Gabrielle Salamanca

Math 448, Tao He

March 10, 2022

Math 448: Project Progress Report I

Introduction

A name is an integral part of someone's identity and can impact their life in various ways. One of those ways is in the job aspect. A name that sounded more masculine or feminine, it may affect how people interact and treat you. There have been personal experiments and much larger ones that have tested it within the hiring process or how workflow affected the person. From Bertrand and Mullainathan's working paper, they have experimented with how ethnicity affected being called back after applying to the job.

"White names receive 50 percent more callbacksfor interviews"; and when it's attached to a high quality resume, it " … elicits 30 percent more callbacks whereas for African Americans, it elicits a far smaller increase." (Bertrand & Mullainathan, 2003)

For this project, I will find if the names are the main reason for not receiving a call back from a job, or if there are other factors that have affected it.

Descritpion of Data

The dataset, "Are Emily and Greg More Employable Than Lakisha and Jamal?", being used in this project was found in Vincent Arel-Bundock's Github projects under Rdatasets. The data was sourced from Stock and Watson's book *Introduction to Econometrics,* 2nd edition (2007). However, the original source is under the Working Papers section of the National Bureau of Economic Research's website. It is titled as *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination* by Bertrand and Mullainathan.

Bertrand and Mullainathan conducted a randomized, controlled experiment to measure racial discrimination in the job market. They have sent put 4,870 fictitious resumes to job advertisements in Chicago and Boston in 2001. Each resume was randomly assigned a name that was either a Caucasian sounding such as Allison or an African-sounding name such as Tyrone. Once sent, Bertrand and Mullainathan waited to see which ones generated a phone call from the employer.

Within the dataset, there are 4,870 rows and 27 columns. Of the 27 columns, 17 concern the applicant's resume, 9 are of the employer and job, and 1 is keeping count of the rows. Of the 27, I may remove a few that may skew the data. The next section will explain all the variables.

Explanation of Variables

The following table contains a brief description of variables of the data set. The descriptions are taken from the DOC provided by Vincent Arel-Bundock's Github projects.

Variable	Description
X	Cell number
Name	Applicant's first name
Gender	Applicant's gender
Ethnicity	Applicant's ethnicity (i.e., Caucasian vs African-American sounding)
Quality	Quality of the resume
Call	Was the applicant called back?
City	Boston or Chicago
Jobs	Number of jobs listed on the resume
Experience	Number of years of work experience on the resume
Honors	Did the resume mention some honors?
Volunteer	Did the resume mention some volunteering experience?
Military	Does the applicant have military experience?
Holes	Does the resume have some employment holes?
School	Does the resume mention some work experience while at school?
Email	Was the e-mail address on the applicant's resume?
Computer	Does the resume mention some computer skills?
Special	Does the resume mention some special skills?

College	Does the applicant have a college degree or more?
Minimum	Minimum experience requirement of the employer
Equal	Is the employer EOE (Equal Opportunity Employment)?
Wanted	Type of position wanted by employer
Requirements	Does the ad mention some requirement for the job?
Reqexp	Does the ad mention some experience requirement?
Reqcomm	Does the ad mention some communication skills requirement?
Reqeduc	Does the ad mention some educational requirement?
Reqcomp	Does the ad mention some computer skills requirement?
Reqorg	Does the ad mention some organizational skills requirement?
Industry	Type of employer industry

Data Entry

While the Resume Names dataset doesn't have any missing values in terms of each cell being filled in or not, it does have values filled as unknown, specifically within the industry variable. It is the third largest subcategory in that variable, and it could skew the data. As of right now, the industry variable won't be removed. It will only be used for visualization purposes. Later on in the project, it will be removed to balance the data.

Other variables that will be dropped later on would be the name and wanted variables, because ethnicity provides the information needed for this project, and there was no use for them. Variables that may be removed would be the ones that don't have a paired req variable. For example, there is a variable reqcomp. To see if the applicant met that requirement, the variable computer can be used. This means these variables would be kept. Now, a variable that may not be kept is reqcomm, where it answers if the ad mentioned some communication skills requirement. Among the 27 variables, there is no variable that can answer if the applicant does have those skills or not.

However, they were all kept for now for the sake of visualization and understanding the data better. So, there was no need for data cleaning right now.

So, the first step that was done was reading the Resume Names data set into R by using the read.csv function. If the summaryfunction was used right now, the variables would only return as characters, because most of the variables are categorical. To fix that, stringsAsFactors = T was added into the arguments, so that there were actual counts of the different responses. After reading in the data set, the dim, names, and summary functions were used to test if it was being properly read and to get the summary of the data.

Summary Results

As mentioned before, a majority of the variables in this data set was categorical. So, stringsAsFactors = T was added into the arguments so the data can be read as numerical. Upon looking into the summary, there were definitely some variables that were skewed, such as gender, call, and honors. This was especially for the call variable, there was an astounding amount of no's. Because of this, the data set will have to be down sampled, and one or more variables will have to be removed to make it more balanced. Others were more balanced such as ethnicity, quality, and holes.

The most notable variables were jobs and experience. They were the only two that had minimum and max values.The variable X doesn't count, because it only serves as counting how many rows there were in the data set. The variable jobs is the number of jobs listed and the variable experience is the number of years of work experience on the resume. These two variables were the only ones that were strictly numerical. Under the jobs variable, most applicants previously had 5 jobs before they applied to the ad. For the experience variable, most had under 10 years of work experience.

The full summary output will be at the end of the Data Visualization/Observations section

Data Visualization/Observations

Let us first take a look what general industries did the applicants apply to.

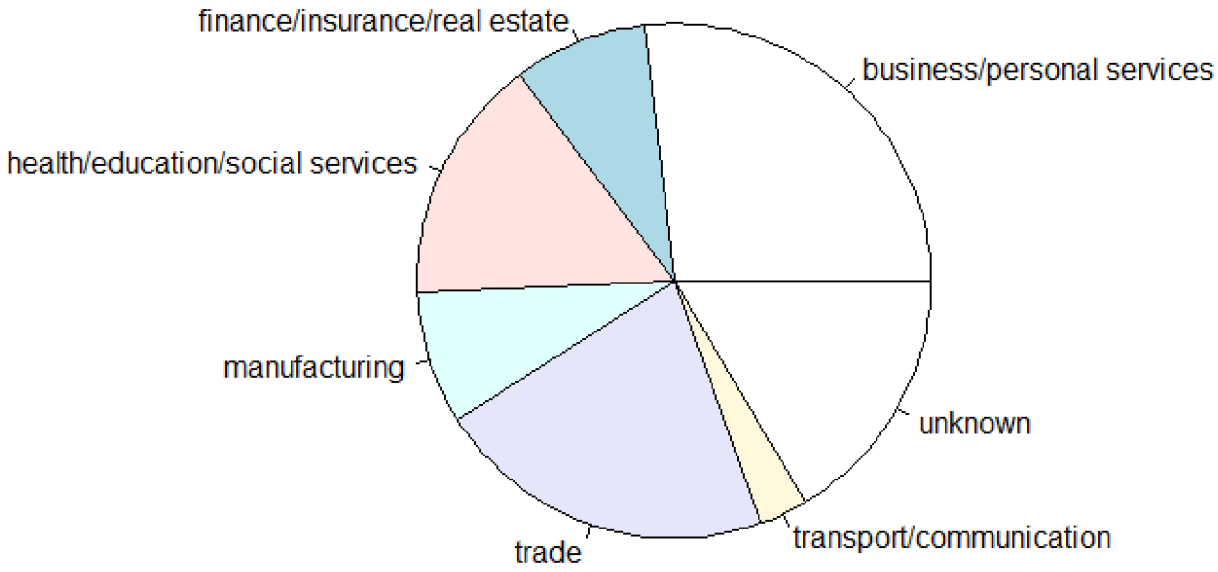


Figure 1

Nearly half of them applied to business/personal services and trade. However, the third largest slice is unknown as mentioned before. While it would be possible to separate callbacks based on industry, it may be best to drop the industry variable because of the amount of missing values.

Next, let's visit the distribution of African American and Caucasian applicants:

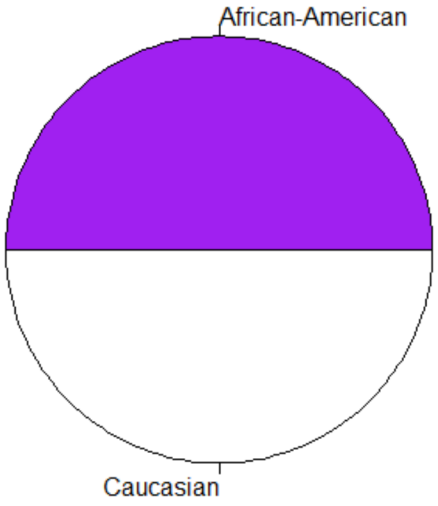


Figure 2

The distribution is exactly 2,435, meaning the data is balanced. We won't have to worry about any skewing. Now, we look at the call back distribution:

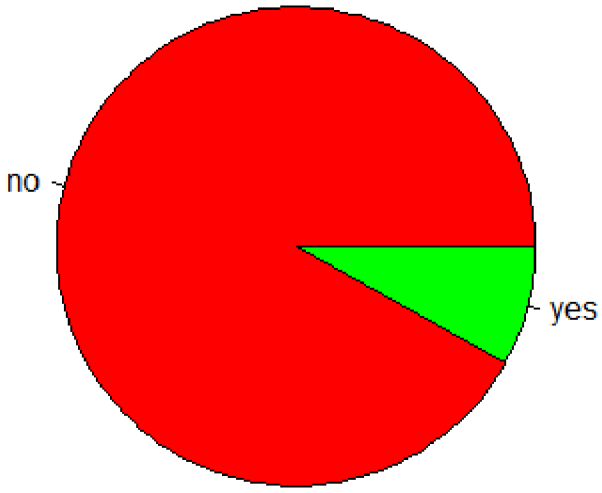


Figure 3

Compared to the ethnicity variable, it is incredibly skewed towards no. To be exact, there are 4,478 no's and 392 yes's. Because of this, we will have to down sample the dataset so it can be balanced.

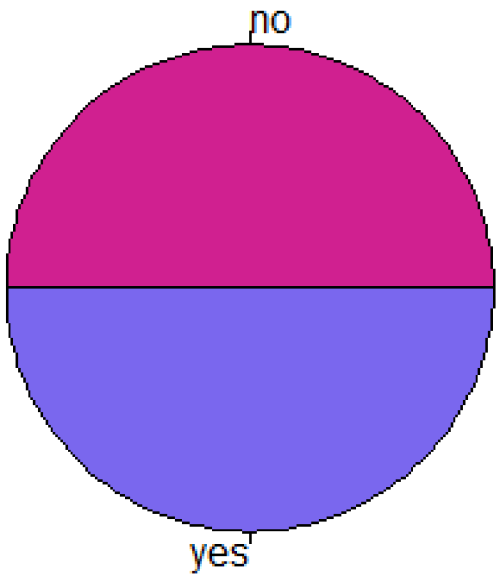


Figure 4

After downsampling, the call variable has been balanced. However, the ethnicity variable was affected.

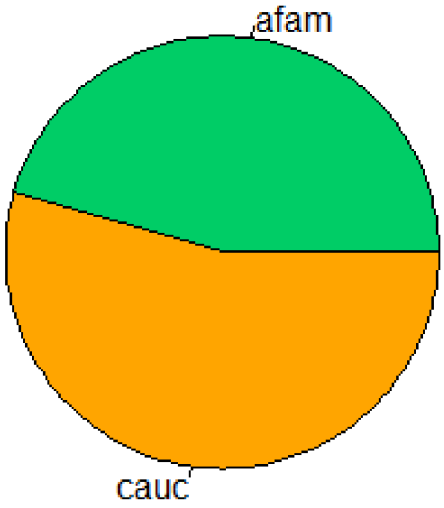


Figure 5

There are more Caucasians than African-Americans now, which may provide some evidence that more Caucasian-sounding names help with the hiring process. This will have to be revisited another time. We will still keep the raw data, minus the X and name columns.

Let us look at some stacked bar charts. Below is comparing callbacks based on ethnicity and gender.

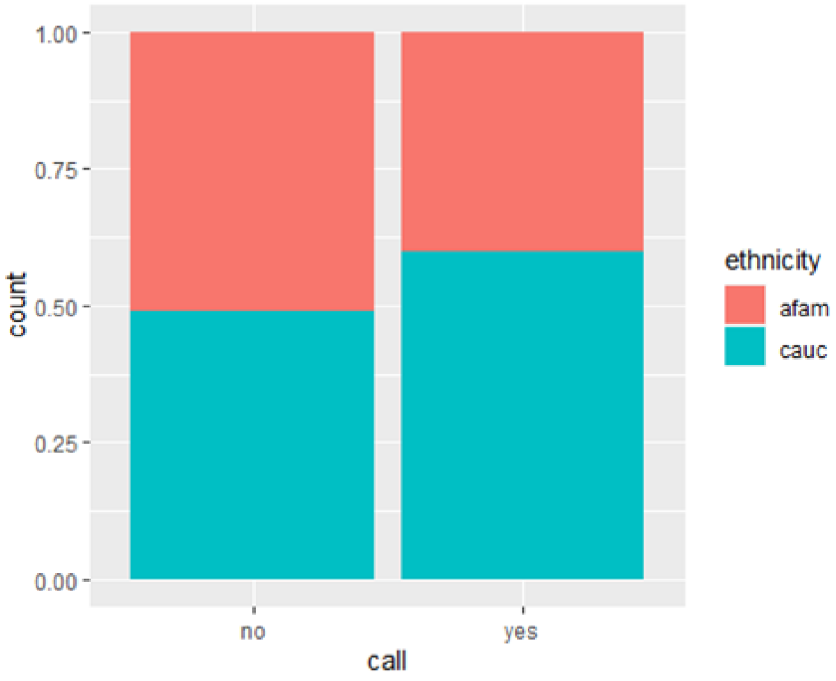


Figure 6

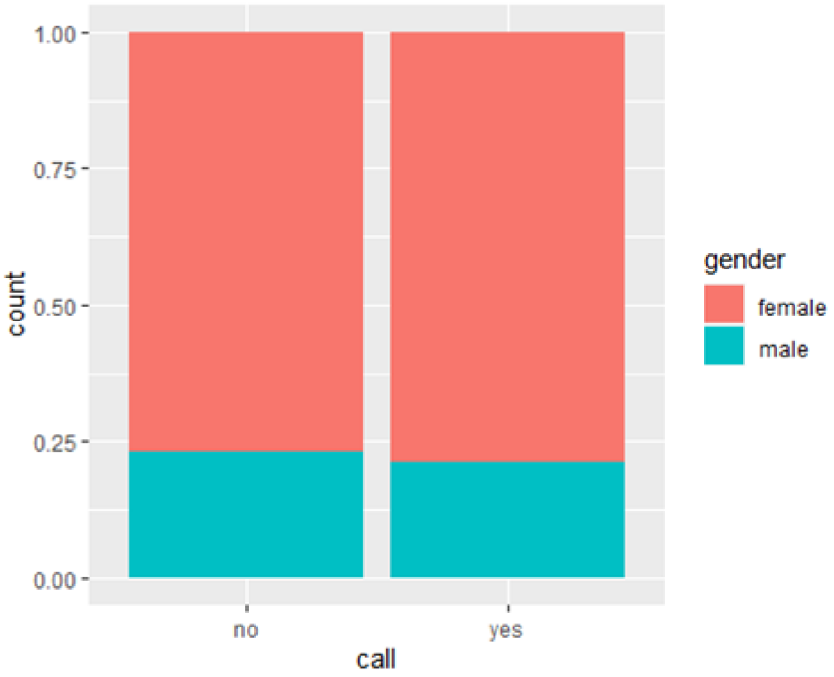


Figure 7

Caucasians received more callbacks compared to African-Americans, but there was significantly more females in receiving callbacks and receiving none, which is interesting. I had to take a look into the gender and equal variables, if perhaps this was an effect to these figures, particularly Figure 7.

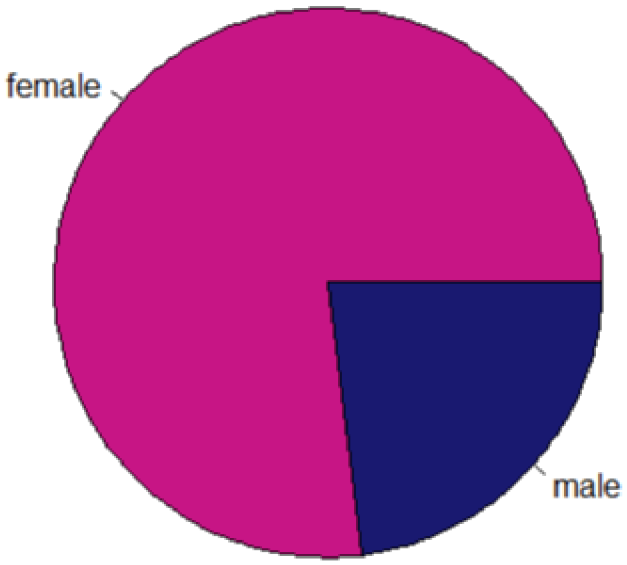


Figure 8

There are significantly more females than males in the dataset, so this is the reason why there are more females in Figure 7. Now this brings in how to down sample the data. There definitely needs to be a down sampling in the callback variables, but it does skew the important variable we are looking at: ethnicity. However, we are also unbalanced with the gender variable.

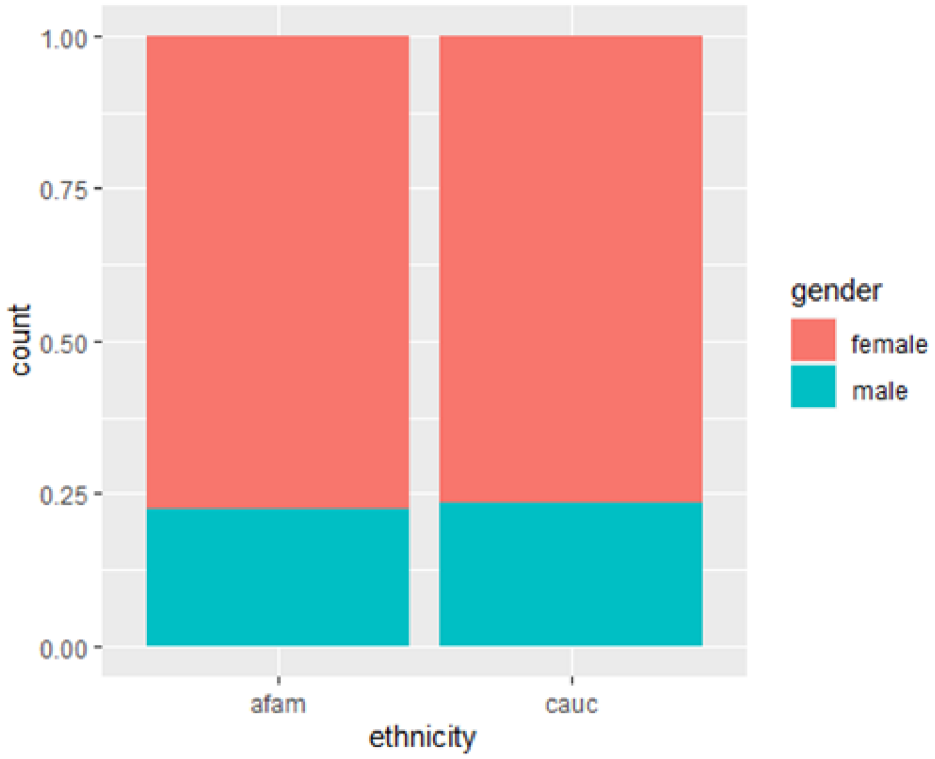


Figure 9

I may have to down sample this data set far more than I thought, but I do worry about removing too much and having a far smaller data set than expected. This will have to be explored at a later date, because downsampling now could greatly affect the graphs and plots.

So, now we take a look at the equal variable in terms of industry. The equal variable answers if the employer is EOE, or Equal Opportunity Employment. If one does not know what EOE is, it is a list of federal laws that " … protect employees and job applicants against employment discrimnation … " according to the U.S. Equal Employment Opportunity Commission. The first bullet on their website is what I want to highlight: " … when it involves unfair treatment because of race, color, religion, sex (incliding pregnancy, gender identity, and sexual orientation), national origin, age (40 or older), disability or genetic information."

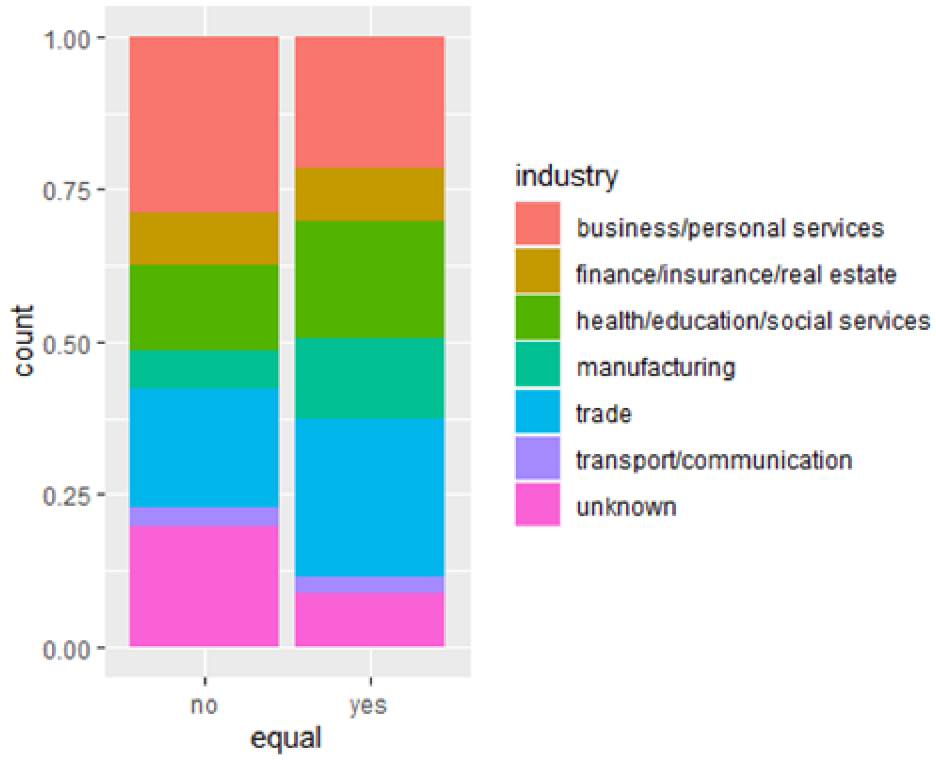
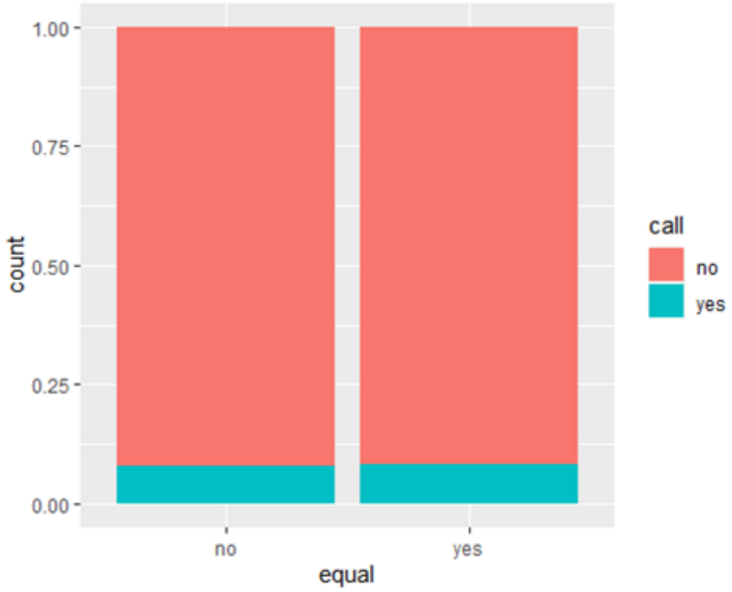
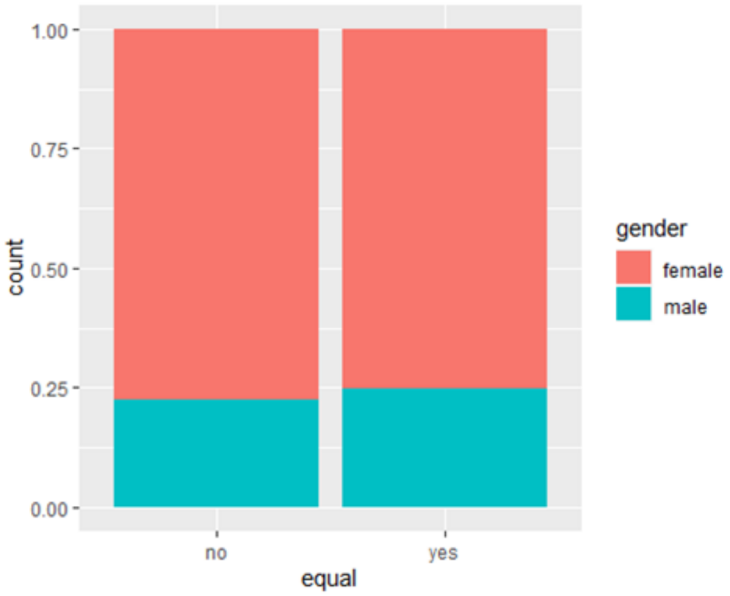
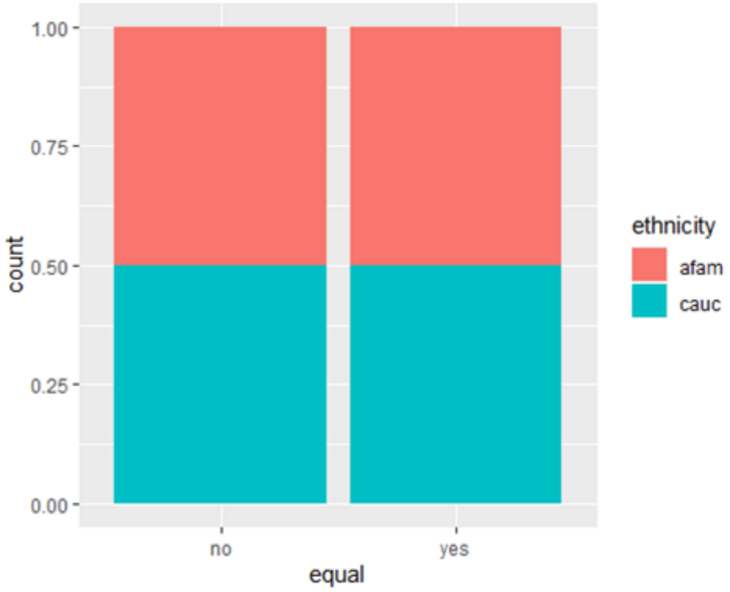


Figure 10

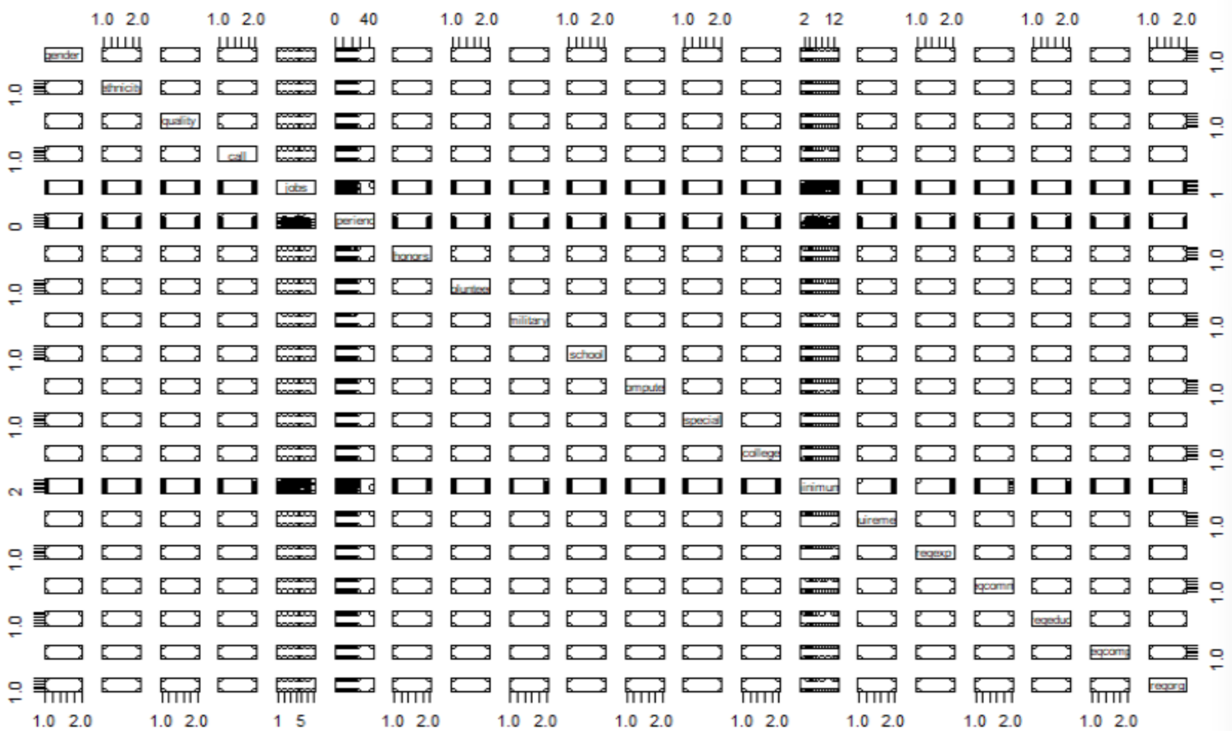
Looking at this bar chart, most industries are either equally halved or lean more towards yes, but there are more no EOE industries in business/personal services and unknown. I then looked at if the employer being EOE has affected callbacks, gender, and ethnicity. Looking at all of them, the employer being EOE had little to no affect on any of these variables.







Because I did not want to go through over 20 variables to see if it had an effect on callbacks, I have tried using the pairs function on the dataset, however the visual did not shed any light in the relationships. This was also after removing the X, name, city, holes, email, equal, wanted, and industry variables.


gender
ethnicity
quality
call
jobs
experience
honors
volunteer
military
school
computer
special
college
minimum
requirement
reqexp
reqcomm
reqeduc
reqcomp
reqorg

The original summary:

```
summary(Rnames)
```

```
##        X             name         gender     ethnicity   quality      call
##  Min.   :   1   Tamika : 256   female:3746   afam:2435   high:2446   no :4478
##  1st Qu.:1218   Anne   : 242   male  :1124   cauc:2435   low :2424   yes: 392
##  Median :2436   Allison: 232
##  Mean   :2436   Latonya: 230
##  3rd Qu.:3653   Emily  : 227
##  Max.   :4870   Latoya : 226
##                 (Other):3457
##
##       city           jobs         experience     honors     volunteer
##  boston :2166   Min.   :1.000   Min.   : 1.000   no :4613   no :2866
##  chicago:2704   1st Qu.:3.000   1st Qu.: 5.000   yes: 257   yes:2004
##                 Median :4.000   Median : 6.000
##                 Mean   :3.661   Mean   : 7.843
##                 3rd Qu.:4.000   3rd Qu.: 9.000
##                 Max.   :7.000   Max.   :44.000
##
##  military   holes      school     email      computer   special    college
##  no :4397   no :2688   no :2145   no :2536   no : 874   no :3269   no :1366
##  yes: 473   yes:2182   yes:2725   yes:2334   yes:3996   yes:1601   yes:3504
##
##     minimum     equal                  wanted     requirements reqexp
##  none   :2746   no :3452   manager       : 741   no :1036     no :2750
##  some   :1064   yes:1418   office support: 578   yes:3834     yes:2120
##  2      : 356              other         : 736
##  3      : 331              retail sales  : 818
##  5      : 163              secretary     :1621
##  1      : 142              supervisor    : 376
##  (Other):  68
##
##  reqcomm    reqeduc    reqcomp    reqorg
##  no :4262   no :4350   no :2741   no :4516
##  yes: 608   yes: 520   yes:2129   yes: 354
##
##          industry
##  business/personal services      :1304
##  finance/insurance/real estate   : 414
##  health/education/social services: 754
##  manufacturing                   : 404
##  trade                           :1042
##  transport/communication         : 148
##  unknown                         : 804
```

The summary after downsampling based on call and removing the X and name columns.

```
summary(Rnames)
```

```
##     gender    ethnicity  quality     call          city          jobs      
##  female:619   afam:355   high:410   no :392   boston :366   Min.   :1.000  
##  male  :165   cauc:429   low :374   yes:392   chicago:418   1st Qu.:3.000  
##                                                            Median :4.000  
##                                                            Mean   :3.662  
##                                                            3rd Qu.:4.000  
##                                                            Max.   :7.000  
##
##    experience     honors    volunteer military  holes     school    email    
##  Min.   : 1.000   no :728   no :462   no :705   no :400   no :339   no:388  
##  1st Qu.: 5.000   yes: 56   yes:322   yes: 79   yes:384   yes:445  yes:396  
##  Median : 7.000                                                            
##  Mean   : 8.186                                                            
##  3rd Qu.:10.000                                                            
##  Max.   :26.000                                                            
##
##  computer  special   college      minimum    equal          wanted       
##  no :147   no :477   no :233   none   :452   no :554   manager       :103  
##  yes:637   yes:307   yes:551   some   :180   yes:230   office support:107  
##                                2      : 59             other         :108  
##                                3      : 46             retail sales  :127  
##                                5      : 20             secretary     :278  
##                                1      : 18             supervisor    : 61  
##                                (Other):  9                                 
##  requirements reqexp    reqcomm   reqeduc   reqcomp   reqorg    
##  no :173      no :452   no :671   no :712   no :443   no :733   
##  yes:611      yes:332   yes:113   yes: 72   yes:341   yes: 51   
##
##
##
##
##
##           industry                        Class
##  business/personal services      :212   no :392  
##  finance/insurance/real estate   : 64   yes:392  
##  health/education/social services:142            
##  manufacturing                   : 50            
##  trade                           :145            
##  transport/communication         : 31            
##  unknown                         :140
```

Code

```
#####################################
#Course Project: Reading & Analyzing Data#
#####################################
#calling libraries
library(caret)
library(ggplot2)
#calling dataset
Rnames = read.csv("Resume Names.csv",header=T,na.strings="?",stringsAsFactors = T)
#checking out the specs
View(Rnames)
dim(Rnames)
names(Rnames)
summary(Rnames)
#industry variable
pie(table(Rnames$industry))
#ethnicity variable
colors = c("purple", "white")
ethnic <- c("African-American", "Caucasian")
pie(table(Rnames$ethnicity), col = colors, labels = ethnic)
#call variable
colors = c("red", "green")
pie(table(Rnames$call),col = colors)
#plots w/ x = call
ggplot(Rnames, aes(x = call, fill = ethnicity)) + geom_bar(position = "fill")
```

```
ggplot(Rnames, aes(x = call, fill = gender)) + geom_bar(position = "fill")
#gender variable
colors = c("mediumvioletred", "midnightblue")
pie(table(Rnames$gender),col = colors)
#plots w/ x = equal
ggplot(Rnames, aes(x = equal, fill = industry)) + geom_bar(position = "fill")
ggplot(Rnames, aes(x = equal, fill = call)) + geom_bar(position = "fill")
ggplot(Rnames, aes(x = equal, fill = gender)) + geom_bar(position = "fill")
ggplot(Rnames, aes(x = equal, fill = ethnicity)) + geom_bar(position = "fill")
#pairs
Names = subset(Rnames, select = -c(X, name, city, holes, email, equal, wanted, industry))
pairs(Names)
#downsample in callback
Rnames <- downSample(Rnames, Rnames$call)
Rnames = subset(Rnames, select = -c(X, name))
summary(Rnames)
colors = c("violetred", "slateblue2")
pie(table(Rnames$call),col = colors)
colors = c("springgreen3", "orange")
pie(table(Rnames$ethnicity),col = colors)
```

Source

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, **94**, 991–1013.

Bertrand, M. and Mullainathan, S. (2003). *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination* (Working Paper No. 9873). National Bureau of Economic Research. http://www.nber.org/papers/w9873

Stock, J.H. and Watson, M.W. (2007). *Introduction to Econometrics*, 2nd ed. Boston: Addison Wesley.

U.S. Equal Employment Opportunity Commission. (2022.). *Employers*. U.S. Equal Employment Opportunity Commission. Retrieved March 12, 2022, from https://www.eeoc.gov/employers