Tej Rai
DS 710
Executive Report

Analysis of U.S. Wheat Market Dynamics

The overarching goal of this project was to apply various statistical and data science techniques to analyze historical data and external factors affecting the price and production of wheat in the United States. With a focus on testing methodologies learned throughout the semester, the project aimed to explore new avenues of analysis and create a comprehensive study suitable for inclusion on my resume.

The research questions guiding this investigation delved into key aspects of the wheat market dynamics. Firstly, the comparison between WASDE (World Agricultural Supply and Demand Estimates) projections and actual wheat prices received by farmers provided valuable insights into the accuracy of agricultural forecasting. Secondly, the exploration of how weather patterns, specifically monthly averages of temperature and precipitation, impact wheat prices sought to uncover correlations between environmental variables and market outcomes. Additionally, the study explored the relationships between U.S. wheat exports, foreign wheat prices, and domestic wheat prices to understand global market dynamics. Lastly, the investigation into the impact of U.S. wheat production on food use and exportation furthered my understanding of the factors influencing domestic agricultural practices. Through these inquiries, the project aimed to contribute to understanding wheat market dynamics while showcasing proficiency in statistical analysis and data-driven decision-making.

Naturally, in a data analysis project, one has to acquire data to start; I chose to use the USDA Economic Research Service and NOAA. The USDA ERS is part of the U.S. Department of Agriculture and provides information on agriculture, food, and economics. The National Oceanic and Atmospheric Administration is an agency tasked with forecasting weather and tracking oceanic and atmospheric conditions. In terms of project requirements, I had the option to download or web scrape data, however I did both because of the difference in data acquisition for both websites. The USDA ERS had data readily available to download in Excel files and in contrast it was easier to web scrape data from the NOAA. In terms of wheat, I focused on obtaining data surrounding domestic and foreign prices, exports, use, and production. While the climate data was regarding precipitation and average temperature values from stations in Kansas for 2017 to 2020. I chose Kansas because much of the wheat data included information from this state, and Kansas is a well-known domestic producer of wheat. I had to limit the amount of data I acquired otherwise the project likely would have been much more daunting.

Additionally, I decided to create my own excel files comprised of tables I copy/pasted from the USDA ERS excel files because the original files contained a lot of descriptive content and unstructured data (sometimes coding everything is not the best solution). For example, the original 'Wheat Data-Recent' excel file contained 35 sheets with descriptive header, merged

columns, and merged rows. Moreover, the original 'futmodwheat' file had 16 unique tables in a sheet vertically and horizontally arranged. Despite creating my own files, I was careful not to take away from the experience of using Python to clean the files.

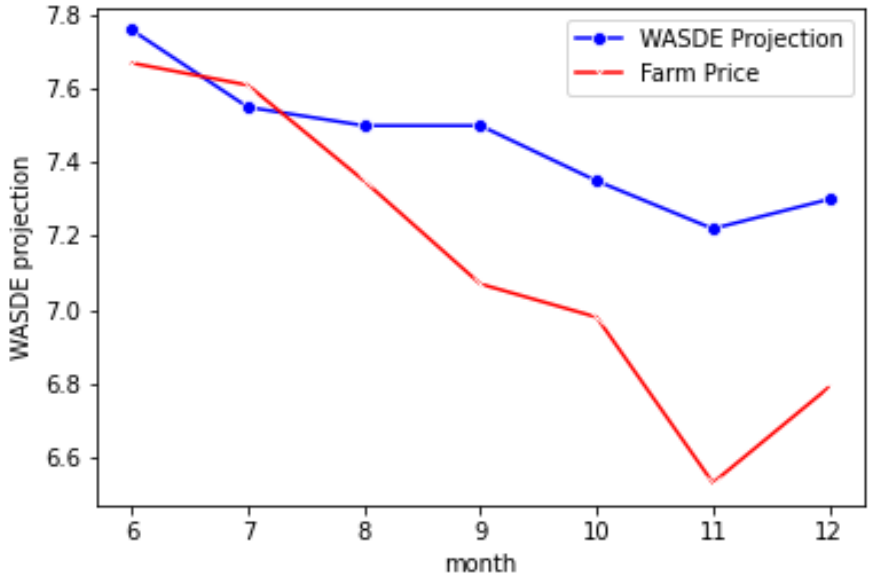
In terms of techniques to collect the data, I created 2 gathering python files. The first file 'wheat_data_gathering_s1' (s1: step 1) imported the excel files I created from the USDA ERS files from my working directory and converted them to CSV files. Each excel file (Futmodwheat and Wheat Data-Recent) contained multiple sheets, so I used a for-loop to create a dataframe for each table on each sheet, which was saved to their own respective CSVs with the prefix of 'futmod_' or 'recent_' depending on which file the data came from. The second file 'climate_data_webscraping_s1' fetched data using a function with the request's module. Based on the documentation provided by the NOAA, I needed an API key and to set parameters (datasetid, stationid, startdate, enddate, datatypeid, units, limit, offset) to obtain the specific data I wanted ('TAVG' and 'PRCP'). Within my function I included some if-else statements for error handling and reporting; sometimes, data from a specific year would not import and I would need to rerun the code because of issues on the NOAA's end I used two different Station IDs for TAVG and PRCP because each respective station collects different types of data. I took data from 2017 to 2020 to have a fair amount of data to analyze with. Using a for-loop, each dataframe for each year for the respective datatype was fetched and concatenated into a single data frame for each datatype. The cumulative dataframes were saved as CSV files 'tavg_kansas_cimmaron.csv' and 'preciptation_kansas.csv'.

Prior to developing code for my analysis, I took an intermediate step and created a Python file that cleaned the data sheets ('data_cleaning_s2.py' (s2: step 2)). The goal of the initial cleaning process was to strip away any unnecessary columns and adjust the 'year' or 'marketing year' columns in each file. To begin, through a function, each '.csv' file was imported and converted to a dataframe. Afterwards, specialized functions were created to handle specific dataframes or groups of dataframes. In reference to adjusting the year, each function converted the year column to a datetime format; certain dataframes had marketing years with a YEAR/YEAR or YEAR-YEAR format that was cleaned with: `.apply(lambda x: x.split('/')[0]), format='%Y')` or `.apply(lambda x: x.split('-')[0])` respectively. I removed unnecessary columns like the 'attributes' column in the weather dataframes or empty columns in the 'futmod_WASDE_projections' dataframe. After the respective cleaning, each dataframe was saved to a CSV file with their respective name and _cleaned as a suffix. I chose to create a new CSV file for the sake of the project and having each processes' output file remain; in reality, it would make sense to overwrite the file to prevent a file overload, consistency, and simplification. It is important to note, this was an initial cleaning to reduce some code required in the analysis portion and prepare the data slightly; during the analysis code, in-depth preparation still occurred based on what I thought was necessary.

Now that the required data was collected and slightly cleaning, I began exploring my research questions by developing code to analyze the datasets:

Q1: How do agricultural forecasting and WASDE projections compare to the actual wheat price by farmers?

- What I used:
  - In order to investigate, I utilized the WASDE projections and the futmod_farm_prices dataframes. The WASDE dataframe contained monthly projections of price received for wheat by week from 2023 to early 2024 and futmod_farmprice dataframe contained the actual monthly price received from mid-1975 to early-2024. The prices were in units of Dollar per Bushel.
- What I did:
  - I decided to compare the overlap in time-period from June - December 2023 with a line plot. To sum up my process, I converted the years to datetime, filtered for my months of interest, took the average projection of each week in the WASDE DF, converted months to integers, and created a dataframe comprised of month (as an integer (6-12)), WASDE projection, and farm price. Utilizing the seaborn module, I created my line plot with the month as my x-axis and price for my y-axis.



- Observations/Analysis:
  - Throughout yearly, there was a general trend of decreasing price as the year progressed, suggesting that the projections may capture the overall market direction.
  - There was a noticeable gap between the WASDE projections and actual farm prices, with the projections consistently higher than the actual prices for the given months. This could indicate that market conditions turned out to be more favorable than expected, perhaps due to better-than-anticipated supply conditions, reduced demand, or other economic factors that drove the price down.

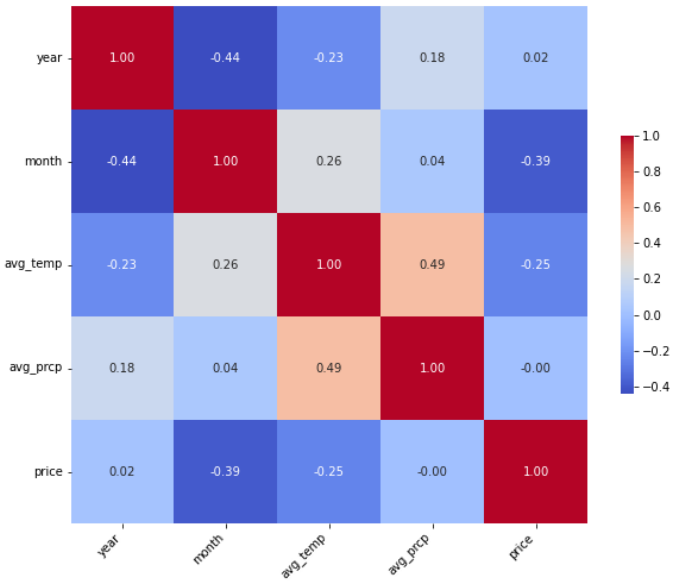
  - It is likely that lower actual prices compared to projections could mean less revenue for farmers and lower purchasing cost for buyers.
  - Though this is small timeframe, it may imply that analysts or market participants may need to adjust their models to better predict future prices by including unaccounted factors or improving their current model.

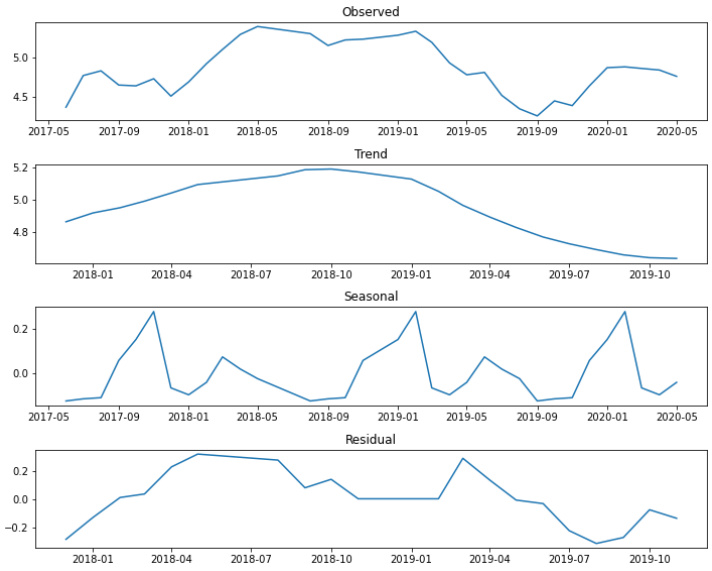
- Conclusion:
  - While the WASDE projections provide a useful benchmark for future prices, it is clear that market prices can deviate significantly from these estimates. Continuous monitoring and model adjustments for improving projection accuracy and understanding the reasons behind discrepancies is crucial for all stakeholders in agricultural commodities.

Q2: How do weather patterns (specifically temperature averages and precipitation) impact wheat prices?

- What I used:
  - I employed the two climate datasets (temperature average and precipitation) and the futmod_farm_prices DF to conduct my analysis. Both climate datasets contained the daily respective value (temperature average (°F) and precipitation (mm)) from 2017 through 2020.

- What I did:
  - I examined the correlation between the monthly average TAVG, monthly average precipitation, monthly wheat price received by farmer through a heatmap with a correlation matrix and time series plot. First, I had to adjust the shape of the farm prices dataframe; the farm prices year column was organized as a marketing year so 2016-2017 was June 2016 to May 2017. I reshaped the dataframe such that a year's row would contain price data from January 2016 to December 2016 and so forth. I did this because originally when I created the final merged dataframe the monthly data for the respective year of the weather datasets were not properly mapping to the corresponding monthly price. Then, I grouped the weather dataframes by year and month, and calculated the average temperature and precipitation. Consequently, the final merged dataset contained the year, month (as an integer), avg_temp, avg_prcp, and price. Using the final dataframe, I created a heatmap from a correlation matrix. Additionally, I used statsmodel's seasonal decompose to help create a time series plot. Numerically, I used scipy.stats's pearsonr and statsmodel's acorr_ljungbox to provide calculations on correlation coefficient, p-value, trend mean, trend standard deviation, seasonal amplitude, residual mean, and residual standard deviation.





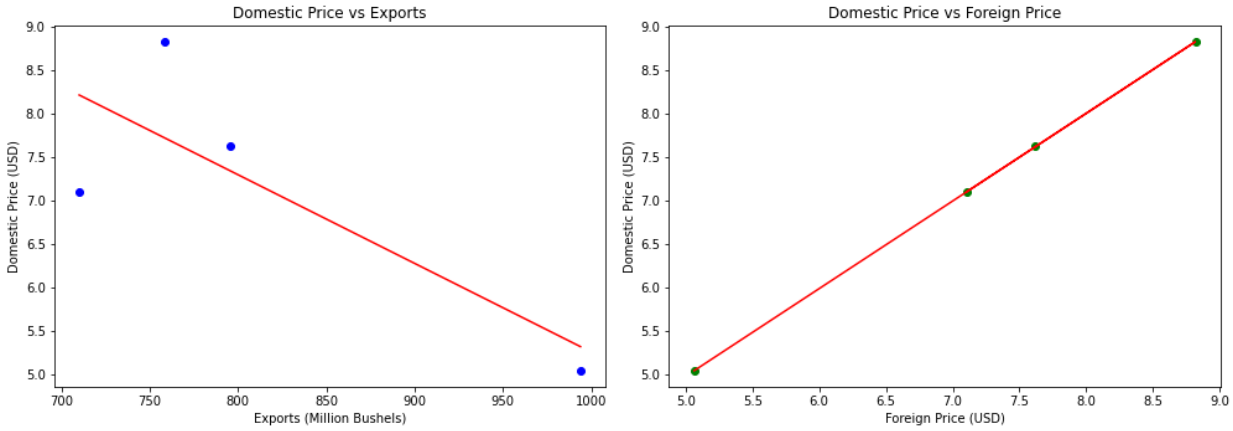
- Observations/Analysis:
  - From the heatmap, I noticed a strong negative correlation between the month and price and a weak negative correlation between the weather variables and price. The correlation coefficient was a weak negative correlation (-0.25), suggesting as average temperature increased, wheat prices tend to decrease slightly. Surprisingly, there was 0 correlation between average precipitation and wheat prices. It is likely average precipitation has no correlation because if precipitation levels are generally within optimal range, then variations might no impact price significantly. Furthermore, in areas with irrigation may reduce the impact of natural precipitation.
  - From the time series plot, the seasonal component aligned with what one might expect that prices fluctuate based on weather patterns and seasonal practices. The trend standard deviation (0.189) was relatively low suggesting that prices do not fluctuate widely from mean or in other words prices did not drastically change over the 3-year course. In addition, the seasonal amplitude (0.20) suggested there is significant seasonal variation in wheat prices.
  - (view Q2_numerical_analysis.txt for numerical data)

- Conclusion
  - Overall, based on the generated visuals and corresponding numerical information the monthly average TAVG and precipitation have little to no effect on wheat prices.

Q3: How do U.S. wheat exports and foreign wheat prices affect domestic wheat prices received by farmers?

- What I used:
  - I utilized the recent Wheat Exports, U.S. & Foreign Wheat Prices, and Wheat Avg Price by Farmer datasets to research my question. The files had information concerning U.S. exports of wheat (million bushels), foreign wheat prices (dollars per metric ton), and domestic wheat prices (dollars per bushel)

- What I did:
  - I filtered for data from 2020 to 2023 because that was the time period overlap and then merged the data frames on marketing year. Also, I made sure to convert the foreign price from dollars per metric ton to dollars per bushel (1 metric ton = 36.74). From the final dataframe, I created simple linear regression graphs of domestic price vs. exports and domestic price vs. foreign price. In addition, I generated a regression model summary for each visual.



- Observations/Analysis:
  - Regarding to domestic price vs exports, the plot and model summary show a negative relationship between U.S. wheat exports (million bushels) and domestic price. The coefficient for Exports_MillionBushels is -0.0102, indicating that for every increase of one million bushels in exports, the domestic price of decreases by 0.0102 USD. Unfortunately, the high p-value for the slope 0.196 suggests that the exports variable may not be a statistically significant predictor of domestic price in this model.
  - In regard to domestic price vs foreign price, the plot and model depict a very strong positive relationship. The coefficient for foreign price is 1.0067, suggesting a nearly 1-to-1 relationship. Both coefficients have a small p-value, suggesting statical significance.

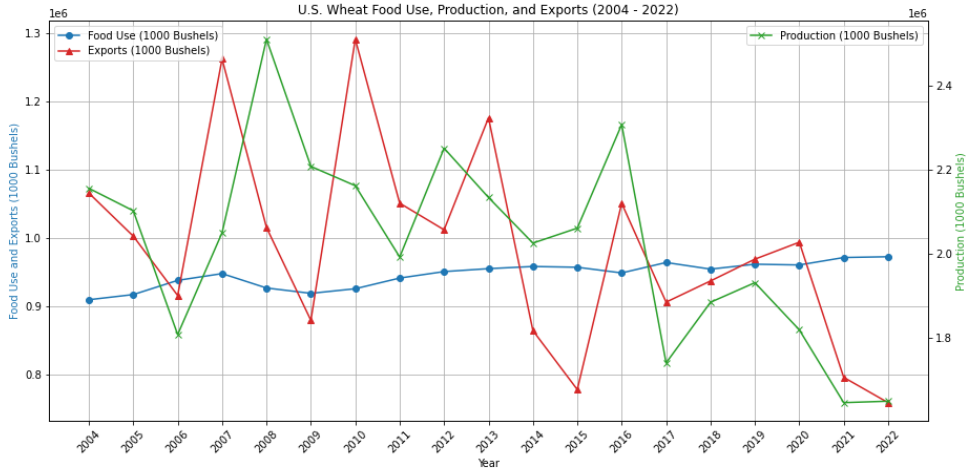
- (view Q3_regression_models for numerical data)

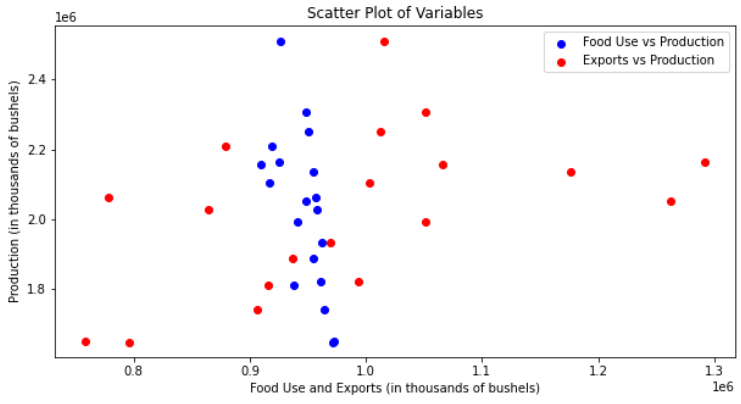
- Conclusion:
  - Despite my initial perception that both U.S. exports and foreign prices affect domestic prices, it is apparent that only foreign price trends align with domestic prices.

Q4: How is U.S. wheat production impacted by U.S. Wheat food use and wheat exportation?

- What I used:
  - I used the recent Wheat Food Use and Wheat Exports datasets to develop my analysis of the research question. The food use dataframe contained annual food use (1000 Bushels) and the exports dataframe contained U.S. production and U.S. exportation of wheat (million bushels). The analysis was conducted on data from 2004 to 2022.

- What I did:
  - Before merging the dataframes, I converted the production and export columns to 1000-bushel units. Afterwards, I excluded 2023 because on dataset did not contain enough data for 2023. In order to analyze this data, I created a time-series line plot, scatter plot of the variables, and a regression model summary.





- Observations/Analysis:
  - The time series plot indicates that food use remains relatively stable, production and exports show more fluctuation over the years with apparent peaks and troughs, and there are years production and exports seem to move in tandem, while in others, they diverge.
  - The scatter plot displays there is no clear trend line between production and food use or production and exports.
  - Despite the scatter plot, the regression model helps shed lights on the significance of the variables. The F-statistic value (6.807) shows that model is statically significant overall, and the independent variables collectively have an effect on production. In terms of relationships, the coefficient for annual food use (-6.4147) suggests as food use increases, wheat production tends to decrease (holding other factors constant) and the coefficient for U.S. exports (0.3996) indicating a slight increase in production with increased exports.
  - (view Q4_regression_model.txt for numerical data)

- Conclusion:
  - Though annual food use might have a negative impact on U.S. wheat production, U.S. exports doesn't appear to have a significant effect on production. This implies that for stakeholders in the wheat industry that while exports are important for the agriculture economy, internal use may impact the production capacity or allocation of wheat in the U.S.

To summarize, despite finding answers to my questions, it has become clear that understanding the effects of different variables on wheat production and domestic pricing is much deeper than at first glance. Furthermore, the results of my analysis have shown that natural assumptions may not always be true, especially in agricultural commodities. For example, I assumed precipitation correlate with domestic prices because of its effect on supply, but my current model has shown that as false.

Looking forward, there are definitely avenues for improvement on my analysis and more questions that can be observed. Overall, the analyzed portion of data should be expanded to a larger time-period in order to improve the accuracy of indicators and significance of trends. When creating these models, I've found that dependent variables like price and production require analysis against multiple variables in tandem (multiple climate patterns or occurrences, supply-chain dynamics, subsidies, technological advancements, etc.). Also, I believe it would be interesting to examine how production and price could be affected in short term by periods of drought, flooding, or natural disasters. New techniques like a SARIMA model or Granger causality test to explore price forecasting and causality instead of correlation could be employed. The analysis can incorporate multiple agriculture commodities for overall agricultural trends or break down by type of wheat to find secret trends. Whether it is expanding the timeframe, number of independent variables or dependent variables, or use of techniques, they all can provide great insight into the dynamics of the agricultural market.

In terms of my experience working on this project, I'm happy with my results as it was my first time working on something like this. I was also glad to practice new techniques like statsmodel, regression, and time-series analysis that I took time to explore out of class. I believe this was an invaluable experience that I can take with me display as an example of my skillset.