A picture containing icon

Description automatically generated

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | Summer 2023 Repeat Assessment – Statistics for Data Analysis & [Machine Learning for Data Analysis](https://moodle.cct.ie/mod/resource/view.php?id=116593) |
| **Assessment Title:** | MSC\_DA\_InterGr\_Repeat Sem1 ver3 |
| **Lecturer Name:** | Marina Iantorno  Muhammad Iqbal |
| **Student Full Name:** | Stephen Somerville |
| **Student Number:** | Sba22232 |
| **Assessment Due Date:** | 26th may 2023 |
| **Date of Submission:** | 5th Of May 2023 |
| **Word Count:** | 3087 – As I only had 2 modules to repeat I have only completed the sections and tasks related to those areas, this is an incomplete report as a result. |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

|  |  |
| --- | --- |
| **Table of Contents:** | |
| **Page 1:** | **Introduction** |
| **Page 2 to 6:** | **Statistics** |
| **Page 7 to 9:** | **Machine Learning** |
| **Page 10:** | **Sentiment Analysis** |

GIT REPO Address - https://github.com/sba22232/Sem1CA1Repeatsba22232

**Introduction**

In this report, I have chosen Dublin House pricing and compare the price Index for Ireland with other countries in Europe. The focus of the analysis will be demonstrating the variance of price within Dublin and then how that compares across Europe. I understand the data I have chosen and subject itself are widely studied and therefore I will be confirming the consensus with my report rather than providing some new element to the conversation.

The other equally important goal of this report is demonstrate my assessment Tasks for this repeat assignment and as such will only be covering areas related to the two modules in question - Statistics for Data Analysis & [Machine Learning for Data Analysis](https://moodle.cct.ie/mod/resource/view.php?id=116593). There are areas from the other modules where there are over lap – Data Preparation etc that will be included however not all areas from those modules will be address as they don’t fall under the assessment tasks as set out.

As I have some experience with Mortgage lending, Financial services and conveyancing I would use this knowledge to provide additional insights and recommendations to potential stakeholders in the Dublin Housing Market, including home owners, renters, and construction companies. Again, to achieve this, we will follow the CRISP-DM methodology and cover the six major steps of the process: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

Throughout the report, I will draw on data and analysis techniques to understand the current state of the Dublin Housing sector and identify trends and opportunities. I demonstrate the tools, steps, and processes that I would follow for actual housing stakeholders.

**Statistics**

My data split across two datasets –

1. Dublin House Price (DHP) by area from 2011 to 2020 and is used under Licence for Re-Use of Public Sector Information adopts CC-BY as the [standard PSI licence](https://data.gov.ie/pages/re-useofpublicsectorinformation).
2. European House price Index (EHPI) by year and quarter from 2010 to 2022 which is supplied under the EU [Creative Commons Attribution 4.0 International (CC BY 4.0) licence](http://creativecommons.org/licenses/by/4.0/)

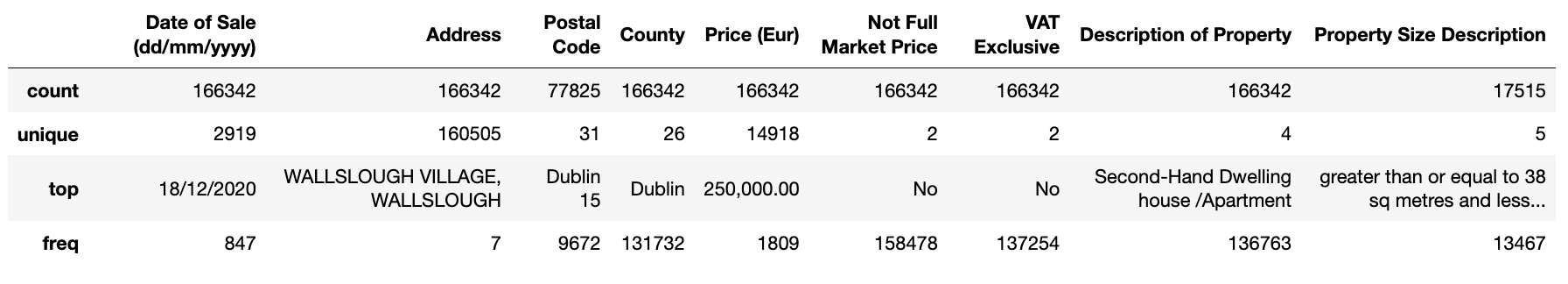
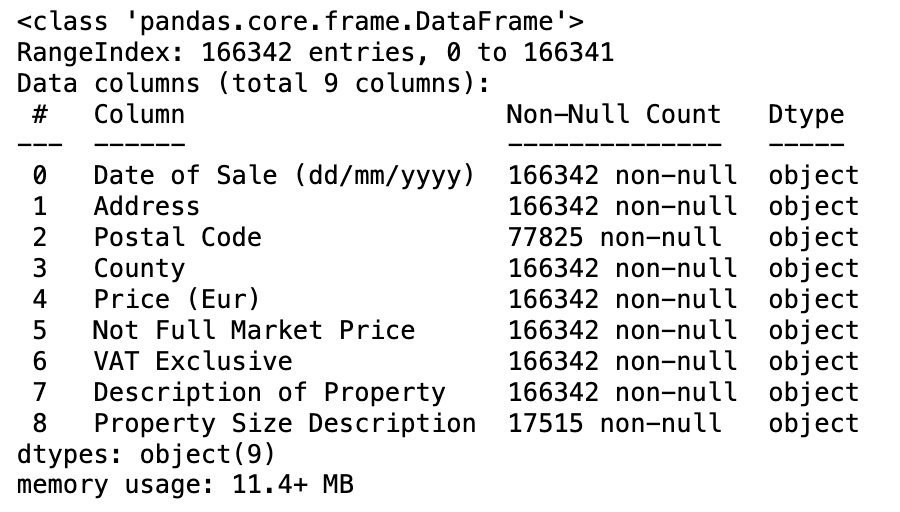
As I am much further into my Data Analysis journey, I have developed a process I like to follow with my Data prep/cleaning and descriptive analysis, as follows –

Structural investigation: Here I explore the general shape of a dataset and the data types of its features. Examine the numbers of rows & columns in the dataset, identifying any missing or duplicate data, checking for outliers and other unusual values, and evaluate the distribution of values for different features. Once I complete these tasks, I have a deeper understanding of the data, its structure and what I can do with it, while also being able to identify potential issues and/or challenges that may need to be addressed in the subsequent stages of the analysis.

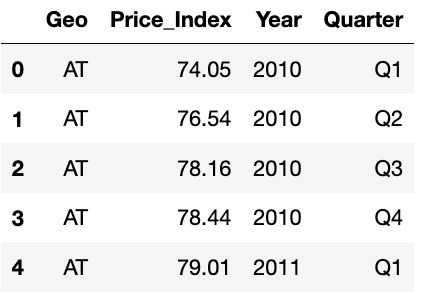
Here are the top 5 rows in the DHP dataset –



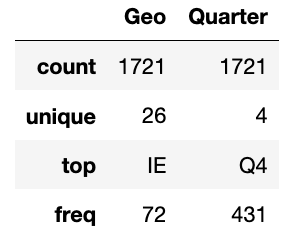
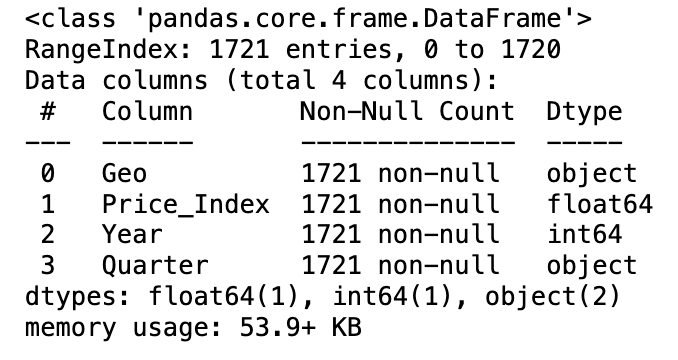
Description of the data and its data types –



Here are the top 5 rows in the EHPI dataset –



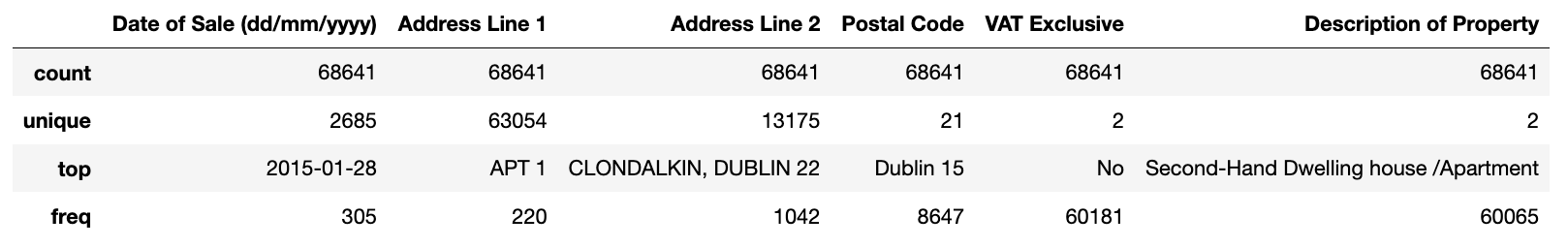
Description of the data and its data types –



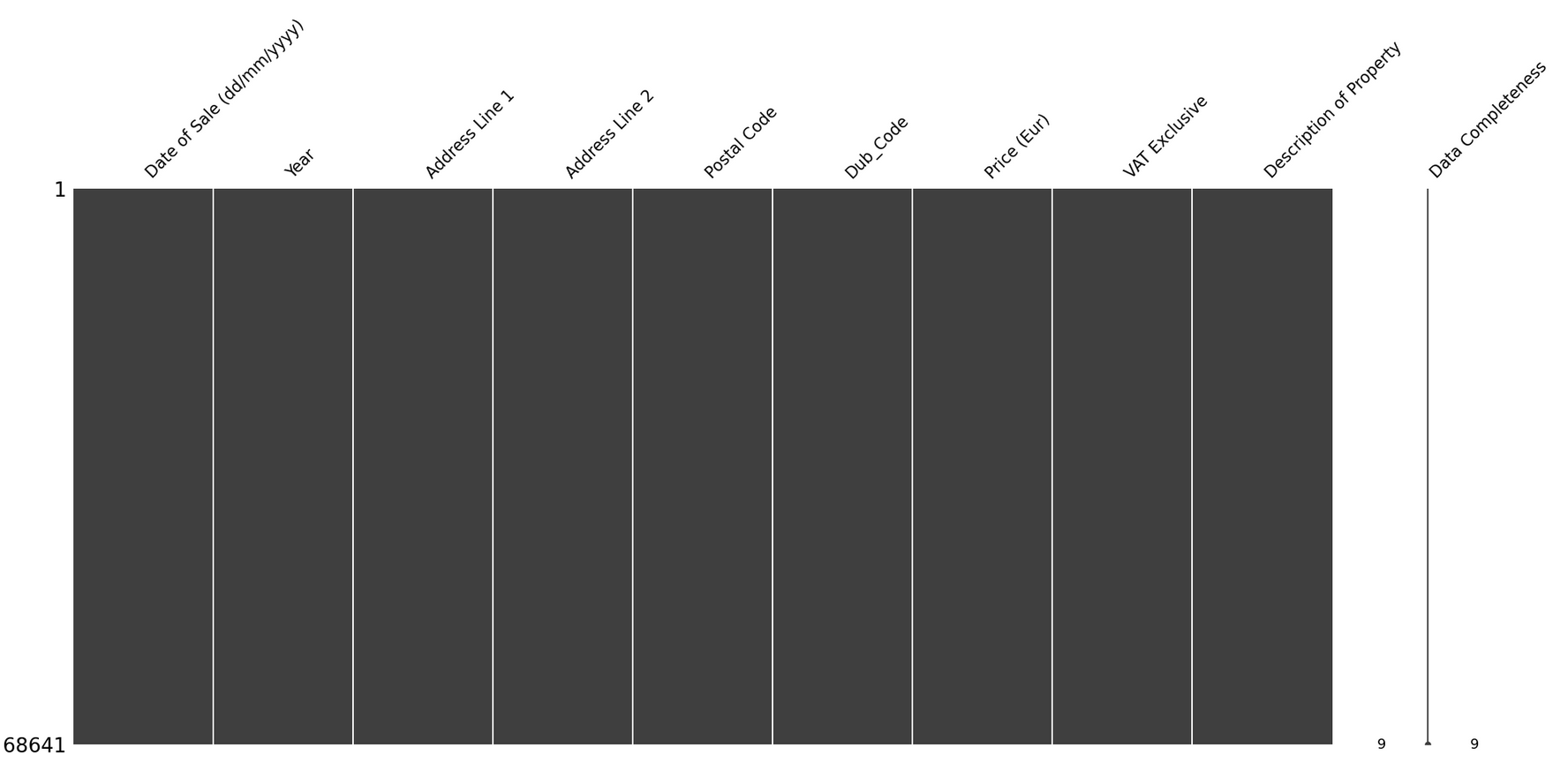
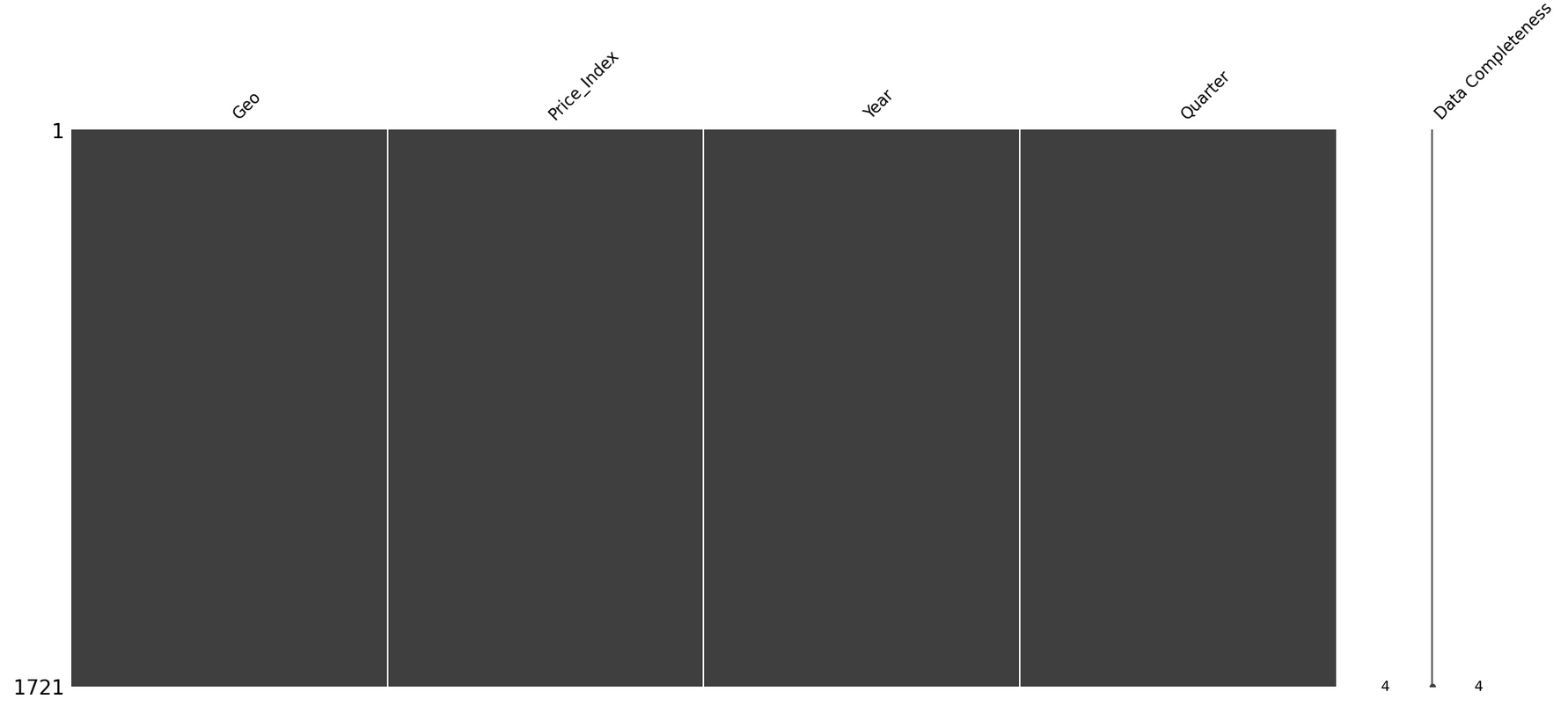
Next I moved into my Qualitative investigation where I decided on the how and what changes were going to be implemented to the table so I had more options with my analysis,

* In the EHPI data, I didn’t need to make any changes as it was laid out pretty simply.
* For the DHP dataset I had to the following -
  + After identifying any entries that had unique, duplicated or Null values – I removed them.
  + Filtered and dropped out Counties outside of Dublin – there wasn’t enough data to track the other counties over the 10 years so Dublin remain consistent.
  + Corrected the formatting on some numerical fields like changing to DateTime for the date or removing the comma so the price was changed to a float from string
  + Separated year from the date into a new column
  + Drops row with properties not sold at market price, these would outliers and sold as estate sales, family to family purchases or heavily discounted corporate sales. Whatever the reason, they are not the norm and I removed them.
  + With the Post Code I split the values so the Dublin number e.g Dub15 = 15 in a new column and dropped ‘6w’.
  + Lastly I removed any sales less than 100k and over 1m – purely arbitrary as I wanted to compare homes that ‘regular’ people can buy.

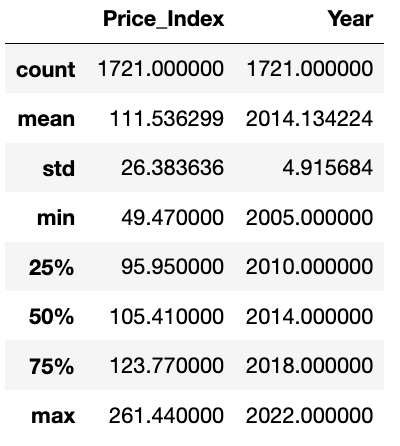
The final text metrics for the DHP dataset are –



Both datasets no longer help any null values –

DHP -  EHPI - 

Content investigation: As mentioned previously the two datasets are a mix of continuous and categorical variables data types. Now the structure and quality of the dataset is understood, I will go ahead and perform a more in-depth exploration on the features values and look at how different features relate to each other. This involved looking at the distribution of values for individual features, as well as examining the relationships between different features. I used different techniques to perform this exploration, such as visualizing the data using plots and charts, calculating summary statistics, and fitting models to the data. Measures of central tendency such as count, mean, median and Standard Deviation were computed to describe the typical values of the data.

EHPI - DHP -

Additionally, I evaluated the dispersion of the data using variance and standard deviation.

* For the EHPI dataset, the variance for Price\_Index was 695.691793.
* For the DHP dataset, the variance for Price (Eur) it was 2.926505

Skewness and kurtosis were also calculated to evaluate the shape of the distribution.

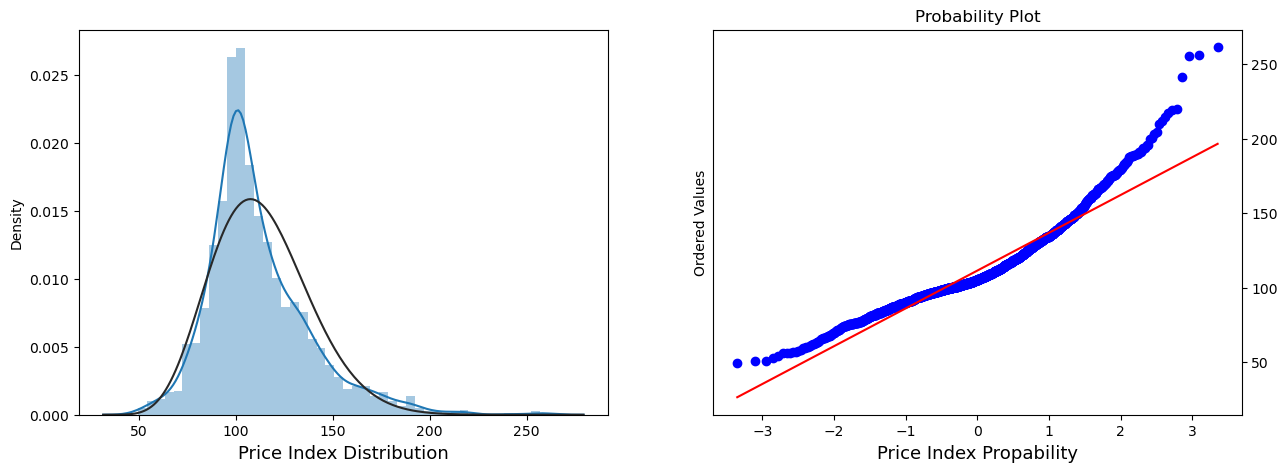
* For the EHPI dataset, the skewness was 1.2608242226001254 and the kurtosis was 2.964501995154232.
* For the DHP dataset, the skewness was 1.2762276870197165 and the kurtosis was 1.6184770676930196.

In both EHPI and DHP datasets - there is a positive skewness, indicating that the datasets distribution is skewed to the right.

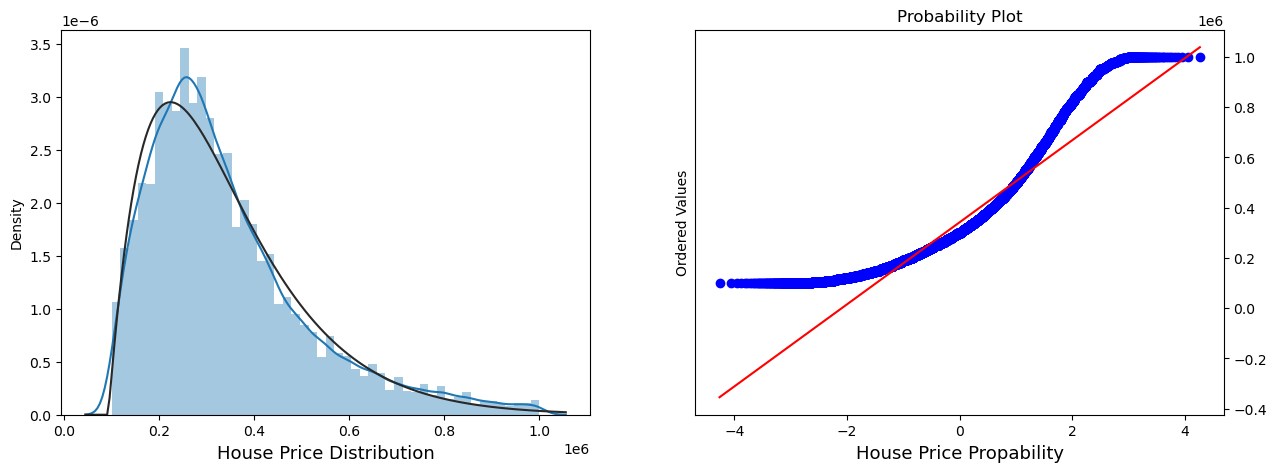
Kurtosis measures the heaviness of the tails of a distribution. In EHPI dataset, the kurtosis is greater than 3, indicating that the distribution has heavier tails than a normal distribution. In DHP dataset, the kurtosis is less than 3, indicating that the distribution has lighter tails than a normal distribution. 3 being the benchmark for normality.

When I compare the skewness and kurtosis values of the two datasets, both have positive skewness, indicating right sided skew. That said, the EHPI dataset has a greater kurtosis value than the DHP dataset, indicating that the EHPI dataset has a heavier tail and a sharper peak than the DHP dataset.

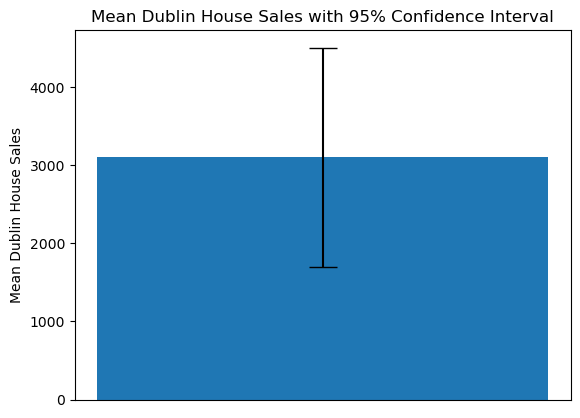
EHPI dataset –



DHP dataset –



Confidence interval based on house sales data per year in Dublin (DHP dataset)



* Mean house sales: 3100.70
* Standard deviation of house sales: 2259.83
* T-statistic: 0.14
* P-value: 0.89
* 95% Confidence interval: (1700.04, 4501.36)

Based on the results above, we can infer that the mean is 3100.70 with a standard deviation of 2259.83.

The t-statistic of 0.14 indicates that there isn’t a significant difference between the sample mean and the null hypothesis mean of 3000. Based on the P-value of 0.89 there is a high probability that the difference between the sample and hypothesis mean occurred by chance. Based on the confidence interval range of (1700.04, 4501.36), I am 95% confident that the true mean lies within this range therefore, I can fail to reject the null hypothesis.

An interesting point to note - When I compare the different property types in the DHP data, There is a significant difference between the means for New and Second hand Homes per year –

* T-statistic of -5.454057810920085
* P-Value of 4.9608953978272

For my inferential test I picked 5 so I had a mix of Parametric and Non-Parametric examples to use. I performed the tests in an order that I felt the fed into each other so I could use the results to guide what variables I would chose next to test. I will add that a lot of the tests are stating the obvious between the influence of time and area over the price of property. I was curious to see to what degree they were impacting the pricing.

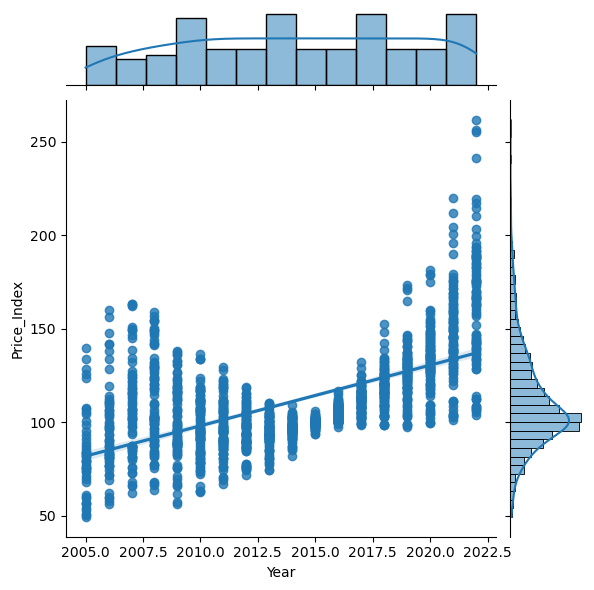
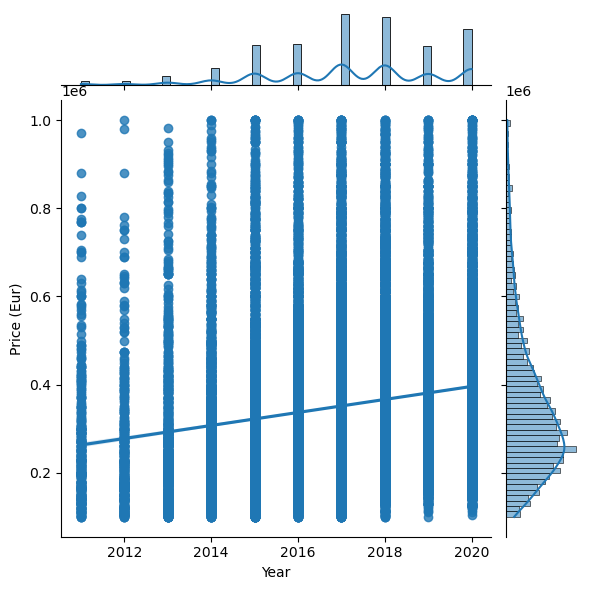
I started with the Shapiro Wilks Test: I wanted to confirm if my data was normally distributed or not so I used the Null Hypothesis that the data is normally distributed. The result was, in both cases, my data wasn’t normally distributed, I can reject the Null Hypothesis.

* EHPI – Statistics=0.923, p=0.000, Sample does not look Gaussian (reject H0)
* DHP – Statistics=0.901, p=0.000, Sample does not look Gaussian (reject H0)

Next I used ANOVA: I used a one way Anova to test the Null Hypothesis – that there was no significant difference in the mean housing prices over the years among the different countries of the EU and separately Ireland. The result suggested that the year had a significant effect on housing prices in both the EU (excluding Ireland) and Ireland datasets, therefore I can reject the Null Hypothesis.

* EU (excluding Ireland) – F-statistic: 138.119, P-value: 4.069 = (reject H0)
* Only Ireland – F-statistic: 138.119, P-value: 4.069 = (reject H0)

\*\* Linear Regression, Pearson Correlation Co-efficient and histogram plot -

EHPI - DHP - 

For the regression Analysis I wanted to confirm the Hypothesis that there is a linear relationship between the Price and Year. I used the Bivariate distribution plot to show the relationship between the variables selected – in both datasets I used the equivalent Price and Year.

The EHPI data shows a linear regression between Year & Price index with a slope of 3.24, meaning that there is an expected increase. The correlation coefficient (r) of 0.6037 shows a moderate positive linear relationship between the two variables, and the p-value of 0 shows that the relationship is statistically significant. There was a standard error of 0.1032 which shows the degree of uncertainty around the estimated regression line.

The DHP data also shows a linear regression between Year & Price index, which alos has an expected positive increase. The correlation coefficient (r) however is smaller at 0.1767, showing a weaker positive linear relationship between the two variables. The p-value is again very low at 0.0000, indicating a statistically significant relationship. The standard error of 466.8038 is much larger than in the EHPI data, which suggests greater uncertainty around the estimated regression line.

In terms of comparison for the Pearson correlation coefficient, both sets of results have positive correlation coefficients. That said, the EHPI results has a higher correlation coefficient than the DHP, indicating a stronger linear relationship. Overall, both datasets have a positive linear regression and Pearson correlation coefficient.

Lastly, I experimented with the Chi-Squared test. To test the alternative hypothesis that there is significant association between two categorical variables – in this case Location and Price. Each dataset had very large the chi-squared test statistic –

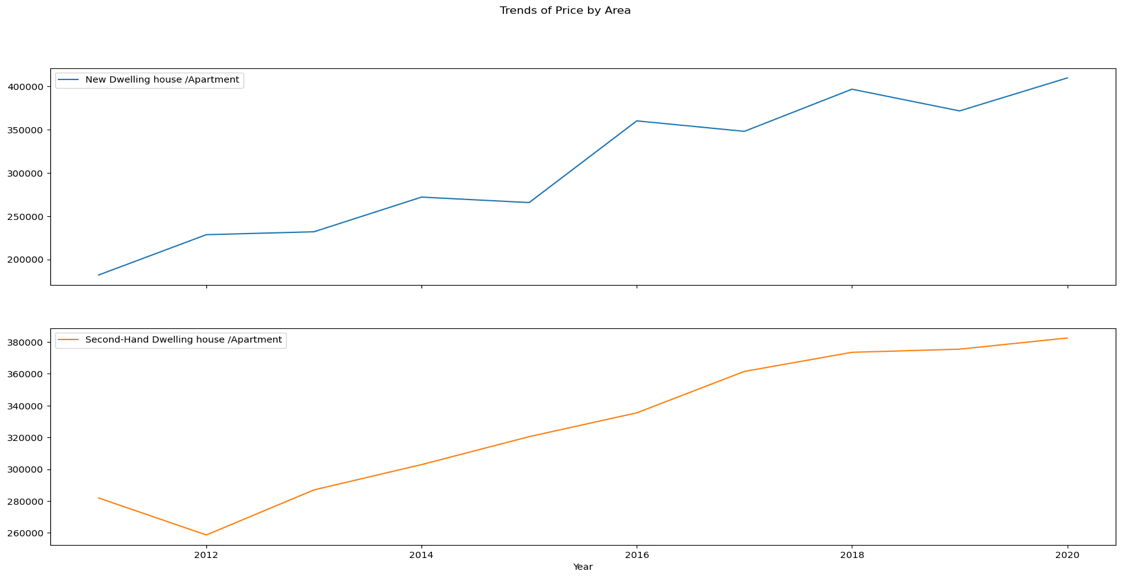
* EHPI = 38462.8429 with a p-value of 0.0208
* DHP = 208204.5440 with a p-value of 0

The high chi-squared test statistic for both shows a significant association between price and location. That said, I was surprised to see EHPI p-value was not as small as the DHP p-value, indicating a slightly weaker association between price and location compared to the Dublin data. However since the p-value is still less than the significance level of 0.05, I can reject the null hypothesis and conclude there is significant association between the chosen variables in both datasets - the price of a house is not independent of its location. Yes, I know its stating the obvious again.

In terms of the outcome of this analysis, I definitely understand the different tests I used a lot better and how I can use them in future research to better understand my data or target my variables to provide better results. I can definitely see me using more advanced statistical methods such as multiple regression or time series analysis to analyse the relationships between variables in my data.

Conclusions from the data itself, I can see two points of interest that would warrant future study for me –

1. There is a clear expected upward trend for both Dublin and EU Price points over the last 10 years.
2. The New Dwelling market price while still increasing showed fluctuations with a year of significant growth followed by a slight dip then growth again. This was unexpected, I would expected the curve to follow similar even though it eventually outperforms the second hand home performance.



Some challenges I experienced were around selecting the appropriate statistical method or test and selecting an appropriate level of significance for hypothesis testing. More research and study will be needed to fully integrate these methods into my day to day approach to Data Analysis.

**Machine Learning**

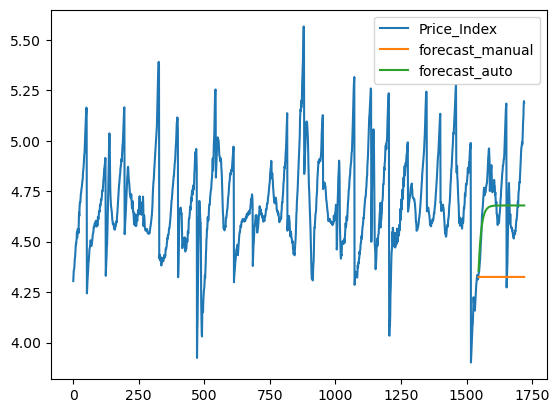
Now my datasets are cleaned, prepped and analysed, I am going to work on predicting changes in the EU datasets house price index. For this, I have chosen two machine learning models –

* ARIMA - Autoregressive Integrated Moving Average.

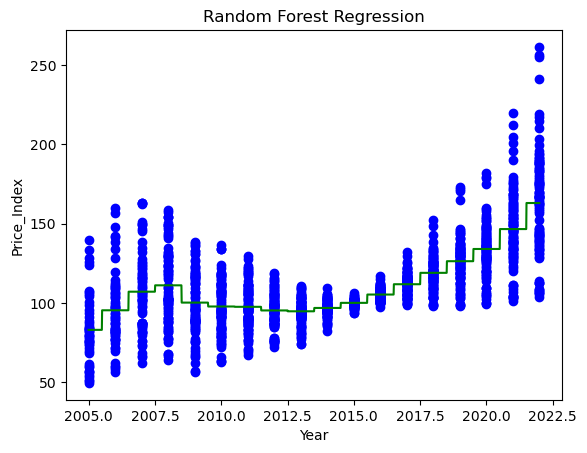
This statistical model is used for analysing and forecasting trends and patterns in time series data. Since house prices tend to exhibit seasonal patterns and trends, I chose ARIMA is an appropriate choice for this problem. Additionally, using AutoARIMA allows for automatic selection of the best ARIMA model parameters, which simplifies the model selection process.

* Random Forest Regression is a supervised learning algorithm used to predict a continuous variable based on a set of features. In this scenario, the target variable is "Price\_Index", and the features are "Year" and " Price\_Index ".

ARIMA Forecast Visual –



Random Forest Prediction visual –



In summary, the choice of ARIMA Time Series forecasting and Random Forest Regression models is justified based on the goal to predict future values of a continuous variable based on past observations using a set of features.

As I was using the ARIMA and Random Forest Regression models, I opted to not use GridsearchCV specifically for the following reason –

* GridsearchCV is designed and most useful for dealing with models that would be difficult to find the optimal combination of hyperparameters or Features that need to be tuned.
* Since my models number of features were so small, it was easy to determine the best combination through trial and error, without the need for an extensive grid search.

That said, the Auto ARIMA could be considered a type of Grid Search as it uses a specific search algorithm to explore and select the hyperparameters. GridSearchCV uses a grid of predefined hyperparameters to be tested exhaustively so their aim is similar but then differ in the approach.

Model Performance and Evaluation table -

|  |  |  |  |
| --- | --- | --- | --- |
| ARIMA - Time Series - | Mean Squared Error | Mean Absolute Error | Mean Absolute Percentage Error |
| - Manual Selected | 0.444704527 | 0.403347512 | 0.083864183 |
| - AutoArima | 0.185387875 | 0.144123523 | 0.029941222 |
| - TSCV - 5 Fold | 944.140311 |  |  |
|  |  |  |  |
| Random Forest Regression - | Mean Squared Error |  |  |
| - Year and Price included | 0.223545693 |  |  |
| - Only Price included | 321.0046885 |  |  |
| - K-Fold (5) Cross Validation | 308.4214032 |  |  |

Starting off with the ARIMA model, I used three metrics to assess its performance. From research I did and a useful article I found explained it was good option for this model. That said, the mean squared error (MSE) will be the primary evaluation metric, the MSE measures the average of the squares of the errors or deviations between predicted values and actual values.

The manual selected ARIMA (M-ARIMA) model had an MSE of 0.44, indicating there is a moderate difference between predicted and actual values. On the other hand, the auto-ARIMA (A-ARIMA) model, the MSE is significantly lower at 0.18, suggesting this model was better at predicting the target variable.

The mean absolute error (MAE) measures the average absolute differences between predicted and actual values. The M-ARIMA model MAE was 0.40, showing the predictions by an average of 0.40 units. The A-ARIMA model performed better, with a lower MAE of 0.14.

The mean absolute percentage error (MAPE) is a percentage-based evaluation metric that measures the average absolute percentage difference between predicted and actual values. The M-ARIMA model showed the predictions deviated from actual values by an average of 8%, whereas the A-ARIMA model performed better, with a lower MAPE of 3%.

The Cross-Validation was used to assess the model's performance on unseen data. In the case of the ARIMA model a 5-fold Time Series Cross Validation was used which gave a MSE is 944.14. That is significantly higher than the MSE of both the M-ARIMA and A-ARIMA models.

Conclusion - The ARIMA model may not generalize well on new data.

The Random Forest Regression (RFR) model, only MSE is used to evaluate this model. I have however generated the MSE twice, once on one features (Price) and a second with two features (Price & Year). In the case of the Random Forest Regression model with both year and price included, MSE was 0.22, indicating the model good at predicting the target variable. Although, when only price was the feature, the MSE jumps to 321.00, this suggests that the year feature plays a significant role in improving model performance.

When I used a 5-fold cross-validation MSE for the RFR model the MSE was 308.42, which was surprising in that it was quite close to the MSE for the RFR model with both features included. This suggests there was consistent performance is across the two for different subsets of the data.

Overall, both models show promise in predicting the target variable.

That said my conclusions are this –

The ARIMA models may not generalise well on new data even based on the MSE from TSCV, even with the A-ARIMA model performing better than the M-ARIMA model. Therefore I would be more inclined to use the RFR based on the performance I got as including the two features improved the model's performance plus based on the cross-validation MSE the model seems to be more consistent across different subsets of the data

**Sentiment Analysis**

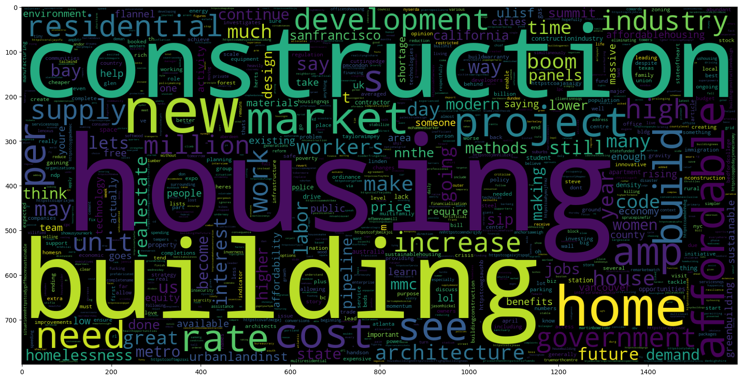
For my Sentiment Analysis I kept this one simple, I attempted to pull 100 based on the restrictions Twitter recently imposed.

My criteria for the tweets was – “housing, building & construction” and I got 98 tweets back. The one caveat I have to add is the data is global and not specific to the EU. The reason for this was the volume of tweets over the 7 days was quite low 6 or 7 depending, this wasn’t enough without including retweet to get any meaningful insights from.

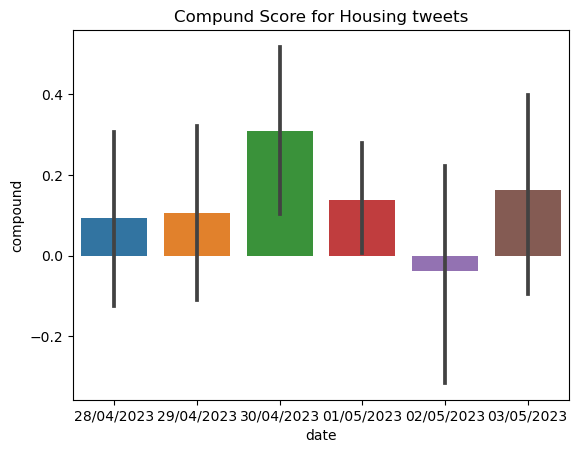
What I found –

* Sunday (30/04) was the most positive day
* Tuesday (02/05) was the most negative day
* The average sentiment for the 6 days was 0.12 which is “Slightly Positive”

Word Cloud –



Compound score for tweet sentiment by day –



Sentiment split by group

