# CS506

# EARI project Final Report

## Team members :

## Nianyi Zhang,

## Haoran Wei,

## Ruikang Wang

## Introduction

*Project description:*

Our mission is mapping and analyzing the relation of education, race, and energy jobs. We are mainly concerned about two questions: First, how have economic benefits, specifically defined as the number of energy efficiency jobs, flowed to different census regions in [Massachusetts](https://cn.bing.com/dict/clientsearch?mkt=zh-CN&setLang=zh&form=BDVEHC&ClientVer=BDDTV3.5.1.4320&q=%E9%A9%AC%E8%90%A8%E8%AF%B8%E5%A1%9E%E5%B7%9E)? How does it vary between advantaged and disadvantaged communities, defined as lower-income, less-educated and enhanced-minority census regions? Second, how does these independent variables like race, ethnicity, education and income influence the dependent variable jobs number in different census regions? What independent variables indicates high correlation with the result?

To address the first question, we want to find out how energy efficiency jobs number would vary as the independent variables changes. Specifically, energy efficiency jobs include highest-paying jobs-architecture and engineering and low-wage jobs like construction and installation trades. We will compare these jobs number independently between majority-white and majority-non-white communities, between Hispanic and non-Hispanic communities, between educated and less-educated communities, between high-income and low-income communities.

To handle the second question, we will build a simple linear regression model on the dataset, and then analyze the statistics to measure the fit of the model and gives some explanations. We will also build a prediction model using KNN, which we could use to predict whether certain job numbers are relatively high in different census regions.

*The datasets description:*

We are working on the 2012-2016 ACS 5-year estimates census tract data for race, ethnic, education, and occupation in [Massachusetts](https://cn.bing.com/dict/clientsearch?mkt=zh-CN&setLang=zh&form=BDVEHC&ClientVer=BDDTV3.5.1.4320&q=%E9%A9%AC%E8%90%A8%E8%AF%B8%E5%A1%9E%E5%B7%9E). These data can be retrieved through the Census Bureau by searching the keywords above. At first, not all of these data are useful to us, so what we have to do is to filter, deleting those data we don't need. And the necessary step for us to do is to give these data proper flags. For instance, we will flag a census region as either educated community or under-educated community depending on what percentage of people in the region has a four-year diploma or higher. We will illustrate all these definitions clearly later in Methodology.

## Methodology

The first question only requires data manipulation skills. At first, we classify these communities by certain definitions. And then we define an ECONOMIC INCLUSION INDEX which is meant to show the comparison between different communities.

**Community Definitions：**

**Income:** Low-income and high-income communities are defined by comparing a region’s median Household Income to the county-specific median Household Income. Median income was used as it better accounts for uneven income distributions, where averages could be skewed to either the high or low end of earnings. A high-income community would be one that has a higher median Household Income compared to the county median Household Income, while a low-income community has a lower median Household Income than county median Household Income.

**Demographics:** There are two comparative groups based on demographics: (a) White communities vs. some other race (African American, American Indian, Asian, Hawaiian, and Pacific Islander, and Other), and (b) Hispanic communities vs. non-Hispanic. The threshold to determine "predominantly" White or "predominantly" Hispanic were based on state-specific averages. A predominantly White community would have a higher proportion of White individuals compared to the statewide average.

A predominantly Hispanic community would have a higher proportion of Hispanic individuals compared to statewide average.

**Education:** Similar to demographics, region-specific baselines were used to identify communities with a higher proportion of individuals with a Bachelor’s degree or higher. An educated community would have a higher proportion of individuals with a Bachelor’s degree or higher.

**Economic Inclusion Index**

The index was generated by comparing resident employed concentration of architecture and engineering occupations, construction and extraction occupations, and installation, maintenance, and repair occupations from the U.S. Census Bureau's American Community Survey. Employment estimates by occupational category were collected for each census tract within the geographies used. The concentration of employment was compared between each set of measures (high income vs. low income, educated vs. under-educated, etc.) for aggregated census tracts. The index measures the comparison of each measure to its opposite category. For example, a 5% employment concentration for architecture and engineering occupations for all employed residents in aggregated high-income census tracts compared to a 2% employment concentration for architecture and engineering occupations for all employed residents in aggregated low-income census tracts would result in a 2.5 index score for aggregated high-income census tracts for that geography.

HIGH INCOME = .05/.02 = 2.5 INDEX – 2.5 times more concentrated in high-income neighborhoods.

As for the second question, we want to build linear regression models on what we have on the first question. Specifically, we want to use the models to explain how factors like race, ethnic, income and education influence the jobs number of “architecture and engineering”, “construction and extraction”, and “installation, maintenance and repair” independently. We want to show how well these data fit the linear regression model and also want to explore the possibility of predicting these jobs number using classification methods we learnt from lectures.

## Summary and Results interpretation

**The first question:**

**In general**, it was found that energy efficiency-related employment is found across all types of communities, advantaged and disadvantaged alike. However, the highest-paying jobs, which are typically found in architecture and engineering are most highly concentrated in advantaged—high-income, educated, White and non-Hispanic communities, while lower-wage energy efficiency jobs in construction and extraction or installation, maintenance, and repair are most often found in disadvantaged communities with lower education and a higher prevalence of ethnic and racial minorities residents. It should be noted that while these are the general trends seen across the state, there is some variation by geography.

**The highest-paying energy efficiency jobs—architecture and engineering—are more likely to be concentrated in high-income, non-Hispanic, educated neighborhoods.** In most of the 14 counties of Massachusetts, energy efficiency-related architecture and engineering jobs are at least one and a half times more likely to belong to residents who live in high-income neighborhoods, except a few exceptions. The overall contrast is most striking in Berkshire County, where high-wage energy efficiency jobs are eight to thirty times more likely to be concentrated in “advantaged” communities (Figures 1 through 4).

**At the same time, low-wage energy efficiency jobs across construction and installation trades are most likely found in disadvantaged communities.** For example, energy efficiency workers living in low-income neighborhoods are most likely to work in construction and installation, particularly in Nantucket, Norfolk and Suffolk (Figures 5 and 6). These construction and installation jobs are more concentrated in the non-white majority, undereducated and Hispanic communities (Figures 7 through 9).

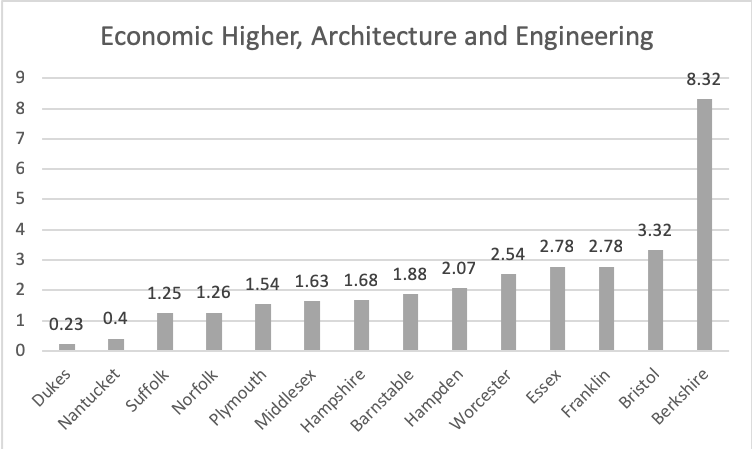


Figure 1

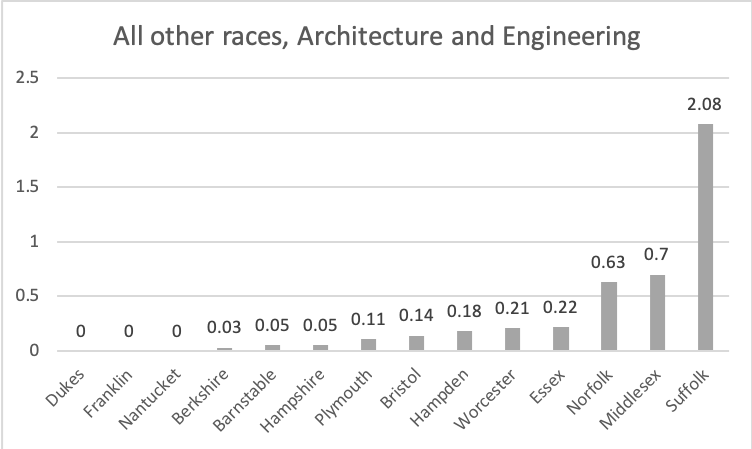


Figure 2

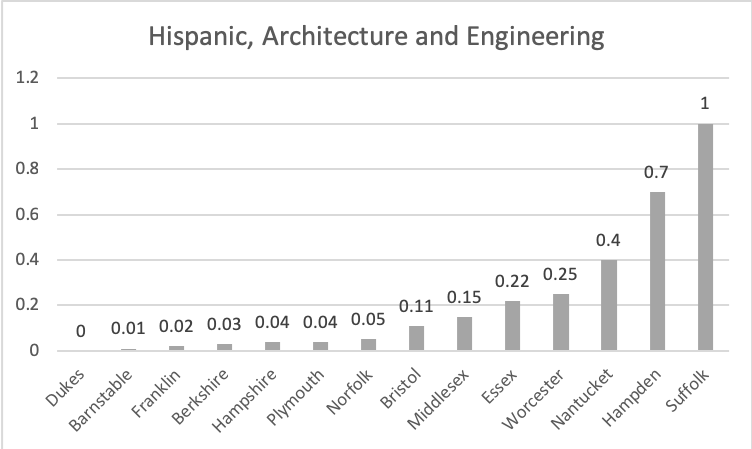


Figure 3

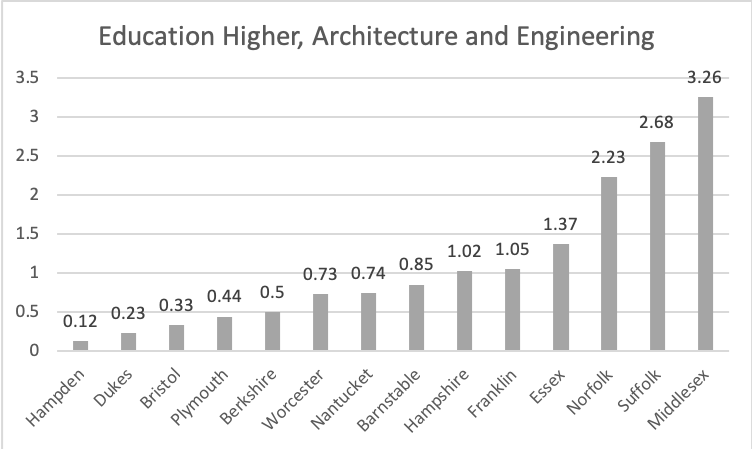


Figure 4

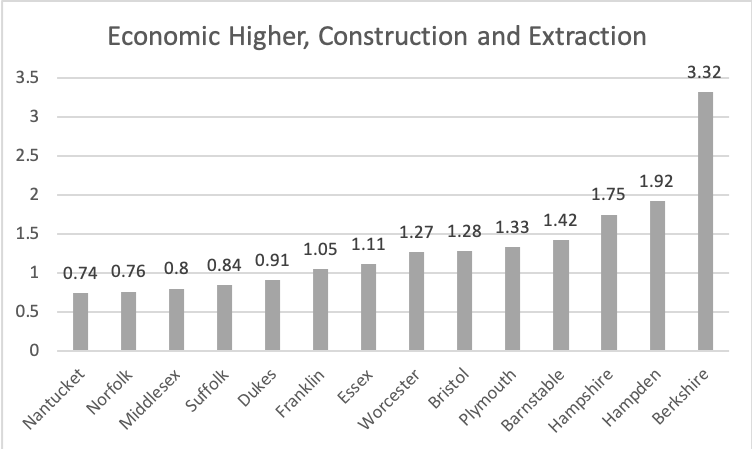


Figure 5

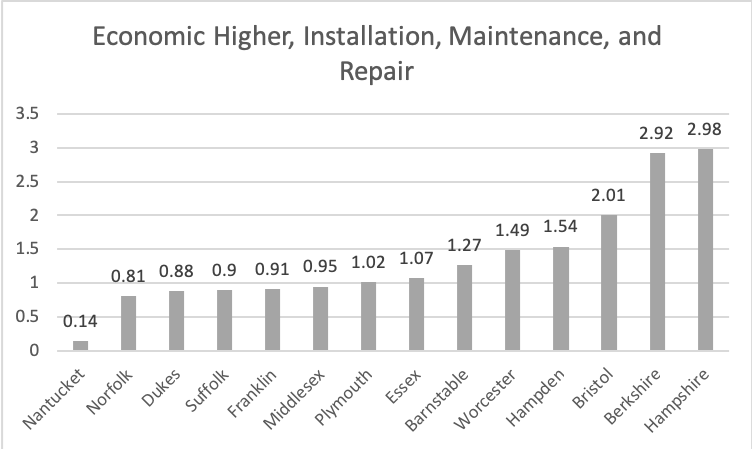


Figure 6

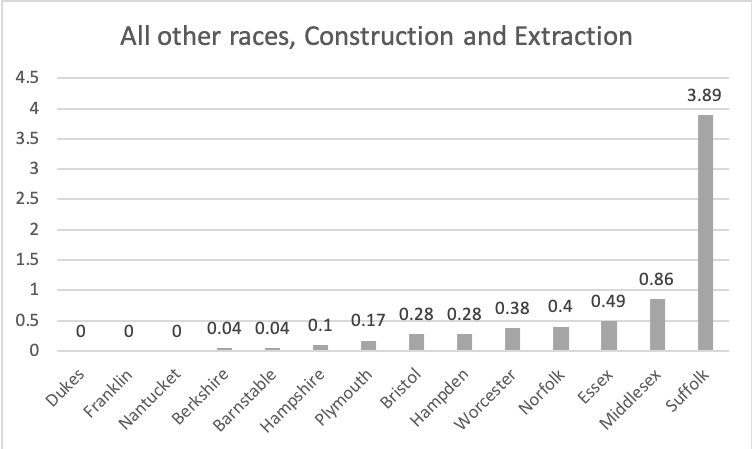


Figure 7

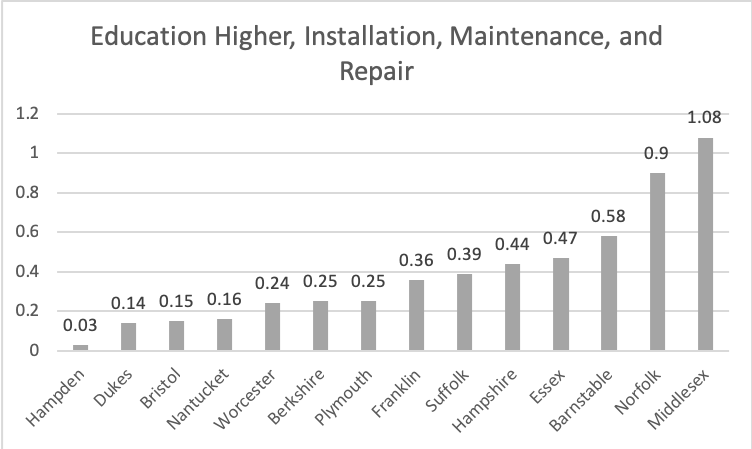


Figure 8

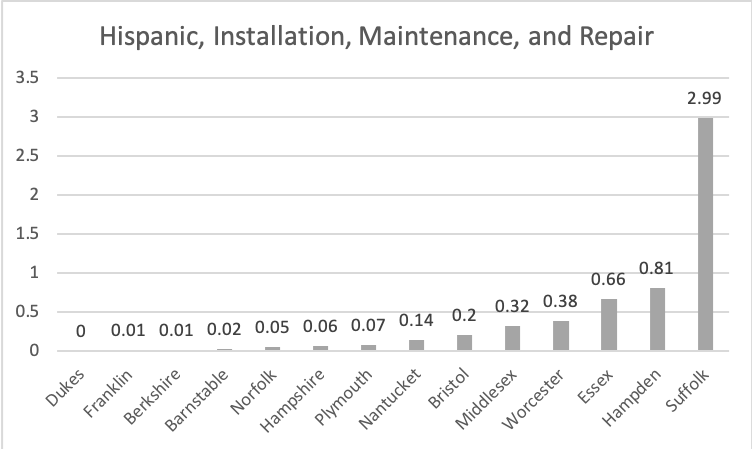


Figure 9

**The Second question**:

We built 3 linear regression models on different type of energy efficient jobs. (Figure10 through 12)

And we use R-squared to measure the fit of a regression model.

This explains how R-squared work:



,,



We know that, the closer the value of is to 1, the better the fit of the regression.

for 3 models are 0.369, 0.427, 0.430. Although it seems that they are not ideal enough, but for explanation problems, they are not bad.

**Figure 10 ['Architecture and Engineering']**

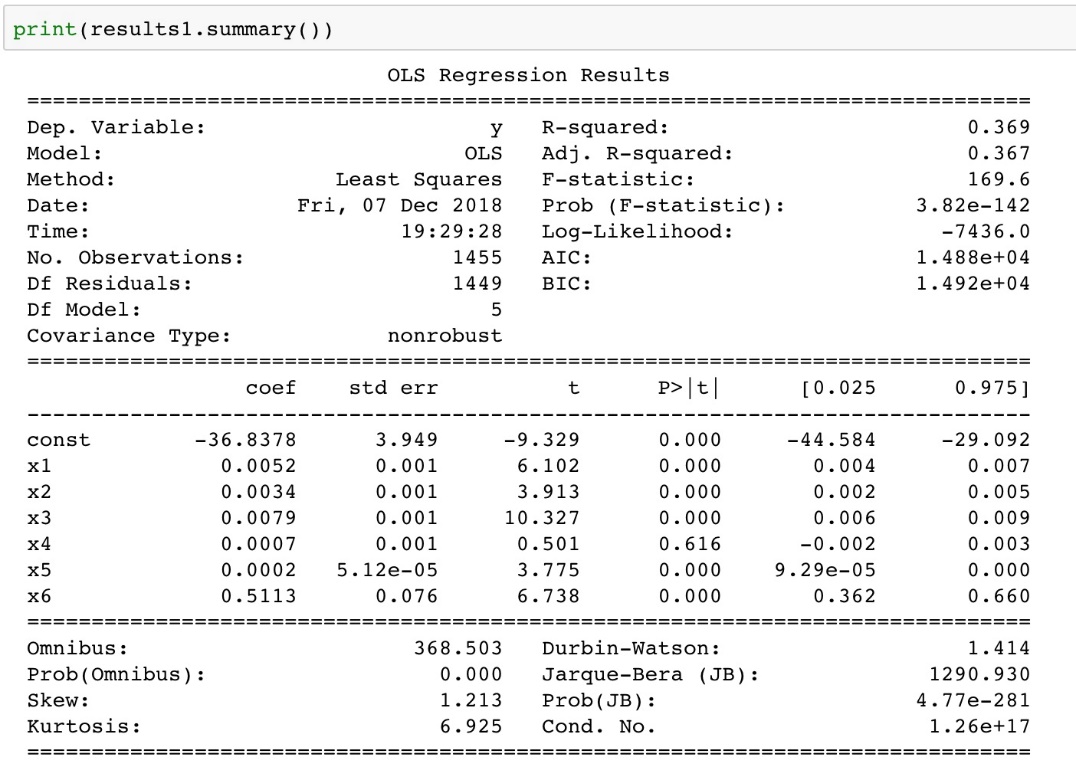


Figure 11 ['Construction and Extraction']

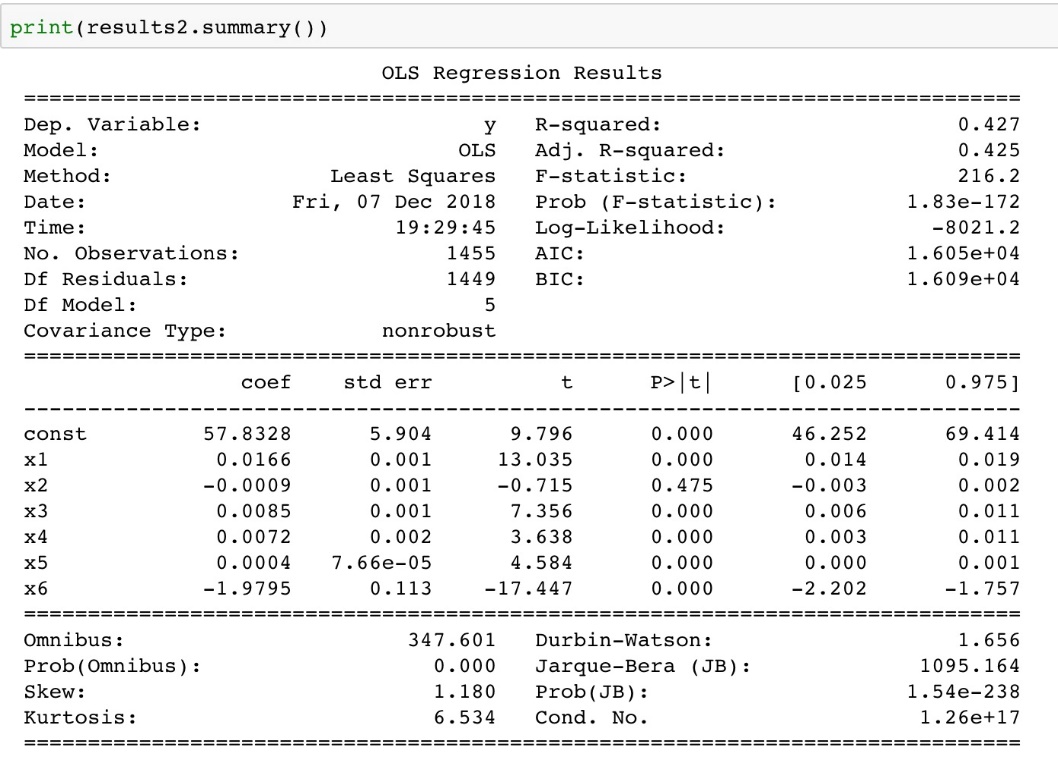
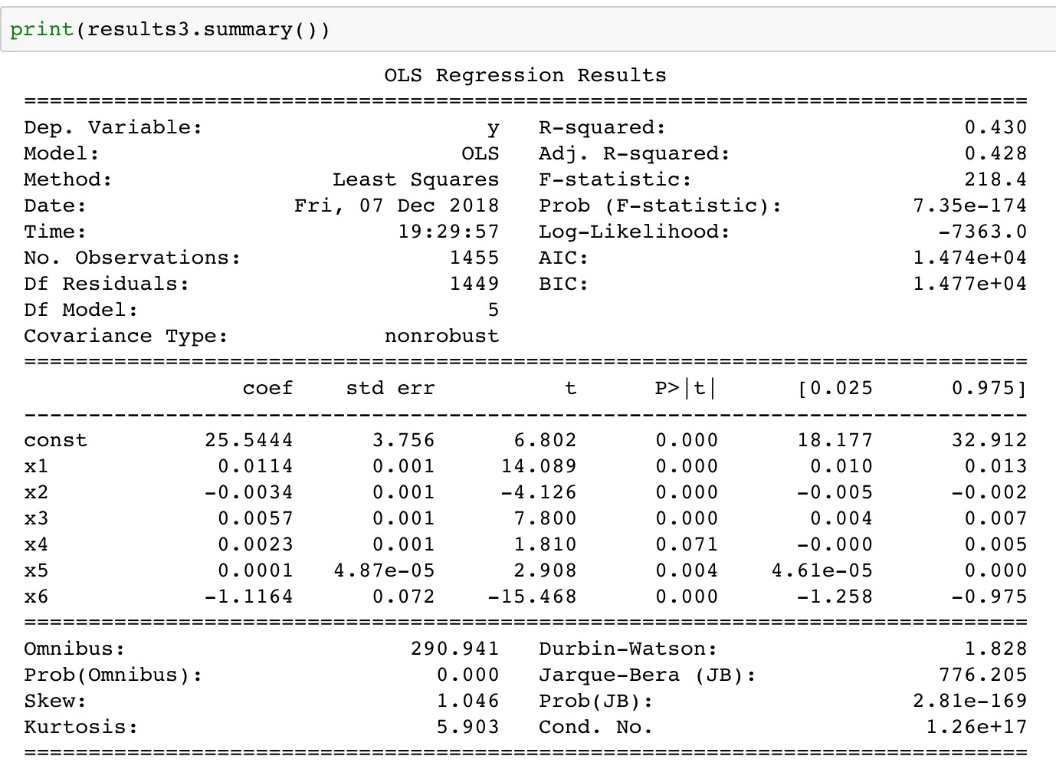


Figure 12 ['Installation, Maintenance, and Repair']



We then compute the confidence intervals and significance of every variables. If the confidence interval for the parameter includes zero, the associated independent variable may not have any predictive value. We can find out for “Architecture and Engineering” and “Installation, Maintenance, and Repair”, “Hispanic” is the only variable which is not significant (Figure 13 and 15). And for “Construction and Extraction”, “Other race” becomes the not significant variable. We can help avoid overfitting by eliminating these not significant variables (Figure 14).

Figure 13

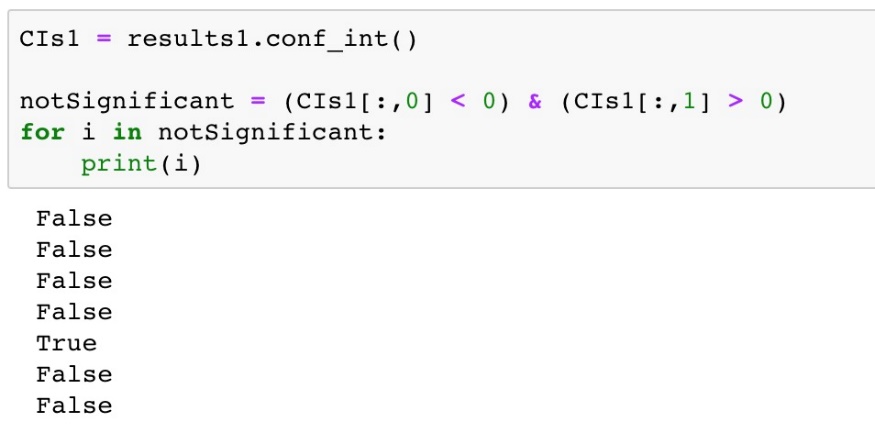


Figure 14

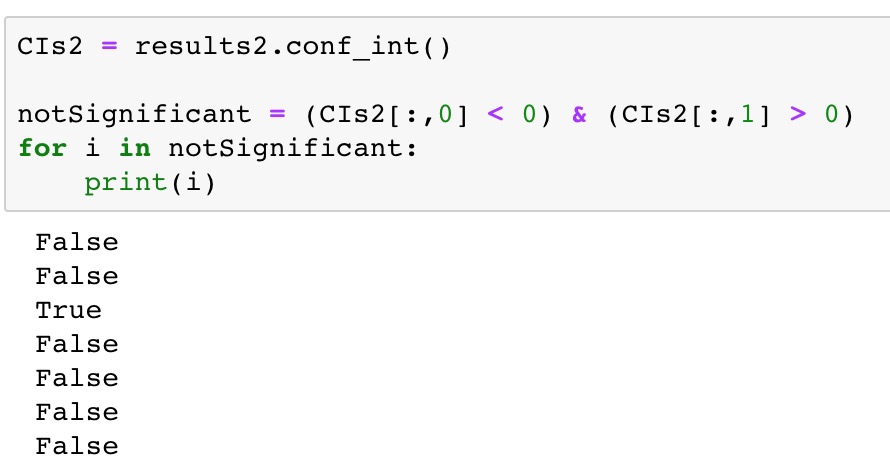
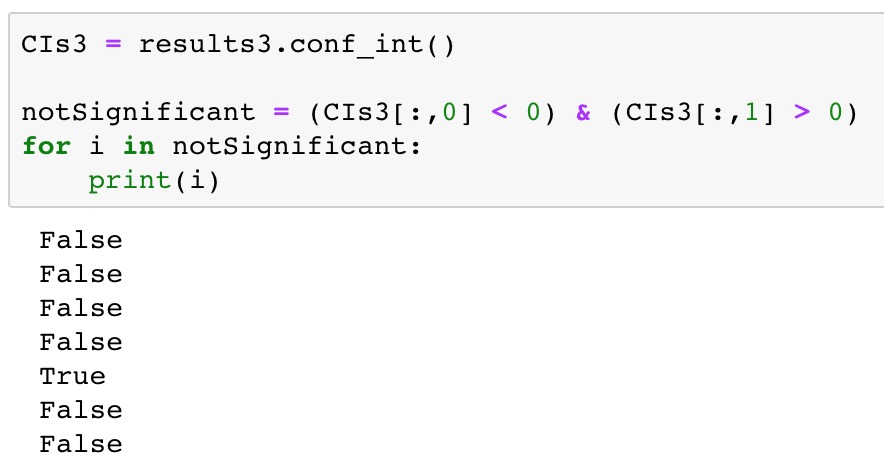
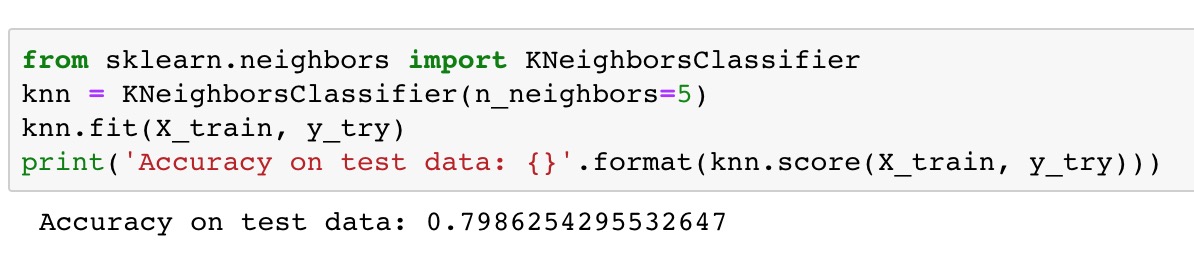


Figure 15



Finally, we built a prediction model, using KNN method. We want to see if we can use these variables like race, ethnic, income and education to predict that whether a certain job is highly concentrated in different regions. The accuracy of our method is 0.799, which is acceptable in prediction. (Figure 16)

Figure 16



## Extensive thoughts

Although the results of linear regression are not good enough, we still could use KNN method to predict whether a certain job is highly concentrated in some regions. Furthermore, as there’re still many kinds of data tables in American Factfinder we didn’t use, such as population density in rural and urban, more detailed analysis could be completed later. And results will be more convincing after American Factfinder update its database to 2018.