UK GENDER PAY GAP LAB

Jonatan Martin | LinkedIn | Contact

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

```
select
count(distinct employername)
from gender_pay_gap_21_22
```

2. How many of them submitted their data after the reporting deadline? **371** companies

```
select
count(distinct employername)
from gender_pay_gap_21_22
where submittedafterthedeadline = true
```

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

```
select
count(distinct employername)
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? **861** nulls in diffmedianhourlypercent and nulls in **99** diffmeanhourlypercent. So would be best to use diffmeanhourlypercent.

```
select
count(diffmedianhourlypercent)
from gender_pay_gap_21_22
where diffmedianhourlypercent = 0

select
count(diffmeanhourlypercent)
from gender_pay_gap_21_22
where diffmeanhourlypercent = 0
```

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.

The UK Standard Industrial Classification of economic activities, abbreviated as UK SIC, is a five-digit classification providing the framework for collecting and presenting a large range of statistical data according to economic activity. Companies within the same sector share the same two digits of their sic code.

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? Used standard deviation for both mean (14.86) and median (16.19) to see how dispersed they are, and considering the median has way more nulls, I will stick to the mean.

```
select
STDDEV_POP(diffmeanhourlypercent),
STDDEV_POP(diffmedianhourlypercent)
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? 13.77 is the average gap using diffmeanhourlypercent. Using the mean because of the conclusions from questions 4 and 5.

```
select
avg(diffmeanhourlypercent)
from gender_pay_gap_21_22
where diffmeanhourlypercent != 0.0
```

7. What are some caveats we need to be aware of when reporting the figure we've just calculated? All data has been self reported, so we are "trusting" companies to have sent truthful data. Also, the UK government asked companies to do their calculations, so we don't have access to raw data.

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
employername,
diffmeanhourlypercent
from gender_pay_gap_21_22
order by diffmeanhourlypercent desc
```

	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?
It makes sense we have so many football teams, and some construction companies, although it does not make sense for the company with the highest gap being a hospitality company, even being a luxury hotel company.

10. Apply some additional filtering to pick out the most significant companies with large pay gaps. Filtering the companies with the highest gaps by company size, we can tell small companies are more likely to have a higher average pay gap in favour of men.

```
select
Avg(diffmeanhourlypercent) as mean,
employersize
from gender_pay_gap_21_22
where diffmeanhourlypercent > 0.0
group by employersize
order by mean desc
```

	mean numeric	employersize character varying
1	19.1700680272108844	Less than 250
2	17.5397320207820618	250 to 499
3	17.1502824858757062	Not Provided
4	16.4271532184950136	500 to 999
5	15.4111052907281299	1000 to 4999
6	15.3407744874715262	5000 to 19,999
7	14.1928571428571429	20,000 or more

```
select
max(diffmeanhourlypercent) as max,
employersize
from gender_pay_gap_21_22
where diffmeanhourlypercent > 0.0
group by employersize
order by max desc
```

	max numeric	employersize character varying
1	100	Less than 250
2	100.0	250 to 499
3	92.0	500 to 999
4	91.6	1000 to 4999
5	63.8	Not Provided
6	53.1	5000 to 19,999
7	45.4	20,000 or more

11. How would you report on the results? Can we say that these companies are engaging in unlawful pay discrimination? If we could have the percentage of female and male employees in every company, we could find out the impact of the pay gap. Because it looks like companies with the highest gap and companies with a higher number of male employees. Even more, football players tend to have big amounts of income, so it would be easy for women in those companies to have a way more smaller pay.

If I could find out that companies with higher numbers of female workers have bigger gaps, that could definitely be unlawful pay discrimination.

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.

12. What's the average pay gap in London versus outside London? Avg pay gap in London is 18.2 compared to 16.3 from outside London

```
select
address,
avg(diffmeanhourlypercent) OVER ()
from gender_pay_gap_21_22
where address not like '%London%' and diffmeanhourlypercent >
0.0
```

13. What's the average pay gap in London versus Birmingham? 16.5 compared to 18.2 in London

```
select
address,
avg(diffmeanhourlypercent) OVER ()
from gender_pay_gap_21_22
where address like '%Birmingham%' and diffmeanhourlypercent >
0.0
```

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? 18.7

```
select
employername,
siccodes,
avg(diffmeanhourlypercent) OVER ()
from gender_pay_gap_21_22
where siccodes like '85%' and diffmeanhourlypercent > 0.0
```

15. What is the average pay gap within banks? 25.5

```
select
employername,
siccodes,
avg(diffmeanhourlypercent) OVER ()
from gender_pay_gap_21_22
where siccodes like '85%' and diffmeanhourlypercent > 0.0
```

16. Is there a relationship between the number of employees at a company and the average pay gap? the companies with the highest gaps by company size, we can tell small companies are more likely to have a higher average pay gap in favour of men.

```
select
max(diffmeanhourlypercent) as max,
employersize
from gender_pay_gap_21_22
where diffmeanhourlypercent > 0.0
group by employersize
order by max desc
```

	max numeric	employersize character varying
1	100	Less than 250
2	100.0	250 to 499
3	92.0	500 to 999
4	91.6	1000 to 4999
5	63.8	Not Provided
6	53.1	5000 to 19,999
7	45.4	20,000 or more

```
select
Avg(diffmeanhourlypercent) a
employersize
from gender_pay_gap_21_22
where diffmeanhourlypercent
group by employersize
order by mean desc
```

	mean numeric	employersize character varying
1	19.1700680272108844	Less than 250
2	17.5397320207820618	250 to 499
3	17.1502824858757062	Not Provided
4	16.4271532184950136	500 to 999
5	15.4111052907281299	1000 to 4999
6	15.3407744874715262	5000 to 19,999
7	14.1928571428571429	20,000 or more