# Department of Computer Science

# Energy and Economic Mobility

bu.edu/cs

CS506FINAL

## Project description:

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? 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, ethnic, 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. 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.

## 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.

**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.

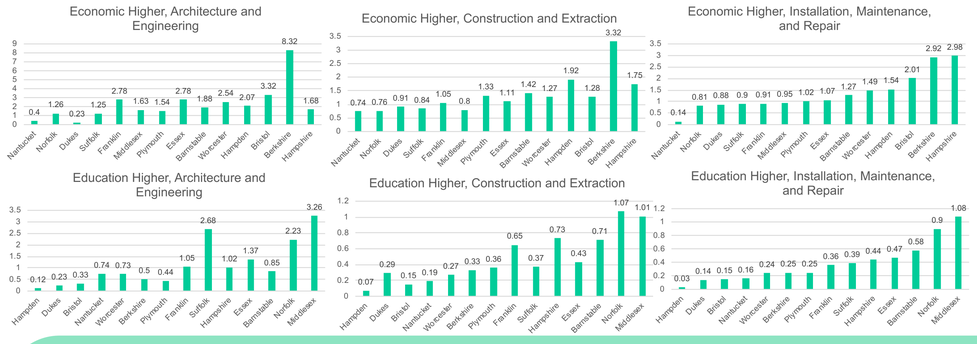
As for the second question, we want to build linear regression models on what we have on the first question. Specifically, we want to use the models to explain how factors like race, ethnic, income and education influence the jobs number of "architecture and engineering", "construction and extraction", 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.

Backg round

Metho dolog y

# Results

First Part




**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.

**At the same time, low-wage energy efficiency jobs across construction and installation trades** are most likely found in disadvantaged communities.

Second Part

We built 3 linear regression models on different type of energy efficient jobs. And we use R-squared to measure the fit of a regression model.

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 .And for "Construction and Extraction", "Other race" becomes the not significant variable. We can help avoid overfitting by eliminating these not significant variables.

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.

```
print(results2.summary())
```

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.427 |
| Model: | OLS | Adj. R-squared: | 0.425 |
| Method: | Least Squares | F-statistic: | 216.2 |
| Date: | Fri, 07 Dec 2018 | Prob (F-statistic): | 1.83e-172 |
| Time: | 19:29:45 | Log-Likelihood: | -8021.2 |
| No. Observations: | 1455 | AIC: | 1.605e+04 |
| Df Residuals: | 1449 | BIC: | 1.609e+04 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 57.8328 | 5.904 | 9.796 | 0.000 | 46.252 | 69.414 |
| x1 | 0.0166 | 0.001 | 13.035 | 0.000 | 0.014 | 0.019 |
| x2 | -0.0009 | 0.001 | -0.715 | 0.475 | -0.003 | 0.002 |
| x3 | 0.0085 | 0.001 | 7.356 | 0.000 | 0.006 | 0.011 |
| x4 | 0.0072 | 0.002 | 3.638 | 0.000 | 0.003 | 0.011 |
| x5 | 0.0004 | 7.66e-05 | 4.584 | 0.000 | 0.000 | 0.001 |
| x6 | -1.9795 | 0.113 | -17.447 | 0.000 | -2.202 | -1.757 |

| | | | |
|---|---|---|---|
| Omnibus: | 347.601 | Durbin-Watson: | 1.656 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 1095.164 |
| Skew: | 1.180 | Prob(JB): | 1.54e-238 |
| Kurtosis: | 6.534 | Cond. No. | 1.26e+17 |

```
print(results3.summary())
```

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | 0.430 |
| Model: | OLS | Adj. R-squared: | 0.428 |
| Method: | Least Squares | F-statistic: | 218.4 |
| Date: | Fri, 07 Dec 2018 | Prob (F-statistic): | 7.35e-174 |
| Time: | 19:29:57 | Log-Likelihood: | -7363.0 |
| No. Observations: | 1455 | AIC: | 1.474e+04 |
| Df Residuals: | 1449 | BIC: | 1.477e+04 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 25.5444 | 3.756 | 6.802 | 0.000 | 18.177 | 32.912 |
| x1 | 0.0114 | 0.001 | 14.589 | 0.000 | 0.010 | 0.013 |
| x2 | -0.0034 | 0.001 | -4.126 | 0.000 | -0.005 | -0.002 |
| x3 | 0.0057 | 0.001 | 7.800 | 0.000 | 0.004 | 0.007 |
| x4 | 0.0023 | 0.001 | 1.810 | 0.071 | -0.000 | 0.005 |
| x5 | 0.0001 | 4.87e-05 | 2.908 | 0.004 | 4.61e-05 | 0.000 |
| x6 | -1.1164 | 0.072 | -15.468 | 0.000 | -1.258 | -0.975 |

| | | | |
|---|---|---|---|
| Omnibus: | 290.941 | Durbin-Watson: | 1.828 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 776.205 |
| Skew: | 1.046 | Prob(JB): | 2.81e-169 |
| Kurtosis: | 5.903 | Cond. No. | 1.26e+17 |


Team members: Ruikang Wang, Nianyi Zhang, Haoran Wei


@BUCompSci