Institute of Digital
Technology Management

PGDM in Big Data Analytics
Term II (Batch 2023-25)

Course Name: Statistics for Data Analysts
Course Code: 210C208

Statistical Analysis of World's Energy Consumption

Submitted to: Ms. Manjari Mundanad

Submitted by:

Name	Enrolment Number
Chirag Malhotra	20231013

Dated: 18th January 2024

CONTENTS

	HEADLINE	PAGE NO
1	INTRODUCTION	1
2	DESCRIPTIVE STATISTICS	2
3	CORRELATION	3
4	COVARIANCE	4
5	REGRESSION	5
6	COEFFICIENT OF VARIANCE	6
7	IQR ANALYSIS	7
8	IMPORTANT GRAPHS	8-10
9	INSIGHTS	11-12
10	CONCLUSION	14

Introduction

Energy use is essential to every culture in the modern world. It is essential to comprehend how we use this valuable resource if we are to ensure sustainable growth and a safe future. This study takes a deep dive into the complex realm of energy use, carefully examining data using a variety of methods. Using Python, EViews software, and Excel's variety of tools, we harness the power of statistical analysis to set out on a quest to find significant insights.

From the complex interweave of data points, not just a compilation of numbers. We use a range of statistical methods, such as regressions and visualizations, to reveal the correlations and hidden patterns within the dataset on energy usage.

We cover a wide range of energy sources, from the conventional fossil fuel powerhouse to the emerging renewable energy frontiers. Our goal is to provide a more comprehensive understanding of our energy use with every analysis by pinpointing important factors, possible inefficiencies, and areas for improvement.

DESCRIPTIVE STATISTICS

	TOTAL	WIND_T...	TRADITION...	SOLAR_T...	OTHER_R...	OIL_TWH_	NUCLEAR_...	HYDROPO...	GAS_TW...	COAL_T...	BIOFUELS...
Mean	134399.0	1035.589	11517.03	416.8790	944.9434	44221.08	6557.433	8405.748	26969.87	33895.80	434.6056
Median	130218.0	183.5335	11441.00	6.541638	674.7783	44652.22	6776.866	7826.347	25729.62	31506.16	178.3269
Maximum	178898.7	5487.600	12500.00	3448.237	2413.808	53512.84	7653.722	11448.03	40670.66	44858.12	1199.207
Minimum	92608.64	0.132342	10430.00	0.018662	236.3207	33667.10	3559.857	5740.620	15902.68	23001.08	57.80642
Std. Dev.	26843.24	1524.442	572.5088	840.0116	659.4545	6101.688	973.9238	1831.112	7527.675	8330.425	396.1006
Skewness	0.122082	1.510848	0.130424	2.253488	0.838329	-0.146053	-1.359133	0.245920	0.268844	0.146404	0.677832
Kurtosis	1.628538	4.160472	2.052576	7.231339	2.416286	1.780198	4.416477	1.737999	1.811717	1.226562	1.811601
Jarque-Bera	3.153353	17.02568	1.569189	62.10271	5.121844	2.556519	15.26749	2.981149	2.764327	5.250081	5.281440
Probability	0.206661	0.000201	0.456305	0.000000	0.077233	0.278522	0.000484	0.225243	0.251035	0.072437	0.071310
Sum	5241560.	40387.97	449164.0	16258.28	36852.79	1724622.	255739.9	327824.2	1051825.	1321936.	16949.62
Sum Sq. Dev.	2.74E+10	88309092	12455119	26813541	16525450	1.41E+09	36044048	1.27E+08	2.15E+09	2.64E+09	5962036.
Observations	39	39	39	39	39	39	39	39	39	39	39

Summarised the data with the help of e-views.From the figure we can get the Mean,Median,Std Deviation, Skewness,Kurtosis etc.

- The table shows descriptive statistics for a dataset of 39 observations.
- The value **26,843.24** indicates a moderate spread of the data around the mean also the data distribution has a little right skew, as indicated by 0.122. This indicates that the tail is longer toward the higher values and that there are more data points clustered near the lower end of the range.
- The skewness of a distribution that is exactly symmetrical would be 0. A right skew is indicated by a positive skewness number, and a left skew is shown by a negative value.
- The kurtosis value is 1.62 that is it is less than the idle value which is 3 so we can say that the data is platokurtic.
- Taking the case of JB test the p-value is 0.2066 that is much greater than 0.05v that is we accept H1 that says the data is normal.

CORRELATION

	WIND_T...	TRADITION...	SOLAR_T...	OTHER R...	OIL_TWH_	NUCLEAR_...	HYDROPO...	GAS_TW...	COAL_T...	BIOFUELS...
WIND...	1.000000	-0.334839	0.963798	0.963565	0.766033	0.219221	0.870094	0.879957	0.772541	0.933447
TRADI...	-0.334839	1.000000	-0.332801	-0.193826	0.148600	0.779891	-0.020328	-0.016105	-0.085366	-0.278649
SOLA...	0.963798	-0.332801	1.000000	0.868335	0.619435	0.142963	0.733427	0.752102	0.603313	0.809257
OTHE...	0.963565	-0.193826	0.868335	1.000000	0.898876	0.394303	0.967187	0.972186	0.900403	0.983294
OIL_...	0.766033	0.148600	0.619435	0.898876	1.000000	0.685933	0.962184	0.969698	0.938615	0.881156
NUCL...	0.219221	0.779891	0.142963	0.394303	0.685933	1.000000	0.547810	0.563809	0.498738	0.323386
HYDR...	0.870094	-0.020328	0.733427	0.967187	0.962184	0.547810	1.000000	0.992176	0.945459	0.959900
GAS_...	0.879957	-0.016105	0.752102	0.972186	0.969698	0.563809	0.992176	1.000000	0.951622	0.957615
COAL...	0.772541	-0.085366	0.603313	0.900403	0.938615	0.498738	0.945459	0.951622	1.000000	0.927071
BIOFU...	0.933447	-0.278649	0.809257	0.983294	0.881156	0.323386	0.959900	0.957615	0.927071	1.000000

The strength of the connection or correlation between two variables. It calculates the likelihood that two variables will change in tandem, both in terms of magnitude and direction. Although it doesn't necessarily suggest a cause-and-effect relationship, it does show a dependence.

- There is a significant positive association between wind energy and other renewable energy sources. Wind power and solar power have a connection coefficient of 0.964, while wind power and other renewable energy sources have a correlation coefficient of 0.963.
- There is a negative relationship between wind power and conventional energy sources. Wind power and oil have a connection coefficient of -0.335, whereas wind power and nuclear power have a correlation coefficient of -0.219. This implies that the output of oil and nuclear power is likely to be low during periods of significant wind power production.
- There is a significant positive association between hydropower and conventional and renewable energy sources. The hydropower and solar power correlation coefficient is 0.733; the hydropower and other renewable energy sources correlation coefficient is 0.967; the hydropower and oil correlation coefficient is 0.962; and the hydropower and nuclear power correlation coefficient is 0.548.
- This means that when other renewable energy sources and conventional energy sources are produced at high levels, hydropower production tends to follow suit. There is a significant positive link between gas power and coal power. The correlation value of 0.952 exists between gas and coal

power. This implies that there is a good chance of high gas power production along with high coal power output.

COVARIANCE

	WIND_T...	TRADITION...	SOLAR_T...	OTHER R...	OIL_TWH_	NUCLEAR_...	HYDROPO...	GAS_TW...	COAL_T...	BIOFUELS...
WIND...	2264336.	-284739.4	1202544.	943834.6	6942683.	317129.6	2366524.	9839033.	9559131.	549193.4
TRADI...	-284739.4	319362.0	-155944.6	-71301.42	505788.4	423701.4	-20763.82	-67627.83	-396690.3	-61569.25
SOLA...	1202544.	-155944.6	687526.7	468680.2	3093498.	113960.1	1099199.	4633851.	4113527.	262359.1
OTHE...	943834.6	-71301.42	468680.2	423729.5	3524145.	246751.1	1137967.	4702340.	4819568.	250260.8
OIL_...	6942683.	505788.4	3093498.	3524145.	36275969	3971694.	10474709	43397654	46486144	2075043.
NUCL...	317129.6	423701.4	113960.1	246751.1	3971694.	924206.4	951895.5	4027512.	3942608.	121554.3
HYDR...	2366524.	-20763.82	1099199.	1137967.	10474709	951895.5	3266999.	13325500	14052188	678368.4
GAS_...	9839033.	-67627.83	4633851.	4702340.	43397654	4027512.	13325500	55212924	58144875	2782123.
COAL...	9559131.	-396690.3	4113527.	4819568.	46486144	3942608.	14052188	58144875	67616602	2980606.
BIOFU...	549193.4	-61569.25	262359.1	250260.8	2075043.	121554.3	678368.4	2782123.	2980606.	152872.7

The intensity and direction of the linear relationship between two variables are measured by covariance. Covariance measures the relationship in the same units as the original data, as contrast to correlation, which expresses the relationship as a dimensionless coefficient between -1 and 1.

- Most pairs of energy sources have a primarily positive covariance, according to the covariance table.
- generally the production of other energy sources tends to increase together with the development of one energy source.
- Negative covariances do exist, though, especially when comparing wind energy to more conventional energy sources like nuclear and oil.
- This implies that the production of these conventional sources typically decreases when wind power generation is strong.
- The covariances' magnitudes differ, with certain pairs such as gas and solar exhibiting stronger positive connections than others, such as hydropower and other renewable energy sources.

REGRESSION

Dependent Variable: FIRST_DIFFERENCE
Method: Least Squares
Date: 01/17/24 Time: 12:44
Sample (adjusted): 4 39
Included observations: 36 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3303.676	728.6035	4.534258	0.0001
FIRST_DIFFERENCE(-1)	-0.239676	0.170929	-1.402196	0.1702
FIRST_DIFFERENCE(-2)	-0.226474	0.192956	-1.173710	0.2489

R-squared	0.077736	Mean dependent var	2279.119
Adjusted R-squared	0.021842	S.D. dependent var	2329.636
S.E. of regression	2304.054	Akaike info criterion	18.40238
Sum squared resid	1.75E+08	Schwarz criterion	18.53434
Log likelihood	-328.2429	Hannan-Quinn criter.	18.44844
F-statistic	1.390765	Durbin-Watson stat	1.975136
Prob(F-statistic)	0.263092		

REGRESSION

- the coefficient for FIRST DIFFERENCE (-1) is 3303.676, indicating statistical significance as the p-value is less than 0.0001. This indicates that, on average, a one-unit rise in the FIRST DIFFERENCE from the prior period corresponds to a 3303.676-unit increase from the current period.
- Although the coefficient for FIRST DIFFERENCE (-2) is -226.474, the p-value of 0.2489 indicates that it is not statistically significant.
- A R-squared of 0.077736 suggests that only 7.77% of the variance in the dependent variable can be explained by the model rest is explained by the error. The dependent variable may be influenced by variables other than the independent variables that were included, as indicated by the comparatively low R-squared value.
- Even less than the R-squared value, the adjusted R-squared value of 0.021842, suggests that the model might be overfitting the data. When a model is overly intricate and begins to fit the data's noise instead of the underlying relationship between the variables, this is known as overfitting.

The regression table indicates that the FIRST DIFFERENCE of the present period and the FIRST DIFFERENCE of the preceding period have a statistically significant association. Nevertheless, it appears that the model may be overfitting the data because it only partially explains the variance in the dependent variable. When interpreting the model's findings, it's critical to exercise caution and consider any additional variables that might be affecting the dependent variable.

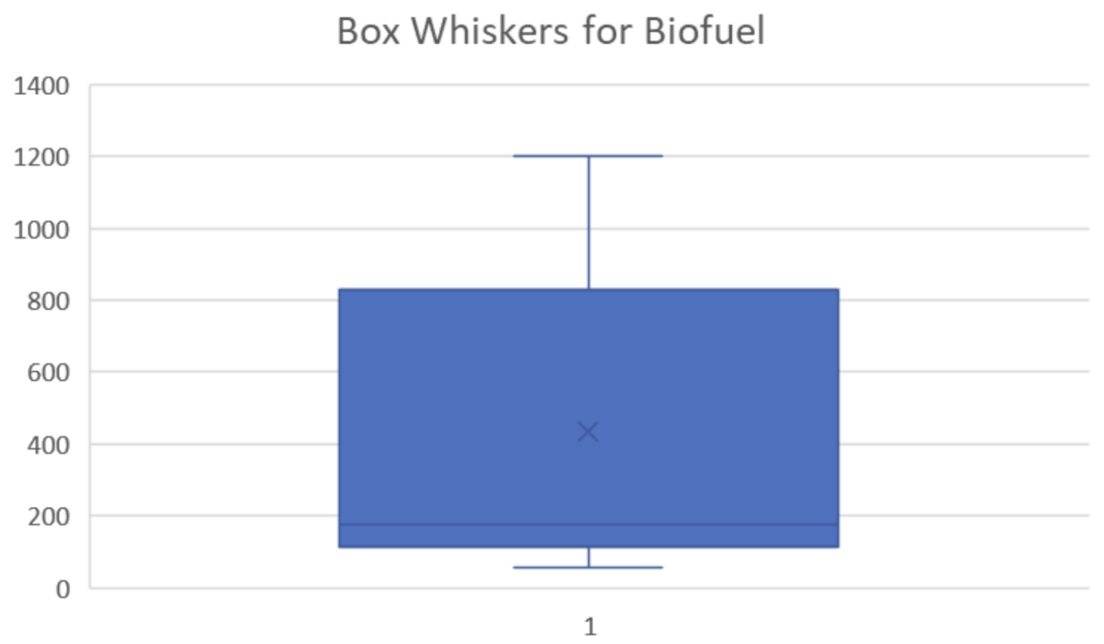
COEFFICIENT OF VARIANCE

Energy Types	Coeff. Of Variance
Other renewables (TWh)	68%
Biofuels (TWh)	91%
Solar (TWh)	201%
Wind (TWh)	147%
Hydropower (TWh)	22%
Nuclear (TWh)	14%
Traditional biomass (TWh)	4%
Gas (TWh)	27%
Oil (TWh)	13%
Coal (TWh)	24%
TWh Total Non-renewables (gasoil & coal)	19%

- The relative variability of different groups or datasets by displaying their coefficient of variation (CV) values. Each row typically represents a specific group or data set, while columns might correspond to different time periods or variables.

IQR ANALYSIS

IQR Analysis		
Q1	9.75	114.4443
Q2	19.5	178.3269
Q3	29.25	828.6506
IQR		714.2063



The distribution of the data set is most likely right-skewed, with a longer tail extending towards higher values and more data points concentrated at the lower end of the range.

The fact that Q3 (29.25) is substantially greater than Q1 (9.75) indicates this.

With an IQR of 714.2063, the data appear to have a substantial gap between the 25th and 75th percentiles.

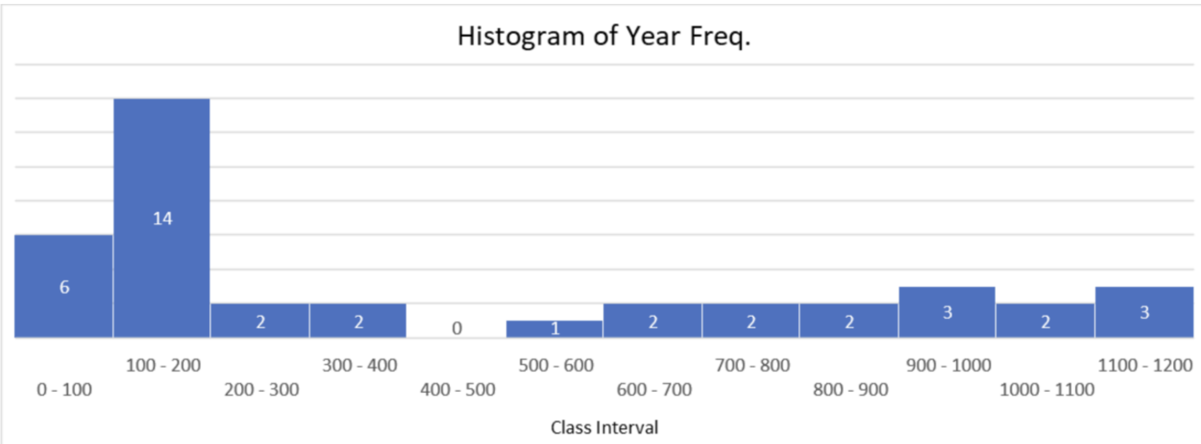
It's hard to evaluate the spread in relation to the whole dataset if you don't know the entire range of values.

IMPORTANT GRAPHS

By deploying Excel, we have plotted other three graphs from which we can analyse our energy conservation data more accurately and efficiently:

- <u>HISTOGRAM OF YEAR FREQUENCY</u>

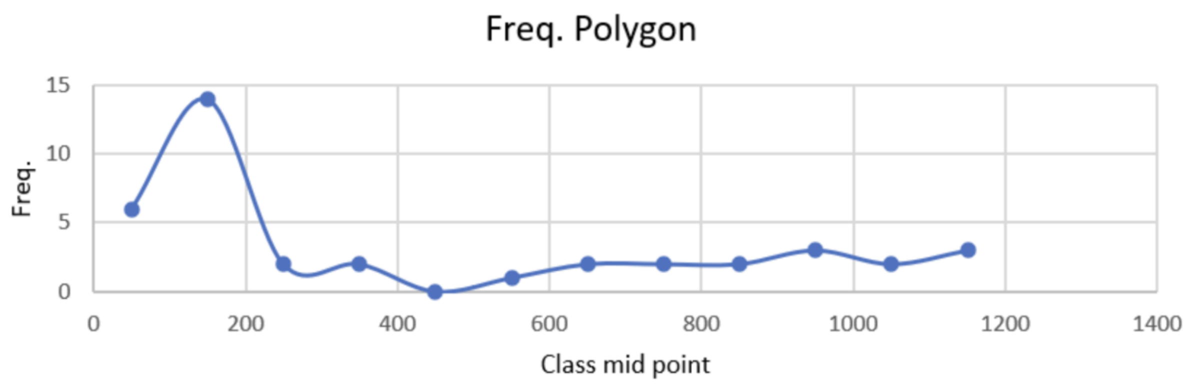
Class Interval	Freq	Cf	Rf
0 - 100	6	6	0.153846154
100 - 200	14	20	0.358974359
200 - 300	2	22	0.051282051
300 - 400	2	24	0.051282051
400 - 500	0	24	0
500 - 600	1	25	0.025641026
600 - 700	2	27	0.051282051
700 - 800	2	29	0.051282051
800 - 900	2	31	0.051282051
900 - 1000	3	34	0.076923077
1000 - 1100	2	36	0.051282051
1100 - 1200	3	39	0.076923077
Total	39		



This graph helps in visualising the data points in this case, which is made of biofuel's data points, by present data points visually by dividing the range of values into "bins" and displaying the frequency of data points within each bin. This creates a visual representation of the data distribution, showing how data points are clustered or spread out across the range.

- FREQUENCY POLYGON

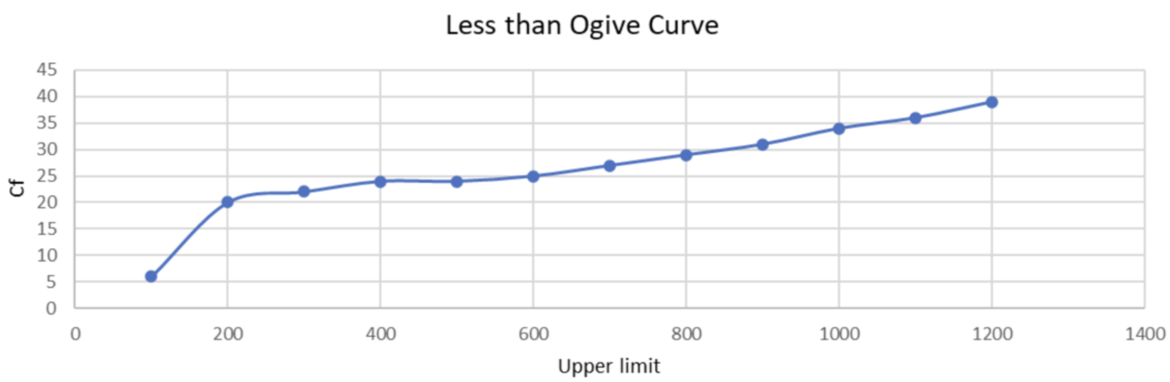
class midpoint	Freq
50	6
150	14
250	2
350	2
450	0
550	1
650	2
750	2
850	2
950	3
1050	2
1150	3



From the fig we can conclude that the trend in frequency depicted with the help of bar graph shows the initial growing stage and then then the deep decrease continued by the static running of biofuel.

- LESS THAN OGIVE CURVE

Upper limit	Cf
100	6
200	20
300	22
400	24
500	24
600	25
700	27
800	29
900	31
1000	34
1100	36
1200	39

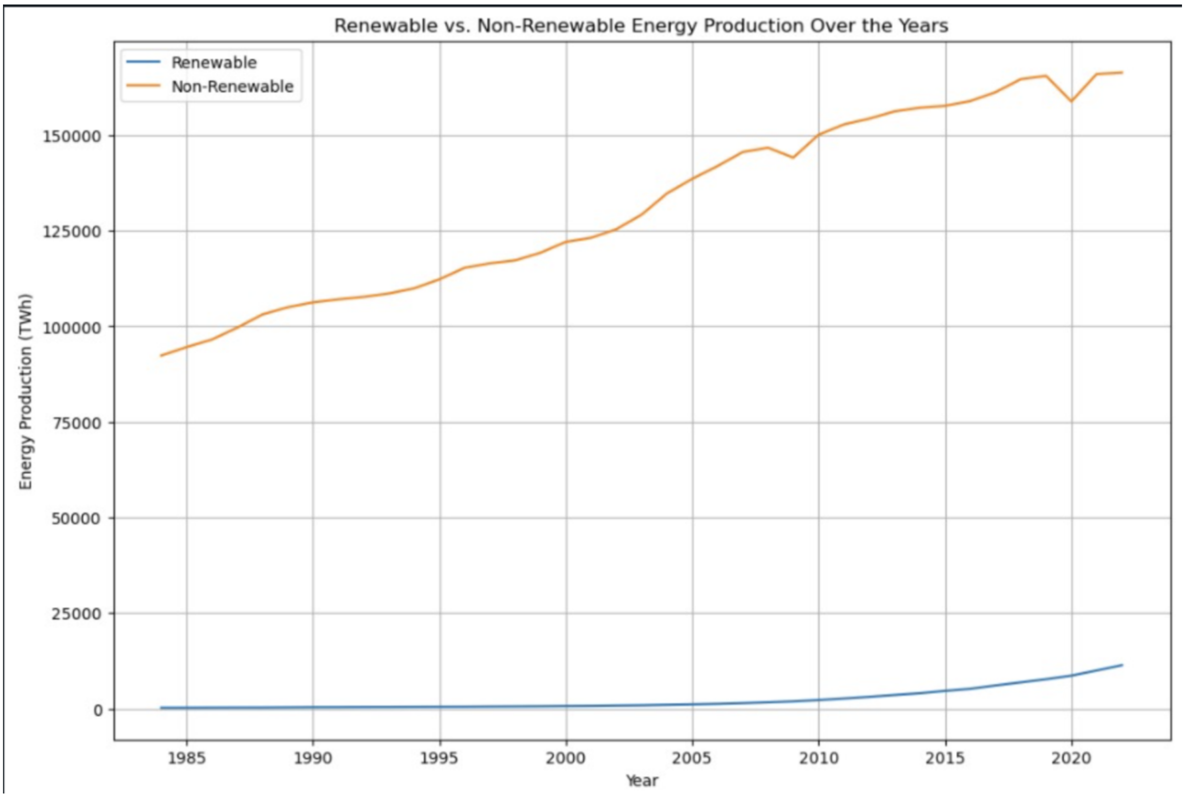


The graphical representation of the cumulative frequencies of a dataset. It visually displays the percentage or proportion of data points that fall below or above a certain value. The data used here is of bio fuels .

.

INSIGHTS

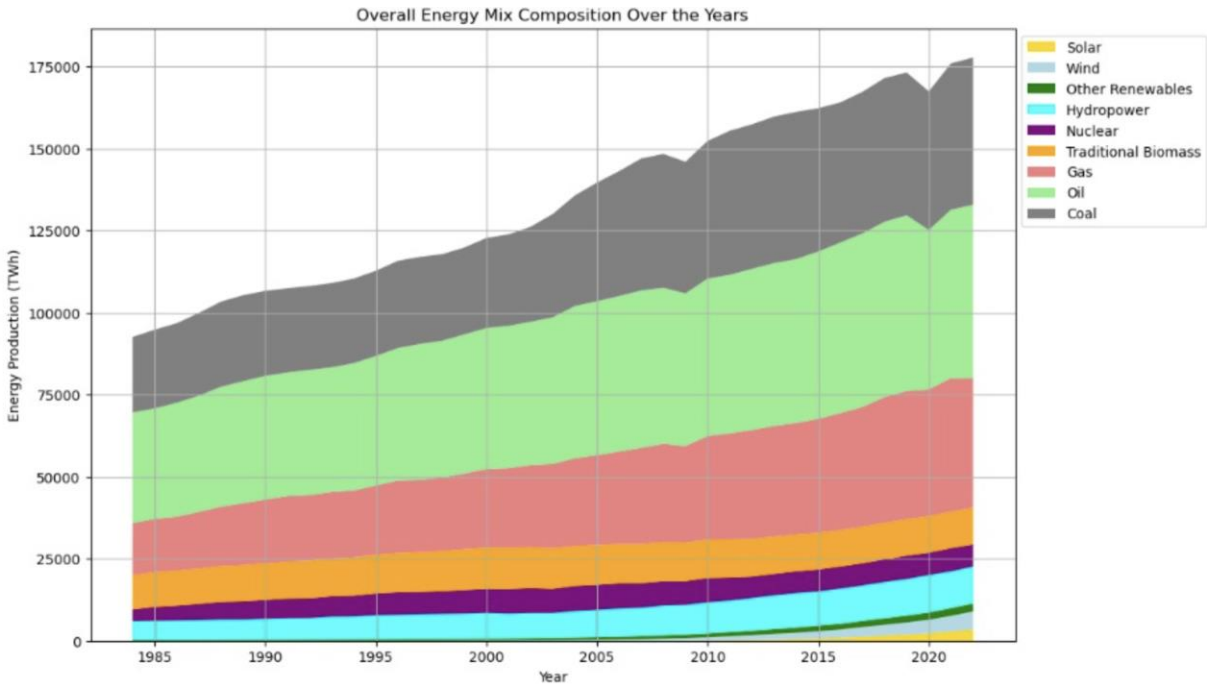
The below insights with the graph is developed with the help of python:



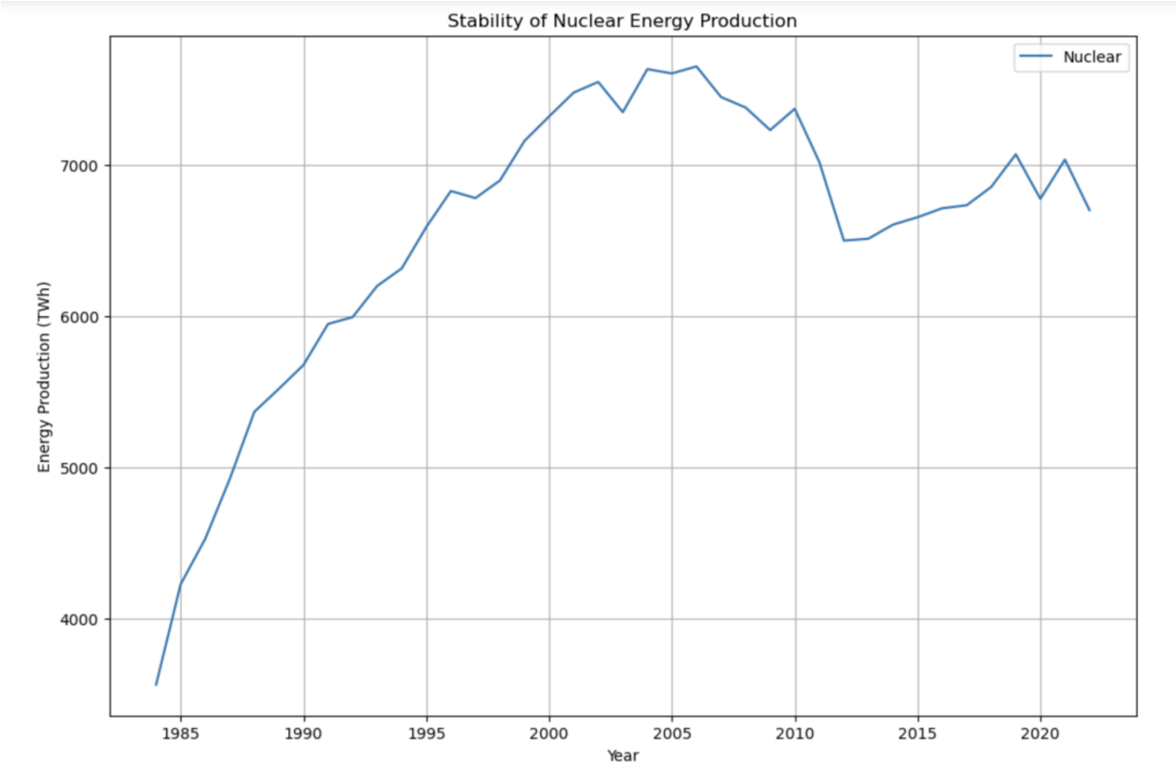
<u>By just observing the trend in the graph we can analyse the following:</u>

- How have renewable and non-renewable energy production changed over time?
- Are there overall increases, decreases, or periods of stagnation in either category?
- Which category has experienced more significant growth or decline?

And here the non-renewable energy has the clear growth as compared to that of renewable energy.



- To the overall energy mic coal is contributing more in comparison with other energy sources and the least is contributed by traditional biomass.
- In comparison the share of renewable energy sources (solar, wind, other renewables, hydropower) evolved less compared to non-renewable sources (nuclear, traditional biomass, gas, oil, coal).
- There is a visible trend in increase and decreased reliance of the sources.



- This graph shows the trend in stability of nuclear energy production.
- From 1985 to 2005 it shows a clear growth in the production
- There is a decline in the production post 2005.
- In 2020 the production of nuclear energy is clearly unstable

CONCLUSION:

From this Regression model we can see that the value for r2 is only 7.7%,and also the P-value for F-statistics is insignificant, which means that the future consumption of energy cannot be predicted on the basis of past or current values of consumption of energy. Also non-renewable energy consumption is much higher than renewable energy consumption, but it has plateaued in recent years, while renewable energy consumption is steadily increasing.