Modern renewable energy consumption in R

INSTRUCTOR: HAMID RAJAEE

By: SARA KHOSRAVI

March 2021

Overview:

The dataset is taken from Kaggle site.

In this project, a dataset include 5095 observations and 7 variables, The dataset is named “Modern renewable energy consumption”.

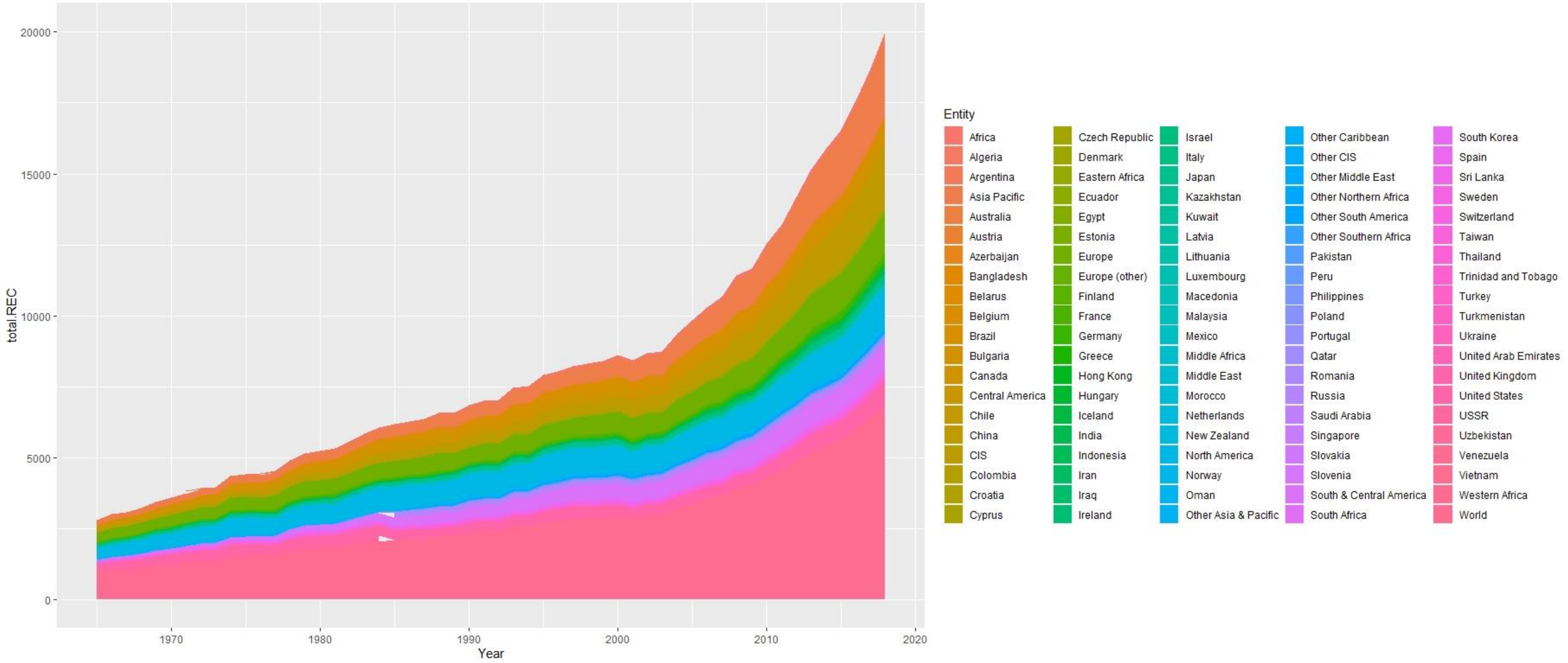
EDA DATA:

Business Understanding

- In this project we looked at, what share renewable technologies collectively accounted for in the energy mix.
- Globally we see that hydropower is by far the largest modern renewable source *[since traditional biomass is not included here]*. But we also see wind and solar power are both growing rapidly.
- The dataset have 7 columns. For understanding the dataset, Analysis and compare the dataset, 3 main columns by calculation have been added the dataset.

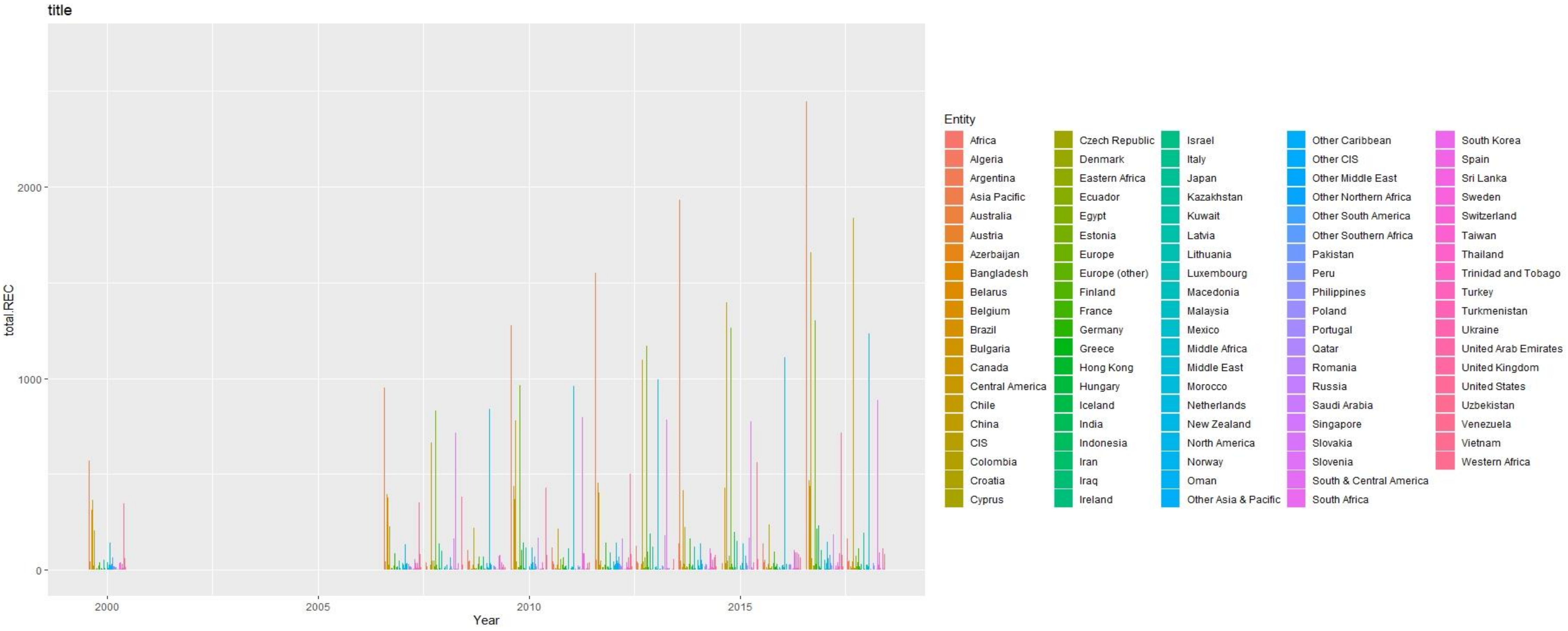
Data Understanding :

The chart shows this as a stacked area chart, which allows us to more readily see the breakdown of the renewable mix, and relative contribution of each.



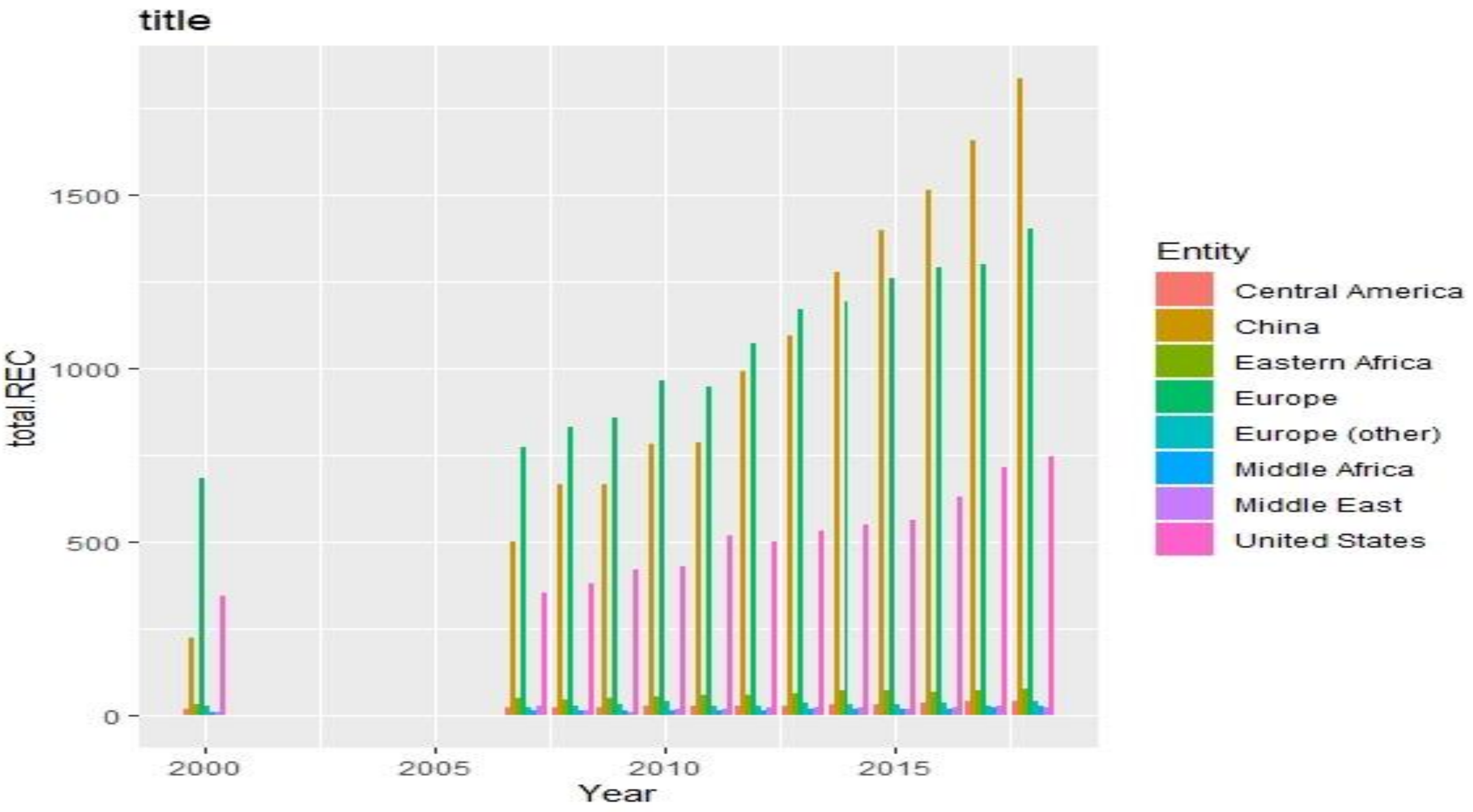
Data Understanding:

❖ This Graph illustrated the Total Renewable Energy during the 1965-2018. But the value before the 2007 is less, Therefore by using CUMSUM for cumulative the total consumption Renewable Energy during the 1965-2007, and after that used this Graph. This graph is creating for visualization of data to understanding better what happen during the year between 19965-2018.



Data Understanding:

- This graph shows that 8 top Renewable Energy Consumer(REC) in the world.
  To obtain this diagram, ***Filter, Subset*** and ***Full Joint*** commands have been used.



Data Understanding:

❖ 10 Top Renewable Energy Consumption in the dataset in 2017. Asia Pacific, North America and Europe are The most important the consumer of Renewable Energy in the world.

```
# A tibble: 1,089 x 8
# Groups:   Entity [99]
   Entity          Year Hydropower   Solar     Wind total.REC GROUPEntity$Entity
   <chr>          <int>      <dbl>   <dbl>    <dbl>     <dbl> <chr>
 1 World           2017      4065. 454.     1.13e+3     6232. World
 2 Asia Pacific    2017      1649. 227.     3.77e+2     2446. Asia Pacific
 3 China           2017      1165. 118.     2.95e+2     1657. China
 4 Europe          2017       585. 125.     3.84e+2     1302. Europe
 5 North Ameri~    2017       725.  82.5    2.97e+2     1204. North America
 6 South & Cen~    2017       720.   7.46   5.61e+1      860. South & Central A~
 7 United Stat~    2017       297.  78.1    2.57e+2      715. United States
 8 Brazil          2017       371.   0.832  4.24e+1      465. Brazil
 9 Canada          2017       397.   3.29   2.91e+1      439. Canada
10 CIS             2017       240.   0.767  5.98e-1      242. CIS
# ... with 1,079 more rows, and 1 more variable: Growth.rate <dbl>
> class(TOP.REC)
[1] "grouped_df" "tbl_df"     "tbl"        "data.frame"
```

Data Understanding:

- Getting familiar with data for Data Understanding in EDA.
  Data frame has a 5059 observation and 7 columns. The missing value can be seen in the dataset. The important column is Entity, Year , Hydropower, Solar and Wind, So by using slice the column of Code dropped at the dataset.

```
> typeof(REC)
[1] "list"
> # Compactly Display the Structure of an Arbitrary R Object
> str(REC)
'data.frame':	5095 obs. of  7 variables:
 $ Entity         : chr  "Africa" "Africa" "Africa" "Africa" ...
 $ Code           : chr  NA NA NA NA ...
 $ Year           : int  1965 1966 1967 1968 1969 1970 1971 1971 1971 1971 ...
 $ Hydropower     : num  14.3 15.6 16.2 18.6 21.6 ...
 $ Solar          : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Wind           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ OtherRenewables: num  0 0 0 0 0 0 0.164 0.164 0.164 0.164 ...
> 
```

Feature Engineering:

❖ For preparation and analysis, the dataset **3 Continues COLUMNS** and **one Categorical Column** are ***added*** to ***dataset*** to make it easy to handle the project.

1. "total.REC“: Total the consumption of Hydropower, Solar, Wind and Other Renewable Energy

```
REC$total.REC <- NA
REC$total.REC <- rowSums(REC[ ,c(3:6)], na.rm=TRUE)
```

2. ” cum_total “: cumulative REC consumption

```
NEWREC$cum_total <- cumsum(NEWREC$total.REC)
```

3. "Growth.rate“: Growth rate per annul

```
RECF <- NEWREC %>% group_by(Entity) %>% mutate(Growth.rate = (total.REC-lag(total.REC))/lag(total.REC))
```

4. “GROUPEntity”:

```
NEWREC$GROUPEntity <- NEWREC %>% group_by(Entity)
```

Data preparation:

Data preparation or Data cleaning is:

1) Handling duplicate data

2) Handling Missing Values

3) Handling outliers

❖ By using frequency in a dataset is observed

That data duplication exists in Africa.

This problem is solved by using the

Duplicated command.

```
sum(duplicated(REC))
RowDuplicate <- which(duplicated(REC))
REC <- REC[-RowDuplicate,]
```

❖ For handling Missing value in project is

converted missing value to NA and after that use some command in R to handle that.

```
REC[REC=='']<-NA # converting Null to Na
sum(is.na(REC))      # 11268 number of missing values
colSums(is.na(REC))
```

❖ This project has outlier but this outlier it

is important for analysis of data. Because this

Outlier happened due to the rapid scientific progress

In this field recently.

```
> # since target is categorical variable, in uni-variate Analysis for summarizing I
> # will find frequency and for visualization I plot: pie chart or bar-chart
> tbl<-table(REC$ Entity)
> tbl

               Africa               Algeria             Argentina          Asia Pacific
                   58                    54                    54                    54
            Australia               Austria            Azerbaijan            Bangladesh
                   54                    54                    34                    54
              Belarus               Belgium                Brazil              Bulgaria
                   34                    54                    54                    54
               Canada       Central America                 Chile                 China
                   54                    54                    54                    54
                  CIS              Colombia               Croatia                Cyprus
                   54                    54                    29                    54
       Czech Republic               Denmark        Eastern Africa               Ecuador
                   54                    54                    54                    54
                Egypt               Estonia                Europe        Europe (other)
                   54                    34                    54                    54
              Finland                France               Germany                Greece
                   54                    54                    54                    54
            Hong Kong               Hungary               Iceland                 India
                   54                    54                    54                    54
            Indonesia                  Iran                  Iraq               Ireland
                   54                    54                    54                    54
               Israel                 Italy                 Japan            Kazakhstan
                   54                    54                    54                    34
               Kuwait                Latvia             Lithuania            Luxembourg
                   54                    34                    34                    54
            Macedonia              Malaysia                Mexico         Middle Africa
                   29                    54                    54                    54
          Middle East               Morocco           Netherlands           New Zealand
                   54                    54                    54                    54
        North America                Norway                  Oman   Other Asia & Pacific
                   54                    54                    54                    54
      Other Caribbean             Other CIS     Other Middle East  Other Northern Africa
                   54                    34                    54                    54
  Other South America  Other Southern Africa              Pakistan                  Peru
                   54                    54                    54                    54
          Philippines                Poland              Portugal                 Qatar
                   54                    54                    54                    54
              Romania                Russia          Saudi Arabia             Singapore
                   54                    34                    54                    54
             Slovakia              Slovenia South & Central America          South Africa
                   54                    29                    54                    54
          South Korea                 Spain             Sri Lanka                Sweden
                   54                    54                    54                    54
          Switzerland                Taiwan              Thailand   Trinidad and Tobago
                   54                    54                    54                    54
               Turkey          Turkmenistan               Ukraine  United Arab Emirates
                   54                    34                    54                    54
       United Kingdom         United States                  USSR            Uzbekistan
                   54                    54                    20                    34
            Venezuela               Vietnam        Western Africa                 World
                   54                    54                    54                    54
```

Descriptive Statistics:

Descriptive statistics comprises three main categories – Frequency Distribution, Measures of Central Tendency, and Measures of Variability.

Descriptive statistics helps facilitate data visualization. It allows for data to be presented in a meaningful and understandable way, which in turn, allows for a simplified interpretation of the data set in question.

	Hydropower	Solar	Wind	Other Renewable Energy	Total Of Renewable Energy
Mean	74.02	1.31	4.7	5.7	85.79
Median	6.03	0	0	0.042	7.53
Standard deviation	284.48	15.3	41.77	29.1	348.9
IQR	29.1	00.2	0.03	1.3	31.98

Hydropower	Solar	Wind	Other Renewable Energy	Total Of Renewable Energy
0.00000	0.000000e+00	0.000000e+00	0.0000	0.000000
0.81007	0.000000e+00	0.000000e+00	0.0000	1.204431
6.03100	0.000000e+00	0.000000e+00	0.0420	7.527449
29.93543	2.052632e-03	3.030303e-02	1.3099	33.187437
4193.10415	5.846309e+02	1.269953e+03	625.8054	6673.493806

Descriptive Statistics:

sapply(NUMdata, quantile, probs = seq(0, 1, 1/10), na.rm = TRUE)

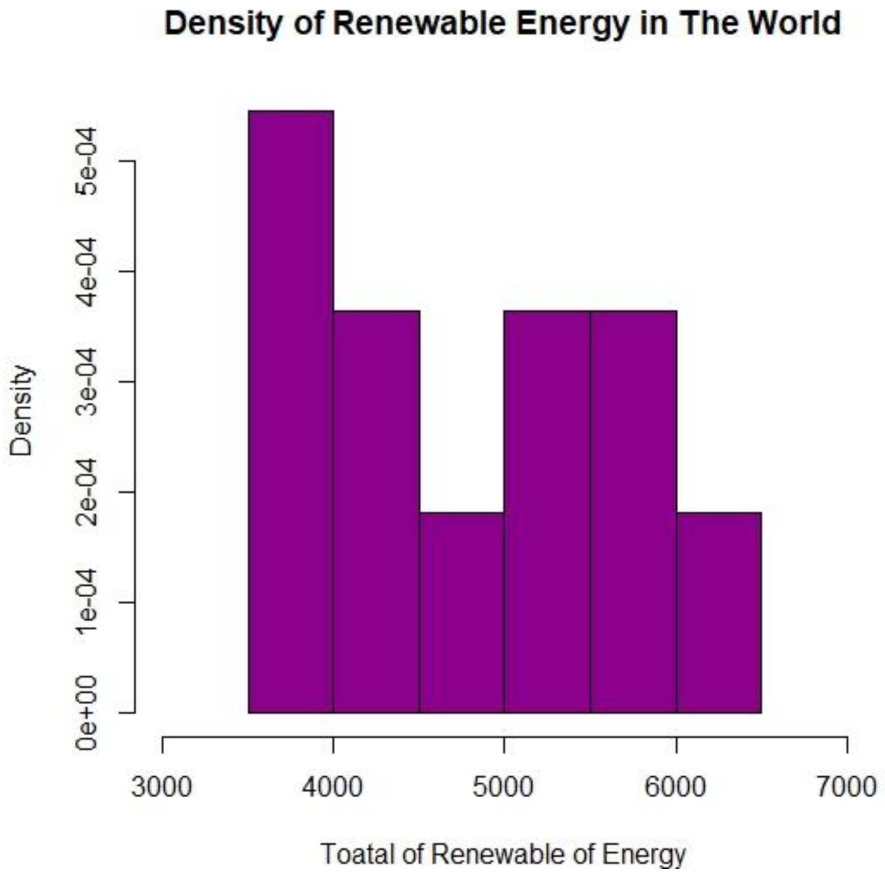
❖For atain quartile is used 1/10 for porobs to get 10 quartile for dataset to accuaracy in distribution of data.

```
> sapply(NUMdata, quantile, probs = seq(0, 1, 1/10), na.rm = TRUE)
       Hydropower        Solar        Wind OtherRenewables   total.REC Rtotal.REC
0%       0.000000   0.00000000    0.000000        0.000000    0.000000      0.000
10%      0.000000   0.00000000    0.000000        0.000000    0.020040      0.020
20%      0.325420   0.00000000    0.000000        0.000000    0.600000      0.600
30%      1.480799   0.00000000    0.000000        0.000000    1.938000      1.940
40%      3.279117   0.00000000    0.000000        0.000000    3.911568      3.910
50%      6.031000   0.00000000    0.000000        0.042000    7.527449      7.530
60%     12.192727   0.00000000    0.000000        0.200792   14.405273     14.408
70%     21.542278   0.00015476    0.006000        0.685000   25.186939     25.184
80%     41.507005   0.00855600    0.122622        2.098000   47.478641     47.480
90%    141.900200   0.17390778    1.720150        8.137519  159.725955    159.730
100% 4193.104151 584.63091780 1269.953375      625.805362 6673.493806   6673.490
> 
```

Univariate analysis:

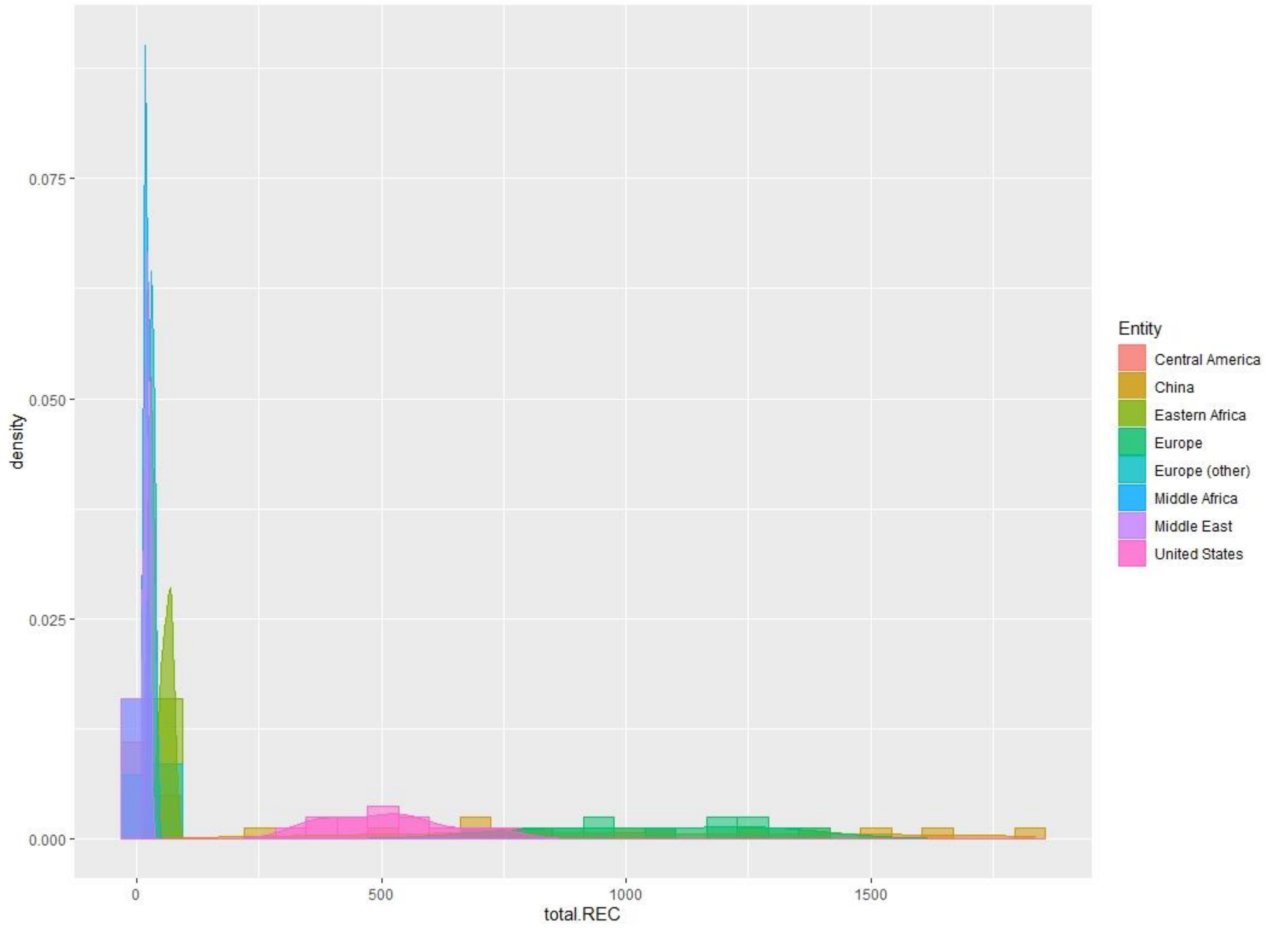
For **Visualization** this **Numerical variable** (Total of Renewable Energy) **plot density** is chosen.

This diagram shows the consumption of renewable energy versus density.



Univariate analysis for Numerical variables :

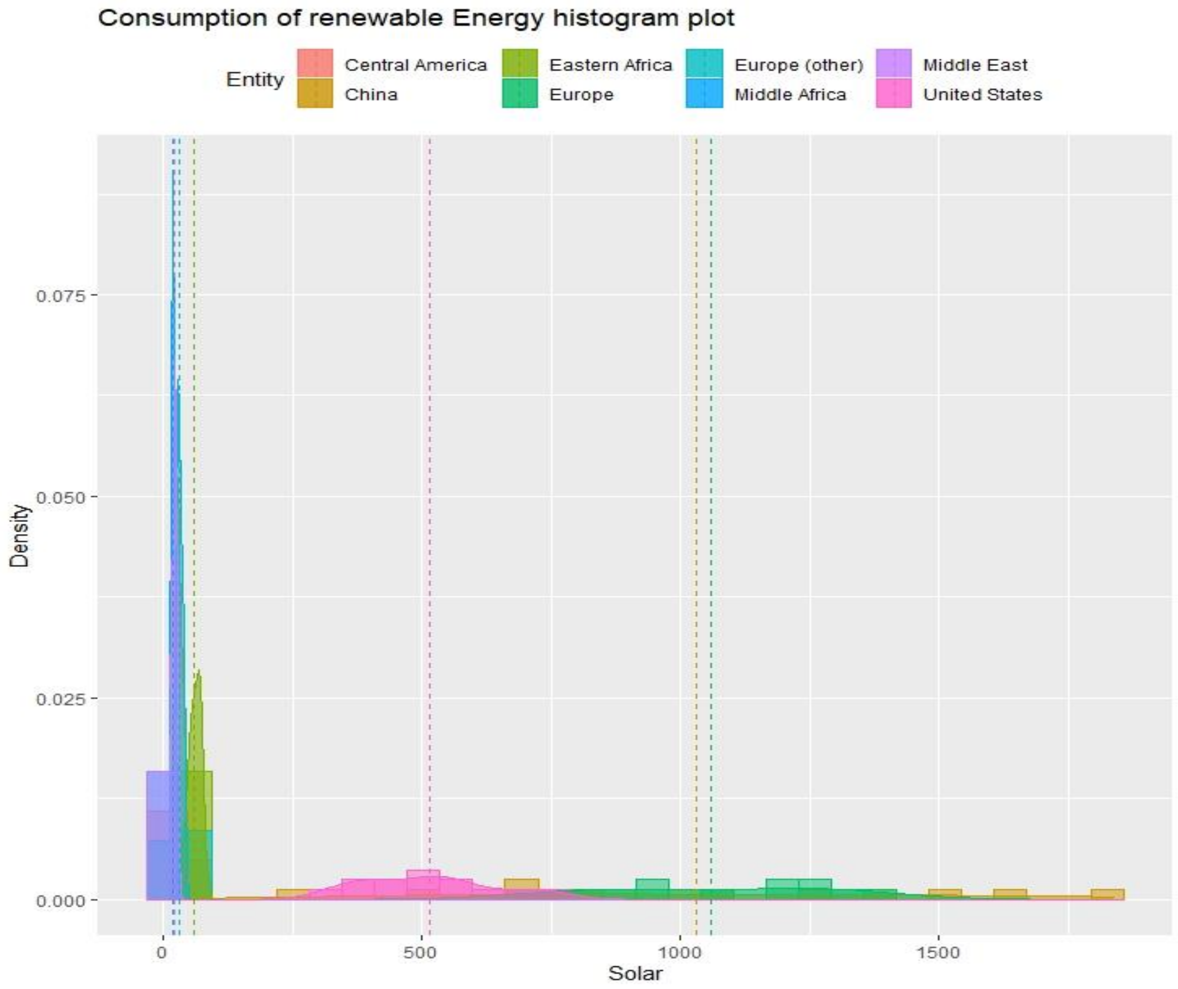
The distribution of "total.REC" shows us , we have mutated recently. Consumption of renewable of energy in the last 10 years. And this graph shows the jump in new energy consumption in recent years.



Univariate analysis for Numerical variables :

For **Visualization** this **Numerical variable** (Total of Renewable Energy) **plot density** is chosen.

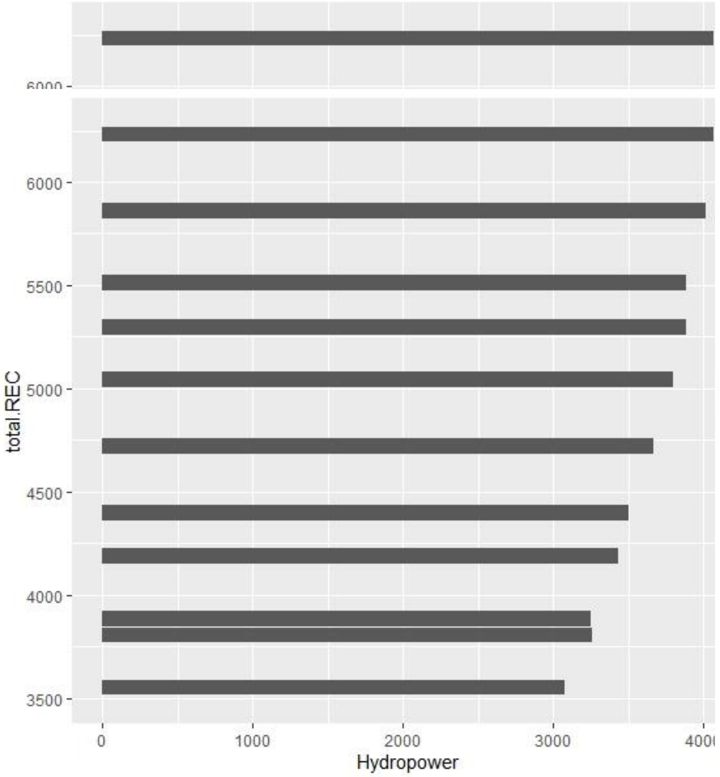
This graph shows , the Solar Energy versus density.

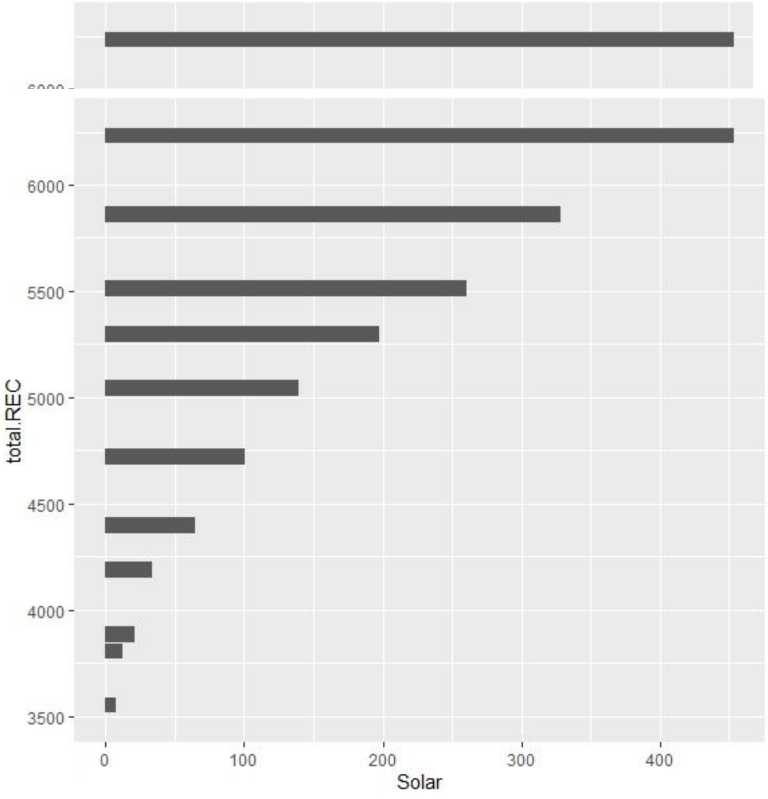


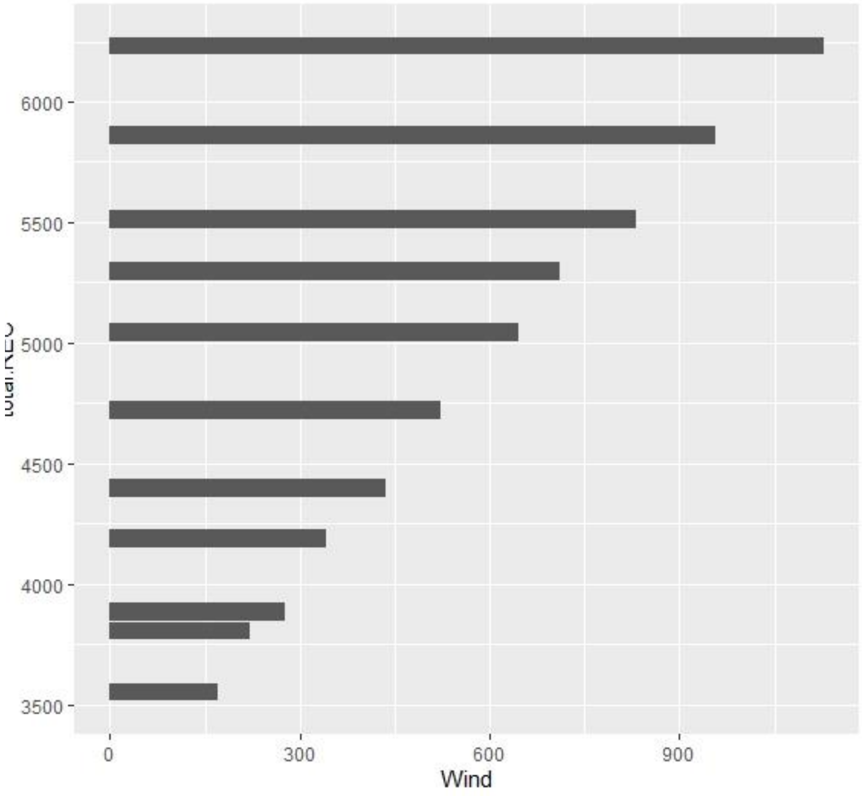
Bi-variate Analysis for Continuous Vs. Continuous:

The amount of Consumption Hydropower, Wind and Solar energy of the total of energy.

In charts shown here we look at the breakdown of renewable technologies by their individual components – hydropower, solar, wind, and others.

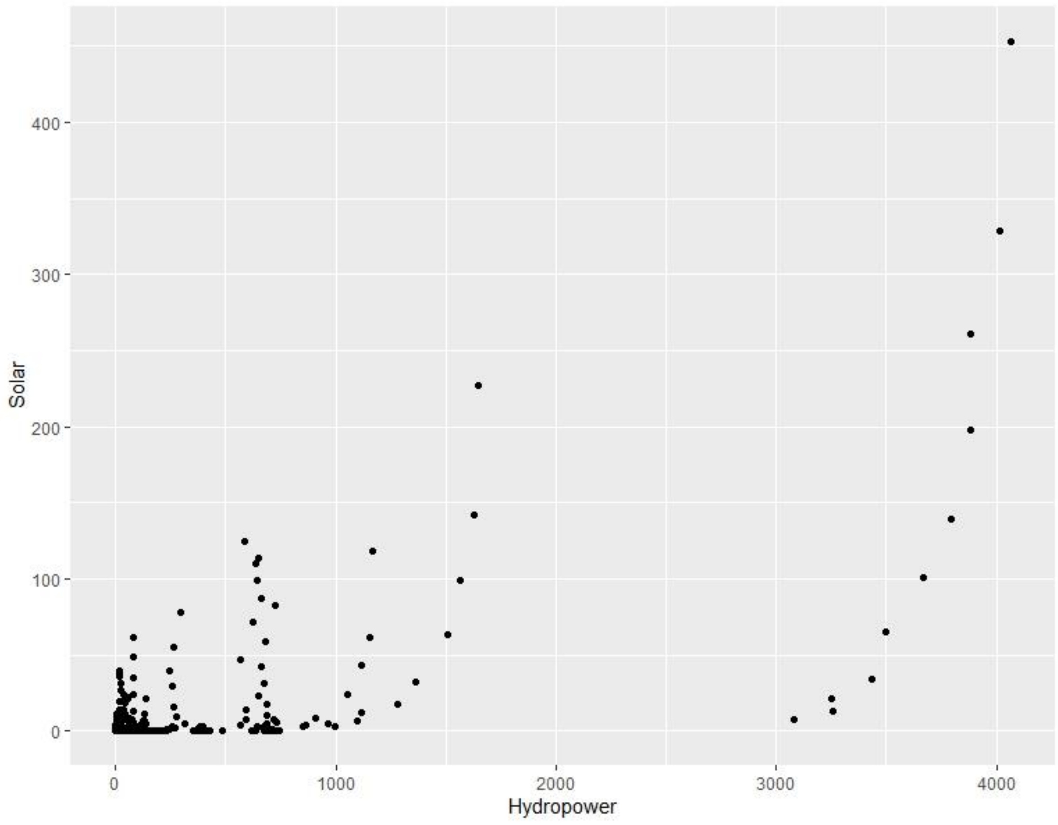


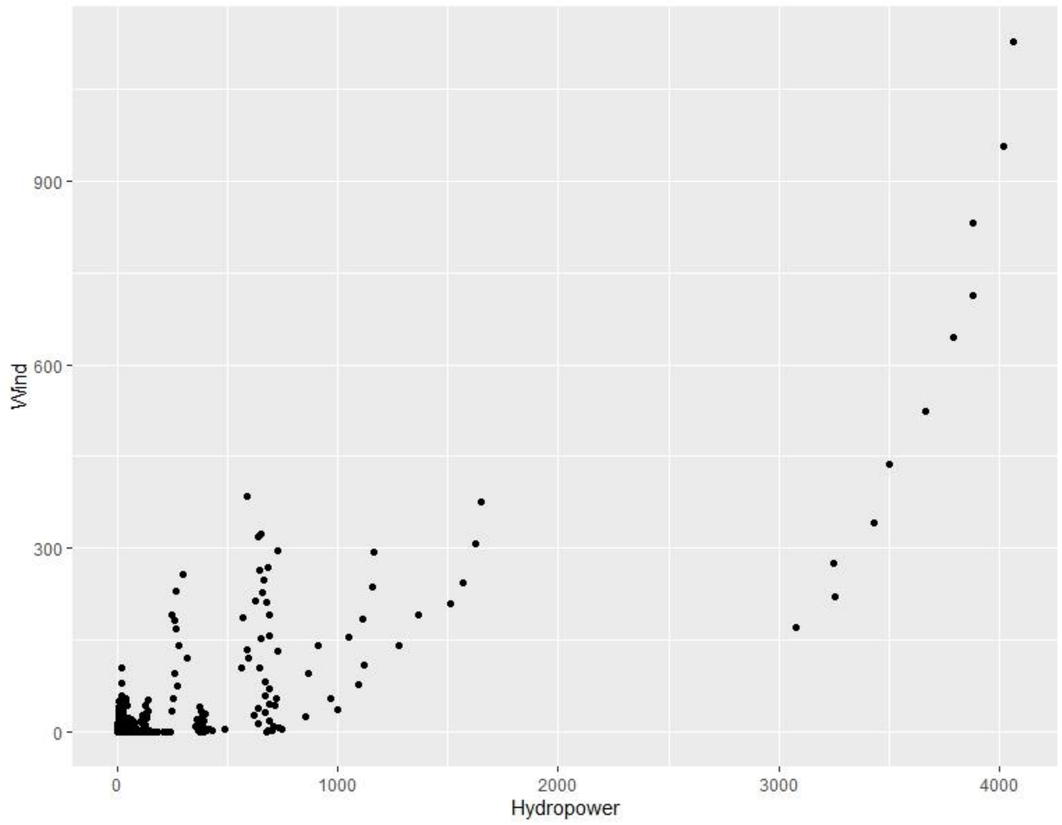




Bi-variate Analysis for Continuous Vs. Continuous:

Using **scatter plot** for showing the relationship between solar versus Hydropower. Also, relationship between Hydropower versus Solar .

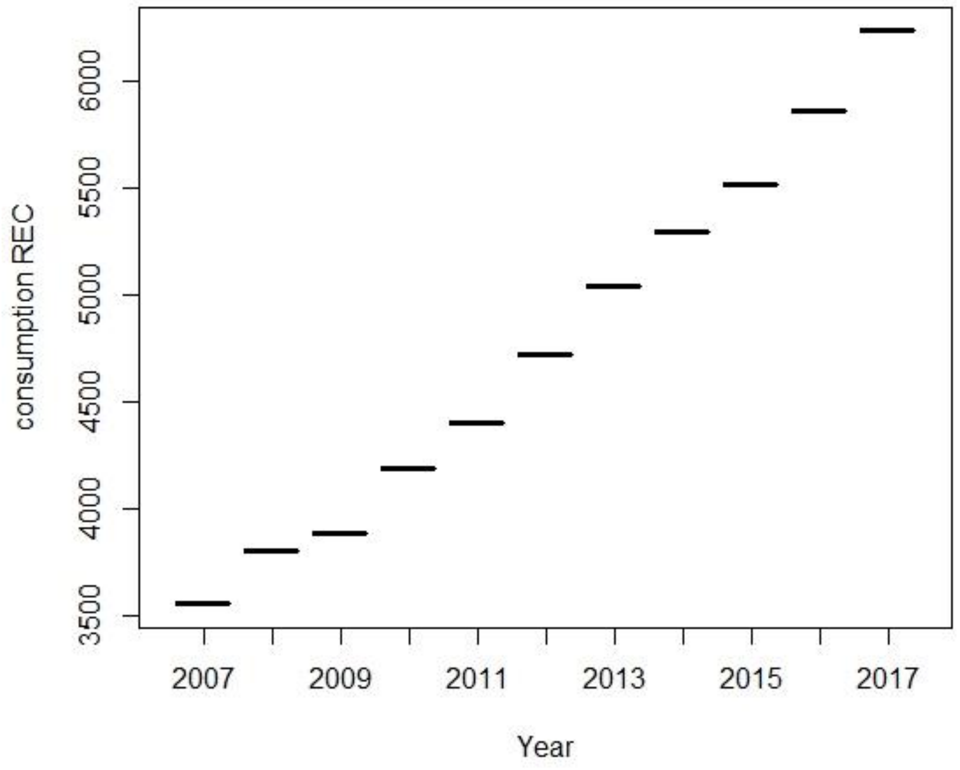




Bi-variate Analysis for Continuous Vs. Categorical:

Consumption Renewable Energy during the 2007-2017 in the word. Target **Year as a categorical** variable in this project.

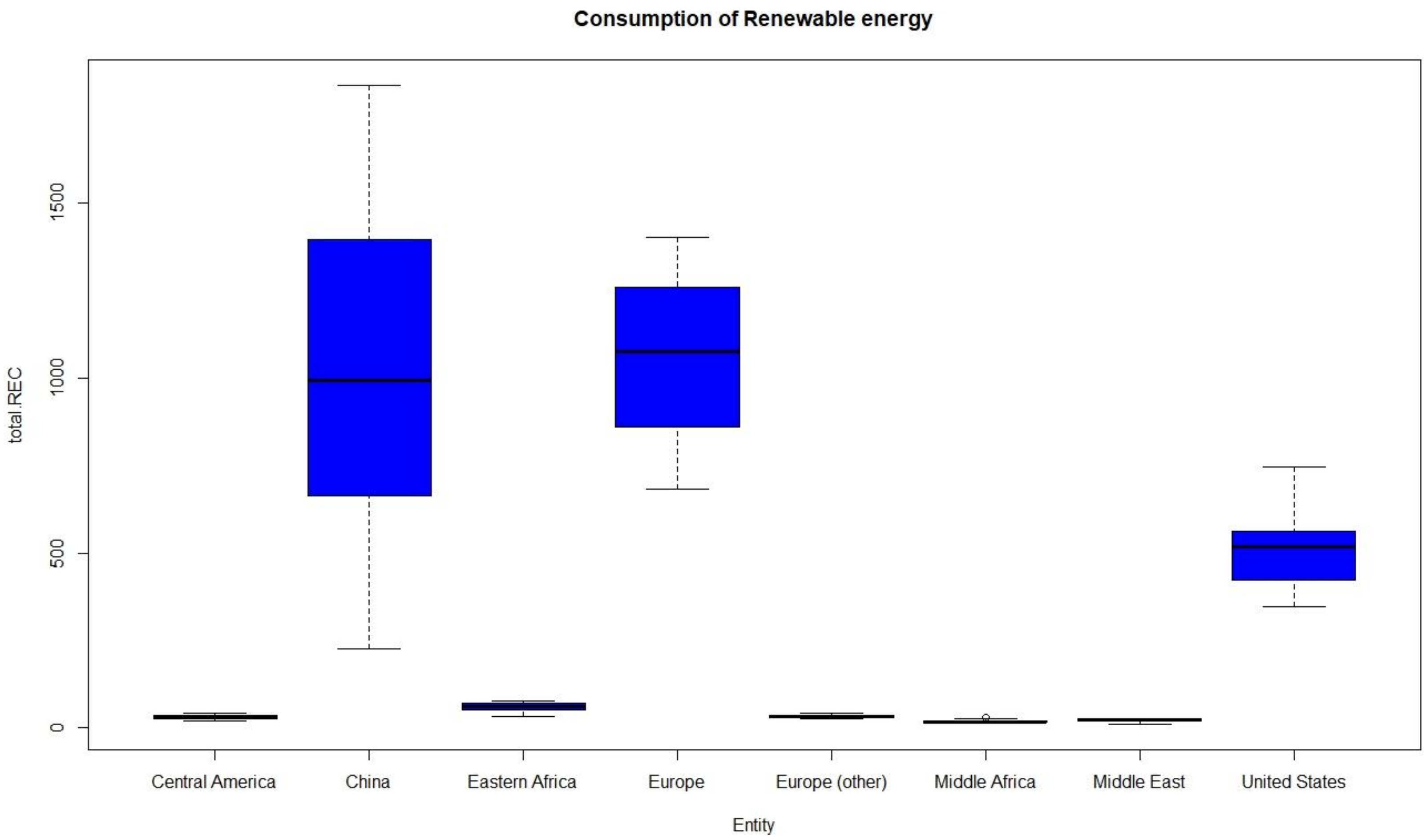
Treating year as a **categorical variable** will calculate effect of each individual **year** - i.e., what impact on the target **variable** was in average each year. On the other hand, including t as **numerical variable** says what happens on average two **years** later.



Bi-variate Analysis for Continuous Vs. Categorical :

visualization: **box plot**

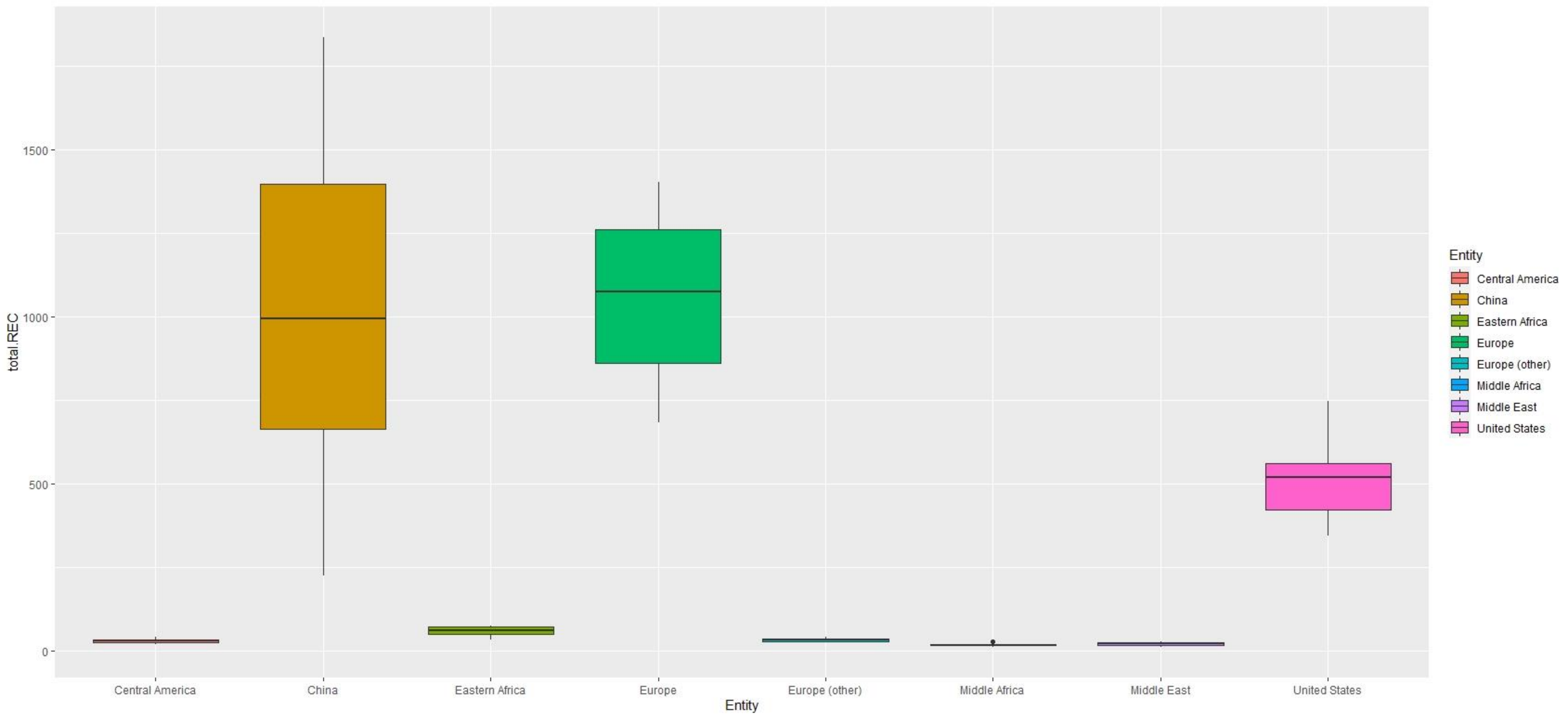
This graph illustrated the most consumption of Renewable Energy in the world are China, Europe and United States.



Bi-variate Analysis for Continuous Vs. Categorical:

visualization: Grouped **box plot**

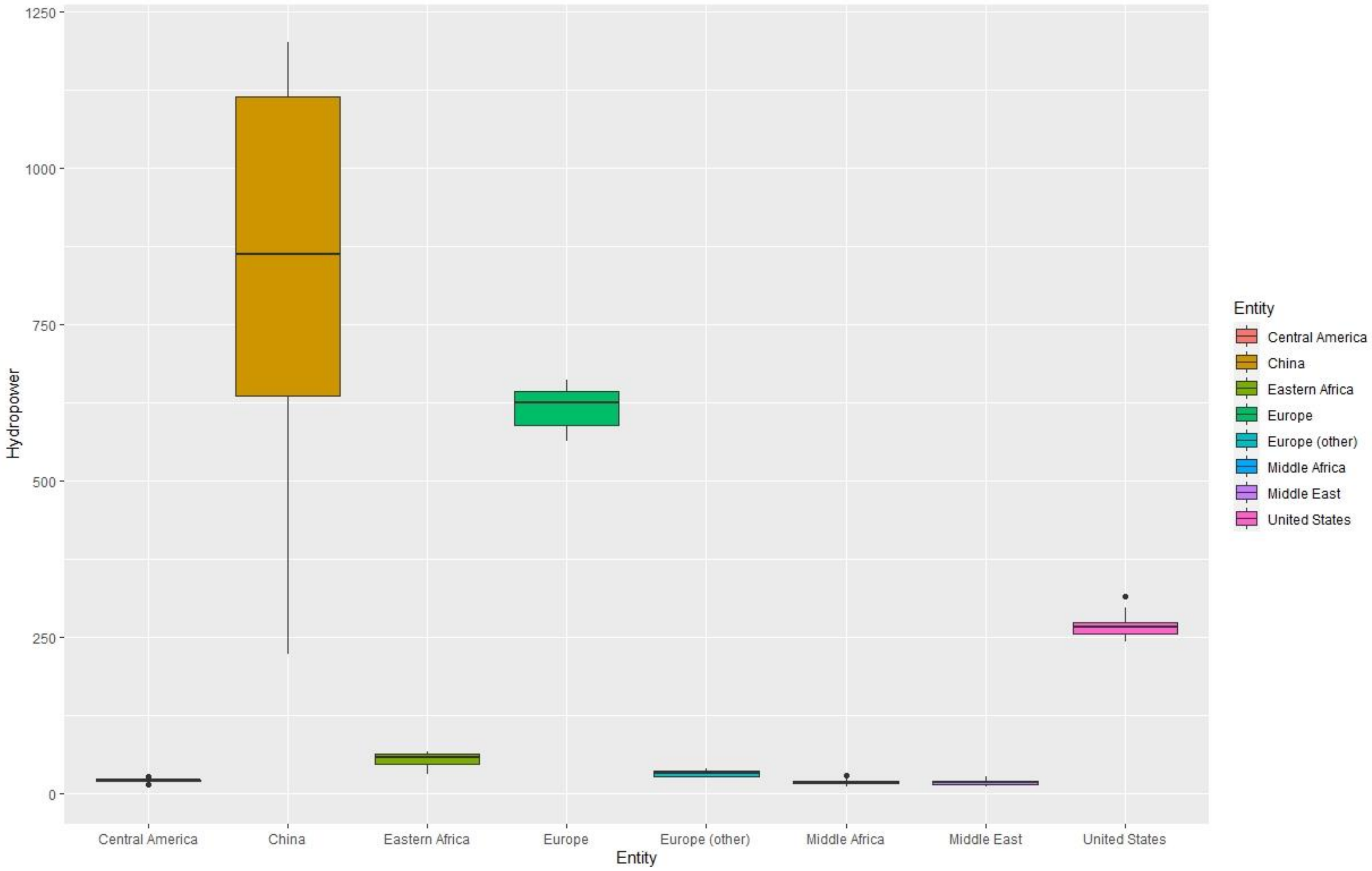
Bi-variate Analysis for continuous(total.REC) Vs. categorical (Entity)



Bi-variate Analysis for Continuous Vs. Categorical:

visualization: Grouped **box plot**

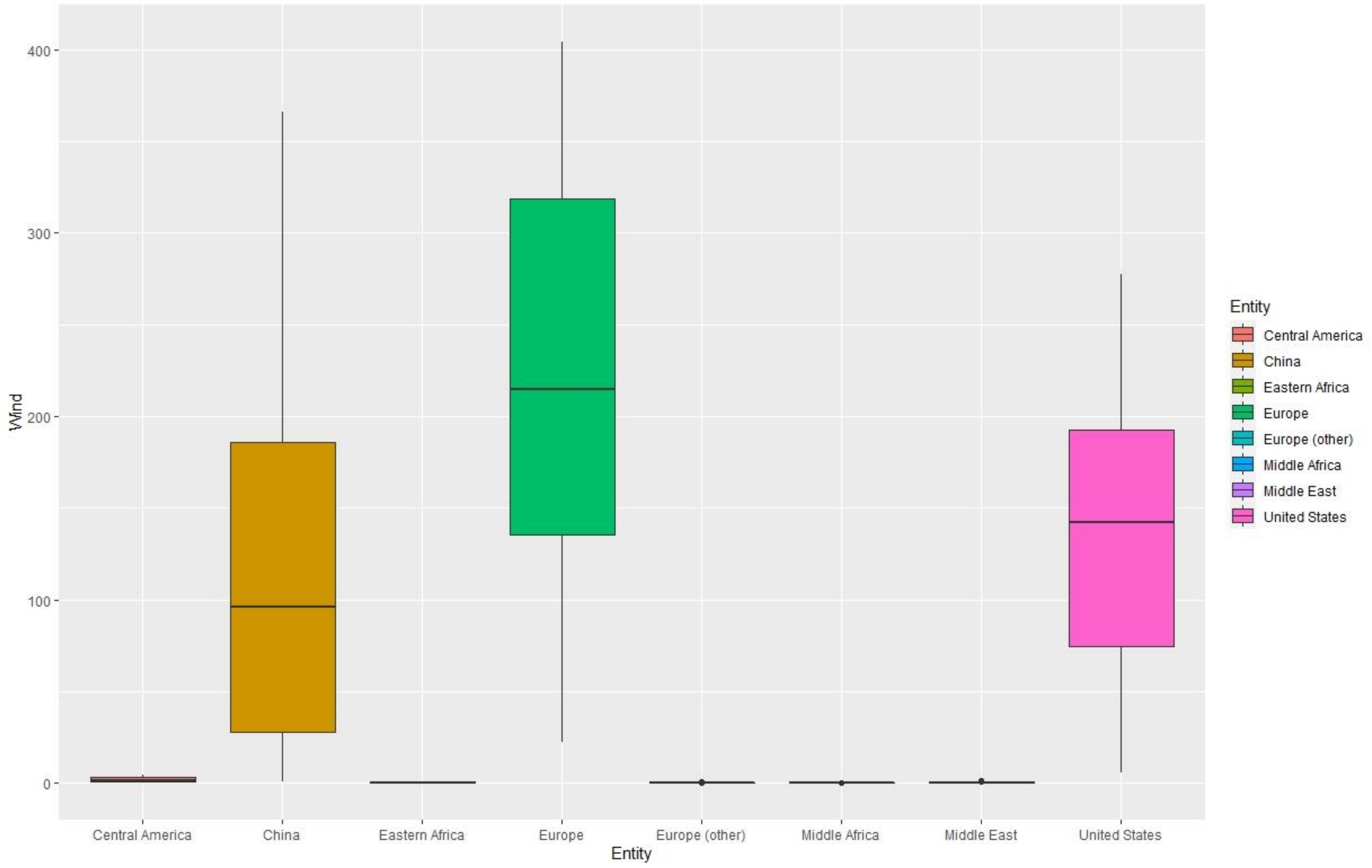
Bi-variate Analysis for continuous(Hydropower) Vs. categorical (Entity)



Bi-variate Analysis for Continuous Vs. Categorical:

visualization: Grouped **box plot**

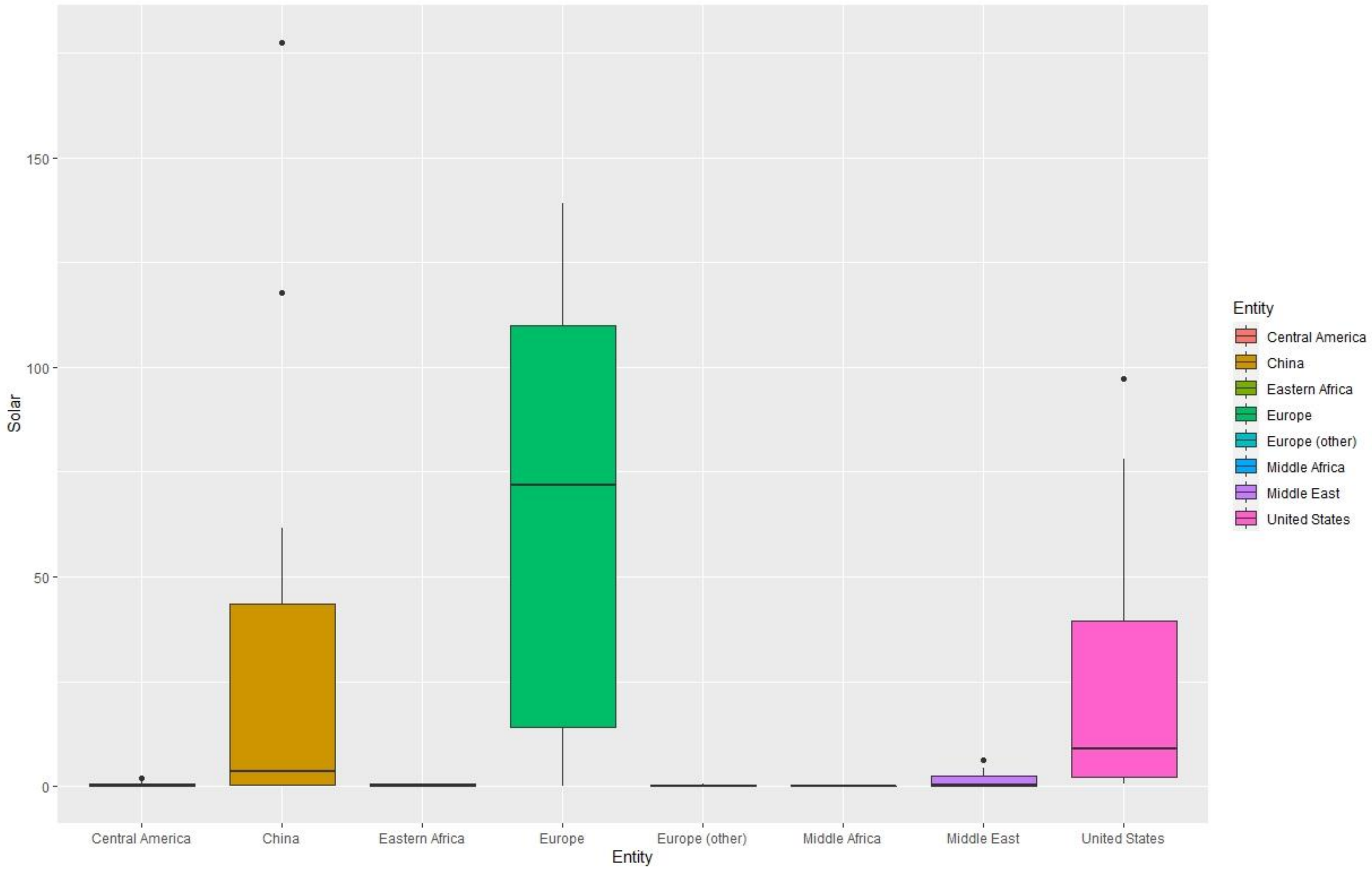
Bi-variate Analysis for continuous (Wind) Vs. categorical (Entity)



Bi-variate Analysis for Continuous Vs. Categorical:

visualization: Grouped **box plot**

Bi-variate Analysis for continuous (Wind) Vs. categorical (Entity)



Bi-variate Analysis for Continuous Vs. Categorical:

visualization: Grouped **box plot**

By using aggregation function compare the numerical variable is comfortable.

```
agg2 <- cbind(aggregate( total.REC ~ Entity , REC , min),
            aggregate( total.REC ~ Entity , REC  , max)[,2],
            aggregate( total.REC ~ Entity , REC  , mean)[,2])

names(agg2) <- c("total.REC","min_REC","max_REC","mean_REC")
agg2

write.table(agg2, file = "agg2.csv",
          sep = "\t", row.names = F)
```

Bi-variate Analysis for Continuous Vs. Categorical:

visualization: Grouped **box plot**

Use aggregation Function for attain total consumption versus Entity, and also calculated min, max, mean for them

```
> agg2 <- cbind(aggregate( total.REC ~ Entity , REC , min),
+               aggregate( total.REC ~ Entity , REC  , max)[,2],
+               aggregate( total.REC ~ Entity , REC  , mean)[,2])
> names(agg2) <- c("total.REC","min_REC","max_REC","mean_REC")
> agg2
                total.REC      min_REC      max_REC     mean_REC
1                  Africa  14.27880557 1.647168e+02 6.589835e+01
2                 Algeria   0.05400000 7.557174e-01 3.221200e-01
3               Argentina   1.21091560 4.546722e+01 2.362641e+01
4            Asia Pacific 152.20159630 2.714488e+03 6.777371e+02
5               Australia   7.62904280 4.915484e+01 1.838341e+01
6                 Austria  16.08300000 5.121165e+01 3.429335e+01
7              Azerbaijan   0.69826520 3.446800e+00 1.856208e+00
8              Bangladesh   0.00000000 1.307640e+00 6.283037e-01
9                 Belarus   0.01600000 8.188784e-01 1.098908e-01
10                Belgium   0.13400000 1.715265e+01 2.792233e+00
11                 Brazil  23.97524500 4.921938e+02 2.345469e+02
12               Bulgaria   1.30372300 8.766185e+00 3.312188e+00
13                 Canada 117.12293880 4.385852e+02 2.979626e+02
14        Central America   1.21829100 4.209269e+01 1.434817e+01
15                  Chile   3.57087520 3.868918e+01 1.625462e+01
16                  China  19.38348840 1.836653e+03 3.433328e+02
17                    CIS  85.32093640 2.473767e+02 1.879055e+02
18               Colombia   3.54394947 5.933623e+01 2.684201e+01
19                Croatia   3.80500000 9.937000e+00 6.650965e+00
20                 Cyprus   0.00000000 4.638000e-01 5.425593e-02
21         Czech Republic   1.08275300 9.618473e+00 3.160531e+00
22                Denmark   0.01900000 2.191709e+01 4.914800e+00
23         Eastern Africa   6.13641138 7.570876e+01 3.031369e+01
24                Ecuador   0.34471320 2.124234e+01 5.738228e+00
25                  Egypt   1.73240480 1.695809e+01 1.052231e+01
26                Estonia   0.00000000 2.048773e+00 4.342091e-01
27                 Europe 305.52508640 1.403121e+03 6.392389e+02
28         Europe (other)   9.41572165 4.057329e+01 2.711962e+01
29                Finland   8.74545454 3.215866e+01 1.783236e+01
30                 France  45.98265740 1.110707e+02 6.817898e+01
31                Germany  13.71347780 2.260910e+02 5.254236e+01
32                 Greece   0.83084720 1.610984e+01 4.943902e+00
```

Bi-variate Analysis for Continuous Vs. Categorical :

Test of independence: Anova

Perform the ANOVA test:

- One-way ANOVA

In the one-way ANOVA example, we are modeling crop total.REC as a function of the type of Entity used. First, we will use aov() to run the model, then we will use summary() to print the summary of the model.

```
one.way <- aov(total.REC~Entity, data = REC.ORGIN)
summary(one.way)
```

Bi-variate Analysis for Continuous Vs. Categorical :

Test of independence: Anova

❖ Two-way ANOVA

In the two-way ANOVA example, we are modeling crop total.REC as a function of type of Entity and Year. First, we use aov() to run the model, then we use summary() to print the summary of the model.

```
two.way <- aov(total.REC~Entity + Year, data = REC.ORGIN)
summary(two.way)
```

❖ Adding interactions between variables

Sometimes you have reason to think that two of your independent variables have an interaction effect rather than an additive effect.

```
interaction <- aov(total.REC~Entity * Year, data =  REC.ORGIN)
summary(interaction
```

Bi-variate Analysis for Continuous Vs. Categorical :

Test of independence: Anova

❖ Adding a Solaring variable

If you have grouped your experimental treatments in some way, or if you have a confounding variable that might affect the relationship you are interested in testing, you should include that element in the model as a Solaring variable. The simplest way to do this is just to add the variable into the

model with a '+'.

```
Solaring <- aov(total.REC~Entity + Year + Solar, data = REC.ORGIN)
summary(Solaring)
```

❖ Find the best-fit model:

There are now four different ANOVA models to explain the data. How do you decide which one to use? Usually, you will want to use the 'best-fit' model -

the model that best explains the variation in the dependent variable.

Bi-variate Analysis for Continuous Vs. Categorical :

Test of independence: Anova

```
install.packages("AICcmodavg")
library("AICcmodavg")

model.set <- list(one.way, two.way, interaction, Solaring)
model.names <- c("one.way", "two.way", "interaction", "Solaring")

aictab(model.set, modnames = model.names)
```

- Check for homoscedasticity

To check whether the model fits the assumption of homoscedasticity, look at the model diagnostic plots in R using the plot() function:

```
par(mfrow=c(2,2))
plot(two.way)
par(mfrow=c(1,1))
```

Bi-variate Analysis for Continuous Vs. Categorical :

Test of independence: Anova

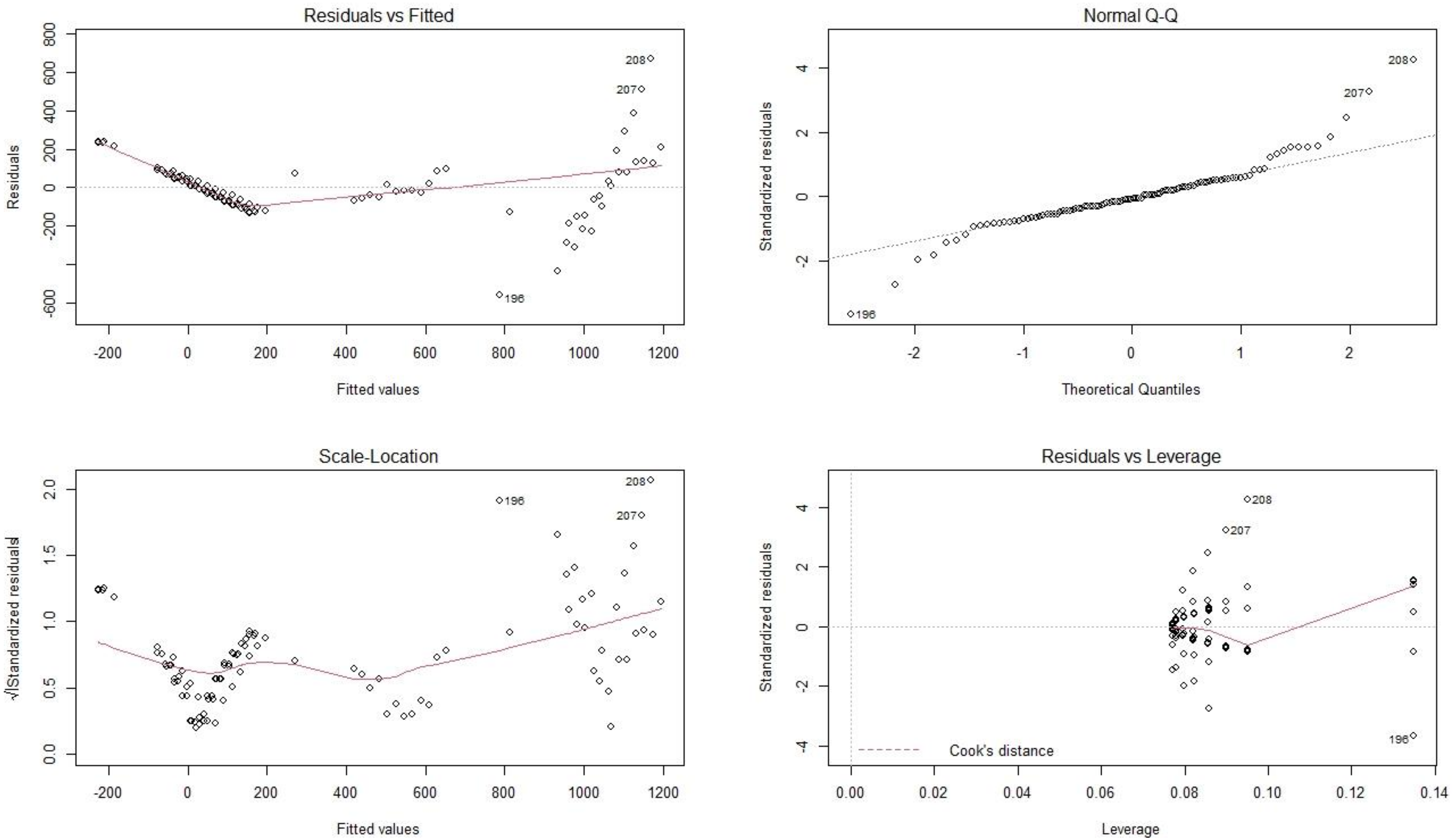
Focus on the column: the probability that F is greater than the listed value from the previous column. This is often called the *p value*. In most cases you put significance at the alpha=.05 level, or *we require the P value to be less then .05* to be considered statistically significant.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> Solaring <- aov(total.REC~Entity + Year + Solar, data =  REC.ORGIN)
> summary(Solaring)
            Df   Sum Sq Mean Sq F value   Pr(>F)    
Entity       7 19461739 2780248  258.00  < 2e-16 ***
Year         1  1032737 1032737   95.83 5.15e-16 ***
Solar        1  1589781 1589781  147.53  < 2e-16 ***
Residuals   94  1012963   10776                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> install.packages("AICcmodavg")
```

- **5.115e-16 < 0.05**
- **Therefore, we fail to reject the null hypothesis**

Bi-variate Analysis for Continuous Vs. Categorical :

Test of independence: Anova



Conclusion:

- We see in this Project the rapid growth of renewable technologies in the World
- This interactive chart shows the amount of energy generated from solar power each year.
- Solar generation at scale – compared to hydropower, for example – is a relatively modern renewable energy source but is growing quickly in many countries across the world.


Thank you for your attention!