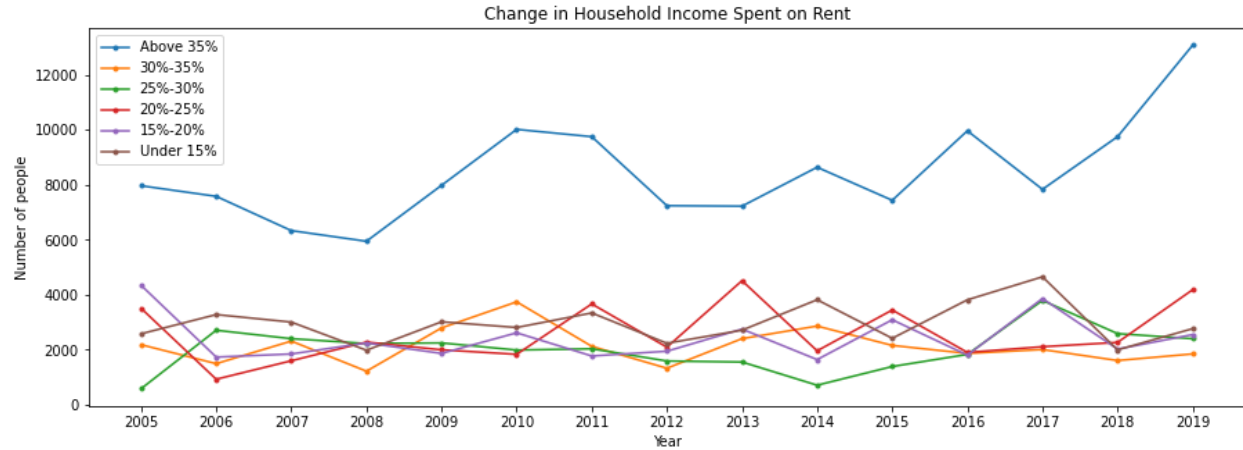


Team 2
Final Report
Draft 1

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

Visualizations

Visualizations - Finance



***See Appendix for large confusion matrices ***

More concise confusion matrices shown here that were formed from these larger matrices :

- Populations and Occupations and Gross Rent as a Percentage of Household Income*
- Populations and Occupations and Owner Costs as a Percentage of Household Income*

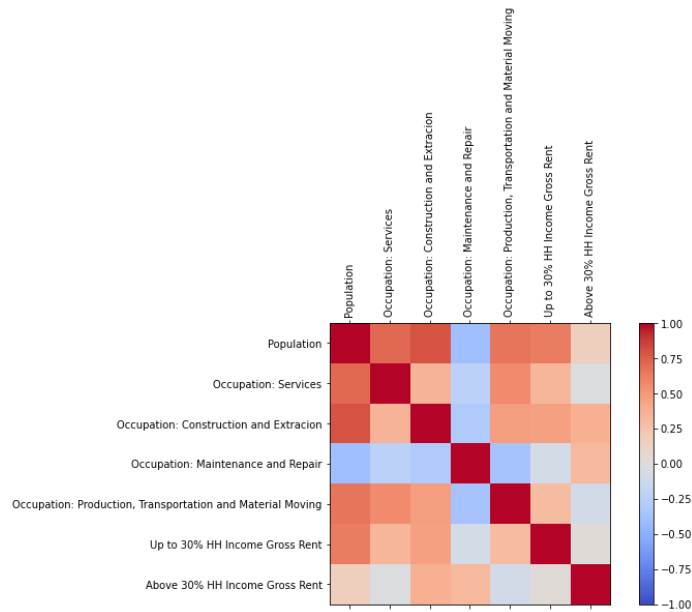
*See Appendix for large confusion matrices

Visualizations - Finance

Population and Occupation and Gross Rent as a Percentage of Household Income

Two Categories for rent:

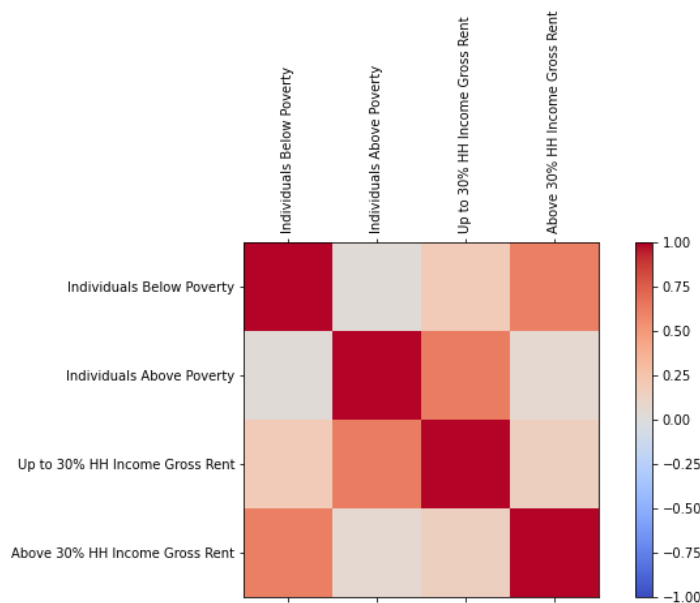
1. Up to 30% monthly household income
2. Above 30% monthly household income



Poverty Level and Gross Rent as a Percentage of Household Income

Two Categories for rent:

1. Up to 30% monthly household income
2. Above 30% monthly household income

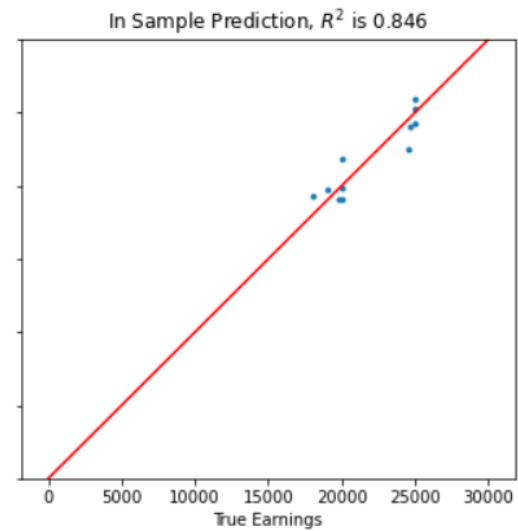
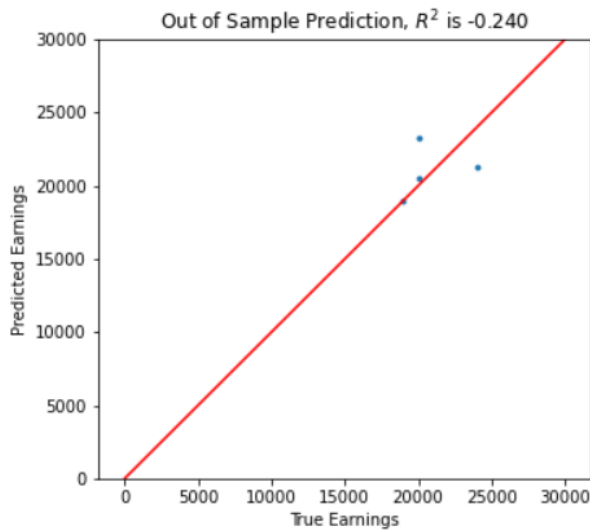


Visualizations - Finance

Median Personal Earnings



LinearRegression

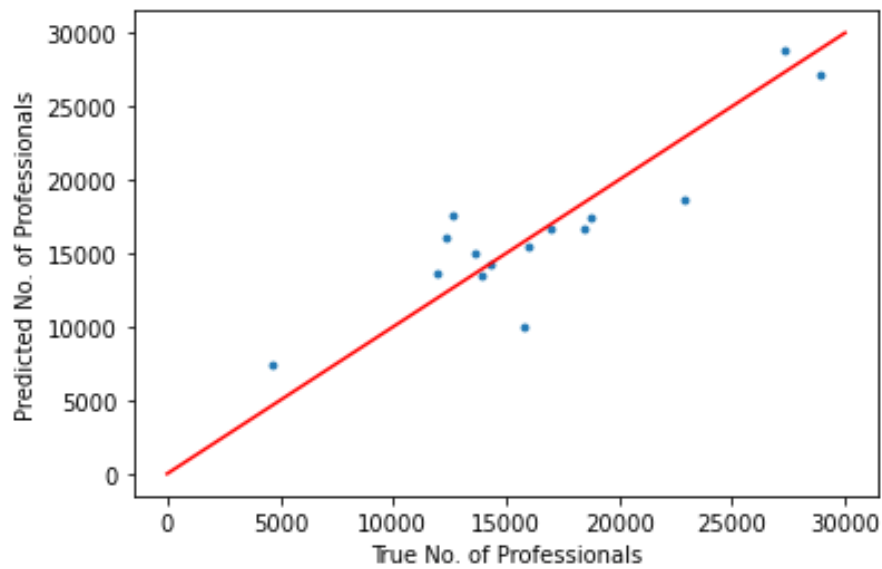


See Appendix for feature engineering

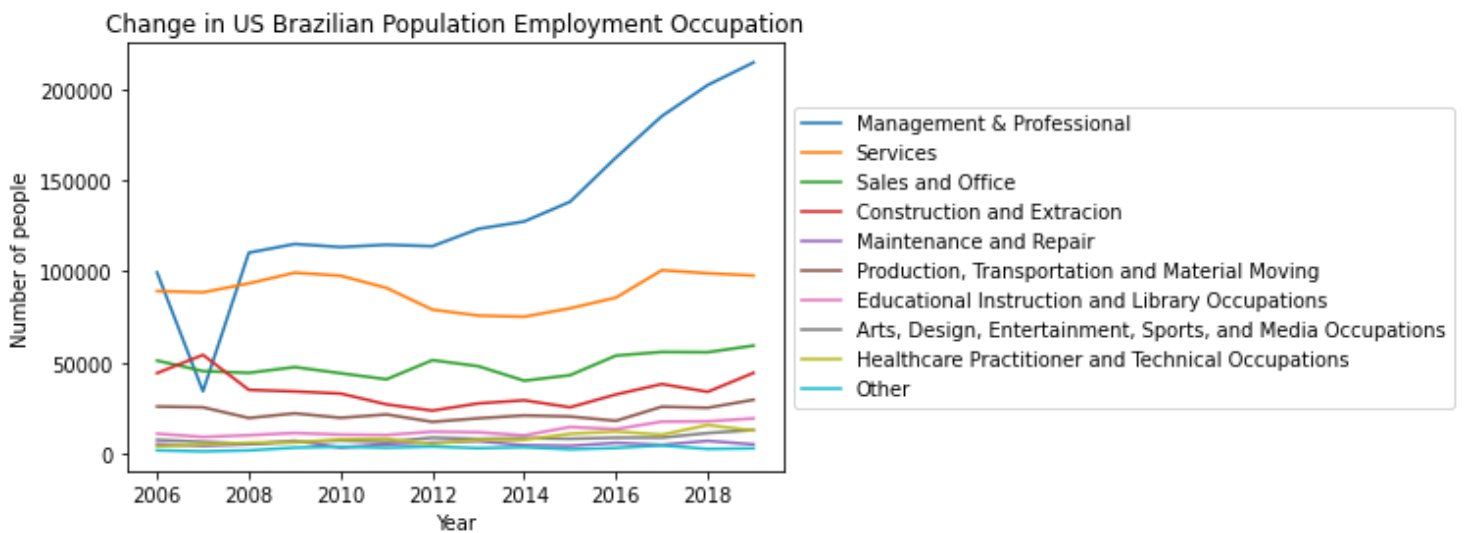
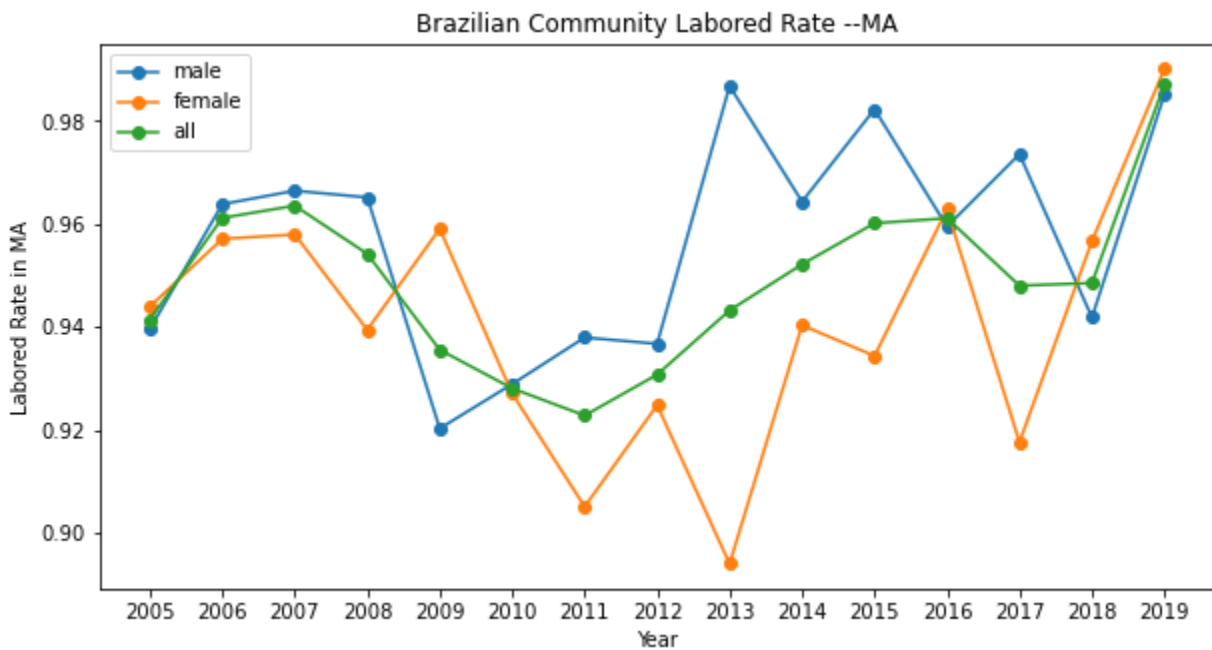
Visualizations - Education



Predictions for No. of Brazilian Professionals and Academics in MA, R^2 : 0.790

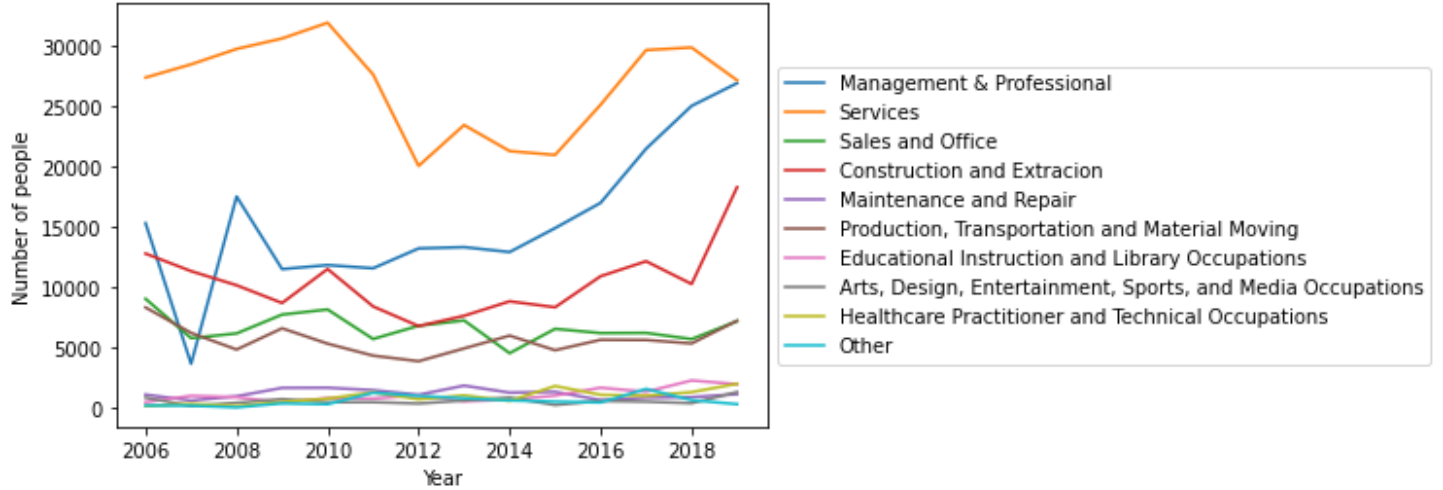


Visualizations - Professions

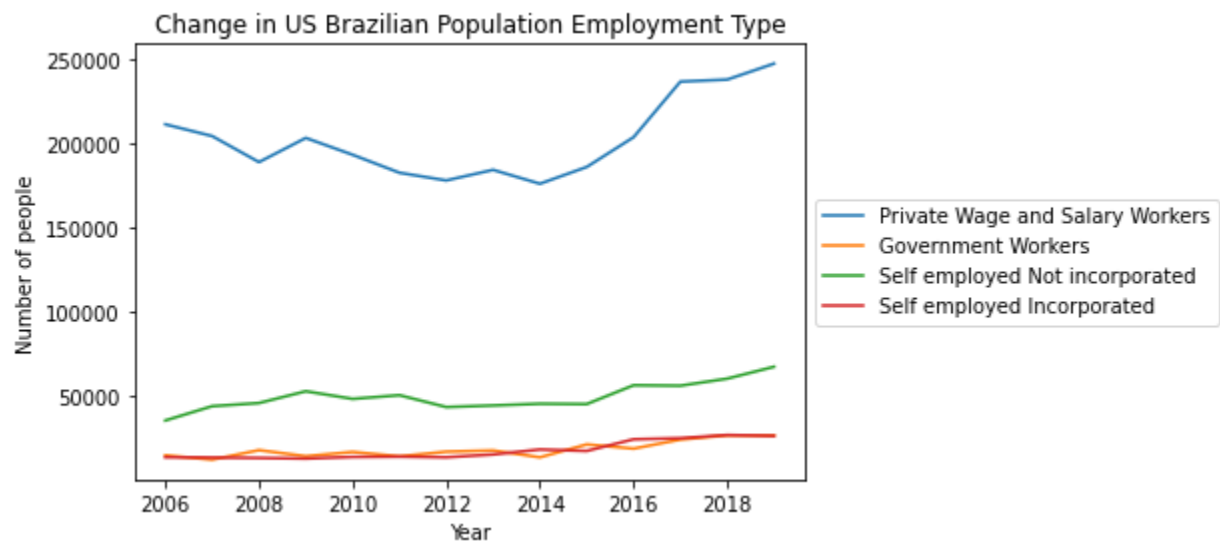
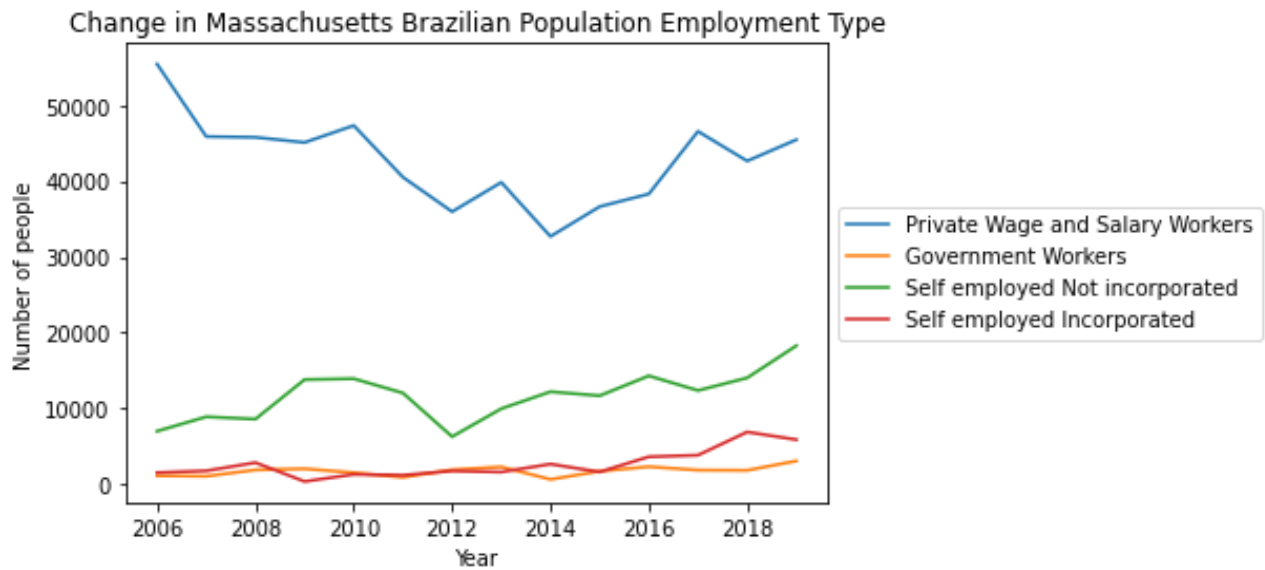


Visualizations - Professions

Change in Massachusetts Brazilian Population Employment Occupation



Visualizations - Professions



Visualizations - Professions

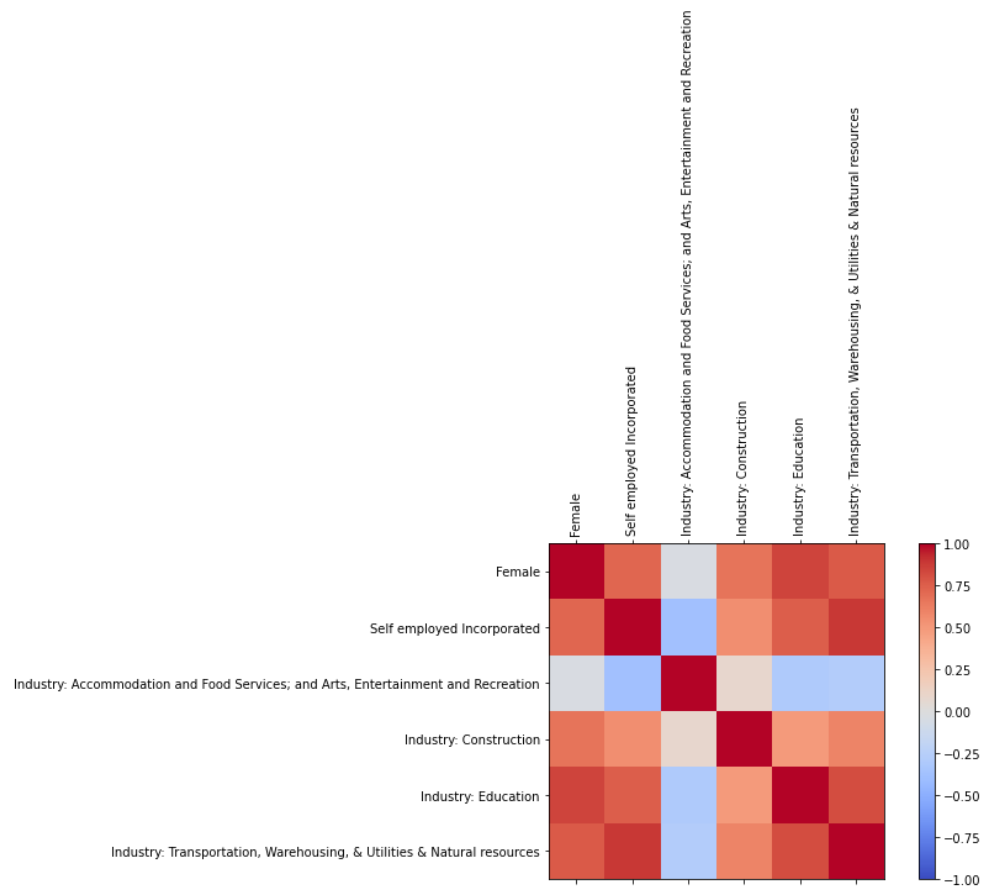
***See Appendix for large confusion matrices ***

More concise confusion matrices shown here that were formed from these larger matrices :

- Females and Industry*
- Females and Occupations*
- Males and Industry*
- Males and Occupations*
- Populations and Occupations and Gross Rent as a Percentage of Household Income*
- Populations and Occupations and Owner Costs as a Percentage of Household Income*

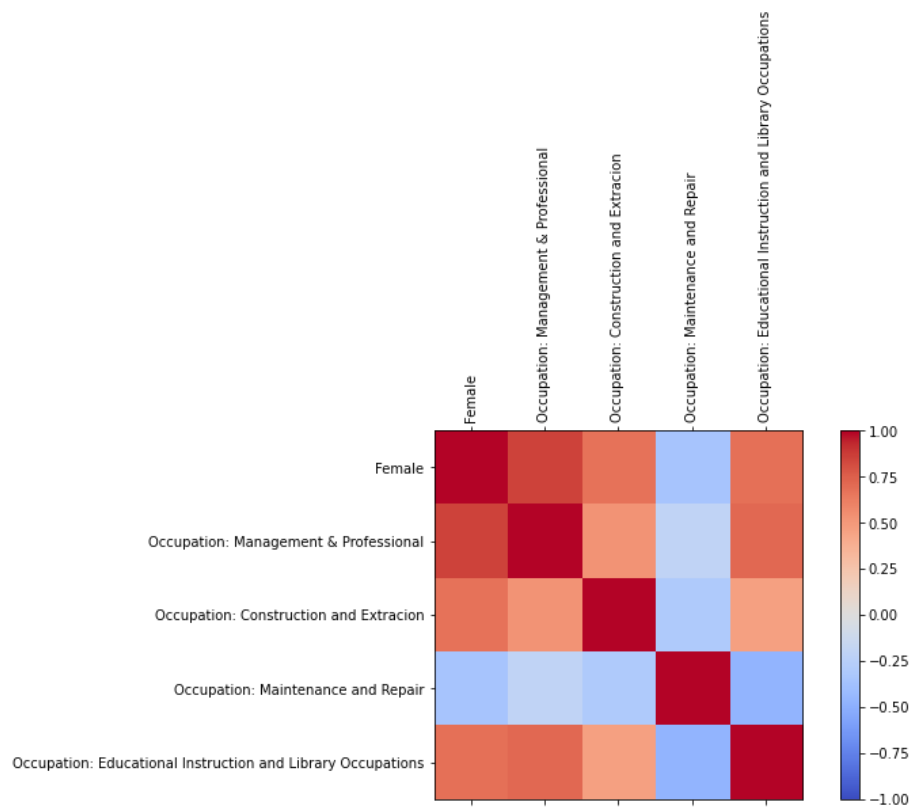
*See Appendix for large confusion matrices

Females and Industry



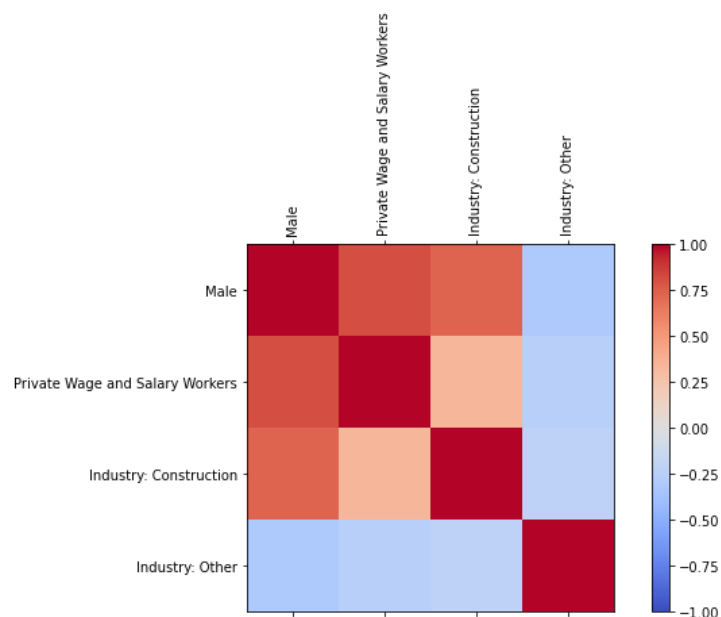
Visualizations - Professions

Females and Occupations

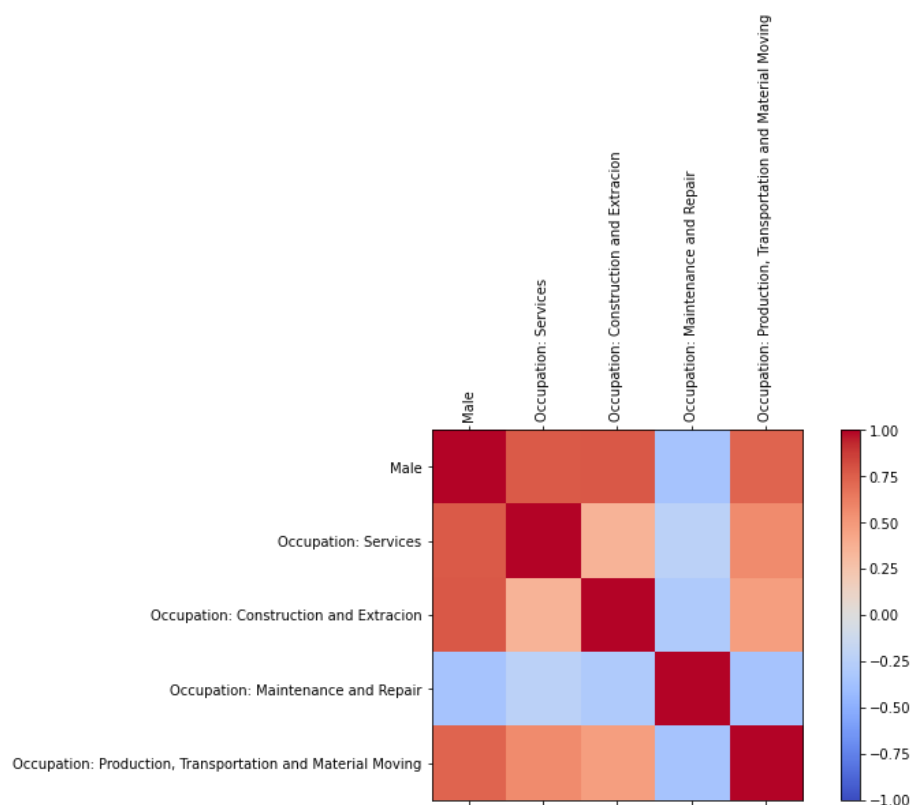


Visualizations - Professions

Males and Industries

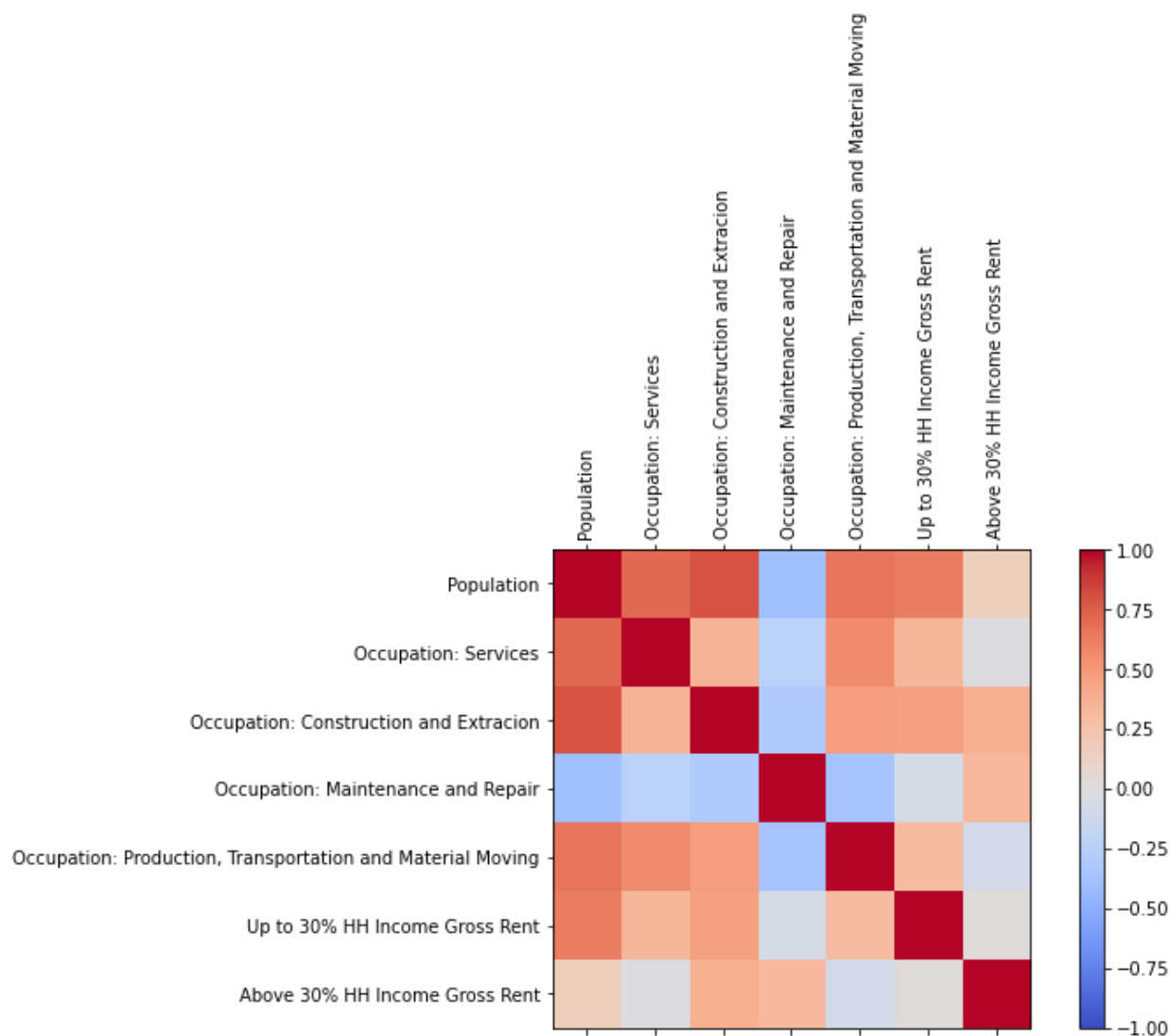


Males and Occupations

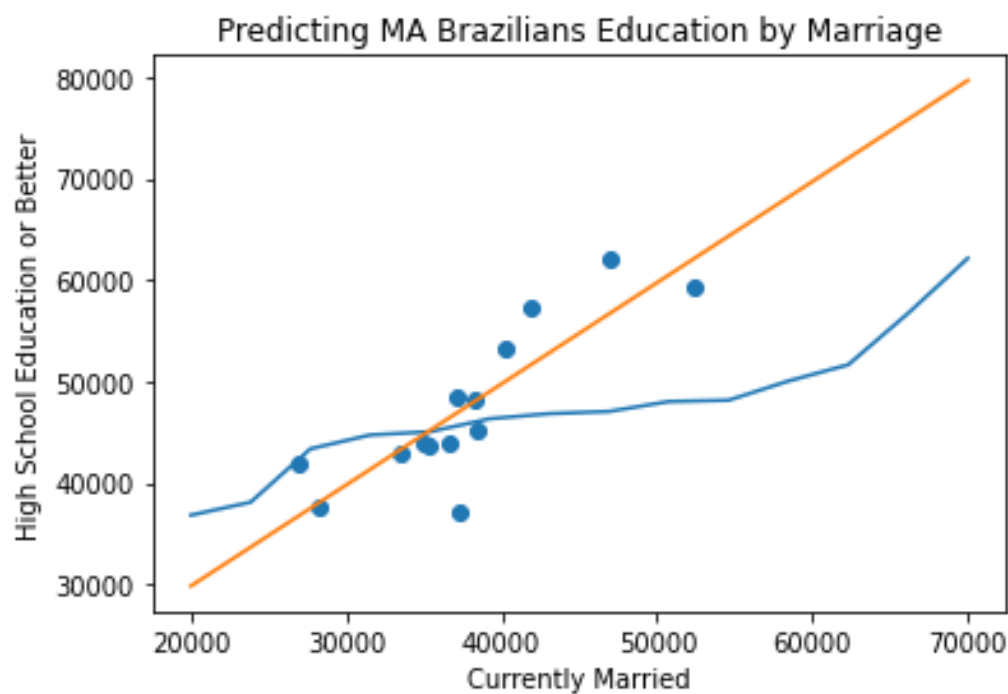
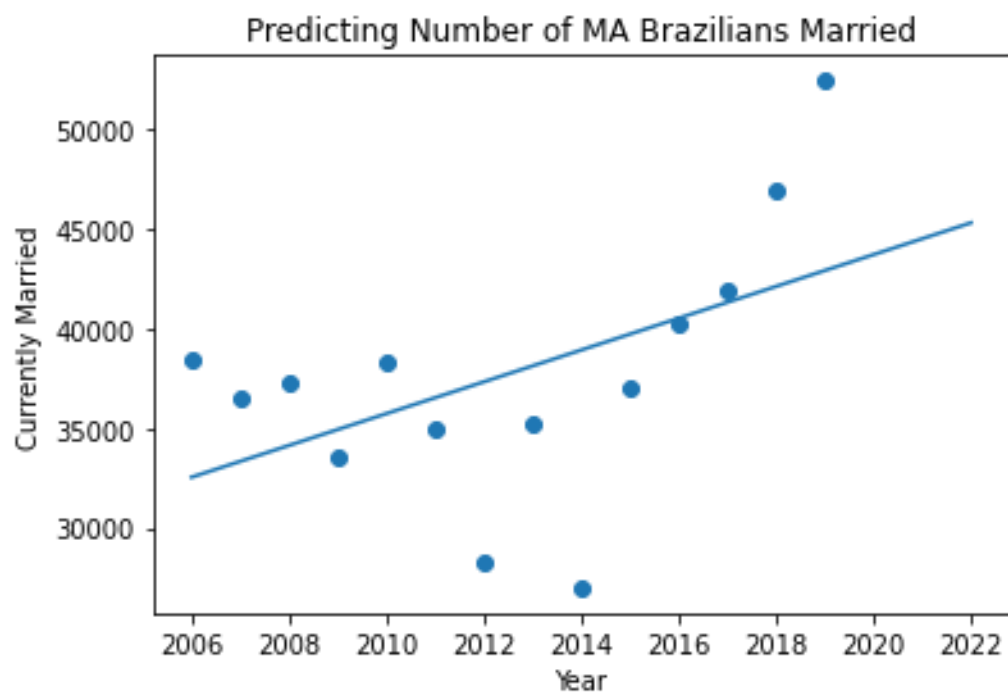


Visualizations - Professions

Population and Occupations and Gross Rent as a Percentage of Monthly Income

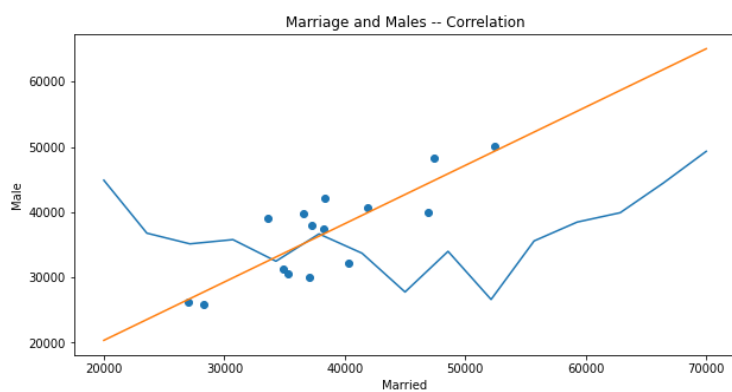
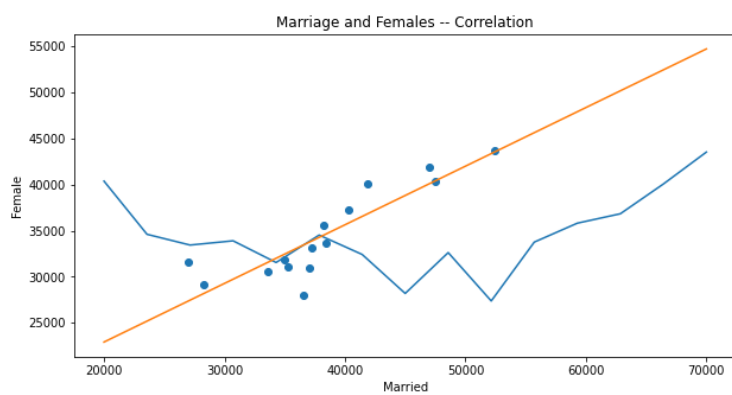
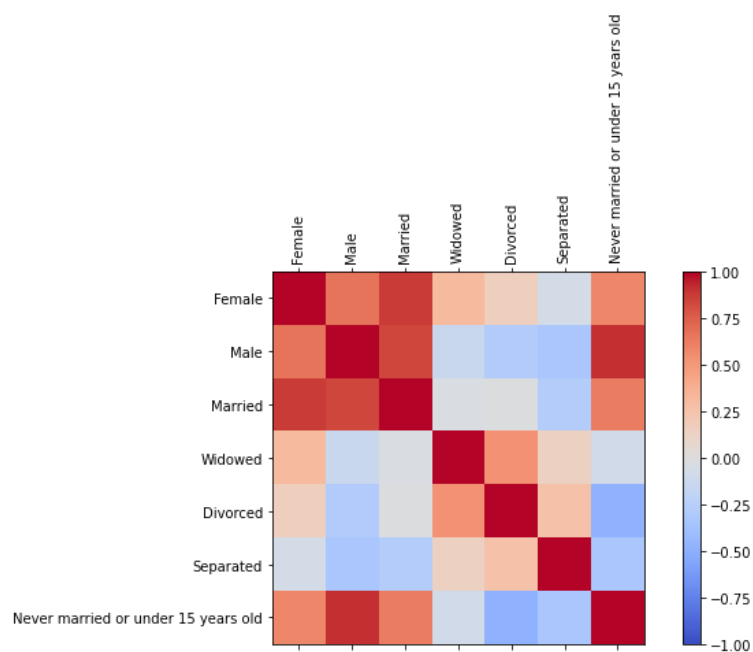


Visualizations - Marriage

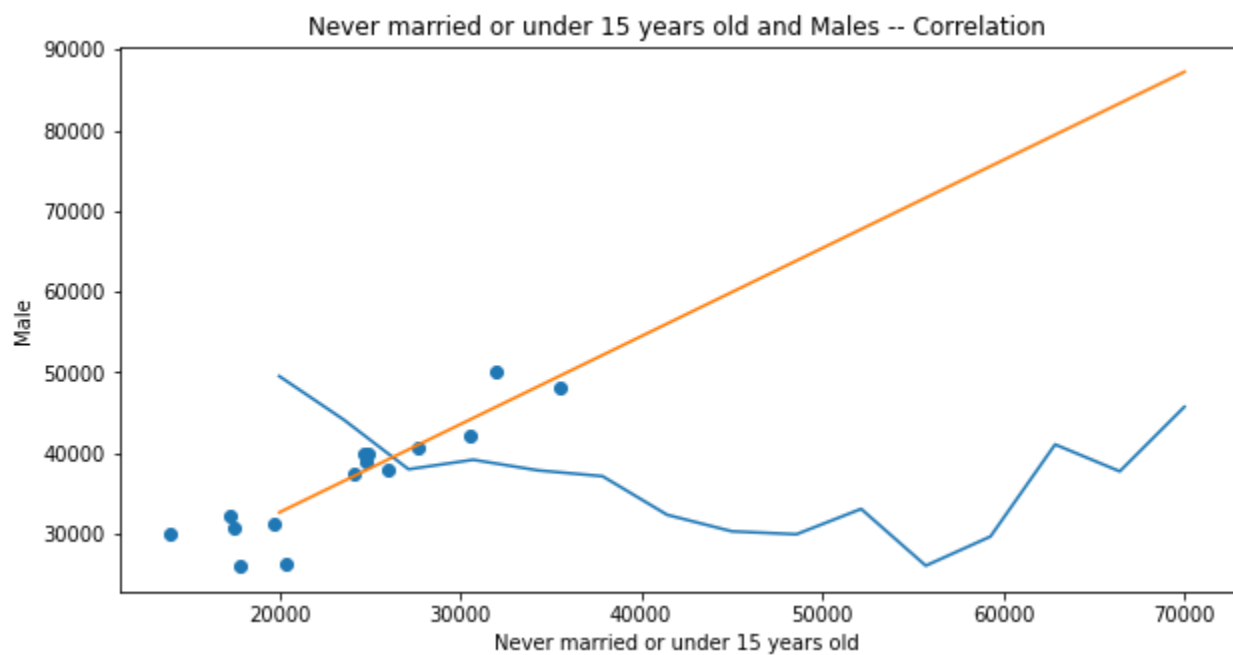
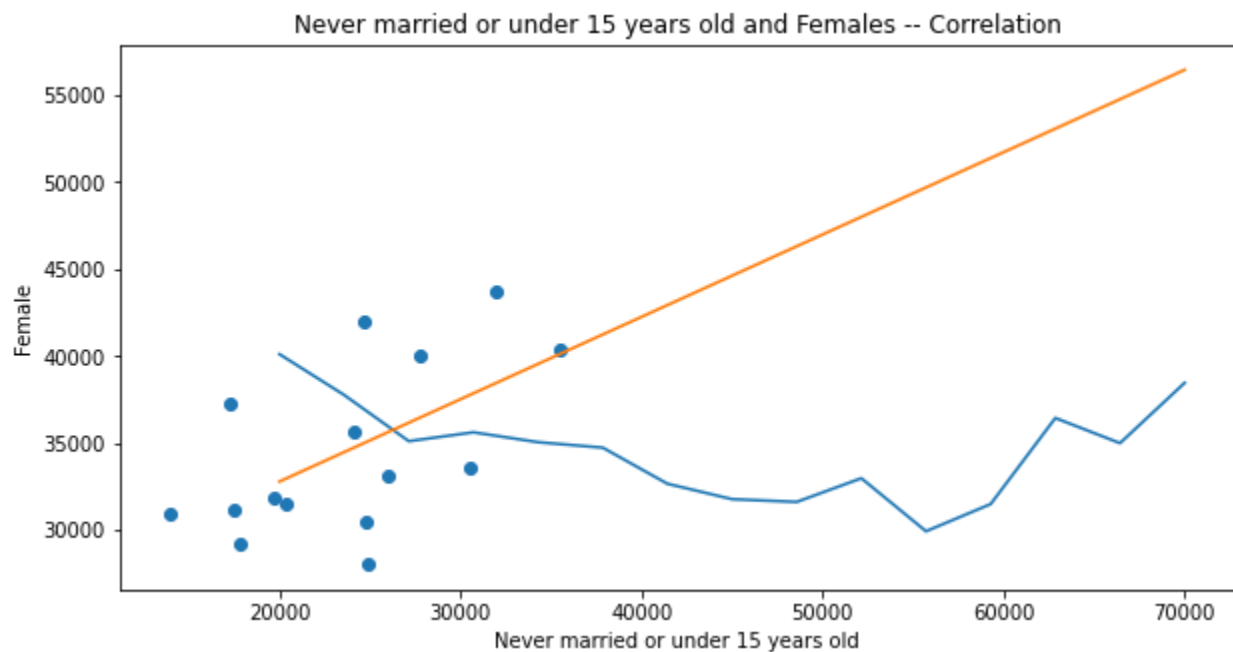


Visualizations - Marriage

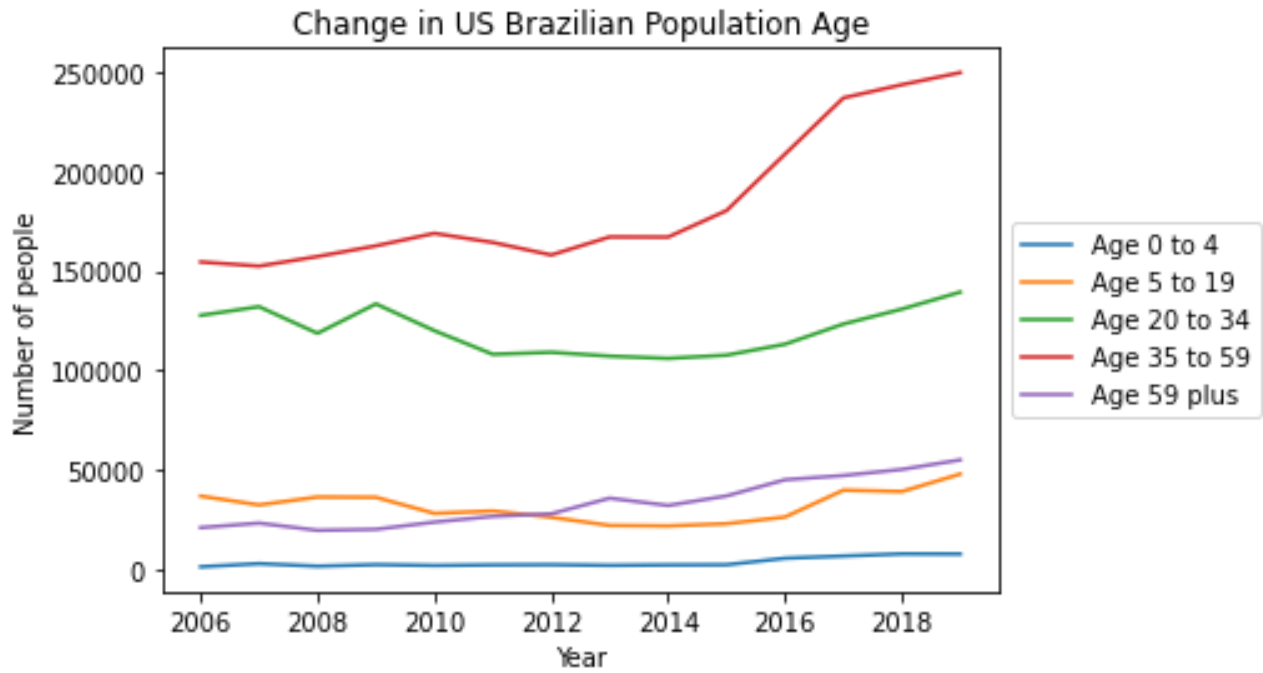
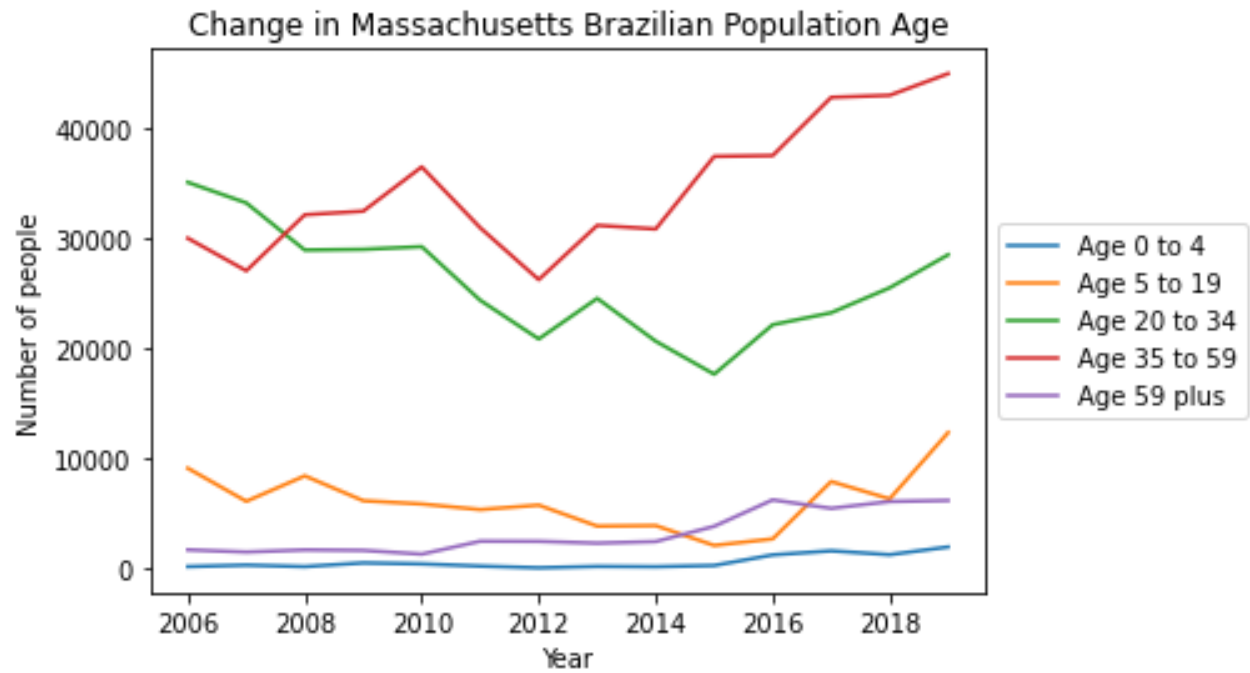
Females and Males and Marital Status



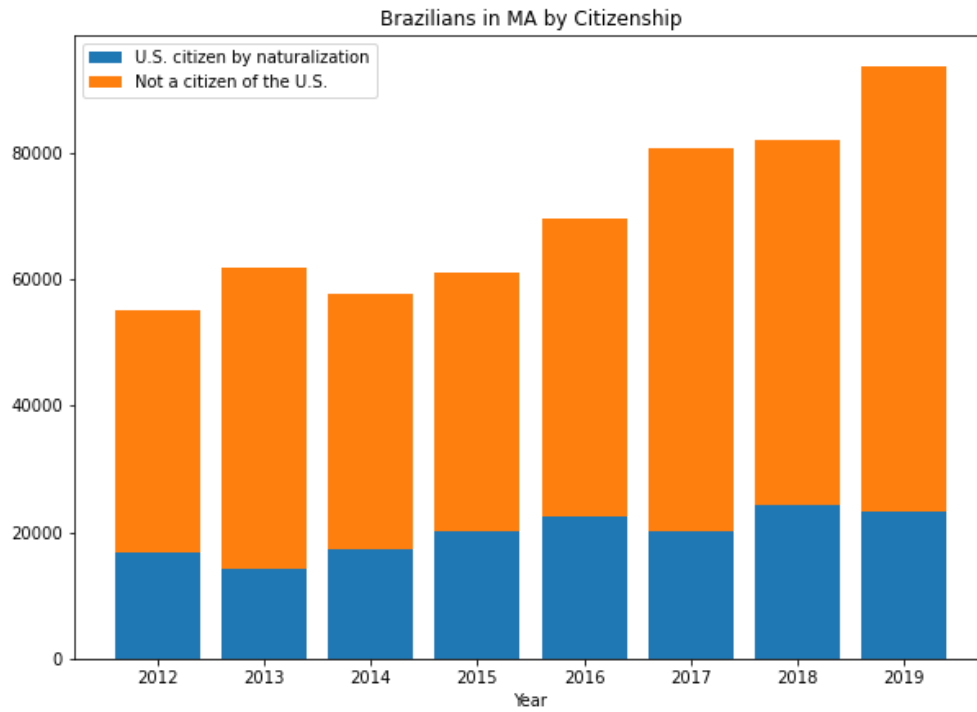
Visualizations - Marriage



Visualizations - Age



Visualizations - Age



Results

Analysis on median personal earnings

In general, personal income in Massachusetts is growing through the years. However, the Median Personal Earnings graph in Visualizations-Finance section shows that median income has dropped sharply in 2014 and rocketed in 2015. What happened in these 2 years is worth surveying.

To analyze the correlation between personal income and other features of the Brazilian community, some features are chosen from the group's extracted data, which contain English proficiency, Citizenship, Year of Entry, employment rate, poverty rate, and percentage of some occupations.

Employment by occupation was not included originally, but after plotting the top 5 occupations graph in the last section, the percentage of people working in Services is extracted as a feature because it is the most engaged industry. The percentage of people working in the tech and healthcare industry is another feature because it is the least engaged one.

For the labored rate, as shown in the Visualizations-Professions section, female and male employment rates are not consistent. Especially in 2013 and 2017, female employment rate drops when male employment rate rises. There may be an explanation for this, but for personal income analysis, employment rate for both genders is included as a feature for model training.

By trying different linear regression models, the conclusion is that: other than year of entry, citizenship and labored rate, personal income can be positively correlated with the proportion of the healthcare and technology industry, and negatively correlated with the proportion of the service industry. And services feature can be more effective as it is the industry most engaged in by the Brazilian community. Contrary to assumptions, English proficiency is not strongly correlated with personal income, indicating that ability to speak English is not very important for the Brazilian community's economy. However, citizens entering after the 2000s that work in the healthcare and technology industry seem to have higher earnings. The work is shown in the LinearRegression graph in Visualizations-Finance section.

Marital status and education comparison

As we can see from the graph labeled under "Visualizations - Marriage," "Predicting the Number of MA Brazilians Married," the number of married Brazilians in the state of MA is growing. The line of best fit created from linear regressions has the formula number of brazilians married = $797 \cdot \text{year} - 1565894$. This regression has an R^2 value of 0.3, indicating that the linear correlation is present, but not very strong. The slope of this line however is almost 800 indicating that the number of married Brazilians in MA is growing also relatively strongly. When we look

at the scatter plot of the points, it's very clear that while linear regression may not be the perfect model to fit this data, there is still some very clear growth throughout the 14 years of data.

With this in mind, the next step was to look for other non trivial attributes that had correlation with the number of marriages. One interesting correlation that was found is with the education rate within and the number of marriages. While a high school degree had strong correlation with marriages, there was a much stronger correlation found when aggregating everyone who had a high school degree or better. In this case, we proceeded with the aggregated data and did a side by side linear comparison. The line of best fit created from linear regressions has the formula $y = 9900 + X$ rounded to the nearest whole number. This regression has an R^2 value of 0.72, indicating that the linear correlation is very strong. The slope of this line is 1 indicating that the number of married Brazilians in MA is growing also relatively strongly. Here we can see that as the number of MA Brazilians currently married grows, so does the number of MA Brazilians who have a high school education or better. This could indicate that Brazilians who are educated have a higher sense of responsibility and as such decide to enter into a formal marriage with a partner.

Effect of English Proficiency on Professions:

According to the correlation heat map under “Visualizations: Education”, there is a strong correlation between the number of Brazilians in Massachusetts who can speak English proficiently and the number of Brazilians in Massachusetts who have occupations categorized as either “Management & Professional” or as “Education Instruction and Library Occupations”. We have combined these occupation categories together as “Professional & Academic”. This suggests that Brazilians who are fluent in English are more likely to obtain a Professional or Academic profession.

Additionally, we have determined that the number of Brazilians with good English proficiency was a good predictor of the number of Brazilians working in Professional or Academic professions in Massachusetts. We have performed a linear regression on this data, which resulted in an R^2 value of 0.79, which indicates that the linear correlation between these two categories is strong and confirms the suggestion stated above.

Analysis on Professions:

From the graph under “Visualizations - Professions” entitled “Brazilian Community Labored Rate -- MA” we can see that in Massachusetts, a longstanding majority of Brazillians worked in service jobs, however, Management & Professional jobs (M&P) grew slowly in popularity and eventually converged with service jobs in 2019. On the country wide scale this happened back in 2008, but it is clear that MA Brazilians are catching up to the trend.

From the confusion matrices, we can see that the strongest positive correlations for females in industries is with Education and there is a negative correlation of females in the Accommodation and Food Services; and Arts, Entertainment and Recreation industries. For

occupations there is a strong correlation between females and Management and Professional occupations. There is a negative correlation between females and Maintenance and Repair occupations. For men, there is a strong correlation between males in construction and private wage and salary workers and a strong negative correlation of males and other types of Industry. It may be of interest to find out what was the most popular “Other” industry that was reported for males that caused a strong negative correlation. For male occupations, there is a positive correlation between males and services, construction and extraction, and production, transportation, and material moving. But the confusion matrix shows a negative correlation between men and the maintenance and repair industry (this trend of negative correlation is the same as the female trend). This observation is consistent with the confusion matrix for population and occupations where there is a strong negative correlation with Maintenance and Repair occupations and a strong positive correlation with Construction and Extraction occupations.

Analysis on Gross Rent:

We can see from graphs under “Visualizations - Finance” “Change in Household Income Spent on Rent” that the most populated category of percentage brackets is the above 35% monthly household income. From the confusion matrix we can see a strong correlation between the number of people spending more than 30% monthly household income on rent with the number of Individuals below the poverty line. The confusion matrix also shows that there is a strong correlation between individuals who spend up to 30% of their monthly income on rent and the number of Individuals above the poverty line. This leads to the conclusion that in order for more people to move above the poverty line there must be a change in how much of their monthly income is spent on rent.

Analysis on Age:

We can see that both nationwide and within MA only, the majority of Brazilians are aged 35-59. One interesting observation is that in MA only there are more people aged 5-19 than there are 59+. However, on the country wide scale, there are more people aged 59+ than there are aged 5-19. In addition while these distributions don’t change too much over the years, they are all growing further showing population growth in general.

Data

See github Team 2 branch -> folder “data” -> two excel sheets

Code

See github -> Team 2 branch -> folder

“Age_EmploymentByType_EmploymentByOccupation-Analysis”

-> “**jamesh_deliverable1.ipynb**”

-> “**jamesh_deliverable2.ipynb**”

See github -> Team 2 branch -> folder “emilygreven”

-> “**CensusProjectDeliverable1.ipynb**”

-> “**CensusProjectDeliverable2.ipynb**”

-> “**CensusProjectDeliverable3.ipynb**”

See github -> Team 2 Branch ->

folder “preprocessing_Qingyang” ->

-> “**PredictionModel.ipynb**”

-> “**DataAnalysis.ipynb**”

-> “**DataExtraction.ipynb**”

See github -> Team 2 Branch -> folder “yichuanma95”->

folder “notebooks”

-> “**Citizenship.ipynb**”

-> “**correlations.ipynb**”

-> “**EducationalAttainment.ipynb**”

-> “**English-Proficiency.ipynb**”

-> “**extract_data.py**”

-> “**Marital-Status.ipynb**”

-> “**PovertyStatus.ipynb**”

-> “**Year-of-Entry.ipynb**”

Folder “preprocessing”:

-> “**Population-2019.ipynb**”

folder “YearOfEntry” -> folder “.ipynb_checkpoints”

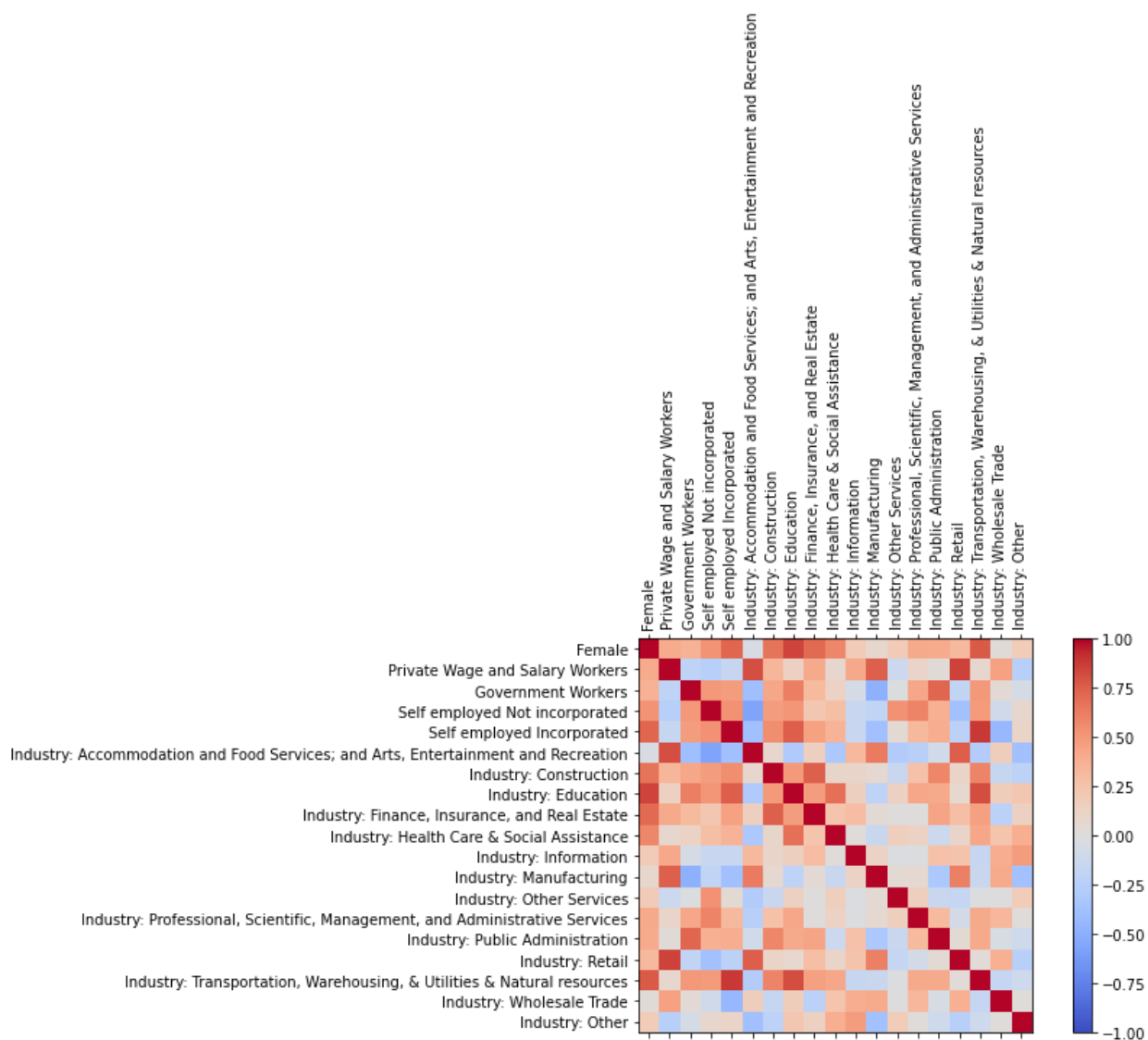
-> “**Year-of-Entry-checkpoint.ipynb**”

Documentation

See Team 2 ReadMe.md

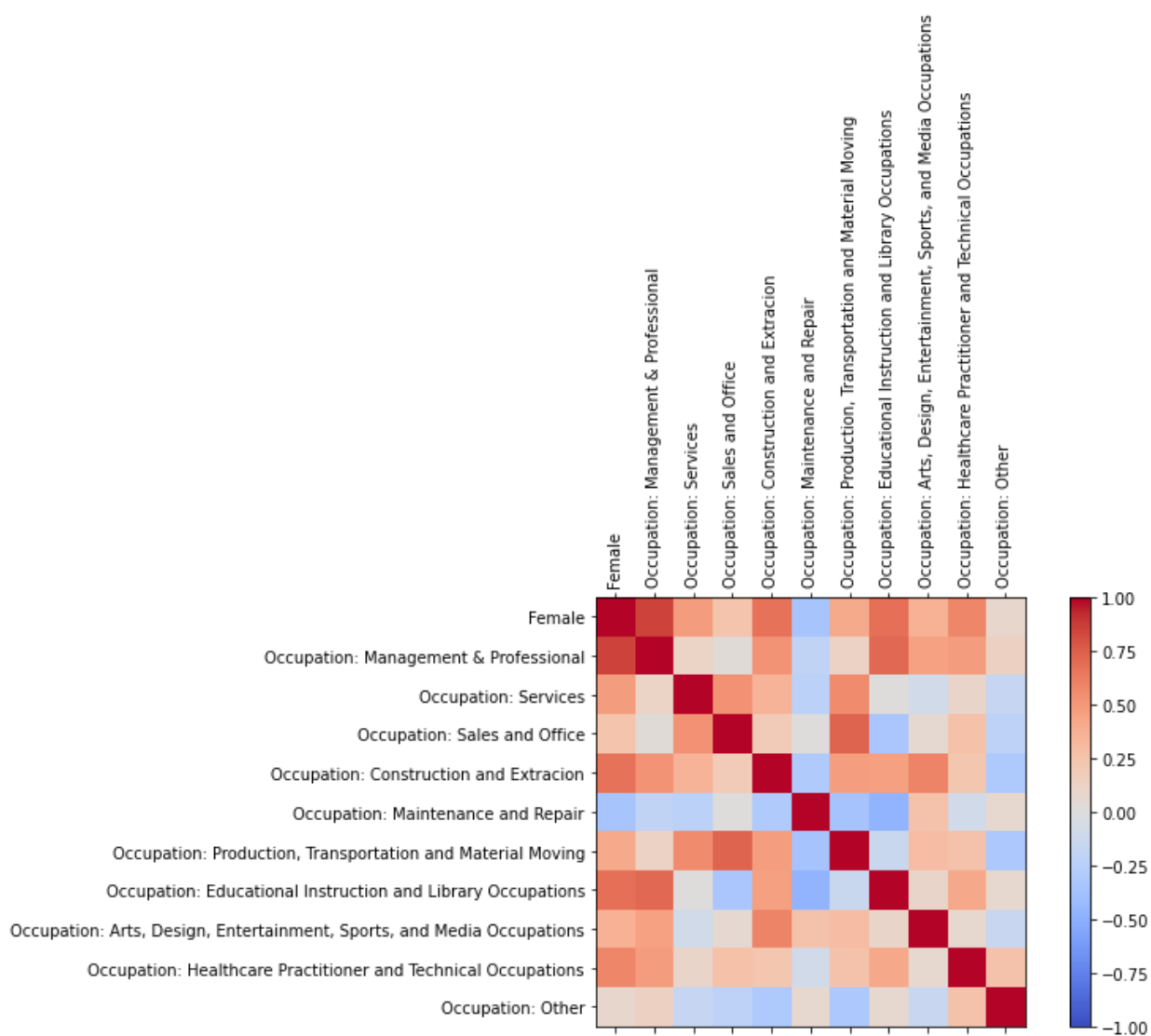
See Team 2 branch -> folder “Preprocessing_Qingyang” -> “README.md”

Females and Industries



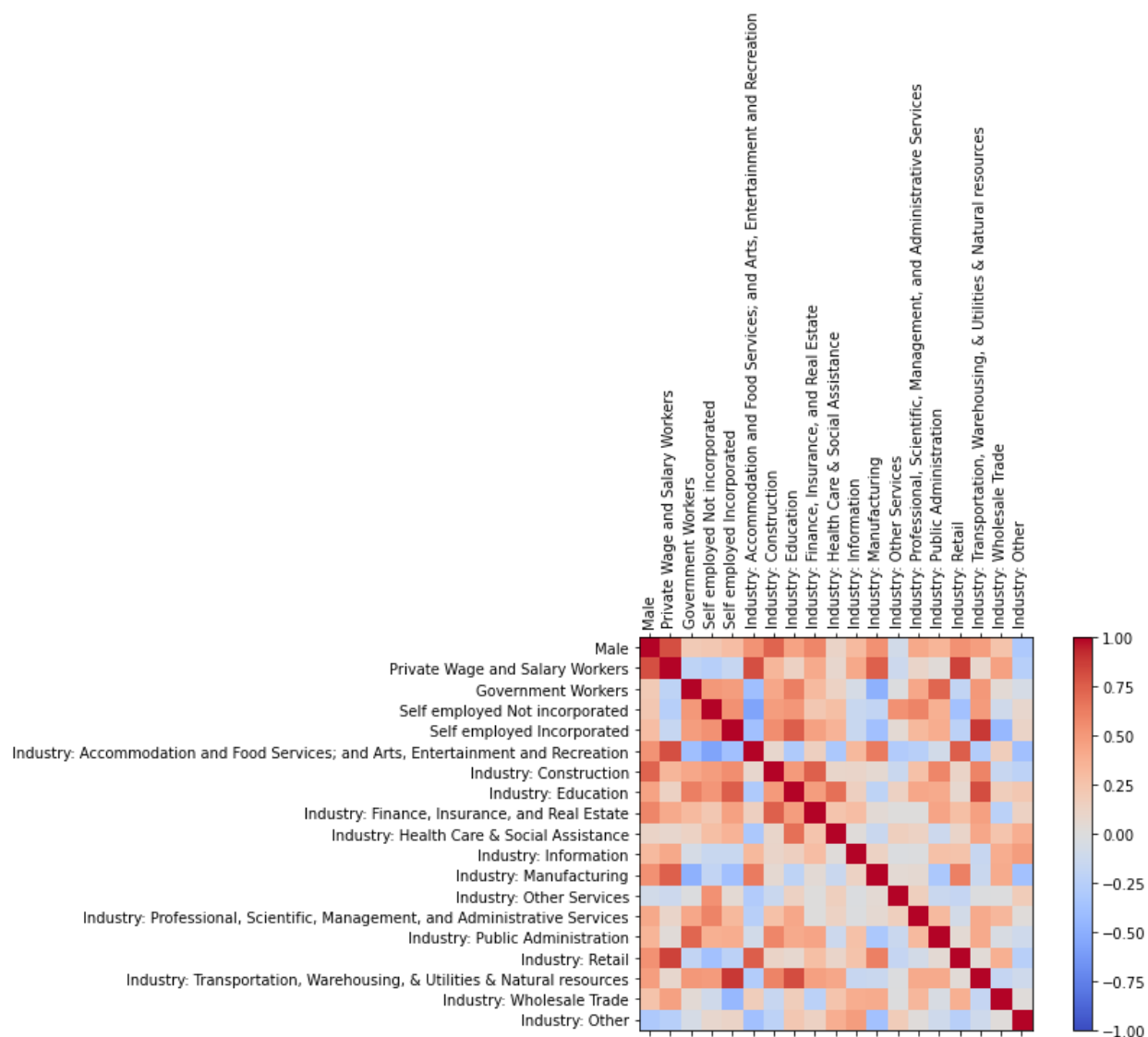
Confusion Matrices:

Females and Occupations



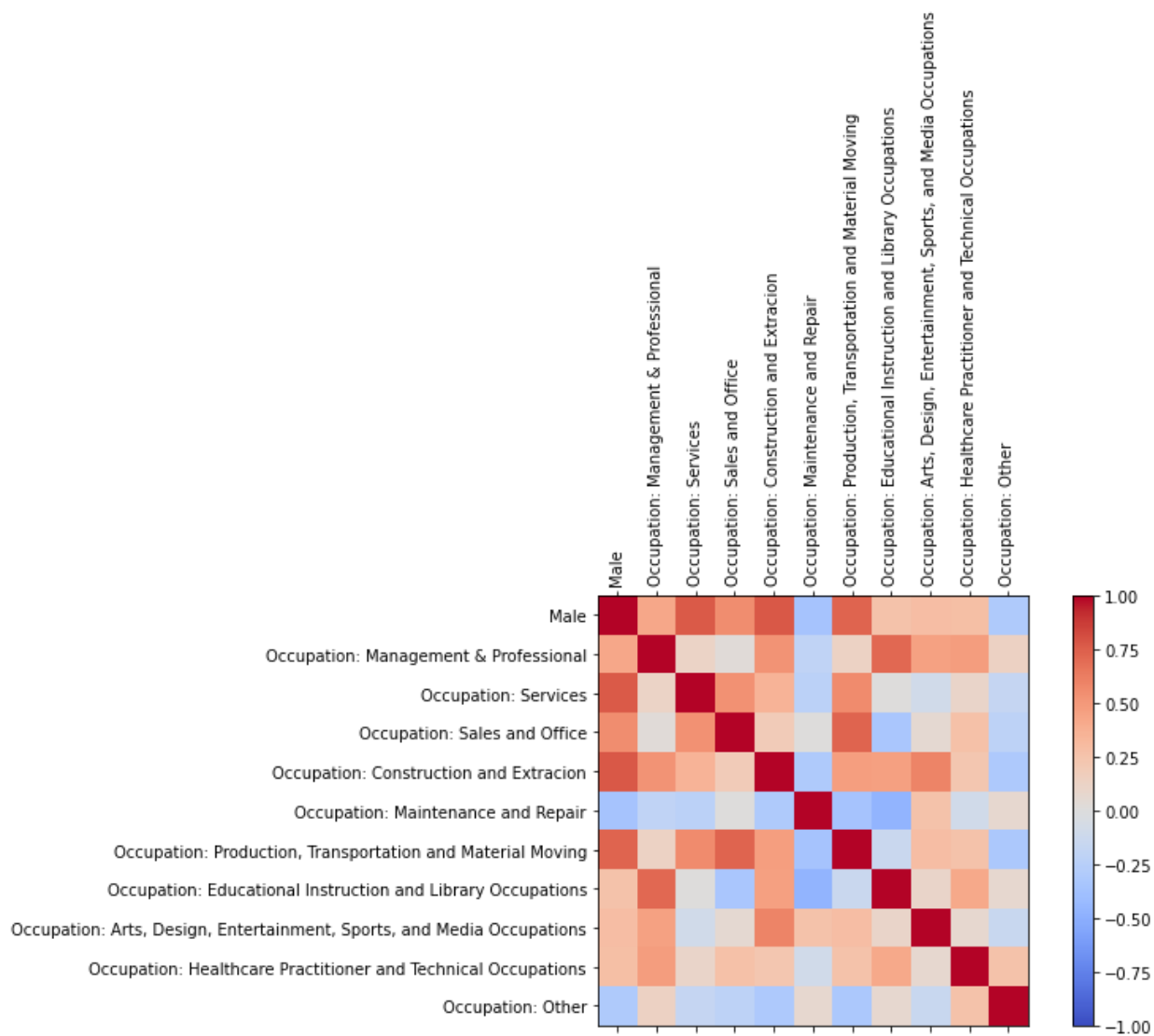
Confusion Matrices:

Males and Industry



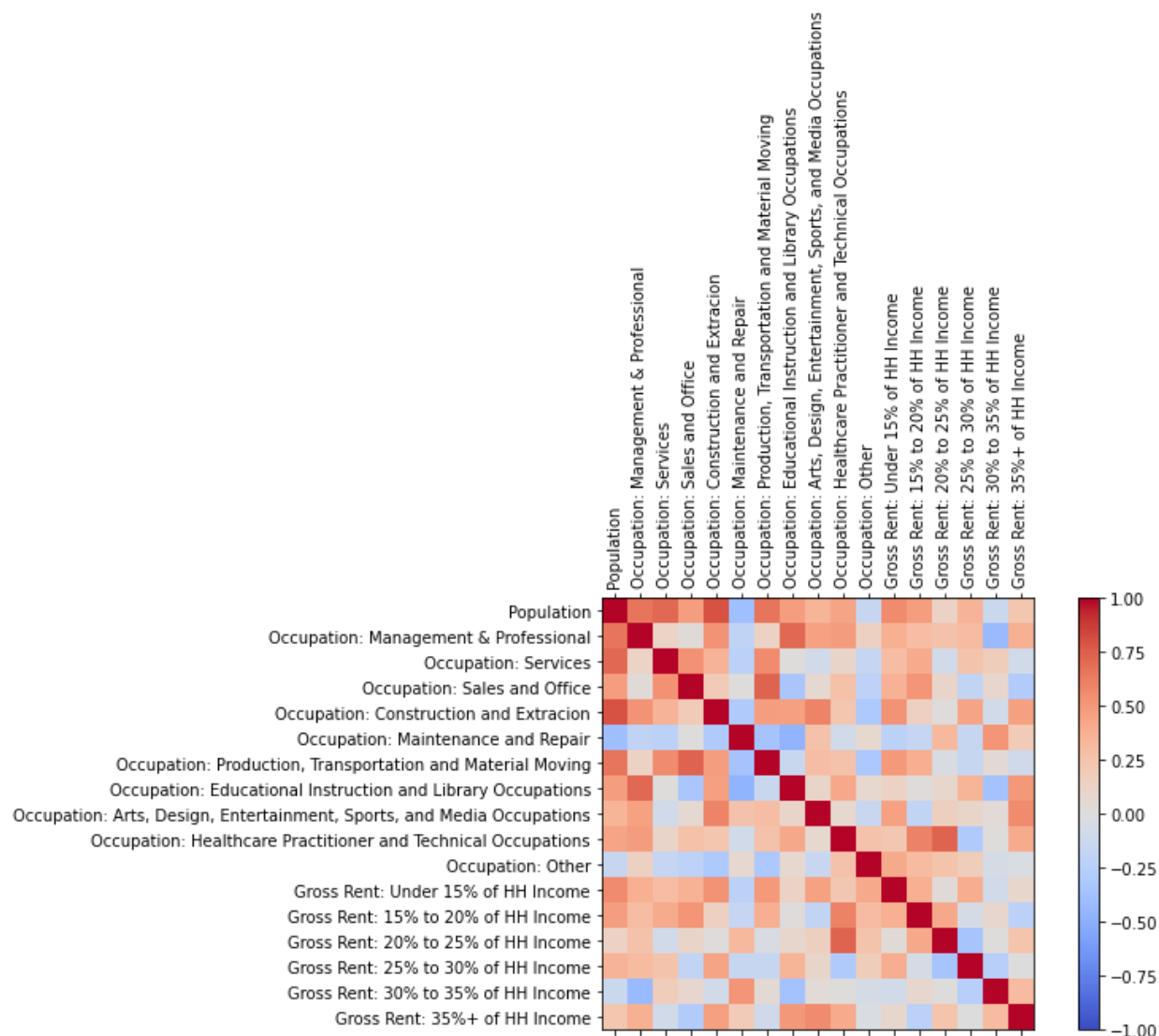
Confusion Matrices:

Males and Occupations



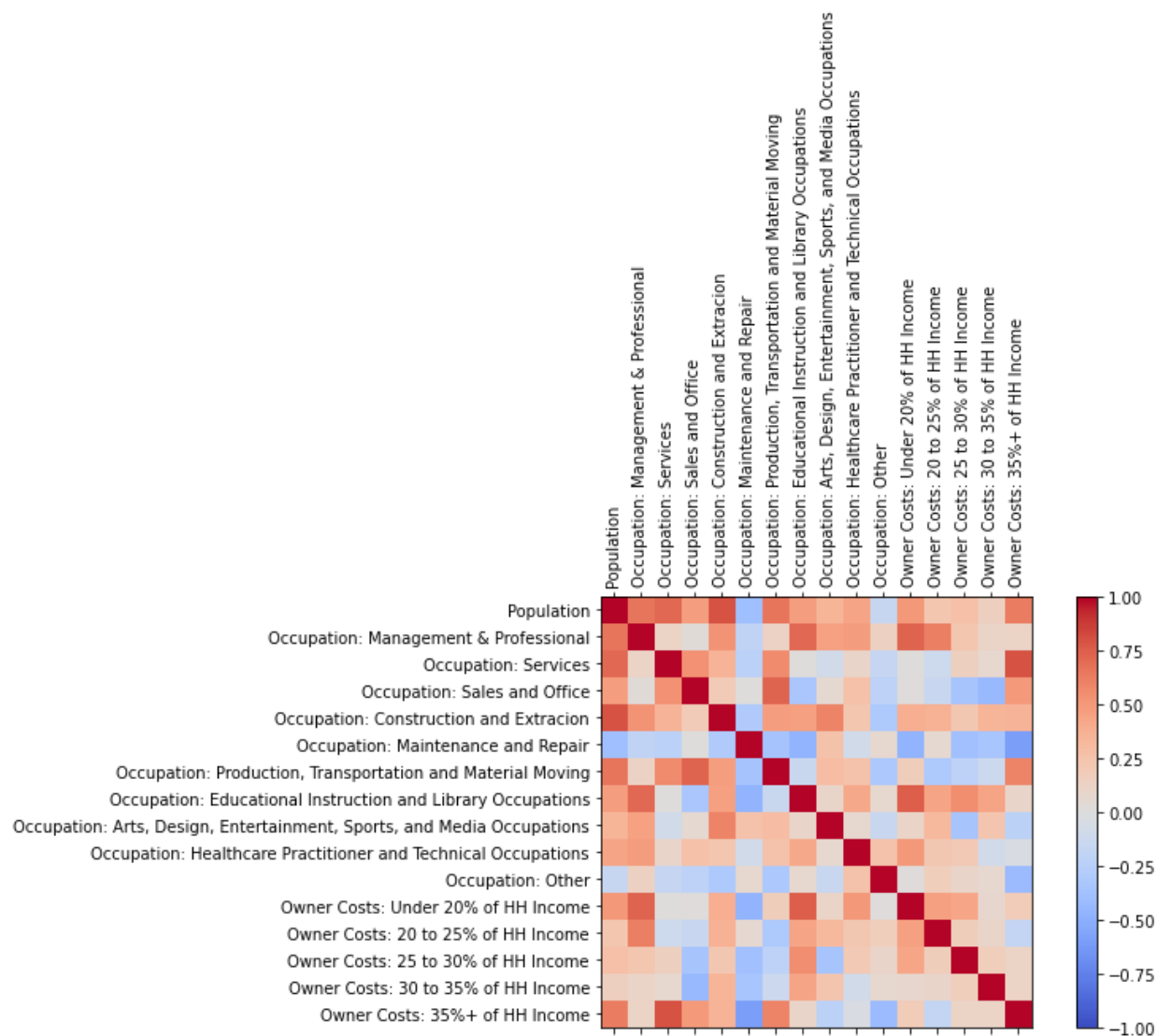
Confusion Matrices:

Populations and Occupations and Gross Rent Percentage of Household Income



Confusion Matrices:

Populations and Occupations and Owner Costs as a Percentage of Household Income



Top 5 Employment by Occupation in MA

