MM916 Project 1 (2024): Exploratory analysis of Irish gender pay gaps

Companies in Ireland with 250+ employees are required by law to report on differences in pay between their male and female employees, but there is no official portal or data format for these reports. Data scientist Jennifer Keane has filled this need with the website `paygap.ie`. The full dataset she has assembled has been provided to you on MyPlace in two forms, a .csv file and an .RData file (which contains identical data).

The "mean pay gap" for one company is defined as the mean pay for men minus the mean pay for women, expressed as a percentage of the mean pay overall. Median pay gaps are defined in an analogous way. See 'Understanding the Data' on `paygap.ie` for more information about the calculations behind the dataset. Note that 'mean' and 'median' in variable names describe the method of summarising across *individuals* within one company—we don't have the individual-level data. Your analysis will involve statistical summaries (mean, median, percentiles, etc.) across sets of *companies*.

Instructions

Write the Results section of a technical report that describes an exploratory analysis of the paygap.ie dataset, including the 8 elements listed on the next page. Make a single Word or PDF document containing figures, tables, and the brief text that accompanies them. Include the R code that supported this analysis at the end as an Appendix.

- A brief summary of results from each element should be included in the main text. This text is important but does not need to be long (our sample solutions are about 4 pages long, mostly figures and tables).
- It is not necessary to write an Introduction or Methods section (this will change in Project 2).
- Figures should have captions. Tables should have descriptive titles, and be formatted professionally: do not simply copy and paste raw output from R.
- The subject of the report is gender pay gaps in Ireland, not the process of programming in R.
- It is not necessary to run any statistical tests of your interpretations of the data. This project is an *exploratory* data analysis and so guesses based on what you see by eye are sufficient (this is how statistical hypotheses are generated, not tested).
- You can make use of everything on MyPlace and also general web searches (Google, Stack Exchange, etc.), but you may not work in groups or copy from a classmate: these are individual projects and everything needs to be your own work.
- Use of AI-based assistants like ChatGPT is not allowed.

Marking

For each of elements 1–7:

- 2 points for correct code approach
- 2 points for code that runs without errors and gives correct output (calculations, figures). The evidence of correct output must be within the report.
- 1 point for figure and table presentation to a high professional standard (labelling; categorical variables sorted and with NA columns removed; effective graphical choices)
- 1 point for clear and sensible written interpretation.

Additionally:

- 4 points for element 8 (the written Summary and Recommendations)
- 4 points for overall presentation (50 points total).

Where the instructions say *for full credit*, this identifies details that may be time-consuming or difficult to get right, worth 1–2 points per element within the rubric above.

What to include

(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.

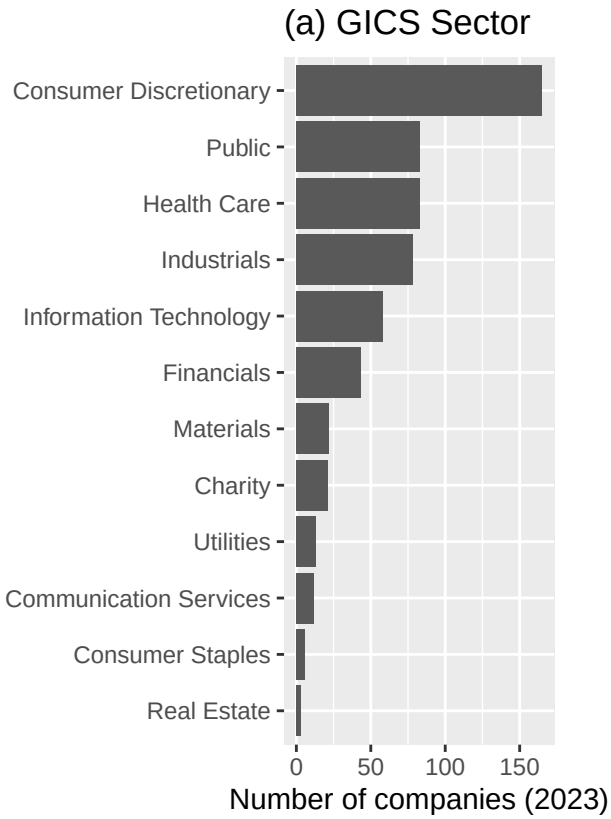
```
pg <- read_delim("paygap.ie.csv", delim=",")
```

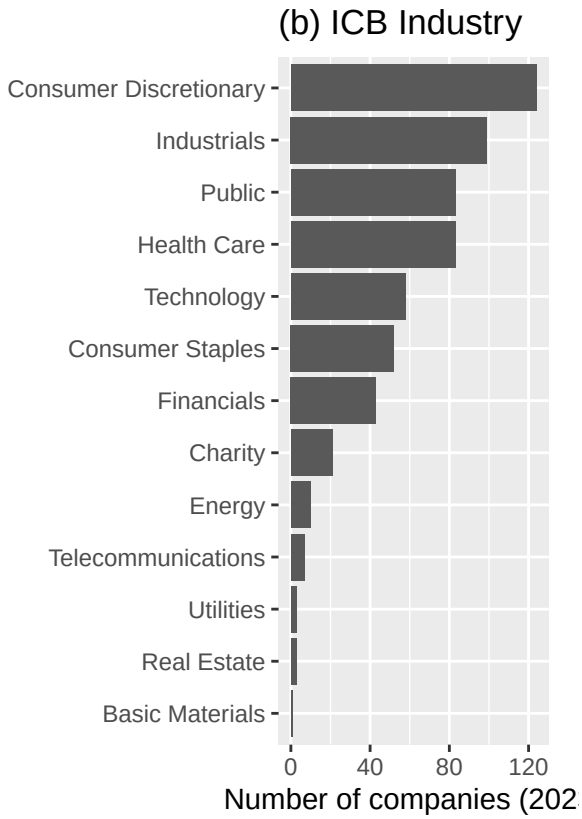
```
## Rows: 1189 Columns: 28
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): Company Name, GICS Sector, ICB Industry, Company Site, Report Link
## dbl (23): Report Year, Mean Hourly Gap, Median Hourly Gap, Mean Bonus Gap, M...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
pg %>% count(`Report Year`)
```

```
## # A tibble: 2 x 2
##   `Report Year`     n
##           <dbl> <int>
## 1          2022   602
## 2          2023   587
```

```
pg %>% filter(`Report Year`==2023) %>%
  count(`GICS Sector`) %>% arrange(n) %>%
  # now that we have the rows in a sorted sequence we want to preserve, coerce the
  # sectors names to a factor and using the names themselves as the levels, to force
  # ggplot to keep that ordering.
  mutate(sector = factor(`GICS Sector`, levels=`GICS Sector`)) %>%
  ggplot() + geom_col(aes(n, sector)) +
  labs(x="Number of companies (2023)", y="", title="(a) GICS Sector") -> p1
pg %>% filter(`Report Year`==2023) %>%
  count(`ICB Industry`) %>% arrange(n) %>%
  mutate(sector = factor(`ICB Industry`, levels=`ICB Industry`)) %>%
  ggplot() + geom_col(aes(n, sector)) +
  labs(x="Number of companies (2023)", y="", title="(b) ICB Industry") -> p2
p1 | p2
```






```
length(unique(c(pg$`GICS Sector`, pg$`ICB Industry`)))
```

```
## [1] 16
```

(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.

```
# a few quantiles to mention in the text
quantile(pg$`Mean Hourly Gap`,c(0.05,0.5,0.95), na.rm=TRUE)
```

```
##     5%    50%    95%
## -6.825  9.700 32.330
```

```
min(pg$`Mean Hourly Gap`, na.rm = TRUE)
```

```
## [1] -27
```

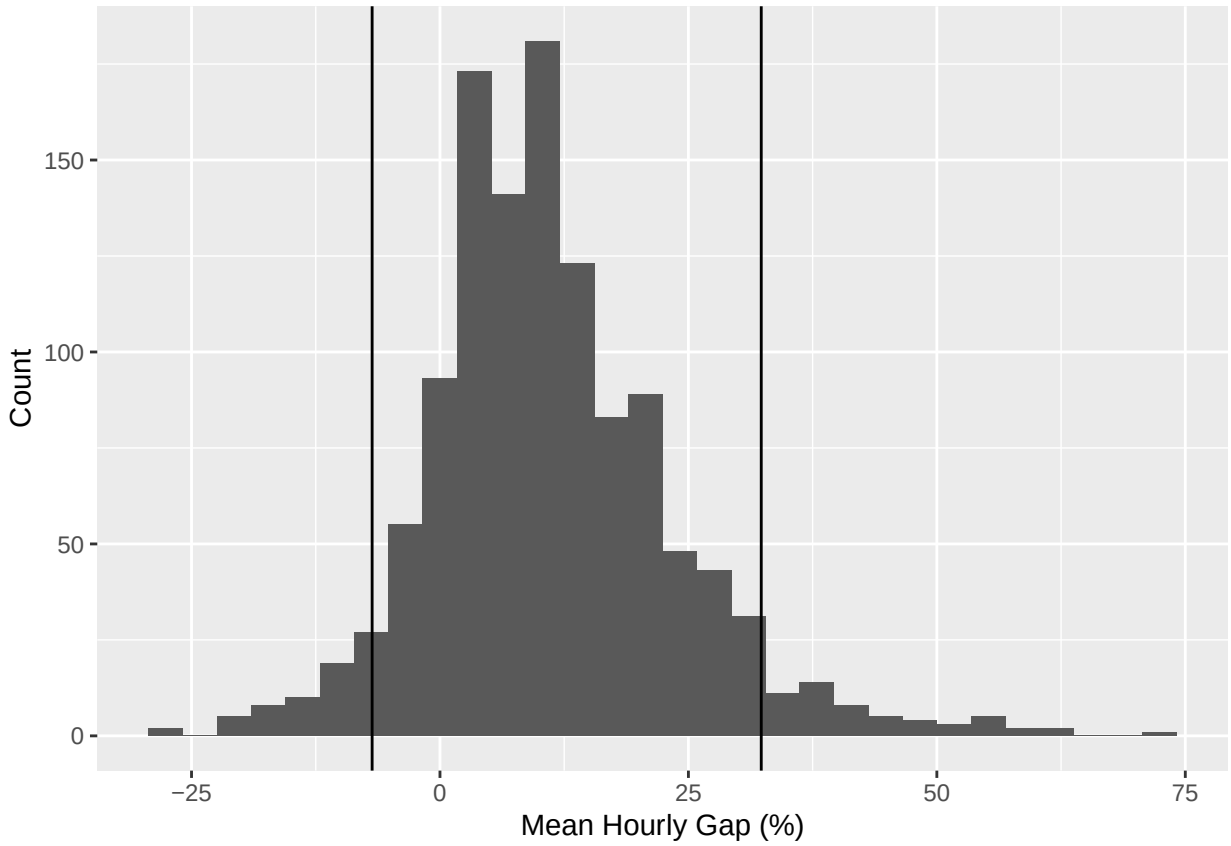
```
max(pg$`Mean Hourly Gap`, na.rm = TRUE)
```

```
## [1] 73
```

```
# individual histograms can be made like this:
ggplot(pg) + geom_histogram(aes(`Mean Hourly Gap`)) +
  labs(x='Mean Hourly Gap (%)', y='Count') +
  geom_vline(aes(xintercept=quantile(pg$`Mean Hourly Gap`,0.05,na.rm=TRUE))) +
  geom_vline(aes(xintercept=quantile(pg$`Mean Hourly Gap`,0.95,na.rm=TRUE)))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
# to identify outliers:
pg %>% arrange(`Mean Hourly Gap`) %>% select(`Company Name`,`ICB Industry`,`Mean Hourly Gap`)
```

```
## # A tibble: 1,189 x 3
##    `Company Name`                    `ICB Industry`         `Mean Hourly Gap`
##    <chr>                             <chr>                             <dbl>
##  1 Google Cloud EMEA Limited         Technology                          -27
##  2 Google Cloud EMEA Limited         Technology                          -27
##  3 Public Appointments Service (PAS) Public                              -21.4
##  4 Public Appointments Service (PAS) Public                              -20.9
##  5 C&C Group Plc (all legal entities) Consumer Staples                   -19.5
##  6 Lifestyle Sports                  Consumer Discretionary              -19.3
##  7 Lifestyle Sports                  Consumer Discretionary              -19.3
##  8 Galco Group                       Industrials                         -18.8
##  9 FMI                               Industrials                         -18.7
## 10 Stripe                            Technology                          -18.1
## # i 1,179 more rows
```

```
pg %>% arrange(desc(`Mean Hourly Gap`)) %>% select(`Company Name`,`ICB Industry`,`Mean Hourly Gap`)
```

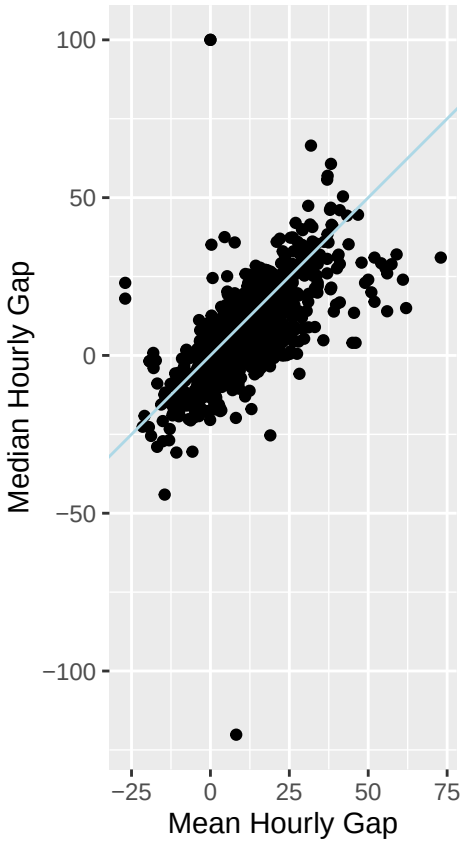
```
## # A tibble: 1,189 x 3
##    `Company Name`                                `ICB Industry` `Mean Hourly Gap`
##    <chr>                                         <chr>                     <dbl>
##  1 Matheson: Support Services (including partn~ Industrials                73
##  2 Matheson: Matheson LLP (including partners)  Consumer Disc~             62
##  3 Matheson: Matheson LLP (including partners)  Consumer Disc~             61
##  4 CRH: CRH Group Services Limited              Industrials                59
##  5 SIG Susquehanna International Group          Financials                 57.4
```

```
##  6 AerCap                                  Industrials              56
##  7 Arthur Cox (including partners)         Consumer Disc~           56
##  8 Mason Hayes & Curran (incl. partners)   Consumer Disc~           56
##  9 CRH: CRH Group Services Limited         Industrials              55.5
## 10 SIG Susquehanna International Group     Financials               54.1
## # i 1,179 more rows
```

(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(pg) +
  geom_point(aes(`Mean Hourly Gap`,`Median Hourly Gap`)) +
  geom_abline(aes(slope=1,intercept=0),color='lightblue') +
  coord_fixed(ratio=1)
```

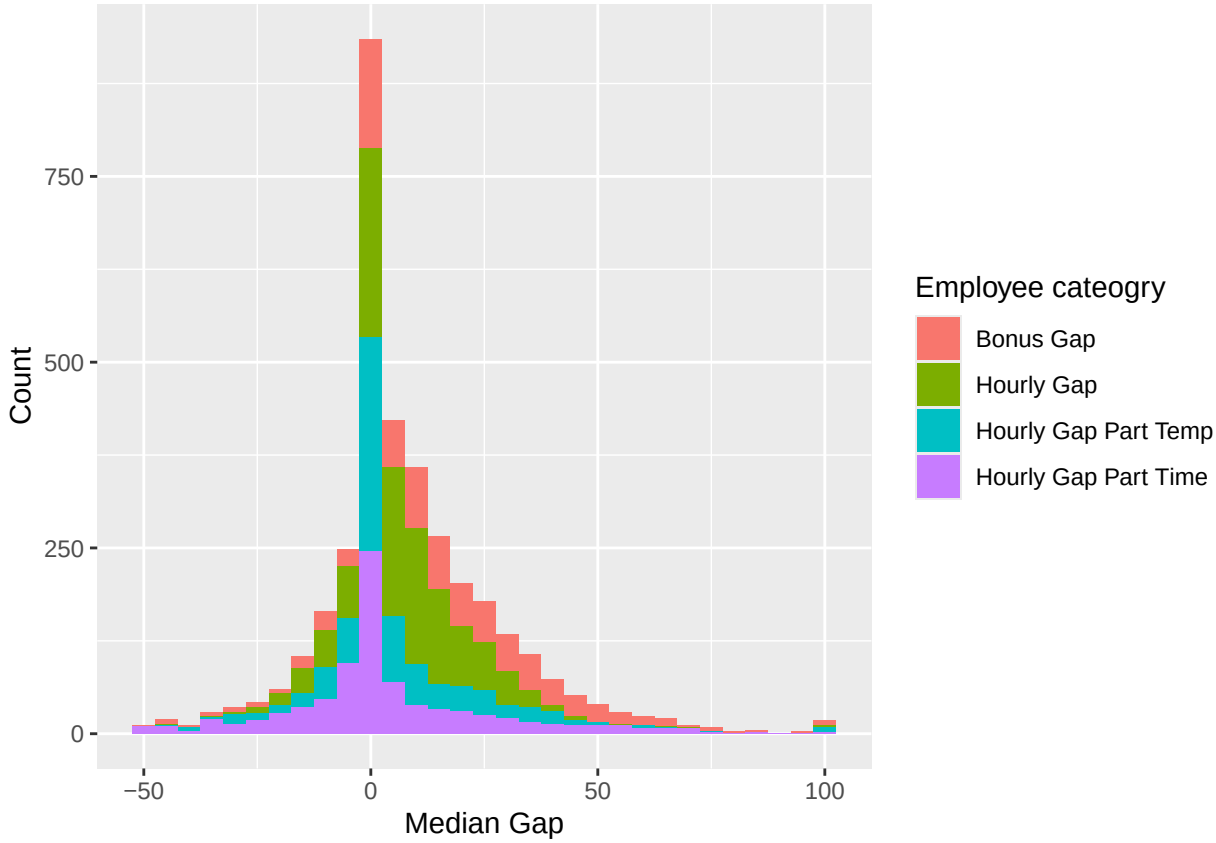
```
## Warning: Removed 5 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



(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 three distributions together on one set of axes.

```
# To put the employee categories together in one plot, we need to re-tidy:
pg %>% pivot_longer(cols=starts_with('Median'),
                    names_to='category',
```

```
                    values_to='median_gap',
                    names_prefix='Median ') -> pg2
pg2 %>% filter(median_gap > -50) %>%    # filter a small number of very negative pay gaps
  ggplot() +
  geom_histogram(aes(median_gap,fill=category),binwidth=5) +
  labs(fill='Employee cateogry', x='Median Gap', y='Count')
```



(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.

```
pg %>% group_by(`ICB Industry`,`Report Year`) %>%
  summarise(gap = median(`Median Hourly Gap`,na.rm=TRUE)) %>%
  arrange(desc(gap)) -> pg_sector_year
```

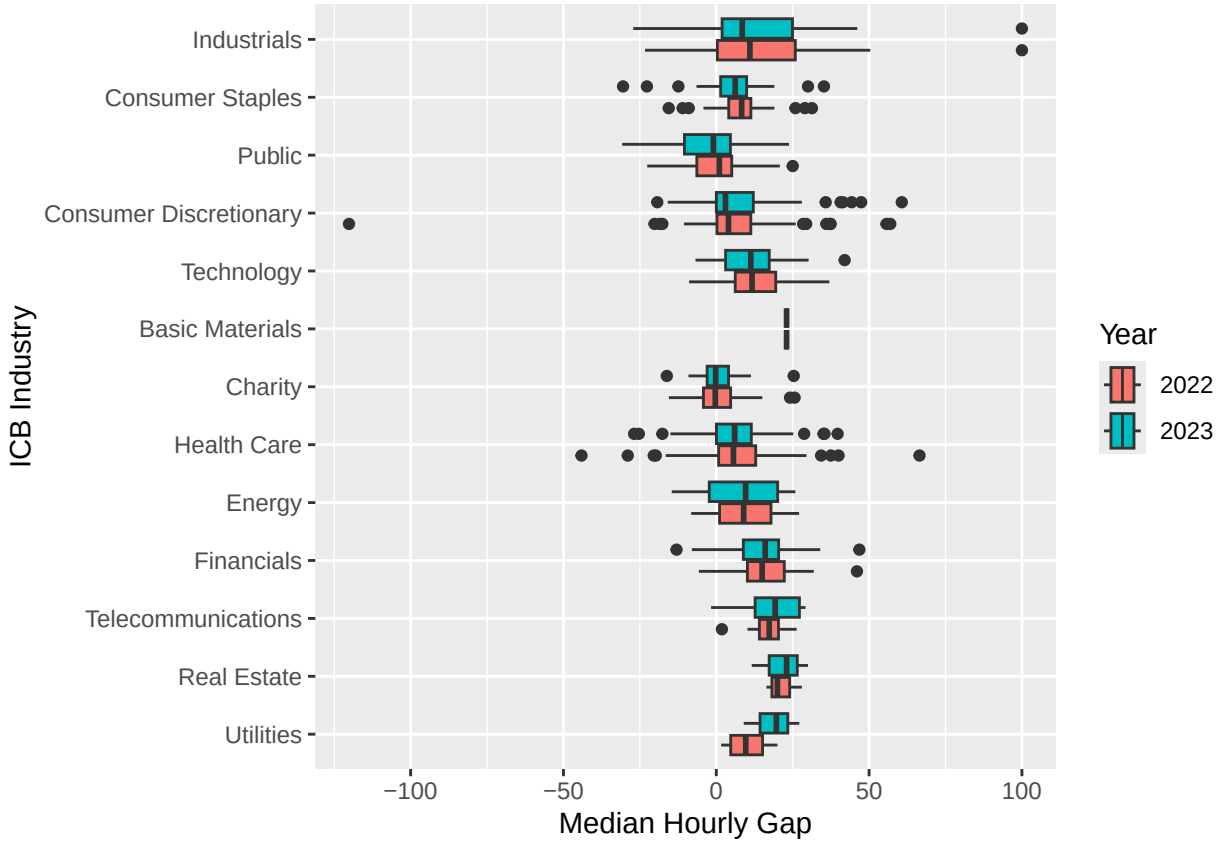
```
## `summarise()` has grouped output by 'ICB Industry'. You can override using the
## `.groups` argument.
```

```
# this contains the required summary values, but to put the table in the requested shape,
# we actually need to untidy it:
pg_sector_year %>% pivot_wider(names_from=`Report Year`,
                               values_from=gap) %>%
  # and once we have this, we can add the difference column with mutate()
  mutate(diff = `2023` - `2022`) %>%
  arrange(desc(diff)) -> pg_sector_year
pg_sector_year
```

```
## # A tibble: 13 x 4
## # Groups:   ICB Industry [13]
##    `ICB Industry`          `2022` `2023`   diff
##    <chr>                    <dbl>  <dbl>  <dbl>
##  1 Utilities                 9.64  19.7  10.0
##  2 Real Estate              20     23     3
##  3 Telecommunications       17.3   19.3   1.94
##  4 Financials               15     16     1
##  5 Energy                    8.95   9.6   0.650
##  6 Health Care               5.61   6     0.390
##  7 Charity                  -0.33  -0.2   0.13
##  8 Basic Materials          23     23     0
##  9 Technology               11.7   11.3  -0.400
## 10 Consumer Discretionary    4      3    -1
## 11 Public                    1     -0.96 -1.96
## 12 Consumer Staples          8.28   6.2  -2.09
## 13 Industrials              11      8.5  -2.5
```

```
# let's make boxplots plots to show the distribution of median hourly gap in 2022 and 2023,
# by sector -- ordering it by the 2023 - 2022 difference in medians
pg %>% ggplot() +
  geom_boxplot(aes(`Median Hourly Gap`,
                   factor(`ICB Industry`,levels = pg_sector_year$`ICB Industry`),
                   fill=factor(`Report Year`))) +
  labs(x='Median Hourly Gap', y='ICB Industry', fill='Year')
```

```
## Warning: Removed 3 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



(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) {
  # q1...q4 are the percentage of female employees at 4 quartiles of individual pay,
  # in one company
  avg = mean(c(q1, q2, q3, q4), na.rm=TRUE)
  return (1 - q4 / avg)
}

# tests:
# highly biased company: expect a high value
lwtp(100,75,25,0)
```

```
## [1] 1
```

```
# completely egalitarian company: expect a minimuum value
lwtp(50,50,50,50)
```

```
## [1] 0
```

```
# edge cases: what if the company has no women at all?
lwtp(0,0,0,0)
```

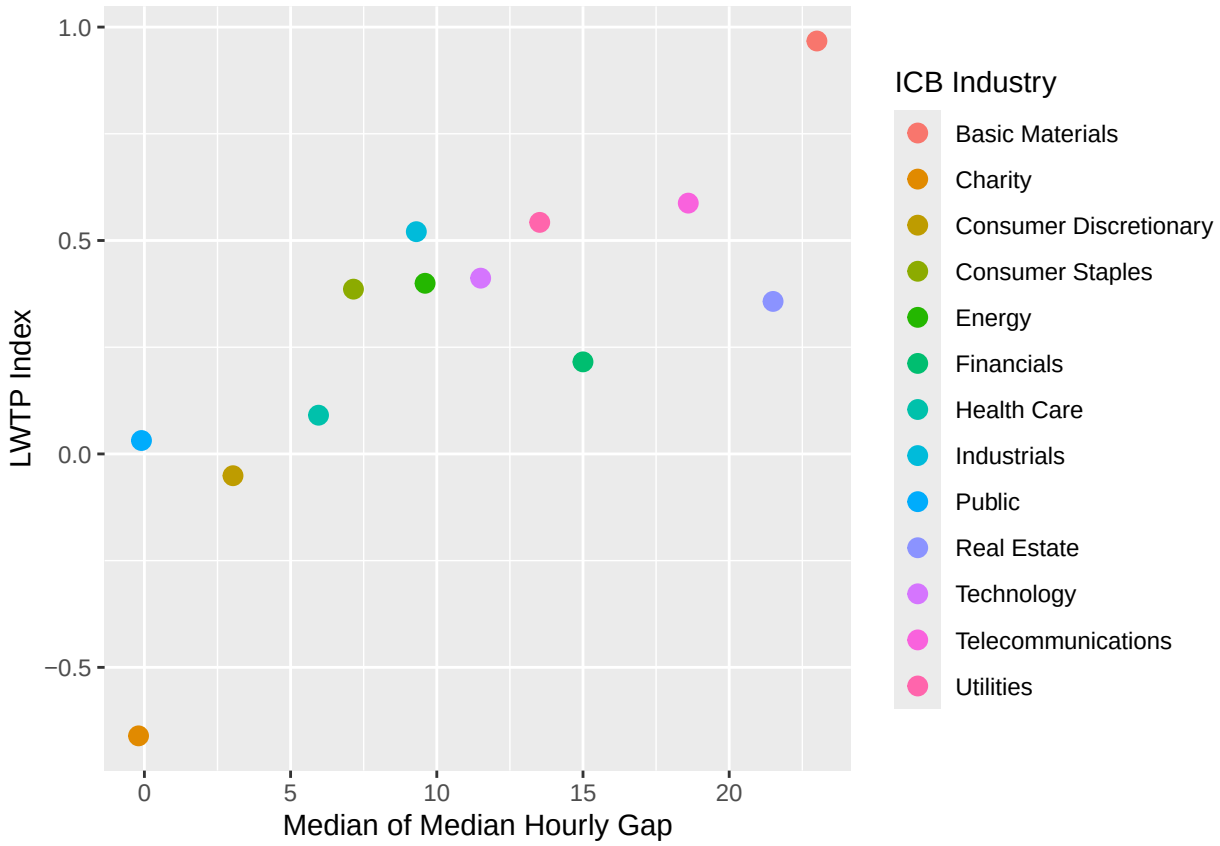
```
## [1] NaN
```

```
# how does the function handle missing data?
lwtp(70,50,NA,30)
```

```
## [1] 0.4
```

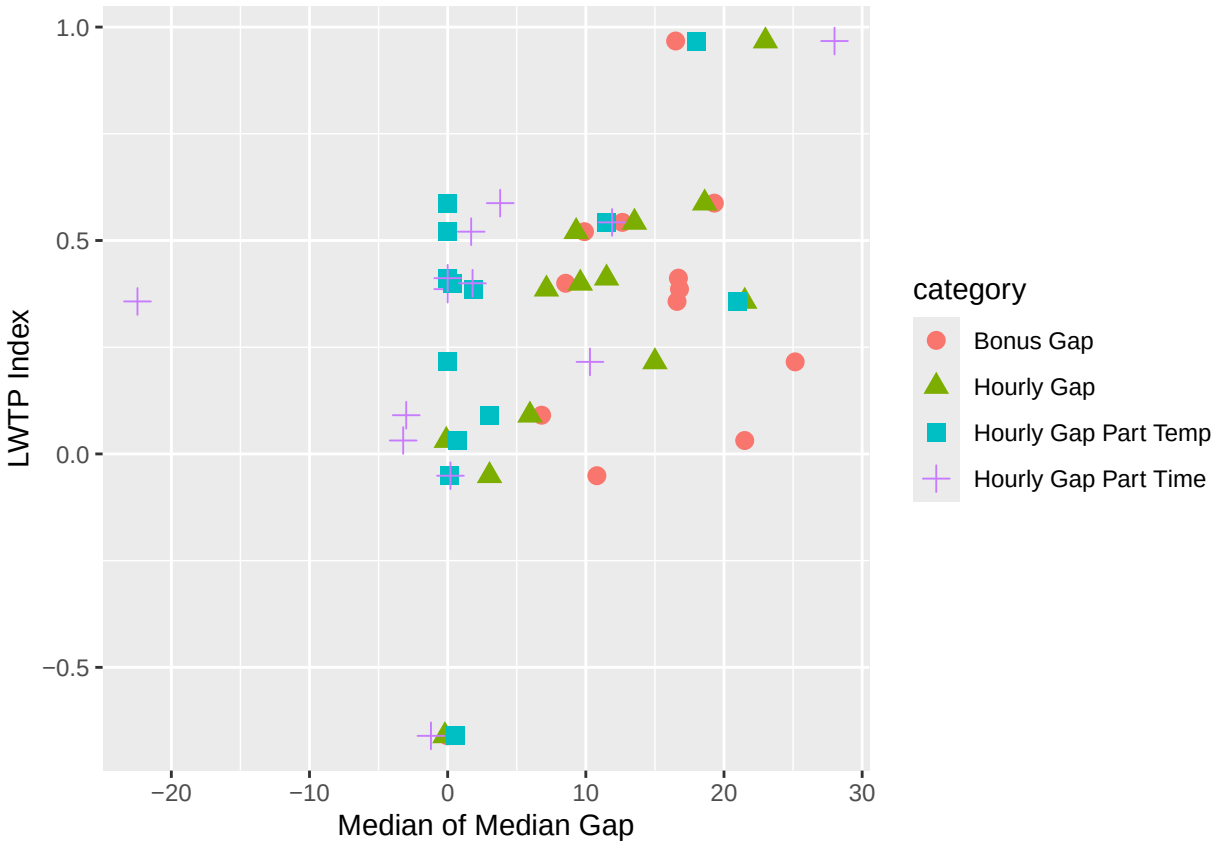
(7) LWTP vs. pay gaps. Now calculate your LWTP index for every company, and summarise the results by industrial sector. Does variation in LWTP seem to be related to pay gaps? There are many ways to approach this. Include the tables or plots that you use to answer the question.

```
# to compare median hourly gap with LWTP, by sector:
pg %>% mutate(lwtp = lwtp(`Q1 Female`,`Q2 Female`, `Q3 Female`, `Q4 Female`)) %>%
  group_by(`ICB Industry`) %>%
  summarise(gap=median(`Median Hourly Gap`,na.rm=TRUE),
            lwtp = median(lwtp,na.rm=TRUE)) %>%
  ggplot() + geom_point(aes(gap,lwtp, color=`ICB Industry`),size=3) +
  labs(x='Median of Median Hourly Gap', y='LWTP Index')
```




```
# this version does the same but for all 4 of the median gap measures we looked at before
# (hourly, part time, temp, bonus)
pg %>% pivot_longer(cols=starts_with('Median'),
                    names_to='category',
                    values_to='median_gap',
                    names_prefix='Median ') %>%
  mutate(lwtp = lwtp(`Q1 Female`,`Q2 Female`, `Q3 Female`, `Q4 Female`)) %>%
  group_by(`ICB Industry`,category) %>%
  summarise(gap = median(median_gap,na.rm=TRUE),
            lwtp = median(lwtp,na.rm=TRUE)) %>%
  ggplot() + geom_point(aes(gap,lwtp, color=category, shape=category),size=3) +
  labs(x='Median of Median Gap', y='LWTP Index')
```

```
## `summarise()` has grouped output by 'ICB Industry'. You can override using the
## `.groups` argument.
```



(8) Summary and recommendations. Conclude your report with a written section that summarises hypotheses you have formed that should be followed up with careful statistical tests, as well as commenting on data limitations, and additional datasets that you think would be useful for interpreting the patterns you saw.