Analysis of the Irish Dairy Industry & Comparison With the EU

Faelan Redmond

SBA22190

sba22190@student.cct.ie

Word count: 6500(3000 would have been very tight)

https://github.com/faelanred/MSc-Data-Analytics.git

January 6, 2023

1 Introduction

The Beef and dairy industry in Ireland accounts for our largest agricultural and food related exports. In this project we will be comparing the dairy industry here to that of the top 5 EU milk producers, in doing so we will be looking at factors such as pastureland usage, export quantities, herd sizes, production volume, as well as manure & fertilizer use over the period of 1991-2020. The goal is to gain some understanding into the similarities and differences among the countries mentioned and identify areas where the total milk production could potentially be improved. An array of analytical techniques will be implemented to realize this goal, culminating in the application of regression modelling to try and forecast the milk production of any of the EU countries, including Ireland. Accompanying this project is an interactive dashboard that will launch in a browser window, this will allow readers, producers, or consumers to gain some hands-on experience with the data and explore it for themselves over a range of countries or years, and this will extend its scope to the top 10 producers in the EU. If at any time further clarification is required on a topic or mode of analysis, the jupyter notebook is included & worth taking a look at as I will be unable to explain every detail here due to certain constraints.

2 Data Permissions

To make this project possible, we first needed to gather data related to every EU member state. Due to the volume of data required, and the number of attributes for which we needed data, it would have been rather difficult to consult the agricultural organizations of every country. This is due to both language barriers and issues that would arise later with data homogeneity and consistency. Conveniently, much of the data we need exists in the Food and Agriculture Organization of the United Nations Database(FAOSTAT) and is available from 1961-2020 as of the time it was retrieved, other data was retrieved from the Climate Change Knowledge Portal.

As mentioned above, FAOSTAT will provide us with data from 1961-2020 for many of the countries around the world, for the purpose of this project we were only interested in EU data. The data that we gathered from here includes data regarding the production of raw cow milk, pastureland area, fertilizer nutrient data, export milk prices & volumes, manure use on pastureland, and the prices paid to producers. All data in the FAOSTAT database, unless stated otherwise, is made available for use under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO)(FAOSTAT 2023). Links to

any dataset used from this database will be included in this assignment submission, as well as the CSV files used.
The data acquired from the Climate Change Knowledge Portal concerns the average weather conditions in each country for each year that we have data available. Specifically, we pulled data related to the average rainfall per year in mm/yr, and the average recorded temperature per year in Celsius, C. This data is not for looking at climate change as that is clearly stated as being disallowed in the project brief. But we will be looking at these factors as they may have an impact on pastureland development, and therefore milk production metrics down the line due to varying nutritional content. Like FAOSTAT, the Climate Change Knowledge Portal data is made available for use unless stated otherwise, under a Creative Commons Attribution 4.0 International License (CC BY 4.0)(bank 2018). Links to the datasets & CSV files used will accompany this report.

3 Analytics Overview Methods

3.1 Data Preparation Visualization

The first objective of this project was to merge all our data into a single dataframe. Thankfully, this process was smooth because FAOSTAT has a consistent data format across its database. We used the raw milk quantity data, "dairy_df," as a base and added additional columns to it. When there were multiple attributes in a single column, we split them into separate dataframes and merged them one at a time. We defined a custom function to speed this process up. In some datasets, rows were missing for every year between 1961 and 1990, most notably in the export data. In this case, we dropped null values and used data from 1991 to 2020 only. In other cases, where there were a few missing years, we used data imputation. To ensure that the distribution of the data remained unchanged, we used a non-parametric Wilcoxon test to compare the distributions before and after imputation. We chose this test because it assumes that the distributions are dependent and continuous, which was the case for our data.
After creating the main dataframe, we made a copy and added calculated columns to it for further analysis. These columns included conversions from tons of milk to liters, calculations of milk produced per cow, calculations of the percentage of land used as pastureland in each country, and export value per liter. These columns allowed us to perform simple analysis and create easy-to-understand visuals for our exploratory data analysis (EDA). We used the seaborn library to create most of our plots because of its widespread industry use and simple theme. We also used a seaborn theme for any matplotlib plots to ensure consistency. Our plots featured gridlines for easy reading of axis values and a "tab20c" color palette to clearly differentiate EU member states.
Finally, we created a Python dashboard to provide a portal for interactive, easily digestible data on the milk production of the top 10 milk producers in the EU. We used the holoviz library and an interactive dataframe wrapper to create interactive plots that could be filtered by country and/or year. We considered similar design elements as before, including the same color palette and gridlines, prominent markers at each data point, consistent country colors, and large, easily readable plots for those with visual impairments. To host the plots and easily serve them, we used the panel library.

3.2 Machine Learning

After preparing and analyzing our data, we were ready to perform predictive analysis. We decided not to include the calculated columns from the previous section in this analysis. We wanted to see if we could use only the data from official sources and minimal preprocessing to build a regression model that could predict the milk production in each EU member state per year. To address multicollinearity between the feature attributes, we used principal component analysis (PCA). We chose a number of principal components that would explain at least 95% of the variance in the data, which was 25 in our case.

To find the best models for our data, we trained eight different models and compared them. We used RandomizedSearchCV to tune the hyperparameters of these models because it is faster than GridSearchCV and still gives us a high probability of finding the best combination of results. The number of iterations used was 500, and the calculations below show our probability of coming within the top 1% of parameters using this method. The probability of falling outside of this interval is,

$$(1\text{-}0.01)^{n}$$

So, the probability that at least one of our models is in the interval desired is,

$$1\text{-}(1\text{-}0.01)^{n}$$

In our random search model we had the number of iterations set to 500 so,

$$1\text{-}(1\text{-}0.01)^{500} = 0.9934$$

With the parameters chosen, we expect to be within the top 1% of models with a 99.34% chance of success. This is a good enough justification for the use of this model and saving on the vast computation time GridSearchCV would require.
Next, we performed sentiment analysis on tweets related to milk and dairy prices using the Twitter API. We prepared the tweets by removing usernames, punctuation, special characters, stop words, and suffixes, and making them lowercase. We used NLTK, textblob, and reg expressions for these tasks. We then labeled the data as positive, negative, or neutral based on sentiment scores and encoded it. The class labels were unevenly distributed, so we used SMOTE to oversample minority classes and balance the distribution. We then created a bag of words model using the sklearn count vectorizer and applied a dictionary of classification models to the data. We tuned these models using RandomizedSearchCV and evaluated them using accuracy, precision, recall, and f1-score.

3.3 Statistics

We calculated descriptive statistics for every section of our data, for each country. To automate this process, we wrote a loop that checked for normality using a Shapiro-Wilks test and calculated the appropriate descriptive statistics accordingly. We also defined a function that allowed us to specify a country and attribute of interest and return a plot based on whether the data was normal or non-normal. This significantly sped up the process and means that all the data is available in dictionaries.
We also calculated inferential statistics for the entire EU, including confidence intervals for the proportion of the population with over 15% pastureland, the median head of cattle in the EU, and the median volume of milk produced in the EU. We chose to form confidence intervals around the median because the data was not normally distributed and using the mean would not accurately represent it.
Finally, we used statistical tests and hypothesis testing to gain further insight into our data and compare Ireland's milk industry to that of France and The Netherlands. These tests included a Wilcoxon test for data before and after imputation, the Shapiro-Wilks test for normality, the Mann-Wittney-U test, Friedman Chi-square test, and the Kruskal-Wallis test. Rationale and justification for each will be given below.

4 Results Discussion

4.1 Exploratory Data Analysis

As we are primarily concerned with the amount of milk produced per year in each EU member state, we would first like to get a feel for the productivity of each country, this way we can narrow our scope of

analysis ang have a good idea of Irelands standing in the rankings. To do this, we can simply produce a barplot of the milk production in litres per country in ascending order.

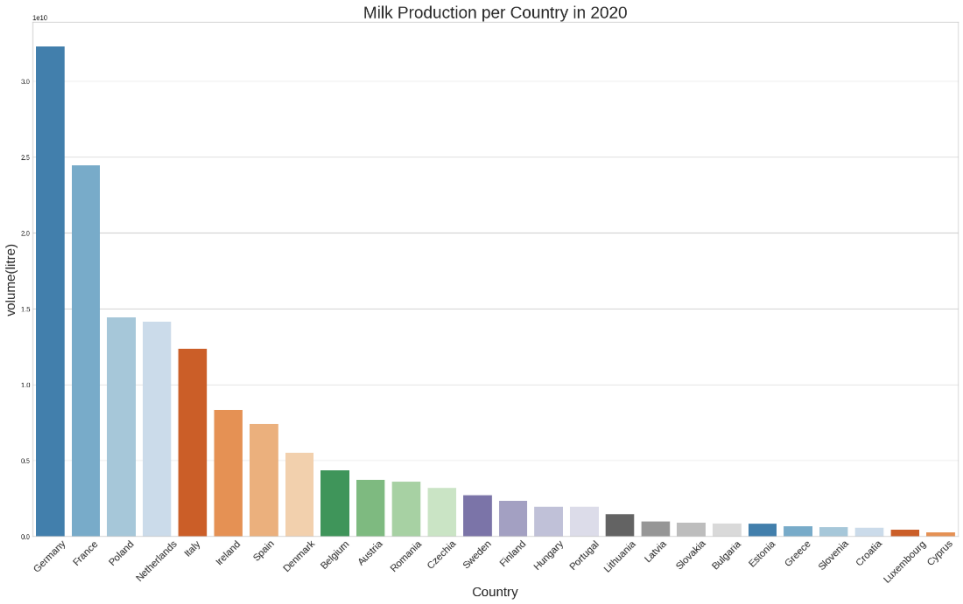


Figure 1: milk production volume by country in 2020

We can see here that there are countries that are clearly outperforming others here, by quite a large margin. While Germany is the clear strong performer here, producing 30% more milk than France, Ireland manages to hold its own here, sitting in sixth place. This isn't so surprising, considering that beef and dairy are Ireland's largest agricultural exports(Bordbia 2021). It's interesting that the reduction in milk production across the plot is so rapid. The answer to this could be rather simple, and the countries that produce the most milk just might have more cows than those that produce the least.

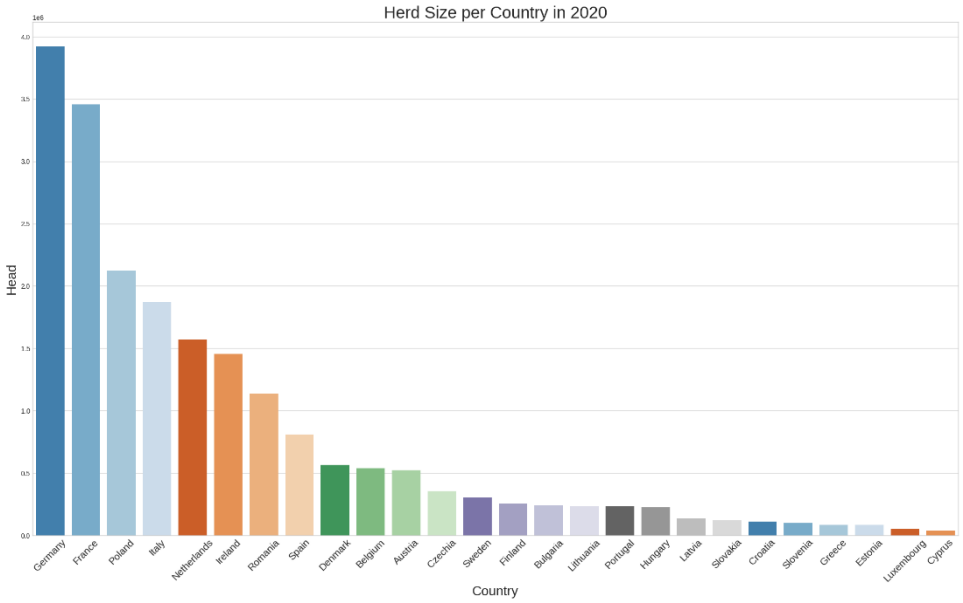


Figure 2: Herd size by country in 2020

As seen above, there is a bar plot showing the number of cattle in each country's collective herd. The plot indicates that the number of cattle is almost directly related to the volume of milk that we would expect

the country to produce. The countries with the highest and lowest milk production also happen to have the highest and lowest number of cattle. The countries in between mostly have the same positions on each plot, with a few exceptions. The correlation coefficient between these attributes is 0.98, which suggests that we could almost use this as a predictor for the milk production in each country in a simple linear regression model.
Moving on from here, we will be looking only at the top 6 producers in the EU. This allows us to keep the following visualizations clean & legible but will also allow us to compare Irelands attributes to those of the top 5 EU countries for milk production. First, we want to look at how the milk production for these 6 countries has changed over the last 29 years.

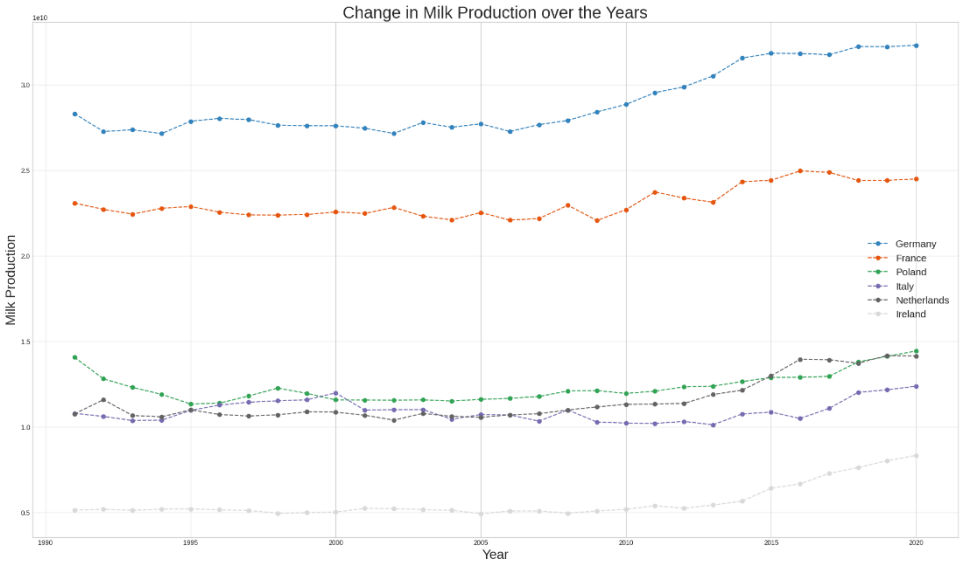


Figure 3: Change in milk production per country 1991 -2020

From the plot, we can see that Ireland's milk production was stable for many years until 2014, when it began to steadily increase every year. It is possible that this sudden increase is due to reforms in the Common Agricultural Policy (CAP) in 2015, such as the abolition of milk quotas in the EU which had been in place since the late 70s(Läpple, Carter, and Buckley 2021). Another notable feature is Germany's increased milk production since 2007. It is difficult to pinpoint the exact cause of this change, as there were no notable changes made to the CAP in 2007. This increase could be due to increased demand for dairy products or improvements in Germany's dairy systems. Further investigation may be required. Aside from these notable increases, the other countries have seen only marginal growth over this period.

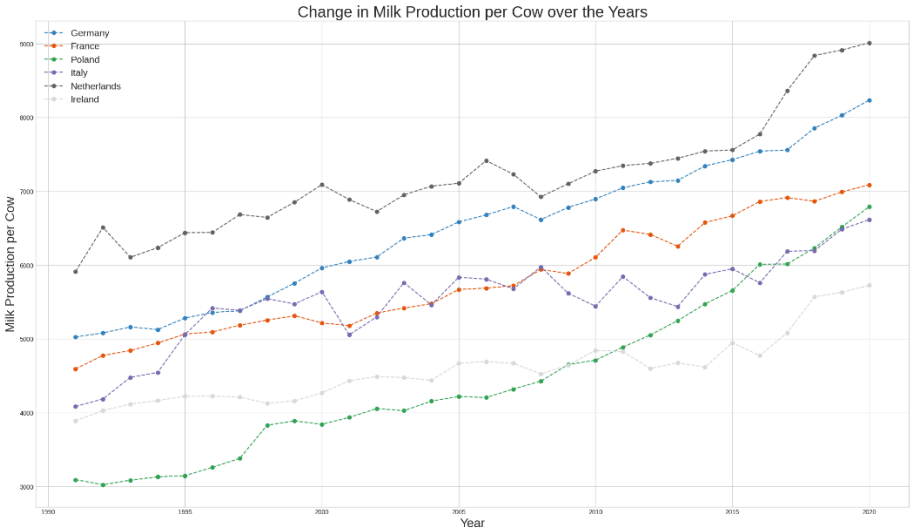


Figure 4: Milk production volume per cow 1991 -2020

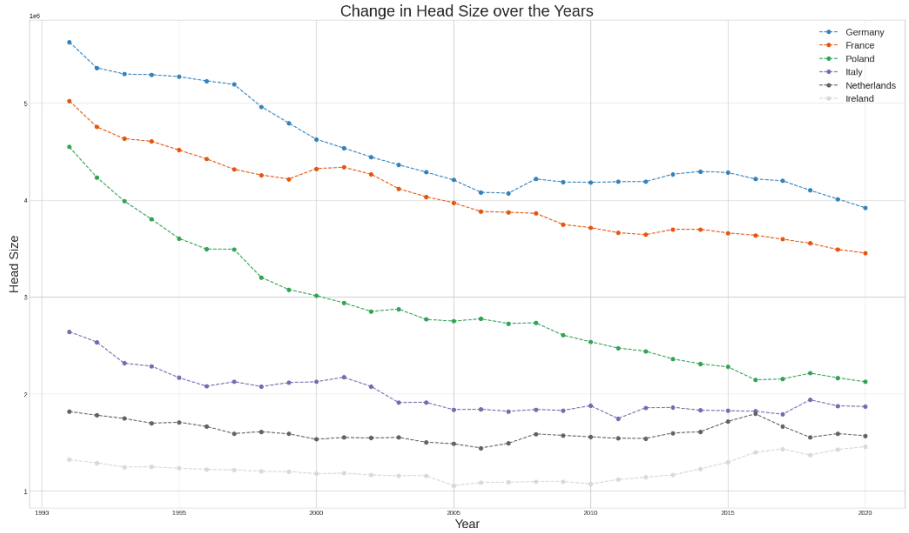


Figure 5: Herd size per country 1991 - 2020

Having looked at the change in milk production over the years, we then went on to look at the change in herd size through the years also. In figure 5 we can see the change in herd size for each country over the last 29 years, it is observed that the herd size in Ireland has seen an increase since 2010, meanwhile the herd size of Germany, France, and Poland has steadily decreased since then. This herd size increase may also be a factor in the increasing milk production in Ireland since 2014 that we saw above. The top three milk producers are actively decreasing their herd sizes year on year. This is explained when we go on to look at the productivity of cows in these countries in figure 5, we see that every country has seen an increase in the productivity of cows over the last 29 years, meaning that each cow produces more milk. The increase seen varies per country, however cows in Ireland in 2020 on average are producing 1.47x the amount of milk they were in 1991, this is the lowest seen out of these countries, meanwhile cows in Poland have the largest increase at 2.2x & the average is 1.67x. From an economic point of view this makes sense, cows are being selectively bred to produce as much milk as possible in doing so this allows farms to reduce their herd size while still producing the same volume or more milk, this allows farmers to save money on feed, housing, and the cattle while maximizing their profits.

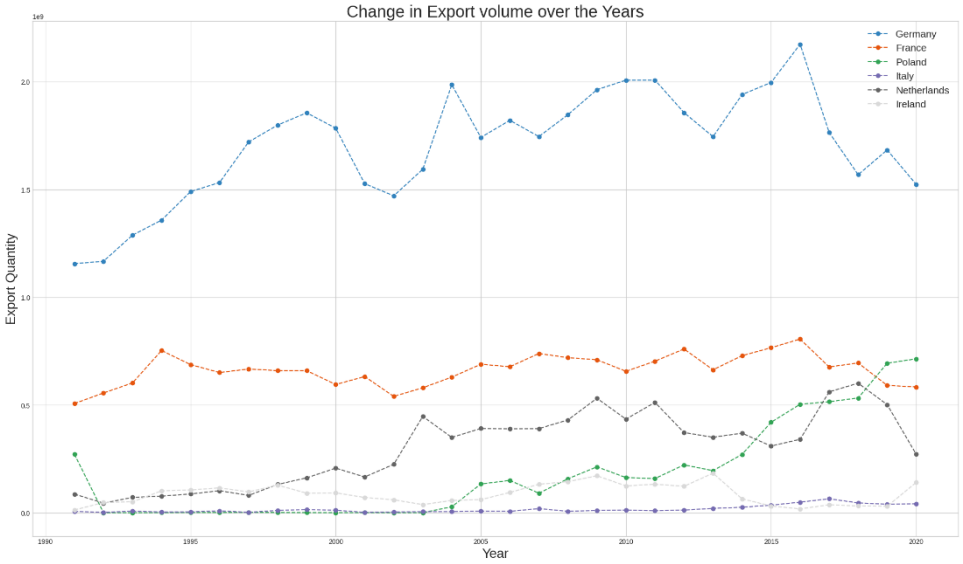


Figure 6: Trend plot of export volumes per year

The exported volume of raw milk is an important factor to consider, and it likely affects the total milk production in each country per year. In 2020 and previous years, the top three spots for milk exports were taken by Germany, France, and Poland. However, in previous years, the Netherlands outperformed Poland in this regard, despite their lower milk production between 1991 and 2015. While Ireland's dairy exports are very large (140807429.406l in 2020), they and Italy's/Poland's are almost overshadowed in this plot due to Germany's astronomical volumes of milk exports. Germany is consistently the highest exporter by far, although there are large variations in specific volumes year to year. Without further research into Germany's dairy industry, it is difficult to pinpoint the cause of this, but it may be as simple as Germany being more sensitive to fluctuations in the global market due to their very high exports.

One major factor that is likely to impact the quantity of milk being produced is the food that the cattle consume. While specifics vary by country, pastureland grazing constitutes a significant percentage of a dairy cow's diet. In addition to a pasture-based dairy system being more beneficial for the cattle than a housed or feedlot system, due to the ability for the animals to roam freely and exhibit natural behavior, thereby reducing the chance of diseases such as mastitis, there is also an apparent increase in the quality of the milk from these animals due to higher concentrations of beneficial fats in the milk that are beneficial to humans(Wedzerai 2020).

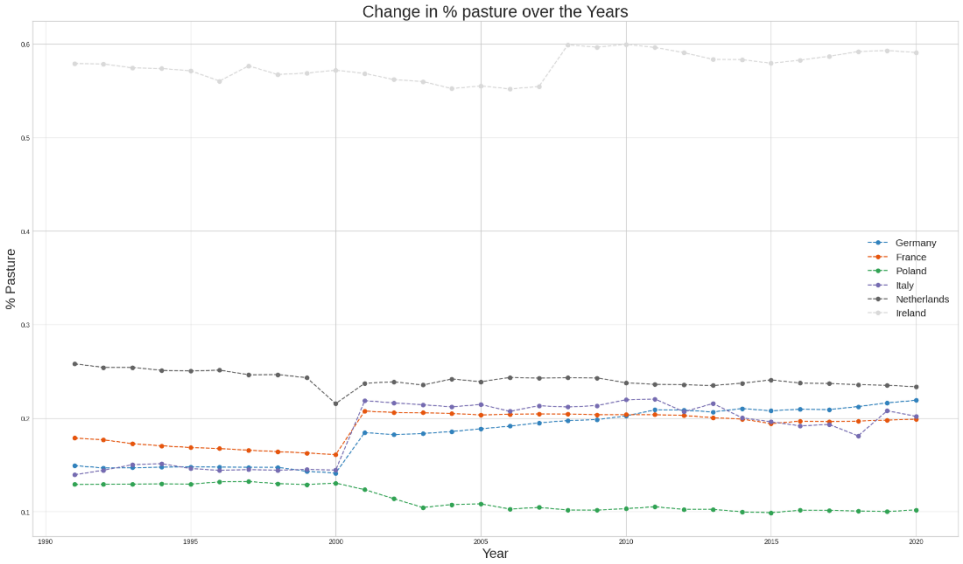
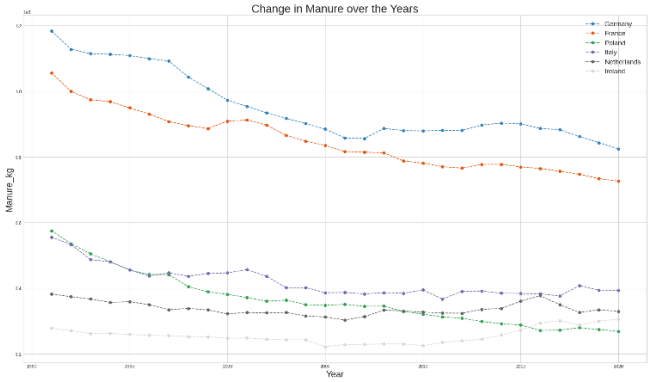
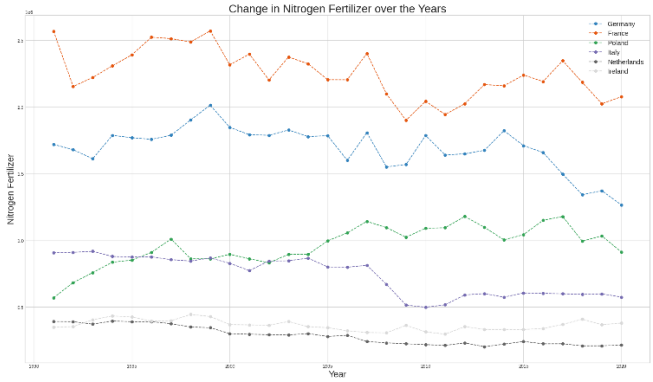


Figure 7: Change in percentage of pasture land

From the plot above, we can see that Ireland has roughly twice the percentage of pastureland compared to the top 5 milk producers and any other country in the EU. This makes sense considering that Ireland has historically employed pasture-based dairy systems to a greater degree than any other country in the EU(Wedzerai 2021), possibly due to Ireland's ideal weather conditions, which are wet and temperate, allowing for consistent grass growth throughout the year. Every other country, aside from Poland, has increased their use of pastureland since 2000. This change may be due to amendments to CAP that allow a payment per animal or per hectare of permanent pastureland. The decrease seen in Poland could be due to changes in land management locally, a preference for other dairy systems, or urbanization of this land.

A key factor in keeping pastureland for dairy production is ensuring luscious growth of healthy, nutritious grass. One major factor in this is the use of manure and modern fertilizers to supplement nutrients drawn from the ground. The most commonly used fertilizers for this purpose are manure, nitrogen, potash, and phosphate fertilizers.





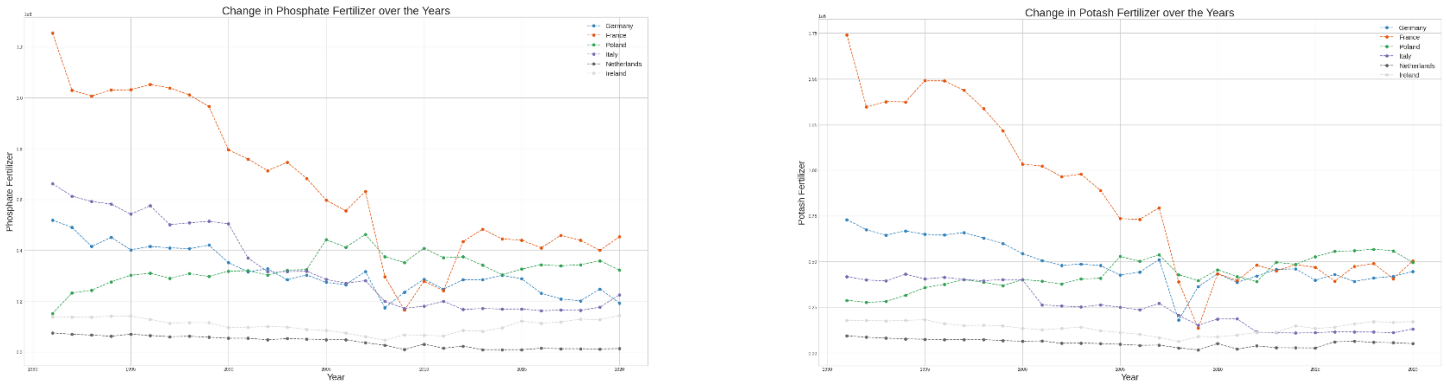


Figure 9: plots showing manure usage, nutrogen usage, phosphate usage, and potash usage

The first plot above is a trendline showing the progression of manure use on pastureland over the years. It shows a clear decrease over time in the use of manure in the top two countries, Germany and France, with it reaching an all-time low in 2020. The rest of the top 5 countries show a similar increase, but at a slower rate. It is likely that these countries are moving away from this natural fertilizer and favoring more modern nitrogen-phosphorus-potassium (NPK) products. These allow farmers to tailor fertilization to specific elements missing in the soil, ensuring the best quality and highest volume of grass. Ireland's manure use has increased in recent years, possibly due to its availability and the fact that buying in NPK fertilizers may not be economically viable given the amount of pastureland. The next three plots show the total kilograms of each elemental fertilizer used per year. The plot for nitrogen shows a relatively stable amount used, with fluctuations from year to year indicating that it is commonly used everywhere, but required amounts vary based on soil conditions. The plots for potash and phosphate show a similar trend, with the amount used in France decreasing rapidly before reaching equilibrium around 2012 or 2010, respectively. Elsewhere, the amount of these fertilizers used has decreased slowly over the years, with only a decrease seen in Poland. These plots may appear to contradict the earlier statement about NPK products but considering the advancements in science and its adoption in agriculture, this trend may make sense. Fertilizers were often used in large quantities without justification, but now we know exactly how much is needed, which saves money and reduces environmental impact.

4.2 Statistics

Due to the constraints of this report, it is unrealistic for us to include descriptive statistics for every attribute in the countries of interest. These can be found stored in our notebook as a dictionary, but for now, we will focus on the quantity of milk produced in these countries because we are interested in examining this variable further in the machine learning section.

Country	Median	Q1	Q3	IQR	Max	Min
France	23352660.5	23016987.25	24270426.0	1253438.75	25627060.0	22653085.0
Germany	28679081.0	28331965.00	31161217.0	2829252.00	33164910.0	27866180.0
Ireland	5326550.0	5235300.00	5571922.5	336622.50	8561470.0	5061250.0
Italy	11120451.5	10687244.50	11526348.5	839104.00	12712480.0	10397465.0
Netherlands	11229955.0	10990575.00	12134942.5	1144367.50	14555000.0	10677000.0
Poland	12419539.0	11937681.75	13111259.0	1173577.25	14821820.0	11642395.0

Running a Shapiro-Wilks test on these 6 countries reveals that all their distributions of milk quantity through the years are non-normal. Therefore, we decided to calculate a 5-figure descriptive statistic,

including the median, first quartile, third quartile, maximum, and minimum values. Using these, we can discuss the distributions in the data, however, our task is made easier if we first plot them as a box plot.

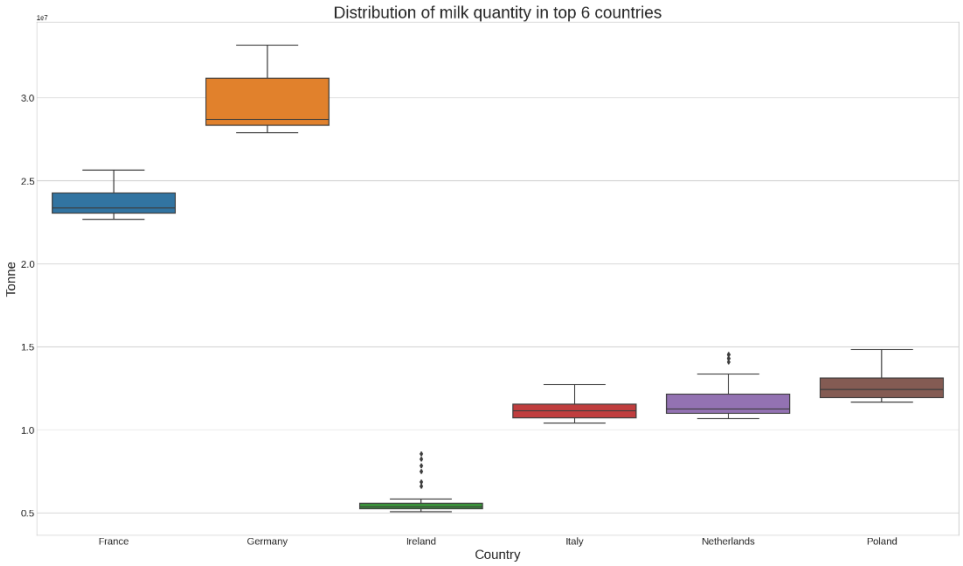


Figure 10: Visualisation of descriptive statistics

These box plots give us a much better idea of the distribution of tones produced. Ireland has the lowest median output by far, at 5326550.0 tones, but the spread around this value is very low in comparison to the other countries when we look at the interquartile range of 336622.50 tones. The exception to this low spread is seen when we observe six years' worth of outliers in the data. This reflects the observed rise in milk production from 2015 to 2020, and I think this represents a new, more profitable era in Irish dairy. Moving on from the lowest producer in this group to the highest, we see that Germany has quite a spread in its data. The median in this case is 28679081 tones, with the highest spread seen in the group, with an IQR of 31161217 tones. It is observed that most of this spread is seen between the 50% and 75% quartiles, though. This tells us that there was a relatively high degree of variability in the quantity of milk produced, but the lack of outliers also tells us that there were no years in this period where the output was uncharacteristically high or low. The rest of the countries in the group all tell a similar story to one another. Italy, the Netherlands, and Poland all have a moderate spread around the median, with a slightly higher spread in values between the 50% and 75% quartiles again, so they have what I would classify as a moderate variability in production values, specifically those above the median. The Netherlands, similar to Ireland, does exhibit three outliers between 2018 and 2020, inclusive, although it's difficult to say if this is a feature observed in the data due to a change in policy or just an anomaly. More data in following years will be required to gain a better understanding of this.

The dataset we put together using FAOSTAT data, while it contains data for 26 EU member states, does lack data for Malta. Due to this fact, what we do have is considered a sample of the population (EU), and as such we can employ confidence intervals (CI) to gain some insight into where the population parameters may lie. The three main areas we identified where it would be nice to have a population parameter for are the proportion of the population that has over 15% of its land dedicated to pastureland, the head of cattle, and finally the volume of milk produced.

The first population parameter that we calculated was a CI around the number of member states that have more than 15% of their land dedicated to pasture. We first calculate the sample proportion that meets these criteria.

$$P_{sample} = \frac{14}{26} = 0.538$$

$$P_{population} = 0.538 \pm 1.96\sqrt{\frac{0.538\,(1-0.538)}{26}} = 0.347, 0.730$$

From this, we can say to a 95% level of confidence that the proportion of the population with a pastureland coverage of greater than 15% lies between the values of (0.347, 0.730). If we wanted to increase the confidence interval, say we wanted to be 99% confident that the true value lies within the range, then this range of numbers would end up wider as a result.

Following this, we calculated the confidence interval around the median head of cattle in the EU. The median was chosen here as a Shapiro-Wilks test revealed that the distribution of herd sizes is not normal, making the mean an inaccurate measure. This was calculated as follows(Zach 2021),

$$CI = \frac{n}{2} \pm z\sqrt{0.5\,(1-0.5)\,n}$$

$$CI = \frac{26}{2} \pm 1.96\sqrt{0.5\,(1-0.5)\,26}$$

$$CI = 8, 17$$

This tells us that there is a 95% chance that the median herd size of the EU population lies somewhere within the 8th and 17th values when they are lined up in ascending order. So, in this case, we can say with a certainty of 95% that the population mean falls between 226000 and 565000 cattle.

The final confidence interval we calculated is that around the median volume of milk produced in the EU each year. Similar to above, the median was the parameter chosen here as the Shapiro-Wilks test revealed that the population was non-normal. The calculation for this is identical to above, so we can skip that step as we already know that the population proportion will be found between the 8th and 17th value when listed in ascending order. So, locating these values tells us that we can be 95% certain that the population mean lies within the range of 2343252190.847 to 435579357.352 litres of milk.

Hypothesis testing, or statistical tests, is very important when it comes to comparing data from various distributions. It gives us a definite, quantitative approach to determining whether two samples come from the same distribution. This is particularly important in this case, where we are comparing various attributes among EU countries. We have already employed the Wilcoxon signed rank test during the data preparation phase when comparing the distributions of imputed data. This test was chosen because it assumes dependent samples and continuous data(Wikipedia$_{C}ontributors$ 2022), which we will see when comparing data pre and post-imputation. The null hypothesis for this test is that the median difference is zero between the samples, and the alternative hypothesis is that the difference is not zero. We reject when $p<0.05$.

The Shapiro-wilks test has been used many times during this project to test the data for normality. The null hypothesis of this test is that the data is normally distributed, meanwhile the alternative hypothesis is that the data is not distributed normally. We reject the null hypothesis on this test when $p<0.05$.

We decided to compare the distribution of herd sizes in Ireland, the Netherlands, and France over the years as the mean values were relatively similar. Initially, we had intended to use an ANOVA test, however the Levene test returned $p<0.05$ indicating that the variances were unequal. Instead, we used the Friedman Chi Squared test. This was suitable due to the independent populations and the data being non-normal(Contributors 2022). The null hypothesis for this test is that the rank order of the treatments is the same across all populations, while the alternative hypothesis is that they are not the same. Upon implementing this test, we got a p value of 9.35e-14 which is far in the rejection region as it is less than the significance level of 0.05. This tells us that the distribution of herd sizes over the years for these countries

is not comparable.

We then looked at the volume of milk produced in these countries. Again, comparing France, Ireland, and the Netherlands, these distributions were not normally distributed. This meant that we again need to use a non-parametric test. This time round we decided to use a Kruskal-Wallis test, this was suitable as the samples are independent and they are not normal(Scholtens et al. 2016). The null hypothesis of this is that the medians are all equal, meanwhile the alternative hypothesis is that at least two of the medians are different. Upon running this test, we got a p value of 6.594e-18, this also falls far into the rejection region as $p<0.05$, se we are unable to say that the medians of these samples are the same and therefore they do not come from the same distribution.

The final comparison we did was between the percentage of pastureland in Ireland and France. This data was also not normally distributed, so we needed to use a non-parametric test. We decided to use a Mann-Whitney-U test, which is suitable for continuous variables, assumes non-normal distributions, and is for independent samples(Cardoso-Moreira and Long 2010). The null hypothesis for this test is that the two samples are equal, while the alternative hypothesis is that they are not equal. Upon running this test, we got a p value of 3.02e-11, which falls well within the rejection region of $p<0.05$. This means that the distribution of pastureland use is not comparable among these countries.

4.3 Machine Learning

The target variable for this machine learning project is the quantity produced per year in a given country. This target variable is continuous, so we needed to use a regression model. We used RandomizedSearchCV to find the best model for our use by tuning the hyperparameters of 8 regression models. For evaluation, we chose to use the median absolute percentage error (MAPE). This was because a few outliers in our test set predictions were far enough off that they heavily influenced the mean absolute percentage error, which is sensitive to outliers.

Model	Best parameters	R2	MAPE
KNeighboursRegression	n_neighbors=2, weights='distance'	0.999032	2.884571
DecisionTreeRegression	ccp_alpha=0.82828	0.995529	3.565786
GradientBoostingRegressor	max_depth=5, max_features='auto', n_estimators=90	0.991133	4.223435
RandomForestRegressor	criterion='absolute_error', max_depth=13, max_features='sqrt', n_estimators=480,	0.997138	5.080359
LinearSVR	C=3.88889, epsilon=2.42242, loss='squared_epsilon_insensitive'	0.994192	8.471146
Lasso	alpha=686.86869	0.994223	8.519573
Ridge	alpha=0.01109, solver='lsqr'	0.994099	8.785302
LinearRegression		0.994135	8.834915

The above table lists the models used, their best parameters, and evaluation metrics for each. We notice

that in some cases the R2 value and MAPE don't match each other. This means that if one model has a higher R2 than another, we would expect the MAPE of that model to be lower than the other model. However, in some cases here, this does not happen. The issue prevailed across the other evaluation metrics we tried to calculate. This may be because R2 only measures the amount of variance in the target that is explained by the model and does not focus on the absolute size of the errors(investopedia 2023). Due to this, our evaluation metric of choice and the table above are sorted by this in ascending order. Based on this, the top three models are KNeighborsRegression, DecisionTreeRegression, and GradientBoostingRegressor. Tree-based models often perform very well at fitting themselves to the data at hand, but there is a risk of overfitting. The use of ensemble-based models or a ccp_alpha pruning parameter is a great way to mitigate this and still ensure high accuracy. The KNN-based regressor here is the top performer. This is because it can "smooth" out noise in the data through the process of aggregating the predictions of close-by points, which will help to reduce the overall error of the model.

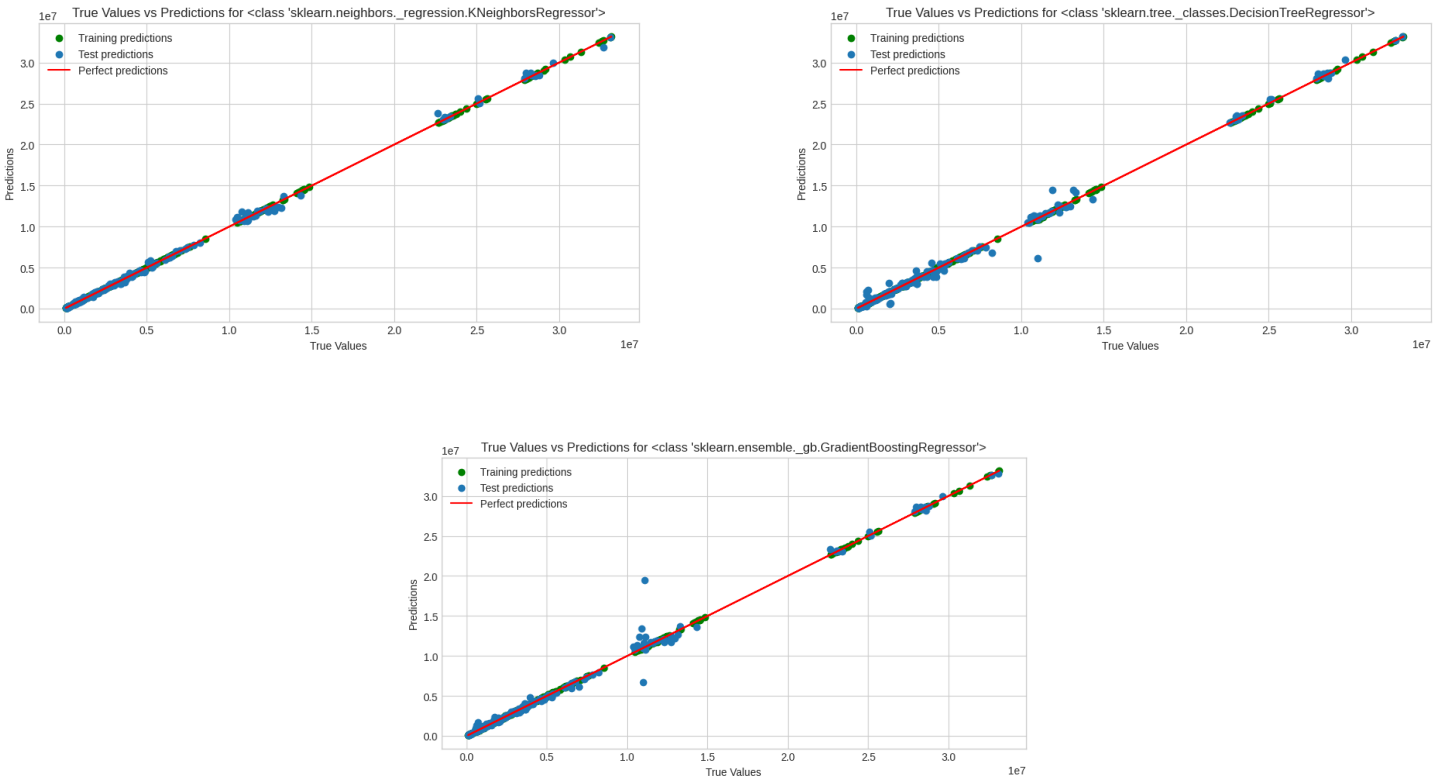
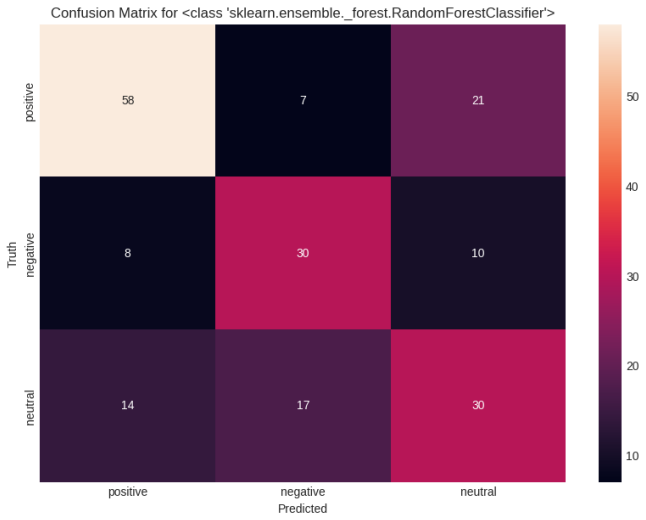


Figure 12: Visualisation of regression results for KNN regressor, Decision tree regressor, and Gradient boosting regressor

The plot above shows the results of these models. One thing that is clear is the lack of outliers in the KNN model, which is where the 'smoothing' effect we mentioned comes into play compared to the other models. The outliers located between 1-1.5e7 on the x-axis are clear deviations from the known values in our models and are likely the culprits that cause the higher MAPE score. These plots can be found in the included Jupyter notebook for every model.
The next task was to carry out natural language processing (NLP) on data collected from Twitter using the Tweepy API regarding the price of milk or dairy. This has been discussed above, and more context on the problem will be given in the code provided. Basic preprocessing was carried out, as specified above, so that we had a dataset containing clean text of tweets and an associated sentiment.

Figure 14: Word cloud broken down by sentiment

Seen above is a word cloud containing the most used words in each of these three sentiment groups. As we can see, many of the same words appear in each group, although this may not matter much as we are more interested in the way in which they are used. In this dataset, negative/positive/neutral sentiment accounts for 24.8%, 42.6%, and 32.6% of the total data, respectively. This indicates that over 75% of the tweets analyzed have positive or neutral opinions of milk prices, and only 24.8% have expressed dissatisfaction. This is rather surprising, as one might expect there to be more negative data related to the topic given the recent price hikes. SMOTE was used to balance out these classes before classification models were employed, although this will decrease the overall accuracy, it will improve the recall, which is a worthwhile tradeoff as we can better predict minority classes(Korstanje 2021). A bag of words model was also created by passing the data through a count vectorizer. The classification models applied were SVC, logistic regression, random forest, multinomial naive Bayes, and a gradient boost classifier. None of these performed particularly well, although the best performer of the bunch was the random forest classifier.



	precision	recall	f1-score	support
0	0.72	0.67	0.70	86
1	0.56	0.62	0.59	48
2	0.49	0.49	0.49	61
accuracy			0.61	195
macro avg	0.59	0.60	0.59	195
weighted avg	0.61	0.61	0.61	195

Figure 15: Random forest classifier results

The figures above show both the confusion matrix & classification report for our random forest classifier. Although this model boasted the greatest accuracy score, it is only 0.61 which is far from ideal, this tells us that of all the predictions made on the test data, only 61% of them were predicted correctly. Although we did apply SMOTE to this, the precision of the model appears to be equally poor. The precision tells us the proportion of positive classifications that were actually correct, for positive, negative, and neutral sentiment the precision score is 0.72, 0.56, and 0.49 respectively. So only 72% of positive predictions are correct, only 56% of negative predictions are correct, and only 49% of neutral predictions are correct. The recall tells us how many of the actual positives were identified, for positive, negative, and neutral sentiment this value is seen to be 0.67, 0.62, and 0.49 respectively. So, we only identified 67%, 62%, and 49% of all the positive, negative, and neutral sentiments correctly. It's difficult to say what has caused this bad accuracy score, and what has resulted in the positive sentiment classifications performing so much better than others. It is possible that the small dataset size of only 971 tweets, this shows one of the big limitations in the tweepy Api as you are unable to scrape historical tweets.
Finally, we ran a topic modelling algorithm on this data. The sklearn LatentDirichletAllocation model was used, this attempts to uncover topics from an assortment of documents and generates a probability distribution over words that belong to each topic. When we implemented this model, the parameters were tuned using GridSearchCV, this found that the optimal parameters were n_components = 1, and learning_decay = 0.9. When we look at the topic extracted from the text we get the following,

- egg dairy amp like gallon food buy year bread

It is hard to say what topic this list of words is supposed to be indicating, if there even is one at all. I find it hard to believe that there is only one topic contained within all these tweets as the scrape is bound to have pull in a variety of tweets, regardless of the key words. This being said, the model seemed confident in this with a relatively good perplexity score of 226.116. This may be an indication of an error in our application of the model or preprocessing, or by some miracle all the tweets do actually have the same topic.

5 Conclusion

During this project we carried out analysis on various attributes related to Irelands dairy industry, and compared them to those of the top 5 milk producers in the EU. Once we had all of the FAOSTAT and World Bank data had been aggregated into a single dataframe and cleaned appropriately, we carried out EDA on the data.

We found that the top 5 producers in the EU were Germany, France, Poland, The Netherlands, and Italy. As Ireland came in sixth on this ranking, it made sense to focus on these 6 countries. The first thing we did was to look at the relationship between herd size and volume of milked produced, upon visual inspection of the two barplots we noticed that there seemed to be a relationship between these variables. Finding the correlation coefficient confirmed this with a value of 0.98. We then saw that Ireland saw a steady increase in milk production since 2014, which we have attributed to amendments to CAP around this time that abolished caps in production. A similar increase was seen earlier in 2007 for Germany that we struggled to identify the reason for, but it could be due to increased demands or amendments to their dairy system. Over the last 29 years there has been an observed decrease in the herd sizes, however since the milk volumes have not gone down in this time we found that individual cows are individually producing more milk each year. The rate of increase was 1.47x in Ireland, 2.2x in Poland, and 1.67x on average. This indicates that there have been significant improvements in genetic breeding during this time. Ireland was found to have the highest percentage of its land used as pastureland of any country in the EU, 4 of the top 5 countries saw a sudden increase from 2000 to 2001 which is again likely due to amendments to CAP that pay producers per hectare that they use for permanent pastureland. Feeding pastureland is essential for providing high quality feed to cattle. The use of manure decreased in the top 5 countries steadily since 1991, however Ireland has seen an increase in this area likely due to the availability of the material and the amount of pastureland we have here. In other countries we suspect that the use of more modern fertilizer is increasing sue to advancements in soil testing and they can more accurately measure the amounts required.

The statistics section focused on descriptive statistics for each of the countries milk production in tone, a 5-figure descriptive statistic was used here due to their non normal distributions. These were visualized on a boxplot so that we could easier discuss them. Ireland had the lowest median, although we observed outliers in the data that can be attributed to an uncharacteristic increase in milk production since 2015, it is possible that these indicate a new era for dairy in Ireland, only time will tell. We also noticed that we are beginning to see a similar trend in the Netherlands. We calculated confidence intervals for EU population parameters based off of the data that we have for 26 of the 27 EU member states. The first was for the proportion of the population that have a pasture land coverage of greater than 15% in 2020, this value was found to lie within the range of (0.347, 0.730). We then calculated the population proportion for median head of cattle in the EU in 2020, it was found that this would lie between the 8^{th} and 17^{th} value when taken at a 95% level of confidence, then the median head of cattle in the EU in 2020 was between (226000, 565000). Finally we calculated the confidence interval for the median volume of milk produced in the EU in 2020, this was found to be between (2343252190.847, 435579357.352)litres again with a 95% level of confidence. Finally, hypothesis testing carried out between Ireland, France, and The Netherlands showed that the distributions of % pasture usage, volume produced, or head size did not compare to each other at all.

Finally machine learning was carried out on the data. Regression modelling was used to predict the Tonnes of milk produced in an EU country on any given year. The best preforming model was the KNN regressor which had an R2 value of 0.999032, and a median absolute percentage error of 2.885%. NLP was then carried out on a tweets that were related to milk/dairy prices. After pre-processing with various text analysis tools and assigning sentiment, we were only able to train a random forest regressor to classify unseen tweets with an accuracy of 61% after applying SMOTE. Unfortunately our topic modelling appears to have been unsuccessful, only assigning 1 topic to all of our tweets. It is possible that we did not have enough tweets to train the model well enough, because this was the best fitting model we found using

GridSearchCV.

It appears as though many of the improvements in dairy production are motivated by amendments to CAP or advancements in agricultural science. This makes it rather difficult to make specific recommendations that could drive innovation in this area. If Ireland were to begin importing cattle or purchasing straws from countries such as the Netherlands and breeding that stock here, we would be able to increase the volume of milk produced per Irish cow. This is likely more cost effective than the alternative of simply increasing the herd size over coming years, doing so would require an adoption of other non-pasture based dairy systems as well as a much higher expenditure which may be unsustainable.

References

bank, world (2018). URL: https://www.worldbank.org/en/about/legal/terms-of-use-for-datasets.

Bordbia (2021). URL: https://www.bordbia.ie/industry/news/press-releases/irish-food-and-drink-exports-enjoyed-a-record-year-as-value-of-sales-up-4-to-13.5bn-in-2021/.

Cardoso-Moreira, Margarida M. and Manyuan Long (June 2010). "Mutational bias shaping fly copy number variation: implications for genome evolution". In: *Trends in Genetics* 26.6, pp. 243–247. DOI: 10.1016/j.tig.2010.03.002. URL: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/mann-whitney-u-test.

Contributors, Wikipedia (Dec. 2022). *Friedman test*. URL: https://en.wikipedia.org/wiki/Friedman_test.

FAOSTAT (2023). URL: https://www.fao.org/contact-us/terms/db-terms-of-use/en/.

investopedia (2023). URL: https://www.investopedia.com/terms/r/r-squared.asp.

Korstanje, Joos (Aug. 2021). *SMOTE | Towards Data Science*. URL: https://towardsdatascience.com/smote-fdce2f605729.

Läpple, Doris, Colin A. Carter, and Cathal Buckley (Aug. 2021). "EU milk quota abolition, dairy expansion, and greenhouse gas emissions". In: *Agricultural Economics* 53.1, pp. 125–142. DOI: 10.1111/agec.12666. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/agec.12666.

Scholtens, Rikie M. et al. (July 2016). "Physiological melatonin levels in healthy older people: A systematic review". In: *Journal of Psychosomatic Research* 86, pp. 20–27. DOI: 10.1016/j.jpsychores.2016.05.005. URL: https://www.sciencedirect.com/topics/medicine-and-dentistry/kruskal-wallis-test.

Wedzerai, Matthew (Nov. 2020). *Grazing: Considering the opportunities for dairy cows - Dairy Global*. URL: https://www.dairyglobal.net/health-and-nutrition/nutrition/grazing-considering-the-opportunities-for-dairy-cows/#:~:text=Better%5C%20animal%5C%20welfare%5C%20and%5C%20health&text=Additionally%5C%2C%5C%20grazing%5C%20systems%5C%20allow%5C%20the,transition%5C%20between%5C%20standing%5C%20and%5C%20lying..

— (July 2021). *EU Grassland-based dairy: Benefits, strategies, key points - Dairy Global*. URL: https://www.dairyglobal.net/health-and-nutrition/nutrition/eu-grassland-based-dairy-benefits-strategies-key-points/.

Wikipedia$_{C}$*ontributors* (Dec. 2022). *Wilcoxon signed-rank test*. URL: https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test#:~:text=The%5C%20Wilcoxon%5C%20signed%5C%2Drank%5C%20test,%5C%2Dsample%5C%20Student%E2%80%99s%5C%20t%5C%2Dtest..

Zach (May 2021). *How to Find a Confidence Interval for a Median (Step-by-Step)*. URL: https://www.statology.org/confidence-interval-for-median/.