Spring
Asia Datathon

Team 1

2024

Prepared by :

Lam Yat Tung Anson
Ng Tsz Hei Jadon
Wang Issam

Non-Technical Executive Summary

Main Questions

Two key questions are being addressed in this study,

1. **Is there any linear relationship between meat production and the unemployment rate? If there is, is meat production a leading indicator or a lagging indicator of the unemployment rate?**
2. **Are meat production yields predictive of restaurant stock prices?**

Key Findings

We discovered that there is a strong linear relationship between meat production and the unemployment rate. In particular, an inverse relationship is demonstrated where a higher meat production number corresponds to a lower unemployment rate. Upon further investigation, we can distinguish the difference in the strength of the relationship between red meat and poultry, where poultry shows a much more significant relationship when compared to red meat. We are also able to discover that meat production is more of a lagging indicator of the unemployment rate during our analysis horizon, meaning that a drop in the unemployment rate today will cause meat production to drop in the future, 12 months into the future to be precise. This relationship is particularly true for red meat which upon more in-depth investigations, this 12-month period also happens to be the average production cycle of beef (red meat) (Walter, 2013). We strongly believe that this relationship is not a coincidence and deserves future research and investigation which is out of the scope of this study. Another key finding in this study is that 2 extreme events were identified for red meat during the 2010-2022 period, which is the surge of corn prices in 2010-2014 and the global pandemic in 2020, while the latter event is also identified as an extreme event for poultry. The key difference between the results of poultry and red meat also reveals the impact of livestock feed, particularly corn in this example, on the two types of meat, demonstrating a key distinction and justification to conduct tests on the two types of meat separately, where we can identify the phenomenon that the surge in corn price lower the red meat production in the 2010-2014 period.

These conclusions drive us to ask the next key question - we discovered the impact of the unemployment rate on meat production yields, how about the impact of meat production on the economy? In particular, the question we attempt to answer is - Are meat production yields predictive of restaurant stock prices?

To answer the problem, we leveraged machine learning models to predict stock prices based on historical stock price data, meat production data, and commodity data. The analysis aimed to enhance predictive models by incorporating features derived from meat production statistics. Features derived from these figures demonstrated improved accuracy, particularly in the technology portfolio, indicating the relevance of certain features in predicting stock prices. However, the inclusion of seasonality indices did not significantly contribute to the models' performance. A more in-depth analysis of the importance of meat production from different animals and the value of commodities was conducted to predict stock prices for the 7 selected companies. The model was only able to outperform the baseline in one of the seven firms analyzed. The findings from this report suggest that specific features play a crucial role in predicting stock prices for different portfolios or individual securities, warranting further investigation.

Technical Exposition

Data Preprocessing and EDA

EDA is first conducted on 2 key label groups from *economic_characteristics_2010-2022.csv* that are thought to be the 2 major areas to gain insight for a macro overview of the US economy and industrial structure from 2010 to 2022. In particular, the unemployment rate and industry group are being investigated for this purpose. Below are some key summaries discovered through the individual analysis of each grouping.

Unemployment rate

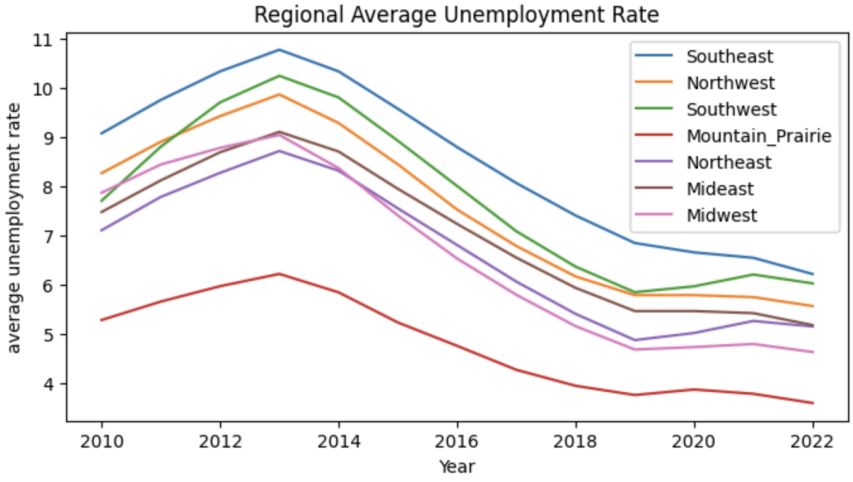
Yearly unemployment data of each state in the United States from 2010 to 2022 is being investigated in this section. It is important to note that we have concatenated the ‘Percent Unemployed’ group, as the unemployment rate data from 2010-2014, and the ‘Unemployment Rate’ label group as the unemployment rate from 2015-2022. Some key summaries are drawn below.

- Puerto Rico has the highest average unemployment rate of 16.76% from 2010-2022, followed by Nevada with 9.02%
- North Dakota has the lowest average unemployment rate of 3.08% from 2010-2022, followed by South Dakota with 4.09%
- The unemployment rate went up from 2010 to 2013 and continuously declined afterward

Another key preprocessing step we took is that we bucket states into their corresponding region, namely *Midwest, Mideast, Mountain Prairie, Northeast, Northwest, Southeast, and Southwest* as per definition (Kuhn et al., 2013). By doing so, we aim to create more understandable and low-dimensional visuals to draw general but specific enough insight to decode the relative relationships among each region. Below is the average unemployment rate of each region from 2010-2022, the mean of the respective state unemployment rate is used to determine the average unemployment rate of each region for simplicity, though it is important to note that it may not be mathematically accurate, a better solution may be to use raw population and unemployment numbers for transformations.

It is also interesting to note that the unemployment rate remains relatively stable from 2020 to 2022 in the provided dataset, which we later found to be somehow inaccurate when we conducted further analysis with external data. We believe that the yearly data from the provided dataset are the corresponding numbers collected at the end of each year, this may lead to some sort of bias especially considering there are indeed extreme cases during the analysis period, which we will discuss more in-depth in later sections.

- The Mountain Prairies region has the lowest average unemployment rate over the study horizon between the 4% to 6% range
- Southeast region has the highest average unemployment rate over the study horizon ranging from 7% to 11%

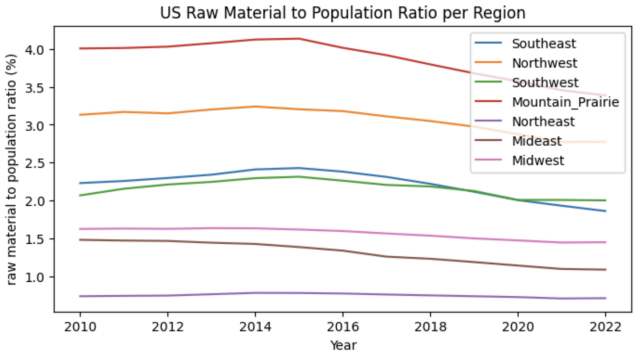


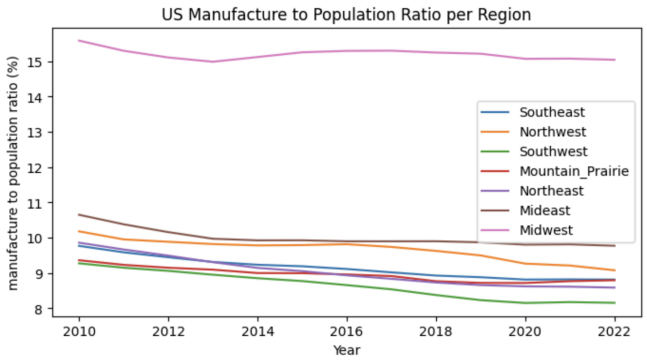
Industry

This section analyzes the estimated employment number of each industry sector per state from 2010-2021. The two major industries we will focus on will be the 'manufacturing' sector and the 'Agriculture, forestry, fishing and hunting, and mining' (which we will refer to as 'raw material' in the study) sector which we believe to be the two major areas relating to the focus of the study. One main feature engineering process that we applied to each sector is the 'sector to population ratio' which aims to quantify the relative impact of the particular sector to the state. Some key summaries are drawn below.

- California and Texas are the top 2 states for raw material sector, employing over 400000 people per state on average from 2010-2021
- Wyoming and North Dakota are the top 2 states for the highest raw material per population ratio, having a ratio of 11.65% and 9.04%
- California is again the top state for the manufacturing sector, employing over 1.6 million people on average
- Indiana, Wisconsin, Michigan, Ohio, and Iowa all have an average of over 15% of people working in the manufacturing sector

Regional transformation is again applied to the dataset, showing that Mountain Prairie has the highest raw material-to-population ratio of around 4% and Midwest has the highest and very dominant manufacturing-to-population ratio of around 15% from 2010-2021.





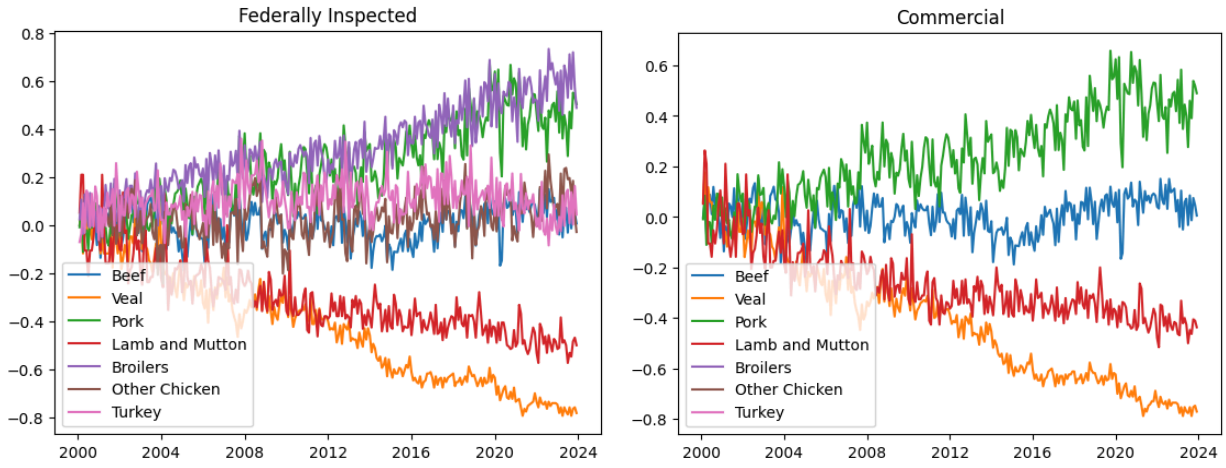
Interestingly, Mountain Prairie is also the region with the lowest unemployment rate among other regions while the Midwest is also the region with the greatest decline in unemployment rate from 2010-2021, suggesting further analysis may be possible for the two regions.

Meat Production

EDA was then performed on the *Meat_Stats_Meat_Production.csv* dataset, covering the years 2000 to 2023. The analysis aims to identify trends in meat production across different types and explore potential relationships between yields and stock prices. Below are some key findings from the analysis:

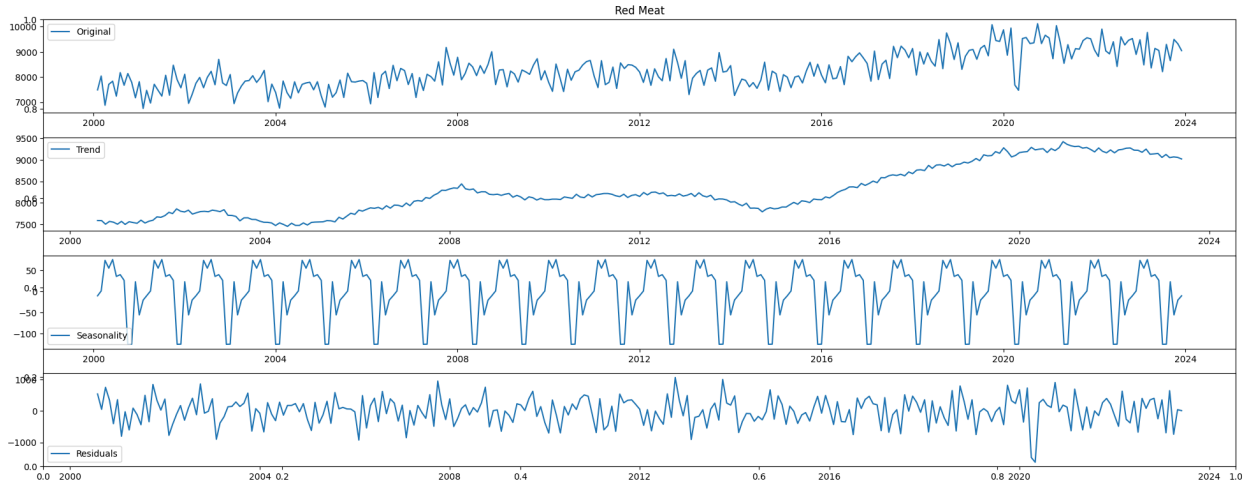
General Category: Red Meat and Poultry

The analysis revealed noticeable production trends for specific meat types. Broilers and pork meat exhibited an upward production trend, while veal, lamb, and mutton experienced a decline. Other meat types demonstrated relatively stable production levels. The below plot shows the cumulative growth in production since 2000.

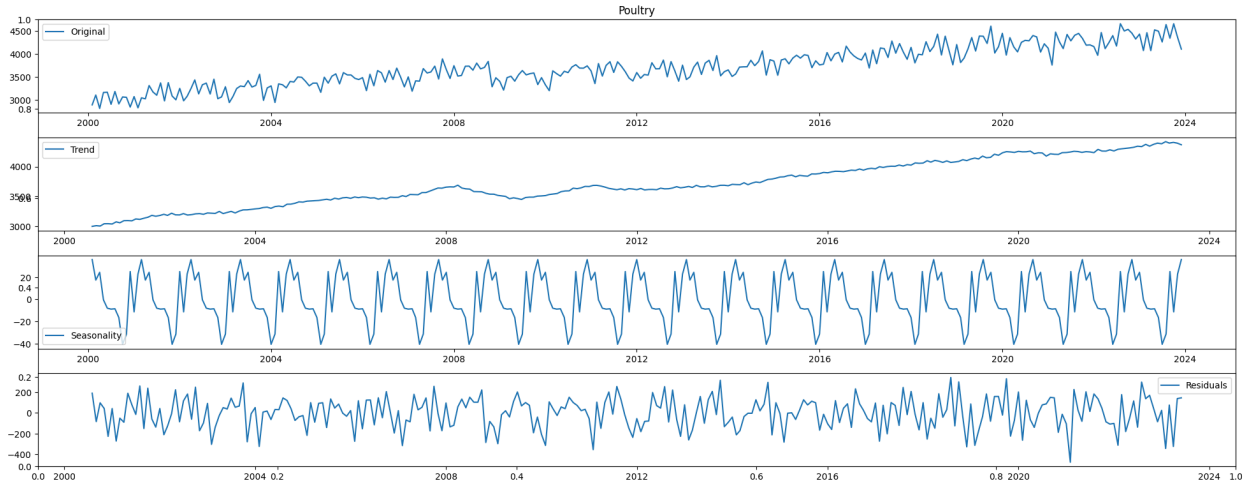


Given the observed fluctuations in meat production, a time series decomposition was conducted using an additive model assumption, assuming constant seasonality factors over time. The results indicated the presence of seasonality in the meat production data. The time series decompositions for total meat production, red meat, and poultry are shown below.

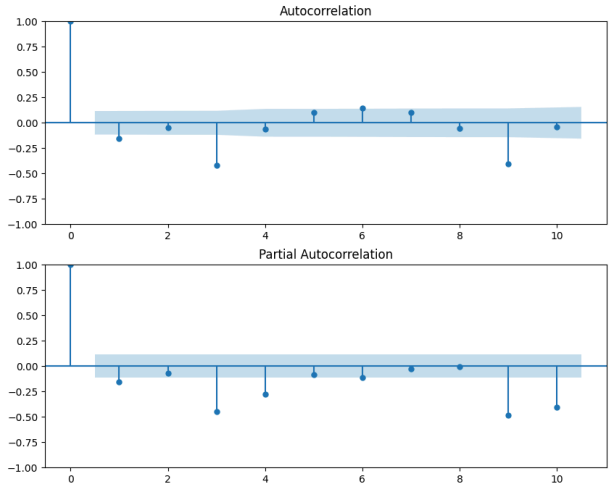
Red Meat



Poultry



The analysis revealed that poultry production exhibits a relatively stable trend compared to red meat production. Residual tests on the decomposition indicated minimal autocorrelation and partial autocorrelation in total meat production and poultry production. However, red meat production displayed statistically significant autocorrelation and partial autocorrelation, suggesting the existence of underlying patterns that warrant further investigation. These findings emphasize the need for in-depth analysis to understand the dynamics and potential factors influencing red meat production.



Are there any linear relationship between unemployment statistics and meat production statistics?

The following section highly leveraged OLS regression to investigate the linear relationship between the dependent variable, *meat production*, and the independent variable, the *unemployment rate* of each state in the United States.

Regression framework

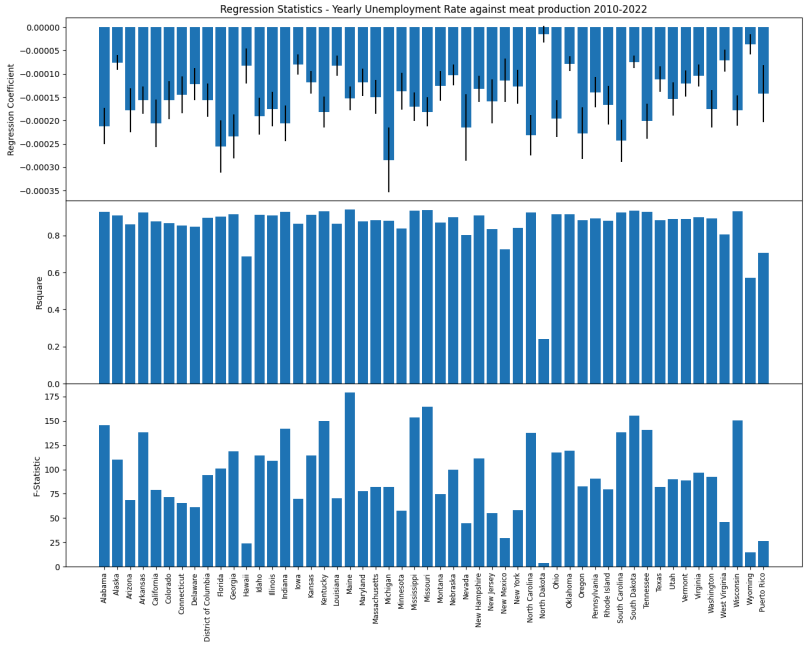
To facilitate the analysis to test for the impact of different hypothesis assumptions, a general framework is developed for consistency. Essentially, an OLS regression will be conducted for each state over the unemployment statistics of each state and the overall yearly aggregated meat production statistics, although meat production of each state will be highly valuable data for investigation which unfortunately are not included in the given dataset. Key results such as the *regression coefficient*, *95% confidence interval of the regression coefficient*, *R-squared*, *F-Statistic,* and *p-value* will then be extracted for each experiment and visualized using a barplot.

It is important to note that there are several underlying assumptions for linear regression to stand, as the study only involves one dependent variable, multivariable considerations such as multicollinearity can be omitted while other assumptions, including the normality of residual, will be asserted using metrics like *Omnibus* and *Prob(Omnibus)*. Hence unless otherwise specified, the assumptions of the following experiment results should be assumed to hold. At the same time, it is important to note that only 13 data points will be used for the regression analysis, so although it still satisfies the 10 data point minimal

requirement as suggested by numerous research, the lack of time series data remains one of the limitations of this study and our solution will be discussed more thoroughly in the later sections.

Regression Analysis by State

The experiment is first conducted on the log-transformed unemployment rate and log-transformed meat production statistics from 2010 to 2022 without any additional constraints as a baseline for comparison. To our surprise, a very statistically significant result is obtained with an average R-squared of 0.86 and an average p-value for the F-Statistic of ~0.0018 for all states, rejecting the null hypothesis that the variable effect is zero. This suggests that there is a statistically significant linear relationship between the unemployment rate and meat production in the United States, in particular, there exists an inverse relationship between the two variables where higher meat production corresponds to a lower unemployment rate.



Interestingly, the only state that does not have a statistically significant result is North Dakota which is the state with the lowest unemployment rate among all states in the study period. This may be one of the areas to further investigate in the future.

From there, there are a lot of directions and questions we investigated and tested. In particular, 2 major areas of focus are the difference between red meat and poultry production and the lagging effect of the meat production cycle.

External Data - Monthly Unemployment Rate Per State

In order to have a more in-depth analysis of the problem, we decided to obtain external data - *monthly unemployment rate per state* (U.S. Bureau of Labor Statistics., 2024) to facilitate the regression analysis which also aims to address the problem of a small regression sample size as mentioned above.

A problem associated with using monthly meat production data for regression analysis is that the seasonal fluctuations of the time series data make the best-fitting line inaccurate, resulting in poor regression results. Hence, instead of using raw monthly meat production data as the dependent variable, we instead apply transformation techniques and replace it with the production trend statistics, which is the meat production statistics after removing the seasonality from the time series data, which is assumed to be constant for the purpose of the study as aforementioned in earlier section.

With more data points, we are able to undergo a more thorough analysis to discover a more in-depth relationship between the two variables. Understanding the difference in the trend between red meat and poultry, we first ran similar regression tests on red meat and poultry separately. A statistically insignificant result is obtained for both red meat and poultry, with an R-Squared of only 18.57% and 38.33% respectively. Upon further analysis, we discover that the unemployment rate exploded between 2020 and 2021, from around 2% to over 12%, which is largely due to the global pandemic during the period. This effect was surprisingly not demonstrated in the provided dataset. Upon different considerations, we decided to drop the affected period entirely for simplicity. That is, the new analysis period will be from the beginning of 2010 to the beginning of 2020.

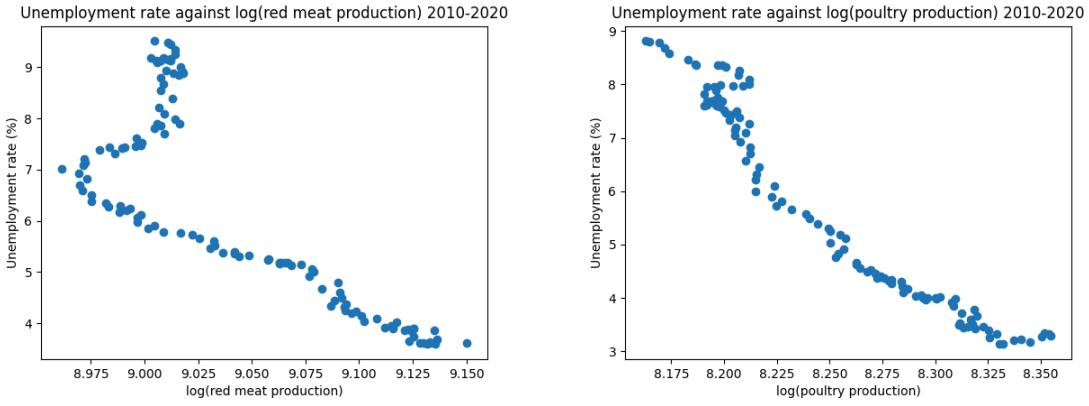
Upon adjustments to the analysis horizon, the result greatly improved for both red meat and poultry, with an R-Squared of 42.59% and 92.87% respectively when grouped by region. A clear distinction is observed between the results of red meat and poultry. By identifying a key difference between the trend between red meat and poultry, and by comparing the dip in red meat production trend in 2014 and the slight increase in the unemployment rate of key beef production states in 2013 we came up with the hypothesis that there may be some sort of lagging effect in the production cycle of red meat and poultry.

In fact, the production cycle of beef is usually around 52 weeks whereas the production cycle of poultry is usually around 7 weeks.

6 different shifting periods are tested for red meat production trends, which are ±3 months, ±6 months, and ±12 months, where a strong statistical result for a negative shifting period (shifting meat production series by a negative window) suggests that meat production may be a lagging indicator of the unemployment rate, conversely, a strong statistical result for a positive shifting period suggest that meat production may be a leading indicator of the unemployment rate. By comparing the R-Squared values among each shifting period, -12 months shifting window gives the highest R-Squared of 69.43% on average with the Northeast region attaining an R-Squared of 78.06%, an improvement of over 25% when compared to the regression analysis without any lagging window considered, suggesting lagging characteristics of meat production statistics when compared to the unemployment rate.

Supply Shock of Red Meat 2010-2014

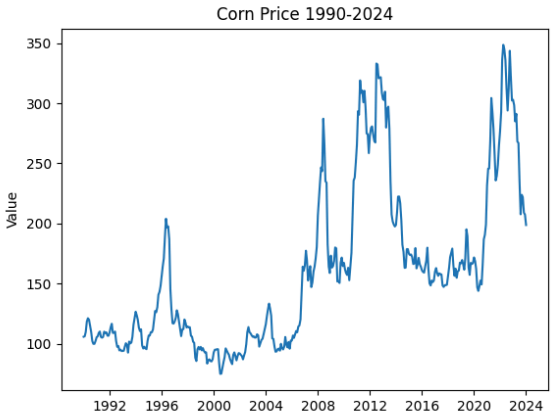
Acknowledging that the statistics shown are still not strong enough to support our hypothesis, we attempt to visualize the relationship with a scatterplot which we have done with the yearly data but not with the monthly data. The results shown (figure below) are shockingly poor, and we quickly realized that a huge mistake had been made, especially for red meat. The scatterplot indeed supports our hypothesis that there is a strong inverse relationship between meat production and unemployment rate data only when the unemployment rate is below 7%.



With the EDA we have done earlier, we successfully identified that the problem lies from 2010 to 2014, instead of the unemployment rate number. Acknowledging the problem is not enough to solve the current situation. Although we can directly change the study time horizon from 2010-2020 to 2014-2020, which

yields an R-Squared of an astonishing 95.03% (with a -12 month shifting window), and come to the conclusion that meat production demonstrates a strong inverse relationship with the unemployment rate from 2014-2020, we are yet to explain the phenomenon from 2010-2014 statistically. As there exists a distinction between red meat and poultry, we use the conclusion drawn from the EDA of meat production dataset to come up with the hypothesis that this distinction is caused by the meat production dip of red meat from 2010-2014.

An exciting discovery is soon revealed from the analysis we conducted for the machine learning task using commodities data (later section), the chart below shows the corn price surge during the 2010-2014 period.



In fact, corn is a key commodity for livestock feed especially for red meat like pork and beef, unlike poultry which is usually fed on cereals like oats. The surge in corn prices in 2010-2014 suggests that the production cost of red meat also surge during The period, which explains the drop in the trend of red meat production during the period as well as explaining the observed phenomenon, providing us with a reasonable justification to separate the 2010-2014 horizon out of the regression analysis horizon for red meat as an extreme event like COVID-19.

Are meat production yields predictive of restaurant stock prices?

Our Machine Learning Solution

To answer this problem, we decided to adopt a machine-learning approach to determine the impact of meat production on the stock price. By training a machine learning model to predict the stock price, we aim to analyze one of the side products of the model - feature importance as our key focus to draw respective insights. Our approach to solving the machine learning task is summarized as follows:

1. Construct a baseline model purely based on autocorrelation features from the stock portfolio as a benchmark for comparison

2. Construct meat production-related features on top of the baseline features to analyze the change in model performance and feature importance
3. Construct a machine learning model trained solely on meat production and commodity-related features on identified companies

To gain a more in depth insight to analyze the relationship between different sectors, we have set up 5 price-weighted portfolios on each sector provided from the given stock dataset, ie. Manufacturing, Technology, Trade & Service, Multiple Portfolio and Market Portfolio, where Market Portfolio is the price-weighted portfolios for all tickers in the dataset.

By doing so, we aim to analyze the difference and corresponding impact of each feature to each sector while reducing the extensive computing resources required to train machine learning models on each individual company. Acknowledging the difference in the business nature of different companies, we also conducted more in-depth analysis on 7 identified companies individually to explore a more one-to-one relationship between the production statistics of different types of meat and commodities on the stock price of the 7 companies as well.

Machine Learning Setting

The model aims to classify the cumulative return holding the portfolio over the next month. A three-class classification task will be the setting of this study, in particular,

- Class 0: The cumulative return over the next month will be less than -0.01
- Class 1: The cumulative return over the next month will be between -0.01 and 0.03
- Class 2: The cumulative return over the next month will be greater than 0.03

The corresponding thresholds are designed particularly to be slightly skewed to create a more balanced dataframe, serving as one of the potential limitations of the study. XGBoost Classifier will be used as the core model for the purpose of this study. The study takes 2000-01-01 to 2016-12-31 as the train period and 2017-01-01 to 2023-12-31 as the test period. While we acknowledge that several extreme market events such as the 2007-2008 global financial crisis and COVID-19 are included in the time horizon, a relatively long period is used to ensure sufficient data points will be used to train and evaluate the result as monthly data will be used for the purpose of this study.

Baseline Model

As mentioned, the baseline model will serve as a benchmark constructed solely through autocorrelation signals of the historical portfolio statistics itself. 4 features are incorporated in the baseline model for the purpose of the study.

1. *log_close_diff*: the difference between log(today's price) and log(price 1 month ago), this is to capture how the price of the portfolio have changed over the last month
2. *1month_return*: the return of holding the portfolio over the last month, this is to capture the profitability of the portfolio over the last month
3. *log_high_low*: difference between log(max. daily price over the previous month) and log(min. daily price over the previous month), this is to capture the range of stock fluctuation in the previous month
4. *vol_ratio*: prev_month_volatility / prev2_month_volatility, is the measure ratio of standard deviation from m_{t-1} to m_t and m_{t-2} to m_{t-1}, this is to capture how the stock volatility change comparing the previous 2 months

The performance of the out-sample data of the trained XGBoost model is similar for the 5 portfolios we constructed, with accuracy scores ranging from 45.9% to 57.6%, which exceed the results of 33.33% from random guessing by a fair margin, a more detailed result of the metrics will be displayed in the next section.

Meat Production + Stock Autocorrelation Model

In the second stage of our machine learning analysis, additional features are added on top of the baseline stock autocorrelation features for the 5 portfolios to capture the contribution of the meat production monthly statistics to the prediction of stock prices. Multiple features derived from the given datasets are tested to see any improvement in the accuracy of the model. The important features identified in each model were carefully noted and subsequently integrated into subsequent models to enhance accuracy. The features examined are the following:

1. *log_prod_diff*: this feature quantifies the change in meat production amount over the last month by calculating the difference between the logarithm of the current month's production and the logarithm of the previous month's production.

2. *{number of months = {1, 3, 6}}_pct_change_production*: these features measure the percentage change in production over fixed intervals of 1, 3, and 6 months, thereby considering the trend observed in the preceding months.
3. *{number of months = {3, 6}}_production_vol*: these features capture the standard deviation of production over the previous 3 or 6 months, providing insights into production volatility.
4. *lag_{window}_production*: this feature incorporates lagged production values to examine the alignment between meat production growth or decline and stock price direction.
5. *rolling_mean:* this feature calculates the simple moving average of production over the previous months, capturing the trend in production levels.
6. *exponential_moving_avg*: this feature computes the exponential moving average of production over the previous months, giving more weight to recent data points.
7. *zscore_{number of months}months*: this feature captures production spikes that are not accounted for by seasonality patterns by calculating the z-score of production over a specific number of months.
8. *lag_{window}_zscore_{number of months}months*: this feature represents a lagged version of feature 7, examining the lagged effects of production spikes.
9. *seasonal_{category}*: these features encompass seasonal indices obtained from the time series decomposition of meat categories, considering both individual meat and type of meat.

By incorporating these features, we aimed to leverage the information within the dataset to improve the accuracy of the predictive models and gain deeper insights into the relationship between meat production and stock prices. The table below shows the summary of the tests conducted:

In models 1 to 4, different clusters of animal meat are tested, and the importance of red meat production statistics is highlighted as seen in the results. Important features related to meat production are in bold.

Model	**Data Investigated**	**Important features**	**Portfolio**	**Accuracy**	**Baseline**
1	Total Meat Production	vol_ratio log_high_low log_close_diff **lag_6_production** **6m_vol_change**	Market	45.9%	48.2%
			Manufacturing	34.1%	44.7%
			Multi-portfolio	48.2%	57.6%
			Technology	48.2%	45.9%
			Trade & Service	44.7%	48.2%

Model	Data Investigated	Important features	Portfolio	Accuracy	Baseline
2	Beef and Veal Production	vol_ratio log_high_low log_close_diff **lag_1_production** **lag_6_production** **3m_vol_change**	Market	41.2%	48.2%
			Manufacturing	42.4%	44.7%
			Multi-portfolio	52.9%	57.6%
			Technology	50.6%	45.9%
			Trade & Service	48.2%	48.2%

Model	Data Investigated	Important features	Portfolio	Accuracy	Baseline
3	Red Meat Production	vol_ratio log_high_low log_close_diff **lag_6_zscore_6months** **lag_6_zscore_3months** **3m_production_vol** **6m_vol_change**	Market	48.2%	48.2%
			Manufacturing	44.7%	44.7%
			Multi-portfolio	57.6%	57.6%
			Technology	49.4%	45.9%
			Trade & Service	40%	48.2%

Model	Data Investigated	Important features	Portfolio	Accuracy	Baseline
4	Poultry Production	3m_production_vol 6m_production_vol 3m_vol_change 6m_production_vol	Market	45.9%	48.2%
			Manufacturing	38.8%	44.7%
			Multi-portfolio	49.4%	57.6%
			Technology	45.9%	45.9%
			Trade & Service	44.7%	48.2%

In models 5 and 6, seasonality indices were added to examine if the unusual residuals could capture market signals. Several features that are deemed not significant (based on F scores) are also eliminated to boost model efficiency. The inclusion of these indices aimed to identify underlying patterns or cyclicality that could impact stock price predictions.

Model	**Data Investigated**	**Important features**	**Portfolio**	**Accuracy**	**Baseline**
5	Red Meat Production + Seasonality Indices (Type of Meat)	log_high_low vol_ratio log_close_diff **log_prod_diff** **3m_production_vol** **3m_vol_change**	Market	48.2%	48.2%
			Manufacturing	40%	44.7%
			Multi-portfolio	50.6%	57.6%
			Technology	54.1%	45.9%
			Trade & Service	47.1%	48.2%

Model	**Data Investigated**	**Important features**	**Portfolio**	**Accuracy**	**Baseline**
6	Red Meat Production + Seasonality Indices (Individual Meat)	vol_ratio log_high_low log_close_diff **3m_rolling_mean** **lag_1_production** **3m_production_vol**	Market	52.9%	48.2%
			Manufacturing	40%	44.7%
			Multi-portfolio	56.5%	57.6%
			Technology	54.1%	45.9%
			Trade & Service	47.1%	48.2%

Despite the generally improved accuracy, seasonality indices are ranked the lowest in the important features in both models 5 and 6. The improved accuracy is deemed to have come from the more relevant features of the previous models; seasonality indicators are concluded to be irrelevant in this predictive model. Although most models do not outperform the baseline, a considerable improvement can be seen in the accuracy of the technology portfolio, which is particularly interesting, considering the undiscernible correlation with the meat production sector.

From the results above, it is believed that the indicators that appear in the top 5 most important features are able to provide relevant information in specific portfolios, or to individual securities. Therefore, a more comprehensive investigation has been conducted to test this hypothesis.

Meat Production and Commodity Model on 7 identified companies

In the third stage of our machine learning analysis, we focused on companies that were from the Retail-eating places and retail-eating and drinking places industry. Instead of using total meat production features in the previous stage, we investigated the importance of the production of meat from different animals and the value of different commodities in predicting the stock prices of the 7 selected companies. The features examined are the following:

1. *{Animal = {Beef, Veal, Lamb and Mutton, Broilers, Turkey, Pork, Other Chicken}}_Production_{number of months = {1, 3, 6}}*: these features measure the percentage change in production of different meat over fixed intervals of 1, 3, and 6 months, thereby considering the trend of meat production observed in the preceding months.
2. *{Commodities = {Coffee, Sugar}}_value_{number of moths = {1, 3, 6}}:* these features measure the percentage in the value of different commodities over fixed intervals of 1, 3, and 6 months, thereby considering the trend of the commodities supply and demand observed in the preceding months.

By incorporating these features, we aimed to gain deeper insights into the relationship between ingredients and restaurants. The table below shows the summary of the tests conducted:

In tests 1 to 7, different restaurants are tested, and the 5 most important and 5 least important features in the prediction with their corresponding F scores are listed as seen in the results. Accuracy of the models vs the accuracy of the baseline model are also shown.

Test	Company Investigated	Accuracy	Baseline	Top 5 Features	F Score
1	Chipotle Mexican Grill, Inc.	52.3%	46.5%	Beef_Production_1m	346
				Veal_Production_6m	323
				Beef_Production_6m	288
				Lamb and Mutton_1m	229
				Veal_Production_3m	221
				Bottom 5 Features	**F Score**
				Other Chicken_Production_6m	75
				Beef_Production_3m	75
				Pork_Production_6m	69
				Other Chicken_Production_1m	16
				Other Chicken_Production_3m	7

Test	Company Investigated	Accuracy	Baseline	Top 5 Features	F Score
2	Darden Restaurants, Inc.	27.3%	37.9%	Lamb and Mutton_Production_3m	232
				Veal_Production_1m	189
				Turkey_Production_1m	169
				Lamb and Mutton_Production_1m	148
				Sugar_value_1m	148
				Bottom 5 Features	**F Score**
				Turkey_Production_6m	81
				Sugar_value_3m	72
				Broilers_Production_1m	69
				Pork_Production_6m	68
				Beef_Production_3m	60

Test	Company Investigated	Accuracy	Baseline	Top 5 Features	F Score
3	McDonald's Corporation	36.3%	55.2%	Lamb and Mutton_Production_3m	666
				Turkey_Production_3m	623
				Sugar_value_6m	612
				Beef_Production_6m	498
				Other Chick_Production_6m	472
				Bottom 5 Features	**F Score**
				Pork_Production_1m	62
				Pork_Production_3m	51
				Coffee_value_3m	44
				Lamb and Mutton_Production_1m	18
				Sugar_value_3m	6

Test	Company Investigated	Accuracy	Baseline	Top 5 Features	F Score
4	Restaurants Brands International Inc.	27.3%	40.9%	Other Chicken_Production_6m	334
				Coffee_value_3m	259
				Turkey_Production_1m	251
				Broilers_Production_6m	179
				Beef_Production_3m	178
				Bottom 5 Features	**F Score**
				Pork_Production_1m	30
				Sugar_value_3m	21
				Beef_Production_6m	18
				Lamb and Mutton_Production_3m	16
				Sugar_value_1m	13

Test	Company Investigated	Accuracy	Baseline	Top 5 Features	F Score
5	Starbucks Corporation	41.8%	56.9%	Beef_Production_3m	302
				Other Chicken_Production_6m	286
				Lamb Mutton_Production_1m	257
				Sugar_value_6m	229
				Lamb and Mutton_Prouction_6m	202
				Bottom 5 Features	**F Score**
				Other Chicken_Production_3m	70
				Sugar_value_1m	53
				Turkey_Production_6m	53
				Coffee_value_3m	44
				Coffee_value_1m	20

Test	Company Investigated	Accuracy	Baseline	Top 5 Features	F Score
6	The Wendy's Company	25.5%	41.4%	Beef_Production_3m	782
				Sugar_value_1m	780
				Other Chicken_Production_6m	747
				Beef_Production_1m	724
				Broliers_Production_1m	700
				Bottom 5 Features	**F Score**
				Pork_Production_1m	302
				Pork_Production_3m	271
				Pork_Production_6m	269
				Turkey_Production_6m	247
				Lamb and Mutton_Production_6m	192

Test	Company Investigated	Accuracy	Baseline	Top 5 Features	F Score
7	Yum! Brands, Inc.	36.4%	58.6%	Sugar_value_3m	124
				Lamb and Mutton_Production_6m	108
				Coffee_value_3m	88
				Sugar_value_1m	87
				Lamb and Mutton_Production_1m	71
				Bottom 5 Features	**F Score**
				Other Chicken_Production_1m	37
				Pork_Production_6m	34
				Broilers_Production_1m	29
				Turkey_Production_6m	22
				Broliers_production_6m	18

Only the accuracy of test 1 outperformed the baseline accuracy. In general, the performance of the models are worse than the baseline model. This shows that solely using percentage change of each type of ingredient cannot effectively predict a restaurant's stock price. Delving deeper into test 5, the investigation in StarBucks showed that the 1-month and 3-month percentage change of the coffee value have the least significance on the prediction, which intuitively contradicts the Starbucks' business model.

From the results above, it is believed that the dynamic of the stock market is much more complicated than we expected. Factors such as macroeconomic conditions, geopolitical issues, and derivatives markets may also affect the stock prices. Hence, the percentage change of ingredients (meats and commodities) cannot be solely used as the features on the model. However, more domain knowledge in the meat production cycle is needed for future features engineering to further investigate the relationship.

Limitations and Future Directions

Several key limitations are mentioned throughout the report including insufficient data points for yearly regression analysis and the skewed classification bucket for the machine learning task. At the same time, the regression analysis for red meat separate the analysis horizon of 2010-2014 away from the period due to the surge of corn price which we regard it as an extreme event, though a better solution may be to conduct statistical analysis on different features such as meat production to corn price ratio to investigate further relationship between corn value, meat production and unemployment rate with the appropriate feature transformation process, which may be a valuable future direction to explore.

Another major limitation of this study is the insufficient features and variety of features tested in this study given the limited domain knowledge and time limitation. The study only mainly used simple statistical transformations such as percentage change and volatility as the feature engineering process due to time constraints, which unfortunately seems to result in an insignificant conclusion regarding the predictiveness of meat production yields on stock price. Future directions include testing of different additional features such as including lag features, using different meat groupings, and even to explore the possibility of the remaining meat relating datasets that are not included in this study.

References

Kuhn, M.T. & Hutchison, Jana & Norman, H.D.. (2005). Characterization of Days Dry for United States Holsteins. Journal of dairy science. 88. 1147-55. 10.3168/jds.S0022-0302(05)72781-8.

U.S. Bureau of Labor Statistics. (2024). Unemployment Rate in the District of Columbia [DCUR], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DCUR,

Walters, K. (2013). *The annual production cycle*. The Beef Site. https://www.thecattlesite.com/articles/3468/the-annual-production-cycle#:~:text=With%20a%20365%20day%20production,time%20from%20calving%20to%20conception).