Data Analytics Immersive

Lab: Gender Pay Gap

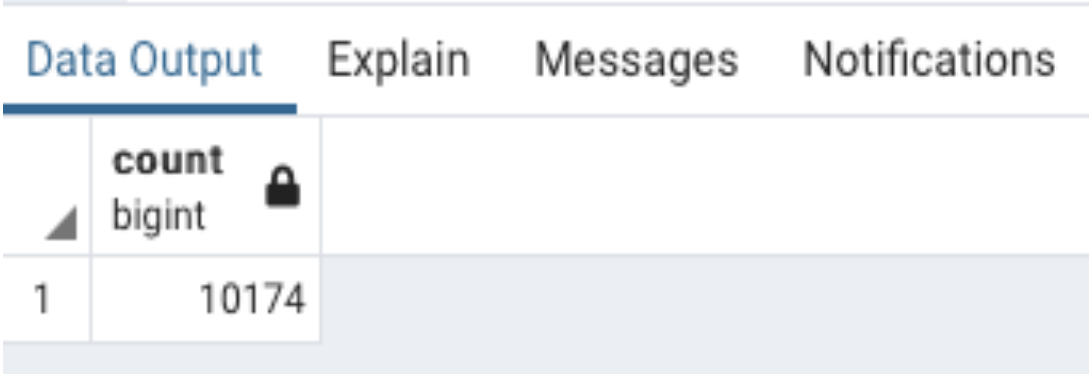
Unit: Data acquisition and cleaning with SQL

Maksim Kuchynski

Initial Inspection

1. How many companies are in the data set?

```
SELECT COUNT(DISTINCT employername)
FROM gender_pay_gap_21_22;
```

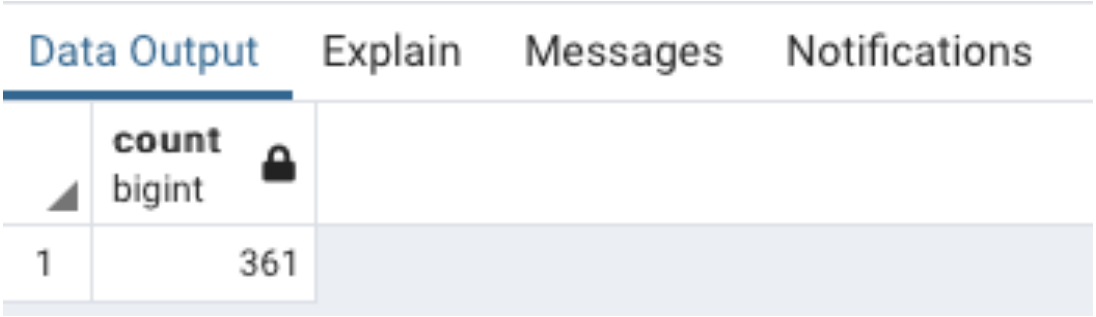


Data Output | Explain | Messages | Notifications

	count bigint
1	10174

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

```
SELECT COUNT(DISTINCT employername)
FROM gender_pay_gap_21_22
WHERE submittedafterthedeadline = TRUE;
```

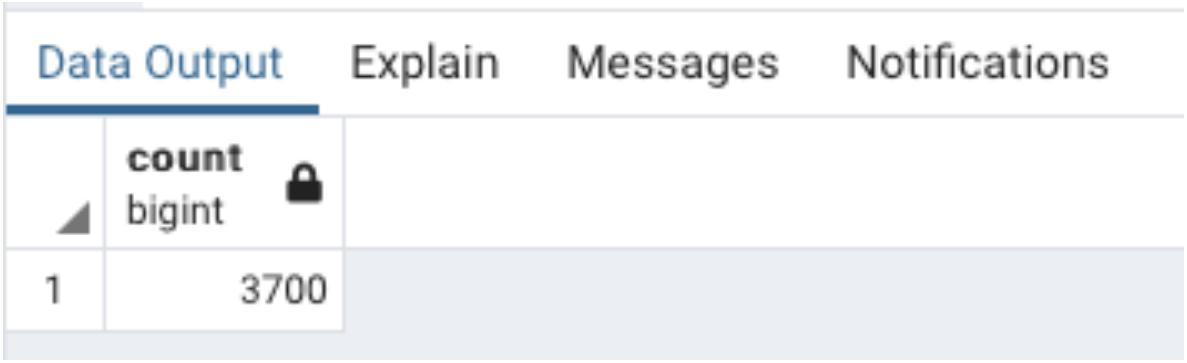


Data Output | Explain | Messages | Notifications

	count bigint
1	361

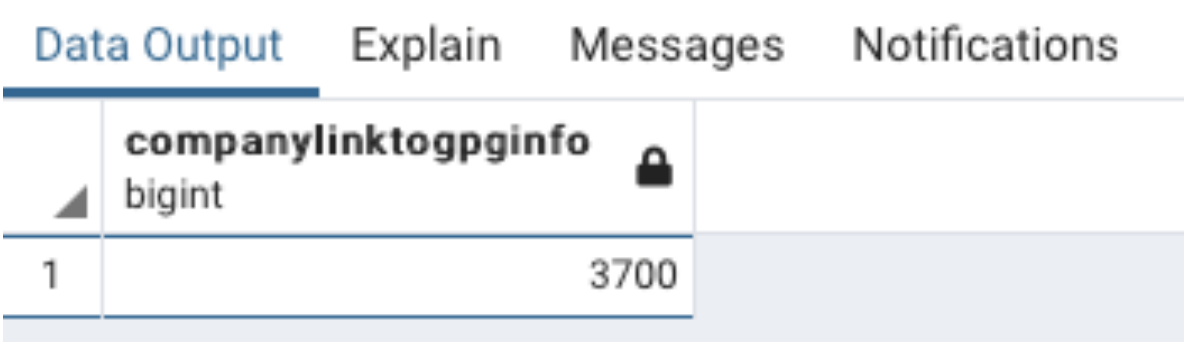
3. How many companies have not provided a URL?

```
SELECT COUNT(CompanyLinkToGPGInfo)
FROM gender_pay_gap_21_22
WHERE CompanyLinkToGPGInfo = '0';
```



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

```
SELECT COUNT(CompanyLinkToGPGInfo) AS
CompanyLinkToGPGInfo
FROM gender_pay_gap_21_22
WHERE CompanyLinkToGPGInfo = '0';
```



	companylinktogpginfo bigint
1	3700

Bonus (optional): Can you find out what the 'SicCodes' column corresponds to? Is there a way we can understand what each SIC code represents? Search online for extra information.

https://www.gov.uk/government/publications/standard-industrial-classification-of-economic-activities-sic

A Standard Industrial Classification (SIC) was first introduced into the UK in 1948 for use in classifying business establishments and other statistical units by the type of economic activity in which they are engaged. The classification provides a framework for the collection, tabulation, presentation and analysis of data, and its use promotes uniformity. In addition, it can be used for administrative purposes and by non-government bodies as a convenient way of classifying industrial activities into a common structure.

There are 21 sections, 88 divisions, 272 groups, 615 classes and 191 subclasses

How to Read a UK SIC Code

The UK SIC is based exactly on NACE but, where it was thought necessary or helpful, a fifth digit has been added to form subclasses of the NACE four digit classes. Thus, the

UK SIC is a hierarchical five digit system. UK SIC is divided into 21 sections, each denoted by a single letter from A to U. The letters of the sections can be uniquely defined by the next breakdown, the divisions (denoted by two digits). The divisions are then broken down into groups (three digits), then into classes (four digits) and, in several cases, again into subclasses (five digits).

Section	**C**	Manufacturing
Division	**10**	Manufacture of Food Products
Group	**10.1**	Manufacture of Dairy Products
Class	**10.51**	Operation of Dairy and Cheese Making
Subclass	**10.51/2**	Butter and Cheese Production

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

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?

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

First, let's see the difference between DiffMeanHourlyPercent and DiffMedianHourlyPercen.

```
SELECT ROUND(AVG(DiffMeanHourlyPercent), 2) AS
avg_DiffMeanHourlyPercent,
ROUND(AVG(DiffMedianHourlyPercent), 2) AS avg_DiffMedianHourlyPercent
FROM gender_pay_gap_21_22;
```

Data Output | Explain | Messages | Notifications

	avg_diffmeanhourlypercent numeric	avg_diffmedianhourlypercent numeric
1	13.64	12.31

MEANs are generally better understood, as they represent what most of us would typically refer to as an “average.” For this reason they appear to be much more quoted in discussions of the gender pay gap than MEDIANs. But there’s a problem with this. MEANs are very easily skewed by outliers, so if we have one or two people in your business who are extremely highly paid compared to the majority, your mean salary will be skewed upwards by these extreme cases.

If we use the MEAN, then, it is important to also understand the standard deviation of our dataset. The standard deviation is a measure of the spread of our dataset, which will help us to understand if we have outliers which may be skewing your result. A low standard deviation means our datapoints are all clustered relatively near to the mean; a high standard deviation means there is a wide spread of datapoints, on other words we may well have outliers.

If we have a dataset with outliers, the median may well prove a more useful statistic to measure as it is much more resistant to distortion and can be viewed as representing a more “usual” number in your dataset, in that it is the salary of the person who falls exactly in the middle when all employees are lined up in salary order.

In this case we have MEAN (13.64) and MEDIAN (12.31), Because there are generally fewer women in higher-paying roles than men, the gender pay gap as measured by MEAN earning is often higher than for median earnings.

Conclusion: Because we don’t know the distribution of gender and age in the companies from the dataset, it’s more accurate to use MEDIAN.

References:

1) https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/genderpaygapintheuk/2021
2) https://www.ft.com/content/9ee99efa-d9d9-11e7-a039-c64b1c09b482
3) https://www.linkedin.com/pulse/gender-pay-gap-1-understanding-means-medians-nancy-roberts/

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

```
SELECT employername, DiffMedianHourlyPercent, employersize
FROM gender_pay_gap_21_22
WHERE DiffMedianHourlyPercent > 0
ORDER BY DiffMedianHourlyPercent DESC
LIMIT 10;
```

Data Output | Explain | Messages | Notifications

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

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

Most of those companies are not well-known companies and have less then 500 employees.

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

```
SELECT employersize

FROM gender_pay_gap_21_22

GROUP BY employersize;
```

```
SELECT employername, DiffMedianHourlyPercent, employersize

FROM gender_pay_gap_21_22

WHERE DiffMedianHourlyPercent > 0 AND employersize = '20,000 or more'

ORDER BY DiffMedianHourlyPercent DESC

LIMIT 10;
```

	employersize character varying
1	20,000 or more
2	250 to 499
3	5000 to 19,999
4	1000 to 4999
5	Less than 250
6	Not Provided
7	500 to 999

Data Output | Explain | Messages | Notifications

	employername character varying	diffmedianhourlypercent numeric	employersize character varying
1	LLOYDS BANK PLC	40.9	20,000 or more
2	LLOYDS BANKING GROUP PLC	34.2	20,000 or more
3	NATIONAL WESTMINSTER BANK PUBLIC LIMITED COMPANY	34.2	20,000 or more
4	J D WETHERSPOON PLC	29.5	20,000 or more
5	HBOS PLC	27.8	20,000 or more
6	BARCLAYS EXECUTION SERVICES LIMITED	27.1	20,000 or more
7	NEXT RETAIL LIMITED	26.9	20,000 or more
8	WPP 2005 LIMITED	23.6	20,000 or more
9	BRITISH AIRWAYS PLC	22.2	20,000 or more
10	Leeds Teaching Hospitals Nhs Trust	21.2	20,000 or more

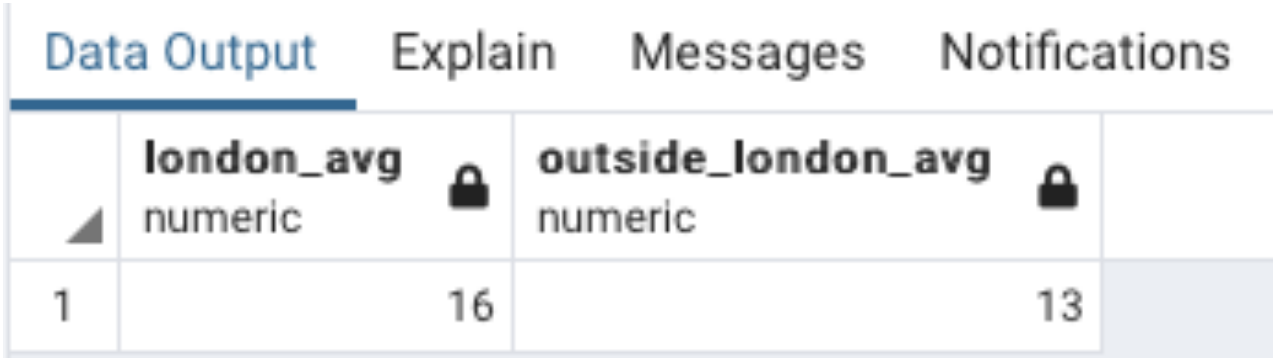
11. How would you report on the results? Can we say that these companies are engaging in unlawful pay discrimination?

The majority of the companies are finance companies with headquarter in London.

By law, we cannot blame any companies for engaging in unlawful pay discrimination without proof. The output shows the gender pay gap, but it does explain the reason.

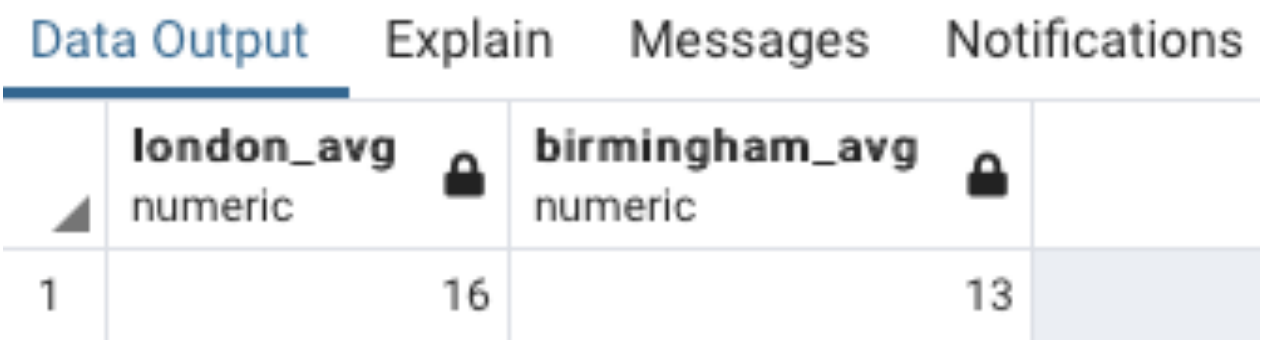
12. What's the average pay gap in London versus outside London?

```
SELECT ROUND(AVG(DiffMeanHourlyPercent)) AS london_avg,

(SELECT ROUND(AVG(DiffMeanHourlyPercent)) AS outside_london_avg

FROM gender_pay_gap_21_22 WHERE address NOT LIKE '%London%')

FROM gender_pay_gap_21_22

WHERE address LIKE '%London%';
```



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

```
SELECT ROUND(AVG(DiffMeanHourlyPercent)) AS london_avg,
(SELECT ROUND(AVG(DiffMeanHourlyPercent)) AS birmingham_avg
FROM gender_pay_gap_21_22 WHERE address LIKE '%Birmingham%')
FROM gender_pay_gap_21_22
WHERE address LIKE '%London%';
```



14. What is the average pay gap within schools?

```
SELECT ROUND(AVG(DiffMeanHourlyPercent)) AS avg_school
FROM gender_pay_gap_21_22
WHERE LOWER(employername) LIKE '%scho%' OR
LOWER(employername) LIKE '%univ%';
```

Data Output | Explain | Messages | Notifications

	avg_school numeric
1	15

15. What is the average pay gap within banks?

```
SELECT ROUND(AVG(DiffMeanHourlyPercent)) AS avg_bank
FROM gender_pay_gap_21_22
WHERE LOWER(employername) LIKE '%bank%';
```

Data Output | Explain | Messages | Notifications

	avg_bank numeric
1	26

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

```
SELECT ROUND(AVG(DiffMeanHourlyPercent),3) AS
av_DiffMeanHourlyPercent, employersize, CASE
        WHEN employersize = 'Less than 250' THEN '1'
        WHEN employersize = '250 to 499' THEN '2'
        WHEN employersize = '500 to 999' THEN '3'
        WHEN employersize = '1000 to 4999' THEN '4'
        WHEN employersize = '5000 to 19,999' THEN '5'
        WHEN employersize = '20,000 or more' THEN '6'
        ELSE '0'
END AS row_number
FROM gender_pay_gap_21_22
GROUP BY employersize
ORDER BY row_number;
```

Data Output | Explain | Messages | Notifications

	av_diffmeanhourlypercent numeric	employersize character varying	row_number text
1	13.032	Not Provided	0
2	14.057	Less than 250	1
3	13.921	250 to 499	2
4	13.667	500 to 999	3
5	12.907	1000 to 4999	4
6	14.121	5000 to 19,999	5
7	12.481	20,000 or more	6

Looking at the data output we can't see a strong correlation between company size and pay gap. In the next exercise I will try to find a correlation between lower/top quartile and company size.

Optional: Is there a correlation between the number of employees at a company and the men/women lower quartile?

hypothesis: As bigger company as more females in the lower quarter.

```
SELECT ROUND(AVG(MaleLowerQuartile),3) AS avg_male_lower_q,
	ROUND(AVG(FemaleLowerQuartile),3) AS avg_female_lower_q,
	ROUND(AVG(FemaleLowerQuartile),3) -
ROUND(AVG(MaleLowerQuartile),3) AS lower_q_diff_female_male,
	employersize,
	CASE
		WHEN employersize = 'Less than 250' THEN '1'
		WHEN employersize = '250 to 499' THEN '2'
		WHEN employersize = '500 to 999' THEN '3'
		WHEN employersize = '1000 to 4999' THEN '4'
		WHEN employersize = '5000 to 19,999' THEN '5'
		WHEN employersize = '20,000 or more' THEN '6'
		ELSE '0'
	END AS row_number
FROM gender_pay_gap_21_22
GROUP BY employersize
ORDER BY row_number;
```

Data Output | Explain | Messages | Notifications

	avg_male_lower_q numeric	avg_female_lower_q numeric	lower_q_diff_female_male numeric	employersize character varying	row_number text
1	37.623	60.949	23.326	Not Provided	0
2	44.64	51.788	7.148	Less than 250	1
3	45.245	53	7.755	250 to 499	2
4	44.843	53.078	8.235	500 to 999	3
5	44.076	54.798	10.722	1000 to 4999	4
6	39.545	58.73	19.185	5000 to 19,999	5
7	41.955	54.819	12.864	20,000 or more	6

The data output (lower_q_diff_female_male) shows strong correlation between the number of employees at a company and percentage of females in the lower hourly pay quarter. As bigger company as more females in the lower quarter.
Hypothesis confirmed

Let's try the top quarter:
hypothesis: As bigger company as less females in the top quarter.

```
SELECT ROUND(AVG(MaleTopQuartile),3) AS avg_male_top_q,
	ROUND(AVG(FemaleTopQuartile),3) AS avg_female_top_q,
	ROUND(AVG(MaleTopQuartile),3) -
ROUND(AVG(FemaleTopQuartile),3) AS top_q_pay_gap_male_female,
	employersize,
	CASE
		WHEN employersize = 'Less than 250' THEN '1'
		WHEN employersize = '250 to 499' THEN '2'
		WHEN employersize = '500 to 999' THEN '3'
		WHEN employersize = '1000 to 4999' THEN '4'
		WHEN employersize = '5000 to 19,999' THEN '5'
		WHEN employersize = '20,000 or more' THEN '6'
		ELSE '0'
	END AS row_number
FROM gender_pay_gap_21_22
GROUP BY employersize
ORDER BY row_number;
```

Data Output | Explain | Messages | Notifications

	avg_male_top_q numeric	avg_female_top_q numeric	top_q_diff_male_female numeric	employersize character varying	row_number text
1	49.571	49	0.571	Not Provided	0
2	58.966	37.462	21.504	Less than 250	1
3	59.311	38.934	20.377	250 to 499	2
4	58.643	39.278	19.365	500 to 999	3
5	57.929	40.944	16.985	1000 to 4999	4
6	52.879	45.397	7.482	5000 to 19,999	5
7	58.724	38.05	20.674	20,000 or more	6

The data output shows correlation between the number of employees at a company and percentage of females in the top hourly pay quarter. Surprisingly, as bigger company, as smaller the difference between the number of men and women. Except for companies with more than 20.000 employees.
Hypothesis refuted.