MSc in Data Analytics

Analysis of Cow’s Milk Production in Ireland and the Development of its Dairy Industry

Author: Mark Conaghan

e-mail: sba22496@student.cct.ie

Student ID: 22496

Word Count: 2907

Abstract

Ireland has a long history of dairy farming and is the sixth largest producer of dairy products in the European Union (EU)(Jack Kennedy, 2022). This analysis explores Ireland’s production of raw cattle milk from the 1960s to current times and compares its rate of production to that of Spain, the next largest milk producer within the EU. The level of impact of factors such as the greenhouse gas emissions produced by cattle, the yield as a measure of process efficiency and the selling price of cow’s milk is evaluated. This evaluation is performed to attempt to aid dairy farmers with general operations and the implementation of proactive initiatives to improve their business standing. Finally, sentiment analysis is performed to explore the popularity of dairy alternatives which may disrupt the established dairy industry in coming years.

Introduction

Ireland's dairy industry is a significant contributor to the country's economy, with milk and dairy products being the largest indigenous food export. The industry is made up of a mix of large and small-scale farms, with most of the milk production coming from family-run operations. In recent years, there has been a shift towards larger-scale, intensive dairy farming as the industry becomes more consolidated. Kerry Group and Glanbia are examples of large, established dairy companies which produce a range of products for domestic and international markets.

Analysis of data describing the production of dairy products benefits small scale and large-scale operations in several ways. Identifying trends in milk production can help farmers and companies optimise their operations to improve efficiency. Additionally, understanding the factors that influence milk production can help businesses make informed decisions about herd management, land usage and feed sourcing. In this report, an analysis was performed using publicly available data for such reasons and evaluated using industry standard methods.

Experimental Methods & Materials

Project management

CRISP-DM is a methodology widely used for data analysis projects in industry settings. This structured process provides a systematic way to organise, execute and assess projects and was utilised in this investigation.

CRISP-DM includes the following six phases.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment.

Business Understanding

Project objectives and requirements from a business perspective were defined as gaining insight into milk production and how the greenhouse gas emissions (GHG) from cattle, the associated yield and the selling price of cow’s milk are correlated. Cattle emissions was determined as a key factor for consideration given the development of green policies by the EU such as the Emissions Trading System (ETS) which caps the amount of GHG emissions produced by an industry. Although this system is not currently applied to the agricultural industry, increased capture of data and further characterisation of the sector will allow for its eventual implementation. Proactive characterisation of one's production process and its carbon footprint will benefit a producer in the long term.

Data Understanding

Available, relevant data was collected and thoroughly examined and characterised to determine its suitability for this analysis and for achieving the identified business objectives. Data was sourced from both the Food and Agriculture Organisation of the United Nations (FAO) and Eurostat, two publicly available databases controlled by reputable, international agencies. All datasets sourced from FAO are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO)(FAO, n.d.) . All other datasets used were sourced from Eurostat (Eurostat, n.d.)

Statistical tests were performed to characterise the available data and gain valuable insight for later modelling.

Data Preparation

Selected datasets were characterised, re-formatted and cleaned to facilitate exploratory data analysis (EDA) and later joining of datasets to allow for modelling using applicable machine learning models. EDA was performed to gain an understanding of the selected data through insightful visualisations and a dashboard. Gaps in datasets, abnormal results and trends in datasets were identified. Care was taken during EDA to ensure the final dataset was sufficiently processed and formatted for optimal model performance.

Modelling

Modelling was performed using Python libraries typically used for data analysis. A range of models were applied using Pandas_Profiling as a means of efficiently exploring the dataset and determining which models performed well. Such information allowed for further insight into the dataset. Where applicable, hyperparameter tuning and cross validation was applied to the dataset to validate model performance and robustness.

Evaluation

Upon completion of modelling, evaluation of the analysis was performed with respect to the identified success criteria. Critical analysis of work performed identified oversights and missed objectives which impacted deliverables to the relevant stakeholders. Upon review of project performance (achievements, resources used, budget), a decision was made to either return to the project to act on the gap assessment performed or to produce the project to stakeholders through deployment.

Deployment

Plans for model deployment and effective communication of results produced from the analysis were developed to maximise stakeholder understanding and satisfaction. Effective communication involved insightful, simple visualisations and a live dashboard which summarised key findings in one location for stakeholders.

Results

Business Understanding

The analysis performed was done to be of use to Irish dairy farmers, large and small. Given the broad scope of the project, a general approach was taken and so Irish milk production was compared to that of Spain's given both countries produced over 8 billion litres of raw milk as of 2022. Milk production was determined to be the key metric for this analysis. Key factors which may have an influence on milk production were identified with dairy farmers in mind to provide them with as much insight as possible. The selling price of milk, emissions produced from cattle and the milk yield were subsequently chosen and their impact on milk production was explored.

Data Understanding, Collection & Cleaning

All datasets were inspected and characterised systematically using statistical methods and visualisations. Key features were identified across three main datasets and later merged into one central dataset to allow for effective modelling. Many features within each dataset were selectively removed based on their relevance to the analysis and the amount of available data.

There were a range of dairy products in the original dataset that may not have been produced from cow's milk. Other milks such as camel and buffalo milk were captured in this dataset so an inspection was performed to determine what data related to cow's milk. By comparing the data of columns in comparison to the data relating to non-cow's milk, it was apparent which data related to cow's milk. This allowed for the effective filtering of the dataset to only capture cow's milk-related data, which would ensure the analysis performed was valid from a data perspective and no false data was used.

Production and yield data was captured within the main dataset. Both production and yield data were determined to be of value for the analysis, and production of raw cow's milk was determined to be the target variable or response for this analysis.

Figure 1 details the production levels of dairy products in Ireland from 1962-2022. Production of raw cows' milk was seen to be the main product produced. Similar analysis of Spain's production levels was performed.

Figure 2 details the spread in production levels of raw cows' milk in Ireland for the period of inspection. A large spread in results was observed as is evident in Figure 1 and Figure 2. A plateau in production was observed from the 1980s to 2015 in all EU countries.

This was the result of EU-placed milk quotas which capped the level of milk production (Läpple et al., 2022). Interestingly, the yield results in Ireland continued to increase in a linear fashion regardless of the implementation of the EU milk quota from the 1980s until 2015. This provided a key insight into the optimisation of farm operations through technological innovations relating to farming equipment and land treatment. Less land was required to produce the same level of milk during the lifetime of this milk quota.

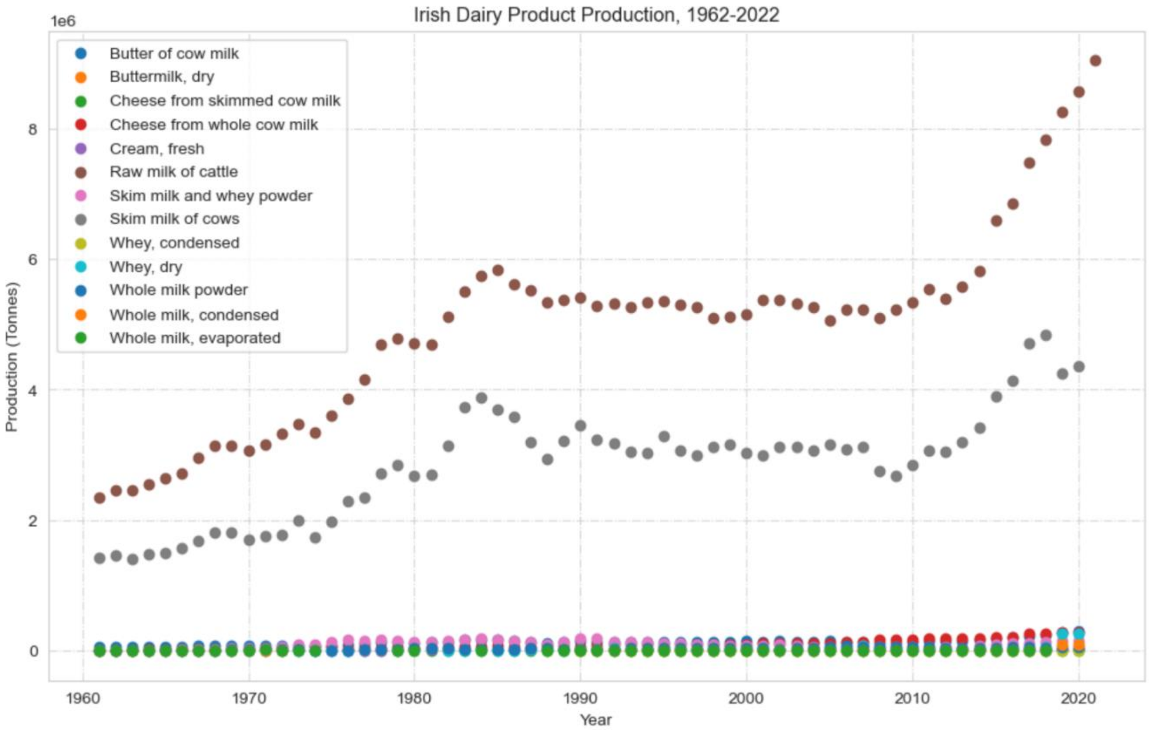


Figure 1: Production levels of dairy products in Ireland, 1962-2022.

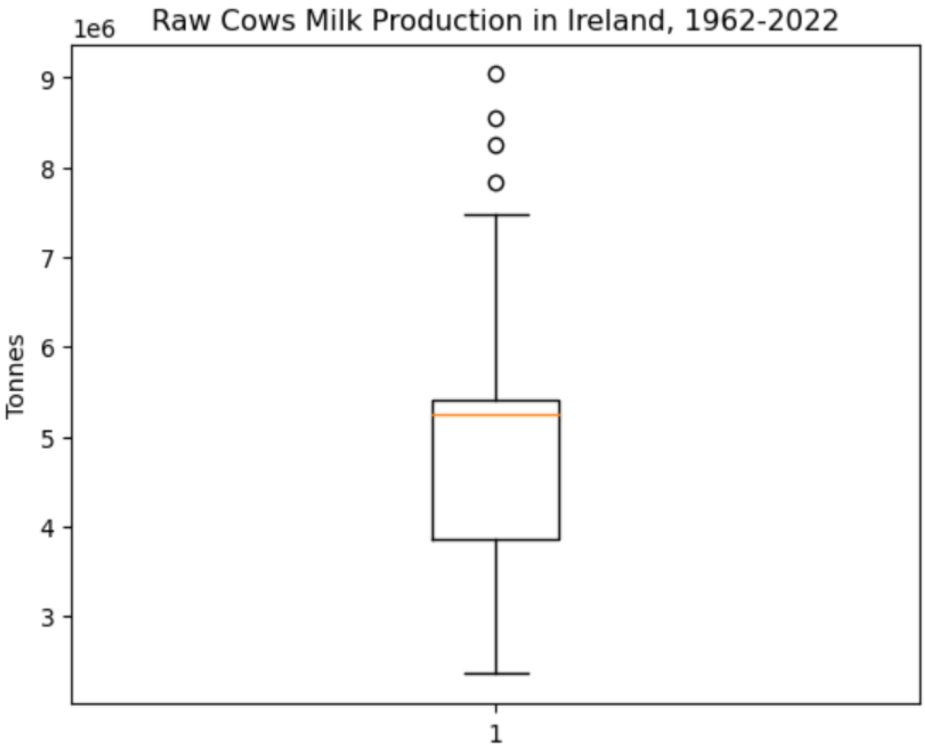


Figure 2: Boxplot detailing the spread of production values of cow's milk in Ireland from 1962-2022.

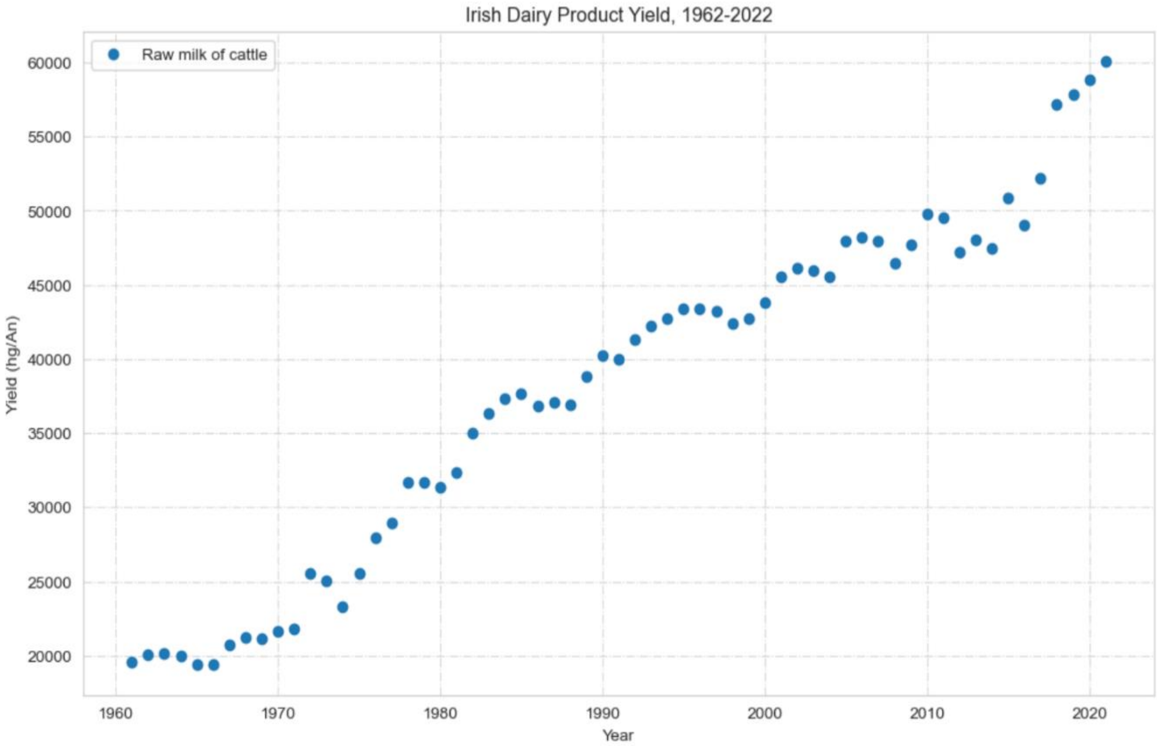


Figure 3: Scatterplot detailing the yield values of cow's milk in Ireland from 1962-2022.

Using the Shapiro Wilk test accompanied by a histogram, it was determined that the distribution of the target data (Irish milk production) was not normal. This was evident upon producing Figure 1 and Figure 4. The EU-imposed quota would have had an impact on the distribution of data given that countries would produce up to the quota cap before it was removed in 2015, as seen by the large number of production results which were around 5 million tonnes of milk or 5 billion litres. Spain's production values showed a similar distribution as seen in Figure 5.

Additionally, this data was determined to be non-stationary using the Augmented Dickey-Fuller test, as expected for time-series data.

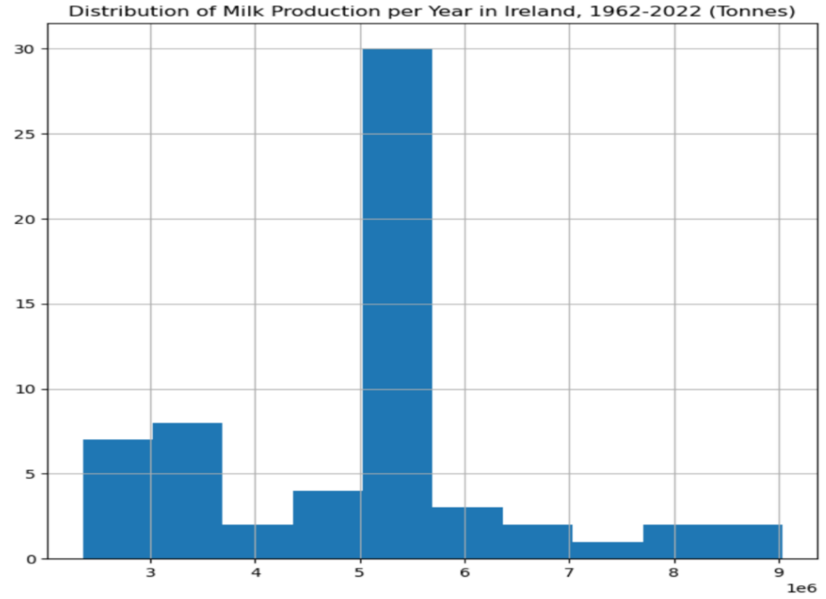


Figure 4: Histogram detailing the distribution of production values of cow's milk in Ireland from 1962-2022.

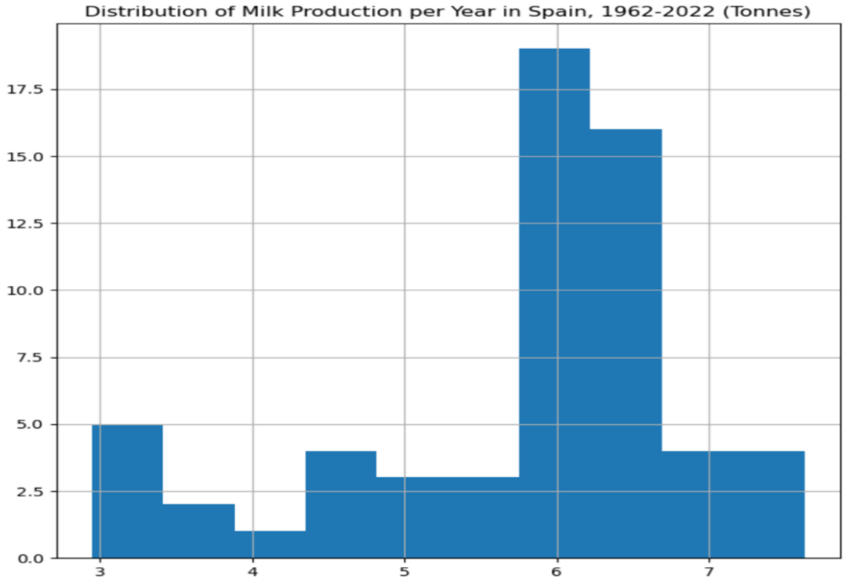


Figure 5: Histogram detailing the distribution of production values of cow's milk in Spain from 1962-2022.

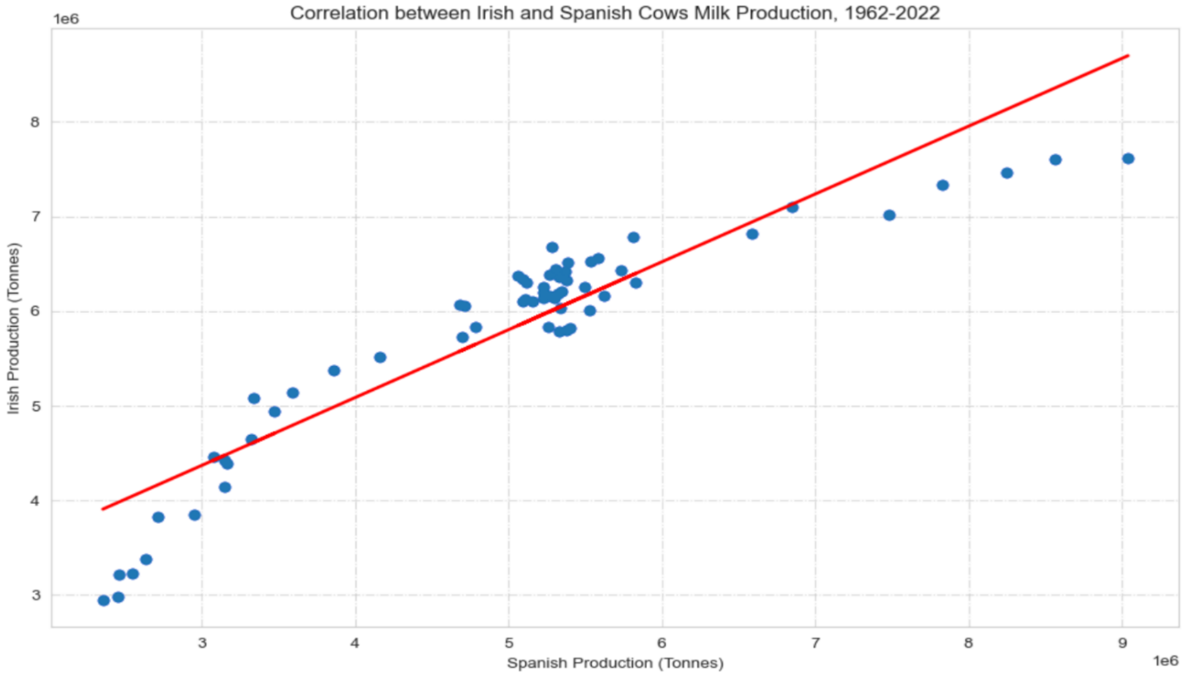


Figure 6: Correlation between Irish and Spanish Milk Production, 1962-2022.

The relationship between Ireland and Spain's milk production levels throughout the years was investigated using a Pearson's correlation test. It was determined that the two features were highly correlated (Correlation Coefficient = 0.927) and that this result was statistically significant (p-value < 0.05). This correlation was plotted and explored visually as a preliminary investigation prior to modelling. Figure 6 details this correlation plot with an applied line of best fit. It is evident that a linear model describes the trend to an extent however it demonstrates non-linear behaviour. Figure 7 reproduces Figure 6 however a third-degree polynomial line of best fit is applied to the data.

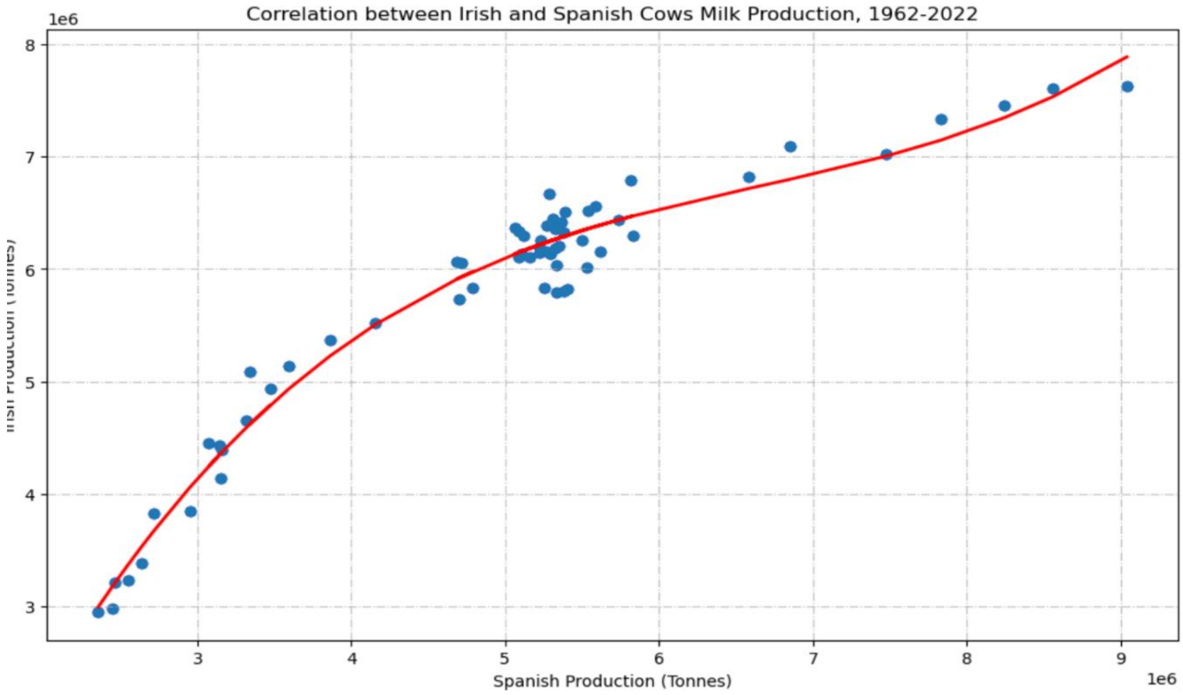


Figure 7: Correlation between Irish and Spanish Milk Production, 1962-2022.

Data describing the selling price of raw cows' milk and greenhouse gas emissions produced via enteric fermentation were sourced from separate datasets and analysed in a similar manner. All data was subsequently parsed and merged into a central dataset to allow for model application. Their correlation with one another was explored using a heatmap as seen in Figure 8 for Ireland and for Figure 9 for Spain.

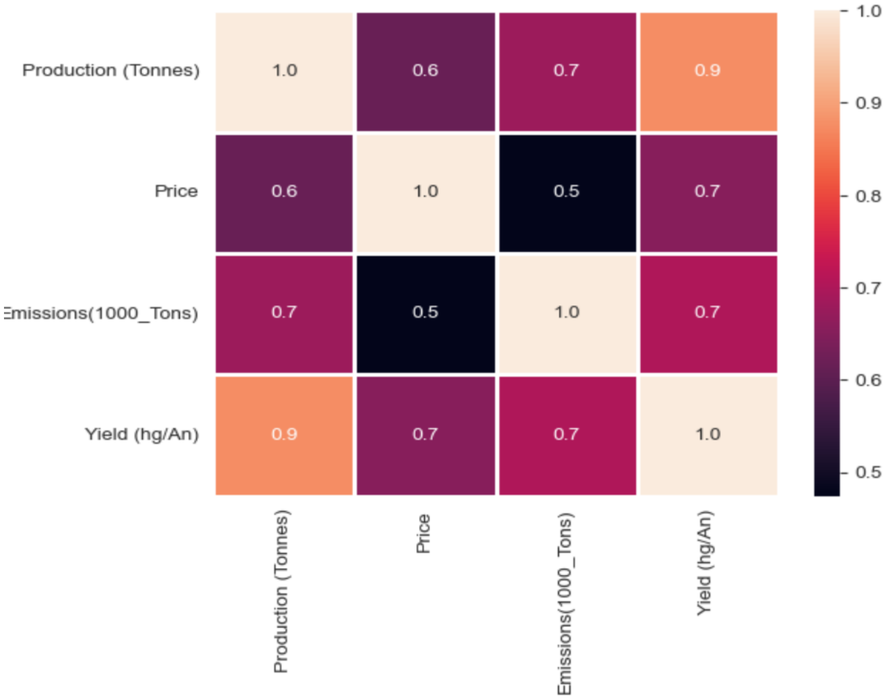


Figure 7: Heatmap of key features influencing milk production in Ireland.

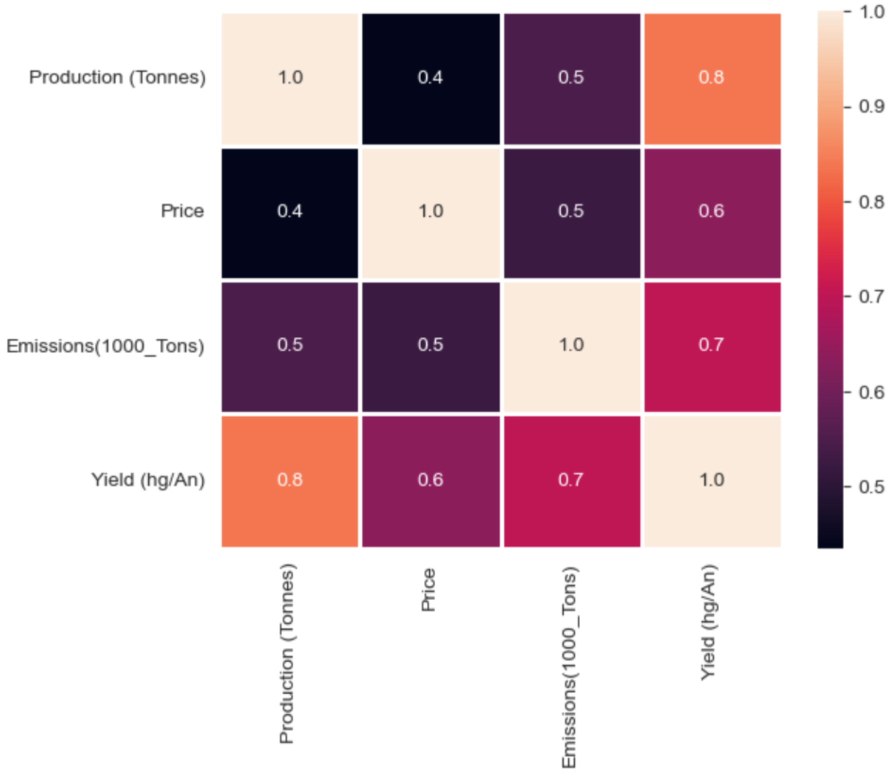
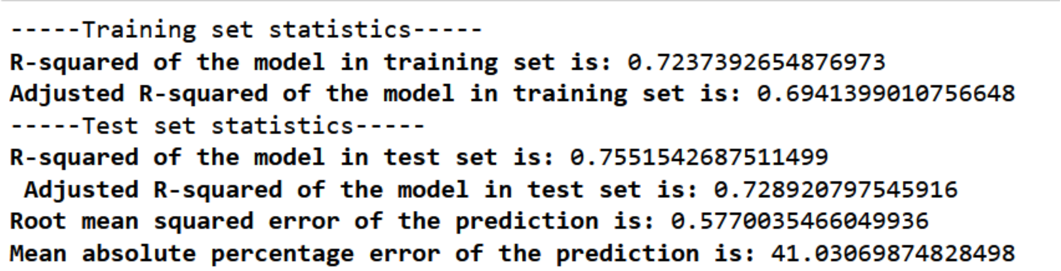


Figure 8: Heatmap of key features influencing milk production in Spain.

Finally, a dashboard was created using the merged dataset and plotly. Plotly is a powerful tool for data exploration and characterisation as it provides the user with insights that may not be immediately apparent.

Modelling

Upon splitting the merged dataset into training and test subsets, a linear regression model was applied given that it was explored previously during data preparation. Figure 9 below details the results of this application.

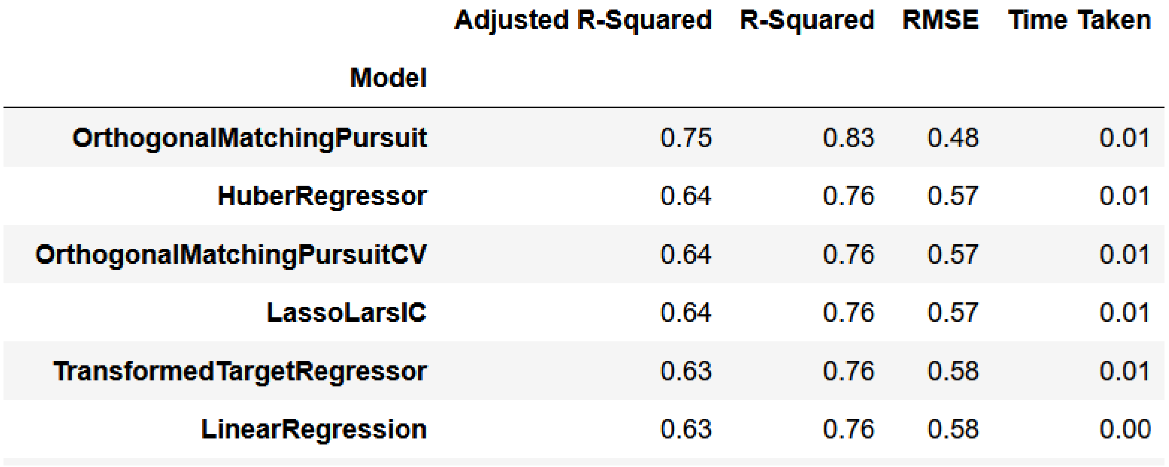


```
-----Training set statistics-----
R-squared of the model in training set is: 0.7237392654876973
Adjusted R-squared of the model in training set is: 0.6941399010756648
-----Test set statistics-----
R-squared of the model in test set is: 0.7551542687511499
 Adjusted R-squared of the model in test set is: 0.728920797545916
Root mean squared error of the prediction is: 0.5770035466049936
Mean absolute percentage error of the prediction is: 41.03069874828498
```

Figure 9: Linear regression results for the production of milk in Ireland

This model was determined to be successful, provided the lack of available. Further validation would be required to confirm this model's suitability if additional data were available in future.

Additionally, a range of models were applied using lazypredict, a Python machine learning library which aids in the exploration and development of machine learning models. This method was applied to efficiently explore the outcome of different model applications to the provided data given limited time and resources. These results are produced in Table 1 below.



	Adjusted R-Squared	R-Squared	RMSE	Time Taken
Model				
OrthogonalMatchingPursuit	0.75	0.83	0.48	0.01
HuberRegressor	0.64	0.76	0.57	0.01
OrthogonalMatchingPursuitCV	0.64	0.76	0.57	0.01
LassoLarsIC	0.64	0.76	0.57	0.01
TransformedTargetRegressor	0.63	0.76	0.58	0.01
LinearRegression	0.63	0.76	0.58	0.00

Table 1: Results of model application using lazypredict library.

Several models were seen to perform similarly on this dataset. This was likely due to the small number of datapoints available for use in this analysis. Gaps in data of key features meant that 32 data points were available. A larger dataset would have likely minimised the success of models such as OMP in modelling the provided data. OMP may be an unsuitable model for this task given that the data collected is time series data. If provided with a larger dataset, one would verify all observations in this analysis and re-apply several successful models listed in Table 1 as verification of applicable models and to eliminate models which happened to operate successfully due to the extremely small dataset.

Sentiment analysis was performed to gain an understanding of public perception and opinion of dairy products and dairy alternatives. Two datasets were created based on sentiment analysis relating to milk and dairy products, and to alternative products like almond milk or soya milk which were the top two milk alternatives in the U.S market as of 2021 (Glanbia Nutritionals, 2021).

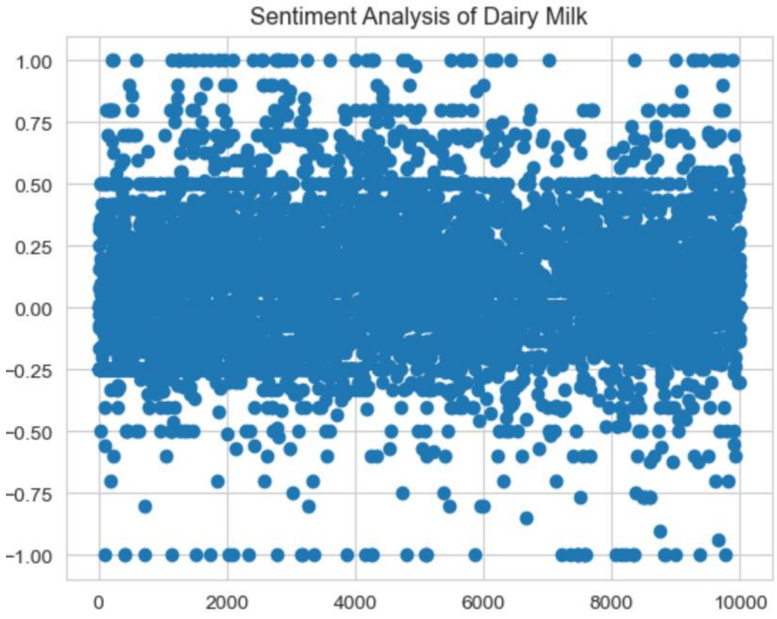


Figure 10: Sentiment analysis of dairy milk, n=10000.

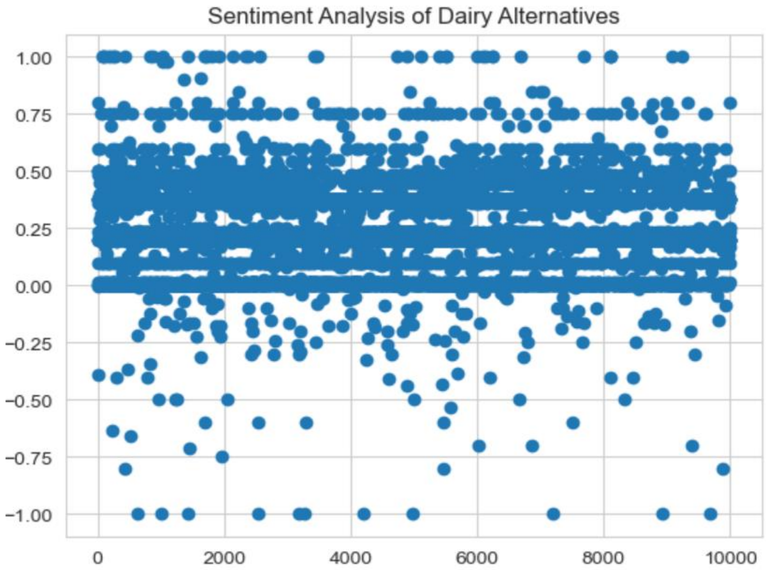


Figure 11: Sentiment analysis of dairy alternatives,

Overall, a more positive sentiment score was produced upon performing sentiment analysis relating to dairy alternatives. The mean sentiment score for dairy-related terms was 0.06, while the sentiment score for dairy alternative terms was 0.25. These results provide an insight into the publics perception and interest in these topics. Due to the nature of the topics in question, sentiment analysis scores likely never reach extremely high or low scores. A more polarising topic that is regularly discussed on public forums like climate change, current political events and current celebrity news would likely produce stronger sentiment scores.

Analysis & Discussion

Business Understanding

The factors and their impact on milk production were chosen for the following reasons:

Yield

Although yield is intrinsically tied to production given that it is defined as the production of milk per unit area used in production through grazing of cows and other means, it also acts as a proxy feature for aspects of farming like improved fertilization methods and more efficient land management leading to more productive land.

Selling Price

Farmers need to make a profit to survive, regardless of size, if they rely on agricultural production as their main source of income. One must explore how correlated the price and therefore the profits are with another aspect of such a business before focusing on this aspect. It is accepted that producing more product means more profit as the quantity available allows a farmer to sell more. However, one may explore the correlation between a more nuanced aspect of dairy production to determine whether selling price is affected by this before investing resources into improving this aspect of their business.

Emissions

The production of greenhouse gasses must be considered when running a farming operation due to the impending implementation of carbon taxes on businesses based on their carbon footprint. It was determined that understanding the relationship between production and a soon-to-be cost of production would be important for business owners. Failing to incorporate the topic of carbon emission in such an analysis would potentially provide farmers with inaccurate data on the market and their position as a business.

Additionally, sentiment analysis was performed to explore the public perception of dairy milk and dairy alternatives given the rise of popularity in dairy alternatives in recent years. As a business, farmers must be able to look forward and based on trends, adjust their business model or practices to optimise and survive.

Data Understanding & Preparation

Data understanding, preparation and cleaning was performed successfully through effective extraction, merging and characterisation of data via visualisations. Statistical analysis and visualisation of production data through plotting allowed for the identification and characterisation of the EU milk quota which drastically altered the behaviour of this data. Additionally, it was possible to identify the constant increase in yield regardless of this production limit which provided valuable insight into the development of the agricultural industry in Ireland as well as Spain. The correlation in production levels of milk in Ireland and Spain was also explored and their performance as similarly sized producers was analysed. In summary, the produced visualisations of available data were determined to be extremely effective in identifying trends and conveying valuable information to both small- and large-scale farming operations.

Modelling

Modelling was performed efficiently using a powerful tool to determine the best fitting model for the provided dataset. Implementing this method of model determination offered additional insight into the data and the relationships between features. It was understood that due to the small amount of available data, the application of machine learning models may have produced inaccurate results. With more data, one could validate the success of certain models suggested as being successful in modelling the provided data.

Regarding sentiment analysis, the produced results were not definitively positive or negative as one may expect, given the broad range of opinions a population will have on such a topic. More robust tuning of the model may produce more significant results.

Evaluation

This analysis successfully explored data relating to the dairy industry in Ireland as well as comparing its performance to that of Spain, its nearest competitor in the EU in terms of production of raw cow's milk. Through effective visualisations in the form of figures, tables and an insightful summary dashboard, insights were delivered to stakeholders describing the relationships of key aspects of successfully operating a dairy farm at all scales.

Limitations to this analysis included the lack of available data meaning modelling and forecasting was negatively impacted. Given additional time, more extensive data collection or a shift in scope would occur to maximise deliverables to stakeholders.

Conclusion

An analysis of the production of dairy milk in Ireland from 1962-2022 was performed to provide dairy farmers in Ireland with valuable information which would allow for data-driven optimisation. Through effective data extraction and cleaning, meaningful information was delivered to such stakeholders through effective visualisations and a simple summary dashboard. With eventual access to additional data, farmers can utilise this report and associated codebase to gain additional insights and continue to improve their operations.

References

Eurostat, n.d. Copyright notice and free re-use of data . Eurostat.

FAO, n.d. Statistical Database Terms of Use [WWW Document]. FAO. URL https://www.fao.org/contact-us/terms/db-terms-of-use/en/ (accessed 12.5.22).

Glanbia Nutrionals, 2021. Which Milk Alternatives Are Most Popular with Consumers? [WWW Document]. URL https://www.glanbianutritionals.com/en/nutri-knowledge-center/insights/which-milk-alternatives-are-most-popular-consumers (accessed 12.20.22).

Jack Kennedy, 2022. Who are the big players in the EU dairy industry? [WWW Document]. Farmer’s Journal. URL https://www.farmersjournal.ie/who-are-the-big-players-in-the-eu-dairy-industry-663914 (accessed 12.10.22).

Läpple, D., Carter, C.A., Buckley, C., 2022. EU milk quota abolition, dairy expansion, and greenhouse gas emissions. Agricultural Economics (United Kingdom) 53, 125–142. https://doi.org/10.1111/agec.12666