Project #7: The Final Project

Harini Shanmugam

21-12-14

1. Introduction

I selected Project Idea #14, “The Salary Gap”, from the list of ideas prepared by Dr. Cudina. Here, I will analyze the salary gap both with respect to gender and with respect to other demographics (like race/ethnicity and age).

For this project will use and analyze data from 2020 as given by the United States Census Bureau from https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-pinc/pinc-05.html#par_textimage_17. Links to the specific sources will be listed under their corresponding sections and visualizations.

2. Data and Analysis

A.

This is initial data of the population of full-time workers, mean annual earnings, and median annual earnings, of men and women.

```
basestats = matrix(c(59653, 84115, 61417, 45866, 63845, 50982), ncol=2)
colnames(basestats) = c("Male", "Female")
rownames(basestats) = c("Full-time Working Population",
                        "Mean Annual Earnings ($)",
                        "Median Annual Earnings ($)")
basestats
##                              Male Female
## Full-time Working Population 59653  45866
## Mean Annual Earnings ($)     84115  63845
## Median Annual Earnings ($)   61417  50982
```

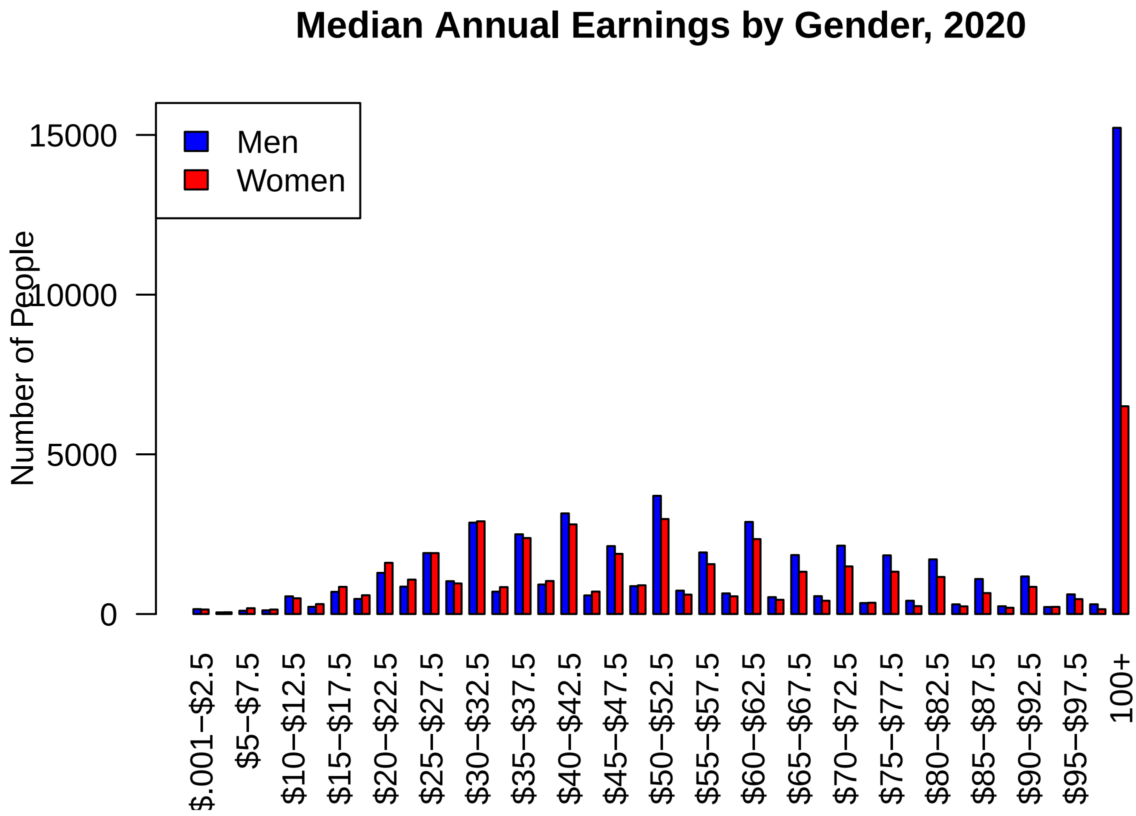
Sources: https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_1_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_1_1.xlsx

When calculating the gender pay gap,
Earnings Ratio $= \frac{\text{Women's median earnings}}{\text{Men's median earnings}} = \frac{\$50{,}982}{\$61{,}417} = 0.8301 = 83\%$
Pay Gap $= \frac{\text{Men's median earnings - Women's median earnings}}{\text{Men's median earnings}} = \frac{\$61{,}417 - \$50{,}982}{\$61{,}417} = 0.1699 = 17\%$

According to these calculations, women were paid 83 cents for every dollar paid to a man and that is a 17% gap in pay. The median, rather than the mean, was used for these calculations to reduce skewness due to jobs lost during the year.

We can also visualize the data by income brackets:

```
men = c(154, 51, 104,116, 554,226, 696, 473, 1289, 858, 1909, 1028, 2862, 701,
        2496, 923, 3150, 583, 2127, 873, 3702, 733, 1928, 646, 2884, 529, 1846,
        560, 2141, 344, 1837, 416, 1711, 303, 1097, 243, 1176, 220, 616, 305,
        15223)
menmu = 84115
menmedian = 61417
women = c(141, 55, 182, 140, 493, 312, 851,588, 1604,1078, 1908, 957, 2903, 843,
          2379, 1034, 2806, 703, 1885, 900, 2975, 608, 1561, 555, 2347, 446,
          1324, 416, 1491, 353, 1326, 247, 1161, 240, 657, 196, 851, 225, 467,
          150, 6505)
womenmu = 63845
womenmedian = 50982
brackets = c("$.001-$2.5", "$2.5-$5", "$5-$7.5",
             "$7.5-$10", "$10-$12.5", "$12.5-$15",
             "$15-$17.5", "$17.5-$20", "$20-$22.5",
             "$22.5-$25", "$25-$27.5", "$27.5-$30",
             "$30-$32.5", "$32.5-$35", "$35-$37.5",
             "$37.5-$40", "$40-$42.5", "$42.5-$45",
             "$45-$47.5", "$47.5-$50", "$50-$52.5",
             "$52.5-$55", "$55-$57.5", "$57.5-$60",
             "$60-$62.5", "$62.5-$65", "$65-$67.5",
             "$67.5-$70", "$70-$72.5", "$72.5-$75",
             "$75-$77.5", "$77.5-$80", "$80-$82.5",
             "$82.5-$85", "$85-$87.5", "$87.5-$90",
             "$90-$92.5", "$92.5-$95", "$95-$97.5",
             "$97.5-$100", "100+")
both = rbind(men, women)
barplot(both, names.arg=brackets,
        beside=T,
        main="Median Annual Earnings by Gender, 2020",
        ylab="Number of People",
        ylim=c(0, 16000),
        las=2,
        col=c("blue", "red"),
        legend.text=c("Men","Women"),
        args.legend=list(x="topleft"))
mtext("Earnings, in thousands ($)", side=1, line=5)
```



Both sets of data mostly follow a normal distribution when excluding the last bracket of $100,000+. That last bracket is also the only bracket where the number of men is more than twice than that of women. This drastically large outlier is an example for why it is better to use the median rather than the mean so that the calculations of the pay gaps are not skewed.

Since we know that there is difference in the mean annual earnings by gender, we can run a two-sample t-test of the means to see if it is a significant different where
$H_0 : \mu_{men} - \mu_{women} = 0$ vs. $H_a : \mu_{men} - \mu_{women} \neq 0$

Significance test: $t = \frac{(\bar{x}_{men}-\bar{x}_{women})-(\mu_{men}-\mu_{women})}{\sqrt{SE^2_{men}+SE^2_{women}}}$
$= \frac{(84115-63845)-(0-0)}{\sqrt{286225+532900}}$
$= \frac{20270}{\sqrt{819125}} = \frac{20270}{\sqrt{819125}}$
$= 22.3964$ with df = 45866 - 1 = 45865

The calculated p-value is nearly 0. Since 0 is lower than 0.05, we reject the null hypothesis. Meaning, there is a statistically significant difference in the pay between men and women and that the difference did not occur by random chance.

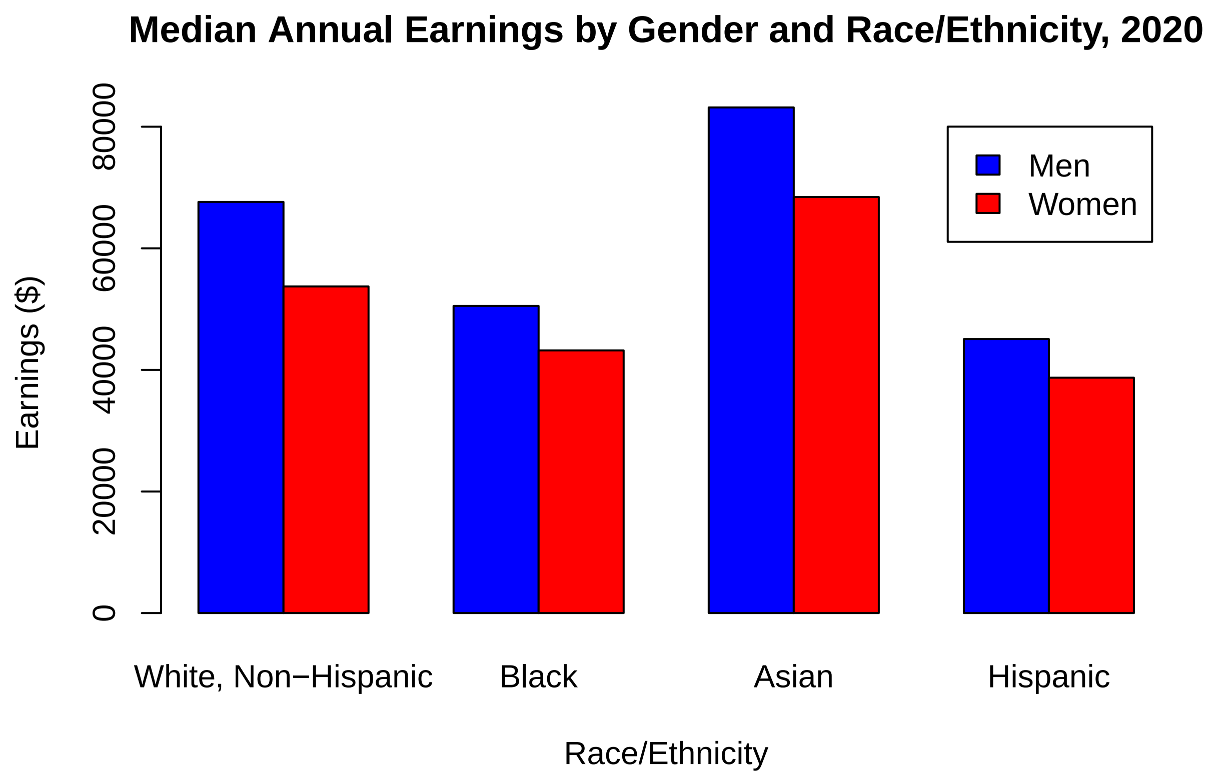
B.

Another demographic to consider is race/ethnicity in addition to gender. This table displays the median annual earnings of men and women of different races in 2020.

```r
race = matrix(c(67629, 50525, 83173, 45074, 53731, 43209, 68442, 38718), ncol=2)
colnames(race) = c("Male", "Female")
rownames(race) = c("White, Non-Hispanic",
                   "Black",
                   "Asian",
                   "Hispanic")
race
##                      Male Female
## White, Non-Hispanic 67629  53731
## Black               50525  43209
## Asian               83173  68442
## Hispanic            45074  38718
```

Sources: https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_1_4.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_1_6.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_1_8.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_1_9.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_1_4.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_1_6.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_1_8.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_1_9.xlsx

```r
menwrace = c(67629, 50525, 83173, 45074)
womenwrace = c(53731, 43209, 68442, 38718)
bothraces = rbind(menwrace, womenwrace)
races = c("White, Non-Hispanic", "Black", "Asian", "Hispanic")
barplot(bothraces, names.arg=races,
        beside=TRUE,
        main="Median Annual Earnings by Gender and Race/Ethnicity, 2020",
        xlab="Race/Ethnicity",
        ylab="Earnings ($)",
        col=c("blue","red"),
        legend.text=c("Men","Women"))
```



Applying the previous method of calculating the pay gap… White, non-Hispanic women were paid 79% of White, non-Hispanic men.
Black women were paid were paid 86% of Black men.
Asian women were paid were paid 82% of Asian men.
Hispanic women were paid were paid 86% of Hispanic men.

The pay gap is consistent among the different races. It is interesting to note that Asian women earn more than men of all of the other races and Asians earn the most in their gender category. This could be indicative of implicit cultural influences that lead Asian people to enter jobs with higher saleries.

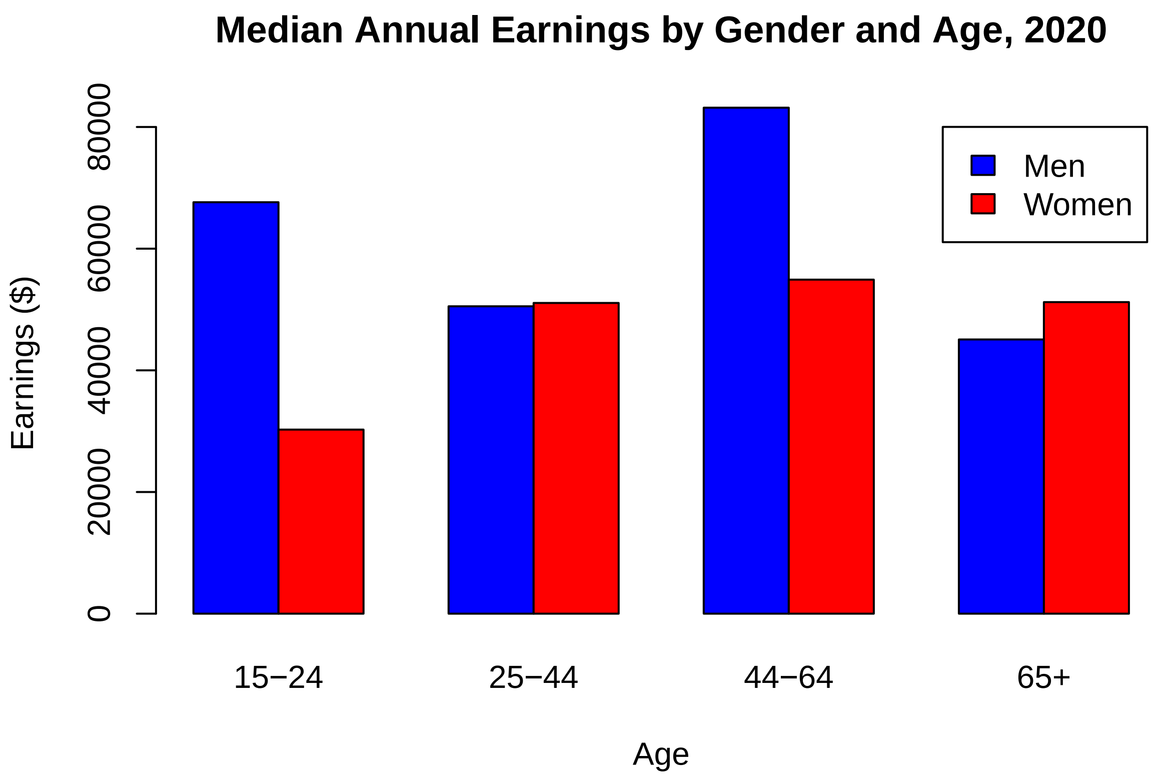
C.

Another key demographic to consider is age. This table displays the median annual earnings of men and women in different age cohorts in 2020.

```
age = matrix(c(31984, 60102, 70551, 70036,
               53731, 43209, 68442, 38718), ncol=2)
colnames(race) = c("Male", "Female")
rownames(race) = c("15-24 years", "25-44 years", "44-64 years", "65+ years")
race
##              Male Female
## 15-24 years 67629  53731
## 25-44 years 50525  43209
## 44-64 years 83173  68442
## 65+ years   45074  38718
```

Sources: https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_3_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_4_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_5_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_2_6_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_3_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_4_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_5_1.xlsx, https://www2.census.gov/programs-surveys/cps/tables/pinc-05/2021/pinc05_3_6_1.xlsx

```r
menwage = c(67629, 50525, 83173, 45074)
womenwage = c(30258, 51074, 54906, 51213)
bothages = rbind(menwage, womenwage)
ages = c("15-24", "25-44", "44-64", "65+")
barplot(bothages, names.arg=ages,
        beside=TRUE,
        main="Median Annual Earnings by Gender and Age, 2020",
        xlab="Age",
        ylab="Earnings ($)",
        col=c("blue","red"),
        legend.text=c("Men","Women"))
```



Men and women between 25 to 44 years of age earn nearly the same amount. It is even more interesting to see that after the age of 65, women make more than men. The largest difference in salary is for young professionals (between 15 to 24 years of age) who are just starting off their careers.

3. Conclusion

In conclusion, there is a clear and significant gap in salaries between men and women. More importantly, the largest disparity occurs in ages 25-44 and 44-64. The pay gap is consistent among different races, but on an unexpected note Asian women earn more than men of other races.

In the future, other demographics that could be considered in relation to gender are level of education, disability, marital status. Sexual orientation and gender identity affecting earnings could also be studied on its own.