Wine Project

Sergio Martinez Barajas Sarah Marie Jennings

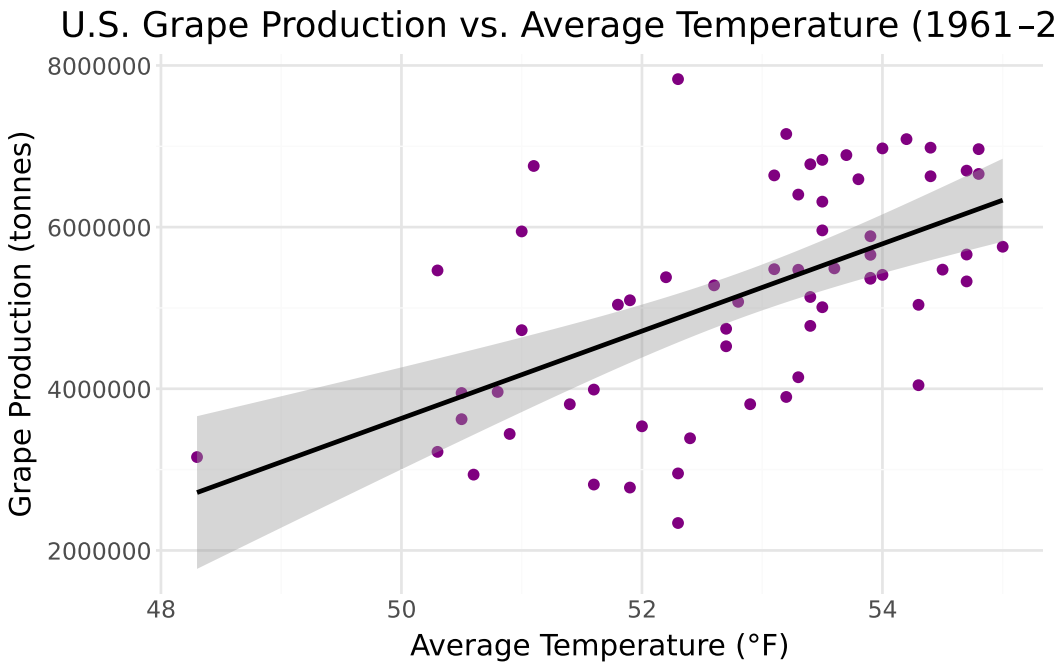
2025-11-28

```
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
import numpy as np

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

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

print("Slope (coefficient):", model.coef_[0])
print("Intercept:", model.intercept_)
print("R² score:", model.score(X, y))
```

```
Slope (coefficient): 539825.7588750448
Intercept: -23357267.297129147
R² score: 0.32630778692770923
```

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_cent

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

```
np.float64(-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.