FINAL REPORT – IRISH RENT ANALYTICS & PREDICTION

Student Name: ARPIT JOSHUA ELIAS

Student ID: 24257567

Module: Data Analytics for Artificial Intelligence

Assessment: CA2 – Data Analytics Project

1. Abstract

This is my Data Analytics for Artificial Intelligence final project in which I chose rental prices in Ireland and what actually moves the market and how it happens. I took the real dataset from the Residential Tenancies Board that is the RTB. In my project I will be cleaning the Raw data taken from the dataset from the RTB then finding patterns that are in it and through that I will be creating features.

I will be trying 2 feature selection methods and then will create a model that can predict rents in different areas of Ireland.
I used 3 models that is MPL model that is the Multi layer perceptron neural network, Random forest model and Linear Regression model.
on top of that I also used K means clustering to get the natural segmentation of the region of Ireland in the rental market and how it behaves.

I got clear signs of growth in the rental market in all of the Ireland, there were regions that had a very low rent patterns and there were a lot that had an incremental growth in the rental price market and showing non linear relationship between the features I used.

I am pretty happy with the work I have done as out of the 3 models I used the MLP model actually gave me the highest accuracy and I got the R^2 of 0.79 that is 79% accuracy, later in the report I have created a table as well to compare the three models and explain what all differences were found and how MLP came on top and performed better than the Forest model and the MLP. In the clustering as well Ireland got divided into clear two major groups that is the Dublin and the rest of the regions of Ireland, In Dublin I found the rents to be increasing in an incremental pace while the other group which is having the rest part of the Ireland are increasing at a much lower pace in terms of the rent as compared to Dublin

So as a whole I learnt a lot from this project firstly how to write down proper code do cleanup, Visualizations, Modelling etc. leading to creating a model which is data driven and can predict residential rents pretty accurately in all the parts of Ireland as taken from the dataset.

2. Introduction

As being a student who recently moved to Ireland living in a single room knows how rent prices dents and takes a major chunks for your savings/earnings if working so this has become a very key economic concern in Ireland affecting everybody from students to working professional so Understanding how the rents fluctuate and what are the patterns of rent considering the

property types in different counties of Ireland is crucial and can help students to plan their stay in advance considering the prices, so being a student and tackling the problem myself my goal and aim of this project was to study and carry out the data analytics to study the Irish rental market trends how do they work and fluctuate and create a model for estimating a monthly rental value in Euros.

So for this project as I mentioned I took a dataset from Central Statistics Office and took the RTB Average Monthly Rent Dataset which is a raw dataset from the year 2008 to the year 2024.
I did the cleaning of the values that were missing in the dataset as the dataset was raw used the feature selection, found many outliers and did the feature engineering as well leading to building a full working ML model

So that report basically guides what I did in the project how I tackled different problems which after tries led me to creating a model that is able to predict the Ireland's rental market.

3. Dataset Description

So, about the dataset, the dataset is taken from the Central Statistics Office that is the CSO and took the RTB average monthly rent series. The dataset is raw and it covers multiple counties and have the data from 2008-2024, the dataset is having columns like Currency, Average rent value, Property type, Year, Location, Report statistic codes as well.

The dataset I took is bit huge within a total of 318,444 rows in it, as the data is raw so it is having missing rent values in many of the records so in the dataset there it is denoted with NaN.
After cleaning the dataset the dataset got reduces to just 109,202 rows. Some of the columns were also just codes and with copied ID's so I removed them as well.

4. Data Cleaning

So for the cleaning of the data part in my project first I removed all the rows that was not showing any data and had NaN as value as the model without data cannot predict.
I also gave the columns proper names from 'Quarter' changed it to 'Year_Quarter' and from 'VALUE' to 'Rent' etc, it was just easier to keep track like this.
I took out the name of the county from the 'Area' and standardised it by trimming the 'Town'.
I also removed columns which had no useful data in it and was having the same RTB code values in every row so to trim and keep the dataset clean those were also removed.
I added a Year_Trend feature so as to show the fluctuation in the rental values, that is Year_Trent=Year it is just the year the data is taken as this makes it easier for created model to understand the year on year trend of the rental values understanding the patterns and then give out predictions accordingly. In the end during the data cleaning, I also made sure to fix all the data types that is making sure that the counties were categories dates were proper and rents is in the numeric.

5. Feature Engineering

The cleaning was not enough so next I did feature engineering on the dataset.

5.1 Year_Trend

This shows how many years have passed as this helps the model in understanding the gradual increase in the renal growth in different parts of the Ireland

5.2 County Extraction and Cleaning

Standardised the different Irish county names.

5.3 County Clusters (from K-Means)

With the help of the K-means clustering I was able to add a new feature called the county_cluster, it gave out values in terms of 1 and 0 for each county if it is 1 then it meant that the county comes in a group which is growing at a rapid pace and had a high rental value the group included counties like Galway, Cork and yes Dublin as well, these were predictable as these are big cities and as it is everywhere the rents are high at these places, whereas if the value comes out to 0 that comes in a group that is growing at a slow pace and have a lower rent as compared to the other group these places usually included smaller towns and countryside places.

The result clearly distributed the counties of Ireland into 2 groups.

6. Exploratory Data Analysis (EDA)

6.1 Rent Distribution

In Figure – 1 the rent distribution is right skewed where Dublin having the most long tail of high rental values between €700–€1200 per month.

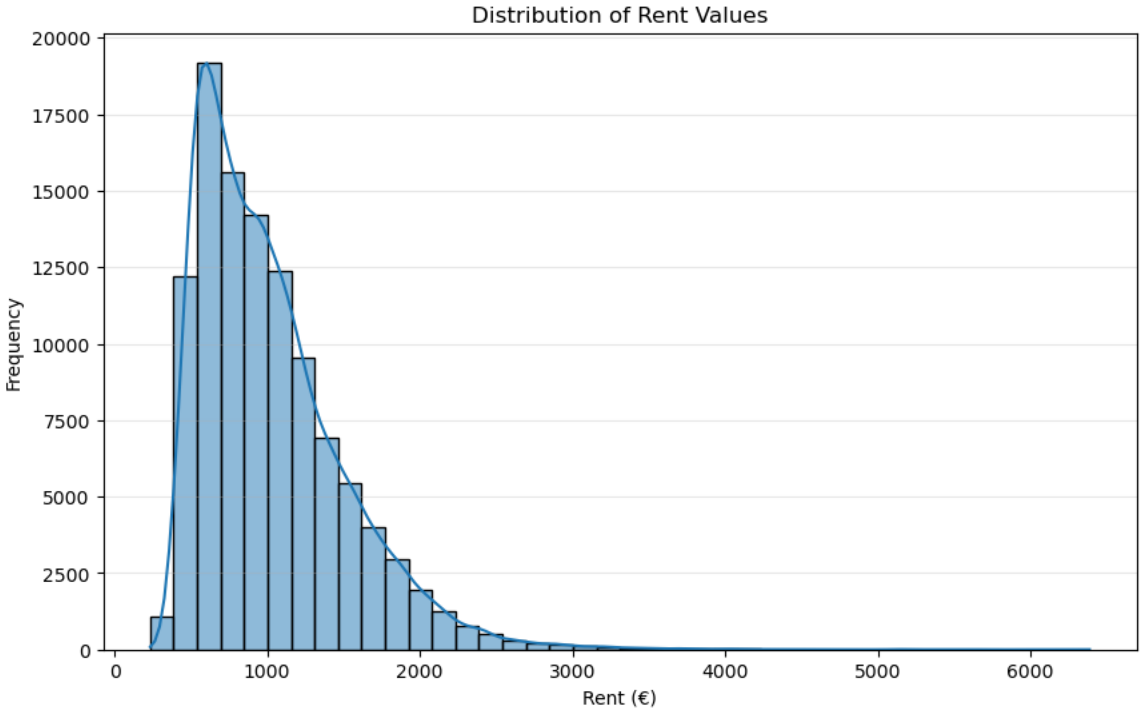

Figure – 1

6.2 Temporal Trends

The below line graph (Figure-2) shows that the by each passing year the monthly average rental value in Ireland has increased, and the sudden dip in the value from 2008 is because of the Global financial crisis in which the market crashed as with it the rental prices.

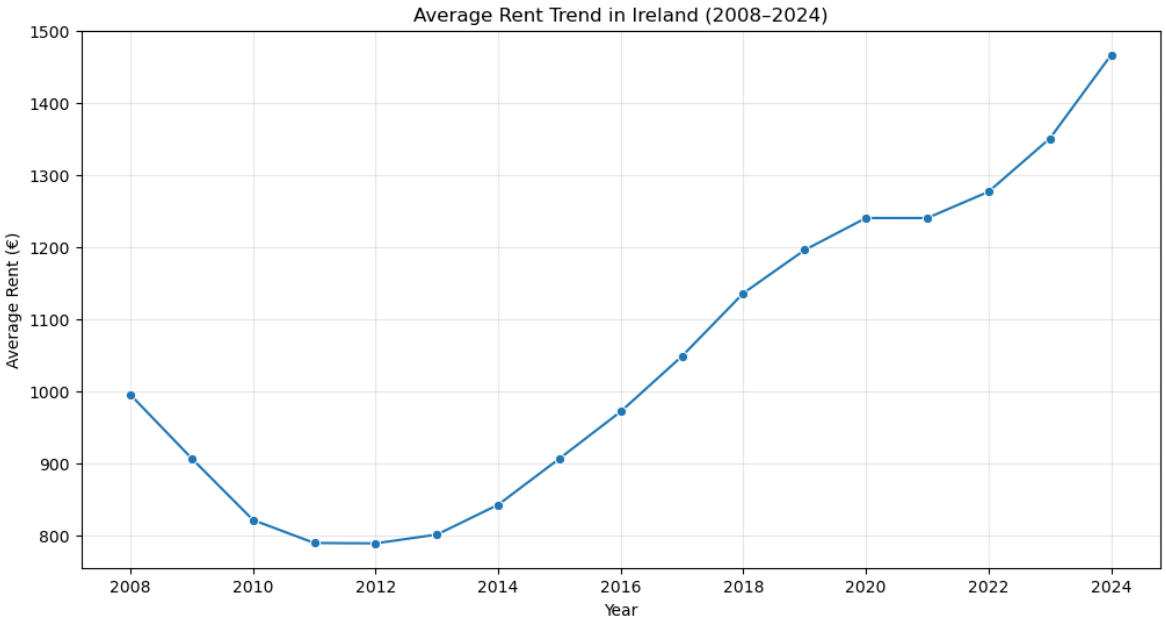

Figure - 2

6.3 County-Level Differences

As expected as we can see in the Figure-3 Dublin 18, Dublin 4 and Dublin 14 are way ahead in terms of rental value meanwhile that counties from the other group had a lower rental value, this clearly showed that how counties like Dublin are experiencing a high pressure and a demand for housing leading to increase in rents whereas the other counties due to low demand and less pressure the prices are not that high compared to the cities.

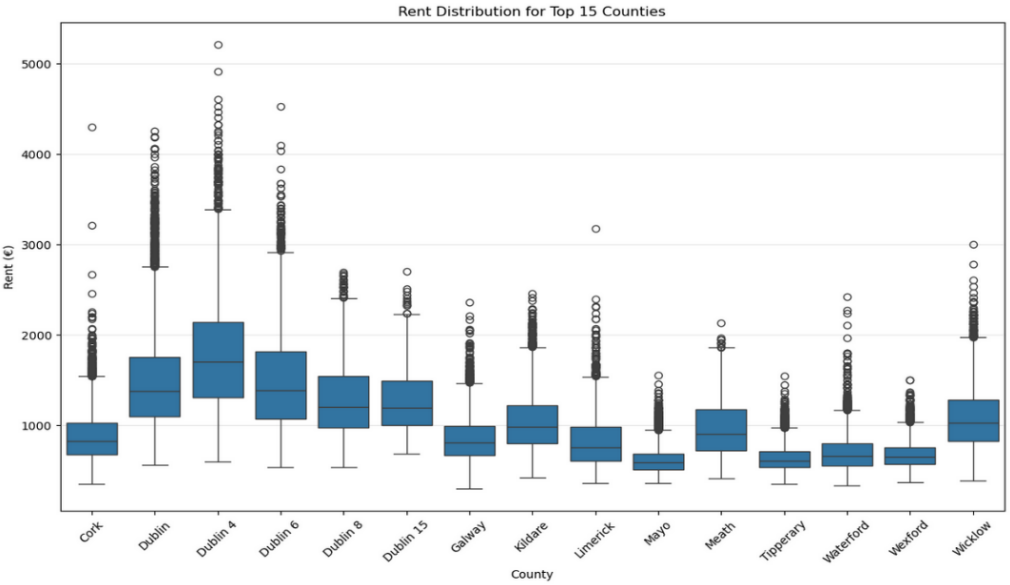

Figure- 3

6.4 Property-Type Differences

Figure-4 demonstrates property type differences and their patterns and through it we got to know that Detached houses had the highest rents overall and then moderate rents for the Apartments and lowest rents for other flats and unspecified property types.

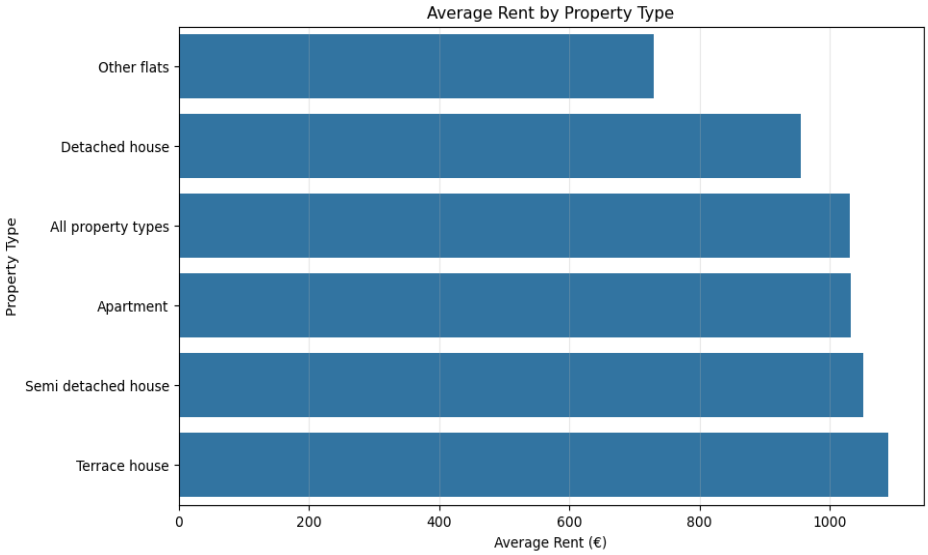

Figure - 4

6.5 Heatmap

Figure-5 shows the county heatmaps with clear geographical clustering, Dublin over the years remained consistently high on rent per area cost with steady upward growth.

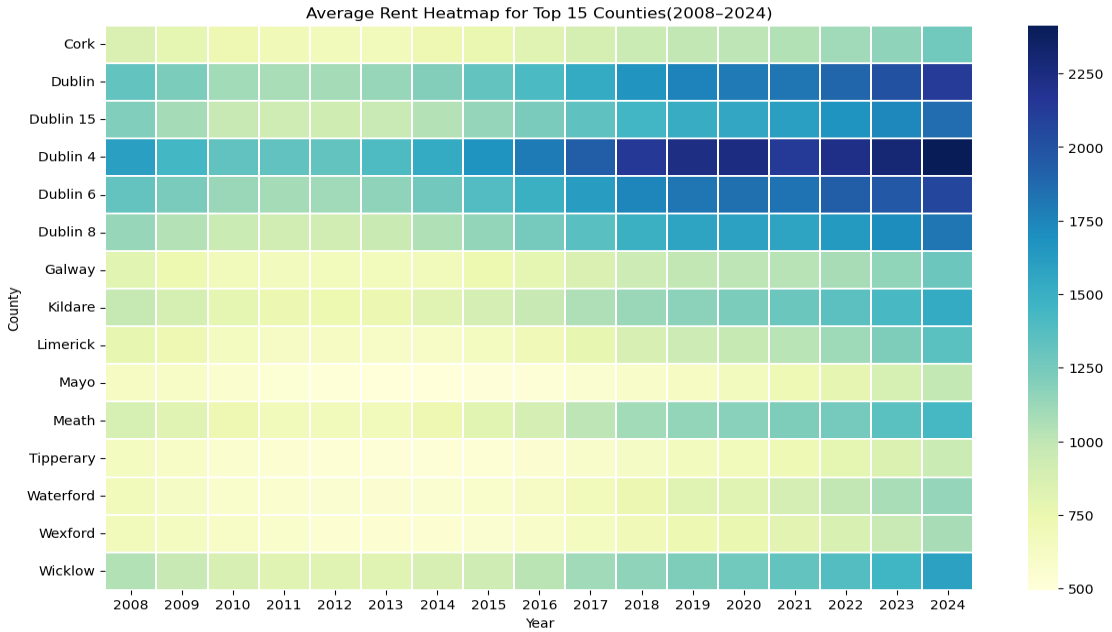

Figure – 5

7. Feature Selection

I used 2 different feature types.

7.1 LassoCV (Coefficient-Based)

Lasso features shrink less important coefficients to zero.
Through this we found out that Semi detached houses, Multiple Dublin districts and Year were the strongest predictors. Dublin dominated in predicting the rent.

7.2 Permutation Importance (Model-Agnostic)

With the help of the MLP model I got to know that which features the neural network relied on to work. The smaller county indicators were the one that impacted and played a major role in changing the predictions of the test set.
Using both I also realised that the permutation captured local sensitivity of the final model but the Lasso actually captured the global linear structure.

I was able to remove the high dimensional one hot encoded space by shrinking the lesser informative coefficients towards the zero mark with the help of the Lasso regression, Year played a major role as being a major predictor along with Dublin county and detached houses / semi detached house categories this all perfectly synced with the EDA findings.

On the other hand, with the help of the Permutation MLP model was able to provide a model agnostic view of feature contribution. Dublin sub counties being the top influential features followed by Year_Trend and selected property types. So this as a whole confirms that the rental price is driven by the time progression and geographical location.

8. Modelling Approach

In total I used 3 models.

8.1 Linear Regression (Baseline)

Keeping the baseline simple by using one hot encoded counties and property types.
Using this the result I got is $R^2 \approx 0.707$, it is good but limited by linearity.

8.2 Random Forest Regressor

10 fold cross validation is used. By using the Random Forest I got the MAE ≈ 242.57 with + and – of 1.94 in which the RMSE ≈ 320.27 with + and – of 3.96 along with $R^2 \approx 0.562$ with + and – of 0.010 so this model struggled with high dimensional one hot encoding.

8.3 MLP Neural Network (Best Model)

With the 10 fold CV the Architecture is 64 and 32 hidden units, Adam optimiser and ReLU activation. I got MAE ≈ 142.89 with + and – 3.51, RMSE ≈ 221.53 with + and – 6.60 and $R^2 \approx 0.791$ with + and – 0.008

This model performed the strongest and retained on all 109,202 rows and then was saved as a reusable pipeline, then to select the final MLP configuration I used a small hyper parameter in which I evaluated 3 architecture that is (32,16) (64,32) and (128,64) each with there Adam optimiser and Relu activation. Validation experiment gave out that (64,32) network provided the best balance between the training time of it and the accuracy. I reduced the step size by changing the learning schedule to adaptive which allowed the optimiser to reduce resulting in automatically reducing the step size when progress was slowed. Convergence was seen after 400 iterations.

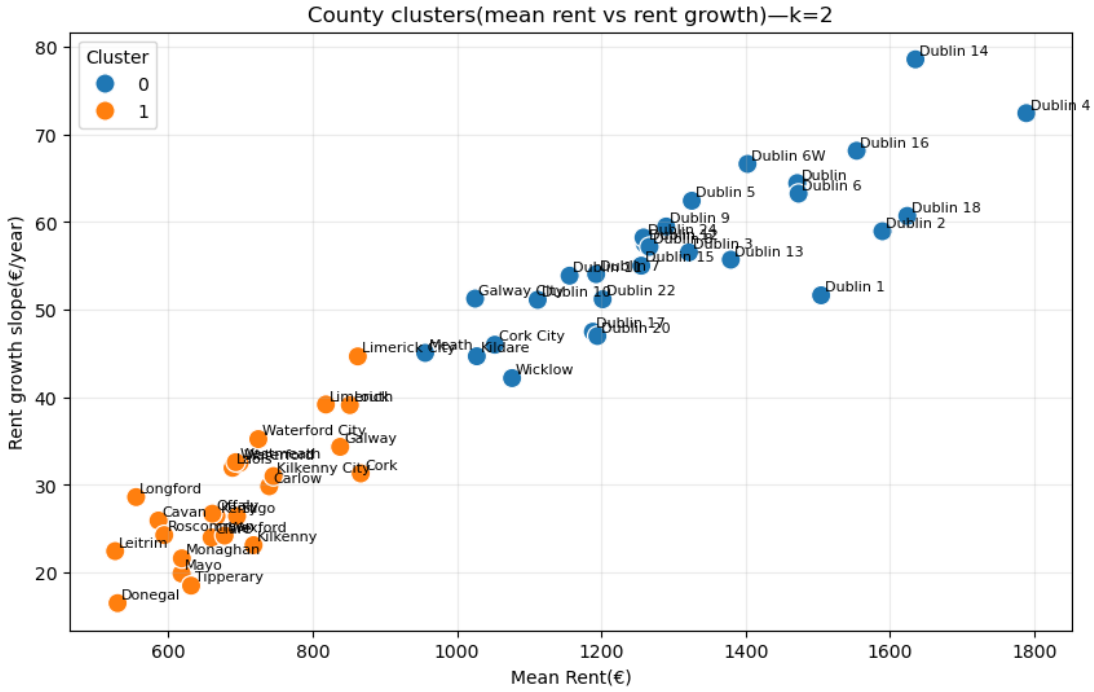

Figure – 6

Figure-6 shows that our predictions align very closely to the actual values and most points even lie very near to the diagonal line as well and due to the right skewed distribution as minor spread indicated that some under predictions in the highest rent areas of the Dublin County.

9. Clustering Analysis (K-Means)

I used K-Means clustering as it was a requirement in the project and it was applied on the county level metrices which were the Rent growth slope, Standard deviation and the Mean rent. After which using the Silhouette and elbow methods k=2 was optimal.

Cluster 1 (High-rent counties)

Dublin and its surrounding areas were characterized by the faster year on year growth along with higher volatility and high average rent.

Cluster 2 (Lower-rent counties)

Whereas Regional counties saw slower growth, more stable patterns and lower rent level.
I added the cluster assignment as feature in the modelling dataset. So the K-means model

identified two county clusters like 2 group, The first group that is the Cluster 1 contained Dublin and its sub regions along with commuter belt counties with high rents and high growth whereas in the second group that is the cluster 2 containing the regional counties that is the small town saw lower rents and slower growth, Clear segmentation demonstrates structural differences in the Irish housing market and was used in the model via the county cluster feature.

10. Final Model Performance

So, my final model got a rent variance of 79% which in average error is around **€140** which is pretty strong considering Irelands strong and diverse housing.

The model was then saved as "mlp_final_all_data_with_cluster.pkl"
The notebook includes a demonstration cell showing how to load it and predict new rent values.

10.1 Model Performance Comparison Table

MODEL	MAE	RMSE	R^2
Linear Regression (test)	186.35	260.49	0.707
Random Forest (10-fold CV)	242.57 ± 1.94	320.27 ± 3.96	0.562 ± .01
MLP (10-fold CV)	142.89 ± 3.51	221.53 ± 6.60	0.791 ± .008

Figure - 7

Figure – 7 which sums up all my work shows all the 3 models I used and clearly shows MLP achieves the strongest performance across all metrics, with the lowest MAE and RMSE and the highest R^2. This confirms that the MLP is the best model that captured non linear relationships in the rental data more effectively than the Random Forest or the linear baseline.

11. Discussion

So, through my project I got to learn that Dublin consistently drives the upper end of the market whereas regional counties behave bit differently, the project showed clear geographical and temporal structure in rent pricing. My neural network performed best as rent is influenced by non linear interaction between geography, property characteristics and time.

Both feature selection methods also complemented each other as on one hand the permutation importance showed what the neural network relied on and on the other hand the Lasso exposed the global linear effects mainly in the Dublin region.

Clustering on top of that added a useful segmentation of the counties making the model even more interpretable.

12. Limitations

With Pro's there were limitations in it as well as the dataset lacked property level details like furnishing, size etc. which limits the predictive accuracy.
County level aggregation also hides within the county variations that is the expensive and the cheap neighbourhoods along with no monthly or seasonal affects were available in the dataset
Despite all this the model performed well.

13. Conclusion

In the end my project of making a working model was successfully implemented and completed a data analytics workflow that is the cleaning, feature engineering, EDA, feature selection, modelling, clustering and evaluation. The neural network also performed well with an $R^2 \approx 0.79$ that is the model accuracy of 79% showing a very reliable rent prediction across the Ireland and its counties.
The major highlight has to be the clustering part as it gave a really clear picture on how the rental market naturally splits across Ireland

A small project with a good dataset is enough to make a good model and we don't need anything fancy to build something useful just with Python, scikit learn and a Jupyter notebook we went from a messy raw dataset to creating a predictive and responsive model along with extra clustering insights.

After completing this project I feel more confident in picking up more projects like these and trying new things in future building something useful.

Thankyou.

14. References

- Central Statistics Office (CSO) – RTB Average Monthly Rent Dataset.
- Scikit learn Documentation.