Zdenko Zahorec

MSc in Data Analytics (SB+) - Feb 2022 - YR1

Lecturers:

Marina Iantorno

David McQuaid

Dr. Muhammad Iqbal

Sirloin Steak price analysisin the USA and Ireland from 1997-2021

Contents

[Abstract 3](#_Toc104079568)

[1. Introduction 4](#_Toc104079569)

[2. Analysis 4](#_Toc104079570)

[I. Sirloin steak price change in the USA in Ireland over the last 25 years 4](#_Toc104079571)

[II. Sentiment Analysis 21](#_Toc104079572)

[3. Conclusion 22](#_Toc104079573)

[4. References 23](#_Toc104079574)

[5. Appendices 25](#_Toc104079575)

[I. **Appendix** Mood’s Median Test 25](#_Toc104079576)

[2. **Appendix 2.** Kruskal-Wallis Test 26](#_Toc104079577)

[*3.* Appendix 3. *Mann-Whitney U test* 27](#_Toc104079578)

[*4.* Appendix 4. *Spearman’s r Test* 27](#_Toc104079579)

[5. **Appendix X –** Project Timeline 28](#_Toc104079580)

[Figure 1 Cordons 4](#_Toc102390571)

# Abstract

Irish Agriculture…

# Introduction

The agricultural sector is characterized by low operational efficiency, a high level of uncertainty because of weather and environmental conditions, and a volatile balance between food supply and demand (Osinga et al., 2022). Through the application of the ‘Common Agricultural Policy (CAP), European Union is trying to promote a sustainable agricultural system by combining several social, economic, and environmental approaches (European Commission - European Commission, 2021). Several EU initiatives we created do help further with the application of sustainable agriculture (‘Agricultural productivity and sustainability (EIP-AGRI) (2012); FOOD 2030 initiative (2016). All these initiatives are trying to find better business models and secure investments for farmers that will increase productivity, improve skills, and provide sustainable high-quality food (European Commission - European Commission, 2021). Cockburn (2020) argues that although large amounts of dairy data are becoming available, a lack of data integration makes it difficult to analyze the data. In their research, where they analyzed peer-reviewed research about machine learning in the dairy sector, findings were that tested algorithms did not perform to that extent where they could be used in the practice. The conclusion states that the main reason for this is the lack of data integration and availability of public data. Osinga et al. (2022) also mentioned that data analysis has not been widely applied in the agricultural sector, and as the blockers, they identify a lack of human resources and expertise, non-reliable infrastructure, and a lack of data standardization and governance.

# Analysis

In our research, we will compare Sirloin steak price change in the USA in Ireland over the last 25 years. We will also review sentiment analysis from tweets containing the words ‘beef production that relates to manufacturing, and we will also review sentiments from the consumer side, analyzing tweets containing the words ‘steak price’. Our initial hypothesis will state that Sirloin steak prices history in Ireland is related to Sirloin steak prices history in the USA. Before starting the data gathering process, we created a project plan timeline to break down tasks into periods (Appendix X.)

## Sirloin steak price change in the USA in Ireland over the last 25 years

In the initial search, we were searching for datasets from Ireland and any other country related to beef production. Central Statistics Office (www.cso.ie) was our first choice to find the data. We found 10+ datasets that relate to beef production (beef exports, average price, meat supply balance, etc.) and decided to focus our research on how steak prices changed over the years (Datasets downloaded from: <https://data.cso.ie/table/CPM12>, <https://data.cso.ie/table/CPM04>, <https://data.cso.ie/table/CPM08> ). The next step was to find related data from a different country. The first database that we investigated was the Eurostat database, but we were not able to find completely matching data. The search was expanded to worldwide data and after reviewing datasets from the UK, Argentina, and the USA, we decided to use the USA dataset (downloaded from: <https://data.bls.gov/pdq> ).   
Data analysis and comparison will be conducted mostly by using Jupyter Notebook code and some manual statistical calculations in the appendix part.

##### Initial Data Analysis

Exploratory Data Analysis (EDA) is one of the best approaches for the initial data analysis. Komorowski et al. (2016) describe EDA as a step where the data is visualized and manipulated without any assumptions.   
We will start EDA by importing relevant libraries and uploading three Irish datasets into the Jupiter notebook using the pandas library. Anthony (2015) says that the benefit of using panda lies in easy data representation via *DataFrames* and data series, easy filtering, and easy code writing. Visual inspection of the first Irish dataset shows that this dataset contains 8 variables and 3953 observations with monthly prices for different products. By using the *‘data set.describe()*’ function we can see the minimum and maximum values for the month column (‘TLIST(M1)’) to check the start and the end date of the data – which are January 1997 -November 2001. We repeat the same process with the second and third Irish datasets.  
The second dataset contains 8 variables and 10527 observations with monthly prices for different products from December 2001 till December 2012.   
The third dataset contains 8 variables and 9348 observations with monthly prices for different products from December 2012 till February 2022. As the second and third datasets contain a duplicate value for December 2011, we will remove that record from the third dataset. After validating that the duplicated record is removed, we need to see how many observations are related to ‘steaks’. We do this by creating *def function* in Python. In the next step, the function will keep only observations containing the string ‘steak’ and in the last step, the function will print those values for all three datasets. Outputs (In [12]) shows that all three datasets have only one steak type in common – Sirloin steak. The next step was to rename all observations from the third dataset from 'Sirloin steak per kg' to 'Sirloin steak per kg.' to have matching names as in the first two datasets. Now we are ready to merge all three datasets into one, restructure the dataset and keep only the data related to Sirloin steak. We used *for loop* function that will iterate through all three datasets with the following steps:

* Rename columns ('Consumer Item':'SteakType','TLIST(M1)':'Daystamp') to work with more meaningful column names.
* Keep only observations that contain string ‘Sirloin’
* Use *pivot.table* function to transform observations into columns and assign them values from the original ‘Values’ column
* Reset indexing and add the Index column name
* Append new data from each dataset into one list using *list.append* function

Once all observations are appended to the list, we *use pandas.concat* function to create a new DataFrame. Using df.info() function we can check the DataFrame format, variable data types, the number of observations and are there any blank values. Our new dataset has 302 observations, two columns, and no empty values. Next, we will transform the integer type column into a DateTime column type that is crucial when dealing with the time-series data analysis. Kirchgssner (2014) describes time series s a set of quantitative observations arranged in chronological order.   
We will also create new ‘Year’ and ‘Month’ columns that will be used to group and visualize data using plots.   
The final data cleaning step is to remove the last two observations from the Irish dataset that are for January and February 2022, to have year start end data only (from January 1997- to December 2021).  
Using the *df. describe function* we will get Pandas to calculate descriptive statistics values.

Text

Description automatically generated with medium confidence  
Figure 1. Irish Sirloin steak prices column description

The description shows that there are 300 observations. The minimum value found is 7.994, and the maximum observed value is 16.982. The average steak price over 25 years was 13.360 Euro, and the standard deviation shows how much data deviate from the mean – what is the variability of the data. In our case, the standard deviation from the mean is 2.065. Percentiles or quartiles tell us where our data exist in the dataset. We can say that 25% of data values are higher than 15.471, and less than 25% of those values are lower than 12.175. Grouping the data by year, we can see that the highest average price was recorded in 2013 and the lowest average price was in 1998.

Kirk (2012, pp.13-14) reasons that data visualization is more successful than data set exploration because human visual functions are faster and more efficient than human cognitive processes.

First, we will visualize the data using a Line plot that will show us prices over years.

Chart, line chart, scatter chart

Description automatically generated  
*Figure 2. Line plot IRL*

The plot shows that steak price was lowest in 1997-1999 and the highest price was recorded in 2013. Over 25 years price went through 3 major cycles of increase/decrease with the latest tendency being price decrease. It would be interesting to investigate other social/economic variables to understand what was driving these price changes.

The next visualization that we will use is a boxplot that will help us identify outliers. Mozaffari et al. (2021) state that a boxplot is an easy to compute and read outlier detection tool that uses the entire sample to determine the cut-off values.

Chart, waterfall chart

Description automatically generated  
*Figure 3. Boxplot IRL*

The next step is to replace outliers from the dataset using *group by* function by and quantiles calculation. This will replace all outlier’s values with their yearly median values.   
We will also plot Poisson distribution to see how many times the event is likely to occur over some time (Hu, 2008).  
To find out are variables correlated we will use a correlation matrix to see correlation values and a pair plot to visualize correlations.

Chart, treemap chart

Description automatically generated Chart

Description automatically generated  
 *Figure 4. Correlation Matrix IRL* *Figure 5. Pair plot IRL*

Matrix and the pair plot show that there is a strong positive correlation between the price and the year variables and a weak positive correlation between the price and the month variable.

Next, we will import the USA dataset and use the *dataset.head(*) function to review first 20 observations. Dataset has 6 columns, but we are interested only in 3 columns that contain data for steak price, year, and a month. The first 9 rows in the dataset contain table information and we can discard them from the model. We re-import the dataset again, but now without the first 9 rows.   
In the ‘Period,’ column first character is ‘M’ – probably short for the month. We remove the ‘M’ character from the string, and then merge the ‘Period’ column with the ‘Year’ column to get an integer type column with month/year values. By using *df.drop()* function, we remove 3 columns from the dataset that we won’t need anymore. Column ‘Year’ is renamed to ‘Daystamp’ to match the Irish dataset column name.   
The next step is to check for blank values (there are not any) and change the integer type column ‘Day stamp’ into a daytime type column. Also, we add back ‘Year’ and ‘Month’ columns to the dataset.   
The USA steak prices are shown in pounds, we multiply price values by 2.2 to match values in kilograms. Last step before we visualize data is to remove unmatching month/year data from the USA dataset. As the Irish dataset holds data from 1997 -to 2021, we remove values that are less than December 1996 or greater than January 2022 from the USA dataset. This will leave us with two matching datasets – they have the same number of columns and rows.

To visualize USA data, we repeat the steps that we did with the Irish dataset :

Chart, line chart, scatter chart

Description automatically generated  
*Figure 6. USA line plot*

The line plot shows that over the 25 years, the Sirloin steak price was almost constantly growing in the USA, with price spikes around 2004, 20015, and 2021. The lowest mean price was recorded in 1998 and the highest in 2021.   
We continue data visualization by plotting boxplots, checking, and replacing the outlier’s values.

Chart, waterfall chart

Description automatically generated  
*Figure 7. USA boxplot*

Finally, we plot the correlation matrix and pair plot to check is there a correlation between the variables.

Chart, treemap chart

Description automatically generated Chart, histogram

Description automatically generated  
 *Figure 8. USA Correlation Matrix Figure 9. USA pair plot*

Matrix show that there is a strong positive correlation between the steak price and the year variables (93%) and almost no correlation between price and the month.

##### Inferential statistics

Inferential statistics utilizes probabilistic techniques to analyze sample information from a certain population to improve our knowledge about that population (Asadoorian and Kantarelis, 2009). We used the following statistical techniques to review and compare data:

* + - 1. Linear Regression

Linear regression is based on linear correlation and the assumption that a change in one variable will result in a proportional change in another variable (Bazdaric et al., 2021). When using time series two features come into consideration: time-step and lag. The time feature creates time dependence and enables values to be predicted from the time that they occurred. We will create an index ‘Time’ column that will be used as an independent variable and the ‘price’ column will be our dependent variable. The fitted linear regression model will be plotted together with the existing time series values. The plotted line will show us the model trend:

Chart, scatter chart

Description automatically generated  
 *Figure 9. IRL linear regression model*

* + - 1. Lag feature

Lag operators are useful in time series analysis as they can show autocorrelation within the values with previous copies of themselves. This can identify patterns within the time series and help determine the seasonality of the data (Matt Dancho, 2017).  
We will use *data.shift()* function to create a new column (‘Lag’) with price values shifted for one place down and predict and plot the model:

Chart

Description automatically generated  
 *Figure 10. IRL lag feature model*

* + - 1. Mood’s Median Test (Appendix 1.)

This test tests do the medians of two or more groups differ and it also can calculate the difference in the range of values between medians. We calculate the overall median that is used to define how many values are above and how many are below the median.

Table

Description automatically generated *Figure 11. Mood’s Median Test*

In **Appendix 1**. we state our null and alternate hypotheses, define rejection criteria, and calculate the Test Statistics value. Calculation shows that we can support our null hypothesis and conclude that the Medians are the same for the differences in monthly price changes for USA and IRL data.

* + - 1. Kruskal-Wallis Test (Appendix 2.)

The Kruskal-Wallis H test (KWt) is a nonparametric statistical procedure frequently used to compare several populations (Vargha, Delaney, and Vargha, 1998). In this test we will observe are the steak prices means equal for samples in the last 3 observed years in the USA.   
First, we will filter the USA dataset and keep only values for the last three years, sort them by price and add ranks in the way that that highest value gets ranked 1. Then, we will sum ranks for every year.

Table

Description automatically generated *Figure 12. Kruskal-Wallis Test*

In Appendix 2. we state our hypotheses, calculate the Test Statistics H value and conclude that we canreject ourNull Hypothesis (H0) and we can say that medians are not the same in 3 selected years and at least one of them is different.

* + - 1. Mann-Whitney U Test (Appendix 3.)

Nachar (2008) states that the Mann‐Whitney U test can be used to answer the concerning the difference between his groups. The test can be also used when we are dealing when measured variables are of an ordinal type and not very precise. In our test, we will test the Null Hypothesis that the mean prices are greater in the USA than in Ireland. First, we merge USA and Ireland datasets and add a column with a country code for each observation. Next, we add ranks to all observations where the smallest value has rank 1.

Table

Description automatically generated  
*Figure 13. Mann- Whitney test*

Then we perform the Shapiro-Wilk test and plot a histogram of the data distribution to confirm that data is not normally distributed and that the Mann-Whitney U test is appropriate. In Appendix 3. We state our hypothesis, calculate the Z value, and conclude that as Z = -0.766 and it is greater than -1.96 and is less than 1.96. We can accept Null Hypothesis H0.

* + - 1. Spearman’s r Test

The Spearman rank correlation coefficient evaluates the correlation between two independent variables (Sedgwick, 2014). The test returns values from -1 to 1 where +1 is a perfect positive correlation and -1 is a perfect negative correlation. We will calculate the correlation between mean prices in the USA and Irish datasets. Using Python, we calculate that coefficient id 0.792 meaning that there is a strong correlation between the means of the two datasets. (Calculation is also done manually in Appendix 4.

##### Machine learning

We will use three machine learning models to predict steak prices in the future. We will use ARIMA, SARIMAX, and LSTM models.

##### ARIMA

ARIMA or Autoregressive Integrated Moving Average is a machine learning model that forecasts a value in a response Time Series which is a linear combination of its related past values, past Errors, and current and past values of alternative Time Series (Jain and Mallick, 2017).   
To start building the model, we use the seasonal decompose function to check for trend and seasonality in the Irish dataset. The trend is the direction of the time series. Seasonality is a periodic behavior.

Timeline

Description automatically generated with medium confidence  
 *Figure 15. Decomposed data*

We can see that we have a growing trend and frequent seasonality with a high residue.   
Using the Augmented Dickey-Fuller test we can determine if our data is stationary by calculating the p-value. If the p-value is less than 0.05, we can say that our data does not have a unit root and is stationary – it doesn’t have time dependency.

Text, letter

Description automatically generated  
*Figure 16. adfuller test*

The p-value is 0.2, is greater than 0.05, and is not stationary. To transform to stationary data, we can use a rolling mean feature that will calculate the average for a window of the data.   
To find out which ARIMA model is best fitted we grid search ARIMA parameters by using a function that will loop through possible combinations of p, s, and q values by calculating the Mean Square Error. Function showed that the best ARIMA model is (8,0,2) with train/test split 90/10 where Mean Square Error is 0.324. Running this model, we get pretty accurate predictions.

##### SARIMAX

SARIMAX is an extension to ARIMA, and it is used on data that have seasonal cycles. We will use SARIMAX in combination with Auto ARIMA. Auto ARIMA can automatically select which model is best to use. The best-suggested model for IRL data is SARIMAX (2,1,3). When we run and plot the model, we can see that the predicted models fit with the actual results.

##### LSTM

LSTM – Long Short-Term Memory is a recurring neural network model that is fit to use with the time series. In this case, we couldn’t find the correct fit that would create accurate predictions. The prediction was always shown as a flat line.

###### Results Ireland

ARIMA Model:

Table

Description automatically generated  
*Figure 17. ARIMA model results*

We can see that AIC is 448, this is an acceptable value because AIC estimates the relative amount of information lost by a given model: the less information a model loses, the higher the quality of that model.  
And BIC value is 462, this is also a good signal of our model, because similarly to the AIC, among various alternative models, the model to be preferred is the one with the minimum BIC value.   
HQ is an alternative to the Akaike information criterion (AIC) and Bayesian information criterion (BIC). It’s 454 in the ARIMA model. It’s approximating AIC and BIC.

Chart, line chart

Description automatically generated  
*Figure 18. ARIMA model IRL plot*

We can see the predicted values fit with the true values, the mean loss of the ARIMA model is very low and we can use the ARIMA model to predict the future’s values.

SARIMAX model

Graphical user interface, table

Description automatically generated with medium confidence  
*Figure 19. SARIMAS model IRL results*

The AIC is 426, it’s lower than the AIC of the ARIMA model, also BIC and HQIC lower than ARIMA’s values so this model is better than ARIMA.

Chart, line chart

Description automatically generated  
 *Figure 20. SARIMAX model IRL plot*

The predicted value of the SARIMAX model is more accurate than the ARIMA model.

Chart, line chart

Description automatically generated  
*Figure 21. SARIMAX model prediction*

LSTM model: Train loss and validation loss:  
 

Chart, line chart

Description automatically generated  
 *Figure 22. LSTM model plot*

The LSTM model predicts results like a linear straight, LSTM is not suitable for our dataset.

Compare ARIMA vs SARIMA:

|  |  |  |
| --- | --- | --- |
|  | ARIMA | SARIMA |
| AIC | 448 | 426 |
| BIC | 462 | 456 |
| HQIC | 454 | 438 |

###### Results USA

ARIMA model

Table

Description automatically generated  
*Figure 23. ARIMA model results in USA*

We can see AIC is 406, and the BIC value is 432, this is also a good signal of our model, because similarly to the AIC, among various alternative models, the model to be preferred is the one with the minimum BIC value.   
HQIC is 417 in ARIMA model. It’s approximate AIC and BIC and lower than the ARIMA model for the IRL dataset. It means this model is more suitable for this dataset than the IRL dataset.

SARIMAX model

Graphical user interface, table

Description automatically generated  
*Figure 24. SARIMAX model results USA*

The AIC, BIC, and HQIC are higher than the ARIMA model, we can see in the IRL dataset SARIMAX model is better than the ARIMA model but in this dataset, the ARIMA model is better than SARIMAX.

**LSTM**



Chart, line chart, scatter chart

Description automatically generated  
 *Figure 25. LSTM model USA*

The LSTM is still not good for this dataset.

## Sentiment Analysis

This analysis aims to find out what Twitter users think about beef products. We will check sentiment from both, the consumer’s, and producer’s sides. To get twitter data, we will use *snscrape.modules.twitter* scraper (social network services scraper). The original plan was to use the official Twitter API, but complications with registration, elevated access, and restrictions on data gathering, turned us to find a different data gathering method.   
\*As at this stage we reached a 3300-word limit, we will focus here only on sentiment analysis results, while the steps will be commented in the Jupyter notebook.

**Producer’s sentiment analysis:**Scraper returned 1342 tweets related to the query: ‘beef production cost’. When we processed data using Natural Language Toolkit, removed stop words, removed special characters, and used *TextBlob* to find sentiment for each tweet, results showed that 675 tweets were positive, and 305 tweets were negative (362 neutral values).

Chart, bar chart

Description automatically generated  
*Figure 26. Sentiment Analysis Production*

**Consumers sentiment analysis:**Scraper returned max (5000) tweets related to query: ‘steak price’. Following the same process as in the first analysis, results showed that 2682 tweets were positive, and 1265 tweets were negative (1062 neutral values).

Chart, bar chart

Description automatically generated*Figure 26. Sentiment Analysis Production*

# Conclusion

##### Compare datasets

Plotting line plots with lines from both datasets we can visually compare how prices were changing over the years.

Chart, line chart

Description automatically generated  
 *Figure 14. Comparing prices from USA and IRL in line plot*

We can see that prices in both countries had similar trends up to 2014 from when the prices in the USA are increasing.

Chart, line chart

Description automatically generated  
 *Figure 14. Comparing monthly price differences*

Correlation matrixes between price differences and monthly prices show us that there is a 75% correlation between UA and Irish prices, but no correlation between the price differences.

Chart, application

Description automatically generated with medium confidence  
*Figure 15. Correlation Matrix* *of the price differences*

Chart

Description automatically generated  
 *Figure 14. Correlation Matrix USA/IRL*

This tells us that although price changes are corralled, monthly price changes in the USA and in the Ireland were following different patterns.   
Our initial Null Hypothesis can be upheld as we can see that prices in both countries are correlated.  
Price increases in Ireland in 2002 and in the USA in the 2014 were probably result of some unusual effect. Quick internet search shows that Irish beef market was impacted by Foot and Mouth Outbreak disease. In the USA in 2014 extreme hot weather and draught caused price spike in agricultural sector. Future research would involve getting more relevant data (yearly beef production, beef export/import, currency prices, etc.) to build more complex model that can make better prediction.

# References

Anthony, F. (2015). *Mastering Pandas*. [online] Birmingham, UK: Packt Publishing. Available at: <https://eds.p.ebscohost.com/eds/ebookviewer/ebook/bmxlYmtfXzEwMTc1ODRfX0FO0?sid=e0802fec-11f1-4eeb-bda5-b450a3609006@redis&vid=4&format=EB&rid=2>.

Asadoorian, M.O. and Kantarelis, D. (2009). *Essentials of inferential statistics*. [online] Lanham: University Press Of America. Available at: <https://books.google.ie/books?hl=en&lr=&id=MUyrEx2ATwkC&oi=fnd&pg=PA1&dq=inferential+statistics&ots=06mdo73B2B&sig=aH2BdoiujPnnnYmWS0I3gJbNOq4&redir_esc=y#v=onepage&q=inferential%20statistics&f=false> [Accessed 10 May 2022].

Bazdaric, K., Sverko, D., Salaric, I., Martinovic, A. and Lucijanic, M. (2021). The ABC of linear regression analysis: What every author and editor should know. *European Science Editing*, [online] 47. doi:[10.3897/ese.2021.e63780](file:///C:\Users\Z\Downloads\10.3897\ese.2021.e63780).

Cockburn, M. (2020). Review: Application and Prospective Discussion of Machine Learning for the Management of Dairy Farms. *Animals*, [online] 10(9), p.1690. doi:[10.3390/ani10091690](file:///C:\Users\Z\Downloads\10.3390\ani10091690).

European Commission - European Commission. (2021). *Sustainable agriculture in the CAP*. [online] Available at: <https://ec.europa.eu/info/food-farming-fisheries/sustainability/sustainable-cap_en> [Accessed 24 Apr. 2022].

<!--[if !supportLists]-->1. <!--[endif]-->Kirk, Andy (2012) Data Visualization: a Successful DesignProcess?: This Book Will Benefit Anyone Who Wants to Discover Effective, Attractive Ways to Visually Analyze and Communicate Data. With No SpecialKnowledge Required, It’s an Inspirational Guide That Teaches Through Examplesand Illustration. Birmingham: Packt Publishing (Community ExperienceDistilled). Available at: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=526173&site=eds-live(Accessed: 3 May 2022).

Jain, G. and Mallick, B. (2017). *A Study of Time Series Models ARIMA and ETS*. [online] papers.ssrn.com. Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2898968> [Accessed 21 May 2022].

Kirchgässner, G., Wolters, J. and Hassler, U., 2012. *Introduction to modern time series analysis*. Springer Science & Business Media.

Komorowski, M., Marshall, D.C., Salciccioli, J.D. and Crutain, Y. (2016). Exploratory Data Analysis. *Secondary Analysis of Electronic Health Records*, [online] pp.185–203. doi:[10.1007/978-3-319-43742-2\_15](file:///C:\Users\Z\Downloads\10.1007\978-3-319-43742-2_15).

Matt Dancho (2017). *Tidy Time Series Analysis, Part 4: Lags and Autocorrelation*. [online] Business Science. Available at: <https://www.business-science.io/timeseries-analysis/2017/08/30/tidy-timeseries-analysis-pt-4.html> [Accessed 18 May 2022].

Mozaffari, A., Chenouri, S., Toyserkani, E. and Ali, U. (2021). Functional Boxplots for Outlier Detection in Additive Manufacturing. *arXiv:2110.10867 [stat]*. [online] Available at: https://arxiv.org/abs/2110.10867.

Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, [online] 4(1), pp.13–20. doi:[10.20982/tqmp.04.1.p013](file:///C:\Users\Z\Downloads\10.20982\tqmp.04.1.p013).

Osinga, S.A., Paudel, D., Mouzakitis, S.A. and Athanasiadis, I.N. (2022). Big data in agriculture: Between opportunity and solution. *Agricultural Systems*, [online] 195, p.103298. doi:[10.1016/j.agsy.2021.103298](file:///C:\Users\Z\Downloads\10.1016\j.agsy.2021.103298).

Sedgwick, P. (2014). Spearman’s rank correlation coefficient. *BMJ*, [online] 349(dec05 1), pp.g7528–g7528. doi:[10.1136/bmj.g7528](file:///C:\Users\Z\Downloads\10.1136\bmj.g7528).

Vargha, A., Delaney, H.D. and Vargha, A. (1998). The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, [online] 23(2), p.170. doi:[10.2307/1165320](file:///C:\Users\Z\Downloads\10.2307\1165320).

# Appendices

### **Appendix** Mood’s Median Test

Null Hypothesis H0: The medians of the populations all are equal

Alternative Hypothesis H1: The medians of the population are not all equal

Known values:

Overall Median =0.02829999999999977 (Jupyter notebook )  
Critical value of χ2 (0.05, 1) = 3.841

Observed values (Jupyter notebook)

|  |  |  |  |
| --- | --- | --- | --- |
| Observed | USA | IRL | Totals |
| > median | 151 | 148 | 299 |
| <= median | 148 | 151 | 299 |
| Totals | 299 | 299 | 598 |

Expected values = (Column total \* Row Total) / N

|  |  |  |  |
| --- | --- | --- | --- |
| Observed | USA | IRL | Totals |
| > median | 149.5 | 149.5 | 299 |
| <= median | 149.5 | 149.5 | 299 |
| Totals | 299 | 299 |  |

(Column total \* Row Total) / N

All Values = 299\*299/588=149.

Critical value of χ2 (0.05, 1) = **3.841**  
Calculated **=0.194**

Since 0.194 is less than 3.841 , we accept the Null Hypothesis. We can conclude that the Medians are the same for the differences in monthly price changes for USA and IRL data.

### **Appendix 2.** Kruskal-Wallis Test

Null Hypothesis H0: the medians (mean on ranks) are equal across the samples

Alternative Hypothesis H1: at least one median is different

Known values:

Critical value of χ2 (0.05, 2) = 5.99  
N=36  
k=3  
Ri=97,211,358  
ni=12,12,12

Observed values (Jupyter notebook)

|  |  |  |  |
| --- | --- | --- | --- |
| Observed | 2019 | 2020 | 2021 |
| Sum of Ranks | 97 | 211 | 358 |

Calculate test statistics H:

H=12N(N+1)∑ki=1R2ini−3(N+1)H=12N(N+1)∑i=1kRi2ni−3(N+1)=

H=\*[ +] – 3(18-1)=

\*[ +] – 3(18-1)=

\*[ + – 51)=

0.009\*15123.499=**136.111**

As **H=136.111** is greater than **χ2(0.05,2)=5.99** we will **reject** ourNull Hypothesis (H0) and we can say that medians are **not the same** in 3 selected years and at least one of them is different.

### Appendix 3. *Mann-Whitney U test*

Null Hypothesis H0: USA mean prices are greater in the USA.  
Alternative Hypothesis H1: The mean prices are the same or greater in IRL  
Known values:α = 0.05  
Results: If z is less than -1.96, or greater than 1.96, reject the null hypothesis.

|  |  |  |
| --- | --- | --- |
| Observed | IRL | USA |
| Sum of Ranks | 598 | 677 |
| Population | 25 | 25 |

U = -∑= - 598= -598 =**352**

U’ = -∑= - 677= -677 =**273**

U = 352 U’ = 273 MIN(U ; U’) = 273

Calculate Z:

Z= = = = -0.766

Results: Z = -0.766 and it is greater than -1.96 and is less than 1.96. We can accept Null Hypothesis H0

### Appendix 4. *Spearman’s r Test*

Use values from the Jupyter notebook to calculate squared difference values



The formula for Spearman’s r is:

*p*= =

### **Appendix X –** Project Timeline

Project Start: Saturday 23rd April:

* Review Project Requirements
* Review individual tasks
* Create a project timeline

Week ending Sunday 1st May:

* Initial literature review
* Twitter API review
* Sentiment analysis task review
* Sentiment analysis task – analysis in Python

Week ending Sunday 8th May:

* Review Agriculture related data sources
* Select suitable data analytics topic from available data
* Clean data

Week ending Sunday 15th May:

* Review Machine Learning models and task
* Create Machine learning models
* Review Statistics tasks
* Calculate Statistics tasks
* Review Literature

Week ending Sunday 22nd May:

* Dashboard
* Complete Report
* Review code and report
* Complete and upload the assignment