EXPLORATORY DATA ANALYSIS OF CROPS GROWN IN INDIA

India ranks second worldwide in farm outputs. As per 2018, agriculture employed more than 50% of the Indian work force and contributed 17–18% to country's GDP.

In [1]:
```
#importing the libraries
import pandas as pd # Data processing, CSV file I/O
import numpy as np #Linear Algebra
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Importing the dataset

In [2]:
```
# Read the data
crop = pd.read_csv('crop1.csv')
```

In [3]:
```
#check first five rows
crop.head()
```

Out[3]:

	Area	Item	Element	Year	Unit	Value
0	Afghanistan	Almonds, with shell	Area harvested	1975	ha	0.0
1	Afghanistan	Almonds, with shell	Area harvested	1976	ha	5900.0
2	Afghanistan	Almonds, with shell	Area harvested	1977	ha	6000.0
3	Afghanistan	Almonds, with shell	Area harvested	1978	ha	6000.0
4	Afghanistan	Almonds, with shell	Area harvested	1979	ha	6000.0

In [4]:
```
#check dataset shape
crop.shape
```

Out[4]:
```
(1895975, 6)
```

In [5]:
```
crop.isna().sum() * 100/crop.shape[0]
```

Out[5]:
```
Area       0.000000
Item       0.000000
Element    0.000000
Year       0.000000
Unit       0.000000
Value      6.830259
dtype: float64
```

In [6]:
```
#check basic info
crop.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1895975 entries, 0 to 1895974
Data columns (total 6 columns):
 #   Column   Dtype  
---  ------   -----  
 0   Area     object 
 1   Item     object 
 2   Element  object 
 3   Year     int64  
 4   Unit     object 
 5   Value    float64
dtypes: float64(1), int64(1), object(4)
memory usage: 86.8+ MB
```

In [7]:
```
#check unique value
crop['Element'].unique()
```

Out[7]:
```
array(['Area harvested', 'Yield', 'Production'], dtype=object)
```

'Area harvested' - Area harvested refers to the total amount of land that is used to produce a particular crop or group of crops.

'Yield' - Yield refers to the amount of a particular product or resource that is produced per unit of land, labor, or capital.

'Production' - Production refers to the total amount of a particular product or resource that is produced.

In [8]:
```
crop.Unit.unique()
```

Out[8]:
```
array(['ha', 'hg/ha', 'tonnes'], dtype=object)
```

ha - A hectare is equal to 10,000 square metres or 2.471 acres land.
hg/ha - hectogram per hectare (Hg/Ha). One hectogram is equal to 100 grams.
tonnes - A tonne, or ton is equal to 1,000 kilograms.

In [9]:
```
crop.nunique()
```

Out[9]:
```
Area          245
Item          118
Element         3
Year           60
Unit            3
Value      420009
dtype: int64
```

In [10]:
```
india = crop[crop['Area'] == 'India']
```

In [11]:
```
india.isna().sum()
```

Out[11]:
```
Area         0
Item         0
Element      0
Year         0
Unit         0
Value      122
dtype: int64
```

Crop Distribution in India in 2020

In [12]:
```
india
```

Out[12]:

	Area	Item	Element	Year	Unit	Value
571594	India	Anise, badian, fennel, coriander	Area harvested	1961	ha	90000.0
571595	India	Anise, badian, fennel, coriander	Area harvested	1962	ha	90000.0
571596	India	Anise, badian, fennel, coriander	Area harvested	1963	ha	100000.0
571597	India	Anise, badian, fennel, coriander	Area harvested	1964	ha	100000.0
571598	India	Anise, badian, fennel, coriander	Area harvested	1965	ha	100000.0
...	...	...	...	...	...	...
585122	India	Wheat	Production	2016	tonnes	92290000.0
585123	India	Wheat	Production	2017	tonnes	98510220.0
585124	India	Wheat	Production	2018	tonnes	99869520.0
585125	India	Wheat	Production	2019	tonnes	103596230.0
585126	India	Wheat	Production	2020	tonnes	107590000.0

13533 rows × 6 columns

In [13]:
```
india = india.dropna()
```

In [14]:
```
harvest_area = india[(india['Element'] == 'Area harvested') & (indi
               by = 'Value', ascending = False)
```

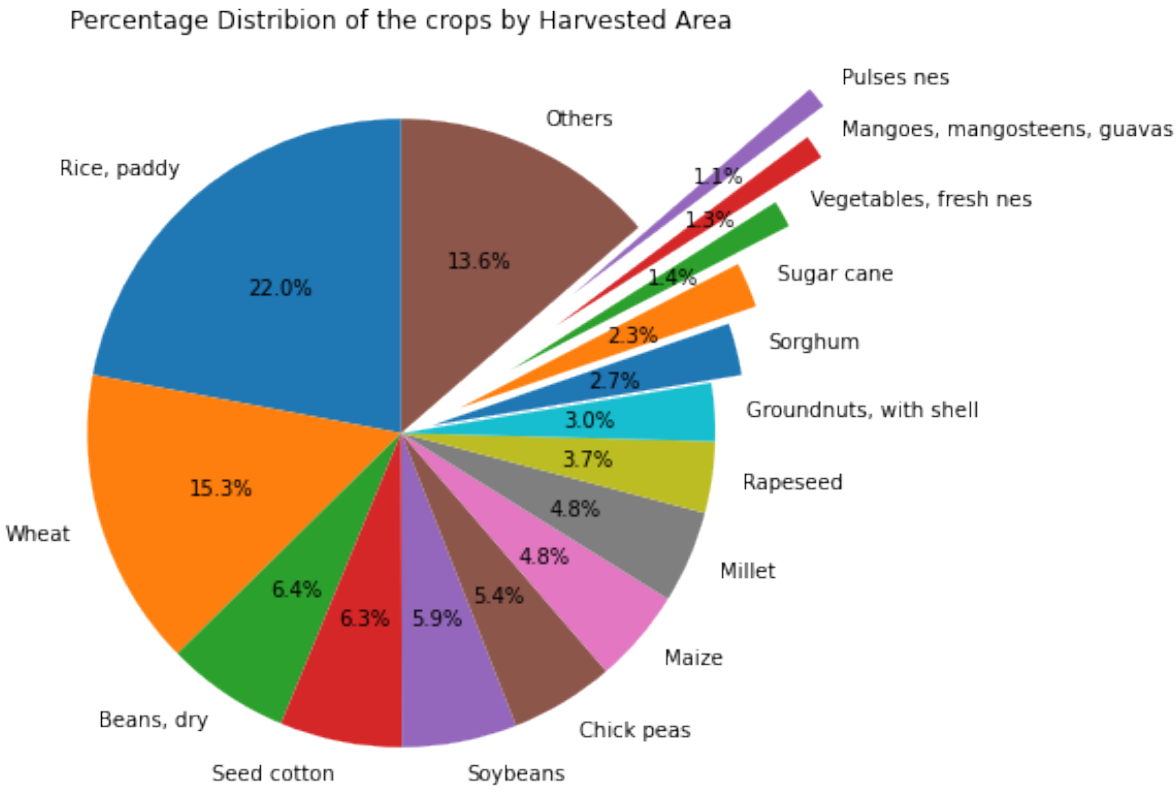
In [15]:
```
top_15 = harvest_area.iloc[:15]
others = pd.DataFrame({'Item':['Others'],'Value':[harvest_area.iloc
Top_15 = pd.concat([top_15, others], ignore_index = True)
```

In [16]:
```
Top_15
```

Out[16]:

	Item	Value
0	Rice, paddy	45000000.0
1	Wheat	31357000.0
2	Beans, dry	13006503.0
3	Seed cotton	12864576.0
4	Soybeans	12100000.0
5	Chick peas	10948882.0
6	Maize	9865000.0
7	Millet	9714019.0
8	Rapeseed	7500000.0
9	Groundnuts, with shell	6100000.0
10	Sorghum	5503062.0
11	Sugar cane	4790094.0
12	Vegetables, fresh nes	2847604.0
13	Mangoes, mangosteens, guavas	2578000.0
14	Pulses nes	2245667.0
15	Others	27873025.0

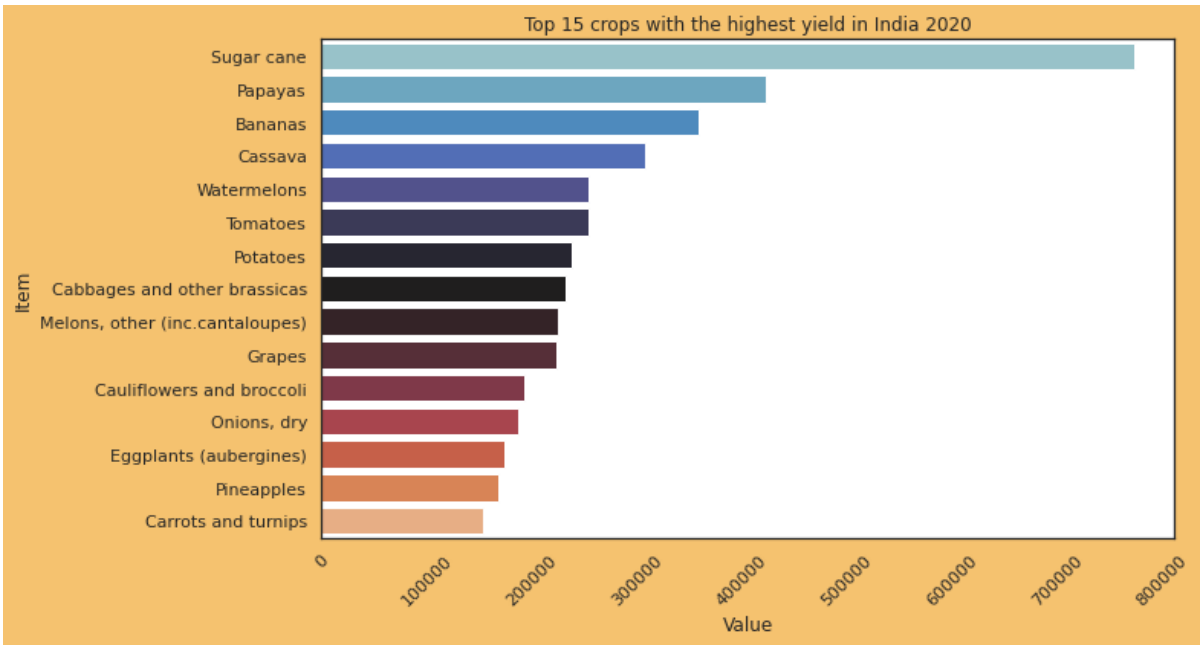
In [17]:
```
fig = plt.figure(figsize =(10, 7))
explode = [0, 0, 0, 0,0,0,0,0,0,0,0.1,0.2,0.4,0.6,0.7, 0]
plt.pie(Top_15['Value'], labels = Top_15['Item'],autopct='%1.1f%%',

plt.title('Percentage Distribion of the crops by Harvested Area')

plt.show()
```



Rice is cultivated in **22.0 %**, Wheat in **15.3%**, Beans,dry in **6.4 %** of the total cultivated area in India in the year 2020.

In [18]:
```
yield_india = india[(india['Element'] == 'Yield') & (india['Year']
                    by = 'Value', ascending = False)
top_15 = yield_india.iloc[:15]
```

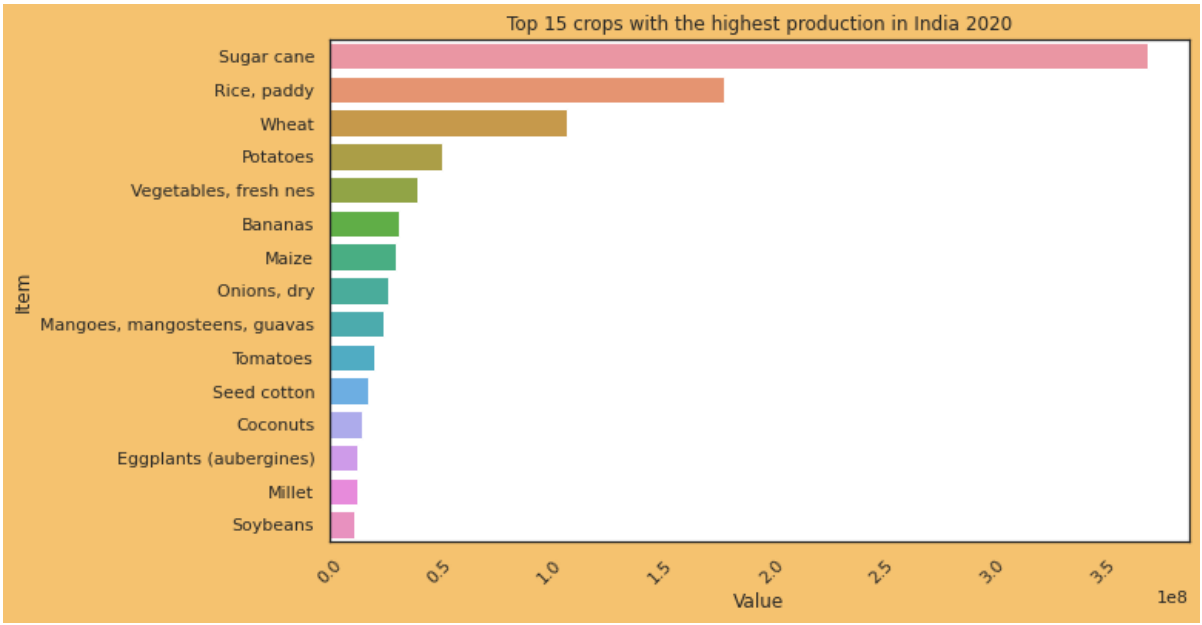
In [19]:
```
sns.set_theme(style = 'white',rc={'axes.facecolor':'white', 'figure
plt.figure(figsize = (10,6))
sns.barplot(data = top_15, y='Item', x = 'Value', palette = 'icefir
plt.title('Top 15 crops with the highest yield in India 2020')
plt.xticks(rotation = 45)
plt.show()
```



Sugarcane, Papayas, Bananas gave the highest yield in India in 2020.

In [20]:
```
production_india = india[(india['Element'] == 'Production') & (indi
                by = 'Value', ascending = False)
top_15 = production_india.iloc[:15]
```

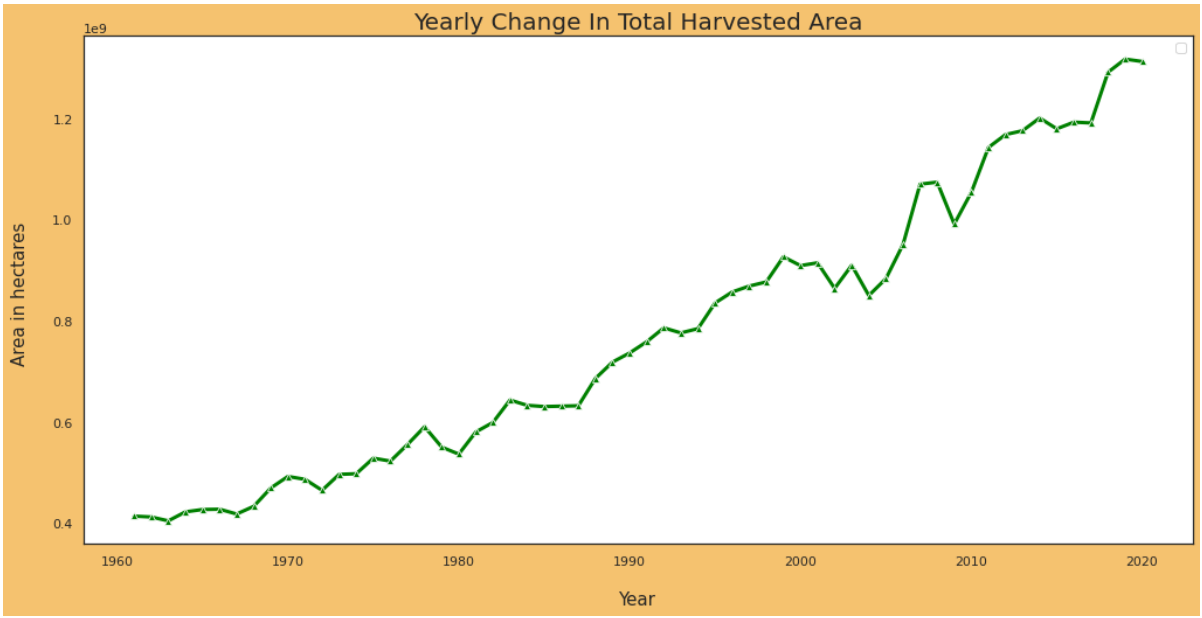
In [21]:
```
plt.figure(figsize = (10,6))
sns.barplot(data = top_15, y='Item', x = 'Value')
plt.title('Top 15 crops with the highest production in India 2020')
plt.xticks(rotation = 45)
plt.show()
```



Sugarcane, Rice, Wheat gave the highest production in India in 2020

Yearly Change In Total Harvested Area

In [22]:
```
total_area = india[(india['Element']=='Area harvested')][['Item','Y
total_area = india.groupby('Year',as_index=False)['Value'].sum()

plt.figure(figsize = (17,8))
sns.lineplot(data = total_area, x = 'Year', y = 'Value', marker = '
plt.title("Yearly Change In Total Harvested Area", size = 20)
plt.ylabel("Area in hectares ", size=15 ,labelpad=20)
plt.xlabel("Year", size=15, labelpad=20)
plt.legend()
plt.show()
```



The total area under cultivation has almost grown over 3 times from 1960 to 2020.

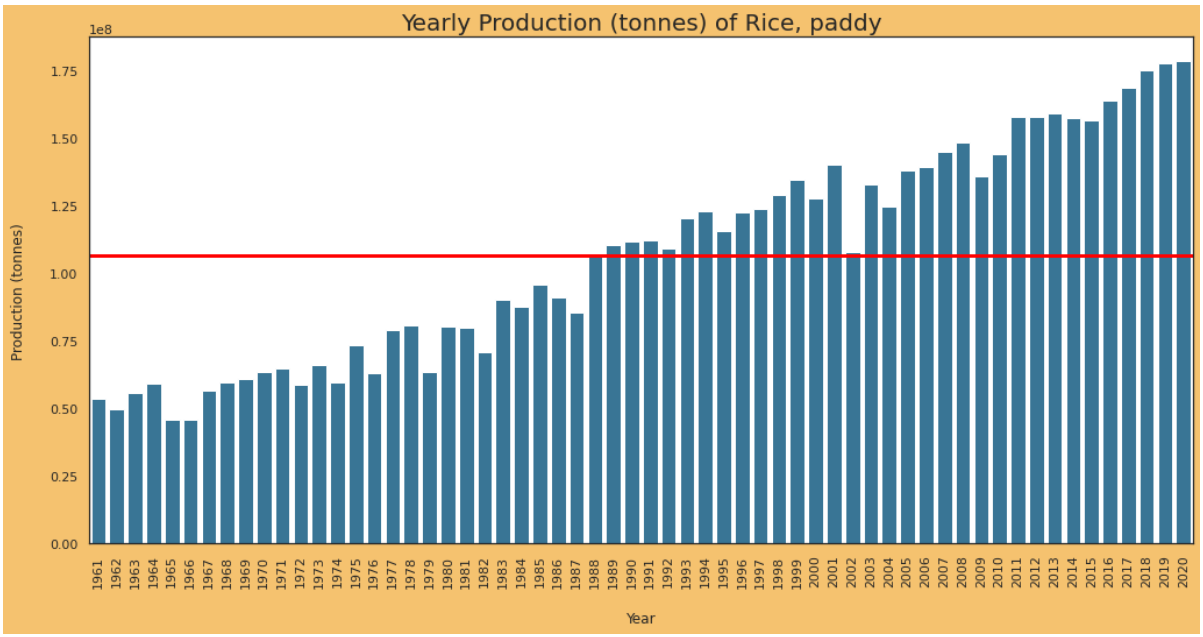
In [23]:
```
def production(cropname):
    if cropname not in india['Item'].unique():
        print('Item not in dataset')
    else:
        crop_prod = india[(india['Element'] == 'Production') & (ind
        plt.figure(figsize = (17,8))
        sns.barplot(data = crop_prod, x = 'Year', y = 'Value', colo
        plt.axhline(crop_prod.Value.mean(), linestyle = 'solid', li
        plt.title("Yearly Production (tonnes) of {}".format(cropnam
        plt.ylabel("Production (tonnes)", size=12 ,labelpad=20)
        plt.xlabel("Year", size=12, labelpad=20)
        plt.xticks(rotation=90)
        plt.show()
```

In [24]:
```
def yield_crop(cropname):
    if cropname not in india['Item'].unique():
        print('Item not in dataset')
    else:
        crop_yield = crop[((crop['Area'] == 'India') | (crop['Area'
                          &(crop['Item'] == cropname) & (crop['Elem
        crop_yield = crop_yield.dropna()
        plt.figure(figsize = (17,8))
        sns.lineplot(data = crop_yield, x = 'Year', y = 'Value', hu
            'red', 'orange','blue'))
        plt.title('Yield of ' + f'{cropname}'+ ' in hectogram per h
        plt.xlabel('Year', size = 15)
        plt.ylabel("Yield('Hg/ha')")
        plt.show()
```

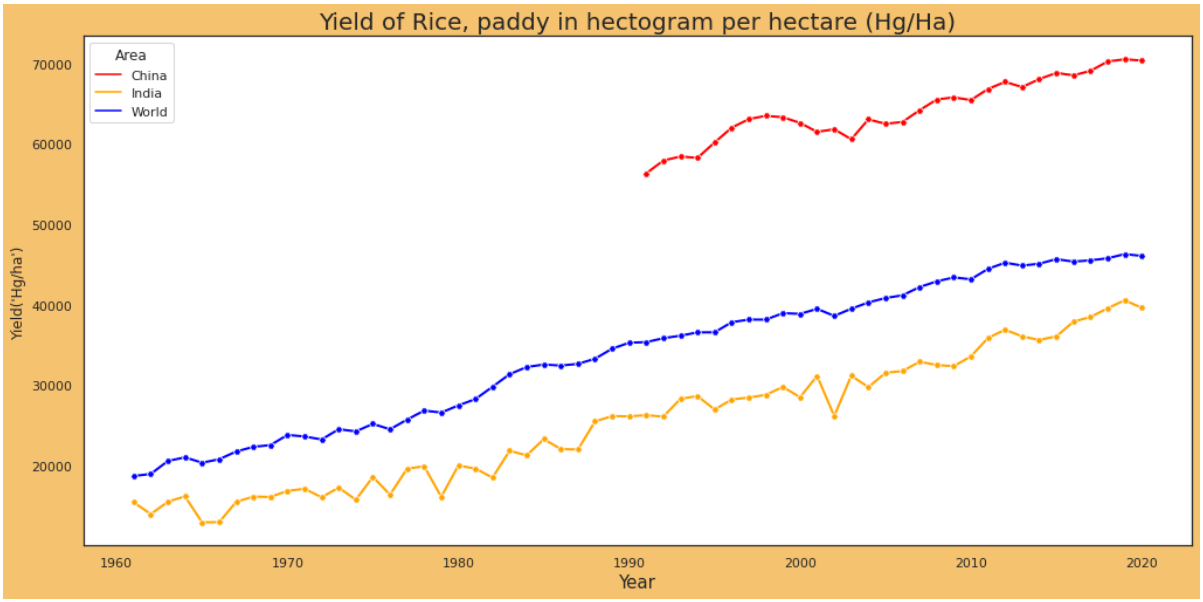
Rice,Wheat, Sugarcane, Papayas, Bananas came in the top 3 in yield and production segments. Let's see how the respective crops varied over years.

Rice

In [25]:
```
production('Rice, paddy')
```



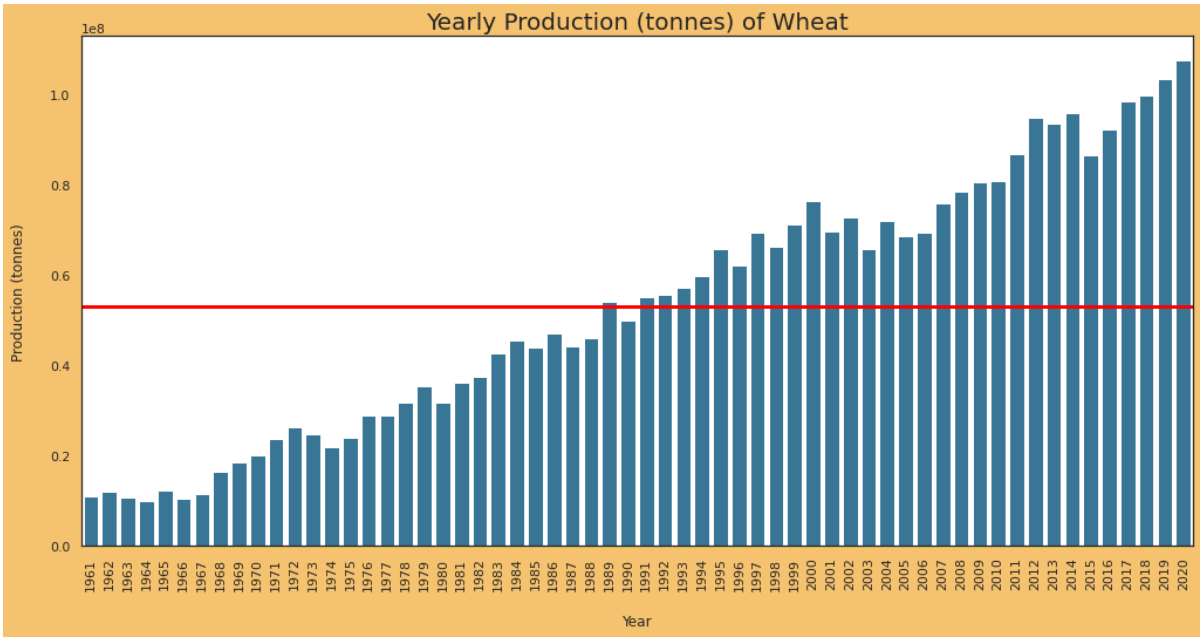
In [26]:
```
yield_crop('Rice, paddy')
```



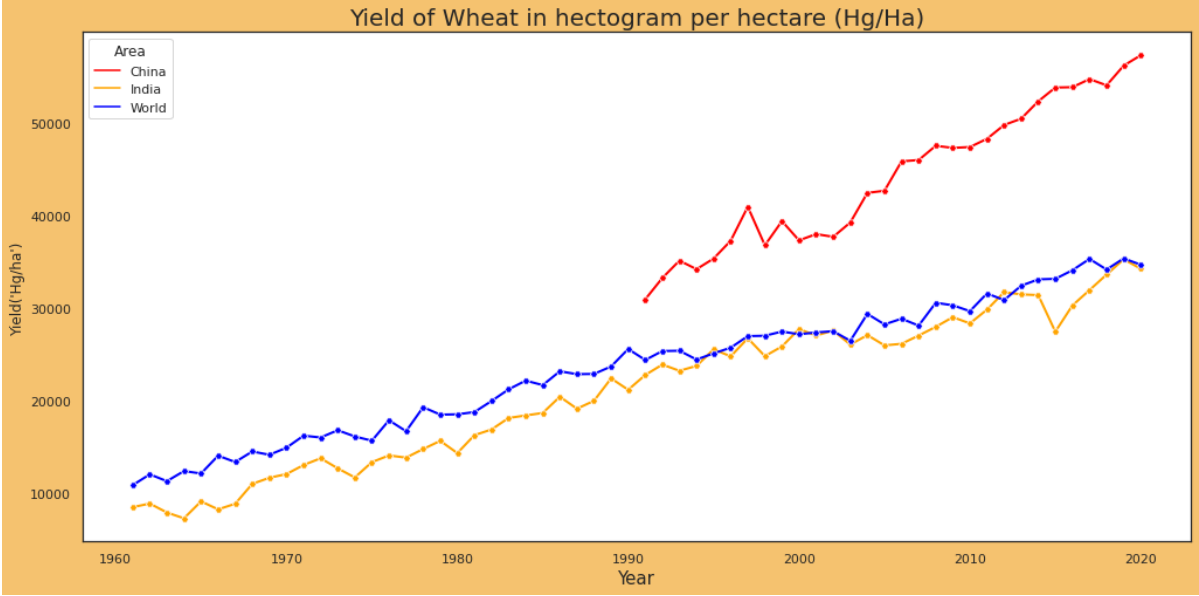
The yield of Rice in China is almost 60 percent more than India. India get to improve its productivity in producing rice.India is lagging behind in world average yield in rice too.

Wheat

In [27]:
```
production('Wheat')
```



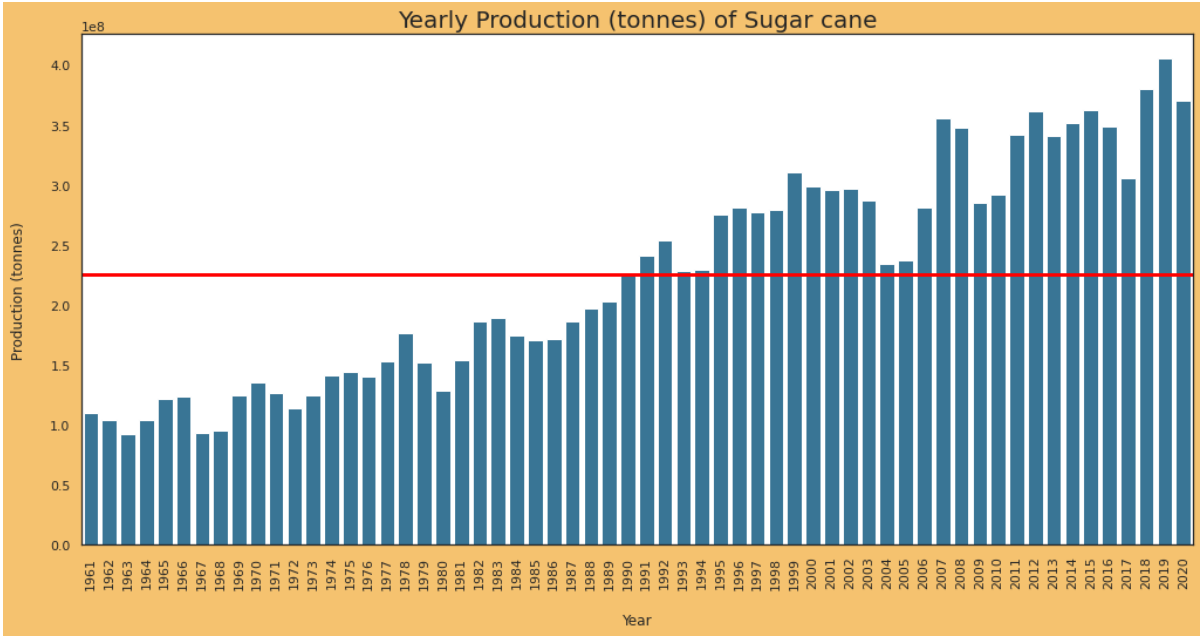
In [28]:
```
yield_crop('Wheat')
```



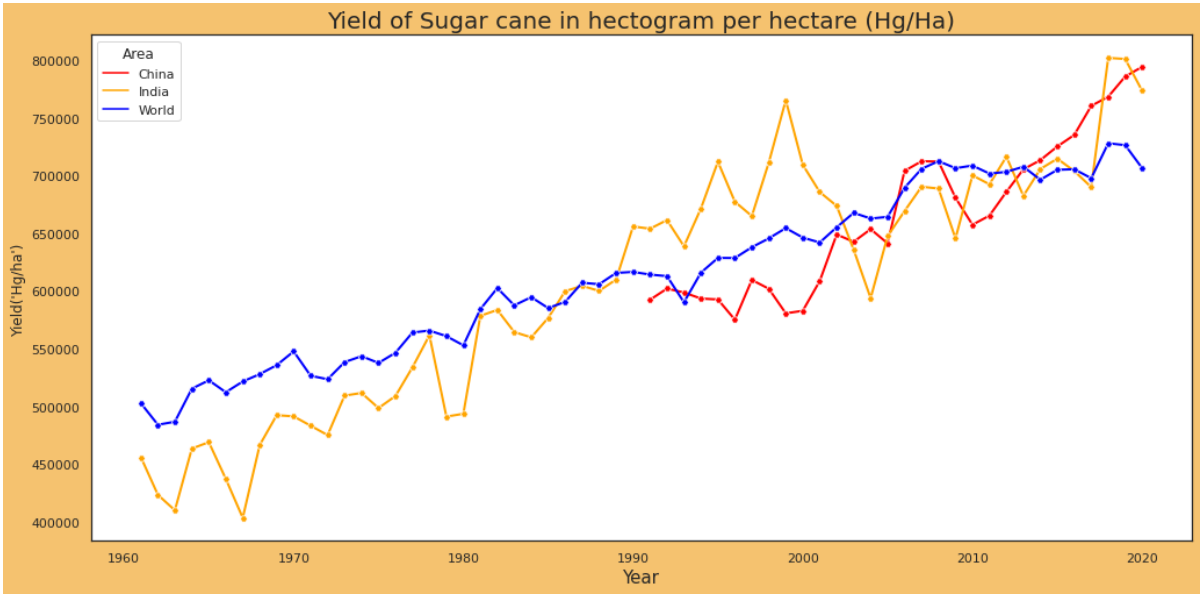
India has improved its yield in wheat with respect to world. But still it's lagging behind China.

Sugar Cane

In [29]:
```
production('Sugar cane')
```



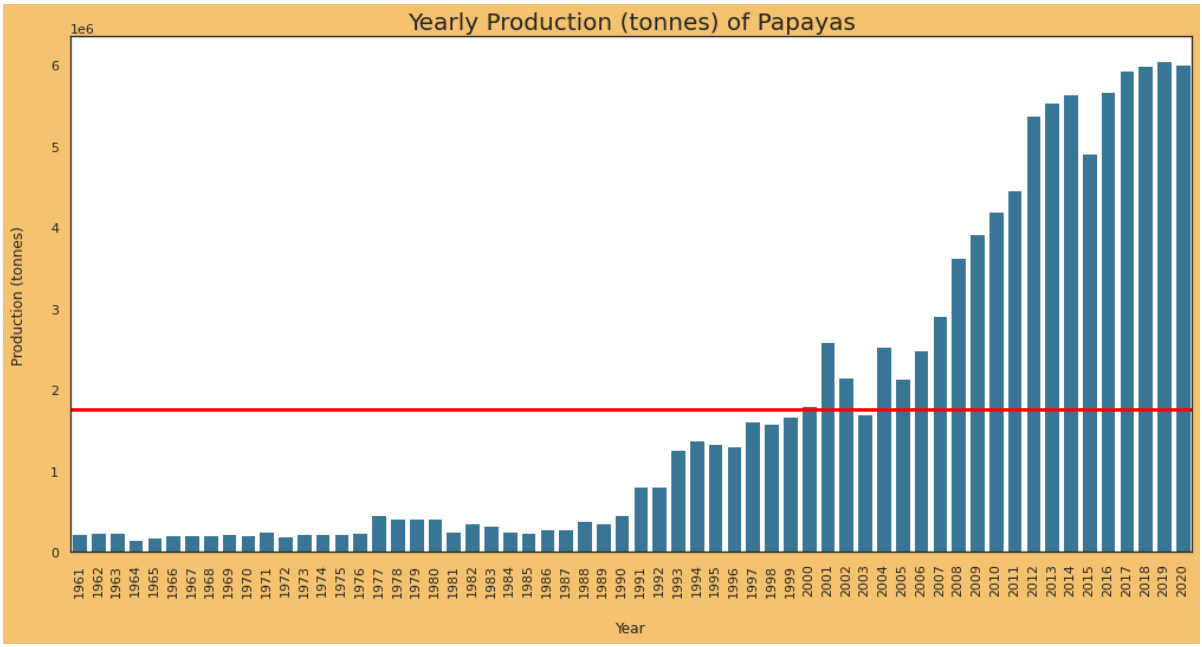
In [30]:
```
yield_crop('Sugar cane')
```



Sugarcane yield in India rised from 1960 to 2000. But after that it's stagnated and dipped till 2017. But it regained it's supremacy in 2018. China has catched up with sugarcane yield of India and even crossed it in 2020

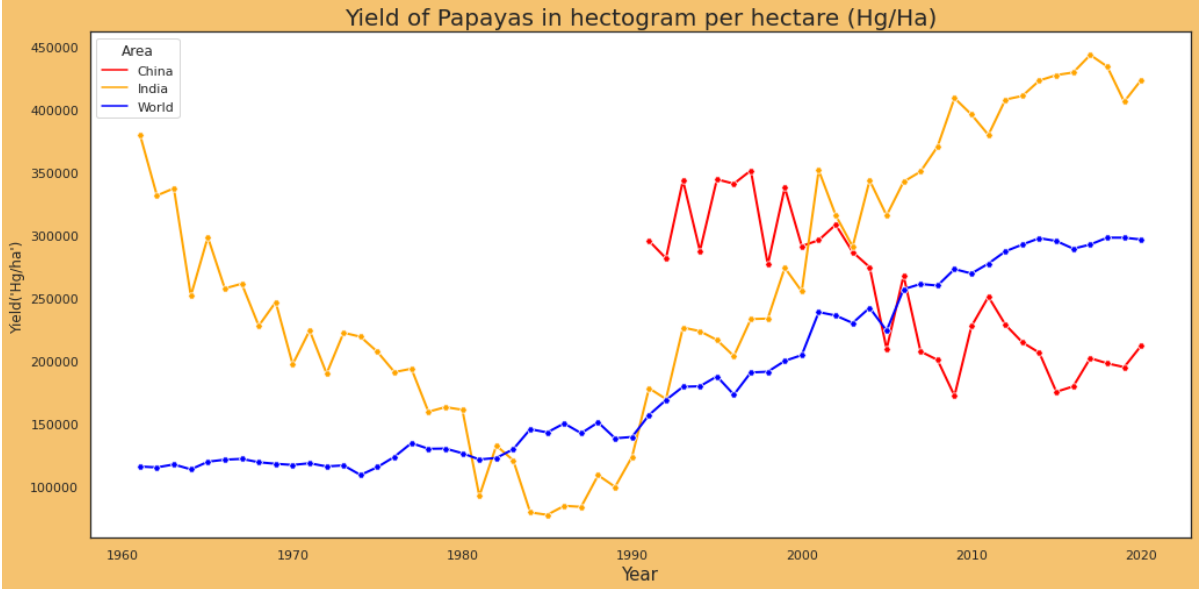
Papayas

In [31]:
```
production('Papayas')
```



Papayas production has remained stagant till 1990 on India. From 1990 the production picked up and production increases rapidly from then onwards.

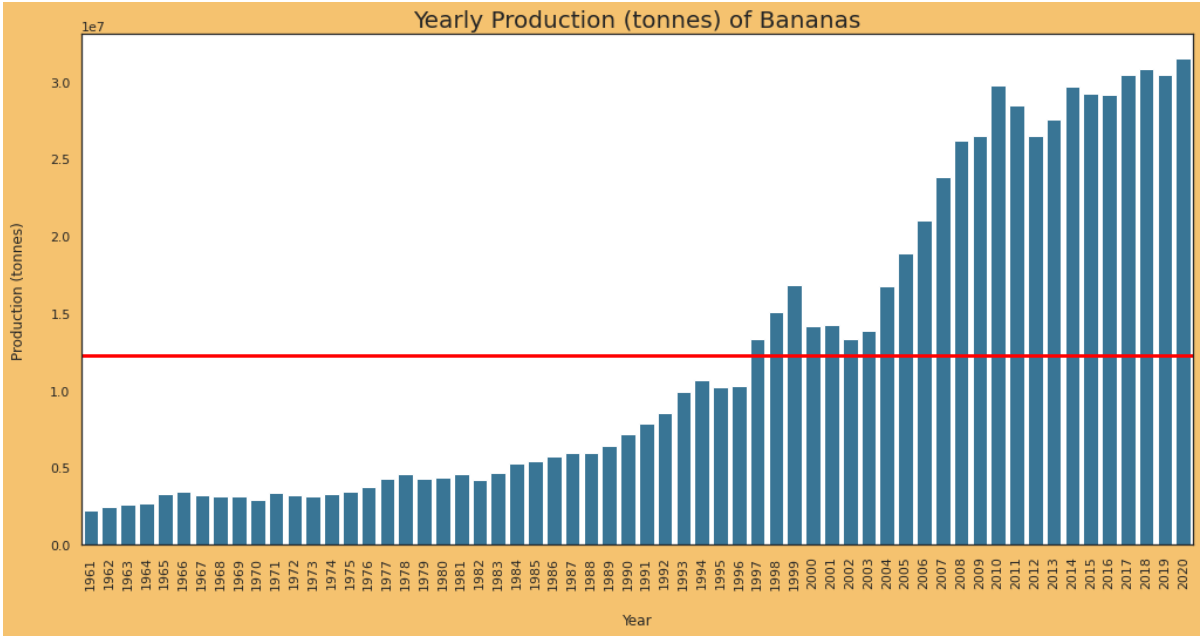
In [32]:
```
yield_crop('Papayas')
```



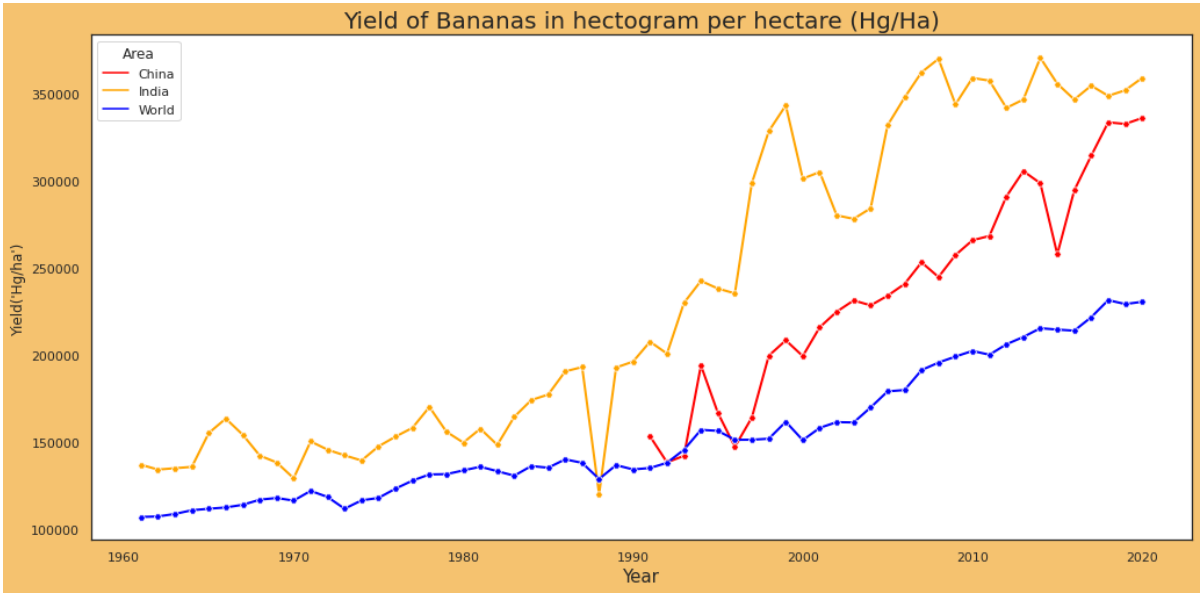
India's Papaya's yield had nosedived from 1960 to 1990. From 1990 it started rising. Now it's yield is higher than China and World. China Papaya's yield is reducing.

Bananas

In [33]:
```
production('Bananas')
```



In [34]:
```
yield_crop('Bananas')
```



India and China are competing in the Banana yield. But India retains it's top position.

In [ ]:
```

```