0.08569140  0.156395642
## OtherGases  0.31529155  0.24235653  0.3581562  0.07103435  0.299533885
## NatGas      0.43346211  0.12856971  0.2138186  0.14781575 -0.019284841
## Petroleum   0.22574333 -0.39260705 -0.2623997  0.37419890  0.090919787
## Nuclear     0.29620920 -0.25075023  0.1669580 -0.44709733 -0.036728562
```

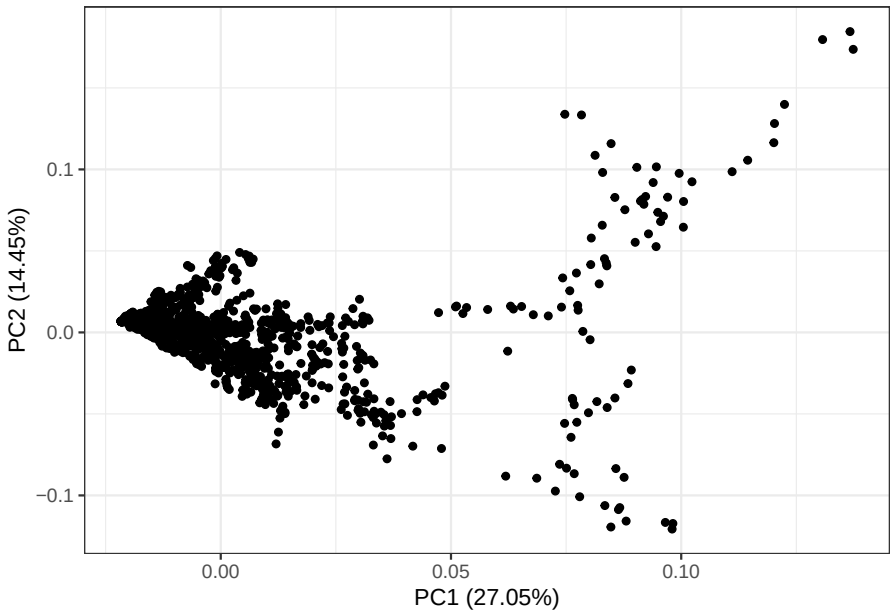
[5] https://towardsdatascience.com/principal-component-analysis-intro-61f236064b38

```
## Other       0.35430958 -0.28512764 -0.1089819  0.35535890  0.006726551
## PSHydro    -0.07524191  0.40018181  0.0315134  0.50574191  0.355491196
##                     PC6           PC7         PC8         PC9        PC10
## Biomass     0.002450888  0.2713645885 -0.16078209  0.20354173  0.04912203
## Geothermal  0.282075430  0.2405045020 -0.35393983  0.24316839 -0.07301742
## ConvHydro  -0.647120630  0.2586922322  0.34195345 -0.21215678  0.11306316
## Solar       0.361837346  0.1166283749  0.43267566 -0.35976429 -0.02885969
## Wind       -0.508078302 -0.1498397421 -0.05403990  0.25287545 -0.39188878
## Wood        0.162046996 -0.6352170548  0.09722670 -0.05725901 -0.42365314
## Coal        0.156283282  0.3295441572  0.31880532 -0.17053121 -0.25928902
## OtherGases  0.072986860  0.1710869958 -0.31711379 -0.34184642  0.09885148
## NatGas     -0.122186235 -0.2756659527 -0.11568478 -0.08206788  0.29983548
## Petroleum  -0.103647119  0.2120007281 -0.01852486 -0.13352544 -0.53630451
## Nuclear     0.026897987  0.1610009747  0.23477881  0.55013132  0.16453826
## Other      -0.012416636 -0.2780555214  0.26300591 -0.05892825  0.39948830
## PSHydro     0.173297022  0.0001269282  0.43834308  0.42562120 -0.06626105
##                    PC11        PC12        PC13
## Biomass     0.20598921  0.43574236  0.56632629
## Geothermal  0.32925651 -0.30065874 -0.38177607
## ConvHydro   0.05704248 -0.03380825 -0.01176702
## Solar      -0.19241975  0.01049434  0.02254972
## Wind        0.09223414  0.17818376 -0.16704726
## Wood        0.01869290  0.15185497  0.08791827
## Coal        0.49662761 -0.10527929  0.07643863
## OtherGases -0.39305649  0.41806601 -0.16658363
## NatGas     -0.07685566 -0.59999653  0.40099549
## Petroleum  -0.34569556 -0.28570645 -0.12304244
## Nuclear    -0.42198608 -0.06234708 -0.16377375
## Other       0.28074409  0.16861190 -0.48786566
## PSHydro    -0.12879729  0.06217991  0.15124703
```

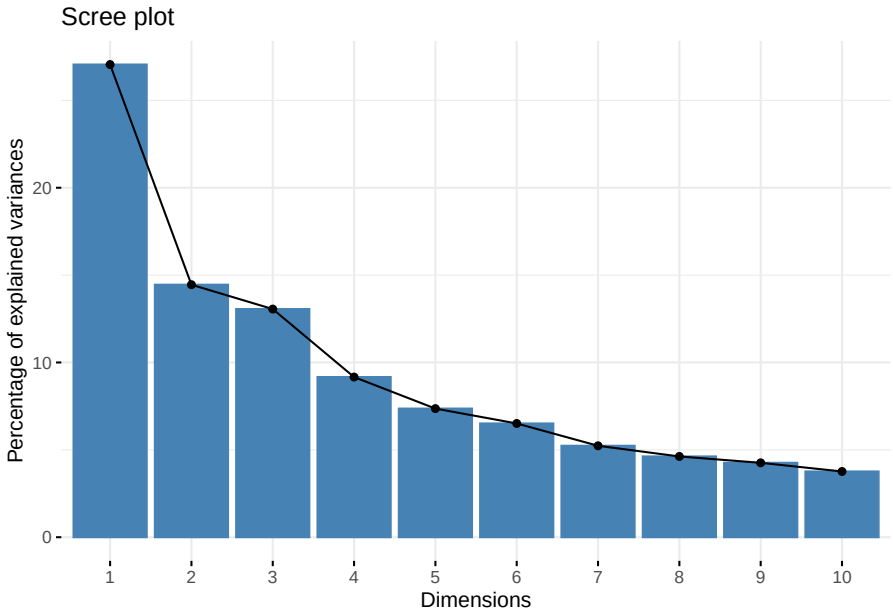
Summary of PCA Analysis

```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.8752 1.3706 1.3028 1.09159 0.97825 0.9200 0.82481
## Proportion of Variance 0.2705 0.1445 0.1306 0.09166 0.07361 0.0651 0.05233
## Cumulative Proportion  0.2705 0.4150 0.5456 0.63723 0.71084 0.7759 0.82827
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.77515 0.74401 0.69942 0.52956 0.40650 0.37841
## Proportion of Variance 0.04622 0.04258 0.03763 0.02157 0.01271 0.01102
## Cumulative Proportion  0.87449 0.91707 0.95470 0.97627 0.98898 1.00000
```

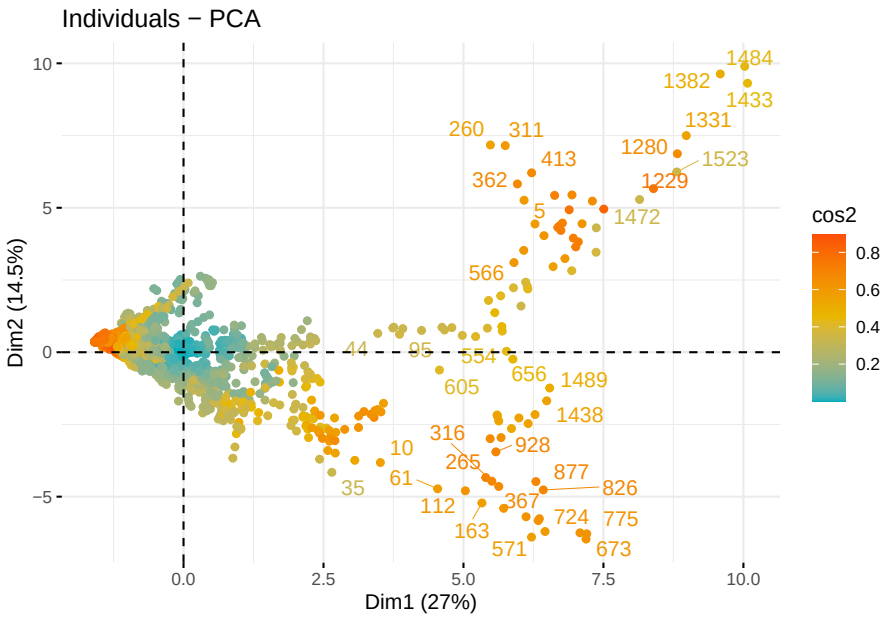
PCA Plots



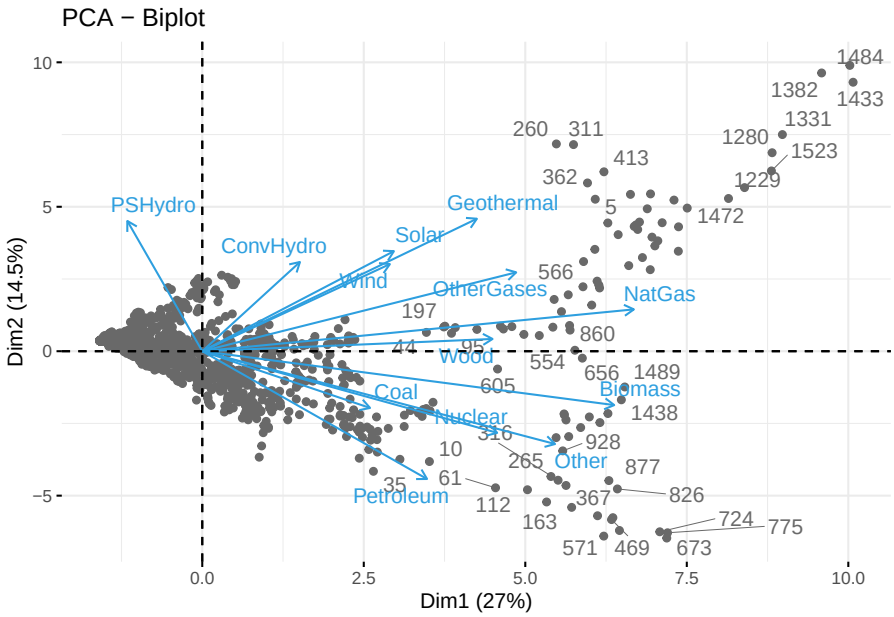
BarPlot with percentage of variances explained by each principal component.



Graph of Sources. Sources with a similar profile are grouped together.



Biplot of sources and variables



3.4 Split of Data

In this project the data is split balanced 80% for training and 20% for testing. The createDataPartition() function in the caret package attempts to balance the class distributions within the splits.

To make the training and test datasets, first set a seed. The advantage of setting a seed is that it provides a reproducible sequence of random numbers in the random number generator.

```
# set random seed to make reproducible results
set.seed(2020, sample.kind="Rounding")

# split the data 80% training and 20% testing
```

```
split_index <- createDataPartition(y = wide_db$Level, times = 1, p = 0.2, list = FALSE)
train_db <- wide_db[-split_index,]
test_db <- wide_db[split_index,]

# little house-keeping
rm(split_index )
```

Check the split of the train and test dataset by passing the dataset into the dim() function.

```
## [1] 1223   22
```

```
## [1] 307  22
```

3.5 Linear Regression

Linear regression will be used to predict an outcome variable based on one or more predictor variables. Model selection is the first step that is accomplished by the following code:

```
# Let's look for the best model
lm_model = step(glm(train_db$LevelNum ~ ., data = train_db %>%
    dplyr::select(Biomass, ConvHydro, Solar, Wind, Wood),
    family=binomial(link='logit')),
    direction = "both")
```

```
## Start:  AIC=296.48
## train_db$LevelNum ~ Biomass + ConvHydro + Solar + Wind + Wood
##
##             Df Deviance    AIC
## - Wind       1   284.49 294.49
## - Solar      1   284.50 294.50
## - Wood       1   286.44 296.44
## <none>           284.48 296.48
## - Biomass    1   311.49 321.49
## - ConvHydro  1   449.81 459.81
##
## Step:  AIC=294.49
## train_db$LevelNum ~ Biomass + ConvHydro + Solar + Wood
##
##             Df Deviance    AIC
## - Solar      1   284.51 292.51
## - Wood       1   286.45 294.45
## <none>           284.49 294.49
## + Wind       1   284.48 296.48
## - Biomass    1   312.91 320.91
## - ConvHydro  1   451.42 459.42
##
## Step:  AIC=292.51
## train_db$LevelNum ~ Biomass + ConvHydro + Wood
##
##             Df Deviance    AIC
## - Wood       1   286.45 292.45
## <none>           284.51 292.51
## + Solar      1   284.49 294.49
## + Wind       1   284.50 294.50
## - Biomass    1   313.74 319.74
## - ConvHydro  1   452.35 458.35
##
```

```
## Step:  AIC=292.45
## train_db$LevelNum ~ Biomass + ConvHydro
##
##             Df Deviance    AIC
## <none>           286.45 292.45
## + Wood       1   284.51 292.51
## + Wind       1   286.44 294.44
## + Solar      1   286.45 294.45
## - Biomass    1   327.55 331.55
## - ConvHydro  1   453.48 457.48
```

Code for stepwise evaluation of various modes using glm (Generalized Linear Model) The glm package, or generalized linear model, is a collection of models when used with the step function allows stepwise evaluation of the various models.

3.5.1 Model comparison with GLM using stepwise evaluation

The models are compared with various variables added or removed using stepwise and subset selection to determine the best model. The "goodness of fit" can be included as an argument. The Akaike Information criterion (AIC) is typically used for selection, using the Bayesian information criterion (BIC) usually results in more parsimonious model.

3.5.2 Code for Linear Regression Model

The model with the best score included Conventional Hydro and Biomass only. The models for the training and test dataset were updated based on the selected model. After the model was run against the training and test dataset a statistical summary explains the goodness of fit.

```
# build model from train data
lm_train = lm(formula = train_db$LevelNum ~ Biomass + ConvHydro + Wood, data = train_db)

# build model from test data
lm_test = lm(formula = test_db$LevelNum ~ Biomass + ConvHydro + Wood, data = test_db)
```

Summary statistics for the Train Linear Model

```
##
## Call:
## lm(formula = train_db$LevelNum ~ Biomass + ConvHydro + Wood,
##     data = train_db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70369 -0.03959 -0.02626  0.00323  0.97874
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.681e-02  6.601e-03   4.062 5.18e-05 ***
## Biomass     -5.262e-08  8.435e-09  -6.238 6.10e-10 ***
## ConvHydro    9.040e-09  3.811e-10  23.721  < 2e-16 ***
## Wood        -1.448e-08  5.060e-09  -2.862  0.00428 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1767 on 1219 degrees of freedom
## Multiple R-squared:  0.3226, Adjusted R-squared:  0.3209
## F-statistic: 193.5 on 3 and 1219 DF,  p-value: < 2.2e-16
```

Summary statistics for the Test Linear Model

```
##
## Call:
## lm(formula = test_db$LevelNum ~ Biomass + ConvHydro + Wood, data = test_db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44038 -0.04751 -0.03105  0.00404  0.89557
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.246e-02  1.401e-02   2.317 0.021181 *
## Biomass     -6.430e-08  1.828e-08  -3.518 0.000502 ***
## ConvHydro    1.134e-08  1.051e-09  10.790  < 2e-16 ***
## Wood        -1.861e-08  1.133e-08  -1.642 0.101664
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1832 on 303 degrees of freedom
## Multiple R-squared:  0.2868, Adjusted R-squared:  0.2798
## F-statistic: 40.62 on 3 and 303 DF,  p-value: < 2.2e-16
```

Anova comparison of Train and Test

```
##
## ================================================
##             Df   Sum Sq Mean Sq F value Pr(> F)
## ------------------------------------------------
## Biomass      1   0.497   0.497  15.918  0.0001
## ConvHydro    1   17.361 17.361  556.342    0
## Wood         1   0.256   0.256   8.192   0.004
## Residuals 1,219 38.040  0.031
## ------------------------------------------------
```

3.6 Decision Tree

For this project the split data that was used with the Linear Regression model is also used for the Decision Tree. The training set is used to build the models and is provided the outcome for each observation. The model is based on the "features": Level, Biomass, and ConvHydro. The test set will be used to see how well the model performs on unseen data. For each observation in the test set, the trained model is used to predict the outcome in the test and scored for accuracy. One of the advantages of decision trees is that they are intuitively very easy to visually plot and explain and closely mirrors a human decision-making approach.

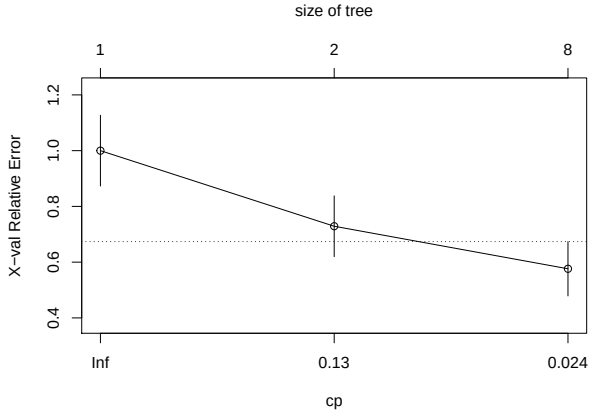
3.6.1 R code for decision tree visualization and summary

Initially grow and visualize the tree. Provide a statistical summary of the tree with confusion matrix and then test accuracy of the two datasets.

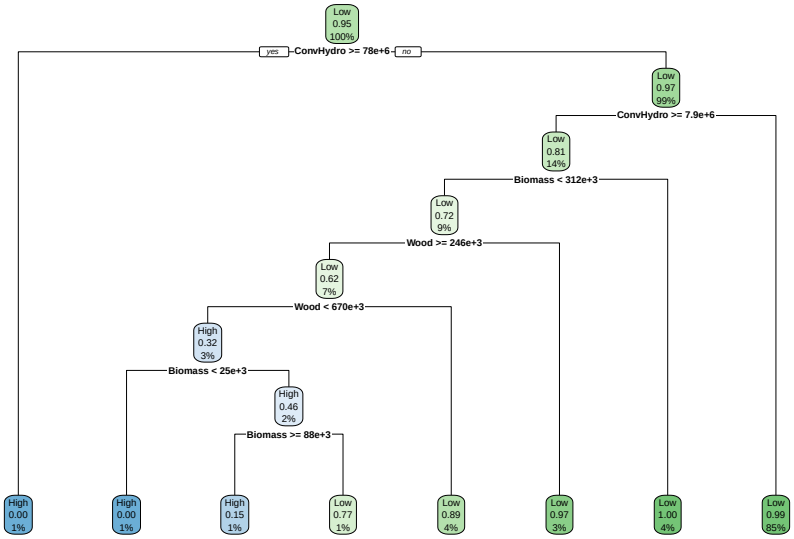
```
# grow tree / build model
# use 'class' method because we are predicting a class
dt_fit <- rpart(Level ~ ConvHydro + Biomass + Wood - Year,
    method="class", data=train_db)

#display the results
printcp(dt_fit)
```

```
##
## Classification tree:
## rpart(formula = Level ~ ConvHydro + Biomass + Wood - Year, data = train_db,
##     method = "class")
##
## Variables actually used in tree construction:
## [1] Biomass   ConvHydro Wood
##
## Root node error: 59/1223 = 0.048242
##
## n= 1223
##
##         CP nsplit rel error  xerror     xstd
## 1 0.288136      0   1.00000 1.00000 0.127010
## 2 0.059322      1   0.71186 0.72881 0.109172
## 3 0.010000      7   0.35593 0.57627 0.097446
```

Decision Tree cross Validation



Decision Tree of the "Renewable" sources



Summary of the decision tree

The summary lists the variables in their order of importance. The decision tree model is shown in the following visualization.

```
## Call:
## rpart(formula = Level ~ ConvHydro + Biomass + Wood - Year, data = train_db,
##     method = "class")
##   n= 1223
##
##           CP nsplit rel error    xerror       xstd
## 1 0.28813559      0 1.0000000 1.0000000 0.12700981
## 2 0.05932203      1 0.7118644 0.7288136 0.10917168
## 3 0.01000000      7 0.3559322 0.5762712 0.09744625
##
## Variable importance
## ConvHydro      Wood   Biomass
##        53        24        23
##
## Node number 1: 1223 observations,    complexity param=0.2881356
##   predicted class=Low   expected loss=0.04824203  P(node) =1
##     class counts:    59  1164
##    probabilities: 0.048 0.952
##   left son=2 (17 obs) right son=3 (1206 obs)
##   Primary splits:
##       ConvHydro < 77895920 to the right, improve=31.232810, (0 missing)
##       Wood      < 348116.5 to the right, improve= 5.082949, (0 missing)
##       Biomass   < 328402.5 to the left,  improve= 2.440618, (0 missing)
##
## Node number 2: 17 observations
##   predicted class=High  expected loss=0  P(node) =0.01390025
##     class counts:    17     0
##    probabilities: 1.000 0.000
##
## Node number 3: 1206 observations,    complexity param=0.05932203
##   predicted class=Low   expected loss=0.03482587  P(node) =0.9860998
##     class counts:    42  1164
##    probabilities: 0.035 0.965
##   left son=6 (169 obs) right son=7 (1037 obs)
##   Primary splits:
##       ConvHydro < 7915654  to the right, improve=9.385834, (0 missing)
##       Wood      < 348116.5 to the right, improve=2.402893, (0 missing)
##       Biomass   < 299528.5 to the left,  improve=1.377066, (0 missing)
##   Surrogate splits:
##       Biomass < 2106619  to the right, agree=0.865, adj=0.036, (0 split)
##       Wood    < 3308976  to the right, agree=0.864, adj=0.030, (0 split)
##
## Node number 6: 169 observations,    complexity param=0.05932203
##   predicted class=Low   expected loss=0.1893491  P(node) =0.1381848
##     class counts:    32   137
##    probabilities: 0.189 0.811
##   left son=12 (114 obs) right son=13 (55 obs)
##   Primary splits:
##       Biomass   < 311699.5 to the left,  improve=5.846569, (0 missing)
##       Wood      < 639870.5 to the left,  improve=4.216971, (0 missing)
```

```
##       ConvHydro < 9033741  to the left,  improve=3.686657, (0 missing)
##   Surrogate splits:
##       ConvHydro < 15603480 to the left,  agree=0.828, adj=0.473, (0 split)
##       Wood      < 2896343  to the left,  agree=0.746, adj=0.218, (0 split)
##
## Node number 7: 1037 observations
##   predicted class=Low   expected loss=0.009643202  P(node) =0.847915
##     class counts:    10  1027
##    probabilities: 0.010 0.990
##
## Node number 12: 114 observations,    complexity param=0.05932203
##   predicted class=Low   expected loss=0.2807018  P(node) =0.09321341
##     class counts:    32    82
##    probabilities: 0.281 0.719
##   left son=24 (82 obs) right son=25 (32 obs)
##   Primary splits:
##       Wood      < 246136.5 to the right, improve=5.536612, (0 missing)
##       Biomass   < 90027    to the right, improve=4.719891, (0 missing)
##       ConvHydro < 34109100 to the right, improve=3.518066, (0 missing)
##   Surrogate splits:
##       Biomass   < 1191.5   to the right, agree=0.912, adj=0.687, (0 split)
##       ConvHydro < 8292040  to the right, agree=0.728, adj=0.031, (0 split)
##
## Node number 13: 55 observations
##   predicted class=Low   expected loss=0  P(node) =0.04497138
##     class counts:     0    55
##    probabilities: 0.000 1.000
##
## Node number 24: 82 observations,    complexity param=0.05932203
##   predicted class=Low   expected loss=0.3780488  P(node) =0.06704824
##     class counts:    31    51
##    probabilities: 0.378 0.622
##   left son=48 (38 obs) right son=49 (44 obs)
##   Primary splits:
##       Wood      < 669799   to the left,  improve=13.276290, (0 missing)
##       Biomass   < 1191.5   to the left,  improve= 5.920976, (0 missing)
##       ConvHydro < 34109100 to the right, improve= 1.564201, (0 missing)
##   Surrogate splits:
##       Biomass   < 1191.5   to the left,  agree=0.622, adj=0.184, (0 split)
##       ConvHydro < 9033741  to the left,  agree=0.610, adj=0.158, (0 split)
##
## Node number 25: 32 observations
##   predicted class=Low   expected loss=0.03125  P(node) =0.02616517
##     class counts:     1    31
##    probabilities: 0.031 0.969
##
## Node number 48: 38 observations,    complexity param=0.05932203
##   predicted class=High  expected loss=0.3157895  P(node) =0.03107114
##     class counts:    26    12
##    probabilities: 0.684 0.316
##   left son=96 (12 obs) right son=97 (26 obs)
##   Primary splits:
##       Biomass   < 24670.5  to the left,  improve=3.497976, (0 missing)
##       Wood      < 441136.5 to the right, improve=1.190283, (0 missing)
```

```
##       ConvHydro < 9522160  to the left,  improve=0.780027, (0 missing)
##   Surrogate splits:
##       ConvHydro < 8750950  to the left,  agree=0.789, adj=0.333, (0 split)
##       Wood      < 461740   to the right, agree=0.711, adj=0.083, (0 split)
##
## Node number 49: 44 observations
##   predicted class=Low   expected loss=0.1136364  P(node) =0.03597711
##     class counts:     5    39
##    probabilities: 0.114 0.886
##
## Node number 96: 12 observations
##   predicted class=High  expected loss=0  P(node) =0.009811938
##     class counts:    12     0
##    probabilities: 1.000 0.000
##
## Node number 97: 26 observations,    complexity param=0.05932203
##   predicted class=High  expected loss=0.4615385  P(node) =0.0212592
##     class counts:    14    12
##    probabilities: 0.538 0.462
##   left son=194 (13 obs) right son=195 (13 obs)
##   Primary splits:
##       Biomass   < 88355.5  to the right, improve=4.923077, (0 missing)
##       ConvHydro < 33831750 to the right, improve=2.617521, (0 missing)
##       Wood      < 422873.5 to the left,  improve=1.034188, (0 missing)
##   Surrogate splits:
##       Wood      < 422873.5 to the left,  agree=0.654, adj=0.308, (0 split)
##       ConvHydro < 9522160  to the left,  agree=0.615, adj=0.231, (0 split)
##
## Node number 194: 13 observations
##   predicted class=High  expected loss=0.1538462  P(node) =0.0106296
##     class counts:    11     2
##    probabilities: 0.846 0.154
##
## Node number 195: 13 observations
##   predicted class=Low   expected loss=0.2307692  P(node) =0.0106296
##     class counts:     3    10
##    probabilities: 0.231 0.769
```

Conventional Hydroelectric, Biomass, Wood (Wood and Wood Derived Fuels) and Wind are selected in the Decision tree. Both Train and Test exhibited good accuracy with the data.

3.6.2 Confusion Matrix and Statistics for Train dataset

```
#Predict for train_db
train_pred <- predict(dt_fit, type = "class", newdata = train_db)
train_tree <- table(train_db$Level, train_pred)

# Display confusionmatrix for train
confusionMatrix(train_tree)
```

```
## Confusion Matrix and Statistics
##
##       train_pred
##        High  Low
##   High   40   19
```

```
##   Low     2 1162
##
##                Accuracy : 0.9828
##                  95% CI : (0.9739, 0.9893)
##     No Information Rate : 0.9657
##     P-Value [Acc > NIR] : 0.0002184
##
##                   Kappa : 0.7834
##
##  Mcnemar's Test P-Value : 0.0004803
##
##             Sensitivity : 0.95238
##             Specificity : 0.98391
##          Pos Pred Value : 0.67797
##          Neg Pred Value : 0.99828
##              Prevalence : 0.03434
##          Detection Rate : 0.03271
##    Detection Prevalence : 0.04824
##       Balanced Accuracy : 0.96815
##
##        'Positive' Class : High
##
```

3.6.3 Confusion Matrix and Statistics for Test dataset

```
## Confusion Matrix and Statistics
##
##       test_pred
##        High Low
##   High   11   4
##   Low     1 291
##
##                Accuracy : 0.9837
##                  95% CI : (0.9624, 0.9947)
##     No Information Rate : 0.9609
##     P-Value [Acc > NIR] : 0.01862
##
##                   Kappa : 0.8064
##
##  Mcnemar's Test P-Value : 0.37109
##
##             Sensitivity : 0.91667
##             Specificity : 0.98644
##          Pos Pred Value : 0.73333
##          Neg Pred Value : 0.99658
##              Prevalence : 0.03909
##          Detection Rate : 0.03583
##    Detection Prevalence : 0.04886
##       Balanced Accuracy : 0.95155
##
##        'Positive' Class : High
##
```

3.6.4 Accuracy of Decision Tree Train and Test dataset

```
# calculate train Accuracy and save to data.frame
train_accuracy <- table(Predicted = train_pred, Actual = train_db$Level)
dtree_train_accuracy <- sum(diag(train_accuracy))/sum(train_accuracy)
results <- data.frame(model="train Data", Accuracy=dtree_train_accuracy)

# calculate test Accuracy and save to data.frame
test_accuracy <- table(Predicted = test_pred, Actual = test_db$Level)
dtree_test_accuracy <- sum(diag(test_accuracy))/sum(test_accuracy)
results <- results %>% add_row(model="Test Data", Accuracy=dtree_test_accuracy)
```

Accuracy is the proportion of true positive and true negative divided by the sum of the matrix. The formula for calucating accuracy:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 1: Decision Tree Accuracy

model	Accuracy
train Data	0.9828
Test Data	0.9837

4 Discussion

Looking at the energy sources identified by the Linear Regression Model, the Decision Tree and Principal Component Analysis Conventional Hydroelectric was identified by all three. Biomass was identified for use in the Linear Regression model and Decision Tree. Principal component analysis identified Conventional Hydroelectric, Solar and Geothermal as explaining the most variability in the data. While this study looks at the renewable generation from a national view, there are a significant number of states that renewable energy makes up less than 10% of their total generation. When reviewing the ratios, simply because renewable may be 70% of their total electric produced that doesn't necessarly mean they can produce the amount that is necessary when demand is high. This study focused on aggregate total numbers and not the detailed daily, weekly, and seasonal high and low numbers for generation and consumption demands.

Table 2: Renewable Top Ratios

Year	State	Ratio
2019	VT	1.00
2018	VT	1.00
2015	VT	1.00
2016	VT	1.00
2017	VT	1.00
1990	ID	0.98

The Ratio of electrial energy produced (Renewable/ Total) in descending order another interesting way to look at the data. Looking at the ratio in descending order to determine the states that are producing a significant amount of electrical power using renewable sources.

Table 3: Renewable Top States

State
VT
ID
OR
WA
SD
ME

Reviewing the ratio of electricity generation (Total/Renewable Sources) for 2019 the top "Renewable" states Vermont (99%), Idaho (76.3%), Washington (69.7%), and South Dakota (73.8%) they are producing a large amount of their total energy with renewable.

5 Limitations and Delimitations

Limitations include time constraints that affect statistical methodologies selected for the analysis as well as the granularity, completeness, and accuracy of the data as it may have been imputed or estimated (EIA, 2018). This project looks only at the electricity generation part of quote by Mark Jacobson "fuel as well as electricity" (Jacobson, M.Z., & Delucchi, M.A., 2009). This project looks at the aggregated annual data which does not have the granularity to analyze the daily, monthly and seasonal peak demands. Also, this project does not look at any of the financial costs for development or land area needed to accommodate such sources of energy (Lyman,2016). Electric generation technologies not currently tracked by EIA include wave devices, tidal turbines, and residential and commercial roof mounted solar panels. Regression projections of necessary available (generated) energy and sales (consumption) will be based on current consumption to generation ratios.

Delimitations include from the available EIA data only annual aggregates for the "Total Electric Power Industry", as type of producer, is used and all "energy source" variables were selected. "US Totals" were not included as this study is looking at individual states. The statistical analysis and focus of research are renewable energy sources: Biomass, Conventional Hydroelectric, Solar, Wind, and Wood and Wood derived Fuels. Biomass includes municipal solid waste from biogenic sources, landfill gas, sludge waste, agricultural byproducts, and other biomass. From 1990 through 2000, biomass also included non-renewable waste (municipal solid waste from non-biogenic sources, and tire-derived fuels). From 1990 through 1989, hydroelectric pumped storage is included in "Conventional Hydroelectric Power."Solar - Electricity net generation from solar thermal and photovoltaic (PV) energy at utility-scale facilities and does not include distributed (small-scale) solar photovoltaic generation. Other Gases includes blast furnace gas, and other manufactured and waste gases derived from fossil fuels. From 1990 through 2010, other-gases also included propane gas. Petroleum includes distillate fuel oil, residual fuel oil, petroleum coke, jet fuel, kerosene, other petroleum, waste oil, and, beginning in 2011, propane. Hydro-electric Pumped Storage is calculated as pumped storage facility production minus energy used for pumping ("Monthly Energy Review", 2019).

6 Conclusions and Future Study

In 2019, electric generation by renewable sources such as hydropower, wind, solar and geothermal made up about 17.9% of the electricity powering our nation, up from 9.3% in 2000. Looking at individual states in 2019, 21 states produced less than 10% of their total using renewable energy sources. In 2019, six states renewable sources produced over 60% of their total generation. While this may appear to be significant, these are total numbers and don't reflect the daily, weekly and seasonal changes in demand. Statistically it may imply that it's possible that a few states may be able to produce enough over the year but may not be able to produce enough on-demand to meet their needs. A study with more fine grained data that would allow the study of generation and demands cycles would be necessary for a more conclusive outcome.

7 References

- Jacobson, M.Z., &Delucchi, M.A. (2009, November). Retrieved December 20, 2020, from https://www.scientificamerican.com/article/a-path-to-sustainable-energy-by-2030/
- Fischetti, M. (2013, April 15). How to Power the World without Fossil Fuels. Retrieved December 20, 2020, from https://www.scientificamerican.com/article/how-to-power-the-world/
- Pulkkinen, L. (2019, May 14). Why Washington Is the Best State in America. Retrieved December 22, 2020, from https://www.usnews.com/news/best-states/articles/2019-05-14/amazon-clean-power-help-fuel-washington
- Simon, R.M., & Hayes, D.J. (2017, June 29). America's Clean Energy Success, by the Numbers. Retrieved December 22, 2020, from https://www.americanprogress.org/issues/green/reports/2017/06/29/435281/americas-clean-energy-success-numbers/
- Huetteman, T., & Martin, L. (2016, June 17). Clean Power Plan accelerates the growth of renewable generation throughout United States. Retrieved December 22, 2020, from https://www.eia.gov/todayinenergy/detail.php?id=26712
- Chevalier, Z. (2018, July 23). These States Use the Most Renewable Energy. December 26, 2020, from https://www.usnews.com/news/best-states/slideshows/these-states-use-the-most-renewable-energy
- Levin, A. (2018, February 28). 2017 Clean Energy By the Numbers: A State-by-State Look. Retrieved December 26, 2020, from https://www.nrdc.org/experts/amanda-levin/2017-clean-energy-by-the-numbers-a-state-by-state-look
- Lyman, Robert (2016, May) Why Renewable Energy Cannot Replace Fossil Fuels by 2050. Retrieved December 27, 2020, from https://friendsofscience.org/assets/documents/Renewable-energy-cannot-replace-FF_Lyman.pdf
- EIA, US Energy Information Administration (2018, March). A Guide to EIA Electric Power Data, Retrieved December 28, 2020, from https://www.eia.gov/electricity/data/guide/pdf/guide.pdf
- (2020, June 20) Renewable Energy Explained. What is renewable energy? Retrieved December 28, 2020, from https://www.eia.gov/energyexplained/?page=renewable_home
- (n.d.) Nonrenewable Energy Explained. Nonrenewable energy sources. Retrieved December 28, 2020, from https://www.eia.gov/energyexplained/index.php?page=nonrenewable_home
- (n.d.) Glossary. Pumped-storage hydroelectric plant. Retrieved December 28, 2020, from https://www.eia.gov/tools/glossary/index.php?id=Pumped-storage%20hydroelectric%20plant
- (Dec, 2020) Monthly Energy Review, Electricity Net Generation: Electric Power Sector (Table 7.2b) Retrieved December 28, 2020, from https://www.eia.gov/totalenergy/data/monthly/pdf/sec7.pdf
- Silverman, Adam (2017, April 30) Vermont ranks No. 2 in US for renewable energy. Retrieved December 30, 2020, from https://www.burlingtonfreepress.com/story/news/2017/04/30/vermont-ranks-high-for-renewable-energy/100955606/

8 Appendix A

The files that can be downloaded are Excel(xlsx) files they need a little manual cleaning and to be converted to csv. The following directions should help in preparing the files.

For the Annual Sales File: URL: https://www.eia.gov/electricity/data/state/sales_annual.xlsx

Row 1 needs to be deleted: 1. Select Row 1 and then left-click and in the context menu select delete 2. Verify the row has been deleted

To save the Excel(xlsx) file in csv format: 1. From the Menu select “File” then scroll down to “Save As” 2. Change the file type to "CSV UTF-8 (Comma delimited) (*.csv) 3. Press “Save”

For the Annual Generation by State file: URL: https://www.eia.gov/electricity/data/state/annual_generation_state.xlsx

Row 1 needs to be deleted: 1. Select Row 1 and then left-click and in the context menu select delete 2. Verify the row has been deleted

There are 3 rows of 2003 that need to be deleted - they have no state or values assigned: 1. Select Year column, the press ctrl+F to open find dialog. 2. Enter 2003 in the “Find What” drop down and press “Find Next” *Note: In the file I downloaded they are rows: 20578, 20579, and 20580* 3. Select the three rows and then left-click and in the context menu select delete 4. Verify the three rows have been deleted

To save the Excel(xlsx) file in csv format: 1. From the Menu select “File” then scroll down to “Save As” 2. Change the file type to "CSV UTF-8 (Comma delimited) (*.csv) 3. Press “Save”

9 Appendix B

Renewable Energy Sources (the major sources)

- Biomass (includes Wood and Wood Derived Fuels)
- Geothermal
- Hydropower
- Wind
- Solar

Non-Renewable Energy Sources

Fossil Fuels - Coal, crude oil, and natural gas are all considered fossil fuels because they were formed from the buried remains of plants and animals that lived millions of years ago.

- Uranium (Nuclear Energy) – uranium is not a fossil fuel but is classified as a nonrenewable fuel.

For further explanation of the various fuel sources:

- Biomass includes biogenic municipal solid waste, landfill gas, sludge waste, agricultural byproducts, other biomass solids, other biomass liquids, and other biomass gases (including digester gases and methane).
- Solar - Electricity net generation from solar thermal and photovoltaic (PV) energy atutility-scale facilities. Does not include distributed (small-scale) solar photovoltaicgeneration.
- Wood and Wood Derived Fuels includes paper pellets, railroad ties, utility poles, wood chips, bark, red liquor, sludge wood, spent sulfite liquor, and black liquor, with other wood waste solids and wood-based liquids.
- Other Gases includes blast furnace gas, propane gas, and other manufactured and waste gases derived from fossil fuels.
- Petroleum includes distillate fuel oil (all diesel and No. 1, No. 2, and No. 4 fuel oils), residual fuel oil (No. 5 and No. 6 fuel oils and bunker C fuel oil), jet fuel, kerosene, petroleum coke, and waste oil.
- Other includes non-biogenic municipal solid waste, batteries, chemicals, hydrogen, pitch, purchased steam, sulfur, tire-derived fuels, and miscellaneous technologies.