Capstone Project Executive Summary

The Gold Standard

Greg Wagner, Chris Nash, Daniel Brickman, and Sawyer Tucker

Dev-10 21-11 Cohort

Introduction

Our group was originally asked to conduct research in the field of retail, marketing, and logistics. As two of our group are starting work in the agricultural field, we decided to focus there. We discovered an API that would allow us to gather real-time data on the current global prices of many commodities traded in the market, and we were curious what historically has affected the changing price. If we could discover what factors contribute to the changing price of commodities, we wanted to develop a machine learning model to help us predict the future prices of these commodities. We also aspired to create a model that would be re-usable and give any other attempt to predict future prices of commodities a solid base to start from.

Research Presentation

We used three datasets in our research. The first and most important data that we collected was from Commodities API. This API was to be called every two minutes and would report the current price in US dollars for certain commodities and conversion rates. This was the most important data set we used and was what started our interest in commodity data to start with. In addition to a near-constant feed of information, this API stored historic data and would allow us to access this data in our efforts to apply machine learning to predict commodity prices. It was here that we narrowed our field of research to three commodities: corn, wheat, and soybeans. These are the three crops that are most produced in America, and we decided that if we were going to make a satisfactory model, we should start with the commodities with the most data associated with them. During our research, we discovered that global supply is the most outstanding factor in determining the price of all three chosen commodities. This ran counter to the way we had originally hypothesized. We had thought that the weather of the previous year's harvest would be the determining factor, but our reasoning was both outdated and too small in scope. Unless hit by an outright natural disaster, modern industrial farming is more resilient in comparison to us growing vegetables in our home gardens or in feudal times. The number of crops produced is affected by weather, but on a smaller scale and on a greater scale than we were imagining. With that misunderstanding resolved and the knowledge that production of crops was sturdier than we had imagined, we became interested in which states produced the most of our chosen commodities.

The second set of data that we collected came from an API called the Economic Research Service API. This API was used to determine the wealth of domestic farms as well as how many there are subdivided by the state and size. This API also reports how much is produced by the farms and what commodity is produced there. This would be important for us in determining the largest domestic producers of our chosen commodities. This data set contained additional commodities such as cattle,

hogs, poultry, and specialty crops, which we decided to bring in as well. While this information would not contribute to the efficacy of our machine learning model or our understanding of our chosen commodities, we determined that being able to see the nation's top producers in these commodities as well would be interesting. Furthermore, these commodities may not concern us now, but this data could be used in the future if we wished to apply our model to other commodities.

Finally, our third set of data was collected from the Foreign Agricultural Service Data API and is responsible for tracking exports of our commodities from the United States. Not only can we see what commodity is being exported, but we see who we are exporting to and this data in terms of US dollars. We thought it would be interesting to see which countries consume the most of America's commodities because we discovered that America is the world's top producer of corn, wheat, and soybeans. We thought to check how much the United States imports of these goods as well but discovered that the United States does not have a need to import much of our chosen commodities; nearly all the corn, wheat, and soybeans we consume here are grown domestically. This data would be used to help determine the world's greatest consumers of our commodities. It was here that we thought it would be useful to relate the US dollar to the currencies of these countries. We returned to the Commodities API and began bringing in this relationship concurrently with the existing commodity data. The currencies we were concerned with was the Brazilian reais, Chinese yuan, euro, Indian rupee, and the Russian ruble. We also decided to relate gold to the US dollar because we thought it to be a universal currency.

To ensure that we would be able to keep these data sets apart from each other, we created a producer/consumer pair for each of these datasets in Azure DataBricks. These datasets were then saved to a .csv file in our data lake and written into our SQL database. We automated this process in our data factory and created a trigger to run every two minutes so that our commodity data is constantly being updated and new data is being appended to our existing commodity table. We started running this data factory on February 10th, 2022, at 5pm cst.

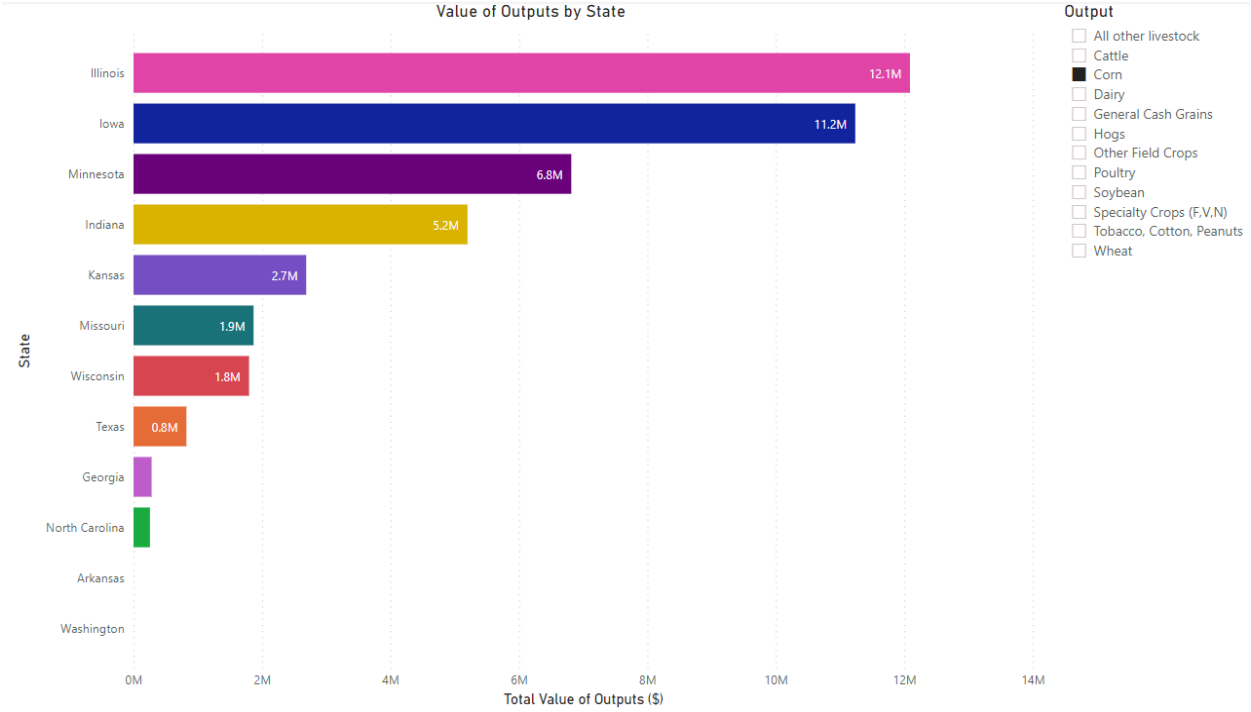
Machine Learning Model

The goal of creating a machine learning model for our data is to be able to predict the price of our commodities in the future. This means that we are working with time series data and would require a specialized model. There are, broadly, two types of time series forecasting. Multi-variate Time Series Forecasting uses predictor variables in order to predict future values and is not suitable for our purposes. We are using Univariate Time Series Forecasting, which means that we are using historic data to predict future values. Specifically, we are using an ARIMA (AutoRegressive Integrated Moving Average) model, which is an algorithm that creates predictions based entirely on historical data. Furthermore, we are receiving new commodity data every two minutes and this data is being integrated into the data set that we will be using in our model to predict future commodity prices.

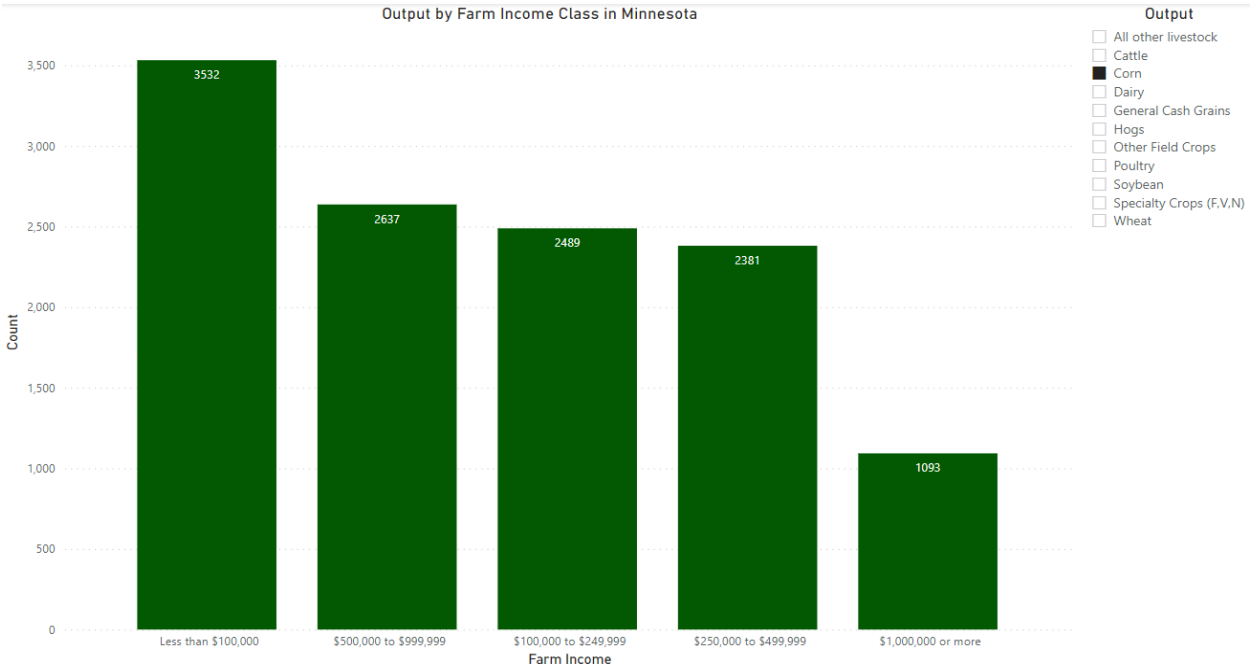
Results

In examining the data mentioned above, we discovered some answers to our questions. Firstly, regarding which states are producing the most of our commodities, we can see that the top producer of corn is Illinois, with $12.1 million dollars being produced. The next three states that are the largest producers of corn are Iowa, Minnesota, and Indiana, with $11.2 million, $6.8 million, and $5.2 million,

respectively.

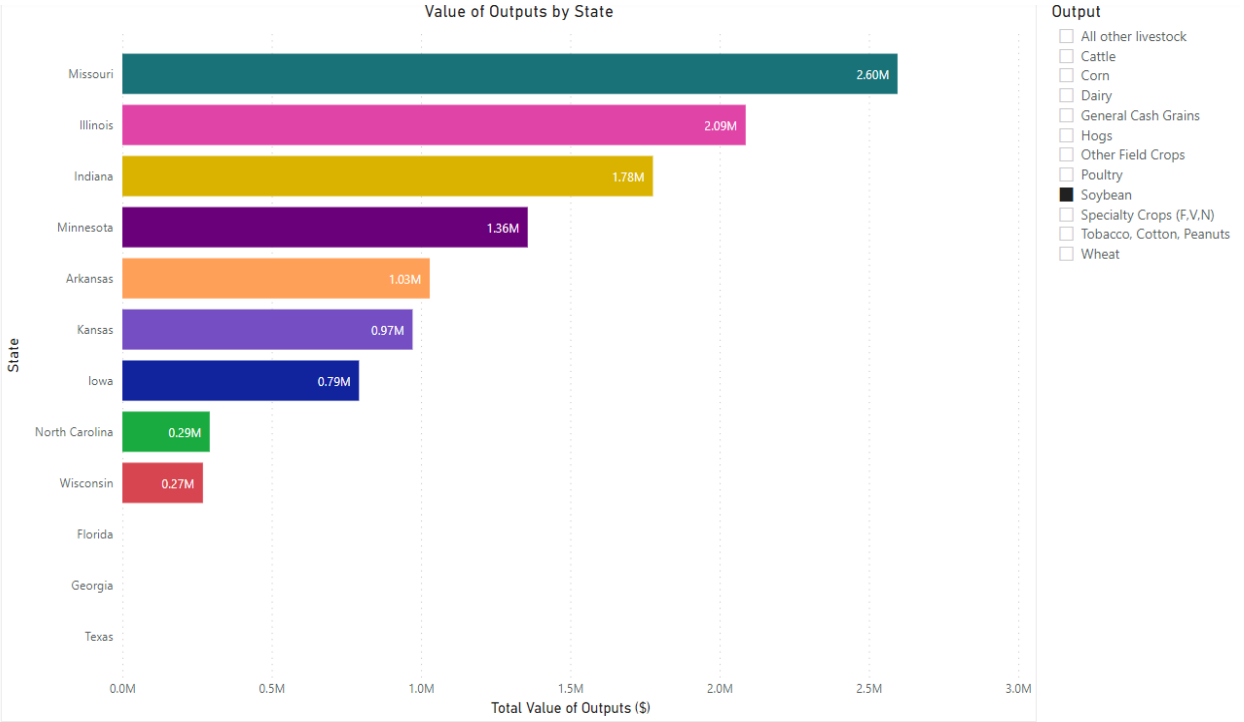


Additionally, we decided to focus on single state, in order to show the information regarding the size and number of farms producing our commodity. The graph below represents the number of farms in Minnesota of each income bracket that produces primarily corn. What is interesting to us in this visualization is that there are over 1,000 farms that make more than $1 million a year, which sounds staggeringly large. This does not match our other commodities.

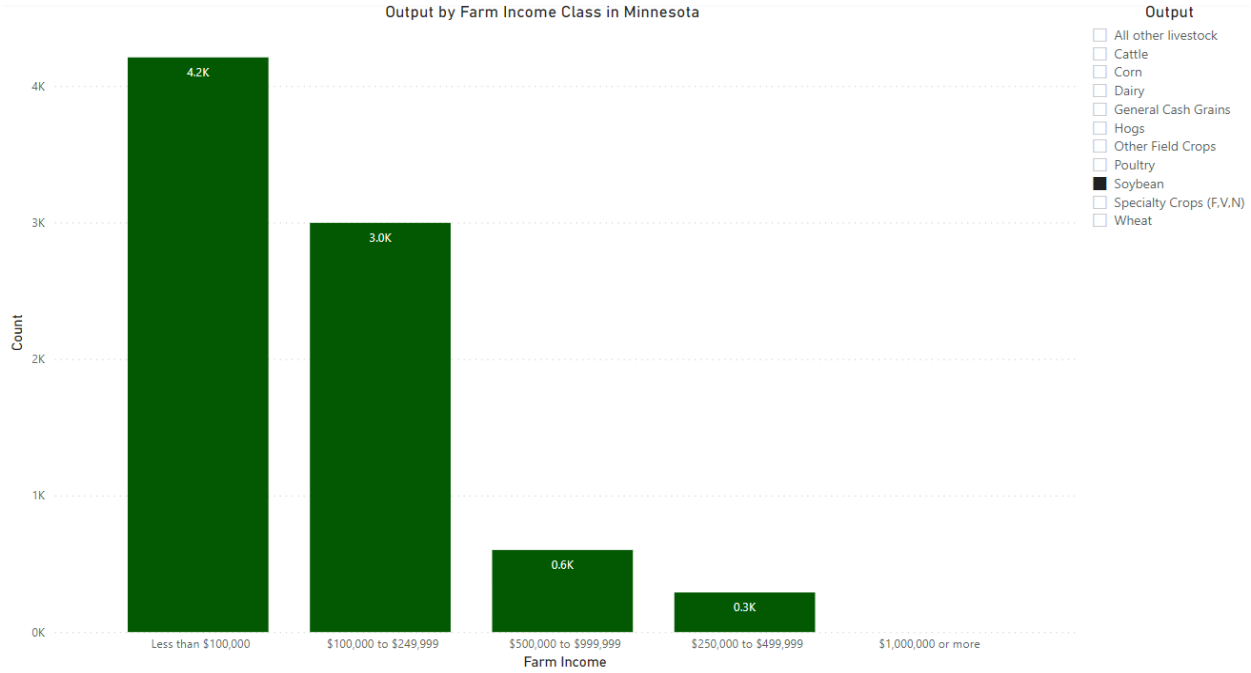


Next, the top producer of soybeans in the United States is Missouri, producing $2.6 million

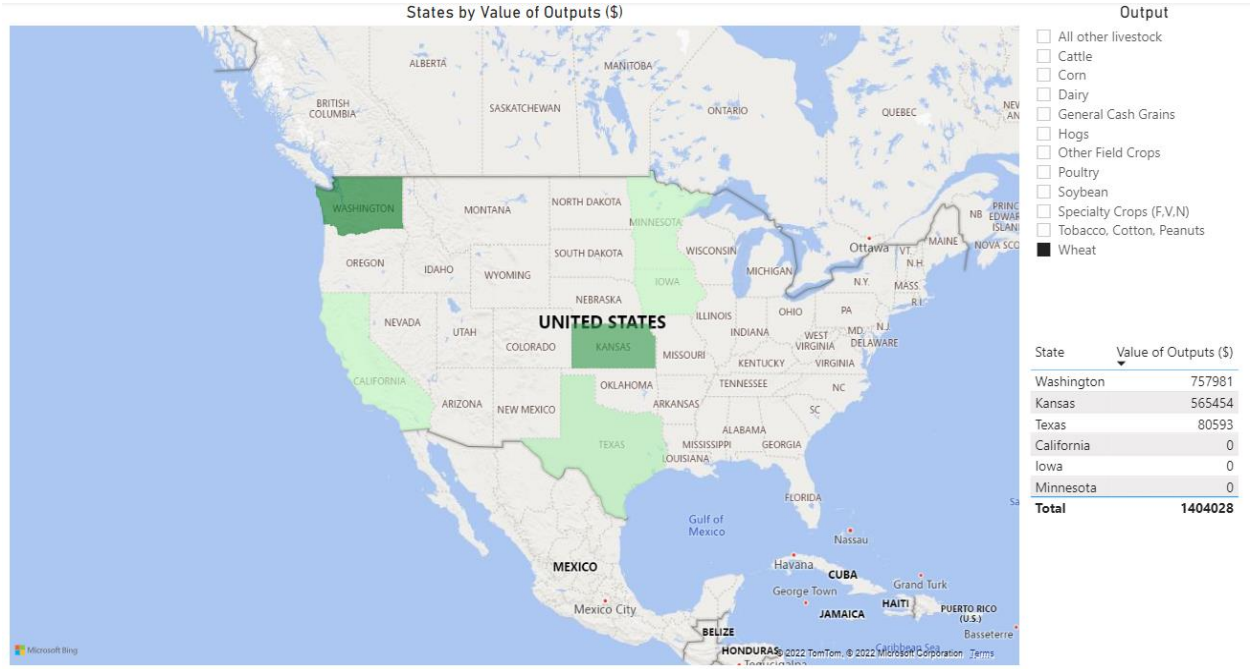
dollars of soybeans. The next three producers are Illinois, Indiana, and Minnesota, producing $2.9, $1.78. and $1.36 million dollars of soybeans.



The graph below is again only focused on the farms in Minnesota and represents the number of farms of each income bracket that produces primarily soybeans. As in the previous graph, the largest section of farms is also the section of farms with the smallest income level. This makes sense to us, because there are logically more small farms than large farms. What was interesting to us was just how many small farms there are, even within a single state. Also, while there are about as many farms that produce soybeans as corn, corn farms have more high-income farms, suggesting to us that corn farms are more generally profitable than soybean farms.

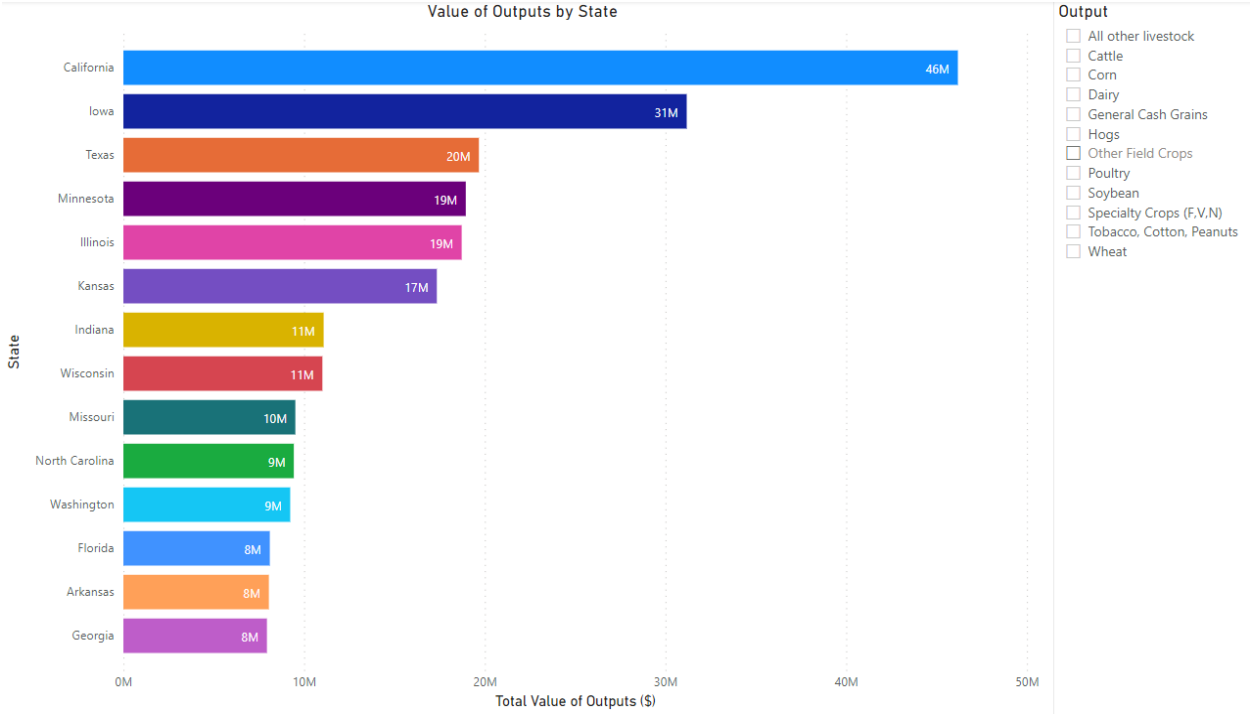


The top producer of wheat in the United States is Washington, producing $760 thousand dollars of wheat. Kansas and Texas trail behind with $570 and $80 thousand dollars in wheat produced. This visualization initially surprised us, as it seems that there are only three states that produce wheat in all the United States. However, this data only categorizes farms based on their dominant crop and is not registered multiple times for each crop a farm may produce. Therefore, it is likely that wheat is still grown and exported at some level in many other places than Washington, Kansas, and Texas, but this will go unrepresented considering how the information is represented in this dataset.

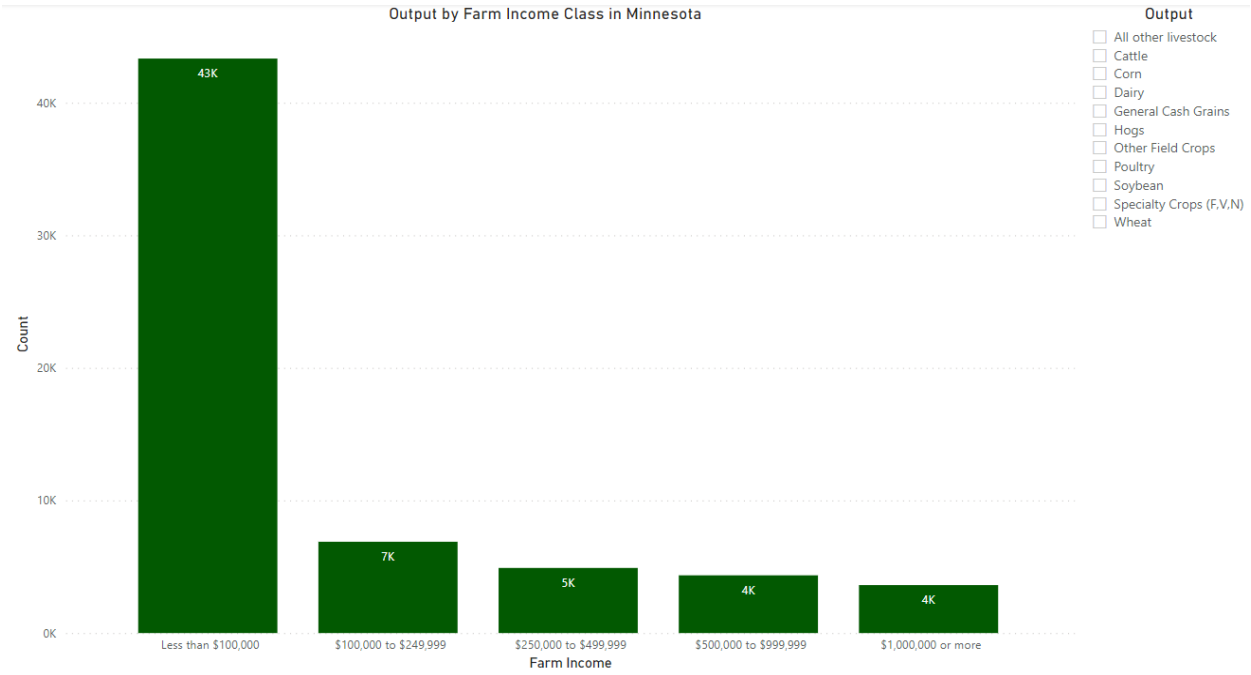


The data that we were using to investigate this information also contained information on other commodities produced in America's farms, and our visualization of the data includes these other

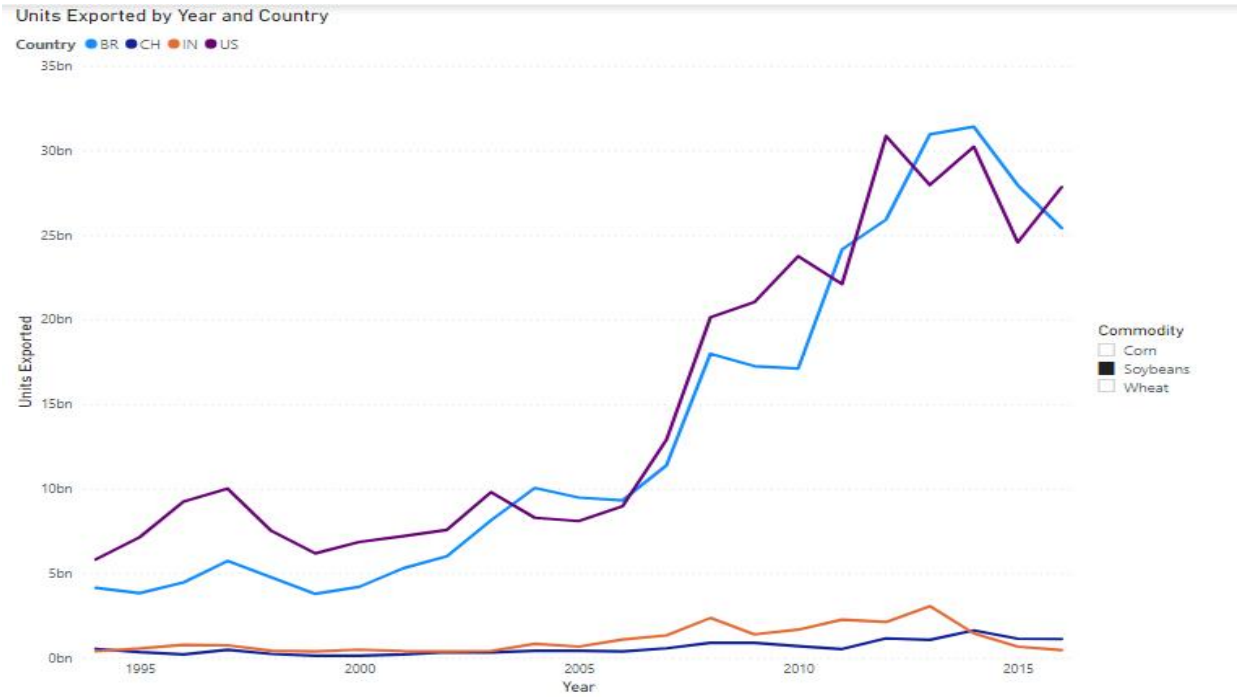
commodities, as seen on the slicer to the right of the graph below. If ever we are interested in expanding our area of interest, we may be able to start here. Currently, the graph shows the states ranked in order of total general goods produced on farms. California is the largest producer, followed by Iowa and Texas. This was interesting to us because while California and Texas are massive states, Iowa is not, but it still ranks with the others. We would expect larger states to have more production, but we concluded that Iowa has a larger *proportion* of its land set aside for farming than California or Texas.



The graph below represents all farms subdivided by income bracket, regardless of the commodities produced. As we can see, smaller farms are much more common than larger farms, which has been echoed throughout all our previous data sets. Farms that have income less than $100,000 per year account for 69% of all farms in America, while only 5.7% are generating over $1 million per year.



The next graph describes the history exporting totals of our commodities between the top agriculture producing countries in the world (US, China, India, Brazil). Setting up this chart could indicate what certain country's economy we wanted to analyze. The results show that the US dominated wheat and corn, while having a major rivalry with Brazil over Soybean production:

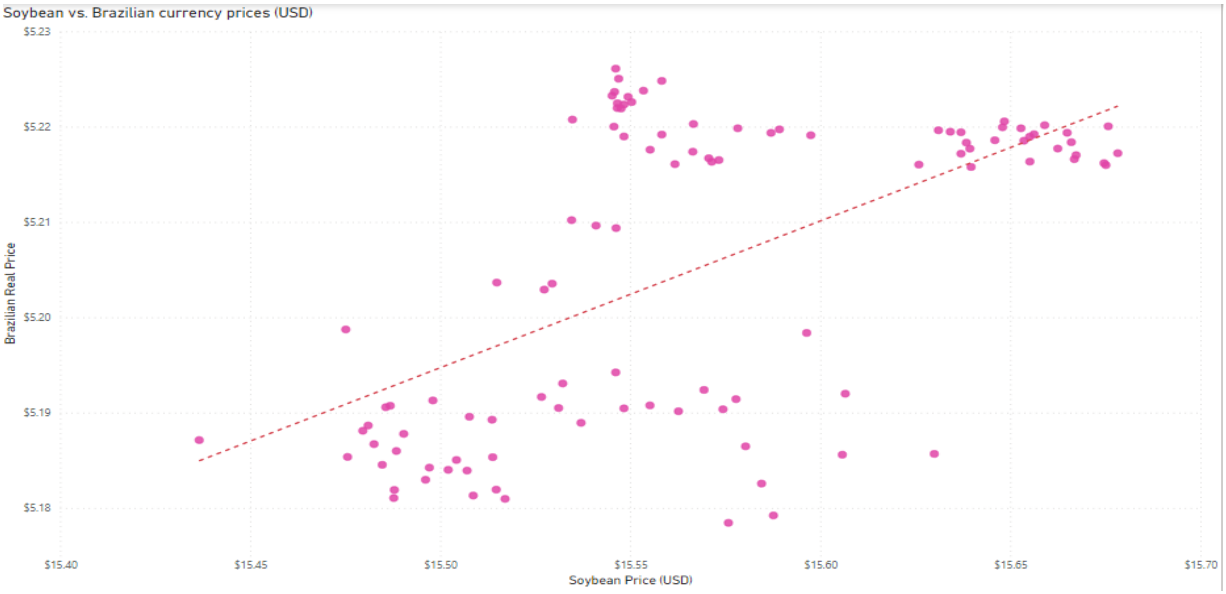


Seeing as this is the only major competitor with the US, we wanted to see how much soybeans could be tied to Brazil's economy. Despite not being our actual regression modeling goal, we did a quick linear correlation matrix on the data to see if there might be a sign of a relationship between soybean price and the strength of the Brazilian *real:*

Correlation Matrix of Commodity API data:

	Brazil	China	Corn	EU	India	Russia	Soybean	Wheat	Gold
Brazil	1.000000	0.292470	0.277658	-0.107523	-0.027667	0.354006	0.727860	-0.165514	-0.613568
China	0.292470	1.000000	0.135128	-0.039133	0.361277	-0.015375	0.209093	-0.117623	-0.274863
Corn	0.277658	0.135128	1.000000	0.585148	0.409206	0.710470	0.409028	0.747494	0.156548
EU	-0.107523	-0.039133	0.585148	1.000000	0.532474	0.566918	-0.288477	0.850292	0.409850
India	-0.027667	0.361277	0.409206	0.532474	1.000000	0.309386	-0.132234	0.470876	0.230484
Russia	0.354006	-0.015375	0.710470	0.566918	0.309386	1.000000	0.437031	0.620987	-0.216702
Soybean	0.727860	0.209093	0.409028	-0.288477	-0.132234	0.437031	1.000000	-0.094772	-0.626773
Wheat	-0.165514	-0.117623	0.747494	0.850292	0.470876	0.620987	-0.094772	1.000000	0.427795
Gold	-0.613568	-0.274863	0.156548	0.409850	0.230484	-0.216702	-0.626773	0.427795	1.000000

Real Price vs. Soybean Price (USD):



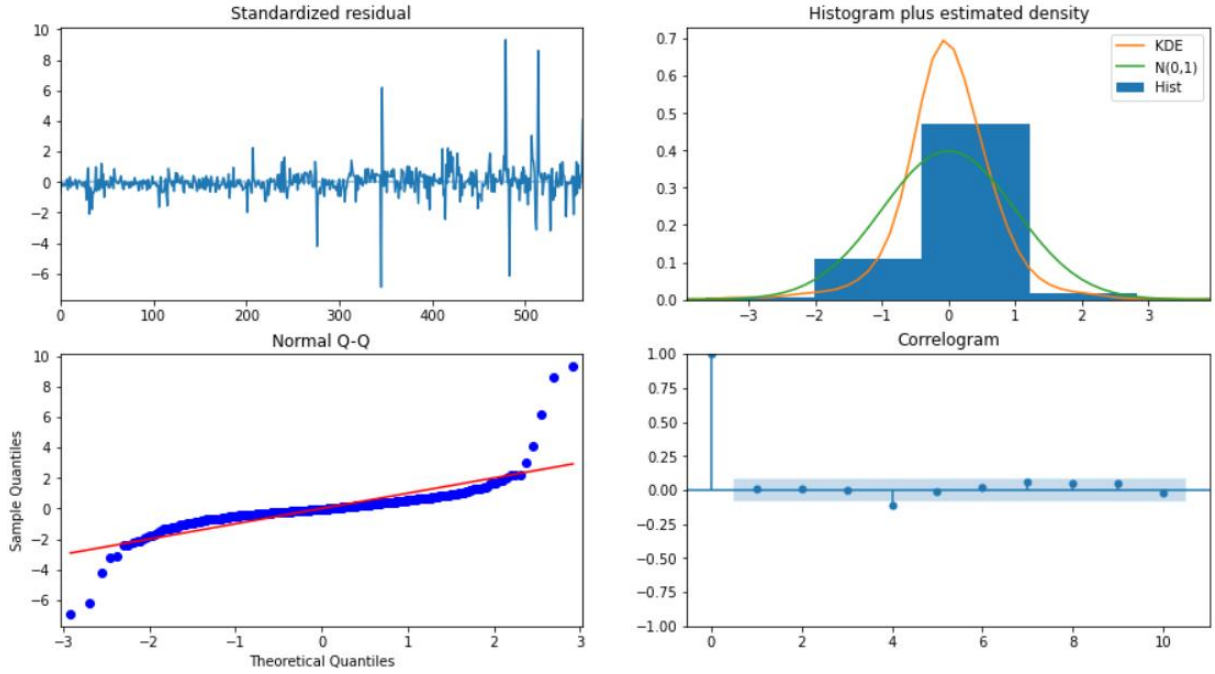
There is a high correlation between soybean prices and the price of Brazil's currency (in USD). However, we must remember what this means: The currency value going up means that it is getting *weaker* against the USD. Based on the laws of supply and demand, low supply or high demand causes prices to rise. Seeing as to how important soybean exports are to Brazil based on the graph, it may be in their best interest to keep production high to create more value for their currency. This pattern follows to other countries as well (like Wheat vs. Russia's *ruble*), so it may be worthwhile to study this relationship further.

Machine Learning Model Results

To run our model, we used the methods provided in a stock market forecasting article since commodity and stock market prices are similar. The data below is focused on predicting corn prices, but the process is the same for wheat and soybeans. We split the data into testing and training data. Our

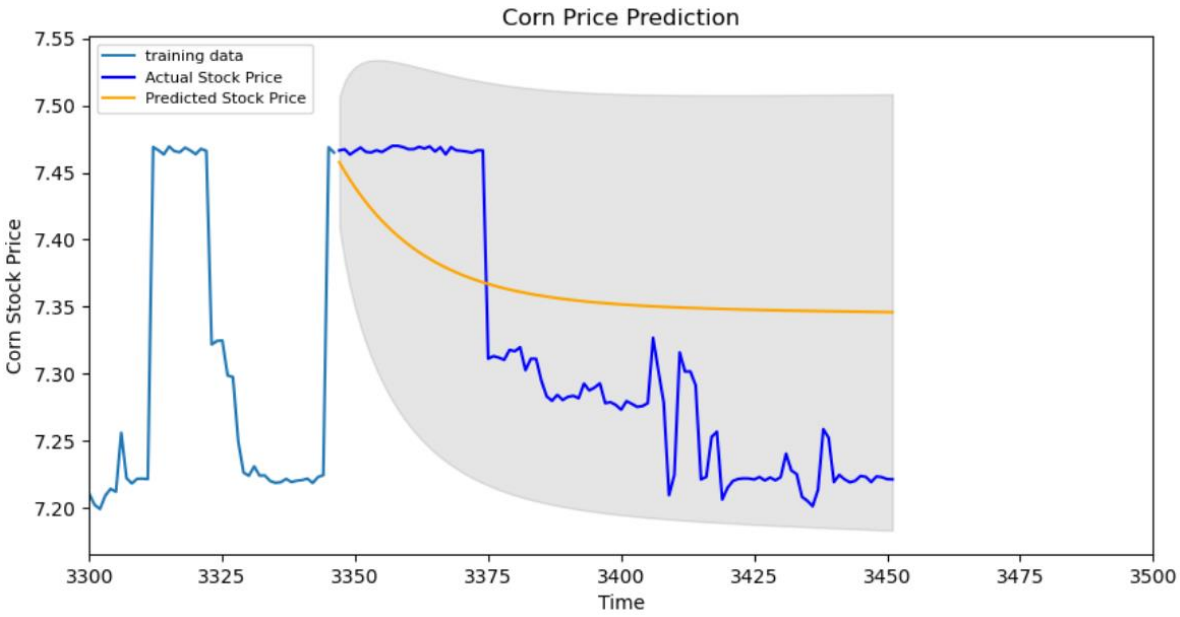
testing data is 90% of our total data set, and our testing data is the final 10%. This split is not done randomly; because this is time series data, our training data is the earliest 90%, and the testing data is the last 10%. Additionally, our data set is being updated every two minutes, so on what date this split is happening is constantly changing. The only constant is the proportion of the data being used. Because we are producing new commodity price data from our API, we hope that this model will only become more accurate the longer time goes on and the larger our training data set becomes.

This model contains three hyperparameters, usually called p, d, and q. The hyperparameter 'p' is the order of the 'Auto Regressive' (AR) term. It refers to the number of lags of Y to be used as predictors. Hyperparameter 'q' is the order of the 'Moving Average' (MA) term. It refers to the number of lagged forecast errors that should go into the ARIMA Model. Hyperparameter 'd' is the minimum number of 'differencing' needed to make the series stationary. To difference a data point we subtract the previous data value from the current data value in order to minimize the amount of correlation between predictors. We used a function called 'autoARIMA' to systematically determine the most accurate complement of hyperparameters to our model. The output of the function showed that the best model was ARIMA (1,0,3) intercept, with p, d, q values of 1, 0, 3 respectively. This model had the lowest AIC score of –5683.77 and took 0.3 seconds to run. The assumptions for this model were also met, as seen by the four graphs below.



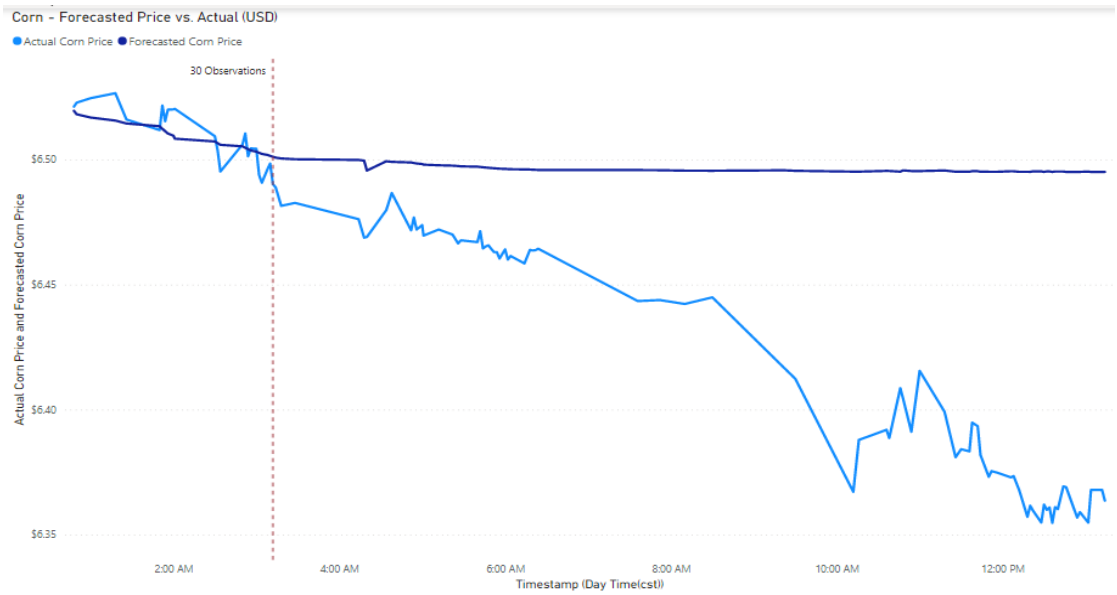
In the Standardized residual plot, there is no real trend for the residuals indicating constant variance throughout the model. In the Histogram plus estimated density plot, the normality assumption is met as our data appears to be normally distributed. In the Normal Q-Q plot, the data follows the line indicating a uniform distribution. In the correlogram, the residuals are not autocorrelated which implies that our data does not have a pattern that is not represented by the model, and there is no need to add more variables to our model.

When applying our model and the parameters that we found above, we were able to create the graph below:



In this graph, our X axis is labeled time which represents entry number since we started collecting data in two-minute intervals on February 10th, 2022, at 5pm CST, and our Y axis is the price of corn per bushel in USD. The light blue line is the training data (going back to when we started collecting data), and the dark blue line is the test data. The yellow line cutting through the dark blue line is our predicted stock price which has a negative trend. The grey shaded region around the yellow line is our 95% confidence interval. So little of the testing data is outside of the 95% confidence interval, which were caused by large positive/negative spikes, so our model was able to produce the low AIC score, a MSE (mean squared error) of 3.506e-05, and a MAPE (mean absolute percentage error) of 0.00235 which tells us that our model is about 99.765% accurate.

However, naturally with time series forecasting, the accuracy fades out quickly the farther out from the training data you go. The following graph charts the forecasted price for corn compared to actual price. After about 30 observations, the forecasted price levels out, while the actual price drops considerably.



Conclusion

As we conducted our research, we discovered that of the three commodities that we were concerned with, corn seems to be the most profitable and widespread, accounting for the most profit over a similar amount of farms as soybeans. Unfortunately, while the data seemed quite representative of corn and soybeans, it appeared to be lacking regarding accurately depicting the amount of wheat produced. The US dominates world exports among world powers for these commodities, and production has skyrocketed since around 2005. There may be a relationship between commodity valuation and certain currencies, based on each country's exports and production.

In using the ARIMA model, we were able to be accurate in predicting several records ahead, marking the model successful. However, as is common regarding time series machine learning models, the accuracy of the model falls off after a time and becomes useless in making long-term predictions. In the last graph in the machine learning section, the upper and lower 95% confidence interval lines have different slopes, indicating we become less confident of the true value of the price of our commodities but still maintain our 95% confidence with our larger range. Another reason why our machine learning model will become less accurate over time is there are real world events that are impossible to predict. In our case when looking at the price of commodities, we think part of the variation in the data is the uncertainty going on between Russia and Ukraine, and we understand that as the possibility of war increases/decreases, our prices will fluctuate as well. Still, barring any unforeseen political or natural changes in the environment, our model is rather accurate for the first thirty data entries ahead.

As for recommendations, something that would make our perception of the commodity market more accurate is more current data. The farm data that we used to determine the country's top producers of our chosen commodities is from a 2017 census and is now approaching five years old. While there is still relevant information to be gathered from this data, the financial and cultural landscape of the world changed considerably due to COVID in 2020, so getting access to some more current data may give us a

better impression of how this change has influenced this market in that regard. Regarding our machine learning model, something to keep in mind is that the model always needs to be updated with current data in order to be accurate and relevant. Furthermore, we only started collecting data from our data source on Friday, February 11th, which means that our data set is rather young. As time goes on and we have more data that we can train our model on, we expect that the model will become more accurate. Right now, our model is only accurate to between thirty and fifty entries ahead, but perhaps this can be extended in the future with a larger training data set.

Sources Cited

Arms data API. USDA ERS - ARMS Data API. (2021, December 16). Retrieved February 11, 2022, from https://www.ers.usda.gov/developer/data-apis/arms-data-api

Crude Oil Price API: Commodities prices and currency conversion JSON API. Commodities. (2022, February 11.). Retrieved February 11, 2022, from https://www.commodities-api.com/

Foreign Agricultural Service Data APIs. OpendataWeb. (n.d.). Retrieved February 11, 2022, from https://apps.fas.usda.gov/opendataweb/home

Prabhakaran, S. (2021, December 19). *Arima model - complete guide to time series forecasting in python: ML+*. Machine Learning Plus. Retrieved February 11, 2022, from https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

Oktoviany, P., Knobloch, R., & Korn, R. (2021, November 18). *A machine learning-based price state prediction model for agricultural commodities using external factors - decisions in economics and Finance*. SpringerLink. Retrieved February 11, 2022, from https://link.springer.com/article/10.1007/s10203-021-00354-7

Stock market forecasting using time series analysis with Arima Model. Analytics Vidhya. (2021, July 11). Retrieved February 11, 2022, from https://www.analyticsvidhya.com/blog/2021/07/stock-market-forecasting-using-time-series-analysis-with-arima-model/