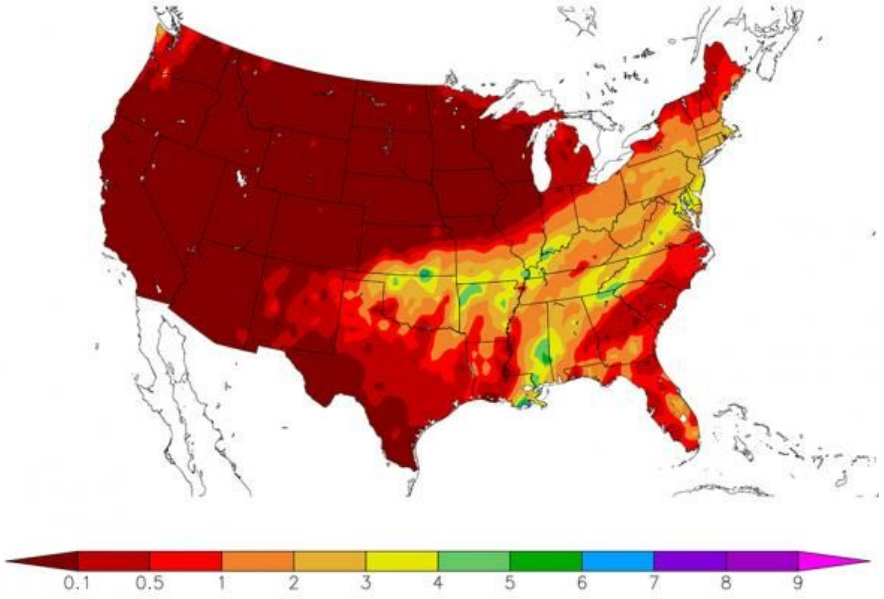
Drought Impact on Maple Syrup Production

Team B4: Chris Chang, Jacinto Lemarroy, Mengxin Li, Shiyu Ye, Ying Li

Table of Contents

Data Overview



Data Sources:

National Oceanic and Atmospheric Administration (NOAA), United States Department of Agriculture (USDA), National Drought Mitigation Center

Dataset Information

- **22** variables

Introduction and Problem Statement

Main Objective

Understand drought factors affecting maple syrup production

Information Sources

NOAA, weather websites, college research websites

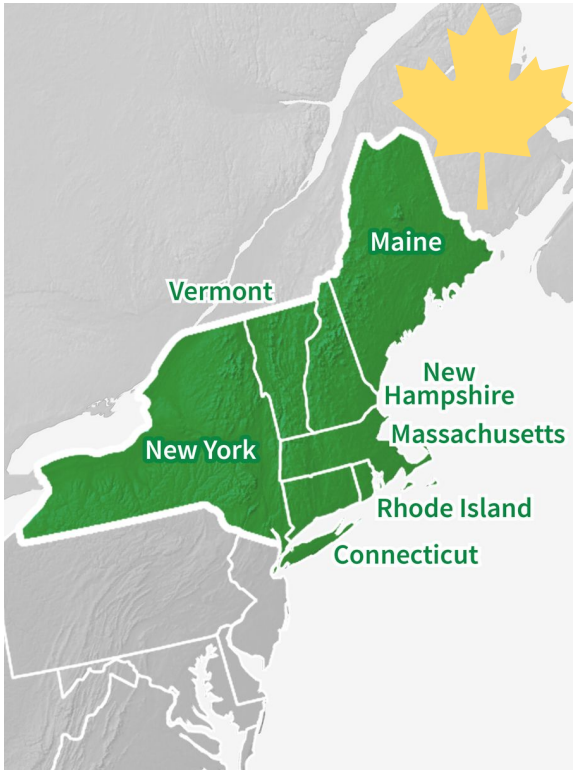
Sectors

Agriculture, Manufacturing, Energy, and Tourism

Initial Ideas

- Golf course
- New England livestock
- Maple trees

Data Overview



01 Drought-related Variables

None, D0, D1, D2, D3, D4, DSCI, SPI drought, SPI wet

02 Climate-related Variables

Average temperature, Extreme Air Temperature (Min & Max), Precipitation

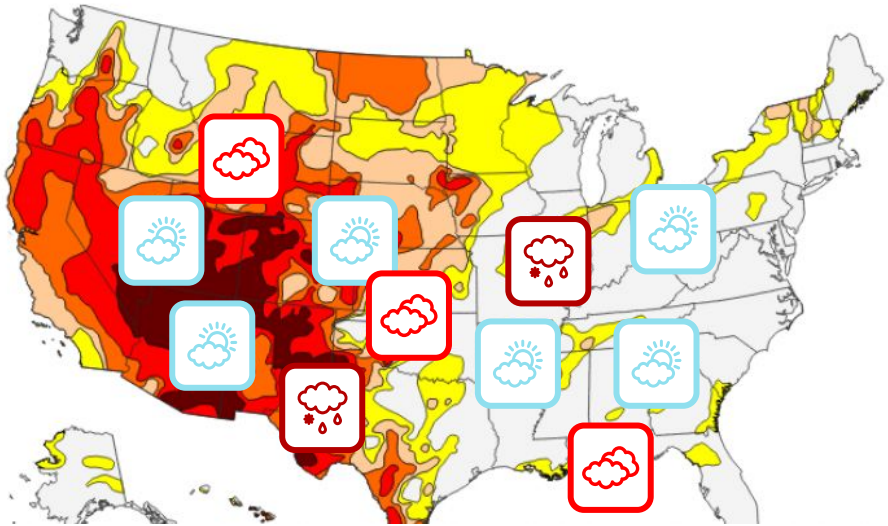
03 Basic Information

State, Year

04 Production Data

Gallons, Value, Number of Taps, Yield per Taps

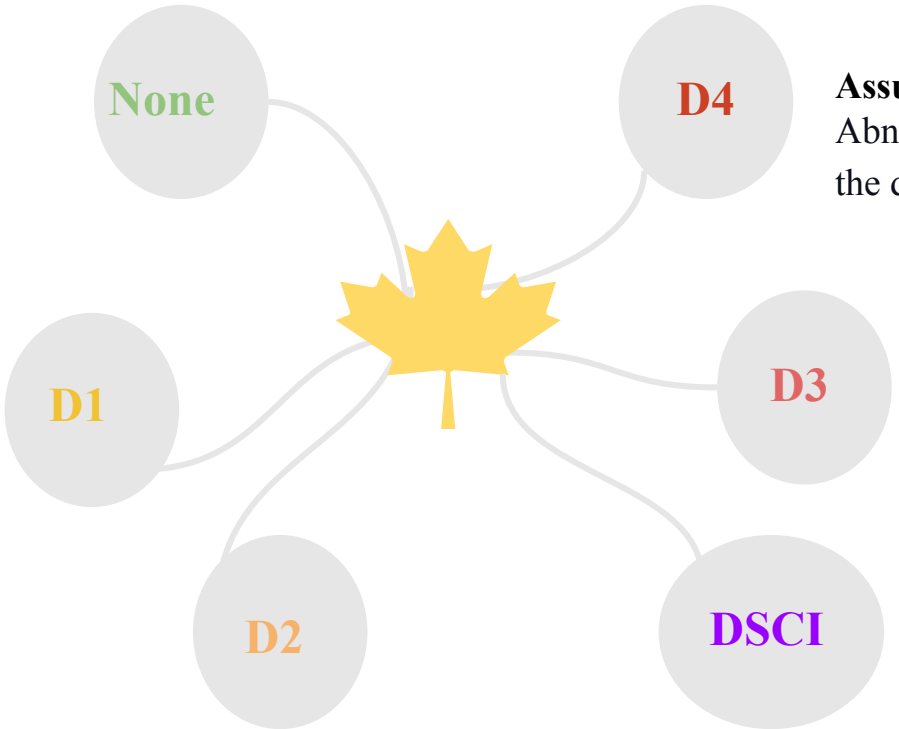
Data Preprocessing



- Missing value: drought related variable for **64** rows (1992 to 1999)
- Data processing goal:
  - Predict “**None**”, “**D0**”, “**D1**”, “**D2**”, “**D3**”, “**D4**”
  - Calculate **DSCI**
- Input: SPI : W0-W4; SPI: D0-D4; ‘State’, ‘Year’, ‘Avg. Temperature’, ‘Precipitation’, ‘SPI-Drought state’, ‘SPI-Wet state’

Feature Engineering

Accuracy Score
Train: 0.864
Test: 0.811



Assumption: D4 is 0
Abnormal Drought is not observed in the dataset

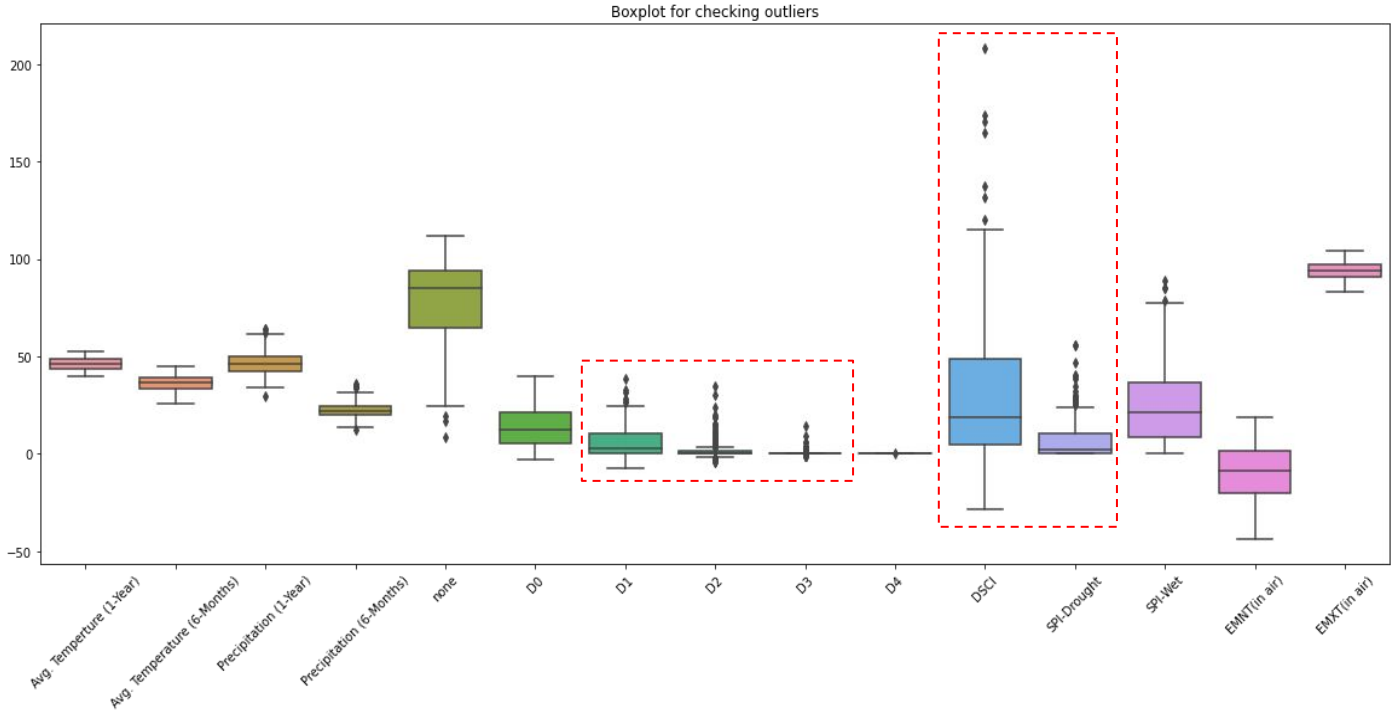
Accuracy Score
Train: 0.732
Test: 0.622

Calculated by:
D3 = 1 - None - D0 - D1 - D2 -D4

Accuracy Score
Train: 0.643
Test: 0.324

Calculated by:
DSCI = 1(D0) + 2(D1) + 3(D2) + 4(D3) + 5(D4)

Outlier Check



Real Value

Did not drop for the further analysis

Exploratory Data Analysis

01 Overall Production of Each State

02 Overall Correlation Heatmap

03 Average Temperature on Gallons

04 Average Temperature on Value

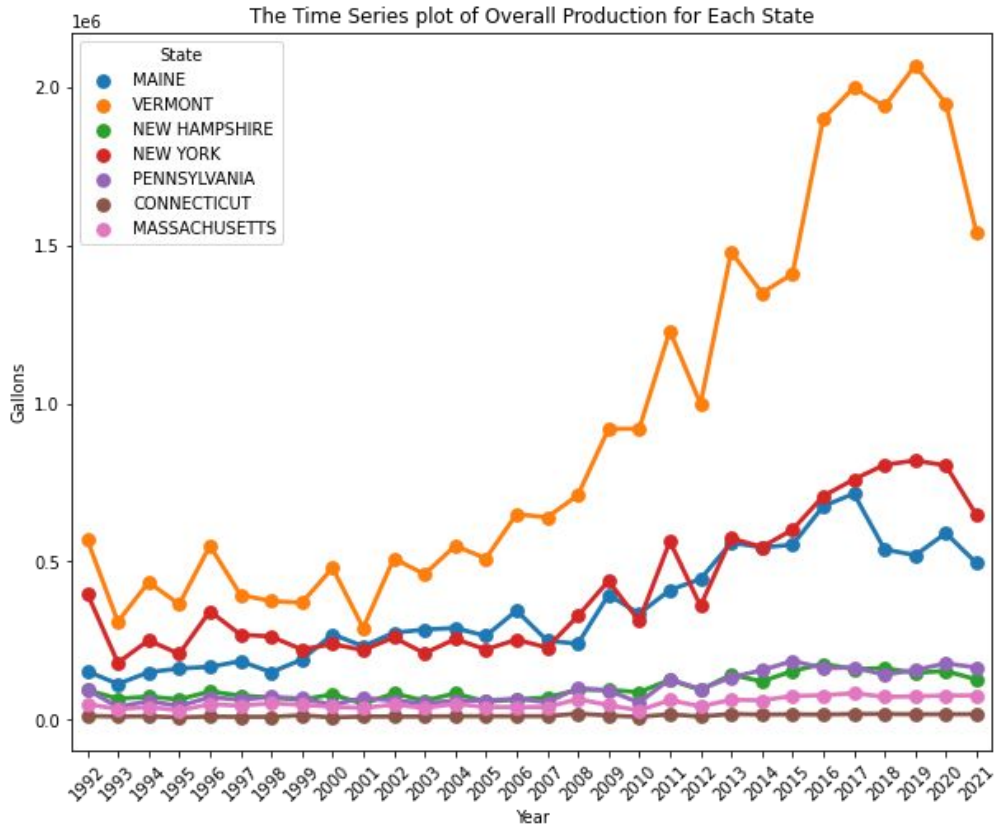
05 Extreme Maximum Air Temperature on Gallons

06 Extreme Minimum Air Temperature on Gallons

07 Average Precipitation on Gallons

08 SPI drought

Overall Production of Each State



Top Three States:

- Vermont
- New York
- Maine

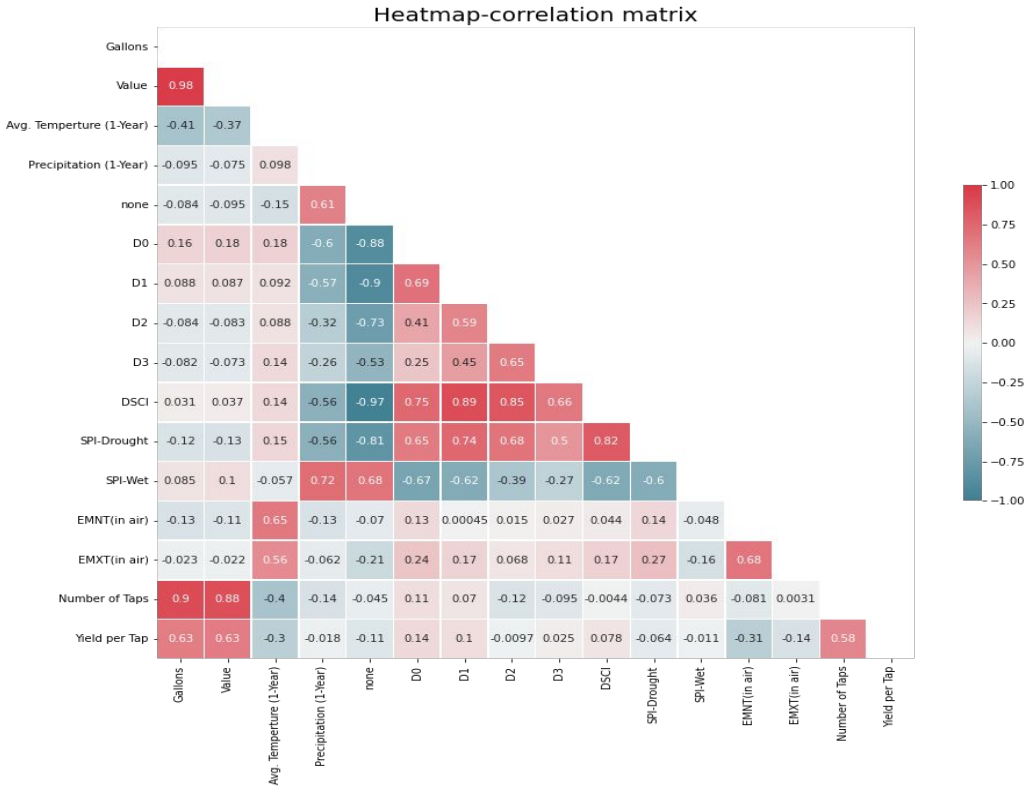
Decline Period:

- 2010, 2012, 2021
- Weather related

Overall Trend:

- Increase
- Stationary

Variable Heatmap - Correlation Matrix



Positive correlation

Gallons vs Value

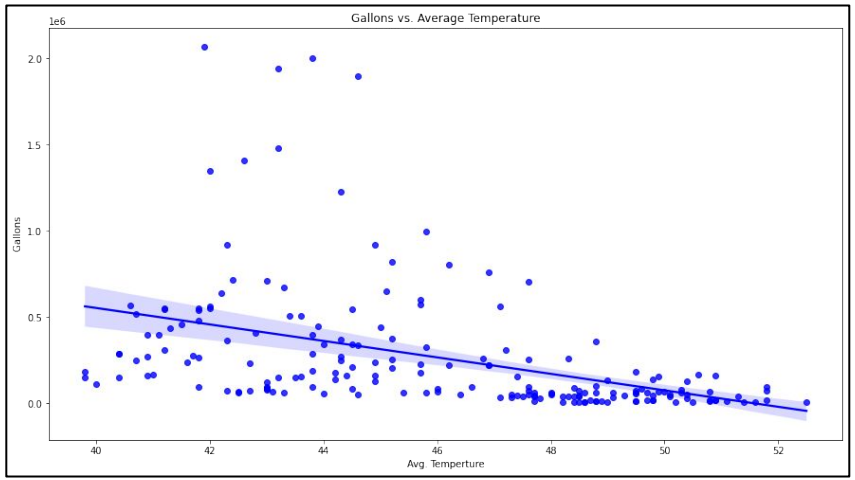
Negative correlation

Average Temperature vs Gallons
Average Temperature vs Value

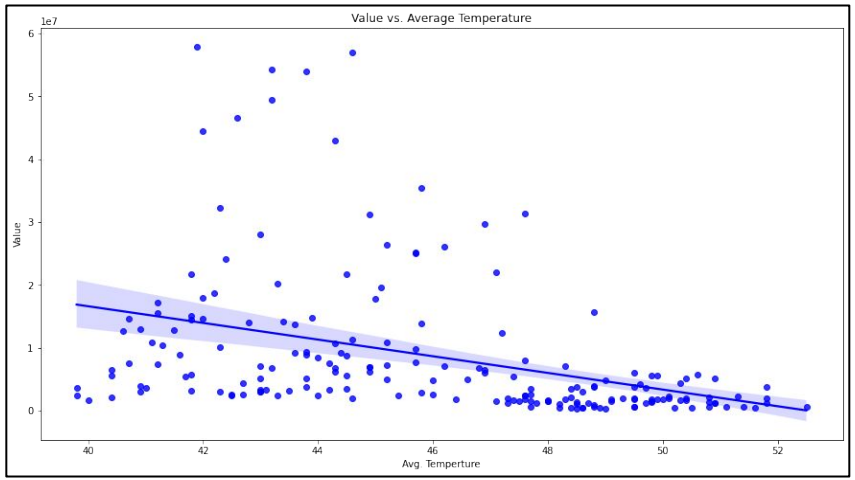
Multicollinearity Issue

Drought-related variables

Average Temperature vs. Gallons/Value

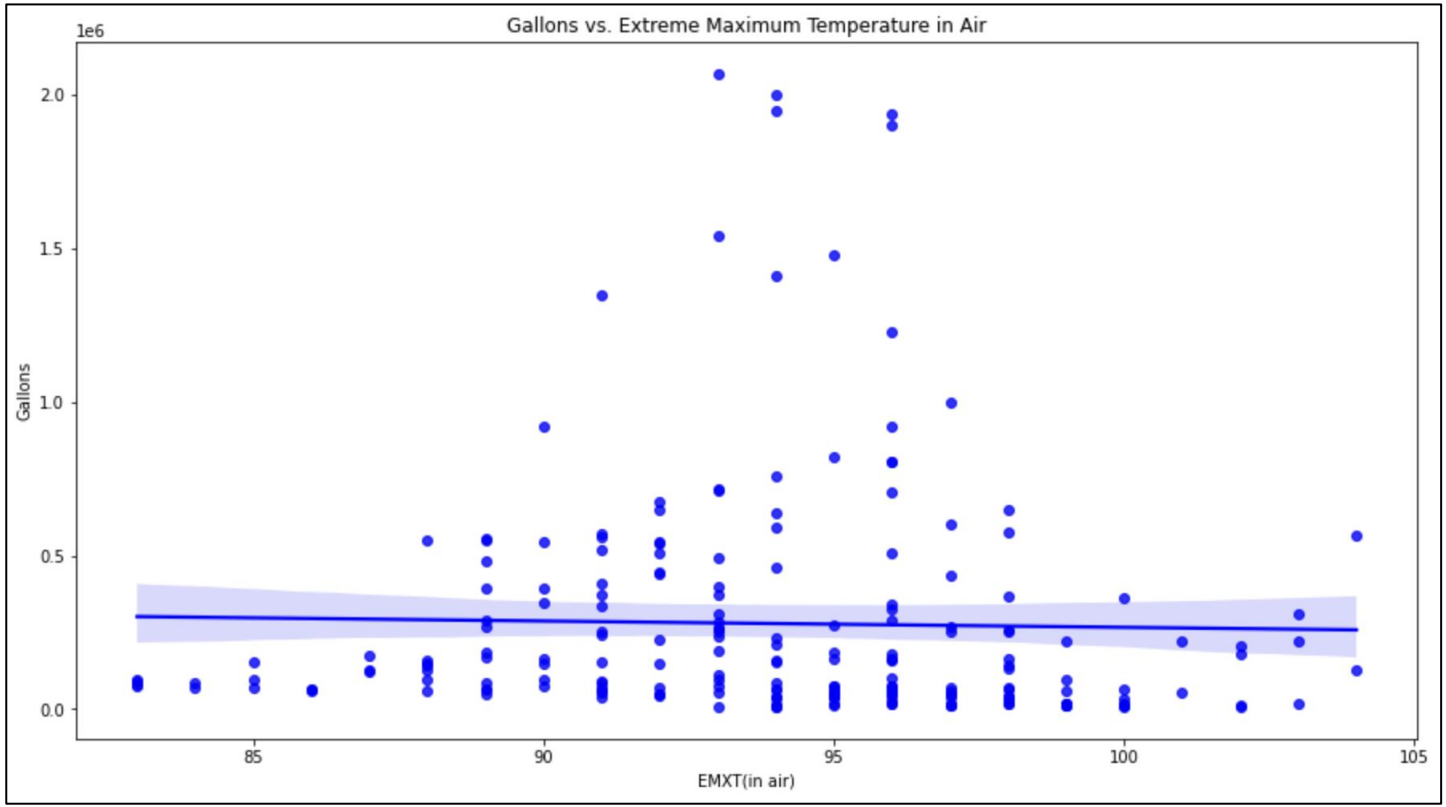


Slope = -3.84

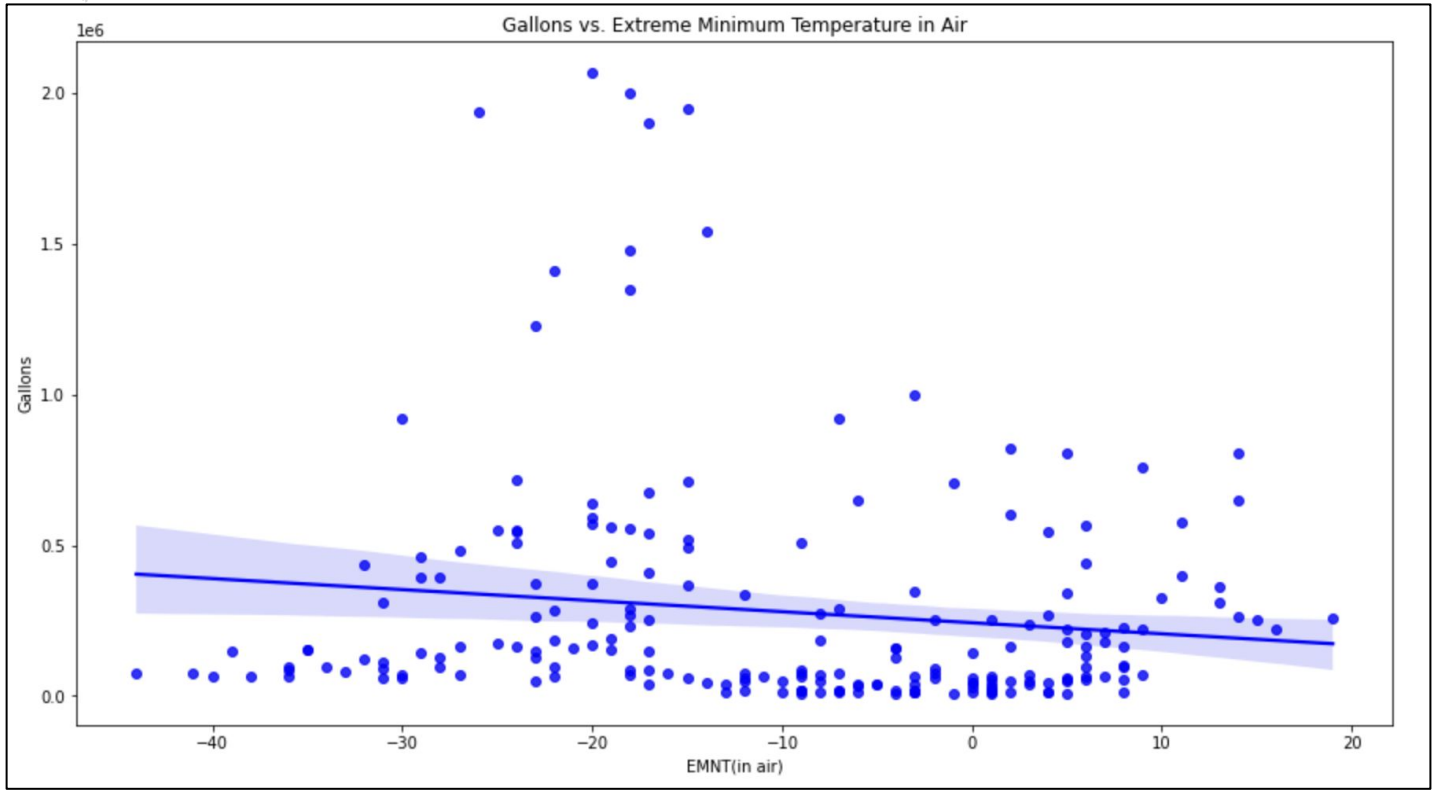


Slope = -1.10

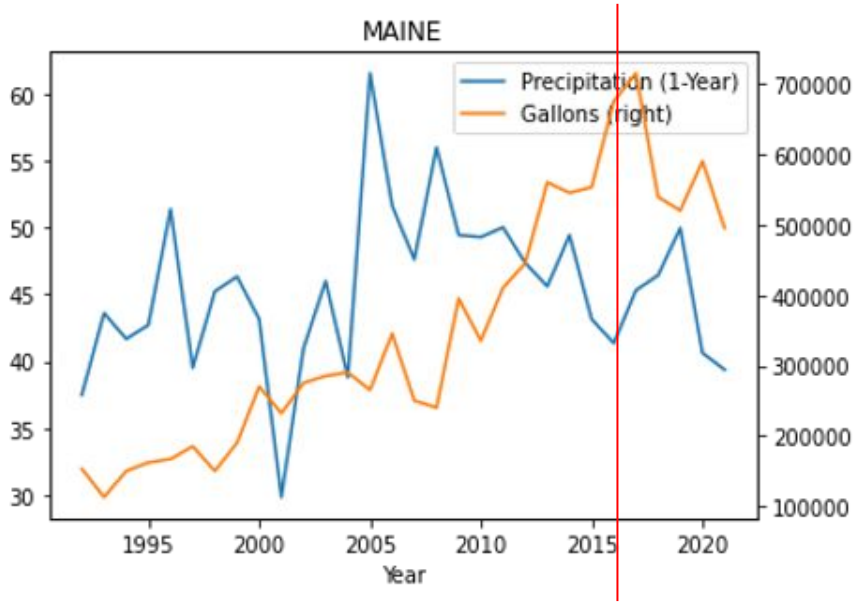
Extreme Maximum Air Temperature on Gallons

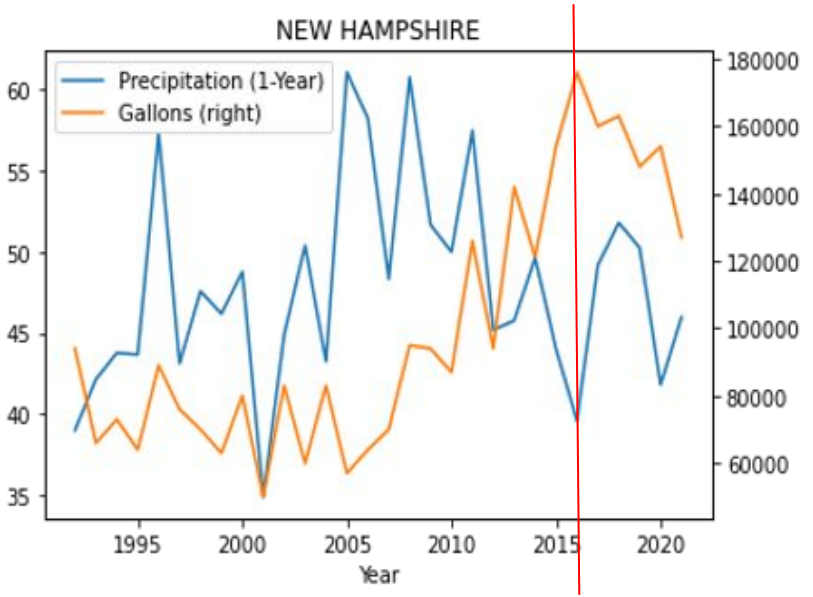


Extreme Minimum Air Temperature on Gallons

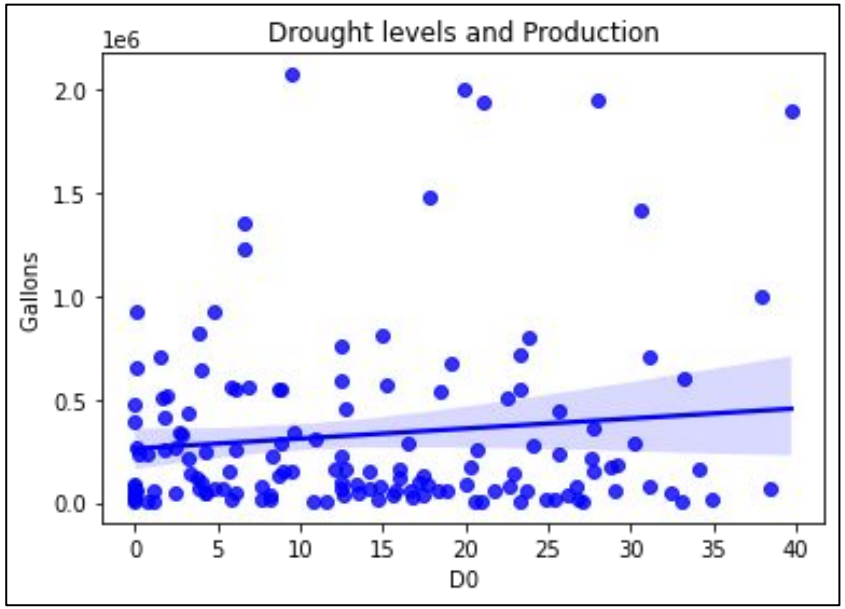


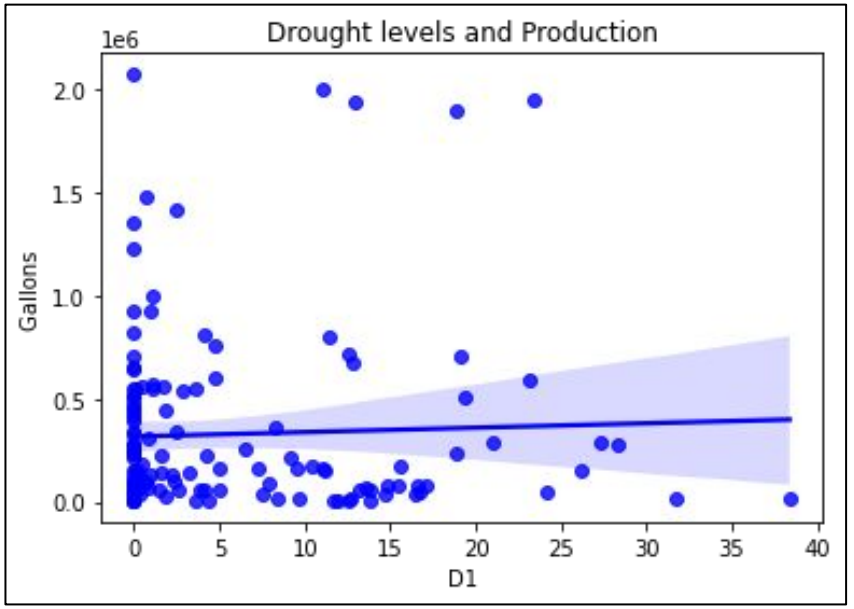
Precipitation vs. Gallons



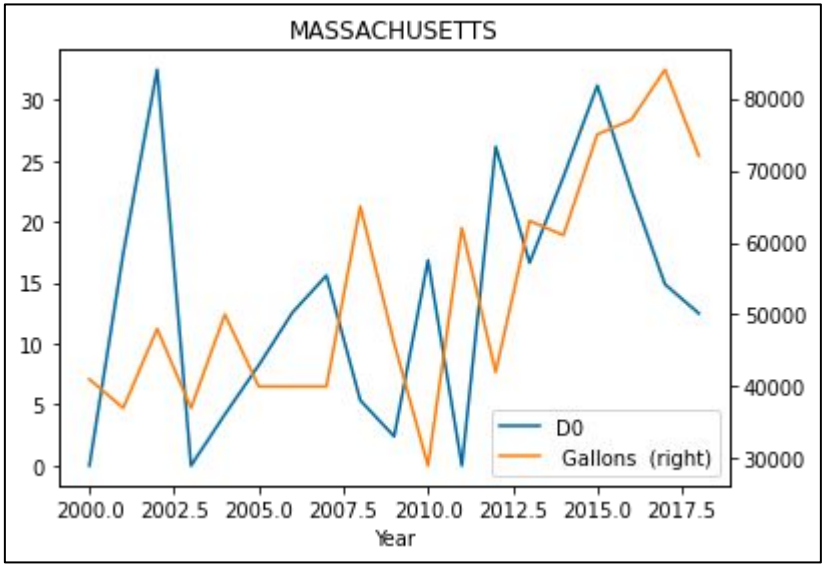


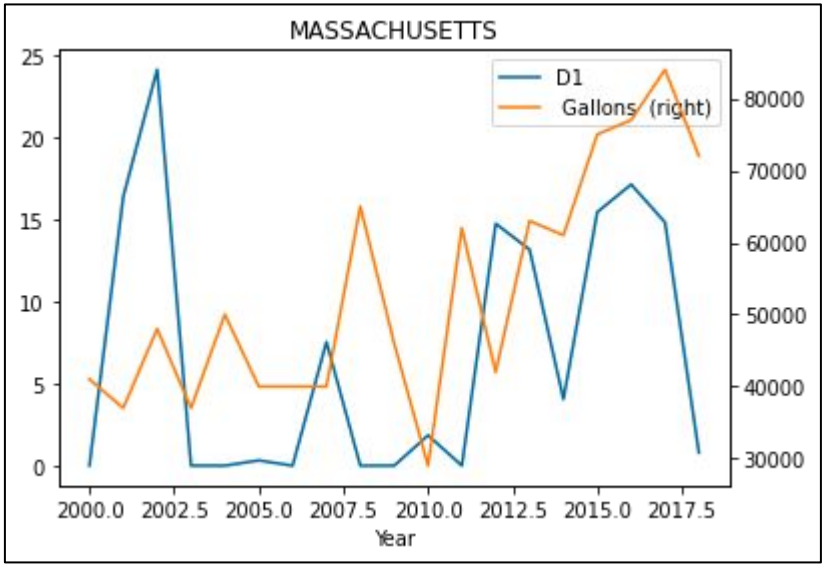
Drought Levels and Gallon Production



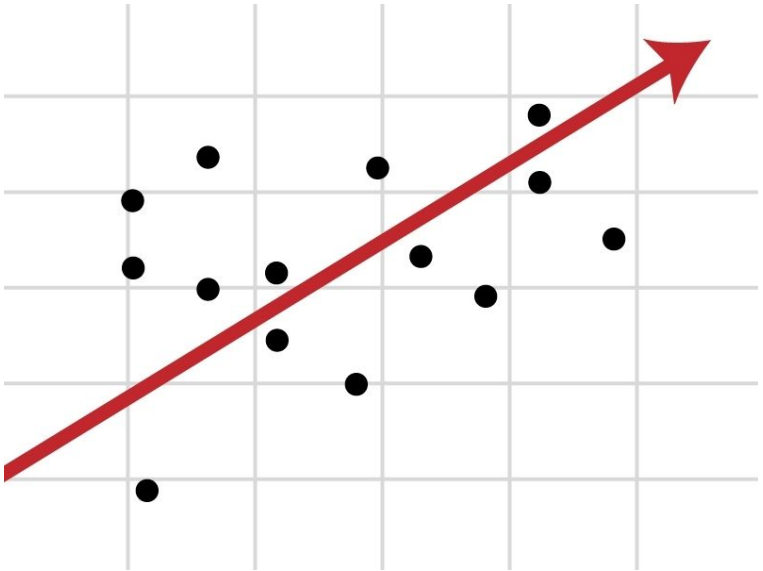


Drought Levels and Gallon Production





Regression Analysis



01 Regression of Each State

02 Single-variate Regression on all states

03 Multivariate Regression on all states

Regression of each state

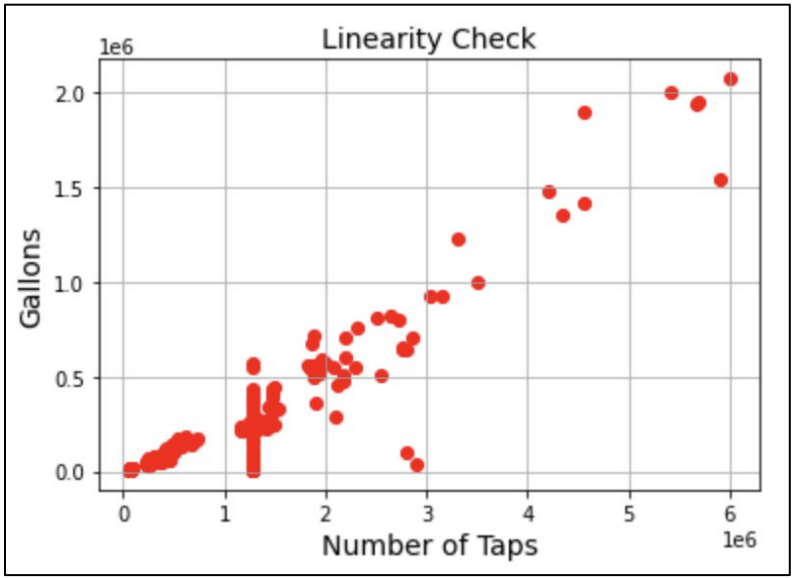
Linearity Check

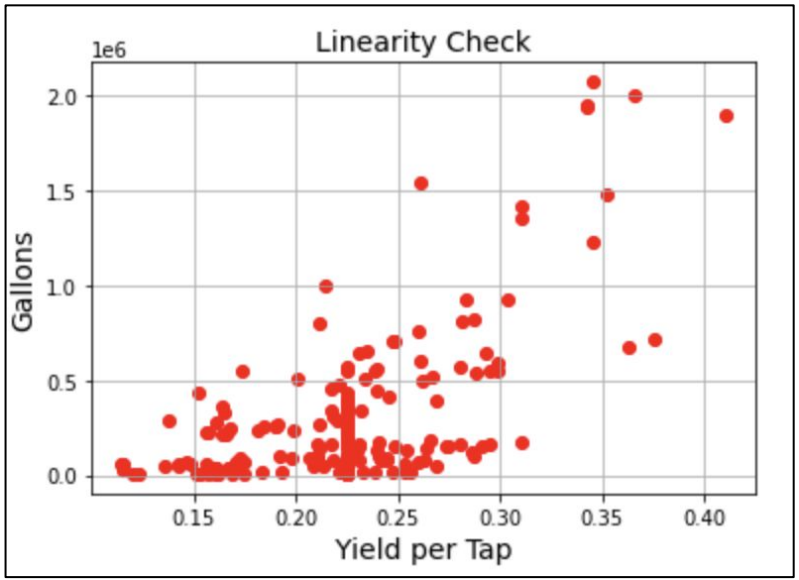
Regression

Feature Selection

Linearity Check

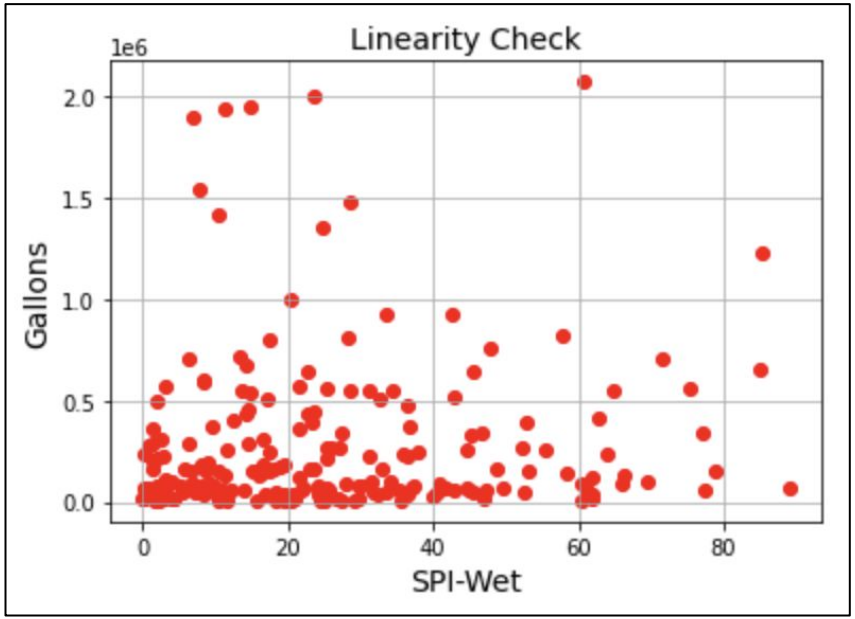
1. Variables '**Yield per Taps**' and '**Number of Taps**' is highly related to target variable '**Gallons**'.
2. On average, a tapped maple will produce **10** to **20** gallons of sap per tap. →Linear relationship?
3. Sugar concentration variability might affect the final production in gallons! → Or not??

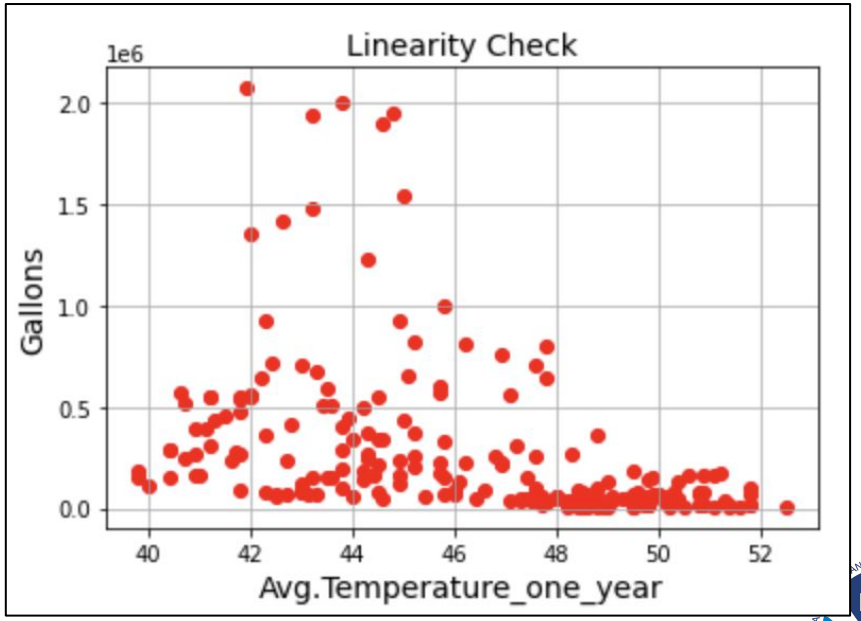




Linearity Check

1. Variables such as '**SPI Wet**' and '**Avg Temperature**' are distracting
2. However, in reality, 'Drought conditions are harder on the trees than wet conditions. The wet ground will supply plenty of moisture for the sap to run but the drought may mean lower sugar.'





Regression - Considering maple growing season

1. Growing season: Last year's September to the second year's February, but only precipitation!
2. Most of the states, R square increased!
3. Connecticut decreased dramatically!

Conclusion: Using precipitation during growing season and average temperature of the whole year!

State	R-square	R-square(growing season)	Difference
ME	52.80%	52.60%	-0.20%
VT	76.40%	82.70%	6.30%
MH	80.10%	80.30%	0.20%
NY	57.50%	58.00%	0.50%
PA	37.60%	41.70%	4.10%
CT	49.90%	42.90%	-7.00%
MA	56.60%	56.60%	0.00%

Feature Selection - Maine as an example

1. Eliminated D1, D2, D3, D4 because of these variables' collinearity with the D0.
2. Kept DSCI for saving the overall drought effect on regressions

VIF Factor	features
1.810565e+11	const
3.800000e+00	Avg.Temperature_one_year
4.600000e+00	Avg.Temperature_six_month
1.140000e+01	Precipitation_one_year
8.800000e+00	Precipitation_six_month
8.530393e+09	none
inf	D0
inf	D1
inf	D2
inf	D3
NaN	D4
inf	DSCI
1.420000e+01	SPI-Drought
1.760000e+01	SPI-Wet
4.100000e+00	EMNT(in air)
3.700000e+00	EMXT(in air)
4.700000e+00	Number of Taps
6.300000e+00	Yield per Tap

OLS Regression Results

Dep. Variable:	y	R-squared:	0.528
Model:	OLS	Adj. R-squared:	0.194
Method:	Least Squares	F-statistic:	1.583
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.188
Time:	19:29:55	Log-Likelihood:	-392.85
No. Observations:	30	AIC:	811.7
Df Residuals:	17	BIC:	829.9
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.2e+08	1.27e+10	0.049	0.962	-2.62e+10	2.74e+10
Avg.Temperature_one_year	-3.086e+04	3.69e+04	-0.836	0.415	-1.09e+05	4.7e+04
Precipitation_one_year	-7331.6186	9886.484	-0.742	0.468	-2.82e+04	1.35e+04
none	-6.188e+06	1.27e+08	-0.049	0.962	-2.74e+08	2.62e+08
D0	1.127e+13	1.17e+13	0.964	0.349	-1.34e+13	3.6e+13
D1	2.255e+13	2.34e+13	0.964	0.349	-2.68e+13	7.19e+13
D2	3.382e+13	3.51e+13	0.964	0.349	-4.02e+13	1.08e+14
D3	4.51e+13	4.68e+13	0.964	0.349	-5.36e+13	1.44e+14
DSCI	-1.127e+13	1.17e+13	-0.964	0.349	-3.6e+13	1.34e+13
SPI-Drought	-1.928e+04	7168.590	-2.689	0.016	-3.44e+04	-4152.966
SPI-Wet	5438.2765	3591.113	1.514	0.148	-2138.309	1.3e+04
EMNT(in air)	3190.4185	6839.786	0.466	0.647	-1.12e+04	1.76e+04
EMXT(in air)	6206.0932	2.3e+04	0.270	0.790	-4.23e+04	5.47e+04

Omnibus:	0.431	Durbin-Watson:	0.962
Prob(Omnibus):	0.806	Jarque-Bera (JB):	0.362
Skew:	0.243	Prob(JB):	0.835
Kurtosis:	2.771	Cond. No.	3.29e+11

Feature Selection - Maine as an example

1. D0 which was previously insignificant now became statistically significant.
2. The SPI drought stayed significant.
3. Variable ‘none’ is negative. → But not significant!

Limitation:

Most of the variables still not significant!

Check all states!!!

OLS Regression Results

Dep. Variable:	y	R-squared:	0.437
Model:	OLS	Adj. R-squared:	0.222
Method:	Least Squares	F-statistic:	2.035
Date:	Sat, 30 Apr 2022	Prob (F-statistic):	0.0918
Time:	19:30:05	Log-Likelihood:	-395.49
No. Observations:	30	AIC:	809.0
Df Residuals:	21	BIC:	821.6
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.712e+05	2e+06	0.285	0.778	-3.59e+06	4.74e+06
Avg.Temperature_one_year	1.444e+04	2.51e+04	0.577	0.570	-3.77e+04	6.65e+04
Precipitation_one_year	-1.099e+04	9317.630	-1.179	0.252	-3.04e+04	8390.840
none	-3808.0987	3415.346	-1.115	0.277	-1.09e+04	3294.503
D0	1.164e+04	6486.454	1.795	0.087	-1846.981	2.51e+04
SPI-Drought	-1.169e+04	4463.465	-2.619	0.016	-2.1e+04	-2409.188
SPI-Wet	5648.9504	3451.947	1.636	0.117	-1529.766	1.28e+04
EMNT(in air)	-2646.5757	5818.757	-0.455	0.654	-1.47e+04	9454.191
EMXT(in air)	-2967.7046	1.98e+04	-0.150	0.882	-4.42e+04	3.83e+04

Omnibus:	1.870	Durbin-Watson:	0.995
Prob(Omnibus):	0.393	Jarque-Bera (JB):	1.432
Skew:	0.341	Prob(JB):	0.489
Kurtosis:	2.175	Cond. No.	1.00e+04

Lag-Effect Regression

- None - No lag effect
- Incorporating previous year's precipitation explains more variance in maple syrup production.
- However, this is not the case for average temperature.

	Lag effect column	R_squared
One_Year	None	0.459
	Temperature	0.370
	Precipitation	0.467
	Both	0.388
Six_month	None	0.317
	Temperature	0.301
	Precipitation	0.303
	Both	0.272

Multivariate Regression (All States)

OLS Regression Results

Dep. Variable:	Gallons	R-squared:	0.473
Model:	OLS	Adj. R-squared:	0.438
Method:	Least Squares	F-statistic:	13.51
Date:	Sun, 01 May 2022	Prob (F-statistic):	3.82e-21
Time:	17:55:32	Log-Likelihood:	-2933.1
No. Observations:	210	AIC:	5894.
Df Residuals:	196	BIC:	5941.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.111e+09	5.51e+09	-0.202	0.840	-1.2e+10	9.75e+09
Avg.Temperature_one_year	-7.574e+04	8818.466	-8.588	0.000	-9.31e+04	-5.83e+04
Precipitation_six_month	-1.864e+04	7552.788	-2.468	0.014	-3.35e+04	-3743.322
none	1.113e+07	5.51e+07	0.202	0.840	-9.75e+07	1.2e+08
D0	-6.437e+11	2.82e+12	-0.228	0.820	-6.21e+12	4.92e+12
D1	-1.287e+12	5.65e+12	-0.228	0.820	-1.24e+13	9.85e+12
D2	-9.656e+11	4.24e+12	-0.228	0.820	-9.32e+12	7.39e+12
D2	-9.656e+11	4.24e+12	-0.228	0.820	-9.32e+12	7.39e+12
D3	-2.575e+12	1.13e+13	-0.228	0.820	-2.48e+13	1.97e+13
D4	-3.219e+12	1.41e+13	-0.228	0.820	-3.11e+13	2.46e+13
DSCI	6.437e+11	2.82e+12	0.228	0.820	-4.92e+12	6.21e+12
SPI-Drought	-1.909e+04	3613.510	-5.284	0.000	-2.62e+04	-1.2e+04
SPI-Wet	8488.4794	1562.595	5.432	0.000	5406.821	1.16e+04
EMNT(in air)	3422.8589	2292.318	1.493	0.137	-1097.916	7943.633
EMXT(in air)	2.243e+04	6767.305	3.315	0.001	9088.080	3.58e+04

Conclusions and Challenges

- Conclusion
  - Average temperature (1-year), Precipitation (6-months), SPI Wet, SPI Drought, and Extreme maximum temperature were significant indicators that explained the variation in gallon production in all states.
  -
- Limitations and Challenges
  - Limited data volume to draw causal inference about how drought affects maple production or make accurate predictions on production with appropriate ML models
  - Data granularity issues
    - Data sources had variables that were on different scale, different level, different years than ours,
  - Lack of response from maple syrup connoisseurs, maple syrup companies, or government agencies, farms
  - Put more effort in learning about factors that influence the growth of maple tree and syrup production

Potential Improvements

- Add more features (Soil, Geological variables etc.) to improve model.
- Find granular data (monthly, weekly, daily) to perform more accurate analysis.
- Design dashboard (Tableau, PowerBI) for data visualization.
- Create database to store and update data since our data is continuous.
- Expand drought impact analysis on other crops - eg. apples, corn, soybeans.

THANK YOU