

MA615-AssignmentF-Strawberry

Group 6 Zhuolin Liu Xiaoyang Hu Sile Wang

1. Background

Strawberries (scientific name: *Fragaria × ananassa*) are an important economic crop in the United States. They are widely consumed in fresh form, as frozen products, or processed into various products. Understanding strawberry planting areas, production volumes, price trends, and profit margins is crucial for agricultural economic research, market analysis, and farm-level decision-making. This report uses historical data from the United States Department of Agriculture (USDA)/National Agricultural Statistics Service (NASS) to analyze the strawberry industry in the United States, examining production trends, price fluctuations, and preliminary profit forecasts, and providing basic future profit forecasts.

2. Data Source and Methods

2.1 Data source

The main data sources for this study include strawberry survey data from USDA/NASS and census data. These sources provide four core indicators:

1. **Strawberries - Acres Bearing**, which is the area of strawberries actually produced and available for market sales in a given year.
2. **Strawberries - Price Received**, measured in \$ / CWT, which is the farm price received by growers for every 100 hundredweight of agricultural products.
3. **Strawberries - Production**, measured in \$, representing the total crop value in US dollars, roughly equivalent to total sales revenue.
4. **Strawberries, Utilized - Production**, measured in CWT, which is the sold physical volume in hundredweight (1 CWT = 100 lb).

These variables are economically meaningful. Acres Bearing is used as the denominator to convert “total scale” into “per-acre intensity” measures such as yield and revenue per acre. Price Received is the price driver of income. Production measured in dollars serves as a proxy for total industry revenue. Utilized Production (in CWT) reflects the amount of fruit that actually entered the market and therefore drives realized sales. They all can collectively help us calculate yield, income per acre, cost risks, and profit per acre.

2.2 Cleaning

	year	data_item	value
1	2022	STRAWBERRIES - ACRES BEARING	70709
2	2017	STRAWBERRIES - ACRES BEARING	58117
3	2024	STRAWBERRIES - PRICE RECEIVED, MEASURED IN \$ / CWT	124
4	2024	STRAWBERRIES - PRODUCTION, MEASURED IN \$	3996863000
5	2024	STRAWBERRIES, UTILIZED - PRODUCTION, MEASURED IN CWT	32225500
6	2023	STRAWBERRIES - PRICE RECEIVED, MEASURED IN \$ / CWT	123

In the original dataset, the annual data records have multiple rows, sometimes including different survey types, and sometimes there are distinctions between “marketing year” and “annual” records. Therefore, in the first step of the cleaning process, we filtered the data set, retaining only the national observation results, that is, (Geo.Level == “NATIONAL” and State == “US TOTAL”), so as to avoid duplicate counting between states and ensure that the data for each year can be presented at the national level in the United States.

Pivot wider

A tibble: 6 x 5

	Year	Acres_Bearing	Price_per_CWT	Production_USD	Utilized_Production_CWT
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2022	70709	114	3259100000	28494300
2	2017	58117	107	2459234000	28850850
3	2024	NA	124	3996863000	32225500
4	2023	NA	123	3543596000	28725400
5	2021	NA	129	3583960000	27854700
6	2020	NA	97.1	2591759000	26684200

And because the USDA data was initially recorded in character string form, these values were processed and converted to numerical form, that is, removing commas and forcing them to be of numeric type. At this stage, all remaining non-numeric or missing values were deleted. Then, we used the `pivot_wider()` function to convert the data from “long” format to “wide” format, making each economic/production concept a separate column. In the generated data frame,

for each year, there are columns such as Acres_Bearing, Price_per_CWT, Production_USD, and Utilized_Production_CWT.

Fill missing values

```
# A tibble: 6 x 5
  Year Acres_Bearing Price_per_CWT Production_USD Utilized_Production_CWT
  <dbl>         <dbl>         <dbl>         <dbl>         <dbl>
1  1997             NA           55.5       903350000             NA
2  1998             NA           61.1      1000254000             NA
3  1999             NA           62.5      1144876000             NA
4  2000             NA           55        1044594000             NA
5  2001             NA           64.7      1068582000             NA
6  2002             NA           61.6      1161630000             NA
```

Finally, for some years, one or more variables were missing. To generate a usable continuous time series, we filled in the missing values through the linear interpolation method (na.approx). This provided smooth estimates for the blank areas and enabled us to calculate derived quantities (such as per-acre yield) and conduct correlation and regression analyses without discarding entire rows of data. After interpolation, the dataset was sorted by year to create a consistent timeline.

Compute key profitability metrics

Year	Acres_Bearing	Price_per_CWT	Production_USD
Min. :1997	Min. :58117	Min. : 55.00	Min. :9.034e+08
1st Qu.:2004	1st Qu.:61265	1st Qu.: 63.02	1st Qu.:1.355e+09
Median :2010	Median :64413	Median : 77.53	Median :2.250e+09
Mean :2010	Mean :64413	Mean : 83.43	Mean :2.173e+09
3rd Qu.:2017	3rd Qu.:67561	3rd Qu.: 99.58	3rd Qu.:2.768e+09
Max. :2024	Max. :70709	Max. :129.00	Max. :3.997e+09
	NA's :22		
Utilized_Production_CWT	Yield_per_Acre	Revenue_Est	
Min. :23958700	Min. :379.4	Min. :2.591e+09	
1st Qu.:27854700	1st Qu.:403.8	1st Qu.:2.755e+09	
Median :28725400	Median :407.4	Median :3.248e+09	
Mean :28385233	Mean :428.9	Mean :3.198e+09	
3rd Qu.:29094850	3rd Qu.:462.0	3rd Qu.:3.533e+09	
Max. :32225500	Max. :496.4	Max. :3.996e+09	
NA's :19	NA's :22	NA's :19	
Cost	Profit	Profit_per_Acre	

Min.	:1.744e+09	Min.	:6.209e+08	Min.	: 9454
1st Qu.	:1.838e+09	1st Qu.	:8.708e+08	1st Qu.	:13937
Median	:1.932e+09	Median	:1.014e+09	Median	:15402
Mean	:1.932e+09	Mean	:1.067e+09	Mean	:16616
3rd Qu.	:2.027e+09	3rd Qu.	:1.289e+09	3rd Qu.	:21006
Max.	:2.121e+09	Max.	:1.548e+09	Max.	:23118
NA's	:22	NA's	:22	NA's	:22

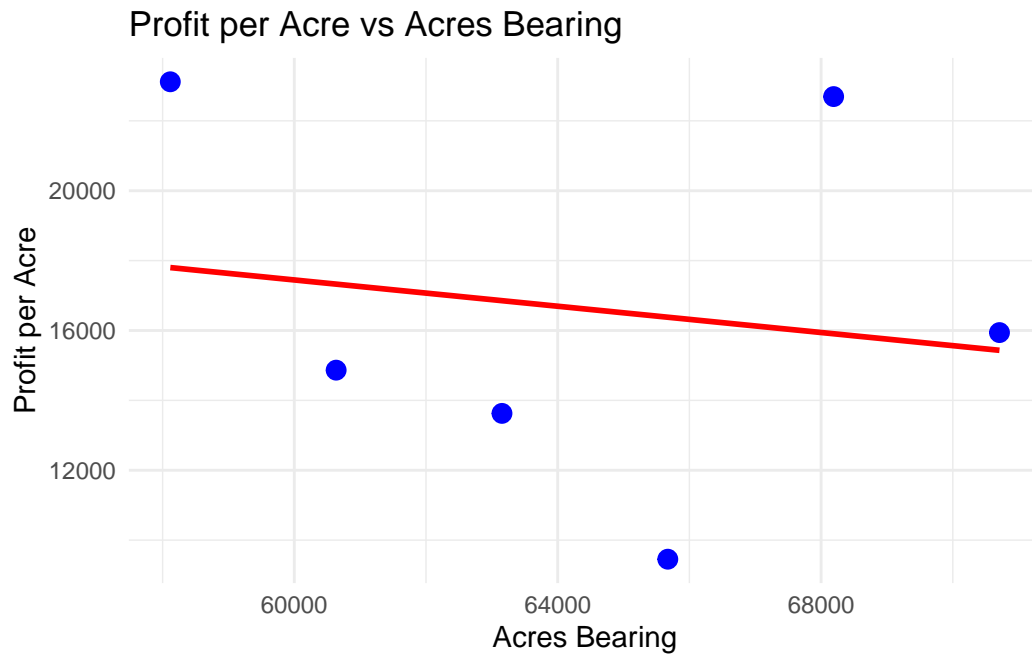
After cleaning, some indicators were calculated:

1. **Yield_per_Acre** = Utilized_Production_CWT / Acres_Bearing.
Shows usable sold production per acre (CWT). This shows how much usable/sold strawberry production we get per acre (in CWT). It measures the land's productivity.
2. **Revenue_Est** = Price_per_CWT × Utilized_Production_CWT. This is an estimate of total strawberry revenue. It uses the selling price and the sold quantity.
3. **Cost** = Acres_Bearing × \\$30,000. We assumed it to be approximately \$30,000 per acre for profit calculations. This is a strong assumption, but it gives us a way to discuss costs.
4. **Profit** = Revenue_Est - Cost. This approximates total gross operating surplus for the industry under the stated cost assumption.
5. **Profit_per_Acre** = Profit / Acres_Bearing. This converts total profit into a per-acre profitability measure, which is easier to compare across years and across different acreage scales. This shows profit per acre. It lets us compare how “good” one year is compared to another, even if total acres change.

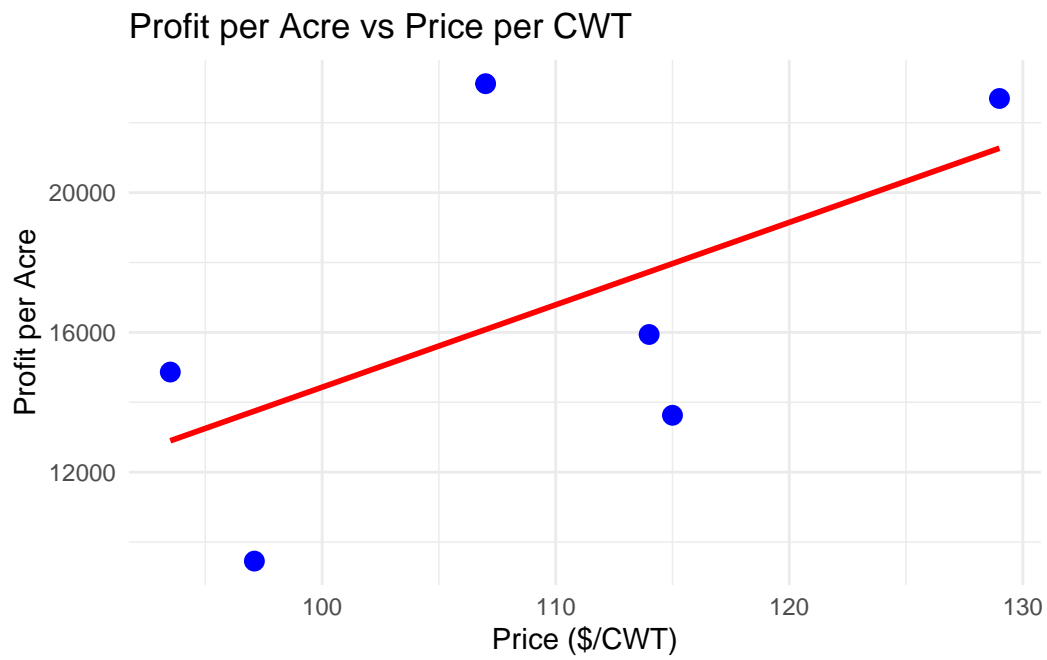
3. Exploratory Data Analysis (EDA) and Profitability Modeling

3.1 Key patterns

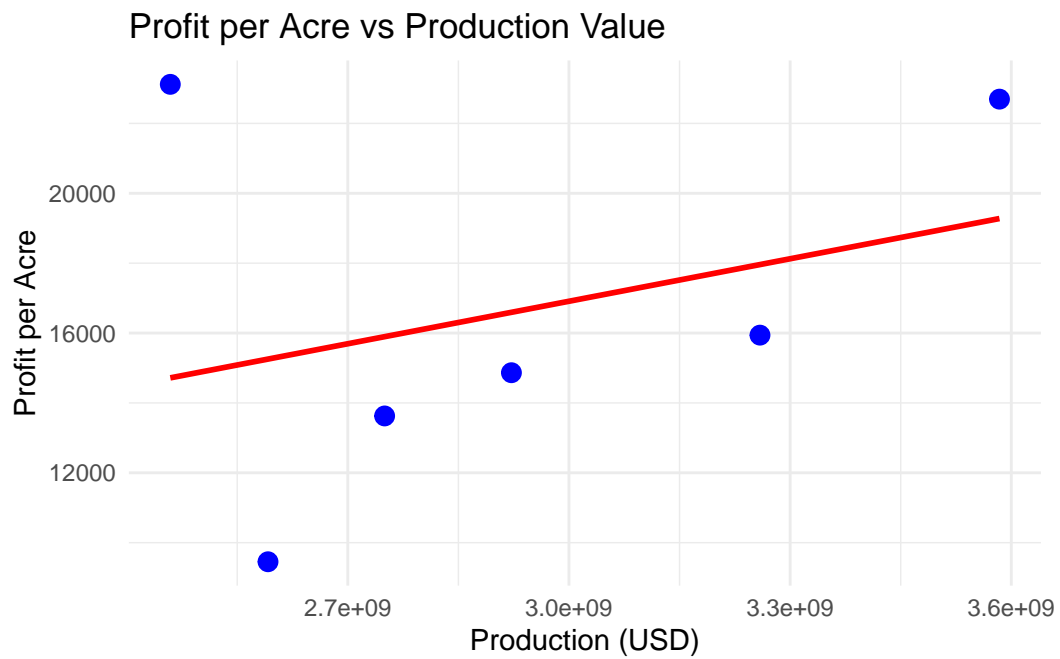
Plot 1



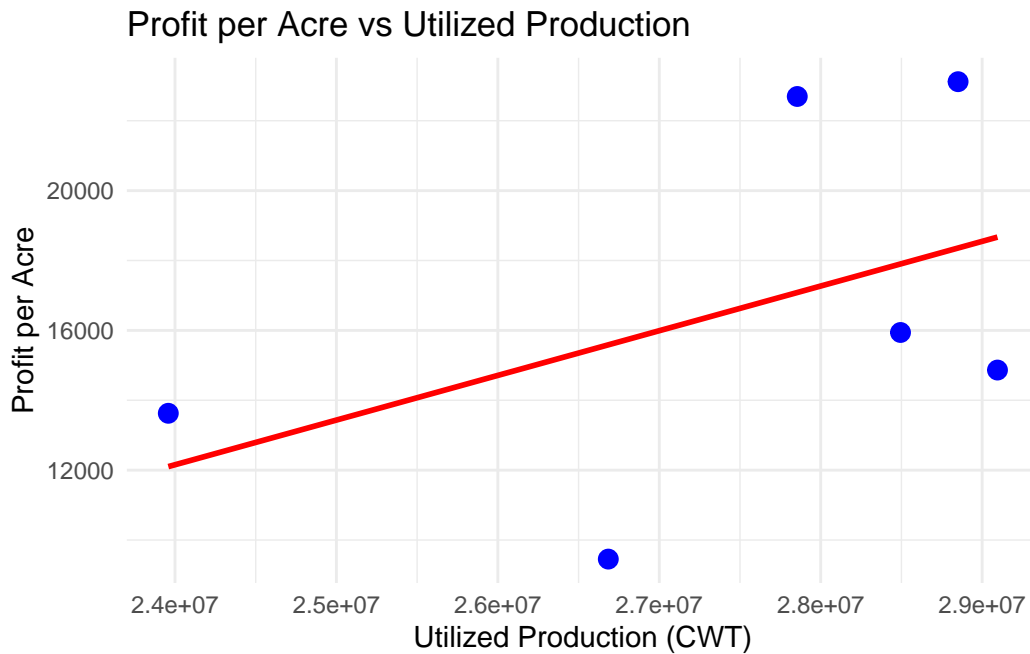
Plot2



Plot3



Plot4



The Exploratory data analysis shows several economic patterns:

1. **Strawberry acreage is significant and supports high output.** Acres_Bearing tells us how many acres are actually producing strawberries. In recent years, the U.S. has had tens of thousands of bearing acres. When we divide Utilized_Production_CWT by Acres_Bearing, we get Yield_per_Acre, which is usually very high. This means each acre of strawberries produces a lot of marketable fruit. Strawberries are very intensive.
2. **Prices are high.** Price_per_CWT is often above \$100 per CWT. This is a high farm-gate price, which shows that strawberries are valuable and not a cheap bulk commodity. Because revenue is basically price x quantity sold, a higher price directly increases total revenue.
3. **Production value is very large.** Production_USD (the value of production in dollars) is in the billions. This represents the industry's economic scale. When the price increases, Production_USD tends to increase as well. So revenue is very sensitive to price, not just to the number of pounds sold.
4. **Profit is positive in many years under our cost assumption.** We estimate the cost at \$ 30,000 per acre. Then we compare that to the estimated revenue. Profit and Profit_per_Acre are often clearly positive in our results.

In some years, Profit_per_Acre is several thousand dollars or more. This means that, under our assumptions, strawberry farming can still be profitable per acre.

3.2 Correlation results

	Acres_Bearing	Price_per_CWT	Production_USD
Acres_Bearing	1.000	0.508	0.732
Price_per_CWT	0.508	1.000	0.946
Production_USD	0.732	0.946	1.000
Utilized_Production_CWT	-0.077	0.164	0.606
Profit_per_Acre	-0.166	0.574	0.322

	Utilized_Production_CWT	Profit_per_Acre
Acres_Bearing	-0.077	-0.166
Price_per_CWT	0.164	0.574
Production_USD	0.606	0.322
Utilized_Production_CWT	1.000	0.462
Profit_per_Acre	0.462	1.000

We also looked at the correlation between variables:

1. **Profit_per_Acre is positively related to Price_per_CWT.** When growers get a better price, they earn more profit per acre.
2. **Profit_per_Acre is also positively related to Utilized_Production_CWT.** Selling more usable fruit helps profit.
3. **Acres_Bearing has a weak or slightly negative relationship with Profit_per_Acre.** This suggests that when the industry becomes very large in total acres, profit per acre can go down. One reason is that the cost increases with the number of acres.

3.3 Model

Call:

```
lm(formula = Profit_per_Acre ~ Acres_Bearing + Price_per_CWT +
    Production_USD + Utilized_Production_CWT, data = strawberry)
```

Residuals:

21	22	23	24	25	26
-4.107	-18.005	-43.277	78.685	65.014	-78.311

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.881e+04	1.750e+03	-22.183	0.0287 *
Acres_Bearing	-6.387e-01	2.056e-02	-31.065	0.0205 *
Price_per_CWT	4.577e+02	7.423e+00	61.653	0.0103 *
Production_USD	-1.982e-06	3.038e-07	-6.524	0.0968 .


```
Utilized_Production_CWT  1.905e-03  3.981e-05  47.853  0.0133 *
```

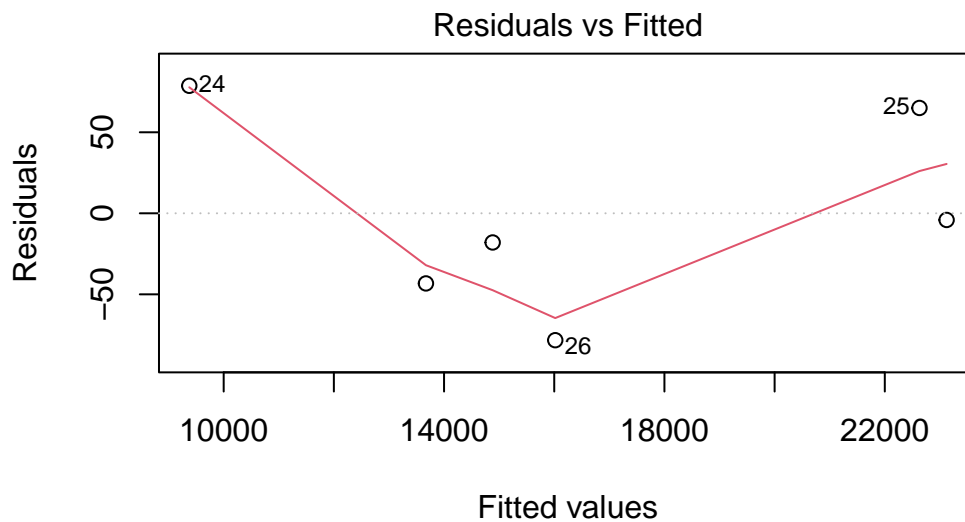
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 137 on 1 degrees of freedom

(22 observations deleted due to missingness)

Multiple R-squared: 0.9999, Adjusted R-squared: 0.9993

F-statistic: 1904 on 4 and 1 DF, p-value: 0.01718



We also ran a simple linear model. In the model, we found that:

1. The dependent variable (the thing we try to explain) is Profit_per_Acre.
2. The independent variables (the predictors) are Acres_Bearing, Price_per_CWT, Production_US, and Utilized_Production_CWT.

The regression shows:

1. Higher Price_per_CWT means higher Profit_per_Acre.
2. Higher Utilized_Production_CWT means higher Profit_per_Acre.
3. Higher Acres_Bearing means lower Profit_per_Acre.

Also, the model's R-squared is extremely high, indicating that these variables together explain almost all of the variation in Profit_per_Acre in our cleaned data. This supports the same story from EDA: profit per acre in strawberries mainly depends on getting a high selling price and getting a lot of marketable product per acre.

4. Conclusion

In summary, in recent years, strawberry planting areas and output in the United States have continued to increase, and prices have also risen slightly. Profitability is affected by both income and costs. Although there are some fluctuations, the overall trend is positive.