PROJECT 1

FORECASTING PRODUCTION OF COCONUT IN TAMILNADU

INTRODUCTION:

Coconut is widely cultivated in various regions of TamilNadu. Almost every part of coconut tree can be used in our daily life. Coconut palm is grown under varying climatic and soil conditions. It provides food, fuel and timber. The palm tolerates wide range in intensity and distribution of rainfall. A well distributed rainfall of about 200cm per year is the best for proper growth and higher yield. In areas of inadequate rainfall with uneven distribution, irrigation is required.

Sampling design and data collection:

1)The broad plan for the survey determined by statistical advisor, Indian Agriculture Statistic Research Institute, New Delhi, utilizing multi-stage stratified random sampling.

2) Blocks within districts served as strata. Sampling occurred in two stages: villages as first-stage units and gardens within villages as second-stage units. 3)Field workers, aided by block statistical inspectors, collected data using predesigned schedules. Selected villages were visited, and two coconut gardens were randomly chosen for enumeration.

ABOUT DATASET:

This dataset contains annual data about the area under cultivation, yield rate per hectare and the production from the year 2003 to 2023 all over TamilNadu. This is the annual data which shows us the details that is collected from all the places in TamilNadu where coconut is cultivated.

This dataset provides information on coconut production over a span of twenty years, from 2003-04 to 2022-23. It includes three main variables for each year:

1. Area under cultivation: This refers to the total land area used for growing coconuts in hectares.

2. Yield rate per hectare: This represents the average number of coconuts produced per hectare of cultivated land.

3. Total production: This indicates the overall quantity of coconuts harvested in a given year.

By analyzing this dataset, one can track trends and patterns in coconut production over time, identify factors influencing productivity, and assess the overall performance of the coconut industry. Additionally, it can be useful for policymakers, researchers, and stakeholders in the agricultural sector for planning and decision-making purposes.

OBJECTIVE:

To forecast the production of coconut in TamilNadu for the next 5 years.

DATA CLEANING

```
> coconut=read.table(file.choose(),header=T,sep=",")
> head(coconut)
     Year Area.under.coconut..in.ha.. Yield.rate.per.hectare.nuts.in.Nos.
1 2003-04                      352710                                7261
2 2004-05                      357056                               11474
3 2005-06                      370515                               13782
4 2006-07                      374604                               14495
5 2007-08                      383366                               14186
6 2008-09                      389429                               13769
  Production
1      25605
2      40970
3      48671
4      54299
5      54386
6      53620
```

Summary of the data:

```
> summary(coconut)
     Year           Area.under.coconut..in.ha..
 Length:20          Min.   :352710
 Class :character   1st Qu.:387913
 Mode  :character   Median :426004
                    Mean   :413942
                    3rd Qu.:436446
                    Max.   :472711
 Yield.rate.per.hectare.nuts.in.Nos.   Production
 Min.   : 7261                       Min.   :25605
 1st Qu.:11086                       1st Qu.:46968
 Median :11674                       Median :51018
 Mean   :12124                       Mean   :50682
 3rd Qu.:13935                       3rd Qu.:54657
 Max.   :14799                       Max.   :67989
```

Structure of the data (descriptive statistics):

```
> str(coconut)
'data.frame':   20 obs. of  4 variables:
 $ Year                           : chr  "2003-04" "2004-05" "2005-06" "2006-07" ...
 $ Area.under.coconut..in.ha..    : int  352710 357056 370515 374604 383366 389429 400463 410149 419400 424165 ...
 $ Yield.rate.per.hectare.nuts.in.Nos.: int  7261 11474 13782 14495 14186 13769 13851 14545 14799 10833 ...
 $ Production                     : int  25605 40970 48671 54299 54386 53620 55471 59656 62009 50753 ...
```

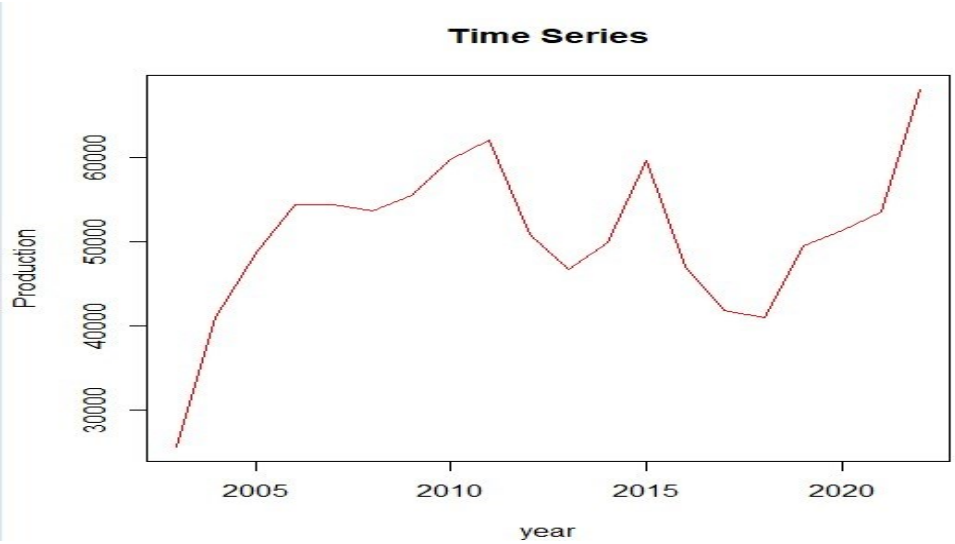
Check for missing values:

```
> colSums(is.na(coconut))
                              Year         Area.under.coconut..in.ha..
                                 0                                   0
Yield.rate.per.hectare.nuts.in.Nos.                          Production
                                 0                                   0
```

There is no missing value in the dataset.

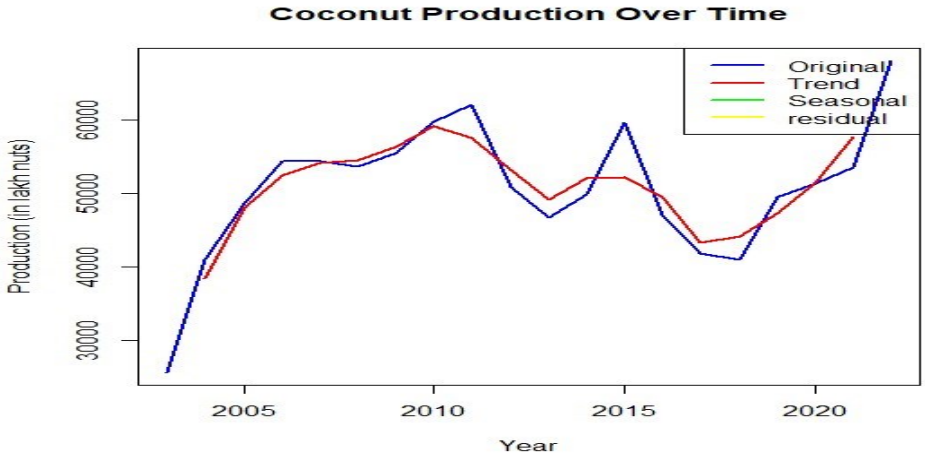
Check for frequency:

```
> #Specify the frequency of the Time index
> coconut_ts =ts(data$Production, start = c(2003), frequency = 1)
> #Plot for time series
> plot(coconut_ts,main="Time Series",xlab="year",ylab="Production",col="red")
```



Exploratory Data Analysis:

```
> #Calculate moving averages for trend ,seasonality and residual
> # Trend:
> trend_window= 3
> trend =filter(coconut_ts, rep(1/trend_window, trend_window), sides = 2)
> # Seasonality:
> seasonal= coconut_ts - trend
> #residual
> residual= coconut_ts - (trend + seasonal)
> # Step 4: Plot the original series, trend, and seasonality
> plot(coconut_ts, type = "l", col = "blue", lwd = 2, main = "Coconut Production Over Time",
+      xlab = "Year", ylab = "Production (in lakh nuts)")
> lines(trend, col = "red", lwd = 2)
> lines(seasonal, col = "green", lwd = 2)
> lines(residual, col = "yellow", lwd = 2)
> legend("topright", legend = c("Original", "Trend", "Seasonal","residual"),
+        col = c("blue", "red", "green","yellow"), lty = 1, lwd = 2)
```



By looking the plot, we can identify that there is no seasonality and some trend occur in the dataset. When dealing with trends in time series data without seasonality, several methods can be employed like moving average, exponential smoothing method, ARIMA model.

Check for stationarity:

Augmented Dickey-Fuller (ADF):

The Augmented Dickey-Fuller (ADF) test is a common statistical test used to determine whether a unit root is present in a time series dataset. It helps in assessing whether a time series is stationary or not. A stationary time series has constant mean and variance over time, making it easier to model and analyse.

Stationarity testing Hypothesis:

Null Hypothesis:

Data has unit roots which means data is not stationary.

Alternative Hypothesis:

Data does not have unit roots which means data is stationary.

```
> stationary_data=adf.test(data$Production)
> stationary_data

        Augmented Dickey-Fuller Test

data:  data$Production
Dickey-Fuller = -1.7995, Lag order = 2, p-value = 0.6488
alternative hypothesis: stationary
```

since the p-value is 0.6488 suggests that fail to reject the null hypothesis, indicating that the time series is likely non stationary. Using first order differencing to achieve stationary.

```
> differencing1=diff(data$Production)
> stationary_data1=adf.test(differencing1)
> stationary_data1

        Augmented Dickey-Fuller Test

data:  differencing1
Dickey-Fuller = -2.2382, Lag order = 2, p-value = 0.4816
alternative hypothesis: stationary
```

After differencing again check for stationarity using ADF test, again the p- value is greater than 0.05 we fail to reject the null hypothesis. Then, go for second order differencing.

```
> differencing2=diff(differencing1)
> stationary_data2=adf.test(differencing2)

> stationary_data2

        Augmented Dickey-Fuller Test

data:  differencing2
Dickey-Fuller = -6.9391, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

After differencing twice the data achieve stationarity.

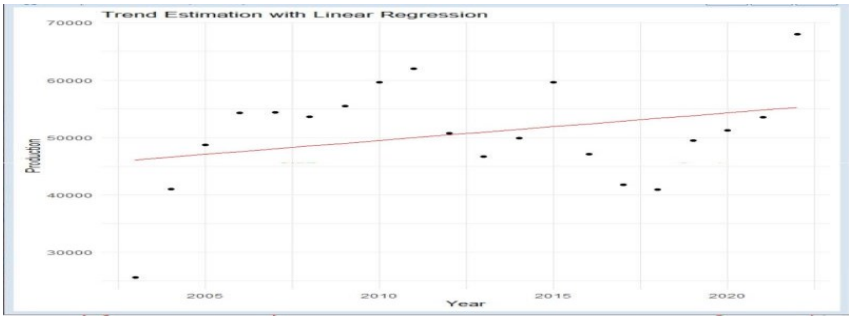
Splitting the dataset into Train set and Test set:

Splitting data is to evaluate model performance on unseen data, prevent data leakage, validate model assumptions, and facilitate effective parameter tuning, ensuring more reliable forecasts and insights. The training set is used to train the model, while the test set is used to evaluate its performance on unseen data. This helps ensure the model can generalize well to new observations and prevents overfitting.

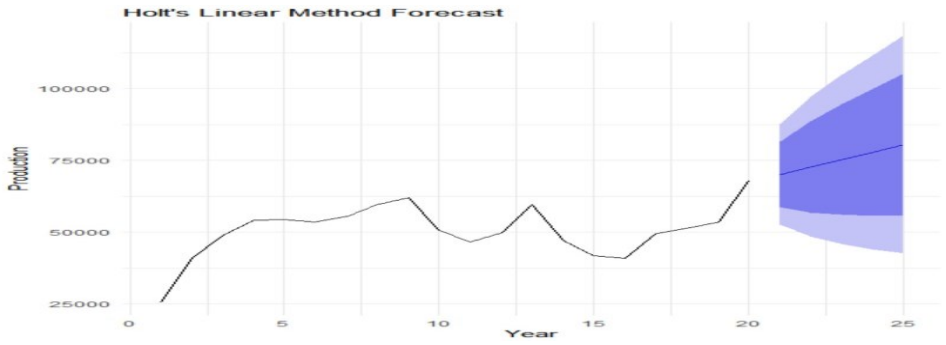
```
> train_index=sample(1:length(coconut_ts),0.8*length(coconut_ts))
> train_set=coconut_ts[train_index]
> length(train_set)
[1] 16
> test_set=coconut_ts[-train_index]
> length(test_set)
[1] 4
```

HOLT LINEAR MODEL:

Holt's linear method is a forecasting technique used for time series data with a linear trend. It involves estimating two components: the level (average value) and the trend (rate of change). These components are updated at each time step using smoothing parameters. The forecast is then generated by combining the level and trend components. This method is suitable for data exhibiting a consistent linear trend over time.



```
> library(forecast)
> #fit holt's linear trend model
> holt_model=holt(coconut$Production,h=5)  #forecasting for next 5 periods
> holt_model
   Point Forecast     Lo 80     Hi 80     Lo 95     Hi 95
21        70149.96 58735.01  81564.91 52692.30  87607.62
22        72761.06 56920.80  88601.31 48535.48  96986.63
23        75372.15 56096.48  94647.82 45892.56 104851.74
24        77983.24 55797.34 100169.14 44052.84 111913.65
25        80594.34 55837.48 105351.19 42731.99 118456.68
> #finding accuracy measures
> accuracy=accuracy(holt_model)
> accuracy
                     ME      RMSE       MAE        MPE     MAPE      MASE
Training set -1272.631 7966.782 6025.114 -4.650825 13.7737 1.022428
                     ACF1
Training set -0.01558853
> library(ggplot2)
> # Plot the original data and the forecast
> autoplot(holt_model) +
+   labs(x = "Year", y = "Production", title = "Holt's Linear Method Forecast") +
+   theme_minimal()
```



From the plot of holt linear method, we can see that the production of coconut and arecanut for next 5 years will move in an upward trend.

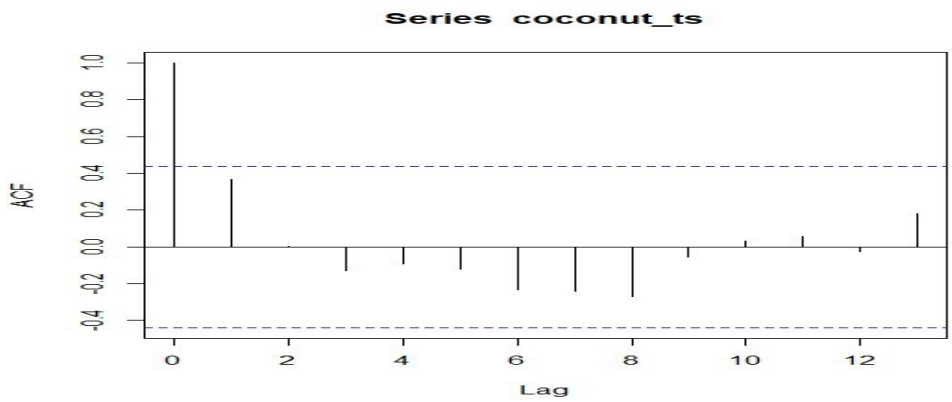
ARIMA MODEL:

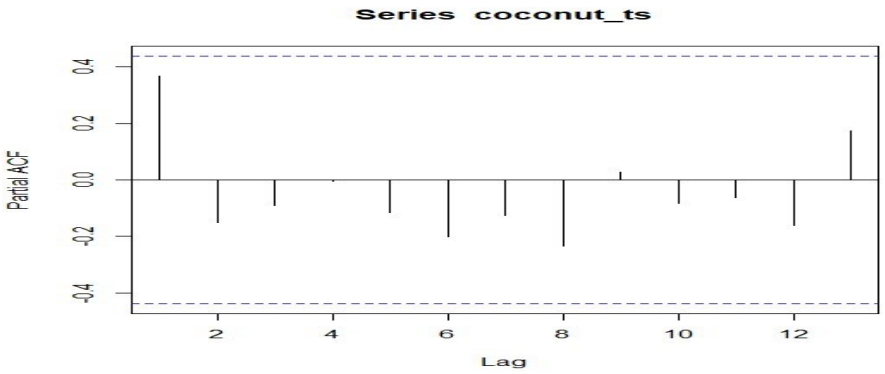
ARIMA, or AutoRegressive Integrated Moving Average, is a widely used time series forecasting method. It combines autoregression (AR), differencing (I), and moving average (MA) components to model complex patterns in data. By selecting appropriate parameters for lag observations, differencing, and moving average window size, ARIMA can provide accurate forecasts.

ACF and PACF Plot for original dataset:

ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) plots are tools used in time series analysis to identify patterns in the data. ACF measures the correlation between a series and its lagged values, while PACF measures the correlation between a series and its lagged values after removing the effects of intermediate lags. These plots help in determining the order of ARIMA models and identifying potential autocorrelation and partial autocorrelation patterns in the data.

```
> acf(coconut_ts)
> pacf(coconut_ts)
```

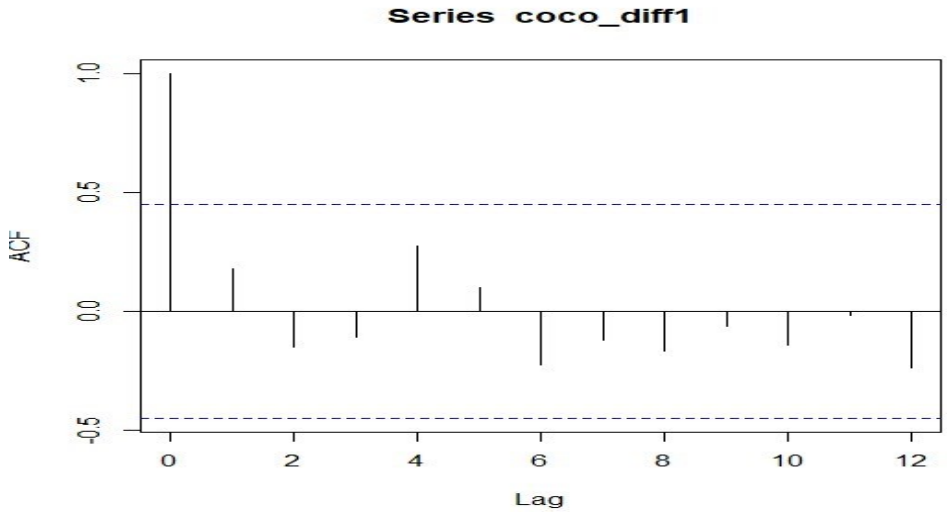


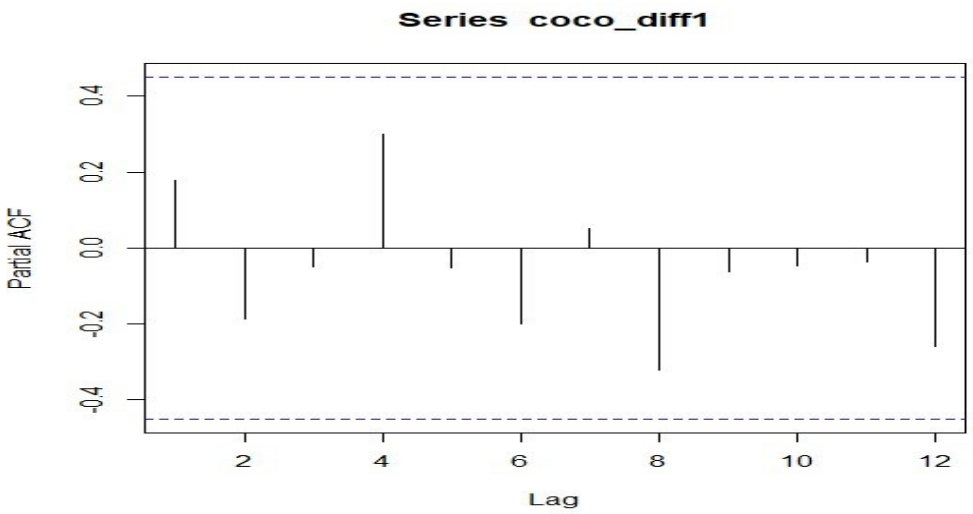


ACF and PACF for differencing data to achieve stationarity:

Since the dataset is non stationarity to achieve stationarity using differencing, acf and pacf of first differencing is given below.

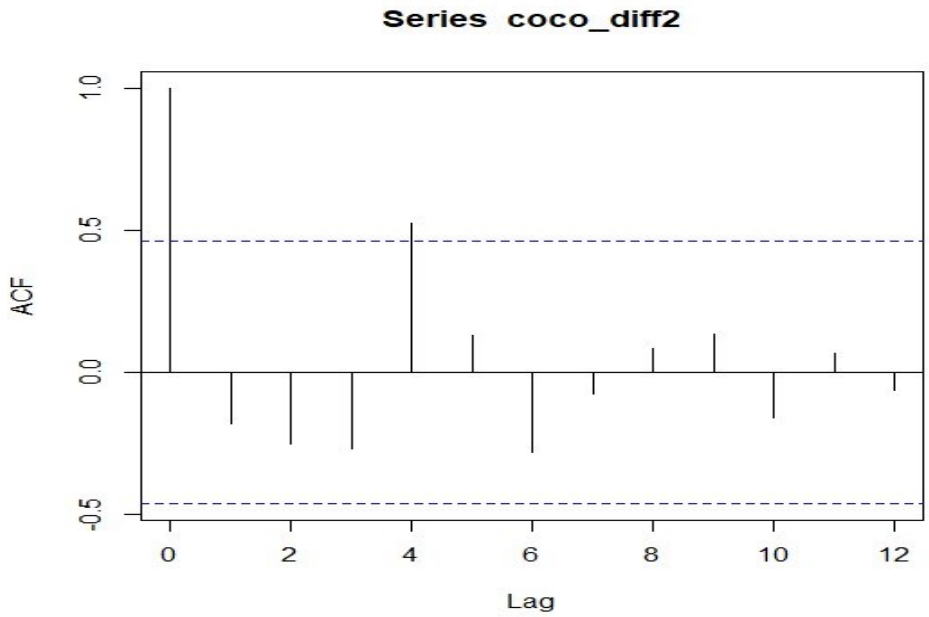
```
> coco_diff1=diff(coconut_ts)
> acf(coco_diff1)
> pacf(coco_diff1)
```

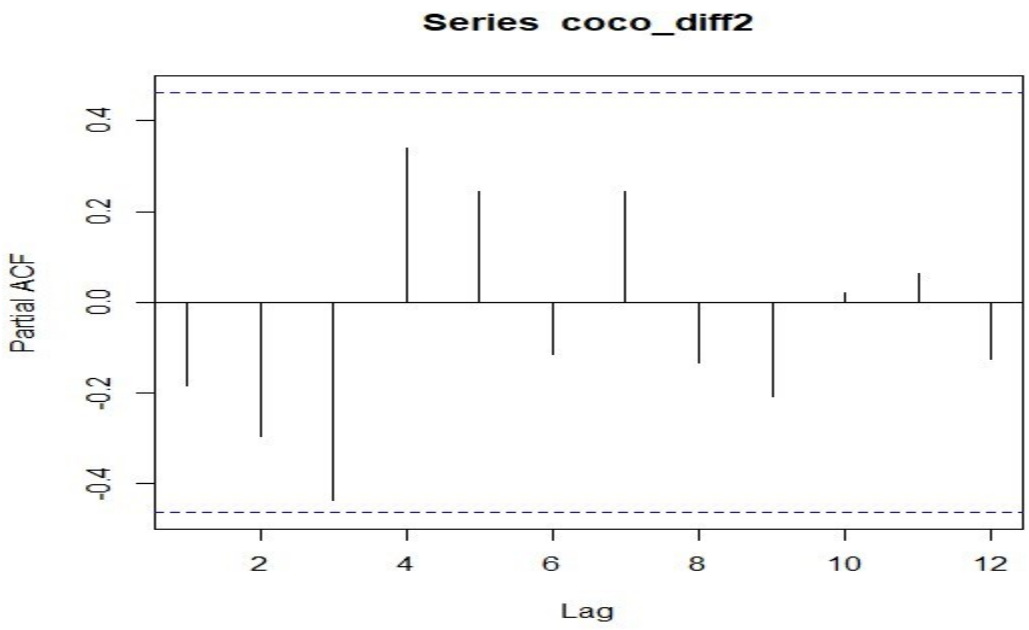




ACF and PACF plot for second order differencing:

```
> coco_diff2=diff(coco_diff1)
> acf(coco_diff2)
> pacf(coco_diff2)
```





Following figure shows the ACF and PACF plot of a simulated time series. We plotted the sample ACF and PACF of the second differenced data. And the ARMA model for the differenced data is ARMA(1,0). (i.e) For original data the model is ARIMA(1,2,0).

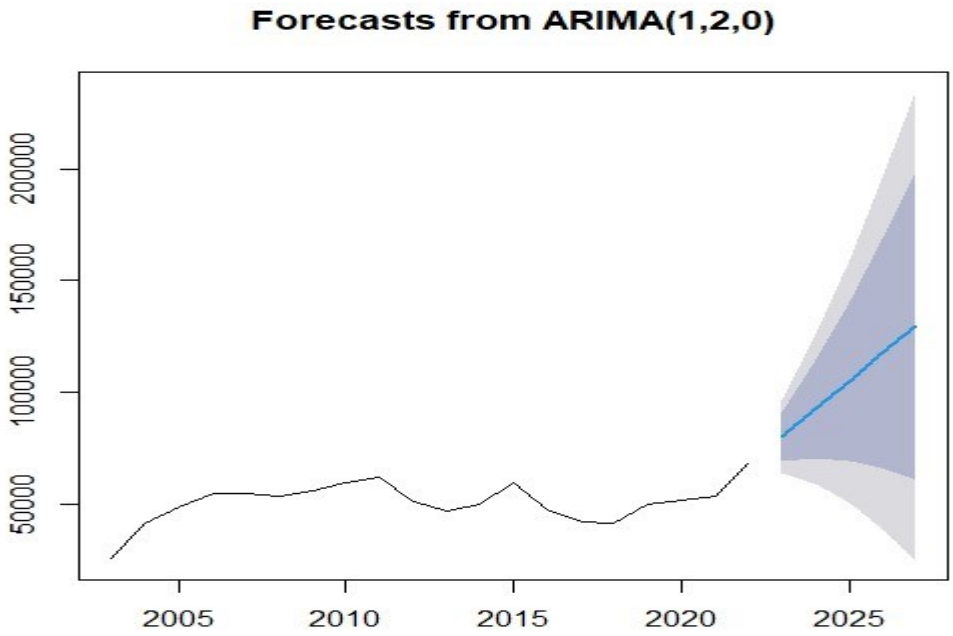
FORECAST FOR NEXT 5 YEARS (using ARIMA(1,2,0)):

```
> library(forecast)
> train_index=sample(1:length(coconut_ts),0.8*length(coconut_ts))
> train_set=coconut_ts[train_index]
> test_set=coconut_ts[-train_index]
> arima_model=arima(train_set,order=c(1,2,0))
> forecast_result=forecast(arima_model,h=5)
> accuracy_result=accuracy(forecast_result,test_set)
> print(accuracy_result)
                     ME      RMSE       MAE       MPE     MAPE      MASE
Training set  -85.55709 17499.987 12592.150 -6.063591 28.22640 1.0454834
Test set     4970.20149  7138.697  5696.852  8.715446 10.25941 0.4729903

> arima_model=arima(coconut_ts,order=c(1,2,0))
> arima_model

Call:
arima(x = coconut_ts, order = c(1, 2, 0))

Coefficients:
          ar1
      -0.2039
s.e.   0.2436

sigma^2 estimated as 69208734:  log likelihood = -188.04,  aic = 380.07
```

This ARIMA model is a second-order differencing (d=2) with an autoregressive term (p=1) and no moving average term (q=0). The coefficient for the autoregressive term is -0.2039 suggests that there is a negative relationship between past observations and future values, as indicated by the negative coefficient for the autoregressive term, with a standard error of

0.2436. The log likelihood is -188.04, and the AIC (Akaike Information Criterion) value is 380.07. The estimated variance is 69208734.

```
> forecast_result=forecast(arima_model,h=5)
> print(forecast_result)
     Point Forecast    Lo 80     Hi 80    Lo 95    Hi 95
2023       79965.71 69304.25  90627.16 63660.41  96271.0
2024       92451.24 70534.42 114368.05 58932.36 125970.1
2025      104833.01 69212.61 140453.41 50356.30 159309.7
2026      117235.94 65933.90 168537.98 38776.24 195695.6
2027      129634.56 60877.04 198392.07 24479.01 234790.1
> accuracy_result=accuracy(forecast_result)
> print(accuracy_result)
                     ME     RMSE      MAE        MPE     MAPE     MASE
Training set -169.8953 7892.266 5947.617 -0.6313029 11.59535 1.009277
                     ACF1
Training set -0.06660315
> plot(forecast_result)
```



INTERPRETATION:

We have forecasted the production of coconut for the next 5 years. By holt linear method we found that RMSE is 7966.782 and MAPE is 13.7737 and by ARIMA method we found that RMSE is 7892.266 and MAPE is 11.59535. The forecasted production of coconut for the next 5 years obtained from holt linear model are 70149.96, 72761.06, 75372.15, 779883.24, 80594.34 and those obtained from ARIMA model are 79965.71, 92451.24, 104833.01, 117235.94, 129634.56. Though the error in ARIMA model is less than the holt linear model we can see that there is sudden increase in production of coconut which is not possible but in holt linear model the production of coconut increases gradually. Therefore, we can apply holt linear method. However, this prediction is only applicable when all the conditions are conventional. If there are any sudden change in weather conditions or any other unpredictable factor occurs this prediction cannot be reliable.