Viswanth Tammana,

viswantht21@gmail.com

Bias Detection from Yelp Dataset.

The analysis started by joining separate JSON files (reviews.json and business.json) to create a unified dataset, which was then converted to CSV format for efficient handling. Reviews were grouped by city, extracting approximately 12,000 entries per city to facilitate focused local analysis. Extensive text cleaning and preprocessing generated a structured clean_text field to ensure clarity during subsequent analysis. Each review was then evaluated using toxicity scoring and bias similarity metrics to identify potentially problematic language. Additionally, an attempt was made to leverage the OpenAI Moderation API to detect subtle and indirect forms of bias; however, practical limitations, including API rate limits and processing speed constraints, prompted a shift toward alternative approaches. Ambiguous or borderline cases underwent manual review, refining and solidifying labeling criteria for the is_biased designation. Individual city-level analyses were conducted initially, revealing nuanced local patterns, before aggregating all city datasets into one comprehensive file to derive broader insights and enable comparative analysis. Future analyses are planned to incorporate sentiment analysis to further categorize and clarify the nature of identified biases, facilitating a deeper understanding of underlying themes and improving overall detection precision.

Key Insights Combined

- Hotels Are Primary Bias Hotspots:
  Hotels consistently exhibit higher bias rates (typically 2–3× more than restaurants and spas), reflecting the longer duration of guest stays, higher service complexity, and greater emotional stakes associated with lodging.
- Bias is Highly Seasonal and Event-Driven:
  Specific months—such as spring in Boise and Edmonton, October in Philadelphia, and holiday periods in Tampa—show distinct peaks in biased language, indicating bias surges often align with high-traffic tourist seasons or local events.
- Bias Concentrates Geographically:
  Rather than evenly spread, bias incidents cluster intensely within certain neighborhoods (e.g., a New Orleans neighborhood showing 14× the city's average bias rate). Targeting moderation efforts geographically could significantly improve efficiency and effectiveness.

Key Insights (Across Cities)

- Star Rating Correlation (Philadelphia, New Orleans):
  - Approximately 58% of biased reviews in Philadelphia were rated 1-star, indicating strong dissatisfaction frequently manifests as biased language.
  - New Orleans displayed an unusual spike in bias for 5-star reviews, suggesting local cultural or service dynamics uniquely contribute to positive yet biased feedback.
- Seasonal Bias Patterns (Multiple Cities):

    - Clear seasonal spikes are evident: Philadelphia peaks in October (~1.38%), Boise and Edmonton during the spring (~1.9% and ~2.0%, respectively), and New Orleans in both January and October.
    - Tampa's bias trends rise toward year-end holidays, aligning with increased visitor volumes and service strain. Conversely, Indianapolis exhibits more stable bias levels year-round, implying different local drivers.
- Toxicity vs. Bias Relationship (Philadelphia, New Orleans):
    - In Philadelphia, bias similarity effectively separates biased from unbiased reviews, with distinct distributions at higher similarity scores (0.25–0.30 and beyond). A threshold near 0.30 is optimal for balancing precision and recall.
    - New Orleans showed toxicity as an extremely strong proxy for bias, as nearly every highly toxic review was flagged as biased, reinforcing the value of toxicity as a screening criterion in specific contexts.
- Engagement Metrics (Philadelphia):
    - Metrics such as "useful," "funny," and "cool" are tightly correlated with each other ($r > 0.9$) but uncorrelated with bias or toxicity, indicating popularity or review engagement doesn't predict biased content.

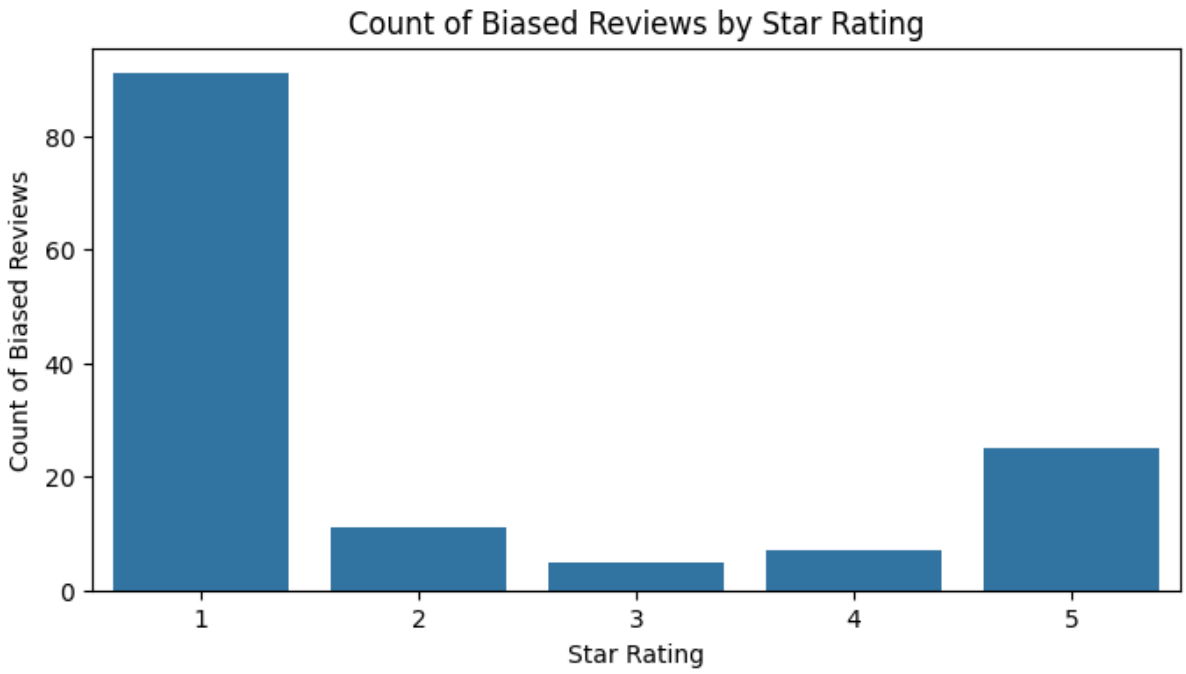
Segment	Typical Monthly Bias Rate (%)	Observations
Hotel	4–6% (up to ~15% during peaks)	High emotional stakes and longer stays amplify bias significantly.
Restaurant	1–2%	Relatively lower bias incidence; shorter interactions and lower emotional involvement.
Spa	1–2% (higher in specific cities)	Occasionally surpass hotels in specific markets (e.g., New Orleans, Philadelphia), pointing to localized service dynamics.

Automation and Ethical Considerations

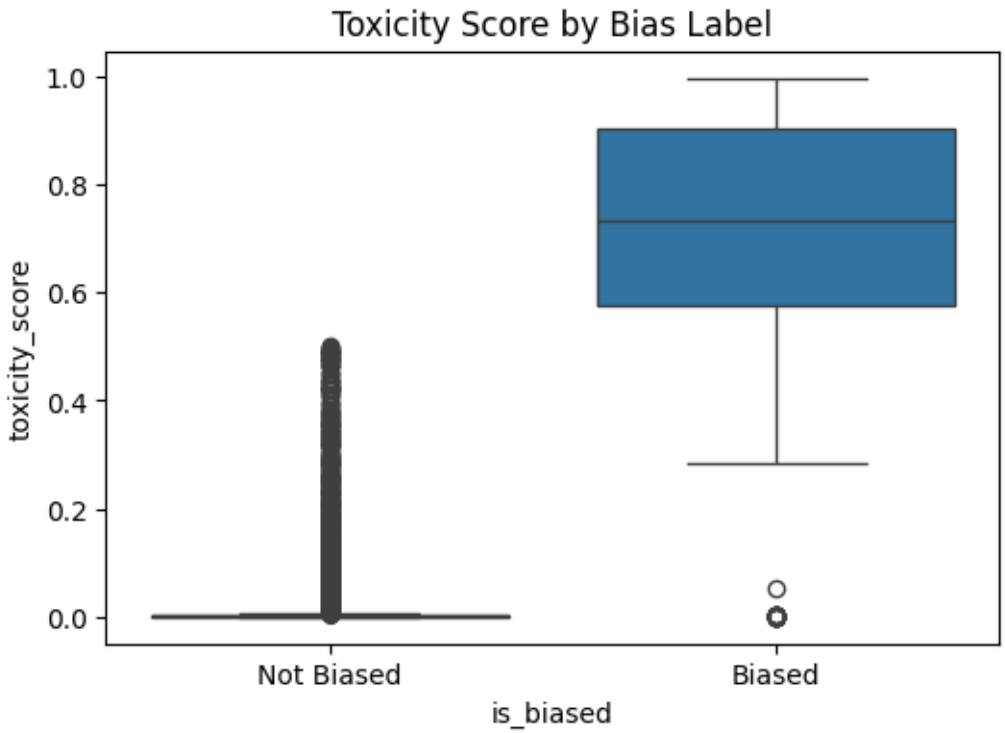
- Real-time Automated Detection:
  Implement a multi-layered real-time detection pipeline leveraging lexicon-based screening, toxicity models, and trained LDA topic classifiers. Human moderators provide oversight to continuously refine detection accuracy.
- Ethical Safeguards:
  Maintain transparency in content moderation, clearly communicating criteria and providing fair appeal processes. Regularly audit the detection algorithms to prevent unintended bias toward particular language styles or user demographics.
- Innovative Extensions:
  Introduce interactive heatmaps or dashboards reflecting bias trends to assist service providers proactively. Publicly available bias insights can empower both consumers and businesses to make informed, fairer choices.

Visual Representation of Analysis and Insights

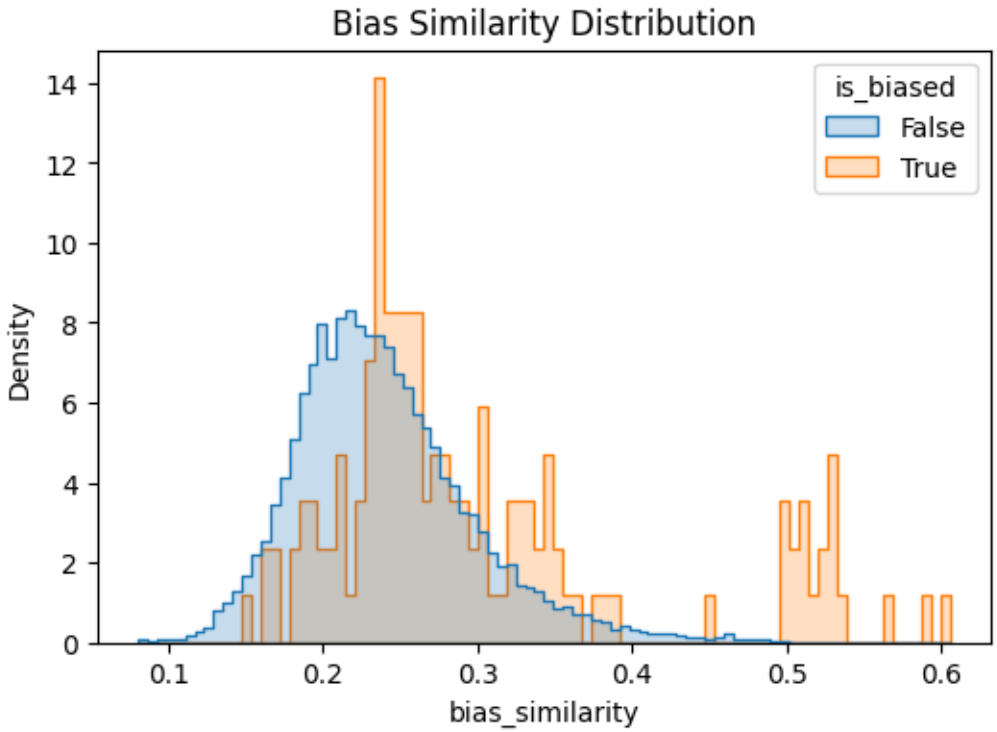
Philadelphia



1. Over half of all biased reviews (≈58%) occur at the 1-star level, showing that extreme dissatisfaction often coincides with biased language.
2. This suggests prioritizing bias-detection and moderation efforts on very low-rated feedback to catch the majority of problematic content.

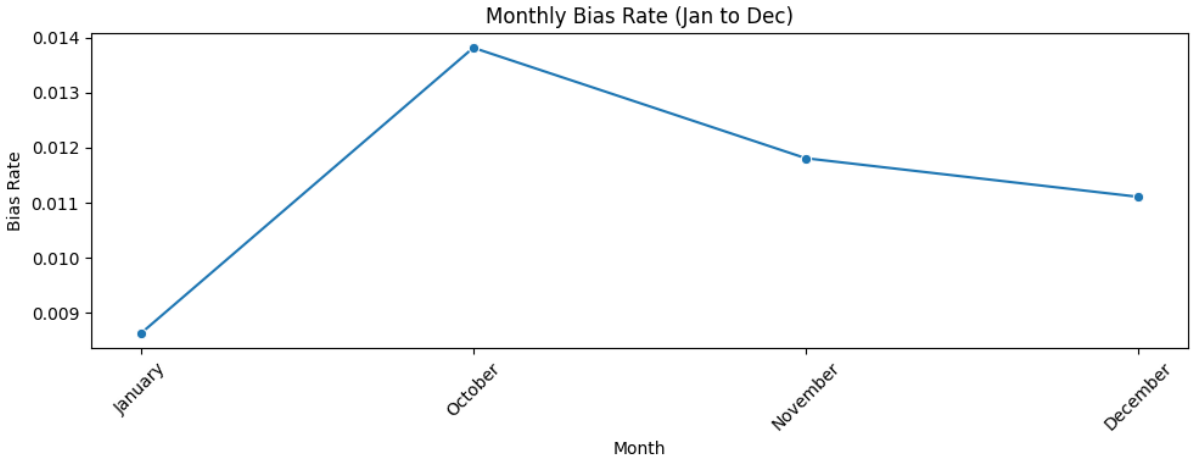


3. Prioritizing high-toxicity reviews for moderation will catch the bulk of biased language efficiently.

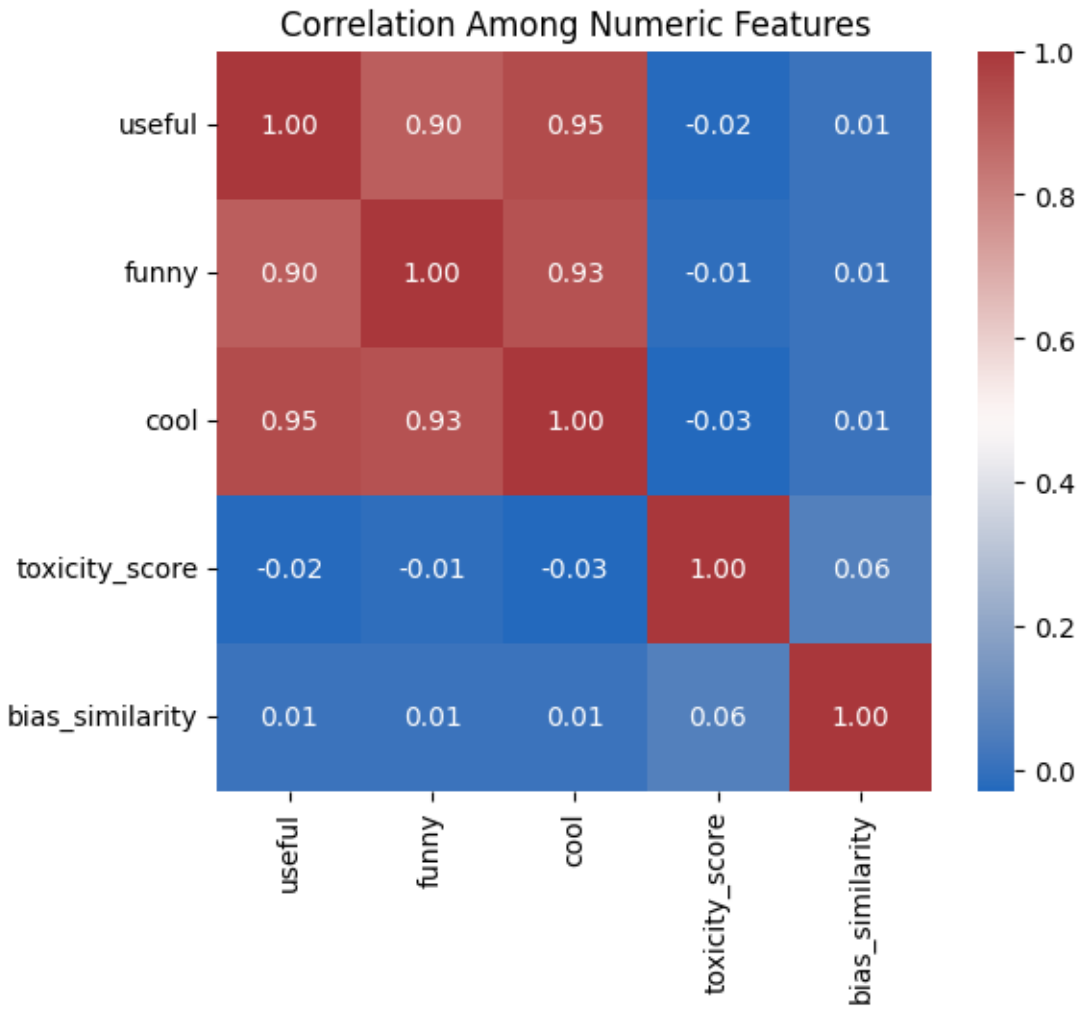


4. Right-shifted distribution for biased reviews
   Biased reviews concentrate at higher bias_similarity values (peak around 0.25–0.30 and a long tail up to 0.6), whereas non-biased reviews peak lower (around 0.20–0.23). This confirms similarity metric is capturing genuine bias signals.

5. Overlap region indicates tuning opportunities
   The overlap between roughly 0.22–0.30 suggests a “gray zone” where thresholds could be optimized or where additional features (e.g., sentiment or toxicity) might help disambiguate borderline cases.
6. Threshold selection for precision
   Placing a cutoff near ~0.30 would capture most biased reviews (high recall) while minimizing false positives, and very high values (>0.45) are virtually guaranteed to be biased—ideal for automated moderation.

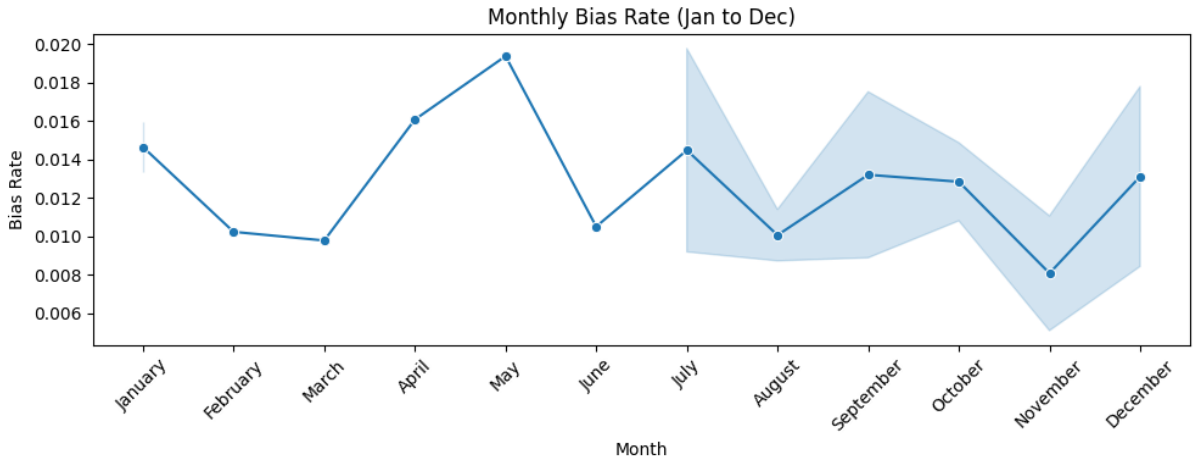


7. The monthly bias rate climbs to its highest in October (~1.38%), then steadily falls through November (1.18%) and December (1.11%) to its lowest in January (0.87%), indicating a seasonal surge in biased language post-summer that subsides over the holiday period into the new year.



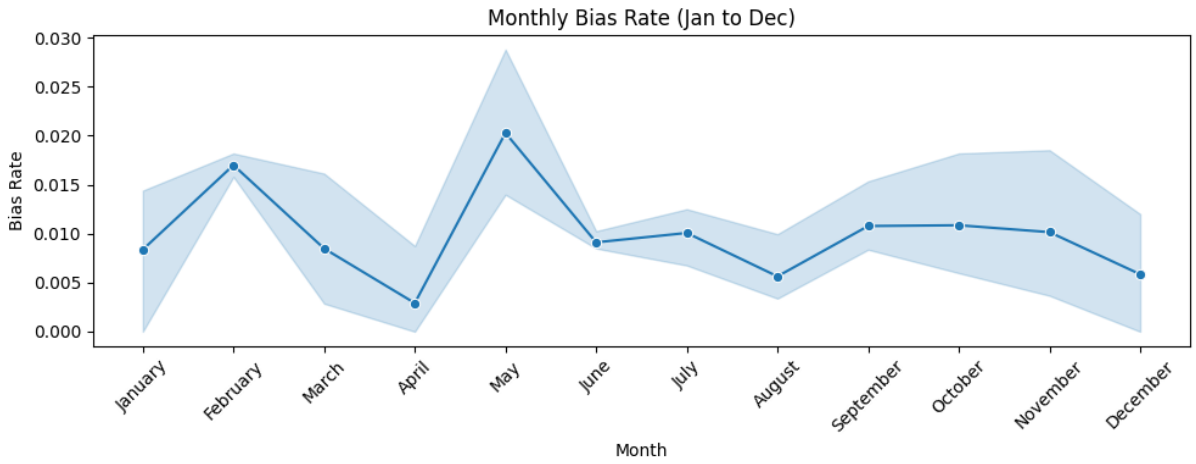
8. The three engagement metrics (“useful,” “funny,” “cool”) form a tight cluster (r>0.9) but are essentially uncorrelated with both toxicity and bias similarity (near zero), while toxicity and bias similarity themselves only weakly correlate (r≈0.06)—showing that popular or highly rated reviews aren’t any more or less biased, and the bias detector is capturing nuances beyond mere toxic language.

Boise



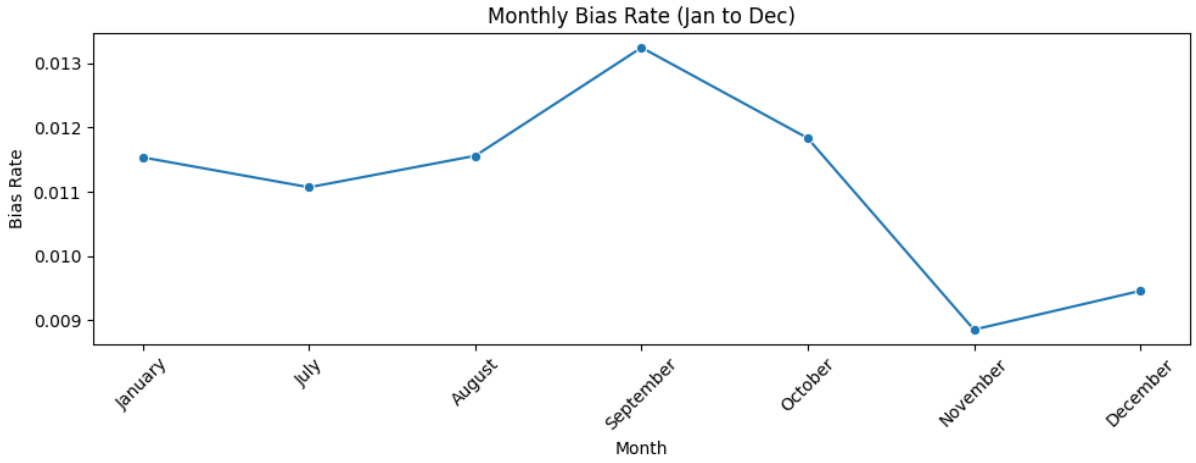
A clear springtime spike (April–May at ~1.9%) stands out against the low in March (~1.0%) and November (~0.8%), suggesting seasonal events (e.g. spring break, festivals) drive more biased feedback.

Edmonton



the Edmonton plot also shows a pronounced spring surge (peaking around May at ~2.0%) with a sharp dip in April, followed by a relatively steady bias rate through the rest of the year. The wider confidence bands in early (Jan–Feb) and mid-year months again flag low-volume periods, so treat those points with caution.

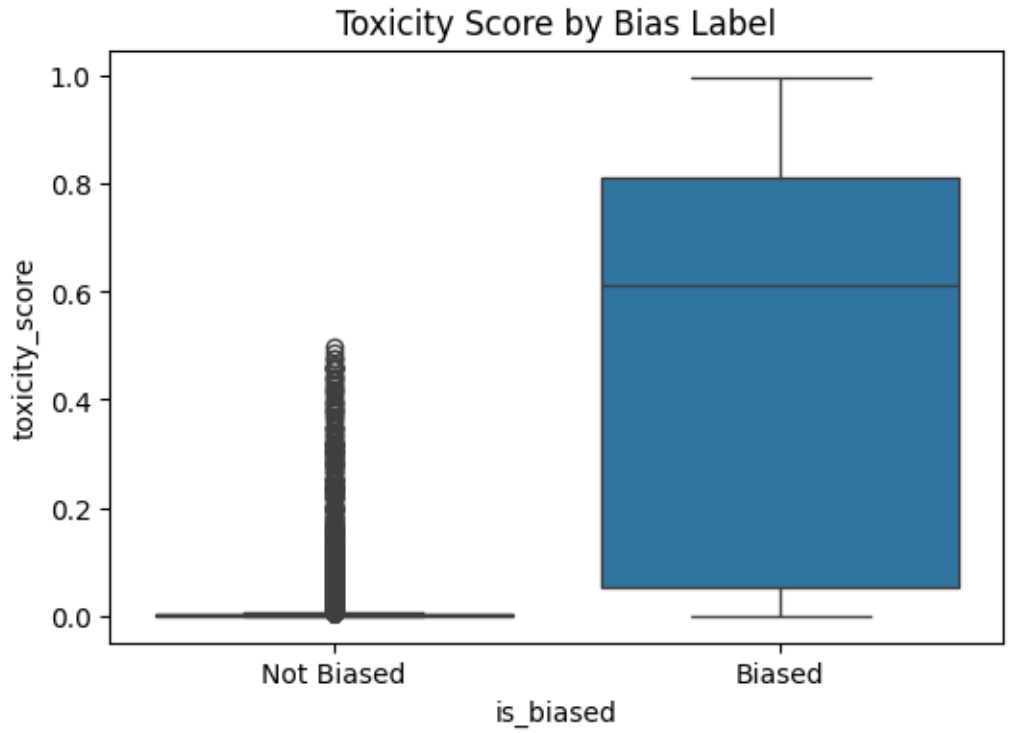
Indianapolis



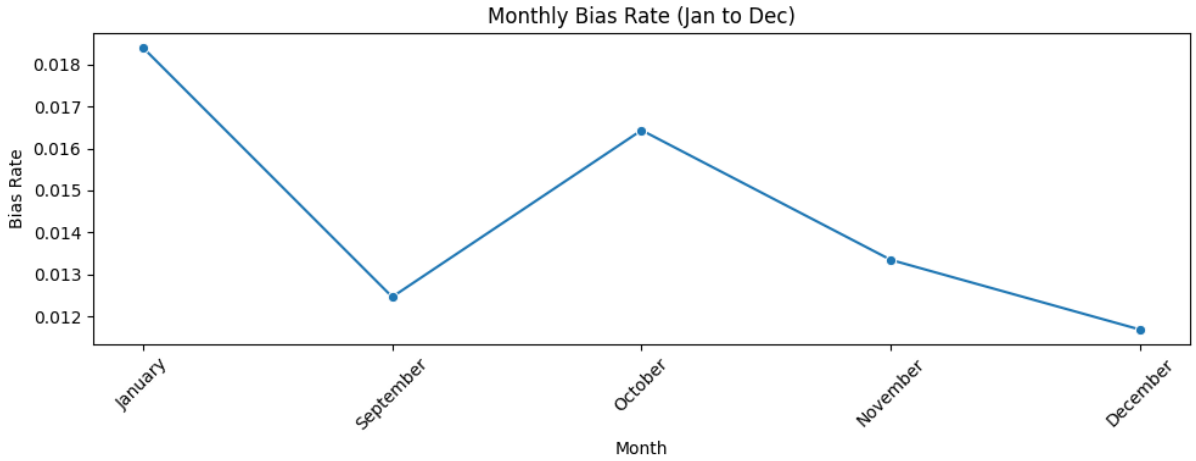
Overall, shows a milder seasonal swing compared with spring-peaking cities—indicating Indy's bias patterns are driven by different local dynamics.

New Orleans

Most of the biased reviews are rated 1 and next considerable amount (~ 50) are 5 star ratings, which is higher in this specific city.

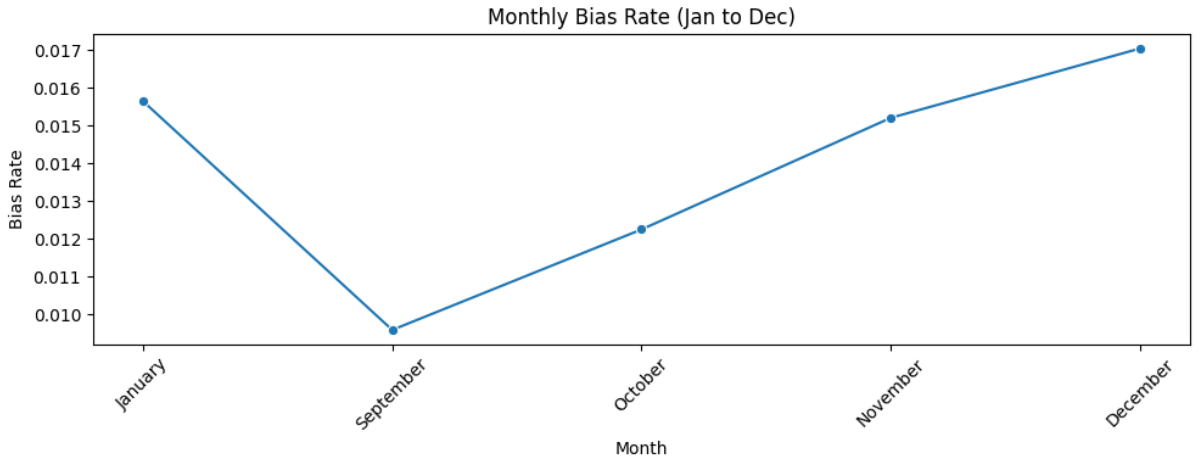


toxicity is a very strong proxy for bias—almost every highly toxic review is flagged, and nearly all unflagged reviews have negligible toxicity.



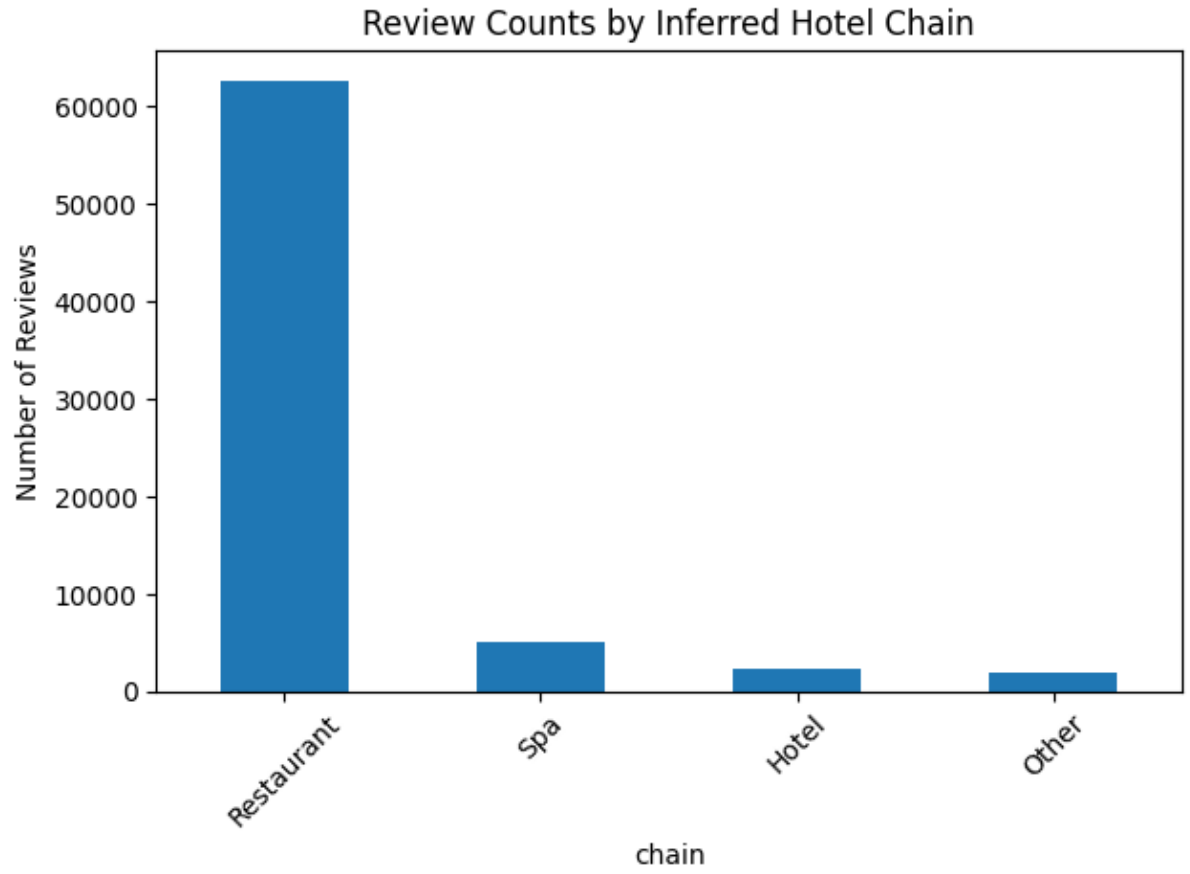
This city shows two clear bias peaks—in January (~1.8%) and again in October (~1.65%)—indicating that winter travel and fall events drive more biased feedback, so ramping up moderation during those specific seasons would be most effective.

Tampa

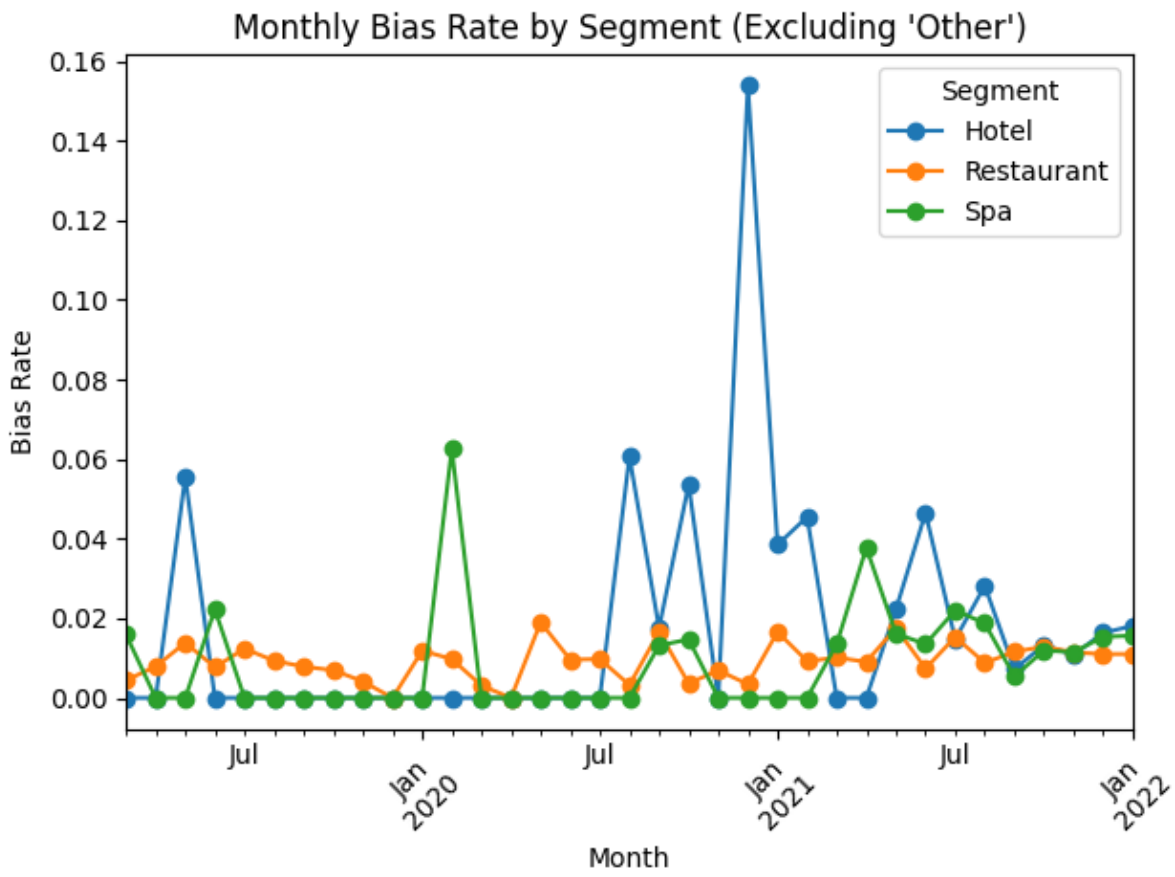


This pattern suggests that as Tampa moves out of its quieter “shoulder” season in early fall into the busy holiday period, higher visitor volumes, peak tourism stress, and service-level strains drive more emotionally charged—and thus more biased—feedback.

Distribution of categories



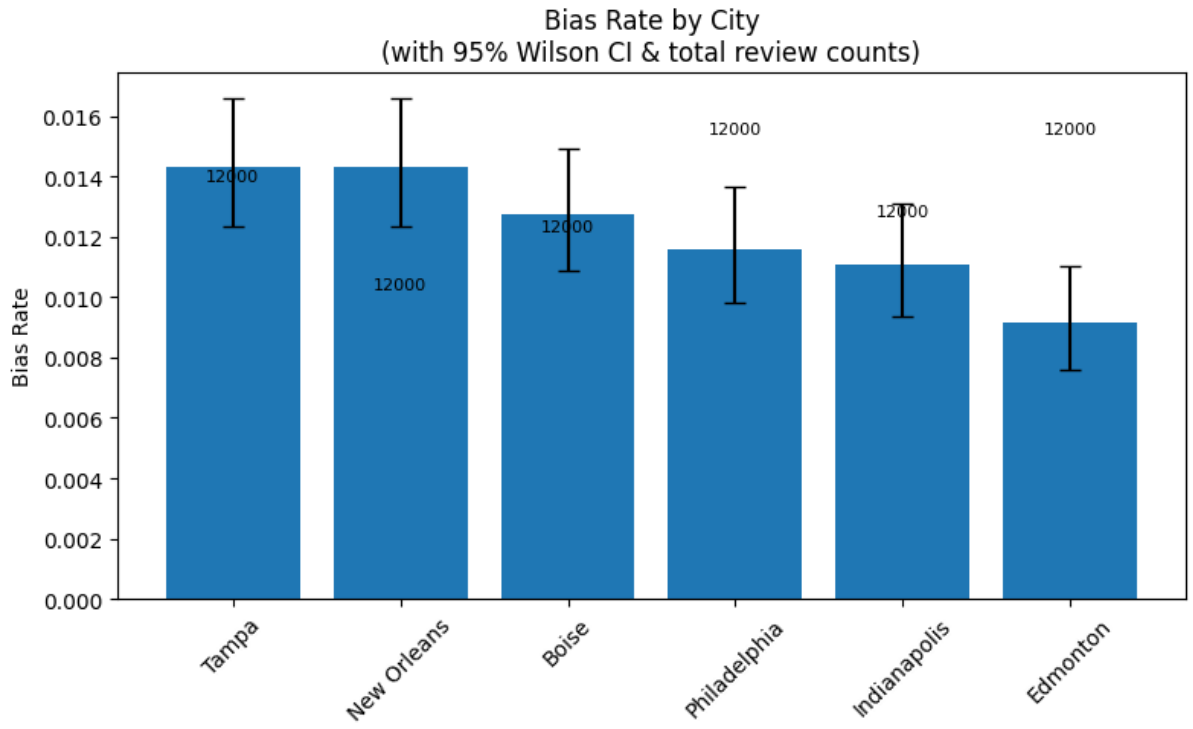
Bias rate by category

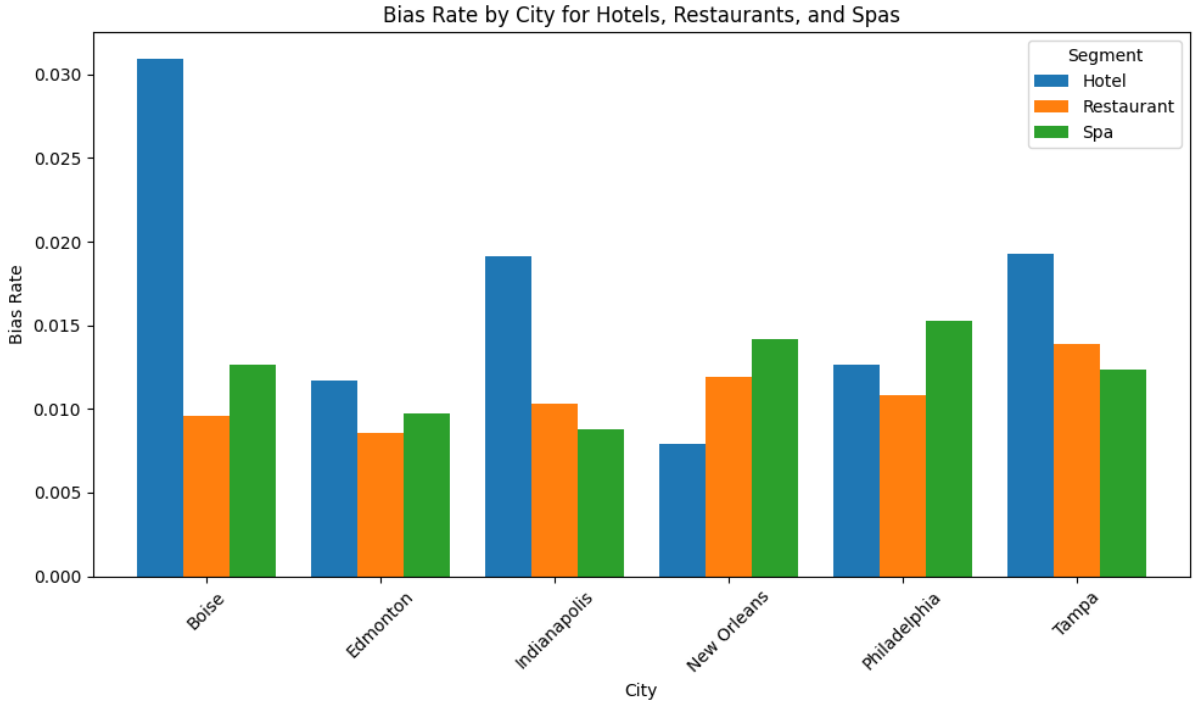


Restaurants and Spas hover at very low bias rates (usually under 1–2% each month), while Hotels occasionally spike much higher (e.g. up toward 6% in mid-2020 and above 15% in early 2021). That tells us:

Lodging carries more emotional stakes. Guests stay longer, encounter more complex services, and expectations run higher—so stress points in hotels lead to disproportionate biased feedback.

Seasonal surges align with travel peaks. The biggest hotel-bias spikes coincide with winter holiday and post-summer rushes, suggesting staff overload or policy changes drive up problematic language.



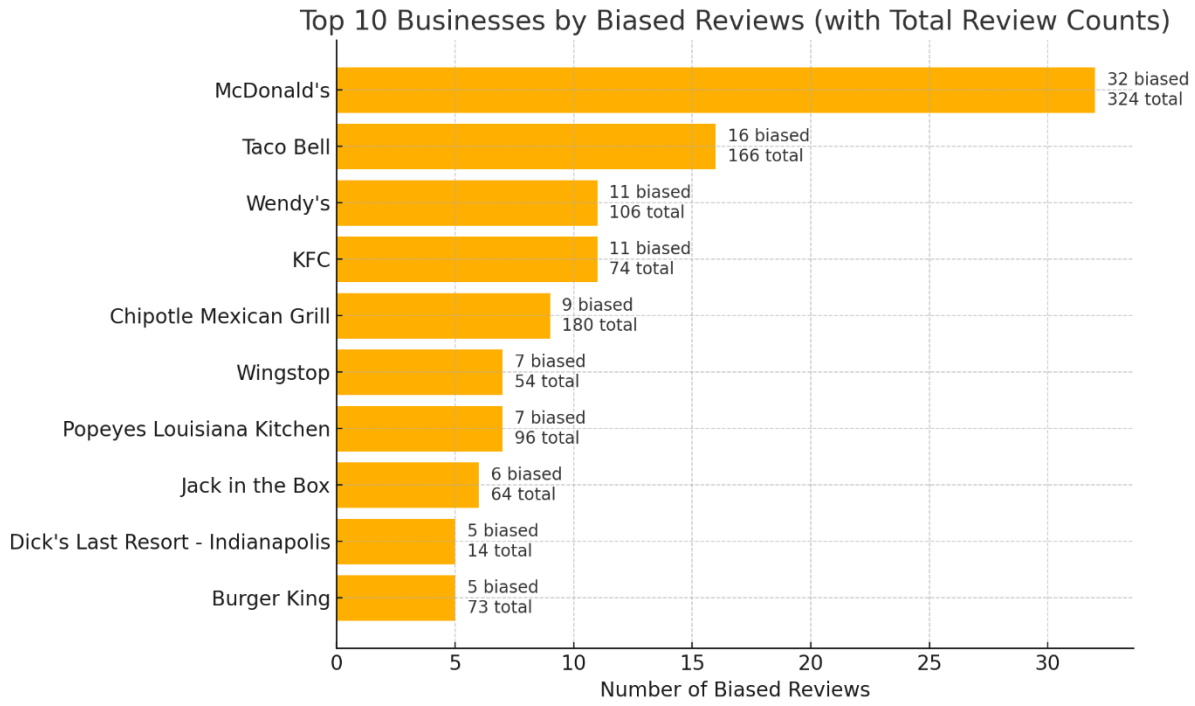


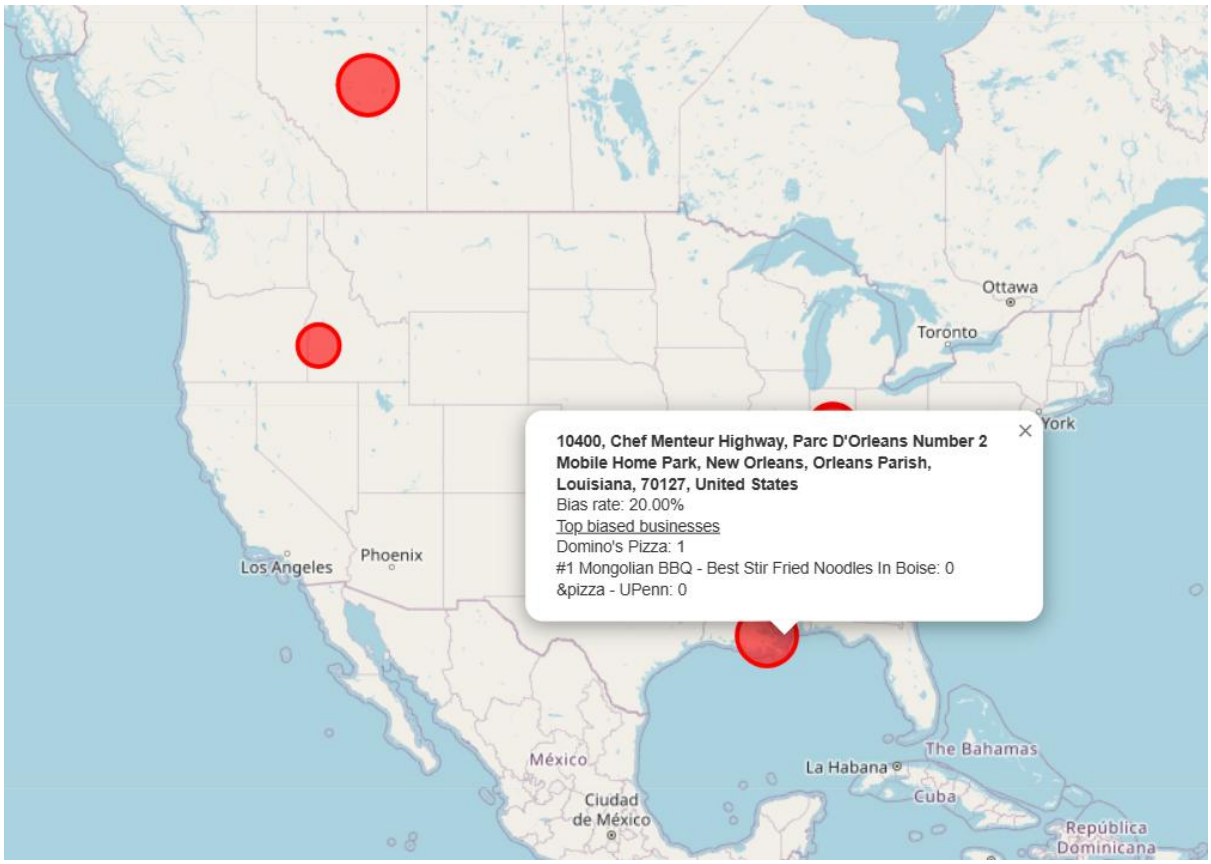
Hotels are the primary bias hotspot

Across most markets, lodging reviews carry the highest bias rates—often 2–3× those of restaurants or spas—reflecting the higher emotional stakes and service complexity of multi-day stays.

Local exceptions demand nuance

In a couple of cities (e.g. New Orleans, Philadelphia), spa reviews actually edge out hotels—showing that regional preferences or service models can flip the pattern.





A hyper-local hotspot in New Orleans—where a single neighborhood cluster shows a 20 % bias rate (over 14× the city average)—reveals that bias isn't evenly spread but concentrated in specific blocks, so moderation resources should be laser-focused on these exact areas.