Final Paper

STOR 320.(01) Group 1

November 21, 2022

INTRODUCTION

Money! The currency that makes the world go round. In American modern society, almost every aspect of living costs money, but we're often desensitized to our money centered culture. In fact, it's almost hard to believe that we can't do anything without spending at least some kind of money. Driving somewhere? Staying in your apartment? Going out to eat? Cooking at home? All these things have a cost associated with them, making money a necessity in our ever changing and growing world. With the importance of money in mind, our analysis focuses on what factors may affect our ability to make money.

Many factors that are considered in this analysis are beyond our individual control, such as race, age, disability, and language spoken. In our initial observations of the data, we were interested to see how these individual factors were related to income, and moved towards looking for a more holistic model to most effectively predict one's income based on given variables. We were also interested in additionally investigating the proportion of the upper income class within groups in order to make observations about societal stereotypes such as the model minority belief and gender roles play out in today's society. Our questions are as follows.

Question 1: Accounting for multicollinearity, what predictors make up the best model to predict income?

Question 2: Are there significant differences in the proportions of the high income class within the different races, education levels and genders?

In our American capitalistic economy, it's important to study the variables that have an effect on an individual's income because it allows us to pinpoint factors that are acting as an unfair disadvantage to some marginalized groups of people. Studies into this topic of income will allow those with power to funnel more funding into providing resources for groups that may be suffering financially due to societal burdens in order to promote equity amongst residents of the United States. Interestingly, in a truly equitable society, none of our explanatory variables should have a significant relationship with income. We wouldn't be able to predict income using these factors in our data set because all groups of people would have a fair chance for any income level, thus erasing the relationship that we currently see between the variables. Our findings emphasize the powerful inequities surrounding money that exist between Americans stemming from past systems of oppression and societal norms.

DATA

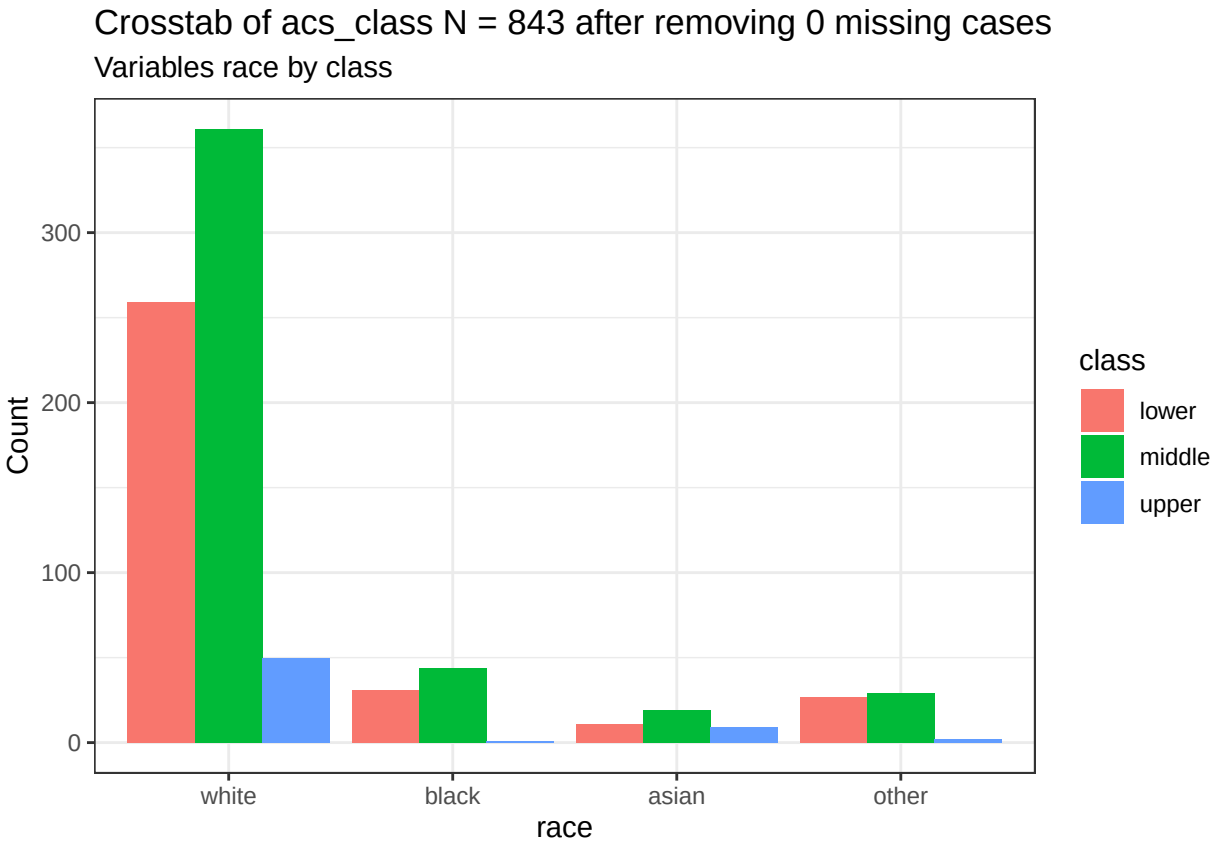
Our data was originally collected by the US Census American Community Survey in 2012 and retrieved from OpenIntro. The purpose of the American Community Survey is to help local officials and leaders examine changes in their communities, and provide detailed data about the nation. The data is available publicly and includes a variety of information about people and housing in every community. Our data set originally had 2000 observations and 13 variables, which should be representative of American communities in 2012.

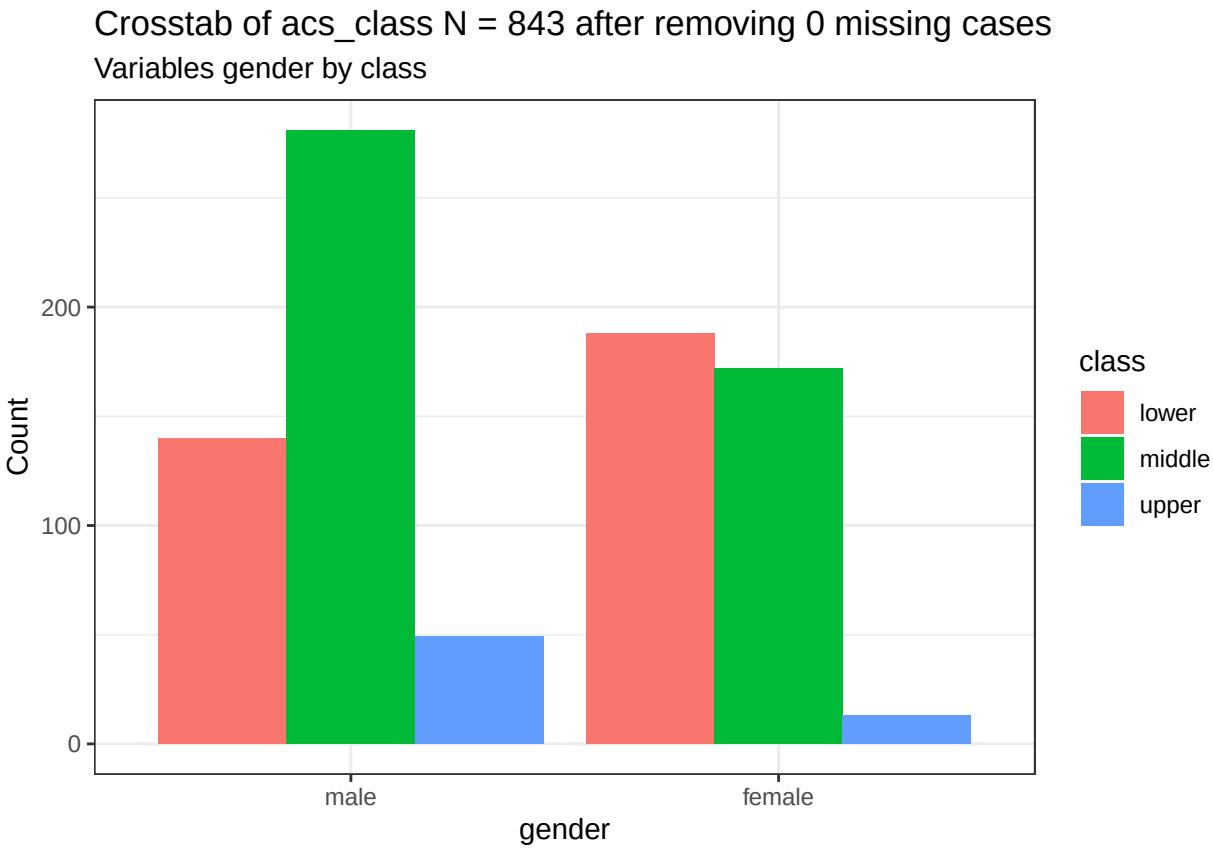
Since we were interested in the factors that contribute to income, we only kept individuals that were employed, resulting in a data set with 843 observations. While this may appear to be a substantial loss in the amount of data, many of those that were omitted were unemployed or not in the labor force. We also eliminated the time_to_work variable due to the significant amount of missing data.

income	employment	hrs_work	race	age	gender	citizen	lang	married	edu	disability	birth_qtr	birth_qtr
1700	employed	40	other	35	female	yes	other	yes	hs or lower	yes	jul thru sep	3
45000	employed	84	white	27	male	yes	english	yes	hs or lower	no	oct thru dec	4
8600	employed	23	white	69	female	yes	english	no	hs or lower	no	jul thru sep	3
33500	employed	55	white	52	male	yes	english	yes	hs or lower	no	apr thru jun	2
4000	employed	8	white	67	female	yes	english	yes	hs or lower	no	apr thru jun	2
19000	employed	35	white	36	female	yes	english	yes	college	no	jul thru sep	3

Above is a snapshot of our data set after data cleaning.

The response variable that we investigated was income, which is in USD. Other variables of interest included the number of hours worked a week, race, age, gender, citizenship status, language spoken at home, and education level, which were used to build the predictive model for income. Additionally, we created a variable that represents income class, based on data from 2012, which classified middle class as those with incomes above $20,592 and below $104,087, which we later used in our investigation of differences between proportions of income classes amongst different races, genders, and education levels.





From an initial observation of the graphs above, one can see stark differences in the proportions of each income class across various demographics.

RESULTS

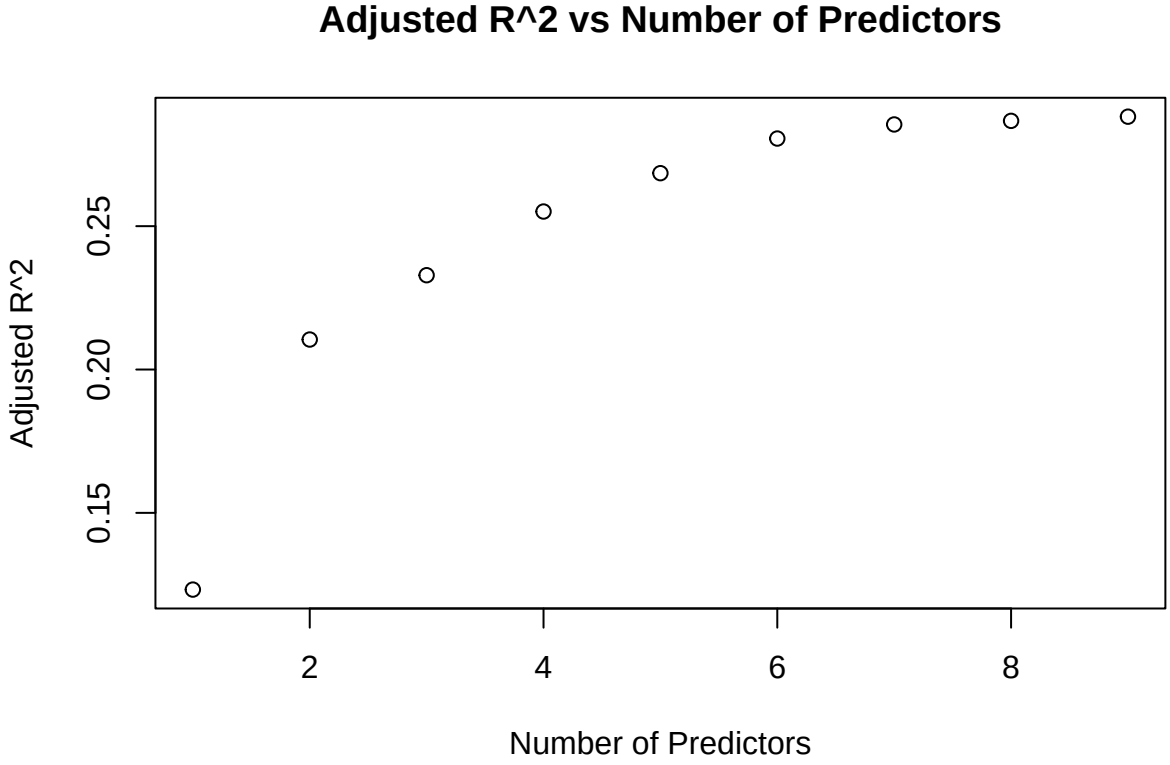
For our first question, we are looking to identify which predictors are the most significant when predicting annual income. We accomplished this through multiple methods of running a stepwise regression, an Akaike information criterion (AIC), and comparing the MSE by model.

In our EDA, we initially built a full equation with all of the relevant predictors. One of our biggest concerns to a model with this many predictors is the presence of multicollinearity. We addressed this issue by running a variance inflation factor (VIF) among each of the independent variables in our regression model. None of the predictors had a VIF value greater than 5, so we could proceed with the model selection like normal without having to modify or remove any predictors. A stepwise regression was then conducted that iteratively added and removed predictors, in the predictive model, to find the subset of variables in the data set resulting in the best performing model. After, it was discovered that the best predicting model uses the following variables as predictors for income: hrs_work, race, age, gender, citizen, lang, married, edu, disability.

Model: income = hrs_work + factor(race) + age + factor(gender) + factor(citizen) + factor(lang) + factor(married) + factor(edu) + factor(disability)

Another check to confirm the best model is using the AIC criteria method. Selecting a proper model according to adjusted R2 should balance complexity and fit. Thus, we utilized regsubsets to find the optimal model and number of predictors that will maximize adjusted R2. In our EDA, we stepped through the full model until AIC was minimized. For a better visual representation, we created a graph that plotted adjusted R2 against the number of predictors. As seen in the figure, the variability of the dependent variable income increases in a logistic manner as more predictors are added to the model. When using the which.max command to explicitly validate the number of predictors that maximizes the adjusted R2, it comes out to be all 9 predictors which

reinforces the findings from the stepwise regression.



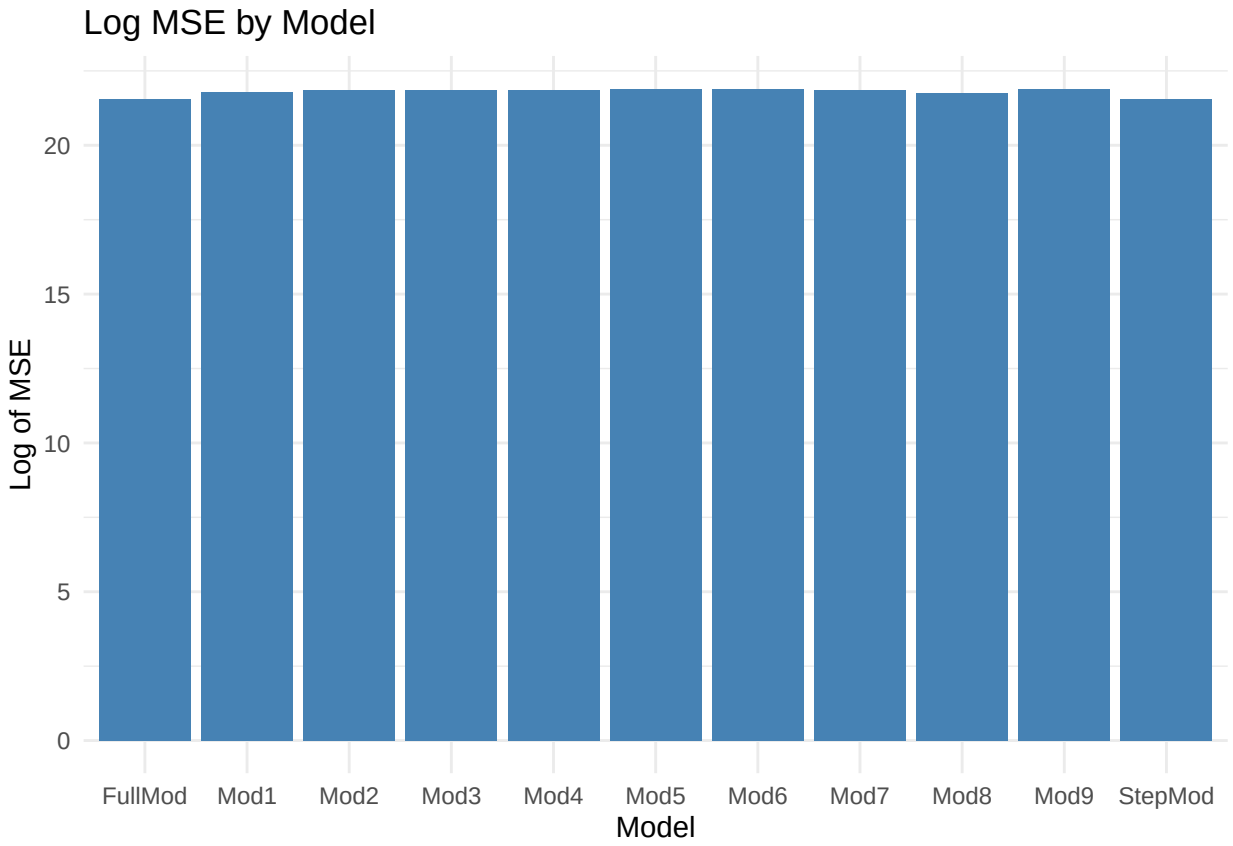
Lastly, we chose to construct models that compared individual regressors against annual income, while the remaining models were the full model and best fit model from the stepwise regression. Thus, a total of 11 models were chosen and listed below:

- FullMod: income = hrs_work + factor(race) + age + factor(gender) + factor(citizen) + factor(lang) + factor(married) + factor(edu) + factor(disability) + factor(birth_qrtr)
- StepMod: income = hrs_work + factor(race) + age + factor(gender) + factor(citizen) +
- factor(lang) + factor(married) + factor(edu) + factor(disability)
- Mod1: income = hrs_work
- Mod2: income = race
- Mod3: income = age
- Mod4: income = gender
- Mod5: income = citizen
- Mod6: income = lang
- Mod7: income = married
- Mod8: income = edu
- Mod9: income = disability

The mean squared error metric (MSE) is a common metric to measure the prediction accuracy of a model. Subsequently, we were looking to find which models have the smallest MSE since the corresponds with a better regression model. Based on the bar plot, there is very little variability between each of the individual regressions. However, looking closely, we noticed that the Model from our stepwise regression does have the lowest value for Log of MSE which indicates that it is still the best model.

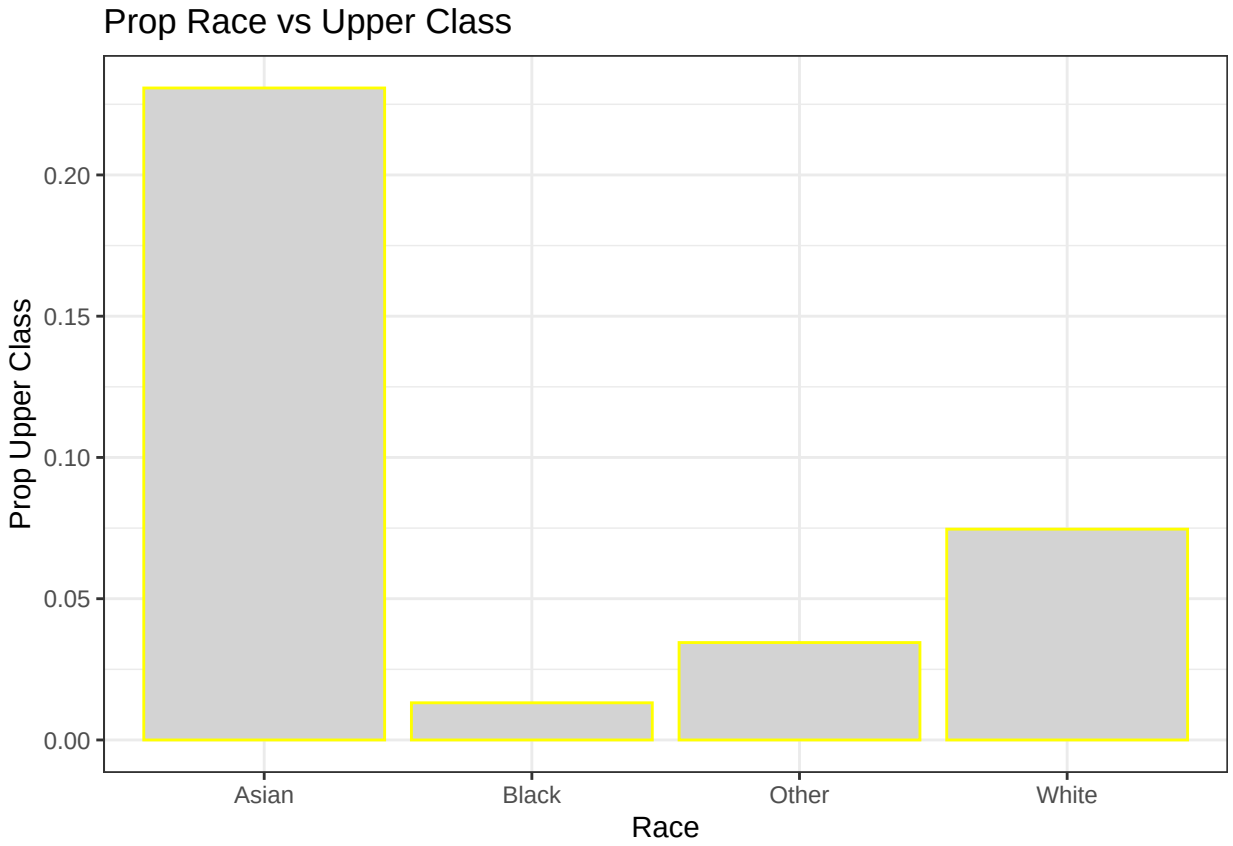
Model	Log_of_MSE
StepMod	21.54880
Mod1	21.78710
Mod2	21.86856

Model	Log_of_MSE
Mod3	21.87051
Mod4	21.84936
Mod5	21.90366

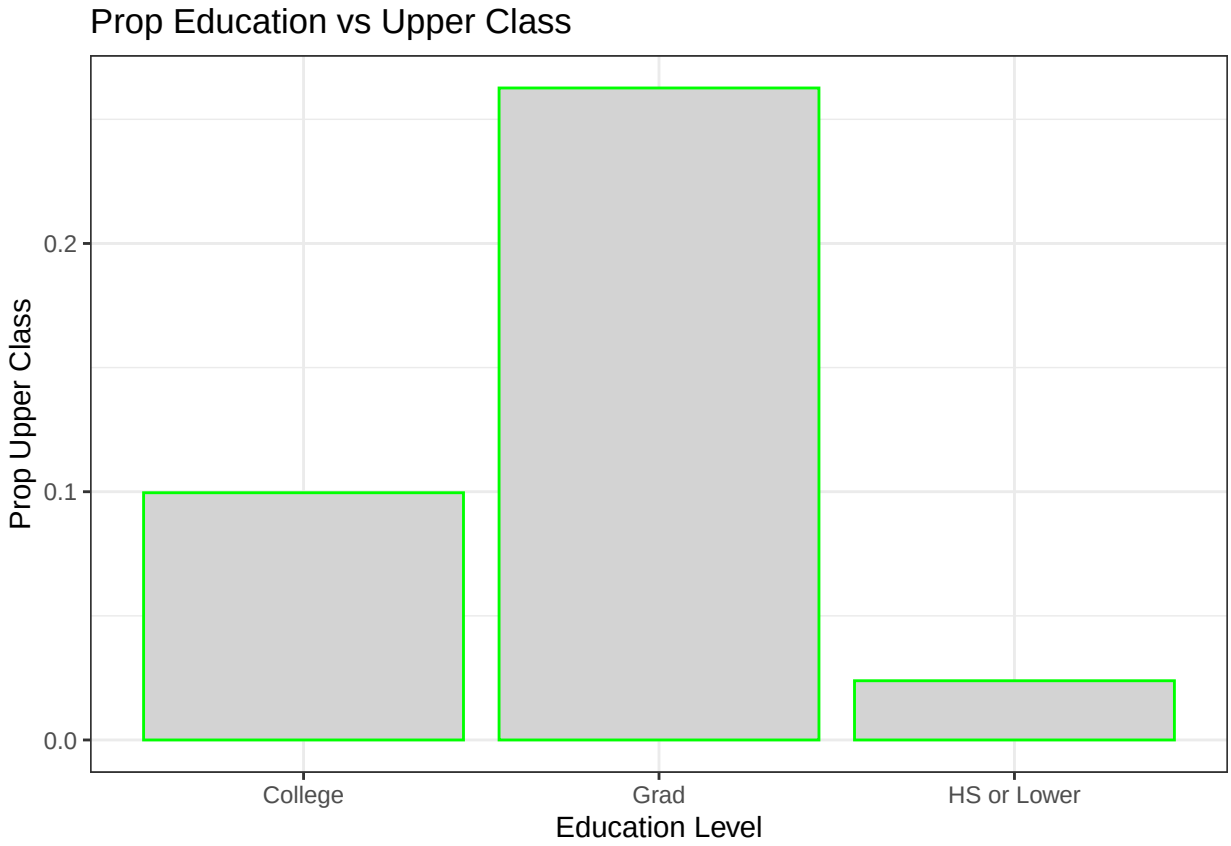


To answer our second question and find out if there is a significant difference in the proportions of each income class within the different races, education levels, and genders, we used a prop test for each combination of race, education level, and gender to see if there is a significant difference between the variables.

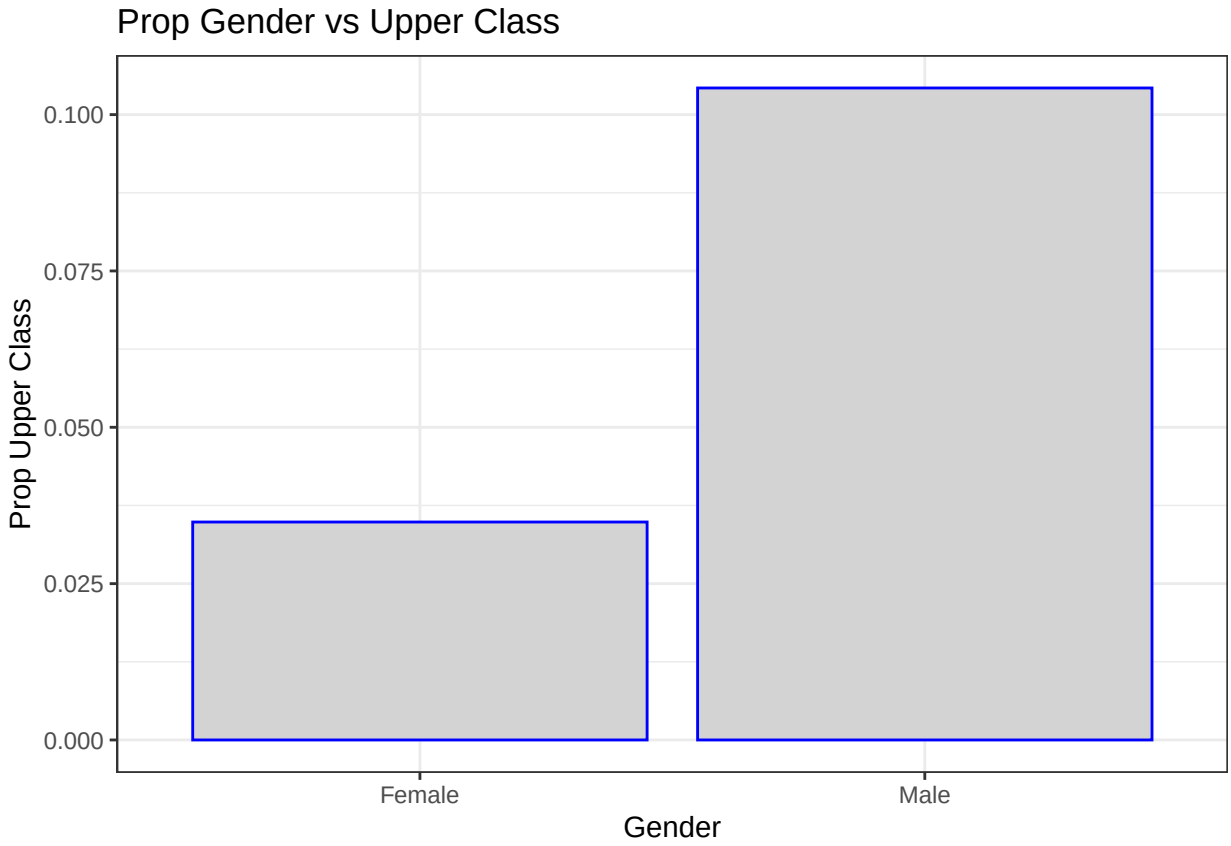
In 2012, the middle class income was below $104,087, so anything above that is consider upper class. We used this fact to create proportions for each race by counting the number of rows that that both are Asian and upper class divided by the number of rows where the race is Asian, the number of rows that that both are white and upper class divided by the number of rows where the race is white, the number of rows that that both are black and upper class divided by the number of rows where the race is black, and the number of rows that that both are other and upper class divided by the number of rows where the race is other. These four calculations gives us the proportion of each race that is of upper class status. We use these values to conduct a prop test and the p values for each of Asian-white, Asian-black, and Asian-other showed p-values of less than our significance level of 0.05, meaning that we reject the null and conclude that there is a difference in the proportions of upper income for the different races compared to Asian. After plotting each proportion of upper class individuals for each race, we found that there is a clear distinction that Asians have a much higher percentage of upper class individuals compared to other races by just looking at the graph.



When viewing our data, it seems that education level plays a significant role in determining if a person is upper class. We created proportions for each education level by counting the number of rows that are both high school or lower and upper class divided by the number of rows that have education level of high school or lower, the number of rows that are both college and upper class divided by the number of rows that have education level of college, the number of rows that are both grad and upper class divided by the number of rows that have education level of grad. This gives us the proportion of each education level that is upper class. We then conduct prop tests using these values for high school or lower - college, high school or lower - grad, and college - grad and found that the p-values were all 3 were below 0.05 so we confirm that there is a significant difference between the education levels and the amount of individuals that can achieve upper class status. The higher your education is, the more likely you will be upper class.



Finally, we created proportions for each gender by counting the number of rows that are both male and upper class divided by the number of rows that are male as well as the number of rows that are both female and upper class divided by the number of rows that are female. This gives us the proportion of each sex that is upper class. We then conduct prop tests using these values for male - female and found that the p-value was below 0.05 so we confirm that there is a significant difference between the gender you are and the amount of individuals that can achieve upper class status. After viewing the graph, we can visually see that the proportion of males that are upper class is significantly higher than the proportion of females that are upper class.



These tests have allowed us to see the influence that different variables can play in determining income, specially upper class income in our case. We used race, education level, and gender in our study, but there are countless other variables out there that can play a role in determining this as well. It's important to take different variables into perspective when interpreting data like this.

CONCLUSION

Our investigation into our first question regarding the best model to predict income revealed that a model with all of our predictors besides birth quarter is the most effective predictive model. These predictors were specifically hours worked, race, age, gender, citizen status, language, marriage status, education level, and disability status. From these findings, it's clear that the relationship between environmental factors and income is significant, which points to a larger concern about the inequities and discrimination that exists in American society that is allowing certain groups to have less opportunity to earn money. In a truly random competitive environment, every person would have an equal opportunity to have any level of income. This proposition should, in turn, show through analysis when there are no significant variables that are related to income, and a significant predictive model isn't able to be created because those relationships do not exist. This can be seen with the removal of the birth quarter predictor, which is a truly random grouping of people and has no significant relationship with income as there is with other factors. Analysis of our second question about significant differences in proportions of income classes (high, middle, and low) in relation to education levels, race groups, and gender groups revealed that higher education levels are correlated with higher income class, Asians are most likely to be upper class between all racial groups, and men are significantly more likely to be in a higher income class. These findings are somewhat consistent with societal ideas of the model minority stereotype, the understanding that education allows for higher income, and the perpetuated patriarchal idea that men need to be the breadwinner of the household.

Although it can be argued that the relationship between education and income is an equitable one, meaning those that have spent longer studying in their fields deserve to be paid more, it's also important to realize

that there are barriers to education that disproportionately impact low income groups, who may not be able to afford any higher education, or need to begin working upon graduation of high school to support their families. Through our findings, it's clear that education can be a bridge out of low income situations; thus, access to education should be available equally to all through government financial support and other avenues to break down current barriers that are in place. The dataset shows that Asians are significantly more likely to be in the high income class, and this could be attributed to societal pressure placed on Asians by the model minority stereotype. From a young age, Asians are often expected to excel in school. Growing up with these pressures may cause a shift towards going above and beyond in education and, in turn, receiving a higher salary. Furthermore, Pew Research found that 57% of Asians in the United States are foreign born and immigrants, which may explain the relationship with higher income class, as most immigrants are only able to enter the country and obtain a visa with evidence of skilled labor. Our findings for the third part of our second question shows that men are significantly more associated with higher income classes than women, and this can possibly be explained by societal expectations that men should be the breadwinners, which is an idea perpetuated from the past, but clearly still lingers in our society as seen in our analysis.

This analysis, when performed on a larger scale, has important implications about government funding and programs to promote equality and fight discriminations that continue to exist in our society. Being able to identify the income disparities between certain groups allows us to take steps towards correcting them and closing the wage gap. Further analyses can be done with more complete data, such as the further subdividing racial and ethnic groups; in our dataset, grouping together all Asians abstracts away from the immense differences between the various ethnicities within Asian population. Follow up research can also be done on datasets that contain more factors that are possibly causing division in our workforce, such as veteran status or sexuality. If disparities are found, effort should be placed into minimizing those gaps by working towards allowing everyone equal opportunity. Our analysis revealed that although many companies market themselves as equal opportunity employers, there are many factors that are often out of an individual's control that can impact their income. To improve these disparities, more work should be done to expose and identify these disparities in order to promote more discussion and create a more equitable America.