CAPSTONE PROJECT
THE BATTLE OF THE NEIGHBORHOODS

Opening a Chinese Restaurant in London

Introduction

- **Business question**

if someone is looking to open a Chinese restaurant in London, can we use data science to help to find out whether a site is good or bad for it?

Is an area with **high density** of exiting Chinese restaurants a good or bad choice?

- **Stakeholders / Target Audience**
  - those who plan to open a Chinese restaurant in London;
  - the landlords of the commercial properties in London.

Data

- The number nearby **Chinese restaurants, similar types of restaurants** and **non-similar types of restaurants**, and their average distance;

  Here similar restaurants refer to Asian type such as Japanese, Korean, Thai, Indian, Malay, etc.

  The above data can be obtained by using the **Foursquare API**.
- **ATM** (i.e. bank automated teller machine) location data obtained from **Overpass API (OpenStreetMap)**.
- **London Underground passenger counts** published by the **London Transportation Department TfL**

TRANSPORT
FOR LONDON

openstreetmap.org

API

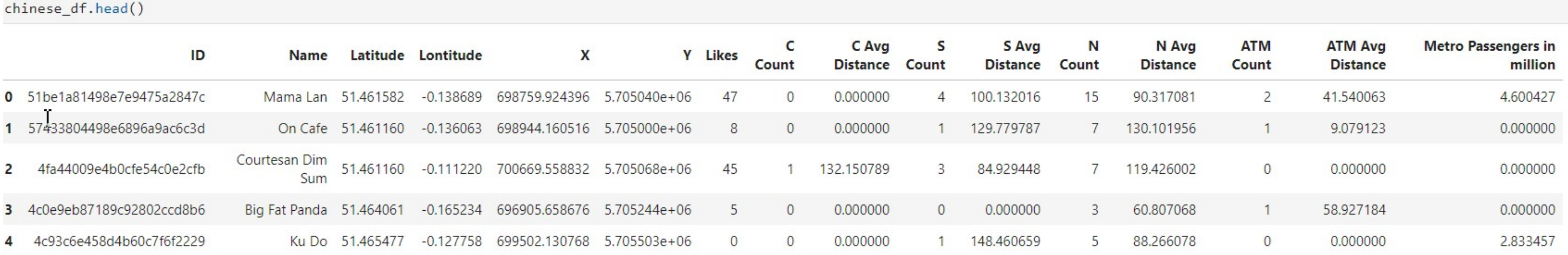
Data

Missing Data Handling

- **GeoPy's geocoder** was used to obtain latitude/longitude of Metro stations.
- Missing data problem was fixed by manually extracting data from **OpenStreetMap**'s website

Data

Marshalling data from different sources into on Pandas dataframe, the first 5 rows of the data are shown here:

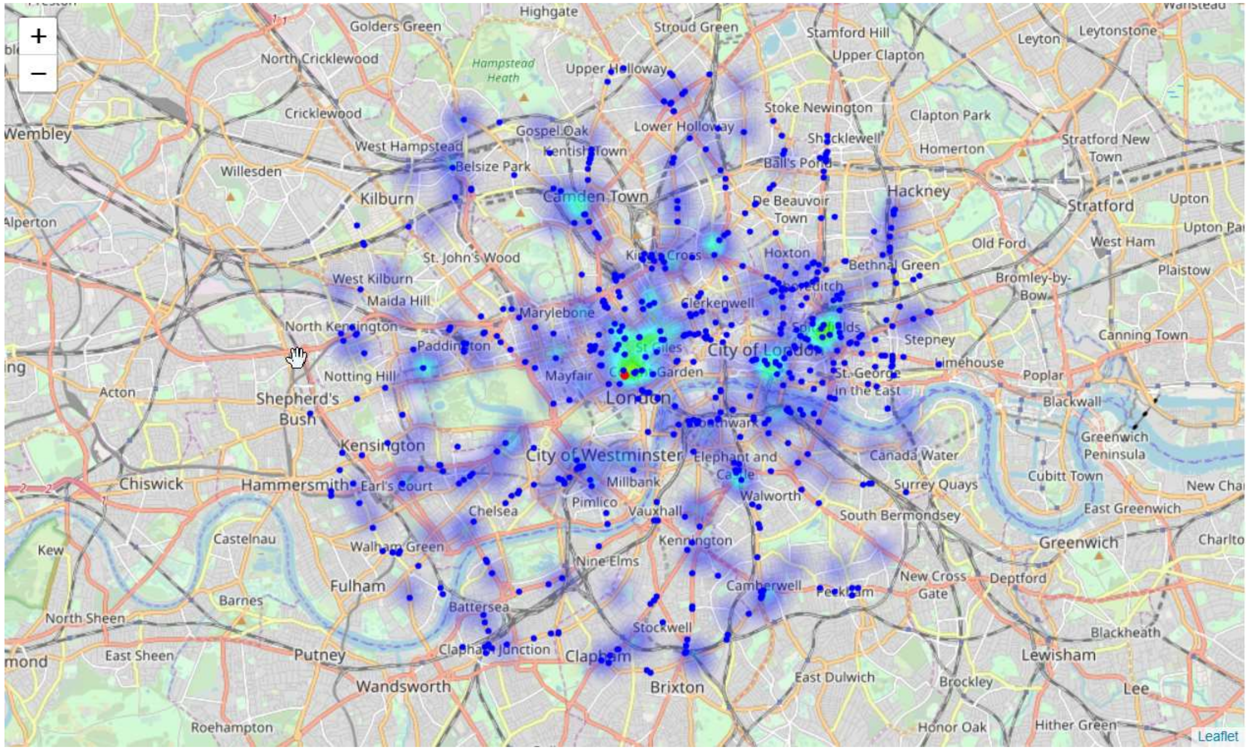


```
chinese_df.head()
```

	ID	Name	Latitude	Lontitude	X	Y	Likes	C Count	C Avg Distance	S Count	S Avg Distance	N Count	N Avg Distance	ATM Count	ATM Avg Distance	Metro Passengers in million
0	51be1a81498e7e9475a2847c	Mama Lan	51.461582	-0.138689	698759.924396	5.705040e+06	47	0	0.000000	4	100.132016	15	90.317081	2	41.540063	4.600427
1	57433804498e6896a9ac6c3d	On Cafe	51.461160	-0.136063	698944.160516	5.705000e+06	8	0	0.000000	1	129.779787	7	130.101956	1	9.079123	0.000000
2	4fa44009e4b0cfe54c0e2cfb	Courtesan Dim Sum	51.461160	-0.111220	700669.558832	5.705068e+06	45	1	132.150789	3	84.929448	7	119.426002	0	0.000000	0.000000
3	4c0e9eb87189c92802ccd8b6	Big Fat Panda	51.464061	-0.165234	696905.658676	5.705244e+06	5	0	0.000000	0	0.000000	3	60.807068	1	58.927184	0.000000
4	4c93c6e458d4b60c7f6f2229	Ku Do	51.465477	-0.127758	699502.130768	5.705503e+06	0	0	0.000000	1	148.460659	5	88.266078	0	0.000000	2.833457

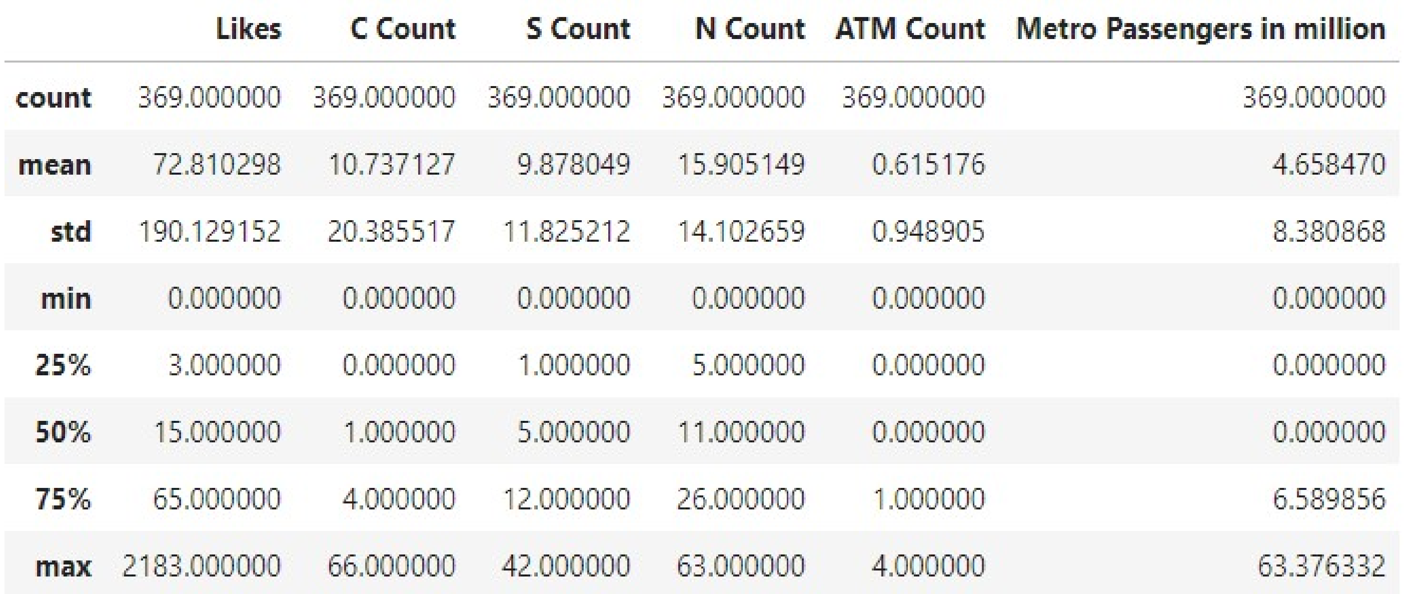
Methodology

Explanatory Analysis - a Heatmap of Chinese restaurants overlaid with ATM to see density and any correlation



Methodology

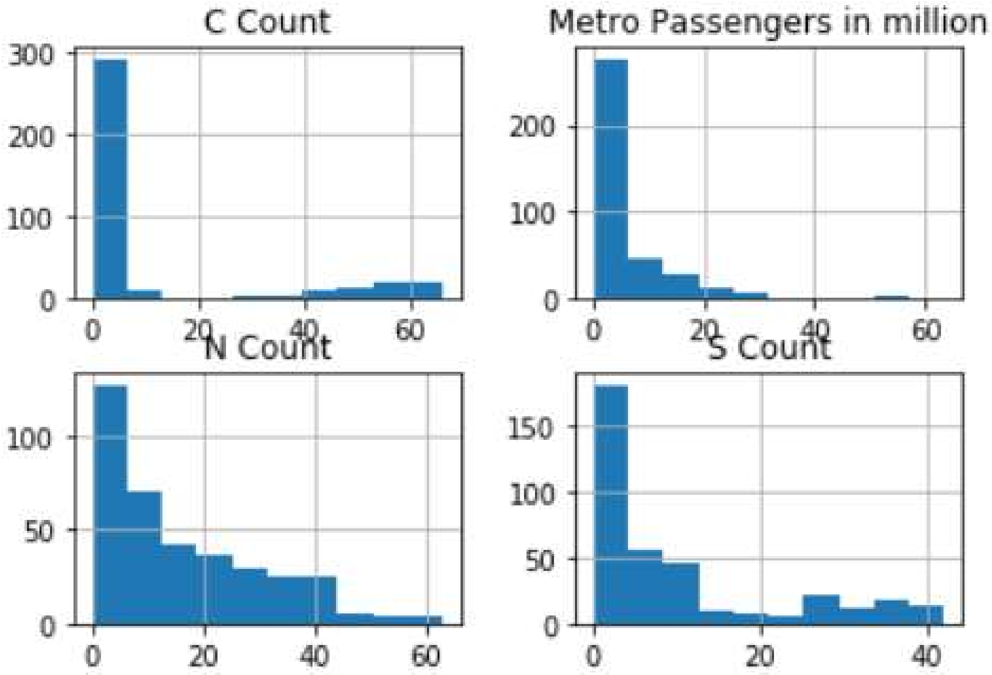
Explanatory Analysis - Descriptive statistics of the dataset



	Likes	C Count	S Count	N Count	ATM Count	Metro Passengers in million
count	369.000000	369.000000	369.000000	369.000000	369.000000	369.000000
mean	72.810298	10.737127	9.878049	15.905149	0.615176	4.658470
std	190.129152	20.385517	11.825212	14.102659	0.948905	8.380868
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	3.000000	0.000000	1.000000	5.000000	0.000000	0.000000
50%	15.000000	1.000000	5.000000	11.000000	0.000000	0.000000
75%	65.000000	4.000000	12.000000	26.000000	1.000000	6.589856
max	2183.000000	66.000000	42.000000	63.000000	4.000000	63.376332

Methodology

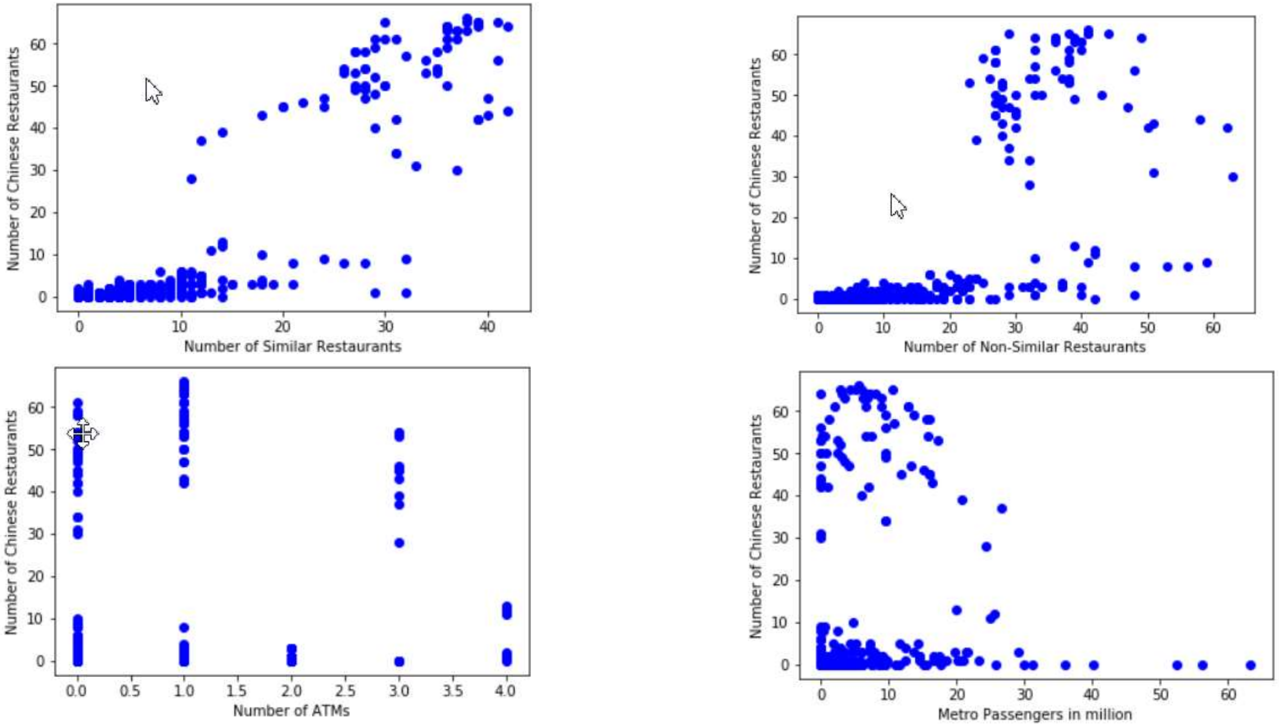
Explanatory Analysis - Histograms of the features



Chinese and similar type of restaurants are in either high or low density but not in-between.

Methodology

Explanatory Analysis – scatter plots of number of nearby Chinese restaurants vs. other features



The Plot of Number of Similar Restaurants vs. Number of Chinese Restaurants shows some linearity.

Methodology

Predictive Modelling

- Fitting the dataset to **multiple linear regression** model:
  - the dependent variable is the number of nearby Chinese restaurants; and
  - The independent variables are
    - the number of nearby similar restaurants and their average distance,
    - the number of nearby non-similar restaurants and their average distance,
    - the number of nearby ATMs and their average distance, and
    - the annual passengers numbers of the nearby Metro stations (weighted with distance)
  - Randomly selected 80% of the dataset is used for training the model.

The resulting model was then tested with remaining 20% of the dataset, the **variance score** of the resulting model is **0.87**, which is **very close to a perfect prediction (score of 1)**

Results

- A multiple linear regression model is then derived using 100% of dataset, and here is the resulting model:
- Coefficients: [2.11546858 -0.03974706 -0.57464468 0.05390643 -0.64187237 0.01702939 0.18028619],
- Intercept: -3.53078879

The variance score of the model is 0.85, which shows the model can be a very good predictive model address the business question set out at the beginning

Results

- To assess whether a given location is suitable for new Chinese restaurant, the model can be used to predict the number of nearby Chinese restaurants for that location:
  - If the number of nearby Chinese restaurants that actually exist is larger than the predicted number, it indicates that the location is 'over-crowded" and may not be suitable for opening yet another new Chinese restaurant;
  - on the other hand, if there are less Chinese restaurants actually exist than predicted, it indicates that location is likely to be able to 'sustain' another Chinese restaurant so the site may be suitable for opening a new Chinese restaurant.

Results

Example:

- Let’s assume the address “119 Newington Causeway, Elephant and Castle, London SE1 6BN” is short listed for opening a new Chinese restaurant.
- Foursquare API shows currently there is no Chinese restaurant nearby, and the predictive model shows 1.74 Chinese restaurant nearby
- This can be interpreted as the location may be a suitable site for opening a new Chinese restaurant

Discussion

- Using better data representing how well a restaurant is performed should make a more compelling predictive model:
  - Such data can be the footfall of the venue. Foursquare's Checkins can be a good source, unfortunately, Foursquare only offers such data to "authorised".
  - Another possible information can be the profit/loss account of the restaurants. However, these records would need to be obtained and interpreted manually, and companies may use accounting tactics to hide the true picture of profit/loss.

Conclusions

The variance score of the resulting multiple linear regression model is 0.85, a score that is very close to a perfect prediction (score of 1), it means this model can provide a very good prediction, and can to be used to address the business question set out at the start of the project.