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.

The data set 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 mostly empty columns because those questions wre provided to allow the respondees to elaborate on their answers if they deemed it appropriate (most people did not feel the need to elaborate).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.

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("./Datasets/salary.csv")
```

Data Cleaning

```
# Clean variable names
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)
```

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

```
# 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),
         is_white = str_detect(.$race, "white"))

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

```
#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(-currency, -country, -city)

#Delete subjective question columns
salary_clean_country <- salary_clean_country %>%
  select(-job_title, -job_title_context, -other_currency, -income_context)
```

Research Questions & Visualizations

question

```
salary_clean_country %>%
  group_by(race) %>%
  summarize(cnt = n()) %>%
  arrange(desc(cnt))
```

```
## # A tibble: 45 x 2
##    race                                                    cnt
##    <chr>                                                 <int>
##  1 "white"                                               17935
##  2 "asian or asian american"                              1038
##  3 "black or african american"                             538
##  4 "hispanic, latino, or spanish origin"                   466
##  5 "another option not listed here or prefer not to answer"  414
##  6 "hispanic, latino, or spanish origin, white"            338
##  7 "asian or asian american, white"                        281
##  8 ""                                                      133
##  9 "black or african american, white"                      100
## 10 "middle eastern or northern african, white"              73
## # ... with 35 more rows
```

```
#%>%
 # ggplot(mapping = aes(x = race, y = cnt)) +
    #geom_bar(stat = "identity")
```

Is there a significant difference in the highest level of education attainment between men and women and does it correlate with annual income?

For our first research question, we are going to determine if there is a significant difference in the highest level of education attainment between the men and women that responded to the survey. We are then going to see if education level and gender have any correlation with annual salary. To accomplish this, we are first going to create a bar graph that is ordered from lowest level of education to highest level of education faceted by binary gender.

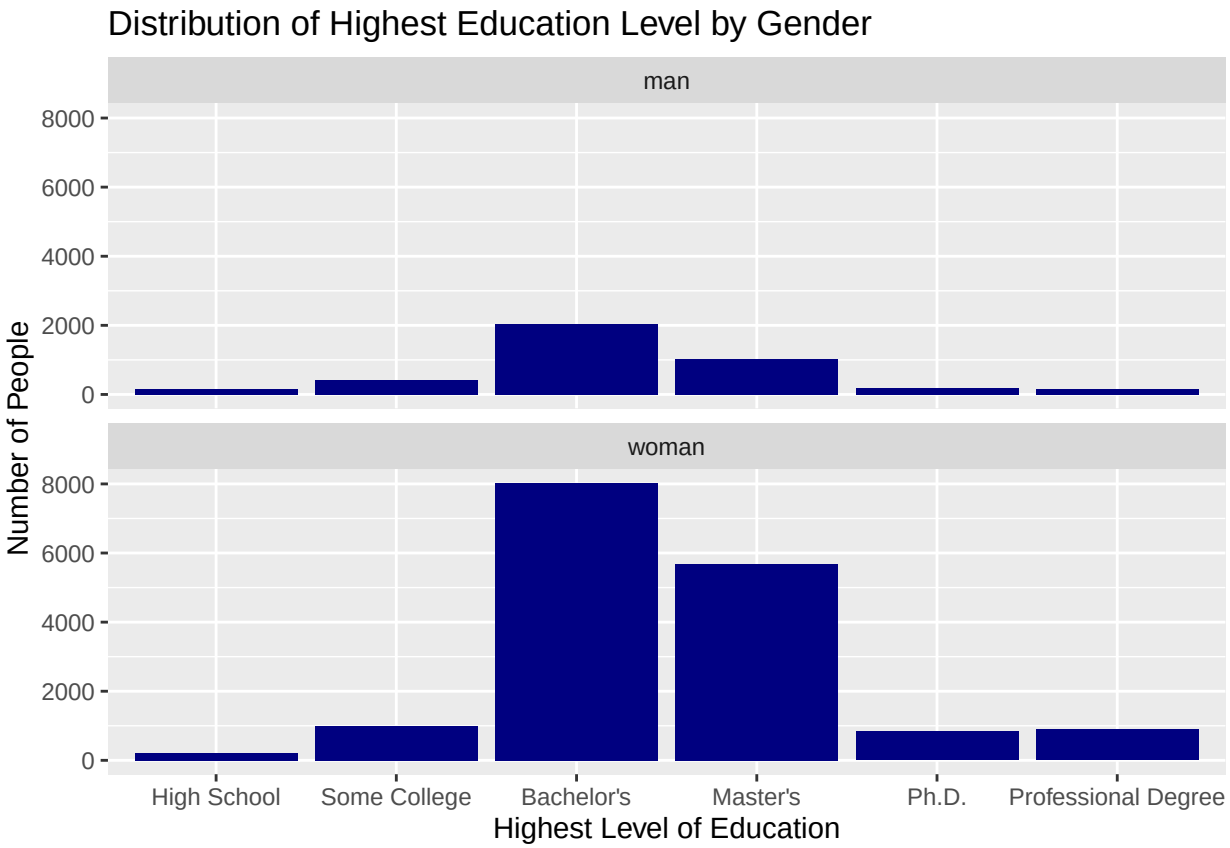
First, we are going to recode the values of the `education` variable to make them more concise and compact. Then, we are going to modify the factor levels for this variable so that the levels are education are in order from lowest level of education (high school diploma) to highest level of education (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 fair assessments with the men and women.

```
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", "Pl
  filter(gender == "man" | gender == "woman",
         education != "")
```

```
## Warning: 1 unknown level in `f`: Other
```

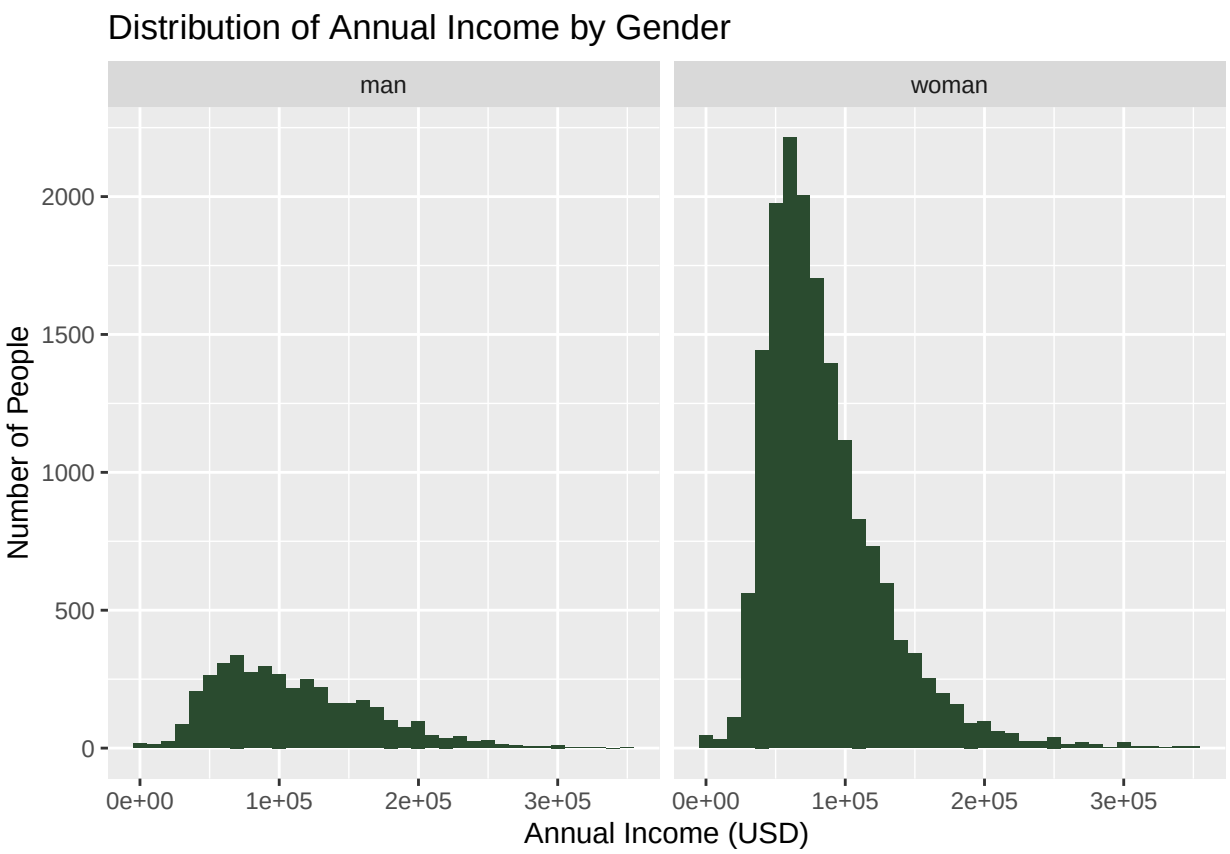
```
education_gender_cleaned %>%
  ggplot(aes(x = education)) +
  geom_bar(fill = "#000080") +
  facet_wrap(~gender, nrow = 2) +
  labs(title = "Distribution of Highest Education Level by Gender", x = "Highest Level of Education", y
```



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.

Now, we are going to create a histogram of annual income, faceted by binary gender. 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.

```
education_gender_cleaned %>%
  filter(extrapolated_annual_salary_unitless <= 350000) %>%
  ggplot() +
  geom_histogram(aes(x = extrapolated_annual_salary_unitless), fill = "#2a4b2f", binwidth = 10000) +
  facet_wrap(~gender) +
  labs(title = "Distribution of Annual Income by Gender", x = "Annual Income (USD)", y = "Number of Peo
```



Both distributions are skewed to the right and there does not seem to be any significant differences between the distributions of annual income between men and women. The mode of the men's distribution appears to be slightly greater than the mode of the women's distribution but the difference is not significant. The modes 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.

Salary Map

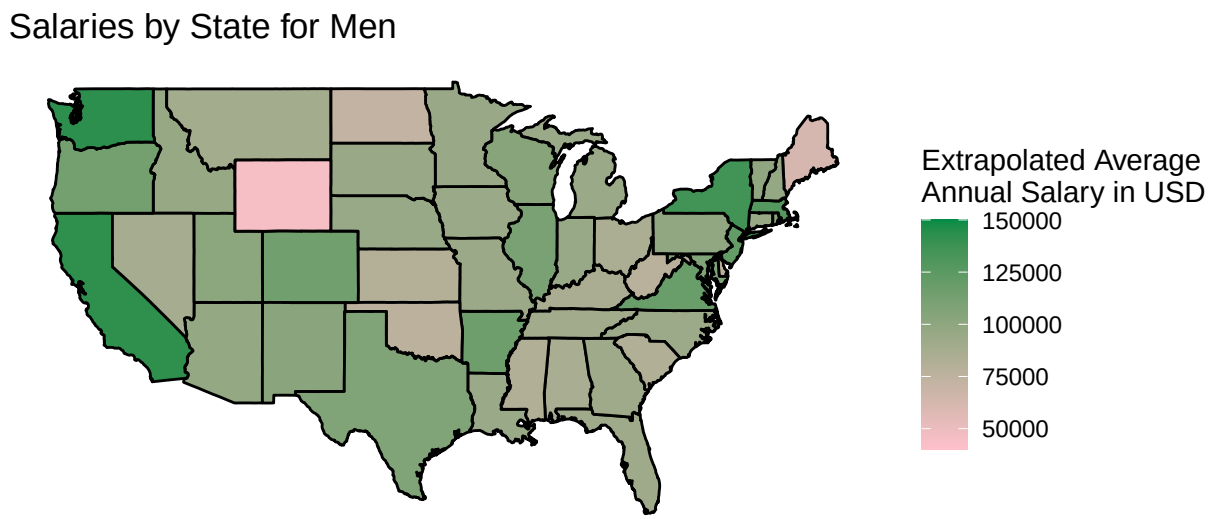
```
numCommas <- salary_clean_country$us_state %>%
  str_replace_all("(hawaii|alaska)(,\\s)*", "") %>%
  str_detect(",") %>%
  sum()

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 != "")
```

The map that we will be creating will not contain Hawaii or Alaska. 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. 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.

```
state_salaries <- salary_clean_map %>%
  group_by(us_state) %>%
  summarize(salary = mean(extrapolated_annual_salary_unitless, na.rm = TRUE))

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

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()
  )

makeGraph <- 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))

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

  ggplot(state_map_df, 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(40000, 150000),
                            name = "Extrapolated Average\nAnnual Salary in USD") +
    clean_map +
    labs(title = paste0("Salaries by State for ", str_replace(g, "a", "e") %>%
                            str_to_title()))
}

plots <- map(.x = c("man", "woman"),
             .f = makeGraph)

plots[[1]]
```




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

Salaries by State for Women