# Technical Report - Gender Wage Gap

Ehram Haque

Sophomore Seminar Project, Spring 2023

**Problem Statement**

The U.S. government's population survey reveals that wages for men and women with the same job-related skills differ.

**Objective**

To identify the wage gap between men and women in the U.S.

**Purpose**

The gender wage gap began in the 1950s and increased over time (Gender Wage Gap, n.d.). Many factors contribute to the pay gap, such as race, ethnicity, education, age, and disability (The Simple Truth About the Pay Gap, 2022). As a result, it's crucial to check whether there are still differences in men's and women's wages.

**Description**

In this project, I analyzed data to observe significant wage differences between men and women with the same job-related skills. I developed the Ordinary Least Squares (OLS), a machine learning algorithm, to predict the wage difference between men and women.

**Data**

The data came from the U.S. Population Survey in the year 2012. I focused on single workers with education levels equal to high school, some college or college graduates. The sample size is approximately 4,000.

The outcome variable Y is an hourly wage, while the X's various characteristics of workers, such as gender, experience, education, and geographical indicators.

**Initial Data Analysis**

- The original data has 3835 rows and 12 columns. Each row corresponds to a single U.S. worker, and each column (variable) contains the corresponding worker's 12 information about their sex, wage, education, job experience, and geographic location.
- There are 8 int64 types, dummies, and binary variables; there are also 4 float64 type variables.

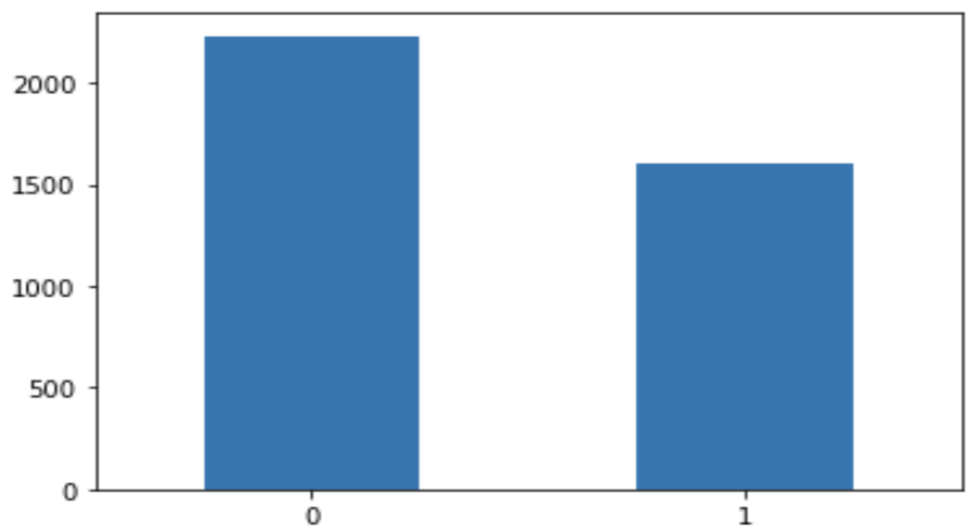
- There are *no* missing values in the data.
- The average wage is about 15 dollars per hour, while the maximum is 348.
- 42% of workers are women.
- The average experience is 13 years, with the minimum and maximum being 2 and 35 years, respectively, indicating that the data is diversified and drawn from various experience groups.
- 38% of the people in the data are college graduates, 32% have gone to some college, and 30% hold only a high school diploma.
- You can also see the geographical distribution of workers across major geographical regions of the states, and they seem to be nearly identical, between 22-28%, which again shows that the data was possibly collected from different regions in a uniform manner.
- Out of 3835 workers, 2232 are men and 1603 are women, as shown below.

```
df['female'].value_counts()
```

```
0     2232
1     1603
Name: female, dtype: int64
```

```
df['female'].value_counts().plot(kind='bar',rot=0)
```

```
<AxesSubplot:>
```




**Male** **Female**

**Proportion of female worker at various wage level**

```
df[df['wage']>20]['female'].value_counts()
```

```
0     475
1     240
```

Using this statement iteratively, the following table was produced:

| Salary per hour | > $10/h | > $20/h | > $30/h | > $40/h | > $50/h | > $100/h |
|---|---|---|---|---|---|---|
| # of Female | 1117 | 240 | 57 | 27 | 13 | 4 |
| # Male | 1634 | 475 | 150 | 63 | 36 | 5 |
| % of female | 41% | 34% | 28% | 30% | 27% | 44% |

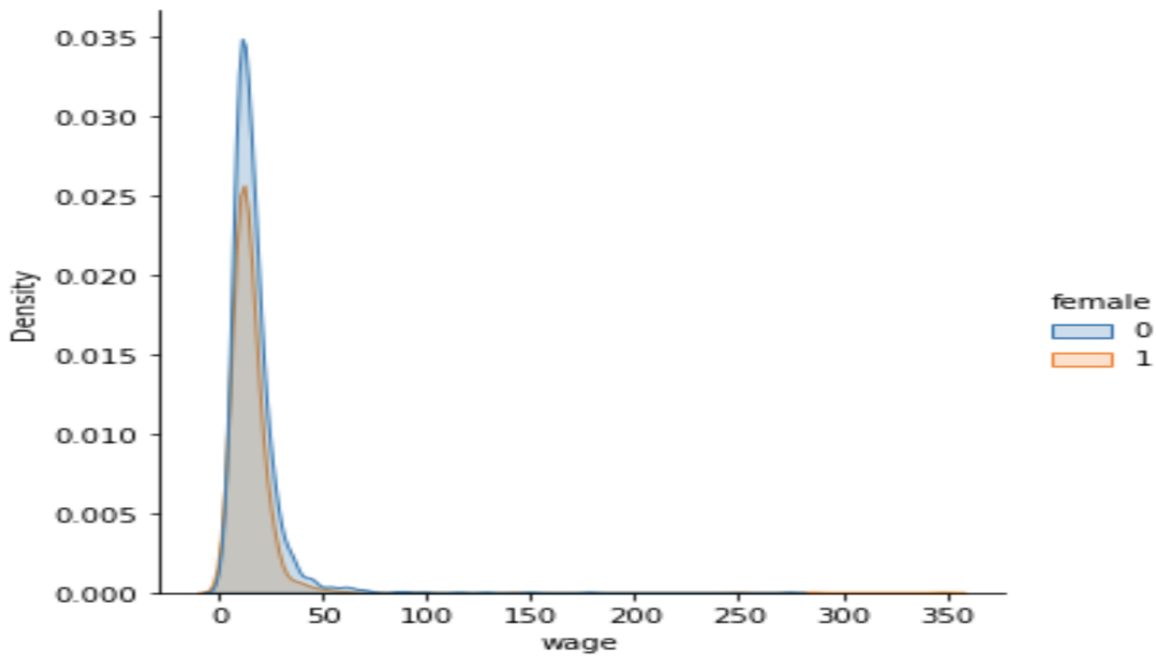
Comparing the above proportions to the original proportion:

| # Female | # Male | Total | % of female |
|---|---|---|---|
| 1603 | 2232 | 3835 | 42% |

Looking at the top table, you'll see that 42% of workers are women. However, almost all levels of the hourly salary are less than 42% (shown in the bottom table). The table indicates that the distribution of women who have the same qualifications as men is disproportional.

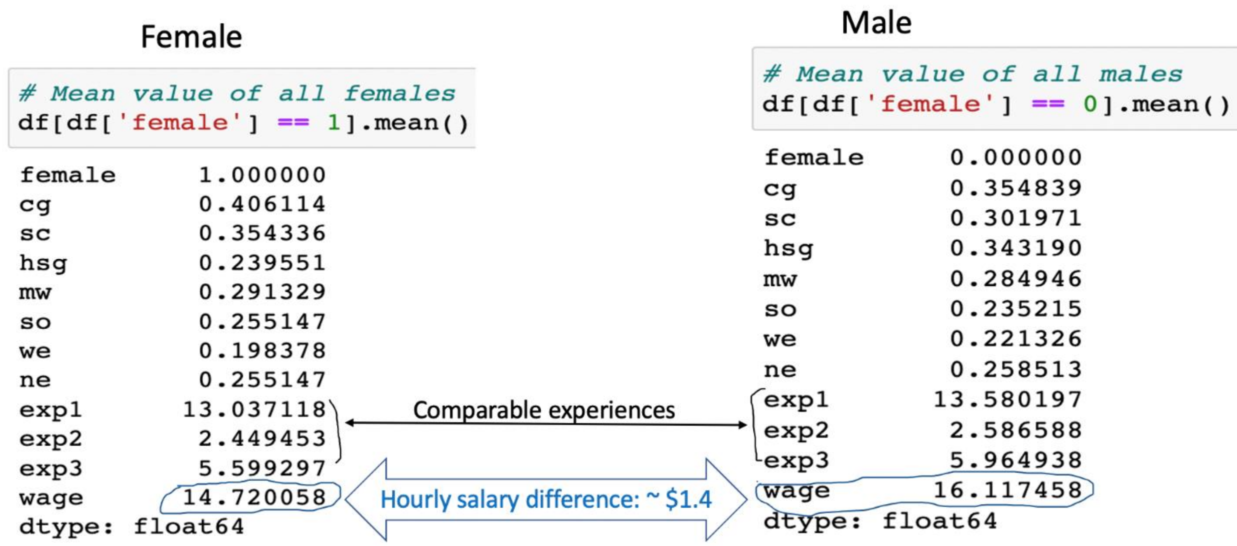
**Salary distribution between men's salaries and women's salary**

```
sns.displot(df, x ="wage", hue = "female", kind = "kde", fill = "True")
```




The wage distribution for men (blue curve) is broader than that for women (orange curve), which indicates that women make less than men.

**Salary comparison by hourly average**




The mean hourly wage is approximately \$16.11 for men and \$14.7 for women. The difference between the two salaries is about \$1.40 per hour.

Therefore, the wage gap exists, as shown above. Now we'll model the ordinary least squares method (OLS).

### Modeling by Ordinary Least Squares (OLS)

```python
import statsmodels.api as sm
Y = df['wage'] # target variable
X = df[['female' , 'sc', 'cg', 'mw' , 'so' , 'we' , 'exp1' , 'exp2' , 'exp3']] #regressors
X = sm.add_constant(X) # adding constant for intercept
model = sm.OLS(Y, X)
results = model.fit() # train the model
print(results.summary()) # summary of the model
```

```
                         OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.095
Model:                            OLS   Adj. R-squared:                  0.093
Method:                 Least Squares   F-statistic:                     44.87
Date:                Sat, 07 May 2022   Prob (F-statistic):           3.17e-77
Time:                        05:58:06   Log-Likelihood:                -15235.
No. Observations:                3835   AIC:                         3.049e+04
Df Residuals:                    3825   BIC:                         3.055e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.9154      1.299      3.784      0.000       2.368       7.462
female        -1.8264      0.425     -4.302      0.000      -2.659      -0.994
sc             2.4865      0.534      4.654      0.000       1.439       3.534
cg             9.8708      0.562     17.567      0.000       8.769      10.972
mw            -1.2142      0.566     -2.146      0.032      -2.323      -0.105
so             0.4046      0.588      0.688      0.491      -0.748       1.558
we            -0.2508      0.611     -0.410      0.682      -1.449       0.947
exp1           1.0965      0.269      4.077      0.000       0.569       1.624
exp2          -4.0134      1.785     -2.248      0.025      -7.514      -0.513
exp3           0.4603      0.344      1.340      0.180      -0.213       1.134
==============================================================================
```

The coefficient of the female indicator is negative, signifying that women are getting lower wages. Based on the given data, the model predicts that women would likely get lower wages.

**Conclusion**

- The ideas discussed were applied to this project and helped me learn about the gender wage gap.
- Based on the data, there is still a gender wage gap.
- The gender wage gap may partly reflect genuine discrimination against women in the labor market.