**How many companies are in the data set?**

SELECT COUNT(DISTINCT employerid) AS Number\_of\_companies

FROM public.gender\_pay\_gap\_21\_22

**A blue and white box with black text

AI-generated content may be incorrect.**

There are a total of 10174 companies in the dataset.

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

SELECT COUNT(DISTINCT employerid) AS Companies\_latereport

FROM public.gender\_pay\_gap\_21\_22

WHERE submittedafterthedeadline IS true

A close-up of a number

AI-generated content may be incorrect.

A total of 361 companies submitted their data after the reporting deadline.

**How many companies have not provided a URL?**

SELECT COUNT(DISTINCT employerid) AS Companies\_NoURL

FROM public.gender\_pay\_gap\_21\_22

WHERE companylinktogpginfo = '0'

A close-up of a number

AI-generated content may be incorrect.

A total of 3700 companies did not provide a URL.

W**hich measures of pay gap contain too much missing data, and should not be used in our analysis?**

SELECT SUM(CASE

WHEN CAST(diffmeanhourlypercent AS text) = '0' THEN 1

ELSE 0

END) AS count\_DMEANHP,

SUM(CASE

WHEN CAST(diffmedianhourlypercent AS text) = '0' THEN 1

ELSE 0

END) AS count\_DMEDHP,

SUM(CASE

WHEN CAST(diffmeanbonuspercent AS text) = '0' THEN 1

ELSE 0

END) AS count\_DMEANBP,

SUM(CASE

WHEN CAST(diffmedianbonuspercent AS text) = '0' THEN 1

ELSE 0

END) AS count\_DMEDBP,

SUM(CASE

WHEN CAST(malebonuspercent AS text) = '0' THEN 1

ELSE 0

END) AS count\_MBP,

SUM(CASE

WHEN CAST(femalebonuspercent AS text) = '0' THEN 1

ELSE 0

END) AS count\_FBP,

SUM(CASE

WHEN CAST(malelowerquartile AS text) = '0' THEN 1

ELSE 0

END) AS count\_MLQ,

SUM(CASE

WHEN CAST(femalelowerquartile AS text) = '0' THEN 1

ELSE 0

END) AS count\_FLQ,

SUM(CASE

WHEN CAST(malelowermiddlequartile AS text) = '0' THEN 1

ELSE 0

END) AS count\_MLMQ,

SUM(CASE

WHEN CAST(femalelowermiddlequartile AS text) = '0' THEN 1

ELSE 0

END) AS count\_FLMQ,

SUM(CASE

WHEN CAST(maleuppermiddlequartile AS text) = '0' THEN 1

ELSE 0

END) AS count\_MUMQ,

SUM(CASE

WHEN CAST(femaleuppermiddlequartile AS text) = '0' THEN 1

ELSE 0

END) AS count\_MUMQ,

SUM(CASE

WHEN CAST(maletopquartile AS text) = '0' THEN 1

ELSE 0

END) AS count\_MTQ,

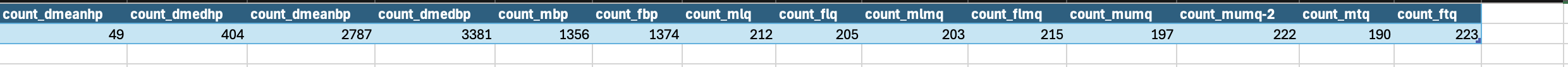
SUM(CASE

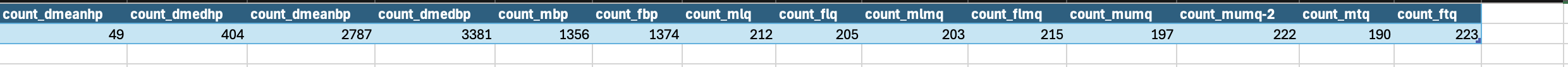
WHEN CAST(femaletopquartile AS text) = '0' THEN 1

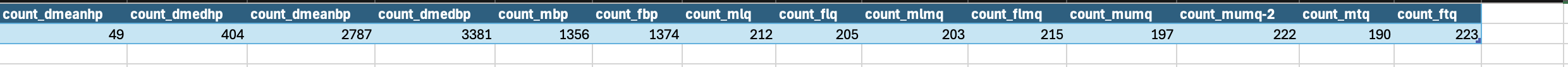
ELSE 0

END) AS count\_FTQ

FROM public.gender\_pay\_gap\_21\_22

****

****



Assume the integer 0 means there is missing data and 0.0 means no bias between men and women. The results shows that diffmeanbonuspercent, diffmedianbonuspercent has the highest number of missing values, followed by malebonuspercent and femalebonuspercent. The mentioned measures should not be used in the analysis.

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

SELECT

(SELECT(3 \* (AVG(diffmedianhourlypercent) - PERCENTILE\_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent)))/STDDEV(diffmedianhourlypercent) AS Skewness\_diffmedianhourlypercent

FROM public.gender\_pay\_gap\_21\_22

WHERE diffmedianhourlypercent::text <> '0'),

(SELECT(3 \* (AVG(diffmeanhourlypercent) - PERCENTILE\_CONT(0.50) WITHIN GROUP (ORDER BY diffmeanhourlypercent)))/STDDEV(diffmeanhourlypercent) AS Skewness\_diffmeanhourlypercent

FROM public.gender\_pay\_gap\_21\_22

WHERE diffmeanhourlypercent::text <> '0')

A screenshot of a cell phone

AI-generated content may be incorrect.

To calculate the pay gap, I would choose the DiffMeanHourlyPercent. Using Mean, Median and Standard Deviation to calculate skewness of data, the skewness of DiffMedianHourlyPercent is 0.41, compared to DiffMeanHourlyPercent of 0.14. DiffMedianHourlyPercent also has more missing data than DiffMeanHourlyPercent, with a ratio of 404 to 49.

**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 public.gender\_pay\_gap\_21\_22

WHERE diffmeanhourlypercent::text <> '0'

SELECT

(3 \* (AVG(diffmeanhourlypercent) - PERCENTILE\_CONT(0.50) WITHIN GROUP (ORDER BY diffmeanhourlypercent)))/STDDEV(diffmeanhourlypercent) AS Skewness

FROM public.gender\_pay\_gap\_21\_22

WHERE diffmeanhourlypercent::text <> '0'

A close-up of a number

AI-generated content may be incorrect.

A close-up of a computer screen

AI-generated content may be incorrect.

The average pay gap across all the companies is 13.7. The mean was used as the averaging metric. Calculating the skewness of the diffmeanhourlypercent data, which is 0.14, shows that the data is slightly positive skewed. Assuming that the skewness of the data is insignificant, the mean measure was used.

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

That there is an assumption that ‘0' means missing data, data is slightly positively skewed and that there are 49 missing data.

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

SELECT employername, employerid, diffmeanhourlypercent

FROM public.gender\_pay\_gap\_21\_22

ORDER BY diffmeanhourlypercent DESC

LIMIT 10

A screenshot of a computer

AI-generated content may be incorrect.

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

SELECT employername, employerid, diffmeanhourlypercent, employersize, siccodes

FROM public.gender\_pay\_gap\_21\_22

ORDER BY diffmeanhourlypercent DESC, employersize

LIMIT 10

A screenshot of a spreadsheet

AI-generated content may be incorrect.

It is a mixed of companies from different industries with varying employersizes. Assume that being a well-known company means that the larger the employersize, the more well known the company is. A minimum of 1000 employee is required for a company to be well known. Out of the top 10 companies, only 3 of them are well known.

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

SELECT employersize

FROM public.gender\_pay\_gap\_21\_22

GROUP BY employersize

SELECT employername, employerid, ABS(diffmeanhourlypercent), employersize, Siccodes

FROM public.gender\_pay\_gap\_21\_22

WHERE employersize = '20,000 or more'

ORDER BY diffmeanhourlypercent DESC

LIMIT 10

**A screenshot of a calculator

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

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

Pay Gap does not equal to Pay discrimination. However, comparing their diffmeanhourlypercent to the average diffmeanhourlypercent of 13.7, shows that these companies have a higher pay gap than most companies.

**What’s the average pay gap in London versus outside London?**

SELECT

(SELECT AVG(diffmeanhourlypercent) AS London\_Average

FROM public.gender\_pay\_gap\_21\_22

WHERE address ILIKE '%london%' AND diffmeanhourlypercent::text <> '0'),

(SELECT AVG(diffmeanhourlypercent) AS Not\_London\_Average

FROM public.gender\_pay\_gap\_21\_22

WHERE address NOT ILIKE '%london%' AND diffmeanhourlypercent::text <> '0')

A close-up of a number

AI-generated content may be incorrect.

**What’s the average pay gap in London versus Birmingham?**

SELECT

(SELECT AVG(diffmeanhourlypercent) AS London\_Average

FROM public.gender\_pay\_gap\_21\_22

WHERE address ILIKE '%london%' AND diffmeanhourlypercent::text <> '0'),

(SELECT AVG(diffmeanhourlypercent) AS Birmingham\_Average

FROM public.gender\_pay\_gap\_21\_22

WHERE address ILIKE '%birmingham%' AND diffmeanhourlypercent::text <> '0')

A close-up of a number

AI-generated content may be incorrect.

**What is the average pay gap within schools?**

SELECT AVG(diffmeanhourlypercent) AS School\_Average

FROM public.gender\_pay\_gap\_21\_22

WHERE siccodes LIKE '85%' AND diffmeanhourlypercent::text <> '0'

A blue and white rectangular sign with black numbers

AI-generated content may be incorrect.

With assumption that Siccodes under Education are all schools.

**What is the average pay gap within banks?**

SELECT AVG(diffmeanhourlypercent) AS Banks\_Average

FROM public.gender\_pay\_gap\_21\_22

WHERE siccodes BETWEEN '64110' AND '66300'

AND diffmeanhourlypercent::text <> '0'

A close-up of a number

AI-generated content may be incorrect.

With assumption that Siccodes under Financial and insurance activities are all Banks.

**Is there a relationship between the number of employees at a company and the average paygap?**

SELECT employersize, ROUND(AVG(diffmeanhourlypercent),2) AS Average\_DMHP

FROM public.gender\_pay\_gap\_21\_22

WHERE employersize <> 'Not Provided'

GROUP BY employersize

ORDER BY Average\_DMHP

A screenshot of a number

AI-generated content may be incorrect.

According to the results, there is a relationship between the number of employees at a company and the average paygap. The more employees at a company, the lower the average paygap.