CAPSTONE PROJECT - THE BATTLE OF NEIGHBORHOODS

1. INTRODUCTION

1.1 BACKGROUND: FOURSQUARE

Foursquare Labs Inc., commonly known as Foursquare, is an American technology company. The company's location platform is the foundation of several business and consumer products, including the Foursquare City Guide and Foursquare Swarm apps. Foursquare built a massive dataset of location through crowd-sourcing their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Communicating with the Foursquare database is really very easy, all thanks to their RESTful API. You simply create a uniform resource identifier, or URI, and you append it with extra parameters depending on the data that you are seeking from the database. Any call request you make is composed of, we can call this base URI, which is api.foursquare.com/v2 and you can request data about venues, users, or tips. But, every time you make a call request, you have to pass your developer account credentials, which are your Client ID and Client Secret as well as what is called the version of the API, which is simply a date.

1.2 BACKGROUND: NOTTINGHAM, UK

Nottingham is a city and unitary authority area in Nottinghamshire, England. Part of the East Midlands region, it is 128 miles (206 km) north of London and 45 miles (72 km) northeast of Birmingham.

Nottingham has links to the legend of Robin Hood and to the lace-making, bicycle (notably Raleigh bikes) and tobacco industries. Nottingham is a popular tourist destination. In 2020,

Nottingham had an estimated population of 330,000. The wider conurbation, which includes many of the city's suburbs, has a population of 768,638.

1.3 BUSINESS PROBLEM

Nottingham city center is very lively with a wide offer of food and entrainment venues, while other areas of the city are also becoming more popular. However, historic data and observations in the city center suggest that there is a very high turn-over among newly opened venues and many of them shut relatively soon after opening. At the same time, there are several examples of venues in different categories that have become very successful and established.

This poses a critical question for potential new owners and investors: **Is there any link between the area and the type of venue that will be successful?**

The target audiences of this case study are existing and prospective new owners of restaurants, bars, arcades, food vans, snack shops and many other types of venues in the food and leisure industry. This case study can inform them which areas of Nottingham might be the best fit for opening their business ideas as well as where there is already a high concentration of similar business, such that there may be a stronger competition.

As a data science project, this case study may also reveal new insights from data, which have not been known and noticed before.

2. DATA

2.1 DATA ACQUISITION AND PROCESSING PLAN

A combination of different datasets will be used to explore the suggested business problem. In order to address the questions outlined in the previous section, a collection of data is required about:

- venue category or type
- its location
- how long it operated
- how popular it was or is

It is proposed to use Foursquare location, reviews and check-ins data to derive the features above. The location data associated with venues can be obtained directly from the Foursquare dataset, which then can be used for spatial clustering. It is more challenging to evaluate how long a venue operated, as it must be recognized that the venue may be **currently open and active** or **already out of business and shut down**.

To deal with this problem, the following method is proposed:

1. Obtain the most recent review/check-in date for the venue - this is the **END DATE**
2. If this date is within the last 3 months, the venue is labelled as **ACTIVE**. If this date is older than the last 3 months, the venue is labelled as **INACTIVE**
3. Obtain the date when the venue was added to Foursquare - this is the **START DATE**
4. Calculate the difference between the start date and the end date - this is the **OPERATION LENGTH**

The final part is to evaluate the how popular the venues was or is. For this, the following values should be extracted:

- Mean review score - this is the **RATING**
- Total number of tips - this is the **REVIEWS** number. Note that ideally the number of check-ins would be used, but this is not available in the free version of Foursquare account

The other features are:

- **NAME**
- venue **TYPE**
- **LONGITUDE**
- **LATITUDE**

Therefore, the objective is to populate a dataframe, which will look like this:

	Name	Type	Latitude	Longitude	Status	StartDate	EndDate	Rating	Reviews
0	Example A	Restaurant	52.125	-1.487	ACTIVE	2017-05-04	2021-03-15	4.6	241
1	Example B	Bar	52.130	-1.975	INACTIVE	2018-10-01	2019-11-30	2.2	148

The dataframe above shows two examples of the data that needs to be extracted from the Foursquare dataset for Nottingham.

2.2 EXTRACTING DATA

2.2.1 VENUES LIST

As the Foursquare location data will be used in this project, the first step is to obtain coordinates of Nottingham using **geopy** module:

Coordinates of Nottingham are: 52.9534193 latitude and -1.1496461 longitude

Foursquare credentials were passed to the Foursquare API and venues nearby the Nottingham coordinates were extracted using 3 km radius. Initially, I only set the limit to 3 items to check the json structure. Once the required information was extracted correctly, a full search for venues was run. There was a total of 247 venues to work with in this project. I populated a dataframe with the information about venues next. I extracted and stored the venues list into a dataframe.

	ID	Name	Type	Latitude	Longitude	Status	StartDate	EndDate	Rating	Reviews
0	5448b19b498ec638a8a752da	200 Degrees Coffee	Coffee Shop	52.953184	-1.148888	None	None	None	None	None
1	4b588c9af964a520685d28e3	Page 45	Bookstore	52.954104	-1.151368	None	None	None	None	None
2	4bc9e8bcfb84c9b668dd1b3e	Aubrey's Traditional Creperie	Creperie	52.954262	-1.153241	None	None	None	None	None
3	5377279f498ea7d99f33b627	Delilah	Deli / Bodega	52.953189	-1.146546	None	None	None	None	None
4	54514d19498e1bb54bc03834	Malt Cross	Bar	52.953068	-1.152378	None	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...
95	4b979daaf964a520cb0b35e3	Royal Concert Hall	Concert Hall	52.955787	-1.151130	None	None	None	None	None
96	57dea566498e8c5236f06855	Heavenly Desserts	Dessert Shop	52.953836	-1.152724	None	None	None	None	None
97	4ba213aff964a52023da37e3	CookieShake	Café	52.955203	-1.148488	None	None	None	None	None
98	4ba640f7f964a520683f39e3	Albert Hall	Concert Hall	52.954316	-1.156111	None	None	None	None	None
99	59b7d054791871507c0b3d37	Fox & Grapes	Pub	52.953572	-1.137641	None	None	None	None	None

100 rows × 10 columns

It can be noted that there is only 100 items on the list. This is due to the limit on the number of venues returned by the call Foursqare API. In order to look for the remaining available venues, I did a grid search around Nottingham and added any items that were not already in the dataframe.

	ID	Name	Type	Latitude	Longitude	Status	StartDate	EndDate	Rating	Reviews
0	5448b19b498ec638a8a752da	200 Degrees Coffee	Coffee Shop	52.953184	-1.148888	None	None	None	None	None
1	4b588c9af964a520685d28e3	Page 45	Bookstore	52.954104	-1.151368	None	None	None	None	None
2	4bc9e8bcfb84c9b668dd1b3e	Aubrey's Traditional Creperie	Creperie	52.954262	-1.153241	None	None	None	None	None
3	5377279f498ea7d99f33b627	Delilah	Deli / Bodega	52.953189	-1.146546	None	None	None	None	None
4	54514d19498e1bb54bc03834	Malt Cross	Bar	52.953068	-1.152378	None	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...
242	53f8b835498e277704e0537f	The Abdication	Pub	52.998669	-1.137892	None	None	None	None	None
243	516bf133e4b0c17e610ac3f7	Arno Vale Recreation Ground	Park	52.994394	-1.124791	None	None	None	None	None
244	57f6e6fb498eacf298264432	A.C.E.S. Ltd	Other Repair Shop	52.991185	-1.131628	None	None	None	None	None
245	4e61fbd862e13e3bce6efa50	Arno Vale Play Park	Playground	52.993740	-1.123585	None	None	None	None	None
246	5b70445d4a1cc0002c360d01	Febuary Gardens	Bar	52.994373	-1.123004	None	None	None	None	None

247 rows × 10 columns

2.2.2 RATINGS, REVIEWS AND DATES

In order to retrieve further data, I made a call to Foursquare API per every row in the venues dataframe and extracted the venue details. Unfortunately, the free version of the API gives me only access to the 2 oldest tips, so I could not populate the end date.

Because it's not possible to obtain the most recent review using free API, I simulated it with random numbers, as this is only a case study for a course capstone project. I found a random number, less than 145, corresponding to the number of days to go back from the reference date (28/03/2021).

The next step was to assign the 'Status' label, where if the end date is within the last 3 months, then the status is ACTIVE and otherwise it should be INACTIVE.

	ID	Name	Type	Latitude	Longitude	Status	StartDate	EndDate	Rating	Reviews
0	5448b19b498ec638a8a752da	200 Degrees Coffee	Coffee Shop	52.953184	-1.148888	Inactive	2014-10-23 07:43:23	2020-12-20 00:00:00	9.2	59
1	4b588c9af964a520685d28e3	Page 45	Bookstore	52.954104	-1.151368	Inactive	2010-01-21 17:19:22	2020-12-11 00:00:00	9.0	7
2	4bc9e8bcfb84c9b668dd1b3e	Aubrey's Traditional Creperie	Creperie	52.954262	-1.153241	Active	2010-04-17 16:58:36	2021-03-18 00:00:00	9.1	16
3	5377279f498ea7d99f33b627	Delilah	Deli / Bodega	52.953189	-1.146546	Active	2014-05-17 09:10:55	2021-01-21 00:00:00	8.8	9
4	54514d19498e1bb54bc03834	Malt Cross	Bar	52.953068	-1.152378	Inactive	2014-10-29 20:24:57	2020-11-18 00:00:00	8.7	11
...	...	...	...	...	...	...	...	...	...	...
242	53f8b835498e277704e0537f	The Abdication	Pub	52.998669	-1.137892	Inactive	2014-08-23 15:50:13	2020-12-16 00:00:00	NaN	1
243	516bf133e4b0c17e610ac3f7	Arno Vale Recreation Ground	Park	52.994394	-1.124791	Inactive	2013-04-15 12:23:15	2020-12-12 00:00:00	NaN	1
244	57f6e6fb498eacf298264432	A.C.E.S. Ltd	Other Repair Shop	52.991185	-1.131628	Active	2016-10-07 00:06:19	2021-01-07 00:00:00	NaN	0
245	4e61fbd862e13e3bce6efa50	Arno Vale Play Park	Playground	52.993740	-1.123585	Active	2011-09-03 10:05:12	2021-03-28 00:00:00	NaN	1
246	5b70445d4a1cc0002c360d01	Febuary Gardens	Bar	52.994373	-1.123004	Active	2018-08-12 14:29:49	2021-02-02 00:00:00	NaN	0

247 rows × 10 columns

2.3 DATA CLEANING AND FEATURE GENERATION

The data has now been fully extracted, so it can be further cleaned. Start with dropping the ID column, as this is no longer needed. Find how long the venue operated using the difference between the end date and the start date. Store as a new column. Now let's drop the start date and end date. Also, convert reviews to float for the data modelling steps. Finally, rearrange the columns of the dataframe and make sure the relevant columns are numeric.

```
Name             object
Type             object
Latitude        float64
Longitude       float64
Status           object
Rating          float64
Reviews           int64
DaysOperated      int64
dtype: object
```

	Name	Type	Latitude	Longitude	DaysOperated	Reviews	Rating	Status
0	200 Degrees Coffee	Coffee Shop	52.953184	-1.148888	2249	59	9.2	Inactive
1	Page 45	Bookstore	52.954104	-1.151368	3976	7	9.0	Inactive
2	Aubrey's Traditional Creperie	Creperie	52.954262	-1.153241	3987	16	9.1	Active
3	Delilah	Deli / Bodega	52.953189	-1.146546	2440	9	8.8	Active
4	Malt Cross	Bar	52.953068	-1.152378	2211	11	8.7	Inactive
...	...	...	...	...	...	...	...	...
242	The Abdication	Pub	52.998669	-1.137892	2306	1	NaN	Inactive
243	Arno Vale Recreation Ground	Park	52.994394	-1.124791	2797	1	NaN	Inactive
244	A.C.E.S. Ltd	Other Repair Shop	52.991185	-1.131628	1552	0	NaN	Active
245	Arno Vale Play Park	Playground	52.993740	-1.123585	3493	1	NaN	Active
246	Febuary Gardens	Bar	52.994373	-1.123004	904	0	NaN	Active

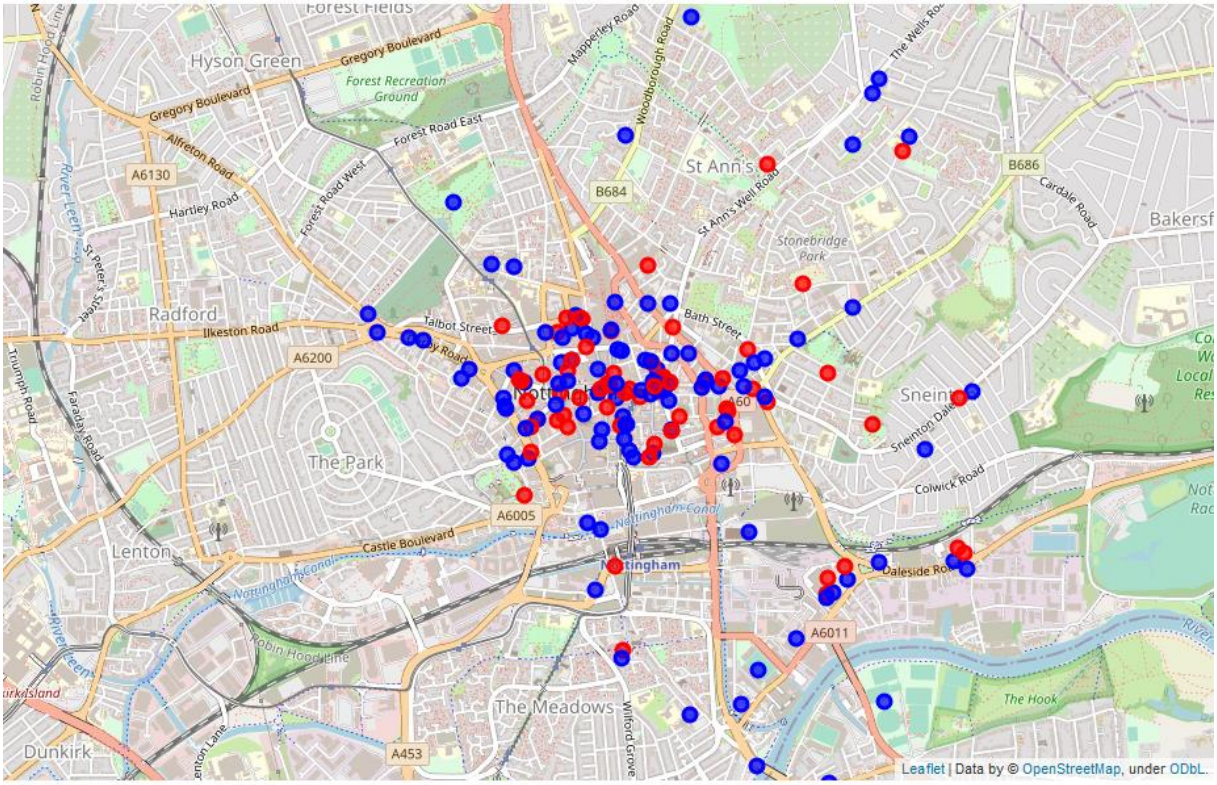
247 rows × 8 columns

3. EXPLORATORY DATA ANALYSIS

3.1 ALL VENUES EXPLORATION

	Name	Type	Latitude	Longitude	DaysOperated	Reviews	Rating	Status
count	247	247	247.000000	247.000000	247.000000	247.000000	197.000000	247
unique	232	100	NaN	NaN	NaN	NaN	NaN	2
top	Co-op Food	Pub	NaN	NaN	NaN	NaN	NaN	Active
freq	7	26	NaN	NaN	NaN	NaN	NaN	152
mean	NaN	NaN	52.954343	-1.140614	3150.757085	7.919028	7.598477	NaN
std	NaN	NaN	0.016055	0.009143	981.862746	11.904495	0.906354	NaN
min	NaN	NaN	52.922995	-1.162756	124.000000	0.000000	5.100000	NaN
25%	NaN	NaN	52.947604	-1.147016	2396.000000	1.000000	7.100000	NaN
50%	NaN	NaN	52.953556	-1.142904	3627.000000	3.000000	7.800000	NaN
75%	NaN	NaN	52.955869	-1.132785	3914.000000	10.000000	8.300000	NaN
max	NaN	NaN	52.998669	-1.122963	4084.000000	89.000000	9.200000	NaN

Let's show all venues on a map and include the details in the pop-ups. However, let's also color the markers according to the status of the venue.



ALL VENUES SHOWN ON THE MAP - ACTIVE IN BLUE, INACTIVE IN RED

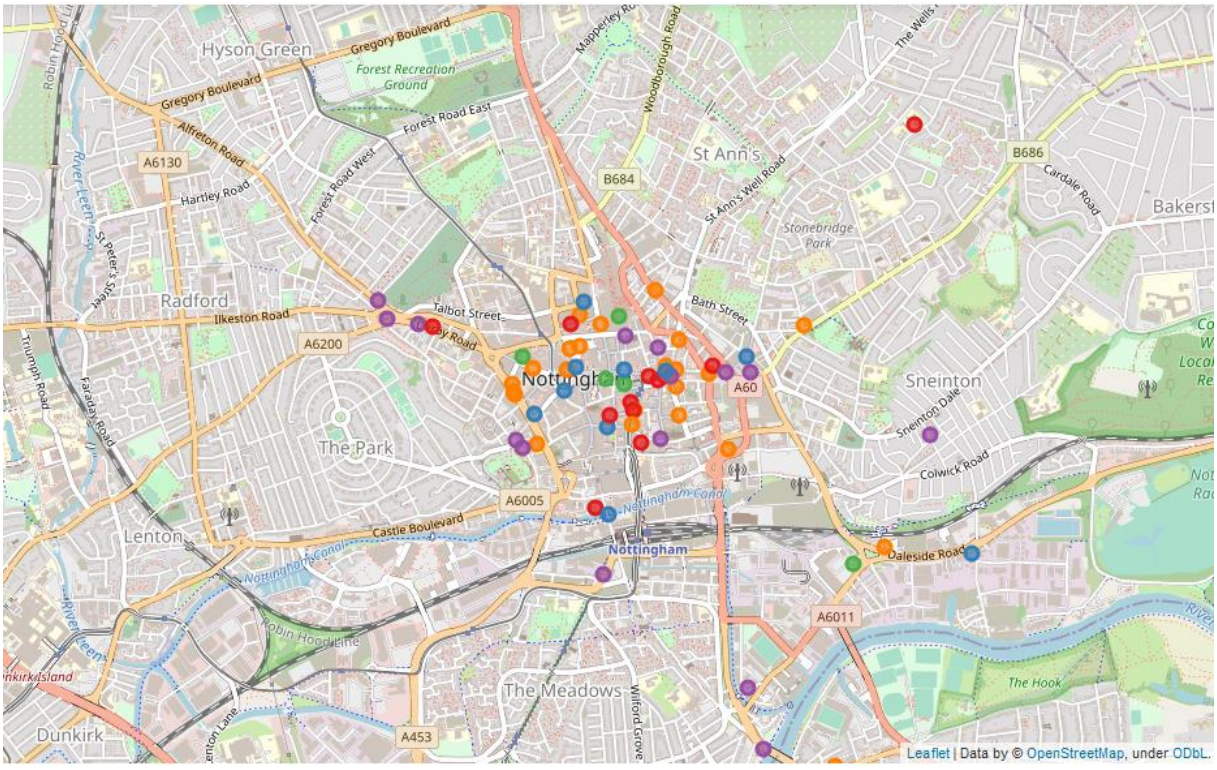
3.2 FOOD VENUES ANALYSIS

In the next step, let's group the venues into more generic categories: Bar, Pub, Cafe, Restaurant, Deli. I'm going to also focus just on these types of venues in the later part of the study and skip all the other ones.

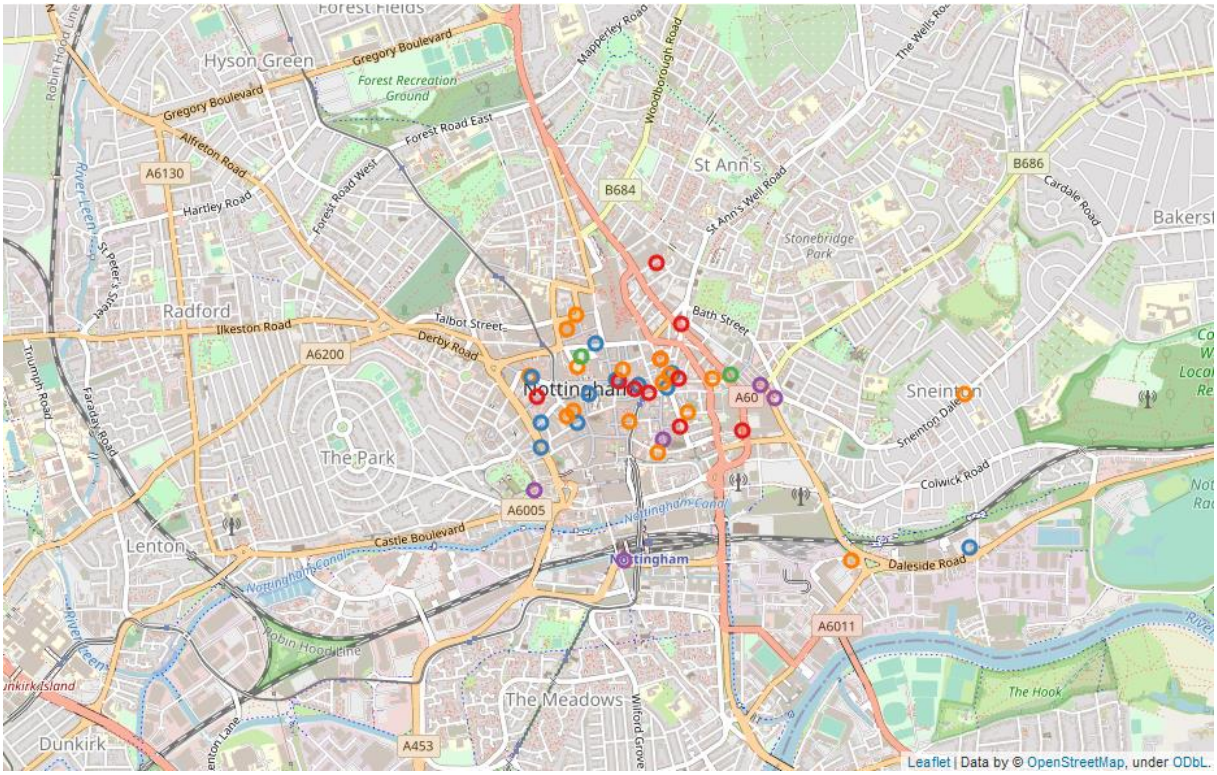
Let's create a separate dataframe for the venues in the chosen high-level categories. It was also seen that 12 out of 144 food venues miss any ratings. For the sake of the exercise and modelling, I inserted the average rating for these venues.

	Name	Type	Category	Latitude	Longitude	DaysOperated	Reviews	Rating	Status
0	200 Degrees Coffee	Coffee Shop	Cafe	52.953184	-1.148888	2249	59	9.2	Inactive
2	Aubrey's Traditional Creperie	Creperie	Deli	52.954262	-1.153241	3987	16	9.1	Active
3	Delilah	Deli / Bodega	Deli	52.953189	-1.146546	2440	9	8.8	Active
4	Malt Cross	Bar	Bar	52.953068	-1.152378	2211	11	8.7	Inactive
6	Fox Cafe	Café	Cafe	52.953770	-1.147148	1982	8	8.7	Inactive
7	Five Guys	Burger Joint	Restaurant	52.953777	-1.150292	2126	11	8.6	Active
8	Wired Cafe Bar	Coffee Shop	Cafe	52.953556	-1.145667	2597	35	8.7	Inactive
10	Kigali	Coffee Shop	Cafe	52.953461	-1.143758	1563	14	8.9	Inactive
11	World Service	Restaurant	Restaurant	52.950793	-1.152312	3843	15	8.8	Active
12	Solo Grano	Italian Restaurant	Restaurant	52.952522	-1.149927	522	3	8.5	Inactive
14	Junkyard Bottle Shop and Pour House	Bar	Bar	52.952447	-1.146109	2309	22	8.6	Active

Let's explore the location of different categories of the venues. Firstly the Active ones, then Inactive ones.



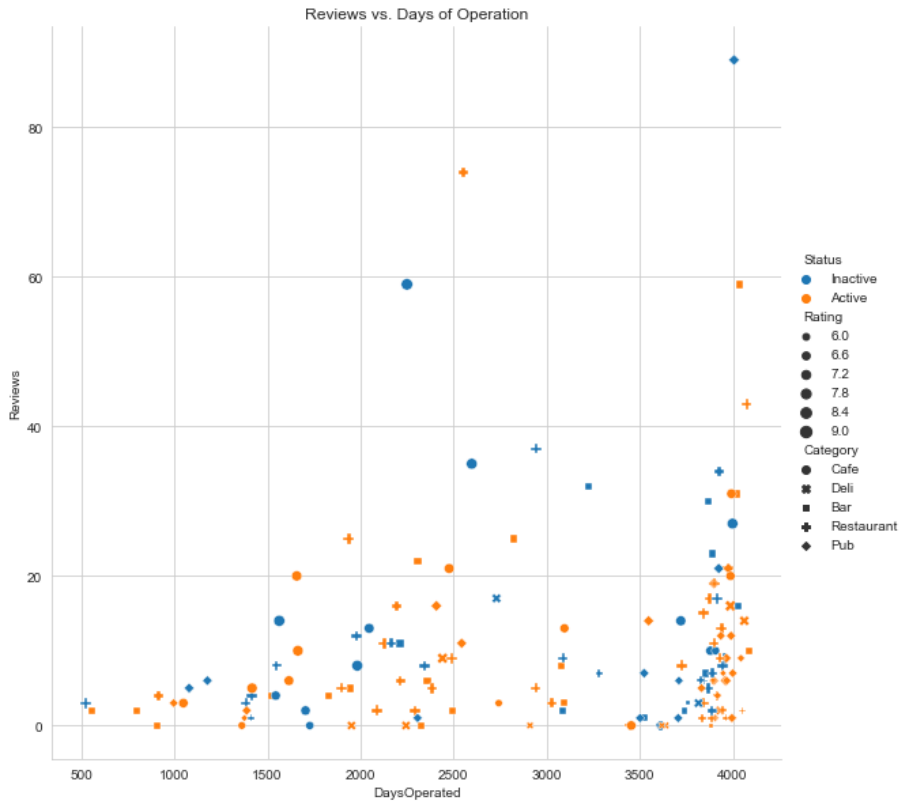
ACTIVE FOOD VENUES: RED - BAR , ORANGE - RESTAURANT, BLUE - CAFE , GREEN - DELI , PURPLE – PUB



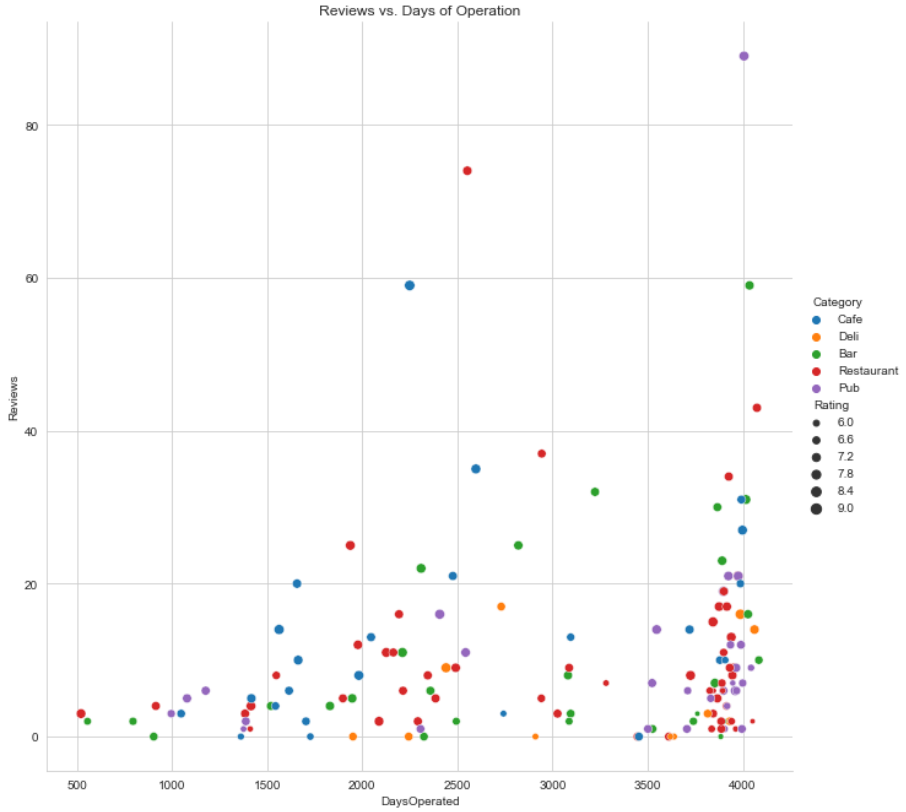
INACTIVE FOOD VENUES: RED - BAR , ORANGE - RESTAURANT, BLUE - CAFE , GREEN - DELI , PURPLE – PUB

3.3 RELATIONSHIPS BETWEEN PARAMETERS

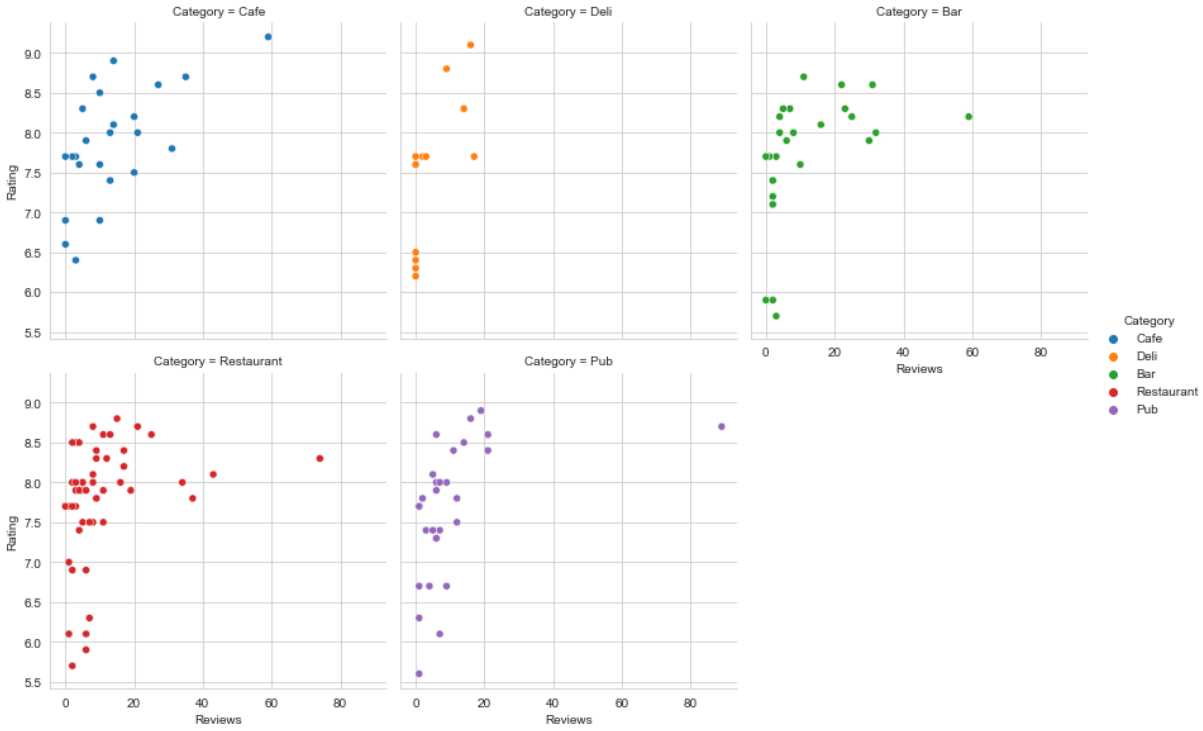
The following relationship plot shows all the features of the dataset in one go. The main purpose of the plot is to show the number of reviews vs. days operated, but additionally the status, rating and venue category.

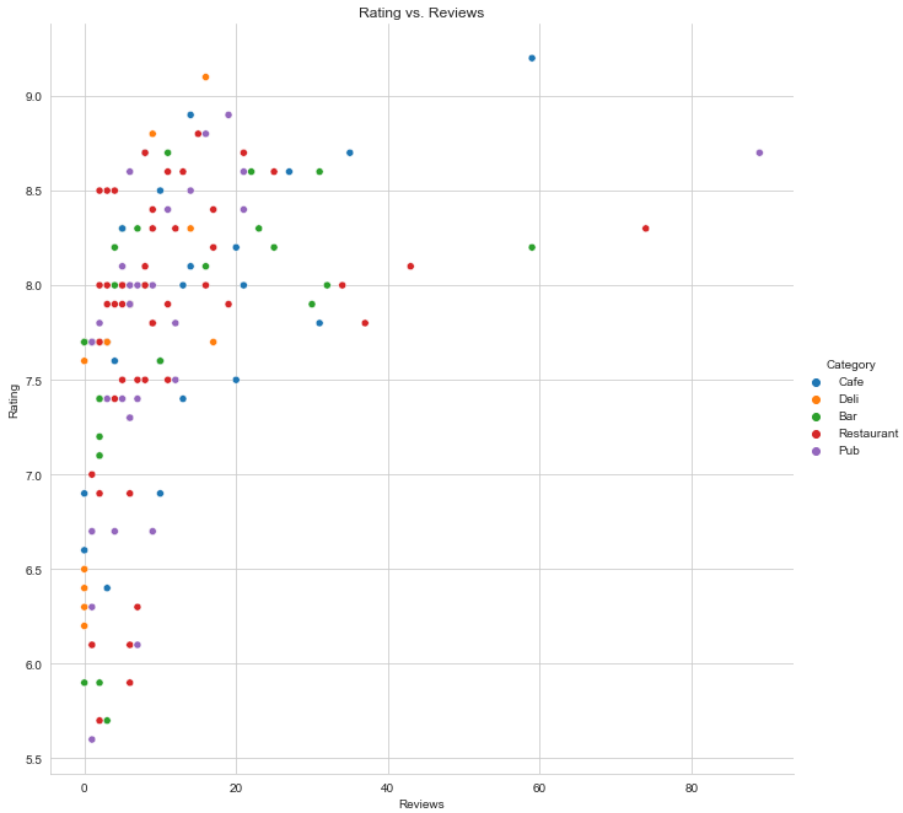


As expected, some of the venues which operated for longer time have more reviews, but at the same time, there is a lot of venues that don't have many reviews. There seem to be no particular useful patterns to review. Let's ignore the status distinction as it was generated from artificial data.

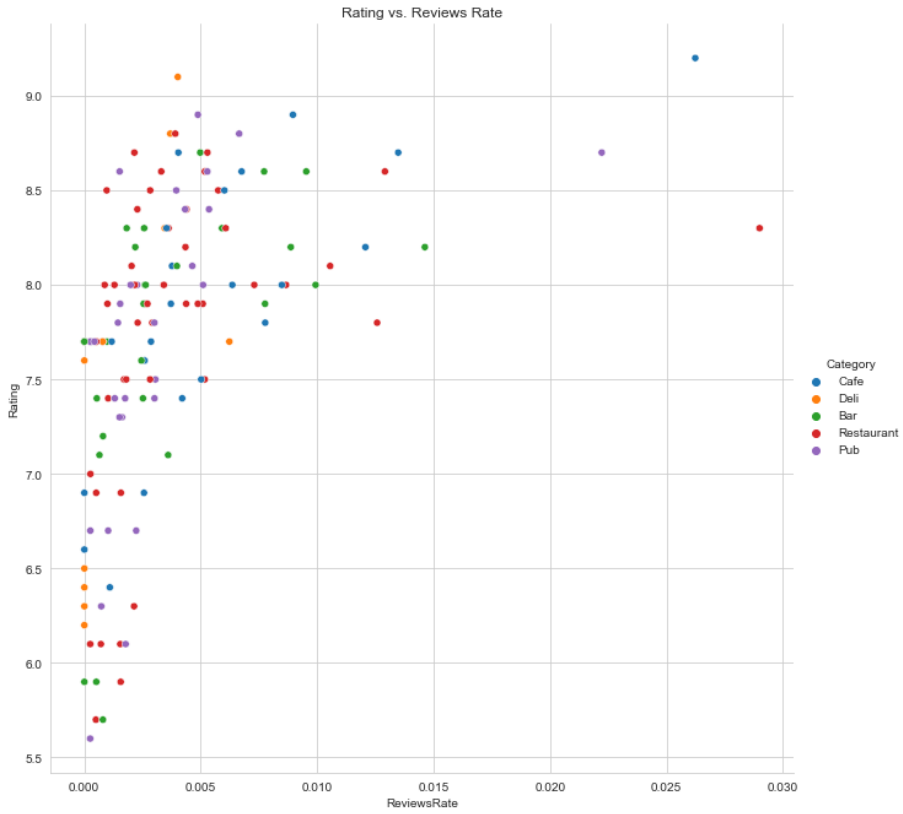


We can also review whether there is a link between the rating and the number of reviews. It could be the case that if the rating is very good, there are also more reviews.



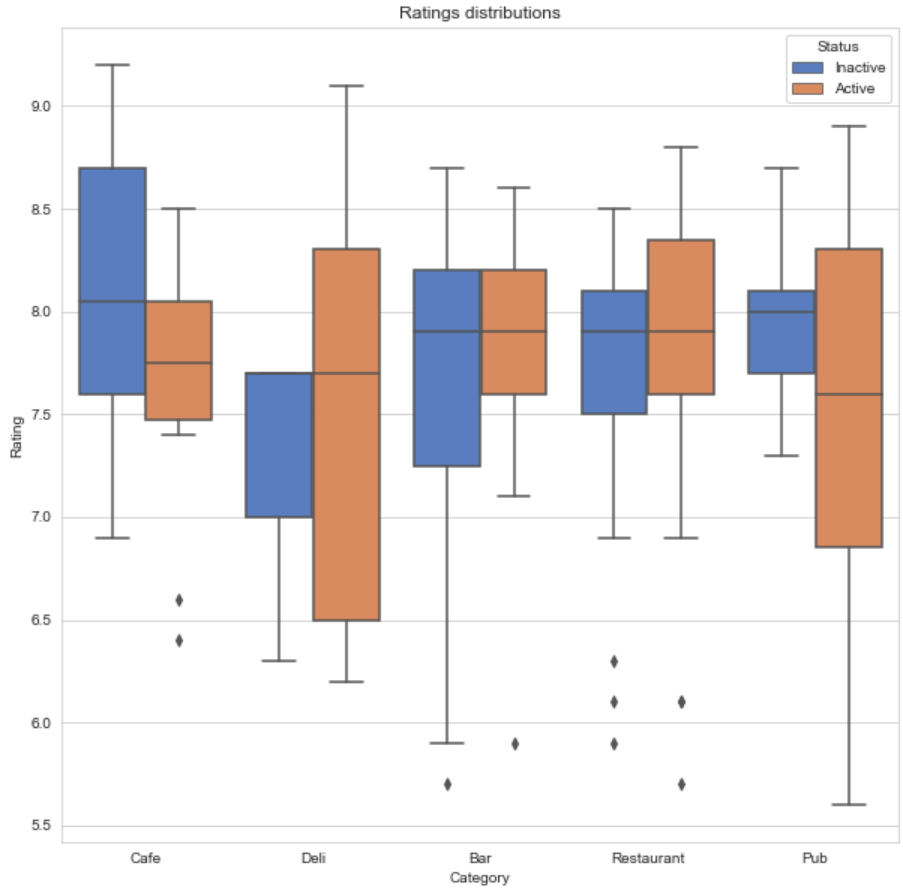


Indeed, from the above plots we can observe that the venues that have a lot of reviews also tend to have high ratings. This could suggest that if customers are particularly satisfied they are also more likely to write a review. Another aspect to check is the rate of reviews.



This confirms that in general the venues that receive very good ratings are also reviewed more frequently.

Let's now inspect the distributions of ratings for the Active and Inactive groups.



In the boxplot above, there is no clear trend between the Active and Inactive venues. Again, this is probably due to the data being populated artificially. It could be expected that if we can access the real data on which venues are closed, the Inactive venues would have lower ratings that the ones that are still in business.

4. PREDICTIVE MODELLING

For this case study, I am going to create a classifier to predict if a venue will be successful based on the location, category, rating and the rate of reviews. This would be a real, appropriate question to explore for the business problem stated at the start of the notebook. However, it should be noted that with some artificial input data, the modelling might not be capable or representative of the true patterns.

4.1 PRE-PROCESS INPUTS

'Category', 'Latitude', 'Longitude', 'Rating', 'ReviewsRate' were extracted as input columns and then category was one-hot encoded:

	Latitude	Longitude	Rating	ReviewsRate	Category_Bar	Category_Cafe	Category_Deli	Category_Pub	Category_Restaurant
0	52.953184	-1.148888	9.2	0.026234	0	1	0	0	0
1	52.954262	-1.153241	9.1	0.004013	0	0	1	0	0
2	52.953189	-1.146546	8.8	0.003689	0	0	1	0	0
3	52.953068	-1.152378	8.7	0.004975	1	0	0	0	0
4	52.953770	-1.147148	8.7	0.004036	0	1	0	0	0
5	52.953777	-1.150292	8.6	0.005174	0	0	0	0	1
6	52.953556	-1.145667	8.7	0.013477	0	1	0	0	0
7	52.953461	-1.143758	8.9	0.008957	0	1	0	0	0
8	52.950793	-1.152312	8.8	0.003903	0	0	0	0	1
9	52.952522	-1.149927	8.5	0.005747	0	0	0	0	1
10	52.952447	-1.146109	8.6	0.009528	1	0	0	0	0

The 'Status' column was used as the target label and converted to numeric:

	Status
0	1
1	0
2	0
3	1
4	1
5	0
6	1
7	1
8	0
9	1

4.2 FIT MODEL

As the data is relatively limited, the cross-validation can be employed to assess the modelling accuracy. As the parameters correspond to very varied factors, decision trees might prove as a good modelling technique:

```
# Set up and evaluate the 1st tree
tree1 = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
scores = cross_val_score(tree1, X, y, scoring='accuracy', cv=cv)
print('Model accuracy: %.3f (+/-%.3f)' % (np.mean(scores), 3*np.std(scores)/scores.size))
```

Model accuracy: 0.540 (+/-0.007)

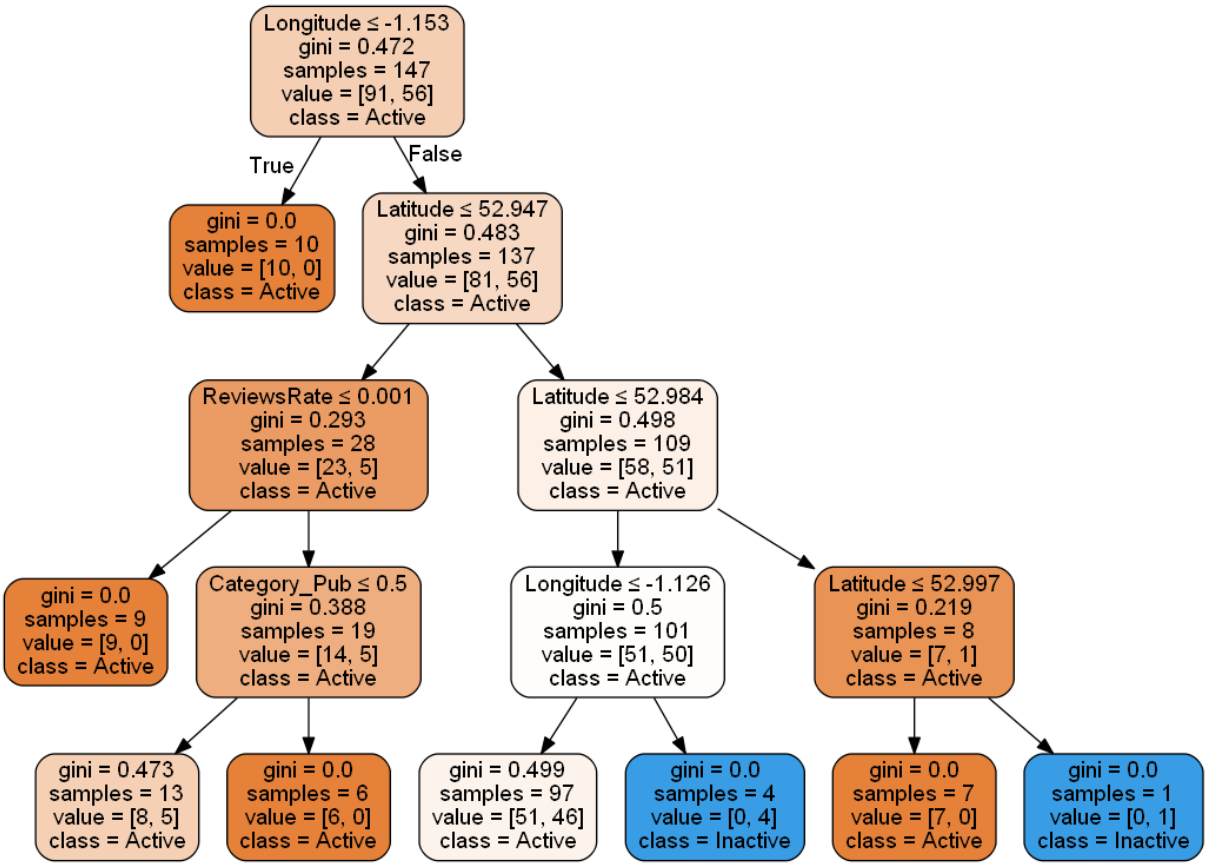
As seen above, the model accuracy over 6 evaluations of 8-fold validation is relatively poor. Different settings of the decision tree can be tried:

```
tree2 = DecisionTreeClassifier(criterion="gini", max_depth = 4)
scores = cross_val_score(tree2, X, y, scoring='accuracy', cv=cv)
print('Model accuracy: %.3f (+/-%.3f)' % (np.mean(scores), 3*np.std(scores)/scores.size))
```

Model accuracy: 0.523 (+/-0.006)

It looks like a very comparable accuracy was obtained using an algorithm with different settings. The modelling predictive capability is poor, not a great deal better than chance. This is likely due to the synthesized target values.

Finally, let's visualize the tree classifier



5. SUMMARY

Nottingham food venues and other types of venues have been explored using Foursquare API and various data science techniques. Based on the mixture of the real and some synthesized data, there was no noticeable relation between the ratings of the venues and whether they are still in business or closed. The only clear trend was that where venues have high ratings, the reviews are also uploaded more often. A decision tree classifier was created and evaluated to predict if a venue will close or stay open based on its location, category, ratings and review rates. In the current case study, the classifier showed very poor predictive accuracy.