# Group ID - MSc in Data Analytics

Author: Luiz Lopes

e-mail: [sba22211@student.cct.ie](mailto:sba22211@student.cct.ie)

Student ID: sba22211

Contents

[Group ID - MSc in Data Analytics 1](#_Toc123930469)

[Abstract 3](#_Toc123930470)

[1. Introduction 4](#_Toc123930471)

[1.1. Version Control 4](#_Toc123930472)

[1.2. Project Framework 4](#_Toc123930473)

[1.3. Project Background 5](#_Toc123930474)

[1.4. Data collection 5](#_Toc123930475)

[2. Data Extraction and Data Cleaning 5](#_Toc123930476)

[2.1. Data Extraction 5](#_Toc123930477)

[2.2. Data Cleaning 7](#_Toc123930478)

[3. Exploratory Data Analysis 9](#_Toc123930479)

[4. Statistical Analysis 15](#_Toc123930480)

[4.1. Descriptive Statistics 15](#_Toc123930481)

[4.2. Inferential Statistical Techniques 17](#_Toc123930482)

[4.2.1. T-Test one Way 17](#_Toc123930483)

[4.2.2. T-Test Two Way 18](#_Toc123930484)

[4.2.3. ANOVA One Way 18](#_Toc123930485)

[4.2.4. ANOVA Two Ways 20](#_Toc123930486)

[4.2.5. U-Mann Whitney 20](#_Toc123930487)

[4.3. Outcome and Findings 20](#_Toc123930488)

[5. Machine Learning 21](#_Toc123930489)

[5.1. Analysing the Models 21](#_Toc123930490)

[5.1.1. Linear Regression 21](#_Toc123930491)

[5.1.2. Polynomial Regression 22](#_Toc123930492)

[5.1.3. Ridge Regression 23](#_Toc123930493)

[5.1.4. Lasso Regression 23](#_Toc123930494)

[5.2. Cross Validation and GridSearchCV 24](#_Toc123930495)

[5.3. Machine Learning Conclusion 25](#_Toc123930496)

[6. Dashboard 25](#_Toc123930497)

[7. Conclusion 27](#_Toc123930498)

[8. References 28](#_Toc123930499)

# Abstract

# Introduction

The Podcast Reviews dataset, available on Kaggle, offers a valuable resource for exploring and analysing reviews from podcast listeners. As podcasts continue to gain popularity as a medium for entertainment, education, and storytelling, understanding listeners' feedback becomes crucial for content creators, researchers, and advertisers. This dataset provides a comprehensive collection of reviews, allowing researchers and data enthusiasts to delve into the sentiments, themes, and preferences expressed by podcast audiences across various genres and topics. By examining this dataset, valuable insights can be gained to improve podcasting experiences, tailor content strategies, and discover trends within the podcasting landscape.

With an ever-growing number of podcasts available, the Podcast Reviews dataset offers an opportunity to explore the diverse range of opinions and perspectives shared by podcast listeners. By analysing this dataset, researchers can gain a deeper understanding of the factors that contribute to listeners' satisfaction, engagement, and preferences. The dataset includes textual reviews, ratings, timestamps, and additional metadata, providing a rich set of information to uncover patterns, sentiment trends, and topics of interest. Whether you are interested in exploring the most popular podcasts, investigating correlations between ratings and reviews, or studying sentiment analysis, this dataset presents a valuable resource to enhance our understanding of the podcasting landscape and its audience.

## Version Control

I have used GitHub for Version control. I have set up a new account under the college email and set up the repository to Public. I have also installed the desktop version and all pull requests where managed through the desktop version.

Project could be found at: <https://github.com/LuizLopes-cct/CA2>

Some of the initial research performed prior to my final choice of data set is stored in a folder called OLD. There are previous downloaded data sets from different sources. After creating the data frames and perform a quick EDA on these data, it was clear that the data was not suited for the project.

The files on this folder are not relevant for this project and can be ignored.

## Project Framework

For the project framework, I have chosen to use CRISP-DM. not just because my personal familiarity with this framework but it because the Agile like approach proposed by CRISP-DM with an Adaptive life cycle.

At the beginning of the project, I was not sure what data to use, CRISP-DM gave me the flexibility to test different datasets. Also, the objectives were not well defined at the beginning, meaning that as I progressed with the Data Understanding I could update the Business Understanding phase accordingly.

This assignment paper is formatted to fit the subjects, but we could clearly perceive the CRISP-DM phases through it.

Data Understanding and Data Preparation phases are divided between chapters 2- Data Cleaning, 3- EDA and chapter 4- Statistics.

The Modelling stage is divided between chapters 2-Data Cleaning, 3-EDA and chapter 5-Machine Learning.

The remaining phases, Evaluation and Deployment concentrated on the Machine Learning chapter

## Project Background

## Data collection

# Data Extraction and Data Cleaning

## Data Extraction

## Data Cleaning

# Exploratory Data Analysis

# Statistical Analysis

Some of the tables and charts regarding the Descriptive Statistics are already presented and explained at the EDA part of the Assignment. I have chosen to repeat them in the Jupyter Notebook CA2\_Statistics.ipynb, so they can be used as a reference to the calculations in the file and to facilitate the readability of the document.

# Machine Learning

For the machine learning models, I have analysed the combination between the Exported Quantity and the Export Value. I want to estipulate, as a potential customer what would be the best price per quantity. We are only analysing the data from the past 20 years for the top 3 exporters (Excluding Poland).

# Dashboard

I have tried different libraries for creating the dashboards, I tried to avoid creating an .exe file for dashboarding, as I can see this technique hard to implement on a day to day. Many companies would block executables to run in collaborators computers due too Security issues. I only tested dashboards libraries with an option to present on the browser.

.

# Conclusion

Mushrooms and Truffles are the 3rd most exported vegetable from Ireland, majority of the Mushroom production in Ireland goes to UK. During our analysis we could see a big difference in price per tonne between the European countries. Ireland average price per Tonne is similar between most countries in the EU, close to Netherlands and Germany. Italy have the highest price and Poland have the lowest price in the top 10 exporters.

# References

Bari, A., Chaouchi, M. and Jung, T. (2014) *Predictive analytics for dummies*. Hoboken, NJ: John Wiley & Sons, Inc.

*Mushroom promotion 2019* (no date) *Bord Bia - The Irish Food Board*. Available at: https://www.bordbia.ie/industry/news/food-alerts/mushroom-promotion-2019/ (Accessed: January 6, 2023).

*Statsmodels.stats.oneway.anova\_oneway¶* (no date) *statsmodels*. Available at: https://www.statsmodels.org/dev/generated/statsmodels.stats.oneway.anova\_oneway.html (Accessed: January 6, 2023).

Theobald, O. (2021) *Machine learning for absolute beginners: A plain English introduction (third edition)*. Independently Published.