In an era of increasing global concern over sustainability and climate change, the pursuit of renewable energy sources has become paramount for nations worldwide. This study delves into the renewable energy landscape of three diverse yet pivotal countries: Egypt, Algeria, and Argentina. By employing advanced statistical modeling techniques, we aim to analyze the current state, trends, and potential trajectories of renewable energy adoption in these nations.Through this exploration, we seek not only to understand the unique challenges and opportunities each country faces but also to offer insights that can inform policy decisions, drive sustainable development, and pave the way towards a cleaner, more resilient energy future.

```
library(ggplot2)
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(readr)
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(reshape)
```

```
##
## Attaching package: 'reshape'

## The following object is masked from 'package:dplyr':
##
##     rename

## The following object is masked from 'package:plotly':
##
##     rename
```

```
energy_data<- read_csv("~/modern-renewable-energy-consumption.csv")
```

```
## Rows: 5695 Columns: 7

## -- Column specification --------------------------------------------------------
```

```
## Delimiter: ","
## chr (2): Entity, Code
## dbl (5): Year, Other renewables (including geothermal and biomass) electrici...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(energy_data)
```

```
## # A tibble: 6 x 7
##   Entity Code   Year Other renewables (including geothe~1 Solar generation - T~2
##   <chr>  <chr> <dbl>                                <dbl>                  <dbl>
## 1 Africa <NA>   1971                                0.164                      0
## 2 Africa <NA>   1972                                0.165                      0
## 3 Africa <NA>   1973                                0.17                       0
## 4 Africa <NA>   1974                                0.175                      0
## 5 Africa <NA>   1975                                0.172                      0
## 6 Africa <NA>   1976                                0.185                      0
## # i abbreviated names:
## #   1: `Other renewables (including geothermal and biomass) electricity generation - TWh`,
## #   2: `Solar generation - TWh`
## # i 2 more variables: `Wind generation - TWh` <dbl>,
## #   `Hydro generation - TWh` <dbl>
```

```
str(energy_data)
```

```
## spc_tbl_ [5,695 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Entity                                                                         : chr [1:5695] "A
##  $ Code                                                                           : chr [1:5695] NA
##  $ Year                                                                           : num [1:5695] 19
##  $ Other renewables (including geothermal and biomass) electricity generation - TWh: num [1:5695] 0.
##  $ Solar generation - TWh                                                         : num [1:5695] 0
##  $ Wind generation - TWh                                                          : num [1:5695] 0
##  $ Hydro generation - TWh                                                         : num [1:5695] 26
##  - attr(*, "spec")=
##   .. cols(
##   ..   Entity = col_character(),
##   ..   Code = col_character(),
##   ..   Year = col_double(),
##   ..   `Other renewables (including geothermal and biomass) electricity generation - TWh` = col_doubl
##   ..   `Solar generation - TWh` = col_double(),
##   ..   `Wind generation - TWh` = col_double(),
##   ..   `Hydro generation - TWh` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

I performed data cleaning and feature selection from my original data.

```
energy_data <- energy_data %>%dplyr:: rename(Other_Sources = `Other renewables (including geothermal an
  select(-Code)
selected_countries<- c('Algeria', 'Egypt', 'Argentina')
sorted_data <- energy_data %>%
  filter(Entity %in% selected_countries) %>%
  select(Entity, Year, Other_Sources, `Solar generation - TWh`, `Wind generation - TWh`, `Hydro generat

# now removing the missing data in the sorted data.
```

```r
sorted_data<- na.omit(sorted_data)
write.csv(sorted_data, "sorted_data.csv", row.names = FALSE)
```

Now that i have the dataset to use, I decided to perform exploratory analysis before determining if i need to do a modelling or perform linear regression.

```r
glimpse(sorted_data) # Viewing the features of my dataset
```

```
## Rows: 99
## Columns: 6
## $ Entity                    <chr> "Algeria", "Algeria", "Algeria", "Algeria", "~
## $ Year                      <dbl> 1990, 1991, 1992, 1993, 1994, 1995, 1996, 199~
## $ Other_Sources             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ `Solar generation - TWh` <dbl> 0.00000000, 0.00000000, 0.00000000, 0.0000000~
## $ `Wind generation - TWh`  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.0~
## $ `Hydro generation - TWh` <dbl> 0.1350, 0.2930, 0.1990, 0.3530, 0.1660, 0.193~
```

```r
#perfomed statistical summary.
summary(sorted_data)
```

```
##     Entity               Year      Other_Sources    Solar generation - TWh
##  Length:99          Min.   :1990   Min.   :0.0000   Min.   :0.000000
##  Class :character   1st Qu.:1998   1st Qu.:0.0000   1st Qu.:0.000000
##  Mode  :character   Median :2006   Median :0.0000   Median :0.000109
##                     Mean   :2006   Mean   :0.3139   Mean   :0.304072
##                     3rd Qu.:2014   3rd Qu.:0.2150   3rd Qu.:0.035300
##                     Max.   :2022   Max.   :2.3409   Max.   :5.077189
##  Wind generation - TWh Hydro generation - TWh
##  Min.   : 0.0000       Min.   : 0.0093
##  1st Qu.: 0.0000       1st Qu.: 0.2580
##  Median : 0.0190       Median :12.9900
##  Mean   : 0.8984       Mean   :13.8547
##  3rd Qu.: 0.5703       3rd Qu.:23.8289
##  Max.   :14.1645       Max.   :38.0152
```

from the summary ambove , the dataset has 99 entries and 6 columns. The year range from 1990 to 2022, Other_Sources: Has a mean of ~0.31 with a standard deviation of ~0.59, indicating some variability and the presence of higher values since the max is 2.34. The mean solar generation is ~0.30 with a standard deviation of ~0.92. The maximum value is significantly higher (5.08) compared to the 75th percentile (0.04), suggesting a right-skewed distribution. Mean wind generation is ~0.90 with a wide range (std = ~2.37), maxing out at ~14.16 TWh. This also suggests variability and potential outliers on the higher end. Hydro generation - TWh: This is the largest source of renewable energy in your dataset, with a mean generation of ~13.85 and a maximum of ~38.02 TWh.

~I plotted heatmap to viausalize the corelation of these variables.

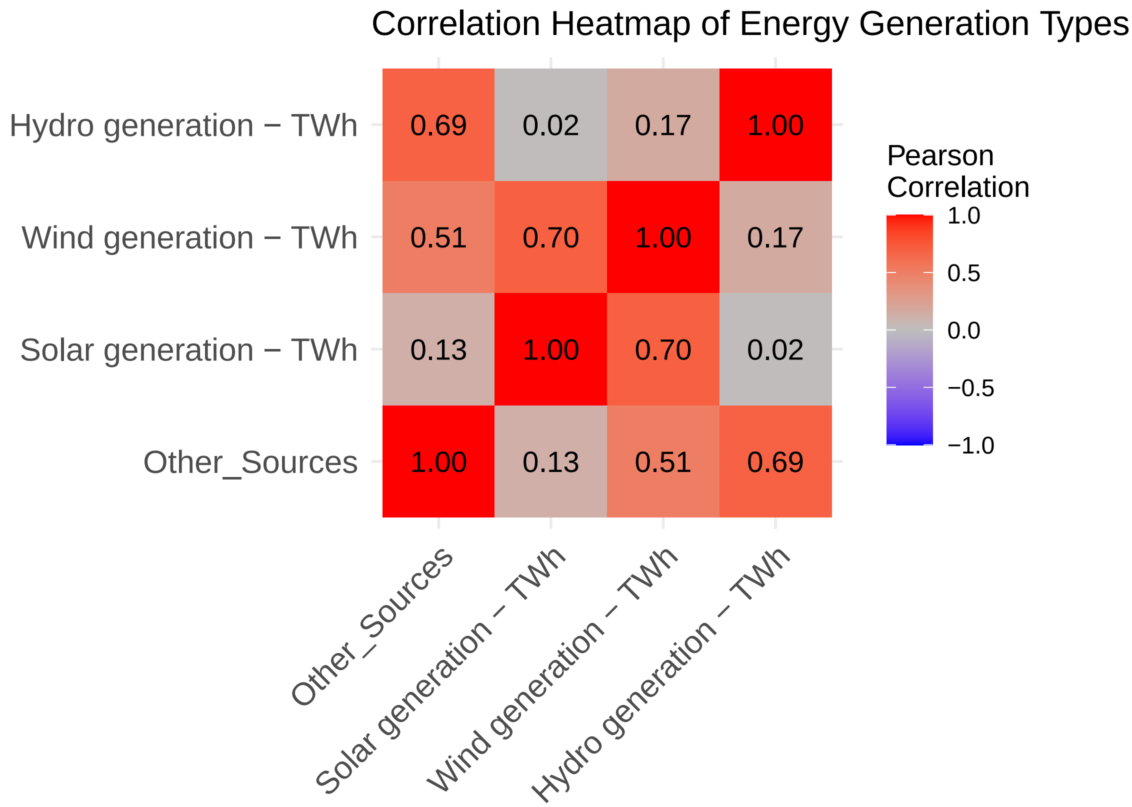
```r
#computing the corelation matrix
correlation_matrix <- cor(sorted_data[,c('Other_Sources', 'Solar generation - TWh', 'Wind generation - '

# Melt the correlation matrix for use with ggplot2
melted_correlation_matrix <- melt(correlation_matrix)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE
```

```r
melted_correlation_matrix
```

```
##                       X1                    X2      value
## 1           Other_Sources         Other_Sources 1.00000000
## 2  Solar generation - TWh         Other_Sources 0.12718627
## 3   Wind generation - TWh         Other_Sources 0.50720916
## 4  Hydro generation - TWh         Other_Sources 0.69004389
## 5           Other_Sources Solar generation - TWh 0.12718627
## 6  Solar generation - TWh Solar generation - TWh 1.00000000
## 7   Wind generation - TWh Solar generation - TWh 0.69510600
## 8  Hydro generation - TWh Solar generation - TWh 0.01542527
## 9           Other_Sources  Wind generation - TWh 0.50720916
## 10 Solar generation - TWh  Wind generation - TWh 0.69510600
## 11  Wind generation - TWh  Wind generation - TWh 1.00000000
## 12 Hydro generation - TWh  Wind generation - TWh 0.16633116
## 13          Other_Sources Hydro generation - TWh 0.69004389
## 14 Solar generation - TWh Hydro generation - TWh 0.01542527
## 15  Wind generation - TWh Hydro generation - TWh 0.16633116
## 16 Hydro generation - TWh Hydro generation - TWh 1.00000000
```

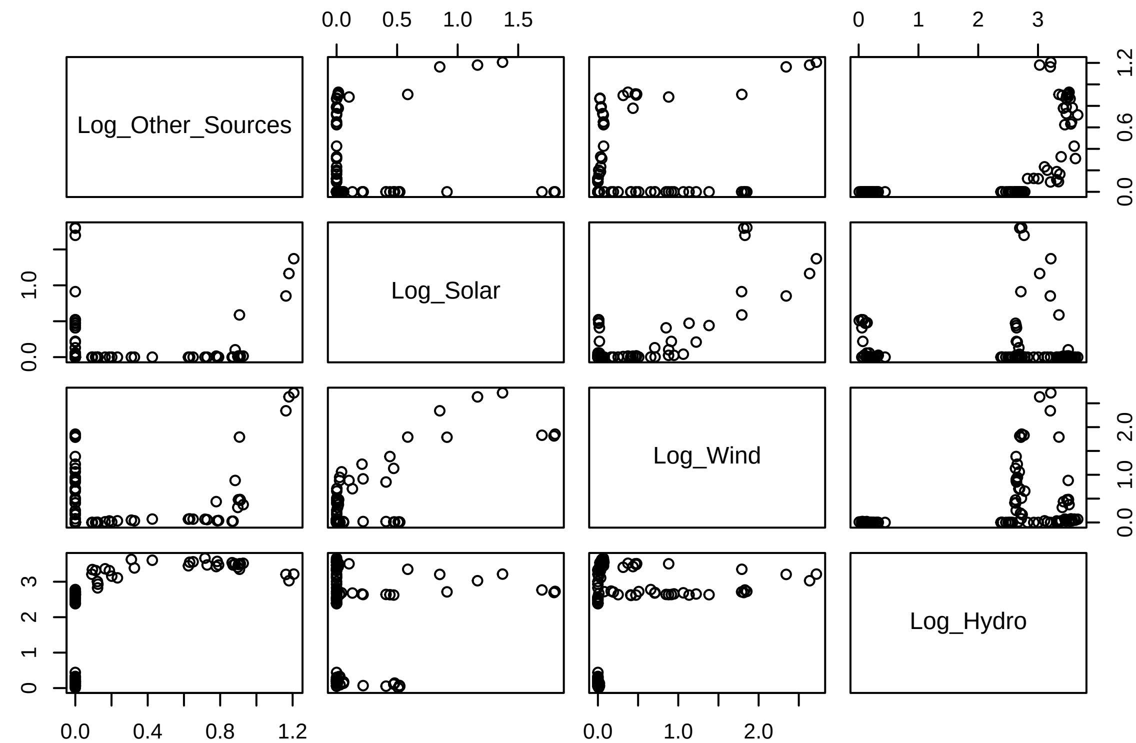
```r
# Ctreating a Heatmap using the ggplot.

ggplot(data = melted_correlation_matrix, aes(X1,X2)) +
  geom_tile(aes(fill = value)) +
  geom_text(aes(label = sprintf("%.2f", value)), color = "black", size = 4) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "grey", midpoint = 0, limit = c(-1,1), space =
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, size = 12, hjust = 1),
        axis.text.y = element_text(size = 12)) +
  labs(title = "Correlation Heatmap of Energy Generation Types", x = "", y = "") +
  coord_fixed()
```



from the above heatmap, we can clearly see that there's a moderate to low positive correlation among different types of renewable energy generation, with the strongest correlation being between Solar and Wind generation (0.61). Other Sources show a moderate positive correlation with Solar (0.49) and Wind (0.47) generation, and a weaker correlation with Hydro generation (0.26). Hydro generation shows relatively lower correlations with other renewable sources, which might be due to its broader use and availability compared to newer technologies like wind and solar.

performed log transformation to further explore the relationships of my variables.

```
sorted_data$Log_Other_Sources <- log(sorted_data$Other_Sources + 1)
sorted_data$Log_Solar <- log(sorted_data$`Solar generation - TWh` + 1)
sorted_data$Log_Wind <- log(sorted_data$`Wind generation - TWh` + 1)
sorted_data$Log_Hydro <- log(sorted_data$`Hydro generation - TWh` + 1)

pairs(~ Log_Other_Sources + Log_Solar + Log_Wind + Log_Hydro, data = sorted_data)
```




```r
fit <- lm(Log_Solar ~ Year + Log_Other_Sources + Log_Wind + Log_Hydro, data = sorted_data)
summary(fit)
```

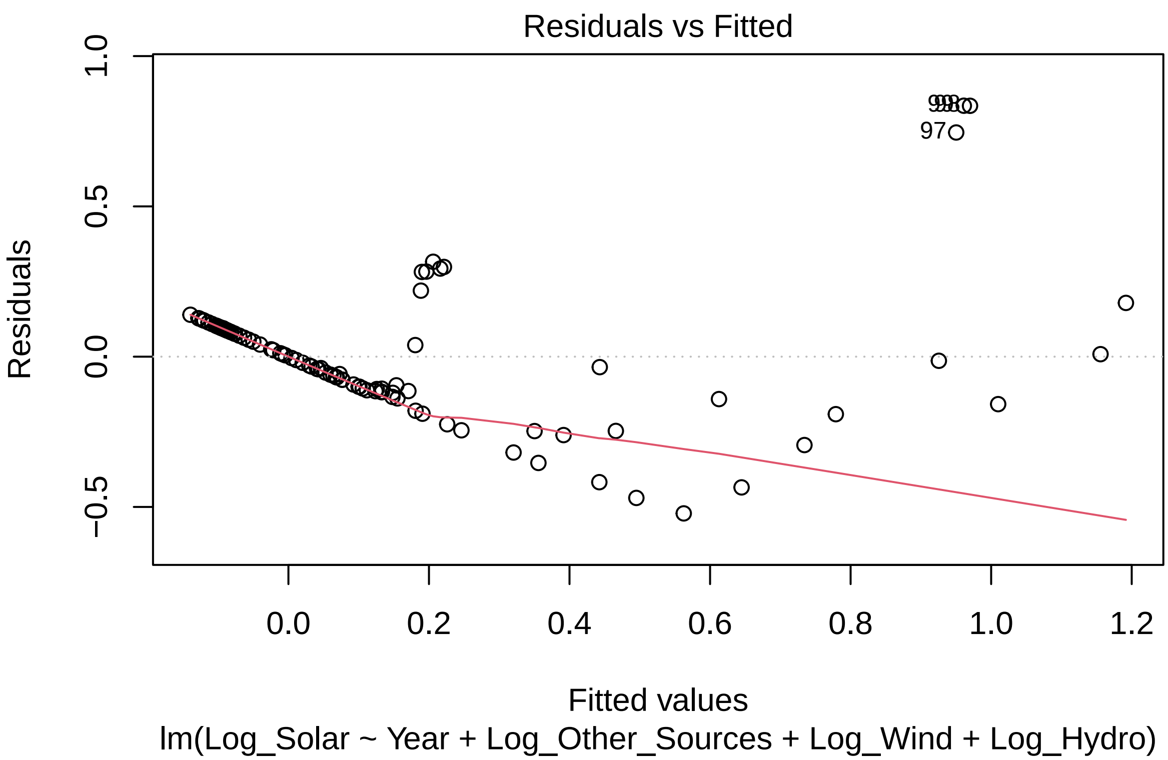
```
##
## Call:
## lm(formula = Log_Solar ~ Year + Log_Other_Sources + Log_Wind +
##     Log_Hydro, data = sorted_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52141 -0.10974 -0.00449  0.09447  0.83487
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -14.894002   6.447667  -2.310   0.0231 *
## Year                0.007474   0.003214   2.326   0.0222 *
## Log_Other_Sources  -0.138651   0.083213  -1.666   0.0990 .
## Log_Wind            0.460438   0.049688   9.267 6.67e-15 ***
## Log_Hydro          -0.034842   0.021635  -1.610   0.1107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2215 on 94 degrees of freedom
## Multiple R-squared:  0.6752, Adjusted R-squared:  0.6614
## F-statistic: 48.85 on 4 and 94 DF,  p-value: < 2.2e-16
```

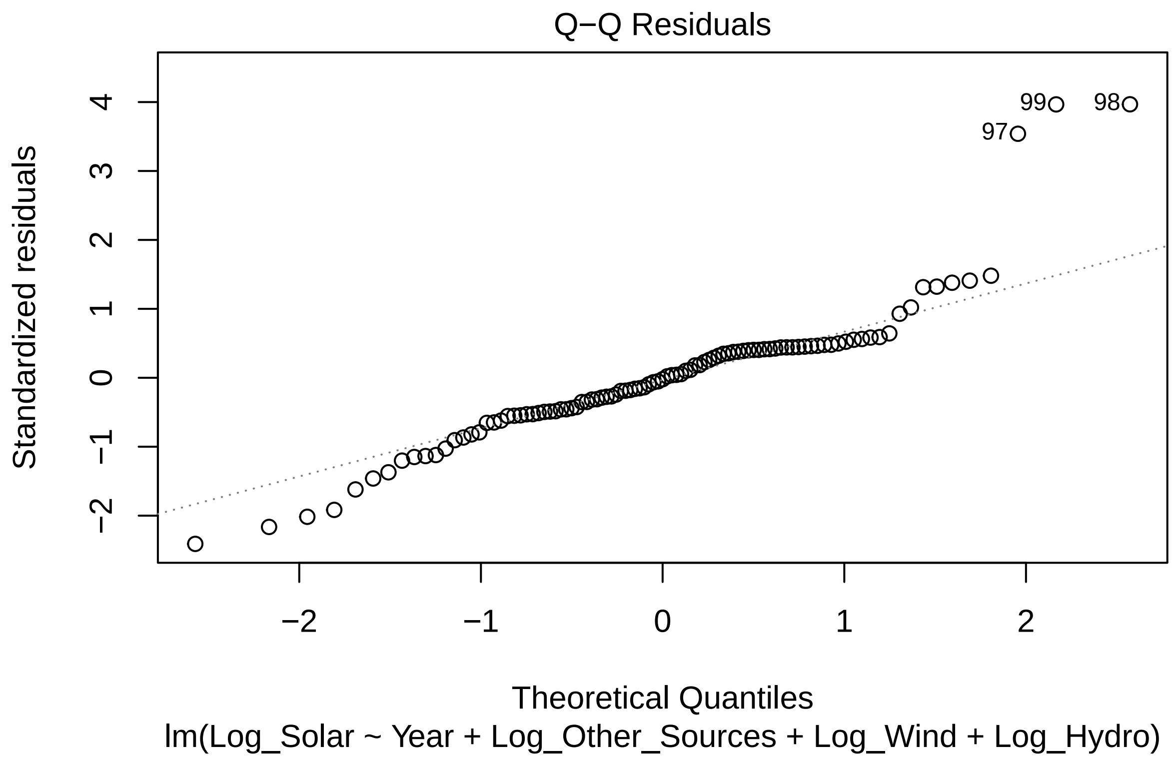
Objective 1: Probability theory provides the language and tools for describing and analyzing

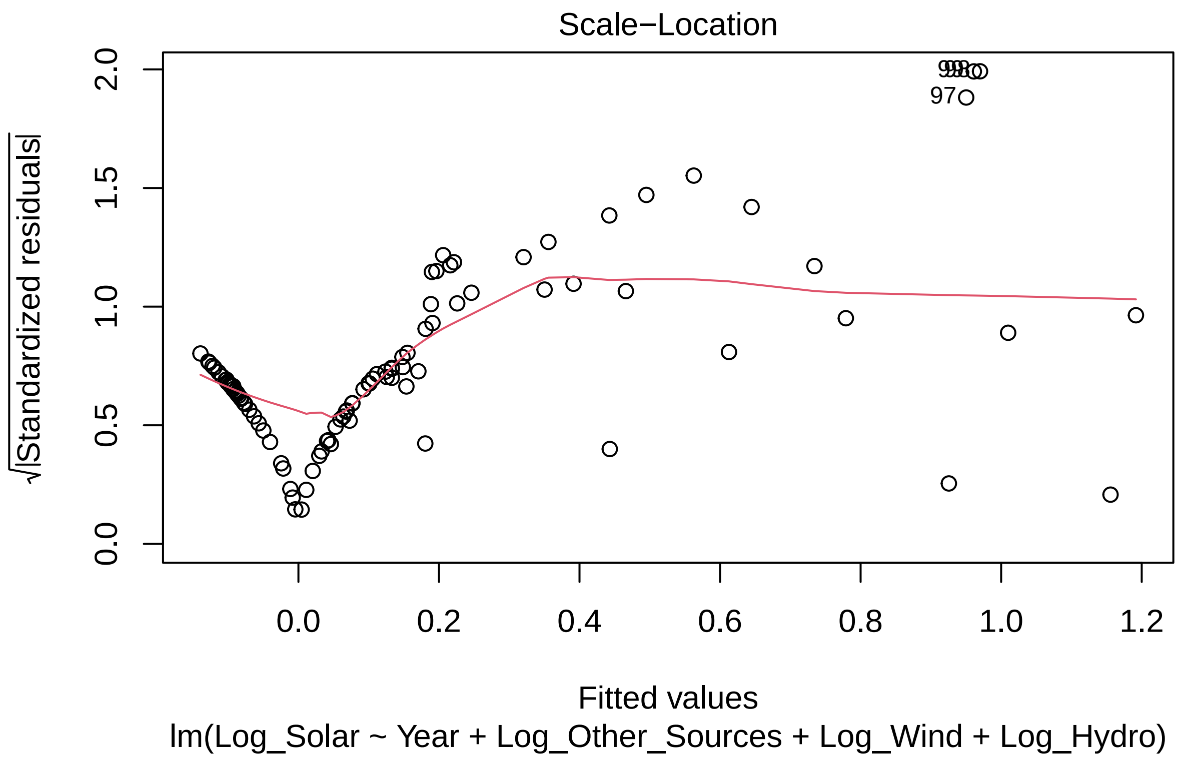
the randomness inherent in the data and the models used to study them. in statistical models are mathematical representation of the real world which is full of randomness. By assuming certain measures in our data we can therefore be in a position to chose models, perform hypothesis testing and regression. In my data set i have performed several processes such as data cleaning, normalization, tranformation to be able to perform statistical analysis. MLE is a method used for estimating the parameters of a statistical model. It is based on the principle of selecting the parameter values that maximize the likelihood of the observed data under the model

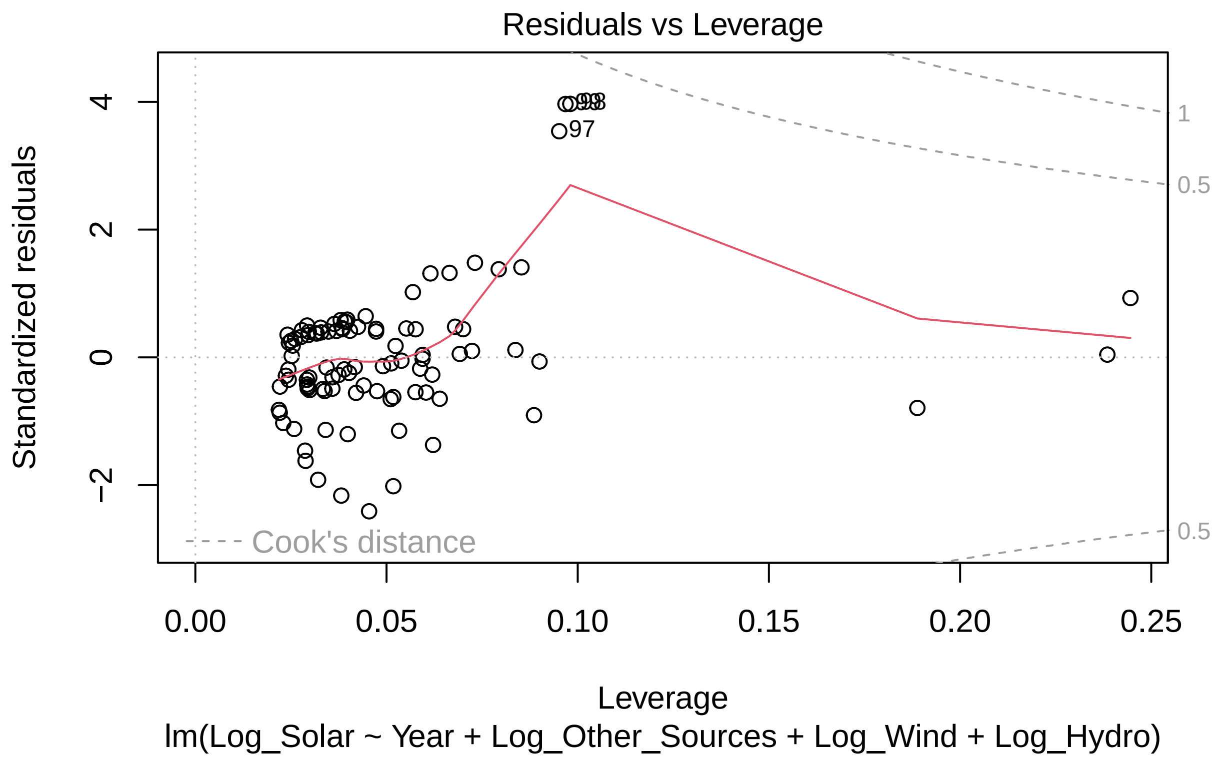
I am going to run a linear model to see if there is any variables that best fits the model.

```
fit <- lm(Log_Solar ~ Year + Log_Other_Sources + Log_Wind + Log_Hydro, data = sorted_data)
plot(fit)
```



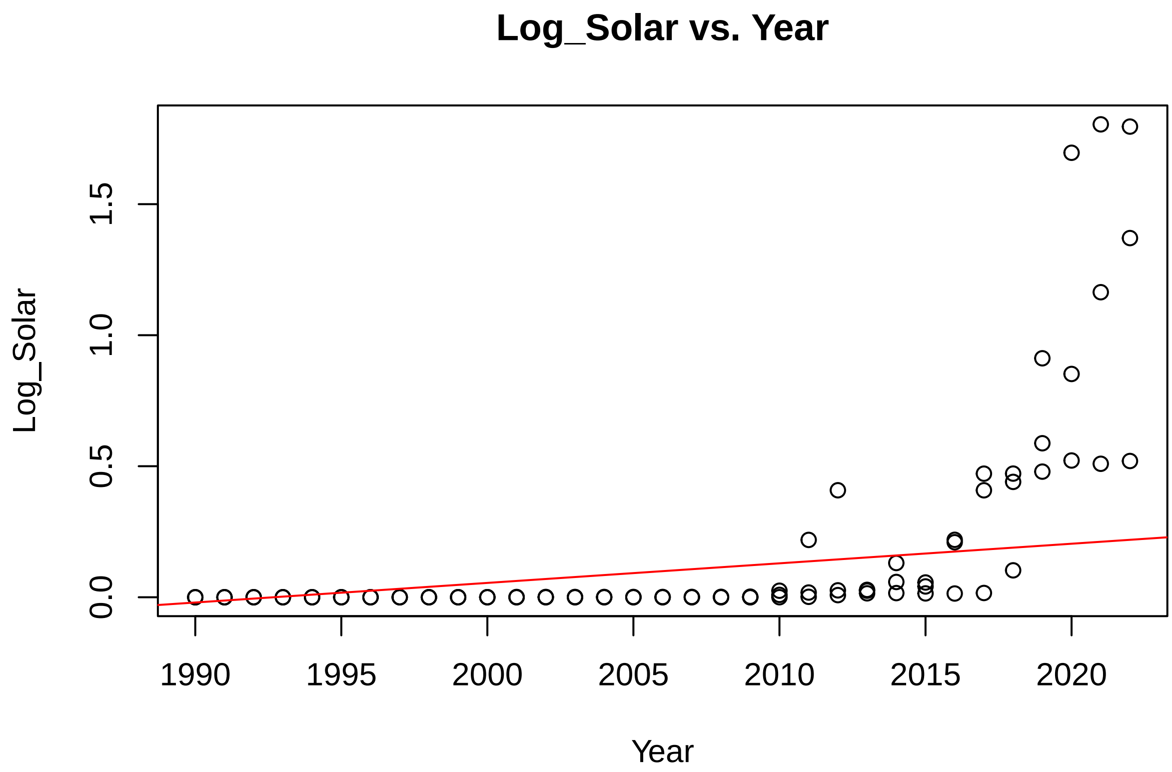

Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Log_Solar ~ Year + Log_Other_Sources + Log_Wind + Log_Hydro)
99
98
97
4
3
2
1
0
−1
−2
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
97
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Fitted values
lm(Log_Solar ~ Year + Log_Other_Sources + Log_Wind + Log_Hydro)




```
plot(sorted_data$Year, sorted_data$Log_Solar, main = "Log_Solar vs. Year", xlab = "Year", ylab = "Log_So
abline(fit, col = "red")
```

```
## Warning in abline(fit, col = "red"): only using the first two of 5 regression
## coefficients
```



From the Linear model above all the assumptions for a lineat regression model are not met except for linearlity of residuals. the scatterplot shows the distribution of transformed values and therefore this is not the best model to use.

Objective 2: Determine and apply the appropriate generalized linear model for a specific data context.

Most of my variables were not linearly correlated and therefore i decided to use the Genaralized Additive Model(GAM) as the model of choice.

```
library(mgcv)
```

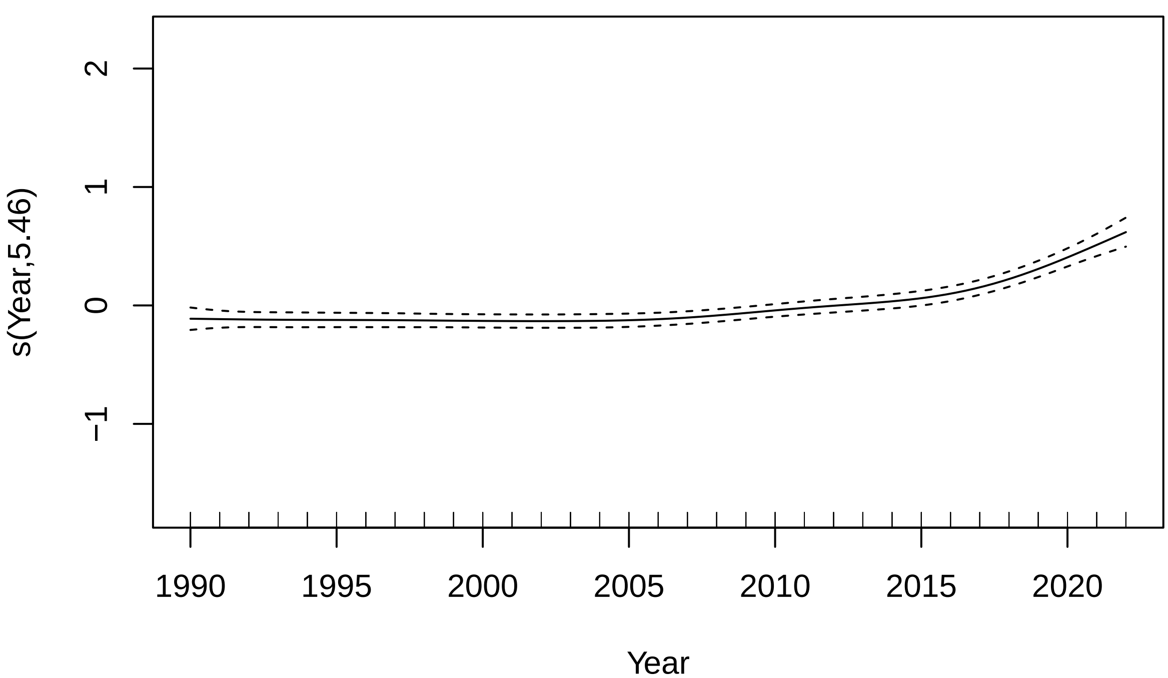
```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.9-0. For overview type 'help("mgcv-package")'.
```

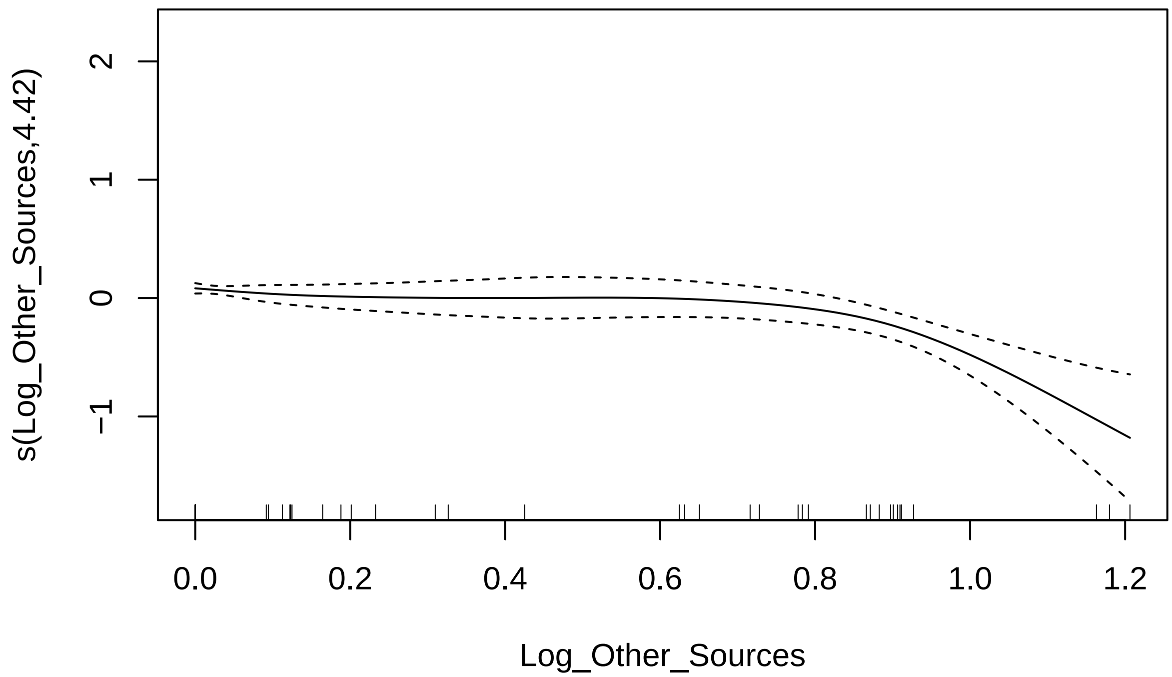
```
gam_fit <- gam(Log_Solar ~ s(Year) + s(Log_Other_Sources) + s(Log_Wind) + s(Log_Hydro), data = sorted_da
summary(gam_fit)
```

```
##
## Family: gaussian
## Link function: identity
##
```

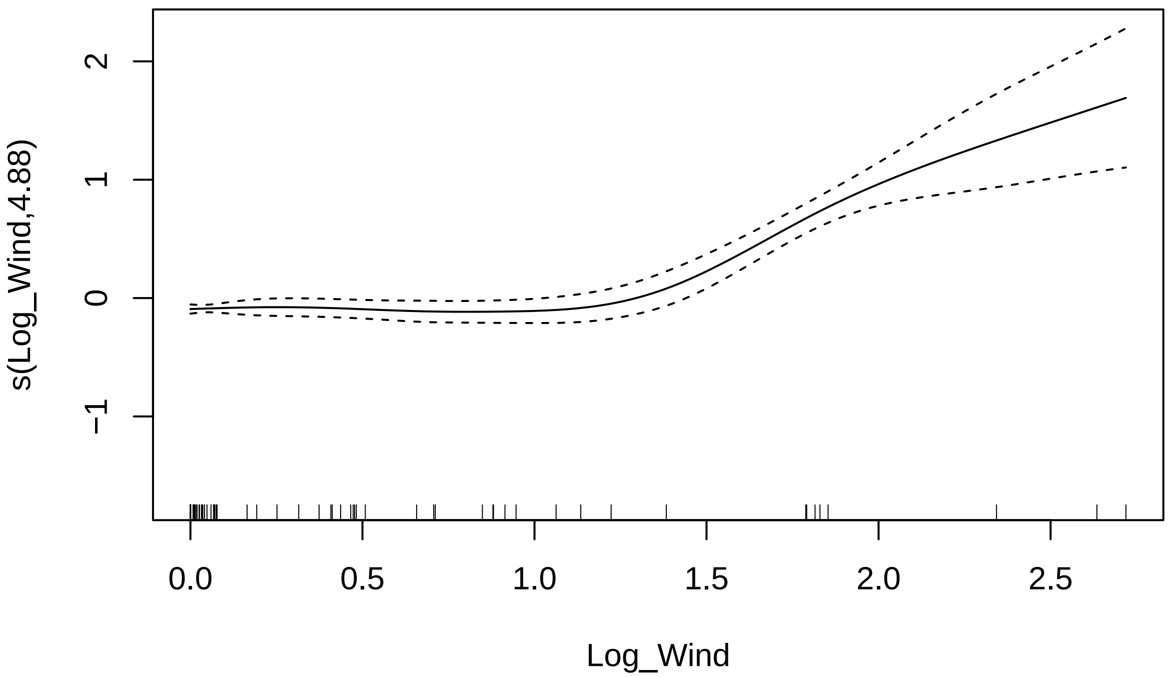
```
## Formula:
## Log_Solar ~ s(Year) + s(Log_Other_Sources) + s(Log_Wind) + s(Log_Hydro)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15837    0.01051   15.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                        edf Ref.df      F  p-value
## s(Year)              5.462  6.580 20.157  < 2e-16 ***
## s(Log_Other_Sources) 4.417  5.303  6.410 3.46e-05 ***
## s(Log_Wind)          4.879  5.809 26.563  < 2e-16 ***
## s(Log_Hydro)         1.462  1.769  1.286    0.322
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.924   Deviance explained = 93.7%
## GCV = 0.013243  Scale est. = 0.01094   n = 99
```

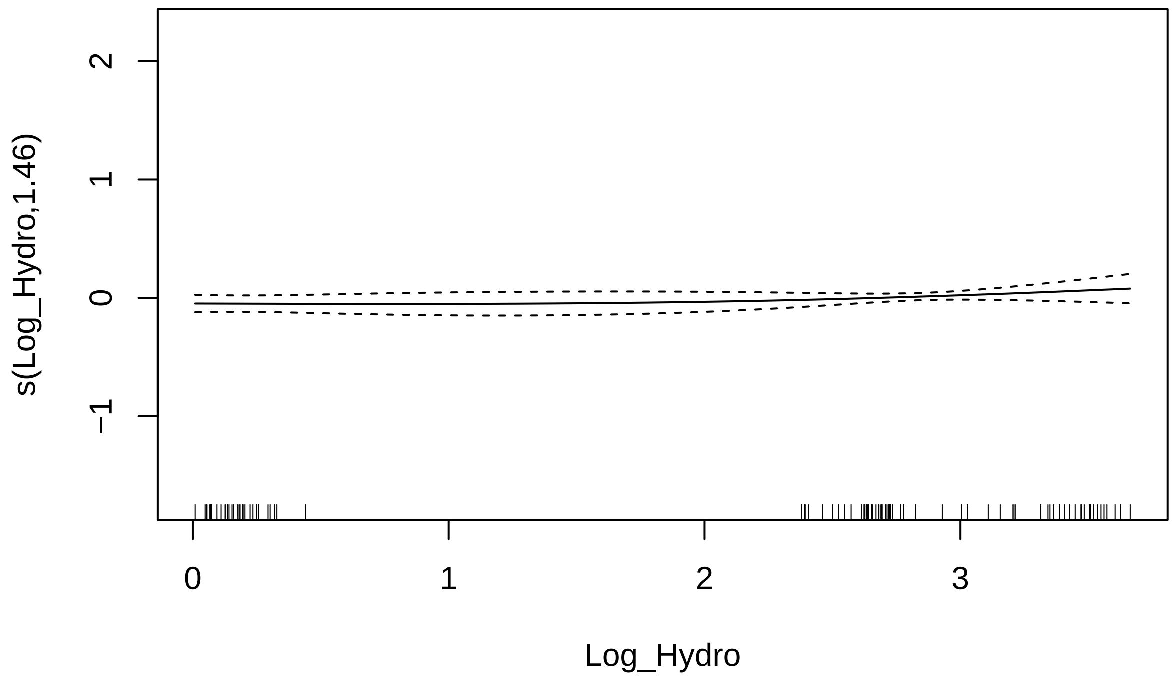
```
plot(gam_fit)
```





2
1
0
−1
s(Log_Other_Sources,4.42)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Log_Other_Sources


s(Log_Wind,4.88)
2
1
0
−1
0.0
0.5
1.0
1.5
2.0
2.5
Log_Wind



Objective 4: Communicate the results of statistical models to a general audience

From the generalized additive model, The relationship between the predictors and the response variable is linear, meaning the expected value of Log_Solar is directly modeled as a linear combination of the predictors. Here the predictors are all the for variables. from the summary, 0.924 proportion of the solar energy is explained by other predictors after the adjustments. the R-squared of 93.7% suggests the model explains a large portion of the variance in the data.

overall the model captures the relationship between Log_Solar and all the predictors, particularly for Year, Log_Other_Sources, and Log_Wind, through non-linear smooth functions, explaining a significant portion of the variance in Log_Solar. The non-significant relationship with Log_Hydro suggests it may not be as important in predicting Log_Solar as the other variables in this model.