FairwAI Hospitality Review Bias Detection Project

Submitted by: VARUN REDDY BHIMAVARAPU

Date: April 17, 2025

## Process Summary

The project began by filtering the Yelp Open Dataset to extract hospitality-focused businesses—specifically hotels, restaurants, and spas. Due to the absence of the FairwAI target cities (e.g., Las Vegas, Toronto) in the current dataset release, I selected cities that were available, including Tampa, Indianapolis, and New Orleans. I then filtered approximately 2.1 million customer reviews associated with these businesses. The reviews were cleaned using standard text normalization techniques, and then analyzed using keyword-based methods to detect potential signals of unfair treatment or bias, specifically in relation to race, gender, and staff behavior. This approach enabled scalable auditing of fairness signals across both geographic and business-type dimensions.

## Key Insights

1. Race-related bias indicators were the most frequent  
   Among the three categories of potential bias (race, gender, treatment), race-related terms were flagged most frequently. Words such as "black," "white," "Mexican," and "racist" appeared disproportionately in negative review contexts. This may suggest that race or perceived ethnicity plays a visible role in how hospitality experiences are communicated and perceived.
2. Over 678,000 reviews contained at least one potential bias signal  
   Roughly one-third of all analyzed reviews contained language aligned with race-, gender-, or treatment-related bias. While some flagged reviews may contain neutral context (e.g., "black tablecloth"), the aggregate volume suggests the presence of systemic behavioral patterns in staff conduct or customer perception.
3. Bias-related language clustered in specific cities  
   After integrating city-level metadata, certain locations consistently demonstrated higher proportions of flagged reviews. This geographic variation indicates potential regional or cultural differences in how bias manifests or is perceived by guests, and may warrant closer compliance review.

## Visual Evidence:

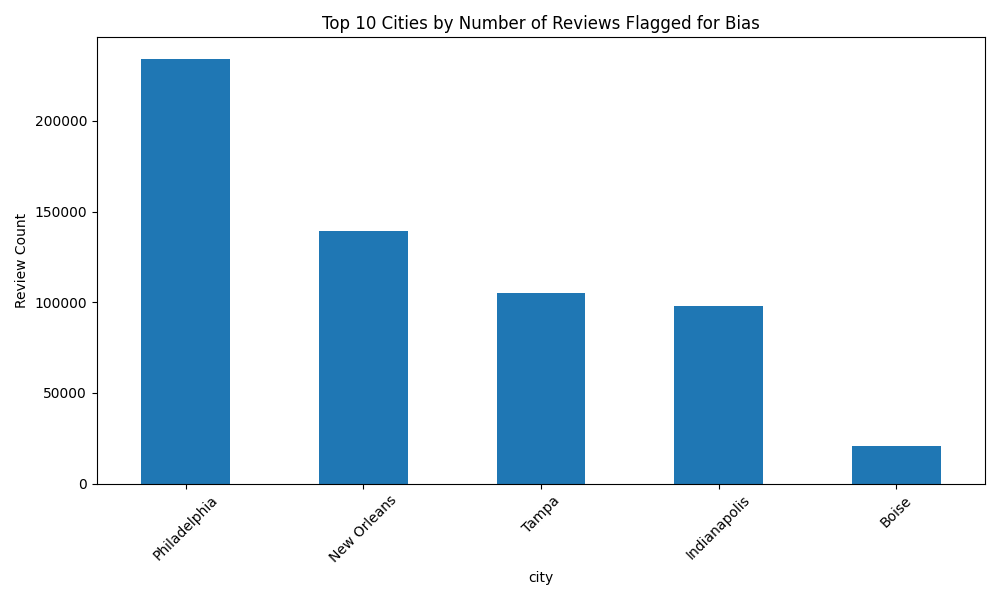
Bias Flags by Type (Race, Gender, Treatment)

A graph of a bar

AI-generated content may be incorrect.

Source: bias\_flag\_counts.png  
  
This bar chart illustrates the relative volume of flagged reviews across each category of potential bias. Race-related language appears most frequently, followed by gender and general treatment-related terms. The distribution highlights that racial identity may be the most visible axis of concern in hospitality-related customer feedback.

**Additional Visual: Top Cities by Flagged Reviews:**

****

This bar chart illustrates the top cities by total number of reviews flagged for potential bias. Philadelphia, New Orleans, and Tampa had the highest number of flagged reviews, suggesting these locations may be experiencing—or at least generating more feedback related to—racial, gendered, or unfair treatment incidents. The variation across cities underscores the importance of geographically targeted fairness audits. While the absolute number of reviews contributes to these totals, the consistency of elevated bias signals in certain regions may point to underlying patterns in service delivery, demographic dynamics, or staff conduct.

## Automation Opportunities and Ethical Considerations

Scalability and Automation:  
This type of fairness audit could be automated using natural language processing (NLP) pipelines incorporating keyword detection, sentiment scoring, and context-aware classifiers. A dashboard could track high-risk locations or brands in real time, allowing internal compliance teams to address issues proactively.  
  
Ethical Considerations:  
Keyword-based systems are prone to false positives and misinterpretation. For example, not every mention of "black" is racially charged. Any automated system should include a human-in-the-loop validation step and prioritize transparency. Bias detection tools should inform corrective action and training—not punishment—and must be culturally aware to avoid perpetuating harm.

## Included Files

- hospitality\_businesses.csv — Filtered hospitality business data  
- hospitality\_reviews.csv — Reviews linked to selected businesses  
- bias\_reviews.csv — Final dataset with cleaned text and flag columns  
- bias\_flag\_counts.png — Visual of review bias types  
- top\_bias\_cities.png — Visual of bias frequency by city  
- Python scripts:  
 - step1\_filter\_businesses.py  
 - step2\_filter\_reviews.py  
 - step3\_bias\_signals.py  
 - step4\_visualize\_bias.py  
- Submission summary (this document)