ST207 Project

London Cafe Database

GitHub repository:
https://github.com/lse-st207/project-at2023-group5.git

Candidate numbers
21630
27065
28717
24852

Introduction

From the Starbucks beside the LSE campus to the Pret across the street, coffee shops are everywhere. Especially in a buzzing city like London, it seems that people couldn’t start their day without a cup of latte, Americano, or a shot of espresso. To understand more about this critical aspect of people’s everyday lives, we hope to collect data about cafes in London to see how it is distributed across different areas and customer opinion towards different coffee shops etc.

Application

Our comprehensive cafe database offers a wealth of information that can be harnessed by three primary end users—consumers, cafe owners, and business users. For consumers, the database provides insights into the highest-rated cafes and their geographical distribution across London, enabling informed decisions about where to enjoy the best coffee experiences. Coffee shop owners can utilise the data to analyse market saturation, identify competitive hotspots, and understand consumer ratings to better position their offerings. Business users, such as potential investors or market analysts, can leverage the database for strategic planning, identifying market trends, and pinpointing areas with potential for development or investment. Together, the database forms a cohesive perspective for the decision-making and strategy development landscape of London's coffee culture.

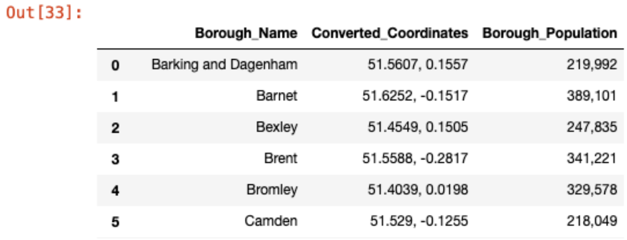
Data Description

London Boroughs and Population

Data Source:

- Borough council coordinates: https://en.wikipedia.org/wiki/List_of_London_boroughs
- Population of each Borough: https://www.statista.com/statistics/381055/london-population-by-borough/

The process began with acquiring the list of London boroughs (including City of London) by web scraping the Wikipedia page titled "List of London boroughs." A script was written to perform an HTTP request to the page and parse the HTML content to extract the table containing the borough data. The extracted data included names, geographic details and others of each borough council, which were then cleaned and processed to isolate the coordinates (latitude and longitude). We also included the population data in 2022 from Statista and combined them in one table, to provide a comprehensive overview of each borough, useful for spatial and demographic analyses.



Out[33]:

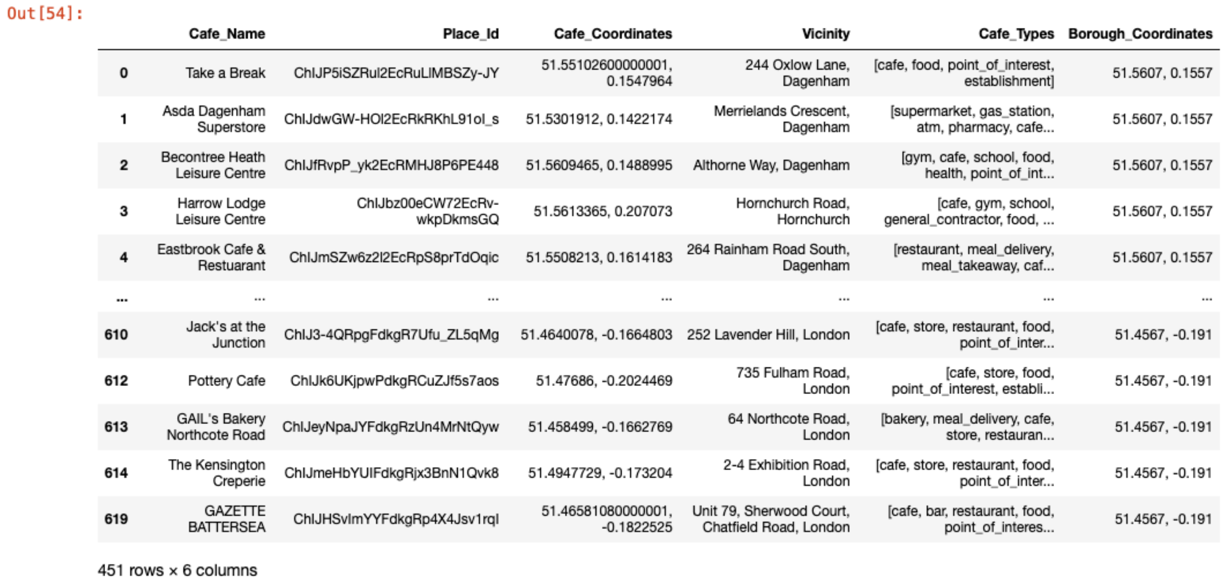
	Borough_Name	Converted_Coordinates	Borough_Population
0	Barking and Dagenham	51.5607, 0.1557	219,992
1	Barnet	51.6252, -0.1517	389,101
2	Bexley	51.4549, 0.1505	247,835
3	Brent	51.5588, -0.2817	341,221
4	Bromley	51.4039, 0.0198	329,578
5	Camden	51.529, -0.1255	218,049

(Figure 1: Borough DataFrame)

Google Map API

- API documentation: https://developers.google.com/maps/documentation/places/web-service

By inputting the coordinates of each London borough council, we used Google Maps API's nearby search (within 5000 metres from each London borough council) to obtain the name, unique identity, geometry, service type, average rating (1.0~5.0), and the total number of ratings of 660 cafes. It is assumed that 5000 metres is the optimum radius to cover every cafe with the least duplications as it is impossible to Lasso Select one particular region in the API. However, because of selection via radius, it leads to potential duplicates. After cleaning up the data to eliminate duplicates using the coordinates, 451 cafes in London were obtained.



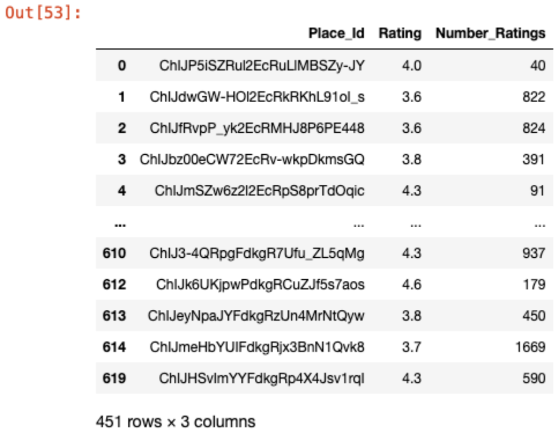
Out[54]:

	Cafe_Name	Place_Id	Cafe_Coordinates	Vicinity	Cafe_Types	Borough_Coordinates
0	Take a Break	ChIJP5iSZRuI2EcRuLIMBSZy-JY	51.55102600000001, 0.1547964	244 Oxlow Lane, Dagenham	[cafe, food, point_of_interest, establishment]	51.5607, 0.1557
1	Asda Dagenham Superstore	ChIJdwGW-HOI2EcRkRKhL91ol_s	51.5301912, 0.1422174	Merrielands Crescent, Dagenham	[supermarket, gas_station, atm, pharmacy, cafe...	51.5607, 0.1557
2	Becontree Heath Leisure Centre	ChIJfRvpP_yk2EcRMHJ8P6PE448	51.5609465, 0.1488995	Althorne Way, Dagenham	[gym, cafe, school, food, health, point_of_int...	51.5607, 0.1557
3	Harrow Lodge Leisure Centre	ChIJbz00eCW72EcRv-wkpDkmsGQ	51.5613365, 0.207073	Hornchurch Road, Hornchurch	[cafe, gym, school, general_contractor, food, ...	51.5607, 0.1557
4	Eastbrook Cafe & Restuarant	ChIJmSZw6z2I2EcRpS8prTdOqic	51.5508213, 0.1614183	264 Rainham Road South, Dagenham	[restaurant, meal_delivery, meal_takeaway, caf...	51.5607, 0.1557
...	...	...	...	...	...	...
610	Jack's at the Junction	ChIJ3-4QRpgFdkgR7Ufu_ZL5qMg	51.4640078, -0.1664803	252 Lavender Hill, London	[cafe, store, restaurant, food, point_of_inter...	51.4567, -0.191
612	Pottery Cafe	ChIJk6UKjpwPdkgRCuZJf5s7aos	51.47686, -0.2024469	735 Fulham Road, London	[cafe, store, food, point_of_interest, establi...	51.4567, -0.191
613	GAIL's Bakery Northcote Road	ChIJeyNpaJYFdkgRzUn4MrNtQyw	51.458499, -0.1662769	64 Northcote Road, London	[bakery, meal_delivery, cafe, store, restauran...	51.4567, -0.191
614	The Kensington Creperie	ChIJmeHbYUIFdkgRjx3BnN1Qvk8	51.4947729, -0.173204	2-4 Exhibition Road, London	[cafe, store, restaurant, food, point_of_inter...	51.4567, -0.191
619	GAZETTE BATTERSEA	ChIJHSvlmYYFdkgRp4X4Jsv1rql	51.46581080000001, -0.1822525	Unit 79, Sherwood Court, Chatfield Road, London	[cafe, bar, restaurant, food, point_of_interes...	51.4567, -0.191

451 rows × 6 columns

(Figure 2: Cafe DataFrame)

Separating the rating data allows for more flexibility in case of database expansion or modification. As the application grows or evolves, we might need to add new attributes or entities related to ratings (like historical rating trends, peak rating periods, etc.). Having a separate table for ratings makes such additions easier and more organised.



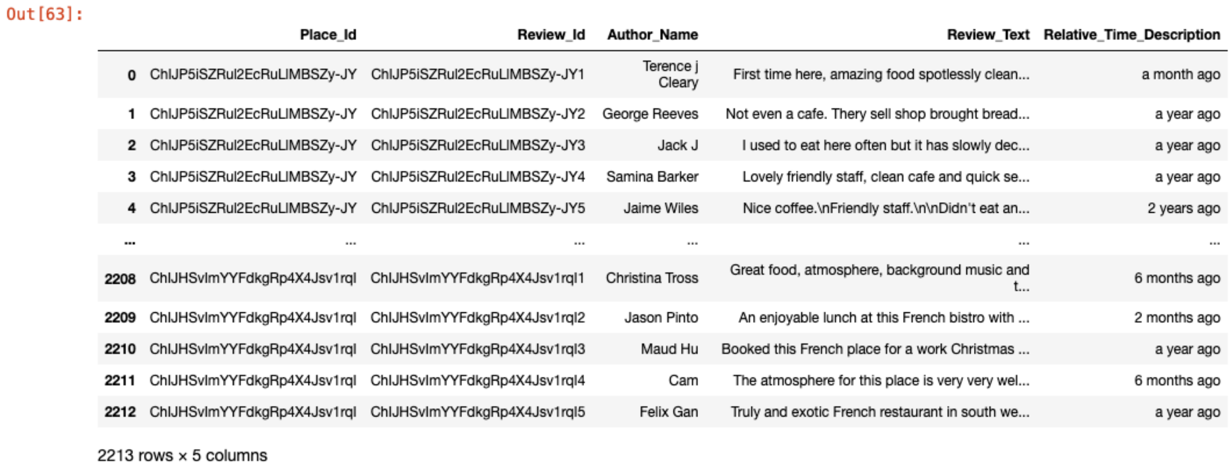
Out[53]:

	Place_Id	Rating	Number_Ratings
0	ChIJP5iSZRuI2EcRuLIMBSZy-JY	4.0	40
1	ChIJdwGW-HOI2EcRkRKhL91ol_s	3.6	822
2	ChIJfRvpP_yk2EcRMHJ8P6PE448	3.6	824
3	ChIJbz00eCW72EcRv-wkpDkmsGQ	3.8	391
4	ChIJmSZw6z2I2EcRpS8prTdOqic	4.3	91
...	...	...	...
610	ChIJ3-4QRpgFdkgR7Ufu_ZL5qMg	4.3	937
612	ChIJk6UKjpwPdkgRCuZJf5s7aos	4.6	179
613	ChIJeyNpaJYFdkgRzUn4MrNtQyw	3.8	450
614	ChIJmeHbYUIFdkgRjx3BnN1Qvk8	3.7	1669
619	ChIJHSvlmYYFdkgRp4X4Jsv1rql	4.3	590

451 rows × 3 columns

(Figure 3: Ratings DataFrame)

Furthermore, the Google Maps API's **Place Details** function provides in-depth information about a specific place, including the top 5 user reviews (ranked by relevance, 5 most informative and helpful for users looking for insights). We focused on extracting comprehensive review data for cafes within London. Utilising the Place ID of each cafe obtained from the initial nearby search, and then employed the function of Place Details. For each of the 451 unique cafes identified in our dataset, we requested detailed information, including reviews by inputting the unique Place ID. For each review, we also generated a unique review identification as a complement to our real datasets. This process resulted in a dataset of user-generated reviews for each cafe in our list, providing valuable insights into customer experiences and preferences.



Out[63]:

	Place_Id	Review_Id	Author_Name	Review_Text	Relative_Time_Description
0	ChIJP5iSZRuI2EcRuLIMBSZy-JY	ChIJP5iSZRuI2EcRuLIMBSZy-JY1	Terence j Cleary	First time here, amazing food spotlessly clean...	a month ago
1	ChIJP5iSZRuI2EcRuLIMBSZy-JY	ChIJP5iSZRuI2EcRuLIMBSZy-JY2	George Reeves	Not even a cafe. Thery sell shop brought bread...	a year ago
2	ChIJP5iSZRuI2EcRuLIMBSZy-JY	ChIJP5iSZRuI2EcRuLIMBSZy-JY3	Jack J	I used to eat here often but it has slowly dec...	a year ago
3	ChIJP5iSZRuI2EcRuLIMBSZy-JY	ChIJP5iSZRuI2EcRuLIMBSZy-JY4	Samina Barker	Lovely friendly staff, clean cafe and quick se...	a year ago
4	ChIJP5iSZRuI2EcRuLIMBSZy-JY	ChIJP5iSZRuI2EcRuLIMBSZy-JY5	Jaime Wiles	Nice coffee.\nFriendly staff.\n\nDidn't eat an...	2 years ago
...	...	...	...	...	...
2208	ChIJHSvlmYYFdkgRp4X4Jsv1rqI	ChIJHSvlmYYFdkgRp4X4Jsv1rqI1	Christina Tross	Great food, atmosphere, background music and t...	6 months ago
2209	ChIJHSvlmYYFdkgRp4X4Jsv1rqI	ChIJHSvlmYYFdkgRp4X4Jsv1rqI2	Jason Pinto	An enjoyable lunch at this French bistro with ...	2 months ago
2210	ChIJHSvlmYYFdkgRp4X4Jsv1rqI	ChIJHSvlmYYFdkgRp4X4Jsv1rqI3	Maud Hu	Booked this French place for a work Christmas ...	a year ago
2211	ChIJHSvlmYYFdkgRp4X4Jsv1rqI	ChIJHSvlmYYFdkgRp4X4Jsv1rqI4	Cam	The atmosphere for this place is very very wel...	6 months ago
2212	ChIJHSvlmYYFdkgRp4X4Jsv1rqI	ChIJHSvlmYYFdkgRp4X4Jsv1rqI5	Felix Gan	Truly and exotic French restaurant in south we...	a year ago

2213 rows × 5 columns

(Figure 4: Review DataFrame)

After that, we exported all Pandas DataFrame to csv files and allowed them to be imported to the DB Browser for further research and analysis.

Limitations and Justification for Data Sources and Usage

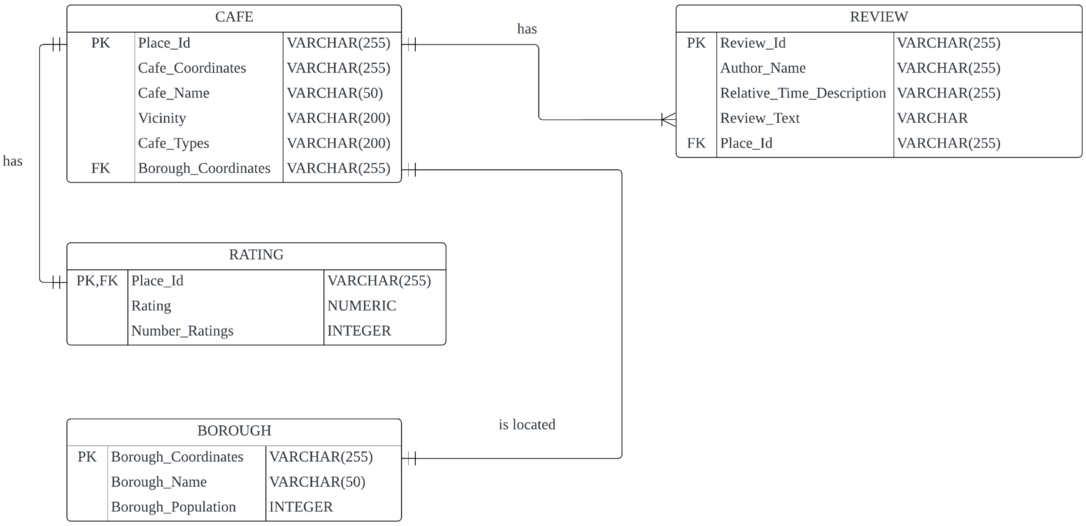
Using the borough council coordinates to select all coffee shops within 5km may not be the best way to cover all cafes in London as it may not comprehensively capture all cafes in London. The 5km radius was chosen to balance coverage and precision, but it inherently has limitations. The density of cafes can vary significantly across different boroughs according to London's varied urban density, which may lead to uneven representation of cafes, with some areas being overrepresented and others underrepresented in the data. For further research, adapting the radius based on population density and urban geography can be considered.

The Google Maps API categorises entities based on types like 'cafe,' 'restaurant,' etc. However, this classification may not always align perfectly with the primary function of the establishment. Integrating data from other sources could help cross-verify and supplement the information obtained from the Google Maps API. For example, the highly popular ‘Pret’ shop is listed as a ‘sandwich’ establishment instead of a ‘cafe’.

Database Creation/Design

ER diagram

We obtained the following ER diagram by using the Lucid Chart:



CAFE

PK	Place_Id	VARCHAR(255)
	Cafe_Coordinates	VARCHAR(255)
	Cafe_Name	VARCHAR(50)
	Vicinity	VARCHAR(200)
	Cafe_Types	VARCHAR(200)
FK	Borough_Coordinates	VARCHAR(255)

REVIEW

PK	Review_Id	VARCHAR(255)
	Author_Name	VARCHAR(255)
	Relative_Time_Description	VARCHAR(255)
	Review_Text	VARCHAR
FK	Place_Id	VARCHAR(255)

RATING

PK,FK	Place_Id	VARCHAR(255)
	Rating	NUMERIC
	Number_Ratings	INTEGER

BOROUGH

PK	Borough_Coordinates	VARCHAR(255)
	Borough_Name	VARCHAR(50)
	Borough_Population	INTEGER

(Figure 5: ER Diagram for London Cafes Dataset)

Entities

CAFE:

- **Place_Id** (Primary Key): a textual identifier that uniquely identifies a place.
- **Cafe_Coordinates**: geographical coordinates of the cafe, formatted as latitude and longitude
- **Cafe_Name**: the exact name of each cafe.
- **Vicinity**: simplified address for the place, including the street name, street number, and locality, but not the province/state, postal code, or country.
- **Cafe_Types**: an array of feature types describing the given cafe, which can also be seen as the service range.
- **Borough_Coordinates**: geographical coordinates of the borough council where the cafe belongs, formatted as latitude and longitude.

BOROUGH:

- **Borough_Coordinates** (Primary Key)
- **Borough_Name:** the exact name of each borough in London.
- **Borough_Population:** population in year of 2022.

RATING:

- **Place_Id** (Primary Key)
- **Rating**: contains the place's rating, from 1.0 to 5.0, based on aggregated user reviews.
- **Number_Ratings**: The total number of reviews for the selected cafe.

REVIEW:

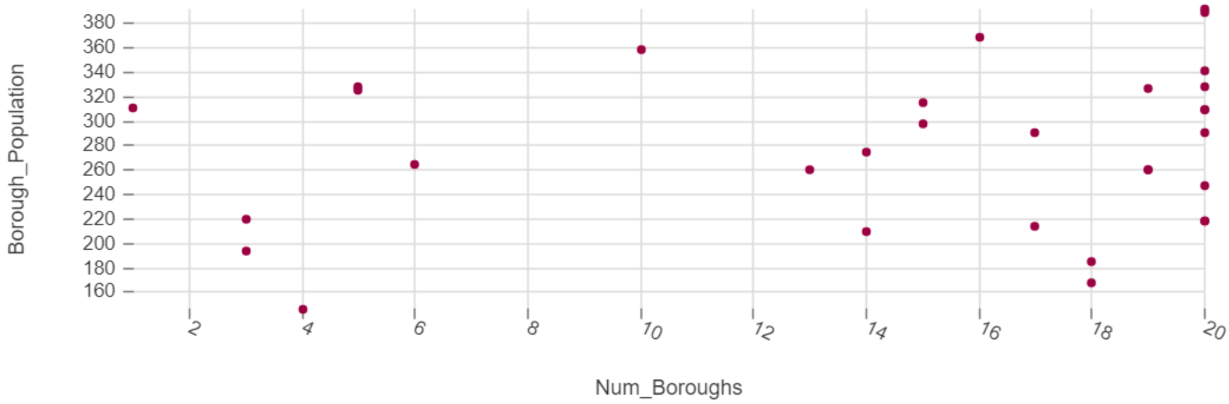
- **Review_Id** (Primary Key): synthetic ID with Place ID as the prefix and the ranking as the suffix.
- **Author_Name**: the name of the user who submitted the review. Anonymous reviews are attributed to "A Google user".
- **Review_Text**: the user's review content.

- **Relative_Time_Description**: the time that the review was submitted in text, relative to the current time.
- **Place_ID** (Foreign Key)

Data Operations

The SQL queries are stored in the following file:
https://github.com/lse-st207/project-at2023-group5/blob/910a5f08b9ad2bbc3187bf444ca1c0826a9aa41f/query.sql
*The instructions for the database technology are in the appendix

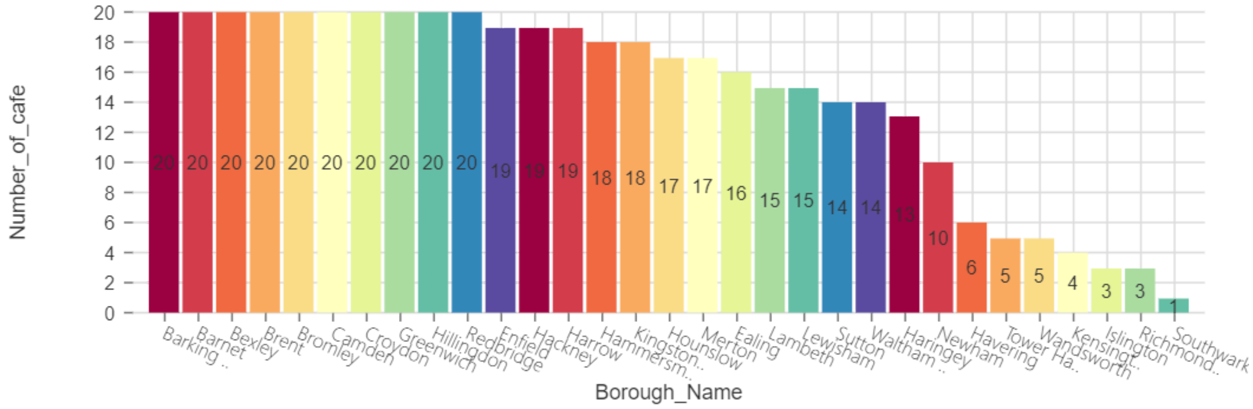
Query 1: How does population distribution relate to the number of coffee shops in each borough?
Standing at the perspective of a user who is a potential entrant/owner of the cafe market, the query helps in understanding the market saturation in various boroughs. If a borough with a high population has fewer cafes, it might represent an untapped market or an opportunity for future expansion.



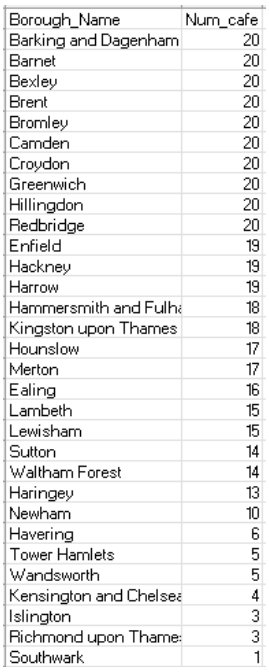
(Figure 6: Kinetica)

The query operates by joining two tables: borough_entity and cafe_entity. The join is made on a common element where the Converted_Coordinates of a borough in the borough_entity table matches the Borough_Coordinates in the cafe_entity table. The outcome of this query is grouped by the Borough_Name. For each borough, it calculates two main pieces of information: the count of cafes and the mean (average) population. The count is determined by the number of times a borough's name appears in the joined table, indicating the number of cafes in that borough. As a result, we found that there is no strong correlation between population distribution and the number of cafes in each borough.

Query 2: Which borough has the highest number of cafes?
From a user perspective, particularly for business owners or potential entrants in the cafe industry, this query provides valuable insights into the most saturated markets. They may choose to avoid such boroughs when they choose the location.



(Figure 7: Kinetica)

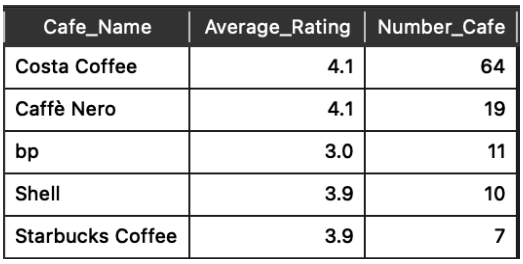


Borough_Name	Num_cafe
Barking and Dagenham	20
Barnet	20
Bexley	20
Brent	20
Bromley	20
Camden	20
Croydon	20
Greenwich	20
Hillingdon	20
Redbridge	20
Enfield	19
Hackney	19
Harrow	19
Hammersmith and Fulha	18
Kingston upon Thames	18
Hounslow	17
Merton	17
Ealing	16
Lambeth	15
Lewisham	15
Sutton	14
Waltham Forest	14
Haringey	13
Newham	10
Havering	6
Tower Hamlets	5
Wandsworth	5
Kensington and Chelsea	4
Islington	3
Richmond upon Thame:	3
Southwark	1

This query involves two primary tables: borough_entity and cafe_entity, which are joined through the coordinates in both entities. This ensures that the cafes are correctly associated with their respective boroughs. The process involves grouping the data by Borough_Name, which consolidates the information for each borough. The query then counts the occurrences of each Borough_Name in the joined dataset, representing the number of cafes in that borough. Finally, the results are sorted in descending order based on this count, positioning Barking and Dagenham, Barnet, Bexley, Brent, Bromley, Camden, Croydon, Greenwich, Hillingdon, and Redbridge as the boroughs with the highest concentration of cafes.

(Figure 8: csv file exported from Kinetica)

Query 3: What are the most common cafes in London, and what is their average rating? It allows consumers to easily identify and compare popular chain cafes in London, giving them a sense of which widely available cafe options are likely to provide the best overall customer experience based on average ratings.



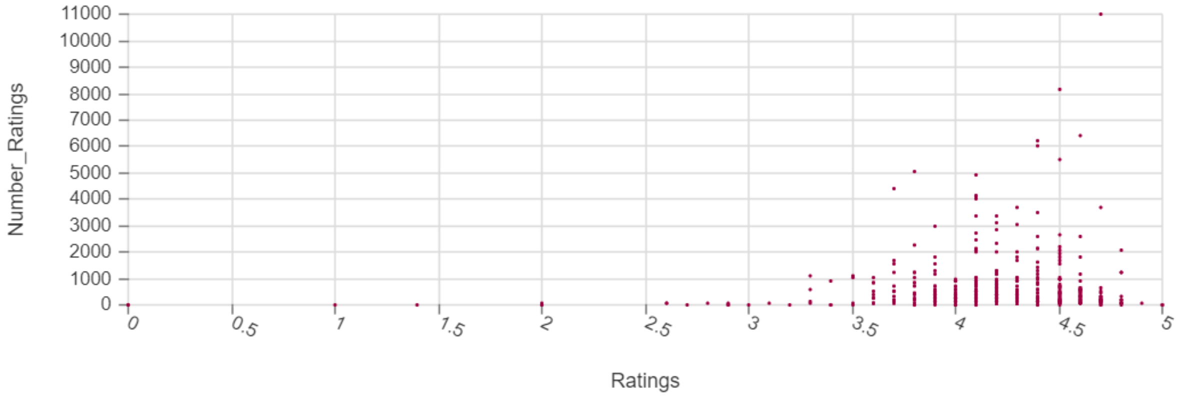
Cafe_Name	Average_Rating	Number_Cafe
Costa Coffee	4.1	64
Caffè Nero	4.1	19
bp	3.0	11
Shell	3.9	10
Starbucks Coffee	3.9	7

(Figure 9: DB Browser result)

The query uses a join operation on the CAFE and RATING entities to match each rating to its respective location, and a GROUP BY operation is used to consider the cafes with the same name. Here it is assumed that all cafes with the same name are chain cafes. Then, the average rating is calculated and filtered to show the average rating of the most common cafes, sorted by descending number of locations. This query is useful for quickly identifying the most common shops and their respective score.

Query 4: How does a cafe rating from 1 to 5 correlate to the total number of ratings for each cafe?

This query provides useful insight for a consumer looking to try out new cafes, as it highlights the tendency for cafes to be rated positively more often than not. The input parameters consist of two main variables: Rating, which represents the individual score each cafe has received, and Number_Ratings, which indicates how many ratings each cafe has accumulated. Since no additional filters are applied, the data encompasses all available ratings for cafes in the dataset.

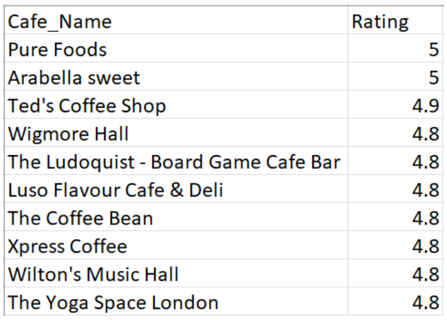


(Figure 10: Kinetica result)

The graph suggests that cafes with ratings closer to 5 have a greater number of ratings, indicative of a positive skew in customer satisfaction. Additionally, the graph might also hint at a sort of normal distribution, with the majority of ratings clustering around the higher middle range, tapering off towards the extreme low and high ends.

Query 5: What are the top 10 highest-rated cafes in London?

As a consumer in London looking for an exceptional coffee experience, this SQL query is a valuable resource to discover the top 10 highest-rated cafes in the city.



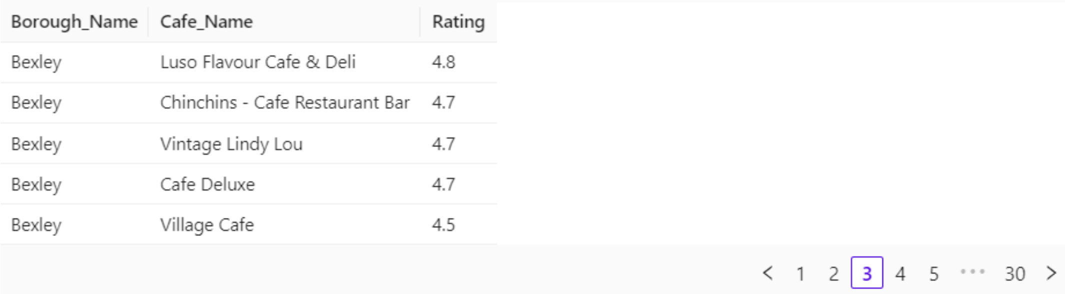
Cafe_Name	Rating
Pure Foods	5
Arabella sweet	5
Ted's Coffee Shop	4.9
Wigmore Hall	4.8
The Ludoquist - Board Game Cafe Bar	4.8
Luso Flavour Cafe & Deli	4.8
The Coffee Bean	4.8
Xpress Coffee	4.8
Wilton's Music Hall	4.8
The Yoga Space London	4.8

(Figure 11: csv file exported Kinetica)

By joining the cafe_kinetica and rating_entity tables, the query effectively pairs each cafe with its respective rating, ensuring an accurate portrayal of customer satisfaction and quality. The selection of Cafe_Name and Rating is sorted in descending order, showing the most highly rated cafes. The limitation to just 10 results makes the list concise and focused, highlighting only the very best as per customer ratings, featuring standout cafes are Pure Foods, Arabella sweet, and Ted's Coffee Shop.

Query 6: What are the top-rated cafes in different boroughs?

Considering that users may not want to travel a long distance just to visit the highest-rating cafe in London, this query allows users to input the borough they are living in and returns the 5 highest-rated cafes in each borough.

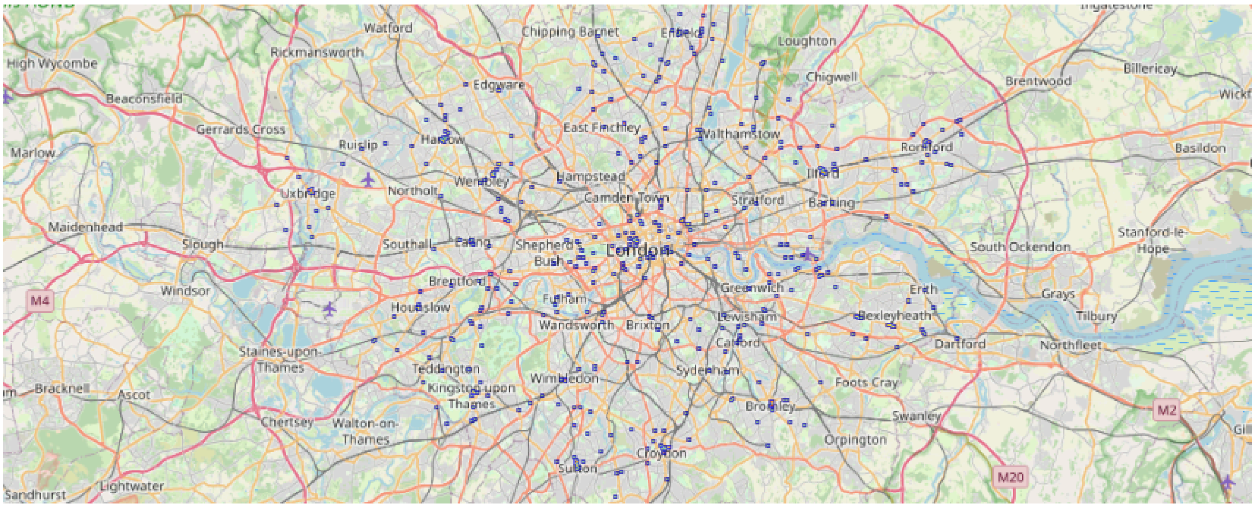


Borough_Name	Cafe_Name	Rating
Bexley	Luso Flavour Cafe & Deli	4.8
Bexley	Chinchins - Cafe Restaurant Bar	4.7
Bexley	Vintage Lindy Lou	4.7
Bexley	Cafe Deluxe	4.7
Bexley	Village Cafe	4.5

(Figure 12: Kinetica result)

We employ a common table expression (CTE) named RankedCafes to organise the data. Within this CTE, the borough_entity, cafe_entity, and rating_entity tables are joined based on borough coordinates and cafe place IDs. For each borough, cafes are ranked based on their ratings in descending order. The main SELECT statement then retrieves the name of the borough, the cafe name, and its rating from the RankedCafes CTE, but only for those cafes that rank in the top five within their respective borough. The final Kinetica result above shows one of the boroughs, in which there are 33 boroughs in total.

From Kinetica: Map of the Distribution of Cafes Across London



(Figure 13: Visualisation of distribution via Kinetica)

From the concentration of points, we can infer that there are more cafes clustered in the centre of the London area, which could also be the centre of the city and the popular part with higher foot traffic and commercial activity.

Justification for Data Operation

For our database, we used DB Browser for SQLite and Kinetica. While DB Browser allows us to create triggers and views, Kinetica supports SQL queries and leverages the power of Graphics Processing Units (GPUs). This not only allows users to easily visualise the cafes across London on a map but also allows users to access cafe information in a specific borough.

Conclusion

In conclusion, our exploration of the cafe landscape in London has provided valuable insights into the city's caffeinated culture. Our cafe database aimed to unravel the nuances of this cultural phenomenon, addressing key questions about cafe distribution, consumer preferences, and variations across different boroughs.

With the inquiry into borough-specific data, we concluded no strong correlation between population distribution and the number of cafes in each borough and summarised the areas with the highest density of cafes. By examining cafe numbers and ratings, we gained valuable perspectives on how cafes are distributed in London and which cafes are most favoured. We found the top 5 most common cafes and the top 10 cafes with the highest ratings. Also, we visualised how cafe ratings from 1 to 5 correlate to the total number of ratings for each cafe.

Furthermore, comparing the ratings of the top 5 most common cafes and the top 5 highest-rated ones, it's notable that one of the most common cafes is not present in the highest-rated group. This indicates that while visiting a chain coffee shop may guarantee a customer a satisfactory experience, it is the independent shops that are likely to offer the best quality.

Lastly, our recommendation query provides a practical tool for customers seeking top-rated cafes within their borough. By inputting their borough, customers can receive a tailored list of the five highest-rated cafes, enhancing their cafe experience and fostering customer loyalty. The inclusion of a visual representation through a map visualisation of cafe distribution further adds an engaging dimension to our analysis. By utilising longitude and latitude data, stakeholders can visually assess the geographical spread of cafes, aiding in location-based decision-making and identifying gaps in the current market of cafes across London.

As cafes continue to serve as social hubs and indispensable morning rituals, our findings contribute to a deeper understanding of how these establishments are woven into the fabric of London's daily life. Although the database still has its limitations, this endeavour serves as a starting point for ongoing research, fostering a greater appreciation for the diverse and dynamic coffee culture that permeates the bustling streets of London.

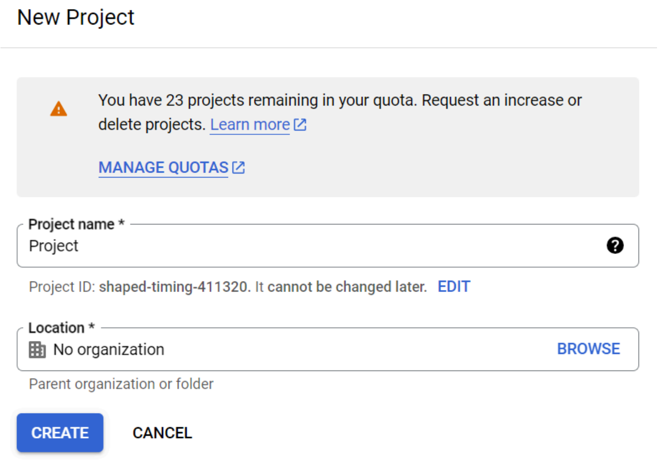
Appendix

Reproducibility

The following provides step-by-step instructions to obtain data and utilise the database tools.

Google Map API:

1. Create Google account
2. Create a project in the Google Cloud console


New Project

You have 23 projects remaining in your quota. Request an increase or delete projects. Learn more

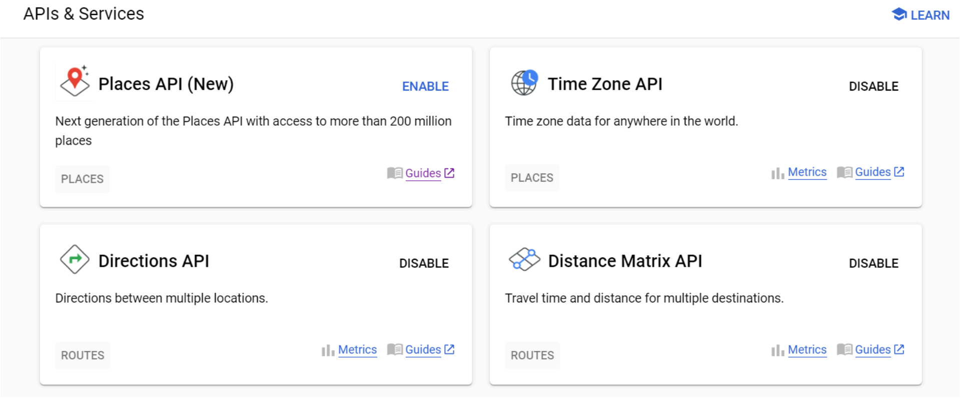
Project name *: Project

Project ID: shaped-timing-411320. It cannot be changed later.

Location *: No organization

Parent organization or folder

3. Select APIs that you want to enable


APIs & Services

Places API (New) — Next generation of the Places API with access to more than 200 million places — PLACES

Time Zone API — Time zone data for anywhere in the world. — PLACES

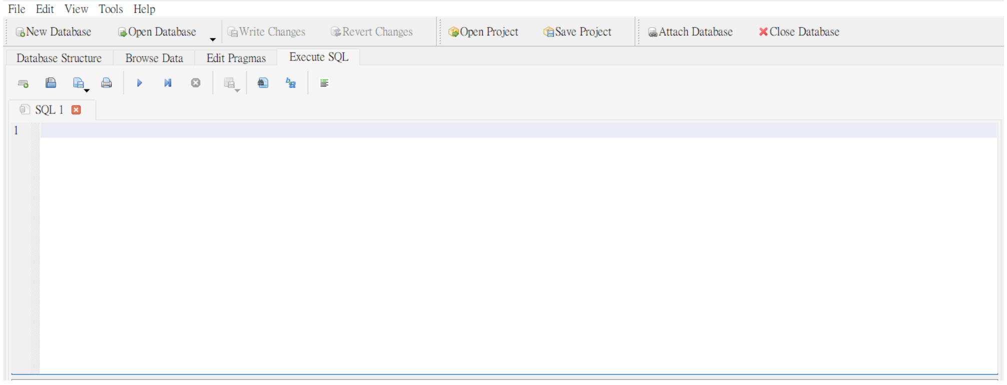
Directions API — Directions between multiple locations. — ROUTES

Distance Matrix API — Travel time and distance for multiple destinations. — ROUTES

4. Go to Credentials page and select your project to obtain your API key
5. Browse through Google API documentation and utilise the API in your preferred coding language

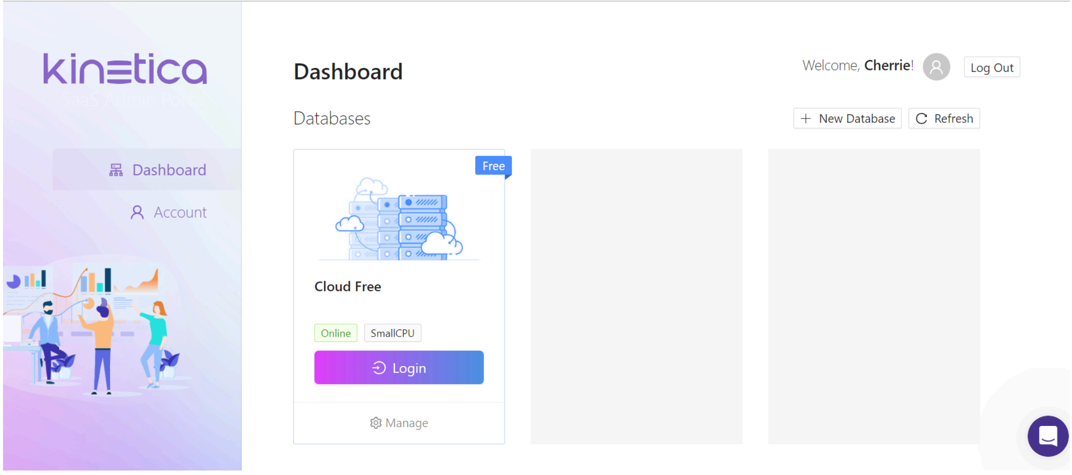
DB Browser:

1. Download DB Browser
2. Click “Open Database” and select Cafe_Databse.db file to load the data
3. Run SQL queries in the “Execute SQL” tab

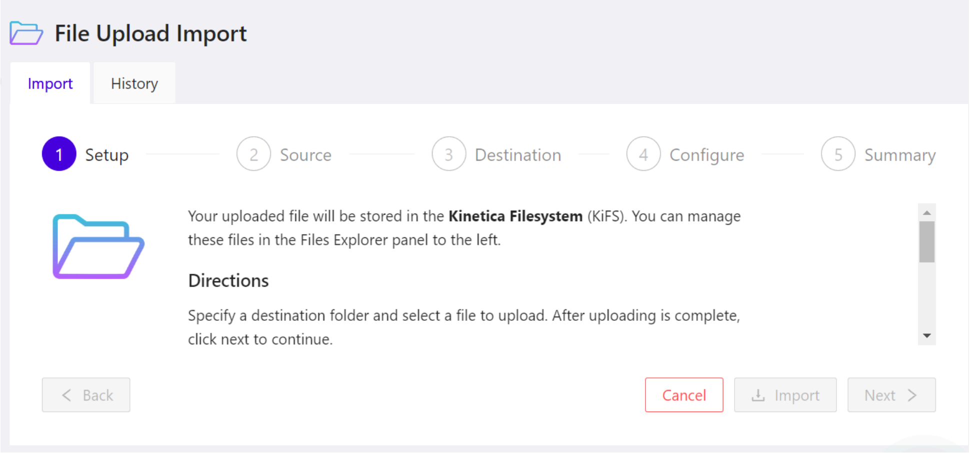


Kinetica:

1. Register for an account at https://www.kinetica.com/
2. Open the dashboard and create a new database (free option)



3. Click login and create a workbook in your database
4. Import csv files from the csv folder

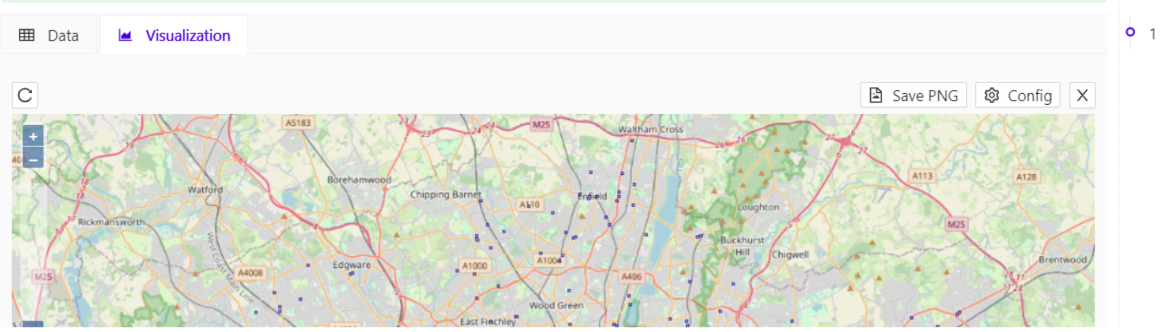


Your uploaded file will be stored in the **Kinetica Filesystem** (KiFS). You can manage these files in the Files Explorer panel to the left.

Directions

Specify a destination folder and select a file to upload. After uploading is complete, click next to continue.

5. Add new block to write SQL queries
6. Visualize the data by pressing Config on the top right corner in the image below and setting the latitude and longitude columns


Data
Visualization
Save PNG
Config
1
A5183
M25
Waltham Cross
Borehamwood
Watford
Chipping Barnet
A110
Enfield
Loughton
A113
A128
Rickmansworth
M25
Buckhurst Hill
Chigwell
Brentwood
Edgware
A1000
A1004
A406
A4008
M25
Wood Green
East Finchley