Predicting rental costs in London using geospatial data

Max Greenwood

Content

What I will talk about

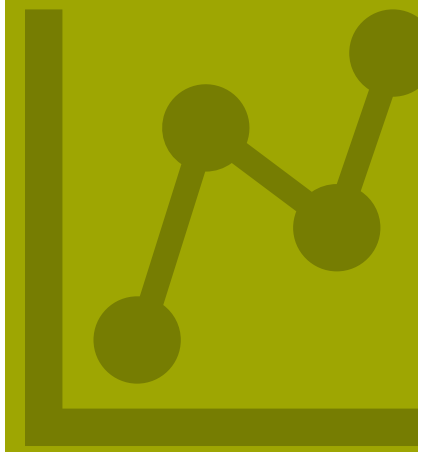
1

The Business Problem

2

The Data Processing

3



The Data Exploration

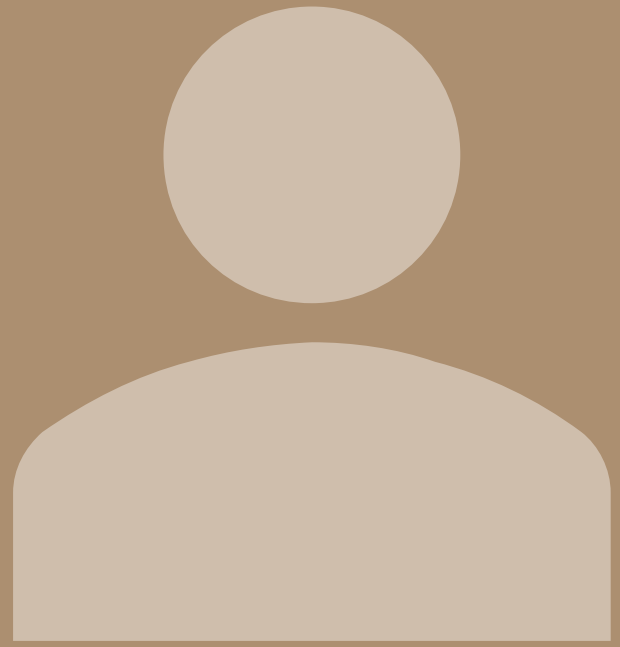
4

The Modelling and Results

Aim: to explore how London rental prices can be predicted using geospatial data.

1

The Business Problem



The Problem

Who is interested and why?

LONDON

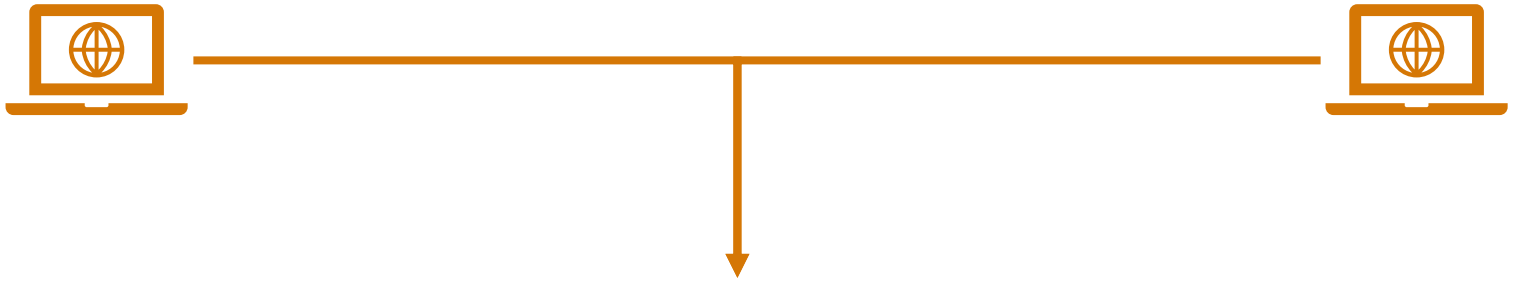
2

The Data Processing

Data Cleaning

Sourcing and concatenating data

Office for National Statistics



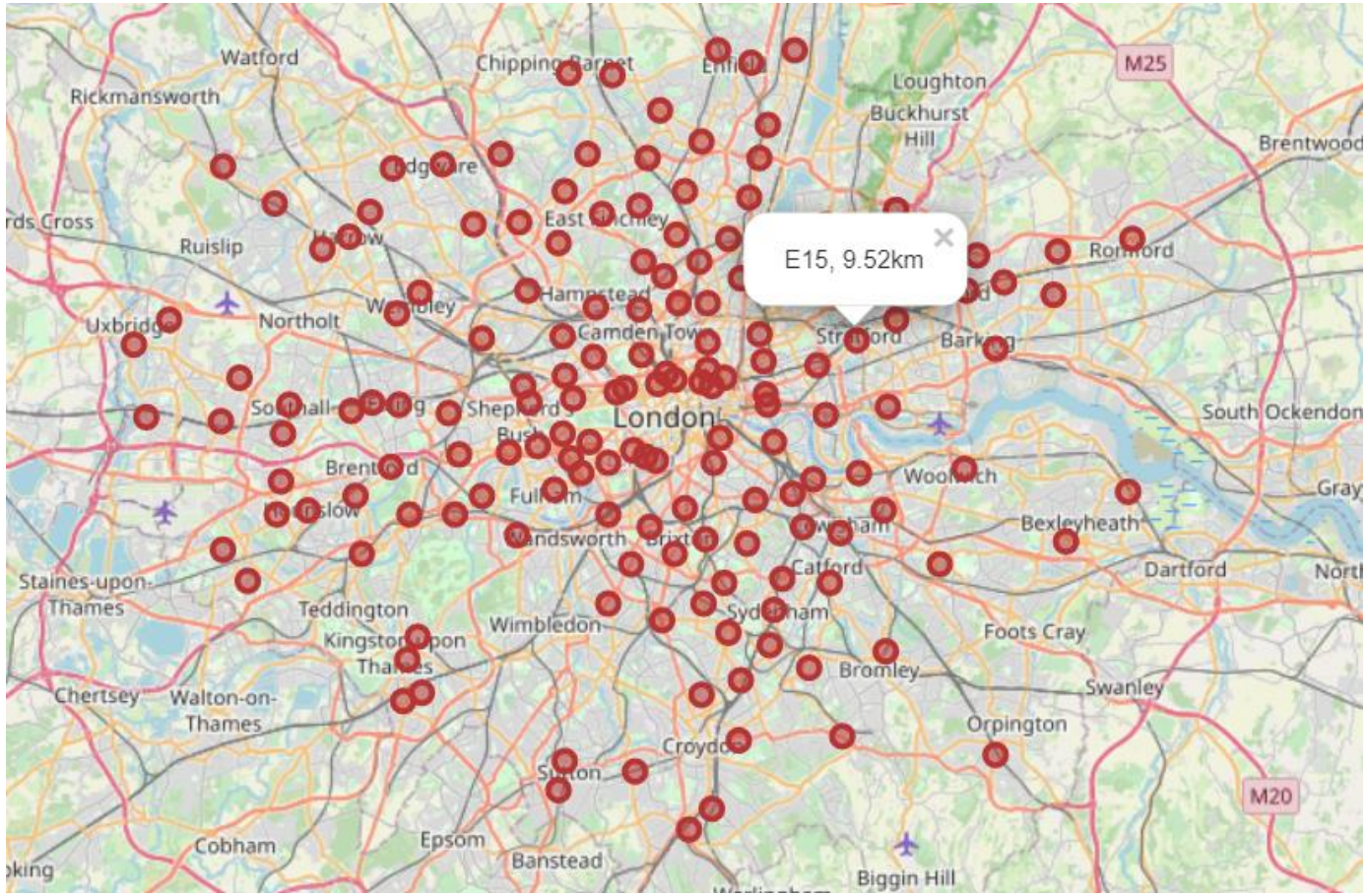
Valuation Office Agency

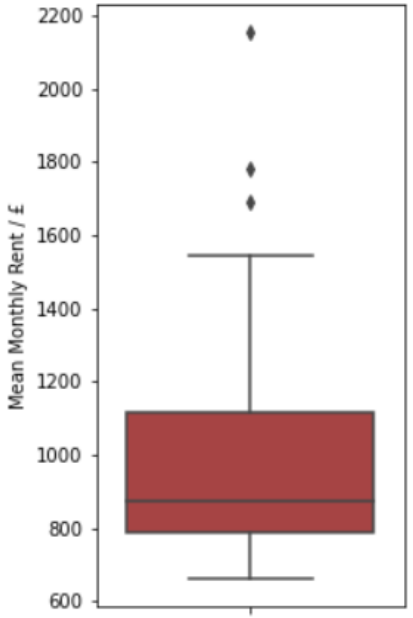
	Postcode	Area	Latitude	Longitude	Population	Mean Monthly Rent / £
0	BR1	Bromley, Bickley, Downham	51.4107	0.019429	55962.0	706.0
1	BR3	Beckenham, Eden Park, Elmers End, Park Langley...	51.4034	-0.031774	47411.0	719.0
2	BR4	West Wickham	51.3756	-0.009892	19367.0	671.0
3	BR6	Orpington, Farnborough, Downe, Pratt's Bottom,...	51.3674	0.092537	46073.0	665.0
4	CR0	Croydon, Addiscombe, Shirley, Addington, New A...	51.3732	-0.078287	153812.0	745.0
...	...	...	...	...	...	...
155	W13	West Ealing, Northfields	51.5131	-0.321423	32428.0	916.0
156	W14	West Kensington, Kensington Olympia, Holland Park	51.4954	-0.211163	37204.0	1153.0
157	WC1E	University College London	51.5207	-0.132398	2256.0	1439.0
158	WC1H	St Pancras	51.5248	-0.126315	8272.0	1418.0
159	WC1N	Great Ormond Street Hospital	51.5220	-0.120627	7084.0	1384.0

160 rows × 6 columns

Data Summary

Overview of processed data





	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

3

The Data Exploration

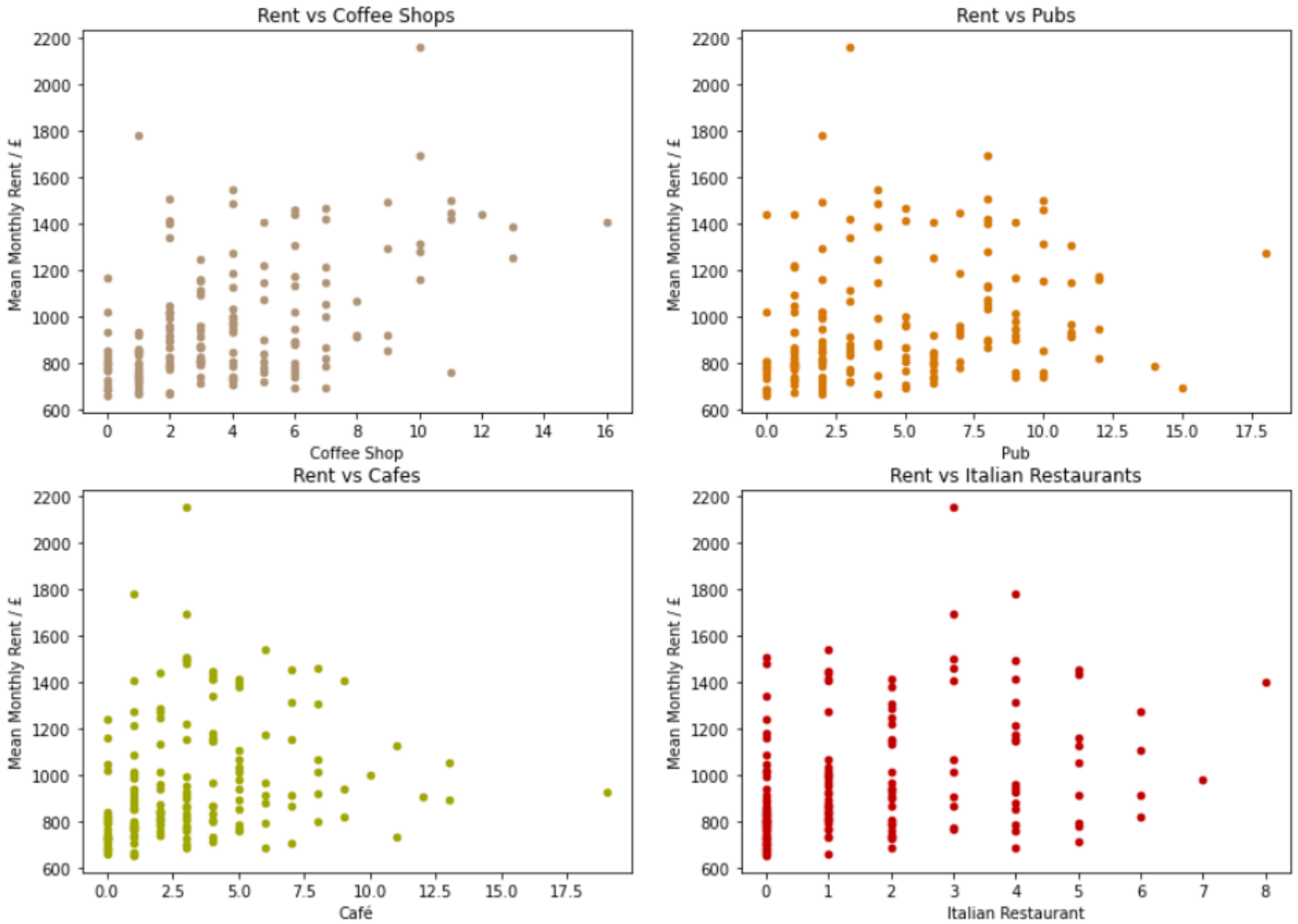
COUNTING VENUES

Identifying the top five venue categories in each postcode district

	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

Rent vs Common Venues

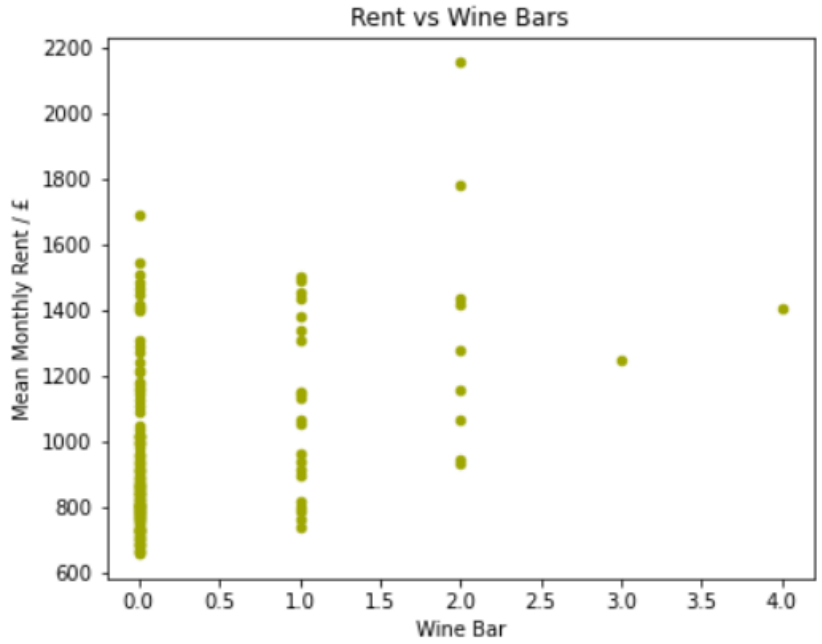
Exploring correlation between common venues and target variable.

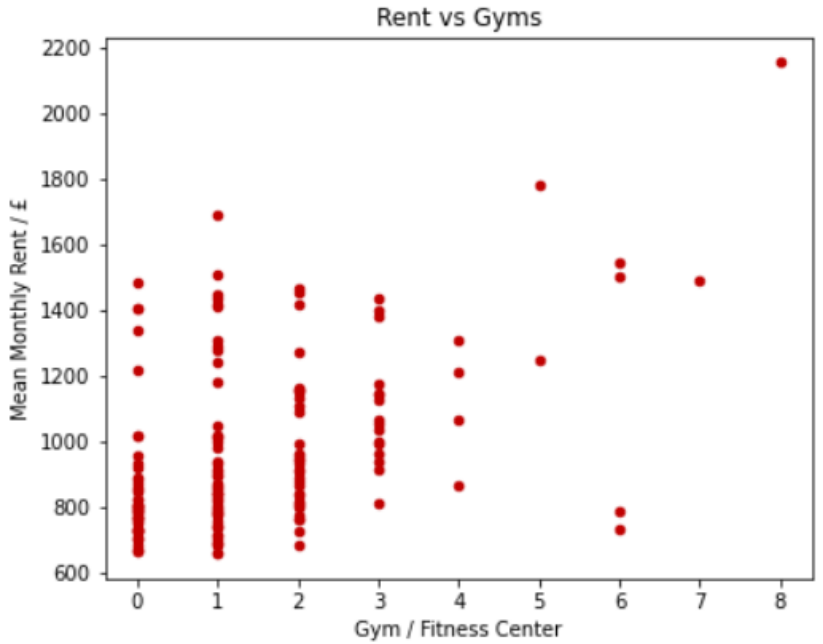


Venue Type	Pearson Correlation	P value
Coffee Shop	0.504530	1.029094e-11
Pub	0.215345	6.243296e-03
Café	0.202145	1.036466e-02
Italian Restaurant	0.332007	1.795047e-05

Rent vs Selected Venues

Exploring correlation between selected venues and target variable.

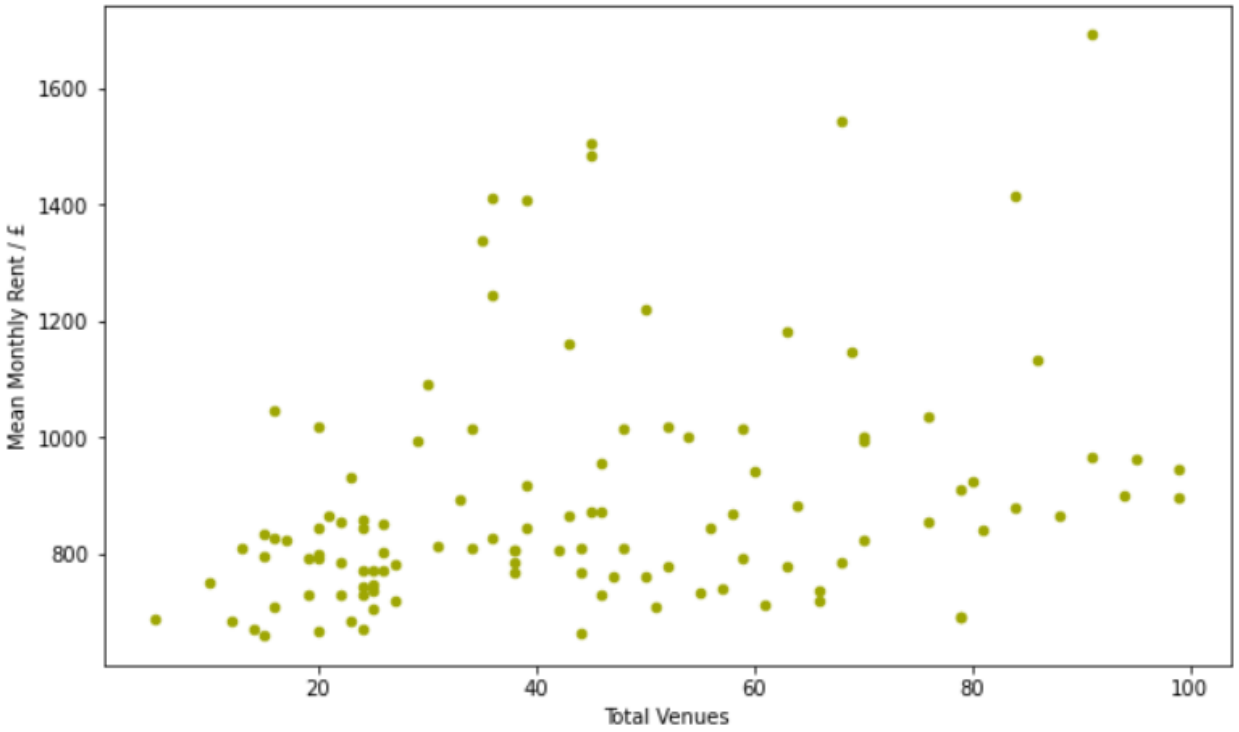




Venue Type	Pearson Correlation	P value
Wine Bar	0.437773	7.086938e-09
Gym / Fitness Center	0.480486	1.275649e-10

Rent vs Total Venues

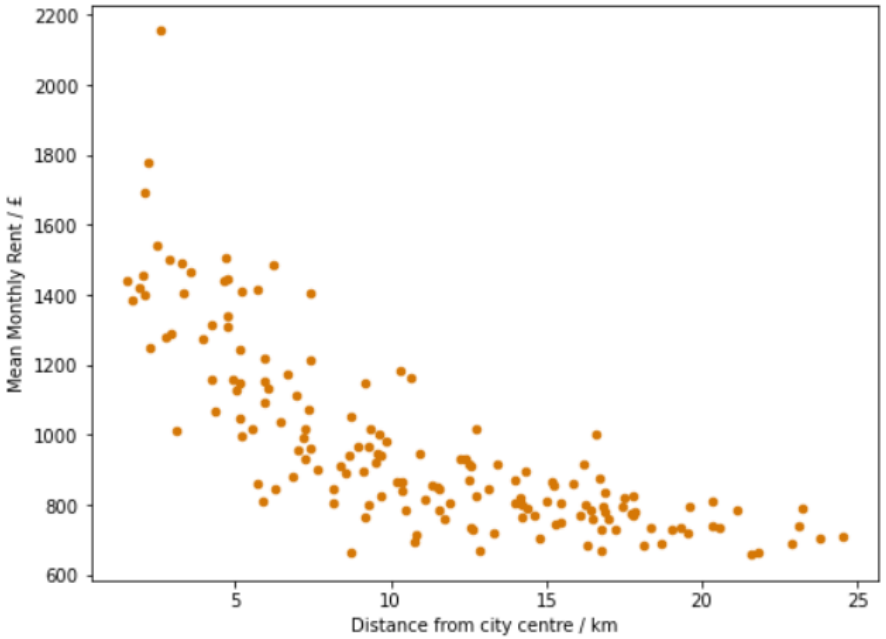
Exploring correlation between total venues and target variable.



Predictive variable	Pearson Correlation	P value
Total Venues	0.311791	0.000733

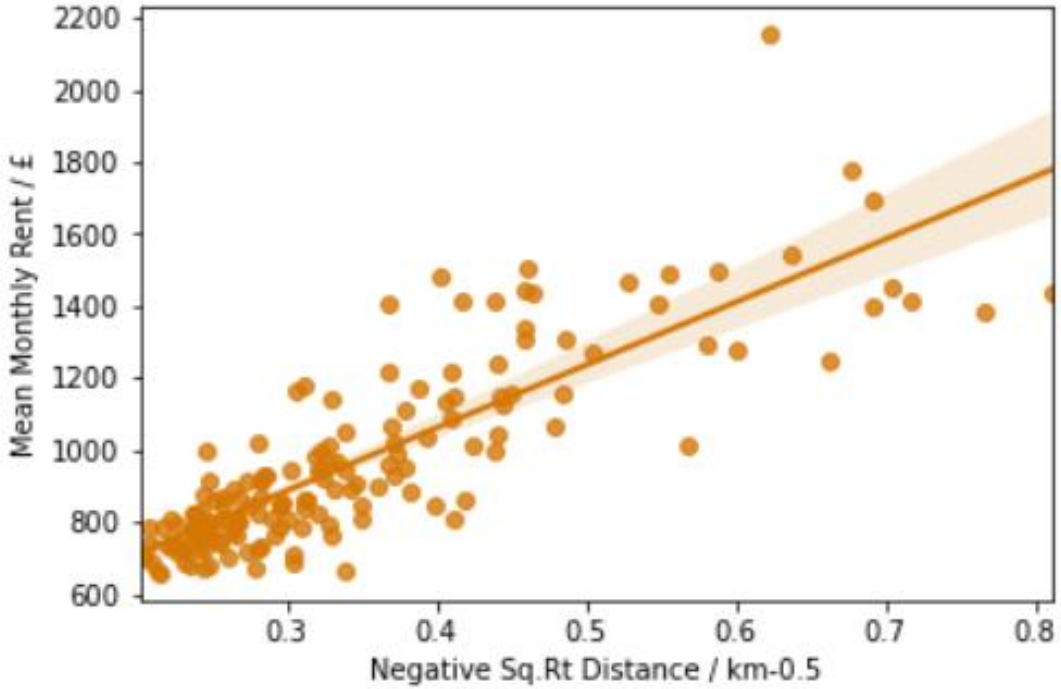
Rent vs Distance from city centre

Exploring correlation between distances and target variable.



$x^{-0.5}$

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



4

The Modelling and Results

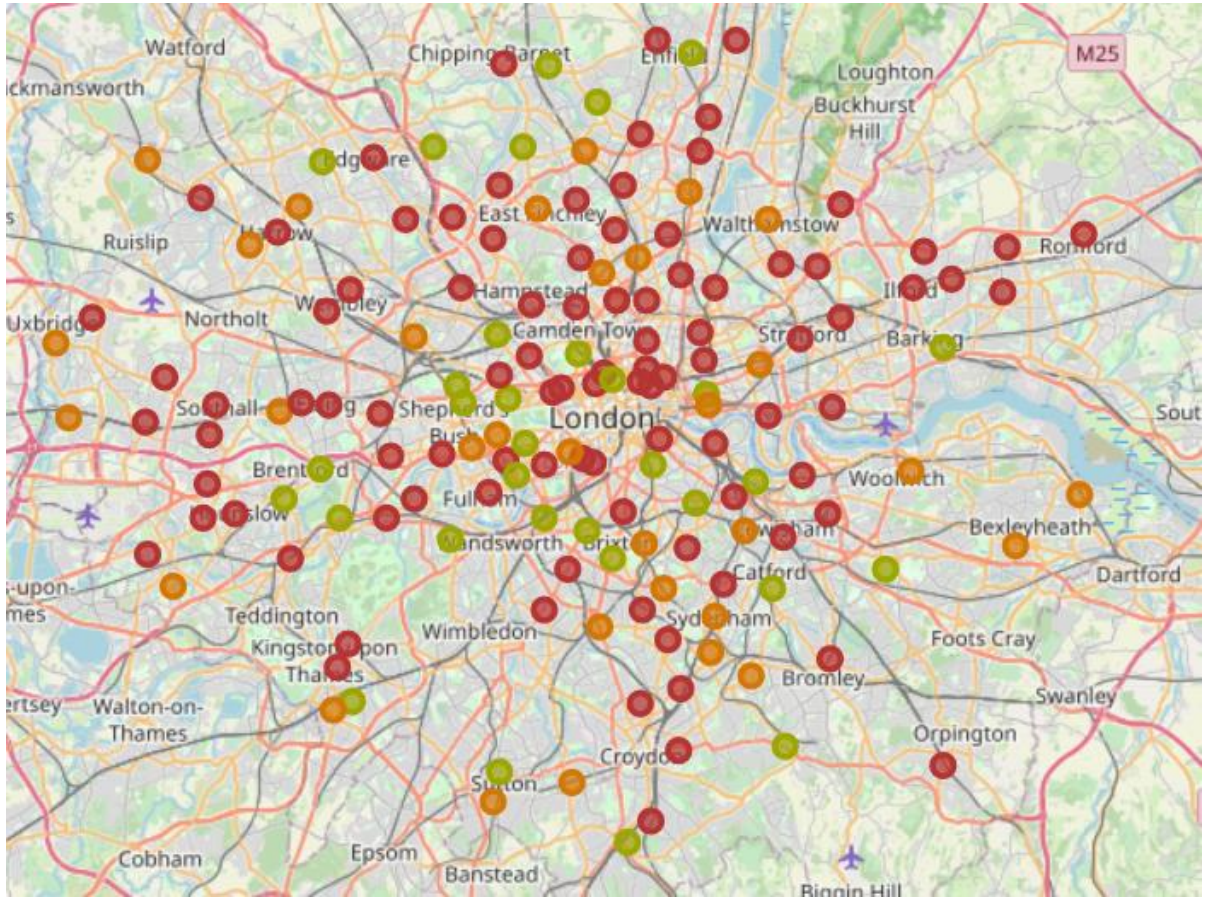
Splitting the Dataset

Creating training, cross-validation and test sets

Train

Cross Validate

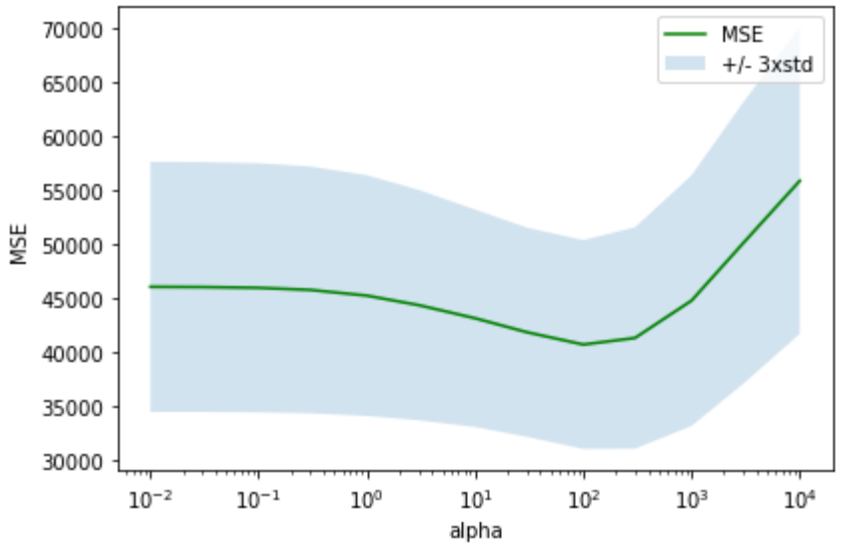
Test



MLR WITH ALL FEATURES

Assessing the MLR model using all venue data

CROSS VALIDATE




```
Lowest MSE value =  40747.0
```

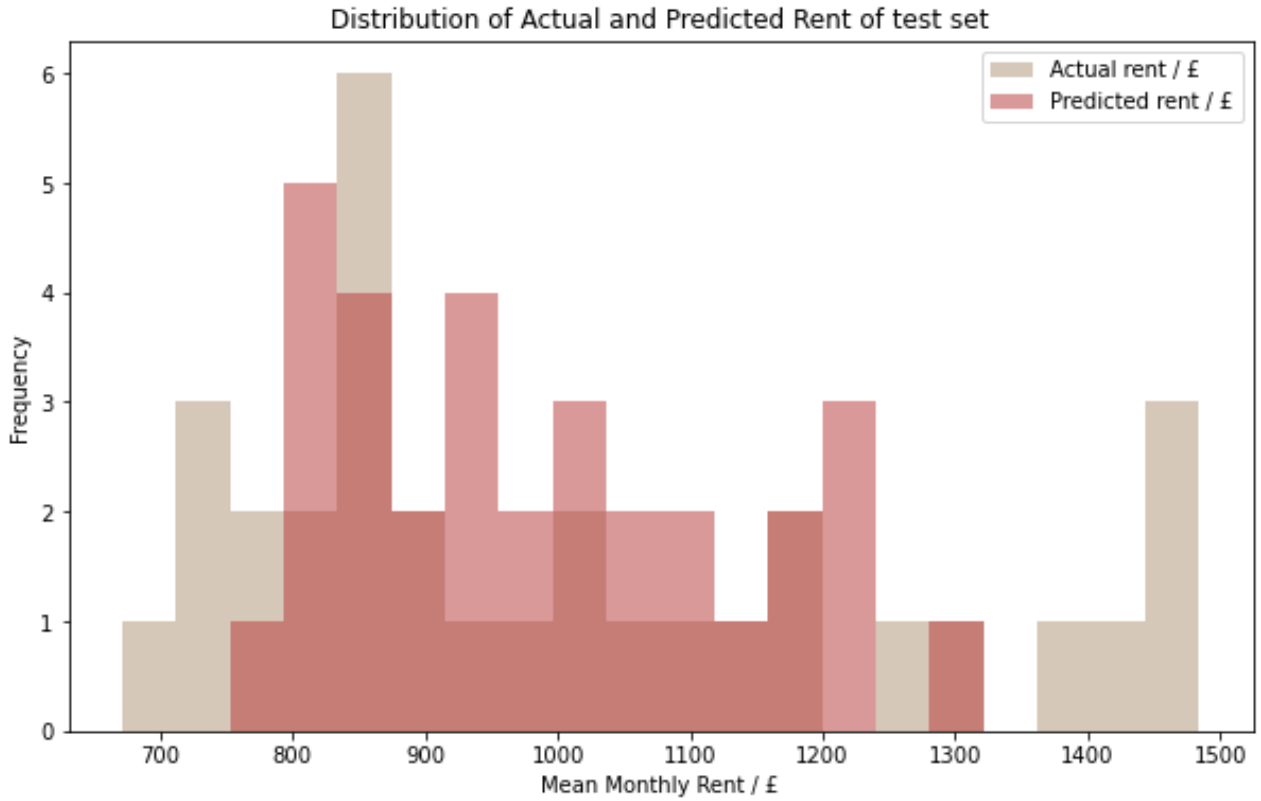
```
Test MSE = 34832.6
```

TEST

	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 Actual vs Predicted Distribution

Assessing the range of target variable predictions



Addressing error magnitude

Reconsidering features used for training model

A Model with all venue features

B Model with no venue features

C Model with selected venue features

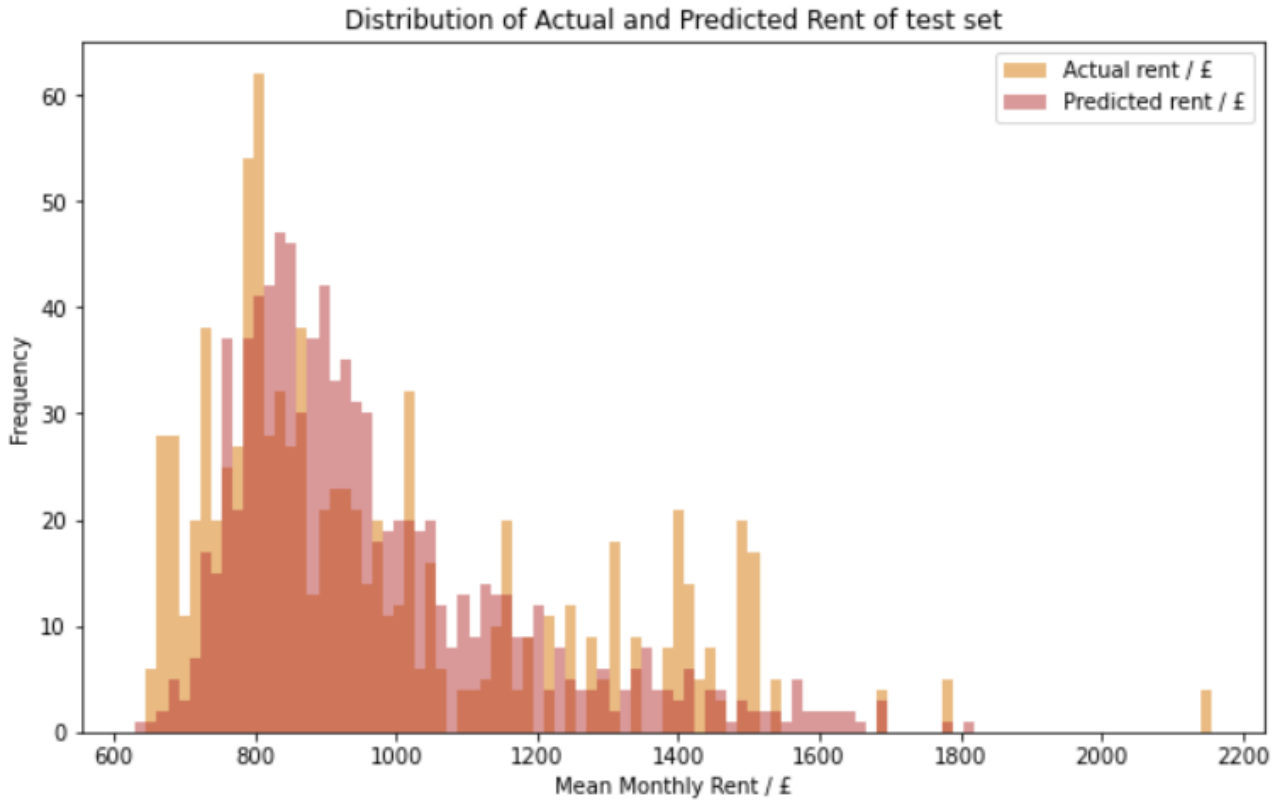
Number of venue features	MSE
384	43907.6
0	24595.2
2	25153.6

Redundant Features?

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

PREDICTED TARGET DISTRIBUTIONS FOR SLR

Assessing the range of target variable predictions for the SLR model over 28 iterations



Conclusions:

- Venue data was detrimental to the accuracy of an MLR model.
- Distance from city center is a key feature for estimating London rent prices at the district level.
- An SLR model would be a good foundation for an application targeting potential tenants exploring the London rental market.

Further Directions:

i. Measure over precise areas
ii. Obtain more complete training examples
iii. Assess venue data more thoroughly
iv. Identify and engineer other features (i.e. demographics, local council data)

Thanks for listening!