Methodology

This document provides an overview of the systematic approach, data sources, and analytical techniques used by our team to extract meaningful insights from customer reviews, focusing on key attributes influencing customer perception of local small businesses.

1. Study Overview

This study investigates core customer experience attributes as reflected in online reviews for small businesses. Leveraging advanced natural language processing (NLP) techniques, we aimed to identify dominant themes, analyze drivers of both satisfaction and dissatisfaction, and uncover emerging trends in customer feedback across a diverse range of industries.

2. Data Acquisition & Scope

Review data was collected from Google Business Listings for small businesses located exclusively in Lancaster County, PA. This deliberate geographic focus allowed our research team to utilize nuanced local knowledge, in an effort to make the results more helpful to small businesses nationwide. The dataset encompasses a mix of business types, including those in urban and suburban settings, catering to both tourist and local clientele, and representing various review volumes (niche industries, low-count, and high-count businesses). The reviews cover a period from February 2006 to July 14th, 2025.

3. Dataset Composition

The raw dataset initially comprised of 1,048,576 million customer reviews. Following a rigorous data preparation process, a total of 718,371 reviews with valid, processable text content were analyzed across general comparisons.

- **Overall Raw Review Distribution (from initial file):**
    - 5-star: 748,639
    - 4-star: 147,682
    - 3-star: 57,246
    - 2-star: 28,332
    - 1-star: 64,927
- **Key Sub-Dataset Compositions (Processed Reviews with Text):**
    - **5-star reviews:** 519,408 reviews (for positive drivers analysis)
    - **1-star reviews:** 55,710 reviews (for dissatisfaction drivers analysis)
- **Time-Series Analysis Datasets:**
    - **Most recent 24 months:** (July 2023 - June 2025): 16,380 reviews
    - **Prior 24 months:** (July 2021 - June 2023): 14,528 reviews
    - *Note: Only businesses with 100 or more reviews per period were included in the time-series comparison to ensure statistical significance of observed trends.*

- **Industry-Specific Analysis Datasets:**
    - Restaurants: 220,987 reviews
    - Hotels: 40,175 reviews
    - Salons: 19,591 reviews
    - Barber Shops: 6,434 reviews
    - Car Dealers: 8,331 reviews
    - Coffee Shops: 7,866 reviews
    - Dentists: 10,357 reviews
    - HVAC Contractors: 6,340 reviews
    - Grocery Stores: 5,864 reviews
    - Retail stores: 4,570 reviews
    - *Note: The dataset naturally skews towards food & beverage/hospitality businesses, reflecting their typically higher online review volumes.*

4. Data Pre-processing

Raw review data, initially in **.xlsx** format, was converted to **.csv** format for efficient loading and processing. The core data preparation and cleaning workflow was executed using Python, leveraging the `pandas` library for data handling and NLTK for linguistic processing:

- **Data Loading:** Reviews were loaded into a Pandas DataFrame from `.csv` files. The `on_bad_lines='skip'` parameter was utilized during loading to automatically bypass any malformed rows (e.g., those with inconsistent column counts or unescaped characters), which contributed to the difference between raw file size and successfully loaded review counts.
- **Text Cleaning:** Each review's text underwent a standardized cleaning pipeline:
    - **Lowercasing:** All characters were converted to lowercase.
    - **Punctuation & Number Removal:** Special characters, punctuation, and numerical digits were removed using regular expressions to isolate meaningful words.
    - **Tokenization:** The cleaned text was segmented into individual words (tokens) using NLTK's `word_tokenize` function.
    - **Stop Word Removal:** Common, non-thematic words (e.g., "a", "the", "is") were removed using NLTK's default English stopwords list, supplemented by a custom list of generic sentiment words and irrelevant industry-specific terms (e.g., "restaurant," "good," "bad").
    - **Lemmatization:** Words were reduced to their base or dictionary form (e.g., "running," "runs," "ran" all became "run") using NLTK's `WordNetLemmatizer` to ensure that variations of the same word were counted consistently.

5. Attribute Definition & Keyword Association

To quantify specific aspects of the customer experience, a comprehensive framework of business attributes was established.

- **Attribute Curation:** Over 150 unique attributes were identified across various types of local businesses. For detailed analyses, specialized lists of approximately 40-60 attributes were developed for each specific business vertical (e.g., restaurants, hotels, dentists). These attributes were designed to capture customer feedback beyond basic product/service quality, focusing on experience factors (e.g., Staff Friendliness, Cleanliness, Speed of Service, Digital Payment options).
- **Keyword Association:** For every defined attribute, a corresponding list of associated keywords was manually compiled. These keywords included both positive and negative terms related to the attribute, all converted to lowercase and their lemmatized form (e.g., 'clean' for Cleanliness, 'rude' for Staff Friendliness). An attribute was counted as "mentioned" in a review if any of its associated keywords appeared in the cleaned text of that review; each review contributed a maximum of one mention per attribute, regardless of how many times keywords for that attribute appeared within it.

6. Analytical Framework

The cleaned and attribute-tagged review data was subjected to several quantitative comparative analyses:

- **Attribute Mention Frequency:** For each analyzed dataset (e.g., overall, filtered by star rating, or by time period), the raw count of mentions for each attribute was tallied and expressed as a percentage of the total reviews within that dataset.
- **Comparative Analysis (Star Ratings):**
    - To pinpoint drivers of satisfaction and dissatisfaction, attribute mentions in 5-star reviews were compared directly against those in 1-star reviews.
    - **Key Metrics:**
        - **Percentage of Reviews Mentioning Attribute:** Calculated for both 1-star and 5-star datasets.
        - **Difference in Percentage Points (1-Star vs. 5-Star):** Highlighted attributes more or less prevalent in negative feedback.
        - **Ratio (1-Star vs. 5-Star):** Indicated how many times more likely an attribute was to be mentioned in a 1-star review compared to a 5-star review. This was used to identify "pain points" (high ratio) and "delighters" (low ratio).
- **Time-Series Trend Analysis:**
    - To identify evolving customer priorities and changes in business performance, attribute mention frequencies from the **Most recent 24 months** were compared against those from the **Prior 24 months**.
    - **Key Metrics:**
        - **Percentage of Reviews Mentioning Attribute:** Calculated for both recent and prior periods.

    - **Difference in Percentage Points (Recent vs. Prior):** Quantified the increase or decrease in an attribute's mention frequency over time.
    - **Ratio (Recent vs. Prior):** Indicated the relative change in mention frequency between the periods.
- **Business Vertical Analysis:** The attribute analysis was applied to filtered datasets specific to different business types (e.g., Restaurants, Hotels), using curated industry-specific attribute lists to provide tailored insights.

7. Key Metrics Calculated

The analyses primarily generated the following quantitative metrics for each attribute:

- **Mentions Count:** The raw number of reviews mentioning a specific attribute.
- **Percentage of Reviews:** The percentage of total reviews (within a given dataset) that mention an attribute.
- **Difference in Percentage Points:** The direct percentage point difference between two groups (e.g., `Percentage_GroupA - Percentage_GroupB`).
- **Ratio of Mentions:** A multiplier indicating how many times more likely an attribute is to be mentioned in one group versus another.

8. Tools & Technologies

The entire analytical process was implemented using:

- **Programming Language:** Python (3.13.5)
- **Key Libraries:**
    - `pandas`: For robust data loading, manipulation, and structuring DataFrames.
    - `nltk` (Natural Language Toolkit): For core text processing functions (tokenization, stop word removal, lemmatization).
    - `collections`: Specifically the `Counter` class for efficient frequency counting.
    - `re`: For regular expressions in text cleaning.
    - `os`: For dynamic handling of file paths across different operating systems.

9. Study Limitations

It is important to acknowledge certain limitations inherent to this study's methodology:

- **Keyword Dependence:** The accuracy of attribute identification is directly dependent on the completeness and precision of the manually compiled keyword lists. Reviewers may use synonyms, slang, or nuanced phrasings not included, potentially affecting counts.
- **Absence of Explicit Sentiment Analysis:** While attributes are counted, their associated sentiment (positive, negative, neutral) is not explicitly analyzed within this framework. An increase in mentions of an attribute does not automatically imply positive feedback, nor does a decrease automatically imply negative.

- **Data Quality & Omissions:** The analysis relies on the quality of the raw review data. Malformed lines in CSV files were automatically skipped, and potential biases inherent in the review collection process (e.g., review platform biases, self-selection bias of reviewers) are assumed to be outside the direct scope of this attribute analysis.
- **Manual Keyword Refinement:** The process of compiling and refining attribute keywords involved human judgment, which, while informed by local knowledge, introduces a degree of subjectivity.

This methodology provides a framework for extracting and quantifying key customer experience attributes from large volumes of review data, offering data-driven insights for strategic decision-making in the small business landscape.