Continuous Assessment 2

Author: Ciara O'Hara

Programme: MSc in Data Analytics May 2023

Student ID: sbs23015

Email: sbs23015@student.cct.ie

May 16, 2023

Contents

0.1 brief

You have been tasked with analysing Ireland's Construction data and comparing the Irish Construction sector with other countries worldwide. This analysis should also include forecasting, sentiment analysis and evidence-based recommendations for the sector as well as a complete rationale of the entire process used to discover your findings. Your Research could include export, import, trade imbalance, house production, material stock, labour/skill pool, etc. (or any other relevant topic EXCEPT Climate change) with Ireland as your base line.

Abstract

Introduction

Info:

https://oecdecoscope.blog/2021/12/13/finlands-zero-homeless-strategy-lessons-from-a-success-story/comment-page-1/

https://www.oecd.org/housing/policy-toolkit/data-dashboard/boosting-efficiency/

https://www.linesight.com/insights/regional-report/europe-2021/

https://www.geeksforgeeks.org/newspaper-scraping-using-python-and-news-api/

https://towardsdatascience.com/web-scraping-news-articles-in-python-9dd605799558

https://www.geeksforgeeks.org/newspaper-article-scraping-curation-python/

Materials and Methods

First steps were to determine and identify an appropriate Irish dataset under the theme of Constrction. There were two datasets found: A House Construction Cost Index from 1975 - 2017, and a social housing construction status reports from 2017 - 2021. Homelessness in Ireland is an issue of major significance and public importance at the moment. In fact, this is an issue across Europe to varying degrees, with the excpetion of Finland. It was decided to try to determine - from publicly available data - what Finland has done differently, the factors that may have impacted that, and to attempt a sentiment analysis around the topic in both Ireland and Finland. As such, the next step was to gather appropriate and complementary Finnish data. Statistics Finland's free-of-charge statistical databases, Tilastokeskus was found, which included ... in JSON format, accessed via APIs.

3.1 Data sources

Population statistics from https://ec.europa.eu/eurostat/databrowser/view/TPS00001/default/table?lang=en

3.2 Programming

3.2.1 Python libraries used

3.3 Data preparation

3.3.1 Cleaning the data

3.3.2 Data visualisation

The colours chosen for the sentiment analysis visualisations were designed to reflect the traditional colours associated with political leanings, for example, red for left-leaning politics, magenta for

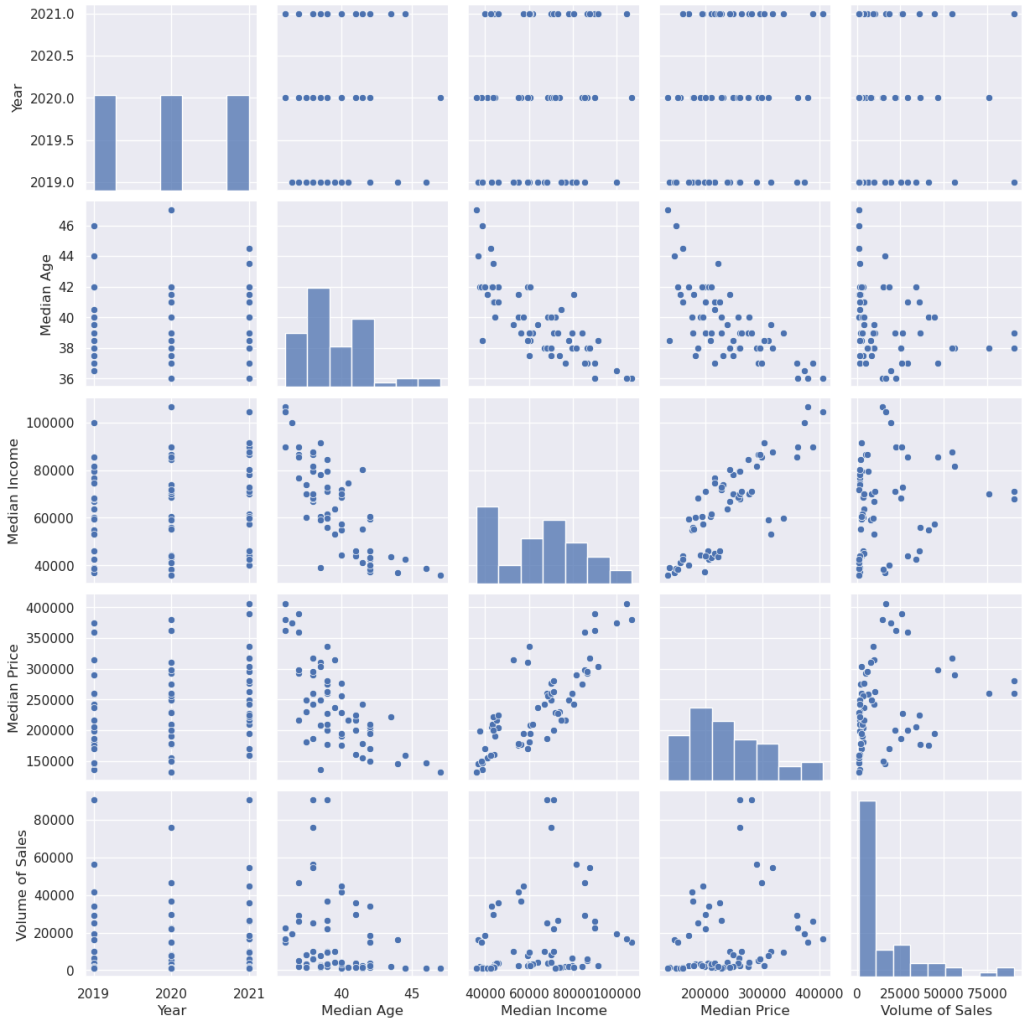

Figure 3.1: *Sample figure.*

centrist, and blue for conservative politics.

3.4 Statistics

3.4.1 Python libraries

3.4.2 Statistical analysis

ARIMA is an appropriate tool for modelling regular-interval, non-seasonal, time-series data, and may be used to predict data in the series into the future (*What Is ARIMA Modeling?* 2023). Univariate analysis was employed here, in that only previous values in the time series were used to predict future ones. First an augmented Dickey-Fuller test was carried out to determine whether the data was

stationary. This test showed that it was not stationary (p = 0.6, DFτ = -1.33), i.e. it had some time-dependent structure (*Augmented Dickey-Fuller Test in Python (With Example)* 2023). Autocorrelation and augmented Dickey-Fuller were used to determine the order of differencing (Raval, 2023), which was 1 (augmented Dickey-Fuller: p = 1.0 x 10^{-5}, DFτ = -5.17), and the most significant degree of lag was determined from the autocorrelation plot (taken to be 50) (Raval, 2023; Nadeem, 2021). Maximising the goodness-of-fit of the ARIMA model is done by determining the best values for p, d and q and this was done by minimising the RMSE (Raval, 2023; *How to do cross validation for time series?* 2023) using cross-validation on a rolling basis, which ensures that no 'future' data is used in the training of the model (Shrivastava, 2020; Hyndman and Athanasopoulos, 2018). An order of differencing of 2 was later found to give better results (based on RMSE) than a value of 1.

Ireland and Finland have roughly similar populations

As there was only four years of Irish social housing construction data available, the irish and finnish cost indexes and construction activity could only be compared over 4 years, which is not a long enough dataset. The inferential analysis was still carried out, but in reality, any results based on such a small dataset would be meaningless. In oder to make valid compaisons acrss both countries the cost index and number of houses built were defined on a per-capita basis. A Shapiro-Wilk test showed that the data could be considered normally distributed (Irish cost index per Capita: W=0.870, pvalue=0.298, Irish houses built per Capita: W=0.827, pvalue=0.161, Finnish cost index per Capita: W=0.763, pvalue=0.051, Finnish houses built per Capita: W=0.839, pvalue=0.193).

Table 3.1: *Results of Shapiro-Wilk tests for normality on the Irish and Finnish construction costs and housing units completed per-capita.*

Variable	W Test statistic	p-value
Irish cost index per Capita	0.870	0.298
Irish houses built per Capita	0.827	0.161
Finnish cost index per Capita	0.763	0.051
Finnish houses built per Capita	0.839	0.193

3.5 Machine learning

The irish data was not split into test and train sets as the dataset was too small.

In splitting the finnish data into test and train a split of 0.3 was taken, as the dataset was small.

3.5.1 Machine learning methods used

Results

Conclusion

References

What Is ARIMA Modeling? (2023). Master's in Data Science. URL: https://www.mastersindatascience.org/learning/statistics-data-science/what-is-arima-modeling/ (visited on 05/11/2023).

Augmented Dickey-Fuller Test in Python (With Example) (2023). Statology. URL: https://www.statology.org/dickey-fuller-test-python/ (visited on 05/11/2023).

Raval, P. (2023). *How to Build ARIMA Model in Python for time series forecasting?* ProjectPro. URL: https://www.projectpro.io/article/how-to-build-arima-model-in-python/544 (visited on 05/11/2023).

Nadeem (2021). *ARIMA: Advanced Time Series Methods: Auto Regression Integrated Moving Average.* Medium. URL: https://medium.com/analytics-vidhya/arima-fc1f962c22d4 (visited on 05/11/2023).

How to do cross validation for time series? (2023). ProjectPro. URL: https://www.projectpro.io/recipes/do-cross-validation-for-time-series (visited on 05/11/2023).

Shrivastava, S. (2020). *Cross Validation in Time Series*. Medium. URL: https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4 (visited on 05/11/2023).

Hyndman, R. J. and G. Athanasopoulos (2018). *Forecasting: Principles and Practice*. Melbourne, Australia: OTexts.