CCT College Dublin

Assessment Cover Page

To be provided separately as a word doc for students to include with every submission

---

Module Title:	Programming for DA Statistics for Data Analytics Machine Learning for Data Analysis Data Preparation & Visualisation
Assessment Title:	MSC_DA_CA2
Lecturer Name:	Marina Iantorno Sam Weiss Muhammad Iqbal David McQuaid
Student Full Name:	Jan Andersson
Student Number:	sba20368
Assessment Due Date:	6th January 2023
Date of Submission:	06/01/2023

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.

Github

https://github.com/JAnderssonCCT/MSC_DA_CA2

Contents

Introduction

Irish dairy farming is based on a system of small, family-owned farms. The majority of Irish dairy farmers are still working the land in this way and producing milk for their own consumption. The main reason why Irish dairy farmers have not diversified into other business areas is because it's too expensive to do so. There are only very few large dairies that can afford to buy directly from small scale farms; and even then, they don't always get what they want as some farmers prefer to sell direct rather than through a middleman who would take a percentage off the top – which often means that consumers end up paying more for their milk products than they should be doing. Irish dairy farmers are very proud of their products and they want to make sure that consumers know exactly what's in the product.

They also want to be able to sell directly to the consumer; and this means that Irish dairy farmers have a great deal of control over their own business. The main problem with direct selling is that it can lead to higher prices for consumers, so some people would prefer if there was a way for small scale farms such as these to sell into larger markets without having all of their costs passed on through middlemen. This is where dairying cooperatives come in – Irish cooperative businesses work by pooling together money from different members who then invest it in order to get more milk production going. The result is that you end up with a much cheaper product than you would otherwise be getting, but you still get quality because each farmer has an equal say in how things should run and they must all abide by the same rules when it comes down to selling direct or selling through middlemen. It's not uncommon for Irish farmers who have joined an agricultural cooperative such as this one – called Coillte Teoranta (the National Dairy Board) – will receive around €50 per tonne more per year than those who produce milk themselves but don't use any other means of distribution besides retail outlets like supermarkets or shops; and many families see this kind of extra cash coming into the farmhouse every month as enough reason alone for them not only continue milking cows but also keep doing so through good times and bad! .

It's said that Ireland is one of the top 10 dairy producing countries in the world, and it's not hard to see why. The country has a long history of co-operatives; this kind of business structure was originally set up by small farmers in the 19th century who wanted to be able to sell their milk directly to consumers for a fair price. It grew from there, with people joining together over time and forming bigger co-ops which would then go on to provide services such as insurance, banking and even legal advice. However, things took a turn for the worse when supermarkets began taking over more and more aspects of our food supply chain – they started buying milk direct from farms at much lower prices than what farmers were being paid before going through middlemen like dairies or cooperatives.

[535 Words]

Project aims

The use of Data Analytics may be quite beneficial in dairy farming. It is beneficial to farmers for a variety of reasons, including:

1- Estimating milk output. It is simpler to arrange your milking routine if you know how much milk your cows make each day and how much they will provide over time. This enables you to improve cow wellness while avoiding having too few or too many cows available during any particular moment.
2- Recognizing issues with your herd. Identifying what's incorrect with your herd, for instance when an animal isn't giving enough dairy, allows you to address concerns before they become larger ones with long-term ramifications for the overall farm business if left unchecked. You may even be able to avoid them from happening in the future by carefully following their activity
3- Herd management. Understanding what really is going on with your cattle might assist you in determining how to govern your herd the best. For instance, if a cow isn't providing enough dairy, it may be time to give her a dry spell or have her bred so she may produce more milk in the future. This might save you cost while also providing you with higher-quality dairy products.
4- Applying data analytics to boost efficiency and productivity. Data analytics may also allow farmers make better judgements regarding their livestock and their business as a whole, which can enhance overall farm operations. Using this type of technology, farmers can track the condition of their pastures over extended periods of time, enabling them to make informed choices while still on the farm, rather than having those decisions made wirelessly from afar by an individual who may not notice all of the specifics of each circumstance at this time.
5- Making better choices. Data analytics may also be utilised to assist farmers in making better informed decisions regarding their livestock and their whole enterprise. Farmers can far more easily monitor the health of their cattle over lengthy periods of time by employing this sort of equipment.
6- Asset management. This is highly relevant if you have a lot going on with your herd. For instance, if you have an old cow that has begun generating less dairy over time or one that is becoming ill due to poor health, data analytics can help you evaluate if she ought to be put down or managed to sell off so you wouldn't make a loss.

[406 Words]

About the datasets

Irish Cattle Births

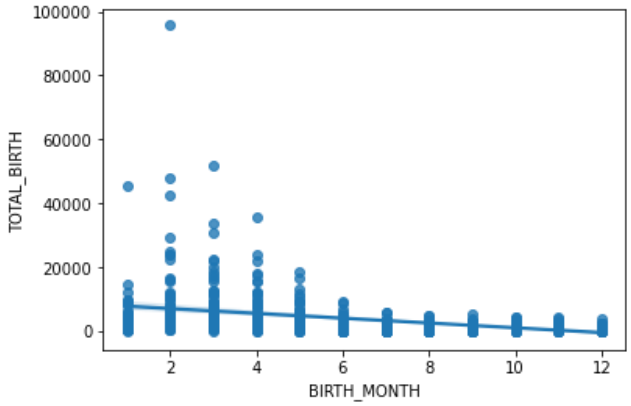
This data was acquired from the EU international database of public statistics, it contained information on the cattle births within Ireland with features describing the county of origin and month and birth year. I have decided to I converted the data into categorical for the month of cattle birth and breed type.
Where:

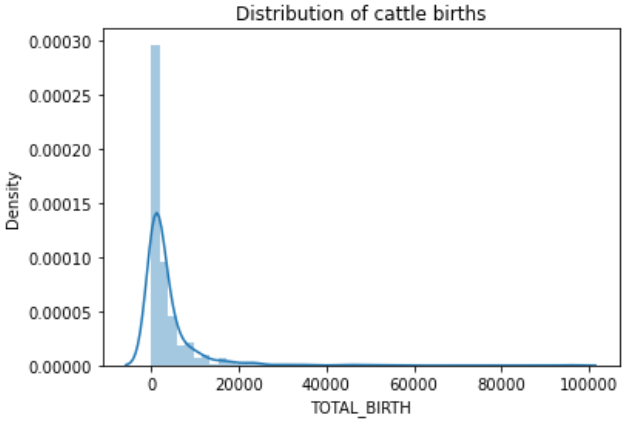
JAN:1 FEB:2 MAR:3 APR:4 MAY:5 JUN:6 JUL:7 AUG:8 SEP:9 OCT:10 NOV:11 DEC:12
and
BEEF:1 DAIRY:2

When measured by confidence interval for population mean size, there is a 95% chance that the confidence interval of [3103.843027719479, 4206.086459460009] contains the true population mean of cattle Births.

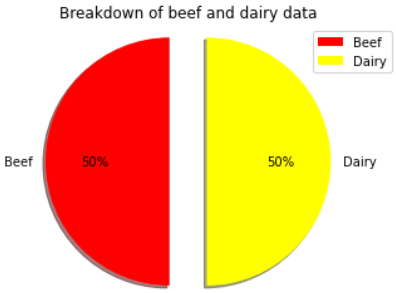
And there is a 99% chance that the confidence interval of [2932.075149970884, 4377.854337208603] contains the true population mean of cattle Births.

I converted the column to an *int* data type after encoding it as a numeric value which allowed me to use seaborn to visualize the confidence interval of the total births against the months we encoded earlier. And also the distribution of the cattle births.





As

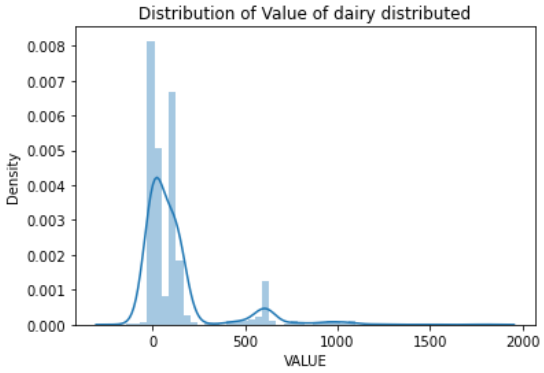


We can see the data for the cattle births is equally distributed for the Dairy and Beef. Meaning the data has a valid number of entrees for both groups of cultivated animals and holds integrity.

[198 Words]

Irish Dairy Production

This data was acquired from the EU international database of public statistics. The data was selected due to its high potential in format and had equal well-preserved data with very few null values that need to be rectified.



Here is the distribution of the dataset value of dairy products.

I have used Stratified Sampling; the population is divided into groups based on characteristics. The two data characteristics I chose for the strata are 'Year' and 'Product Type.

I dropped the unnecessary column as we have no background information on what data that feature represents or a reference to transform the data into a usable source. For instance the feature titled ‘C02064V02491’ had no information on the data.

Within the dataset for the sales of dairy throughout the years in Ireland we have 5 different categories representing the items as displayed with the value counts of the product type feature each one of an integer data type.

```
Cheese                        220
Butter                        220
Cream                         220
Milk Powder                   220
Drinking Milk & Buttermilk    220
```

[198 Words]

EU Dairy cattle supply

Country Key ID's for Numeric Dataframe

This data was acquired from the EU international database of public statistics, I have chosen it as it is the most reliable source and has the most relevant and up to date historical data.

I have supplied the country codes for reference with use with the numeric only dataframe of EU Dairy cattle stocks.

- 0: Austria
- 1: Belgium
- 2: Bulgaria
- 3: Croatia
- 4: Cyprus
- 5: Czechia
- 6: Denmark
- 7: Estonia
- 8: Finland
- 9: France
- 10: Germany
- 11: Greece
- 12: Hungary
- 13: Ireland
- 14: Italy
- 15: Lithuania
- 16: Luxembourg
- 17: Malta
- 19: Netherlands
- 20: Poland
- 21: Portugal
- 22: Romania
- 23: Slovakia
- 24: Slovenia
- 25: Spain
- 26: Sweden

As we can see the data for the EU countries is formatted where the year is the column featuring rows with the corresponding countries within the EU. I have removed the unnecessary and unformatted data as features such as *Unnamed: 0* have no use in any data analysis without the documentation that is not present in this case.

I then created a separated dataframe which contained the EU dairy cattle distribution and formatted it to remove special characters any values that contained strings and converted the whole dataframe to an integer. As the year and country wasn't in the correct format, I transposed the dataframe to make the feature selection the countries within the dataframe rather than the year.

Here I have displayed the averages for the selected countries to get a better understanding of what country might have a similar distribution of the cattle quantity.

```
0      557.826087
1      548.695652
2      320.913043
3      103.608696
4       26.260870
5      416.695652
6      585.260870
7      105.652174
8      301.913043
9     3830.695652
10    4267.173913
11     136.956522
12     279.956522
13    1155.869565
14    1927.869565
15     175.217391
16     369.086957
17      46.391304
18       5.956522
19    1568.565217
20    2570.086957
21     269.608696
22    1104.304348
23     176.739130
24     118.826087
25     948.565217
26     368.521739
```

[200 Words]

Irish and EU dairy market research

The European Union is one of the world's greatest economic areas. It is the world's most populous area, with a population of approximately 510 million people. The EU is made up of 27 member nations spread over three continents: Europe, Asia, and Africa. This region's economy is quite diversified and is dependent on the specific requirements and resources of each country.

The European Union produces about $1 trillion in commodities each year, making it one of the world's top dairies producing areas (European Commission, 2019). Belgium, France, Germany, Italy, the Netherlands, Poland, Portugal, and Spain are the ten member countries that export dairy products. Sweden United Kingdom, Ireland, and Norway Greece (European Commission) (European Commission).

Other three nations with considerable exporters are Ireland ($18 billion), Greece ($17 billion), and Norway ($13 billion) (European Commission). These five nations generate more than 80% of the dairy products sold in Europe.

Second, four additional EU states have large amounts to export: the Netherlands (7%), Poland (6%), and Greece (4%). These eight nations account for around 23% of total EU total export, or almost half of total EU export value (European Commission). This demonstrates that two-thirds come from just five countries, while one-third comes from eight other countries, making them equivalent benefactors to the income of this economic zone.

[215 Words]

EU Supply data analysed and compared

In order to find the correct country with similar supply of dairy cattle numbers I must run a few inferential statistical techniques to compare them.

Hypothesis

The population distribution of Dairy cattle has been equal between the country of Romania and Ireland according to thew EU data statistics collected from 1998 to 2020.

Z-test

The Z Test can be used to ascertain if there is a significant difference in amounts between the countries. In this scenario, the null hypothesis is that the growth and quantity of dairy cattle in both countries are equal. The hypothesis test would enable us to support or refute this claim. Usually, for hypothesis tests, a 5% level of significance is applied and the claim is rejected if the p-value is less than the level of significance.

Result: `(0.0, 1.0)`

T-test

The T-Test has a similar purpose as the Z-Test. However, it is applied when the population standard deviation is not known, or for samples with small sample sizes usually where the sample size is lower than 30.

Result: Ttest_1sampResult(statistic=1.7809032442726092, pvalue=0.08874237530230429)

Pearson Correlation Test

The correlation test tests if the relationship between these variables is statistically significant. The Pearson Correlation Coefficient is a popular correlation coefficient that measures the linear relationship between 2 variables. In this case I used Pearson Correlation.

In this case I will be testing the correlation of dairy cattle quantity between the two countries of Ireland and Spain for analysis in patterns between the country's growth in Dairy farming.

The `pearsonr` function on Scipy returns the correlation coefficient and tests if the correlation is significant

Conclusions

Unfortunately, the test resulted in a negative correlation for unknown reasons as the average and general series of data is the most similar between these 2 countries in both given cases as proven by the dataset data itself and research conducted.

It has however shown us that the correlation between the Irish and Spanish data is greater than the Irish and Romanian data.

Wilcoxon Test

The Wilcoxon signed-rank test is the non-parametric univariate test which is an alternative to the dependent t-test. It also is called the Wilcoxon T test, most commonly so when the statistic value is reported as a T value. Which scipy.stats.wilcoxon() uses for it's calculation. This is the recommended test to use when the data violates the assumption of normality; which is the case with this data so far.

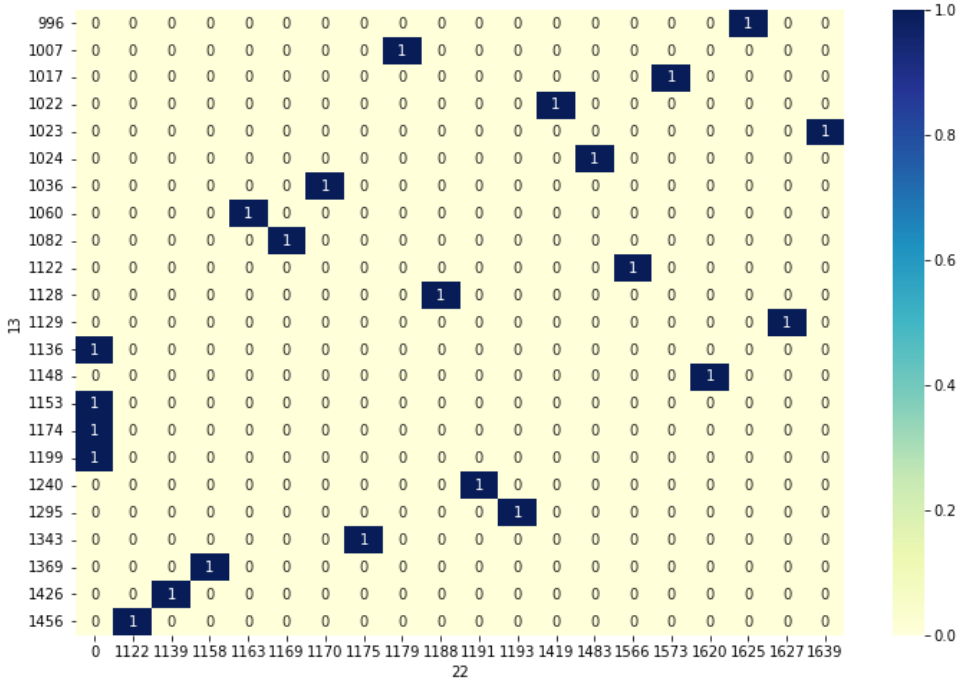
Both of the variables have a significant p-value which means each variable violates the assumption of normality. Therefore, the Wilcoxon signed rank test, a.k.a the Wilcoxon T test, is the correct test to use to analyse this data.

Conclusions

As we can see the test proved more successful than the last series executed for analysis with a P-Value of `0.776785135269165`. This in my opinion derives from thew lack of normal distribution within the dataset structure.

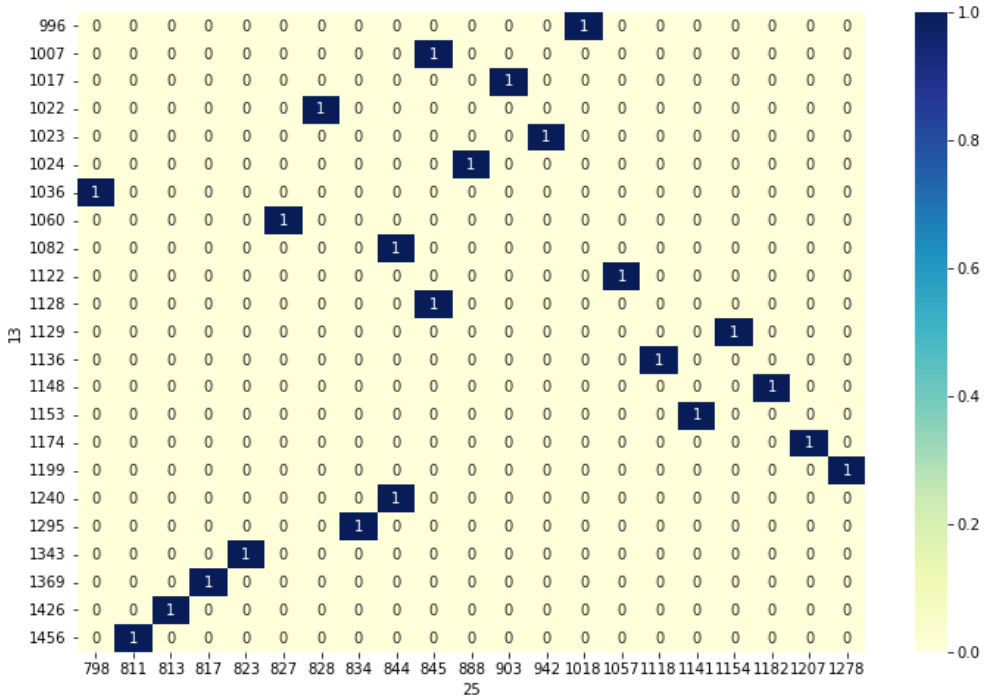
```
Ireland:
ShapiroResult(statistic=0.8945345878601074, pvalue=0.01953703537583351)
Romania:
ShapiroResult(statistic=0.7398244738578796, pvalue=4.8182504542637616e-05)
WilcoxonResult(statistic=128.0, pvalue=0.776785135269165)
```

Chi-Square (Ireland and Romania)



P-Value: 0.2513008139441771

Chi-Square (Ireland and Spain)



P-Value: 0.24615506857308492

[498 Words]

References

Awojide, Margaret. “Statistics for Data Analysts: Inferential Statistics with Python.” *CodeX*, 14 Sept. 2022, medium.com/codex/statistics-for-data-analysts-inferential-statistics-with-python-de8b7f49cfa. Accessed 5 Jan. 2023.

Duca, Angelica Lo. “How to Build a Dataset from Twitter Using Python Tweepy.” *Medium*, 22 Apr. 2022, towardsdatascience.com/how-to-build-a-dataset-from-twitter-using-python-tweepy-861bdbc16fa5. Accessed 6 Jan. 2023.

Hirsch, Stefan, et al. “Dataset on the Dairy Processing Industry in France, Italy, Poland and Spain.” *Www.research-Collection.ethz.ch*, 2019, www.research-collection.ethz.ch/handle/20.500.11850/333174, 10.3929/ethz-b-000333174. Accessed 4 Jan. 2023.

“How to Get Tweets Using Python and Twitter API.” *Quantitative Finance & Algo Trading Blog by QuantInsti*, 11 July 2022, blog.quantinsti.com/python-twitter-api/. Accessed 6 Jan. 2023.

“K-Means Clustering with Python.” *Kaggle.com*, www.kaggle.com/code/prashant111/k-means-clustering-with-python.

“Python Pandas - Descriptive Statistics - Tutorialspoint.” *Www.tutorialspoint.com*, www.tutorialspoint.com/python_pandas/python_pandas_descriptive_statistics.htm.

“Python Z Test | When to Perform Z Test in Python with Examples?” *EDUCBA*, 13 Feb. 2022, www.educba.com/python-z-test/.

RAHMAN, KALILUR. “How to Avoid KERAS Import Errors in Your Notebooks - a Solution | Data Science and Machine Learning.” *Www.kaggle.com*, www.kaggle.com/general/274656. Accessed 6 Jan. 2023.

Shane. “Delete Rows & Columns in DataFrames Using Pandas Drop.” *Www.shanelynn.ie*, www.shanelynn.ie/pandas-drop-delete-dataframe-rows-columns/.

Stojiljković, Mirko. “Python Statistics Fundamentals: How to Describe Your Data – Real Python.” *Realpython.com*, realpython.com/python-statistics/.

Thu, et al. “Irish Dairy’s Elevated Position across Range of Global Metrics.” *Irish Examiner*, 7 Oct. 2021, www.irishexaminer.com/farming/arid-40715282.html.

“Twitter Sentiment Analysis.” *Kaggle.com*, www.kaggle.com/code/paoloripamonti/twitter-sentiment-analysis. Accessed 6 Jan. 2023.

“UK and EU Cow Numbers | AHDB.” *Ahdb.org.uk*, 2019, ahdb.org.uk/dairy/uk-and-eu-cow-numbers.

Zach. “How to Calculate Confidence Intervals in Python.” *Statology*, 16 July 2020, www.statology.org/confidence-intervals-python/.

---. “Pandas: How to Remove Special Characters from Column.” *Statology*, 10 Oct. 2022, www.statology.org/pandas-remove-special-characters/. Accessed 4 Jan. 2023.