Q1. How many companies are in the data set?

A1.

```
Query  Query History
1  -- Counting of unique companies in data set.
2  SELECT COUNT(DISTINCT employername) as Companies_Count
3  FROM public.gender_pay_gap_21_22
```

Data output | Messages | Notifications

	companies_count bigint
1	10174

There are a total of 10174 unique companies in the data set.

Q2. How many of them submitted their data after the reporting deadline?

A2.

```
Query  Query History
1  -- Determination of errors in column "submittedafterthedeadline"
2  SELECT datesubmitted, duedate, submittedafterthedeadline
3  FROM public.gender_pay_gap_21_22
4  WHERE datesubmitted > duedate AND submittedafterthedeadline = 'FALSE'
5
```

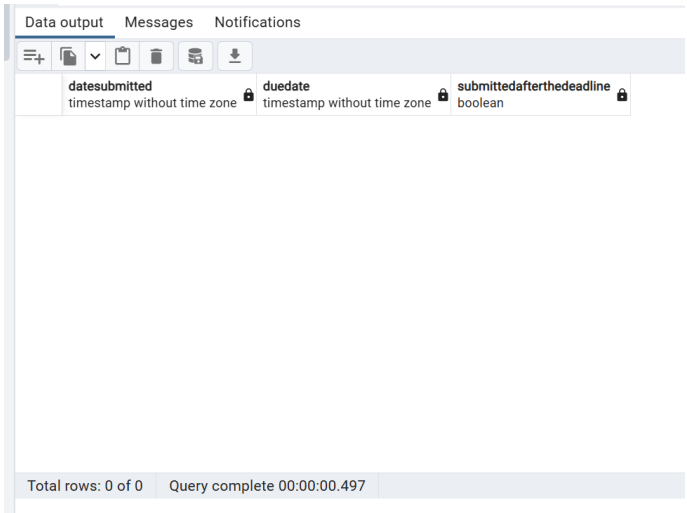
Data output | Messages | Notifications

	datesubmitted timestamp without time zone	duedate timestamp without time zone	submittedafterthedeadline boolean
1	2022-04-06 10:12:42	2022-04-05 00:00:00	false
2	2022-04-06 09:57:33	2022-04-05 00:00:00	false
3	2022-04-05 09:34:08	2022-04-05 00:00:00	false
4	2022-04-06 12:20:47	2022-04-05 00:00:00	false
5	2022-04-05 14:06:30	2022-04-05 00:00:00	false
6	2022-04-01 10:55:07	2022-03-31 00:00:00	false
7	2022-04-05 10:12:55	2022-04-05 00:00:00	false
8	2022-04-08 11:14:30	2022-04-05 00:00:00	false
9	2022-04-05 16:12:09	2022-04-05 00:00:00	false
10	2022-04-08 13:47:01	2022-04-05 00:00:00	false
11	2022-04-06 16:36:04	2022-04-05 00:00:00	false
12	2022-04-11 09:54:38	2022-04-05 00:00:00	false
13	2022-04-07 10:00:12	2022-04-05 00:00:00	false
14	2022-04-07 17:30:16	2022-04-05 00:00:00	false

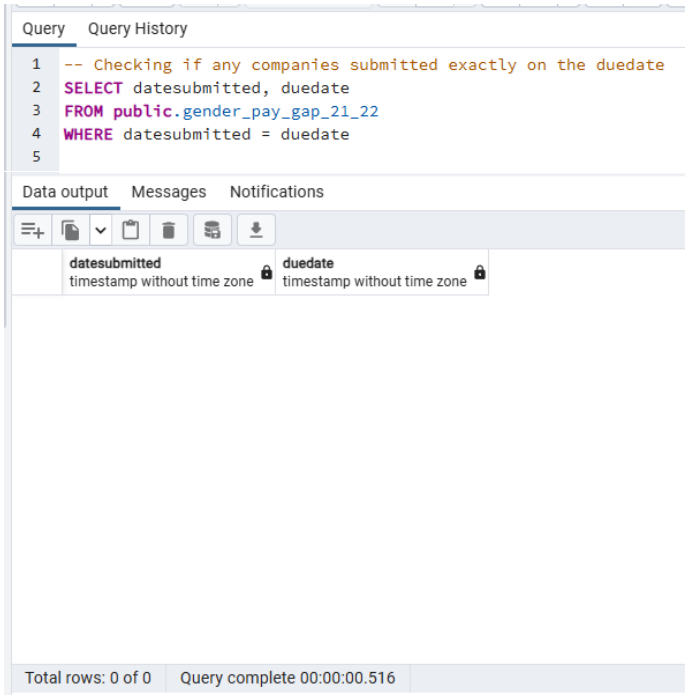
Total rows: 247 of 247 Query complete 00:00:00.500

There are 247 rows with both a ‘false’ entry in the submittedafterthedeadline column and datesubmitted > duedate, indicating 247 rows were wrongly indicated as ‘false’ in submittedafterthedeadline column.

```
Query  Query History
1  -- Determination of errors in column "submittedafterthedeadline"
2  SELECT datesubmitted, duedate, submittedafterthedeadline
3  FROM public.gender_pay_gap_21_22
4  WHERE datesubmitted < duedate AND submittedafterthedeadline = 'TRUE'
5
```



There are 0 rows with both a ‘true’ entry in the submittedafterthedeadline column and datesubmitted < duedate, indicating no rows were wrongly indicated as ‘true’ in the submittedafterthedeadline column.

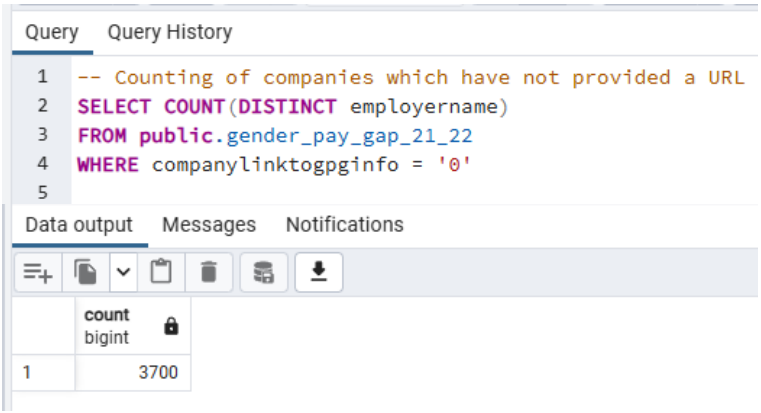


There are no entries with datesubmitted = duedate, indicating no companies submitted exactly on the duedate.

Using datesubmitted > duedate as an indicator, 608 unique companies were counted to have submitted data after the reporting deadline.

Q3. How many companies have not provided a URL?

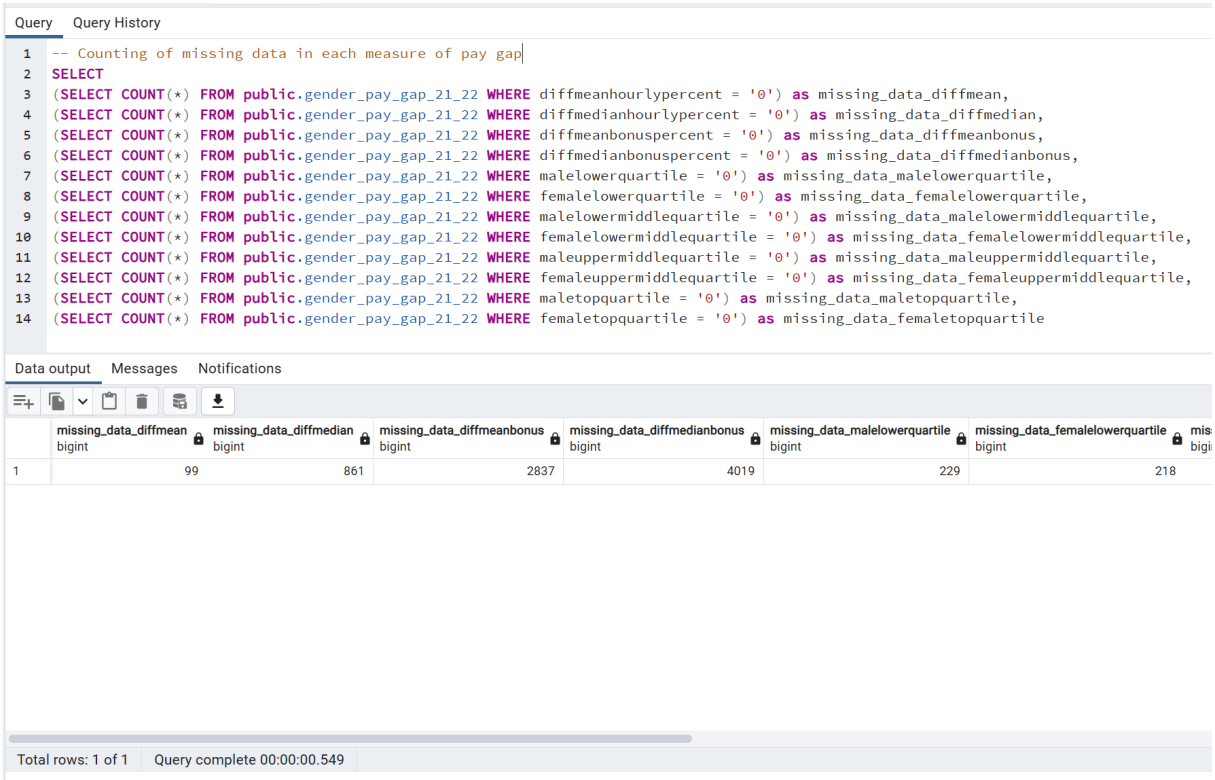
A3.



Assuming the '0' values in the companylinktogpginfo column are due to companies not providing a URL, 3700 unique companies were counted.

Q4. Which measures of pay gap contain too much missing data, and should not be used in our analysis?

A4.



Any entry in the dataset with a value of '0' or '0.0' were considered missing data. The SQL dataset from the SQL database was compared with the dataset from the original website and there was no correlation between the '0' values in the SQL dataset and the null values in the original dataset.

The top 2 highest amount of missing data came from the diffmeanbonuspercent and diffmedianbonuspercent columns with 2837 and 4019 missing values, respectively, which converts to 27.9% and 39.5%, respectively. These 2 measures of pay gap have more than 25% of their data missing and should not be used for analysis.

Q5. Choose which column you will use to calculate the pay gap. Will you use DiffMeanHourlyPercent or DiffMedianHourlyPercent? Can you justify your choice?

A5. DiffMedianHourlyPercent. There are extreme outliers which exist on the data when it comes to the income of both men and women in the company. For example, the salary of the CEO or directors of the company will be several times higher than a junior level executive in the company. Therefore, the median will be a more accurate representation of central tendency of the data because it is not affected by extreme outliers.

Q6. 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?

A6.

Query | Query History

```
1  -- Checking of duplicate values in dataset
2  SELECT diffmeanhourlypercent, diffmedianhourlypercent, diffmeanbonuspercent, diffmedianbonuspercent, malebonuspercent, femalebonuspercent,
3         malelowerquartile, femalelowerquartile, malelowermiddlequartile, femalelowermiddlequartile, COUNT(*)
4  FROM public.gender_pay_gap_21_22
5  GROUP BY diffmeanhourlypercent, diffmedianhourlypercent, diffmeanbonuspercent, diffmedianbonuspercent, malebonuspercent, femalebonuspercent,
6           malelowerquartile, femalelowerquartile, malelowermiddlequartile, femalelowermiddlequartile
7  HAVING COUNT(*) > 1
8
```

Data output | Messages | Notifications

	ercent numeric	diffmedianhourlypercent numeric	diffmeanbonuspercent numeric	diffmedianbonuspercent numeric	malebonuspercent numeric	femalebonuspercent numeric	malelowerquartile numeric	femalelowerquartile numeric	malelowermiddlequartile numeric	femalelowermiddlequartile numeric	count bigint
1	0	0	0	0	0	0	0	0	0	0	9
2	16.1	14.7	54.5	14.2	26.8	35.2	73.0	27.0	82.0	18.0	2
3	5.1	6.2	-0.9	6.6	81.0	81.0	49.0	51.0	51.0	49.0	2
4	24.6	-1.9	51.0	60.8	10.1	7.0	31.4	68.6	37.1	62.9	2
5	22.5	10.2	54.6	13.8	100	100	26.9	73.1	19.2	80.8	2
6	27	18	51	17	27	23	80	20	80	20	2
7	12.6	16.1	32.8	0.0	91.9	92.1	83.8	16.2	79.7	20.3	2
8	28.6	35.0	55.9	63.3	7.5	23.3	50.9	49.1	90.4	9.6	2
9	24.0	23.0	-51.0	-210.0	18.0	3.0	64.0	36.0	78.0	22.0	2
10	-6.7	-9.8	34.7	0	18.8	36.6	93	7	95	5	2
11	-4.1	-5.2	0	0	0	0	93.8	6.2	89.5	10.5	2
12	27.0	21.8	69.5	60.0	100.0	100.0	78.6	21.4	88.6	11.4	2

Total rows: 49 of 49 | Query complete 00:00:00.523 | Ln 8, Col 1

49 groups of repeats were found in the dataset. 1 group was found to have 9 rows containing ‘0’ values across 10 columns used for computing pay gap measures. 47 groups were found to have duplicates, and 1 group was found to have triplicates.

Query | Query History

```
1  -- Identifying duplicate rows by employerid
2  SELECT dataset.*
3  FROM public.gender_pay_gap_21_22 AS dataset
4  JOIN (SELECT diffmeanhourlypercent, diffmedianhourlypercent, diffmeanbonuspercent, diffmedianbonuspercent, malebonuspercent, femalebonuspercent,
5        malelowerquartile, femalelowerquartile, malelowermiddlequartile, femalelowermiddlequartile, COUNT(*)
6        FROM public.gender_pay_gap_21_22
7        GROUP BY diffmeanhourlypercent, diffmedianhourlypercent, diffmeanbonuspercent, diffmedianbonuspercent, malebonuspercent, femalebonuspercent,
8        malelowerquartile, femalelowerquartile, malelowermiddlequartile, femalelowermiddlequartile
9        HAVING COUNT(*) > 1) AS duplicated
10 USING (diffmeanhourlypercent, diffmedianhourlypercent, diffmeanbonuspercent, diffmedianbonuspercent, malebonuspercent, femalebonuspercent,
11        malelowerquartile, femalelowerquartile, malelowermiddlequartile, femalelowermiddlequartile)
12 ORDER BY dataset.employerid
```

Data output | Messages | Notifications

	employername character varying	employerid character varying	address character varying	postcode character varying	companynumber character varying
1	RAGLETH LIMITED	10243	Great Rybugrh, Fakenham, Norfolk, NR21 7AS	NR21 7AS	05498655
2	SALTIRE FACILITIES MANAGEMENT LIMITED	10838	Carnbroe House Finch Way, Strathclyde Business Park, Bellshill, Scotland, ML4 3PE	ML4 3PE	SC211524
3	SCOTTISH WATER BUSINESS STREAM LIMITED	10947	7 Lochside View, Edinburgh, EH12 9DH	EH12 9DH	SC294924
4	SECURIGROUP LIMITED	10986	Venlaw, 349 Bath Street, Glasgow, Lanarkshire, G2 4AA	G2 4AA	SC346167
5	ALHCO GROUP LIMITED	1106	80 Hill Street Hill Street, Hilperton, Trowbridge, England, BA14 7RS	BA14 7RS	06842801
6	ORANGE GENIE COVER LIMITED	111	3rd Floor Buckingham House Buckingham Street, Aylesbury, Buckinghamshire, HP20 2LA	HP20 2LA	04896166
7	SHIELD ENVIRONMENTAL HOLDINGS LIMITED	11140	Shield House Caxton Business Park, Crown Way, Warmley, Bristol, BS30 8XJ	BS30 8XJ	05938654
8	SHIELD ENVIRONMENTAL SERVICES LIMITED	11141	Shield House Caxton Business Park, Crown Way, Warmley, Bristol, BS30 8XJ	BS30 8XJ	01889657
9	SILENTNIGHT GROUP LIMITED	11184	Long Ing Business Park, Long Ing Lane, Barnoldswick, Lancashire, BB18 6BJ	BB18 6BJ	07525259
10	SILENTNIGHT HOLDINGS LIMITED	11185	Long Ing Business Park, Long Ing Lane, Barnoldswick, Lancashire, BB18 6BJ	BB18 6BJ	07627383
11	SPINKO LIMITED	11475	The Innovation Centre, Westland Road, Leeds, England, LS11 5SB	LS11 5SB	04231536

Total rows: 106 of 106 | Query complete 00:00:00.545 | ✓ Successfully run. Total query runtime: 545 msec. 106 rows

The rows with ‘0’ values, duplicates and triplicates were identified using their employerid column.

Query | Query History

```
-- Calculation of percentiles for diffmedianhourlypercent column, excluding missing and duplicate data.
SELECT
    PERCENTILE_CONT(0.00) WITHIN GROUP(ORDER BY diffmedianhourlypercent) AS min,
    PERCENTILE_CONT(0.25) WITHIN GROUP(ORDER BY diffmedianhourlypercent) AS percentile25,
    PERCENTILE_CONT(0.50) WITHIN GROUP(ORDER BY diffmedianhourlypercent) AS median,
    PERCENTILE_CONT(0.75) WITHIN GROUP(ORDER BY diffmedianhourlypercent) AS percentile75,
    PERCENTILE_CONT(1.00) WITHIN GROUP(ORDER BY diffmedianhourlypercent) AS max
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent != 0 AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                                                          '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                                                          '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                                                          '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381',
                                                          '20520')
```

Data output | Messages | Notifications

	min double precision	percentile25 double precision	median double precision	percentile75 double precision	max double precision
1	-275.9	3	11.5	23	100

Query | Query History

```
-- Calculation of no. of outliers for diffmedianhourlypercent column, excluding missing and duplicate data.
SELECT COUNT(diffmedianhourlypercent) as num_of_outliers
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 23+1.5*(23-3)
   OR diffmedianhourlypercent < 3-1.5*(23-3)
  AND diffmedianhourlypercent != 0
  AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                         '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                         '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                         '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','2
```

Data output | Messages | Notifications

	num_of_outliers bigint
1	204

The calculation of percentiles for the diffmedianhourlypercent column, excluding missing and duplicate data. The number of outliers found in the diffmedianhourlypercent column was 204.

Query | Query History

```
-- Calculation of percentiles for diffmeanhourlypercent column, excluding missing and duplicate data.
SELECT
    PERCENTILE_CONT(0.00) WITHIN GROUP(ORDER BY diffmeanhourlypercent) AS min,
    PERCENTILE_CONT(0.25) WITHIN GROUP(ORDER BY diffmeanhourlypercent) AS percentile25,
    PERCENTILE_CONT(0.50) WITHIN GROUP(ORDER BY diffmeanhourlypercent) AS median,
    PERCENTILE_CONT(0.75) WITHIN GROUP(ORDER BY diffmeanhourlypercent) AS percentile75,
    PERCENTILE_CONT(1.00) WITHIN GROUP(ORDER BY diffmeanhourlypercent) AS max
FROM public.gender_pay_gap_21_22
WHERE diffmeanhourlypercent != 0 AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                                                        '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                                                        '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                                                        '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381',
                                                        '20520')
```

Data output | Messages | Notifications

	min double precision	percentile25 double precision	median double precision	percentile75 double precision	max double precision
1	-184.2	4.7	13	22.1	100

Query | Query History

```
-- Calculation of no. of outliers for diffmeanhourlypercent column, excluding missing and duplicate data.
SELECT COUNT(diffmeanhourlypercent) as num_of_outliers
FROM public.gender_pay_gap_21_22
WHERE diffmeanhourlypercent > 22.1+1.5*(22.1-4.7)
   OR diffmeanhourlypercent < 4.7-1.5*(22.1-4.7)
  AND diffmeanhourlypercent != 0
  AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                         '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                         '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                         '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20
```

Data output | Messages | Notifications

	num_of_outliers bigint
1	218

The calculation of percentiles for the diffmeanhourlypercent column, excluding missing and duplicate data. The number of outliers found in the diffmeanhourlypercent column was 218.

```
-- Calculation of average gender pay gap across all companies, excluding outliers, missing, and duplicate data.
SELECT ROUND(AVG(ABS(diffmedianhourlypercent)),2) as average_diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent <= 23+1.5*(23-3)
   OR diffmedianhourlypercent >= 3-1.5*(23-3)
  AND diffmedianhourlypercent != 0
  AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                         '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                         '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                         '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20
```

Data output | Messages | Notifications

	average_diffmedianhourlypercent numeric
1	14.77

The diffmedianhourlypercent column was used as the averaging metric because it has less outliers and because median is a more accurate representation of central tendency of the data when extreme outliers exist.

The average gender pay gap across all the companies, excluding outliers, missing, and duplicate data, was calculated to be 14.77%.

Q7. What are some caveats we need to be aware of when reporting the figure we've just calculated?

A7. The average gender pay gap across all companies completely disregards the differences in average gender pay gaps between companies of different sizes, of different industries, and across different locations.

Q8. What are the 10 companies with the largest pay gaps skewed towards men?

A8.

```
-- Top 10 companies with the largest pay gaps skewed towards men
SELECT employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 0
ORDER BY diffmedianhourlypercent DESC
LIMIT 10
```

Data output | Messages | Notifications

	employername character varying	diffmedianhourlypercent numeric
1	HPI UK HOLDING LTD.	100
2	ATFC LIMITED	100
3	M. ANDERSON CONSTRUCTION LIMITED	100.0
4	PSJ FABRICATIONS LTD	100
5	HULL COLLABORATIVE ACADEMY TRUST	93
6	SERVICE INNOVATION GROUP-UK LIMITED	90.4
7	BRAND ENERGY & INFRASTRUCTURE SERVICES UK, LTD.	89.0
8	ROBINSON WEBSTER (HOLDINGS) LIMITED	85.6
9	THE LEARNING FOR LIFE PARTNERSHIP	82.6
10	GREENBROOK HEALTHCARE (HOUNSLOW) LIMITED	77.1

Q9. What do you notice about the results? Are these well-known companies?

A9. The top 10 companies with largest pay gaps skewed towards men are mostly small and unknown companies. 6 out of 10 of them are small companies with less than 500 employees, 1 company has between 500 to 999 employees, and only 1 of them has between 1000 to 4999 employees. Another 2 companies did not report their number of employees.

Q10. Apply some additional filtering to pick out the most significant companies with large pay gaps.

A10.

Query | Query History

```
-- Top 3 companies of different company sizes with the largest pay gaps skewed towards men
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 0 AND employersize = 'Less than 250'
ORDER BY diffmedianhourlypercent DESC
LIMIT 3)
UNION
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 0 AND employersize = '250 to 499'
ORDER BY diffmedianhourlypercent DESC
LIMIT 3)
UNION
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 0 AND employersize = '500 to 999'
ORDER BY diffmedianhourlypercent DESC
LIMIT 3)
UNION
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 0 AND employersize = '1000 to 4999'
ORDER BY diffmedianhourlypercent DESC
LIMIT 3)
```

Data output | Messages | Notifications

	employersize character varying	employername character varying	diffmedianhourlypercent numeric
1	1000 to 4999	BRAND ENERGY & INFRASTRUCTURE SERVICES UK, LTD.	89.0
2	1000 to 4999	BAM CONSTRUCT UK LIMITED	59.8
3	1000 to 4999	NICHOLAS POSTGATE CATHOLIC ACADEMY TRUST	58.2
4	20,000 or more	LLOYDS BANK PLC	40.9
5	20,000 or more	NATIONAL WESTMINSTER BANK PUBLIC LIMITED COMPA...	34.2
6	20,000 or more	LLOYDS BANKING GROUP PLC	34.2
7	250 to 499	HPI UK HOLDING LTD.	100
8	250 to 499	M. ANDERSON CONSTRUCTION LIMITED	100.0
9	250 to 499	ATFC LIMITED	100
10	5000 to 19,999	EASYJET AIRLINE COMPANY LIMITED	63.6
11	5000 to 19,999	INDEPENDENT VETCARE LIMITED	49.4
12	5000 to 19,999	CVS (UK) LIMITED	42.9
13	500 to 999	GREENBROOK HEALTHCARE (HOUNSLOW) LIMITED	77.1
14	500 to 999	THE FALLIBROOME TRUST	74.4
15	500 to 999	MOTORLINE LIMITED	71.4
16	Less than 250	PSJ FABRICATIONS LTD	100
17	Less than 250	P. D. HOOK (GROUP) LIMITED	69.0
18	Less than 250	WYBOSTON LAKES LIMITED	66.1

Total rows: 18 of 18 | Query complete 00:00:00.574

The companies were filtered based on employer size and the top 3 companies with the largest pay gaps skewed positively towards men were selected.

Query | Query History

```sql
-- Top 3 companies of different company sizes with the largest pay gaps skewed towards women
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent < 0 AND employersize = 'Less than 250'
ORDER BY diffmedianhourlypercent
LIMIT 3)
UNION
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent < 0 AND employersize = '250 to 499'
ORDER BY diffmedianhourlypercent
LIMIT 3)
UNION
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent < 0 AND employersize = '500 to 999'
ORDER BY diffmedianhourlypercent
LIMIT 3)
UNION
(SELECT employersize, employername, diffmedianhourlypercent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent < 0 AND employersize = '1000 to 4999'
ORDER BY diffmedianhourlypercent
LIMIT 3)
```

Data output | Messages | Notifications

	employersize character varying	employername character varying	diffmedianhourlypercent numeric
1	1000 to 4999	FORTEL SERVICES LIMITED	-128.8
2	1000 to 4999	PAYSTREAM MY MAX 2 LIMITED	-83.0
3	1000 to 4999	RULLION ENGINEERING LIMITED	-64.1
4	20,000 or more	NHS PROFESSIONALS LIMITED	-14.4
5	20,000 or more	OPENREACH LIMITED	-14.2
6	250 to 499	RLC (UK) LIMITED	-121.5
7	250 to 499	RAGDALE HALL (1990) LIMITED	-97.6
8	250 to 499	WEST HAM UNITED FOOTBALL CLUB LIMIT...	-68
9	5000 to 19,999	G4S SECURE SOLUTIONS (UK) LIMITED	-275.9
10	5000 to 19,999	FDM GROUP LIMITED	-9.6
11	5000 to 19,999	PULSE HEALTHCARE LIMITED	-9.1
12	500 to 999	NCR UK GROUP LIMITED	-104
13	500 to 999	PICTUREHOUSE CINEMAS LIMITED	-59.3
14	500 to 999	The National Farmers' Union	-51.8
15	Less than 250	INS-SURE SERVICES LIMITED	-61
16	Less than 250	SPORTING INDEX LIMITED	-59.4
17	Less than 250	NICHOLAS ASSOCIATES GROUP LIMITED	-46.2

Total rows: 17 of 17 | Query complete 00:00:00.520

The companies were filtered based on employer size and the top 3 companies with the largest pay gaps skewed positively towards women were selected. For companies with 20,000 or more employees, only 2 companies had pay gaps which were skewed towards women.

Q11. How would you report on the results? Can we say that these companies are engaging in unlawful pay discrimination?

A11. In the UK, the Equality Act 2010 creates a legal obligation for equal pay for men and women, performing equal work. We need to compare pay gap between men and women performing equal work. In order to establish equal work, we need to differentiate between workers with different years of experience, job function, and titles. Unfortunately, the data does not provide that level of granularity.

Therefore, we are not able to claim that these companies are engaging in unlawful pay discrimination despite the large gender pay gaps.

Q12. What's the average pay gap in London versus outside London?

A12.

Query | Query History

```sql
-- Average pay gap in London VS outside London, excluding outliers, missing, and duplicate data.
SELECT
CASE
    WHEN postcode LIKE 'EC%'
    OR postcode LIKE 'E1%'
    OR postcode LIKE 'E2%'
    OR postcode LIKE 'E_ ___'
    OR postcode LIKE 'W1%'
    OR postcode LIKE 'W_ ___'
    OR postcode LIKE 'WC%'
    OR postcode LIKE 'SE%'
    OR postcode LIKE 'SW%'
    OR postcode LIKE 'N_ ___'
    OR postcode LIKE 'N1%'
    OR postcode LIKE 'N2%'
    OR postcode LIKE 'NW%'
    THEN 'London'
    ELSE 'Outside London'
END AS location, ROUND(AVG(ABS(diffmedianhourlypercent)),2) as average_pay_gap
FROM gender_pay_gap_21_22
WHERE diffmedianhourlypercent != 0
AND diffmedianhourlypercent <= 23+1.5*(23-3)
AND diffmedianhourlypercent >= 3-1.5*(23-3)
AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                       '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                       '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                       '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20520')
GROUP BY location
ORDER BY location
```

Data output | Messages | Notifications

	location text	average_pay_gap numeric
1	London	16.39
2	Outside London	14.95

Q13. What's the average pay gap in London versus Birmingham?

A13.

Query | Query History

```sql
-- Average pay gap in London VS Birmingham, excluding outliers, missing, and duplicate data.
SELECT
CASE
    WHEN postcode LIKE 'EC%'
    OR postcode LIKE 'E1%'
    OR postcode LIKE 'E2%'
    OR postcode LIKE 'E_ ___'
    OR postcode LIKE 'W1%'
    OR postcode LIKE 'W_ ___'
    OR postcode LIKE 'WC%'
    OR postcode LIKE 'SE%'
    OR postcode LIKE 'SW%'
    OR postcode LIKE 'N_ ___'
    OR postcode LIKE 'N1%'
    OR postcode LIKE 'N2%'
    OR postcode LIKE 'NW%'
    THEN 'London'
    WHEN postcode BETWEEN 'B1%' AND 'B99%'
    THEN 'Birmingham'
    ELSE 'Others'
END AS location, ROUND(AVG(ABS(diffmedianhourlypercent)),2) as average_pay_gap
FROM gender_pay_gap_21_22
WHERE diffmedianhourlypercent != 0
AND diffmedianhourlypercent <= 23+1.5*(23-3)
AND diffmedianhourlypercent >= 3-1.5*(23-3)
AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                       '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                       '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                       '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20520')
GROUP BY location
ORDER BY location
```

Data output | Messages | Notifications

	location text	average_pay_gap numeric
1	Birmingham	14.64
2	London	16.39
3	Others	14.97

Q14. What is the average pay gap within schools?

A14.

```
-- Average pay gap within schools, excluding outliers, missing, and duplicate data.
SELECT ROUND(AVG(ABS(diffmedianhourlypercent)),2) as average_pay_gap
FROM gender_pay_gap_21_22
WHERE (siccodes LIKE '%85__0%' OR siccodes LIKE '%85__1%' OR siccodes LIKE '%85__2%')
AND diffmedianhourlypercent <= 23+1.5*(23-3)
AND diffmedianhourlypercent >= 3-1.5*(23-3)
AND diffmedianhourlypercent != 0
AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                       '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                       '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                       '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20520')
```

Data output | Messages | Notifications

	average_pay_gap numeric
1	23.06

Q15. What is the average pay gap within banks?

A15.

```
-- Average pay gap within banks, excluding outliers, missing, and duplicate data.
SELECT ROUND(AVG(ABS(diffmedianhourlypercent)),2) as average_pay_gap
FROM gender_pay_gap_21_22
WHERE (siccodes LIKE '%64191%' OR siccodes LIKE '%64110%')
AND diffmedianhourlypercent <= 23+1.5*(23-3)
AND diffmedianhourlypercent >= 3-1.5*(23-3)
AND diffmedianhourlypercent != 0
AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                       '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                       '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                       '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20520')
```

Data output | Messages | Notifications

	average_pay_gap numeric
1	31.70

Q16. Is there a relationship between the number of employees at a company and the average pay gap?

A16.

```
-- Relationship between the number of employees at a company and the average pay gap, excluding outliers, missing, and duplicate data.
SELECT employersize, ROUND(AVG(ABS(diffmedianhourlypercent)),2) as average_gender_pay_gap
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent <= 23+1.5*(23-3)
  AND diffmedianhourlypercent >= 3-1.5*(23-3)
  AND employersize != 'Not Provided'
  AND diffmedianhourlypercent != 0
  AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                         '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                         '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                         '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20520')
GROUP BY 1
ORDER BY 2
```

Data output | Messages | Notifications

	employersize character varying	average_gender_pay_gap numeric
1	5000 to 19,999	11.78
2	20,000 or more	12.01
3	1000 to 4999	13.91
4	500 to 999	15.52
5	250 to 499	16.08
6	Less than 250	16.47

There is a general trend of average pay gap decreasing as the number of employees at a company increases. The only exception are companies with more than 20,000 employees.

Q17. What's the average pay gap in London, Birmingham, Manchester, Liverpool, and Leeds, considering only companies that are skewed towards women?

A17.

Query | Query History

```
-- Average pay gap in London, Birmingham, Manchester, Liverpool, and Leeds considering only companies that are skewed towards women,
-- excluding outliers, missing, and duplicate data.
SELECT
CASE
    WHEN postcode LIKE 'EC%'
    OR postcode LIKE 'E1%'
    OR postcode LIKE 'E2%'
    OR postcode LIKE 'E_ ___'
    OR postcode LIKE 'W1%'
    OR postcode LIKE 'W_ ___'
    OR postcode LIKE 'WC%'
    OR postcode LIKE 'SE%'
    OR postcode LIKE 'SW%'
    OR postcode LIKE 'N_ ___'
    OR postcode LIKE 'N1%'
    OR postcode LIKE 'N2%'
    OR postcode LIKE 'NW%'
    THEN 'London'
    WHEN postcode BETWEEN 'B1%' AND 'B99%'
    THEN 'Birmingham'
    WHEN postcode BETWEEN 'M1%' AND 'M99%'
    THEN 'Manchester'
    WHEN postcode BETWEEN 'L1%' AND 'L99%'
    THEN 'Liverpool'
    WHEN postcode BETWEEN 'LS1%' AND 'LS99%'
    THEN 'Leeds'
    ELSE 'Others'
END AS location, ABS(ROUND(AVG(diffmedianhourlypercent),2)) as average_pay_gap
FROM gender_pay_gap_21_22
WHERE diffmedianhourlypercent != 0
AND diffmedianhourlypercent < 0
AND diffmedianhourlypercent <= 23+1.5*(23-3)
AND diffmedianhourlypercent >= 3-1.5*(23-3)
AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                       '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
```

Data output | Messages | Notifications

	location text	average_pay_gap numeric
1	Birmingha...	6.45
2	Leeds	4.61
3	Liverpool	6.34
4	London	6.59
5	Manchester	5.97
6	Others	6.77

Q18. What's the average pay gap in London, Birmingham, Manchester, Liverpool, and Leeds, considering only companies that are skewed towards men?

A18.

Query | Query History

```sql
-- Average pay gap in London, Birmingham, Manchester, Liverpool, and Leeds considering only companies that are skewed towards men,
-- excluding outliers, missing, and duplicate data.
SELECT
CASE
    WHEN postcode LIKE 'EC%'
    OR postcode LIKE 'E1%'
    OR postcode LIKE 'E2%'
    OR postcode LIKE 'E_ ___'
    OR postcode LIKE 'W1%'
    OR postcode LIKE 'W_ ___'
    OR postcode LIKE 'WC%'
    OR postcode LIKE 'SE%'
    OR postcode LIKE 'SW%'
    OR postcode LIKE 'N_ ___'
    OR postcode LIKE 'N1%'
    OR postcode LIKE 'N2%'
    OR postcode LIKE 'NW%'
    THEN 'London'
    WHEN postcode BETWEEN 'B1%' AND 'B99%'
    THEN 'Birmingham'
    WHEN postcode BETWEEN 'M1%' AND 'M99%'
    THEN 'Manchester'
    WHEN postcode BETWEEN 'L1%' AND 'L99%'
    THEN 'Liverpool'
    WHEN postcode BETWEEN 'LS1%' AND 'LS99%'
    THEN 'Leeds'
    ELSE 'Others'
END AS location, ROUND(AVG(diffmedianhourlypercent),2) as average_pay_gap
FROM gender_pay_gap_21_22
WHERE diffmedianhourlypercent != 0
AND diffmedianhourlypercent > 0
AND diffmedianhourlypercent <= 23+1.5*(23-3)
AND diffmedianhourlypercent >= 3-1.5*(23-3)
AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                       '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
```

Data output | Messages | Notifications

	location text	average_pay_gap numeric
1	Birmingha...	16.01
2	Leeds	15.26
3	Liverpool	14.77
4	London	17.66
5	Manchester	14.22
6	Others	16.50

Q19. Is there a relationship between the number of employees at a company and the average pay gap, considering only companies that are skewed positively towards women?

A19.

```
-- Relationship between the number of employees at a company and the average pay gap considering only companies that are skewed towards women,
-- excluding outliers, missing, and duplicate data.
SELECT employersize, ABS(ROUND(AVG(diffmedianhourlypercent),2)) as average_gender_pay_gap
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent < 0
  AND diffmedianhourlypercent <= 23+1.5*(23-3)
  AND diffmedianhourlypercent >= 3-1.5*(23-3)
  AND employersize != 'Not Provided'
  AND diffmedianhourlypercent != 0
  AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                         '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                         '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                         '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20520')
GROUP BY 1
ORDER BY 2
```

Data output | Messages | Notifications

	employersize character varying	average_gender_pay_gap numeric
1	5000 to 19,999	3.67
2	1000 to 4999	5.29
3	500 to 999	6.27
4	250 to 499	7.18
5	Less than 250	8.64
6	20,000 or more	14.30

There is a general trend of average pay gap skewed positively towards women decreasing as the number of employees at a company increases. The only exception are companies with more than 20,000 employees, the pay gap jumps up to 14.30% from 3.67%.

Q20. Is there a relationship between the number of employees at a company and the average pay gap, considering only companies that are skewed positively towards men?

A20.

```
-- Relationship between the number of employees at a company and the average pay gap considering only companies that are skewed towards men,
-- excluding outliers, missing, and duplicate data.
SELECT employersize, ROUND(AVG(diffmedianhourlypercent),2) as average_gender_pay_gap
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 0
  AND diffmedianhourlypercent <= 23+1.5*(23-3)
  AND diffmedianhourlypercent >= 3-1.5*(23-3)
  AND employersize != 'Not Provided'
  AND diffmedianhourlypercent != 0
  AND employerid NOT IN ('16771','18859','5684','3857','10986','7299','3958','12634','15788','16954','17413','17463','17770','21263',
                         '5053','5946','20745','1259','12937','10947','13484','5872','11475','9626','17467','12713','11184','5328',
                         '20672','19313','625','20152','18916','10243','12938','19792','1617','19722','4775','15039','11940','11140',
                         '8636','12744','19463','1646','7061','15944','11644','19681','4745','10838','20329','1958','20360','111','2381','20520')
GROUP BY 1
ORDER BY 2
```

Data output | Messages | Notifications

	employersize character varying	average_gender_pay_gap numeric
1	20,000 or more	11.92
2	5000 to 19,999	12.29
3	1000 to 4999	15.02
4	500 to 999	17.04
5	250 to 499	17.69
6	Less than 250	18.51

There is a general trend of average pay gap skewed positively towards men decreasing as the number of employees at a company increases.