Analyzing the Income Disparities between Foreign-Born and Native-Born U.S. workers and Factors that Affect the Income Levels of Immigrants

Authors: Jing Chen, Dvija Muktesh Shah, Jaini Chetan Gala

Summary

Background of Project and Related Works: According to the U.S. Census Bureau American Community Survey(ACS) 5-year estimates, the foreign-born population in the United States was estimated to be 44,011,870 in 2019, which accounts for 14% of the total population in the U.S. The research conducted by Barry R. Chiswick (1978) suggests that, for white men, earnings are unrelated to whether they are foreign-born or U.S. citizens. Stacey Fitzsimmons et al. (2020) study the effect of gender and race on earnings and suggest that white men received more in annual pay than women of color. The 2019 ACS 5-Year Estimates indicates that the estimated median personal income of native-born workers is $38,000, and the estimated median income of foreign-born workers is $32,000. This report studies whether native-born workers still have higher personal income than foreign-born workers after controlling for demographic factors, educational level, work hours, and occupation.

Description of the Data: Every year, The U.S. Census Bureau conducted the American Community Survey(ACS) to gather information such as place of birth, citizenship status, educational attainment, household income, English language proficiency, occupation, and housing characteristics. The ACS survey covers approximately one percent of the United States population. In this project, we use the 2019 ACS 1-Year Estimates Public Use Microdata Sample (PUMS) data which includes variables for nearly every question on the ACS survey.

Project Goals and Methods: This project aims to provide a better understanding of the difference in income levels between native-born and foreign-born workers in the U.S. We also want to find the factors that are strongly correlated with the immigrants' income levels and provide recommendations on how to improve income levels for low-income immigrant families. The project also aims to provide a helpful tool for people to better understand the spatial distribution of immigrants and their basic information, such as median income level and their origin.

Brief Description of Methods and Results: We visualize and perform regression on PUMS data to study income disparity between native-born and foreign-born workers. We reject the null hypothesis after controlling the factors for age, race, gender, the number of hours worked per week, educational attainment, state of residence, occupation, and English speaking ability. Therefore, we conclude there is a difference in personal income level between native-born and foreign-born workers. We also performed the regression analysis on the log income and the potential factors affecting income levels. We found that higher English speaking skill helps foreign-born workers to improve their income level. Foreign-born workers with a college/associate degree have a higher income than those with only a high school degree. Foreign-born workers with at least a Bachelor's degree have a higher income than those with a college/associate degree. We've successfully created an interactive map to visualize the percentage of the immigrant population and basic information of immigrants in each Public Use Microdata Area (PUMA) using 2015-2019 5-year American Community Survey estimates.

Methods

For data collection and processing, we create a program to load 2019 U.S Census PUMS data of all states into data frames using ***get_pums()*** function under the ***tidycensus*** package and grouping 530 four-digit OCCP codes into nine groups. Since our analysis focuses on the income level, especially the wage level of workers, the data has been filtered to only contain people who are employed, 25 years old or above, do not work from home, do not work in the military, and with wage greater than 0. We created one additional variable, “nativity,” based on the place of birth(POBP) and citizenship status(CIT) of workers, which contains two categories: native-born and foreign-born.

We perform spatial analysis to study the distribution of the immigrants in the U.S. and how the income level of immigrants differs among areas. By grouping the data using subjects' place of birth, we study how the income level of workers differs among immigrant communities. We import the PUMA shapefile into R for data visualization, merge the PUMA shapefile with data using geo_join by PUMA geography, and create interactive maps using ***leaflet*** package.

We use the ACS PUMS data to study if nativity would impact the income level of workers. The dataset was randomly separated into training and testing sets. The training set contains 20% of the data used for EDA and preliminary analysis. The hypothesis and testing model were generated based on the related published studies and the EDA results. We use linear regression models to examine the difference in income between native-born workers and foreign-born workers with similar demographic characteristics, education background, and occupation type by controlling the factors and covariates, such as age, gender, total hours worked per week, and educational attainment of workers, and median income in the area where the workers live. PUMS data contains uniquely identified observations and person weight(PWGTP) associated with each observation. Weighting the data using PWGTP would bring the PUMS estimates closer to the published ACS estimates/Decennial Census numbers. *svyglm* in the survey package is used for building linear regression models on survey-based data with sample weight. According to *survey* package documentation, these functions perform weighted estimation, with each observation being weighted by the inverse of its sampling probability, which gives precision estimates that incorporate the effects of stratification and clustering. Please see the model and definition of variables below:

Model 1 tests the relationship between income and nativity without controlling for other factors

$$Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity$$

Model 2 tests the relationship between income and nativity while controlling for other factors

$$Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \beta_6 English$$
$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

Model 3 adds an additional control factor for geography (state)

$$Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \beta_6 English\ +\ \beta_7 ST$$
$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

Model 4 is designed for studying the impact of English speaking ability and educational attainment on income level.

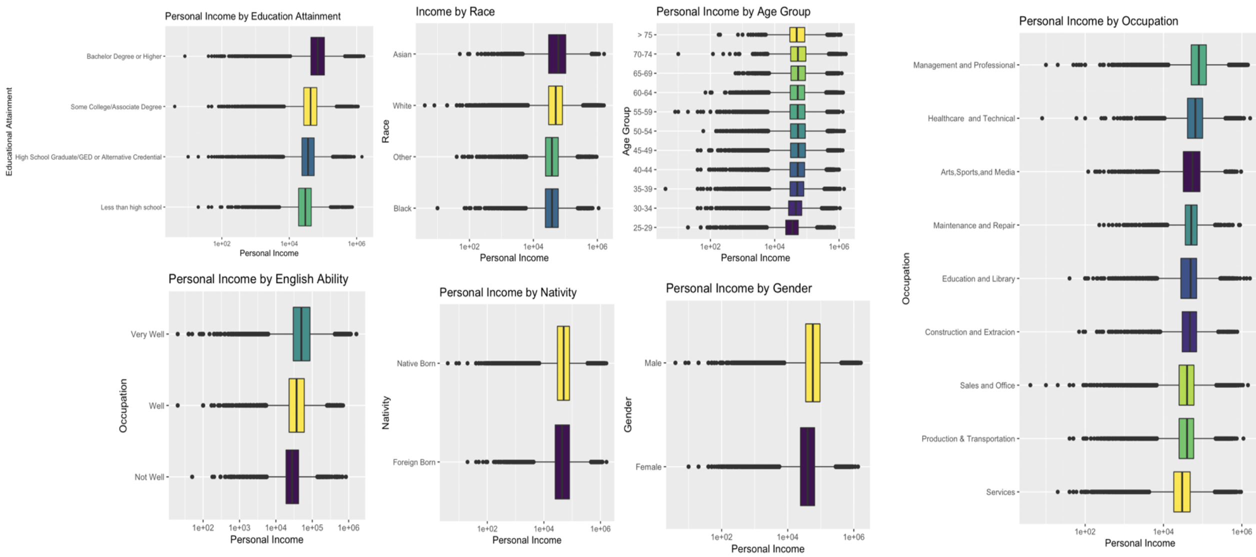
$$Log\ Income\ =\ \beta_0\ +\ \beta_1 English\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \beta_6 ST$$
$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

- $Income$: log10 of personal income in U.S. dollar
- $Nativity$: Native-born refer to the population who were U.S. citizens at birth, including U.S. citizens born in the U.S. or born aboard of American parents. Foreign-born workers refer to the population who were not U.S. citizens at birth, including naturalized U.S. citizens and non-citizens.
- $AGEP$: reported age
- $Gender$: Male and female
- $WKHP$: total number of hours worked per week
- ST: State Code
- $EducationalAttainment\ Factors$: Reported educational attainment levels of subject: Less than high school, high school graduate/ GED or Alternative Credential, Some College/ Associate Degree, and Bachelor's degree or higher.
- $Engligh$: Reported English speaking ability of the subject: Not well, well, very well
- $Occupation\ Factors$: Reported occupation codes combined into 9 categories: "Management & Professional", "Services", "Sales and Office", "Construction and Extraction", "Maintenance and Repair", "Production and Transportation", "Educational and Library ", "Arts, Sports, and Media", and "Healthcare and Technical".

Results

***EDA Results*:**

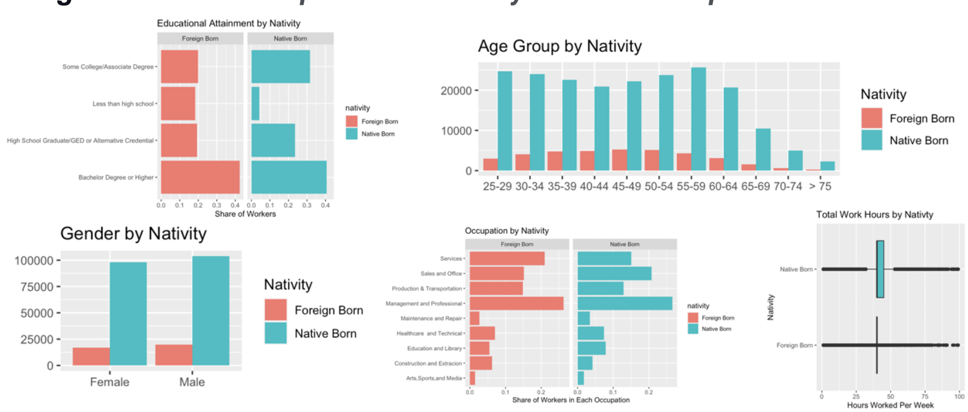
***Figure 1:* Relationship between independent and dependent variables**



(See Appendix for Details)

According to the analysis, Foreign-born workers have higher personal income than Native-born workers. Various factors affect the income level of the workers, and the relationship between Personal income level and the independent factors have shown in ***Figure 1***. People having Bachelor degrees or higher and the people in Management and Professional occupation or Healthcare and Technical occupations earn comparatively more than the other workers. We can also see that Asian workers tend to have a higher median personal income than workers of other races. Personal income level tends to increase with the increase in age, but we can see a slight decrease in the income level after the age of 60. We can also see that male workers' income level is slightly higher than female workers. Workers having better English speaking ability tend to earn more than other workers.

***Figure 2:* Relationship between nativity and other independent variables**



(See Appendix for Details)

Based on the EDA, there are more native-born workers than foreign-born workers in the U.S. The native-born workers have a higher share of people who are Aged 25-29 and 55-59 than foreign-born workers, whereas the foreign-born workers have a higher share of people who are Aged 45-54. Native-born workers tend to work for more hours per week compared to foreign-born workers. Native-born workers seem to have a lower share of people without high-school degrees compared with foreign-born workers.

Regression Analysis Results

Null Hypothesis (H0)*: There is no difference in personal income level between native-born and foreign-born workers*
Alternative Hypothesis (Ha)*: There is a difference in personal income level between native-born and foreign-born workers*

$$Model\ 1:\ Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity$$

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         10.602093   0.002788 3803.22   <2e-16 ***
nativityNative Born  0.123783   0.003086   40.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If we don't control other factors, our team rejects the null hypothesis at the 0.001 significance level. There is a difference in personal income level between native-born and foreign-born workers. The log personal income of native-born workers is 0.12 units greater than the log personal income of foreign-born workers.

$$Model\ 2: Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \beta_6 English$$
$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                      8.507e+00  1.141e-02 745.274  < 2e-16 ***
nativityNative Born                             -1.157e-02  3.122e-03  -3.705 0.000468 ***
AGEP                                             1.444e-02  7.289e-05 198.137  < 2e-16 ***
genderMale                                       2.531e-01  2.347e-03 107.857  < 2e-16 ***
WKHP                                             2.793e-02  1.203e-04 232.216  < 2e-16 ***
educational_attainmentBachelor Degree or Higher  4.687e-01  2.790e-03 167.996  < 2e-16 ***
educational_attainmentLess than high school     -1.218e-01  3.949e-03 -30.854  < 2e-16 ***
educational_attainmentSome College/Associate Degree  1.115e-01  2.538e-03  43.926  < 2e-16 ***
english_abilityVery Well                         1.358e-01  5.262e-03  25.819  < 2e-16 ***
english_abilityNot Well                         -9.076e-02  7.219e-03 -12.571  < 2e-16 ***
raceAsian                                        4.472e-02  4.381e-03  10.208 1.18e-14 ***
raceBlack                                       -1.216e-01  3.172e-03 -38.336  < 2e-16 ***
raceOther                                       -3.738e-02  3.576e-03 -10.452 4.79e-15 ***
(Full results can be found in Appendix)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Reference level for Categorical Variables:
Educational Attainment (High School Degree or GED); Race (White);and English Ability(Well).
```

After controlling for additional variables, we still reject the null hypothesis at the 0.001 significance level. There is a difference in personal income level between native-born and foreign-born. However, after controlling for other factors, such as age, gender, race, the number of hours worked per week, English speaking ability, educational attainment, and occupation, the result shows that the income level for native-born is lower than foreign-born workers.

The Log personal income of Native-Born workers is 0.0115 units lower than the personal income of foreign-born workers. The Log personal income of male workers is approximately 0.253 units more than that of female workers. With a unit change in the number of hours worked per week, the log income changes by 0.0279 units. English speaking ability shows a strong impact on the log income levels. If a worker is fluent in English, he or she earns 0.1358 units more in terms of their log income, and if a worker does not speak English well, he or she makes 0.0907 units lower in log income. Log income level differs depending on the race of the workers, compared to a white worker, if the person is an Asian, he or she earns 0.0447 units more in terms of log income.

$$Model\ 3: Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \beta_6 English\ +\ \beta_7 ST$$
$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      8.342e+00  1.440e-02 579.261  < 2e-16 ***
nativityNative Born                              1.709e-02  3.183e-03   5.369 0.000451 ***
AGEP                                             1.445e-02  7.408e-05 195.091  < 2e-16 ***
genderMale                                       2.500e-01  2.330e-03 107.307 2.69e-15 ***
WKHP                                             2.819e-02  1.189e-04 237.119  < 2e-16 ***
educational_attainmentBachelor Degree or Higher  4.535e-01  2.678e-03 169.391  < 2e-16 ***
educational_attainmentLess than high school     -1.191e-01  3.909e-03 -30.466 2.16e-10 ***
educational_attainmentSome College/Associate Degree  1.082e-01  2.503e-03  43.244 9.44e-12 ***
english_abilityVery Well                         1.422e-01  5.245e-03  27.119 6.11e-10 ***
english_abilityNot Well                         -9.022e-02  7.163e-03 -12.596 5.09e-07 ***
raceAsian                                        3.634e-03  4.457e-03   0.815 0.435902
raceBlack                                       -1.169e-01  3.395e-03 -34.415 7.29e-11 ***
raceOther                                       -7.479e-02  3.817e-03 -19.593 1.09e-08 ***
(Full results can be found in Appendix)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Reference level for Categorical Variables:
Educational Attainment (High School Degree or GED); Race (White);and English Ability(Well).
```

We still reject the null hypothesis at the significance level of 0.001. Adding one more predictor that is State to model 2 shows that there is a difference in the personal income level between native-born workers and foreign-born workers. In model 2, the personal income of native-born workers is 0.0115 units lower than the personal income of foreign-born workers. And after adding the State predictor, it changes to 0.0170 units higher than the personal income of foreign-born workers. The coefficients of other factors in model 2 are also slightly different from the results from model 2.

$$Model\ 4\text{: } Log\ Income\ =\ \beta_0\ +\ \beta_1 English\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ \ +\ \beta_6 ST$$

$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

```
Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      8.5367580  0.0431203 197.976  < 2e-16 ***
AGEP                                             0.0118354  0.0002141  55.275 9.11e-14 ***
genderMale                                       0.2507487  0.0052639  47.636 4.01e-13 ***
raceAsian                                        0.0165391  0.0055412   2.985 0.013697 *
raceBlack                                       -0.1134013  0.0083379 -13.601 8.93e-08 ***
raceOther                                       -0.0470693  0.0064063  -7.347 2.46e-05 ***
WKHP                                             0.0265202  0.0003263  81.281 1.94e-15 ***
english_abilityVery Well                         0.1610192  0.0059581  27.025 1.11e-10 ***
english_abilityNot Well                         -0.1013082  0.0074150 -13.663 8.55e-08 ***
When using 'High School Degree/GED' as reference level
educational_attainmentBachelor Degree or Higher  0.3221086  0.0080104  40.211 2.16e-12 ***
educational_attainmentLess than high school     -0.0579717  0.0074223  -7.811 1.45e-05 ***
educational_attainmentSome College/Associate Degree  0.0547292  0.0070706   7.740 1.57e-05 ***
When using 'Some College/Associate Degree'as reference level
educational_attainmentBachelor Degree or Higher  0.2673794  0.0079332  33.704 1.25e-11 ***
educational_attainmentHigh School Graduate/GED  -0.0547292  0.0070706  -7.740 1.57e-05 ***
educational_attainmentLess than high school     -0.1127010  0.0082371 -13.682 8.43e-08 ***
When using 'Management and Professional Occupation'as reference level
occupationMaintenance and Repair                -0.4797778  0.0143415 -33.454 1.35e-11 ***
occupationConstruction and Extracion            -0.4855983  0.0116120 -41.819 1.47e-12 ***
occupationProduction & Transportation           -0.6250248  0.0089122 -70.132 8.47e-15 ***
occupationServices                              -0.7322400  0.0087137 -84.034 1.39e-15 ***
occupationEducation and Library                 -0.5739185  0.0120991 -47.435 4.18e-13 ***
occupationArts,Sports,and Media                 -0.3845876  0.0218576 -17.595 7.48e-09 ***
occupationHealthcare  and Technical              0.0311675  0.0113838   2.738 0.020909 *
occupationSales and Office                      -0.5412648  0.0092614 -58.443 5.22e-14 ***
When using 'Service'as reference level
occupationSales and Office                       0.1909752  0.0082967  23.018 5.41e-10 ***
occupationProduction & Transportation            0.1072152  0.0074763  14.341 5.38e-08 ***
occupationManagement and Professional            0.7322400  0.0087137  84.034 1.39e-15 ***
occupationMaintenance and Repair                 0.2524622  0.0127368  19.821 2.34e-09 ***
occupationHealthcare  and Technical              0.7634074  0.0125013  61.066 3.37e-14 ***
occupationArts,Sports,and Media                  0.3476524  0.0228300  15.228 3.02e-08 ***
occupationEducation and Library                  0.1583215  0.0144480  10.958 6.83e-07 ***
occupationConstruction and Extracion             0.2466417  0.0104045  23.705 4.05e-10 ***
(Full results can be found in Appendix)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

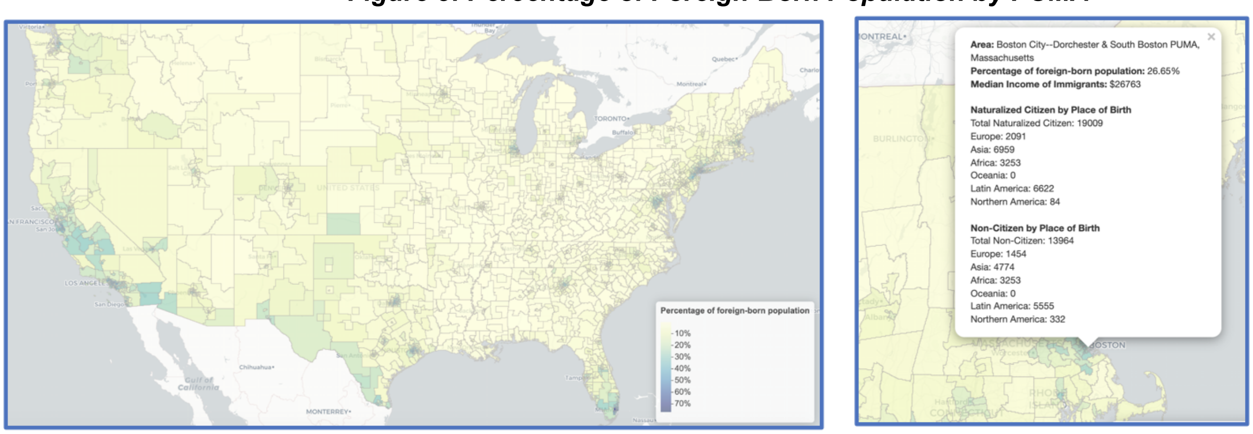
For this model, we are still considering predictors of English speaking ability, age, gender, race, the number of hours worked per week, state, educational attainment factors, and occupation as control factors or covariates. Based on model 4, at the 0.001 level significance level, we found that English speaking ability improves the log personal income levels. Log income of foreign-born workers who speak English very well is expected to be 0.161 units higher than those who speak English just well. Educational attainment shows a strong impact on the earnings of a worker. The log personal income of a worker with a High School Degree/GED is 0.3221 units less than the log personal income of a worker with Bachelor's Degrees or Higher, 0.0547 units less than the log personal income of those with Some College/Associate Degree, and only 0.0579 units higher than the log personal income of a worker with Less than High school degrees. The log personal income of a worker with a Some College/Associate Degree is 0.2673 units less than the log personal income of a worker with Bachelor's Degrees or Higher, 0.0547 units more than the log personal income of a worker with High School Degree/GED, and only 0.1127 units more than the log personal income of a worker with Less than High school degrees.

At the significance level of 0.001, there is a difference in the income level between workers in the Management and Professional occupation and other occupations (except Healthcare and Technical occupation). On comparing the occupations based on the log personal income, the log personal income of workers in Management and Professional Occupation is 0.479 units higher than workers in Maintenance and Repair, 0.4855 units higher in Construction and Extraction, 0.6250 units higher in Production and Transportation, 0.7322 higher in Services, 0.5739 higher in Education and Library, 0.3845 higher in Arts, Sports and Media, 0.5412 lower in Sales and Office. At the significance Level 0.001, we don't see any difference in the Management and Professionals worker's log income levels and Healthcare and Technical worker's log personal income levels. At the significance level of 0.001, there is a difference in the income level between workers in the Service occupation and other occupations. Compared to the log personal income of workers, the log income foreign-born workers in Service is 0.19 units higher in Sales and Office occupation, 0.1 units higher in production and transportation, 0.732 units higher in Management and Professional occupations, 0.25 units higher in Maintenance and Repair, 0.763 units higher in Healthcare and Technical occupation, and 0.246 units higher than Construction and Extraction occupation.

<u>Project Product: Interactive Map</u>

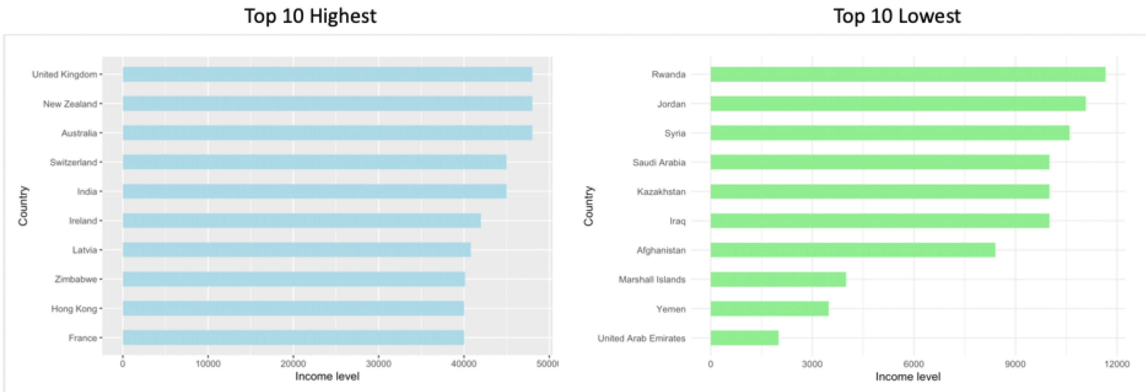
We created an interactive map that shows the percentage of the immigrant population by PUMA. When clicking the map, a pop-up text box will display the basic information about the PUMA, including the area name, percentage of the immigrant population, the median income of the foreign-born population, counts by place of birth (shown in ***Figure 3***). Immigrants are more concentrated in cities, especially coastal cities, than suburbs and rural areas. Coastal cities, such as Miami, Los Angeles, and San Jose, have the highest share of the immigrant population.

Figure 3: Percentage of Foreign-Born Population by PUMA



Our team also compared the median income level of immigrants by their country of birth. Personal income level varies among immigrants from different countries of the world. In 2019, among all immigrants in the U.S., the immigrants from the United Kingdom, New Zealand, and Australia had the highest median personal income level. Immigrants from the United Arab Emirates had the lowest median personal income level (approximately $2,000), followed by and Yemen, Saudi Arabia, Kazakhstan, and Iraq (shown in ***Figure 4***).

Figure 4: Top 10 Immigrant communities with highest/lowest median personal income



Discussion

Meaning of the Results

There is a difference in the income level between the native-born workers and the foreign-born workers. Based on our analysis, we can say that the income level of the native-born workers is comparatively higher than the immigrant workers. Various factors affect the income level of immigrants. This project helps the Immigrant community with recommendations on improving the income levels for low-income immigrant families.

The income level of the immigrants depends on various factors such as location, educational attainment, occupation, gender, age, and the number of hours worked per week. Personal income levels vary among immigrants from different countries in the world. The United Kingdom, New Zealand and Australia have the highest income level, whereas the United Arab Emirates has the lowest income level. The parameters such as Education Attainment and Occupation majorly affect the income level, People with the degree of Bachelors or Higher and Occupation like Management and Professional tend to earn more than people with different degrees and occupations. Income level also depends on the number of hours worked per week. It has also been noticed that Male tend to earn comparatively higher than women and also income increases with age as people try to achieve higher degrees and have some kind of experience in their field of work that can bolster their income levels.

Impact and Importance of the Study

Our study can be beneficial to the immigrant community and immigrant organizations to improve their financial situations. In order for immigrants to increase their income level, they should get a Bachelor’s or some higher degree and also improve their English speaking skills. We would recommend the Mayor’s Office of Immigrant Advancement to provide pilot programs for immigrants to improve their English speaking ability and their educational levels. For example, the City of Boston currently has a Restaurant Revitalization Fund to offer tuition waivers to eligible restaurant workers in the participating colleges and programs. Immigrants, especially non-citizen count for a high share of the restaurant's workforce. A program like this can be helpful for immigrant groups. Also, The Immigrant Learning Centre (ILC) offers free English language programs to immigrants and refugees to meet their individual needs and help them achieve their goals. These programs not only improve their English speaking skills but also help students gain leadership, problem-solving, organizational, and job skills.

Future Work

In the future, we will explore the methodology for sharing the interactive map online. We also want to inform immigrant organizations, such as the Chinese Progressive Association and Mayor's Office Immigrant Advancement, about the existence of this project and check with them to see how we can put this interactive map in use or if it is possible to share this on their website. Because converting city-level income data to PUMA is too time-consuming, our model only considers the difference in income level by state. In the future, we can still add a covariate for the median income by area to count that cities have a higher share of immigrants than rural areas.

Statement of contributions

Jing Chen developed the functions to pull and map data related to the immigrant population, personal income level, and their characteristics using API by Public Use Microdata Area (PUMA) to study the spatial distribution of immigrants and studies how income levels differ among immigrant communities and regions. Jing scoped the project and developed the hypothesis and linear regression models with teammates to identify the potential factors that affect income levels of immigrants, such as educational attainment and age. Jing designed and create an interactive map that visualizes the demographics of immigrants by PUMA. Jing worked with other team members to write the project report and prepare the presentation slides.

Dvija Muktesh Shah performed EDA on the data with teammates to identify the potential factors that affect income levels of immigrants, such as age and educational attainment, and scoped the project and did regression using linear regression models to study the relationship between income levels and potential factors that would affect the income levels, such as Educational Attainment factors, Age, Gender, No of Hours worked per week, Median income in PUMA of residence born workers and Occupation factors and tested the hypotheses in this project. Dvija worked with other team members to write the project report and prepare the presentation slides.

Jaini Chetan Gala developed a methodology for combining multiple years of data with inconsistent variable names in the dataset, checked the feasibility of predicting based on the PUMS data, and performed EDA on the data with teammates to identify the explanatory factors. Jaini worked with other teammates to prepare the final report and the presentation slides.

References

1. Chiswick, Barry R. “The Effect of Americanization on the Earnings of Foreign-Born Men.” Journal of Political Economy, vol. 86, no. 5, University of Chicago Press, 1978, pp. 897–921, http://www.jstor.org/stable/1828415.
2. Fitzsimmons, S. R., Baggs, J., & Brannen, M. Y. (2020). “The Immigrant Income Gap.” Harvard Business Review Digital Articles, 2-8. https://hbr.org/2020/05/research-the-immigrant-income-gap
3. U.S. Census Bureau, 2005-2019 American Community Survey 1-Year Estimates, Public Use Microdata Sample (PUMS)
4. U.S. Census Bureau, 2015-2019 American Community Survey 5-Year Estimates, Public Use Microdata Sample (PUMS)
5. R survey package documentation, retrieved from https://cran.r-project.org/web/packages/survey/survey.pdf

Appendix

Regression Results of Model 2:

$$Model\ 2: Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \beta_6 English$$
$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

```
Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                          8.507e+00  1.141e-02 745.274  < 2e-16 ***
nativityNative Born                                 -1.157e-02  3.122e-03  -3.705 0.000468 ***
AGEP                                                 1.444e-02  7.289e-05 198.137  < 2e-16 ***
genderMale                                           2.531e-01  2.347e-03 107.857  < 2e-16 ***
WKHP                                                 2.793e-02  1.203e-04 232.216  < 2e-16 ***
educational_attainmentBachelor Degree or Higher      4.687e-01  2.790e-03 167.996  < 2e-16 ***
educational_attainmentLess than high school         -1.218e-01  3.949e-03 -30.854  < 2e-16 ***
educational_attainmentSome College/Associate Degree  1.115e-01  2.538e-03  43.926  < 2e-16 ***
english_abilityVery Well                             1.358e-01  5.262e-03  25.819  < 2e-16 ***
english_abilityNot Well                             -9.076e-02  7.219e-03 -12.571  < 2e-16 ***
raceAsian                                            4.472e-02  4.381e-03  10.208 1.18e-14 ***
raceBlack                                           -1.216e-01  3.172e-03 -38.336  < 2e-16 ***
raceOther                                           -3.738e-02  3.576e-03 -10.452 4.79e-15 ***
occupationConstruction and Extracion                -3.575e-02  9.526e-03  -3.753 0.000401 ***
occupationEducation and Library                     -1.893e-01  7.931e-03 -23.866  < 2e-16 ***
occupationHealthcare  and Technical                  3.095e-01  8.265e-03  37.447  < 2e-16 ***
occupationMaintenance and Repair                     3.987e-03  8.061e-03   0.495 0.622693
occupationManagement and Professional                2.982e-01  7.562e-03  39.432  < 2e-16 ***
occupationProduction & Transportation               -1.892e-01  7.612e-03 -24.853  < 2e-16 ***
occupationSales and Office                          -9.150e-02  7.681e-03 -11.913  < 2e-16 ***
occupationServices                                  -2.766e-01  7.648e-03 -36.168  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Reference level for Categorical Variables:
Educational Attainment (High School Degree or GED); Race (White); English Ability(Well); and Occupation(Arts,Sports,and Media).
```

Regression Results of Model 3:

$$Model\ 3: Log\ Income\ =\ \beta_0\ +\ \beta_1 Nativity\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \beta_6 English\ +\ \beta_7 ST$$
$$\alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

```
Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                          8.342e+00  1.440e-02 579.261  < 2e-16 ***
nativityNative Born                                  1.709e-02  3.183e-03   5.369 0.000451 ***
AGEP                                                 1.445e-02  7.408e-05 195.091  < 2e-16 ***
genderMale                                           2.500e-01  2.330e-03 107.307 2.69e-15 ***
WKHP                                                 2.819e-02  1.189e-04 237.119  < 2e-16 ***
educational_attainmentBachelor Degree or Higher      4.535e-01  2.678e-03 169.391  < 2e-16 ***
educational_attainmentLess than high school         -1.191e-01  3.909e-03 -30.466 2.16e-10 ***
educational_attainmentSome College/Associate Degree  1.082e-01  2.503e-03  43.244 9.44e-12 ***
english_abilityVery Well                             1.422e-01  5.245e-03  27.119 6.11e-10 ***
english_abilityNot Well                             -9.022e-02  7.163e-03 -12.596 5.09e-07 ***
raceAsian                                            3.634e-03  4.457e-03   0.815 0.435902
raceBlack                                           -1.169e-01  3.395e-03 -34.415 7.29e-11 ***
raceOther                                           -7.479e-02  3.817e-03 -19.593 1.09e-08 ***
occupationConstruction and Extracion                -1.170e-02  9.352e-03  -1.251 0.242393
occupationEducation and Library                     -1.686e-01  7.845e-03 -21.491 4.81e-09 ***
occupationHealthcare  and Technical                  3.378e-01  8.004e-03  42.200 1.18e-11 ***
occupationMaintenance and Repair                     3.113e-02  7.874e-03   3.954 0.003336 **
occupationManagement and Professional                3.129e-01  7.456e-03  41.970 1.23e-11 ***
occupationProduction & Transportation               -1.582e-01  7.415e-03 -21.338 5.12e-09 ***
occupationSales and Office                          -7.117e-02  7.517e-03  -9.468 5.63e-06 ***
occupationServices                                  -2.578e-01  7.471e-03 -34.511 7.11e-11 ***
ST10                                                 1.367e-01  1.840e-02   7.428 3.98e-05 ***
```

ST11	3.476e-01	1.782e-02	19.502	1.13e-08	***
ST12	1.414e-02	9.965e-03	1.419	0.189729	
ST13	6.641e-02	1.067e-02	6.225	0.000154	***
ST15	1.957e-01	1.680e-02	11.644	9.95e-07	***
ST16	-1.527e-02	1.845e-02	-0.828	0.429156	
ST17	1.383e-01	1.056e-02	13.106	3.62e-07	***
ST18	4.270e-02	9.627e-03	4.436	0.001634	**
ST19	3.802e-02	1.159e-02	3.280	0.009535	**
ST2	1.537e-01	2.552e-02	6.023	0.000197	***
ST20	2.494e-02	1.148e-02	2.172	0.057960	.
ST21	-9.579e-03	1.178e-02	-0.813	0.436967	
ST22	2.056e-02	1.282e-02	1.603	0.143288	
ST23	9.517e-03	1.552e-02	0.613	0.554847	
ST24	2.422e-01	1.165e-02	20.783	6.47e-09	***
ST25	2.539e-01	1.096e-02	23.170	2.47e-09	***
ST26	5.096e-02	1.183e-02	4.307	0.001970	**
ST27	1.657e-01	1.152e-02	14.380	1.63e-07	***
ST28	-5.102e-02	1.461e-02	-3.493	0.006801	**
ST29	2.524e-02	1.184e-02	2.131	0.061903	.
ST30	-4.344e-02	1.988e-02	-2.185	0.056698	.
ST31	4.435e-02	1.541e-02	2.879	0.018213	*
ST32	1.528e-01	1.307e-02	11.692	9.60e-07	***
ST33	1.474e-01	1.405e-02	10.490	2.40e-06	***
ST34	2.650e-01	1.167e-02	22.707	2.96e-09	***
ST35	-4.528e-03	1.606e-02	-0.282	0.784396	
ST36	2.542e-01	1.068e-02	23.790	1.96e-09	***
ST37	1.338e-02	9.976e-03	1.341	0.212691	
ST38	5.661e-02	1.967e-02	2.877	0.018260	*
ST39	5.154e-02	9.866e-03	5.224	0.000547	***
ST4	6.135e-02	1.117e-02	5.491	0.000385	***
ST40	1.222e-02	1.225e-02	0.997	0.344669	
ST41	1.196e-01	1.193e-02	10.026	3.50e-06	***
ST42	9.148e-02	1.109e-02	8.251	1.73e-05	***
ST44	1.616e-01	1.784e-02	9.056	8.11e-06	***
ST45	1.124e-02	1.143e-02	0.984	0.350940	
ST46	1.662e-02	1.902e-02	0.874	0.404899	
ST47	1.565e-02	1.136e-02	1.377	0.201701	
ST48	7.785e-02	9.620e-03	8.092	2.02e-05	***
ST49	5.904e-02	1.371e-02	4.307	0.001971	**
ST5	-2.862e-02	1.288e-02	-2.222	0.053408	.
ST50	4.999e-02	2.472e-02	2.023	0.073805	.
ST51	1.513e-01	9.905e-03	15.274	9.64e-08	***
ST53	2.318e-01	1.075e-02	21.559	4.68e-09	***
ST54	-2.124e-02	1.320e-02	-1.609	0.142089	
ST55	8.337e-02	1.131e-02	7.371	4.23e-05	***
ST56	7.676e-02	2.110e-02	3.638	0.005415	**
ST6	2.434e-01	9.629e-03	25.275	1.14e-09	***
ST8	1.086e-01	1.217e-02	8.924	9.15e-06	***
ST9	2.430e-01	1.144e-02	21.239	5.34e-09	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Regression Results of Model 4:

$$Model\ 4\text{:}\ Log\ Income\ =\ \beta_0\ +\ \beta_1 English\ +\beta_2\ AGEP\ +\ \beta_3 Gender\ +\ \beta_4 Race\ +\beta_5 WKHP\ +\ \alpha\ EducationalAttainment\ Factors\ +\ \gamma\ Occupation\ Factors$$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.5367580	0.0431203	197.976	< 2e-16	***
AGEP	0.0118354	0.0002141	55.275	9.11e-14	***
genderMale	0.2507487	0.0052639	47.636	4.01e-13	***
raceAsian	0.0165391	0.0055412	2.985	0.013697	*
raceBlack	-0.1134013	0.0083379	-13.601	8.93e-08	***
raceOther	-0.0470693	0.0064063	-7.347	2.46e-05	***
WKHP	0.0265202	0.0003263	81.281	1.94e-15	***
educational_attainmentBachelor Degree or Higher	0.3221086	0.0080104	40.211	2.16e-12	***
educational_attainmentLess than high school	-0.0579717	0.0074223	-7.811	1.45e-05	***
educational_attainmentSome College/Associate Degree	0.0547292	0.0070706	7.740	1.57e-05	***
english_abilityVery Well	0.1610192	0.0059581	27.025	1.11e-10	***
english_abilityNot Well	-0.1013082	0.0074150	-13.663	8.55e-08	***
occupationConstruction and Extracion	-0.1010107	0.0231899	-4.356	0.001431	**
occupationEducation and Library	-0.1893309	0.0245358	-7.717	1.61e-05	***
occupationHealthcare and Technical	0.4157551	0.0231616	17.950	6.16e-09	***
occupationMaintenance and Repair	-0.0951902	0.0273904	-3.475	0.005968	**
occupationManagement and Professional	0.3845876	0.0218576	17.595	7.48e-09	***
occupationProduction & Transportation	-0.2404372	0.0228271	-10.533	9.86e-07	***
occupationSales and Office	-0.1566772	0.0223935	-6.997	3.73e-05	***

```
occupationServices                                          -0.3476524  0.0228300 -15.228 3.02e-08 ***
ST10                                                         0.2497148  0.0710177   3.516 0.005573 **
ST11                                                         0.3996689  0.0533302   7.494 2.08e-05 ***
ST12                                                         0.0626834  0.0382760   1.638 0.132532
ST13                                                         0.1449518  0.0407684   3.555 0.005220 **
ST15                                                         0.2863918  0.0447574   6.399 7.84e-05 ***
ST16                                                         0.1360703  0.0672190   2.024 0.070474 .
ST17                                                         0.2037594  0.0393284   5.181 0.000412 ***
ST18                                                         0.0883542  0.0492938   1.792 0.103323
ST19                                                         0.1136397  0.0591078   1.923 0.083452 .
ST2                                                          0.2414110  0.0910742   2.651 0.024284 *
ST20                                                         0.1197867  0.0487871   2.455 0.033949 *
ST21                                                         0.1359674  0.0564828   2.407 0.036857 *
ST22                                                        -0.0273675  0.0617519  -0.443 0.667066
ST23                                                         0.0216484  0.0988049   0.219 0.830977
ST24                                                         0.2536730  0.0432038   5.872 0.000157 ***
ST25                                                         0.2970360  0.0372266   7.979 1.20e-05 ***
ST26                                                         0.1318747  0.0453167   2.910 0.015563 *
ST27                                                         0.2587129  0.0430051   6.016 0.000129 ***
ST28                                                         0.0030443  0.0831471   0.037 0.971514
ST29                                                         0.0869860  0.0494179   1.760 0.108866
ST30                                                        -0.1930767  0.1639925  -1.177 0.266317
ST31                                                         0.1333977  0.0569857   2.341 0.041276 *
ST32                                                         0.2365626  0.0438349   5.397 0.000303 ***
ST33                                                         0.2563708  0.0641750   3.995 0.002539 **
ST34                                                         0.2931219  0.0399488   7.337 2.49e-05 ***
ST35                                                         0.0526725  0.0517495   1.018 0.332749
ST36                                                         0.3216337  0.0390509   8.236 9.12e-06 ***
ST37                                                         0.0926664  0.0413633   2.240 0.048979 *
ST38                                                        -0.1078229  0.0988706  -1.091 0.301053
ST39                                                         0.1412791  0.0410630   3.441 0.006326 **
ST4                                                          0.1424635  0.0396887   3.590 0.004933 **
ST40                                                         0.0902076  0.0448677   2.011 0.072112 .
ST41                                                         0.2656162  0.0420531   6.316 8.72e-05 ***
ST42                                                         0.1741679  0.0454040   3.836 0.003287 **
ST44                                                         0.1755403  0.0577693   3.039 0.012491 *
ST45                                                         0.0708087  0.0482889   1.466 0.173279
ST46                                                        -0.0009675  0.1027672  -0.009 0.992673
ST47                                                         0.0767428  0.0505348   1.519 0.159822
ST48                                                         0.1119975  0.0386018   2.901 0.015797 *
ST49                                                         0.0573403  0.0469312   1.222 0.249805
ST5                                                          0.2007629  0.0572619   3.506 0.005669 **
ST50                                                         0.1486555  0.1401241   1.061 0.313693
ST51                                                         0.2060535  0.0400932   5.139 0.000438 ***
ST53                                                         0.3231661  0.0420938   7.677 1.69e-05 ***
ST54                                                        -0.0952570  0.1276700  -0.746 0.472771
ST55                                                         0.1163524  0.0476600   2.441 0.034772 *
ST56                                                         0.2063115  0.0822966   2.507 0.031076 *
ST6                                                          0.2936506  0.0395899   7.417 2.27e-05 ***
ST8                                                          0.2249194  0.0418441   5.375 0.000312 ***
ST9                                                          0.2663542  0.0440850   6.042 0.000125 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Most relevant code:

Code for retrieving data

```
st_list<-c('AL','AK','AZ','AR','CA','CO','CT','DE','DC','FL','GA','HI','ID','IL','IN','IA','KS','KY','LA','ME','MD','MA','MI','MN','MS','MO','MT','NE','NV','NH','NJ','NM','NY','NC','ND','OH','OK','OR','PA','RI','SC','SD','TN'
,'TX','UT','VT','VA','WA','WV','WI','WY')

for (state_index in st_list) {
  df <- get_pums(
  variables = c("AGEP","SEX","CIT","ENG","SCHL","ESR","COW","OCCP","WAGP",
              "PINCP","SERIALNO","POBP","PUMA","WKHP","JWTRNS","RAC1P",
              "HISP","NAICSP"),
  state = state_index,
  survey = "acs1",
  year = 2019,
  recode = TRUE,
  rep_weights = "person",
  key = "census api key"
  )
  write_csv(df,paste0(state_index,"_data.csv"))
}
```

```
df <- data.frame()

for (state_index in st_list) {
  temp <- read_csv(paste0("acs1_data/",state_index,"_data.csv"),col_types = cols(
    ST = col_double(),
    PUMA = col_double()
  ))
  temp <- temp%>%filter(
    (ESR==1|ESR==2|ESR==4|ESR==5),   # civilian employed
    JWTRNS != 11, # does not work at home
    PINCP > 0,   # income is above 0
    WAGP > 0,
    AGEP>=25,# 25 years old or over
    (OCCP<6000 | OCCP>= 6200) & (OCCP<9800)
    )
  df <- dplyr::bind_rows(df, temp)
}
```

Code for processing data

```
df_processed <- df %>%
  mutate(SCHL1 = as.numeric(SCHL))%>%
  mutate(
    educational_attainment = case_when(
      (SCHL1>=1) & (SCHL1<=15) ~ 'Less than high school',
      (SCHL1>=16) & (SCHL1<=17) ~ 'High School Graduate/GED or Alternative Credential',
      (SCHL1>=18) & (SCHL1<=20) ~ 'Some College/Associate Degree',
      (SCHL1>=21) & (SCHL1<=24) ~ 'Bachelor Degree or Higher',
      (is.na(SCHL1) ~ 'Less than high school')),
    naics2 =  substr(NAICSP,1,2),
    industry = case_when(naics2=="72"  ~ 'Accommodation and Food Services',
                         naics2=="71" ~ 'Arts, Entertainment and Recreation',
                         naics2=="23" ~ 'Construction',
                         naics2=="61" ~ 'Education',
                         naics2=="52" | naics2=="53" ~ 'Finance, Insurance, and Real Estate',
                         naics2=="62" ~ 'Health Care & Social Assistance',
                         naics2=="51" ~ 'Information',
                         naics2=="31" | naics2=="32" | naics2=="33" | naics2=="3M" ~ 'Manufacturing',
                         naics2=="81" ~ 'Other Services',
                         naics2=="54" | naics2=="55" | naics2=="56" ~ 'Professional, Scientific, Management, and Administrative Services',
                         naics2=="92" ~ 'Public Administration',
                         naics2=="44" | naics2=="45" | naics2=="4M" ~ 'Retail',
                         naics2=="22" | naics2=="11" | naics2=="21"  | naics2=="48" | naics2=="49" ~ 'Transportation, Warehousing, & Utilities & Natural resources',
                         naics2=="42" ~ 'Wholesale Trade',
                         naics2=="99" ~ 'Other'),
    occupation = case_when((OCCP>0 & OCCP<2000) | (OCCP>=2100 & OCCP<2200) ~ 'Management and Professional',
                          (OCCP>=3600 & OCCP<4700) | (OCCP>=2001 & OCCP<2100) ~ 'Services',
                          (OCCP>=4700 & OCCP<6000) ~ 'Sales and Office',
                          (OCCP>=6200 & OCCP<7000) ~ 'Construction and Extracion',
                          (OCCP>=7000 & OCCP<7700) ~ 'Maintenance and Repair',
                          (OCCP>=7700 & OCCP<9800) ~ 'Production & Transportation',
                          (OCCP>=2200 & OCCP<2600) ~ 'Education and Library',
                          (OCCP>=2600 & OCCP<3000) ~ 'Arts,Sports,and Media',
                          (OCCP>=3000 & OCCP<3600) ~ 'Healthcare  and Technical',
                          (OCCP>=6000 & OCCP< 6200) | (OCCP>=9800) ~ 'Other'
),
english_ability = case_when(   (is.na(ENG)|ENG=='b') ~ 'Very Well',
                             (ENG==1) ~ 'Very Well',
                             (ENG==2) ~ 'Well',
                             (ENG==3) ~ 'Not Well'),
nativity = case_when(
  CIT>=4&CIT<=5 ~ "Foreign Born",
  CIT>=1&CIT<=3 ~ "Native Born"),
gender = case_when(
      SEX == 1  ~ "Male",
      SEX == 2  ~ "Female"
    ),
race = dplyr::case_when(
      RAC1P==1  ~ "White",
      RAC1P==2  ~ "Black",
      RAC1P==6  ~ "Asian",
      RAC1P!=1 & RAC1P!=2 & RAC1P!=6 ~ "Other"),
income = PINCP
)

df_processed$educational_attainment <- as.factor(df_processed$educational_attainment)

df_processed$english_ability <- as.factor(df_processed$english_ability)
```

Code for visualization and modeling

```
set.seed(1)
train <- createDataPartition(df_processed$nativity, p=0.2, list=FALSE)

df_part_1 <- df_processed[as.integer(train),]
df_part_2 <- df_processed[-as.integer(train),]

ggplot(df_part_1, aes(x=reorder(educational_attainment,income,na.rm = TRUE), y=income,fill=educational_attainment)) +
  geom_boxplot(show.legend = FALSE) +
 labs(x="Educational Attainment",  y="Personal Income", title="Personal Income by Education Attainment") +
  scale_y_log10()+
  coord_flip() +
  scale_fill_viridis_d()

ggplot(df_part_1, aes(x=reorder(race,income,na.rm = TRUE), y=income,fill= race)) +
  geom_boxplot(show.legend = FALSE) +
 labs(x="Race",  y="Personal Income", title="Income by Race") +
```

```
  scale_y_log10()+
  coord_flip() +
  scale_fill_viridis_d()

df_part_1 <- df_part_1 %>%
  mutate(
    # Create categories
    Age_Group = dplyr::case_when(
      AGEP <= 15            ~ "0-15",
      AGEP > 15 & AGEP <= 19 ~ "16-19",
      AGEP > 19 & AGEP <= 24 ~ "20-24",
      AGEP > 24 & AGEP <= 29 ~ "25-29",
      AGEP > 29 & AGEP <= 34 ~ "30-34",
      AGEP > 34 & AGEP <= 39 ~ "35-39",
      AGEP > 39 & AGEP <= 44 ~ "40-44",
      AGEP > 44 & AGEP <= 49 ~ "45-49",
      AGEP > 49 & AGEP <= 54 ~ "50-54",
      AGEP > 54 & AGEP <= 59 ~ "55-59",
      AGEP > 59 & AGEP <= 64 ~ "60-64",
      AGEP > 64 & AGEP <= 69 ~ "65-69",
      AGEP > 69 & AGEP <= 75 ~ "70-74",
      AGEP > 75             ~ "> 75"
    ),
    # Convert to factor
    Age_Group = factor(
      Age_Group,
      level = c("0-15","16-19","20-24", "25-29", "30-34", "35-39", "40-44", "45-49", "50-54", "55-59", "60-64", "65-69", "70-74", "> 75")
    )
  )

ggplot(df_part_1, aes(x=Age_Group, y=income,fill= Age_Group)) +
  geom_boxplot(show.legend = FALSE) +
  scale_y_log10()+
  coord_flip() +
 labs(x="Age Group",  y="Personal Income", title="Personal Income by Age Group") +
  scale_fill_viridis_d()

ggplot(df_part_1, aes(x=nativity, y=income, fill = nativity)) +
  geom_boxplot(show.legend = FALSE) +
  scale_y_log10()+
  coord_flip() +
 labs(x="Nativity",  y="Personal Income", title="Personal Income by Nativity")+
  scale_fill_viridis_d()

ggplot(df_part_1, aes(x=gender, y=income,fill= gender)) +
  geom_boxplot(show.legend = FALSE) +
  scale_y_log10()+
  coord_flip() +
 labs(x="Gender",  y="Personal Income", title="Personal Income by Gender") +
 scale_fill_viridis_d()

ggplot(df_part_1[!is.na(df_part_1$english_ability),], aes(x=reorder(english_ability,income,na.rm = TRUE), y=income,fill= english_ability)) +
  geom_boxplot(show.legend = FALSE) +
  scale_y_log10()+
  coord_flip() +
 labs(x="Occupation",  y="Personal Income", title="Personal Income by English Ability") +
  scale_fill_viridis_d()

ggplot(df_part_1, aes(x=reorder(industry,income,na.rm = TRUE), y=income,fill=industry)) +
  geom_boxplot(show.legend = FALSE) +
  scale_y_log10()+
  coord_flip() +
 labs(x="Industry",  y="Personal Income", title="Personal Income by Industry") +
  scale_fill_viridis_d()

ggplot(df_part_1, aes(x=reorder(occupation,income,na.rm = TRUE), y=income,fill=occupation)) +
  geom_boxplot(show.legend = FALSE) +
  scale_y_log10()+
  coord_flip() +
 labs(x="Occupation",  y="Personal Income", title="Personal Income by Occupation") +
  scale_fill_viridis_d()

ggplot(df_part_1, aes(x=WKHP, y=income)) +
  geom_point(alpha=0.3) +
  geom_smooth(aes(color = 'blue'),show.legend = FALSE)+
  scale_y_log10()+
 labs(x="Hours Worked Per Week",  y="Personal Income", title="Income by Hours Worked per Week") +theme_minimal()

ggplot(df_part_1, aes(y=WKHP, x=Age_Group,fill=Age_Group)) +
  geom_boxplot(show.legend = FALSE) +
  scale_y_log10()+
 labs(x="Age Group",  y="Hours Worked Per Week", title="") +coord_flip()

ggplot(df_part_1, aes(y=WKHP, x=nativity,fill=nativity)) +
  geom_boxplot(show.legend = FALSE) +
 labs(x="Nativity",  y="Hours Worked Per Week", title="Total Work Hours by Nativty") +coord_flip()

ggplot(df_part_1, aes(Age_Group, ..count..)) +
  geom_bar(aes(fill = nativity), position = "dodge")+
  labs(x="",  y="", title="Age Group by Nativity") +
  guides(fill=guide_legend("Nativity"))


ggplot(df_part_1, aes(gender, ..count..)) +
  geom_bar(aes(fill = nativity), position = "dodge")+
  labs(x="",  y="", title="Gender by Nativity") +
  guides(fill=guide_legend("Nativity"))

ggplot(df_part_1, aes(educational_attainment, ..count..)) +
  geom_bar(aes(fill = nativity), position = "dodge")+
  labs(x="",  y="", title="Educational Attainment by Nativity") +
  guides(fill=guide_legend("Nativity")) + theme(axis.text.x = element_text(angle = 40, vjust = 0.5, hjust=1)) +
  coord_flip()

ggplot(df_part_1, aes(occupation, ..count..)) +
  geom_bar(aes(fill = nativity), position = "dodge")+
  labs(x="",  y="", title="Occupation by Nativity") +
```

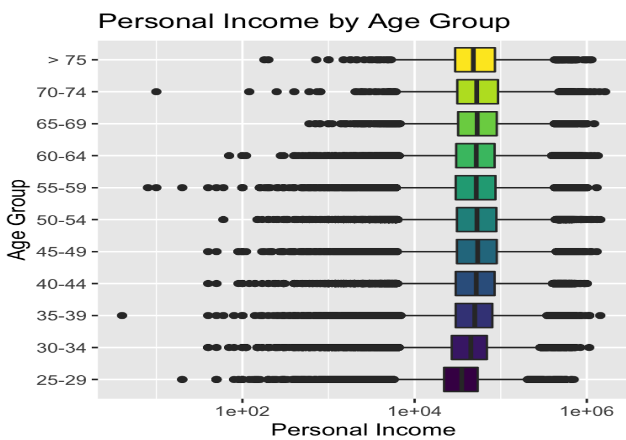
```
  guides(fill=guide_legend("Nativity")) +
  coord_flip()

ggplot(df_part_1, aes(x=educational_attainment, y=income,fill= educational_attainment)) +
  geom_boxplot(show.legend = FALSE) +
  labs(x="Educational Attainment",  y="Personal Income", title="Personal Income by Education Attainment and Occupation") +
  scale_y_log10() +
  coord_flip() + facet_wrap(~occupation,ncol=5) +
  scale_fill_viridis_d()

df_part_2$race <- as.factor(df_part_2$race)
df_part_2<- within(df_part_2, race<-relevel(race,ref=4))
df_part_2<- within(df_part_2, english_ability<-relevel(english_ability,ref=3))
df_part_2$ST <- as.character(df_part_2$ST)
model_sd <- df_part_2%>%
  to_survey()
model_1 <- survey::svyglm(log(income) ~ nativity, design = model_sd)
summary(model_1)
model_2 <- survey::svyglm(log(income) ~  nativity  + AGEP + gender+ WKHP + educational_attainment +english_ability + race + occupation, design = model_sd)
summary(model_2)
model_3 <- survey::svyglm(log(income) ~  nativity  + AGEP + gender+ WKHP + educational_attainment +english_ability + race + occupation +ST, design = model_sd)
summary(model_3)

fb_worker <- df_part_2%>%filter(nativity=="Foreign Born")

model_sd_fb <- fb_worker%>%
  to_survey()


model_4 <- survey::svyglm(log(income) ~  AGEP + gender + race + WKHP + educational_attainment + english_ability + occupation + ST, design = model_sd_fb)
summary(model_4)
```

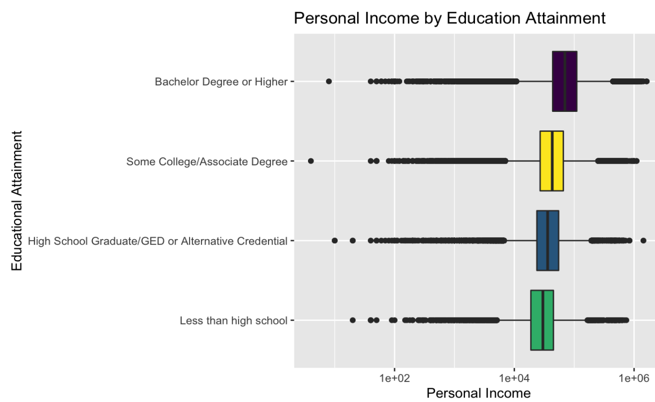
Code for Interactive Map

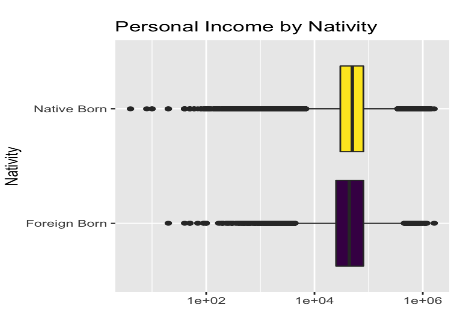
```
endyr=2019
sp=5
tabnum = "B06011"
data <- acs.fetch(endyear=endyr,span=sp,geography=geo,
                    table.number=tabnum,col.names="pretty")

endyr=2019
sp=5
tabnum = "B07004FPR"
data <- acs.fetch(endyear=endyr,span=sp,geography=geo,
                    table.number=tabnum,col.names="pretty")

# convert to a data.frame for merging
temp_df <- data.frame(data@geography$NAME,
                        paste0(str_pad(data@geography$state,2,"left",pad="0"),
                              str_pad(data@geography$publicusemicrodataarea,5,"left",pad="0")),
                        data@estimate[,], stringsAsFactors=FALSE)

write_csv(temp_df,paste0(tabnum,'.csv'))

B05002 <- read_csv('B05002.csv')
colnames(B05002)[2]<- 'GEOID'

B07004FPR <- read_csv('B07004FPR.csv')
colnames(B07004FPR)[2]<- 'GEOID'

B06011 <- read_csv('B06011.csv')

# convert to a data.frame for merging
colnames(B06011)[2]<- 'GEOID'

B05001 <- read_csv('B05001.csv')
colnames(B05001)[2]<- 'GEOID'
}
combined_df <- B05001%>%left_join(B05002,c("GEOID"))%>%left_join(B06011,c("GEOID"))%>%left_join(B07004FPR,c("GEOID"))


data.df <- combined_df[,c(1,2,3,7,8,24:29,31:36,42)]

# do some cleaning of the data frame and create variables of interest
colnames(data.df) <- c("name","GEOID","total","naturalized_citizen",
                         "not_citizen","naturalized_citizen_europe",
                         "naturalized_citizen_asia","naturalized_citizen_africa",
                         "naturalized_citizen_Oceania","naturalized_citizen_latin_america",
                         "naturalized_citizen_northern_america","fb_europe",
                         "fb_asia","fb_africa",
                         "fb_Oceania","fb_latin_america",
                         "fb_northern_america","median_income")

data.df$immigrants<-data.df$naturalized_citizen+data.df$not_citizen
data.df$pct.immigrants <- 100*(data.df$immigrants/data.df$total)

# do the merge
data.df.merged <- geo_join(puma_boundary,data.df,"GEOID","GEOID")

popup <- paste0("<B>", "Area: ", "</B>",data.df.merged$name,
              "<br>", "<B>", "Percentage of foreign-born population: ","</B>", paste0(round(data.df.merged$pct.immigrants,2),"%"),
              "<br>","<B>", "Median Income of Immigrants: ","</B>",paste0("$",data.df.merged$median_income),
              "<br>","",
              "<br>", "<B>","Naturalized Citizen by Place of Birth ","</B>",
               "<br>", "Total Naturalized Citizen: ",data.df.merged$naturalized_citizen,
               "<br>", "  Europe: ",data.df.merged$naturalized_citizen_europe,
              "<br>", "  Asia: ",data.df.merged$naturalized_citizen_asia,
              "<br>", "  Africa: ",data.df.merged$naturalized_citizen_africa,
              "<br>", "  Oceania: ",data.df.merged$naturalized_citizen_Oceania,
              "<br>", "  Latin America: ",data.df.merged$naturalized_citizen_latin_america,
              "<br>", "  Northern America: ",data.df.merged$naturalized_citizen_northern_america,
              "<br>","",
              "<br>", "<B>","Non-Citizen by Place of Birth ","</B>",
               "<br>", "Total Non-Citizen: ",data.df.merged$not_citizen,
               "<br>", "  Europe: ",data.df.merged$fb_europe,
              "<br>", "  Asia: ",data.df.merged$fb_asia,
```

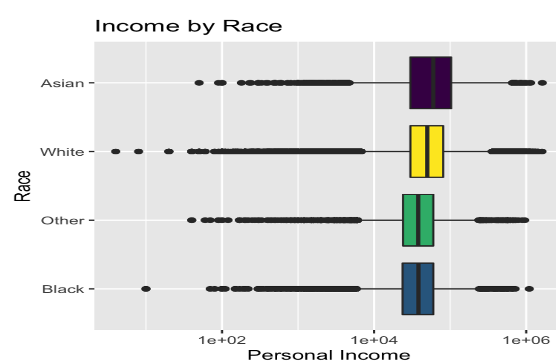
```
          "<br>", "  Africa: ",data.df.merged$naturalized_citizen_africa,
          "<br>", "  Oceania: ",data.df.merged$naturalized_citizen_Oceania,
          "<br>", "  Latin America: ",data.df.merged$fb_latin_america,
          "<br>", "  Northern America: ",data.df.merged$fb_northern_america)

pal <- colorNumeric(
  palette = "YlGnBu",
  domain = data.df.merged$pct.immigrants

immigrant_share<-leaflet() %>%
  addProviderTiles("CartoDB.Positron") %>%
  addPolygons(data = data.df.merged,
          fillColor = ~pal(pct.immigrants),
          color = "#b2aeae", # you need to use hex colors
          fillOpacity = 0.7,
          weight = 1,
          smoothFactor = 0.2,
          popup = popup) %>%
  addLegend(pal = pal,
         values = data.df.merged$pct.immigrants,
         position = "bottomright",
         title = "Percentage of foreign-born population",
         labFormat = labelFormat(suffix = "%"))
```

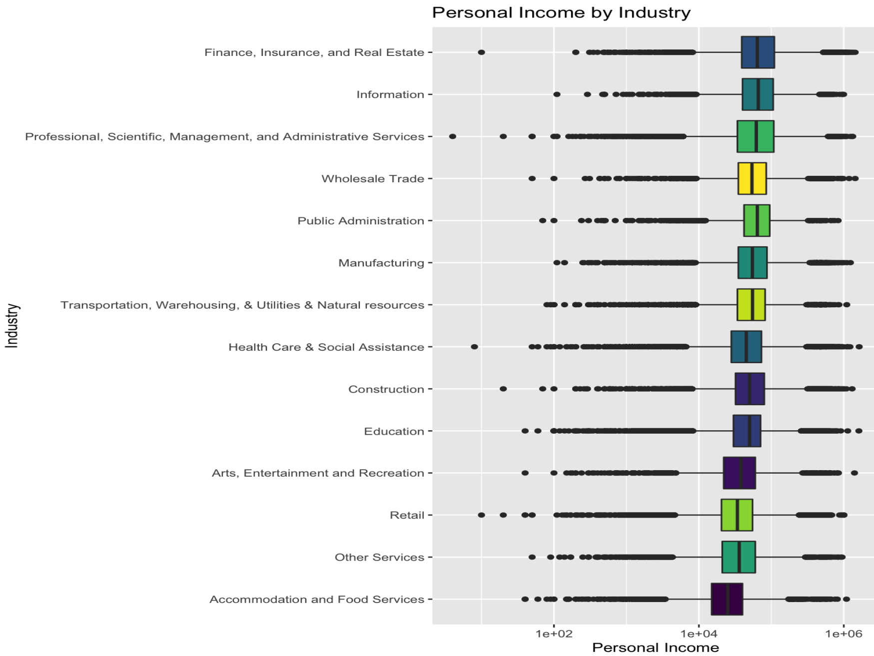
Relationship between independent and dependent variables

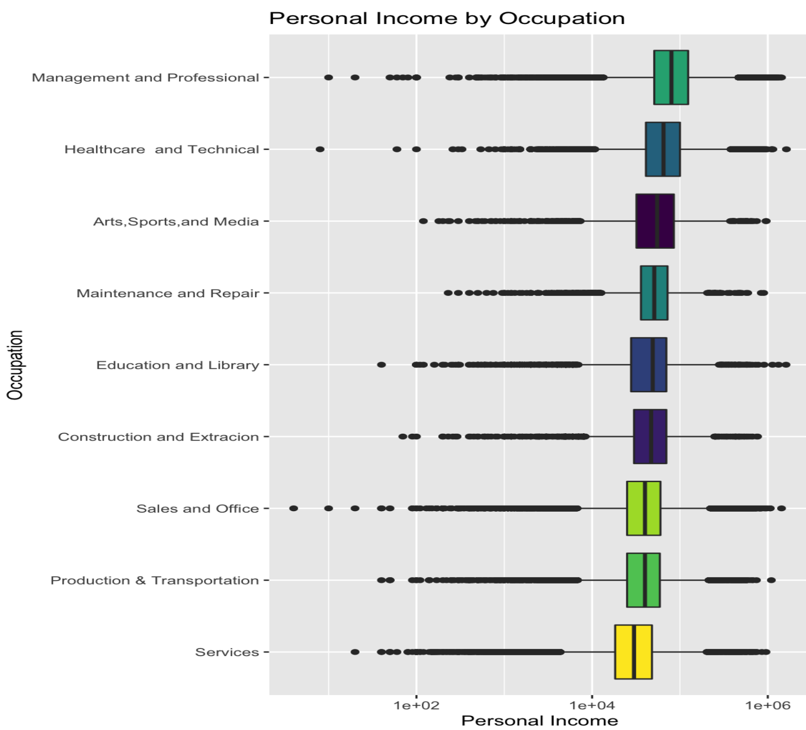









Personal Income by Industry
Finance, Insurance, and Real Estate
Information
Professional, Scientific, Management, and Administrative Services
Wholesale Trade
Public Administration
Manufacturing
Transportation, Warehousing, & Utilities & Natural resources
Health Care & Social Assistance
Construction
Education
Arts, Entertainment and Recreation
Retail
Other Services
Accommodation and Food Services
Industry
1e+02
1e+04
1e+06
Personal Income


Personal Income by Occupation
Management and Professional
Healthcare and Technical
Arts,Sports,and Media
Maintenance and Repair
Education and Library
Construction and Extracion
Sales and Office
Production & Transportation
Services
Occupation
1e+02
1e+04
1e+06
Personal Income

Relationship between Nativity and other Independent variables