Exploratory Data Analysis Report - UK

Dataset: GPG_2017_2024_cleaned_v2.csv

Introduction

The dataset contains employer-reported statistics for UK gender pay gap filings. This report focuses on data quality, distributions, temporal trends, and relationships among reported metrics. To assess the quality and structure of the UK Gender Pay Gap dataset (2017–2024) and extract decision-ready insights by:

- **Validating data quality & coverage:** detect missingness, anomalies, and year-to-year reporting consistency.
- **Describing pay gaps:** quantify distributions and trends of mean/median hourly and bonus gaps over time.
- **Linking drivers:** examine how quartile representation, bonus eligibility, and employer size relate to the gaps.
- **Identifying outliers & biases:** flag implausible values and years with coverage risk (e.g., 2019).
- **Informing action:** provide robust, sensitivity-checked findings to guide deeper modeling or policy analysis.

Method

We analyzed `GPG_2017_2024_cleaned_v2.csv` (2017–2024) in Python with pandas/numpy/matplotlib, coercing types and treating percentages as point values. Coverage is profiled overall and by year (using non-null employer fields), then we summarize key metrics: hourly and bonus gaps, bonus eligibility, and quartile composition, via unweighted means/medians, boxplots, and year-over-year trends. For structure, we inspect the latest year: employer-size mix, mean-vs-median scatter with correlations, and a numeric correlation heatmap linking gaps to quartile shares and eligibility. Data quality is handled by clipping implausible percentages, normalizing booleans, and running sensitivities (exclude 2019, exclude “Unknown” size, optional headcount weighting and stable-cohort checks).

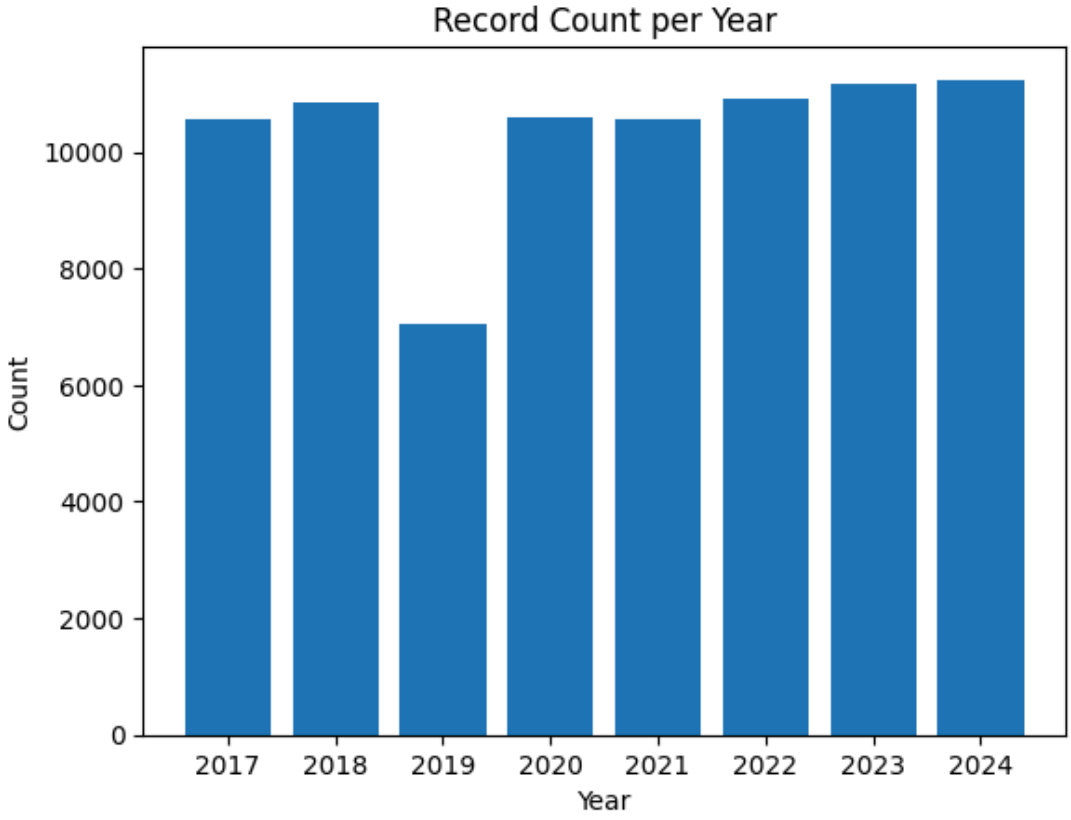
Insights

1. Dataset Overview & Quality

Rows	15203
Columns	209
Overall Missingness	33.6%

Dtype	Count
float64	112
object	96
int64	1

2. Record Count per Year



The 2019 drop likely reflects **coverage/ingestion differences** (under-reporting, schema changes, or missing 2019-prefixed fields) rather than a real shift in employer population.

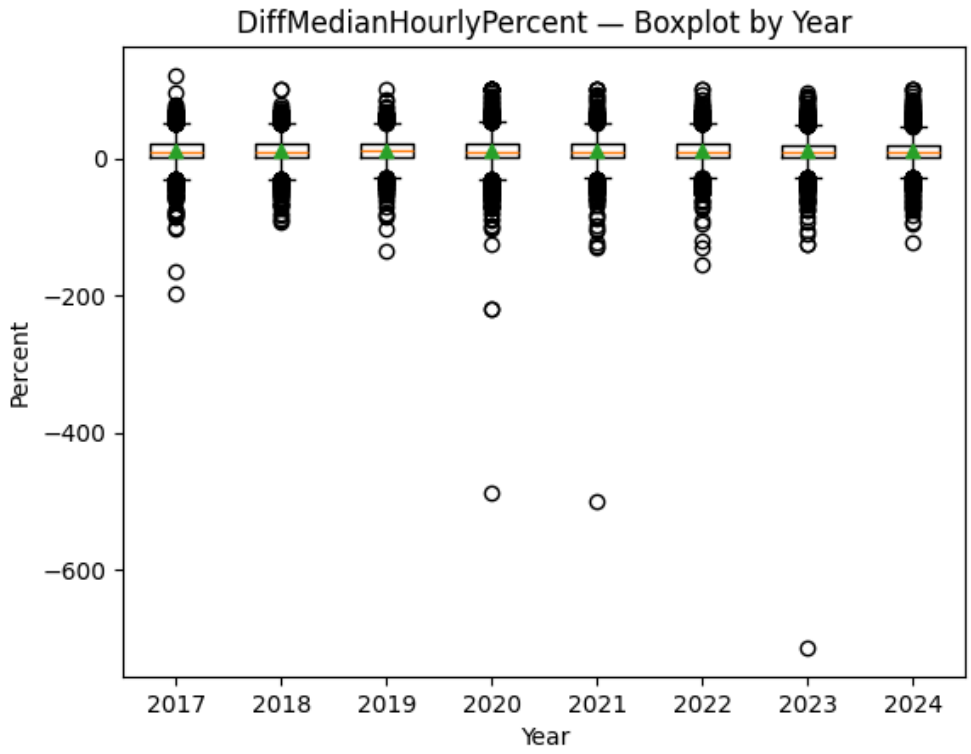
3. Distribution of *DiffMedianHourlyPercent* (by Year)

- **Central tendency:** Medians are consistently > 0% in every year, indicating men earn more than women at the median in most employers.

- **Spread:** Interquartile ranges are fairly stable year-to-year, with a *gradual leftward shift* (narrowing gap) that matches the mean-trend analysis.

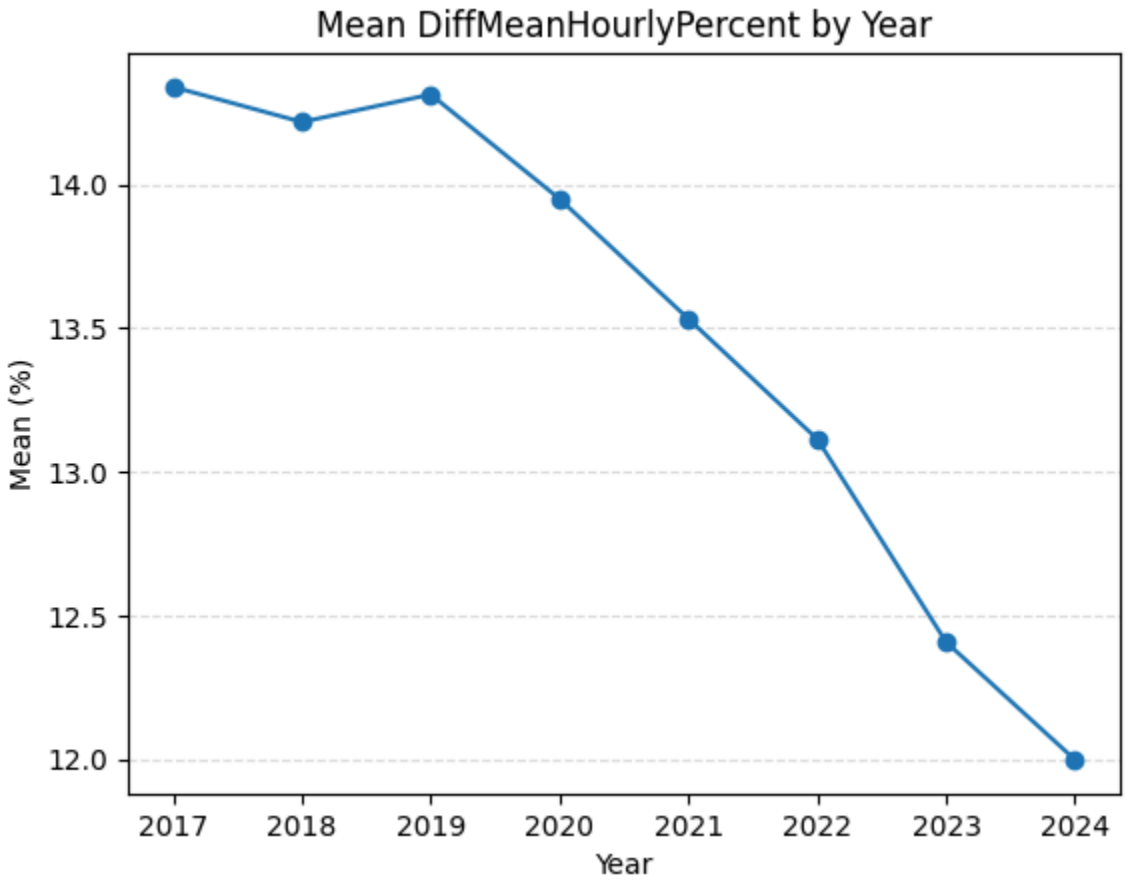
- **Outliers:** There are extreme negative values (e.g., below –200% and one point near –600%), which are not plausible for percentage pay gaps and almost certainly reflect data entry or parsing errors (e.g., wrong sign, wrong unit, or malformed rows).

- **Practical implication:** Use robust statistics (median, trimmed mean) and winsorize or clip implausible values for clearer year-over-year comparisons. Always report sensitivity with/without outliers.



4. Mean Hourly Gender Pay Gap (DiffMeanHourlyPercent) by Year

- The average *DiffMeanHourlyPercent* shows a steady, monotonic decline from 14.34% (2017) to 12.00% (2024)
- This equates to a –2.34 percentage point improvement between 2017 and 2024 (≈ –16.3% relative to 2017).
- The shape is gradual rather than stepwise, suggesting incremental progress rather than a one-off shift.



Insight:

We can see a clear coverage dip in 2019. While the 2019 mean aligns with 2017–2018, we’’ll report results with and without 2019 as a robustness check. We'll also verify completeness for 2024 fields, if many employers submitted partial data, the 2024 mean could be optimistically biased.

5. Mean Bonus Gap (*DiffMeanBonusPercent*) by Year

Interpretation

- Bonus gaps are consistently positive and substantially exceed the hourly gap, implying men are more likely to receive bonuses and/or receive larger bonuses when they do.

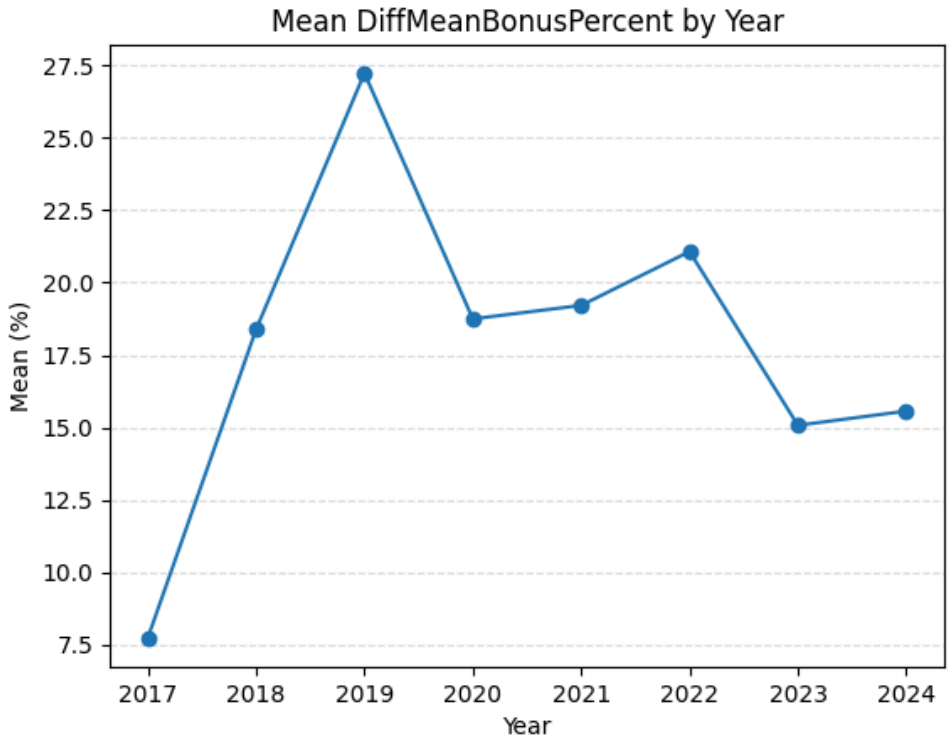
- The 2019 spike stands out. Given the coverage dip in 2019 (fewer valid employer rows), this peak may be *partly a data/coverage artifact* rather than a real system-wide jump.

- Post-2019, the bonus gap trends downward but remains ***elevated (15–21%)*** versus the hourly gap, consistent with bonuses being more skewed at the top of pay distributions.

Caveats

- Bonus metrics depend heavily on *eligibility reporting and non-missing* bonus fields; uneven missingness can bias means.

- Outliers can inflate the average.



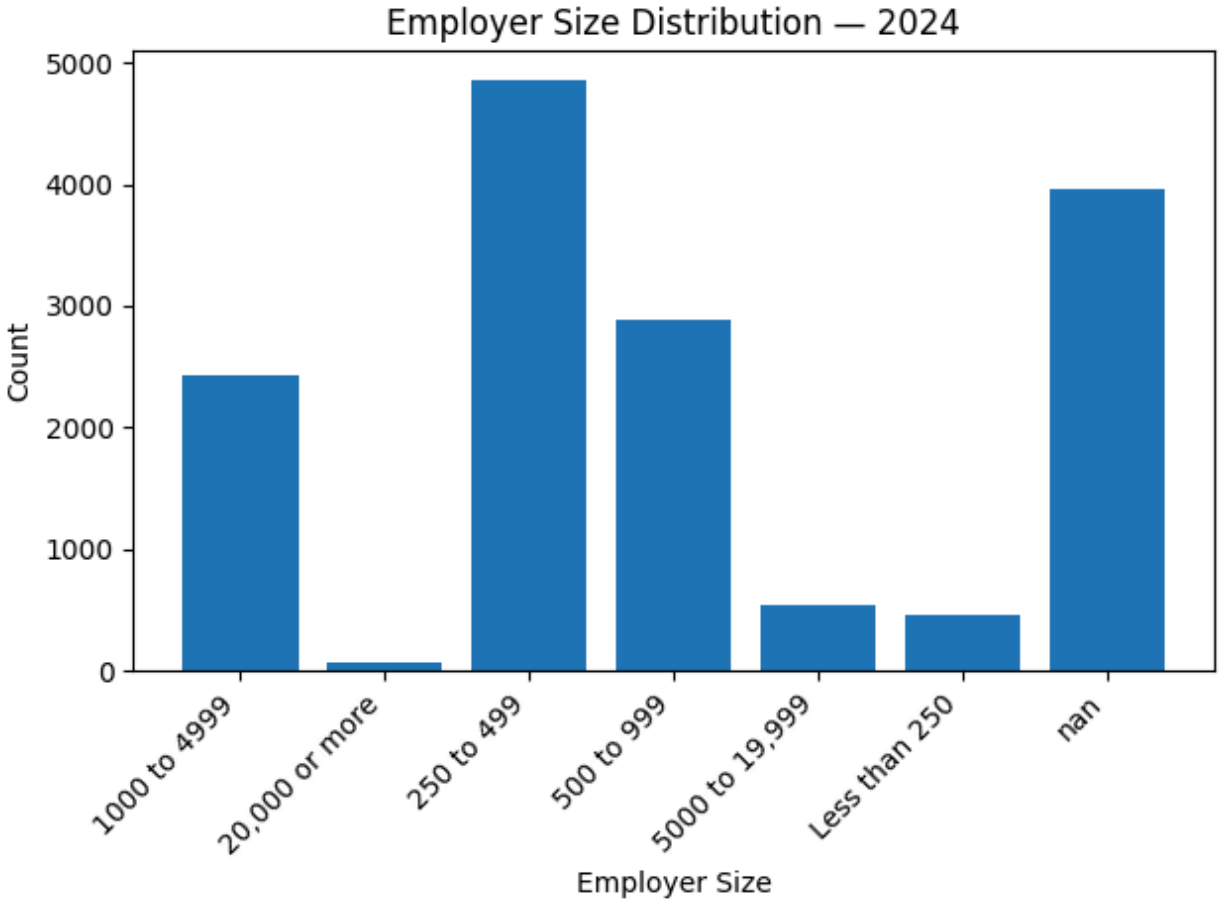
6. Employer Size Distribution - 2024

- Reporting is concentrated in mid-sized employers: the largest bucket is **"***250 to 499", followed by "500 to 999."*

- There is meaningful representation of small firms *("Less than 250")* and large firms *("1000 to 4999," "5000 to 19,999," "20,000 or more")*, but these buckets are smaller than the mid-size categories.

- A **large *"NaN/Unknown"*** bar indicates many submissions did not include employer size.

  ~ Implication: any size-stratified pay-gap comparisons may be **biased** if missingness is systematic (e.g., small firms more likely to omit size).
  ~ **Action:** We can treat "Unknown" as a separate group, and run sensitivity analyses excluding unknowns.



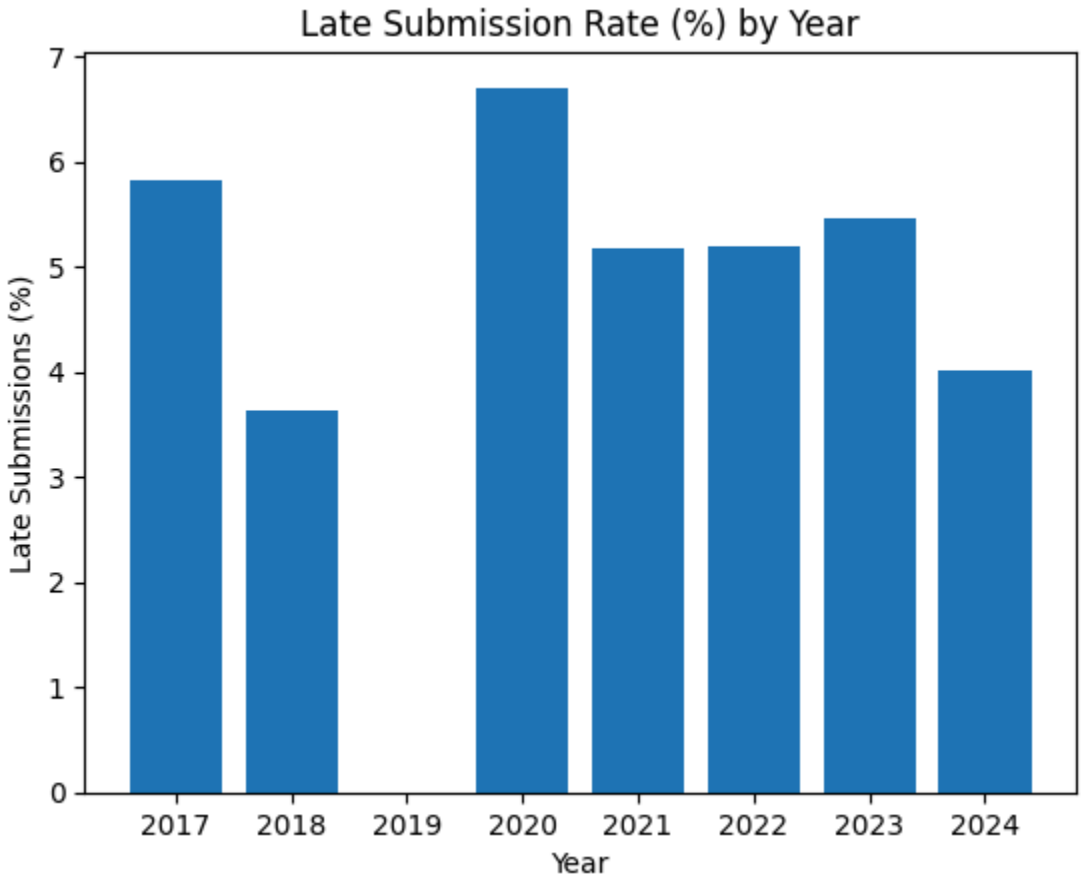
7. Late Submission Rate (by Year)

Late filings are generally low (≈4–6%), with a peak in 2020 (~6.7%) and moderate levels in 2017–2018 and 2021–2023. 2024 is lower (~4.0%), likely reflecting improved timeliness or partial-year effects.
2019 shows **0%**, which is likely an artifact (e.g., missing/uncoded booleans) given the 2019 coverage dip, we need to treat that with caution.

Implications

- A small but non-trivial share of employers file late, which can shift availability of some metrics year-to-year.

- Analyses sensitive to reporting windows (e.g., trend cut-offs) should either exclude late submissions or report sensitivity with/without them.



8. Employer Size (Latest Year - 2024)

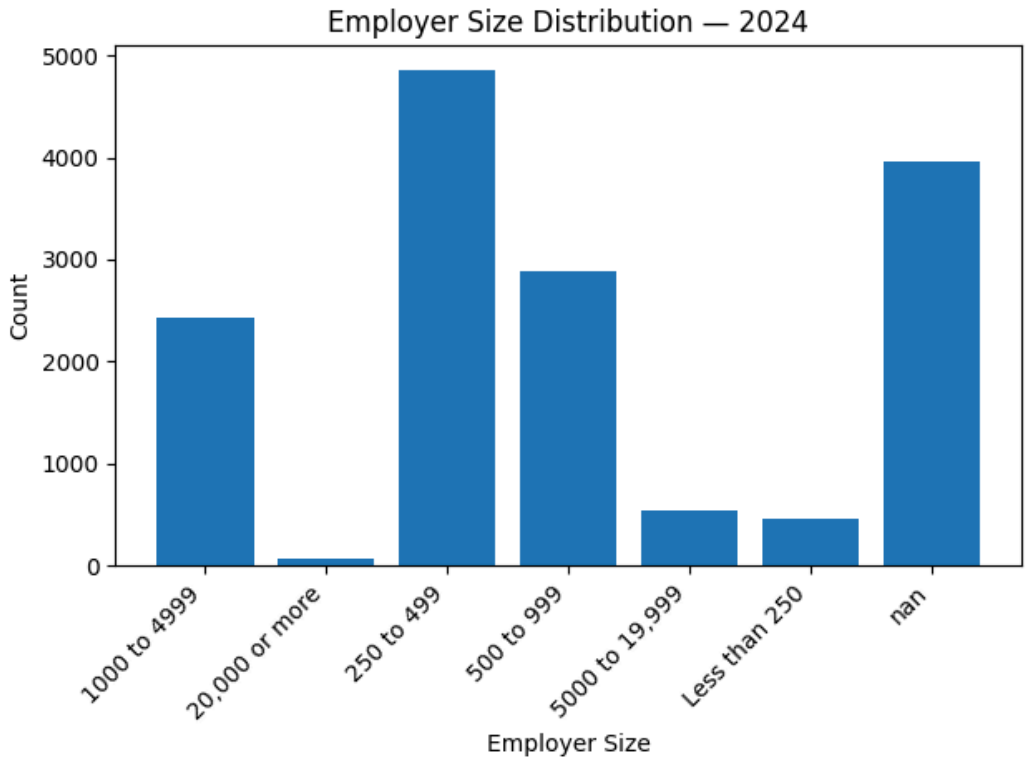
- Reporting is concentrated in mid-sized firms, led by *“250 to 499” and then “500 to 999.”*

- Large employers (“1000 to 4999”, “5000 to 19,999”, “20,000 or more”) are present but fewer than mid-size categories.

- A sizable “Unknown/NaN” group indicates many submissions omit size.

    - **Implication:** size-stratified comparisons may be **biased** if missingness is systematic (e.g., smaller firms less likely to report size).

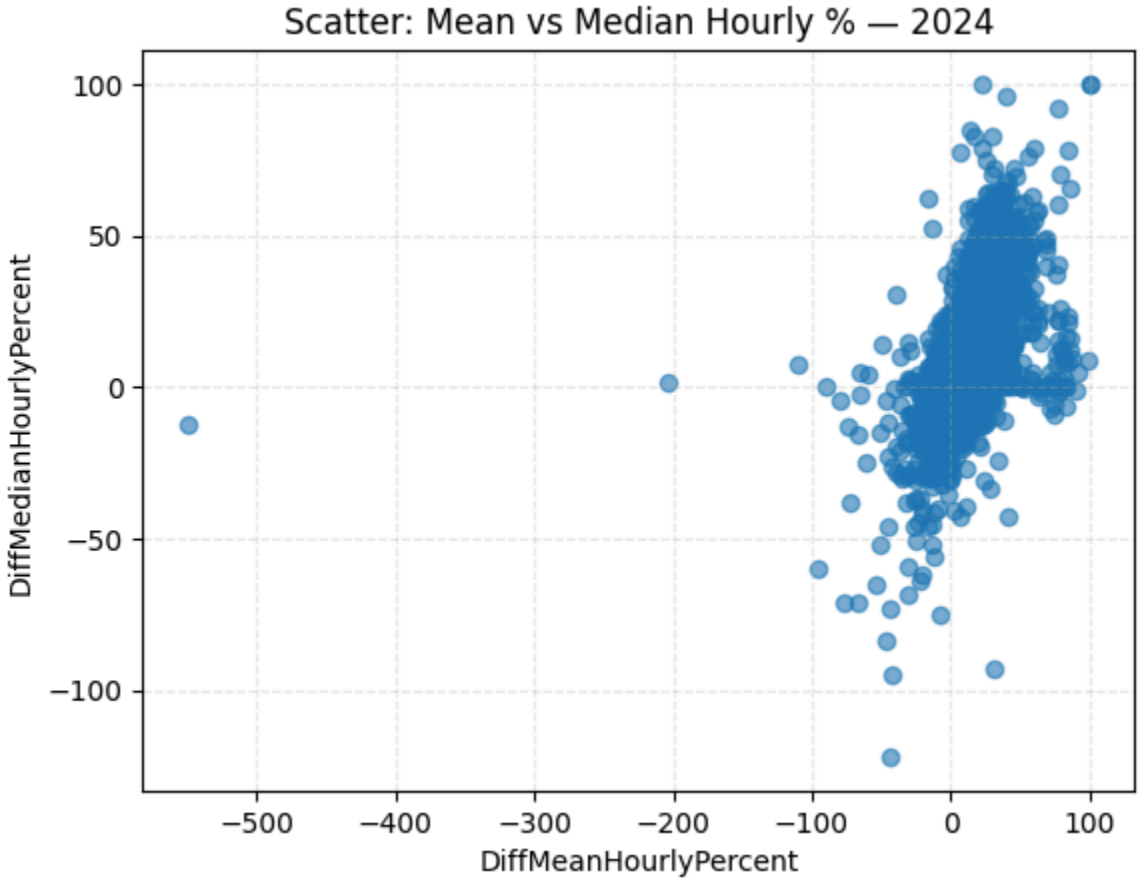
Takeaway: Mid-size employers dominate the 2024 reporting base, but the Unknown share must be handled explicitly to avoid skew in size-based analyses.



9. Mean vs Median Hourly Gap - Latest Year

- The scatter shows a strong positive relationship between *DiffMeanHourlyPercent* and *DiffMedianHourlyPercent*: employers with higher mean gaps also report higher median gaps.
- Most observations cluster in the **0–60%** mean range with broadly similar medians, indicating *consistent direction and magnitude* across both measures.
- A handful of *extreme negatives* (e.g., ≤ –100%) are implausible for percentage gaps and likely reflect data entry/parsing errors.

Implication: conclusions are unchanged whether you use mean or median, but reporting the *median (and/or trimmed mean)* is advisable because it's more robust to outliers. Also, we can Include a sensitivity view that clips to [-100%, 100%].



10. Female Top Quartile (%), 2017–2024

The average share of women in the top pay quartile rose steadily from 39.20% (2017) to 41.75% (2024), a +2.55 pp increase (≈ +6.5% relative). The rise accelerates after 2021, reaching the series high in 2024.

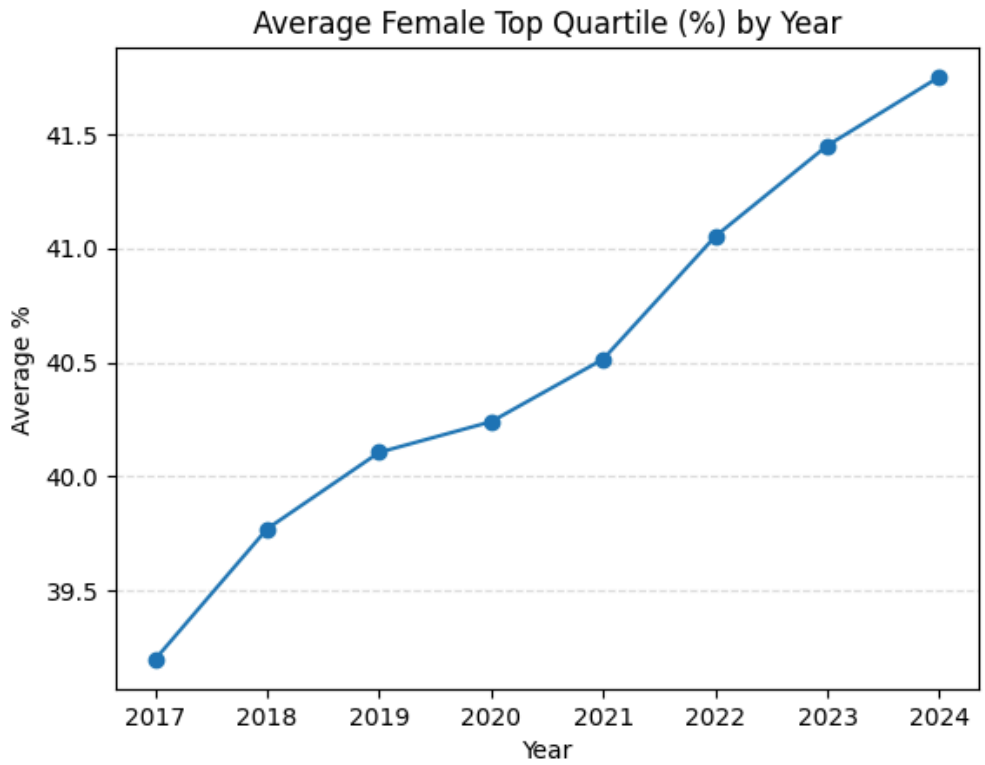
Interpretation

- Women remain *under-represented* at the top: 2024 is still 8.25 pp below parity (50%).

- The upward trend is directionally consistent with the observed narrowing in hourly pay gaps over the same period, suggesting improved representation at senior levels is one contributor.

Caveats

- These are means across employers; improvements may be uneven. We can try to include median and IQR to show distribution, and segment by size/industry.
- The 2019 coverage dip may slightly bias that year's average; the multi-year upward trend is nonetheless persistent.

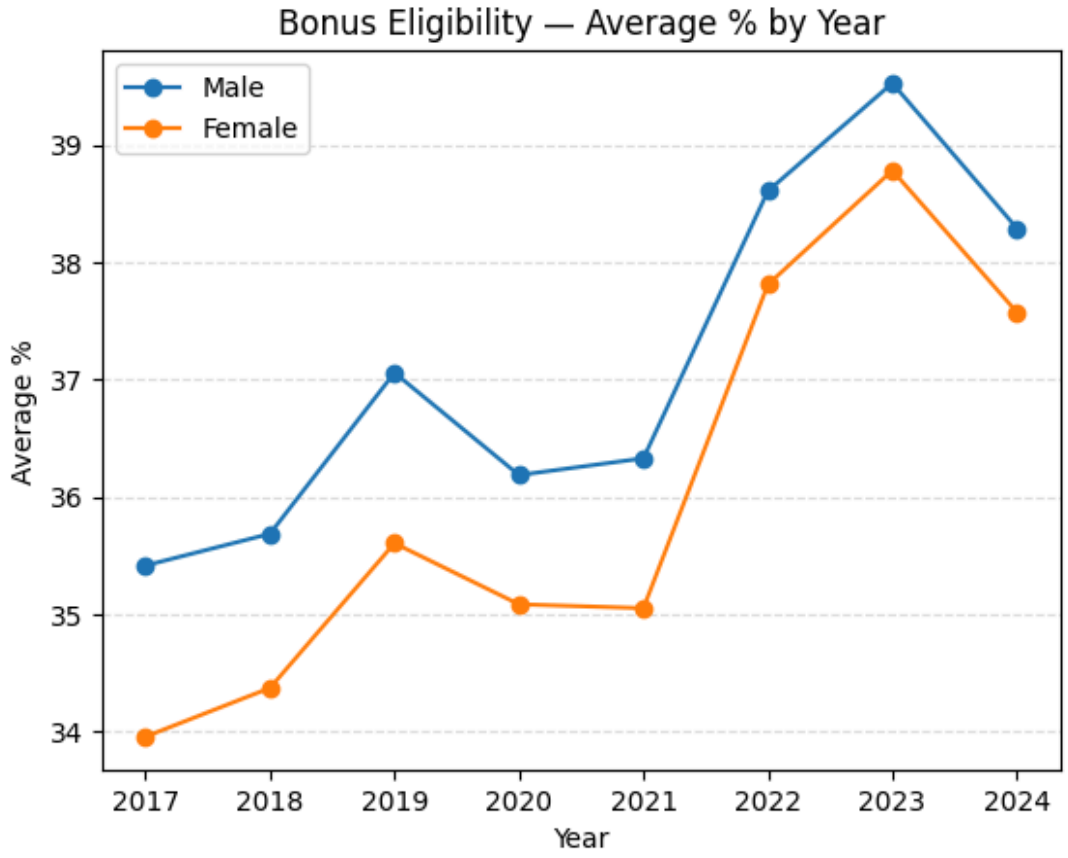


11. Bonus Eligibility (Average %) by Year

- Average eligibility rates are in the mid-30s to ~40% for both genders (most employees are *not* bonus-eligible).
- Men are consistently more likely to be bonus-eligible, but the eligibility gap narrows over time:
    - Early years: ~1.3–1.5 pp (e.g., 2018–2019).
    - Recent years: ~0.7–0.8 pp (2022–2024).
- A step up in eligibility occurs around 2022–2023 for both genders, peaking near 39–40% before easing in 2024.

Interpretation

- The persistent (though shrinking) eligibility gap helps explain why bonus pay gaps exceed hourly gaps: if men are more frequently eligible, and (separately) receive larger bonuses when eligible, the aggregate *DiffMeanBonusPercent* remains high even as hourly gaps narrow.
- The recent convergence in eligibility may be contributing to the post-2019 decline in the mean bonus gap.



12. Correlation Structure (2024)

Key patterns

- **Hourly gaps co-move:** *DiffMeanHourlyPercent* and *DiffMedianHourlyPercent* are *strongly positively correlated*, confirming both metrics tell a consistent story about pay disparity.

- **Quartiles ↔ gaps:**

    - Higher male share in upper quartiles (e.g., *MaleTopQuartile*, *MaleUpperMiddleQuartile*) is ***positively*** correlated with hourly/bonus gaps.

    - Higher female share in upper quartiles (e.g., *FemaleTopQuartile*) is ***negatively*** correlated with the gaps.

    - Within-year, the top-quartile female share tends to align with smaller hourly and bonus gaps.

- **Bonus vs hourly:** *DiffMeanBonusPercent* is ***moderately*** correlated with hourly gap metrics, same direction but noisier (bonuses are more volatile and eligibility-dependent).

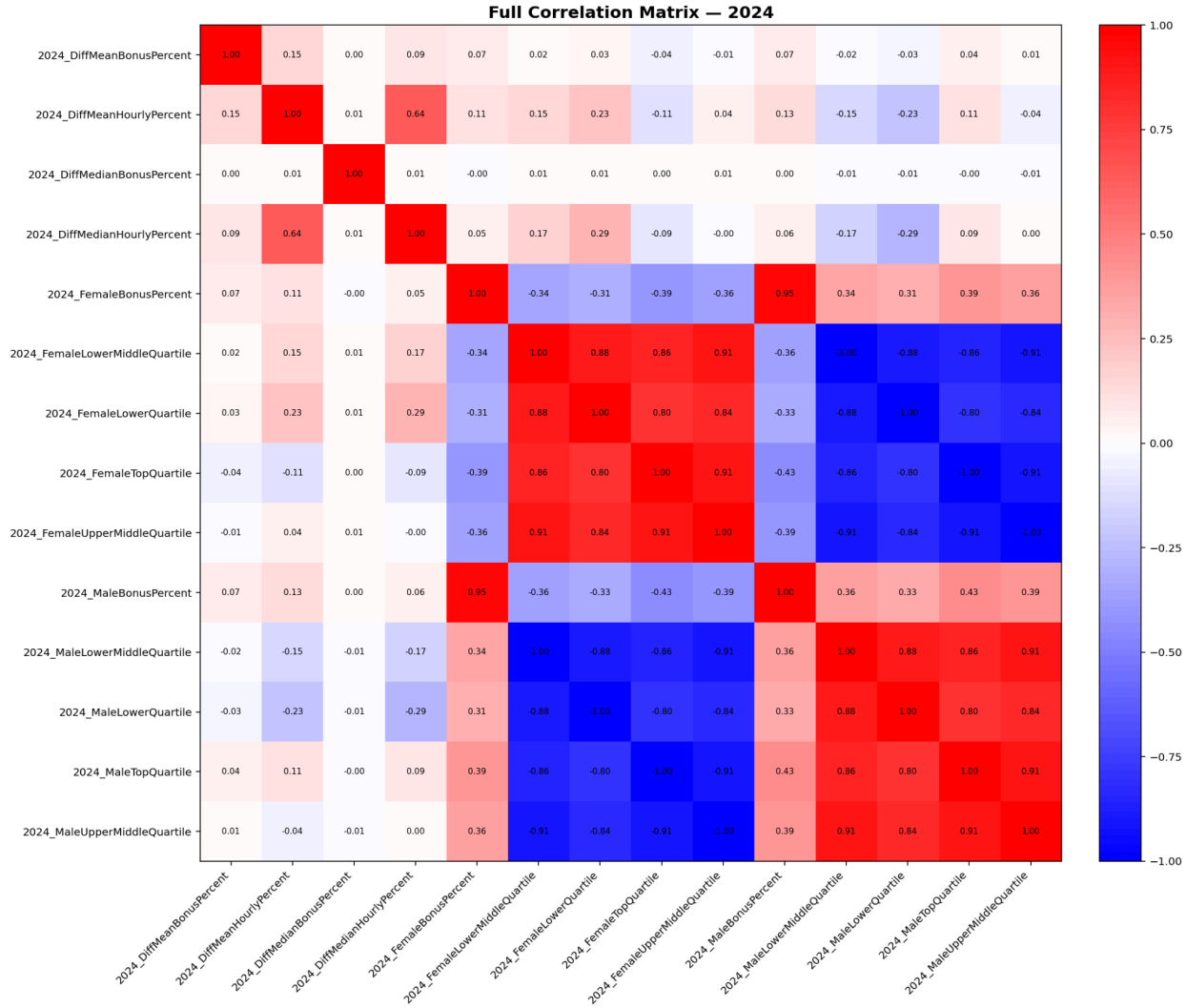
- **Eligibility fields:** Male–female bonus eligibility percentages correlate with the bonus gap in the expected directions (higher male eligibility → larger bonus gap), but magnitudes are below the quartile–gap relationships.

- **Administrative/meta fields** (e.g., *SubmittedAfterTheDeadline*, ID-like numeric columns) show near-zero correlation with pay gap metrics, as expected.

Implications

- Pay gaps are *structurally linked* to representation at the top of pay distributions.

- When modeling gaps, we can expect *multicollinearity* among quartile variables and between mean/median gap measures, we must/will prefer dimensionality reduction or pick one representative feature from each concept group.


Full Correlation Matrix — 2024
2024_DiffMeanBonusPercent
2024_DiffMeanHourlyPercent
2024_DiffMedianBonusPercent
2024_DiffMedianHourlyPercent
2024_FemaleBonusPercent
2024_FemaleLowerMiddleQuartile
2024_FemaleLowerQuartile
2024_FemaleTopQuartile
2024_FemaleUpperMiddleQuartile
2024_MaleBonusPercent
2024_MaleLowerMiddleQuartile
2024_MaleLowerQuartile
2024_MaleTopQuartile
2024_MaleUpperMiddleQuartile
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Use of AI:

Used Gen AI to debug the code and draw in depth insights from the EDA.

References

Gender Pay Gap Service. (n.d.). *Download gender pay gap data*. GOV.UK. Retrieved [Date you accessed it], from https://gender-pay-gap.service.gov.uk/viewing/download