Objective

Using data about wards and reviews estimate the demand for restaurants in an area (zip code) to find out which area has a need for restaurants and the type of cuisine people are looking for.

Steps:

1. Clean the data to include zip code and cuisine type
2. Find outside data and make the data correlate to zip code
3. Calculate metrics

Edinburgh

We decided to use Edinburgh as we found a lot of other demographic data on Edinburgh that we can use with the yelp data. Further some of us has visited the company and therefore could provide local insights.

Cleaning the Data

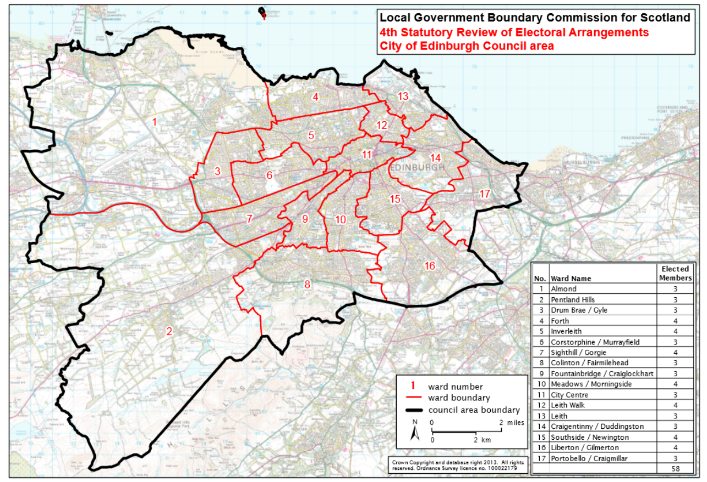
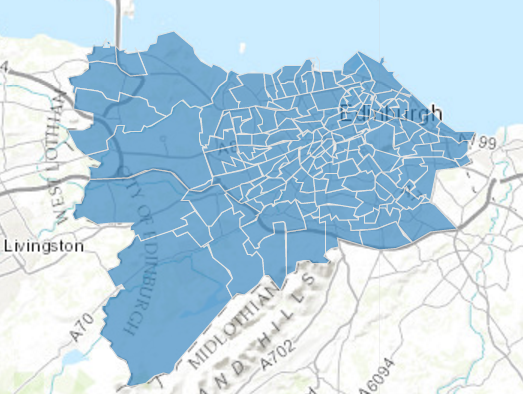
The data from yelp comes in a JSON file with nested objects. By researching online, we are able to find a function to import the JSON files into pandas for python. Using pandas, we filtered out the Edinburgh data and noticed that the businesses contains not only restaurants but all businesses.

We therefore made the assumption that the business is a restaurant if it states restaurant in the category. Based on the filtered out data, using joins, we were able to extract the relevant tables for Checkins, Reviews, Tips and Users. In general there are 1215 restaurants registered on Yelp.

We also noticed that in the categories, it mentions a lot of different attributes to explain what category it is. An example of a category would be “Gastropubs,Bars,Scottish,Nightlife,Restaurants”. For our analysis, we would like to focus on cuisine, so we made a list of cuisine keywords such as Italian or Scottish. Then we search the category for such keywords and if there is such keyword, the cuisine column will state the keyword.

As we would also like to combine with outside data, we need a way to state what area the restaurant should be in. Unfortunately the neighborhood column from the business dataset does not correspond to the data that we got from the Edinburgh government website. Therefore we decided to use the zip code of the restaurant’s address instead. We created a function that extracts the zip code from the data and we are only focusing on the general big area, therefore the first three or four letters of the zip code. An example would be EH6 or EH12.

The outside data that we would like to use is the Locality and ward data profiles from <http://www.edinburgh.gov.uk/info/20247/edinburgh_by_numbers/1393/locality_and_ward_data_profiles>. It contains data from Population - gender and age, Housing, Employment, Education and professions, Income, Benefits, Health & disability, Lifestyle, Satisfaction with Services, Scottish Index of Multiple Deprivation data. In general we would like to compare the population, income and property prices with the amount of restaurants there are. However the data is organized by ward. Therefore we have to give each ward a zip code. Through another set of data that we collected the pooling districts from the city of Edinburgh Mapping portal (<http://data.edinburghcouncilmaps.info/datasets/2cee9b18a21344b0879c3c51d71fd2c6_28>), we noticed that there is a zip code for each polling place. Therefore for each ward, we used the zip code of the polling districts. As a ward might have numerous polling districts with different zips, we use the highest count of zip for that ward as long as the zipcode is from EH1-16 which are where the restaurants from yelp is from. Therefore if the ward Almond has 4 zips from EH4 and only 1 from EH2 we will considered Almond to be part of EH4 and the population will be from there.



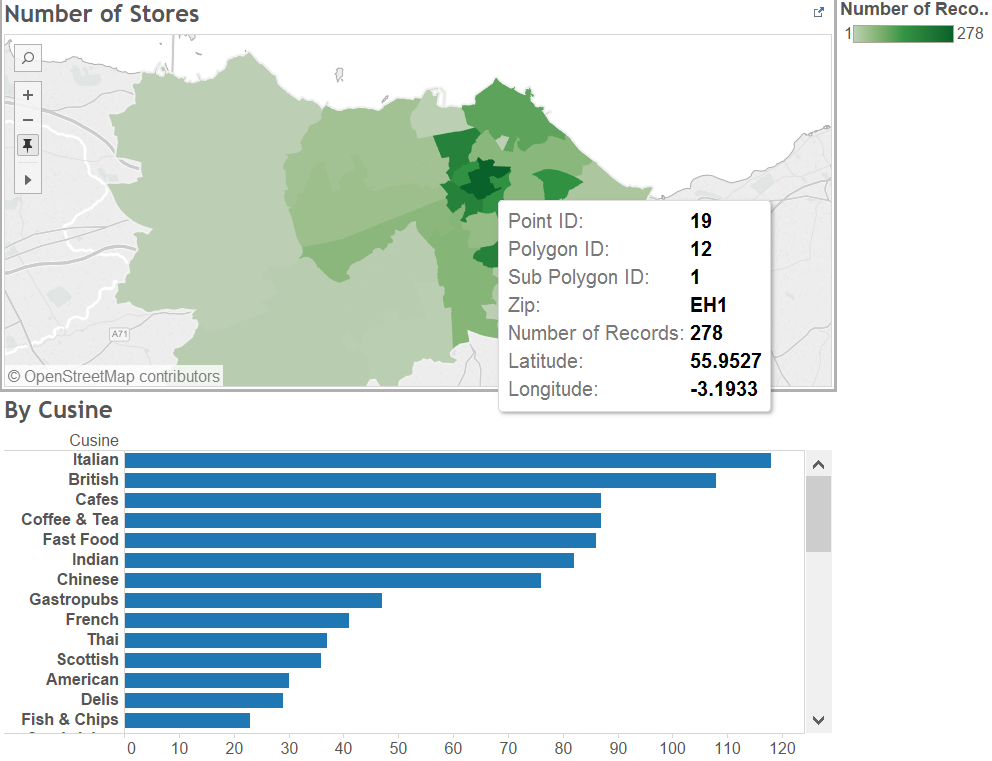
*Polling Districts Wards*

However due to the fact that some wards encompass two zip codes, there will be some zip codes that does not belong to any ward using the method above. Therefore for those zip codes, we have found wards that do belong to that zip and for that ward divided the population numbers by half and set it for the two zip codes. Income and property prices remain the same.

Further from booking.com, we have downloaded all the hotels that they have in Edinburgh. We have done some cleansing to it to extract the zip code so that the data will be useful.

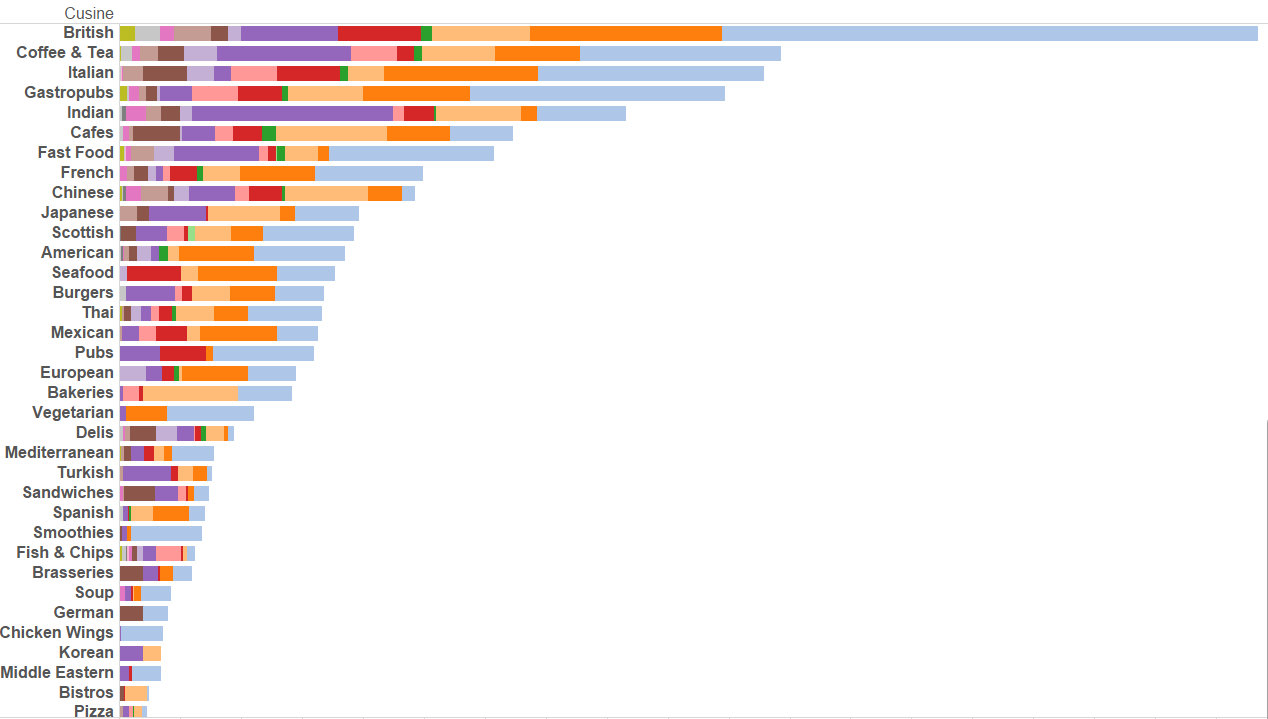
Analysis

We put the information into Tableau and noticed that in general, Italian and British Cusine are the most popular. The most stores are located in City Centre which is EH1 and EH2 as shown in the table below

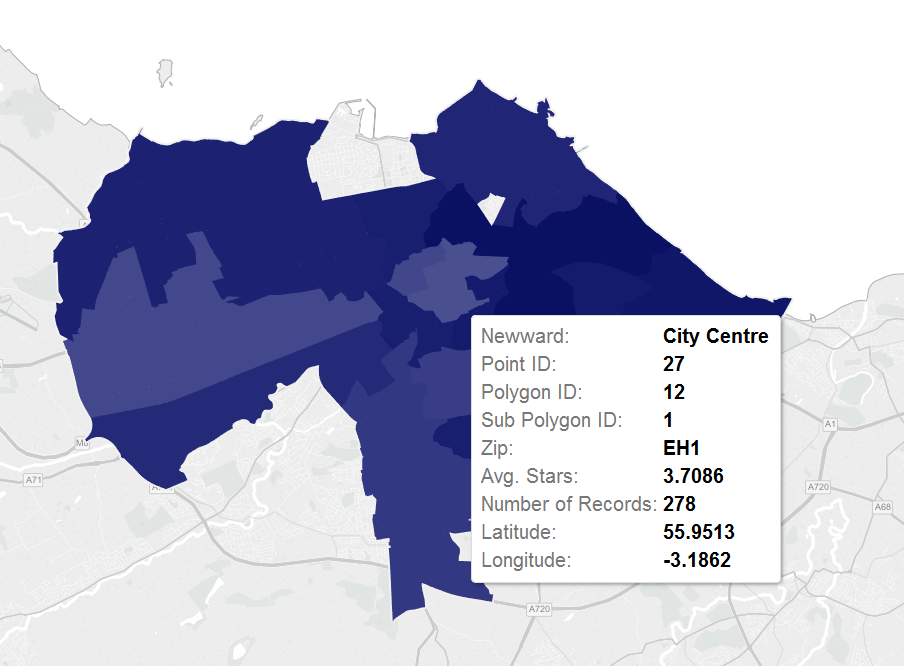


We also looked at the type of cusine that has the most reviews. we decide to use a ratio of the amount of reviews for a restaurant as an indicator of visitors. We decided against using users as an indicator as a user can visit a restaurant twice and give two reviews which will be counted as two visits. This allows to show the demand for each zip.

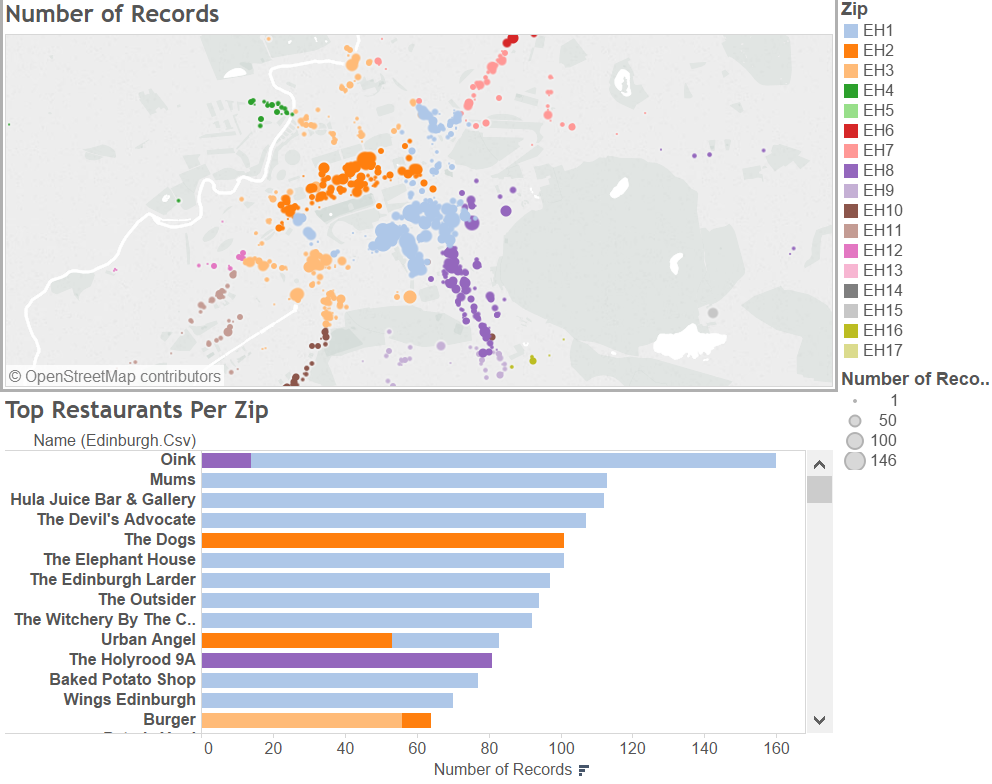
Here it is shown that British food is very popular in Edinburgh. In certain places, such as EH1 which is shown above, shows that thought here are more Italian than British restaurant and so we can captalize on the lack of British food to open a British restaurant there.



However very interestingly, the stores at city centre EH1 and EH2 has also the lowest rating. The chart below shows the average star ratings by color. The darker the color the higher average score it has. It could be because there are many stores in the city centre and hence the low score pulls down the high score. It also means if our restaurant wants to focus on ratings, we should not try to put our store in the city centre.



We also decided that we want to calculate the demand via looking at the number of reviews. We understand the number of reviews is not the best indicator of how many customers has but it generally serves as a good proxy for the number of customers. As seen at the graph below the location with the most reviews are EH1, EH2, EH3 and EH8. So we will based our analysis on that.



From the data we have also generated numerous metrics such as population divide by restaurant, income divided by restaurants, property divided by restaurants and hotel divided by restaurants.

The statistics are follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| zip | pop/rest | income/rest | property/rest | hotels/rest | num of rest |
| EH1 | 45.912548 | 158.756654 | 775.768061 | 0.441065 | 263 |
| EH2 | 85.035211 | 294.035211 | 1436.809859 | 0.338028 | 142 |
| EH3 | 130.402174 | 206.195652 | 961.570652 | 0.478261 | 184 |
| EH4 | 2272.921875 | 4068.375 | 18528.375 | 0.40625 | 32 |
| EH5 | 5535.166667 | 10830.66667 | 45313.66667 | 2.333333 | 3 |
| EH6 | 251.828283 | 342.929293 | 1421.535354 | 0.515152 | 99 |
| EH7 | 1028.803571 | 1210.053571 | 5167.160714 | 1.017857 | 56 |
| EH8 | 235.624113 | 294.120567 | 1495.035461 | 0.347518 | 141 |
| EH9 | 369.144444 | 921.577778 | 4684.444444 | 0.866667 | 45 |
| EH10 | 551.048387 | 759.354839 | 3911.274194 | 0.241935 | 62 |
| EH11 | 690.703704 | 488.685185 | 1928.814815 | 0.277778 | 54 |
| EH12 | 1158.666667 | 2265.230769 | 10028.74359 | 1.102564 | 39 |
| EH13 | 6059.5 | 12448.75 | 58974.5 | 0.25 | 4 |
| EH14 | 3952.5 | 7143.5 | 31195.5 | 0.5 | 6 |
| EH15 | 1335.789474 | 1735.526316 | 7157.684211 | 0.368421 | 19 |
| EH16 | 2782.666667 | 2779 | 10926.83333 | 3.083333 | 12 |

Assuming that we would like to build a restaurant at a central location which as mention above is EH1,EH2,EH3 and EH8. With EH1 and EH2 being considered central city. When considering EH8, it seems that it has the largest population over restaurant metric compared to the other three. However it also has the lowest hotel per restaurant. Therefore there might be less tourist in the area. EH3 though ahs a decent population to restaurant ratio and a hotel to restaurant ratio. Further it’s property value/restaurant ratio is low so it should not be too expensive to buy a restaurant there.

Potential Revenue

In order to calculate potential revenue we have to estimate the amount of customers a store would go to. We looked at the check-in dataset and realized that it is not a good estimate of how many customers will go to a store as the amount of check-ins are based on the amount of offers. As mentioned on the yelp site, <http://www.yelp.com/topic/walnut-creek-what-is-this-yelp-check-in-thing> it states that “Certain businesses offer discounts when yelpers check in to that business” and further it states “ You checkin with the yelp app on an iPhone/iPad or Android device. You have to be within a close proximity to a location to check-in and the app used your phones GPS to measure your location.” Hence most of the customer visits will not have a check-in even if the customer came from yelp as the customer will most likely not open the yelp application and clicking in check-in.

Again we decide to use a ratio of the amount of reviews for a restaurant as an indicator of visitors. We give the number of reviews a meaning such as that it indicates .1% of people that will visit the store. Finally, we found the average number of users per restaurant in that zipcode which will give reviews.

The results are as follows

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| EH1 | EH2 | EH3 | EH4 | EH5 | EH6 | EH7 | EH8 |
| 17.36331 | 14.98649 | 10.47644 | 4.625 | 4.333333 | 9.038462 | 8.844828 | 12.71622 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| EH9 | EH10 | EH11 | EH12 | EH13 | EH14 | EH15 | EH16 |
| 8.108696 | 9.603175 | 6.263158 | 4.589744 | 1.75 | 3 | 5.9 | 5.666667 |

As expected EH1, EH2, EH3 and EH8 has the highest user base. As we decided that we want to focus on EH3 where we expect there to be more traffic due to the amount of hotels and population, we can expect that for a year there will be 10.47655 \* 1000 which is around 10000 customers.

We also looked at the different price range from this site <http://www.yelp.com/topic/san-diego-can-anyone-give-me-the-actual-dollar-range-for-the-dollar-sign-symbols-in-rrgards-to-pricing> and noticed the following $= under $10, $$= $11-$30, $$$= $31-$60, $$$$= above $61. We will use the average price of each for the calculation.

If we open a restaurant that is around three dollars signs. It will be 10000 \* 45 which is $450000 per month.