Gender Pay Gap Data Analysis with SQL

Phasiri Honsa

*****I have use Sqlite for all my queries*****

What the 'SicCodes' column corresponds to?

SIC code or Standard Industrial Classification are 5 digit codes that groups companies into the type of business. For instance, 20301 is a code for manufacture of paints, varnishes and similar coatings.

1) How many companies are in the data set?

There are 10174 companies

SQL Code:

```
SELECT COUNT(DISTINCT employerid)

FROM GenderPayGap21;
```

Using the employerid to make sure that I have the correct number of companies.

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

There are 361 companies who submitted their reporting after the deadline.

```
SELECT COUNT (employerid)

FROM GenderPayGap21

WHERE submittedafterthedeadline ='True';
```

3) How many companies have not provided a URL?

There were 3700 companies did not submit their gender pay gap information link

```
SELECT COUNT (employerid)

FROM GenderPayGap21

WHERE companylinktogpginfo = 0 ;
```

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

We should not include all the bonus columns which are ;

diffmedianhourlypercent = 861,

diffmeanbonuspercent contains 2837 zeros,

diffmedianbonuspercent contains 4019 zeros,

malebonuspercent contains 2704 zeros ,

and femalebonuspercent which contains 2754 zeros.

I did the formula for each columns one by one to find the columns that contain most zero.

```
SELECT COUNT (*)

FROM GenderPayGap21

WHERE (insert different columns) = 0;
```

5)Which is better DiffMeanHourlyPercent or DiffMedianHourlyPercent? Can you justify your choice?

I would choose the DifferentMedianHourlyPercent because there are a lot of influential outliners to this data. The median is the difference in pay between all the middle ranking pays. Which means the median estimate is more standardized.

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 median as your averaging metric? Can you justify your choice?

Similar to the reasoning in question 5, I use the median to find my average because the result will be more harmonized. The average for the gender paygap is 12.3.

```
SELECT AVG(diffmedianhourlypercent)
FROM GenderPayGap21;
```

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

The unemployment and pay cuts as the result from Covid during the time of data set. More specifically, this median figure is a standardized figure and are not specific to company or sectors.

8) What are the top 10 companies with the largest pay gap skewed towards men?

To find the largest skewed, we must find the most positive numbers within the diffmedianhourlypercent column.

```
SELECT employername, diffmedianhourlypercent, employersize
FROM GenderPayGap21
ORDER BY diffmedianhourlypercent desc
LIMIT  10;
```

The top 10 companies which have the largest pay gap leaning towards men are listed in the table below :

Company names	Different Median Hourly Percent	Employee Numbers
ATFC LIMITED	100	250 to 499
HPI UK HOLDING LTD.	100	250 to 499
M. ANDERSON CONSTRUCTION LIMITED	100	250 to 499
PSJ FABRICATIONS LTD	100	Less than 250
HULL COLLABORATIVE ACADEMY TRUST	93	Not provided
SERVICE INNOVATION GROUP-UK LIMITED	90.4	250 to 499
BRAND ENERGY & INFRASTRUCTURE SERVICES UK, LTD.	89	1000 to 4999
ROBINSON WEBSTER (HOLDINGS) LIMITED	85.6	250 to 499

THE LEARNING FOR LIFE PARTNERSHIP	82.6	Not provided
GREENBROOK HEALTHCARE (HOUNSLOW) LIMITED	77.1	500 to 999

9) Result Analysis

From the table above most of these are smaller companies with less than 500 employees. I also notices that 4 of these companies men have 100% more pay than women, which seems very unlikely. Maybe some data given by the company is incorrect. There’s only one big company, which is Brand Energy & Infrastructure services UK,LTD

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

In my opinion the most significant companies would be companies with largest number of employees. Because they have more influences on the pay gap as a whole and society. I have filter the results to only showing me companies with 20,000 or more employees. All of these companies are well-known companies. The average pay gaps for these companies are also well above the total average median, which is 12.3 in this case. Which mean the companies with more than 20,000 employees are skewed towards men.

```
SELECT employername, diffmedianhourlypercent
FROM GenderPayGap21
WHERE employersize = '20,000 or more '
ORDER BY diffmedianhourlypercent DESC
LIMIT 10;
```

Employername	Different Median Hourly Percent
LLOYDS BANK PLC	40.9
LLOYDS BANKING GROUP PLC	34.2
NATIONAL WESTMINSTER BANK PUBLIC LIMITED COMPANY	34.2
J D WETHERSPOON PLC	29.5
HBOS PLC	27.8
BARCLAYS EXECUTION SERVICES LIMITED	27.1
NEXT RETAIL LIMITED	26.9
WPP 2005 LIMITED	23.6
BRITISH AIRWAYS PLC	22.2
Leeds Teaching Hospitals Nhs Trust	21.2

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

Although it is illegal to provide unequal pay due to gender discrimination it is difficult to proof it. If the employers can justify why they are paying men more then they will not be penalized. The pay difference can be justify through numerous reasons such as experience, performance or location.

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

The average pay gap for companies in London is 13.62. While the average for the rest of the companies outside of London is 12.31

```
SELECT AVG(diffmedianhourlypercent)
FROM GenderPayGap21
WHERE address like '%london%';

SELECT AVG (diffmedianhourlypercent)
FROM  GenderPayGap21
WHERE address != '%london%';
```

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

The average pay gap for companies in London is 13.62. While the average for companies in Birmingham is 10.76.

```
SELECT AVG(diffmedianhourlypercent)
FROM GenderPayGap21
WHERE address like '%Birmingham%';
```

14)What is the average pay gap within schools?

The average pay gap is 24.6 among schools and education institutions.

```
SELECT AVG(diffmedianhourlypercent)
FROM GenderPayGap21
WHERE employername like '%school%' or employername like '%education%';
```

15)What is the average pay gap within banks?

The average pay gap withing banks is 25.66.

```
SELECT AVG(diffmedianhourlypercent)
FROM GenderPayGap21
WHERE employername like '%bank%';
```

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

From the table below, I have notice that the relationship between average pay gap and company sizes goes opposite ends. As the number of employees decreases, the average pay gap goes up except for companies with less than 250 employees. Which have average pay gap of 11.3.

size_cat	avg(diffmedianhourlypercent)
small	12.7764857276556
medium	12.6457816873251
large	11.56175504458
extra small	11.3451127819549
Larger	10.2519396551724
extra large	9.73064516129032

```
WITH company_size
AS
(
SELECT employername, diffmedianhourlypercent,

CASE
WHEN employersize = 'Less than 250' THEN ' extra small'
WHEN employersize = '250 to 499' THEN  'small'
WHEN employersize = '500 to 999' THEN 'medium'
WHEN employersize = '1000 to 4999' THEN 'large'
WHEN employersize = '5000 to 19,999' THEN 'larger'
WHEN employersize = '20,000 or more' THEN 'extra large'
ELSE 'Not provided'
END AS size_cat
FROM GenderPayGap21)

SELECT size_cat, AVG(diffmedianhourlypercent)
FROM company_size
GROUP BY size_cat
ORDER BY AVG(diffmedianhourlypercent) DESC;
```