Answer Sheet - Advanced Python (Major)

Name: Vinod Kumar

Phone No.: +91-7983059757

Email: vinodkmr51@gmail.com

Topic: To visualize how honey production has changed over the years **(1998-2021)** in the **United States.**

Using the **Honey Production In USA Dataset** to solve the following questions.

- Before solving the questions, let us import the necessary modules/libraries in the Jupyter notebook and perform pre-processing steps.

```
import numpy as np
# numpy is aliased as np
import pandas as pd
# pandas is aliased as pd
import seaborn as sns
# seaborn is aliased as sns
import matplotlib.pyplot as plt
# matplotlib.pyplot is aliased as plt
import warnings
warnings.filterwarnings('ignore')
```

- **Storing the csv file in USA_df variable**

```
USA_df = pd.read_csv('honeyproduction 1998-2021.csv')
USA_df.head()
```

Input Screenshot:

```
# Storing the csv file in USA_df variable
USA_df = pd.read_csv('honeyproduction 1998-2021.csv')
USA_df.head()
```

✓ 0.1s

Output Screenshot:

	State	numcol	yieldpercol	totalprod	stocks	priceperlb	prodvalue	year
0	Alabama	16000.0	71	1136000.0	159000.0	0.72	818000.0	1998
1	Arizona	55000.0	60	3300000.0	1485000.0	0.64	2112000.0	1998
2	Arkansas	53000.0	65	3445000.0	1688000.0	0.59	2033000.0	1998
3	California	450000.0	83	37350000.0	12326000.0	0.62	23157000.0	1998
4	Colorado	27000.0	72	1944000.0	1594000.0	0.70	1361000.0	1998

```
# Pre- processing the dataframe
```

- **Checking the shape of the dataframe**

```
USA_df.shape
```

Input Screenshot:

```
# Pre- processing the dataframe
# Checking the shape of the dataframe
USA_df.shape
```

Output Screenshot:

```
(985, 8)
```

(985,8)- Means that the data has 986 Rows and 8 columns

- **Checking the columns of the dataframe**

```
USA_df.columns
```

Input screenshot:

```
USA_df.columns
```

Output Screenshot:

```
Index(['State', 'numcol', 'yieldpercol', 'totalprod', 'stocks', 'priceperlb',
       'prodvalue', 'year'],
      dtype='object')
```

- **Checking the data types of the columns**

```
USA_df.dtypes
```

Input Screenshot:

```
USA_df.dtypes
✓ 0.0s
```

Output Screenshot:

```
State          object
numcol        float64
yieldpercol     int64
totalprod     float64
stocks        float64
priceperlb    float64
prodvalue     float64
year            int32
dtype: object
```

- **Checking for the null values in the dataframe.**

```
USA_df.isnull().sum()
```

Input screenshot:

```
USA_df.isnull().sum()
✓ 0.0s
```

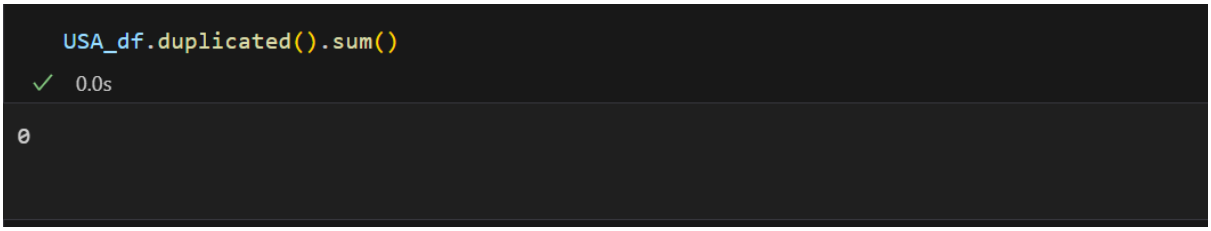
Output screenshot:

```
State          0
numcol         0
yieldpercol    0
totalprod      0
stocks         0
priceperlb     0
prodvalue      0
year           0
dtype: int64
```

- **Checking for the duplicated values**

```
USA_df.duplicated().value_counts()
```

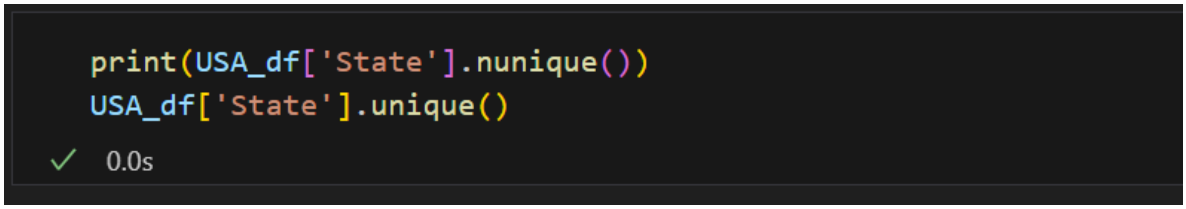
Output Screenshot:



- **Checking for the unique values of State column and their names:**

```
print(USA_df['State'].nunique())
USA_df['State'].unique()
```

Input Screenshot:



Output Screenshot:

- **Adding a new column namely: totalprod, which is the product of numcol and yieldpercol**

```
USA_df['totalprod'] = USA_df['numcol'] *
USA_df['yieldpercol']
USA_df['totalprod'].head()
```

Input Screenshot

```
USA_df['totalprod'] = USA_df['numcol'] * USA_df['yieldpercol']
USA_df['totalprod'].head()
```

✓ 0.0s

Output Screenshot:

```
0     1136000.0
1     3300000.0
2     3445000.0
3    37350000.0
4     1944000.0
Name: totalprod, dtype: float64
```

- Since, we have done the pre-processing of the dataset, we can continue with our questions.

Question 1) How has honey production **yield changed** from **1998 to 2021**?

Solution)

```
q1=USA_df.groupby(['year'])['yieldpercol'].agg(
{'sum','mean'})
q1.head()
```

Input Screenshot:

```
q1 = USA_df.groupby(['year'])['yieldpercol'].agg({'sum','mean'})
q1.head()
```

✓ 0.0s

Output Screenshot:

year	mean	sum
1998	69.953488	3008
1999	65.465116	2815
2000	67.581395	2906
2001	64.545455	2840
2002	66.795455	2939

Grouping the dataframe with respect to year and then finding the sum and average of the yield per colony to plot bar and line chart using the subplots.

```
fig,(ax1,ax2) =
plt.subplots(1,2,figsize=(10,4))

ax1.bar(q1.index,q1['sum'])
ax1.set_title('Year wise total of yield')

ax2.plot(q1.index,q1['mean'])
ax2.set_title('Year wise average of yield per
colony', color = 'DarkBlue')
ax2.grid()

plt.savefig('Year wise total of Yield per
Colony.jpg')

plt.show()
```

Input Screenshot

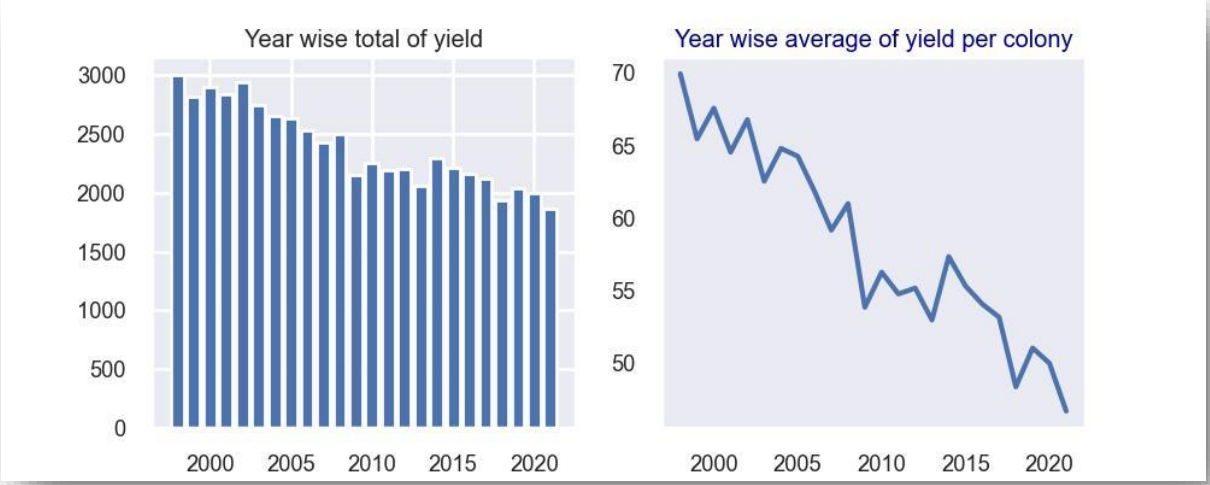
```
fig,(ax1,ax2) = plt.subplots(1,2,figsize=(10,4))

ax1.bar(q1.index,q1['sum'])
ax1.set_title('Year wise total of yield')

ax2.plot(q1.index,q1['mean'])
ax2.set_title('Year wise average of yield per colony', color = 'DarkBlue')
ax2.grid()

plt.savefig('Year wise total of Yield per Colony.jpg')
plt.show()
```

✓ 0.7s



- The output plot shows that the Total yield has decreased per year, and the average of total yield per colony is also decreasing as we move from year 1998 to 2021.

Question 2) Over time, what are the **major production** trends across the states?

Solution) To get the output as desired we must change the data type of 'year' column to category, so that we can plot using the catplot() function of Seaborn.

```
USA_df.year = USA_df.year.astype('category')
```

```
USA_df.year = USA_df.year.astype('category')
✓ 0.0s
```

```
sns.set_style('darkgrid')
sns.set_context("notebook", font_scale = 1,
rc={"line.linewidth": 0.3})

fig2 = sns.catplot(data = USA_df, x='year',
y='totalprod',col = 'State', kind =
'point',lw=1.5,height =3,col_wrap =4)

fig2.set_ylabels('Total production',
color='Red')

plt.savefig('Year wise total production per
state.jpg')
plt.show()
```

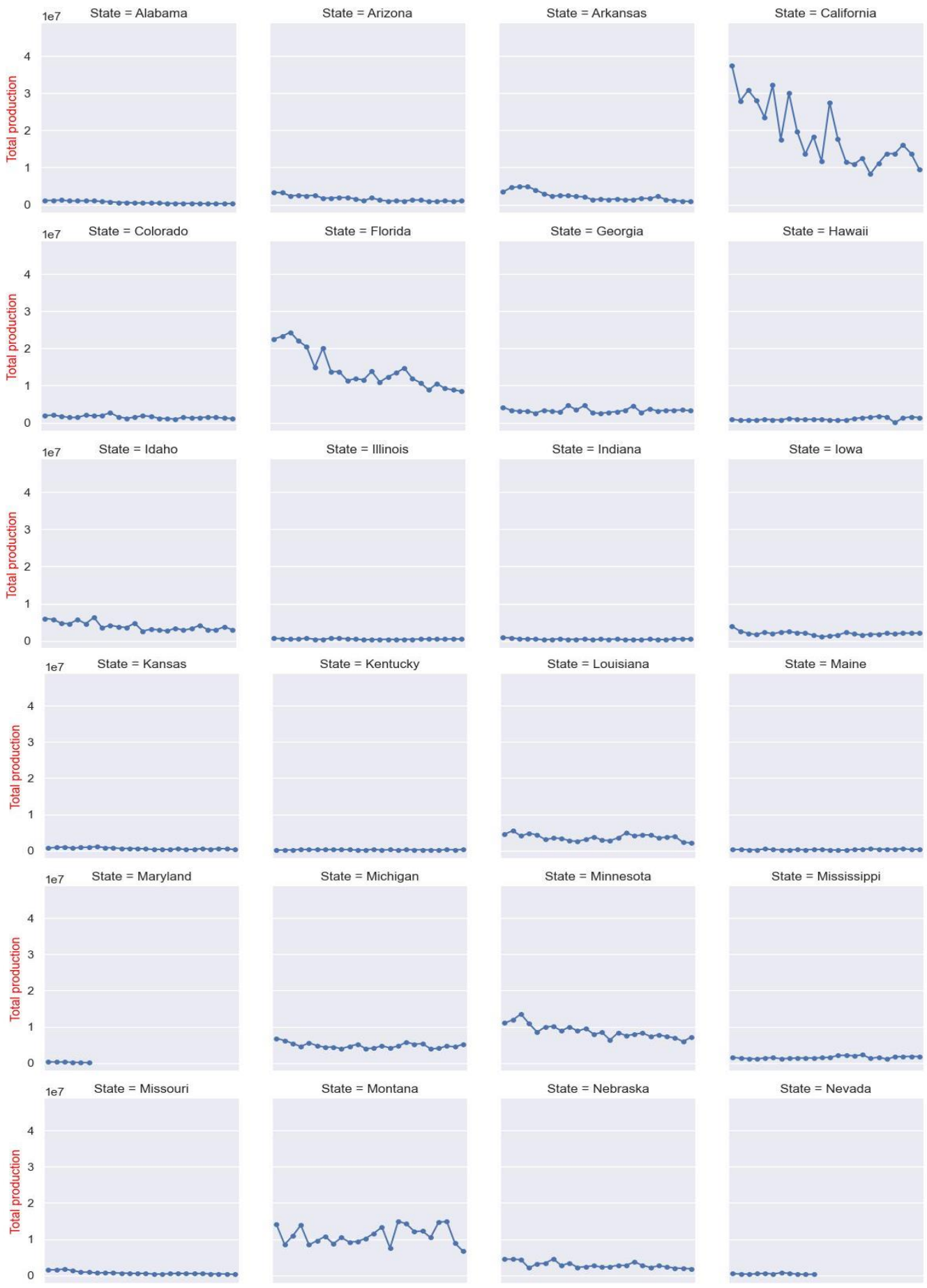
```
sns.set_style('darkgrid')
sns.set_context("notebook", font_scale = 1, rc={"line.linewidth": 0.3})

fig2 = sns.catplot(data = USA_df, x='year', y='totalprod',
                   col = 'State', kind = 'point',lw=1.5,
                   height =3,col_wrap =4)

fig2.set_ylabels('Total production', color='Red')

plt.savefig('Year wise total production per state.jpg')
plt.show()
✓ 17.6s
```

The output of the above code would give us a catplot, where year wise total production is plotted in X and Y axis, and each State is displayed in different plot.


State = Alabama
State = Arizona
State = Arkansas
State = California
State = Colorado
State = Florida
State = Georgia
State = Hawaii
State = Idaho
State = Illinois
State = Indiana
State = Iowa
State = Kansas
State = Kentucky
State = Louisiana
State = Maine
State = Maryland
State = Michigan
State = Minnesota
State = Mississippi
State = Missouri
State = Montana
State = Nebraska
State = Nevada
1e7
Total production
0
1
2
3
4


State = New Jersey
State = New Mexico
State = New York
State = North Carolina
State = North Dakota
State = Ohio
State = Oklahoma
State = Oregon
State = Pennsylvania
State = South Dakota
State = Tennessee
State = Texas
State = Utah
State = Vermont
State = Virginia
State = Washington
State = West Virginia
State = Wisconsin
State = Wyoming
State = South Carolina
1e7
Total production
0
1
2
3
4
year

Observations from the above plot:

- The most prominent honey producing states of US are - **California, Florida, North Dakota and South Dakota and Montana**
- Unfortunately, the honey production in **California** has seen a steep decline over the years.
- **Florida's** total production also has been on a decline.
- **South Dakota** has more of less maintained its levels of production. **North Dakota** has seen an impressive increase in the honey production.
- Remaining states of the US has maintained its levels of production.

Question 3) Does the data show any trends in terms of the number of **honey-producing colonies** and **yield per colony** before **2006**, which was when concern over Colony Collapse Disorder spread nationwide?

Solution) We must revert the data type of 'year' column back to 'int' for this question.

```
USA_df['year'] = USA_df['year'].astype(int)
```

Output Screenshot:

```
USA_df['year'] = USA_df['year'].astype(int)
USA_df.dtypes
✓ 0.0s
```

```
USA_df.dtypes
```

```
USA_df.dtypes
✓ 0.0s
```

Output Screenshot:

```
State           object
numcol         float64
yieldpercol      int64
totalprod      float64
stocks         float64
priceperlb     float64
prodvalue      float64
year             int64
dtype: object
```

```
q3 = USA_df[(USA_df['year']<=2006)]
q3.head()
```

```
q3 = USA_df[(USA_df['year']<=2006)]
q3.head()
```
✓ 0.0s

Output screenshot:

	State	numcol	yieldpercol	totalprod	stocks	priceperlb	prodvalue	year
0	Alabama	16000.0	71	1136000.0	159000.0	0.72	818000.0	1998
1	Arizona	55000.0	60	3300000.0	1485000.0	0.64	2112000.0	1998
2	Arkansas	53000.0	65	3445000.0	1688000.0	0.59	2033000.0	1998
3	California	450000.0	83	37350000.0	12326000.0	0.62	23157000.0	1998
4	Colorado	27000.0	72	1944000.0	1594000.0	0.70	1361000.0	1998

```
q3=q3.groupby('year')[['numcol','yieldpercol']]
.sum().reset_index()
q3.head(10)
```

Input Screenshot:

```
q3 = q3.groupby('year')[['numcol','yieldpercol']].sum().reset_index()
q3.head()
```
✓ 0.0s

Output Screenshot:

	year	numcol	yieldpercol
0	1998	2621000.0	3008
1	1999	2637000.0	2815
2	2000	2604000.0	2906
3	2001	2542000.0	2840
4	2002	2565000.0	2939

- Dividing the ‘numcol’ by 1000 gives a relatable value so that we can plot yield per colony and number of colonies in a single line plot.

```
q3['numcol'] = q3['numcol']/1000
q3.head()
```

Input screenshot:

```
q3['numcol'] = q3['numcol']/1000
q3.head()
```

✓ 0.0s

Output screenshot:

	year	numcol	yieldpercol
0	1998	2621.0	3008
1	1999	2637.0	2815
2	2000	2604.0	2906
3	2001	2542.0	2840
4	2002	2565.0	2939

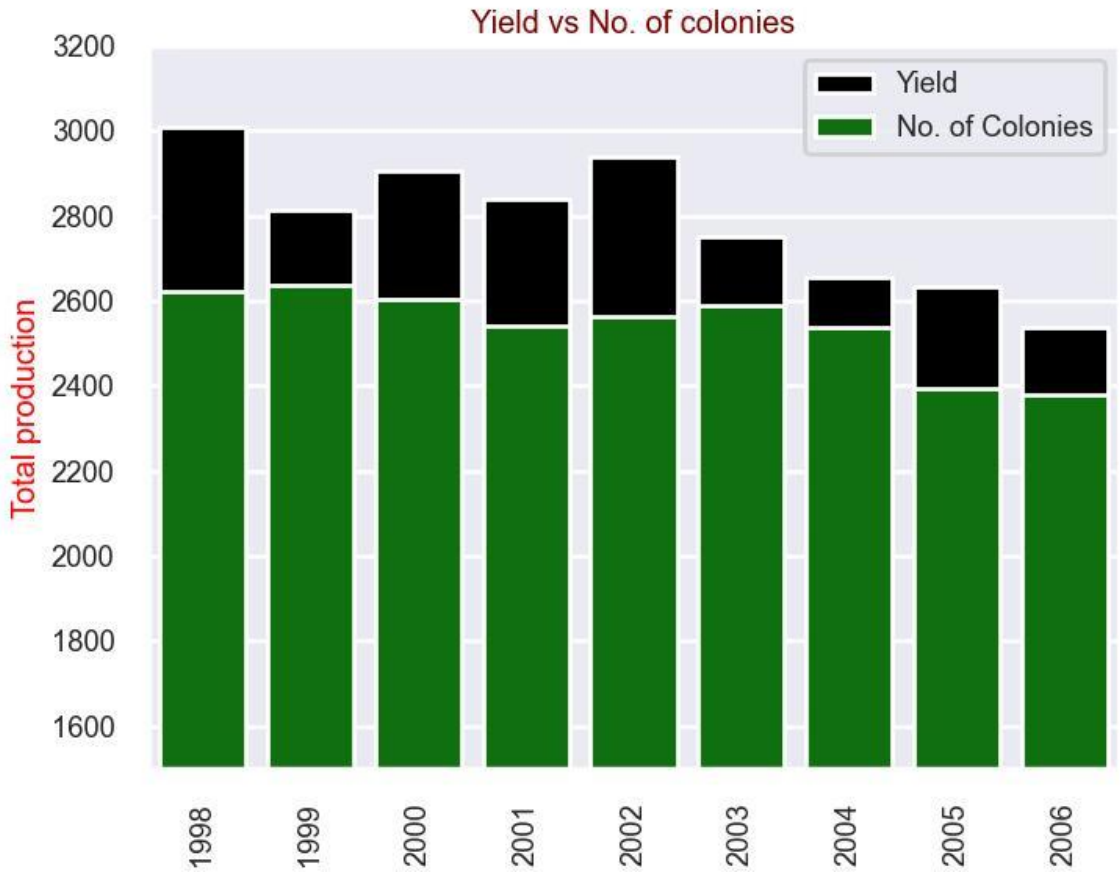
```
sns.set(rc={'figure.figsize':(8,6)})
sns.set_style('darkgrid')
sns.set_context("poster", font_scale = .6,
rc={"line.linewidth": 0.8})

sns.barplot(data =q3, x='year',
y='yieldpercol', color= 'black',label =
'Yield')
sns.barplot(data =q3, x='year',
y='numcol',color='green', label='No. of
Colonies')

plt.ylim(1500,3200)
plt.title('Yield vs No. of colonies', color =
'maroon')
plt.xlabel('Year', color='red')
plt.ylabel('Total production', color='Red')
plt.xticks(rotation = 90)
plt.savefig('Yield vs No. of colonies.jpg')
plt.show()
```

Input screenshot:

Output screenshot:



- Let us plot another line plot to see the relation between yield per colony and the number of honey producing colonies.

```
plt.figure(figsize=(8,6))
sns.lineplot(data=q3,x='year',y='yieldpercol',l
abel='Yield_per_Colony', color
='Blue',errorbar=None)
sns.lineplot(data=q3,x='year',y='numcol',label=
'No_of Colonies', color ='Red',errorbar=None)

plt.xlabel('States')
plt.xticks(rotation = 85)
```

```
plt.title(' No of colonies vs Yield per
colony')
plt.savefig('No of colonies vs Yield per
colony.jpg')
plt.show()
```

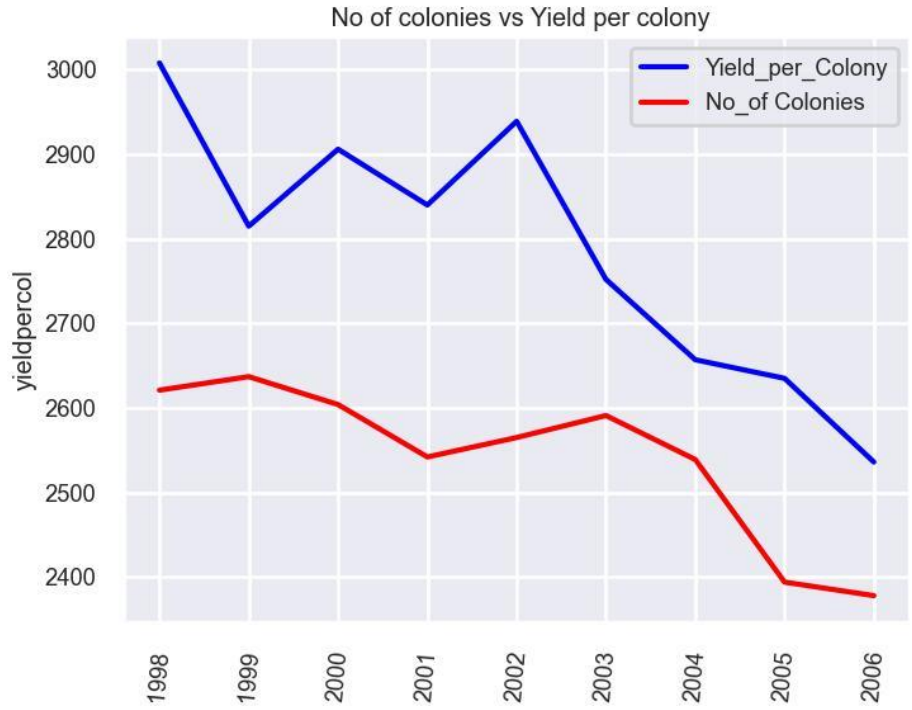
```
plt.figure(figsize=(8,6))

sns.lineplot(data=q3,x='year',y='yieldpercol',label='Yield_per_Colony',
             color ='Blue',errorbar=None)
sns.lineplot(data=q3,x='year',y='numcol',label='No_of Colonies',
             color ='Red',errorbar=None)

plt.xlabel('States')
plt.xticks(rotation = 85)
plt.title(' No of colonies vs Yield per colony')
plt.savefig('No of colonies vs Yield per colony.jpg')
plt.show()
```

✓ 0.5s

Output screenshot:



Conclusion:

- From the above two plots, we can conclude that the number of honey producing colonies are on a decline and therefore the total yield from the colonies is also declining.

Question 4) Are there any **patterns** that can be observed between **total honey production** and the **value of production** every year?

Solution) Plotting year wise sum of total production and production value.

```
q4=USA_df.groupby('year')[['totalprod','prodvalue']].sum().reset_index()
q4.head()
```

```
q4 = USA_df.groupby('year')[['totalprod','prodvalue']].sum().reset_index()
q4.head()
✓ 0.0s
```

Output screenshot:

	year	totalprod	prodvalue
0	1998	219519000.0	146091000.0
1	1999	202387000.0	123657000.0
2	2000	219558000.0	131568000.0
3	2001	185748000.0	132282000.0
4	2002	171265000.0	227302000.0

```
q4['totalprod'] = q4['totalprod']/100000
q4['prodvalue'] = q4['prodvalue']/100000

q4.head()
```

```
q4['totalprod'] = q4['totalprod']/100000
q4['prodvalue'] = q4['prodvalue']/100000
q4.head()
```

✓ 0.0s

	year	totalprod	prodvalue
0	1998	2195.19	1460.91
1	1999	2023.87	1236.57
2	2000	2195.58	1315.68
3	2001	1857.48	1322.82
4	2002	1712.65	2273.02

```
sns.set(rc={'figure.figsize':(8,6)})
sns.set(rc={'figure.figsize':(8,6)})
sns.set_style('darkgrid')
sns.set_context("poster", font_scale = .6,
rc={"line.linewidth": 0.8})
```

```
sns.lineplot(data =q4, x='year', y='totalprod',
color= 'green',label = 'Total Honey Prod')
sns.lineplot(data =q4, x='year',
y='prodvalue',color='red', label='Production
value')

plt.title('Total production Vs Total Production
Value', color = 'maroon')
plt.xlabel('Year', color='red')
plt.ylabel('Total production', color='Red')
plt.xticks(rotation = 90)

plt.savefig('Total production Vs Total
Production Value.jpg')

plt.show()
```

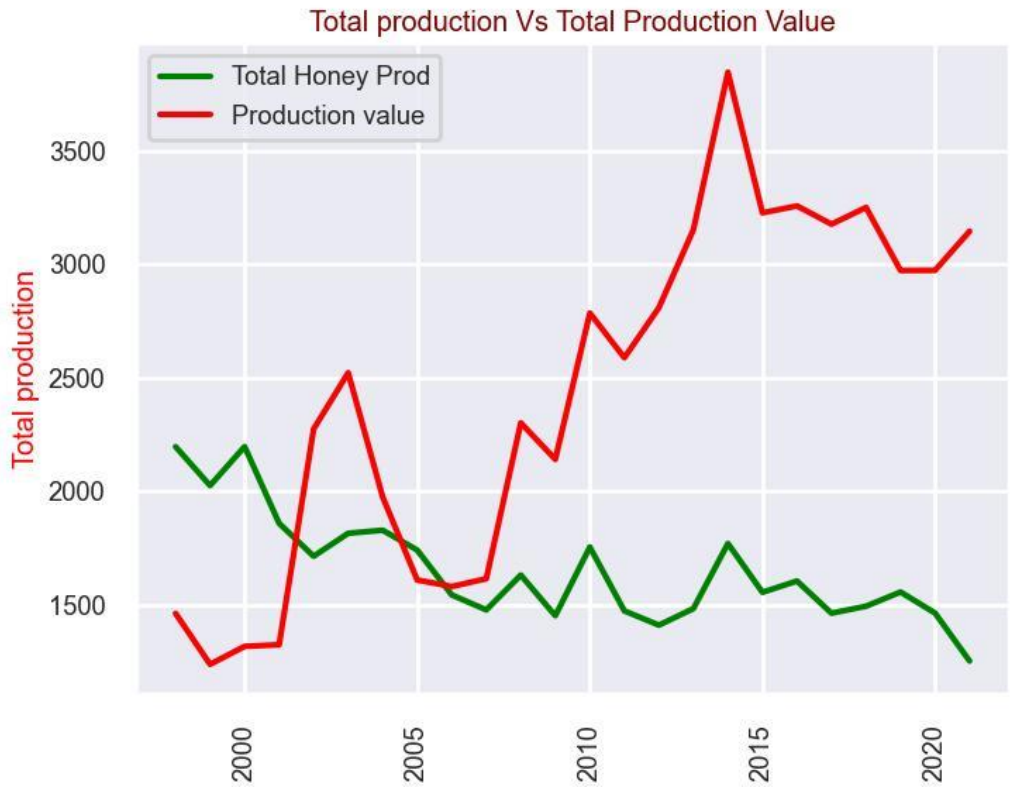
```
sns.set(rc={'figure.figsize':(8,6)})
sns.set_style('darkgrid')
sns.set_context("poster", font_scale = .6, rc={"line.linewidth": 0.8})

sns.lineplot(data =q4, x='year', y='totalprod', color= 'green',label = 'Total Honey Prod')
sns.lineplot(data =q4, x='year', y='prodvalue',color='red', label='Production value')

plt.title('Total production Vs Total Production Value', color = 'maroon')
plt.xlabel('Year', color='red')
plt.ylabel('Total production', color='Red')
plt.xticks(rotation = 90)

plt.savefig('Total production Vs Total Production Value.jpg')
plt.show()
```

0.4s



Conclusion:

- From the plot above, we can see that the Total production of honey is declining as we move from year 1998 to 2021 and on the other hand the value of production is increasing year by year.

Question 5) How has the **value of production**, which in some
sense could be tied to **demand**, changed every year?

Solution) To find the change in the value of production I have taken the year wise average of the production.

```
q5=USA_df.groupby('year')['prodvalue'].mean().reset_index()
```

q5.head()

```
q5 = USA_df.groupby('year')['prodvalue'].mean().reset_index()
q5.head()
```
✓ 0.0s

Output Screenshot:

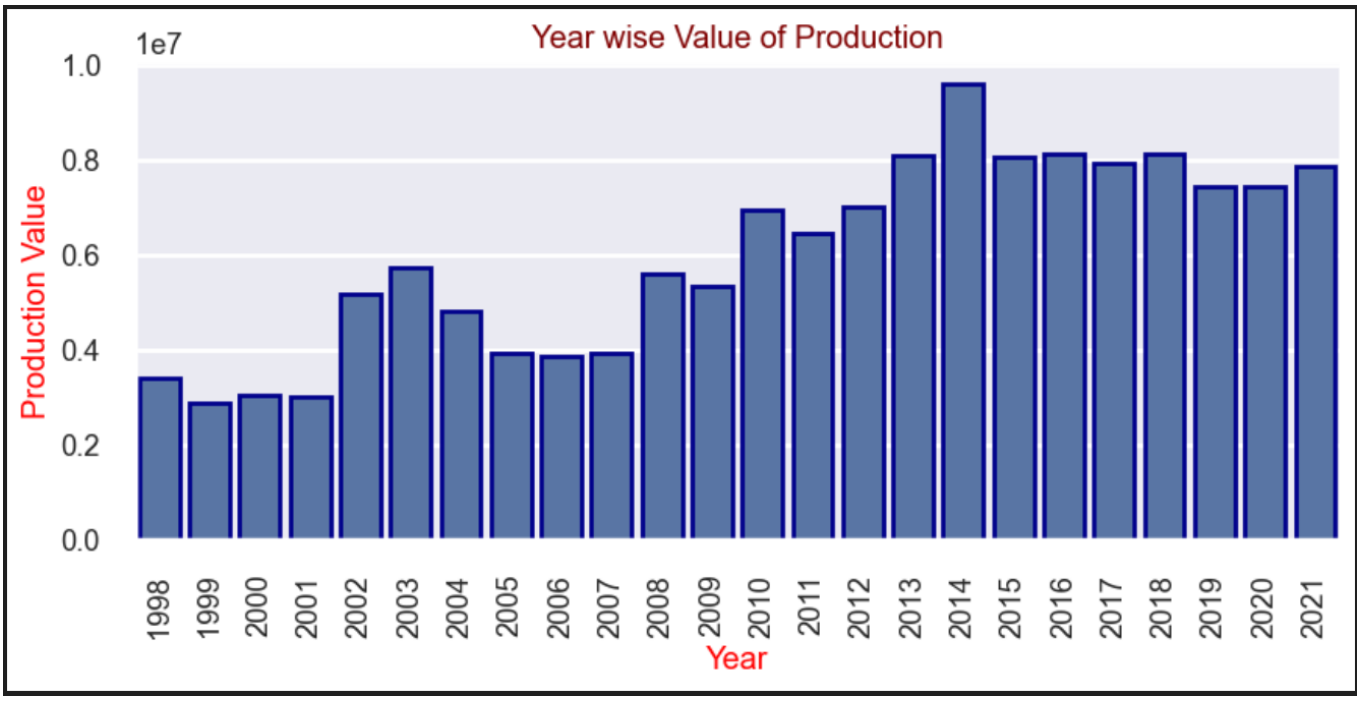
	year	prodvalue
0	1998	3.397465e+06
1	1999	2.875744e+06
2	2000	3.059721e+06
3	2001	3.006409e+06
4	2002	5.165955e+06

```
sns.set(rc={'figure.figsize':(10,4)})
sns.set_style('darkgrid')
sns.set_context("poster", font_scale = .6,
rc={"line.linewidth": 0.8})

sns.barplot(data =q5, x='year',
y='prodvalue',edgecolor='darkblue')

plt.title('Year wise Value of Production',
color = 'maroon')
plt.xlabel('Year', color='red')
plt.ylabel('Production Value', color='Red')
plt.xticks(rotation = 90)

plt.savefig('Year wise Value of
Production.jpg')
plt.show()
```



Conclusion:

- From the above bar plot, we can see that the average value of production is increasing year by year.

Overall Conclusion:

The Global concern which was raised in 2006 over the rapid decline in the honeybee population seems to be valid, as we have seen from the above plots, we can conclude that the demand and value of honey production has increased but the number of honey producing colonies and yield is decreasing.