Differentially private boxplots

Jairo Diaz-Rodriguez
York University, Canada

joint work with
Kelly Ramsay

International Conference of Machine Learning.

June 14, 2025

Differential Privacy (DP)

Average public salary

Day 1: $8666

Differential Privacy (DP)

Average public salary

Day 1: $8666

Day 2: $8666

Differential Privacy (DP)

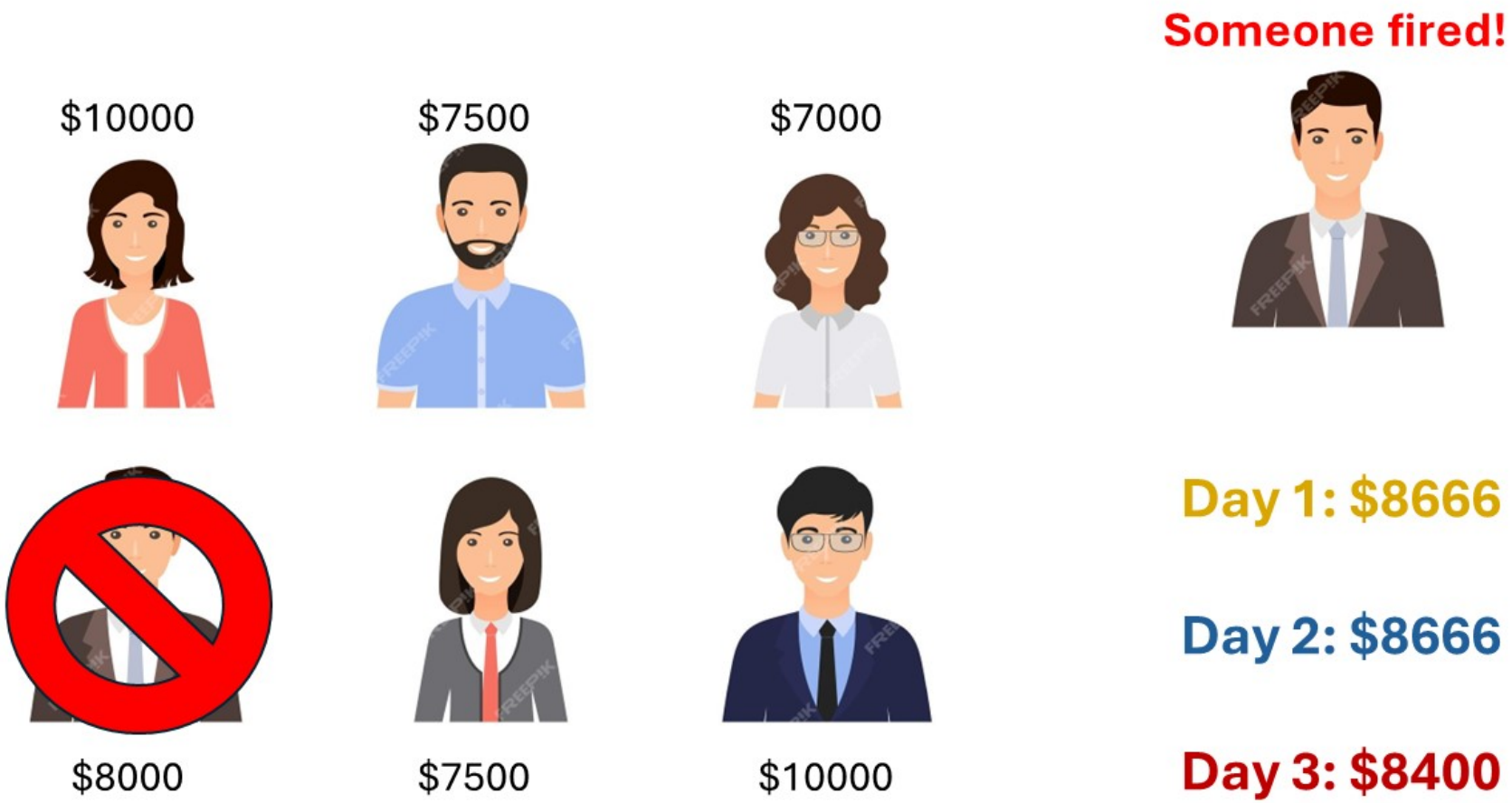
Average public salary

Day 1: $8666

Day 2: $8666

Day 3: $8400

Differential Privacy (DP)



Differential Privacy (DP)

DP Average public salary

Day 1: $8666 + *noise*

Day 2: $8666 + noise

Day 3: $8400 + *noise*

Differential Privacy (DP)

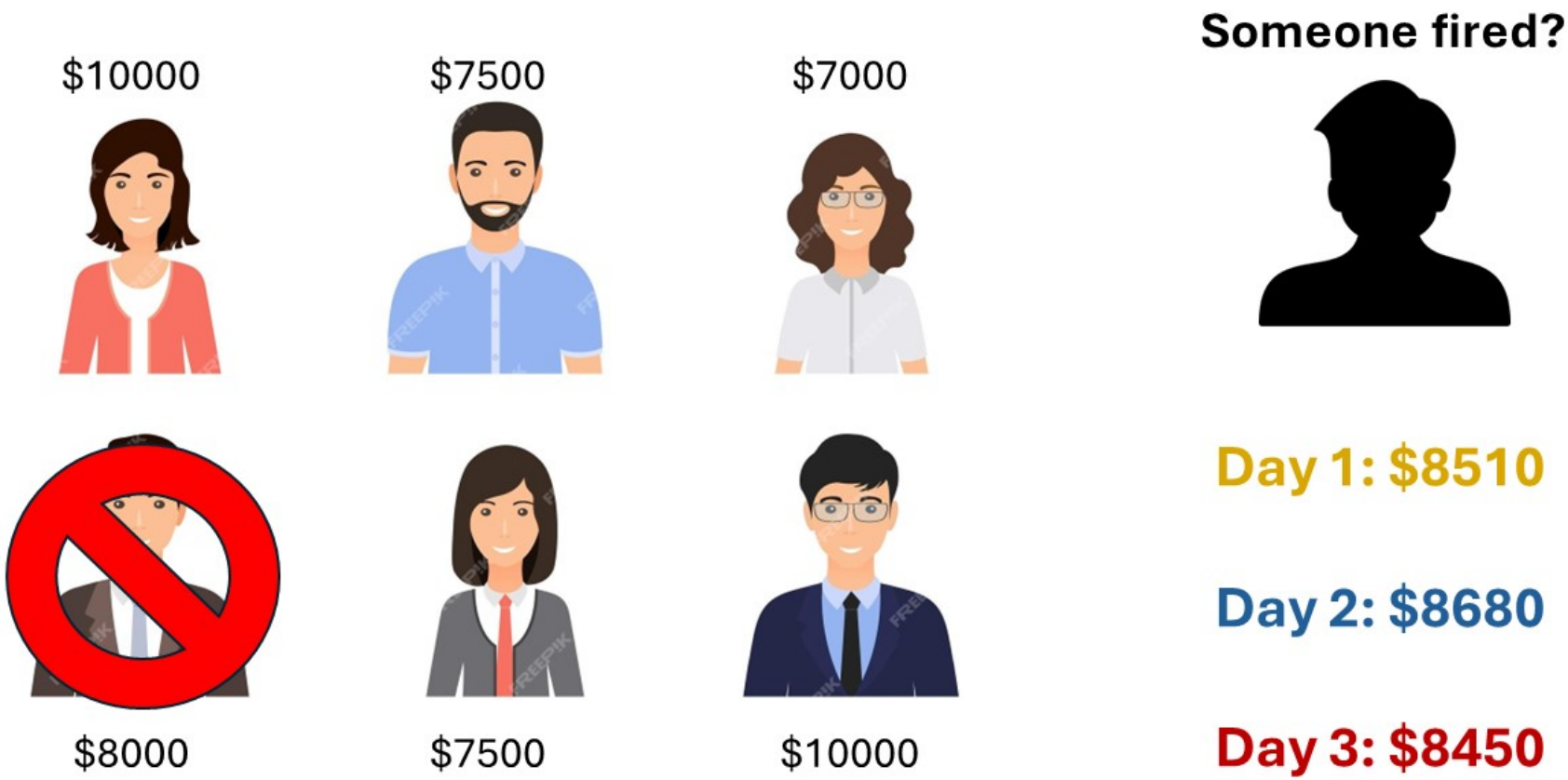
DP Average public salary

Day 1: $8510

Day 2: $8680

Day 3: $8450

Differential Privacy (DP)



Differential Privacy (DP)

A randomized function $\mathcal{A}$, which operates on datasets, is $\epsilon-$differentially private if for any two datasets D and D' that differ in a single observation (these are often called adjacent datasets), and for all sets S of possible outputs of $\mathcal{A}$:

$$P(\mathcal{A}(D) \in S) \leq e^{\epsilon} \times P(\mathcal{A}(D') \in S)$$

Differential Privacy (DP)

Low privacy budget (small ε)	High privacy budget (big ε)
Day 1: $8590	**Day 1: $8650**
Day 2: $8680	**Day 2: $8675**
Day 3: $8550	**Day 3: $8420**

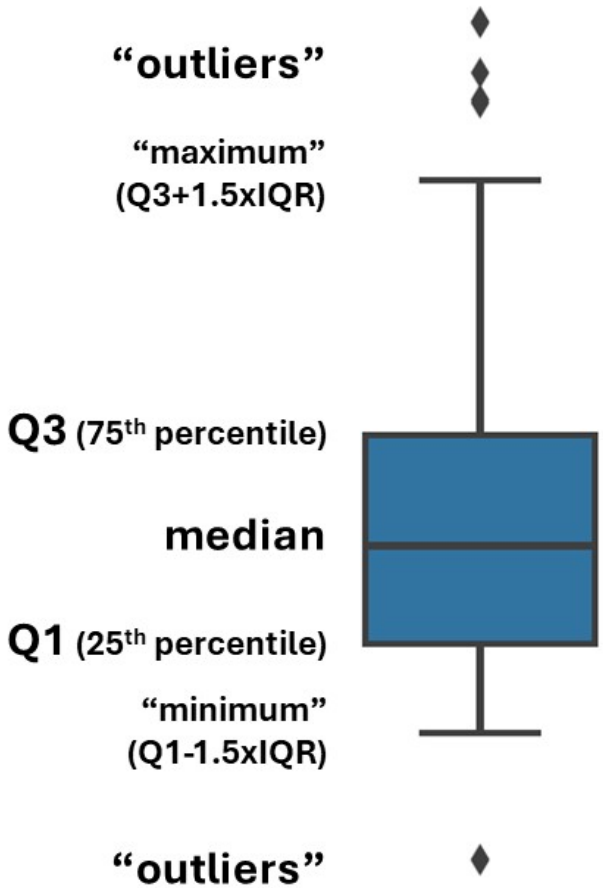
Differential Privacy (DP)

Low privacy budget (small ε)	High privacy budget (big ε)
Day 1: $8590	**Day 1: $8650**
Day 2: $8680	**Day 2: $8675**
Day 3: $8550	**Day 3: $8420**

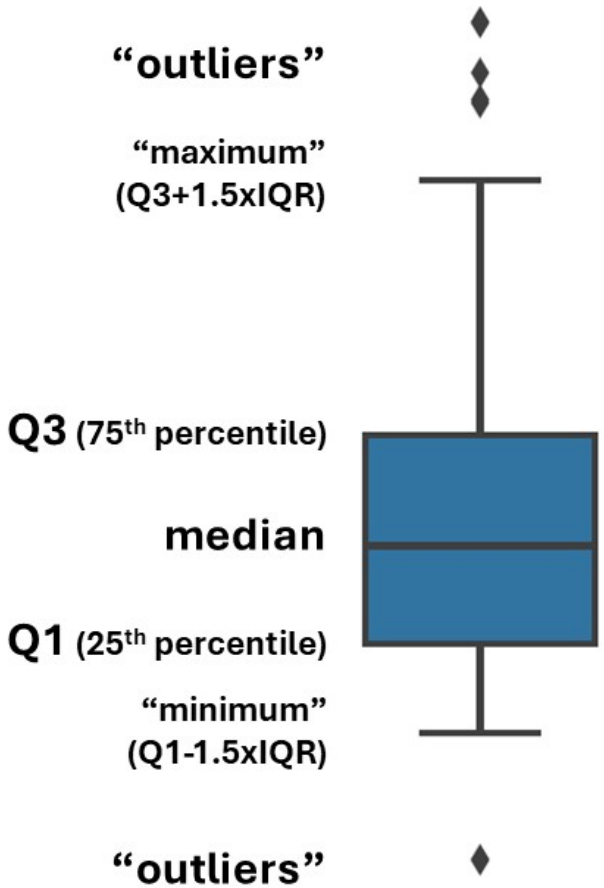
Challenge

Draw useful insights **from datasets while** ***protecting individual privacy***

Boxplots


“outliers”
“maximum”
(Q3+1.5xIQR)
Q3 (75th percentile)
median
Q1 (25th percentile)
“minimum”
(Q1-1.5xIQR)
“outliers”

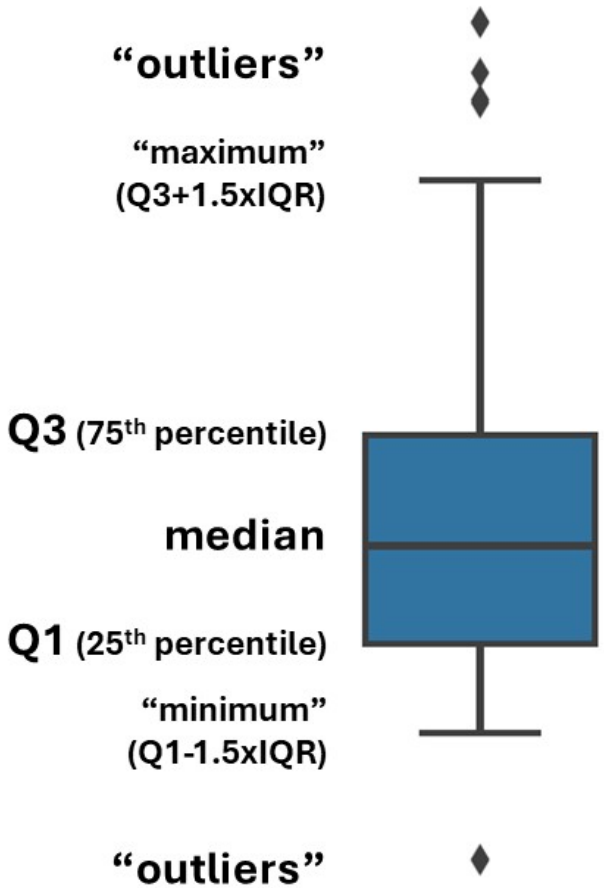
Boxplots



Data analysis with boxplots

- Location
- Scale
- Skewness
- Tails

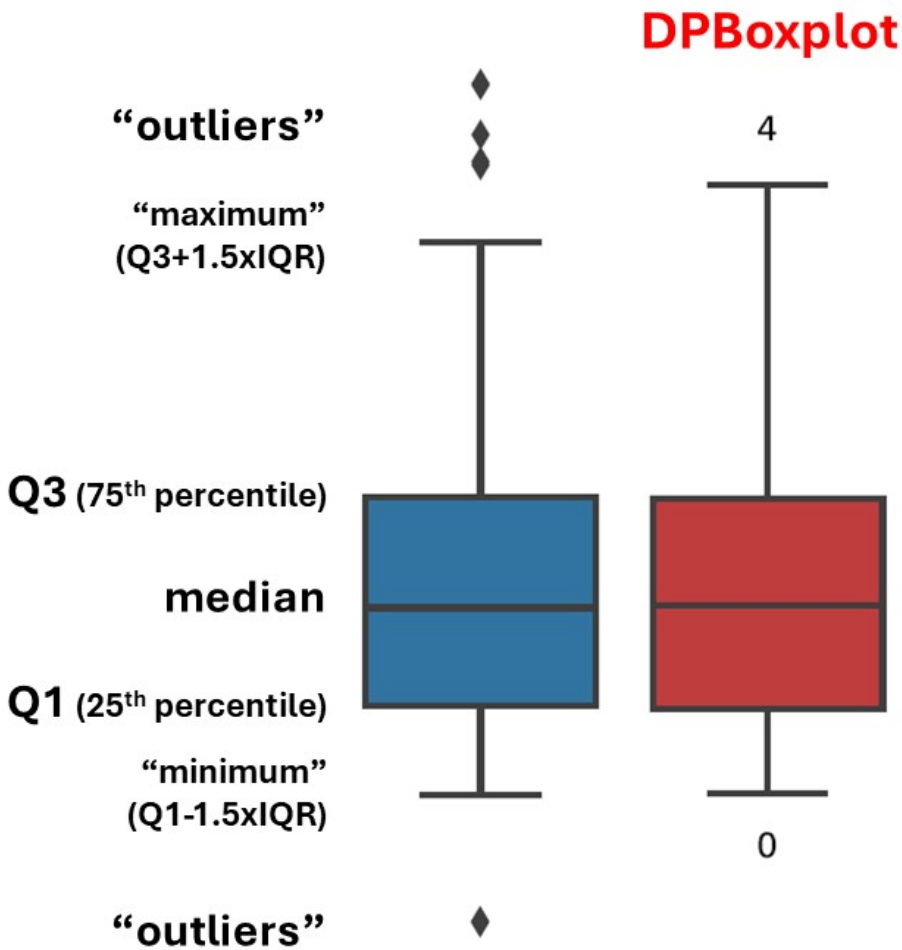
Boxplots



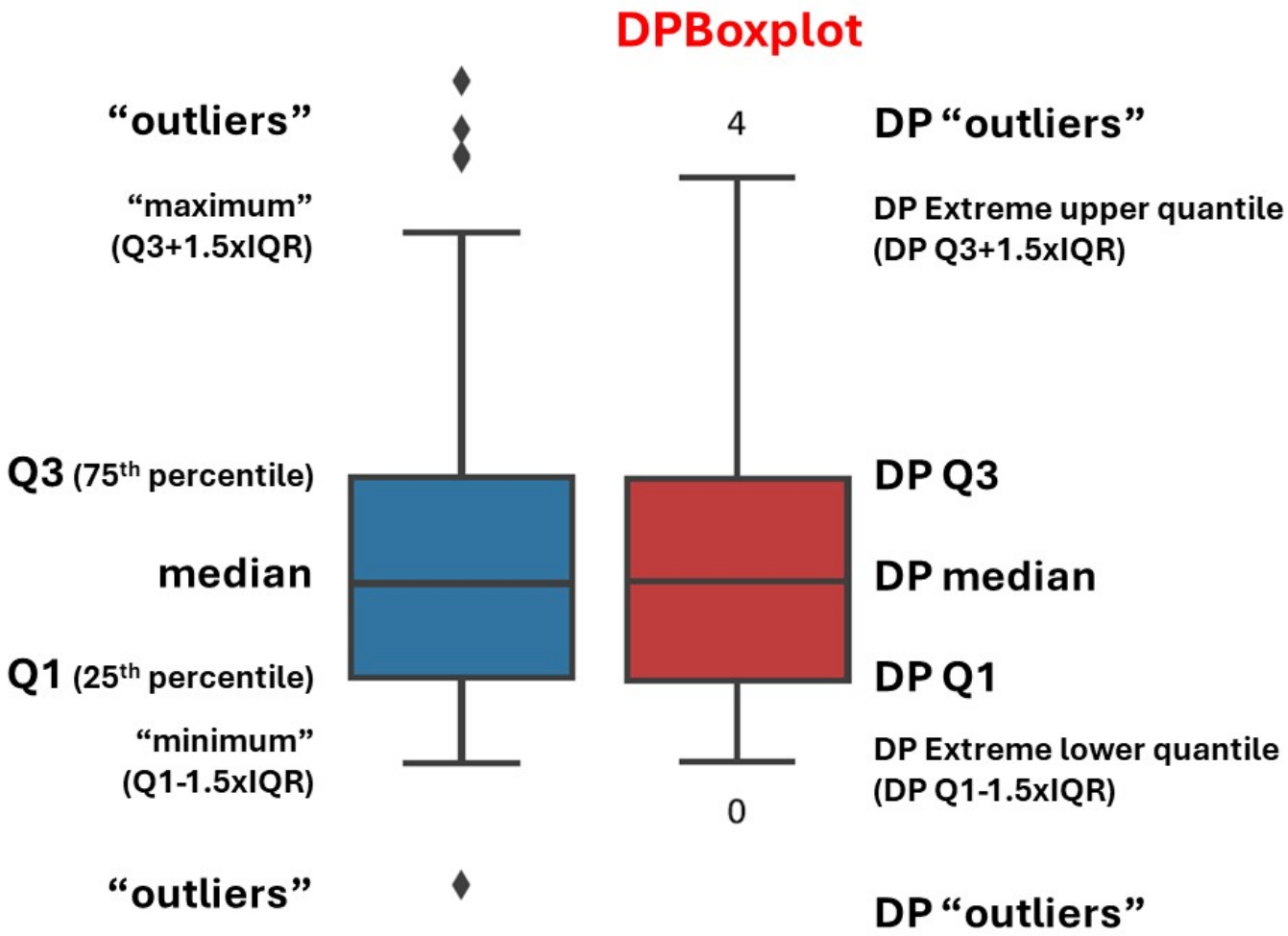
Challenge for differentially private boxplots

Draw useful insights **from datasets while** ***protecting individual privacy***

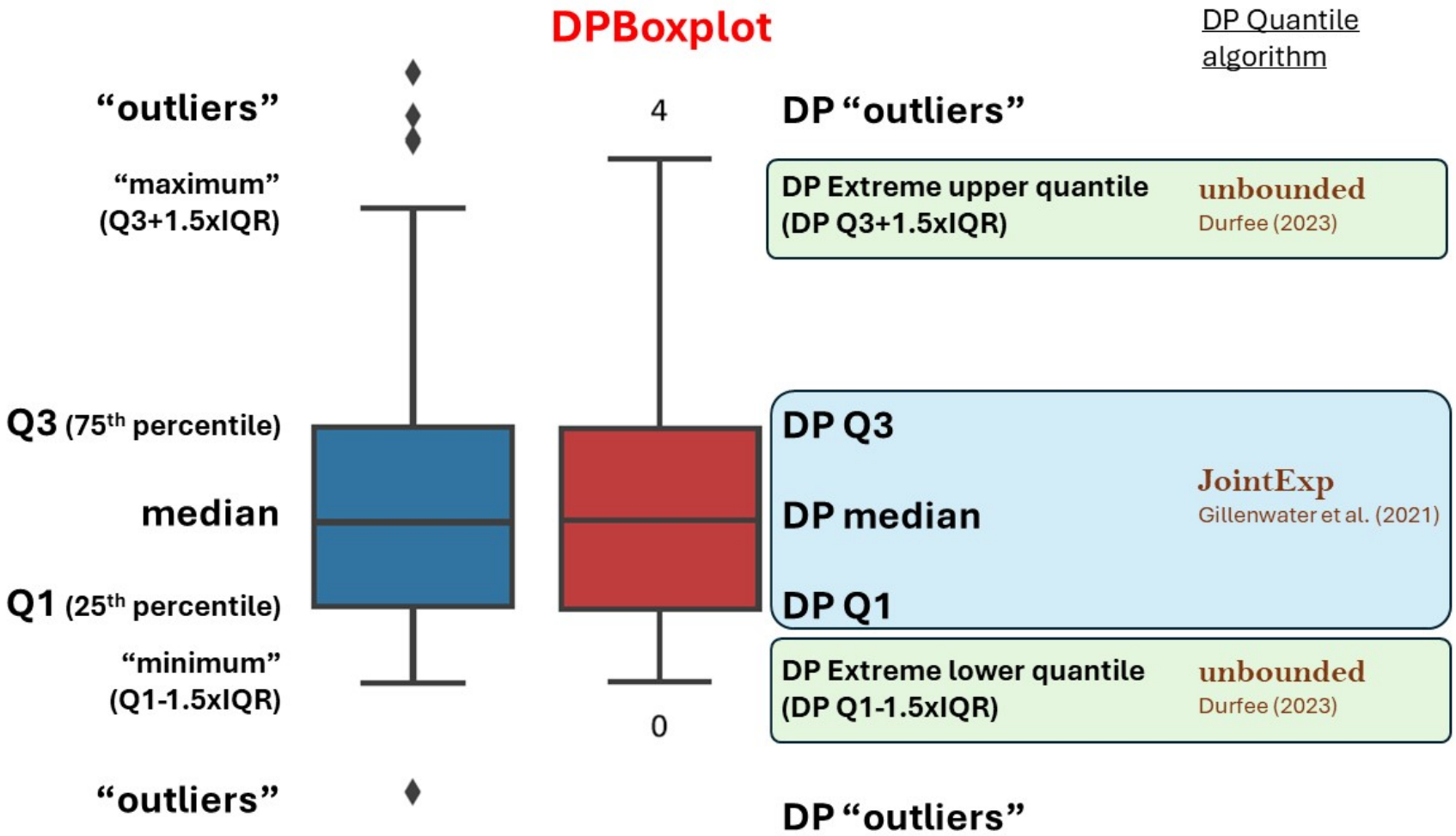
Differentialy private boxplots



Differentialy private boxplots



Differentialy private boxplots



On DP Quantile estimation

Box

On DP Quantile estimation

Box

- We used `JointExp` quantile estimation (Gillenwater et al., 2021).

On DP Quantile estimation

Box

- We used `JointExp` quantile estimation (Gillenwater et al., 2021).
- **Main theoretical contributions:**
  - We proved a minimax lower bound for privately estimating a quantile which matches the upper bound.

On DP Quantile estimation

Box

- We used `JointExp` quantile estimation (Gillenwater et al., 2021).
- **Main theoretical contributions:**
  - We proved a minimax lower bound for privately estimating a quantile which matches the upper bound.
  - This implies that the **scale** and **location** of the proposed private boxplot are estimated optimally, up to logarithmic factors.

On DP Quantile estimation

Whiskers and outliers

On DP Quantile estimation

Whiskers and outliers

- We used `unbounded` quantile estimation (Durfee, 2023).

On DP Quantile estimation

Whiskers and outliers

- We used `unbounded` quantile estimation (Durfee, 2023).
- **Main theoretical contributions:**
  - `JointExp` is inconsistent for extreme quantiles.

On DP Quantile estimation

Whiskers and outliers

- We used `unbounded` quantile estimation (Durfee, 2023).
- **Main theoretical contributions:**
  - `JointExp` is inconsistent for extreme quantiles.
  - We proved that whiskers and number of outliers are weakly consistent for their population counterparts when using `unbounded`.

On DP Quantile estimation

Whiskers and outliers

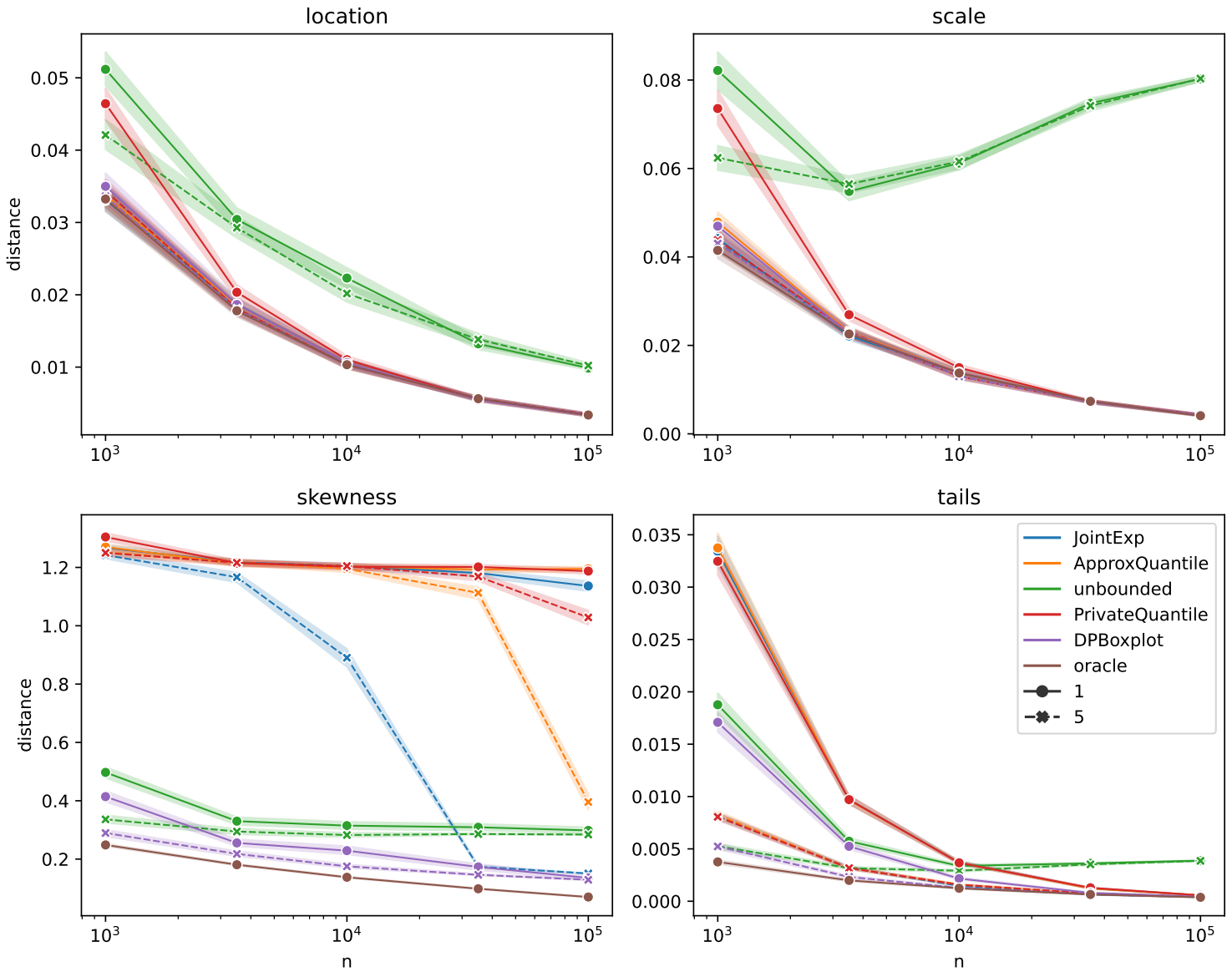
- We used `unbounded` quantile estimation (Durfee, 2023).
- **Main theoretical contributions:**
  - `JointExp` is inconsistent for extreme quantiles.
  - We proved that whiskers and number of outliers are weakly consistent for their population counterparts when using `unbounded`.
  - This implies that **skewness** and **tails** will be correctly portrayed by our proposed differentially private boxplot.

Simulation studies

We performed an exhaustive simulation study under multiple settings:

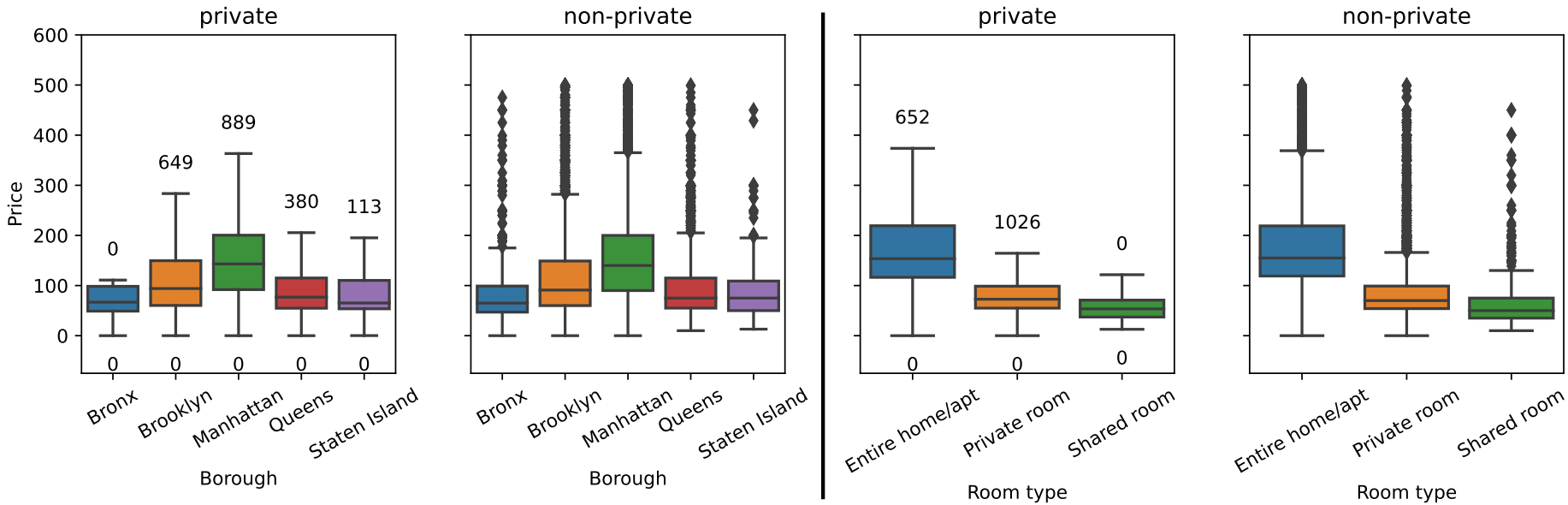
- Generating distributions: Normal, skew-normal, uniform, beta, real data.
- ϵ: 0.5, 1, 5, 10
- sample sizes: 1000, 2000, 10000, 20000, 10000
- We quantified similitude in **location**, **scale**, **skewness**, and **tails** between constructed private boxplot and the population counterpart.
- We compared with naive private boxplots constructed by using different private quantile estimation methods.

Simulation studies (summary)



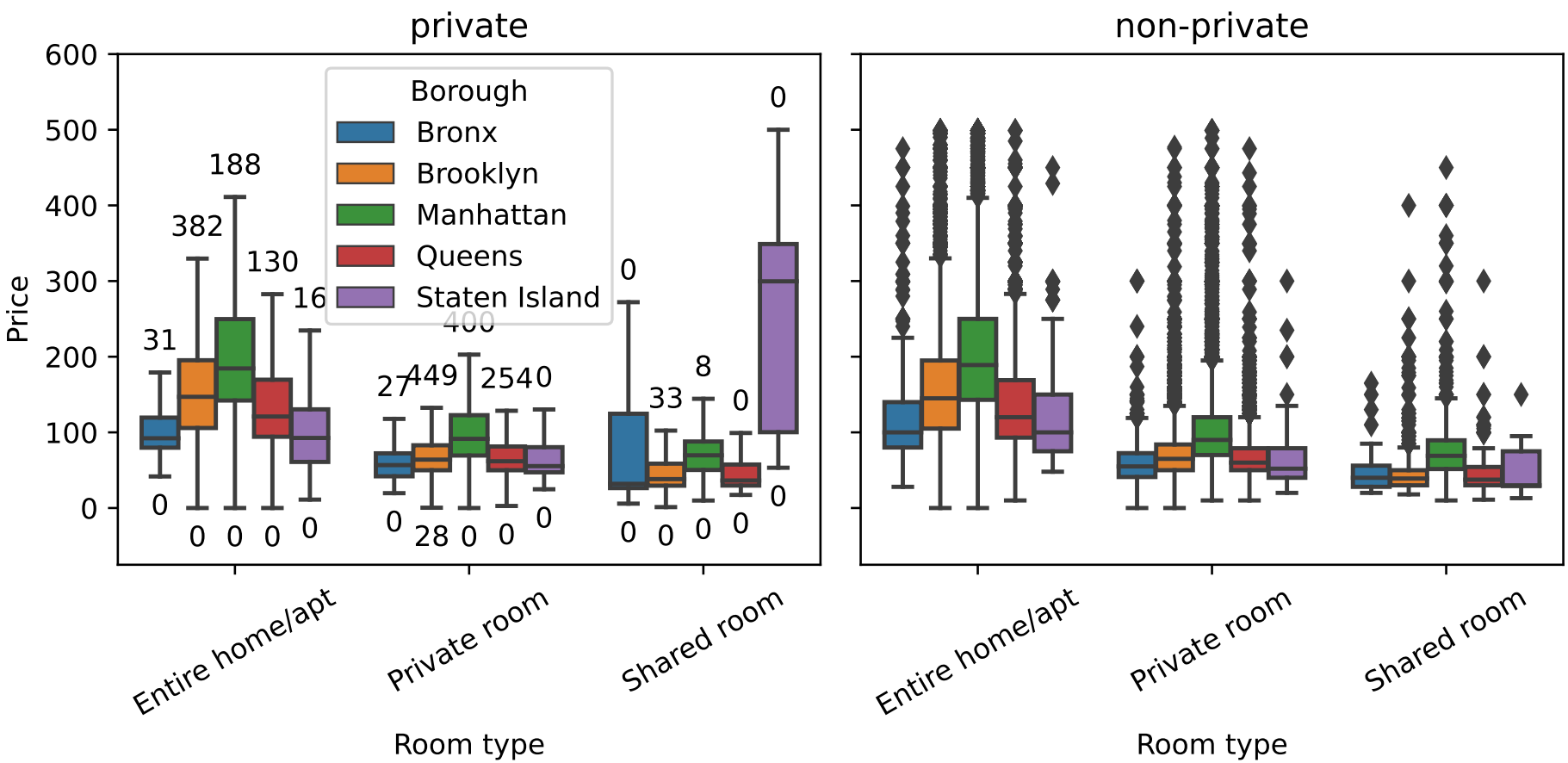
Real data

Do discernible patterns emerge in Airbnb listing prices across various boroughs in New York City and differing room types?



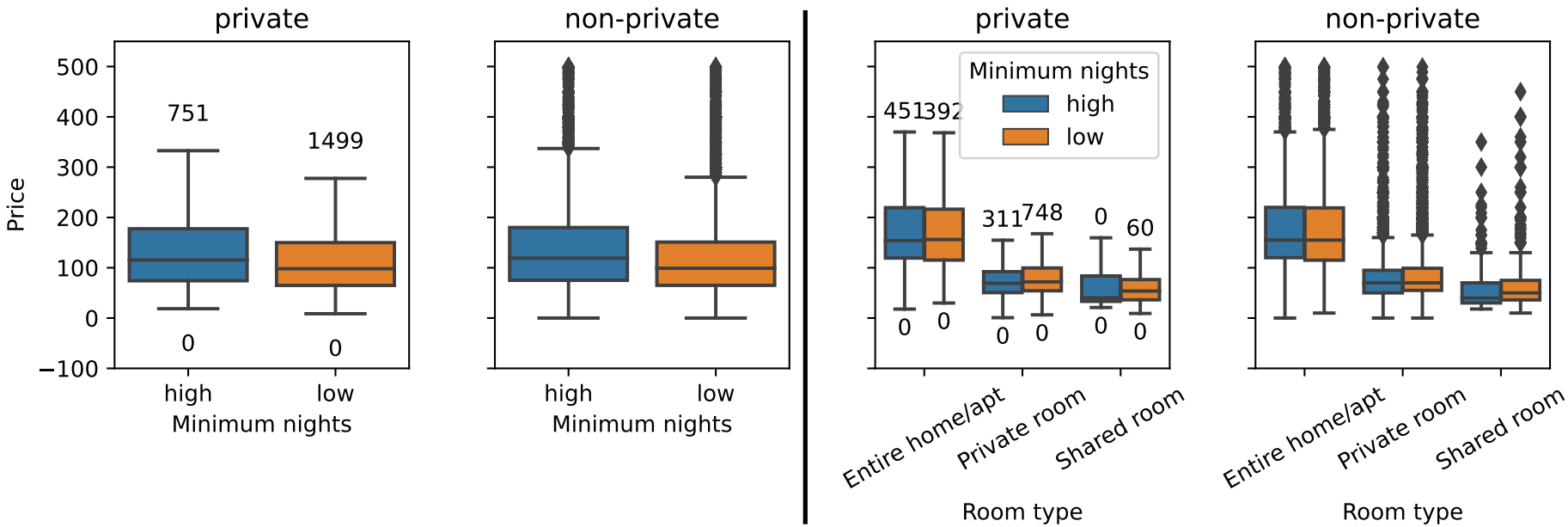
Real data

Do discernible patterns emerge in Airbnb listing prices across various boroughs in New York City and differing room types?



Real data

Are there observable trends in Airbnb listing prices concerning minimum nights required for reservation and the types of rooms offered?



Real data

Conclusions

- We observe relatively consistent patterns between the private and non-private boxplots.

Real data

Conclusions

- We observe relatively consistent patterns between the private and non-private boxplots.
- The primary visual disparities pertains to the positioning of whiskers on the boxplots and number of outliers. This underscores the recognized challenge associated with differentially private estimation of extreme quantiles and small sample sizes.

Real data

Conclusions

- We observe relatively consistent patterns between the private and non-private boxplots.
- The primary visual disparities pertains to the positioning of whiskers on the boxplots and number of outliers. This underscores the recognized challenge associated with differentially private estimation of extreme quantiles and small sample sizes.
- Discrepancies does not materially impede the analytical value of the visual findings.

Real data

Conclusions

- We observe relatively consistent patterns between the private and non-private boxplots.
- The primary visual disparities pertains to the positioning of whiskers on the boxplots and number of outliers. This underscores the recognized challenge associated with differentially private estimation of extreme quantiles and small sample sizes.
- Discrepancies does not materially impede the analytical value of the visual findings.
- This case study demonstrates the potential of differentially private, exploratory data analysis and confirms the efficacy of the differentially private boxplot, in accordance with our theoretical and simulated results.

Thank you.

Differentially private boxplots.
code: https://github.com/jairoadiazr/DPBoxplot