# 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](#_Toc135127163)

[Abstract 2](#_Toc135127164)

[Introduction 2](#_Toc135127165)

[1. Data Preparation & Visualisation Tasks 3](#_Toc135127166)

[Acquiring Raw Data 3](#_Toc135127167)

[Exploratory Data Analysis 3](#_Toc135127168)

[Data Cleansing and Preparation 3](#_Toc135127169)

[Interactive Dashboard 4](#_Toc135127170)

[2. Statistical Analysis 4](#_Toc135127171)

[Descriptive Statistics 4](#_Toc135127172)

[Analysis of Variables 4](#_Toc135127173)

[Inferential statistical techniques for comparison 4](#_Toc135127174)

[Outcome to deepen research 4](#_Toc135127175)

[3. Machine Learning Tasks 5](#_Toc135127176)

[Rationality of choice of machine learning models 5](#_Toc135127177)

[Webscraping and sentiment analysis 5](#_Toc135127178)

[Testing models that were developed 5](#_Toc135127179)

[4. Programming for DA Tasks [0-100] 5](#_Toc135127180)

[Programming 5](#_Toc135127181)

[Data structures: 6](#_Toc135127182)

[Documentation: 6](#_Toc135127183)

[Testing & Optimisation: 6](#_Toc135127184)

[Data manipulation: 6](#_Toc135127185)

# Abstract

# Introduction

This project uses a dataset published by The Department of Housing which contains applications for planning permission in Ireland between 1987 and 2023. The goal of the project is to assess whether statistical and machine learning methods can be applied to predict whether an application for a “one off house” was granted or refused. The dataset itself comes in the form of a CSV file, and requires exploratory data and statistical analysis and visualisation to help gain insight in data quality and usability.

# Data Preparation & Visualisation Tasks

## Acquiring Raw Data

*Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition. This should include the relevance and implications of any and all licensing/permissions associated with the data. [0-15]*

### Eurostat Database

Eurostat is an online resource which offers datasets and insights into EU countries. 4 of the 6 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 **GUI engine** to 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 analysis 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 structured differently, whereby the years and data were stored vertically as individual observations. It is well known that when performing time series or forecasting that the time range is best stored vertically.

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 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.

## Exploratory Data Analysis

*Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.****[0-25]***

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 show 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. 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.

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 each other, 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.

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

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

### Initial look at the Irish dataset

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 is faster.  
  
For example, “Employment\_Expectation\_Score” has many null values in comparison to the other columns. It can be 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) shows the first 15 records in the table appear 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 it can be seen that 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

*Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.[0-30]*

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

As seen during EDA, there are malformatted 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.

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

## Interactive Dashboard

*Modern construction has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern construction industries, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development.* ***Note you may not use Powerbi, rapidminer, tableau or other such tools to accomplish this (at this stage).[0-30]***

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 are 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 it’s data to Eurostat.

# Statistical Analysis

## Descriptive Statistics

*Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. [0-20]*

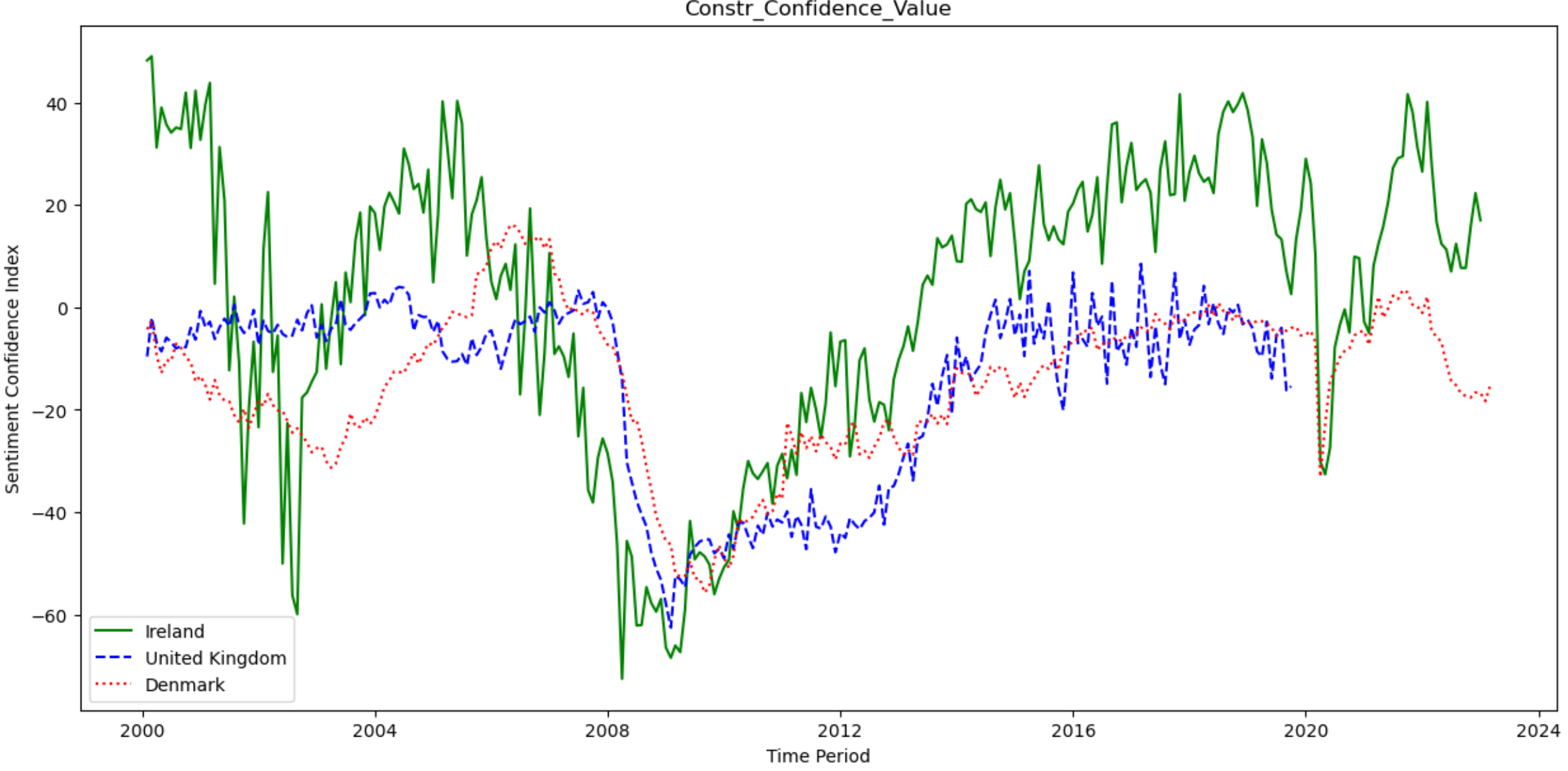
### Coefficient of Variation and Standard Deviation

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

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, 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.

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.



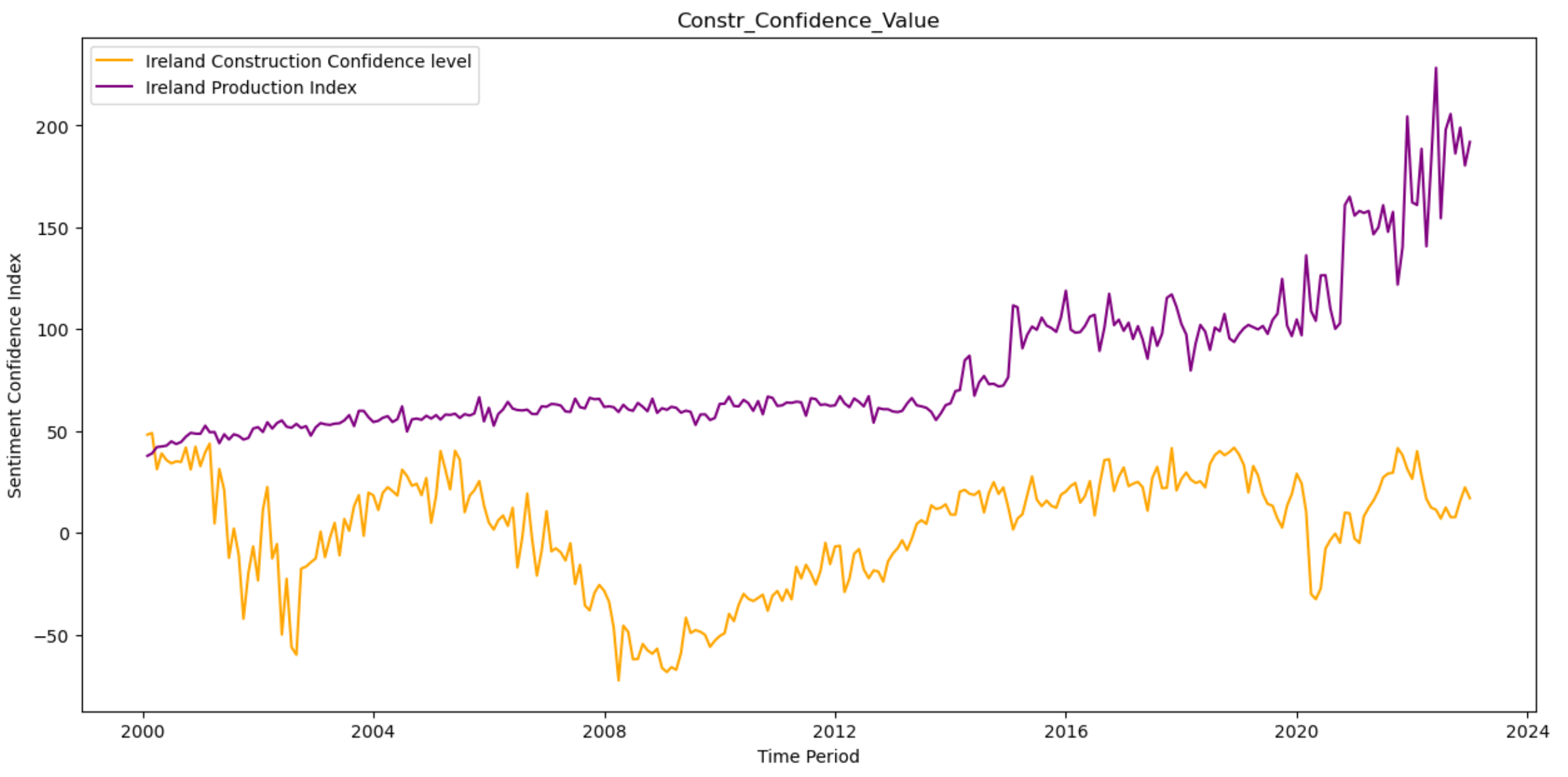
The volatility of Ireland’s score is clearly seen in green in the graph above.

## Analysis of Variables

*Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with international home building, you could find a confidence interval for the population proportion of yearly apartment builds out of all home builds).* ***[0-20]***

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).

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 tends to be a tendency for the values of "Industry\_Prod\_index" to also increase, although the relationship may not be extremely strong. The below graph illustrates this:



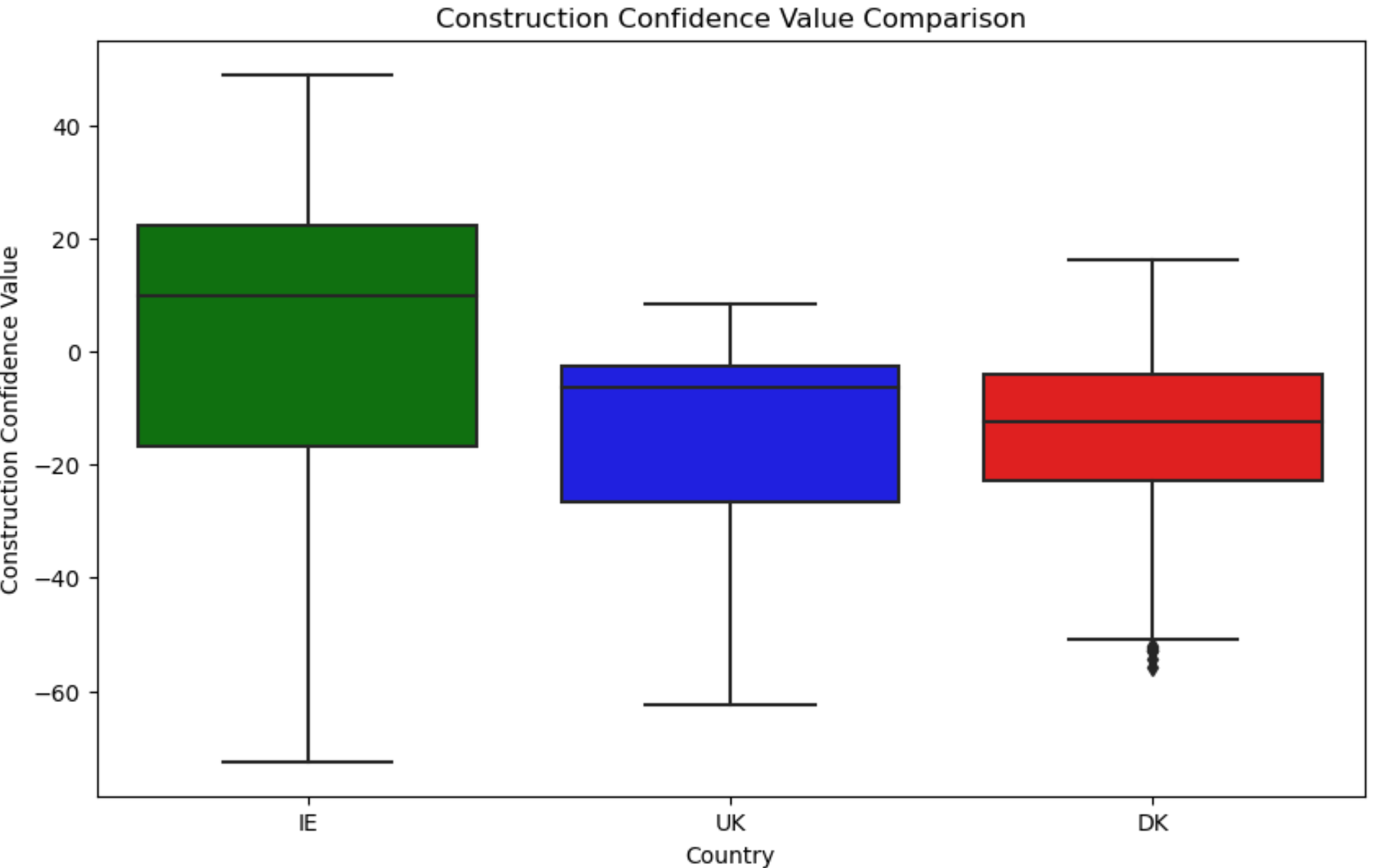
## Inferential statistical techniques for comparison

*Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests.* ***[0-40]***

There are two variables in the dataset which are populated across all three datasets.

The first is the Construction Confidence Index and the second is the Industry\_Prod\_index. The graph in the section above shows Ireland as having a lot more volatility than the UK and DK countries.

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.



## Outcome to deepen research

*Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process.* ***[0-20]***

# Machine Learning Tasks

## Rationality of choice of machine learning models

*Use of multiple models (at least two) to compare and contrast results and insights gained.*

*Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.*

***[0 - 30]***

## Webscraping and sentiment analysis

*Collect and develop a dataset based on the construction topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate construction topic (e.g., house price, availability of labour etc…) for producers and consumers point of view in Ireland.*

***[0 - 25]***

Sentiment analysis done but I need to build out the output a bit more. Maybe try to incorporate Bigrams and identify themes in the comments that way to find the comments most relevant to my topic.

## Testing models that were developed

*You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.*

***[0 - 30]***

***Graphics and Table that demonstrates the ML outcomes***

*A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.*

***[0 - 15]***

# Programming for DA Tasks [0-100]

## Programming

***Programming:*** *The project must be explored programmatically: this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook.* ***[0-20]***

## Data structures:

*You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format.* ***[0-20]***

## Documentation:

*The project documentation must include sound justifications and explanation of your code choices. Code quality standards should also be applied.* ***[0-20]***

## Testing & Optimisation:

*You are required to document and evaluate a testing and optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). Note any trade-offs that you've made in these areas.* ***[0-20]***

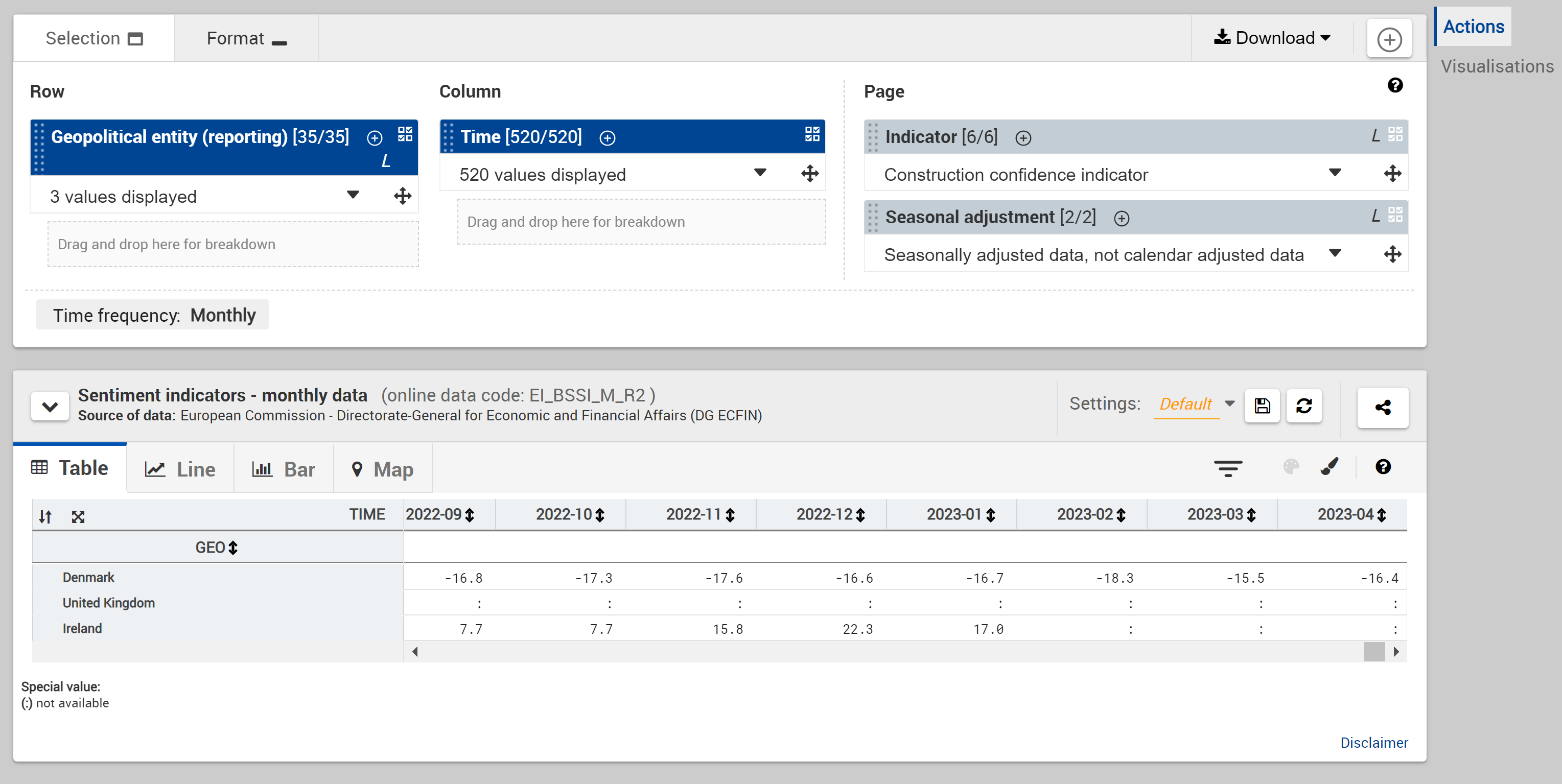
## Data manipulation:

*For each of the different data sources, compare and contrast at least two relevant libraries and techniques for a) processing and b) aggregating the respective data, in order to justify your chosen libraries/techniques.* ***[0-20]***

# Appendix

Dataset downloaded from :

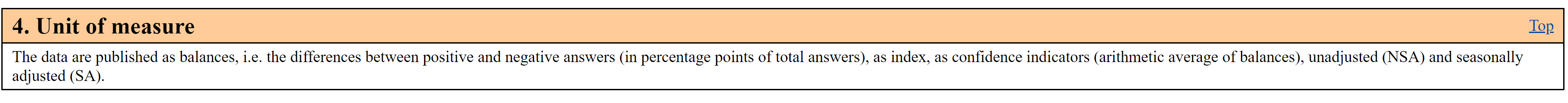
https://ec.europa.eu/eurostat/databrowser/view/EI\_BSSI\_M\_R2\_\_custom\_6284756/default/table?lang=en



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).



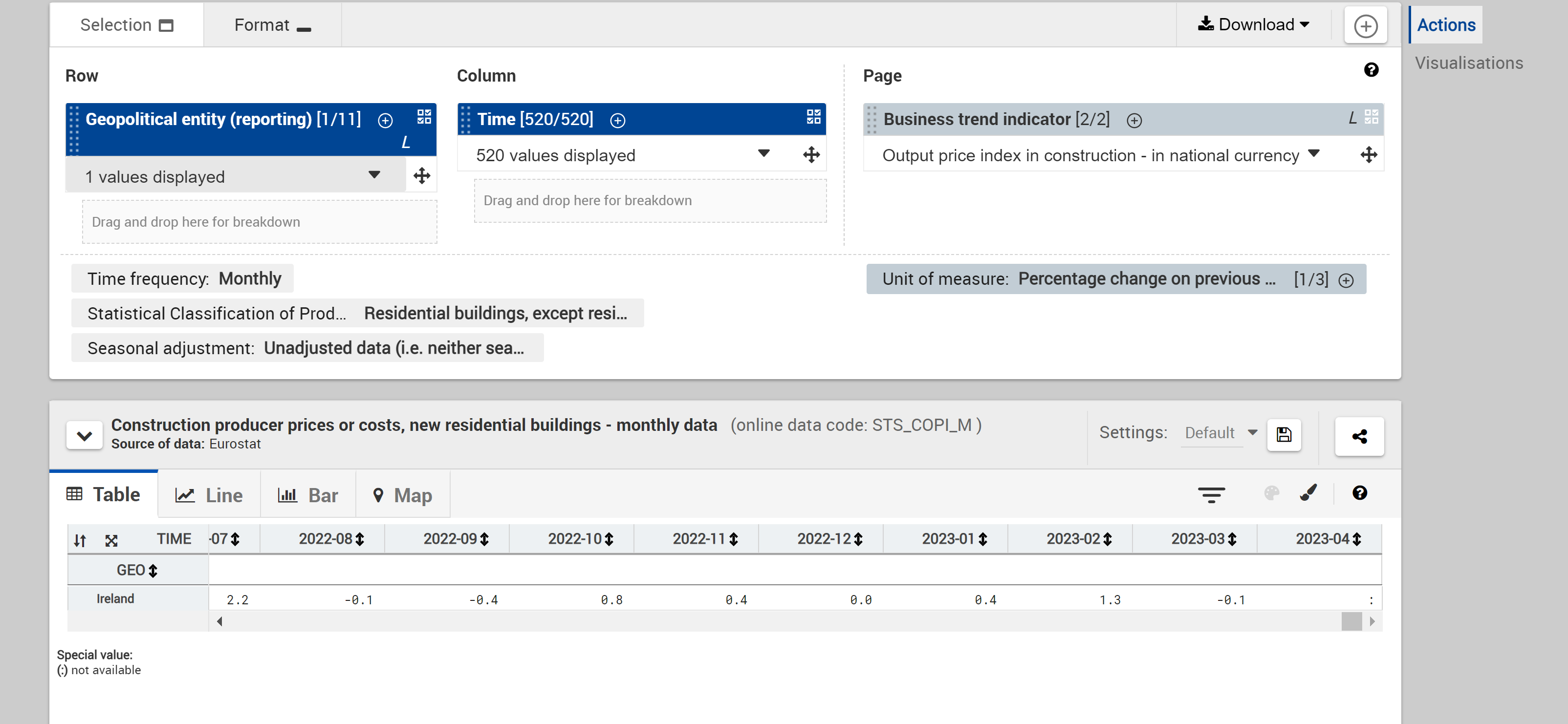
Balances refer to the differences between positive and negative answers. For example, if 60% of people gave a positive answer and 40% gave a negative answer, the balance would be +20 percentage points.

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

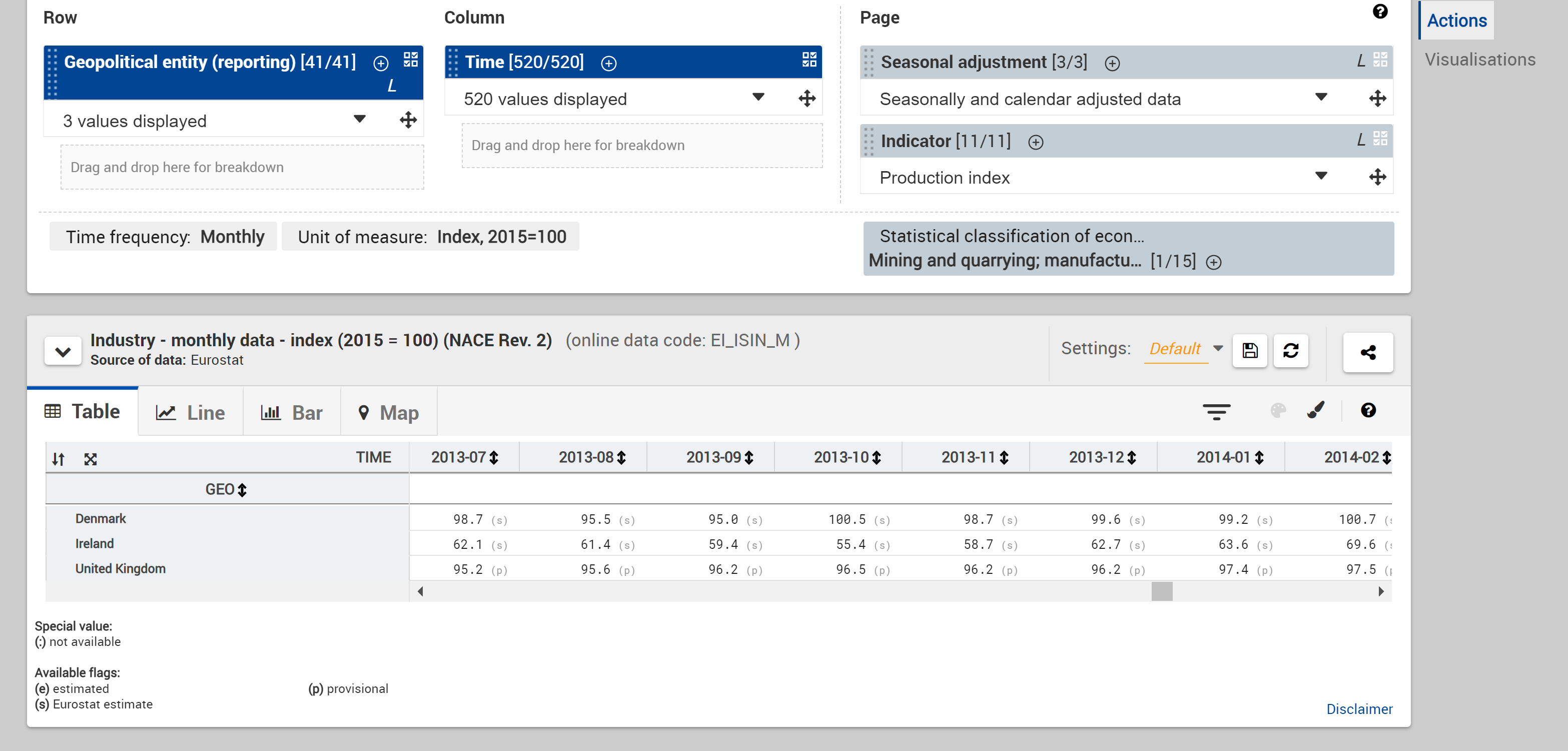


Data set 3:

Industry - monthly data - index (2015 = 100)

Production Index

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

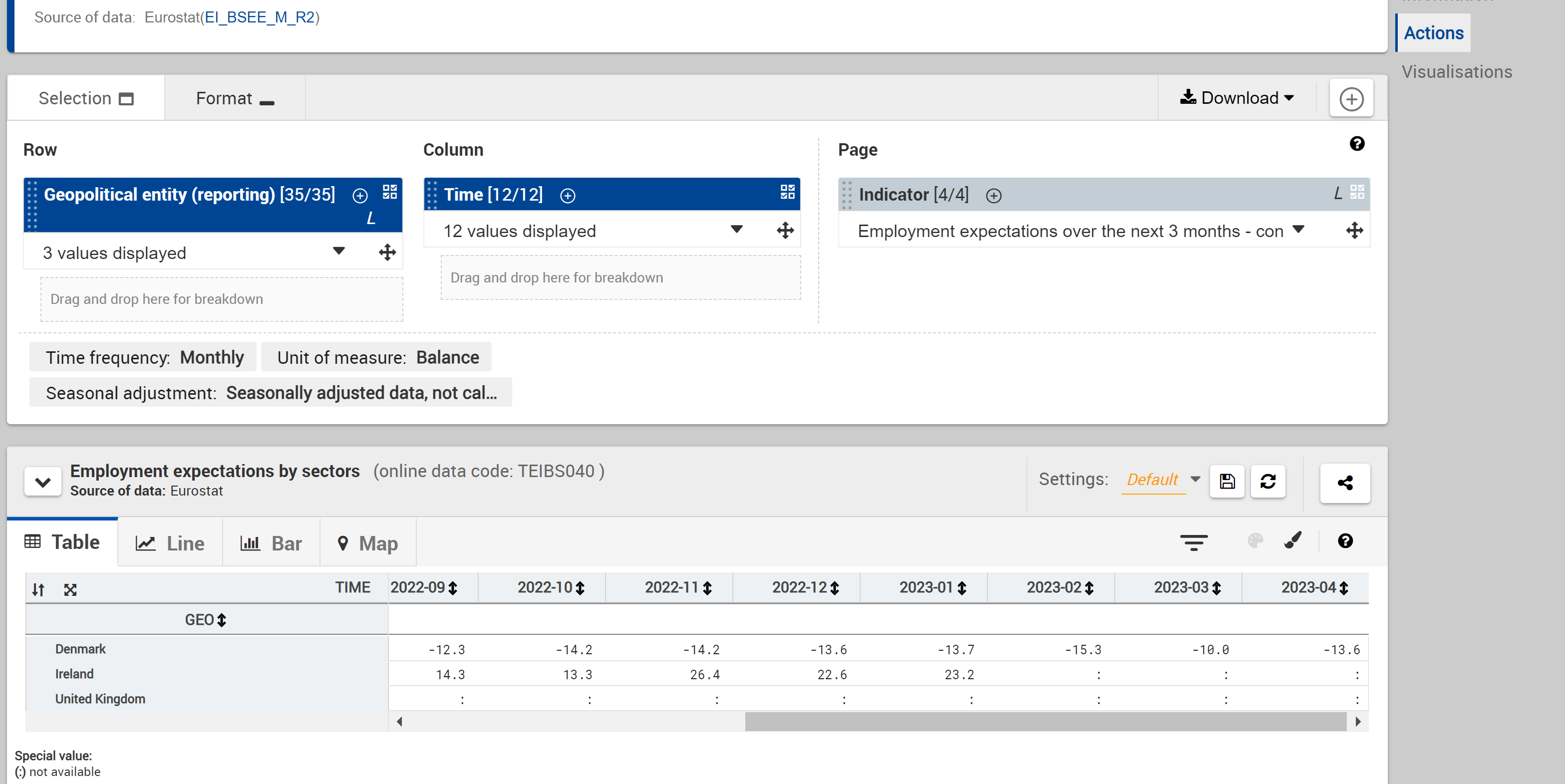


<https://ec.europa.eu/eurostat/databrowser/view/EI_ISIN_M__custom_6285603/default/table?lang=en>

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.