**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Statistics for Data Analytics  Programming for Data Analytics  Data Preparation and Visualisation  Machine Learning for Data Analytics |
| **Assessment Title:** | [CA2 50%](https://moodle.cct.ie/mod/assign/view.php?id=115129) Integrated Assessment |
| **Lecturer Name:** | John O'Sullivan, Sam Weiss, David McQuaid, Muhammad Iqbal |
| **Student Full Name:** | Zhongjie Fei |
| **Student Number:** | 2022173 |
| **Assessment Due Date:** | 06/01/2023(11/01/2023 after PMC) |
| **Date of Submission:** | 11/01/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Icon

Description automatically generated Author: Zhongjie Fei

Student ID: 2022173

**Report for Beef Price Analysis in Ireland**

**and**

**a quick comparison to Spain**

MSc in Data Analytics

*e-mail:* [*2022173@student.cct.ie*](mailto:2022173@student.cct.ie)

***Abstract***

*Keywords: beef, beef prices, data analysis, statistics, visualisation, machine learning*

**0. Introduction**

In

**1. Agriculture and beef production**

The emergence of agriculture, dating back thousands of years ago, marked a significant milestone in the historical development of human civilisation. It is the major source of food and income for majority of people around the world (Branco, 2020).

According to Branco (2020), food production in agriculture is a complex and multi-faceted process that involves a range of factors including farming size, soil management, water and electricity usage, pest and disease control, food security monitoring etc. Data science, for example artificial intelligence, now plays an important role in modern agriculture management, which have led to significant increases in food production, allowing for the feeding of growing global populations. However, the fast development of modern agriculture consequently raises concerns about the environmental and the potential loss of biodiversity (ibid.). Additionally, the use of chemicals for fertilising and pest control can have negative effects on soil and water quality but also on food safety. Therefore, organic farming and organic food production are being promoted in order to maintain sustainability.

Animal production is one of the two major branches in agriculture, alongside crop production. Within animal production, there are various sub-sectors, including beef production, dairy production, pig production, and poultry production, to name a few. Despite the increasing demand for chicken and other meats, beef production continues to play a crucial role in meeting the protein needs of the global population, providing essential nutrients and energy for human health and development (Lewis Kahn and David Cottle, 2014).

The price of beef is determined by a variety of factors, including supply and demand, production costs, and government policies. To elaborate, the demand for beef is affected by consumer preferences, population growth, and income levels, and the supply of beef is influenced by factors such as weather conditions, disease outbreaks, and production costs as well as government policies (Croxton, 1905). An increase in demand will lead to a higher price, however, an increase in supply will lead to a lower price. The production costs of beef encompass a wide range of expenses, e.g. equipment and facilities, veterinary care, feeding and labour.

**1.1 Common Agriculture Policy (CAP)**

The European Union's Common Agriculture Policy (CAP), first established in 1962 and regularly amended, plays a significant role in shaping the agricultural industry in Europe. It was proposed to, according to official website of European Commission:

1. Ensure the continuity of food production and distribution;
2. Ensure a supply of safe and affordable food;
3. Encourage younger generations to take up farming and promote environmentally friendly farming;
4. Practice LEADER method to thrive remote and mountainous areas with disadvantages;
5. Reduce power imbalance and help small farms etc.

The latest publication of the CAP places a stronger emphasis on environmental protection and organic farming, while also promoting fairness and innovation within the industry (European Commission, 2022).

As reported by the European Commission (2021), in Ireland in 2020, the allocation of CAP expenditure was 76.4% for direct payments, 19.9% for rural development, and 3.8% for market measures. Additionally, the beef (cattle) sector accounted for 29% of the output component, making it the second largest sector behind dairy. Feedingstuffs represented the highest percentage of intermediate consumption, at 47.2%.

**1.2 Beef carcass classification**

In

**2. Study objective**

The main purpose of this study was to investigate and visualise the dynamic changes in beef prices in Ireland and to analyse the relationship with other potential influencing factors such as feedingstuffs and beef production. Moreover, the study aimed to compare the beef prices and production in Ireland to that of Spain, another major beef-producing country in Europe. Given the assumption that beef prices in Ireland may differ significantly from those in Spain due to its high consumption level, several hypothesis tests were used to test this assumption. To further understand the data, two machine learning models were applied and compared. Finally, a sentiment analysis was implemented using Twitter API to observe people's recent (past 7 days) comments on beef prices worldwide.

Jupyter notebook version 6.4.8 and Python version 3.9.12 were used for exploratory data analysis (EDA), visualisation and machine learning in our report.

**3. Data collection and preparation**

Due to the complexity of beef prices and the study limitation, four datasets were collected to

**3.1 Processing of beef price dataset**

Table 1: Details of features in original data

|  |  |  |
| --- | --- | --- |
| FN | Attribute name | Description |
| 1 | type\_stands | types of stands in Dublin (7 types) |
| 2 | X | longitude |
| 3 | Y | latitude |
| 4 | Easting | easting |
| 5 | Northing | northing |
| 6 | location\_stand | address where stands locate |
| 7 | no\_stands | numbers of stands in the location |

**3.2 Processing of beef production dataset**

In regards

**3.3 Processing of beef feeding price dataset**

**3.4 Processing of pig meat price dataset**

**4. Statistics**

In this section, we firstly briefly have an overview of . Statistical analysis is integrated in between.

**4.1 Basic statistics**

T

Table

Description automatically generated

**Figure 1: First half** **of the basic statistics in relation to Ireland**

Table

Description automatically generated

**Figure 2: Second half of the basic statistics in relation to Ireland**

*Overview:*

* The average bull

Graphical user interface

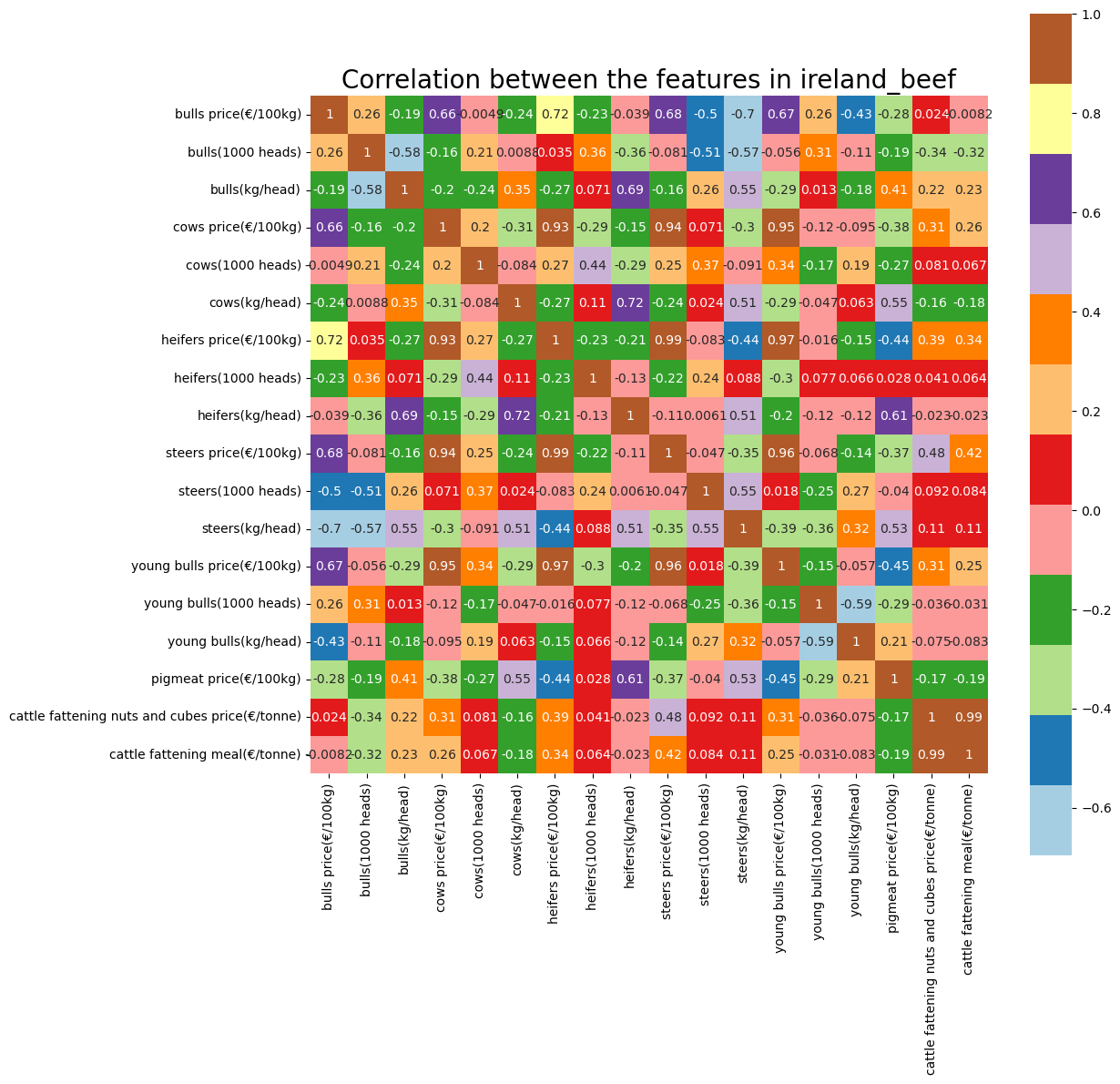
Description automatically generated with low confidence

**4.2 Hypothesis test**

A circle map function was implemented after. The advantage of this visualisation is

**5. Visualisation**

In this section,



Chart, histogram

Description automatically generated

Graphical user interface, chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart, line chart, histogram

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

**6. Machine Learning and sentimental analysis**

In our study, we applied two supervised learning models to train and test our data and make predictions. It is worth mentioning that since the original categorical data stand\_type is practically distinctive, there is no point in learning on this feature. So we brought in the five areas of Dublin as our dependent learning feature. Before starting machine learning, we encoded feature stand\_type using installed package.

**6.1 Linear regression**

KNN

**6.2 Sentimental analysis**

before scaling was 0.44, and after it increased to 0.92. Besides, there are two important arguments (i.e. C and gamma) that can potentially affect our model. GridSearchCV was used to test and find out the best hyperparameter. However, after we implement the best scores of these two arguments, the accuracy stayed the same.

**7. Conclusion**

In order to promote cycling, the assembly of abundant and safe bike parking facilities is extremely important. Through our study, we found out that in Dublin there is not only a lack of diversity in the type of bike parking stands but also imbalanced distribution across the city. The analysis may address the required attention for other areas, northside especially, as well as the openness into other type of stands.

**8. Discussion**

In this section, the author discusses and reflects on this study.

The author is aware of the possible biases generated by the specific handling method we chose to deal with missing value in our data. Deletion method is plausible since the percentage of the missing values is lower than 5% in the whole dataset.

The target audience for this study is aiming mainly to the Dublin City Council. If extra data is collected considering the cyclists preference. For instance, the promotion of Bike Locker ([*https://www.bikelocker.ie/*](https://www.bikelocker.ie/)) and Cyc-Lok([*https://cyc-lok.ie/*](https://cyc-lok.ie/)), different insights may be received consequently.

Lastly, the author is also aware the difference of handling geospatial data than normal data and the variety of tools that can be used to generate better analysis, such as GDAL and so on.

**References**

Ali, M.M. *et al.* (2021) ‘Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison’, *Computers in Biology and Medicine*, 136, pp. 1-10 Available at: <https://doi.org/10.1016/j.compbiomed.2021.104672>.

Aslanyan, T. (2021) *Fundamentals Of Statistics For Data Scientists and Analysts Towards Data Science*. Available at: <https://towardsdatascience.com/fundamentals-of-statistics-for-data-scientists-and-data-analysts-69d93a05aae7> (Accessed: 10 November 2022).

Dublin City Council(2021) *Cycle Parking.* Dublin City Council. Available at: <https://www.dublincity.ie/residential/transportation/active-travel/cycling-dublin-city/cycle-parking> (Accessed: 14 October 2022).

Dublin City Council (2022a) Public Cycle Parking Stands DCC. DATA.GOV.IE. <https://data.gov.ie/dataset/dcc_public_cycle_parking_stands?package_type=dataset>.

Dublin City Council (2022b) Administrative Areas DCC. DATA.SMARTDUBLIN.IE. <https://data.smartdublin.ie/dataset/administrative-areas-dcc>.

Dún Laoghaire-Rathdown County Council Municipal Services Department (2018) *Standards for Cycle Parking and associated Cycling Facilities for New Developments*. Dún Laoghaire-Rathdown County Council Municipal Services Department. <https://www.dlrcoco.ie/sites/default/ffles/atoms/ffles/dlr_cycle_parking_standards_0.pdf>

El Naqa, I. and Murphy, M.J. (2015) ‘What Is Machine Learning?’, in I. El Naqa, R. Li, and M.J. Murphy (eds) *Machine Learning in Radiation Oncology: Theory and Applications*. Cham: Springer International Publishing, pp. 3–11. Available at: <https://doi.org/10.1007/978-3-319-18305-3_1>.

Egan, R., Dowling, C.M. and Caulfield, B. (2022) ‘Planning by Cycle Parking Type: A Cycle Parking Preference Typology for Cyclists’, *Case Studies on Transport Policy*, 10, pp. 1930–1944. Available at: <https://doi.org/10.1016/j.cstp.2022.08.007>.

Faisal, S. (2018) *Nearest Neighbor Methods for the Imputation of Missing Values in Low and High-Dimensional Data*. Göttingen: Cuvillier Verlag.

Holdgraf, C. and Wasser, L. (2018) *Geographic vs projected coordinate reference systems - GIS in Python*, *Earth Data Science - Earth Lab*. Available at: <https://www.earthdatascience.org/courses/use-data-open-source-python/intro-vector-data-python/spatial-data-vector-shapefiles/geographic-vs-projected-coordinate-reference-systems-python/> (Accessed: 26 October 2022).

*Irish Grid Reference Finder* (no date). Available at: <https://irish.gridreferencefinder.com/> (Accessed: 27 October 2022).

Janssen, V. (2009) ‘Understanding coordinate reference systems, datums and transformations’, *International Journal of Geoinformatics*, 5(4), pp. 41–53.

Kanani, B. (2020) ‘Pandas - How to remove DataFrame columns with only one distinct value?’, *Machine Learning Tutorials*. Available at: <https://studymachinelearning.com/pandas-how-to-remove-dataframe-columns-with-only-one-distinct-value/> (Accessed: 31 October 2022).

Klokan Technologies, G. (2022) *EPSG.io: Coordinate Systems Worldwide*. Available at: <https://epsg.io> (Accessed: 27 October 2022).

Li, X.-B. (2009) ‘A Bayesian Approach for Estimating and Replacing Missing Categorical Data’, *Journal of Data and Information Quality*, 1(1), Article 3, pp. 1-11. Available at: <https://doi.org/10.1145/1515693.1515695>.

McKinney, W. (2018) *Python for data analysis: data wrangling with pandas, NumPy, and IPython Wes McKinney*. Second edition. Beijing: O’Reilly.

Mukhiya, S.K. and Ahmed, U. (2020) *Hands-On Exploratory Data Analysis with Python : Perform EDA Techniques to Understand, Summarize, and Investigate Your Data*. Birmingham, UK: Packt Publishing.

Nakai, M., Chen, D.-G., Nishimura, K. and Miyamoto, Y. (2014) ‘Comparative Study of Four Methods in Missing Value Imputations under Missing Completely at Random Mechanism’, *Open Journal of Statistics*, 4(1), pp. 27–37. Available at: <https://doi.org/10.4236/ojs.2014.41004>.

Peng, J., Wu, W., Lockhart, B., Bian, S., Yan, J.N., Xu, L., Chi, Z., Rzeszotarski, J.M. and Wang, J. (2021) ‘DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python’, in *Proceedings of the 2021 International Conference on Management of Data*. New York, NY, USA: Association for Computing Machinery (SIGMOD ’21), pp. 2271–2280. Available at: <https://doi.org/10.1145/3448016.3457330>.

Sebastian Raschka and Vahid Mirjalili (2019) *Python Machine Learning : Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow 2, 3rd Edition*. [S.l.]: Packt Publishing.

Squire, M. (2015) *Clean Data*. Birmingham, UK: Packt Publishing.