# strawberry

## 1.Import packages

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(janitor)
```

```

Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
library(zoo)
```

```

Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

## 2.Read and clean data

```
df <- read.csv("~/Documents/MA_615/Group6/ACRES BEARING+PRICE RECEIVED+PRODUCTION+UTILIZED -
head(df)
```

```
  Program Year          Period Week.Ending Geo.Level    State State.ANSI
1  CENSUS 2022            YEAR          NA  NATIONAL US TOTAL         NA
2  CENSUS 2017            YEAR          NA  NATIONAL US TOTAL         NA
3  SURVEY 2024 MARKETING YEAR           NA  NATIONAL US TOTAL         NA
4  SURVEY 2024            YEAR          NA  NATIONAL US TOTAL         NA
5  SURVEY 2024            YEAR          NA  NATIONAL US TOTAL         NA
6  SURVEY 2023 MARKETING YEAR           NA  NATIONAL US TOTAL         NA
  Ag.District Ag.District.Code County County.ANSI Zip.Code Region
1          NA                NA     NA          NA       NA     NA
2          NA                NA     NA          NA       NA     NA
3          NA                NA     NA          NA       NA     NA
4          NA                NA     NA          NA       NA     NA
5          NA                NA     NA          NA       NA     NA
6          NA                NA     NA          NA       NA     NA
  watershed_code Watershed    Commodity
1              0        NA STRAWBERRIES
2              0        NA STRAWBERRIES
3              0        NA STRAWBERRIES
4              0        NA STRAWBERRIES
5              0        NA STRAWBERRIES
6              0        NA STRAWBERRIES
                                        Data.Item Domain Domain.Category
1                         STRAWBERRIES - ACRES BEARING  TOTAL   NOT SPECIFIED
2                         STRAWBERRIES - ACRES BEARING  TOTAL   NOT SPECIFIED
3   STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT  TOTAL   NOT SPECIFIED
4              STRAWBERRIES - PRODUCTION, MEASURED IN $  TOTAL   NOT SPECIFIED
5 STRAWBERRIES, UTILIZED - PRODUCTION, MEASURED IN CWT  TOTAL   NOT SPECIFIED
6   STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT  TOTAL   NOT SPECIFIED
          Value CV....
1        70,709   15.4
```

```
2        58,117    4.3
3           124     NA
4 3,996,863,000     NA
5    32,225,500     NA
6           123     NA
```

```
new_df <- df %>%
  clean_names() %>%
  filter(df$Geo.Level == "NATIONAL", df$State == "US TOTAL") %>%
  select(year, data_item, value) %>%
  mutate(
    year = as.numeric(year),
    value = as.numeric(gsub(",", "", value))
  ) %>%
  filter(!is.na(value))

head(new_df)
```

```
  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
```

### 3.Pivot wider to make each variable a column

```
strawberry <- new_df %>%
  pivot_wider(
    names_from = data_item,
    values_from = value,
    values_fn = mean
  )

colnames(strawberry) <- c("Year", "Acres_Bearing", "Price_per_CWT",
                          "Production_USD", "Utilized_Production_CWT")

head(strawberry)
```

```
# 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
```

```
write_csv(strawberry, "clean_strawberry.csv")
```

### 4.Fill missing values

```
strawberry <- strawberry %>%
  arrange(Year) %>%
  mutate(
    Acres_Bearing = na.approx(Acres_Bearing, na.rm = FALSE),
    Price_per_CWT = na.approx(Price_per_CWT, na.rm = FALSE),
    Production_USD = na.approx(Production_USD, na.rm = FALSE),
    Utilized_Production_CWT = na.approx(Utilized_Production_CWT, na.rm = FALSE)
  )

head(strawberry)
```

```
# 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
```

### 5.Compute key profitability metrics

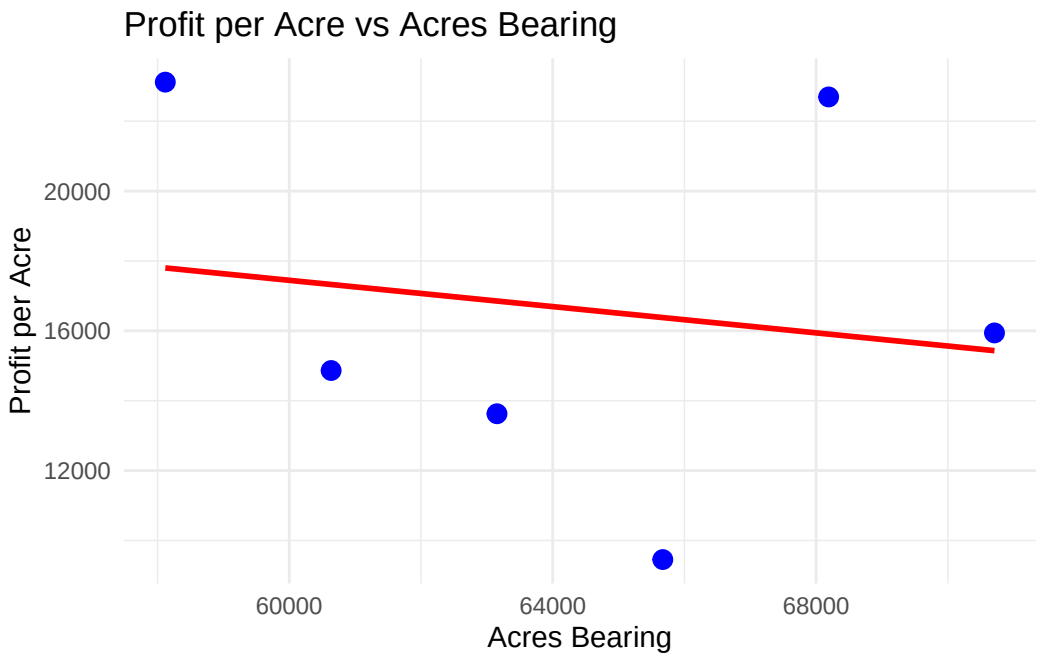
```r
cost_per_acre <- 30000
strawberry <- strawberry %>%
  mutate(
    Yield_per_Acre = Utilized_Production_CWT / Acres_Bearing,
    Revenue_Est = Price_per_CWT * Utilized_Production_CWT,
    Cost = Acres_Bearing * cost_per_acre,
    Profit = Revenue_Est - Cost,
    Profit_per_Acre = Profit / Acres_Bearing
  )
summary(strawberry)
```

```
      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
```

## Plot1

```
ggplot(strawberry %>%
         filter(!is.na(Acres_Bearing), !is.na(Profit_per_Acre)),
       aes(x = Acres_Bearing, y = Profit_per_Acre)) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Profit per Acre vs Acres Bearing",
       x = "Acres Bearing", y = "Profit per Acre") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



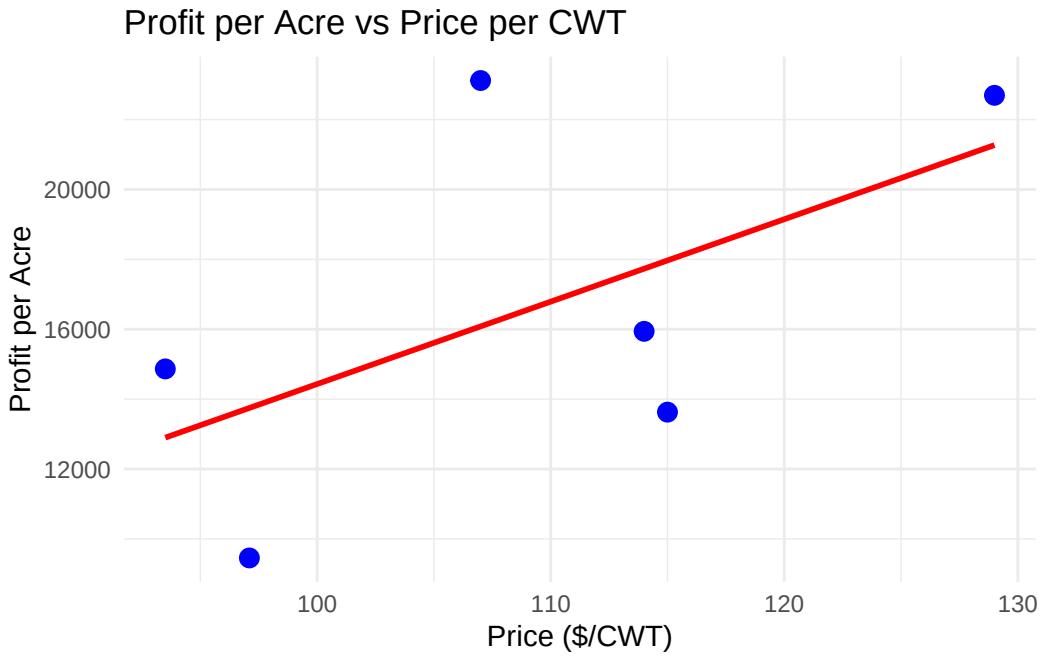

Explanation: This plot shows a slight negative relationship between Acres Bearing and Profit per Acre. As the total strawberry area increases, the profit per acre tends to decrease a little. This suggests that expanding the growing area might raise total cost faster than revenue. It means efficiency per acre may drop when production scale becomes larger.

### Plot2

```
ggplot(
  strawberry %>%
```

```
    filter(!is.na(Price_per_CWT), !is.na(Profit_per_Acre)),
  aes(x = Price_per_CWT, y = Profit_per_Acre)
) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Profit per Acre vs Price per CWT",
    x = "Price ($/CWT)",
    y = "Profit per Acre"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

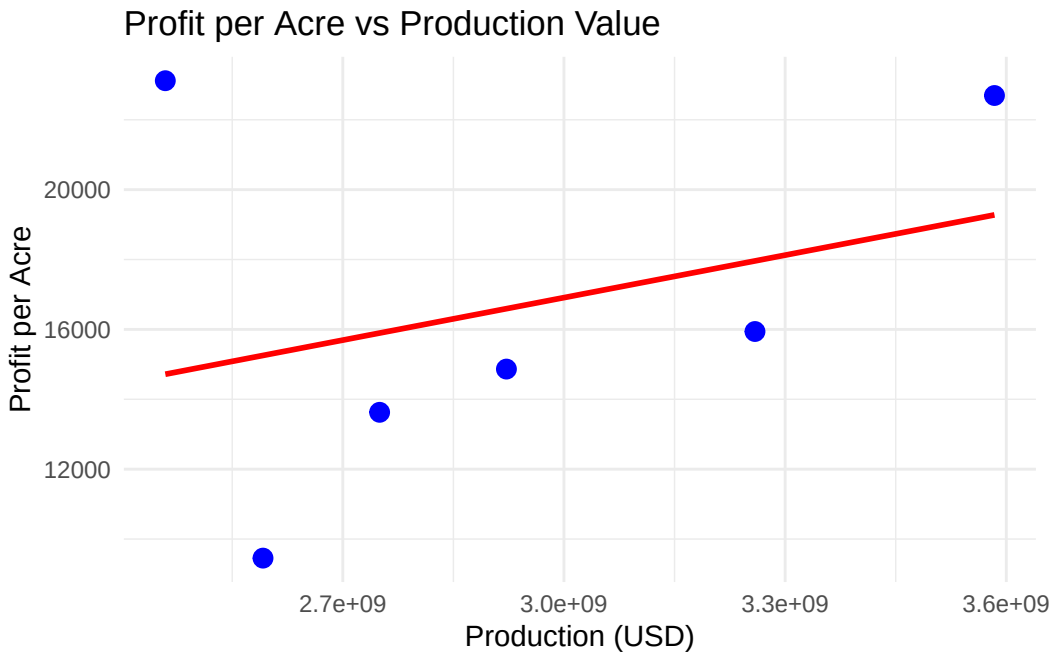



Explanation: This plot shows a clear positive relationship between Price per CWT and Profit per Acre. When the selling price of strawberries increases, the profit per acre also goes up. This means that market price is an important factor driving profitability. Higher prices directly improve revenue and help farmers earn more profit per unit of land.

### Plot3

```
ggplot(
  strawberry %>%
    filter(!is.na(Production_USD), !is.na(Profit_per_Acre)),
  aes(x = Production_USD, y = Profit_per_Acre)
) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Profit per Acre vs Production Value",
    x = "Production (USD)",
    y = "Profit per Acre"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```

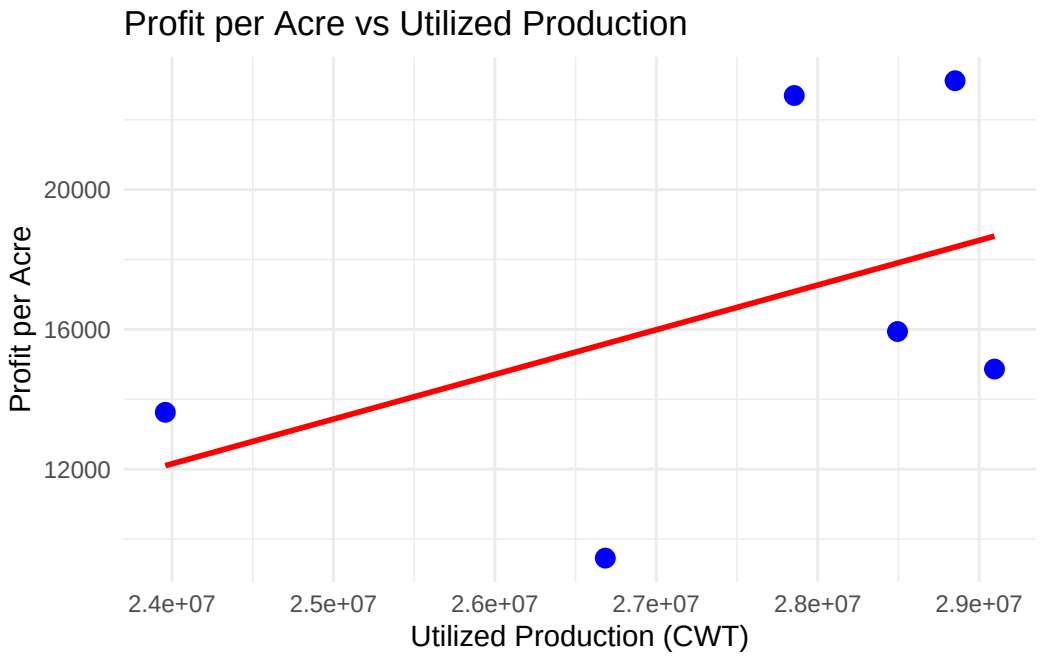



Explanation: This plot shows a positive relationship between Production Value and Profit per Acre. When the total production value increases, farms tend to earn more profit per acre. It means that higher sales or output usually lead to better profitability.

### Plot4

```
ggplot(
  strawberry %>%
    filter(!is.na(Utilized_Production_CWT), !is.na(Profit_per_Acre)),
  aes(x = Utilized_Production_CWT, y = Profit_per_Acre)
) +
  geom_point(color = "blue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Profit per Acre vs Utilized Production",
    x = "Utilized Production (CWT)",
    y = "Profit per Acre"
  ) +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```




Explanation: This plot shows a positive relationship between Utilized Production and Profit per Acre. As the amount of strawberries sold or used increases, the profit per acre also rises. This means that selling more products helps farmers gain higher efficiency and income.

## 6.Correlation

```
corr <- strawberry %>%
  select(Acres_Bearing, Price_per_CWT, Production_USD,
         Utilized_Production_CWT, Profit_per_Acre) %>%
  cor(use = "pairwise.complete.obs")
print(round(corr, 3))
```

```
                       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
```

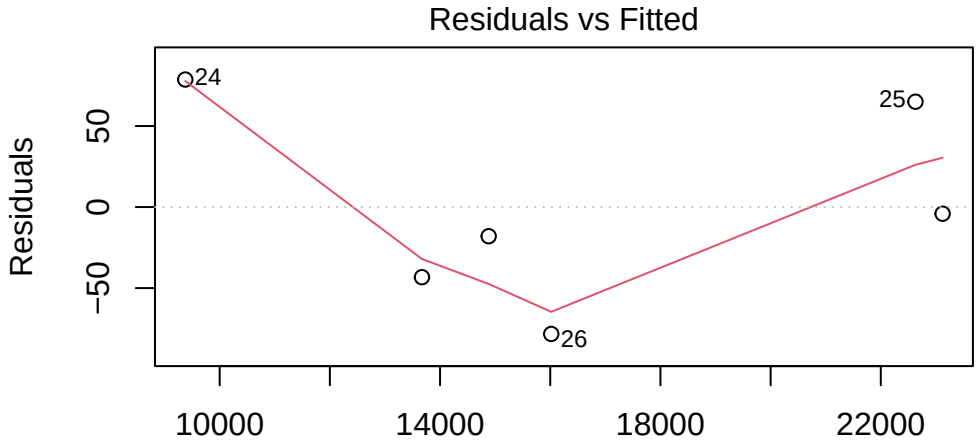
## 7.Model

```
model <- lm(Profit_per_Acre ~ Acres_Bearing + Price_per_CWT +
              Production_USD + Utilized_Production_CWT,
            data = strawberry)
summary(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
```

```r
plot(model, which = 1)
```




ı(Profit_per_Acre ~ Acres_Bearing + Price_per_CWT + Production_USD +