```
--Understanding the table
        -- PRAGMA table_info(UK_Gender_Pay_Gap_Data_2023_2024);

-- 1. How many companies are in the data set?
        10,852


CREATE VIEW '2023_2024_clean' AS

 WITH a (ID, name) AS(
        SELECT
                EmployerId,
                EmployerName
        FROM UK_Gender_Pay_Gap_Data_2023_2024
        WHERE EmployerId != '23038' AND EmployerId != '5269' AND EmployerId != 16169
        AND EmployerId != 23200 AND EmployerId != 20186 AND EmployerId != 21148 AND
        EmployerId != 14427 AND EmployerId != 22619),

 b (ID, late) AS(
        SELECT
                EmployerID,
                CASE
                        WHEN DateSubmitted > DueDate THEN 'True'
                        WHEN DateSubmitted <= DueDate THEN 'False'
                END
        FROM UK_Gender_Pay_Gap_Data_2023_2024)

        SELECT
                a.name AS Employer_Name,
                a.ID AS Employer_ID,
                Address,
                PostCode,
                CompanyNumber,
                SicCodes,
                DiffMeanHourlyPercent,
                DiffMedianHourlyPercent,
                DiffMeanBonusPercent,
                DiffMedianBonusPercent,
                MaleBonusPercent,
                FemaleBonusPercent,
                MaleLowerQuartile,
                FemaleLowerQuartile,
                MaleLowerMiddleQuartile,
                FemaleLowerMiddleQuartile,
```

```
        MaleUpperMiddleQuartile,
        FemaleUpperMiddleQuartile,
        MaleTopQuartile,
        FemaleTopQuartile,
        CompanyLinkToGPGInfo,
        ResponsiblePerson,
        EmployerSize,
        CurrentName,
        b.late AS Submitted_Late,
        DueDate,
        DateSubmitted
    FROM UK_Gender_Pay_Gap_Data_2023_2024
    JOIN a
    ON UK_Gender_Pay_Gap_Data_2023_2024.EmployerID = a.ID
    JOIN b
    ON UK_Gender_Pay_Gap_Data_2023_2024.EmployerID = b.ID;
```

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

765

```
SELECT count(*) FROM '2023_2024_clean' WHERE Submitted_Late = 'True';
```

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

4,869

```
SELECT DISTINCT COUNT(*)
FROM '2023_2024_clean'
WHERE CompanyLinkToGPGInfo IS null;
```

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

No measures of pay gap are missing data except for DiffMeanBonusPercent and DiffMedianBonusPercent. In those cases, NULL is a result of one gender not having

been paid a bonus, thus this information is still relevant and will be included in the analysis.

```
SELECT
	COUNT(DiffMeanHourlyPercent),
	COUNT(DiffMedianHourlyPercent),
	COUNT(DiffMeanBonusPercent),
	COUNT(DiffMedianBonusPercent),
	COUNT(MaleBonusPercent),
	COUNT(FemaleBonusPercent),
	COUNT(MaleLowerMiddleQuartile),
	COUNT(FemaleLowerMiddleQuartile),
	COUNT(MaleUpperMiddleQuartile),
	COUNT(FemaleUpperMiddleQuartile),
	COUNT(MaleTopQuartile),
	COUNT(FemaleTopQuartile)
FROM UK_Gender_Pay_Gap_Data_2023_2024;
```

```
SELECT
	EmployerName,
	DiffMeanBonusPercent,
	MaleBonusPercent,
	FemaleBonusPercent
FROM UK_Gender_Pay_Gap_Data_2023_2024
WHERE DiffMeanBonusPercent IS null;
```

-- BONUS: SicCodes = List of comma-separated SIC codes used to describe the employer's purpose and sectors of work at the time of reporting

-- 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 because the data is largely continuous between -100 and 100. The only employers outside of those bounds (6) have been removed. In this case as well, I want to capture all data points to give an overall picture of the gender pay gap. The mean and median are very similar (12.62% vs. 11.59%).

-- 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?

-- On average, companies in the UK pay male employees 12.62% more than they pay women.

-- Of companies that had a bias towards men in their average hourly pay (86%), the average bias was 15.68%.

-- Of companies that had a bias towards women in their average hourly pay (13%), the average bias was only 7.02%.

-- Less than 1% of companies analyzed had no bias.

```
SELECT avg(DiffMeanHourlyPercent) FROM '2023_2024_clean';


SELECT avg(DiffMedianHourlyPercent) FROM '2023_2024_clean';


CREATE VIEW bias AS
	SELECT
		Employer_Name,
		CASE
			WHEN DiffMeanHourlyPercent > 0 THEN 'male'
			WHEN DiffMeanHourlyPercent < 0 THEN 'female'
			WHEN DiffMeanHourlyPercent = 0 THEN 'no bias'
		END AS Bias
	FROM '2023_2024_clean';


SELECT
	count(*),
	avg(data.DiffMeanHourlyPercent),
	avg(data.DiffMedianHourlyPercent)

FROM '2023_2024_clean' data
JOIN bias
ON bias.Employer_Name = data.Employer_Name

WHERE bias = 'male';


SELECT
	count(*),
	abs(avg(data.DiffMeanHourlyPercent)),
```

```
        abs(avg(data.DiffMedianHourlyPercent))

 FROM '2023_2024_clean' data
 JOIN bias
 ON bias.Employer_Name = data.Employer_Name

 WHERE bias = 'female';


select count(*) from '2023_2024_clean' where DiffMeanHourlyPercent BETWEEN -1 AND 1;
```

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

-Caveat: This data is self-reported and the calculations are done by each company themselves. This means the information is unverified, is prone to errors, and may be duplicative in some places.

-Caveat: This data only represents companies with at least 250 employees.

-Caveat: Position information is not reported.

-Caveat: The latest duedate was April 5, 2024, so this data only captures companies who reported on time or about a month late. In 21-22 the max datesubmitted was April 11, 2022 and in 22-23 the max datesubmitted was May 13,2023. So we can be reasonably confident the majority of companies who are going to submit, have submitted.

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

1. FOX RECRUITMENT LINCS LTD (Other activities of employment placement agencies)
2. AIR PRODUCTS PUBLIC LIMITED COMPANY (Manufacture of industrial gases)
3. AFC BOURNEMOUTH LIMITED (Activities of sport clubs)
4. JOINERY AND TIMBER CREATIONS (65) LIMITED (Manufacture of kitchen & other furniture)
5. CHELSEA FOOTBALL CLUB LIMITED (Activities of sport clubs)
6. WEST HAM UNITED FOOTBALL CLUB LIMITED (Activities of sport clubs)
7. STOCKS HALL CARE HOMES LIMITED (Residential nursing care facilities)
8. BRENTFORD FC LIMITED (Activities of sport clubs)
9. LEEDS UNITED FOOTBALL CLUB LIMITED (Operation of sports facilities)

10. NOTTINGHAM FOREST FOOTBALL CLUB LIMITED (Activities of sport clubs)

```
SELECT
        data.Employer_Name,
        data.SicCodes,
        EmployerSize,
        ResponsiblePerson,
        data.DiffMeanHourlyPercent,
        data.DiffMeanBonusPercent,
        data.MaleBonusPercent,
        data.FemaleBonusPercent,
        data.MaleLowerQuartile,
        data.FemaleLowerQuartile,
        data.MaleLowerMiddleQuartile,
        data.FemaleLowerMiddleQuartile,
        data.MaleUpperMiddleQuartile,
        data.FemaleUpperMiddleQuartile,
        data.MaleTopQuartile,
        data.FemaleTopQuartile

FROM '2023_2024_clean' data
JOIN bias
ON bias.Employer_Name = data.Employer_Name

WHERE bias.bias = 'male'

ORDER BY DiffMeanHourlyPercent DESC

LIMIT 10;
```

-- 9. What do you notice about the results? Are these well-known companies?
60% of them are football clubs.

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

Company	*Mean Bias*
Google	*13.0%*
Netflix	*11.1%*
Apple	*9.0%*
Amazon	*8.6%*
Facebook	*0.2%*

```
WITH FAANG AS(
SELECT
	Employer_name,
	EmployerSize,
	DiffMeanHourlyPercent,
	DiffMedianHourlyPercent
FROM '2023_2024_clean'
WHERE Employer_name LIKE '%Amazon%' OR Employer_name LIKE '%Facebook%'  OR
Employer_name LIKE '%Apple%' AND Employer_name != 'MCLEAN & APPLETON
(HOLDINGS) LIMITED' OR Employer_name LIKE '%Netflix%' OR Employer_name LIKE
'%Google%'
)

SELECT
	CASE
		WHEN Employer_name LIKE '%Amazon%' THEN 'Amazon'
		WHEN Employer_name LIKE '%Facebook%' THEN 'Facebook'
		WHEN Employer_name LIKE '%Apple %' THEN 'Apple'
		WHEN Employer_name LIKE '%Netflix%' THEN 'Netflix'
		WHEN Employer_name LIKE '%Google%' THEN 'Google'
		ELSE 'n/a'
	END AS FAANG,
	Round(Avg(DiffMeanHourlyPercent),1) as 'Mean Bias',
	Round(Avg(DiffMedianHourlyPercent),1) as 'Median Bias'
FROM FAANG
GROUP BY FAANG
ORDER BY 'Median Bias' ASC;
```

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

TWO BIG CAVEATS:

1. *The data has been self-calculated and self-reported by the companies.*

   This means the information is unverified, is prone to errors, and may be duplicative in some places.

2. *Position information is not required as part of the reporting.*

   This means that we cannot conclude whether a company is paying a man and a woman different wages for the same work, which would be unlawful.

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

One way to improve data collection is to create standardized fields for address instead of it being free-form.

Of companies that provided an address:

London: 14.43%
Not London: 12.12%

```
SELECT count(*), avg(diffmeanhourlypercent) FROM '2023_2024_clean' WHERE address LIKE '%, London,%' OR address LIKE '%,London,%' OR address LIKE '% London,%' OR address LIKE '% London.%';
```

```
SELECT count(*), avg(diffmeanhourlypercent) FROM '2023_2024_clean' WHERE address NOT LIKE '%, London,%' AND address NOT LIKE '%,London,%' AND address NOT LIKE '% London,%' AND address NOT LIKE '% London.%';
```

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

Of companies that provided an address:

London: 14.43%
Birmingham: 11.86%

```
SELECT avg(diffmeanhourlypercent) FROM '2023_2024_clean' WHERE address LIKE '%, Birmingham,%'  OR address LIKE '% Birmingham,%'  OR address LIKE '%,Birmingham,%' OR address LIKE '% Birmingham.,%' OR address LIKE '%, Birmingham %';
```

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

Schools = pre-primary, primary, general secondary, technical and vocational secondary, post-secondary non-tertiary, first-degree level higher, post-graduate level higher, sports and recreation, cultural, other

Avg pay gap = 14.76%

```
SELECT employer_name, SicCodes, DiffMeanHourlyPercent FROM '2023_2024_clean' WHERE SicCodes IN (85200, 85100, 85310, 85320, 85410, 85421, 85422, 85510, 85520, 85590);
```

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

Banks = banks and central banking (bank of england)

Avg pay gap = 30.04%

```
SELECT avg(DiffMeanHourlyPercent) FROM '2023_2024_clean' WHERE SicCodes LIKE '%64110%' OR SicCodes LIKE '%64191%';
```

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

No

```
select
	EmployerSize,
	round(avg(diffmeanhourlypercent),2) as bias
 from '2023_2024_clean'
 group by employersize
 order by bias desc;
```

```
SELECT
	EmployerSize,
	DiffMeanHourlyPercent
FROM '2023_2024_clean'
ORDER BY EmployerSize;
```

ADDITIONAL ANALYSIS

```
CREATE VIEW SicCodes AS

WITH RECURSIVE cte AS(
	SELECT
		Employer_name,
		SicCodes,
		SicCodes + 0 col
	FROM
		'2023_2024_clean'
	UNION ALL
	SELECT
		Employer_name,
		SUBSTR(SicCodes, length(col) +2),
		SUBSTR(SicCodes,length(col) +2) + 0
	FROM cte
	WHERE instr(SicCodes, ',')
)

SELECT
	Employer_name,
	col SicCodes
 FROM cte
 ORDER BY Employer_Name;


 WITH a AS (
 SELECT
 s.siccodes,
 count(s.siccodes),
 round(avg(c.DiffMeanHourlyPercent),2) AS bias
 from SicCodes s
 join '2023_2024_clean' c
```

```
on s.employer_name = c.employer_name
group by s.SicCodes
having s.SicCodes != 0 AND s.SicCodes != 1 AND count(s.SicCodes) >= 23
order by bias DESC
limit 40)

SELECT
	CASE
		WHEN siccodes LIKE '93%' THEN 'sport'
		WHEN SicCodes LIKE '64%' OR SicCodes LIKE '66%' OR SicCodes LIKE '65%'
		THEN 'finance'
		WHEN SicCodes LIKE '68%' THEN 'real estate'
		WHEN SicCodes LIKE '41%' OR SicCodes LIKE '43%' THEN 'construction'
		WHEN SicCodes LIKE '51%' THEN 'air travel'
		WHEN SicCodes LIKE '91%' THEN 'gas mining'
		WHEN SicCodes LIKE '47%' OR SicCodes LIKE '46%' THEN 'cosmetics retail'
		WHEN SicCodes LIKE '33%' THEN 'equipment repair'
		WHEN SicCodes LIKE '75%' THEN 'veterinary services'
		WHEN SicCodes  LIKE '45%' THEN 'car sales'
		WHEN SicCodes LIKE '74%' THEN 'environmental consulting'
		WHEN SicCodes LIKE '69%' THEN 'solicitors'
			ELSE 'new industry'
	END AS 'industry',
	SicCodes,
	bias
FROM a
GROUP BY industry
ORDER by BIAS asc;


WITH a AS (
SELECT
	a.SicCodes,
	count(a.siccodes),
	avg(b.DiffMeanHourlyPercent) as Bias
FROM SicCodes a
JOIN '2023_2024_clean' b
ON a.Employer_name = b.Employer_name
GROUP BY a.siccodes
HAVING a.siccodes != 0 AND a.SicCodes !=1 AND count(a.siccodes) >= 23
ORDER by avg(b.diffmeanhourlypercent) ASC
LIMIT 30
)
```

```
SELECT
	CASE
		WHEN SicCodes LIKE '38%' THEN 'Waste Collection'
		WHEN SicCodes LIKE '80%' THEN 'Private Security'
		WHEN SicCodes LIKE '93%' THEN 'Fitness Facilities'
		WHEN SicCodes LIKE '56%' THEN 'Unlicenced Restaurants and Cafes'
		WHEN SicCodes LIKE '87%' OR SicCodes LIKE '88%' THEN 'Social Work'
		WHEN SicCodes LIKE '46%' THEN 'Building Materials Sales'
		WHEN SicCodes LIKE '29%' THEN 'Manufacture of Motor Vehicles'
		WHEN SicCodes LIKE '49%' OR SicCodes LIKE '52%' THEN 'Land Travel'
		WHEN SicCodes LIKE '84%' THEN 'Public Administration'
		WHEN SicCodes LIKE '90%' THEN 'Arts Facilities'
		ELSE 'new industry'
	END AS 'Industry',
	round(bias,2)
FROM a
GROUP BY industry
ORDER BY bias ASC;



SELECT
	Employer_name,
	address,
	DiffMeanHourlyPercent,
	avg(DiffMeanHourlyPercent) OVER (PARTITION BY SicCodes)
FROM
	'2023_2024_clean'
WHERE SicCodes = 93120
ORDER BY DiffMeanHourlyPercent ASC;


SELECT
	avg(old.DiffMeanHourlyPercent),
	avg(new.diffmeanhourlypercent),
from UK_Gender_Pay_Gap_Data_2022_2023 old
JOIN UK_Gender_Pay_Gap_Data_2023_2024 new
ON old.EmployerId = new.EmployerId;


WITH cte AS(
	SELECT
		Employer_name,
		SicCodes,
```

```
            CASE
                WHEN SicCodes IN (93110, 93120, 93199) THEN 'TRUE'
                ELSE 'FALSE'
            END AS 'FC'
        FROM SicCodes),

 cte2 AS(
 SELECT
        cte.Employer_name,
        cte.SicCodes,
        cte.FC,
        data.DiffMeanHourlyPercent
 FROM cte
 JOIN '2023_2024_clean' data
 ON cte.Employer_name = data.Employer_name
 WHERE FC = 'TRUE' AND cte.Employer_name LIKE '%FC%' OR cte.Employer_name LIKE
'%Football Club%')

 SELECT
        Employer_name,
        SicCodes,
        FC,
        DiffMeanHourlyPercent,
        avg(DiffMeanHourlyPercent) OVER ()
 FROM cte2
 WHERE Employer_name != 'CHELSEA FC FOUNDATION'
 GROUP BY Employer_name;


SELECT
        Employer_name,
        DiffMeanHourlyPercent,
        DiffMedianHourlyPercent,
        FemaleTopQuartile,
        MaleTopQuartile
FROM '2023_2024_clean'
WHERE FemaleTopQuartile > MaleTopQuartile;
```