PREDICTING LONDON RENT USING GEOSPATIAL DATA

Max Greenwood

1 INTRODUCTION

1.1 Background

London is a large city with a population of almost 9 million people. The money which flows through it has immense pulling power, attracting businesses, investors, and young professionals like myself looking to enter the workforce. With so much capital and wealth at the center of it, the demand for accommodation around the city has driven housing prices significantly higher than anywhere else in the UK. London, therefore, is an economic microcosm where higher wages and greater professional opportunity is matched by inflated prices for property, which might cost half as much anywhere else in the country. The churn of accommodation in London is faster as well, with renters moving through properties rapidly and often across large distances.

Each district of London has a unique character defined by its people, its natural surroundings, and the businesses and services within it. Rent prices also vary a significant amount within the city, with flats in the highly gentrified inner-city being notoriously costly and flats in the suburbs around London markedly less. The diverse nature of London living raises interesting questions about how well one can predict it might cost to live in a certain area, even before delving into property specifics.

1.2 Project aim

The aim of this project is to explore how London rental prices can be predicted using only geospatial data and information about surrounding venues. For simplicity, this investigation will only encompass prices of studio flats as a barometer for trends of other sorts of accommodation. The geographical scope will be the London metropolitan area which includes Greater London and its commuting belt. This area will be sub-divided by postcode district - the greatest level of granularity for which data is available. Supervised learning regression techniques will be used to build models using a training set of the data, and subsequently assessed with a test set.

1.3 Interest

Young professionals moving around London and first-time renters are the primary groups of stakeholders who would benefit from a predictive analysis involving venue data. By understanding how their preferred surroundings impacts their rent, they will be able to assess the likelihood of finding a place in a certain part of the city which aligns with their lifestyle and budget.

By extension, landlords and housing agents would be interested in whether a certain density of nearby venues affects the average price of a property in the area. They would therefore be able to match the needs of their potential customers more precisely and secure more business.

2 DATA

2.1 Variables and Data Sources

The predictive variables for the investigation will include:

- Venue data, such as the number of venues and types of venues in each district;
- Geographical data, such as the location, population, size, and distance from city centre of each district

The target variable for the investigation will be the average monthly rent of a studio flat.

2.1.1 Venue Data

Venue data will be gathered using the Foursquare API.

These data will help establish how the number of certain types of venues may correlate to the cost of living in London. For example, we might expect a district with a large number of expensive wine bars to reflect affluence of its inhabitants which could correspond to higher property prices. Alternatively, we might expect areas with a higher venue density to reflect a more urban environment, and therefore more expensive properties in a closer proximity to workplaces.

2.1.2 Geographical Data

Geographical data will be gathered using census data from the Office of National Statistics. The latest data available for postcode district geographies is from August 2019. Since population censuses are only conducted every ten years at the district level, the best available statistics from 2011 are a little out of date. However, for this experimental purpose they should serve reasonably well.

These data will help establish how fundamental geographical aspects of districts may correlate to the cost of living in them. For example, it is reasonable to assume that there is a relationship between rent and distance from city centre, but what does this relationship look like and how much weight does this factor have when other factors are considered in conjunction? Also, will population density indicate anything about how expensive a district might be to live in?

2.1.3 Rental Data

Rental data will be gathered from the open-source statistics provided by the Valuation Office Agency. The latest data available on the UK government website for the private rental market in London dates between July 2018 to June 2019.

3 METHODOLOGY

3.1 Data Cleaning and Feature Extraction

The data from the two three sources above were compiled into a single table. To begin with the geographical and rental data were combined. The geographical data contained information on all postcode districts in the UK; the first task was to cut all the districts which were absent from the rental dataset. Cleaning the rental data meant removing districts which had no information about studio flat costs. The two cleaned up dataframes were then concatenated and any districts with missing information were removed. This presented a total of 160 districts with complete information to be used for modelling.

3.1.1 Calculating distance from the city centre

From the longitude and latitude information and the central point of London obtained from the $geolocator$ library, the approximate distance of each district from the city centre was calculated using the Haversine Formula (Equation 3.1). In this equation, d is the distance between two points, ϕ_i is the latitude of point i, λ_i is the longitude of point i, and r is equal to the radius of the earth in kilometers.

$$d = 2r\ arcsin\left(\sqrt{sin^2\left(\frac{\phi_1-\phi_2}{2}\right)} + cos(\phi_1)\ cos\ (\phi_2)\ sin^2\left(\frac{\lambda_1-\lambda_2}{2}\right)\right) \qquad \text{Equation 3.1}$$

The districts were subsequently plotted on a Folium map to sense check the calculated distances and visualise the data accumulated up to that point. Figure 3.1 displays the 160 district centers. Since no geojson files were readily available to plot a choropleth map, postcode districts are approximated as circles with 1 km radii for the remainder of this investigation.

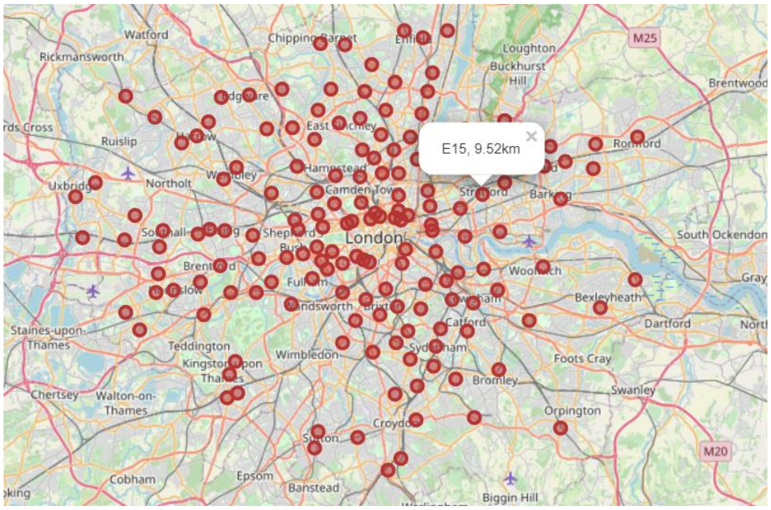


Figure 3.1: Folium map of London postcode district center points. Each point contains the district postcode and distance from city center.

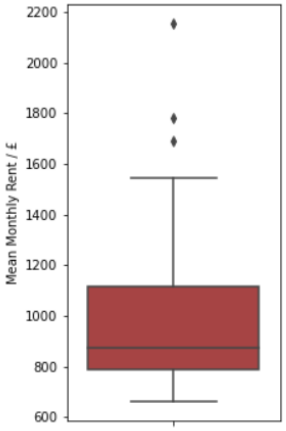
3.1.2 Obtaining venue data

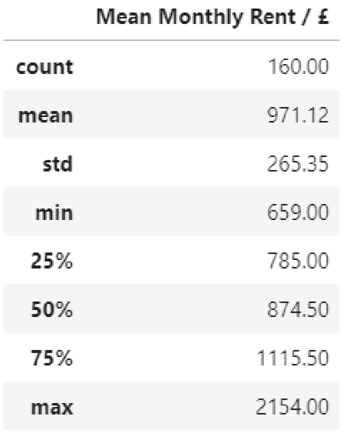
For each of the districts, the Foursquare API was called to obtain information about all venues within a 1 km radius of the centre point. One-hot encoding was used to identify how many of each venue category are contained within each approximated district. Subsequently, the total number of each venues per district could also be calculated. These data were appended to the dataframe containing geographical and rental data; surplus columns such as area names and population were also removed in preparation for modelling.

3.2 Exploratory Data Analysis

3.2.1 Overviewing rent prices

The first aspect of data exploration involved only the target variable. A boxplot was plotted (Figure 3.2) to visually convey key information about the dataset. The features of this plot are also detailed in Table 3.1.





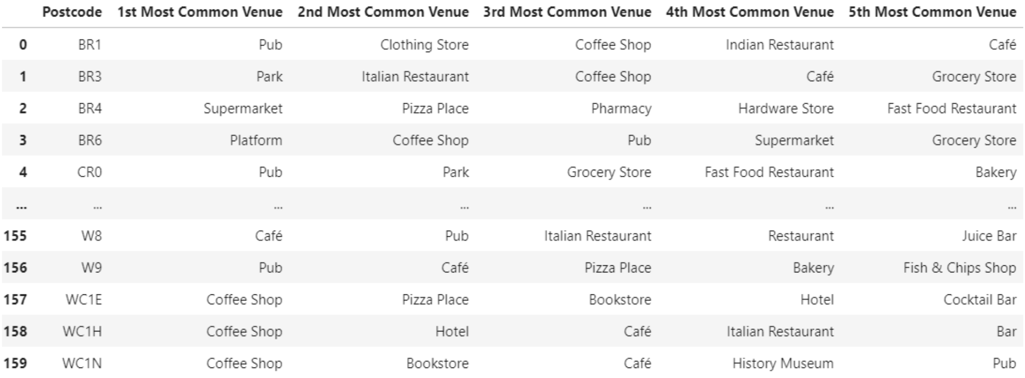
	Mean Monthly Rent / £
count	160.00
mean	971.12
std	265.35
min	659.00
25%	785.00
50%	874.50
75%	1115.50
max	2154.00

Figure 3.2: Box plot of mean monthly rent prices. *Table 3.1. Descriptive statistics of mean monthly rent prices.*

The data is skewed towards bottom, as implied by the median average rent (£874.50) being significantly lower than the mean average rent (£971.12). More districts fall below the mean average rent, which is what one might expect given the concentration of wealth in a small area at the centre of London. There are three outliers which did not warrant omission from modelling since it would not be as reflective of the true distribution of rent prices.

3.2.2 Rent vs. common venue types

To get a sense of how indicative venue densities might be to predicting rent prices, an additional dataframe was produced to identify the most common venue categories within the city. By surveying the top five most common venues, those with the most data points could then be used to assess correlation with rent. An overview of the dataframe is displayed in Table 3.2.



	Postcode	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	BR1	Pub	Clothing Store	Coffee Shop	Indian Restaurant	Café
1	BR3	Park	Italian Restaurant	Coffee Shop	Café	Grocery Store
2	BR4	Supermarket	Pizza Place	Pharmacy	Hardware Store	Fast Food Restaurant
3	BR6	Platform	Coffee Shop	Pub	Supermarket	Grocery Store
4	CR0	Pub	Park	Grocery Store	Fast Food Restaurant	Bakery
...	...	...	...	...	...	...
155	W8	Café	Pub	Italian Restaurant	Restaurant	Juice Bar
156	W9	Pub	Café	Pizza Place	Bakery	Fish & Chips Shop
157	WC1E	Coffee Shop	Pizza Place	Bookstore	Hotel	Cocktail Bar
158	WC1H	Coffee Shop	Hotel	Café	Italian Restaurant	Bar
159	WC1N	Coffee Shop	Bookstore	Café	History Museum	Pub

Table 3.2: Top five most common venues per district.

From brief observation of the data, four venue categories were chosen to explore further. These categories were: Coffee Shops, Pubs, Cafes and Italian Restaurants. Figure 3.3 shows scatter plots of average rent against the incidence of these four venues, respectively.

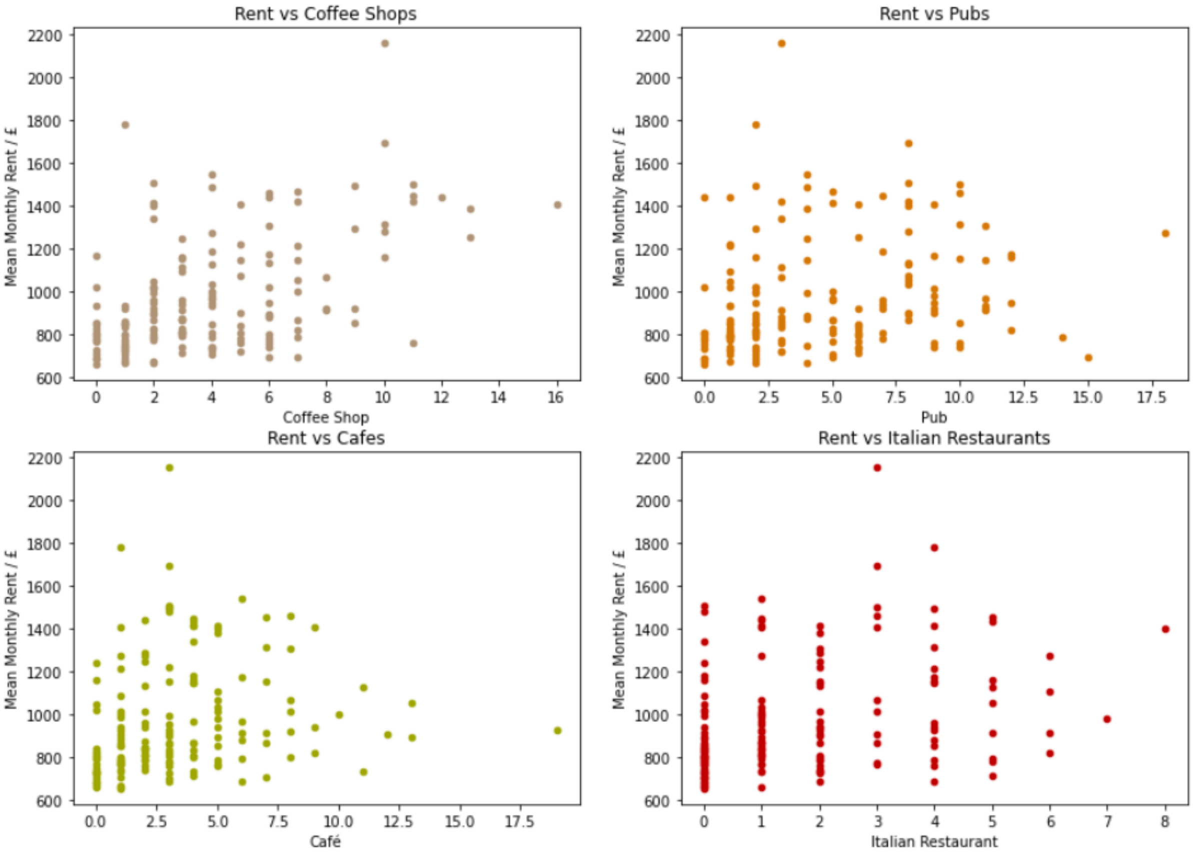
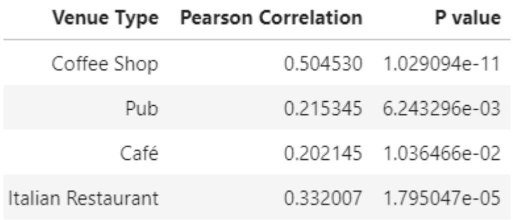


Figure 3.3: Scatter graphs of rent against commonly occuring venue incidence.

From these scatter plots there is no obvious positive or negative correlation between venue incidence and rent. Increasing number of coffee shops nearby perhaps correspond to a general increase in mean monthly rent, but it is not clear purely from visualization. Table 3.3 displays the Pearson correlation and P values for each of these features.

Table 3.3: Correlation between common venue incidence and rent.



Venue Type	Pearson Correlation	P value
Coffee Shop	0.504530	1.029094e-11
Pub	0.215345	6.243296e-03
Café	0.202145	1.036466e-02
Italian Restaurant	0.332007	1.795047e-05

The Pearson coefficient for Coffee Shops is indeed higher than the others, but in all cases the positive correlation is weak and the confidence in asserting this fact is high given the small P values.

3.2.3 Rent vs. selected venue types

Two more venue categories were selected to investigate further. Wine bars and Gyms were chosen to represent a class of venue which might be assumed to be more partial towards those of a higher economic status. That is, the more wine bars and gyms there are in a neighbourhood, the more likely it is residents in that area have disposable income and the more they will have to spend on property. Figure 3.4 shows the scatter graphs of rent against the incidence of these two venues.

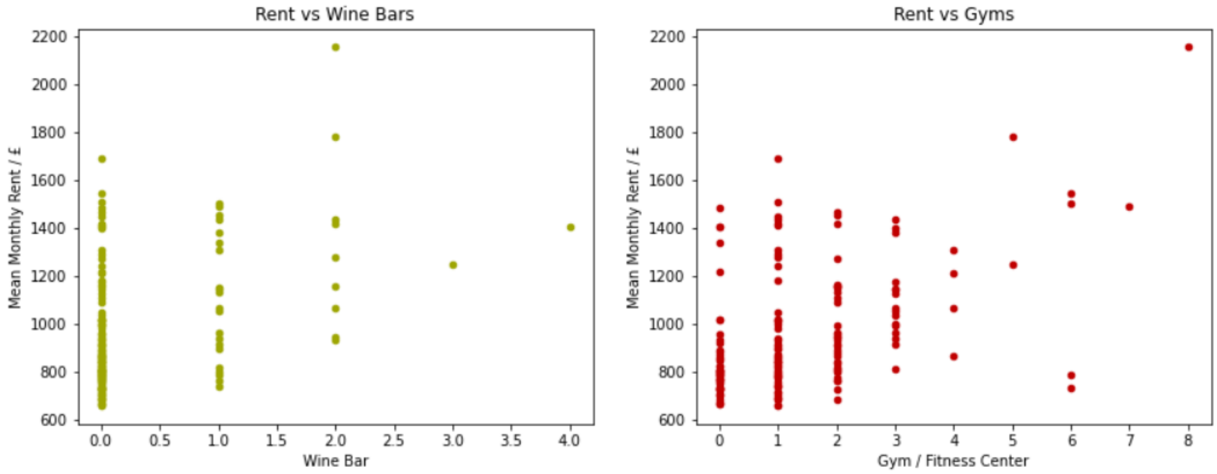
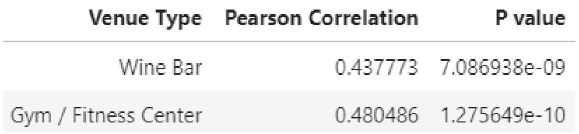


Figure 3.4: Scatter graphs of rent against selected venue incidence.

Both venues seem to show a broad positive correlation, albeit in with few occurrences in total in the case of Wine Bars. This intuition is supported by the Pearson coefficients seen in Table 3.4. However, the coefficients are still well below 1 and therefore far from proving strong positive correlations between these pairs of variables.

Table 3.4: Correlation between selected venue incidence and rent.

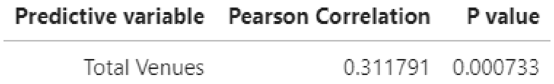


Venue Type	Pearson Correlation	P value
Wine Bar	0.437773	7.086938e-09
Gym / Fitness Center	0.480486	1.275649e-10

3.2.4 Rent vs. total venues

The total number of venues in an area was suspected to be a potential indicator of its rent prices. A higher density of commercial and public buildings could suggest a more developed or popular urban environment which might in turn correspond to higher rent on average. The total number of venues per area is plotted in a scatter graph shown in Figure 3.5. The theory that it might be a strong indicator of rent is not vindicated by the Pearson coefficient of 0.31 seen in Table 3.5, however.

Table 3.5: Correlation between total venue incidence and rent.



Predictive variable	Pearson Correlation	P value
Total Venues	0.311791	0.000733

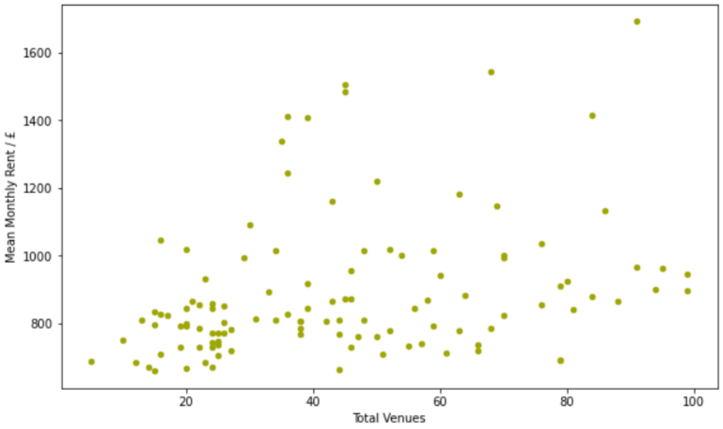


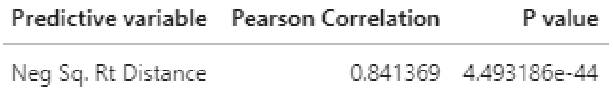
Figure 3.5: Scatter graph of rent against total venue incidence.

Not all districts were featured on this plot due to restrictions on how many venues could be called for each point. Total Venues was removed as a feature before modelling given the incomplete data and the weak correlation.

3.2.5 Rent vs. distance from city center

The relationship between average rent and distance from the city center is much more evident from visualisation. Inspecting the data without modification, an inverse relationship between the two variables is apparent, as evidenced in Figure 3.6. Transforming distance by a power of -0.5 presented a clear linear correlation between the two. This stronger correlation is reinforced by a reasonably large Pearson coefficient of 0.84 seen in Table 3.6.

Table 3.6: Correlation between the negative square root of distance from city center and rent.



Predictive variable	Pearson Correlation	P value
Neg Sq. Rt Distance	0.841369	4.493186e-44

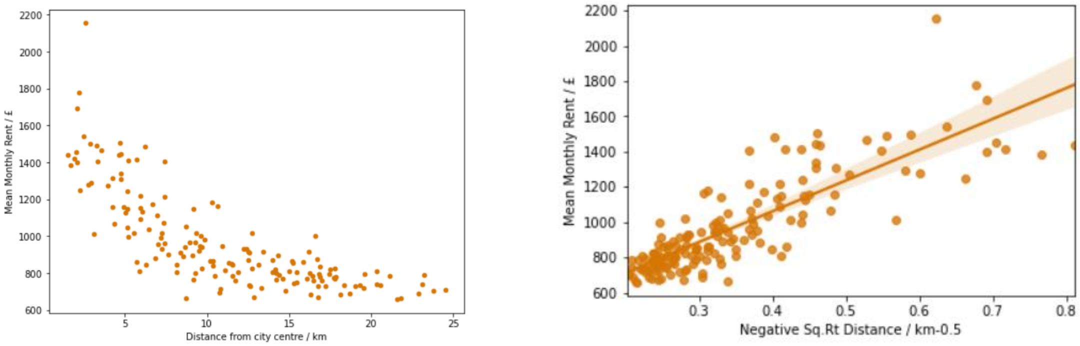


Figure 3.6: Relationship between rent and: unmodified distance from city centre (left); negative square root of distance (right).

3.3 Predictive Modelling with Multiple Linear Regression

To predict the average rent of studio flats within a subset of these districts, a multiple linear regression technique was employed using the aforementioned features of venue incidences and the transformed distance to the city center. The venue incidence linear correlations are largely weak as noted previously, but combining multiple venue features in a learning algorithm was assessed to see if together they would increase the predictive power of the subsequent model.

In advance of modelling, the dataset was mean normalised and split into a training set (60%; 96 districts), a cross-validation set (20%; 32 districts) and a test set (20%; 32 districts). The divide is visualised on a Folium map for one random state in Figure 3.7.

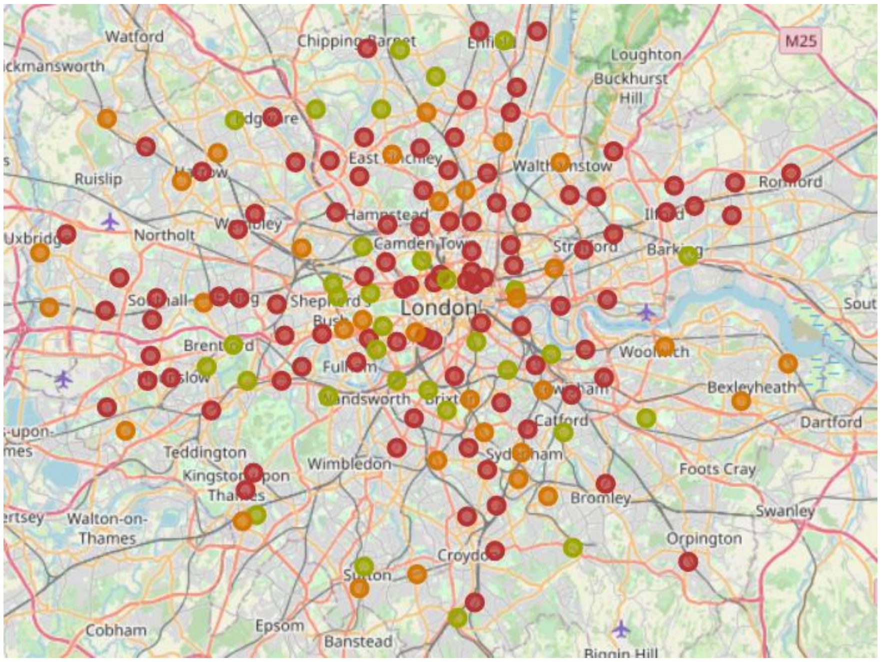


Figure 3.7: Folium map of training set, cross-validation set and test districts in red, orange and green respectively.

The dataset input involves a relatively small number of data points, $m = 160$, and a larger number of features, $n = 385$. This complexion makes it vulnerable to overfitting, where the algorithm will fit a hypothesis function too precisely to all the features of the training dataset and fail to make accurate predictions for new samples. Therefore, the Ridge regression technique was employed to introduce a regularisation parameter, α, which weakens the influence of certain feature parameters which contribute to overfitting. The cross-validation set is used for mapping this α regularisation parameter separate of the training data. With a close to optimum value of α calculated, the Ridge regression model can then be employed on the test set. In this investigation, α was trialed at a range of values between 0.01 and 10000, increasing by a factor approximately 3 on each iteration. The accuracy of the model predictions were assessed using mean squared error (MSE).

4 Results and Discussion

4.1 Predictive Modelling with Regression

4.1.1 Multiple Linear Regression with all features

A multiple linear regression (MLR) model was trained using 384 venue features and the modified distance from city center feature. The optimal regularisation parameter was identified to be 100 using the cross-validation set, as depicted in Figure 4.1. An extract of the subsequent predictions using the test set is shown in Table 4.1. The MSE for the cross validation and test set predictions were 40474 and 34832 respectively.

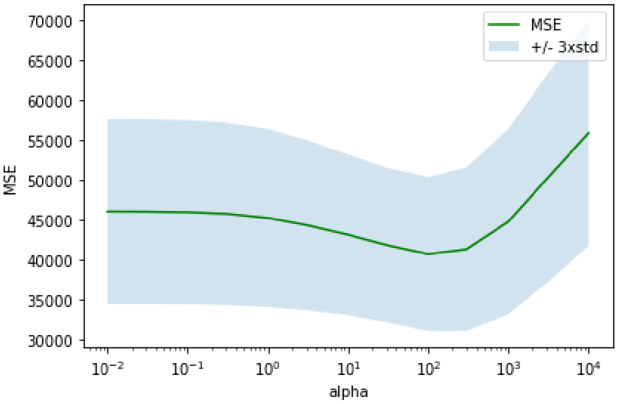
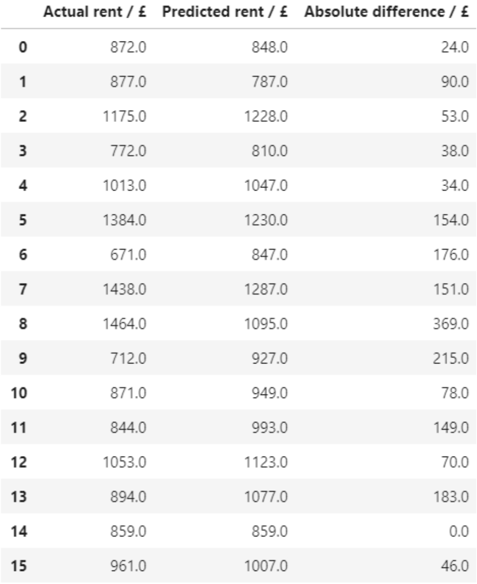


Figure 4.1: A line graph of mean squared error against α for one iteration of regularised MLR using a cross-validation set.

Table 4.1: Actual rent vs. predicted rent for an initial trial of an MLR model.



	Actual rent / £	Predicted rent / £	Absolute difference / £
0	872.0	848.0	24.0
1	877.0	787.0	90.0
2	1175.0	1228.0	53.0
3	772.0	810.0	38.0
4	1013.0	1047.0	34.0
5	1384.0	1230.0	154.0
6	671.0	847.0	176.0
7	1438.0	1287.0	151.0
8	1464.0	1095.0	369.0
9	712.0	927.0	215.0
10	871.0	949.0	78.0
11	844.0	993.0	149.0
12	1053.0	1123.0	70.0
13	894.0	1077.0	183.0
14	859.0	859.0	0.0
15	961.0	1007.0	46.0

Surveying the differences between the actual rent and predicted rent, it is apparent that the model performs reasonably well on districts with rents close to the mean across London, as is the case with Prediction 14 where there is no difference between actual and predicted target variable, but struggles when the rent is further out from the centre of the distribution, as is the case with Prediction 8 where the difference is £369. Indeed, the predicted values lie in a much narrower range than the actual values, as shown in the histogram in Figure 4.2. The reason for this is the small weight the regression learning algorithm affords to districts with large average rent; MSE is minimised when the majority of districts around the mean are prioritised over the small number a few standard deviations away.

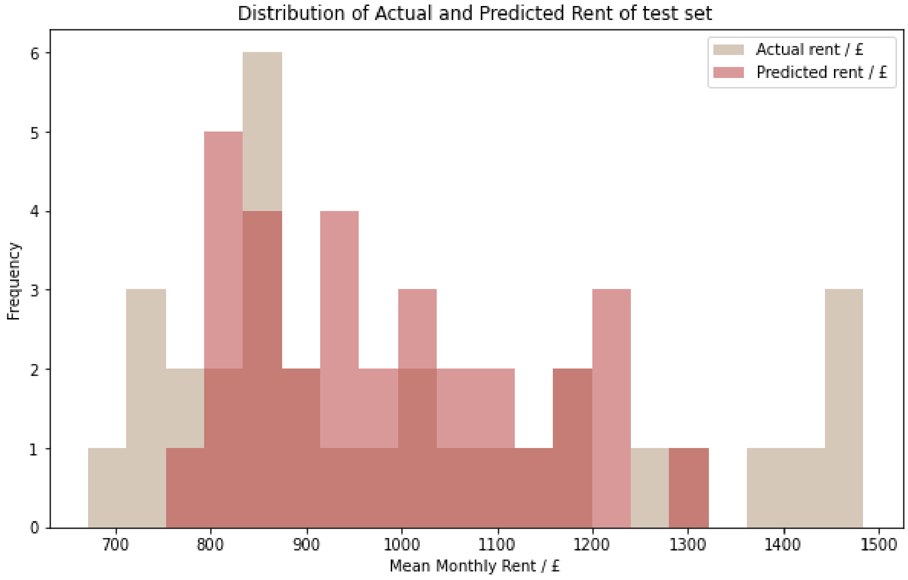


Figure 4.2: Histogram displaying the distribution of actual and predicted rent using an MLR model.

Another issue is the magnitude and difference between the MSE values for the cross validation and test set predictions, which were 40474 and 34832 respectively. The difference between the values might be attributed to a large variability of error in small subsets of the data which can be mitigated by iterating the train-test and averaging, but the magnitudes of the error hint at poor predictive power, possibly due to continued overfitting or redundant features.

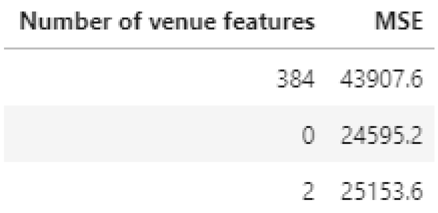
The poor predictive power can be addressed by reconsidering the number of features used for modelling. The narrow predictive range can be addressed by playing with the error weighting for districts in the various regions of the target variable distribution.

4.2 Addressing problems

4.2.1 Addressing predictive power by removing features

The average test set MSE of the Ridge linear regression model over 28 train-test cycles was calculated with three inputs of varying feature quantities. The first included all 384 venue features as before. The second included only the distance to city center feature and no venue features. The third included the distance feature and the two selected features discussed in Section 3.2.3. The MSE values are shown in Table 4.2.

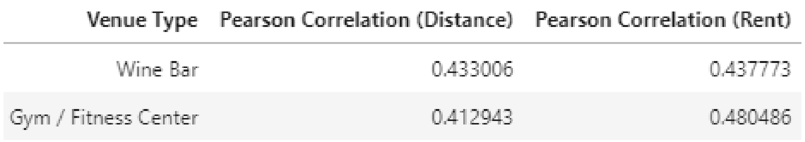
Table 4.2: Mean MSE of models involving the distance feature and a certain number of venue features.



Number of venue features	MSE
384	43907.6
0	24595.2
2	25153.6

Given that the best MSE was recorded from a simple linear regression model (SLR) using only distance data, it is apparent that the venue incidence features are obfuscating our learning algorithm. Even when the two venue features Wine Bar and Gym / Fitness Center, with weak positive correlations are reincorporated, the mean MSE increase by over 400. It is clear that these features are not providing additional useful information. Checking the correlation between these two features and the distance variable, the Pearson coefficients are similar to those involving the target variable. A comparison is shown in Table 4.3.

Table 4.3: Correlations between selected venue incidences and both distance and rent.



Venue Type	Pearson Correlation (Distance)	Pearson Correlation (Rent)
Wine Bar	0.433006	0.437773
Gym / Fitness Center	0.412943	0.480486

Where venue incidence has a similar or more correlation to distance as it does to the target variable, it is not valuable for training a learning algorithm. More work could be done to identify which, if any, of the venues would be useful to include in the modelling phase.

4.2.2 Addressing narrow prediction range by altering sample error weights

Going forward with just an SLR model, the narrow prediction range may not be as much of a problem as it was for the MLR model. Predicted and actual values from test sets for the SLR model were obtained over 28 iterations as before and plotted as a histogram displayed in Figure 4.3.

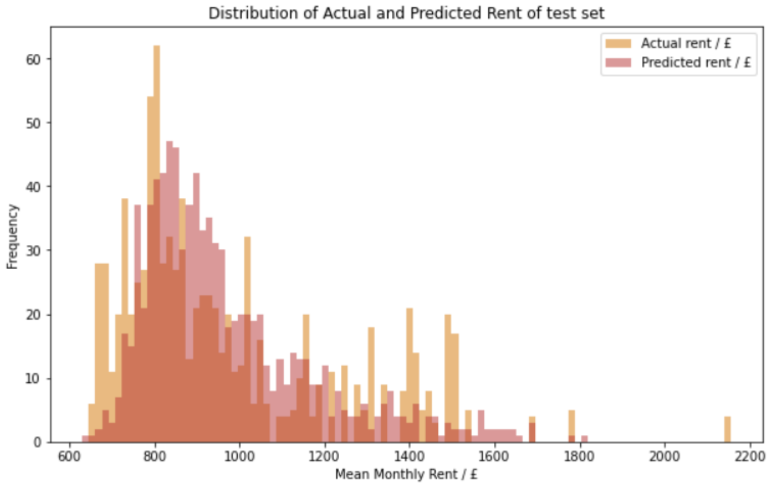


Figure 4.3: Histogram displaying the distribution of actual and predicted rent using an SLR model over 28 train-test cycles.

From this, it is apparent that the prediction range is close to the actual range. The prediction distribution is smoother than the actual rent distribution, partly due to the small number of districts being worked with which create localised maxima bins across the chart. There is no hope of matching these statistical nuances with an SLR prediction model. Therefore, there is no need to alter sample error rates; the issue has resolved itself by removing redundant features.

5 Conclusion

This study aimed to explore how mean London rental prices can be predicted using only geospatial data and information about surrounding venues. Venue incidence data was, overall, discovered to have only weak correlations with rent and was detrimental to the accuracy of a regularised MLR prediction model. In comparison, geographical data used in calculating the distance of London districts from the city center was a valuable asset for the learning algorithms. The SLR model including only the distance feature outperformed the MLR model. This simple model would act as a good foundation for an application targeted at potential tenants exploring the London rental market at a district level.

6 Further Directions

There are four immediately evident actions which would improve or build upon this preliminary study:

i. Measure over precise areas. The districts in this study were approximated as circular with lots of overlap and unscanned space. Using geojson data to build a map with area boundaries to identify venues in and allow per capita data to be produced might facilitate more productive feature engineering.

ii. Obtain more training examples. The dataset used for training and testing machine learning algorithms was quite small. By identifying missing variables for each district and compiling a more complete London district dataset, the learning algorithm will perform better so long as overfitting is mitigated against.

iii. Assess venue data more thoroughly. Although altogether the 384 venue features hampered the algorithms, perhaps there are some which would be useful for modelling.

iv. Identify and engineer other features to use. There are other features at the district level relevant to the business problem which could be sourced and used to improve predictive modelling. For example, population demographics and local council data may be helpful in creating more accurate models.