**EDS 6397 Information Visualization**

**An Analytical Approach for rating and Visitor Sentiment Analysis in Disneyland Reviews**

**Group 5:**

Vengal Rao Pachava - 2156575

Amulya Pemmasani - 2211641

Mahitha Kistipati – 2147691

FNU Syed Sohaib Ali - 2209809

Fahad Mohammed Abdul - 2191130

Abdul Aziz Mohammed - 2215231

Rebecca Guradigudda - 2150227

Deekshitha Teegala -2247690

Venkata Pranay Kumar Goud Etikela - 2191851

|  |  |  |
| --- | --- | --- |
| **Name** | **PSID** | **Question** |
| Vengal Rao Pachava | 2156575 | 2,3,4,7,8 |
| Amulya Pemmasani | 2211641 | 1,3,4,5,8 |
| Mahitha Kistipati | 2147691 | 2,3,4,5,8 |
| FNU Syed Sohaib Ali | 2209809 | 1,5,6,7,8 |
| Fahad Mohammed Abdul | 2191130 | 3,5,6,7,8 |
| Abdul Aziz Mohammed | 2215231 | 1,5,6,8 |
| Rebecca Guradigudda | 2150227 | 3,4,5,8 |
| Deekshitha Teegala | 2247690 | 3,4,8 |
| Venkata Pranay Kumar Goud Etikela | 2191851 | 3,4,8 |

|  |  |  |
| --- | --- | --- |
| **Question** | **Description** | **Techniques used** |
| 1 | Data Cleaning | Python |
| 2 | Finding positive, common and negative words and their frequency | Python |
| 3 | Branch-wise visitor sentiment analysis | Bar graph |
| 4 | Temporal sentiment trends | Line chart |
| 5 | Geographical analysis of visitor sentiments | Symbol map |
| 6 | Sentiment in review text | Symbol map |
| 7 | Popular aspects and areas of improvement | Word clouds |
| 8 | Preparing report | Word document |

**Introduction**:

**Objective**:

The primary objective of this project is to conduct a comprehensive data analysis using Tableau on the Disneyland Reviews dataset. The focus is on understanding visitor sentiments and geographical patterns across three Disneyland branches: Paris, California, and Hong Kong.

The project aims to achieve the following specific objectives:  
**Sentiment Analysis**: Utilize the dataset to analyze and interpret the sentiments expressed in visitor reviews, focusing on positive and negative aspects of their experiences.  
**Rating Trends:** Explore and identify trends in ratings given by visitors across different Disneyland parks. Examine variations in ratings to uncover factors influencing visitor satisfaction or dissatisfaction.  
**Visitor Comments Analysis:** Conduct a detailed analysis of visitor comments to gain insights into specific aspects of the Disneyland experience that contribute to positive or negative sentiments.  
**Geographical Patterns:** Investigate the geographical origin of visitors and assess how sentiments and ratings vary based on the visitors' home locations. Identify any discernible patterns or differences among visitors from different regions.  
**Recommendations for Improvement:** Based on the analysis, provide recommendations for each Disneyland branch to enhance visitor satisfaction. Address specific areas of improvement and suggest strategies to capitalize on positive trends.  
  
By achieving these objectives, the project aims to offer valuable insights into the visitor experience at Disneyland parks, enabling each branch to make informed decisions to enhance overall satisfaction. The findings will contribute to a better understanding of the dynamics of visitor sentiments and geographical patterns, facilitating targeted improvements for each Disneyland branch.

**Dataset Overview**: Disneyland Reviews  
  
Source: The dataset is derived from customer reviews of Disneyland, specifically focusing on Disneyland Hong Kong. The reviews were likely collected from an online platform or customer feedback system associated with the theme park.  
  
**Scope**:  
The dataset includes reviews with various details, such as Review\_ID, Rating, Year\_Month, Reviewer\_Location, Review\_Text, and Branches. The reviews cover a range of aspects, including visitor satisfaction, experiences with rides, attractions, food, staff interactions, and overall impressions of the park.  
**Time Frame**:  
The dataset spans multiple years, with the latest entry in 2019.  
The time frame is specified at a monthly level, allowing for a temporal analysis of visitor sentiments.

**Variables**:  
Review\_ID: A unique identifier for each review.  
Rating: Numeric rating given by the reviewer (on a scale of 1 to 5).  
Year\_Month: The month and year when the review was made, providing a temporal dimension.  
Reviewer\_Location: The geographic location of the reviewer.  
Review\_Text: The text of the review, containing detailed feedback and comments.  
Branch: Specifies the Disneyland branches from Hongkong, Paris and California.  
**Exciting Facts**:  
The dataset captures reviews from visitors around the world, providing a global perspective on Disneyland Branches. Reviews include mentions of various attractions, rides, and experiences within the park. Visitors share insights into the park's strengths, such as entertainment options, food choices, and staff interactions.

**Key Observations**:  
Reviews highlight aspects like ride experiences, park size, food options, and the behaviour of staff. Some reviews express concerns about park maintenance, crowd management, and prices. Visitors from diverse locations, including Australia, Philippines, United States, India, etc., share their experiences.  
**Analysis Potential**:  
Analysis can be conducted to identify trends in visitor satisfaction over time. Geographical patterns can be explored to understand how sentiments vary based on the reviewer's location. Specific themes within reviews, such as comments on rides, food, and staff behavior, can be analyzed for insights.

**Data Cleaning:**

Data cleaning is a crucial step in the data analysis process that involves identifying and correcting errors, inconsistencies, and inaccuracies in a dataset. The primary aim is to ensure that the data is accurate, complete, and ready for analysis. It involves various steps such as handling missing values, validating the data type, Standardizing the data, Handling duplicates and outliers to get meaningful insights.

Steps taken into cleaning and pre-processing it:

* The Column Year\_Month consists of missing values to deal with we have used the mean/ mode/ median to impute the missing values as the dates are irregularly spaced.
* After examining the related "Review Text" field, duplicate items in the "Review ID" column were found and eliminated.
* For "Branch" column the prefix "Disneyland\_" was eliminated using delimiters. For uniformity “Hongkong” is changed to “Hong Kong”.

1. **Sentiment Analysis:**

We used text mining techniques to perform sentiment analysis based on the "Rating" column. We created a new column named “Sentiment” categorize reviews as positive, negative, and neutral, providing an additional dimension for analysis.

1. **Text Cleaning and Keyword Extraction:**

We performed text mining on the "Review Text" field, which involved the following steps:

* Unwanted characters were removed from the "Review Text" field.
* Stop words were Tokenized and removed from the "Review Text" field using Python to focus on meaningful words., Considered lemmatization to reduce words to their base form. Removed special characters, emojis, and symbols to facilitate text analysis.
* Then the new column was delimited, and the extracted keywords were distributed across separate columns with respect to the branches i.e., all the common words for the branch ‘California’ was assigned to the column ‘CF\_common\_words’ along with that its word frequency and top 200 positive and negatives words were also determined, similarly all the words for other branches were also extracted.

A screenshot of a computer code

Description automatically generated

The removal of duplicates, standardization of date and branch columns, sentiment analysis, and text mining provided a more refined dataset.

**Data Exploration**

Data exploration is an important stage in the data analysis process. It aims to identify the underlying patterns, trends, and possible insights in the dataset. Exploring a Disneyland reviews dataset that includes three different locations: Paris, California, and Hong Kong allow to understand visitor sentiments and preferences across different geographical environment.

To show patterns in visitor comments, we used data visualization tools to build visualizations such as geographic maps, sentiment analysis, and word clouds. By evaluating the data, patterns about popular attractions, reoccurring problems, or distinctive characteristics that contribute to the overall visitor experience in each Disneyland location may emerge.

Using Tufte's data-ink ratio approach to investigate the Disneyland reviews dataset in Paris, California, and Hong Kong entails emphasizing the representation of relevant information while eliminating unrelated graphic features. By using this technique, we constructed visualizations that convey findings without clutter, providing for a better understanding of trends between the three Disneyland locations.

**Visualizations:**

**Branch-wise visitor sentiment analysis:**

**Comparative Analysis:** Using this bar plot we compare the distribution of ratings between the three Disneyland branches.

**Axis Labeling:** We added “Rating” along the x-axis to convey the spectrum from 1 to 5. Meanwhile, the y-axis is labeled as "Number of Ratings," providing a straightforward representation of each rating level.

**Use of Color, Size, and Shape:** This dynamic visualization utilizes a color scheme—red, yellow, and green—to distinctly represent positive, neutral, and negative sentiments expressed by visitors respectively. The choice of red, green, and yellow hues serves a purpose beyond aesthetics. This color coding aids quick comprehension of sentiment distribution.

**Interactive Features:** The dropdown feature, enabling users to effortlessly switch between Disneyland locations in California, Hong Kong, and Paris. As the selected branch changes, the bar plot dynamically adapts to reflect the corresponding data. Furthermore, the dropdown functionality elevates user engagement and allows for a more focused exploration on comparison of ratings across different Branches.

**Temporal sentiment trends:**

**Comparative Analysis:** To unveil the temporal shifts in sentiment, we've crafted a dynamic line graph, aligning months along the columns and sum of ratings along the rows. This visualization effectively captures the flow of ratings throughout the year for all the three Disney Land branches, offering a comprehensive view of sentiment evolution.

**Axis Labeling:** We designated the x-axis as "Sum of Ratings" to succinctly convey the cumulative ratings for each branch throughout the month. Simultaneously, the y-axis is labeled as "Year Month," offering a clear representation of the specific month within the chosen year. This labeling strategy aims to enhance the interpretation of the data, emphasizing the aggregated ratings over time for each Disney Land branch.

**Use of Color, Size, and Shape:** Three vibrant lines, each adorned with distinctive colors, represent the sentiment trajectories of three distinct Disneyland locales. The blue line signifies California, orange embodies Hong Kong, and red characterizes Paris. This color differentiation allows viewers to quickly identify and associate each line with a specific location, reducing confusion and enhancing the overall readability of the graph.

**Interactive Features:** Our user-friendly design extends to a dropdown feature for the year as well as branch to showcase the entire chronological spectrum of the dataset. Users can effortlessly toggle between years, witnessing the nuanced sentiment dynamics over time. This interactive approach empowers users to delve into the temporal nuances across diverse locations and years. Also, this graph shows the change of ratings over time with animation.

**Geographical Analysis of Visitor Sentiments:**

**Comparative Analysis:** We leveraged symbol maps to visualize the distribution of ratings across various reviewer locations.

**Axis Labeling:** We placed branches in the y-axis so that it serves as a title for the map.

**Use of Color, Size, and Shape:** Employing a spectrum of five distinct colors in a pie chart, we have taken red to green color palette to indicate 5 ratings in the pie chart.

**Interactive Features:** We've enhanced user interactivity by implementing a dropdown feature, facilitating the selection of three distinct Disneyland locations. As users choose different options from the dropdown, the maps dynamically adjust to display the selected location, offering a seamless and personalized exploration experience.

**Sentiment in Review Text:**

**Comparative Analysis:** We employed maps to illustrate the ratings across various reviewer locations, shedding light on the sentiment distribution within these locations. Additionally, we conducted a comparative analysis across the three Disneyland locations, providing insights into how sentiments vary among these destinations.

**Axis Labeling:** We enhanced the visual presentation by incorporating the "Branch" information into the columns, aligning the title of the map with the specific name of the Disneyland branch.

**Use of Color, Size, and Shape:** We employed a gradient of light blue to represent ratings ranging from 5 to 1, providing a visual spectrum that intuitively conveys the rating scale. In addition, we utilized colored circles to depict sentiment distribution among reviewer locations: red indicating negative sentiment, yellow for neutral sentiment, and green for positive sentiment.

**Interactive Features:** We implemented a dropdown menu for seamlessly switching between different Disneyland branches.

**Popular Aspects and Areas for Improvement:**

**Comparative Analysis:** We employed word clouds to visually depict positive, neutral, and negative sentiments conveyed in the comments across all Disneyland branches.

**Axis Labeling:** We integrated word clouds for three distinct locations into a comprehensive dashboard, complemented by specific titles for each word cloud. Employing three separate dashboards, we distinguished between positive, negative, and common words, ensuring a structured and focused representation of sentiment across the locations.

**Use of Color, Size, and Shape:** We employed a diverse color palette to distinguish between words, coupled with dynamic sizing that adjusts based on the frequency of each term.

**Interactive Features:** We implemented a dropdown feature across all dashboards, streamlining the process of selectively viewing the relevant word clouds corresponding to different Disneyland branches.

**Hypotheses**

Key insights inform the examination of the Disneyland review dataset, which spans three locations: Paris, California, and Hong Kong. Notably, Australia leads the world in reviews, with the UK and the USA following closely behind. A notable increase in reviews is visible in April 2019. The investigation is motivated by a few hypotheses, which include investigating regional differences in the distribution of reviews, identifying seasonal trends—with an emphasis on the April 2019 surge—demographic influences on review content, investigating possible relationships between higher review activity and higher visitor satisfaction, and closely examining relationships between review peaks and particular events or promotions. All together, these theories offer a strong foundation for exploring the complex dynamics of consumer reviews over different geographic and temporal scales.

**Methodology**

We used a variety of charts, such as bar charts, line charts, symbol maps, and word clouds, to analyze and visualize different aspects of visitor sentiment across different dimensions.

1. **Branch-wise Visitor Sentiment Analysis:** We conducted a nuanced analysis of visitor sentiment across different branches using a bar chart for rating vs count of rating. Leveraging a calculated field, "Filter by Branch," and “Branch Parameter" we implemented a filter to refine data by 3 branches. Utilized "show parameter" for dropdown visibility, we streamlined the process. "Review with sentiment" which consists of positive, negative, and neutral rating is placed in color.

2**. Temporal Sentiment Trends:** We employed a dynamic line chart to illustrate the evolving trend of ratings over time. To distinguish between the three Disneyland branches, we assigned different colors to the lines by incorporating the "Branch" dimension into the color scheme. We introduced a dropdown showcasing all available years. This was achieved through the creation of a calculated field, "Filter by year,". "Filter by year” is shown in worksheet with the help of “Show parameter" using dropdown visibility, users can dynamically select specific years of interest. The graph also allows users to focus on selected branches. Adding "Month" to pages created an animated effect, illustrating the evolution of ratings from January to December for each year. These dynamic changes were made more accessible using "Show history," providing users with a comprehensive visual journey of ratings over time.

**3. Geographical Analysis of Visitor Sentiments:** We employed symbol maps to visually convey visitor satisfaction levels across diverse reviewer locations. In the initial map, we utilized the map as marks, keeping reviewer location into the color scheme to distinguish between countries. In the second map, we employed the pie as a mark, forming a pie chart that illustrates the distribution of ratings from 1 to 5. Rating values were assigned distinct colors, ranging from red for 1 to green for 5, enhancing visual clarity. To facilitate interactive exploration, we introduced the "Filter by branch" to filters and employed "Show parameter" for branch parameter visibility, presenting users with a dropdown for branch selection. To seamlessly integrate both maps, we implemented the "Dual axis" feature.

4. **Sentiment in Review Text:** It is a similar mapping approach as mentioned earlier, we have generated two maps. In the first map, we incorporated the count of sentiments, utilizing color to represent variations, and selected map as the mark type. In the second map, we assigned sentiment to color, where red corresponds to negative sentiment, yellow to neutral, and green to positive sentiment. To enhance user interaction, we showcased a branch dropdown. To provide clarity and context, we included the branch in the columns, effectively displaying the branch name as the map's title. The integration of a dual axis seamlessly combined these two maps, offering a comprehensive representation of sentiment distribution across different branches and locations.

5. **Popular Aspects and Areas for Improvement:** We used word clouds to visualize the common aspects and areas of improvement. As we mentioned in data cleaning, we used text mining techniques to perform word clouds. Totally we used 9 different word clouds, 3 for visualizing the common words, 3 for visualizing the positive words, 3 for visualizing the negative across three Disneyland branches. Then we created 3 dashboards to add all sheets.

To create word cloud first we need to drop common words dimension in the text label and for coloring drop common word dimension in the color label, now we have frequency measure which holds frequency of words that appeared based on common words. In mark section instead of considering automatic choose text from the drop down, now we constructed the word cloud by all this steps and same steps we follow for all the remaining word clouds.

**6. Dashboard Creation:** To make visualizations more interactive we have created a dashboard for geographical analysis. It consists of Geographical Analysis of Visitor Sentiments, Branch wise sentiments, Branch wise Ratings, Total Reviews and Avg Ratings. Hovering over one visualization triggers corresponding changes in the other four visualizations within the dashboard.

**Insights and Patterns**

**Branch-wise Visitor Sentiment Analysis using Bar Chart:**

Observing the ratings across Disneyland branches, Disneyland California exhibits a notable preference for higher ratings, particularly 5 stars. The limited occurrence of lower ratings (1 and 2 stars) implies a few areas that could benefit from minor enhancements. Disneyland Hong Kong stands out with an overwhelming majority of 5-star ratings. However, the presence of a small proportion of 1-star ratings suggests areas that may not align with certain visitors' expectations. Disneyland Paris displays a balanced spread of ratings. The relatively higher frequency of lower ratings (1 and 2 stars) compared to other parks indicates substantial opportunities for improvement. In our observations, reviewers express greater satisfaction with the California Disneyland branch when compared to all the branches.

A screenshot of a graph

Description automatically generated

**Temporal Sentiment Trends using Line Chart:**

The patterns may be indicative of various external influences such as seasonal climate variations, major holidays, or unique events hosted by the parks. This year-over-year analysis can help understand how these changes resonate with visitors, potentially influencing their overall experience and satisfaction levels.

From this Line chart we have observed that ratings of each month with respect to every year shows some similar patterns. Since July, August and December are holiday seasons, ratings in those months are more compared to other months. It shows that most of the reviewers are visiting Disneyland in the summer and winter holidays.

A graph of different colored lines

Description automatically generated

**Geographical Analysis of Visitor Sentiments using Maps:**

The visitors from specific countries or regions exhibit unique satisfaction levels or preferences. These trends could stem from cultural norms or their prior experiences with theme parks. By analyzing the geographical map, we observed that visitors to Disneylands are more from European countries. And visitors from the same regions have similar satisfaction levels.

A map of the world

Description automatically generated

**Sentiment in Review Text using Maps:**

Positive feedback often centers around certain attractions, special events, or experiences with customer service, while negative or neutral comments may point to issues such as lengthy wait times, pricing concerns, or the quality of food offerings. It's insightful to juxtapose the findings from sentiment analysis of reviews with the quantitative ratings, to discern if there's a match or mismatch between the verbal expressions of visitors and their assigned ratings.

A map of the world

Description automatically generated

**Popular Aspects and Areas for Improvement using Word cloud:**

The most used positive and negative words and phrases in the feedback often relate to attractions, performances, interactions with staff, or the general ambiance of the park. Focus especially on the repeated negative themes, as they serve as clear indicators of the aspect’s visitors believe require prompt attention and enhancement.

Common words across three branches: Disneyland, park, ride, time, adventure, ticket, place, etc., by see these few words, we can assume that people are more speaking about Disneyland Park and the rides available in the park and how adventure those rides and ticket prices for those rides.

Positive words across three branches: Food, stay, fireworks, fast, family, restaurant and photos, etc., by see these few words, we can assume that people had given positive reviews about the food provided across all three branches and accommodation provided and their services are fast, and these places are good for family accommodation and all these places are good for photo.

Negative words across three branches: disappointed, haunted, staff, expensive, crowded, queuing, etc., by seeing these few words, we can assume that people had negative opinion related to staff and they are disappointed with the services provides and some people thought that it is very expensive and some people had bad opinion related to handling the group of people and some are scared about these kind of places.

A group of text on a white background

Description automatically generated

A close up of text

Description automatically generated

A close up of text

Description automatically generated

To make visualizations interactive we have used dashboards and merged some individual plots and combinations 2 or more potential direction points. We can use branch drop to see the dynamic changes that will take place in all plots parallelly for the dashboard with Ratings and Sentiment analysis across branches. In another dashboard we can hover on any plot which shows the plots of others changing dynamically.

Below Visualizations shows interactive dashboards:

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Ethical Considerations**

* We maintained the accurate representation without any misleading in the scales or data.
* Our visualizations provide clear labelling, context, and less clutters.
* Adherence to privacy laws, user empowerment, peer reviews, and continuous monitoring further contribute to responsible visualization practices.

**User Experience and Engagement**

Colin Ware's Engagement concept highlights the importance of developing visually appealing and approachable designs to improve user friendliness.

We made use of different font sizes to emphasize important information while maintaining consistency. Organized visual elements logically and methodically. Align text, charts, and labels for better readability. The color we choose for every visualization is harmonious and enhances the text.

We included some interactive elements like dropdowns, parameters, calculated fields and added some filters and pages to make the visualizations more user friendly. Dashboards are created to merge distinct visualizations and it makes users more engaging towards the visuals.

**Conclusion**

**Overview of Findings**

The analysis of Disneyland Reviews using Tableau revealed distinct patterns in visitor sentiments across the Disneyland parks in California, Hong Kong, and Paris. Positive sentiments were frequently associated with entertainment options, staff interactions, and food choices. In contrast, negative sentiments focused on park maintenance, crowd management, and pricing issues. This dichotomy highlights areas of strength and opportunities for improvement within the parks.

**Rating and Sentiment Trends**

Visitor ratings varied across the different Disneyland locations, with Paris generally receiving the highest percentage of positive reviews. The fluctuation in ratings over the years, peaking in 2018 and dipping in 2012, suggests the influence of various operational and external factors on visitor satisfaction. This trend underscores the dynamic nature of visitor experiences and the need for ongoing monitoring and adaptation.

**Geographical Insights**

The analysis also uncovered geographical variations in visitor sentiments, indicating that visitor experiences and expectations differ based on their region of origin. This finding is crucial for tailoring marketing strategies and enhancing visitor experiences, ensuring that the parks meet the diverse needs and expectations of their global audience.

**Strategic Implications**

The insights from this data analysis provide Disneyland with a roadmap for operational improvements and strategic decision-making. By focusing on areas highlighted by visitor feedback, Disneyland can enhance overall visitor satisfaction and tailor experiences to meet the evolving needs of their diverse visitor base. This approach not only addresses immediate concerns but also sets the stage for sustained growth and visitor engagement.