
Egypt University of Informatics
Computing and Information Sciences
Data Analysis Course

# A Study on the Gender Pay Gap in Egypt's Tech Market

Supervised by:
Dr. Mohamed Taher Alrefaie

Submitted by:
Mohamed Elmosalamy 23-101283
Omar Mohammad 23-101288
Omar Hesham 23-101302
Mohamed Kotb 24-101222

2025-05-24

# Contents

# 1 Abstract

The gender pay gap remains a contentious and widely debated issue worldwide. This analysis investigates whether women earn less than their peers in **Egypt's tech industry**.

The study aims to quantify salary differences between genders using publicly available data, offering insights into whether a significant gender-based wage gap exists.

# 2 Introduction

The gender pay gap has been a global issue across labor markets. While considerable research has been done in Western countries, limited data-driven analysis exists in the Middle East. In Egypt, this remains a largely unstudied topic.

To address this, we define three methods of investigation:

1. **A straightforward assessment of the gender pay gap**

   The first hypothesis tests for the presence of a gender pay gap in Egypt's tech industry by conducting a **Welch's independent t-test** on male and female salaries. This analysis aims to answer: **"Is there a pay gap between men and women in Egypt's tech industry?"**

2. **Assessing the gender pay gap after controlling for all contributing factors**

   This second analysis aims to control for years of experience between the two genders, and additionally, check for any pay gap per bracket. This analysis will use **Blinder–Oaxaca decomposition** to determine the pay gap per each experience bracket.

3. **Estimating The Cost of Being a Woman**

   *The Cost of Being a Woman* is defined as the monthly salary disparity a woman faces compared to a man with identical skills, title, and experience. We aim to provide a 95% confidence interval for this cost.

# 3 Research Questions

1. Is there a statistically significant difference in mean salaries between genders?

2. After controlling for years of experience and other contributing factors, does the pay gap persist/vanish?
3. What is the cost of being a woman—how much does a woman gain or lose per month compared to an equally qualified man?

# 4 Hypothesis

## 4.1 Difference in mean salaries

- Null hypothesis $(H_o)$: There is no significant pay gap between men and women.
- Alternative hypothesis $(H_\alpha)$: There is a significant pay gap between men and women.
- Significance level $(\alpha)$: 0.05

## 4.2 Controlled difference in mean salaries

- Null hypothesis $(H_o)$: After controlling for contributing factors (experience, title, level, etc.), there is no significant gender pay gap.
- Alternative hypothesis $(H_\alpha)$: A significant pay gap persists even after controlling for these factors.
- Significance level $(\alpha)$: 0.05

# 5 Population of Interest

All professionals working in Egypt's tech field.

# 6 Dataset

The dataset used comes from the Egyptian Tech Market Survey API, conducted in 2024.

**Dataset columns include:**

- **Gender:** Male, Female
- **Degree:** Bachelor's degree (Yes, No)
- **Title:** Professional title (e.g., Data Analyst, Scrum Master)
- **Level:** Professional level (e.g., Junior, Senior, Team Lead)
- **YearsOfExperience:** Number of years in the tech field
- **Salary:** Monthly salary in EGP
- **IsEgp:** Currency used (EGP, foreign, hybrid)
- **ProgrammingLanguages:** Languages the subject can write
- **BusinessMarket:** Scope (Local, Regional, Global)
- **BusinessSize:** Company size (Start-up, SME, Large)

- **WorkSetting:** Working environment (Office, Remote, etc.)
- **CompanyLocation:** City/state of the company

Sample size: **2649**

# 7 Analysis

For the rest of our analysis, we are going to split the data into two parts. A part where the salary is in EGP, and another where the salary is in any other currency.

*However, we are going to be mainly focusing on **the salaries in EGP** during this study.*

## 7.1 Five-number summary

Firstly, we want to do a Five-number summary for our quantitative variables in order to work out the **IQR**, and do our **outlier analysis**.

| | Salary (EGP) | Years of Experience |
|---|---|---|
| Minimum | 12.00 | 0.0 |
| Q1 | 15,000.00 | 1.00 |
| Median (Q2) | 24,000.00 | 2.00 |
| Q3 | 40,000.00 | 4.00 |
| Maximum | 330,000.00 | 31.00 |

From that summary, we can see that the middle 50% of the salaries lies **between 15,000.00 EGP and 40,000.00 EGP**, and the middle 50% of the years of experience lies **between 1 and 4 years**.

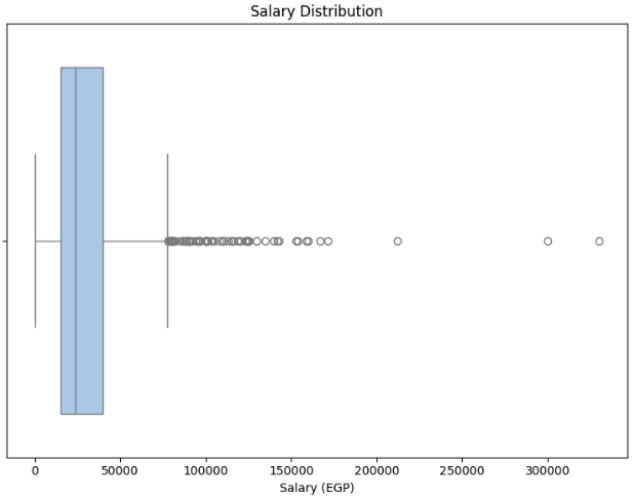



Figure 1: A box plot of Salary in EGP. It shows a slight right-skewness, in addition to a big number of outliers.

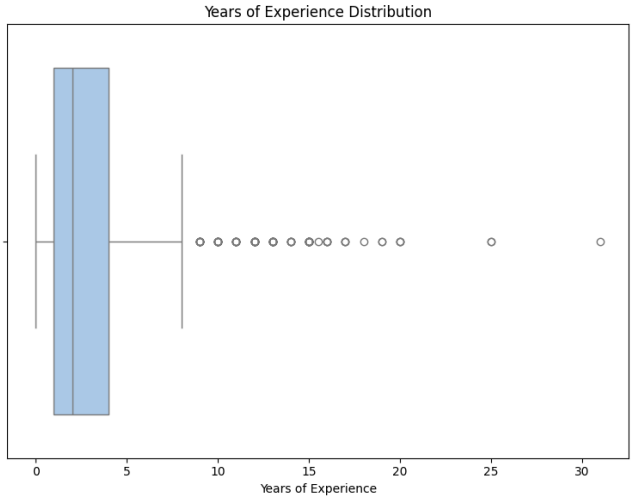



Figure 2: A box plot of Years of Experience. It shows a slight right-skewness as well, and a significantly smaller number of outliers.

## 7.2 Outlier Analysis

From the box plots and the five-number summary, it is evident that there are several outliers in both the Salary and Years of Experience distributions. However, the severity and frequency of these outliers differ:

- **Salary (EGP)** shows a significant number of outliers on the higher end, confirming the right-skewness observed in the box plot. These are likely due to a small number of professionals earning exceptionally high salaries, possibly in senior or specialized roles. To ensure our analysis remains representative and interpretable, salaries above the 99th percentile were removed.
- **Years of Experience**, on the other hand, exhibits far fewer outliers. The majority of participants fall within a narrow range (1–4 years), with a few outliers representing professionals with notably long careers (up to 31 years).

These outliers, particularly in the Salary variable, can distort statistical measures like the mean and standard deviation. By identifying and handling them appropriately, the analysis focuses on the central trends that are more reflective of the general population.

## 7.3 IQR Visualizations

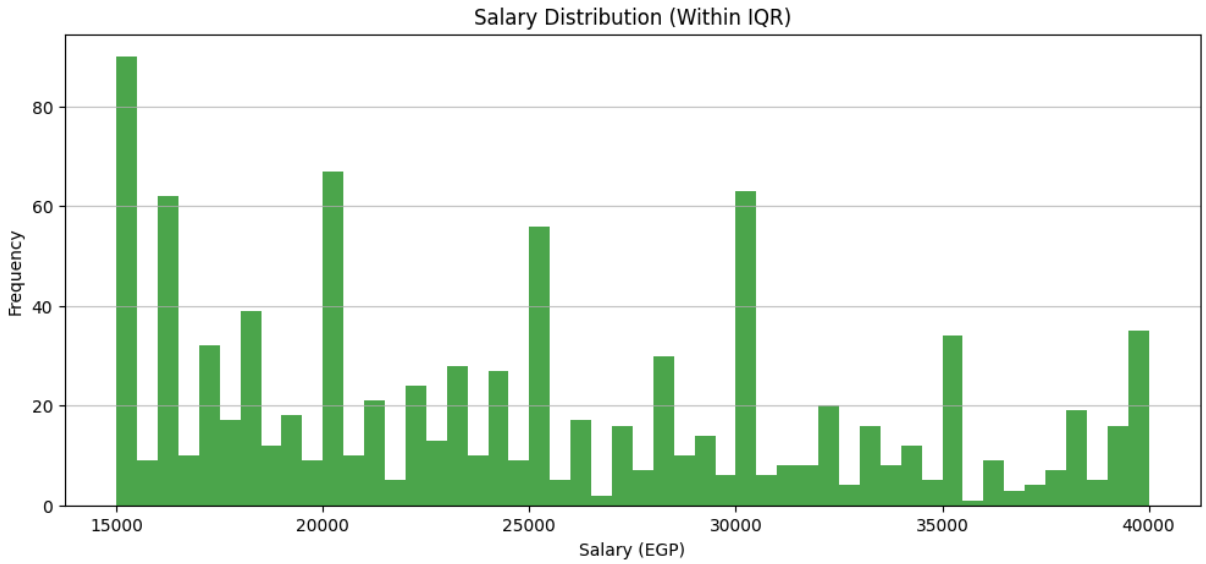



Figure 3: A histogram of the IQR of the salary column reveals a multi-modal representation, which reflects the existence of cofactor variables like years of experience, job level, etc.

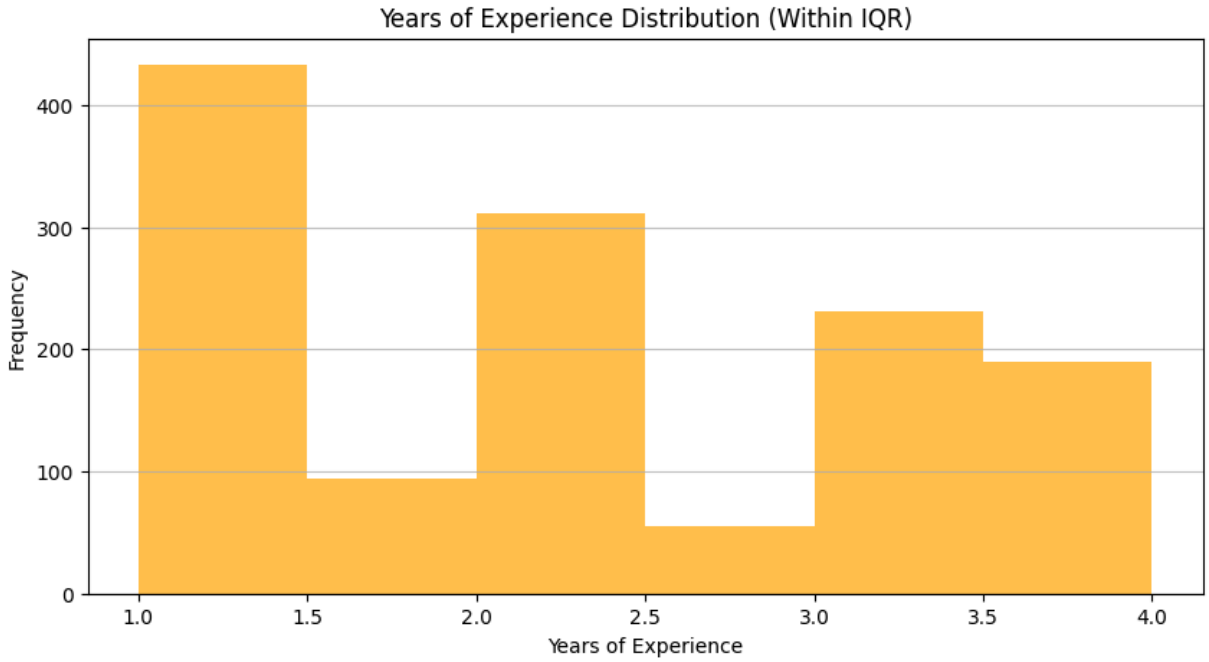



Figure 4: A histogram of the IQR of the years of experience column reveals a multi-modal representation as well.

## 7.4 IQR Gender Visualizations

To have a better understanding of the difference in Salaries and Years of Experience between males and females, we represent them using side-by-side histograms.

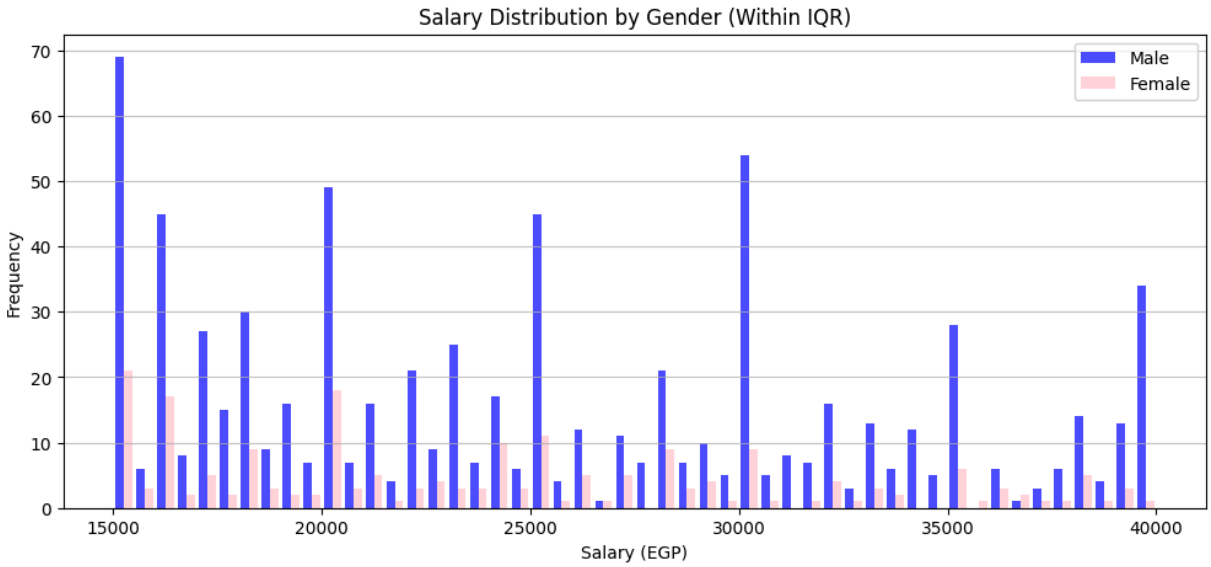



Figure 5: A histogram of the IQR of the salary per gender. It is observed that men earn higher salaries .

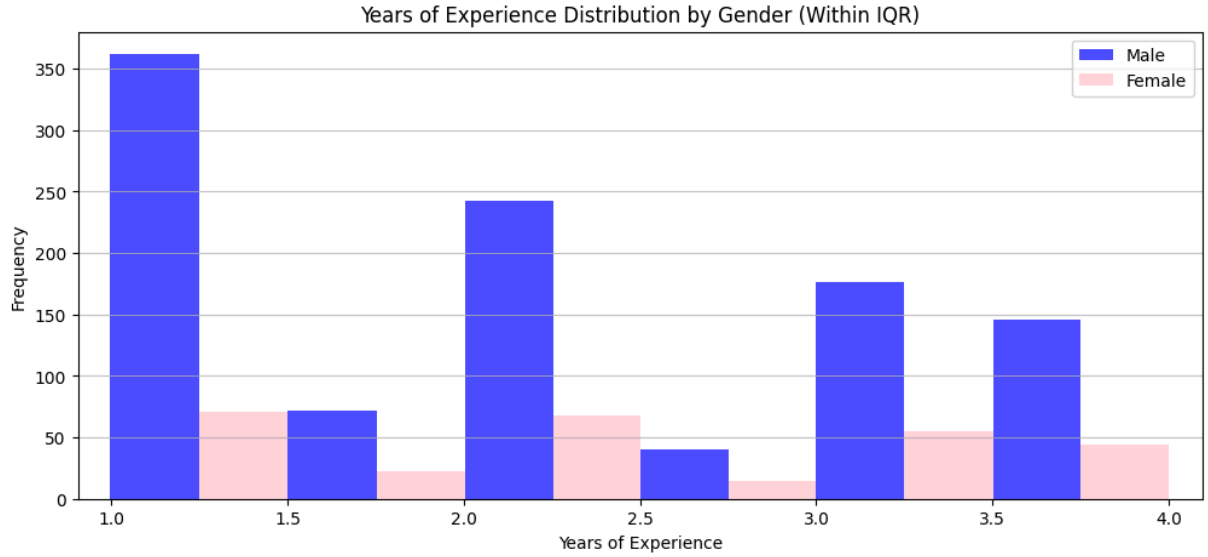



Figure 6: A histogram of the IQR of the years of experience per gender, women seem to be evenly spread across all years of experience, while men seem to be present in the 1-2 years of experience bracket more than the rest.

## 7.5 Gender Distribution

An important aspect of our study is **the number of females and males present in our dataset**, knowing that count will help us determine whether this sample is representative of the population or not.

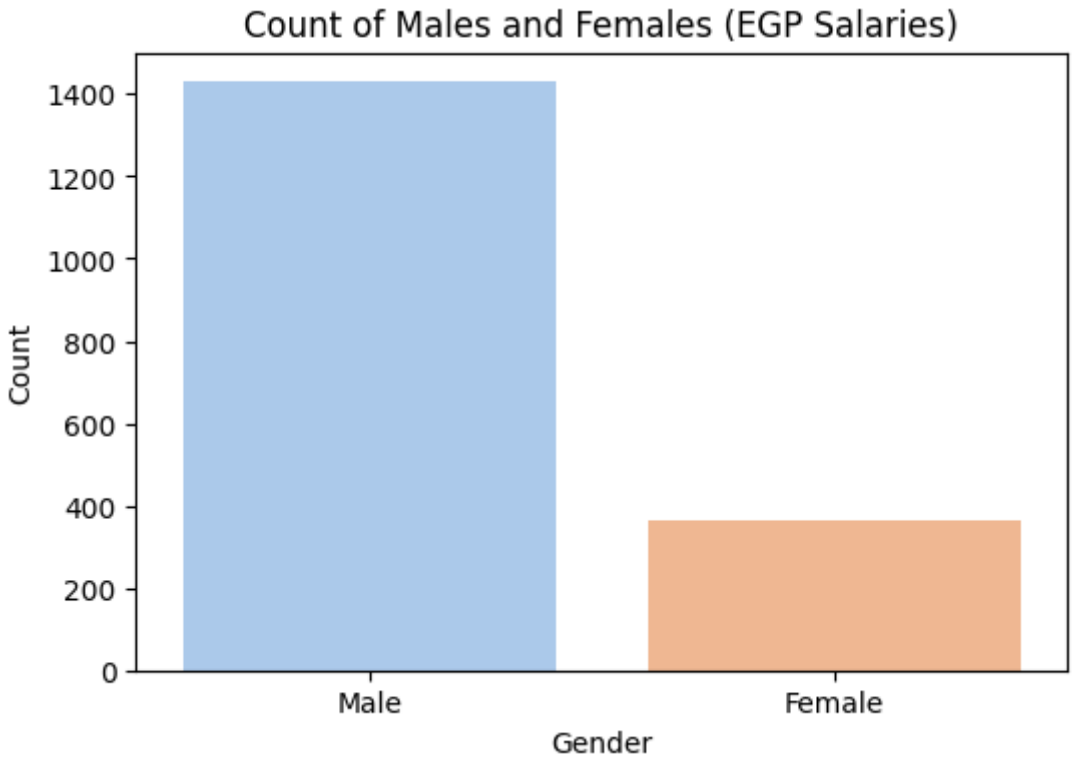



Figure 7: A histogram of the count of Males and Females whose Salaries are in EGP (The subset of the market we are studying). It is noticed that the **count of Males is almost 4 times more than that of Females**.

The ratio between working men and women in the tech industry is reflected in the diagram. as according to this report, **women represent 20% of the tech and engineering industry**.

## 7.6 Correlation between Salary and Years of Experience

To find the correlation between two quantitative variables, salary and years of experience, we are going to use **Pearson's Correlation**, but first we need to check if the two variables observe a linear relationship or not.

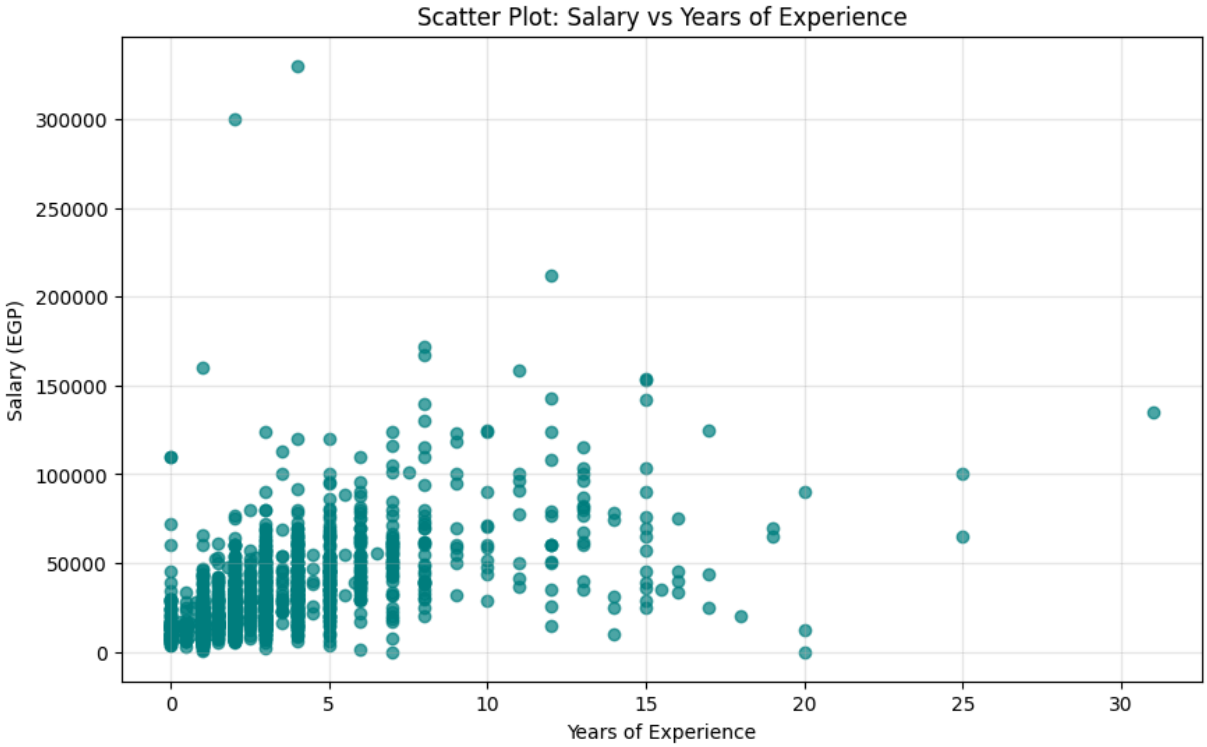



Figure 8: A scatter plot of salary and years of experience. It is noticed that there is a slight positive linear relationship between them.

Then, calculating the **Pearson correlation coefficient** yields:

$$r = 0.55$$

Which means there is **moderate positive linear correlation** between Salary and Years of Experience.

## 7.7 Salaries per Job Level

Next, we want to check the Salaries per each Job Level in our dataset.

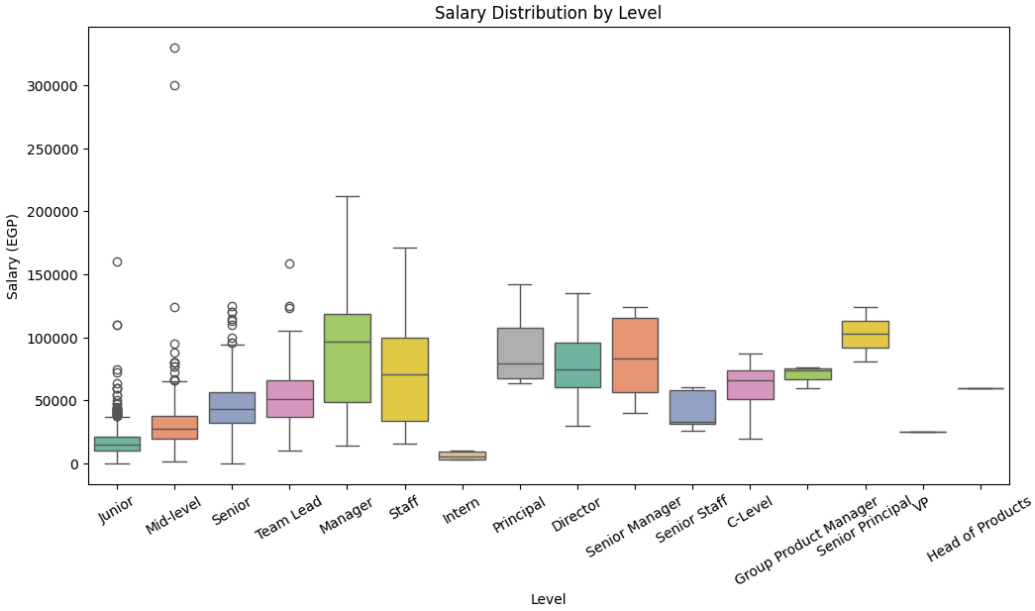



Figure 9: A series of box plots for each Job Title and the average Salary associated with it. The data seems to be following a normal trend where Management Jobs have the highest salaries, while Interns are paid the lowest.

## 7.8 Salaries per Job Level and Gender

Now let's check how these box plots look when taking gender into consideration.

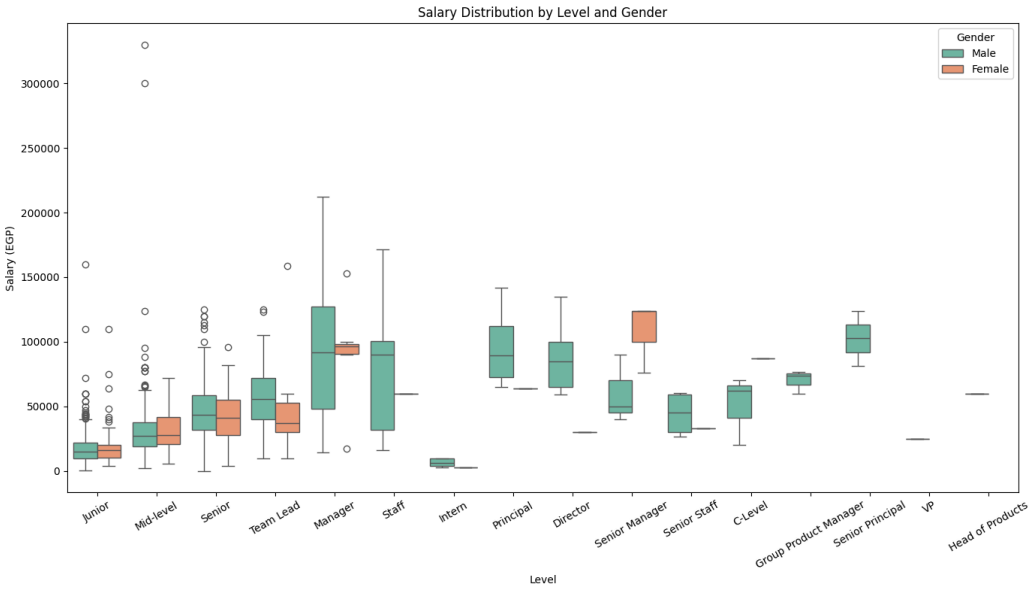



Figure 10: Box plots for each Gender's Job Title and the average Salary associated with it. It is observed that in the majority of Job Levels, men earn a more rewarding Salary than women.

## 7.9 Top 10 Job Titles with Highest Average Salaries by Gender

Lastly, we want to check what are the top 10 most paying jobs in the dataset, and how different is the average pay of these jobs per gender.

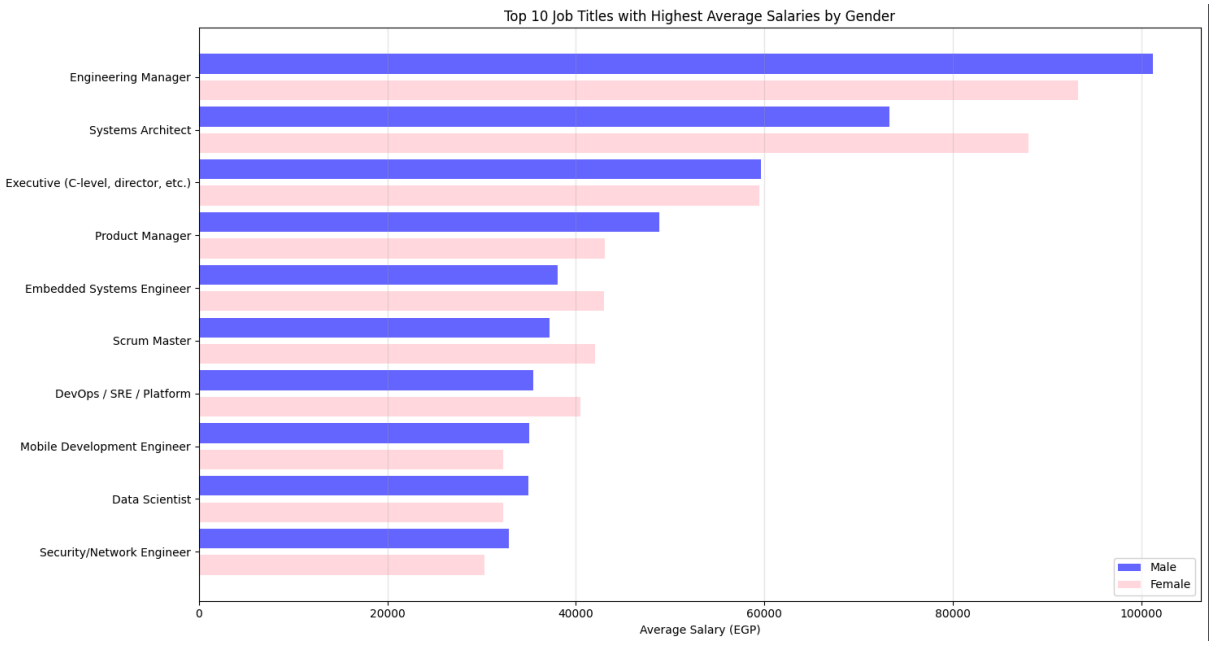



Figure 11: A series of bar plots for top 10 most paying jobs in the dataset, and the average Salary associated with them for each gender.

Although men seem to have a bigger average salary in the top paying job across the dataset (Engineering Manager), women secure that spot in the second most paying job which is Systems Architect, as well as a few other jobs like Embedded Systems Engineer, Scrum Master, and DevOps in which women seem to have the bigger average salary.

# 8 Hypothesis Testing Steps

## 8.1 Difference in mean salaries

We used **Welch's Independent T-Test**:

1. Hypotheses defined (see Section 4.1)
2. Significance level set to 0.05
3. Data cleaned and prepared
4. Used `scipy.stats.ttest_ind(equal_var=False)`
5. Decision made based on resulting p-value vs. alpha

## 8.2 Controlled difference in mean salaries

For this, we applied the **Blinder–Oaxaca decomposition**:

1. Hypotheses defined (see Section 4.2)
2. Significance level set to 0.05
3. Data cleaned and outcome/explanatory variables defined
4. Ran group regressions, decomposed results using `statsmodels`
5. Interpreted contributions of each factor (explained/unexplained)

## 8.3 Cost of Being a Woman

Cost is defined as:

$$\text{Cost} := \text{Expected Salary based on objective factors} - \text{Actual Salary}$$

**Objective factors used:**

- Years of Experience
- Title
- Level

We trained a regression model on male data using all variables except gender, then applied it to female employees to estimate expected salaries. This allowed us to construct a 95% confidence interval around the cost of being a woman.

# 9 Conclusions

## 9.1 Difference in mean salaries

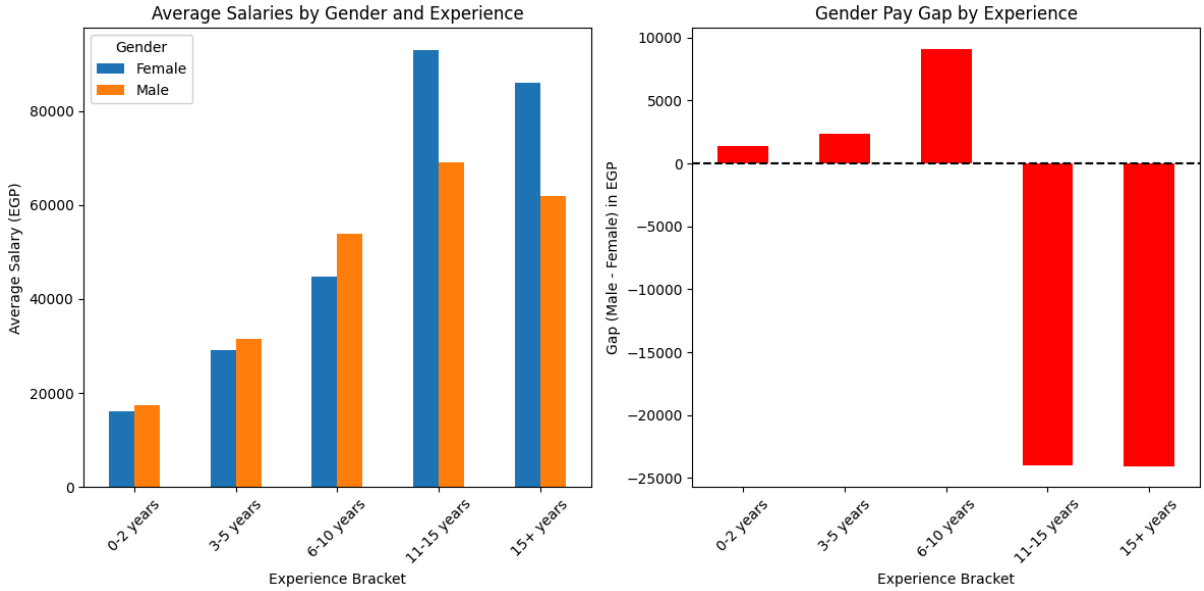
**P-value:** 0.2668 — It's above the 0.05 threshold. We fail to reject the null hypothesis. No significant evidence of a gender-based pay gap exists.

To further verify that conclusion, we run a **Practical Significance** test using **Cohen's d effect size**. The results were:

$$d = 0.0592$$

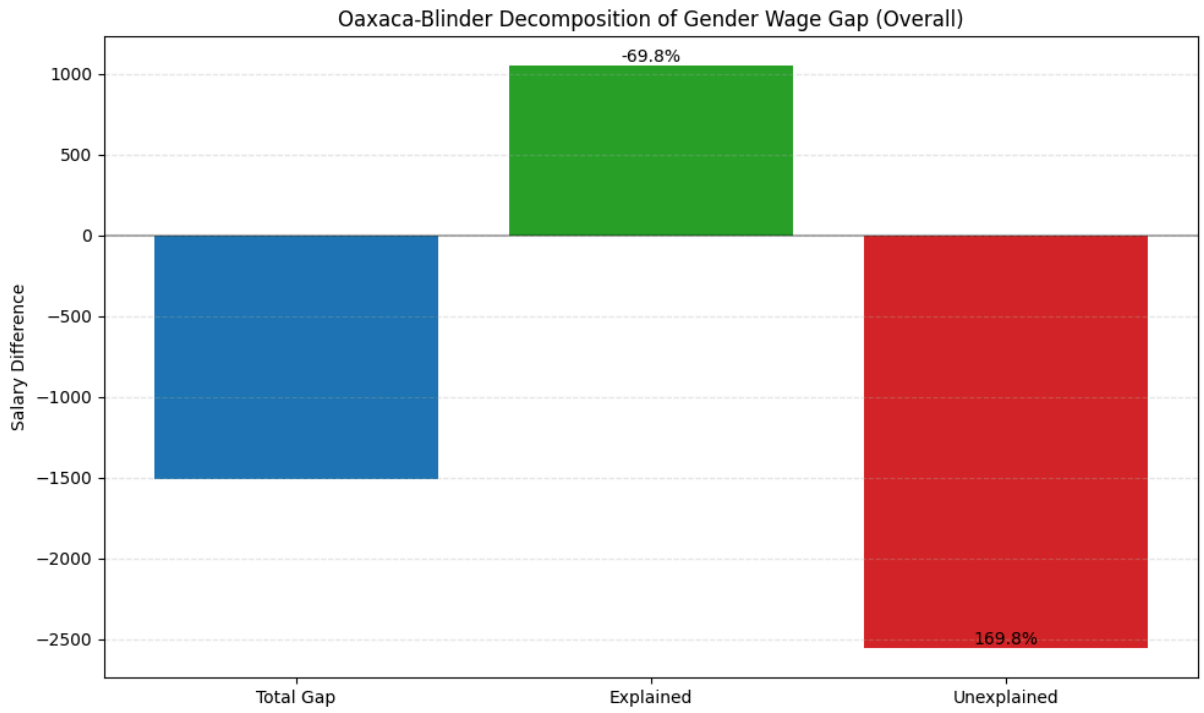
This effect size means that the gender pay gap has a negligible practical significance.

## 9.2 Controlled difference in mean salaries




**Insights:**

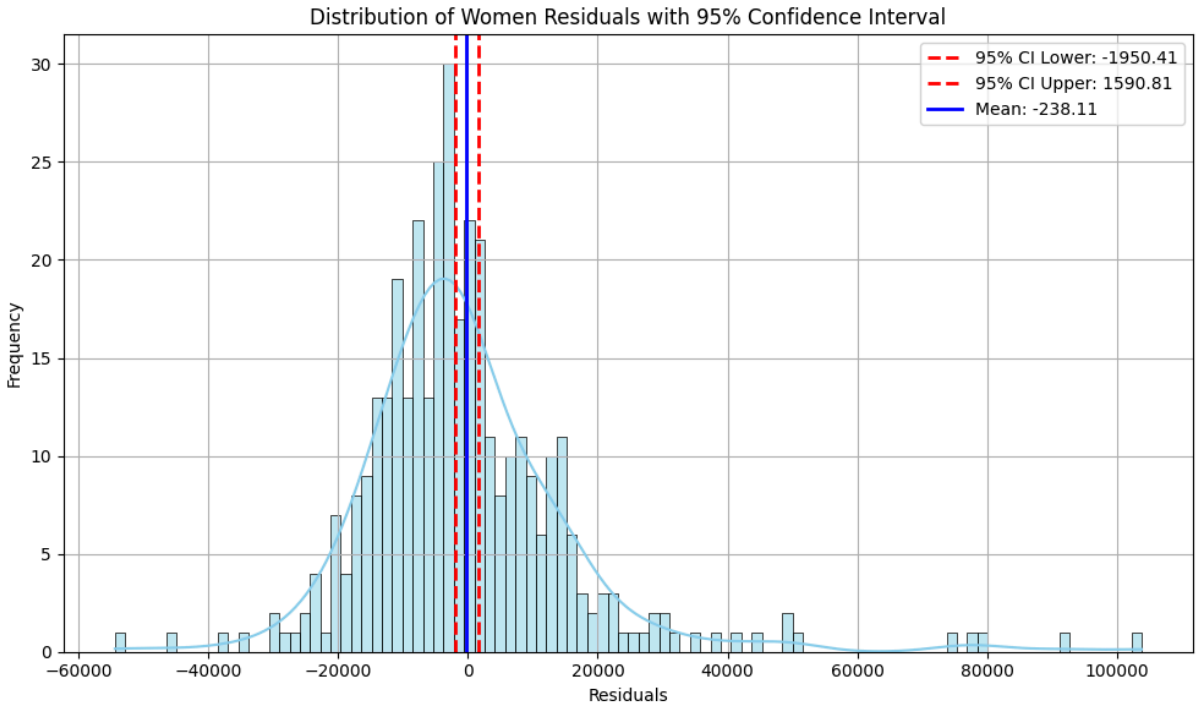
- Average salaries rise with experience for both genders
- Men earn slightly more in the first 10 years
- After 10 years, women's average salaries exceed those of men




To further investigate the sources of this weird behavior, we performed a Oaxaca-Blinder decomposition. However, due to the relatively small sample size, the decomposition yielded unstable and inconclusive estimates regarding each portion. These results should not be interpreted as strong

evidence but rather as exploratory insights that warrant further investigation with larger datasets.

## 9.3 The Cost of Being a Woman




The 95% confidence interval ranges from **−1950.41 to 1590.81 EGP**. Since zero lies within this interval, it suggests that the "cost of being a woman" may, statistically, be zero.

# 10 Any Potential Issues

The dataset included a big amount of very extreme outliers that significantly altered our results and distorted our visualizations. We dealt with that by trimming the largest 1% Salaries, in other words, we used the 99th percentile of the dataset in our entire analysis. While that didn't ultimately get rid of all outliers, but it helped reduce their effect a bit.

Another issue we faced, was that scarcity of female records in the dataset. This limited us from reaching a clear consensus regarding the **Blinder–Oaxaca decomposition**, as it required more female records to yield accurate results. However, we backed that off with a regression analysis that supported the results of the **Blinder–Oaxaca decomposition** test.