Final Project Step 2

Jahedur Rahman

2/18/2022

How to import and clean my data and how do you plan to slice and dice the data?

Out of the three data sets I have one of the data sets "Glassdoor-Gender-Pay-Gap.csv" has almost all the variables and information I need. So this will be my main data set and the other two data sets will be used as supporting data sets whenever and wherever they are needed. I will rename and select columns that I need.

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

```
##                           Occupation All_workers All_weekly M_workers M_weekly
## 1                     ALL OCCUPATIONS      109080        809     60746      895
## 2                          MANAGEMENT       12480       1351      7332     1486
## 3                   Chief executives        1046       2041       763     2251
## 4     General and operations managers         823       1260       621     1347
## 5                           Legislators           8         Na         5       Na
## 6 Advertising and promotions managers          55       1050        29       Na
##   F_workers F_weekly
## 1     48334      726
## 2      5147     1139
## 3       283     1836
## 4       202     1002
## 5         4       Na
## 6        26       Na
```

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

```
##   age         workclass fnlwgt  education education.num      marital.status
## 1  39         State-gov  77516  Bachelors            13       Never-married
## 2  50  Self-emp-not-inc  83311  Bachelors            13  Married-civ-spouse
## 3  38           Private 215646    HS-grad             9            Divorced
## 4  53           Private 234721       11th             7  Married-civ-spouse
## 5  28           Private 338409  Bachelors            13  Married-civ-spouse
## 6  37           Private 284582    Masters            14  Married-civ-spouse
##           occupation   relationship   race     sex capital.gain capital.loss
## 1       Adm-clerical  Not-in-family  White    Male         2174            0
## 2    Exec-managerial        Husband  White    Male            0            0
## 3  Handlers-cleaners  Not-in-family  White    Male            0            0
## 4  Handlers-cleaners        Husband  Black    Male            0            0
## 5     Prof-specialty           Wife  Black  Female            0            0
## 6    Exec-managerial           Wife  White  Female            0            0
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             13  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40           Cuba  <=50K
## 6             40  United-States  <=50K
```

```r
# 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)
head(orig_glassdoor_pay_df)
```

```
##              JobTitle gender age PerfEval education           Dept experience
## 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
##   annual_income Bonus
## 1         42363  9938
## 2        108476 11128
## 3         90208  9268
## 4        108080 10154
## 5         99464  9319
## 6         70890 10126
```

```
# 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)
head(glassdoor_pay_df)
```

```
##   gender age education experience annual_income
## 1 Female  18   College          2         42363
## 2   Male  21   College          5        108476
## 3 Female  19       PhD          5         90208
## 4   Male  20   Masters          4        108080
## 5   Male  26   Masters          5         99464
## 6 Female  20       PhD          4         70890
```

```
# 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
```

```
head(glassdoor_pay_df)
```

```
## # A tibble: 6 x 5
##     age education   experience gender mean_annual_income
##   <int> <chr>            <int> <chr>                <dbl>
## 1    18 College              1 Female               41603
## 2    18 College              2 Female               42363
## 3    18 College              3 Female               62759
## 4    18 College              3 Male                 80355
## 5    18 College              5 Male                 85306
## 6    18 High School          1 Male                 51296.
```

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

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

```r
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)
head(orig_weekly_income_df)
```

```
##                           Occupation All_workers All_weekly
## 1                     ALL OCCUPATIONS     109080        809
## 2                          MANAGEMENT      12480       1351
## 3                    Chief executives       1046       2041
## 4     General and operations managers        823       1260
## 5                           Legislators          8         Na
## 6 Advertising and promotions managers         55       1050
##   number_male_workers male_median_weekly_income number_female_workers
## 1               60746                       895                 48334
## 2                7332                      1486                  5147
## 3                 763                      2251                   283
## 4                 621                      1347                   202
## 5                   5                        Na                     4
## 6                  29                        Na                    26
##   female_median_weekly_income
## 1                         726
## 2                        1139
## 3                        1836
## 4                        1002
## 5                          Na
## 6                          Na
```

```r
# 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
head(weekly_income_df)
```

```
##   number_male_workers male_median_weekly_income number_female_workers
## 1               60746                       895                 48334
## 2                7332                      1486                  5147
## 3                 763                      2251                   283
## 4                 621                      1347                   202
## 5                   5                        Na                     4
## 6                  29                        Na                    26
##   female_median_weekly_income
## 1                         726
```

```
## 2                       1139
## 3                       1836
## 4                       1002
## 5                         Na
## 6                         Na
```

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

```
##   number_male_workers male_median_weekly_income number_female_workers
## 1               60746                       895                 48334
##   female_median_weekly_income
## 1                         726
```

```
# 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")
head(orig_income_evaluation_df)
```

```
##   age         workclass fnlwgt  education education.num      marital.status
## 1  39         State-gov  77516  Bachelors            13       Never-married
## 2  50  Self-emp-not-inc  83311  Bachelors            13  Married-civ-spouse
## 3  38           Private 215646    HS-grad             9            Divorced
## 4  53           Private 234721       11th             7  Married-civ-spouse
## 5  37           Private 284582    Masters            14  Married-civ-spouse
## 6  52  Self-emp-not-inc 209642    HS-grad             9  Married-civ-spouse
##           occupation   relationship   race     sex capital.gain capital.loss
## 1       Adm-clerical  Not-in-family  White    Male         2174            0
## 2    Exec-managerial        Husband  White    Male            0            0
## 3  Handlers-cleaners  Not-in-family  White    Male            0            0
## 4  Handlers-cleaners        Husband  Black    Male            0            0
## 5    Exec-managerial           Wife  White  Female            0            0
## 6    Exec-managerial        Husband  White    Male            0            0
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             13  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40  United-States  <=50K
## 6             45  United-States   >50K
```

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

```
##   age         workclass fnlwgt  education education.num      marital.status
## 1  39         State-gov  77516  Bachelors            13       Never-married
## 2  50  Self-emp-not-inc  83311  Bachelors            13  Married-civ-spouse
## 3  38           Private 215646    HS-grad             9            Divorced
## 4  53           Private 234721       11th             7  Married-civ-spouse
```

```
## 5  37           Private 284582    Masters            14  Married-civ-spouse
## 6  52  Self-emp-not-inc 209642    HS-grad             9  Married-civ-spouse
##           occupation   relationship   race  gender capital.gain capital.loss
## 1       Adm-clerical  Not-in-family  White    Male         2174            0
## 2    Exec-managerial        Husband  White    Male            0            0
## 3  Handlers-cleaners  Not-in-family  White    Male            0            0
## 4  Handlers-cleaners        Husband  Black    Male            0            0
## 5    Exec-managerial           Wife  White  Female            0            0
## 6    Exec-managerial        Husband  White    Male            0            0
##   hours.per.week native.country income
## 1             40  United-States  <=50K
## 2             13  United-States  <=50K
## 3             40  United-States  <=50K
## 4             40  United-States  <=50K
## 5             40  United-States  <=50K
## 6             45  United-States   >50K
```

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

```
##    education   race
## 1  Bachelors  White
## 2  Bachelors  White
## 3    HS-grad  White
## 4       11th  Black
## 5    Masters  White
## 6    HS-grad  White
```

```
# 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")
head(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
```

What does the final data set look like?

Since it is not possible to combine the three data sets, there are three final data sets. One of them will be my main data set which will address most of the questions, and the other two will be supporting data sets which will be used whenever and wherever needed.

```
# MAIN DATA SET
glassdoor_pay_df_IQR
```

```
## # A tibble: 386 x 5
##      age education   experience gender mean_annual_income
##    <int> <chr>            <int> <chr>                <dbl>
##  1    18 College              3 Male                80355
##  2    18 College              5 Male                85306
##  3    18 High School          4 Male                77820.
##  4    18 Masters              2 Male                88482
##  5    18 Masters              4 Male                79664.
##  6    18 Masters              5 Male               103174.
##  7    18 PhD                  3 Female              78462
##  8    18 PhD                  4 Male                78270
##  9    18 PhD                  5 Male                97523
## 10    19 College              3 Female              84007
## # ... with 376 more rows
```

```
#SUPPORTING DATA SETS
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
```

```
weekly_income_df
```

```
##   number_male_workers male_median_weekly_income number_female_workers
## 1               60746                       895                 48334
##   female_median_weekly_income
## 1                         726
```

Questions for future steps.

The data sets have been condensed as much as possible. However, this has me worried that the accuracy of the analysis has decreased. The supporting data sets, "race_education_df" and "weekly_income_df", should not affect the accuracy much. However, the main data set, "glassdoor_pay_df_IQR", will have the biggest impact in my analysis. Initially this data set had 1000 rows of data.

What information is not self-evident?

Information that is not self-evident is information of the location of the data. I have made an assumption that the data is from USA. So for the "race_education_df" I have only included data that has United-States in the rows in the native-country column. In addition, I would like to have compared race and income, but the income column in "orig_income_evaluation_df" would only tell if income was greater than 50k or less than 50k.

What are different ways you could look at this data?

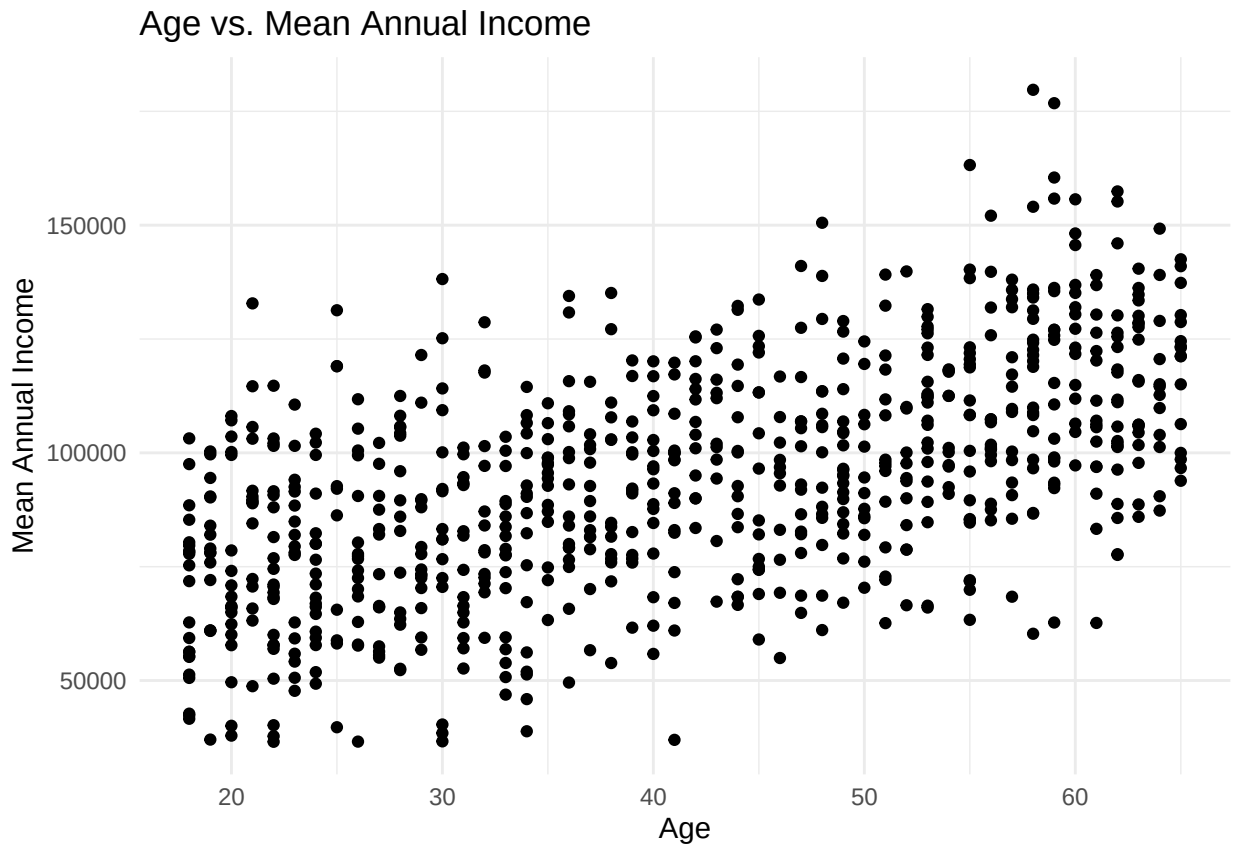
Different ways I could look at this data is by comparing different variables to each other, other than the ones I am already comparing. I could compare age and education, age and experience, education and experience, and gender and income.

How could you summarize your data to answer key questions?

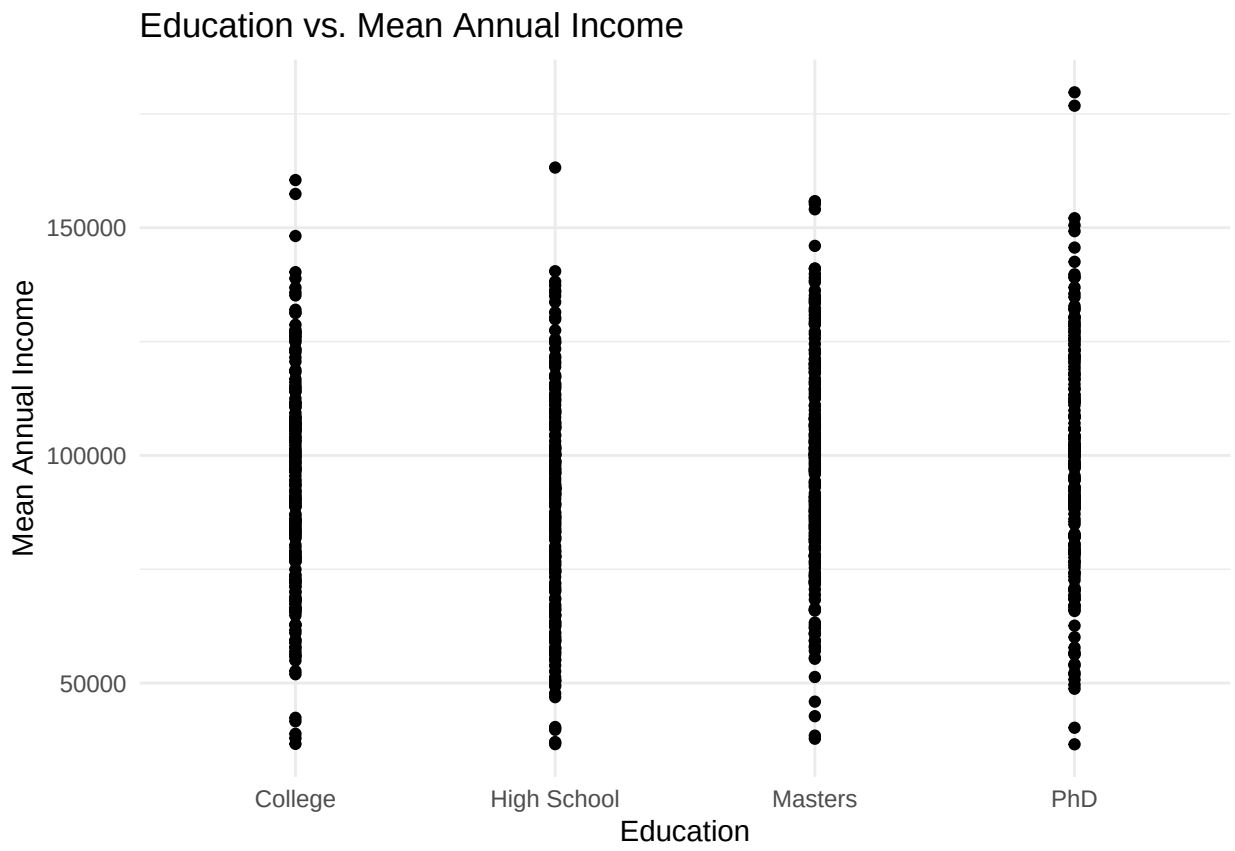
Calculating the correlation and covariance are great ways to summarize my data to answer key questions. Results from the summary function would also help. In addition, finding the maximum, minimum, mean, and median values will provide some more information.

What types of plots and tables will help you to illustrate the findings to your questions?

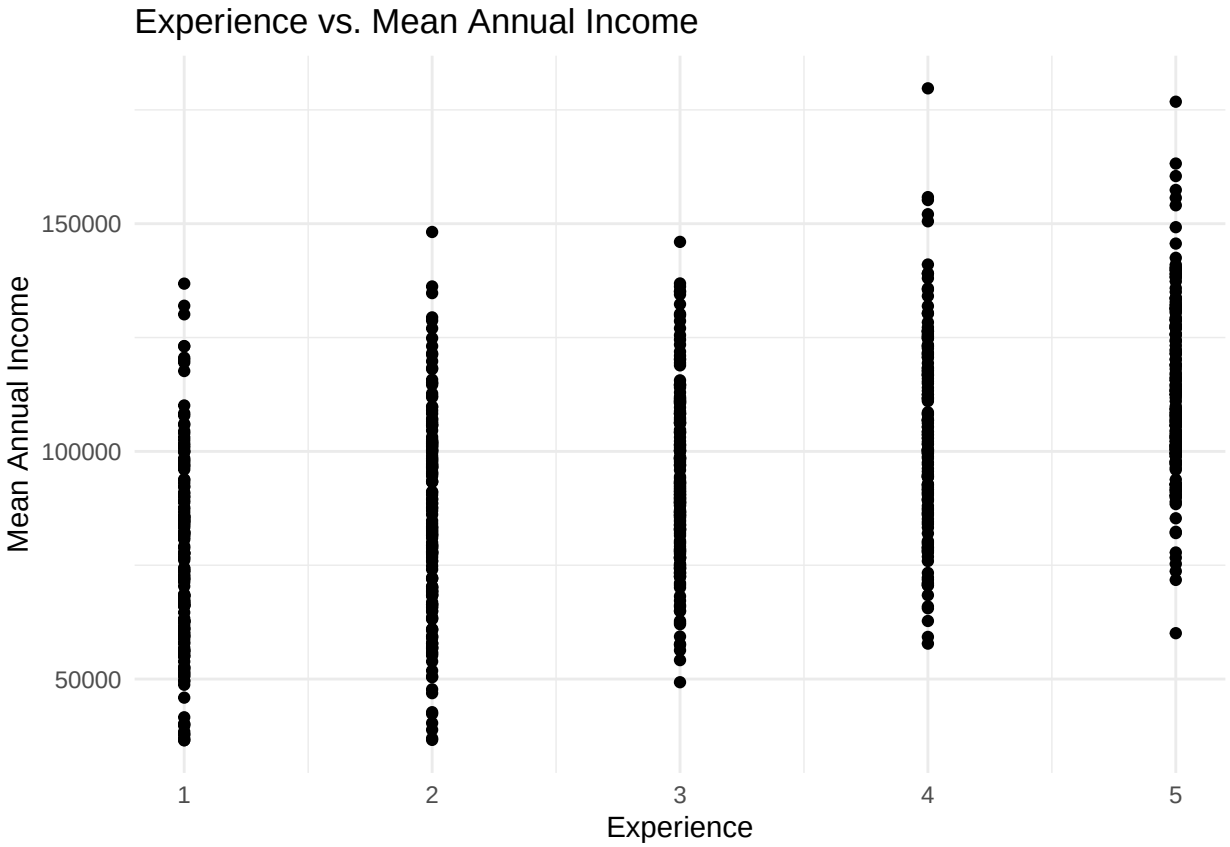
```
## Load the ggplot2 package
library(ggplot2)
theme_set(theme_minimal())

## 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
```



Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

For now I do not plan on incorporating any machine learning techniques. However, after learning how to use them and if I do see if they are useful then I will decide to use them or not.

Questions for future steps.

Is there a different way to condense the data so the data can be as accurate as possible? Should any of the questions be changed?