A Look Into the Reasons Behind the Gender Wage Gap

By Kritika, JP, and Nick

Summary of questions and results:

1. How have gender wage gaps changed over time?
   - Gender wage gaps have generally lessened over time. However, the specific trends differ by state.

2. How are gender wage gaps affected across the different jobs and within the same jobs?
   - Gender wage gaps are most severe in the legal and financial fields. There was less of a pattern to fields with the least severe gaps, those being occupations in engineering and clerical positions.

3. What factors affect gender wage gaps? We’re going to research various possible factors such as age, department, education levels, and seniority.
   - The most significant factor affecting the gender wage gap is age, with older female employees suffering a greater gap than younger employees. In addition, the wage gap was slightly more severe for employees with college degrees more advanced than a bachelor’s.

Motivation:

The average yearly wage received by women in the United States is estimated to be 80% of that received by men. This trend can be observed across various fields, ages, states, seniority, and education levels. This wage gap leaves women at a disadvantage when trying to support themselves and their families. By investigating the causes and underlying trends behind the gender wage gap, we hope to spread awareness of the discrimination and disadvantages that the audience or anybody they know may be experiencing.

Dataset:

We had a total of 7 datasets:

Glassdoor income for specific jobs use Glassdoor Gender Pay Gap.csv

The average income for all jobs use CurrentPopulationSurvey.csv and PanelStudyIncomeDynamics.csv
https://www.kaggle.com/datasets/fedesoriano/gender-pay-gap-dataset
Geospatial dataset used for US state boundaries shape files *Cb_2018_us_state_500k.zip* https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html


**Method:**

1. **How have gender wage gaps changed over time?**

   To track changes in the wage gap over time, we used the *CurrentPopulationSurvey.csv* dataset. This dataset stores the year of recording in column 1 and the wage and salary income for that year in column 50. We used a bar plot to track the mean wages of male and female employees by year, then created a second bar plot to track the ratio of female wages to male wages by year.

2. **How are gender wage gaps affected across the different jobs and within the same jobs?**

   To track changes in the wage gap over different jobs, we used the *inc_occ_gender.csv* dataset. We plotted the linear regression between the female and male mean wages across 142 occupations. Using the same data, we calculated the ratio of female to male wages for each occupation on a percentage scale. We then created a histogram to display the data.

3. **What factors affect gender wage gaps? We’re going to research various possible factors such as age, department, education levels, and seniority.**

   To track the factors that most affect the wage gap, we primarily used the *Glassdoor.csv* dataset. We used column 3 for age, column 5 for rough education level, column 6 for department, and column 7 for seniority in years. To verify these results, we used the *CurrentPopulationSurvey.csv* and *PanelStudyIncomeDynamics.csv* datasets to track the differences in the wage cap by age and education levels across all occupations.

**Results:**

1. **How have gender wage gaps changed over time?**
In regards to how gender wage gaps have changed over time, the overall trend is that they have decreased, something that we expected. The first way we analyzed this was by looking at two datasets: a collection of Census Population Survey data and a collection of Panel Study of Income Dynamics data (with data from years between 1980 and 2013). We had about 5000 different rows of data per year. For each year with data available, we found women’s wages as a percentage of men’s wages on average and plotted it. Our visualization is below. The CPS data is on the left subplot, while the PSID data is on the right subplot:

To prove this general visually upward trend mathematically, we calculated Pearson's coefficient of correlation between both of these separate datasets which was around 0.98, which is very high. This validates our findings of a general upward trend as women’s wages are becoming closer to men’s wages between 1980 and 2013.

In fact, we decided to showcase this change over time in another manner as well - using geopanda plots to show the change of gender wage gaps over time per state. Below are the visualizations we created for 2010 and 2020.
Looking at these visualizations, the general trend of the wage gap decreasing is visible, as the colors are lighter in the second visualization, indicating that women’s wages are overall closer to men’s wages now than 10 years ago. We couldn’t find data from 2021 or 2022, but we found it interesting that despite the COVID-19 pandemic (at the heart of 2020), which had a drastic impact on the economy, the gender wage gap had still visibly lessened in 2020. Furthermore, it was interesting to see how while the wage gap had generally decreased, some specific states don’t follow this trend (specifically Idaho and Utah, for instance, actually become darker, signaling that the wage gap is increasing).
This upward trend is very important, as it shows how over time, this issue of inequality in the workforce is being actively accounted for and targeted. While in 1980, the average woman’s wages were about 50% of the average man’s wages, that percentage has risen to about 90% in many states today.

2. How are gender wage gaps affected across the different jobs and within the same jobs?

Our results showed the wage gap varied significantly between the 142 occupations we analyzed. Over all of these occupations, the data followed a roughly normal distribution, with a median wage ratio of 80% and with most of the data falling between 65% and 96%. These results are based on data taken as of the year 2015.

*Average female weekly income by average male income for each occupation.*
The occupations with the highest income ratios included food services (94.76%), hotel desk clerks (96.09%), electronics and electromechanics assemblers (96.11%), counselors (99.34%), police officers (100.80%), and wholesale and retail buyers (111.17%). The occupations with the lowest ratios included financial sales agents (52.50%), sales representatives (60.94%), and financial managers (65.24%).

Within most occupations, both male and female employees saw wages increase with age, education level, and seniority, but the ratios remained generally fixed.

3. What factors affect gender wage gaps? We’re going to research various possible factors such as seniority, age, and education levels.

To analyze the trends that could influence the wage gap, we used a variety of datasets and focused primarily on the employees’ ages, the amount of time spent in their positions, and their highest level of education completed. We found that the age gap was likely to increase over time for employees between the ages of 25 and 40, then remained roughly constant for employees between the ages of 40 and 64. The final wage ratio the data settled to was around 60%.
As we can see from the above plots, female employees suffer from a reduced mean income compared to male employees, and tend to stop receiving pay increases earlier. When using the glassdoor dataset to determine the wage gap by age, we received an increased female to male wage ratio of roughly 80%.
The female to male wage ratio by age, calculated using an additional dataset.

To test the validity of our data, we calculated the Pearson Coefficient and received a value of 78.00%.

When analyzing the wage gaps by seniority using the Glassdoor dataset, we found that employees with higher seniority received higher wages, but we saw a far less drastic change in the wage ratio than we did for age. The ratios we received were from seniority levels between one year and five years, and all fell between 80% and 90%.
Female to male wage ratio by seniority in years

These plots suggest an employee who is paid less in their first year of employment will be unlikely to make up the wage gap by remaining in their current position.

When analyzing our data by college education level, we found that the wage gap increased slightly as employees obtained more advanced degrees. The ratios of female to male wages ranged between 60% and 65%. When using the *Panel Study of Income Dynamics* dataset, we found that on average, a female employee with an advanced degree would earn less than a male employee with a bachelor’s degree.
Female to male wage ratios by college degree. 0.0 represents no degree, 1.0 represents a bachelor’s degree, and 2.0 represents an advanced degree beyond a bachelor’s.

Male and female wages by college degree. 0.0 represents no degree, 1.0 represents a bachelor’s degree, and 2.0 represents an advanced degree beyond a bachelor’s.

When using the Glassdoor dataset to analyze the wage gap by more broad departments, we found that the wage gap was not significantly affected. All of the departments included in the data showed a female to male wage ratio of roughly 85%.
These plots suggest that the department employees work for is less significant in determining the gap in their wages, compared to the specific occupations they hold within their departments.

**Impact and Limitations:**
A significant stakeholder in our research topic that would benefit from our analysis is public + private companies. Knowing the disparity between how much women make and how much men make for the exact same job will bring awareness of this problem to these companies which have the power to bring changes on the financial side of the business. Women looking for general trends in salaries would also benefit as our visualizations show general trends and correlations regarding the gender wage gap in regards to state, age, seniority, profession, and department. This can be helpful for women wanting to move to a new state or join a specific industry or job.

Some limitations of our results are that we didn’t have data for all years between 1980 and 2013. For that reason, we had to use data from only the years with data available to generalize trends of how gender wages were changing over time. Similarly, when we were analyzing the impact of different factors on the gender wage gap such as age, education level, and departments, we only used data from a specific year: 2015. This may not reflect the current trend or trends in other years. Additionally, several of our datasets did not have data on races, so we were not able to ensure that the datasets we used did in fact accurately reflect all races fairly, especially minority races for which there may not be as much data, especially in past years. For the same reason, we also did not take into account poverty levels to ensure that communities with socioeconomic difficulties were accurately represented. For this reason, our analysis may not accurately represent all subgroups of American people. Another limitation of our analysis is that we did not have the data to analyze the impacts of the COVID-19 pandemic on gender wage gap trends. The pandemic definitely shook the American economy, leading to many people losing jobs or making a limited amount of their regular salary. It is very possible that these sudden changes have changed what wages by sex look like today.

Our analysis may harm companies that are currently doing their part in lowering the wage gap for their positions. Because our visualizations show the wage gap in different professions and industries in general, we are shedding negative light on them without distinguishing the specific companies for which this is a real issue and the specific companies that are paying men and women equal wages for the same jobs.

Challenge Goals:

1. **Multiple Datasets**: We combined datasets twice within our project, both to create the geospatial representation of gender wage gaps by state in 2010 and 2020. We found two datasets provided by the US Bureau of Labor Statistics regarding women’s wages. We also found a US Census Bureau dataset with the administrative boundaries of the United states. We filtered the shapefile to include only the geometry and state name. For both the US Bureau of Labor Statistics datasets, we filtered them down to State and women’s
wages as a percentage of men’s wages. We then combined these 2 datasets with the shape file so we could use geopandas to plot 2 visualizations displaying how the gender wage gap changed across states. Overall, we used 7 different datasets in total throughout our project to analyze the various factors that contributed to the wage disparity.

2. **Result Validity:** We wanted to show that the visualizations we were creating and the trends we were establishing were not specific to the dataset we used, but rather were truly applicable. To do this, we tried searching for several datasets that used similar metrics to compare factors that contributed to the wage gap. We were able to find 2 separate datasets with data about wage gaps over time. We also found 2 separate datasets with data about age in relation to income by sex. For both factors, we cleaned up the data and created 2 visualizations to prove the similarity between trends. To validate the trend statistically, we found Pearson’s coefficient of correlation between the datasets for both these variables and printed it out. Both were above 75%, proving that our results were not by chance and that they do display real trends.

3. **New Libraries:** We used plotly for several of our visualizations to create visualizations that the user could interact with. We also used Scipy to calculate Pearson’s coefficient of correlation for several datasets when approaching the result validity challenge goal.

**Changes to challenge goals:**

Our challenge goals changed quite a lot over the course of this project. We changed the way we approached using multiple datasets. Initially we had planned to merge two specific datasets (Current Population Survey + Panel Survey of Income Dynamics). However, we realized that merging them did not give us any helpful new data that we could analyze. Another challenge goal initially planned to have was ML. However, we soon realized that training an ML model would not work well for our project, as we only had data for specific years. If we were to feel the model data only from a few years, it would likely not be able to fill in the gaps and recognize the general trends for years in between. Instead, we used result validity as we found some datasets that overlapped in data, and we thought this was a good opportunity to show the similarity in trends across multiple datasets and validate our findings. We also used new libraries to support result validity and so we could produce more unique visualizations that captured our findings well.

**Work Plan Evaluation:**

For the most part, our work plan estimates were relatively accurate. However, because we removed our challenge goal of using ML, we instead spent a lot more time working on visualizations. We spent around a week on visualizations, unlike the 3-5 days we had planned.
This left us with slightly less time to formally answer our research questions and work on the report (about 2 days as opposed to the 5 days we had planned). Additionally, we had initially expected to clean and filter all our csv files before we started creating the visualizations. However, once we got started, we found that it was more efficient to do them simultaneously, as each visualization required the datasets to be filtered differently. In the end, most of the steps we outlined in our work plan still applied and were executed as we worked on the project.

**Testing:**

The way we decided to test our code was to utilize reduced versions of our CSV files and clean those files instead. We then compared them to an expected output file. The files that showcase differences between the file we clean and the expected output for the first question are called CPS Reduced Question 1 Differences and PSID Reduced Question 1 Differences. The file that shows the differences for the second question is called Glassdoor reduced gender differences. For the third question, the files that show the differences are called CPS Reduced Question 3 Differences and PSID Reduced Question 3 Differences.

**Collaboration:**

We did not seek much assistance from people other than the course staff and team members. We did get some advice on how to collaborate on code (through VS Code liveshare) and some possible websites from where we could get datasets from people who have previously taken the class. Other than that generic advice, we consulted online resources when we needed help with a new library or a specific bug.