notebook

June 25, 2024

1 Visualizing The Health of The Bees

Honey is a good source of food. Honey = Happy. That is why we need to examine and analyze the health of the bees because no bees = no honey. Aside from that, and most importantly, **we cant live without bees!**

1.1 Analyzing the health of the bee population.

Lets load the honey dataset found in this repository. In the honey.csv dataset, there is a column called numcol, which contains the number of bees in a colony

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
honey = pd.read_csv('../../data/honey.csv')
honey.head()
```

[1]:

	state	numcol	yieldpercol	totalprod	stocks	priceperlb	\
0	AL	16000.0	71	1136000.0	159000.0	0.72	
1	AZ	55000.0	60	3300000.0	1485000.0	0.64	
2	AR	53000.0	65	3445000.0	1688000.0	0.59	
3	CA	450000.0	83	37350000.0	12326000.0	0.62	
4	CO	27000.0	72	1944000.0	1594000.0	0.70	

	prodvalue	year
0	818000.0	1998
1	2112000.0	1998
2	2033000.0	1998
3	23157000.0	1998
4	1361000.0	1998

Let's create a scatterplot that shows the relationship between the number of bees in a colony and the state using sns.relplot

[2]:
```
sns.relplot(data = honey, x = "numcol", y = "state", height = 15, aspect= .5)
```

[2]:
```
<seaborn.axisgrid.FacetGrid at 0x7cf902688370>
```

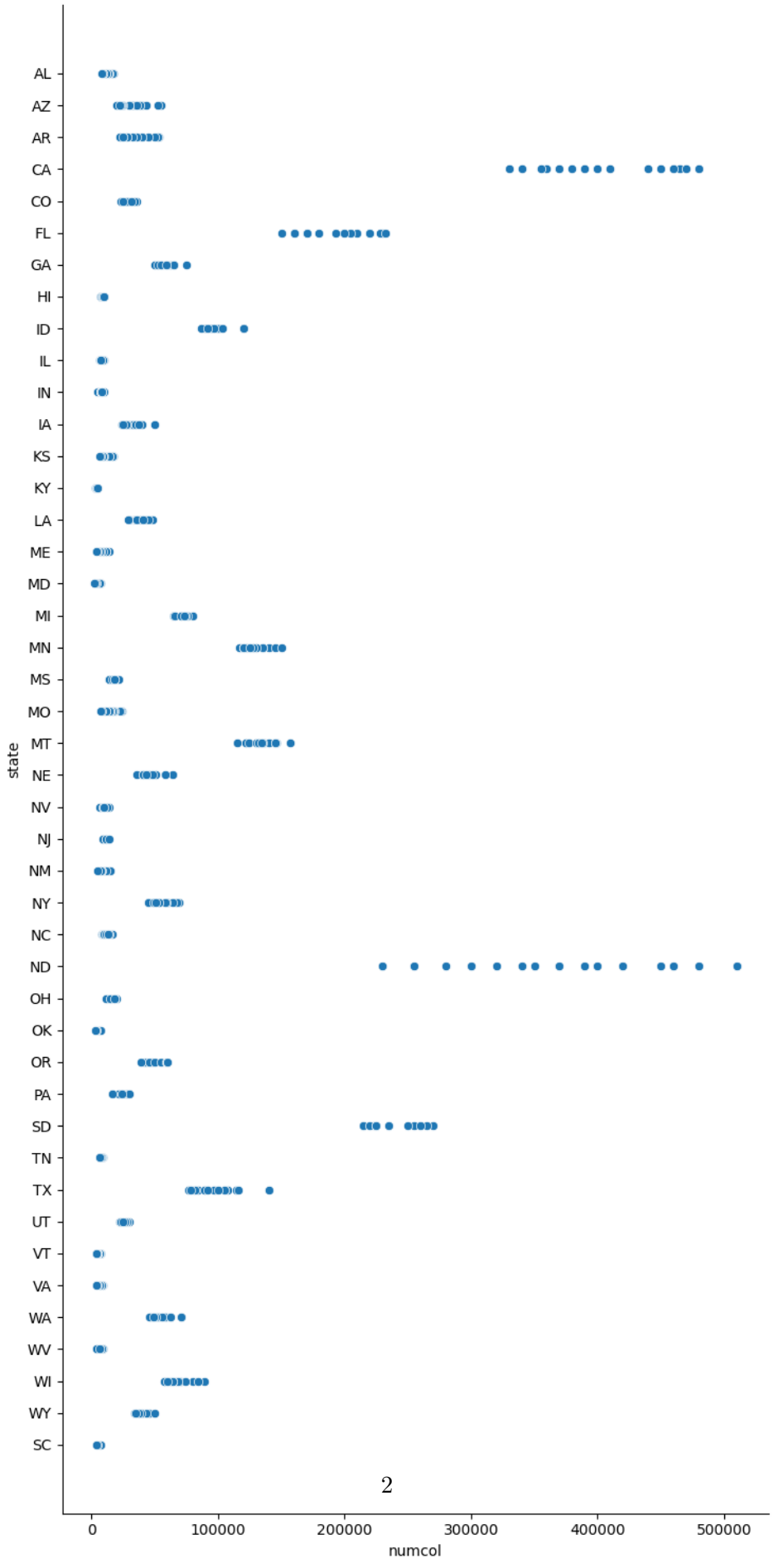

AL
AZ
AR
CA
CO
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
MI
MN
MS
MO
MT
NE
NV
NJ
NM
NY
NC
ND
OH
OK
OR
PA
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY
SC
state
0
100000
200000
300000
400000
500000
numcol

Figure 1

The plot only shows the distribution of the number of bees in a colony in each state. It does not show the accompanying date for each circle.

[6]:
```
sns.relplot(data = honey, x = "numcol", y = "state", height = 10, aspect= .5, ␣
↪hue = "year")
```

[6]:
```
<seaborn.axisgrid.FacetGrid at 0x7cf8ff99ece0>
```

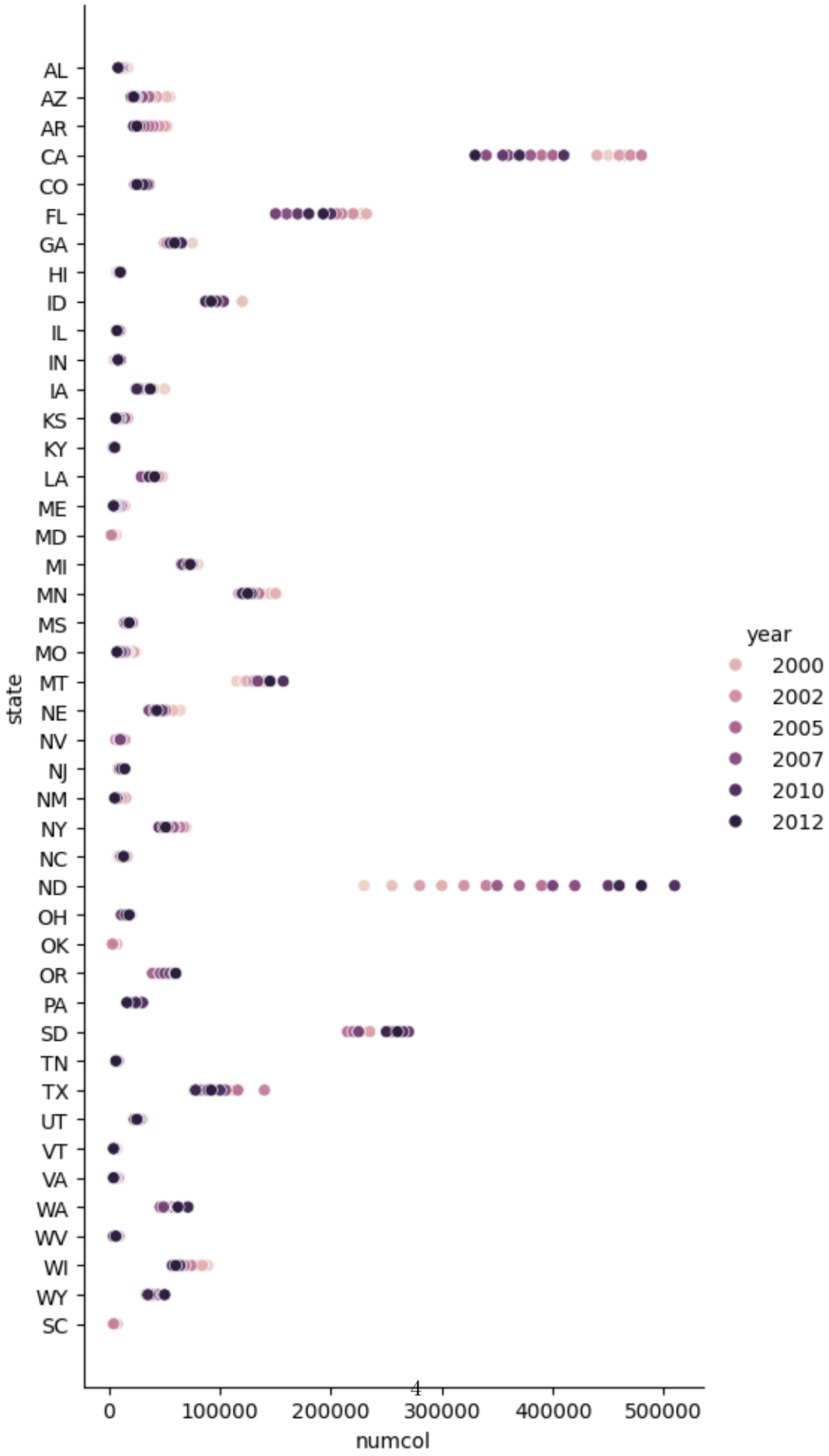

AL
AZ
AR
CA
CO
FL
GA
HI
ID
IL
IN
IA
KS
KY
LA
ME
MD
MI
MN
MS
MO
MT
NE
NV
NJ
NM
NY
NC
ND
OH
OK
OR
PA
SD
TN
TX
UT
VT
VA
WA
WV
WI
WY
SC
state
0
100000
200000
300000
400000
500000
numcol
year
2000
2002
2005
2007
2010
2012

Figure 2

We can see that the number of colonies over time differ as time passes by, either they are dwindling, or they are growing, as depicted by the hue that signifies the year. For example, the data for the number of colonies of bees in California in the year 2000 is much larger than the following years. This is very alarming!

We might want to look at the line chart of the number of bees in a colony in each state for each year.

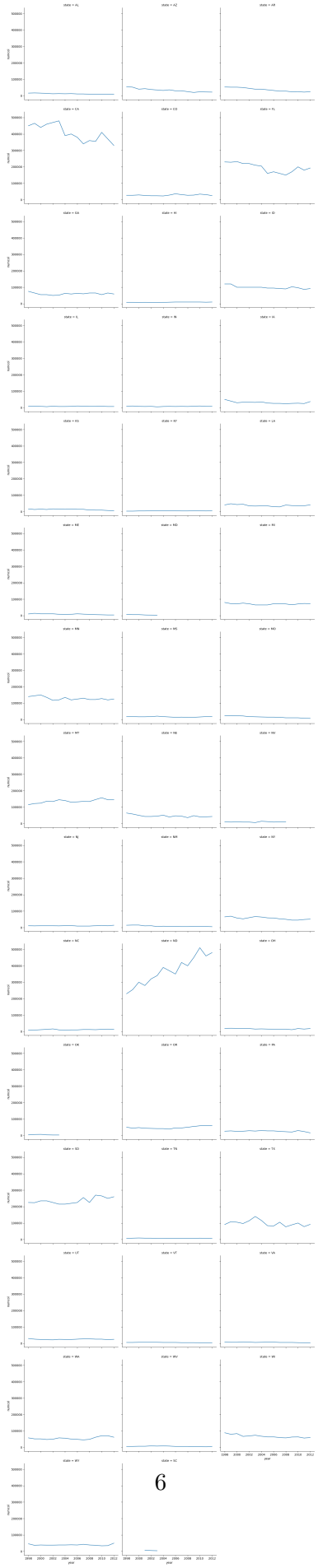
[12]:
```
len(pd.unique(honey['year']))
```

[12]:
```
15
```

since the number of unique states is 15, we can use facit grid to plot all of them.

[14]:
```
sns.relplot(
    data = honey,
    x = "year",
    y = "numcol",
    col = "state",
    col_wrap = 3,
    kind = "line"
)
```

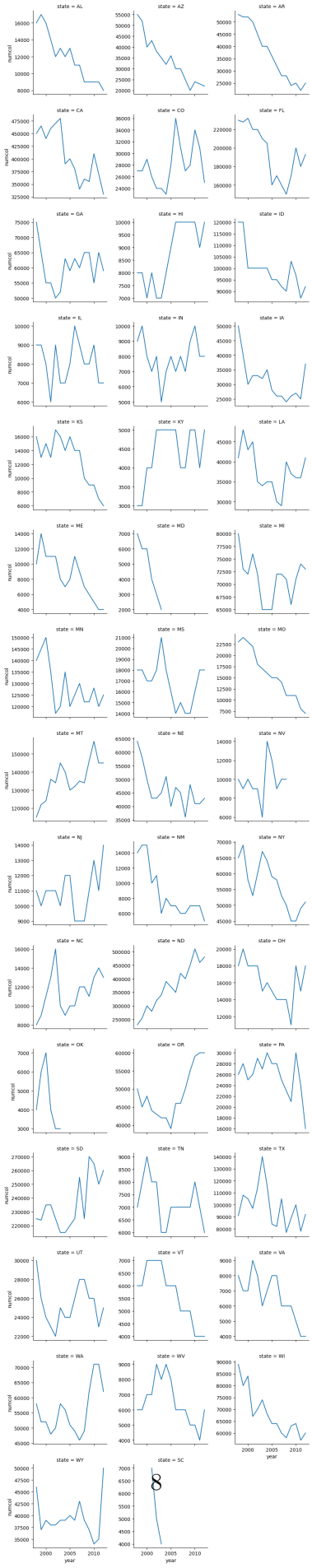
[14]:
```
<seaborn.axisgrid.FacetGrid at 0x7cf8f7a220e0>
```


state = AL
state = AZ
state = AR
state = CA
state = CO
state = FL
state = GA
state = HI
state = ID
state = IL
state = IN
state = IA
state = KS
state = KY
state = LA
state = ME
state = MD
state = MI
state = MN
state = MS
state = MO
state = MT
state = NE
state = NV
state = NJ
state = NM
state = NY
state = NC
state = ND
state = OH
state = OK
state = OR
state = PA
state = SD
state = TN
state = TX
state = UT
state = VT
state = VA
state = WA
state = WV
state = WI
state = WY
state = SC
count
year

We can see that the plot is not really helpful for some state as the y-axis is scaled from 0 to 500000, leaving other states with relatively low changes in the number of colonies look like it is not changing much. Let's change that.

[21]:
```
g = sns.FacetGrid(
    data = honey,
    col = "state",
    col_wrap = 3,
    sharey=False
)
g.map_dataframe(sns.lineplot, x = "year", y = "numcol")
```

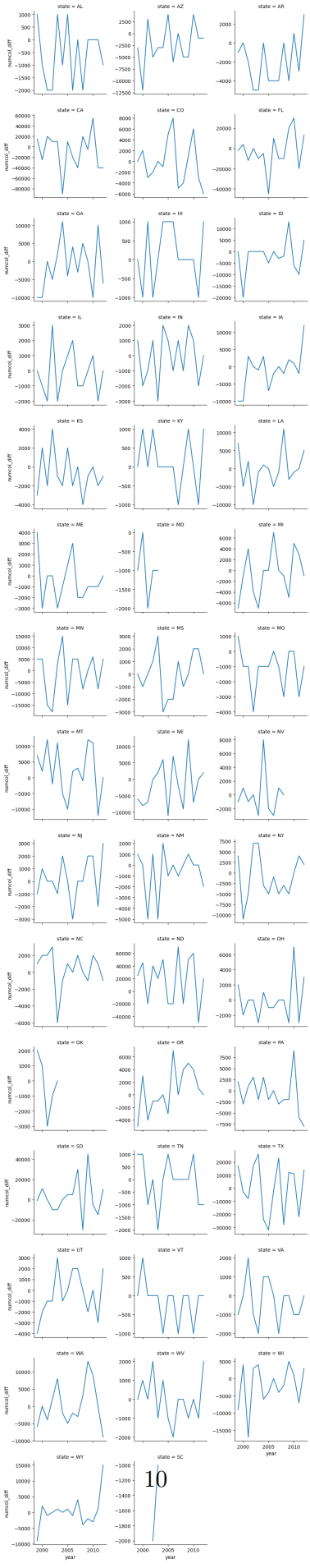
[21]:
```
<seaborn.axisgrid.FacetGrid at 0x7cf8f0722950>
```


state = AL
state = AZ
state = AR
state = CA
state = CO
state = FL
state = GA
state = HI
state = ID
state = IL
state = IN
state = IA
state = KS
state = KY
state = LA
state = ME
state = MD
state = MI
state = MN
state = MS
state = MO
state = MT
state = NE
state = NV
state = NJ
state = NM
state = NY
state = NC
state = ND
state = OH
state = OK
state = OR
state = PA
state = SD
state = TN
state = TX
state = UT
state = VT
state = VA
state = WA
state = WV
state = WI
state = WY
state = SC
numcol
year

We observe that many states show a decline in the number of bee colonies over time. To perform a more detailed analysis, let's plot the year-over-year difference in the number of bee colonies for each state.

[24]:
```
honey['numcol_diff'] = honey.groupby('state')['numcol'].diff()
g = sns.FacetGrid(
    data = honey,
    col = "state",
    col_wrap = 3,
    sharey=False
)
g.map_dataframe(sns.lineplot, x = "year", y = "numcol_diff")
```

[24]:
```
<seaborn.axisgrid.FacetGrid at 0x7cf8ee3f84f0>
```

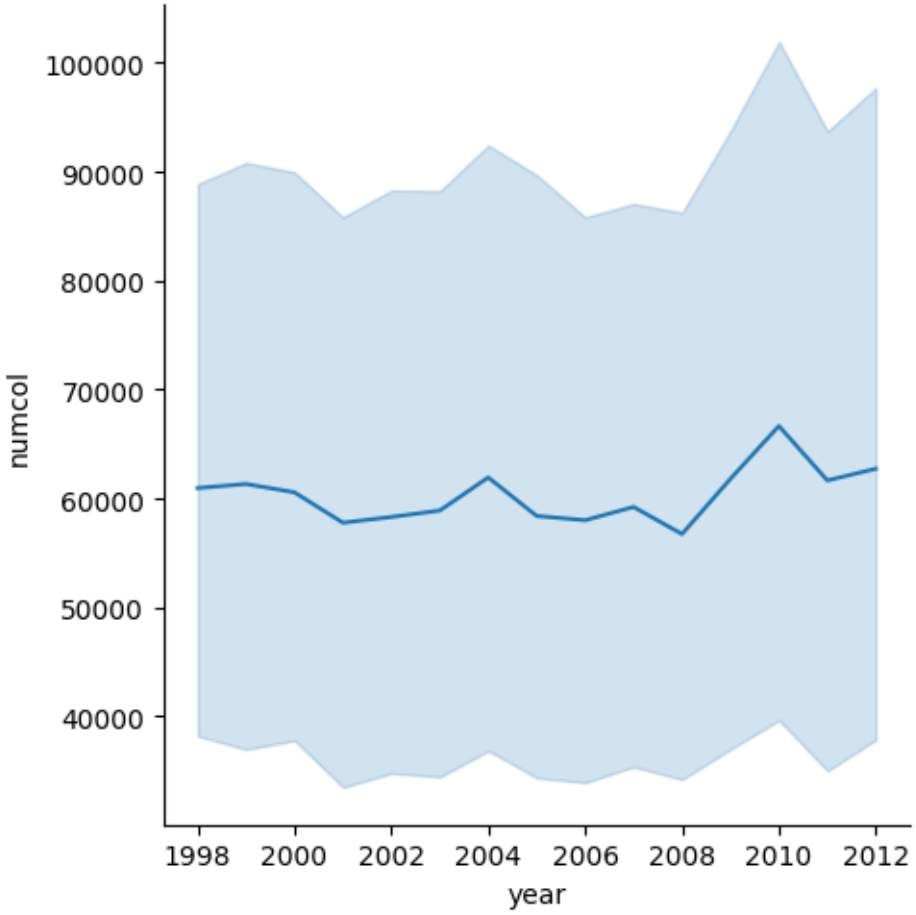

state = AL
state = AZ
state = AR
state = CA
state = CO
state = FL
state = GA
state = HI
state = ID
state = IL
state = IN
state = IA
state = KS
state = KY
state = LA
state = ME
state = MD
state = MI
state = MN
state = MS
state = MO
state = MT
state = NE
state = NV
state = NJ
state = NM
state = NY
state = NC
state = ND
state = OH
state = OK
state = OR
state = PA
state = SD
state = TN
state = TX
state = UT
state = VT
state = VA
state = WA
state = WV
state = WI
state = WY
state = SC
numcol_diff
year

We can see a lot of states having negative rate of change in the number of bee colonies over time. That is indeed bad!

1.1.1 Analyzing the average number of bees in a colony for each year.

[15]:
```
sns.relplot(data = honey, x = "year", y = "numcol", kind = "line")
```

[15]:
```
<seaborn.axisgrid.FacetGrid at 0x7cf8f3cda0e0>
```



Because Seaborn is aggregating data around one line, it displays "the multiple measurements at each x value by plotting the mean and the 95% confidence interval around the mean". We can see that the confidence interval (depicted by the light blue shaded region) is very wide. This is important as it tell us how confident we are in our estimate of the mean (the blue line), which is not very good. This is evident from figure 2, showing different values and the difference is massive.