2023

BDA105 DATA MINING AND MACHINE LEARNING CA1

BERNADETTE O’GRADY 40030457

Contents

1) Problem Selection

My chosen problem is predicting future house prices in Dublin, a topic that I am genuinely interested in. Dublin is the capital city of Ireland and has a rapidly growing housing market due to its population growth. According to estimates from April 2021, the city's population is 1.43 million persons, which is 28.5% of the total population, and it's expected to reach 2.2 million by 2031.

Predicting house prices accurately can benefit different stakeholders, including home buyers, sellers, and real estate agents. For instance, buyers can use the predictions to make informed decisions about purchasing properties, while sellers can determine the optimal selling price. Also, real estate agents can provide accurate pricing information to their clients, and investors can use the predictions to make informed decisions about investing in the Dublin housing market.

The project has the potential to provide valuable insights into the Dublin housing market and can be a useful tool for various stakeholders involved in the market.

Dataset Selection:

I chose to get my data from daft.ie as it is considered the most popular and widely used platform in Ireland. It also has the features I am looking for and the data is current. The main features are price, location, size of property, number of bedrooms, number of bathrooms, type of property.

2) Understanding the nature of the problem

The main issue is that property prices in Dublin have become unaffordable for most people due to the high demand exceeding the supply of properties. The root cause of this problem is attributed to the city's population growth, which has been fuelled by immigration in recent years. Additionally, there has been limited new construction in the city, which has further compounded the issue.

To address this problem, accurately predicting future house prices in Dublin is crucial. However, this task is challenging as several factors influence the market value of residential properties in the city, such as the property's size, location, number of bedrooms, age, and other features.

Data mining can help solve this problem by using statistical and machine learning techniques to analyse historical data on property sales, along with data on various property features and market trends. By identifying patterns and relationships within the data, data mining can be used to create predictive models that can estimate the market value of a property based on its features.

From a business perspective, various stakeholders would be interested in the project of predicting future house prices in Dublin. These stakeholders may include real estate agents, property developers, homeowners, and potential homebuyers. By leveraging the predictive model developed through data mining, real estate agents and property developers can determine the optimal selling price for a property, while homeowners can estimate the value of their property. Potential homebuyers can use the model to determine if a property is priced fairly, and investors can use it to make informed decisions about investing in the Dublin housing market.

The aim of this project is to create a precise and reliable predictive model that can estimate the market value of a residential property in Dublin based on a variety of features. The developed solution should be user-friendly and provide dependable and accurate estimates that are reflective of current market trends. By offering stakeholders precise pricing information, the project endeavours to support them in making informed decisions and reaching their business objectives.

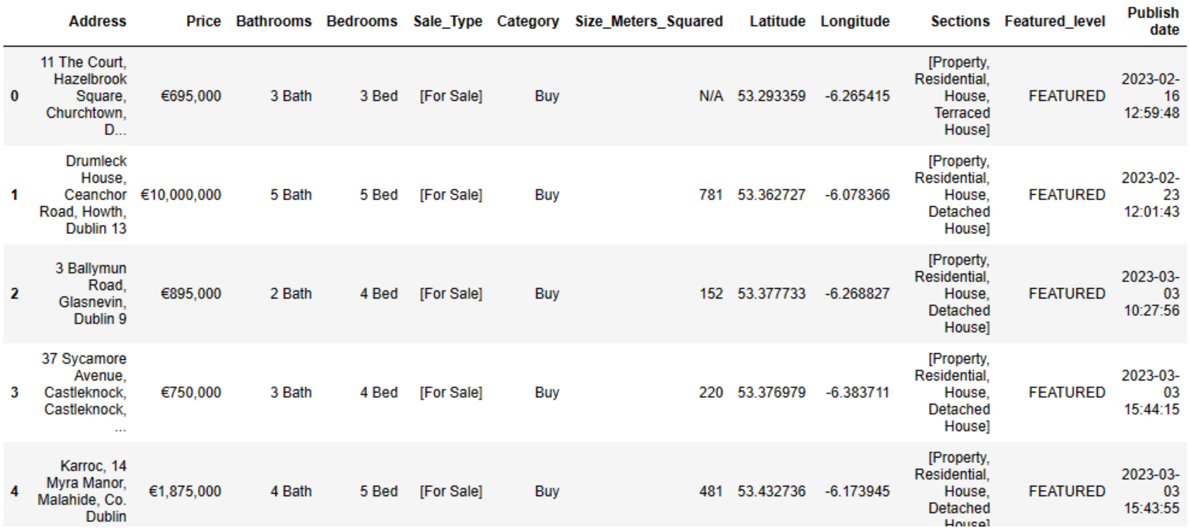
3) Understanding the data

1. Collect initial data: Acquire the necessary dataset and (if necessary) load it into your analysis/programming (Python/R/Weka or other) tool.

I am getting my data from Daft.ie website for Dublin Residential Properties on Sale. The daftlistings package (Bloomer, 2023) only collects the last 6 months of data so my project is limited to this. There are limited features to select from daftlistings package and I have included all the relevant features for this dataset. However, I do think there is enough data for machine learning model and it is an exercise worth doing.

2. Describe data: Examine the data and document its surface properties like data format, number of records, or field identities.

The initial DataFrame has **3690** rows and **12** columns. Latitude and Longitude are float64 and the remaining fields are objects. The field identities are shown in the following figure.



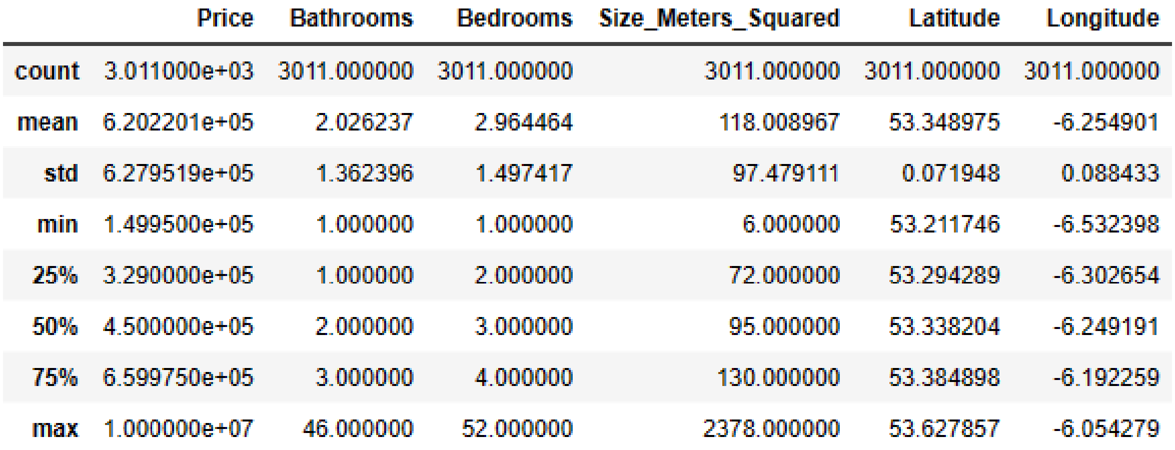
	Address	Price	Bathrooms	Bedrooms	Sale_Type	Category	Size_Meters_Squared	Latitude	Longitude	Sections	Featured_level	Publish date
0	11 The Court, Hazelbrook Square, Churchtown, D...	€695,000	3 Bath	3 Bed	[For Sale]	Buy	N/A	53.293359	-6.265415	[Property, Residential, House, Terraced House]	FEATURED	2023-02-16 12:59:48
1	Drumleck House, Ceanchor Road, Howth, Dublin 13	€10,000,000	5 Bath	5 Bed	[For Sale]	Buy	781	53.362727	-6.078366	[Property, Residential, House, Detached House]	FEATURED	2023-02-23 12:01:43
2	3 Ballymun Road, Glasnevin, Dublin 9	€895,000	2 Bath	4 Bed	[For Sale]	Buy	152	53.377733	-6.268827	[Property, Residential, House, Detached House]	FEATURED	2023-03-03 10:27:56
3	37 Sycamore Avenue, Castleknock, Castleknock, ...	€750,000	3 Bath	4 Bed	[For Sale]	Buy	220	53.376979	-6.383711	[Property, Residential, House, Detached House]	FEATURED	2023-03-03 15:44:15
4	Karroc, 14 Myra Manor, Malahide, Co. Dublin	€1,875,000	4 Bath	5 Bed	[For Sale]	Buy	481	53.432736	-6.173945	[Property, Residential, House, Detached House]	FEATURED	2023-03-03 15:43:55

Figure 1 First 5 rows of Dublin Property dataset

3. Explore data: Dig deeper into the dataset. Query it, visualize it, and identify relationships among the data variables (attributes or columns).

To query the data, I am reformatting the fields as the majority are objects. I reformatted the following columns to numeric fields in order to do boxplots and histograms on them.

- Price, Bathrooms, Bedrooms, Size Meters Squared.



	Price	Bathrooms	Bedrooms	Size_Meters_Squared	Latitude	Longitude
count	3.011000e+03	3011.000000	3011.000000	3011.000000	3011.000000	3011.000000
mean	6.202201e+05	2.026237	2.964464	118.008967	53.348975	-6.254901
std	6.279519e+05	1.362396	1.497417	97.479111	0.071948	0.088433
min	1.499500e+05	1.000000	1.000000	6.000000	53.211746	-6.532398
25%	3.290000e+05	1.000000	2.000000	72.000000	53.294289	-6.302654
50%	4.500000e+05	2.000000	3.000000	95.000000	53.338204	-6.249191
75%	6.599750e+05	3.000000	4.000000	130.000000	53.384898	-6.192259
max	1.000000e+07	46.000000	52.000000	2378.000000	53.627857	-6.054279

Figure 2 Describes Statistics Data on Numeric Fields

The scatterplot of latitude and longitude shows the geographic distribution of the properties in Dublin. It shows the areas with high concentrations of properties, and the properties that are geographically isolated.

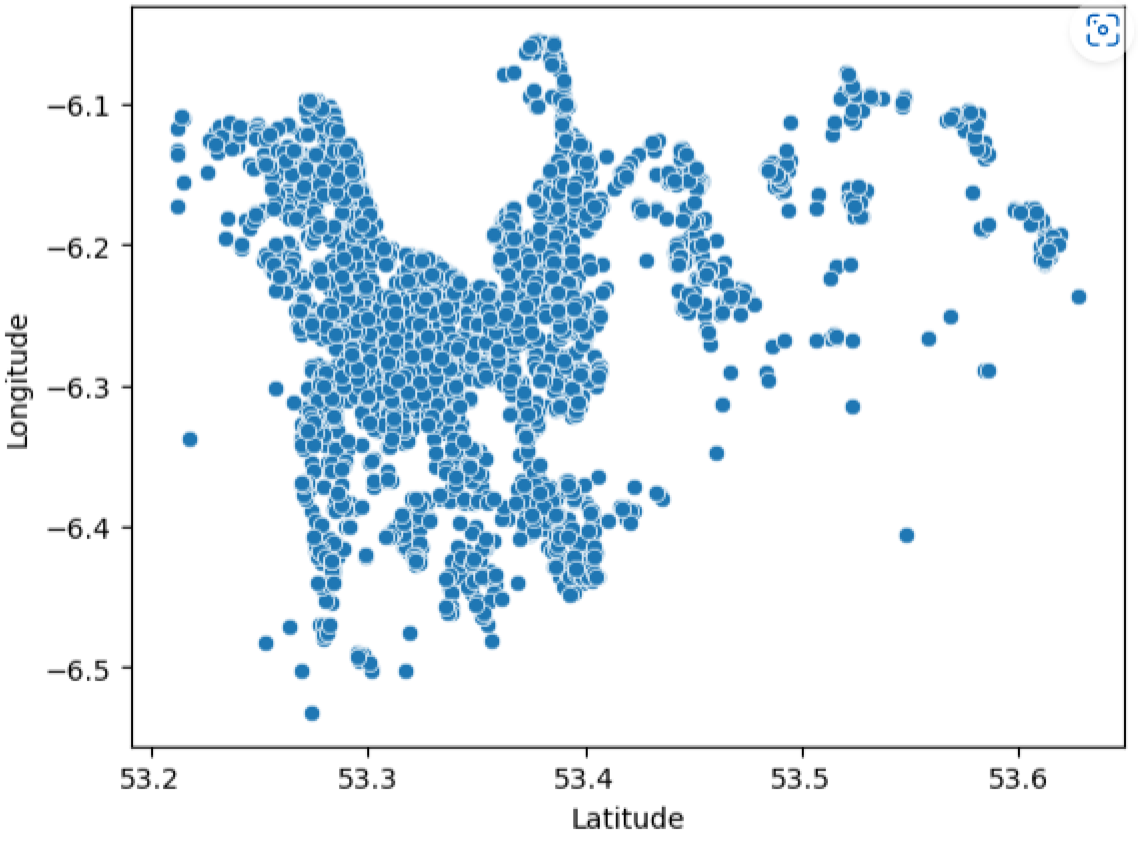


Figure 3 Scatterplot for Latitude and Longitude of Dublin Properties for Sale

I converted Sections to String type and renamed the data in Sections column to more simpler data in order to do a Countplot on type of property for sale shown in the following figure.

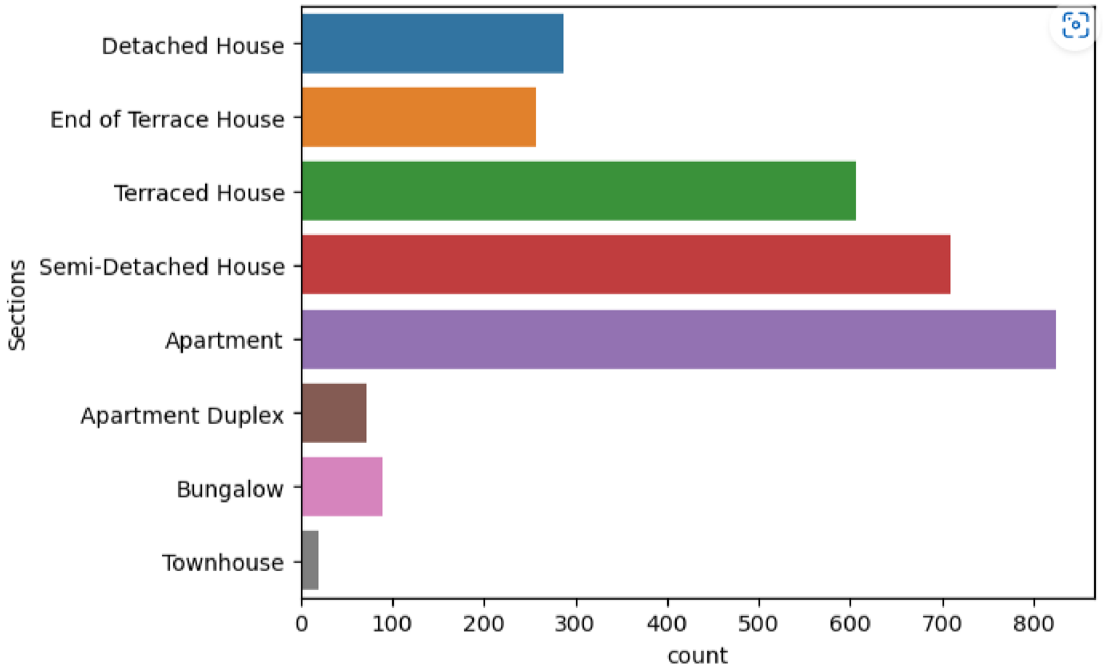


Figure 4 Countplot on the type of property for sale in Dublin

The Boxplot for Price, Bedrooms, Bathrooms and Size Meter Squared shows that there are outliers. I am creating a new boundary for outliers using this formula

- New Boundary = third quartile + 3*IQR
- Num Outliers = count of field name > new boundary

Table 1 Boxplot data

FIELD NAME	UPPER FENCE	Q3	MEDIAN	Q1	MIN	NEW BOUNDARY	NUM OUTLIERS
PRICE	1.15M	660k	450k	329k	149.95k	1652950.0	146
BEDROOMS	7	4	3	2	1	10.0	3
BATHROOMS	6	3	2	1	-	9.0	3
SIZE METER SQUARED	214	130	95	72	6	304	102

I have decided to cap the outliers to the new boundary value for all the fields listed in the table and keep the rows as I want my machine learning model to be realistic and not biased.

- Histogram for Price is right skewed in that the tail of the distribution extends to the right, meaning that there are a few extreme values that are much larger than the majority of the values. This results in a histogram that has a long tail to the right and a peak that is shifted to the left. In a right-skewed distribution, the mean is typically greater than the median which in this case is true as the mean is **6202201** and the median is **450000**.
- Histogram for Bedrooms is showing the central tendency of the data to be **3** bedrooms per property.
- Histogram for Bathrooms is showing the central tendency of the data to be **2** bathrooms per property.

- Histogram for Size Meters Squared is showing the central tendency of the data to be on the lower side approx. from **95** square meters (Median) to **118** square meters (Mean).

4. Verify data quality: How clean/dirty is the data? Document any quality issues (missing or duplicate data).

There is a fair amount of cleaning to be done on the dataset. The majority of the fields are object type so I have converted Price, Bathrooms, Bedrooms, Size meters Squared to numeric type.

There are **76** rows with Price on Application for column Price. There are **53** rows with Blanks for column Bathrooms. There are **597** rows with N/A for column Size Meters Squared.

I check for duplicates using Address Field. There are **16** duplicates.

4) Preparing data for machine learning models

1. Data selection: Determine which datasets (or if you are using a single dataset which variables/columns) will be used and document reasons for inclusion/exclusion.

I am using a single dataset derived from last 6 months on daftlistings for Residential Sales in Dublin.

I am going to drop the following columns

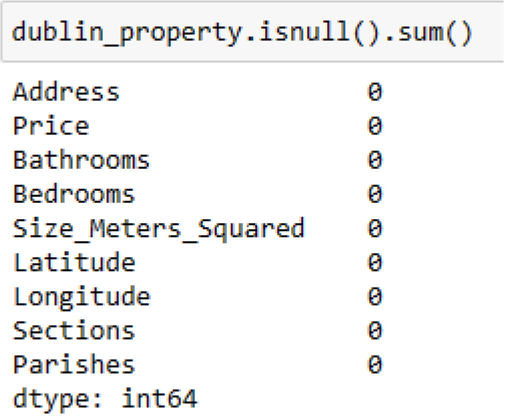
1. Sale_type as they are all for Sale so not required for further Analysis
2. Category as they are all Buy except for 1 (New Home) so not required for further Analysis
3. Publish Date not required as all houses for sale are last 6 months
4. Featured_Level is not required either as it is just used for the type of advertisement on Daft and I can see Standard is used for the majority of ads
5. I will drop the Address column at a later stage I want to use it for verification on the new Parishes Column.

2. Data cleaning: This is the lengthiest task! Without it, you'll end up with "garbage-in, garbage-out". A common practice during this task is to correct, impute, or remove erroneous values from rows or columns.

The data cleaning involves the following steps

1. Converting these fields to numeric Price, Bathrooms, Bedrooms, Size Meters Squared
2. Dropping all the records that have Size_Meters_Squared = N/A and Price = Price on Application and Bathrooms = Blank
3. Removing all duplicate rows based on Address
4. Capping all outliers on Price, Bathrooms, Bedrooms and Size Meters Squared to the new Boundary value as outlined in 3.3
5. Converting Sections to String type and renaming the data in Sections column to more simpler data as outlined in 3.3

The end result there will be a clean dataset with no isnull data and no extreme outliers.

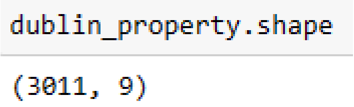


```
dublin_property.isnull().sum()
```

```
Address                0
Price                  0
Bathrooms              0
Bedrooms               0
Size_Meters_Squared    0
Latitude               0
Longitude              0
Sections               0
Parishes               0
dtype: int64
```

Figure 5 Checking for presence of isnull on the data

The shape of the data will have



```
dublin_property.shape
```

```
(3011, 9)
```

Figure 6 Shape of Dataset

3. Data construction: Derive new attributes that will be helpful for the modelling process.

I am creating a new Feature called Parishes which are the Parishes of Dublin. There are **166** unique Parishes. This is created using the Latitude and Longitude values. This new feature enables me to show a Map of Dublin with the Average House Price with the darker colours showing the most expensive area for properties and the lighter showing the least expensive. (LYU, 2021). I used this citation to help with the code to generate the Map.

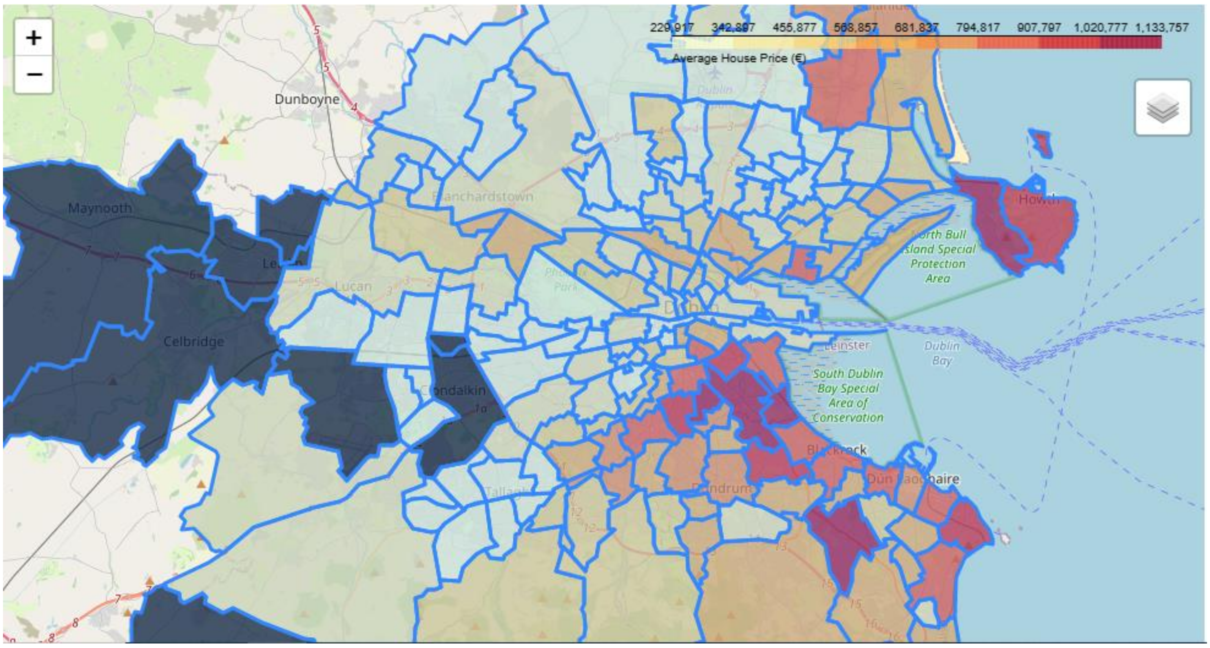


Figure 7 Map of Dublin showing Average House Price

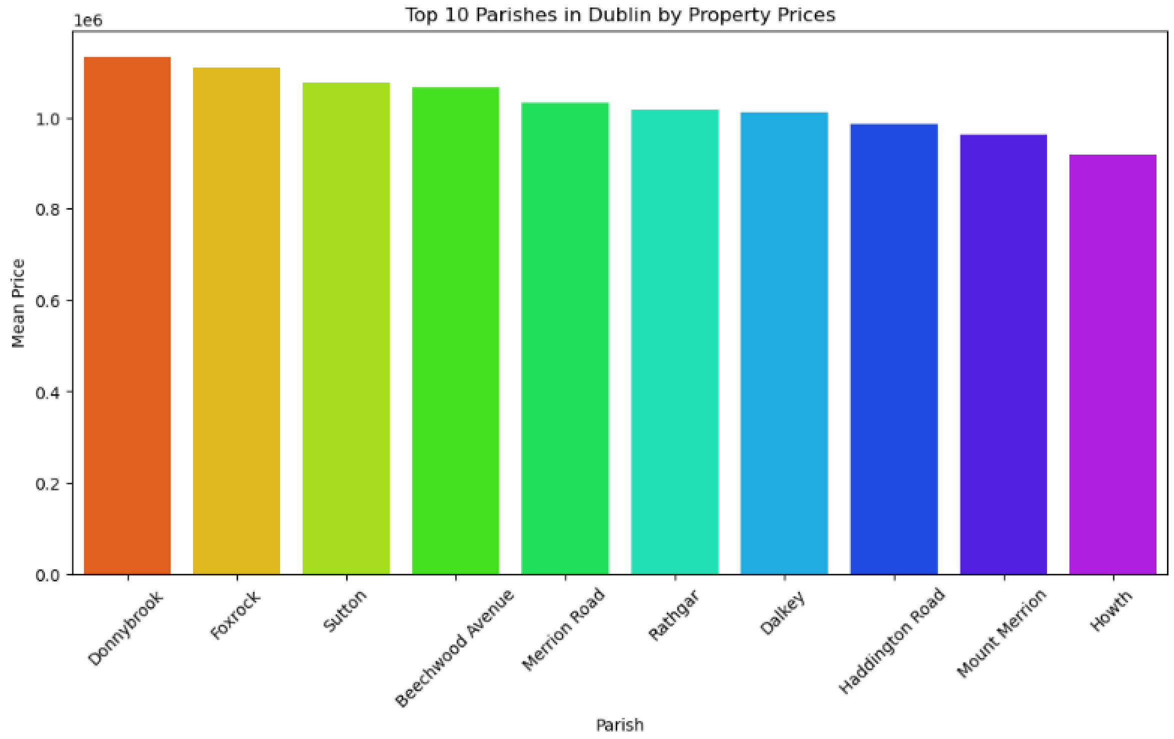


Figure 8 Top 10 Most expensive places to live in Dublin based on average house prices

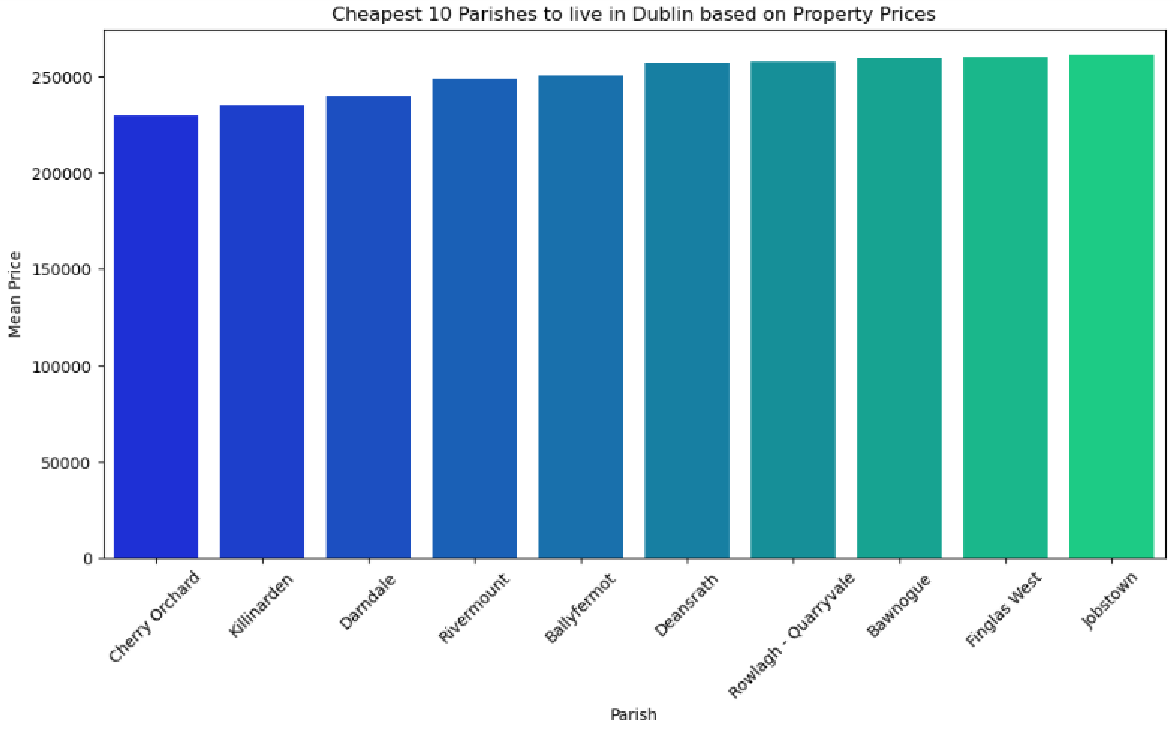


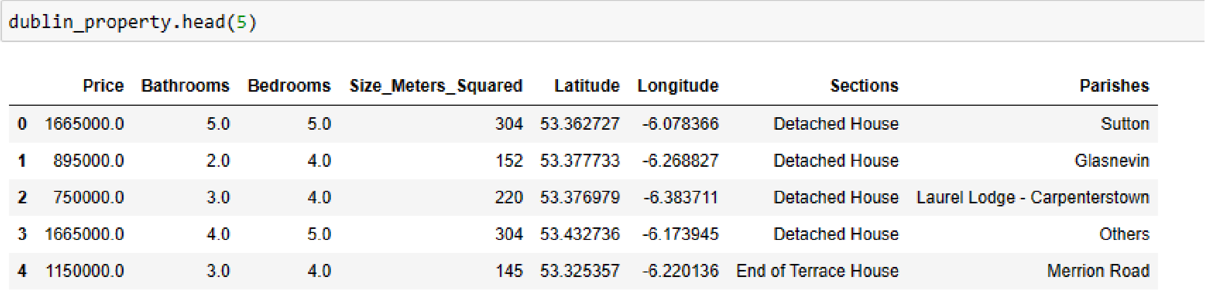
Figure 9 The 10 cheapest places to live in Dublin based on average house prices

4. Data integration: Create new data sets by combining data from multiple sources (if applicable)

Not applicable.

5. Data formatting and encoding

The following figure details the first 5 rows of the final dataset before One Hot Encoding is applied. I have dropped column Address as it is not required for the machine learning model. The Parishes column with Latitude and Longitude covers the address.



```
dublin_property.head(5)
```

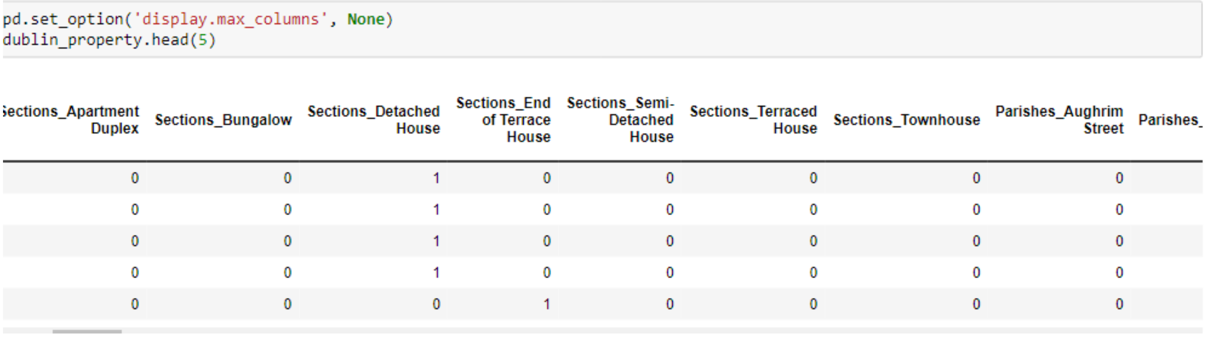
	Price	Bathrooms	Bedrooms	Size_Meters_Squared	Latitude	Longitude	Sections	Parishes
0	1665000.0	5.0	5.0	304	53.362727	-6.078366	Detached House	Sutton
1	895000.0	2.0	4.0	152	53.377733	-6.268827	Detached House	Glasnevin
2	750000.0	3.0	4.0	220	53.376979	-6.383711	Detached House	Laurel Lodge - Carpenterstown
3	1665000.0	4.0	5.0	304	53.432736	-6.173945	Detached House	Others
4	1150000.0	3.0	4.0	145	53.325357	-6.220136	End of Terrace House	Merrion Road

Figure 10 First 5 rows pre-One Hot Encoding

I will apply One Hot Encoding to the two categorical columns Sections and Parishes. There are **166** unique Parishes. Parishes with less than 10 properties will be converted to Others before One Hot Encoding.

The following figure details the format of the dataset after One Hot Encoding is applied with each column having numeric data.

The dataset is now primed for the next phase of the CRISP - DM Model.



```
pd.set_option('display.max_columns', None)
dublin_property.head(5)
```

ections_Apartment Duplex	Sections_Bungalow	Sections_Detached House	Sections_End of Terrace House	Sections_Semi-Detached House	Sections_Terraced House	Sections_Townhouse	Parishes_Aughrim Street	Parishes_
0	0	1	0	0	0	0	0	
0	0	1	0	0	0	0	0	
0	0	1	0	0	0	0	0	
0	0	1	0	0	0	0	0	
0	0	0	1	0	0	0	0	

Figure 11 First 5 Rows showing One Hot Encoding on the dataset

The shape of the data is (3011 Rows and 140 Columns after One Hot Encoding)

Bibliography

Bloomer, A., 2023. *Anthony Bloomer Daft Listings.* [Online]
Available at: https://github.com/AnthonyBloomer/daftlistings
[Accessed 27 February 2023].

LYU, G., 2021. *Dublin Rental: EDA.* [Online]
Available at: https://www.kaggle.com/code/d17129765/dublin-rental-eda#3.Maps
[Accessed 27 February 2023].