Answer Sheet – Advance Python [Major]

Name: Isha Garg

Phone No.: 7053017594

Email: ishagarg989@gmail.com

Importing Libraries

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

Load Data

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

Data Exploration

```
# basic info about data
df.info()


# check for missing values
df.isnull().sum()


# summary statistics of the data
print(df.describe())
```

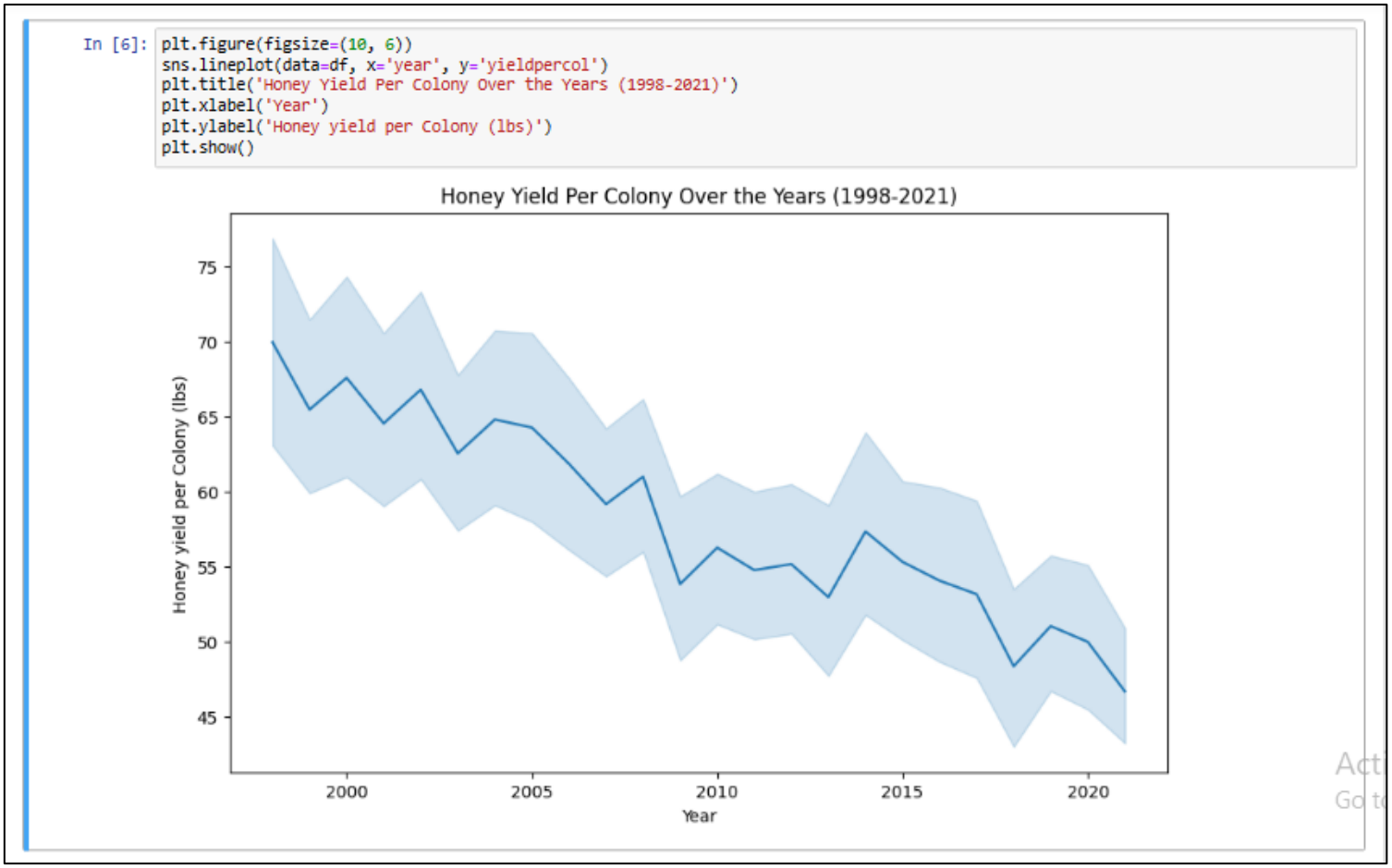
Key questions to be answered:

1. How has honey production yield changed from 1998 to 2021?

Code:

```
plt.figure(figsize=(10, 6))
sns.lineplot(data=df, x='year', y='yieldpercol')
plt.title('Honey Yield Per Colony Over the Years (1998-2021)')
plt.xlabel('Year')
plt.ylabel('Honey yield per Colony (lbs)')
plt.show()
```

Output:



Insights:

There is an overall decreasing trend in the yearly honey yield per colony from 1998 to 2021 with some fluctuations.

The highest honey yield per colony was in 1998

The lowest honey yield per colony was in 2021

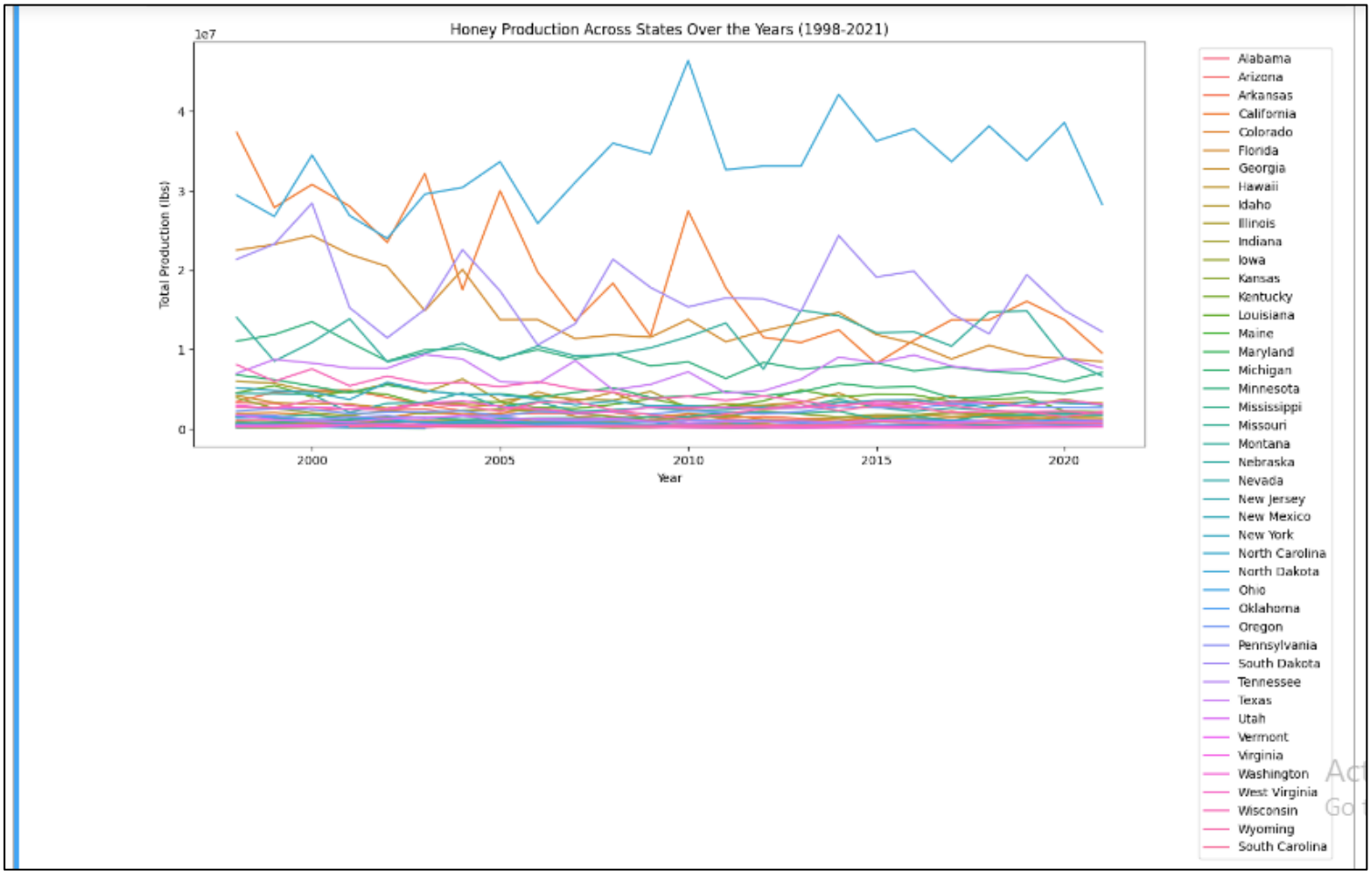
Overall, this chart gives us a clear idea of the trend and variations in the honey production yield over the years.

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

Code:

```
plt.figure(figsize=(14, 6))
sns.lineplot(df, x='year', y='totalprod', hue='State')
plt.title('Honey Production Across States Over the Years (1998-2021)')
plt.xlabel('Year')
plt.ylabel('Total Production (lbs)')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

Output:



Insights:

The top honey-producing state in the dataset is North Dakota, followed by South Dakota and California.

From the initial months of 2003 till 2012, North Dakota has been consistently the highest producer of honey.

The trend of honey production seems to be decreasing for most of the top 10 states over time, with North Dakota as the only exception.

North Dakota, South Dakota, California, Florida, Minnesota, Montana, Texas are the top 7 honey producing states.

The overall trend of honey production for California is different compared to the other states. Its production has fluctuated over the years, with some variations from year to year. Its production decreases with the greatest number of sharp dips and peaks in between

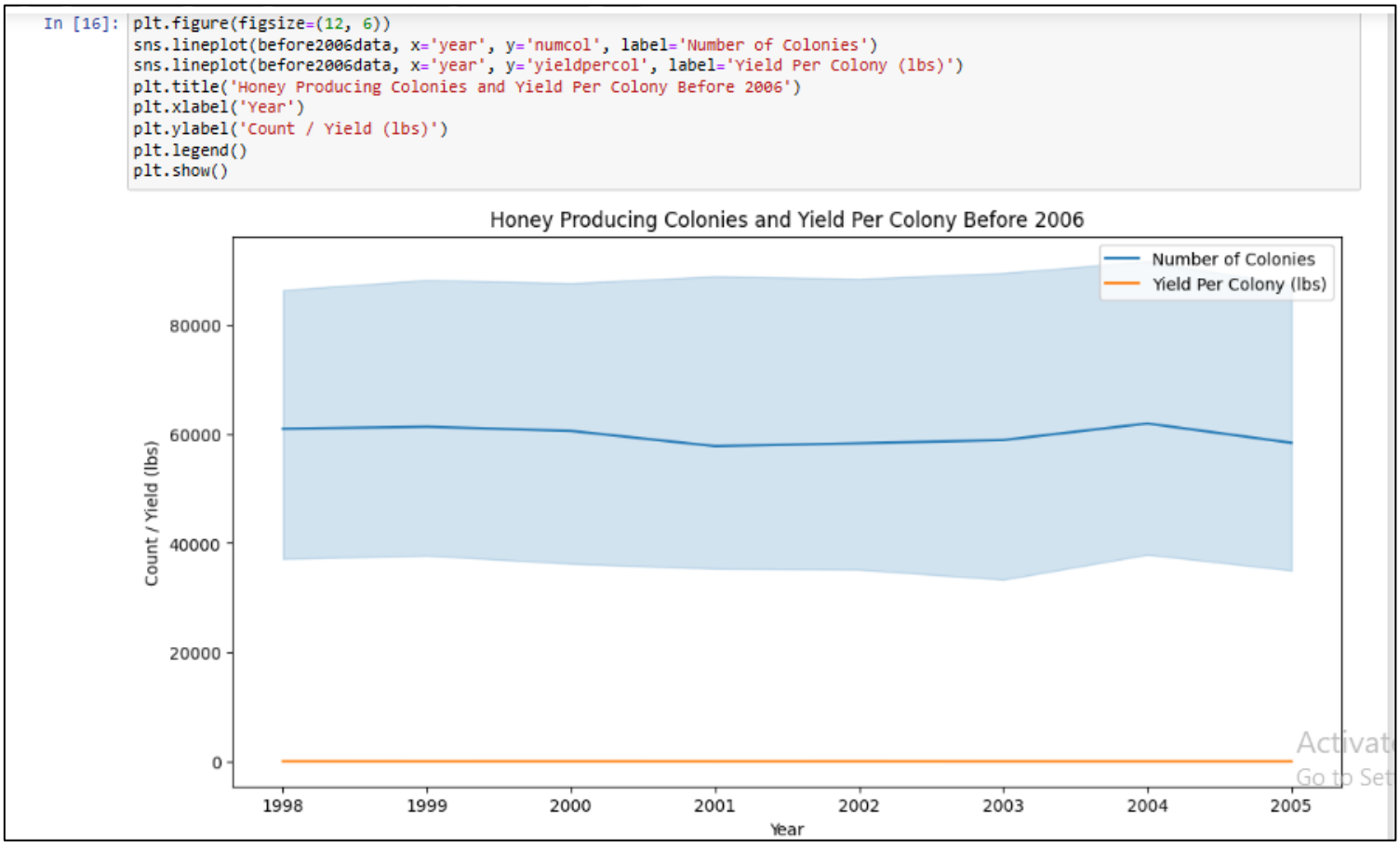
3. Does the data show any trends in terms of the number of honeys producing colonies and yield per colony before 2006, which was when concern over Colony Collapse Disorder spread nationwide?

Code:

```
# Data before 2006

before2006data = df[df['year'] < 2006]

before2006data['year'].unique()
```

```
plt.figure(figsize=(12, 6))

sns.lineplot(before2006data, x='year', y='numcol', label='Number of Colonies')

sns.lineplot(before2006data, x='year', y='yieldpercol', label='Yield Per Colony (lbs)')

plt.title('Honey Producing Colonies and Yield Per Colony Before 2006')

plt.xlabel('Year')

plt.ylabel('Count / Yield (lbs)')

plt.legend()

plt.show()
```

Output:



Insights:

Before 2006, Yield per colony is stable between 25,50,000 to 27,50,000 but Number of colonies has fluctuated, with some variations from year to year

Highest yield per colony before 2006 was in the year 1998, and the value was 3008

Lowest yield per colony before 2006 was in the year 2005, and the value was 2635

Highest number of colonies before 2006 was in the year 1999, and the value was 26,37,000

Lowest number of colonies before 2006 was in the year 2005, and the value was 23,94,000

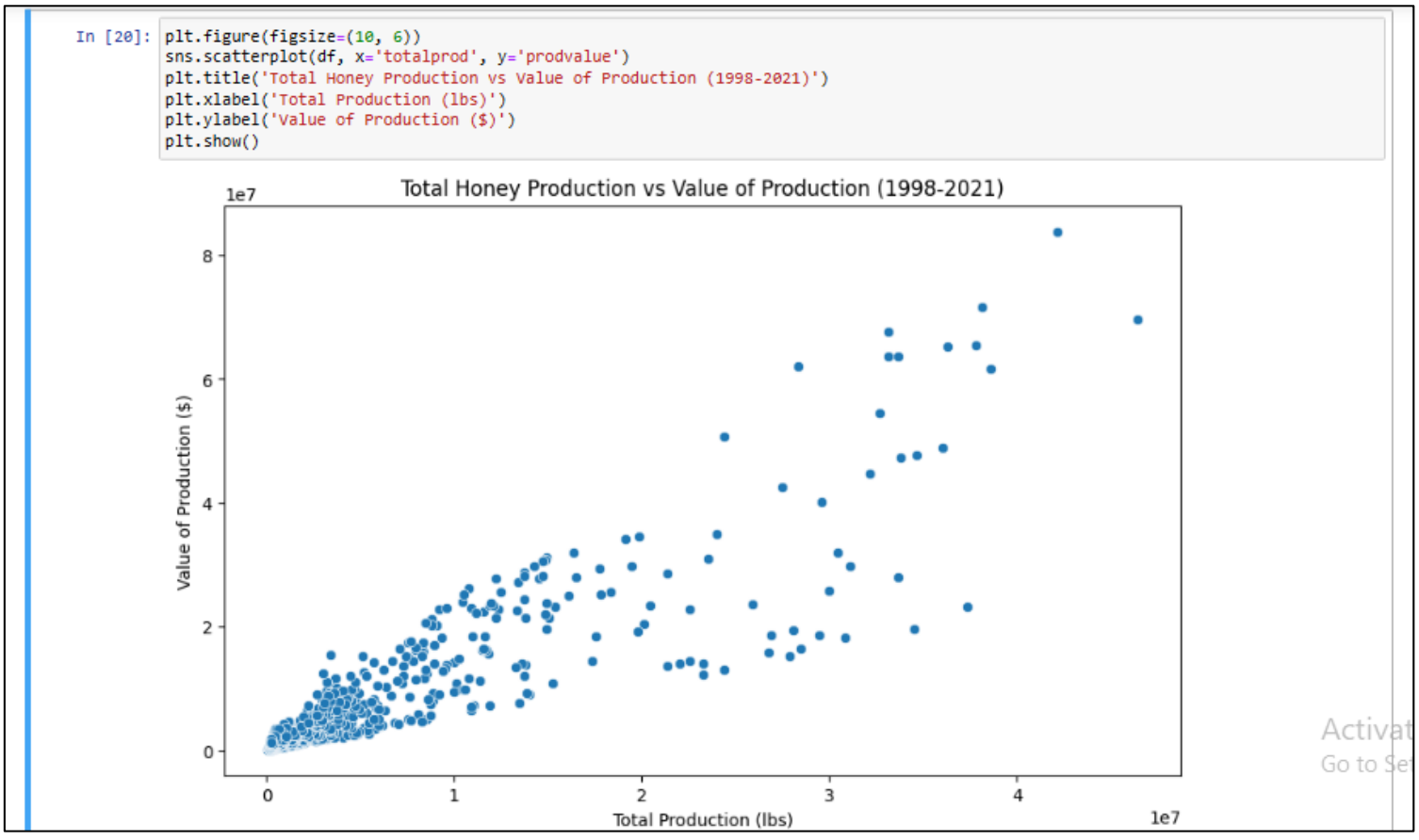
There was little bit fluctuation in number of colonies and some fluctuation in Yield per colony

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

Code:

```
plt.figure(figsize=(10, 6))

sns.scatterplot(df, x='totalprod', y='prodvalue')

plt.title('Total Honey Production vs Value of Production (1998-2021)')

plt.xlabel('Total Production (lbs)')

plt.ylabel('Value of Production ($)')

plt.show()
```

Output:



Insights:

There is a positive correlation between Total production and value of production, it means both the variables are changes in the same direction. When Total production keeps increasing and value of production keeps increasing too.

When total production increases then the value of production also increases and vice versa

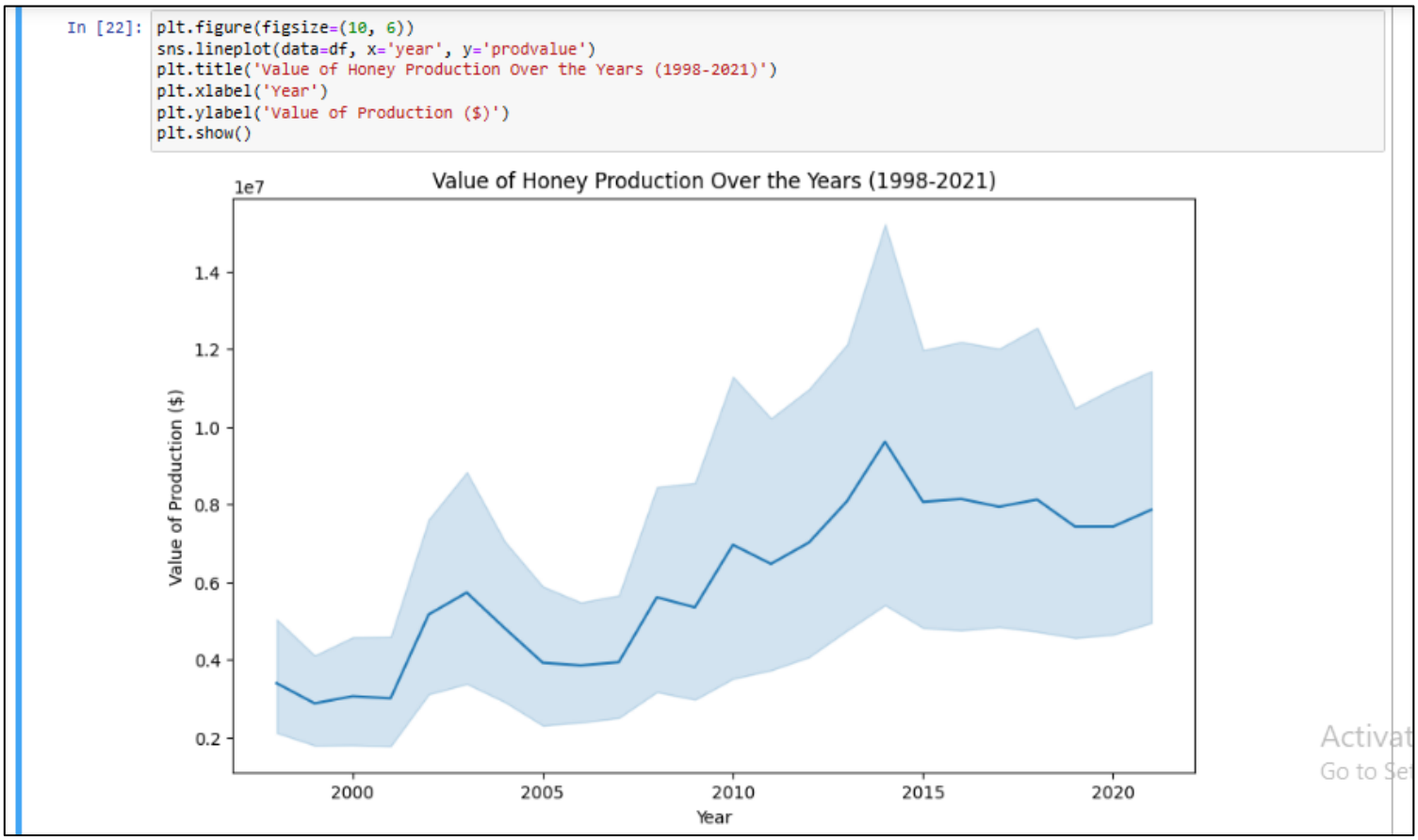
We can also observe an outlier point, which are much higher or much lower values.

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

Code:

```
plt.figure(figsize=(10, 6))

sns.lineplot(data=df, x='year', y='prodvalue')

plt.title('Value of Honey Production Over the Years (1998-2021)')

plt.xlabel('Year')

plt.ylabel('Value of Production ($)')

plt.show()
```

Output:



Insights:

The production value of honey in the US has generally increased over time, with a few notable dips and peaks.

The largest peaks occurred from 2013 to 2018.

Honey production is growing after 2013 with some minor fluctuations in between.

The highest production value was in 2014.

The lowest production value was in 1999.

Overall, the chart suggests that honey production is an important and growing industry in the United States.

6. Constructs the related plots using Seaborn and Matplot apply customization and derive insights from the visualization.

Code:

```
# Creating new Dataframe of numerical columns, (except state)
df1 = df.iloc[:,1:]


correlation_matrix = df1.corr()
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Heatmap')
plt.show()
```

Output: