**MSC\_DA\_CA2**

**Ireland Milk Production Comparison with Selected EU Countries**

**Mahender Kumar**

**SBA22212**

Table of Contents

[**1. Report Description** 3](#_Toc123144398)

[**2. Abstract** 3](#_Toc123144399)

[**3. Introduction** 3](#_Toc123144400)

[**3. Exploratory data analysis (EDA)** 3](#_Toc123144401)

[**4. Conclusion** 5](#_Toc123144402)

[**4. References** 5](#_Toc123144403)

[**5. Appendix** 6](#_Toc123144404)

[5.1 Gantt Chart 6](#_Toc123144405)

Table of Figures

[Figure 3.1 Display first five rows of the livestock products data frame 4](#_Toc123202914)

[Figure 3.2 Display first five rows of the world population data frame 4](#_Toc123202915)

[Figure 3.3 Common methods to get the insight of data 4](#_Toc123202916)

[Figure 3.4 Merging two datasets 5](#_Toc123202917)

# **1. Report Description**

word count and

other stuff

why use colour on the graph

**Project outcome**Create a datasheet of dairy product over time (1960 to 2020) for Ireland and other EU nations. Data frame will be created from different dairy products, types of animals where dairy products are collected and total population of the Countries.

**Data Description**This data, which is in a csv file format, consists of Year, Ireland and other European countries. These columns contain information about the columns mentioned from 1960 to 2020.

**How Was This Data Collected?**Crop and Food data were obtained from the Food and Agricultural Organisation of the United Nations. FAO collects the data from all the countries in the world. FAO is an international organisation and it is a part of the United Nations. I am using raw data to form the base of my data frames. FAO is a website that provides reliable data in various fields.

For more detailed information, please click the link below. Link:

<https://www.fao.org/faostat/en/#data/QCL>

Total yearly population data are collected from EUROSTAT which is freely available and there is no licensing requirement to access these data. This population is then compared to the other 27 countries in the EU. Four countries are identified as having a similar (+/- 1m people) population in 2020. The population data for these countries is downloaded. They can only be downloaded separately.

For more detailed information, please click the link below. Link:

https://data.worldbank.org/indicator/SP.POP.TOTL?end=2020&start=1960

**Is There a License for the Data Used?**The FAO, the website from which the data was obtained, has a Creative Commons Attribution 4.0 International license. This license type makes shared data publicly available according to open data standards and license datasets. Data are freely available and no licence is required to access or download the data.

# **2. Abstract**

# **3. Introduction**

Add introduction here

Why choose this data

Data explanation

# **3. Exploratory data analysis (EDA)**

EDA is first step in the data analysis process, it is an approach to analysing data sets to summarise their main characteristics. It allows to check and analyse the data before we make any conclusion or assumption. EDA also ensures that the results produced are valid and applicable to the business needs (Wes McKinney 2017).

Following the benefits, we begin using EDA on the data sets.

* It helps to understand the dataset variable and the relationship among the data
* It provides better knowledge of the data set
* Helps to identify if there are any errors in our dataset. I.e., Duplicates, missing data
* Helps detect outliers or anomalies

In my project, data are collected from 3 datasheets. These datasheets consist milk production data from all around the world. Data are gathered over a period of 61 years (1961-2021).

The following are 3 types of datasets.

1. Livestock products
2. World population
3. World population in farming

After loading the data from csv files into the data frame using Pandas library, simple tasks were performed to get an insight of data.

Initially, I worked on following the two data sets  
a. lp\_df (Livestock products)

b. wpt\_df (World population)

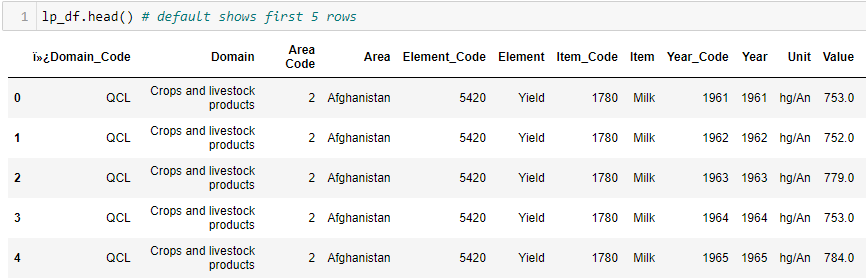


Figure 3.1 Display first five rows of the livestock products data frame

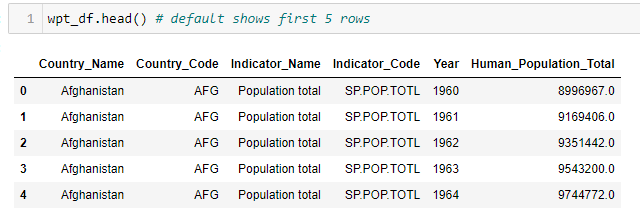


Figure 3.2 Display first five rows of the world population data frame

I also performed the following methods on the DataFrame to get more information

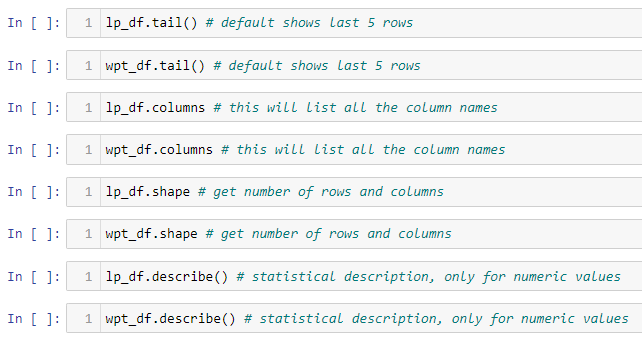


Figure 3.3 Common methods to get the insight of data

I also did some cleaning of data fin data frame lp\_df (Livestock products). I merged all unique “Element” row into column. [Please see section 2.0.3 in jupyter notebook]

*Merging the data frames*

Data frame lp\_df (Livestock products) and wpt\_df (World population) were merged into one dataframe (Please see section 2.0.4 in jupyter notebook)

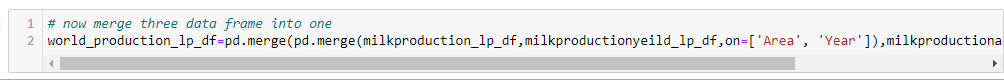


Figure 3.4 Merging two datasets

*Extraction of European country data from the main data frame*

This project evaluates milk production only in European countries so I need to extract European countries data from the main data frame world\_production\_inc\_lp\_df. Library countrygroups was used to get the list of countries in EU union. (Please see section 2.0.5 in jupyter notebook)

## 3.1 Dealing with missing data and data cleaning

One of the most common issues in dataset are missing values. Data can either be missing during data extraction or collection due to corruption or data not being recorded. I checked both datasheets for missing and empty(zero) data.

Below are some formats that could be in the missing data:

* n/a
* NA
* —
* na
* NaN

Why is it important to deal with missing data?

Missing data are important because, depending on the type, they can sometimes bias your results. This means results may not be generalisable outside of study because data come from an unrepresentative sample.

**5. Conclusion**

# **6. References**

***Books***Andreas, C. M. and Sarah, G. and O'Reilly (2016) Introduction to Machine Learning with Python, Media, Inc.

Hands on Machine Learning with Scikit Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, September 2019, ISBN: 9781492032649.

Python Machine Learning Third Edition, Sebastian Raschka , Vahid Mirjalili , Copyright © 2017 Packt Publishing.

Discovering Knowledge In Data: An Introduction To Data Exploration, Second Edition, By Daniel Larose And Chantal Larose, John Wiley And Sons, Inc., 2014.

Chun-houh, C. and Wolfgang, H. and Antony, U. (2008) Handbook of Data Visualization, Springer publishing.

Wes, M. (2017). Python for Data Analysis, 2nd Edition.

Perez (2021). Data Mining. The CRISP-DM Methodology. The CLEM language and IBM SPSS MODELER.

Ojala and Garriga (2010) Permutation Tests for Studying Classifier Performance.

***Websites***

Google (2022) Machine Leaning. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values. (Accessed: 17 Nov 2022).

Google (2022) Exploratory data analysis. Available at: https://businessanalyst.techcanvass.com/objective-of-exploratory-data-analysis/ (Accessed: 11 Nov 2022).

Google (2022) Handling outliers. Available at: https://towardsdatascience.com/exploratory-data-analysis-topic-that-is-neglected-in-data-science-projects-9962ae078a56. (Accessed: 12 Nov 2022).

Google (2022) Normal Distribution. Available at: https://www.askpython.com/python/normal-distribution (Accessed: 12 Nov 2022).

Google (2022) Pedestrian footfall. Available at: https://www.retailsensing.com/people-counting/count-pedestrian-footfall-town-centre/ (Accessed: 22 Nov 2022).

Google (2022) Linear regression. Available at: https://realpython.com/linear-regression-in-python/(Accessed: 20 Nov 2022).

Google (2022) Supervised Leaning. Available at: https://www.sciencedirect.com/topics/computer-science/supervised-learning (Accessed: 15 Nov 2022).

Google (2022) Data Science project management methodologies. Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb (Accessed: 13 Nov 2022).

Google (2022) Descriptive analysis. Available at: https://www.analyticssteps.com/blogs/overview-descriptive-analysis (Accessed: 18 Nov 2022).

Google (2022) Descriptive analysis. Available at: https://www.labxchange.org/library/items/lb:LabXchange:10d3270e:html:1 (Accessed: 22 Nov 2022).

# **5. Appendix**

## 5.1 Gantt Chart