Sizing the Pie: Decoding Economic & Societal Drivers Behind U.S. Food Industry Segments - 2024 Citadel Summer Invitational Datathon Final Report

Team 2: Bo Chi, Calvin Huang, Tanish Kumar, Tony Zhang

August 4, 2024

1 Executive Summary

The production and processing of food is one of the largest industries in the United States, with its various sectors contributing significantly to the economy. In the chain of production from the food source to the consumer lie numerous stakeholders including farmers, processors, distributors, retailers, and ultimately, consumers.

Evaluating the behavior of financial markets could provide us with deep insight about the economic health and food sectors' stability. Moreover, fluctuations in the financial market could induce a great impact on the prices of downstream supply chain such as retailers and food industry.

In summary, we hope to explore the following:

Topic Question: Which factors have the greatest impact on the core segments of the processed food economy's production-consumption cycle? Where may these sector-level shifts be reflected in broader socioeconomic indicators?

To examine the first question, we aggregated relevant stocks into four clusters representing different industry segments (Agriculture & Machinery, Food & Beverages, Restaurants & Fast Food, and Retail). To understand consumer-side dynamics in a structured way, we dive into the U.S. Supplemental Nutrition Assistance Program (SNAP). Specifically, we explore enrollment rates across the country, and determine how the USDA's decisions regarding which food retailers are SNAP-eligible influence and shape consumer purchasing decisions.

In order to classify the relative contributions of different drivers toward food market performance, we constructed a LASSO linear regression model to map potential features to the yearly performance of these clusters and preferentially drop variables contributing less to the predictions. This model identified important features to be the production flows of various meat classes (especially with lamb, beef, and turkey volumes) and the prices of key commodities (sugar, coffee, and corn). With this initial list, we sought to further study how these top factors influence our target stock cluster pricing at more precise timescales. We designed an XGBoost model to calculate month-to-month links between the features and targets of interest and were able to predict pricing with relatively high accuracy $0.78 \sim 0.83$. Our principal components analysis additionally revealed that these trends were largely driven by shifts in meat production (with chicken and pork being strong positive drivers and lamb and veal being negative drivers), as well as higher commodity prices, which were were also linked to increases in industry stocks.

Finally, to answer the second question, we studied the downstream effects on overall market performance (through index funds), historical US unemployment rates in relevant sectors, such as agriculture and leisure/hospitality, and health metrics such as obesity rates in both children and adults.

Our results suggest a strong link between the factors selected and the strength of the food supply chain as a whole. Further study on regional-level economic differences will allow for better predictive models of economic performance and guide future research on potential causal relationships with the industry-relevant metrics used.

Contents

2 Technical Exposition

2.1 Data Approaches

In order to systematically characterize the effects different drivers display on the US processed foods economy, we made use of a wide array of data measuring the changes over time in those public- and industry-level metrics. Table 2.1.1 shows the metrics being considered.

Category	**Potential Features**
Supply Chain Metrics	Meat Production (lbs) - split by animal Commodity Prices ($)
Sector Employment	People Employed in Key Industry Segments (#)
Health-Related Community Metrics & Programs	Health Insurance Status SNAP Commitments (# Households / Persons & Spending ($))
Stock & Index Prices (*Target Variable*)	Food and Beverage Stocks ($) Restaurant and Fast Food Stocks ($) Retailer Stocks ($) Agricultural and Machinery Stocks ($) Market Index Funds ($)

Figure 2.1.1: Data considered in our modeling approach

With this initial broad set of data points across different US economy and community health segments, we sought to consider effective aggregation methods to reduce individual swings and adequately represent the state of this US industry.

2.1.1 Measuring Supply Chain Performance Using Synthetic ETFs

There are 29 distinct stocks and ETFs provided in the `all_stock_and_etfs.csv`, along with their time series data (i.e open, close, low, high prices.) For illustration purposes, we provide the visualization of the daily mean prices of these 29 securities in Figure 2.1.2. Individual stocks, however, may not exhibit strong enough correlations with broader impacts—such as obesity—to yield meaningful insights. **Furthermore, using individual stock performance as proxies for broader supply chain economies can introduce significant noise and bias.** This is because the multifaceted nature of supply chain dynamics involves numerous factors beyond the scope of individual stock performance.

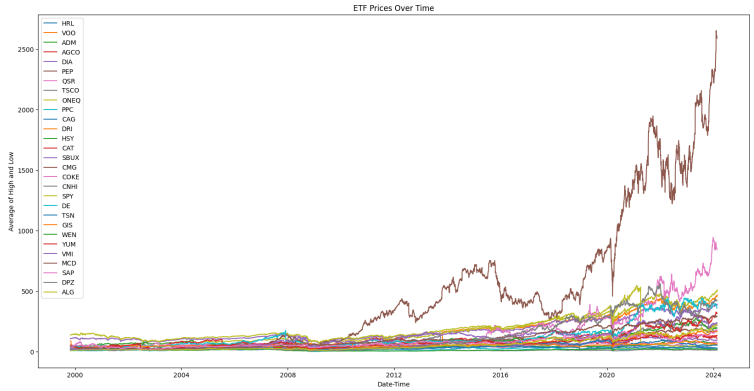

Figure 2.1.2: ETFs & Stocks Prices Over Time

Hence, we propose the following methodology for quantifying the distinct components of the processed food economy and optimizing feature selection.

We first grouped all the ticker symbols present within `all_stock_and_etfs.csv` into 5 clusters according to their different roles as market participants, with details provided in Table 2.1.3. We calculated the VWAP (Volume-Weighted Average Price) of each security by multiplying the closing price ('Close') by the volume ('Volume') to get the total traded value, grouping the data by year, month, and ticker symbol, and then dividing the total dollar volume by the total volume. We then summed the VWAPs of the securities in each ETF to create a baseline value for analysis. Furthermore, we calculate a 5-year average close price for each ticker symbol by using a rolling mean with a 60-month window.

Category	**Companies**
Food and Beverage Manufacturers	HRL (Hormel Foods Corporation) ADM (Archer-Daniels-Midland Company) PEP (PepsiCo, Inc.) CAG (Conagra Brands, Inc.) HSY (The Hershey Company) TSN (Tyson Foods, Inc.) GIS (General Mills, Inc.) COKE (Coca-Cola Consolidated, Inc.) PPC (Pilgrim's Pride Corporation)
Restaurant and Fast Food Chains	QSR (Restaurant Brands International) DRI (Darden Restaurants, Inc.) SBUX (Starbucks Corporation) CMG (Chipotle Mexican Grill, Inc.) WEN (The Wendy's Company) YUM (Yum! Brands, Inc.) MCD (McDonald's Corporation) DPZ (Domino's Pizza, Inc.)
Retailers	TSCO (Tractor Supply Company)
Agricultural and Machinery Companies	AGCO (AGCO Corporation) CAT (Caterpillar Inc.) DE (Deere & Company) CNHI (CNH Industrial N.V.) VMI (Valmont Industries, Inc.) ALG (Alamo Group Inc.)
Investment Funds and ETFs	VOO (Vanguard S&P 500 ETF) DIA (SPDR Dow Jones Industrial Average ETF Trust) ONEQ (Fidelity Nasdaq Composite Index ETF) SPY (SPDR S&P 500 ETF Trust)

Figure 2.1.3: Companies and ETFs categorized by type

However, results with deep insights cannot be obtained without fully exploiting the dataset. With the clusters of relevant stock indicators, we decided to further normalize the ETF movements by using algorithm 1, and obtain the normalized ETF movement parametrized by the VWAP and 5Y_Avg_Close (5-Year Average Close price) for downstream models and statistical analysis, see Figure 2.1.4.

Algorithm 1 Calculate_Normalized_ETF(DataFrame[Year-Month-Day, Open, High, Low, Close, Volume, Ticker_Symbol]):

Initialize Normalized_DataFrame
for (year, month, ticker_symbol) ∈ DataFrame **do**
    subdataframe ← DataFrame[(year, month, ticker_symbol)]
    total_volume ← sum(subdataframe[Volume])
    total_traded_value ← sum(subdataframe[Volume] * subdataframe[Close])
    Normalized_DataFrame[(year, month, ticker_symbol)].VWAP ← total_traded_value / total_volume
    Normalized_DataFrame[(year, month, ticker_symbol)].5Y_Avg_Close ← avg. of previous 60 months' Normalized_DataFrame[(year, month, ticker_symbol)].VWAP
end for
return Normalized_DataFrame

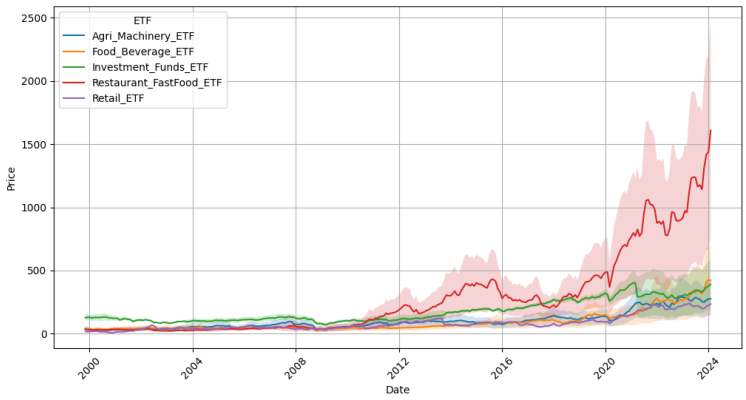

Figure 2.1.4: Monthly ETF Prices Over Time with Volatility

2.2 Exploratory Data Analysis

In order to evaluate the effects each of our individual driver classes may have on the prices of our synthetic ETFs, we performed comparative analyses to form an initial understanding about the relationships between these data points over time.

2.2.1 Meat Production Trends

Meat production is a central component of the US food economy, with changes in its supply being meaningfully reflected down the line, from the farmers to the final restaurants. The baseline metric to evaluate the activity of the US meat industry is the total weight of meat produced at any given time. Figure 2.2.1 displays the changes in monthly meat production over time between different meat classes.

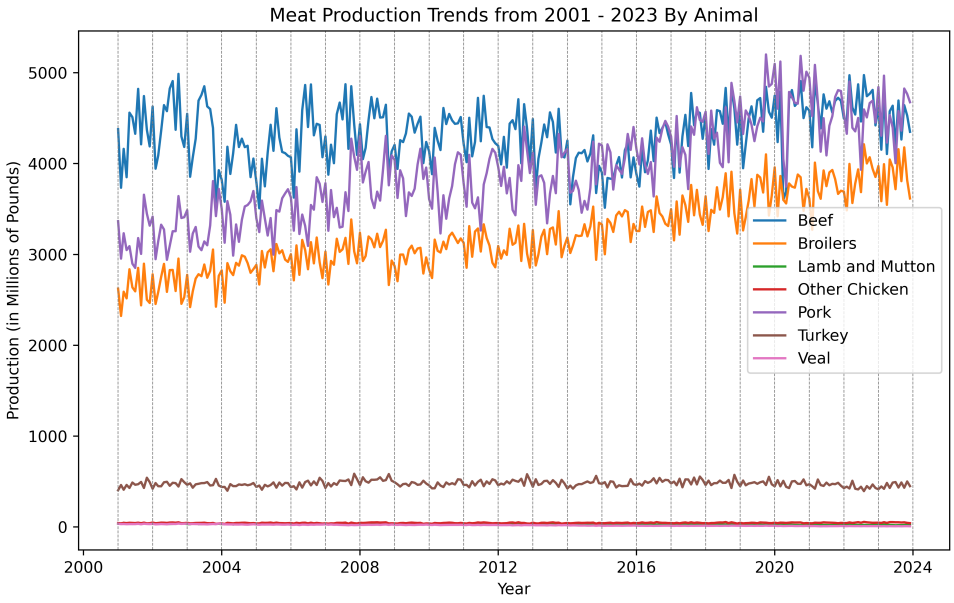

Figure 2.2.1: Meat Production Values from 2001 - 2023

This chart suggests that meat production is relatively stable in the US, with certain classes, namely broilers (primary type of chicken) and pork, on a steady rise over the last 20 years. Additionally, the oscillations in the graph suggest highly seasonal patterns drive US meat production flows. The relationships between these meat production trends were combined with synthetic ETF performance in Figure 2.2.2.

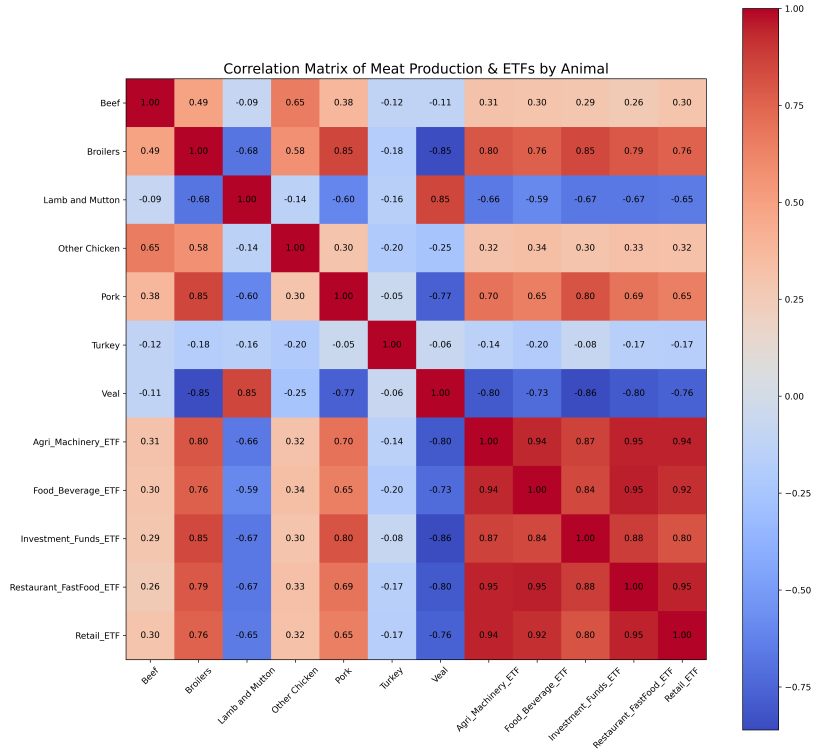

Figure 2.2.2: Meat Production vs. Supply Chain Indicators

These results suggest relatively strong links between certain meat classes and synthetic ETF performance and highlight meat production as a potentially useful factor to link to us.

2.2.2 ETF vs Commodities

The prices of ETFs and commodities are often interlinked, since the production of commodities could potentially have a profound influence of downstream supply chain industry. First we exhibit a line graph for the 3 different kinds of commodities[1] provided in `all_commodities.csv`: Corn, Coffee, Sugar. See Figure 2.2.3.

[1]There were 3 different kinds of commodities provided in the dataset: Coffee, Sugar, nan. The "nan" is confirmed to be Corn after verification with the technical staff of Correlation One.

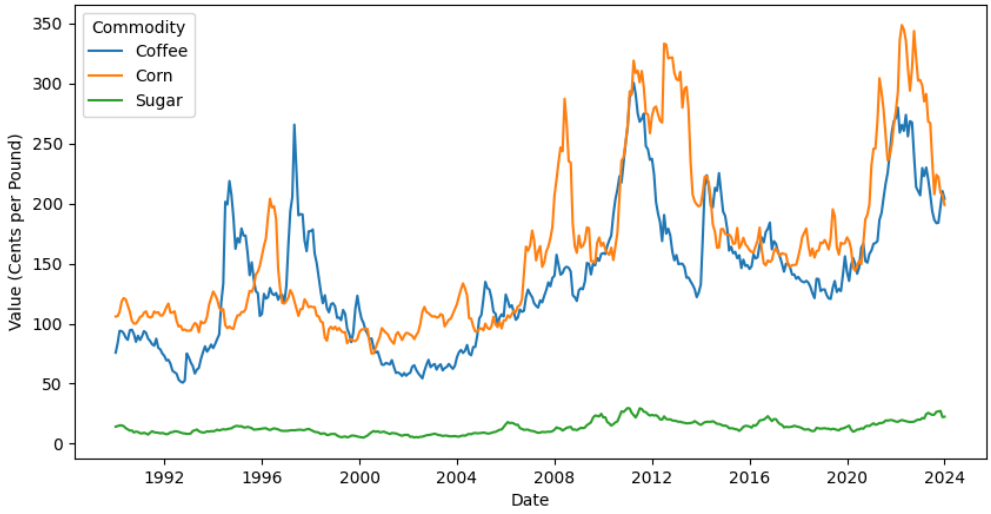

Figure 2.2.3: Values of Commodities from 1990 to 2024

Now we illustrate the interplay between ETFs and Commodities by computing the correlation matrix of them, which serves as motivation for features selection of future models as well as shedding some insight for the individual relations between items in these two datasets. See Figure 2.2.4.

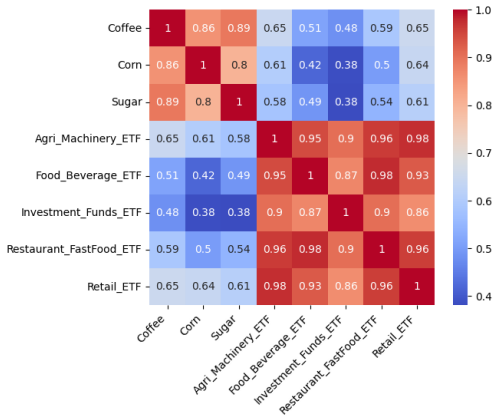

Figure 2.2.4: Yearly Commodities vs. Supply Chain Indicators

2.2.3 Federal Nutrition Assistance Programs

The Food Stamp Program (FSP) in the United States begin around the 1940s, the support farmers and the poor in the aftermath of the Great Depression. Since then, the program has transformed into what is known today as the Supplemental Nutrition Assistance Program (SNAP) with the goal of providing "food benefits to low-income families to supplement their grocery budget so they can afford the nutritious food essential to health and well-being" [1]. Those who qualify for the program are given an EBT card which can be used at eligible food retailers to purchase allowed items. In 2023, the federal government spent a total of $113

billion on the SNAP program, $107 billion of which went directly as households as monthly benefit checks. With an average of about $73 billion per month in food retailer revenuees, we estimate that nearly 12% of food retailer revenues are supported by government spending on SNAP [3]. Given this, we aim to understand whether USDA fuels the consumption of highly processed foods, contributing to obesity, or whether the USDA's stated goal for SNAP holds true.

To begin, we started by looking at US enrollment percentages in SNAP from 2000 to 2020, as seen in 2.2.5. This is calculated by taking the total number of people enrolled in SNAP per year divided by the estimated nationwide population. We decided to use 2000-2020 as the time frame to avoid COVID-19 pandemic disruptions and ensure appropriate data availability.

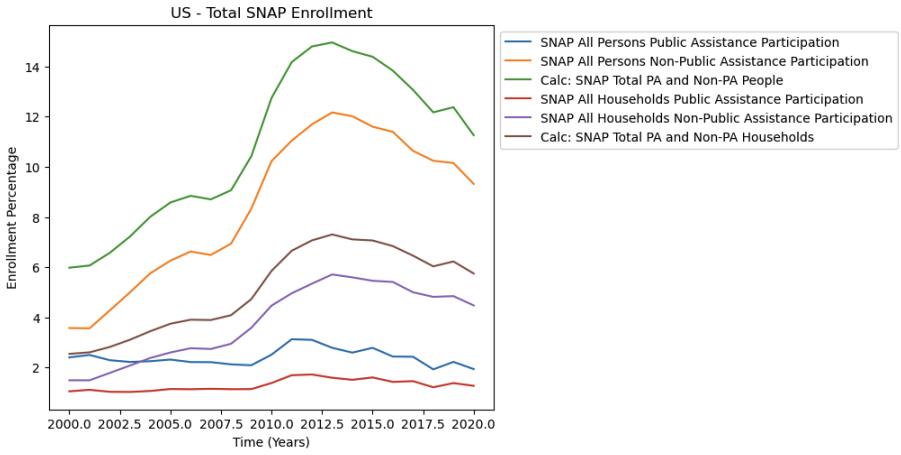

Figure 2.2.5: Nationwide SNAP Enrollment Through the Years

As we can see, enrollment percentages have nearly doubled over the time period, with a notable spike starting around 2008. We believe this uptick can be attributed to the passing of the Food, Conservation, and Energy Act of 2008, which significantly expanded enrollment eligibility and benefits for U.S. citizens.

We further visualized SNAP enrollment by state, as seen in Figure 2.2.6. To take into account varying state populations, we computed the SNAP enrollment per capita, by dividing the total enrollment of people in SNAP per state, and dividing by the estimated state population. In this section, we decided to use 2000-2020 as the time frame to avoid COVID-19 pandemic disruptions with regards to government policy and ensure appropriate data availability.

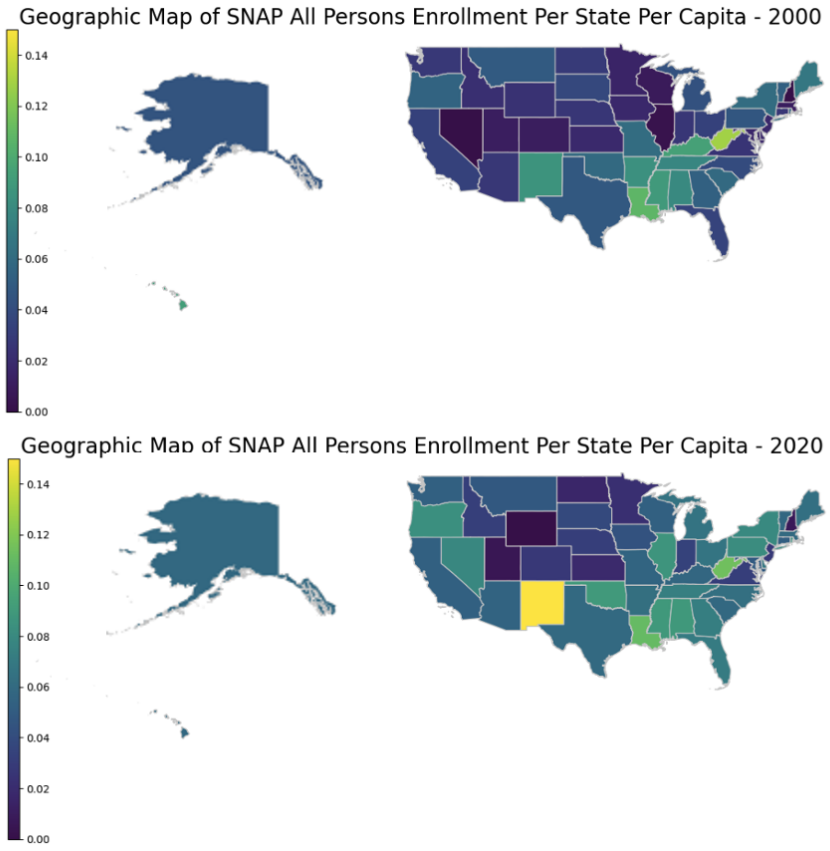

Figure 2.2.6: 2000 and 2020 Heatmaps for SNAP Enrollment Nationwide

For nearly all states, we see the enrollment per capita increasing, leading to an average individual enrollment to about 13% of the entire population nationwide in 2020. To understand how the over 43 million US citizens receiving SNAP benefits are using them, we turn to data published by the USDA on authorized SNAP-eligible retailers [5]. We argue that the retailers the USDA classifies as SNAP-eligible shapes and directs the purchasing of those using SNAP benefits. Therefore, analyzing SNAP-eligible retailers helps us understand the choices that enrolled consumers will ultimately make. The USDA classifies SNAP-eligible food retailers into 17 categories, described in full here: [6]. Based on these definitions, we find the following assumption reasonable. The following categories offer the highest concentration of highly processed foods, as compared with the rest of the categories: Convenience Stores, Combination Grocery/Other, Supermarket, and Super Store/Chain Store. This is due to the fact that there categories of establishments generally tend to lie at the end of a longer supply chain, wherein raw goods are repeatedly processed before reaching the consumer. While these stores can, and often do, offer organic and local products, there are a few caveats. For one, the majority of food stock at these retailers is processed, meaning that while fresh and nutritious food sections exist, they don't make up a significant portion of what is offered for purchase. Furthermore, organic and fresh foods have large price premiums, as seen here: [4]. For an individual enrolled in SNAP, maximizing the amount of food purchased given the fixed benefits is a goal. Therefore, this individual is automatically further incentivized to

purchased processed foods, which come at a discount compared to organic foods. Overall, we believe studying the food retailer categories listed above is crucial for understanding whether SNAP-enrolled individuals are made more vulnerable to purchasing processed foods.

To understand the growth and movement of the SNAP-eligible food retail segment, we divide the number of locations by the estimated population of the U.S. This allows us to control for the expansion in retail locations simply due to growing populations. The data is summarized in Figure 2.2.7.

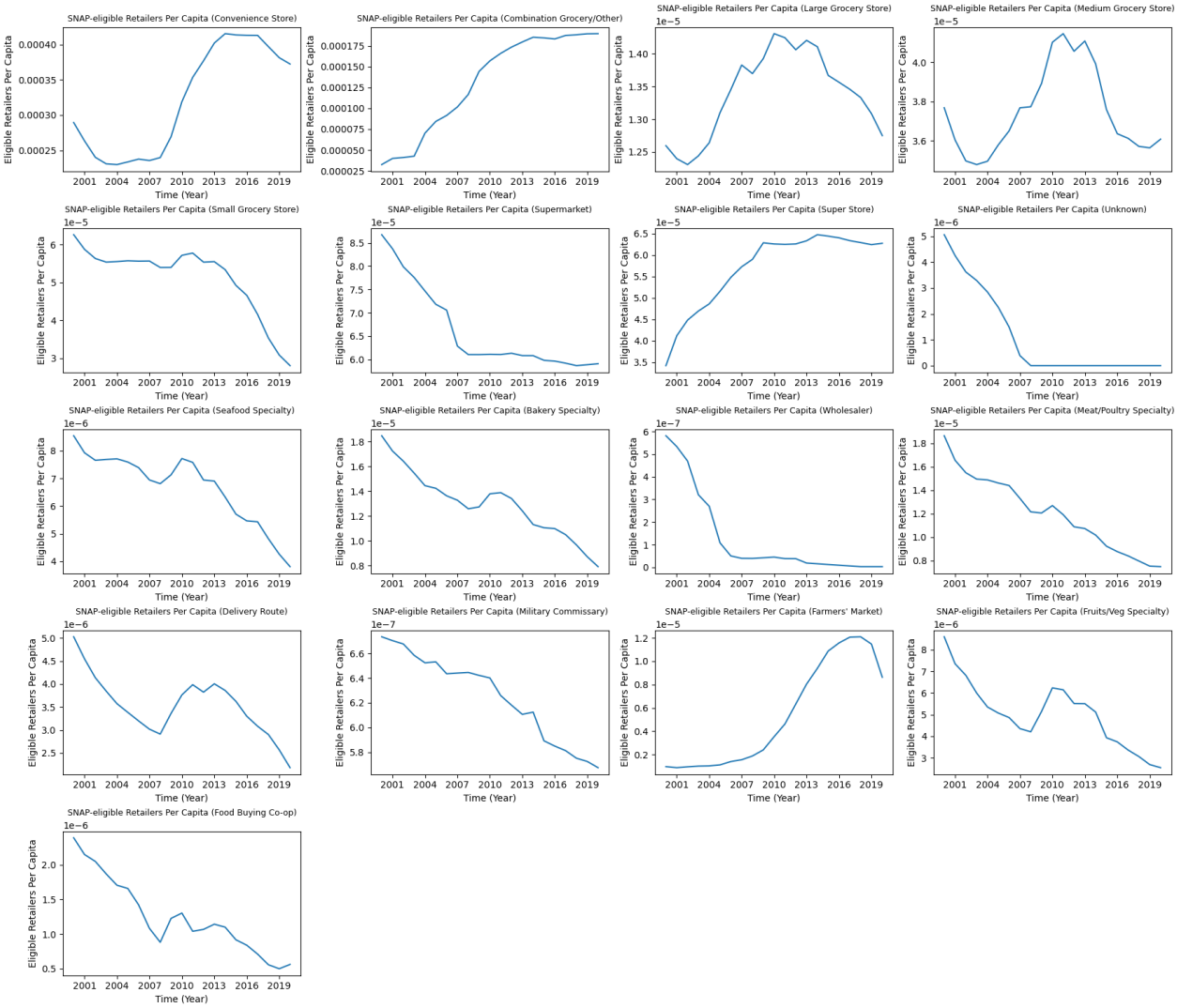

Figure 2.2.7: Prevalence of SNAP-eligible establishments by Retailer Type

From 2000 to 2020, the number of SNAP-eligible Super Stores and Convenience Stores has nearly doubled while the number of SNAP-eligible Combination Grocery stores per capita has risen by 7-fold. Meanwhile, the number of SNAP-eligible grocery stores and fresh-food oriented establishments per capita have fallen significantly. One may also notice that the SNAP-eligible Farmers' Markets, almost always a source of fresh foods and nutritious items, has also risen significantly in this time period. However, we attribute this to the rise of

online and electronic banking in areas without official storefronts, allowing consumers to use EBT cards where once, only using cash was possible. Overall, note that while types of retail stores are expected to rise and fall in number over time as populations expand, grow, and innovate, we have largely controlled for that. The USDA is allowing the density of SNAP-eligible establishments carrying highly processed food to rise. For consumers enrolled in the SNAP program, their food shopping options become increasingly oriented towards big-box retailers and chains carrying highly processed food, as opposed to places with fresher and more nutritious options.

2.3 Comparative Modeling of Drivers with LASSO

To properly select the most important variables as predictors of our synthetic ETF pricing, we perform a LASSO regression mapping our entire collection of data to the target prices. We also selected the LASSO due to its relative ability to handle potentially multicollinear/correlated data (as may be observed between our buckets of predictors) and generate useful insights about the impact each of our features may have on the predictive model. As desired, the model shrinks most coefficients to 0, leaving 8 nonzero coefficients for Agriculture & Machinery, 7 for Food & Beverages, 16 for Restaurants & Fast Food, and 7 for Retail. The common selected features observed between the target variable are types of meat production, with commodity pricing (namely coffee for restaurants and corn for agriculture and retail) also playing a major role. Comparisons of relevant coefficients and directions for each target variable are shown in Figure 2.3.1.

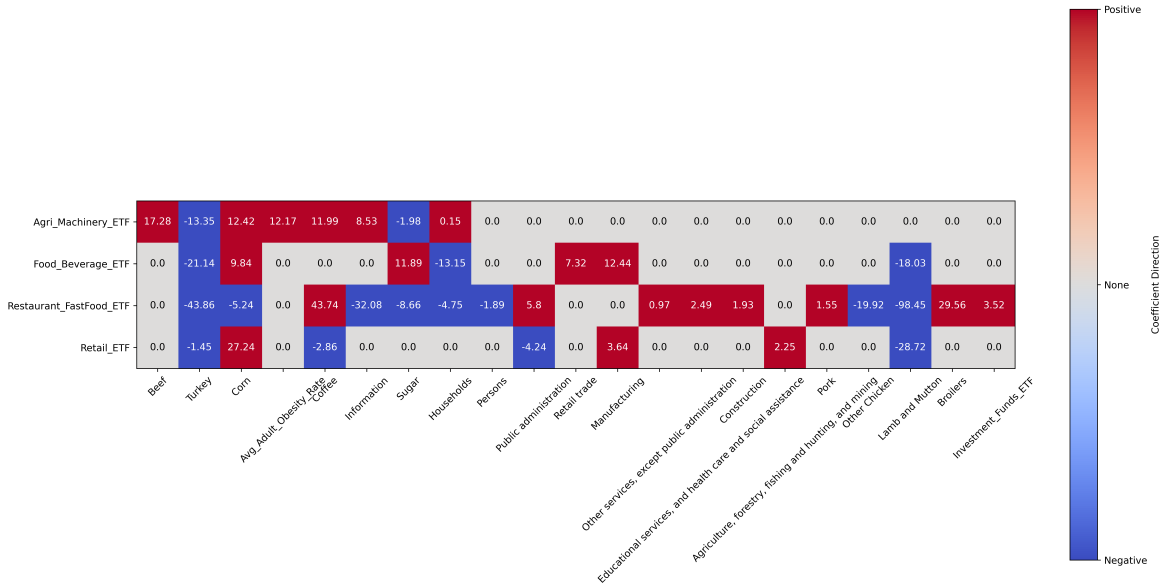

Figure 2.3.1: Heatmap of Variable Coefficients in LASSO

Given these results, we aim to show that **meat production flows and commodity prices can be adequate predictors for our synthetic market ETFs**.

2.4 Are they adequate predictors? More fine-grained testing

2.4.1 Linking Meat Production and Commodity Prices to ETF Trends

As our Potpourri LASSO Regression model suggests, meat production—particularly animal type—and commodity prices seem to have a connection with the performance of the synthetic ETFs.

To provide evidence in support of this theory, we consolidated data on monthly average commodity prices and monthly meat production yields, merging these datasets with monthly ETF prices. We then applied StandardScaler to ensure uniformity in data scaling, with the goal of fitting a Random Forest Regressor to determine feature importance. However, as identified during EDA, meat production—when disaggregated by animal type—exhibits significant multicollinearity. Additionally, there is a high degree of multicollinearity within the commodity and ETF price datasets. **This inflates the variance of the coefficient estimates, making it difficult to determine the true importance of each feature.**

Therefore, we employed Principal Component Analysis to transform the original features into orthogonal principal components, reducing dimensionality and ensuring no correlation between the components. As seen from Figure 2.4.1, the first principal component (PC1) alone captures approximately 50% of the total variance in the data. The second principal component (PC2) captures an additional 20%, and together, the first three components alone can capture around 80% of the variance.

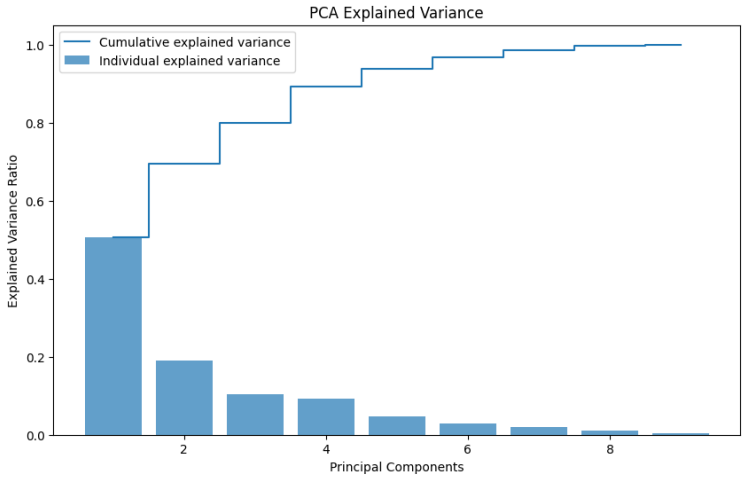

Figure 2.4.1: Explained Variance Ratios Across Principal Components

Furthermore, Figure 2.4.2 provides deeper insights into the contributions of each original feature to the principal components. The high positive loadings of Pork and Broilers on PC1 suggest that this component predominantly represents variations in meat production. Conversely, the negative loading of Veal on PC1 indicates an inverse relationship compared to Pork and Broilers. Similarly, other features like Corn, Sugar and Coffee show significant loadings on multiple components, highlighting their substantial impact on the dataset's variability.

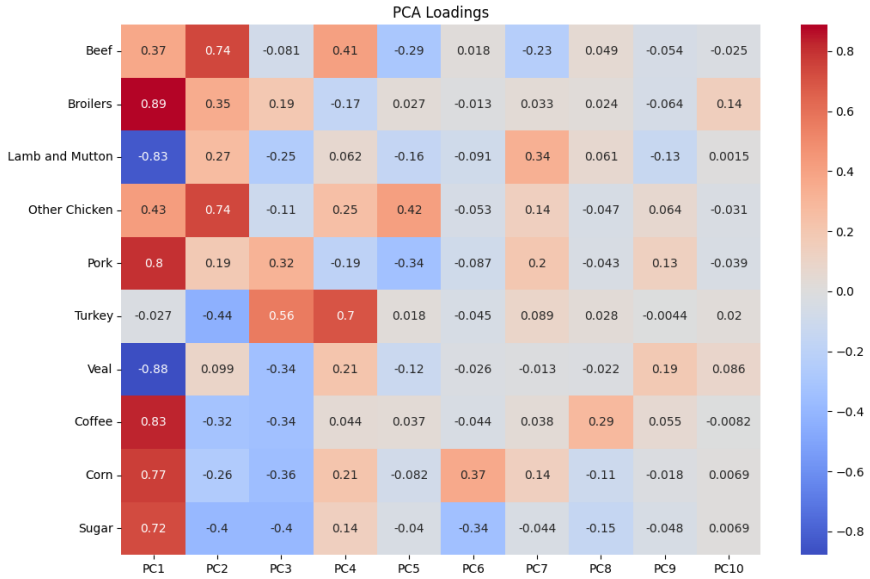

Figure 2.4.2: PCA Loadings of Various Meat and Commodity Prices

We proceed to fit Random Forest Regression models to evaluate the impacts of each principle component on the performance of the Synthetic ETFs. As shown in Figure 2.4.3, across all the synthetic ETFs (Agri Machinery, Food & Beverage, Restaurant & Fast Food, and Retail ETFs), **PC1 emerged as the most significant predictor, accounting for over 90% of the feature importance of every ETF.**

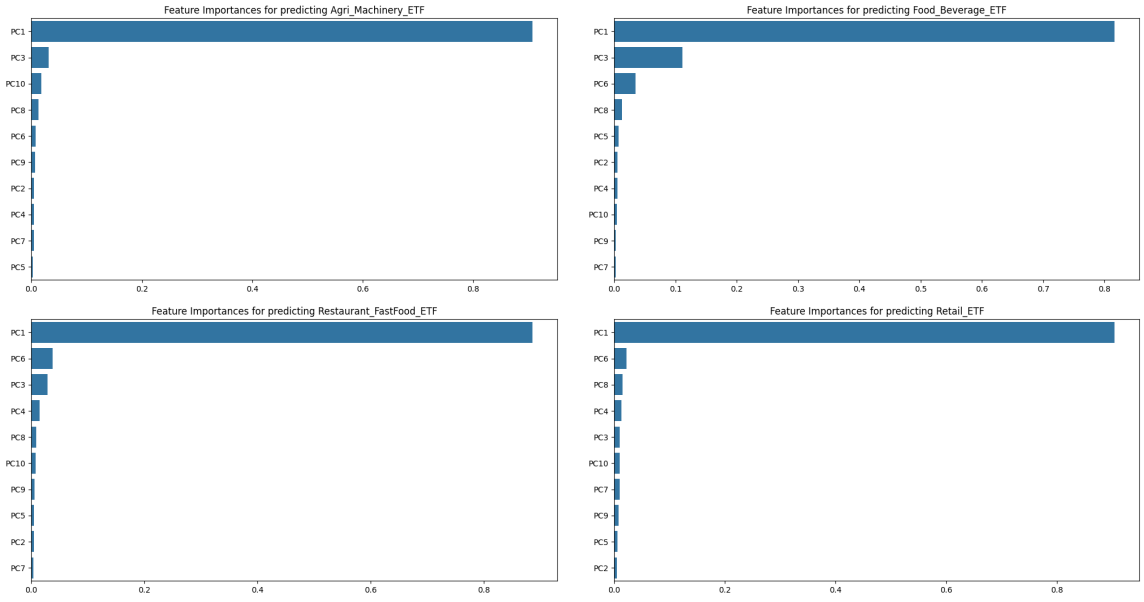

Figure 2.4.3: Random Forest Regression Results for ETFs: Agriculture & Machinery (top left), Food & Beverage (top right), Restaurants & Fast Food (bottom left), Retail (bottom right)

Further analysis on the loadings shown in Figure 2.4.4 indicate that **broilers and pork are strong positive drivers whereas lamb and veal are negative drivers across all ETFs.** We speculate that high national demand and production efficiency make broilers and pork economically significant, contributing to stable and profitable markets that positively influence performance. On the contrary, niche market demand, higher production costs, and ethical concerns may lead to less market stability and profitability for lamb and veal. On the commodities side, **loadings for coffee, corn, and sugar prices are also strong positive drivers**. We speculate that this can be attributed to coffee's widespread consumption and corn being essential in food production and animal feed.

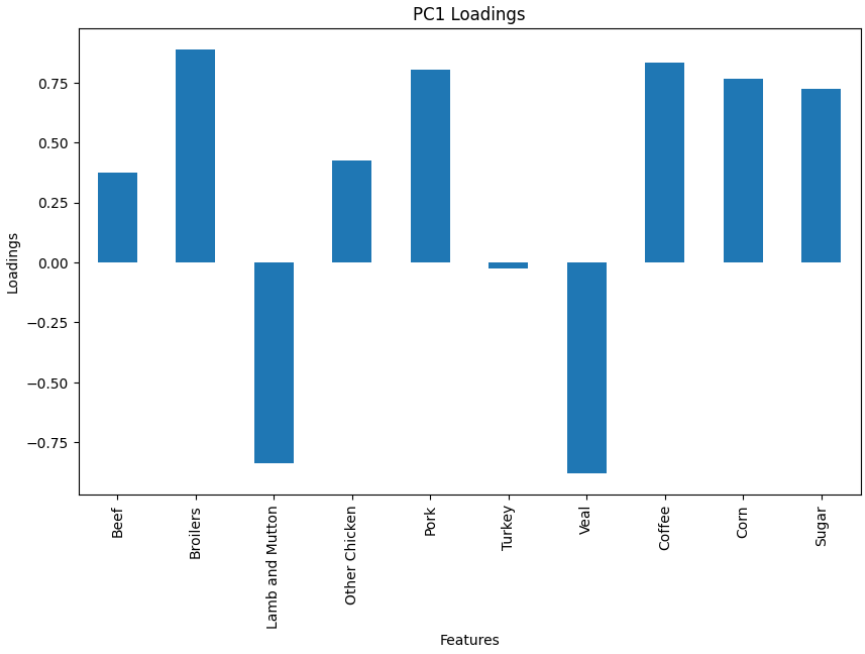

Figure 2.4.4: Loadings of PC1 for Meat Production & Commodities

However, a potential weakness of Random Forest is its lack of ability to capture historical influences. In other words, by building each tree independently, the model is unable to account for past data when handling current outcomes. Thus, the model may not be suitable for determining how previous values of a feature influence future values.

2.4.2 Forecasting ETF Trends with XGBoost

To better understand how past values of Meat Production and Commodity Prices influence the performance of the processed food economy's production flow, we design a robust forecasting model with XGBoost tailored for time series data.

We first preprocess the dataset utilized to train the Random Forest Regression, using RobustScaler to scale features by centering on the median and scaling to the IQR. This reduces outlier influence as we anticipate events like COVID-19 to impact our results. We

additionally create lagged features corresponding to 1-month, 2-month, and 3-month intervals for each predictor variable to capture local trends. We further integrate a custom walk-forward validation strategy, which handles scenarios where future information may influence training. We begin by splitting into training and test sets with a training window of 36 months and a testing window of 12 months. This process repeats, moving the training window forward by 12 months each time, ensuring the model is always validated on future data. As shown in 2.4.5, our model achieves high R-squared values and reasonable RMSE across the ETFs given the corresponding growths of each ETF over time.

Category	RMSE	R-squared
Agri_Machinery_ETF	27.08	0.778
Food_Beverage_ETF	27.065	0.809
Restaurant_FastFood_ETF	119.61	0.83
Retail_ETF	22.585	0.8224

Figure 2.4.5: RMSE and R-Squared for Each ETF

We additionally display model predictions against actual ETF values over time in Figure 2.4.6. Given the limited temporal resolution of our dataset (only 276 time points could be aligned during data merging), our model reasonably tracks actual prices.

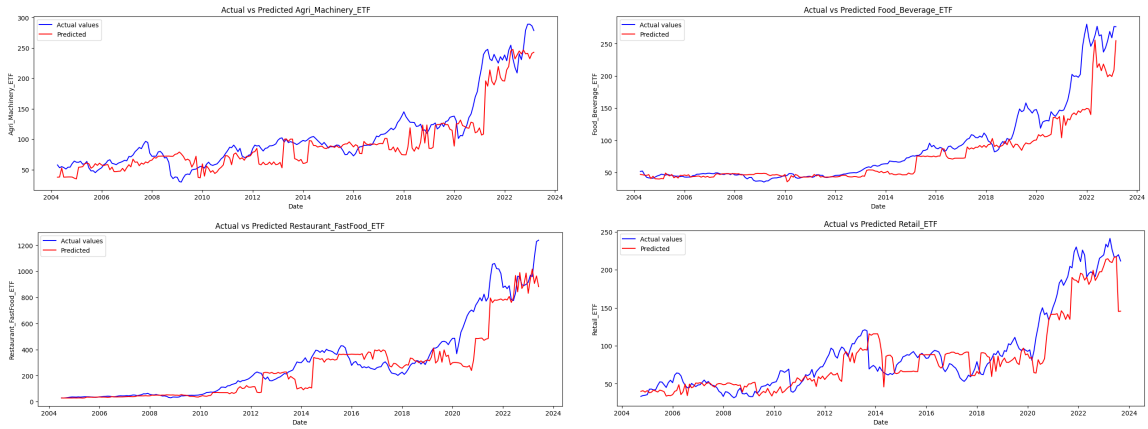

Figure 2.4.6: Predicted vs. actual average price for ETFs: Agriculture & Machinery (top left), Food & Beverage (top right), Restaurants & Fast Food (bottom left), Retail (bottom right)

However, as shown by 2.4.7, anomalous dates were disproportionately found within the COVID-19 Pandemic era, across the months of 2020-2022. We speculate that unprecedented market volatility followed by rapid recoveries may have lead to erratic spikes in ETF values. Sector-based impacts may have also played a role, with Restaurant and Fast Food ETFs initially declining due to lockdowns but rebounding strongly as consumer behavior adapted. Furthermore, global supply chain disruptions caused delays and shortages, which may have affected sectors like Agriculture, Machinery and Retail, which then saw increased demand.

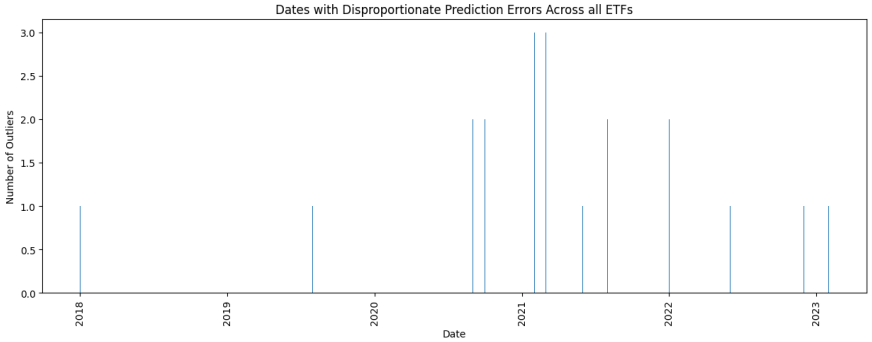

Figure 2.4.7: Dates with Disproportionate Prediction Errors by 2 Standard Deviations

As an extension, we extracted feature importance scores from the trained XGBoost model to better determine the contribution of each lagged feature to the ETFs. Figure 2.4.8 depicts the feature importance scores of each feature lag on each individual ETF.

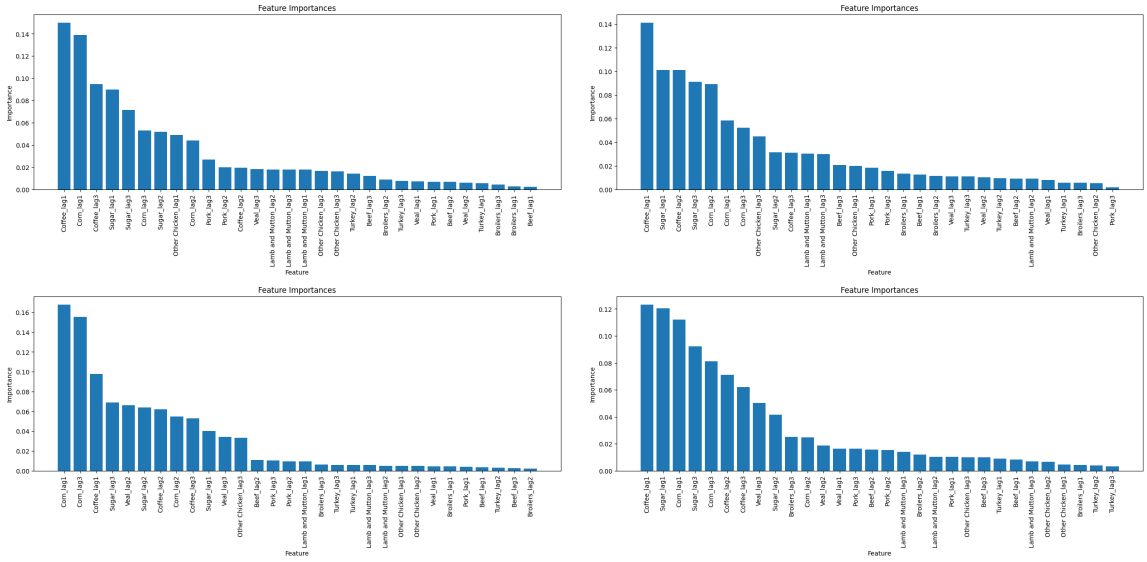

Figure 2.4.8: XGBoost Feature Lag Results for ETFs: Agriculture & Machinery (top left), Food & Beverage (top right), Restaurants & Fast Food (bottom left), Retail (bottom right)

We derive three key insights:

1. **Coffee, Corn and Sugar Volatility Outshine Meat Volatility**

   Meat production features, such as Veal Lag 2 and Other Chicken Lag 3, exert moderate impact on various ETFs like Restaurant & Fast Foods, indicating that meat volatility does have an effect on the overall processed food economy's production flow. However, Coffee, Corn, and Sugar exhibit the highest feature importance scores across all ETFs. For instance, in the Agriculture and Machinery ETFs, Coffee lag1, Corn lag1, and Sugar lag1 hold importance values of 0.149956, 0.139075, and 0.089906 respectively, significantly higher than any Meat-related features. Similarly, for the Retail ETFs,

Coffee lag1, Sugar lag1, and Corn lag1 score 0.123154, 0.120431, and 0.112199. The data indicates that **fluctuations in the commodities may have a significant effect throughout the supply entire chain**, affecting factors like procurement and production across diverse sectors of the economy.

2. **Corn Prices are the Leading Predictors for Restaurant & Fast Food Sectors**

   The notable importance of lagged corn prices within the Restaurant & Fast Food ETF suggests that corn prices serve as key indicators for this sector. This is evident from the high feature importance values for these lagged variables, with Corn Lag 1 and Corn Lag 3 achieving the highest importance scores of 0.167760 and 0.164904, respectively. **Thus, we posit that the impact of corn price changes can be felt almost immediately, and these effects can persist for several months.** Consequently, we speculate that fluctuations in corn prices can directly influence the cost of a wide range of menu items and key supply chain components in the restaurant and food service industry.

3. **Coffee Prices are Immediate Indicators for Multiple Sectors**

   Coffee Lag 1 achieves the highest importance score across Retail ETF, Food & Beverage ETF, and Agriculture and Machinery ETF. **Thus, we hypothesize that coffee prices from the previous month have a significant and immediate impact on multiple sectors of the processed food economy**, as fluctuations in coffee prices can quickly influence consumer spending patterns, production costs, and overall profitability in retail, food and beverage, and agricultural industries. The immediate impact likely stems from coffee's role as a widely consumed commodity and its integration into various products and supply chains

2.5 Relevance of the Food Economy

Through our analysis thus far, we have strove to answer our first research question and show that meat production and commodity pricing can come together to adequately predict the performance of the US food economy. In this section, we will explore how our food market indicators (the synthetic ETFs) can be used to predict overall economic and community health indicators.

2.5.1 Relation to the broader US Economic State

Predictions for Index Funds by Synthetic ETFs - In our previous discussion of designing synthetic ETF clusters, we left out the buckets of market index as an initial target, due to the much broader and more volatile nature of the US economy at large and our desired scope of just the food industry. However, given the pivotal role food plays in the US economy and our stated second goal, we sought to evaluate how well our synthetic ETFs of the food economy could predict overall US market indicators. Using a similar XGBoost approach, the results and feature importances are shown in Figure 2.5.1.

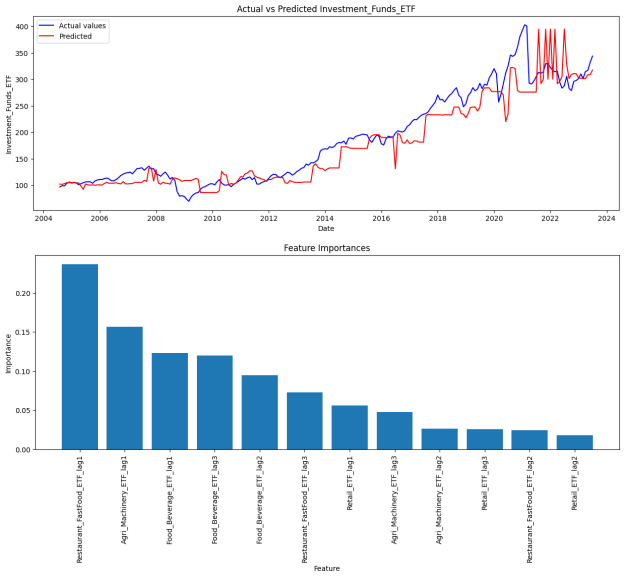

Figure 2.5.1: Index fund predictions (top) & relative feature importance (bottom)

Our model performs well in predicting historical index fund performance ($R^2 = 0.867$), although the sudden drop following the COVID-19 pandemic appear to temporarily cause issues for the forecasting methods, which we hope to smooth out with more performance information in the coming years.

Predictions for Unemployment Rates by Synthetic ETFs - Expanding from merely looking at stock & fund prices as economic indicators, we also sought to explore the impact changes in synthetic ETF prices display on national unemployment rates, a key metric for the health of the US economy at a personal level [2]. Again harnessing the XGBoost approach to examine monthly aggregated unemployment rate across all national industries, this model appears to perform significantly worse than others, with $R^2 = -0.2215$. Actual predictions and feature trends are shown in figure 2.5.2.

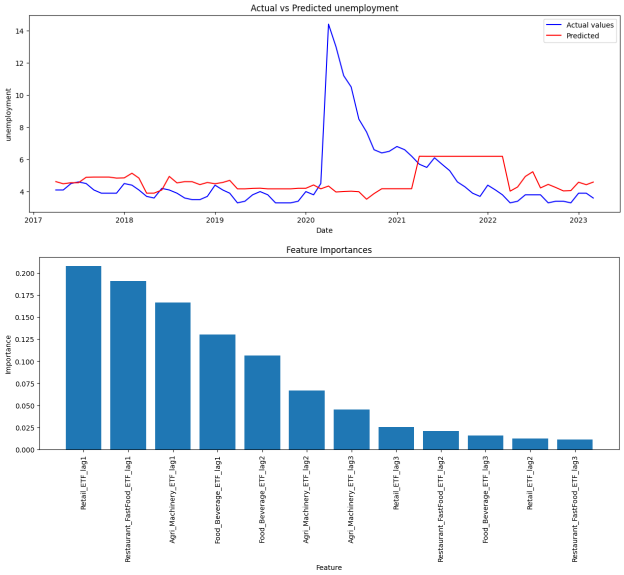

Figure 2.5.2: Unemployment - Aggregated (top), Feature importance (bottom)

Looking more closely at the trends over time shows the abysmal numerical result is primarily the result of the model's inability to predict the spike in unemployment as a result of COVID-19, especially without similar shocks seen in the synthetic ETFs, over the limited timescale provided in the data. Otherwise, the XGBoost model appears to be reasonably good at predicting overall unemployment numbers for other years. Expanding the timescale to include other potential anomalous events of note in future analyses may allow this model system to recognize economic indicators that can be linked to such unusual events that are linked to swings in unemployment.

We also trained the model to predict month-by-month unemployment rates for each of four relevant economic sectors: Retail, Manufacturing of Nondurable Goods (relating to foods & beverages), Leisure & Hospitality, and Agriculture-related Work. Predictions are shown in Figure 2.5.3.

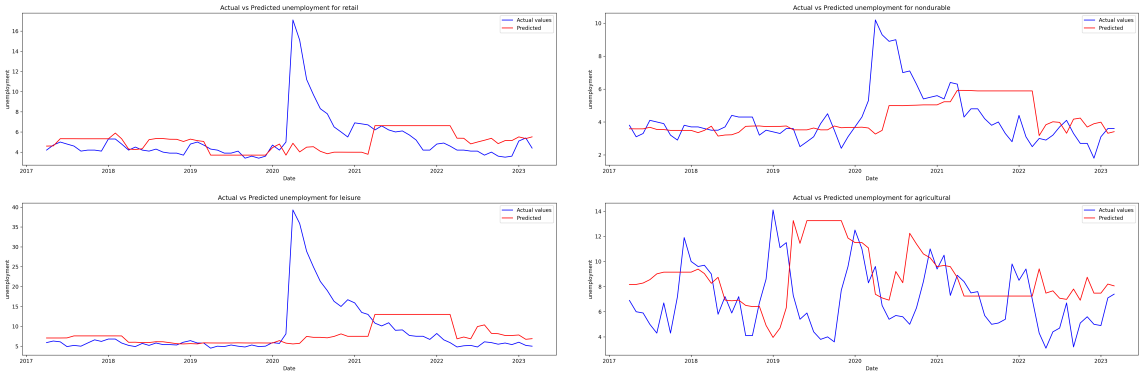

Figure 2.5.3: Unemployment model predictions for retail (top left), manufacturing (top right), hospitality (bottom left), and agriculture (bottom right)

These sector-specific models perform similarly poorly: **Retail** - $R^2 = -0.240$, **Manufacturing (Nondurable Goods)** - $R^2 = -0.060$, **Leisure & Hospitality** - $R^2 = -0.161$, **Agriculture** - $R^2 = -1.492$, matching trends seen in aggregated data. One deviation of note occurs in the agriculture sector, which, even without a similarly pronounced unemployment shock during COVID-19, displayed seasonal cycles that the model was unable to predict. For reference, the feature importance lists for each of our targets are also shown in Figure 2.5.4.

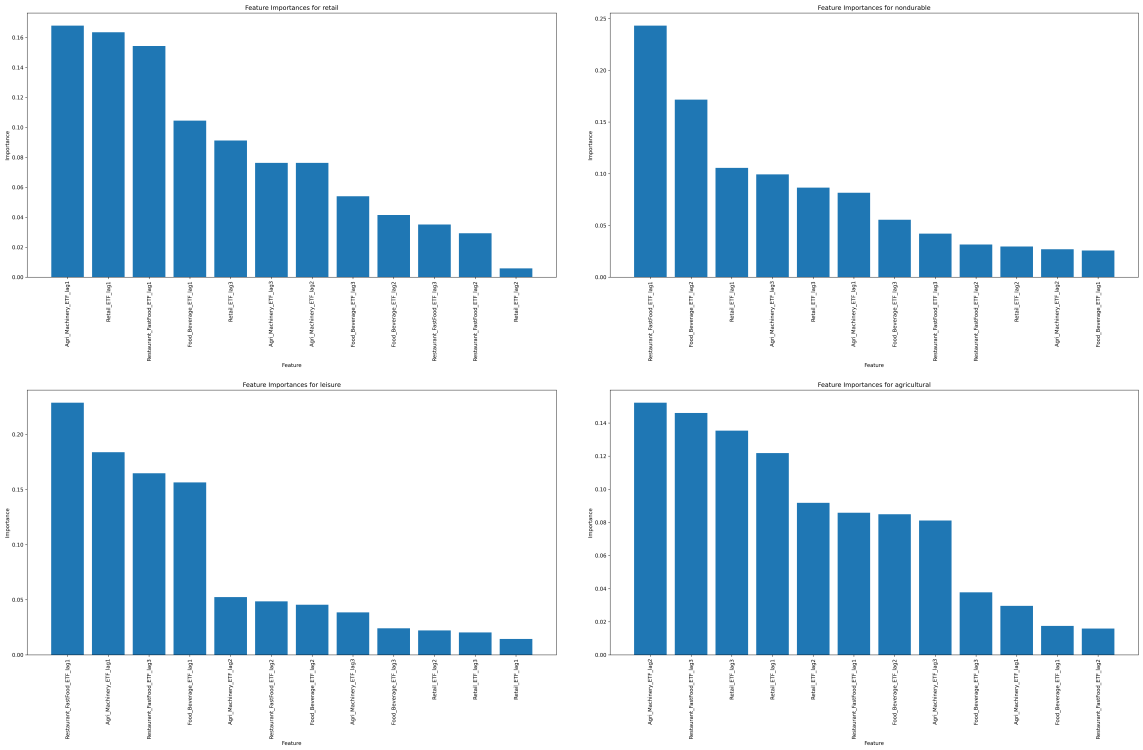

Figure 2.5.4: Feature importance analyses for retail (top left), manufacturing (top right), hospitality (bottom left), and agriculture (bottom right)

In conclusion, while our synthetic ETFs can reasonably predict index fund performance and baseline unemployment rates, they are currently unable to account for the major economic shocks that arguably impact people's livelihoods the most. Additional explorations and data will reveal more insights about which metrics could reasonably follow external market swings and predict the path of the economy as it stabilizes.

2.5.2 Quantifying Impact on Health Outcomes

We perform further data analysis on the combined ETF prices and the obesity rate within each ethnicity group (Two or More Races, American Indian, Asian, Hawaiian, Hispanic, non-Hispanic Black, non-Hispanic White). From the preliminary EDA with raw ETF prices, it turns out that each ethnicity group's obesity rate roughly aligns with the growth of ETF, despite some minor fluctuation, see Figure 2.5.5.

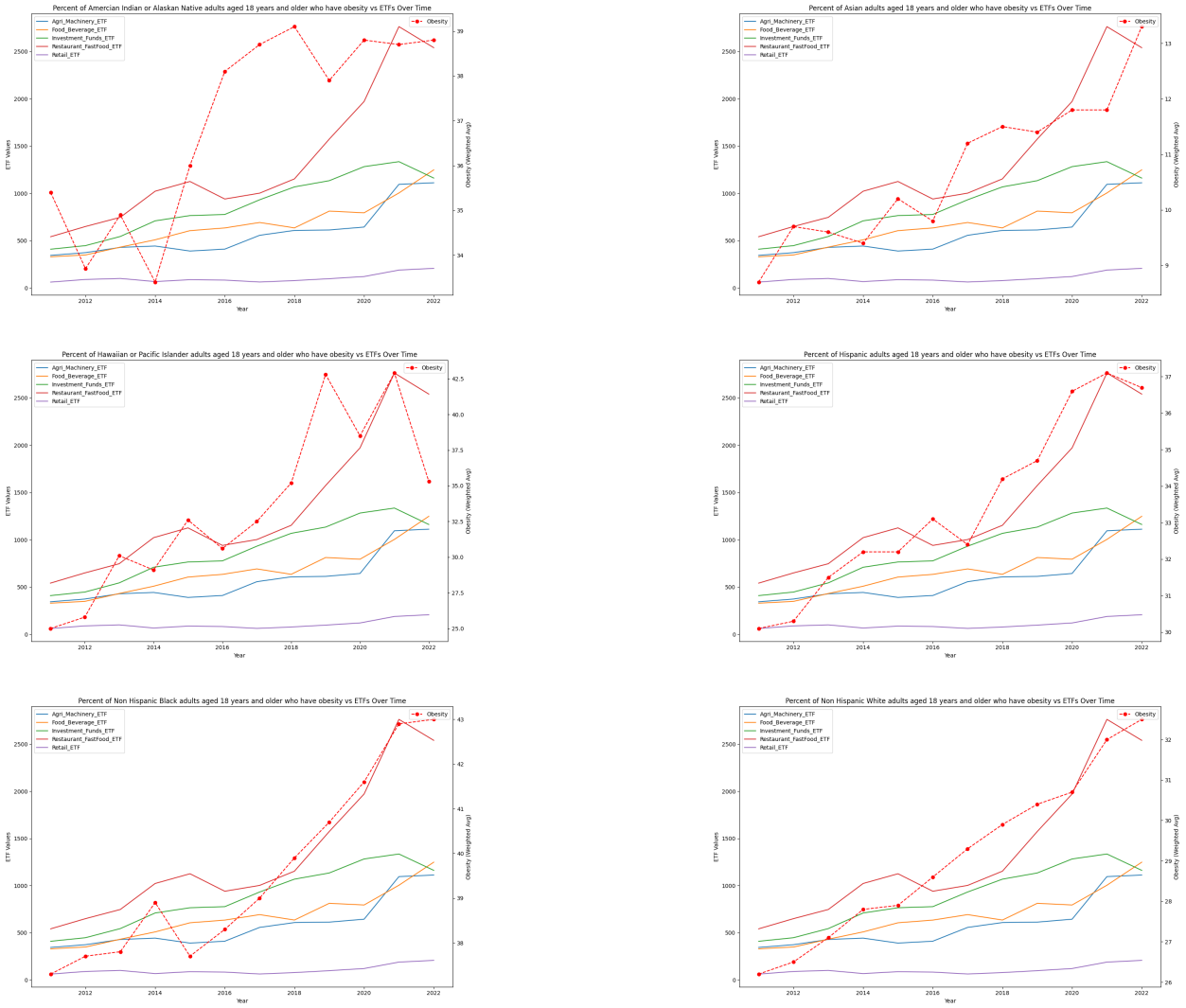

Figure 2.5.5: ETF Movements with Different Races and Ethnicity groups

This provides us with the intuition of combining the obesity data together by taking the weighted summation of each group with weights being their population. Figure 2.5.6 exhibits the visualization of the relationship between the normalized ETF prices and the weighted obesity rate.

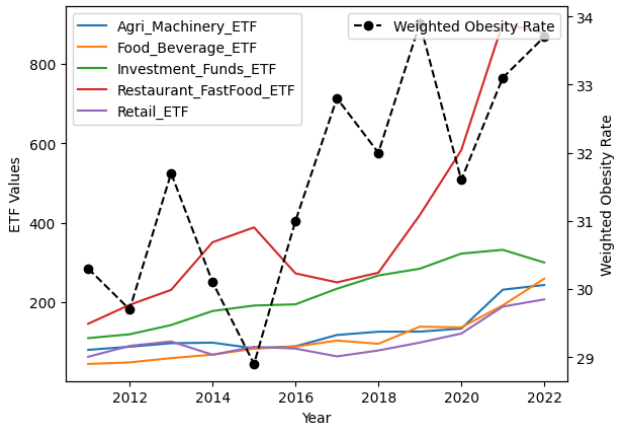

Figure 2.5.6: Various ETFs and Weighted Obesity Rate

From the visualizations of the trends shown in Figure 2.5.6, it can be seen that different features exhibit some rough linear relationship with others, which motivates us to build up a linear regression model and perform some regression analysis. For the sake of robustness and reducing the multicollinearity of the model, we computed the correlation matrix as shown in Figure 2.5.7. It can be seen that some of most of the features are high-correlated with each other, this suggests us some of the features will be redundant and detrimental when building up a linear regression model, since multicollinearity can make the regression coefficients unstable and inflate the standard errors, which can lead to unreliable statistical tests.

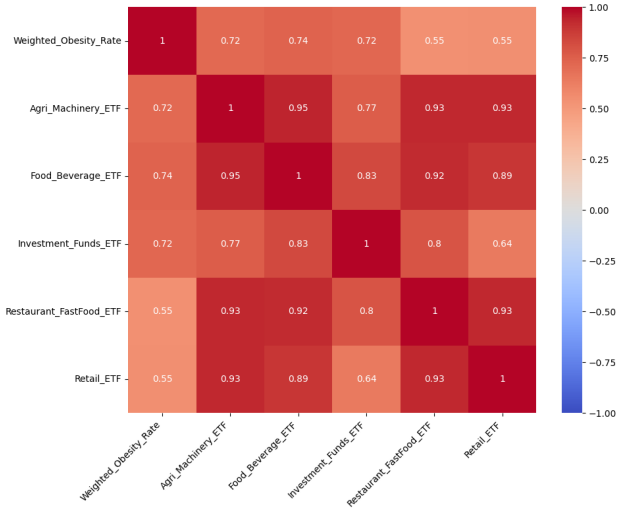

Figure 2.5.7: Correlation Matrix of Various ETFs and Weighted Obesity Rate

To determine whether to include a specific feature into the model or not, deeper statistical analysis must be conducted. We made some further discovery by plugging the features into a linear regression model with `Obesity_Rate` as the response variable and all the other ETFs as features (after adding a constant term). The summary of the model is shown in

Figure 2.5.8, where we deployed the package `statsmodels`.

Figure 2.5.8: Regression Coefficients and Statistics

Variable	Coefficient	Std Err	t	$P > \|t\|$	[0.025	0.975]
Agri_Machinery_ETF	0.0865	0.126	0.689	0.513	-0.210	0.383
Food_Beverage_ETF	-0.1265	0.094	-1.350	0.219	-0.348	0.095
Investment_Funds_ETF	0.1307	0.031	4.246	0.004	0.058	0.203
Restaurant_FastFood_ETF	-0.0598	0.022	-2.669	0.032	-0.113	-0.007
Retail_ETF	0.2789	0.120	2.234	0.052	-0.004	0.562

We set up a threshold for P values as $P = 0.05$, in other words, we identify a feature as statistically significant if its P value is less than 0.05 and vice versa. For example, `Investment_Funds_ETF` and `Restaurant_FastFood_ETF` both have P values less than 0.05, which means that they have a statistically significant impact on the dependent variable `Obesity_Rate`.

2.6 Strengths & weaknesses of our approach

The central model behind our approach, following careful variable selection, is a series of XGBoost models to adequately predict the target economic indicators from our predictive factors. Here are some key strengths and weaknesses behind our methodology:

Strength 1: Look Ahead Validation Strategy handles scenarios where future information may influence training ensuring the model is validated on future.

Strenght 2: Built-in L1 and L2 regularization to help prevent overfitting and handle colinearity

Weakness 1: Hyperparameter Sensitivity: Demands careful tuning of numerous hyperparameters, which is not good because we have limited time to tune.

Weakness 2: XGBoost may not perform optimally with small datasets due to its complexity and the risk of overfitting.

3 Concluding Thoughts

3.1 Summary of Results

Our LASSO regression models suggest that across the four synthetic ETFs, the best variables for predicting ETF prices are meat production volume and the prices of key commodities (namely coffee and corn).

Afterward, we aimed to show that these features selected could be used to design an adequate model of our synthetic ETF pricing. Using extreme gradient-boosted regression and careful consideration of data, we were able to construct effective predictors for our four prices, with each having $R^2 > 0.77$ and two being above 0.8.

Lastly, we demonstrated links between the prices of our synthetic ETFs and market performance, unemployment rates, and obesity rates, constructing appropriate models to predict each under the general case (although anomalous events such as COVID-19 can significantly reduce performance, as seen in unemployment rates). This applied analysis will be the start of further exploration of how health and market movements are intertwined, as well as how they can be improved synergistically.

3.2 Future Directions

Our initial model structures provide a broad baseline by which other societal indicators can predict economic performance in the food industry. Here are two potential directions for exploration to better align these predictions with the true state of US food and potentially apply the results.

- **Regional-Level Study** - Food production and processing is an especially heterogenous industry in the United States. Understanding food economy metrics past national-level stock prices to localized supply chains and consumption can lead to a more nuanced understanding of the regional centers driving national industry changes.
- **Health Explorations** - Supplementing our current analyses with additional time-series data on more varied community health metrics will allow us to better apply our models and truly examine the downstream meaning of food economy changes.

Categorizing these potential causal chains will provide targets for future policy promoting positive community outcomes while ensuring stability in the constituent industries.

4 Appendix

References

[1] U.S. Department of Agriculture. *SNAP*. https://www.fns.usda.gov/snap/supplemental-nutrition-assistance-program. Accessed: 2024-08-04. 2024.

[2] U.S. Bureau of Labor Statistics. *Table A-14. Unemployed persons by industry and class of worker, not seasonally adjusted*. Accessed: 2024-08-04. 2024. URL: https://www.bls.gov/webapps/legacy/cpsatab14.htm.

[3] Federal Reserve Bank of St. Louis. *MRTSSM4451USS: Retail Sales: Food Services and Drinking Places*. Accessed: 2024-08-04. 2024. URL: https://fred.stlouisfed.org/series/MRTSSM4451USS.

[4] Economic Research Service U.S. Department of Agriculture. *Investigating Retail Price Premiums for Organic Foods*. Accessed: 2024-08-04. 2024. URL: https://www.ers.usda.gov/amber-waves/2016/may/investigating-retail-price-premiums-for-organic-foods/.

[5] Food U.S. Department of Agriculture and Nutrition Service. *SNAP Retailer Historical Data*. Accessed: 2024-08-04. 2024. URL: https://www.fns.usda.gov/snap/retailer/historical-data.

[6] Food U.S. Department of Agriculture and Nutrition Service. *SNAP Store Definitions*. Accessed: 2024-08-04. 2024. URL: https://www.fns.usda.gov/snap/store-definitions.