**Introduction**

The audience is a hospitality company that manages many restaurants around the US. This company wants to open a new restaurant and maximize the likelihood of the success of this restaurant. This analysis will recommend a location to open up a restaurant. This project takes on a broad scope in that is considers all major cities in the country. I will first identify the vibrant, up and coming cities as the first filter for narrowing the candidates. Of those cities, I will identify which cities may be underserved, from a gastronomic perspective. Next, I will choose one city, and identify which types of restaurants might add an exciting new element to their restaurant scene. Finally, I will conduct location-based clustering and recommend a specific cluster associated with a location to open in the target city.

**Data**

Three different data sources will be utilized. First, data from the largest cities in the US is scraped from Wikipedia. Next, the zip codes from the filtered cities is downloaded from a Zip Code API web service. Once all the zip codes for each city have been obtained, they will be appended with their corresponding latitude and longitude. Finally, the venues associated with these zip codes will be extracted from Foursquare. The data from Foursquare will be analyzed to make the final recommendation.

**Methodology**

First, the list of potential cities is compiled by scraping a Wikipedia page. The data is then cleaned such that it can be filtered based on population, percent change in population and density. The python package Geocoders is used to assign the latitude and longitude of each city. Next, the data is filtered to identify the target cities. I determine the city must contain at least 50000 people and contain a density above the 25th percentile to be a sufficiently large and dense city. From this filtered data, I sort the rows according the fastest growth rate. I decide to analyze the top 5 cities with the fastest growth rate. I consider growth rate to be the most important factor because it provides useful information about what the city might look like in the future. If the city is growing fast, it will be larger and have more potential consumers for this restaurant. This filtering yields the cities Seattle, Denver, Charlotte, Austin and Fort Worth.

Next, I use these city names and a zip code API to return all the relevant zipcodes for each city. I use a python package “zipcodes” to add a latitude and longitude to each zip code. Now that I have a latitude and longitude for each zip code, I can download the venue information for each zip. I communicate with foursquare to download all the venue information for each of the five cities. I downloaded the venue information; I inspect all the unique venue name to create a filter to classify each venue as a food related venue or not. Once I have updated this classifier to the data set, I determine what percentage of the venues are food related venues. I determine the city with the lowest percentage of food related venues is the city where it makes the most sense to open a restaurant. There will be less competition and therefore, hopefully more business. I identified Fort Worth has having the lowest composition of restaurants at 41%. However, Fort Worth is very close to Dallas. If Dallas had a glut of restaurants, Fort Worth could lose potential customers to Dallas restaurants. Therefore, the restaurant composition of Dallas is also analyzed. Dallas appears to fall roughly in the middle for percentage of restaurants, therefore, the proximity of Dallas will not deter me from recommending Fort Worth.

Next, I classify all the venues in Fort Worth based on K Means classifying. I chose this method because I knew I needed an unsupervised classifier given I did not have labeled data. I wanted to create these clusters according to location, because I want to recommend a specific location in Fort Worth, therefore, I only fed in the latitude and longitude into the clustering algorithm. I chose 8 clusters because it provided a few viable different clusters relatively close to the center of Fort Worth. Then I determined the percentage of food venues by cluster and found cluster 5 to contain the least amount of food venues at 23.8%. Therefore, I chose cluster 5 as the location of the restaurant. Finally, I need to recommend a type of restaurant. I determined the least frequently occurring venues for cluster 5 and Fort Worth as a whole. I found category of Japanese Restaurant to form an intersection of these two subsets, therefore, I chose to recommend a Japanese Restaurant.

**Results and Discussion**

Based on the results of the analysis I recommend opening a Japanese Restaurant in Fort Worth. Specifically, I recommend opening this restaurant within one of 5 zip codes: 76114, 76116, 76135, 76179, 76127. This conclusion was reached as a result of identifying that Japanese restaurants had a small presence in cluster 5 and in Fort Worth as whole. It is important to note the limitations and uncertainties of this analysis. The primary source of uncertainty was the number of venues returned by Fort Worth. Fort Worth returned 3 to 5 times less venues than the other cities considered. This confidence depends on the confidence we have in the foursquare data. We are confident in the foursquare data so it is possible we missed zip codes that would have provided more venues. Additionally, the consideration of the proximity of Fort Worth to Dallas. Dallas was included in the restaurant ratio analysis because I determined if Dallas as overserved from a restaurant perspective, that would distort the low ratio for Fort Worth because restaurants in Fort Worth would have to compete with restaurants in Dallas.

Given Dallas, fell in the middle of the pack, I determined Dallas to be a wash and not affect the decision to choose Fort Worth. I decided to use a location-based Clustering algorithm because my recommendation involved the physical location within Fort Worth. Therefore, I used KMeans to group the venues by location. I chose 8 clusters in order to provide multiple options for the location of this restaurant. Note, this choice was relatively arbitrary and could have benefitted from additional scrutiny. Doubts about sufficient venue data was the primary reason the number of clusters was not scrutinized further. Finally, I needed to decide what type of restaurant to recommend to build in cluster 5. I made the simplifying assumption not to recommend a type of restaurant that was not already present in Fort Worth. I also assumed the residents of Fort Worth have roughly as similar preference for all restaurants in Fort Worth. This is obviously a huge, unrealistic assumption; however, significant steps would have to be taken to ascertain the restaurant preferences of Fort Worth residents. This would require some sort of survey or restaurant performance data to determine which restaurants perform the best in Fort Worth. Given my assumptions, I looked for restaurants that were underrepresented in Cluster 5 and Fort Worth as a whole. The restaurant of type "Japanese" satisfied this union of low occurrence in cluster 5 and Fort Worth overall. This union was decided so this restaurant would have less competition to consumers in cluster 5 and Fort Worth. It must be noted that there is reasonable possibility that a type of restaurant has a low number of occurrences because this type of restaurant may be unfavorable to its residents. However, this cannot be concluded with more data.

**Conclusion**

This analysis consisting of filtering the largest cities in America and identifying the 5 fastest growing cities that contained as least 500000 residents and met the 25th percentile of population density within potential cities. The cities I identified were Seattle, Charlotte, Austin, Fort Worth and Denver. For each city, I obtained a list of Zip codes by accessing a Zip Code API. Next, I used a python package called "zipcodes" to add a latitude and longitude to each zip code. Using these coordinates, I accessed foursquare to return all the venues near these coordinates. I defined a list of venues which fell into the criteria of food. I then determined what percentage of the total venues were food venues for each city. Fort Worth contained the lowest ratio, at about 41%, therefore, I chose this city as the target. I took all the venues and used KMeans to create 8 clusters. I identified 5 clusters near the center of the city to determine was the final location for the restaurant. I again determined the food venues as a percentage of the total venues for each cluster to determine the optimal location for this restaurant within Fort Worth. I identified the fifth cluster as containing the least number of restaurants, at 23.8%. Next, I determined the least frequently occurring venues with at least one occurrence in Cluster 5 and Fort Worth as a whole. I found Japanese Restaurants as one of the least frequently occurring food category in both subsets, therefore, I made the recommendation to open a Japanese Restaurant within Cluster 5 in Fort Worth, Texas.