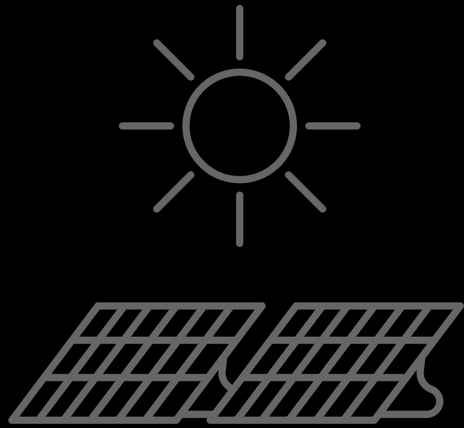
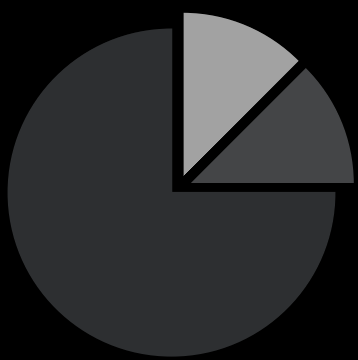
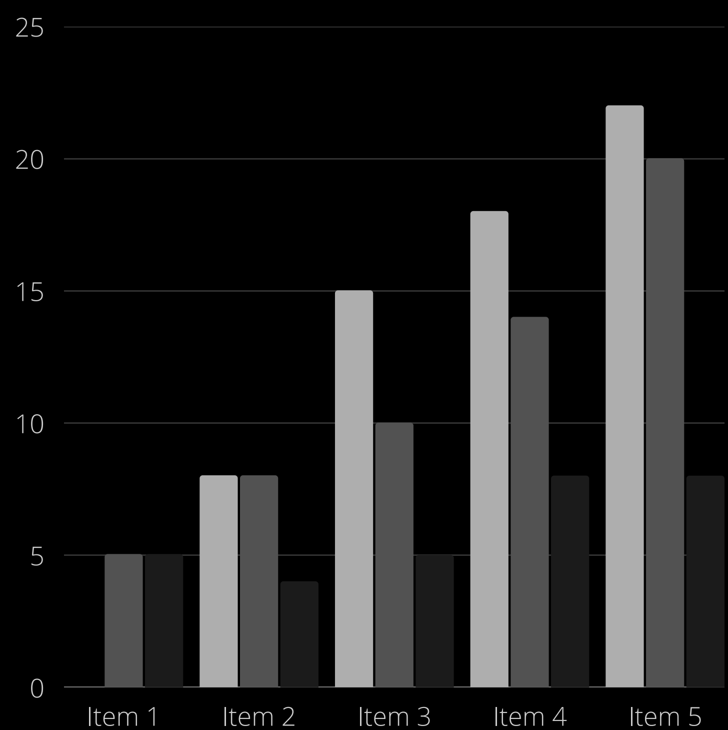
FERGUSSON COLLEGE PUNE
DEPARTMENT OF STATISTICS



The Study of Conventional Energy Source Coal and Solar as an Alternative





Deccan Education Society's
Fergusson College (Autonomous), Pune
Department of Statistics

STS3609: Statistics Practical III

CERTIFICATE

This is to certify that Mr./Ms.______________________________________

Roll No.________, has satisfactorily completed the project work entitled

"____" towards the

partial fulfillment of B.Sc. (Statistics), Semester VI during the academic year

2021-22.

Place: Pune

Date: / /2022

Prof. Deepa Kulkarni **Dr. Subhash. S. Shende**

Contents

Group Members

Name	Roll No.
Aaryan Mallayanmath	4565
Rohit Fuke	4530
Sudhansu Bhadra	4537
Tanmay Shah	4541
Mukta Wagh	4508
Rutvik Aglawe	4568
Chaitanyakumar Devkar	4546

Project Guide :- Prof. Deepa Kulkarni

Acknowledgement

We wish to thank the Department of Statistics, Fergusson College (Autonomous), Pune for giving us an opportunity to do this project .

This project has been completed under the valuable guidance of Prof. Deepa Kulkarni.We would like to express our profound gratitude towards her for her patient guidance, valuable and constructive suggestions for this research project. We appreciate her willingness for motivating and keeping our progress on schedule.

We would also like to extend a special thanks to the Head of the Department of Statistics, Prof. Subhash S. Shende for his encouragement and support throughout the course of the project.

We would also like to thank the other staff members for the support, suggestions and encouragement. Finally, we would like to express our gratitude to our family and friends for offering us tremendous assistance and contributing to this project during the process of data collection.

Objectives and Statistical Tools Used

Objectives

- To draw various graphs and check relationship between different attributes

- To check the relationship between Education Level and Familiarity with Solar Energy and Policies related to Solar Development.

- To check the relationship between Occupation and willingness to Install Solar Panels at Home.

- To check the relationship between Education Level and Awareness of Government schemes related to usage Solar Energy.

- To check the relationship between CO2 Emission (Dependent variable/Response Variable) and Coal Production in India (Independent Variable/Regressor)

- To check the relationship between Electricity Generation (Dependent variable/Response Variable) and Coal Production in India (Independent Variable/Regressor)

- To develop a forecasting model using **ARIMA Modelling** technique to predict the future trend in Consumption of Coal in India.

- To develop a forecasting model using **Holt-Winters** method and **ARIMA** technique to predict the future trend in Production of Coal in India and check the accuracy of both the models.

Tools Used:

1] Exploratory Data Analysis

2] Chi Square Test of Independence of Attributes

3] Linear Regression

4]Time Series Analysis

Introduction and Overview

Coal in India has been mined since 1774, and India is the second largest producer and consumer of coal after China, mining 716 million metric tons (789 million short tons) in 2018. Coal supplies over 40% of energy in India. Around 30% of coal is imported. Due to high demand and poor average quality, India imports coking coal to meet the requirements of its steel plants. Dhanbad, the largest coal producing city, has been called the coal capital of India. State-owned Coal India had a monopoly on coal mining between its nationalization in 1973 and 2018. Most of the coal is burned to generate electricity and most electricity is generated by coal, but coal-fired power plants have been criticized for breaking environmental laws. The health and environmental impact of the coal industry is serious, and phasing out coal would have short-term health and environmental benefits greatly exceeding the costs. Electricity from new solar farms in India is cheaper than that generated by the country's existing coal plants.

The production of coal was 716.08 million metric tons (789.34 million short tons) in 2020–21, a decline of 2.02% over the previous year primarily due to disruptions caused by the COVID-19 pandemic. The production of lignite was 36.61 million metric tons (40.36 million short tons) in 2020–21, a decrease of 13.04% over the previous fiscal. Production of coal grew by an compound annual growth rate (CAGR) of 3.19%, and production of lignite declined by a CAGR of 1.60% over the last 10 years. Coal mining is one of India's most dangerous jobs. India targets to increase its coal production to 1,200 million metric tons (1,300 million short tons) by 2023-24. Washing Coal is an integral part of the coal production process in which raw coal from mines is washed to remove the ash content to make it fit for feeding into boilers such as those in steel plants. Coal washeries are generally not a part of coal mines in India, with some exceptions. There were 60 coal washeries (19 coking and 41 non-coking) in India as on 31 March 2021 with a total installed capacity of 138.58 million tonnes per year, of which 108.60 million tonnes are non-coking and 29.98 million tonnes are coking coal washeries.

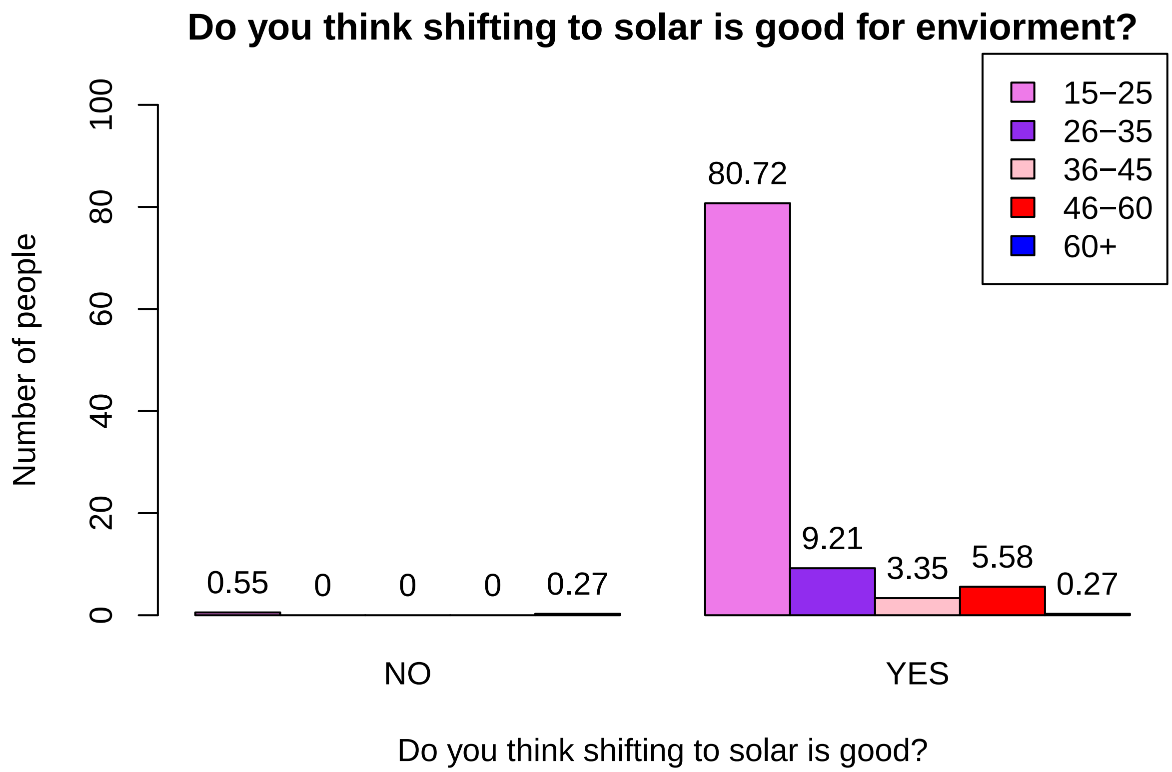
Motivation

Coal is the most important and abundant fossil fuel in India. It accounts for 55% of the country's energy need. The country's industrial heritage was built upon indigenous coal. Commercial primary energy consumption in India has grown by about 700% in the last four decades. Driven by the rising population, expanding economy and a quest for improved quality of life, energy usage in India is expected to rise. Considering the limited reserve potentiality of petroleum & natural gas, eco-conservation restriction on hydel project and geo-political perception of nuclear power, coal will continue to occupy centre-stage of India 's energy scenario.

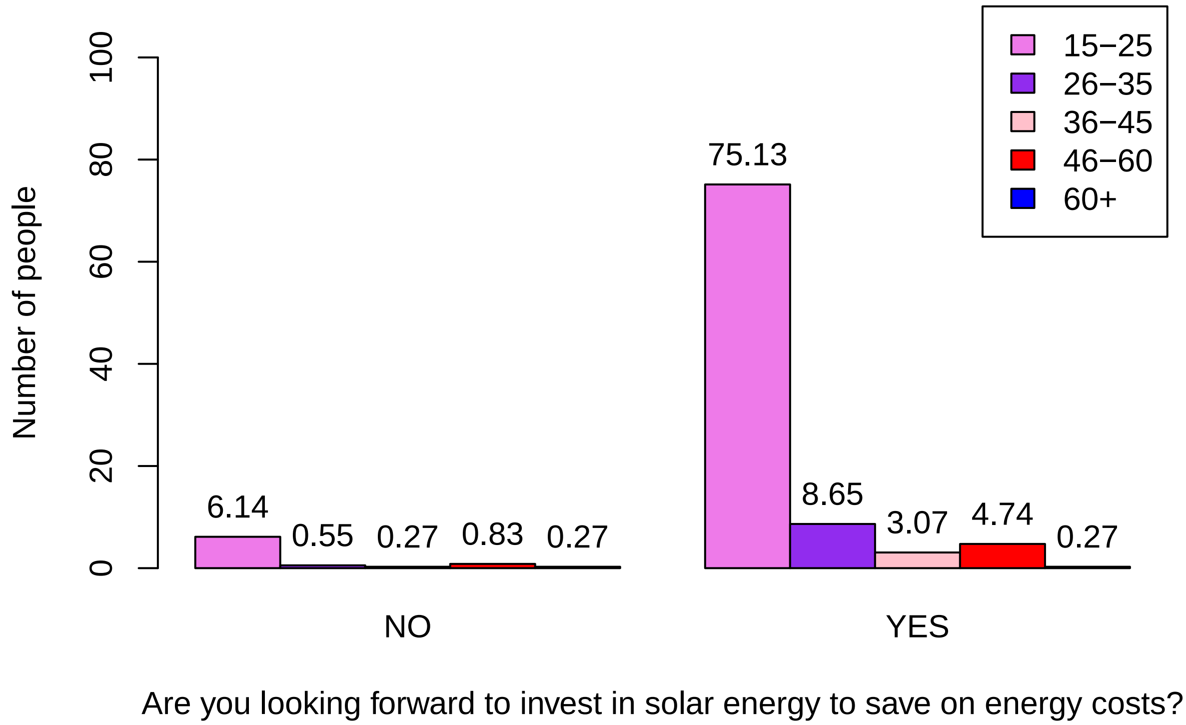
In October 2021, as the economy recovered from the pandemic's second wave, a sharp surge in energy demand triggered a spiraling fuel shortage across coal-fired thermal stations. Due to several factors, India is staring at a coal crisis again, and the Indian Railways has cancelled trains to prioritize delivery of coal rakes across the country.

The motivation for doing this project is to the understand the behavioral pattern of the production of coal and it's demand within the country. We would also like to know the effects production of coal has on the emission of CO2 gas and generation of electricity in the country. Finally with the help of a short survey, we want to check the people's knowledge and opinions towards Solar as an alternative source of energy.

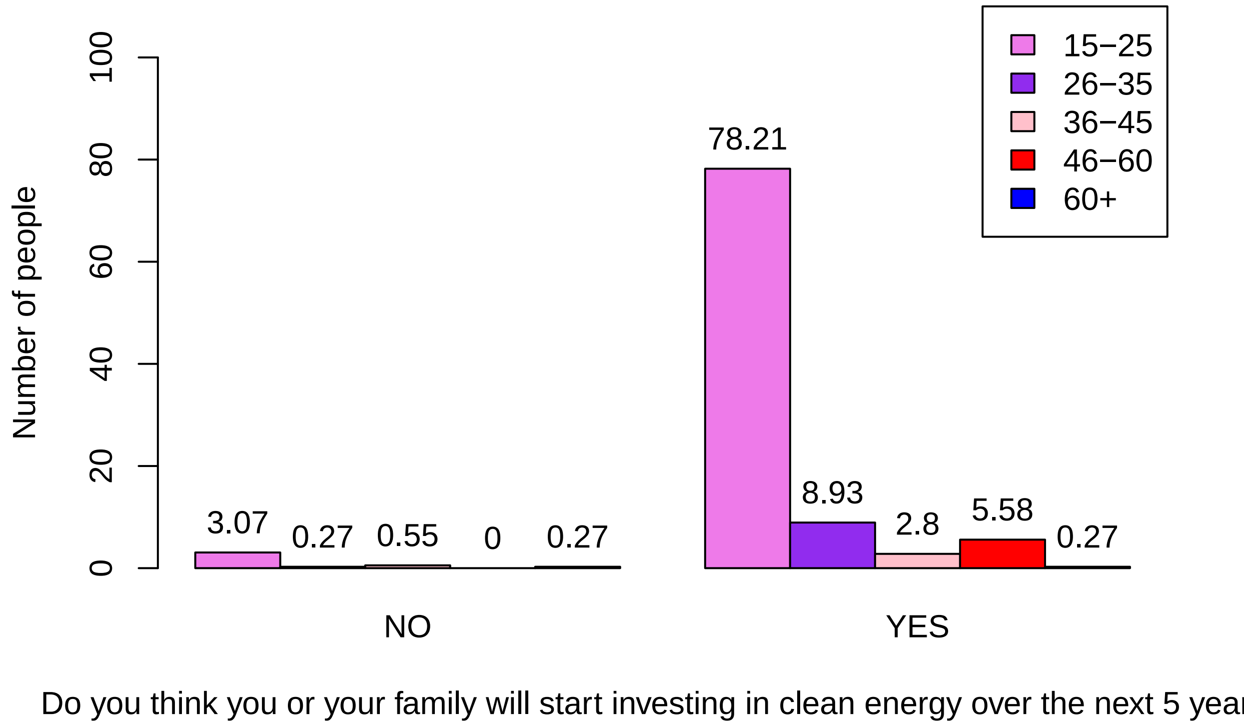
Exploratory Data Analysis



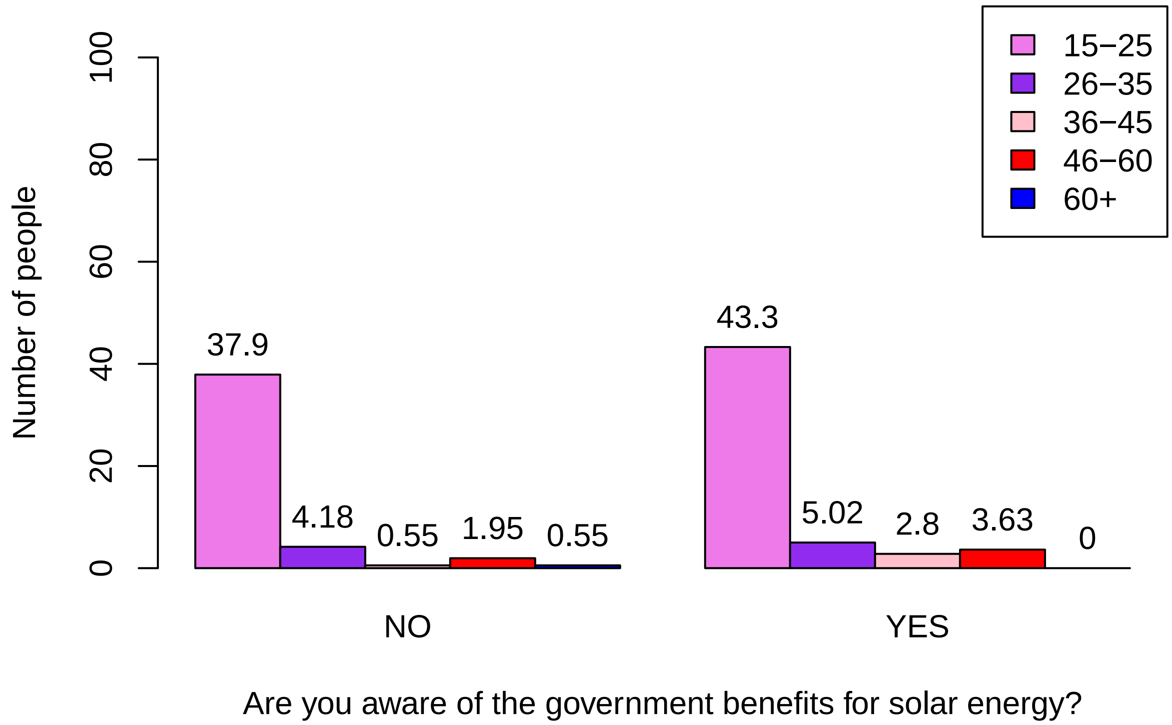
INTERPRETATION: From the graph it is easy to see that almost all people from the different age groups think that it is good to shift towards Solar as primary source of energy.



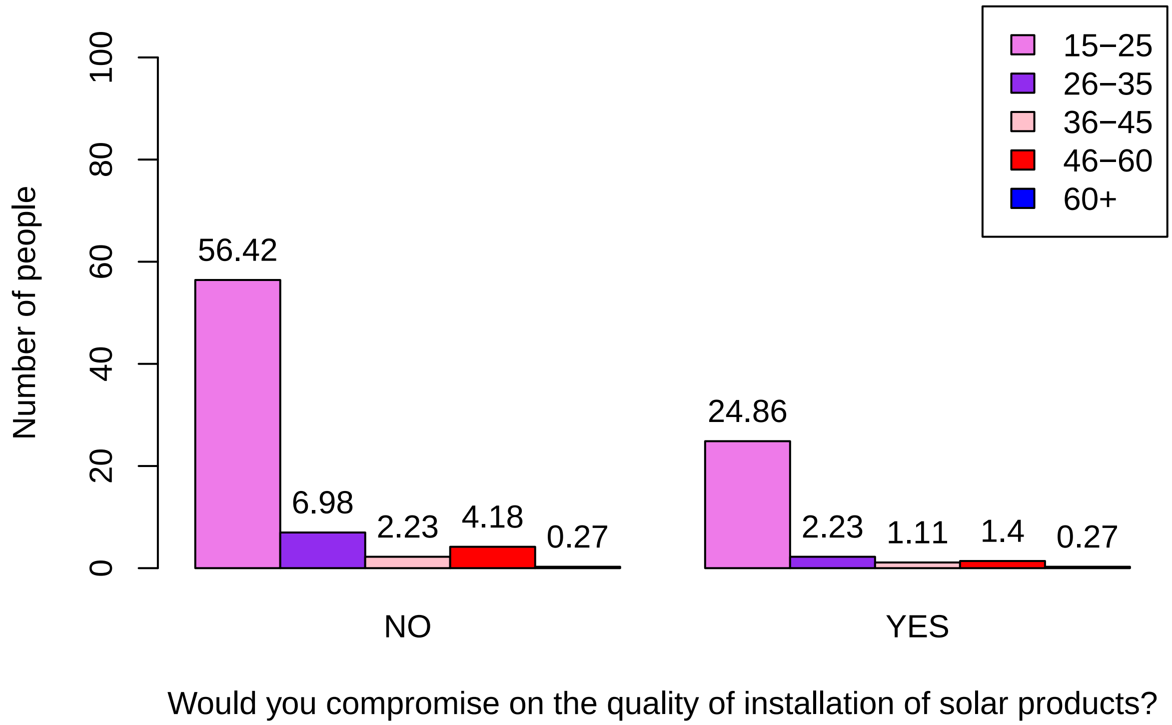
INTERPRETATION: Here we can see that people from all most all age groups are looking forwards to invest in solar energy to save on energy costs. A very small amount of people from the age group 15-25 are against this thought.



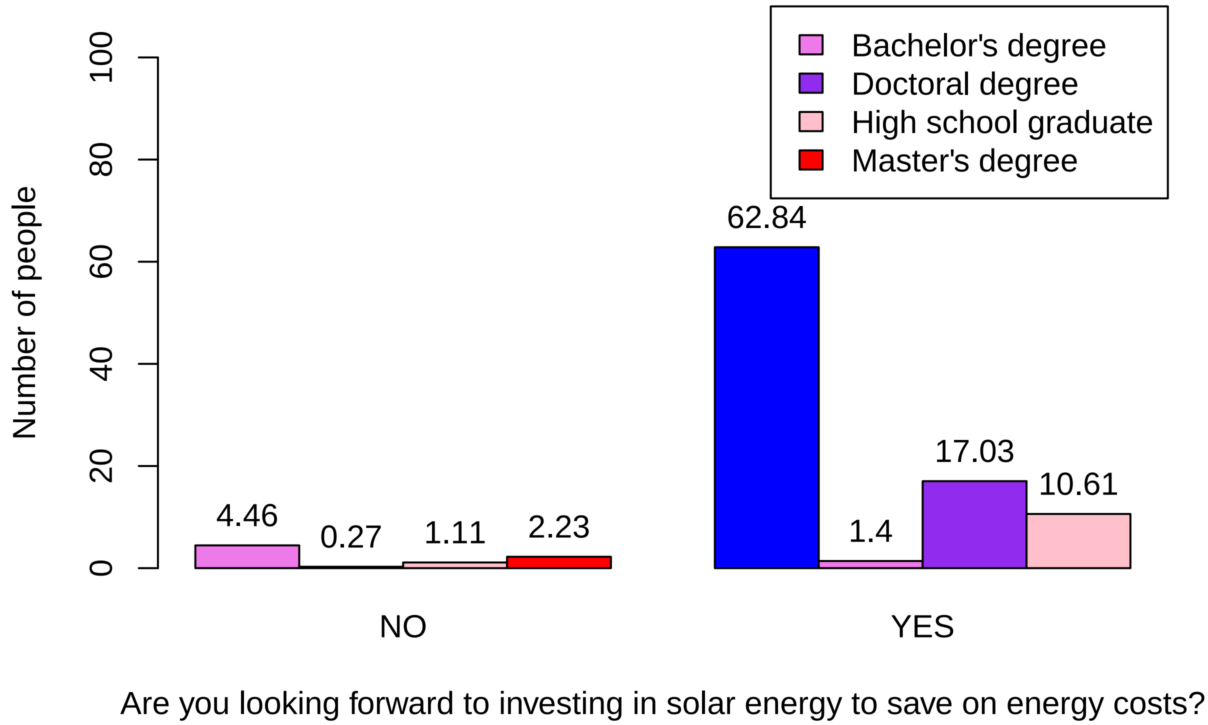
INTERPRETATION: Almost 100% of people from different age groups are positive about investing in clean energy over the next 5 years.



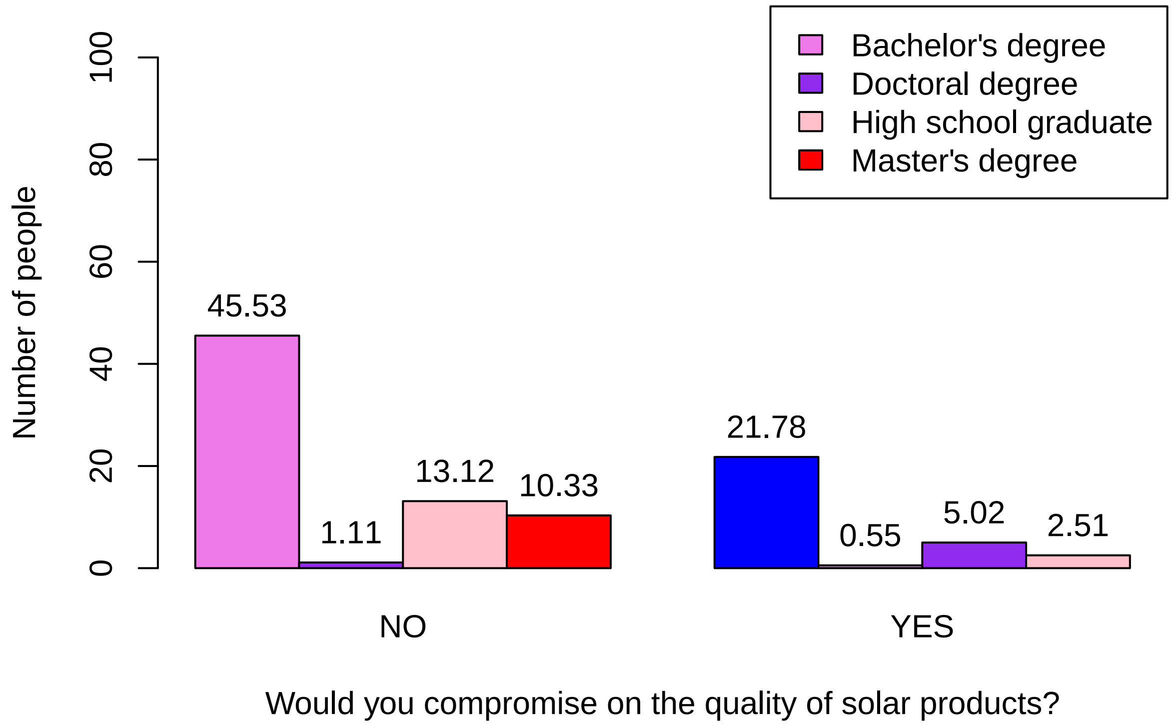
INTERPRETATION: Amount of people knowing and not knowing about the government benefits for solar energy is almost same. We can see that a greater number of people within the age group 15-25 know about government benefits.



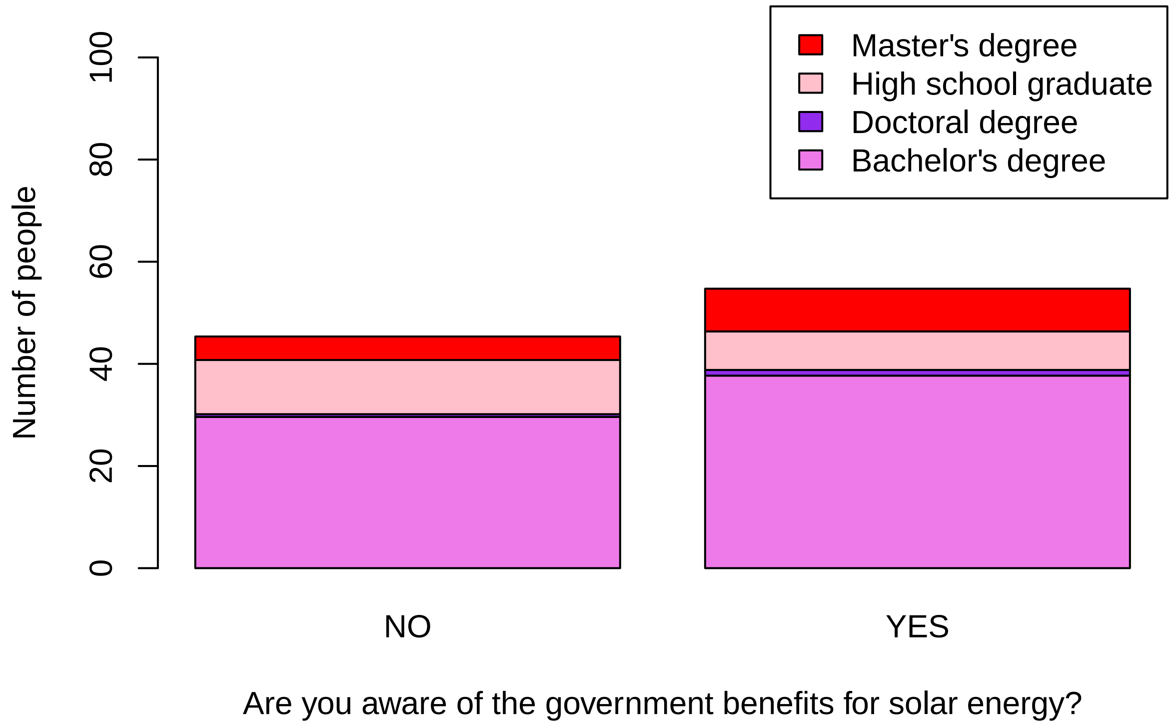
INTERPRETATION: From the bar plot it is easy to make out that a greater number of people won't compromise on the quality of installation of solar products.



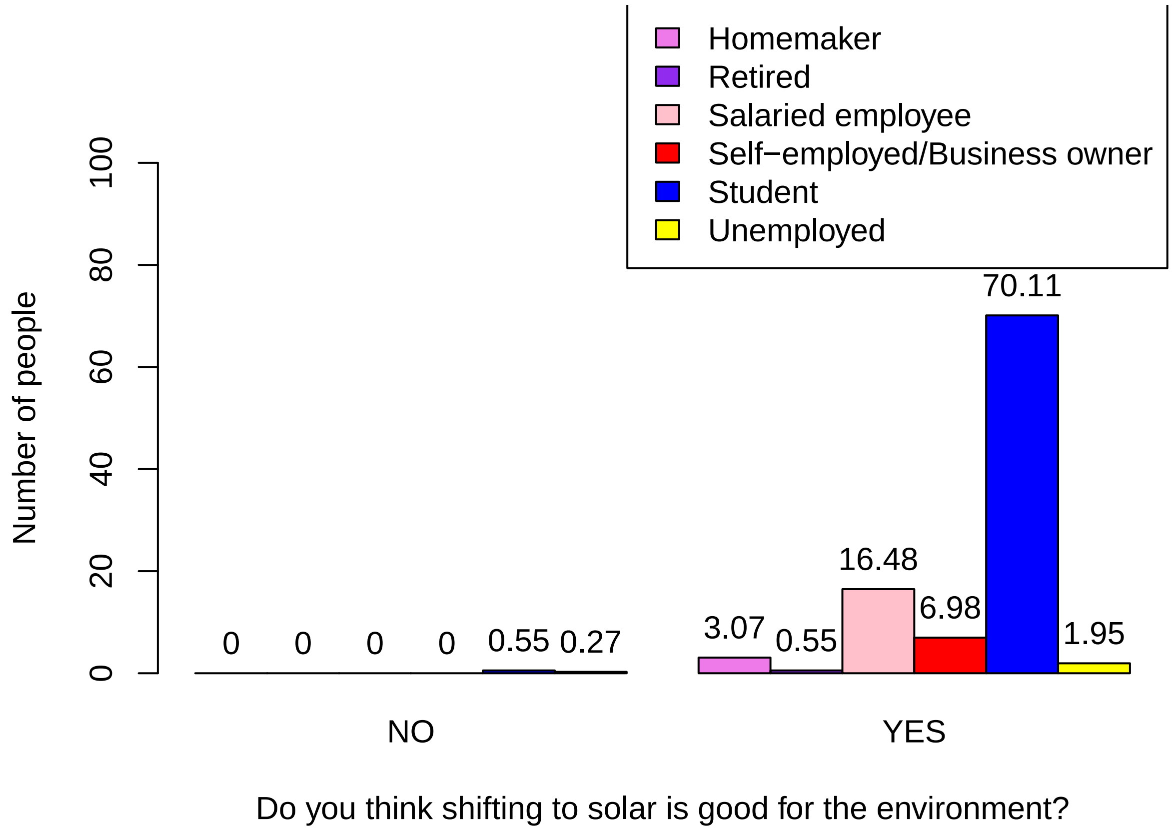
INTERPRETATION: According to the graph, most of people with different education levels are looking for to investing in Solar to save on energy costs.



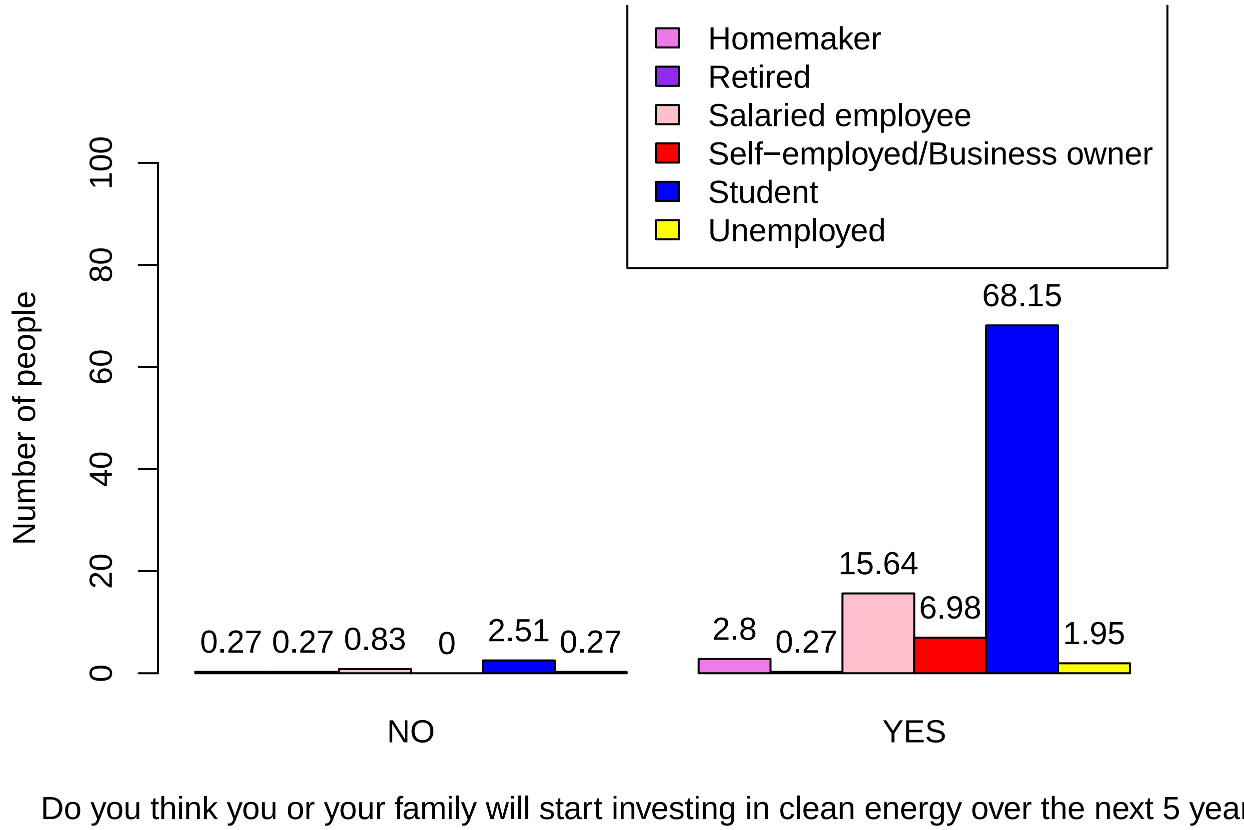
INTERPRETATION: Most people with a Bachelors Degree will not compromise on the quality of solar products. Whereas, more people with Doctoral Degree are willing to compromise.



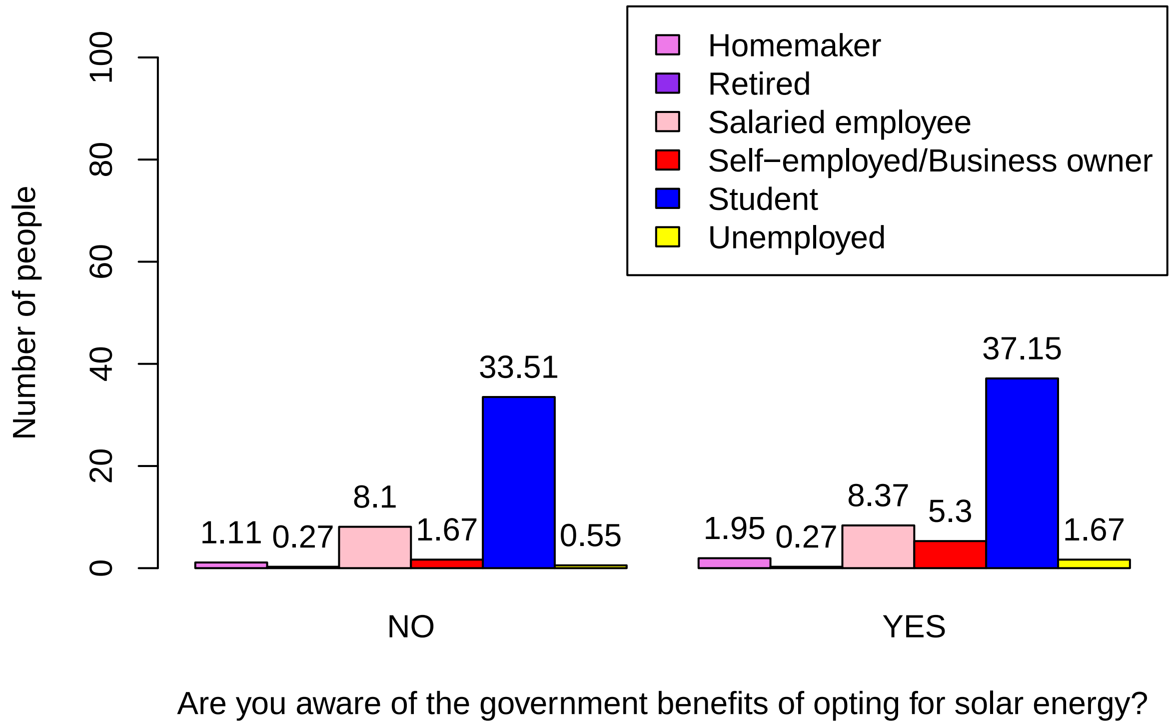
INTERPRETATION: The number of people with different education levels knowing about goverenment benefits is almost equal to the number of people not knowing about the government benefits.



INTERPRETATION: We can say that different occupations do not hamper the peoples opinion about solar energy being good for the environment.



INTERPRETATION: People from all various kinds of occupation are looking forward to investing in Solar energy over the next 5 years.



INTERPRETATION: The number of people with different occupations levels knowing about government benefits is almost equal to the number of people not knowing about the government benefits.

Confirmatory Data Analysis

χ^2 Test for Independence of Attributes

The Chi Squared statistics is commonly used for testing relationships between categorical variables. The null hypothesis for a chi squared test of independence is that two variables are independent. This test is one -tailed. The degrees of freedom are (m-1)*(n-1). The formula is as follows:

$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

where O_i = Observed frequency
E_i = Expected frequency
Degrees of freedom = (m-1)(n-1)
m = Number of columns
n = Number of rows

Decision Rule :
Reject Ho if $p - value \leq \alpha$
OR
Reject H_o at α % l.o.s if $\chi^2_{calc} \geq \chi^2_{(m-1)(n-1),\alpha}$

1]How familiar are you with solar energy (i.e. solar panels or solar PV) and policies related to solar development.
Let,
A = Education level
Classes of A=3
A1=Bachelor's degree
A2=Master's degree
A3=High school graduate
And
B=How familiar are you with solar energy (i.e. solar panels or solar PV) and policies related to solar development
Classes of B=3
B1=Moderately familiar
B2=Not at all familiar
B3=Very familiar

Hypothesis:
To Test
H_o: A and B are Independent
Vs
H_1: A and B are Associated

```
##
##  Pearson's Chi-squared test
##
## data:  A
## X-squared = 2.586, df = 4, p-value = 0.6293
```

Here, as $p - value \geq\ 0.05$
AND
$\chi^2_{calc} \leq \chi^2_{4;0.05} = 9.487729$

Therefore we may ACCEPT Ho

CONCLUSION: We conclude that people's knowledge about solar energy and their education levels are independent.

2]Occupation of people and willingness to install solar panels at home
Let
A=Occupation
Classes of A=4
A1=Students
A2=Salaried employee
A3=Self-employed/Business owner
A4=Unemployed
And
B= installment of solar panel at home
B1=Yes, I have solar panels at my home
B2=No, I don't have enough rooftop space
B3=No,I find it costly
B4=No,it is high maintenance

Hypothesis:
To Test
H_o: A and B are Independent
Vs
H_1: A and B are Associated

```
##
##  Pearson's Chi-squared test
##
## data:  C
## X-squared = 4.3016, df = 9, p-value = 0.8905
```

Here, as $p - value \geq\ 0.05$
AND

$\chi^2_{calc} \leq \chi^2_{9;0.05} = 16.91898$

Therefore we may ACCEPT Ho

CONCLUSION: We conclude that occupation of people and their willingness to install solar panels at home are independent.

3]Education of people and awareness of government policies related to solar energy .
Let
A=Education
Classes of A=3
A1=Bachelor's degree
A2=Master'degree
A3=High school graduate
And
B= Are you aware of government scheme
Classes of B=2
B1=Yes
B2=No

Hypothesis:
To Test
H_o: A and B are Independent
Vs
H_1: A and B are Associated

```
##
##  Pearson's Chi-squared test
##
## data:  B
## X-squared = 1.8441, df = 2, p-value = 0.3977
```

Here, as $p-value \geq$ 0.05
AND
$\chi^2_{calc} \leq \chi^2_{2;0.05} = 5.99146$

Therefore we may ACCEPT Ho

CONCLUSION: We conclude that awareness of government policies related to solar energy and education of people are independent.

Time Series Analysis

A time series is a series of data points indexed or listed in time order. It can be continuous trace or discrete set of observations. Here we are dealing with observations taken at discrete time periods. By appropriate choice of origin and scale we can take the time periods to be 1, 2,. . .

Modelling: We observe the pattern in data and fit a time series model to data. This model depends on unknown parameters which need to be estimated.

Forecasting: On the basis of available observations we predict the further observations with the help of different time series forecasting techniques which suit our data.

Analysis of data:
Here for forecasting we have used two techniques:
1] Triple Exponential Smoothing(Holt-winters)
2] ARIMA Modelling

HOLTWINTERS TRIPLE EXPONENTIAL SMOOTHING: Exponential smoothing is a very popular technique to produce a smoothed time series. Exponential smoothing assigns exponentially decreasing weights as the observation get older. Recent observations are given relatively more weights in forecasting than the older observations. It is usually used to make short term forecasts.

ARIMA MODELLING:

ARIMA stands for Autoregressive Integrated Moving Average. This model is fitted to the time series data either to better understand the data or to predict future points in the series. The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged values. The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. Non seasonal ARIMA Models are generally denoted by ARIMA(p,d,q) where parameters p,d and q are non negative integers.

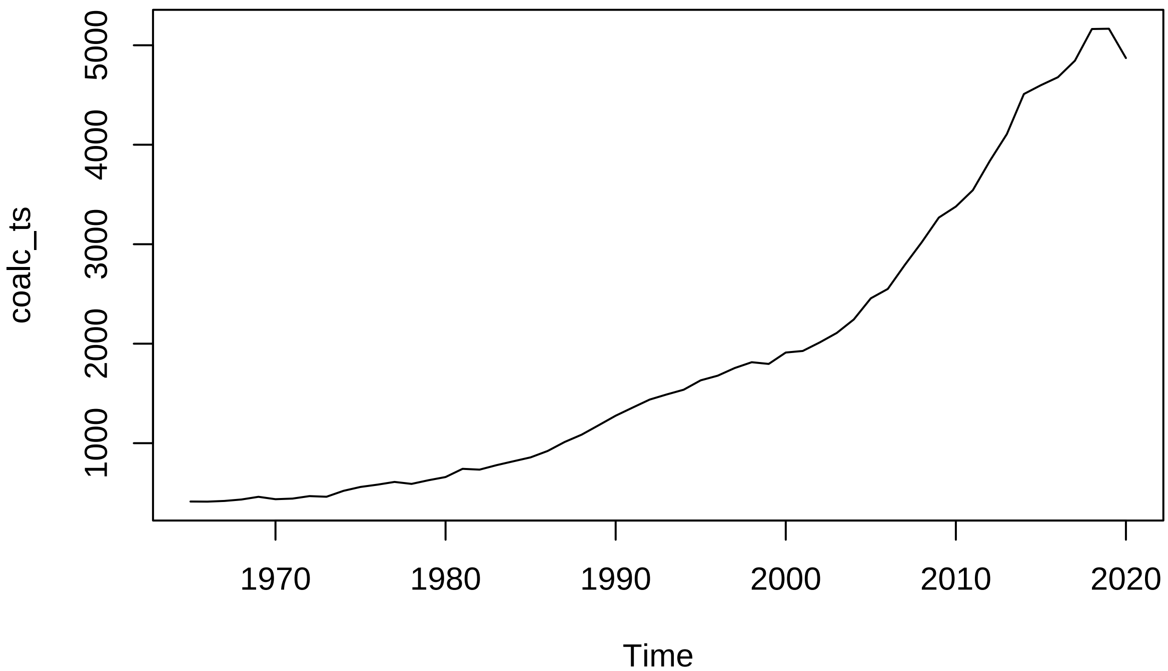
P = order of autoregressive model

q = order of the moving average model

d = degree of differencing

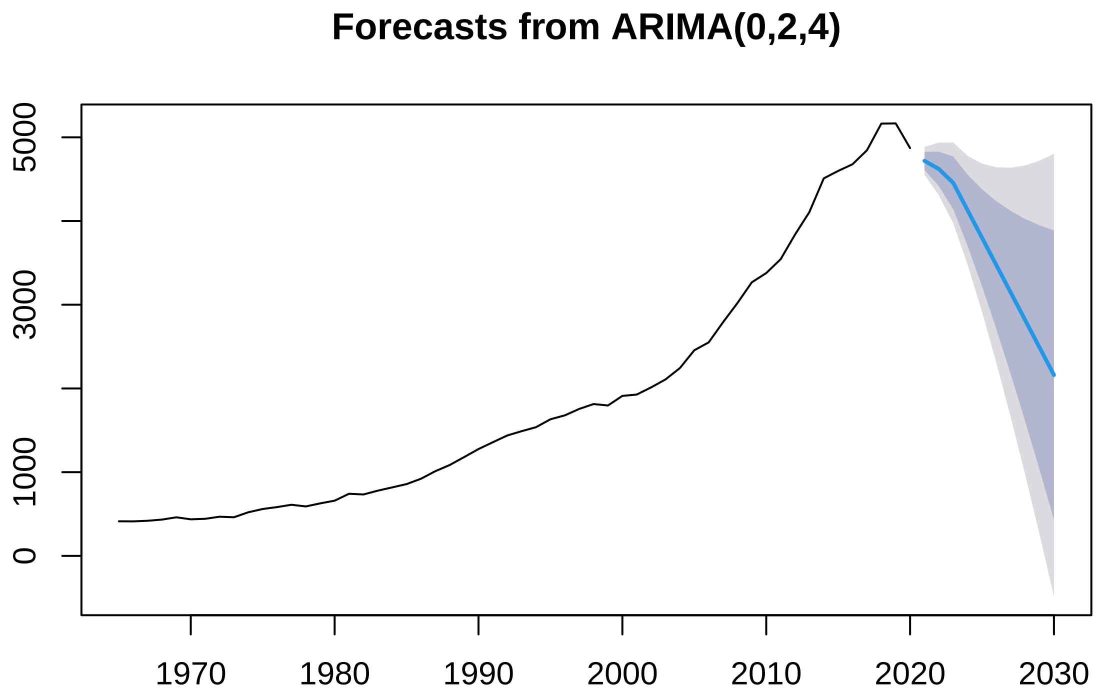
Differencing: In statistics, differencing is the transformation applied to the time series data in order to make it stationary. In order to difference the data, the difference between consecutive observations is computed. Differencing removes the changes in the level of time series, eliminating trend and seasonality and consequently stabilizing the mean of the time series

1] Analysis of Coal Consumption in India.




```
##
##  ARIMA(2,2,2)                    : Inf
##  ARIMA(0,2,0)                    : 647.7614
##  ARIMA(1,2,0)                    : 648.6055
##  ARIMA(0,2,1)                    : 646.4098
##  ARIMA(1,2,1)                    : Inf
##  ARIMA(0,2,2)                    : 645.6606
##  ARIMA(1,2,2)                    : 647.653
##  ARIMA(0,2,3)                    : 645.3616
##  ARIMA(1,2,3)                    : 646.5562
##  ARIMA(0,2,4)                    : 639.8219
##  ARIMA(1,2,4)                    : 641.8212
##  ARIMA(0,2,5)                    : 641.621
##  ARIMA(1,2,5)                    : Inf
```

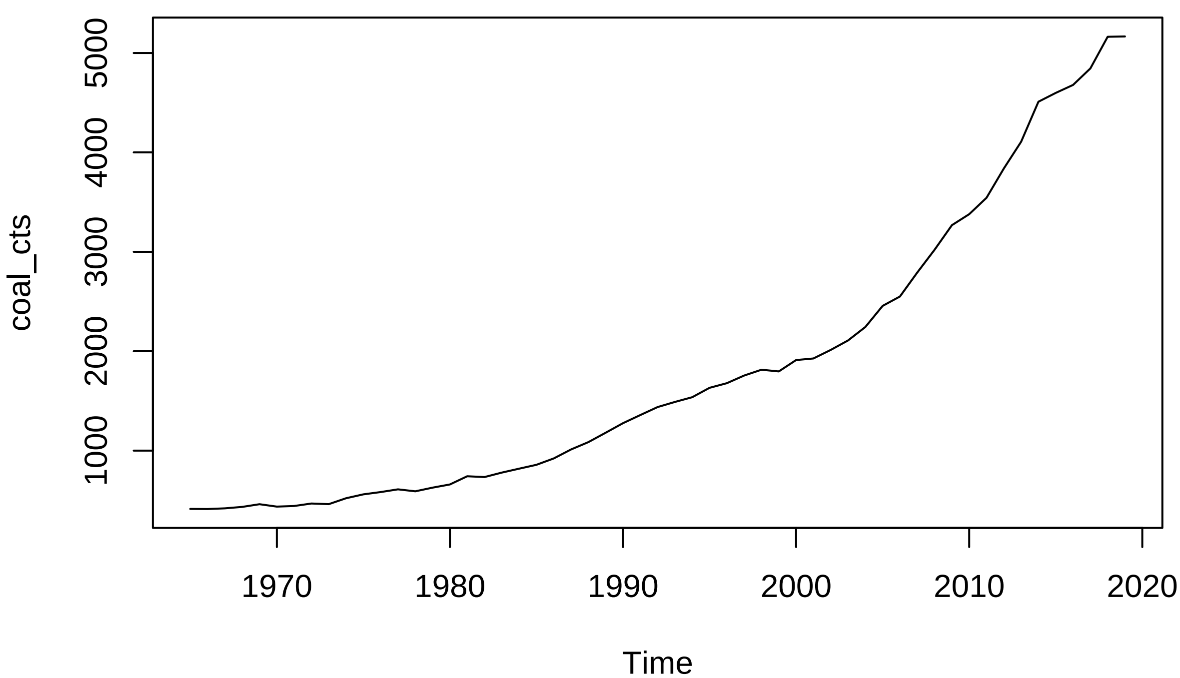
```
##
##  Best model: ARIMA(0,2,4)
```



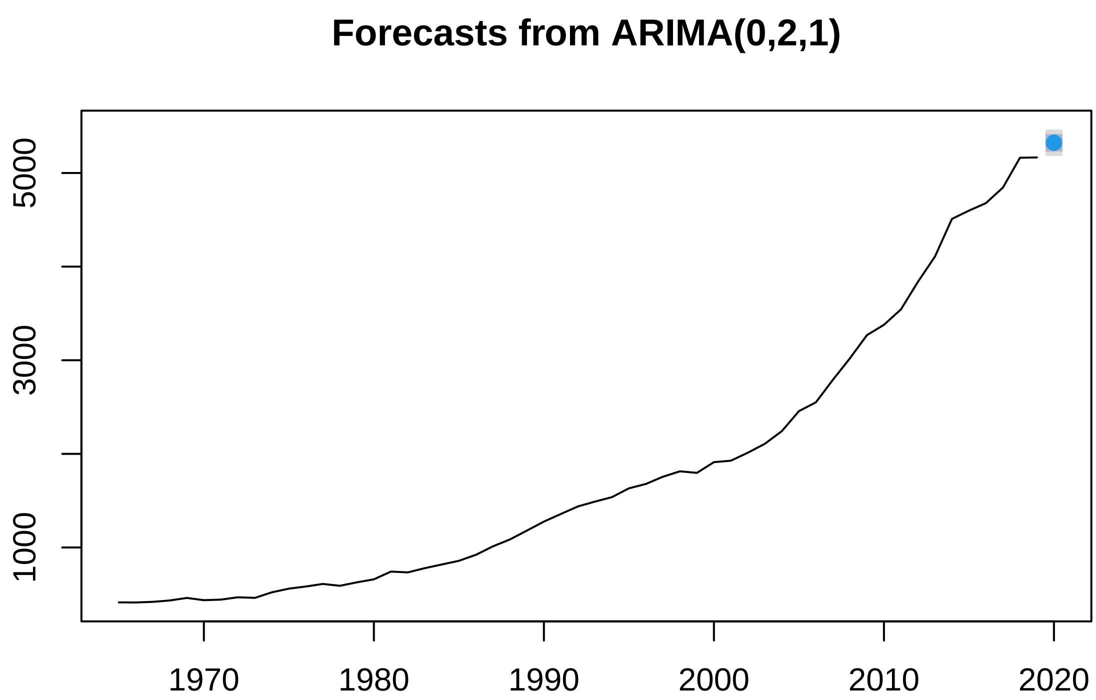

```
##
## Forecast method: ARIMA(0,2,4)
##
## Model Information:
## Series: coalc_ts
## ARIMA(0,2,4)
##
## Coefficients:
##           ma1      ma2      ma3     ma4
##       -0.3791  -0.0376  -0.1415  0.5421
## s.e.   0.1685   0.2283   0.2305  0.1218
##
## sigma^2 = 7133:  log likelihood = -314.91
## AIC=639.82   AICc=641.07   BIC=649.77
##
## Error measures:
##                     ME     RMSE      MAE       MPE    MAPE      MASE
## Training set -5.892991 79.80697 53.15302 0.1100585 3.21307 0.5622333
```

```
##                      ACF1
## Training set -0.00616414
##
## Forecasts:
##      Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## 2021       4718.447 4610.2077 4826.687 4552.9091 4883.985
## 2022       4621.575 4415.4230 4827.728 4306.2925 4936.858
## 2023       4455.127 4139.8056 4770.448 3972.8845 4937.369
## 2024       4127.361 3701.3697 4553.353 3475.8635 4778.859
## 2025       3799.595 3220.0521 4379.139 2913.2603 4685.930
## 2026       3471.830 2706.7659 4236.893 2301.7657 4641.894
## 2027       3144.064 2168.0956 4120.032 1651.4492 4636.679
## 2028       2816.298 1607.9377 4024.658  968.2704 4664.326
## 2029       2488.532 1028.7810 3948.284  256.0352 4721.029
## 2030       2160.767  432.3507 3889.183 -482.6175 4804.151
```

We can see that the forecasts show a downward trend for the next 10 years. In the original time series plot, we can see that there is a sudden dip in the value of coal consumption. We assume that this is due to the COVID 19 Pandemic. To see whether this irregularity affects the forecast, we try and extrapolate the value for the year 2020. We do this by plotting a time series till year 2019 and then forecast the year 2020's value and then replacing this value with the original value.



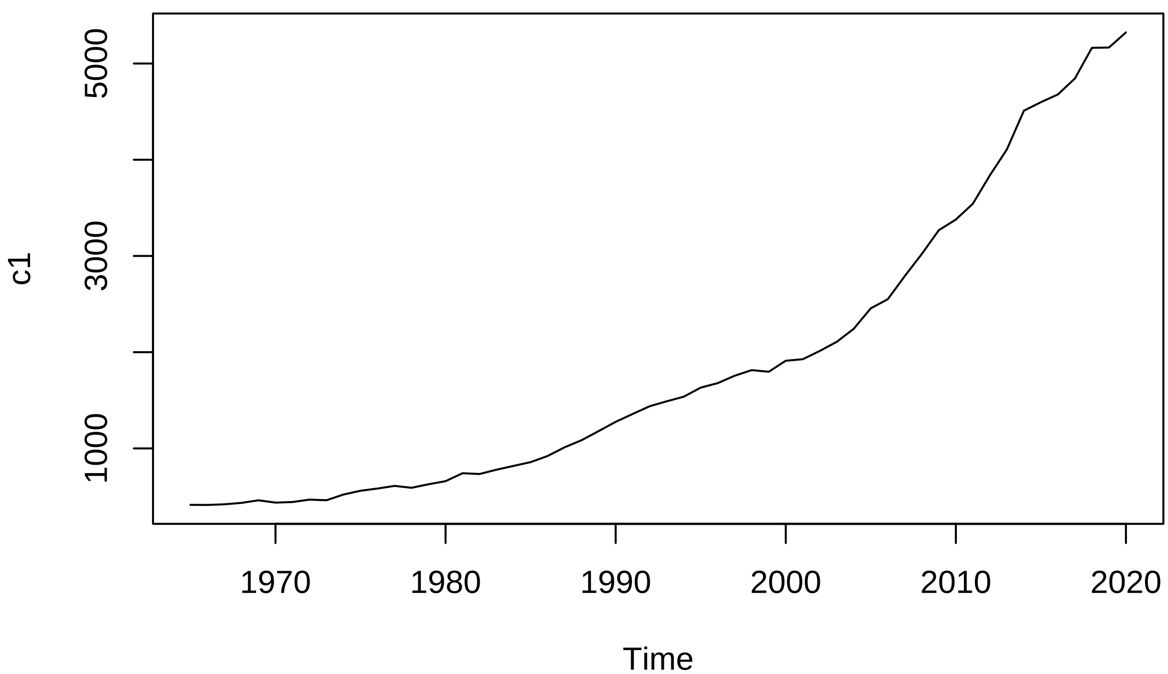
```
##
##  ARIMA(2,2,2)                    : Inf
##  ARIMA(0,2,0)                    : 626.3239
##  ARIMA(1,2,0)                    : 616.3943
##  ARIMA(0,2,1)                    : 607.1129
##  ARIMA(1,2,1)                    : 608.9175
##  ARIMA(0,2,2)                    : 608.8423
##  ARIMA(1,2,2)                    : 610.8436
##
##  Best model: ARIMA(0,2,1)
```




```
##
## Forecast method: ARIMA(0,2,1)
##
## Model Information:
## Series: coal_cts
## ARIMA(0,2,1)
##
## Coefficients:
##           ma1
##       -0.7226
```

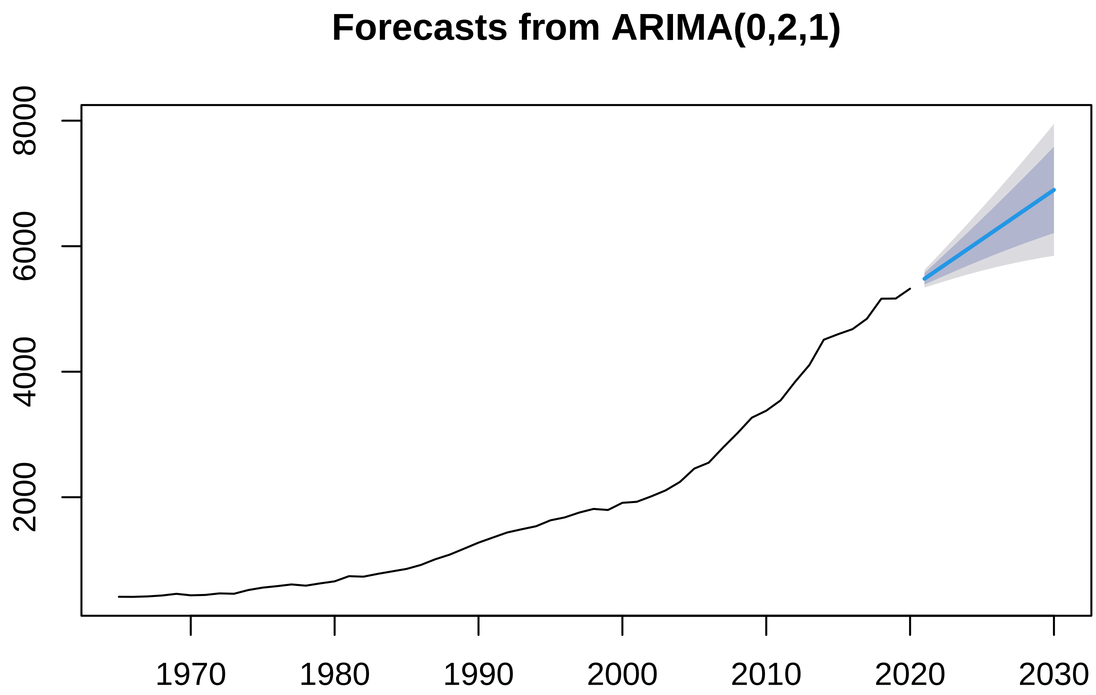
```
## s.e.   0.0923
##
## sigma^2 = 5150:  log likelihood = -301.56
## AIC=607.11   AICc=607.35   BIC=611.05
##
## Error measures:
##                    ME     RMSE     MAE       MPE     MAPE      MASE        ACF1
## Training set 10.09605 69.77806 47.7975 0.8055216 2.784597 0.5262177 -0.08650412
##
## Forecasts:
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2020        5323.38 5231.412 5415.347 5182.727 5464.032
```

Here we can see that the forecasted Value of year 2020 is 5323.38. We replace the original value with the forecasted value as 'Extrapolated Observation'




```
##
##  ARIMA(2,2,2)                    : Inf
##  ARIMA(0,2,0)                    : 640.1888
##  ARIMA(1,2,0)                    : 626.939
##  ARIMA(0,2,1)                    : 617.469
```

```
##  ARIMA(1,2,1)                    : 619.2972
##  ARIMA(0,2,2)                    : 619.2373
##  ARIMA(1,2,2)                    : Inf
##
##  Best model: ARIMA(0,2,1)
```




```
##
## Forecast method: ARIMA(0,2,1)
##
## Model Information:
## Series: c1
## ARIMA(0,2,1)
##
## Coefficients:
##           ma1
##       -0.7228
## s.e.   0.0894
##
## sigma^2 = 5053:  log likelihood = -306.73
## AIC=617.47   AICc=617.7   BIC=621.45
##
```

```
## Error measures:
##                    ME    RMSE      MAE       MPE     MAPE      MASE        ACF1
## Training set 9.925301 69.1518 46.94579 0.7916639 2.734933 0.5100485 -0.07765312
##
## Forecasts:
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2021       5480.768 5389.673 5571.863 5341.450 5620.086
## 2022       5638.156 5490.389 5785.923 5412.166 5864.146
## 2023       5795.544 5590.885 6000.203 5482.545 6108.543
## 2024       5952.932 5688.878 6216.986 5549.096 6356.768
## 2025       6110.320 5783.781 6436.859 5610.921 6609.719
## 2026       6267.708 5875.444 6659.972 5667.792 6867.624
## 2027       6425.096 5963.871 6886.321 5719.713 7130.479
## 2028       6582.484 6049.120 7115.848 5766.775 7398.193
## 2029       6739.872 6131.275 7348.469 5809.103 7670.641
## 2030       6897.260 6210.423 7584.097 5846.834 7947.686
```

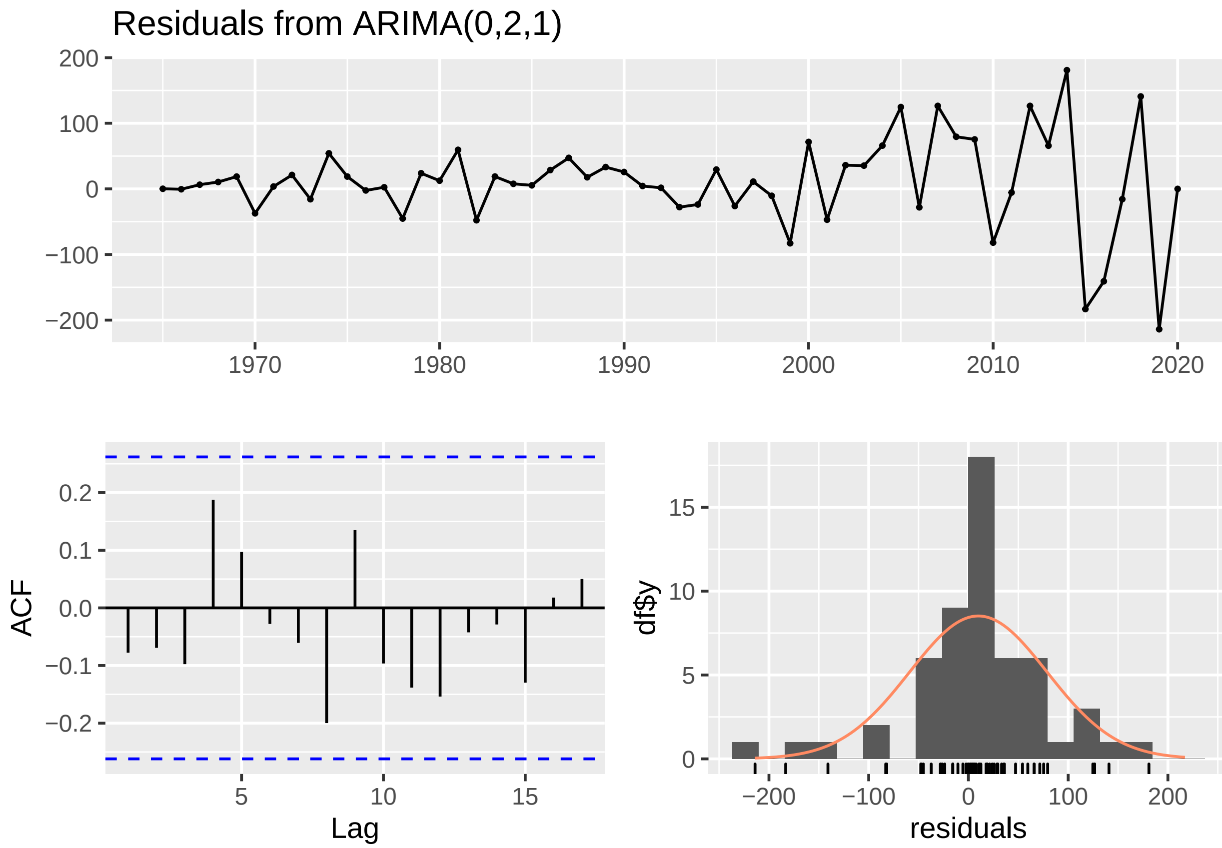
Here we can see that after extrapolating, the time series forecast shows an upward trend. Hence, we can say that the Irregularity is removed and to check this we will perform Residual Analysis.

Ljung Box or Box Pierce Test to check the Autocorrelation between Residuals

To Test:
H_o:Residuals follow i.i.d rvs or White Noise
vs
H_1:Residual have serial dependance

```
##
##  Box-Ljung test
##
## data:  resid(cmodel)
## X-squared = 0.3561, df = 1, p-value = 0.5507
```

Again, we see that the p-value is greater than $\alpha = 0.05$, thus we can accept the null hypothesis,indicating the time series does not contain any Autocorrelation.



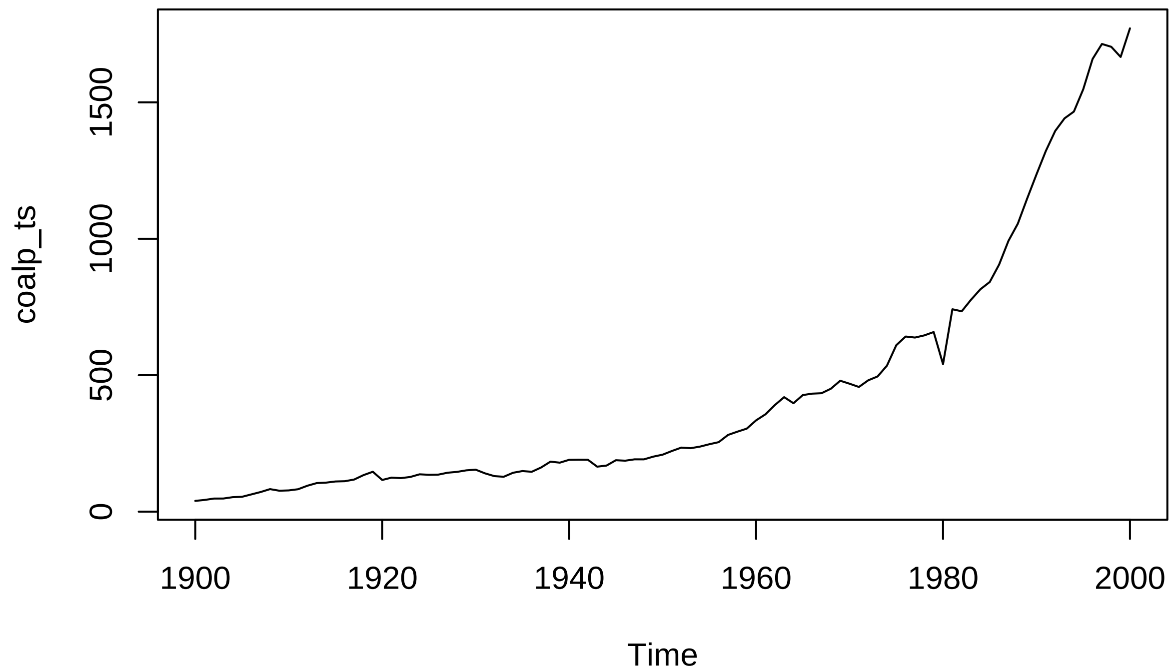

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,2,1)
## Q* = 8.9404, df = 9, p-value = 0.4428
##
## Model df: 1.   Total lags used: 10
```

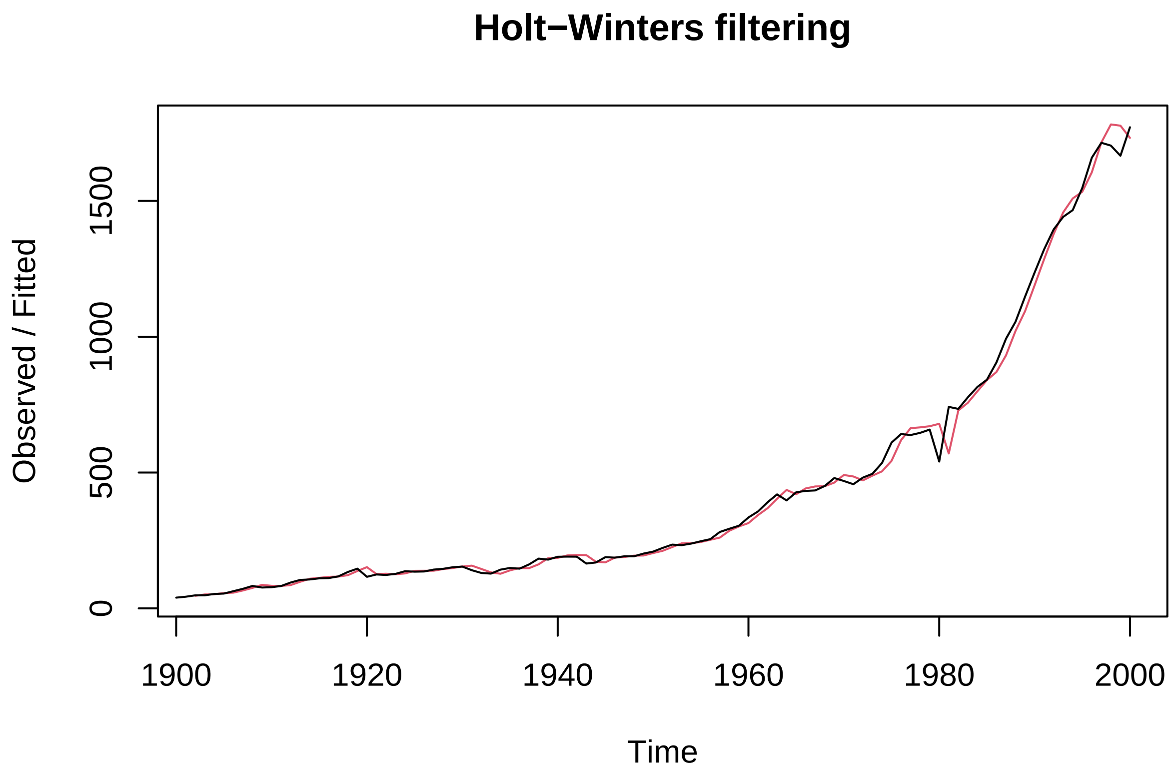
From the ACF vs Lag graph, we can see that there is autocorrelation between the residuals. The Histogram also shows us that the residuals follow Normal Distribution.

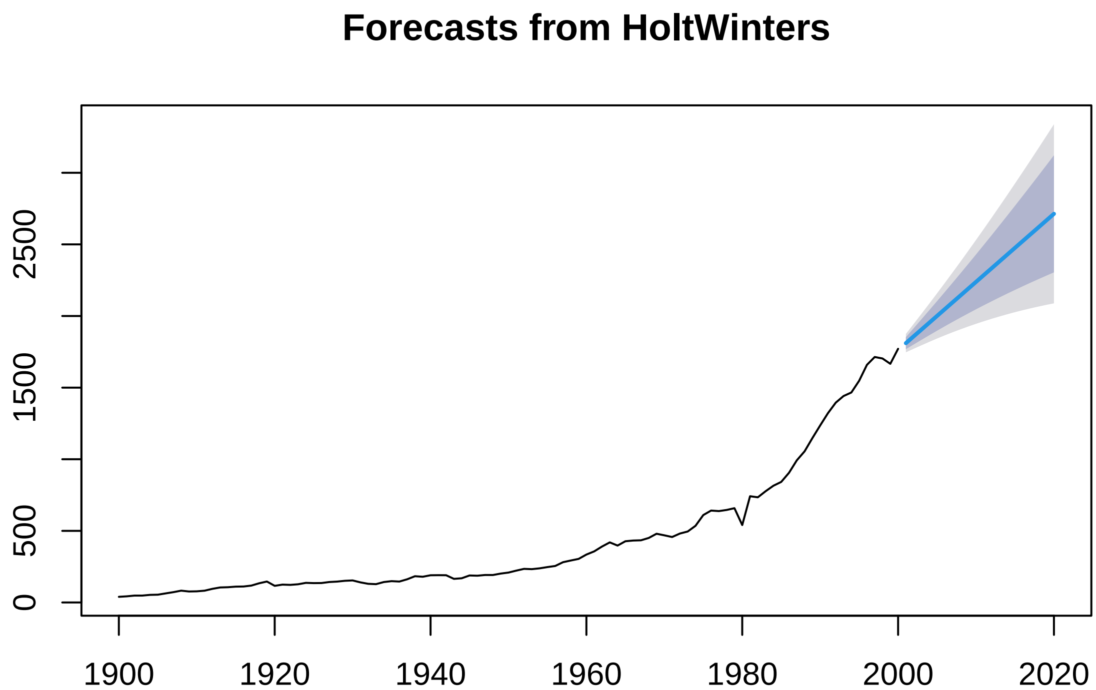
2] Analysis of Coal Production in India.

In this section, we check which model is fits better to the data, Holt-Winters Exponential Smoothing or ARIMA Model.

Holt-Winters Exponential Smoothing:


coalp_ts
0
500
1000
1500
1900
1920
1940
1960
1980
2000
Time


Holt−Winters filtering
Observed / Fitted
0
500
1000
1500
1900
1920
1940
1960
1980
2000
Time




```
##
## Forecast method: HoltWinters
##
## Model Information:
## Holt-Winters exponential smoothing with trend and without seasonal component.
##
## Call:
## HoltWinters(x = coalp_ts, gamma = FALSE)
##
## Smoothing parameters:
##  alpha: 0.7886826
##  beta : 0.1677496
##  gamma: FALSE
##
## Coefficients:
##          [,1]
## a 1762.89653
## b   47.53059
##
## Error measures:
##                       ME     RMSE       MAE        MPE      MAPE       MASE
```

```
## Training set 3.373813 32.61609 18.23873 0.4661984 4.842943 0.7668284
##                      ACF1
## Training set -0.006511473
##
## Forecasts:
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2001       1810.427 1768.641 1852.214 1746.520 1874.334
## 2002       1857.958 1801.149 1914.766 1771.077 1944.839
## 2003       1905.488 1833.625 1977.352 1795.583 2015.394
## 2004       1953.019 1865.734 2040.304 1819.528 2086.510
## 2005       2000.549 1897.344 2103.755 1842.710 2158.389
## 2006       2048.080 1928.399 2167.761 1865.044 2231.117
## 2007       2095.611 1958.876 2232.345 1886.493 2304.728
## 2008       2143.141 1988.769 2297.514 1907.049 2379.233
## 2009       2190.672 2018.080 2363.264 1926.715 2454.629
## 2010       2238.202 2046.814 2429.591 1945.500 2530.905
## 2011       2285.733 2074.983 2496.484 1963.418 2608.048
## 2012       2333.264 2102.594 2563.933 1980.485 2686.042
## 2013       2380.794 2129.660 2631.928 1996.718 2764.871
## 2014       2428.325 2156.191 2700.458 2012.132 2844.517
## 2015       2475.855 2182.198 2769.513 2026.745 2924.966
## 2016       2523.386 2207.691 2839.081 2040.572 3006.200
## 2017       2570.917 2232.680 2909.153 2053.628 3088.205
## 2018       2618.447 2257.175 2979.720 2065.929 3170.966
## 2019       2665.978 2281.185 3050.770 2077.488 3254.467
## 2020       2713.508 2304.719 3122.297 2088.320 3338.697
```

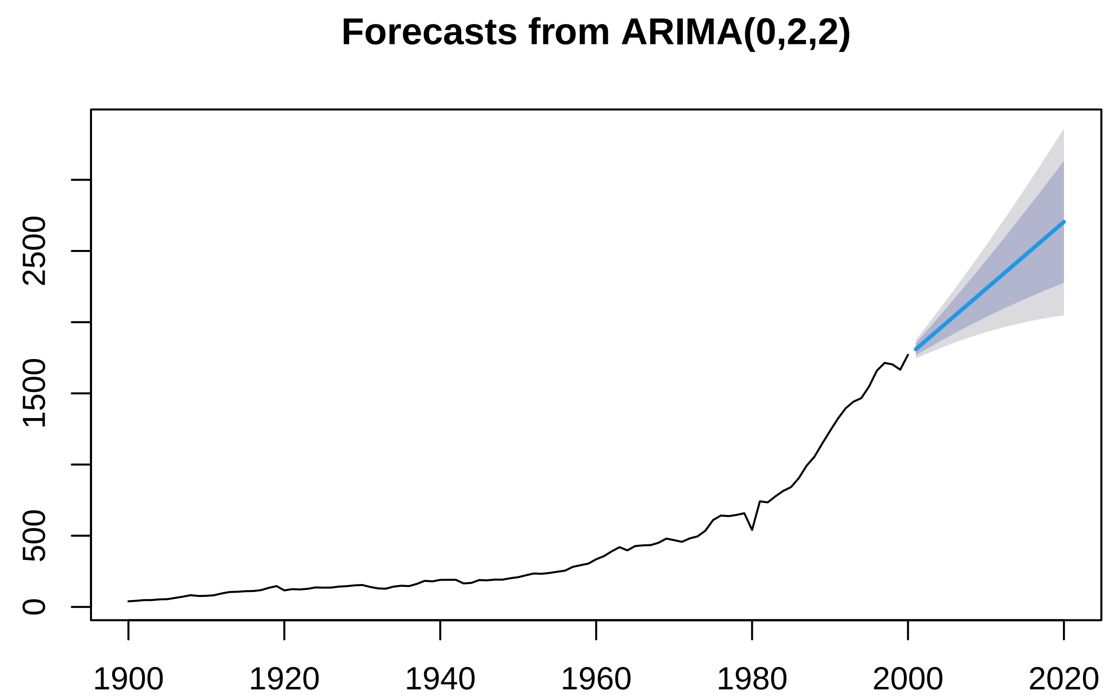
ROOT MEAN SQUARE ERROR(RMSE):

```
## [1] 597.076
```

ARIMA Model:

```
##
##  ARIMA(2,2,2)                    : Inf
##  ARIMA(0,2,0)                    : 1049.11
##  ARIMA(1,2,0)                    : 1009.684
##  ARIMA(0,2,1)                    : 980.5135
##  ARIMA(1,2,1)                    : 978.7752
##  ARIMA(2,2,1)                    : 980.666
##  ARIMA(1,2,2)                    : 980.5053
##  ARIMA(0,2,2)                    : 978.5784
##  ARIMA(0,2,3)                    : 980.546
##  ARIMA(1,2,3)                    : Inf
```

```
##
##  Best model: ARIMA(0,2,2)
```



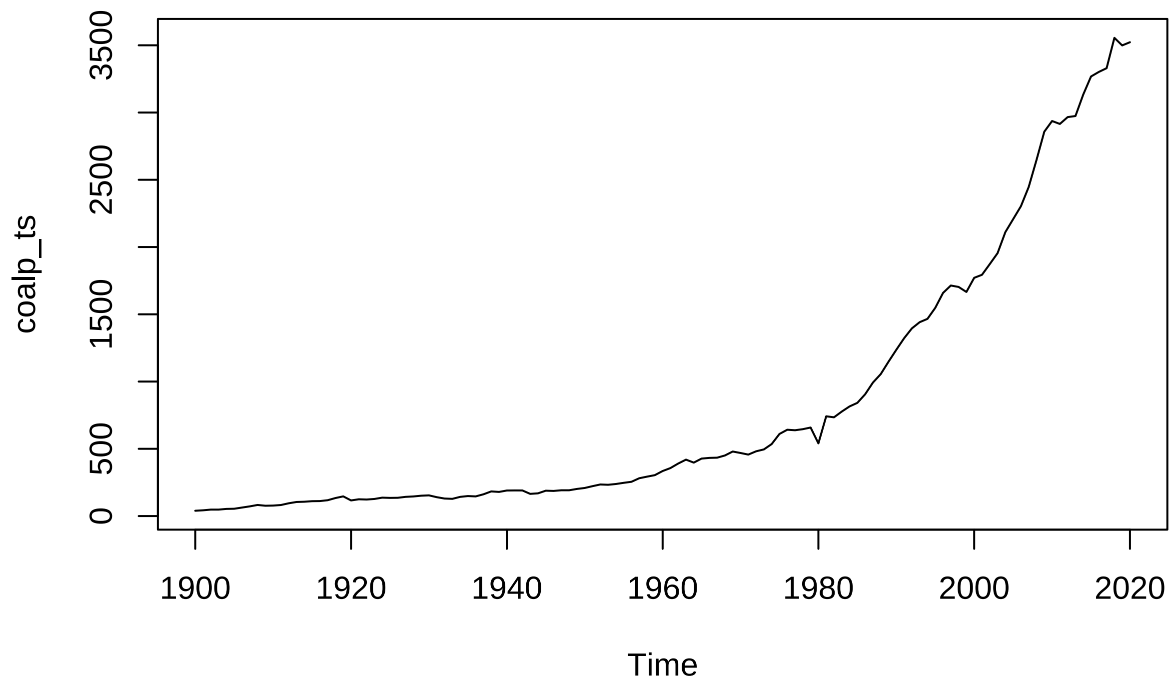

```
##
## Forecast method: ARIMA(0,2,2)
##
## Model Information:
## Series: coalp_ts
## ARIMA(0,2,2)
##
## Coefficients:
##           ma1     ma2
##       -1.0683  0.2081
## s.e.   0.1013  0.1054
##
## sigma^2 = 1086:  log likelihood = -486.29
## AIC=978.58   AICc=978.83   BIC=986.36
##
## Error measures:
##                    ME     RMSE      MAE       MPE     MAPE      MASE
## Training set 3.074871 32.29568 17.81282 0.3966507 4.733281 0.7489215
```

```
##                     ACF1
## Training set -0.01572053
##
## Forecasts:
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2001       1809.852 1767.619 1852.085 1745.262 1874.442
## 2002       1856.952 1799.229 1914.675 1768.672 1945.231
## 2003       1904.052 1830.705 1977.398 1791.878 2016.225
## 2004       1951.151 1861.728 2040.574 1814.391 2087.912
## 2005       1998.251 1892.175 2104.327 1836.022 2160.480
## 2006       2045.351 1921.997 2168.705 1856.697 2234.005
## 2007       2092.451 1951.176 2233.726 1876.390 2308.512
## 2008       2139.551 1979.711 2299.390 1895.097 2384.005
## 2009       2186.650 2007.607 2365.694 1912.827 2460.474
## 2010       2233.750 2034.874 2432.626 1929.596 2537.905
## 2011       2280.850 2061.524 2500.176 1945.421 2616.280
## 2012       2327.950 2087.570 2568.330 1960.321 2695.579
## 2013       2375.050 2113.024 2637.076 1974.316 2775.784
## 2014       2422.150 2137.898 2706.401 1987.425 2856.875
## 2015       2469.249 2162.206 2776.293 1999.667 2938.832
## 2016       2516.349 2185.958 2846.740 2011.060 3021.639
## 2017       2563.449 2209.167 2917.731 2021.621 3105.277
## 2018       2610.549 2231.843 2989.255 2031.368 3189.730
## 2019       2657.649 2253.996 3061.301 2040.316 3274.982
## 2020       2704.749 2275.637 3133.860 2048.480 3361.018
```

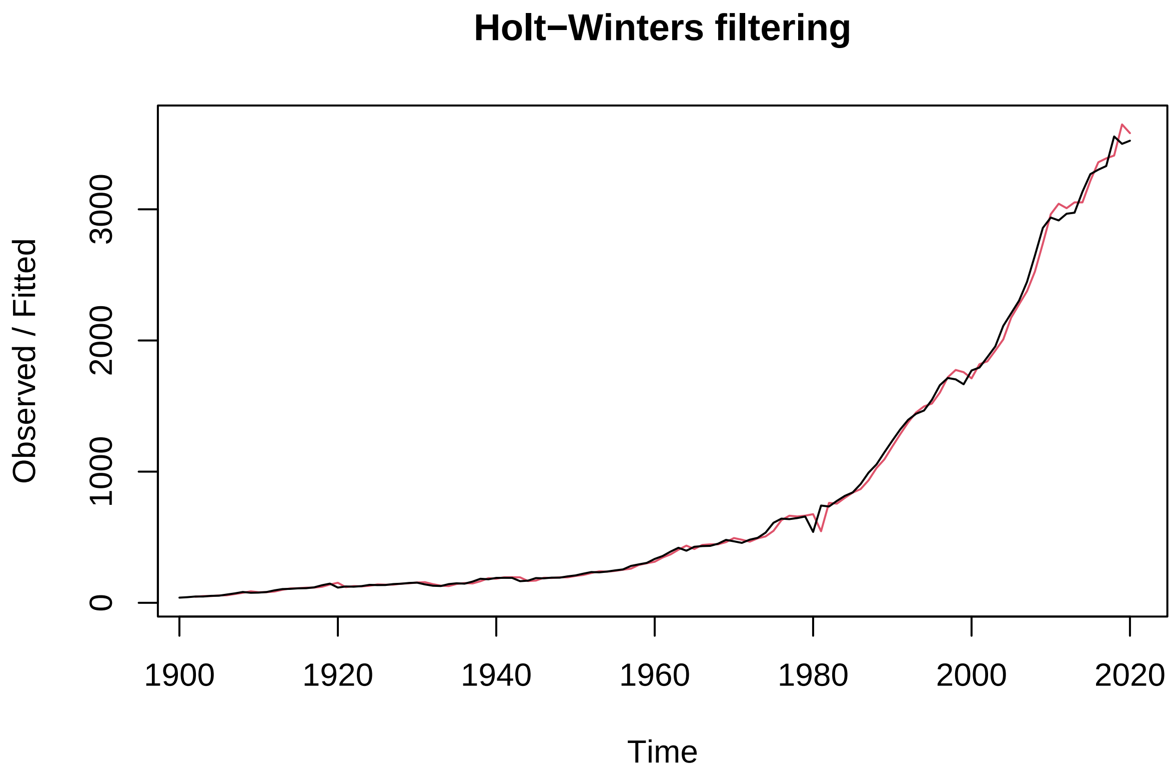
ROOT MEAN SQUARE ERROR(RMSE):

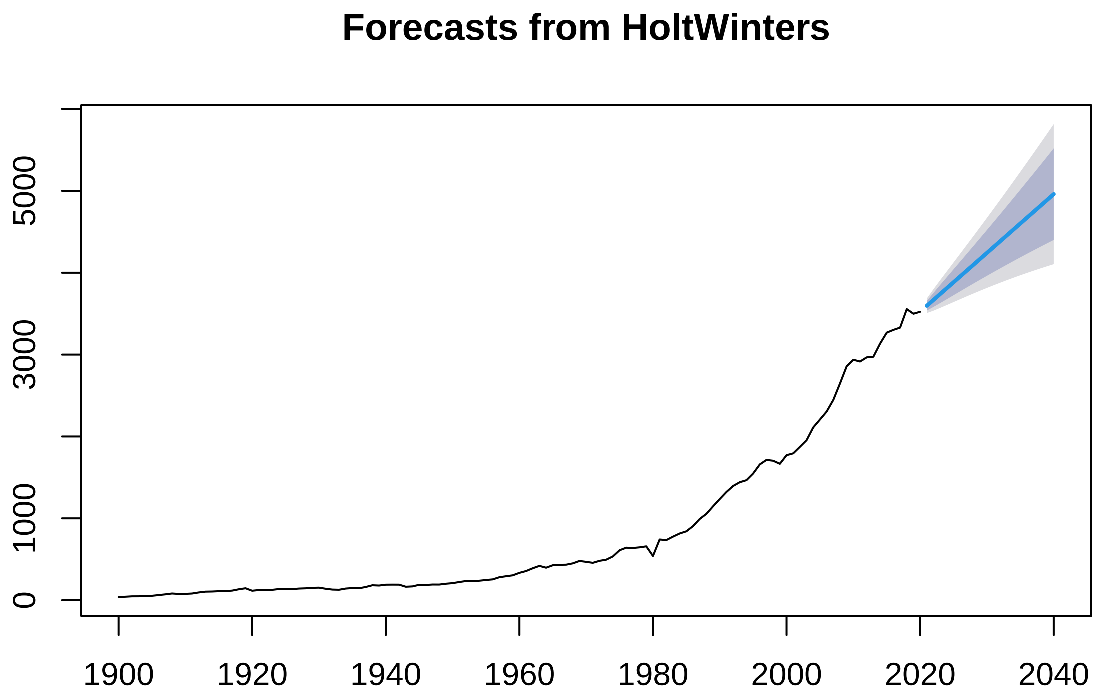
```
## [1] 602.291
```

As we can see that the RMSE of Holt Winters Model is less than the RMSE of ARIMA Model, we conclude saying that the Holt-Winters Model is a better fit than ARIMA Model for our data.
Hence, we use the Holt-Winters Model to forecast the value of coal production for the next 10 years




```
## Holt-Winters exponential smoothing with trend and without seasonal component.
##
## Call:
## HoltWinters(x = coalp_ts, gamma = FALSE)
##
## Smoothing parameters:
##  alpha: 0.9770477
##  beta : 0.1158967
##  gamma: FALSE
##
## Coefficients:
##          [,1]
## a 3524.29106
## b   71.74971
```


Holt−Winters filtering
Observed / Fitted
0
1000
2000
3000
1900
1920
1940
1960
1980
2000
2020
Time




```
##
## Forecast method: HoltWinters
##
## Model Information:
## Holt-Winters exponential smoothing with trend and without seasonal component.
##
## Call:
## HoltWinters(x = coalp_ts, gamma = FALSE)
##
## Smoothing parameters:
##  alpha: 0.9770477
##  beta : 0.1158967
##  gamma: FALSE
##
## Coefficients:
##         [,1]
## a 3524.29106
## b   71.74971
##
## Error measures:
##                      ME     RMSE      MAE        MPE     MAPE       MASE        ACF1
```

```
## Training set 5.076657 45.38516 27.2851 0.5316051 4.432046 0.7638095 -0.01375229
##
## Forecasts:
##      Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
## 2021       3596.041 3537.998 3654.084 3507.272 3684.810
## 2022       3667.790 3581.920 3753.661 3536.463 3799.118
## 2023       3739.540 3628.844 3850.236 3570.246 3908.835
## 2024       3811.290 3676.773 3945.807 3605.564 4017.016
## 2025       3883.040 3724.976 4041.103 3641.303 4124.776
## 2026       3954.789 3773.110 4136.469 3676.935 4232.644
## 2027       4026.539 3820.988 4232.091 3712.175 4340.903
## 2028       4098.289 3868.501 4328.077 3746.858 4449.719
## 2029       4170.038 3915.584 4424.493 3780.884 4559.193
## 2030       4241.788 3962.197 4521.380 3814.190 4669.387
## 2031       4313.538 4008.313 4618.763 3846.737 4780.339
## 2032       4385.288 4053.919 4716.657 3878.503 4892.073
## 2033       4457.037 4099.004 4815.070 3909.473 5004.602
## 2034       4528.787 4143.566 4914.008 3939.642 5117.932
## 2035       4600.537 4187.602 5013.471 3969.008 5232.066
## 2036       4672.286 4231.114 5113.459 3997.572 5347.001
## 2037       4744.036 4274.105 5213.967 4025.339 5462.734
## 2038       4815.786 4316.578 5314.994 4052.313 5579.259
## 2039       4887.536 4358.537 5416.534 4078.502 5696.569
## 2040       4959.285 4399.987 5518.583 4103.913 5814.658
```

Regression Analysis

Simple Linear Regression Model

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them. Here, we have fitted Simple Linear Regression for two models This analysis considers the simple linear regression model, that is, a model with a single regressor x that has a relationship with a response y that is a straight line. This simple linear regression coefficient. $Y = \beta_0 + \beta_1 X$ where the intercept β_0 and the slope β_1 are unknown constants and ϵ is a random error component. The errors are assumed to have mean zero and unknown variance σ^2. Additionally, we usually assume that the errors are uncorrelated. This means that the value of one error does not depend on the value of any other error

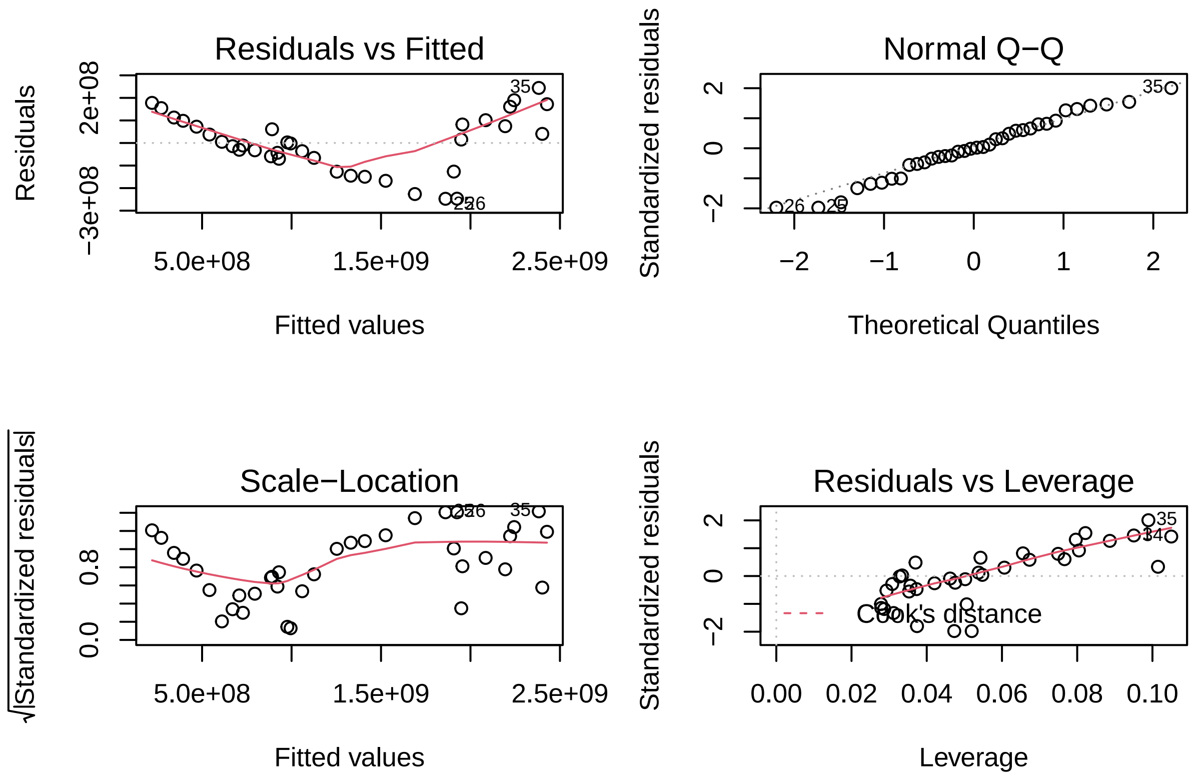
1]To check the relationship between CO2 Emission and Coal Production in India.

Here,
Dependent Variable Y = CO2 Emission
Independent Variable X = Coal Production.

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -247413131  -66990642     270578   86136719  244084732
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -464585396   58671826  -7.918 3.19e-09 ***
## x               813506      25393  32.037  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128300000 on 34 degrees of freedom
## Multiple R-squared:  0.9679, Adjusted R-squared:  0.967
## F-statistic:  1026 on 1 and 34 DF,  p-value: < 2.2e-16
```

From the output we can see that,
1) The model is, Y = -464585396 + 813506*X
2) R-squared value or strength of the model is 0.9679

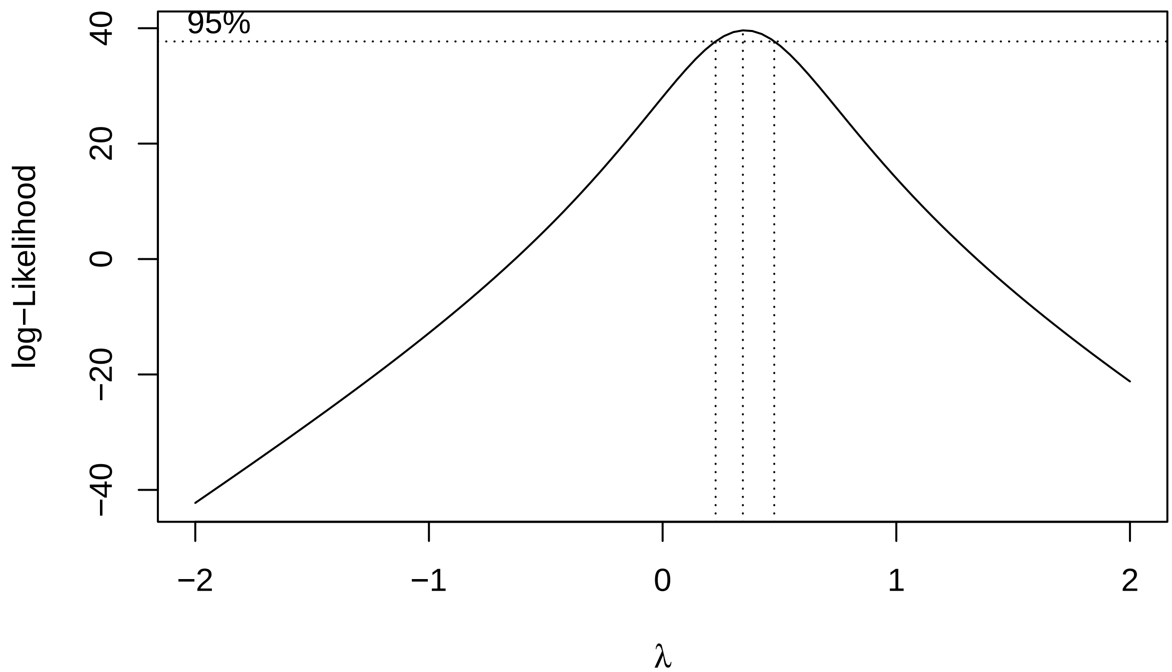
Diagnostic Plots



Conclusion:
From Residual Vs. Fitted Plot we conclude that relationship between CO2 Emission and Coal Production is not linear.

Thus, to overcome this problem we proceed to use the *Box-Cox transformation.*

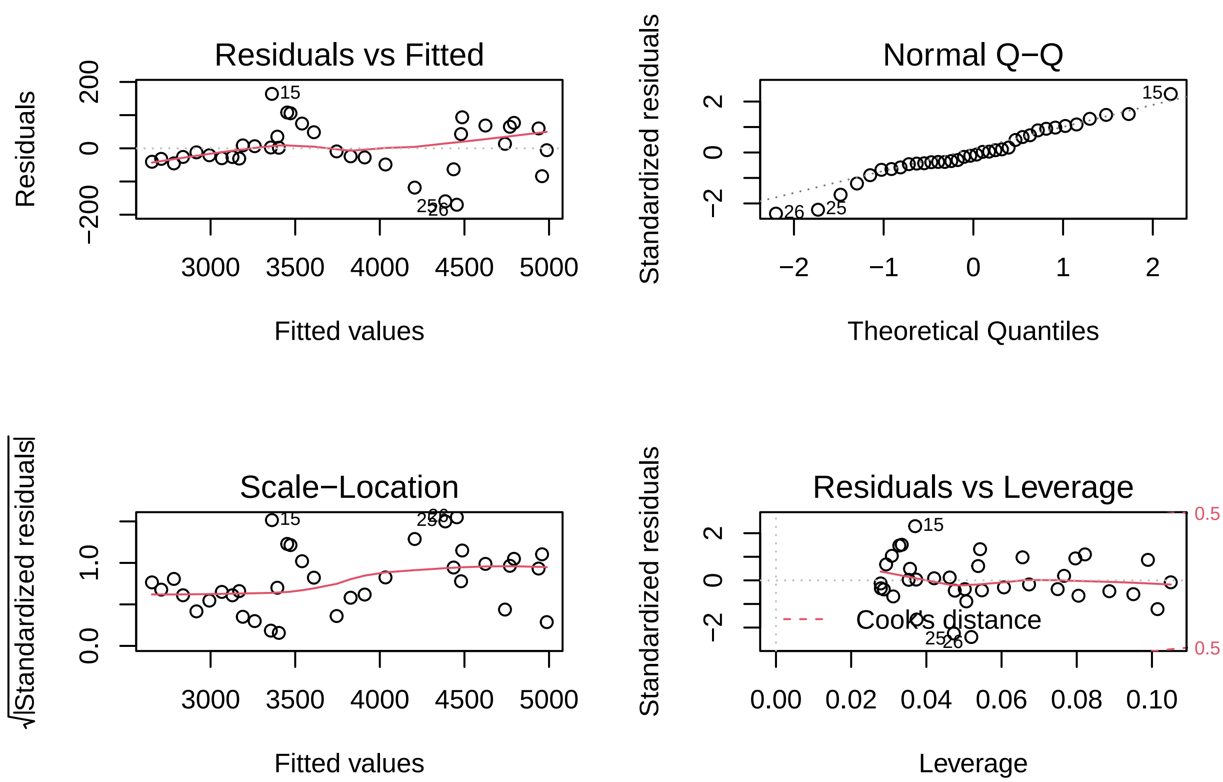
Box Cox Transformation: A Box Cox transformation is a Transformation of non-normal dependent variables into a normal shape. At the core of the Box Cox transformation is an exponent,λ which varies from -5 to 5. All values of λ are considered and the optimal value for your data is selected; The "optimal value" is the one which results in the best approximation of a normal distribution curve. The transformation of Y has the form:




```
## [1] 0.3434343

##
## Call:
## lm(formula = ((y^lamda - 1)/lamda) ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -170.017  -31.173   -7.466   51.311  163.940
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1929.9212    33.2769    58.0   <2e-16 ***
## x              0.8598     0.0144    59.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 72.74 on 34 degrees of freedom
## Multiple R-squared:  0.9905, Adjusted R-squared:  0.9903
## F-statistic:  3564 on 1 and 34 DF,  p-value: < 2.2e-16
```



Hypothesis Testing
To Test
H_o: $\beta_1 = 0$
Vs
H_1: $\beta_1 \neq 0$

Test Criteria: Reject H_o if $p - value \leq 0.05$

Decision:
Here $p - value = 2.2e^{-16} \leq 0.05$
Therefore ,Reject H_o

Conclusion:
1) There is relationship between CO2 Emission and Coal Production.
2) From the above graph of Box-Cox Transformation (Residuals Vs. Fitted), we can now conclude that the relationship between CO2 Emission and Coal Production is linear.
R-Squared: 0.9679 (for Simple Regression Model)
R-Squared: 0.9905 (for Box Cox Transformation)
The Coefficient of Determination is Relatively High for Box Cox Transformation which implies that regression model captures most of the variability expressed by CO2 Emission while using Box-Cox Transformation.

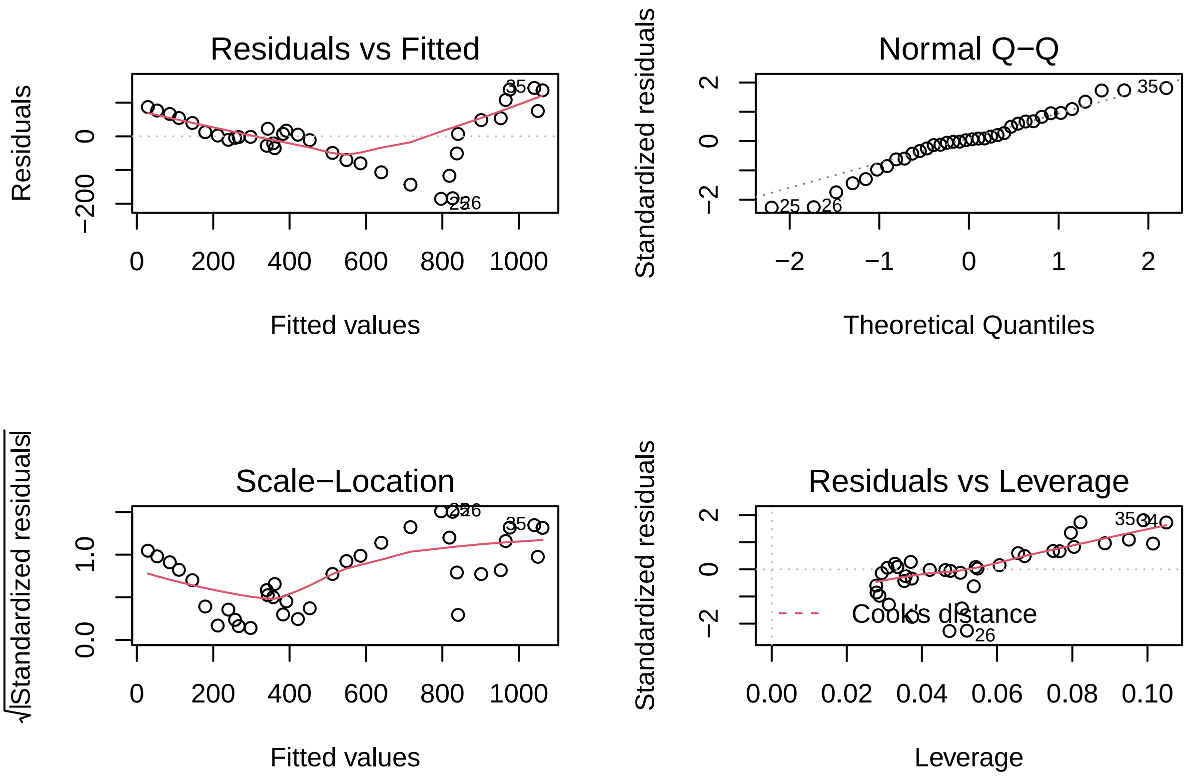
2]To check the relationship between Electricity Generation and Coal Production in India.

Here,
Dependent Variable Y = Electricity Generation
Independent Variable X = Coal Production

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -185.540  -38.776    3.613   53.848  143.716
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -291.50003   38.24981  -7.621 7.42e-09 ***
## x              0.38077    0.01655  23.001  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.61 on 34 degrees of freedom
## Multiple R-squared:  0.9396, Adjusted R-squared:  0.9378
## F-statistic:   529 on 1 and 34 DF,  p-value: < 2.2e-16
```

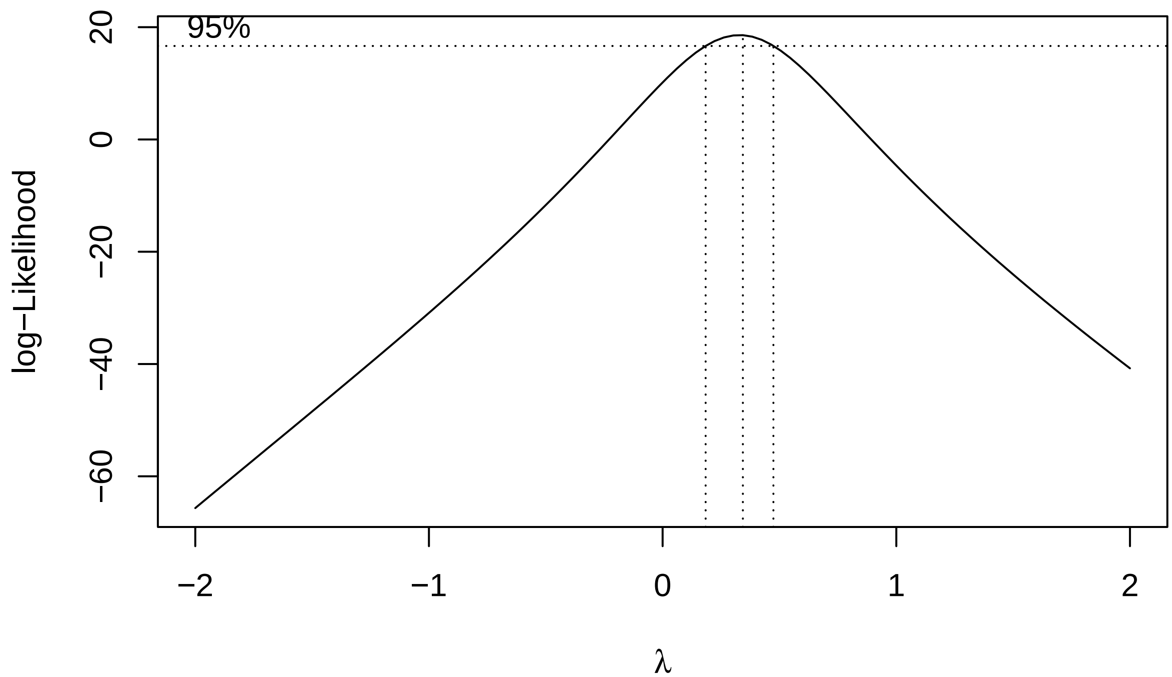
From the output we can see that,
1) The model is, Y = Y = -291.50003 + 0.38077*X
2) R-squared value or strength of the model is 0.9396

Diagnostic Plots

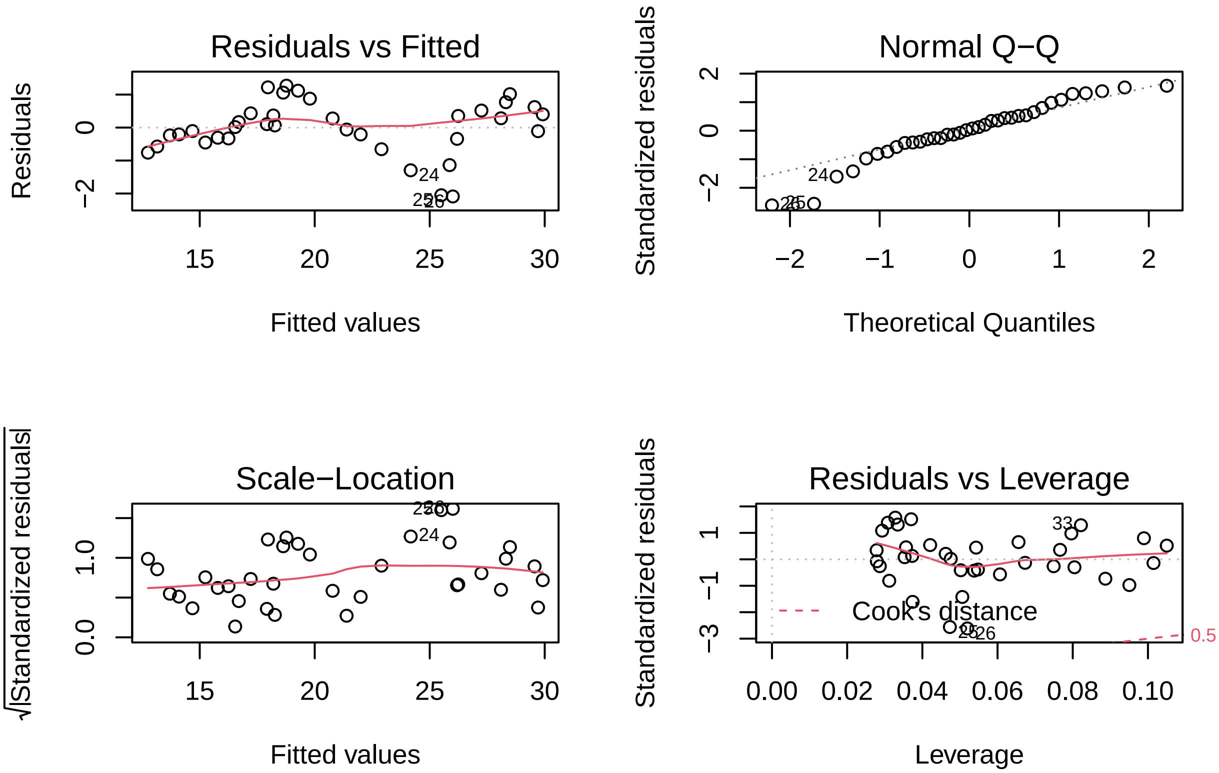


Conclusion:
From Residual Vs. Fitted Plot we conclude that relationship between Electrcity Generation and Coal Production is not linear.
Thus, to overcome this problem we proceed to use the *Box-Cox transformation.*

Box Cox Transformation:




```
##
## Call:
## lm(formula = ((y^lamda - 1)/lamda) ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08826 -0.33342  0.03946  0.45318  1.27034
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.4244594  0.3753566   19.78   <2e-16 ***
## x           0.0063259  0.0001625   38.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8205 on 34 degrees of freedom
## Multiple R-squared:  0.9781, Adjusted R-squared:  0.9774
## F-statistic:  1516 on 1 and 34 DF,  p-value: < 2.2e-16
```



Hypothesis Testing

To Test

H_o: $\beta_1 = 0$

Vs

H_1: $\beta_1 \neq 0$

Test Criteria: Reject H_o if $p - value \leq 0.05$

Decision:

Here $p - value = 2.2e^{-16} \leq 0.05$

Therefore ,Reject H_o

Conclusion:

1)From the above graph of Box-Cox Transformation (Residuals Vs. Fitted), we can now conclude that the relationship between Electricity Generation and Coal Production is linear.

R-Squared: 0.9396 (for Simple Regression Model)

R-Squared: 0.9781 (for Box Cox Transformation)

The Coefficient of Determination is Relatively High for Box Cox Transformation which implies that regression model captures most of the variability expressed by Electricity Generation while using Box-Cox Transformation.

Conclusion

Through the time series plot, we can see that there has always been an increase in Consumption of Coal. With the help of extrapolation, we tried to remove the irregularity observed in the data and then forecast using the ARIMA Model.

We also fit, both Holt-Winters Exponential Smoothing and ARIMA Model for the data of coal production and compared the accuracy of the models. We conclude that the Holt-Winters Exponential Smoothing is a better model for our data.

With the help of Linear Regression, we compare the relation between Coal Production and CO2 Emission, and Coal Production and Electricity Generation. Although at first we don't find there to be a linear relationship, after using the Box-Cox Transformation, we conclude by saying that there is a Linear Relationship in both the models.

With the help of forecasting, we can say that the demand of coal in the country will on increasing with time. But as we now that there is a shortage of Coal in recent times and hence depending on Coal entirely as the primary source of energy may not be the best option. Hence we have studied of Solar Energy as an alternative in particular as it in an upcoming source of Energy and it's installation is quite easy as compared to tidal energy or wind energy.

We have done Exploratory Data analysis using primary data. Here we analyse the data through various Bar Graphs. These graphs help us understand the awareness of Solar Energy among people and their willingness to move towards the usage of clean energy with respect to different attributes such as Age, Occupation and Education.

Finally, with the help of Chi-Square Test of Independence, we were able to determine the relation between Education and awareness of solar energy and its policies. It was found that they are independent and hence Education does not affect the level of awareness of Solar among people. Similarly, there was no relationship found between Occupation and people's willingness to install Solar panels at home. Finally, we also came to know that education level does not affect the awareness of government policies related to solar energy.

Limitations and Scope

1]As the data of Indian Coal Prices was not available, we could not perform Multiple Linear Regression to check relation between dependent variable Indian Coal Prices and independent variables International Coal Prices, Import and Export of Coal in India.
2]Due to unavailability of data of coal consumption and production for the year 2021, we could not compare the forecasted value to check the accuracy of the model as compared with the real life data.
3]Due to restriction of time, we were not able to collect large amount of primary data from Google forms and were restricted to 358 responses.

We have studied Solar Energy as the non conventional alternative of Coal as it can be easily utilised in our day to day lives. The installation and use of solar energy is quite easy. But we would also like to tap into the different non-conventional sources of energy such as tidal energy, wind energy, geothermal energy , etc.

Softwares used in project

1]R-SOFTWARE:

R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. R is used among data miners and statisticians for data analysis and developing statistical software. Users have created packages to augment the functions of the R language.

2]R- Markdown: R-Markdown makes use of Markdown syntax. Markdown is a very simple 'markup' language which provides methods for creating documents with headers, images, links etc. from plain text files, while keeping the original plain text file easy to read. You can convert Markdown documents to many other file types like.

3]EXCEL:

Microsoft excel is powerful data visualization and analysis software, which uses spreadsheets to store, organize, and track data sets with formulas and functions.

References

Websites:
1)https://coal.gov.in
2)https://www.statisticshowto.com
3)thetoprated.in
4)https://rmarkdown.rstudio.com
5)https://ourworldindata.org

Books:
1) Introduction to Linear Regression Analysis by Douglas C. Montgomery
2) The Analysis of Time series An Introduction by Chris Chatfield