Harvesting Resilience: Navigating Agriculture in a Changing Climate

In [1]:
```
import numpy as np
import pandas as pd
import seaborn
from matplotlib import pyplot as plt
import duckdb
from scipy.stats import ttest_ind
from scipy.stats import pearsonr
from scipy.stats import spearmanr
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
```

Introduction

Key Terms

Key terms in our report that we would like to clarify are quantity, quality, yield, and emissions. When we discuss the quantity of crops produced, the relevant unit of measurement is in either pounds or CWT, a standard unit of weight or mass used in certain commodity markets. When we discuss the quality of agricultural goods, we are referring to whether a crop meets particular standards such as taste, size, and appearance. Crop yield is measured in area harvested per acre planted. Fianlly, CO2 emissions relate to the amount of carbon dioxide in the air, which is affected by burning fossil fuels (coal, natural gas, and oil), solid waste, trees and other biological materials, and chemical reactions (e.g., cement production).

Research Questions

Our project hopes to answer 2 reasearch questions:

---

1. What are the effects of climate change (CO2 emissions, rising surface temperatures, droughts) on the quality and quantity of crop production in the US?

2. Which crops are most negatively impacted by climate change? More specifically, how drastically do changes in climate effect the quantity and quality of agricultural yields?

To address these questions, we have gathered data from diverse and reputable sources. Our climate change data encompasses information on CO2 emissions, surface temperature changes, and drought occurrences in the United States. Additionally, our crop data sources provide insights into corn, tobacco, potatoes, coffee, and rice.

Climate change is a major problem for our world right now. The air is being polluted, ecosystems are dying, and weather patterns are changing and becoming more extreme. All of these issues correlate to the factors of climate change we are analyzing in this study. It is hard for people to make change when they cannot physically see the effects of climate change in their community. This report aims to inform how climate change is effecting our planet, and specifically crops in the United States. People understand that food is an important part of daily routines, culture and survival, so showing data on how climate change is effecting crop growth in the US can help people realize how badly we need to make change when it comes to the climate.

The US is a major source of agriculture, epsecially in the crops we are looking at in our study. If our findings indicate that climate change is negatively effecting the quality and quantity of crop growth, we know it will not be long until the access to various crops decreases for those supplied by agriculture in the US. The only way to stop this from happening is by spreading awareness of the dangers of climate change to encourage people to try and help the planet by reducing waste and pollution. Our report does this by presenting data, models, and interpretation of analysis to show the actual numbers behind the issue of climate change related to crop growth.

Summary of Findings

From our analysis we have determined the following:

---

1. CO2 emissions have a significantly negative effect on coffee bean yield in the US. Because of this, as the climate continues to get worse, coffee prices are likely to continuously rise as domestic supply becomes more scarce.

1. There appears to be no negative impact between CO2 emissions and the yields of our other crops of interest. However, this does not mean that climate change is irrelevant; It only indicates that with the data provided, our team did not find any negative correlation. This suggests that coffee beans are much more sensitive to the effects of climate change compared to other agricultural goods

Data Description

Our World In Data https://ourworldindata.org/co2-emissions (https://ourworldindata.org/co2-emissions)

What are the observations (rows) and the attributes (columns)?

Each observation (row) represents a country during a specific year. The attributes for each observation represent metics related to CO2 emissions such as Population, CO2 level, CO2 growth, CO2 per capita, and more.

Why was this dataset created?

This dataset was developed to illustrate global CO2 emissions over several centuries in a comprehensible manner, aiming to make this information accessible for research and analysis. Our World In Data recognizes the potential of data in addressing various global challenges. Through this dataset, they seek to explore solutions related to CO2 emissions and other world problems.

Who funded the creation of the dataset?

Our World in Data is a collaborative effort between researchers at the University of Oxford, who are the scientific editors of the website content; and the non-profit organization Global Change Data Lab (GCDL), who publishes and maintains the website and the data tools that make their work possible.

What processes might have influenced what data was observed and recorded and what was not?

Several factors could have influenced the selection of observed and recorded data, as well as what data was excluded. One such factor could be the fact that the US is this organization's main source of funding, which may result in a higher concentration of data from the US compared to other countries. Additionally, historical data from distant periods might be missing due to the lack of accurate documentation of events and statistics during those times. Finally, the accessibility of particular regions/countries could significantly impact data collection. After all, remote areas are often harder to access, making it challenging to gather comprehensive and representative data from those regions leading to gaps in information.

What preprocessing was done, and how did the data come to be in the form that you are using?

We were able to download this data off the website by going to a GitHub repository from a link and then downloading the file as a csv. We then cleaned the data to get the rows and columns according to the focus of our research question.

If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

Our World In Data acknowledges the significance of data collection. The mission behind Our World in Data is grounded in the belief that existing research and data is often hidden behind academic jargon and paywalls. Our World in Datais emphasizes the need to unlock this knowledge, making data accessible and understandable to the world in order to address major global issues such poverty, disease, hunger, climate change, war, existential risks, and inequality. This specific dataset is intended for combating climate change, aiming to alleviate global suffering.

Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted on Github, in a Cornell Google Drive or Cornell Box).

The raw data can be found by visiting the website in the top URL link. Clicking 'download data' on that page will redirect you to a github repository where you can download the updated CO2 data and codebook as a csv file.

https://github.com/owid/co2-data (https://github.com/owid/co2-data)

https://github.com/OrenSte/ORIEOS/blob/main/owid-co2-data.csv (https://github.com/OrenSte/ORIEOS/blob/main/owid-co2-data.csv)

US Drought Monitor https://droughtmonitor.unl.edu/DmData/DataDownload/ComprehensiveStatistics.aspx (https://droughtmonitor.unl.edu/DmData/DataDownload/ComprehensiveStatistics.aspx)

What are the observations (rows) and the attributes (columns)?

This dataset presents each observation as a unique combination of state and date spanning from 1990 to the present day. The attributes in this dataset represent various categories: "None", indicating the percentage of state area unaffected by drought, "abnormally dry" (D0), the percentage of state area potentially entering or exiting drought "moderate" (D1), "severe" (D2), "extreme" (D3), and "exceptional" (D4). This classification system allows for a detailed analysis of drought conditions, capturing various levels of severity across different states and dates.

Why was this dataset created?

The U.S. Drought Monitor is a map released every Thursday displaying where drought is and how bad it is across the U.S. and its territories. The USDA uses the Drought Monitor to trigger disaster declarations and eligibility for low-interest loans. The Farm Service Agency uses it to help determine eligibility for their Livestock Forage Disaster Program, and the Internal Revenue Service uses it for tax deferral on forced livestock sales due to drought. State, local, tribal and basin-level decision makers use it to trigger drought responses or declare drought emergencies, ideally along with other local indicators of drought. Though these various entities use the USDM, it does not in itself trigger any political actions.

Who funded the creation of the dataset?

The U.S. Drought Monitor is produced through a partnership between the National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture and the National Oceanic and Atmospheric Administration.

What processes might have influenced what data was observed and recorded and what was not?

This dataset appears to be very complete for filling in values for all the attributes and observations, however we notice that the dates do not cover every day of the year and instead cover intervals of days. The data for certain days might not have been recorded because they do not have enough resources to be able to record drought levels for every state for every day so they used period of times. This might mean the information represents an average as oppose to exact numbers every day.

What preprocessing was done, and how did the data come to be in the form that you are using?

On the website you can download specific data as a csv file. We chose what data we think would be good to analyze for our report, downloaded the file, and subsequently cleaned the data.

If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

They do want/allow us to use this data, the state: If you reproduce the U.S. Drought Monitor map, please use this wording: The U.S. Drought Monitor is jointly produced by the National Drought Mitigation Center at the University of Nebraska-Lincoln, the United States Department of Agriculture, and the National Oceanic and Atmospheric Administration. Map courtesy of NDMC.

Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted on Github, in a Cornell Google Drive or Cornell Box).

This data can be found by going to the website and clicking the data tab and selecting data downlaod and choosing percent area from 1/1/2000 to 10/13/2023 where area type is national. All other markdowns left as they were.

https://github.com/OrenSte/ORIEOS/blob/main/dm_export_19900101_20231011.csv (https://github.com/OrenSte/ORIEOS/blob/main/dm_export_19900101_20231011.csv)

National Agriculture Statistics Servive (Corn Quality) https://quickstats.nass.usda.gov/ (https://quickstats.nass.usda.gov/)

What are the observations (rows) and the attributes (columns)?

This dataset presents each observation as a unique combination of state and date spanning from 2014 to the present day. The attributes in this dataset represent various categories: "Data Item" is the condition of the corn measured in terms of excellent, fair, good, poor and very poor, "value" is the percent of corn that falls in each of the data item categories.

Why was this dataset created?

NASS dataset are created and used to support research, education, and advocacy for the future of agriculture anyone's your region.

Who funded the creation of the dataset?

The National Agricultural Statistics Service (NASS) offers Quick Stats, an on-line database containing official published aggregate estimates related to U.S. agricultural production. NASS develops these estimates from data collected through: hundreds of sample surveys conducted each year covering virtually every aspect of U.S. agriculture and the Census of Agriculture conducted every five years providing state- and county-level aggregates.

What processes might have influenced what data was observed and recorded and what was not?

This data set gives the percentage of corn that falls into the various categories, but not the actual number of ears or weight of the corn. This is proabbaly because it is hard to quantify corn production in large quantities like all of the US, so they use percetages to be able to not have to count/weigh it all.

What preprocessing was done, and how did the data come to be in the form that you are using?

On the website you can download specific data as a csv file. We choose what data we think would be good to analyze for our report and then downloaded the file and cleaned the data.

If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

Yes they are aware of data collection. They have a page called quick stats which is meant to be used for data collection for the public. They want people to use the data in order encourage research and advocacy about about agriculture.

Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted on Github, in a Cornell Google Drive or Cornell Box).

This raw data can be found by going to the NASS website and then clicking on quickstat and selecting crops for field sector, corn for field commodity, the five levels of conditions states above for field data item and state for field geographic level. Then download as a spreadsheet.

https://github.com/OrenSte/ORIEOS/blob/main/99AD9A39-78E2-3A59-9604-11CF6CEFB6B8.csv (https://github.com/OrenSte/ORIEOS/blob/main/99AD9A39-78E2-3A59-9604-11CF6CEFB6B8.csv)

National Agriculture Statistics Service (Coffee, Rice, Potatoes, Tobacco - Acerage, Yield, Production, Price) https://quickstats.nass.usda.gov/ (https://quickstats.nass.usda.gov/)

What are the observations (rows) and the attributes (columns)?

These datasets present each observation as a year spanning all the way back from 1946, 1895, 1866, and 1866 for Coffee, Rice, Potatoes, and Tobacco respectivley to present day. The attributes in this dataset represent various categories: "Commodity" - the crop, "PRODUCTION in $" - production quantity in dollars, "AREA HARVESTED in ACRES" - the area used to harvest in acres, "PRODUCTION in CWT" / "PRODUCTION in LB" - production in the relevant unit of measurement, and "YIELD in LB / ACRE" - yield in pounds per acre.

Why was this dataset created?

NASS dataset are created and used to support research, education, and advocacy for the future of agriculture anyone's region.

Who funded the creation of the dataset?

The National Agricultural Statistics Service (NASS) offers Quick Stats, an on-line database containing official published aggregate estimates related to U.S. agricultural production. NASS develops these estimates from data collected through: hundreds of sample surveys conducted each year covering virtually every aspect of U.S. agriculture and the Census of Agriculture conducted every five years providing state- and county-level aggregates.

What processes might have influenced what data was observed and recorded and what was not?

While this data set provides metrics that focus on the quantity of a particular commodity, it lacks attributes highlighting the percieved quality of said commodity (whether the commodity tatses good, if it is damaged, etc). The reason for this could be that it is easier to collect quantitative rather than qualitative data as it is not easy to objectivley assess quality through taste - that would involve having many groups of experts tasting everything that is produced.

What preprocessing was done, and how did the data come to be in the form that you are using?

Quick Stats Lite allows users to filter data from a selection of observations and attributes before downloading the query as a csv file. Using this tool we filtered for data that we thought would be interesting to analyze for our report and then downloaded the resulting file as a csv. Afterward we cleaned the data via python.

If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

Yes, NASS is aware of data collection. Quick stats is a tool accessible to the public and was designed for anybody to easily access and manipulate data. The goal of this would be to encourage people to conduct agricultural research.

Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted on Github, in a Cornell Google Drive or Cornell Box).

The raw data can be found by visiting the NASS website, clicking on Quickstats Lite, and then filtering with the tool by selecting crops as the field sector, field crops as the field group, coffee, potatoes, rice and tobacco as the field commodity, acerage, yield, production and price as the field view, and national as the field geographic level. Then download the resulting query. The output should provide a csv file.

https://github.com/OrenSte/ORIEOS/blob/main/Agriculture%20data%20sets.zip
(https://github.com/OrenSte/ORIEOS/blob/main/Agriculture%20data%20sets.zip)

NASA/GISS Global Temperature https://climate.nasa.gov/vital-signs/global-temperature/ (https://climate.nasa.gov/vital-signs/global-temperature/)

“Global Surface Temperature.” NASA, NASA, 26 July 2023, climate.nasa.gov/vital-signs/global-temperature/.

What are the observations (rows) and the attributes (columns)?

The dataset presents each observation as a year spanning from 1880 to 2022. The attributes are the surface temperatures in celcius for each year, along with the temperatures for each year with lowess smoothing applied.

Why was this dataset created?

This data was created in efforts to make a climate-literate society working together to build a more sustainable future.

Who funded the creation of the dataset?

As a federal agency, the National Aeronautics and Space Administration (NASA) receives its funding from the annual federal budget passed by the United States Congress.

What processes might have influenced what data was observed and recorded and what was not?

The temperature for every yeear was recorded, so there is no missing data. However, it is an overall temperature when in reality temperature can vary throught. They most likely took an average becuase it would have been difficult to measure temperature at every location it was different.

What preprocessing was done, and how did the data come to be in the form that you are using?

First we saved the page at a .txt file as the page was a .txt file, then opened the file and imported into excel via Get Data (Power Query) feature while leveraging the text import wizard, Deleted 4 rows pertaining to redundant cells, and exported from excel as a csv file.

If people are involved, were they aware of the data collection and if so, what purpose did they expect the data to be used for?

Yes people are involved and they expected the data to be used to engage the world with accurate, accessible, and actionable information about our rapidly changing climate, from the global perspective of NASA.

Where can your raw source data be found, if applicable? Provide a link to the raw data (hosted on Github, in a Cornell Google Drive or Cornell Box).

The data is easily accessable on the NASA web page as described above.

https://github.com/OrenSte/ORIEOS/blob/main/Land-Ocean%20Temperature%20Index%20(C).csv (https://github.com/OrenSte/ORIEOS/blob/main/Land-Ocean%20Temperature%20Index%20(C).csv)

Works Cited

“Comprehensive Statistics.” Comprehensive Statistics | U.S. Drought Monitor, National Drought Mitigation Center, University of Nebraska- Lincoln, United States Department of Agriculture, National Oceanic and Atmospheric Administration, droughtmonitor.unl.edu/DmData/DataDownload/ComprehensiveStatistics.aspx. Accessed 15 Nov. 2023.

“Global Surface Temperature.” NASA, NASA, 26 July 2023, climate.nasa.gov/vital-signs/global-temperature/.

Ritchie, Hannah, et al. “CO2 Emissions.” Our World in Data, 11 May 2020, ourworldindata.org/co2-emissions.

“USDA/Nass QuickStats AD-Hoc Query Tool.” USDA/NASS QuickStats Ad-Hoc Query Tool, quickstats.nass.usda.gov/. Accessed 15 Nov. 2023.

Preregistration Statement

Experiment 1

Increasing global surface temperatures, a proxy for global warming, have negativley impacted overall corn quality in the U.S over the years.

We plan on using the Pearson correlation coefficient, a correlation measure, to determine the relationship between global surface temperatures and corn quality, and most importantly, to test for its significance. From this analysis we will first determine whether we succeed or fail in rejecting the null hypothesis. If we fail in rejecting the null hypothesis ($p < 0.05$), the two attributes do not have a statistically relevant relationship.

Otherwise, we will conclude our experiement by performing a linear regression on the two variables and analyzing its statistical summary. From this we will determine whether we can confidently say that global surface temperatures influence crop quality in either positive (contradicting our hypothesis) or negative (supporting our hypothesis) way.

Experiment 2

Co2 emissions, a proxy for global warming, negativley impact national coffee, tobacco, potato, and rice yields.

We plan on using the Pearson correlation coefficient, a correlation measure, to determine the relationship between Co2 emissions and the selected crop yields and to test for its significance. From this analysis we will first determine whether we succeed or fail in rejecting the null hypothesis. If we fail in rejecting the null hypothesis ($p < 0.05$), the two attributes do not have a statistically relevant relationship.

Otherwise, we will conclude our experiement by performing a linear regression and analyzing the statistical summaries for the 4 crops. From this we will determine whether we can confidently say that Co2 emissions influence crop yields in either positive (contradicting our hypothesis) or negative (supporting our hypothesis) way.

Experiment 3

Droughts, a proxy for global warming, tend to raise the prices for corn products in Iowa (the largest producer of corn in the U.S.).

We plan on using the Pearson correlation coefficient, a correlation measure, to determine the relationship between Drought percentage (the % area in the state affected by drought) and corn prices and to test for its significance. From this analysis we will first determine whether we succeed or fail in rejecting the null hypothesis. If we fail in rejecting the null hypothesis ($p < 0.05$), the two attributes do not have a statistically relevant relationship.

Otherwise, we will conclude our experiement by performing a linear regression on the two variables and analyzing its statistical summary. From this we will determine whether we can confidently say that droughts influence corn prices in either positive (supporting our hypothesis) or negative (contradicting our hypothesis) way.

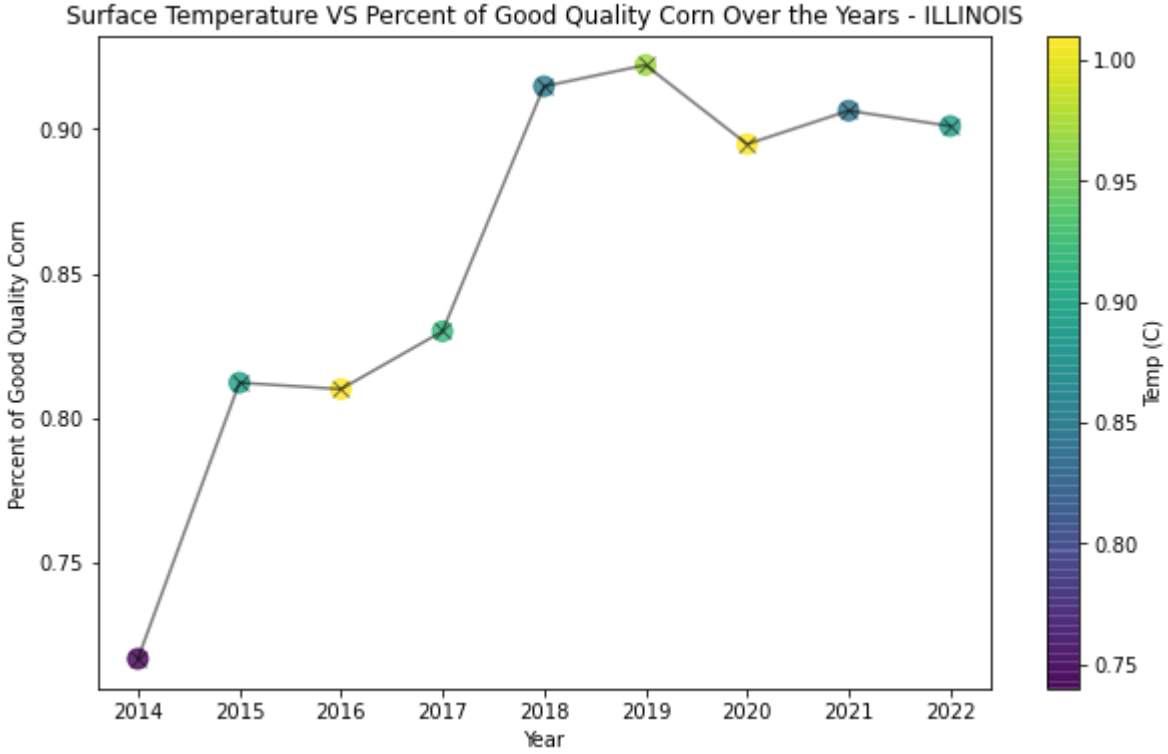
Data Analysis and Evaluation of Significance

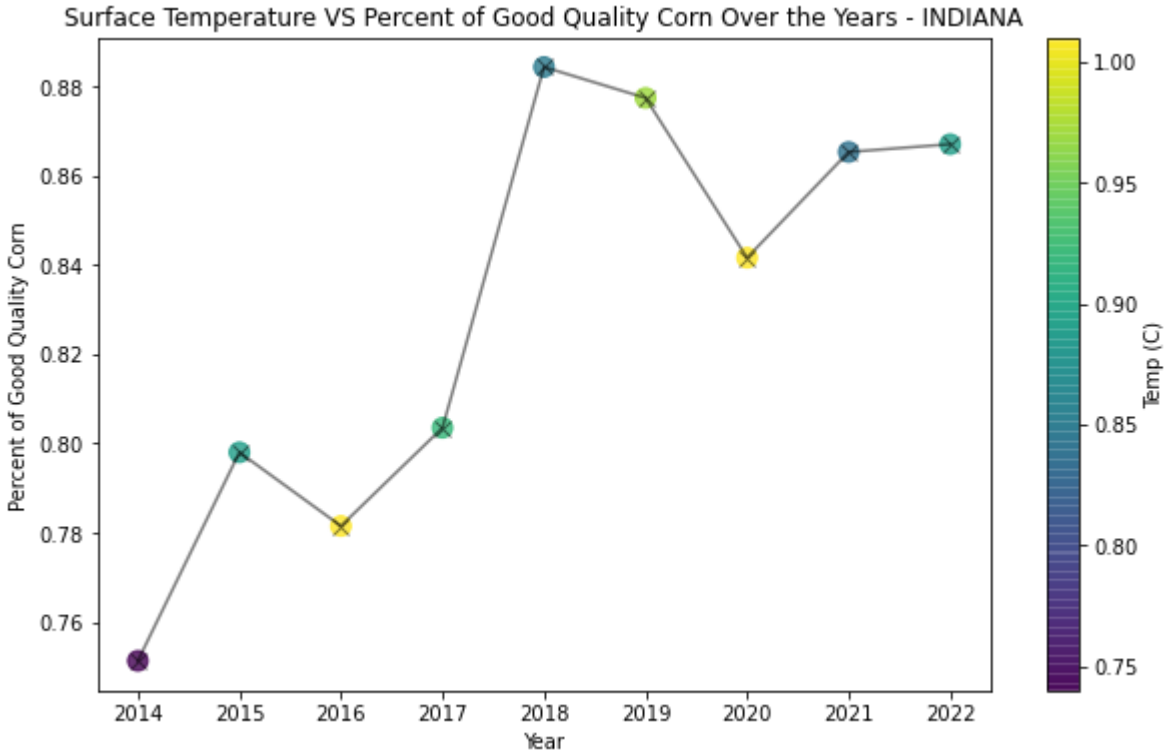
In [2]:
```
co2_df = pd.read_csv('co2_df.csv')
drought_df = pd.read_csv('drought_df.csv')
corn_df = pd.read_csv('corn_df.csv')
temp_df = pd.read_csv('temp_df.csv')
coffee_df = pd.read_csv('coffee_df.csv')
tobacco_df = pd.read_csv('tobacco_df.csv')
rice_df = pd.read_csv('rice_df.csv')
potatoes_df = pd.read_csv('potatoes_df.csv')
corn_vs_temp = pd.read_csv('corn_vs_temp.csv')
```

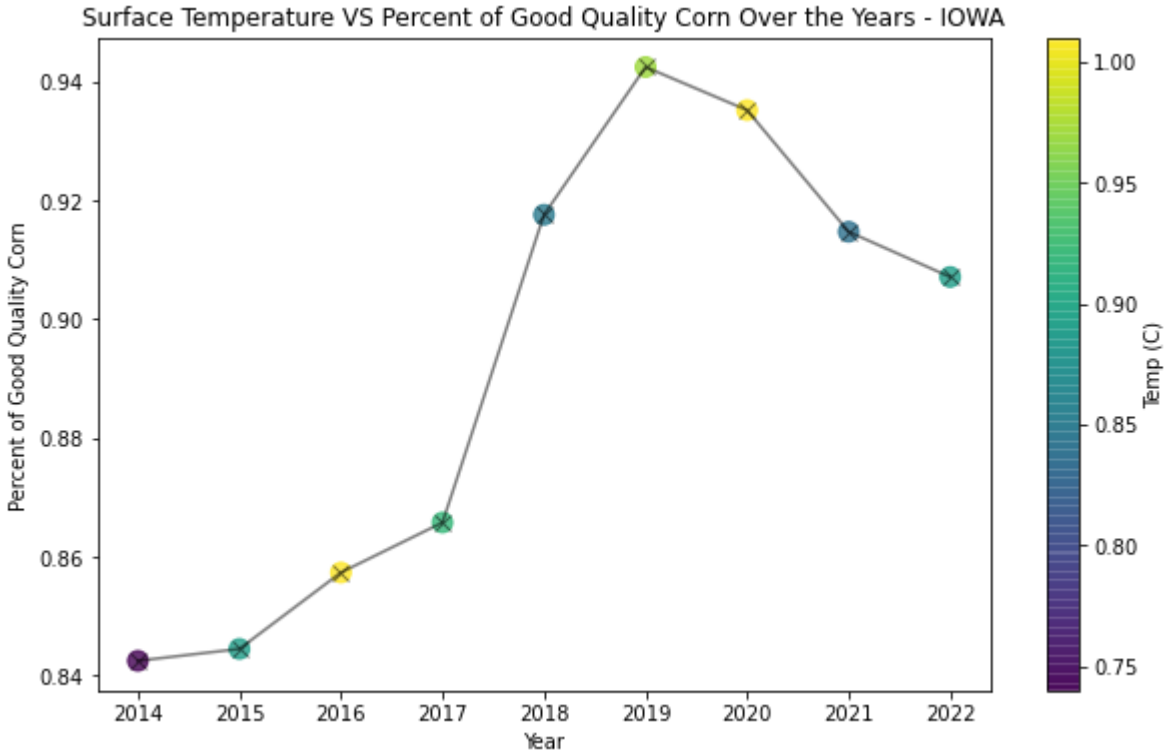
Corn Quality vs. Surface Temperature

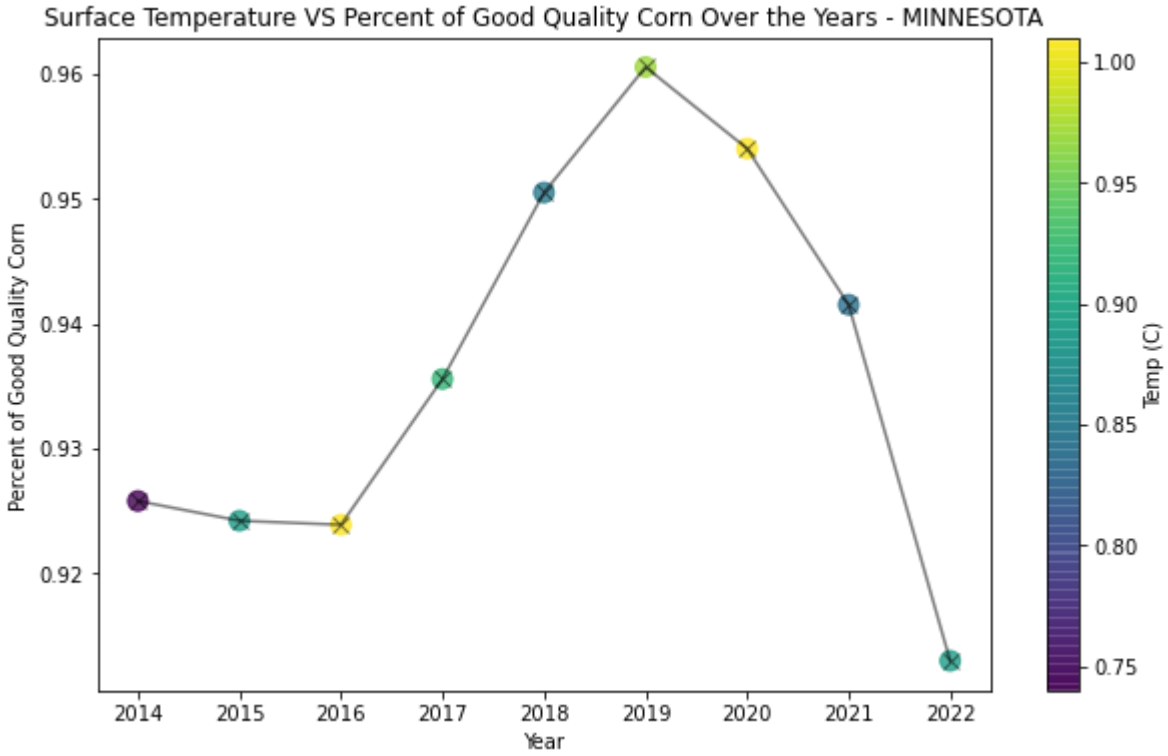
Merge the temperature df and the corn df together

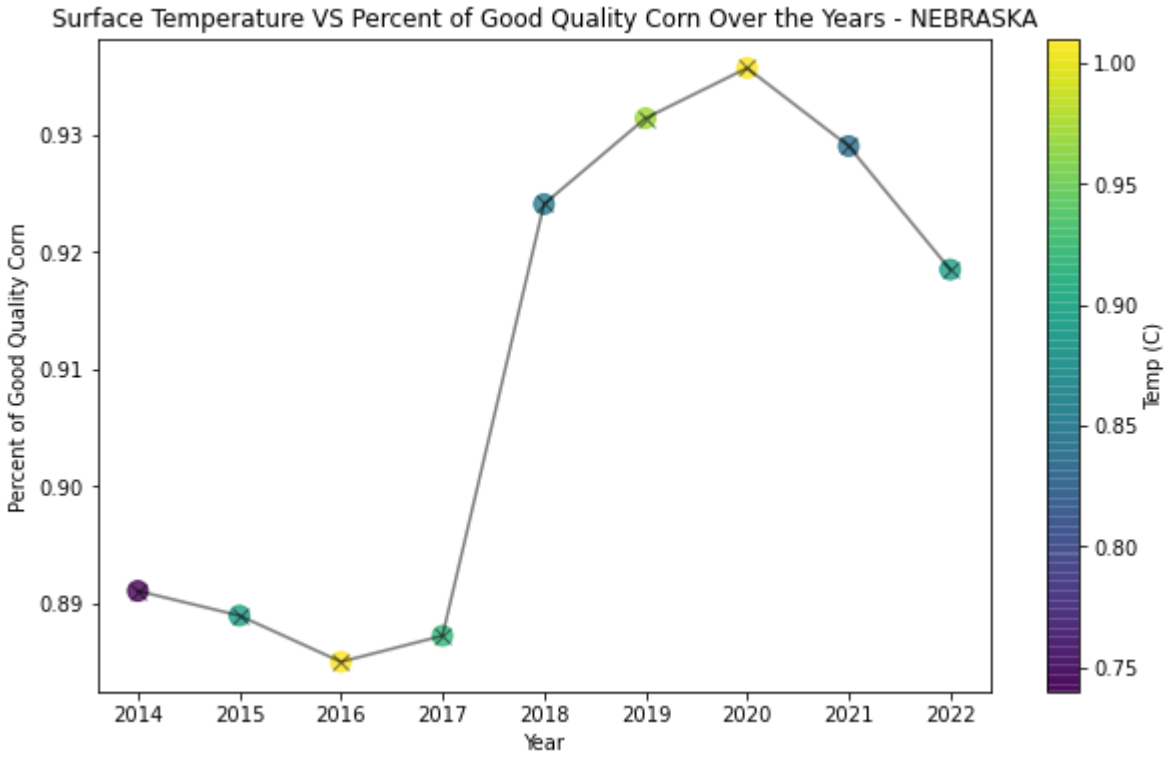
In [3]:
```
states = corn_vs_temp['State'].unique()
for state in states:
    plt.figure(figsize=(10, 6))
    state_data = corn_vs_temp[corn_vs_temp['State'] == state]
    scatter = plt.scatter(data=state_data, x='Year', y='PercentGood',
                          c='Temp', cmap='viridis', s=100, alpha=.8)
    plt.plot(state_data['Year'], state_data['PercentGood'], marker='x', linestyle='-',
             markersize=8, color='black', alpha=0.5)
    plt.xlabel('Year')
    plt.ylabel('Percent of Good Quality Corn')
    plt.title(f'Surface Temperature VS Percent of Good Quality Corn Over the Years - {state}')
    cbar = plt.colorbar(scatter, label='Temp (C)')
    plt.show()
```


Surface Temperature VS Percent of Good Quality Corn Over the Years - ILLINOIS
Percent of Good Quality Corn
0.75
0.80
0.85
0.90
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year
Temp (C)
0.75
0.80
0.85
0.90
0.95
1.00


Surface Temperature VS Percent of Good Quality Corn Over the Years - INDIANA
Percent of Good Quality Corn
0.88
0.86
0.84
0.82
0.80
0.78
0.76
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year
Temp (C)
1.00
0.95
0.90
0.85
0.80
0.75


Surface Temperature VS Percent of Good Quality Corn Over the Years - IOWA
Percent of Good Quality Corn
0.94
0.92
0.90
0.88
0.86
0.84
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year
Temp (C)
1.00
0.95
0.90
0.85
0.80
0.75


Surface Temperature VS Percent of Good Quality Corn Over the Years - MINNESOTA
Percent of Good Quality Corn
0.96
0.95
0.94
0.93
0.92
2014
2015
2016
2017
2018
2019
2020
2021
2022
Year
Temp (C)
1.00
0.95
0.90
0.85
0.80
0.75



These graphs show the relationship between the percent of good quality grown for each year compared to the surface temperature that year for each of the 5 states we chose. We chose these 5 states because they are the biggest producers of corn in the united states. This analysis is valuable to our report because the increase of surface temperature is due to climate change and we can see that it effects the production of corn, which is food that feeds and is very important to the entire country.

Now we want to perform a preliminary analysis of the correlation between surface temperature and the quality of corn yields. To do this, we decided to use a Pearson correlation coefficient as follows:

In [4]:
```
correlation_coefficient, p_value_correlation = pearsonr(corn_vs_temp['Temp'], corn_vs_temp['PercentGood'])
print(f'Pearson Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')
alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Pearson Correlation Coefficient: 0.20004780426157465
P-Value: 0.18765014132421057
Fail to reject the null hypothesis: There is no significant correlation.
```

In [5]:
```
correlation_coefficient, p_value_correlation = spearmanr(corn_vs_temp['Temp'], corn_vs_temp['PercentGood'])
print(f'Spearman Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')
alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Spearman Correlation Coefficient: 0.07585146367217757
P-Value: 0.6204426774111793
Fail to reject the null hypothesis: There is no significant correlation.
```

In [6]:
```
high_temp = corn_vs_temp[corn_vs_temp['Temp'] > 0.87]
low_temp = corn_vs_temp[corn_vs_temp['Temp'] <= 0.87]
t_statistic, p_value_ttest = ttest_ind(high_temp['PercentGood'], low_temp['PercentGood'], equal_var=False)
print(f'T-Statistic: {t_statistic}')
print(f'P-Value (t-test): {p_value_ttest}')

alpha_ttest = 0.05
if p_value_ttest < alpha_ttest:
    print('Reject the null hypothesis: There is a significant \
    difference in corn quality between high and low global temperatures.')
else:
    print('Fail to reject the null hypothesis: There is no significant \
    difference in corn quality between high and low global temperatures.')
```

```
T-Statistic: -0.0023363826803916967
P-Value (t-test): 0.9981567921479912
Fail to reject the null hypothesis: There is no significant     difference in corn quality between high and
low global temperatures.
```

As shown, there is no significant linear or non-linear correlation between global surface temperatures and the overall quality of corn products, and we fail to find any significant difference in corn quality between high and low global temperatures from the T test. This result may seem surprising at first, but upon further inspection, global surface temperatures, an aggregation of many different climates, may significantly differ from U.S. surface temperatures. Furthermore, there are many drivers of corn quality that have not been considered in this data set, particularly technological advancements, economic conditions, rainfall, global pandemics, and much more. For these reasons, the correlation that is drawn between these attributes may not paint the full picture.

A limitation of the significance metric of surface temperature is that this study only considers global surface temperatures, which might not accurately represent the local climate conditions affecting corn quality. It acknowledges that global temperatures are an aggregation of various climates, and there might be significant differences from U.S. surface temperatures.

Crop Yields vs. CO2 Emisssions

Tobacco:

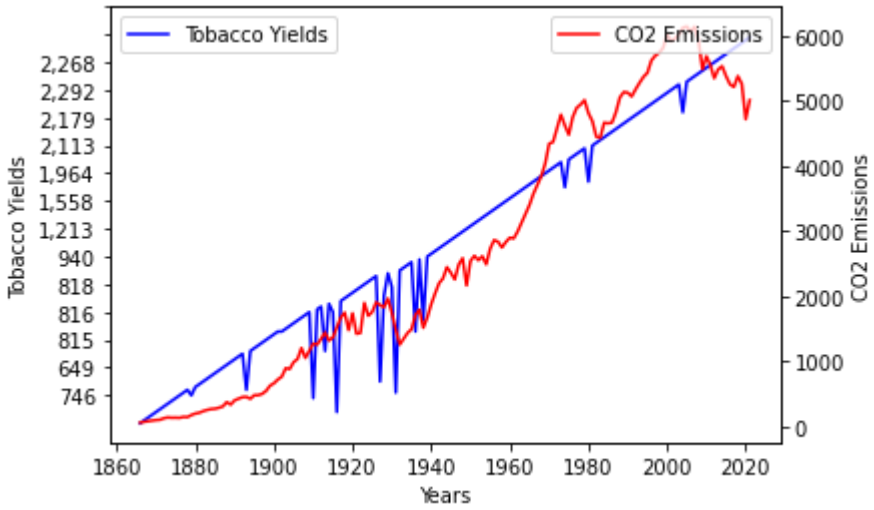
In [7]:
```
co2_df_1946 = duckdb.sql('SELECT * FROM co2_df WHERE year >= 1946 AND year <= 2014').df()
co2_df_1895 = duckdb.sql('SELECT * FROM co2_df WHERE year >= 1895').df()
co2_df_1866 = duckdb.sql('SELECT * FROM co2_df WHERE year >= 1866').df()
```

For each crop, data begins at different years. For coffee, data starts at 1946, rice data starts at 1895, and for potatoes and tobacco data starts at 1866. To account for this, we then created a new CO2 data set, starting at the various years listed above.

In [8]:
```
tobacco_df_2021 = duckdb.sql('SELECT * FROM tobacco_df WHERE YEAR <= 2021').df()
```

Since CO2 data only goes up to 2021, we accounted for this by only selecting years up to 2021

In [9]:
```
plt.plot(tobacco_df_2021['YEAR'],tobacco_df_2021['Yield'], color='blue',label='Tobacco Yields')
plt.xlabel('Years')
plt.ylabel('Tobacco Yields')
plt.legend(loc='upper left')
y_tick_locations = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150]
plt.yticks(y_tick_locations)
plt.twinx()
plt.plot(tobacco_df_2021['YEAR'], co2_df_1866['co2'], color='red',label='CO2 Emissions')
plt.ylabel('CO2 Emissions')
plt.legend(loc='upper right')
plt.show()
```



This is a graph of tobacco yields and C02 emissions over the years

Now we want to perform a preliminary analysis of the correlation between Co2 emissions and tobacco yields. To do this, we decided to use a Pearson correlation coefficient as follows:

In [10]:
```
tobacco_df_2021['Yield'] = pd.to_numeric(tobacco_df_2021['Yield'].str.replace(',', ''), errors='coerce')
```

In [11]:
```
correlation_coefficient, p_value_correlation = pearsonr(tobacco_df_2021['Yield'], co2_df_1866['co2'])
print(f'Pearson Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')

alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Pearson Correlation Coefficient: 0.9515750974807646
P-Value: 8.683629842087667e-81
Reject the null hypothesis: There is a significant correlation.
```

In [12]:
```
high_co2 = tobacco_df_2021[co2_df_1866['co2'] > 4000]
low_co2 = tobacco_df_2021[co2_df_1866['co2'] <= 4000]
t_statistic, p_value_ttest = ttest_ind(high_co2['Yield'], low_co2['Yield'], equal_var=False)
print(f'T-Statistic: {t_statistic}')
print(f'P-Value (t-test): {p_value_ttest}')

alpha_ttest = 0.05
if p_value_ttest < alpha_ttest:
    print('Reject the null hypothesis: There is a significant difference \
    in tobacco yields between high and low CO2 emissions.')
else:
    print('Fail to reject the null hypothesis: There is no significant \
    difference in tobacco yields between high and low CO2 emissions.')
```

```
T-Statistic: 27.796449033291495
P-Value (t-test): 2.4304598402943386e-59
Reject the null hypothesis: There is a significant difference     in tobacco yields between high and low CO
2 emissions.
```

As shown, there is a linear correlation between tobacco yields and the gradual increase of CO2 emissions over the years. Performing a T-Test, we can see that we can reject the null hypothesis, suggesting that there is a significant difference in tobacco yields between low and high CO2 emissions. From this we can observe that rising CO2 emissions, a proxy for global warming, has a nearly proportional relationship with tobacco yields. Does this make sense? Are CO2 emissions a positive driver for U.S. tobacco growth? Unlikely. There may be other variables at play that have had more of an impact on tobacco production.

In [13]:
```
X = sm.add_constant(co2_df_1866['co2'])
Y = tobacco_df_2021['Yield']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y)
model = sm.OLS(Y_train, X_train).fit()
print(model.summary())
Y_pred = model.predict(X_test)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Yield   R-squared:                       0.908
Model:                            OLS   Adj. R-squared:                  0.907
Method:                 Least Squares   F-statistic:                     1134.
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           2.09e-61
Time:                        16:39:21   Log-Likelihood:                -778.26
No. Observations:                 117   AIC:                             1561.
Df Residuals:                     115   BIC:                             1566.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        567.2956     29.652     19.132      0.000     508.560     626.031
co2            0.2932      0.009     33.681      0.000       0.276       0.310
==============================================================================
Omnibus:                        1.656   Durbin-Watson:                   2.099
Prob(Omnibus):                  0.437   Jarque-Bera (JB):                1.354
Skew:                           0.072   Prob(JB):                        0.508
Kurtosis:                       2.493   Cond. No.                     5.78e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.78e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

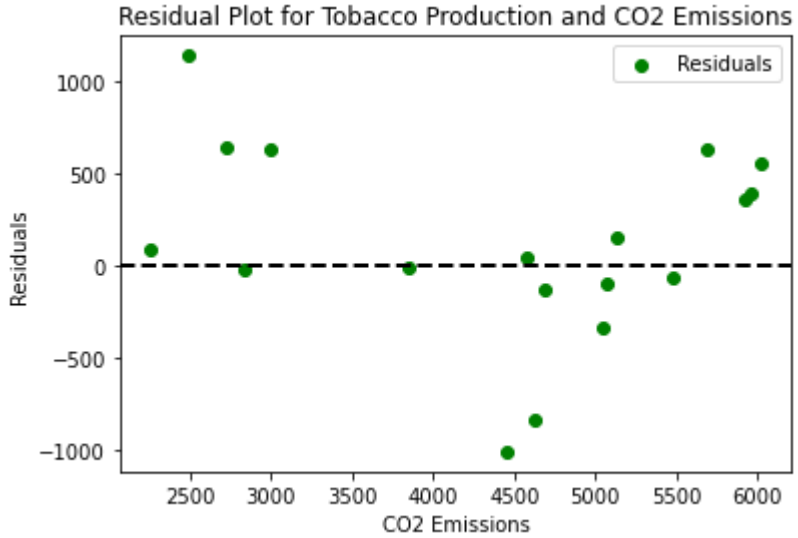
R^2 represents the proportion of the variability in the dependent variable (Yield) that is explained by the independent variable(s) (co2) in the model. In this case, about 89.8% of the variability in 'Yield' is explained by the model.

The coefficients represent the estimated impact of the independent variable 'co2' on the dependent variable 'Yield.' The positive coefficient for 'co2' (0.2916) suggests a positive relationship between 'co2' and 'Yield.'

The p-values associated with the coefficients test the null hypothesis that the corresponding coefficient is equal to zero. In this case, the coefficients both co2 and the y-intercept are statistically significant.

Overall, the analysis suggests that the linear model is statistically significant, as indicated by the low p-value for the coefficients. The R^2 value of 0.898 indicates a high level of explanatory power, and the positive coefficient for 'co2' suggests that as 'co2' increases, 'Yield' tends to increase.

In [43]:
```
residuals = Y_test - Y_pred
plt.scatter(X_test['co2'], residuals, color='green', label='Residuals')
plt.axhline(y=0, color='black', linestyle='--', linewidth=2)
plt.xlabel('CO2 Emissions')
plt.ylabel('Residuals')
plt.title('Residual Plot for Tobacco Production and CO2 Emissions')
plt.legend()
plt.show()
```



After further exaimining the graph of tobacco yields and C02 emissions over the years, we noticed that although there is a positive linear trend over time for both CO2 emissions and Tobacco yields, on the years where there are spikes in CO2 emissions, there are dips in Tobacco yields. Other factors at play may contribute to the increased yields over time, but it seems that increased CO2 levels may hinder that growth in a specific year.

Potatoes:

In [15]:
```
potatoes_df_2021 = duckdb.sql('SELECT * FROM potatoes_df WHERE YEAR <= 2021').df()
potatoes_df_2021 = potatoes_df_2021.drop_duplicates(subset = ['YEAR'])
```

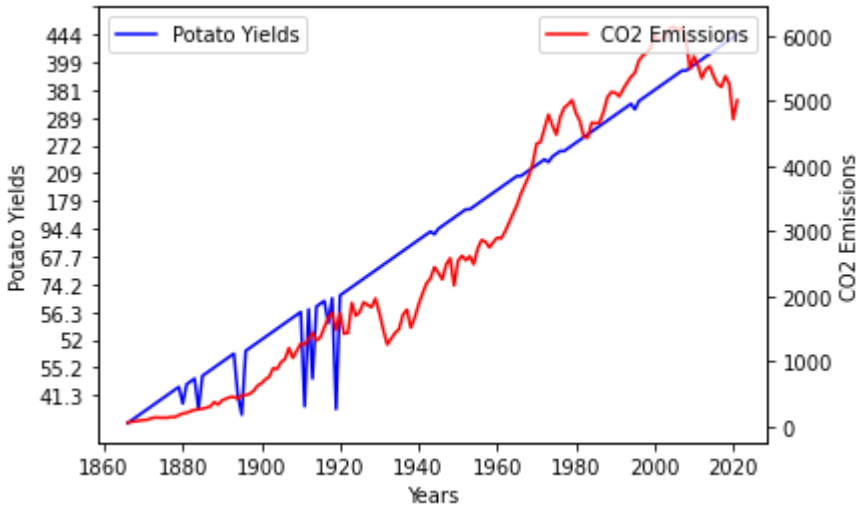
In [16]:
```
potatoes_df_2021.head()
```

Out[16]:

	YEAR	AreaHarvestedInAcres	ProductionDollars	Yield	ProductionLbs
0	1866	1,225,000	74,080,000	54.7	66,969,000
1	1867	1,289,000	90,097,000	46.4	59,798,000
2	1868	1,400,000	94,808,000	51.5	72,175,000
3	1869	1,479,000	73,574,000	58.7	86,759,000
4	1870	1,443,000	76,330,000	44.9	64,725,000

As above, but there are some duplicate years so we removed them.

In [17]:
```
plt.plot(potatoes_df_2021['YEAR'],potatoes_df_2021['Yield'], color='blue',label='Potato Yields')
plt.xlabel('Years')
plt.ylabel('Potato Yields')
plt.legend(loc='upper left')
y_tick_locations = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150]
plt.yticks(y_tick_locations)
plt.twinx()
plt.plot(potatoes_df_2021['YEAR'], co2_df_1866['co2'], color='red',label='CO2 Emissions')
plt.ylabel('CO2 Emissions')
plt.legend(loc='upper right')
plt.show()
```



This is a graph of potato yields and C02 emissions over the years

Now we want to perform a preliminary analysis of the correlation between Co2 emissions and potato yields. To do this, we decided to use a Pearson correlation coefficient as follows:

In [18]:
```
potatoes_df_2021['Yield'] = pd.to_numeric(potatoes_df_2021['Yield'].str.replace(',', ''),
                                          errors='coerce')
```

In [19]:
```
correlation_coefficient, p_value_correlation = pearsonr(potatoes_df_2021['Yield'], co2_df_1866['co2'])
print(f'Pearson Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')

alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Pearson Correlation Coefficient: 0.951169543787073
P-Value: 1.6249122036179434e-80
Reject the null hypothesis: There is a significant correlation.
```

In [20]:
```
high_co2 = potatoes_df_2021[co2_df_1866['co2'] > 4000]
low_co2 = potatoes_df_2021[co2_df_1866['co2'] <= 4000]
t_statistic, p_value_ttest = ttest_ind(high_co2['Yield'], low_co2['Yield'], equal_var=False)
print(f'T-Statistic: {t_statistic}')
print(f'P-Value (t-test): {p_value_ttest}')

alpha_ttest = 0.05
if p_value_ttest < alpha_ttest:
    print('Reject the null hypothesis: There is a significant \
    difference in potato yields between high and low CO2 emissions.')
else:
    print('Fail to reject the null hypothesis: There is no significant \
    difference in potato yields between high and low CO2 emissions.')
```

```
T-Statistic: 23.23378265277846
P-Value (t-test): 1.6505120198789342e-37
Reject the null hypothesis: There is a significant     difference in potato yields between high and low CO2
emissions.
```

There is a significant linear correlation between potato yields and the gradual increase of CO2 emissions over the years. Furthermore, performing a T-test, we can reject the null hypothesis to suggest that there is a significant difference in potato yields between low and high CO2 emissions. Rising CO2 emissions, a proxy for global warming, has a nearly proportional relationship with potato yields. Does this make sense? Are CO2 emissions a positive driver for U.S. potato growth? Once again, correlation does not always indicate causation as there could be other variables at play that have had more of an impact on long-term potato production.

In [21]:
```
X = sm.add_constant(co2_df_1866['co2'])
Y = potatoes_df_2021['Yield']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y)
model = sm.OLS(Y_train, X_train).fit()
print(model.summary())
Y_pred = model.predict(X_test)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Yield   R-squared:                       0.904
Model:                            OLS   Adj. R-squared:                  0.903
Method:                 Least Squares   F-statistic:                     1079.
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           2.87e-60
Time:                        16:39:22   Log-Likelihood:                -602.42
No. Observations:                 117   AIC:                             1209.
Df Residuals:                     115   BIC:                             1214.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.6036      6.572      0.548      0.585      -9.414      16.621
co2            0.0625      0.002     32.846      0.000       0.059       0.066
==============================================================================
Omnibus:                       26.875   Durbin-Watson:                   2.091
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.027
Skew:                           1.093   Prob(JB):                     1.23e-09
Kurtosis:                       4.907   Cond. No.                     5.84e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.84e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R^2 represents the proportion of the variability in the dependent variable (Yield) that is explained by the independent variable(s) (co2) in the model. In this case, about 91.0% of the variability in 'Yield' is explained by the model.

The coefficients represent the estimated impact of the independent variable 'co2' on the dependent variable 'Yield.' The positive coefficient for 'co2' (0.0643) suggests a positive relationship between 'co2' and 'Yield.'

The p-values associated with the coefficients test the null hypothesis that the corresponding coefficient is equal to zero. In this case, the coefficient for co2, not the intercept, is statistically significant.

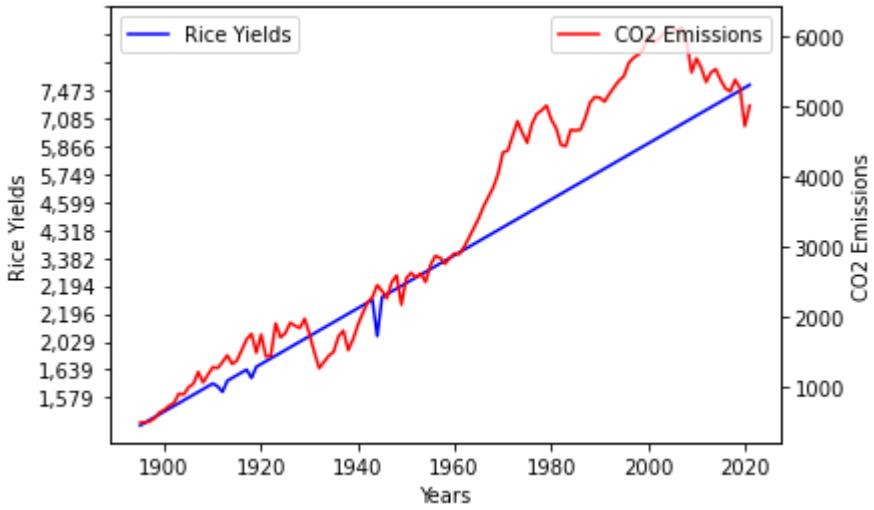
Overall, the analysis suggests that the linear model is statistically significant, as indicated by the low p-value for the co2 coefficient. The R^2 value of 0.910 indicates a high level of explanatory power, and the positive coefficient for 'co2' suggests that as 'co2' increases, 'Yield' tends to increase.

Rice:

In [22]:
```
rice_df_2021 = duckdb.sql('SELECT * FROM rice_df WHERE YEAR <= 2021').df()
rice_df_2021 = rice_df_2021.drop_duplicates(subset = ['YEAR'])
```

As above, but there are some duplicate years so we removed them.

In [23]:
```
plt.plot(rice_df_2021['YEAR'],rice_df_2021['Yield'], color='blue',label='Rice Yields')
plt.xlabel('Years')
plt.ylabel('Rice Yields')
plt.legend(loc='upper left')
y_tick_locations = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150]
plt.yticks(y_tick_locations)
plt.twinx()
plt.plot(rice_df_2021['YEAR'], co2_df_1895['co2'], color='red',label='CO2 Emissions')
plt.ylabel('CO2 Emissions')
plt.legend(loc='upper right')
plt.show()
```



This is a graph of rice yields and C02 emissions over the years

Now we want to perform a preliminary analysis of the correlation between Co2 emissions and rice yields. To do this, we decided to use a Pearson correlation coefficient as follows:

In [24]:
```
rice_df_2021['Yield'] = pd.to_numeric(rice_df_2021['Yield'].str.replace(',', ''), errors='coerce')
```

In [25]:
```
correlation_coefficient, p_value_correlation = pearsonr(co2_df_1895['co2'], rice_df_2021['Yield'])
print(f'Pearson Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')

alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Pearson Correlation Coefficient: 0.9538638167070546
P-Value: 3.599888751763009e-67
Reject the null hypothesis: There is a significant correlation.
```

In [26]:
```
high_co2 = rice_df_2021[co2_df_1895['co2'] > 4000]
low_co2 = rice_df_2021[co2_df_1895['co2'] <= 4000]
t_statistic, p_value_ttest = ttest_ind(high_co2['Yield'], low_co2['Yield'], equal_var=False)
print(f'T-Statistic: {t_statistic}')
print(f'P-Value (t-test): {p_value_ttest}')

alpha_ttest = 0.05
if p_value_ttest < alpha_ttest:
    print('Reject the null hypothesis: There is a significant \
    difference in rice yields between high and low CO2 emissions.')
else:
    print('Fail to reject the null hypothesis: There is no significant \
    difference in rice yields between high and low CO2 emissions.')
```

```
T-Statistic: 20.40593940551996
P-Value (t-test): 8.068804000159146e-36
Reject the null hypothesis: There is a significant     difference in rice yields between high and low CO2 e
missions.
```

As shown there is a significant linear correlation between rice yields and the gradual increase of CO2 emissions over the years. Performing a T-test, we can reject the null hypothesis to suggest that there is a significant difference in rice yield between low and high CO2 emissions.

Overall, it appears that rising CO2 emissions, a proxy for global warming, has a nearly 1-1 relationship with rice yields. Are CO2 emissions therefore a positive driver for U.S. rice growth? Context may suggest otherwise as it tough to argue that global warming has had a positive effect on agriculture. Furthermore, correlation does not always indicate causation; There could be other representations of global warming that affect yield more directly.

In [27]:
```
X = sm.add_constant(co2_df_1895['co2'])
Y = rice_df_2021['Yield']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2)
model = sm.OLS(Y_train, X_train).fit()
print(model.summary())
Y_pred = model.predict(X_test)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Yield   R-squared:                       0.902
Model:                            OLS   Adj. R-squared:                  0.902
Method:                 Least Squares   F-statistic:                     916.4
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           7.60e-52
Time:                        16:39:22   Log-Likelihood:                -795.25
No. Observations:                 101   AIC:                             1594.
Df Residuals:                      99   BIC:                             1600.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        172.2502    129.662      1.328      0.187     -85.026     429.527
co2            1.0954      0.036     30.272      0.000       1.024       1.167
==============================================================================
Omnibus:                       25.659   Durbin-Watson:                   1.734
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               39.742
Skew:                           1.136   Prob(JB):                     2.34e-09
Kurtosis:                       5.068   Cond. No.                     7.27e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.27e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R^2 represents the proportion of the variability in the dependent variable (Yield) that is explained by the independent variable(s) (co2) in the model. In this case, about 89.9% of the variability in 'Yield' is explained by the model.

The coefficients represent the estimated impact of the independent variable 'co2' on the dependent variable 'Yield.' The positive coefficient for 'co2' (1.0956) suggests a positive relationship between 'co2' and 'Yield.'

The p-values associated with the coefficients test the null hypothesis that the corresponding coefficient is equal to zero. In this case, the coefficient for co2, not the intercept, is statistically significant.

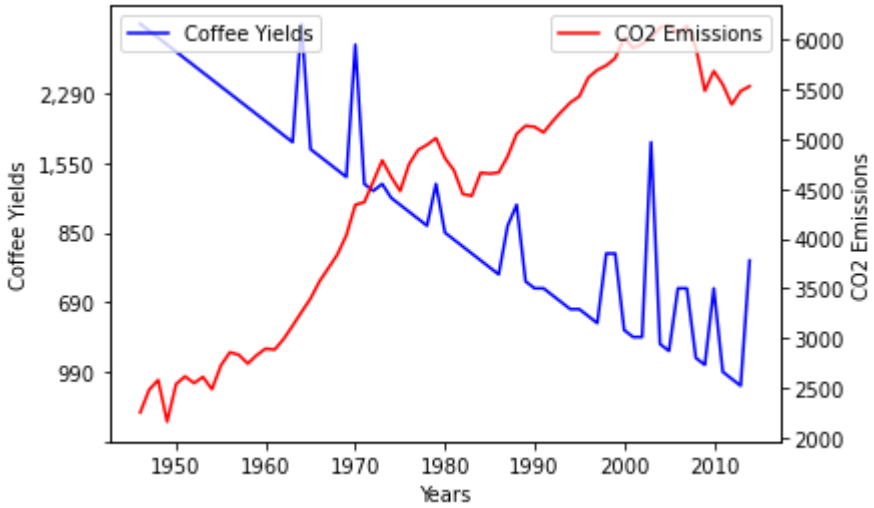
Overall, the analysis suggests that the linear model is statistically significant, as indicated by the low p-value for the co2 coefficient. The R^2 value of 0.899 indicates a high level of explanatory power, and the positive coefficient for 'co2' suggests that as 'co2' increases, 'Yield' tends to increase.

Coffee:

In [28]:
```
coffee_df_2021 = duckdb.sql('SELECT * FROM coffee_df WHERE YEAR <= 2014').df()
coffee_df_2021 = coffee_df_2021.drop_duplicates(subset = ['YEAR'])
```

As above, but coffee yield data stops at 2014, so accounted for that

In [29]:
```
plt.plot(coffee_df_2021['YEAR'],coffee_df_2021['Yield'], color='blue',label='Coffee Yields')
plt.xlabel('Years')
plt.ylabel('Coffee Yields')
plt.legend(loc='upper left')
y_tick_locations = [10, 20, 30, 40, 50, 60]
plt.yticks(y_tick_locations)
plt.gca().invert_yaxis()
plt.twinx()
plt.plot(coffee_df_2021['YEAR'], co2_df_1946['co2'], color='red',label='CO2 Emissions')
plt.ylabel('CO2 Emissions')
plt.legend(loc='upper right')
plt.show()
```



This is a graph of rice yields and C02 emissions over the years

Now we want to perform a preliminary analysis of the correlation between Co2 emissions and coffee yields. To do this, we decided to use a Pearson correlation coefficient as follows:

In [30]:
```
coffee_df_2021['Yield'] = pd.to_numeric(coffee_df_2021['Yield'].str.replace(',', ''),
                                        errors='coerce')
```

In [31]:
```
correlation_coefficient, p_value_correlation = pearsonr(co2_df_1946['co2'],
                                                        coffee_df_2021['Yield'])
print(f'Pearson Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')

alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Pearson Correlation Coefficient: -0.764457138447713
P-Value: 2.115313933377246e-14
Reject the null hypothesis: There is a significant correlation.
```

In [32]:
```
high_co2 = coffee_df_2021[co2_df_1946['co2'] > 4000]
low_co2 = coffee_df_2021[co2_df_1946['co2'] <= 4000]
t_statistic, p_value_ttest = ttest_ind(high_co2['Yield'], low_co2['Yield'], equal_var=False)
print(f'T-Statistic: {t_statistic}')
print(f'P-Value (t-test): {p_value_ttest}')

alpha_ttest = 0.05
if p_value_ttest < alpha_ttest:
    print('Reject the null hypothesis: There is a significant \
    difference in coffee yields between high and low CO2 emissions.')
else:
    print('Fail to reject the null hypothesis: There is no significant \
    difference in coffee yields between high and low CO2 emissions.')
```

```
T-Statistic: -9.306819978297076
P-Value (t-test): 8.164957416835339e-10
Reject the null hypothesis: There is a significant     difference in coffee yields between high and low CO2
emissions.
```

As shown, there is a significant linear correlation between coffee bean yield and the gradual increase of CO2 emissions over the years. Furthermore, performing a T-test, we can reject the null hypothesis. This suggests a significant difference in coffee bean yield between low and high CO2 emissions. More interestingly, rising CO2 emissions, a proxy for global warming, has an strong inverse relationship with coffee bean yield. Because of this, coffee may be sensitive to global warming in ways that are much more detrimental than other agricultural goods.

Now performing a linear regression:

In [33]:
```
X = sm.add_constant(co2_df_1946['co2'])
Y = coffee_df_2021['Yield']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y)
model = sm.OLS(Y_train, X_train).fit()
print(model.summary())
Y_pred = model.predict(X_test)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Yield   R-squared:                       0.586
Model:                            OLS   Adj. R-squared:                  0.577
Method:                 Least Squares   F-statistic:                     69.25
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           6.18e-11
Time:                        16:39:22   Log-Likelihood:                -382.76
No. Observations:                  51   AIC:                             769.5
Df Residuals:                      49   BIC:                             773.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3428.4510    236.136     14.519      0.000    2953.918    3902.984
co2           -0.4275      0.051     -8.322      0.000      -0.531      -0.324
==============================================================================
Omnibus:                        5.256   Durbin-Watson:                   1.787
Prob(Omnibus):                  0.072   Jarque-Bera (JB):                4.158
Skew:                           0.568   Prob(JB):                        0.125
Kurtosis:                       3.815   Cond. No.                     1.73e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.73e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

R^2 represents the proportion of the variability in the dependent variable (Yield) that is explained by the independent variable(s) (co2) in the model. In this case, about 61.0% of the variability in 'Yield' is explained by the model.

The coefficients represent the estimated impact of the independent variable 'co2' on the dependent variable 'Yield.' The negative coefficient for 'co2' (-0.4391) suggests a negative relationship between 'co2' and 'Yield.'

The p-values associated with the coefficients test the null hypothesis that the corresponding coefficient is equal to zero. In this case, both coefficients are statistically significant.

Overall, the analysis suggests that the linear model is statistically significant, as indicated by the low p-values for the F-statistic and individual coefficients. The R^2 value of 0.610 indicates a moderate level of explanatory power, and the negative coefficient for 'co2' suggests that as 'co2' increases, 'Yield' tends to decrease.

Summary of CO2 Emissions and Crop Produciton

As we can see, yields in tobacco, potatoes, and rice products have trended positively as global CO2 emissions has increased in the U.S. Coffee however, yields have decreased throughout the years. This is because coffee production has decreased in the U.S. as more favorable climates in Latin America have taken over global coffee production.

A limitation to the significance metric of CO2 emissions is that the growth of emissions alone do not provide a complete picture of the greenhouse gas dynamics in the atmosphere. It doesn't differentiate between emissions that are quickly absorbed or sequestered and those that remain in the atmosphere for an extended period, contributing to the greenhouse effect. This means even with small growth in CO2, there could be a lot left in the atmosphere from previous years emissions.

Drought vs. Price of Corn in Iowa

In [34]:
```
corn_iowa_df = pd.read_csv('CORN-AcreageYieldProductionandPrice-2023-11-13.csv')
corn_iowa_df= corn_iowa_df[(corn_iowa_df['REFERENCE PERIOD'] == 'MARKETING YEAR') &
                           (corn_iowa_df['YEAR']>= 2000)]
corn_iowa_df.rename(columns={'PRICE RECEIVED in $ / BU': 'PRICE'}, inplace=True)
```

We had to only find when the reference period is a marketing year because those values give us the price of the corn, whcih we then renamed the column to something simpler. We also only want from 2000 onwards because we do not have droguht data for anytime before 2000

In [35]:
```
drought_df_analysis = drought_df.groupby(['Year', 'StateAbbreviation'])['PercentAreaNotInDrought'].mean()
drought_df_analysis = pd.DataFrame(drought_df_analysis)
drought_df_analysis['PercentAreaInDrought'] = 100 - drought_df_analysis['PercentAreaNotInDrought']
drought_df_analysis.reset_index(inplace=True)
```

We wanted to group the Years and States and to take the mean of how much in drought each state was by each year. Then we added a new column, PercentAreaInDrought to measure how much area is in drought.

In [36]:
```
drought_df_analysis_iowa = drought_df_analysis[
    (drought_df_analysis['StateAbbreviation'] =='IA')&(drought_df_analysis['Year']<2023)]
drought_df_analysis_iowa=duckdb.sql("SELECT * FROM drought_df_analysis_iowa \
ORDER BY Year DESC").df()
```

We are only focused on Iowa so we selected just Iowa, and then we had to limit onlt to 2022 because we do not have corn data for 2023. We also wanted the dataframes to be ordered the same way

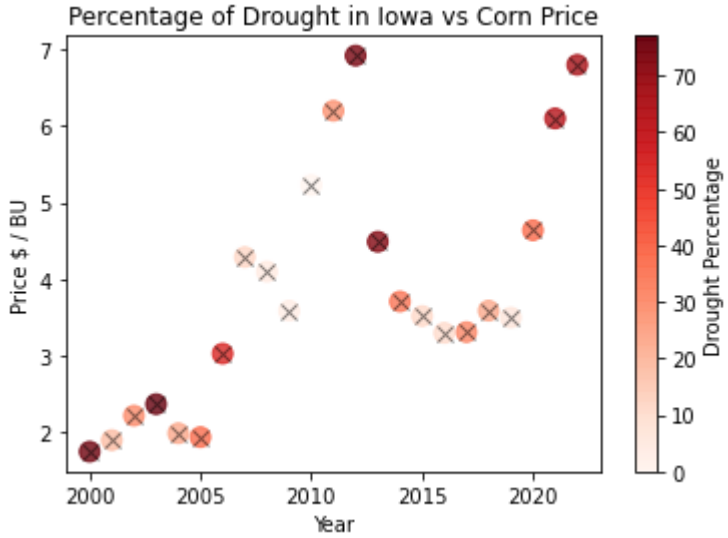
In [37]:
```
iowa_analysis_df = duckdb.sql('SELECT * FROM corn_iowa_df LEFT JOIN \
drought_df_analysis_iowa ON corn_iowa_df.YEAR = drought_df_analysis_iowa.Year').df()
iowa_analysis_df = duckdb.sql("SELECT YEAR, PRICE, \
PercentAreaInDrought FROM iowa_analysis_df").df()
```

We then join the data sets together on year, then we only selected the columns that we are interested in.

In [38]:
```
iowa_analysis_df['PRICE'] = iowa_analysis_df['PRICE'].astype(float)
```

The PRICE columns for some reason as an object so we had to convert it to a float

In [39]:
```
plt.xlabel('Year')
plt.ylabel('Price $ / BU')
scatter = plt.scatter(data=iowa_analysis_df, x='YEAR', y='PRICE',
                      c='PercentAreaInDrought', cmap='Reds', s=100, alpha=.8)
plt.plot(iowa_analysis_df['YEAR'], iowa_analysis_df['PRICE'], marker='x',
         linestyle=' ', markersize=8, color='black', alpha=0.5)
plt.title('Percentage of Drought in Iowa vs Corn Price')
cbar = plt.colorbar(scatter, label='Drought Percentage')
plt.show()
```



In [40]:
```
correlation_coefficient, p_value_correlation = pearsonr(
    iowa_analysis_df['PercentAreaInDrought'], iowa_analysis_df['PRICE'])
print(f'Pearson Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')

alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Pearson Correlation Coefficient: 0.17676671391460583
P-Value: 0.4197522051292436
Fail to reject the null hypothesis: There is no significant correlation.
```

In [41]:
```
correlation_coefficient, p_value_correlation = spearmanr(
    iowa_analysis_df['PercentAreaInDrought'], iowa_analysis_df['PRICE'])
print(f'Spearman Correlation Coefficient: {correlation_coefficient}')
print(f'P-Value: {p_value_correlation}')

alpha_correlation = 0.05
if p_value_correlation < alpha_correlation:
    print('Reject the null hypothesis: There is a significant correlation.')
else:
    print('Fail to reject the null hypothesis: There is no significant correlation.')
```

```
Spearman Correlation Coefficient: 0.020261923030226423
P-Value: 0.926886468551548
Fail to reject the null hypothesis: There is no significant correlation.
```

In [42]:
```
drought_years = iowa_analysis_df[iowa_analysis_df['PercentAreaInDrought'] > 40]
non_drought_years = iowa_analysis_df[iowa_analysis_df['PercentAreaInDrought'] <= 40]
t_statistic, p_value_ttest = ttest_ind(drought_years['PRICE'], non_drought_years['PRICE'], equal_var=False)
print(f'T-Statistic: {t_statistic}')
print(f'P-Value (t-test): {p_value_ttest}')

alpha_ttest = 0.05
if p_value_ttest < alpha_ttest:
    print('Reject the null hypothesis: There is a significant\
    difference in corn prices between drought and non-drought years.')
else:
    print('Fail to reject the null hypothesis: There is no significant \
    difference in corn prices between drought and non-drought years.')
```

```
T-Statistic: 1.0697401740645565
P-Value (t-test): 0.3172358774883401
Fail to reject the null hypothesis: There is no significant     difference in corn prices between drought a
nd non-drought years.
```

As shown, there is no significant linear or non-linear correlation between drought and the overall quality of corn products. Furthermore, we cannot reject the null hypothesis, suggesting no significant difference in corn prices between drought and non-drought years. This result may seem surprising at first. Dought, a proxy for global warming, may cause scarcity, driving up prices. Nevertheless, other economic factors such as inflation and other agricultural policies may outclass local droughts in terms of their impact on price. For these reasons, the correlation that is drawn between these attributes may not paint the full picture.

A limitation to the significance metric of drought area percentage is that even with a drought, crop production could be fueled by man made watering and not just from rain. So droughts might actually have a significant correlation to corn production, but we cannot see this in the data if humans are watering the crops in place of rain. This means other factors such as price of growing crops could show us this extra expense of buying more water, but we would need to do further research to test this theory. We also must note that we are only doing analysis on Iowa, one of the largest producers of corn in the United States. This could be a limitation because we therefore have less data about variation in dorught levels as different states see more or less droughts.

Interpretation and Conclusion

Firstly, we studied if corn quality was impacted over time due to rising global temperatures. Our analysis showed that there is no significant linear or non-linear correlation between the global surface temperature and the quality of corn produced. The Pearson correlation produced a p-value of 0.128, which exceeds our threshold of 0.05, and therefore we fail to reject the null hypothesis. The Spearman correlation produced a p-value of 0.620, which again exceeds our threshold so we fail to reject the null hypothesis. We ran a T-test to see if there was any difference in the quality of corn between higher and lower temperatures, which then produced a p-value of 0.998 which exceeded our threshold, thus there is no significant difference.

For our next analysis, we examined whether there was a relationship between CO2 levels and yields of tobacco, potataoes, rice, and coffee. We found that each crop did have a significant linear correlation between rising CO2 levels. Each analysis produced p-values at or near 0, satisfying our threshold. We ran a linear regression on each analysis to deepen our understanding of the relationship between crop yields and CO2 levels. For tobacco, our r^2 was 0.909, meaning 90.9% of the yield can be explained by the model. CO2 has a positive coefficient of 0.292, suggesting a positive correlation. The p-values associated with the coefficient test and the F-statistic is equal to 0, implying that the model is statistically significant. We then performed a T-test to assess whether there was a significant difference in tobacco yields for higher and lower CO2 emmissions. This prodcued a p-value near 0, so we found that there is a significant difference of tobacco yields between higher and lower CO2 emmissions. We saw a similar result for potato yields. We had an r^2 of 0.902, as well as a positive coefficient for CO2 of 0.064. We had low p-values for the coefficient and the F-statistic, shwoing that our model was statistically significant. Then we performed a T-test again to assess whether there was a significant difference in potato yields for higher and lower CO2 emmissions. This prodcued a p-value near 0, so we found that there is a significant difference of potato yields between higher and lower CO2 emmissions. We got similar results for rice as well. The linear regression produced an r^2 of 0.912, so 91.2% of the variability in yield can be explained by our model. CO2 had a positive coefficient of 1.080, and when we tested the coefficient and the F-statistic, the p-vaules suggested that our model was indeed statistically significant. We performed another T-test to assess whether there was a significant difference in rice yields for higher and lower CO2 emmissions. This prodcued a p-value near 0, so there is a significant difference of rice yields between higher and lower CO2 emmissions. Finally, our linear regression model of coffee produced an r^2 of 0.648, which implies a moderate level of explanatory power. There was a negative coefficient for CO2, suggesting a negative correlation between CO2 levels and coffee yields. The p-values for for the coefficient test and the F-statistic implies that our linear model is statistically significant. We performed another T-test to assess whether there is a significant difference in coffee yields for higher and lower CO2 emmissions. This prodcued a p-value near 0, so we found that there is a significant difference of coffee yields between higher and lower CO2 emmissions. Overall, we found that rising CO2 levels are statistically correlated with crop yields.

For our last analysis, we tested to see if there is any correlation between droughts and corn prices in the state of Iowa. We chose Iowa because it is the largest producer of corn in the United States. Our Pearson analysis produced a p-value of 0.420, which is larger than our threshold, and our Spearman analysis procued a p-value of 0.927 which also exceeds our threshold and therefore there is no statistical significance related to linear or non-linear correlation. We performed a T-test to see if there is a significant difference in corn prices in drought years vs. non-drought years. The T-test produced a p-value of 0.317 which exceeds our threshold so we found that there is no significant differences in corn prices in drought years vs non-drought years.

Limitations

Our World In Data

There is a lot of missing data in this dataset, so we would have to work with NaNs which would be a downside. However we believe that most of this missing data corrosponds to countries outside the US, which could be a plus as the scope of our project only focuses on the US. Additionally, the units of measurement for the CO2 column in the data set is unclear. This may affect the sensitivity of our analysis.

US Drought Monitor

Not all of the dates from 1/1/2000 to 10/13/2023 are listed for each state. We can fix this by working with the valid start and end dates of the periods to make much more consistent intervals.

National Agriculture Statistics Service (Corn Quality)

This data only contains percentages for each condition of corn. This is a limitation to our data because we are looking at how climate change effects corn production in quantity and quality. For example, let's say that 50% of corn for one year is of poor quality, and the next year only 25% of it is poor. If this is the case, data would indicate that the second year has been more successful, whereas in reality, the first year could have produced significantly more corn of good quality. To address this, we should merge this data set with another one about quantity of corn that is harevsted in each period.

The dataset in question can be found on the NASS website by clicking on quickstats, selecting crops for field sector, corn for field commodity, acres harvested for field data item, and then state for field geographic level. We downloaded this as a spreadsheet. https://github.com/OrenSte/ORIEOS/blob/main/D066FF92-5109-3381-8D5B-DE552C048617.csv (https://github.com/OrenSte/ORIEOS/blob/main/D066FF92-5109-3381-8D5B-DE552C048617.csv)

National Agriculture Statistics Service (Other Crops)

There are many missing values in the datasets that we have pulled from this website. Some of these missing values are in rows that we would like to analyze for our report. This is a problem because we don't want to have any NaN's in our analysis. We could solve this issue by analyzing smaller time spans or taking averages for time periods.

NASA/GISS Global Temperature

This data contains Earth’s global average surface temperature for every year since recordkeeping began in 1880. In our case, we plan on using the data to determine whether higher temperatures in the United States negatively affect corn quality. Therin lies the limitation: we are using global average surface temperatures as a proxy for U.S. surface temperatures, a metric which may not capture crop quality.