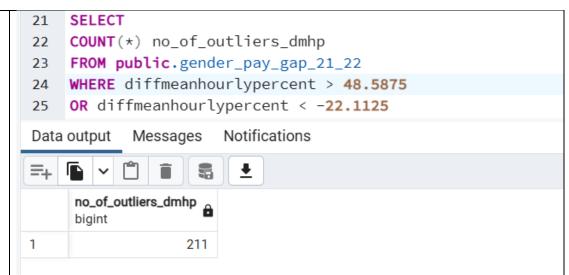


FROM public.gender\_pay\_gap\_21\_22 WHERE diffmedianhourlypercent::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_dmbp FROM public.gender\_pay\_gap\_21\_22 WHERE diffmeanbonuspercent::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_dmdbp FROM public.gender\_pay\_gap\_21\_22 WHERE diffmedianbonuspercent::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_mbp FROM public.gender\_pay\_gap\_21\_22 WHERE malebonuspercent::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_fbp FROM public.gender\_pay\_gap\_21\_22 WHERE femalebonuspercent::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_mlq FROM public.gender\_pay\_gap\_21\_22 WHERE malelowerquartile::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_flq FROM public.gender\_pay\_gap\_21\_22 WHERE femalelowerquartile::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_mlmq FROM public.gender\_pay\_gap\_21\_22 WHERE malelowermiddlequartile::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_flmq FROM public.gender\_pay\_gap\_21\_22 WHERE femalelowermiddlequartile::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_mumq FROM public.gender\_pay\_gap\_21\_22 WHERE maleuppermiddlequartile::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_fumq FROM public.gender\_pay\_gap\_21\_22 WHERE femaleuppermiddlequartile::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_mtq FROM public.gender\_pay\_gap\_21\_22 WHERE maletopquartile::text = '0'), (SELECT COUNT(\*) no\_of\_missing\_ftq FROM public.gender\_pay\_gap\_21\_22 WHERE femaletopquartile::text = '0') Data output Messages Notifications no\_of\_missing\_dmhp bigint no\_of\_missing\_dmhp bigint no\_of\_missing\_dmbp a bigint no\_of\_missing\_dmbp a bigint no\_of\_missing\_dmbp a bigint no\_of\_missing\_mbp a no\_of\_missing\_mbp a bigint no\_of\_missing\_mbp a no\_of\_missing\_mbp a bigint no\_of\_m 2787 404 3381 212 205 no\_of\_missing\_mlmq a no\_of\_missing\_flmq a no\_of\_missing\_flmq a no\_of\_missing\_mumq a no\_of\_missing\_mtq a no 223 Assumption: value of 0.0 is an input provided whereas 0 is no input provided/missing data. To only count 0 and not 0.0, I used cast to convert data type to TEXT, and then only counted '0'.

Diffmeanbonuspercent and diffmedianbonuspercent were found to have the most missing data - 2787 (>27%) and 3381 (>33%) out of 10174 rows respectively. Hence, they should not be used in the analysis. Additionally, malebonuspercent and femalebonuspercent were found to have significant missing data – 1356 (>13%) and 1374 (>13%) respectively. Hence they should also not be used in the analysis. Choose which column you will use to calculate the pay gap. Will you use 5 DiffMeanHourlyPercent or DiffMedianHourlyPercent? Can you justify your choice? 12 SELECT 13 PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) + 1.5\* 14 (PERCENTILE CONT(0.75) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) -15 PERCENTILE\_CONT(0.25) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC)) AS Q3\_plus1\_5\_IQR, 16 PERCENTILE\_CONT(0.25) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) - 1.5\* 17 (PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) 18 PERCENTILE\_CONT(0.25) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC)) Q1minus1\_5\_IQR 19 FROM public.gender\_pay\_gap\_21\_22 Data output Messages Notifications =+ 6 ~ 6 = 5 q3\_plus1\_5\_iqr
double precision

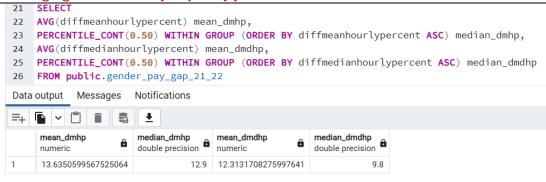
q1minus1\_5\_iqr
double precision 52.75 -30.04999999999999 SELECT 21 COUNT(\*) no\_of\_outliers\_dmdhp 22 23 FROM public.gender\_pay\_gap\_21\_22 WHERE diffmedianhourlypercent > 52.75 24 OR diffmedianhourlypercent < -30.05 25 Notifications Data output Messages no\_of\_outliers\_dmdhp bigint 1 199 12 SELECT 13 PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) + 1.5\* 14 (PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) 15 PERCENTILE\_CONT(0.25) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC)) AS Q3\_plus1\_5\_IQR, 16 PERCENTILE\_CONT(0.25) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) - 1.5\* 17 (PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) 18 PERCENTILE\_CONT(0.25) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC)) Q1minus1\_5\_IQR 19 FROM public.gender\_pay\_gap\_21\_22 Data output Messages Notifications =+ 6 ~ 6 1 3 4 q1minus1\_5\_iqr q3\_plus1\_5\_igr double precision double precision

48.587500000000006 -22.112500000000004



While there are more missing values for diffmeanhourlypercent column, based on the IQR, there are more outliers. Additionally, as only percentage values are reported, there is no data on the absolute difference in monetary value for wages. And as income can vary largely, there tends to be extreme outliers that skew the mean salary. As such I choose to use median salary as the preferred basis of calculating pay gap, and hence choose **diffmedianhourlypercent**.

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?



Due to the variable being a calculated percentage, where the value can be positive or negative, while also not being calculated linearly, I find median to be a better way to find the average instead of mean.

Additionally, based on the mean and median of diffmeanhourlypercent and diffmedianhourlypercent, it shows that data is more skewed diffmedianhourlypercent. Which further supports to use of median for diffmedianhourlypercent.

As such, median of diffmedianhourlypercent is used as the appropriate metric to find average gender pay gap = 9.8%

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

- Difference in dollar value for pay gap is not captured in the data.
   With only percentage difference being given, this may not be a fair/accurate representation of the actual pay gap.
- ii. Percentage may be a good way of understanding the situation at a specific company as other variables like no. of staff is known. However, aggregating percentage across companies with varying sizes will likely lead to an outcome that is inaccurate from the actual pay gap situation (compared to aggregating using dollar pay gap, treating the whole nation as a big 'company'), as company size is not constant across companies and reported across a range



public.gender\_pay\_gap\_21\_22

Messages

3 ORDER BY diffmedianhourlypercent DESC

Notifications

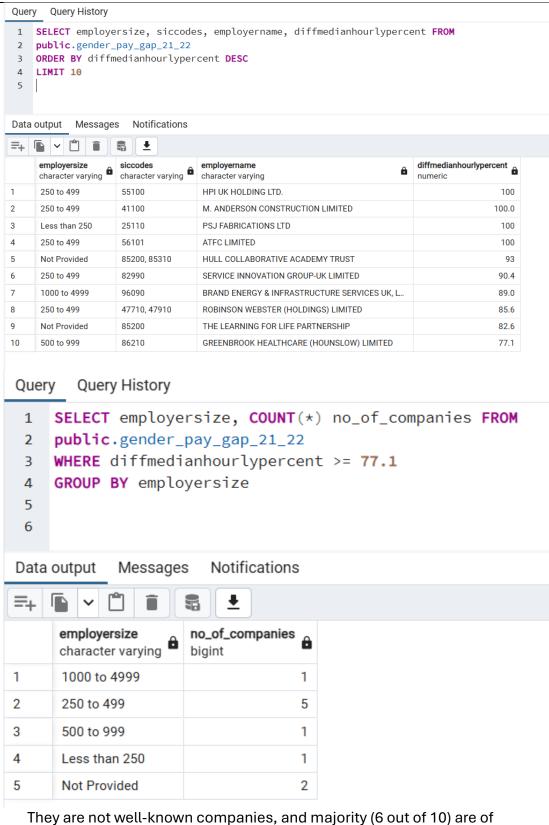
4 LIMIT 11

Data output

5 -- used LIMIT 11 to look for top 10 to investigate if there is a tie

=+ <u>\*</u> diffmedianhourlypercent employername character varying numeric 1 M. ANDERSON CONSTRUCTION LIMITED 100.0 2 ATFC LIMITED 100 HPI UK HOLDING LTD. 100 3 PSJ FABRICATIONS LTD 100 4 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 q THE LEARNING FOR LIFE PARTNERSHIP 82.6 10 GREENBROOK HEALTHCARE (HOUNSLOW) LIMITED 77.1 11 THE FALLIBROOME TRUST 74.4

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



They are not well-known companies, and majority (6 out of 10) are of relatively smaller companies (499 or less employees). This could be that with less employees there will be less salary data. Companies with 100% diffmedianhourlypercent are also assumed to not have any female employees, hence (M-W)/M \* 100% would be 100%.

