# Objective

The objective of this assignment was to use data provided by Yelp in order to estimate the demand for restaurants in an area (zip code) of a chosen city, and to find out which type of cuisine people are looking for.

In order to undertake this analysis, we performed the following steps:

1. Data cleaning to extract zip codes and restaurant types;
2. Integrate external data sources with the Yelp data in order to gather more detail on the zip code demographics; and
3. Calculate metrics for each area in order to address the main objective.

# Analysing Restaurants in Edinburgh

We decided to perform this analysis for Edinburgh. This was because we found extensive demographic data that could be easily integrated with the Yelp data. We also has better knowledge of the zip codes as Edinburgh is still in the UK.

## Cleaning the Data

The Yelp data were provided as JSON files. We used the pandas Python package to process this data, and filtered the larger data to just that relevant to Edinburgh. Noting that the data contained many business types, we made the assumption that the business was a restaurant if it category field stated as such. We then used this filtered data to extract the relevant tables for check-ins, reviews, tips and users. This provided us with data on 1215 restaurants in Edinburgh that were registered on Yelp.

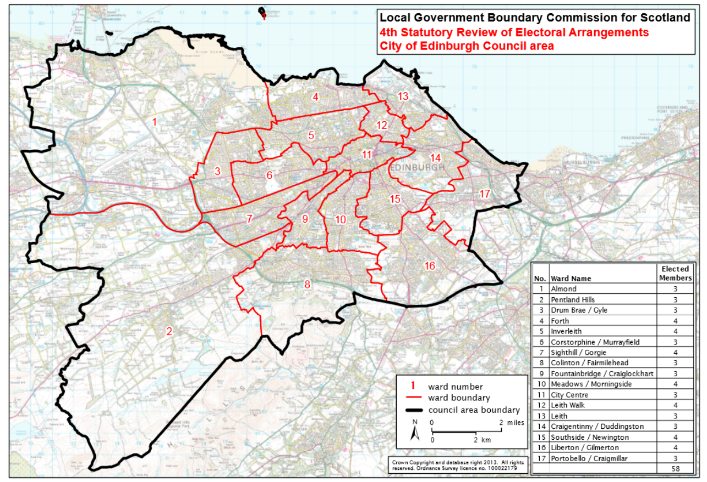
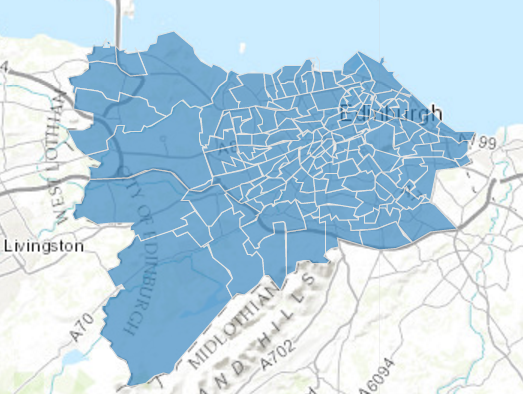
Regarding restaurant cuisine type, we noticed that many different attributes were used to explain the restaurant category was . For example: “Gastropubs,Bars,Scottish,Nightlife,Restaurants”. As we were primarily interested in cuisine, we made a list of cuisine keywords such as “Italian” or “Scottish”. We searched the category for such keywords in order to assign each restaurant to a cuisine type. If there are more than one cuisine detected, we will focus on the cuisine that is the most expressive of the type of food, which is usually the country type of the food.

In order to combine the external demographic data with the Yelp data, we extracted the zip code from each restaurant’s address. In doing so we only focussed on the general area, the first three or four letters of the zip code. For example EH6 or EH12.

The data obtained from the Edinburgh government [website](http://www.edinburgh.gov.uk/info/20247/edinburgh_by_numbers/1393/locality_and_ward_data_profile) provided locality and ward demographic profiles. It contained data on –gender, age, housing, employment, education and professions, income, benefits, health and disability, lifestyle, satisfaction with services, and Scottish Index of Multiple Deprivation data.

We used this data in order to compare the population, income and property prices with the amount of restaurants there are. However, this data was organized by ward. Therefore it was necessary for us to map each ward to a zip code.

We used a second external data set of polling districts from the [City of Edinburgh Mapping portal](http://data.edinburghcouncilmaps.info/datasets/2cee9b18a21344b0879c3c51d71fd2c6_28). Figure 1 below is a demonstration of the different polling districts and figure 2 are the wards districts.



*Figure 1:Polling Districts Figure 2: Wards*

We noticed that each polling location was assigned to a zip code. We therefore used the zip code of the polling districts in order to map back to ward. As some wards had numerous polling districts within them, we used the zip code with the largest number of occurrences in the ward (as long as the zip code was from EH1-16 – our area of interest).. For example if the ward “Almond” has 4 zip codes in EH4 and only 1 in EH2 we considered Almond to be part of EH4.

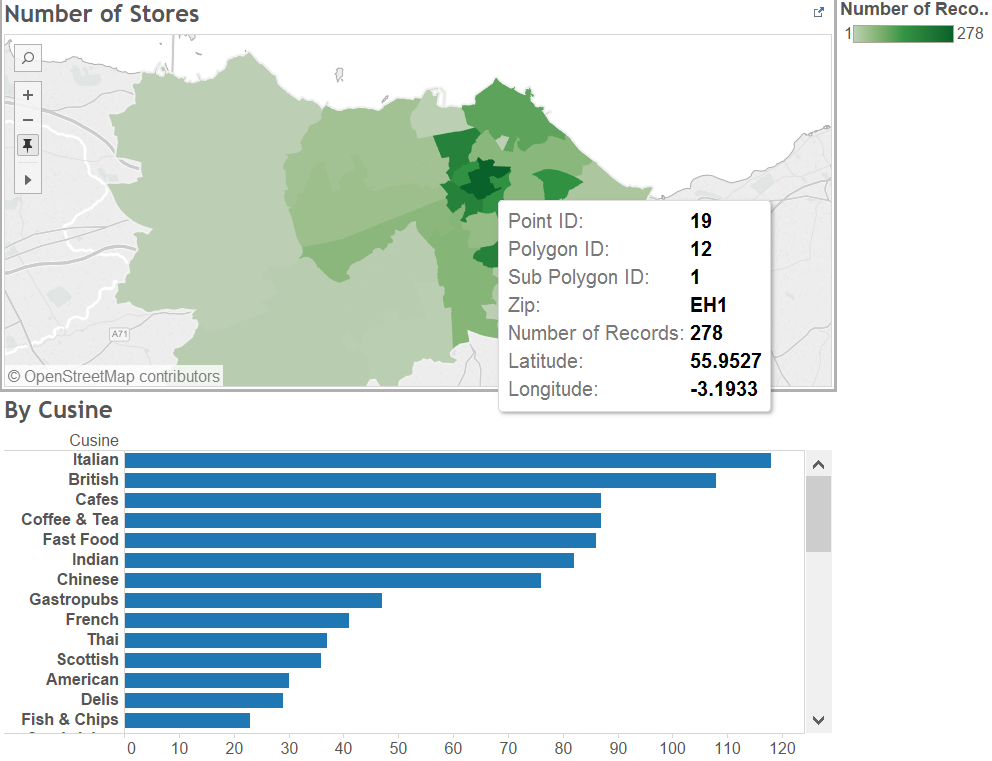
As some zip codes are really small such as EH1 and EH2, there are some zip codes that were not assigned to any single ward using the method above. For those zip codes we found wards that contained that zip divided the population numbers in two and assigned half of the population in the ward to each of the two zip codes. For example the ward central city in general encompass EH1 and EH2. So we split the population for central city into two by half and assign it to the population of EH1 and EH2. Income and property prices remained the same.

We also downloaded data on all the hotels in Edinburgh from booking.com. We cleansed this data to extract the zip code so that the data will be useful.

## Analysis

### Exploratory Data Analysis

We used Tableau to undertake an initial exploratory data analysis. Figure 3 below shows the number of stores in Edinburgh. Figure 4 shows the type of cuisine that has the most stores. We noticed that in general, Italian and British cuisine are the most popular. The city centre (zip codes EH1 and EH2) contained the greatest number of restaurants.



*Figure 3 Number of Stores Figure 4 Amount of Cuisines in Edinburgh.*

We also looked at the relationship between cuisine type and the number of reviewed. We decided to use the amount of reviews for a restaurant as an indicator of the number of visitors. We decided against using the number of users as an indicator because a user can visit a restaurant twice and give two reviews which will be counted as two visits. This information was then used as an estimate of demand within each zip code.

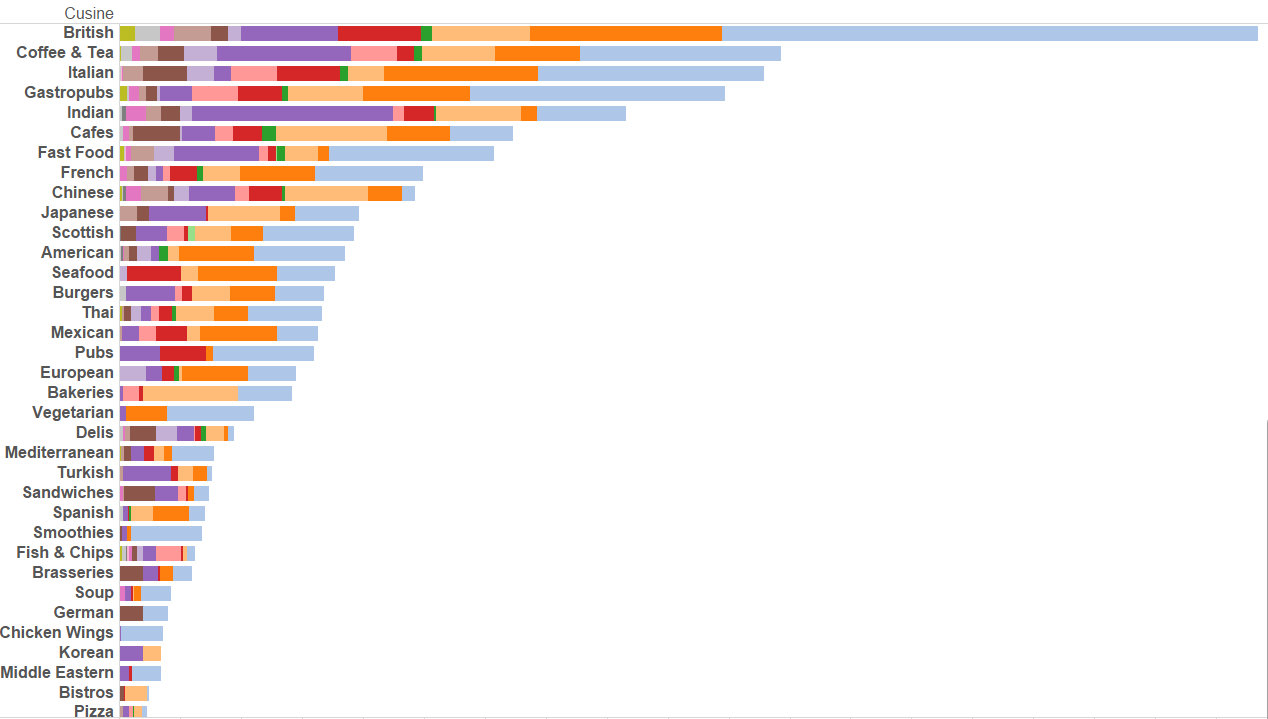


Figure 5: Cuisines with Reviews.

Figure 5 shows that British food has a very high number of reviews in Edinburgh, combined with the knowledge that there are more Italian than British restaurant in general, it might show that British restaurants are quite popular.

In certain places, such as EH2 which is shown below in Figure 6 shows that thought here are more Italian than British restaurant and so we can capitalize on the lack of British food to open a British restaurant there.

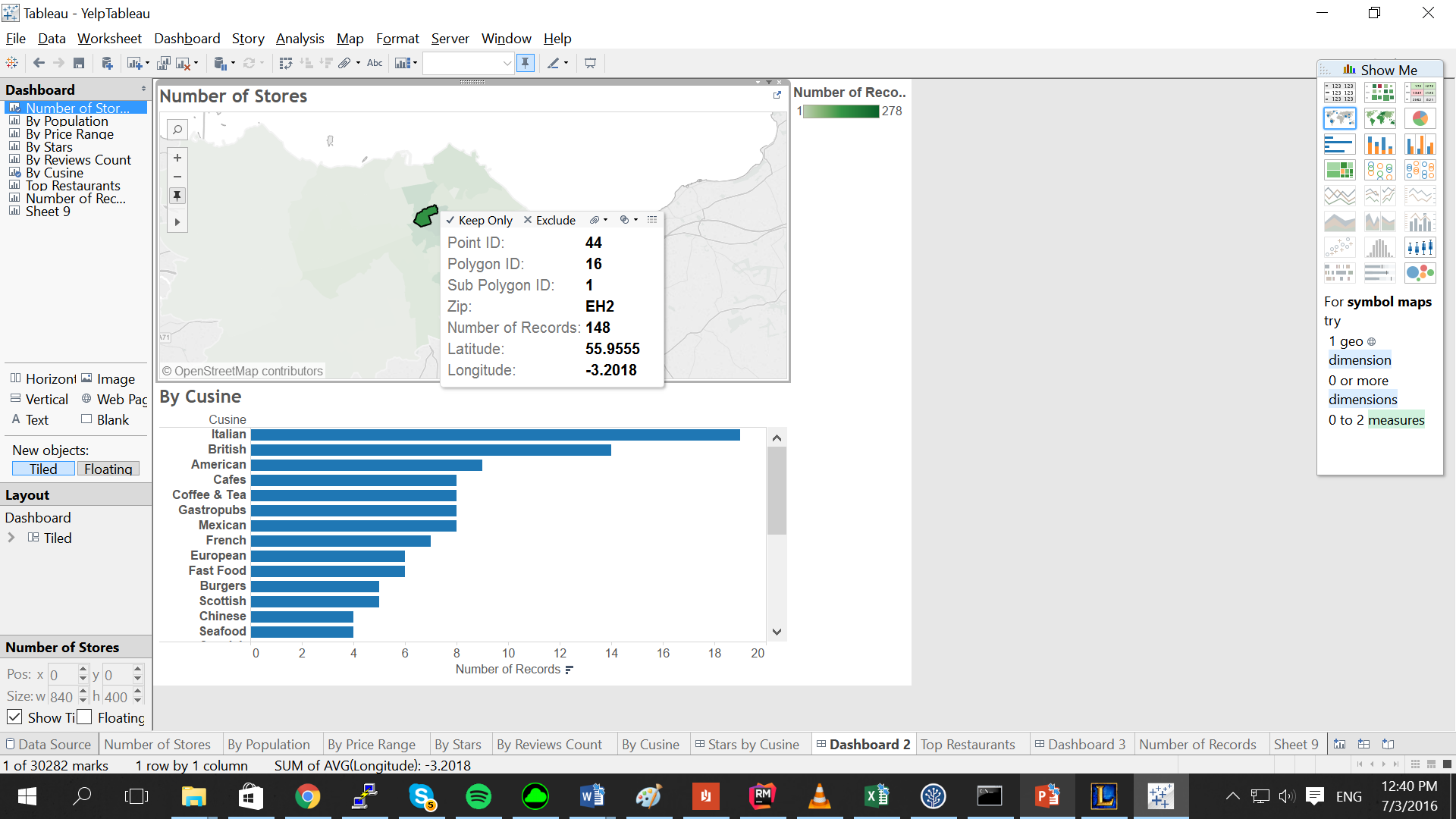


Figure 6 EH2 *3 Number of Stores and Amount of Cuisines in Edinburgh.*

Interestingly, the restaurants in the city centre (EH1 and EH2) had the lowest rating average rating (Figure [7]). An explanation for this is that there are many restaurants in the city centre and hence many low scores pull down the overall average. From this we can potentially infer that if our restaurant wants to focus on ratings, we should not try to put our restaurant in the city centre.

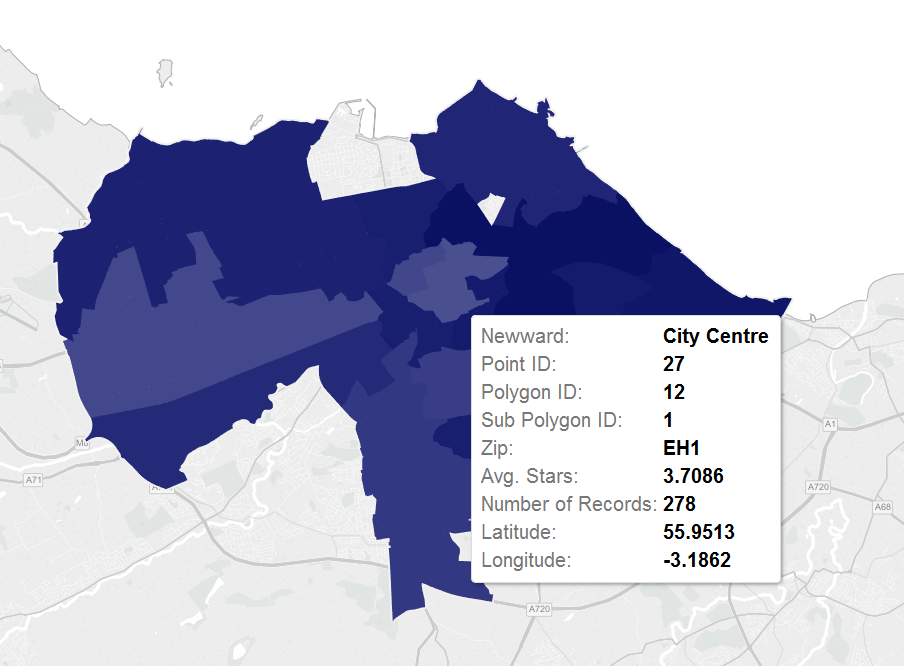
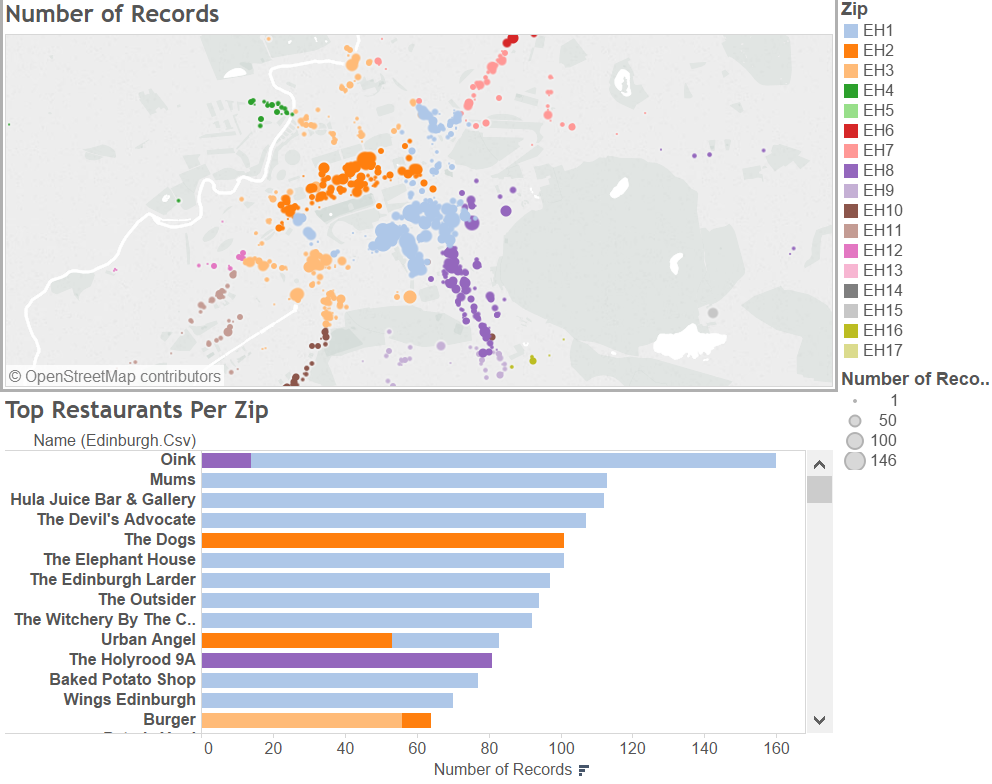


Figure 7: Average rating per zip

We decided to calculate the demand using the number of reviews. We understand the number of reviews may not be the best indicator of how many customers a restaurant had, but it served as a good proxy in the absence of more concrete data. Figure [X] shows the locations with the most reviews are EH1, EH2, EH3 and EH8. We used this information in order to focus our analysis on these areas.



We generated metrics for each zip code (table 1). We considered the ratios of: total population per zip to number of restaurants per zip; average income per zip to the number of restaurants per zip; the total number of properties to the number of restaurants per zip; and the total number of hotels to the number of restaurants per zip.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| zip | pop/rest | income/rest | property/rest | hotels/rest | num of rest |
| EH1 | 45.9 | 158.8 | 775.8 | 0.441 | 263 |
| EH2 | 85.0 | 294.0 | 1436.8 | 0.338 | 142 |
| EH3 | 130.4 | 206.2 | 961.6 | 0.478 | 184 |
| EH4 | 2272.9 | 4068.4 | 18528.4 | 0.406 | 32 |
| EH5 | 5535.2 | 10830.7 | 45313.7 | 2.333 | 3 |
| EH6 | 251.8 | 342.9 | 1421.5 | 0.515 | 99 |
| EH7 | 1028.8 | 1210.1 | 5167.2 | 1.018 | 56 |
| EH8 | 235.6 | 294.1 | 1495.0 | 0.348 | 141 |
| EH9 | 369.1 | 921.6 | 4684.4 | 0.867 | 45 |
| EH10 | 551.0 | 759.4 | 3911.3 | 0.242 | 62 |
| EH11 | 690.7 | 488.7 | 1928.8 | 0.278 | 54 |
| EH12 | 1158.7 | 2265.2 | 10028.7 | 1.103 | 39 |
| EH13 | 6059.5 | 12448.8 | 58974.5 | 0.250 | 4 |
| EH14 | 3952.5 | 7143.5 | 31195.5 | 0.500 | 6 |
| EH15 | 1335.8 | 1735.5 | 7157.7 | 0.368 | 19 |
| EH16 | 2782.7 | 2779.0 | 10926.8 | 3.083 | 12 |

*Table 1: Metrics on zip code*

The results of our exploratory analysis led us to chose to build a restaurant at a central location (EH1, EH2, EH3 and EH8). We refer to these areas as our “candidate areas”. EH8 has the largest ratio of population to restaurants compared to the other candidate areas. However, it also has the lowest number of hotels per restaurant.This may mean that there are fewer tourists in the area (who we may wish to attend our restaurant). In contrast, EH3 has attractive population to restaurant and hotel to restaurant ratios. Furthermore, its property value to restaurant ratio is low potentialy making it not too expensive to buy a restaurant there.

## Potential Revenue

In order to calculate potential revenue, we have estimated the amount of customers a restaurant may receive.

In order to do this we first considered the check-in dataset. However, we realised that it did not provide a good estimate of the number of customers as check-ins are based on the amount of offers and so represent a biased value for our analysis. The Yelp [site](http://www.yelp.com/topic/walnut-creek-what-is-this-yelp-check-in-thing), states that “Certain businesses offer discounts when yelpers check in to that business” and further states “You check in 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 it is likely that most of the customer visits will not have a check-in event (even if the customer came from Yelp) as the customer will most likely not open the application and check-in.

We therefore decided to use a mutliple of the number of reviews for a restaurant as an indicator for the number of visitors. We assumed that the number of reviews indicate approximately .1% of total visits. Finally, we found the average number of reviews per restaurant in each zip code which will provided a review (table 2).







|  |  |
| --- | --- |
| Zip | Reviews/ Restaurant |
| EH1 | 17.36 |
| EH2 | 14.99 |
| EH3 | 10.48 |
| EH4 | 4.63 |
| EH5 | 4.33 |
| EH6 | 9.04 |
| EH7 | 8.84 |
| EH8 | 12.72 |
| EH9 | 8.11 |
| EH10 | 9.60 |
| EH11 | 6.26 |
| EH12 | 4.59 |
| EH13 | 1.75 |
| EH14 | 3.00 |
| EH15 | 5.90 |
| EH16 | 5.67 |

*Table 2: Average number of reviews per restaurant by zip*

As expected our candidate areas had the highest user base. As we had decided 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 around 10000 customers per year (10.47655 \* 1000).

We used data on prices obtained from the Yelp [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) in order to calculate our expected revenue at a given level of price. The mapping of Yelp price symbols to dollar ($) value is given in table [x].

|  |  |
| --- | --- |
| Yelp Symbol | Actual Value ($) |
| $ | < 10 |
| $$ | 11 – 30 |
| $$$ | 31 – 60 |
| $$$$ | > 61 |

*Table 3: Yelp Symbol to Value*

Using the average price for each Yelp price range, we calculated that if we open a restaurant with a Yelp price rating of $$$, . we can expect to make 10,000 \* 45 which is $450,000 per year.

# References

1. Edinburgh ward and locality demographics. Sourced on June 9th 2016 from <http://www.edinburgh.gov.uk/info/20247/edinburgh_by_numbers/1393/locality_and_ward_data_profiles>
2. City of Edinburgh Mapping Portal. Source on June 9th 2016 from <http://data.edinburghcouncilmaps.info/datasets/2cee9b18a21344b0879c3c51d71fd2c6_28>
3. Yelp price symbol mapping to real values. Sources on [add date] from <http://www.yelp.com/topic/san-diego-can-anyone-give-me-the-actual-dollar-range-for-the-dollar-sign-symbols-in-rrgards-to-pricing>