Gender Gap Analysis in Workplace

Group Number: 9

Group Members:

- Vaibhav Bothra (22116101) (Course Code: 46)
- V.E. Sudarsan (22116098) (Course Code: 43)
- Vaibhav Prajapati (22116102) (Course Code: 47)
- Yash Gehlot (22116109) (Course Code: 50)
- Kuldeep (22116048) (Course Code: 26)

1. **Aim:**
   To analyze the gender gap by modelling US National data consisting of different fields from the period 1981 to 2013.

2. **Motivation:**
   The gender gap continues to be a significant issue on a global scale. To gain a deeper understanding of the gap, it is essential to analyze the qualitative aspects of the workplace. By exploring the data of varied domains, valuable insights can be revealed like how men and women get pay, the proportions of men and women in different segments.
   While much of the focus has traditionally been on comparing the average salaries of men and women, but a deeper analysis of factors such as education level, working hours is also important to understand the gender-based disparities.
   The report outlines the key findings from the analysis of dataset and the goal is to identify the underlying factors that contribute to the gender pay gap, using both quantitative and qualitative approaches to gain valuable insights.

3. **Dataset:**
   a. The dataset that has been used is:
   "Gender Pay Gap Dataset" (from Kaggle)

   Description:
   Time Period: 1981 to 2013
   Important Columns that are considered:
   i) Year: Survey year
   ii) Sex: Male (1) and Female (2)
   iii) Age
   iv) Marst: Marital Status (Married, spouse present=1, Married, spouse absent=2, Separated=3, Divorced=4, Widowed=5, Never mar=6)
   v) Educ99: Educational attainment, (No school=1, 1st-4th grade=4, 5th-8th grade=5, 9th grade=6, 10th grade=7, 11th grade=8, 12th grade=9, High school=10, Some college=11, Associate=13, Associate=14, Bachelors=15, Masters d=16, Profession=17, Doctorate=18)

vi) Occ: Occupation
vii) Wkswork1: Weeks worked last year
viii) Uhrswork: Usual hours worked per week (last year)
ix) Incwage: Wage and Salary Income

4. Methodology:

It is designed to offer a comprehensive approach to understanding the various factors. It combines quantitative statistical techniques with visualization to provide a holistic view of the gender pay gap, based on variables such as gender, education level, marital status and working hours.

a) Data Collection and Preprocessing:

The data was collected from the Kaggle Dataset, and it was cleaned to remove outliers and missing values to ensure that the analysis would be based on complete, accurate information. Important features, like education, working hours, age, gender, marital status, salary income, were only considered to align with the objectives.

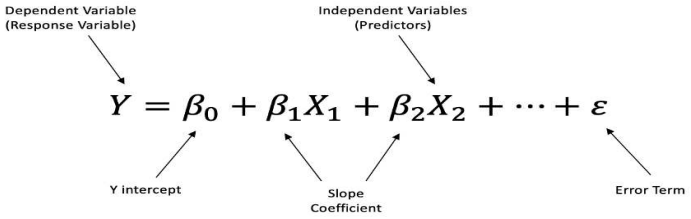
	year	sex	age	marst	educ99	occ	wkswork1	uhrswork	incwage
1	2009	Male	28	6	10.0	5120	52	40	17680.0
5	1999	Male	37	1	11.0	424	52	40	42000.0
6	2007	Male	44	1	15.0	7750	52	80	33000.0
8	1999	Male	41	1	10.0	308	52	40	30000.0
9	2011	Male	55	1	10.0	9620	52	40	67000.0
...	...	...	...	...	...	...	...	...	...
344280	1999	Female	34	1	10.0	337	52	40	17000.0
344281	2013	Female	45	3	11.0	7800	52	32	23000.0
344283	1999	Female	27	1	10.0	13	9	40	3200.0
344285	2007	Female	49	1	18.0	2860	25	10	8800.0
344286	2013	Female	36	1	11.0	4760	52	24	15000.0

256875 rows × 9 columns

b) Regression Analysis:

It was employed to define the relationship between gender and income based on factors such as education, marital status, hours worked per week, weeks worked last year.

Linear Regression: The multiple linear regression model has been used as income was dependent variable and other factors listed above were independent variables to calculate R^2 for knowing how better the fit is.



Logistic regression: It has also been used to get the data fit with the same factors, but the gender was dependent while others were independent, as it could take only two values, that is, 0 for “Male” and 1 for “Female” as Logistic regression also gives only two values either 0 or 1.

$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

where usually $b = e$.

The regression analysis was used to calculate the coefficients that reveal how income is influenced by each of these factors.

c) **Hypothesis Testing:**

It was conducted to assess whether the observed differences in income between men and women are statistically significant. The main motive was to evaluate if gender influences income across various factors like education level, marital status etc. T-tests were performed to compare the mean income between men and women within specific groups like education, working hours etc. assuming the variance of both the incomes were different.

Null Hypothesis:
No significant difference in income between males and females.
Alternate Hypothesis:
Significant difference in income between males and females.

A confidence level of 95% (**alpha=0.05**) was used to determine whether gender-based disparities were statistically significant.

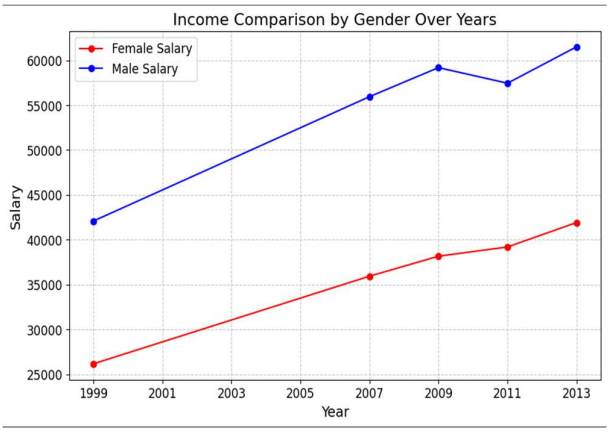
d) **Principal Component Analysis (PCA):**

It was used to reduce the dimensionality of the dataset and identify the primary components contributing to income variations. It allows us to mark the most important features to explain most of the variance in income data.
It transformed the data into components, sorted by the amount of the variance they incorporate. This helps to isolate and assess how income and other factors are correlated.

5. Observations and Results:

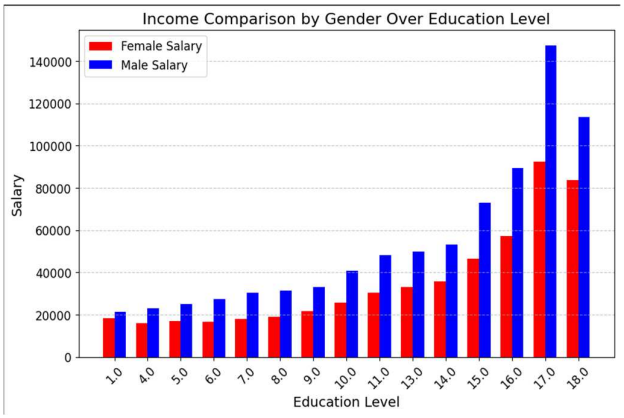
a) Visualization of Income Disparities by Gender Across Key Factors:

i) Income Comparison by Gender over Years



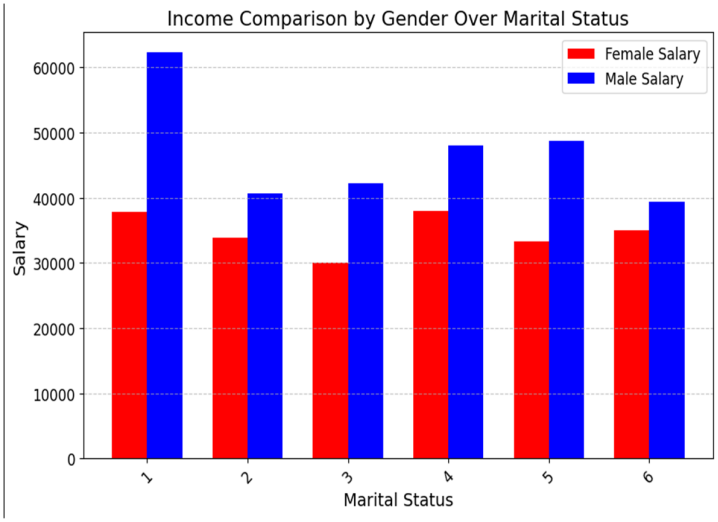
The increase in wages for both genders might reflect periods of economic growth and possibly inflation. However, benefits of this growth are not equally distributed between the genders. The graph indicates a persistent gender pay gap that does not appear to narrow significantly over the period.

ii) Income Comparison by Gender over Education Level



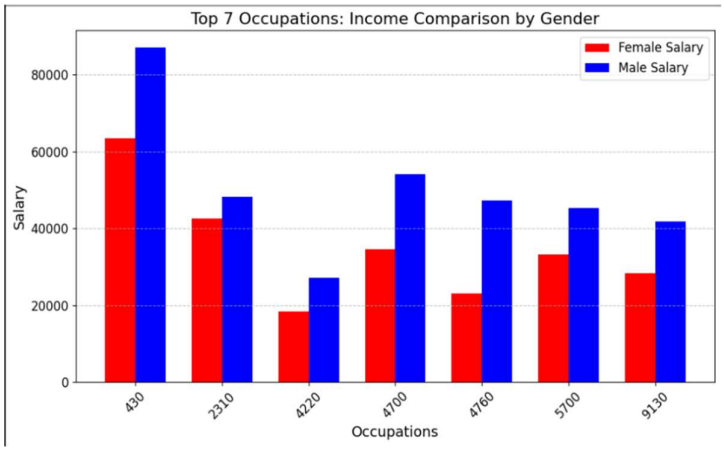
It shows that higher education generally correlates with higher income for both genders. However, it also suggests that the gender pay gap persists across different educational achievements.

iii) Income Comparison by Gender over Marital Status



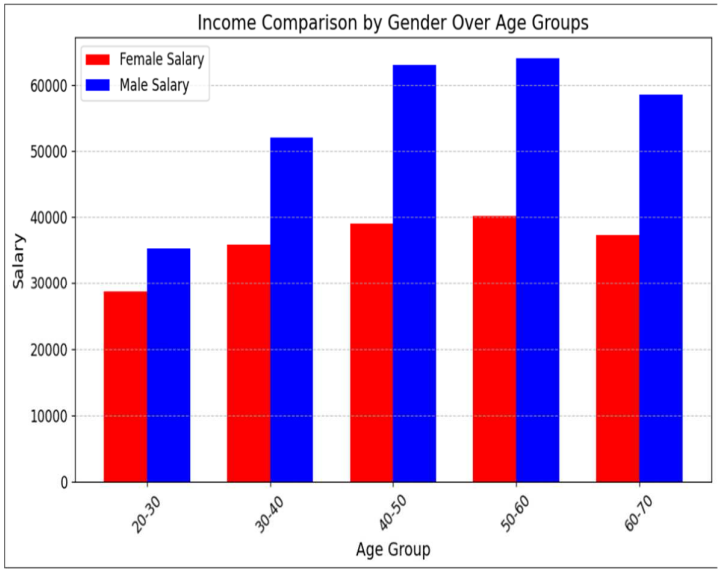
The bar chart shows that men earn higher salaries than women across all marital statuses, with the largest gap in status 1 where males earn over 60,000. The gap narrows for statuses 3 to 6, while female salaries remain relatively stable with less variation. These disparities emphasize the need for policies to address gender-based income inequality.

iv) Income Comparison by Gender over Top 7 Occupations



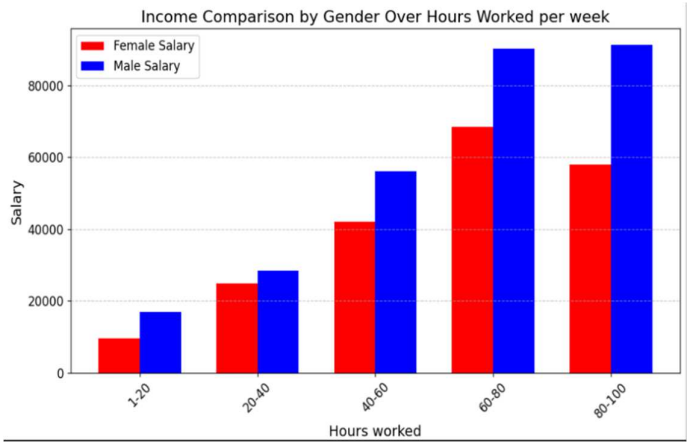
For every occupation, there is a visible pay gap between the two genders. This gap varies in size across different occupations, with some showing a relatively small difference and others showing a substantial disparity.

v) Income Comparison by Gender over Age Groups



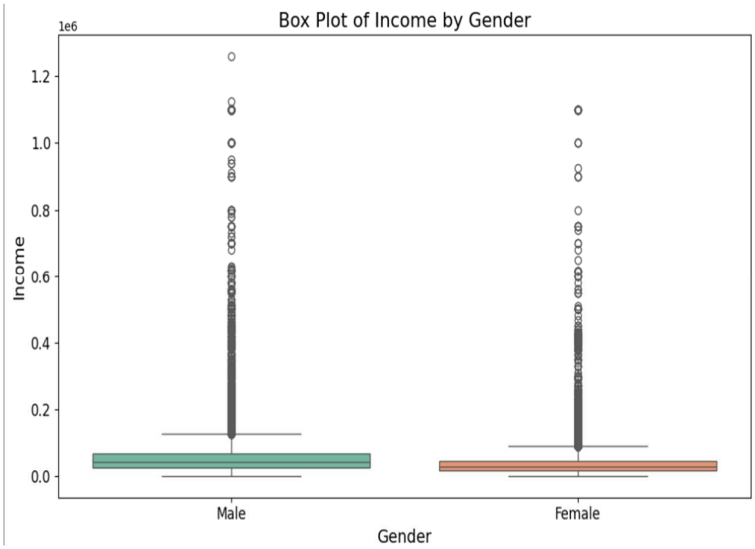
The bar chart shows men consistently earning more than women across all age groups. Salaries peak at 50–60 before declining at 60–70. The smallest gap is in the 20–30 group, widening significantly by 30–50. Female salary growth plateaus after 40–50, while males continue to rise, highlighting persistent gender income disparities and the need for workplace equality policies.

vi) Income Comparison by Gender over Hours Worked per Week



The graph shows that while incomes rise with weekly hours, men consistently earn more. The gap is smaller at 1-40 hours but widens significantly at 60-100 hours, indicating men benefit more from high-hour work categories, highlighting persistent pay disparities.

vii) Box Plot of Income by Gender

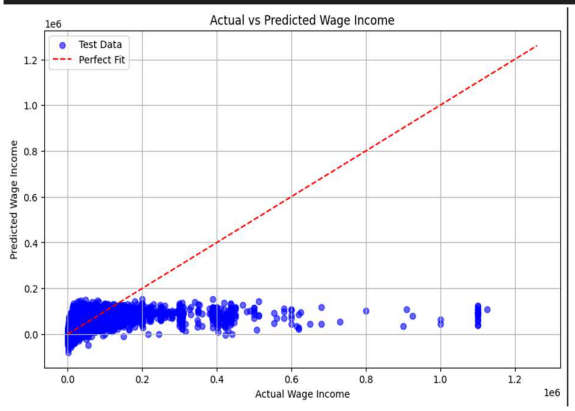


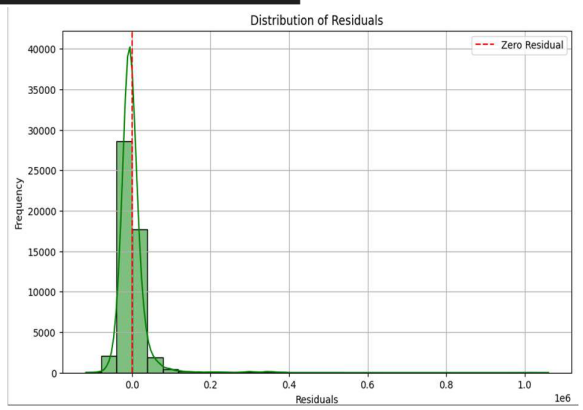
The box plot shows men have a higher median income than women, with both genders displaying a similar interquartile range (IQR). However, men have more high-income outliers, indicating a broader income range at the top and a slight overall wage discrepancy favouring men.

b) Regression Analysis:

i) Linear Regression

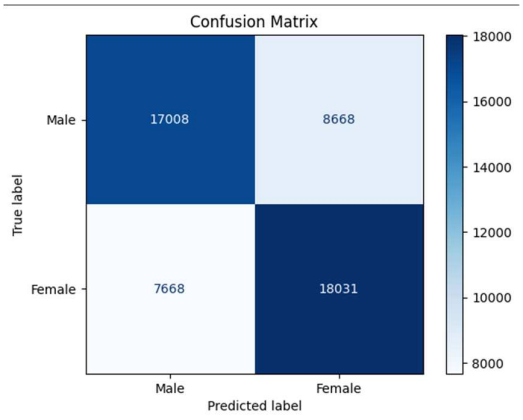
```
Linear Regression Results:
Coefficients: [ 7462.90368887 -7777.38561083  4362.64196014 -3056.18763971
 11978.39633704 -6988.62419165  4926.81967616 11366.47041901]
Intercept: 46548.64976155731
Train R-Squared: 0.25122186778869626
Test R-Squared: 0.260632860285852
```

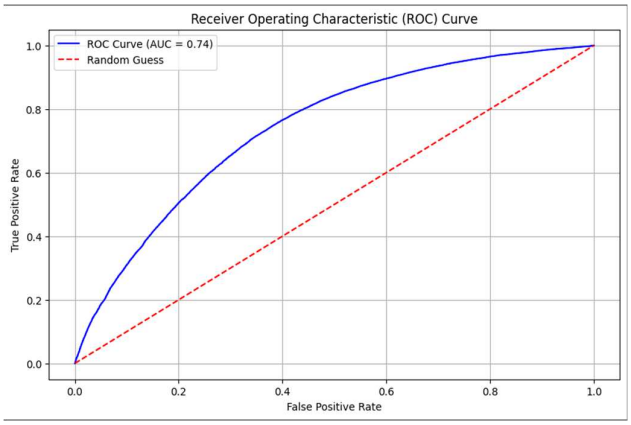




The actual vs. predicted income graph shows that predictions are more accurate for lower incomes, while accuracy declines for higher incomes due to the presence of outliers and non-linear patterns. Residuals are generally centred around zero but exhibit a long tail, indicating significant prediction errors at higher income levels. These issues highlight the limitations of linear regression, as it oversimplifies complex relationships, leading to poor generalization for high-income predictions. To improve accuracy, especially for higher income levels, non-linear models could be explored to better capture the underlying patterns in the data.

ii) Logistic Regression






```
Logistic Regression Results:
Coefficients: [[ 0.25738673  0.13227291  0.09234548  0.24775215 -0.4741471   0.03332919
  -0.46953673 -0.87148976]]
Intercept: [-0.06249606]
Train Accuracy: 0.6753138686131387
Test Accuracy: 0.6820048661800486
```

The objective of predicting gender using logistic regression shows moderately effective results, with the model being able to classify genders based on the given features. However, a significant number of misclassifications are observed, suggesting that alternative models such as Support Vector Machines (SVM) or ensemble methods

could improve accuracy. The AUC of 0.74 indicates a moderate ability to distinguish between genders, which highlights the need for model refinement to achieve higher classification performance.

c) Hypothesis Testing:

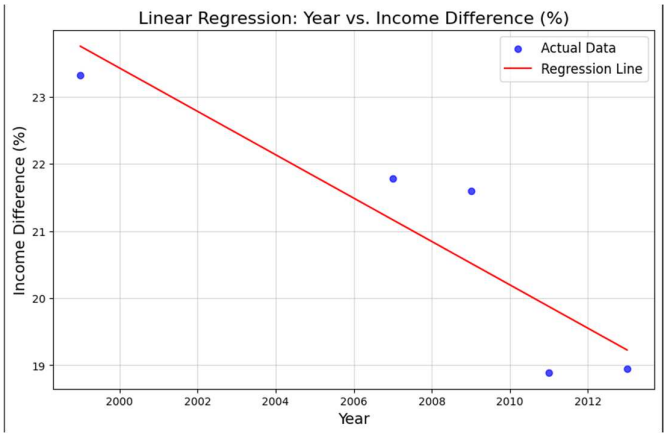
```
T-Statistic: 617.6782138093707
P-Value: 0.0
Reject the null hypothesis: There is a significant difference in income between males and females.
```

The t-statistic measures the difference between two groups relative to the standard error, with a larger t-value indicating a greater difference. In this case, the high t-statistic suggests a significant wage disparity. The p-value, representing the probability of observing such results under the null hypothesis, is 0.0—well below the 0.05 threshold. Since the p-value is lower than alpha, we reject the null hypothesis, confirming a significant income difference between males and females, supporting the presence of a gender pay gap in the dataset.

d) Principal Component Analysis:

The PCA scatterplot demonstrates that the first two components explain 42.97% of the variation, with Component 1 accounting for the most and aligning with characteristics such as income, education, and work hours. Despite this, there is a large overlap between the male (green) and female (orange) clusters, showing minimal separation and potential gender differences in these factors.

e) Difference in Income (%) over the years:



The graph presents the trend of income disparity (%) across the years, accompanied by a linear regression line that has been applied to the data. The negative slope of the line signifies a consistent decline in income disparity over time. According to the regression analysis, it is anticipated that the income difference will diminish to zero by approximately the year 2072. This indicates a gradual advancement in the effort to mitigate income inequalities, while also underscoring the sluggish rate of progress.

6. Conclusion:

The gender pay gap continues to be a significant challenge across various industries, with men earning more than women even when factors such as education, marital status, and years of experience are considered. The analysis reveals that these disparities are particularly pronounced in higher income groups, where men disproportionately benefit from longer work hours and greater experience. While linear regression provides useful insights, alternative approaches like non-linear models, Support Vector Machines (SVM), and ensemble methods could improve prediction accuracy. The moderate AUC score of 0.74 suggests that the gender classification models can be refined. The data must be properly updated from time to time to get better prediction results. To address this ongoing issue, systemic changes are needed, including policies that promote equality in the workplace, eliminate biases in hiring and promotions, and create more career development opportunities for all genders. This gives an idea to take drastic steps in the advancement of policies to reduce gender gap in the workplace where skills matter more than gender. By fostering greater diversity and inclusivity, organizations can help in reducing gender gap and promoting a good ecosystem for both men and women. Government interventions and social awareness are essential to encourage long-term changes that emphasize on the gender equality in the workplace.

7. References:

i) https://www.kaggle.com/datasets/fedesoriano/gender-pay-gap-dataset/suggestions?status=pending&yourSuggestions=true

ii) https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/

8. Uploading Documents:

i) PDF file named: G9.pdf

ii) Jupyter Notebook named: G9.ipynb

iii) Filtered dataset which are used by us named: data.csv