In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.simplefilter(action='ignore')
```

In [2]:

```
df=pd.read_csv(r'C:\Users\hp\Downloads\honeyproduction 1998-2021.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	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

In [4]:

```
df.shape
```

Out[4]:

```
(985, 8)
```

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 985 entries, 0 to 984
Data columns (total 8 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   State        985 non-null    object 
 1   numcol       985 non-null    float64
 2   yieldpercol  985 non-null    int64  
 3   totalprod    985 non-null    float64
 4   stocks       985 non-null    float64
 5   priceperlb   985 non-null    float64
 6   prodvalue    985 non-null    float64
 7   year         985 non-null    int64  
dtypes: float64(5), int64(2), object(1)
memory usage: 61.7+ KB
```

In [6]:

```
df.isnull()
```

Out[6]:

	State	numcol	yieldpercol	totalprod	stocks	priceperlb	prodvalue	year
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
980	False	False	False	False	False	False	False	False
981	False	False	False	False	False	False	False	False
982	False	False	False	False	False	False	False	False
983	False	False	False	False	False	False	False	False
984	False	False	False	False	False	False	False	False

985 rows × 8 columns

In [7]:

```
df.isnull().sum()
```

Out[7]:

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

In [8]:

```
df.columns
```

Out[8]:

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

In [9]:

```
df.dtypes
```

Out[9]:

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

Key questions to be answered

question 1

How has honey production yield changed from 1998 to 2021?

In [8]:

```
df.head()
```

Out[8]:

	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

In [10]:

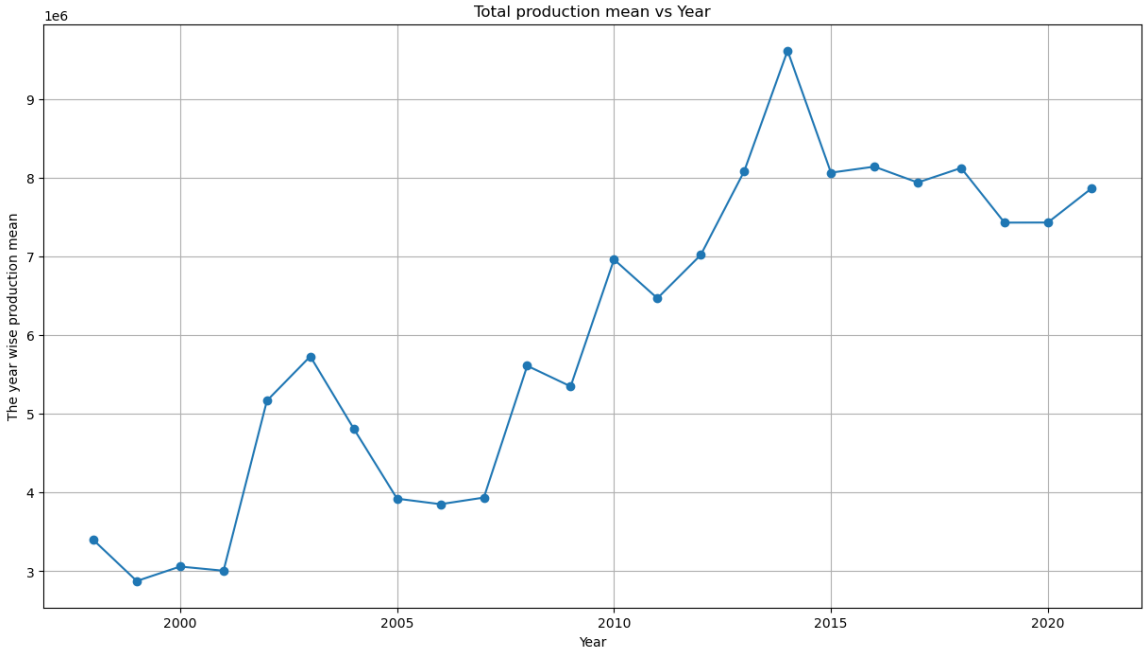
```
df1=df.groupby('year')['prodvalue'].mean()
df1
```

Out[10]:

```
year
1998    3.397465e+06
1999    2.875744e+06
2000    3.059721e+06
2001    3.006409e+06
2002    5.165955e+06
2003    5.729068e+06
2004    4.812366e+06
2005    3.921780e+06
2006    3.851805e+06
2007    3.935512e+06
2008    5.609561e+06
2009    5.348000e+06
2010    6.959250e+06
2011    6.467200e+06
2012    7.018125e+06
2013    8.079949e+06
2014    9.612075e+06
2015    8.062625e+06
2016    8.138925e+06
2017    7.937550e+06
2018    8.122075e+06
2019    7.428000e+06
2020    7.429425e+06
2021    7.860325e+06
Name: prodvalue, dtype: float64
```

In [11]:

```
plt.figure(figsize=(15,8))
plt.plot(df1,marker='o')
plt.ylabel('The year wise production mean')
plt.xlabel('Year')
plt.grid()
plt.title("Total production mean vs Year")
plt.show()
```



conclusion:

. The total honey production mean was most 9.61 in year of 2014

. The total honey production mean is less 2.67 in year of 1999

question 2

Over time , what are the major production trends across the states ?

In [15]:

```
df.head(10)
```

Out[15]:

	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
5	Florida	230000.0	98	22540000.0	4508000.0	0.64	14426000.0	1998
6	Georgia	75000.0	56	4200000.0	307000.0	0.69	2898000.0	1998
7	Hawaii	8000.0	118	944000.0	66000.0	0.77	727000.0	1998
8	Idaho	120000.0	50	6000000.0	2220000.0	0.65	3900000.0	1998
9	Illinois	9000.0	71	639000.0	204000.0	1.19	760000.0	1998

In [12]:

```
df['State'].duplicated().sum()
```

Out[12]:

```
941
```

In [13]:

```
df['State'].value_counts()
```

Out[13]:

```
Alabama           24
Montana           24
New Jersey        24
New York          24
North Carolina    24
North Dakota      24
Ohio              24
Oregon            24
Pennsylvania      24
South Dakota      24
Tennessee         24
Texas             24
Utah              24
Vermont           24
Virginia          24
Washington        24
West Virginia     24
Wisconsin         24
Wyoming           24
Arizona           24
Nebraska          24
Missouri          24
Iowa              24
Arkansas          24
California        24
Colorado          24
Florida           24
Georgia           24
Hawaii            24
Idaho             24
Illinois          24
Indiana           24
Kansas            24
Mississippi       24
Kentucky          24
Louisiana         24
Maine             24
Michigan          24
Minnesota         24
New Mexico        15
Nevada            11
South Carolina    11
Oklahoma           6
Maryland           6
Name: State, dtype: int64
```

In [35]:

```
df_state=df[df['State']=='Alabama'][['totalprod','year']]
df_state.head()
```

Out[35]:

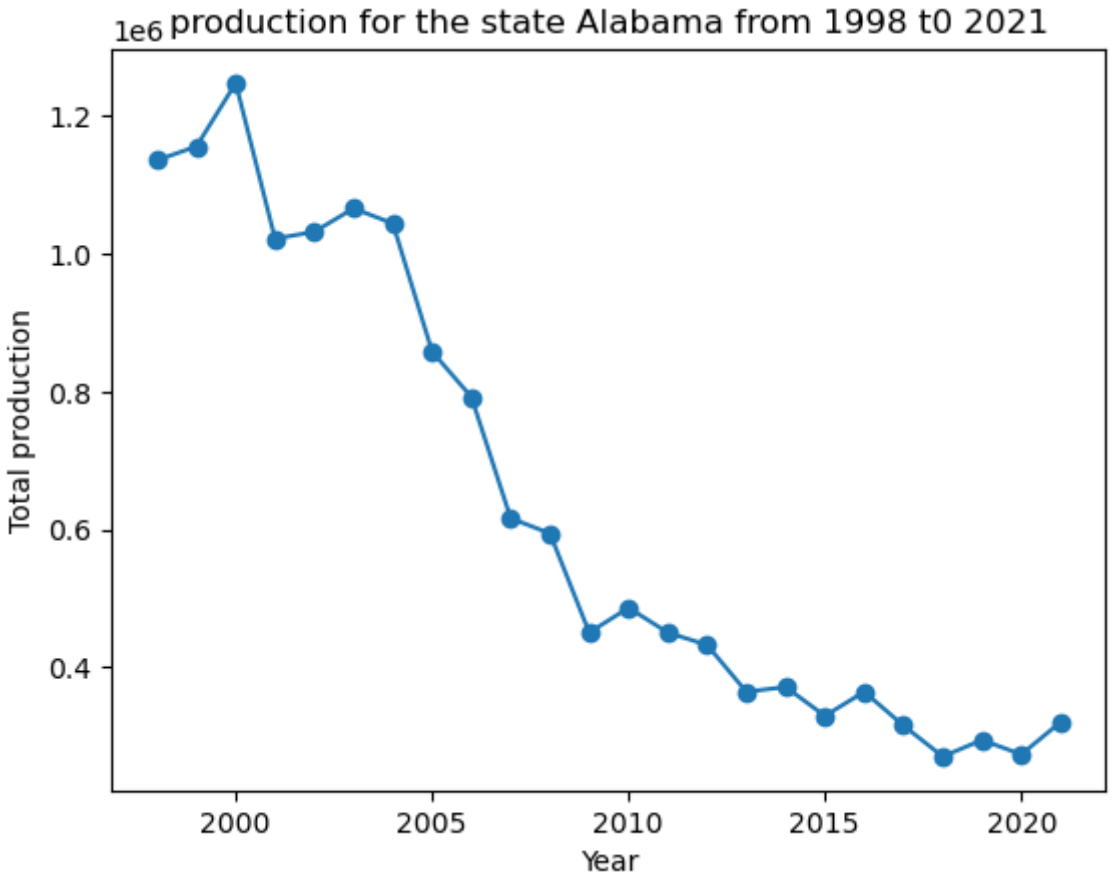
	totalprod	year
0	1136000.0	1998
43	1156000.0	1999
86	1248000.0	2000
129	1022000.0	2001
173	1032000.0	2002

In [45]:

```
def prod_state(state):
    df2=df[df['State']==state][['totalprod','year']]
    plt.plot(df2['year'],df2['totalprod'],marker='o')
    plt.xlabel('Year')
    plt.ylabel('Total production')
    plt.title(f'production for the state {state} from 1998 t0 2021')
    plt.show()
```

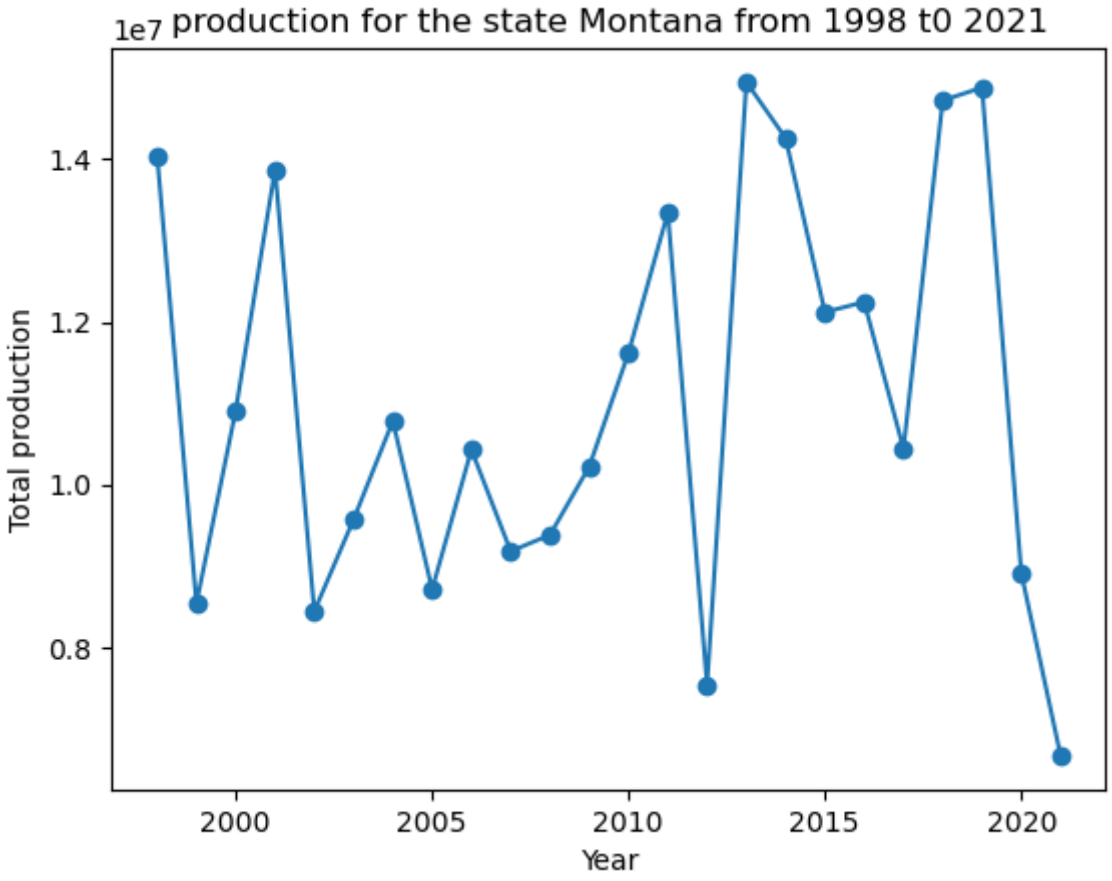
In [47]:

```
prod_state("Alabama")
```



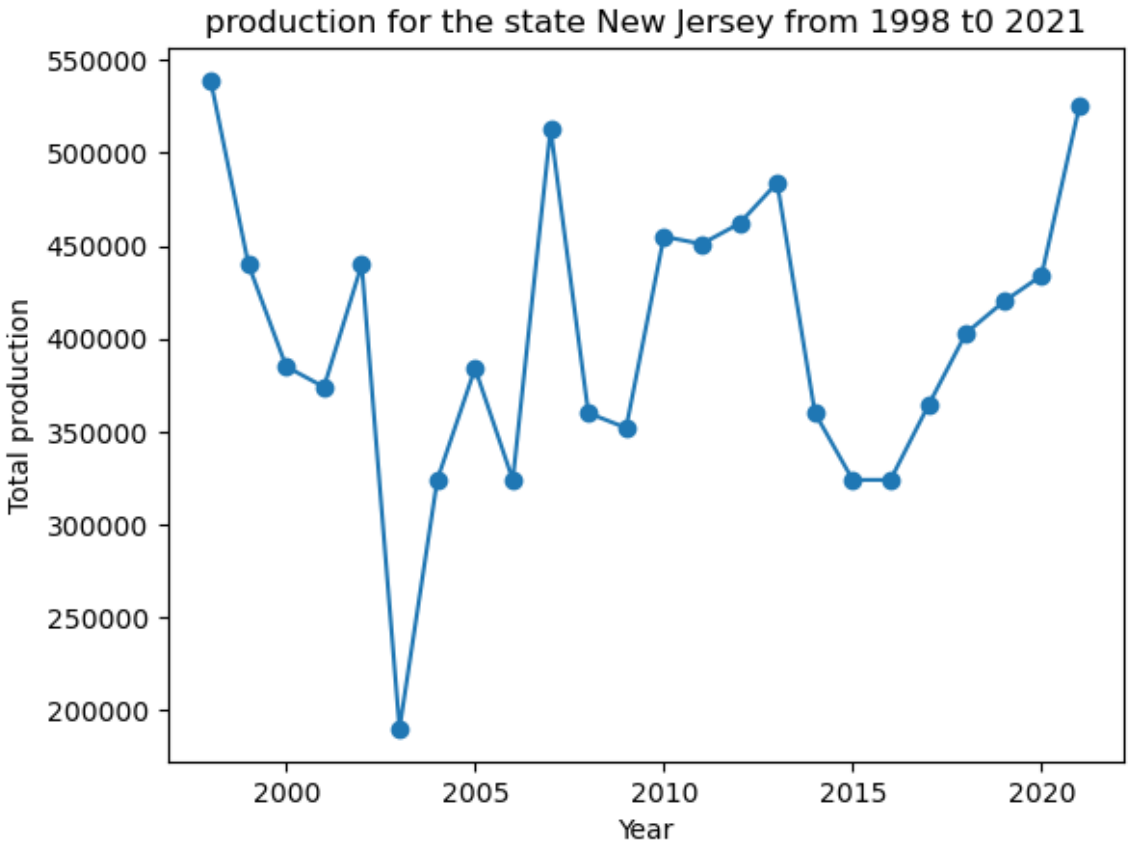
In [48]:

```
prod_state("Montana")
```



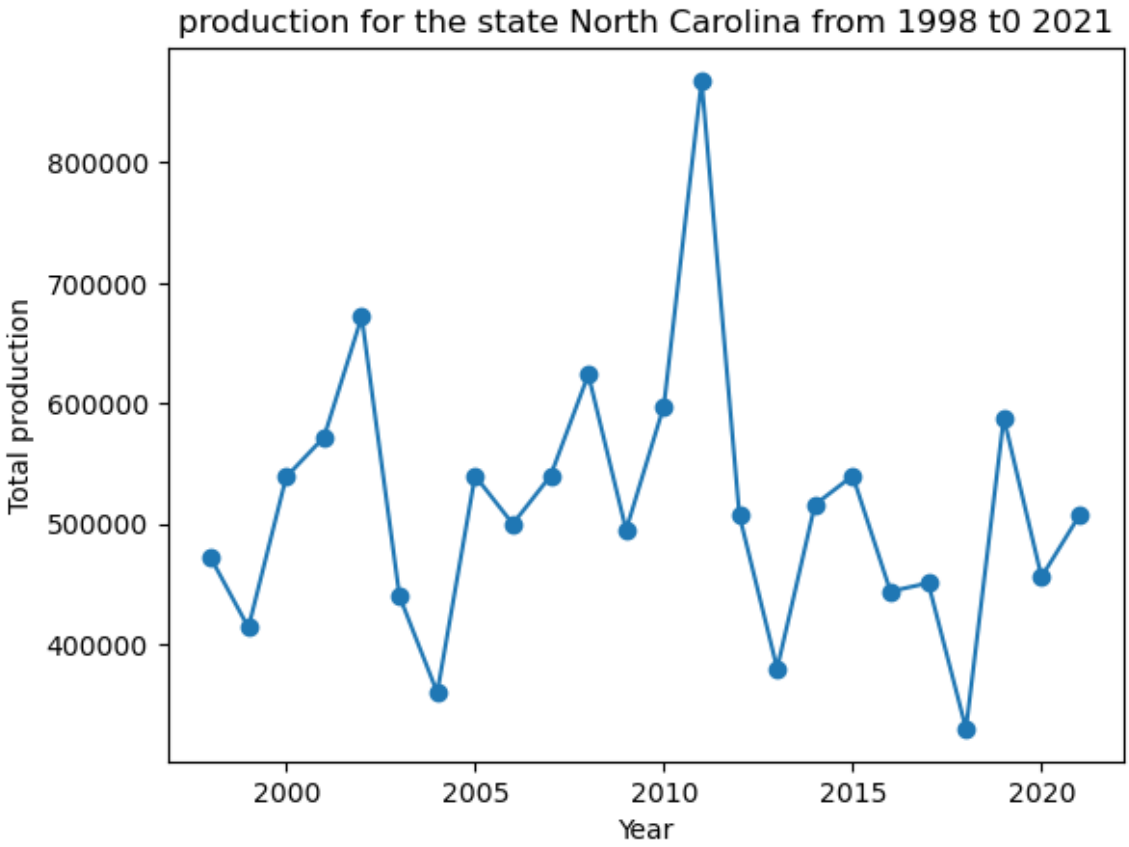
In [49]:

```
prod_state("New Jersey")
```



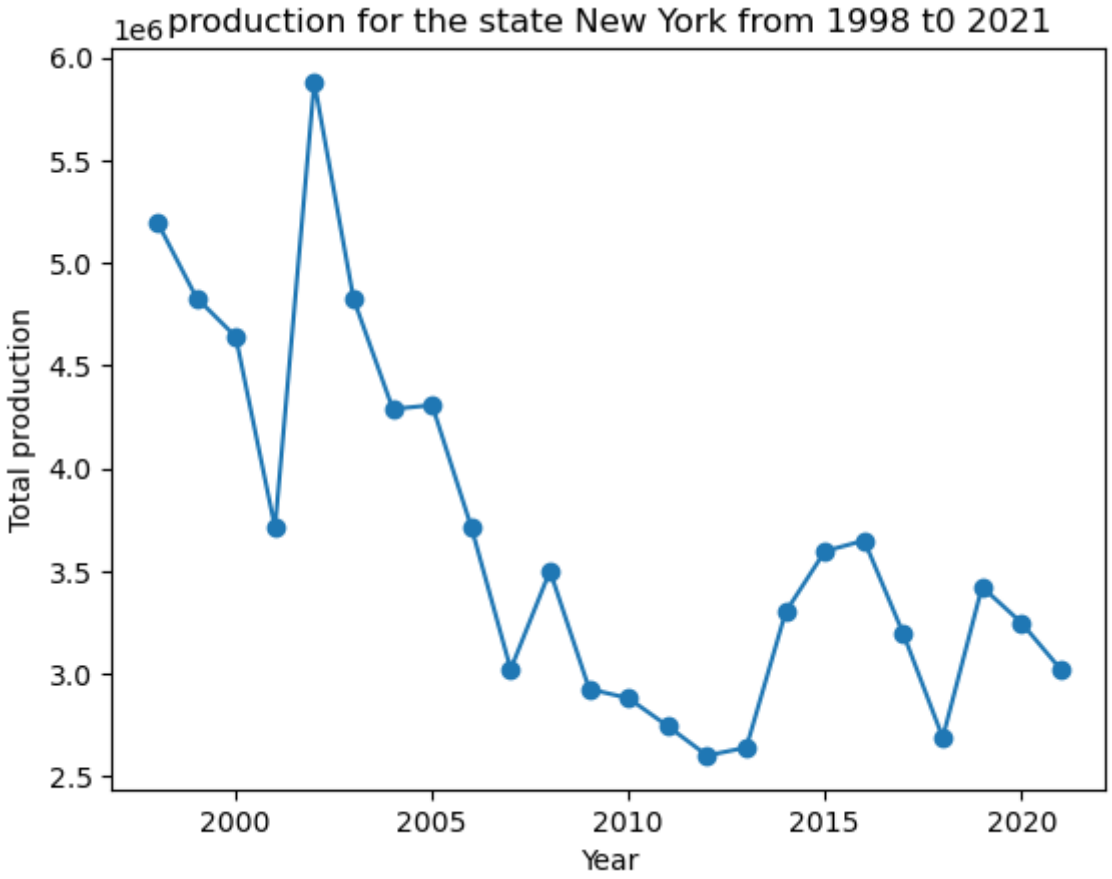
In [50]:

```
prod_state("North Carolina")
```



In [51]:

```
prod_state("New York")
```



In [ ]:

```

```

question 3

Does the data show any trends in terms of the number of honey producing colonies and yieldpercolony before 2006,which was when concernover Colony Collapse Disorder spread nation wide?

In [84]:

```
df.head()
```

Out[84]:

	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

In [ ]:

```
df.reset_index(drop=True,inplace=True)
```

In [24]:

```
df3=df[df['year']<=2006]
df3
```

Out[24]:

	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
...	...	...	...	...	...	...	...	...
379	Virginia	8000.0	42	336000.0	114000.0	2.20	739000.0	2006
380	Washington	49000.0	52	2548000.0	1605000.0	1.24	3160000.0	2006
381	West Virginia	6000.0	42	252000.0	68000.0	2.02	509000.0	2006
382	Wisconsin	64000.0	93	5952000.0	2500000.0	1.12	6666000.0	2006
383	Wyoming	39000.0	85	3315000.0	497000.0	0.98	3249000.0	2006

384 rows × 8 columns

In [33]:

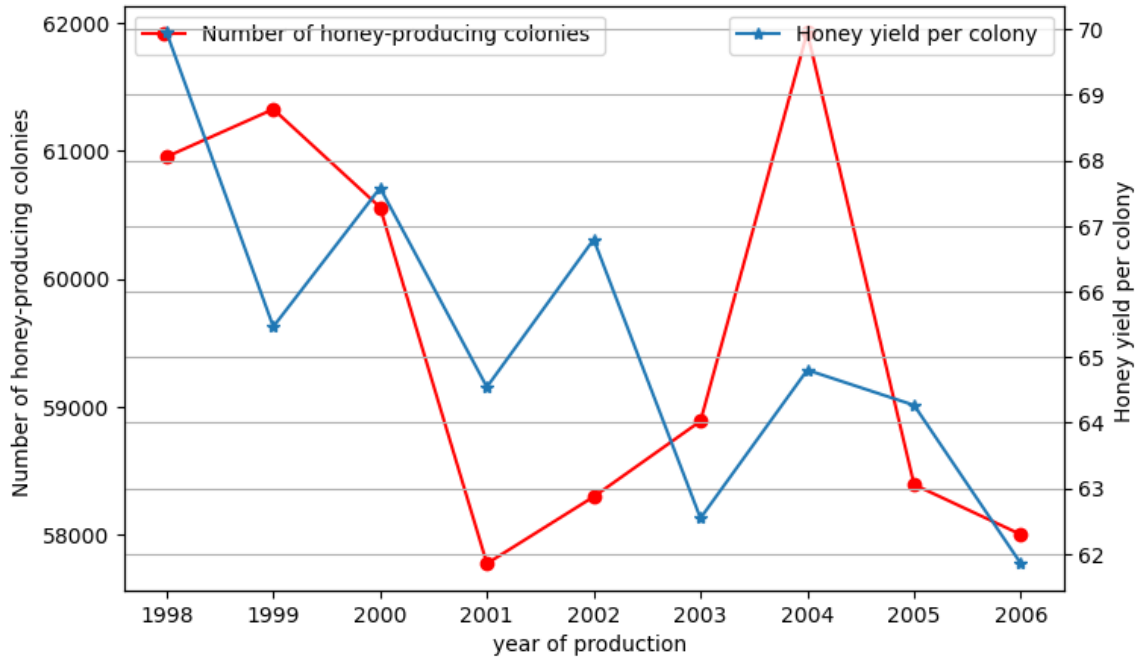
```
df4=df3.groupby('year')[['numcol','yieldpercol']].mean().reset_index()
df4.head()
```

Out[33]:

	year	numcol	yieldpercol
0	1998	60953.488372	69.953488
1	1999	61325.581395	65.465116
2	2000	60558.139535	67.581395
3	2001	57772.727273	64.545455
4	2002	58295.454545	66.795455

In [63]:

```
fig=plt.figure(figsize=(8,5))
ax1=plt.subplot(111)
ax1.plot(df4.year,df4.numcol,'red',marker='o',label="Number of honey-producing colonies "
ax2=ax1.twinx()
ax2.plot(df4.year,df4.yieldpercol, marker='*',label='Honey yield per colony ')
ax1.set_xlabel("year of production ")
ax1.set_ylabel("Number of honey-producing colonies ")
ax2.set_ylabel('Honey yield per colony ')
ax1.legend()
ax2.legend()
plt.grid()
plt.show()
```



In [41]:

```
# converting the numcol in million
df4['numcol']=df4['numcol']*1000000
df4.head()
```

Out[41]:

	year	numcol	yieldpercol
0	1998	60953.488372	69.953488
1	1999	61325.581395	65.465116
2	2000	60558.139535	67.581395
3	2001	57772.727273	64.545455
4	2002	58295.454545	66.795455

In [ ]:

question 4

Are there any patterns that can be observed between total honey production and value of production every year?

In [88]:

```
df.head()
```

Out[88]:

	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

In [54]:

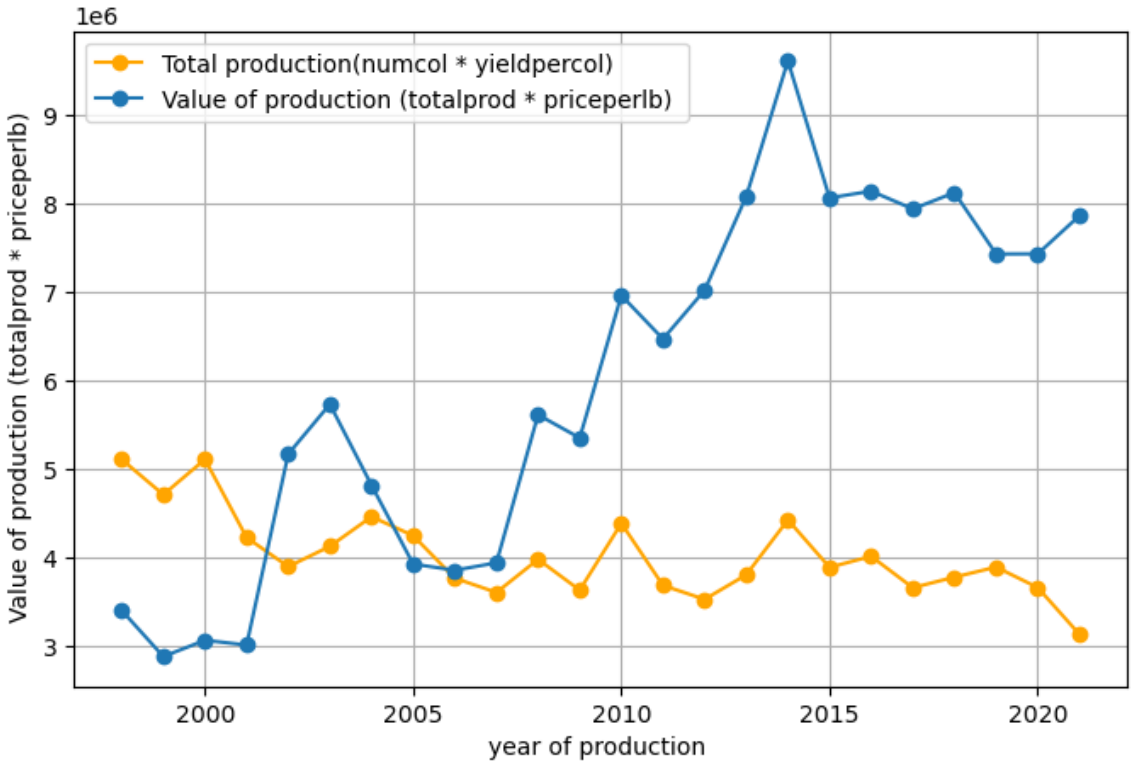
```
df5=df.groupby('year')['totalprod','prodvalue'].mean().reset_index()
df5
```

Out[54]:

	year	totalprod	prodvalue
0	1998	5.105093e+06	3.397465e+06
1	1999	4.706674e+06	2.875744e+06
2	2000	5.106000e+06	3.059721e+06
3	2001	4.221545e+06	3.006409e+06
4	2002	3.892386e+06	5.165955e+06
5	2003	4.122091e+06	5.729068e+06
6	2004	4.456805e+06	4.812366e+06
7	2005	4.243146e+06	3.921780e+06
8	2006	3.761902e+06	3.851805e+06
9	2007	3.600512e+06	3.935512e+06
10	2008	3.974927e+06	5.609561e+06
11	2009	3.626700e+06	5.348000e+06
12	2010	4.382350e+06	6.959250e+06
13	2011	3.680025e+06	6.467200e+06
14	2012	3.522675e+06	7.018125e+06
15	2013	3.800103e+06	8.079949e+06
16	2014	4.421650e+06	9.612075e+06
17	2015	3.884400e+06	8.062625e+06
18	2016	4.008925e+06	8.138925e+06
19	2017	3.654125e+06	7.937550e+06
20	2018	3.773725e+06	8.122075e+06
21	2019	3.887600e+06	7.428000e+06
22	2020	3.655475e+06	7.429425e+06
23	2021	3.127925e+06	7.860325e+06

In [64]:

```
fig=plt.figure(figsize=(8,5))
ax1=plt.subplot(111)
ax1.plot(df5.year,df5.totalprod,'orange',marker='o',label="Total production(numcol * yiel
ax2=ax1
ax2.plot(df5.year,df5.prodvalue, marker='o',label='Value of production (totalprod * price
ax1.set_xlabel("year of production ")
ax1.set_ylabel('Total production(numcol * yieldpercol)')
ax2.set_ylabel('Value of production (totalprod * priceperlb)  ')
ax1.legend()
ax2.legend()
plt.grid()
plt.show()
```



question5

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

In [93]:

```
df.head()
```

Out[93]:

	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

In [71]:

```
df6=df.groupby('year')['totalprod','priceperlb'].mean().reset_index()
```

In [72]:

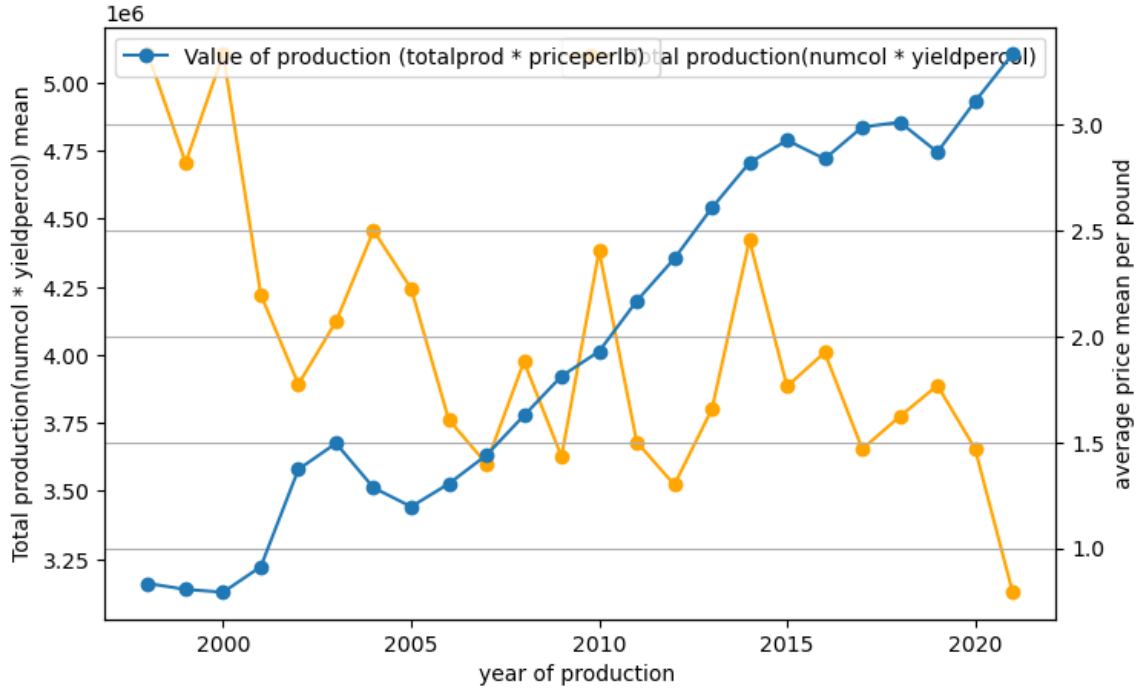
```
df6
```

Out[72]:

	year	totalprod	priceperlb
0	1998	5.105093e+06	0.832558
1	1999	4.706674e+06	0.804186
2	2000	5.106000e+06	0.791395
3	2001	4.221545e+06	0.911818
4	2002	3.892386e+06	1.371364
5	2003	4.122091e+06	1.494773
6	2004	4.456805e+06	1.284634
7	2005	4.243146e+06	1.195122
8	2006	3.761902e+06	1.303659
9	2007	3.600512e+06	1.438293
10	2008	3.974927e+06	1.625610
11	2009	3.626700e+06	1.812000
12	2010	4.382350e+06	1.929000
13	2011	3.680025e+06	2.167250
14	2012	3.522675e+06	2.367000
15	2013	3.800103e+06	2.607179
16	2014	4.421650e+06	2.820250
17	2015	3.884400e+06	2.926250
18	2016	4.008925e+06	2.838500
19	2017	3.654125e+06	2.988750
20	2018	3.773725e+06	3.011750
21	2019	3.887600e+06	2.870500
22	2020	3.655475e+06	3.109250
23	2021	3.127925e+06	3.334250

In [74]:

```
fig=plt.figure(figsize=(8,5))
ax1=plt.subplot(111)
ax1.plot(df6.year,df6.totalprod,'orange',marker='o',label="Total production(numcol * yiel
ax2=ax1.twinx()
ax2.plot(df6.year,df6.priceperlb, marker='o',label='Value of production (totalprod * pric
ax1.set_xlabel("year of production ")
ax1.set_ylabel('Total production(numcol * yieldpercol) mean')
ax2.set_ylabel('average price mean per pound')
ax1.legend()
ax2.legend()
plt.grid()
plt.show()
```



question 6

US_Honey_Production-EDA

Explporatory data analysis on "Honey Production" data in US from the year 1998 to 2021

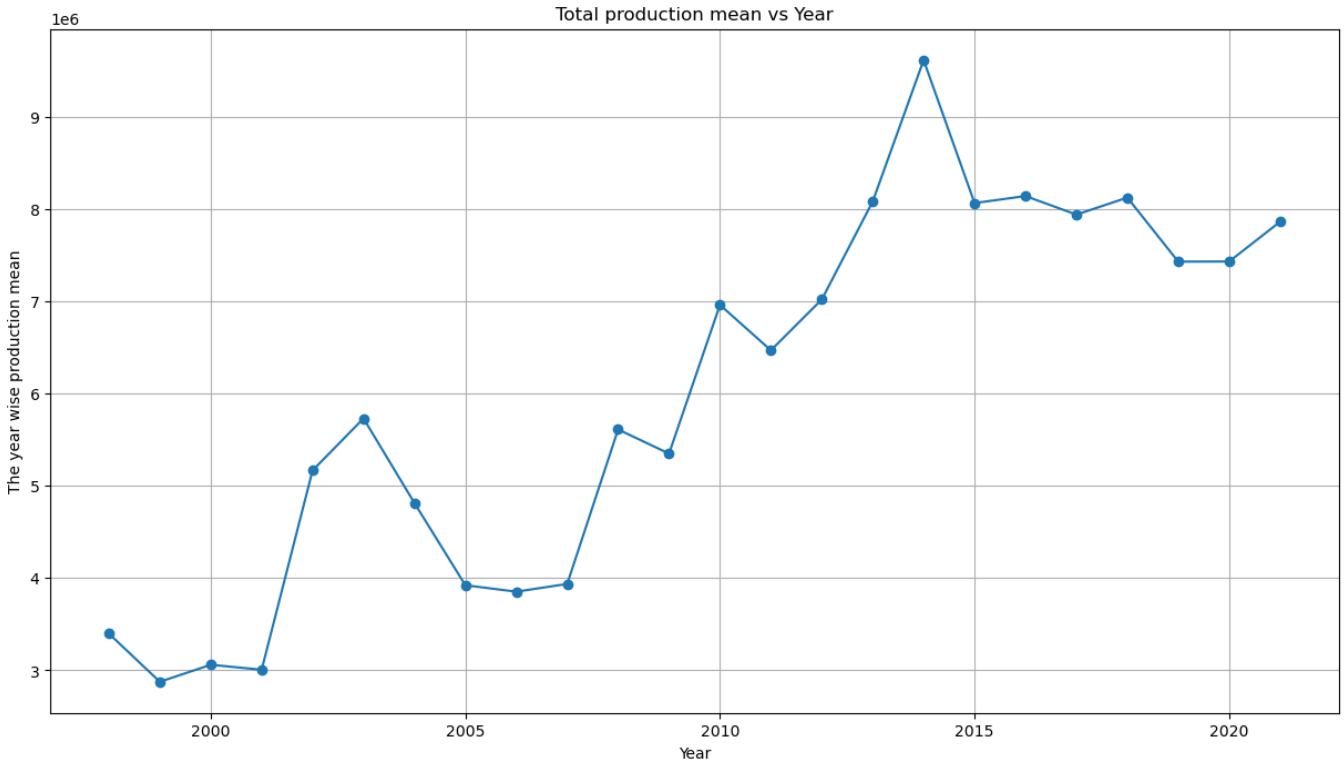
Objectives of this project

To get insights about the United States Honey Production from 1998 to

2021

Insights from Data

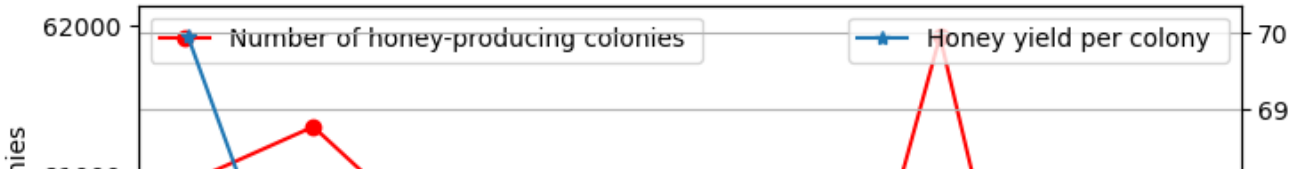
In this data, there is no missing value spotted. So lets talk about the data. The data is all about honey production, supply and demand in United States from 1998 to 2021.



. The total honey production mean was most 9.61 in year of 2014
. The total honey production mean is less 2.67 in year of 1999

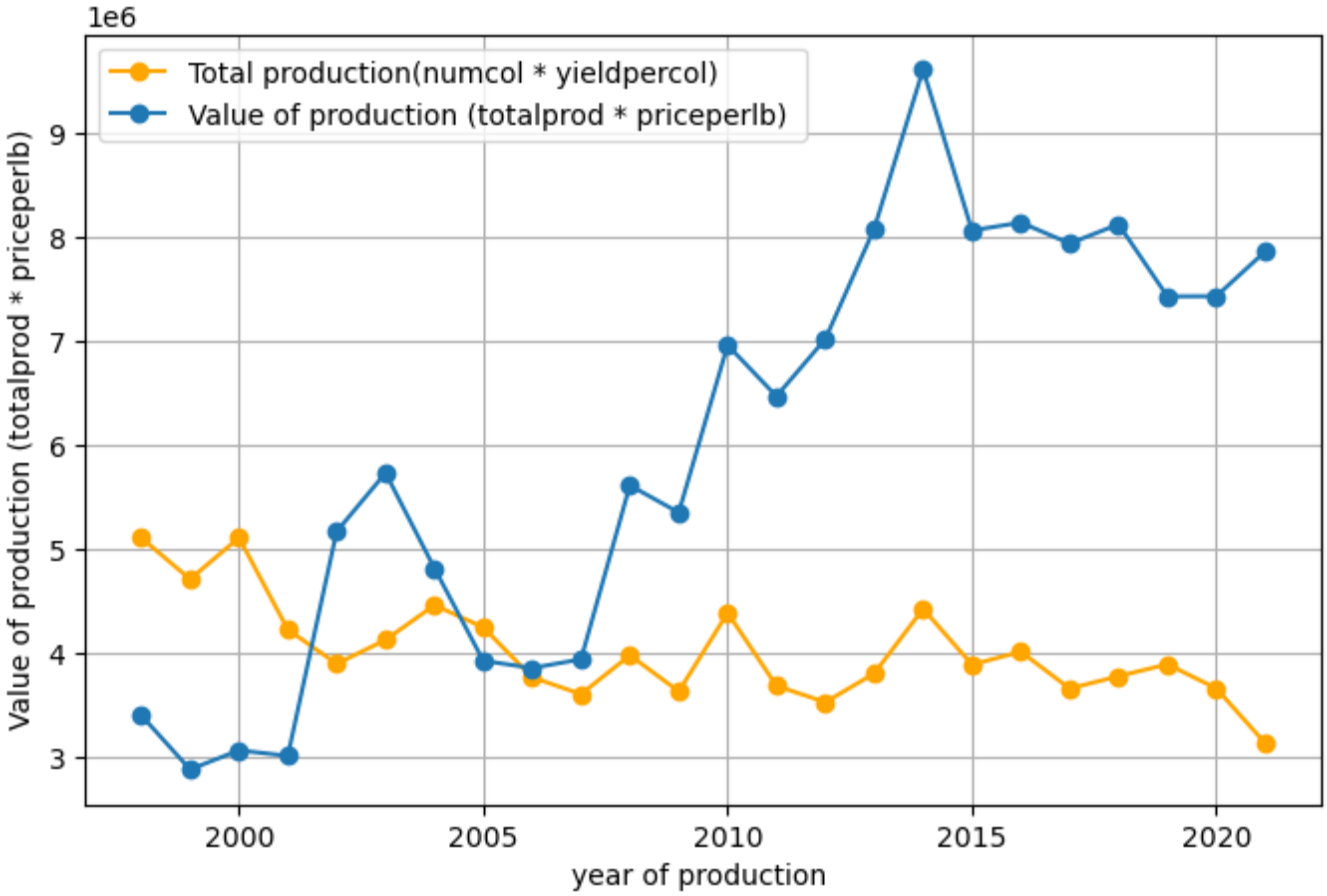
Over time , what are the major production trends across the states ?

Does the data show any trends in terms of the number of honey producing colonies and yieldpercolony before 2006,which was when concernover Colony Collapse Disorder spread nation wide?



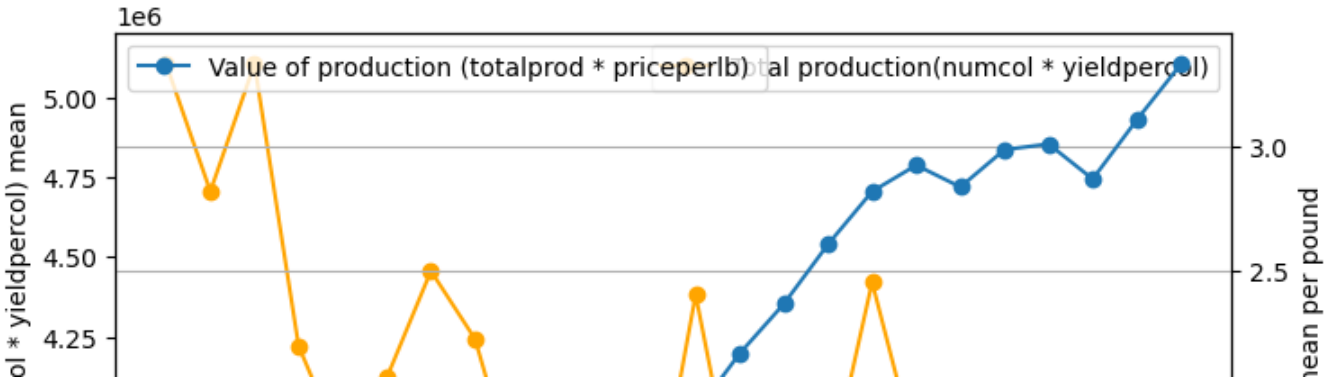
. Honey yield per colony are going less and less in number, so downward trend created towards 2006.
. honey producing colony highest in 2004 and lowest in 2006 , created toward 2006

Are there any patterns that can be observed between total honey production and value of production every year?



. total production (numcol * yieldpercol) going less and less in number , so downward
. value of production(totalprod * priceperlb) is higher in 2014

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



. total production (numcol * yieldpercol) mean is upward every year of production

In [ ]: