Answers to 16 Questions –

Initial Inspection

```
select *
from public.gender_pay_gap_21_22
limit 1000;
```

--Q.1 How many companies are in the data set? Ans.10174

```
select count(distinct employerid)
from gender_pay_gap_21_22;

select count(distinct employername)
from gender_pay_gap_21_22;
```

-- Q.2 How many of them submitted their data after the reporting deadline? Ans 361

```
select count(submittedafterthedeadline) as count_after_deadline
from gender_pay_gap_21_22
where submittedafterthedeadline = 'true';
```

--Q.3 How many companies have not provided a URL? Ans 3700

```
select count(employername)
from gender_pay_gap_21_22
where companylinktogpginfo IS NULL OR companylinktogpginfo = '0' or companylinktogpginfo = ' ';
```

--Q.4 Which measures of pay gap contain too much missing data, and should not be used in our analysis? Ans. there are no null values in dataset. Based on Q3 and Q16, employer name with '0' values and employer size with 'Not Provided' as values should not be considered for analysis. Also 794 values for sicodes don't seem correct as they have 0 or 1 (query below) which cannot be used for analysis.

```
select *
FROM gender_pay_gap_21_22
```

```
WHERE coalesce(diffmeanhourlypercent, diffmedianhourlypercent, diffmeanbonuspercent,
diffmedianbonuspercent, malebonuspercent, femalebonuspercent, malelowerquartile,

        femalelowerquartile, malelowermiddlequartile, femalelowermiddlequartile,
maleuppermiddlequartile, femaleuppermiddlequartile,

        maletopquartile, femaletopquartile) IS NULL;


select *

from public.gender_pay_gap_21_22

where siccodes = '0' or siccodes = '1' or siccodes = '1,'
```

--Q.5 Choose which column you will use to calculate the pay gap. Will you use DiffMeanHourlyPercent or DiffMedianHourlyPercent? Can you justify your choice

Ans. On an initial view of the mean and median values, there seems to be lot of variation among them, so I chose to calculate standard deviation and variance of both measures. The std deviation and variance of mean hourly percent is lower than median so I chose mean hourly percent to calculate pay gap.

dev_mean_hr_percent numeric	dev_median_hr_percent numeric	var_pop numeric	var_pop numeric
14.8645401269567651	16.1869647952114101	220.9545531859078417	262.0178292814135671

```
select employername, employersize, diffmeanhourlypercent, diffmedianhourlypercent

FROM gender_pay_gap_21_22;


SELECT

stddev_pop(DiffMeanHourlyPercent) as dev_mean_hr_percent,
stddev_pop(diffmedianhourlypercent) as dev_median_hr_percent,
var_pop(DiffMeanHourlyPercent), var_pop(diffmedianhourlypercent)

FROM gender_pay_gap_21_22;
```

--Q.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?

Ans. Based on my approach for Q4, diffmeanhourly percent seems to be an appropriate metric to find gender pay gap across all companies. 13.63 % is the mean across all companies in UK.

```
SELECT

avg(DiffMeanHourlyPercent)
```

FROM gender_pay_gap_21_22;

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

Ans. The mean across all companies is generalising all the data and not considering gender pay gap across regions, industries, employer size etc. so probably categorising the data into regions/ postcodes, industry type (siccodes are unclean, need to clean first) and then working out the averages would give abetter picture.

```
select distinct postcode, avg(diffmeanhourlypercent)

from public.gender_pay_gap_21_22

group by postcode
```

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

Ans.- 10 companies with the largest pay gaps.

```
SELECT *

FROM gender_pay_gap_21_22

WHERE diffmeanhourlypercent >1

ORDER BY diffmeanhourlypercent DESC

LIMIT 10;
```

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

Ans. Looking at the employername, currentname and sicodes column, we can see that largest pay gaps exist mostly in construction/ energy companies or football clubs. The size of all these companies are less than 4999.

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

Ans.checked the mean bonus percent and male top quartile percent >90, in addition to mean hourly percent, to see the largest pay gaps. The result was a list of 6 companies (construction and football) with big pay gaps across all metrics.

```
SELECT *

FROM gender_pay_gap_21_22
```

```
WHERE diffmeanhourlypercent >90 AND diffmeanbonuspercent >90 AND maletopquartile > 90
ORDER BY diffmeanhourlypercent DESC;
```

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

The result should be reported as per the industry bucket. The nature of work in the construction industry employs more males than females, hence larger gap. Same with Football, except this analysis can be used to encourage more female football clubs to start up, encouraging females to play more football as a sport.

--Q.12 What's the average pay gap in London versus outside London

Ans. London : 15.7%, Outside London : 13.04%

```
SELECT avg(diffmeanhourlypercent) as avg_pay_gap_London
FROM gender_pay_gap_21_22
WHERE address LIKE '%London%'

SELECT avg(diffmeanhourlypercent) as avg_pay_gap_London
FROM gender_pay_gap_21_22
WHERE address NOT LIKE '%London%'
```

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

Ans. London : 15.7%, Birmingham : 13.22%

```
SELECT avg(diffmeanhourlypercent) as avg_pay_gap_London
FROM gender_pay_gap_21_22
WHERE address LIKE '%London%'

SELECT avg(diffmeanhourlypercent) as avg_pay_gap_Birmingham
FROM gender_pay_gap_21_22
WHERE address LIKE '%Birmingham%'
```

--Q.14 What is the average pay gap within schools

Ans. 17.088% Checked by running the query with employername and current name, both have same result. Didn't use siccodes since data is not clean and contains '0' or '1' or ',1' values.

```
SELECT avg(diffmeanhourlypercent) as avg_pay_gap_school

FROM gender_pay_gap_21_22

WHERE employername LIKE '%School%'


select employername, siccodes,diffmeanhourlypercent

from public.gender_pay_gap_21_22

WHERE employername LIKE '%School%'

ORDER BY siccodes desc;


select currentname, siccodes,diffmeanhourlypercent

from public.gender_pay_gap_21_22

WHERE currentname LIKE '%School%'

ORDER BY siccodes desc;
```

--Q.15 What is the average pay gap within banks?

Ans. 19.2% Checked by running the query with employername and current name, both have same result. Didn't use siccodes since data is not clean and contains '0' or '1' or ',1' values. However there is only one bank in the data- Bank of England. To include other financial institutions can run the query with siccodes starting with '64' , '65' , or '66' .

```
select AVG(diffmeanhourlypercent)

from public.gender_pay_gap_21_22

WHERE employername LIKE '%Bank%'


select employername, currentname, diffmeanhourlypercent,siccodes

from public.gender_pay_gap_21_22

WHERE employername LIKE '%Bank%' OR currentname LIKE '%Bank%';
```

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

Ans. Meanhourly paygap is highest in companies with 5000 – 19999 employees. And bonus gap, male top quartile and female lower quartile gap is highest in companies with 20000 or more employees.

```
select employersize, round(AVG(diffmeanhourlypercent),2) AS avg_paygap,
ROUND(AVG(diffmeanbonuspercent),2) AS avg_bonusgap, ROUND(AVG(maletopquartile),2) AS
avg_maletop_quartile, round(AVG(femalelowerquartile),2) as avg_female_lower_quartile

from public.gender_pay_gap_21_22

WHERE employersize NOT LIKE 'Not Provided'

GROUP BY employersize

ORDER BY 2,3,4,5
```

Bonus Question (optional)

Gender pay gap across regions

1. Across industries
2. Age of company- date of incorporation
3. Companies with largest pay or bonus gap top 10
4. Smallest pay/bonus gap bottom 10