# Group ID - MSc in Data Analytics

Author: Conor McKenna

e-mail: [sbs23014@student.cct.ie](mailto:sbs23014@student.cct.ie)

Student ID: sbs23014

Table of Contents

[Group ID - MSc in Data Analytics 1](#_Toc136032380)

[Abstract 3](#_Toc136032381)

[Introduction 3](#_Toc136032382)

[1. Data Preparation & Visualisation Tasks 3](#_Toc136032383)

[Acquiring Raw Data 3](#_Toc136032384)

[Eurostat Database 3](#_Toc136032385)

[Data.Gov.ie 4](#_Toc136032386)

[Exploratory Data Analysis 4](#_Toc136032387)

[Initial look at the Irish dataset 5](#_Toc136032388)

[Data Cleansing and Preparation 6](#_Toc136032389)

[Interactive Dashboard 7](#_Toc136032390)

[2. Statistical Analysis 7](#_Toc136032391)

[Descriptive Statistics 7](#_Toc136032392)

[Coefficient of Variation and Standard Deviation 8](#_Toc136032393)

[Check for normal distribution 9](#_Toc136032394)

[Analysis of Variables 10](#_Toc136032395)

[Inferential statistical techniques for comparison 11](#_Toc136032396)

[Anova Test 12](#_Toc136032397)

[T-Test 12](#_Toc136032398)

[Wilcoxon’s Test 13](#_Toc136032399)

[Pearson’s Correlation Coefficient (PCC) Test 13](#_Toc136032400)

[Outcome to deepen research 14](#_Toc136032401)

[3. Machine Learning Tasks 14](#_Toc136032402)

[Rationality of choice of machine learning models 14](#_Toc136032403)

[Testing models that were developed 15](#_Toc136032404)

[Webscraping and sentiment analysis 17](#_Toc136032405)

[Sentiment analysis 17](#_Toc136032406)

[4. Programming for DA Tasks [0-100] 18](#_Toc136032407)

[Data structures: 18](#_Toc136032408)

[Testing & Optimisation: 18](#_Toc136032409)

[Data manipulation: 18](#_Toc136032410)

[Appendix 19](#_Toc136032411)

[References 22](#_Toc136032412)

# Abstract

The construction industry has been through much turmoil since records began. There are many datasets which reflect this including material cost indexes, production indexes, and labour cost datasets. This project takes a closer look at a dataset relating to construction confidence/sentiment and examines whether other key indicators can be used to forecast the confidence in the construction industry.

# Introduction

This project primarily uses a confidence indicator dataset sourced from Eurostat which was then supplemented with 4 separate datasets. The confidence indicator for Ireland is compared against the UK’s confidence indicator because they are our nearest neighbour, and also Denmark’s confidence indicator since they are a comparable country to Ireland insofar as it has a similar population and size, it is a developed country like Ireland and is also mostly bordered by the sea.

# Data Preparation & Visualisation Tasks

## Acquiring Raw Data

### Eurostat Database

Eurostat is an online resource which offers datasets and insights into EU countries. 4 of the 5 datasets were sourced from Eurostat.

The first step in the process was to identify the datasets that would be useful for the project. This required utilising the website’s GUI engineto search for construction datasets and downloading them directly as zipped CSV files. Eurostat offers a user-friendly interface which allows the user to tailer the dataset as appropriate. For example, since this project focuses mainly on Ireland, United Kingdom, and Denmark, those specific countries were selected prior to downloading the CSV, which reduced the workload post-download.

The website offers both a CSV and TSV (Tab separated values) options, and the reason why CSV was favoured above TSV is because the TSV files had the years and data stored horizontally as features, whereas the CSV had the data stored vertically as individual observations which is preferable when performing time series or forecasting.

Eurostat's data is generally available under the open data policy, allowing for free usage, sharing, and redistribution.

The reason why Eurostat was investigated as an initial dataset source is because it offers access to comprehensive and harmonised EU-wide statistics that facilitates cross-country comparisons, trend analysis, and evidence-based decision-making.

### Data.Gov.ie

The National House Construction Cost Index was sourced from the Data.Gov.ie website. The reason why this website was utilised was because, while Eurostat is an impressive and useful repository of datasets for EU countries, it was found that Ireland’s contribution was lacking in many areas. Many other EU countries had contributed to would-be useful datasets, however Ireland’s observations were either entirely null, or Ireland was missing from the listing altogether.

Examples of such datasets include:

* Production in Construction
* Construction – Monthly growth rates
* Labour Input – monthly rates
* Building Permits – monthly data
* Monthly labour costs

### Reddit API Datascraping

Reddit is a popular online platform with a wealth of user-generated content. It provides an API that allows programmatic access to its data. By using the API, it’s possible to retrieve data in a structured and authorised way, ensuring compliance with Reddit's terms of service. To access the Reddit API, a client ID and secret was needed. A .getenv file was created and saved in the local computer folder. The reason why a .getenv file was used is because vital credentials such as username and login, and App\_Secret are required and it’s best practice to ensure these credentials remain hidden.

Jupyter Notebook (jn) 1.1

## Exploratory Data Analysis

The 4 datasets from Eurostat and 1 dataset from Data.gov.ie were imported as dataframes using the Pandas library. A .head(5) was performed on each dataset to get an initial look at the column names and content. A .shape was performed to understand the general structure of the datasets. The .head() and .shape outputs showed that the 4 Eurostat datasets had a similar structure (9-11 columns) with observation counts ranging from just 21 observations to 1,455 observations.

The Construction Cost dataset from the Data.gov.ie website is very different to the other datasets, in that the months and years are stored horizontally as opposed to vertically.

*JN 1.3*

In order to ensure the datasets are more uniform, the Construction Cost dataset was ‘melted’ so that the time data was stored as observations vertically rather than features horizontally.

*JN 1.4*

Before bringing the datasets together, it was prudent to rename the fields to avoid confusion after the concatenation.

*JN 1.5*

It was seen that each dataset had a two digit iso country code and a year and month “Time\_Period” in the same format which allowed them to be merged together as one (per country). The overarching goal is to compare the countries against one another, so it was decided to bring the data together at a country level early, rather than a subject level, and to perform the EDA on each country level dataset.

Once the datasets were merged, subsets for each country were created. Since forecasting and machine learning will be applied against the Irish dataset, it will be the main focus during the EDA phase.

### Initial look at the Irish dataset

*JN 1.7*

Doing a .shape of the dataset showed that there are 531 rows with 7 columns. Normally a .unique function would be run to identify the number of unique values within each column. However this is not required since all columns contain continuous data.

A .info() was performed to confirm the datatypes within the dataframe. This is also important to get an initial view in terms of null values within each column. There are other functions to help search for null value counts, such as *.dtypes* or *.isna()* but *.info* was used because it provides faster insight into all columns at once.

*JN 1.8*

For example, “Employment\_Expectation\_Score” has many null values in comparison to the other columns. It was seen that the continuous features are all float64 and the reason why this is an important detail is because it suggests that there is no rogue *character* data in these numeric columns.

The .duplicated() function was used to identify any duplicate values in the dataset and none were found. This is expected since there should be one single observation per month and year in this dataset.

A .head(15) showed the first 15 records in the table appeared to be very clean with no irregularities, but the .tail(15) shows some irregularities in the Time\_Period feature. During the data cleaning and preparation phase those values which don’t comply to the YYYY-MM date range will be removed since they carry little or no value (all the features corresponding to these malformities contain Null values).

Column Employment\_Expectation\_Score might be removed from the dataset since it contains just 9 non-null observations.

## Data Cleansing and Preparation

A melt operation was carried out earlier in the process to allow for Exploratory Data Analysis to be performed on the reformatted table.

*JN 1.9*

As seen during EDA, there are badly formatted Time\_Period entries which were removed. There are numerous ways to carry out this operation, including the building of regex, however it is less complex to build a For, Try and If statement that loops through and removes the badly formatted records. The reason why the removed records were also stored in another dataframe called “removed\_df” is so that they could be reviewed afterwards to ensure the intended outcome was reached.

It could have been prudent to delete “removed\_df” from memory once the records within it were verified, but it was kept since it is a very small size and should not impact on processing power.

*JN 1.10*

It was seen during EDA that column Producer\_Price\_Percent\_change had a large number of nulls which suggested that the data didn’t go as far back as the other columns such as Constr\_Confidence\_Value or Industry\_Prod\_index. A query was run to find the first non-null record in the table.

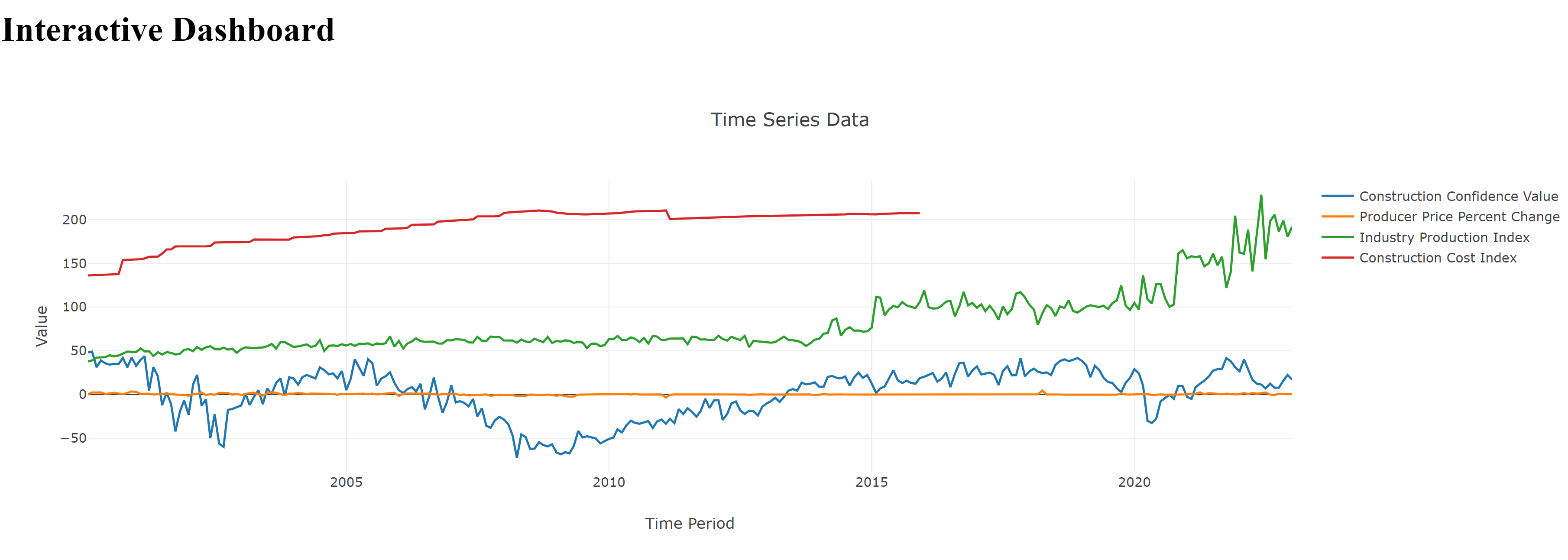
*JN 1.11*

Since the project centres around forecasting and time\_series, the table was backed up before any records were removed, so that the observations for the other features could be accessed later during the time\_series analysis.

## Interactive Dashboard

*JN 4.1*

A dashboard was built with a timeline of years along the x-axis, and values representing each variable along the y-axis. There are 4 colour coded lines within the dashboard itself which allows the user to see The Construction Confidence value, Producer Price Percent Change, Industry Production Index, and Construction Cost Index all in one place. Colour red was chosen as the most appropriate colour to represent cost, whereas colour green was chosen to represent a positive value “Industry Production Index”.



The yellow line representing Producer Price Percent change was barely visible in comparison to others so the user was given the ability to zoom in by highlighting a section with their mouse they want to view more closely.

The dashboard was built with Tuft’s principles in mind. Firstly the dashboard shows all the data, and it illustrates it in a clear and concise manner. The viewer is encouraged to see not only how each variable singularly performs over time, but how they perform in comparison to each other. The fact the viewer can zoom in on specific ranges helps to “reveal the data at several levels of detail”. The dashboard is designed to very quickly reveal all the data within the dataset in a quick and coherent way.

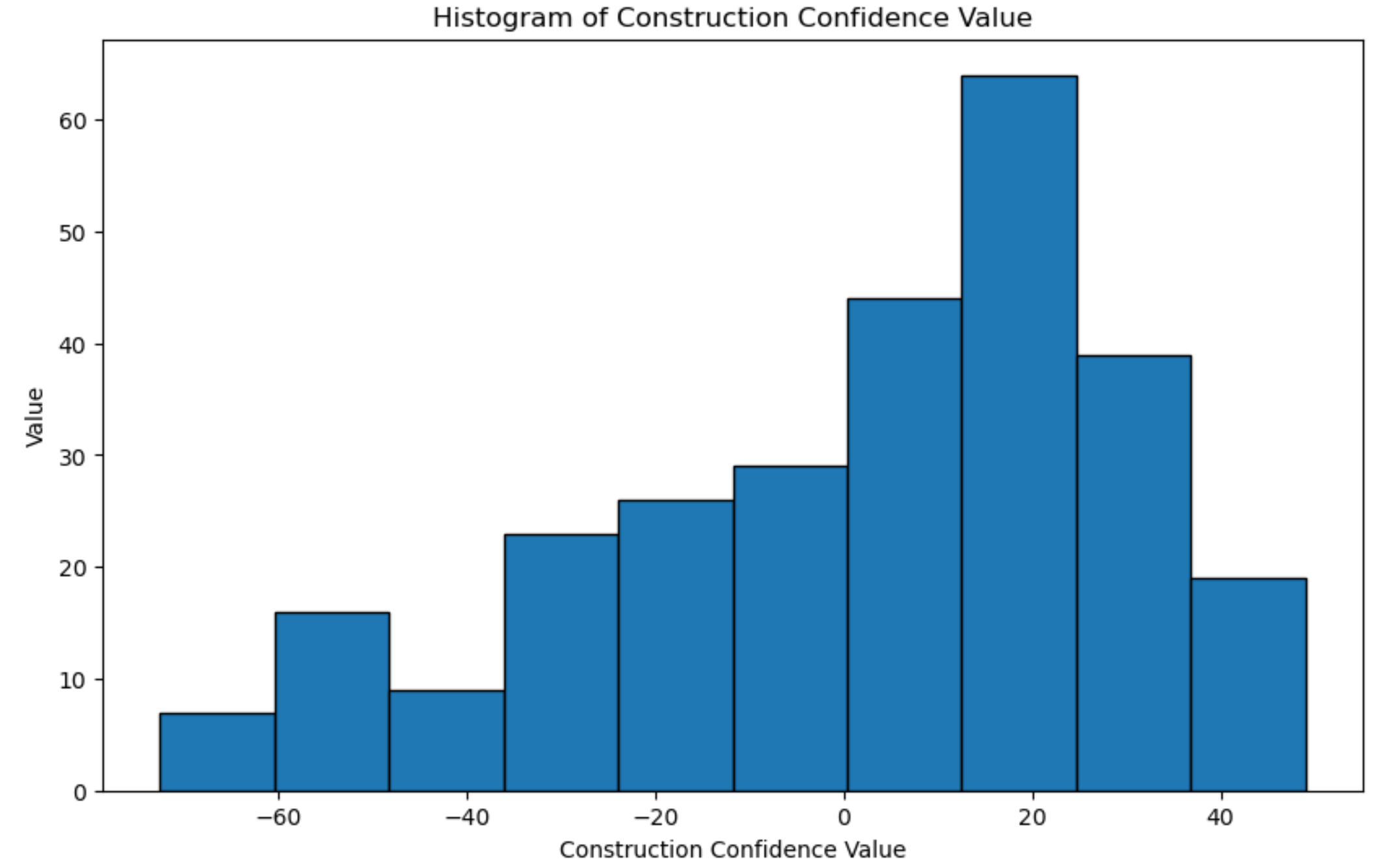
There were a small number of null values within feature 'Constr\_Confidence\_Value' that were contained within each dataset. These were analysed and removed when it was determined that they were likely due to a delay from the given country in sending the data, or due to Brexit and the UK no longer needing to share its data to Eurostat.

# Statistical Analysis

## Descriptive Statistics

*JN 2.1*

A histogram of the dependent variable was created, showing the data within it is slightly negatively skewed.



### Coefficient of Variation and Standard Deviation

The coefficient of variation is derived by dividing the standard deviation (std) by the mean, and ideally the resulting outcome would be less than 1 (Sharma, 2007).

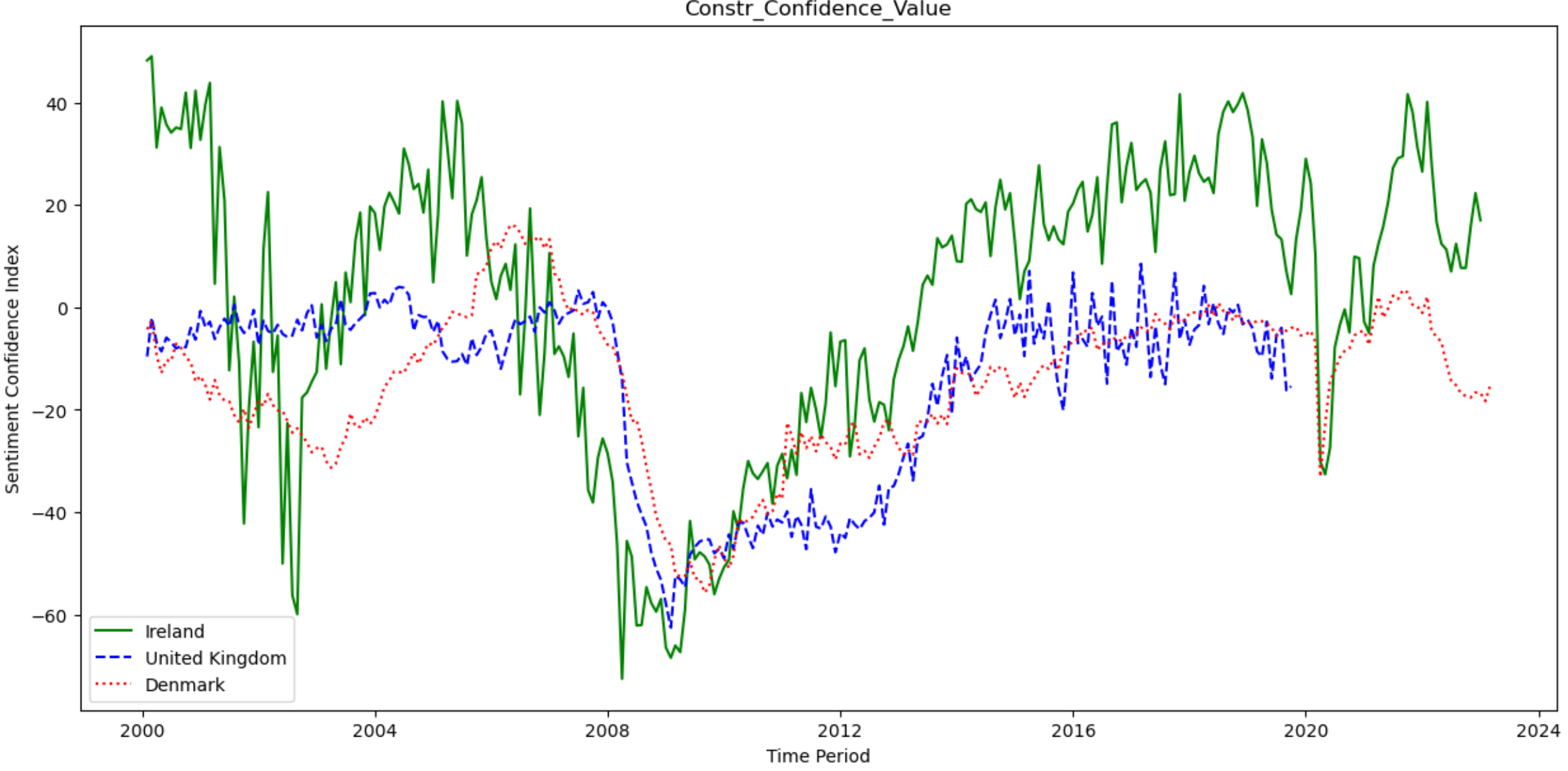
*JN 2.2*

It can be seen that the std for Constr\_Confidence\_Value is a lot higher than the mean which implies great variation within that feature. By stark comparison, the Constr\_cost\_index has a mean which is far greater than the std implying there is a lot less variance within the data. Some data points have narrow ranges such as the Producer\_Price\_Percent\_change, where as others have wide ranges such as Industry\_Prod\_index. The reason why these are important measurements to analyse, is because they give an insight into which columns have a great degree of volatility, and which columns are more stable in terms of data changes over time.

For example, the Constr\_Confidence\_Value column would appear to be quite volatile and potential difficult to work with in terms of regression analysis and forecasting.

*JN 2.3*

When a .describe is performed on the data that came from other countries (Denmark and the UK), a stark difference was immediately noticed. The sentiment average in both countries is a lot lower. The score very rarely goes above 0 and is a lot less volatile.



The volatility of Ireland’s score is clearly seen in green in the graph above.

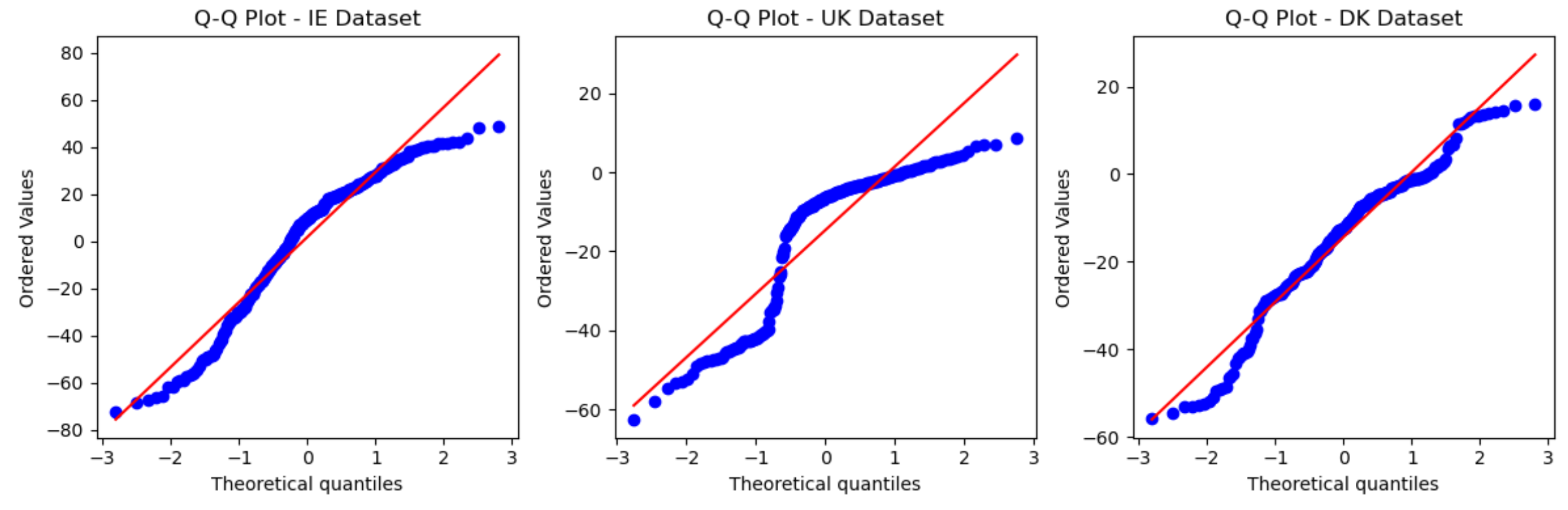
### Check for normal distribution

*JN 2.4*

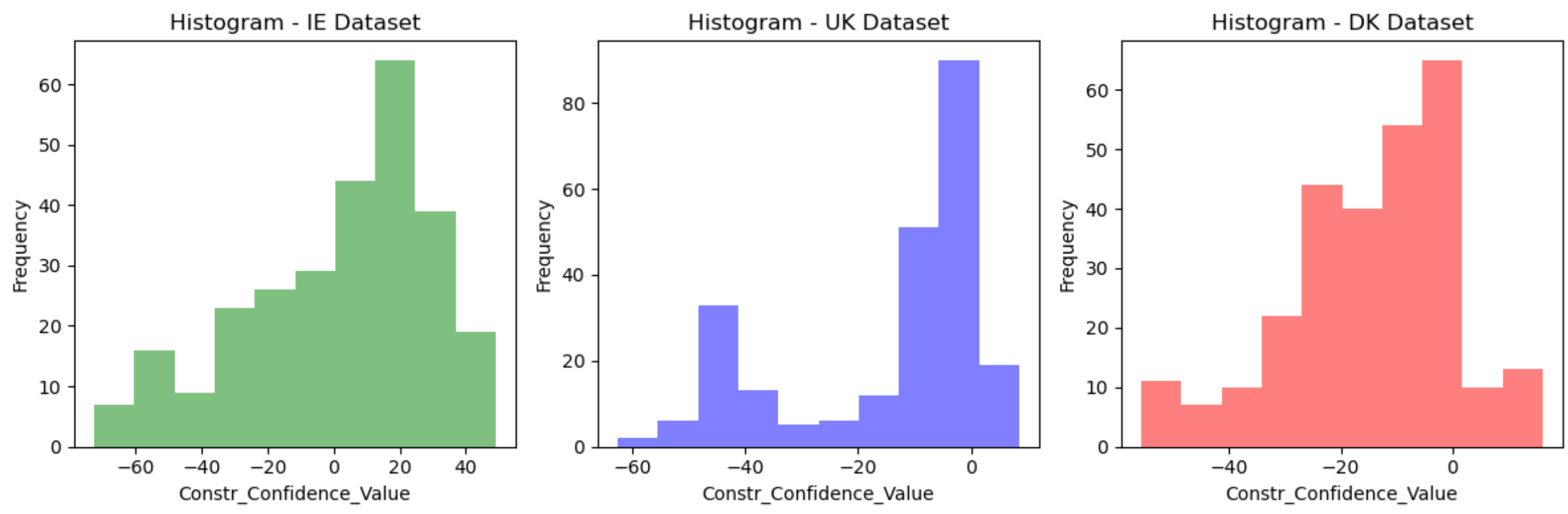
A Shapiro-Wilk Test was performed to test the datasets for normal distribution. The null-hypothesis of this test is that a given population is normally distributed. The p-value for all countries was close to zero which is less than the given alpha level of 0.05. This implies that the data does not follow a strict normal distribution. However it is prudent to visualise the data to check the degree to which the data is not “normally distributed”.

*JN 2.5*

A Q-Q plot helped to visualise the data in this way – here the quartiles of each dataset are plotted along a theoretical line of normal distribution. The plots show that Ireland and DK do tend to follow the line quite well, with the UK straying the furthest.



Another way to perform a visual check for a normal distribution is the histogram:



It can be seen in the plots above that again while no dataset follows a perfectly normal distribution, Ireland and Denmark exhibit a normal tendency, whereas the UK follows what is sometimes referred to as a Bimodal distribution (Weiss et al., 2017).

The outcome of the above analysis around distribution helped to select which statistical tests were most appropriate. For example, T-tests are better suited for normally distributed datasets, whereas the Wilcoxon test doesn’t rely as much on the dataset being strictly “normal” in its distribution.

## Analysis of Variables

*JN 2.6*

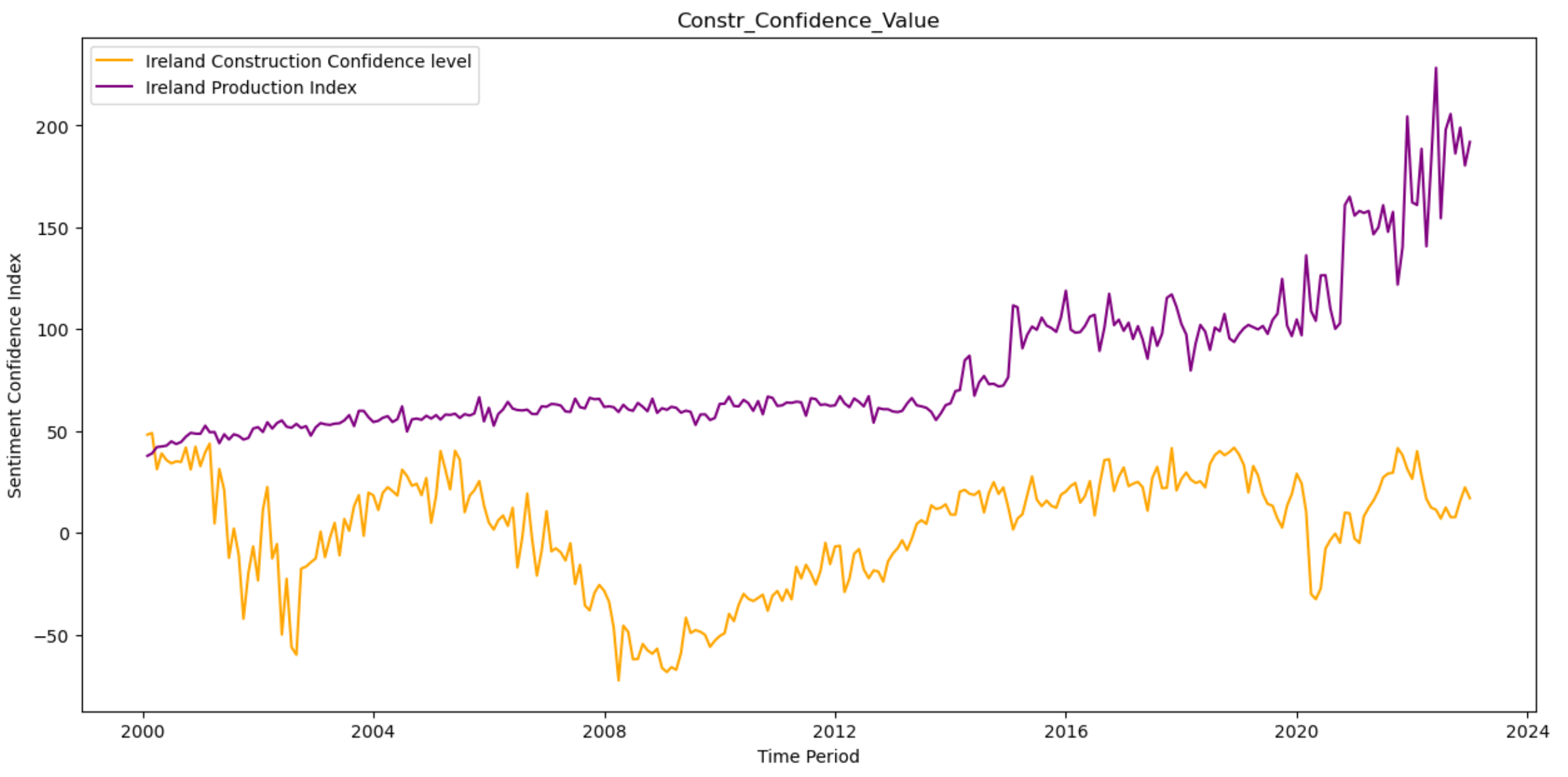
To analyse Ireland’s confidence interval for the future based on the sample population available, the first step was to choose a confidence level. In this case 95% was chosen as it is the leading confidence level in most business domains. Scipy.stats was imported into the Jupyter notebook which helped obtain the standard error, mean, and margin of error from the dataset. A confidence interval (95%) was determined for Ireland which was between -1.58 and 5.11.

Both the UK’s confidence interval and Denmark’s confidence interval reflected the negative values on the plots, having confidence ranges of -16.89 and -12.36 (UK) and -16.14 and -12.60 (Denmark).

*JN 2.7*

A correlation coefficient was derived using the .corr function within Pandas. A correlation coefficient of 0.31 suggests a moderate positive correlation between the two variables. This means that as the values of "Constr\_Confidence\_Value" increase, there is a tendency for the values of "Industry\_Prod\_index" to also increase, even though the relationship may not be extremely strong.

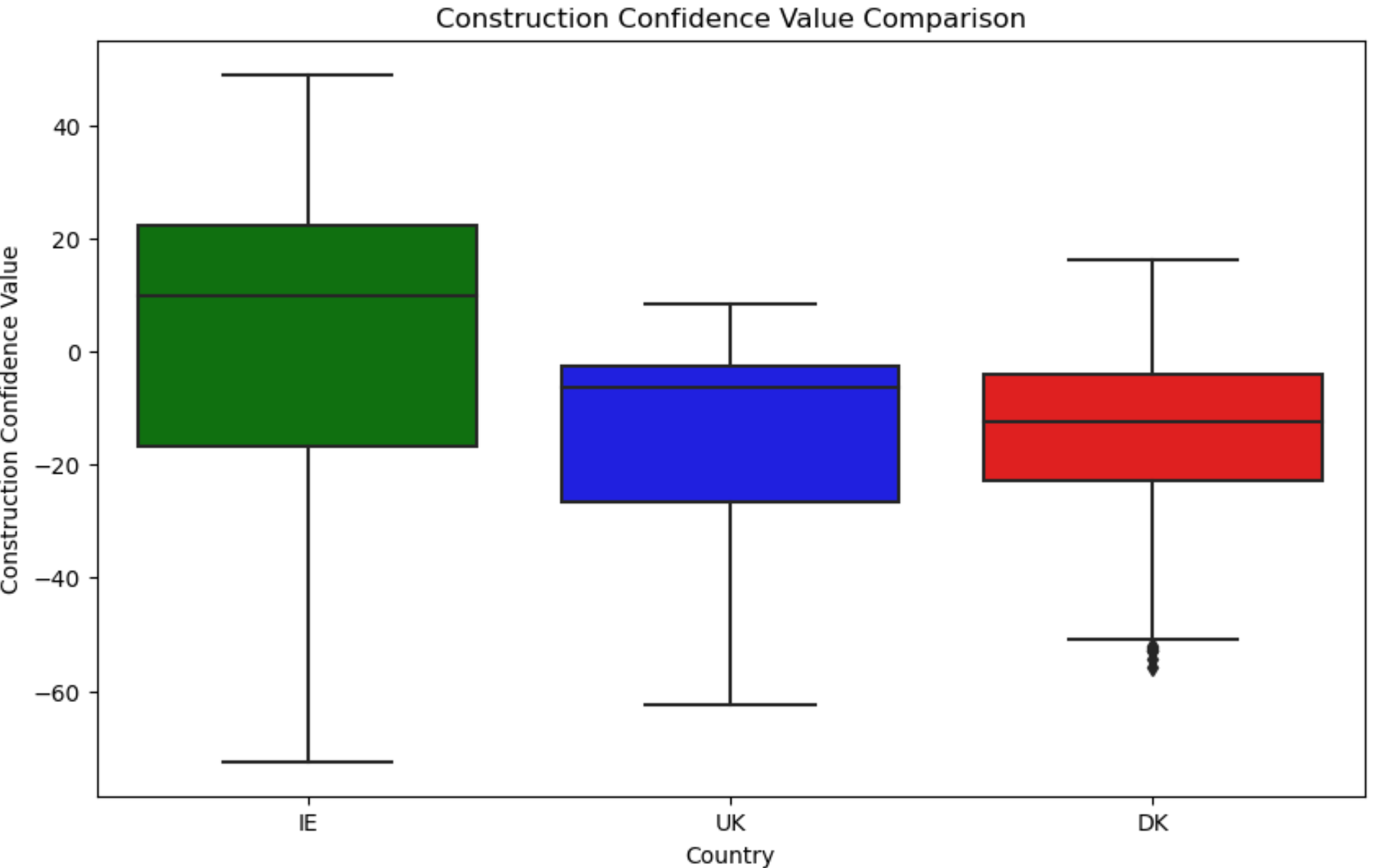
The below graph illustrates this:



## Inferential statistical techniques for comparison

*JN 2.8*

The below boxplot also shows that Ireland’s range is far greater than UK and Denmark, and only Denmark has notable outlier on the negative side of the plot.



### Anova Test

*JN 2.9*

An Anova test was used to check for variance between the 3 datasets. First the null hypothesis was declared which is that there is no significant difference between the means:

H₀: μ₁ = μ₂ = μ3

An F statistic of 52.438 and a p-value close to zero indicate that there is a statistically significant difference among the means of the three datasets and the null hypothesis should be rejected.

### T-Test

*JN 2.10*

To investigate the variance highlighted by the Anova test in more detail, t-tests were performed between each of the three datasets to find out where the major point of variance lies.

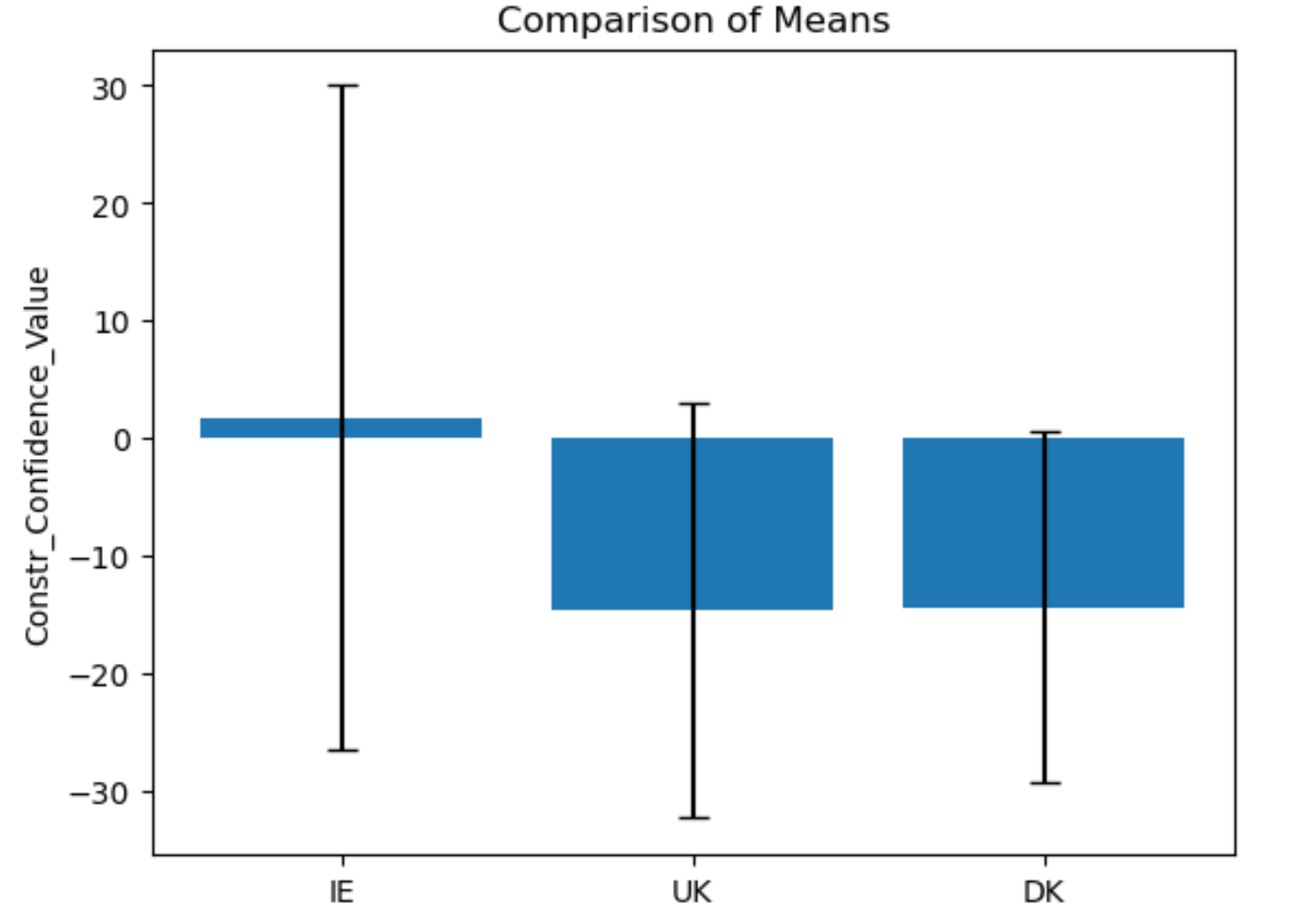
A t-test was performed to check whether there is a statistical difference between the means of Ireland and the UK. The t-statistic of 7.73 suggests a substantial difference between the means – they are almost 8 times as different from each other as they are the same as each other. The p-value of 0.000000000000059 (or 5.9e-14 in scientific notation) indicates an extremely small probability of observing a large difference in the means by chance alone. In this case, since the p-value is significantly lower than the conventional significance level of 0.05, it provides strong evidence to reject the null hypothesis.

A similar test was also performed to check for a statistical difference between Ireland and Denmark and the outcome was the same – The T-statistic of 8.38 suggests a substantial difference between the means and a p-value which is near zero provides strong evidence to reject the null hypothesis.

A final t-test was performed between the UK’s mean and Denmark’s means and a very different outcome was found - the t-stat of -0.187 suggests a small difference between the means. The negative t-stat value suggests that the mean of the UK is slightly smaller than the mean of the Denmark (which can be seen in the .describe results earlier) and in this case, since the p-value (0.85) is greater than the conventional significance level of 0.05, there is insufficient evidence to reject the null hypothesis.

*JN 2.11*

A boxplot of the means between the Ireland, The UK, and Denmark does reflect the above T-Test findings, insofar as there appears to be a significant difference between Ireland and Denmark/UK, but also that Denmark and The UK are quite similar.



### Wilcoxon’s Test

*JN 2.12*

Between the Shapiro-Wilk test and the data visualisation carried out in the first section, it could be argued that both t-test and Wilcoxon tests are appropriate tests to carry out since there is indication that the data is both normally distributed (as per the graphs) but also not perfectly normally distributed (As suggested by the Shapiro-Wilk test)

The Wilcoxon result again show there is a statistical difference between the means of Ireland and the UK and to Denmark. The null hypothesis is the same as the previous tests:

H₀: μ₁ = μ₂

With p-values for Ireland and UK/Denmark being so close to zero, the null hypothesis can be rejected. As with previous tests, there is not enough evidence to reject this null hypothesis between Denmark and the UK.

### Pearson’s Correlation Coefficient (PCC) Test

*JN 2.13*

This test was carried out to further test the correlation between these variables.

The null hypothesis for the PCC is that there is no correlation between the two variables

H₀: μ₁ ≠ μ₂

The difference between this test and the others is that the PCC test results show a very low (near zero) P value between all countries, including Denmark and the UK which indicates strong evidence against the null hypothesis. This is significant since it is the first statistical evidence that the data is correlated in some way.

# Machine Learning Tasks

## Rationality of choice of machine learning models

Within this dataset, both the dependent variable (**Constr\_Confidence\_Value**) and independent variables (**Producer\_Price\_Percent\_change** and **Industry\_Prod\_index**) are all continuous data types. Therefore regression-based machine learning algorithms are more appropriate than classification algorithms.

*JN 3.1*

The reason why 5 algorithms (Linear Regression, Decision Trees, Random Forest, Gradient Boosting and SVR) were used was because quite low R2 scores were produced by each algorithm and a more desirable result was being sought.

SVM are great for small to medium datasets like the one being used for this project. They perform a lot better when StandardScaler() is incorporated (Geron 2017) which is why that technique was applied to the code.

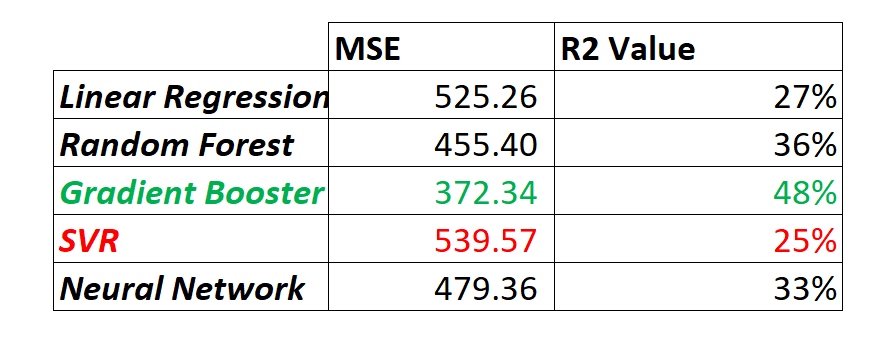
Decision trees also form a major part of random forests which are said to be among the most powerful machine learning algorithms available (Geron 2017).

Since Random Forests build upon Decision Trees (George, 2021) and are a very powerful algorithmic tool, they were incorporated into the project as a follow on to Decision Trees. It could be argued that they are more appropriate for larger and more complex datasets being an “ensemble method” (ie a collection of different models).

In “Neural Networks Projects For Python”, James Loy explains how Neural Networks are highly scalable and flexible, and can be very powerful when used as a regression technique (Loy, 2017).

## Testing models that were developed

In all cases the data was split up into training and testing sets, and an initial 20% of data was set as testing size.



The above table provides the summary results of each of the analyses. All results suggest that the dependent variable Constr\_Confidence\_Value can not be affectively predicted using variables Producer\_Price\_Percent\_change and Industry\_Prod\_index alone. The highest R2 value was achieved by the Gradient Booster, from which we can infer that 48% of the Constr\_Confidence\_Value can be explained by the independent variables. The worst performing algorithm was the SVR analysis with just 25% R2 value in spite of using StandardScaler to improve its performance.

*JN 3.3*

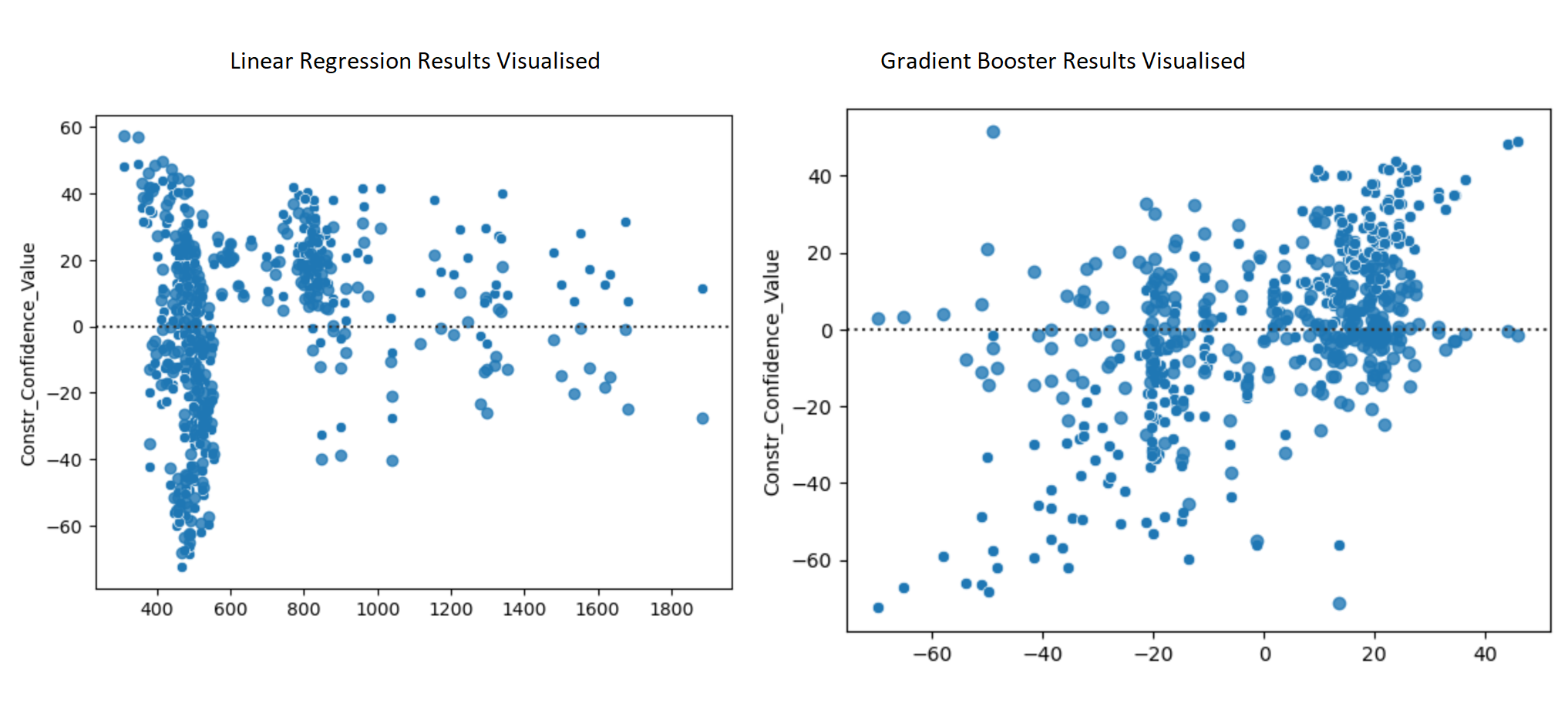
GridsearchCV was applied to the Linear Regression model to find the best hyper parameters but the R2 score fell from 27% to just 14%.

*JN 3.6*

Another hyperparameter tuning technique called RandomizedSearchCV was also applied but this time to the GradientBooster and again the R2 score reduced from 48% to 43%. The reason why these scores should be compared against one another is to check whether the hyperparameters proposed by these tuning techniques are better than the default values that come with the algorithm. In this case the hyperparameter tuning technique’s output does not help the algorithm’s performance and can therefore be brought back out of the analysis.

From the results it can be inferred that further analysis would be required around variables before a good forecasting model could be selected to reliably forecast Constr\_Confidence\_Value

### Graphics and Table that demonstrates the ML outcomes



## Webscraping and sentiment analysis

Python-dotenv was installed and a .env file was saved in the folder along with all the required private credentials for logging into the Reddit API.

An authorisation request was sent to the API and status\_code ‘200’ was printed to confirm it was a success. The access token was assigned to variable TOKEN and the res variable was set to the webpage of results for the Irish reddit page (r/ireland).

The first step in terms of scraping data was to print the titles, and then the id for that post was extrapolated from the Json file. The post ids were saved into a list and referred to when scrapping the comments for each title.

### Sentiment analysis

It was decided to use two different sentiment functions (SentimentIntensityAnalyzer from NTLK and Textblob. The reason why both sentiment functions were used is because both provide a result which is between -1 and +1 so they are easily comparable.

# Programming for DA Tasks [0-100]

## Data structures:

The two datasets that were gathered and processed were CSV (the 5 datasets) and the data accessed for the sentiment analysis was saved in the web API in JSON format.

## Testing & Optimisation:

The testing strategy for this project differed depending on the section. For example within the initial section, as the CSV files were loaded and manipulated into a more favourable format, It was pertinent to constantly perform .head() and .tail() functions to ensure the layout of the dataset was as expected. This approach showed up the badly formatted date data that was addressed in Jupter Notebook section 1.9. It also helped confirm the TIME\_PERIOD column had been sorted correctly in section 1.10.

When badly formatted records were removed by a loop in section 1.9, they were stored in a removed\_Df dataframe so that they could be reviewed afterwards to ensure no ‘good’ data was erroneously taken out of the dataframe.

Since the datasets were quite small, and time was limited, the dataframe was frequently exported as a .csv file and opened within Excel purely to visualise the structure in a more complete sense. The trade off with this is that with larger datasets this would not be possible. In those cases there would be a need to rely more on .info, .unique, and .isna functions to ensure the data remained clean as expected.

During the webscraping exercise, when looping through each of the posts, the post ID was printed as part of the loop to confirm loop was working effectively and hadn’t stalled.

Throughout the statistics sections, the data was visualised in as many ways as possible to ensure it looked as expected.

## Data manipulation:

Pandas was used as the primary library for data manipulation, but another popular library for this is NumPy. While Pandas is built on top of NumPy and provides additional functionalities for working with structured data, NumPy focuses on numerical operations and provides powerful array-based computing capabilities. The main advantage with working with NumPy is that it has efficient array operations – it provides a powerful n-dimensional array object, which allows for efficient manipulation and computation on large datasets (Chin & Dutta, 2016). Pandas is built on top of NumPy though and does provide a lot more data manipulation techniques such as handling missing data and a timeseries functionality (DateTimeIndex and PeriodIndex).

Requests was used to scrape information from Reddit’s API, but BeautifulSoup is another popular library which could have been used. Primarily built to work with HTML and XML it does have functionality to handle JSON files insofar as the JSON file can be parsed and information can be extracted from it. Requests was more appropriate for this project because there was a need to send HTTP requests to a URL and retrieve the JSON response.

# References

Weiss, N.A. and Weiss, C.A. (2017) “Chapter 3” in *Introductory statistics*. Boston etc.: Pearson, pp. 150–150.

www.oreilly.com. (n.d.). 5. Support Vector Machines - Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition [Book]. [online] Available at: https://learning.oreilly.com/library/view/hands-on-machine-learning/9781098125967/ch05.html#:-:text=SVMs%20shine%20with%20small [Accessed 26 May 2023].

George, N. (2021) Practical data science with python, O’Reilly Online Learning. Available at: https://learning.oreilly.com/library/view/practical-data-science/9781801071970/Text/Chapter\_15.xhtml (Accessed: 26 May 2023).

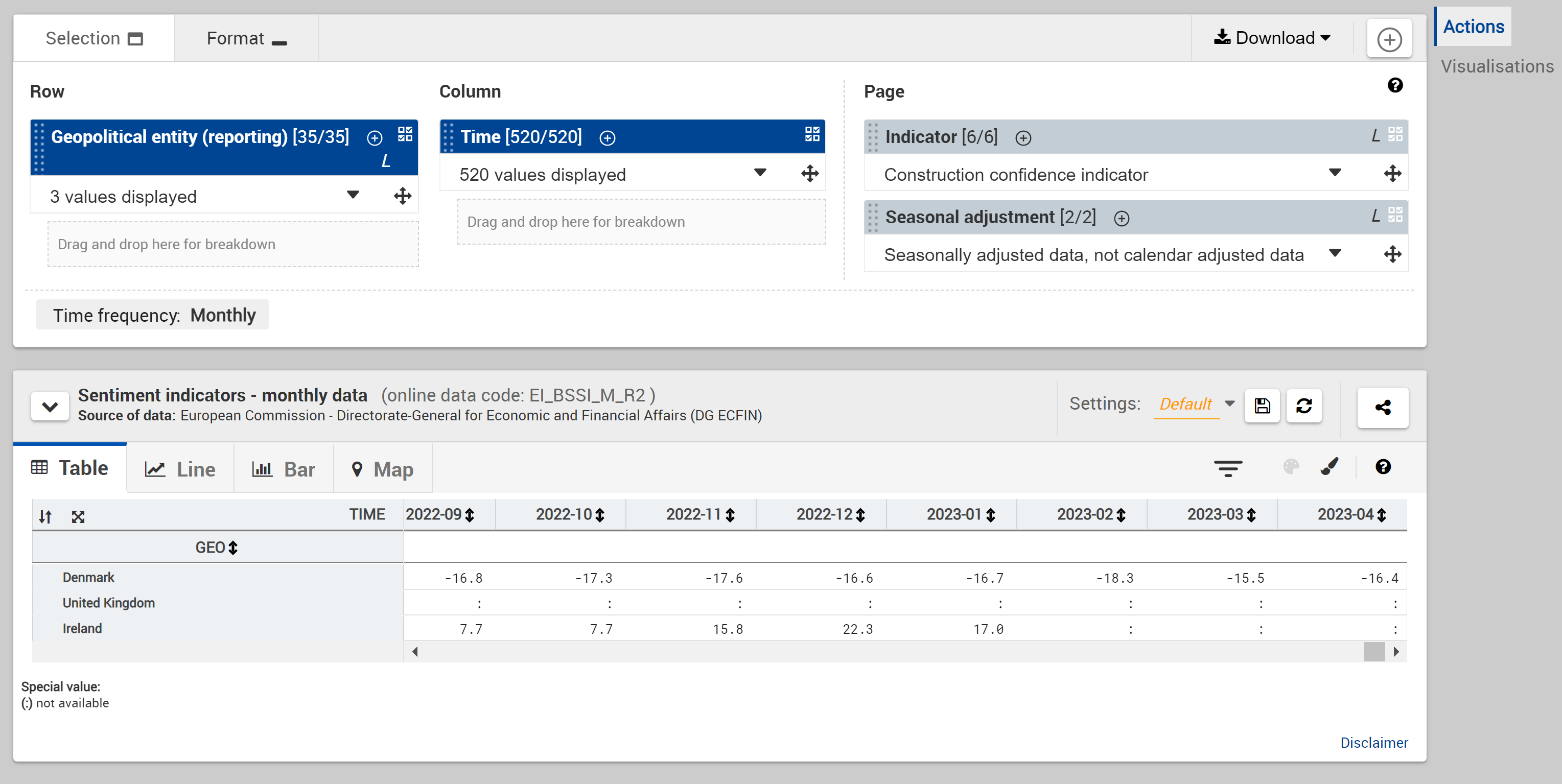
Loy, J. (2017) Neural network projects with python, O’Reilly Online Learning. Available at: https://learning.oreilly.com/library/view/neural-network-projects/9781789138900/9d599ee3-52a3-4c8f-b11b-e6ad402e39a3.xhtml#:-:text=Why%20neural%20networks%3F (Accessed: 26 May 2023).

Chin, L. and Dutta, T. (2016) *NumPy Essentials*, *O’Reilly Online Learning*. Available at: https://learning.oreilly.com/library/view/numpy-essentials/9781784393670/ch01.html#:-:text=Chapter%201.%20An%20Introduction%20to%20NumPy (Accessed: 26 May 2023).

# Appendix

Dataset downloaded from :

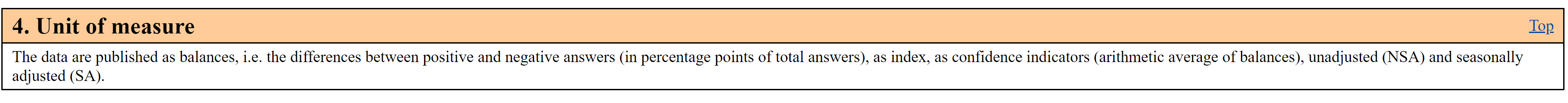
https://ec.europa.eu/eurostat/databrowser/view/EI\_BSSI\_M\_R2\_\_custom\_6284756/default/table?lang=en



The unit of measure for the Confidence/Sentiment indicator:

<https://ec.europa.eu/eurostat/cache/metadata/en/ei_bcs_esms.htm#unit_measure1678715053148>

The data are published as balances, i.e. the differences between positive and negative answers (in percentage points of total answers), as index, as confidence indicators (arithmetic average of balances), unadjusted (NSA) and seasonally adjusted (SA).



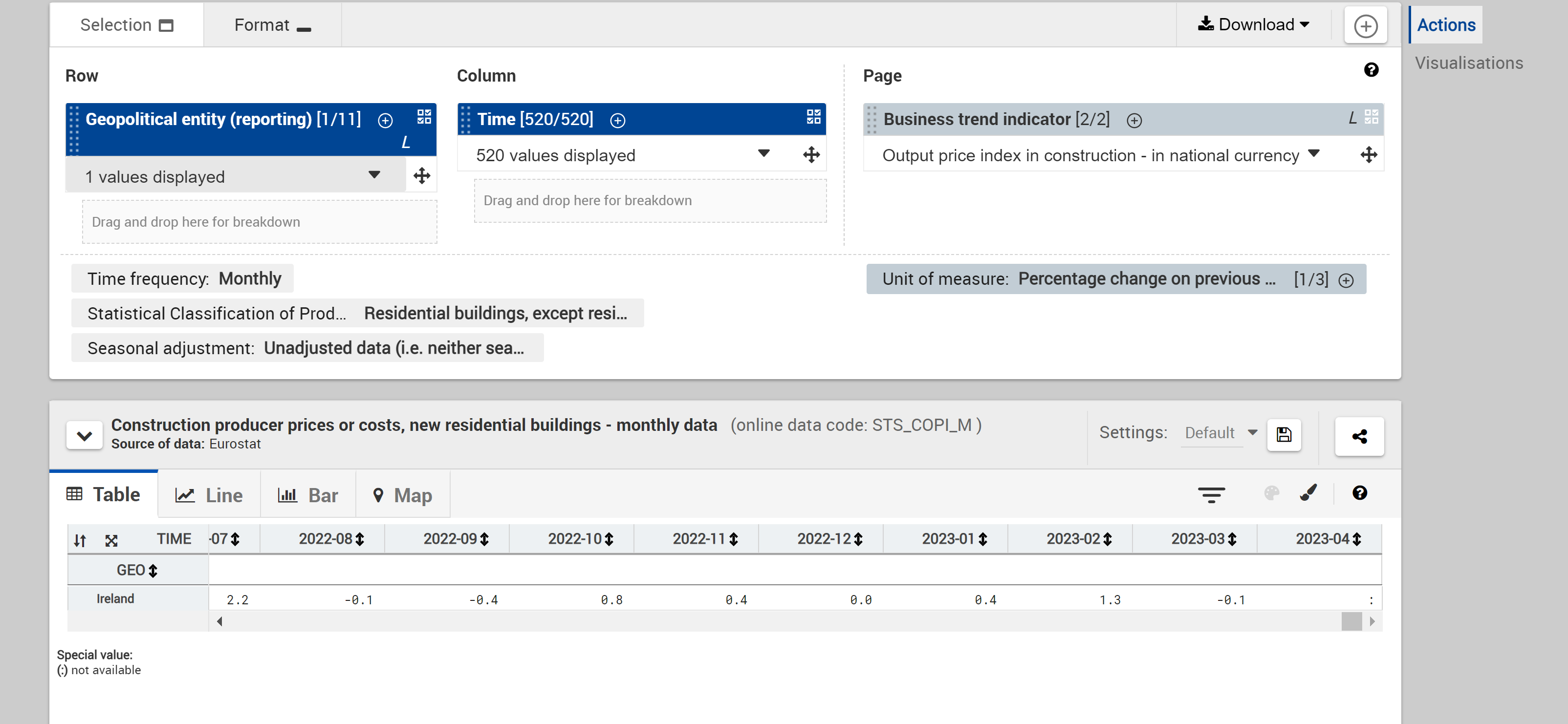
Balances refer to the differences between positive and negative answers. For example, if 60% of people gave a positive answer and 40% gave a negative answer, the balance would be +20 percentage points.

Dataset 2:

Producer prices new residential buildings. Only available for Ireland

https://ec.europa.eu/eurostat/databrowser/view/STS\_COPI\_M\_\_custom\_6285414/default/table?lang=en

Unit is the percentage change on previous period

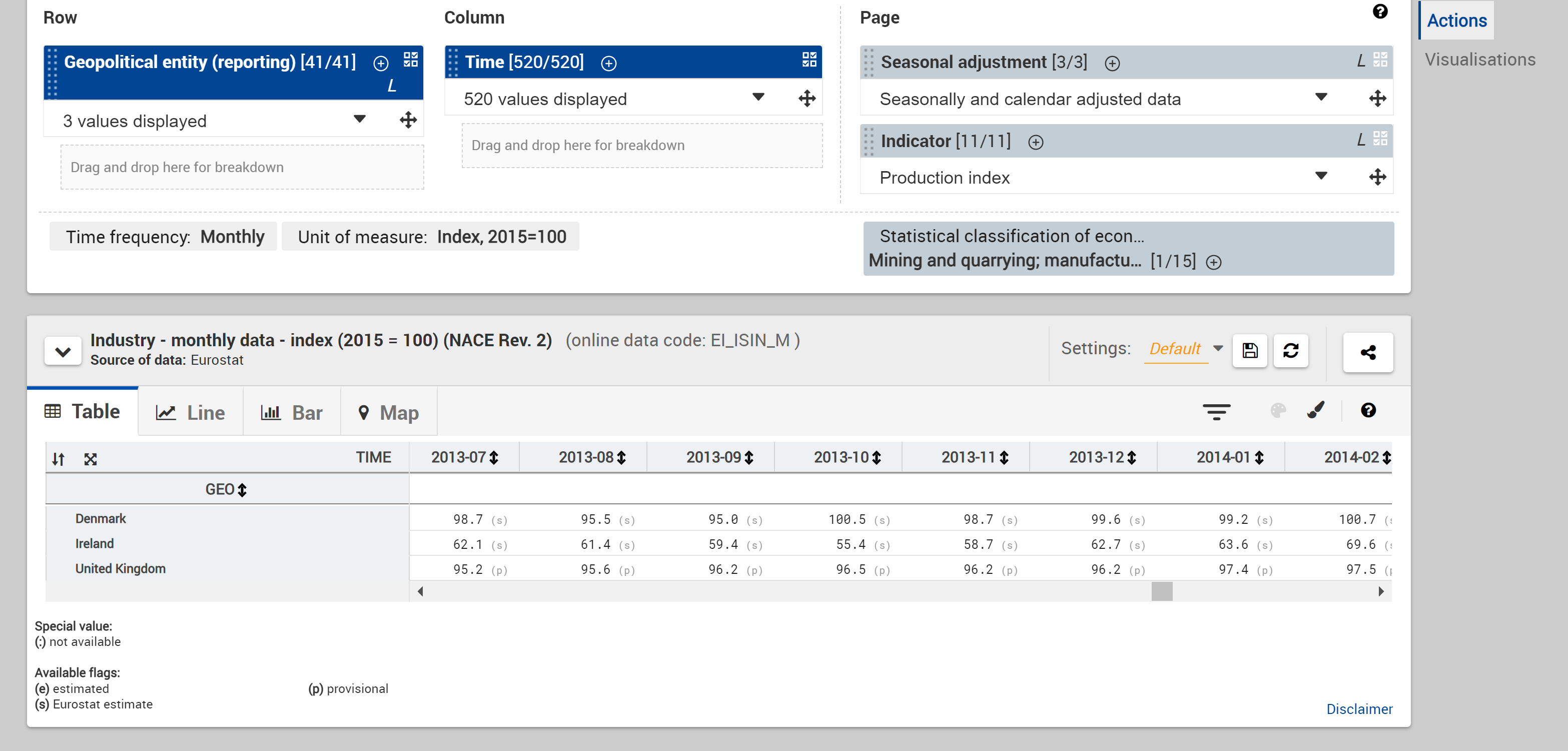


Data set 3:

Industry - monthly data - index (2015 = 100)

Production Index

Mining and quarrying; manufacturing; electricity, gas, steam and air conditioning supply; construction

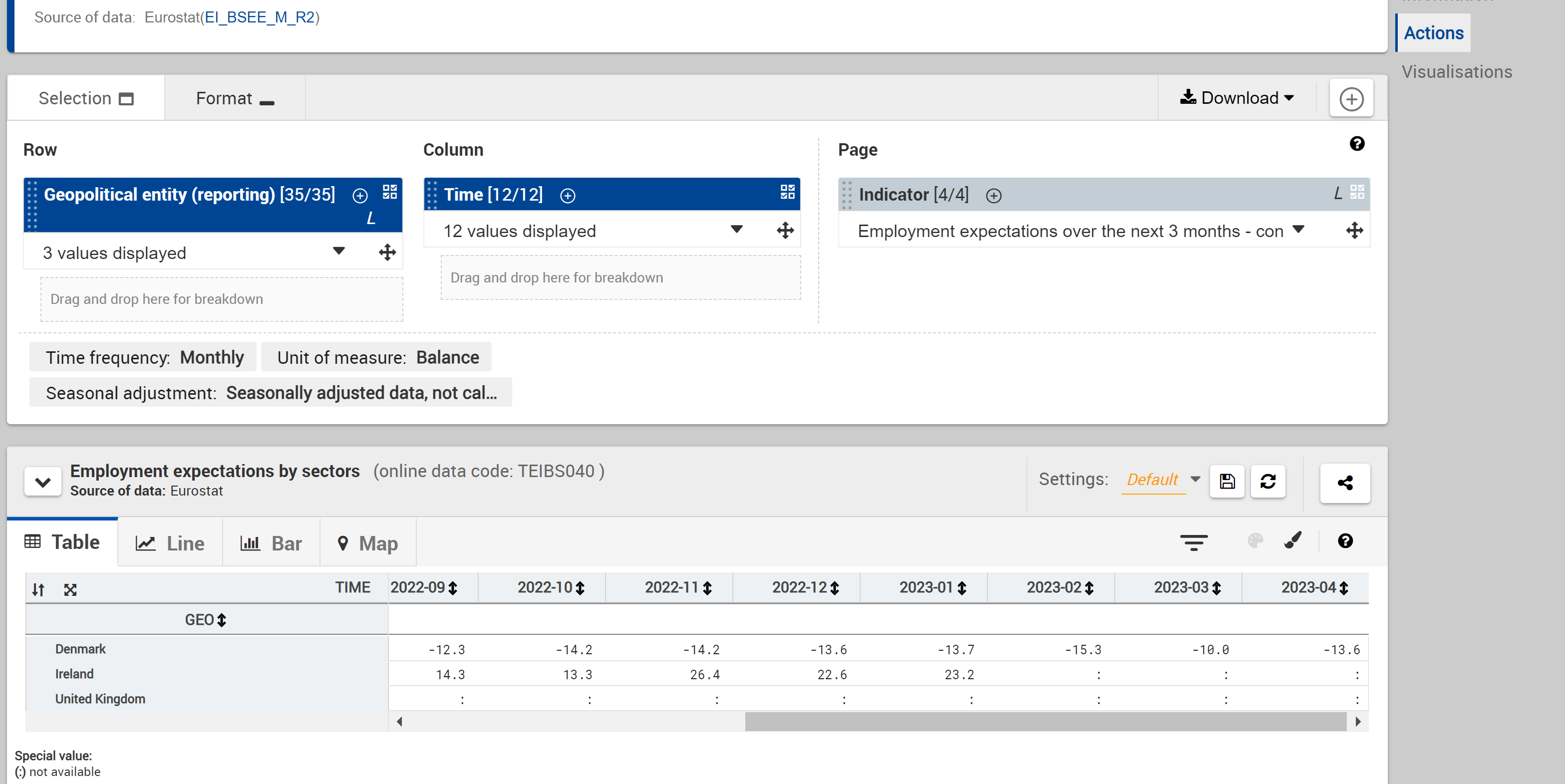


<https://ec.europa.eu/eurostat/databrowser/view/EI_ISIN_M__custom_6285603/default/table?lang=en>

Dataset 4

Employment Expectations

Balance unit



Not available for UK

https://ec.europa.eu/eurostat/databrowser/view/TEIBS040\_\_custom\_6285718/default/table?lang=en

Dataset 5

National House Construction Cost Index

https://data.gov.ie/dataset/national-house-construction-cost-index?package\_type=dataset

Required melting the data

One thing I noticed during research for gathering datasets is that Ireland is not producing the same level of statistics as many other countries.