Citadel Datathon 2024

Life, Liberty, and the Pursuit of Healthy Living: Unraveling the Hidden Costs of Ultra Processed Foods in the United States

Daniel Kim[a,1], **Mark Zhu**[a,2], **Krishna Karthikeya Chemudupati**[b,3] **and Advit Ranawade**[b,4]

[a] *Stanford University*
[b] *University of Pennsylvania*

Abstract—This report examines the footprint of processed foods in the United States through the lens of environmental factors in an effort to illustrate their complex relationships with societal, economic, and geographic variables. krishna one sentence summary of project. We track and analyze correlations between processed-food-related stocks with carbon dioxide emissions and other air pollutants by state. We then specifically demonstrate how the relationships between environmental and societal factors with the consumption of processed foods can converge on the similar results; in our case we examine how both food desert severity and carbon dioxide emissions are lower in states with high dairy and grain consumption. Building off this analysis, we then examined the individualized supply chains of minimally processed, processed, and ultra processed foods (UPFs) and their diverging impacts on climate and air quality. In order to do so, we interpreted the supply chain as a directed graph composed of nodes (sources of raw materials) and edges (transportation networks).

Contents

1. Introduction

One of the largest challenges of the 21st century is the environmental crisis. Amidst rising global temperatures and the steady output of greenhouse gases, we need to find sustainable methods of reducing our carbon footprint without completely compromising our existing economy. But for many years, environmentally-friendly policies haven't been prioritized as much as they should.

At the same time, the US has a unique problem of its own regarding processed foods. The United States has much higher obesity and overweight rates than comparable countries, and this is largely attributed to the prevalence of low-cost, low-nutrition, processed foods. The dominance of processed foods over fresh produce and other costlier yet healthier alternatives has been examined in many lenses. However, the environmental impact of processed foods has not been conclusively analyzed, despite estimates that the food industry contributes to over 25% of humanity's greenhouse emissions (Paraskevi, 2020).

While common perceptions suggest that processing foods is more energy – and thus fossil fuel – intensive, there are have studies that suggest otherwise and argue that minimally processed are even more taxing for the planet (Brown, 2021). In short, the complex relationship between processed foods and environmental factors in the US must still be thoroughly examined, especially when they coincide with socioeconomic characteristics.

2. CO2 Emissions and Fast-Food Stocks

2.1. Introduction

The relationship between environmental factors and corporate performance has become an increasingly important area of study as businesses strive to balance profitability with sustainability. The fast food industry is known for its significant environmental footprint, and is particularly under scrutiny (Hernandez, 2023). This section aims to analyze the correlation between fast food stock prices and national CO2 emissions. By examining this relationship, we can gain insights into how environmental impacts influence the financial performance of major fast food companies. This analysis also helps in understanding whether corporate sustainability initiatives are effectively contributing to both environmental preservation and economic success.

2.2. Data Collection and Preparation

In our analysis, we used the stocks and ETFs data provided by the datathon to analyze fast food stock prices. The provided data lists in great detail the open, high, low, and close prices for stocks across a variety of sectors. Crucially, they included data on six major fast food companies: Starbucks (SBUX), Chipotle (CMG), Wendy's (WEN), Yum! Brands (YUM), McDonald's (MCD), and Domino's Pizza (DPZ). Furthermore, we used the external dataset listing the national CO2 emissions data (EIA Data). The included data spans from 1980 to 2023, recording annual CO2 emissions in millions of metric tons.

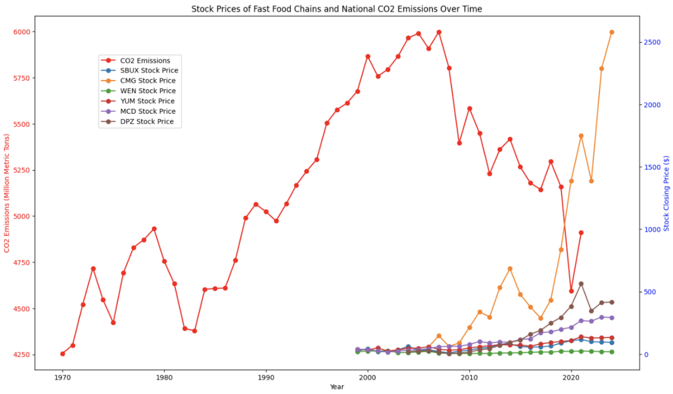

Figure 1. Highlights trends and potential correlations between CO2 emissions and stock prices over the years.

2.3. Data Cleaning and Preprocessing

In order to effectively use our data, we needed to carry out initial cleaning and preprocessing steps. The CO2 emissions data values were converted to numeric types, ensuring any commas were removed and values were properly formatted. The stocks and ETFs dataset was filtered to include only the six specified fast food chains along with their end of year close prices in order to align with the annual nature of the CO2 emissions values.

Missing values in the stock prices were addressed by forward filling to maintain a consistent time series. The stock prices data was then merged with the CO2 emissions data based on the year to create a unified dataset for correlation analysis.

2.4. Visual Analysis

To visualize the relationship between CO2 emissions and fast food stock prices, we created a combined plot with two y-axes. The left y-axis represents CO2 emissions, while the right y-axis shows the stock closing prices for each fast food chain.

2.5. Correlation Analysis

We investigated the overall correlation using all available data points and then examine how these correlations have evolved over different time periods. The first step was to merge the CO2 emissions data with the stock prices data based on the year. This combined dataset allowed us to calculate the correlation coefficients for each fast food company.

The overall correlation coefficients for each fast food company are shown below.

Table 1. Correlation coefficients for fast food companies

Company	Correlation Coefficients
Starbucks (SBUX)	-0.771412
Chipotle (CMG)	-0.768297
Wendy's (WEN)	-0.051483
Yum! Brands (YUM)	-0.821215
McDonald's (MCD)	-0.879937
Domino's Pizza (DPZ)	-0.768541

2.5.1. Historical Correlational Analysis

We also analyzed the correlation coefficients for two distinct time periods: 2000-2009 and 2010-2019. This historical analysis helps identify any shifts in the relationship between CO2 emissions and stock prices over time. The results are shown in the tables below.

Table 2. Correlation coefficients for fast food companies from 2010-2019

Company	Correlation Coefficients
Starbucks (SBUX)	0.357691
Chipotle (CMG)	0.250424
Wendy's (WEN)	0.260377
Yum! Brands (YUM)	0.310383
McDonald's (MCD)	-0.299190
Domino's Pizza (DPZ)	0.540345

Table 3. Correlation coefficients for fast food companies from 2000-2009

Company	Correlation Coefficients
Starbucks (SBUX)	-0.409569
Chipotle (CMG)	-0.249886
Wendy's (WEN)	-0.711469
Yum! Brands (YUM)	-0.632642
McDonald's (MCD)	-0.678731
Domino's Pizza (DPZ)	-0.690555

2.6. Findings

In the period from 2000 to 2009, the correlations were generally positive and weak, indicating that during this time, the growth of fast food companies did not necessarily align with environmental improvements. From 2010 to 2019, the correlations became significantly negative, suggesting a shift in either company practices or broader environmental policies and market conditions.

The observed negative correlation does not imply causation. It is possible that other external factors are influencing both CO2 emissions and stock prices simultaneously. For example, economic downturns could reduce both industrial activity (and thus CO2 emissions) and consumer spending on fast food, leading to lower stock prices. While current sustainability efforts may lead to short-term reductions in CO2 emissions, the long-term environmental impact of the fast food industry, including deforestation, water usage, and pollution, remains significant.

On the other hand, there are several supporting arguments that come to mind. Many fast food companies have increasingly adopted sustainability practices to reduce their environmental impact. This includes sourcing sustainable ingredients, reducing energy consumption, minimizing waste, and improving supply chain efficiency. For instance, McDonald's has implemented initiatives to source sustainable beef, reduce greenhouse gas emissions, and use renewable energy sources. Starbucks has focused on reducing its carbon footprint through measures such as using recycled materials in its packaging and investing in renewable energy.

The shift towards more environmentally friendly practices may be driven by regulatory pressure, consumer demand for sustainable products, and the need to maintain a positive corporate image. Governments and international bodies have implemented stricter environmental regulations and targets, which companies must comply with. key events such as the 2010 United Nations Climate Change Conference may have affected the rate at which CO2 emissions were reduced through policy-making.

Consumers are increasingly aware of environmental issues and are more likely to support companies with strong sustainability credentials. Similarly, investors are becoming more environmentally conscious and are factoring in environmental, social, and governance (ESG) criteria when making investment decisions. Companies that demonstrate strong environmental performance may attract more investment, leading to higher stock prices.

These ideas highlight that a far more nuanced perspective is needed to understand the footprint of processed food in an environmental context.

3. CO2 Emissions and General Agriculture Stocks

3.0.1. Introduction

We then investigated the impact of CO2 emissions on agricultural stock prices, focusing on AGCO Corporation (AGCO), Tractor Supply Company (TSCO), and Deere & Company (DE). Our aim was to determine if there is a correlation between national CO2 emissions and these companies' stock prices, providing insights for investors and policymakers in sustainable investing.

Using the same dataset of stocks and national CO2 emissions, we merged, preprocessed, and analyzed the data. Our exploratory analysis involved calculating correlation coefficients, visualizing trends, and performing seasonal decomposition. For predictive analysis, we employed ARIMA models to forecast stock prices and CO2 emissions, and used Random Forest regression to predict stock prices based on CO2 emissions and other features.

The analysis revealed a nuanced relationship: while there were negative correlations between CO2 emissions and stock prices, CO2 emissions were not strong predictors of stock prices. This comprehensive study using data science and machine learning techniques enhanced our understanding of how environmental factors influence market performance in the agriculture sector.

3.1. Data Collection and Preparation

The preparation of the data involved several key steps to ensure consistency, accuracy, and suitability for analysis. The stock price data was loaded after filtering to contain only the annual closing prices for the selected agricultural stocks. The data was then merged based on the year, creating a combined dataset that aligns stock prices with corresponding CO2 emissions. The combined dataset was filtered to include only the stocks of interest: AGCO, TSCO, and DE. This step ensured that the analysis focused on the agricultural sector.

Missing values in the dataset were addressed by using forward fill and backward fill techniques. This approach ensured that the dataset was complete and ready for analysis without introducing bias.

To capture the temporal dynamics and potential lagged effects of stock prices and CO2 emissions, lagged features were created for up to three previous periods. These features help in understanding how past values influence current values. Additional features, such as year-on-year percentage change, moving averages, and seasonal components (month and quarter), were created to enrich the dataset and provide more context for the analysis. After creating new features, rows with any remaining missing values were dropped to ensure data integrity.

The data collection and preparation process ensured that the dataset was comprehensive, clean, and well-structured, allowing for robust analysis. The combined dataset now includes not only the stock prices and CO2 emissions but also additional derived features that enhance the ability to uncover patterns and relationships. This thorough preparation laid the groundwork for the subsequent exploratory and predictive analyses.

3.2. Exploratory Analysis

Initial bivariate analysis was performed to examine the correlation between CO2 emissions and the stock prices of AGCO, TSCO, and DE.

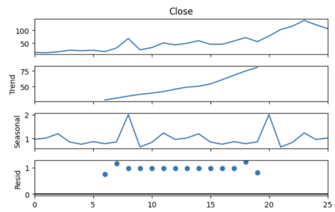

Figure 2. Seasonal decomposition of AGCO stock prices

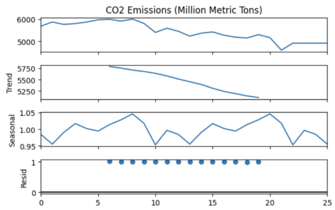

Figure 3. Seasonal Decomposition of CO2 Emissions for AGCO

They were computed to determine the strength and direction of their linear relationship:

Table 4. Correlation coefficients for Agricultural Stocks against CO2 Emissions

Ticker Symbol	Correlation Coefficients
AGCO	-0.807202
TSCO	-0.785975
DE	-0.747529

These negative correlations suggest that higher CO2 emissions are associated with lower stock prices for these agricultural companies.

Seasonal decomposition was performed on the stock prices and CO2 emissions to identify the trend, seasonal, and residual components. This analysis helps in understanding the underlying structure and periodic behavior of the data.

The trend component of AGCO stock prices shows a gradual upward movement over the years. This indicates a general increase in AGCO's stock price, reflecting long-term growth in the company's value. The seasonal component reveals periodic fluctuations within each year. These seasonal variations can be attributed to recurring factors such as agricultural cycles, market demand, and financial reporting periods that influence the company's stock price. The residual component captures the irregular fluctuations that are not explained by the trend or seasonal components. These residuals represent random noise or unexpected events affecting the stock prices. The relatively small and stable residuals suggest that the majority of the variations in AGCO's stock prices are captured by the trend and seasonal components.

The trend component of CO2 emissions exhibits a slight downward trend over the years. This indicates a gradual reduction in national CO2 emissions, possibly due to increased regulatory measures, technological advancements, and shifts towards more sustainable practices. The seasonal component shows regular fluctuations within each year, which could be related to seasonal economic activities, industrial production cycles, and energy consumption patterns that affect CO2 emissions. The residual component highlights the random variations in CO2 emissions that are not explained by the trend or seasonal patterns. The residuals appear to be relatively small and consistent, suggesting that the trend and seasonal components adequately capture the majority of the variations in CO2 emissions.

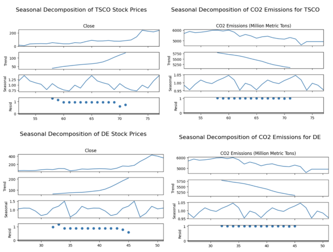

Figure 4. Seasonal decomposition of stock and CO2 for stock, for TSCO and DE

Comparing these two seasonal decomposition plots to those of TSCO and DE, we see almost identical patterns.

The presence of similar upward trends in the stock prices of AGCO, TSCO, and DE indicates that these companies are experiencing industry-wide growth. The consistent seasonal patterns across the three companies suggest that the agricultural industry as a whole is subject to similar seasonal influences. The similar downward trends in CO2 emissions for AGCO, TSCO, and DE imply that external factors such as regulatory measures, technological improvements in emissions reduction, and a shift towards more sustainable practices are uniformly affecting the agricultural sector. This uniformity highlights the effectiveness of industry-wide initiatives aimed at reducing carbon footprints.

3.3. Modeling and Forecasting

In this section, we applied ARIMA models and machine learning techniques to forecast stock prices and analyze the relationship between CO2 emissions and the stock performance of agricultural companies. Our approach involved fitting ARIMA models to historical data and leveraging Random Forest Regressors to predict stock prices based on various features, including CO2 emissions data.

We divided the historical data into training and testing sets, using 50% of the data for training and the remaining 50% for testing. This split was chosen to account for recent anomalies such as the COVID-19 pandemic, as nearly 40% of the most recent available data fell under the period of the pandemic.

For each agricultural stock (AGCO, TSCO, DE), we fitted an ARIMA model to the training data. The ARIMA model parameters (p, d, q) were set to (5, 1, 0), based on preliminary analysis and grid search optimization.

Using the fitted ARIMA models, we generated forecasts for both stock prices and CO2 emissions over the testing period. The forecasts were compared against the actual values to assess the model performance.

The ARIMA model performed reasonably well in capturing the general trend and seasonality in stock prices. However, the forecasts for stock prices showed larger deviations from actual values, indicating potential limitations in using ARIMA for this variable. Meanwhile, the model performed fairly well in forecasting CO2 emissions which was an expected result given that CO2 emissions were not as drastically affected by recent anomalies anyways.

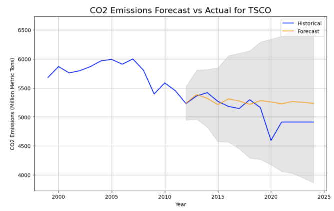

Figure 5. CO2 Emissions Forecast vs Actual for an example stock of TSCO

While ARIMA captured the overall patterns, its predictive power was limited by the high volatility and external shocks in the agricultural market. In order to seek a more robust predictive model, we set out to develop a machine learning model.

We began by engineering a comprehensive set of features, including lagged stock prices, lagged CO2 emissions, year-over-year changes, moving averages, and seasonal indicators (month and quarter). Additional features such as rolling means, rolling standard deviations, and first and second differences of CO2 emissions were included to enhance the model's ability to capture temporal dependencies. We utilized a Random Forest Regressor to model the relationship between the engineered features and stock prices. The data was split into training and testing sets, maintaining the 50% split ratio. The model was trained on the training set and evaluated on the testing set using metrics such as Mean Squared Error (MSE) and r^2 score. The initial Random Forest model achieved a Mean Squared Error of 213.85 and an r^2 score of 0.80. This indicated that the model had good explanatory power, capturing a significant portion of the variability in stock prices.

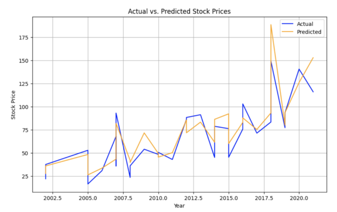

Figure 6. Actual vs Predicted Stock Prices from Random Forest Regressor Model

Feature importance analysis revealed that stock-related features (e.g., lagged stock prices, moving averages) had higher predictive power compared to CO2-related features. This suggested that stock prices were more strongly influenced by their own historical values and trends than by CO2 emissions data.

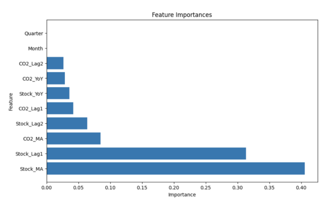

Figure 7. Initial model's feature importances

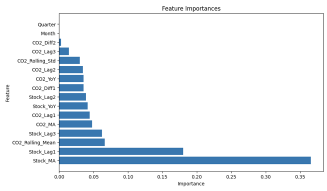

Figure 8. Hyperparametrized model's feature importances

To improve the model, we employed GridSearchCV to perform hyperparameter tuning on the Random Forest Regressor. This involved testing various combinations of parameters to identify the optimal configuration for the model. In an attempt to increase the weight of CO2 emissions data, we introduced additional features such as rolling means, rolling standard deviations, and first and second differences of CO2 emissions. After hyperparameter tuning and adding new CO2-related features, the model's performance deteriorated, with a Mean Squared Error of 462.50 and an r^2 score of 0.65. The feature importance analysis for the tuned model still showed lower importance for CO2-related features compared to stock-related features.

The results suggest that the additional CO2 emissions features did not significantly improve the model's performance, and the feature importance of the CO2-related features remains relatively low. This indicates that the CO2 emissions data might not be strongly predictive of stock prices for the agricultural stocks in our dataset.

In an effort to better understand the nuances of this sector and the impact of processed foods, we continue by analyzing further distinguishing types of pollutants within specific regions in the context of these stocks.

4. Air Pollutants and Stocks at the State Level

To further analyze the relationship between different air pollutants and different processed-food-related stocks, we selected a few key stocks from the the provided list most related to processed foods:

- PepsiCo (PEP)
- Coca-Cola (COKE)
- Tyson Foods (TSN)
- Archer-Daniels-Midlands (ADM)
- ConAgr (CAG)
- Hormel Foods (HRL)
- Pilgrims Pride (PPC)
- Tractor Supply (TSCO)
- Starbucks (SBUX)
- Restaurant Brands Intl (QSR)
- McDonalds (MCD)
- Yum! Brands (YUM)

We processed their closing prices by year and merged this data with the AQI dataset, which contains AQI levels for hundreds of different atmospheric pollutants, categorized by state, over the last 20 years. This data was collected from the United States Environmental Protection Agency.

4.1. Data Selection

First, we focused on key atmospheric pollutants that included Particulate Matter 2.5 (PM 2.5), Sulfur dioxide (SO2), Carbon monoxide (CO), Nitrogen dioxide (NO2), Benzene, Formaldehyde, Acetaldehyde, and Ozone. These pollutants were selected due to their known environmental impact and potential to affect public health and, by extension, economic activities related to the sectors of the selected stocks (Center for Disease Control and Prevention, 2024). For each stock, we aggregated state-wise AQI levels into national averages and created correlation matrices to analyze the relationships between annual stock closing prices and pollutant levels.

Then, we standardized the AQI levels for each selected pollutant to normalize the data across different scales and units. We aggregated these standardized AQI values for each state and created a correlation matrix between the state-wise AQI levels and stock prices for the selected stocks.

Notably, we dropped the two pollutants with the lowest correlations across stocks from the list being analyzed - formaldehyde and acetaldehyde. This is justifiable because they are mostly indoor air pollutants and also are generated at much lower levels than the others (California Air Resources Board, 2024).

4.2. Correlation Examination

The correlations between most of the pollutants and stock price data across states and different stocks tend to be negative. This suggests that higher pollutant levels could potentially correspond to lower stock performance, possibly due to the negative impact on public perception, increased operational costs, or regulatory penalties that affect the financial performance of these companies.

Key pollutants such as SO2, CO, NO2, and Benzene were consistently well-correlated with the stock prices. The relative strengths of these correlations were highly consistent across each of the stocks, with the top correlators—SO2 and Benzene—exhibiting correlation coefficients in the 0.70-0.90 range for all of the stocks analyzed. These pollutants are known to be significant due to their prevalent emissions from industrial and manufacturing processes and their substantial role in air quality degradation (cite some report or study or metric about these being the key pollutants).

State-wise analysis provided intriguing insights: nearly all states showed negative correlations with stock prices as expected. However, states with minimal agricultural activity such as Nevada, Alaska, and Hawaii showed weak positive or almost no correlations, possibly due to less impact from agricultural pollutants or localized economic conditions. Conversely, states known for more intensive agriculture or manufacturing, such as Utah, Vermont, California, Nebraska, and South Carolina, exhibited some of the highest correlations. This pattern underscores the significant impact of localized industrial activities on stock performance in environmentally sensitive sectors like food and beverages. Section 5.4 revisits this concept, underscoring how the economic structure of a state can greatly impact the relationship between processed food and air pollutants.

We initiated our analysis by processing annual closing prices for select stocks, identified due to their relevance in the processed foods sector, including PEP, COKE, ADM, CAG, GIS, HRL, PPC, TSN, TSCO, SBUX, QSR, MCD, and YUM. We then merged this data with the AQI dataset sourced from the United States Environmental Protection Agency. This dataset comprehensively tracks AQI levels for various atmospheric pollutants across U.S. states over the past twenty years, providing a rich backdrop to explore environmental impacts on economic activities.

Our primary focus settled on critical atmospheric pollutants known for their significant environmental and health impacts—Sulfur dioxide (SO2), Carbon monoxide (CO), Nitrogen dioxide (NO2), and Benzene. These pollutants were selected based on their prevalence in industrial emissions and their potential to influence public health and business operations directly, especially in sectors related to our chosen stocks. Each stock was analyzed by aggregating state-wise AQI levels into national averages, and correlation matrices were then developed to discern the relationships between these aggregated pollutant levels and stock prices.

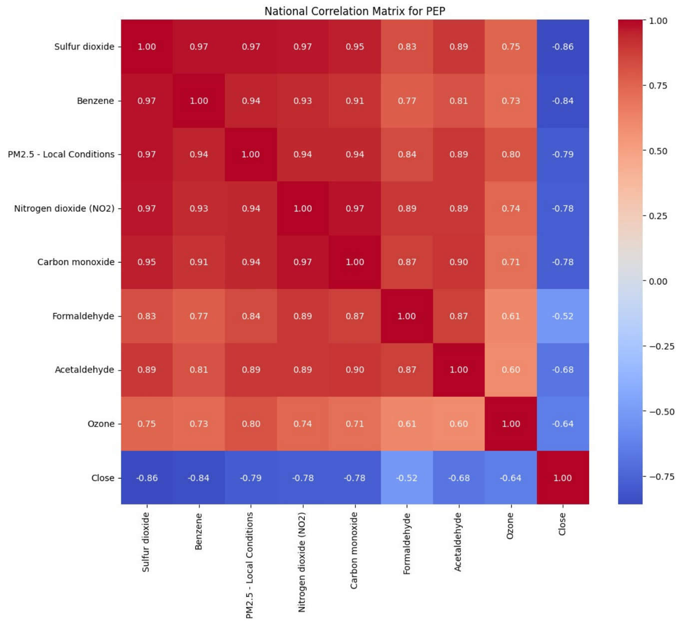

Figure 9. The correlation matrix between stock closing prices and the parts per million of different air pollutants.

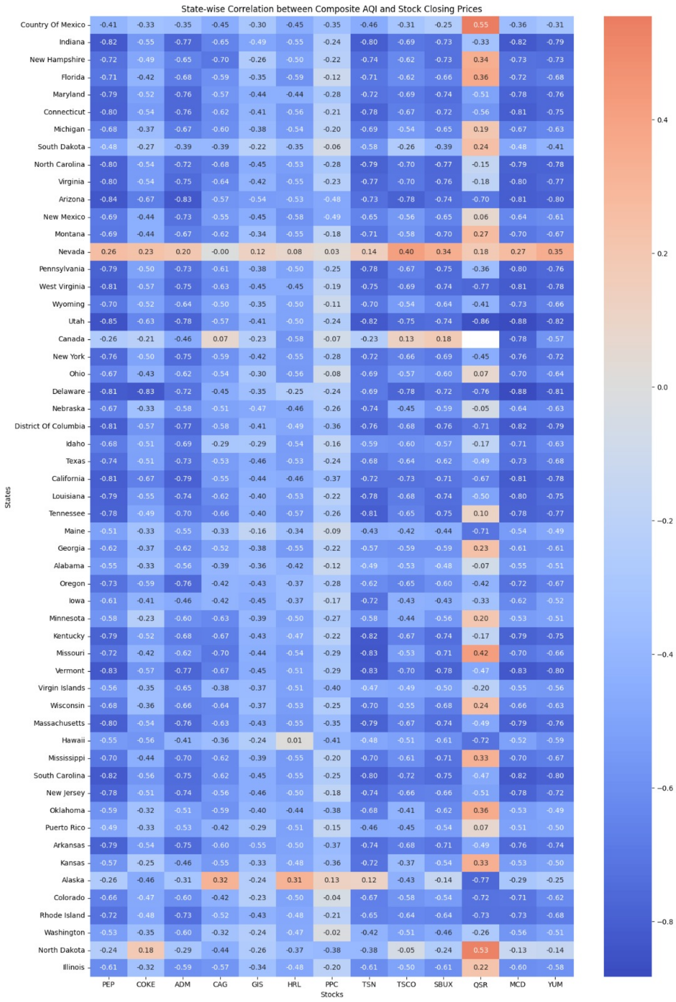

Figure 10. The correlation matrix between the AQI of states and stock closing prices.

To ensure a robust comparative analysis, we standardized the AQI data for these pollutants, normalizing the figures across various scales and units to prevent disparities in measurement from skewing results. This standardization process allowed us to aggregate standardized AQI values for each state, leading to a detailed state-wise correlation matrix between AQI levels and stock prices. In our further analysis, formaldehyde and acetaldehyde were excluded due to their minimal correlation impact, highlighting the necessity to focus on pollutants with more significant economic implications.

Our findings revealed predominantly negative correlations across most pollutants with stock prices, reinforcing the hypothesis that higher levels of environmental pollutants can adversely affect stock performance—potentially due to negative public perceptions, increased operational costs, or regulatory penalties. Particularly, pollutants such as SO2 and Benzene consistently showed strong negative correlations (ranging from -0.70 to -0.90) with stock prices across the board. These pollutants are critically relevant as they are major byproducts of industrial activities that directly degrade air quality, underlining their importance in environmental monitoring. According to the Environmental Protection Agency's air quality indices, these pollutants have been identified as significant due to their widespread impact on urban and industrial air quality.

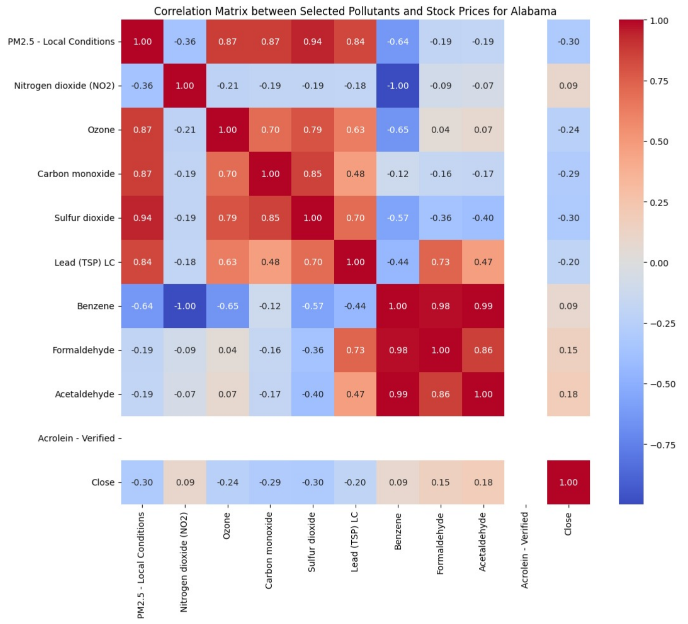

Figure 11. The correlation matrix between select air pollutants and stock prices in Alabama.

In examining the data on a state-by-state basis, nearly all states demonstrated negative correlations as anticipated, with the notable exceptions of Nevada, Alaska, and Hawaii—states known for minimal agricultural output and unique economic structures that might mitigate the typical impacts of industrial pollutants. On the contrary, states with intensive agricultural or manufacturing activities such as California, Nebraska, and South Carolina showed the highest negative correlations, suggesting a direct link between industrial pollutants and lower stock values in companies heavily reliant on environmental quality.

5. Converging Results From Environmental and Socio-Geographic Factors

As stated in our introduction, one of the primary motivations of studying the environmental impact of processed foods is to supplement current research in its socioeconomic and health effects. In previous sections, we've examined the large-scale relationships between processed food stock and environmental pollutants. From those studies, we've found that not only is there a complex and multi-faceted connection between processed foods and environmental markers, but that it varies depending on the state, the year, and the marker we choose to examine. In doing so, we've established that there are a myriad of rich relationships to study, and that they warrant deeper analysis.

In this section, we will illustrate how the relative prevalence of ultra processed, processed, and minimally processed foods in a state has a clear connection to its carbon footprint and food desert prominence. Specifically, we will show that states which consume a high amount of 'processed' food, as opposed to 'ultra processed' and 'minimally processed,' have both low carbon footprints and less severe food deserts. This demonstrates how the socioeconomic and environmental impact of processed foods can converge upon the same result, suggesting a nuanced connection between the two.

5.1. Representing Processed Food Consumption By State

In order to study the relative prevalence of ultra processed, processed, and minimally processed foods, we first needed to gather a dataset which represents the unit sales of different food categories by state. From the U.S Department of Agriculture, we found the weekly sales of alcohols, fats and oils, sugar and sweeteners, fruits, vegetables, and other food groups for each state over a period of five years.

Table 5. A sample from the USDA dataset showing the weekly unit sales of alcohol in Alabama on a weekly timescale. The complete dataset includes many more food groups across 43 states (including the District of Columbia).

Date	State	Category	Unit sales
2019-10-06	Alabama	Alcohol	2974221
2019-10-13	Alabama	Alcohol	2895573
2019-10-20	Alabama	Alcohol	2730939
2019-10-27	Alabama	Alcohol	2764970
2019-11-03	Alabama	Alcohol	2872683
2019-11-10	Alabama	Alcohol	2787906
2019-11-17	Alabama	Alcohol	2715427
2019-11-24	Alabama	Alcohol	2765197
2019-12-01	Alabama	Alcohol	3220147

We then decided to define three different stratifications of processed food: ultra processed, processed, and minimally processed. From previous studies, we found which food groups were most prevalent in different types of processed food and used that to inform our decision. We then assigned each category of food given to us by the USDA dataset to these three stratifications.

Table 6. A chart showing the food processing stratification of each food category listed in the USDA dataset.

Food Category	Processed Level
Alcohol	Ultra Processed
Beverages	Ultra Processed
Commercially Prepared Items	Ultra Processed
Dairy	Processed
Fats and oils	Ultra Processed
Fruits	Minimally Processed
Grains	Processed
Meats, eggs, and nuts	Minimally Processed
Sugar and sweeteners	Ultra Processed
Vegetables	Minimally Processed

Next, we condensed the timescale of the dataset so that it was annual rather than weekly by simply summing the unit sales of every week within a given year. Finally, we calculated the ratios of unit sales of different food stratifications. This gave us the dataset we would use for analysis in the rest of this section.

5.2. Relationship Between Food Consumption and Food Desert Severity

To gather information on food desert severity, we used data from the 2015 U.S Census which for each state tracked both the number of people and percentage of the state's total population living in food deserts. Food deserts – also known as Low Income Low Access tracts – were defined as where the average distance between a residence and its nearest market was over a mile, and where the median income of the region was in the lower 20% of the state's income distribution.

Although the food consumption dataset we created from USDA data did not track information from 2015, we used its earliest year to isolate each state's processed food consumption in order to compare it with their food desert severity.

To search for correlations, we created a correlation matrix between a state's food consumption information and absolute food desert population. We then did the same with a state's food consumption information and percent of population in food deserts.

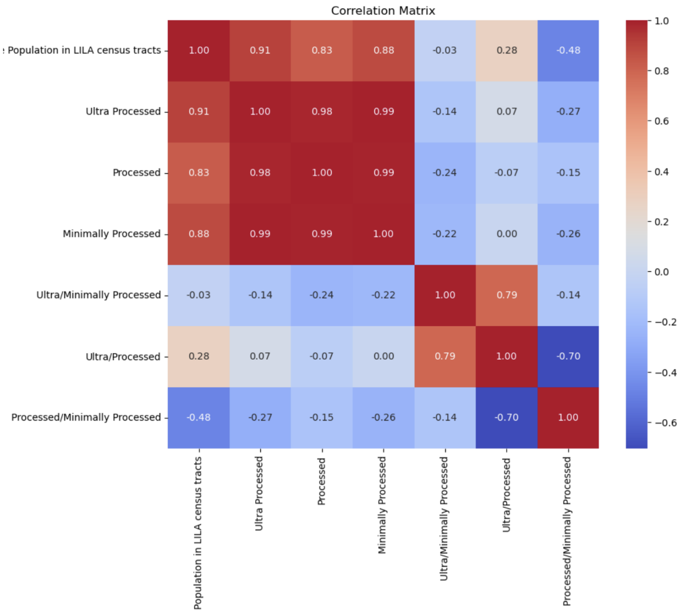

Figure 12. The correlation matrix between food consumption information with food desert absolute population.

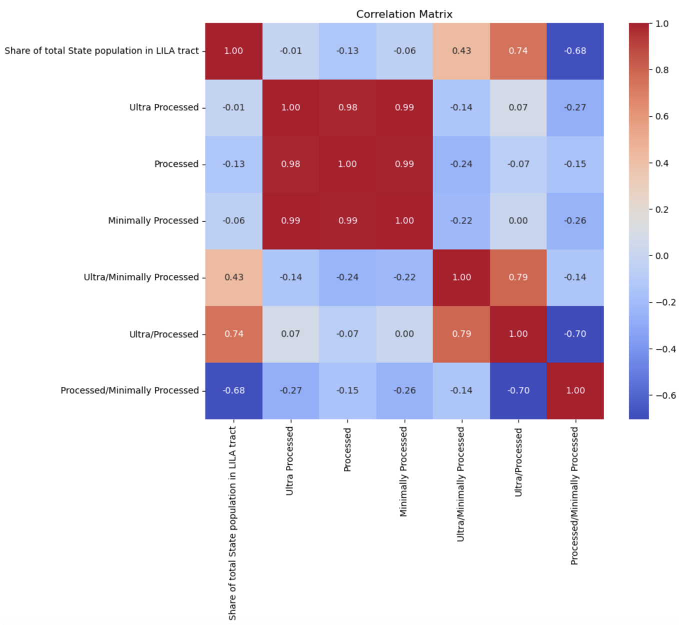

Figure 13. The correlation matrix between food consumption information with food desert relative population.

In conjunction with these matrices, we then identified which correlations had statistically significant Pearson P-Values, which are shown in the tables below.

Table 7. This table shows the correlation coefficient R and Pearson P-Value of correlations between food consumption information with food desert absolute population.

Processed Level	R	P-Value
Ultra	0.91	2.36E-17
Processed	0.82	8.99E-12
Minimally	0.88	3.57E-15
Ultra/Minimally	-0.03	0.84
Ultra/Processed	0.28	0.07
Processed/Minimally	-0.48	0.001

Table 8. This table shows the correlation coefficient R and Pearson P-Value of correlations between food consumption information with food desert relative population.

Processed Level	R	P-Value
Ultra	-0.01	0.94
Processed	-0.13	0.39
Minimally	-0.06	0.72
Ultra/Minimally	0.43	0.004
Ultra/Processed	0.74	1.35E-08
Processed/Minimally	-0.68	5.41E-07

From the tables, we noticed that there are strong and statistically significant correlations between a state's food consumption and the absolute population living in food deserts. We can say that the correlations are strong because the R values are high, and that they are significant because the P-Values are low. However, this set of correlations is likely due to the fact that states with larger populations tend to have both higher food consumption and more people inhabiting any geographic area. The fact that all three food stratifications have about equally strong correlations supports this hypothesis, since only the quantity of food, not the type, seems to matter.

But we do have interesting relationships between a state's food consumption and the percentage of their population living in food deserts. This is because the ratios of 'ultra processed to processed' and 'processed to minimally processed' food consumption have moderately strong yet statistically significant correlations. Specifically, states consuming more 'processed' food in comparison to 'ultra processed' and 'minimally processed' are linked to having a smaller share of their population in food deserts. This is the first incidence of what will be a general trend throughout the rest of this section, where having a higher relative consumption of 'processed' foods is connected to a more ideal state.

5.3. Relationship Between Food Consumption and CO2 Emissions

We have just shown that there is a socioeconomic relationship between the types of foods a state consumes and the severity of food deserts within it. We will now illustrate how the same factors that link states to less severe food deserts also link them to lower carbon emissions.

We first gathered data on each state's carbon dioxide emissions from the year 1970 onwards, using files from the U.S Energy Information Administration. We merged this with our existing information on food consumption, allowing us to compare the two datasets for any year and state.

Just as with the food desert information, we first identified significant correlations with a correlation matrix and Pearson P-Values.

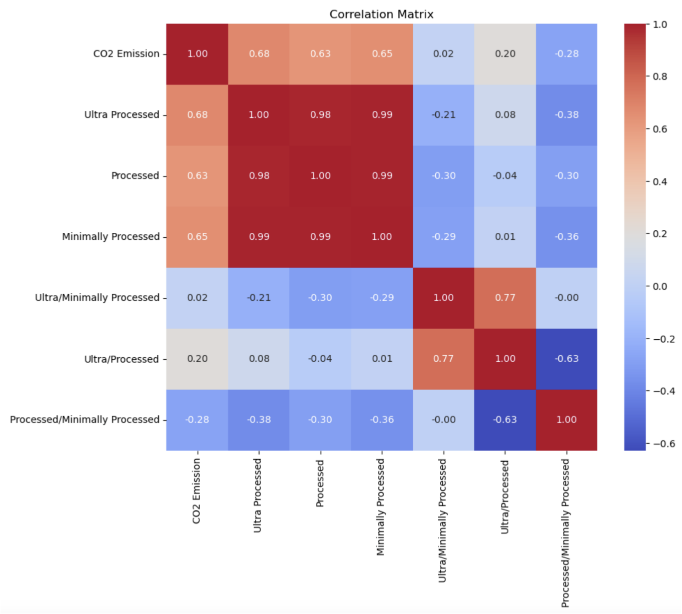

Figure 14. The correlation matrix between food consumption information with carbon dioxide emissions.

Table 9. This table shows the correlation coefficient R and Pearson P-Value of correlations between food consumption information with carbon dioxide emissions.

Processed Level	R	P-Value
Ultra	0.68	7.74E-19
Processed	0.63	2.33E-15
Minimally	0.65	6.81E-17
Ultra/Minimally	0.016	0.85
Ultra/Processed	0.20	0.02
Processed/Minimally	-0.28	0.001

Similar to our analysis of food desert data, we see that there is a strong and significant correlation between the amount of food consumption (regardless of food stratification) and carbon dioxide emissions. This is again likely attributable to populous states requiring both more food and more energy, so it isn't unexpected.

But we also find two weak yet significant correlations between the ratios of 'ultra processed to processed' (which we can denote as UP/P) and 'processed to minimally processed' (P/MP) foods consumed with carbon emissions. States with a higher relative consumption of 'processed' foods tend to have a smaller carbon footprint. This is exactly analogous to our analysis of a state's relative population in food deserts, and yet we arrived at the same result while examining a very different set of factors.

However, we cannot be entirely confident that this relationship between food consumption and carbon footprint is secure due to its low correlation coefficients. We don't expect the value for R to be high because we are likely looking at a very indirect relationship between these variables, but that also means we can't draw any conclusions from these basic characteristics alone. To further investigate, we will use ANOVA and Tukey HSD tests.

5.4. Reinforcing Food Consumption and Carbon Dioxide Emission Relationship

The goal is to check whether or not there truly is a significant difference in the carbon dioxide emissions of a state during a given year if the state has higher or lower values of UP/P and P/MP. This means that we'll need to stratify each state and each year into having either a 'high,' 'medium,' or 'low' value of UP/P. We'll need to do the same sort of stratification for P/MP as well.

We will determine which category a given state and year fall into by finding the tertiles of the distributions of UP/P and P/MP. The region below the first tertile is designated as 'low,' between the first and second tertiles is 'medium,' and above the second tertile is 'high.'

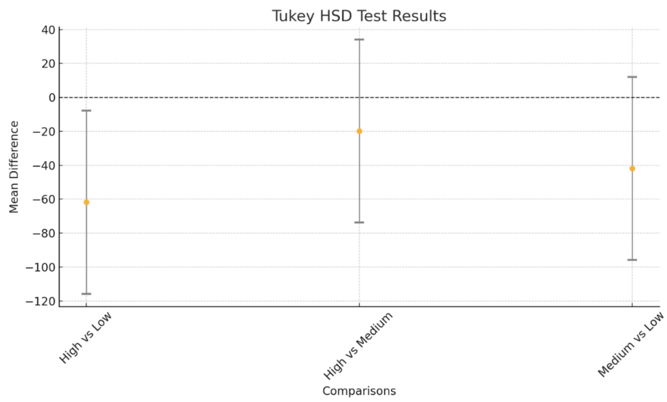

Figure 15. Histogram showing the distribution and tertiles of P/MP.

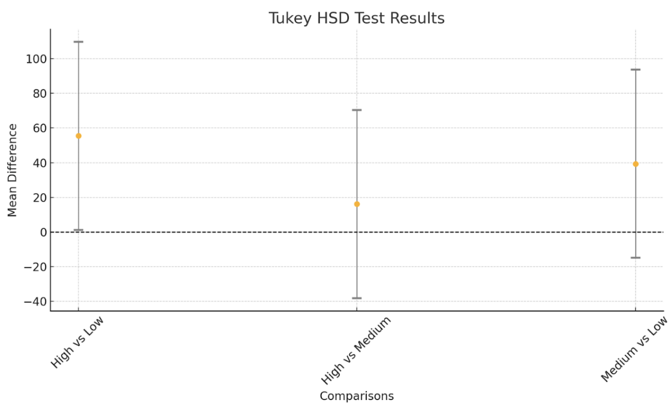

Figure 16. Histogram showing the distribution and tertiles of UP/P.

With these stratifications, we could conduct ANOVA and Tukey HSD tests to see if there was a significant difference in carbon dioxide emissions depending on whether or not a state had a high, medium, or low P/MP during a given year. From the ANOVA result we found that there is a statistically significant difference, and from the Tukey HSD we found that the difference is between the states with 'high' and 'low' values of P/MP.

Table 10. The Tukey HSD results from comparing the carbon dioxide emissions of states with different stratifications of P/MP values.

Strat 1	Strat 2	Mean Diff	P-Adj
High	Low	-61.8	0.0204
High	Medium	-19.9	0.658
Medium	Low	-41.9	0.16

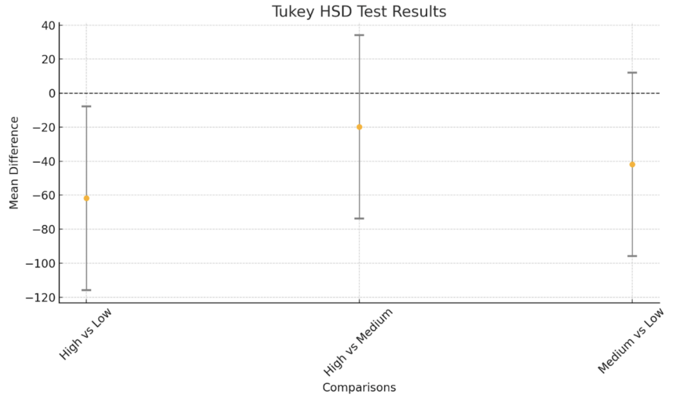

Figure 17. Visualization of the mean difference in carbon dioxide emissions between P/MP stratifications. Higher P/MP values are correlated with lower CO2 emissions with varying levels of confidence.

We ran the same test to compare the carbon dioxide emissions of states with different UP/P stratifications as well, with analogous results. The ANOVA result was statistically significant, and the Tukey HSD result showed that the significant difference was found between states with 'high' and 'low' UP/P values.

Table 11. The Tukey HSD results from comparing the carbon dioxide emissions of states with different stratifications of UP/P values.

Strat 1	Strat 2	Mean Diff	P-Adj
High	Low	55.5	0.044
High	Medium	16.1	0.76
Medium	Low	39.4	0.20

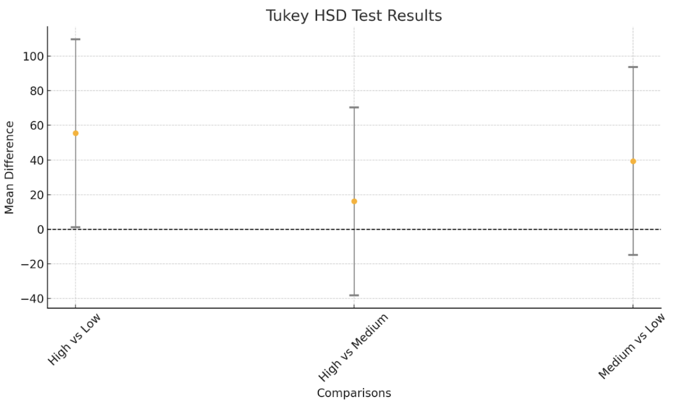

Figure 18. Visualization of the mean difference in carbon dioxide emissions between UP/P stratifications. Lower UP/P values are correlated with lower CO2 emissions with varying levels of confidence.

Together, these two tests support the initial hypothesis that states with a higher relative consumption of 'processed' foods tend to have smaller carbon footprints. There are statistically significant differences in the mean carbon emissions of states in the 'high' and 'low' strata of both P/MP and UP/P. The mean difference is negative for P/MP implying that higher levels of 'processed' relative to 'minimally processed' food consumption is better for low carbon emissions. Meanwhile, the mean difference is positive for UP/P suggesting that higher levels of 'processed' relative to 'ultra processed' food consumption is also better. So once again it seems that a relative prevalence of 'processed' food in a state is linked not only to less severe food deserts, but also a smaller carbon footprint.

To add further nuance in our examination, we finally want to test if this hypothesis holds equally true for all states, or if certain ones exhibit a stronger correlation than others. In particular, we wish to test if states with a more agricultural economy better exemplify this relationship between 'processed' food prevalence and carbon dioxide.

5.5.

We will make one final level of stratification to determine whether or not more agricultural states have a stronger correlation with 'processed' food consumption and carbon dioxide emissions. Using census data, we compiled the labor force breakdown of each state over the years and found the average percentage of the workforce dedicated to agriculture for each one. We determined the tertiles of this agricultural workforce distribution and used them to classify each state as having a 'high,' 'medium,' or 'low' level of agriculture relative to the rest of its economy.

This allowed us to run the same ANOVA and Tukey HSD experiments on each agricultural stratification of states. What we found was that states with a 'medium' or 'low' level of agriculture did not pass the ANOVA test, meaning that there was no significant difference in carbon dioxide emissions depending on P/MP or UP/P values.

However, the 'highly' agricultural states did have significant results. Not only was there a significant mean difference between 'high' and 'low' values of P/MP or UP/P, there were also significant mean differences between 'medium' and 'low' values of P/MP and between 'high' and 'medium' values of UP/P. These additional results further support the hypothesis that states with higher levels of 'processed' food consumption tend to emit less carbon dioxide. But now, we also know that this trend is mostly exhibited by the states with highly agricultural economies.

Table 12. The Tukey HSD results from comparing the carbon dioxide emissions of highly agricultural states with different stratifications of P/MP values.

Strat 1	Strat 2	Mean Diff	P-Adj
High	Low	-170.2	0.0089
High	Medium	-14.4	0.962
Medium	Low	-155.84	0.0177

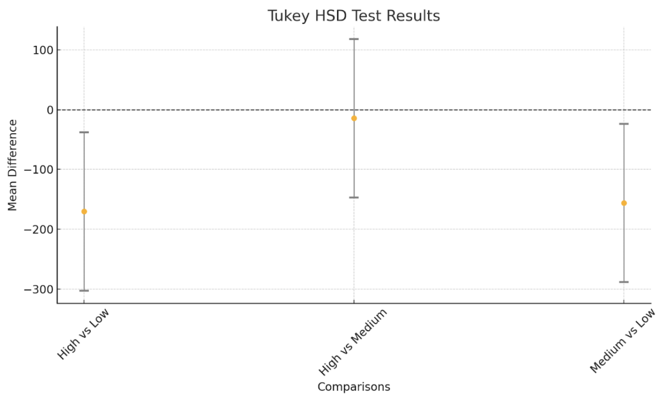

Figure 19. Visualization of the mean difference in carbon dioxide emissions between P/MP stratifications in highly agricultural states. Higher P/MP values are correlated with lower CO2 emissions with varying levels of confidence.

Table 13. The Tukey HSD results from comparing the carbon dioxide emissions of highly agricultural states with different stratifications of UP/P values.

Strat 1	Strat 2	Mean Diff	P-Adj
High	Low	183.5	0.0036
High	Medium	168.2	0.008
Medium	Low	15.39	0.955

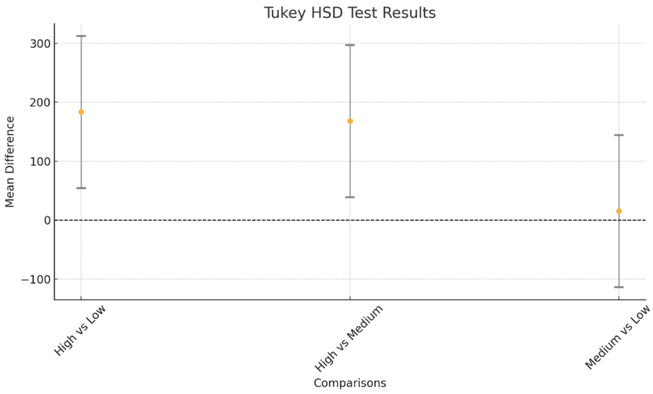

Figure 20. Visualization of the mean difference in carbon dioxide emissions between UP/P stratifications in highly agricultural states. Lower UP/P values are correlated with lower CO2 emissions with varying levels of confidence.

5.6. Take-Aways

From our examination of food consumption and its relation with carbon emissions and food deserts, we've seen how the socioeconomic and environmental impacts of processed foods can be closely linked. States that consume high amounts of 'processed foods,' which we define as mostly dairy and grain, in comparison to 'ultra processed' and 'minimally processed' foods tend to have a smaller carbon footprint and less severe food deserts. Furthermore, the connection between carbon emissions and food consumption is mostly found among highly agricultural states.

We hypothesize that a prevalence of 'processed' foods could be connected to better food transportation and preservation infrastructure. For example, grain can be shipped in trucks like most foodstuffs, but it is also a noteworthy cargo of freight trains due to its long shelf life and high demand (Perry, 2019). So states that consume large amounts of grain would need robust transportation networks, and potentially several different methods of transportation, which could be used to mitigate food deserts. Similarly, dairy is easily spoiled so a state that consumes large amounts of it must have substantial cold storage facilities and efficient transportation mechanisms. These may also reduce areas in which food deserts could occur.

At the same time, 'processed' foods may also be less energy intensive to cultivate en masse. Compared to fresh produce and meats, which make up the 'minimally processed' category, this holds true. And compared to fats, alcohol, sugar, and the other foods that make up the 'ultra processed' category, dairy and grains take far less manufacturing (and therefore less energy) to create. This could potentially explain why states consuming high amounts of 'processed' food have lower carbon emissions.

These explanatory hypotheses could be the subject of future research, in order to build upon the connections drawn in this section.

6. Supply Chain Footprint Analysis

6.1. Introduction

In any large civilization, the infrastructure built to transport large quantities of specialized goods from one place to another represents its blood vessels. As a result, there exists a particular importance in understanding each vessel and each artery. Indeed, these lines can be seen in Figure 1, which depicts one of the end product visualizations depicting the flow of commodities across America. As seen in our previous analysis, prevalence in agriculture and manufacturing can act as proxies for pollutants per capita, meaning foods that require a high number of steps in order to produce (i.e., UPFs). In fact, according to climate research in *Future Foods*, UPFs make up for a disproportionately large percent of the emissions related to food production (Shabir et al., 2025). As a result, attention to detail regarding the supply chain and the manufacturing costs of each category is of utmost importance.

6.2. Methodology

Our methodology for analyzing the supply chain footprint involves several key steps, each incorporating specific data collection, preprocessing, and analytical techniques.

6.2.1. Data Collection and Preprocessing

We begin by gathering ultra-processed food (UPF) data. This data includes information on the number of ingredients, sourcing locations, and the steps involved in the manufacturing process on a state wide level. To predict whether local sourcing is possible for each ingredient, we use logistic regression:

$$P(\text{local sourcing possible}) = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n}}$$

where $X_1, X_2, \ldots, X_n$ are the predictor variables.

To estimate the quantity of each ingredient needed, we apply gamma regression:

$$E(Y) = \frac{1}{\alpha \cdot \beta} \quad \text{where} \quad Y \sim \text{Gamma}(\alpha, \beta)$$

In conjunction with this cut of the data, we pull in from several Google Maps API, including GeoLocator, Distance Matrix, and Route Finder. By using the GeoLocator API, we were able to separate out the land mass within the United States into 2086 50x50 mile squares, which allowed us to compute vast numbers of source nodes, routes, and sink nodes.

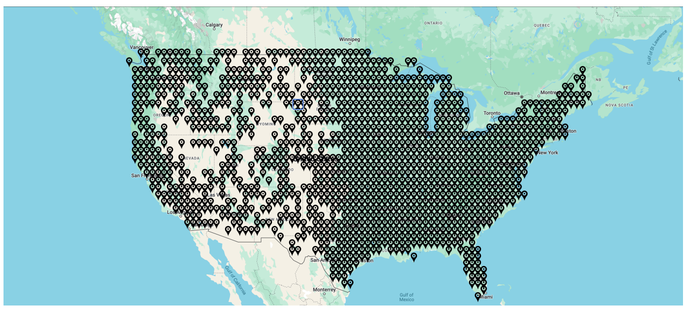

Figure 21. Visualization of sqaure centers, which will act as the POIs of the following code and algorithms.

6.2.2. Sourcing Efficiency

Next, we analyze the efficiency of sourcing routes. Given the cost data, we aim to minimize the total cost using the following optimization formula:

$$\min \sum_{i=1}^{n} C_i \cdot X_i \quad \text{subject to constraints}$$

where C_i represents the cost of sourcing route i and X_i is a binary variable indicating if route i is chosen. For example, in Figure 22, we see that within the radius of our POI, we have successfully identified several key characteristics of the supply chain (farms, food processing plants, distribution centers, and end markets). In the product dataframe, we are able to write in stratifications such as whether it is a fruit farm, vegetable farm, or cattle ranch when analyzing source nodes.

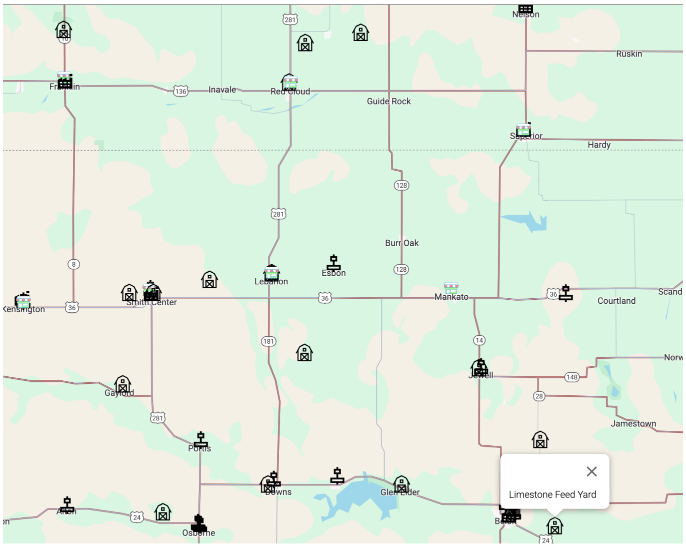

Figure 22. Visualization of key characteristics in radius of POI

6.2.3. Distribution Algorithm

To distribute the processed foods efficiently, we develop a distribution algorithm that considers population data, store locations, as well as more macro data on county to county transfer of food products. The weight for distribution is calculated as:

$$W = \frac{P_i}{D_{ij}}$$

where P_i is the population of area i and D_{ij} is the distance from area i to store j. Through this, we are able to produce Figure 23, which clearly depicts the aforementioned "blood vessels" of the United States.

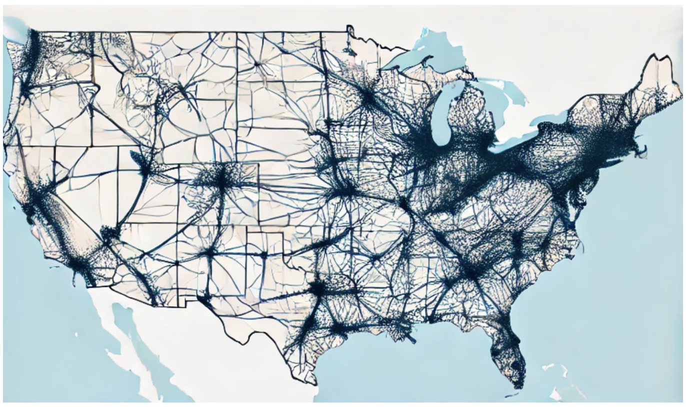

Figure 23. Visualization of flow of products from source, distribution networks, and end markets through our algorithms and validation methods based on concrete data regarding county to county transfers by freight

6.2.4. Environmental Impact Analysis

To understand the environmental impact, we calculate CO2 emissions and other pollutants across the supply chain. The total emissions are calculated as:

$$E = \sum_{i=1}^{n} (Q_i \cdot D_i \cdot E_i)$$

where Q_i is the quantity of ingredient i, D_i is the distance traveled for ingredient i, and E_i is the emissions factor for ingredient i.

We also assign weights to each food type based on the number of processes involved and the distance traveled, as well as most likely mode of transportation. These modes of transportation may include freight by cargo train, boats, trucks, and plane:

$$W_i = f(\text{Number of Processes, Distance})$$

6.2.5. PageRank Algorithm for Local Sourcing

To assess the likelihood of local sourcing, we implement a PageRank algorithm:

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{PR(j)}{L(j)}$$

where d is the damping factor, N is the total number of nodes, $M(i)$ is the set of nodes linking to i, and $L(j)$ is the number of outbound links from node j.

6.2.6. Graph Search for Distance Estimation

Finally, we use graph search algorithms (BFS/DFS) to estimate the average distance traveled for each ingredient:

$$\text{Distance}(s,t) = \text{Shortest path from } s \text{ to } t$$

We then sum the distances and emissions to get the total impact:

$$\text{Total Emissions} = \sum_{\text{all paths}} (\text{Distance} \times \text{Emissions factor})$$

6.3. Results and Limitations of Analysis

Through this analysis, we are able to bucket out the results into the emission impacts of minimally processed, moderately processed, and ultra processed foods as previously delineated. In doing so, we find that UPFs yield a 2.4x coefficient in terms of amount of emissions per percent of daily food intake, while moderately processed foods yield a 1.7 coefficient using minimally processed foods as a baseline. This is a result of previously explored behaviors shifted by lower costs of processed goods particularly impacting demographics such as those in food deserts, those of low income, and those in rural areas.

7. Conclusions and Further Study

In this report, we investigated the environmental impact of processed foods in the US. We began by examining stock prices related to processed foods and their correlations with carbon dioxide emissions, air quality index, and other air pollutant concentrations. We illustrated how these relationships are complex and can vary over many factors, such as states, state economies, and potency of air pollutant. We then examined the relative consumption of processed foods in various states over time, showing how the consumption of dairy and grains correlated with lower carbon dioxide emissions. Furthermore, the same factor of high dairy and grain consumption also correlated with an apparently unrelated societal factor, the lower severity of food deserts. Finally, we created a methodology to study the transportation of processed food across the country, using API software to create a graph. From emissions extrapolations and transportation trends, we can predict the hidden environmental cost of processed foods.

8. References

BNSF Railway Company. (n.d.). BNSF agricultural shuttle map. Retrieved August 4, 2024, from https://www.bnsf.com/bnsf-resources/pdf/ship-with-bnsf/maps-and-shipping-locations/bnsf-agricultural-shuttle-map.pdf

BNSF Railway Company. (n.d.). BNSF network map. Retrieved August 4, 2024, from https://www.bnsf.com/bnsf-resources/pdf/ship-with-bnsf/maps-and-shipping-locations/bnsf-network-map.pdf

Association of American Railroads. (n.d.). Freight rail and the agriculture industry. Retrieved August 4, 2024, from https://www.aar.org/issue/freight-rail-agriculture-industry/

Federal Highway Administration. (n.d.). National highway system (NHS) major truck routes: 2040. Retrieved August 4, 2024

Clark, M. A., Springmann, M., Hill, J., & Tilman, D. (2020). Multiple health and environmental impacts of foods. The Lancet Planetary Health, 4(10), e426-e437. https://doi.org/10.1016/S2542-5196(20)30177-7

Lawrence, F. (2021, June 17). The truth about processed foods' environmental impact. BBC Future. Retrieved August 4, 2024, from https://www.bbc.com/future/article/20210617-the-truth-about-processed-foods-environmental-impact

Garnett, T., & Godde, C. (2020). The fruit and vegetable supply chain: Impacts and opportunities for climate mitigation. Retrieved August 4, 2024, from National Center for Biotechnology Information. (2019). Grains and their health impacts. Retrieved August 4, 2024

Bureau of Transportation Statistics. (n.d.). Transportation delivers harvest from field and farm to family tables. Retrieved August 4, 2024, from https://www.bts.gov/data-spotlight/transportation-delivers-harvest-field-and-farm-family-tables

East Carolina University Physicians. (2021). NOVA classification reference sheet. Retrieved August 4, 2024, from https://ecuphysicians.ecu.edu/wp-content/pv-uploads/sites/78/2021/07/NOVA-Classification-Reference-Sheet.pdf

Federal Highway Administration. (n.d.). Freight analysis: State information. Retrieved August 4, 2024

Economic Research Service, USDA. (n.d.). Weekly retail food sales documentation. Retrieved August 4, 2024, from https://www.ers.usda.gov/data-products/weekly-retail-food-sales/documentation/

U.S. Energy Information Administration. (n.d.). State carbon dioxide emissions. Retrieved August 4, 2024, from https://www.eia.gov/environment/emissions/state/

Perry, T. (2019, October 28). The first map of America's food supply chain is mind-boggling. Fast Company. Retrieved August 4, 2024, from https://www.fastcompany.com/90422553/the-first-map-of-americas-food-supply-chain-is-mind-boggling

Economic Research Service, USDA. (n.d.). Food access research atlas: State-level estimates of low-income and low-access populations. Retrieved August 4, 2024, from https://www.ers.usda.gov/data-products/food-access-research-atlas/state-level-estimates-of-low-income-and-low-access-populations/