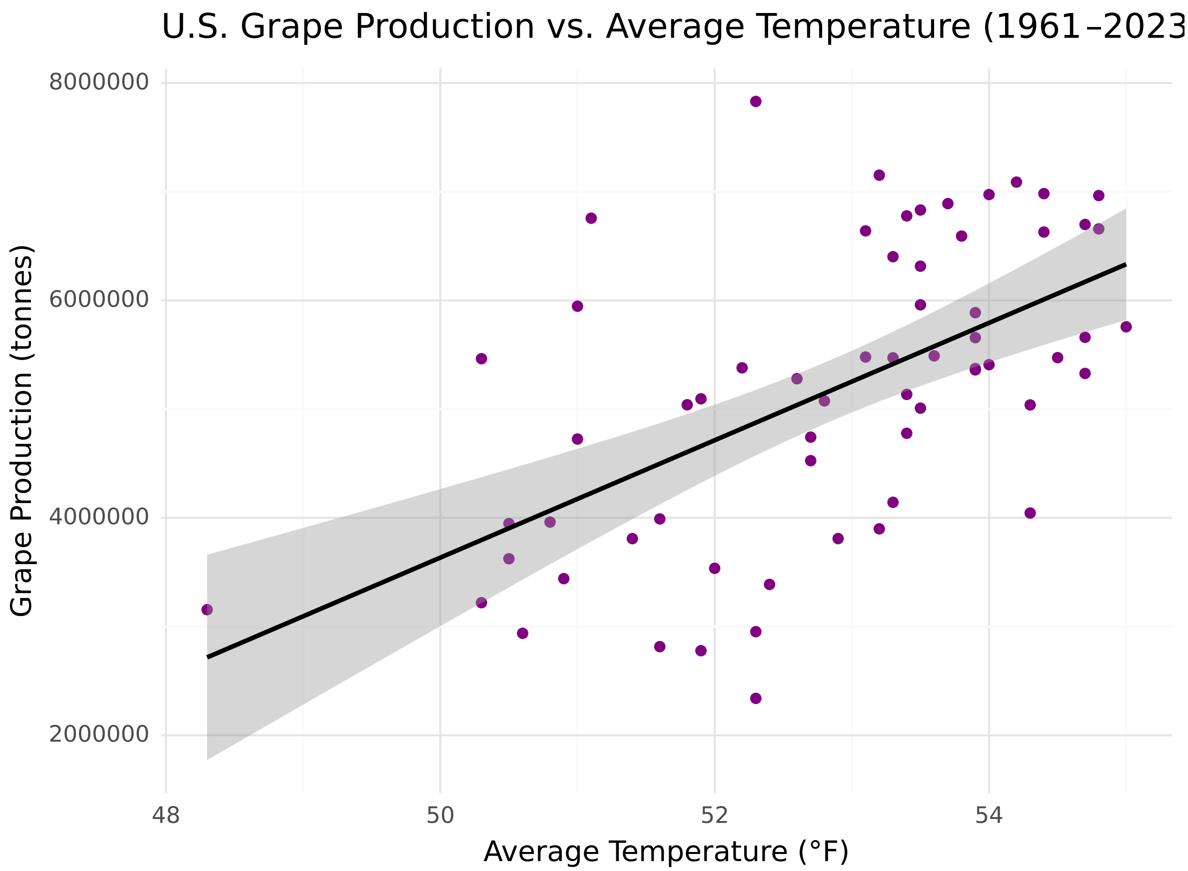
Wine Project

Sergio Martinez Barajas Sarah Marie Jennings

2025-12-01

```
grapes = pd.read_csv("us-grape-production.csv")
temps = pd.read_csv("filtered-temp.csv")

grapes = grapes.rename(columns={
"Grapes | 00000560 || Production | 005510 || tonnes": "Grape_Production_tonnes"
})

grapes = grapes[(grapes["Year"] >= 1961) & (grapes["Year"] <= 2023)]
temps = temps[(temps["Year"] >= 1961) & (temps["Year"] <= 2023)]

merged = pd.merge(grapes, temps, on="Year")
(
ggplot(merged, aes(x="Average_Fahrenheit_Temperature", y="Grape_Production_tonnes")) +
geom_point(color="purple") +
geom_smooth(method="lm", color="black") +
labs(
title="U.S. Grape Production vs. Average Temperature (1961-2023)",
x="Average Temperature (°F)",
y="Grape Production (tonnes)"
) +
theme_minimal()
)
```




```
#Correlation analysis
corr = merged["Average_Fahrenheit_Temperature"].corr(merged["Grape_Production_tonnes"])
print("Correlation:", corr)
```

```
Correlation: 0.5712335660023041
```

what the correlation means is that there tends to be possitive correlation between average yealy temeprture and grape yield, but it's not very strong.

```
#linear regression
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import numpy as np

X = merged["Average_Fahrenheit_Temperature"].values.reshape(-1, 1)
Y = merged["Grape_Production_tonnes"].values
```

```
X_train, X_test, Y_train, Y_test = train_test_split( X, Y, test_size = 0.2, shuffle = True, r

model = LinearRegression()
model.fit(X_train, Y_train)

Y_pred = model.predict(X_test)

mae = mean_absolute_error(Y_test, Y_pred)
rmse = np.sqrt(mean_squared_error(Y_test, Y_pred))
r2 = r2_score(Y_test, Y_pred)

print("MAE: ", mae)
print("RMSE: ", rmse)
print("R^2: ", r2)
```

```
MAE:  1019018.7867519022
RMSE:  1162452.7731372928
R^2:  0.32537166473166856
```

slope means that for every 1 F increase we get 540k tonnes increase yield but propably because temperature is average across all US, grape production is national total, there are other possible factors at work R^2 = 0.326 means that temperture explains 32% of the variation in grape yields, which means that a whole 2/3rds of variablility is caused by something else.

Both temp and yield increase overtime so this might make false impressions. It's not that more heat = more grapes, it could be more time = more grapes and more heat. So if we detrend the data to see if the correlation still exisits.

```
merged["Year_centered"] = merged["Year"] - merged["Year"].mean()

# detrend
from sklearn.linear_model import LinearRegression

# detrend temperature
m1 = LinearRegression().fit(merged[["Year_centered"]], merged["Average_Fahrenheit_Temperature
merged["Temp_detrended"] = merged["Average_Fahrenheit_Temperature"] - m1.predict(merged[["Yea

# detrend yield
m2 = LinearRegression().fit(merged[["Year_centered"]], merged["Grape_Production_tonnes"])
merged["Yield_detrended"] = merged["Grape_Production_tonnes"] - m2.predict(merged[["Year_cen
```

```python
merged["Temp_detrended"].corr(merged["Yield_detrended"])
```

```
-0.005950884596132765
```

After detrending the data we see that correlation is bascially 0. So temp and grape yield have almost nothing to do with each other.

```python
####################################################
#### From here on out we're only using Cal Data ####
####################################################

grapes = pd.read_csv("Californa_Wine_Production_1980_2020.csv")
temp = pd.read_csv("California-avg-temp-1980-2021.csv")
rain = pd.read_csv("California-rain-1980-2021.csv")

# trim to overlapping years
grapes = grapes[(grapes["Year"] >= 1980) & (grapes["Year"] <= 2020)]
temp = temp[(temp["Year"] >= 1980) & (temp["Year"] <= 2020)]
rain = rain[(rain["Year"] >= 1980) & (rain["Year"] <= 2020)]

# merge
merged = grapes.merge(temp, on="Year").merge(rain, on="Year")

merged = merged.rename(columns={
    "Temp": "AvgTemp_F",
    "rain_in": "Rain_in"
})
merged[["Yield(Unit/Acre)", "AvgTemp_F", "Rain_in"]].corr()
```

	Yield(Unit/Acre)	AvgTemp_F	Rain_in
Yield(Unit/Acre)	1.000000	0.022516	0.002424
AvgTemp_F	0.022516	1.000000	-0.267420
Rain_in	0.002424	-0.267420	1.000000

What this means is that temp and percipitaion have almost no impact on yield. Although let it be known that in our lit review it did say that percipitaion almost never has an impact on yields, so that's at least 1 thing proven. Temp and rain have a negative correlation, when it rains it's colder and when hot there's less rain, but it's really weak

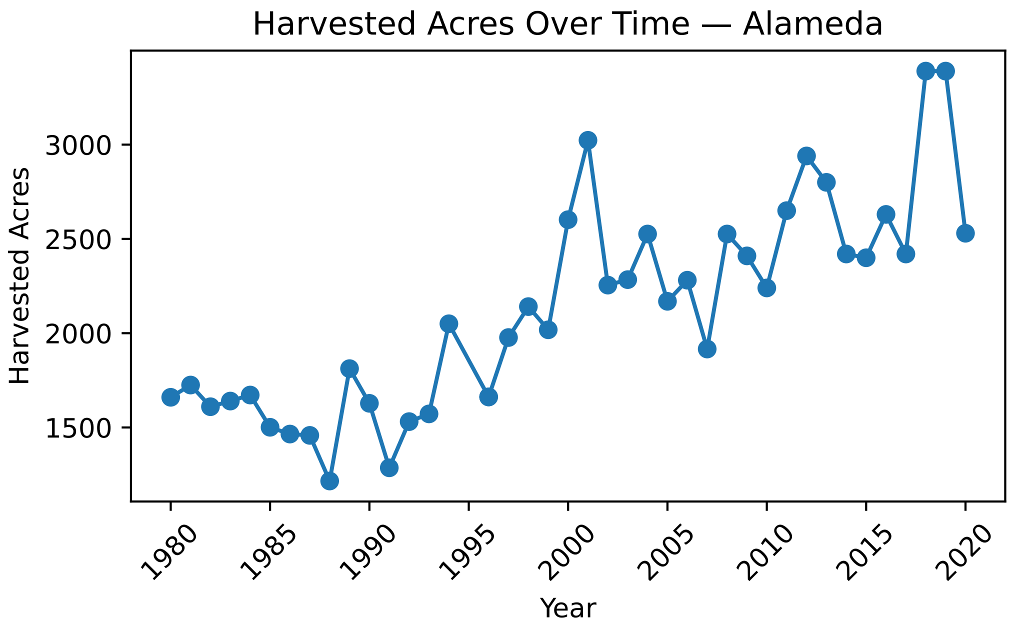
```python
county_stats = merged.groupby("County").agg({
    "Yield(Unit/Acre)": "mean",
    "HarvestedAcres": "mean",
    "AvgTemp_F": "mean",
    "Rain_in": "mean"
}).sort_values(by="HarvestedAcres", ascending=False)

print(county_stats)
```

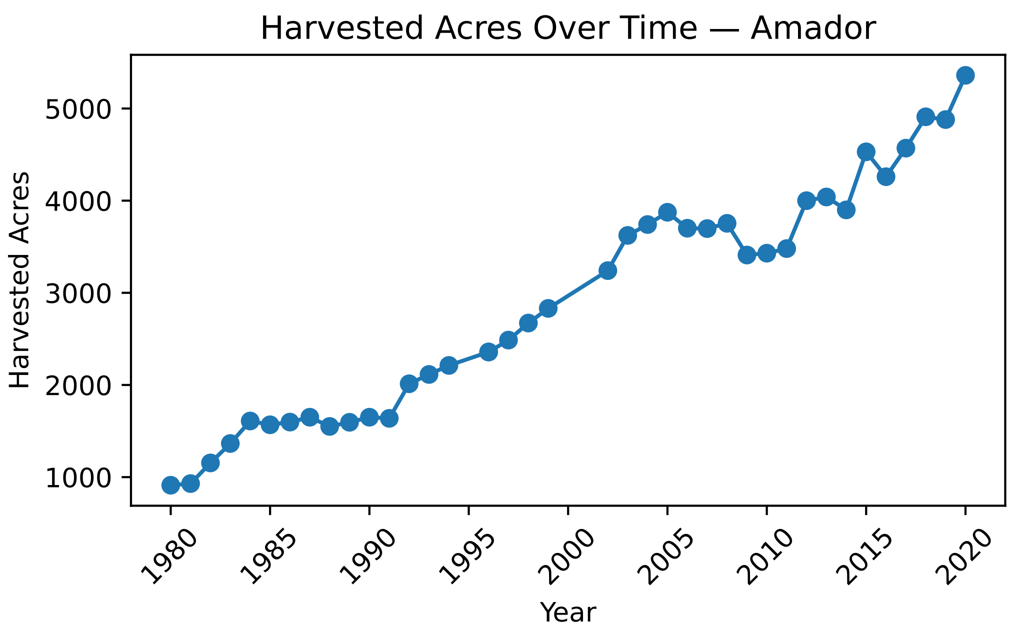
```
                Yield(Unit/Acre)  HarvestedAcres  AvgTemp_F    Rain_in
County
SanJoaquin              6.849268    70829.780488  58.717073  22.571463
Fresno                 10.505854    61157.951220  58.717073  22.571463
Madera                  9.824634    46011.853659  58.717073  22.571463
Sonoma                  3.847073    42911.219512  58.717073  22.571463
SanLuisObisp            4.100000    36600.000000  57.500000  20.630000
Napa                    3.647073    35201.341463  58.717073  22.571463
Kern                    9.046341    35094.560976  58.717073  22.571463
Monterey                4.161389    35093.777778  58.830556  22.370000
SanLuisObispo           4.308889    22972.861111  58.766667  22.243333
Sacramento              7.032683    18104.365854  58.717073  22.571463
Tulare                 11.260244    17996.536585  58.717073  22.571463
Stanislaus              8.937619    15855.428571  58.428571  23.613333
SantaBarbara            3.540556    14581.805556  58.730556  22.233333
Mendocino               4.051282    13917.717949  58.746154  22.495897
Merced                  9.039024    13003.463415  58.717073  22.571463
Yolo                    6.908537     7654.000000  58.717073  22.571463
Lake                    4.030000     5392.731707  58.717073  22.571463
SanBenito               3.976098     3223.390244  58.717073  22.571463
Amador                  3.440256     2888.820513  58.741026  22.727949
Solano                  4.739756     2855.853659  58.717073  22.571463
Kings                  11.305610     2615.341463  58.717073  22.571463
Riverside               3.841951     2266.341463  58.717073  22.571463
Alameda                 3.678049     2124.097561  58.717073  22.571463
Colusa                  8.633000     1896.700000  59.500000  21.298000
SanBernardino           2.461951     1791.341463  58.717073  22.571463
ContraCosta             3.699091     1616.818182  59.018182  24.374545
SantaClara              3.150833     1615.416667  58.711111  21.843889
ElDorado                3.022727     1458.424242  58.806061  23.114848
Glenn                   6.750000      875.000000  57.750000  29.885000
Calaveras               2.694000      443.200000  58.737500  22.665750
SantaCruz               2.090833      422.305556  58.791667  22.632222
```

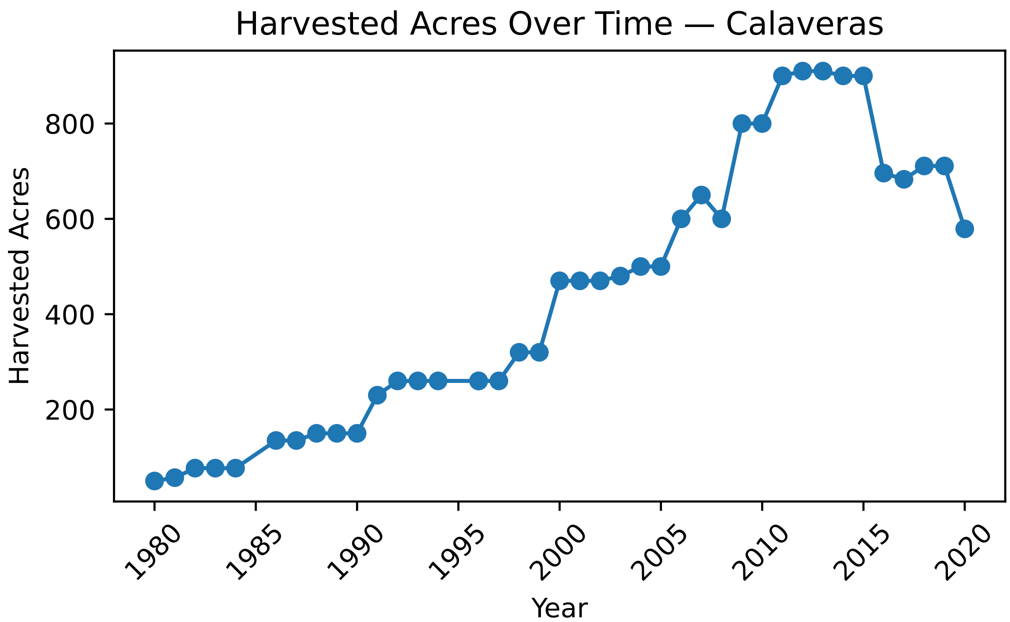
SanDiego	2.550244	416.097561	58.717073	22.571463
Nevada	4.037586	309.533333	58.744737	22.401316
Placer	2.457500	190.392857	58.864286	21.974643
Yuba	2.008571	172.500000	58.607143	22.031429
Shasta	2.358333	167.500000	59.425000	20.604167
Marin	1.485455	154.409091	59.027273	21.613182
Tehama	4.250000	147.000000	59.700000	22.506667
SanMateo	NaN	101.962963	59.014286	22.725000
Trinity	2.130000	94.687500	58.910000	22.110000
Mariposa	1.264000	80.400000	58.950000	21.981000
Mono	13.755000	14.250000	58.250000	22.597500

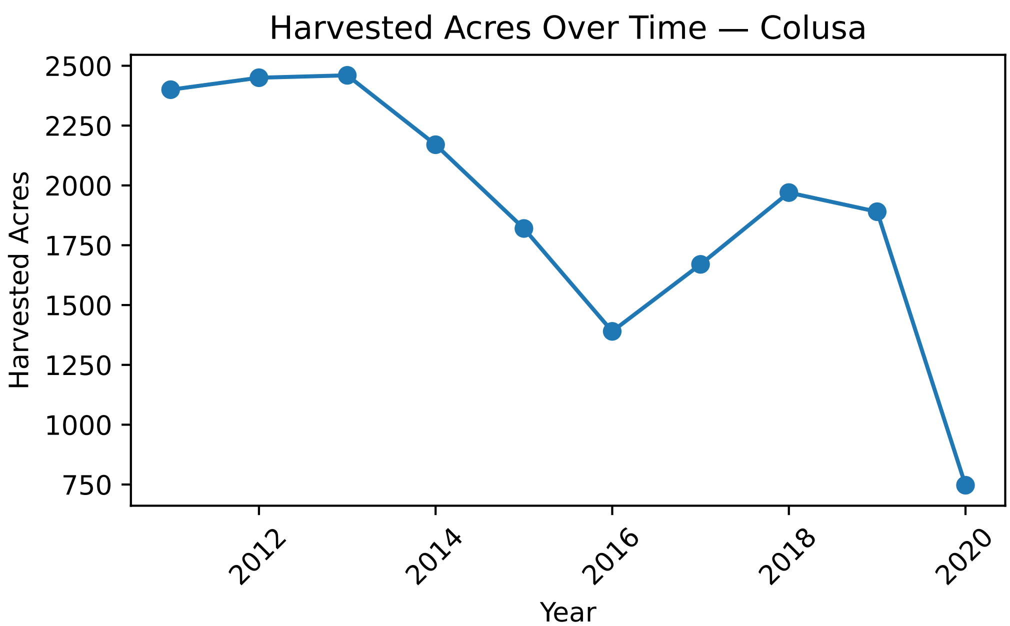
So right here we're looking at the counties that harvest the most winegrapes. Doesn't really tell us anything

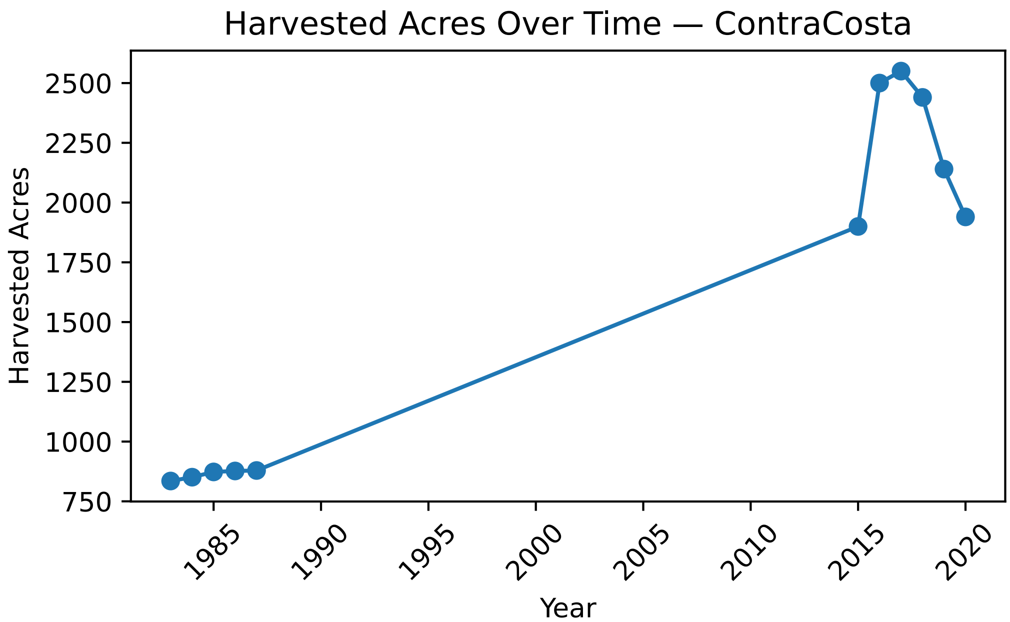
```
for county in merged["County"].unique():
    sub = merged[merged["County"] == county]

    if len(sub) > 1:  # must have multiple years to plot
        plt.figure()
        plt.plot(sub["Year"], sub["HarvestedAcres"], marker="o")
        plt.title(f"Harvested Acres Over Time - {county}")
        plt.xlabel("Year")
        plt.ylabel("Harvested Acres")
        plt.xticks(rotation=45)
        plt.tight_layout()
        plt.show()
```

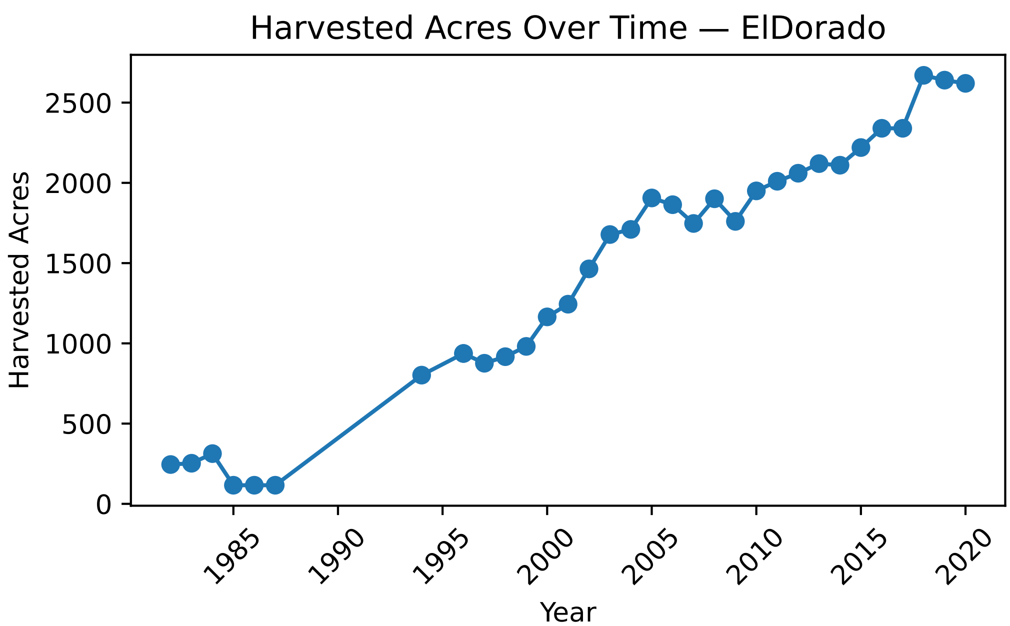

Harvested Acres Over Time — Alameda
Harvested Acres
3000
2500
2000
1500
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

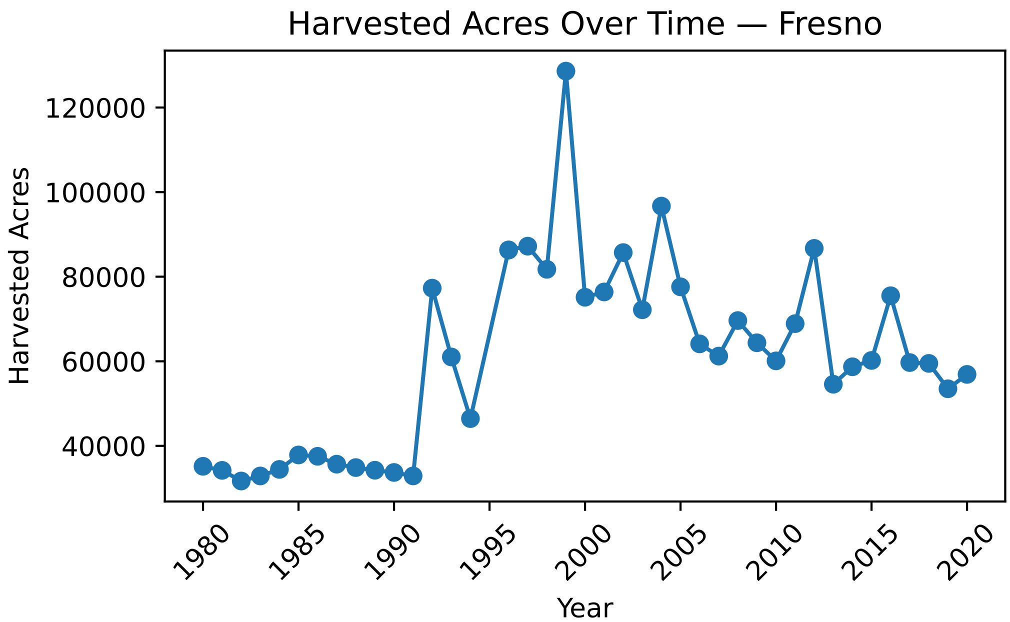

Harvested Acres Over Time — Amador
Harvested Acres
5000
4000
3000
2000
1000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

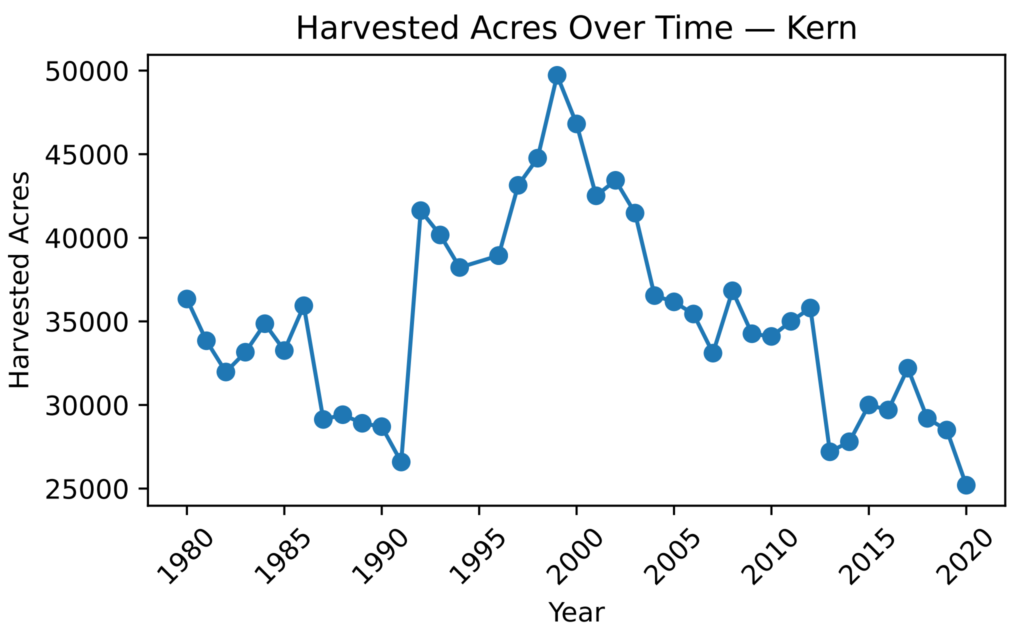

Harvested Acres Over Time — Calaveras
Harvested Acres
200
400
600
800
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

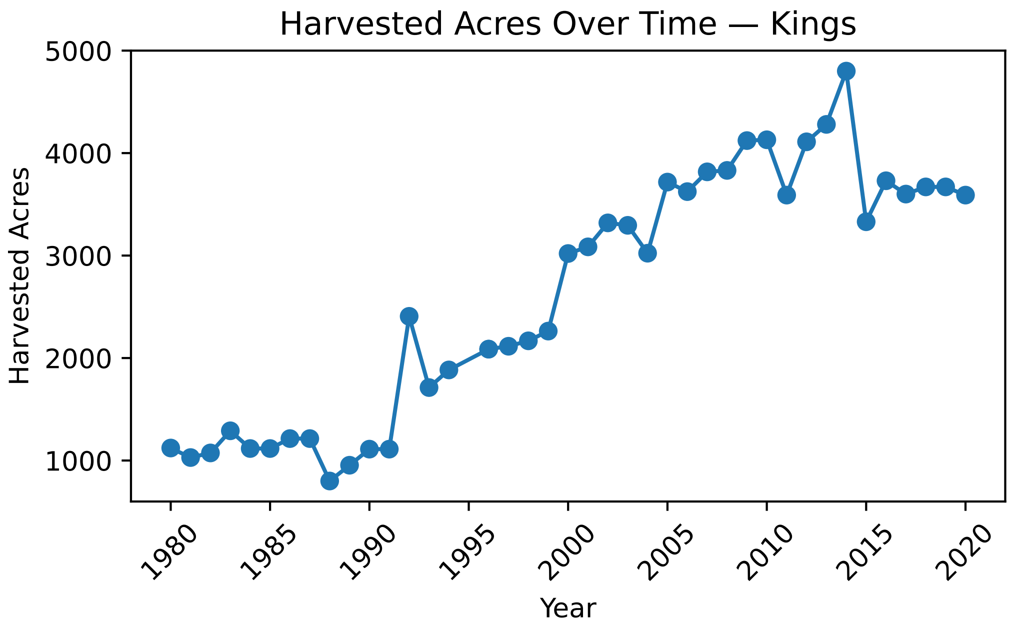

Harvested Acres Over Time — Colusa
Harvested Acres
750
1000
1250
1500
1750
2000
2250
2500
2012
2014
2016
2018
2020
Year

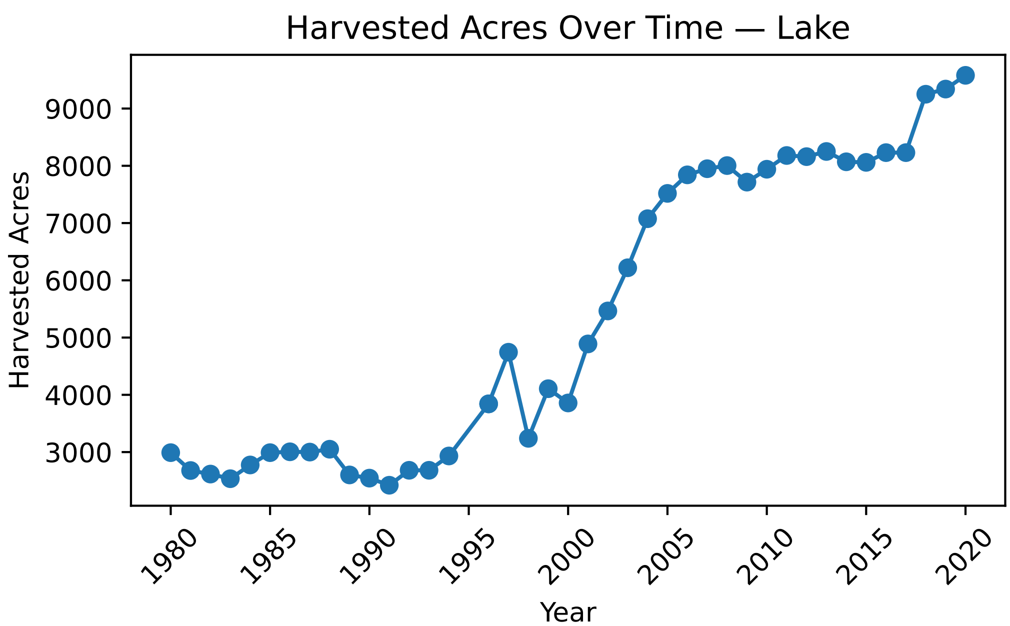

Harvested Acres Over Time — ContraCosta
Harvested Acres
2500
2250
2000
1750
1500
1250
1000
750
1985
1990
1995
2000
2005
2010
2015
2020
Year

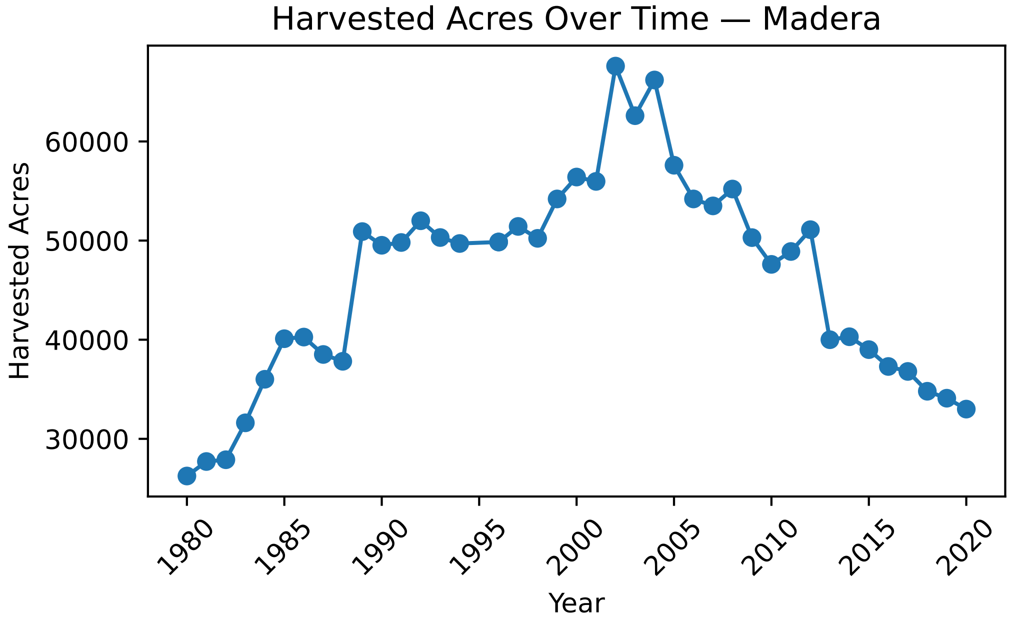

Harvested Acres Over Time — ElDorado
Harvested Acres
2500
2000
1500
1000
500
0
1985
1990
1995
2000
2005
2010
2015
2020
Year

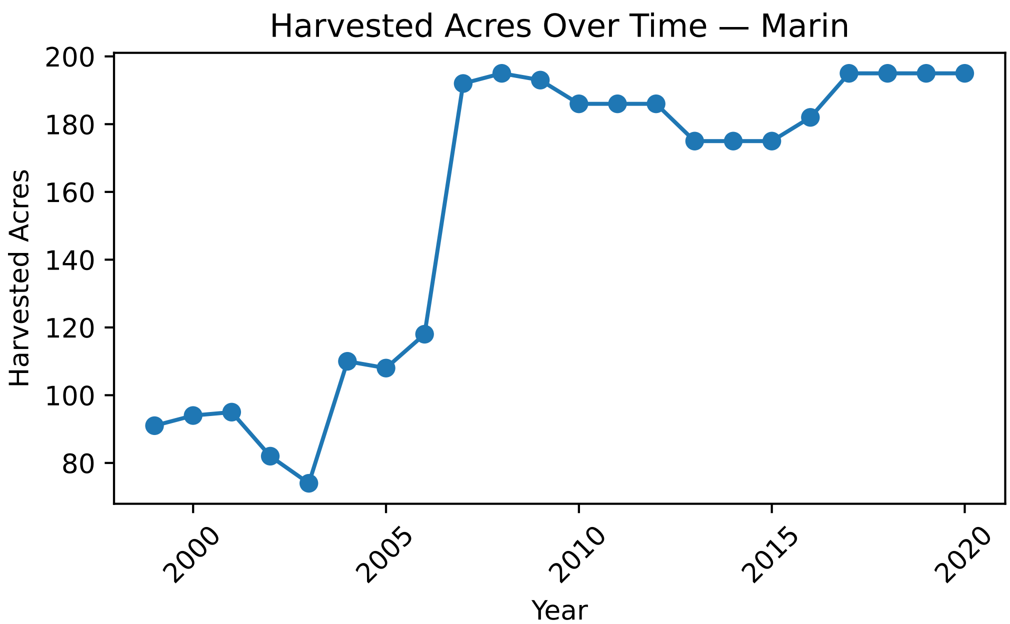

Harvested Acres Over Time — Fresno
Harvested Acres
120000
100000
80000
60000
40000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

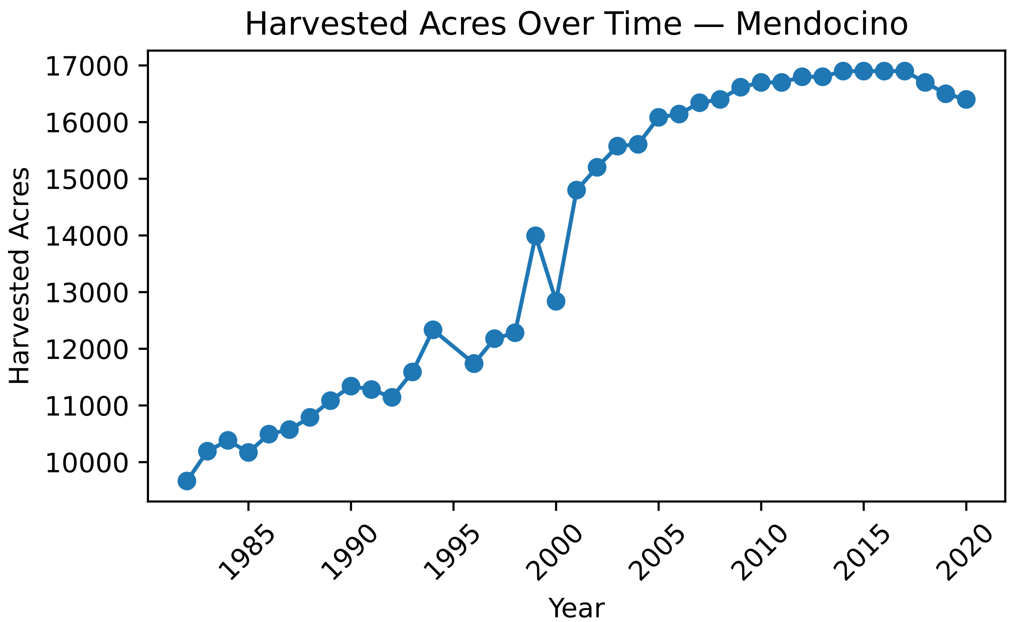

Harvested Acres Over Time — Kern
Harvested Acres
50000
45000
40000
35000
30000
25000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year


Harvested Acres Over Time — Kings
Harvested Acres
Year
5000
4000
3000
2000
1000
1980
1985
1990
1995
2000
2005
2010
2015
2020

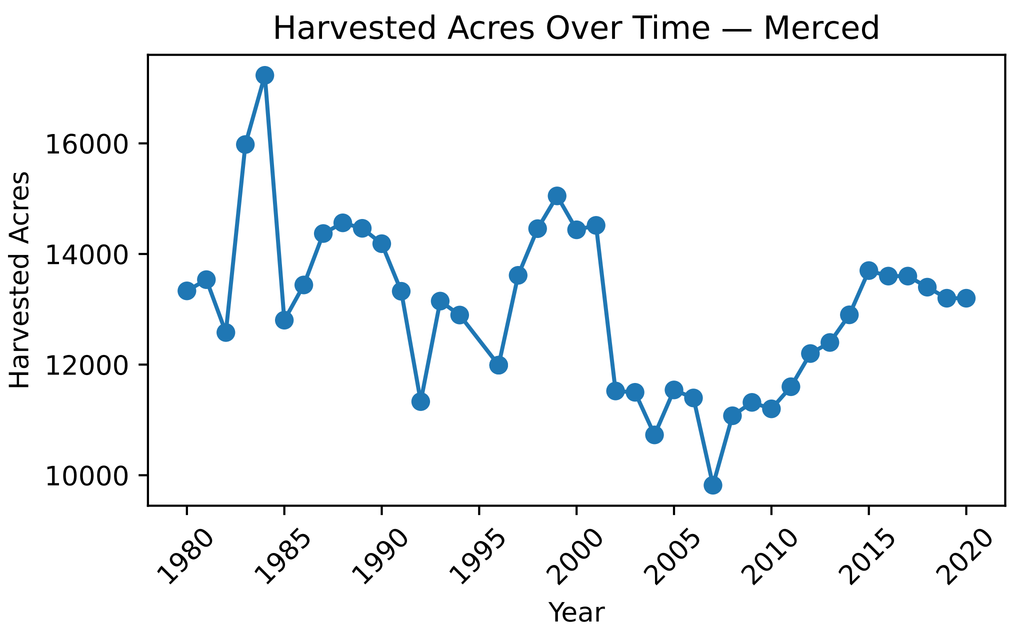

Harvested Acres Over Time — Lake
Harvested Acres
Year
9000
8000
7000
6000
5000
4000
3000
1980
1985
1990
1995
2000
2005
2010
2015
2020


Harvested Acres Over Time — Madera
Harvested Acres
60000
50000
40000
30000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

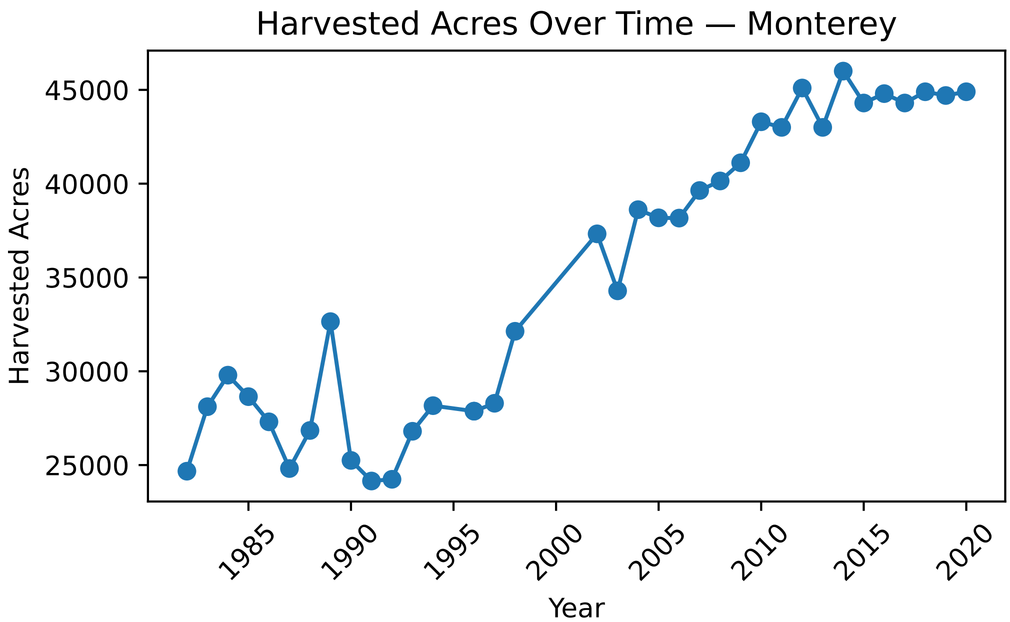

Harvested Acres Over Time — Marin
Harvested Acres
200
180
160
140
120
100
80
2000
2005
2010
2015
2020
Year

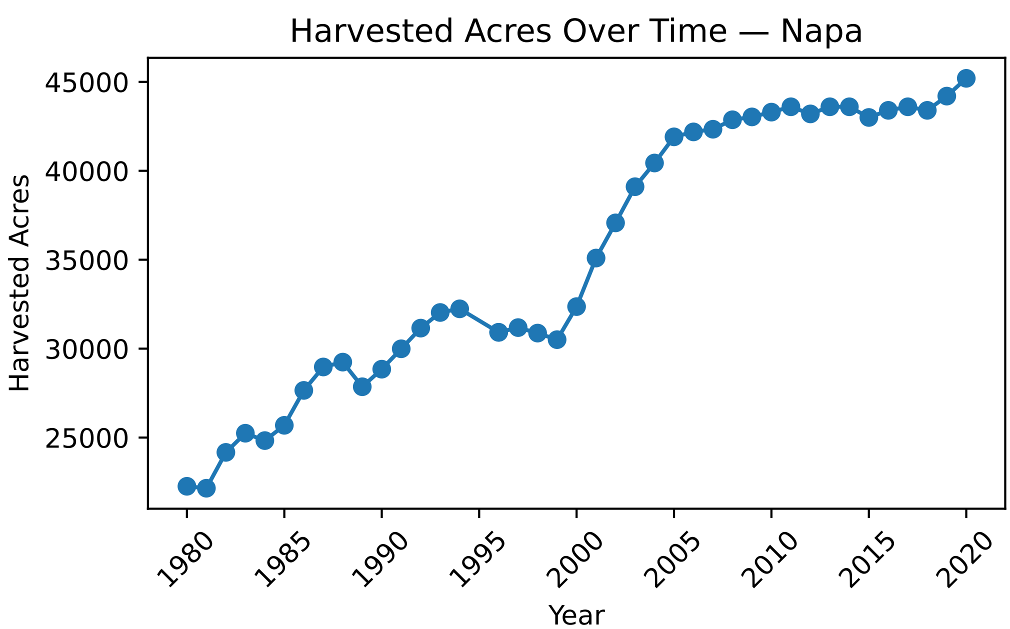


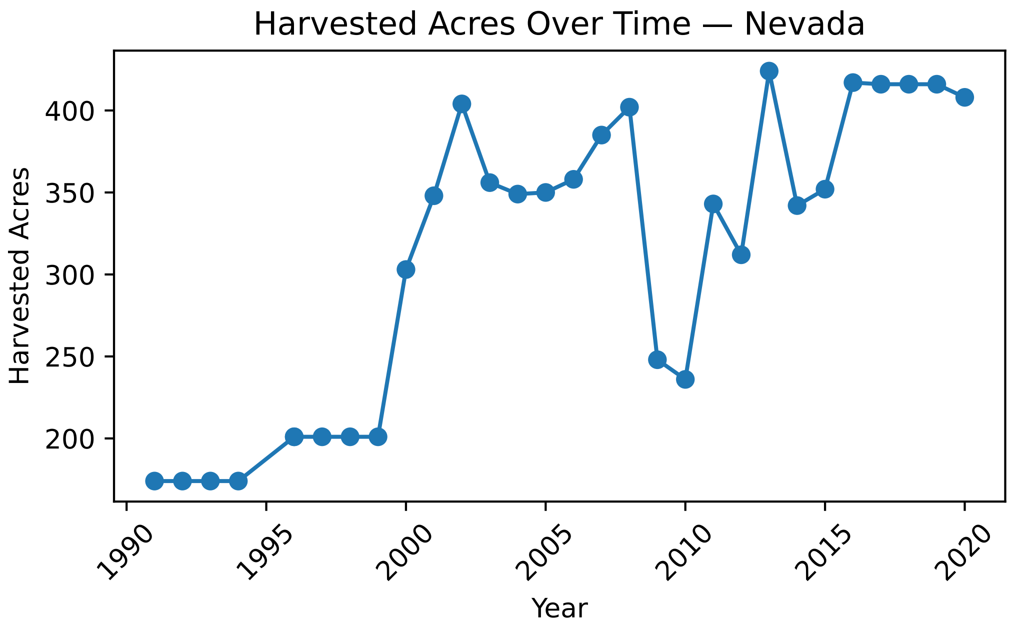
Harvested Acres Over Time — Mendocino
Harvested Acres
Year
17000
16000
15000
14000
13000
12000
11000
10000
1985
1990
1995
2000
2005
2010
2015
2020

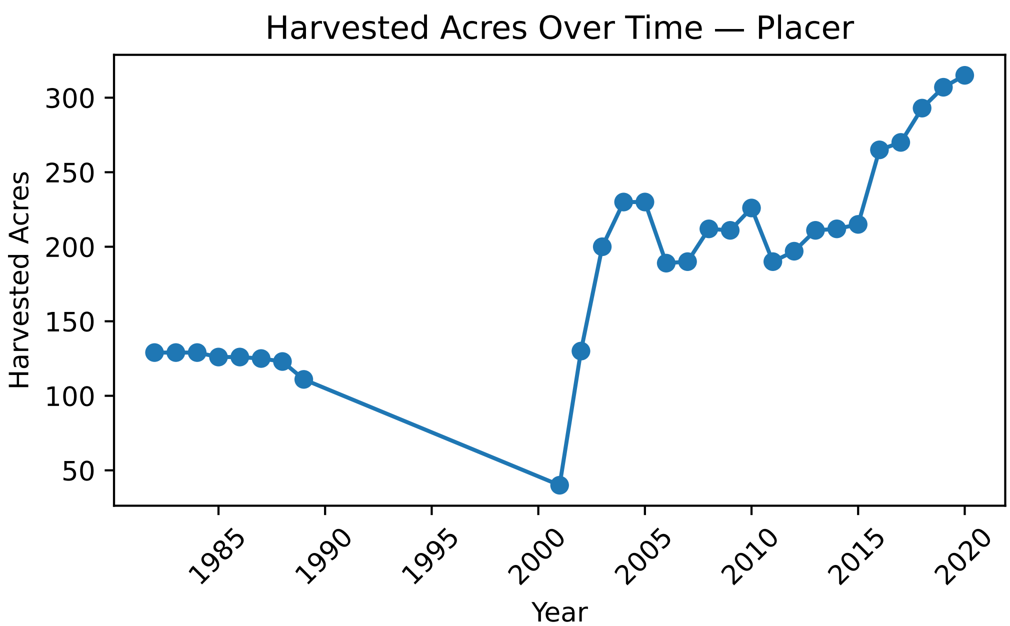


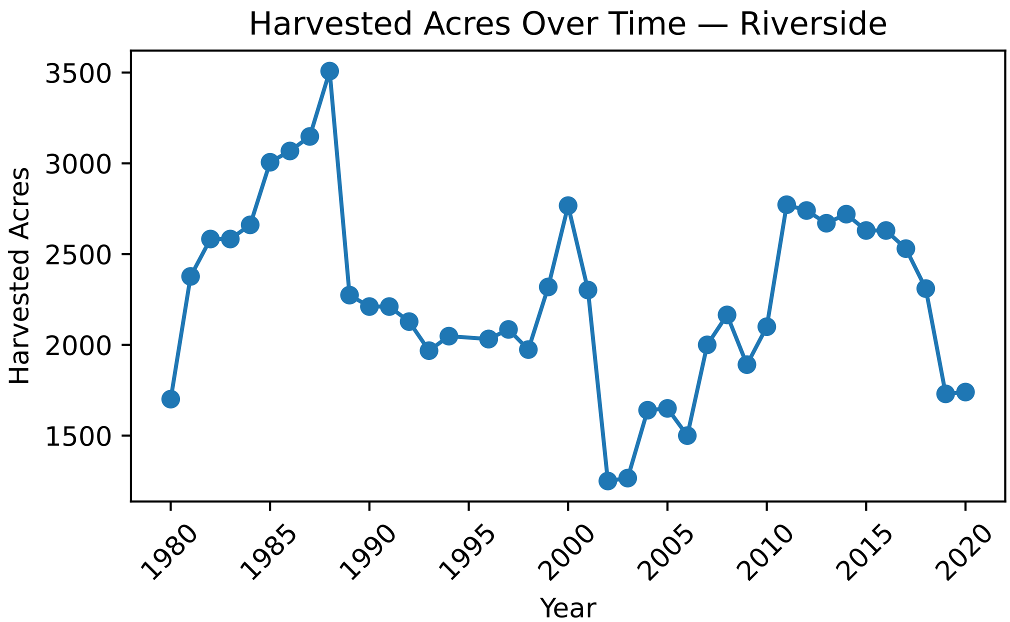
Harvested Acres Over Time — Merced
Harvested Acres
Year
16000
14000
12000
10000
1980
1985
1990
1995
2000
2005
2010
2015
2020

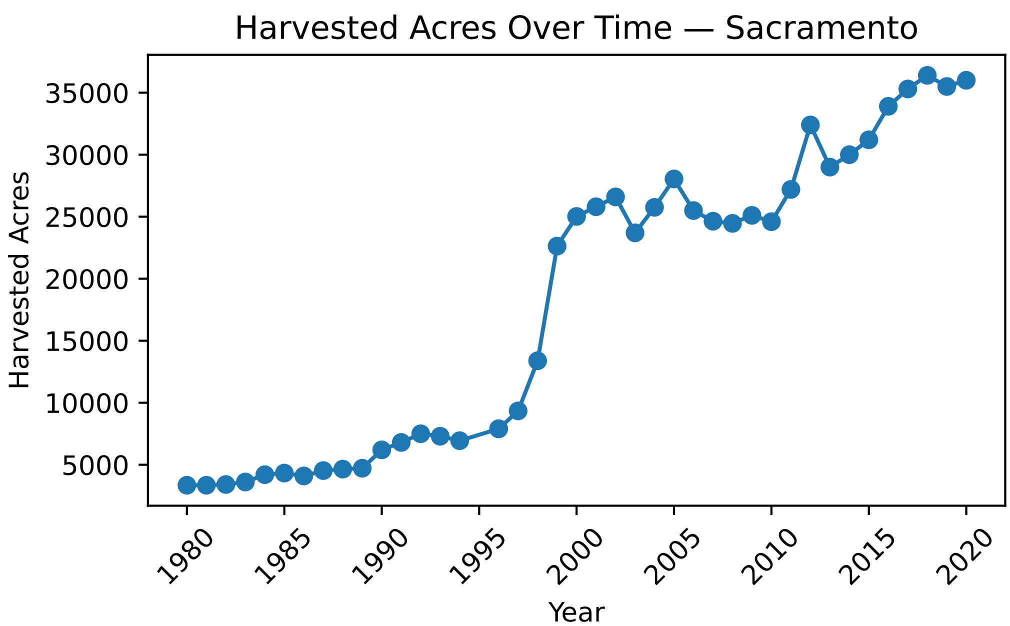

Harvested Acres Over Time — Monterey
Harvested Acres
45000
40000
35000
30000
25000
1985
1990
1995
2000
2005
2010
2015
2020
Year

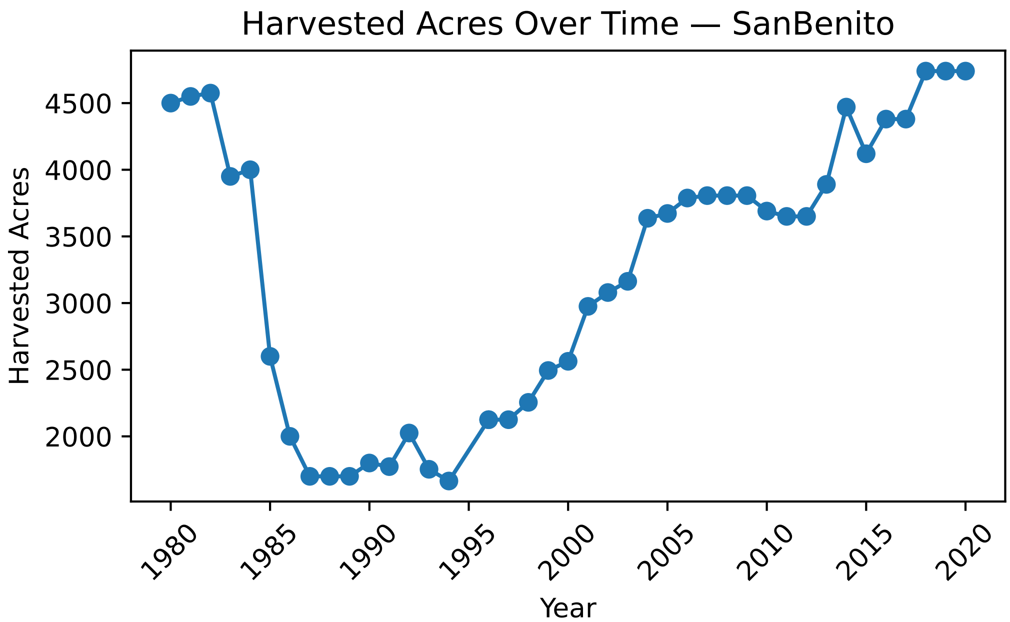

Harvested Acres Over Time — Napa
Harvested Acres
45000
40000
35000
30000
25000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

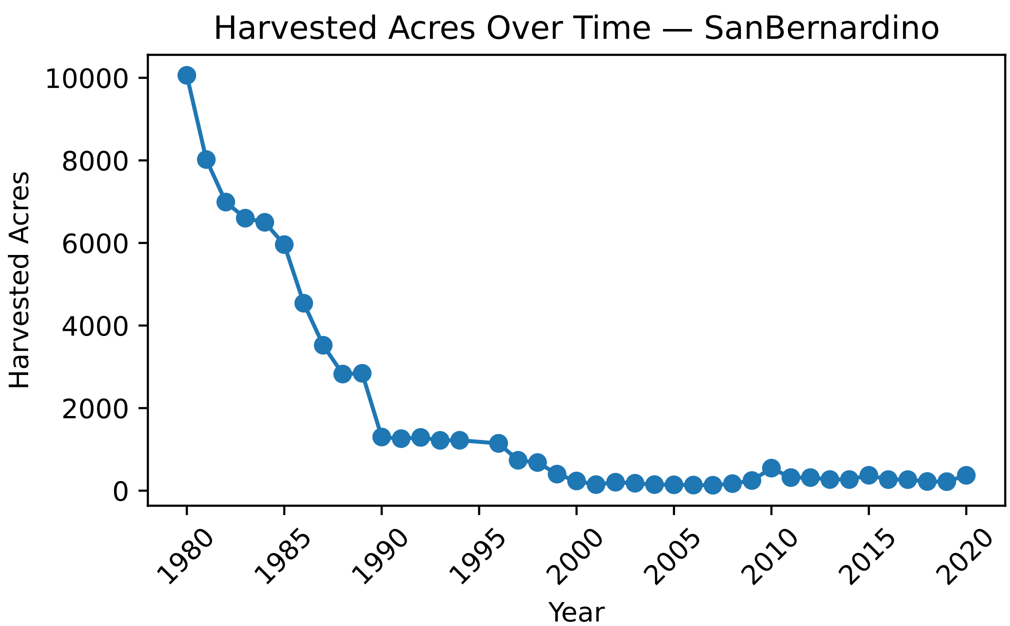

Harvested Acres Over Time — Nevada
Harvested Acres
400
350
300
250
200
1990
1995
2000
2005
2010
2015
2020
Year

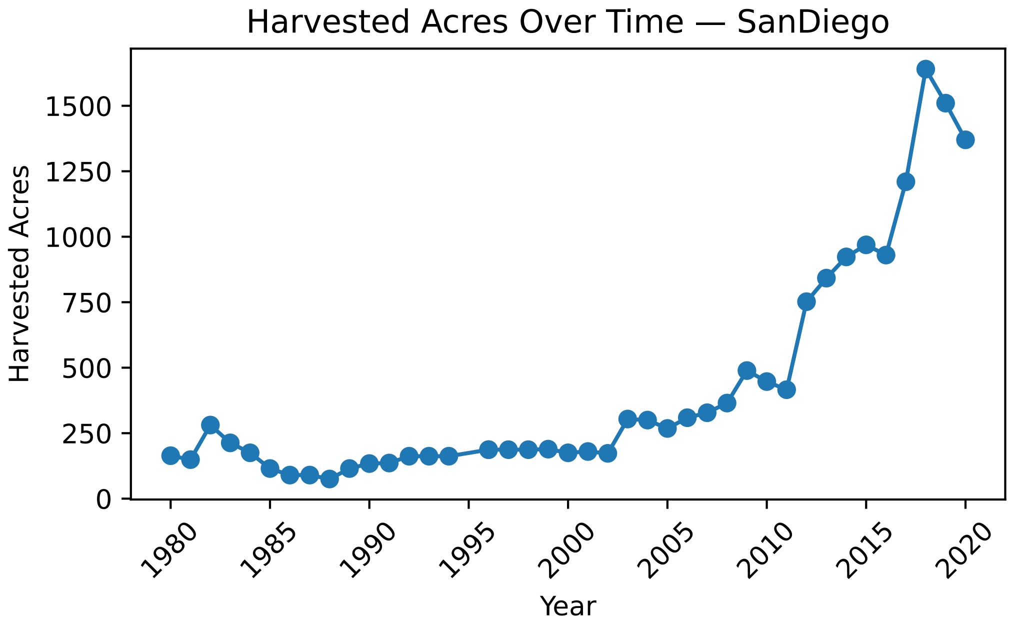

Harvested Acres Over Time — Placer
Harvested Acres
300
250
200
150
100
50
1985
1990
1995
2000
2005
2010
2015
2020
Year

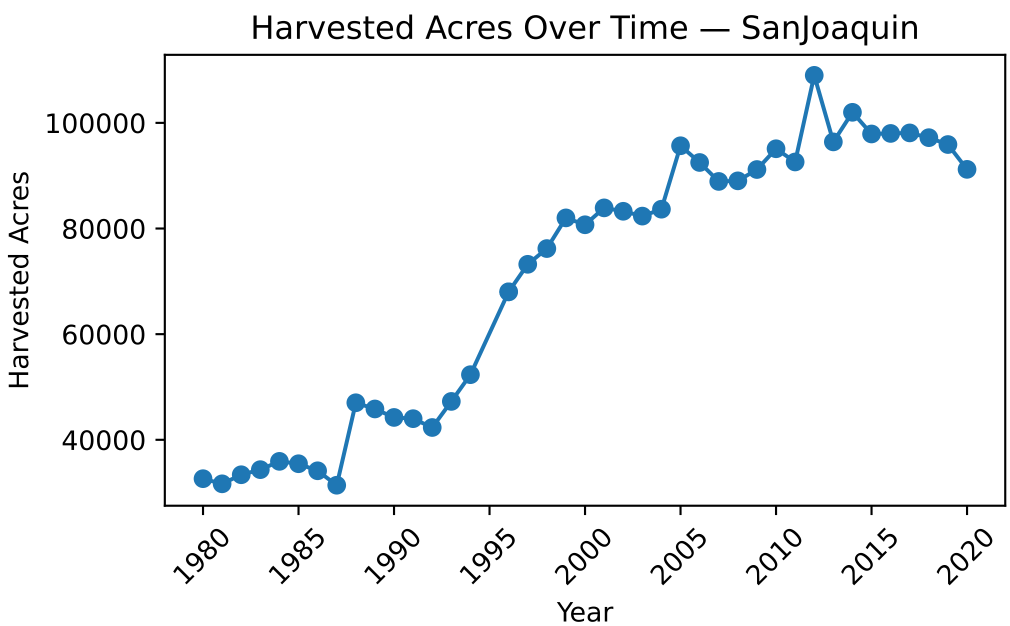

Harvested Acres Over Time — Riverside
Harvested Acres
3500
3000
2500
2000
1500
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

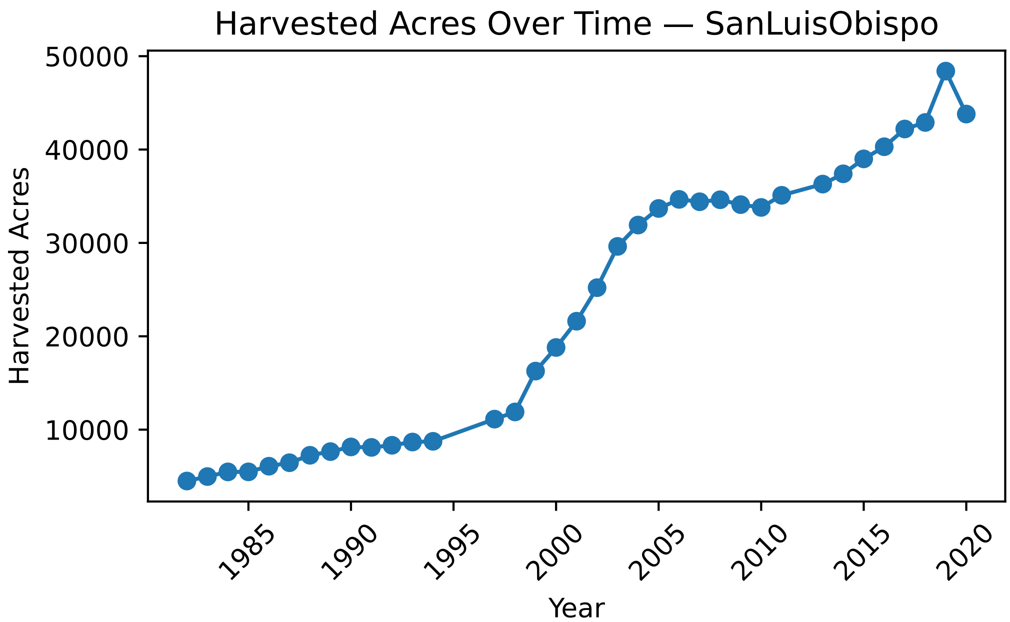

Harvested Acres Over Time — Sacramento
Harvested Acres
35000
30000
25000
20000
15000
10000
5000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year


Harvested Acres Over Time — SanBenito
Harvested Acres
4500
4000
3500
3000
2500
2000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

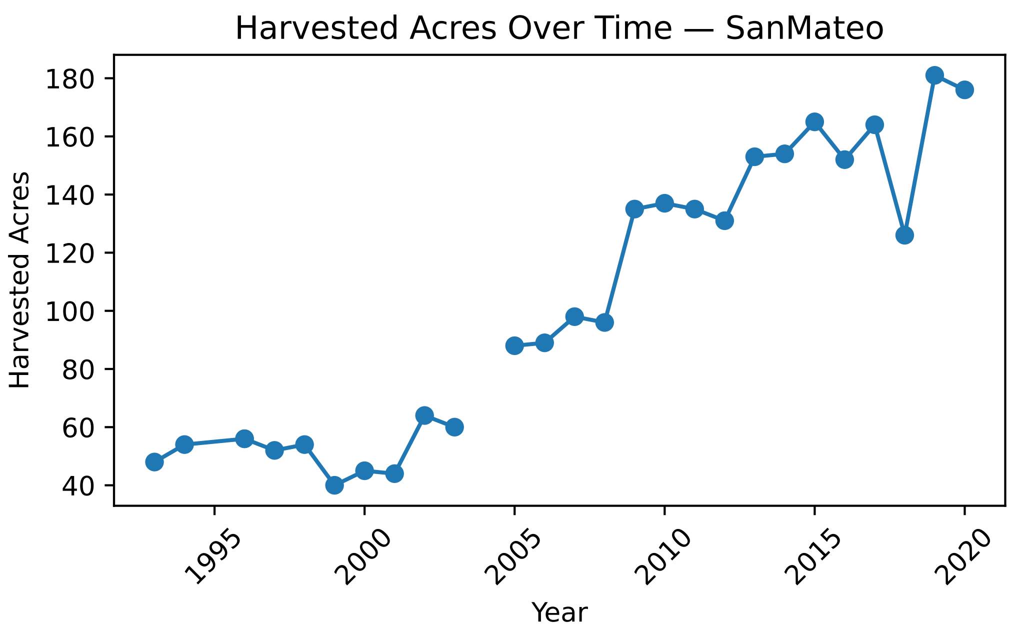

Harvested Acres Over Time — SanBernardino
Harvested Acres
10000
8000
6000
4000
2000
0
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year


Harvested Acres Over Time — SanDiego
Harvested Acres
0
250
500
750
1000
1250
1500
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

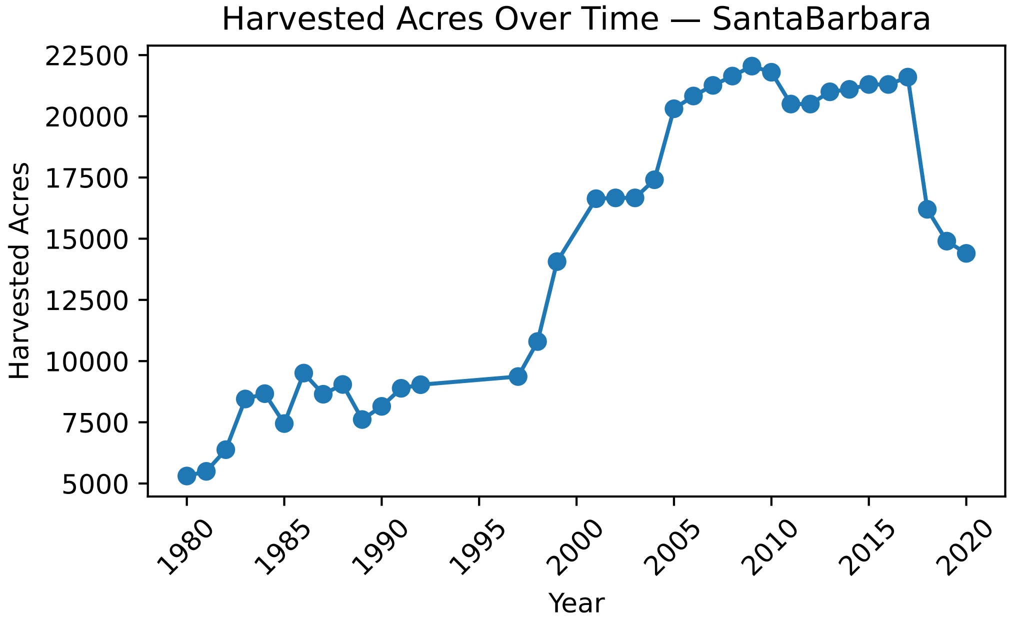

Harvested Acres Over Time — SanJoaquin
Harvested Acres
40000
60000
80000
100000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

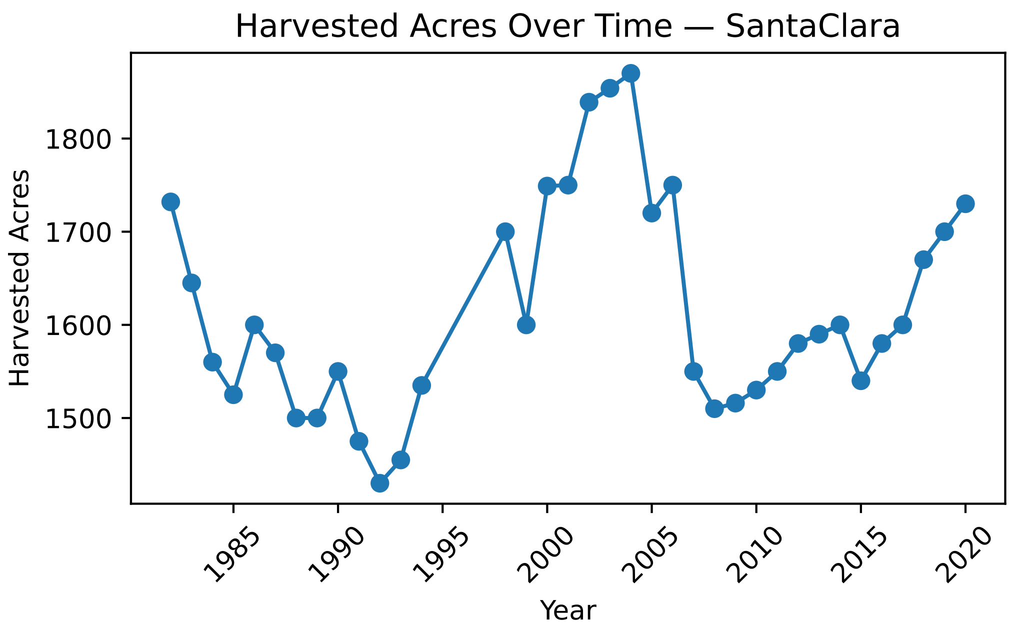


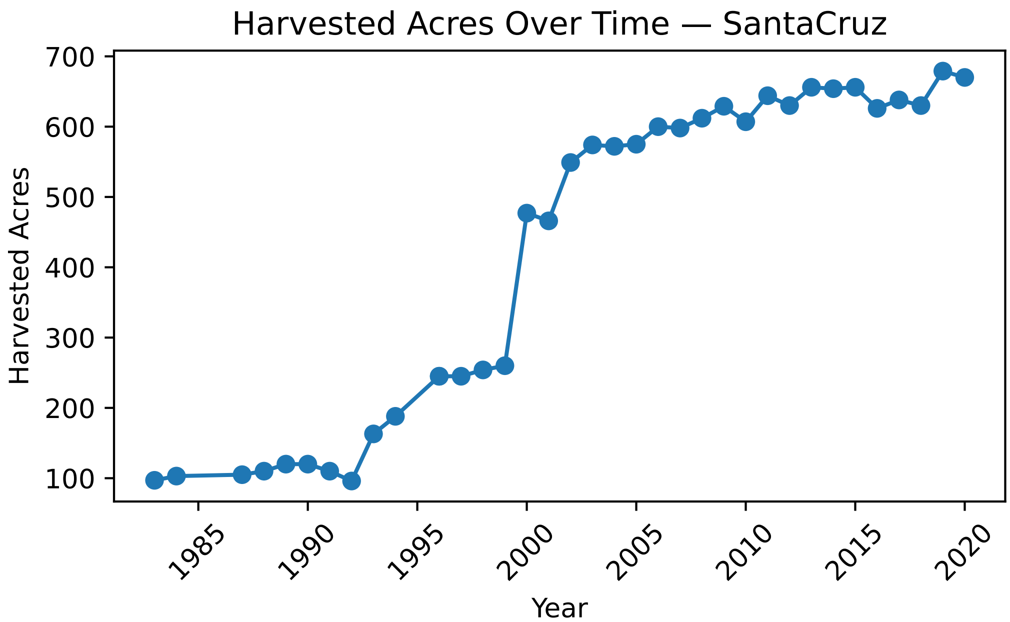
Harvested Acres Over Time — SanLuisObispo
Harvested Acres
50000
40000
30000
20000
10000
1985
1990
1995
2000
2005
2010
2015
2020
Year

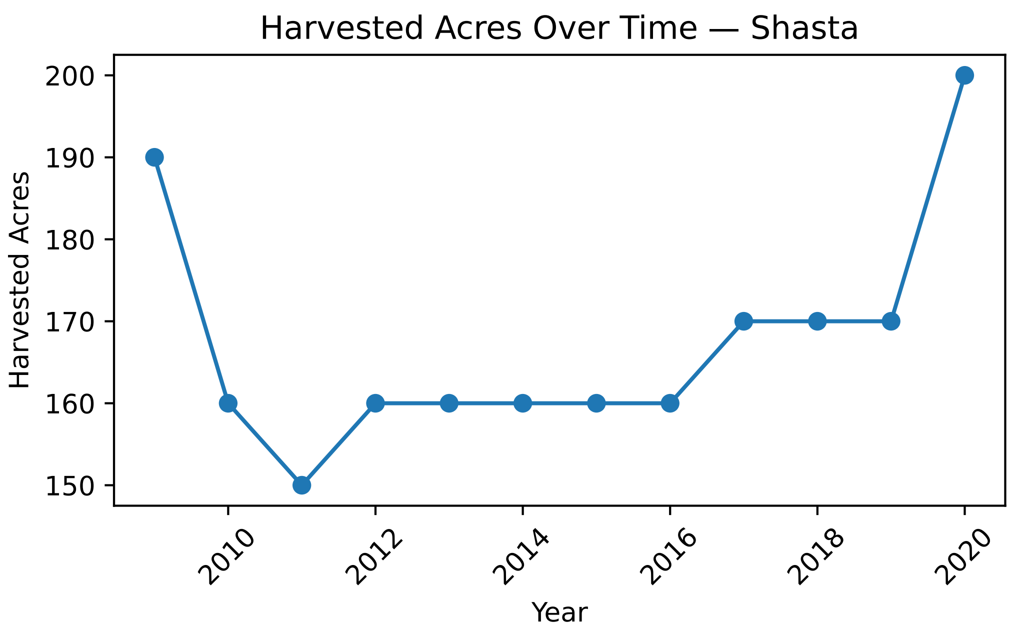


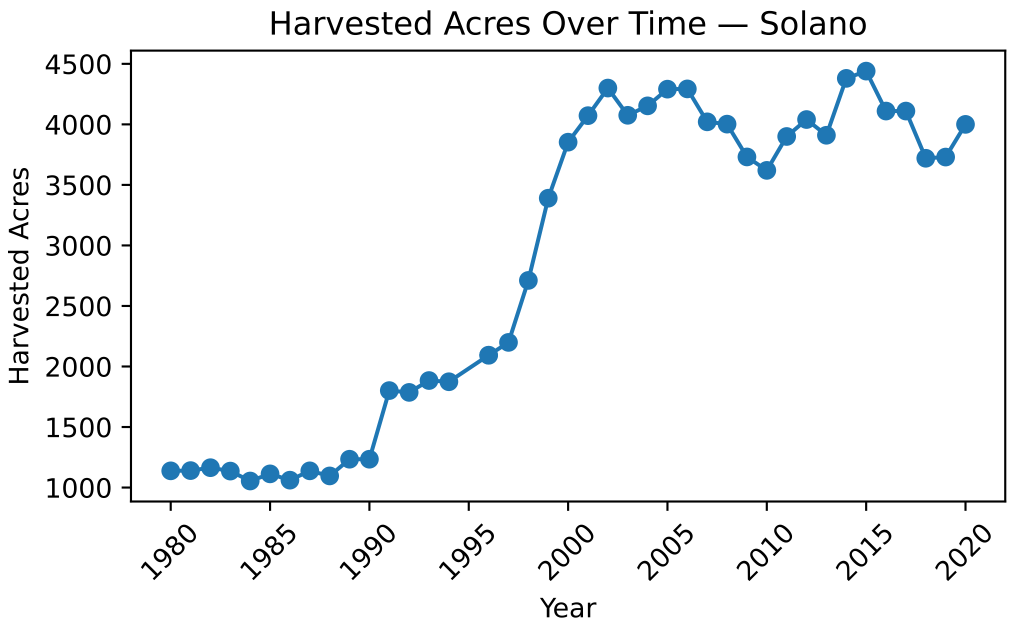
Harvested Acres Over Time — SanMateo
Harvested Acres
180
160
140
120
100
80
60
40
1995
2000
2005
2010
2015
2020
Year


Harvested Acres Over Time — SantaBarbara
Harvested Acres
Year
22500
20000
17500
15000
12500
10000
7500
5000
1980
1985
1990
1995
2000
2005
2010
2015
2020

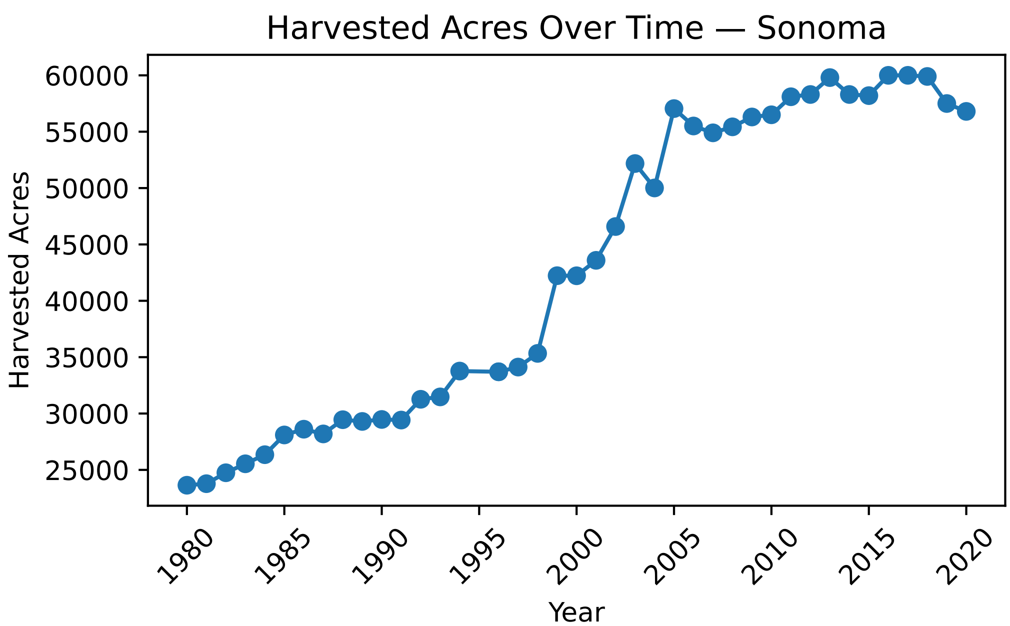

Harvested Acres Over Time — SantaClara
Harvested Acres
Year
1800
1700
1600
1500
1985
1990
1995
2000
2005
2010
2015
2020


Harvested Acres Over Time — SantaCruz
Harvested Acres
Year
100
200
300
400
500
600
700
1985
1990
1995
2000
2005
2010
2015
2020

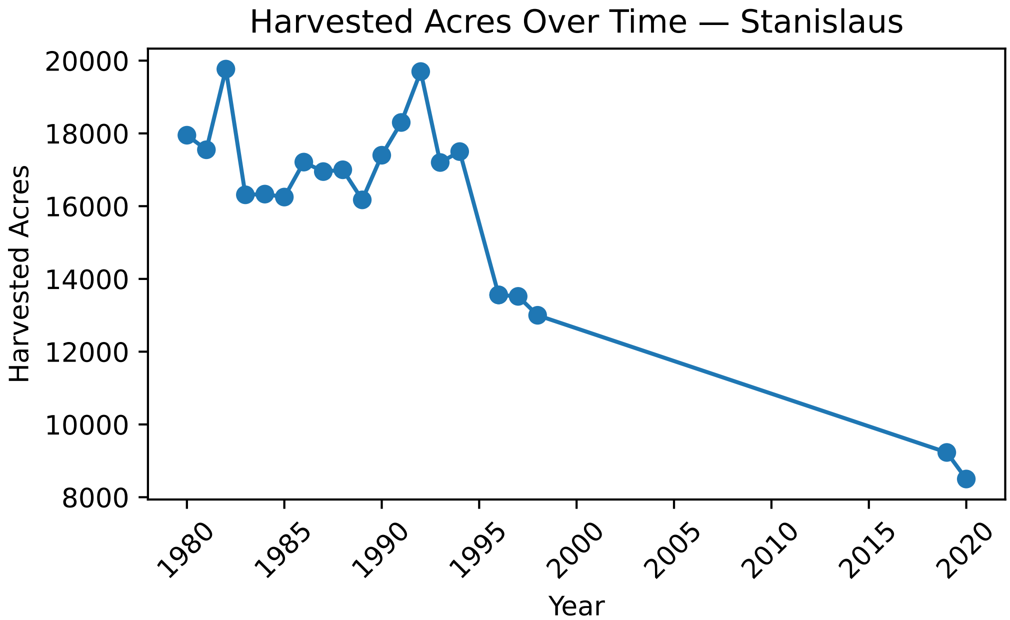

Harvested Acres Over Time — Shasta
Harvested Acres
Year
150
160
170
180
190
200
2010
2012
2014
2016
2018
2020

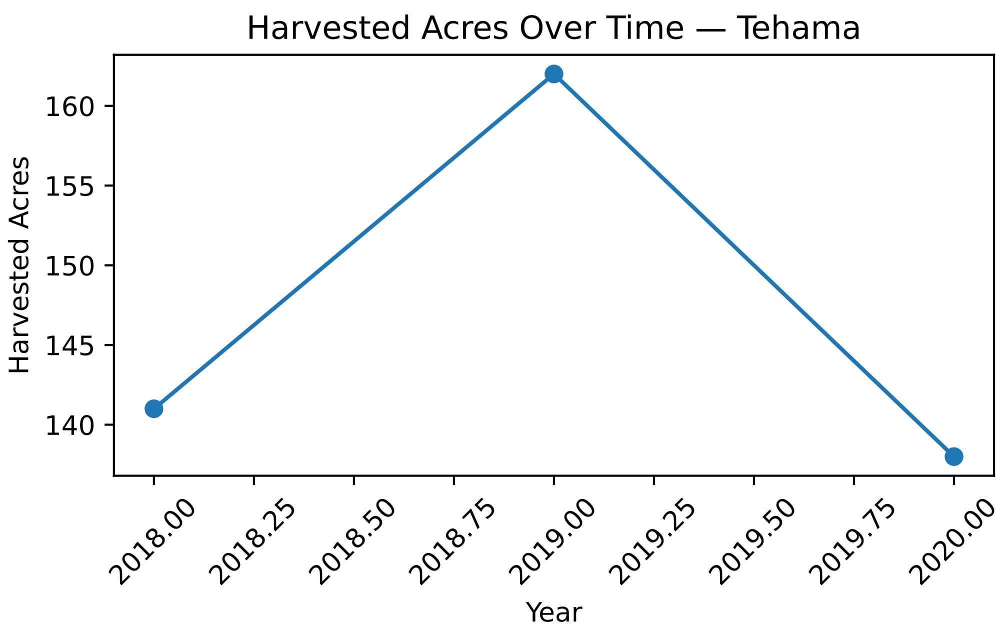


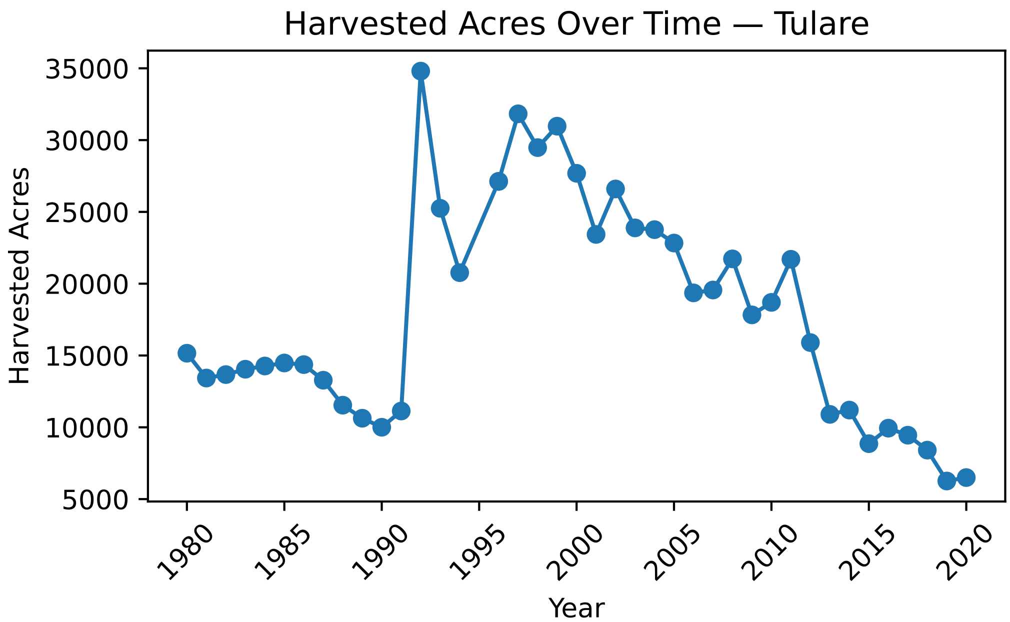
Harvested Acres Over Time — Solano
Harvested Acres
Year



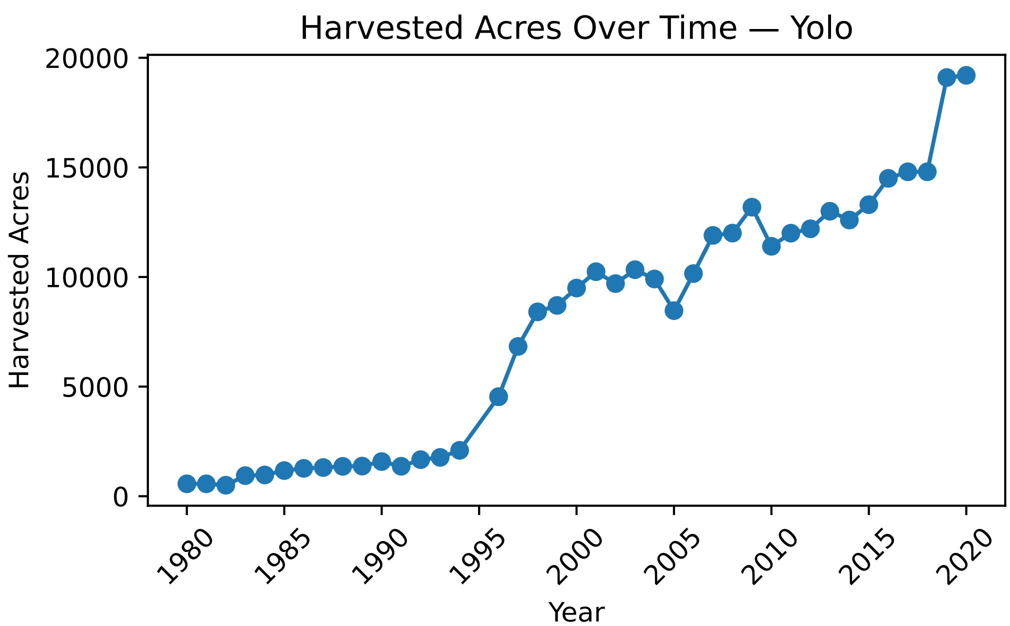
Harvested Acres Over Time — Sonoma
Harvested Acres
Year


Harvested Acres Over Time — Stanislaus
Harvested Acres
20000
18000
16000
14000
12000
10000
8000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

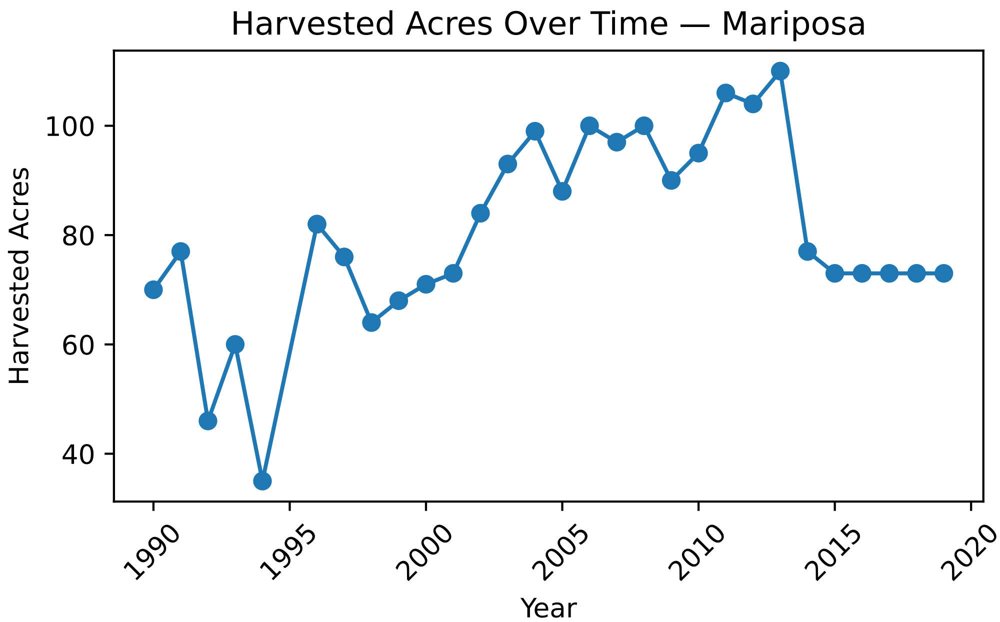

Harvested Acres Over Time — Tehama
Harvested Acres
160
155
150
145
140
2018.00
2018.25
2018.50
2018.75
2019.00
2019.25
2019.50
2019.75
2020.00
Year

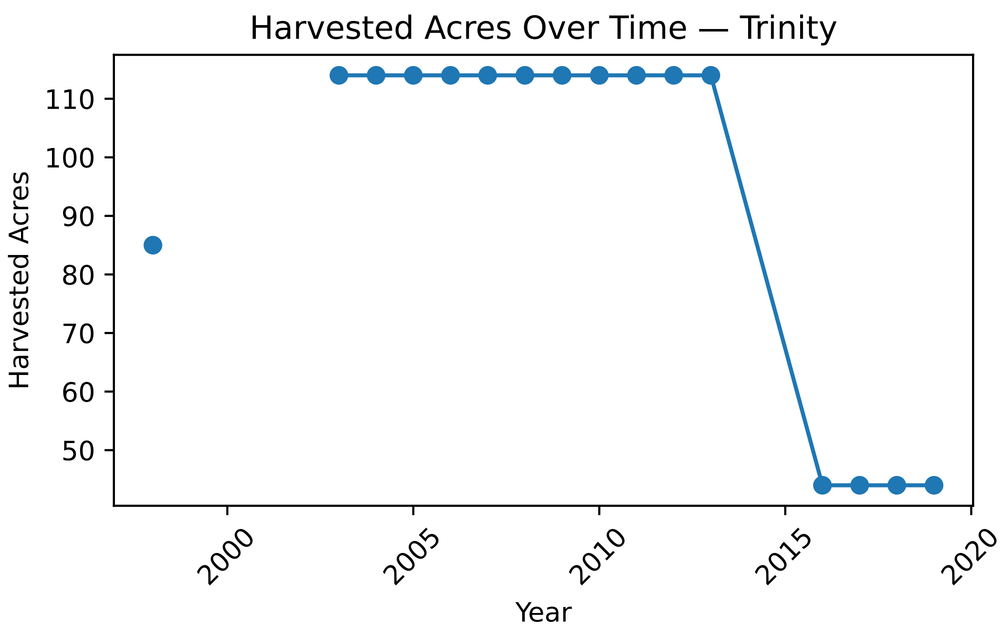


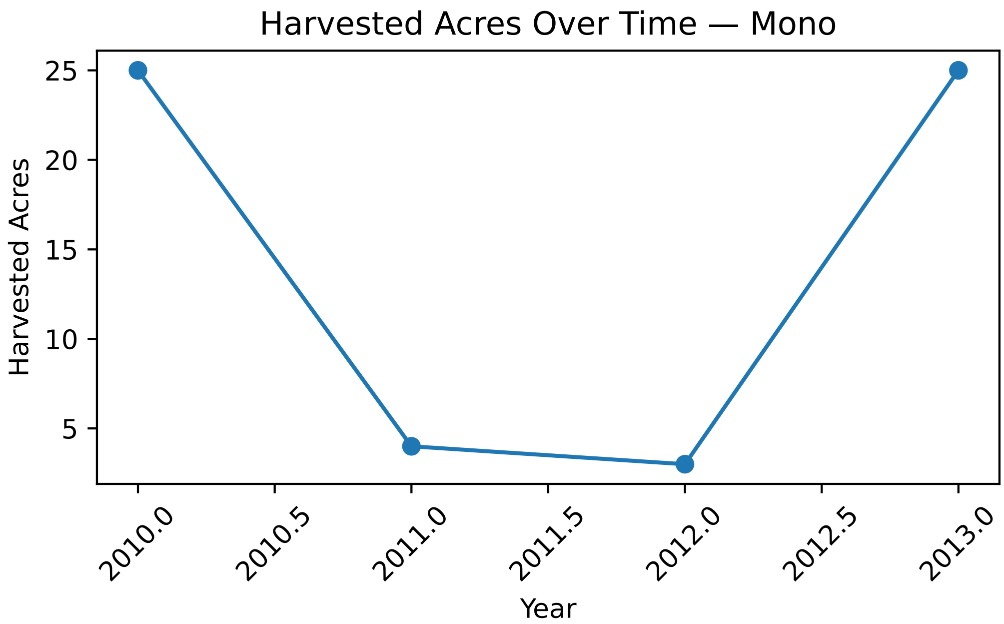
Harvested Acres Over Time — Tulare
Harvested Acres
35000
30000
25000
20000
15000
10000
5000
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

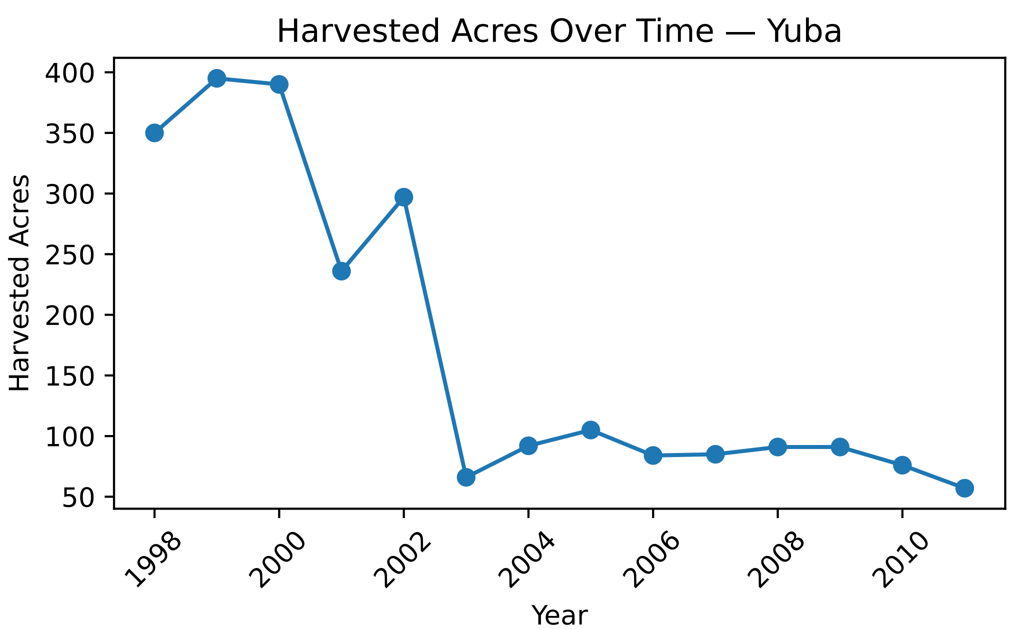


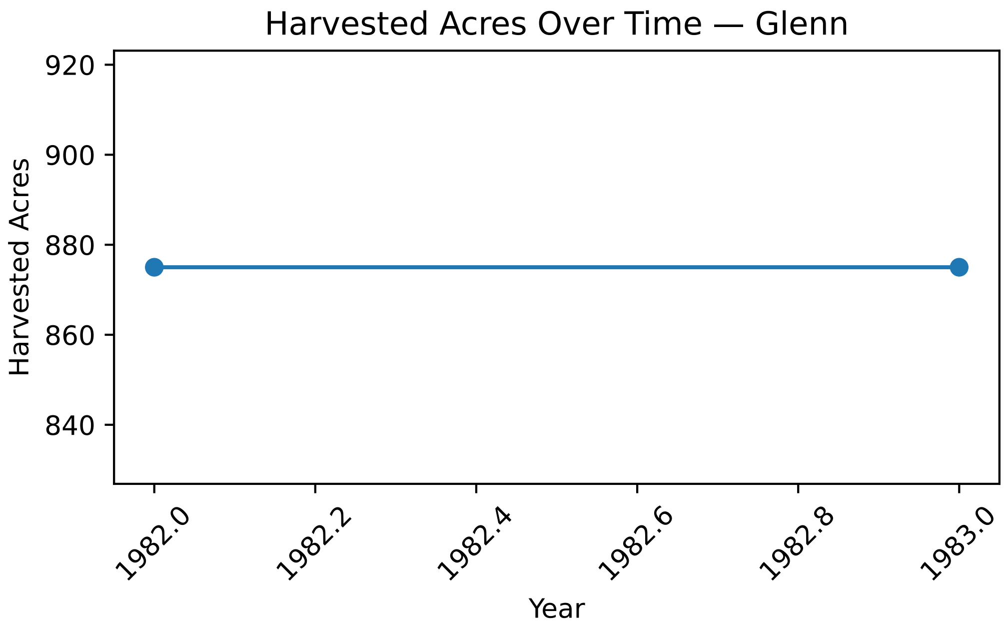
Harvested Acres Over Time — Yolo
Harvested Acres
20000
15000
10000
5000
0
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year


Harvested Acres Over Time — Mariposa
Harvested Acres
Year
40
60
80
100
1990
1995
2000
2005
2010
2015
2020


Harvested Acres Over Time — Trinity
Harvested Acres
Year
50
60
70
80
90
100
110
2000
2005
2010
2015
2020


Harvested Acres Over Time — Mono
Harvested Acres
Year
25
20
15
10
5
2010.0
2010.5
2011.0
2011.5
2012.0
2012.5
2013.0


Harvested Acres Over Time — Yuba
Harvested Acres
Year
400
350
300
250
200
150
100
50
1998
2000
2002
2004
2006
2008
2010



My theory that maybe somehow droughts and heatwaves are effecting yield is based on nothing. There is no patter or corelaion that I can see between the graphs and the dates of major droughts and heatwaves that I can see. At this point i have to a question. Is everything we've heard anecdotal? I mean to say that grape farmers are saying that it's getting harder to keep farming grapes, but so long as they keep watering their grapes, does the heat really impact them at all? I'm gonna ask my farmer friend and see what he tells me

```
temps = pd.read_csv("yearly_temps.csv")

data = grapes[['Year', 'Yield(Unit/Acre)']]

df = pd.merge(grapes[['Year', 'Yield(Unit/Acre)']], temps, on='Year', how='inner')

df = df.dropna(subset=['Yield(Unit/Acre)', 'MaxTemp'])

X = df[['MaxTemp']]
Y = df['Yield(Unit/Acre)']

model = LinearRegression()
model.fit(X,Y)

y_pred = model.predict(X)
```

```
print("Slope:", model.coef_[0])
print("Intercept:", model.intercept_)
print("R^2:", r2_score(Y, y_pred))
print("RMSE:", np.sqrt(mean_squared_error(Y, y_pred)))
print("MAE:", mean_absolute_error(Y, y_pred))
```

```
Slope: 0.024954742311580007
Intercept: 3.6478943125562777
R^2: 8.276083750624608e-05
RMSE: 3.26588280009236
MAE: 2.6054558101420753
```

```
import statsmodels.api as sm
from scipy.stats import pearsonr

r, p_value_corr = pearsonr(df['MaxTemp'], df['Yield(Unit/Acre)'])

print("Correlation r:", r)
print("p-value:", p_value_corr)

X = sm.add_constant(df['MaxTemp'])
Y = df['Yield(Unit/Acre)']

model = sm.OLS(Y, X).fit()

print(model.summary())
```

```
Correlation r: 0.009097298363053438
p-value: 0.746409781244134
                            OLS Regression Results
==============================================================================
Dep. Variable:       Yield(Unit/Acre)   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                     -
0.001
Method:                 Least Squares   F-statistic:                    0.1046
Date:                Mon, 01 Dec 2025   Prob (F-statistic):              0.746
Time:                        13:02:43   Log-Likelihood:                    -
3294.7
No. Observations:                1266   AIC:                             6593.
Df Residuals:                    1264   BIC:                             6604.
Df Model:                           1
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.6479      4.726      0.772      0.440      -5.624      12.920
MaxTemp        0.0250      0.077      0.323      0.746      -0.126       0.176
==============================================================================
Omnibus:                      294.907   Durbin-Watson:                   1.676
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              640.243
Skew:                           1.302   Prob(JB):                     9.40e-
140
Kurtosis:                       5.314   Cond. No.                     3.15e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.15e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```