CS591 Final Report
Happy Hours in Boston
Yingfeng Chen, Daniel Qiu, Peter Gunawan, Xavier Bohorquez

Note* on running the web service is also specified within this paper.

1. Introduction

"Happy Hour" are two words that people who reside in cities are always glad to hear -- whether it be about food or about drinks. And as we are happy residents of the city of Boston, it would be quite fortunate to enjoy such occasions. When it comes to the latter of happy hour drinking though, Massachusetts is one of the many states that have a ban for such a thing because of a major incident involving a young female. The state decided to ban Happy Hours on alcoholic beverages, following the death of 20 year-old Kathleen Barry in a drunk driving incident. Since then, Kansas and Illinois have lifted their ban while states such as Massachusetts, Alaska, Hawaii, and Rhode Island maintain their stance against Happy Hours. The goal or solution that we propose is to determine a set of restaurants and bars that would theoretically be safe to host the happy hours. Using a set of safety constraints determined by correlations with crime, restaurants are measured for safety relative to other restaurants in Boston. With this project, it will be determined which, if any, areas would be able to host Happy Hours for their customers with minimal adverse consequences.

2. Datasets

As a quick background on how this project came to be - initially project one started with our group as two independent groups. One of our groups had already began the happy hour idea though and started off by retrieving data about restaurants with alcohol permits around the greater city of Boston. We did a lot of transformations regarding the restaurants as we clustered the restaurants' addresses with the addresses of other datasets we retrieved about college campuses and police departments. With these datasets, we originally planned to use the restaurant zip codes as a key and then the aggregate number of colleges, public high schools, or police departments within the same zip code as the values. The second group had many datasets retrieved that related to crimes, properties, and incomes of different areas.

Our intention when choosing these datasets was that we would find the "best safety constraints" that would allow the allocated restaurants to host happy hours. Initially we thought of where in general most people would drink and also thought what situations that drinking, or drinking too much, would be a bad situation. So we decided that a prime place to target would be around college campuses because that is where a lot of people would be interested in having happy hours. From the initial datasets we found that there were not enough public high schools around the areas we were targeting to be concerned and decided to drop that dataset entirely. The dataset of properties would show how densely populated a certain area is and the general economic wealth of the community. This was a very significant dataset as we decided that it was a good complement to be transformed with the crimes dataset later on -- which will be explained later.

.

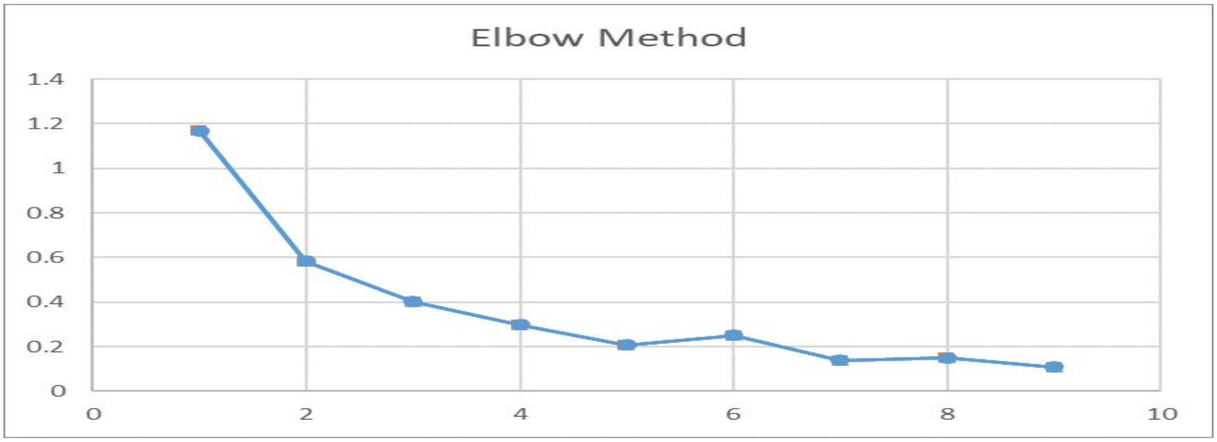
In the end, we came together and decided that only a select few of both our datasets would benefit the idea of happy hour locations fully. We also decided that it would be better to retrieve and use a restaurant dataset created by another fellow classmate because that set contained actual coordinates rather than just zip codes. We believed that using actual coordinates would be more beneficial as zip codes are a bit awkward in terms of how they are seperated. We also discovered the wonders of “GeoPy”, which needed coordinates -- so that fit perfectly into our project. We also did a new transformation for “Crimes vs. Properties” in which we projected a new set using the unique street addresses of the properties as keys and then grouping the aggregated numbers of crimes under each street as values. The rest of the datasets from both our groups were scrapped due to the scope of our project (but can be contributed to future works). Below is a list of our finalized datasets:

- Dataset of Crimes around Boston (retrieved from city of Boston datasets)
- Dataset of Properties around Boston (retrieved from city of Boston datasets)
- Transformed Dataset of Restaurant locations (retrieved from another group)
- Transformed Dataset of Crimes vs Properties (as explained above)

3. Algorithms, Techniques, and Analyses

After the above steps and datasets were completed, we decided that the most appropriate way to solve this problem would be to treat it as an optimization problem by choosing a set of restaurants based on a defined set of safety constraints created by correlations in our dataset.

Our first goal was to cluster the restaurants based on their coordinate locations because we would later be applying metrics based on location on them. Using K-means in this case was the most appropriate and efficient way to cluster them. In order to decide what the most appropriate amount of k-means needed was, we used the “elbow test” to find the optimal amount of clusters for the datasets we were dealing with. After performing the elbow test, we saw from the graph that the most optimal amount of k-means needed would be five (where the graph plateaued).



Using k-means with k=5, we created a new dataset in which it contained the coordinates of all five means -- which we will call the center or centroids of the clusters. To complete this part of the algorithm in terms of our project, we wanted to create a radius for each of these centroids. The radius would determine the maximum distance a restaurant can be from the centroid to be allowed to host happy hours. We created these radii by first modifying the k-means algorithm with Scipy -- which allowed us to store and keep track of which restaurant coordinates were associated with which centroid. In the end, we calculated the radii by simply taking the distance of the centroid to the farthest point associated with this centroid. The only problem that we encountered with this method of radius calculation was that this created a lot overlapping radii. We realized that by determining radii based on centroid to its farthest point, we were creating the possibility for an outlier point to drag a centroid's radius to large amounts.

We then moved onto our safety constraints. We had plenty of different correlations that we wanted to try and then apply as metrics, but we thought the most feasible was the streets vs. crimes. We analyzed the relationship between the amount of properties on the streets to the number of crimes that occurred there. It was discovered that the greater the number of properties, the more crime. This makes sense, as more densely populated areas will have more crime, because more people will lead to higher chances of people commiting crimes. However, this difference was relatively small. Improvements on this could still be made, which will be discussed later on in the paper.

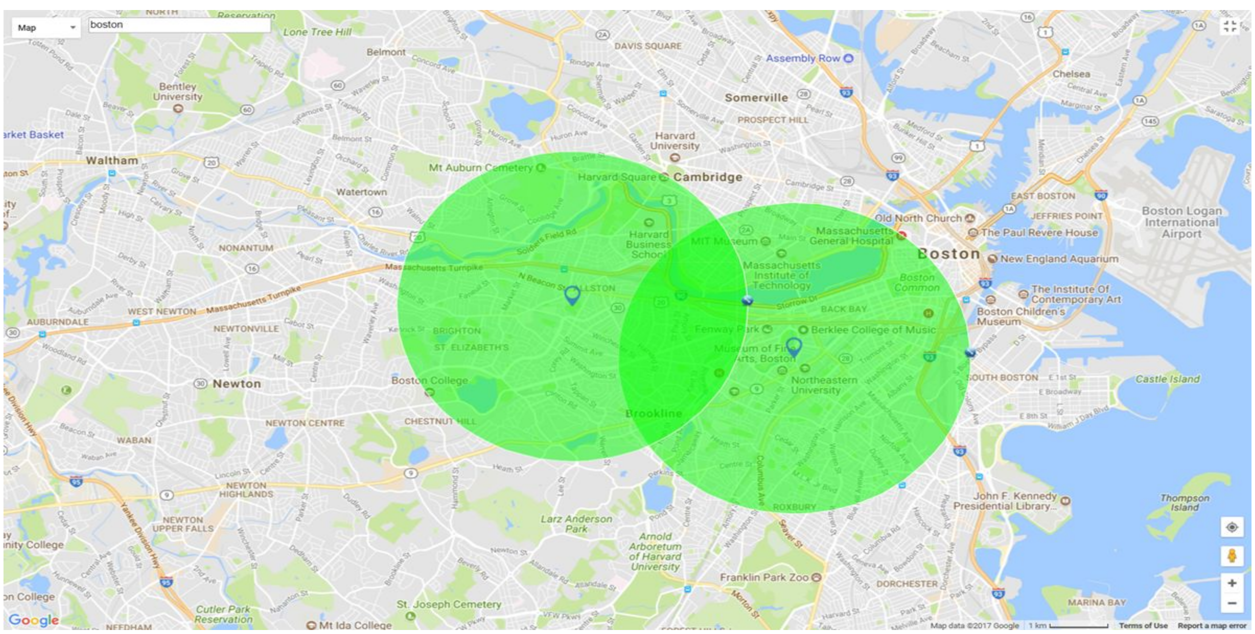
After deciding that the data showed a correlation between increasing number of properties to the increasing number of crimes, we used that as a metric to further narrow down the five clusters we had. We would choose the two optimal clusters that had the smallest radii, as that would give us a smaller inclusion of people / properties -- in hopes that these clusters would mean less chances of crime.

Generated K-Means	
Centroid (latitude, long)	Radius (km)
(42.34444150597612, -71.06143274501994)	7.056903575
(42.34152172340425, -71.09143748226946)	3.715178918
(42.286093032967045, -71.11360539560437)	6.622821797
(42.368445129629634, -71.04660057407406)	5.14957234
(42.350705047619044, -71.13836019047618)	3.505063108

Correlation Coefficients		
vs.	Low Population Density	High Population Density
Crimes	0.691001073	0.720337136

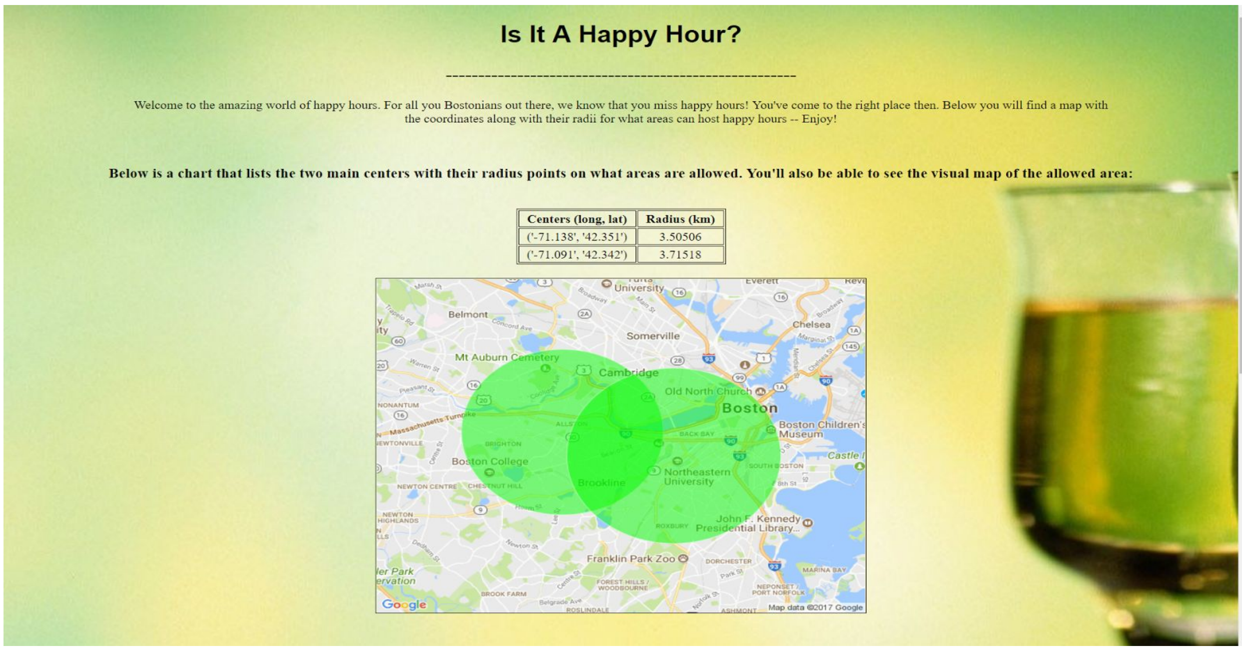
3. Conclusions and Web Service

Two clusters were identified to be the safest out of the five clusters because they had less densely populated areas. The reasoning is that if Happy Hours can potentially be dangerous in more populated areas, then hosting Happy Hour in a less populated area should reduce the chance of a crime or incidence occurring.



After concluding this, we decided that it would be useful to create an interactive web service with two components for users to see. The first component would simply just be a visual map with the radii of the two clusters on it along with their actual coordinates. This first part would allow users to see exactly what locations might possibly be able to host happy hours (good for both customers and restaurant managers to find out).

Our second feature or component on this web service would allow the user to enter in a full address into a form. This form would then use GeoPy to transform the address into coordinates on the backend and then check whether this coordinate is within either of the two specified centroids. If so, it will say so, otherwise, unfortunately the address is not allowed. (See below for picture examples)



Note* on running the web service:

In order to run this web service, we assume that the repo is already initialized and contains the datasets generated and used from project #2. Otherwises, the datasets and methods from project 2 must be generated and executed first. Once the mongoDB repo is up to date and running, the “webapp.db” must be run. This will connect and open a localhost address - presumably http://127.0.0.1:5000/, which can then be entered into the browser.

4. Future Works

The “safety” metric that we applied was chosen because of the scope of the project. To further improve the safety metric, more factors, such as the presence of schools, hospitals, and police stations, will have to be analyzed for a strong correlation to crime levels. Expanding on the set of factors that could potentially affect safety will provide a more comprehensive evaluation of each restaurants’ and bars’ eligibility to host Happy Hours.

As previously mentioned, the radii for each cluster was determined by using the distance between the centroid to its farthest associated point. As explained, this creates problems when there are outlier points. There are many other ways that we could go about this, but all with their own problems. One other way we could have considered is to take the average distance between the centroid to all of its associated points as a radius.

The web service itself could definitely be heavily improved. The first component involving the map with the two circles could probably be dynamic and possibly changed depending on which factors that the users want to apply. The second component could be improved by possibly connecting Yelp APIs to it in order to accurately display whether there actually are happy hours in that specific address.