Factors affecting Crop Production in Central and Southern Asia

Abstract

This project addresses the critical issue of food security in Central and Southern Asia, a region of 12 countries heavily reliant on wheat, rice, and potatoes as staple crops. Fluctuating yields due to complex interactions of climate, operational, and economic factors threaten food stability. My research aimed to identify key factors influencing yield variations and develop a robust model to predict yields for these vital crops. Employing four comprehensive datasets and advanced machine learning models, particularly Random Forest, I achieved exceptional accuracy in yield forecasting. This empowers policymakers to proactively manage resources and ensure food availability, while enabling farmers to optimize practices and maximize productivity. My project aligns with the UN's Zero Hunger goal, paving the way for a food-secure future in the region.

Introduction

Central and Southern Asia, encompassing 12 countries (India, Pakistan, Sri Lanka, Maldives, Bangladesh, Tajikistan, Turkmenistan, Uzbekistan, Kyrgyz Republic, Kazakhstan, Nepal), faces a precarious food security situation. Wheat, rice, and potatoes are the region's lifeblood, but their yields are subject to unpredictable swings due to a complex interplay of climate variability, operational limitations, and economic factors. This precariousness poses a constant threat to food stability for millions.

My research addressed this challenge by pursuing two key objectives:

1. **Identifying the key factors:** I delved deep into the intricate dance between climate (precipitation, temperature, etc.), operational practices (irrigation, fertilization, etc.), and economic forces (market prices, subsidies, etc.) to pinpoint their precise influence on wheat, rice, and potato yields across the region.

2. **Building a robust prediction model:** I explored various machine learning models using R, ultimately selecting Random Forest due to its superior accuracy and interpretability. This model empowers me to accurately forecast yields for these crucial crops, enabling proactive interventions and informed decision-making.

My research was motivated by a profound understanding of the critical role of yield prediction in ensuring food security. By empowering policymakers and farmers with this knowledge, I hope to achieve the following:

- Guide informed policymaking: Precise yield forecasts enable policymakers to plan strategically, allocate resources effectively, and implement targeted interventions to safeguard food availability for millions.

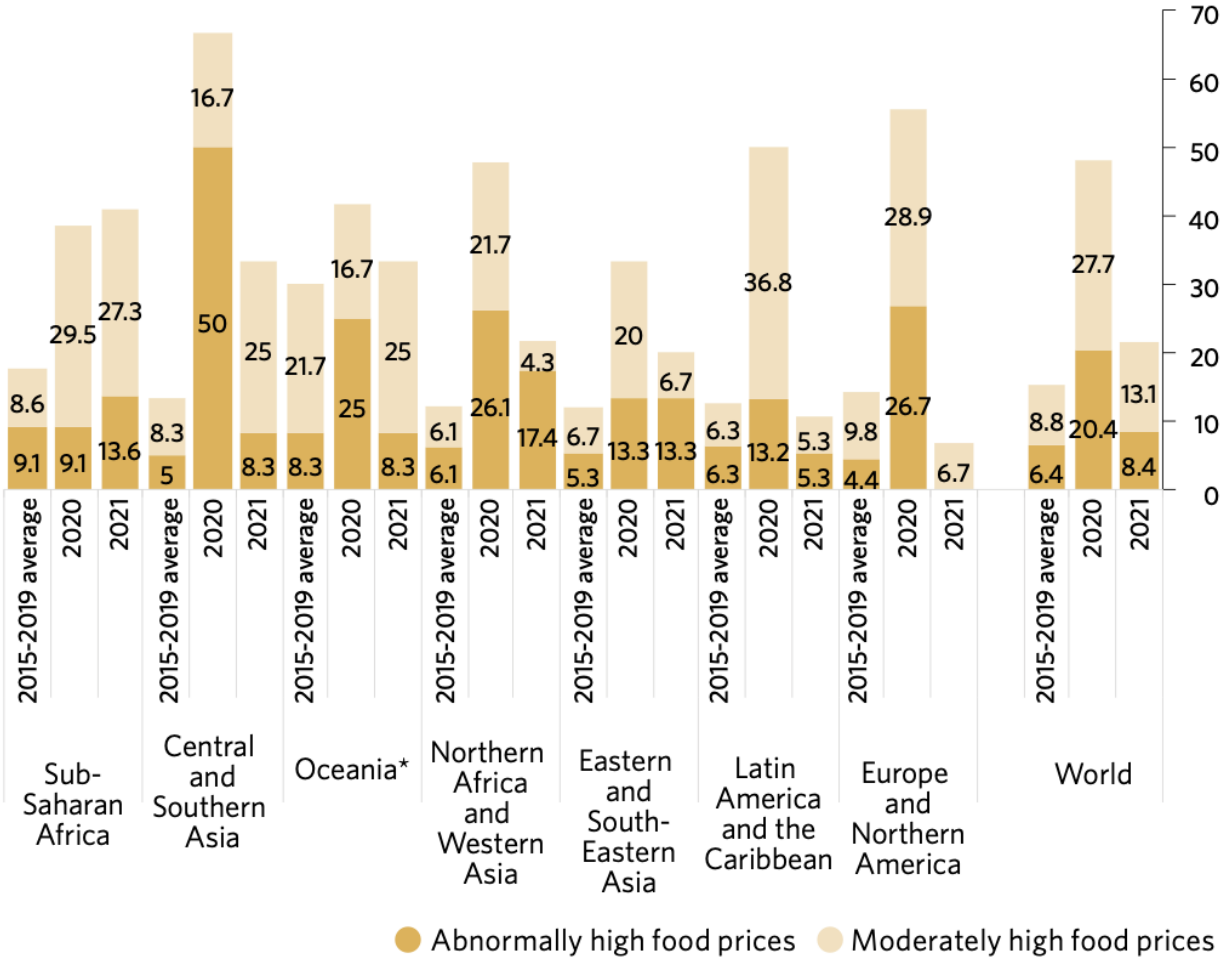
- Support farmers' livelihoods: Farmers equipped with yield insights can optimize field operations, manage resources better, and ultimately enhance their productivity, leading to improved incomes and well-being.

- Align with SDG 2: My project contributes directly to the UN's Zero Hunger goal by promoting stable food production and strengthening food security throughout the region.

By unraveling the secrets of crop yields in Central and Southern Asia, I pave the way for a future where food security is not a distant aspiration but a tangible reality for all.

Additionally, the previous researches have been done in more developed region, And not on these specific crops, especially talking about Wheat, Rice and Potato which are staples for central and southern Asia.

According to UN sustainable goals report, this was the area affected with higher prices of food, which might lead to food insecurity in a longer run.



Also, all the other studies have only focused on economic factors, mine focuses on both economic and environmental factors.

Data Description and Collection

For my data, I collected four datasets as described in the following:

1) World Development Index(WDI) from World Bank

2) Crop and Livestock Products from Food and Agriculture Organization (FAO)
3) Observed Temperature data from Climate Change Knowledge Portal (CCKP) by World Bank Group
4) Observed Precipitation data from Climate Change Knowledge Portal (CCKP) by World Bank Group

All the data was collected from the years 1960 to 2021 for 12 countries in the Central and Southern Asia as mentioned in the Introduction. The WDI data contains the all the economic factors, having:

- - Agricultural Land (% of land area): The share of land that is arable, under permanent crops and under permanent pastures.
- - Agricultural irrigated land (% of total agricultural land): Agricultural land purposefully provided with water
- - Agricultural Machinery, tractors: Number of wheel and crawler tractors used specifically for agricultural purposes.
- - Agricultural raw material exports: Percentage share of exported crude fertilizers and minerals out of total exports
- - Agricultural raw material imports: Percentage share of imported crude fertilizers and minerals out of total imports
- - Annual agricultural freshwater withdrawal: Percentage of freshwater withdrawals for irrigation out of total freshwater withdrawals
- - Arable land: Percentage of land area that is defined by FOA under temporary crops including kitchen gardens
- - Inflation, consumer prices: Annual percentage inflation as measured by the consumer price index
- - Total CO2 emissions: thousand metric tons of CO2 emissions stemming from burning of fossil fuels and manufacturing of cement.
- - Droughts, flood, extreme temperatures: Annual average percentage of population that is affected by natural disasters.
- - Employment in agriculture: Percentage of total employment involved in agriculture
- - Fertilizer consumption: Kilograms of plant nutrients used per hectare of arable land
- - GDP growth: Annual percentage growth rate of GDP at market prices
- - Health Capital Index: Score from 0 to 1 that indicates productivity of a worker relative

to the benchmark of full health and complete education

- - Population: Midyear estimates of total population
- - Total greenhouse gas emissions: % change of greenhouse gas emissions to the chosen base year, 1990.

For the Crop Yield Data from FAO, I picked up three staple crops as stated, Rice, Wheat, and Potato.

Then from CCKP portal, picked up Temperature (in Degree Celsius) and Precipitation(in mm).

When I performed the data collection, the data was not in a proper format, so I had to modify the format, using both excel and Rstudio.

Finally, I merged the data using the common columns as “Country Name”, “Country Code”, and “Year”.

Data Preprocessing

It was the longest step in the whole project. The data had several missing values, especially the WDI data. So, I decided I would do the following steps to clean up the data:

1) Remove attributes having missing values above a certain threshold:

   I decided to remove the attributes which had 30% or more missing values for more than 8 countries. (See Pivot Table 1, highlighted attributes were removed as per the stated threshold.)

2) Interpolation

   For the remaining values, I used interpolation to predict the values that might be not captured in the real world.

Attributte	Missing Values more than 30% for 8+ countries
Agricultural.irrigated.land.of.total.agricultural.land.	Yes
Agricultural.land.of.land.area.	No
Agricultural.machinery, tractors	No
Agricultural.raw.materials.exports.of.merchandise.exports.	Yes
Agricultural.raw.materials.imports.of.merchandise.imports.	Yes
Annual.freshwater.withdrawals.	Yes
Child.employment.in.agriculture(Ages 7-14)	Yes
Employment.in.agriculture.of.total.employment	Yes
Employment.in.agriculture..female....of.female.employment...modeled.ILO.estimate.	yes
Employment.in.agriculture.male..of.male.employment...modeled.ILO.estimate.	yes
Arable.land.of.land.area.	No
Fertilizer.consumption.kilograms.per.hectare.of.arable.land.	No
CO2.emissions..kt.	No
Total.greenhouse.gas.emissions....change.from.1990.	Yes
Inflation.consumer.prices.annual.	No
GDP.growth.annual.	No
Average.precipitation.in.depth..mm.per.year.	Yes
Droughts.floods..extreme.temperatures..of.population.average.1990.2009.	Yes
Human.capital.index.HCI.(scale 0-1)	Yes

So, I modified and cleaned the missing values for WDI, picked up only the features whose values were available and interpolated the remaining values in a separate R file, and saved it as “Cleaned WDI Data.csv”

There were no missing values for Climate data i.e, Temperature (in deg C) and Precipiation (in mm). Their data was also collected in two separate files, so I merged them in a file called “climatePcpt.csv”.
Kept only required attribute for all the files

The similar interpolation was done for Rice, Wheat and Potato yield.

Then I merged these two datasets as follows:

```
setwd('/Users/ananya/Desktop/DA_FALL23_ANANYA_UPADHYAY')
wdi.data<-read.csv('Cleaned WDI Data.csv')
climate.pcpt.data<-read.csv('climatePcpt.csv')
data1<-
left_join(wdi.data,climate.pcpt.data,by=c('Country_Code','Country_Name','Year'))
```

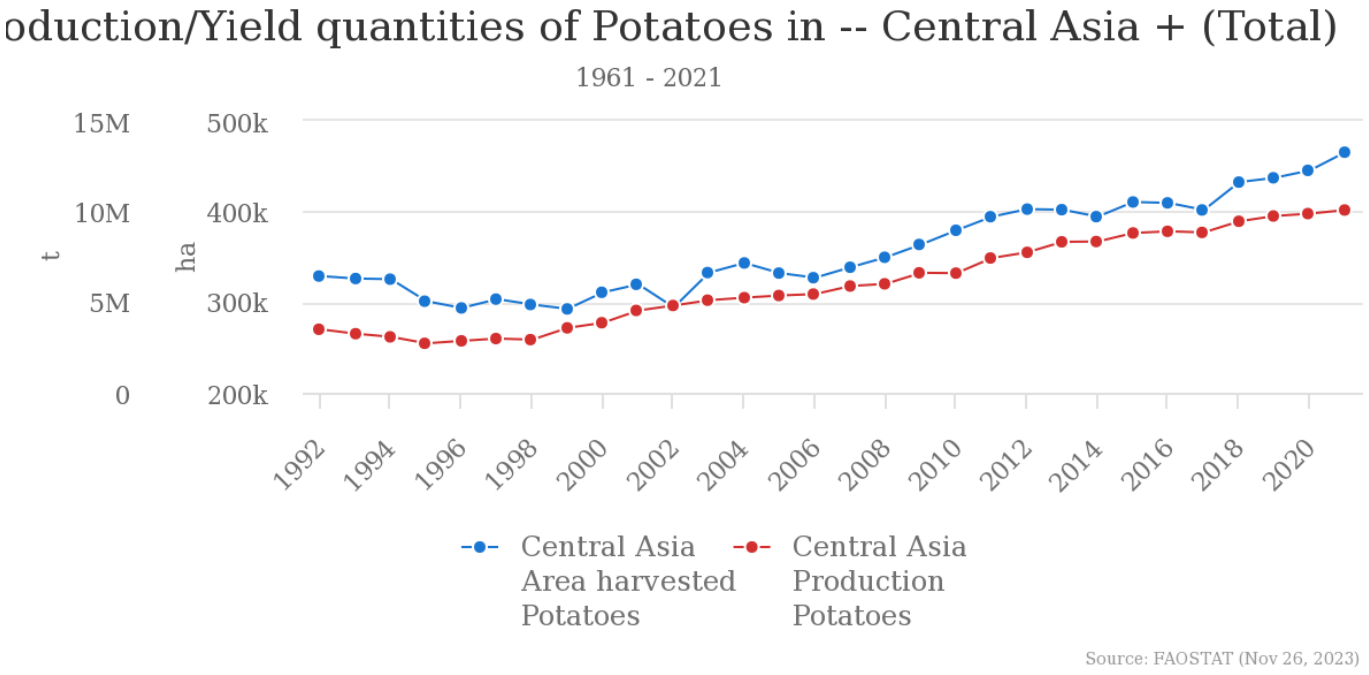
Now, I had three separate file for staple crops yield, potato.csv, rice.csv and wheat.csv, merged them as follows:

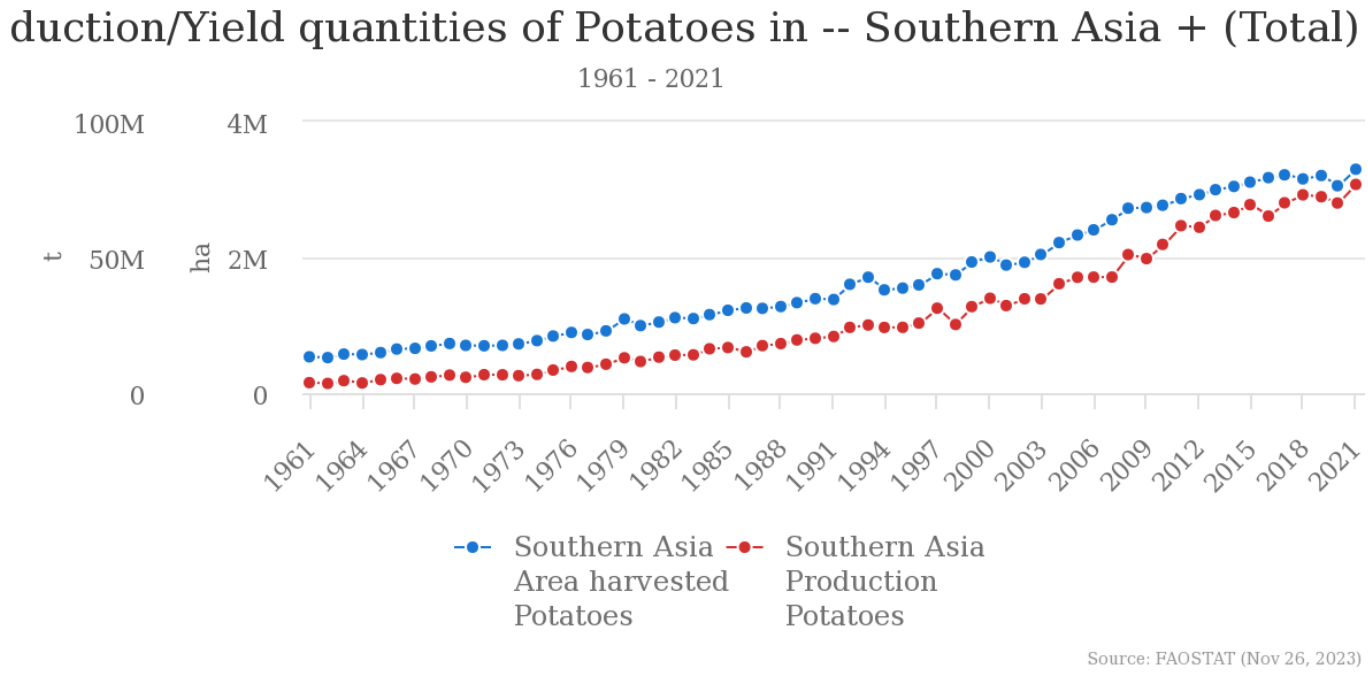
```
potato<-read.csv('yield_potato.csv')
rice<-read.csv('yield_rice.csv')
wheat<-read.csv('yield_wheat.csv')

data2<-left_join(potato,rice,by=c('Country_Name','Year'))
data2<-left_join(data2,wheat,by=c('Country_Name','Year'))
```

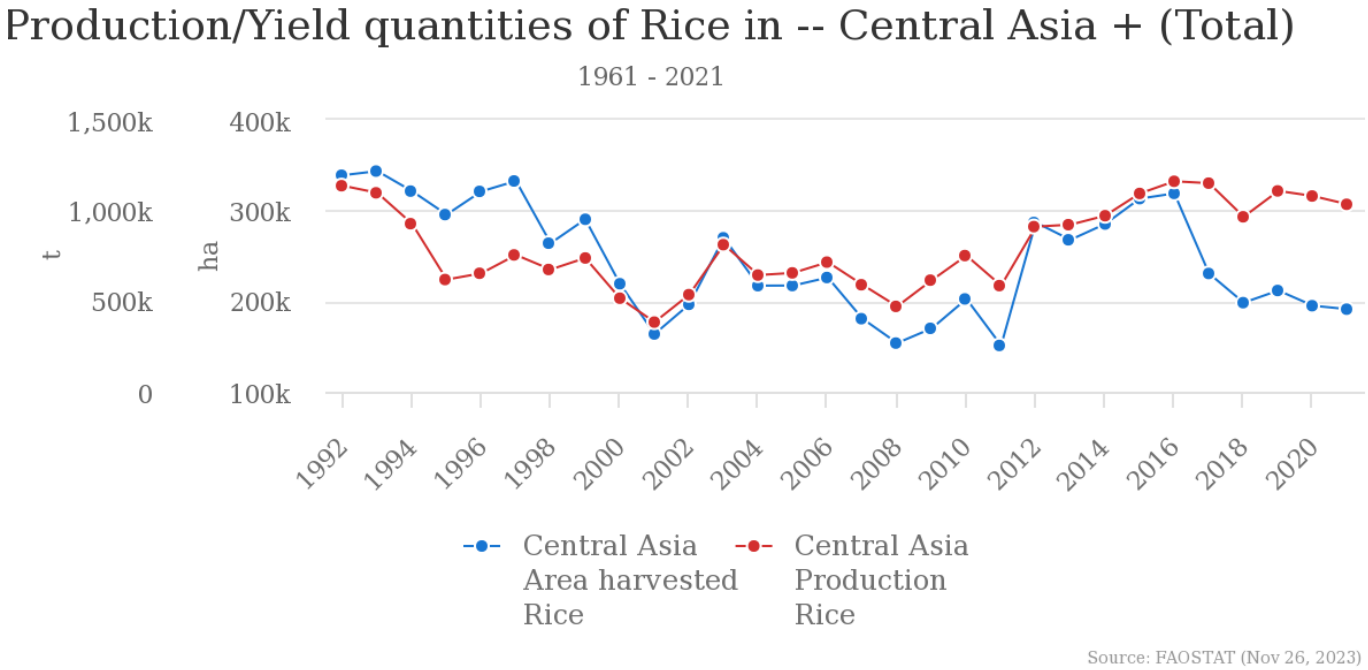
Finally, merged them all together as follows.

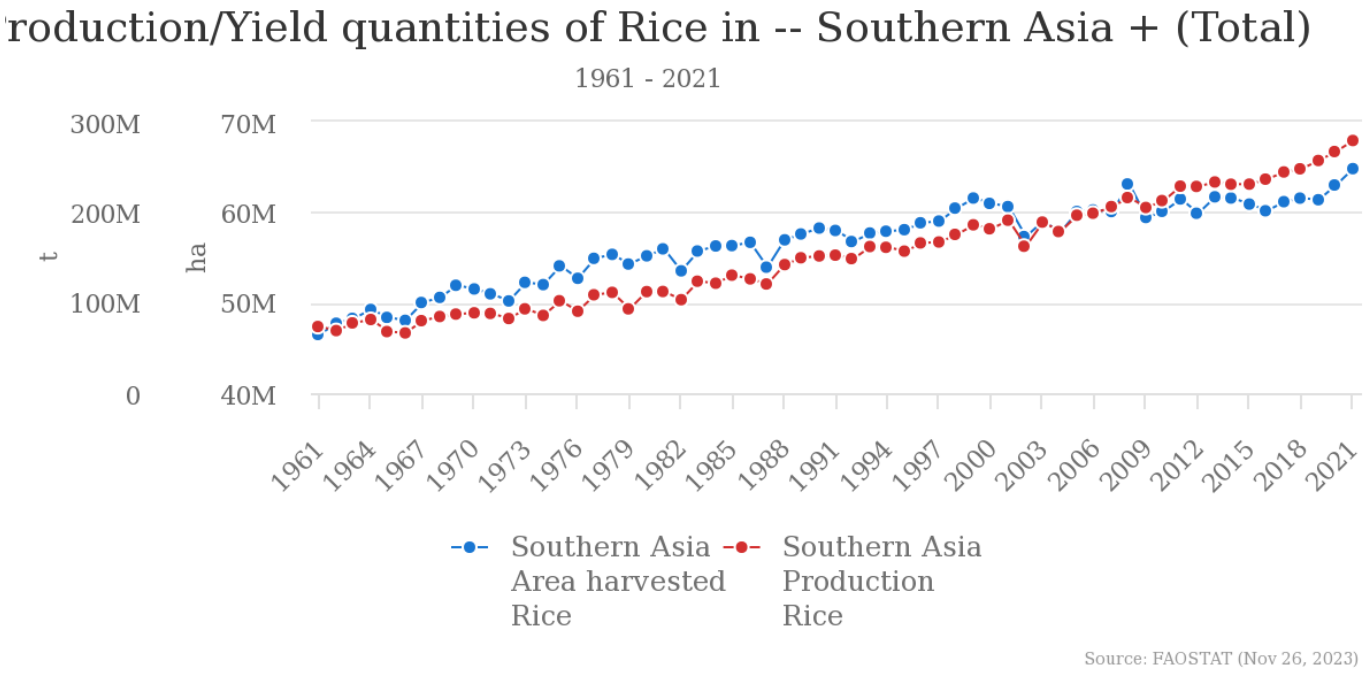
```
data<-left_join(data1,data2,by=c('Country_Name','Year'))
```

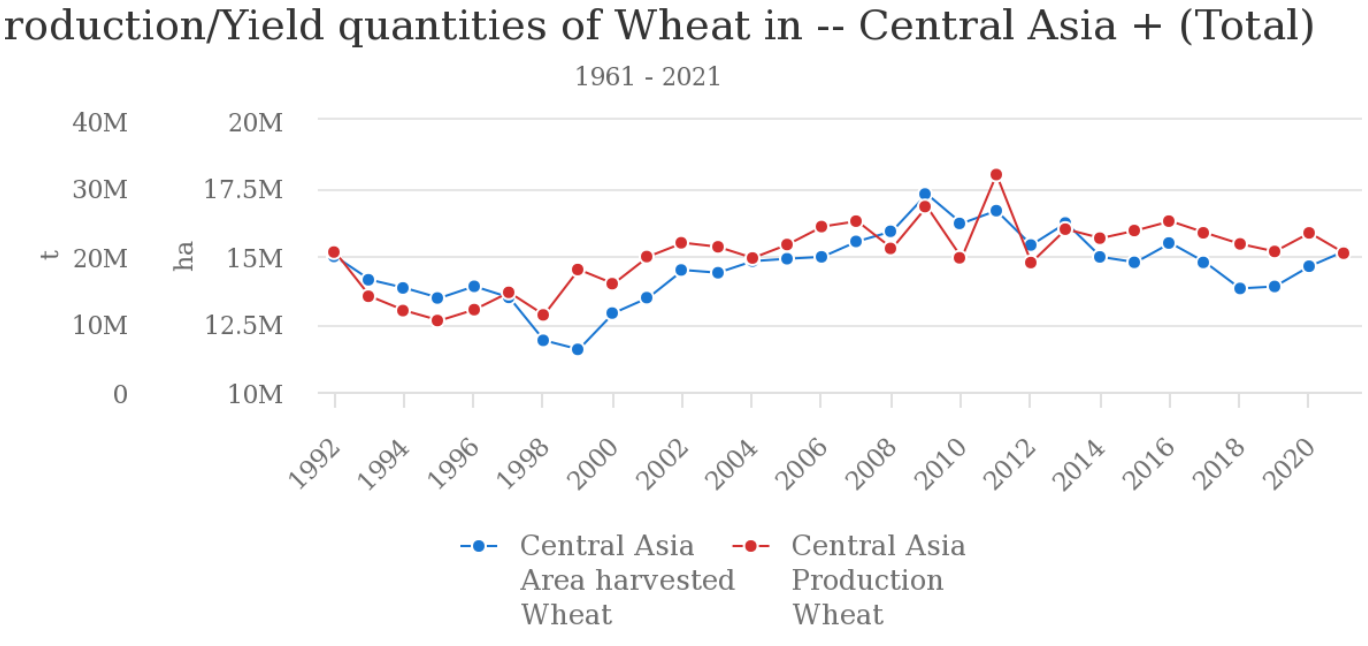
The following charts represent the data for central and southern asia for the staple crops as stated for the years 1960 to 2021.

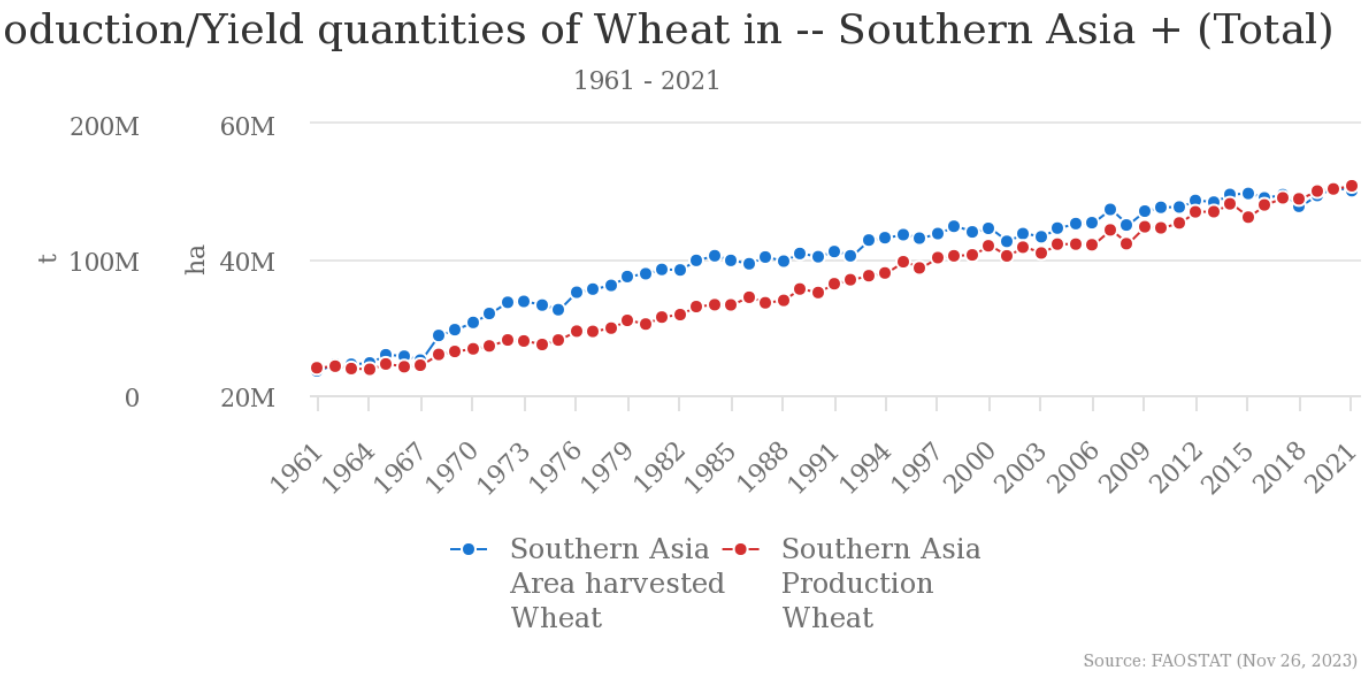



duction/Yield quantities of Potatoes in -- Southern Asia + (Total)
1961 - 2021
100M
50M
0
t
4M
2M
0
ha
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
2015
2018
2021
Southern Asia Area harvested Potatoes
Southern Asia Production Potatoes
Source: FAOSTAT (Nov 26, 2023)


Production/Yield quantities of Rice in -- Central Asia + (Total)
1961 - 2021
1,500k
1,000k
500k
0
t
400k
300k
200k
100k
ha
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Central Asia Area harvested Rice
Central Asia Production Rice
Source: FAOSTAT (Nov 26, 2023)


roduction/Yield quantities of Rice in -- Southern Asia + (Total)
1961 - 2021
300M
200M
100M
0
t
70M
60M
50M
40M
ha
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
2015
2018
2021
Southern Asia Area harvested Rice
Southern Asia Production Rice
Source: FAOSTAT (Nov 26, 2023)


roduction/Yield quantities of Wheat in -- Central Asia + (Total)
1961 - 2021
40M
30M
20M
10M
0
t
20M
17.5M
15M
12.5M
10M
ha
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018
2020
Central Asia Area harvested Wheat
Central Asia Production Wheat
Source: FAOSTAT (Nov 26, 2023)



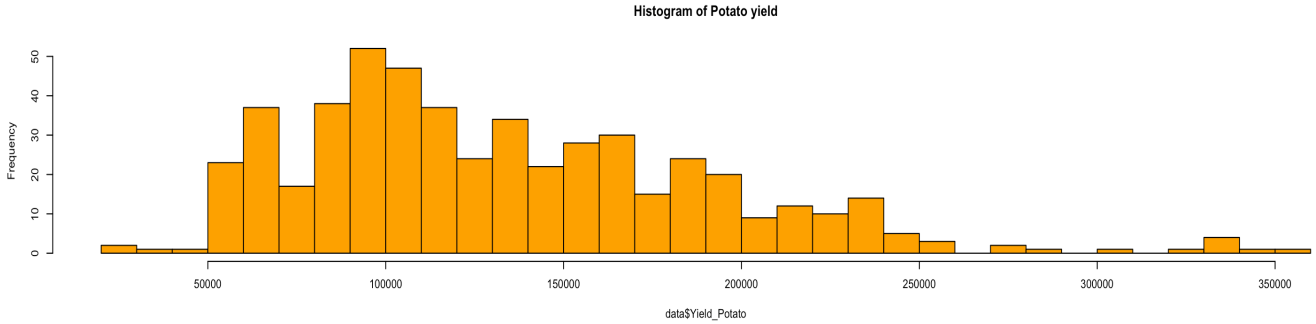
Unfortunately, I didn't have a lot of data for the staple crops mentioned in Maldives, so I had to filter that out.

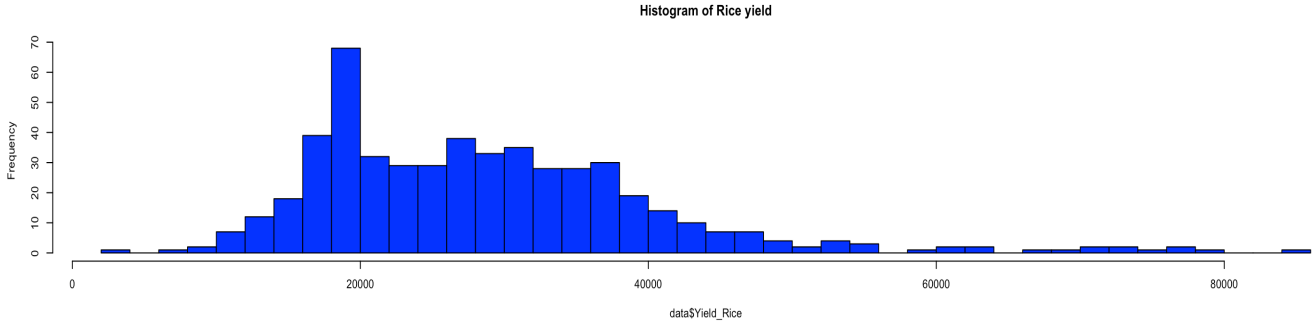
Also had to remove the year 1960 so that it doesn't cause an issue while handling the data.

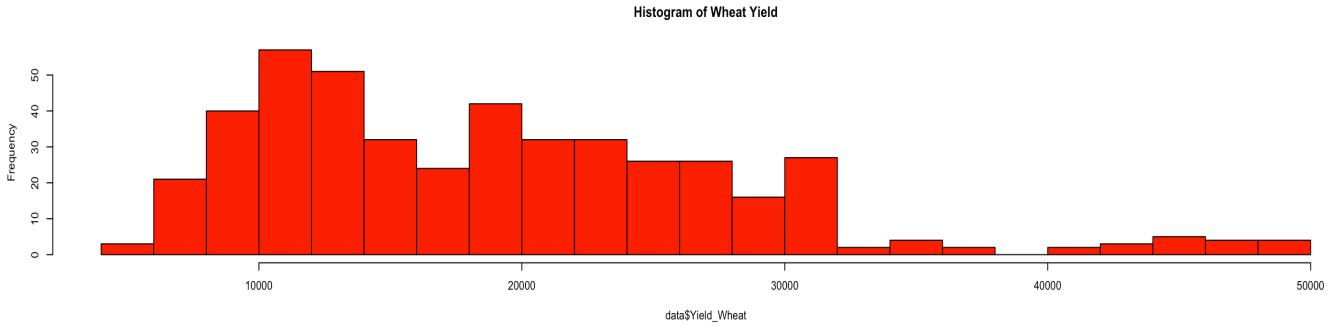
This is how the final dataset looked like:

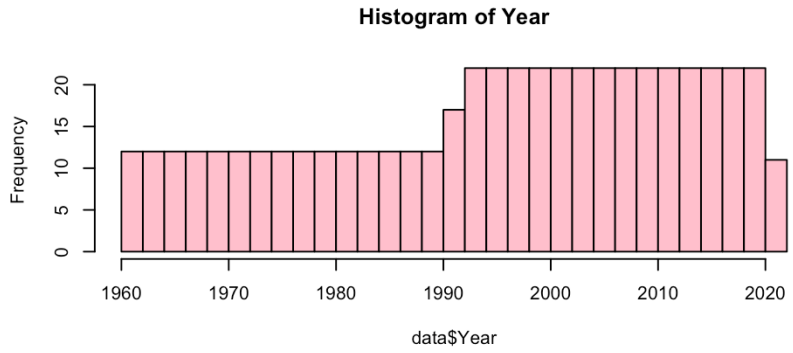
	Year	Country_Name	Country_Code	Agricultural.land....of.land.area.	Agricultural.machinery..tractors	Arable.land....of.land.area.	Fertilizer.consumption..kilograms.per.hectare.of.arable.land.	CO2.emissions..kt.	Inflation..consumer.prices..annual...	GDPgrowth..annual...	Climate.deg.C.	Precipitation.mm.	Yield
1	1961	India	IND	58.84319536	31016	52.4036473	2.17128994	277224	1.69521294	3.72274253	24.37	1188.03	
2	1962	India	IND	59.35779415	35000	52.7043344	2.88576899	290872.6	3.63221497	2.93112774	24.37	1157.69	
3	1963	India	IND	59.24747493	40000	52.9633155	3.45399124	304521.2	2.94616136	5.99435326	24.37	1126.18	
4	1964	India	IND	59.43515214	44000	53.0514363	4.90198565	318169.8	13.3552612	7.45295012	24.38	1098.61	
5	1965	India	IND	59.59154982	48000	53.2142245	4.95904333	331818.4	9.47475859	-2.6357701	24.39	1079.62	
6	1966	India	IND	59.71027751	54012	53.3400153	6.93990794	345467	10.8018484	-0.0553288	24.4	1069.39	
7	1967	India	IND	59.80344344	66000	53.7402588	9.63199399	359115.6	13.0622025	7.82596303	24.4	1066.68	
8	1968	India	IND	59.99078431	78000	53.8374608	10.9996314	372764.2	3.23741243	3.38792918	24.4	1070.22	
9	1969	India	IND	59.79839835	90000	53.8061812	12.3893584	386412.8	-0.5841366	6.5397003	24.4	1078.75	
10	1970	India	IND	59.88584652	1E+05	54.0026033	14.054559	400061.4	5.09226162	5.15722974	24.39	1090.97	
11	1971	India	IND	59.80680683	143000	53.7321866	16.6303613	413710	3.07993868	1.64293038	24.38	1105.33	
12	1972	India	IND	59.85019457	170000	53.8768125	17.2792878	359115.6	6.44209746	-0.5533013	24.37	1120.13	
13	1973	India	IND	60.12094753	184293	54.1653914	17.6256178	372764.2	16.940816	3.29552114	24.36	1133.53	
14	1974	India	IND	60.29954359	203351	54.3190983	15.9336475	386412.8	28.5987341	1.18533626	24.36	1143.81	
15	1975	India	IND	60.43811529	227668	54.4324446	21.5882549	468304.4	5.7484303	9.14991202	24.38	1149.24	
16	1976	India	IND	60.42298003	250884	54.5061029	21.0839396	481953	-7.6339476	1.66310364	24.41	1149.28	
17	1977	India	IND	60.40347236	294313	54.5067756	26.5076299	454655.8	8.30747009	7.25476459	24.46	1145.21	
18	1978	India	IND	60.74014779	334138	54.9036557	31.4324396	464582.0545	2.52304876	5.71253209	24.5	1138.47	
19	1979	India	IND	60.73611172	378714	54.9726052	32.2612026	474508.3091	6.27568337	-5.2381827	24.54	1130.45	
20	1980	India	IND	60.63251928	382869	54.7741651	33.9727609	484434.5636	11.3460735	6.73582153	24.57	1122.58	
21	1981	India	IND	60.67590702	417769	54.7936728	37.3489982	494360.8182	13.1125469	6.00620363	24.59	1115.89	
22	1982	India	IND	60.80035248	461567	54.90601	35.9304363	504287.0727	7.89074279	3.47573324	24.59	1110.09	
23	1983	India	IND	60.69642371	502581	54.7331318	40.7756311	514213.3273	11.8680813	7.2888929	24.59	1104.69	
24	1984	India	IND	61.0293994	553555	55.0311282	48.8907089	524139.5818	8.31890712	3.82073786	24.59	1099.15	
25	1985	India	IND	60.99172942	607773	55.0048937	52.9864253	534065.8364	5.55642423	5.25429922	24.58	1093	
26	1986	India	IND	60.94195124	648932	54.9608333	58.8559994	543992.0909	8.72972073	4.77656417	24.58	1085.91	
27	1987	India	IND	61.06000626	697568	54.8922201	51.1179192	553918.3455	8.80112581	3.96535564	24.59	1079.01	
28	1988	India	IND	60.90798099	750935	54.7788739	66.0964708	563844.6	9.38347186	9.62778292	24.6	1074	
29	1989	India	IND	61.04588001	925365	54.983368	69.210771	573770.8545	7.07428003	5.94734333	24.6	1072.62	
30	1990	India	IND	61.01628218	988070	54.9776503	73.5230241	563575.4	8.9712325	5.53345456	24.61	1076.58	
31	1991	India	IND	61.07446884	1063012	54.8844843	77.9993198	607224	13.8702462	1.05683143	24.61	1087.39	
32	1992	India	IND	60.97995755	1136160	54.7243869	74.7016644	626293.3	11.787817	5.48239602	24.6	1103.47	

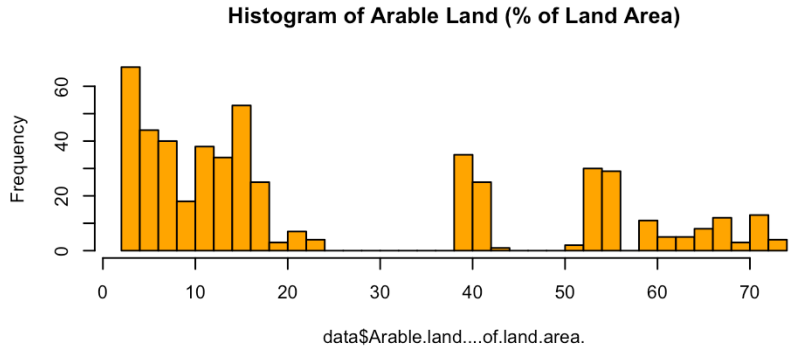
EDA

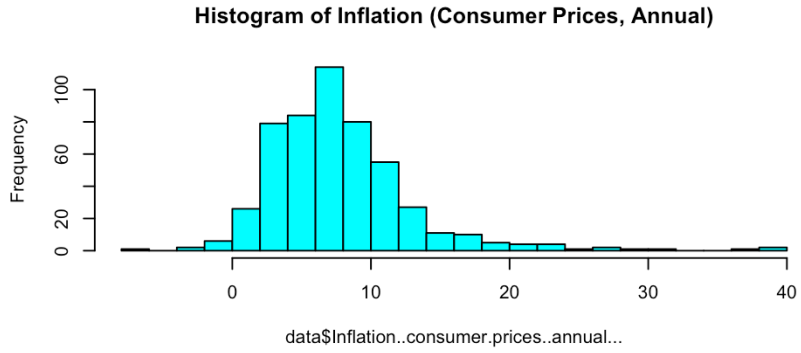

Histogram of Potato yield
Frequency
0
10
20
30
40
50
50000
100000
150000
200000
250000
300000
350000
data$Yield_Potato

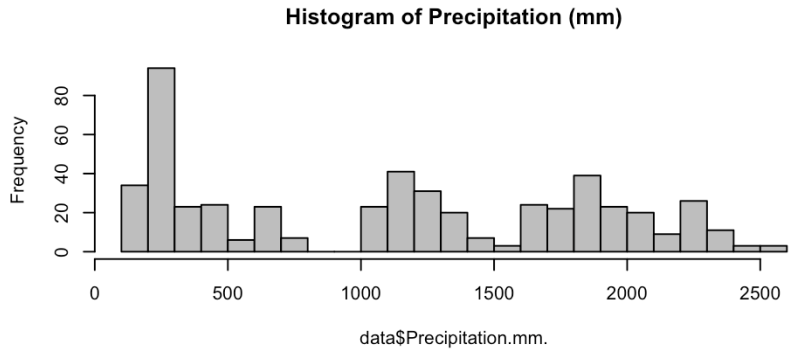

Histogram of Rice yield
Frequency
0
10
20
30
40
50
60
70
0
20000
40000
60000
80000
data$Yield_Rice

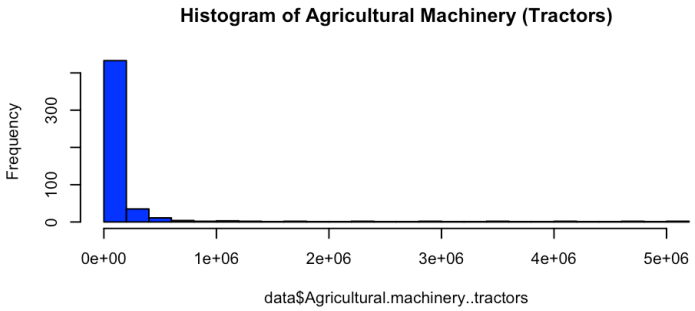

Histogram of Wheat Yield
Frequency
0
10
20
30
40
50
10000
20000
30000
40000
50000
data$Yield_Wheat

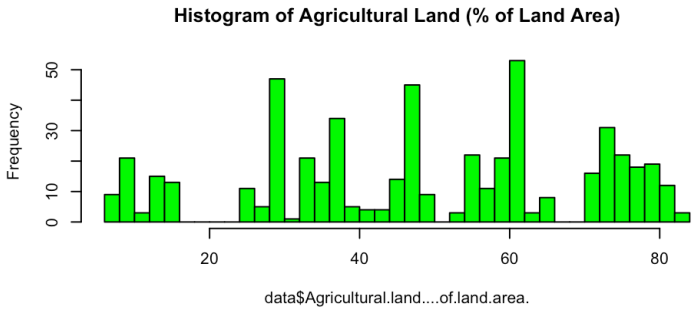

Histogram of Year
Frequency
0
5
10
15
20
1960
1970
1980
1990
2000
2010
2020
data$Year

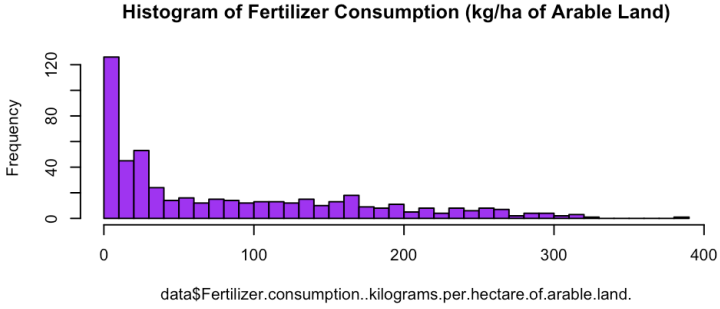

Histogram of Arable Land (% of Land Area)
Frequency
0
20
40
60
0
10
20
30
40
50
60
70
data$Arable.land....of.land.area.

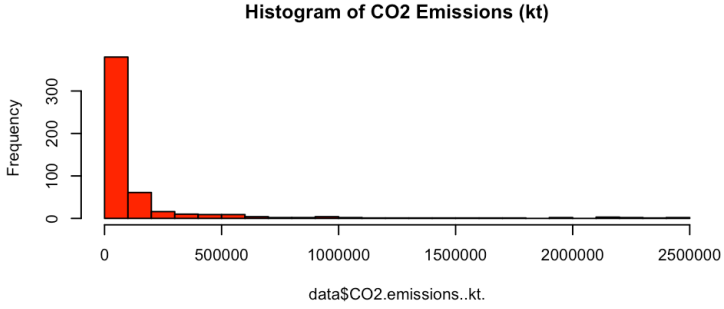

Histogram of Inflation (Consumer Prices, Annual)
Frequency
0
20
60
100
0
10
20
30
40
data$Inflation..consumer.prices..annual...

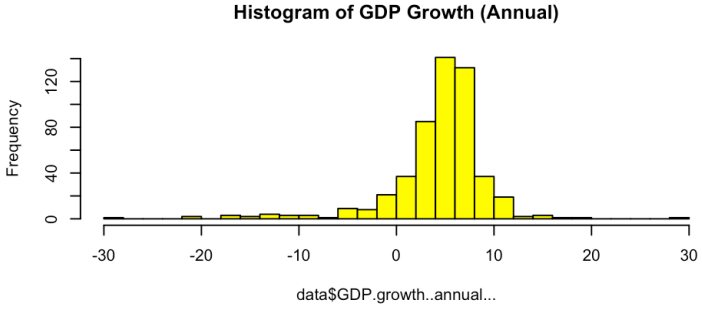

Histogram of Precipitation (mm)
Frequency
0
20
40
60
80
0
500
1000
1500
2000
2500
data$Precipitation.mm.

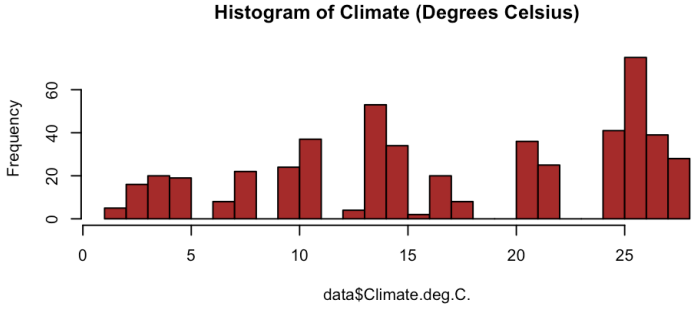

Histogram of Agricultural Machinery (Tractors)
Frequency
0
100
300
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
data$Agricultural.machinery..tractors

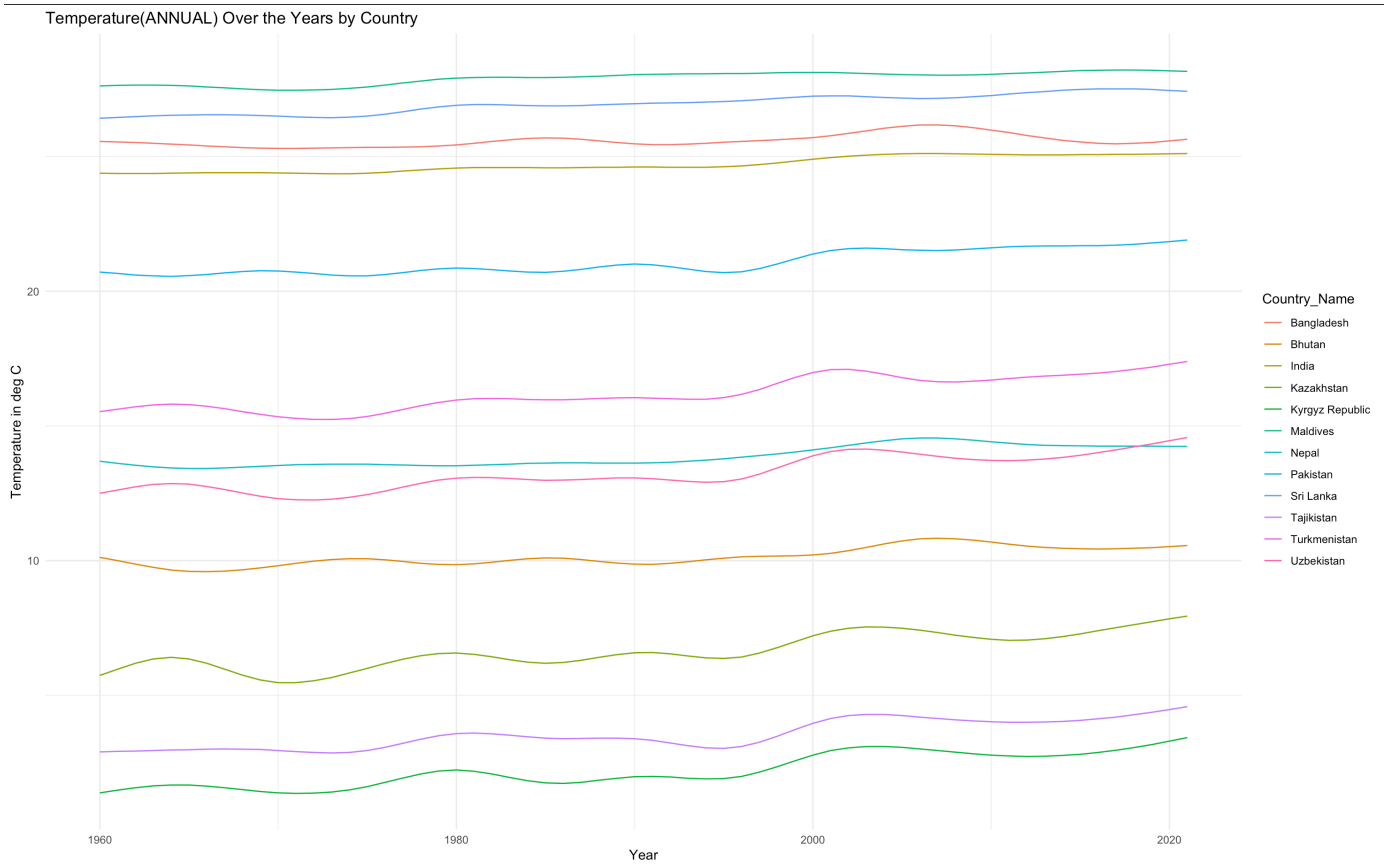

Histogram of Agricultural Land (% of Land Area)
Frequency
0
10
30
50
20
40
60
80
data$Agricultural.land....of.land.area.

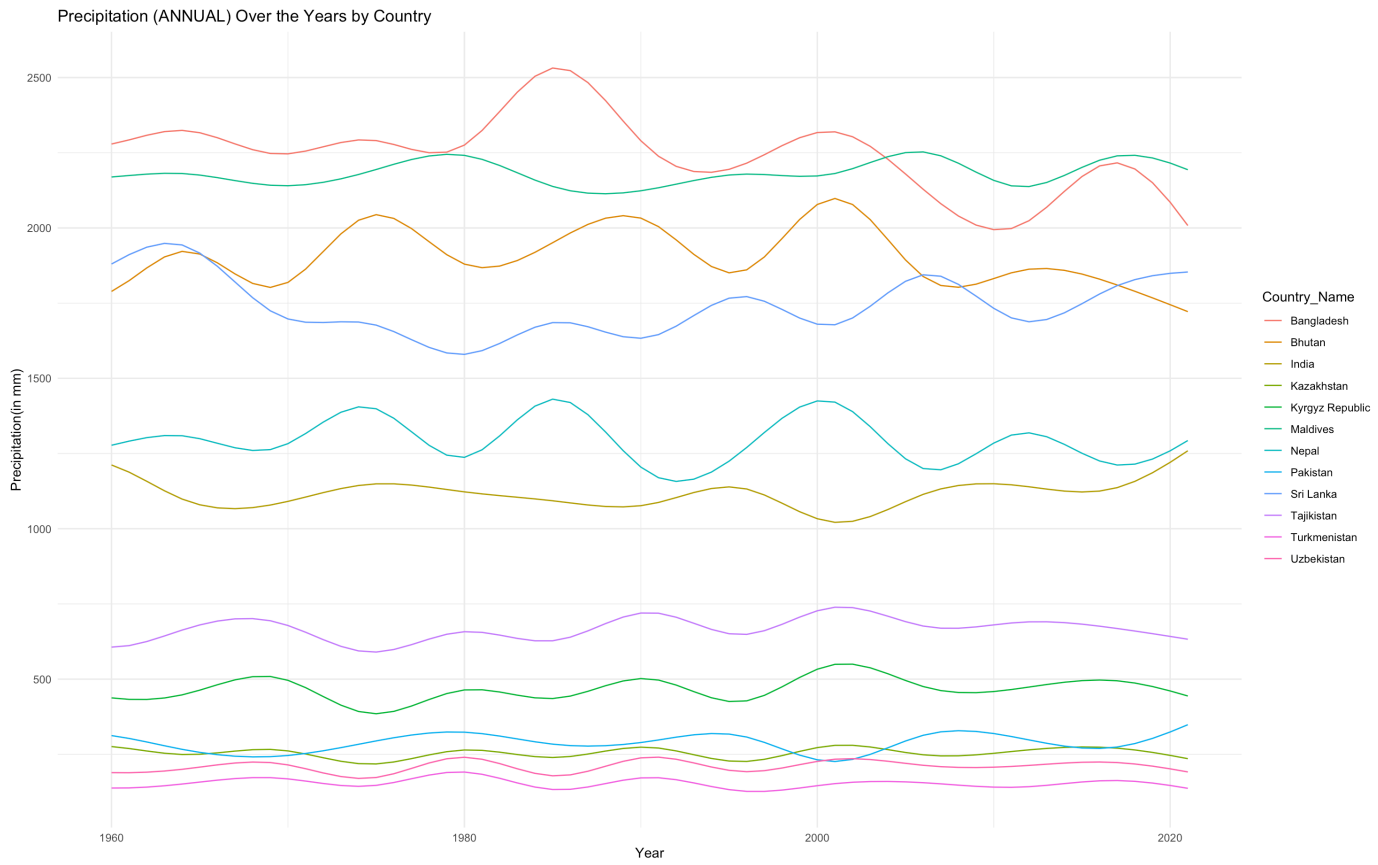

Histogram of Fertilizer Consumption (kg/ha of Arable Land)
Frequency
0
40
80
120
0
100
200
300
400
data$Fertilizer.consumption..kilograms.per.hectare.of.arable.land.


Histogram of CO2 Emissions (kt)
Frequency
0
100
200
300
0
500000
1000000
1500000
2000000
2500000
data$CO2.emissions..kt.


Histogram of GDP Growth (Annual)
Frequency
0
40
80
120
-30
-20
-10
0
10
20
30
data$GDP.growth..annual...


Histogram of Climate (Degrees Celsius)
Frequency
0
20
40
60
0
5
10
15
20
25
data$Climate.deg.C.

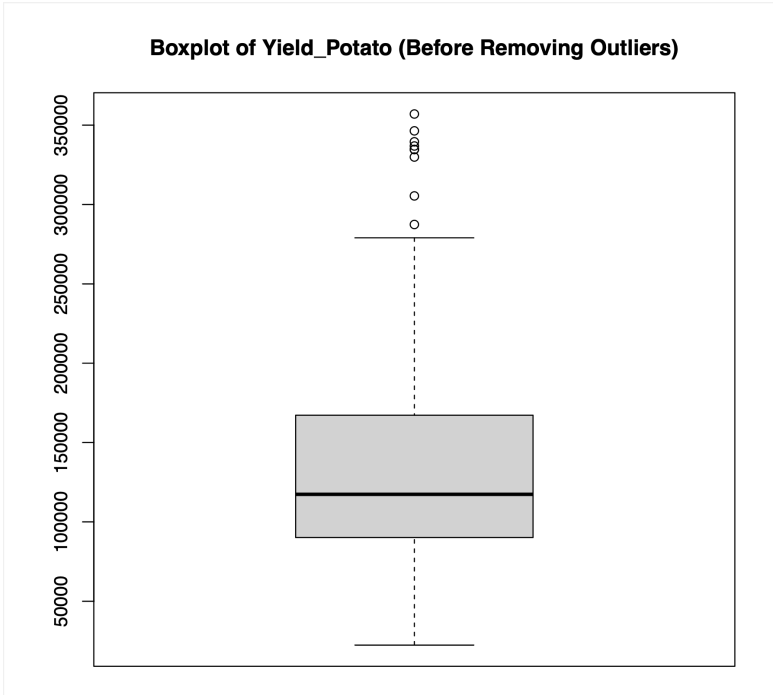

Temperature(ANNUAL) Over the Years by Country
Temperature in deg C
20
10
1960
1980
2000
2020
Year
Country_Name
Bangladesh
Bhutan
India
Kazakhstan
Kyrgyz Republic
Maldives
Nepal
Pakistan
Sri Lanka
Tajikistan
Turkmenistan
Uzbekistan

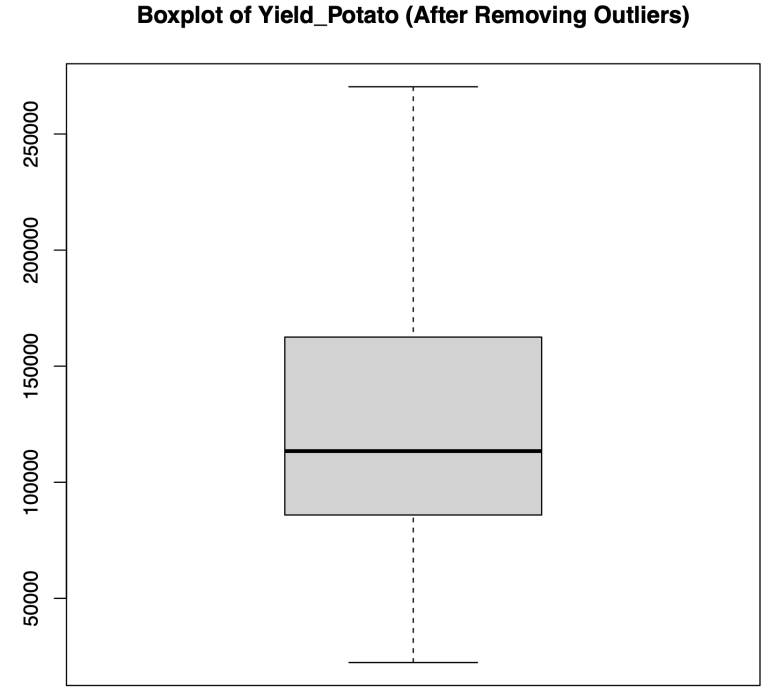


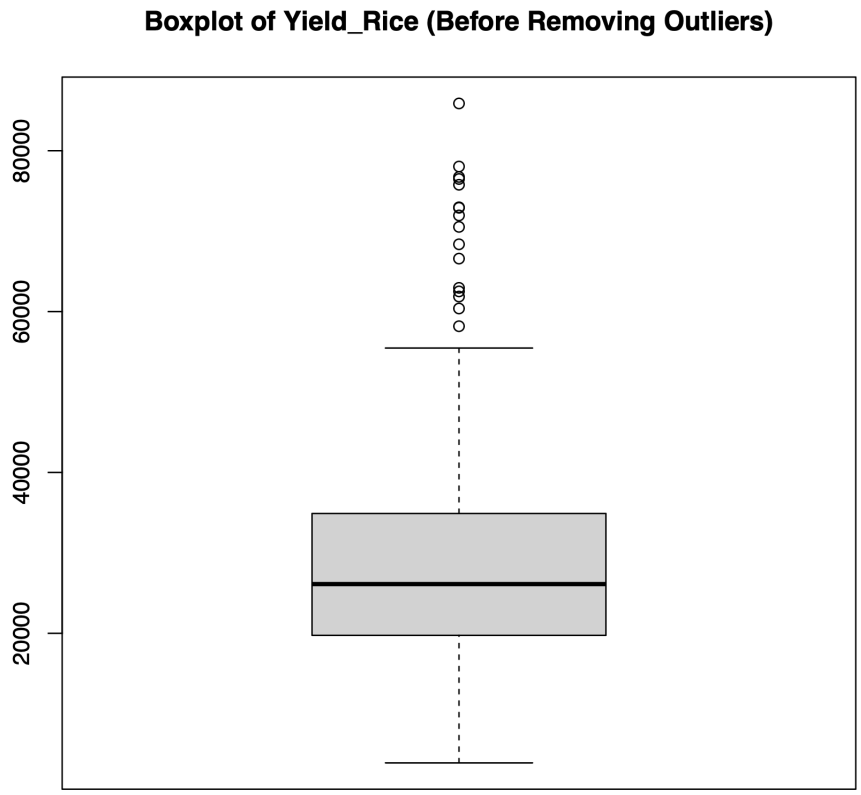

```
Number of outliers in Year : 0
Number of outliers in Agricultural.land....of.land.area. : 0
Number of outliers in Agricultural.machinery..tractors : 108
Number of outliers in Arable.land....of.land.area. : 0
Number of outliers in Fertilizer.consumption..kilograms.per.hectare.of.arable.land. : 5
Number of outliers in CO2.emissions..kt. : 69
Number of outliers in Inflation..consumer.prices..annual... : 26
Number of outliers in GDP.growth..annual... : 59
Number of outliers in Climate.deg.C. : 0
Number of outliers in Precipitation.mm. : 0
Number of outliers in Yield_Potato : 9
Number of outliers in Yield_Rice : 16
Number of outliers in Yield_Wheat : 14
```

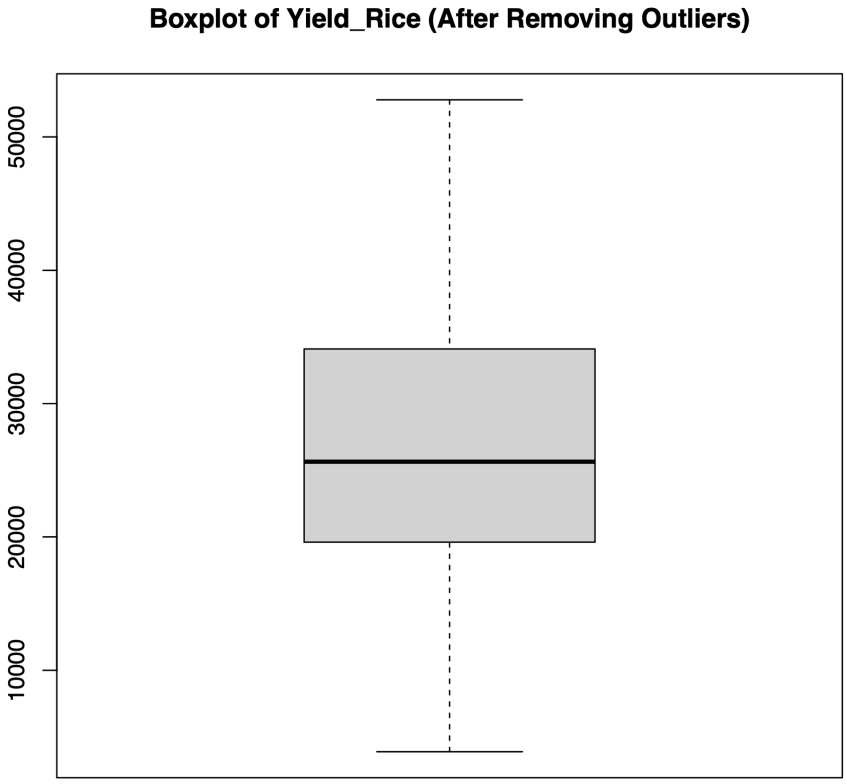
That was my EDA, I cleaned the data, plotted some of the graphs. In some of which contained outliers which are stated as above,

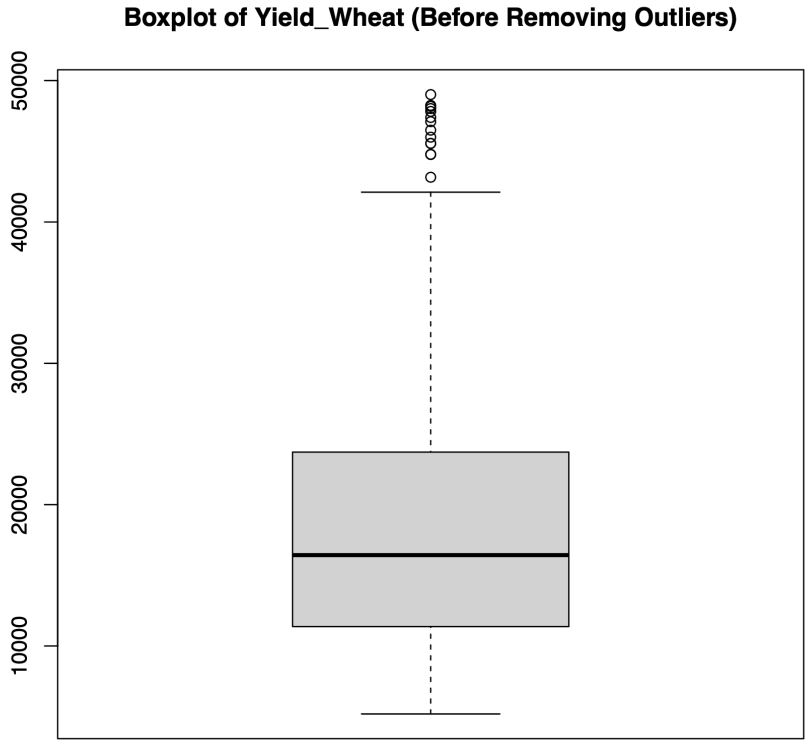
removed the outliers(using IQR method, the pdf can be found as ‘outliers.pdf’). For example, Yields had outliers, this is how their boxplots looked like before and after removal of the outliers.

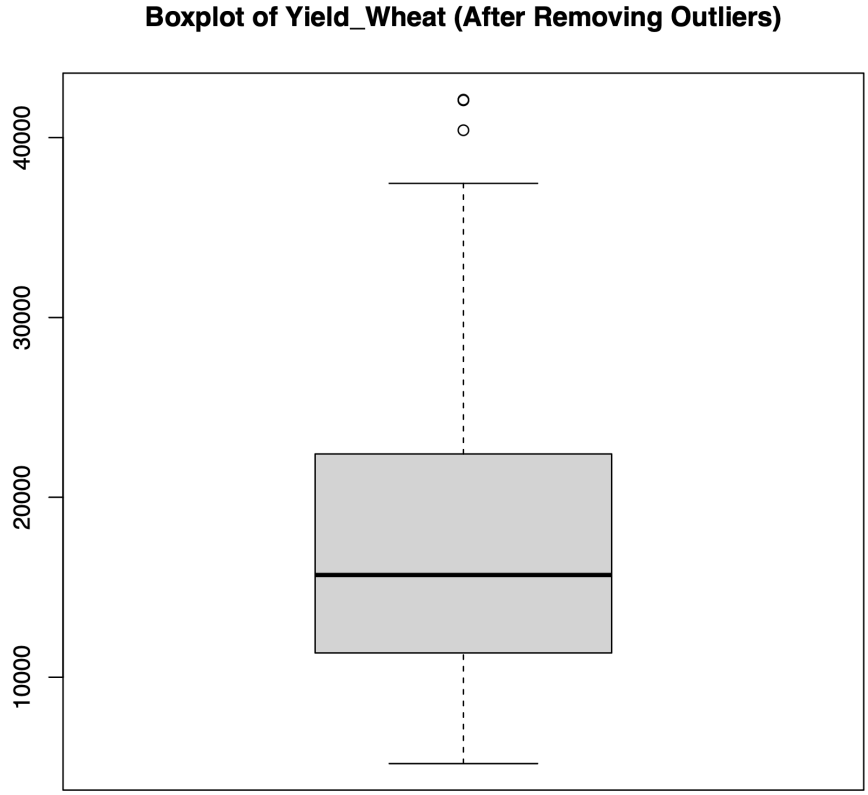





Boxplot of Yield_Rice (Before Removing Outliers)
80000
60000
40000
20000


Boxplot of Yield_Rice (After Removing Outliers)
50000
40000
30000
20000
10000


Boxplot of Yield_Wheat (Before Removing Outliers)
50000
40000
30000
20000
10000


Boxplot of Yield_Wheat (After Removing Outliers)
40000
30000
20000
10000

Significance Test (F- Statistics Test)

For my response variables of the crop yields(rice, wheat and potato) I had to decide which economical and environmental data according to country in the cleaned data were significant to perform my analysis and build models, so I chose to do a F-test.

After all this was done, I performed a significance test on all the continuous variables from the cleaned data decided which features to choose to train the models and the following results came out:

```
Year is significant (p-value = 3.420831e-74 )
Agricultural.land....of.land.area. is significant (p-value = 0.02170632 )
Agricultural.machinery..tractors is significant (p-value = 0.01067103 )
Arable.land....of.land.area. is significant (p-value = 0.02946193 )
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. is significant (p-value = 1.518561e-21 )
CO2.emissions..kt. is significant (p-value = 0.02008196 )
Inflation..consumer.prices..annual... is not significant (p-value = 0.6206307 )
GDP.growth..annual... is significant (p-value = 0.005107327 )
Climate.deg.C. is significant (p-value = 0.00266868 )
Precipitation.mm. is significant (p-value = 0.000401127 )
```

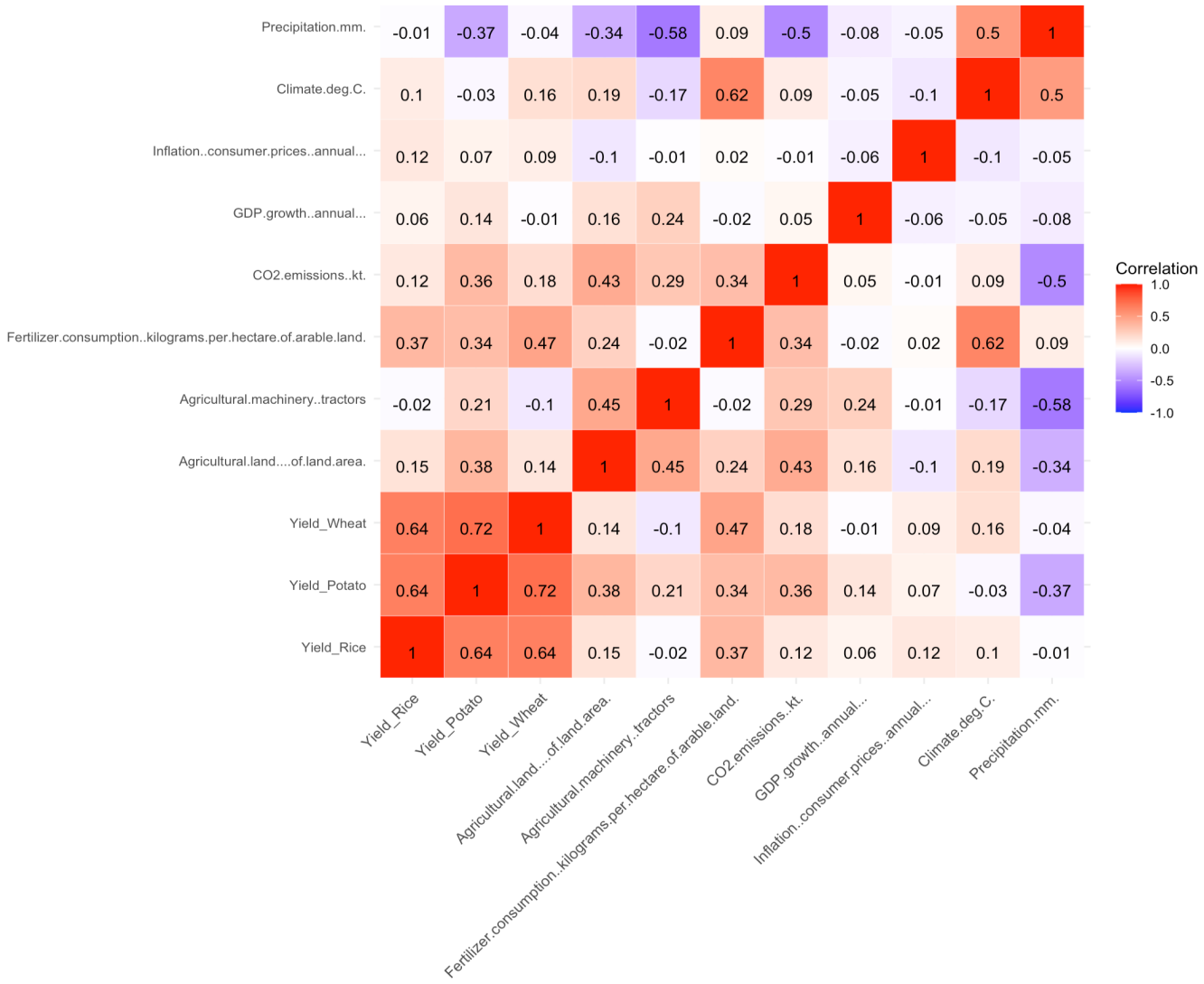
And made a contingency table for the categorical variables, country name and code, and found out they both had a positive significance on the Yields.

Correlation Matrix

1. **Yield Relationships:**
    - Yield_Rice and Yield_Potato have a moderate positive correlation of 0.6411.
    - Yield_Potato and Yield_Wheat also have a moderate positive correlation of 0.7158.
    - Yield_Rice and Yield_Wheat have a slightly lower but still significant positive correlation of 0.6376.
2. **Agricultural Factors:**
    - Agricultural.land....of.land.area. shows a weak positive correlation with crop yields.
    - Agricultural.machinery..tractors has a moderate positive correlation with Agricultural.land....of.land.area.

- Fertilizer.consumption..kilograms.per.hectare.of.arable.land. has moderate positive correlations with all three crop yields.

3. **Environmental Factors:**
   - CO2.emissions..kt. has a moderate positive correlation with Agricultural.land....of.land.area.
   - Climate.deg.C. shows a weak positive correlation with Agricultural.land....of.land.area.
   - Precipitation.mm. has a negative correlation with Agricultural.machinery..tractors and a weak negative correlation with Agricultural.land....of.land.area.



4. **Economic Factors:**
    - GDP.growth..annual... has weak positive correlations with Agricultural.land....of.land.area. and CO2.emissions..kt.
    - Inflation..consumer.prices..annual... has a weak negative correlation with Agricultural.land....of.land.area.
5. **Climate Factors:**
    - Climate.deg.C. has a moderate positive correlation with Fertilizer.consumption..kilograms.per.hectare.of.arable.land.
    - Precipitation.mm. has a moderate positive correlation with Climate.deg.C.

Data Modelling

Firstly, the data was divided into training and testing sets and then, I built the following regression models(as my targeted variables were all continuous):

1) Multiple Linear Regression Model
2) Decision Tree Model
3) Random Forest Regressor
4) Polynomial Regression Model

1) Multiple Linear Regression Model

This model assumes a linear relationship between the yield and all the independent variables. It simply fits a straight line through the data points, minimizing the squared error between the predicted and actual values.

Results(Rice):

Root Mean Squared Error: 6444.398

-The RMSE for Rice is a measure of the average magnitude of the errors between the predicted and actual Rice yields. A lower RMSE indicates that, on average, the model's predictions are closer to the true values. In this case, the RMSE of 6444.398 suggests a moderate level of prediction error for Rice.

R-squared: 0.65081

-The R-squared value for Rice is 0.65081, indicating that approximately 65.08% of the variability in Rice yields can be explained by the variables included in the model. This suggests a moderate level of explanatory power, with room for improvement.

Results(Potato):

Root Mean Squared Error: 23622.58

- The RMSE for Potato is considerably higher than for Rice, indicating a larger average difference between the model's predictions and the actual Potato yields. This suggests that the model has more difficulty accurately predicting Potato yields compared to Rice.

R-squared: 0.7790014

-The R-squared value for Potato is 0.7790014, signifying that around 77.90% of the variability in Potato yields is explained by the model. This is a relatively high R-squared, suggesting that the model is better at explaining the variance in Potato yields compared to Rice.

Results(Wheat):

Root Mean Squared Error: 4401.186

-The RMSE for Wheat is the lowest among the three crops, suggesting that the model's predictions for Wheat yields are, on average, closer to the actual values. A lower RMSE is generally desirable, indicating better predictive performance, and in this case, Wheat appears to be the crop with the most accurate predictions.

R-squared: 0.7582456

-The R-squared value for Wheat is 0.7582456, indicating that approximately 75.82% of the variability in Wheat yields is explained by the model. This is a reasonably high level of explanatory power, although it is slightly lower than that for Potato.

MLR	Rice	Potato	Wheat
RMSE	6444.398	23622.58	4401.186
RSQUARE	0.65081	0.7790014	0.7582456

In summary, while Potato has the highest R-squared value, indicating strong explanatory power, Wheat stands out with the lowest RMSE, suggesting more accurate predictions.

2) Decision Tree Model

This model splits the data into smaller and smaller subsets based on the values of the independent variables. It then fits a simple prediction rule (e.g., average yield) within each subset. This can capture non-linear relationships and interactions between variables.

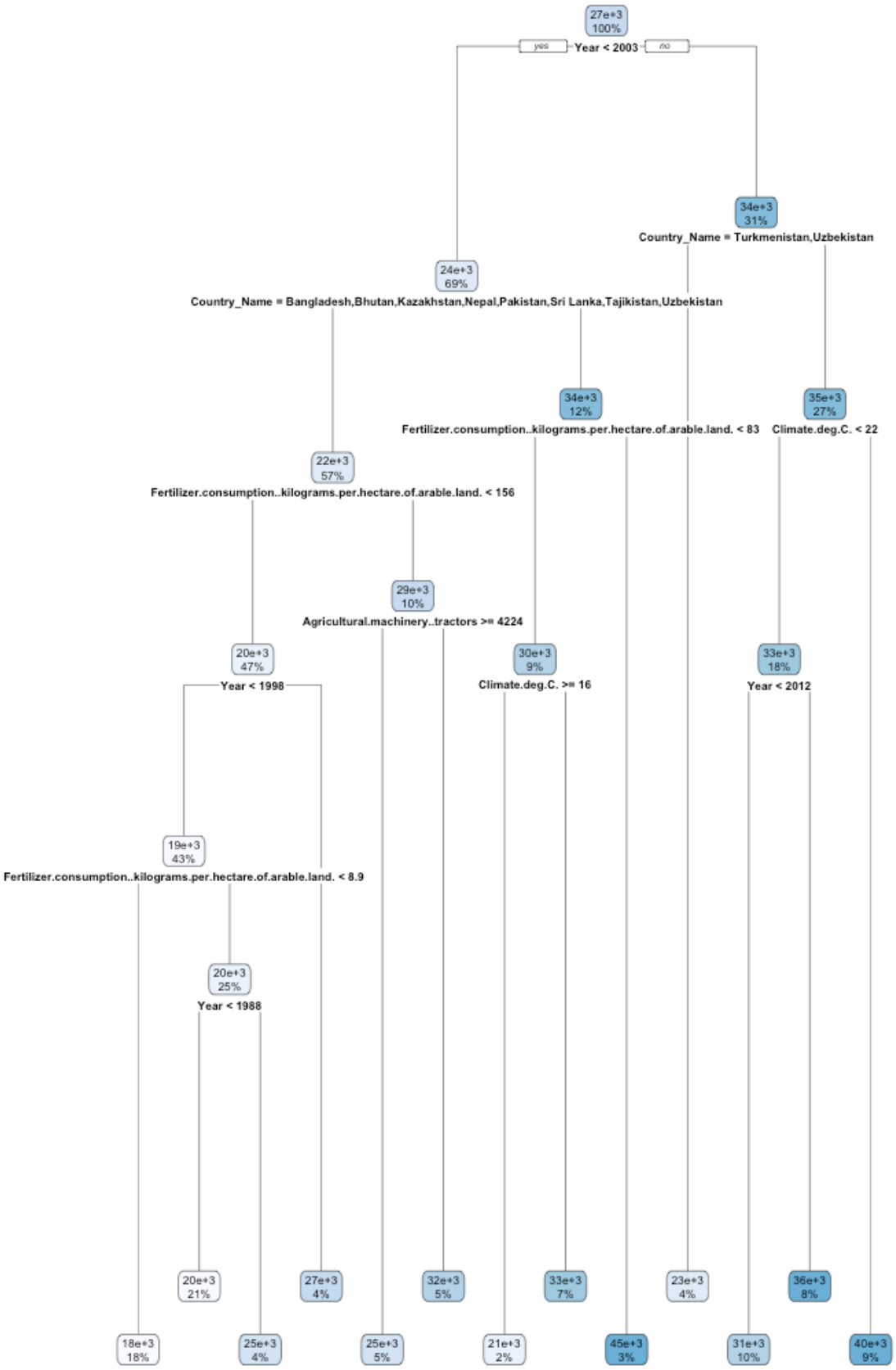
Results(Rice):

Decision Tree Root Mean Squared Error: 5691.204

-The RMSE of 5691.204 for Rice indicates the average magnitude of errors between the predicted and actual Rice yields. This value is lower than the RMSE for Rice in the Multiple Linear Regression (MLR) model (6444.398), suggesting that the Decision Tree model performs slightly better in predicting Rice yields compared to the MLR model.

Decision Tree R-squared: 0.6312062

-The R-squared value of 0.6312062 indicates that approximately 63.12% of the variability in Rice yields is explained by the Decision Tree model. This is slightly lower than the R-squared value for Rice in the MLR model (0.65081), suggesting that the MLR model has a slightly higher explanatory power for Rice yields


27e+3
100%
yes
Year < 2003
no
34e+3
31%
Country_Name = Turkmenistan,Uzbekistan
24e+3
69%
Country_Name = Bangladesh,Bhutan,Kazakhstan,Nepal,Pakistan,Sri Lanka,Tajikistan,Uzbekistan
34e+3
12%
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. < 83
35e+3
27%
Climate.deg.C. < 22
22e+3
57%
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. < 156
29e+3
10%
Agricultural.machinery..tractors >= 4224
20e+3
47%
Year < 1998
30e+3
9%
Climate.deg.C. >= 16
33e+3
18%
Year < 2012
19e+3
43%
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. < 8.9
20e+3
25%
Year < 1988
20e+3
21%
27e+3
4%
32e+3
5%
33e+3
7%
23e+3
4%
36e+3
8%
18e+3
18%
25e+3
4%
25e+3
5%
21e+3
2%
45e+3
3%
31e+3
10%
40e+3
9%

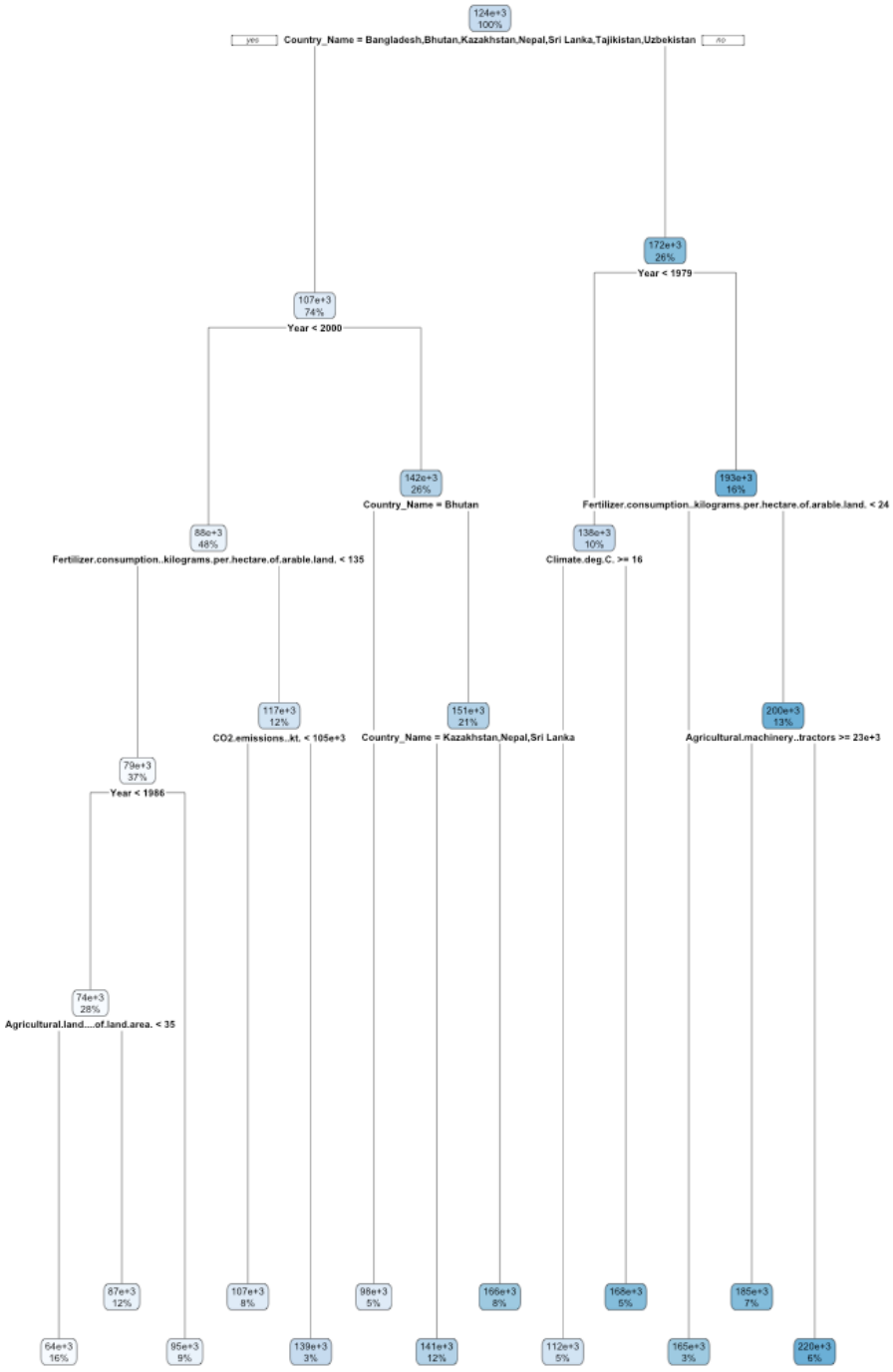
Results(Potato):

Decision Tree Root Mean Squared Error: 25895.34

-The RMSE of 25895.34 for Potato is higher than the RMSE for Potato in the MLR model (23622.58), indicating that the Decision Tree model performs less well in predicting Potato yields compared to the MLR model.

Decision Tree R-squared: 0.7363685

- The R-squared value of 0.7363685 indicates that approximately 73.64% of the variability in Potato yields is explained by the Decision Tree model. This is slightly lower than the R-squared value for Potato in the MLR model (0.7790014), suggesting that the MLR model has a slightly higher explanatory power for Potato yields.


124e+3
100%
yes
Country_Name = Bangladesh,Bhutan,Kazakhstan,Nepal,Sri Lanka,Tajikistan,Uzbekistan
no
107e+3
74%
Year < 2000
172e+3
26%
Year < 1979
88e+3
48%
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. < 135
142e+3
26%
Country_Name = Bhutan
138e+3
10%
Climate.deg.C. >= 16
193e+3
16%
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. < 24
79e+3
37%
Year < 1986
117e+3
12%
CO2.emissions..kt. < 105e+3
151e+3
21%
Country_Name = Kazakhstan,Nepal,Sri Lanka
200e+3
13%
Agricultural.machinery..tractors >= 23e+3
74e+3
28%
Agricultural.land....of.land.area. < 35
64e+3
16%
87e+3
12%
95e+3
9%
107e+3
8%
139e+3
3%
98e+3
5%
141e+3
12%
166e+3
8%
112e+3
5%
168e+3
5%
165e+3
3%
185e+3
7%
220e+3
6%

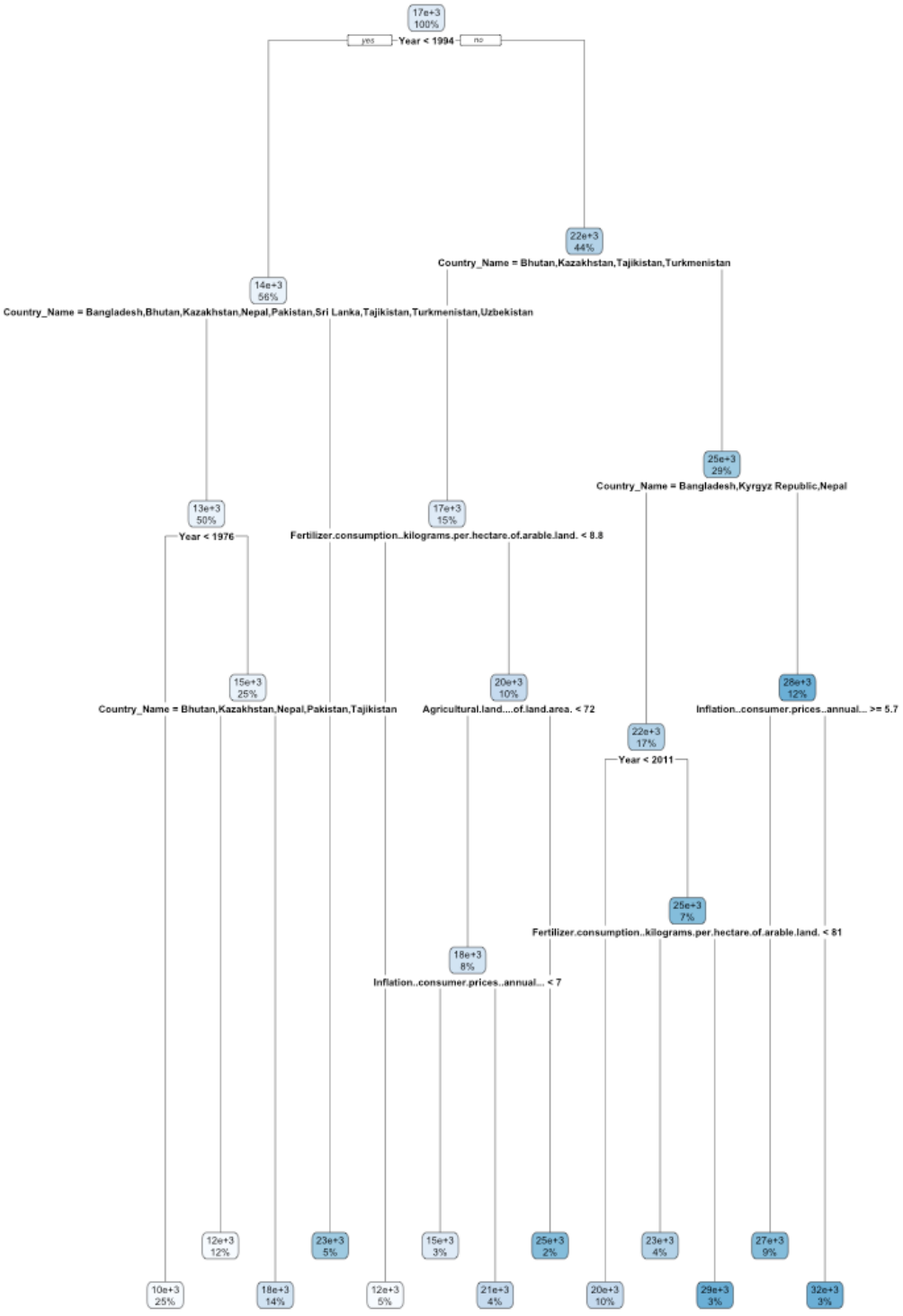
Results(Wheat):

Decision Tree Root Mean Squared Error: 4651.095

- The RMSE of 4651.095 for Wheat is lower than the RMSE for Wheat in the MLR model (4401.186), suggesting that the Decision Tree model performs slightly better in predicting Wheat yields compared to the MLR model.

Decision Tree R-squared: 0.6522271

- The R-squared value of 0.6522271 indicates that approximately 65.22% of the variability in Wheat yields is explained by the Decision Tree model. This is similar to the R-squared value for Wheat in the MLR model (0.7582456), suggesting that both models have comparable explanatory power for Wheat yields.


17e+3
100%
yes
Year < 1994
no
14e+3
56%
Country_Name = Bangladesh,Bhutan,Kazakhstan,Nepal,Pakistan,Sri Lanka,Tajikistan,Turkmenistan,Uzbekistan
22e+3
44%
Country_Name = Bhutan,Kazakhstan,Tajikistan,Turkmenistan
13e+3
50%
Year < 1976
17e+3
15%
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. < 8.8
25e+3
29%
Country_Name = Bangladesh,Kyrgyz Republic,Nepal
15e+3
25%
Country_Name = Bhutan,Kazakhstan,Nepal,Pakistan,Tajikistan
20e+3
10%
Agricultural.land....of.land.area. < 72
28e+3
12%
Inflation..consumer.prices..annual.... >= 5.7
22e+3
17%
Year < 2011
25e+3
7%
Fertilizer.consumption..kilograms.per.hectare.of.arable.land. < 81
18e+3
8%
Inflation..consumer.prices..annual.... < 7
12e+3
12%
23e+3
5%
15e+3
3%
25e+3
2%
23e+3
4%
27e+3
9%
10e+3
25%
18e+3
14%
12e+3
5%
21e+3
4%
20e+3
10%
29e+3
3%
32e+3
3%

Decision Tree	Rice	Potato	Wheat
RMSE	5691.204	25895.34	4651.095
RSQUARE	0.6312062	0.7363685	0.6522271

In summary, the Decision Tree model shows varying performance across different crops, with improvements in some cases (e.g., Rice and Wheat) and degradation in others (e.g., Potato) compared to the MLR model

3) **Random Forest Regressor**

This is an ensemble of multiple decision trees. Each tree is trained on a different bootstrapped sample of the data and different random subsets of features. This reduces variance and helps prevent overfitting.

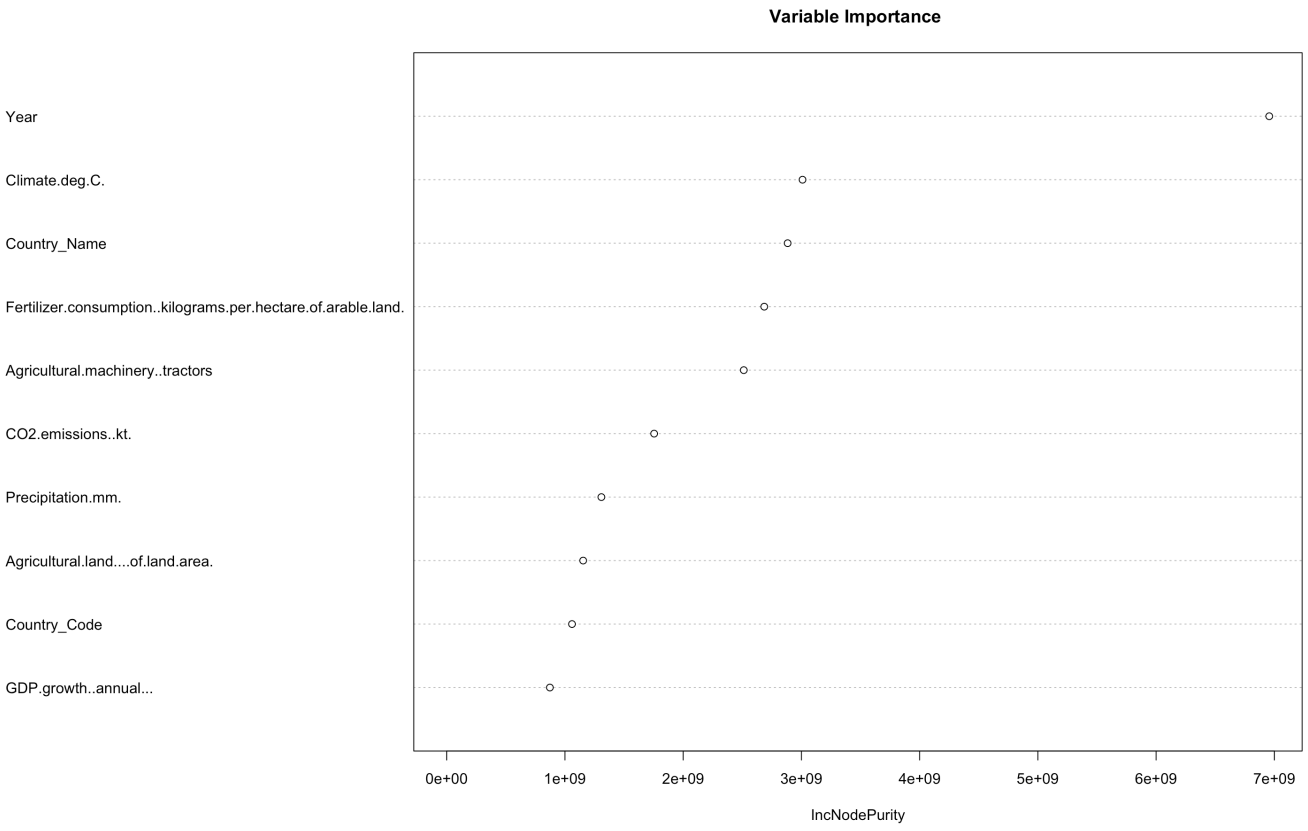
Results(Rice):

Random Forest Root Mean Squared Error: 4672.501

-The RMSE of 4672.501 for Rice indicates the average magnitude of errors between the predicted and actual Rice yields. This value is lower than the RMSE for Rice in both the Multiple Linear Regression (MLR) model (6444.398) and the Decision Tree model (5691.204), suggesting that the Random Forest model performs better in predicting Rice yields.

Random Forest R-squared: 0.7514155

-The R-squared value of 0.7514155 indicates that approximately 75.14% of the variability in Rice yields is explained by the Random Forest model. This is an improvement compared to both the MLR model (0.65081) and the Decision Tree model (0.6312062), suggesting that the Random Forest model has higher explanatory power for Rice yields.



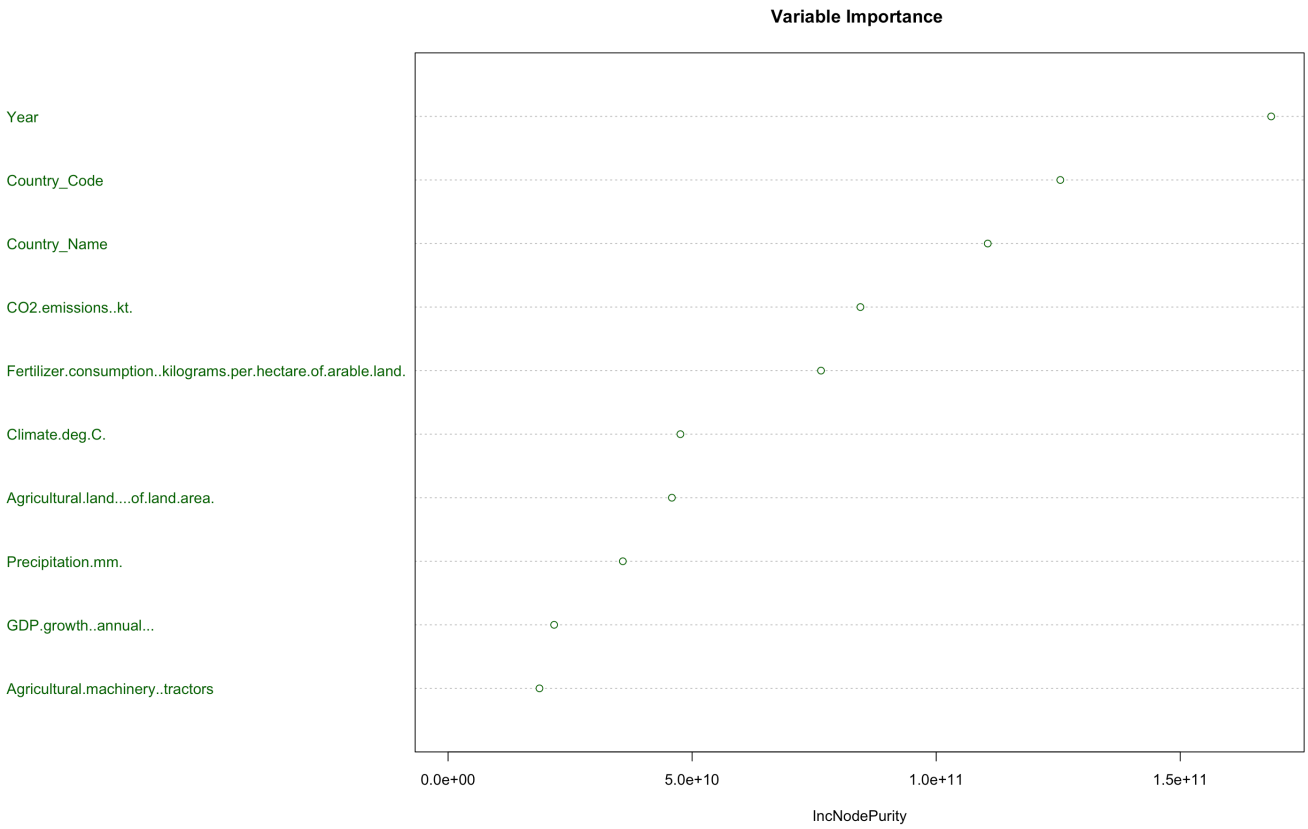
Results(Potato):

Random Forest Root Mean Squared Error (Yield_Potato): 18394.53

-The RMSE of 18394.53 for Potato is higher than the RMSE for Potato in both the MLR model (23622.58) and the Decision Tree model (25895.34), indicating that the Random Forest model performs less well in predicting Potato yields compared to the MLR and Decision Tree models.

Random Forest R-squared (Yield_Potato): 0.8669755

-The R-squared value of 0.8669755 indicates that approximately 86.70% of the variability in Potato yields is explained by the Random Forest model. This is an improvement compared to both the MLR model (0.7790014) and the Decision Tree model (0.7363685), suggesting that the Random Forest model has higher explanatory power for Potato yields.



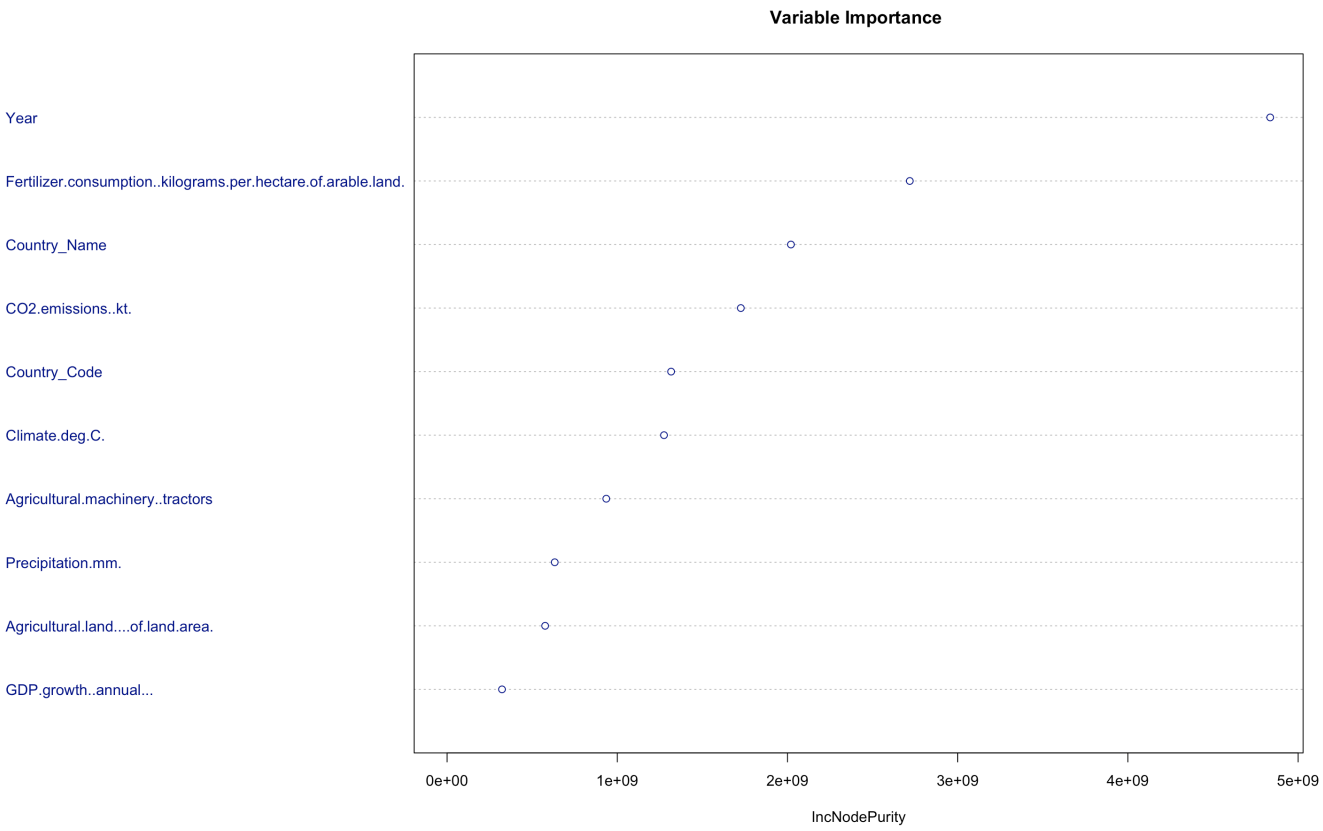
Results(Wheat):

Random Forest Root Mean Squared Error (Yield_Wheat): 3175.575

-The RMSE of 3175.575 for Wheat is lower than the RMSE for Wheat in both the MLR model (4401.186) and the Decision Tree model (4651.095), suggesting that the Random Forest model performs better in predicting Wheat yields.

Random Forest R-squared (Yield_Wheat): 0.8378824

- The R-squared value of 0.8378824 indicates that approximately 83.79% of the variability in Wheat yields is explained by the Random Forest model. This is an improvement compared to both the MLR model (0.7582456) and the Decision Tree model (0.6522271), suggesting that the Random Forest model has higher explanatory power for Wheat yields.



Random Forest	Rice	Potato	Wheat
RMSE	4672.501	18394.53	3175.575
RSQUARE	0.7514155	0.8669755	0.8378824

In summary, the Random Forest model outperforms both the Multiple Linear Regression and Decision Tree models across all three crops (Rice, Potato, and Wheat) in terms of both RMSE and R-squared. The Random Forest model demonstrates improved predictive accuracy and explanatory power for the given agricultural yield prediction task.

4) Polynomial Regression Model

This model extends MLR by including terms that capture non-linear relationships, such as squared or cross-product terms of the variables. It can be more flexible than MLR.

Results(Rice):
Polynomial Model - Root Mean Squared Error: 5018.033

-The RMSE of 5018.033 for Rice indicates the average magnitude of errors between the predicted and actual Rice yields. This value is higher than the RMSE for Rice in the Multiple Linear Regression (MLR) model (6444.398) but lower than that of the Polynomial model (5018.033), suggesting that the Polynomial model performs better in predicting Rice yields than the MLR model.

Polynomial Model - R-squared: 0.6764281

-The R-squared value of 0.6764281 indicates that approximately 67.64% of the variability in Rice yields is explained by the Polynomial model. This is an improvement compared to the MLR model (0.65081), suggesting that the Polynomial model has slightly higher explanatory power for Rice yields.

Results(Potato):

Polynomial Model for Yield_Potato - Root Mean Squared Error: 23494.38

-The RMSE of 23494.38 for Potato is higher than the RMSE for Potato in both the MLR model (23622.58) and the Polynomial model (23494.38). This suggests that the Polynomial model does not significantly improve the predictive accuracy for Potato yields compared to the MLR model.

Polynomial Model for Yield_Potato - R-squared: 0.7813935

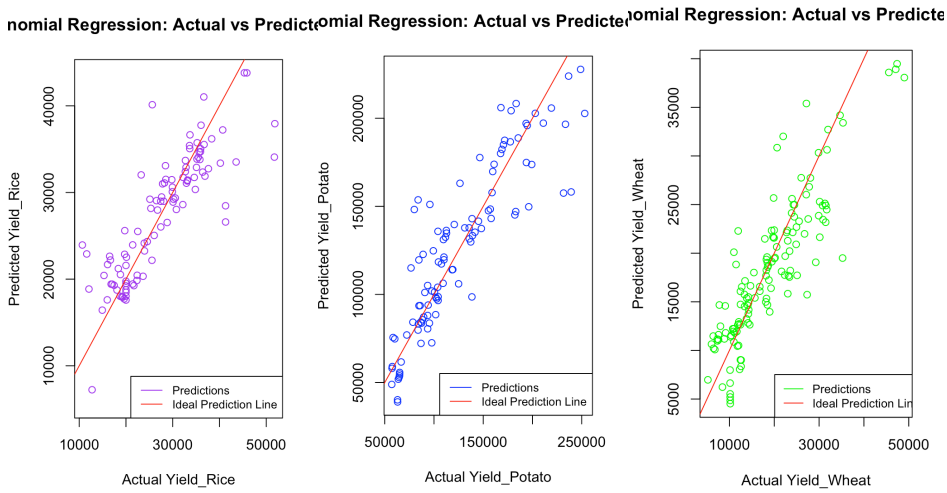
-The R-squared value of 0.7813935 indicates that approximately 78.14% of the variability in Potato yields is explained by the Polynomial model. This is an improvement compared to the MLR model (0.7790014), suggesting that the Polynomial model has slightly higher explanatory power for Potato yields.

Results(Wheat):
Polynomial Model for Yield_Wheat - Root Mean Squared Error: 4462.389

- The RMSE of 4462.389 for Wheat is lower than the RMSE for Wheat in the MLR model (4401.186) but higher than that of the Polynomial model (4462.389). This suggests that the Polynomial model does not significantly improve the predictive accuracy for Wheat yields compared to the MLR model.

Polynomial Model for Yield_Wheat - R-squared: 0.7514752

-The R-squared value of 0.7514752 indicates that approximately 75.15% of the variability in Wheat yields is explained by the Polynomial model. This is similar to the R-squared value for Wheat in the MLR model (0.7582456), suggesting that both models have comparable explanatory power for Wheat yields.



Polynomial Regression	Rice	Potato	Wheat
RMSE	5018.033	23494.38	4462.389
RSQUARE	0.6764281	0.7813935	0.7514752

In summary, the Polynomial Regression model shows mixed results across different crops. It appears to improve the performance for Rice but does not significantly enhance predictive accuracy for Potato and Wheat compared to the Multiple Linear Regression model.

Conclusion

In conclusion, this research aimed to address the critical issue of food security in Central and Southern Asia by investigating factors affecting crop production, with a focus on wheat, rice, and potatoes. The study employed advanced machine learning models, including Multiple Linear Regression (MLR), Decision Tree, Random Forest Regressor, and Polynomial Regression, to predict crop yields based on various influencing factors. The analysis was conducted separately for each crop, revealing distinct performance characteristics for each model.

The MLR model demonstrated moderate predictive accuracy across all crops, with varying levels of explanatory power. While Potato exhibited the highest R-squared value, indicating strong explanatory power, Wheat stood out with the lowest Root Mean Squared Error (RMSE), suggesting more accurate predictions.

The Decision Tree model showed mixed performance across crops, with improvements in some cases (e.g., Rice and Wheat) and degradation in others (e.g., Potato) compared to the MLR model. The model's ability to capture non-linear relationships and interactions between variables contributed to its varied performance.

The Random Forest Regressor, an ensemble of decision trees, generally outperformed both the MLR and Decision Tree models across all crops in terms of both RMSE and R-squared. The Random Forest model demonstrated improved predictive accuracy and explanatory power, showcasing its effectiveness in addressing the complex interactions influencing crop yields.

Lastly, the Polynomial Regression model exhibited mixed results, showing improvement for Rice but not significantly enhancing predictive accuracy for Potato and Wheat compared to the MLR model. The inclusion of non-linear terms did not consistently improve the model's performance across all crops.

In summary, the Random Forest Regressor emerged as the most effective model for predicting crop yields in Central and Southern Asia, providing valuable insights for policymakers and farmers to proactively manage resources and ensure food security. The study contributes to the broader goal of achieving the UN's Zero Hunger initiative and lays the groundwork for a more food-secure future in the region.

References

https://sdgs.un.org/goals
https://www.worldbank.org/en/home
https://www.fao.org/faostat/en/

Related Work:

-Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003

-Bowman, M. S., & Zilberman, D. (2013). Economic factors affecting diversified farming systems. Ecology and society, 18(1).

-Nigam, A., Garg, S., Agrawal, A., & Agrawal, P. (2019, November). Crop yield prediction using machine learning algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 125-130). IEEE

-Everingham, Y., Sexton, J., Skocaj, D., & Inman-Bamber, G. (2016). Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for sustainable development, 36(2), 27