Deepnote Rajesh Kumar / Python Published at Apr 24, 2023 Unlisted

Honey Production Case Study 1998-2021

Background: In 2006, global concern was raised over the rapid decline in the honeybee population, an integral component of American honey agriculture. Large numbers of hives were lost to Colony Collapse Disorder, a phenomenon of disappearing worker bees causing the remaining hive colony to collapse. Speculation as to the cause of this disorder points to hive diseases and pesticides harming the pollinators, though no overall consensus has been reached. The U.S. used to locally produce over half the honey it consumes per year. Now, honey mostly comes from overseas, with 350 of the 400 million pounds of honey consumed every year originating from imports. This dataset provides insight into honey production supply and demand in America from 1998 to 2021.

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

Key questions to be answered: 1. How has honey production yield changed from 1998 to 2021? 2. Over time, what are the major production trends across the states? 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? 4. Are there any patterns that can be observed between total honey production and value of production every year? 5. How has the value of production, which in some sense could be tied to demand, changed every year? 6. Constructs the related plots using Seaborn and Matplot apply customization and derive insights from the visualization.

Dataset: state: Various states of the U.S. numcol: Number of honey-producing colonies. Honey producing colonies are the maximum number of colonies from which honey was taken during the year. It is possible to take honey from colonies that did not survive the entire year yieldpercol: Honey yield per colony. Unit is pounds totalprod: Total production (numcol x yieldpercol). Unit is pounds stocks: Refers to stocks held by producers. Unit is pounds priceperlb: Refers to average price per pound based on expanded sales. The unit is dollars. prodvalue: Value of production (totalprod x priceperlb). The unit is dollars. year: Year of production.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

Read the dataset

```
honeyprod = pd.read_csv("honeyproduction 1998-2021.csv")
```

First few rows of the dataset

```
honeyprod.head()
```

	State object	numcol float64	yieldpercol int64	totalprod float64	stocks float64	priceperlb float64	prodvalue float64	year int64
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	0.7	1361000.0	1998

Observation:- The dataset looks clean and consistent with the Data Dictionary.

Checking the shape of the dataset

```
honeyprod.shape
```

```
(985, 8)
```

Observation:- In this Dataset 985 Rows of 8 columns.

Check the datatype of each column to make sure the data is read in properly.

```
honeyprod.dtypes
```

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

Observations: 1)state is object or string data type. 2)All the other variables are numerical and their python data types are float64 and int64.

Analyse the quantitative variables in the dataset

```
honeyprod.describe()
```

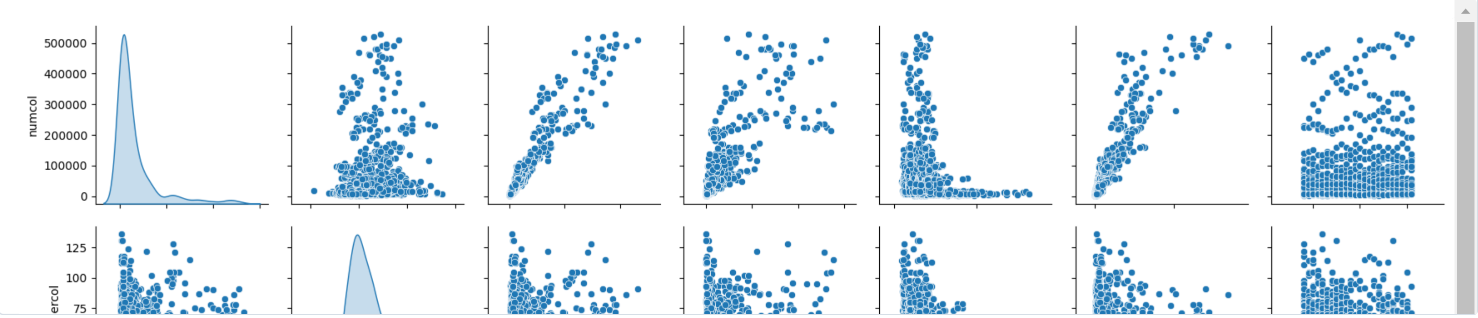
	numcol float64	yieldpercol float64	totalprod float64	stocks float64	priceperlb float64	prodvalue float64	year float64
count	985.0	985.0	985.0	985.0	985.0	985.0	985.0
mean	62892.38578680203	58.40203045685279	4035131.9796954314	1167186.802030457	1.9695939086294416	5939910.659898478	2009.2913705583755
std	94163.79191719607	19.291695313480805	6752289.667882031	2088588.6070600604	1.1779216563159631	9806594.833674308	6.962251342040307
min	2000.0	3.0	84000.0	8000.0	0.49	162000.0	1998.0
25%	9000.0	45.0	469000.0	108000.0	1.2	1037000.0	2003.0
50%	26000.0	55.0	1488000.0	360000.0	1.7	2409000.0	2009.0
75%	68000.0	70.0	3780000.0	1217000.0	2.36	5897000.0	2015.0
max	530000.0	136.0	46410000.0	13800000.0	8.23	83859000.0	2021.0

Observations: 1)Number of colonies in every state are spread over a huge range. Ranging from 2000 to 530000. 2)The mean numcol is close to the 75% percentile of the data, indicating a right skew. 3)As expected, standard deviation of numcol is very high 4)Yield per colony also has spread ranging from 3 pounds to 136 pounds. 5)Infact, all the variable seem to have a huge range, we will have to investigate further if this spread is mainly across different states or varies in the same state over the years.

Relationship between numerical variables using pair plots.

```
sns.pairplot(honeyprod, diag_kind="kde")
```

```
<seaborn.axisgrid.PairGrid at 0x7f0fca878ac0>
```

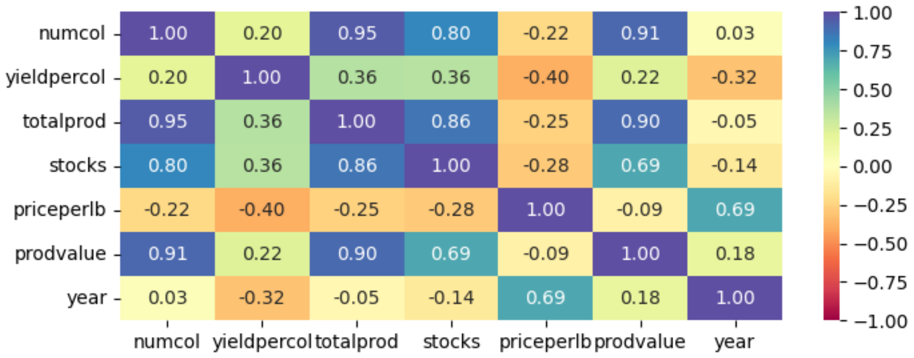


Relationship between numerical variables using correlation plots.

```python
correlation = honeyprod.corr()
# creating a 2-D Matrix with correlation plots
correlation
```

	numcol float64	yieldpercol float64	totalprod float64	stocks float64	priceperlb float64	prodvalue float64	year float64
numcol	1.0	0.19857614085812222	0.9496400898773083	0.7968966692622902	-0.22336015981409435	0.9126010141640781	0.03436490289295102
yieldpercol	0.19857614085812222	1.0	0.36439228981431165	0.35629385958359333	-0.3982469631997554	0.22470392220060167	-0.3183180552873403
totalprod	0.9496400898773083	0.36439228981431165	1.0	0.8643334740641427	-0.25267804829860246	0.9007201224627923	-0.04883749867794546
stocks	0.7968966692622902	0.35629385958359333	0.8643334740641427	1.0	-0.28489356530961263	0.6933722989536647	-0.1375903361956744
priceperlb	-0.22336015981409435	-0.3982469631997554	-0.25267804829860246	-0.28489356530961263	1.0	-0.08852908709640848	0.6947380111463096
prodvalue	0.9126010141640781	0.22470392220060167	0.9007201224627923	0.6933722989536647	-0.08852908709640848	1.0	0.17803966051475836
year	0.03436490289295102	-0.3183180552873403	-0.04883749867794546	-0.1375903361956744	0.6947380111463096	0.17803966051475836	1.0

```python
plt.figure(figsize=(8, 3))
sns.heatmap(correlation, annot=True, vmin=-1, vmax=1, fmt=".2f", cmap="Spectral")
plt.show()

# following code for information of the arguments
# help(sns.heatmap)
```



Observations:- 1)Number of colonies have a high positive correlation with total production, stocks and the value of production. As expected, all these values are highly correlated with each other. 2)Yield per colony does not have a high correlation with any of the features that we have in our dataset. 3)Same is the case with priceperlb. 4)Determining the factors influencing per colony yield and price per pound of honey would need further investigation.

Explore the number of state and year.

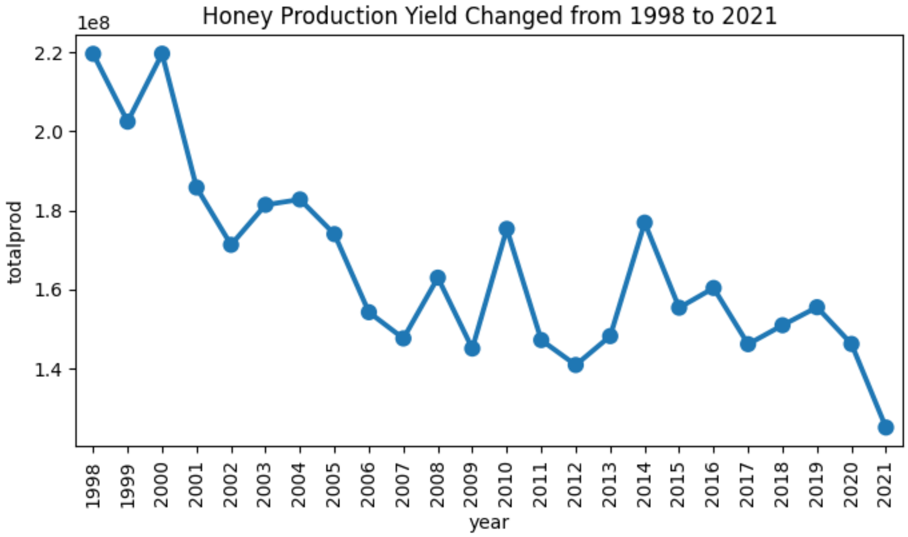
```python
print(honeyprod.State.nunique())
print(honeyprod.year.nunique())
```

```
44
24
```

Observation:- We have honey production data for 44 US states over a span of 24 years, from 1998 to 2021.

Let us honey production yield changed from 1998 to 2021 .

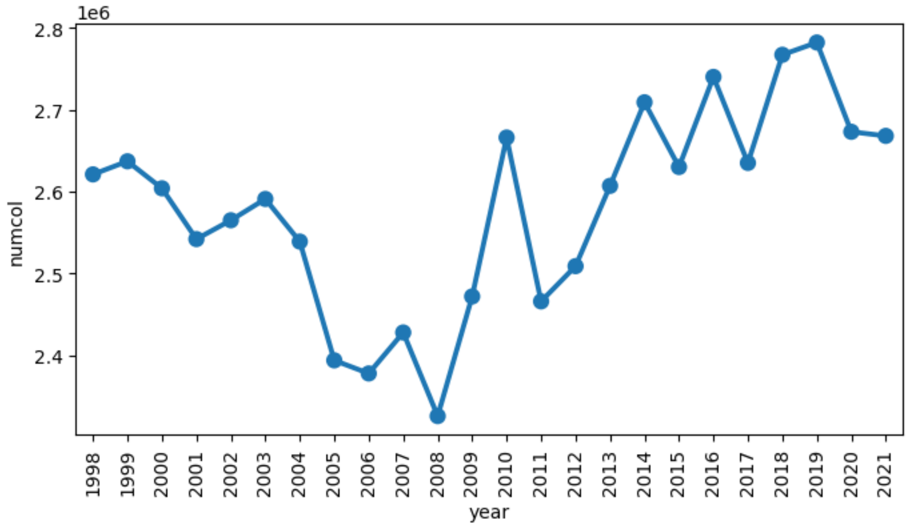
```
plt.figure(figsize=(8, 4))
sns.pointplot(x='year', y='totalprod', data=honeyprod, estimator=sum, errorbar= None)
plt.xticks(rotation=90) # To rotate the x axis labls
plt.title('Honey Production Yield Changed from 1998 to 2021')
plt.show()

# following code to check the actual values
# honeyprod.groupby(['year'])['totalprod'].sum().reset_index()
```



Observations: 1)The overall honey production in the US has been decreasing over the years. 2)Total honey production = number of colonies * yield per colony. Let us check if the honey production is decreasing due to one of these factors or both.

Show the variation in number of colonies with years.

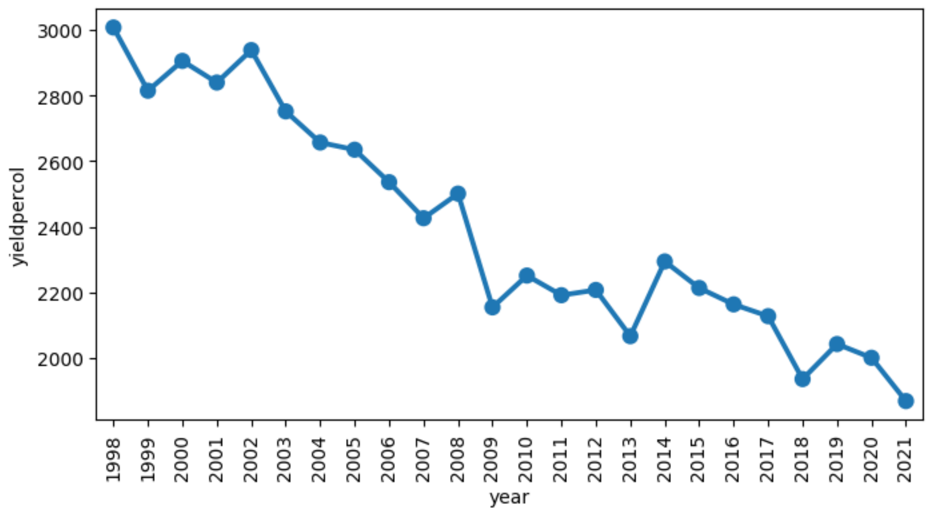
```
plt.figure(figsize=(8, 4))
sns.pointplot(x='year', y='numcol', data=honeyprod, errorbar=None, estimator=sum)
plt.xticks(rotation=90) # To rotate the x axis labls
plt.show()
```



Observations: 1)The number of colonies across the country shows a declining trend from 1998-2008 but has seen an uptick since 2008. 2)It is possible that there was some intervansion in 2008 that help in increasing the number of honey bee colonies across the country.

Show the variation in yield per colony with years.

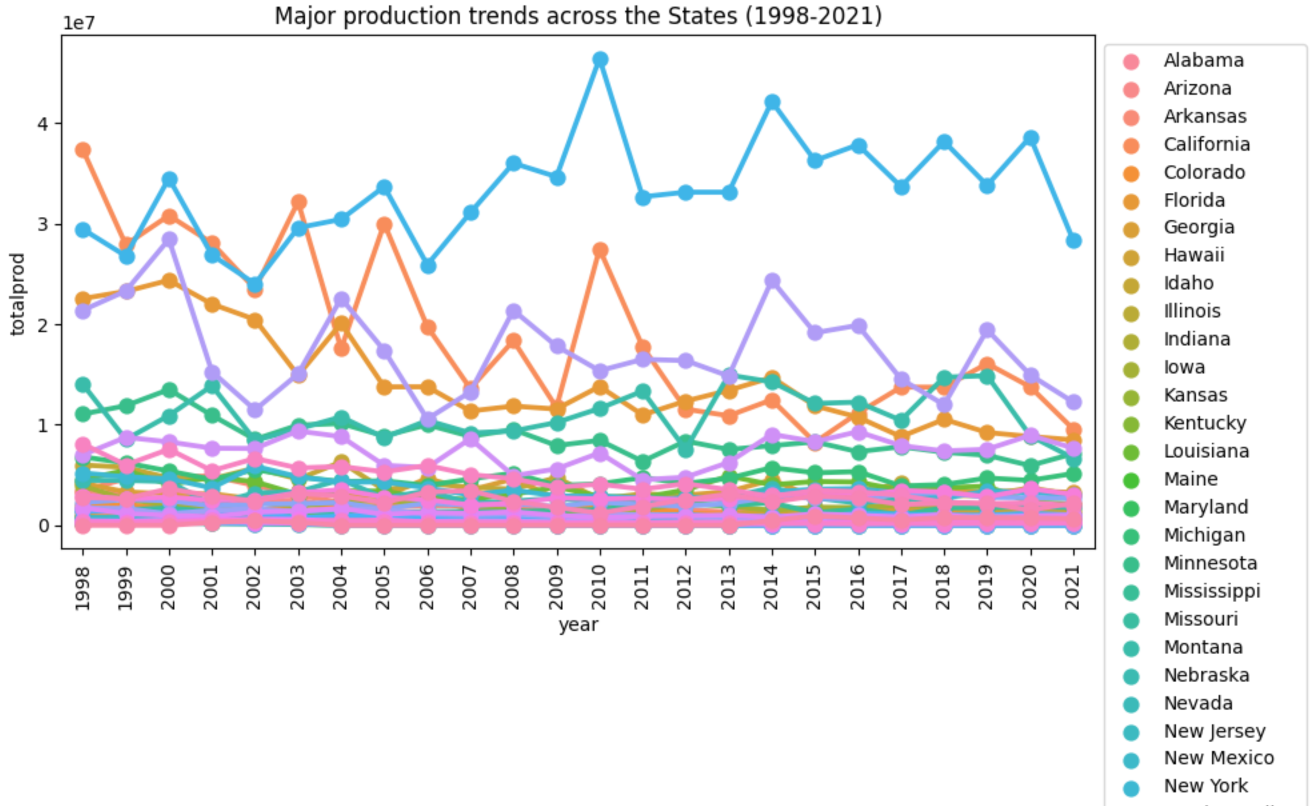
```
plt.figure(figsize=(8, 4))
sns.pointplot(x='year', y='yieldpercol', data=honeyprod, estimator=sum, errorbar=None)
plt.xticks(rotation=90) # To rotate the x axis lables
plt.show()
```



Observation: 1)In contrast to number of colonies, the yield per colony has been decreasing since 1998. 2)This indicates that, it is not the number of colonies that is causing a decline in total honey production but the yield per colony.

Let us look major production trends across the states.

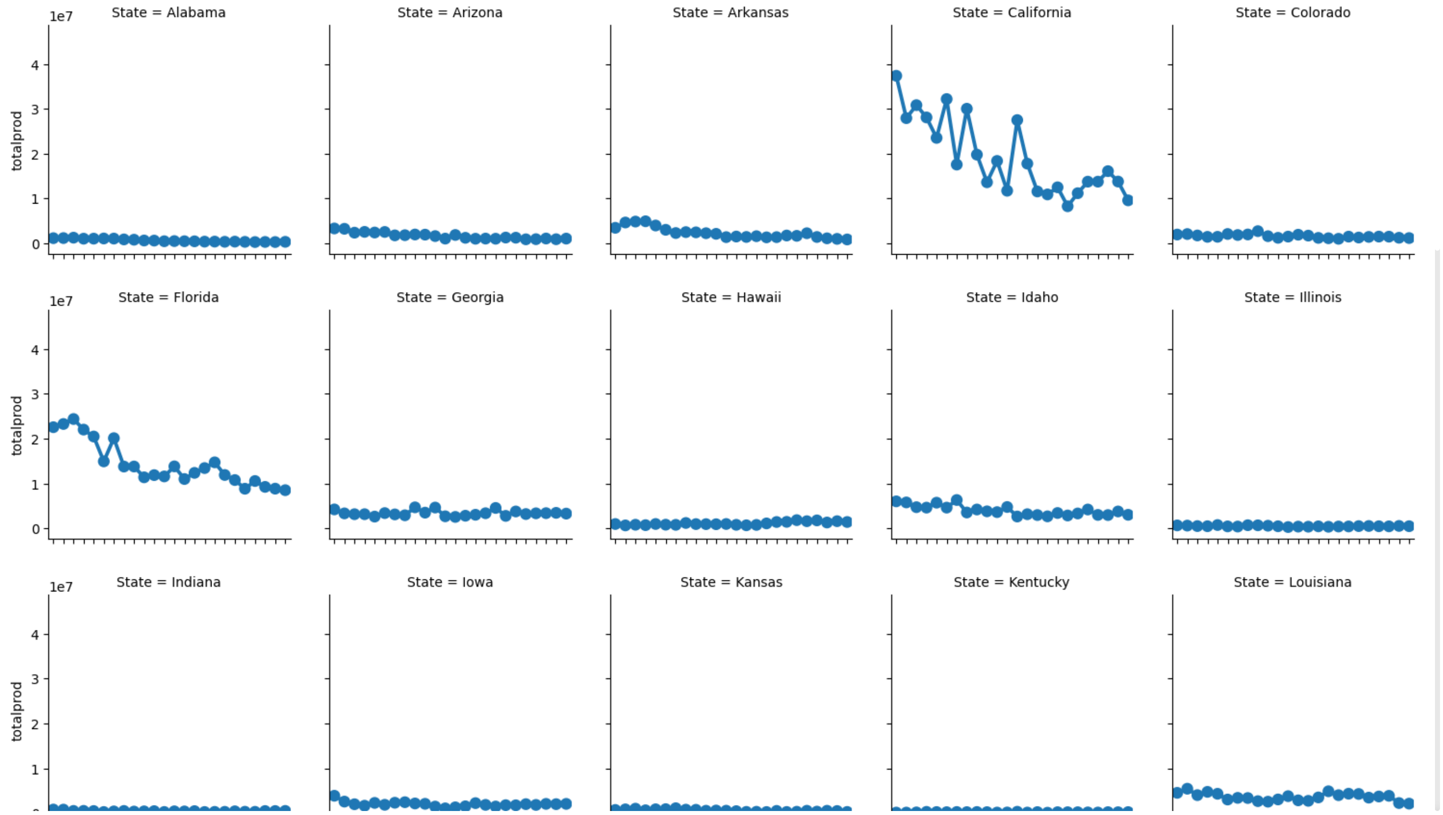
```
# Add hue parameter to the pointplot to plot for each state
plt.figure(figsize=(10, 5)) # To resize the plot
sns.pointplot(x='year', y='totalprod', data=honeyprod, estimator=sum, errorbar=None, hue = 'State')
plt.legend(bbox_to_anchor=(1, 1))
plt.xticks(rotation=90) # To rotate the x axis labls
plt.title("Major production trends across the States (1998-2021)")
plt.show()
```



Observation:- There are some states that have much higher productions than the others but this plot is a little hard to read. Let us try plotting each state separately for a better understanding.

Let us plotting each state separately for a better understanding.

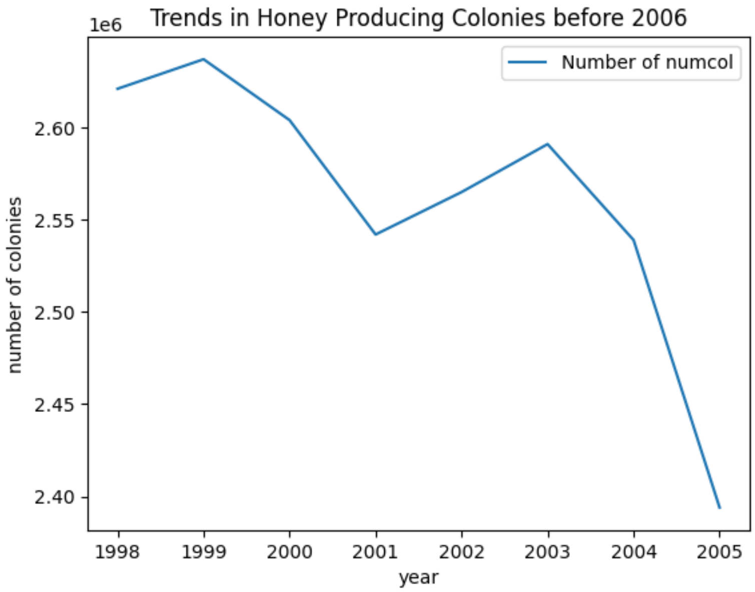
```
sns.catplot(x='year', y='totalprod', data=honeyprod,
                estimator=sum, col='State', kind="point",
                height=3,col_wrap = 5)
plt.xticks(rotation=90)
plt.show()
```



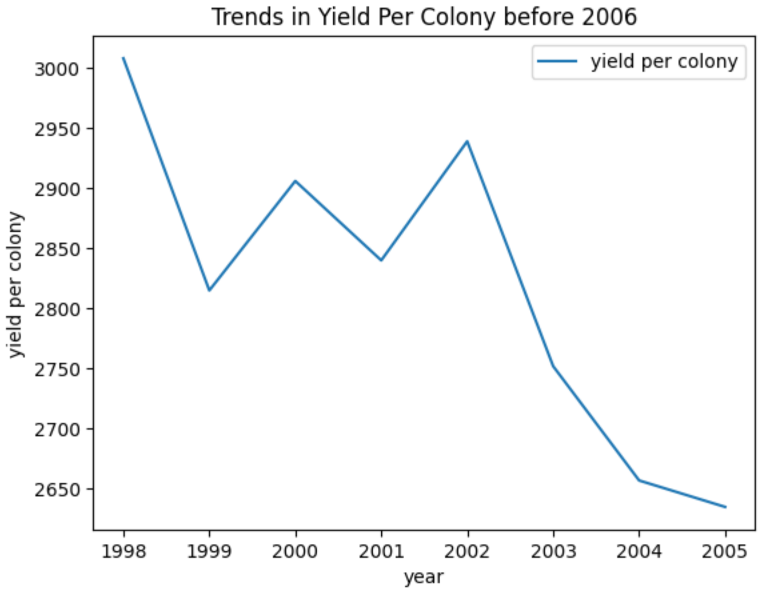
Observations:- 1)The most prominent honey producing states of US are - California, Florida, North Dakota and South Dakota and Montana 2)Unfortunately, the honey production in California has seen a steep decline over the years. 3)Florida's total production also has been on a decline. 4)South Dakota has more of less maintained its levels of production. 5)North Dakota has actually seen an impressive increase in the honey production.

Data show the 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.

```
honey_data = pd.read_csv('honeyproduction 1998-2021.csv')
pre_2006_data = honey_data[honey_data['year'] < 2006]
sum_numcol = pre_2006_data.groupby('year')['numcol'].sum()

plt.plot(sum_numcol.index, sum_numcol.values, label='Number of numcol')
plt.xlabel('year')
plt.ylabel('number of colonies')
plt.title('Trends in Honey Producing Colonies before 2006')
plt.legend()
plt.show()
```

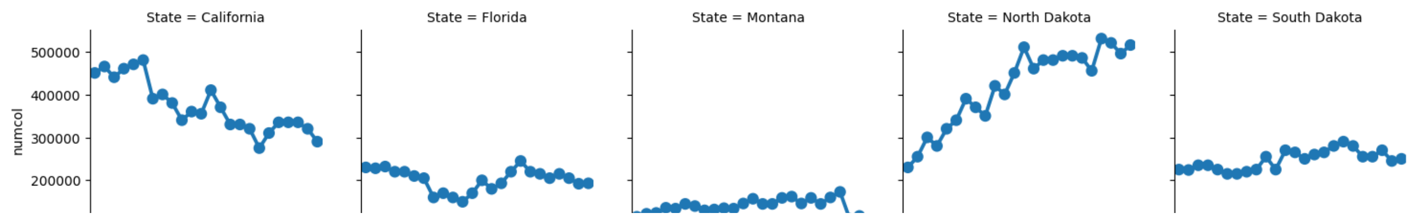


```
pre_2006_data = honey_data[honey_data['year'] < 2006]
sum_yieldpercol = pre_2006_data.groupby('year')['yieldpercol'].sum()

plt.plot(sum_yieldpercol.index, sum_yieldpercol.values, label='yield per colony')
plt.xlabel('year')
plt.ylabel('yield per colony')
plt.title('Trends in Yield Per Colony before 2006')
plt.legend()
plt.show()
```

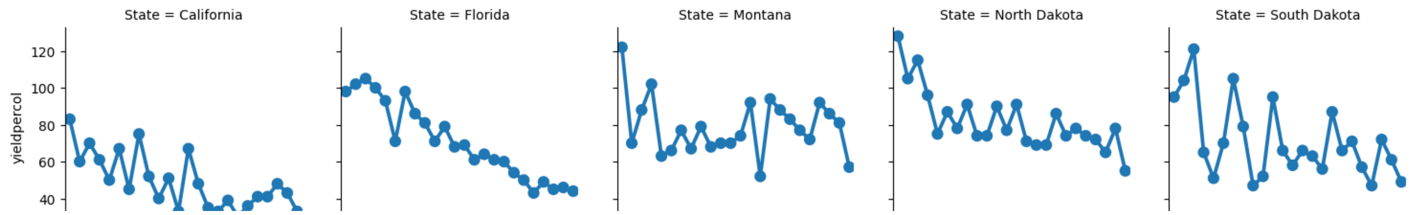


Let us the trend in number of colonies and yield per colony in these 5 states over the year.

```
cplot1=sns.catplot(x='year', y='numcol',
            data=honeyprod[honeyprod["State"].isin(["North Dakota","California","South Dakota","Florida","Montana"])],
                estimator=sum, col='State', kind="point",
                height=3,col_wrap = 5)
cplot1.set_xticklabels(rotation=90)
plt.show()

# Following code look at the top 5 honey producing states in the US
# honeyprod.groupby(['State'])['totalprod'].mean().sort_values(ascending = False).reset_index().head()
```



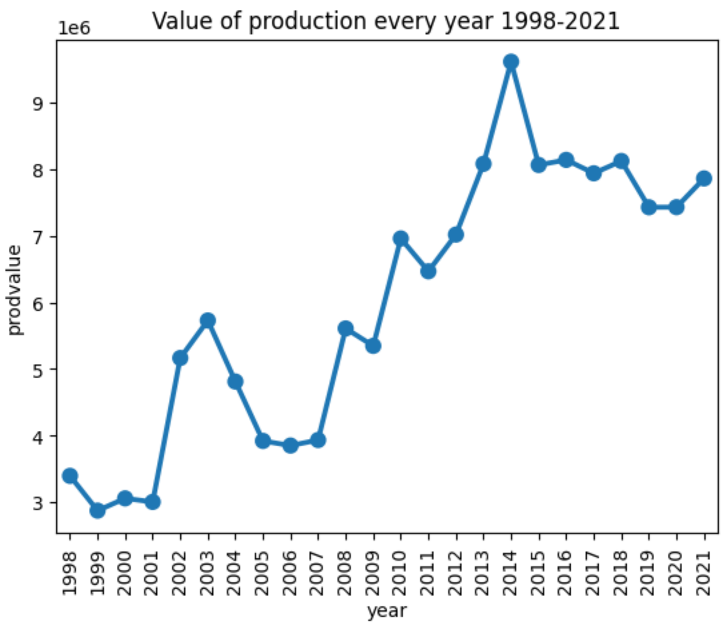
```
cplot2=sns.catplot(x='year', y='yieldpercol',
            data=honeyprod[honeyprod["State"].isin(["North Dakota","California","South Dakota","Florida","Montana"])],
                estimator=sum, col='State', kind="point",
                height=3,col_wrap = 5)
cplot2.set_xticklabels(rotation=90)
plt.show()
```



Observation:- 1)In North Dakota, the number of colonies has increased significantly over the years as compared to the other 4 states 2)If we check the yield per colony, it has been in an overall decreasing trend for all the 5 states over the years

Let us the declining production trend has had on the value of production over the every year.

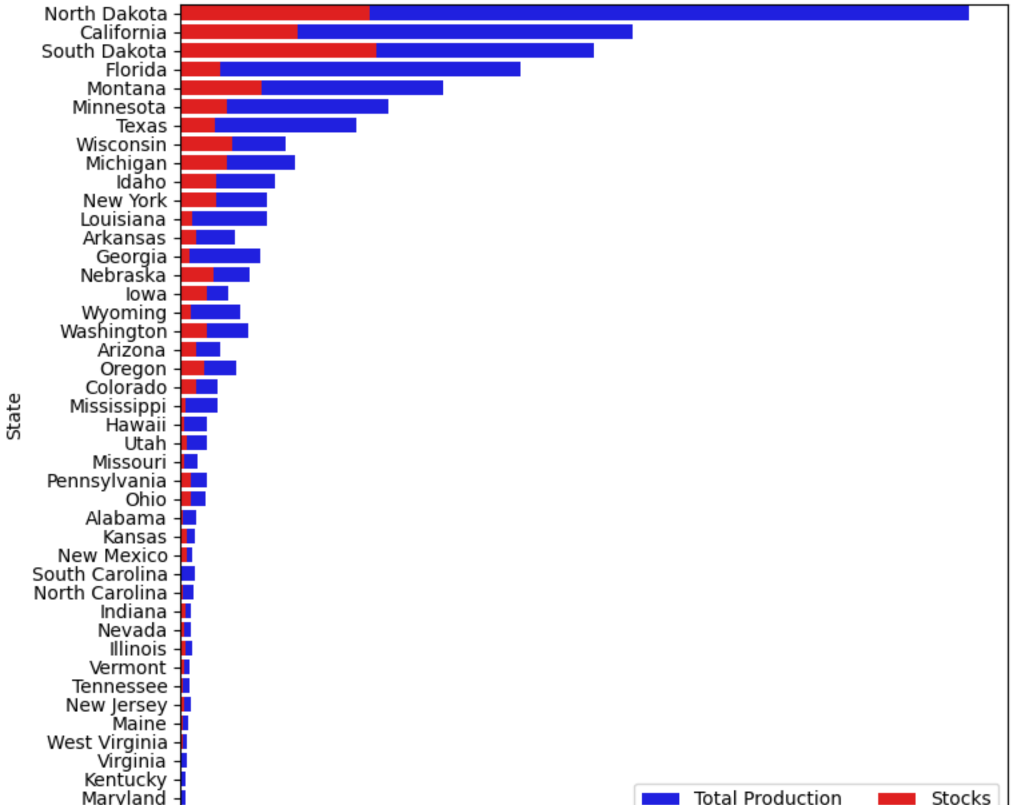
```
sns.pointplot(x="year", y="prodvalue", data=honeyprod, errorbar=None)
plt.xticks(rotation=90) # To rotate the x axis lables
plt.title('Value of production every year 1998-2021')
plt.show()
```



Observations:- 1)This is an interesting trend. As the total production has declined over the years, the value of production per pound has increased over time. 2)As the supply declined, the demand has added to the value of honey.

Let us check which of the states have been capitalising on this trend. We can compare the total production with the stocks .

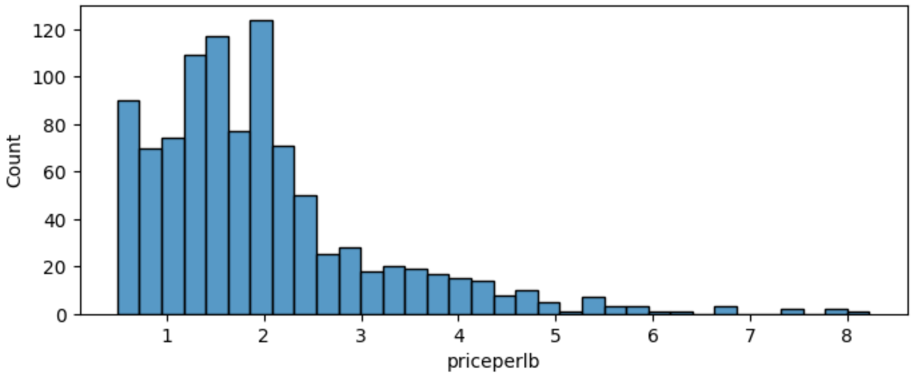
```
plt.figure(figsize = (8,8)) # To resize the plot

# Plot total production per state
sns.barplot(x="totalprod", y="State", data=honeyprod.sort_values("totalprod", ascending=False),
            label="Total Production", color="b", errorbar=None)

# Plot stocks per state
sns.barplot(x="stocks", y="State", data=honeyprod.sort_values("totalprod", ascending=False),
            label="Stocks", color="r", errorbar=None)

# Add a legend
plt.legend(ncol=2, loc="lower right", frameon=True)
plt.show()
```



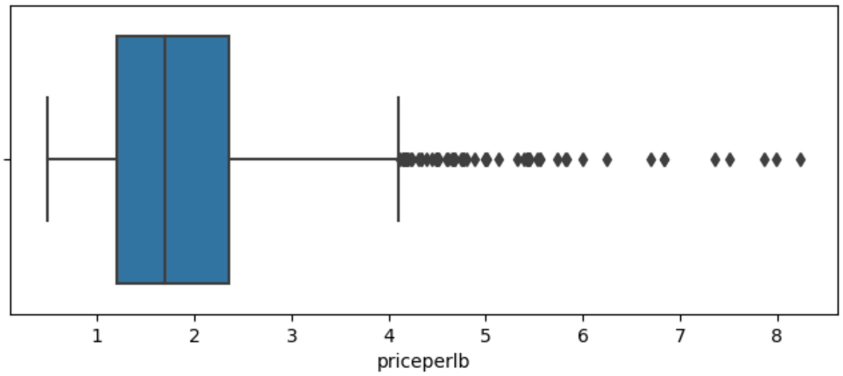
Observations:- 1)North Dakota has been able to sell more honey as compared to South Dakota despite having the highest production value. 2)Florida has the highest efficiency among the major honey producing states 3)Michigan is more efficient than Wisconsin in selling honey.

Let us look at the spread of average price of a pound of honey

```
plt.figure(figsize=(8, 3))
sns.histplot(honeyprod.priceperlb)
plt.show()
```



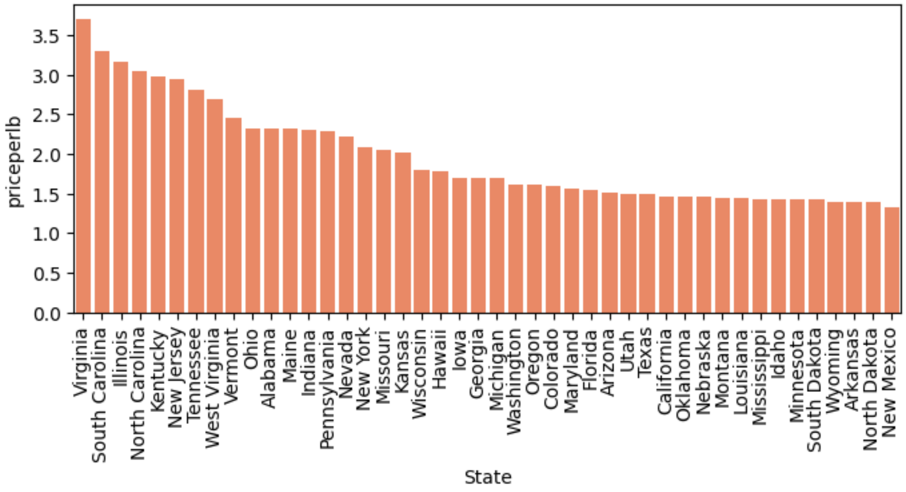
```
plt.figure(figsize=(8, 3))
sns.boxplot(data = honeyprod, x = 'priceperlb')
plt.show()
```



Observations:- 1)Price per pound of honey has a right skewed distribution with a lot of outliers towards the higher end. 2)The median price per pound of honey is 1.5 .

Let us look at the average price per pound of honey across states.

```
plt.figure(figsize=(8, 3)) # To resize the plot
sns.barplot(data = honeyprod, x = "State", y = "priceperlb", errorbar=None, color = "coral",
            order=honeyprod.groupby('State').priceperlb.mean().sort_values(ascending = False).index)
plt.xticks(rotation=90) # To rotate the x axis lables
plt.show()
```



Observations:- 1)Virginia has the highest price per pound of honey. 2)The average price per pound of honey in the major honey producing states is towards the lower end.

Conclusion:- 1)We can conclude that the total honey production has declined over the years whereas the value of production per pound has increased. 2)The reason for the declined honey production is the decrease in the yield per colony over the years. 3)The major honey producing states are California, Florida, North Dakota, South Dakota and Montana. 4)Among these, Florida has been very efficient in selling honey.