

Gender Pay Gap Data Analysis with SQL

1. How many companies are in the data set? **10,174 companies**

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

2. How many of them submitted their data after the reporting deadline? **361 companies submitted their data after the reporting deadline.**

```
SELECT COUNT (DISTINCT employerid)
FROM gender_pay_gap_21_22
WHERE SubmittedAfterTheDeadline = 'true';
```

3. How many companies have not provided a URL? **3,700 companies did not provide a URL.**

```
SELECT COUNT (DISTINCT employerid)
FROM gender_pay_gap_21_22
WHERE CompanyLinkToGPGInfo = '0';
```

4. Which measures of pay gap contain too much missing data, and should not be used in our analysis? **The columns about bonuses had a lot of missing info, so probably best not to use those.**

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.

SicCodes are a list of comma-separated SIC codes used to describe the employer's purpose and sectors of work at the time of reporting. It corresponds to the The Company Number of the employer as listed on Companies House (null for public sector) via CoHo API.

Let's work out the average gender pay gap across the UK.

5. Choose which column you will use to calculate the pay gap. Will you use DiffMeanHourlyPercent or DiffMedianHourlyPercent? Can you justify your choice? **The values for the mean and median are both very close to each other (mean: 13.64 and median: 12.31). But we should use a 'diffmedianhourlypercent' since it's not skewed by any outliers. Mean: 13.6350599567525064. Median: 12.3131708275997641**

```
SELECT
AVG(diffmeanhourlypercent) as mean,
```

```
AVG(diffmedianhourlypercent) as median
FROM gender_pay_gap_21_22;
```

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? **I used the average of the median (as presented in the query above). Per the Data Dictionary, I liked the Median pay gap definition better than the mean pay gap to answer these questions. The Median pay gap is the difference in pay between the middle-ranking women and men and that if you line up all the women/men working at a company in two separate lines in order of salary, the median pay gap will be the difference in salary between the women in the middle of her line and the man in the middle of his.**
7. What are some caveats we need to be aware of when reporting the figure we've just calculated? **'ResponsiblePerson' tells us that the name of the responsible person who confirms that the published information is accurate - Employers covered by the private sector regulations only. Also the source website tells us that the data is self-reported, by all companies with more than 250 employees and that it's mandatory for these companies.**

Now, let's look at some of the companies with the largest pay gaps.

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

```
SELECT *
FROM gender_pay_gap_21_22
ORDER BY diffmedianhourlypercent DESC
LIMIT 10;
```

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

```

SELECT employername, maletopquartile, malebonuspercent,
diffmeanhourlypercent, diffmedianhourlypercent
FROM gender_pay_gap_21_22
WHERE maletopquartile >= '50.0'
ORDER BY maletopquartile DESC
LIMIT 10;

```

	employername character varying	maletopquartile numeric	malebonuspercent numeric	diffmeanhourlypercent numeric	diffmedianhourlypercent numeric
1	COLLINS RIVER ENTERPRISES LIMITED	100	10	54	48.5
2	CGC EVENTS LIMITED	100	1.7	36.3	35.4
3	CAPRICE HOLDINGS LIMITED	100	3	40	-7.9
4	BREWHOUSE & KITCHEN LIMITED	100	11.8	25.8	19.5
5	BUILD-A-BEAR WORKSHOP UK LIMITED	100	29	20	13
6	BRENTFORD FC LIMITED	100	59	74	23
7	ALTRAD ENGINEERING SERVICES LIMIT...	100.0	39.0	31.0	29.0
8	BILLS RESTAURANTS LTD.	100	96.7	40.9	15.9
9	BOSWELLS COFFEE COMPANY LIMITED	100.0	0.0	44.0	51.0
10	CRERAR HOTEL GROUP LIMITED	100	0	50.1	9.6

```

SELECT employername, maletopquartile, malebonuspercent,
diffmeanhourlypercent, diffmedianhourlypercent
FROM gender_pay_gap_21_22
WHERE maletopquartile >= '50.0'
ORDER BY maletopquartile DESC
LIMIT 10;

```

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

```

SELECT employername, maletopquartile, malebonuspercent,
diffmeanhourlypercent, diffmedianhourlypercent
FROM gender_pay_gap_21_22

```

```
WHERE maletopquartile >= '50.0'
ORDER BY maletopquartile DESC
LIMIT 10;
```

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

9. What do you notice about the results? Are these well-known companies? **There are not well known companies and the size of the companies are on the smaller side.**

```
SELECT employername, employersize
FROM gender_pay_gap_21_22
ORDER BY diffmedianhourlypercent DESC
LIMIT 10;
```

	employername character varying	employersize character varying
1	HPI UK HOLDING LTD.	250 to 499
2	M. ANDERSON CONSTRUCTION LIMITED	250 to 499
3	PSJ FABRICATIONS LTD	Less than 250
4	ATFC LIMITED	250 to 499
5	HULL COLLABORATIVE ACADEMY TRUST	Not Provided
6	SERVICE INNOVATION GROUP-UK LIMITED	250 to 499
7	BRAND ENERGY & INFRASTRUCTURE SERVICES UK, L..	1000 to 4999
8	ROBINSON WEBSTER (HOLDINGS) LIMITED	250 to 499
9	THE LEARNING FOR LIFE PARTNERSHIP	Not Provided
10	GREENBROOK HEALTHCARE (HOUNSLOW) LIMITED	500 to 999

10. Apply some additional filtering to pick out the most significant companies with large pay gaps. **(I did below)**
11. How would you report on the results? Can we say that these companies are engaging in unlawful pay discrimination? **This data set doesn't tell us why there is a gender gap at a specific company versus another, therefore we don't know if an employer's gender gap is caused by unequal gender pay.**

Let's see if there are differences in the average pay gaps in different parts of the country.

Think about where you might be able to find this information, since there's no 'city' column in our data set.

```
SELECT address, postcode
FROM gender_pay_gap_21_22
LIMIT 5;
```

12. What's the average pay gap in London versus outside London? **Average pay gap is 11.9356709628506444 (Outside of London) and 13.6265342163355408 (In London).**

```
SELECT
AVG(diffmedianhourlypercent) AS average_median_difference
FROM gender_pay_gap_21_22
WHERE address NOT LIKE '%London%';
```

```
SELECT
AVG(diffmedianhourlypercent) AS average_median_difference
FROM gender_pay_gap_21_22
WHERE address ILIKE '%London%';
```

13. What's the average pay gap in London versus Birmingham? **Average pay gap is 10.7639593908629442 (Inside Birmingham). The average pay gap is less in Birmingham (10.76) than in London (13.62).**

```
SELECT
AVG(diffmedianhourlypercent) AS average_median_difference
FROM gender_pay_gap_21_22
WHERE address ILIKE '%Birmingham%';
```

Let's see if there are differences in the average pay gaps across different industries.

Think carefully about where you might be able to find this information in your data set — there are a couple of different approaches to this task.

14. What is the average pay gap within schools? **Using the employer name in the where clause I got 215 employers that are schools and their average pay gap is 15.71. Alternatively, using the address in the where clause, I got 363 schools and their average pay gap is 19.68.**

I filtered in the WHERE clause via employer name:

```
SELECT
AVG(diffmeanhourlypercent) AS average_median_difference
FROM gender_pay_gap_21_22
WHERE employername ILIKE '%School%';
```

Another way to filter is through the address:

```
SELECT AVG(diffmeanhourlypercent) AS average_median_difference
FROM gender_pay_gap_21_22
WHERE address ILIKE '%School%';
```

15. What is the average pay gap within banks? **Using the employer name in the where clause I got 69 total banks and their average pay gap is 26.31.**

```
SELECT
AVG(diffmeanhourlypercent) AS average_median_difference
FROM gender_pay_gap_21_22
WHERE employername ILIKE '%Bank%';
```

16. Is there a relationship between the number of employees at a company and the average pay gap? **Per the output below, it appears that the largest company size actually has the lowest gender pay gaps (20,000 or more in size = 12.48 average pay gap).**

```
SELECT
employersize,
AVG(diffmeanhourlypercent) AS average_median_difference
FROM gender_pay_gap_21_22
GROUP BY employersize
ORDER BY average_median_difference DESC;
```

	employersize character varying	average_median_difference numeric
1	5000 to 19,999	14.1213362068965517
2	Less than 250	14.0567669172932331
3	250 to 499	13.9206832007487131
4	500 to 999	13.6666133546581367
5	Not Provided	13.0319047619047619
6	1000 to 4999	12.9070389488503050
7	20,000 or more	12.4806451612903226

Industry Analysis:

```
SELECT siccodes, AVG(diffmedianhourlypercent) as
diffmedianhourlypercent, employername
FROM gender_pay_gap_21_22
GROUP BY siccodes, employername
ORDER BY diffmedianhourlypercent DESC
LIMIT 10;
```

	siccodes character varying	diffmedianhourlypercent numeric	employername character varying
1	55100	100.0000000000000000	HPI UK HOLDING LTD.
2	25110	100.0000000000000000	PSJ FABRICATIONS LTD
3	56101	100.0000000000000000	ATFC LIMITED
4	41100	100.0000000000000000	M. ANDERSON CONSTRUCTION LIMITED
5	85200,85310	93.0000000000000000	HULL COLLABORATIVE ACADEMY TRUST
6	82990	90.4000000000000000	SERVICE INNOVATION GROUP-UK LIMITED
7	96090	89.0000000000000000	BRAND ENERGY & INFRASTRUCTURE SERVICES UK, LTD.
8	47710,47910	85.6000000000000000	ROBINSON WEBSTER (HOLDINGS) LIMITED
9	85200	82.6000000000000000	THE LEARNING FOR LIFE PARTNERSHIP
10	86210	77.1000000000000000	GREENBROOK HEALTHCARE (HOUNSLOW) LIMITED

```
SELECT siccodes, AVG(diffmedianhourlypercent) as
diffmedianhourlypercent
FROM gender_pay_gap_21_22
GROUP BY siccodes
ORDER BY diffmedianhourlypercent DESC
LIMIT 10;
```

	siccodes character varying	diffmedianhourlypercent numeric
1	41100, 55100, 56101, 82990	66.1000000000000000
2	41201, 64203, 70100, 82990	59.8000000000000000
3	61200, 61300, 61900	58.0000000000000000
4	62030, 78109, 78200, 78300	55.0500000000000000
5	41100, 41202	55.0000000000000000
6	33190	53.1333333333333333
7	51102	50.6333333333333333
8	5102, 68209, 78300	49.9000000000000000
9	46510, 62030, 62090	48.0000000000000000
10	33170, 71129	47.9000000000000000




Region Analysis favoring women:

```
SELECT postcode, AVG(diffmedianhourlypercent) as
diffmedianhourlypercent, employername
FROM gender_pay_gap_21_22
GROUP BY postcode, employername
ORDER BY diffmedianhourlypercent ASC
LIMIT 5;
```

	postcode character varying	diffmedianhourlypercent numeric	employername character varying
1	SM1 1JB	-275.9000000000000000	G4S SECURE SOLUTIONS (UK) LIMIT...
2	WS2 8DS	-128.8000000000000000	FORTEL SERVICES LIMITED
3	BB5 5AY	-121.5000000000000000	RLC (UK) LIMITED
4	W2 1BQ	-104.0000000000000000	NCR UK GROUP LIMITED
5	CV23 0UZ	-97.6000000000000000	RAGDALE HALL (1990) LIMITED

Regional analysis favoring men:

```
SELECT postcode, AVG(diffmedianhourlypercent) as
diffmedianhourlypercent, employername
FROM gender_pay_gap_21_22
GROUP BY postcode, employername
ORDER BY diffmedianhourlypercent DESC
LIMIT 5;
```


	postcode character varying 	diffmedianhourlypercent numeric 	employername character varying 
1	CM2 5PW	100.0000000000000000	M. ANDERSON CONSTRUCTION LIMI...
2	NE11 0BL	100.0000000000000000	ATFC LIMITED
3	W1S 4HQ	100.0000000000000000	HPI UK HOLDING LTD.
4	WA16 8QZ	100.0000000000000000	PSJ FABRICATIONS LTD
5	HU7 6AH	93.0000000000000000	HULL COLLABORATIVE ACADEMY T...