Who eats What?

Machine Learning Capstone Project

Sohrab Rahimi
September 23rd, 2017

I. Definition

Project Overview

In this project I will discover different demographic factors that have to do with people's taste in different zip codes across the United States. To this end, I will first identify restaurants with similar clientele/purchased food and divide them up into a few groups. Next I will focus on zip codes with at least 10 restaurants and see how the composition of the restaurants are in terms of the previously defined groups (e.g. 10% group 1, 25% group 2, etc.). This information will help us to see which demographic factors are associated with the composition of restaurants in different zip codes.

Previous research has discussed the association of demographic factors and taste with. In his well-known book, Distinction: A Social Critique of the Judgment of Taste, [1] argues that difference in social class are most obvious in the routine everyday choices such as taste of food, furniture, and clothing as they are representative of the pure taste. Many other studies use Bourdieu's argument for designing recommender system (i.e. algorithms made for recommending products to users) [2], [3]. In this study, I will specifically focus on taste of food, in order to see which demographic factors are associated with that. This information helps urban planners and restaurant owners to understand which factors are most associated with restaurant allocation.

Problem Statement

In this study I will try to investigate socio-economic factors that influence the restaurant allocation in the USA. The major purpose of this study is to see which demographic factors affect the type of taste practiced in a neighborhood. In this study, we assume that the type of restaurants in a given area representative of people's food preference in that area. The results of this study helps us better understand different socio-economic factors associated with taste. Studying the types of restaurants and practiced taste in different areas is also useful for health studies related to diabetes and obesity.

Metrics

The first step of this research will be to find the clusters of restaurants with similar clientele/purchased food. We can then quantify the restaurant composition in a given zip code as a list of continuous variables where every variable represents the proportion of restaurants in every zip code that falls in previously defined clusters. Now we would want to see what demographic factors are associated with these composition of restaurants in different zip codes. The simplest answer to this question would be a multivariate regression where our response variables are the continuous variables that we just calculated and our independent variables will be a list of demographic factors that we extract for each zip code. The evaluate different models, we can use R-square and P- value that are conventionally used to evaluate regression models.

Since we have more than one response variables, chances are that the regression model will end up with low R-square, that is, the regression model fail to explain the variations in data-set, properly. As an alternative solution, we can represent all the response variables as one categorical variable by using clustering techniques on the Y matrix (i.e. response variables). We can then use the resulting labels as our targets for a classification algorithm. We can then analyze the importance of every factor by looking at its importance. The measure of importance may differ in definition depending on the kind of classification algorithm that we end up choosing. For example, if we choose random forest, feature importance will be mean decrease impurity.

II. Analysis

Data Exploration

Yelp Data-set:

Our Yelp review table looks like below:

Business_id	User_id	Review	Name
_2HFbrYZFp2p59hptZeDbw	2MfOxr17iC_jRqERBUQNAg	I actually swore off subway for awhile after I...	Subway
_2HFbrYZFp2p59hptZeDbw	DInIFTmJWQR-SYL7EcPdbA	I don't know why I don't think about Subway mo...	Subway

Table1. Yelp reviews table

Using this table, we can tell which users have stopped by which businesses. To this end, we will n eed to use one-hot coding to apply 1 if a user has stopped by a business and 0 otherwise. It is im portant to note that the original Yelp reviews table includes 2,984,624 observations for 6,140 rest aurants. However, the majority of businesses have only a few reviews (less than 5) which is not us eful for our purposes. After filtering those, 44,051 reviews will remain for 1,872 businesses.

ACS Data-set:

This data includes the following variables for all US zip codes: 'churches', 'colleges', 'govOffices', 'hospitals', 'libraries', 'museums', 'recAreas', 'schools', 'shoppingCenters', 'Total Pop', 'White', 'Black', 'Asian', 'Other Race', 'Male Pop', 'Female Pop', 'sixty_plus', 'below_eighteen', 'eighteen_to_sixty', 'household_size', 'non_family', 'median_income', and 'education'.

First, any row with "NaN" values were dropped. Second, some variables were normalized on the total population in that zip code, these variables are: 'churches', 'colleges', 'govOffices', 'hospitals', 'libraries', 'museums', 'recAreas', 'schools', 'shoppingCenters','White', 'Black', 'Asian', 'Other Race', 'sixty_plus', 'below_eighteen', 'eighteen_to_sixty', 'household_size', 'non_family', 'education'.

Also, two new variables were created: sex_ratio which is the Male population divided by Female population, and population density, which is the total population divided by zip code area. After this step, the following variables were deleted in order to finalize the independent variables matrix: 'Total Pop','Male Pop','Female Pop','area'. We also removed any row with NA values. Of 14575 rows, only 4,622 remained.

	churches	colleges	govOffices	hospitals	libraries	museums	recAreas	schools	shoppingCenters	White	Black
count	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000
mean	0.099899	0.006269	0.041191	0.008327	0.009712	0.015474	0.014076	0.064810	0.002794	79.225425	10.875235
std	0.144480	0.026643	0.406435	0.029887	0.022422	0.095800	0.574238	0.150557	0.065156	20.220493	17.335094
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.873407	0.000000
25%	0.018131	0.000000	0.005009	0.000000	0.000000	0.000000	0.000000	0.025318	0.000000	71.108868	0.851701
50%	0.054755	0.000000	0.018396	0.000000	0.004630	0.003704	0.003306	0.044668	0.000000	86.076741	3.530378
75%	0.132885	0.006211	0.042623	0.006825	0.011785	0.013989	0.010978	0.073669	0.002301	94.303145	12.593483
max	5.882353	1.077375	47.058824	1.581028	1.169591	7.692308	69.230769	10.650888	7.692308	99.978388	98.667611

	Asian	Other Race	sixty_plus	below_eighteen	eighteen_to_sixty	household_size	non_family	education	pop_density	sex_ratio
count	14575.000000	14575.000000	14575.000000	14575.000000	14575.000000	14505.000000	14575.000000	14575.000000	14575.000000	14575.000000
mean	3.241905	3.070439	21.577641	22.586765	55.815856	0.044986	13.012463	27.083500	0.083356	1.007867
std	6.050717	5.719058	7.761143	5.602588	7.530407	0.194083	6.348691	16.224199	0.233843	0.553017
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.002649	0.000000	0.000000	0.000018	0.317647
25%	0.304457	0.259576	17.099335	19.899465	52.187093	NaN	9.341410	15.198289	0.003156	0.916788
50%	1.161698	1.055638	21.099613	22.799075	55.197518	NaN	12.237186	22.296192	0.014956	0.966183
75%	3.439808	3.217514	25.196970	25.699151	58.398433	NaN	15.358747	35.299024	0.087160	1.019956
max	73.317148	73.137698	99.818182	60.030000	99.973205	12.121212	99.218750	99.975388	5.778294	33.000000

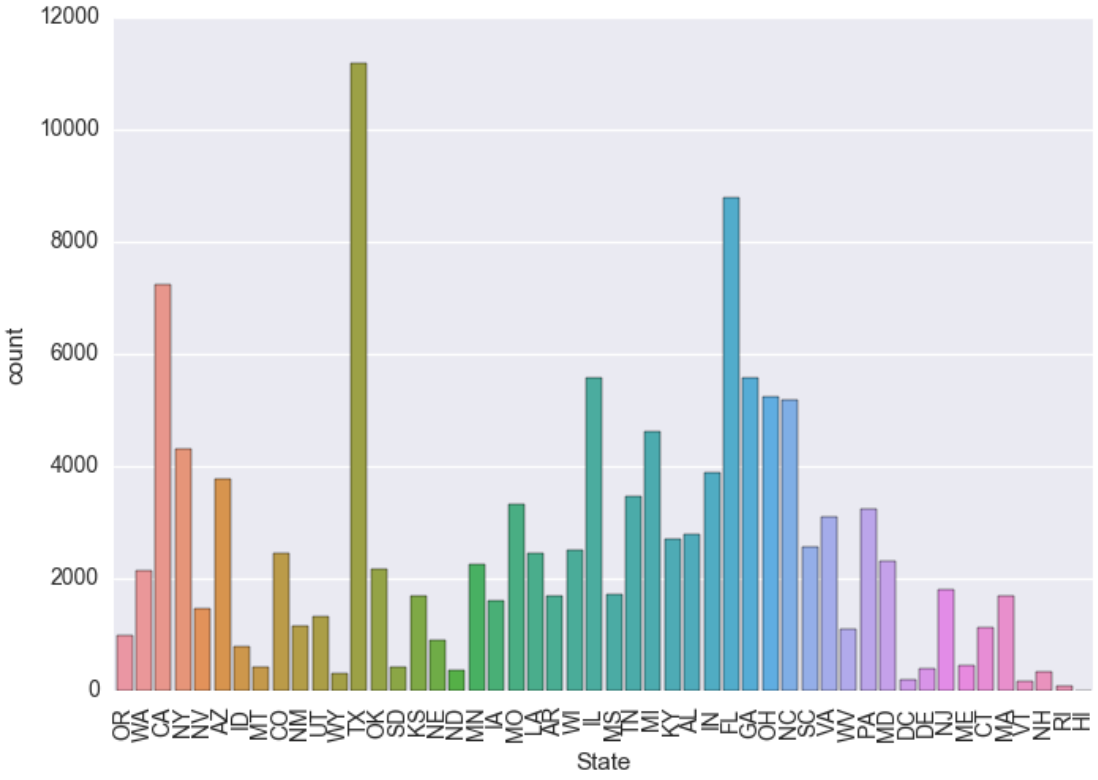
Yellow Pages Data-set:

This data-set includes the following variables for 796,763 restaurants: unique ID, Name, Category (e.g. Japanese, Mexican), Zip Code, Address, Longitude, and Latitude. I used ArcGIS 10.2. and the Zip code shapefile downloaded from here. After importing both data-sets in the Arcmap environment, I conducted a spatial join to see which zip code each restaurant belongs to. This join also a

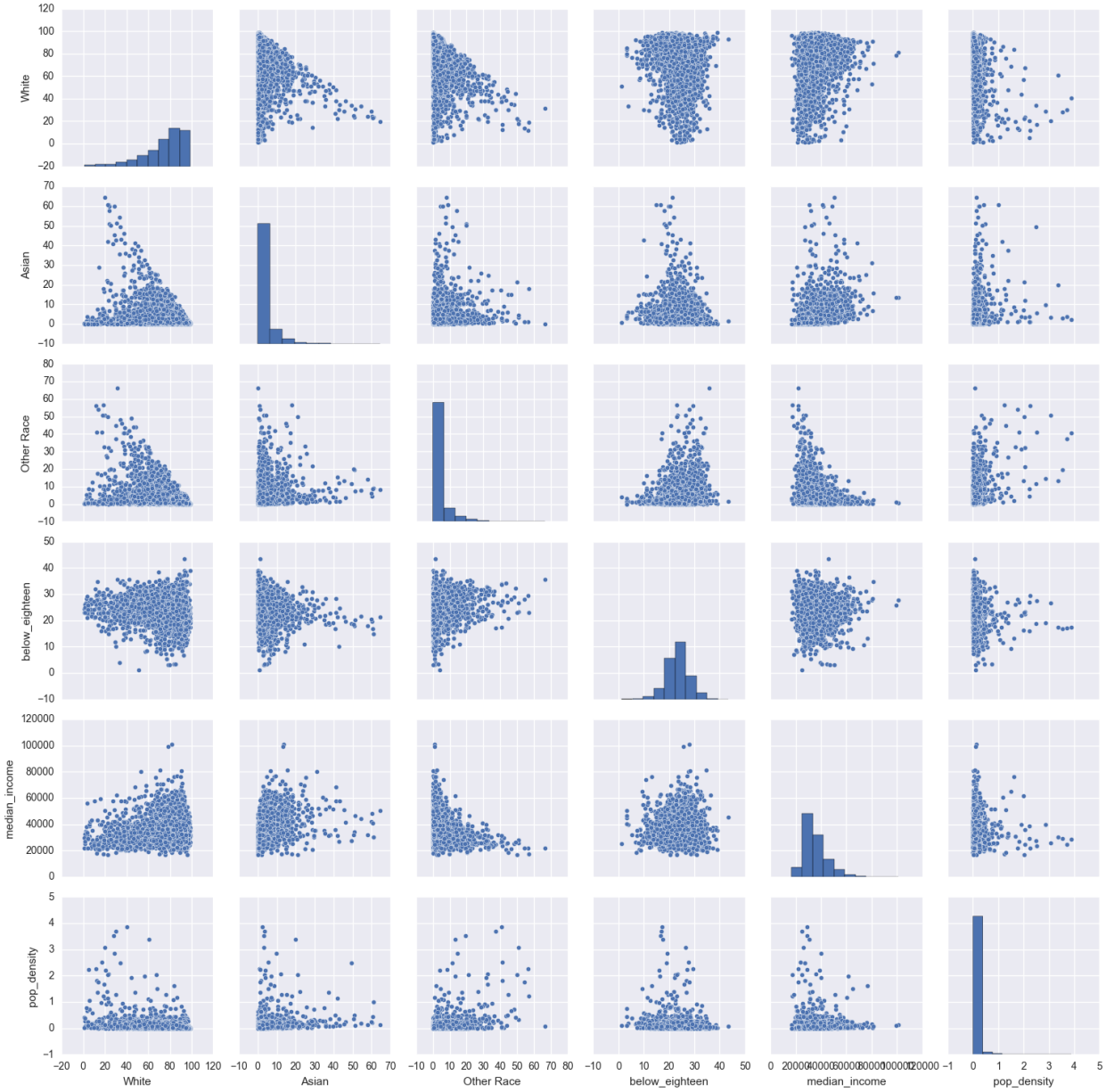
ssociated the geometric attributes of every zip code, including land area to the Yellow Pages dat a set.

Exploratory Visualization

- *Exploratory visualization on Yellow Pages data-set:* Our yellow Pages data-set includes mostly categorical columns. One interesting observation about this data-set could be the number of restaurants in different states. As we can see, Texas has the largest number of restaurants with more than 100k restaurants, and California and Florida are the next two.



- *Exploratory visualization on demographic factors:* The ACS data-set has lots of continuous variables which enables us to see some pairwise correlations between these variables. As plotted below, we can see some of these correlations for income, age, population density and different racial groups. For example, we can see that as the population of "Other Race" increases the median income decreases, whereas, "White" shows an opposite trend. We can also see that there is a positive correlation between median income and "Asian".



Algorithms and Techniques

Finding similar Clientele: The first step to this study is to find groups of restaurants with similar clientele. To do this, I first pivoted the review table to a business-user matrix with 0 and 1 values where 1 means that the user has put a review on that restaurant and 0 means otherwise. After this step, I calculated the pairwise cosine similarity matrix to be use as an input for a clustering algorithm. I used Spectral Clustering to cluster this sparse matrix as similar studies have used the same method for similar problems [4]. To find the best number of clusters, I used the Eigengap Heuristic method [5] which finds the largest difference between two consecutive eigenvalues of the Laplacian matrix and set the number of clusters equal to the rank of the eigenvalues.

Characterizing resulting clusters: After identifying the best number of clusters, we can see whic h restaurants fall into which clusters. Since we have the comments provided by different users fo r all these restaurants, we can use NLP to see if we can extract any attributes regarding food habi ts in these clusters. To this end, I will first find the most frequent words used in the entire corpus of reviews. Since some common food types are not within the frequent words, I will train a Word 2Vec model first, and use the foods that are already in the frequent words as input. This enables me to find words that are close to those foods (e.g. "Salmon" returns 'tuna', 'trout', scallop' etc.). I will then search within the returned words and generate 12 features that are: 'meat_types','vegi e_types','hardliq_types','softliq_types','sweets_types','fast_food','latin_types', 'italian_types','asian_t ypes','soda_types','seafood_types','ethnic_food'. It is expected that these generated features sum marize the practiced food habits in every cluster. After this step, for every zip code with more tha n 10 restaurants, we the proportion of different clusters and these will be our response variable.

Finding associations between variables: Once we prepared the response and independent vari ables, we will be ready to run a multivariate regression model. For this study, I will use amultitud e of models and choose the best one using the Grid Search method. I will use Lasso regression, D ecision Tree regression, Random Forest Regression and OLS regression. I will then choose the be st model which gives the highest R-square.

Alternative method to find associations between variables: depending on the optimum num ber of clusters that we find during using the heuristic eigengap method, a multivariate regressio n may not be the best answer since as we increase the number of response variables, R-square d ecreased significantly. To remedy this, one option is to cluster the resulting response variable ma trix first and then use a classifier instead by inserting the resulting labels from clustering as the ta rget variable.

Benchmark

Many studies have previously shown that the abovementioned factors are important factors for store allocation. Although, often times, these studies have not been conducted on a nationwide data-set with all types of restaurants.

Important Factor	Reference	Our measures
Income and ethnic diversity	[6],[7],[8]	Median Income in the Past 12 Months, Race
Family structure	[9]	Household and Families , Marital status
Education	[10]	Educational Attainment
Age	[10]	Age and Sex
Land use and zoning	[9]	Number of Churches, Colleges, Government offices, Hospitals, Libraries, Museums, Recreational Areas, Schools, and Shopping Centers.

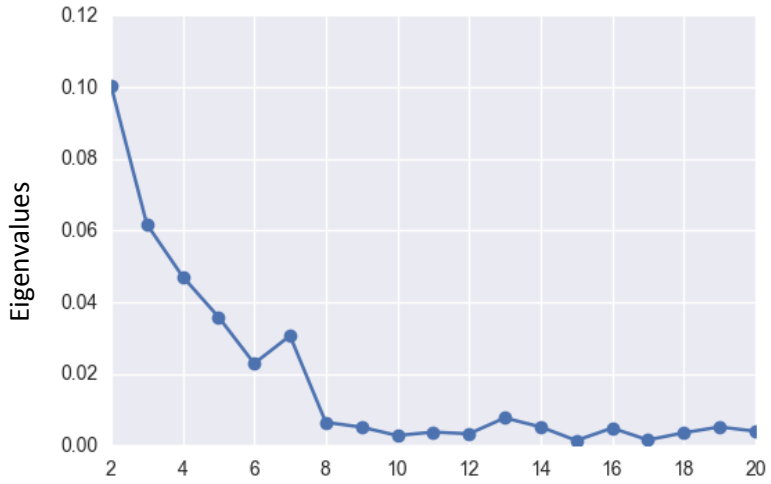
I intend to see if I will get the same results discussed in the previous studies. Is the allocation of restaurants in different zip codes influenced by these factors? Since I am using regression models, I will be able to investigate which factors are associated with the composition of different restaurants in U.S. zip codes.

III. Methodology

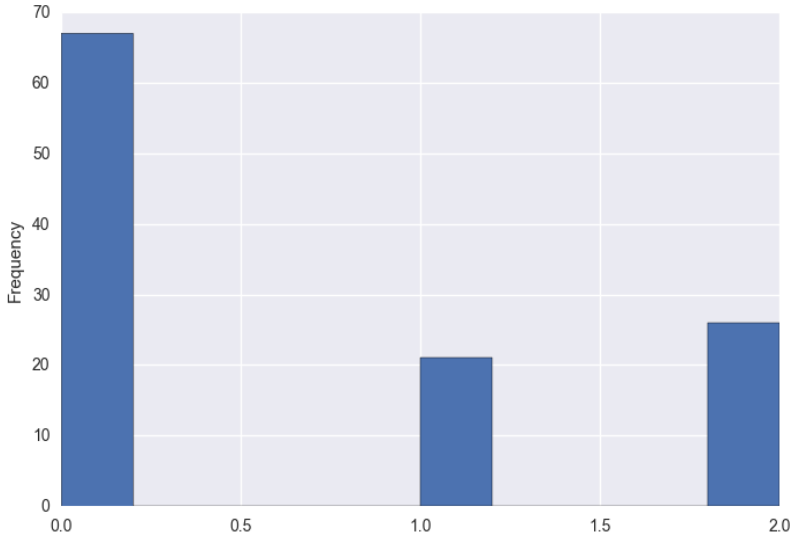
Data Preprocessing

Finding similar Clientele:

This step includes analyzing the Yelp review table in order to first create a 0,1 matrix indicating which clients chose which restaurants, generating a cosine pairwise similarity matrix, and clustering the resulting matrix. As explained earlier, after removing restaurants with less than 5 visitors, we ended up with 1872 restaurants and 44051 reviews and 27285 unique users. Accordingly, our final matrix would be a 1872 by 27285 matrix. I next calculated pairwise cosine similarity matrix from the check-in matrix where the similarities between businesses were measured. The resulting matrix was a 1872 by 1872 matrix, showing the similarities between businesses in terms of their clientele. My next step was to use the Eigengap Heuristic method to find the best number of clusters. Figure below shows the results from this step:



From the figure above, we can see that actually the best number of clusters is two. The figure below shows the changes in the eigenvalues in different number of clusters. For example, we can see that the eigenvalues decrease from 0.10 to 0.6 by choosing three clusters over two. This is the largest reduction in the eigenvalue (i.e. y axis) which is desirable. Accordingly, according to this graph, three is the best number of clusters and therefore, I divided restaurants into three clusters. Histogram below shows the proportion of restaurants falling in each cluster:



As we can see more than 60% of restaurants fall in cluster 1, about 27% fall in cluster 2 and the rest fall in cluster 1. Now we will need to see what sorts of foods are consumed in these clusters and how can we characterize them.

Characterizing the clusters:

In order to characterize each cluster, I used a bag-of-words model since only the frequency of words matter to us. This model makes sense because we assume that when a user talks about a certain food or drink he/she has purchased it or at least considered it. In this case we only care about the types of foods and drinks that the user considers. However, this raises up an important question: which foods should we consider?

To find relevant foods and drinks, first, I used English stop-words to remove commonly-used words [11] and then, chose features among the top 1000 frequent words. Fourty-five features of the three categories (i.e. foods and drinks, and food adjectives) were selected at this step.

Although frequent features can provide much information for restaurants, I expect to get more specific words from the comments. For example, different types of fish (e.g. haddock, tilapia) or different adjectives used to describe an ambience (e.g. divey, hipster) are not among frequent words. To address this problem, I used theWord2Vec model. This open-source model was developed by Google in 2013 which transforms words in a document to high-dimensional spatial vectors by using a Neural Network Language Model (NNLM) [12], [13]. I trained our Yelp corpus with this model and every word was turned into a 100-dimensional vector. As an example, table below shows the closest words to the word “classy”. It is noteworthy that the model does not necessarily return synonyms of classy but rather, it considers the way word classy is used in a sentence and therefore, it returns all adjectives that are used to describe an ambience. The 45 words chosen in the last step were given as input to this model to find the top 20 most relevant words. At the end of this step, a total of 189 features were selected.

Word2Vec Output	Similarity to classy
Swank	0.87688
Trendy	0.86152
Chic	0.85917
Posh	0.84972
elegant	0.84592

189 features is too many for us to learn the overall practiced taste in a restaurant. Therefore, I generated different features each representing a category of food, drink or adjective. Below are the least of food categories that I generated: 'meat_types', 'vegie_types', 'hardliq_types', 'softliq_types', 'sweets_types', 'fast_food', 'latin_types', 'italian_types', 'asian_types', 'soda_types', 'seafood_types', 'ethnic_food'. To find out which features each of these categories are comprised of, please refer to the python code. After this step, I was able to actually compare different clusters.

	meat_types	vegie_types	hardliq_types	softliq_types	sweets_types	fast_food	latin_types	italian_types	asian_types	soda_types
Cluster										
0	1.000000	1.000000	0.000000	0.336019	0.41576	0.359059	1.00000	0.991885	1.000000	0.000000
1	0.898167	0.000000	0.059089	0.000000	1.00000	1.000000	0.00000	1.000000	0.000000	0.944017
2	0.000000	0.580992	1.000000	1.000000	0.00000	0.000000	0.13815	0.000000	0.209864	1.000000

In the table above, I have normalized these features with a max-min scaler to exaggerate the differences. Figures below show the differences between different clusters:

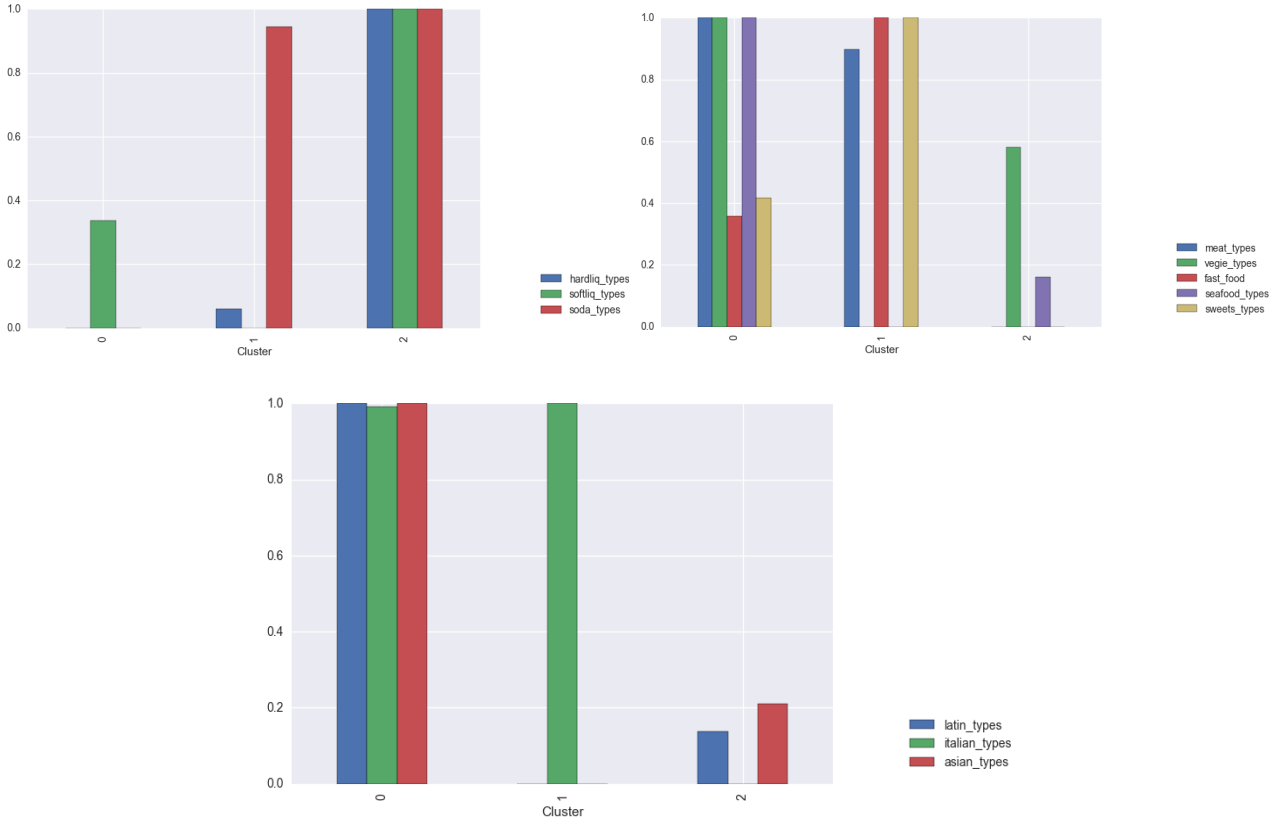


Table below shows examples of restaurants that fall in every cluster:

	Names_cluster_0	Count_0	Names_cluster_1	Count_1	Names_cluster_2	Count_2
0	Subway	364	Pizza Hut	155	McDonald's	345
1	Taco Bell	180	Wendy's	144	Burger King	147
2	Panda Express	119	Jimmy John's	95	Domino's Pizza	97
3	Chipotle Mexican Grill	95	KFC	93	Papa John's Pizza	84
4	Jack in the Box	79	Roberto's Taco Shop	44	Arby's	63
5	Denny's	77	Panera Bread	44	Chick-fil-A	55
6	Sonic Drive-In	67	PT's	33	Jersey Mike's Subs	47
7	Del Taco	60	Olive Garden Italian Restaurant	30	Quiznos	32
8	Applebee's	50	Culver's	25	Qdoba Mexican Grill	30
9	Port of Subs	46	Fatburger	22	Chili's Grill & Bar	29

Cluster 0 looks to include restaurants with mostly soft liquor, a balance between all types of food and ethnic groups. Cluster 1 is the highest on soda types, fast food, sweet types, meat types, and Italian types and smallest in every other features. Cluster 2, is highest on all sorts of drinks, and moderate on vegies and sea food. Table above shows the clusters of restaurants with similar clientele. Generally, looks like cluster 2 includes many burgers and grills. Cluster 0 includes many Mexican food types and non-burger fast foods and cluster 1 covers the rest. Intuitively, however, many of these restaurants do not seem to be similar in terms of the types of foods offered. For example, cluster 2 and 1 both have lots of pizza and chicken places and 0 and 1 both have Mexican foods.

Finding associations between response and independent variables:

In the next step I filtered out zip codes with less than 10 restaurants (i.e. Yelp restaurants). After this step, 4622 zip codes remain. For every zip code I calculated the proportion of restaurants in each cluster such as below:

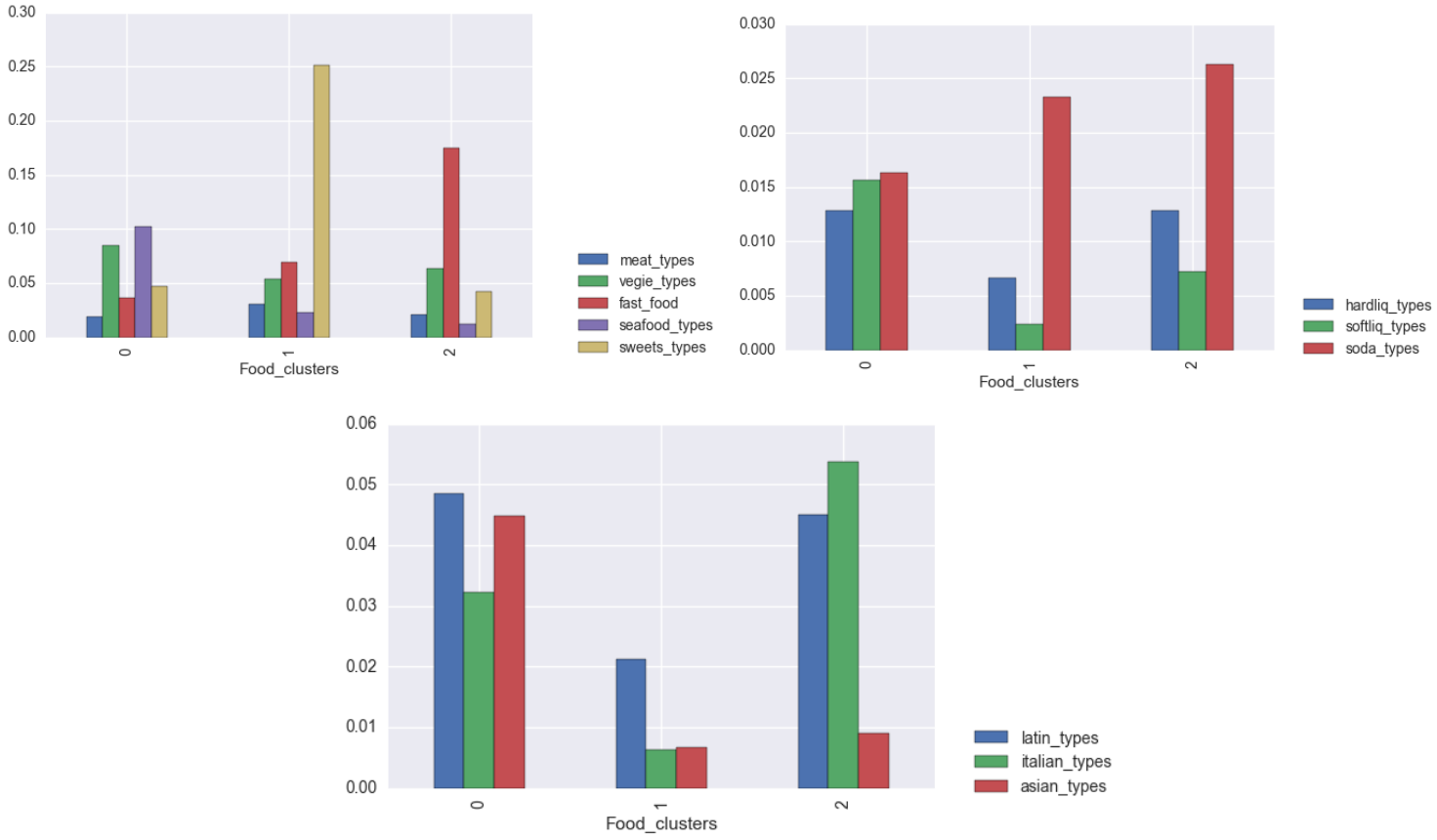
	ZIP	Cluster0	Cluster1	Cluster2
0	01020	0.857143	0.047619	0.095238
1	01035	0.833333	0.083333	0.083333
2	01040	0.869565	0.043478	0.086957
3	01060	0.900000	0.100000	0.000000
4	01085	0.875000	0.000000	0.125000

This matrix will constitute the response variable table. With our independent variable matrix, we can now proceed to split the dataset to train, test and validation sets. To this end, I first separate 20% for the test set and then 10 –fold CV for the remaining data points. Four different regressors were chosen: OLS, Lasso, Decision Tree, and Random Forest. All these regressors accept more than one response variables. For every regressor, I created a dictionary of multiple hyper parameters. The algorithm, chooses one regressor and performs a grid search on the corresponding hyper parameters for each regressor, with a 10-fold cross validation. The performance function, in this case, is R-square. For every regressor, the algorithm chooses those sets of hyper parameters that maximize the r-square. The highest r-square that was achieved was 0.167 for Random Forest algorithm. This number is pretty low. One possible reason is that we have multiple response variables and this increases the chance for higher MSE.

A possible solution to change the question from regression to classification. That is, we perform a clustering on the response variables matrix first, so that we can convert it to a single column of classes. To find the best number of clusters, I calculated the silhouette score for multiple number of clusters. The highest silhouette score achieved when the number of clusters were two.

We can now perform a classification task, in much the same way that we did the regression. We choose four different classifiers first: Random Forest, Adaboost, Logistic Regression, and Decision tree classifier. We then create a dictionary of corresponding hyper parameters. We save 20 percent of the data for testing and perform grid search and 10 fold cross validation on the rest. This time, I added a loop in the grid search process where we examine different number of features as independent variables using the k-best method. The reason for choosing k-best over other dimension reduction methods is that we will need to avoid any transformation on the independent variables to be able to interpret the importance of different factors. For every classifier, therefore, we will end up with best number of features and best hyper parameters. The best classifier was Random Forest with f1 score of 0.576 and 17 features. The accuracy of this model was 0.49. Given the fact that we only have two labels, a random guess would have been 50% accurate. This means that our classifier does not perform well.

At this point, I consider another scenario to see if we can improve the results. In this scenario, instead of finding clusters of similar clientele, we will focus on the types of foods that people in their reviews. To this end, I will use the same features that I used earlier to describe the characteristics of every client cluster. This time, I will weight these features using TF-IDF, which considers the frequency of terms in comparison to the total number of restaurants and weights them accordingly. I will then cluster different restaurants based on the types of foods discussed in each. We can see the differences between these clusters for certain food categories in figures below:



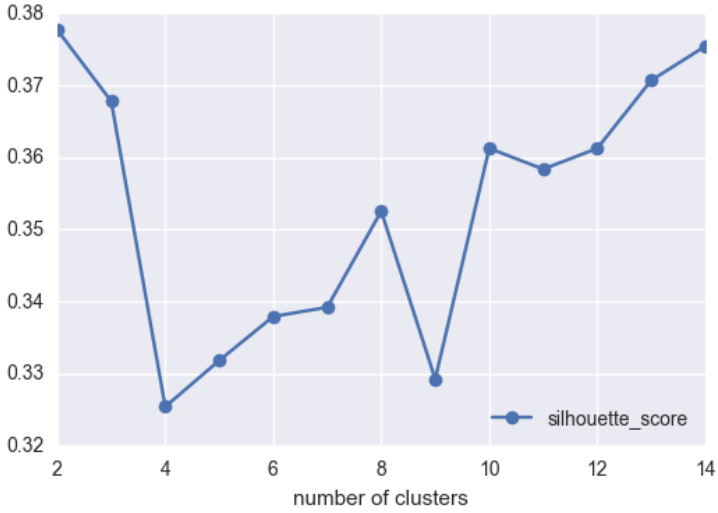
The restaurants in each cluster, in this case, are as follows:

	Names_cluster_0	Count_0	Names_cluster_1	Count_1	Names_cluster_2	Count_2
0	Subway	364	Dairy Queen	45	Del Taco	60
1	McDonald's	345	IHOP	40	Roberto's Taco Shop	44
2	Taco Bell	180	Carl's Jr.	34	Panera Bread	44
3	Pizza Hut	155	Church's Chicken	28	Capriotti's Sandwich Shop	41
4	Burger King	147	Popeyes	25	Carl's Jr	41
5	Wendy's	144	Culver's	25	Barro's Pizza	40
6	Panda Express	119	Waffle House	24	Little Caesars Pizza	37
7	Domino's Pizza	97	Village Inn	19	Rosati's Pizza	36
8	Chipotle Mexican Grill	95	Whataburger	19	PT's	33
9	Jimmy John's	95	Boston Market	18	Quiznos	32

Not surprisingly, the differences between food categories in the figures are more obvious, although they have not been scaled like the last time, using max-min scaler. Cluster 0 represents sea food and vegies, cluster 1 excels the others in sweet types and cluster 2 are mostly fast food. Again, some exceptions can be seen for example, McDonalds is in cluster 0 while it would have made more sense if it was in cluster 2.

The next steps are exactly the same as described earlier. This time for the regression processes, I found that Random forest was the best with R-square of 5%. Again the R-square was considerably low, so I

proceeded to reframe the question as a classification problem. The number of clusters were determined as 2 because 2 returned the highest silhouette score:



number of clusters	silhouette_score
2	0.377708
3	0.367789
4	0.325393
5	0.331813
6	0.337918
7	0.339198
8	0.352537
9	0.329230
10	0.361258
11	0.358377
12	0.361247
13	0.370693
14	0.375425,

After repeating those steps once again, we find that Random Forest, again, returns the best F1 score. This time, the F1 score, gets to 0.73, however, the higher f1-score is due to high recall (i.e. recall = 1.0). Our precision in this case is 0.58 and accuracy is slightly better than random (i.e. 0.58). Although the scores are not ideal, this is the highest that we have reached so far.

IV. Results

Model Evaluation and Validation

Now we can perform a sensitivity test to see if the results in the final model are consistent under different samplings. We calculate the f1 scores with different test sizes, once stratified on the response variable (i.e. classification labels) and once without stratification. We calculate the f1 score for training set using 10-fold cross validation and for test set. The f1 results show that the results were not affected by sampling bias.

	stratification = Falsetest size: 0.1	stratification = Truetest size: 0.1	stratification = Falsetest size: 0.15	stratification = Truetest size: 0.15	stratification = Falsetest size: 0.2	stratification = Truetest size: 0.2	stratification = Falsetest size: 0.25	stratification = Truetest size: 0.25	stratification = Falsetest size: 0.3	stratification = Truetest size: 0.3	stratification = Falsetest size: 0.35	stratification = Truetest size: 0.35	stratification = Falsetest size: 0.4	stratification = Truetest size: 0.4	stratification = Falsetest size: 0.45	stratification = Truetest size: 0.45	sets
0	0.731141	0.732558	0.731006	0.733607	0.726477	0.732026	0.725995	0.733179	0.725000	0.733167	0.725806	0.732620	0.725146	0.732558	0.722045	0.734177	train1
1	0.731518	0.732558	0.731006	0.733607	0.726477	0.732026	0.725995	0.733179	0.726817	0.733167	0.725806	0.732620	0.725146	0.732558	0.722045	0.734177	train2
2	0.731518	0.732558	0.729897	0.732510	0.726477	0.733624	0.725995	0.731935	0.726817	0.733167	0.725806	0.732620	0.725146	0.732558	0.722045	0.734177	train3
3	0.731518	0.732558	0.729897	0.732510	0.726477	0.733624	0.725995	0.731935	0.726817	0.733167	0.726287	0.733154	0.725146	0.732558	0.722045	0.732484	train4
4	0.731518	0.732558	0.729897	0.732510	0.728070	0.733624	0.725995	0.731935	0.726817	0.733167	0.726287	0.733154	0.725146	0.732558	0.722045	0.732484	train5
5	0.731518	0.733981	0.729897	0.732510	0.728070	0.733624	0.725995	0.731935	0.726817	0.733167	0.726287	0.733154	0.725146	0.732558	0.722045	0.732484	train6
6	0.731518	0.733981	0.729897	0.732510	0.726872	0.733624	0.725995	0.733645	0.725441	0.733167	0.726287	0.733154	0.723529	0.732558	0.724359	0.732484	train7
7	0.731518	0.732943	0.729897	0.732510	0.726872	0.733624	0.725995	0.733645	0.725441	0.733167	0.726287	0.733154	0.723529	0.734694	0.718447	0.732484	train8
8	0.731518	0.732943	0.729897	0.732510	0.726872	0.732456	0.724706	0.733645	0.725441	0.731830	0.726287	0.733154	0.723529	0.733138	0.722581	0.732484	train9
9	0.731518	0.732943	0.731405	0.734021	0.726872	0.732456	0.724706	0.733645	0.725441	0.733668	0.726287	0.733154	0.725664	0.733138	0.722581	0.732484	train10
10	0.746116	0.727861	0.747987	0.732618	0.757081	0.728870	0.752407	0.732109	0.748706	0.726231	0.745428	0.734001	0.745132	0.732283	0.745580	0.732919	test

We ran the classification once more, this time the performance function is set on accuracy-score, and not f1-score. The highest accuracy was achieved by a random forest algorithm at 0.58, which was not significantly different from the once that we reached at using f1-score. At this point, we can trust the final model as our best model.

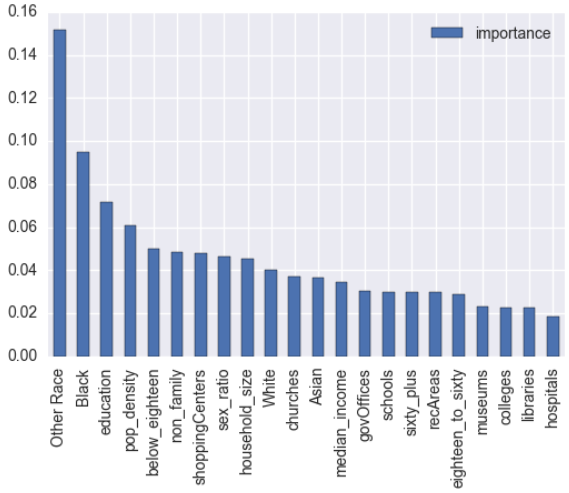
Justification

At this point we can see the importance of different factors. Since in all case, we found the Random Forest model, either as regressor or classifier, to be the best model, it is easier to draw a comparison between different regressors and classifiers, in terms of factor importance. Table below summarizes our results.

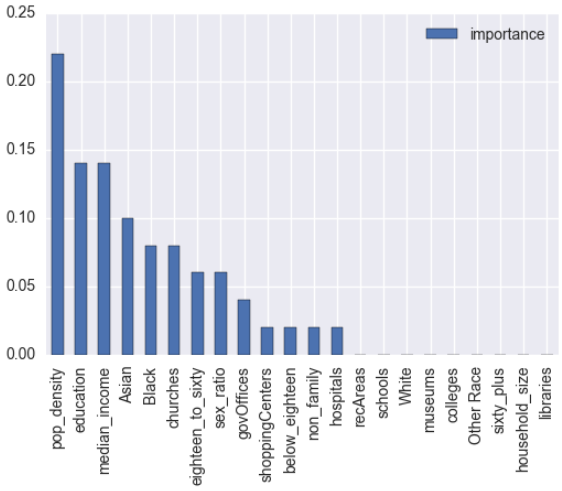
Model	**Cluster type used**	**Number of clusters**	**Score**	**Top 10 important factors**
Random Forest Regressor	Check-ins	3	R-square: 0.167	Other Race, Black, Pop-density, Non-family, Below-eighteen, Schools, White, Education, Asian
Random Forest Classifier	Check-ins	2	F1-score:0.576 Precision: 0.563 Recall: 0.553 Accuracy: 0.49	Other Race, Black, Pop-density, Asian, Non-family, White, Sex-ratio, Below-eighteen, Churches, Education
Random Forest Regressor	Food names	3	R-square: 0.05	Pop-density, Black, Household-size, Shopping centers, Asians, Schools, Other Race, Eighteen-to-sixty, White, Non-family, Sixty-plus
Random Forest Classifier	Food names	2	F1-score: 0.732 Precision: 0.577 Recall: 1.0 Accuracy: 0.58	Pop-density, Education, Median income, Asian, Black, Churches, Eighteen-to-sixty, Sex-ratio, GovOffices, Shopping centers
Random Forest Regressor	Check-ins	2	R-square: 0.197	Other Race, Black, Education, Pop-density, Below-eighteen, Non-family, ShoppingCenters, Sex-ratio, Household Size, White

From the table above, we can see that among regressors, the check-in matrix with two clusters as response variables performs the best. Among classifiers, food names matrix with two clusters perform the best. In all classifiers and regressors, "Other Race", "Black", "Education", and "Pop-

density" seem to be among the most important factors. Age and Sex are both important in the best classifier and regressor.



Factor importance for the best regressor



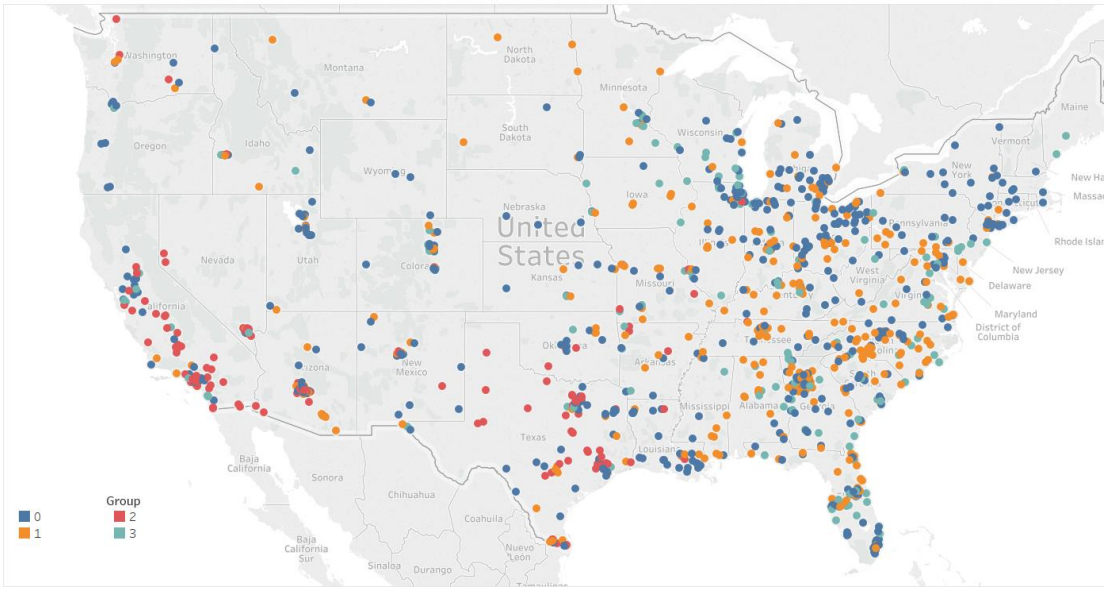
Factor importance for the best classifier

Recall that in our bench mark we discussed that the previous research has found the following factors to be important: Income, Race, Family structure, Education Age, Sex, and land use. Our models show that all these factors are important although, given the generally low scores, there seems to be other important factors that affect restaurant choice, which begs further investigation.

To study the geographical aspect, I defined another function (i.e. reg_state()) to limit the geographical scope of the analysis to certain part of the data. This function assists us to only focus on one state at a time. For example, by using this function for Texas, we will see that our R-square increases slightly to 0.20 although this increase is not as substantial.

V. Conclusion

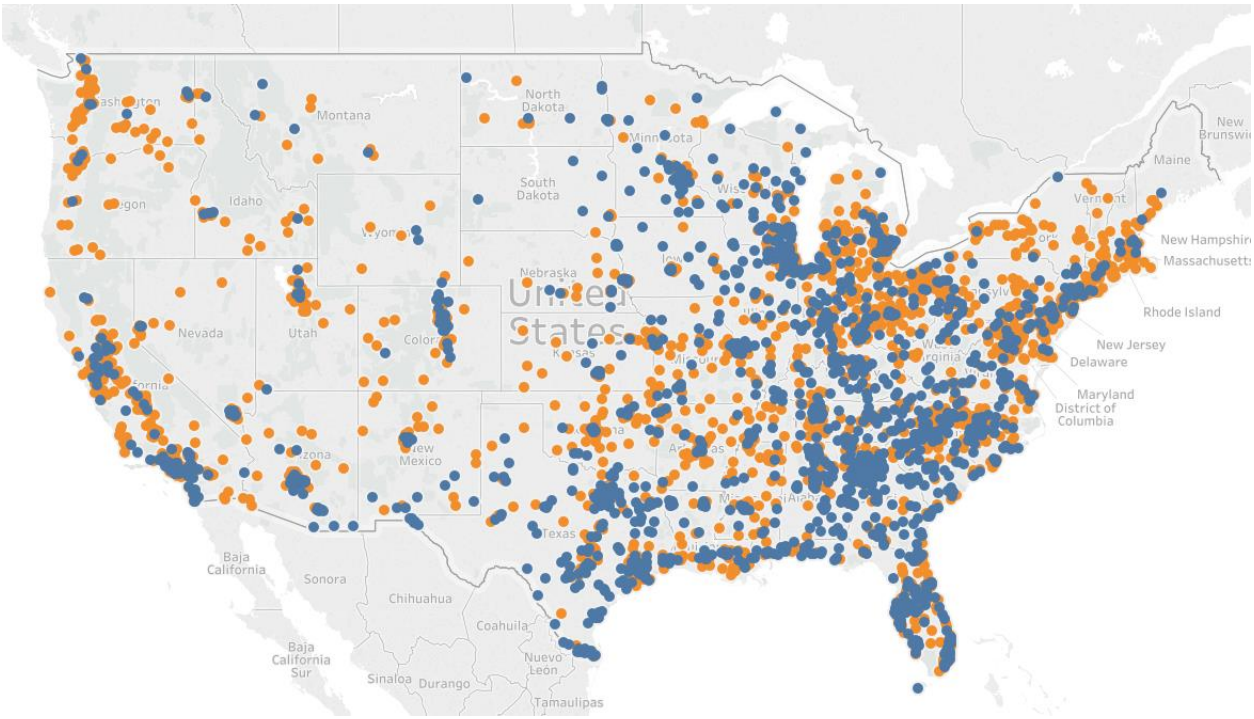
Free-Form Visualization



Clusters based on similar clientele (4 clusters)

One possibility for getting low scores could be the importance of geographic factors. Since we have focused on restaurants names, simply, some chain restaurants may not exist in all areas, but this does not mean that restaurant types are not actually representative of taste. In order to see this, I created a number of maps. The figure above shows the distribution of different zip codes once we cluster them based on clientele. We can see that the group 2 denoted by color red, are all in south and west and almost no zip code of that kind exists in the east coast. Group 1 denoted by orange are mostly in the east and south, group 0, are mostly concentrated in north and east. Therefore, there seems to be some geographical factors affecting our models.

Figure below indicates the clustering results based on consumed food. This case seems to be less geographically affected. We can clearly see the impact of population density in this one. As we can see the blue points are mostly located in metropolitan areas.



Clusters based on consumed foods (2 clusters)

Reflection

In this project, I asked a simple question: is the taste of people in different zip codes in the U.S. a function of the area's demography? I characterized the practiced taste in different regions using the Yelp data-set by having two different hypotheses in mind: First, restaurants with similar clientele are reflective of similar tastes. Therefore, by clustering restaurants based on their clientele we can investigate the restaurant composition of different zip codes. Second, the practiced taste in a region is merely a function of the types of foods that are consumed in that region. Both these hypotheses were tested in multiple ways.

To test the two hypotheses, I clustered businesses twice: first, based on the types of clients each had and second, based on the types of foods consumed at each. Using Eigengap Heuristics method, I found the best number of clusters for each case. Next, I asked this question: what proportion of every zip code belongs to these clusters. I only focused on zip codes with at least 10 restaurants. This matrix became my response variable matrix which I used in two different models: first, a regression model which takes multiple response variables in, second, a classification model which takes the resulting labels from clustering the continuous response variable that we used as an input for the regression model.

For every classification and regression task, I tested a multitude of classifiers and regressors. In each operation, I chose the best model given a performance function: for regressors I used R-square and for Classifiers I used F1-score. The best model for the first hypotheses (i.e. taste defined as similar clientele) was a Random Forest Classifier with the second hypothesis (i.e. taste defined as similar types of foods consumed at restaurants) with F1-score of 0.73 and accuracy of 0.58. Among regressors, a Random Forest regressor with R-square of 0.19 was the best with the first hypothesis (i.e. taste defined as similar clientele). In both cases, models aligned with studies introduced in the benchmark section.

Improvement

This project can use improvements in many respects: first, our independent variables may not have been comprehensive enough. There seems to be other factors affecting restaurant allocation that are not necessarily demographic, for example, proximity to major roads, attraction areas etc. Capturing these variables requires one to do some extensive GIS processes. Second, we only focused on restaurant names derived from 6 cities and obviously we lost a whole lot of restaurants. Many restaurants may have different restaurants with names that are not included in our yelp data-set. This greatly affects our understanding of restaurant composition in a given zip code. Additional review data set for different cities across the U.S. could definitely reduce the bias in our response variables. Third, some part of the unwanted variance may be due to the fact that we aggregated everything on "restaurant names". A restaurant brand may not be representative of practiced taste in a region. For example, people may purchase burgers from Macdonald's in one region and buy fried chicken from the same restaurant in another. That is, these two regions may both be fans of McDonald's in quite different ways. Lastly, the foods that we looked for in the yelp reviews may not be sufficient for quantifying taste in a given restaurant. Also, many zip codes, may have a few number of reviews in their restaurants and that affects our understanding of the types of foods and drinks that are consumed in those zip codes.

References:

[1] P. Bourdieu, *Distinction: A social critique of the judgment of taste*, vol. 1, no. 3. 1984.

[2] J. He and W. W. Chu, *A Social Network-Based Recommender System (SNRS)*, vol. 12. 2010.

[3] P. Bonhard, C. Harries, J. McCarthy, and M. Sasse, "Accounting for taste: using profile similarity to improve recommender systems.," *CHI 06 Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 1057–1066, 2006.

[4] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, "Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks.," *Soc. Mob. Web*, pp. 32–35, 2011.

[5] U. von Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[6] J. Beaulac, E. Kristjansson, and S. Cummins, “A systematic review of food deserts, 1966-2007,” *Prev. Chronic Dis.*, vol. 6, no. 3, p. A105, 2009.

[7] K. Morland, S. Wing, A. Diez Roux, and C. Poole, “Neighborhood characteristics associated with the location of food stores and food service places,” *Am. J. Prev. Med.*, vol. 22, no. 1, pp. 23–29, 2002.

[8] L. M. Powell, S. Slater, D. Mirtcheva, Y. Bao, and F. J. Chaloupka, “Food store availability and neighborhood characteristics in the United States,” *Prev. Med. (Baltim).*, vol. 44, no. 3, pp. 189–195, 2007.

[9] D. Levinson, “Zoned Out: Regulation, Market, and Choice In Transportation and Metropolitan Land-Use – Jonathan Levine.,” *Growth Chang.*, vol. 37, no. 3, pp. 492–494, 2006.

[10] F. Li *et al.*, “Built environment and 1-year change in weight and waist circumference in middle-aged and older adults: Portland neighborhood environment and health study,” *Am. J. Epidemiol.*, vol. 169, no. 4, pp. 401–408, 2009.

[11] S. Bird and E. Loper, “NLTK: The Natural Language Toolkit,” in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, pp. 1–4.

[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv Prepr. arXiv1301.3781*, 2013.

[13] X. Rong, “word2vec Parameter Learning Explained,” *arXiv:1411.2738*, pp. 1–19, 2014.