UK Gender Pay Gap Analyses with SQL

1. How many companies are in the data set?

```
--count number of distinct employername--
SELECT COUNT (DISTINCT employername)
FROM gender_pay_gap_21_22;
```

Answer: 10174

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

```
--count number of rows where submittedafterthedeadline is TRUE--
SELECT COUNT (submittedafterthedeadline)
FROM gender_pay_gap_21_22
WHERE submittedafterthedeadline = 'TRUE';
```

Answer: 361

3. How many companies have not provided a URL?

```
--count number of rows where companylinktogpginfo is 0--
SELECT COUNT (companylinktogpginfo)
FROM gender_pay_gap_21_22
WHERE companylinktogpginfo = '0';
```

Answer: 3700

4. 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?

```
--calculate average of diffmeanhourlypercent, rounded to 2 decimal places--
SELECT ROUND (AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22;

--calculate 10th, 50th, 90th percentile of diffmeanhourlypercent--
SELECT PERCENTILE_CONT(0.1) WITHIN GROUP (ORDER BY diffmeanhourlypercent),
    PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY diffmeanhourlypercent),
    PERCENTILE_CONT(0.9) WITHIN GROUP (ORDER BY diffmeanhourlypercent)
FROM gender_pay_gap_21_22;
```

Answer: The average gender pay gap is 13.64%, biased towards men.
Mean was used as the measure of central tendency. Looking at the percentiles (10th: -1, 50th: 12.9, 90th: 30.5), the distribution of DiffMeanHourlyPercent appears to follow a relatively normal distribution with Kelly's skewness coefficient[1] of 0.11.

[1] Kelly's coefficient of skewness is calculated by $[(P_{90} + P_{10} - 2P_{50}) / (P_{90} - P_{10})]$

Since skewness is not of concern and mean is an easier concept for stakeholders to understand, mean was chosen as the averaging metric.

5. What are some caveats we need to be aware of when reporting the figure we've just calculated?
   Answer: First, the mean is sensitive to extreme outliers and may not be an accurate measure of central tendency. Second, it is assumed that '0' values in DiffMeanHourlyPercent (0.97% of data) reflects a true zero difference in mean hourly pay, instead of missing values.

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

```
--display top 10 employernames by diffmeanhourlypercent--
SELECT employername, diffmeanhourlypercent
FROM gender_pay_gap_21_22
ORDER BY 2 DESC
LIMIT 10;
```

Employer Name	Pay Gap (%)
HPI UK HOLDING LTD.	100
PSJ FABRICATIONS LTD	100
M. ANDERSON CONSTRUCTION LIMITED	100
BIRMINGHAM CITY FOOTBALL CLUB PLC	99
ACUSHNET EUROPE LTD	96.8
HOOK 2 SISTERS LIMITED	92
CHELSEA FOOTBALL CLUB LIMITED	91.6
BRAND ENERGY & INFRASTRUCTURE SERVICES UK, LTD.	91
MANCHESTER CITY FOOTBALL CLUB LIMITED	91
NEWCASTLE UNITED FOOTBALL COMPANY LIMITED	90.4

7. What do you notice about the results? Are these well-known companies?
   Answer: Four of these 10 companies are famous men's football clubs in the UK. The large pay gap in these football clubs could be a result of differences in roles and responsibilities (i.e., football players vs. support staff).

8. What's the average pay gap in London versus outside London?

```
--calculate average of diffmeanhourlypercent based on location, rounded to 2d.p.--
SELECT CASE
    WHEN address LIKE '%London,%' THEN 'London'
    ELSE 'Outside London'
    END AS location,
    ROUND(AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22
GROUP BY 1;
```

Location	Average Pay Gap (%)
London	15.88
Outside London	13.04

Answer: The average pay gap in London is 2.84 percentage points higher than elsewhere, biased towards men.

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

```
--calculate average of diffmeanhourlypercent based on location, rounded to 2d.p.--
SELECT CASE
    WHEN address LIKE '%London,%' THEN 'London'
    WHEN address LIKE '%Birmingham,%' THEN 'Birmingham'
    END AS location,
    ROUND(AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22
GROUP BY 1;
```

Location	Average Pay Gap (%)
London	15.88
Birmingham	13.15

Answer: The average pay gap in London is 2.73 percentage points higher than Birmingham, biased towards men.

10. What is the average pay gap within schools?

Note: Schools are defined as pre-primary to higher education institutions.

```
--calculate average of diffmeanhourlypercent for schools, rounded to 2d.p.--
SELECT ROUND(AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22
WHERE siccodes ~ ANY(ARRAY['85100','85200','85310','85320',
                           '85410','85421','85422']);
```

Answer: The average pay gap within schools is 16.59%, biased towards men.

11. What is the average pay gap within banks?

```
--calculate average of diffmeanhourlypercent for banks, rounded to 2d.p.--
SELECT ROUND(AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22
WHERE siccodes LIKE '%64191%';
```

Answer: The average pay gap within banks is 31.74%, biased towards men.

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

```sql
--display average diffmeanhourlypercent grouped by employersize, rounded to 2d.p.--
SELECT employersize, ROUND(AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22
GROUP BY 1;
```

Number of Employees	Average Pay Gap (%)
Less than 250	14.06
250 to 499	13.92
500 to 999	13.67
1000 to 4999	12.91
5000 to 19,999	14.12
20,000 or more	12.48

Answer: Average pay gap appears consistent (~12-14%) across company sizes, biased towards men.

13. How many companies are biased towards men, women, or neither?

```sql
SELECT COUNT(*), CASE
    WHEN diffmeanhourlypercent > 1 THEN 'men'
    WHEN diffmeanhourlypercent < -1 THEN 'women'
    ELSE 'neither'
    END AS bias
FROM gender_pay_gap_21_22
GROUP BY 2;
```

Bias	Number of Companies
Towards men	8621
Towards women	1009
Neither	544

14. What is the average pay gap in each of the 4 regions in the UK?

```sql
--calculate average of diffmeanhourlypercent based on country, rounded to 2d.p.--
SELECT CASE
    WHEN postcode = '0' THEN NULL
    WHEN LEFT(postcode,2) = 'BT' THEN 'Northern Ireland'
    WHEN LEFT(postcode,2) IN ('CF','CH','GL','HR','LD','LL','NP','SA','SY') THEN 'Wales'
    WHEN LEFT(postcode,2) IN ('AB','DD','DG','EH','FK','HS','IV','KA','KW','KY','ML',
                              'PA','PH','TD','ZE','G1','G2','G3','G4','G5','G6','G7',
                              'G8','G9','G0') THEN 'Scotland'
    ELSE 'England'
    END AS location,
    ROUND(AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22
GROUP BY 1
ORDER BY 2 DESC;
```

Region	Average Pay Gap (%)
England	13.93
Northern Ireland	12.53
Scotland	10.99

Wales	10.34

15. What is the average pay gap in the Top 10 cities in England?

```
--calculate average of diffmeanhourlypercent based on location, rounded to 2d.p.--
SELECT CASE
    WHEN address LIKE '%London,%' THEN 'London'
    WHEN address LIKE '%Birmingham,%' THEN 'Birmingham'
    WHEN address LIKE '%Liverpool,%' THEN 'Liverpool'
    WHEN address LIKE '%Bristol,%' THEN 'Bristol'
    WHEN address LIKE '%Manchester,%' THEN 'Manchester'
    WHEN address LIKE '%Sheffield,%' THEN 'Sheffield'
    WHEN address LIKE '%Leeds,%' THEN 'Leeds'
    WHEN address LIKE '%Leicester,%' THEN 'Leicester'
    WHEN address LIKE '%Coventry,%' THEN 'Coventry'
    WHEN address LIKE '%Bradford,%' THEN 'Bradford'
    END AS location,
    ROUND(AVG(diffmeanhourlypercent),2)
FROM gender_pay_gap_21_22
GROUP BY 1
ORDER BY 2 DESC;
```

City	Average Pay Gap (%)
Leicester	16.87
London	15.88
Birmingham	13.15
Coventry	13.14
Leeds	12.93
Bristol	12.74
Liverpool	12.61
Manchester	12.51
Bradford	12.43
Sheffield	10.53

16. What is the average pay gap in each industry?

```
/* calculate average of diffmeanhourlypercent for each industry by siccodes,
rounded to 2d.p.*/
SELECT ROUND(AVG(diffmeanhourlypercent),2) AS mining_and_quarrying
FROM gender_pay_gap_21_22
WHERE siccodes ~ ANY(ARRAY['5101','5102','5200','6100','6200','7100',
                           '7210','7290','8110','8120','8910','8920',
                           '8930','8990','9100','9900']);
```

Answer: 14.98 for Mining and Quarrying

Note: This query is repeated for other industries using the list of SIC codes in Qn 5.

Section	Name	Average Pay Gap (%)
A	Agriculture, Forestry and Fishing	17.06
B	Mining and Quarrying	14.98

C	Manufacturing	10.70
D	Electricity, gas, steam and air conditioning supply	15.29
E	Water supply, sewerage, waste management and remediation activities	6.23
F	Construction	19.83
G	Wholesale and retail trade; repair of motor vehicles and motorcycles	14.11
H	Transportation and storage	8.93
I	Accommodation and food service activities	8.39
J	Information and communication	17.70
K	Financial and insurance activities	24.94
L	Real estate activities	16.57
M	Professional, scientific and technical activities	18.26
N	Administrative and support service activities	10.90
O	Public administration and defence; compulsory social security	6.72
P	Education	15.81
Q	Human health and social work activities	9.60
R	Arts, entertainment and recreation	19.57
S	Other service activities	12.32
T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use	3.44
U	Activities of extraterritorial organisations and bodies	4.43

17. What is the average percentage of male and female employees who were paid bonuses, by bias?

```
--calculate average of each bonus by bias, rounded to 2d.p.--
SELECT CASE
    WHEN DiffMeanHourlyPercent > 1 THEN 'men'
    WHEN DiffMeanHourlyPercent < -1 THEN 'women'
    ELSE 'neither'
    END AS bias,
    ROUND(AVG(malebonuspercent),2) AS avg_mbp,
    ROUND(AVG(femalebonuspercent),2) AS avg_fbp
FROM gender_pay_gap_21_22
GROUP BY 1;
```

Bias	Average Malebonuspercent (%)	Average Femalebonuspercent (%)
Neither	25.58	26.19
Men	38.02	36.47
Women	30.22	30.42

18. For companies that showed a bias towards men, what is the average percentage of employees in each quartile by gender?

```
/*calculate average of each quartile for companies biased towards men,
rounded to 2d.p.*/
SELECT
    ROUND(AVG(malelowerquartile),2) AS avg_mlq,
    ROUND(AVG(femalelowerquartile),2) AS avg_flq,
    ROUND(AVG(malelowermiddlequartile),2) AS avg_mlmq,
    ROUND(AVG(femalelowermiddlequartile),2) AS avg_flmq,
    ROUND(AVG(maleuppermiddlequartile),2) AS avg_mumq,
    ROUND(AVG(femaleuppermiddlequartile),2) AS avg_fumq,
    ROUND(AVG(maletopquartile),2) AS avg_mtq,
    ROUND(AVG(femaletopquartile),2) AS avg_ftq
FROM gender_pay_gap_21_22
WHERE diffmeanhourlypercent > 1;
```

Quartile	Average
MaleLowerQuartile	42.47
FemaleLowerQuartile	56.10
MaleLowerMiddleQuartile	47.73
FemaleLowerMiddleQuartile	50.83
MaleUpperMiddleQuartile	52.88
FemaleUpperMiddleQuartile	45.68
MaleTopQuartile	59.45
FemaleTopQuartile	39.11