Final Project Step 3

Logan Quandt

3/2/2022

Introduction and Problem Statement

> Many people still extol the American dream that used to be a single income supporting a suburban home with a picket fence, two cars in the driveway and a family. However, for many there appears to be inequality in the US in who shares in this dream. There is income and wealth inequality by gender, race and even age which may seem to be caused by inequity. Personally, I am very interested in financial ideas due to my work experience and my experience in trying to be as financially successful as my older family members. In this project I hope to explore some of this income inequality to see if I can identify relationships between variables that may also help to explain some of the differences.

How I will Approach This

> My hypothesis is that there are multiple variables that go into the differences in pay between genders, races, and age groups. In my approach I will be looking at the correlation and strength of correlation between variables to explore the relationships between them along with performing numerous other statistical tests. I will also explore the data visually using plots and graphs.

Libraries

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(purrr)
```

Loading Datasets

```
setwd("/Users/logan/Documents/GitHub/DSC520LQ/Final Project")

#Dataset one
```

```r
census_df <- read.csv("data/census.csv")
head(census_df)
```

```
##   age         workclass  education       maritalstatus         occupation
## 1  39         State-gov  Bachelors       Never-married       Adm-clerical
## 2  50  Self-emp-not-inc  Bachelors  Married-civ-spouse    Exec-managerial
## 3  38           Private    HS-grad            Divorced  Handlers-cleaners
## 4  53           Private       11th  Married-civ-spouse  Handlers-cleaners
## 5  28           Private  Bachelors  Married-civ-spouse     Prof-specialty
## 6  37           Private    Masters  Married-civ-spouse    Exec-managerial
##     relationship   race     sex capitalgain capitalloss hoursperweek
## 1  Not-in-family  White    Male        2174           0           40
## 2        Husband  White    Male           0           0           13
## 3  Not-in-family  White    Male           0           0           40
## 4        Husband  Black    Male           0           0           40
## 5           Wife  Black  Female           0           0           40
## 6           Wife  White  Female           0           0           40
##    nativecountry over50k
## 1  United-States   <=50K
## 2  United-States   <=50K
## 3  United-States   <=50K
## 4  United-States   <=50K
## 5           Cuba   <=50K
## 6  United-States   <=50K
```

```r
census_df_final <- select(census_df, age, sex, race, education, occupation, over50k)

#Dataset two
pinc_df_whites <- read.csv('data/pinc_white_test_csv.csv')

pinc_df_afr_amer <- read.csv('data/pinc_african_amer_test_csv.csv')

#Dataset three
glassdoor_pay_df <- read.csv('data/Glassdoor Gender Pay Gap.csv')
head(glassdoor_pay_df)
```

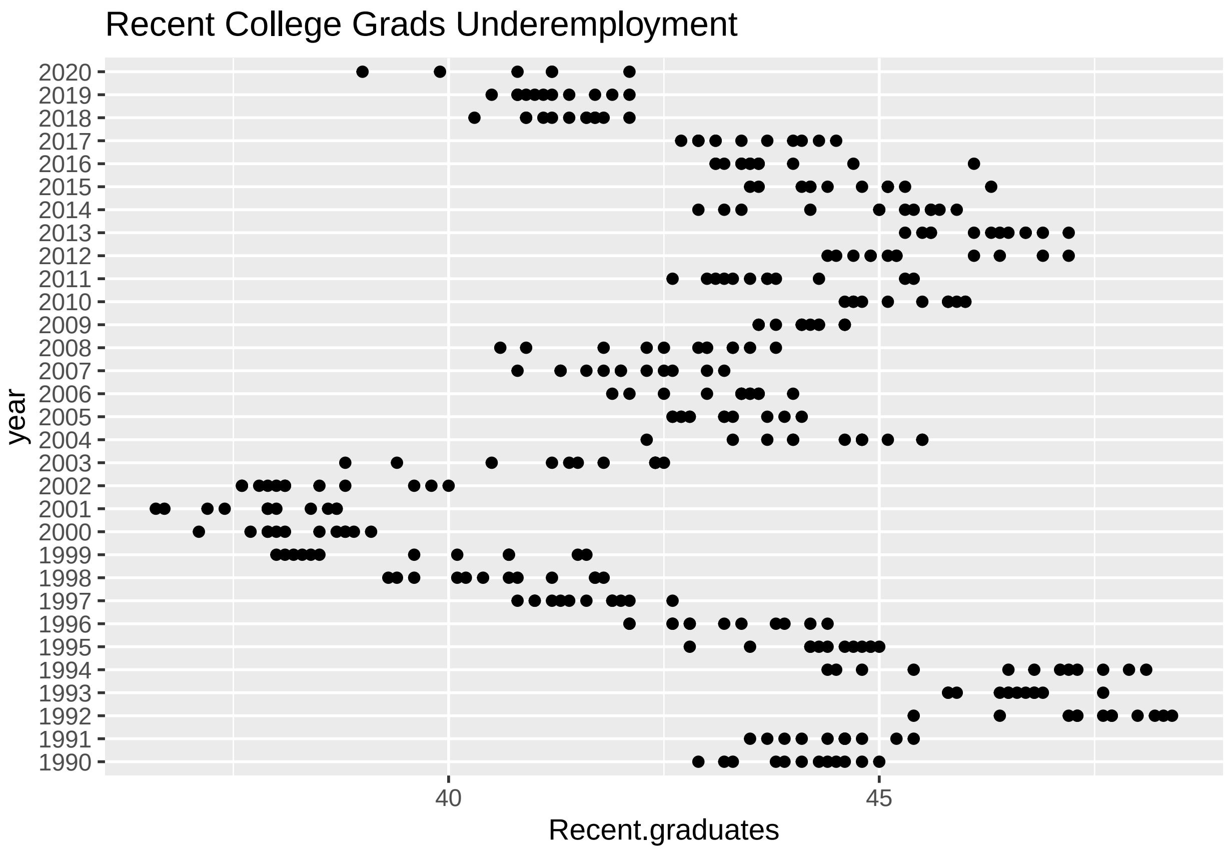
```
##              JobTitle Gender Age PerfEval Education          Dept Seniority
## 1    Graphic Designer Female  18        5   College    Operations         2
## 2   Software Engineer   Male  21        5   College    Management         5
## 3 Warehouse Associate Female  19        4       PhD Administration         5
## 4   Software Engineer   Male  20        5   Masters         Sales         4
## 5    Graphic Designer   Male  26        5   Masters   Engineering         5
## 6                  IT Female  20        5       PhD    Operations         4
##   BasePay Bonus
## 1   42363  9938
## 2  108476 11128
## 3   90208  9268
## 4  108080 10154
## 5   99464  9319
## 6   70890 10126
```

```r
glassdoor_pay_df_final <- select(glassdoor_pay_df, JobTitle, Gender, Age, Education,Seniority, BasePay,

#Dataset four
college_df_wages <- read.csv('data/wages.csv')
```

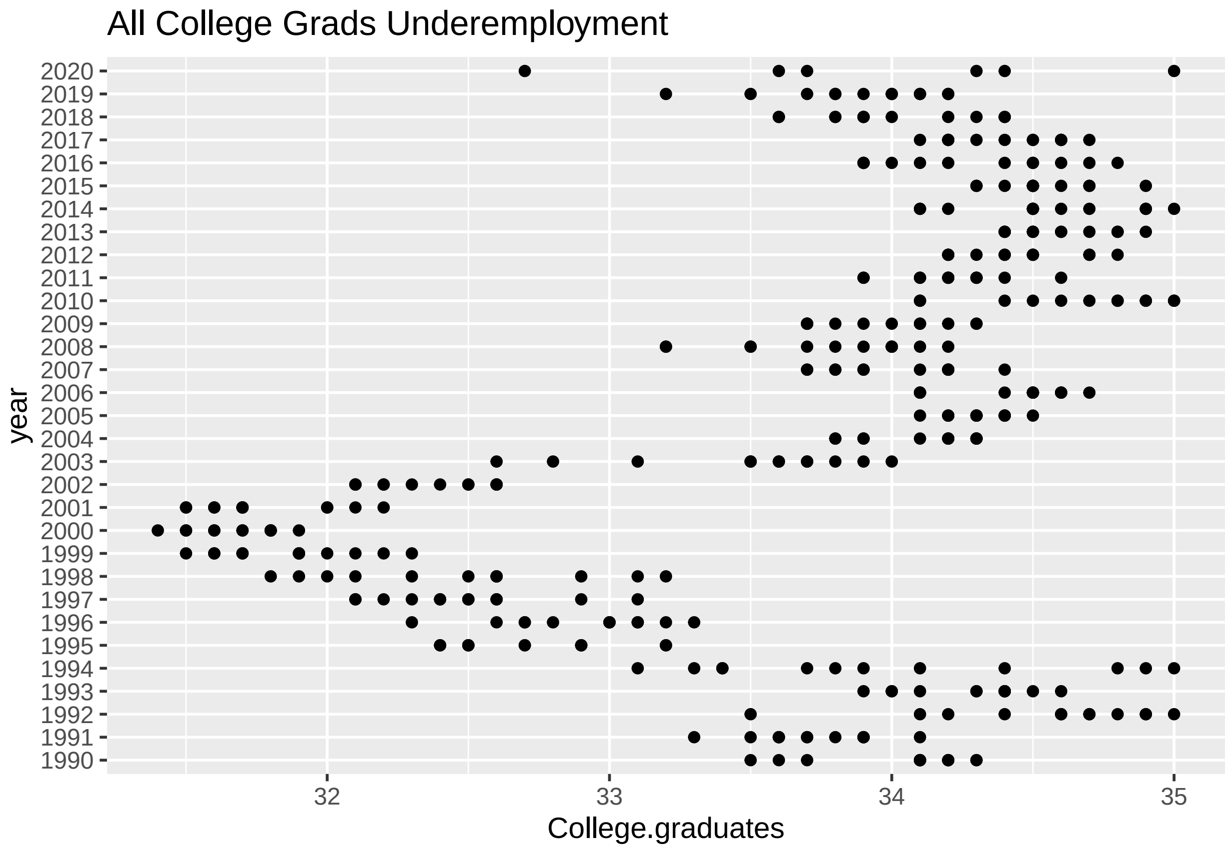
```
college_df_unemployment <- read.csv('data/Unemployment_rate.csv')
college_df_underemployment <- read.csv('data/under_employment_college_grads.csv')
college_df_labor_market <- read.csv('data/labor_market_college_grads.csv')
```

College Datasets Analysis

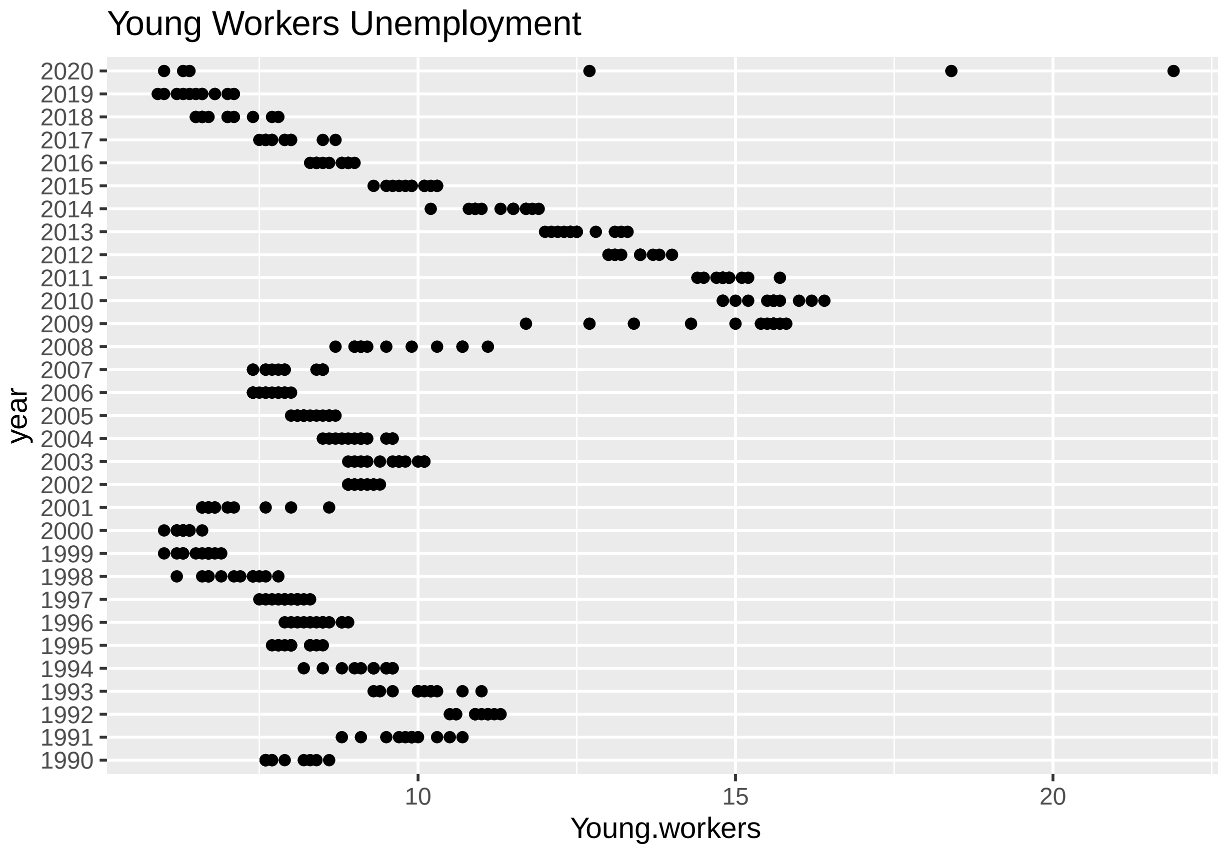
```
#format years
college_df_underemployment$year <-format(as.Date(college_df_underemployment$Date, format="%d/%m/%Y"),"%

#graphs for undermployment
underplot1 <- ggplot(college_df_underemployment, aes(x=Recent.graduates, y=year)) + geom_point() + ggti
underplot2 <- ggplot(college_df_underemployment, aes(x=College.graduates, y=year)) + geom_point() + ggt

underplot1
```



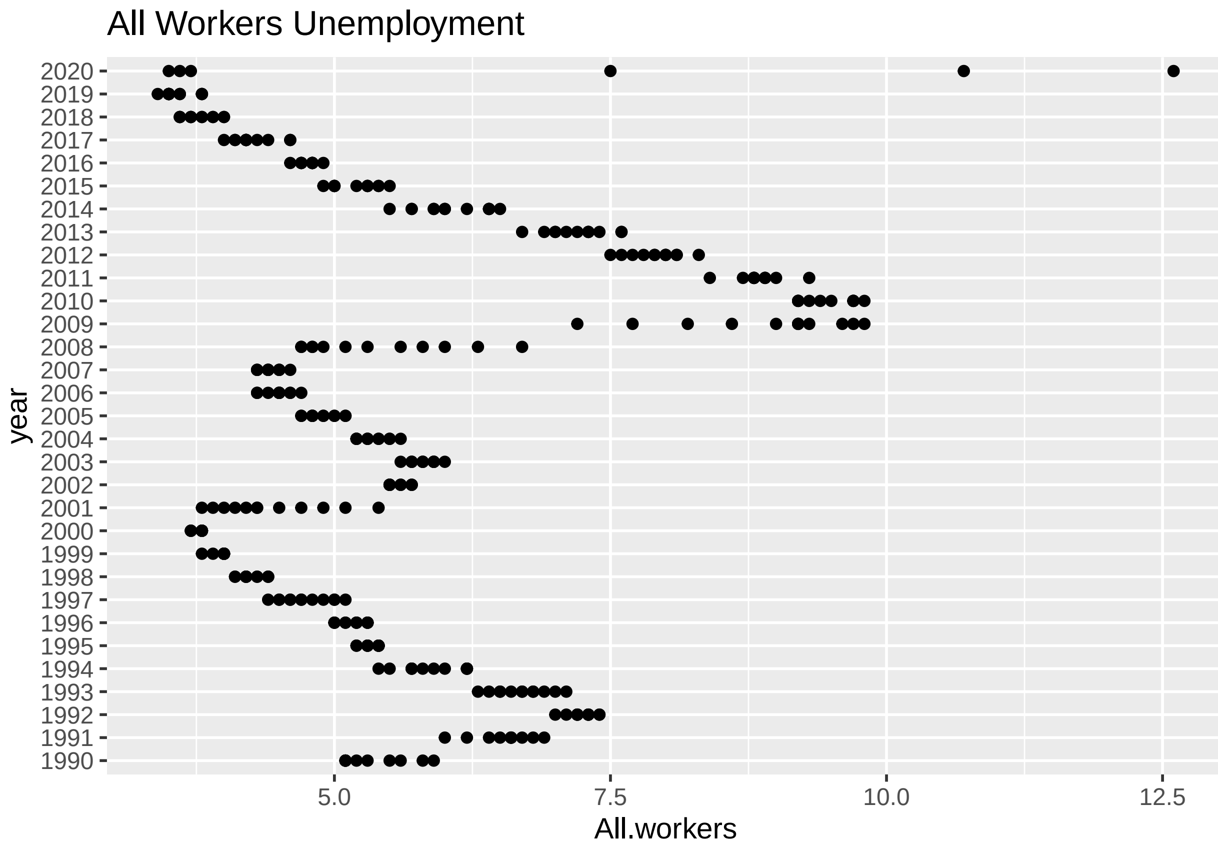

```
underplot2
```



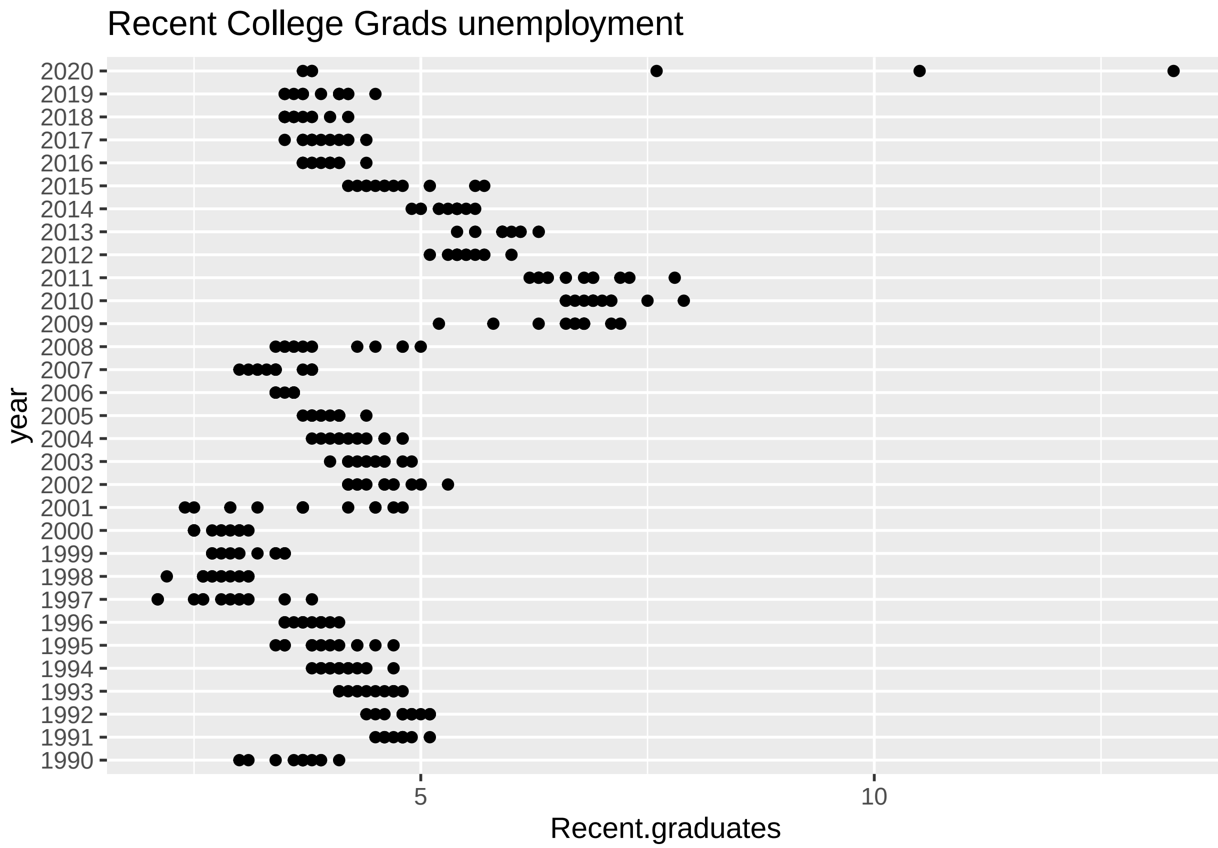

```
#graphs for unemployment
college_df_unemployment$year <- format(as.Date(college_df_unemployment$Date, format="%d/%m/%Y"),"%Y")

unplot1 <- ggplot(college_df_unemployment, aes(x=Young.workers, y=year)) + geom_point() + ggtitle("Young
unplot2 <- ggplot(college_df_unemployment, aes(x=All.workers, y=year)) + geom_point() + ggtitle('All Wo
unplot3 <- ggplot(college_df_unemployment, aes(x=Recent.graduates, y=year)) + geom_point() + ggtitle("R
unplot4 <- ggplot(college_df_unemployment, aes(x=College.graduates, y=year)) + geom_point() + ggtitle("

unplot1
```


Young Workers Unemployment
year
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
1996
1995
1994
1993
1992
1991
1990
10
15
20
Young.workers

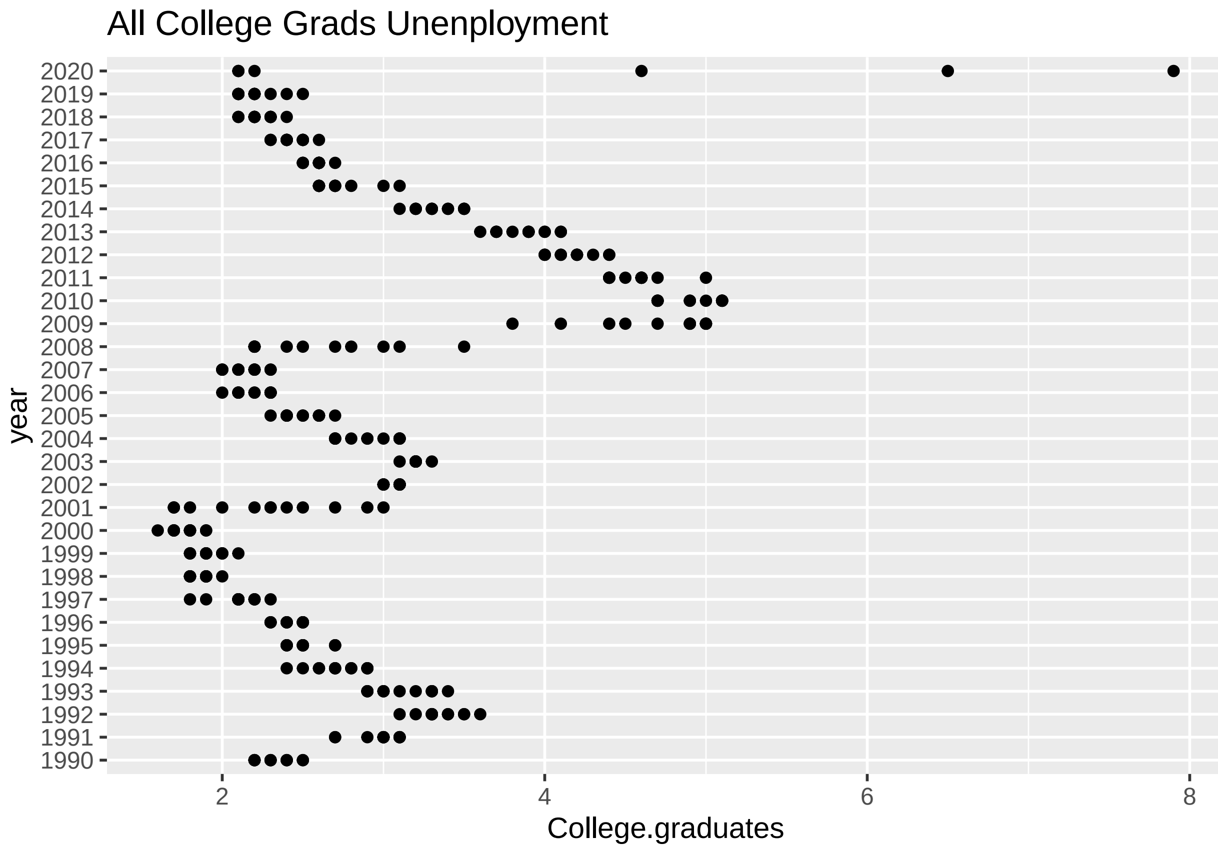
```
unplot2
```


All Workers Unemployment
year
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
1996
1995
1994
1993
1992
1991
1990
5.0
7.5
10.0
12.5
All.workers

```
unplot3
```


Recent College Grads unemployment
year
2020
2019
2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
2006
2005
2004
2003
2002
2001
2000
1999
1998
1997
1996
1995
1994
1993
1992
1991
1990
5
10
Recent.graduates

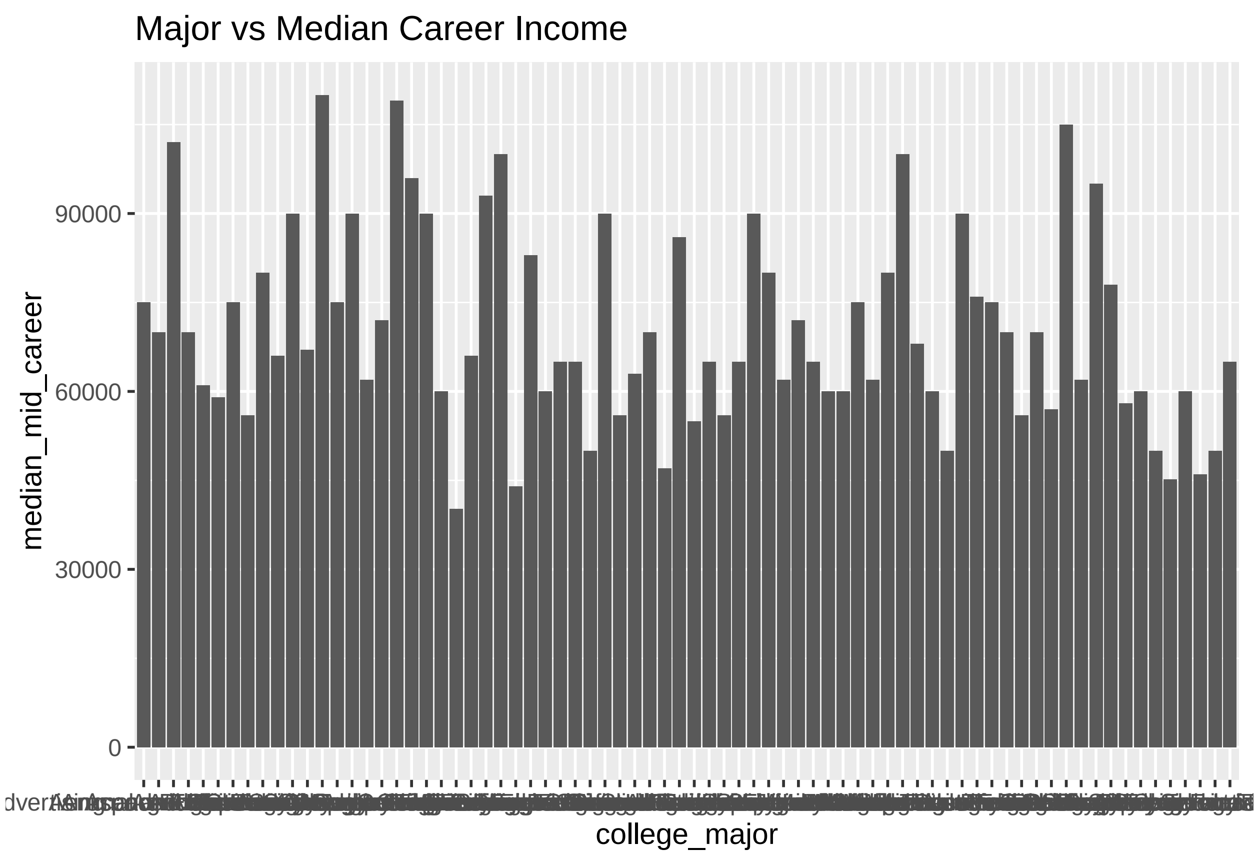
```
unplot4
```




```
#change df names and select to show differences by major
names(college_df_labor_market)[names(college_df_labor_market) == 'Median.Wage.Mid.Career'] <- 'median_mi
names(college_df_labor_market)[names(college_df_labor_market) == 'Major'] <- 'college_major'

college_df_labor_market_final <- select(college_df_labor_market, median_mid_career, college_major)

head(college_df_labor_market_final)
```

```
##   median_mid_career              college_major
## 1             70000                Agriculture
## 2             61000 Animal and Plant Sciences
## 3             65000     Environmental Studies
## 4             75000               Architecture
## 5             65000            Ethnic Studies
## 6             72000            Communications
```

```
#labor market graphs
ggplot(college_df_labor_market_final, aes(x=college_major, y=median_mid_career)) + geom_col() + ggtitle
```




```
#correlation between share with Grad degree and median career income
cor(college_df_labor_market_final$median_mid_career, college_df_labor_market$Share.with.Graduate.Degree)
```

```
## [1] -0.001777715
```

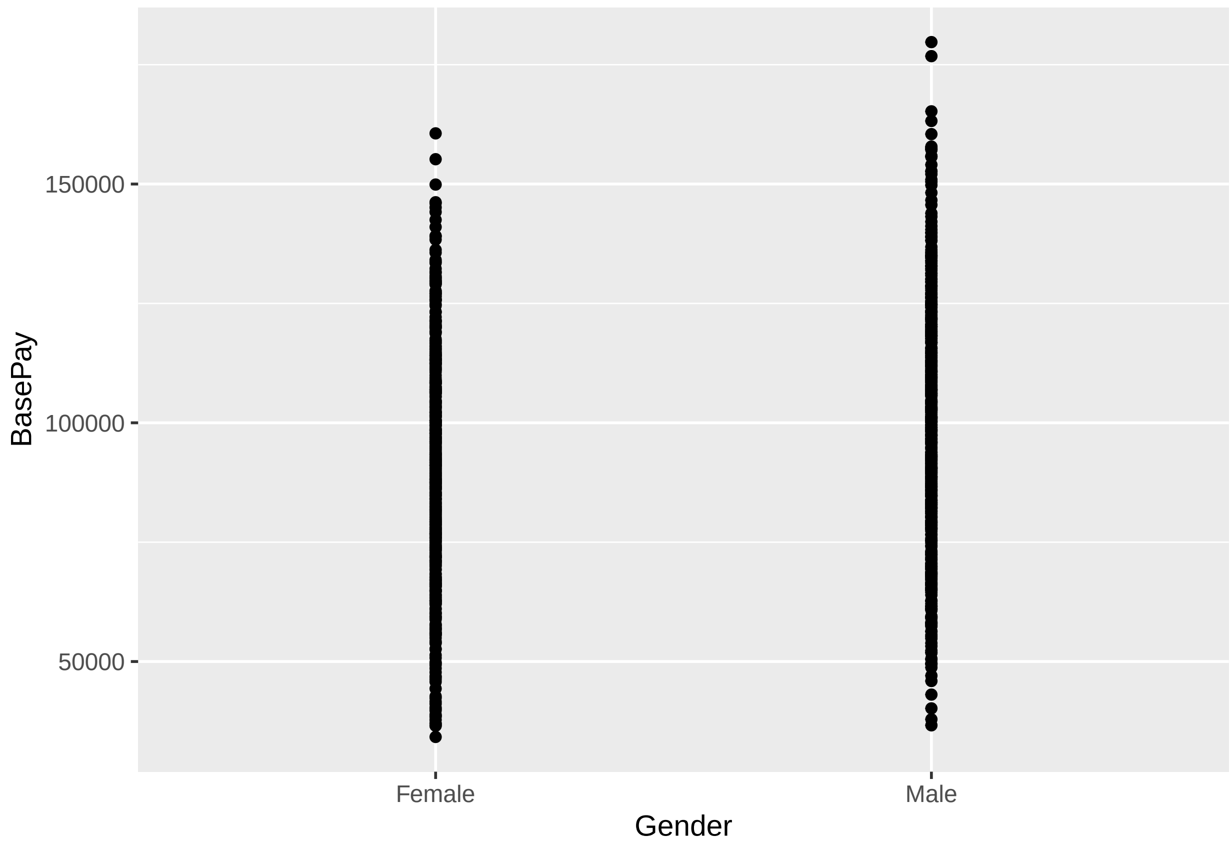
```
#Regression to check significance of correlation
graduate_lm <- lm(median_mid_career ~ Share.with.Graduate.Degree, college_df_labor_market)
summary(graduate_lm)
```

```
##
## Call:
## lm(formula = median_mid_career ~ Share.with.Graduate.Degree,
##     data = college_df_labor_market)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30159 -10379  -4324   9653  39661
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               70440.320   5868.021  12.004   <2e-16 ***
## Share.with.Graduate.Degree    -2.101    139.304  -0.015    0.988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16650 on 72 degrees of freedom
## Multiple R-squared:  3.16e-06,   Adjusted R-squared:  -0.01389
```

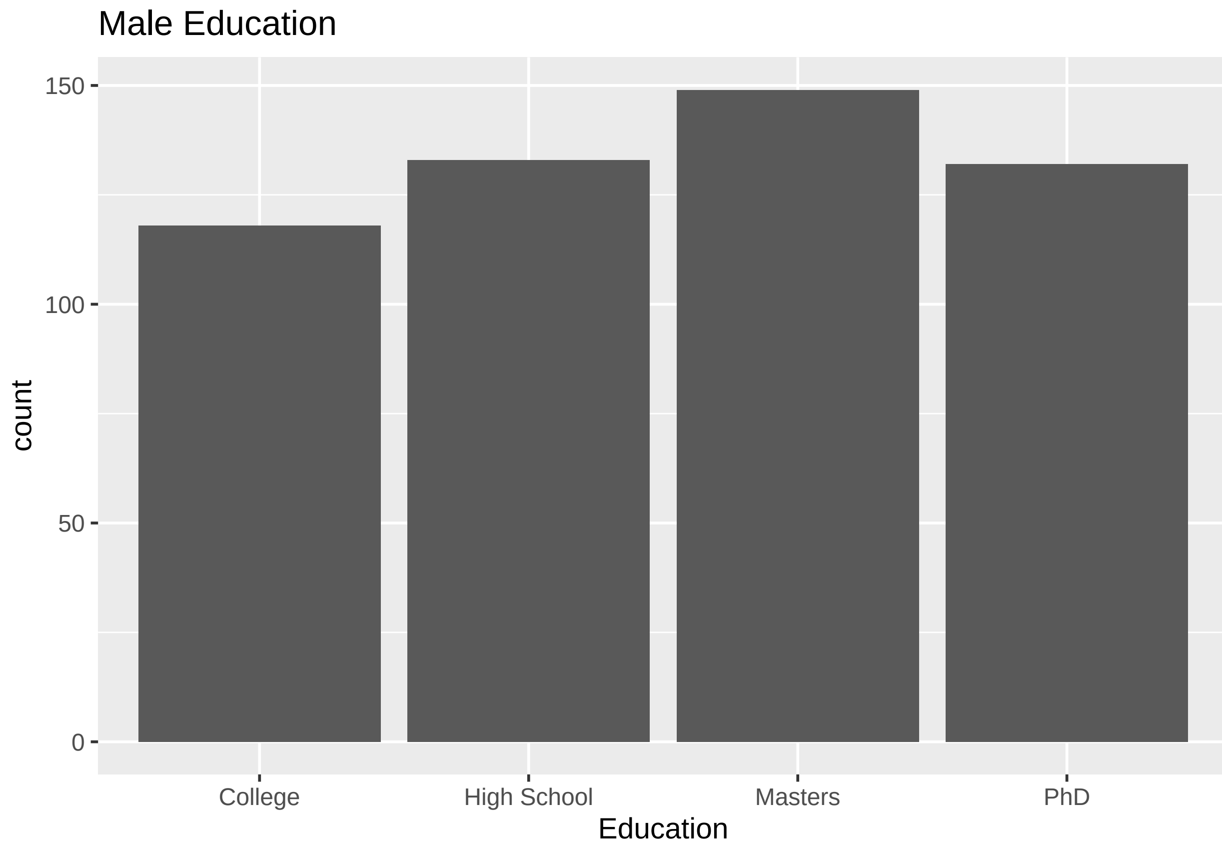
```
## F-statistic: 0.0002275 on 1 and 72 DF,  p-value: 0.988
```

Gender Analysis

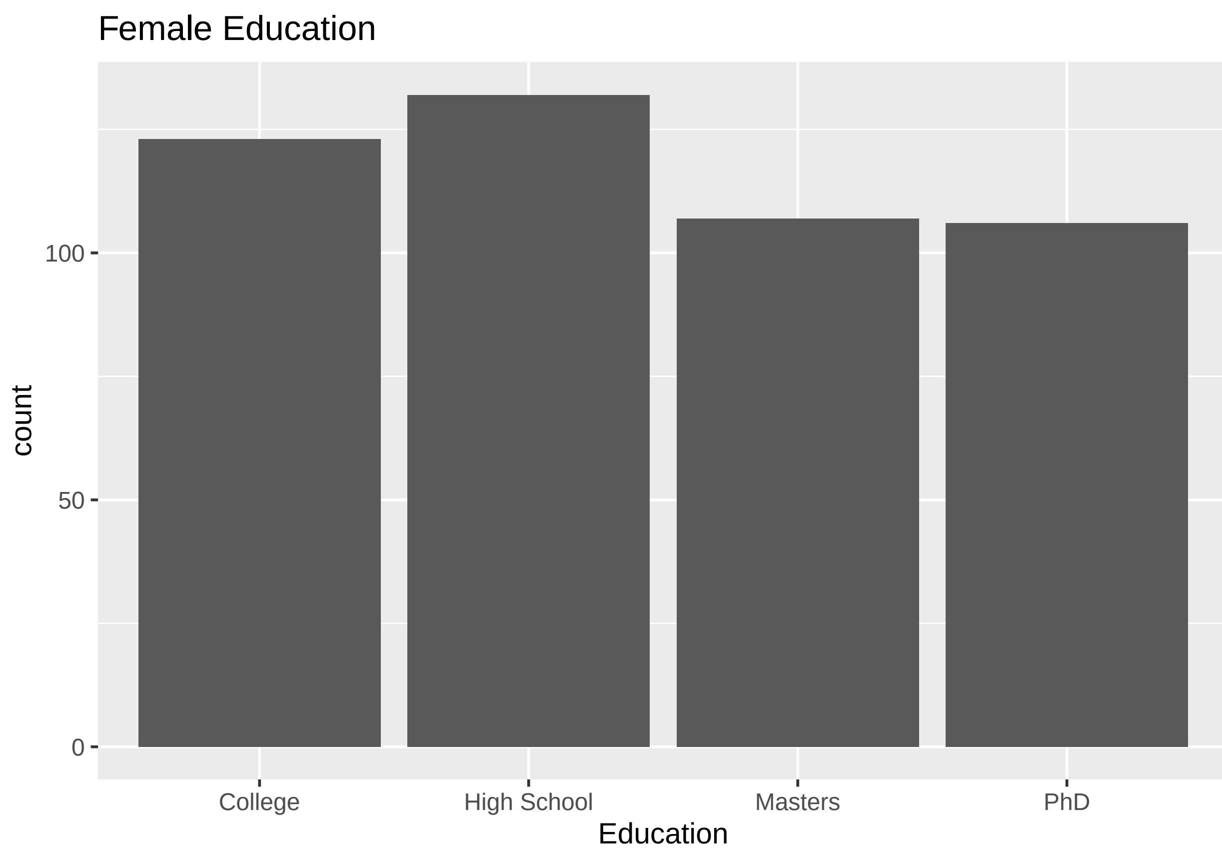
```r
#graph showing Gender vs Basepay
ggplot(glassdoor_pay_df_final, aes(x=Gender, y= BasePay)) + geom_point()
```



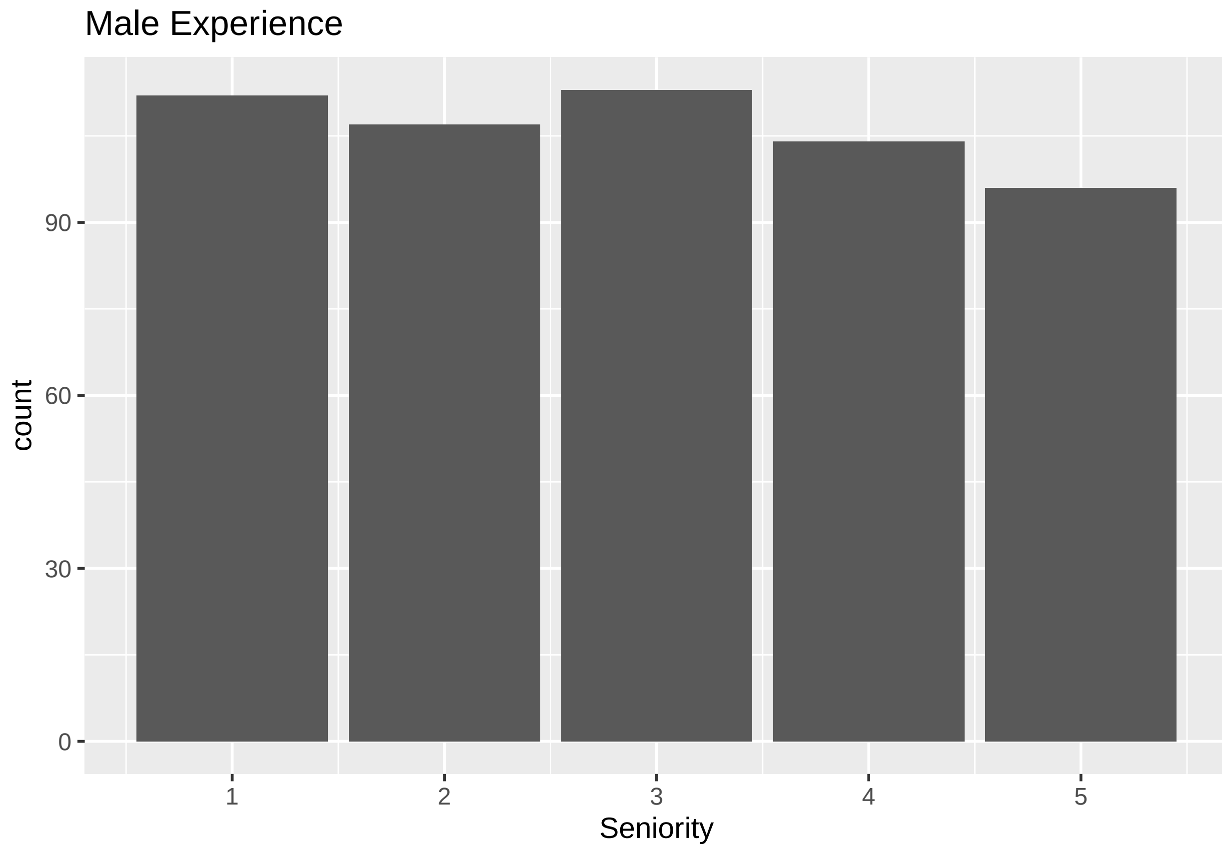

```r
#filter dataset by Gender
glassdoor_pay_df_final_m <- glassdoor_pay_df_final %>% filter(Gender == 'Male')
glassdoor_pay_df_final_f <- glassdoor_pay_df_final %>% filter(Gender == 'Female')

#graphs
glass_m_edplot <- ggplot(glassdoor_pay_df_final_m, aes(x=Education)) + geom_bar() + ggtitle('Male Educa
glass_f_edplot <- ggplot(glassdoor_pay_df_final_f, aes(x=Education)) + geom_bar() + ggtitle('Female Edu
glass_m_Senplot <- ggplot(glassdoor_pay_df_final_m, aes(x=Seniority)) + geom_bar() + ggtitle('Male Expe
glass_f_Senplot <- ggplot(glassdoor_pay_df_final_f, aes(x=Seniority)) + geom_bar() + ggtitle("Female Ex

glass_m_edplot
```



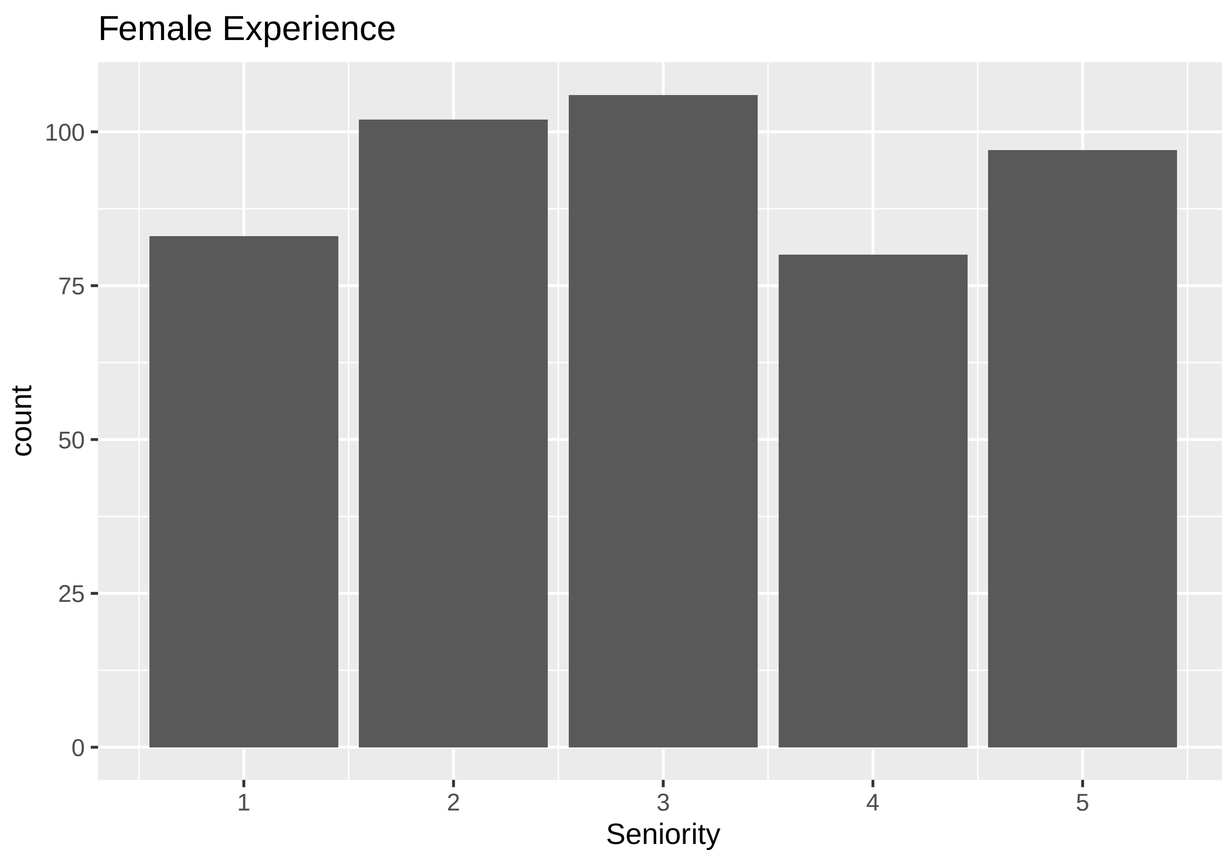

```
glass_f_edplot
```




```
glass_m_Senplot
```




```
glass_f_Senplot
```




```r
#correlation between Age, Seniority and basepay
cor(glassdoor_pay_df$BasePay, glassdoor_pay_df$Age)
```

```
## [1] 0.5626813
```

```r
cor(glassdoor_pay_df$BasePay, glassdoor_pay_df$Seniority)
```

```
## [1] 0.5110963
```

```r
#calculate percentage of managerial jobs
table(glassdoor_pay_df_final_m$JobTitle)
```

```
##
##      Data Scientist              Driver   Financial Analyst    Graphic Designer
##                  54                  45                  58                  50
##                  IT             Manager Marketing Associate     Sales Associate
##                  46                  72                  11                  51
##   Software Engineer Warehouse Associate
##                 101                  44
```

```r
table(glassdoor_pay_df_final_f$JobTitle)
```

```
##
##      Data Scientist              Driver   Financial Analyst    Graphic Designer
##                  53                  46                  49                  48
##                  IT             Manager Marketing Associate     Sales Associate
##                  50                  18                 107                  43
##   Software Engineer Warehouse Associate
##                   8                  46
```

```r
72/nrow(glassdoor_pay_df_final_m)
```

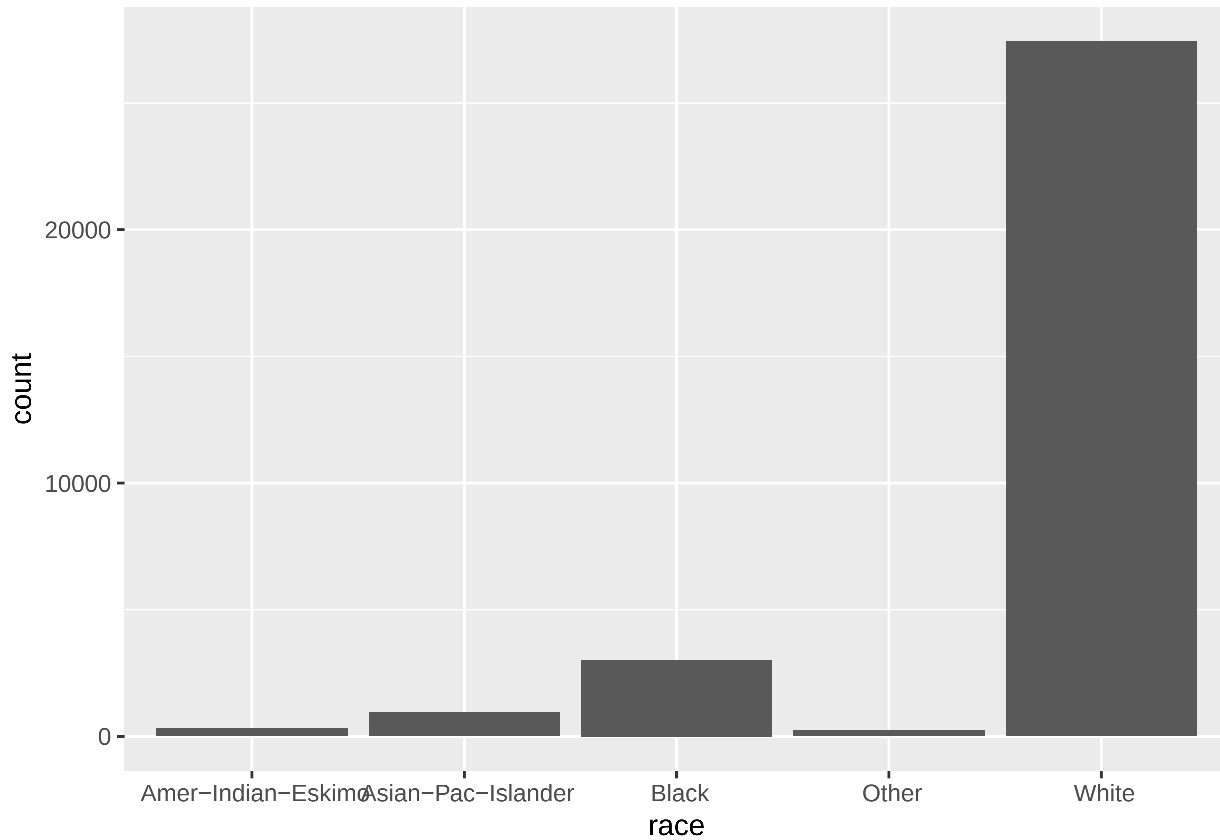
```
## [1] 0.1353383
```

```r
18/nrow(glassdoor_pay_df_final_f)
```

```
## [1] 0.03846154
```

##Racial Analysis

```r
ggplot(census_df_final, aes(x=race)) + geom_bar()
```




```r
#separate df by race
census_df_final <- data.frame(lapply(census_df_final, trimws), stringsAsFactors = FALSE)
census_df_final_w <- census_df_final %>% filter(sex == 'Male')
census_df_final_b <- census_df_final %>% filter(race == 'Black')
census_df_final_o <- census_df_final %>% filter(race == c('Asian-Pac-Islander','Amer-Indian-Eskimo'))

#percentage of population with income over 50k
table(census_df_final_w$over50k)
```

```
##
## <=50K  >50K
## 14837  6533
```

```r
6533/nrow(census_df_final_w)
```

```
## [1] 0.3057089
```

```r
table(census_df_final_b$over50k)
```

```
##
## <=50K  >50K
##  2654   374
```

```r
374/nrow(census_df_final_b)
```

```
## [1] 0.1235139
```

```r
#education including incomplete college
table(census_df_final_w$education)
```

```
##
##         10th         11th         12th      1st-4th      5th-6th      7th-8th
##          630          736          277          117          239          473
##          9th   Assoc-acdm    Assoc-voc    Bachelors    Doctorate      HS-grad
##          363          640          874         3625          308         7018
##      Masters    Preschool  Prof-school Some-college
##         1150           34          470         4416
```

```r
(3625+308+1150+4416+874+640)/nrow(census_df_final_w)
```

```
## [1] 0.5153486
```

```r
table(census_df_final_b$education)
```

```
##
##         10th         11th         12th      1st-4th      5th-6th      7th-8th
##          130          153           64           16           21           55
##          9th   Assoc-acdm    Assoc-voc    Bachelors    Doctorate      HS-grad
##           86          103          111          308            8         1144
##      Masters    Preschool  Prof-school Some-college
##           81            5           15          728
```

```r
(308+8+81+728+103+111)/nrow(census_df_final_b)
```

```
## [1] 0.4422061
```

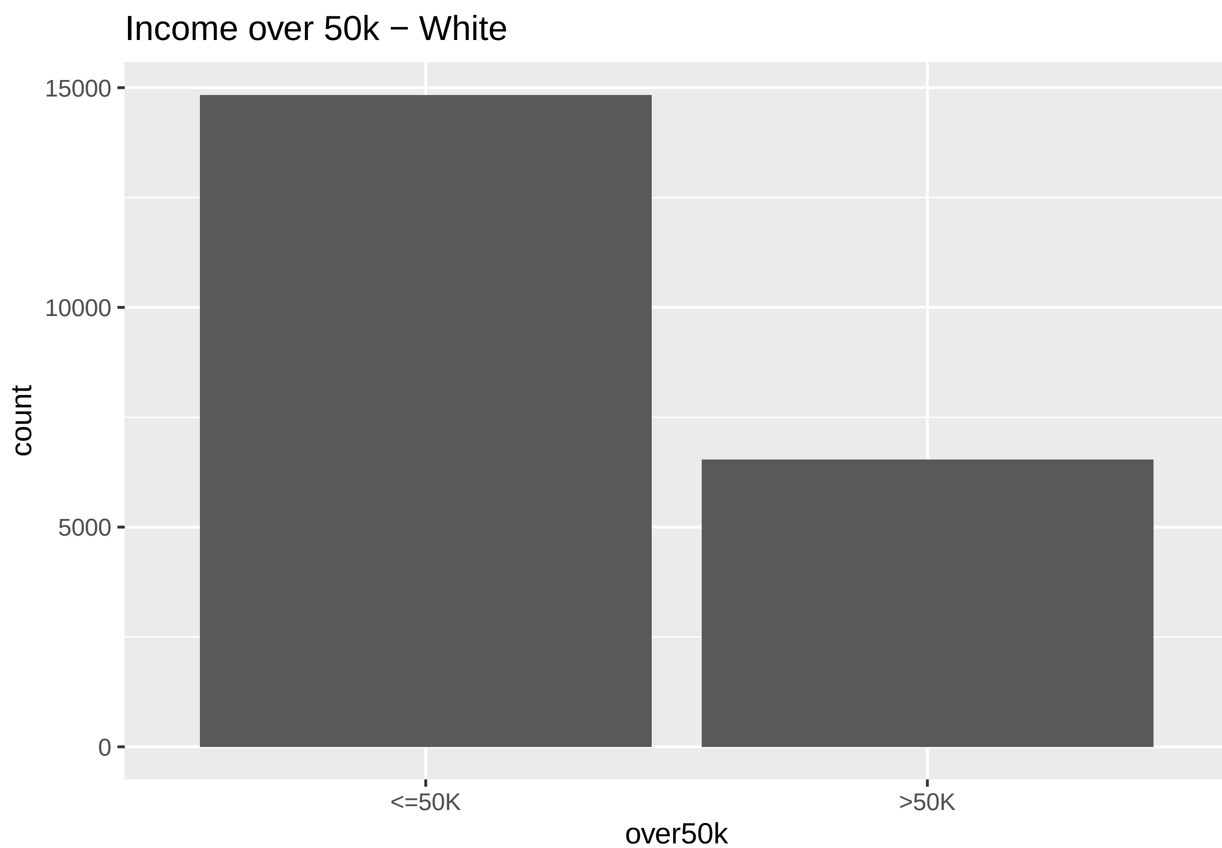
```r
#education with college grads
(3625+308+1150+874+640)/nrow(census_df_final_w)
```

```
## [1] 0.3087038
```

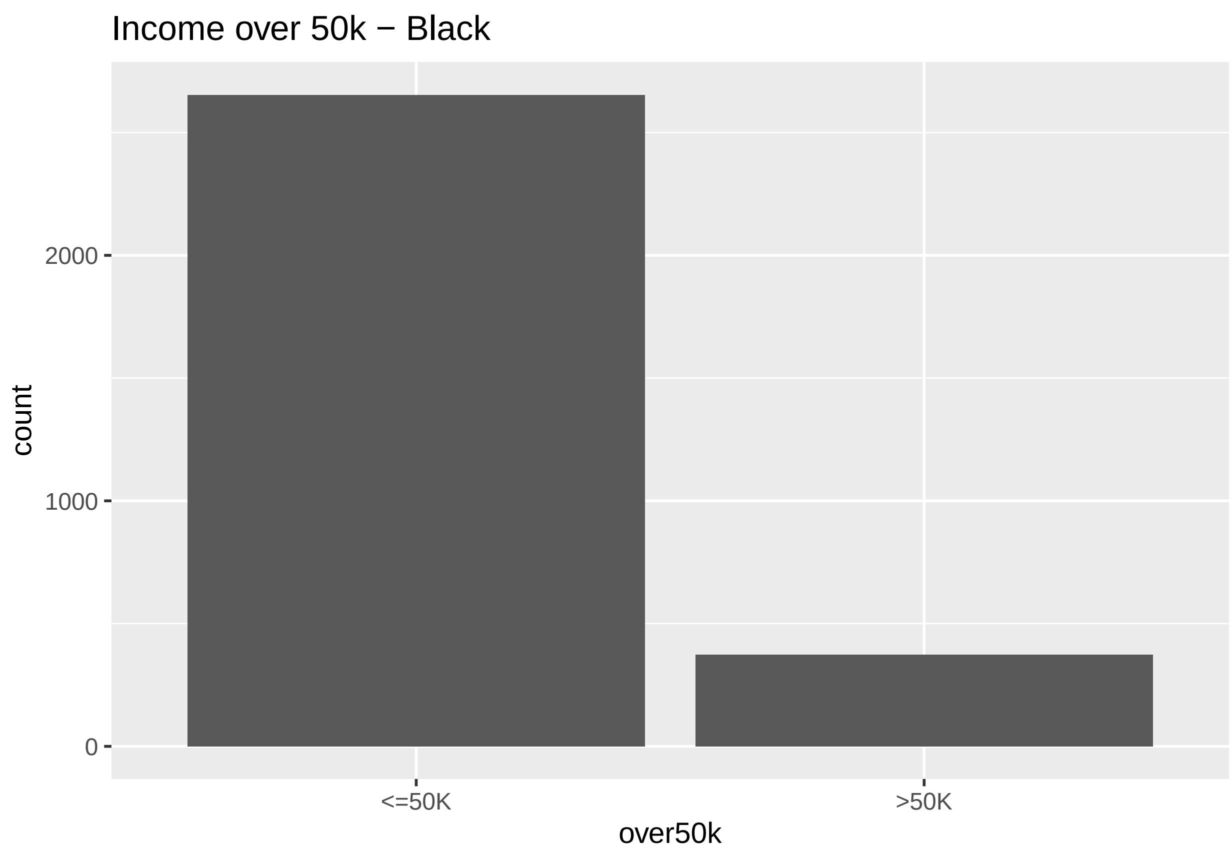
```r
(308+8+81+111+103)/nrow(census_df_final_b)
```

```
## [1] 0.2017834
```

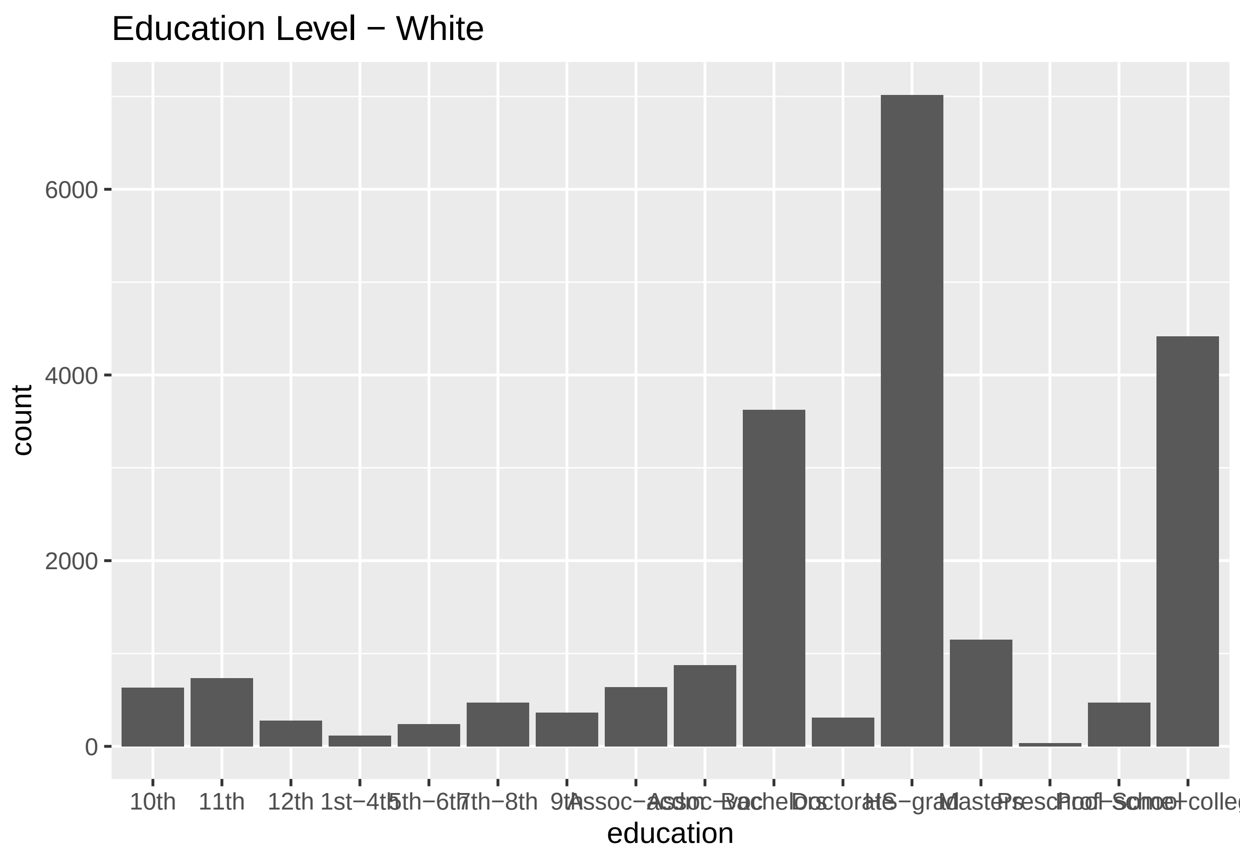
```r
#graphs
census_50k_wplot <- ggplot(census_df_final_w, aes(x=over50k)) + geom_bar() + ggtitle('Income over 50k -
census_Ed_wplot <- ggplot(census_df_final_w, aes(x=education)) + geom_bar() + ggtitle('Education Level -
census_50k_bplot <- ggplot(census_df_final_b, aes(x=over50k)) + geom_bar() + ggtitle('Income over 50k -
census_Ed_bplot <- ggplot(census_df_final_b, aes(x=education)) + geom_bar() + ggtitle("Education Level -

census_50k_wplot
```



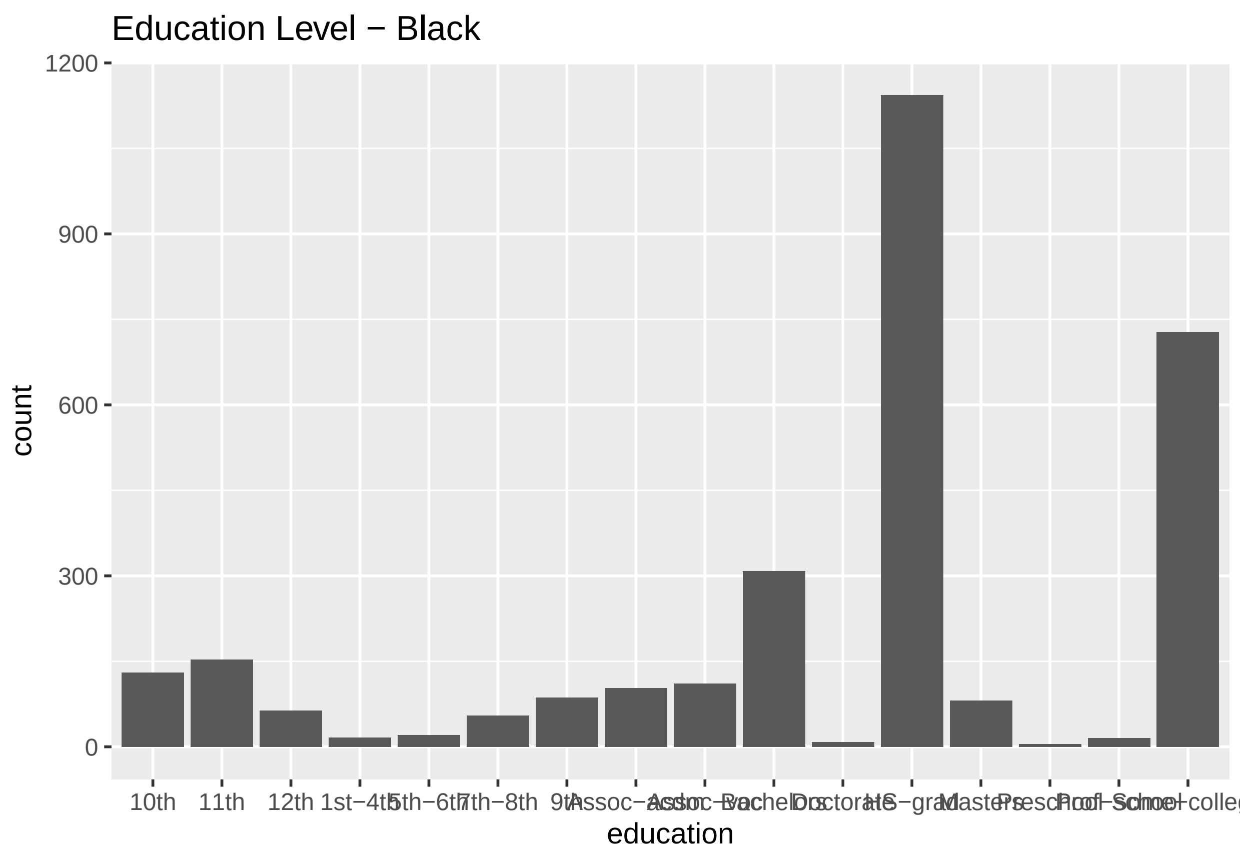

```
census_50k_bplot
```



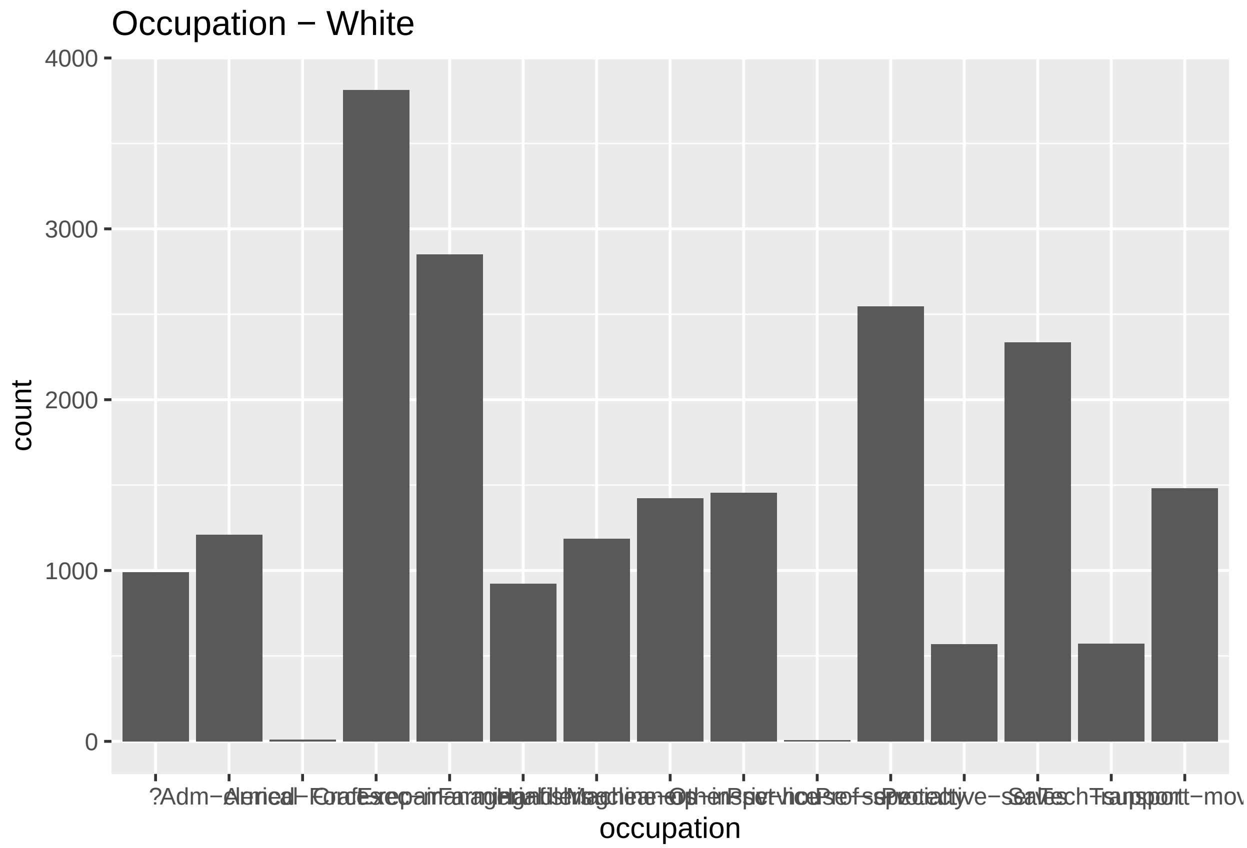

```
census_Ed_wplot
```



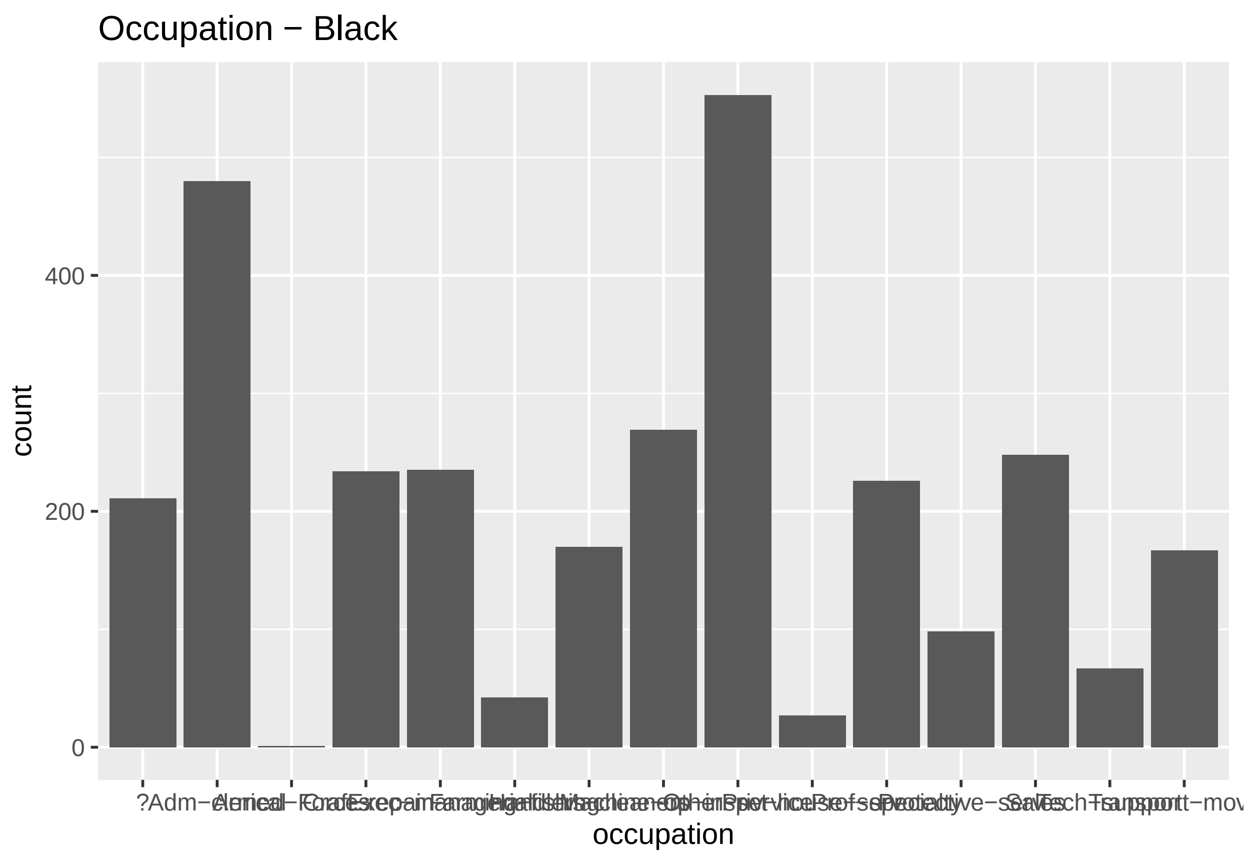

```
census_Ed_bplot
```




```
ggplot(census_df_final_w, aes(x=occupation)) + geom_bar() + ggtitle("Occupation - White")
```




```
ggplot(census_df_final_b, aes(x=occupation)) + geom_bar() + ggtitle('Occupation - Black')
```




```
#economic inequality ratio
pinc_df_whites$GINI.Ratio[1]
```

```
## [1] "0.52"
```

```
pinc_df_afr_amer$GINI.Ratio[1]
```

```
## [1] "0.506"
```

#Analysis Review > My analysis broke down differences in economic equality between Ages, Education Level, Seniority, Gender and Race.

> In my analysis on the college datasets it showed that younger workers and recent college graduates have much higher underemployment and unemployment rates compared to the total workforce and college graduates of all ages. The unemployment rate for recent college graduates peaked at around 13% in 2020 due to COVID while older graduates had an unemployment rate below 8%. Younger workers in general peaked at over 20% and the total workforce peaked at 12.5 again highlighting the difficulties young Americans have in entering the workforce. My analysis on the college data sets also showed there is a strong difference in median career income dependent on the college major chosen. A share of workers of a graduate degree had a very weak correlation to median career income but was not statistically significant based on this dataset.

> Looking at the Glass door Data set, it confirmed the common knowledge that men make more than women looking at the Scatter plot comparing their base pay. The Scatter plot on Education shows that men have a higher share of Master's and PHD degrees compared to women. Looking at Seniority it appears Women have a disproportionately small share of workers with 4 years of experience. I also looked at the correlation between age and seniority with base pay. There is a positive correlation (0.56 and 0.51) but it is not a strong positive correlation. Men also had a higher share of managers compared with women at 13.5% compared to 3.8% for women.

Looking at racial income differences I founded on the differences between Caucasians and African Americans as it is a common example used in racial discrepancies. This data showed that 30% of Whites make over 50k while only 12% of African Americans do. There is also a discrepancy in the education levels between races. White Americans have 51.5% who at least have attended some college compared to 30.5% of African Americans. Looking at those who completed college the difference drops somewhat to 30.8% for Caucasians compared to 20.1% for African Americans. The GINI ratio between the two also reflects some of the income differences with whites having a score of 0.52 (1.0 being perfect equality) compared to 0.50 for African Americans.

Implications

The implications of this analysis show that beyond gender and racial differences there are other factors that are impacting economic equality in America. Knowing some of these factors that also impact economic inequality and their levels of difference will help lawmakers and employers by setting up programs to help mitigate the differences. Some examples of this may be setting up scholarships specifically for minorities to help them achieve higher education or if the lack of women with four years of seniority is caused by pregnancies related absences perhaps setting up programs to help them navigate a career and motherhood at the same time.

Limitations

Limitations I encountered included the fact that the Pinc data sets had 50 columns which was difficult to work with so I ended up using it as a supplemental piece. I also struggled to calculate the correlation for some variables as they were categorical which could have been solved by replacing the string data with numerical values that represent them. I was able to convert them originally but could not find a way for them to be used accurately so I left them as categorical. These data sources may also be limited in whether they accurately reflect the total US population.

Concluding Remarks

In conclusion, there are multiple variables that have an effect on income. Age and Education both have a positive correlation to income levels. Experience or Seniority in your job also plays a role in income levels. Gender and Racial differences may be attributed partially to some of these differences.