Exploring Salary Information from *Ask a Manager's* 2021 Survey

Lance Ding, Grayson Stark, & Kristy Tran

April 2023

For our final project, we are going to do an analysis on a data set that contains the responses from a survey conducted by *Ask a Manager* in 2021. This survey asked about the salary, demographic information, occupations, and education levels of the individuals who responded. In this project, we will import the data set and any necessary packages, perform extensive cleaning on the data set, and use visualizations to answer self-proposed research questions.

Importing Packages & the Dataset

Before we begin, we are going to import the packages that we will be using for our analysis: `tidyverse`, `lubridate`, and `ggplot2`. We will also import the `salary` data set, which we will be using for this analysis, after we downloaded it from the *Ask a Manager* website and saved it to our working directory.

```
pacman::p_load(tidyverse, lubridate, ggplot2)
salary <- read.csv("D:/Projects/QTM150/Datasets/salary.csv")
```

Exploring the Dataset

The uncleaned data looks like this:

```
head(tibble(salary))
```

```
## # A tibble: 6 x 24
##   Timestamp          How.old.are.you. What.industry.do.you.work.in. Job.title
##   <chr>              <chr>            <chr>                         <chr>
## 1 4/27/2021 11:02:10 25-34            Education (Higher Education)  Research an~
## 2 4/27/2021 11:02:22 25-34            Computing or Tech             Change & In~
## 3 4/27/2021 11:02:38 25-34            Accounting, Banking & Finance Marketing S~
## 4 4/27/2021 11:02:41 25-34            Nonprofits                    Program Man~
## 5 4/27/2021 11:02:42 25-34            Accounting, Banking & Finance Accounting ~
## 6 4/27/2021 11:02:46 25-34            Education (Higher Education)  Scholarly P~
## # i 20 more variables:
## #   If.your.job.title.needs.additional.context..please.clarify.here. <chr>,
## #   What.is.your.annual.salary...You.ll.indicate.the.currency.in.a.later.question..If.you.are.part.ti
## #   How.much.additional.monetary.compensation.do.you.get..if.any..for.example..bonuses.or.overtime.in
## #   Please.indicate.the.currency <chr>,
## #   If..Other...please.indicate.the.currency.here.. <chr>,
## #   If.your.income.needs.additional.context..please.provide.it.here. <chr>, ...
```

As you can see, the data is organized in a long sort of format, with observations going down the dataset, and characteristics of the observations going across the dataset. To take a closer, more technical look at the dataset, we employ the help of `str()`.

```
str(salary)
```

```
## 'data.frame':    28043 obs. of  24 variables:
##  $ Timestamp
##  $ How.old.are.you.
##  $ What.industry.do.you.work.in.
##  $ Job.title
##  $ If.your.job.title.needs.additional.context..please.clarify.here.
##  $ What.is.your.annual.salary...You.ll.indicate.the.currency.in.a.later.question..If.you.are.part.ti
##  $ How.much.additional.monetary.compensation.do.you.get..if.any..for.example..bonuses.or.overtime.in
##  $ Please.indicate.the.currency
##  $ If..Other...please.indicate.the.currency.here..
##  $ If.your.income.needs.additional.context..please.provide.it.here.
##  $ What.country.do.you.work.in.
##  $ If.you.re.in.the.U.S...what.state.do.you.work.in.
##  $ What.city.do.you.work.in.
##  $ How.many.years.of.professional.work.experience.do.you.have.overall.
##  $ How.many.years.of.professional.work.experience.do.you.have.in.your.field.
##  $ What.is.your.highest.level.of.education.completed.
##  $ What.is.your.gender.
##  $ What.is.your.race...Choose.all.that.apply..
##  $ X
##  $ X.1
##  $ X.2
##  $ X.3
##  $ X.4
##  $ X.5
```

We now know that the dataset contains 28,043 rows and 24 columns. Each row corresponds to the answers of one respondee to the survey, while each column corresponds to the different questions that were asked in the survey. Of the 24 columns, there are 17 character columns, 1 numeric column, and 6 logical columns. A quick glance at the actual data reveals that the 6 logical columns are empty columns, probably generated from the porting of the dataset. The 1 numeric column asks the individuals about their annual income which was reported as a number. The other 17 character columns correspond to the other 17 questions that asked about the respondees' age range, occupation, location of residence, experience, education level, gender, and race.

Data Dictionary

This is the data dictionary of our original, pre-cleaning dataset:

- **Timestamp** (`character`)
  - The submission date and time of the survey. Comes in the form *M/D/Y 24H:M:S*
- **How old are you?** (`character`)
  - The age of the respondee in years, divided into age categories. Possible values are:
    - under 18

    * 18-24
    * 25-34
    * 35-44
    * 45-54
    * 55-64
    * 65 or over
- **What industry do you work in?** (`character`)
  - Field that the respondee works in. Possible values are:
    * Accounting, Banking & Finance
    * Agriculture or Forestry
    * Art & Design
    * Business or Consulting
    * Computing or Tech
    * Education (Primary/Secondary)
    * Education (Higher Education)
    * Engineering or Manufacturing
    * Entertainment
    * Government & Public Administration
    * Government Affairs & Lobbying
    * Health care
    * Hospitality & Events
    * Insurance
    * Law
    * Law Enforcement & Security
    * Leisure, Sport & Tourism
    * Marketing, Advertising & PR
    * Media & Digital
    * Nonprofits
    * Property or Construction
    * Recruitment or HR
    * Retail
    * Sales
    * Science
    * Social Work
    * Transport or Logistics
    * Utilities & Telecommunications
    * <Other>
- **Job title** (`character`)
  - The formal title of the position that the respondee holds.
- **If your job title needs additional context, please clarify here:** (`character`)
  - The additional context that may be needed for a job title.
- **What is your annual salary? (You'll indicate the currency in a later question. If you are part-time or hourly, please enter an annualized equivalent – what you would earn if you worked the job 40 hours a week, 52 weeks a year.)** (`character`)
  - The unitless annual salary of the respondee. For respondees who work a part-time job or are paid hourly, the value is an estimate of an annualized equivalent of their pay based on an assumption of a 40 hour work week, 52 weeks a year. The currency unit is specified in **Please indicate the currency**
- **How much additional monetary compensation do you get, if any (for example, bonuses or overtime in an average year)? Please only include monetary compensation here, not the value of benefits.** (`numeric`)

  - The unitless monetary compensation that the respondee receives outside the value of benefits (for example, bonuses or overtime in an average year).

- **Please indicate the currency** (`character`)
  - 3-letter currency code for the type of currency that the respondee is paid with. For example, United States dollars are abbreviated as USD. Possible values are:
    * USD
    * EUR
    * JPY
    * GBP
    * CHF
    * CAD
    * AUD/NZD
    * ZAR
    * HKD
    * SEK
    * <Other>

- **If "Other," please indicate the currency here:** (`character`)
  - 3-letter currency code for the type of currency that the respondee is paid with, if the currency is not a listed choice for **Please indicate the currency**.

- **If your income needs additional context, please provide it here:** (`character`)
  - Additional context that may be important to the understanding of the respondee's income.

- **What country do you work in?** (`character`)
  - Country that the respondee works in. Possible values are:

- **If you're in the U.S., what state do you work in?** (`character`)
  - Name of the state that the respondee works in, if they work in the United States. Values include full state names

- **What city do you work in?** (`character`)
  - Name of the city that the respondee works in.

- **How many years of professional work experience do you have overall?** (`character`)
  - The amount of professional work experience the respondee has overall (in years). Possible values are:
    * 1 year or less
    * 2-4 years
    * 5-7 years
    * 8-10 years
    * 11-20 years
    * 21-30 years
    * 31-40 years
    * 41 years or more

- **How many years of professional work experience do you have in your field?** (`character`)
  - The amount of professional work experience the respondee has in their current field (in years). Possible values are:
    * 1 year or less
    * 2-4 years
    * 5-7 years
    * 8-10 years

        * 11-20 years
        * 21-30 years
        * 31-40 years
        * 41 years or more

- **What is your highest level of education completed?** (`character`)
    - The highest level of education that the respondee has attained. Possible values are:
        * High School
        * Some college
        * College degree
        * Master's degree
        * PhD
        * Professional degree (Md, JD, etc.)
- **What is your gender?** (`character`)
    - Gender of the respondee. Possible values are:
        * Man
        * Woman
        * Non-binary
        * Other or prefer not to answer
        * Prefer not to answer
- **What is your race? (Choose all that apply.)** (`character`)
    - Race of the respondee. Choices may be a combination of these possible values:
        * Asian or Asian American
        * Black or African American
        * Hispanic, Latino, or Spanish origin
        * Middle Eastern or Northern African
        * Native American or Alaska Native
        * White
        * Another option not listed here or prefer not to answer
- **X, X.1, ... X.5** (`Logical`)
    - Empty columns left over from the porting of the original google sheet to a `.csv` format

Data Cleaning

Before we begin answering any research questions, we need to perform some extensive cleaning on the dataset.

Most of the variable names in this dataset list out the entire question that was asked, making them long and unnecessary. We are going to use the `clean_names()` function from the `janitor` package to convert the names to `snake_case` format. Then, we are going to rename all of the columns to make them more compact and concise. The last few columns on the right side of the dataset are blank and unnecessary so we will also delete them.

```
salary_clean_names <- salary %>%
  select(-(X:X.5)) %>%
  janitor::clean_names() %>%
  rename(age = how_old_are_you,
         industry = what_industry_do_you_work_in,
         job_title_context = if_your_job_title_needs_additional_context_please_clarify_here,
         extrapolated_annual_salary_unitless = what_is_your_annual_salary_you_ll_indicate_the_currency_
```

```
        compensation = how_much_additional_monetary_compensation_do_you_get_if_any_for_example_bonuses_
        currency = please_indicate_the_currency,
        other_currency = if_other_please_indicate_the_currency_here,
        income_context = if_your_income_needs_additional_context_please_provide_it_here,
        country = what_country_do_you_work_in,
        us_state = if_you_re_in_the_u_s_what_state_do_you_work_in,
        city = what_city_do_you_work_in,
        overall_exp = how_many_years_of_professional_work_experience_do_you_have_overall,
        field_exp = how_many_years_of_professional_work_experience_do_you_have_in_your_field,
        education = what_is_your_highest_level_of_education_completed,
        gender = what_is_your_gender,
        race = what_is_your_race_choose_all_that_apply)

head(salary_clean_names)
```

```
##            timestamp   age                       industry
## 1 4/27/2021 11:02:10 25-34  Education (Higher Education)
## 2 4/27/2021 11:02:22 25-34             Computing or Tech
## 3 4/27/2021 11:02:38 25-34 Accounting, Banking & Finance
## 4 4/27/2021 11:02:41 25-34                    Nonprofits
## 5 4/27/2021 11:02:42 25-34 Accounting, Banking & Finance
## 6 4/27/2021 11:02:46 25-34  Education (Higher Education)
##                                 job_title job_title_context
## 1       Research and Instruction Librarian
## 2 Change & Internal Communications Manager
## 3                     Marketing Specialist
## 4                          Program Manager
## 5                        Accounting Manager
## 6           Scholarly Publishing Librarian
##   extrapolated_annual_salary_unitless compensation currency other_currency
## 1                              55,000            0      USD
## 2                              54,600         4000      GBP
## 3                              34,000           NA      USD
## 4                              62,000         3000      USD
## 5                              60,000         7000      USD
## 6                              62,000           NA      USD
##   income_context        country       us_state        city  overall_exp
## 1                 United States  Massachusetts      Boston    5-7 years
## 2                United Kingdom                  Cambridge 8 - 10 years
## 3                            US      Tennessee Chattanooga  2 - 4 years
## 4                           USA      Wisconsin   Milwaukee 8 - 10 years
## 5                            US South Carolina  Greenville 8 - 10 years
## 6                           USA  New Hampshire     Hanover 8 - 10 years
##     field_exp       education     gender  race
## 1   5-7 years Master's degree      Woman White
## 2   5-7 years  College degree Non-binary White
## 3 2 - 4 years  College degree      Woman White
## 4   5-7 years  College degree      Woman White
## 5   5-7 years  College degree      Woman White
## 6 2 - 4 years Master's degree        Man White
```

We will also convert all of the column names and individual values to lowercase letters.

```
# Everything lowercase
for (n in colnames(salary_clean_names)){
  salary_clean_names[[n]] <- str_to_lower(salary_clean_names[[n]])}
```

We are going to convert some of the categorical variables into factors to make them easier to work with when we answer our research questions.

```
# Change variable data type
salary_clean_dtype <- salary_clean_names %>%
  mutate(timestamp = mdy_hms(timestamp),
         age = factor(age),
         overall_exp = factor(overall_exp),
         field_exp = factor(field_exp),
         currency = factor(currency),
         education = factor(education),
         gender = factor(gender))
head(salary_clean_dtype)
```

```
##             timestamp   age                       industry
## 1 2021-04-27 11:02:10 25-34  education (higher education)
## 2 2021-04-27 11:02:22 25-34             computing or tech
## 3 2021-04-27 11:02:38 25-34 accounting, banking & finance
## 4 2021-04-27 11:02:41 25-34                    nonprofits
## 5 2021-04-27 11:02:42 25-34 accounting, banking & finance
## 6 2021-04-27 11:02:46 25-34  education (higher education)
##                                  job_title job_title_context
## 1       research and instruction librarian
## 2 change & internal communications manager
## 3                     marketing specialist
## 4                          program manager
## 5                        accounting manager
## 6           scholarly publishing librarian
##   extrapolated_annual_salary_unitless compensation currency other_currency
## 1                              55,000            0      usd
## 2                              54,600         4000      gbp
## 3                              34,000         <NA>      usd
## 4                              62,000         3000      usd
## 5                              60,000         7000      usd
## 6                              62,000         <NA>      usd
##   income_context        country       us_state        city  overall_exp
## 1                 united states  massachusetts      boston    5-7 years
## 2                united kingdom                  cambridge 8 - 10 years
## 3                            us      tennessee chattanooga  2 - 4 years
## 4                           usa      wisconsin   milwaukee 8 - 10 years
## 5                            us south carolina  greenville 8 - 10 years
## 6                           usa  new hampshire     hanover 8 - 10 years
##     field_exp       education     gender  race
## 1   5-7 years master's degree      woman white
## 2   5-7 years  college degree non-binary white
## 3 2 - 4 years  college degree      woman white
## 4   5-7 years  college degree      woman white
## 5   5-7 years  college degree      woman white
## 6 2 - 4 years master's degree        man white
```

The entries for the question asking about the respondents' annual income contain commas to separate the thousands place from the hundreds place, making them character strings instead of numeric integers. We are going to eliminate those commas and convert the values to doubles, so they can be viewed as and dealt with as numbers.

```
# Get rid of thousands separators (",")
salary_clean_dtype$extrapolated_annual_salary_unitless <- salary_clean_dtype$extrapolated_annual_salary_
  str_replace_all(",", "") %>%
  as.double()
```

Most of the respondents are from the United States, but there are some people who are from other countries such as Canada and the United Kingdom. We are going to filter out those individuals, leaving only the American respondents. The question asking about the individuals' country of residence appears to be a short answer questions because the answers include different variations of the United States. To make everything consistent, we are going to recode all of these variations to "USA". After doing so, all of the data that is left pertains to the USA, meaning that we no longer need the country and currency information. For that matter, for the rest of the analysis, we only require salary information, state information, education information and gender information - meaning that we can drop everything else and just keep `extrapolated_annual_salary_unitless`, `us_state`, `education`, and `gender`.

```
#Recode all variations to "USA"
salary_clean_country <- salary_clean_dtype %>%
  mutate(country = recode(country,
                          "united states" = "USA",
                          "us" = "USA",
                          "usa" = "USA",
                          "u.s." = "USA",
                          "united states of america" = "USA"
                          )) %>%
  # Filter only US responses
  filter(country == "USA") %>%
  # Country and Currency are now irrelevant - we only have US entries.
  # Drop city as well because we will not use it in our analysis
  select(extrapolated_annual_salary_unitless,
         us_state,
         education,
         gender)

head(salary_clean_country)
```

```
##   extrapolated_annual_salary_unitless       us_state       education gender
## 1                               55000  massachusetts master's degree  woman
## 2                               34000      tennessee  college degree  woman
## 3                               62000      wisconsin  college degree  woman
## 4                               60000 south carolina  college degree  woman
## 5                               62000  new hampshire master's degree    man
## 6                               33000 south carolina  college degree  woman
```

Now that we are done with the data cleaning, we can begin to make visualizations to answer our research questions.

Research Questions & Visualizations

The research questions we intend to answer with the data set are **Is There a Significant Difference in the Highest Level of Education Attainment and/or Salary Between Men and Women?** and **How Does the Gender Wage Gap Vary Throughout Different States Within the U.S?** To answer the first question, we will create a side-by-side bar graph to visualize the difference in proportion of the highest education level between men and women, and a superpositioned histogram to visualize the differences in salary. To answer the second question, we will create a choropleth map to visualize the difference in average annual income between men and women in each of the contiguous U.S. states.

Is There a Significant Difference in the Highest Level of Education Attainment and/or Salary Between Men and Women?

We are interested in the differences between the two binary genders. Our cleaned dataset has two variables that we can very easily compare - `education`, corresponding to the highest level of education attained by the respondee, and `extrapolated_annual_salary_unitless`, corresponding to the extrapolated salary of the respondee. We will create graphs to help us visualize the potential differences in education and salary.

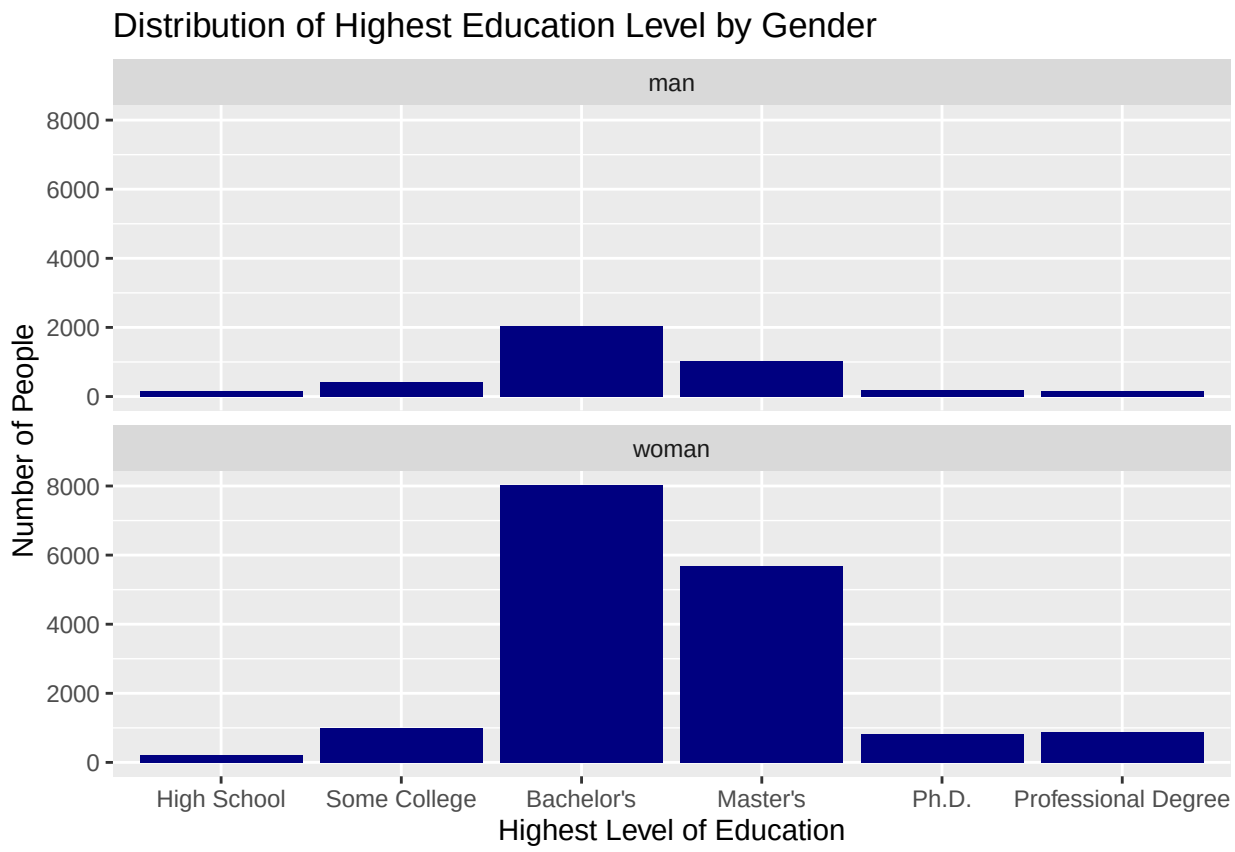
First, we are going to recode the values of the `education` variable to make them more concise and compact. Then, we are going to modify the factor levels for this variable so that the levels for `education` are in order from lowest level of `education` (high school diploma) to highest level of `education` (PhD or professional degree). To facet and compare the distribution by binary gender, we are going to use the `filter()` function to only include individuals who selected "man" or "woman" as their gender. There were not enough non-binary individuals to make conclusive analyses.

```
# Recode education
education_gender_cleaned <- salary_clean_country %>%
  mutate(education = recode(education,
        "high school" = "High School",
        "some college" = "Some College",
        "college degree" = "Bachelor's",
        "master's degree" = "Master's",
        "phd" = "Ph.D.",
        "professional degree (md, jd, etc.)" = "Professional Degree"
        )) %>%
  mutate(education = fct_relevel(education,
                                 "High School",
                                 "Some College",
                                 "Bachelor's",
                                 "Master's",
                                 "Ph.D.",
                                 "Professional Degree")) %>%
  filter(gender == "man" | gender == "woman",
         education != "")
head(education_gender_cleaned)
```

```
##   extrapolated_annual_salary_unitless       us_state  education gender
## 1                               55000  massachusetts   Master's  woman
## 2                               34000      tennessee Bachelor's  woman
## 3                               62000      wisconsin Bachelor's  woman
## 4                               60000 south carolina Bachelor's  woman
## 5                               62000  new hampshire   Master's    man
```

```
## 6                              33000 south carolina Bachelor's  woman
```

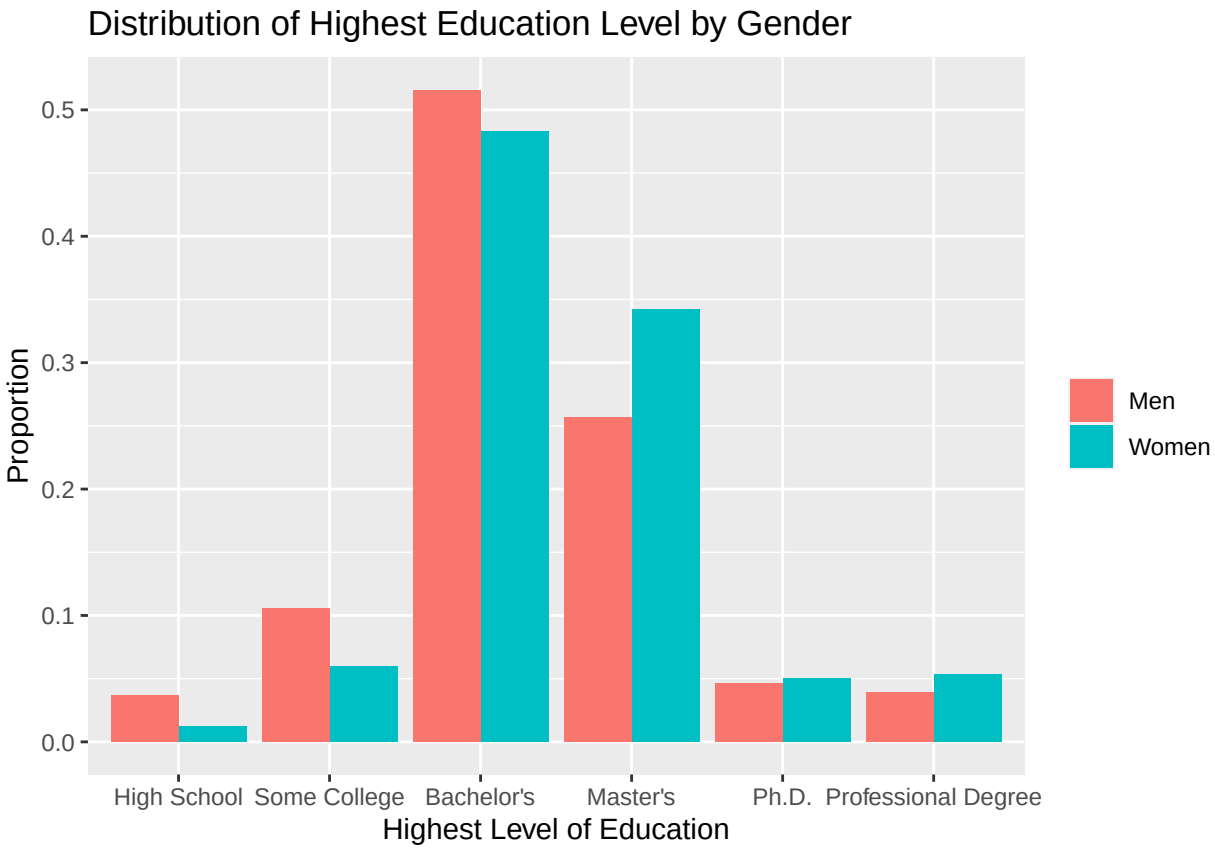
```
# Generate bar graph
education_gender_cleaned %>%
  ggplot(aes(x = education)) +
  geom_bar(fill = "#000080") +
  facet_wrap(~gender, nrow = 2) +
  labs(title = "Distribution of Highest Education Level by Gender",
       x = "Highest Level of Education",
       y = "Number of People")
```



From this plot, it is clear that women were the majority in those who answered the survey. Because of the vast difference in number of responses, it isn't really easy to see how the two genders compare. We introduce a normalized version of these charts in a side-by-side form to solve this. This new plot utilizes normalization via `group_by()` and `summarize()`, and gives a much better picture of the difference between the education levels.

```
# Relative
education_gender_cleaned %>%
  group_by(education) %>%
  summarize(mcnt = (sum(!is.na(education) & gender == "man")),
            wcnt = (sum(!is.na(education) & gender == "woman"))) %>%
  mutate(men = mcnt/sum(mcnt),
         women = wcnt/sum(wcnt)) %>%
  pivot_longer(cols = c("women", "men"),
               names_to = "gender",
```

```
              values_to = "proportion") %>%
ggplot(mapping = aes(x = education,
               y = proportion,
               fill = gender)) +
geom_bar(position = "dodge",
         stat = "identity") +
theme(legend.title = element_blank()) +
scale_fill_discrete(labels = c("Men", "Women")) +
labs(title = "Distribution of Highest Education Level by Gender",
     x = "Highest Level of Education",
     y = "Proportion")
```

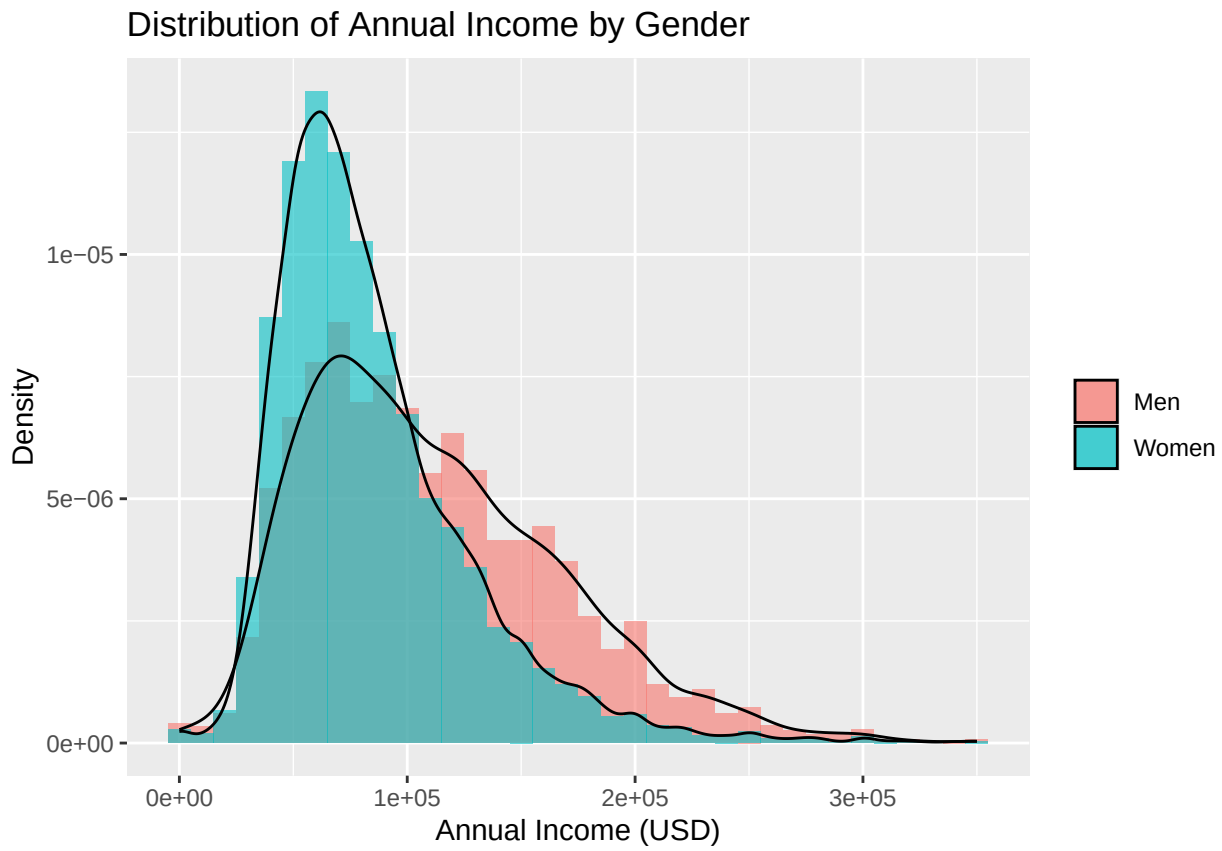


There does not seem to be any significant differences in the distributions of highest education level between men and women. For both men and women, the highest level of education attained by the most individuals was a Bachelor's degree followed by a Master's degree. The proportion of individuals with some college education was greater for the women than for the men. For both groups, the number of individuals with a Ph.D. and the number of individuals with a Professional Degree were roughly equal. We now shift our analysis to comparing the average income between men and women.

Now, we are going to create a histogram of annual income with both genders. There is one individual whose reported annual income is over $5 million when the median annual income is only $79,000 which heavily skews our data to the right. To fix this issue and eliminate any other values that may heavily skew the data, we are going to filter out any reported annual incomes that are over $350,000. We create a `ggplot` graph on top of it, using the density of the salaries instead of actual counts to avoid the same problem encountered with the education analysis.

```r
education_gender_cleaned %>%
  filter(extrapolated_annual_salary_unitless <= 350000) %>%
  ggplot(aes(x = extrapolated_annual_salary_unitless)) +
  geom_histogram(aes(y = ..density..,
                     fill = gender),
                 binwidth = 10000,
                 position = "identity",
                 alpha = 0.6) +
  geom_density(aes(group = gender),
               alpha = 0.3) +
  labs(title = "Distribution of Annual Income by Gender",
       x = "Annual Income (USD)",
       y = "Density") +
  theme(legend.title = element_blank()) +
  scale_fill_discrete(labels = c("Men", "Women"))
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Both distributions are skewed to the right but there does seem to be a difference between the distributions of annual income between men and women. Specifically, the mean of the men's distribution appears to be slightly greater than the mean of the women's distribution, and the men's distribution is flatter, with more

higher salary values. The means for both distributions appear to be in the $60,000 to $70,000 range which is reasonable since the median annual salary for individuals with only a Bachelor's degree in the United States is approximately $70,000.

How Does the Gender Wage Gap Vary Throughout Different States Within the U.S?

As our previously constructed figure on men's and women's salaries showed, there is a difference in the distributions. Specifically, it looks like there are more men with higher salaries. We will investigate this difference with a geographical approach - we will take the average salaries for each binary gender in each U.S. state and compare them via a map visualization. To do so, we must first prepare our data for a `ggplot` map.

Because the map information given by `map_data("state")` only maps out the continental U.S., it does not include information to map out Alaska and Hawaii. Therefore, we can remove these states from our analysis. Additionally, we must deal with the entries with multiple states. These represent people who have worked in multiple states, and we do not want to double count these people, so we remove these entries altogether.

```
numCommas <- salary_clean_country$us_state %>%
  str_replace_all("(hawaii|alaska)(,\\s)*", "") %>%
  str_detect(",") %>%
  sum()

numEmpty <- salary_clean_country$us_state %>%
  str_replace_all("(hawaii|alaska)(,\\s)*", "") %>%
  str_replace_all(",.*$", "") %>%
  str_detect("^$") %>%
  sum()

salary_clean_map <- salary_clean_country

salary_clean_map$us_state <- salary_clean_map$us_state %>%
  str_replace_all("(hawaii|alaska)(,\\s)*", "") %>%
  str_replace_all(",.*$", "")

salary_clean_map <- salary_clean_map %>%
  filter(us_state != "")

head(salary_clean_map)
```

```
##   extrapolated_annual_salary_unitless       us_state       education gender
## 1                               55000  massachusetts master's degree  woman
## 2                               34000      tennessee  college degree  woman
## 3                               62000      wisconsin  college degree  woman
## 4                               60000 south carolina  college degree  woman
## 5                               62000  new hampshire master's degree    man
## 6                               33000 south carolina  college degree  woman
```

While this method may cause us to lose some information, there are only 95 entries with multiple states after removing Hawaii and Alaska, while we have 21617 non Hawaii/Alaska entries in total, meaning that removing these entries probably will not impact our analysis too much. Furthermore, there are only 153

empty values - indicating the surveyed did not fill out a state - in our cleaned dataset to begin with, meaning we can also drop these without impacting our analysis too much.

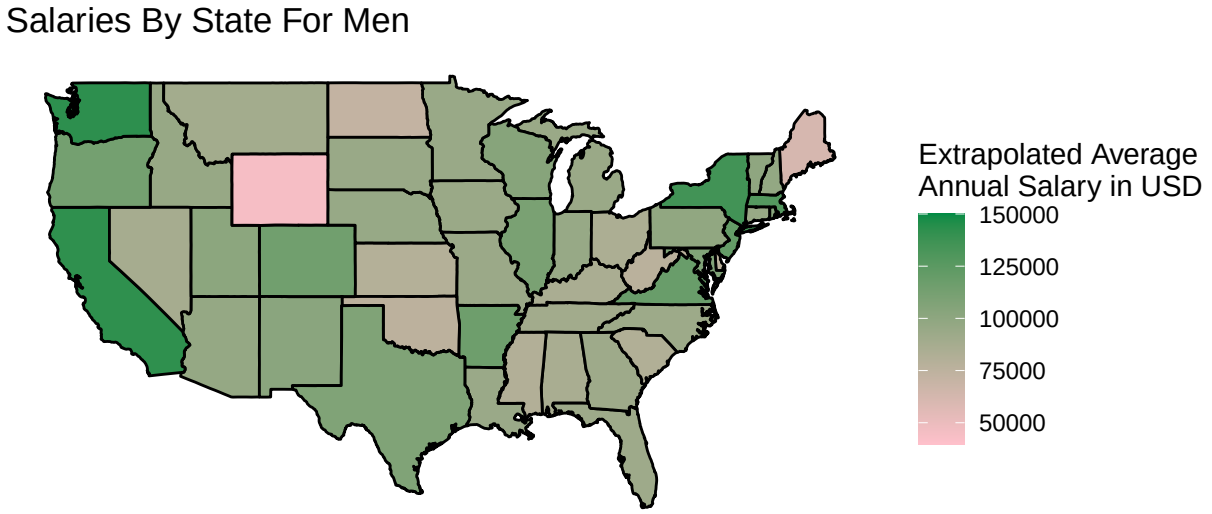
We can now proceed to the actual creation of the map. We first define `clean_map` as a theme for our map to use. By default, `ggplot` canvases are lined with a background as well as a grid, both of which we do not need. This theme gets rid of all these extraneous elements and allows us to create a clean and concise map. We then define several functions: `getSalary()`, `makeGraph()` and `makeGraphGender()`. They do these things respectively:

- `getSalary(g)`
  - Calculates the state-average salary for the given gender `g`, and returns the result as a dataframe with 2 columns - 1 for state and 1 for the state-average salary for the given gender in that state.
- `makeGraph(d, t, l, lo, hi)`
  - Creates a `ggplot` map given the dataset `d`. It will title the graph `t` and have the legend title as `l`, with lower and upper bounds of the color scale at `lo` and `hi` respectively.
- `makeGraphGender(g)`
  - Creates the map that we want. This function incorporates `getSalary()` and `makeGraph()` to 1) create a dataframe containing both the map coordinate information as well as the salary information for the given gender `g` and 2) creates the map corresponding to that gender.

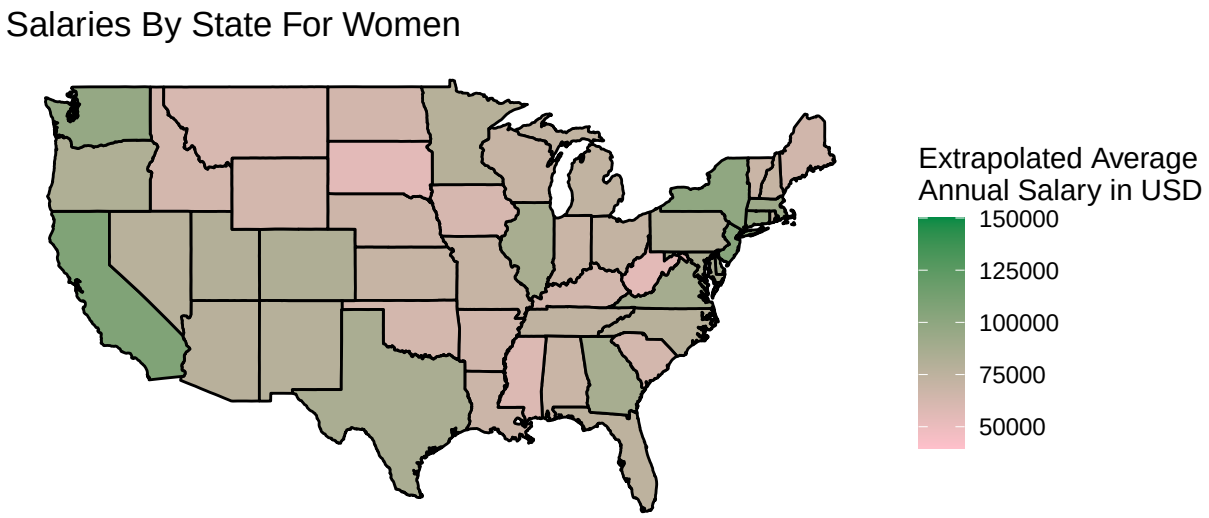
Using these functions, we plot the average salaries of the two binary genders for each state.

```
# Clean map theme
clean_map <- theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.background = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank()
  )


# Function to compute state-average salaries
getSalary <- function(g){
  gendered_state_salaries <- salary_clean_map %>%
    filter(gender == g) %>%
    group_by(us_state) %>%
    summarize(salary = mean(extrapolated_annual_salary_unitless, na.rm = TRUE))

    return (gendered_state_salaries)
}


# Function for creating maps
makeGraph <- function(d, t, l, lo, hi){
ggplot(d, aes(x = long, y = lat, fill = salary)) +
    geom_polygon(aes(group = group),
                 color = "black") +
    coord_fixed(1.3) +
    scale_fill_continuous(low = "pink",
```

```
                          high = "springgreen4",
                          limits = c(lo, hi),
                          name = l) +
    clean_map +
    ggtitle(t %>% str_to_title())
}

# Function for creating maps for inputted gender g: c("man", "woman")
makeGraphGender <- function(g){
  gendered_state_salaries <- getSalary(g)

  state_map_df <- map_data("state") %>%
    left_join(gendered_state_salaries, by = c("region" = "us_state"))

  makeGraph(d = state_map_df,
            t = paste0("Salaries by State for ", str_replace(g, "a", "e")),
            l = "Extrapolated Average\nAnnual Salary in USD",
            lo = 40000,
            hi = 150000)
}


# Generate maps for men and women
plots <- map(.x = c("man", "woman"),
             .f = makeGraphGender)

plots[[1]]
```


Salaries By State For Men
Extrapolated Average
Annual Salary in USD
150000
125000
100000
75000
50000

```
plots[[2]]
```



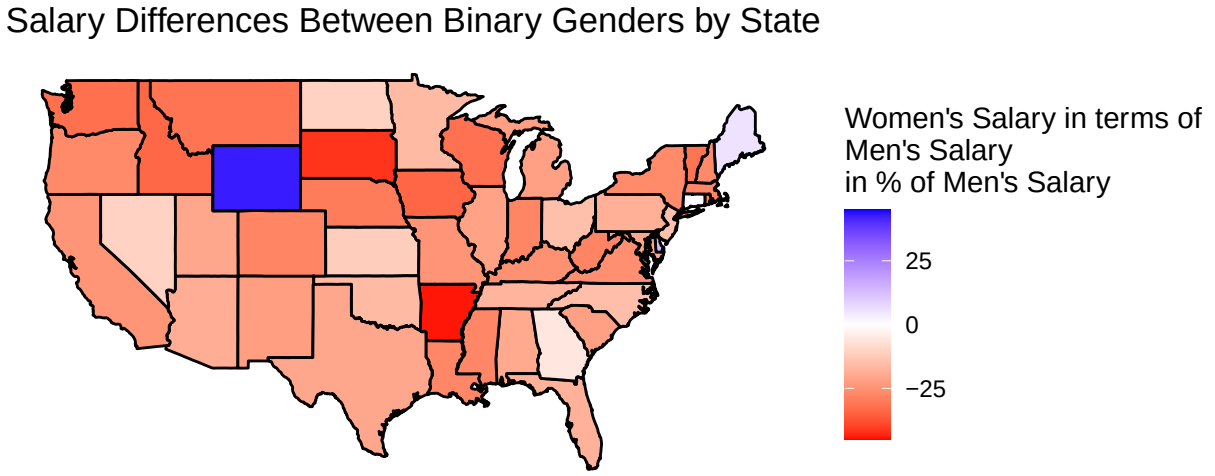
Looking at these graphs, we can already see that the women's graph seems pinker than the men's in almost every state, indicating a lower average salary. But this isn't the most easily comparable set of graphs - we need something relative. To that end, we can create a map of the relative wage difference. We calculate the relative wage difference with the percent difference formula:

$$\text{Women's Wages} = \frac{\text{Women - Men}}{\text{Men}} * 100$$

We then plot a map using this information. This new map, in conjunction with the absolute measures from the first set of maps, can give us a much better picture of the salary differences in the United States.

```
# Compute relative differences in state-level salaries for binary gender
salary_diff_relative <- getSalary("man") %>%
  rename("men" = "salary") %>%
  left_join(getSalary("woman"),
            by = c("us_state")) %>%
  rename("women" = "salary") %>%
  group_by(us_state) %>%
  summarize(diff = (100 * (women - men)/(men)))

# Create the map
map_data("state") %>%
  left_join(salary_diff_relative,
            by = c("region" = "us_state")) %>%
  ggplot(aes(x = long, y = lat, fill = diff)) +
    geom_polygon(aes(group = group),
                 color = "black") +
```

```
    coord_fixed(1.3) +
    scale_fill_gradient2(low = "red",
                         mid = "white",
                         high = "blue",
                         limits = c(-45, 45),
    name = "Women's Salary in terms of\nMen's Salary\nin % of Men's Salary") +
    clean_map +
    ggtitle("Salary Differences Between Binary Genders by State")
```



As this map shows, in most states, women's salaries are less than men's salaries, as indicated by the reddish color. In fact, visually examining the graph, we can only see two or three states that are obviously blue, indicating women are higher paid than men in those states. This shows that the gender wage gap, as of 2021 in the United States, is still a very real and pressing issue.

Conclusion

From our visualization and analyses, we can conclude a few things. Firstly, there was no significant difference between the education levels of the two binary genders. However, there was a noticeable difference in the distribution of salaries between the two binary genders. Furthermore, the gender wage gap is significant, with men making more money than women throughout the continental U.S. This highlights the gender bias in the United States - despite having a similar (in fact, on average higher) level of education, there exists a substantial wage difference in favor of men.