Jonothan Meyer
Data Science - Midterm
03/19/21

Data Wrangling:

Initially one of the more challenging parts of this data set was tiding it in a way to conduct analysis on all of the business owners rather than just the first owner of small businesses. This was a necessary step to have an accurate picture and understanding of the research question, especially when looking at something like whether the business is the primary or secondary source of income. To conduct analysis on only the first owner would skew the analysis in favor of a higher proportion of owners who see it as their primary source of income, either because they have the highest majority stake in the ownership, or they are the sole owner and are unable to partition the responsibilities related to the buisness.

The process for grouping all owners together was time-consuming and redundant, but allowed for an accurate way to view all small buisness owners contained within the data. The first step was to use the PIP operator to selectively filter, summarize, mutate, group, and arrange selective owner related variable columns into a data frame that could be additively combined with the adjacent other owner variables that contained similar data. For example, 'PRMINC1' and 'SEX1' were pulled, filtered, grouped, and summarized as such:

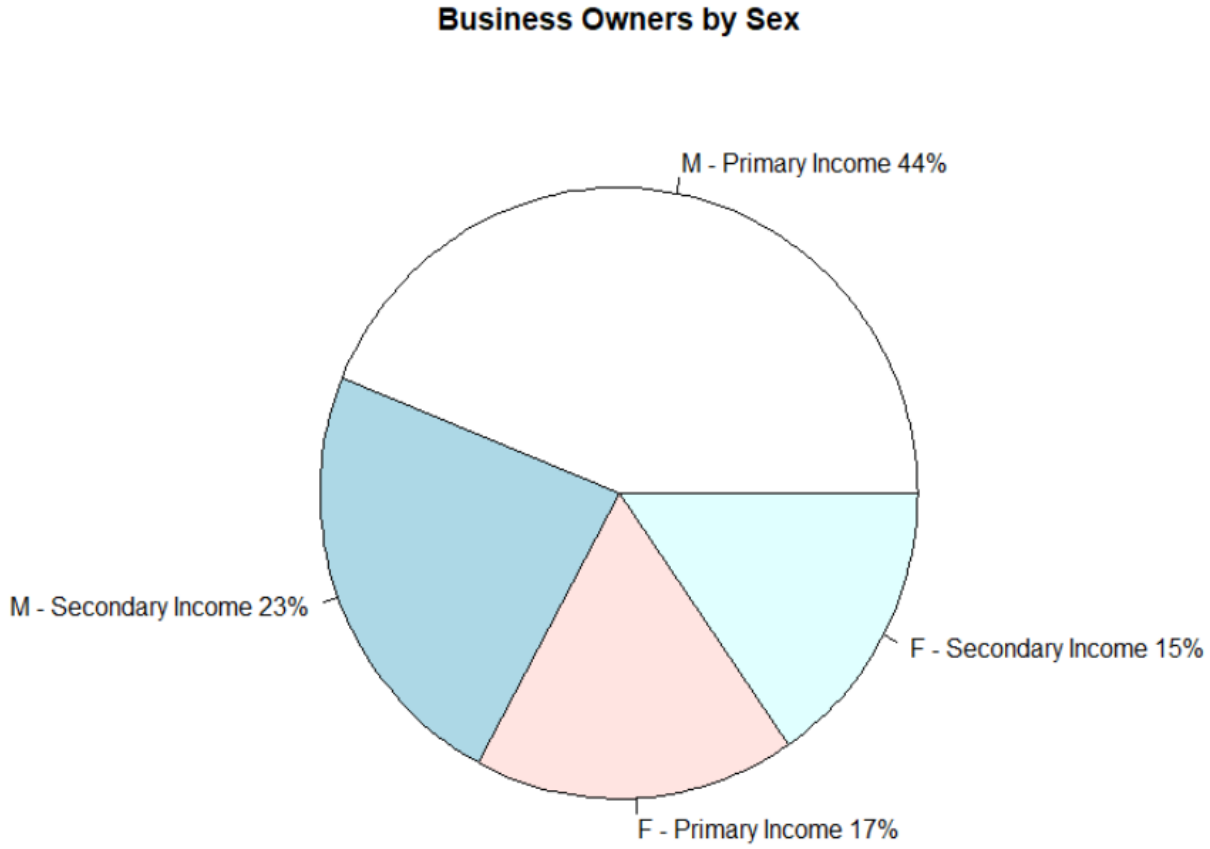
```
#sb_sex = 'small buisness sex'
sb_sex <- small_business %>% select(SEX1:SEX4, PRMINC1:PRMINC4) %>%
                filter(!is.na(SEX1) & !is.na(PRMINC1)) #basic filter of relevant columns

sb_sex1 <- sb_sex %>% rename(SEX = SEX1, PRMINC = PRMINC1) %>%
                select(SEX, PRMINC) %>% group_by(PRMINC, SEX) %>% summarize(n = n())
                #Partition and summary of only the first owners of a business. This process was repeated
                #4 times for each owner SEX1:4 column, each stored in a different variable.
...

sb_sex1$n <- sb_sex1$n + sb_sex2$n + sb_sex3$n + sb_sex4$n #Data from all owners was combined
                                                           #into a single variable for greater analysis.
```

This process was used for all variables analyzed in the data set that were partitioned into multiple columns, one to represent each owner. The biggest limitation with this method of wrangling is it treats all owners as equal within a company, regardless of the observed proportion of ownership between them. However, my goal was to gather an accurate view of the total amount of owners **included** in business ownership based off various variables, and for that purpose I believe this method to be appropriate.

Furthermore, a code variable was made named 'CodeVals' which contained numeric codes for all catagorical values in the variables that were used. The purpose of this was to be able to easily switch out categorical values with numeric values and vise versa using the left_join() function for consistent and easy arrangement and plotting. Lastly, in at least one instance a data point was removed entirely using the filter() function because it was an extreme outlier that drastically misrepresented the mean and distribution of a data group.

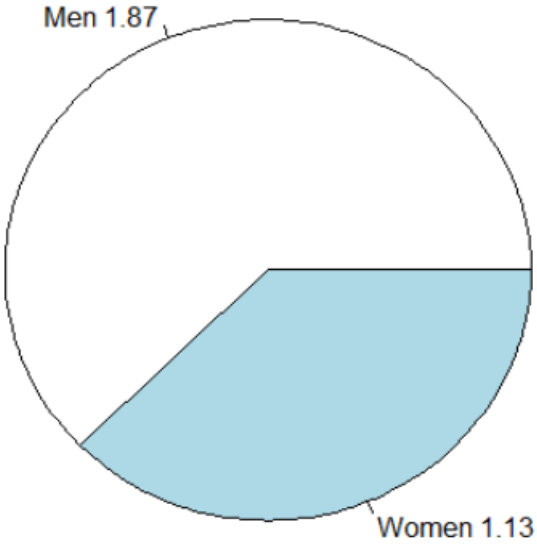
Q1 How does the extent to which income derived from the small business is not the primary source of income for a business owner vary by owner's sex, ethnicity, race, and veteran status and by business characteristics (e.g., size, sector, location)?

To answer this question I wanted to present the data in the most easily readable and understandable way possible that is also accurately able to answer this question. The graphical display's included below for both research questions primarily look at the owners within a single predictor that use their business as a primary form of income in proportion to a secondary form, or look at how large the amount of owners in a certain group compares to all other groups. Although tempted to reach for additional precise information from outside data sets and statistics, I decided to keep my analysis and conclusion within the confines of this data set, other than generally known perceptions of the racial, ethnic, and sex makeup of The United States.



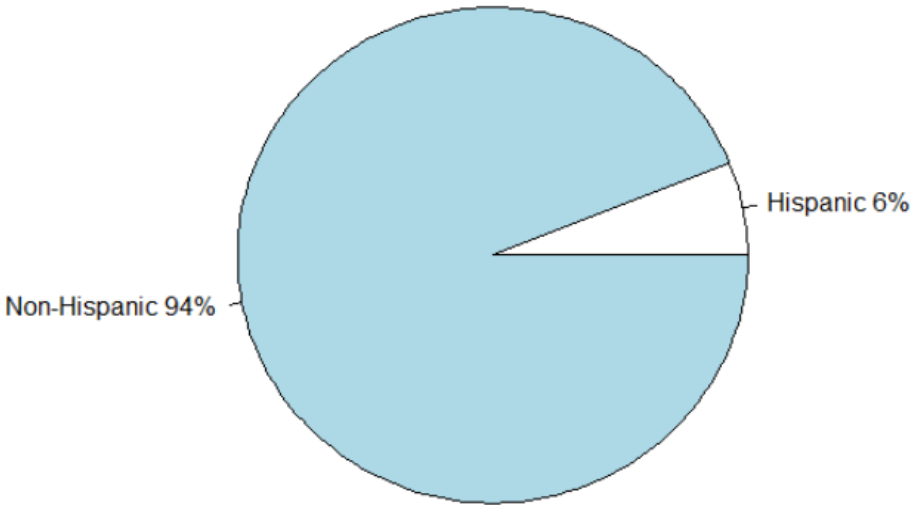
This chart illustrates that in the word of small bushiness owners a significant majority are men, accounting for 3/4 of all small business owners. Also, the proportion of owners that own a business as a primary form of income over a secondary is significantly higher as well; This concept is demonstrated more explicitly in the chart below.

Proportion of Primary Business Owners by Sex

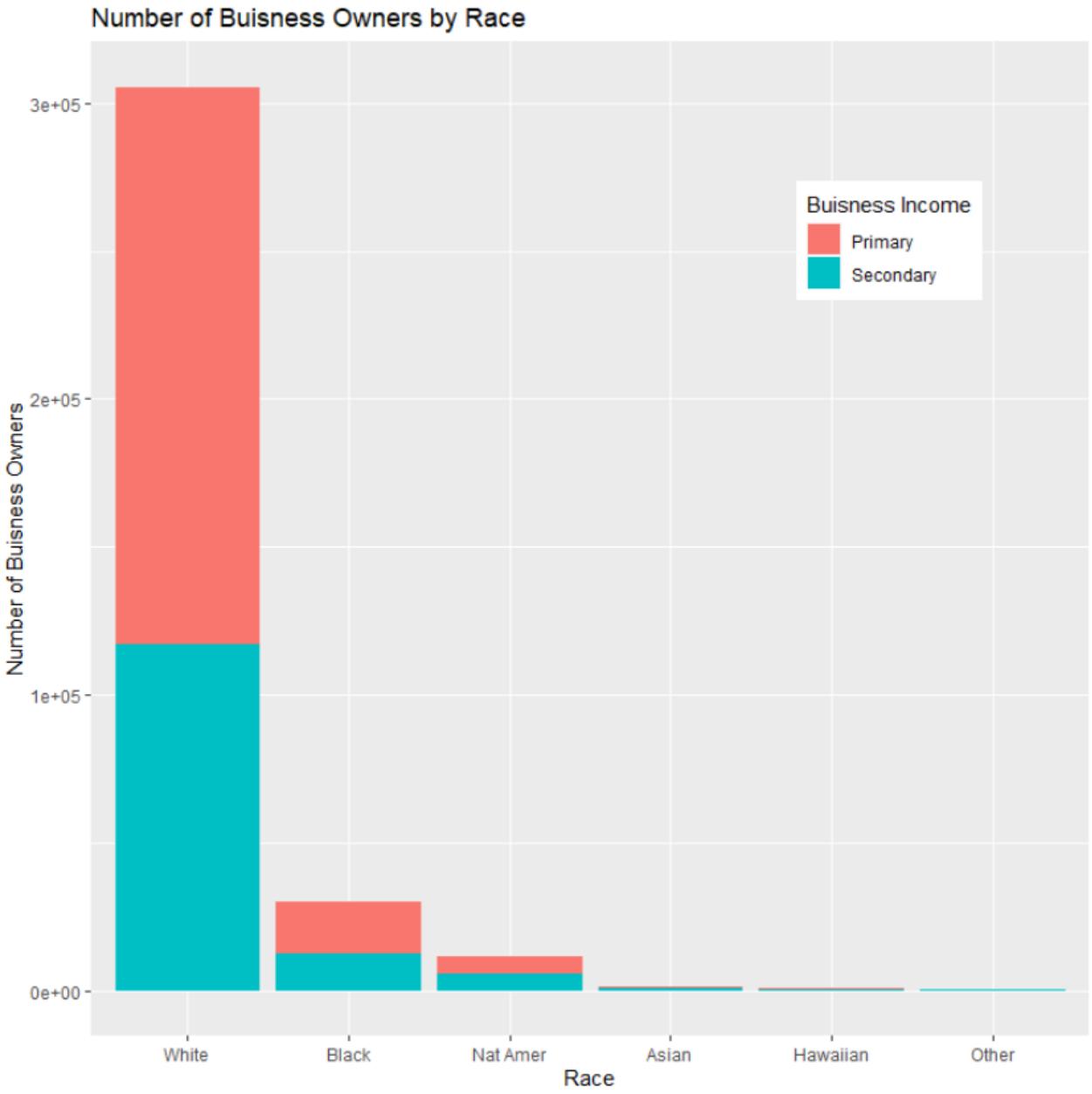


For every Man that owns a business as a secondary form of income 1.87 men own a business as a primary form of income. More interestingly, Women have close to a 1-1 ratio for primary and secondary income. This seems like a rather large difference between the sexes in business. I speculate it at least partially results from pregnancy, and the challenges contained therein in conjunction with the amount of time and work necessary to run a business full time.

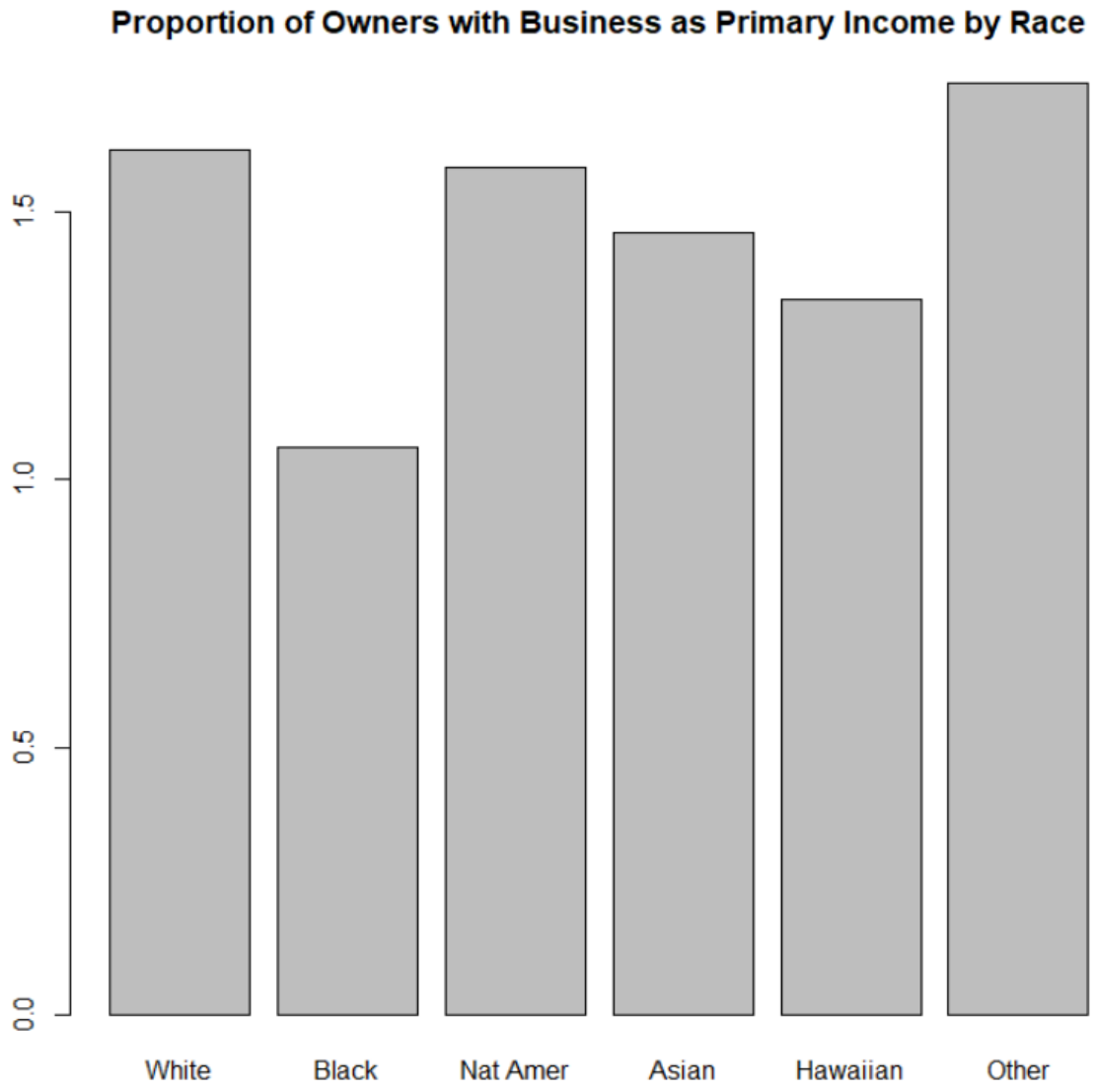
Business Ownership as Secondary Income by Ethnicity



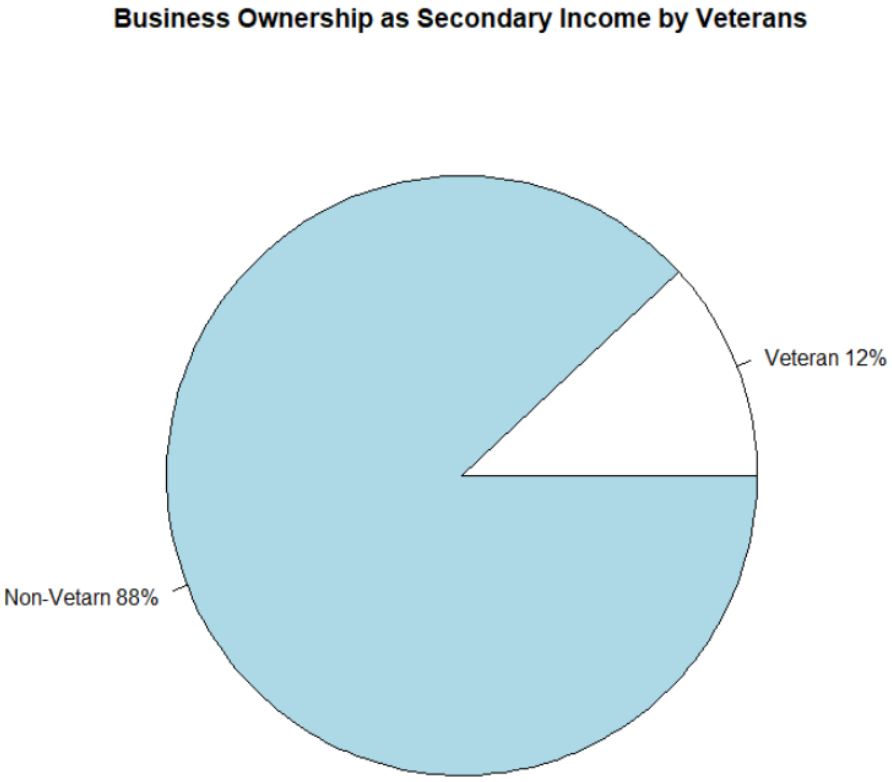
When only looking at the ethnic makeup of small business owners that own a business as their secondary form of income surprisingly only 6% are Hispanic. This percentage only increases to 7% when looking at primary income business owners, which could suggest ethnic constraints for Latinos looking to start a business.



Considering the racial proportions of people within American society this bar graph is fairly unsurprising. However, even though white people make up the majority of society they do seem to be disproportionately over-represented in small business ownership.



When we look at the proportion of primary/secondary business owners by race some interesting data emerge. All racial groups have a fairly equal amount of owners that have business as a primary form of income vs. secondary, at ~1.5 primary owners for every 1 secondary owner. However, the black population stands out as having a near 1-1 ratio. This is curious, and even atypical of minority groups since other racial groups do not share this quality.

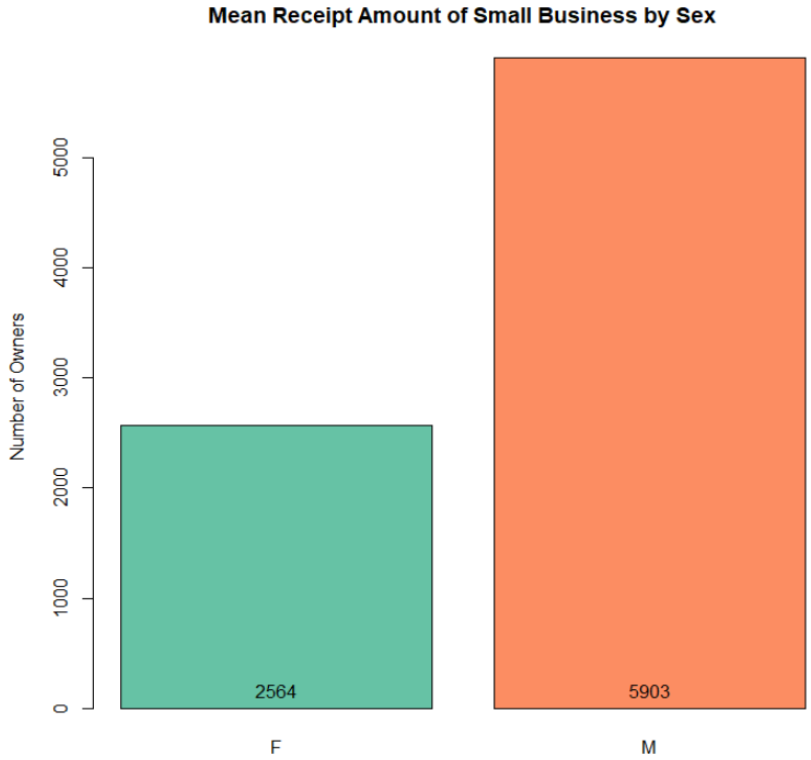


Veterans compose 12% of those who own a small business as a secondary form of income. I expected this number to be higher, and speculate that it could be because few skills learned in the military are applicable as marketable businesses in the private sector.

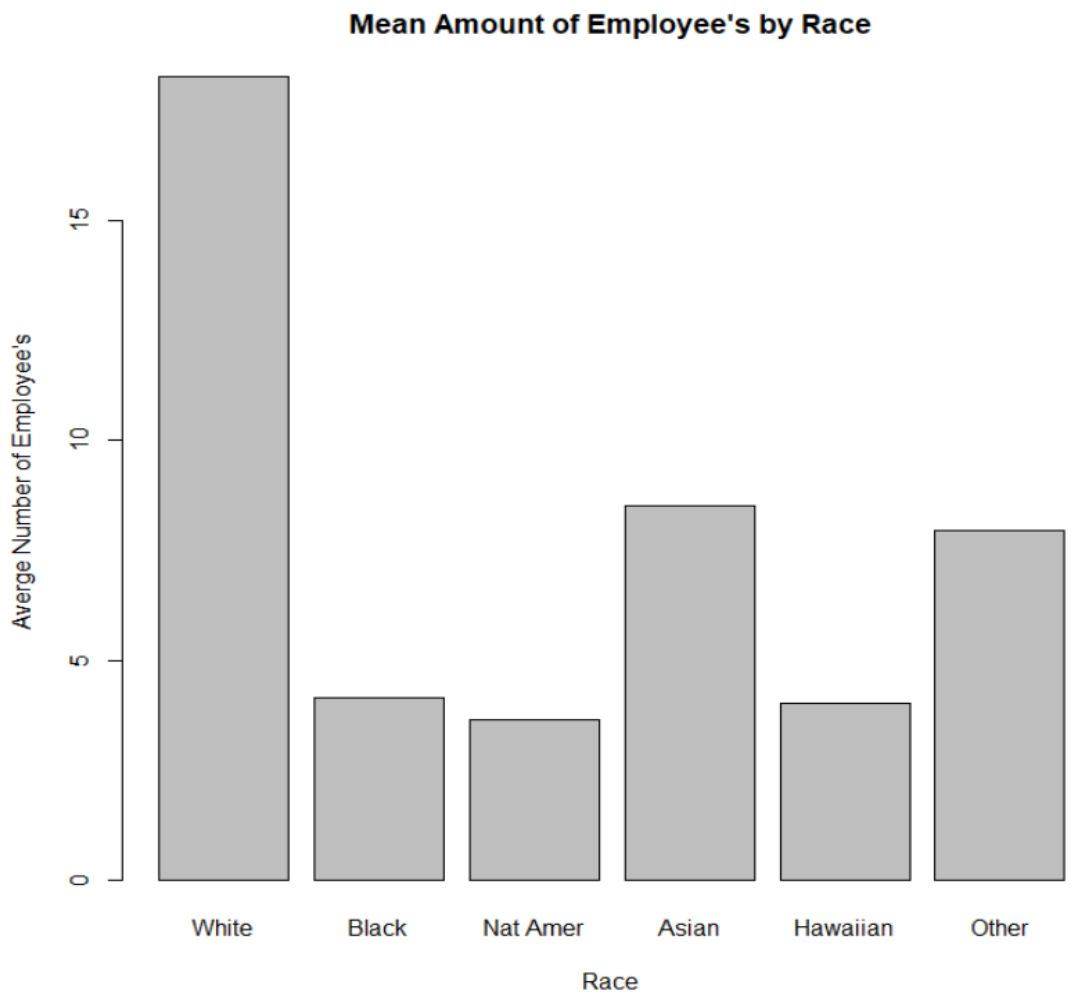
Q2 How does the business size (establishment employment, establishment payroll, establishment receipts) vary by owner's sex, ethnicity, race, and veteran status and by other business characteristics (e.g., sector, location)?

Payroll of Small Business Ownership by Sex				
Sex	Number of Owners	Summation of Payroll	Standard Deviation of Payroll	Mean Payroll
Female	207237	92975810	3419.78	448.65
Male	384830	415023900	6435.37	1078.46
Men have a much wider distribution of payroll amounts amoungst small business owners, and also have nearly double the size of mean payroll. This suggests that men own the larger of the small businesses, with a wider sd that is possibly explained by men being represented larger in a more broad amount of industries potentially.				

Amount of Employee's by Sex			
Sex	Summation of Employees	Standard Deviation of Amount of Employees	Mean Amount of Employees
Female	2557435	100.9	12.34
Male	9999659	231.54	25.98
In terms of proportions between the sexes this summary is nearly identical to the previous. Men have more than double the distribution in the amount of employee's than women, as well as about double the amount of employees. This suggests that the amount of employee's is a very strong predictor for payroll, at least among sexes.			

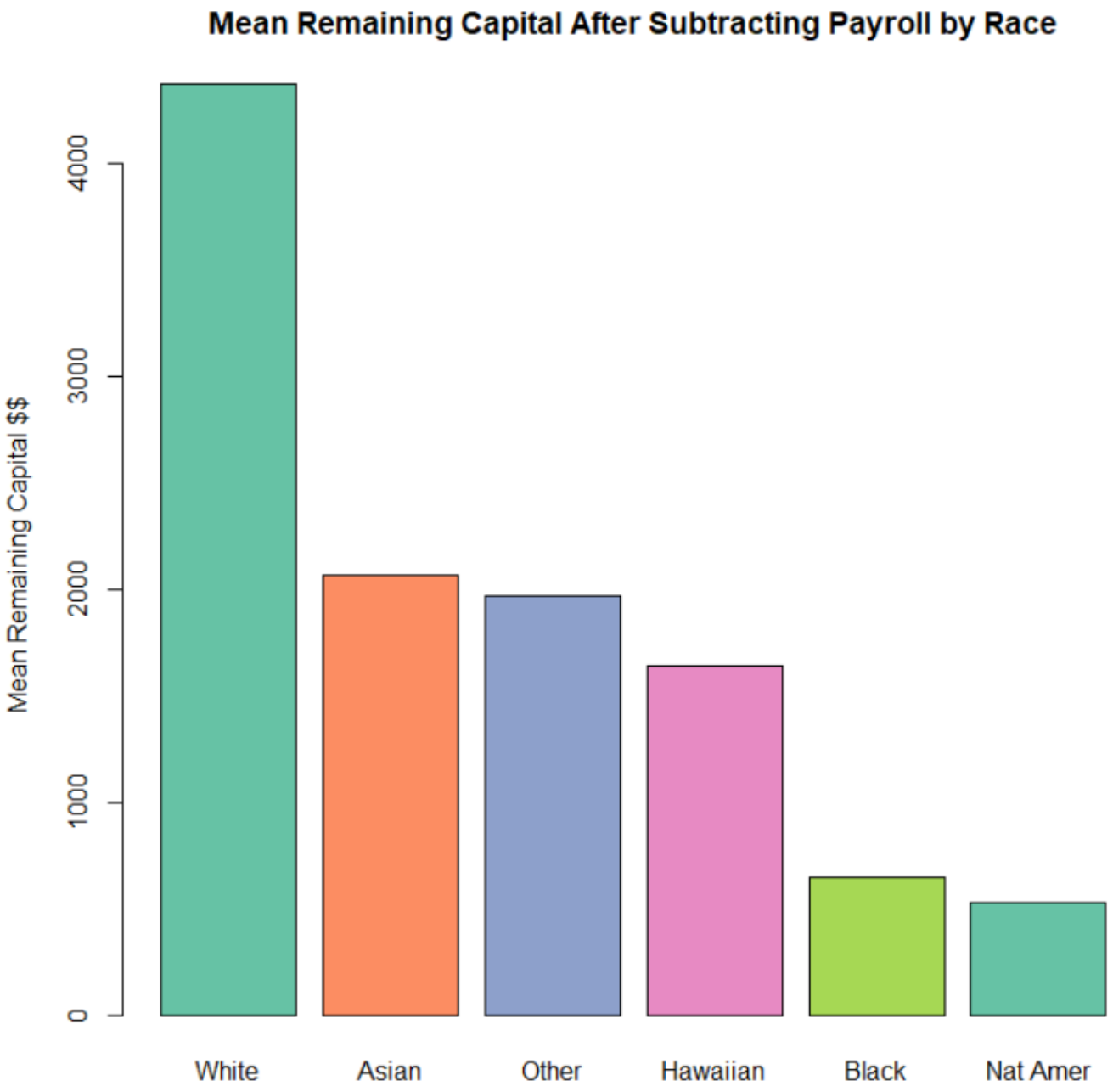


Male owned small businesses generate more than double the revenue of female owned businesses.



White small business owners, on average, have >3x the amount of employees than Black, Native American, and Hawaiian business owners, relatively. Asians might potentially be performing better than other minority groups in this category because of the high amount of educated Asian immigrants in The United States.

Amount of Employees by Race				
Race	Num of Owners	Num of Employees	Standard Deviation	Mean
White	489950	8956558	76.08	18.28
Black	127394	127394	41.01	4.15
Native American	3098	11302	18.72	3.64
Asian	56978	484239	50.91	8.5
Hawaiian	1025	4131	26.59	4.03
Other	1312	10430	31.14	7.95
Due to the high number of white small business owners, they employ 14x total more people than all other groups combined.				



Mean remaining capital was found by subtracting the sum of payroll expense for all businesses by race from receipts. Businesses certainly have additional expenses outside of payroll, but this graph gives a good idea how much an average small business has available to spend on such expenses, and also pay themselves a profit. Considering previously derived information regarding average payroll and employment it comes as no surprise that the white demographic drastically over preforms in this category. Note: A single data point was identified as an outlier and removed prior to analysis and construction of this chart. Including this data indicated that the Hawaiian group drastically over preformed in this category, and I believe was a misrepresentation of the data.

Conclusion:

Overall, not too much inference can be made about the reason for the differences in proportion between predictors and response within the confines of the research questions, and the data itself. However, this data does allow for a colorful and precise overview of the breakdown of demographics within small business ownership. Since the American racial majority is white it was expected that the majority of small bushiness ownership would be white, and this hypothesis fails to be rejected after dissecting this data. A similar thought process was initially hypothesized with regard to ownership relating to sex, and the assumption that men would hold the majority failed to be rejected as well. Women as a demographic started to significantly join the workforce less than a century ago, so it is to be expected that they would be observed in business ownership in lower numbers than men, despite representing half of the population.

Surprisingly there appears to be a positive correlation between the number of people a business employs, and remaining capital after subtracting payroll. Intuition leads you to think that since more employees necessarily means a higher payroll cost that employing more people would result in a lower amount of remaining capital; however, the data paints a different picture. The top three employers by race (White, Asian, Other) also held the highest remaining capital after deducting for employment. This leads me to believe that amount of employees could serve as an even better predictor for business revenue rather than cost.

Black communities also appear to under participate and perform in small business ownership despite making up the second largest racial demographic in The United States. Even compared to other minority groups this category is surprisingly low in measurements that could indicate financial independence and success such as mean remaining capital after deducting payroll, total number of small business owners, mean amount of employees, proportion whom's business is primary income as opposed to secondary. Without exploring additional datasets I believe these could be big indicators of systemic racism, lack of opportunity, and income inequality. Regardless of the reasons for this under representation by blacks it is clear that small business ownership inequality does exist between ethnic, racial, and sex lines. This is not just among the black community, but all minority communities to comparable degrees. These groups also under preform even when accounting for different sizes of groups within the population, as can be seen with potential business success indicators such as number of employees and remaining capital after deducting payroll.