Gender Pay Gap Analysis – SQL Project

Q/N	Question & Answer
1	How many companies are in the data set?

Query Query History

```
select count(employerid) no_of_companies
from public.gender_pay_gap_21_22
```

Data output Messages Notifications

	no_of_companies bigint
1	10174

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

Query Query History

```
SELECT COUNT(employerid) no_of_companies_submitted_late
-- counts number of companies that submitted late where submittedafterdeadline column is 'true'
from public.gender_pay_gap_21_22
where submittedafterthedeadline = true
```

Data output Messages Notifications

	no_of_companies_submitted_late bigint
1	361

3 How many companies have not provided a URL?

Query Query History

```
SELECT COUNT(employerid) companies_without_url FROM public.gender_pay_gap_21_22
-- counts number of employers that does not have 'http' in the companylinktogpginfo column
WHERE companylinktogpginfo NOT ILIKE 'http%'
-- ILIKE is used as a few URLs have capitalised 'HTTP' instead
```

Data output Messages Notifications

	companies_without_url bigint
1	3700

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

```
SELECT
(SELECT COUNT(*) no_of_missing_dmhp
FROM public.gender_pay_gap_21_22
WHERE diffmeanhourlypercent::text = '0'),
(SELECT COUNT(*) no_of_missing_dmdhp
```

```
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_dmdhp bigint	no_of_missing_dmbp bigint	no_of_missing_dmdbp bigint	no_of_missing_mbp bigint	no_of_missing_fbp bigint	no_of_missing_mlq bigint	no_of_missing_flq bigint
1	49	404	2787	3381	1356	1374	212	205

no_of_missing_mlmq bigint	no_of_missing_flmq bigint	no_of_missing_mumq bigint	no_of_missing_fumq bigint	no_of_missing_mtq bigint	no_of_missing_ftq bigint
203	215	197	222	190	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.
5	Choose which column you will use to calculate the pay gap. Will you use DiffMeanHourlyPercent or DiffMedianHourlyPercent? Can you justify your choice?

```sql
SELECT
PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) + 1.5*
(PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) -
PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC)) AS Q3_plus1_5_IQR,
PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) - 1.5*
(PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) -
PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC)) Q1minus1_5_IQR
FROM public.gender_pay_gap_21_22
```

Data output Messages Notifications

	q3_plus1_5_iqr double precision	q1minus1_5_iqr double precision
1	52.75	-30.049999999999997

```sql
SELECT
COUNT(*) no_of_outliers_dmdhp
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent > 52.75
OR diffmedianhourlypercent < -30.05
```

Data output Messages Notifications

	no_of_outliers_dmdhp bigint
1	199

```sql
SELECT
PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) + 1.5*
(PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) -
PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC)) AS Q3_plus1_5_IQR,
PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) - 1.5*
(PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) -
PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC)) Q1minus1_5_IQR
FROM public.gender_pay_gap_21_22
```

Data output Messages Notifications

	q3_plus1_5_iqr double precision	q1minus1_5_iqr double precision
1	48.587500000000006	-22.112500000000004

```sql
SELECT
COUNT(*) no_of_outliers_dmhp
FROM public.gender_pay_gap_21_22
WHERE diffmeanhourlypercent > 48.5875
OR diffmeanhourlypercent < -22.1125
```

Data output | Messages | Notifications

	no_of_outliers_dmhp bigint
1	211

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**.

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?

```sql
SELECT
AVG(diffmeanhourlypercent) mean_dmhp,
PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmeanhourlypercent ASC) median_dmhp,
AVG(diffmedianhourlypercent) mean_dmdhp,
PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_dmdhp
FROM public.gender_pay_gap_21_22
```

Data output | Messages | Notifications

	mean_dmhp numeric	median_dmhp double precision	mean_dmdhp numeric	median_dmdhp double precision
1	13.6350599567525064	12.9	12.3131708275997641	9.8

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%

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

i. 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

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

Query Query History

```
SELECT employername, diffmedianhourlypercent FROM
public.gender_pay_gap_21_22
ORDER BY diffmedianhourlypercent DESC
LIMIT 11
-- used LIMIT 11 to look for top 10 to investigate if there is a tie
```

Data output Messages Notifications

	employername character varying	diffmedianhourlypercent numeric
1	M. ANDERSON CONSTRUCTION LIMITED	100.0
2	ATFC LIMITED	100
3	HPI UK HOLDING LTD.	100
4	PSJ FABRICATIONS LTD	100
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
9	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?

Query Query History

```
SELECT employersize, siccodes, employername, diffmedianhourlypercent FROM
public.gender_pay_gap_21_22
ORDER BY diffmedianhourlypercent DESC
LIMIT 10
```

Data output Messages Notifications

	employersize character varying	siccodes character varying	employername character varying	diffmedianhourlypercent numeric
1	250 to 499	55100	HPI UK HOLDING LTD.	100
2	250 to 499	41100	M. ANDERSON CONSTRUCTION LIMITED	100.0
3	Less than 250	25110	PSJ FABRICATIONS LTD	100
4	250 to 499	56101	ATFC LIMITED	100
5	Not Provided	85200, 85310	HULL COLLABORATIVE ACADEMY TRUST	93
6	250 to 499	82990	SERVICE INNOVATION GROUP-UK LIMITED	90.4
7	1000 to 4999	96090	BRAND ENERGY & INFRASTRUCTURE SERVICES UK, L...	89.0
8	250 to 499	47710, 47910	ROBINSON WEBSTER (HOLDINGS) LIMITED	85.6
9	Not Provided	85200	THE LEARNING FOR LIFE PARTNERSHIP	82.6
10	500 to 999	86210	GREENBROOK HEALTHCARE (HOUNSLOW) LIMITED	77.1

Query Query History

```
SELECT employersize, COUNT(*) no_of_companies FROM
public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent >= 77.1
GROUP BY employersize
```

Data output Messages Notifications

	employersize character varying	no_of_companies bigint
1	1000 to 4999	1
2	250 to 499	5
3	500 to 999	1
4	Less than 250	1
5	Not Provided	2

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%.

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

Query Query History

```
SELECT employersize, employername, diffmedianhourlypercent FROM
public.gender_pay_gap_21_22
WHERE employersize IN ('5000 to 19,999', '20,000 or more')
ORDER BY diffmedianhourlypercent DESC
LIMIT 10
```

Data output Messages Notifications

	employersize character varying	employername character varying	diffmedianhourlypercent numeric
1	5000 to 19,999	EASYJET AIRLINE COMPANY LIMITED	63.6
2	5000 to 19,999	INDEPENDENT VETCARE LIMITED	49.4
3	5000 to 19,999	CVS (UK) LIMITED	42.9
4	5000 to 19,999	H&M HENNES & MAURITZ UK LIMITED	42.6
5	5000 to 19,999	SAVILLS (UK) LIMITED	41.1
6	5000 to 19,999	VETPARTNERS PRACTICES LIMITED	41
7	20,000 or more	LLOYDS BANK PLC	40.9
8	5000 to 19,999	STONEGATE PUB COMPANY LIMITED	39.2
9	5000 to 19,999	CIVICA UK LIMITED	36.5
10	5000 to 19,999	EDF ENERGY LIMITED	36.3

11	How would you report on the results? Can we say that these companies are engaging in unlawful pay discrimination?
	As these companies are large (5000+ employees), there should be sufficient data points in the salary data to see an accurate representation of the pay gap. With that in mind, before drawing any conclusions in terms of unlawful pay discrimination, I would investigate further into the individual companies as to what are the reasons the pay gap is this large. For example, in airline companies where typically Pilots tend to be male and cabin crew tend to be female, it could explain the large pay gap.
12	What's the average pay gap in London versus outside London?

Query Query History

```
SELECT
(SELECT PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_paygap_percent_London
FROM public.gender_pay_gap_21_22
WHERE address ILIKE ('%London%')
AND diffmedianhourlypercent::text != '0'),
-- filters away values that are deemed to be missing
(SELECT PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_paygap_percent_outside_London
FROM public.gender_pay_gap_21_22
WHERE address NOT ILIKE ('%London%')
AND diffmedianhourlypercent::text != '0')
-- filters away values that are deemed to be missing
```

Data output Messages Notifications

	median_paygap_percent_london double precision	median_paygap_percent_outside_london double precision
1	12.9	10

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

Query Query History

```sql
SELECT
(SELECT PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_paygap_percent_London
FROM public.gender_pay_gap_21_22
WHERE address ILIKE ('%London%')
AND diffmedianhourlypercent::text != '0'),
-- filters away values that are deemed to be missing
(SELECT PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_paygap_percent_Birmingham
FROM public.gender_pay_gap_21_22
WHERE address ILIKE ('%Birmingham%')
AND diffmedianhourlypercent::text != '0')
-- filters away values that are deemed to be missing
```

Data output Messages Notifications

	median_paygap_percent_london double precision	median_paygap_percent_birmingham double precision
1	12.9	8.8

14 What is the average pay gap within schools?

Section P	Education
85100	Pre-primary education
85200	Primary education
85310	General secondary education
85320	Technical and vocational secondary education
85410	Post-secondary non-tertiary education
85421	First-degree level higher education
85422	Post-graduate level higher education
85510	Sports and recreation education
85520	Cultural education
85530	Driving school activities
85590	Other education n.e.c.
85600	Educational support services

My assumption is Schools refer to any / group of education institutes. To filter these companies, I used the uk gov website for siccodes (https://resources.companieshouse.gov.uk/sic/) under Section P – Education.

Query Query History

```sql
SELECT PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_paygap_percent_schools
FROM public.gender_pay_gap_21_22
WHERE siccodes LIKE '%85100%'
OR siccodes LIKE '%85200%'
OR siccodes LIKE '%85310%'
OR siccodes LIKE '%85320%'
OR siccodes LIKE '%85410%'
OR siccodes LIKE '%85421%'
OR siccodes LIKE '%85422%'
OR siccodes LIKE '%85510%'
OR siccodes LIKE '%85520%'
OR siccodes LIKE '%85530%'
OR siccodes LIKE '%85590%'
OR siccodes LIKE '%85600%'
```

Data output Messages Notifications

	median_paygap_percent_schools double precision
1	21

15 What is the average pay gap within banks?

Section K	Financial and insurance activities
64110	Central banking
64191	Banks

My assumption is Banks refer to the 2 above siccodes (https://resources.companieshouse.gov.uk/sic/)

Query Query History

```
SELECT PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_paygap_percent_banks
FROM public.gender_pay_gap_21_22
WHERE siccodes LIKE '%64110%'
OR siccodes LIKE '%64191%'
```

Data output Messages Notifications

	median_paygap_percent_banks double precision
1	32.4

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

Query Query History

```
SELECT employersize, PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_paygap_percent
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent::text != '0'
GROUP BY employersize
HAVING employersize != 'Not Provided'
ORDER BY median_paygap_percent DESC
```

Data output Messages Notifications

	employersize character varying	median_paygap_percent double precision
1	250 to 499	11.5
2	500 to 999	10.350000000000001
3	Less than 250	10.2
4	1000 to 4999	9.8
5	5000 to 19,999	9.5
6	20,000 or more	7.55

Query Query History

```
SELECT employersize, PERCENTILE_CONT(0.50) WITHIN GROUP (ORDER BY diffmedianhourlypercent ASC) median_payg
FROM public.gender_pay_gap_21_22
WHERE diffmedianhourlypercent::text != '0'
GROUP BY employersize
HAVING employersize != 'Not Provided'
ORDER BY median_paygap_percent DESC
```

Data output Messages Notifications

	employersize character varying	median_paygap_percent double precision
1	250 to 499	11.5
2	500 to 999	10.350000000000001
3	Less than 250	10.2
4	1000 to 4999	9.8
5	5000 to 19,999	9.5
6	20,000 or more	7.55