

```
select*
```

```
from gender_pay_gap_21_22
```

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

```
select
```

```
count(employerid)
```

```
from gender_pay_gap_21_22;
```

→ 10174

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

```
select
```

```
submittedafterthedeathline,
```

```
count(employerid)
```

```
from gender_pay_gap_21_22
```

```
group by submittedafterthedeathline;
```

→ True (361)

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

```
select
```

```
count(companylinktogpginfo)
```

```
from gender_pay_gap_21_22
```

```
where companylinktogpginfo = '0'
```

→ 3700

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

```
select
```

```
diffmeanhourlypercent
```

```
from gender_pay_gap_21_22
```

```
where diffmeanhourlypercent is NULL;
```

→ No NULLS

```
select
```

```
count(DiffMedianHourlyPercent)
```

```
from gender_pay_gap_21_22
```

```
where DiffMedianHourlyPercent = 0.0 ;
```

→ 861 (too many missing values)

```
select
```

```
count(DiffMeanHourlyPercent)
```

```
from gender_pay_gap_21_22
```

```
where DiffMeanHourlyPercent = 0.0;
```

→ 99

```
select
```

```
STDDEV_SAMP(diffmeanhourlypercent)
```

```
from gender_pay_gap_21_22;
```

→ 14

```
select
STDDEV_SAMP(DiffMedianHourlyPercent)
from gender_pay_gap_21_22;
→ 16
```

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

DiffMeanHourlyPercent

Even though the mean can have outliers and can skew the results the median has too many missing values to be reliable.

-- 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
avg(diffmeanhourlypercent)
from gender_pay_gap_21_22;
```

→ 13.6350599567525064

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

→ that there might be outliers and the sample size the results are out of the given data set there might not be representative of the population.

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

```
select
employername,
diffmeanhourlypercent
from gender_pay_gap_21_22
where diffmeanhourlypercent >0
order by diffmeanhourlypercent desc
limit 10;
```

	employername character varying	diffmeanhourlypercent numeric
1	HPI UK HOLDING LTD.	100
2	PSJ FABRICATIONS 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	BRAND ENERGY & INFRASTRUCTURE SERVICES UK, LTD.	91.0
9	MANCHESTER CITY FOOTBALL CLUB LIMITED	91
10	NEWCASTLE UNITED FOOTBALL COMPANY LIMITED	90.4

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

These are male dominant industries.
Football clubs mostly so yes.

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

```
select
employername,
diffmeanhourlypercent,
responsibleperson,
employersize,
malebonuspercent,
femalebonuspercent
from gender_pay_gap_21_22
where diffmeanhourlypercent >0
and malebonuspercent>80
order by diffmeanhourlypercent desc
limit 10;
```

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

We don't know as the different types of jobs are not specified is all included under the same aggregation so we can not know the nuances of seniority and the responsibility of different roles within each company.

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

```
select
```

```
avg(diffmeanhourlypercent)
from gender_pay_gap_21_22
where address like '%London%'
;
```

→ 15.700486725663717

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

```
select

avg(diffmeanhourlypercent)

from gender_pay_gap_21_22

where address like '%Birmingham%'

;
```

→ 13.222448979591837

14. What is the average pay gap within schools?

```
SELECT

avg(diffmeanhourlypercent) AS avg_siccodes,

CASE

    WHEN siccodes = '85100' THEN 'Pre-primary education'

    WHEN siccodes = '85200' THEN 'Primary education'

    WHEN siccodes = '85310' THEN 'Secondary education'

    WHEN siccodes = '85320' THEN 'TechVoc Secondary education'

    WHEN siccodes = '85410' THEN 'Post secondary education'

    WHEN siccodes = '85421' THEN 'undergraduate'

    WHEN siccodes = '85442' THEN 'Masters'

    ELSE 'Other'

END AS category

FROM gender_pay_gap_21_22

GROUP BY category
```

15. What is the average pay gap within banks?

```
SELECT  
  
    avg(diffmeanhourlypercent),  
  
    CASE  
  
        WHEN siccodes like '64%' THEN 'Banks'  
  
        ELSE 'Other'  
  
    END AS category  
  
FROM gender_pay_gap_21_22  
  
GROUP BY category;
```

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

```
SELECT  
  
COUNT(employerid) AS Companies_Count, avg(DiffMeanHourlyPercent) AS  
AVG_DiffMeamHourlyPercent, employersize  
  
FROM gender_pay_gap_21_22  
  
GROUP BY employersize;
```

