Team 2 --- Project Deliverable 2

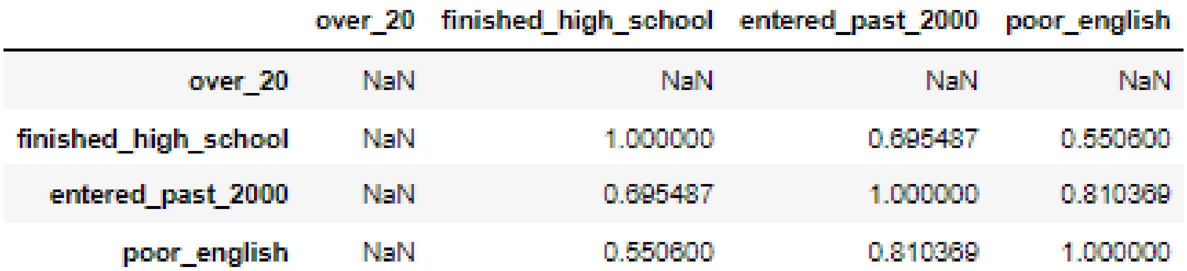
Emily Greven, James Heilberg, Yichuan Philip Ma, Qingyang Xu

Collect and pre-process a secondary batch of data

We preprocessed all the data in deliverable 1 and we have not received any additional data from the client.

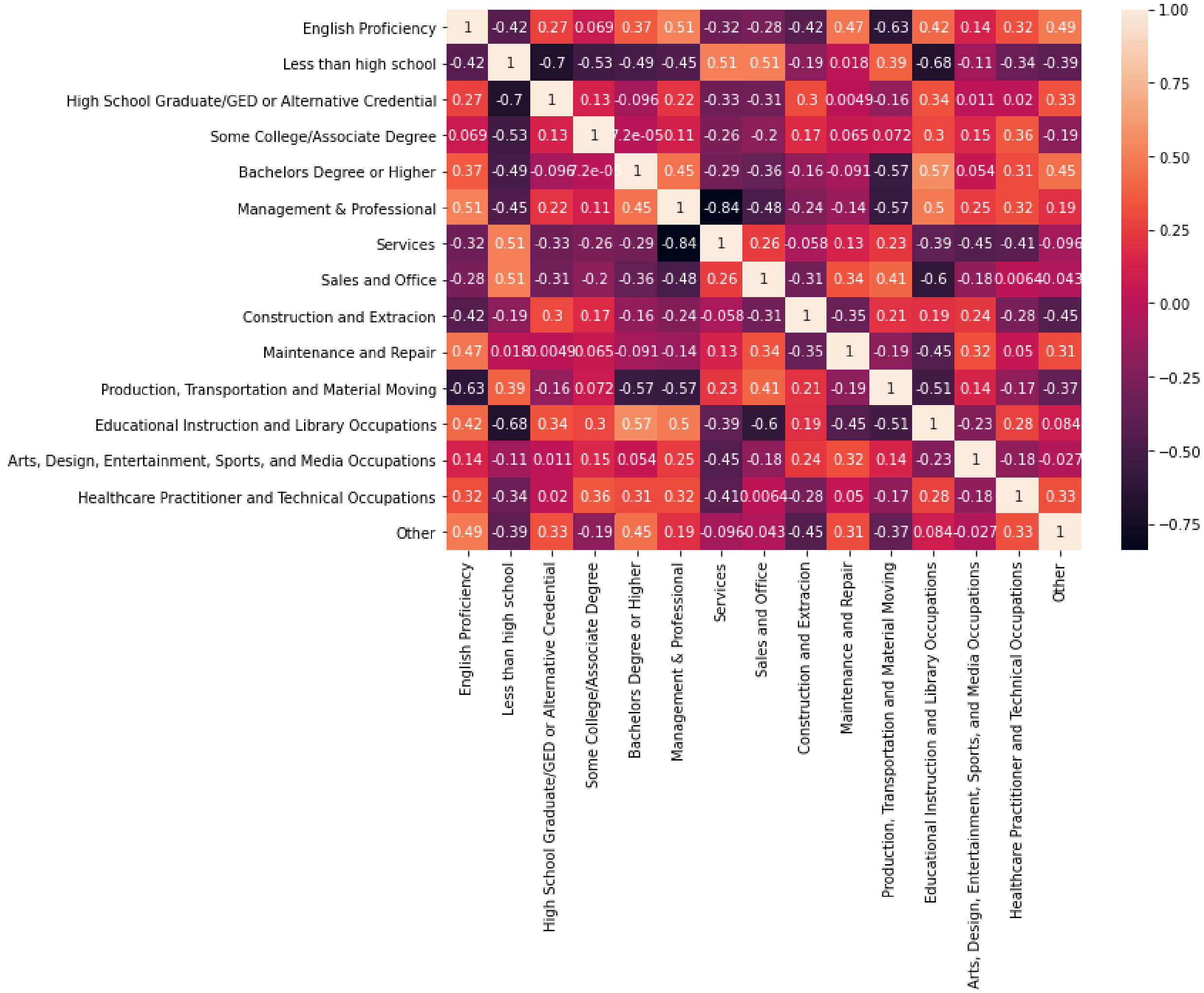
Refine Preliminary Analysis performed in PD1 (figure and analysis below):

Many more visualizations/charts/analysis can be found on the Team 2 branch

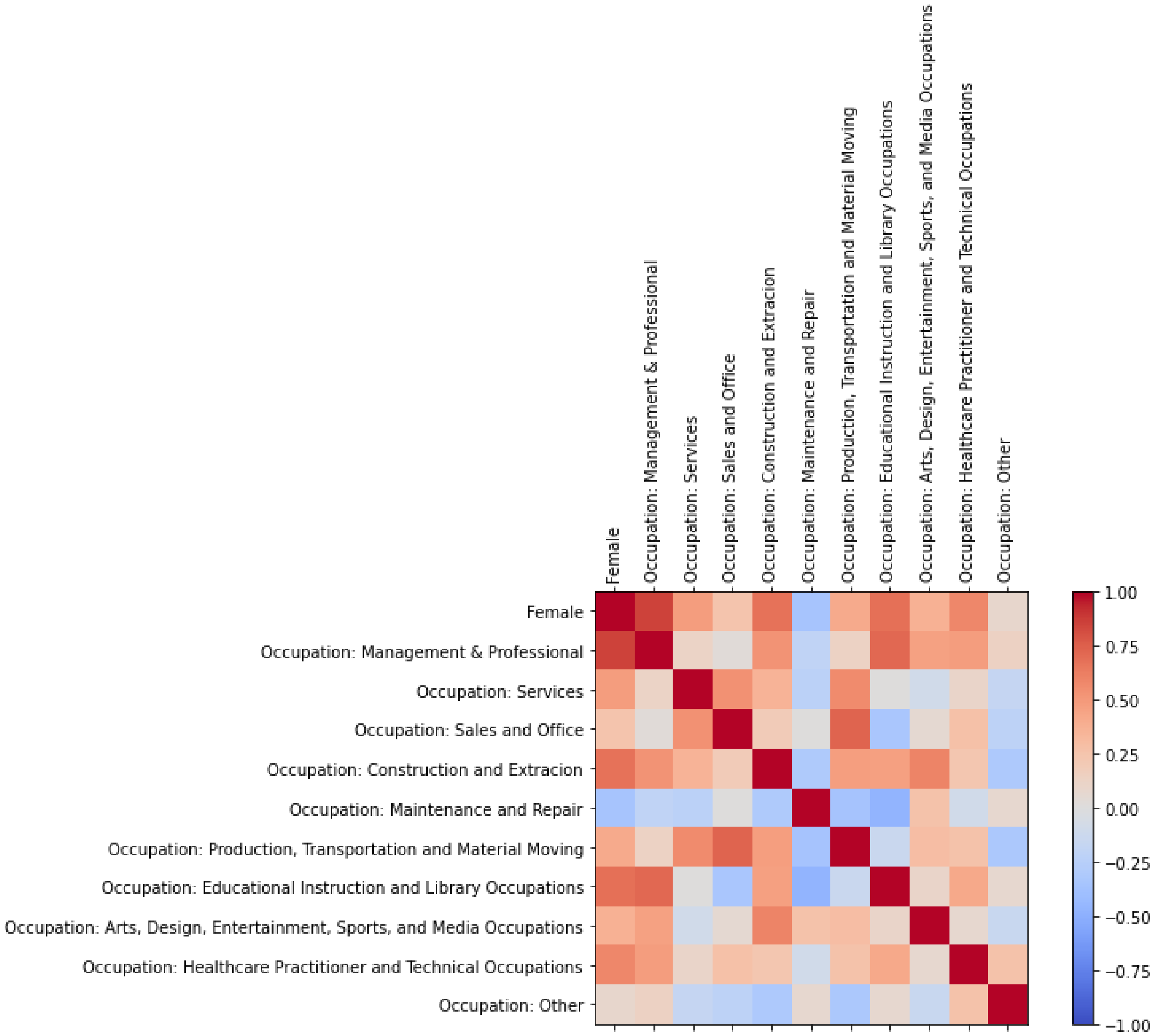


	over_20	finished_high_school	entered_past_2000	poor_english
over_20	NaN	NaN	NaN	NaN
finished_high_school	NaN	1.000000	0.695487	0.550600
entered_past_2000	NaN	0.695487	1.000000	0.810369
poor_english	NaN	0.550600	0.810369	1.000000

In Massachusetts, we have a very strong correlation between entering the US after the year 2000 and not having excellent English speaking skills. However, this is as expected as many immigrating Brazilians do not speak English as their first language. However, if we dive deeper into education we can see that those who entered past 2000 have a very strong linear correlation with those who finished high school or received their GED. This means that while newly immigrating Brazilians may not be super well versed in English, they are still educated to at least the minimum standards of the US. Finally we can see that those with poor English speaking skills have a correlation though not weak or strong with finishing high school or receiving a GED. This could indicate that while Brazilians in the US are getting a high school education, it may mean that they are getting an English education, but not necessarily.

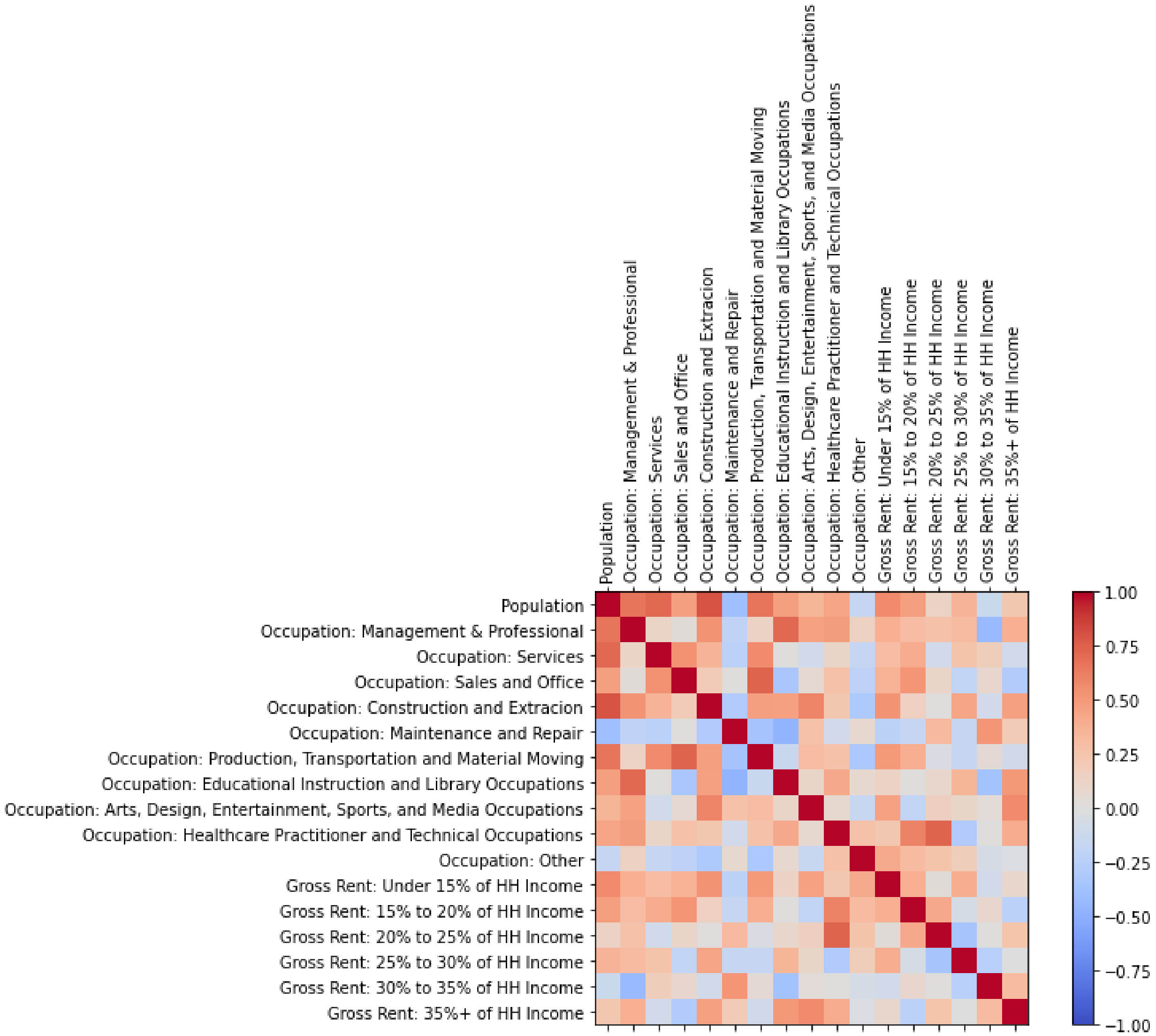


The figure above is a correlation heatmap between English Proficiency, Educational Attainment, and Employment by Occupation for the Brazilian population in Massachusetts. The correlation between English proficiency and individuals with a Bachelor's degree or higher is 0.365, which is lower than what we expect because we suspect there may be other factors affecting the relationship between English proficiency and college education. The correlation between English proficiency and individuals with Management or Professional occupations is 0.51, which is pretty strong as expected. With that being said, we hope to predict English proficiency among the Brazilian population using features for Bachelor's degrees and Management/Professional jobs.

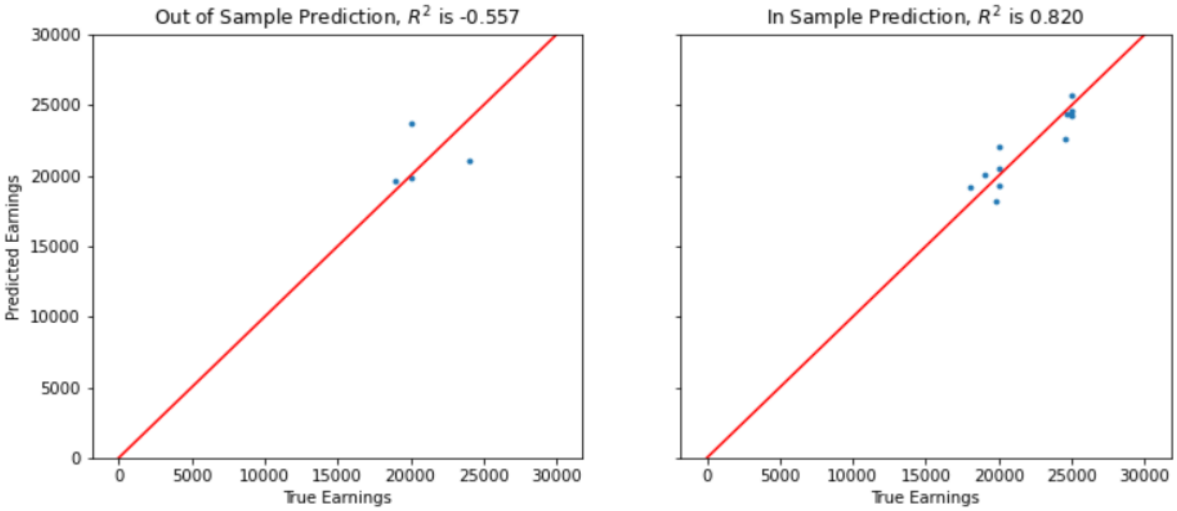


This is a correlation matrix showing female population and Occupations. It seems like there may be a correlation between Females and Management & Professional occupations over the years 2005-2019. Also, a negative correlation between female population and Maintenance and Repair occupations. It may be worth doing a linear regression analysis on Females in Management & Services, Females in Educational Instruction and Library Occupations, and Females in Construction and Extraction occupations.

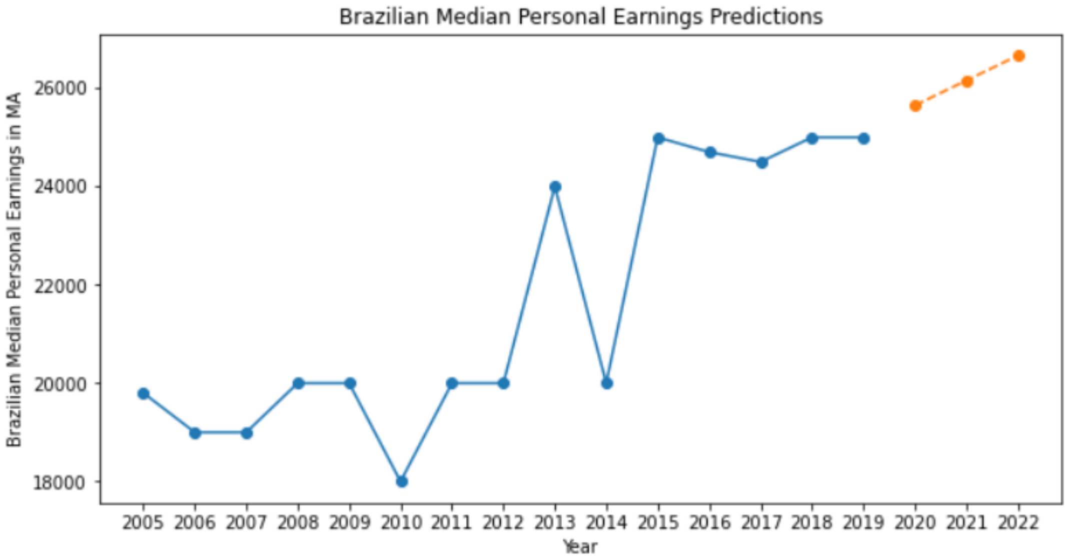
A similar correlation matrix was made for males. This can be found under the github and “Team 2” then “emilygreven” folder. There was a strong correlation between males in the construction and extraction occupations as well as the production, transportation and moving occupations. There is a strong negative correlation between males and the maintenance and repair occupations.



From the above coefficient matrix we can see some correlations between gross rent as a percentage of household income but in general these correlations are weak because they are spread across six sub classes of percentages of gross rent of Household income. We are proposing for Deliverable 3 to combine the gross rent percentages to three subclasses thereby allowing correlations to be seen more clearly.



By analyzing the correlation between Median Personal Earnings and other features, Citizenship, Year of Entry and Labored Rate seems to be mostly strongly correlated to personal income. By training a linear regression model using 15 years of data in Massachusetts, the model produced fine results. Contrary to previous conjecture, Ability to Speak English is not a major contributor to personal income.



Linear regression model is also used to predict Median Personal Earnings in the next 3 years. The results show that the personal income grows at a rate of about 1.9%.

Additional Key Questions

More than 1 question answered in explanations above as data was collected by the client.

Refine project scope and list of limitations with data and potential risks of achieving project goal

- In Household.xlsx Median Household income cannot be used because of invalid values
- Scope remains the same as deliverable 1 scope

5. Submit a PR with the above report and modifications to original proposal

- We have discussed with the client changes to the original requirements and received approval for our new plan.
- Based on the analysis we have performed in determining correlations and some linear regression models, we can use that to make more linear regression models and learn more about potential correlations in the data.