UNIVERSITY OF
HOUSTON

EDS – 6397 INFORMATION VISUALIZATION

Disneyland Review Analysis Using NLP & Tableau

Professor: Dr. Lucy Nwosu

Group - VIII

Varun Vaddi (2347481)

Surya Vardhan Reddy (2311061)

Akhila Vemuru (2349951)

Poojitha Reddy Bommu (23111344)

Harshini Nimmala (2316672)

Sanjana Ponaganti (2312535)

Monika Nanjappa (2267139)

Bindhu Sri Chinta (231689)

Shreyas Mysore Narayana (2307777)

Objective

The primary objective of this study is to conduct a comprehensive analysis of over 40,000 publicly available Disneyland reviews from parks in California, Paris, and Hong Kong using Natural Language Processing (NLP) and interactive data visualization techniques. Specifically, the study aims to assess visitor sentiment, identify recurring themes such as cost, cleanliness, staff behavior, and crowd management, and examine regional and temporal variations in guest satisfaction. Through advanced text analytics— including sentiment scoring and keyword extraction— and the engineering of over 45 features, the analysis seeks to uncover actionable insights that reflect the lived experiences of park visitors. Furthermore, the development of interactive Tableau dashboards enables dynamic exploration of patterns by location, year, and sentiment, ultimately supporting data-driven decision-making and strategic enhancements to the guest experience across Disneyland's global operations.

Dataset Overview

Dataset: DisneylandReviews.csv

This dataset consists of online reviews collected from visitors to Disneyland theme parks across various international locations. Each entry in the dataset represents a single review submitted by a user, along with associated metadata such as the rating, date, location, and the park branch visited.

Dimensions:

- Total Records: 42,656
- Total Features: 6

Feature Descriptions:

Review_ID:

- Datatype: Integer
- Unique identifier for each review.

Rating:

- Datatype: Integer
- Rating provided by the reviewer (range:1 to 5).

Year_Month:

- Datatype: Date
- The year and month when the review was posted, in YYYY-MM format.

Reviewer_Location:

- Datatype: String
- Country or region of the reviewer.

Review_Text:

- Datatype: String
- Free-form text of the review, often descriptive of the visitor experience.

Branch:

- Datatype: String
- Disneyland branch reviewed (e.g., Disneyland_Paris, Disneyland_HongKong).

Data - Preprocessing:

To prepare the data for analysis, extensive preprocessing was performed. This included text normalization (lowercasing, punctuation removal), date parsing, and the elimination of null or duplicate entries. Feature engineering was a core component, resulting in the creation of 45 new columns covering temporal patterns, text structure, sentiment scores, and keyword flags. These enhancements allowed for a multi-dimensional exploration of guest behavior.

Datetime Conversion: Parsed Year_Month into proper datetime format; extracted year, month, day_of_week, Review_Quarter.
Text Cleaning: Lowercased text removed special characters, extra spaces, and punctuation using regex.
Duplicates & Nulls: Removed duplicate rows and invalid/missing Year_Month entries.

Feature Engineering:

- **Total Features Created**: 45
- **Types of Features**:
    - **Temporal**: year, month, Review_Quarter
    - **Text Metrics**: word/character counts, punctuation, digit, and uppercase usage
    - **Boolean Flags**: e.g., presence of exclamations, numbers, adjectives
    - **Encodings**: Reviewer and branch locations
    - **Ratings**: Categorical rating levels and high-rating indicator
    - **Reviewer Context**: Local reviewer flag, review counts per location/branch
    - **Branch Info**: Country, continent, monthly review count

NLP Features:

Natural Language Processing (NLP) techniques were employed using Python libraries such as TextBlob for sentiment analysis and NLTK for part-of-speech tagging. These tools enabled the extraction of key opinion words (adjectives, adverbs) and the categorization of reviews into sentiment labels (Positive, Negative, Neutral, Mixed).

- **Sentiment Analysis (Text Blob)**:
    - Polarity (Sentiment_Score), subjectivity, and sentiment labels (Positive, Neutral, Negative, Mixed)
- **Keyword Extraction**:
    - Top 5 words per review using POS tagging (nouns, adjectives, adverbs)
- **Opinion Word Frequency**:
    - Top adjectives/adverbs identified per sentiment type

Thematic Flags:

Flags are set based on the presence of concern-related terms in the review text:

- **Mentions_Staff**: e.g., “staff”, “employee”, “service”
- **Mentions_Crowd**: e.g., “crowd”, “busy”, “packed”
- **Mentions_Cost**: e.g., “price”, “expensive”, “cheap”
- **Mentions_Cleanliness**: e.g., “clean”, “dirty”, “hygiene”

New Dataset:

- **Original**: 42,656 reviews with 6 raw features.
- **Transformed**: 45 enriched features.

Attribute	Meaning and Usage
Review_ID	Unique identifier for each review; used for indexing.
Rating	Numerical rating given by the reviewer; primary target for sentiment or satisfaction analysis.
Year_Month	Time period of the review; useful for temporal trends analysis.
Reviewer_Location	Geographical location of the reviewer; helps identify regional trends or biases.
Review_Text	Full text of the review; primary source for NLP-based analysis.
Branch	Indicates which Disneyland branch the review is about; useful for branch-specific insights.
year	Year extracted from review date; supports time-based analysis.
month	Month extracted from review date; used for seasonal trend analysis.
month_name	Name of the month; same as above but in readable format.
day_of_week	Day of the week when review was posted; useful for weekly patterns.
review_length	Character length of the review; may correlate with review detail or sentiment.
review_word_count	Total number of words in the review; similar relevance to review_length.
review_sentence_count	Number of sentences; used in linguistic or readability analysis.
word_density	Average word length or richness; used in textual complexity analysis.
uppercase_count	Count of uppercase letters; may indicate emphasis or emotion.
digit_count	Count of numerical digits; might indicate mention of dates, prices, or counts.
punctuation_count	Total punctuation marks; might correlate with expressiveness.
review_has_exclamatior	Boolean indicating presence of exclamation; proxy for emotional tone.
review_has_question	Boolean indicating if review has questions; can show uncertainty or inquiry.
review_has_numbers	Boolean indicating if numbers are mentioned; might point to data or metrics.
location_encoded	Numerical encoding of reviewer location; used in ML models.
location_length	Character length of location name; minor feature, sometimes useful in text patterning.

Attribute	Meaning and Usage
branch_encoded	Numerical encoding of branch; required for model input.
rating_category	Categorical version of rating (e.g., high/low); used in classification models.
is_high_rating	Boolean indicating if rating is high; simplifies prediction targets.
top_5_words	Most frequent words in the review; used for keyword-based analysis.
review_has_adjectives	Boolean indicating if adjectives are present; key for sentiment and descriptiveness.
review_has_comparison	Shows presence of comparative language; useful in opinion mining.
Review_Quarter	Quarter of the year; supports time-series segmentation.
Local_Reviewer	Indicates if reviewer is local; may impact expectations and ratings.
Total_Reviews_Per_Revie	Review frequency per user; used for trust or bias evaluation.
Branch_Location	Physical location of the branch; geographical segmentation.
Branch_Country	Country of the branch; high-level location feature.
Continent	Continent of the branch; broader regional grouping.
Monthly_Reviews_Per_Bı	Volume of reviews per branch per month; useful in trend and popularity analysis.
Avg_Word_Length	Average word length in the review; indicator of writing style or readability.
Sentiment_Score	Numerical sentiment score from text; central to opinion analysis.
Sentiment_Category	Categorical sentiment (positive/neutral/negative); derived from sentiment score.
Review_Subjectivity	Measure of subjectivity in review; helps separate fact vs opinion.
Mentions_Staff	Boolean flag for mentions of staff; important for service quality analysis.
Mentions_Crowd	Indicates crowd-related mentions; relates to comfort, experience.
Mentions_Cost	Mentions of cost or price; important for value analysis.
Mentions_Cleanliness	Mentions of cleanliness; key for hygiene and satisfaction.
sentiment2	Alternative or secondary sentiment label; used for comparison or ensemble models.
sentiment_keywords	Specific keywords that indicate sentiment; useful for interpretable NLP models.

Visualization Description

The data visualization component of this project was central to uncovering patterns, trends, and anomalies in Disneyland guest reviews. Guided by Edward Tufte's principles of graphical integrity and Colin Ware's principles of perception and clarity, the dashboards were designed for accurate, clear, and engaging storytelling.

Exploratory visuals such as bar charts, stacked area graphs, and heat maps were used to show sentiment trends, visitor counts, and complaint volumes over time. Scatter plots and matrix charts were employed for comparative performance analysis between Disneyland branches.

Key comparative visualizations include:

- Rating distribution by branch and sentiment
- Year-over-year trends in mentions of crowd, cost, cleanliness, and staff
- Visitor origin maps showing regional satisfaction
- A branch performance matrix comparing average rating and standard deviation

Tufte's principles were applied through accurate axis labeling, the preservation of outliers, and careful avoidance of visual distortion. Each graph's scale was chosen to match the data's natural variation, ensuring interpretability and fairness in comparisons.

Colin Ware's perceptual design principles were implemented using:

- Color to distinguish sentiment and regions
- Size and shape to emphasize performance variability
- Layout and spacing to reduce cognitive load

Design Justification

A key success factor of this project was the design and delivery of an intuitive and engaging user experience through Tableau dashboards. By applying Colin Ware's perceptual and engagement principles, the dashboard was crafted to maximize clarity, interactivity, and user exploration.

User-Friendly Design:

- Clean and consistent layout with readable fonts and logical sectioning.
- Use of whitespace and alignment to minimize clutter and enhance comprehension.
- Color palettes chosen to emphasize key differences (e.g., sentiment polarity, branch performance) without overwhelming the user.

Interactive Features:

- Filters for Year, Continent, Branch, and Sentiment Category allow users to customize views and analyze subsets of data.
- Hover effects dynamically update charts and tooltips, helping users trace influences and drill down by region.
- Forecast lines and dynamic performance matrices support real-time hypothesis testing and trend spotting.

Exploratory Capability:

- Users can observe how ratings, review volume, and complaint patterns evolve across time and space.
- The dashboard supports business storytelling by letting stakeholders explore sentiment shifts during specific quarters, like peak holiday periods

Accessibility and Engagement:

- The dashboard was published to Tableau Public to ensure open access and cross-platform usability.
- Load times and responsiveness were tested to ensure smooth navigation and responsiveness on both desktop and mobile.

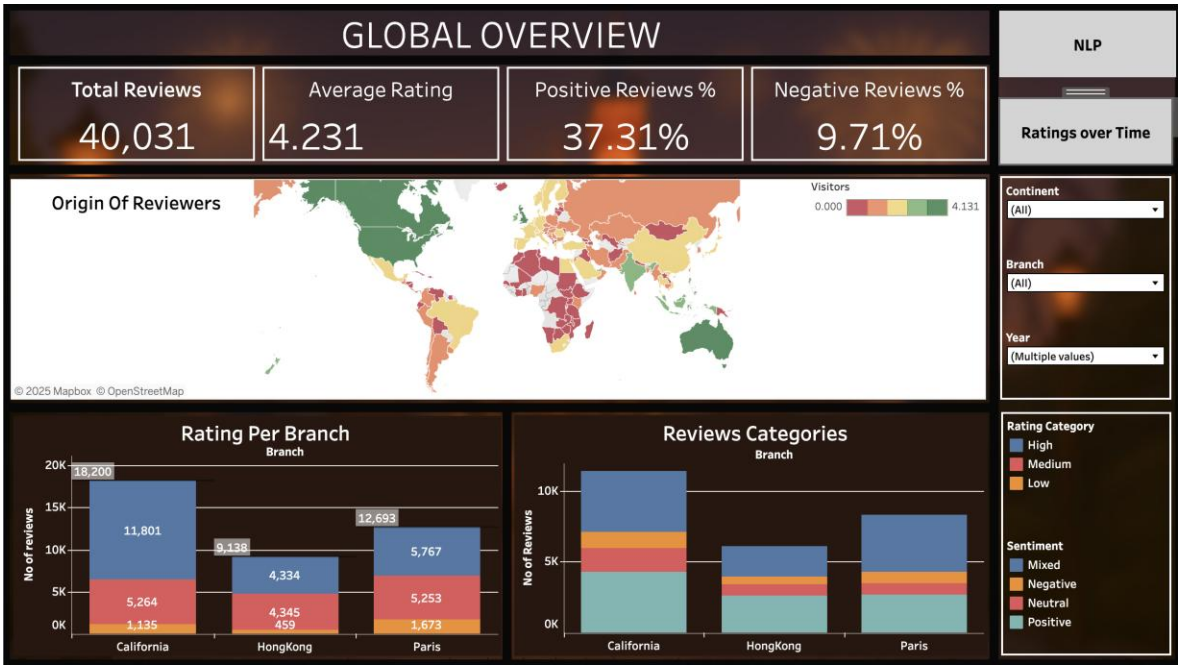
Ethical Considerations

Ethical considerations play a critical role in ensuring that data visualizations and analytical results are both honest and responsible. This project adheres to best practices in data ethics, focusing on transparency, fairness, and accurate representation.

- All visualizations were carefully constructed to **avoid distortion**. Axes were appropriately scaled, and zero baselines were used where applicable.
- **Outliers and anomalies** were retained to preserve the integrity of the analysis and prevent cherry-picking of favorable data.
- Graphs and dashboards included **clear titles, labels, and legends** to ensure user understanding and reduce the risk of misinterpretation.
- Reviews were not **altered or removed** based on sentiment. Both positive and negative feedback were treated equally in sentiment scoring.
- Complaint categories were derived using **transparent keyword matching** logic, ensuring replicability of results.
- The **Influence Score formula** was disclosed and explained to prevent the misuse or overemphasis of subjective metrics.
- The dataset used was publicly available and anonymized, ensuring that **no personally identifiable information (PII)** was exposed.
- Insights and interpretations were framed objectively, **avoiding bias or assumptions** not supported by the data.
- The **log scale** was transparently applied and clearly documented to avoid misleading interpretations.
- No countries were removed based on low review counts—ensuring the visualization maintains **complete global representation**.

1. Global Overview

The Global Overview dashboard offers a comprehensive snapshot of Disneyland's global sentiment landscape. With over 40,000 reviews analyzed, it's evident that California attracts the most attention, particularly from North America, while Paris and Hong Kong show region-specific visitor trends. The distribution highlights how proximity, cost, and regional identity influence park selection. California's high review volume is balanced by strong positivity, but Paris and Hong Kong offer more nuanced engagement, calling for tailored improvements. This view sets the foundation for understanding the broader sentiment ecosystem.



Insights

- **Overall Satisfaction Trends**
  The dashboard reveals consistent satisfaction levels over time, indicating sustained performance in delivering quality guest experiences.
- **Category-Specific Feedback**
  Analysis of specific feedback categories, such as cleanliness, staff friendliness, and ride experiences, highlights areas of excellence and opportunities for improvement.
- **Temporal Patterns**
  Temporal analysis uncovers seasonal variations and the impact of specific events or changes on visitor satisfaction.

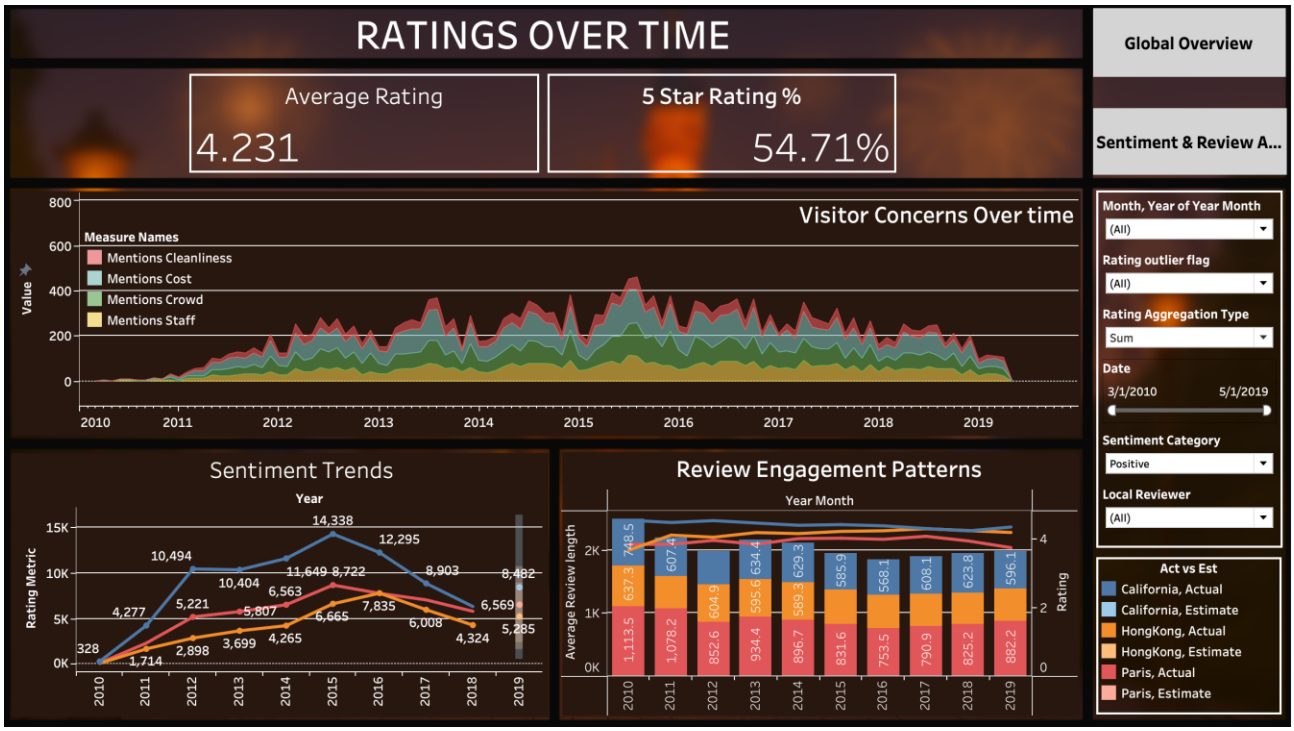
Observations

- High satisfaction scores in areas like staff interaction suggest effective training and customer service strategies.
- Lower scores in certain categories may point to areas requiring attention, such as ride maintenance or food services.

- Seasonal dips in satisfaction could correlate with higher crowd levels, indicating a need for crowd management strategies.

2. Ratings Over Time

This dashboard showcases how visitor satisfaction and concern areas have evolved from 2010 to 2019. While the overall average rating remains strong at 4.23, certain operational pain points—like mentions of crowd, cost, and cleanliness—spike during peak seasons. The rise and fall of 5-star ratings, along with changing review volumes, reflect how external factors (holidays, promotions, economic shifts) shape guest perception. The trend of shorter, sentiment-rich reviews also signals a shift in how guests communicate their experiences—valuable for improving review analysis methods.



Insights

1. **Consistent Performance Across Years:**
   Visitor ratings have shown relatively stable patterns across the decade, suggesting that Disneyland has maintained a consistent level of service and experience.
2. **Variations Between Parks:**
   Filtering by branch reveals that some locations perform slightly better or worse in specific years. This indicates park-specific factors—such as infrastructure upgrades, seasonal events, or crowd management—that influence satisfaction.
3. **Notable Rating Peaks and Dips:**
   Certain years show visible increases or decreases in average ratings. These shifts may align with internal changes (e.g., new attractions, policy adjustments) or external factors (e.g., travel restrictions, economic changes).
4. **Empowered Exploration:**
   The dashboard's interactivity allows users to generate personalized insights. Whether analyzing a

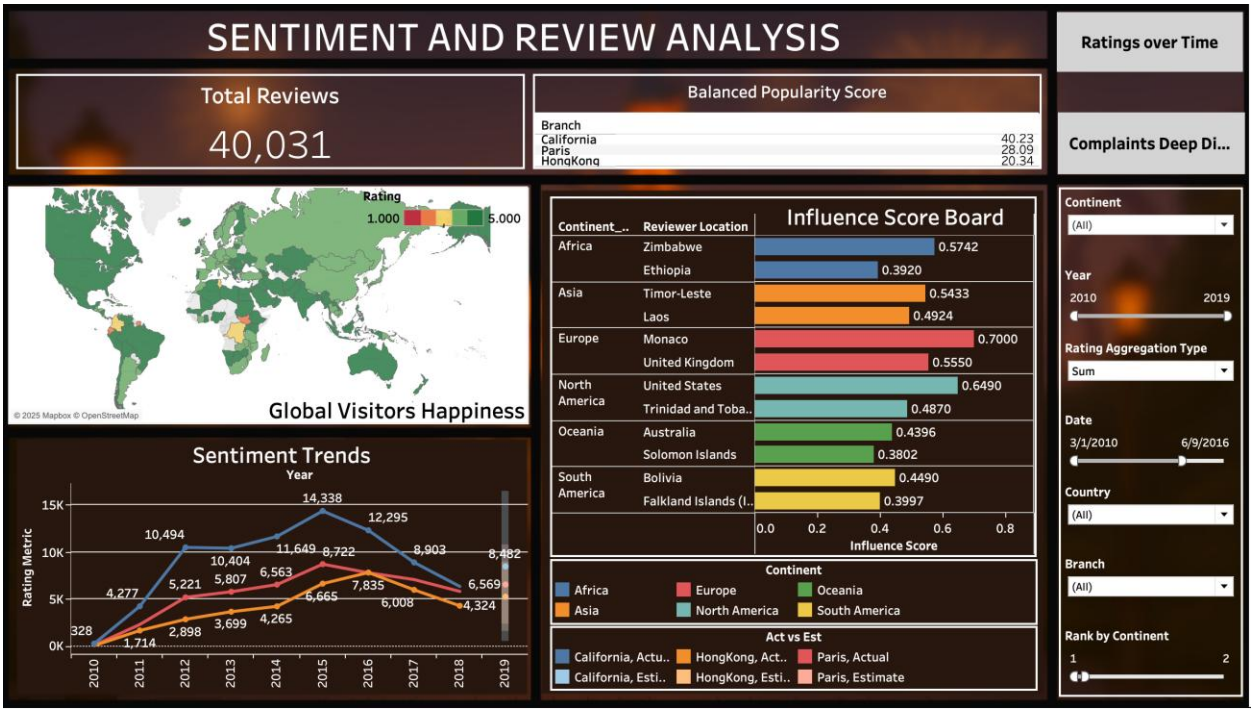
specific year or comparing parks side by side, users can tailor the dashboard view to their analytical needs.

Observations

- Ratings have **remained positive overall**, indicating strong brand loyalty and consistently satisfying guest experiences.
- **Short-term fluctuations** are evident and merit further investigation to understand contributing factors.
- **The California and Paris parks** tend to show slightly higher average ratings than the Hong Kong branch in certain time frames, which could reflect differences in visitor expectations, crowd sizes, or regional preferences.
- The **intuitive layout and interactivity** make the dashboard accessible and user-friendly, even for non-technical stakeholders.

3. Sentiment and Review Analysis

This dashboard dives deeper into where reviews come from and how influential they are. It confirms that European and North American reviewers—especially from countries like Monaco, the US, and the UK—leave longer, more impactful reviews. These regions hold higher influence scores, suggesting that Disney should strategically prioritize their feedback for brand perception and messaging. Sentiment strength remains highest among visitors from North America and Oceania, while other regions show potential sentiment gaps tied to cost or unmet expectations.



Insights

- **Demographic Trends**
  Analysis of visitor demographics reveals patterns in age distribution, visit frequency, and other key characteristics, helping to tailor marketing and operational strategies.

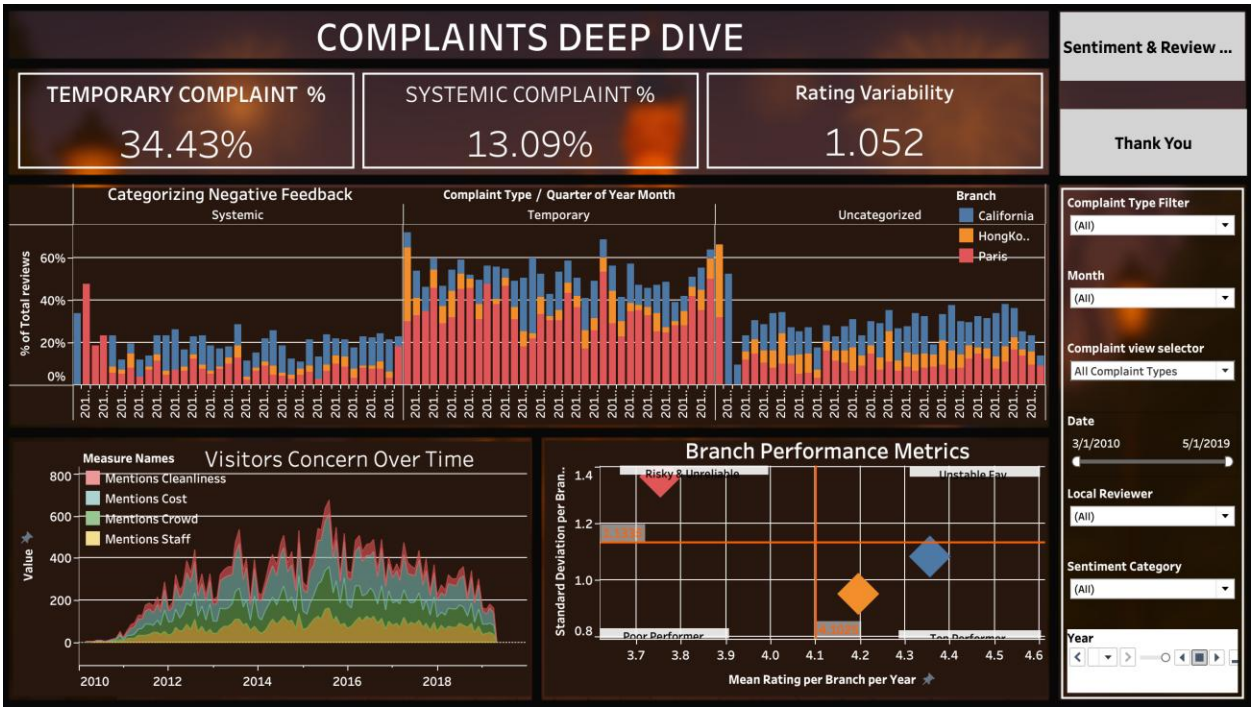
- **Attendance Patterns**
  The dashboard highlights fluctuations in park attendance over time, identifying peak periods and potential areas for capacity management.
- **Satisfaction Correlations**
  By correlating satisfaction scores with other variables, stakeholders can pinpoint factors that contribute to positive or negative guest experiences.

Observations

- Certain age groups may exhibit higher satisfaction levels, indicating successful engagement strategies for those demographics.
- Attendance peaks often align with specific events or seasons, suggesting opportunities for targeted promotions or resource allocation.
- Variations in satisfaction scores across different parks or time periods may highlight areas for improvement or successful initiatives worth replicating.

4. Complaints Deep Dive

This section effectively categorizes negative reviews into systemic (e.g., crowding, poor value) and temporary (e.g., ride closures, weather). California experiences a larger proportion of systemic issues, suggesting deeper operational improvements are needed. Paris sees more temporary complaints, indicating situational dissatisfaction. Hong Kong stands out with the least negative feedback overall, highlighting its operational consistency. By breaking down complaints in this structured way, the dashboard empowers teams to target improvements based on root cause and recurrence.



Insights

- **Wait Time Analysis**
  The dashboard highlights the impact of wait times on visitor satisfaction, emphasizing the importance of efficient queue management.
- **Ride Availability**
  Data on ride downtimes and maintenance schedules reveal their influence on overall guest experiences.
- **Crowd Level Correlations**
  Analysis shows how varying crowd levels affect satisfaction, providing guidance for capacity planning and crowd control measures.

Observations

- Extended wait times are associated with lower satisfaction scores, underscoring the need for effective queue management systems.
- Frequent ride downtimes negatively impact visitor experiences, suggesting the importance of proactive maintenance.
- High crowd levels correlate with decreased satisfaction, indicating a need for strategies to manage park capacity during peak times.

Key Insights

Name	Title	Key Insight
Varun	*Origin of Visitors*	Visitors typically choose the Disneyland location closest to their region — Europeans favor Paris, Asians prefer Hong Kong, and North Americans opt for California. Proximity and cost drive this behavior.
Surya	*Visitor Concerns Over Time*	While most reviews are positive, crowd and cost concerns spike during peak seasons. Cleanliness and staff feedback remain consistent, highlighting areas for operational improvement.
Shreyas	*Temporal Sentiment Trends*	Average ratings and footfall trends can be forecasted for future years, enabling better planning and resource allocation.
Akhila	*Categorizing Negative Feedback*	Paris faces more temporary complaints, while California has more systemic issues. Hong Kong has the fewest negative comments, indicating relatively higher consistency.
Bindhu	*Sentiment Distribution Across Locations*	California has consistent, mostly positive feedback and high volume. Paris shows instability in sentiment. Hong Kong has fewer but more balanced reviews.
Poojitha	*Influencer Scoreboard*	European and North American reviewers, especially from Monaco, are the most influential due to detailed, high-quality reviews that shape broader park perceptions.
Harshini	*Happiness Based on Ratings*	Visitors from Western Europe and North America are the happiest. Lower ratings from Africa, South America, and Southeast Asia may stem from unmet expectations or cost sensitivity.
Sanjana	*Review Engagement Patterns*	Review lengths are becoming more concise yet sentiment-rich. This helps sentiment analysis tools extract clearer feedback, reflecting evolving visitor communication styles.
Monika	*Branch Performance Over Time*	California consistently performs well. Paris shows volatility and risk. Hong Kong has improved steadily, showing potential for growth

Conclusion

The Disneyland data analysis project successfully leveraged text analytics, sentiment classification, and interactive dashboards to derive actionable insights from over 40,000 visitor reviews. By analyzing feedback across global branches — California, Paris, and Hong Kong — the project highlighted key differences in customer satisfaction, complaint patterns, and regional sentiment trends.

Summary of Findings:

- California Disneyland emerged as the top-performing branch in both volume and sentiment of reviews.

- Paris Disneyland revealed greater inconsistencies, with elevated systemic complaints around cost and crowding.

- Hong Kong displayed stable but modest engagement, with room to improve visitor experience and outreach.

- Seasonal and regional sentiment trends aligned with major holidays and economic conditions, supporting predictive modeling.

Implications:

The findings provide strategic value by pinpointing where operational changes are needed and which visitor groups are most influential. Complaint tracking and sentiment scores offer a framework for continuous experience monitoring.

Future Enhancements:

- Extend the analysis to newer data beyond 2019 to monitor post-pandemic sentiment.

- Integrate pricing, weather, and park event schedules for deeper correlation insights.

- Build automated alerts for sudden spikes in complaints by topic or branch.

- Implement a Machine Learning model enhanced with Deep Learning techniques which can be integrated with Tableau for better predictions and trend analysis.