Minority Action Project Report

Divya Veerapaneni
Steve Kim
Arundeep Singh

Introduction

The Minority Action Project is a research project designed to identify the resources available for minority-led startups in innovative cities in the United States. For this project, we will refer specifically to women and African-Americans as minorities. The objective is to determine what factors contribute to the success of minority-led tech companies in innovative cities. The factors taken into account for this study include gender, income, and ethnicity of entrepreneur. The specific objectives of this project are the following:

1. Determine the distribution of businesses and resources in Boston to gain insight into the distribution of businesses and resources in a typical innovative American city
2. Measure and graph demographics such as ethnicity, income, gender, and education level of entrepreneurs throughout the United States.
3. Measure the change in entrepreneurial growth throughout the years in the U.S. to see whether it matches minority entrepreneurial growth/decline.
4. After creating a metric for entrepreneurial success, measure the success rate throughout the years in the U.S. Identify what factors (especially ethnicity and gender) contribute to success. Build a model to predict success based on demographic information for each entrepreneur.

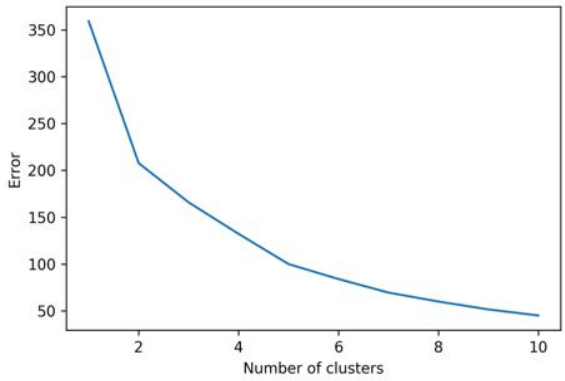
To obtain statistics regarding the demographics of American entrepreneurs and variables that contribute to success of entrepreneurs, we used microdata from the Kauffman Index of Entrepreneurship in America (KIEA), which surveyed 650,000 Americans each year from 1996 to 2015 for entrepreneurial data. To obtain statistics related to businesses and resources in Boston, we used a property assessment document from data.boston.gov.

Methods

The initial intention of this project was to correlate funding with businesses and entrepreneurial success, with particular emphasis on minority entrepreneurship. However, we were unable to find data relevant to this goal. We especially had trouble finding demographic microdata on entrepreneurships that we could connect directly to funding microdata. Furthermore, we originally intended on focusing specifically on Boston, but the sources we found lacked an adequate amount of data on that, so we expanded our scope to the entire country.

Objective 1

To create an accurate representation of the distribution of businesses in Boston, we utilized Boston's census data on property, which included type and value of property. The data had lists of addresses, which we converted into longitude and latitude data using the python package *geopy* to access Google's API. This was somewhat problematic, as the maximum requests we could send per day was around 1250. But we eventually managed to obtain the longitude and latitude data of the dataset (approximately 6000 locations). We generated density maps based on either location and property type or location and property value, making sure to filter out data on governmental and residential locations. We then clustered based on these parameters using *k*-means as our algorithm. Based on the error plot for the location and property value scatter plot shown below, the ideal number of clusters was 5 for that given density map. Upon creating the respective scatter plots, we overlaid a map of Boston provided by Google Maps onto our original plots. We also created a scatter plot displaying the density map for location and property values in terms of ranges of property values.



Objective 2

To determine the demographics of entrepreneurs, we chose to analyze the 2015 KIEA dataset since that is the most recent dataset. Using *pandas* and *matplotlib*, we generated graphs showing the breakdown of entrepreneurs with regard to education level, gender, family income, and race.

Objective 3

In order to take objective 2 further, we used the KIEA dataset to track entrepreneurial activity throughout 1996-2015 on a national and minority level. To do so, we plotted the percentage of U.S. adults who are entrepreneurs over the years. Then, we did the same for minority entrepreneurs. This was done to gain insight into how minorities are represented in entrepreneurship on a national level.

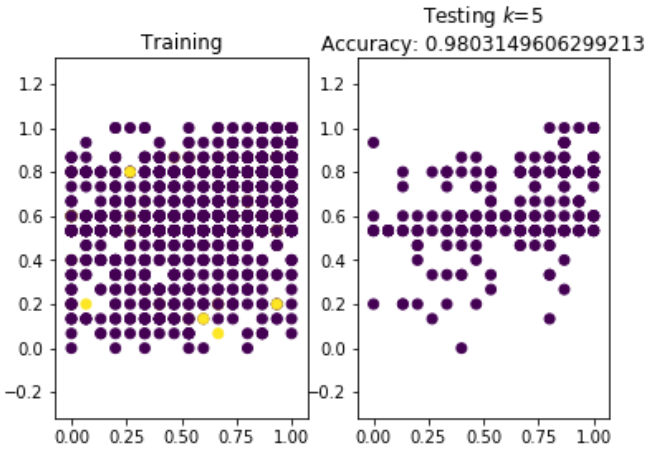
Objective 4

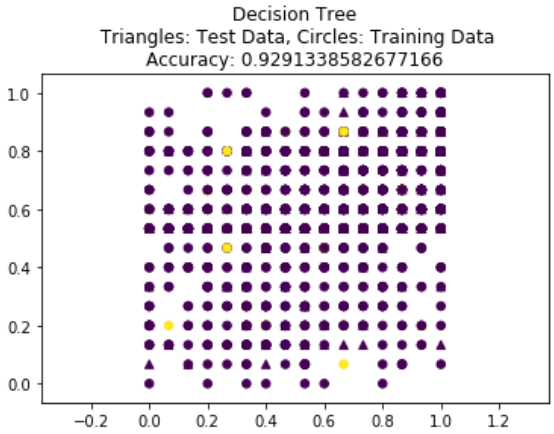
Based on the KIEA dataset, we determined that an appropriate definition of a successful entrepreneur would be the turnover rate of entrepreneurs to the following

year. Originally, we planned to define success as individuals whose businesses survived for 5 years, but we found that the sample size given by the KIEA was too small to accurately portray this. Future efforts to conduct this study should use more comprehensive data from the U.S. Bureau of Labor Statistics, just as other similar studies have done so. First, we identified the entrepreneurs in one year. Then, we created a field for each entrepreneur to see if they still were considered entrepreneurs in the next year's dataset. We plotted success rates for each year from 1996-2014 to see if the data correlated with data obtained in objective 3.

Since the majority of our data consisted of categorical values, we used dummies to create new feature dimensions with binary values for each possible category. The only values that we treated as numerical values were education level and family income (which were given as hierarchical categories). We normalized those values at a range of [0,1] to match our binary values for our categorical data. We then created a linear regression model using *statsmodels.api* for each field and success to determine the most contributing factor to success based on our definition of success. The most contributing factors were determined by the models with the highest coefficient of determinations (R^2).

Finally, we used *k*-nearest neighbors and decision trees from *sklearn* to build models that predict the target feature dimension (if the entrepreneur was successful) given the entrepreneur's demographics. We determined our *k* value by plotting accuracy vs *k* value, choosing the *k* value with the highest accuracy on the test data. The following images show training data and testing data for both *k*-nearest neighbors and decision trees:

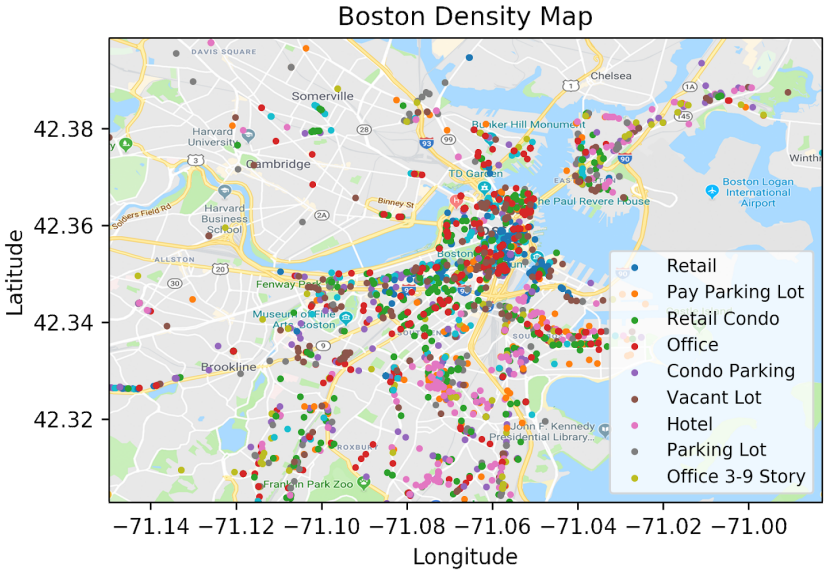




Analysis and Results

Objective 1

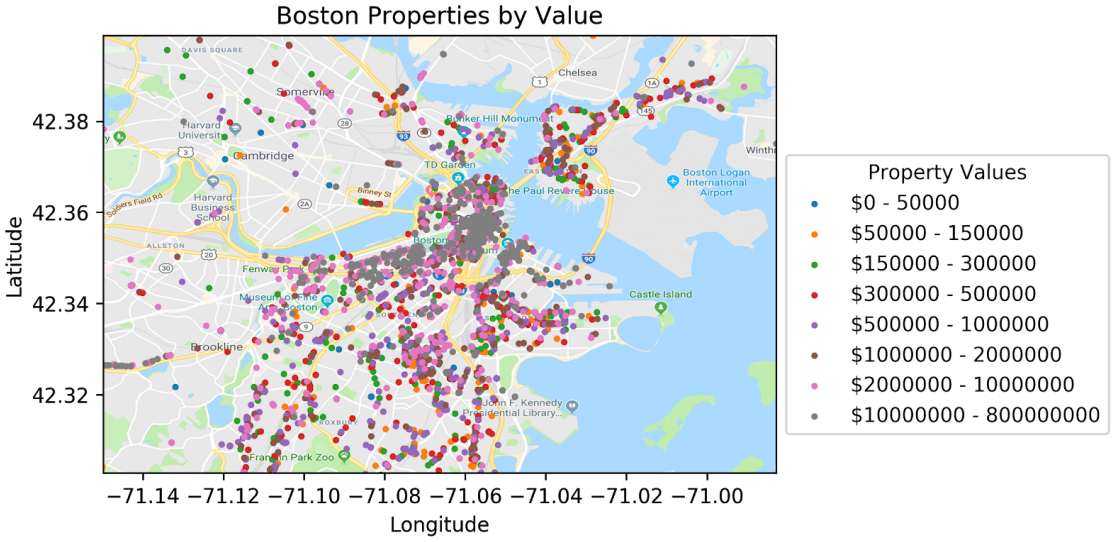
Upon clustering by property types, we found that the clusters were quite mixed and the plot was densely packed. However, certain clusters were filled with retail and office locations while the other clusters were filled with parking lot properties. This led to little insight in regards to our project goals, so we decided to leave out the cluster map for property types and location. We did, however, included a scatter plot of the most common property types in Boston to give an overview of how businesses and properties are distributed locally.

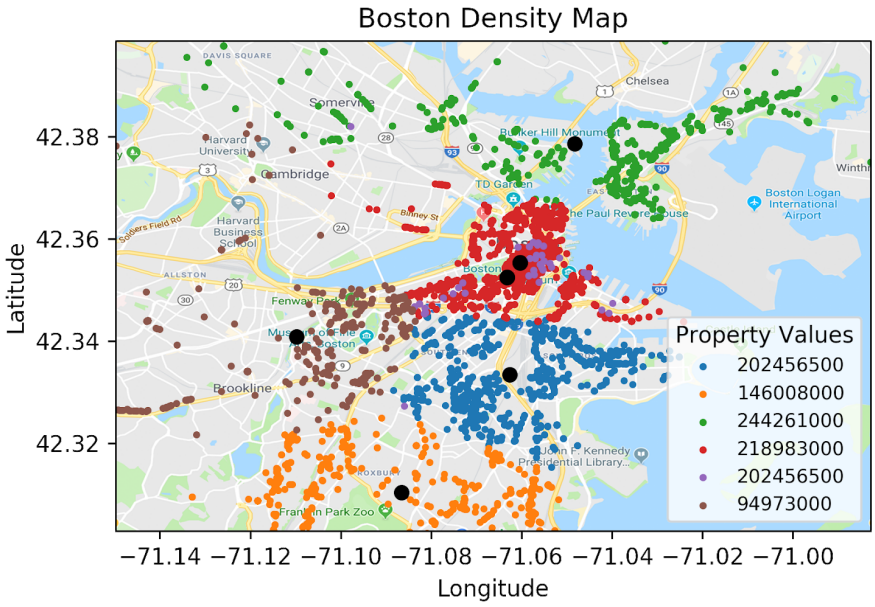


Clustering by property value and location led to more meaningful results with respect to our project goal. The ideal number of clusters for property values and locations, based on our error plot, was 5. However, since our focus is on minority areas as described by our project goal, we expanded our clusters to 6. Doing so pushed the cluster locations further out and gave us information on the Roxbury/Dorchester areas, which have higher African American and Hispanic populations. We found that the centroid locations fell on Park St. Station, West End near Bowdoin, East Boston near Central Square, south side Brookline, West End near Government Center, and Boylston near Copley/Prudential. Since the property value clusters were very homogeneous in terms of value, we suspect that property values in Boston tend to fluctuate proportionally with location.

The regions with the lowest average property value were Brookline/Longwood ($95,000) followed by Roxbury/Dorchester ($145,000). Both Roxbury and Dorchester have a high percentage of African Americans (about 50%) and Hispanics (20%) according to US Census Bureau. The wealthiest regions, in terms of average property

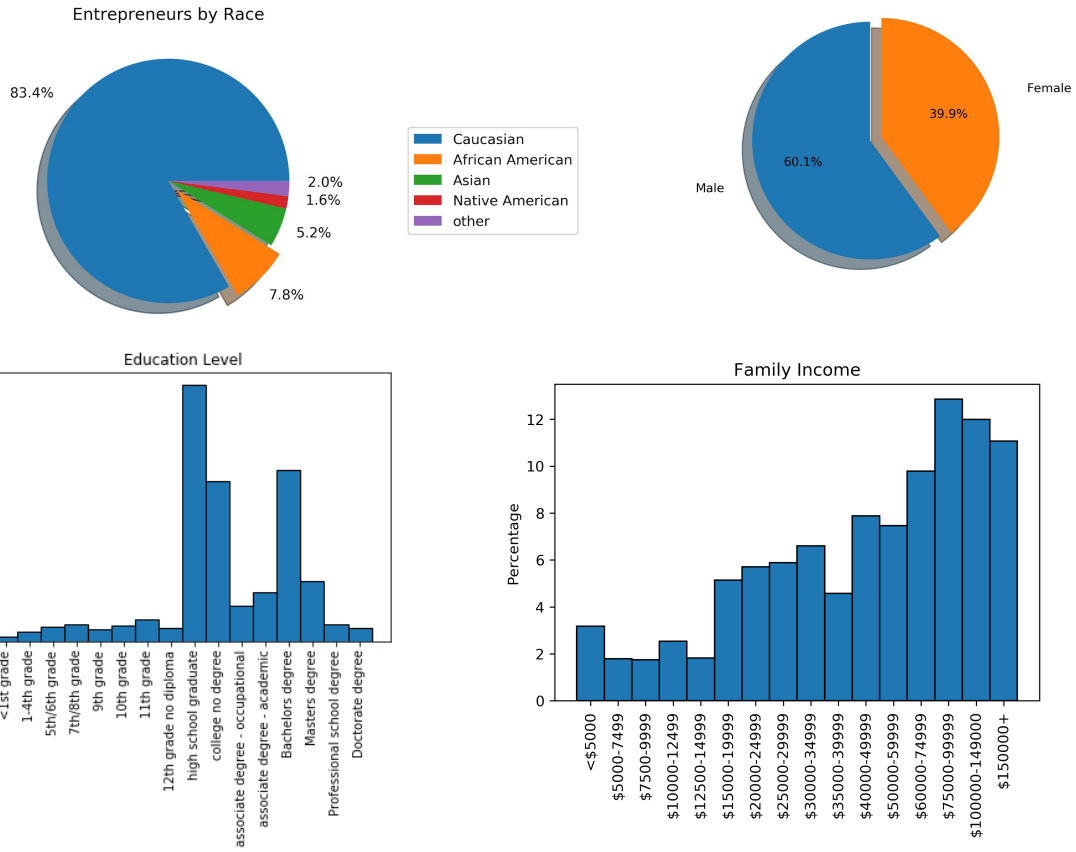
value, were East Boston, Boston Commons, and Chinatown. The scatter maps below report our findings. The first map shows points based on property value ranges, and the second shows points based on clustering.





Objective 2

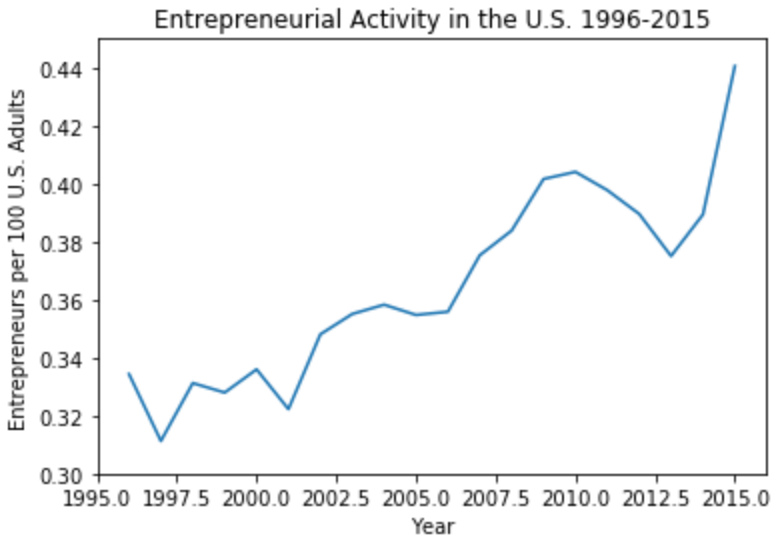
After breaking down the entrepreneurs by each category, we found the following distributions:



Using the KIEA 2015 dataset, we discovered that the majority of entrepreneurs were of the following category: college dropouts or graduates, family incomes greater than $75,000, male, and white. This is expected as most Americans are white and male. In addition, it makes sense that having high capital (family income greater than $75,000) and being highly educated (at least some college education) contribute to success as entrepreneurs.

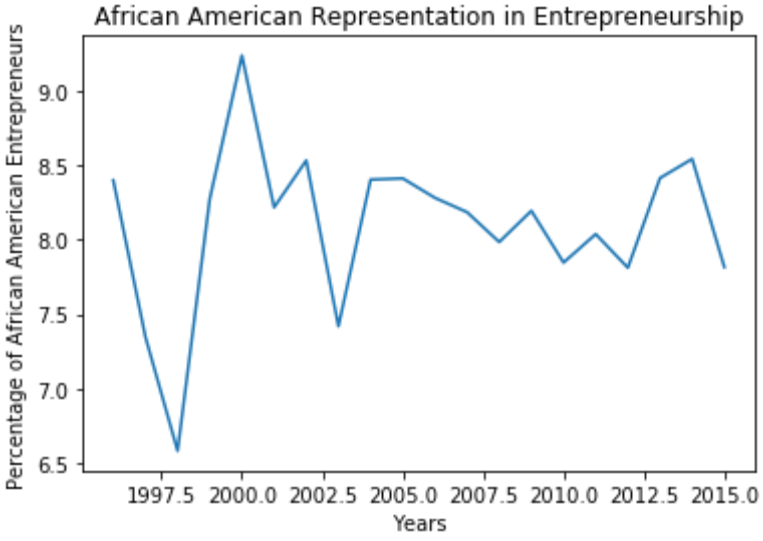
Objective 3

The KIEA datasets between the years 1996 and 2015 have shown an overall increase in entrepreneurial activity in the United States. The plot below shows the change in the amount of entrepreneurs per capita in the United States for the aforementioned years.



As shown, the percentage of U.S. adults who are entrepreneurs has positively evolved over the years. A net change of 0.11% is observed between 1996 and 2015.

Furthermore, using the demographics KIEA data, we were able to chart the change in minority representation in entrepreneurship within the United States. The point in doing this is to see if the trend for a particular minority group's representation matches that of the entire country. Below are graphs that track the minority groups we are considering for this project: African Americans and women.

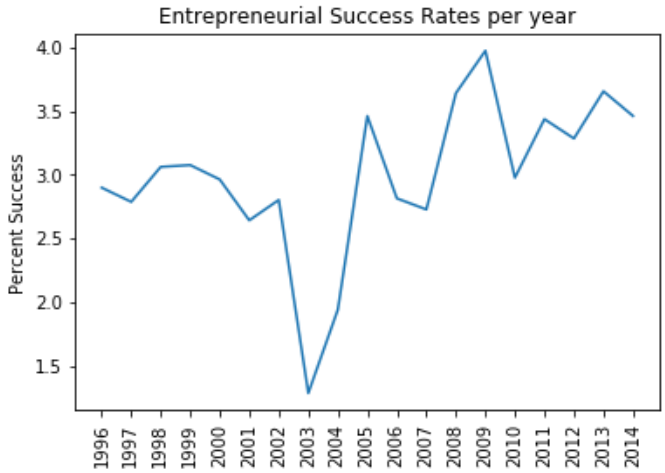


The graph for African American representation shows that their representation in the entrepreneurial community has remained generally stagnant with a few variations between 1996 and 2015. The graph for female representation shows that their representation in the entrepreneurial community has seen an overall decline.

Further analysis shows that for the most recent year, 2015, both African Americans and females are underrepresented in the entrepreneurial community when compared to the total U.S. population demographic. According to the U.S. Census Bureau, for the year 2015, 13.2% of the population was African American and 50.8% of the population was female. This means that there was 5.8% discrepancy for African Americans and a 10.9% discrepancy for females.

Objective 4

After defining success as the turnover rate Here is the graph showing the success rate over the years 1996-2014 as per our definition.



Based on our definition of success, the rate of success generally increased over time from the years 1996 to 2014. The rate of success was approximately 3%. A sharp decline was observed in the years 2003-2004. Though outside the scope of our project, we suspect that the decline in 2003 is due to the aftermath of the 2002 Stock Market crash.

The variables with the highest correlation to success, according to our definition, were as follows:

Variable	***R^2* score**
Education Level	0.03029
Family Income	0.02690
Urbana-Champaign, IL	0.01149
Connecticut	0.01126

This study shows that family income and education level are the two most important factors that contribute to success of entrepreneurs. This is expected since one would expect that being a successful entrepreneur requires a certain amount of capital and professional expertise; however, because the R^2 values for both of these factors are less than 0.1, these factors do not have a significant correlation with success overall.

The correlation values for the demographics of interest (gender and ethnicity) are as follows:

Variable	***R^2* score**
Male	-4.865
Female	-3.196
Caucasian	-39.80
African American	0.00191
Native American	0.0
Asian	0.00274

In terms of minorities, both gender and race do not seem to have any direct correlation with success. Some of the R^2 values did not make sense for gender or race factors since they were out of the range of -1 to 1. This can be attributed to the fact that we had a very few data points that were considered successful (~ 3%). This can also

help explain why the correlations for graduation and family income were also low, as even though more successes on average seemed to be attributed to those two variables, the vast majority of the responses would still have been unsuccessful.

The classification models we created that used the combined factors of race, gender, family income, region and state to predict success were successful, as the accuracy levels of the decision trees model and *k*-nearest neighbors model on the test data were 0.93 and 0.98, respectively. Using a *k* value of 5, the *k*-nearest neighbors algorithm provided the better model for predicting entrepreneurial success.

Conclusion

Our results and analysis show interesting revelations concerning minority representation in the entrepreneurial community. By giving a breakdown of Boston with respect to property values and location in objective 1, we were able to find that areas with lower average property values were in the outskirts of Boston. The Dorchester/Roxbury area, which was the second lowest in property value, was heavily populated by African Americans and Hispanics. In objectives 2 and 3, we found that individuals who came from wealthier family backgrounds and had some form of college education were more likely to be entrepreneurs. In addition, we found that females and African Americans were underrepresented in the entrepreneurial community relative to the American population. Lastly, in objective 4, we found that the demographics presented in the KIEA data set such as family income, education level, and race had little to do with success in entrepreneurship. Given our findings, we are confident that we have given the Minority Action Project a path to further pursue their objective of identifying the resources available for minority-led startups.