Escape Room Profiling and Saturation

By Keith Jones, for Coursera Applied Data Science Capstone, December 12th, 2018

Introduction

One of the most recent rising trends in the world of so-called ‘experiential entertainment’ is the increase in popularity of going to ‘escape rooms’. Channeling the video game experience of searching through a virtual location and solving puzzles to collect needed objects, escape room experiences physically put small groups of people into a themed room that is filled with locks and puzzles that need to be solved somewhat sequentially so that the group can officially “escape the room”. Though still unknown to a large portion of the population, escape rooms are gaining in popularity. Due to low start-up/variable costs and strong margins, they can be an excellent business to enter into, as well. However, with growing popularity of them, these businesses seem to be popping up all over, and for someone to enter into this business market now the core question is one of whether or not the geographic market is already saturated, and what the level of saturation is.

This analysis will delve into several major US locations (Phoenix, Salt Lake City, and Houston) to analyze the venues in the neighborhoods of the most successful escape rooms (as measured by # of likes, since many escape rooms have not yet been rated on Foursquare), and determine what zip codes are similar to the zip codes of successful escape rooms, and how saturated those neighborhoods are relative to each other. This analysis can then be used to find if there are other neighborhoods that would be equally supportive of successful escape rooms that are not as saturated with escape rooms or are not currently being serviced by escape rooms, to help determine what might be strategic locations to open up a new escape room.

# Data

This project will utilize the following sources of data:

Foursquare API

The Foursquare database will be used to provide information of 3 primary types:

1. Search data about the various escape rooms within a given metropolitan area
2. Venue data about different escape rooms within that metropolitan area
3. Explore data about each zip code in the metropolitan area

Escape Room Search Data: FourSquare’s search functionality will be used to gather listings of venues in a given metropolitan area that contain the word ‘escape’ in some form in their listing. It will use query strings of the form: https://api.foursquare.com/v2/venues/explore?client\_id=ase8sf8a9823alf&client\_secret=23pouf2893u24p&v=20181120&ll= 33.6050991,-112.4052495 &radius=100000&limit=100&query=escape

Once obtained and converted into dataframe form (and unnecessary columns are removed), data will be of the form:

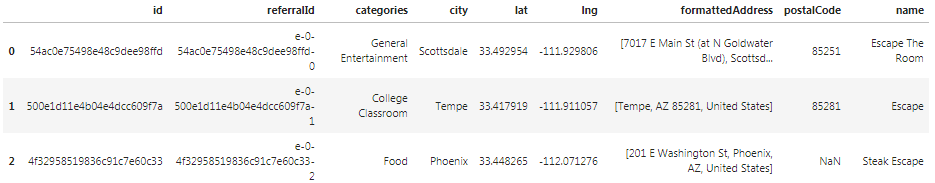


Figure 1: Sample of Escape Room Search Query Data

This data will need to be cleaned up in a few ways, most notably in that it includes venue information for venues not in the escape room industry. More about this process will be included in the Methodology section.

Escape Room Rating Data: Because many escape rooms do not yet have ratings on Foursquare, ‘likes’ will suffice for purposes of determining venue successfulness. This information is gained through queries of the following type: https://api.foursquare.com/v2/venues/189732975983?client\_id=ase8sf8a9823alf&client\_secret=23pouf2893u24p&v=20181120

Once obtained, the query results will be picked through to obtain the number of ‘likes’ if any (using the ‘response’🡪’venue’🡪’likes’🡪’count’ dictionary key progression) and the number of ‘dislikes’ if any (using the ‘response’🡪’venue’🡪’dislike’🡪’count dictionary key progression).

Zip Code Data: The analysis will require venue exploration information from the various zip codes in the chosen metropolitan area. Each zip code will require a query of the following type: https://api.foursquare.com/v2/venues/explore?client\_id=ase8sf8a9823alf&client\_secret=23pouf2893u24p&v=20181120&ll=33.4764,-112.2980&radius=500&limit=100

Once obtained and converted into dataframe form (and unnecessary columns are removed), data will be of the form:



Figure 2: Sample of Zip Code Venue Query Data

www.BestPlaces.net

Data from BestPlaces.net will be used to obtain lists of zip codes for appropriate metropolitan areas/cities. The principal section of the applicable websites that contains the needed data looks like this:

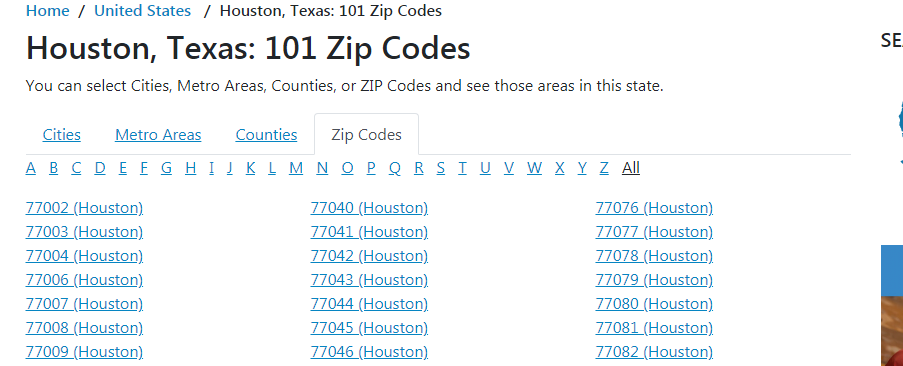


Figure 3: Sample Data from www.BestPlaces.net

The Beautiful Soup package will be used to scrape the zip codes from the data to provide us with the needed zip codes for venue profiling.

Splitwise Blog Posting on “Free US Population Density And Unemployment Rate By Zip Code”

Available at <https://blog.splitwise.com/2014/01/06/free-us-population-density-and-unemployment-rate-by-zip-code/> (accessed 12/12/18). This blog hosts free access to csv and Excel files containing population and population density information for every zip code in the United States (see example below). This information will help to indicate which zip codes in a given metropolitan area should be removed from consideration due to low population density, as well as provide helpful context regarding population densities of the remaining zip codes in the final analysis of each metropolitan area.

|  |  |  |  |
| --- | --- | --- | --- |
| Zip/ZCTA | 2010 Population | Land-Sq-Mi | Density Per Sq Mile |
| 80622 | 268 | 13.179 | 20.33538 |
| 80623 | 1028 | 0.666 | 1543.544 |
| 80624 | 946 | 31.361 | 30.16485 |
| 80631 | 48603 | 102.846 | 472.5804 |

*Table 1: Sample data from the “Free US Population Density and Unemployment Rate By Zip Code” blog post*

# Methodology

This section will discuss the following efforts that were made during the analysis:

* Initial Escape Room Data Analysis
* Population Density Analysis
* Aggregation of Venues
* K-Means Clustering by Zip Code
* Automation

## Initial Escape Room Data Analysis

One of the first challenges of getting the data I needed was trying to find as many of the escape rooms in a given area as I could. There is no specific one category that only escape rooms fall into, but not all escape rooms use the term “escape room” in their names. While not all escape rooms use the word “escape” in their title, the vast majority do, and so I used that as the query search term, to try and strike a middle ground between finding too many venues and finding not enough escape rooms. However, this means that many other venues with “escape” in their titles came through in the query data as well—for example, “Steak Escape”, a steak restaurant in Phoenix (see the picture in the *Escape Room Search Data* subsection of the **Data** section for an example of some of the types of venues that were also showing up). So one of the initial tasks was to filter this data down to only the escape room venues. While there was no cohesive category that contained all escape rooms, after performing a visual review of the data that came back, I found that the actual escape rooms tended to fall into one of five categories:

* Theme Park
* General Entertainment
* Theme Park Ride / Attraction
* Recreation Center
* Arts & Entertainment

After filtering out the other categories, I was left with a DataFrame of only escape rooms.

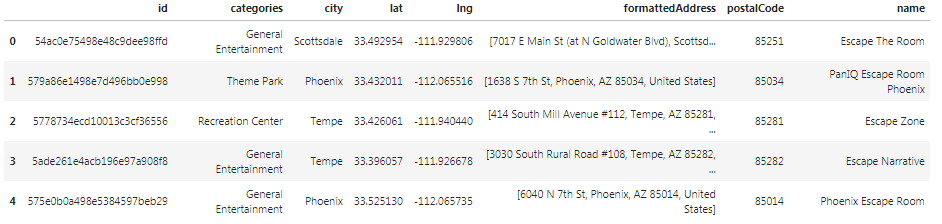


Figure 4: Sample DataFrame of Escape Rooms in the Phoenix Area

I also found, though, that due to errors in Foursquare’s latitudinal/longitudinal classifying, there was an escape room from Connecticut included in with the Phoenix data. This was fixed by screening out any rows that did not have a zip code above a certain level.

## Population Density Analysis

Data about zip code population and population density was used in two different ways:

* With relation to the escape room data to understand the lowest population density of any zip code that contained an escape room in the metropolitan area
* With relation to zip code venue data, to screen out zip codes with population densities too low to effectively support the existence of an escape room

Initial analysis revealed that if all zip codes within the metropolitan area were left in, the clustering skewed heavily because the profiles of those zip codes were so dramatically different that the vast majority of remaining zip codes were all placed in the same cluster. Implementing population density analysis helped alleviate this problem to some extent. The population density data was added to the escape room data, as well as to the broad zip code data. Using population density on the escape room data enabled visibility to the population densities of the zip codes already containing escape rooms, which enabled me to obtain a ‘recommended minimum population density’ (essentially, if we want to capture all the zip codes that contain escape rooms currently, the required zip code population density needs to be at least as high as the recommended minimum population density). This number could then be used in conjunction with the population density data attached to the zip code list to screen out the zip codes with population densities that were too low. In certain cities, the minimum population density was actually rather high, due to a lack of escape rooms in lower-density areas, so a threshold was used that was lower than the recommended minimum, because those lower thresholds still screened out more rural zip codes. Typically, 500 people per square mile was used as the threshold if the recommended minimum wasn’t lower than that number.

## Aggregation of Venues

There were several steps involved with aggregating the venue data in preparation for clustering analysis. These steps were as follows:

1. Obtain zip code list for the selected metropolitan area (see the *BestPlaces.net* subsection of the **Data** section for more on this). This involved using the Beautiful Soup python package for web scraping to get the correct table of data from the website.
2. Add latitudinal/longitudinal data and population density data to the zip code list, and then remove zip codes with population densities lower than the desired threshold (see the *Population Density Analysis* subsection of this section). After this step, we had a zip code dataframe that looked like the following:

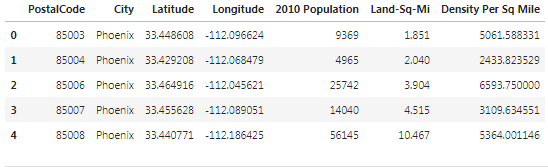


Figure 5: Sample DataFrame of Zip Code and Population Data

1. For each zip code on the list, query the Foursquare database to get names and categories of the closest venues to each of those zip codes. These venues were all aggregated into a new dataframe along with basic zip code data.
2. One-hot encoding was used on the categories column to create a dataframe of the category encoding for each venue. This data was then grouped and averaged by zip code. At this point, the data was ready for cluster analysis

## K-Means Clustering by Zip Code

The overarching goal of this analysis is to determine what the profile is of the zip codes that contain the most successful existing escape rooms, so that other zip codes with similar profiles can be identified for potential development of future escape rooms in zip codes not already containing escape rooms. The best machine learning method for this type of problem is K-Means Clustering.

While trying to strike a balance between obtaining useful categorization and not overfitting the model, depending on the number of zip codes in the data set I chose either 10 or 20 clusters. The one-hot encoding data from the previous section was used as the data set, and the clustering grouped them into the appropriate number of clusters. The labels were then added back in to the main zip code dataframe, as a record of which zip code belonged to which profile group.

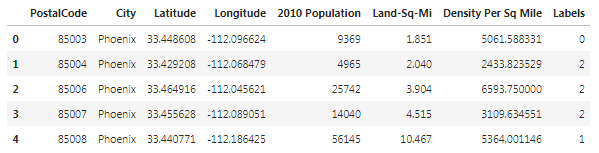


Figure 6: Sample Final Zip Code Data with Labels Column Containing Profile Groupings

This data was then connected back to the escape room data to categorize each of the escape rooms by what profile group they belonged to. Finally, the profile groups of the top 2 most successful escape rooms were identified and the list of zip codes was filtered down to only those belonged to profile groups (clusters) that the top 2 escape rooms represented. To round out the analysis, the escape room data was grouped and counted by zip code to determine how many escape rooms are currently in each zip code, and that information was added on to final analysis. The resulting dataframe is a list of zip codes that match the profile of the zip codes that house the top 2 most successful escape rooms, along with information about population, population density, and the number of existing escape rooms (see below). A potential escape room developer could take this list and determine what zip codes to target for development, based off of their profile for supporting successful escape rooms and the current saturation of escape rooms in those zip codes.

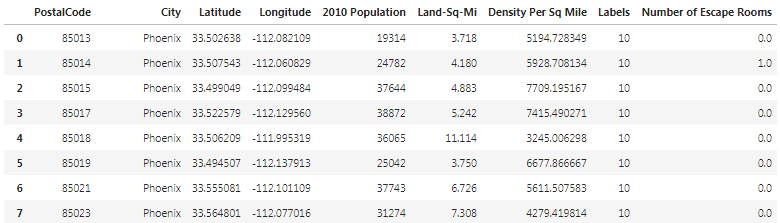


Figure 7: Sample DataFrame Containing Final Zip Code Analysis Data + Escape Room Saturation Data

## Automation

After going through the initial analysis for the Phoenix area, it seemed that there was significant opportunity to automate a lot of the analysis and package it all in a small number of functions. I was already using two functions during my initial analysis: one that extracted the venue category id from a table entry that contained lingering JSON features, and another to extract the venue category from similar table entries. After some planning and analysis of the connections between the various code cells, I consolidated everything into five additional generalized functions:

1. **latLong –** a function to convert a text representation of a location into a list of the latitude and longitude value of that location
2. **getEscapeRooms –** a function to return a dataframe with initial data about the escape rooms in a given metropolitan area (see Figure 8, below)
3. **getZCData –** a function to return a dataframe with data about all zip codes 1) in a given metropolitan area and 2) with population densities above a certain threshold, as well as the group labels for the clusters that these zip codes were partitioned into, using K-Means Clustering (see Figure 6 for an example of this output)
4. **analyzeEscapeRooms –** a function that takes the results of functions 2 and 3 above and returns a dataframe containing the final analysis of the zip codes that correspond to the profile groups of the top 2 most successful escape rooms in the area, as described in the *K-Means Clustering by Zip Code* subsection of this **Methodology** section (see Figure 7 for an example of this output)
5. **mapEscapeRooms –** a function that takes the results of functions 2 and 3 above and returns a folium map object that maps all the zip codes, color-coded by profile group, and contains pop-up labels for all of the identified escape rooms in the area (see Figure 9 below for an example of this output)



Figure 8: Sample DataFrame with Escape Room Data for the Phoenix Area

The addition of these functions made it significantly easier to analyze other metropolitan areas with just a few lines of code.

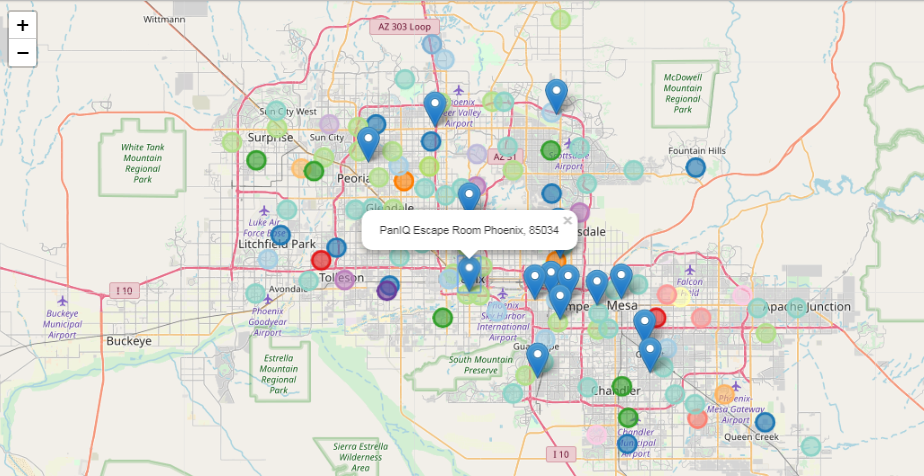


Figure 9: Sample Map of Color-Coded Phoenix Zip Codes and Popup Labels for Escape Room Locations

# Results

## Phoenix Analysis

In the Phoenix area, there were 16 existing escape rooms. There was a great deal of stratification within the escape room data set, as one escape room had 32 likes, one had 5 likes, and the remaining 14 all had 2 or fewer likes (most had 0). Some of the escape rooms were in somewhat low-population-density areas, so the minimum population density that I used for filtering was the recommended minimum: 488.706 people/square mile. Out of the 155 zip codes in the Phoenix/Mesa metropolitan area, only 110 made the final cut for analysis. For the K-Means Clustering of the Phoenix data, I used 20 clusters, and the groupings were as follows (the groups that corresponded to the top 2 escape rooms are highlighted in green):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Profile Group | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| # of Zip Codes in Group | 10 | 9 | 23 | 7 | 3 | 2 | 1 | 2 | 1 | 1 | 39 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 3 |

*Table 2: Phoenix Cluster groups and the quantity of zip codes in each group*

There were 42 zip codes identified by the final analysis as being comparable profiles to the profiles belonging to the top 2 escape rooms. Of these zip codes, 6 already have 1 escape room, and 1 zip code already has 2 escape rooms. For those who are interested, the 42 zip codes were: 85013, 85014, 85015, 85017, 85018, 85019, 85021, 85023, 85031, 85032, 85035, 85048, 85050, 85051, 85086, 85120, 85140, 85143, 85201, 85203, 85208, 85210, 85224, 85233, 85251, 85257, 85258, 85259, 85260, 85284, 85295, 85296, 85298, 85302, 85303, 85305, 85307, 85310, 85323, 85338, and 85375.

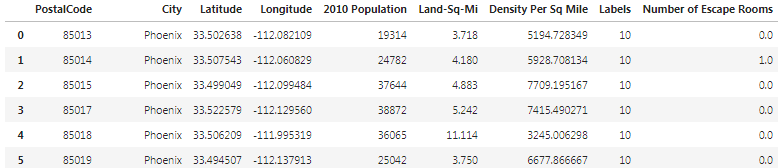


Figure 10: Top 6 Rows of the Phoenix Analysis Data

## Salt Lake City Analysis

In the Salt Lake City area, there were 4 existing escape rooms that the Foursquare search found. There was a good distribution of likes amongst the four—1, 3, 4, and 0 likes respectively among the escape rooms in the data set. The recommended minimum population density was 1,578, which is plenty high, so I went with a lower value of 500 people/square mile. Out of the 46 zip codes in the Salt Lake City metropolitan area, only 28 made the final cut for analysis. For the K-Means Clustering of the Salt Lake City data, I used 10 clusters, and the groupings were as follows (the groups that corresponded to the top 2 escape rooms are highlighted in green):

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Profile Group | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| # of Zip Codes in Group | 18 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 |

*Table 3: Salt Lake City Cluster groups and the quantity of zip codes in each group*

There were 20 zip codes identified by the final analysis as being comparable profiles to the profiles belonging to the top 2 escape rooms. Of these zip codes, only 3 already have 1 escape room. Once again, if you’re interested, the 20 recommended zip codes were: 84020, 84047, 84070, 84088, 84093, 84101, 84102, 84103, 84104, 84105, 84106, 84107, 84111, 84112, 84113, 84115, 84117, 84119, 84120, and 84124.

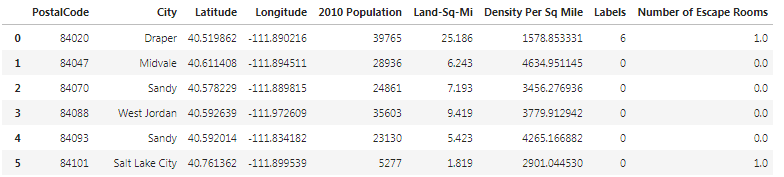


Figure 11: Top 6 Rows of the Salt Lake City Analysis Data

## Houston Analysis

In the city of Houston, there were 8 existing escape rooms that the Foursquare search found. There was wide stratification within this escape room data set, as well: the top 2 escape rooms had 19 and 17 likes respectively, one escape room had 7 likes, and the rest had 1 or 2. The recommended minimum population density was plenty high at 2,543 people/square mile, so I used 500 people/square mile for this threshold as well. That meant that out of the 101 zip codes in Houston, most (98) made the final cut for analysis. For the K-Means Clustering of the Houston data, I used 20 clusters, and the groupings were as follows (the group [only one in this case] that corresponded to the top 2 escape rooms is highlighted in green):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Profile Group | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| # of Zip Codes in Group | 1 | 10 | 4 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 65 |

*Table 4: Houston Cluster groups and the quantity of zip codes in each group*

There were 65 zip codes identified by the final analysis as being comparable profiles to the profiles belonging to the top 2 escape rooms. Of these zip codes, 7 already have 1 escape room. And if you’re interested in the recommended zip codes, they’re: 77002, 77003, 77004, 77006, 77007, 77008, 77009, 77010, 77011, 77012, 77016, 77018, 77019, 77020, 77021, 77023, 77024, 77025, 77026, 77027, 77031, 77032, 77035, 77036, 77037, 77040, 77041, 77042, 77044, 77045, 77046, 77051, 77054, 77056, 77057, 77058, 77059, 77060, 77061, 77063, 77064, 77067, 77068, 77070, 77072, 77074, 77077, 77079, 77081, 77082, 77084, 77085, 77086, 77087, 77088, 77090, 77092, 77093, 77096, 77098, 77339, 77345, 77388, 77429, and 77598.

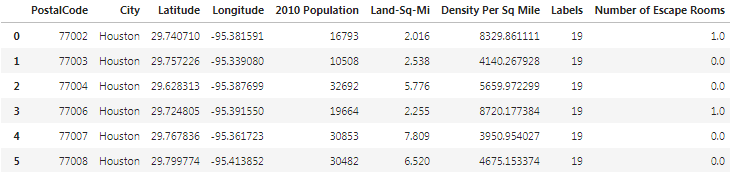


Figure 12: Top 6 Rows of the Houston Analysis Data

# Discussion

On the surface, it would appear that there are 3 lists from 3 different cities of good candidate zip codes to build future escape rooms in, there are some important considerations to keep in mind.

**Foursquare Data Set:** The Foursquare data set is not particularly robust with regards to escape rooms. Of the initial Phoenix data set, only one escape room had an actual rating generated by user reviews. ‘Quantity of Likes’ became a reasonable proxy for this data, but it raises the question of whether or not better ratings data could have been obtained using a more utilized location database, like Yelp.

**Blindness of Zip Code Profiles:** While the zip code profiles that were generated by the K-Means Clustering were helpful, they are likely blind to certain one-shot types of things. In an analysis like this, the types of categories that hold the most sway are going to be the kind of categories that have multiple venues of that category, which will typically be things like restaurants or stores. For some types of locations, such as an airport or prison or capitol building, there’s likely only going to be one in a given zip code, and most zip codes won’t have any, and K-Means clustering would be effectively blind to this. It might change a potential proprietor’s mind to find out that a given recommended zip code also contains a maximum security prison, or a nuclear power plant, or a haunted graveyard (though that last one, depending on the theme of your escape room, could actually help you).

**K-Means Clustering Consistency**: K-Means Clustering is not wholly consistent. You can run the same data through it multiple times and get different clusterings, depending on where the initial centroids are chosen. This means that any given set of recommendations is not necessarily the ‘golden answer’.

**Zip Code Variance:** Despite screening out zip codes with low population densities, the cluster groups still tended to be very lop-sided in their distribution: one (or maybe two) cluster groups contained most of the zip codes. It is unfortunately difficult to tell whether this indicates 1) that there are large numbers of zip codes that are functionally similar and would provide the same type of ‘neighborhood experience’, or 2) that there were still too many ‘wildcard’ zip codes that were skewing data so that the urban zip codes all still looked basically the same. Deeper analysis in the future could do much solve out some of this ambiguity.

**Escape Room Variability**: This analysis presumes much about the role that the surrounding neighborhood plays in helping make an escape room successful. And while that no doubt has a strong impact, this analysis does nothing to take into account the quality of the actual escape room and its staff, or the level of marketing that the escape room does. While a lack of marketing could potentially be overcome by choosing your location well, failure in creating a good experience for patrons could easily outweigh the benefits of a well-chosen and beneficial location. Location is merely one part of the equation: you also have to execute well.

**Recommendations for Future Analysis:** A couple things could help improve the analysis conducted in this project. Utilizing location data that has better ratings data for escape rooms would help with accuracy. To improve the problems with certain clusters containing the majority of the zip codes, I would recommend doing initial analysis as I did it here, but then taking those core majority-holding clusters and re-running the clustering on just those zip codes. Then the model might result in even better analysis and ability to recommend helpful locations. Finally, it might help to determine in advance the location of structures (like prisons) that the proprietor wants to avoid, and overlay that on the analysis in some form. These additional steps would help strengthen the potential results from this analysis.

# Conclusion

The escape room industry is increasing in popularity and decreasing in obscurity, and someone with only a modest amount of capital can start up their own escape room, but it can be difficult to know where to put one. What locations in a given city will be conducive to facilitating patronage of the room, and what locations will consign the business to geographic obscurity? The analysis carried out in this project has aimed to analyze the locations in which successful escape rooms are located, and determine what other locations within that city are of a similar profile of neighborhood (and ideally, don’t already have escape rooms in them). This analysis was carried out over three metropolitan areas, Phoenix/Mesa Arizona, Salt Lake City Utah, and Houston Texas. Though there is helpful analysis contained within the results, there are other things that a potential proprietor should consider (such as their business plan, and what other types of ‘one-off’ structures might also be in a given zip code) before they lock down on a location. With a few improvements, the analysis can become even stronger and provide more helpful information to potential escape room proprietors.