An Analysis of Potato Crop Yields Within the EU

Justin Flannery (sba22206)
Email: sba22206@student.cct.ie

06 January 2023

Abstract

This report examined potato crop yields within the EU. In particular it sought to develop models for predicting yields. To this end a number of different models and features were explored. The best model among those considered was a Lasso regression as measured by R^2 on out of sample predictions.

Contents

1 Introduction

This report examines potato crop yields for the 27 member states of the European Union. Crop yields are the harvested production per unit of harvested area (Eurostat 2023). Although Ireland is principally a livestock producer (ITA 2023), the country does produce some crops including potatoes. I chose potatoes as the crop has significant cultural importance in Ireland (Mac Con Iomaire and Gallagher 2008).

1.1 Project Management Framework

The project management methodology followed was Cross-Industry Standard Process for Data Mining or CRISP-DM (Chapman et al. 2000). I will provide some comments on the various steps in the process in this section. Also here is the link to the GitHub repository: link. Word Count of this report: 3295. This number is approximate (and does not include this sentence) because I had to copy and paste a *LaTeX* document into a word document to calculate it.

1. The Research Understanding phase began by reading the brief for this assignment and considering what would be required in light of the feedback from our first assignment. I then explored the Eurostat website to get a sense of the data that was available in the agricultural area which informed my research question.

2. In the data understanding phase I had to decide which variables I was going to include as potential explanatory variables. I also had to decide which subsets of the data I was going to take. During this phase I downloaded all the data I would need. For the Eurostat data I used the Eurostat (Cazzaniga 2021) library, did some basic processing and wrote out csv files for further processing.

3. Data preparation phase is outlined in detail throughout the report but particularly Section 3..

4. The Modeling Phase is outlined in Section 4 & 5.

5. The Evaluation Phase is outlined in Section 4 & 5.

6. The Deployment phase involved developing a dashboard and writing this report.

I'll include a note on the planning and delivery of the assignment. For the first couple weeks after the release of the assignment, I was in step 1 of CRISP-DM. The project began in earnest two weeks ahead of the deadline and over an intensive two weeks I stepped through the process of CRISP-DM.

1.2 Visualisations

I am mostly using the seaborn library for data visualisation. I chose this library because of its ease of use. For common plots, it provides a high level interface which can produce attractive plots with minimal code. In addition, it's integrated with matplotlib so where greater customisation is required, all the options available in matplotlib can be utilised.

Note that the colours I've used throughout the report are from the seaborn 'colorblind' palette. I chose this to make the plots more accessible to people with colour blindness, a very common condition, particularly among men (NHS 2021). However, as this palette only contains eight colours, I've used seaborn defaults where there are more than 8 colours. This is typically only where colour is just for aesthetic purposes. For example, in figure 2, colour is not necessary to distinguish the bars because they are labelled.

2 Potato Yields in the EU, some Statistics

2.1 Country Variation

Yields over time for a selection of countries are shown in figure 1. I plotted the data from 2010 since many countries have missing data before then. For comparison I chose western European countries with similar living standards to Ireland although not all such countries so as to not clutter the graph. From the graph it can be seen that there is significant variation in yields, both across countries and within countries across time. This indicated to me that predicting potato crop yields might be an interesting research question.

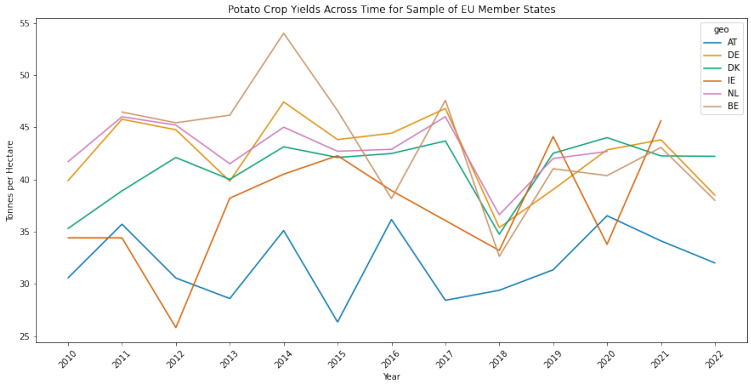

Figure 1: Potato Crop Yields within the EU.

To consider all 27 countries, I decided to aggregate the data by taking the mean across all years for each country. I sorted the data in ascending order and plotted in a bar plot. See figure 2.

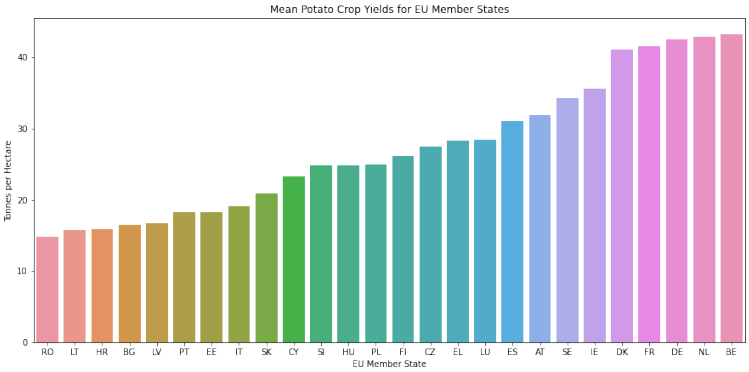

Figure 2: Mean Potato Crop Yields: 2000-2022.

Normal	T	Bootstrap	Bound
33.524913	33.389401	33.512286	Lower
37.740801	37.876314	37.708095	Upper

Table 1: Parametric & Non-Parametric Confidence Intervals

As can be seen from the plot there is considerable variation across countries. Corroborating what we saw in figure 1

2.2 Confident about Irish Yields?

To further explore Irish yields I constructed a 95% confidence interval. I did this in multiple ways, both parametric and non-parametric. I produced the first parametric approach using the scipy.stats.norm function. I did this because the shapiro wilks test for normality failed to reject the null hypothesis that the data is normally distributed. However, as the population standard deviation is unknown, I would get more robust estimates using a t-distribution so I also did this.

Finally, as can be seen in figure 3, the histogram is not clearly normal from visual inspection. While the kernel density estimate reasonably approximates the theoretical normal curve I've overlayed on the plot, it does deviate. So for completeness I also conducted the confidence interval using the bootstrap method, which unlike the above parametric methods does not depend on the data being normally distributed. I've included the confidence intervals for all 3 methods in table 1. So for example, taking the Bootstrap interval, there is 95% probability that the true population median value would be between 33.5 & 37.7.

2.3 Are the differences statistically significant?

Next I want to examine the differences between countries. As most countries had no data prior to 2010, I only considered yields from 2010 onwards. However there were still some missing values. I imputed these using the mean of each country. I used this simple approach because I'm interested in examining the mean differences between countries, as such, I don't have to worry about preserving the temporal structure of the data.

2.4 Parametric Test

As I'm comparing many countries, I will consider an ANOVA hypothesis test. The assumptions for ANOVA are that the data is independent and normally distributed with constant variance. To test for independence I used the pearson correlation test in the scipy.stats library using a p-value of 0.05, a standard value to use. Of the 26 EU countries in the dataset, Ireland had a significant relationship only with Spain. As there were 25 relationships to consider and we were using a p-value of 1 in 20, it's possible that the significant result with Spain is due to chance. Therefore, I will consider the data to be independent.

Next I tested for normality with the Shapiro-wilk test. Again I used a p-value of 0.05. Ireland's data was normally distributed, as was most other countries data. However, for four countries the null hypothesis of normality was rejected.

Finally I used the levene test to test the null hypothesis that all input samples are from populations with equal variances. I again used a p-value of 0.05. The results were very mixed, of the 25 pairs between Ireland and another member state, 9 of them had the null hypothesis rejected, ie we can only consider 16 of them to have equal variance.

We've seen that some countries data is likely not normal and that the samples are from populations with unequal variances in many cases. I'll proceed with the ANOVA for completeness but then for robustness I'll conduct a non-parametric test.

I conducted the ANOVA using the f_oneway test from the scipy library. The null hypothesis is that the groups have the same population mean. The p – value was vanishingly small ($1.385700249808896e-123$) and so we can rule out the results being a matter of chance. There are statically significant differences between the mean value of potato yields across EU countries.

2.5 Non-Parametric

As the assumptions for ANOVA were not met, I'll use the Kruskal-Wallis test. This test does not require the data to be normally distributed (Lantz 2013). While it does require the variances of the groups to be similar, it's not as important an assumption as in the case of the one-way ANOVA. Kruskal-Wallis also requires the data to be independent which I have confirmed above and it requires the data to be on an ordinal scale. As my data are floats, they can of course be ordered and are therefore on an ordinal scale. The null hypothesis,

that the population medians of all the groups are equal, is rejected with a p-value of (1.5463164959760256e-48).

I've shown that the potato yields for EU countries have statistically significant differences between their measures of central tendency. This report will seek to develop machine learning models to explain and predict yields.

2.6 Challenges

There were a number of challenges I faced conducting the statistical analysis of my dataset. For example, a number of countries were missing data and I had to consider how best to handle this as discussed.

Another challenge was testing the assumptions of the models I was using. For example, in testing independence between countries, for all pairs of EU member states that involved Ireland, only 1 pair was found to likely have a dependent relationship. Do I conclude that the data are not independent? In the end I concluded the data were independent for the reasons I outline above.

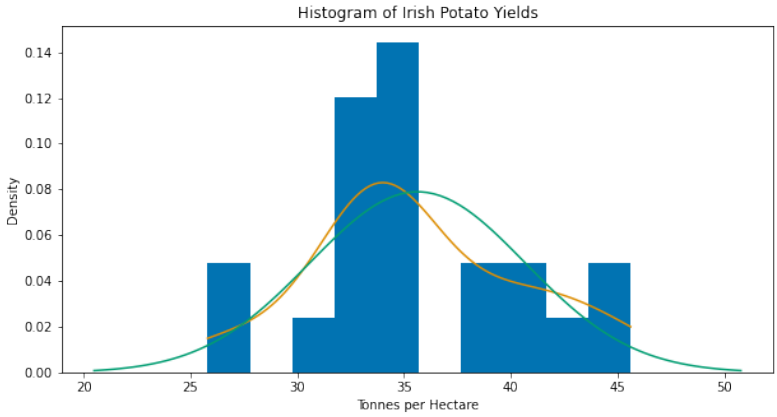

Figure 3: Histogram of Irish Potato Crop Yields: 2000-2022.

3 Exploring the Data

3.1 What are our features?

This report attempts to predict potato crop yields for the member states of the EU. To this end I have gathered data on a number of different indicators which may be useful features in machine learning models.

For the variables sourced from Eurostat there was a considerable amount of overlap in the processing of the data. My first step was to apply a function I created to rename some columns, filter some rows and drop some columns. For

example, all the datasets had country information stored in a column with the name ‘geo
TIME_PERIOD’ so I would rename this to geo seen as it contained no year information. I would also filter the data so the rows only related to the 27 member states of the EU. Finally I would drop some superfluous columns.

The data would also typically have the various years as columns so I would convert the data to a long format where the various years would be in a column and the associated values would be in another column.

3.2 Crop Rotation - An Example of EDA

The first potential explanatory variable I looked at was crop rotation for arable land where farm structure was FT16_SO, ie general field cropping, which would include potatoes. The data is broken down by total land, land where 0 percent of it is part of crop rotation, 1-24 percent of it is involved in crop rotation etc. The values of the data are hectares of land so larger countries would naturally have larger values. This would be of limited use in predicting crop yields which are defined on a per area of land basis. So I decided to engineer the data so that it showed the share of a countries total land which was in each of the aforementioned categories. As can be seen in figure 4, most countries have most of their land (general field cropping) in the category PC_GE75.

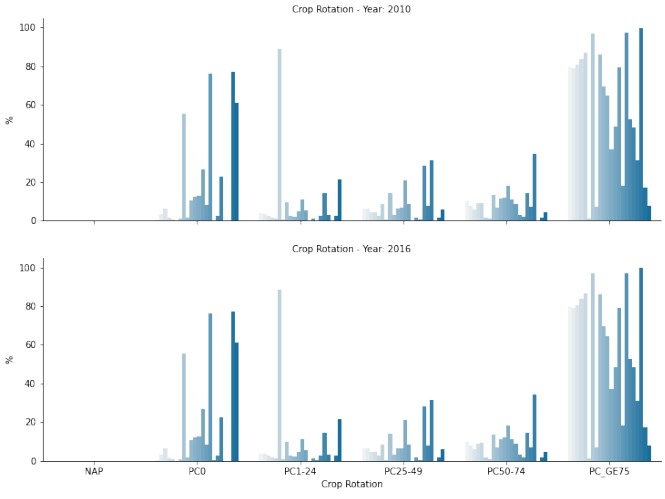

Figure 4: Crop Rotation.

This means that over 75% of the land in this category is part of crop rotation. As there was only data for 2010 and 2016, I decided to convert this numerical data to categorical data. Otherwise I would be losing most of the yields data observations during a merge. I did this by defining a categorical_converter_function

which assigned a value from 'very low' to 'very high'. There were missing values for the category 'NAP', however these are actually 0 values. I know this because the sum of the sub-categories match the total figure if the NAP values are 0. Finally, as can be seen from figure 4, the 2013 and 2016 information contains almost indistinguishable information. As a result, I simply decided to take the more recent 2016 values when producing the categories. The final step is to pivot the data so the categories are their own column which will be necessary later for one hot encoding.

Most of the above steps were repeated for the other variables. Rather than having considerable repetition, I will move the discussion around exploring the rest of my explanatory variables to the appendix, see appendix A. For one variable I made use of the sklearn (Pedregosa et al. 2011) MinMaxScaler function to scale the data to values between 0 and 100 so I could use my categorical_converter_function.

3.3 Exploration of Final Data Set

Once I had all my data clean, the next step was to merge all the datasets together. My merged dataframe had some missing values because the weather and gross nutrient balance data didn't have values for 2022 while other NAs were introduced because not all countries with data on yields had data for each of the explanatory variables. I've used forward fill for the gross nutrient balance data for reasons discussed in appendix A. I've also used forward fill for the temperature data because it has trended upwards over time so a measure of central tendency, for example, would change the structure of the time series. I've also used forward fill for precipitation seen as the values don't exhibit much year to year fluctuation. Greece had missing weather data. I decided to assign the missing Greek data the mean of the Italian data. I did this because Italy has a similar Mediterranean climate as Greece and the two countries are geographically close.

For countries with missing crop rotation data I've assigned a value of 'Other'. I did this because I don't know that missing data here isn't a feature of the countries crop sector that I should be trying to capture.

Germany had no data on legal form. Unlike crop rotation, I doubt this indicates anything about the German crop sector. As a result, I simply assigned the modal value, ie, the legal form which is most prevalent among the other countries.

The final step in getting my dataset ready was to do a one hot encoding on my categorical variables because the sklearn library expects categorical variables to be binary for modelling. I did this using pandas get_dummies function.

In figure 21 I consider a correlation heatmap for my numerical variables. This tells us that precipitation and the various measure of soil nutrients are positively correlated with yields. Although correlations are quite weak. Temperature is slightly negatively correlated although it's essentially no correlation at all. I could possibly drop this variable but as I'll be formally using feature engineering methods later, I'll leave it in.

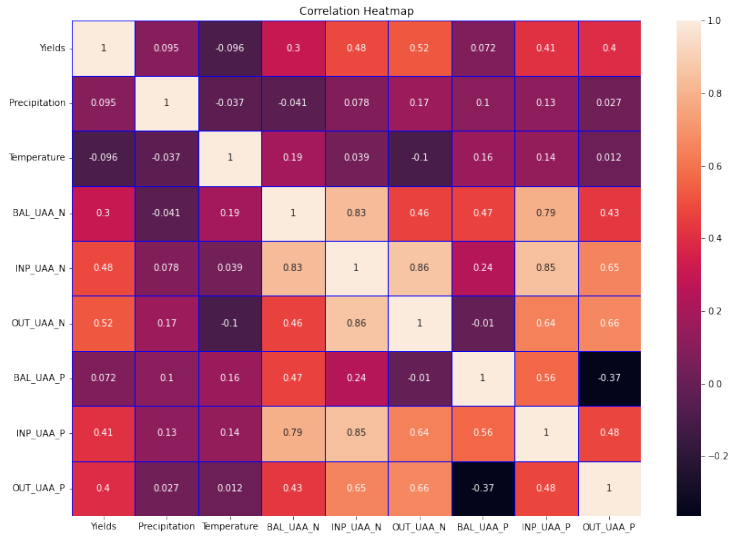

Figure 5: Correlation Heatmap.

I also plotted box plots to show the distribution of my data. As can be seen from figure 21 in appendix A, most variables have a number of outliers. However, as this is official data which has gone through checks not only by Eurostat but also by member states national statistical institutes, I've concluded that the outliers are likely actual observations. While leaving them in will likely effect the performance of machine learning models (Soehono and Pratomo 2017), throwing them out would give a misleading picture of the predictability of yields.

The final step before running machine learning models was to use the StandardScaler function from the sklearn library on my numerical variables. I did this because some of the values for my data are on different scales which can impact on the performance of models, particularly for the dimensionality reduction and clustering (James et al. 2013).

Note that I've not taken account of the CAP. This is for two reasons. Firstly, CAP is largely concerned with income supports for farmers and it's dubious whether it would have any impact on crop yields. Secondly, I'm only considering EU countries.

3.4 Challenges

There were a number of challenges in acquiring the raw data. Most of my data was from Eurostat and found that the metadata was not very accessible. Weather data for many countries is not widely available and was difficult to find.

Accessing twitter data is also not trivial. Regarding licenses and permissions, all the data I used is free for non-commercial use. See appendix B for more details on this and for more discussion around challenges.

3.5 Dashboard

In the dashboard I chose the appropriate visual encoding for the different data types. For example, I used a scatter plot for numerical data and bar plots for categorical data. I've also labelled the axes accurately and clearly while also providing informative titles. Colour has been used to distinguish data but it's not the only means of encoding data. For example, the height of bars is being used to convey the count of categorical information. The app also utilises interactivity effectively, allowing the user to choose different variables, graph types, clusters in the Kmeans algorithm and the alpha slope parameter in the ridge regression. As I've demonstrated, the choices made during the development of the app are in line with Tufts principles, resulting in an effective and informative visualisation tool.

The dashboard can be viewed in the Jupyter notebook 'EDA + Dashboard.ipynb'.

4 Predicting Potato Crop Yields

4.1 Supervised

As part of the modelling and evaluation phase of the CRISP-DM framework, I began by using the train_test_split function from sklearn to produce test and training sets. I then trained a number of regression models as I'm trying to predict a continuous variable. The first model I ran was linear regression. This gives a baseline for more advanced methods. See figure 6 for a graph of actual vs predicted and table 2 for the results.

Figure 6: Actual vs Observed Predictions.

Model	Group	Train R^2	Test R^2	Train RMSE	Test RMSE
Lasso	Lag	0.73	0.8	0.52	0.46
LinearRegression	Feature Engineering	0.71	0.8	0.53	0.47
LinearRegression	Lag	0.73	0.79	0.52	0.48
RandomForestRegressor	Lag	0.94	0.79	0.25	0.48
Ridge	Feature Engineering	0.73	0.79	0.52	0.47
Ridge	Lag	0.71	0.78	0.54	0.48
RandomForestRegressor	Feature Engineering	0.85	0.75	0.38	0.51
Ridge	Baseline	0.75	0.7	0.5	0.55
Lasso	Baseline	0.75	0.7	0.5	0.54
Lasso	Feature Engineering	0.58	0.7	0.65	0.57
LinearRegression	Baseline	0.75	0.69	0.5	0.55
DecisionTreeRegressor	Feature Engineering	0.73	0.67	0.52	0.59
RandomForestRegressor	Baseline	0.94	0.66	0.25	0.58
DecisionTreeRegressor	Lag	0.6	0.59	0.63	0.66
DecisionTreeRegressor	Baseline	0.75	0.53	0.5	0.68

Table 2: Regression Results

Next I created a function for doing grid search cross validation for a given model while also printing summary information on the model and plotting a graph as in figure 6. I called this function with Ridge, Lasso, Decision Tree and Random Forest models. This allowed for hyperparameter tuning of relevant parameters for the various models in order to improve performance. See results in table 2. The cross-validation gives us additional confidence in the results because the models were trained and tested on multiple train-test sets and averaged. In my case I used 5 fold cross validation.

To try and improve model performance I added a 1 year lag of Yields. Given that this is annual, ie not seasonal data, a 1 period lag is a natural choice. I ran all the above models again with a lag. In general this improved performance as measured by R^2 on the test set.

Next I used feature selection with the RFECV function from the sklearn library. I again ran all models listed above. The results of this analysis can be seen in table 2. The best model was a Lasso regression which included a lag.

4.2 Unsupervised

I also used unsupervised learning as I wanted to visualise my dataset and see how similar or different countries are. To achieve this I performed dimensionality reduction on my explanatory variables using the sklearn PCA function with the number of components set to 1. I then performed Kmeans clustering using the sklearn library. Through experimentation, I believe 3 is the appropriate number of clusters from visual inspection. See figure 7. I've marked the Irish data on the graph and you can see that Ireland straddles the first two clusters.

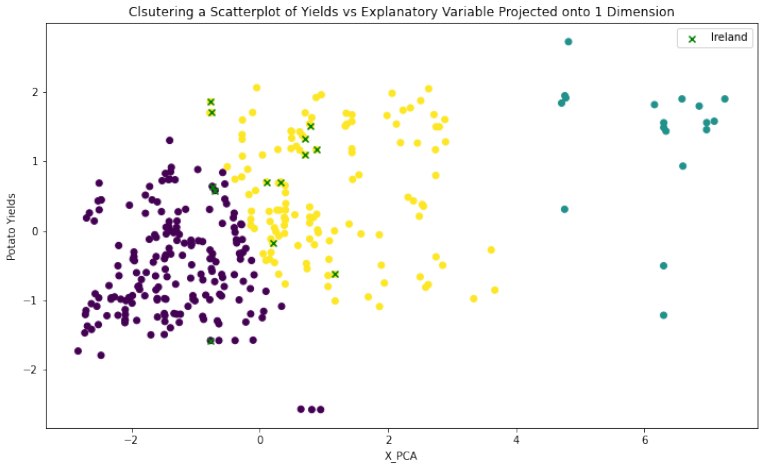

Figure 7: Clustering of Yields and dimension reduced explanatory variables.

5 Sentiment Analysis

I also carried out a sentiment analysis on twitter data. I queried the Twitter API for tweets that referenced 'potatoes'. After a considerable amount of processing, see appendix A for details, the sentiment was produced using the TextBlob library. Overall the sentiment around potatoes from consumers was quite neutral. The lesson for potato crop producers is not to sell the farm just yet.

I also carried out classification modelling with my sentiment data. I began by classifying sentiment as either being positive or negative. I then produced train and test splits. As I had only binary outcomes, I used logistic regression. See confusion matrix in figure 8 for results. As can be seen from the confusion matrix, the accuracy of the model was $\frac{15}{20} = 0.75$ or 75%

6 Conclusion

We've seen in this report that there are statistically significant differences between potato yields across EU countries. A dataset with a large number of features was developed to train machine learning models to predict yields. The best model based on R^2 in the test set was a Lasso Regression. The report also utilised the unsupervised learning methods of dimensionality reduction and Kmeans clustering. A sentiment analysis was also performed on twitter data which mentioned 'potatoes'. The findings were that consumer sentiment on 'potatoes' was quite neutral. These findings suggest that potato crop producers in the EU should not be overly concerned about the market.

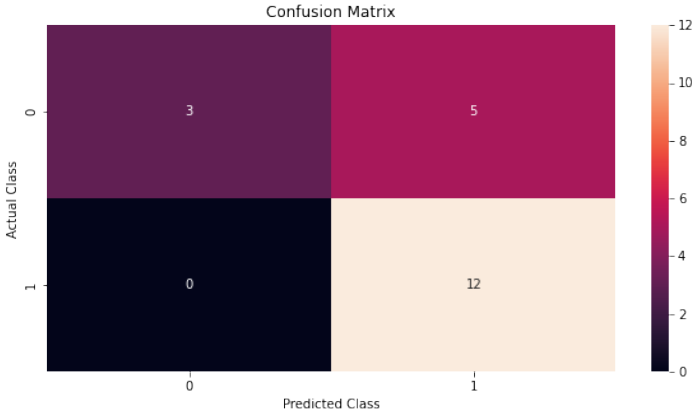

Figure 8: Confusion Matrix.

References

Chapman, Pete et al. (Aug. 2000). *CRISP-DM 1.0 Step-by-step data mining guide*. Tech. rep. The CRISP-DM consortium. URL: https://maestria-datamining-2010.googlecode.com/svn-history/r282/trunk/dmct-teorica/tp1/CRISPWP-0800.pdf.

Mac Con Iomaire, M. and P. Gallagher (2008). "The History of the Potato in Irish Cuisine and Culture". In: *Vegetables: Proceedings of the Oxford Symposium on Food and Cookery 2008*. Ed. by S. Friedland. Devon: Prospect Books.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

James, G. et al. (2013). *An Introduction to Statistical Learning (with Applications in R)*. Springer.

Lantz, Björn (2013). "The impact of sample non-normality on ANOVA and alternative methods". In: *British Journal of Mathematical and Statistical Psychology* 66.2, pp. 224–244. DOI: https://doi.org/10.1111/j.2044-8317.2012.02047.x. eprint: https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8317.2012.02047.x. URL: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8317.2012.02047.x.

Soehono, L.A. and Devanto Pratomo (July 2017). "Does outlier need to be removed from regression analysis? Case study in economics research". In: *Journal of Applied Economic Sciences* 12, pp. 1141–1147.

Bank, World (2021). *Climate Knowledge Portal*. URL: https://climateknowledgeportal.worldbank.org/.

Cazzaniga, Noemi (2021). *Eurostat Python Library*. URL: https://pypi.org/project/eurostat/.

NHS (2021). *Colour vision deficiency*. URL: https://www.nhs.uk/conditions/colour-vision-deficiency/.

Eurostat (2023). *eurostat*. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Crop_yields#:~:text=Crop%5C%20yields%5C%20mean%5C%20harvested%5C%20production,harvested%5C%20area%5C%20for%5C%20crop%5C%20products..

ITA (2023). *International Trade Administration*. URL: https://www.trade.gov/country-commercial-guides/ireland-agricultural-sector#:~:text=Agricultural%5C%20production%5C%20is%5C%20a%5C%20key,and%5C%20dairy%5C%20products%5C%20are%5C%20exported..

A Appendix

Another variable explored was the gross nutrient balance of soil. There are many indicators here which go into detail on what has been added or removed from the soil. For my purposes, I believe the aggregate is what's of interest. So I include the aggregate input (INP_UAA), output (OUT_UAA) and balance (BAL_UAA) in my analysis. I examined the null values across available years, 1985-2019, and discovered that there were many missing values in the 80's and 90's and again from 2015. The missing values from the earlier period were of no concern because my yield data does not got back that far. As can be seen from figure 9, the data doesn't exhibit large year to year fluctuations, although there do appear to be level changes for Ireland and Belgium in more recent years.

Figure 9: Gross Nutrient Balance

Note I chose a small sample of countries for the graph because it would be too cluttered to include all 27 member states. I chose the countries I did because they're similar to Ireland in that they are small, western European countries. Given the lack of volatility, I filled missing values using simple forward fill. To get a sense of the data for all countries, I aggregated the data by calculating mean values across years for each country. I also broke it down by nutrient where N is nitrogen and P is phosphorous. As can be seen from figure 10, Austria appears to be a significant outlier for nitrogen.

Next I considered data on irrigation. See figure 11, note that UAAIB and UAAIT are acronyms for 'Irrigable utilised agricultural area' and 'Irrigated utilised agricultural area' respectively. Note as I understand it the terms can

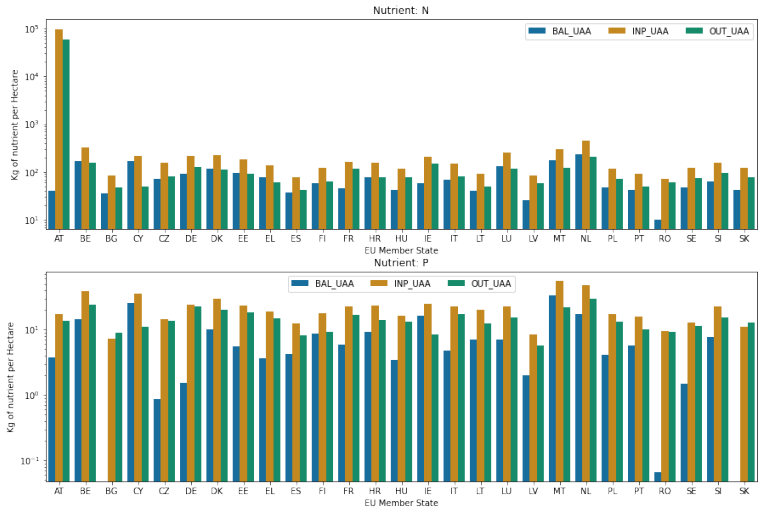

Figure 10: Gross Nutrient Balance

be considered potential and actual irrigation respectively.

As with crop rotation we just have values for two years and the data are similar in each year. Processing was almost identical to crop rotation except this time I averaged the years rather than choosing the most recent. I did this because although the data was very similar in each year, I could at least through visual inspection identify differences which is more than can be said for the crop rotation data. These differences can be seen clearer when the data are seen as shares of the total rather than as absolute values, see figure 12.

Next I considered data on the legal form of farms across countries. The processing was almost identical to crop rotation so I won't go into details. See figure 13.

Next I considered pesticides but as so few countries had data I decided not to include it as potential explanatory variable.

Next I considered soil erosion. There was no missing values. There is only data available for 3 years, see figure 14. Note that I chose a font of 8.5 so that the x-axis labels wouldn't overlap yet still be legible. As with irrigation, I averaged the data and converted it to a categorical variable. However, while with the other data I could readily convert to shares, this data I could not. To apply my catgories_converter function I need values between 0 and 100. To achieve this I used the MinMaxScaler from the sklearn library.

Next I considered tenure status. Processing was very similar to crop rotation except I blended the data across the years so as to minimise missing values. I felt this was appropriate because tenure status is not a variable which would

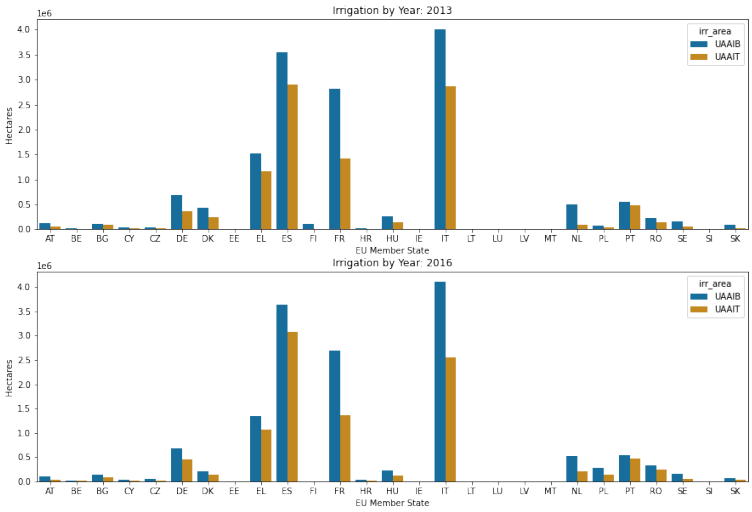

Figure 11: Irrigation

have much year to year variation. See figure 15 where a countries total hectares in general field crop farms are broken down by country and tenure status and represented as shares of the total.

Finally, as it relates to the Eurostat explanatory variables, I considered training levels of farmers. Processing is very similar to crop rotation.

Next I turned to the weather data. First up is precipitation. The only unique step here is that the country labels are the full country names so to make it consistent with the Eurostat data (where countries are labelled with a 2 digit code), I had to create a dictionary which defined the relevant mapping and made the changes with the pandas rename method. Mean precipitation levels for each country can be seen in figure 16. Not surprisingly, Ireland is almost the wettest countries in the EU, second only to Slovenia.

Finally, as it relates to my explanatory variables, I considered temperature. Figure 17 shows the distribution of temperatures across all countries. The data is slightly skewed to the right. To make this clearer I mark the mean and median values on the graph. The mean being above the median confirms the skewness. Figure 18 plots the median, on account of the skewness identified in the distribution, Ireland is in the middle which is what you might expect for our temperate climate.

Finally I considered potato yields which is the variable I'm trying to model. The previously discussed analysis highlighted the statistically significant variation across countries. Figure 19 displays a histogram with a density curve. Note the slight skew in the data. Figure 20 shows a choropleth graph for median crop

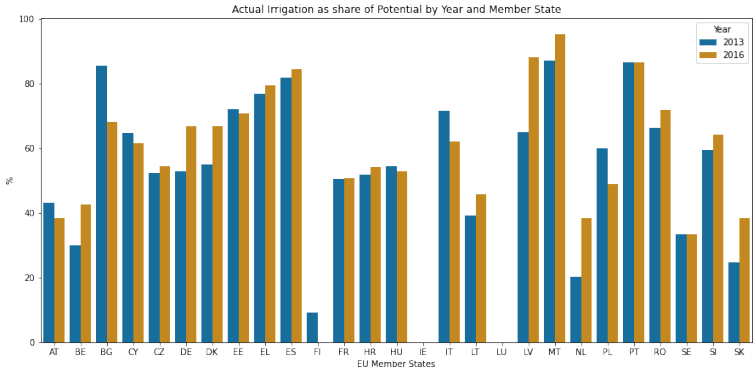

Figure 12: Irrigation

yields across time for each member state of the EU. I chose the median because the distribution had a slight skew.

Seperately, I conducted a number of steps to clean my Twitter data ahead of sentiment analysis and classification. Steps included the following:

- Make words lower case.
- Remove Twitter handles, punctuation and commonly occuring words.
- I examined frequently occuring words and from this analysis removed 'rt' seen as it's not relevant for determining sentiment
- By examining infrequent words I was able to remove some irrelevant 'words'
- Finally, spelling mistakes were removed with the TextBlob library and the words were lematized

These steps were all taken to reduce the noise in the data and increase the signal for the sentiment analysis.

For the classification of the sentiment, I defined a lambda function to convert the sentiment, which ranged from -1 to 1 into a binary categorical variable which could take the value 'Positive' or 'Negative'. The sklearn CountVectorizer function was used to transform the text strings into numerical features.

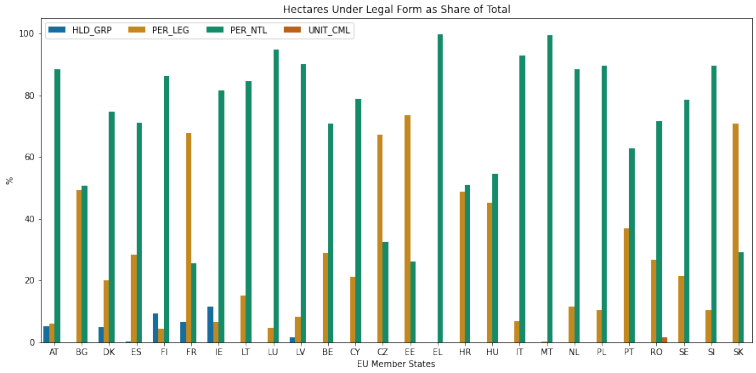

Figure 13: Legal Form of Farms

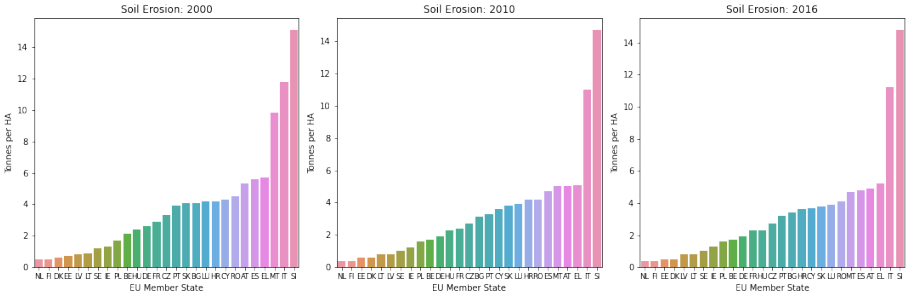

Figure 14: Soil Erosion

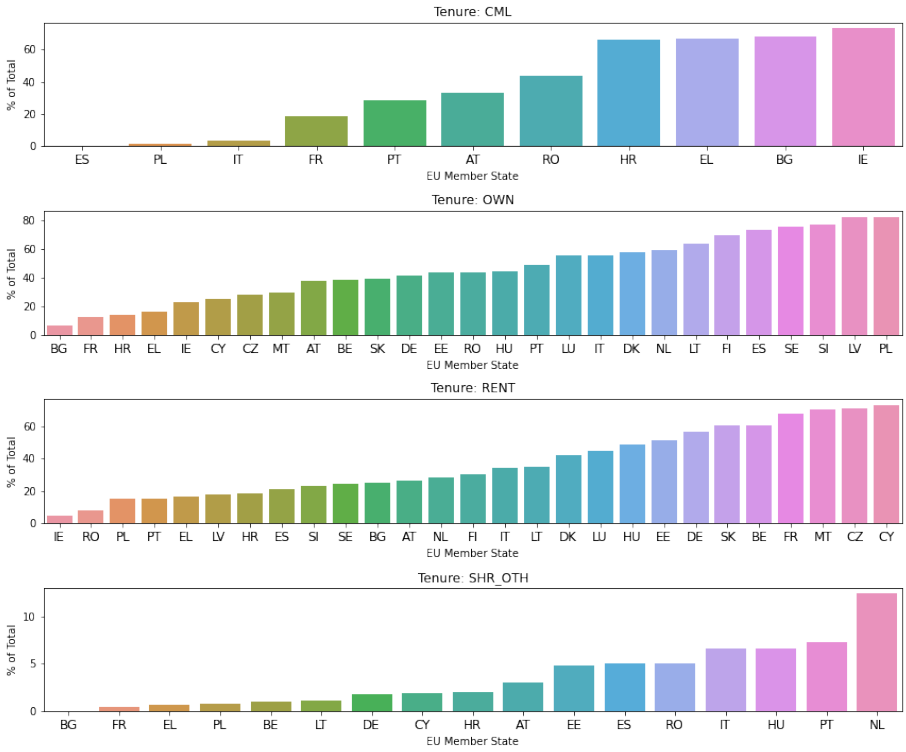

Figure 15: Tenure Status.

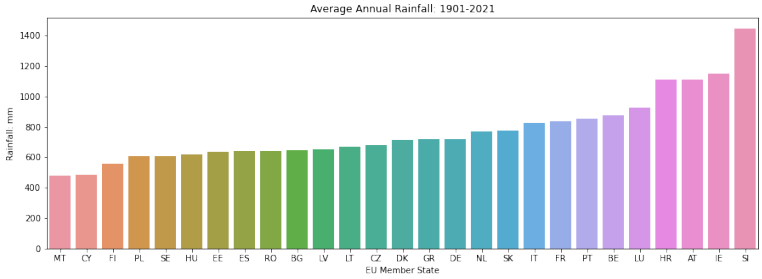

Figure 16: Mean Precipitation

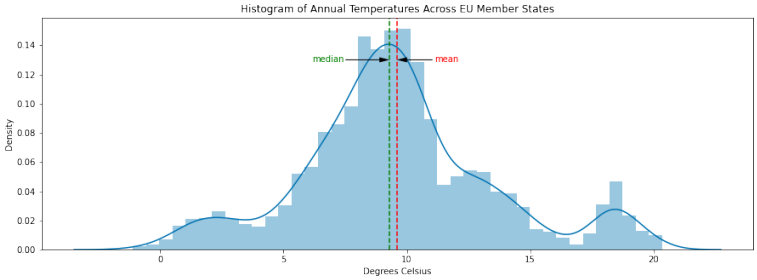

Figure 17: Temperature Distribution

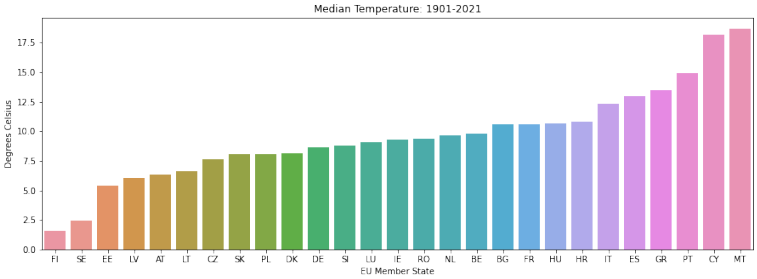

Figure 18: Temperature Mean by Country

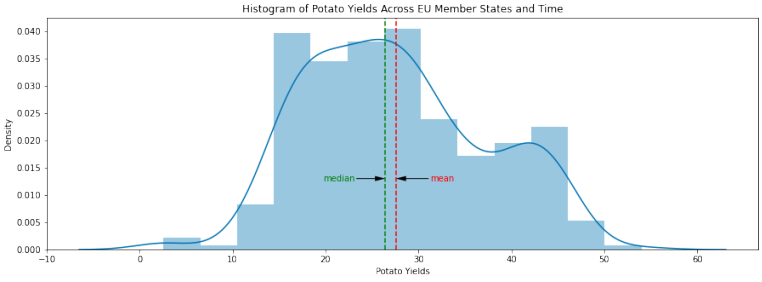

Figure 19: Yields Distribution

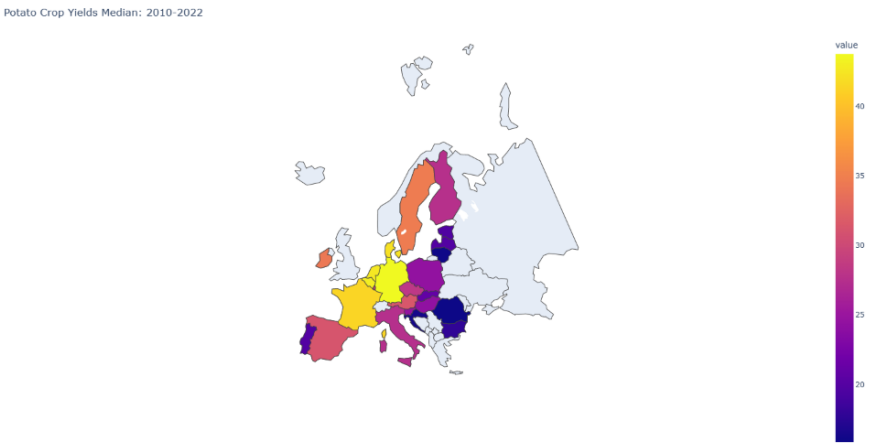

Figure 20: Median Temperature

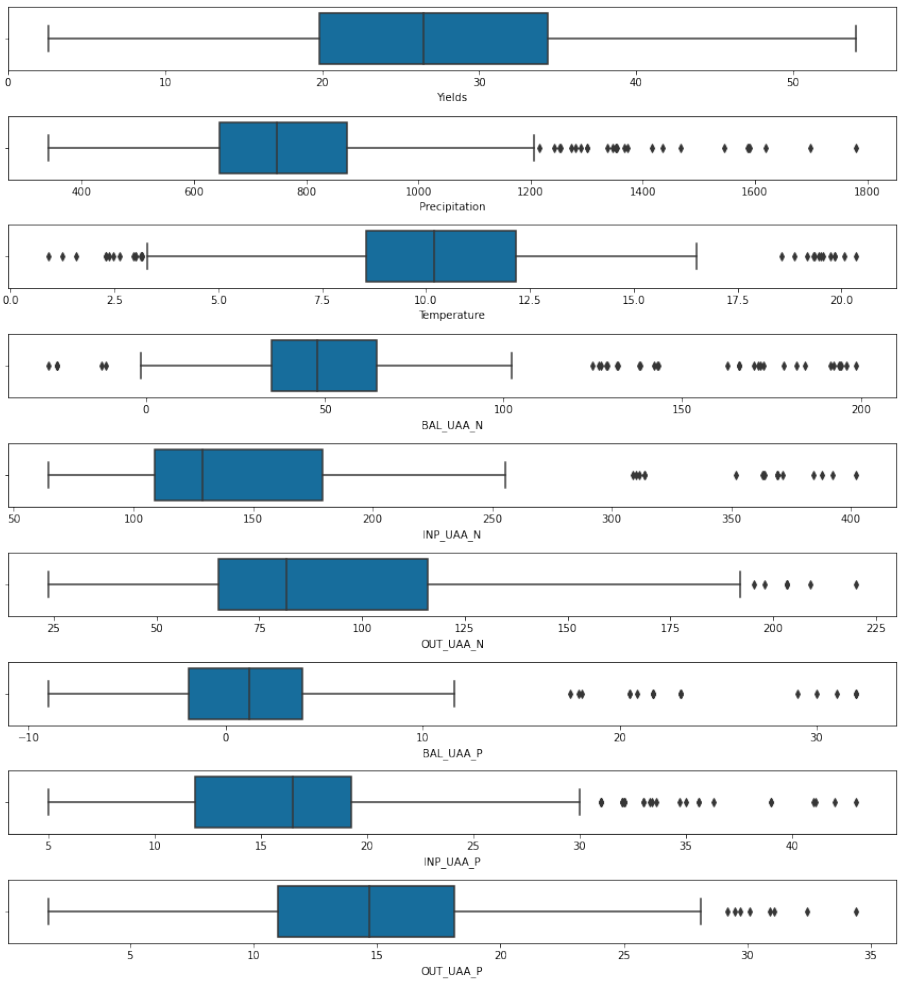

Figure 21: Boxplots.

B Appendix

There were a number of challenges in acquiring the raw data. Most of my data was from Eurostat and I found that the metadata was not very accessible.

As I was trying to predict agricultural yields, I reasoned that climate might be important. I was surprised at how difficult it was to access weather data for EU countries. I eventually found data on the World Banks Climate Change Knowledge Portal (Bank 2021). However it was not accessible seamlessly. Registration was required. I also had to manually download separate csv files for both temperature and precipitation for each of the 27 member states of the EU.

Finally, I had to access data from Twitter for sentiment analysis. This was challenging as there are a number of steps in accessing the data. I had to join Twitter and then register for a developer account. I had to learn about good programming standards around not keeping private keys etc in notebooks. I achieved this through a .env file. Ingesting data from the twitter API using the requests library was not trivial. The final challenge was taking the nested json data structure returned by the requests.get method and converting it to a simple dataframe.

Regarding licensing/permissions, the Eurostat data is freely available for both commercial and non-commercial purposes. The only conditions on its use is that the source is indicated and if the data is modified, this must be made clear to the end user. Regarding the World Bank data, this is again freely available for download, however unlike with Eurostat, the data is not intended for commercial purposes. Permissions around Twitter are more complicated. In signing up for a developer account, you are required to sign a lengthy 'Developer Agreement and Policy' which outlines the rules and guidelines for using the Twitter API. In general, you are required to respect the privacy and rights of Twitter users, and you are not allowed to use the data for any illegal or malicious purposes.