Report

Alan Wu, Oscar Su, Sean Gee, Shelby Jackson, Alton Law

2025-02-25

Introduction

This research project aims to develop a predictive model of pay gaps among different Irish companies to show how factors such as report years, hourly pay gaps, and gender composition of each company affect the pay gap within Ireland. Our data set was sourced from PayGap.ie, which contained 1917 observations and 29 variables to analyze. From these variables, we chose to analyze:

- **Report.Year**: The specified year in which data was collected
- **Mean.Hourly.Gap**: The difference between the mean hourly remuneration of employees of the male gender and that of employees of the female gender expressed as a percentage of the mean hourly remuneration of employees of the male gender
- **Median.Hourly.Gap**: The difference between the median hourly remuneration of employees of the male gender and that of employees of the female gender expressed as a percentage of the median hourly remuneration of employees of the male gender
- **Q1.Female**: The percentage of all female employees who fall within the lower remuneration quartile
- **Q2.Female**: The percentage of all female employees who fall within the lower-middle remuneration quartile
- **Q3.Female**: The percentage of all female employees who fall within the upper-middle remuneration quartile
- **Q4.Female**: The percentage of all female employees who fall within the upper remuneration quartile

From the perspective of large companies, it is of the utmost importance to minimize pay gaps between the genders to maintain egalitarian principles. In this project, we explore whether pay gap trends exist between the genders within Irish Companies throughout three years. To perform this, we let Mean.Hourly.Gap act as a response variable, while Report.Year, Q1.Female, Q2.Female, Q3.Female, and Q4.Female act as predictor variables. The analysis of this study was conducted using R Studio. Multiple linear regression was used to observe the relationship between the latter variables.

This study reports summary statistics of all utilized variables, scatterplot matrices, diagnostic plots, and density plots. Our data was initially fitted using a generalized full model. However, through a series of descriptive tests, we saw that the relationship between mean pay gap and gender/year lacked heteroscedasticity. To mitigate this issue our team used a weighted least squares (WLS) regression model to preserve constant variance. Effectively, allowing us to better understand the relationship between pay gaps among

different Irish companies while taking into account gender and year. Ultimately, we unveiled that gender was statistically significant to mean pay gaps, whereas year was not. These results provide modern insights into the persistence of gender-based pay disparities across Irish companies.

```
library(tidyverse)
library(car)
library(GGally)
library(MASS)
library(vtable)
data <- read.csv("cleaned_dataset.csv", header=TRUE) %>% dplyr::select(-X)
attach(data)
```

Summary statistics and data exploration

Distribution of categorical variables

```
par(mfrow = c(1, 2))
company_type <- table(NACE.Letter)
barplot(company_type, xlab="Company Type by NACE Letters", ylab="Count")

report_years <- table(Report.Year)
barplot(report_years, xlab="Year Reported", ylab="Count")
```

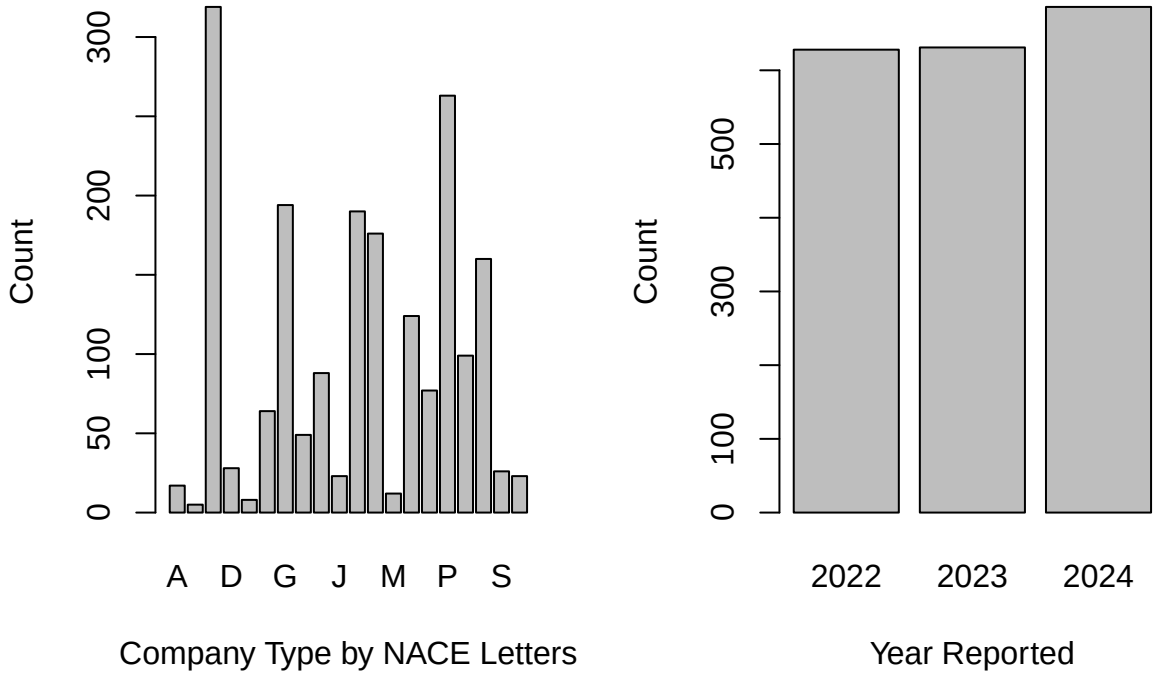


Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Report.Year	1945	2023.03	0.821606	2022	2022	2024	2024
Mean.Hourly.Gap	1945	11.1995	12.5103	-52.1	3.4	17.96	73
Median.Hourly.Gap	1945	8.12264	13.6991	-120.17	0.49	15.5	100
Q1.Female	1945	50.9757	18.237	0	39.9	62.5	100
Q2.Female	1945	47.8691	20.8093	0	33	61	99
Q3.Female	1945	45.5544	21.4042	0	31	59	100
Q4.Female	1945	39.9133	19.4888	0	26.5	50	99
Percentage.Employees.Female	1945	46.0782	18.4675	0	34	56.25	98.25

Extract variables

There are many variables that will not contribute to our research question, so we removed them from our exploration and analyses. Our research question is interested in the relationship of gender pay gaps and gender composition in companies, so we excluded company names and types even though they could be affecting pay gaps under other contexts.

We also excluded types of pays, including bonus and BIK (benefit in kind), since we are only interested one response variables and hourly gaps are the most commonly recognized meausre of comparing wage gaps.

Finally, we also excluded percentage of males since it's generally a reciprocal of percentage of females. While the percentages do not correlated with each other 100%, for practicality we are only using percentage of female employees as gender composition of companies.

In the end, we examined report years, hourly pay gaps, and gender composition of each company to see which variables are the most sensible to construct a model from.

```
new_table <- data.frame(Report.Year, Mean.Hourly.Gap, Median.Hourly.Gap, Q1.Female, Q2.Female, Q3.Female
```

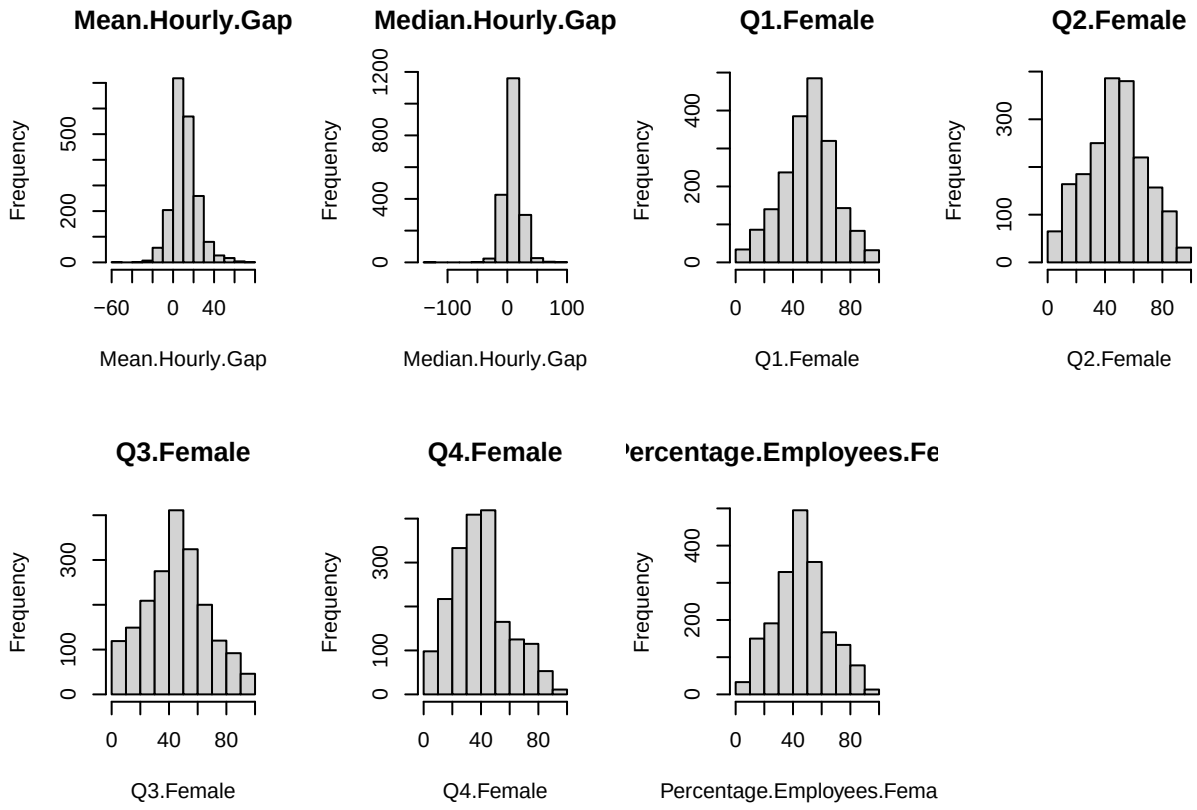
Summary Table

```
st(new_table, digits=6)
```

From the summary table, we can see that there are companies with very few or even no female employees. Upon further examination, we discovered that some of these companies reported 0 in the pay gaps. Including these companies in our model will make no sense as 0 for these companies does not mean there are no difference in the hourly pay between gender, but rather *since there are no female employees, there doesn't exist a pay gap*. Since these companies will **NOT** contribute in answering our research question, we will remove them from our subsequent model analyses.

Distribution of variables

```
par(mfrow = c(2,4))
for (col in 2:ncol(new_table)){
  hist(new_table[, col], main=colnames(new_table)[col], xlab=colnames(new_table)[col])
}
```



All the quantitative variables look approximately normally distributed, with the pay gaps centered around zero and the composition variables centered around 50%.

The rationale of reporting median pay gaps is that it is resistant to the influence of outliers. For example, if a company has a high median pay gap but a low mean, this means that only a few female employees are receiving similar pays as their male counterparts while the rest get paid a lot less. However, judging from the summary table and distribution graphs, the mean and median pay gap of the data set share similar distribution, which shows that both measures capture a similar pay gap situation, so it is pratical to just use one of them. Since we're only interested in one response variable and the mean is a more commonly used measure, we picked mean hourly gap as our response variable for our model.

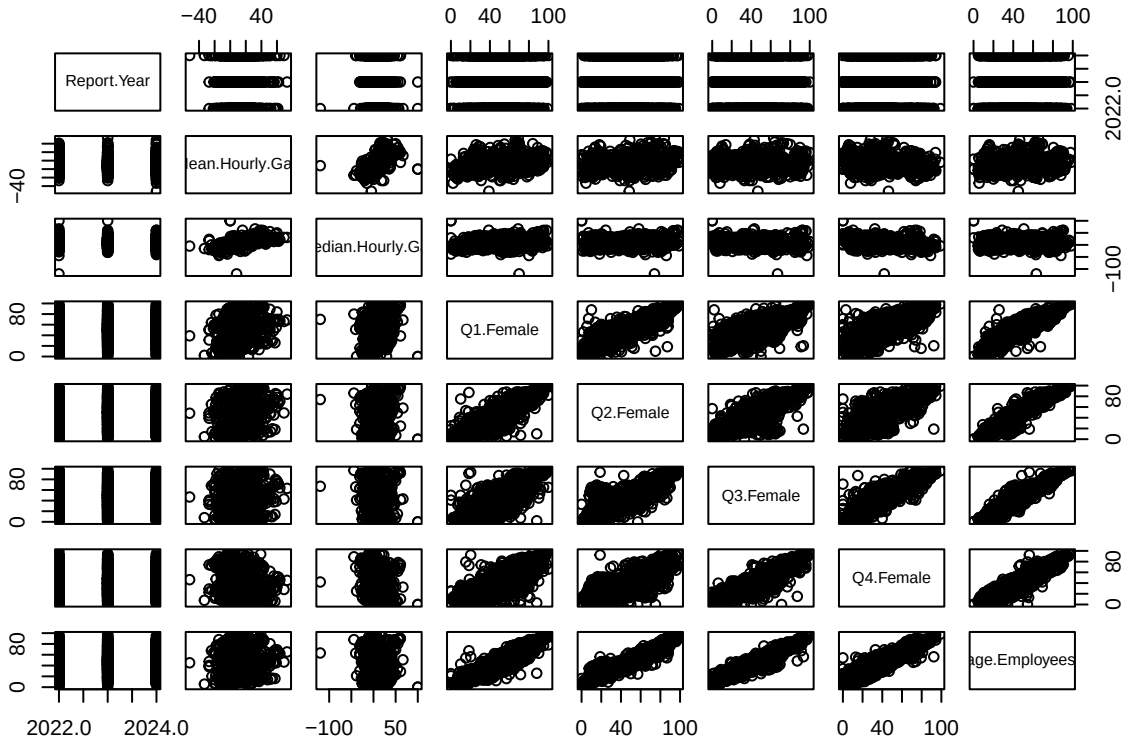
Correlation table

```
cor(new_table)
```

```
##                                Report.Year Mean.Hourly.Gap Median.Hourly.Gap
## Report.Year                   1.000000000     -0.02348382       -0.01124693
## Mean.Hourly.Gap              -0.023483824      1.00000000        0.69990044
## Median.Hourly.Gap            -0.011246931      0.69990044        1.00000000
## Q1.Female                    -0.013744612      0.24688464        0.13511161
## Q2.Female                    -0.022871926      0.20630734        0.06206905
## Q3.Female                    -0.014419243      0.00346460       -0.23752267
## Q4.Female                    -0.002373881     -0.21686709       -0.33463570
## Percentage.Employees.Female -0.014640714      0.06285680       -0.10626813
##                               Q1.Female   Q2.Female   Q3.Female    Q4.Female
```

```
## Report.Year                  -0.01374461 -0.02287193 -0.01441924 -0.002373881
## Mean.Hourly.Gap               0.24688464  0.20630734  0.00346460 -0.216867090
## Median.Hourly.Gap             0.13511161  0.06206905 -0.23752267 -0.334635702
## Q1.Female                     1.00000000  0.84595646  0.76697197  0.722243571
## Q2.Female                     0.84595646  1.00000000  0.83660128  0.762222025
## Q3.Female                     0.76697197  0.83660128  1.00000000  0.881509926
## Q4.Female                     0.72224357  0.76222202  0.88150993  1.000000000
## Percentage.Employees.Female   0.89797049  0.93405758  0.94734516  0.912277376
##                              Percentage.Employees.Female
## Report.Year                                   -0.01464071
## Mean.Hourly.Gap                                0.06285680
## Median.Hourly.Gap                             -0.10626813
## Q1.Female                                      0.89797049
## Q2.Female                                      0.93405758
## Q3.Female                                      0.94734516
## Q4.Female                                      0.91227738
## Percentage.Employees.Female                    1.00000000
```

```
pairs(new_table)
```



Judging from the correlation table and plot, it seems like the pay gaps (mean and median) are somewhat positively correlated with each other, but they don't seem to correlate with the gender composition.

On the other hand, gender composition variables are positively correlated with each others but not correlated with the pay gap variables.

One thing to note that since the overall Percentage Female Employees are highly correlated with the quartile

compositions and that the percentage can be directly calculated by combining the quartiles, we believe that it would be redundant to include the variable in our model, so we dropped it.

Variable selection

After all the exploration and discussion, we have selected,for our initial model, **Mean Hourly Gap** as our response variable (see reasons above) and **Report Year** (to examine if there is an overall trend in pay gaps over the years) and **Q1 - Q4 Female percentages** as our predictor variables.

Results and Interpretation

```
full_model <- lm(Mean.Hourly.Gap ~ Report.Year+Q1.Female+Q2.Female+Q3.Female+Q4.Female, data = data)
summary(full_model)
```

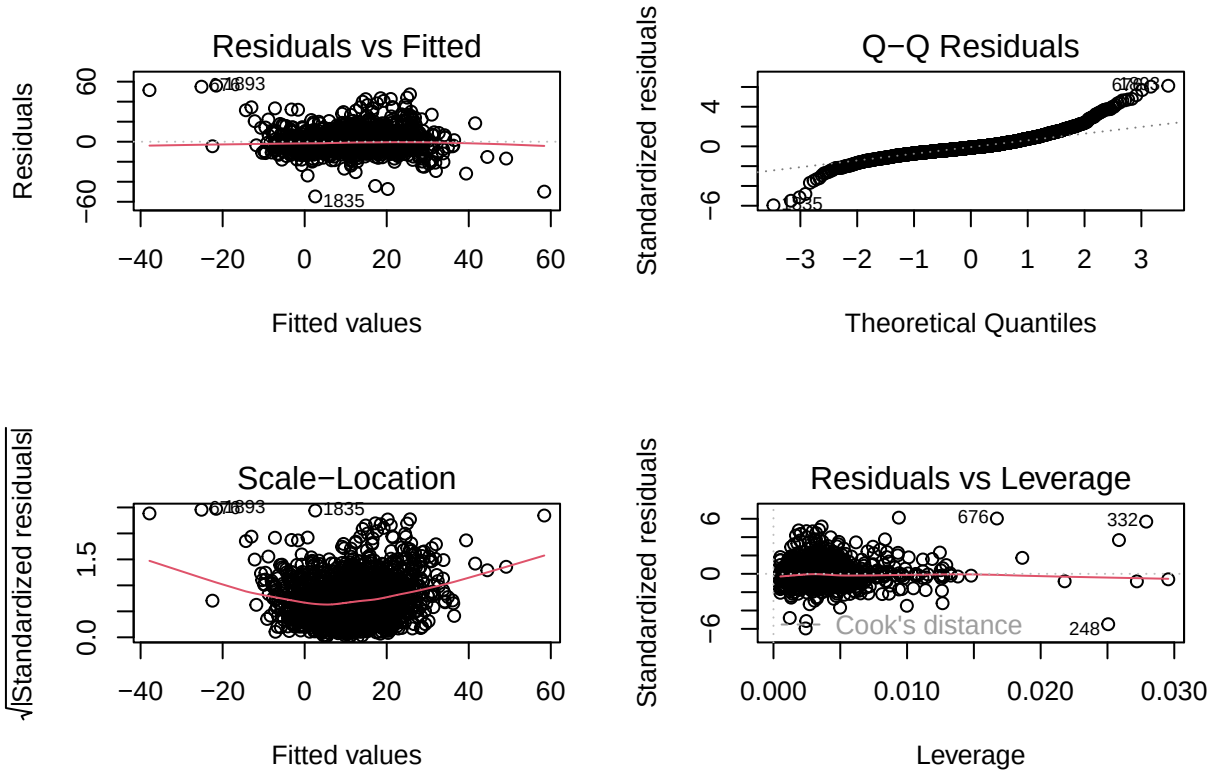
```
##
## Call:
## lm(formula = Mean.Hourly.Gap ~ Report.Year + Q1.Female + Q2.Female +
##     Q3.Female + Q4.Female, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.679  -4.849  -1.194   3.552  55.987
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 179.19896  513.65412   0.349    0.727
## Report.Year  -0.08691    0.25390  -0.342    0.732
## Q1.Female     0.35007    0.02205  15.876  < 2e-16 ***
## Q2.Female     0.25497    0.02248  11.342  < 2e-16 ***
## Q3.Female     0.15263    0.02455   6.218 6.16e-10 ***
## Q4.Female    -0.73109    0.02293 -31.883  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.192 on 1939 degrees of freedom
## Multiple R-squared:  0.4615, Adjusted R-squared:  0.4601
## F-statistic: 332.4 on 5 and 1939 DF,  p-value: < 2.2e-16
```

From the full model, we derive the equation:

$$Mean.Hourly.Gap = -0.087(Report.Year)+0.35(Q1.Female)+0.255(Q2.Female)+0.153(Q3.Female)+-0.731(Q4.Female$$

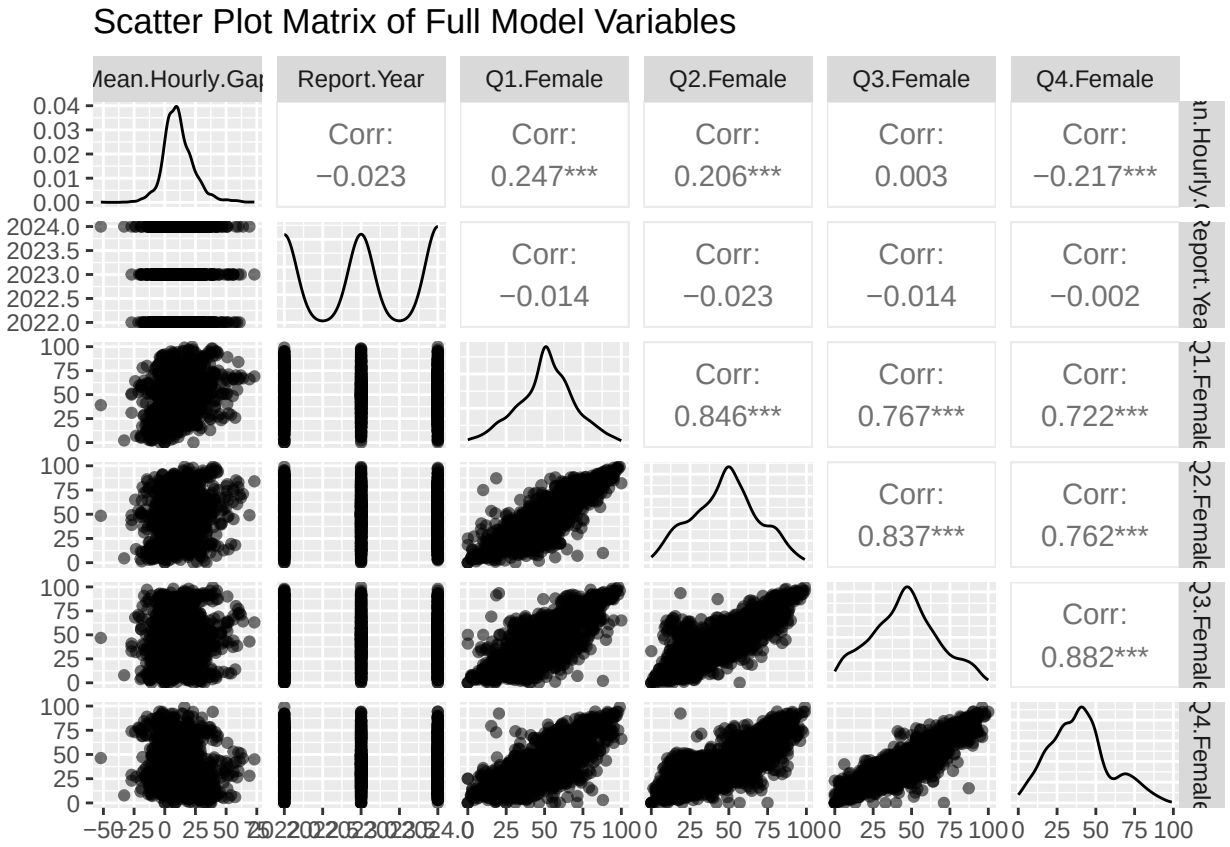
The summary indicates an R^2 value of 0.4615, meaning 46.15% of the variation in mean hourly gap can be explained by the variables: Report.Year, Q1.Female, Q2.Female, Q3.Female, Q4.Female The p-value is 2.2e-16 which is less than 0.05, giving significant evidence to reject the null hypothesis meaning of the predictors in the model is significant.

```
par(mfrow = c(2,2))
plot(full_model)
```



The Residuals vs Fitted plot shows the relationship is linear at mean 0. The Q-Q plot shows heavy tailing indicating deviation from normality. There also seems to be non-constant variance shown in the standardized residuals plot which violates the linear model assumption. Residuals vs Leverage doesn't show any outliers or leverage points that need inspection.

```
scatter_data <- data[, c("Mean.Hourly.Gap", "Report.Year", "Q1.Female", "Q2.Female", "Q3.Female", "Q4.F
ggpairs(scatter_data,
        title = "Scatter Plot Matrix of Full Model Variables",
        aes(alpha = 0.6))
```



The density plots on the diagonal show that Mean.Hourly.Gap is still right-skewed, reinforcing the need for a log transformation to improve normality. The correlation values confirm multicollinearity among Q1.Female, Q2.Female, Q3.Female, and Q4.Female, with correlations reaching 0.881, indicating that some of these predictors may be redundant and could inflate variance in the model. The scatter plots illustrate strong positive linear relationships between the quartile-based female representation variables, further supporting the presence of collinearity. Meanwhile, Report.Year exhibits weak correlations with all variables, implying that it likely does not play a significant role in predicting the gender pay gap.

```
# VIF
vif(full_model)  # If VIF > 5, consider removing redundant variables
```

```
## Report.Year   Q1.Female   Q2.Female   Q3.Female   Q4.Female
##    1.001207    3.720716    5.034822    6.351447    4.594843
```

The VIF indicate moderate multicollinearity among Q1.Female, Q2.Female, Q3.Female, and Q4.Female, with Q3.Female exceeding the threshold of 5, suggesting that some predictors may be redundant and could inflate standard errors.

Weighted Least Solutions

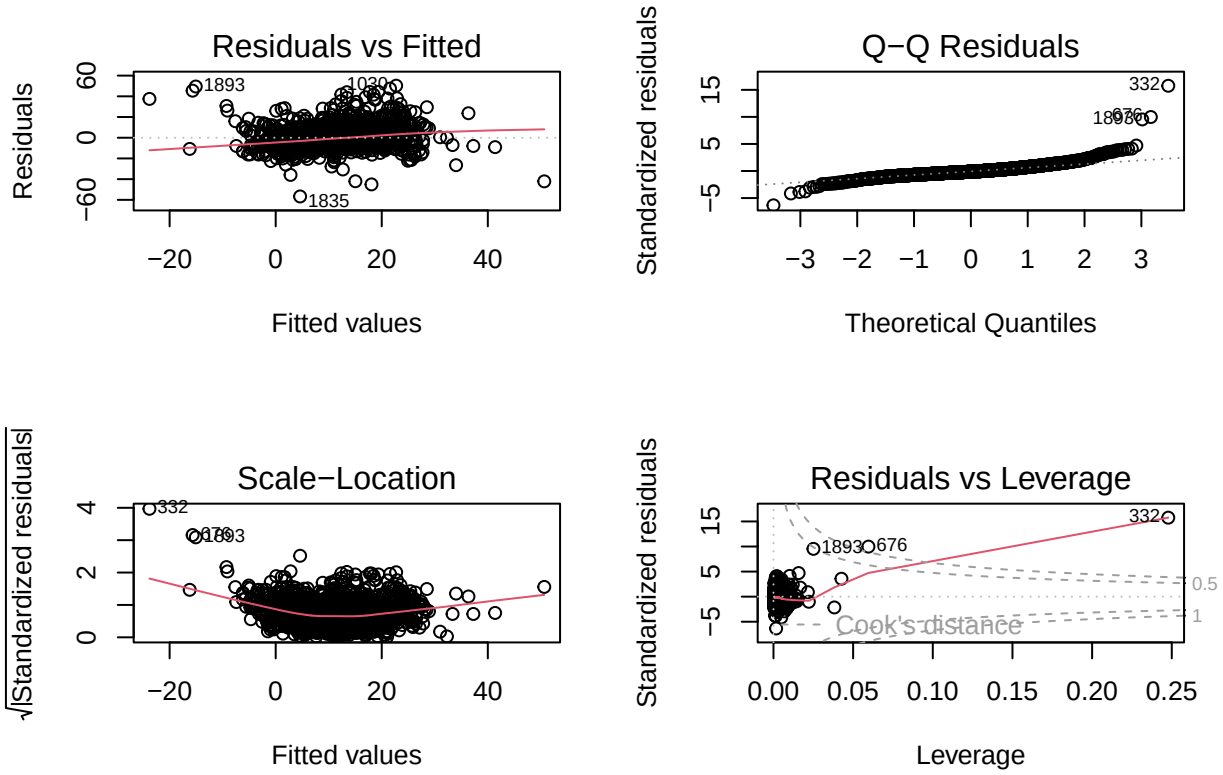
```
ols_model <- lm(Mean.Hourly.Gap ~ Q1.Female + Q2.Female + Q3.Female + Q4.Female, data = data)
residuals_ols <- abs(resid(ols_model))
variance_model <- lm(residuals_ols ~ fitted(ols_model))
```

```r
weights <- 1 / (fitted(variance_model)^2)

wls_model <- lm(Mean.Hourly.Gap ~ Q1.Female + Q2.Female + Q3.Female + Q4.Female,
                data = data, weights = weights)
summary(wls_model)
```

```
##
## Call:
## lm(formula = Mean.Hourly.Gap ~ Q1.Female + Q2.Female + Q3.Female +
##     Q4.Female, data = data, weights = weights)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3813  -0.8480  -0.2056   0.6448  22.3870
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.52814    0.63613   7.118 1.53e-12 ***
## Q1.Female    0.24045    0.02294  10.480  < 2e-16 ***
## Q2.Female    0.18945    0.02305   8.218 3.73e-16 ***
## Q3.Female    0.20759    0.02567   8.088 1.06e-15 ***
## Q4.Female   -0.60871    0.02439 -24.961  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.639 on 1940 degrees of freedom
## Multiple R-squared:  0.3458, Adjusted R-squared:  0.3444
## F-statistic: 256.3 on 4 and 1940 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2))
plot(wls_model)
```



The Weighted Least Squares (WLS) regression improved heteroscedasticity, stabilizing residual variance while keeping Q1-Q4 Female percentages statistically significant, confirming their impact on the gender pay gap. However, Report.Year remains insignificant (p = 0.39008), suggesting no significant time trend in the pay gap. Observation 332 still has high leverage, along with 189 and 676, indicating strong influence on the model.