University of Strathclyde Glasgow

DEPARTMENT OF

MATHEMATICS AND STATISTICS

Exploratory Analysis of Irish Gender Pay Gaps

by

202490978

MSc Data Analytics

Contents

List of Figures

List of Tables

List of Equations

Results

1 Dataset Overview

First looking at the Number of Companies by Industry Sector classifications in 2023. There were 587 different companies total in 2023 where the GICS Sector has 12 classifications and the ICB industry has 13 classifications:

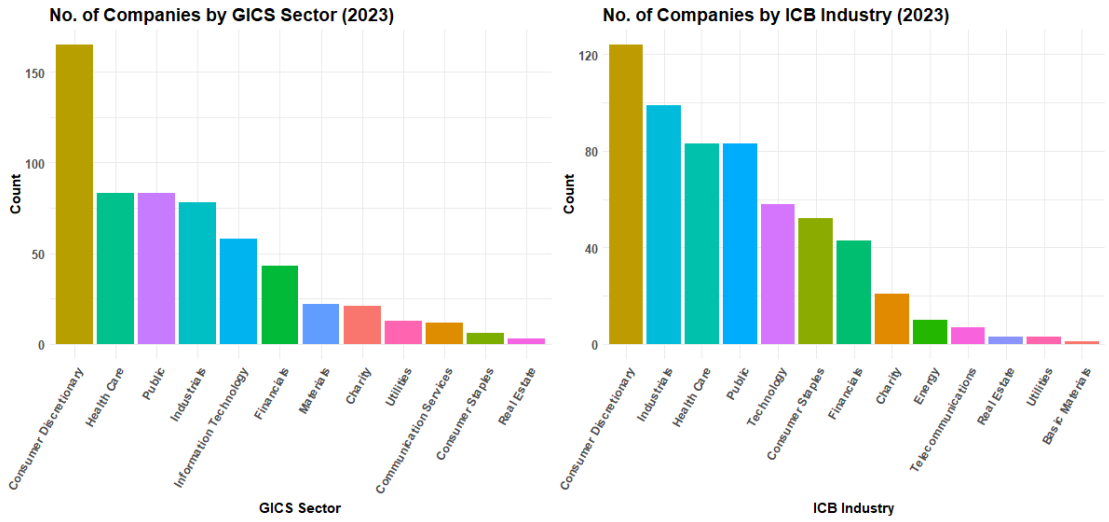

Figure 1.1 No. of Companies by GICS Sector and ICB Industry classifications for 2023

From figure 1.1, we can see in both classifications, the Consumer Discretionary sector has the highest representation, followed by Health Care and Public sectors in the GICS Sector classification and Industrials and Health Care in the ICB Industry classification.

2 Range of Pay Gaps by Company

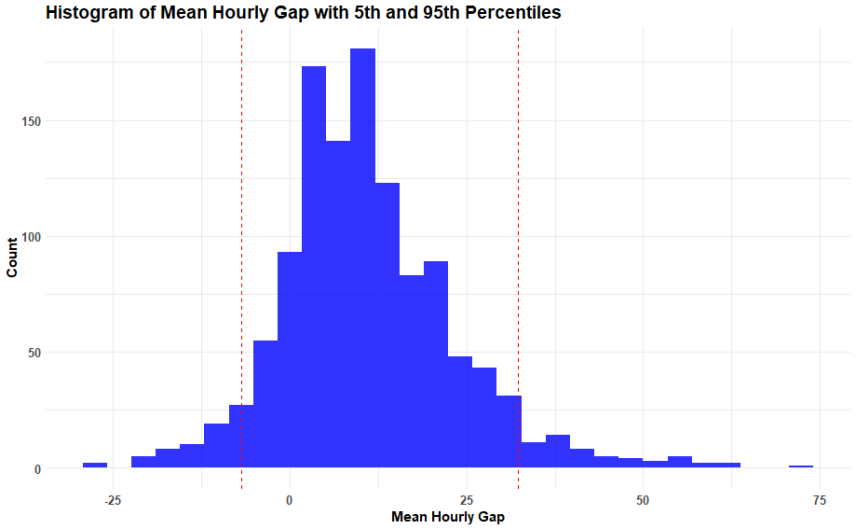

Figure 2.1 Histogram of Mean Hourly Gap

Figure 2.1 shows that most companies have a mean hourly gap greater than 0 where the median figure is 9.7, indicating that there is a significant mean hourly pay gap in Ireland. The pay gaps look to be approximately normally distributed.

3 Means vs. Medians

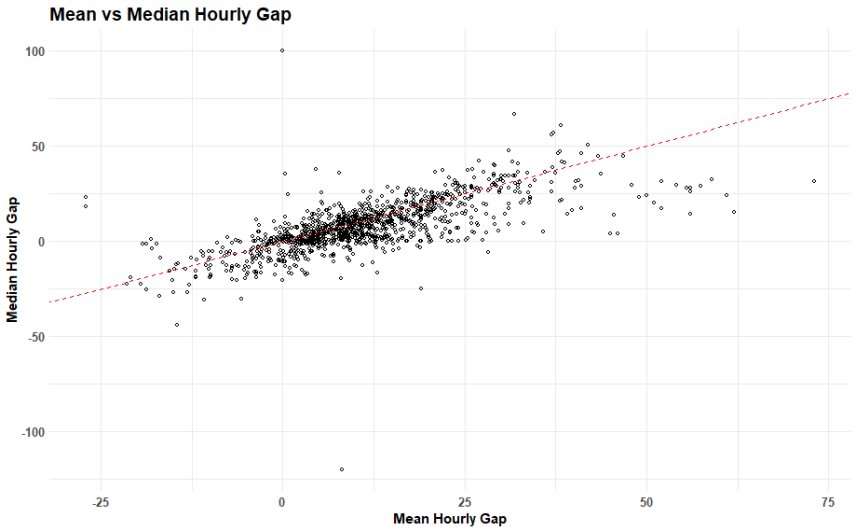

Figure 3.1 Scatterplot of Mean vs Median Hourly Gap

From figure 3.1, the plot shows a positive linear relationship between the Mean and the Median Hourly Gap across companies. This suggests that, for most companies, mean and median pay gaps are similar. However, some companies are some slight outliers, where the mean gap is higher than the median, possibly indicating the influence of a few highly paid individuals affecting the mean more than the median.

4 Variation Among Employee Categories and Types of Pay

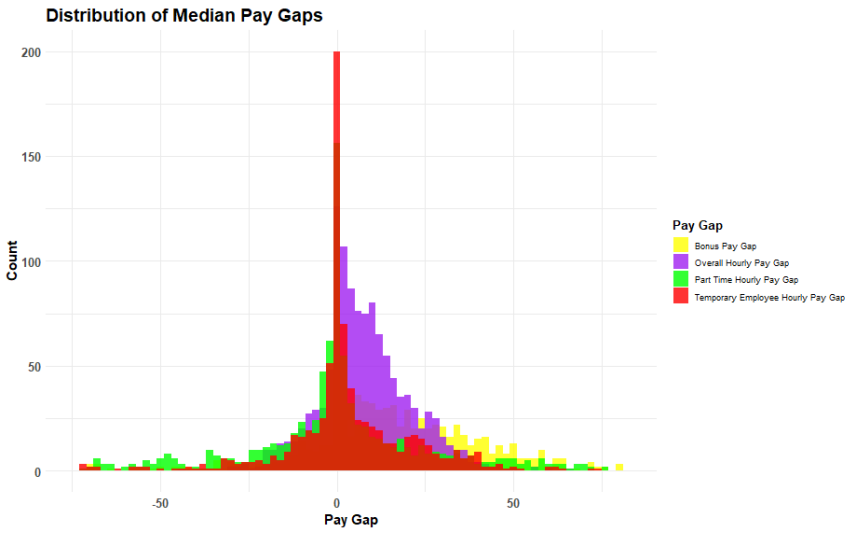

Figure 4.1 Histogram of Median Pay Gaps by Type of Pay Gap

Figure 4.1 shows that Part Time and Temporary Employee Hourly Pay Gap are tightly centred around zero, indicating small pay gaps for most companies in these categories. In contrast, Over Hourly Pay Gap and Bonus Pay Gap are skewed to the right, suggesting more significant pay disparities in these categories.

5 Breakdown by Industrial Sector and Change Over Time

Table 5.1 Median Hourly Pay Gap by GICS Sector from 2022 to 2023

GICS Sector	***2022***	***2023***	***Difference***
Charity	-0.33	-0.2	0.13
Communication Services	18.6	16.035	-2.565
Consumer Discretionary	4.99	3.49	-1.5
Consumer Staples	4.1	5.85	1.75
Financials	15	16	1
Health Care	5.61	6	0.39
Industrials	13.07	9.25	-3.82
Information Technology	11.7	11.3	-0.4
Materials	3.87	7.575	3.705
Public	1	-0.96	-1.96
Real Estate	20	23	3
Utilities	11.5	10.3	-1.2

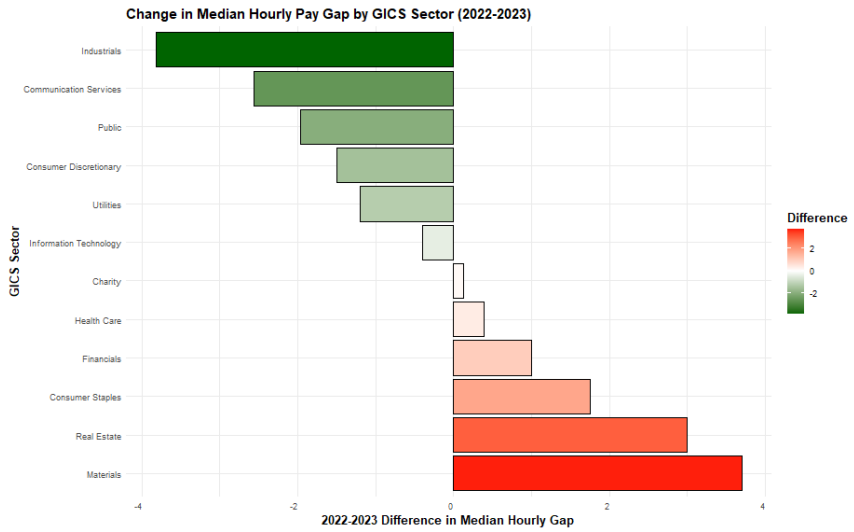

Figure 5.1 Change in Median Hourly Pay Gap by GICS Sector from 2022 to 2023

Table 5.1 and figure 5.1 show notable decreases in sectors like Industrials and Communication Services, while Materials and Real Estate show significant increases, with slight changes in sectors like Charity and Health Care.

6 Lack of Women in Top Positions: Defining an Index

The following formula was used to generate a "lack of women in top positions" (LWTP) index:

Equation 6.1 LWTP Index

$$LWTP = \frac{(Q1 + Q2) - (Q3 + Q4)}{(Q1 + Q2)}$$

This index measures the proportional difference in female representation between the lower pay quartiles (Q1 and Q2) and the upper pay quartiles (Q3 and Q4):

- A positive LWTP indicates that female representation decreases from the lower to the upper quartiles, suggesting fewer women in top pay positions.
- A zero LWTP implies equal representation across lower and upper pay quartiles.
- A negative LWTP suggests higher female representation in the upper quartiles, indicating more women in higher-paying roles.

7 LWTP vs. Pay Gaps

Figure 7.1 Scatterplot of Median Houry Pay Gap vs LWTP Index

Figure 7.1 shows each company's LWTP index against Median Hourly Pay Gap, which shows a positive linear relationship. Suggesting, that generally as pay gaps increase there is fewer women in higher earning positions, however, there are some outliers.

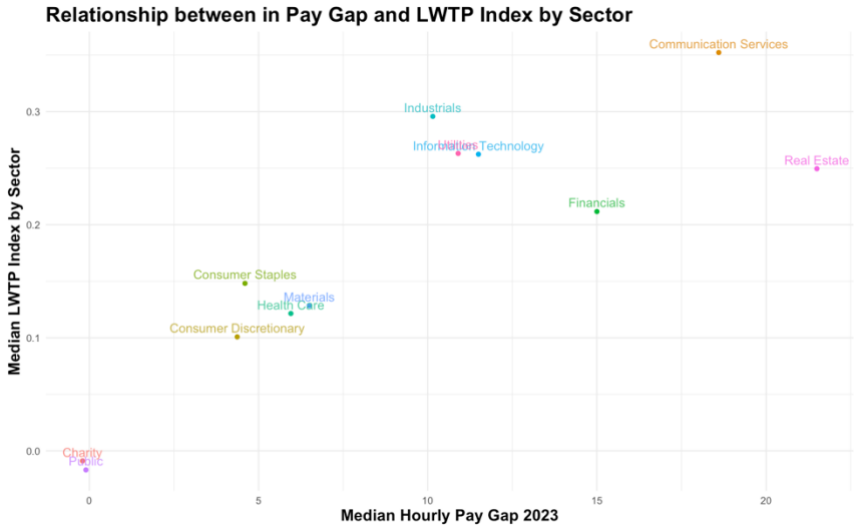

Figure 7.2 Scatterplot of Median Hourly Pay Gap

Figure 7.2 summarises this relationship by sector, once again showing a positive linear relationship. Also, looking at sectors like Communication Services and Real Estate exhibit both high pay gaps and high LWTP scores, suggesting larger gender disparities. In contrast, sectors like Public and Charity have low pay gaps and LWTP scores, indicating more equal representation and pay.

8 Summary and Recommendations

This dataset provides a comprehensive view of gender pay gaps across large companies in Ireland. The data highlights notable disparities in some sectors, like Communication Services and Real Estate.

The Lack of Women in Top Positions (LWTP) index also reveals that sectors with high pay gaps often have lower female representation in high-paying roles, indicating potential barriers to advancement for women.

However, data limitations restrict the depth of interpretation, particularly the absence of individual-level data to understand each employees' circumstances. Access to detailed data on this would be invaluable for identifying the root causes of these gaps.

Overall, focusing on policies tailored to inequal sectors, enhancing female representation in top roles, and improving data collection practices are key recommendations for advancing gender pay equity in Ireland.

Appendix

This analysis was conducted using R:

```
setwd("~/Library/CloudStorage/OneDrive-UniversityofStrathclyde/Fifth Year/MM916/Project 1")

library(tidyverse)
library(ggplot2)
library(patchwork)

# (1) Dataset overview. Read the dataset into R. For full credit, do this directly from the file
paygap.ie.csv,
# rather than paygap.RData (which can be loaded with load('paygap.RData')). Summarise the number of
# companies by report year and make two bar charts that give number of companies by sector, for the
two sector
# classifications (GICS Sector and ICB Industry). Use just one Report Year for this to avoid double-
counting.
# For full credit, get R to display the two bar plots side by side in a single image, and order the bars by
number
# of companies, not alphabetically by sector.

paygap <- read.csv('paygap.ie.csv')

colnames(paygap)

paygap23 <- paygap %>%
  filter(Report.Year == '2023')

length(paygap$Company.Name[paygap$Report.Year == 2023])

gics <- paygap23 %>%
  group_by(GICS.Sector) %>%
  summarise(`No. of Companies` = n())

icb <- paygap23 %>%
  group_by(ICB.Industry) %>%
  summarise(`No. of Companies` = n())

write.csv(gics, "~/Library/CloudStorage/OneDrive-UniversityofStrathclyde/Fifth Year/MM916/Project
1/gics.csv", row.names = FALSE)

write.csv(icb, "~/Library/CloudStorage/OneDrive-UniversityofStrathclyde/Fifth Year/MM916/Project
1/icb.csv", row.names = FALSE)

p1 <- ggplot(paygap23, aes(x = fct_infreq(GICS.Sector), fill = GICS.Sector)) +
  geom_bar() +
  labs(x = "GICS Sector",
```

```
      y = "Count",
      title = "No. of Companies by GICS Sector (2023)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1, face = "bold"),
        plot.title = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        legend.position = 'none')

p2 <- ggplot(paygap23, aes(x = fct_infreq(ICB.Industry), fill = ICB.Industry)) +
  geom_bar() +
  labs(x = "ICB Industry", y = "Count", title = "No. of Companies by ICB Industry (2023)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 60, hjust = 1, face = "bold"),
        plot.title = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        legend.position = 'none')

p1 + p2

# (2) Range of pay gaps by company. Make a histogram of Mean Hourly Gap. Mark the 5th and 95th
# percentiles of the data with vertical lines. Comment on the range of values and identify the companies
with
# the maximum and minimum overall pay gaps.

quantile(paygap$Mean.Hourly.Gap, c(0.05, 0.5, 0.95), na.rm = TRUE)

ggplot(paygap, aes(x = Mean.Hourly.Gap)) +
  geom_histogram(fill = 'blue', alpha = 0.8, bins = 30) +
  labs(x = "Mean Hourly Gap",
       y = "Count",
       title = "Histogram of Mean Hourly Gap with 5th and 95th Percentiles") +
  geom_vline(xintercept = -6.825, colour = "red", linetype = "dashed", size = 0.5) +
  geom_vline(xintercept = 32.330, colour = "red", linetype = "dashed", size = 0.5) +
  theme_minimal()

# (3) Means vs. medians. Companies are required to report both means and medians, since if the gender
# pay gap is driven by a small number of highly paid men, the mean will be much higher than the
median, with
# the median better reflecting the typical employee. Use a scatter plot to form a hypothesis about
whether this
# is an important issue in this dataset.

ggplot(paygap, aes(x = Mean.Hourly.Gap, y = Median.Hourly.Gap)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
```

```
  labs(x = 'Mean Hourly Gap',
       y = 'Median Hourly Gap',
       title = 'Mean vs Median Hourly Gap') +
  theme_minimal()

# (4) Variation among employee categories and types of pay. The median hourly pay gap is given for
# i) employees overall, ii) part-time employees, and iii) temp employees, along with iv) the median gap
for
# bonus pay. Compare the four distributions using histograms and form a hypothesis about whether
differences
# exist. You will probably need to filter some outliers out of the plot or control the axis scale in order to
see
# the relevant detail. For full credit, put the four distributions together on one set of axes.

mediangap <- paygap %>%
  select('Company.Name', 'GICS.Sector', 'Report.Year', 'Median.Hourly.Gap',
'Median.Hourly.Gap.Part.Time',
         'Median.Hourly.Gap.Part.Temp', 'Median.Bonus.Gap') %>%
  pivot_longer(cols = c('Median.Hourly.Gap', 'Median.Hourly.Gap.Part.Time',
                        'Median.Hourly.Gap.Part.Temp', 'Median.Bonus.Gap'),
               names_to = "Pay_Gap",
               values_to = "Value") %>%
  na.omit()

summary(mediangap$Value)

mean <- mean(mediangap$Value)
s <- sd(mediangap$Value)

quantile(mediangap$Value, seq(0,1,0.01))

mediangap$Pay_Gap <- recode(mediangap$Pay_Gap, Median.Hourly.Gap = 'Overall Hourly Pay Gap',
                            Median.Hourly.Gap.Part.Time = 'Part Time Hourly Pay Gap',
                            Median.Hourly.Gap.Part.Temp = 'Temporary Employee Hourly Pay Gap',
                            Median.Bonus.Gap = 'Bonus Pay Gap')

mediangap1 <- mediangap %>%
  filter(Pay_Gap == 'Overall Hourly Pay Gap')

mediangap2 <- mediangap %>%
  filter(Pay_Gap == 'Part Time Hourly Pay Gap')

mediangap3 <- mediangap %>%
  filter(Pay_Gap == 'Temporary Employee Hourly Pay Gap')

mediangap4 <- mediangap %>%
  filter(Pay_Gap == 'Bonus Pay Gap')
```

```
p1 <- ggplot(mediangap1, aes(x = Value)) +
  geom_histogram(alpha = 0.8, position = "identity", binwidth = 2, fill = 'purple') +
  labs(title = "Distribution of Median Overall Hourly Pay Gap",
       x = "Pay Gap", y = "Count") +
  xlim(mean - s, mean + s) +
  ylim(0, 200) +
  theme_minimal()

p2 <- ggplot(mediangap2, aes(x = Value)) +
  geom_histogram(alpha = 0.8, position = "identity", binwidth = 2, fill = 'green') +
  labs(title = "Distribution of Median Part Time Hourly Pay Gaps",
       x = "Pay Gap", y = "Count") +
  xlim(mean - s, mean + s) +
  ylim(0, 200) +
  theme_minimal()

p3 <- ggplot(mediangap3, aes(x = Value)) +
  geom_histogram(alpha = 0.8, position = "identity", binwidth = 2, fill = 'red') +
  labs(title = "Distribution of Median Temporary Employee Pay Gaps",
       x = "Pay Gap", y = "Count") +
  xlim(mean - s, mean + s) +
  ylim(0, 200) +
  theme_minimal()

p4 <- ggplot(mediangap4, aes(x = Value)) +
  geom_histogram(alpha = 0.8, position = "identity", binwidth = 2, fill = 'yellow') +
  labs(title = "Distribution of Median Bonus Pay Gaps",
       x = "Pay Gap", y = "Count") +
  xlim(mean - s, mean + s) +
  ylim(0, 200) +
  theme_minimal()

(p1 + p2)/(p3 + p4)

ggplot(mediangap, aes(x = Value, fill = Pay_Gap)) +
  geom_histogram(alpha = 0.8, position = "identity", binwidth = 2) +
  scale_fill_manual(values = c("Overall Hourly Pay Gap" = "purple", "Part Time Hourly Pay Gap" =
"green",
                               'Temporary Employee Hourly Pay Gap' = 'red', 'Bonus Pay Gap' = 'yellow')) +
  labs(title = "Distribution of Median Pay Gaps",
       x = "Pay Gap", y = "Count", fill = 'Pay Gap') +
  xlim(mean - s, mean + s) +
  theme_minimal() +
  theme(plot.title = element_text(size = 18, face = "bold"),
        axis.title.x = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 14, face = "bold"),
```

```
      axis.text.x = element_text(size = 12, face = "bold"),
      axis.text.y = element_text(size = 12, face = "bold"),
      legend.title = element_text(size = 12, face = "bold"))

# (5) Breakdown by industrial sector and change over time. Make a table that summarises the median
# hourly gap data by industrial sector (either GICS or ICB) and year. Use rows for sectors, two columns
for
# 2022 and 2023 median, and a third column for the 2022–2023 difference. (For full credit, make R
produce the
# table in this format, instead of rearranging it by hand for your report.) Are there industrial sectors
where it
# seems that pay gaps are increasing or decreasing over time? Make a plot that allows you to explore
this.

industrygap <- paygap %>%
  select('GICS.Sector', 'Report.Year', 'Median.Hourly.Gap') %>%
  group_by(GICS.Sector, Report.Year) %>%
  summarise(Median = median(Median.Hourly.Gap, na.rm = TRUE)) %>%
  ungroup() %>%
  pivot_wider(names_from = 'Report.Year', values_from = 'Median') %>%
  mutate(Difference = `2023` - `2022`)

write.csv(industrygap, "~/Library/CloudStorage/OneDrive-UniversityofStrathclyde/Fifth
Year/MM916/Project 1/Industry Gap.csv", row.names = FALSE)

industrygap <- industrygap %>%
  arrange(desc(Difference))

ggplot(industrygap, aes(x = fct_reorder(GICS.Sector, Difference, .desc = TRUE), y = Difference, fill =
Difference)) +
  geom_bar(stat = "identity", colour = "black") +
  coord_flip() +
  labs(x = "GICS Sector",
       y = "2022-2023 Difference in Median Hourly Gap",
       title = "Change in Median Hourly Pay Gap by GICS Sector (2022-2023)") +
  scale_fill_gradient2(low = "darkgreen", high = "red") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1),
        plot.title = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        axis.title.x = element_text(size = 12, face = "bold"),
        legend.title = element_text(size = 12, face = "bold"))

# (6) Lack of women in top positions: defining an index. The dataset also contains the percentage of
# female employees by quartile of individual pay (Q1 Female . . . Q4 Female). If the gender pay gap is
driven
```

```
# by a lack of women in top positions, we would expect to see that the Q4 percentage was low
compared
# with the company-wide value, or that there was a gradual decline across the Q1. . . Q4 percentages.
Write a
# function that takes the four quartile values as arguments and returns a single value that summarises
them as
# a “lack of women in top positions” (LWTP) index.
#
# There is no standard way of defining this: you will need to invent something and test it. (Google the
“Glass
# Ceiling Index” if you would like to see an example of something similar.) It is okay if your index is quite
# simple mathematically. Write test cases that can be used as examples of how to interpret values of
this
# LWTP index, as well as verifying that your function is working.

LWTP <- function(Q1, Q2, Q3, Q4) {
  ifelse(Q1 == 0, NA, ((Q1 + Q2) - (Q3 + Q4)) / (Q1 + Q2))
}

LWTP(80, 50, 50, 20)

LWTP(55, 50, 50, 45)

LWTP(50, 50, 50, 50)

LWTP(40, 50, 50, 55)

# (7) LWTP vs. pay gaps. Now calculate your LWTP index for every company, and summarise the results
# by industrial sector. Does variation in LWTP seem to be related to pay gaps? There are many ways to
# approach this. Include the tables or plots that you use to answer the question.

paygap <- paygap %>%
  mutate(LWTP = LWTP(Q1.Female, Q2.Female, Q3.Female, Q4.Female))

ggplot(paygap, aes(x = Median.Hourly.Gap, y = LWTP)) +
  geom_point(aes(colour = GICS.Sector), alpha = 0.6) +
  labs(x = "Median Hourly Pay Gap",
       y = "Median LWTP Index",
       title = "Relationship between Pay Gap and LWTP by Sector",
       colour = "Sector") +
  theme_minimal()

ggplot(paygap, aes(x = Median.Hourly.Gap, y = LWTP)) +
  geom_point(aes(colour = GICS.Sector), alpha = 0.6) +
  labs(x = "Median Hourly Pay Gap",
       y = "Median LWTP Index",
       title = "Relationship between Pay Gap and LWTP by Sector",
```

```
    colour = "Sector") +
  ylim(-1.5, 1) +
  xlim(-50, 75) +
  theme_minimal()

femalerep <- paygap %>%
  select(Company.Name, GICS.Sector, Q1.Female, Q2.Female, Q3.Female, Q4.Female) %>%
  na.omit() %>%
  mutate(LWTP = LWTP(Q1.Female, Q2.Female, Q3.Female, Q4.Female)) %>%
  group_by(GICS.Sector) %>%
  summarise(median_LWTP = median(LWTP, na.rm = TRUE)) %>%
  ungroup()

gicsgap <- paygap %>%
  select('GICS.Sector', 'Median.Hourly.Gap') %>%
  group_by(GICS.Sector) %>%
  summarise(Median = median(Median.Hourly.Gap, na.rm = TRUE)) %>%
  ungroup()

final <- left_join(gicsgap, femalerep, by = "GICS.Sector")

ggplot(final, aes(x = Median, y = median_LWTP, label = GICS.Sector, colour = GICS.Sector)) +
  geom_point() +
  geom_text(vjust = -0.5, hjust = 0.5, alpha = 0.8) +
  labs(x = "Median Hourly Pay Gap 2023",
       y = "Median LWTP Index by Sector",
       title = "Relationship between in Pay Gap and LWTP Index by Sector") +
  theme_minimal() +
  theme(legend.position = 'none',
        plot.title = element_text(size = 18, face = "bold"),
        axis.title.x = element_text(size = 14, face = "bold"),
        axis.title.y = element_text(size = 14, face = "bold"))
```