# Group ID - MSc in Data Analytics

Author: Sinéad Duffy

e-mail: [sba22229@student.cct.ie](mailto:sba22229@student.cct.ie)

Student ID: sba22229

* You have been tasked with analysing Ireland's Agricultural data and comparing the Irish Agri sector with other countries worldwide.
* This analysis should also include forecasting, sentiment analysis and evidence based recommendations for the sector as well as a complete rationale of the entire process used to discover your findings.
* Your Research could include export, import, trade imbalance, arable production, animal stock, medicinal input, organic, gm products etc. (or any other relevant topic EXCEPT Climate change) with Ireland as your base line.

## Abstract

*Short abstract of the report on organic farming in Ireland*

## Word Counts

Full Document

Report only

# Introduction

This report analyses Ireland’s Organic Sector and compares it against other countries in the European Union.

## Methodology -CRISP DM Project Management Plan

The author utilised the cross-industry standard process for data mining or CRISP DM methodology tool in order to structure the analysis. CRISP DM is a project management tool that allows the developer to work in a circular pattern, revisiting steps as knowledge of the business and the understanding of the data evolve (Wijaya. 2021, Holt. 2022).

Diagram

Description automatically generated

Figure - Wijaya 2021

Appendix 1 includes the authors plan to incorporate the CRISP-DM methodology to the deliverable sections of this report.

Analysis of the dataset will be carried out in three phases namely:

* Exploratory Data Analysis,
* Statistical Analysis,
* Machine Learning

Each phase is described in more detail throughout the report below keeping the research questions in mind.

## Business Understanding

The Central Statistics Office (CSO) estimates that the Agricultural sector in Ireland was estimated to be valued at €8.7 billion euros (CSO, 2022). Of that, Bord Bia as reported by Teagasc, that the organic retail food market was €260 million euros (Teagasc, 2022). Bord Bia’s 2020 report into Attitudes towards Organic Food outlines that consumers are purchasing more organic food since the onset of Covid in 2020. Locally produced food is more popular by a 2 to 3 ratios (Reaper, 2020).

This report will analyse the production of organic food in Ireland and compare against other countries within the European Union.

### Research Question

As stated above, this report will explore the production of organic food in Ireland compared with fellow member countries is Europe.

### Project Goal

This project aims to analysis the production of organic animal products within Ireland, as compared with other countries in the European Community.

## Data Understanding

### Data Collection

The data for this assignment was selected from Eurostat.eu. Eurostat describes themselves as being ‘The home of high-quality statistics and data on Europe’ and is a key source for data and datasets (Eurostat.eu, 2022). The data was sourced from the agricultural section of the datasets, specifically the database section (Eurostat.eu, 2022). The accompanying metadata file from Eurostat (see appendix 2), outlines that the data for the organic farming section from 2012 is broken into seven distinct sections – see page 2. The data is collected by questionnaire which was revised in 2012. Of the seven listed, the dataset relating to organic products of animal origin – dataset org\_aprod is selected.

This dataset was chosen by the author as it had relatively clean data and should allow for a comprehensive evaluation to be carried out on it including machine learning and statistical analysis (Anjai, 2021).

### Data Description

A full description of the data used in this report can be found in the data dictionary appearing in appendix 3.

### Exploring the Data

*Chapter 3 of Predictive Analytics for Dummies – pg. 51 to*

Shape of the Dataframe

A key step before beginning analysis is to explore the dataset. In this instance, commends such as .head() and .shape() are used to see the structure of the dataframe.

Contents of the Dataframe

The contents of the dataframe is reviewed using .isnull() to identify any empty rows, and .duplicated() to locate duplicate rows

The unique values in each column is identified using the .nuique() function. The .groupby() function helps to show the number of groups withing the agriprod and geo columns.

### Data Quality

*Chapter 7 of Python for Data Analysis, pg. 195 to 223*

Based on the initial analysis completed in the previous section it’s clear that some columns need action to allow the data to be formatted in such as way to make it useable for analysis. The analysis and outcome of each column is listed in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name** | **Analysis of Column** | **Action needed** | **Justification** |
| DATAFLOW | Source of dataflow – same value for all entries | Remove column | Data not needed to complete analysis |
| LAST UPDATE | Date and time of last update to dataframe – same value for all entries | Remove column | Data not needed to complete analysis |
| freq | Frequency of data input - same value for all entries | Remove column | Data not needed to complete analysis |
| agriprod | 23 unique values – listed as alpha numeric values | Analysis of categorical values   1. Update value names 2. Categorical variable analysis | Use one-hot, melt or another form of categorical variable analysis |
| unit | 2 unique values displaying the method of weight |  | Use one-hot, melt or another form of categorical variable analysis to allow for data analysis |
| geo | 33 unique values outlining geographical area reporting the data |  |  |
| TIME\_PERIOD |  |  |  |
| OBS\_VALUE |  |  |  |
| OBS\_FLAG |  |  |  |

Table - Analysis of actions needed post initial review

## Data Preparation

### Data Selection

### Data Cleaning

### Feature Engineering

### Data Integration

### Data Formatting

## Modelling

### Model Selection

### Test Design

### Model Development

### Model Assessment

## Evaluation

### Evaluation Results

### Review Process

### Determine next steps

# Data Preparation and Visualisation

## Selecting and loading the data frames

## Begin Exploratory Data Analysis in

## Reviewing the variables

## Exploring the Variables using Graphs

# Statistics

## Descriptive Statistics

## Poisson Distribution

## Binomial Distribution

## Normal Distribution

1. Machine Learning

## Overview

## Algorithm 1 - Logistic Regression

## Algorithm 2 - Naive Bayes (Gaussian)

## Algorithm 3 - Decision Trees

# Results and Discussion

## Exploratory Data Analysis

## Statistics

## Machine Learning

# Conclusions

# Reference

# Appendix

## Appendix 1

Detailed steps undertaken as part of the CRISP-DM are outlined in the diagram below. Also included are intersection points with repot deliverables.

Diagram, timeline

Description automatically generated

## Appendix 2