Investigating the Gender Wage-Gap

Elliot Spears

2022-05-09

Introduction

As of 2021, women earned 83 cents for every dollar that men earned in the United States. This disparity has created a lot of controversy in recent years, with many people demanding an explanation. Some say that the best explanation for this is discrimination. If they weren't discriminated against by their employers, we would expect them to earn the same amount as men. Here we have the smoking gun, which demonstrates that women don't earn the same amount as men, so they must be systematically discriminated against by employers.

Economists have taken up interest in this question in recent years. A team of economists studied the gender pay-gap as it exists among Uber drivers.[1] Another economist looked at data on gender wage-gap differences among individuals completing tasks on Amazon Mechanical Turk.[2] These are just two papers that make up a sea of research on this issue. The economists see this issue a little differently than those who appeal to the discrimination explanation. The economists from the aforementioned papers, for instance, cite interruptions in labor force participation and hours worked as major explanations for the disparity in incomes between the sexes. This is typical of much of the economic literature on this issue, at as of right now.

Perhaps the most heavily weighed factor that virtually all economists agree to be the main contributor to the gender pay-gap is the fact that women birth children and men do not. We can attribute to this fact much of the blame for why a gender pay gap continues to persist, according to them.

That brings us to the point of this study: to gain insight into some of the causes of the gender pay-gap. Of course, this brief analysis will be fairly limited due to time constraints, but we can survey the data in order to get some clarity on this issue that has troubled so many people for decades. I would like to draw attention to the fact that the gender wage-gap has been shrinking over the years, but it would be helpful to know what could be done, if anything, about the wage-gap on both the individual and the national level. We now turn to the methods used to conduct this analysis.

Methods

I decided to extract data from IPUMS USA, which is a website and database that contains an enormous amount of information on the American public. I gathered information and variables from the website that will help us in the task at hand. I decided to use data from the year 2018, even though the website contains information up until 2021. The reason I decided this was that I wanted to completely eliminate any disruptions that COVID-19 may have caused in the data. Although the stay-at-home orders went into effect in 2020 in the United States, the coronavirus did emerge in 2019, so for good measure I pushed back to 2018.

[1] Cody Cook, Rebecca Diamond, Jonathan V Hall, John A List, Paul Oyer, The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers, The Review of Economic Studies, Volume 88, Issue 5, October 2021, Pages 2210–2238

[2] Adams-Prassl, Abi, The Gender Wage Gap on an Online Labour Market: The Cost of Interruptions (January 2020). CEPR Discussion Paper No. DP14294

Table 1: Avg Income by State

State	Avg Yearly Income	SD
11	69590.83	93823.15
9	56217.45	91630.33
34	54618.08	82642.58
24	53964.57	73634.03
25	53021.82	78784.68

The variables are named and coded in ways that are not immediately obvious. The variable that identifies the state (US state) that the observation was made in is coded numerically, with Alabama being equal to 1, and so on down the line to Wyoming, which is coded as 56 (the numerical coding omitted some numbers, which is why Wyoming isn't coded as 50). The variable "OCC" is a variable that designates the occupation that the individual works at the time of the census. There hundreds of different occupations found in the data, all coded by a number, which can be matched to the occupation title that IPUMS provides on their website.

One major drawback of the dataset, and of the IPUMS data in general, is that none of the datasets that come after 1990 contain information on whether or not an individual has ever had children in their lifetime. This is one of the most important variables at play, according to other studies. This limits the scope of my analysis, but I am still able to use other variables that have a major impact upon the observed outcomes. Ultimately, I feel that the findings of this study are substantial enough, even without the variable for children, that I can proceed with this analysis and produce important insights that merit a project of this kind.

Also, there were several observations where total income for the year was not recorded. Instead of putting N/A where there was no income reported, IPUMS put 9999999 into the total income column for that particular observation. I simply removed those observations from the data set.

I begin by looking at averages across the population as differentiated by sex. I make several comparisons, some with no controls, and others with several controls in order to finally get into a position where the large discrepancies in the wage-gap nearly disappear. I then run several regressions using tree models that predict total income using a train-test split in order to obtain a model with the lowest root-mean-squared-error possible. I provide partial dependence plots below that help us gauge how relevant each variable is to total predicted income. I then provide several tables and plots that compare men and women of the same lucrative profession in order compare apples to apples and see how much of the wage-gap still persists.

Results

We start by looking at the average income of the top 5 states in **Table 1** by average earnings. Again, the states are coded numerically. The states from top to bottom are ordered as follows: Washington DC, Connecticut, New Jersey, Maryland, and Massachusetts. This will provide us with a good baseline for comparison later on.

None of these states are particularly surprising in terms of making it into the top 5 for average income. DC is quite far ahead of the other states, which makes sense given the sheer amount of politicians, judges and lawyers that reside there. Next we turn to the national average income as determined by sex. **Table 2** contains no controls for any of the other variables. We are simply comparing incomes purely on the basis of sex.

Here there is a clear disparity in income. In this sample, women make 60 cents for every dollar that men earn. It is estimated that the true disparity is closer to 83 cents for every dollar, but this is a smaller data set than what is used to calculate that statistic. However, the fact remains the same: women on average earn less than men by a large margin.

Table 2: National Average Income by Sex

Sex	Avg Yearly Income	SD
1	53238.76	77088.32
2	32364.76	47550.10

Table 3: National Average Income by Sex

Sex	Avg Yearly Income	SD
1	139318.2	111783.1
2	138787.8	127538.0

So, do we have a clear-cut case of discrimination here? This is what we want to find out in the next table. I ran several more tests and produced several more tables that are not shown here in this paper. For each additional table I produced, I included an additional control variable, which invariably led to a reduction in the wage-gap disparity. I began by introducing an education variable in order to compare highly educated men to highly educated women. In our sample, the highest level of education recorded is 5+ years of college. This will include men and women that stayed in their undergraduate studies for 5+ years, as well as master's degree holders, PhD holders, MDs, JDs, and so on.

I then decided to control for marital status as this seemed to be a good proxy for the missing child variable. I then chose to control for both men and women that have never been married. Although there are many men and women that beget and raise children outside of marriage, this is a minority of the population. Hence, this control serves as a relatively reliable substitute for the control I would like to have been able to introduce, which would have been men and women that also never had children.

The next control I introduced was prime working-age individuals. I decided to limit the scope of the analysis to individuals between the ages of 25 and 55. I decided to provide a buffer in terms of when individuals enter the workforce, which I arbitrarily chose to be 25, and 55 is when many people begin to retire. If there is a large disparity in terms of age of retirement, capping the age range at 55 should help take care of that difference.

Lastly, I wanted to compare men and women that work the same number of hours. Men, on average, work more hours than women do. Hence, this is a very relevant control. In the United States, 35 hours per week is the official cutoff for full-time employment, according to the IRS, but the usual understanding of full-time work is 40 hours per week, which is what I decided to use here, which means we are only comparing men and women that work 40 hours per week or more on average.

Lest anyone think that this will drastically reduce the sample size to such a point as to skew the results, the hours worked are only coded with whole numbers. There are no decimal points, and a vast portion of the data is made up of men and women that work exactly 40 hours per week or more.

Table 3 contains all of the aforementioned controls.

Clearly, much of the difference has disappeared with the controls we've implemented. Now women earn 99.6 cents for every dollar that men earn, a 66 percent increase from the original 60 cents mark we saw earlier! What can be concluded from this is that women who are highly educated, never married, prime working-age and full-time workers earn an almost identical amount of money that men of the *same description* earn.

So how much should we weigh each of these factors? To get an idea I've decided to turn to tree modeling to predict income on the basis of some of these different variables. I used a train-test split with the training data making up 80% of the observations and the testing split making up the other 20%. I used three different approaches: CART, random forests, and gradient-boosted trees to predict total yearly income. I used the testing data to come up with a root-mean-squared-error estimate so as to provide me with the most explanatory model, which ended up being the gradient-boosted tree.

To accomplish this I had to drastically reduce the data set as the original data set I've been using up until this point has over 3,000,000 observations. I decided to run these test only on the state of Washington as it isn't too large for R to run these tests, and it has a fairly diverse population demographically. The model contains total income as the left-hand side variable, and age, hours worked, and education as predictor variables.

Here are the Root Mean Squared Errors of the three different models tested. In order they are the CART, random forests, and gradient-boosted trees models:

```
## [1] 57682.36
```

```
## [1] 54683.5
```

```
## Using 500 trees...
```

```
## [1] 53270.71
```

Below are two partial-dependence plots that isolate the predicted change in yearly income with education and average weekly hours worked using the gradient-boosted trees model.

Figure 1: Predicted Income by Years of Education

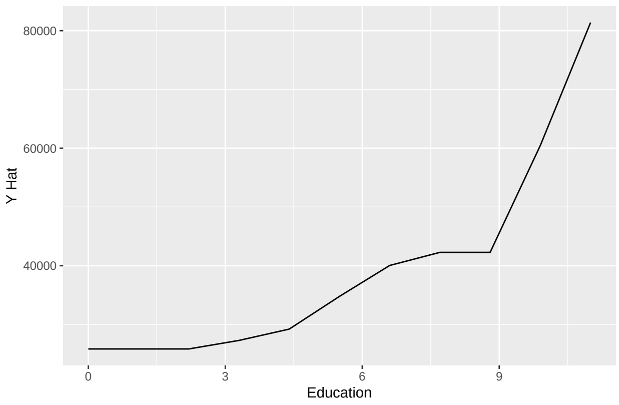
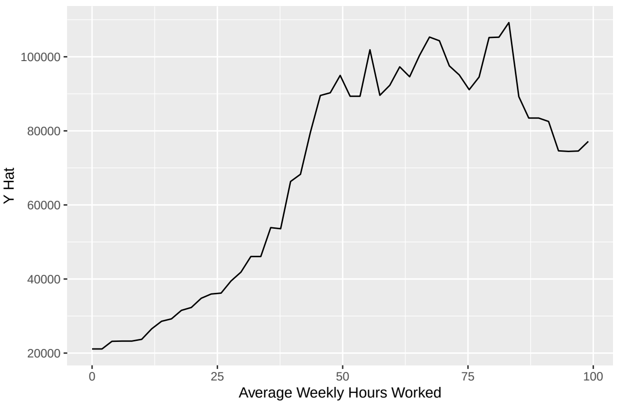


Figure 2: Predicted Income by Average Weekly Hours Worked



The outcome is unsurprising, there are strong positive relationships between education and total income, as well as between weekly hours worked and total income. In both cases the effects begin to wear off at the extreme ends of the spectrum. Hours worked having a very large impact on predicted total yearly earnings is probably the least surprising fact, since more hours worked almost always means more pay.

I should note that the way the education variable is coded is such that no schooling = 0, nursery school to grade 4 = 01, grade 5, 6, 7, or 8 = 02, and then each additional year has its own code number all the way up to the 5+ years of college mark, which is coded as 11.

Age was also included in the model with a partial-dependence plot, which appears in the appendix. It also has a positive relationship to yearly earnings and it appears to be ever increasing without much tapering off until we get into the 80 years of age range. This initially sticks out as relevant for the following reason: if, on average in the US, men as a whole are much older than women as a whole, then this could also go a long way in explaining the disparity. Women do live five years longer than men on average, but the difference in the average age of men and women on the whole is not large enough to warrant suspicion that this is playing a large role in driving the gender wage-gap.

Education has long been thought to be a major driving factor in explaining differences in income, not just between men and women, but in general. Economists have devoted great resources into isolating just how much of an effect education has upon lifetime income. It is universally acknowledged among professional

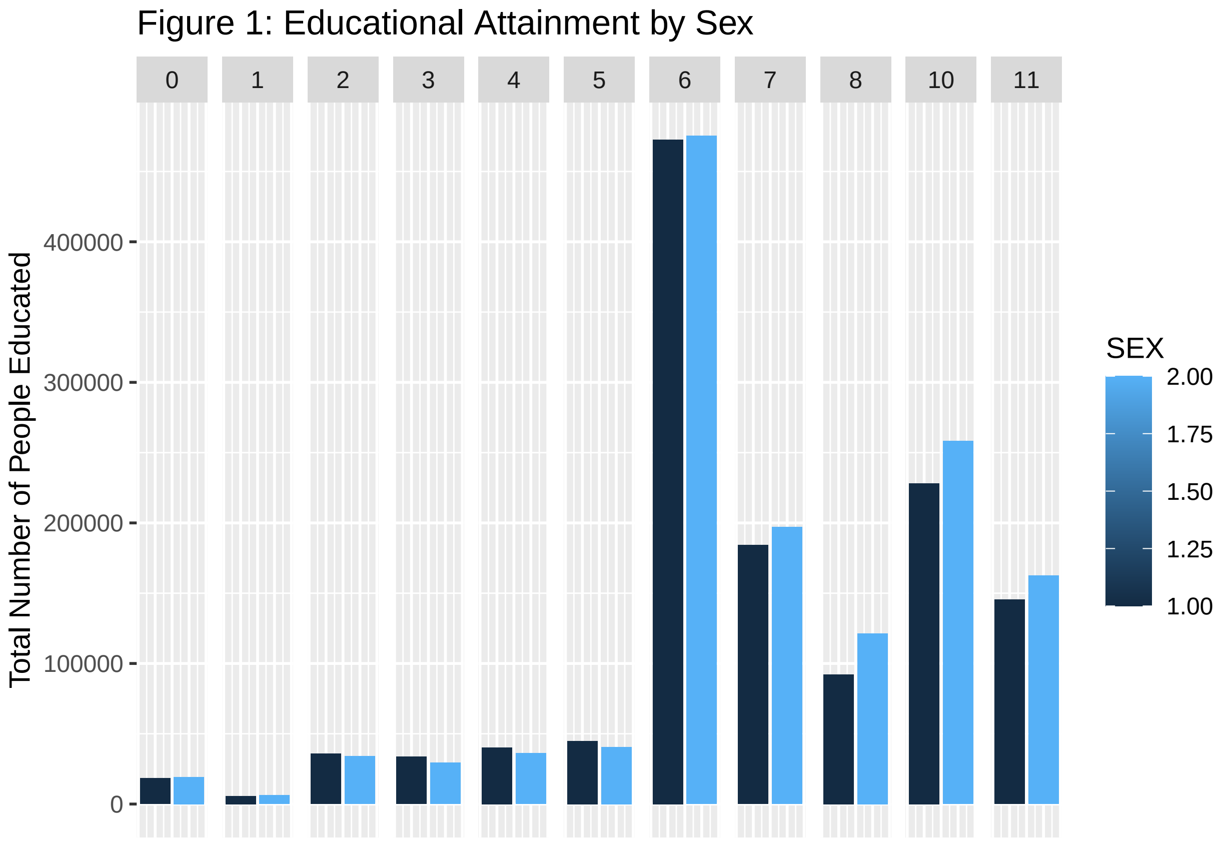
Table 4: Average Education by Sex

Sex	Avg Education	Std Dev
1	7.151297	2.479773
2	7.304732	2.463081

Table 5: Average Hours Worked By Sex

Sex	Avg Hrs Worked	SD
1	27.19017	21.94856
2	20.65302	20.04782

economists that education has a positive relationship with income. This being the case, let's see how men and women stack up in terms of educational attainment:



According to Figure 1, it appears that women have a slight advantage in educational attainment over men. Facet number 6 indicates high school graduate, and facet 10 indicates 4 years of college. If there had been a substantial difference in educational attainment favoring men in either of these facets, then we likely could've closed up shop with a definitive answer as to why the gender wage-gap exhibits the disparity is does at the national level. Unfortunately, things aren't going to be so easy for us here as **Table 4** indicates that women have slightly more education on average than men in the United States.

So, what about hours worked? We turn to that next in **Table 5**.

Here we have a large difference between men and women in terms of average hours worked per week. Men work about 31% more hours per week than women do on average. Since we lack data on whether subjects have

Table 6: Top 5 Jobs by Average Salary

Job	Avg Salary	Std Dev
3100	329874.9	223718.5
3090	239887.4	185100.9
3010	191973.9	163602.6
10	184314.7	178626.9
2100	173194.4	160732.6

ever had children and the subjects' labor-force continuity, the next best step explanation for the disparity might well be the difference in hours worked.

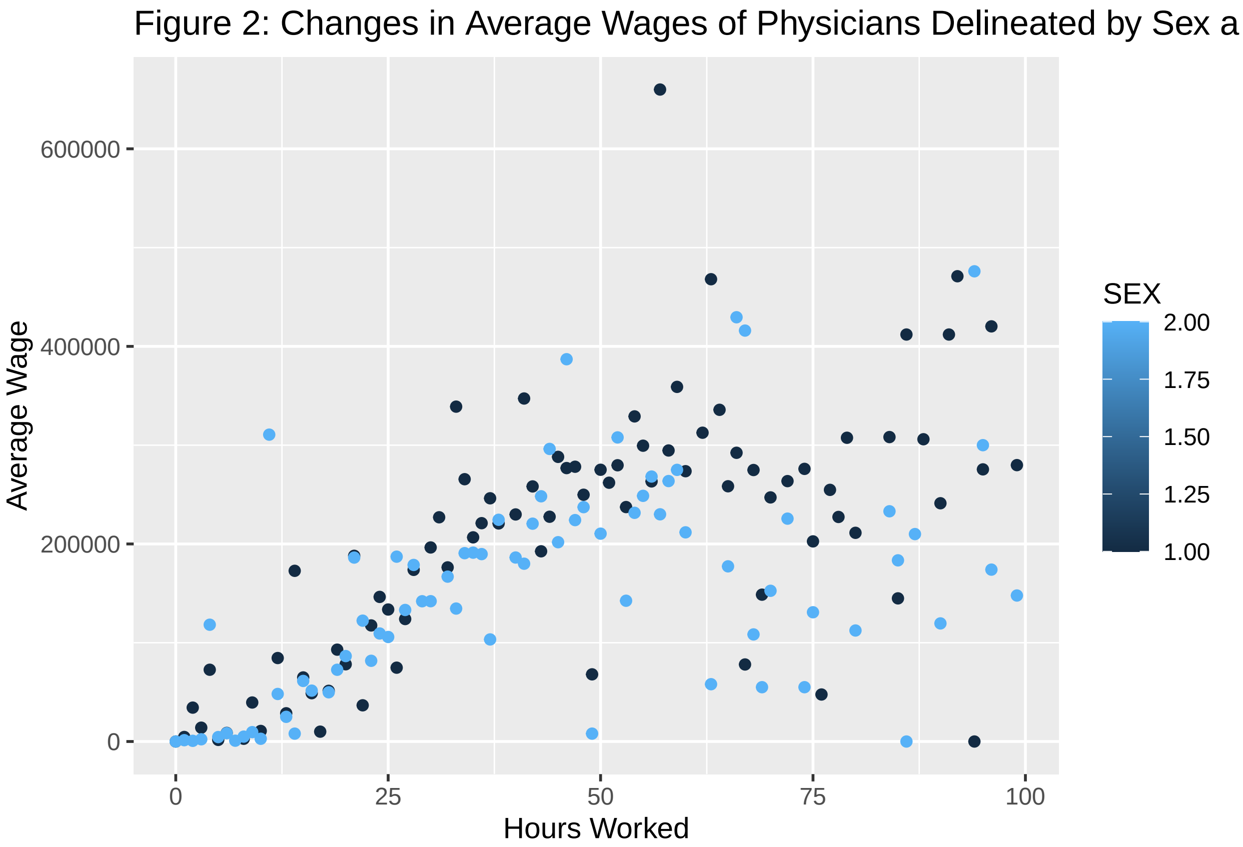
Another factor that we should look into is how many women are filtering into the highest paying jobs are in the United States. This could go a long way in providing insight as to what is going on behind the scenes with the pay-gap. **Table 6** ranks the top 5 professions in the United States by average salary. The professions are coded in the following way: 3100 = "Surgeon", 3090 = "Physician", 3010 = "Dentist", 10 = "Chief Executives and Legislators", and 2100 = "Lawyers, Judges, Magistrates, and Other Judicial Workers."

Across the board I found that men dominate these fields. I've included plots in the appendix that demonstrate this for each profession listed in **Table 6**. This is fairly explanatory. If men disproportionately fill up the jobs that are the highest paying jobs in the economy, then we should expect men to earn more. But why do men tend to outnumber women in these professions? Is there discrimination going on here? The explanation for this has largely been attributed to the fact that becoming and Surgeon, Physician, or Dentist requires intense and lengthy graduate school work that, for most women, would take place during prime childbearing years, which pushes them away from pursuing careers in those fields.

This doesn't rule out discrimination though. If there is systematic discrimination, we should see women who are in the same lucrative field as men, who have the same educational attainment as men, who work the same hours as men, with the same experience as men, making less money than men. We can't control for experience in this data set, but we can control for hours worked as well as job function. That is where we turn to in Figure 2:

Table 7: Physician Pay by Sex

Sex	Average Pay	Average Work Hours
1	269775.3	46.58024
2	191124.2	45.11139



Based upon the pattern we see in Figure 2, it looks like there is a large amount of overlap between male and female physicians up until about the 50 hour per week mark, at which point there is much more variance for both men and women and the benefits of additional hours worked begin to taper off. **Table 7** shows that male physicians earn about 41% more yearly earnings than female physicians on average, but that men work 3% more hours per week on average than women do. Since male physicians work 3% more hours per week, and since this compounds itself over the year to yield a higher total income for men, a major component in the explanation of the 41% yearly income disparity between male and female physicians can be found in hours worked.

Conclusion

The purpose of this study was to gain insight into what explains the gender pay-gap. If it were the case that systematic discrimination against women on the part of employers exists, we could easily see this if we found that women working in the same field, with the same education, and the same hours as men, earn less money than men.

There were several limitations to this study due to constraints in the data. We didn't have access to information on whether a mother or father ever had children, and we don't have access to years of experience in the field of interest. However, we made due with what we had and came away with very interesting results.

When controlling for profession, marital status, education, and age, we found that the difference in incomes between men and women were reduced to a trivial amount. We found that education plays an important role in explaining income, but that women have a higher educational attainment on average than men, which means that education *per se* is not relevant to the issue at hand. It might well be the case that the *types* of disciplines studied predominantly by men or women plays a role, but we don't have data on that, so we had to move on.

We then controlled for hours worked and found that men work 31% more hours per week on average than women. When we looked at the most lucrative professions in the United States, we found that men dominated the top 5 fields. When looking at the difference in earnings between male and female physicians, which was the second most lucrative job in the US, we found that male physicians earn about 41% more than women do on average, but that men work 3% more hours per week. Although the disparity in the hours work doesn't explain all of the difference in earnings between male and female physicians, it certainly goes a long way since working more hours tends to earn one more money.

Upon analyzing the data, we've provided some insight as to what is driving the gender pay-gap and what can be done to help close the gap between male and female workers.

Appendix

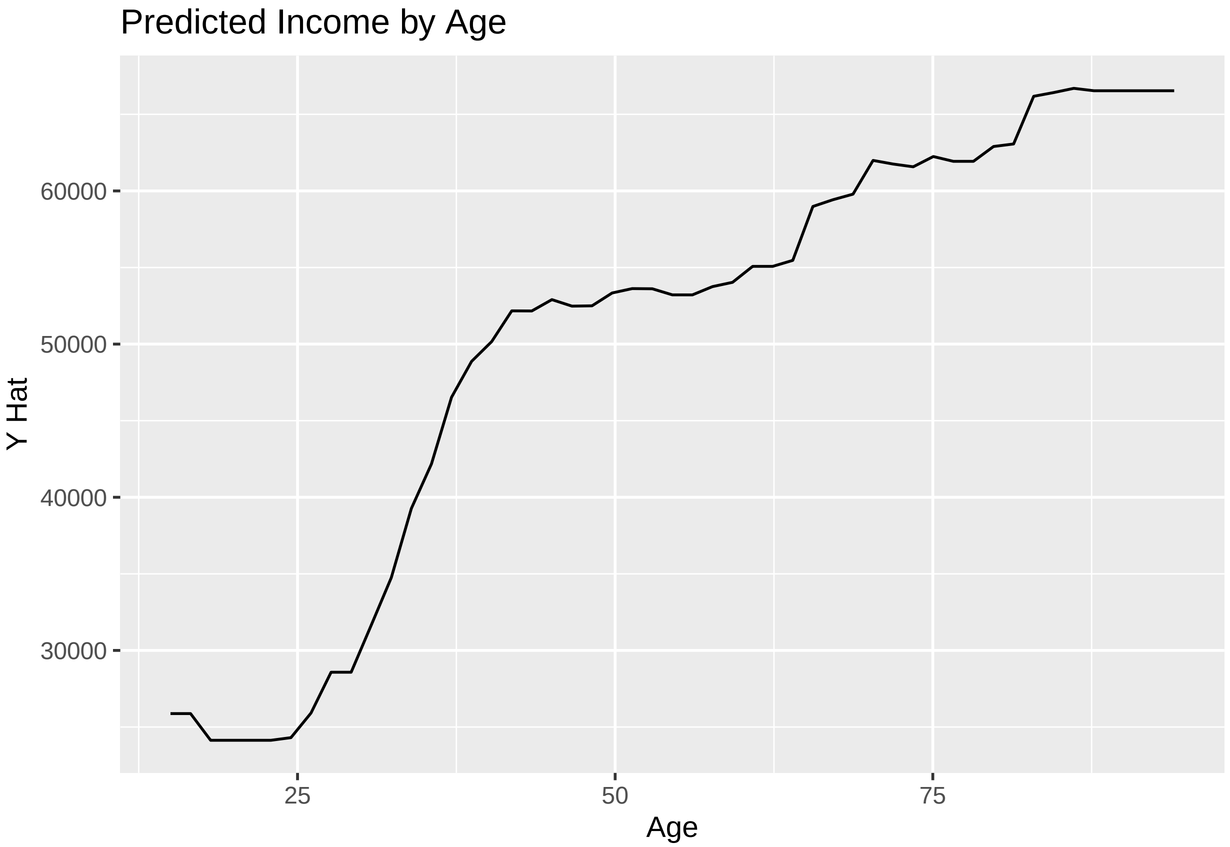
The following are miscellaneous tables and graphs that weren't included in the body of the paper.

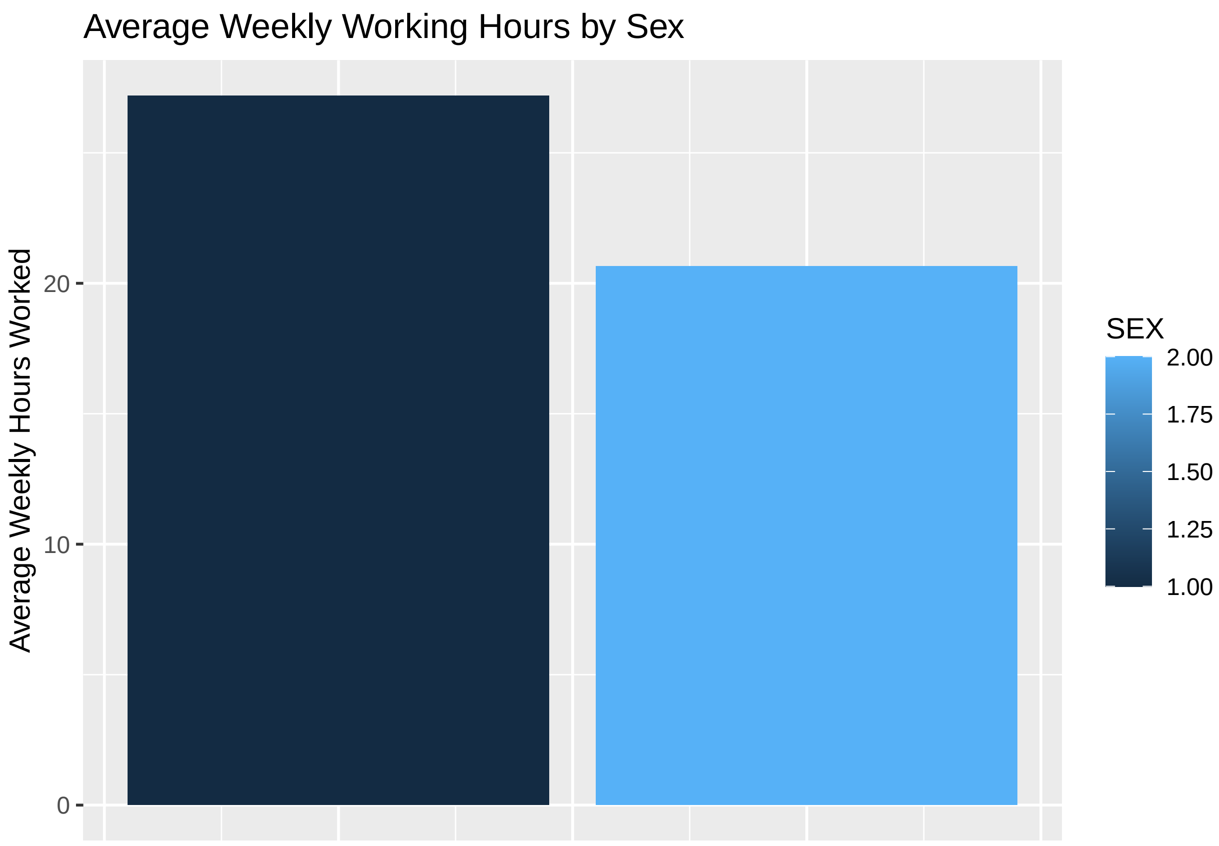
SEX	Average_Income	Standard_Deviation
1	123051.37	131139.12
2	70948.22	77842.51

SEX	Average_Income	Standard_Deviation
1	73975.23	83469.14
2	63622.46	63470.61

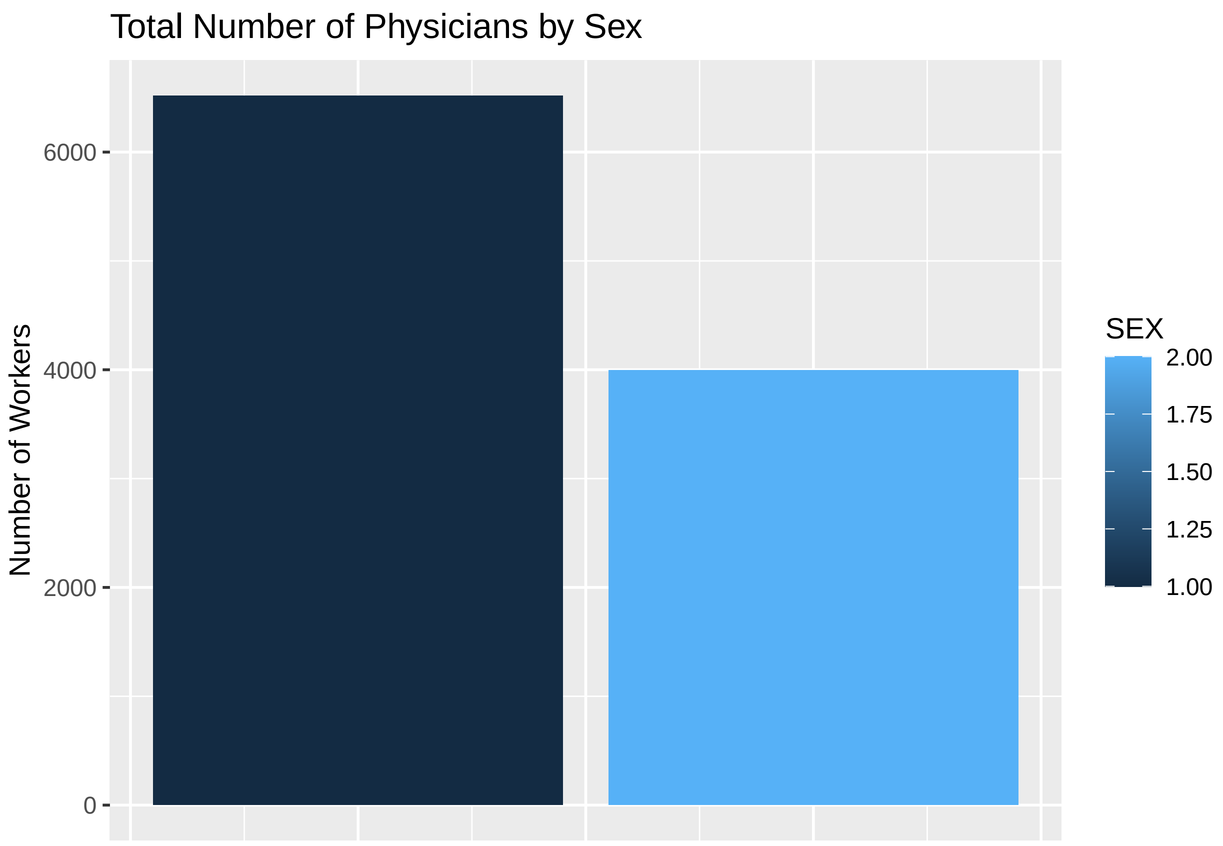
SEX	Average_Income	Standard_Deviation
1	77312.70	82917.29
2	66074.98	61861.22

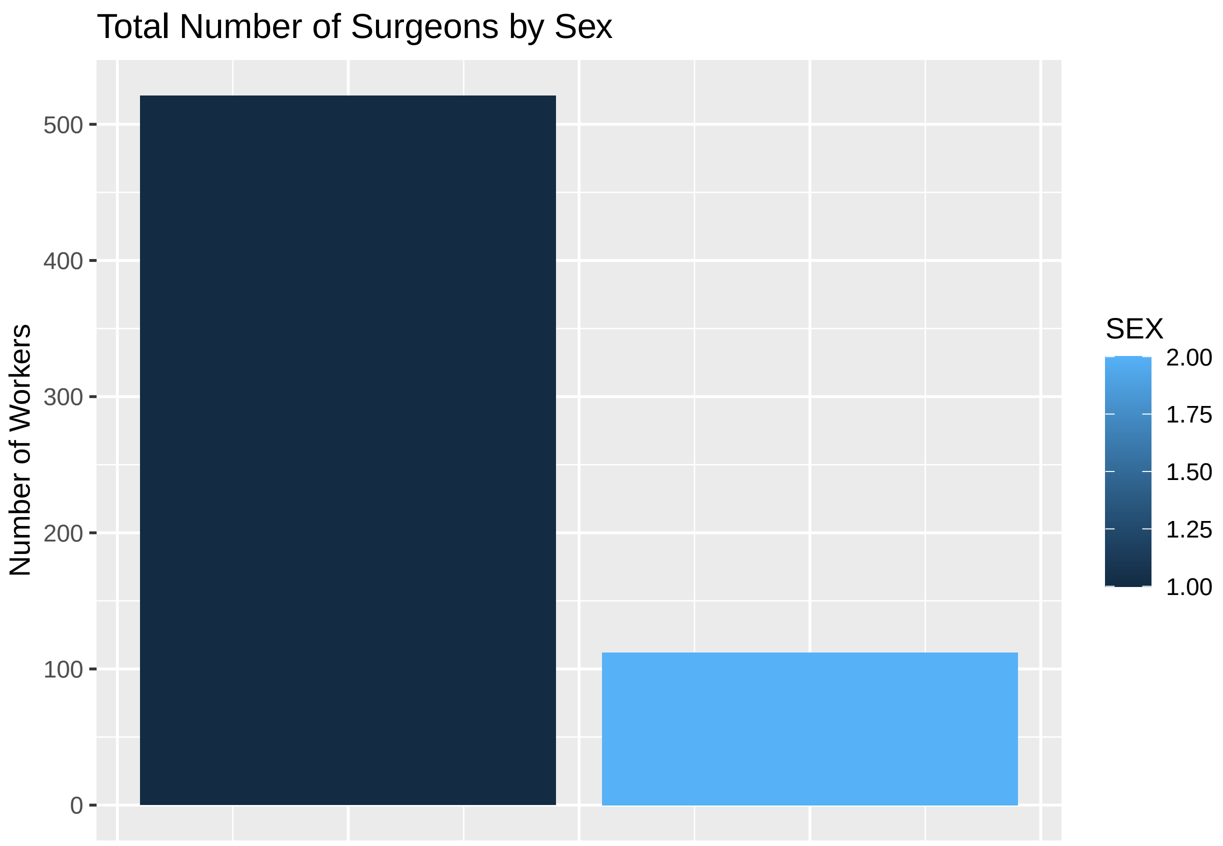
SEX	Average_Income	Standard_Deviation
1	76225.99	65391.90
2	65618.01	47333.92

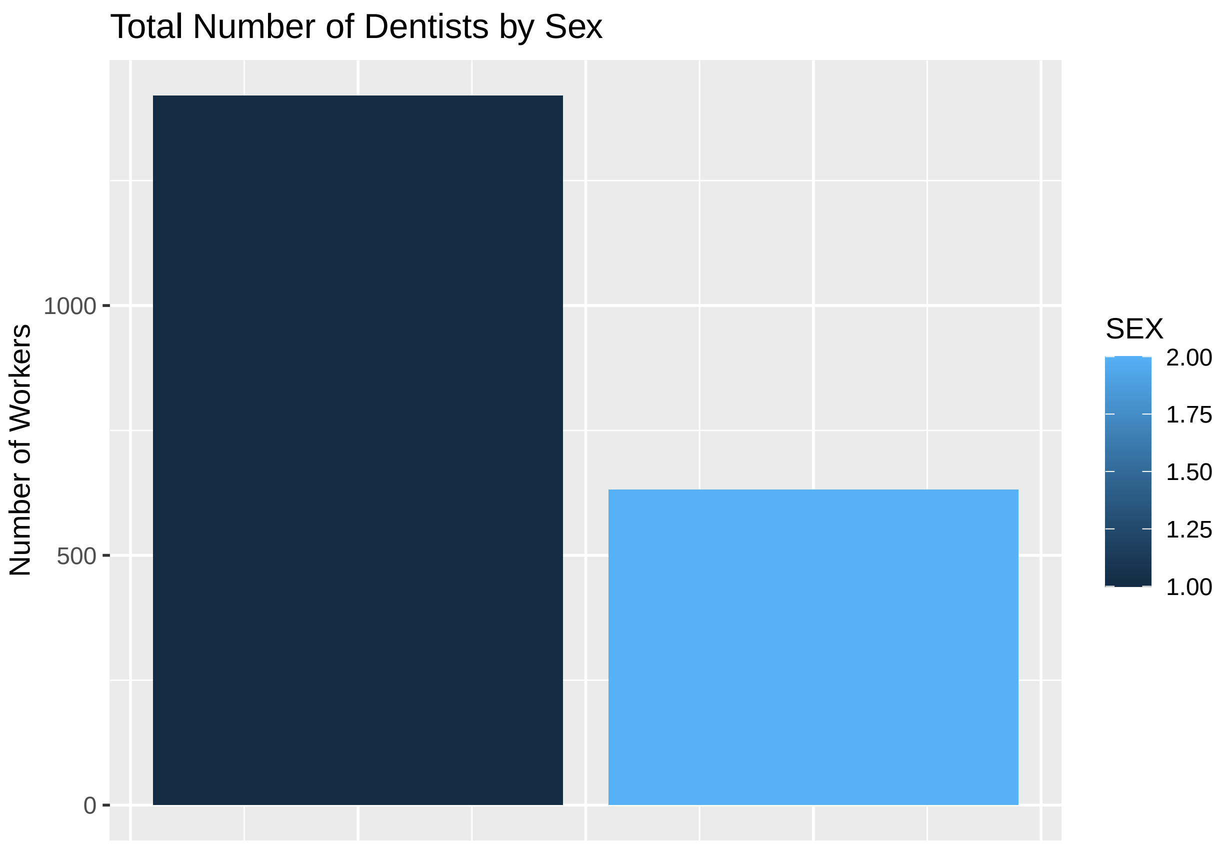

Predicted Income by Age
Y Hat
60000
50000
40000
30000
25
50
75
Age

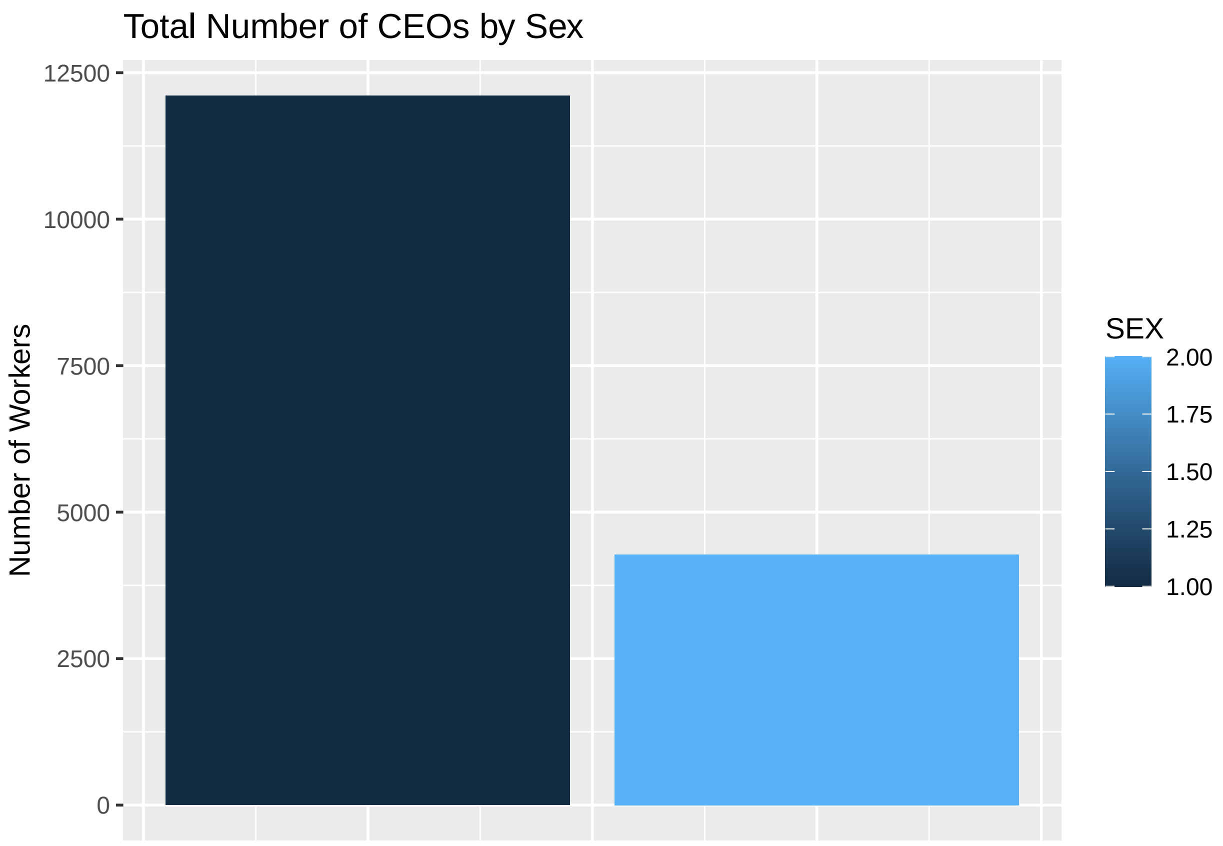


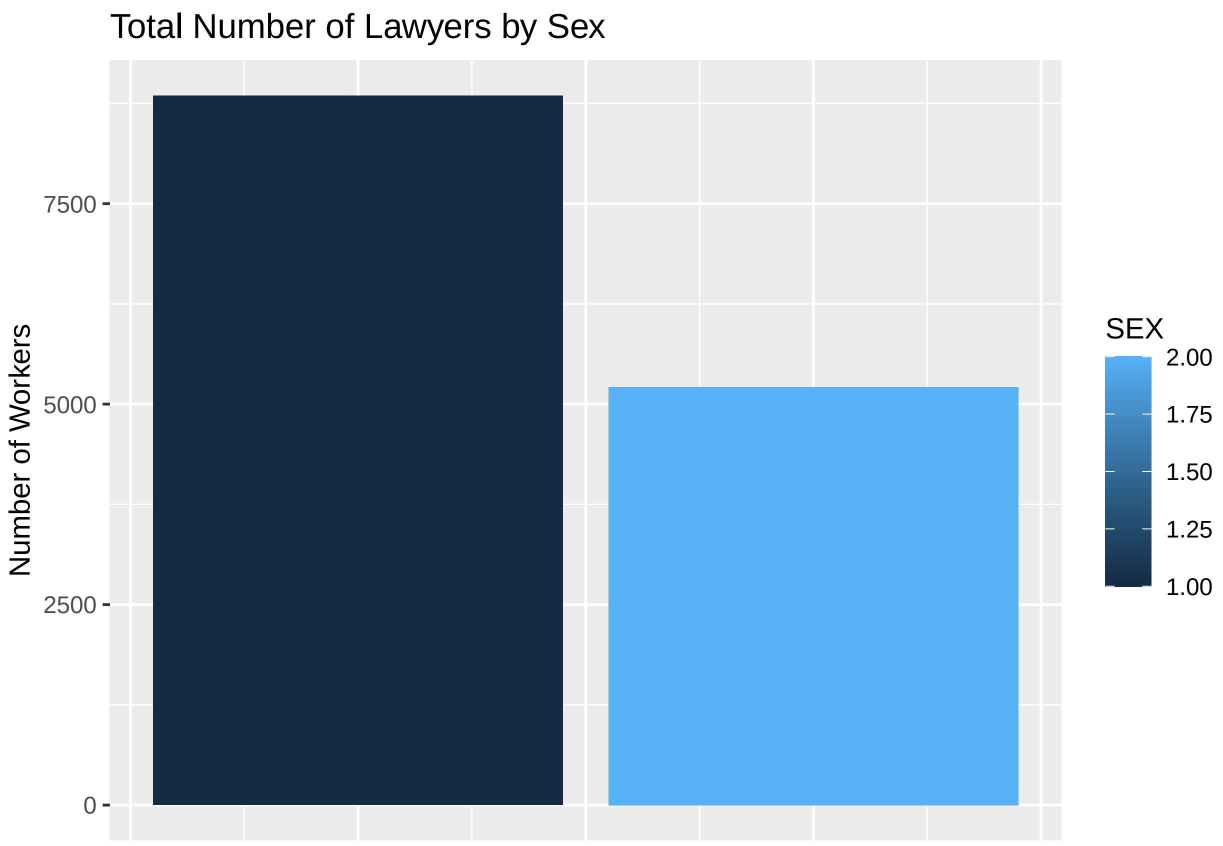
Average Weekly Working Hours by Sex


Total Number of Physicians by Sex
Number of Workers
0
2000
4000
6000
SEX
2.00
1.75
1.50
1.25
1.00


Total Number of Surgeons by Sex
Number of Workers
500
400
300
200
100
0
SEX
2.00
1.75
1.50
1.25
1.00


Total Number of Dentists by Sex
Number of Workers
1000
500
0
SEX
2.00
1.75
1.50
1.25
1.00


Total Number of CEOs by Sex
Number of Workers
12500
10000
7500
5000
2500
0
SEX
2.00
1.75
1.50
1.25
1.00


Total Number of Lawyers by Sex
Number of Workers
7500
5000
2500
0
SEX
2.00
1.75
1.50
1.25
1.00