

**1. How many companies are in the data set?**

```
SELECT COUNT(DISTINCT(employerid))  
FROM gender_pay_gap_21_22
```

Answer: 10174

**2. How many of them submitted their data after the reporting deadline?**

```
SELECT COUNT(employerid) AS employerlatesubmission, DateSubmitted  
FROM gender_pay_gap_21_22  
WHERE DateSubmitted > Duedate  
GROUP BY 2  
ORDER BY 2
```

Answer: 608

**3. How many companies have not provided a URL?**

```
SELECT COUNT(employerid) AS no_url  
FROM gender_pay_gap_21_22  
WHERE CompanyLinkToGPGInfo = '0'
```

Answer: 3700

**4. Which measures of pay gap contain too much missing data, and should not be used in our analysis?**

```
SELECT COUNT(employerid) AS missingquartile  
FROM gender_pay_gap_21_22  
WHERE maleuppermiddlequartile = '0' AND femaleuppermiddlequartile = '0'
```

*The assumption is that 0 is equal to a NULL value. SQL database has converted these null values to zeros. I have checked this via excel*

The male and female quartile information is missing for over 185 companies

**Bonus (optional): Can you find out what the 'SicCodes' column corresponds to? Is there a way we can understand what each SIC code represents? Search online for extra information.**

SIC Codes correspond to industries.

**5. Choose which column you will use to calculate the pay gap. Will you use DiffMeanHourlyPercent or DiffMedianHourlyPercent? Can you justify your choice?**

I will use DiffMeanHourlyPercent

I have compared histograms of the data and believe the data has a central tendency of both the mean and the median data of around 20. The standard deviation of the mean and medians are quite similar/close to each other's values. I will make an assumption that based on the above information there are no significant outliers

```
SELECT stddev(DiffMeanHourlyPercent ) AS standard_deviation_DiffMean
FROM gender_pay_gap_21_22
= 14.8652706968403290
```

```
SELECT stddev(DiffMedianHourlyPercent ) AS standard_deviation_DiffMedian
FROM gender_pay_gap_21_22
= 16.1877603602867295
```

**6. Use an appropriate metric to find the average gender pay gap across all the companies in the data set. Did you use the mean or the median as your averaging metric? Can you justify your choice?**

```
SELECT ROUND(avg(diffmeanhourlypercent),2) as avg_pay_gap
FROM gender_pay_gap_21_22
```

ANSWER : 13.64%

I chose the mean as the averaging metric. There is no median function in Postgres

**7. What are some caveats we need to be aware of when reporting the figure we've just calculated?**

It's important to consider that taking the average of averages may skew the results. Ideally, we would want to take the average of all the raw data, but that would be very difficult to do for this analysis.

**8. What are the 10 companies with the largest pay gaps skewed towards men?**

```
SELECT EmployerId,EmployerName,DiffMeanHourlyPercent
FROM gender_pay_gap_21_22
Order by 3
DESC
LIMIT 10
```

	employerid character varying	employername character varying	diffmeanhourlypercent numeric
1	6357	HPI UK HOLDING LTD.	100
2	19413	PSJ FABRICATIONS LTD	100
3	20315	M. ANDERSON CONSTRUCTION LIMITED	100.0
4	15371	BIRMINGHAM CITY FOOTBALL CLUB PLC	99
5	883	ACUSHNET EUROPE LTD	96.8
6	6290	HOOK 2 SISTERS LIMITED	92.0
7	3148	CHELSEA FOOTBALL CLUB LIMITED	91.6
8	2321	BRAND ENERGY & INFRASTRUCTURE SERVICES UK, L...	91.0
9	8049	MANCHESTER CITY FOOTBALL CLUB LIMITED	91
10	8911	NEWCASTLE UNITED FOOTBALL COMPANY LIMITED	90.4

### 9. What do you notice about the results? Are these well-known companies?

40% of these companies are well-known professional male football clubs. We can assume that the clubs consist of male players which would skew the results towards males

### 10. Apply some additional filtering to pick out the most significant companies with large pay gaps.

```
SELECT EmployerId,EmployerName,DiffMeanHourlyPercent
FROM gender_pay_gap_21_22
Order by 3
ASC
LIMIT 10
```

The company within this filter are heavily skewed towards women, with Fortel Services at a -184%

```
SELECT EmployerId,EmployerName, diffmeanbonuspercent
FROM gender_pay_gap_21_22
WHERE diffmeanbonuspercent != 0
Order by 3 DESC
LIMIT 10
```

Companies with 100% diffmeanbonuspercent also consist of male football clubs and Abingdon School, which is an all boys school. We can assume that the composition of staff is made up of males more than females

### 11. How would you report on the results? Can we say that these companies are engaging in unlawful pay discrimination?

I don't think I would report on these general results. It doesn't take into consideration the industries where it is male dominated because of the nature of their work. I do not believe these companies are engaging in unlawful pay discrimination because the majority of their workforce is likely largely male. For example, the Chelsea Football Club must only employ male players due to the rules of the league they play within.

**12. What's the average pay gap in London versus outside London?**

In London: 15.68%

Outside London: 13.04%

**QUERY 1:**

```
WITH location AS (
  SELECT
    employerid,
    employername,
    address,
    CASE
      WHEN address ILIKE '%London%' THEN 'london'
      ELSE 'outside_london'
    END AS london_location
  FROM
    gender_pay_gap_21_22
  GROUP BY
    employerid, employername, address
)

SELECT
  london_location,
  COUNT(gpg.employerid) AS employer_count,
  AVG(diffmeanhourlypercent) AS london_pay_gap_average
FROM location
JOIN gender_pay_gap_21_22 gpg ON location.employerid = gpg.employerid
WHERE london_location = 'london'
GROUP BY london_location;
```

**QUERY 2:**

```
WITH location AS (
  SELECT
    employerid,
    employername,
    address,
    CASE
      WHEN address ILIKE '%London%' THEN 'london'
      ELSE 'outside london'
    END AS london_location
  FROM
    gender_pay_gap_21_22
  GROUP BY
    employerid, employername, address
```

)

```
SELECT
    london_location,
    COUNT(gpg.employerid) AS employer_count,
    AVG(diffmeanhourlypercent) AS london_pay_gap_average
FROM location
JOIN gender_pay_gap_21_22 gpg ON location.employerid = gpg.employerid
WHERE london_location = 'outside london'
GROUP BY london_location;
```

### 13. What's the average pay gap in London versus Birmingham?

Birmingham Pay gap average: 13.21%

In London: 15.68%

```
WITH location AS (
    SELECT
        employerid,
        employername,
        address,
        CASE
            WHEN address ILIKE '%London%' THEN 'london'
            WHEN address ILIKE '%Birmingham%' THEN 'birmingham'
            ELSE 'neither location'
        END AS region_location
    FROM
        gender_pay_gap_21_22
    GROUP BY
        employerid, employername, address
    ORDER BY 4 ASC
)
```

```
SELECT
    region_location,
    COUNT(l.employerid) AS employer_count,
    AVG(diffmeanhourlypercent) AS birmingham_pay_gap_average
FROM location l
JOIN gender_pay_gap_21_22 gpg ON l.employerid = gpg.employerid
WHERE region_location = 'birmingham'
GROUP BY region_location
```

**14. What is the average pay gap within schools?**

Average pay gap for schools: 16.58%

This was calculated using the following SIC Codes:

- 85100 Pre-primary education (such as nurseries)
- 85200 Primary education
- 85310 General secondary education
- 85320 Technical and vocational secondary education
- 85410 Post secondary non-tertiary education
- 85421 First-degree level higher education (such as undergraduate)
- 85422 Postgraduate level higher education (such as Master's programmes)

```
WITH education_industry AS
(
SELECT
    employerid,
    EmployerName,
    SicCodes
FROM gender_pay_gap_21_22
WHERE SicCodes ILIKE '%85100%' OR
SicCodes ILIKE '%85200%' OR
SicCodes ILIKE '%85310%' OR
SicCodes ILIKE '%85320%' OR
SicCodes ILIKE '%85410%' OR
SicCodes ILIKE '%85421%' OR
SicCodes ILIKE '%85422%'
Group by 1,2,3
)
SELECT AVG(gpg.diffmeanhourlypercent) AS avg_pay_gap_education
FROM education_industry ei
JOIN gender_pay_gap_21_22 gpg ON ei.employerid = gpg.employerid
```

**15. What is the average pay gap within banks?**

Average pay gap for banks: 31.68%

Assuming that banks are grouped by their SicCodes 64110 (Banks) or 64191 (Central banking)

```
WITH banking_industry AS (
SELECT DISTINCT
    employerid,
    EmployerName,
    SicCodes
FROM gender_pay_gap_21_22
```

```
WHERE SicCodes ILIKE '%64110%' OR SicCodes ILIKE '%64191%'
)
```

```
SELECT AVG(gpg.diffmeanhourlypercent) AS avg_pay_gap_banking
FROM banking_industry bi
JOIN gender_pay_gap_21_22 gpg ON bi.employerid = gpg.employerid
```

**16. Is there a relationship between the number of employees at a company and the average pay gap?**

20,000 or more: avg = 12.48% across 62 companies  
 5000 to 19,999: avg = 14.12% across 464 companies  
 1000 to 4999: avg = 12.90% across 2131 companies  
 500 to 999: avg = 13.66% across 2501  
 250 to 499: avg = 13.92% across 4274  
 Less than 250: avg = 14.05% across 532  
 Not provided: avg = 13.03% across 210

```
SELECT avg(diffmeanhourlypercent), COUNT(employerid)
FROM gender_pay_gap_21_22
WHERE Employersize = '20,000 or more'
```

```
SELECT avg(diffmeanhourlypercent), COUNT(employerid)
FROM gender_pay_gap_21_22
WHERE Employersize = '5000 to 19,999'
```

```
SELECT avg(diffmeanhourlypercent), COUNT(employerid)
FROM gender_pay_gap_21_22
WHERE Employersize = '1000 to 4999'
```

```
SELECT avg(diffmeanhourlypercent), COUNT(employerid)
FROM gender_pay_gap_21_22
WHERE Employersize = '500 to 999'
```



```
SELECT avg(diffmeanhourlypercent), COUNT(employerid)
FROM gender_pay_gap_21_22
WHERE Employersize = '250 to 499'
```

```
SELECT avg(diffmeanhourlypercent), COUNT(employerid)
FROM gender_pay_gap_21_22
WHERE Employersize = 'Less than 250'
```

```
SELECT avg(diffmeanhourlypercent), COUNT(employerid)
FROM gender_pay_gap_21_22
WHERE Employersize = 'Not Provided'
```

**Extra:**

I wanted to look at the gender distribution across quartiles, but for the sake of time, I looked just at the top quartile distribution as a sample

gender_distribution_status 	status_count 
Female Dominant	3567
Equal Distribution	273
Male Dominant	6334

```

WITH top_quartile_status AS(
SELECT
  employerid,
  femaletopquartile,
  maletopquartile,
  CASE
    WHEN maletopquartile > femaletopquartile THEN 'Male Dominant'
    WHEN maletopquartile < femaletopquartile THEN 'Female Dominant'
    ELSE 'Equal Distribution'
  END AS gender_distribution_status
FROM
  gender_pay_gap_21_22)

SELECT
  gender_distribution_status,
  COUNT(gender_distribution_status) AS status_count
FROM
  top_quartile_status tq
JOIN
  gender_pay_gap_21_22 gpg ON tq.employerid = gpg.employerid
GROUP BY
  gender_distribution_status

```