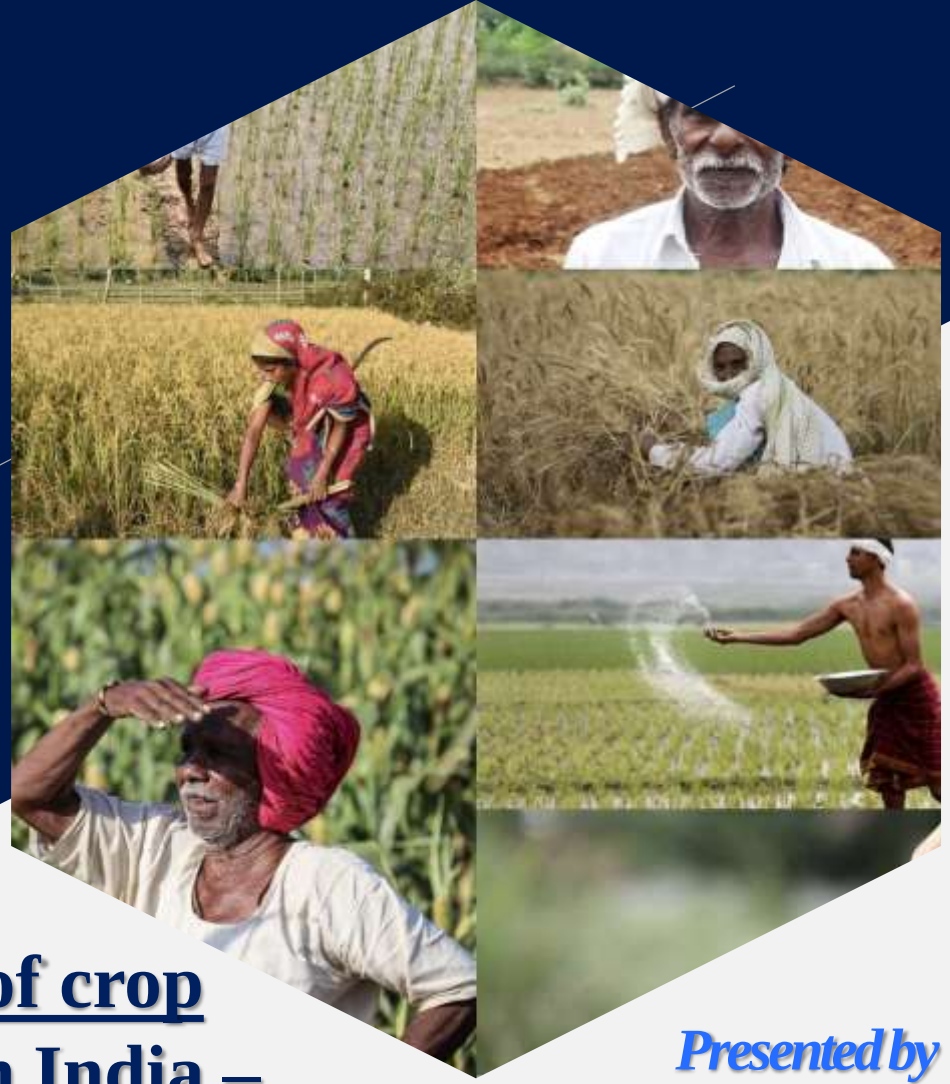


Two types of crop production in India – A case study and detailed data analysis

Presented by

Papu Mahanta

Sem-Vi, Roll no- B.Sc.(Sem-Vi) Stat-07

Department of Statistics, Visva-Bharati University

Contents

Introduction:

In India, agriculture is the backbone of the economy, and it is the primary source of livelihood for a majority of the population. India has a diverse climate, which allows for two crop seasons, Kharif and Rabi.

Kharif is the summer crop season, which starts in June and ends in October. It is also known as the monsoon season, as it is heavily reliant on rainfall.

On the other hand, Rabi is the winter crop season, which starts in October and ends in March. Rabi crops are sown after the monsoon season and rely on irrigation.

Kharif crops in India :-

Rice, Maize, Millet, Cotton, Sugarcane, Groundnut, Soybean, Moong dal, Urad dal

Rabi crops in India :-

Wheat, Barley, Mustard, Peas, Lentil, Rapeseed, Flaxseed, Cumin

DATA DESCRIPTION

My data is about agriculture land use, production, Productivity of crop and percentage coverage under irrigation. It's a time series data. My data given below. We collect the data from year 1962 to 2021. Here we have considered two types of crops, kharif and rabi.

Data source:-
://www.indiastat.com/table/agriculture/season-wise-area-production-productivity-rice-indi/7264

C:/Users/USER/OneDrive/Desktop/PROJECT/crop production in india.csv

In this slide we have shown the first part of the data which we have analyzed to prepare our report:-

Year	Area (In ' 000 Hectare)			Production (In ' 000 Tonne)			Productivity (In Kg./Hectare)			%age Coverage under Irrigation
	Kharif	Rabi	Total	Kharif	Rabi	Total	Kharif	Rabi	Total	
1962-1963	34999	696	35695	32340	877	33217	924	1260	931	37.4
1963-1964	35176	633	35809	36175	823	36998	1028	1300	1033	37.1
1964-1965	35792	670	36462	38388	920	39308	1073	1373	1078	37.3
1965-1966	34826	643	35470	29429	1161	30590	845	1806	862	36.5
1966-1967	33933	1318	35251	28622	1816	30438	843	1378	863	37.9
1967-1968	34913	1524	36437	35313	2299	37612	1011	1509	1032	38.6
1968-1969	35307	1660	36967	37127	2634	39761	1052	1587	1076	38.4
1969-1970	35828	1852	37680	37591	2839	40430	1049	1533	1073	38.2
1970-1971	35958	1641	37592	39559	2666	42225	1100	1625	1123	38.4
1971-1972	36087	1671	37758	39992	3076	43068	1108	1841	1141	37.2
1972-1973	35075	1612	36688	36324	2921	39245	1036	1811	1070	39.1
1973-1974	36488	1798	38286	40904	3147	44051	1121	1750	1151	38.4
1974-1975	35958	1931	37889	35926	3653	39579	999	1892	1045	38.8
1975-1976	37441	2034	39475	44745	3995	48740	1195	1964	1235	38.7
1976-1977	37107	1404	38511	39266	2651	41917	1058	1888	1088	38.4
1977-1978	38416	1866	40282	48947	3724	52671	1274	1995	1308	40.2

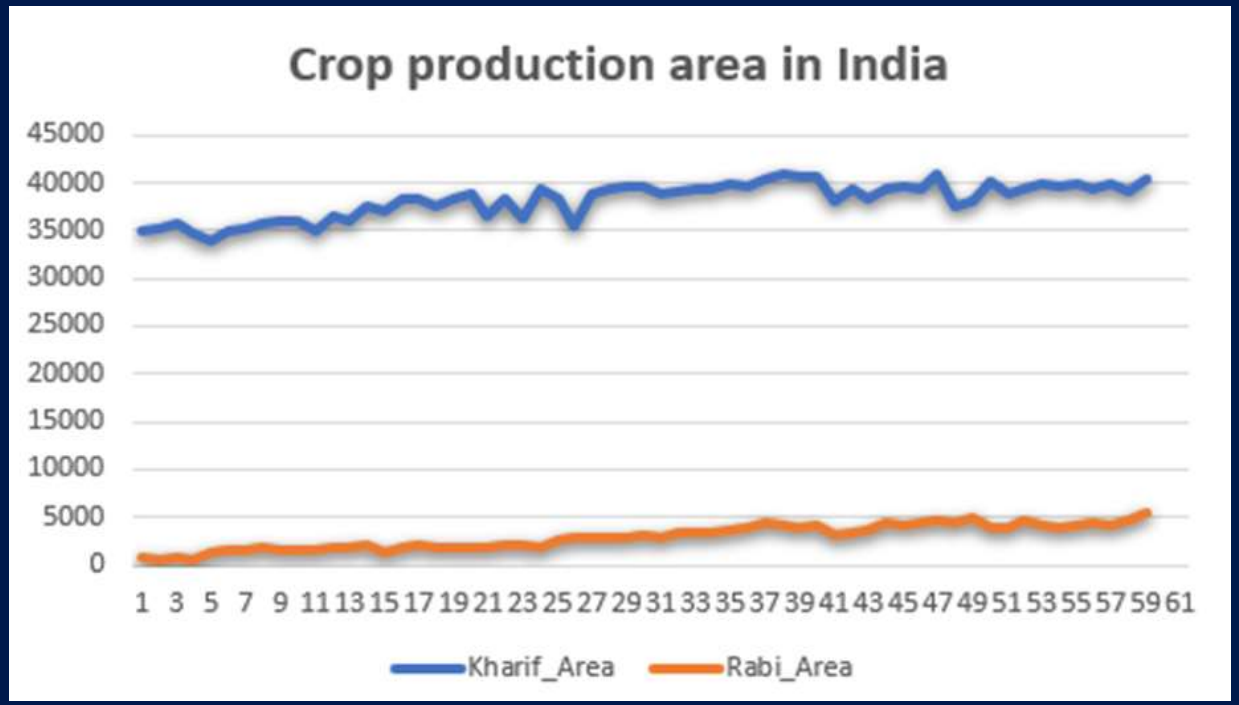
Objective

1. Analysis Trend by Moving Average
2. Autocorrelation and partial autocorrelation function
3. Model fitting
4. Forecasting
5. Simple Regression on total productivity and percentage coverage under irrigation

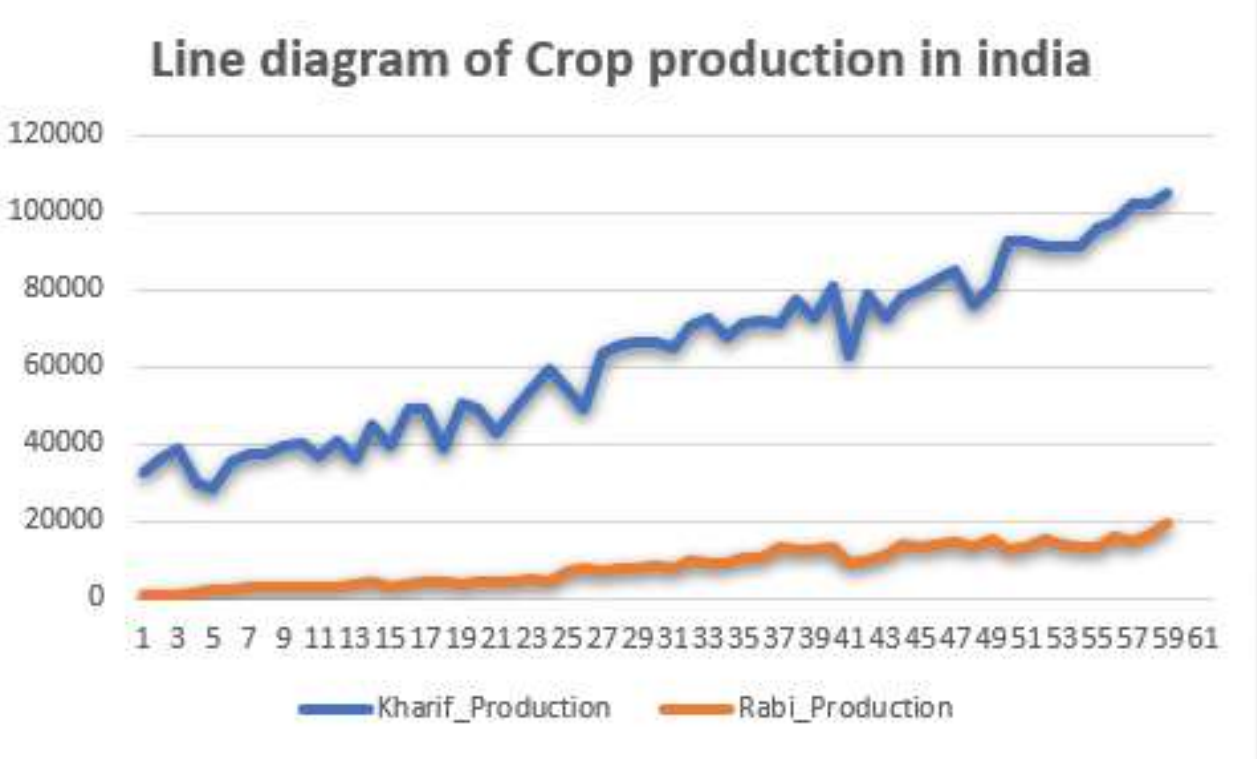
Descriptive Analysis:-

Diagrams like graphs, charts, maps, pictures etc. are attractive and effective means for presentation of statistical data. It is more effective than tabular representation, being easily intelligible to a layman.

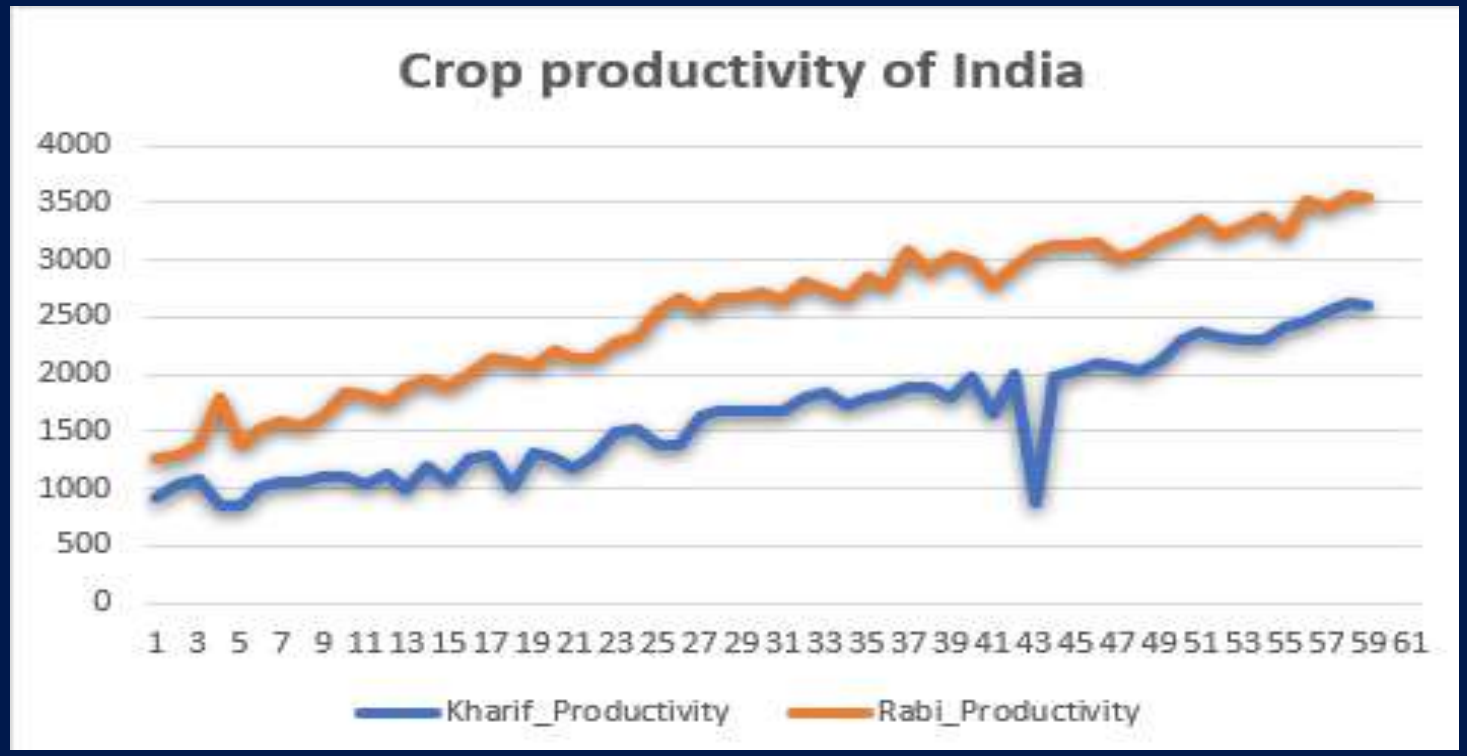
Line Diagram:



INTERPRETATION- The following line graphs shows that the kharif production area is constant with respect to time but Rabi crop production area is slightly increasing with respect to time.

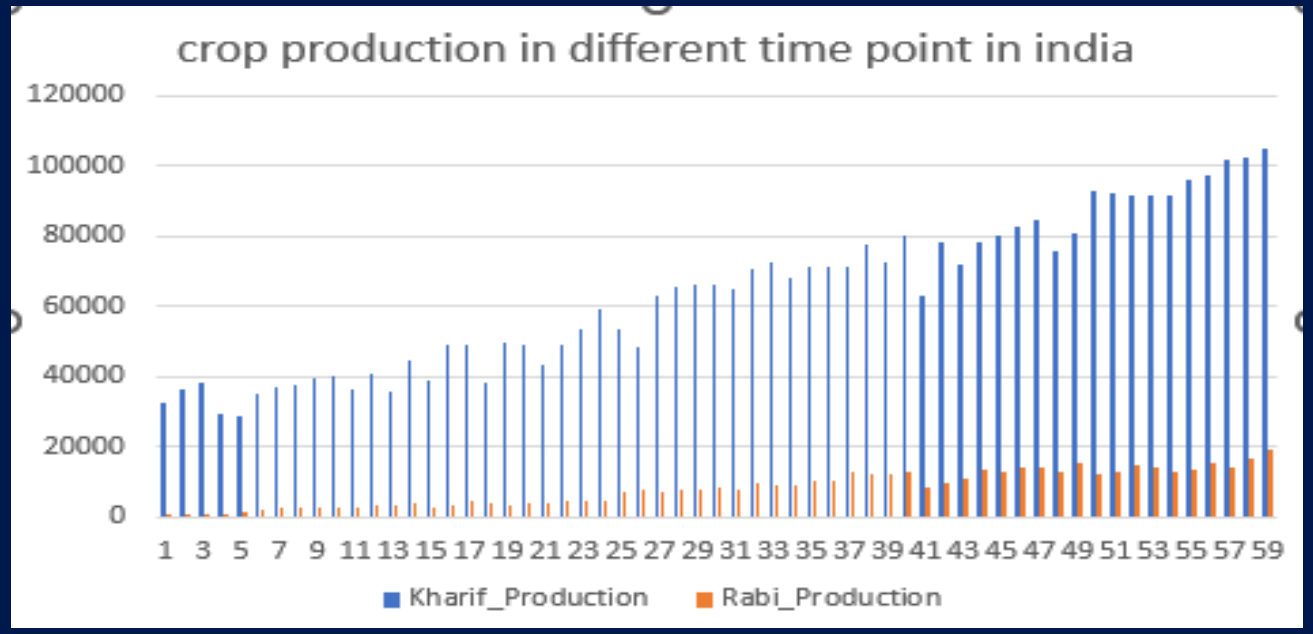


INTERPRETATION- - From the graphical representation of Crop production in INDIA we can see that the Kharif Crop production in INDIA is gradually increasing with respect to time. AS long as time is increasing production of kharif and Rabi crop is also increasing.

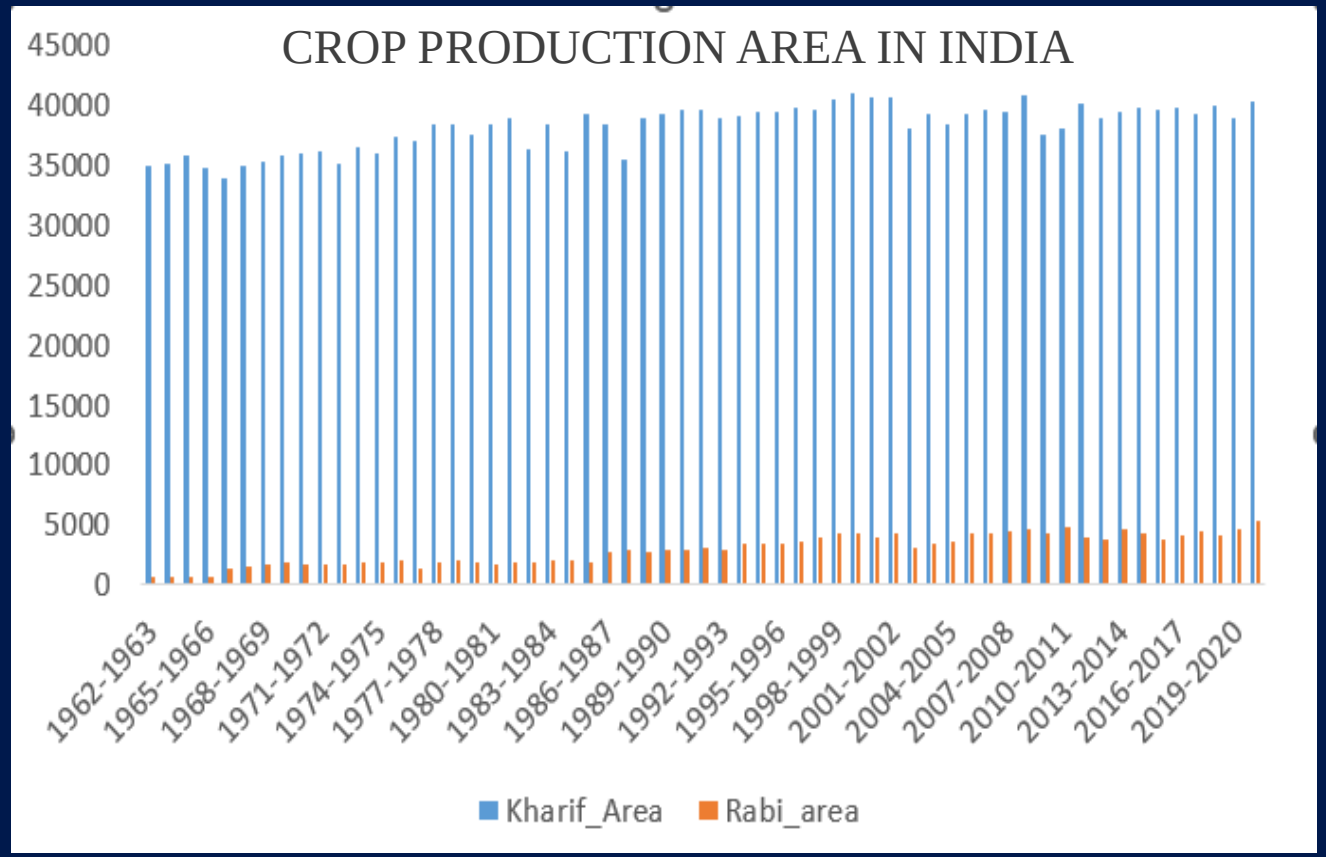


INTERPRETATION- From this diagram of Crop productivity in INDIA we can see that the Kharif Crop productivity in INDIA is gradually increasing with respect to time. AS long as time is increasing productivity of kharif and Rabi crop is also increasing.

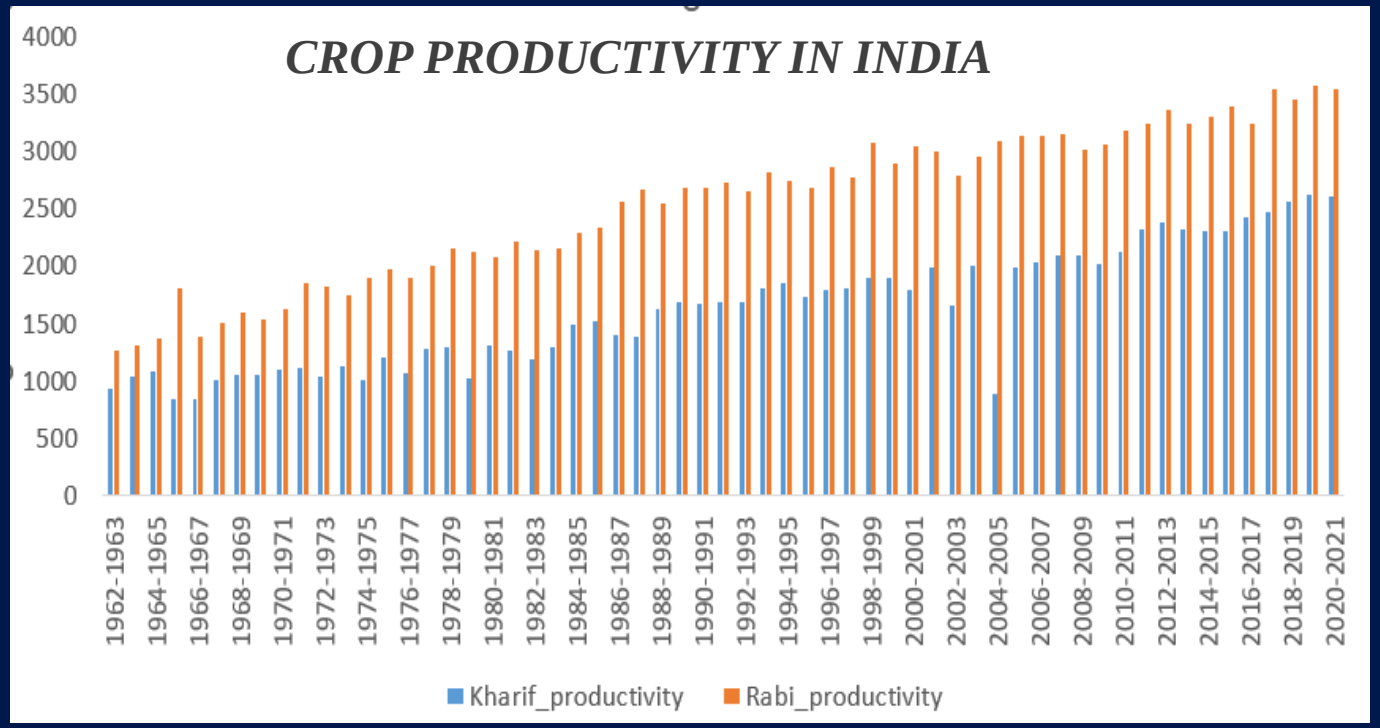
ii. Bar Diagram: Another mode of diagrammatic representation of data is the use of bar diagrams.



→INTERPRETATION- - Following ber graphs shows that the Crop production in INDIA is gradually increasing with respect to time. AS long as time is increasing production of kharif and Rabi crop is also increasing.

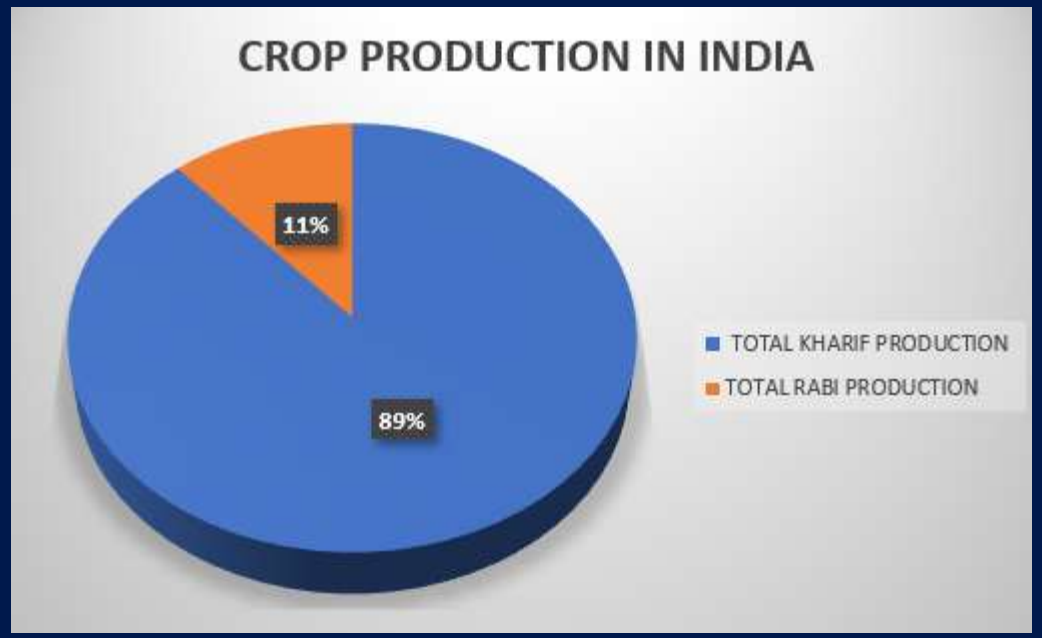


INTERPRETATION- From the graphical representation of Crop production area in INDIA we can see that the Kharif Crop production area in INDIA is constant with respect to time but Rabi crop production area is slightly increasing with respect to time.



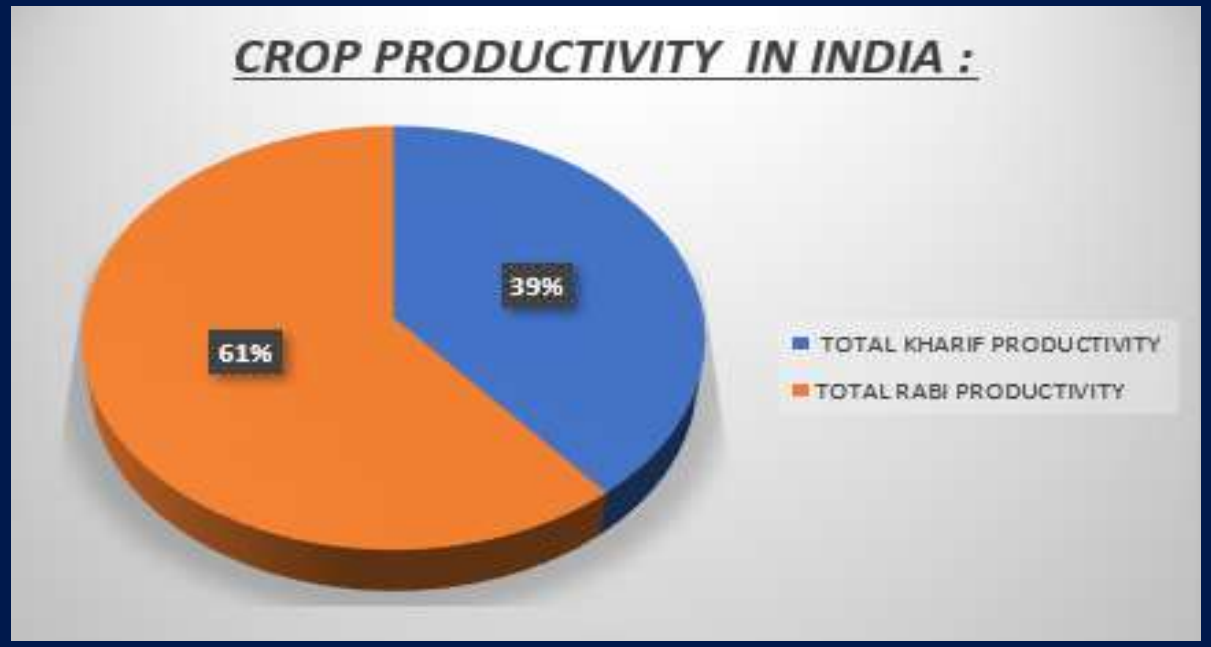
INTERPRETATION- - From the ber graph we have found that the Crop productivity in INDIA is gradually increasing with respect to time. AS long as time is increasing productivity of kharif and Rabi crop is also increasing.

iii. Pie diagram: A pie chart is an appropriate diagram used for exhibiting the relative sizes of the different parts of a whole.



INTERPRETATION- - The following pie graph shows that 11%is the Rabi and 89% is the Kharif production out of the total average crop production of 60 years.

INTERPRETATION - - Here we see that 61% is the Rabi and 39% is the Kharif productivity out of the total average crop productivity of 60 years.



Statistical Tools and Techniques

. ***Augmented Dickey Fuller test (ADF Test):***

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not. It is one of the most commonly used statistical test when it comes to analysing the stationary of a series. Here, the null hypothesis that a unit root is present in a time series sample i.e. The time series is non-stationary. The alternative hypothesis the time series is stationary.

The testing procedure for ADF test is the same as for the Dickey Fuller test but it is applied to the model-

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \cdots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t$$

Where α is a constant, β the coefficient on a time trend and p the lag order of the autoregressive process. Imposing the constraints $\alpha = 0$ and $\beta = 0$ corresponds to modelling a random walk and using the constraints $\beta = 0$ corresponds to modelling a random walk with a drift.

By including lags of the order p the ADF formulation allows for higher-order autoregressive processes.

Data analysis

Augmented Dickey-Fuller Test

data: diff of kharif production

Dickey-Fuller = -7.1721, Lag order = 3, p-value = 0.01

P -value is smaller than printed p-value so the data is stationary, we can analysis the data by time series component.

Augmented Dickey-Fuller Test

data: diff of rabi production

Dickey-Fuller = -4.1137, Lag order = 3, p-value = 0.01091

alternative hypothesis: stationary

P -value is smaller than printed p-value so the data is stationary, we can analysis the data by time series component.

Analysis Trend by Moving Average:

Moving Average method is a very commonly used method for the computation of Trend in a time series. This method consists in measurement of Trend by smoothing out the fluctuations present in the data by means of a Moving Average. Moving Average may be defined as an arithmetic mean of a given number of observations, each time in a time series. The period of it is adjusted successively by replacing the first observations of the previously averaged groups by the next observations below the relevant group in calculating such successive average.

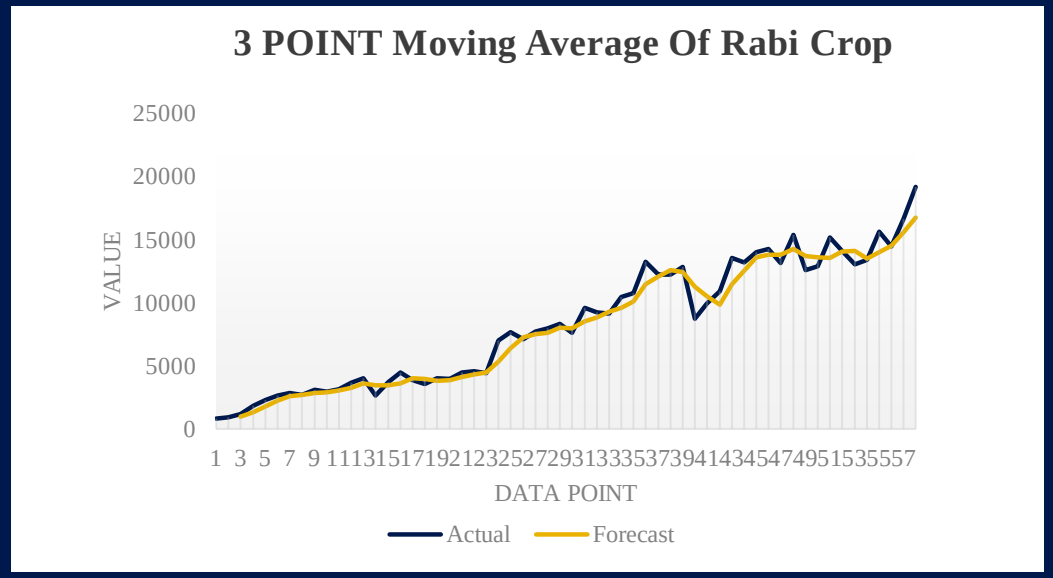
Moving average method is flexible in the sense that any change in the Trend is faithfully reflected by the Moving Averages. Here we are going to plot 3, 5 & 10 years moving average for the production of two types of crops i.e., Kharif & Rabi .

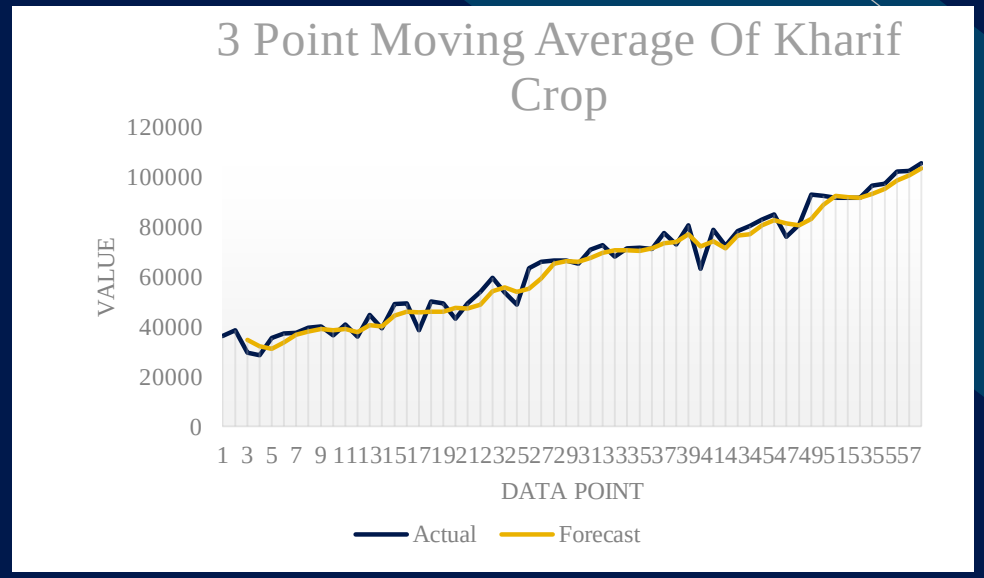
By Moving Average Method we can smooth out the random fluctuations in the trend of the production of three types of crops, individually:-

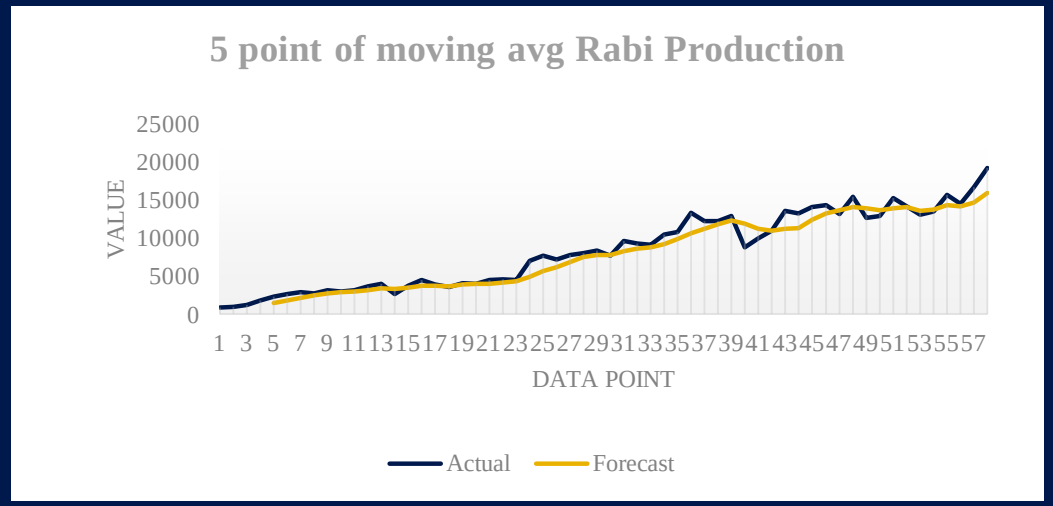
YEAR	Kharif_production	Rabi_production	3 point of moving avg Kharif_production	3 point of moving avg Rabi_production	5 point of moving avg Kharif_production	5 point of moving avg Rabi_production	10 point of moving avg Kharif_production	10 point of moving avg Rabi_production
1962-1963	32340	877						
1963-1964	36175	823						
1964-1965	38388	920	34664	968				
1965-1966	29429	1161	32146.33	1299				
1966-1967	28622	1816	31121.33	1758.667	33585.4	1403.8		
1967-1968	35313	2299	33687.33	2249.667	33775.8	1766		
1968-1969	37127	2634	36677	2590.667	33616.4	2149.8		
1969-1970	37591	2839	38092.33	2713	35642.4	2450.8		
1970-1971	39559	2666	39047.33	2860.333	37916.4	2702.8		
1971-1972	39992	3076	38625	2887.667	38118.6	2827.2	35852	2115.5
1972-1973	36324	2921	39073.33	3048	38874	2929.8	36324.9	2347.9
1973-1974	40904	3147	37718	3240.333	38541	3092.6	36078.7	2621.2
1974-1975	35926	3653	40525	3598.333	39578.2	3358.4	37610.3	2904.6
1975-1976	44745	3995	39979	3433	39433	3273.4	38674.7	2988.1
1976-1977	39266	2651	44319.33	3456.667	41957.6	3434	40038.1	3130.6
1977-1978	48947	3724	45850	3603.667	43644.2	3691.8	41259.1	3310.8
1978-1979	49337	4436	45590	4001.333	44156.2	3730	41348.6	3411.3
1979-1980	38486	3844	45970.67	3940.667	45225	3639.4	42401.6	3498.9
1980-1981	50089	3542	45940	3796.333	47220.8	3909.8	43326.9	3591.6
1981-1982	49245	4003	47499.33	3832.333	46064.2	3955.4	44010.9	3694.7
1982-1983	43164	3952	47248.67	4130.333	46064.2	3955.4	44854.2	3823.6
1983-1984	49337	4436	48761	4314.333	49123.4	4097.6	46639.8	3913.8
1984-1985	53782	4555	54170.33	4474.667	50984	4275.8	48104.5	3957.6
1985-1986	59392	4433	55578.33	5328	51847.2	4874.4	49534	4392.1
1986-1987	53561	6996	53905.33	6367	52967	5618.4	49515.6	4786.9
1987-1988	48763	7672	55233.33	7260	55774.8	6153.6	50919.5	5054.5

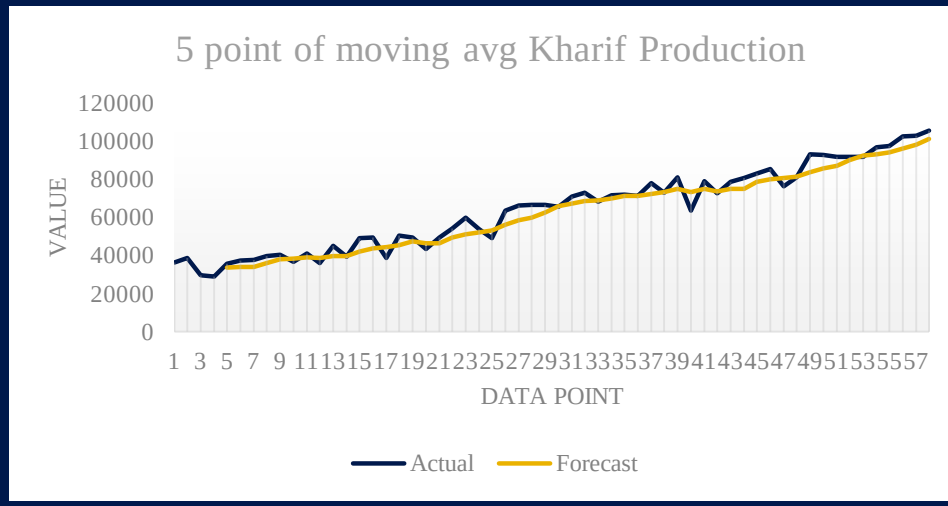
1988-1989	63376	7112	59339	7493	58194	6781.6	53658.7	5439.6
1989-1990	65878	7695	65190.33	7593.667	59579	7489.8	55281.5	5882.8
1990-1991	66317	7974	66187.67	7993	62140.4	7752.6	56993.8	6313.5
1991-1992	66368	8310	65976	7969.667	65436.4	7743.2	59201.7	6680.8
1992-1993	65243	7625	67444.67	8503.333	66905.8	8235.8	61340.3	7194.7
1993-1994	70723	9575	69523	8803.667	68250.8	8539	63222.4	7660.3
1994-1995	72603	9211	70401.67	9294.333	68563.2	8763.6	64071.1	8126.7
1995-1996	67879	9097	70601.67	9574	69554.2	9184.4	65847.3	8468.5
1996-1997	71323	10414	70257.67	10080	70819.8	9805.2	68128.1	8774.2
1997-1998	71571	10729	71328.67	11460.33	70893.6	10537.8	68899.7	9386.8
1998-1999	71092	13238	73381	12056.33	71869	11136	70059.9	9837.5
1999-2000	77480	12202	73783.33	12546	72848.8	11756.2	70706	10259.9
2000-2001	72778	12198	76926.67	12406	74688.6	12237	72121.4	10710.7
2001-2002	80522	12818	72128	11251	72991.2	11838.6	71905.5	10821.9
2002-2003	63084	8737	74075	10487.33	74496.6	11172.4	72695.1	10855.1
2003-2004	78619	9907	71311	9848.667	73446.6	10912.4	72657.8	11024.2
2004-2005	72230	10902	76373.67	11443.67	74545.4	11177.2	73697.1	11466.7
2005-2006	78272	13522	76891	12536.33	74475.2	11250.6	74581.9	11743.8
2006-2007	80171	13185	80382	13565.67	78399	12301.2	75695.1	12069.9
2007-2008	82703	13990	82608.33	13802	79665.4	13166	77081	12169.2
2008-2009	84951	14231	81204.33	13785	80411.2	13612.4	76928.9	12262.4
2009-2010	75959	13134	80505.67	14242.67	80878.2	13980.6	77711.8	12578.9
2010-2011	80607	15363	83101.33	13686.67	83391.6	13856.2	78933.4	12553.4
2011-2012	92738	12563	88571	13599.67	85324.6	13632.8	81861.8	12967
2012-2013	92368	12873	92201	13528.33	86633.8	13816.4	83149.6	13491.2
2013-2014	91497	15149	91752	14037.67	89720.2	14007.8	85065.7	13810.1
2014-2015	91391	14091	91433.67	14078.33	91881.4	13534.2	86379.8	13757.4
2015-2016	91413	12995	93035.67	13494	92594.4	13700.8	87993	13778.5
2016-2017	96303	13396	94950.33	14004.33	93547.8	14250.6	89436.2	13941.7
2017-2018	97135	15622	98492.67	14485.33	95656.4	14108.4	91145.1	13962.4
2018-2019	102040	14438	100484	15551.33	97833.6	14609	93776.9	14308.4
2019-2020	102277	16594	103175	16730.67	100592.6	15842	96237	14688.1
2020-2021	105208	19160						

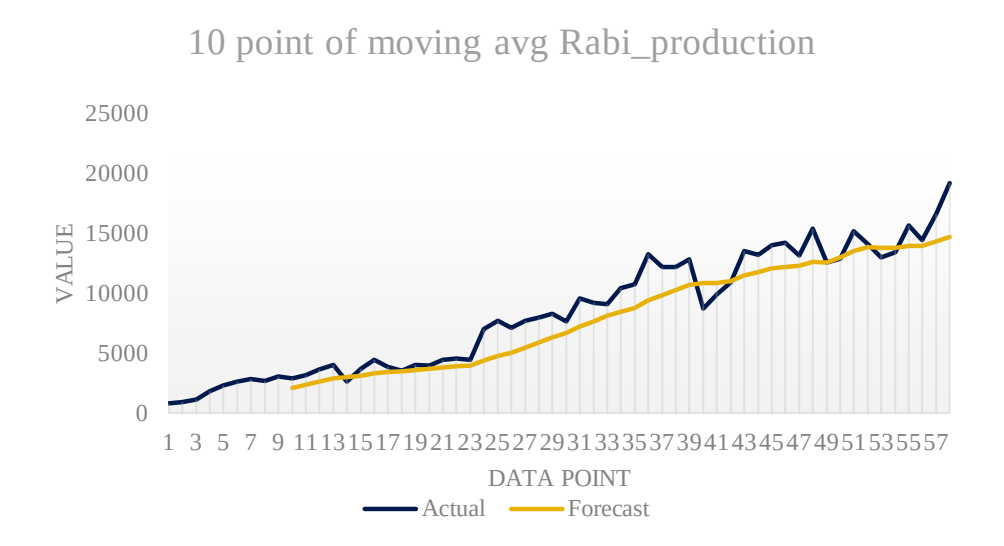
Here are the graphs of 3 types moving average -

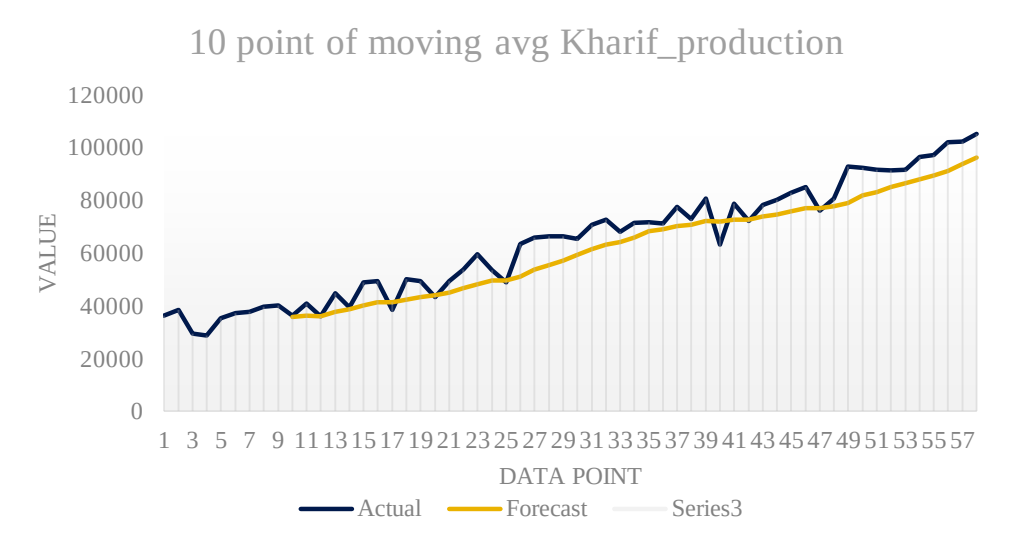












INITIAL INTERPRETATION:- The moving averages of 3 years, 5 years and 10 years have smoothen our data which makes the trend clearer,10 year moving average makes the data smoothest .The figures shows that our time series data has overall an upward trend.

AUTOCORRELATON AND PERTIAL AUTOCORRELATION FUNCTION:

AUTOCORRELATION FUNCTION(ACF)

Autocorrelation is the correlation between two observations at different points in a time series. For example, values that are separated by an interval might have a strong positive or negative correlation. When these correlations are present, they indicate that past values influence the current value.

Autocorrelation function is defined by-

$$\rho_k = \mathrm{corr}\big(\rho(t), \rho(t+k)\big) = \frac{\gamma_k}{\gamma_0}$$

PACF stands for Partial Autocorrelation Function, which is a statistical tool used in time series analysis to identify the relationship between variables. Autocorrelation refers to the correlation between a time series and its own past values, while partial autocorrelation represents the correlation between a time series and its lagged values after removing the effects of the intermediate lags.

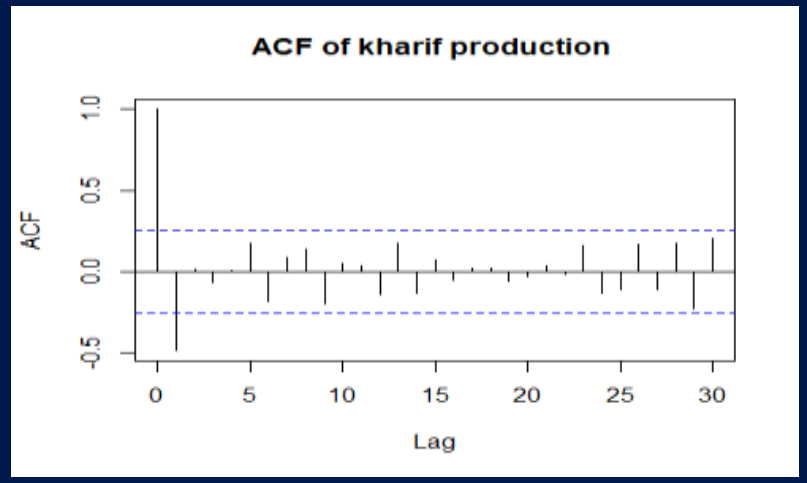
The ACF can be graphed as a function of lag and is often used to identify the order of an autoregressive (AR) model.

. The PACF can also be graphed as a function of lag and is often used to identify the order of a moving average (MA) model.

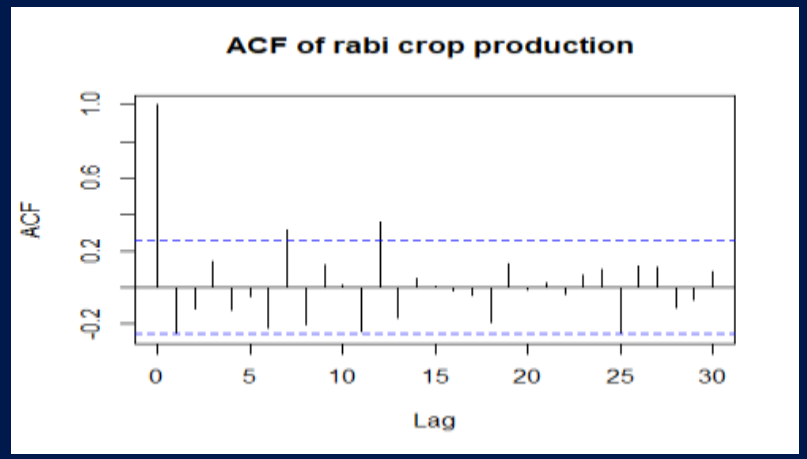
If both the ACF and PACF decay slowly, it may indicate the need for a mixed ARMA model.

Data analysis:

According our dataset we get the ACF as follows-
For kharif and Rabi crop –

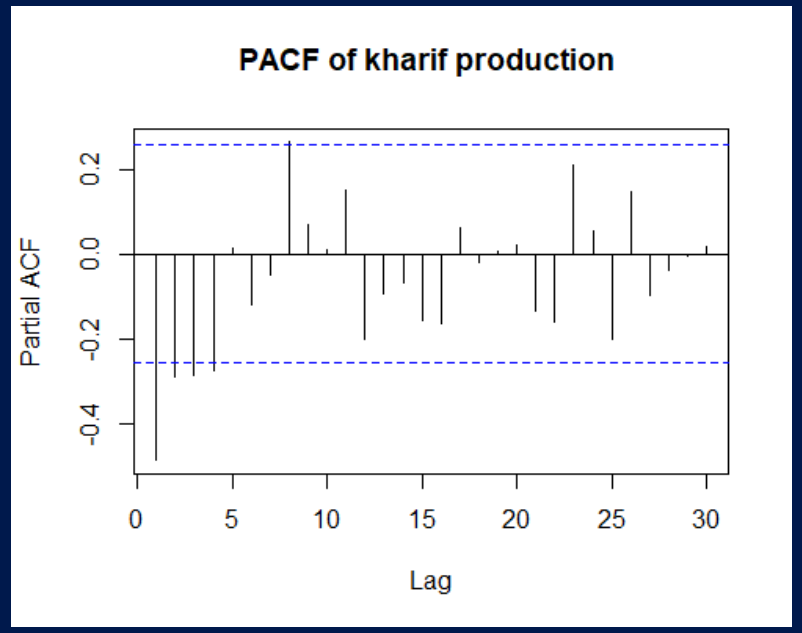


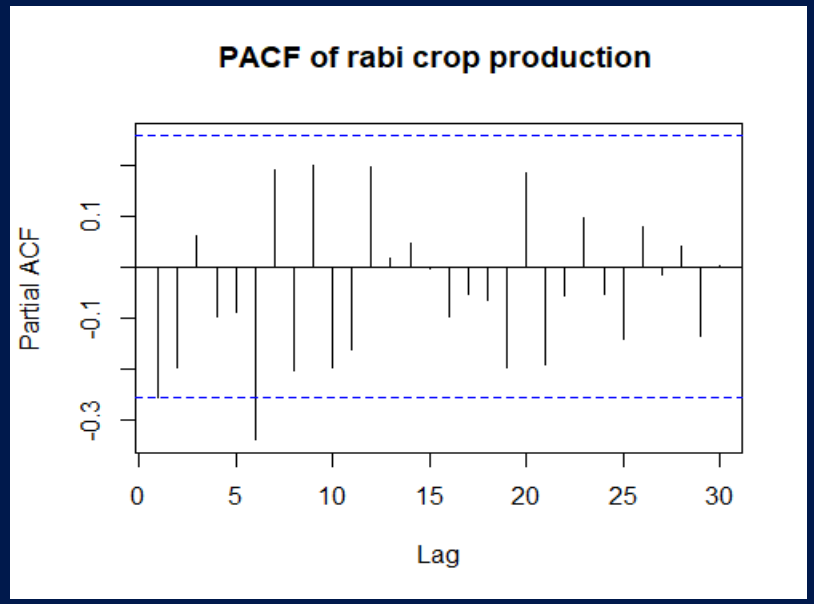
- From visual inspection, the ACF is towards negative axis. Here lag1 is significant with θ being the negative values. It is MA(1).



- From the diagram, here is Two significant spike lag6 and lag12 , so the data can be modelled as MA(2).

According our dataset we get the PACF as follows





Model fitting on the data: -

ARIMA (Autoregressive Integrated Moving Average) is a widely used statistical model for time series analysis and forecasting. It combines three key elements: autoregression (AR), differencing (I), and moving average (MA) to capture the underlying patterns and dependencies in the data.

ARIMA model combines the above three elements and is denoted as ARIMA (p, d, q). It can be used for forecasting future values of a time series, assuming that the underlying patterns and dependencies captured by the model remain unchanged.

Evaluating the model by checking the residuals for randomness and examining performance metrics such as **the mean squared error (MSE).** After that we calculate the **RMSE (Root Mean Squared Error)** of the predicted values with respect to the observed values in the testing data. From RMSE we can understand how good the fit is.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_{obs} - y_{pred})^2}{n}}$$

Data analysis: -

Here are the results of the ARIMA model

Series: y-kharif production

Summary of the model is given below -

arima(x = kharif production, order = c(0, 1, 1))

Coefficients:

	ma1
	-0.4376
s.e.	0.0958

Series: y-rabi production

Summary of the model is given below -

arima(x = y rabi production , order = c(1, 1, 2))

Coefficients:

	ar1	ma1	ma2
	-0.6821	0.5606	-0.2939
s.e.	0.1558	0.1681	0.1232

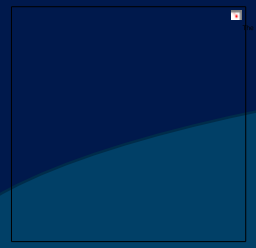
Training set error measures: Kharif Production ARIMA Model is mentioned below:-

Training set

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
2125.954	5474.994	4540.134	2.740958	8.056011	0.9831093	-0.251324

Training set error measures: Rabi Production ARIMA Model is mentioned below:-

ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
387.1899	1190.704	823.123	5.207842	11.35013	0.9058867	-0.1489476



Interpretation: -

From inspection we can see that prediction of the testing data along with the observed values. As we can see that the RMSE is in both cases so we can say that the prediction of our testing data is not good enough.

Holt Winters exponential smoothing:

we have applied the fitted **Holt-Winters model** to the testing data and predict the present values of the testing data. After that we calculate **the RMSE (Root Mean Squared Error)** of the predicted values with respect to the observed values in the testing data. From **RMSE** we can understand how good the fit is.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_{obs} - y_{pred})^2}{n}}$$

Data analysis: -

- The summary of Holt Winters exponential smoothing for kharif crop production –

Holt Winters(x = kharif crop production , gamma = FALSE)
Smoothing parameters:
alpha: 0.3102772
beta: 0.3461922
gamma: FALSE
Coefficients:
[,1]
a 104514.945
b 2140.187

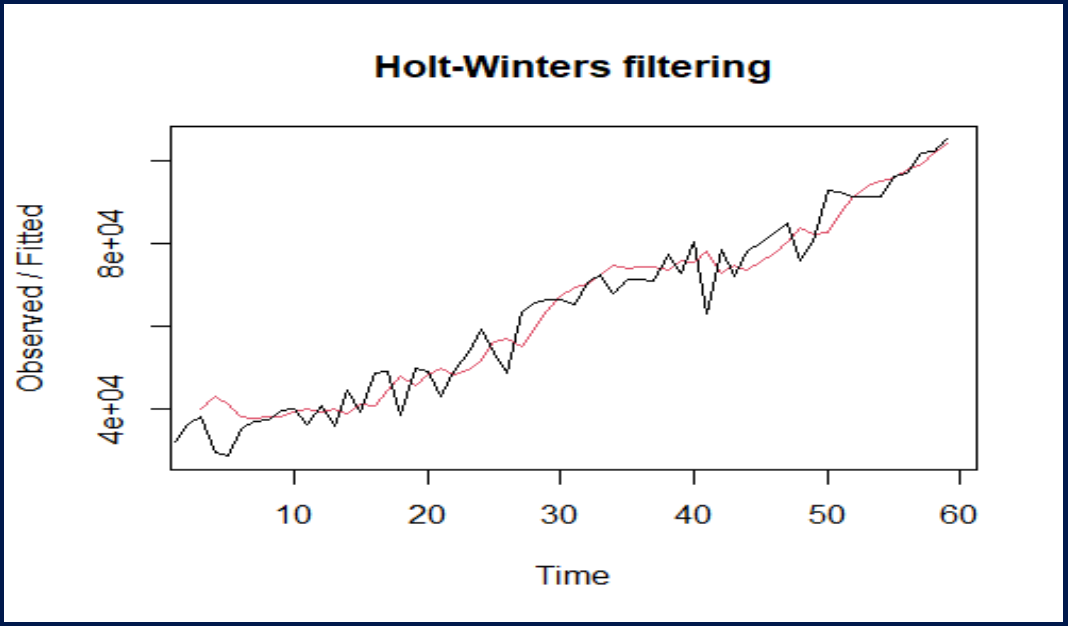
- The summary of Holt Winters exponential smoothing for Rabi crop production –

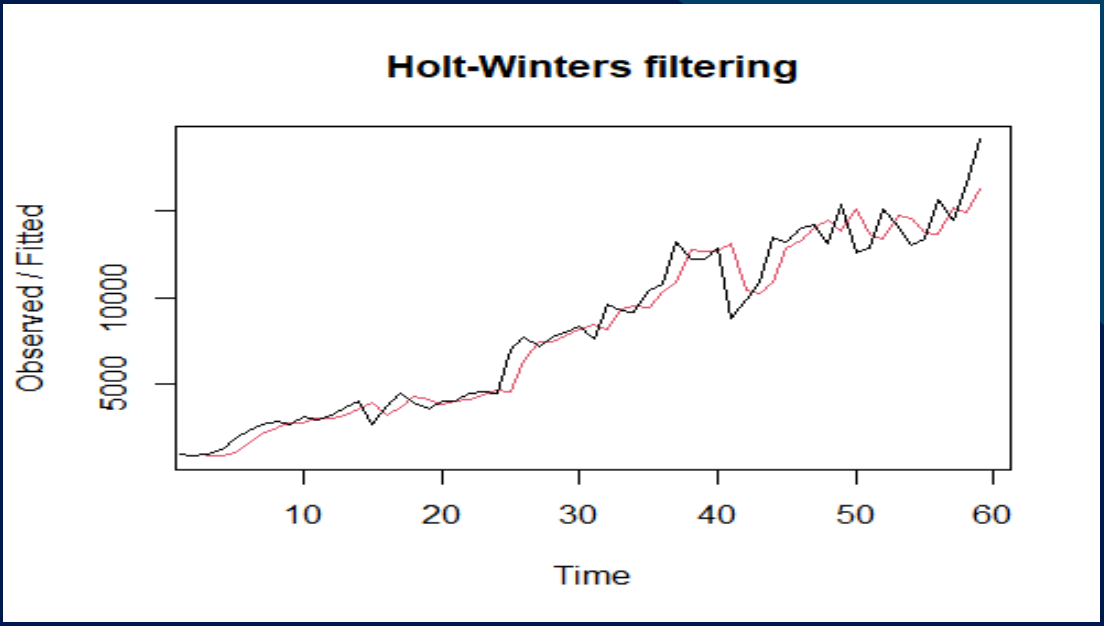
HoltWinters(x =Rabi crop production , gamma = FALSE)

Smoothing parameters:
alpha: 0.6489699
beta: 0.04915577
gamma: FALSE

Coefficients:
[,1]
a 18143.5277
b 355.4084

The graph is given bellow, the red line is smoothing line -





Interpretation: -

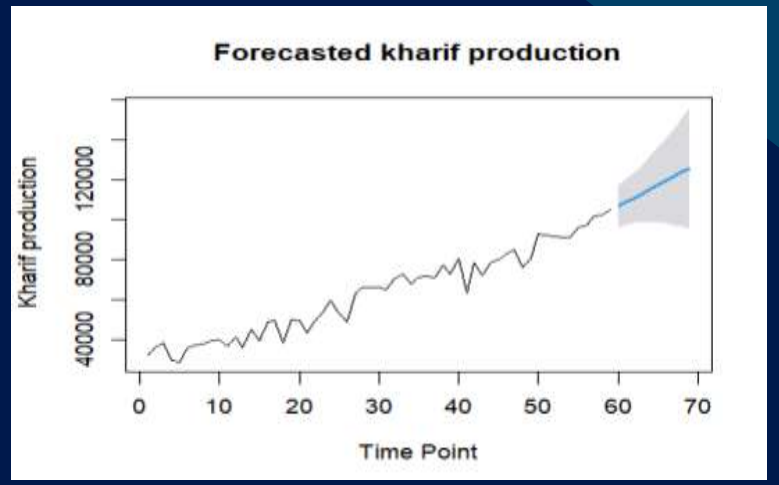
From the visual inspection we can see that prediction of the testing data along with the observed values, we can say that the prediction of our testing data is good enough. Now we will perform further process accordingly to predict the data.

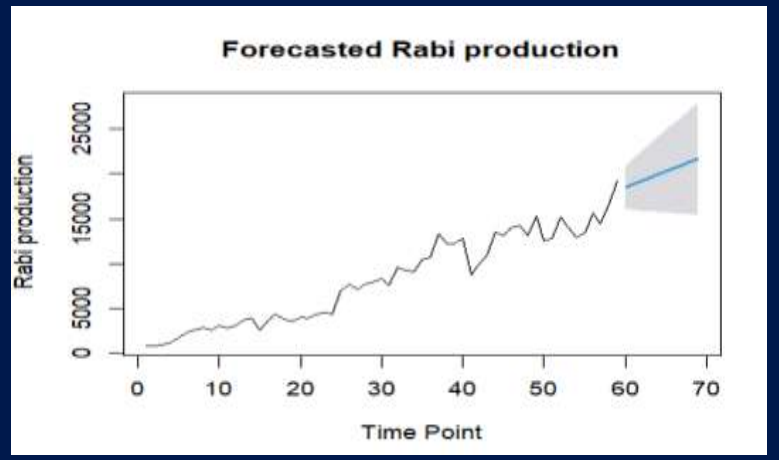
Forecasting :-

Here is the RMSE of seasonal ARIMA is greater than the RMSE of Holt-winters exponential smoothing. So, we will predict the future crop production using Holt-winters exponential smoothing.

The forecasted production values and graphs from 2021 to 2030 is as follows:

Year	forecasted kharif crop production	forecasted rabi production
2021	106655.1	18498.94
2022	108795.3	18854.34
2023	110935.5	19209.75
2024	113075.7	19565.16
2025	115215.9	19920.57
2026	117356.1	20275.98
2027	119496.3	20631.39
2028	121636.4	20986.8
2029	123776.6	21342.2
2030	125916.8	21697.61

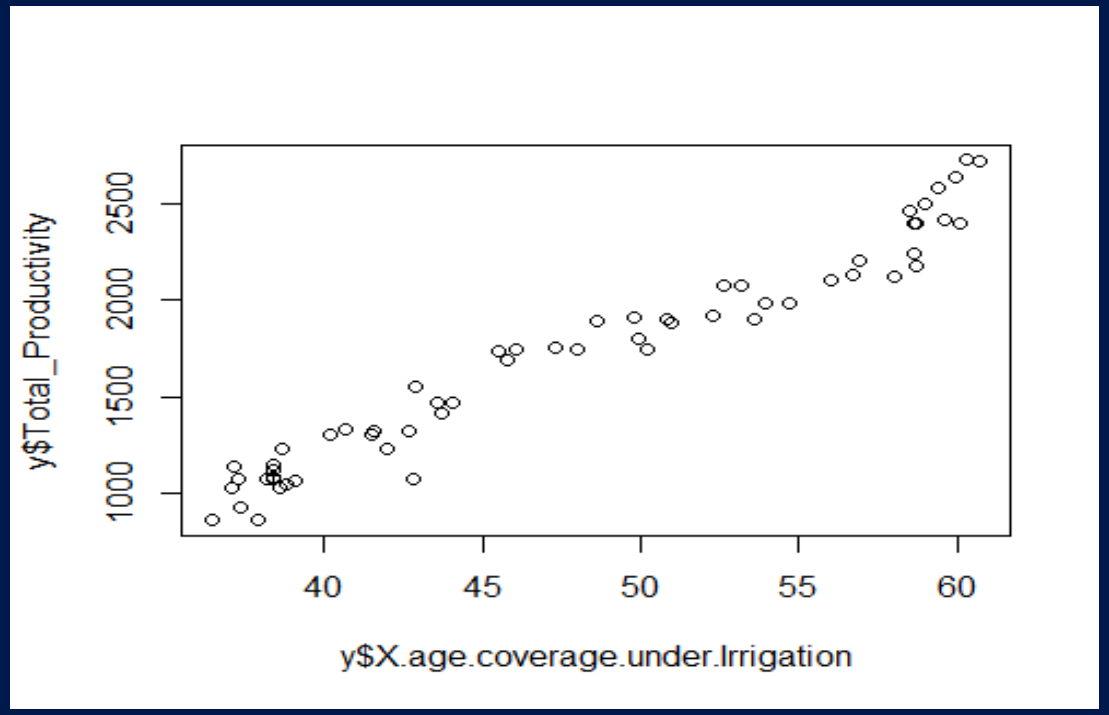




Simple Regression on total productivity and percentage coverage under irrigation:

Data Visualization

Regression analysis is used to analyze and determine the relationship between response variable and explanatory variable. The variable considered for analysis in this study is percentage cover under irrigation. total productivity is a dependent variable which depends on these factors.
The main data visualization tool used here is Scatterplot,



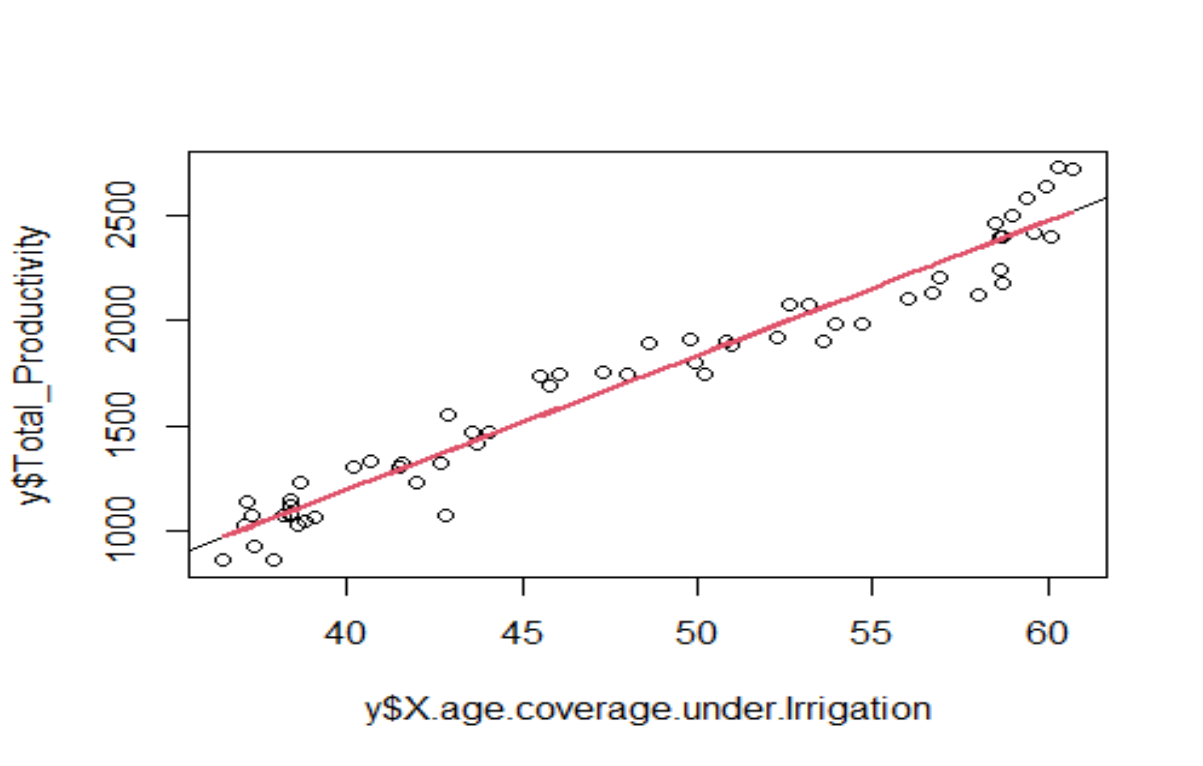
- An increasing Trend found. Linear Trend may be fitted.
- correlation between two variables is 0.976 (i.e highly correlated).

simple Linear regression fitting

we have fitted the model, got a summary of the model given here

Dependent Variable	Independent Variable	Regression Equation	t Value	R Squared	Decision
Total productivity	Area coverage under irrigation	y= -1351.80423 +63.73125x	-14.93 34.24	Multiple R-squared: 0.9536, Adjusted R-squared: 0.9528	X has an effect in the total productivity as the R square value is = 0.9528(quite high) so the regression line fit properly.

Here is the model fitted graph with abline –



Conclusion:-

Crop Production in India is an essential aspect that plays a significant role in the country's economy and food security. Though the production of Kharif and Rabi crops has been increasing steadily over the years, the productivity of Indian agriculture is relatively low compared to the other countries. However, there are several challenges that Indian farmers face and it is essential to address these issues to improve the productivity and ensure sustainable agriculture. With the governments support and increased adoption of technology, Indian agriculture has the potential to become more productive and efficient, Which helps in improving the conditions of the Indian farmers and ensures the food security of the country .

References :

Books:

Notes on Time Series Analysis provided by
Dr. Saran Ishika Maiti (Asst. Prof. department
of Statistics , Visva-Bharati University)

Fundamental of Statistics(vol-1) by Gun Gupta Dasgupta

Fundamental of Statistics(vol-2) by Gun Gupta Dasgupta

Website:

Geeks for Geeks

Stack overflow

Stack Exchange

Slide model

Wikipedia

Unesco

Image Source: Google, Bing, Internet Resource

Software: R Console, R Studio, Microsoft Word, Excel, Power Point, Adobe Express, Google Slide

Thank you