Unraveling Wage Disparities: Gender and Other Determinants of Compensation in the Workplace

Authors: Federico Di Gennaro, Elsa Farinella, Marco Scialanga

Data and Problem Description

In an era of increasing awareness and emphasis on equal opportunities in the workplace, the issue of wage discrimination continues to be a subject of paramount concern. What are the factors that unfairly affect workers' compensation? With the goal of addressing this issue, we dive into a subset of the *1985 Current Population Survey*, containing information about personal details and professional backgrounds of 534 workers in the United States.

We aim to investigate potential correlations between hourly wages and specific worker attributes. Specifically, our analysis will primarily address potential gender-related wage disparities, while also accounting for variables such as educational attainment, work experience, and marital status.

```
# Load data, select columns we are interested in
df <- read.csv("data.csv", header = TRUE, sep = ",")
df <- df[c("WAGE", "OCCUPATION", "EDUCATION",
           "EXPERIENCE", "AGE", "SEX", "MARR", "RACE")]
```

```
# Transform categorical variables to factors for easier visualization
df$SEX <- as.factor(df$SEX)
df$MARR <- as.factor(df$MARR)
df$RACE <- as.factor(df$RACE)
df$OCCUPATION <- as.factor(df$OCCUPATION)

# Check presence of missing values
print(c('Is there any missing value? ', any(is.na(df))))
```

```
[1] "Is there any missing value? " "FALSE"
```

The variables we are interested in are:

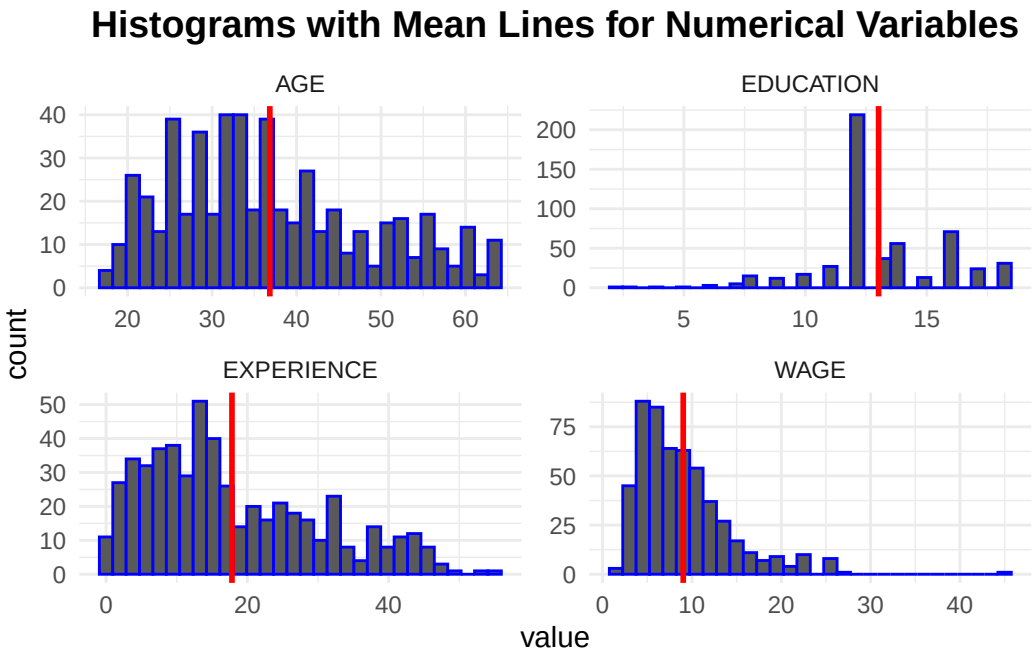
- `WAGE`: hourly wage in dollars (numerical);
- `OCCUPATION`: worker's occupation category (categorical - 1=Manager, 2=Salespeople, 3=Office workers, 4=Manual workers, 5=Professionals, 6=Other);
- `EDUCATION`: level of education measured in years of schooling/university education (numerical);
- `EXPERIENCE`: level of professional experience measured in years of full-time employment (numerical);
- `AGE`: worker's age measured in years (numerical);
- `SEX`: worker's gender (binary - 1 if female, 0 if male);
- `MARR`: worker's marital status (binary - 1 if married or has a steady partner, 0 otherwise);
- `RACE`: worker's ethnicity (categorical - 3 if Caucasian, 2 if Hispanic, 1 otherwise).

To get an idea of how our data is distributed, we take a look at each variable individually using histograms, barplots, and boxplots.

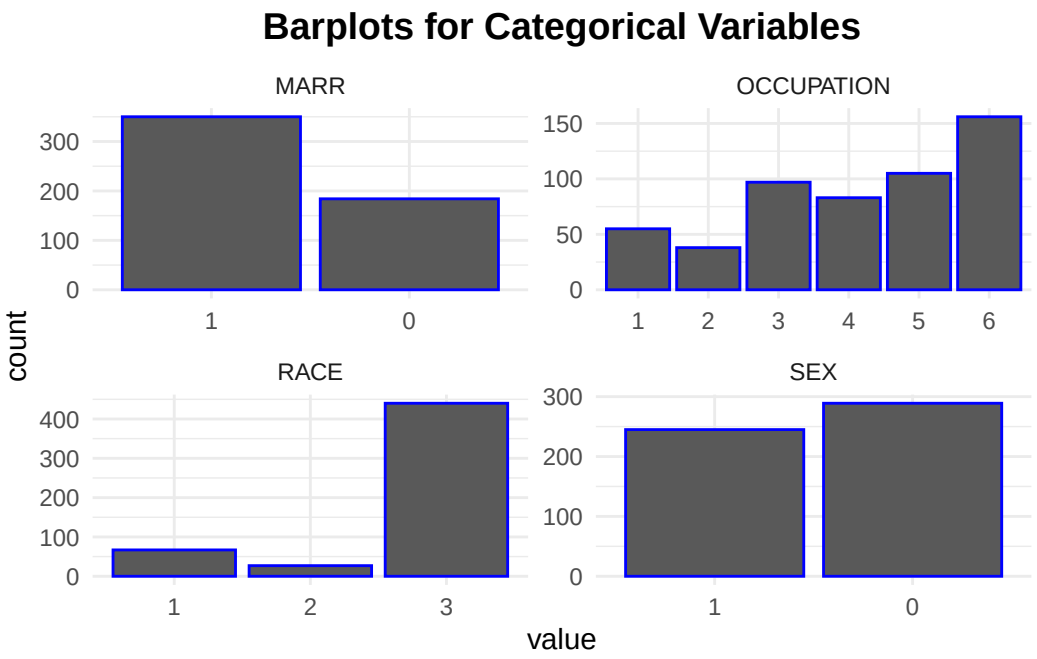
```
# Divide data in numerical and categorical for better visualization
numerical_data <- df[, sapply(df, is.numeric)]
factor_data <- df[, sapply(df, is.factor)]
```

```
# Calculate mean for each variable
stats_data <- numerical_data %>%
  pivot_longer(everything()) %>%
  group_by(name) %>%
  summarize(
    mean_value = mean(value, na.rm = TRUE),
  )

# Histograms for numerical variables
numerical_data %>%
  pivot_longer(everything()) %>%
  ggplot(aes(value)) +
  geom_histogram(bins = 30, color="blue") +
  geom_vline(data = stats_data, aes(xintercept = mean_value), color = "red",
             linetype = "solid", linewidth = 1) +
  facet_wrap(~ name, scales = "free") +
```

```
  labs(title = "Histograms with Mean Lines for Numerical Variables") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14, face = "bold", hjust = 0.5))
```




```
# Barplots for categorical variables
factor_data %>%
  pivot_longer(everything()) %>%
  ggplot(aes(x = value)) +
  geom_bar(color="blue") +
  facet_wrap(~ name, scales = "free") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14, face = "bold", hjust = 0.5)) +
  labs(title = "Barplots for Categorical Variables")
```



From the histograms above, we can see that our outcome variable WAGE is skewed right with a mean of around $9 per hour. Below we will see if and how this distribution varies for different subsets of our data (e.g. males / females, married / not married). From the histograms of the variables AGE and EXPERIENCE, we note that we have a pretty good representation of both young and experienced workers, with the majority being 25-35 years old and having 5-20 years of work experience. Furthermore, we can see that the most common education level in the data is high school, which corresponds to 12 years of education, with some university students as well.

The two genders are represented quite evenly. There are about twice as many married individuals than not married. The majority of occupations is in the "other" category, with the rest of the jobs represented somewhat equally. On the other hand, the vast majority of the individuals in the dataset is Caucasian, with much fewer representants of other ethnicities. This is a potential weakness of our data: it would be useful to have a more even representation of ethnicities for a better analysis.

Below are the descriptive statistics of the numerical variables of our dataset to get more specific information regarding each one.

```
# Descriptive statistics of numerical variables
summary(numerical_data)
```

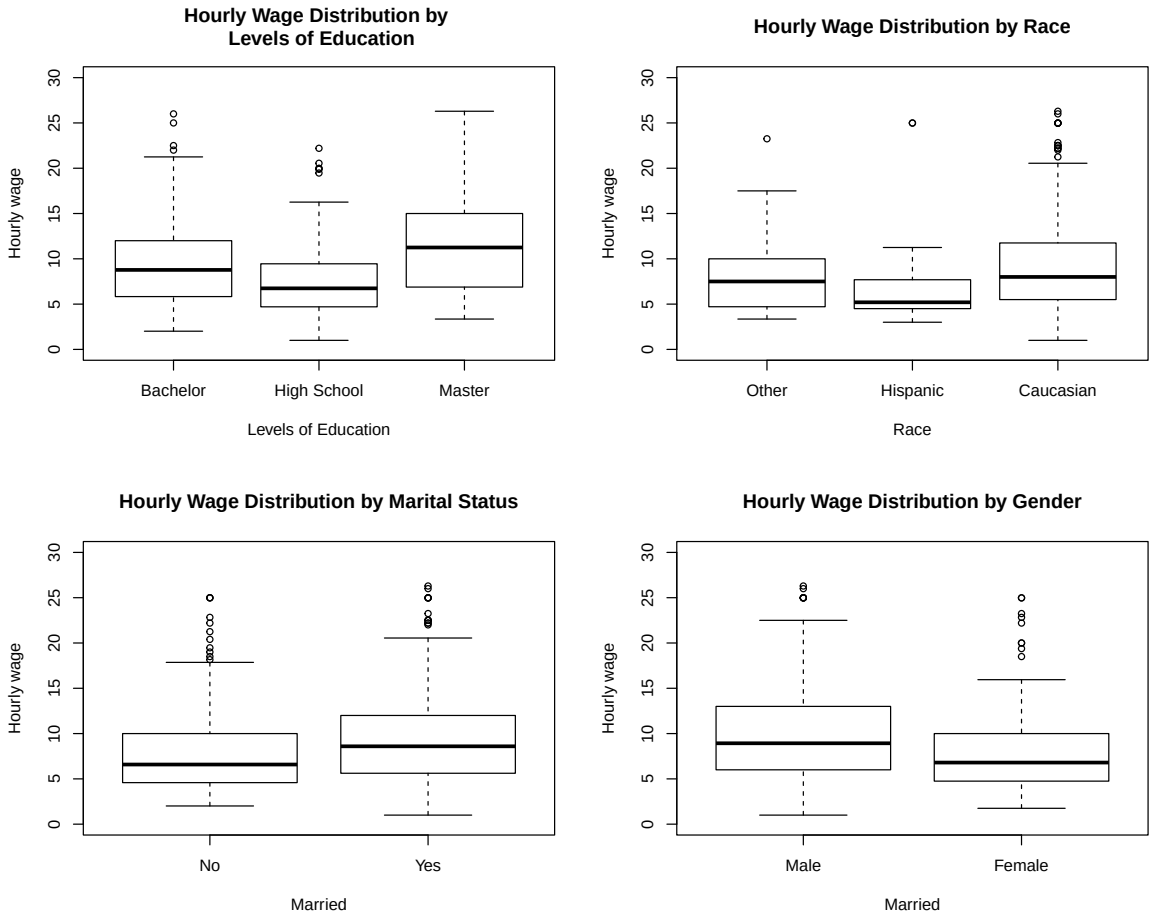
```
     WAGE          EDUCATION       EXPERIENCE          AGE
```

```
Min.   : 1.000   Min.   : 2.00   Min.   : 0.00   Min.   :18.00
1st Qu.: 5.250   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:28.00
Median : 7.780   Median :12.00   Median :15.00   Median :35.00
Mean   : 9.024   Mean   :13.02   Mean   :17.82   Mean   :36.83
3rd Qu.:11.250   3rd Qu.:15.00   3rd Qu.:26.00   3rd Qu.:44.00
Max.   :44.500   Max.   :18.00   Max.   :55.00   Max.   :64.00
```

Analysis of Correlation between Wage and Other Variables

We now begin investigating how hourly wage is correlated with personal details of each workers, such as their gender, ethnicity, education, etc.

```r
# New categorical feature for level of education
df$level_edu <- ifelse(df$EDUCATION<=12, "High School",
                ifelse(df$EDUCATION>12 & df$EDUCATION<=15, "Bachelor",
                       "Master"))

# Boxplots
par(mfrow=c(2,2)) # better visualizations
boxplot(df$WAGE ~ df$level_edu, col="white", ylim=c(0,30),
        xlab="Levels of Education", ylab="Hourly wage",
        main="Hourly Wage Distribution by \n Levels of Education")

boxplot(df$WAGE ~ df$RACE, col="white", ylim=c(0,30),
        names=c("Other", "Hispanic", "Caucasian"), xlab="Race",
        ylab="Hourly wage", main="Hourly Wage Distribution by Race")

boxplot(df$WAGE ~ df$MARR, col="white", ylim=c(0,30),
        names=c("No", "Yes"), xlab="Married", ylab="Hourly wage",
        main="Hourly Wage Distribution by Marital Status")

boxplot(df$WAGE ~ df$SEX, col="white", ylim=c(0,30),
        names=c("Male", "Female"), xlab="Married", ylab="Hourly wage",
        main="Hourly Wage Distribution by Gender")
```



Through the four boxplots above, we can see how the distribution of the WAGE variable differs depending on education level, ethnicity, marital status, and gender.

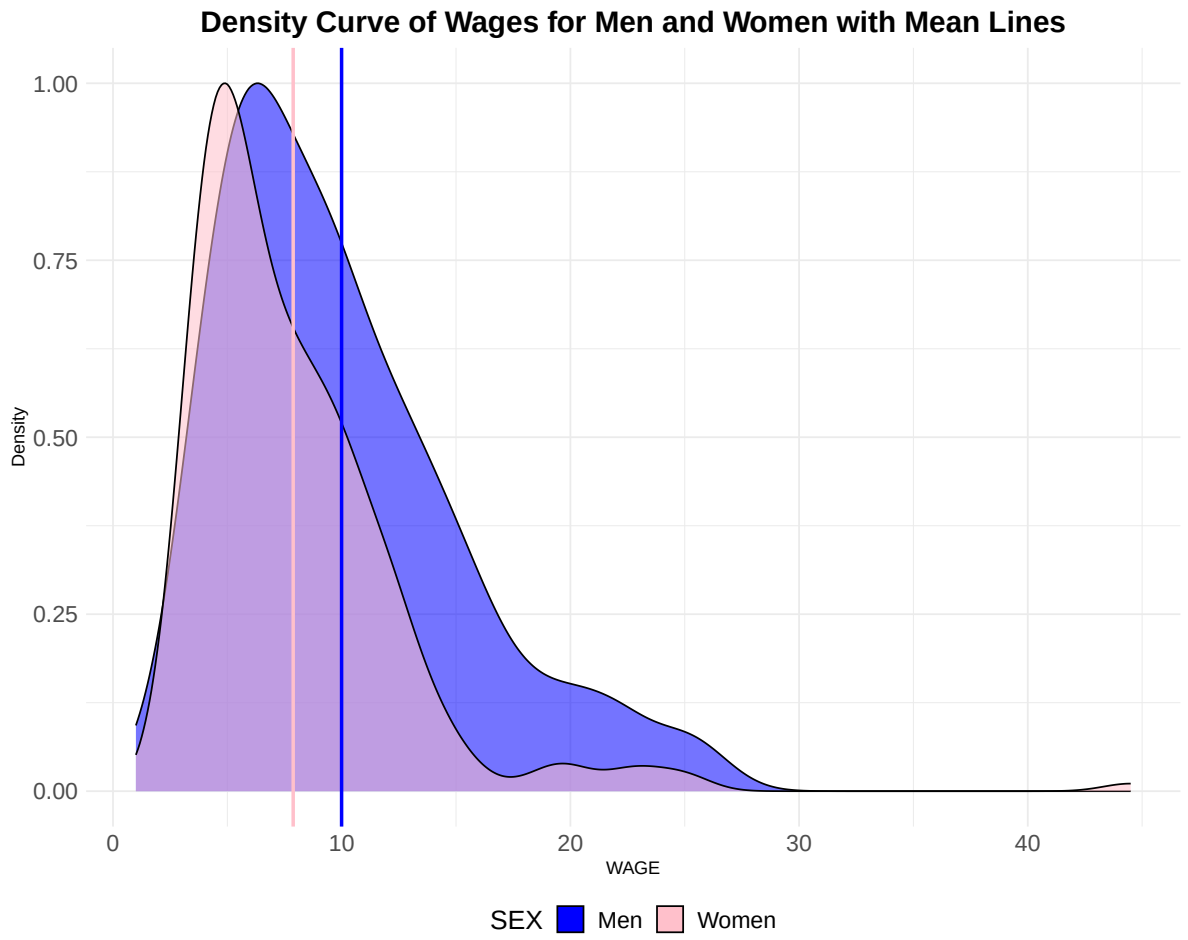
To generate the first boxplot on the topleft, we built a new categorical feature for better interpretability: "high school" (12 years or less of education), "bachelor" (12 - 15 years), "master" (more than 15 years). The boxplot illustrating the distribution of hourly wages across the three educational levels indicates that individuals with a Master's degree typically have a marginally higher median wage than those with either a Bachelor's degree or a high school education, as expected.

Furthermore, the boxplot representing the hourly wage distribution across three racial categories reveals that Caucasians possess the highest median wage. They are followed closely by the Hispanic group, while the "Other" category registers the lowest median wage. A possible explanation is that, in the United States, certain minorities are often living in worse socioeconomic conditions with fewer job opportunities and a lower access to high-level education.

Upon analyzing the wage distribution based on marital status, we note that individuals who are either married or have a stable partner have a higher median wage compared to their unmarried counterparts. A possible explanation would be that older people generally earn more and people who are married are, on average, older than people who are not. Further analysis with more data would be needed to make final conclusions on this topic.

The boxplot in which the wage distribution for the two different genders is compared suggests that the median wage for males exceeds that of females, highlighting a potential gender wage disparity. To further investigate this behavior, below we plot the distribution of the numerical variables and we get their correlation by disaggregating data with respect to the sex of the individual.

```r
# Compute mean for graph
mean_men <- mean(df$WAGE[df$SEX == "0"], na.rm = TRUE)
mean_women <- mean(df$WAGE[df$SEX == "1"], na.rm = TRUE)

# Density curves for men's and women's wages
ggplot(df, aes(x=WAGE, fill=SEX, alpha=0.5)) +
  geom_density(aes(y=after_stat(scaled)), position="identity") +
  scale_alpha(guide='none') +
  scale_color_manual(values = c("0" = "blue", "1" = "pink"),
                     labels = c("0" = "Men", "1" = "Women")) +
  scale_fill_manual(values = c("0" = "blue", "1" = "pink"),
                     labels = c("0" = "Men", "1" = "Women")) +
  theme_minimal() +
  geom_vline(aes(xintercept=mean_men), color="blue", linetype="solid", linewidth=1) +
  geom_vline(aes(xintercept=mean_women), color="pink", linetype="solid", linewidth=1) +
  labs(y="Density", title = "Density Curve of Wages for Men and Women with Mean Lines") +
  theme(
    axis.text.x = element_text(size = 14),
    axis.text.y = element_text(size = 14),
    legend.position = "bottom",
    legend.text = element_text(size = 14),
    legend.title = element_text(size = 16),
    plot.title = element_text(face = "bold", hjust = 0.5, size = 18))
```



- **Correlation of Wage with Age and Experience by Gender**

	Age	Experience
Men	0.284	0.186
Women	0.092	0.003

The difference between the two distributions above is clear. Although the two density curves have the same shape (resembling a log normal distribution), the men's curve is moved to the right of the women's, a discrepancy that is well summarised by the distance between the two means.

In addition, from the table above, we can see that variable WAGE for the two genders behaves quite differently with respect to EXPERIENCE and AGE: for men, wage is positively correlated with experience and age, while for women the correlation is absent. This indicates that

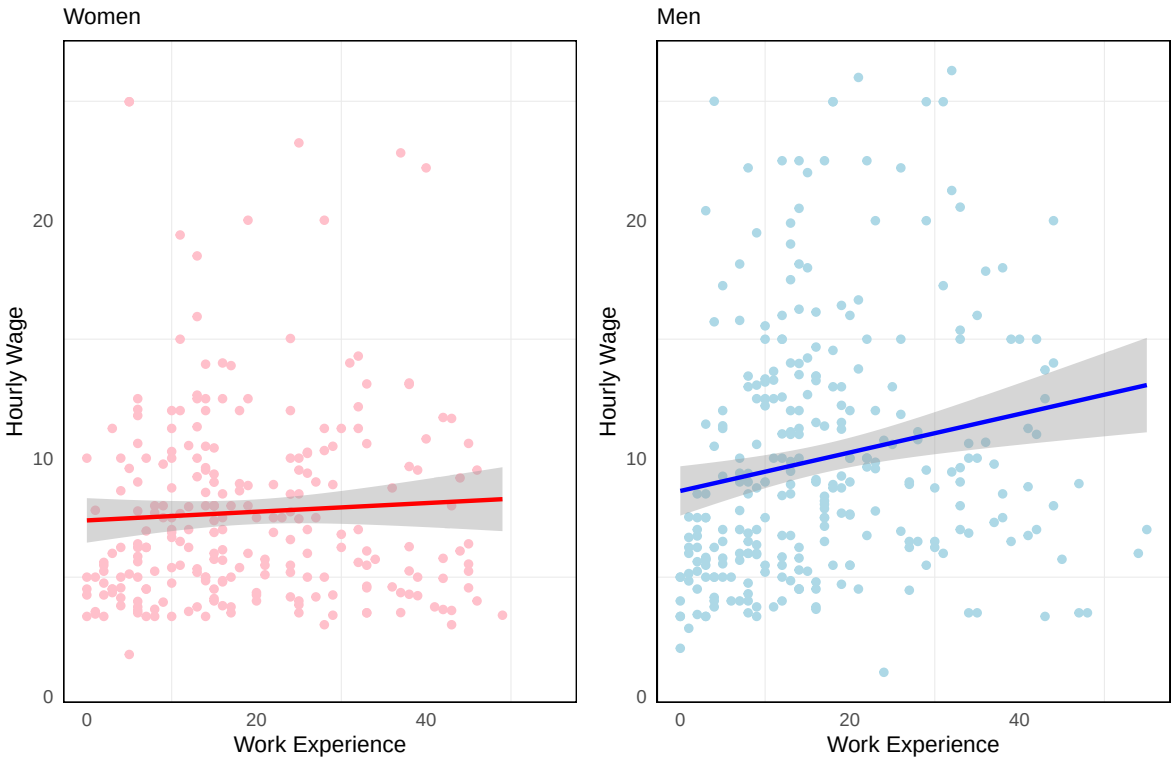
women's wages, unlike mens', tend not to increase as female workers grow older and become more experienced at their jobs.

In the analysis ahead, we will mainly focus on this disparity, looking deeper into factors that might cause it.

```
# Remove the outlier
df <- df[df$WAGE <40, ]
wage_women <- df[df$SEX==1, ]
wage_men <- df[df$SEX==0, ]

# Calculate common axis limits
y_limits <- range(wage_women$WAGE, wage_men$WAGE)
x_limits <- range(0, max(wage_women$EXPERIENCE, wage_men$EXPERIENCE))

# Scatterplot with regression lines
plot1 <- ggplot(data = wage_women, aes(x = EXPERIENCE, y = WAGE)) +
  geom_point(colour="pink") +
  geom_smooth(method="lm", se=T, colour="red") +
  labs(x="Work Experience", y="Hourly Wage", subtitle = "Women") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.border = element_rect(fill = NA, color = "black"),
        plot.title = element_text(face = "bold", hjust = 0.5)) +
  ylim(y_limits) +
  xlim(x_limits)

plot2 <- ggplot(data = wage_men, aes(x = EXPERIENCE, y = WAGE)) +
  geom_point(colour="lightblue") +
  geom_smooth(method="lm", se=T, colour="blue") +
  labs(x="Work Experience", y="Hourly Wage", subtitle = "Men") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.border = element_rect(fill = NA, color = "black"),
        plot.title = element_text(face = "bold")) +
  ylim(y_limits) +
  xlim(x_limits)

grid.arrange(plot1, plot2, ncol = 2,
             top = textGrob("Hourly Wage by Work Experience:
                            \n A Comparison Between Men and Women",
                            gp = gpar(fontface = "bold", fontsize = 14)))
```

Hourly Wage by Work Experience:

A Comparison Between Men and Women

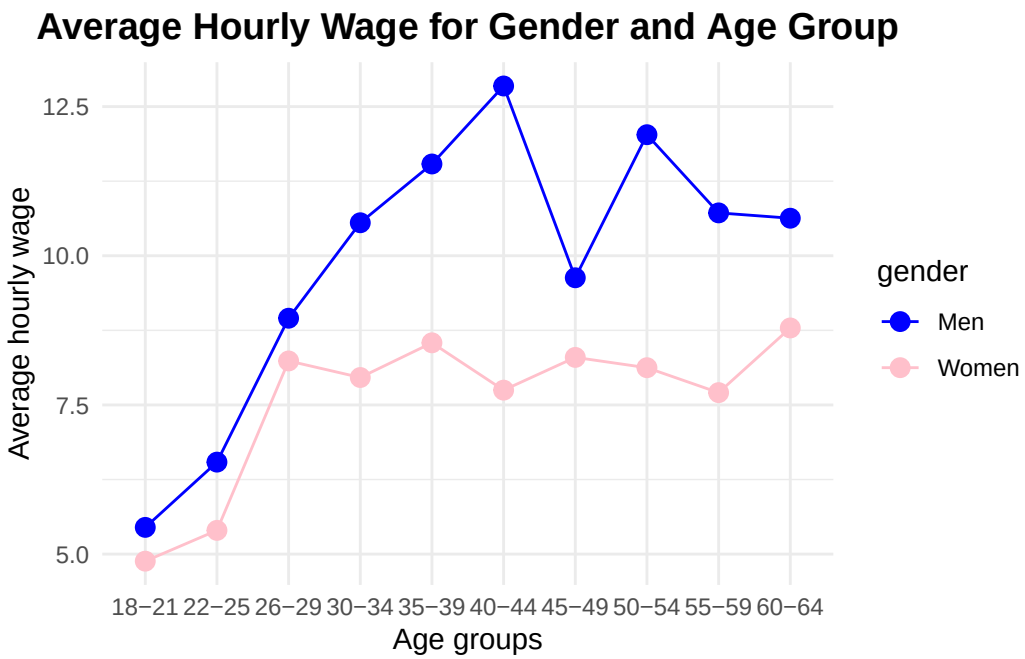


In the scatterplots above, the imposed regression lines with 95% confidence bands suggest that as work experience increases, the hourly wage also tends to rise for both genders. However, the slope of the trend line for men is much steeper, indicating a higher rate of wage increase with experience compared to women. At the same levels of work experience, women seem to have, on average, a lower hourly wage compared to men, as indicated by the trend lines.

In summary, while both genders experience a rise in hourly wages with increased work experience, there appears to be a gender wage gap with men potentially earning more than women, especially as work experience increases.

Next, we will look into wage disparities between the two genders, across different age groups, to check for a similar trend as that found above. First, we created a new column indicating the age group of each individual, using a range of 3 years for young workers (to better detect career evolutions that might be fairly quick at the beginning) and 4 years for older workers. Using this new feature, we generated the following plot.

```r
# New categorical variable for better interpretability
df$group_age <- ifelse(df$AGE<18, "Under 18",
                ifelse(df$AGE>=18 & df$AGE<22, "18-21",
                ifelse(df$AGE>=22 & df$AGE<26, "22-25",
                ifelse(df$AGE>=26 & df$AGE<30, "26-29",
                ifelse(df$AGE>=30 & df$AGE<35, "30-34",
                ifelse(df$AGE>=35 & df$AGE<40, "35-39",
                ifelse(df$AGE>=40 & df$AGE<45, "40-44",
                ifelse(df$AGE>=45 & df$AGE<50, "45-49",
                ifelse(df$AGE>=50 & df$AGE<55, "50-54",
                ifelse(df$AGE>=55 & df$AGE<60, "55-59",
                ifelse(df$AGE>=60 & df$AGE<65, "60-64"," "))))))))))

# Build matrix for plot below
mat_wage_sex <- aggregate(WAGE ~ SEX + group_age, data=df, FUN=mean)
colnames(mat_wage_sex) <- c("gender", "group_age","avg_salary")
mat_wage_sex <- mat_wage_sex %>%
  spread(key = gender, value = avg_salary)

matr_wage_sex <- tidyr::gather(mat_wage_sex, key = "gender",
                               value = "avg_salary", -group_age)

# Lineplot
ggplot(matr_wage_sex, aes(x = group_age, y = avg_salary,
                          color = gender, group = gender)) +
  geom_point(size = 3) + geom_line() +
  scale_color_manual(values = c("0" = "blue", "1" = "pink"),
                     labels = c("0" = "Men", "1" = "Women")) +  # Replace labels
  labs(title = "Average Hourly Wage for Gender and Age Group",
       x = "Age groups", y = "Average hourly wage") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14, face = "bold", hjust = 0.5))
```

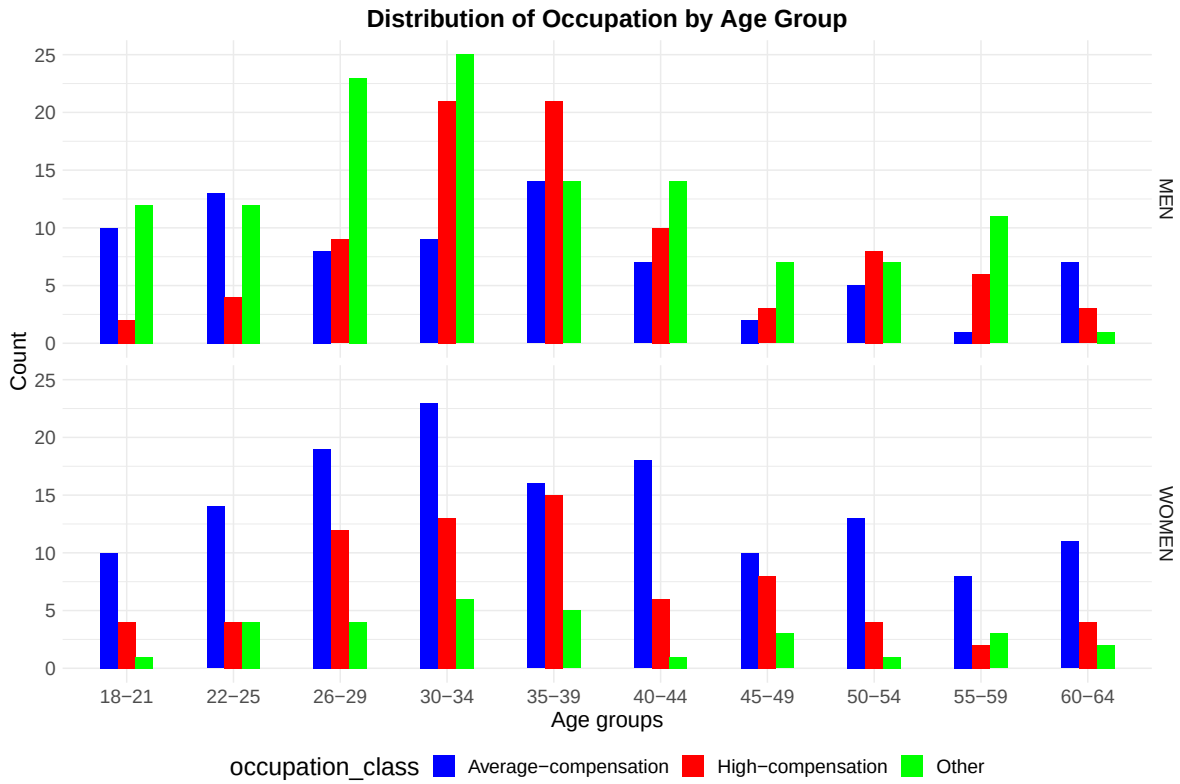


The chart illustrates that, across all age groups examined, men's average hourly wages consistently surpass those of women. In particular, it emerges that in the subgroup of women the average hourly wage tends to have a growth until the age of 30 while then it tends to stabilize for the remaining groups considered. In contrast, for the male subgroup, the upward trajectory in average hourly wages is evident until the age of 45, post which it remains mainly stable.

Note that this could be due to unbalanced classes in terms of type of occupation at different ages for men and women. For this reason, we first investigate, for every age group we used in the previous plot, the difference in type of occupation for men and women.

```
# New categorical variable for better interpretability
df$occupation_class <- ifelse(df$OCCUPATION==1 | df$OCCUPATION==5,
                              "High-compensation", ifelse
                              (df$OCCUPATION==2 | df$OCCUPATION==3
                              |df$OCCUPATION==4,"Average-compensation", "Other"))

# Create a custom facet variable
df$facet_label <- ifelse(df$SEX == 0, "MEN", "WOMEN")

# Create separate grouped bar charts for men and women
# stacked vertically with custom titles
```

```r
ggplot(df, aes(x = group_age, fill = occupation_class)) +
  geom_bar(width = 0.5, position = position_dodge(width = 0.5)) +
  scale_fill_manual(values = c("blue", "red", "green")) +
  facet_grid(facet_label ~ .) +
  labs(
    title = "Distribution of Occupation by Age Group",
    x = "Age groups",
    y = "Count"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 18, face = "bold", hjust = 0.5),
    axis.title.x = element_text(size = 16),
    axis.title.y = element_text(size = 16),
    axis.text.x = element_text(size = 14),
    axis.text.y = element_text(size = 14),
    legend.position = "bottom",
    legend.text = element_text(size = 14),
    legend.title = element_text(size = 18),
    strip.text = element_text(size = 14)
  )
```

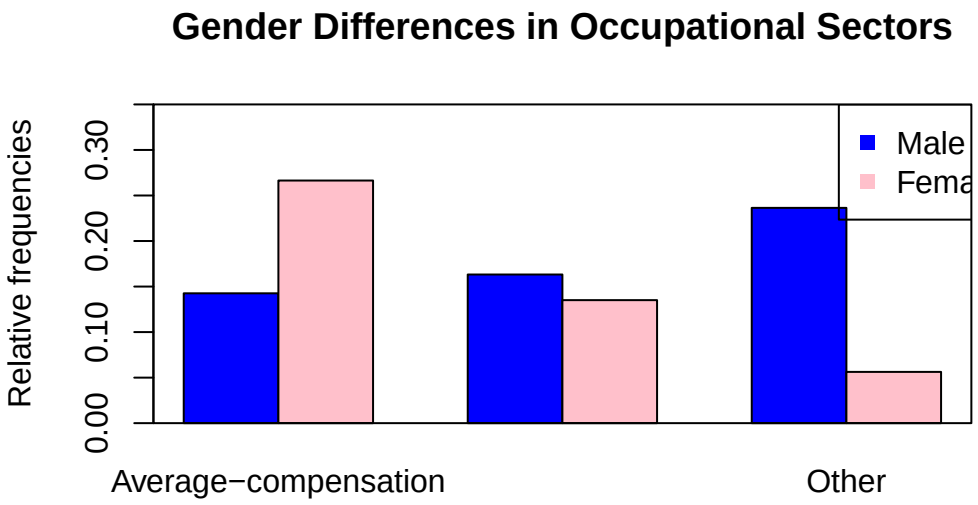


First, we aggregated what we considered "average-compensation" jobs (salespeople, office workers, manual workers) and what we considered "high-compensation" jobs (managers, professionals) to see whether men and women have a difference in proportion within them or not.

From the plot above, we can see that the huge difference in wage for the category 40-44 that we noticed (and to a lesser extent for the others as well) can be associated with the different distribution of the occupations for men and women. In particular, it is noticeable that the majority of women in our data are employed in what we called "average-compensation" jobs, and the majority of men instead are doing what we referred to as "high-compensation" jobs. For this reason, it would be better to investigate the difference in salary within categories of age by further dividing by occupation category, to eliminate the confounding factor of different occupations between men and women. Further, it is possible from the plot above to notice that a lot of men in our data have a "high-compensation" job early in their careers, which made us wonder: do men's careers advance faster than women's careers? To answer such a multifaceted and complicated question, more specific data and a deeper analysis would be needed.

```
# Barplot
barplot(prop.table(table(df$SEX, df$occupation_class)), beside=T,
        col=c("blue", "pink"), horiz = F,
```

```
        ylim=c(0,0.35), main="Gender Differences in Occupational Sectors",
        ylab="Relative frequencies")
legend(x=7.92, y=0.35, legend = c("Male", "Female"),
       col=c("blue", "pink"), pch=15)
box()
```



To conclude, this final graph illustrates gender differences across various occupational sectors. In "average-compensation" jobs, a higher proportion of females are represented compared to males. In contrast, "high-compensation" jobs show a relatively even distribution between the genders, with females slightly outnumbering males. For the "Other" category, males dominate significantly over their female counterparts. On one hand, the higher proportion of women in "average-compensation" jobs might explain why they are paid less, in general, than men. On the other hand, the fact that women and men are more or less equally represented in "high-compensation" jobs might suggest that women are compensated unfairly for this type of jobs compared to men.

Conclusion

The problem of wage disparity is very complex and affects a great amount of workers around the world. In this report, we highlighted certain trends related to this topic that we found in our data. First, we noticed that there are evident differences in hourly wage when considering

different categories, such as sex, marital status and education. Women are, on average, paid less than males, although this could be explained by other factors. However, the absence of a significant positive correlation between experience and hourly wage for women highlights a serious problem. Furthermore, we found that married and highly educated individuals tend to be better compensated than those who aren't.

These or similar trends might very well still be present in today's society, and an analysis of more modern data would be necessary to look for similar patterns in the modern workplace. For this reason, it is important to highlight that the results we have obtained should not be regarded as definitive, but rather as a starting point for deeper analysis.