ANALYSIS OF LONDON HOUSING MARKET

GROUP 1

NITHYASHREE BALAJI

CLAUDIA BONON

MATTHEW PAYTE

SARAH POPAL

DANIEL TERENYI

Matt

MIS 6341.001 - Applied Machine Learning

1. Introduction:

Housing prices are rising worldwide, but because it is difficult to efficiently compare prices across wide geographical areas without considering location-based nuance, we are choosing to closely examine London. With the inflated and inhumane price of housing in London, it is important for people to understand the market and find what is affordable.

We want to predict housing prices based on factors such as house prices, number of houses sold, salary, crime, and population density to classify affordability based on mean/median income, as well as cluster housing into different categories to better understand the nuances of the data and see the broad categories of housing choices.

2. About the Data Set:

a. Data Set Name: Housing in London

   i. https://www.kaggle.com/datasets/justinas/housing-in-london/data

b. Description

   This dataset includes monthly and yearly housing price information for mostly London boroughs with additional information that relates/might affect those prices, found on Kaggle and gathered from the London government (https://data.london.gov.uk/), 7 columns/13,550 rows for Monthly data, 12 columns/1,072 rows for yearly data and any unique characteristics.

c. Objective

   With London being one of the most expensive housing markets in the world, we want to understand the scale to which prices, indicated directly in the dataset, might affect housing availability for the average person.

3. Software and Platforms used:

- Software/Tools: Python, Excel, Jupyter Notebook, R, maybe Tableau
- Platforms: R Studio and Anaconda

4. Roles of Members:

- Data Cleaning: Claudia
- Classification: Sarah
- Regression: Matthew
- Clustering: Daniel
- Data Visualization: Nithya

5. Data Cleaning

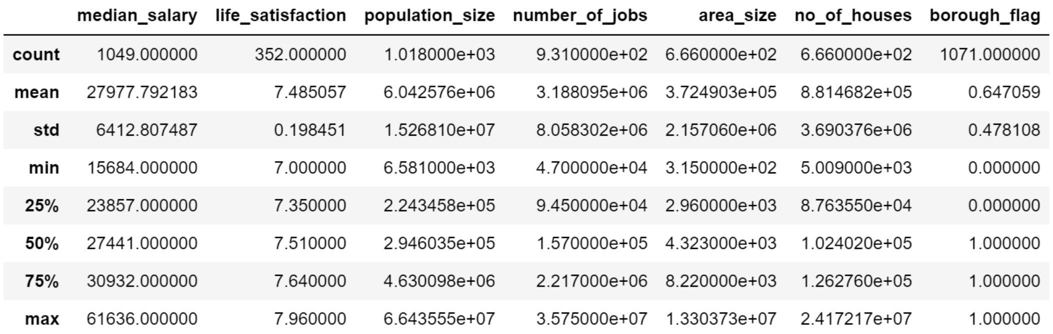
We have 2 datasets: one containing yearly information and the other containing monthly information. We analyzed both datasets and found that the data was quite clean. For the yearly dataset, we created a code to remove blank spaces and 'NaN' values. Then, we proceeded to understand each dataset.

The shape of the Yearly Dataset is: (1071 rows, 12 columns)

The shape of the Monthly Dataset is: (13549 rows, 7 columns)

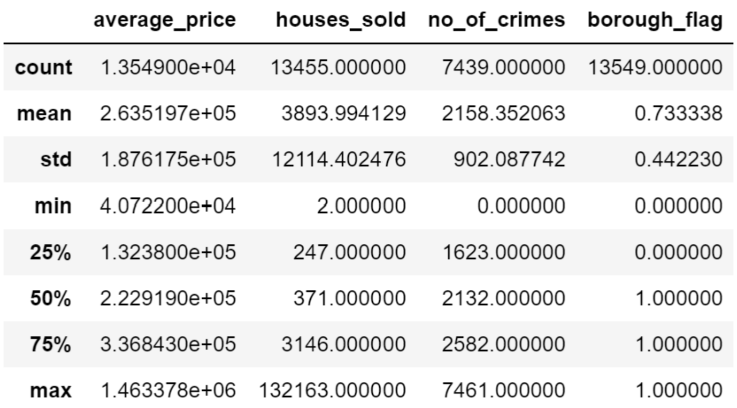
We applied summary statistics to understand both datasets and continue with the Visualizations

- Yearly Dataset



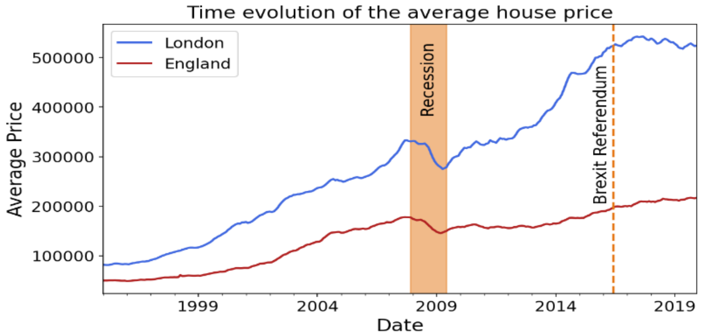
	median_salary	life_satisfaction	population_size	number_of_jobs	area_size	no_of_houses	borough_flag
count	1049.000000	352.000000	1.018000e+03	9.310000e+02	6.660000e+02	6.660000e+02	1071.000000
mean	27977.792183	7.485057	6.042576e+06	3.188095e+06	3.724903e+05	8.814682e+05	0.647059
std	6412.807487	0.198451	1.526810e+07	8.058302e+06	2.157060e+06	3.690376e+06	0.478108
min	15684.000000	7.000000	6.581000e+03	4.700000e+04	3.150000e+02	5.009000e+03	0.000000
25%	23857.000000	7.350000	2.243458e+05	9.450000e+04	2.960000e+03	8.763550e+04	0.000000
50%	27441.000000	7.510000	2.946035e+05	1.570000e+05	4.323000e+03	1.024020e+05	1.000000
75%	30932.000000	7.640000	4.630098e+06	2.217000e+06	8.220000e+03	1.262760e+05	1.000000
max	61636.000000	7.960000	6.643555e+07	3.575000e+07	1.330373e+07	2.417217e+07	1.000000

- Monthly Dataset

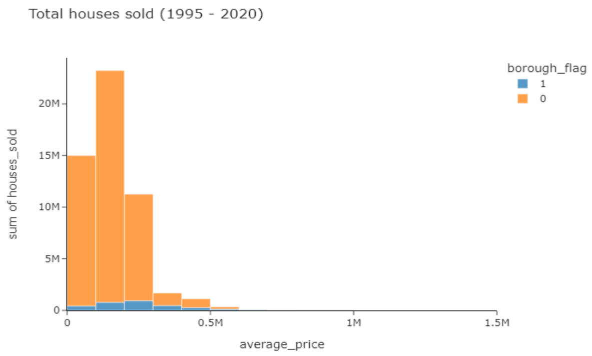


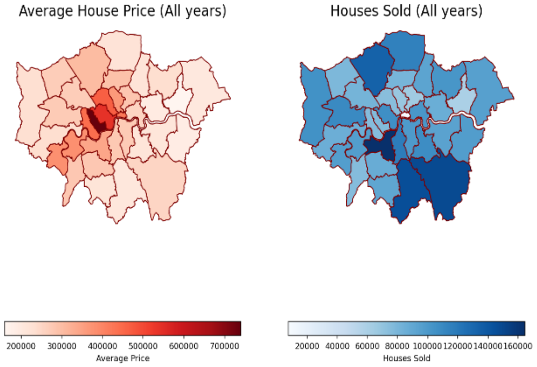
	average_price	houses_sold	no_of_crimes	borough_flag
count	1.354900e+04	13455.000000	7439.000000	13549.000000
mean	2.635197e+05	3893.994129	2158.352063	0.733338
std	1.876175e+05	12114.402476	902.087742	0.442230
min	4.072200e+04	2.000000	0.000000	0.000000
25%	1.323800e+05	247.000000	1623.000000	0.000000
50%	2.229190e+05	371.000000	2132.000000	1.000000
75%	3.368430e+05	3146.000000	2582.000000	1.000000
max	1.463378e+06	132163.000000	7461.000000	1.000000

6. Data Visualizations



This graph shows a vivid picture of soaring house prices in London boroughs, fueled by strong demand and limited supply, a trend mirrored, albeit more gradually, outside the capital. The impact of major events, such as the recession, is evident, leading to a temporary decline in house prices across both London and the rest of England. However, London's swiftly recovering and experiencing a surge in average prices until 2016.





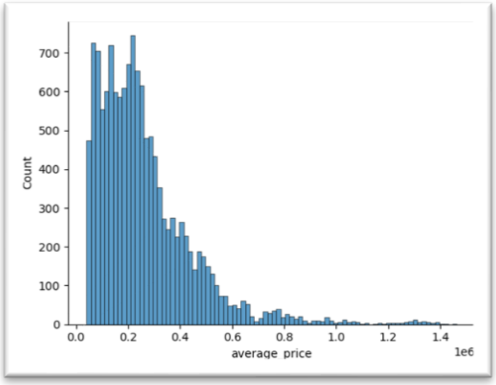
The total houses sold in London boroughs amount to approximately 3.2 million, while outside London, this number reaches around 49.6 million. This stark difference in the metrics can be attributed to geographical factors where the size of a borough plays a crucial role in the number of houses sold, with larger boroughs generally seeing more transactions. Interestingly, the most affluent boroughs tend to be smaller in size, showcasing an inverse relationship between size and turnover.

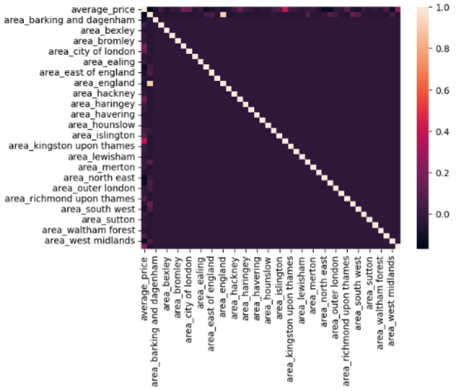
7. Regression Analysis

Further steps were taken in Regression Analysis to tweak the monthly housing dataset to better accommodate regression analysis. Specifically, the “date” “code” “borough_flag” and “no_of_crimes” predictors were dropped. The first three had questionable appropriateness in helping to predict a continuous outcome focusing on London housing on a basis of regression, and the “no_of_crimes” predictor had 6110 null values in the original dataset. 94 records from the original dataset in which “houses_sold” contained null values were also dropped, leaving a total of 13549 records to consider.

The “Area” categorical predictor was then transformed into k-1, in this case 45, dummy predictors. In total, 47 predictor variables were applied to predict continuous values of average monthly housing price.

A frequency histogram of that average_price outcome in the dataset, as well as a heatmap of predictor correlations, may be seen below (see Powerpoint for full size images):





The data was then split, with 70% of the original data partitioned into the training set, and 30% of the original data partitioned into the testing set. From there, the following models were tested: Linear Regression, Polynomial Regression, Polynomial Regression with Gridsearch, Ridge

Regression with Gridsearch, Lasso Regression with Gridsearch, Decision Tree Regressor, and Random Forest Regressor. Polynomial Regression and Polynomial Regression with Gridsearch were attempted but were too computationally inefficient for consideration. A basic polynomial regression with n_jobs = 8 on a computer whose n_cpu = 12 was not completed within an hour of executing that code.

The best performing regression model was Random Forest Regression with 500 trees and 1000 maximum samples. Its R^2 accuracy was printed as `0.6620678697391713 for the training data, and 0.581369297386636 for the testing data.  An overall snapshot of each model's R^2 abbreviated accuracy scores may be seen below:`

Model	**Training R^2 Score**	**Test R^2 Score**	**Best Gridsearch**
Linear Regression	0.4098995	0.3805054	N/A
Polynomial	N/A	N/A	N/A
Ridge	0.4098994	0.3805083	0.4034508
Lasso	0.4098995	0.3805054	0.4034507
Decision Tree	0.4470087	0.4370932	0.4282369
Random Forest	**0.6620677**	**0.5813693**	**N/A**

8. Classification Analysis

A) Data Cleaning

The analysis of the London housing market through classification methods involved applying advanced machine learning algorithms to decipher patterns and predict crucial metrics. Beforehand, the data was meticulously cleaned to enhance dataset usability, addressing inconsistencies from multiple sources.

- Non-numeric fields were converted to appropriate numeric formats
- irrelevant or unusable fields, such as "life_satisfaction" and "area_size," were eliminated
- Missing values were addressed by imputing numerical fields with the median and categorical fields with the most frequent value
- The 'date' column was transformed to extract the 'year', which became a pivotal feature for our modeling

- Engineered the "affordability" feature -- It was derived from mean salary data, categorizing areas into "low," "medium," and "high" based on salary distribution quartiles at the 30th and 70th percentiles.

B) Feature Selection

Following data cleaning, our project proceeded with feature selection to enhance model performance, which involved selecting the most relevant features that could significantly impact the predictions. A correlation matrix was used to identify key features that were used in our final analysis, and feature selection was further refined by employing multiple Random Forest models to assess the impact of various features on model accuracy, by iteratively training models with different combinations of features and comparing their performance.

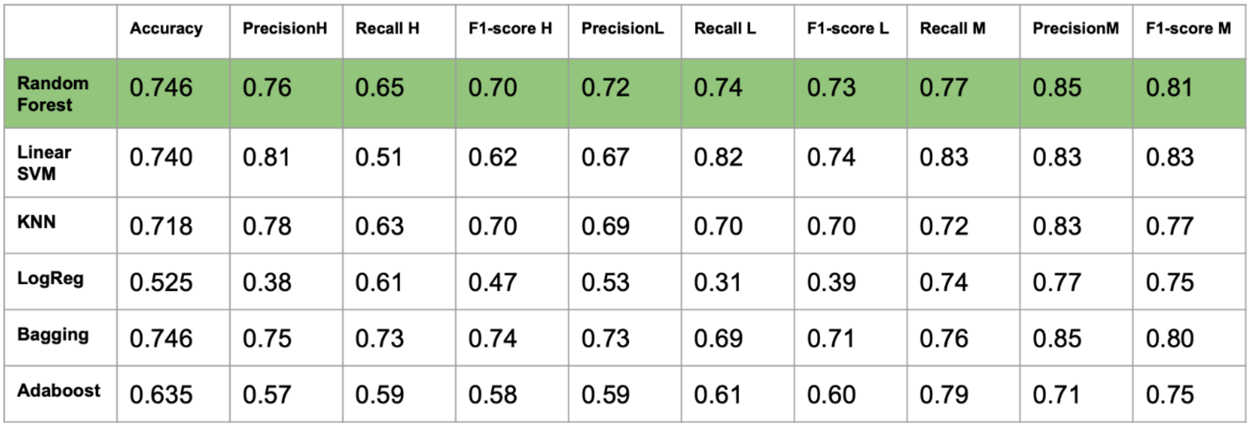
C) Model Training and Hyperparameter Tuning

After selecting the most informative features, we then transitioned to the training phase. We trained seven different models, Random Forest Classifier, Linear SVM, KNN, LogReg, Bagging, Adaboost, and Gaussian Naive Bayes. GridSearchCV was used to find the best parameters for each of the models, and the best n for KNN was identified as the "n" that produced the highest accuracy score. Random Forest had the highest accuracy and superior precision and recall scores, while Naive Bayes had the lowest and was thus not included in the final model evaluation. For our best model, Random Forest, GridSearch CV had returned the best parameters as follows:

```
Best Parameters: {'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 2,
'n_estimators': 200}
```

D) Model Evaluation

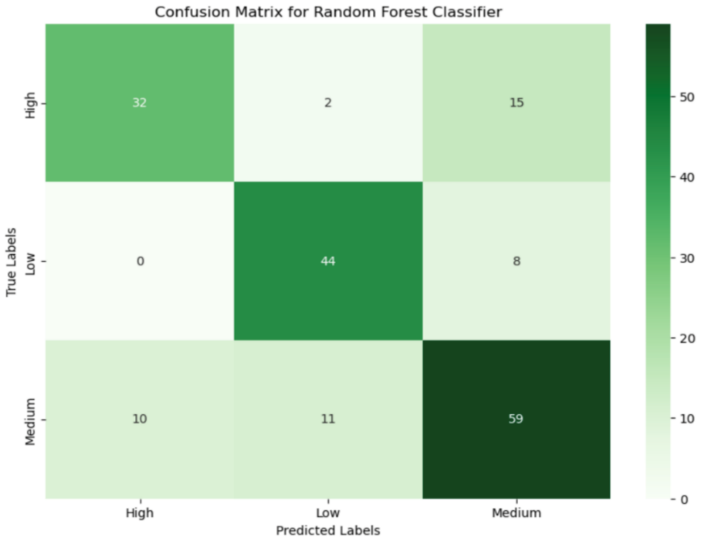
After tuning the models with GridSearchCV, we conducted a comparative analysis of their performances using key metrics like accuracy, precision, recall, and F1 scores, in order to pinpoint the most effective model. This analysis was crucial for selecting the algorithm that best balanced accuracy and robustness, ensuring reliable predictions.



	Accuracy	PrecisionH	Recall H	F1-score H	PrecisionL	Recall L	F1-score L	Recall M	PrecisionM	F1-score M
Random Forest	0.746	0.76	0.65	0.70	0.72	0.74	0.73	0.77	0.85	0.81
Linear SVM	0.740	0.81	0.51	0.62	0.67	0.82	0.74	0.83	0.83	0.83
KNN	0.718	0.78	0.63	0.70	0.69	0.70	0.70	0.72	0.83	0.77
LogReg	0.525	0.38	0.61	0.47	0.53	0.31	0.39	0.74	0.77	0.75
Bagging	0.746	0.75	0.73	0.74	0.73	0.69	0.71	0.76	0.85	0.80
Adaboost	0.635	0.57	0.59	0.58	0.59	0.61	0.60	0.79	0.71	0.75

Through this in-depth analysis, we successfully identified the Random Forest Classifier as the most accurate model with the ability to effectively capture the complex interrelations within the housing data, with an accuracy of 74.6%. We then looked at the results of this model more in depth with a confusion matrix, and identified key areas of improvement:

E) Visualization

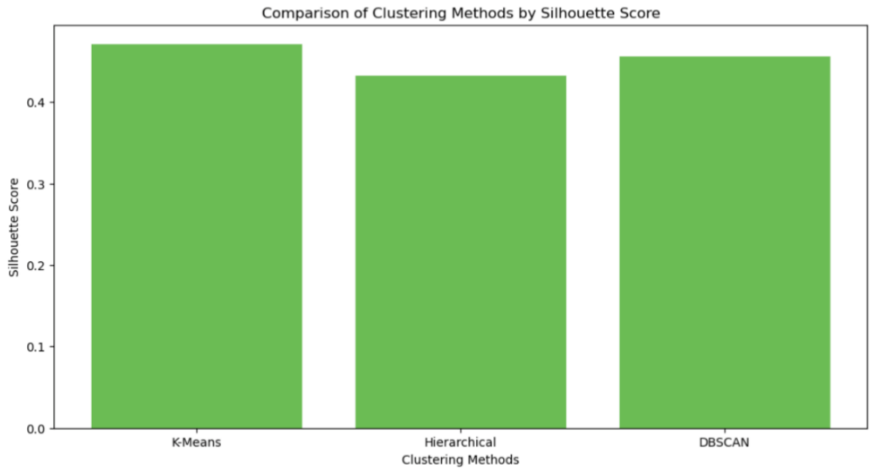


This confusion matrix for the Random Forest Classifier highlights a strong predictive accuracy for low and medium categories, with counts of 44 and 59 correct predictions, respectively. However, there is some misclassification between high and medium categories, indicating areas where model performance could potentially be improved, leading to future steps to improve that accuracy through more predictive features and potential model tuning.

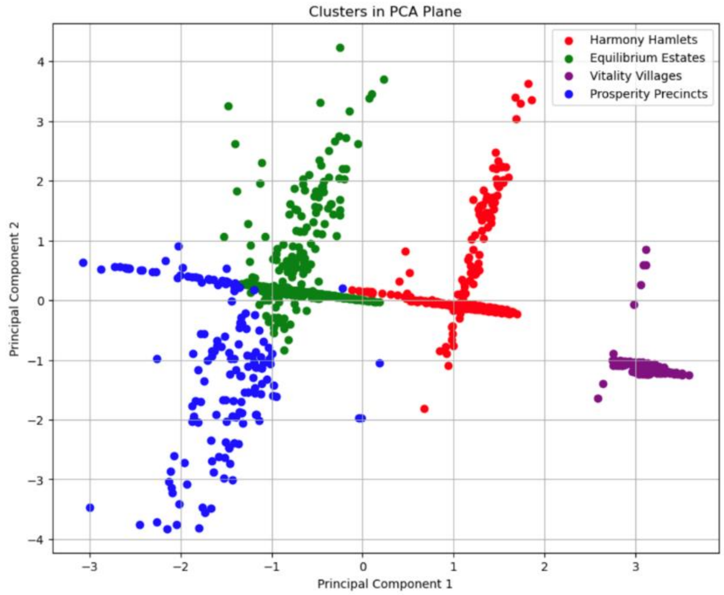
9. Clustering Analysis - Dániel

In our examination of the London housing market through advanced clustering techniques, we uncover nuanced patterns diverging from typical income based residential analyses. Our exploration employs several clustering methods to dissect the underlying structure of housing data, with a particular emphasis on K-Means clustering. Segmenting the London housing market reveals not just economic divisions, but also variations in lifestyle and community engagement across different boroughs.

Using Python and Jupyter notebook, we began with data loading, cleaning (duplicate and missing value handling), and transformation (date and categorical variable conversion). Next, we explored the data using correlation matrices and clustered using k-means with four clusters. For the deep dive into clustering, we examined centroids and calculated PCA loadings to interpret the influence of original features on the principal components to combine data management, exploratory analysis, clustering, and dimensionality reduction in a cohesive analysis framework.



KMeans was determined to be the most effective clustering method from the silhouette scores, although DBSCAN excelled in handling outliers and Hierarchical was useful for revealing nested data structures.



- **Red, Harmony Hamlets:** Encompasses lower-income areas outside borough zones with unexpectedly high life satisfaction. These moderately populated areas might represent suburban or semi-rural settings where community strength offsets economic limitations.
- **Green, Equilibrium Estates:** Features middle-income, moderately satisfied neighborhoods within boroughs, not densely populated. These areas likely represent balanced residential zones offering stable living conditions.
- **Purple, Vitality Villages:** Represents densely populated, low-income urban cores outside boroughs, yet exhibiting high life satisfaction. This could indicate vibrant communities where social and cultural dynamics positively impact well-being.
- **Blue, Prosperity Precincts:** Comprises high-income, low-satisfaction regions within boroughs, sparsely populated. These areas might face challenges like high living costs or social isolation, suggesting affluence does not equate to happiness.

Our findings challenge traditional perceptions linking income levels to life satisfaction and community strength, offering fresh insights into the complexities of urban living in a rapidly evolving global landscape, where new challenges brought by the vast and growing trenches of class disparities must be taken heavily into consideration by those it affects most: the workers of the world. This comprehensive analysis prompts a reevaluation of how urban environments are structured and the profound impacts they have on their residents, emphasizing the need for a broader understanding of urban wellbeing beyond economic metrics.

10. Conclusion

In London's housing market, the data reveals that higher income does not necessarily mean higher satisfaction, nor does lower income imply weak community bonds. This underscores the role social cohesion and community amenities play in residents' wellbeing, challenging conventional assumptions about affluent areas and highlighting new economic trends as worldwide disparities increase.