6000 level Data Analytics Project

Introduction

Motivation

Up to 2016, 13% of total population in the world do not have access to electricity. I was born in urban area, and electricity certainly impacts my life. I could not image how to live a life without electricity, which drives me to find out which part of the population do not have access to electricity. And, in the recent year, human being is focusing more on the environmental problems. One term that interests me is "Clean Energy". I start to think and research which countries have start to develop clean energy, and if those countries with more international investment perform better at developing clean energy technologies.

Initial Hypothesis

1. There is a positive relationship between the amount of international investment and access to electricity, clean energy, or renewable energy.
2. There is a positive relationship between the access to electricity and portion of renewable energy used in a country.
3. Access to electricity is a determent factor for any countries to create and output more energy or clean energy.

Related Work

Global patterns of renewable energy innovation, 1990–2009 (1)

Review of renewable energy investment and financing in China: Status, mode, issues and countermeasures (2)

Data Description and Exploratory Data Analytics

The data we used in this project comes from SDG database. Specifically, the data are pulled from Goal 7 and Goal 12. In total, we downloaded 4 data sheets in the format of xlsx. Each of the excel sheet contains information of one SDG indicator, such as access to electricity and portion of clean energy used. All 4 dataset are merged by GeoAreaName.

1.

Name: EG_ACS_ELEC.xlsx

Description: Proportion of population with access to electricity, by urban/rural (%)

Column:

SDG Goal: an indicator for which goal is the data addressing, such as 7.1

GeoAreaName: the name of the country, such as Algeria

Location: this column indicates rural/urban or total

Unit: Percentage %

Data: data point from year 2000 to 2020

2.

Name: EG_EGY_RNEW.xlsx

Description: Proportion of population with primary reliance on clean fuels and technology (%)

Column:

SDG Goal: an indicator for which goal is the data addressing, such as 7.1

GeoAreaName: the name of the country, such as Algeria

Unit: Percentage %

Data: data point from year 2000 to 2020

3.

Name: EG_FEC_RNEW.xlsx

Description: Renewable energy share in the total final energy consumption

Column:

SDG Goal: an indicator for which goal is the data addressing, such as 7.1

GeoAreaName: the name of the country, such as Algeria

Unit: Percentage %

Data: data point from year 2000 to 2020

4.

Name: EG_IFF_RANDN.xlsx

Description: International financial flows to developing countries in support of clean energy research and development and renewable energy production, including in hybrid systems

Column:

SDG Goal: an indicator for which goal is the data addressing, such as 7.1

GeoAreaName: the name of the country, such as Algeria

Location: this column indicates rural/urban or total

Unit: millions of constant United States dollars

Data: data point from year 2000 to 2017

Analysis

Before we start to analyze the data, we must do some data cleaning work to make sure that the data we work on are correct.

The very first step is to load the correct data from excel sheets. Each of the excel sheets contain not only the data but also the explanation for the SDG goal, Reporting type and Unit. When importing the data, we need to make sure we are importing data from sheet 2.

Then, we find that some data has value "N" which is NA in R. We need to replace "N" with NA.

Then we would merge our dataset by GeoAreaName, and we are only interested in four columns, which are RNEW, ACS, FEC, and IFF. Therefore, only these 4 columns are taken from the merge result.

ACS means Proportion of population with access to electricity, by urban/rural (%)

RNEW means Proportion of population with primary reliance on clean fuels and technology (%)

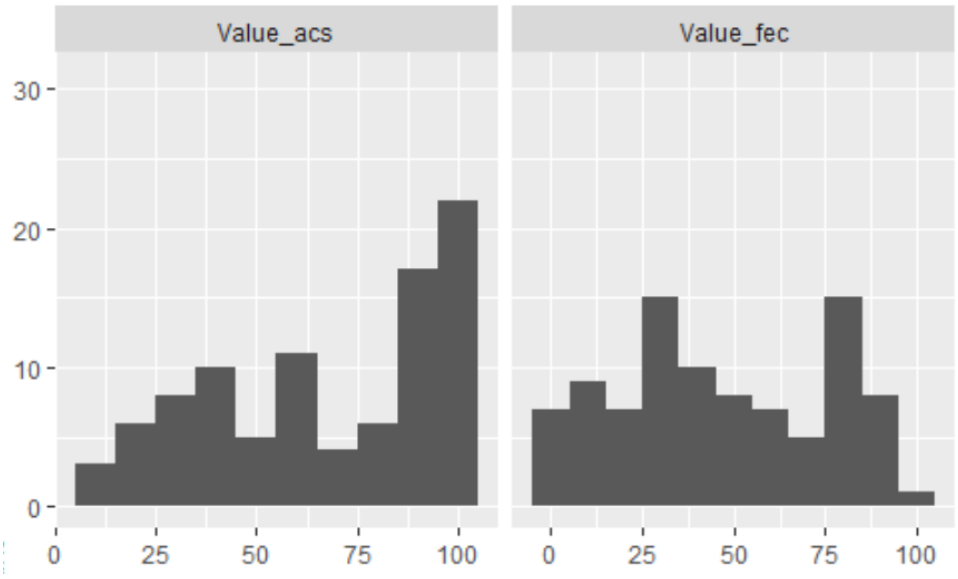
FEC means Renewable energy share in the total final energy consumption

IFF means International financial flows to developing countries in support of clean energy

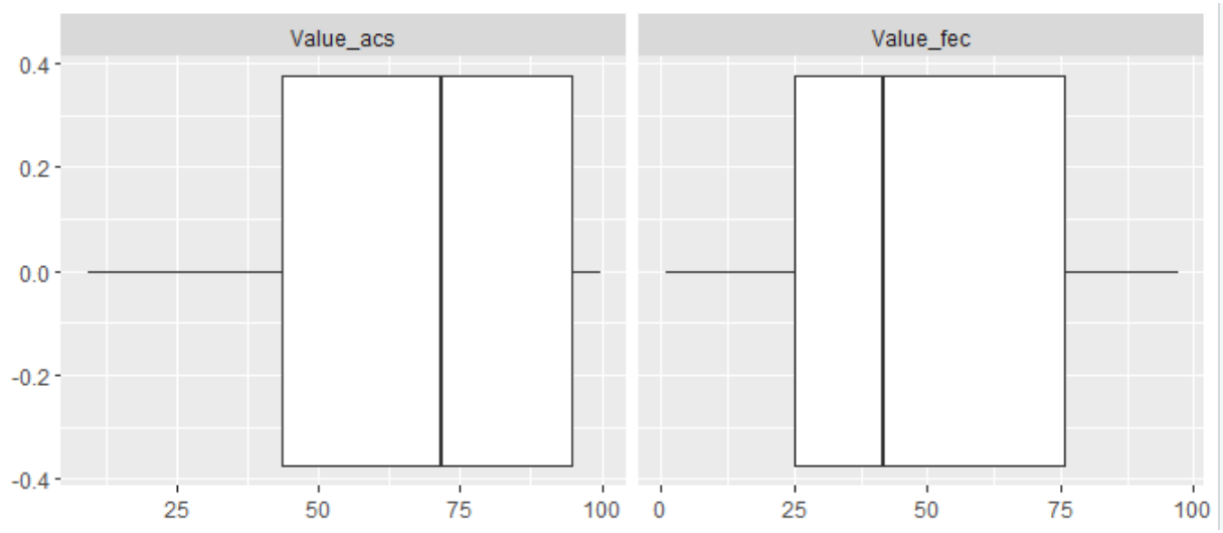
After we have obtained the data column that we need, we perform another check on the merged dataset to test if NA value is found using **any(is.na(data)).** The result is FALSE, indicating that all data records do not have any NA value.

Finally, we specify the data type for each column as numeric.

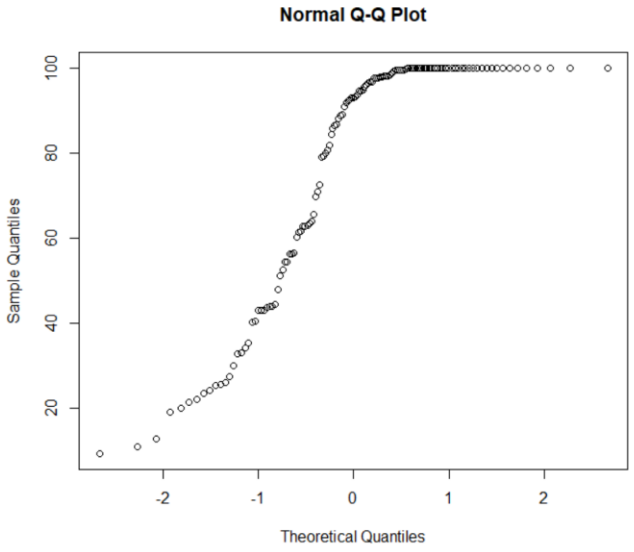
With the acquisition of desired data, we could finally start to analyze the dataset.



A histogram on access to electricity and portion of energy rely on renewable energy. We can see that most countries either have more than 80% access to electricity or less than 40%. Very few countries fall into the middle of the histogram. This suggests that the world we are living isverypolarized. And no clear pattern can be identified for the renewable energy share in the total final energy consumption.



A boxplot on access to electricity and portion of energy rely on renewable energy. We can find a difference here that, most country have around 40 – 95 % access to electricity. However, on average only 25 – 75% are relied on clean energy.



The qqnorm plot of access to electricity also suggests that it is not a normal distribution, and very likely it is an exponential distribution.

Summary function is also called to check if there is any eye-observable error in the dataset, such like a percentage that is more than 100%. We can see the highest access rate is 100% and the highest portion rely on renewable energy is 97.12% (In case you are wondering, it is Democratic Republic of the Congo).

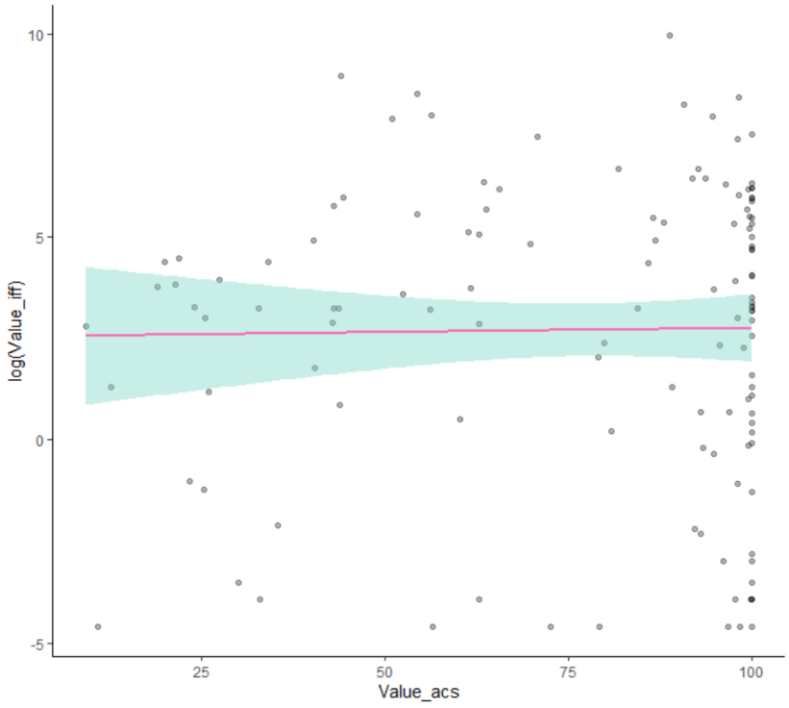
Model Development and Application of Models

1.

Linear Regression Model 1 is built to test hypothesis 1 if there is a relationship between international investment and access to electricity.

lm3 <- lm (log(d$Value_iff) ~ d$Value_acs)

A log transform is used on the investment variable because the value span from 0 to 42000, which is too larger and may influence the analysis result. Notice that, the data points on the right edge do not have value 100. Many of them are between 99 to 100. Data point with 100 percent access, are not used for this analysis since the goal is to test if the investment help countries improve the access rate.

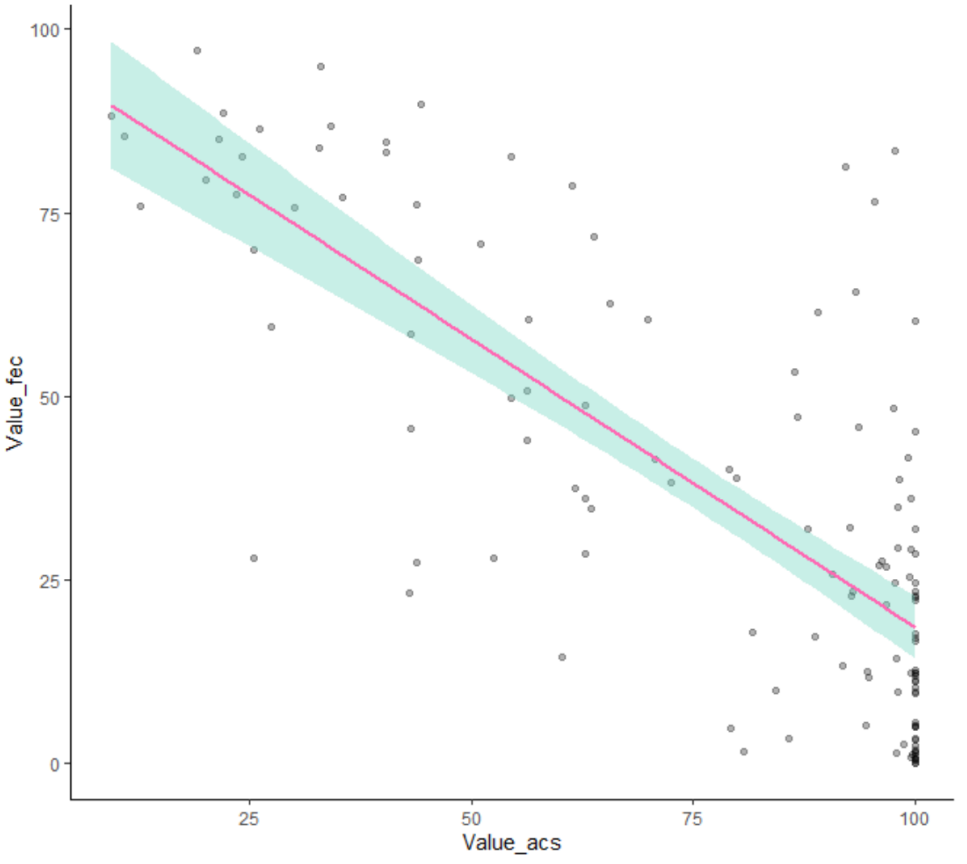


This is the result of the model we built. From a statistical perspective, the performance is very poor, with a p-value of 85.7%. Therefore, there we fail to reject the null hypothesis for hypothesis 1.

Linear Regression Model 2 is built to test hypothesis 2 that there is a positive relationship between access rate and portion rely on renewable energy.

lm2 <- lm(d$Value_acs ~ d$Value_fec, data = d)

Log transform is used for both variables in this case for proper scaling. Notice that the data points on the right edge includes data points with 100 access rates.




```
Call:
lm(formula = d$Value_acs ~ d$Value_fec, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-58.043  -9.800  -0.701  11.340  55.121

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 103.97645    2.61552   39.75   <2e-16 ***
d$Value_fec  -0.73583    0.05614  -13.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The statistical result of this model is pretty good. The p value is less than 0.001 for both intercept and the variable. Therefore, we can say that there is a negative relationship between access rate and portion of energy consumed in renewable energy.

The formula could be written down as

Y (portion rely on renewable energy) = 103.97 -0.736 x (access rate)

2.

The second model used is KNN modeling. KNN (K-Nearest Neighbor) is one of the simplest machine learning algorithms, which can be used for classification and regression, and is a supervised learning algorithm. The idea is that if most of the K most similar (ie, the nearest neighbors in the feature space) samples of a sample in the feature space belong to a certain category, the sample also belongs to this category. That is to say, this method only determines the category of the sample to be classified based on the category of the nearest one or several samples in the classification decision. In this project it is crucial to see if countries are clustered in term of energy investment or development.

To have correct result from the KNN Modeling, first of all we need to normalize the data. In this project, MinMax scaling is used to make sure that all variables have equal length of maximum 1.

```
normalize <- function(x){
  return((x-min(x)) / (max(x)-min(x)))
}
```

Then we normalize all variables in the dataset. Then to make sure the model is working correctly, we divide the dataset to two subsets for cross validation.

```
m <- dim(data.values)[1]

val <- sample(1:m, size = round(m/3), replace = FALSE,
              prob = rep(1/m, m))

#corss validation
data.learn <- data.values[-val,]
data.valid <- data.values[val,]
```

Then we build the model, and use the model to test how many are fitted for data.valid

```
kknnm <- kknn(value_iff ~., data.learn, data.valid, distance = 1,
              kernel = "triangular", label)
summary(kknnm)
fit <- round(as.integer(fitted(kknnm)), -2)
table(round(as.integer(data.valid$value_iff, data.valid$value_rnew),-2), fit)
```

```
   fit
     0
 0  43
```

The result is that all testing cases are fitted with this model, which means the KNN prediction accuracy is around 100%. This means that given the historical data of portion rely on the renewable energy, international investment and energy per capita, access rate of a country could be predicted.

3. The third model we use is random forest.

The model we build use all variables to predict the portion of energy rely on renewable energy.

To develop the model, we need to take a subset of the dataset as training dataset. Therefore, 50% of data was randomly selected for training purpose.

Random Forest (random forest) is an extended variant of Bagging. Based on the decision tree-based learner to build Bagging integration, it further introduces random feature selection in the training

process of the decision tree. Therefore, it can be summarized that RF includes four section, Random selection of samples (replacement sampling), Randomly select features, Build a decision tree and Random forest voting (average). In this case, the access to electricity and portion of energy rely on renewable energy is positively related. Using the bagging method can add randomness to our model can increase the accuracy of the model. The first model we build using bagging classifier look like this.

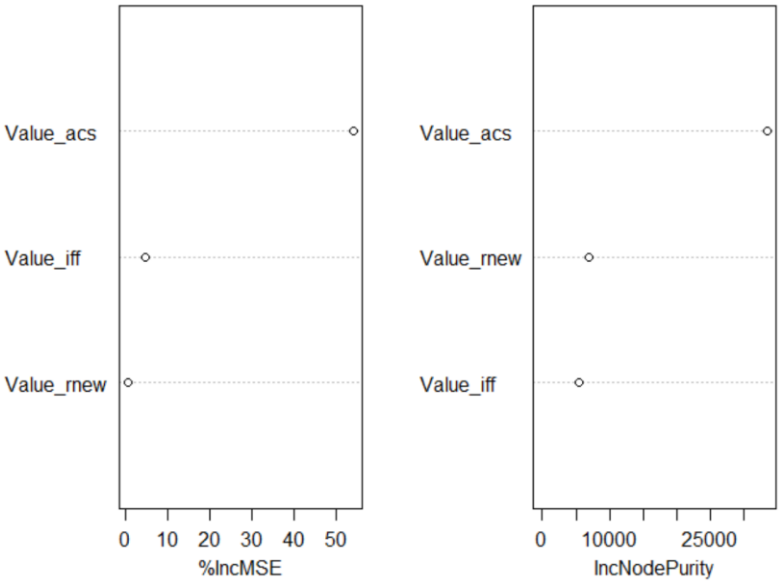
bag.data <- randomForest(Value_fec ~ ., data = d, subset = train, mtry=3,ntrees=500, importance=T)

mtry = 3 is used to select 3 dependent columns, which is all columns in this project. And the result is that

```
Call:
 randomForest(formula = Value_fec ~ ., data = d, mtry = 3, ntrees = 500,      importan
ce = T, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 354.5326
                    % Var explained: 52.6
> importance(bag.data) # importance of each predictor
             %IncMSE IncNodePurity
Value_rnew  0.5509058      6866.586
Value_acs  54.0796661     33500.510
Value_iff   4.7367226      5422.246
> varImpPlot(bag.data)
> yhat.bag=predict(bag.data, newdata = d[-train,])
> d.test=d[-train,"Value_fec"]
> mean((yhat.bag-d.test)^2)
[1] 430.0269
```

From the result, we can see that the access rate is the most important variable to influence the portion of energy rely on renewable energy with a significance of 54%. And the prediction error on column FEC is around 430.



This result plot visualizes the importance of each variables in this model. Again, we can tell that the access rate is the most influential factor.

Then we want to improve this model, we build another model with same parameter except for mtry value. Based on a empirical conclusion, we use square root of total number of columns, round 2, instead of 3.

rf.d<-randomForest(Value_fec ~ ., data=d, subset = train, mtry=2, ntrees=500, importance=T)

The result looks like this

```
Call:
 randomForest(formula = Value_fec ~ ., data = d, mtry = 2, ntrees = 500,      importan
ce = T, subset = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 359.1311
                    % Var explained: 51.99
```

By applying the random forest classifier, we can see that the prediction error of FEC of decreases from 430 to around 400, which is a improvement in the modeling.

```
> mean((yhat.rf-d.test2)^2)
[1] 399.6594
```

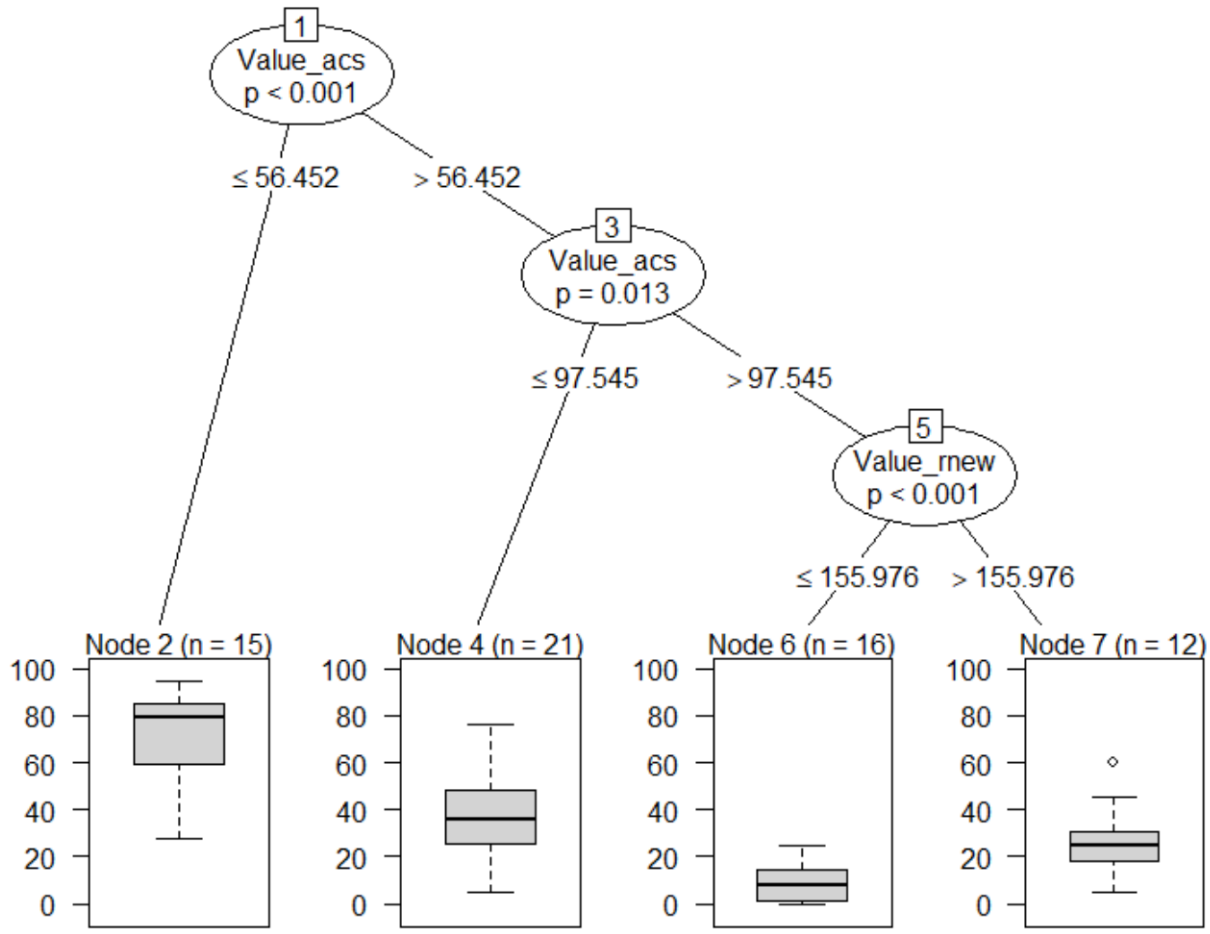
When we look at the importance of each variable, we get something similar that the access is the most influential factor to determine the FEC.

```
> importance(rf.d) # importance of each predictor
             %IncMSE IncNodePurity
Value_rnew  2.470782      9016.505
Value_acs  46.880581     30060.981
Value_iff   3.499034      6056.694
```

4.

With the analysis result from random forest modeling, we would like to further confirm the idea that access rate is the most influential variables to determine FEC value. A decision tree model is used.

tr<-ctree(d$Value_fec ~ . , data=d, subset = train)



This is the plotting result from the decision tree model. The training dataset is used in the fitting process to improve the accuracy. From the result, we can see that to classify or predict the FEC value, the most significant variable is access rate to electricity. This makes a lot of sense, because only with more access rate, the government could spend more money on renewable energy research. And electricity is one of the most important infrastructure for any countries. According the result, if the access rate is less than 56.4 %, this country will likely have 60% to 85% energy consumption relied on renewable energy. And if the access rate is between 56% to 97%, this country will have around 25% to 50% of energy consumption of renewable energy. This classification should represent most developing countries. And if the access rate is more than 98%, there are two more classes, and the determent factor is Value_RENW, which is the measurement of proportion of population with primary reliance on clean

fuels and technology. And because of this, if a country has more population that rely on the clean energy, then it should have more energy consumption in renewable energy.

Conclusion

After the analysis on the dataset, I have a better understanding of the global energy area. And it makes me think more about the energy field, sometimes things do not go as we plan. The initial thought and hypothesis were that the more international investment a country receives, the better access to electricity, clean energy, or renewable energy. However, after the 4 models were ran, I am sad to find that international investment seems not to be an important factor in any of those models. Potential cause is that the money received is not properly used or corruption in the government steals the money.

However, we do find something that could be helpful. The hypothesis 2 is wrong. Instead of a positive relationship, a negative relationship is found between access rate and portion of renewable energy used. And the random forest model and decision tree model all implies that the access rate to electricity is crucial determent factor for portion of renewable energy. Therefore, the suggestion for developing countries is to increase the access rate to electricity. It is the most influential factor for renewable energy development.

Throughout this project, I find the KNN modeling seems to have the highest prediction accuracy. And linear models are very helpful to determine if two variables are corelated. Random forest classification and predict integrates randomness to the dataset which makes it also very accurate. Decision tree model is very good at classification visualization.

Future Work

Only four datasets are used in this project. And in the future, we could gather more information about each countries and used many of them to build new models or improve existing models.

Citation

https://www.sciencedirect.com/science/article/pii/S0973082613000094?casa_token=BYwEyACw2xYAAAAA:X7RKemDkkCQ7rH6bT-B2_rn1RPyy-YDRfrRAE59h3iVkPmANuMWDqP2HDVjBtMq7qla62ipV

https://www.sciencedirect.com/science/article/pii/S1364032113007752?casa_token=1Mao_Se_UBgAAAAA:Cq9PgWaj8W7gIUe8i-rdfMt5MyBO0ip-109hZ85OEHBVho_JLwymANvYPBpeyDepr9IhzaGk