Final Project Step 3

Jahedur Rahman

2/22/2022

Introduction.

For many years people have been arguing over inequality in many different areas. Some things that are talked about are gender inequality or race inequality. The gender pay gap is something that is debated and argued every year for many years. It has been researched and studied that males have a higher income than females. However, other variables may or may not be affecting this difference in income. Employers look at a number of different information of a person before hiring them. The relationships between the different information can show how much does each variable affect the income. This research project and the data obtained tells the story of how different variables have an affect on income.

The problem statement you addressed.

Variables such as age, education, and experience have an affect on how much a person earns.

How you addressed this problem statement

Glassdoor Gender Pay Gap: https://www.kaggle.com/nilimajauhari/glassdoor-analyze-gender-pay-gap
Incomes by Career and Gender: https://www.kaggle.com/jonavery/incomes-by-career-and-gender
Income Classification: https://www.kaggle.com/lodetomasi1995/income-classification

I used the data from the three sources above. I renamed and selected the columns I would be comparing. Also, I extracted the rows of data in Income Classification that have United States as the native country. After cleaning up and creating the final data sets I plotted the variables to compare the relationships between each variable. In addition, I found the correlation and covariance between the variables.

Analysis.

```
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
```

```
# Load "Glassdoor-Gender-Pay-Gap.csv" to orig_glassdoor_pay_df
orig_glassdoor_pay_df <- read.csv('Glassdoor-Gender-Pay-Gap.csv')

# Load "inc_occ_gender.csv" to orig_weekly_income_df
orig_weekly_income_df <- read.csv('inc_occ_gender.csv')

# Load "income_evaluation.csv" to orig_income_evaluation_df
orig_income_evaluation_df <- read.csv('income_evaluation.csv')

# rename columns of orig_glassdoor_pay_df
orig_glassdoor_pay_df <- orig_glassdoor_pay_df %>%
  rename(gender = Gender,
         age = Age,
         education = Education,
         experience = Seniority,
         annual_income = BasePay)

# select columns from orig_glassdoor_pay_df to glassdoor_pay_df
glassdoor_pay_df <- orig_glassdoor_pay_df %>%
  select(gender, age, education, experience, annual_income)

# find the mean annual_income based on all the other columns
glassdoor_pay_df <- glassdoor_pay_df %>%
  group_by(age, education, experience, gender) %>%
  summarize(mean_annual_income = mean(annual_income)) %>%
  ungroup()
```

```
## ‘summarise()‘ has grouped output by 'age', 'education', 'experience'. You can override using the ‘.g
```

```
# further condense the data by looking for the outliers and IQR
boxplot.stats(glassdoor_pay_df$mean_annual_income)$out
```

```
## [1] 163208 179726 176789
```

```
lower_lim = quantile(glassdoor_pay_df$mean_annual_income, 0.25)

upper_lim = quantile(glassdoor_pay_df$mean_annual_income, 0.75)

glassdoor_IQR <- which(glassdoor_pay_df$mean_annual_income > lower_lim & glassdoor_pay_df$mean_annual_i

glassdoor_pay_df_IQR <- glassdoor_pay_df[glassdoor_IQR,]

# rename columns of orig_weekly_income_df
orig_weekly_income_df <- orig_weekly_income_df %>%
  rename(number_male_workers = M_workers,
         male_median_weekly_income = M_weekly,
         number_female_workers = F_workers,
         female_median_weekly_income = F_weekly)
```

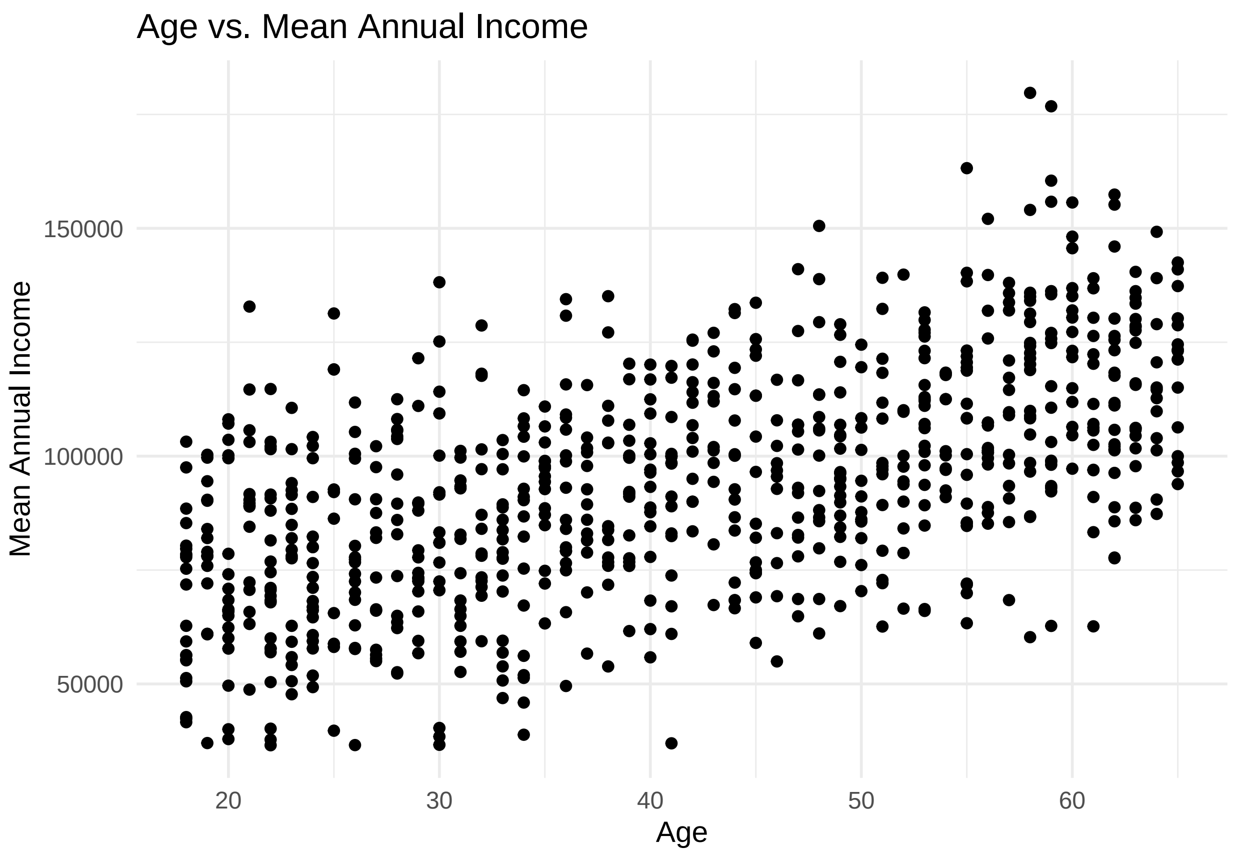
```
# select columns from orig_weekly_income_df to weekly_income_df
weekly_income_df <- orig_weekly_income_df %>%
  select(number_male_workers, male_median_weekly_income, number_female_workers, female_median_weekly_in

# only need the first row because they are the total number
weekly_income_df <- weekly_income_df[1,]

# in orig_income_evaluation_df extract rows where native-country = " United-States" since the other dat
orig_income_evaluation_df <- orig_income_evaluation_df %>%
  filter(native.country == " United-States")

# rename columns of orig_income_evaluation_df
orig_income_evaluation_df <- orig_income_evaluation_df %>%
  rename(gender = sex)

# select columns from orig_income_evaluation_df to race_education_df
race_education_df <- orig_income_evaluation_df %>%
  select(education, race)

# there is a leading white space on all the values in the data, so this removes it
race_education_df <- data.frame(lapply(race_education_df, trimws), stringsAsFactors = FALSE)

# there are some values under education that does not apply to this analysis, so this removes them
race_education_df <- race_education_df %>%
  filter(!education %in% c("Preschool", "1st-4th", "5th-6th", "7th-8th", "Prof-school"))

# change all 9th, 10th, 11th, and 12th education values to Some-HS and Assoc-acdm and Assoc-voc to Asso
race_education_df <- race_education_df %>%
  mutate(education = recode(education, "9th" = "Some-HS", "10th" = "Some-HS", "11th" = "Some-HS", "12th

# tally the total based on race and education
race_education_df <- race_education_df %>%
  count(race, education, name = "total")

cor(glassdoor_pay_df$age, glassdoor_pay_df$mean_annual_income)
```

```
## [1] 0.5706051
```

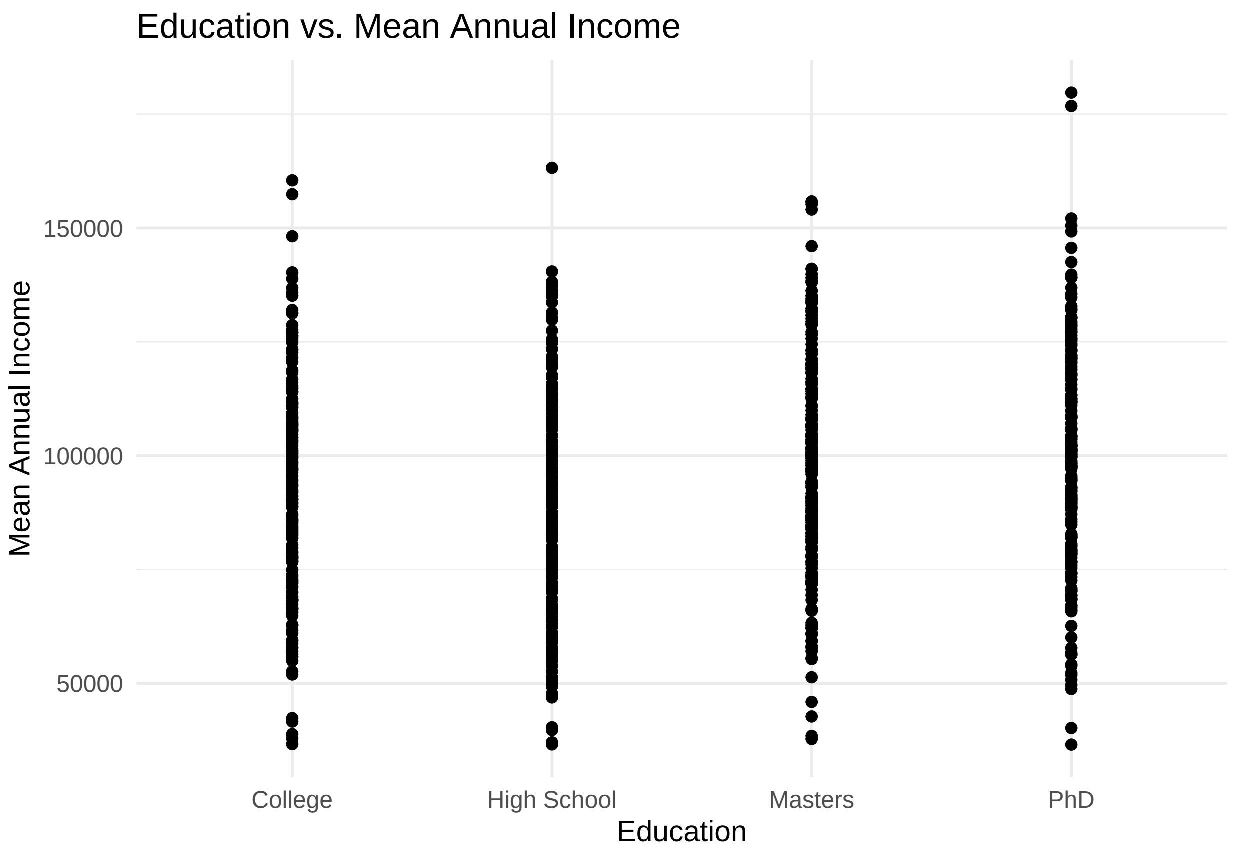
```
cor(glassdoor_pay_df$experience, glassdoor_pay_df$mean_annual_income)
```

```
## [1] 0.5271802
```

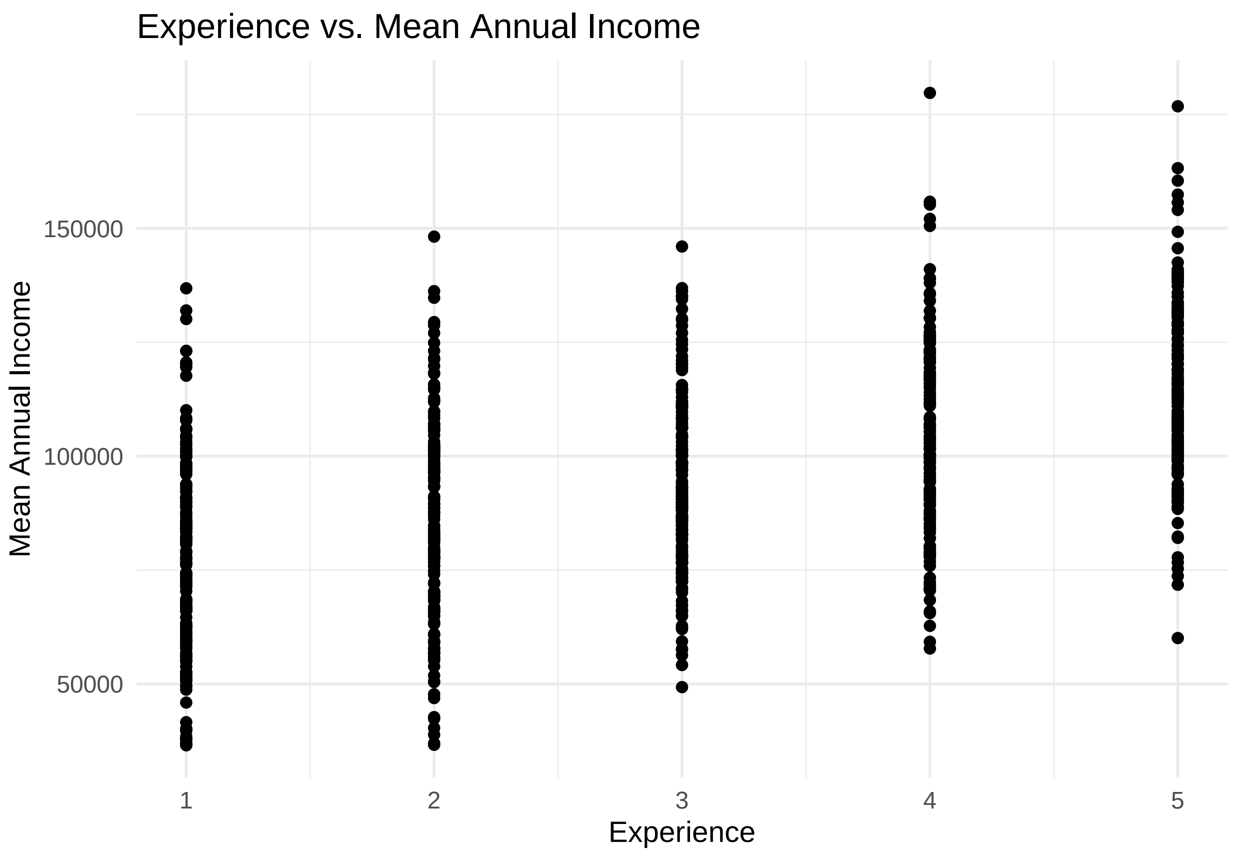
```
## Load the ggplot2 package
library(ggplot2)
theme_set(theme_minimal())

## Using `geom_point()` create scatterplots for
## `age` vs. `mean_annual_income`
ggplot(glassdoor_pay_df, aes(x=age, y=mean_annual_income)) + geom_point() + ggtitle("Age vs. Mean Annual
```



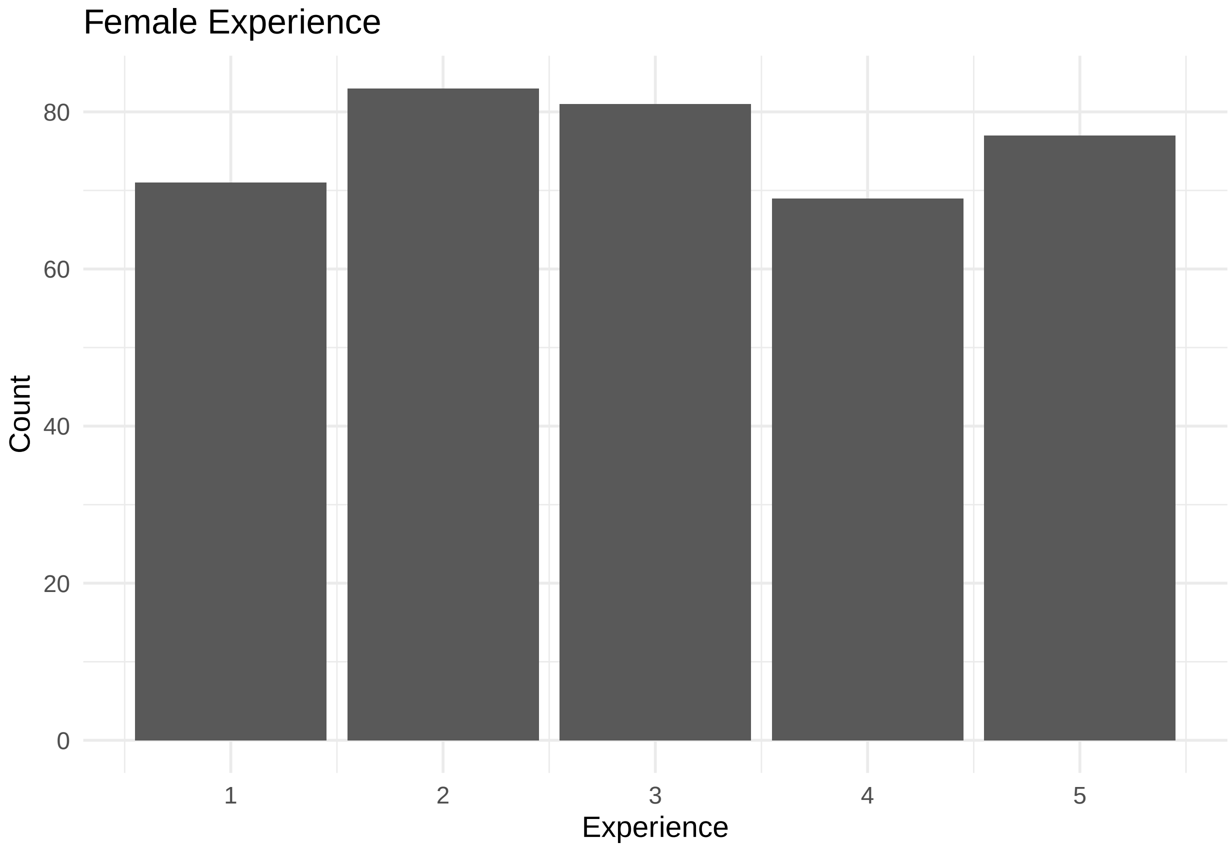
```
## `education` vs. `mean_annual_income`
ggplot(glassdoor_pay_df, aes(x=education, y=mean_annual_income)) + geom_point() + ggtitle("Education vs
```



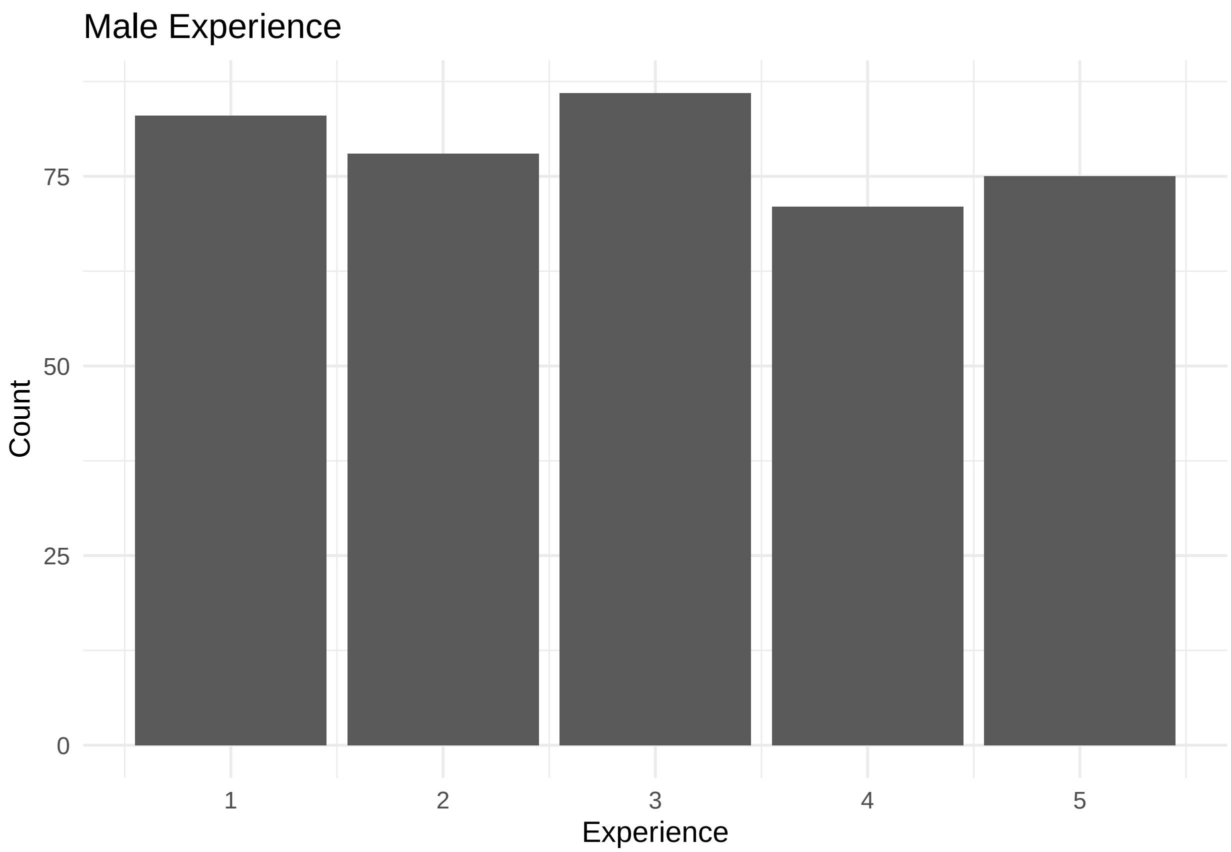
```
## `experience` vs. `mean_annual_income`
ggplot(glassdoor_pay_df, aes(x=experience, y=mean_annual_income)) + geom_point() + ggtitle("Experience v
```



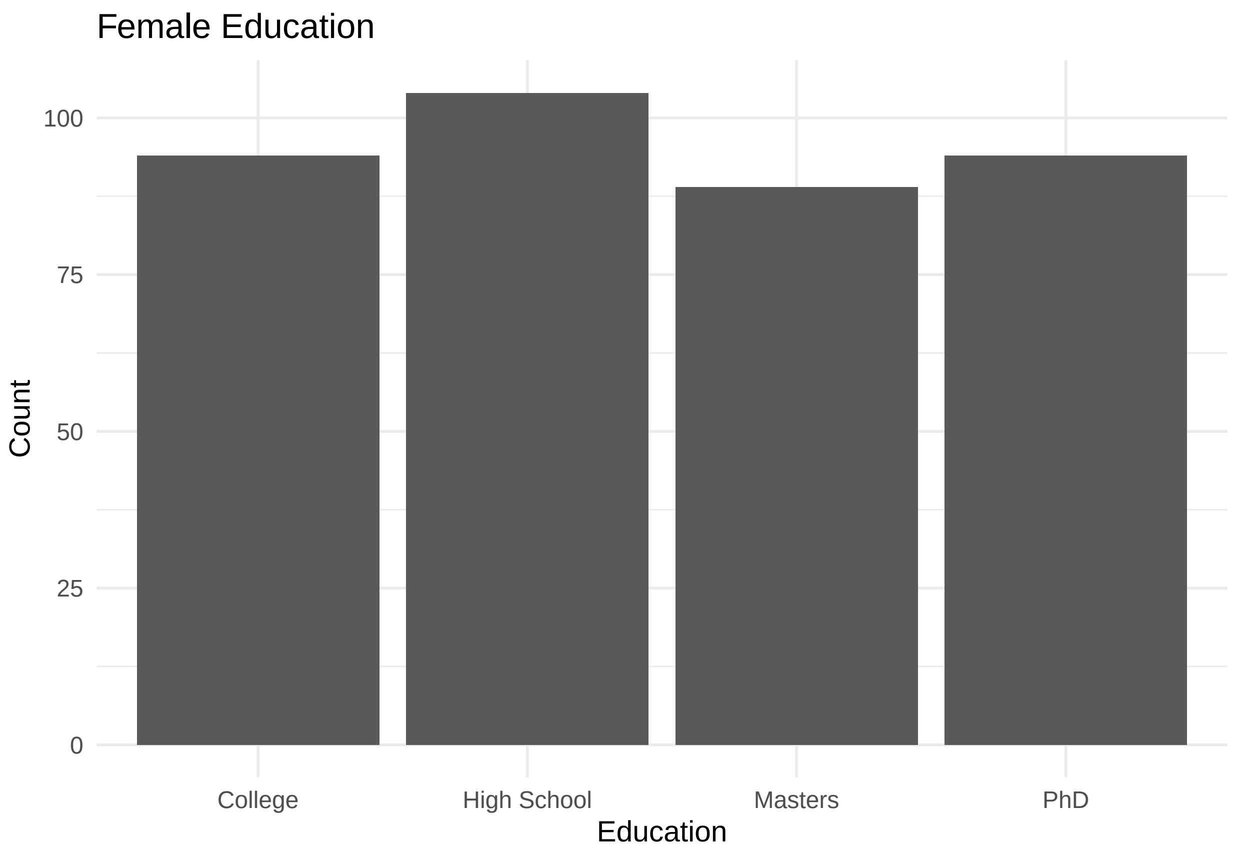
```
# Creating separate data frames based on gender
glassdoor_pay_df_female <- glassdoor_pay_df %>%
  filter(gender == "Female")
glassdoor_pay_df_male <- glassdoor_pay_df %>%
  filter(gender == "Male")

# Bar graphs for gender vs experience
ggplot(glassdoor_pay_df_female, aes(experience)) + geom_bar() + ggtitle("Female Experience") + xlab("Exp
```



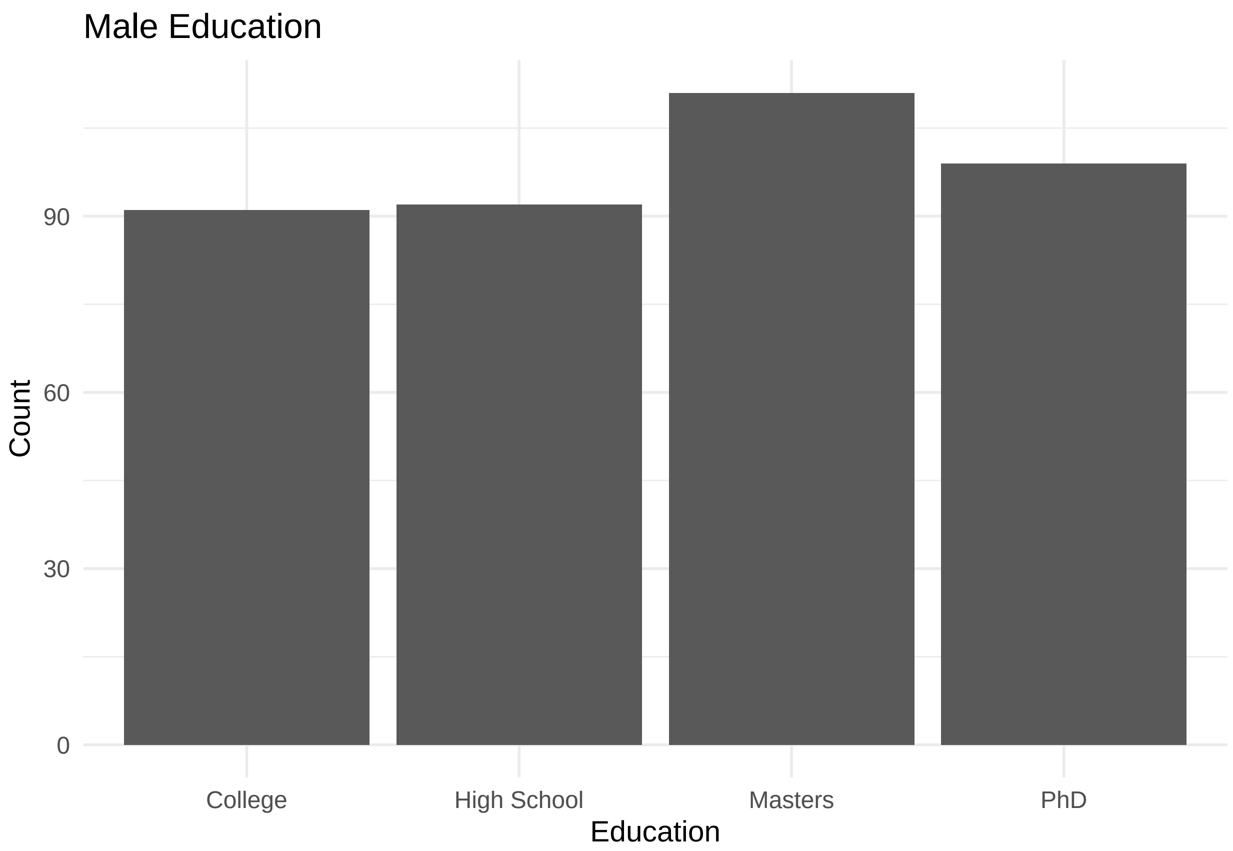
```
ggplot(glassdoor_pay_df_male, aes(experience)) + geom_bar() + ggtitle("Male Experience") + xlab("Experi
```



```
# Bar graphs for gender vs education
ggplot(glassdoor_pay_df_female, aes(education)) + geom_bar() + ggtitle("Female Education") + xlab("Educa
```



```
ggplot(glassdoor_pay_df_male, aes(education)) + geom_bar() + ggtitle("Male Education") + xlab("Educatio
```



```
race_education_df
```

```
##                   race    education total
## 1  Amer-Indian-Eskimo   Associates    25
## 2  Amer-Indian-Eskimo    Bachelors    19
## 3  Amer-Indian-Eskimo    Doctorate     3
## 4  Amer-Indian-Eskimo      HS-grad   117
## 5  Amer-Indian-Eskimo      Masters     5
## 6  Amer-Indian-Eskimo Some-college    78
## 7  Amer-Indian-Eskimo      Some-HS    36
## 8  Asian-Pac-Islander   Associates    36
## 9  Asian-Pac-Islander    Bachelors    66
## 10 Asian-Pac-Islander      HS-grad    76
## 11 Asian-Pac-Islander      Masters    12
## 12 Asian-Pac-Islander Some-college    85
## 13 Asian-Pac-Islander      Some-HS     7
## 14              Black   Associates   198
## 15              Black    Bachelors   286
## 16              Black    Doctorate     7
## 17              Black      HS-grad  1087
## 18              Black      Masters    76
## 19              Black Some-college   673
## 20              Black      Some-HS   411
## 21              Other   Associates    11
## 22              Other    Bachelors    15
## 23              Other    Doctorate     1
## 24               Other      HS-grad   38
## 25               Other      Masters    3
## 26               Other Some-college   32
## 27               Other      Some-HS   21
## 28               White   Associates 2001
## 29               White    Bachelors 4380
## 30               White    Doctorate  317
## 31               White      HS-grad 8384
## 32               White      Masters 1431
## 33               White Some-college 5872
## 34               White      Some-HS 2200
```

The correlation for age and income is 0.57, and for experience and income is 0.53. They are positively correlated but it is not a strong correlation. The scatter plot for Age vs Mean Annual Income shows that as age increases income increases. The scatter plot for Experience vs Mean Annual Income shows that more experience gives higher ranges of income. The scatter plot for Education vs Mean Annual Income doesn't show us significant information. The only thing that really stands out is the 2 high incomes for PhD level of education. The bar charts for gender vs experience don't show a big difference. One thing that stands out is that most females have 2 years of experience and most males had 3 years of experience. The bar charts for gender vs education shows us that most females had high school level education, and most males had masters level education. This is a significant difference. Looking at the race_education_df we can see that clearly that education of the white race is higher than any other race by a lot. This analysis tells us that age and experience do have some affect on income. The older and/or more experienced person earns more. In addition, gender does have an affect on education. Males have a higher chance of getting a Master's Degree while Females have a higher chance of getting a High School Degree.

Implications.

This research uses data with no intended bias. The data highlights how certain information can affect income.

Limitations.

The limitations of this analysis is that it is difficult to calculate the covariance and correlation of all the data. This is because both require numerical data. This can be fixed by replacing the string values with numerical values that can represent the string values. In addition, the data for race_education_df may not be an accurate representation of the relationship. This can be fixed by finding another data source that may have more accurate information.

Concluding Remarks

In conclusion, the analysis does show that there are other variables that have an affect on income. Age and experience were obvious variables to have an affect on income. Education may have a small affect on income. The analysis that surprised me the most is the gender vs education. While most females go up to receiving High School degrees, most males receive Master's Degrees. Even though the analysis for education doesn't have much of an affect on income, this different is puzzling. It can lead to a greater analysis of gender vs education. Additionally, the analysis for race vs education shows that the white race have a great amount of people with any sort of education. However, I feel the information may not be accurate because of the very large differences in numbers. To sum everything up, age and experience definitely have an affect on income earned.