ANALYSIS OF THE “RESUME” DATASET USING R

Adel Broussard-Sanchez
C00162321
INFX 512
Dr. Tozal
Spring 2024

Contents

Dataset

Description

This dataset comes from a study done by Marianne Bertrand and Sendhil Mullainathan titled *Are Emily And Greg More Employable Than Lakisha And Jamal? A Field Experiment On Labor Market Discrimination.* The data was collected from 2001 to 2002, and the study was published in 2003. The researchers sought to uncover whether race and gender influenced the rate of application callbacks. To obtain the *resume* dataset, the researchers used resumes from real job seekers in the Boston and Chicago areas, changed the names to reflect stereotypical African American and White names, changed the addresses to reflect either "good" or "bad" neighborhoods in the areas, and then counted how many callbacks these resumes received (Bertrand and Mullainathan, p. 1). The dataset consists of 4,870 observations (resumes) over 30 variables (Arel-Bundock). The original dataset is composed of 10 categorical variables (10 character variables) and 20 numerical variables (10 numeric variables and 10 integer variables). A more in-depth description of these variables can be found below in the section titled *Variable Descriptions.*

Cleaning the Data

In order to view the original dataset, I first downloaded it in CSV format Github and viewed it in Excel. I immediately saw that there were many missing values for the variables job_fed_contractor and for the observation job_req_min_experience. Because these missing values indicate that the information was unavailable in the original data collection, there was not much I could do to fix this issue. I did not want to remove any observations completely as that would skew the results when analyzing the other variables. So, I left them as is and, as I will show in my analysis, will work around this issue. Once I made sure there were no spelling errors, I loaded the dataset into R. It is also worth noting that this dataset is labeled a 'tibble' as opposed to the more common 'data frame'. According to the R-Project, "Tibbles are a modern take on data frames" (R-Project). Put simply, tibbles are a type of data frame that addresses and fixes some problematic behaviors of data frames that have not aged well.

Loading the Data

```
> install.packages("openintro")
> library("openintro")
> data(resume)
```

Changing the Data Types

As is stated above and is seen in the below image, the dataset contains both numeric and integer values. I decided to change all the integer variables to numeric for easier analysis. This will prevent me from running into problems regarding this in the future. I used the as.numeric() command on all integer variables, then used the str() command to view the changes.

```
> str(resume)
tibble [4,870 × 30] (S3: tbl_df/tbl/data.frame)
 $ job_ad_id             : num [1:4870] 384 384 384 384 385 386 386 385 386 386 ...
 $ job_city              : chr [1:4870] "Chicago" "Chicago" "Chicago" "Chicago" ...
 $ job_industry          : chr [1:4870] "manufacturing" "manufacturing" "manufacturing" "manufacturing" ...
 $ job_type              : chr [1:4870] "supervisor" "supervisor" "supervisor" "supervisor" ...
 $ job_fed_contractor    : num [1:4870] NA NA NA NA 0 0 0 0 0 0 ...
 $ job_equal_opp_employer: num [1:4870] 1 1 1 1 1 1 1 1 1 1 ...
 $ job_ownership         : chr [1:4870] "unknown" "unknown" "unknown" "unknown" ...
 $ job_req_any           : num [1:4870] 1 1 1 1 1 0 0 1 0 0 ...
 $ job_req_communication : num [1:4870] 0 0 0 0 0 0 0 0 0 0 ...
 $ job_req_education     : num [1:4870] 0 0 0 0 0 0 0 0 0 0 ...
 $ job_req_min_experience: chr [1:4870] "5" "5" "5" "5" ...
 $ job_req_computer      : num [1:4870] 1 1 1 1 1 0 0 1 0 0 ...
 $ job_req_organization  : num [1:4870] 0 0 0 0 1 0 0 1 0 0 ...
 $ job_req_school        : chr [1:4870] "none_listed" "none_listed" "none_listed" "none_listed" ...
 $ received_callback     : num [1:4870] 0 0 0 0 0 0 0 0 0 0 ...
 $ firstname             : chr [1:4870] "Allison" "Kristen" "Lakisha" "Latonya" ...
 $ race                  : chr [1:4870] "white" "white" "black" "black" ...
 $ gender                : chr [1:4870] "f" "f" "f" "f" ...
 $ years_college         : int [1:4870] 4 3 4 3 3 4 4 3 4 4 ...
 $ college_degree        : num [1:4870] 1 0 1 0 0 1 1 0 1 1 ...
 $ honors                : int [1:4870] 0 0 0 0 0 1 0 0 0 0 ...
 $ worked_during_school  : int [1:4870] 0 1 1 0 1 0 1 0 0 1 ...
 $ years_experience      : int [1:4870] 6 6 6 6 22 6 5 21 3 6 ...
 $ computer_skills       : int [1:4870] 1 1 1 1 1 0 1 1 1 0 ...
 $ special_skills        : int [1:4870] 0 0 0 1 0 1 1 1 1 1 ...
 $ volunteer             : int [1:4870] 0 1 0 1 0 0 1 1 0 1 ...
 $ military              : int [1:4870] 0 1 0 0 0 0 0 0 0 0 ...
 $ employment_holes      : int [1:4870] 1 0 0 1 0 0 0 1 0 0 ...
 $ has_email_address     : int [1:4870] 0 1 0 1 1 0 1 1 0 1 ...
 $ resume_quality        : chr [1:4870] "low" "high" "low" "high" ...
```

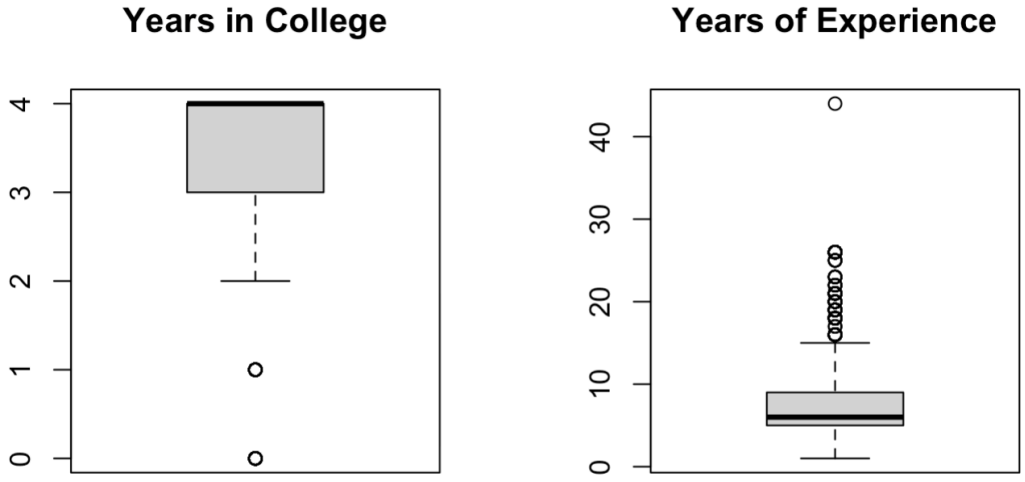
```
> resume$years_college <- as.numeric(resume$years_college)
> resume$honors <- as.numeric(resume$honors)
> resume$worked_during_school <- as.numeric(resume$worked_during_school)
> resume$works_during_school <- as.numeric(resume$worked_during_school)
> resume$worked_during_school <- as.numeric(resume$worked_during_school)
> resume$years_experience <- as.numeric(resume$years_experience)
> resume$computer_skills <- as.numeric(resume$computer_skills)
> resume$special_skills <- as.numeric(resume$special_skills)
> resume$volunteer <- as.numeric(resume$volunteer)
> resume$military <- as.numeric(resume$military)
> resume$employment_holes <- as.numeric(resume$employment_holes)
> resume$has_email_address <- as.numeric(resume$has_email_address)
```

```
> str(resume)
tibble [4,870 × 31] (S3: tbl_df/tbl/data.frame)
 $ job_ad_id             : num [1:4870] 384 384 384 384 385 386 386 385 386 386 ...
 $ job_city              : chr [1:4870] "Chicago" "Chicago" "Chicago" "Chicago" ...
 $ job_industry          : chr [1:4870] "manufacturing" "manufacturing" "manufacturing" "manufacturing" ...
 $ job_type              : chr [1:4870] "supervisor" "supervisor" "supervisor" "supervisor" ...
 $ job_fed_contractor    : num [1:4870] NA NA NA NA 0 0 0 0 0 0 ...
 $ job_equal_opp_employer: num [1:4870] 1 1 1 1 1 1 1 1 1 1 ...
 $ job_ownership         : chr [1:4870] "unknown" "unknown" "unknown" "unknown" ...
 $ job_req_any           : num [1:4870] 1 1 1 1 1 0 0 1 0 0 ...
 $ job_req_communication : num [1:4870] 0 0 0 0 0 0 0 0 0 0 ...
 $ job_req_education     : num [1:4870] 0 0 0 0 0 0 0 0 0 0 ...
 $ job_req_min_experience: chr [1:4870] "5" "5" "5" "5" ...
 $ job_req_computer      : num [1:4870] 1 1 1 1 1 0 0 1 0 0 ...
 $ job_req_organization  : num [1:4870] 0 0 0 0 1 0 0 1 0 0 ...
 $ job_req_school        : chr [1:4870] "none_listed" "none_listed" "none_listed" "none_listed" ...
 $ received_callback     : num [1:4870] 0 0 0 0 0 0 0 0 0 0 ...
 $ firstname             : chr [1:4870] "Allison" "Kristen" "Lakisha" "Latonya" ...
 $ race                  : chr [1:4870] "white" "white" "black" "black" ...
 $ gender                : chr [1:4870] "f" "f" "f" "f" ...
 $ years_college         : num [1:4870] 4 3 4 3 3 4 4 3 4 4 ...
 $ college_degree        : num [1:4870] 1 0 1 0 0 1 1 0 1 1 ...
 $ honors                : num [1:4870] 0 0 0 0 0 1 0 0 0 0 ...
 $ worked_during_school  : num [1:4870] 0 1 1 0 1 0 1 0 0 1 ...
 $ years_experience      : num [1:4870] 6 6 6 6 22 6 5 21 3 6 ...
 $ computer_skills       : num [1:4870] 1 1 1 1 1 0 1 1 1 0 ...
 $ special_skills        : num [1:4870] 0 0 0 1 0 1 1 1 1 1 ...
 $ volunteer             : num [1:4870] 0 1 0 1 0 0 1 1 0 1 ...
 $ military              : num [1:4870] 0 1 0 0 0 0 0 0 0 0 ...
 $ employment_holes      : num [1:4870] 1 0 0 1 0 0 0 1 0 0 ...
 $ has_email_address     : num [1:4870] 0 1 0 1 1 0 1 1 0 1 ...
 $ resume_quality        : chr [1:4870] "low" "high" "low" "high" ...
 $ works_during_school   : num [1:4870] 0 1 1 0 1 0 1 0 0 1 ...
```

Outliers

Finally, I wanted to determine if there existed any outliers and if they should be removed. Intuitively, the only predictors that could contain outliers would be years_college and years_experience.

```
> par(mfrow=c(1,2))
> boxplot(resume$years_college, main="Years in College")
> boxplot(resume$years_experience, main="Years of Experience")
```



It is clear that most values fall within range for the years_college predictor. The two that do not will not have much effect on the analysis. For years_experiene, however, there seem to be quite a large amount of observations in which the applicant listed many more years of experience on their resume as compared to other applicants. This is good information to have moving forward, especially when analyzing the relation between years of experience and race. These outliers will not be removed from the dataset, but will serve as a tool to help us understand if having more years of experience can overcome racial bias.

Variable Description

I obtained descriptions of the variables from R-Project at https://vincentarelbundock.github.io/Rdatasets/doc/openintro/resume.html. I have created a table including the variable name, its description, and mode. The modes included in this table are representative of the original dataset, and all integer variables have been changed to numeric ones.

Variable Name	Description	Mode
job_ad_id	Job advertisement ID	numeric
job_city	Location of job	character
job_industry	Industry of job	character
job_type	Role title	character
job_fed_contractor	Indicates if employer is a federal contractor	numeric
job_equal_opp_employer	Indicates if employer is an equal opportunity employer	numeric
job_ownership	Indicates if company is a nonprofit or private company	character
job_req_any	Indicates if job requirements are listed in posting	numeric
job_req_communication	Indicates if communication skills are required	numeric
job_req_education	Indicates if some level of education is required	numeric
job_req_min_experience	Amount of required experience	character
job_req_computer	Indicates if computer skills are required	numeric

job_req_organization	Indicates if organization skills are required	numeric
job_req_school	Level of education required	character
received_callback	Indicates if applicant received a callback	numeric
firstname	Named used on resume	character
race	Inferred race of applicant	character
gender	Inferred gender of applicant	character
years_college	Years of education listed on resume	integer
college_degree	Indicates if the resume lists a college degree	numeric
honors	Indicates if resume lists honors	integer
worked_during_school	Indicates if resume lists working while in school	integer
years_experience	Years of experience listed on resume	integer
computer_skills	Indicates if computer skills were listed on resume	integer
special_skills	Indicates if any special skills were listed on resume	integer
volunteer	Indicates if volunteer experience was listed on resume	integer
military	Indicates if military experience was listed on resume	integer
employment_holes	Indicates if gaps in employment appeared on resume	integer
has_email_address	Indicates if email address was provided on resume	integer
resume_quality	Classification of resume as either “low” or “high” quality	integer

Purpose and Expectations

The primary purpose of this project is to determine whether race and/or gender has an effect on the amount of callbacks one receives from job applications. Using this dataset, I plan to uncover other connections. For example, can one's years of experience overcome racial bias of job recruiters? How much does gender actually impact one's ability to get an interview? Do white men with less experience and lower quality resumes still get more callbacks than women of color with the necessary qualifications? This dataset has the potential to answer all of these questions and more.

My educated hypothesis is that racial and gender bias play a large role in the hiring process, even more so than qualifications. I expect that individuals with stereotypical "white sounding" names will, overall, receive more callbacks than those with stereotypical "black sounding" names. I also expect that individuals with names that sound more masculine will receive more callbacks and more feminine-sounding names, overall. I do expect experience, skills, education, and quality of the resume to increase one's chances of receiving a callback; however, I do not expect these variables to overcome racial or gender bias. I also do not expect to see much correlation between the applicant's address and their chances of receiving a callback, as I do not believe recruiters spend enough time looking at resumes to notice if the applicant lives in a "bad" or "good" area.

One thing to note is that this study was published in 2003. We can argue that our society, as a whole, is more accepting than it was twenty years ago. In some ways, I believe that is true. However, racial and gender biases and discrimination do persist today. Because of this, I believe the outcome of this research to still be relevant and pertinent in creating equal opportunities for all.

Data Analysis

Categorical Variable Analysis

Barplots

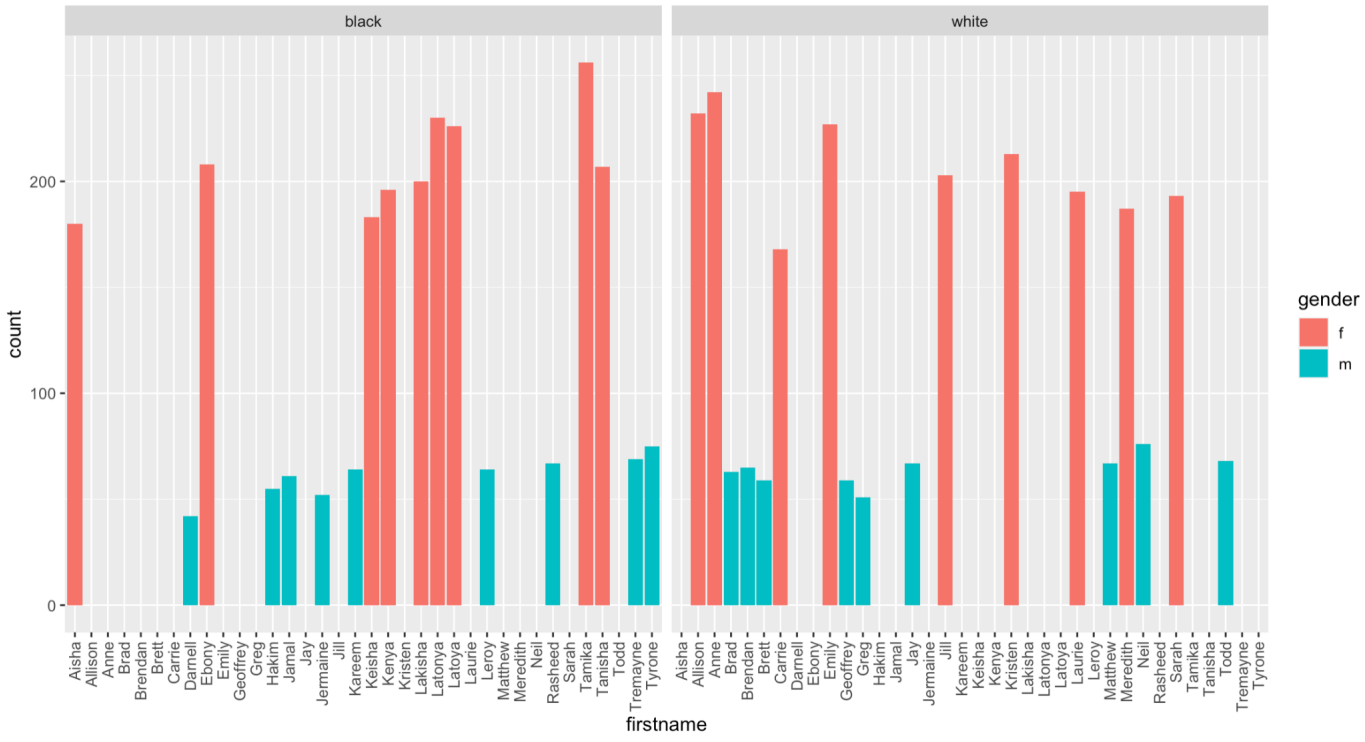
The first step in any analysis should be to first become comfortable with the data and to get a better understanding of how it is distributed among variables. For this dataset, that meant looking at the splits between race and gender. I used the table() function to do this. We can see that the race of the applicants is split evenly between white and black. However, there appears to be about three times as many female applicants as there are male. I will keep this in mind moving forward.

```
> table(resume$race)
black white
 2435  2435
```

```
> table(resume$gender)
   f    m
3746 1124
```

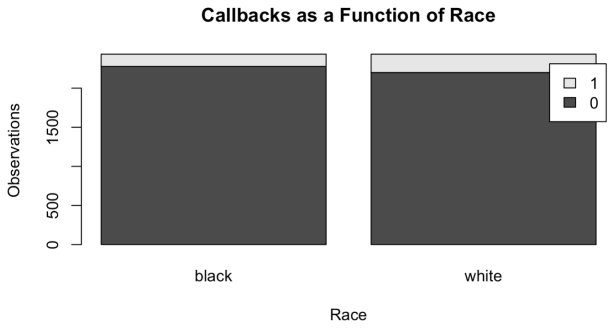
Next, I wanted to visualize the data in terms of both gender and race. To do this, I created a barplot that grouped the names into these two categories. We can see that the female names were used more (as the table also shows). I think it is also interesting to note the actual names used.

```
> ggplot(resume, aes(firstname, fill=gender))+geom_bar(position="dodge") + facet_wrap(~ race)
+ theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

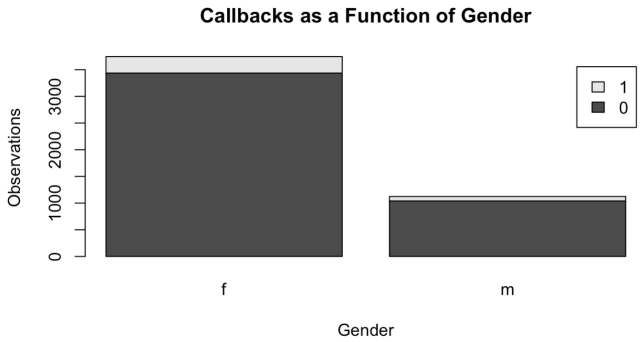


Finally, I wanted to visualize the amount of callbacks as a function of race, gender, years of experience, and resume quality. I felt like this would be a good place to start as I hypothesized that individuals would receive more callbacks if they had a “white-male-sounding” name, had more years of experience, and had a better resume. I created barplots for these predictors separately for now, as further analysis later on will provide me the opportunity to determine the relationship between more than two predictors at a time.

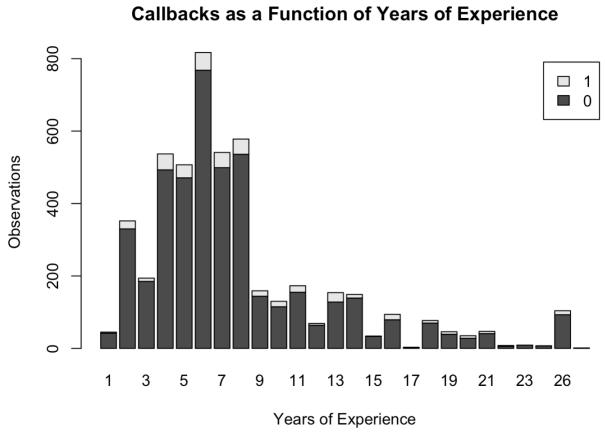
```
> barplot(table(resume$received_callback, resume$race), legend.text=TRUE, xlab="Race", ylab="Observations", main="Callbacks as a Function of Race")
```



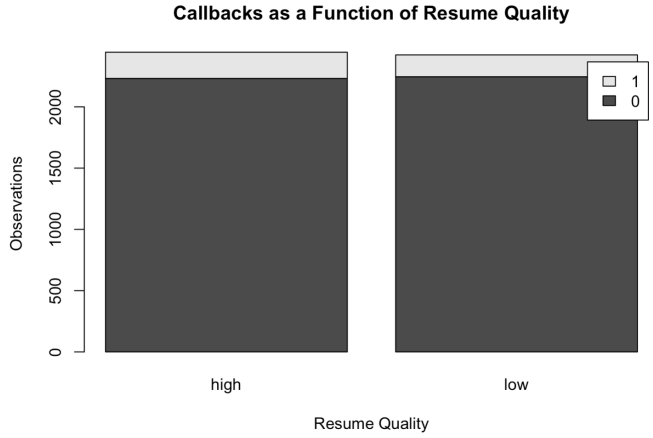

```
> barplot(table(resume$received_callback, resume$gender), legend.text=TRUE, xlab="Gender", ylab="Observations", main="Callbacks as a Function of Gender")
```




```
> barplot(table(resume$received_callback, resume$years_experience), legend.text=TRUE, xlab="Years of Experience", ylab="Observations", main="Callbacks as a Function of Years of Experience")
```




```
> barplot(table(resume$received_callback, resume$resume_quality), legend.text=TRUE, xlab="Resume Quality", ylab="Observations", main="Callbacks as a Function of Resume Quality")
```



Please note that 1 means the applicant received a callback, whereas 0 means they did not. As we can see above, the white applicants received only slightly more callbacks than the black applicants. The ratio of gender to callbacks is about equal for both men and women. While it may seem that an individual receives more callbacks if they have more years of experience, the ratio is, again, about equal when taking into account the number of observations. Finally, the number of callbacks is nearly the same for both high- and low-quality resumes. Already I can see that my hypothesis is off. I can tell that the likelihood of an individual receiving a callback is not simply a function of one variable but will most probably be a function of multiple variables at the same time. It is at this point that I change my hypothesis. I think that, if an individual has multiple compounding variables (e.g. female, black, only a few years of experience, and low-quality resume), then their likelihood of receiving a callback decreases. I think that the more "negative" traits this individual has, the less likely they are to receive a callback. It is clear that logistic regression will be useful in this analysis.

Heatmaps

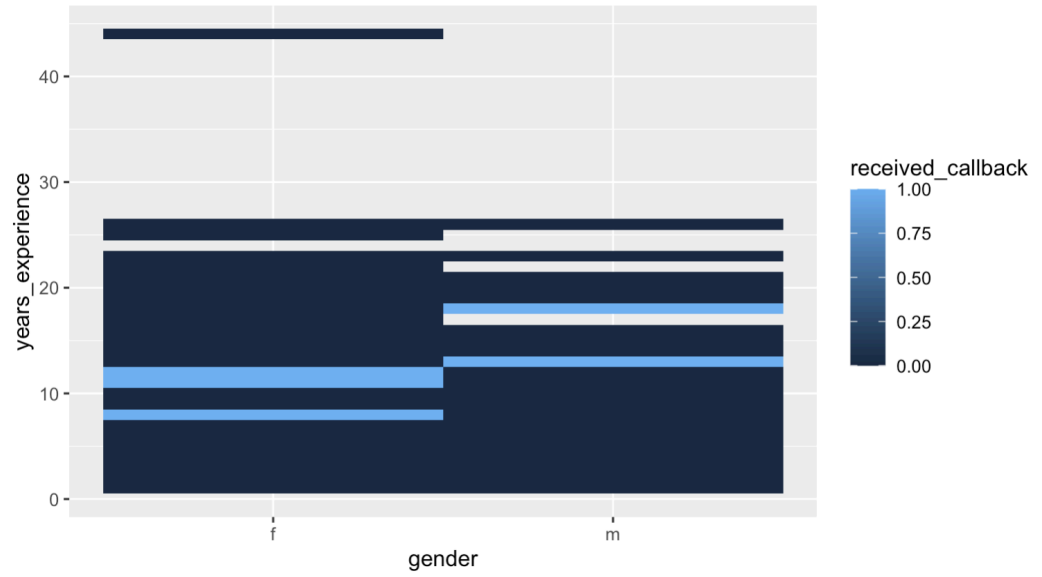
To further familiarize myself with the data and also to corroborate the above results, I created a few heatmaps. I wanted to look at recieved_callbacks as a function of the following predictors:

gender x race
gender x years_experience
gender x resume_quality
race x years_experience
race x resume_quality
years_experience x resume_quality

What I found was that (gender x race), (gender x resume_quality), and (race x resume_quality) did not produce any interpretable results. Below are the results of the remaining functions.

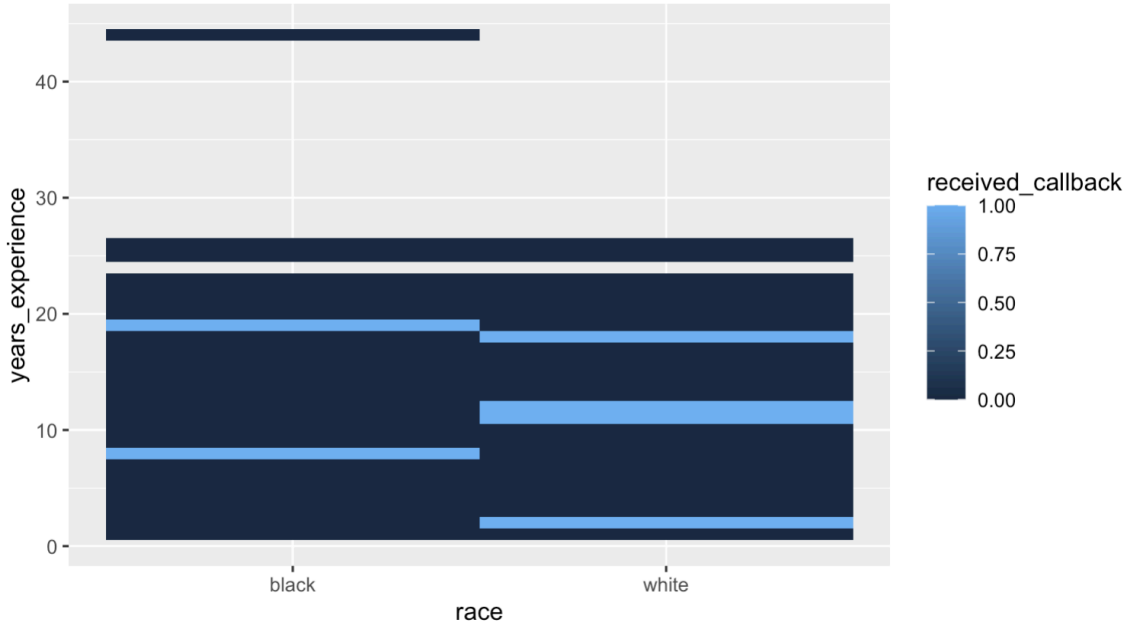
Gender x Years of Experience

```
> ggplot(resume, aes(x=gender, y=years_experience, fill=received_callback)) +geom_tile()
```



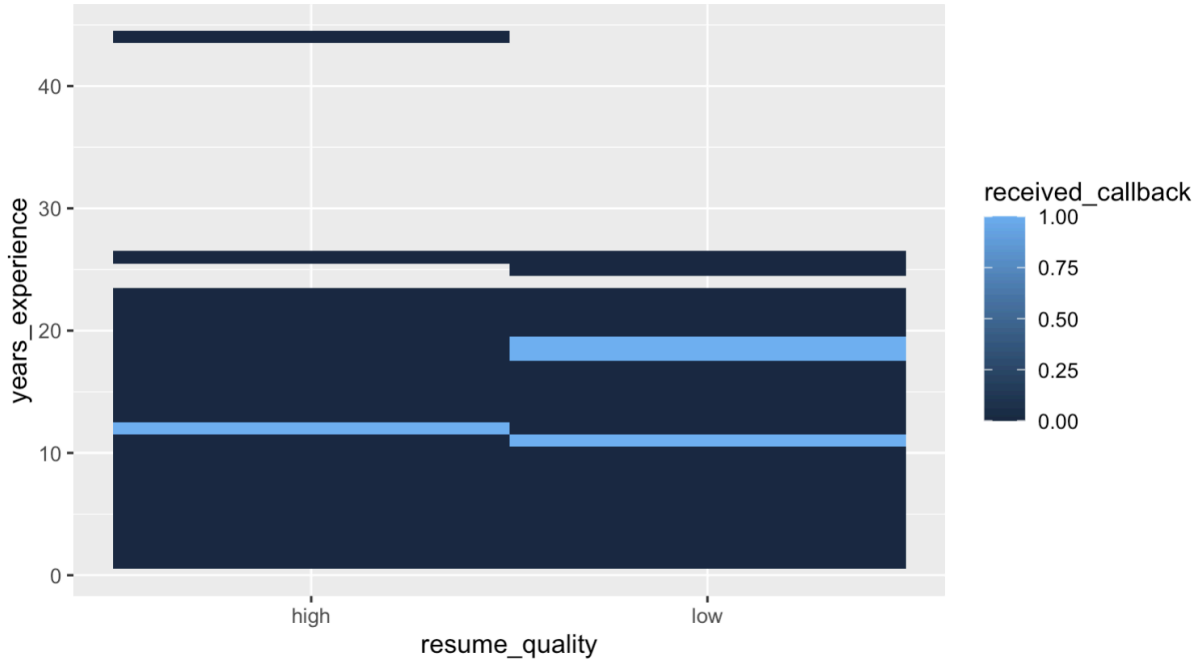
Race x Years of Experience

```
> ggplot(resume, aes(x=race, y=years_experience, fill=received_callback)) +geom_tile()
```



Resume Quality x Years of Experience

```
> ggplot(resume, aes(x=resume_quality, y=years_experience, fill=received_callback)) +geom_tile()
```



What is interesting here is that the data shows that, for the same number of years of experience, women are more likely to receive a callback than men. As expected, though, white individuals are more likely to receive a callback despite having less experience than their black counterparts. Finally, it seems as though resume quality doesn't have a huge impact at all on the likelihood that one might receive a callback. In fact, having more years of experience tends to make up for a low-quality resume.

Numerical Variable Analysis

Correlation Matrices

My first step in analyzing the numerical data was to view the correlation matrices. First, I had to create a dummy variable for gender, race, and resume quality. For gender, the applicant was assigned a 1 if they were female and a 0 if they were male. For race, they were assigned a 1 if they were black and a 0 if they were white. For resume quality, they were assigned a 1 if their resume was of high quality and a 0 if it was of low quality. Next, I created a new data frame with these new variables and removed the unnecessary ones or ones I couldn't use (name, job_req_school, job_req_min_experience, job_ownership, job_type, job_industry, job_city, and job_ad_id). Finally, I removed all NA values. The results are shown below, with the highlighted areas being the variables that had significant correlations.

Converting gender and race to numeric:

```
> resume$gender_numeric <- ifelse(resume$gender == "f", 1, 0)
> resume$race_numeric <- ifelse(resum$race == "black", 1, 0)
> resume$quality_numeric <- ifelse(resume$resume_quality == "high", 1, 0)
```

Removing columns for correlation:

```
> resume_cor <- resume[,-30][,-18][,-17][,-16][,-14][,-11][,-7][,-4][,-3][,-2][,-1]
```

Correlation Matrix without NAs

```
> cor(resume_cor,use="complete.obs")
```

	job_fed_contractor	job_equal_opp_employer	job_req_any	job_req_communication	job_req_education	job_req_computer
job_fed_contractor	1.000000e+00	1.241234e-01	-3.549769e-02	0.0134374451	6.667620e-02	0.0075785817
job_equal_opp_employer	1.241234e-01	1.000000e+00	1.106209e-01	-0.0318230831	2.254438e-01	0.0504381868
job_req_any	-3.549769e-02	1.106209e-01	1.000000e+00	0.1833288704	1.841224e-01	0.4137666914
job_req_communication	1.343745e-02	-3.182308e-02	1.833289e-01	1.0000000000	-4.374490e-02	0.1232040801
job_req_education	6.667620e-02	2.254438e-01	1.841224e-01	-0.0437449026	1.000000e+00	-0.0196660867
job_req_computer	7.578582e-03	5.043819e-02	4.137667e-01	0.1232040801	-1.966609e-02	1.0000000000
job_req_organization	-2.699570e-02	-5.505104e-02	1.242795e-01	0.4002626867	-2.085138e-02	0.0968235612
received_callback	1.447112e-02	1.737151e-02	-2.755005e-02	0.0169583338	-2.013631e-02	0.0162234320
years_college	2.142231e-02	8.086355e-02	1.093440e-02	-0.0036711000	7.792352e-02	-0.0537820472
college_degree	2.369684e-02	9.089458e-02	-2.350921e-02	-0.0350004684	9.150543e-02	-0.1662503377
honors	9.301360e-03	-2.871748e-02	6.436581e-03	0.0166968279	-1.185552e-03	0.0244321585
worked_during_school	1.256857e-02	2.787740e-02	5.421573e-02	0.0145776695	7.452894e-03	0.0738370570
years_experience	-3.097144e-02	5.233821e-02	7.835336e-02	0.0075772004	3.018771e-02	0.0229018936
computer_skills	-1.767589e-02	2.914448e-02	1.713048e-01	0.0903911276	4.044116e-02	0.2908400460
special_skills	-1.124762e-02	-6.828365e-02	-5.959804e-02	-0.0213808615	-3.772063e-02	0.0563263222
volunteer	1.615504e-02	3.155362e-02	-4.573399e-02	-0.0418185950	8.443166e-03	-0.0587779336
military	-9.449360e-03	-1.881524e-02	-2.660915e-02	-0.0220331029	-6.711897e-03	-0.0847609449
employment_holes	-1.093812e-02	-8.180831e-02	3.703442e-02	0.1003101750	-3.685198e-02	0.1709612032
has_email_address	1.461674e-02	-1.914430e-02	-1.091928e-02	0.0133202172	-1.642454e-02	0.0647080496
works_during_school	1.256857e-02	2.787740e-02	5.421573e-02	0.0145776695	7.452894e-03	0.0738370570
gender_numeric	1.521026e-02	-2.155460e-02	1.577626e-01	0.0996630960	-4.989938e-03	0.3583702480
race_numeric	7.018026e-19	5.938794e-19	-2.573710e-18	-0.0009649862	3.203751e-19	0.0006515764
quality_numeric	6.466271e-03	1.528208e-02	-2.807823e-03	-0.0026940251	2.111103e-03	0.0058713893

	job_req_organization	received_callback	years_college	college_degree	honors	worked_during_school
job_fed_contractor	-0.026995701	0.01447112	0.021422311	0.023696844	0.009301360	0.012568572
job_equal_opp_employer	-0.055051040	0.01737151	0.080863547	0.090894584	-0.028717476	0.027877400
job_req_any	0.124279493	-0.02755005	0.010934404	-0.023509209	0.006436581	0.054215727
job_req_communication	0.400262687	0.01695833	-0.003671100	-0.035000468	0.016696828	0.014577670
job_req_education	-0.020851379	-0.02013631	0.077923522	0.091505427	-0.001185552	0.007452894
job_req_computer	0.096823561	0.01622343	-0.053782047	-0.166250338	0.024432159	0.073837057
job_req_organization	1.000000000	-0.04476689	0.001315197	-0.019977818	0.003391860	0.025684254
received_callback	-0.044766888	1.00000000	0.011315197	0.011355759	0.083243827	-0.034244599
years_college	0.001315197	0.01131520	1.000000000	0.852272682	0.044613300	0.146267278
college_degree	-0.019977818	0.01135576	0.852272682	1.000000000	0.043263453	0.052314089
honors	0.003391860	0.08324383	0.044613300	0.043263453	1.000000000	-0.055540253
worked_during_school	0.025684254	-0.03424460	0.146267278	0.052314089	-0.055540253	1.000000000
years_experience	0.003797400	0.05016915	0.028221867	-0.012265511	0.126317509	-0.170571715
computer_skills	0.087235898	-0.02149785	0.007618168	-0.051487228	-0.039877823	0.321052668
special_skills	-0.008144611	0.10394436	-0.180183154	-0.191945228	0.092555450	-0.168224384
volunteer	-0.041981230	-0.01458327	-0.019403893	-0.034420306	-0.001671085	0.287340865
military	-0.031889634	-0.01696370	0.035974597	0.003488836	0.009111093	0.171400218
employment_holes	0.073041652	0.06858143	-0.119306121	-0.136039535	0.163794487	-0.482101805
has_email_address	0.017581496	0.02260912	-0.028745866	-0.077575587	0.075494873	0.313951206
works_during_school	0.025684254	-0.03424460	0.146267278	0.052314089	-0.055540253	1.000000000
gender_numeric	0.121255273	0.02307727	-0.082915912	-0.189850164	0.032880110	0.036050138
race_numeric	0.000000000	-0.05946475	0.003132587	0.005131329	-0.027293527	0.012319223
quality_numeric	0.004128038	0.02175991	0.028695907	-0.003907733	0.063573476	0.323713459

	years_experience	computer_skills	special_skills	volunteer	military	employment_holes	has_email_address
job_fed_contractor	-0.03097144	-0.017675889	-0.011247623	0.016155043	-0.009449360	-0.01093812	0.014616736
job_equal_opp_employer	0.05233821	0.029144476	-0.068283648	0.031553620	-0.018815240	-0.08180831	-0.019144295
job_req_any	0.07835336	0.171304798	-0.059598043	-0.045733988	-0.026609150	0.03703442	-0.010919282
job_req_communication	0.00757720	0.090391128	-0.021380862	-0.041818595	-0.022033103	0.10031017	0.013320217
job_req_education	0.03018771	0.040441158	-0.037720628	0.008443166	-0.006711897	-0.03685198	-0.016424538
job_req_computer	0.02290189	0.290840046	0.056326322	-0.058777934	-0.084760945	0.17096120	0.064708050
job_req_organization	0.00379740	0.087235898	-0.008144611	-0.041981230	-0.031889634	0.07304165	0.017581496
received_callback	0.05016915	-0.021497847	0.103944359	-0.014583274	-0.016963698	0.06858143	0.022609118
years_college	0.02822187	0.007618168	-0.180183154	-0.019403893	0.035974597	-0.11930612	-0.028745866
college_degree	-0.01226551	-0.051487228	-0.191945228	-0.034420306	0.003488836	-0.13603954	-0.077575587
honors	0.12631751	-0.039877823	0.092555450	-0.001671085	0.009111093	0.16379449	0.075494873
worked_during_school	-0.17057171	0.321052668	-0.168224384	0.287340865	0.171400218	-0.48210180	0.313951206
years_experience	1.00000000	-0.079173052	-0.035875705	0.032853692	-0.245085580	0.06526017	-0.020401814
computer_skills	-0.07917305	1.000000000	0.012395550	0.169756496	0.106540030	-0.09368960	0.228171354
special_skills	-0.03587571	0.012395550	1.000000000	0.016006409	-0.091170906	0.19268731	0.077546153
volunteer	0.03285369	0.169756496	0.016006409	1.000000000	0.284175981	-0.19346444	0.684653834
military	-0.24508558	0.106540030	-0.091170906	0.284175981	1.000000000	-0.16699205	0.341946208
employment_holes	0.06526017	-0.093689601	0.192687312	-0.193464442	-0.166992050	1.00000000	-0.124754505
has_email_address	-0.02040181	0.228171354	0.077546153	0.684653834	0.341946208	-0.12475450	1.000000000
works_during_school	-0.17057171	0.321052668	-0.168224384	0.287340865	0.171400218	-0.48210180	0.313951206
gender_numeric	0.02318817	0.211282985	0.082593316	-0.052925856	-0.092084750	0.15916226	0.044284168
race_numeric	-0.01466630	0.034235844	-0.008978412	0.003916843	0.029318450	-0.01827483	0.003874857
quality_numeric	0.08368092	0.233716596	0.083230280	0.795190353	0.318849526	-0.20880011	0.883089906

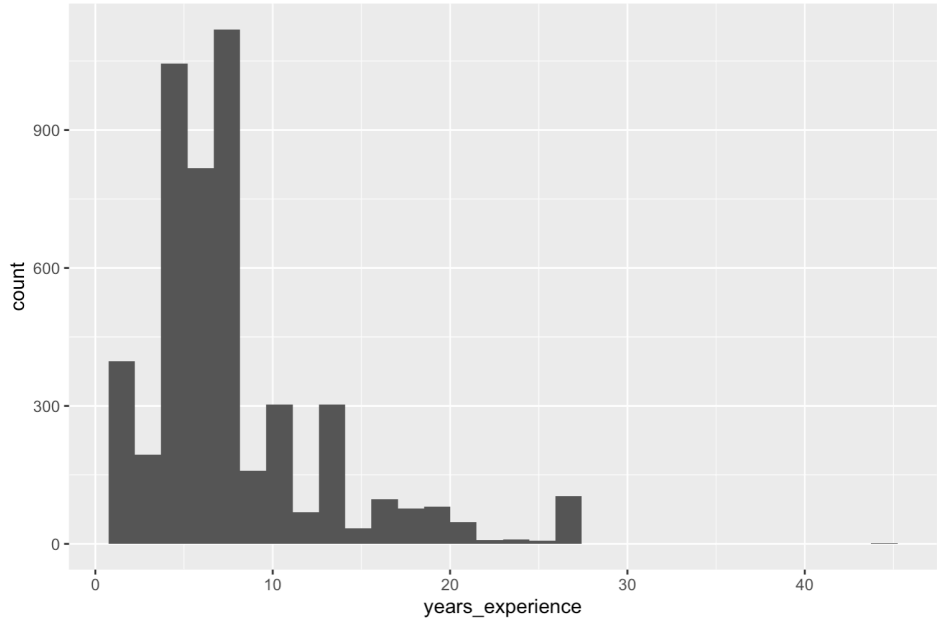
	works_during_school	gender_numeric	race_numeric	quality_numeric
job_fed_contractor	0.012568572	0.015210259	7.018026e-19	6.466271e-03
job_equal_opp_employer	0.027877400	-0.021554600	5.938794e-19	1.528208e-02
job_req_any	0.054215727	0.157762580	-2.573710e-18	-2.807823e-03
job_req_communication	0.014577670	0.099663096	-9.649862e-04	-2.694025e-03
job_req_education	0.007452894	-0.004989938	3.203751e-19	2.111103e-03
job_req_computer	0.073837057	0.358370248	6.515764e-04	5.871389e-03
job_req_organization	0.025684254	0.121255273	0.000000e+00	4.128038e-03
received_callback	-0.034244599	0.023077268	-5.946475e-02	2.175991e-02
years_college	0.146267278	-0.082915912	3.132587e-03	2.869591e-02
college_degree	0.052314089	-0.189850164	5.131329e-03	-3.907733e-03
honors	-0.055540253	0.032880110	-2.729353e-02	6.357348e-02
worked_during_school	1.000000000	0.036050138	1.231922e-02	3.237135e-01
years_experience	-0.170571715	0.023188175	-1.466630e-02	8.368092e-02
computer_skills	0.321052668	0.211282985	3.423584e-02	2.337166e-01
special_skills	-0.168224384	0.082593316	-8.978412e-03	8.323028e-02
volunteer	0.287340865	-0.052925856	3.916843e-03	7.951904e-01
military	0.171400218	-0.092084750	2.931845e-02	3.188495e-01
employment_holes	-0.482101805	0.159162256	-1.827483e-02	-2.088001e-01
has_email_address	0.313951206	0.044284168	3.874857e-03	8.830899e-01
works_during_school	1.000000000	0.036050138	1.231922e-02	3.237135e-01
gender_numeric	0.036050138	1.000000000	2.442798e-02	-6.987268e-03
race_numeric	0.012319223	0.024427984	1.000000e+00	1.145309e-18
quality_numeric	0.323713459	-0.006987268	1.145309e-18	1.000000e+00

As we can see above, there were quite a few variables that were correlated to one another. However, there were only a few that were *significantly* correlated. Those are (years_college x college_degree), (volunteer x has_email_address), (quality_numeric x volunteer), and (quality_numeric x email). We can ignore the first one - (years_college x college_degree) - since it intuitively makes sense that the more years someone would attend college, the more likely they are to have a college degree. The next - (volunteer x has_email_address) - also doesn't tell us much other than the fact that someone is more likely to have an email address on their resume if they have volunteer experience. The last two tell us a bit more about how recruiters view resumes. We see that a resume with an email address on it is more likely to be indicated as high quality. We also see that those with volunteer experience listed tend to have higher quality resumes as well. We also see that those that list military experience, computer skills, and working during school on their resume have higher quality resumes than those that do not. What this matrix does not tell us, though - and what I hope to find out - is if the quality of one's resume can overcome gender and racial bias.

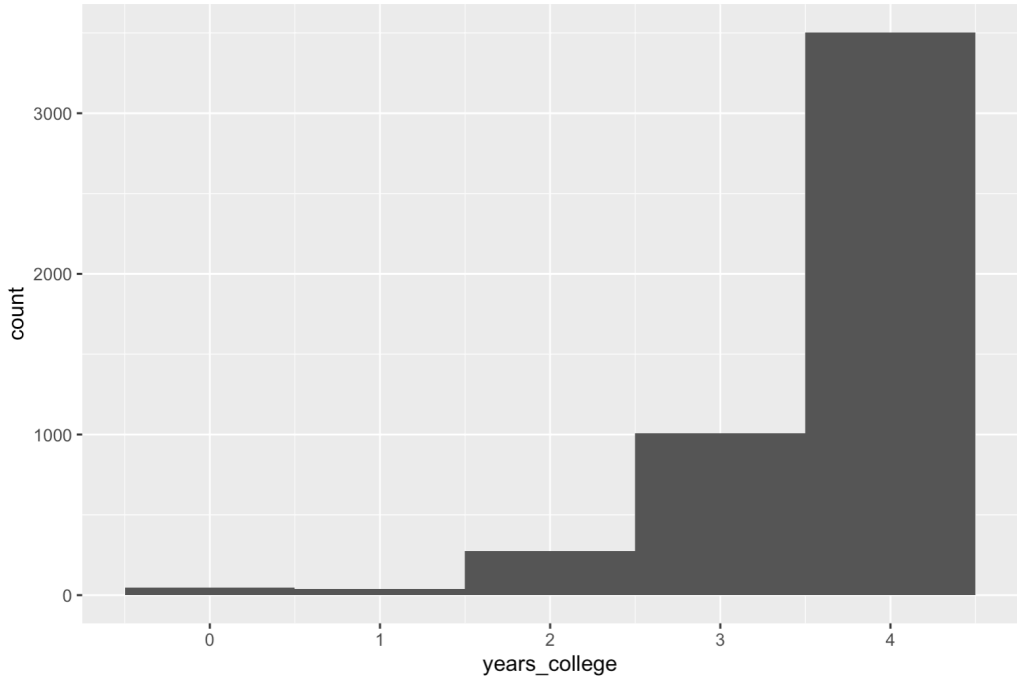
Histograms

So far, I have talked a bit about the distribution of certain variables. However, I have not shown them directly. Below are histograms for the continuous variables in this dataset. Both years_experience and years_college were simple enough to plot as they contained no NA values. I did want to see the distribution of job_req_min_experience, too, though. The first issue that I ran into was that, in the original dataset, this is a character variable as some values are listed as "some". The second issue was that there were quite a bit of NA values. To remedy this, I simply went into the dataset in Excel and removed all of the values that were not numerical. Doing this brought the original dataset of 4,870 observations down to 1,060 observations. That is a significant loss in data. However, I only wanted to see the distribution for the available values for this singular variable, so I created a separate file and used it only for creating the third histogram.

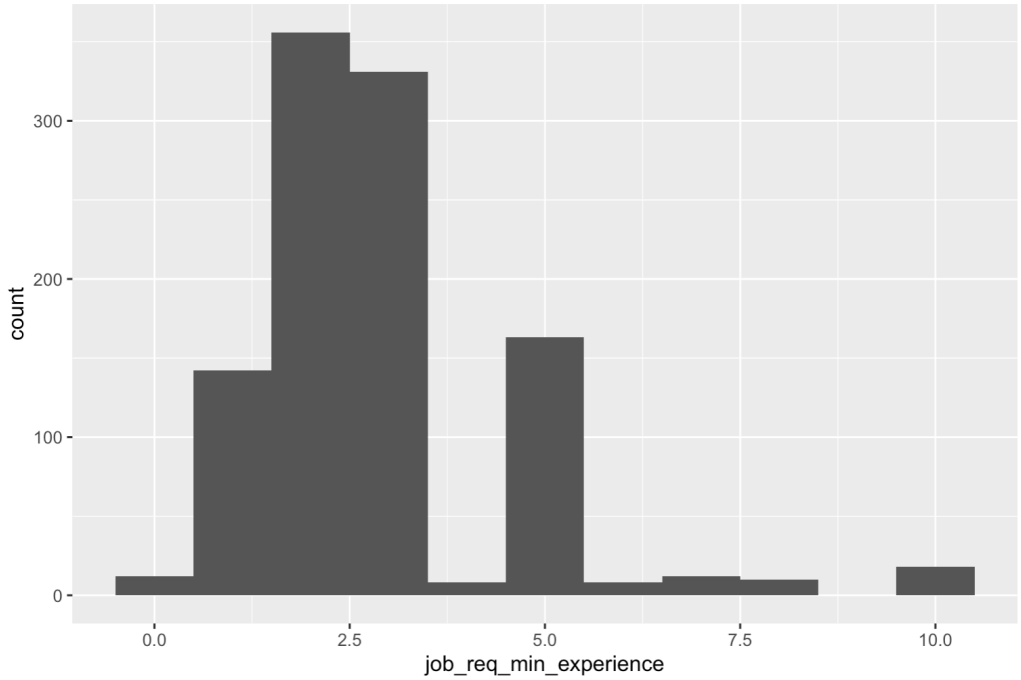
```
> ggplot(resume, aes(years_experience)) + geom_histogram()
```



```
> ggplot(resume, aes(years_college)) + geom_histogram(binwidth=1)
```




```
> resume_jobminexp <- read.csv("~/Desktop/resume.jobminexp.csv")
> ggplot(resume_jobminexp, aes(job_req_min_experience)) +geom_histogram(binwidth = 1)
```



We can see that the years of experience listed on applicant resumes are positively skewed, with most applicants having around 3-7 years of experience. The years of college listed on applicant resumes are negatively skewed. This makes sense intuitively, as most individuals will

have a bachelor degree. What is interesting to note, though, is that no applicant listed more than four years of college on their resume. Finally, we see that job_req_min_experience is, overall, positively skewed.

Density Graphs

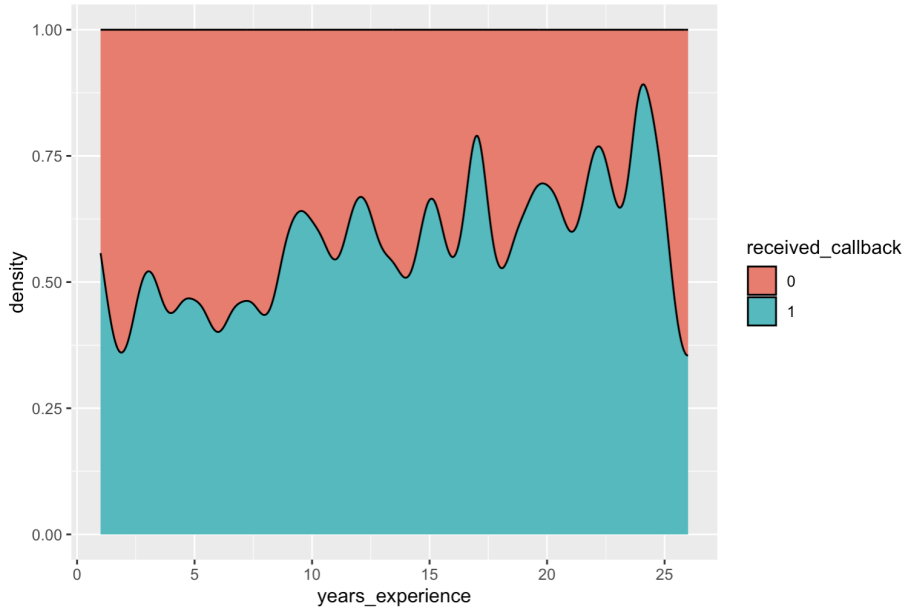
My main goal in creating the following density graphs was to visualize the following relationships:

Job industry/type and required years of experience
Job industry/type and listed years of experience

To begin, I wanted to visualize the overall density of callbacks as they relate to years of experience. First, I needed to create a file that removed the singular outlier of 44 years of experience. I did that in excel and loaded the file into R. I also needed to transform the received_callback variable to a factor, so I did that as well. We can see that the rate of callbacks does slightly increase with years of experience. I keep this graph in mind as I create and analyze the next density graphs.

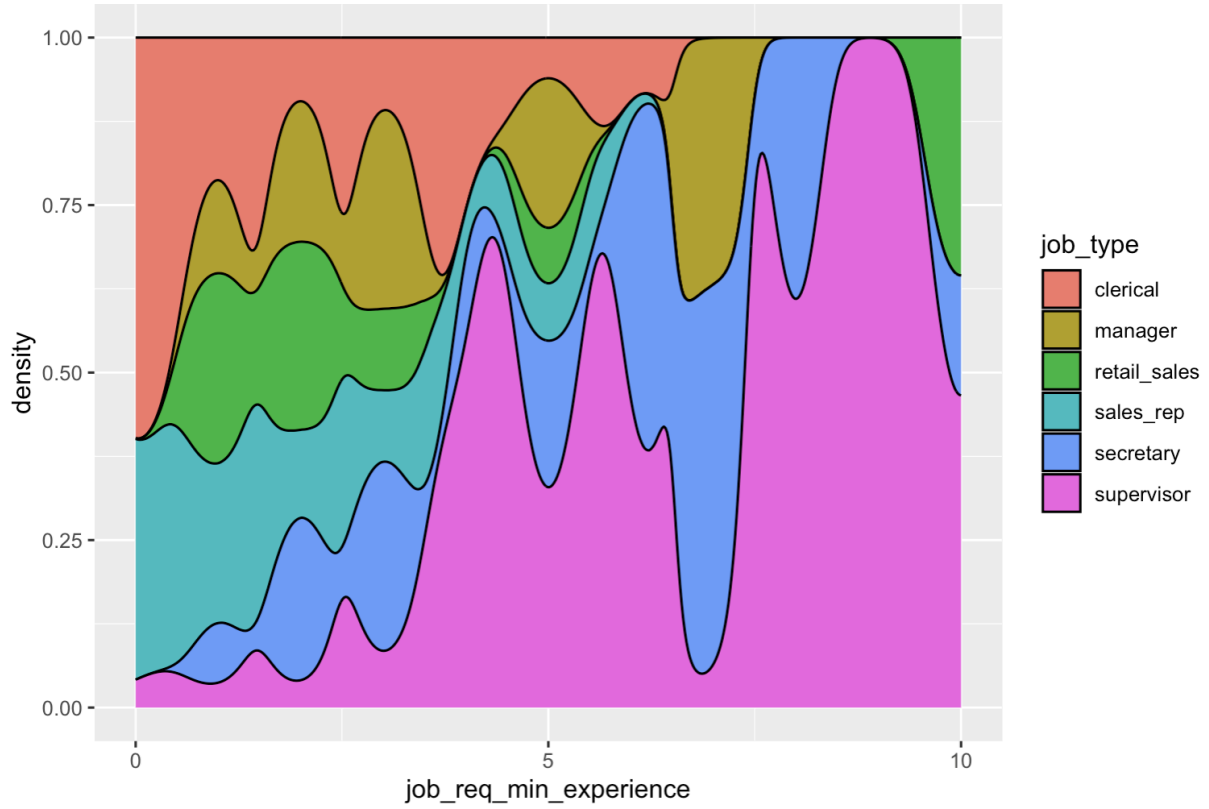
years_experience vs. received_callback

```
> resume_minus_44years <- read.csv("~/Desktop/resume_minus_44years.csv")
> resume_minus_44years$received_callback <-
as.factor(resume_minus_44years$received_callback)
> ggplot(resume_minus_44years, aes(x=years_experience, fill=received_callback)) +
geom_density(position="fill") + scale_x_continuous(breaks=c(0,5,10,15,20,25))
```



job_req_min_experience vs. job_type

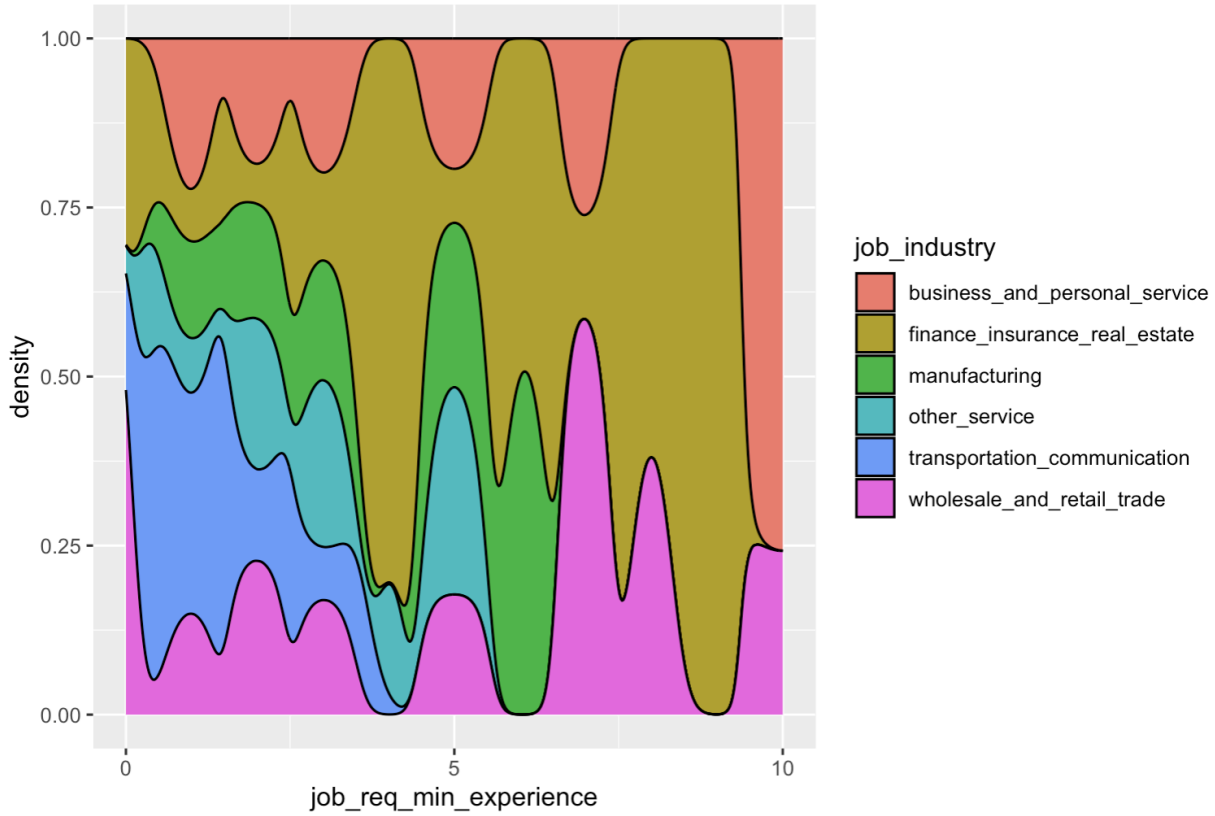
```
> ggplot(resume_jobminexp, aes(x=job_req_min_experience, fill=job_type)) +
geom_density(position="fill") + scale_x_continuous(breaks = c(0, 5, 10))
```



For jobs like *supervisor* and *manager,* it makes sense that they would require more years of experience. What is surprising to me, though, is that jobs like *secretary* also seem to require a high number of years of experience. As we will see in the next graph, this can be explained by a third variable - *job industry*.

job_req_min_experience vs. job_industry

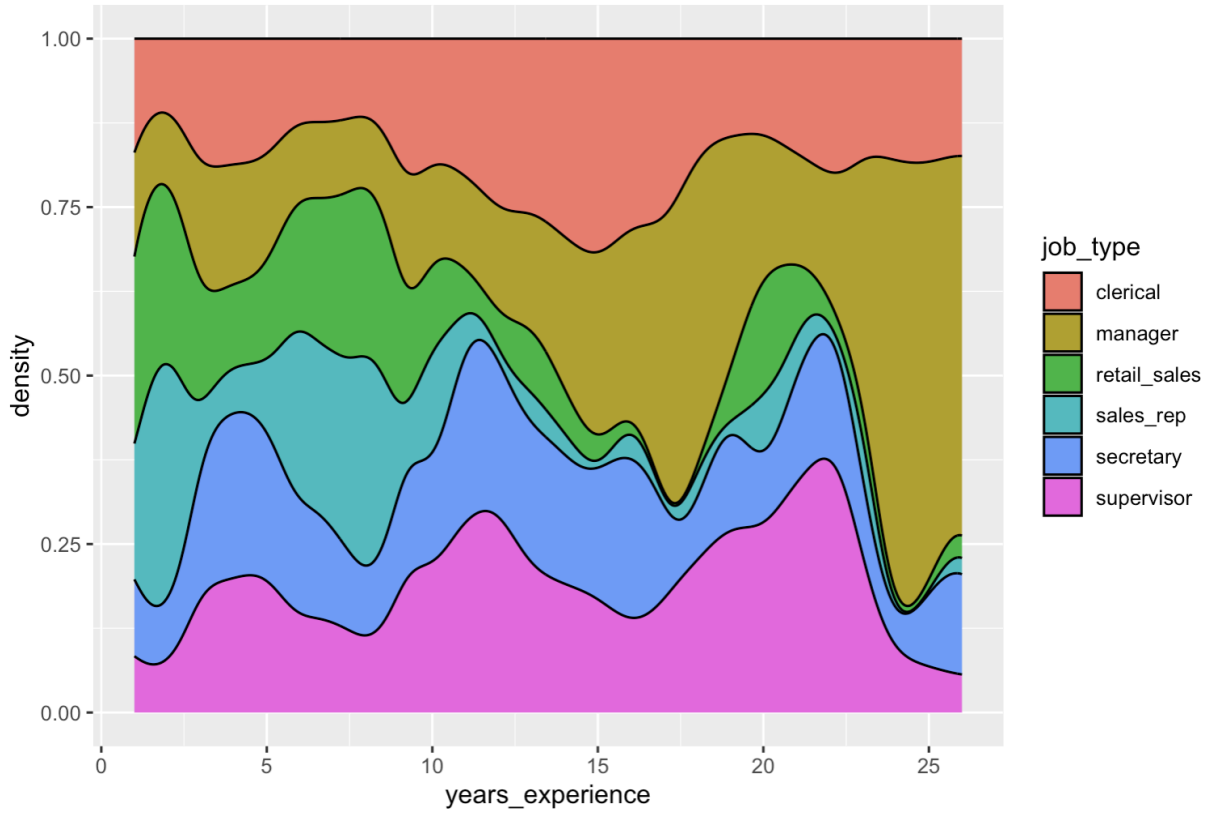
```
ggplot(resume_jobminexp, aes(x=job_req_min_experience, fill=job_industry)) +
geom_density(position="fill") + scale_x_continuous(breaks = c(0, 5, 10))
```



Recruiters in the business and finance sector are requiring more experience for all positions than those in other fields. We can infer that a manager at a restaurant might be required to have the same amount of years of experience as a secretary at a finance or insurance firm.

years_experience vs. job_type

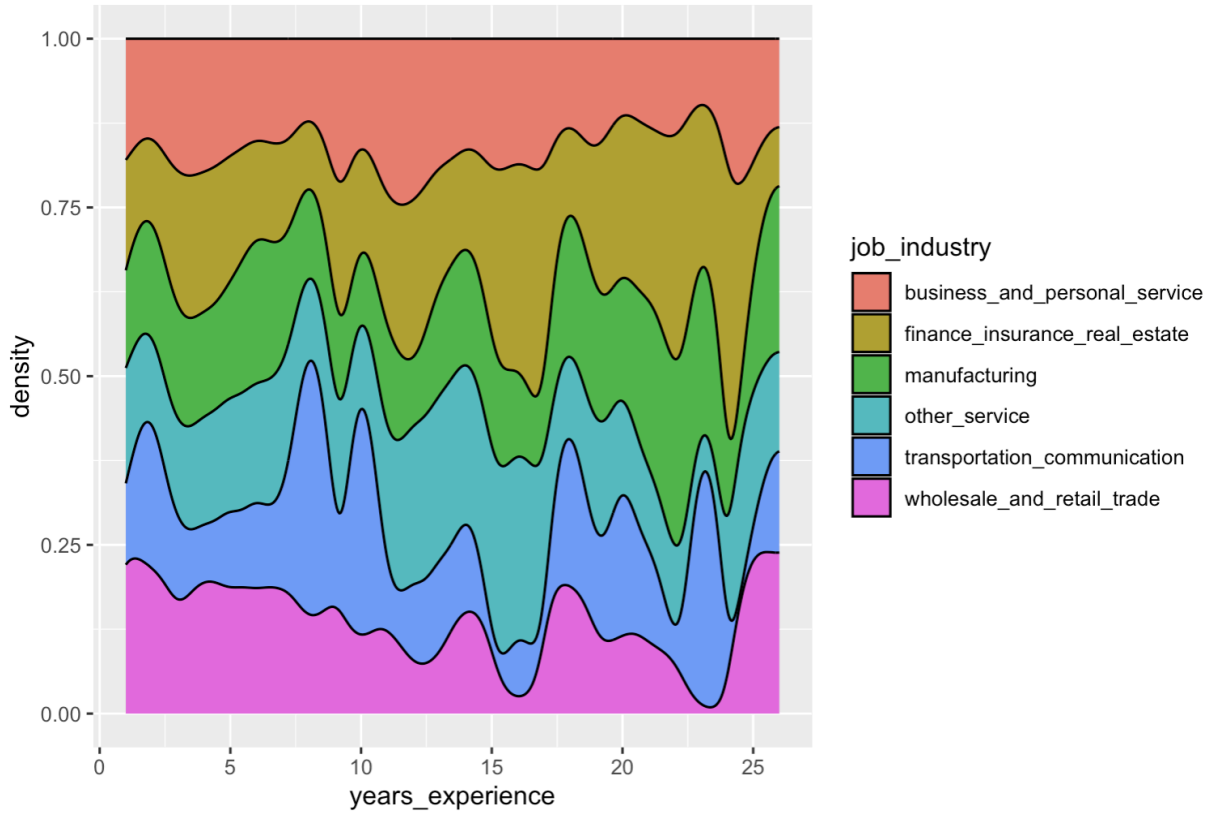
```
> ggplot(resume_minus_44years, aes(x=years_experience, fill=job_type)) +
geom_density(position="fill") + scale_x_continuous(breaks = c(0, 5, 10, 15, 20, 25))
```



We can see here that applicants are applying for all positions at relatively equal rates regardless of their years of experience. So, someone with 20 years of experience will apply for a manager position at around the same rate as someone with 10 years of experience.

years_experience vs. job_industry

```
> ggplot(resume_minus_44years, aes(x=years_experience, fill=job_industry)) +
geom_density(position="fill") + scale_x_continuous(breaks = c(0, 5, 10, 15, 20, 25))
```



The same is true here as it was for the above graph: regardless of years of experience, applicants seem to apply for jobs in all industries at the same rate. Whether someone has 5 years of experience or 20 years of experience - that does not influence their decision to apply for jobs in certain industries.

Best Subset Selection

In order to perform best subset selection, and thus find the best subset of variables to model the data, I first needed to remove all NA values. I also removed certain variables that I knew would have no impact on the model and would only create more noise and difficulty in calculating the best subset. The variables I chose to keep were those that provided information on the applicant, not the job itself.

```
> library(leaps)
> resume <- resume[-c(1:14)]
> resume$years_college <- na.omit(resume$years_college)
.
```

.

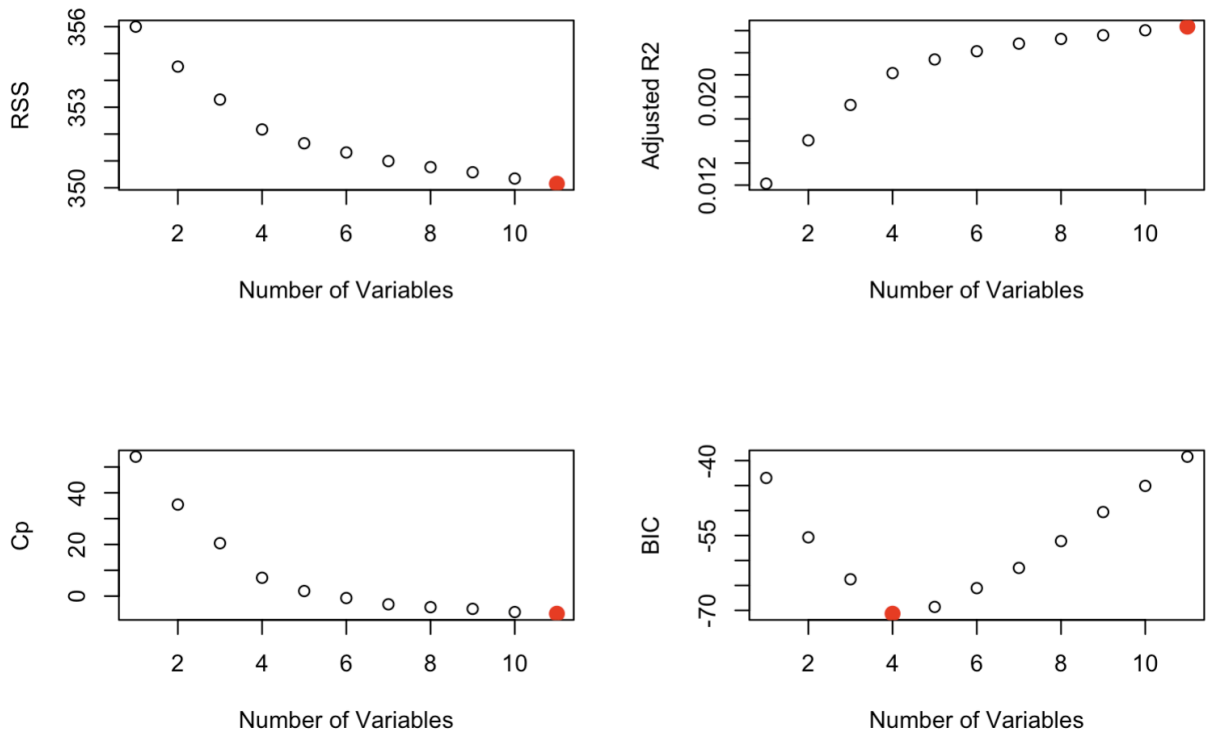
.

```
> resume$works_during_school <- na.omit(resume$works_during_school)
>regfit.full <- regsubsets(received_callback ~ ., resume, really.big = T,  nvmax=10)
>summary(regfit.full)
```

```
1 subsets of each size up to 11
Selection Algorithm: exhaustive
         firstnameAllison firstnameAnne firstnameBrad firstnameBrendan firstnameBrett firstnameCarrie firstnameDarnell firstnameEbony
1  ( 1 )  " "              " "           " "           " "              " "            " "             " "               " "
2  ( 1 )  " "              " "           " "           " "              " "            " "             " "               " "
3  ( 1 )  " "              " "           " "           " "              " "            " "             " "               " "
4  ( 1 )  " "              " "           " "           " "              " "            " "             " "               " "
5  ( 1 )  " "              " "           " "           " "              " "            " "             " "               " "
6  ( 1 )  " "              " "           "*"           " "              " "            " "             " "               " "
7  ( 1 )  " "              " "           "*"           " "              " "            " "             " "               " "
8  ( 1 )  " "              " "           "*"           " "              " "            "*"             " "               " "
9  ( 1 )  " "              " "           "*"           " "              " "            " "             " "               " "
10  ( 1 ) " "              " "           "*"           " "              " "            "*"             " "               " "
11  ( 1 ) " "              " "           "*"           " "              " "            "*"             " "               "*"
         firstnameEmily firstnameGeoffrey firstnameGreg firstnameHakim firstnameJamal firstnameJay firstnameJermaine firstnameJill
1  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
2  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
3  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
4  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
5  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
6  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
7  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
8  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
9  ( 1 )  " "            " "               " "           " "            " "            " "           " "               " "
10  ( 1 ) " "            " "               " "           " "            " "            " "           " "               " "
11  ( 1 ) " "            " "               " "           " "            " "            " "           " "               " "
         firstnameKareem firstnameKeisha firstnameKenya firstnameKristen firstnameLakisha firstnameLatonya firstnameLatoya firstnameLaurie
1  ( 1 )  " "             " "             " "           " "              " "              " "              " "             " "
2  ( 1 )  " "             " "             " "           " "              " "              " "              " "             " "
3  ( 1 )  " "             " "             " "           " "              " "              " "              " "             " "
4  ( 1 )  " "             " "             " "           " "              " "              " "              " "             " "
5  ( 1 )  " "             " "             " "           " "              " "              " "              " "             " "
6  ( 1 )  " "             " "             " "           " "              " "              " "              " "             " "
7  ( 1 )  " "             " "             " "           "*"              " "              " "              " "             " "
8  ( 1 )  " "             " "             " "           "*"              " "              " "              " "             " "
9  ( 1 )  " "             " "             " "           "*"              " "              " "              " "             " "
10  ( 1 ) " "             " "             " "           "*"              " "              " "              " "             " "
11  ( 1 ) " "             " "             " "           "*"              " "              " "              " "             " "
         firstnameLeroy firstnameMatthew firstnameMeredith firstnameNeil firstnameRasheed firstnameSarah firstnameTamika firstnameTanisha
1  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
2  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
3  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
4  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
5  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
6  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
7  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
8  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
9  ( 1 )  " "            " "              " "              " "           " "              " "             " "              " "
10  ( 1 ) " "            " "              " "              " "           " "              " "             " "              " "
11  ( 1 ) " "            " "              " "              " "           " "              " "             " "              " "
         firstnameTodd firstnameTremayne firstnameTyrone racewhite genderm years_college college_degree honors worked_during_school
1  ( 1 )  " "           " "               " "             " "       " "     " "           " "            " "    " "
2  ( 1 )  " "           " "               " "             " "       " "     " "           " "            " "    " "
3  ( 1 )  " "           " "               " "             "*"       " "     " "           " "            " "    " "
4  ( 1 )  " "           " "               " "             "*"       " "     " "           " "            "*"    " "
5  ( 1 )  " "           " "               " "             "*"       " "     " "           " "            "*"    " "
6  ( 1 )  " "           " "               " "             "*"       " "     " "           " "            "*"    " "
7  ( 1 )  " "           " "               " "             "*"       " "     " "           " "            "*"    " "
8  ( 1 )  " "           " "               " "             "*"       " "     " "           " "            "*"    " "
9  ( 1 )  " "           " "               " "             "*"       " "     " "           " "            "*"    "*"
10  ( 1 ) " "           " "               " "             "*"       " "     " "           " "            "*"    " "
11  ( 1 ) " "           " "               " "             "*"       " "     " "           " "            "*"    " "
```

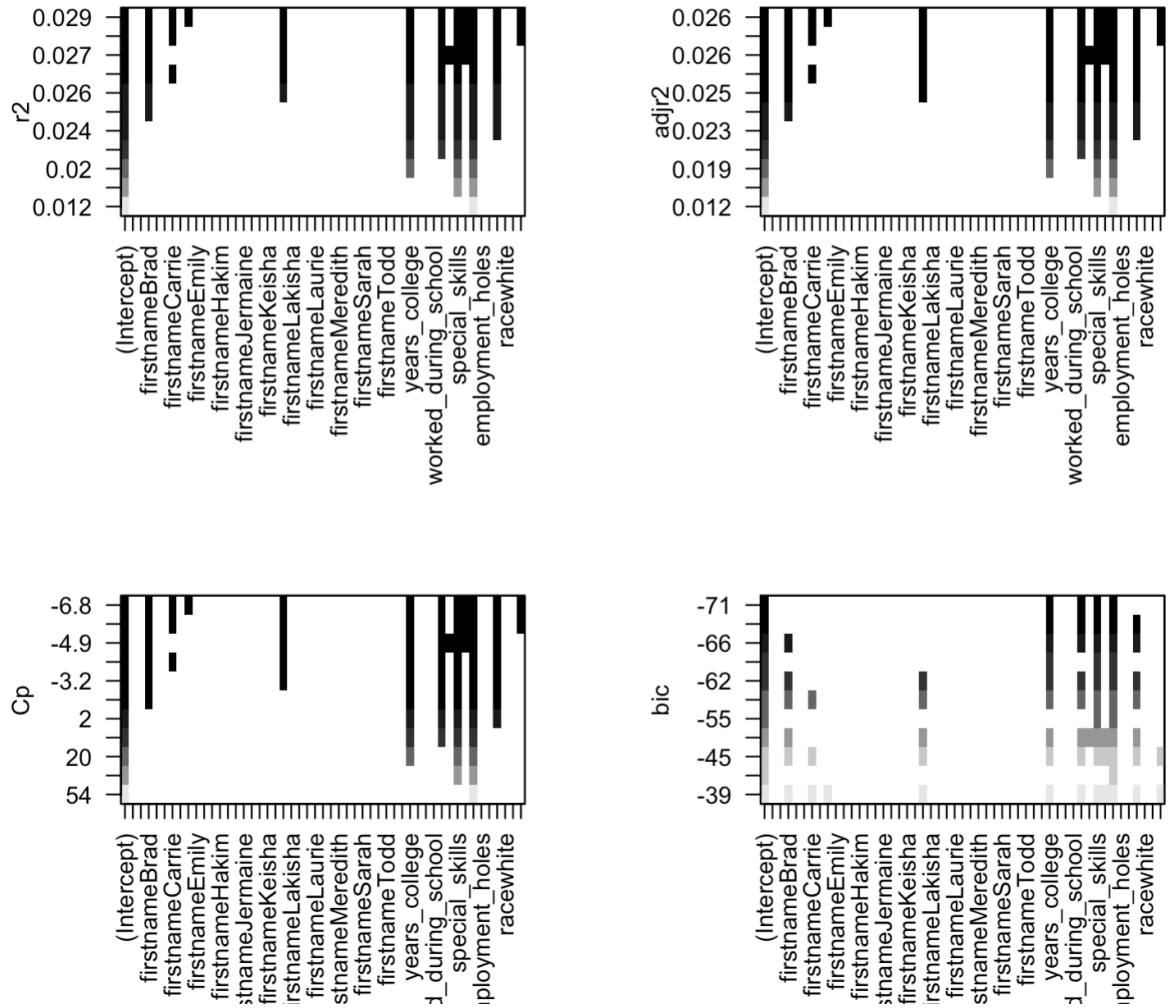
```
         years_experience computer_skills special_skills volunteer military employment_holes has_email_address resume_qualitylow
1  ( 1 )  " "              " "             "*"            " "       " "      " "              " "               " "
2  ( 1 )  "*"              " "             "*"            " "       " "      " "              " "               " "
3  ( 1 )  "*"              " "             "*"            " "       " "      " "              " "               " "
4  ( 1 )  "*"              " "             "*"            " "       " "      " "              " "               " "
5  ( 1 )  "*"              " "             "*"            " "       " "      "*"              " "               " "
6  ( 1 )  "*"              " "             "*"            " "       " "      "*"              " "               " "
7  ( 1 )  "*"              " "             "*"            " "       " "      "*"              " "               " "
8  ( 1 )  "*"              " "             "*"            " "       " "      "*"              " "               " "
9  ( 1 )  "*"              "*"             "*"            " "       " "      "*"              " "               " "
10  ( 1 ) "*"              "*"             "*"            " "       " "      "*"              " "               " "
11  ( 1 ) "*"              "*"             "*"            " "       " "      "*"              " "               " "
         works_during_school
1  ( 1 )  " "
2  ( 1 )  " "
3  ( 1 )  " "
4  ( 1 )  " "
5  ( 1 )  " "
6  ( 1 )  " "
7  ( 1 )  " "
8  ( 1 )  " "
9  ( 1 )  " "
10  ( 1 ) "*"
11  ( 1 ) "*"
```

```
> reg.summary <- summary(regfit.full)
> par(mfrow=c(2,2))
> plot(reg.summary$rss, xlab = "Number of Variables", ylab = "RSS")
> which.min(reg.summary$rss)
[1] 11
> points(11,reg.summary$rss[11],col="red", cex=2, pch=20)
> plot(reg.summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted R2")
> which.max(reg.summary$adjr2)
[1] 11
> points(11,reg.summary$adjr2[11],col="red", cex=2, pch=20)
> plot(reg.summary$cp, xlab="Number of Variables", ylab = "Cp")
> which.min(reg.summary$cp)
[1] 11
> points(11, reg.summary$cp[11], col="red", cex=2, pch=20)
> plot(reg.summary$bic, xlab="Number of Variables", ylab="BIC")
> which.min(reg.summary$bic)
[1] 4
> points(4, reg.summary$bic[4], col="red", cex=2, pch=20)
```



Looking at the above plots, we can see one commonality: at around 4 variables, the models do not get significantly "better". For example, with Cp, it is true that 11 variables result in the best fit. However, this fit begins to level off around 4 to the point where 11 variables are not significantly better than 4. The same is true for RSS and Adjusted R-Squared. For the BIC, the best option is clearly 4 variables. We can see below that these variables are years_college, years_experience, special_skills, and military.

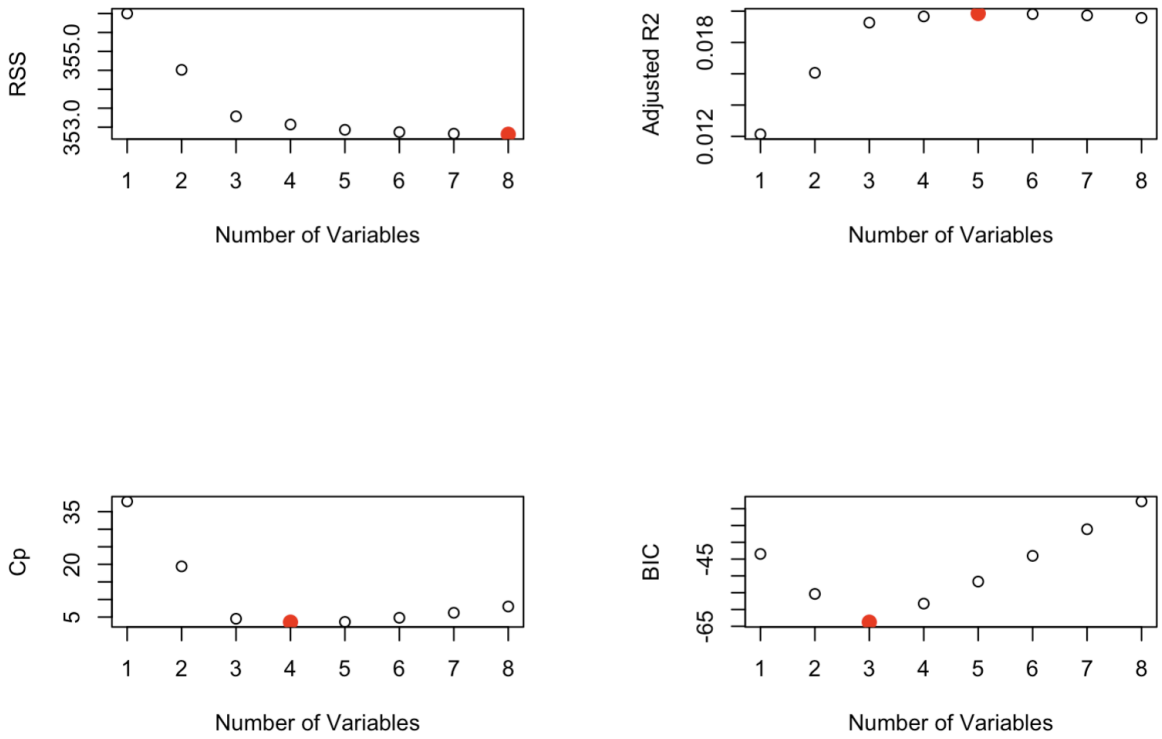
```
> plot(regfit.full, scale="r2")
> plot(regfit.full, scale="adjr2")
> plot(regfit.full, scale="Cp")
> plot(regfit.full, scale="bic")
```




```
> coef(regfit.full, 4)
    (Intercept)    years_college years_experience   special_skills
    0.013656734      0.004755745      0.003500676      0.066605050
       military
    0.002839199
```

It is at this point that I realized that I needed to make adjustments to the received_callback variable if I wanted to see the connection between race/gender and receiving a callback. The above results show that an applicant is more likely to receive a callback if their resume lists a higher number of years in college, a higher number of years of experience, special skills, and military experience. I want to know if they are *less* likely to receive a callback if they are black and/or female. To do this, I changed all 0s to 1s - and all 1s to 0s - in the Excel file as they relate to the received_callback variable. I then did the above steps again.

```
> resume_new <- read.csv("~/Desktop/resume_new.csv")
> resume_new <- resume_new[-c(1:15)]
> resume_new$received_callback <- na.omit(resume_new$received_callback)
.
.
.
> resume_new$works_during_school <- na.omit(resume_new$works_during_school)
> resume_new_new <- resume_new[-c(2, 7, 8, 13, 14, 15, 17)]
> regfit.full.new.new <- regsubsets(received_callback ~ ., resume_new_new, really.big = T)
> reg.summary.new <- summary(regfit.full.new.new)
> par(mfrow=c(2,2))
> plot(reg.summary.new$rss, xlab = "Number of Variables", ylab = "RSS")
> reg.summary.new.new <- summary(regfit.full.new.new)
> par(mfrow=c(2,2))
> plot(reg.summary.new.new$rss, xlab = "Number of Variables", ylab = "RSS")
> which.min(reg.summary.new.new$rss)
[1] 8
> points(8,reg.summary.new.new$rss[8],col="red", cex=2, pch=20)
> plot(reg.summary.new.new$adjr2, xlab = "Number of Variables", ylab = "Adjusted R2")
> which.max(reg.summary.new.new$adjr2)
[1] 5
> points(5,reg.summary.new.new$adjr2[5],col="red", cex=2, pch=20)
> plot(reg.summary.new.new$cp, xlab="Number of Variables", ylab = "Cp")
> which.min(reg.summary.new.new$cp)
[1] 4
> points(4, reg.summary.new.new$cp[4], col="red", cex=2, pch=20)
> plot(reg.summary.new.new$bic, xlab="Number of Variables", ylab="BIC")
> which.min(reg.summary.new.new$bic)
[1] 3
> points(3, reg.summary.new.new$bic[3], col="red", cex=2, pch=20)
```




```
> coef(regfit.full.new.new, 8)
     (Intercept)        racewhite          genderm
     0.957919206     -0.031324746      0.006879430
  college_degree  years_experience   computer_skills
    -0.008720180     -0.003203722      0.022322857
  special_skills       volunteer resume_qualitylow
    -0.065397757      0.005848495      0.015749880
```

The results were both expected and surprising. What they tell us is that *not* getting a callback is positively correlated with a low-quality resume. This makes sense; an applicant is less likely to receive a callback if they have a low-quality resume. *Not* getting a callback is also negatively correlated with being white. Unfortunately, this did not surprise me and also confirmed one part of my hypothesis - being black is positively correlated with *not* receiving a callback, meaning the white applicants had a better chance of getting a callback. What surprised me is that being male was positively correlated with *not* receiving a callback. That is, until I recalled that the number of female applicants was nearly triple that of male applicants. I decided to move forward in my analysis with the below predictors, as the plots above show that they best fit the model.

```
> coef(regfit.full.new.new, 3)
    (Intercept)       racewhite years_experience   special_skills
    0.983915080    -0.031753279    -0.003456968    -0.065151985
```

Hypothesis 1: All 3 predictors are statistically significant.

```
> hypothesis1 <- glm(received_callback ~ race+years_experience+special_skills, data =
resume_new_new, family = binomial)
> summary(hypothesis1)

Call:
glm(formula = received_callback ~ race + years_experience + special_skills,
   family = binomial, data = resume_new_new)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.353614   0.128953  26.006  < 2e-16 ***
racewhite       -0.444416   0.108236  -4.106 4.03e-05 ***
years_experience -0.041143   0.009211  -4.467 7.95e-06 ***
special_skills   -0.821789   0.106592  -7.710 1.26e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 2726.9  on 4869  degrees of freedom
Residual deviance: 2634.8  on 4866  degrees of freedom
AIC: 2642.8

Number of Fisher Scoring iterations: 5
```

Looking at the p-values, we can see that all are below the accepted 0.05 value. Thus, we can now reject the null hypothesis and say that all three variables are statistically significant when relating to the received_callback variable.

The Validation Set Approach

To evaluate the above model, I will start by using the validation set approach. I separated the training and test data into 70% and 30%, respectively.

```
> set.seed(101)
> n_train <- round(nrow(resume_new_new)* 0.7)
> train <- rep(FALSE, nrow(resume_new_new))
> train[sample(nrow(resume_new_new),n_train)] <- TRUE
> test <- !train
```

```
> training_data <- resume_new_new[train, ]
> test_data <- resume_new_new[test, ]
> glm_model <- glm(received_callback ~ race+years_experience+special_skills, data =
training_data, family = binomial)
> glm_pred_prob <- predict(glm_model, newdata = test_data, type = "response")
> glm_pred <- ifelse(glm_pred_prob >= 0.5, 1, 0)
> table(glm_pred, test_data$received_callback)

glm_pred    0    1
       1  124 1337
```

To visualize this output better, let's rewrite it as follows:

glm_pred	0	1
0	-	124
1	-	1337

What this tells me is that the above model correctly predicted that the applicant would receive a callback based on the predictors provided for 1337 instances. However, in the cases where the applicant did not receive a call back, it incorrectly predicted that the applicant would receive a callback in 124 instances. This means that this model correctly predicted the outcome 92% (resulting in a test error rate of 8%) of the time, as shown by the calculation below.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} = \frac{1337}{1461} = 0.92 * 100 = 92\%$$

I believe this error rate is acceptable for a few reasons. First, this is a large dataset. If the dataset were smaller, I believe an error rate of 8% might be unacceptable; however, for a dataset this large, I believe it is actually pretty good. Second, error rate acceptability depends on context. In order to achieve this calculation, I had to do a number of modifications to the data: there were a huge amount of NA values that I had to remove and I had to remove quite a few variables altogether from the dataset. With all of this being said, I believe an 8% error rate is perfectly acceptable. Thus, this model, in my opinion, is a good fit.

LOOCV

While I believe 8% is an acceptable test error rate in this context, I wanted to try out LOOCV since it does usually produce an improved test error rate.

```
> library(caret)
> train.control <- trainControl(method = "LOOCV")
> model <- train(received_callback ~ race+years_experience+special_skills,
data=resume_new_new, method="glm", trControl=train.control)
Warning message:
In train.default(x, y, weights = w, ...) :
  You are trying to do regression and your outcome only has two possible values Are you trying
to do classification? If so, use a 2 level factor as your outcome column.
```

```
> print(model)
Generalized Linear Model

4870 samples
   3 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 4869, 4869, 4869, 4869, 4869, 4869, ...
Resampling results:

  RMSE       Rsquared    MAE
  0.2695859  0.01810767  0.1452213
```

This was not the output I was expecting, as I did not obtain an "accuracy" or "Kappa" value. As suspected, this had to do with the warning error I received above. To fix this, I converted received_callback to a factor with two levels.

```
> resume_new_new$received_callback <- factor(resume_new_new$received_callback, levels =
c(0, 1))
> model <- train(received_callback ~ race + years_experience + special_skills,data =
resume_new_new,method = "glm")
> loocv_results <- trainControl(method = "LOOCV")
> cv_results <- train(received_callback ~ ., data = resume_new_new,method = "glm",trControl =
loocv_results)
> cv_results$results
  parameter  Accuracy Kappa
1      none 0.9195072     0
```

Finally I was able to obtain the accuracy and Kappa values. As we see above, I get a similar accuracy of around 92% as I did when performing the validation set approach. What was interesting was the Kappa value of 0 (accuracy is no better than chance). I do believe that this is due to an imbalance in the dataset, specifically in receieved_callback (Covidence). There are significantly more "no callback" observations than "yes callback"observations.

K-fold Cross-Validation

Because of the imbalance in the dataset, I have decided to use a resampling technique. I decided upon K-fold Cross Validation since it is best used for larger datasets.

```
> model.k <- train(received_callback~race+years_experience+special_skills,
data=resume_new_new, trControl = train_control, method="glm")
> print(model.k)
Generalized Linear Model

4870 samples
```

3 predictor
2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4383, 4382, 4384, 4383, 4383, 4383, ...
Resampling results:

Accuracy Kappa
0.9195079 0

As we can clearly see, both LOOCV and K-fold Cross Validation both result in a 92% accurate and a Kappa of 0. This suggests that the low Kappa is not due to the imbalance in the receieved_callback predictor. Rather, it might be due to some other issue. I tried a few different approaches, such as using the original dataset (resume), using the second-iteration dataset (resume_new), and even applying cross-validation across all predictors, not just the three selected by best subset. The results are shown below.

Best Subset Predictors, resume_new_new	**Best Subset Predictors, resume_new**	**Best Subset Predictors, resume**
> model.k1 <- train(received_callback~race +years_experience+special_skills, data=resume_new_new, trControl = train_control, method="glm") > print(model.k1) Generalized Linear Model 4870 samples 3 predictor 2 classes: '0', '1' No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 4382, 4383, 4383, 4383, 4384, 4383, ... Resampling results: Accuracy Kappa 0.9195075 0	> model.k2 <- train(received_callback~race +years_experience+special_skills, data=resume_new, trControl = train_control, method="glm") > print(model.k2) Generalized Linear Model 4870 samples 3 predictor 2 classes: '0', '1' No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 4382, 4383, 4383, 4382, 4384, 4384, ... Resampling results: Accuracy Kappa 0.9195079 0	> model.k3 <- train(received_callback~race +years_experience+special_skills, data=resume, trControl = train_control, method="glm") > print(model.k3) Generalized Linear Model 3102 samples 3 predictor 2 classes: '0', '1' No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 2791, 2793, 2793, 2792, 2791, 2792, ... Resampling results: Accuracy Kappa 0.9235999 0

9 predictors, resume_new_new	16 predictors, resume_new	30 predictors, resume
> model.k4 <- train(received_callback~., data=resume_new_new, trControl = train_control, method="glm") > print(model.k4) Generalized Linear Model 4870 samples 9 predictor 2 classes: '0', '1' No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 4383, 4383, 4383, 4383, 4383, 4383, ... Resampling results: Accuracy Kappa 0.9195075 0	> model.k5 <- train(received_callback~., data=resume_new, trControl = train_control, method="glm") > print(model.k5) Generalized Linear Model 4870 samples 16 predictor 2 classes: '0', '1' No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 4384, 4383, 4384, 4383, 4383, 4383, ... Resampling results: Accuracy Kappa 0.9195079 0	> model.k6 <- train(received_callback~., data=resume, trControl = train_control, method="glm") > print(model.k6) Generalized Linear Model 3102 samples 30 predictor 2 classes: '0', '1' No pre-processing Resampling: Cross-Validated (10 fold) Summary of sample sizes: 2791, 2791, 2792, 2793, 2792, 2792, ... Resampling results: Accuracy Kappa 0.9226313 -0.001868565

Looking at the table, we see all the results are the same except for the last, which actually had a *worse* Kappa value (likely due to including all the predictors). What this tells me is that the low Kappa value is not due to poor calculations in the best subset selection. I don't believe this is caused by a poor model. I think that things might look better if there were no NA values for any of the observations, though I can't say for certain.

Summary

Through my analysis, I struggled to find as clear of an answer to my hypothesis as I thought I would. When I began this project, I thought all methods of analysis would point toward one clear answer: being black and/or female lowers your chances of receiving a callback despite experience, education, or skill set. I immediately ran into a hurdle with the dataset. There were three times as many female applicants as male applicants. This threw a wrench in analyzing the gender portion of my hypothesis. While performing best subset selection, I found that being male was correlated with *not* receiving a callback - the total opposite of what I had hypothesized. I can say with near certainty that this is due to the uneven split in data based on gender. I was unable to make any valid conclusions when it came to gender since the data did not adequately represent the population and skewed the results based on this predictor.

I moved forward in my analysis without this predictor. I included race, years of experience, and special skills to attempt to predict the likelihood of applicants receiving callbacks. The model I used based on these predictors actually did a very good job as it relates to accuracy. It was accurate around 92% of the time. I suspected it would be since it intuitively made sense that a model using data on race, years of experience, and special skills would have enough information to accurately predict whether the applicant received a callback. Despite the accuracy of the model, all variations produced a Kappa of 0 (and a negative value, in one instance). So, despite the models' accuracies, these accuracies were due no more to chance than any other model. Intuitively and logically, this does not add up. Throughout this project, I have had to make significant adjustments to the data in order to perform my calculations in R. I believe this has had a huge effect on my results. The data contained a large amount of NA values across many predictors. Some variables had to be removed altogether. Simply put, I think a more complete and evenly split dataset would have produced different results.

Works Cited

Arel-Bundock, V. *Which resume attributes drive job callbacks?* Github. https://vincentarelbundock.github.io/Rdatasets/doc/openintro/resume.html

Bertrand, M. & Mullainathan, S. (2002, July). *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination.* (Working Paper No. 9873). https://www.nber.org/system/files/working_papers/w9873/w9873.pdf

Covidence. (2023, April 5). *Why is my Cohen's Kappa value low and what can I do to improve it?* Covidence. https://support.covidence.org/help/why-is-my-cohen-s-kappa-value-low-and-what-can-i-do-to-improve-it

R-Project. *Tibbles.* The R-Project for Statistical Computing. https://cran.r-project.org/web/packages/tibble/vignettes/tibble.html#:~:text=Unlike%20data%20frames%2C%20tibbles%20don,coerced%20to%20a%20data%20frame.