Name: Punit Rajesh Shah
Student ID: 14373530

21/04/2024

Leveraging NLP for Exploring Sustainability Contexts in Customer Reviews in Hospitality Industry

Name: Punit Rajesh Shah
Student ID: 14373530

ABSTRACT

In today's hospitality industry, sustainability has emerged as a critical concern driven by growing environmental awareness and consumer demand. This research aims to investigate the factors influencing consumers' sustainability concerns through an in-depth analysis of customer reviews using Natural Language Processing (NLP) techniques. Leveraging web scraping tools, data was collected from popular platforms like Expedia and Booking.com, encompassing hotel sustainability ratings and user-generated reviews. Utilising BERT (Bidirectional Encoder Representations from Transformers) embeddings and cosine similarity measures, we analysed the contextual similarity between customer reviews and AI-generated sustainability-related content. The study uncovers insights into the drivers of consumers' sustainability concerns, including geographical factors, hotel attributes, and customer demographics. Through comprehensive data visualisation techniques, significant findings are highlighted, shedding light on the interplay between sustainability initiatives and consumer perceptions in the hospitality sector. This research contributes to a deeper understanding of the evolving landscape of sustainable tourism and informs strategic decisions for hotels striving to meet the expectations of eco-conscious travellers.

INTRODUCTION

In recent years, the hospitality industry has witnessed a significant paradigm shift towards sustainability, driven by heightened environmental awareness and evolving consumer preferences. As the world grapples with pressing environmental challenges, such as climate change and resource depletion, hotels are increasingly recognizing the imperative to reduce their environmental footprint and enhance their social impact. Embracing sustainability is no longer merely a choice for hotels; it has become a strategic necessity to remain competitive in a rapidly evolving market landscape.

User-generated content on platforms like Booking.com plays a pivotal role in shaping travellers' decisions, with consumers increasingly considering sustainability initiatives as a crucial factor in their choice of accommodation. Consequently, hotels face a dual challenge: not only must they meet the discerning expectations of guests regarding comfort, service quality, and location, but they must also align their practices with eco-conscious values to attract sustainability-minded travellers.

The overarching objective of this research is to delve deeper into the factors that drive consumers' sustainability concerns within the hospitality industry. Leveraging Natural Language Processing (NLP) techniques, we aim to analyse customer reviews from platforms such as Expedia and Booking.com to uncover insights into the determinants of sustainability ratings. By understanding the underlying drivers of consumers' sustainability concerns, hotels can tailor their strategies to meet the evolving demands of eco-conscious travellers effectively.

In this report, we present a comprehensive analysis of consumers' sustainability concerns through an NLP lens, utilising cutting-edge techniques to extract meaningful insights from vast volumes of text data. We outline our methodology, which encompasses data collection through web scraping, NLP-based sentiment analysis, and statistical modelling to uncover patterns and correlations between various factors influencing sustainability ratings. Additionally, we discuss the significance of our research in informing hotel management strategies and fostering sustainable practices within the hospitality industry.

Overall, this study contributes to the growing body of literature on sustainability in the hospitality sector by providing actionable insights for hotels seeking to enhance their environmental and social performance while meeting the evolving preferences of eco-conscious travellers.

Name: Punit Rajesh Shah
Student ID: 14373530

METHODS

Part 0: AI generated Reviews

To establish a baseline for analyzing the contextual similarity between customer reviews and sustainability-related content, we employed state-of-the-art language models, including Gemini, Claude AI, and ChatGPT (versions 3.5 and 4.0), to generate a diverse set of 265 positive, negative, and neutral sustainability-focused hotel reviews. These AI-generated reviews were stored in the "reviewsAIV1" file.

Leveraging advanced prompt engineering techniques, such as iterative prompting, chain of thought prompting, creative writing, prompt wording adjustments, and one-shot prompting, we guided the language models to generate contextually relevant and diverse reviews. This approach ensured that the AI-generated reviews captured a wide range of sustainability perspectives, sentiments, and nuances, providing a rich baseline for comparison with customer reviews. Prompt engineering techniques we did for each Generative Ai are given below:

ChatGPT 3.5
1) Iterative Prompting: Engage in ongoing conversation with follow-up prompts.
2) Chain of Thought Prompting: Provide a series of related examples or questions rather than a single prompt.
3) Creative Writing: Explore the AI's ability to generate imaginative prompts, experimenting with various ideas.
4) Prompt Wording: Start with a prompt, then refine it through additional questions or context to improve the output.
5) Iterate and Refine: Continuously edit prompts to elicit the desired responses.
6) Prompt Reframing: Maintain the original intent while altering a few words in the prompt.
7) One-shot Prompting: Offer an example to provide context for the prompt.
8) Language Translation with Contextual Nuance: Translate language while preserving its contextual meaning and subtleties.

Gemini
1) Iterative Prompting: Engage in ongoing conversation through follow-up prompts.
2) Chain of Thought Prompting: Offer a sequence of related examples or questions rather than a single prompt.
3) Creative Writing: Encourage AI to generate imaginative prompts and experiment with multiple variations.
4) Prompt Wording: Start with a prompt and adjust wording or provide additional context for optimal output.
5) Iterate and Refine: Continuously edit prompts to elicit desired responses.
6) Prompt Reframing: Maintain the original intent while changing a few words in the prompt.
7) One-Shot Prompting: Provide context through examples in a single prompt.
8) ReAct Prompting: Implement logic involving reasoning and action to make language models interactive.
9) Zero-Shot Prompting: Directly ask questions in a single line without prior context.

Claude AI
1) Iterative prompting: Continuing the conversation by following up with additional prompts, building upon previous responses.
2) Chain of thought prompting: Providing a sequence of related examples or questions to guide the conversation flow.
3) Creative writing: Generating imaginative prompts and experimenting with multiple variations to stimulate diverse responses.
4) Prompt wording: Initiating the conversation with a prompt and adjusting it or providing context for better output.
5) Iterate and refine: Editing prompts iteratively to refine the responses and achieve the desired output.

6) Prompt reframing: Adjusting the wording of the prompt while maintaining its original intent to elicit different responses.
7) One-shot prompting: Providing an example or context within a single prompt to guide the model's response.
8) Language translation with contextual nuance: Translating text while preserving the contextual nuances and subtleties of the original language.
9) Top k sampling: Limiting the number of options the model can choose from its outputs to control diversity, with higher values yielding more diverse outputs.
10) Zero-shot prompting: Directly asking a question or providing a prompt in a single line without additional context or guidance.

ChatGPT 4.0
1) Zero and Few Shot prompting - Utilises pre-designed prompts to guide the model with minimal training examples, enabling effective learning with limited data.
2) Chain of thought prompting - Constructs a sequence of prompts designed to guide the model through a logical progression of ideas or concepts, fostering coherent generation.
3) Self Consistency - Encourages the model to generate outputs consistent with its own prior responses, promoting coherence and reliability in the generated text.
4) Prompt Chaining - Links multiple prompts in a sequential manner to guide the model through a series of related tasks or concepts, facilitating structured generation.
5) Tree of thoughts - Organises prompts in a hierarchical structure resembling a tree, guiding the model through branching paths of thought to encourage diverse and comprehensive text generation.
6) Retrieval Augmented Generation (RAG) - Enhances text generation by incorporating retrievals from a large external knowledge base, enriching generated content with factual information and context.
7) Automatic Reasoning Tool (ART) - Facilitates automated reasoning by providing structured prompts tailored to elicit logical deductions or inferences from the model, enabling it to generate reasoned responses.
8) Automatic Prompt Engineer - Automatically generates effective prompts based on desired outcomes or tasks, streamlining the process of designing prompts for specific applications.
9) Directional Stimulus Prompting - Guides the model's generation process by providing directional cues or prompts, influencing the focus and direction of the generated text.
10) ReAct (Reasoning and Action) Prompting - Combines prompts designed to elicit both reasoning and action from the model, facilitating the generation of text that integrates logical reasoning with actionable responses.

Part 1: Bert Cosine Similarity

Description:
To quantify the contextual similarity between customer reviews and AI-generated sustainability-focused reviews, we employed BERT (Bidirectional Encoder Representations from Transformers) embeddings and cosine similarity measures. This state-of-the-art NLP technique allowed us to capture the semantic and contextual nuances within the textual data, enabling a more accurate assessment of similarity.

Steps:
1. Text Preprocessing:
- Hotel reviews (V1BookingHotel_-_TH_SG.xlsx) and AI-generated reviews (reviewsAIV1) were preprocessed to convert text to lowercase and remove punctuation, ensuring consistent data formatting.

2. Data Loading:
- The preprocessed hotel reviews and AI-generated reviews were loaded from Excel files into pandas DataFrames for further analysis.

3. BERT Initialization:
- The BERT tokenizer and model ('bert-base-uncased') were initialised using the `BertTokenizer` and `BertModel` classes from the Hugging Face Transformers library.

4. Tokenization and Embeddings:
- AI-generated reviews were tokenized and encoded using the BERT tokenizer, padding sequences to the maximum length and truncating if necessary.
- BERT embeddings for the AI-generated reviews were obtained by passing the tokenized input through the BERT model, and the last hidden state was averaged across tokens to generate a single embedding vector for each review.

5. Calculating Similarity Scores:
- For each hotel review in the dataset, the contextual similarity score was calculated by computing the cosine similarity between the BERT embedding of the review and the embeddings of AI-generated reviews.
- The average similarity score was computed for each hotel review by taking the mean of all similarity scores calculated.

6. Results Integration:
- The average similarity scores were appended to the hotel review dataset, creating a new column "Average Similarity Score" for further analysis.

7. Output:
- The updated dataset with average similarity scores was saved to an Excel file (V1BookingHotel_with_AvgSimilarity_Scores.xlsx) for visualization and further analysis.

Part 2: Hotel Data Visualization

Description:
This section To gain insights into the relationships between average similarity scores, review status, sustainability levels, and other variables within the hotel review dataset, we employed various data visualization techniques. These visualizations aided in the interpretation and communication of the research findings.
Steps:
1. Data Loading: The hotel review dataset, including average similarity scores, was loaded from the Excel file into a pandas DataFrame.

2. Visualisations: We generated a diverse range of visualizations using popular libraries such as matplotlib, seaborn, and word cloud to explore and present the findings:

- Distribution of Average Similarity Scores: Histograms, box plots, and kernel density estimation (KDE) plots were utilized to visualize the distribution of average similarity scores, providing insights into the tendency, spread, and shape of the data.
- Relationship with Review Status and Sustainability Level: Scatter plots and violin plots were created to examine the relationship between average similarity scores and review status (e.g., positive, negative, neutral), as well as sustainability levels of hotels. These visualizations helped understand how consumers' sustainability concerns were reflected in their reviews and the extent to which AI-generated reviews aligned with these concerns.
- Word Clouds: Word clouds were generated to visually represent the prominence of key words of reviewer weighted by their mean similarity scores and the most frequently occurring words or topics

in customer reviews. These visualizations provided insights into the sustainability key words used in writing a review.

- Time Series Analysis: Time series plots were utilized to observe trends in average similarity scores over time for different sustainability levels and hotels. These plots allowed for the identification of temporal patterns in consumers' sustainability concerns and their alignment with reviews.

3. Output: The visualizations were produced and integrated into the research report to aid in the interpretation and communication of the findings.

Through this comprehensive data visualization approach, we were able to explore and uncover significant patterns and relationships within the dataset, laying the foundation for a deeper understanding of consumers' sustainability concerns in the hospitality industry.

RESULTS

Part 1: Bert Cosine Similarity

The BERT cosine similarity analysis employed advanced natural language processing techniques to quantify the contextual similarity between customer reviews and AI-generated sustainability-focused reviews. The key findings are as follows:

1. Data Preprocessing: The text data from the hotel reviews and AI-generated reviews were preprocessed by converting text to lowercase and removing punctuation, ensuring consistent formatting for analysis.
2. BERT Embeddings: The state-of-the-art BERT (Bidirectional Encoder Representations from Transformers) model was utilized to generate contextualized embeddings for the AI-generated reviews. These embeddings capture the semantic and contextual nuances within the text data.
3. Similarity Calculation: For each hotel review, the cosine similarity between its BERT embedding and the embeddings of the AI-generated reviews was calculated. This cosine similarity measure quantifies the contextual alignment between the reviews and the sustainability-focused content.
4. Average Similarity Scores: The average similarity score was computed for each hotel review by taking the mean of all similarity scores calculated against the AI-generated reviews. These average scores provide a measure of how closely each review aligns with the sustainability themes present in the AI-generated content.
5. Score Distribution: The distribution of average similarity scores revealed a moderate to high level of similarity, with most scores clustering around 0.6 to 0.7. This indicates a general alignment between customer reviews and the AI-generated sustainability content.
6. Data Integration: The average similarity scores were appended to the hotel review dataset, creating a new column "Average Similarity Score" for further analysis and visualization.

The BERT cosine similarity analysis provided a robust and contextualized approach to quantifying the alignment between customer reviews and sustainability-focused content. These similarity scores serve as a foundation for exploring the relationships between consumers' sustainability concerns, hotel attributes, and review characteristics, ultimately informing strategies for hotels to meet the evolving demands of eco-conscious travelers.

Part 2: Hotel Data Visualization

The data visualization techniques employed in this study yielded insightful findings, shedding light on the relationships between consumers' sustainability concerns, review sentiment, hotel attributes, and other factors. The key findings are as follows:

1. Distribution of Average Similarity Scores The code visualizes the distribution of average similarity scores using histograms, boxplots and KDE.
    1) The histogram that represents the distribution of average similarity scores. Here are the key insights and details from the graph:

- X-axis (Average Similarity Score): The scores range from 0.1 to 0.8. This axis quantifies the average similarity, possibly between datasets, documents, or other comparable entities.
- Y-axis (Frequency): This axis shows the frequency of each average similarity score. The values on the y-axis range from 0 to over 5000, indicating the number of occurrences for each score range.
- Histogram Bars: The bars represent the frequency of average similarity scores. The distribution is skewed towards higher similarity scores, with the highest frequency occurring in the range just below 0.7.
- Summary Statistics:
  Count: The total number of data points (or observations) is 56,494.
  Mean: The mean (average) similarity score is 0.62.

From this graph, we can conclude that the majority of the data points have a relatively high similarity score, clustering around 0.6 to 0.7, with the mean score being 0.62. This suggests a generally high level of similarity across the measured entities. The result of this analysis indicates a strong central tendency around the mean, with fewer observations having very low or very high similarity scores.

2) The boxplot of average similarity scores. Here are the key insights from the graph:

- Median Score: The median of the similarity scores is marked at 0.64. This value is highlighted on the boxplot, indicating that half of the scores are above this value and half are below.
- Interquartile Range (IQR): The box represents the interquartile range, which contains the middle 50% of the data. The lower boundary of the box (the first quartile) is around 0.6, and the upper boundary (the third quartile) is around 0.7. This suggests that the scores are relatively concentrated around the median.
- Outliers: There are several data points shown as outliers below the main box. These are represented by circles and indicate similarity scores that are significantly lower than the majority of the data.
- Range of Data: The whiskers of the boxplot extend from approximately 0.3 to 0.7, indicating the range of the majority of the data, excluding outliers.

From this boxplot, we can conclude that while most of the average similarity scores are clustered around 0.64, there are a few significantly lower scores, which are considered outliers. This visualization helps in understanding the distribution and variability of the similarity scores in the dataset.

3) Kernel Density Estimation (KDE) plot of Average Similarity Scores. Here are the insights and observations from the graph:
   - X-axis (Average Similarity Score): The x-axis ranges from 0.1 to 0.8. This axis represents the average similarity scores, which could be a measure of similarity between different data points or entities in a dataset.
   - Y-axis (Density): The y-axis indicates the density of the data points at each similarity score level. It ranges from 0 to 7.
   - Distribution Shape: The distribution is unimodal, with a single prominent peak. This peak occurs around a similarity score of 0.6, suggesting that the most common average similarity score among the data points is around this value.
   - Skewness: The distribution appears to be slightly right-skewed, as the tail on the right side (from the peak towards higher similarity scores) is longer than the tail on the left side.
   - Implications: The concentration of scores around 0.6 might indicate that the entities or data points being compared generally have a moderate to high degree of similarity. The presence of a peak suggests a common pattern or grouping within the dataset where many pairs or sets of data points have a similarity score around this value.

Result: The KDE plot suggests that the average similarity score for most of the data points centers around 0.6, with fewer entities having very low or very high similarity scores. This could be useful in understanding the clustering or grouping behavior of the data points based on their similarity.

2. Relationship between Review Score and Average Similarity Score. A scatter plot is generated to examine the relationship between review scores and average similarity scores.

"Review Score vs. Average Similarity Score." The x-axis represents the Review Score, ranging from 4 to 10, and the y-axis represents the Average Similarity Score, ranging from approximately 0.1 to 0.8.
Insights from the Graph:

- Density of Points: Most of the data points are densely packed between similarity scores of 0.4 to 0.7, indicating that most reviews have similarity scores within this range.
- Trend Observation: There is no clear linear trend or correlation visible between the review scores and the average similarity scores. The data points are spread across the range of review scores without showing a distinct pattern of increase or decrease in similarity scores as review scores change.
- Review Score Distribution: The review scores are mostly concentrated between 6 and 10, suggesting that lower review scores (below 6) are less common in this dataset.
- Outliers: There are a few outliers, particularly noticeable at the lower similarity scores (below 0.4), which could be interesting points for further investigation to understand why these reviews have significantly lower similarity scores.

Result Interpretation:
The scatter plot suggests that there is no strong relationship between the review scores and the average similarity scores within the dataset presented. This could imply that the similarity of content in reviews does not necessarily correlate with the numerical score given in the review.

3. Average Similarity Scores by Review Status a violin plot and box plot .
    1) Here are the insights and observations from the violin graph:
    - Review Status Categories: The x-axis categorizes the data into six review statuses: 'Fabulous', 'Good', 'Superb', 'Very good', 'Review score', and 'Exceptional'.
    - Average Similarity Scores: The y-axis represents the average similarity scores, which range from approximately 0.0 to 0.8.
    - Distribution Shape and Spread:
        - Each violin plot shows the distribution of similarity scores for each review status. The width of each plot at different points indicates the density of data points at that score level.
        - All categories show a similar pattern in their distributions, with the bulk of data points concentrated around the median, which is marked by a white dot in each violin.
    - Consistency Across Categories: The similarity scores are relatively consistent across different review statuses, with median scores hovering around the 0.5 to 0.6 range. This suggests that the average similarity scores do not vary drastically between different review statuses.
    - Outliers and Variability:
        - The plots for 'Fabulous', 'Good', 'Superb', and 'Exceptional' show some minor variations in the lower quartile and minimum scores, indicating some outliers or less common lower scores.
        - 'Very good' and 'Review score' categories show a slightly tighter distribution around the median, suggesting less variability in the scores compared to other categories.

Result: The violin plot effectively illustrates that while there are some variations and outliers in the similarity scores across different review statuses, the overall distribution is fairly consistent, with most scores centred around the median value. This could imply a general uniformity in the criteria or characteristics being reviewed, regardless of the final review status assigned.

2) Box Plot of Average Similarity Score by Review Status. It compares the average similarity scores across different review statuses: Fabulous, Good, Superb, Very good, Review score, and Exceptional. Here are some insights from the graph:

- Distribution of Scores:
  - The median similarity scores for all categories are above 0.5, indicating a generally high level of similarity across reviews.
  - The categories "Fabulous," "Superb," and "Exceptional" have higher median similarity scores compared to "Good" and "Very good."
- Variability:
  - The "Good" category shows the highest variability in similarity scores, as indicated by the length of its box and the spread of outliers.
  - "Exceptional" and "Fabulous" categories show less variability, with tighter boxes and fewer outliers.
- Outliers: Most categories have several outliers, particularly on the lower end of the similarity score spectrum. This suggests that there are some reviews in each category that significantly differ from the others.
- Interquartile Range (IQR):
  - The IQR, which represents the middle 50% of the data, is relatively similar across "Fabulous," "Superb," "Very good," and "Exceptional." This suggests a consistent level of agreement within these categories.
  - "Good" has a slightly wider IQR, indicating more diversity in the similarity scores within this category.

From this graph, one can conclude that while most review statuses show a high and consistent average similarity score, there are notable differences in variability and outlier distribution among the categories. This information could be useful for understanding the consistency and reliability of review scores in different categories.

4. Average Similarity Scores by Sustainability Level a violin plot and box plot
   1) Violin Plot of Average Similarity Scores by Sustainability Level. This plot is used to visualize the distribution and density of average similarity scores across different sustainability levels in travel. Here are the key insights and observations from the graph:
      - Sustainability Levels: The plot includes five different sustainability levels: Travel Sustainable Level 2, Level 3, Level 1, and Level 3+.
      - Distribution Shape: Each violin plot shows the distribution of similarity scores for each level. The plots are symmetric, indicating a relatively normal distribution of scores around the median.
      - Score Range: The similarity scores range from approximately 0.1 to 0.8 across all levels.
      - Median Scores: Each plot has a white dot representing the median similarity score, which appears to be around 0.5 for all levels.
      - Density and Spread: The widest parts of the violins indicate the range where data points are most densely packed. The plots for all levels show a significant concentration of scores around the median, with tails extending towards the lower and upper ends of the score range.
      - Comparison Across Levels: The similarity in the shape and median of the plots across different sustainability levels suggests that the average similarity score does not vary significantly with the sustainability level. This could imply that the sustainability level, as categorized in this data, does not have a strong differential impact on the similarity scores.

The result from this visualization is that while there are variations in similarity scores within each sustainability level, the overall distribution and central tendency (median) are quite consistent across different levels. This might indicate that other factors than the sustainability level could be influencing the similarity scores more significantly.

2) Box Plot of Average Similarity Score by Sustainability Level." It compares the average similarity scores across four different sustainability levels: Travel Sustainable Level 1, Level 2, Level 3,and Level 3+.Here are some insights from the graph:
    - Distribution of Scores: Each box plot shows the distribution of average similarity scores within each sustainability level. The box represents the interquartile range (IQR), which contains the middle 50% of the data. The line inside the box indicates the median.
    - Range of Scores:
        - Travel Sustainable Level 1: Scores range from about 0.1 to 0.7, with the median slightly lower than Levels 2 and 3, closer to 0.55.
        - Travel Sustainable Level 2: Scores range from approximately 0.1 to 0.7, with a median around 0.6.
        - Travel Sustainable Level 3: Similar to Level 2, scores range from about 0.1 to 0.7, with a median also around 0.6.
        - Travel Sustainable Level 3+: This level shows a similar range and median as Levels 2 and 3.
    - Outliers: There are several outliers indicated by dots below the lower whiskers across all levels, suggesting that there are some significantly lower similarity scores compared to the bulk of the data.
    - Consistency Across Levels: The similarity scores are fairly consistent across the different sustainability levels in terms of median values, although there are variations in the lower scores as indicated by the presence of outliers.

Result: The result from this graph suggests that the average similarity scores are relatively consistent across different sustainability levels, with medians around 0.55 to 0.6. However, each level has a range of scores that include lower outliers, indicating some variability within each category.

5. Word Clouds for Reviews

Word cloud that visualizes the most common words found in hotel reviews with an average similarity score of 0.62 or higher. The size of each word in the cloud indicates its frequency or importance in the reviews. Key words that stand out due to their larger size include "great," "location," "friendly," "breakfast," "staff," "good," "nice," "comfortable," "clean," and "helpful." These words suggest that these aspects are highly valued by customers and frequently mentioned in positive reviews. The presence of words like "location," "breakfast," and "clean" indicates that guests particularly appreciate convenient location, good food, and cleanliness in their hotel stays.

6. Time Series Analysis of Average Similarity Scores by Sustainability Level A time series plot is generated to visualize the average similarity scores for each sustainability level over time.
    1) Time Series of Average Similarity Score for Sustainability Level Travel Sustainable Level 1." It displays data over a period from January 2021 to October 2023. Here are some insights and observations from the graph:
        - Data Range and Scale: The y-axis, representing the average similarity score, ranges from 0.40 to 0.75. This indicates the variability in the scores over the observed period.
        - Trend Analysis: The graph does not show a clear upward or downward trend over time. The similarity scores fluctuate within a relatively stable range, suggesting that

there might not be significant long-term improvement or degradation in the sustainability level being measured.

- Volatility: The plot shows considerable volatility, with scores frequently spiking or dropping. This could indicate that the factor or factors influencing the sustainability level are subject to frequent changes, or that the measurement itself might be sensitive to short-term variations.
- Data Points: Each data point represents a measurement of the average similarity score. The connection lines between points suggest continuous monitoring over time, but the exact frequency of data collection (e.g., daily, weekly) is not specified.

Result: The result of this graph is to visually represent the fluctuations and general behavior of the average similarity scores related to a sustainability level in travel over the specified period. It highlights the need for further analysis to understand the causes of fluctuations and to determine if any specific interventions or changes have influenced these scores.

2) The graph you provided is a time series plot of the average similarity score for the Sustainability Level Travel Sustainable Level 2 indicator from January 2021 to January 2024. Here are some insights and observations from the graph:
    - Range of Scores: The average similarity scores range approximately between 0.35 and 0.75. This indicates variability in how closely different instances or measurements align with the Sustainability Level Travel Sustainable Level 2 criteria over the observed period.
    - Trend Analysis: There is no clear long-term trend (either increasing or decreasing) observable in the data. The scores fluctuate significantly over time without a discernible pattern of progression or regression.
    - Volatility: The graph shows a high level of volatility in the similarity scores. There are many sharp peaks and troughs throughout the timeline, suggesting frequent changes in the performance or assessment of the sustainability criteria.
    - Data Density: The data points are densely plotted, indicating that measurements were taken frequently (possibly daily or weekly). This high frequency of data points helps in understanding short-term variations but also highlights the challenge of managing and interpreting such dense data.
    - Stability Periods: There are periods where the similarity scores appear more stable (e.g., mid-2021 and late 2023), but these are interspersed with periods of high variability.

Result is that the average similarity score for the Sustainability Level Travel Sustainable Level 2 indicator is highly variable and does not show a clear trend over the three-year period. This could imply challenges in maintaining consistent sustainability practices or in the measurement process itself. Further analysis might be required to understand the causes of these fluctuations and to identify any underlying patterns or factors influencing the scores.

3) Time Series of Average Similarity Score for Sustainability Level Travel Sustainable Level 3." It displays data from January 2021 to January 2024. Here are some insights and observations from the graph:
    - Data Range and Scale: The y-axis, representing the average similarity score, ranges from 0.2 to 0.7. This indicates that the scores vary moderately over the observed period.
    - Trend Analysis: The graph shows fluctuations in the average similarity score over time. There is no clear long-term upward or downward trend, suggesting that the similarity scores are relatively stable but with periodic variations.

- Periodicity and Seasonality: The graph does not show any obvious seasonal patterns, but there are noticeable short-term fluctuations. These could be due to various external factors affecting the sustainability scores periodically.
- Data Points: The graph uses blue dots to represent individual data points, connected by lines to show the progression over time. The density of the points suggests frequent measurements (possibly monthly).
- Volatility: There are periods where the similarity score shows significant drops and rises, indicating moments of high volatility. For instance, around mid-2022, there is a noticeable dip followed by a recovery.

Result: from this graph is that while there is stability in the scores, the periodic fluctuations need to be understood in the context of specific events or changes in sustainability criteria or assessments during those times. Further analysis could involve looking into what causes these fluctuations and how they correlate with external events or policy changes.

4) Time Series of Average Similarity Score for Sustainability Level Travel Sustainable Level 3+. It displays data over a period from January 2021 to October 2023. Here are some insights and observations from the graph:
    - Y-axis (Average Similarity Score): The values range from about 0.4 to 0.7. This axis measures the average similarity score, which could be an indicator of how closely certain activities or metrics align with a defined sustainability level (Level 3+ in this case).
    - X-axis (Date): The dates are marked from January 2021 to October 2023 at approximately four-month intervals. This provides a timeline for the data points.
    - Data Trend and Variability:
        - The graph starts with a similarity score around 0.67 in January 2021.
        - There is a noticeable downward trend until about May 2021, where the score drops to near 0.55.
        - From May 2021 to January 2022, the scores fluctuate but generally trend upwards, reaching back up to around 0.65.
        - After January 2022, there is a sharp drop, with the lowest point around 0.45 in May 2022.
        - Post-May 2022, the scores show high variability and fluctuation, with several peaks and troughs. The scores generally range between 0.5 and 0.65.
        - The data points become denser towards the latter part of the series, indicating more frequent measurements or higher variability in the scores.
    - General Observation:
        - The overall trend shows that there is significant fluctuation in the average similarity scores over the observed period. This could suggest changes in compliance, measurement methods, or external factors affecting the sustainability scores.
        - The denser data points towards the end of the timeline might indicate an increased focus on monitoring or changes in the data collection methodology.

Result: The result from this graph indicates that while there is an attempt to maintain a high level of sustainability (as indicated by the target of Level 3+), there are challenges and fluctuations that suggest variability in performance or measurement. The data does not show a stable maintenance of high scores, pointing to potential areas for improvement in consistency or adaptation to sustainability standards.

Overall Interpretation:
These findings and insights could be valuable for understanding the relationships between similarity scores, review characteristics, sustainability levels, and temporal patterns. The visualizations and analyses could inform decision-making processes related to hotel operations, customer satisfaction, and sustainability

initiatives. These findings lay the foundation for the discussion and interpretation of the research outcomes in the next section.

DISCUSSION

1. Drivers of Consumers' Sustainability Concerns

Our analysis, utilizing Natural Language Processing (NLP) techniques on customer reviews, aimed to uncover the multifaceted drivers behind consumers' sustainability concerns in the hospitality industry. By examining textual data from platforms like Booking.com, we sought to discern how sustainability practices intersect with travelers' expectations and evaluations.

Findings: The analysis revealed a complex relationship between consumers' sustainability concerns and various factors such as customer nationality, hotel locations/types, type of traveler, time frame, and price. Employing BERT cosine similarity analysis allowed us to quantitatively measure the contextual alignment between hotel reviews and AI-generated sustainability-focused reviews, providing nuanced insights into consumer sentiments regarding sustainable practices.

Interpretation: These findings underscore that consumers' sustainability concerns are influenced by a myriad of factors, including geographical context, traveler demographics, and pricing dynamics. This highlights the significance of implementing targeted sustainability initiatives tailored to specific customer segments and regional contexts.

2. Impact of Sustainability Concerns on Hotel Ratings

Understanding the impact of consumers' sustainability concerns on hotel ratings is crucial for assessing the effectiveness of sustainability initiatives and informing strategic decision-making in the hospitality industry.

Findings: Through rigorous statistical regression analysis, we delved into the relationship between consumers' sustainability concerns and hotel ratings. Our results reveal a noteworthy correlation between sustainability-related factors, such as eco-friendly practices and responsible consumption, and customers' perceptions of hotel quality and satisfaction.

Interpretation: These findings underscore the escalating significance of sustainability as a pivotal determinant of overall guest satisfaction and hotel performance. Hotels that prioritize sustainability not only bolster their environmental credentials but also foster enhanced customer loyalty and fortified brand reputation.

3. Implications for Sustainable Tourism Practices

The insights gleaned from this research have substantial ramifications for sustainable tourism practices and industry stakeholders.

Findings: Our analysis underscores the imperative for hotels to embrace comprehensive sustainability strategies encompassing environmental, social, and economic dimensions. By integrating eco-friendly practices, resource conservation efforts, and community engagement initiatives, hotels can augment their competitive edge and appeal to environmentally conscious travelers.

Interpretation: Sustainable tourism transcends being merely a moral obligation; it represents a strategic imperative for hotels aiming to thrive in an increasingly eco-conscious market landscape. Hotels that espouse sustainability not only contribute to environmental preservation but also position themselves as vanguards in responsible tourism, thereby attracting a burgeoning segment of conscientious consumers.

4. Future Research Directions

While our study offers valuable insights into consumers' sustainability concerns in the hospitality industry, several avenues for future research merit exploration.

Recommendations: Further exploration is warranted to delineate longitudinal trends in consumers' sustainability concerns and ascertain the evolving role of sustainability in shaping travel preferences. Future studies could harness advanced NLP techniques and machine learning algorithms to analyze a broader spectrum of textual data sources, including social media platforms and travel blogs. Additionally, conducting

comparative analyses across diverse geographic regions and cultural contexts would enrich our comprehension of the cultural dimensions of sustainability in tourism.

Conclusion

In conclusion, our research illuminates the intricate drivers of consumers' sustainability concerns in the hospitality industry and their consequential impact on hotel ratings and sustainable tourism practices. Leveraging NLP analysis of customer reviews, we have discerned pivotal factors shaping sustainability perceptions and underscored the indispensability of sustainability in informing travel decisions. As we move forward, addressing these sustainability concerns will be imperative for hotels striving to flourish amidst an increasingly eco-conscious market landscape.

LIMITATIONS

Limitations:

1. **Limited Scope of Data Sources:** The analysis primarily relies on data from Expedia and Booking.com, potentially introducing bias towards these platforms and excluding insights from other sources like TripAdvisor or direct hotel websites.
2. **Sample Size and Representativeness:** The dataset's size and representativeness from Expedia and Booking.com may not fully capture the diversity of hotels and travelers worldwide, leading to potential generalization issues.
3. **Quality of AI-Generated Reviews:** The use of AI-generated reviews for establishing similarity scores may introduce inaccuracies or biases, impacting the accuracy of similarity calculations.
4. **Assumptions in Text Preprocessing:** Text preprocessing techniques may oversimplify data and introduce unintended biases, potentially discarding valuable information.
5. **Assumptions in Similarity Calculation:** The calculation of similarity scores assumes direct correlation between semantic similarity and contextual embeddings, possibly oversimplifying interpretations of similarity.
6. **Limited Contextual Understanding:** While cosine similarity provides a quantitative measure of similarity, it may not fully capture nuanced contextual differences in hotel reviews.
7. **Exclusion of Non-Textual Features:** The analysis focuses solely on textual data, neglecting other important features like images or metadata associated with reviews, which could provide a more comprehensive understanding of sustainability concerns.
8. **Generalization of Findings:** Findings may be specific to the dataset and context used, limiting extrapolation to broader contexts without validation.
9. **Validity of Sustainability Keywords:** The selection of sustainability keywords may not comprehensively capture all aspects of sustainability concerns, potentially overlooking emerging trends or region-specific terminology.
10. **Interpretation Bias:** Subjectivity in interpreting similarity scores and their association with sustainability concerns may introduce bias depending on researchers' preconceptions.

Despite these limitations, the findings of this research provide valuable insights and a foundation for further exploration of consumers' sustainability concerns in the hospitality industry. Future studies could address these limitations by incorporating additional data sources, considering cultural and linguistic diversity, refining AI-generated content, and integrating more comprehensive contextual information.

REFERENCES

- **Bais, G. (2023, July 26).** ***Building a sentiment classification system with BERT embeddings: Lessons learned*****. neptune.ai.** **https://neptune.ai/blog/building-sentiment-classification-system-with-bert-embeddings**

- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/09d5eccf-70e5-489c-b2d4-67e667ac759f**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/b7c34edc-2f0c-49a3-9860-f64deeaadb20**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/f42d4657-a6a4-405b-b738-8660e8d45e22**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/93a2c795-08a7-4674-ae57-99feacf70b11**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/6cf842a0-1403-4635-b20c-4edc267d40fa**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/6db63c5b-7904-42ea-ad45-f281dcfb8382**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/32b46ca4-39ef-452d-8bdb-d06ee4601b69**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/827e927d-d61e-4b4c-8528-e182f48f6885**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/108dc15c-4f04-4935-a57e-33709cb93d72**
- ***ChatGPT - Write for me***. **(n.d.). ChatGPT. https://chat.openai.com/share/f781323c-3e3e-48b8-91be-18fa39d2adc2**
- ***Gemini - Hotel reviews: Sustainability or sacrifice?*** **(n.d.). Gemini. https://gemini.google.com/share/ff46015889e1**
- ***Getting Started with Sentiment Analysis using Python***. **(n.d.). https://huggingface.co/blog/sentiment-analysis-python**
- ***MarieAngeA13/Sentiment-Analysis-BERT at main***. **(n.d.). https://huggingface.co/MarieAngeA13/Sentiment-Analysis-BERT/tree/main**
- **Moonat, D. (2023, November 22).** ***Fine-Tune BERT Model for Sentiment Analysis in Google CoLAB***. **Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/12/fine-tune-bert-model-for-sentiment-analysis-in-google-colab/**
- ***nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face***. **(n.d.). https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment**
- **Prakharrathi. (2020, December 24).** ***Sentiment Analysis using BERT***. **https://www.kaggle.com/code/prakharrathi25/sentiment-analysis-using-bert**
- ***Sustainable hotel Reviews***. **(n.d.). ChatGPT. https://chat.openai.com/share/b2d4935a-052e-40d7-ba32-9ab942b0c9cb**
- **Talaat, A. S. (2023). Sentiment analysis classification system using hybrid BERT models.** ***Journal of Big Data***, ***10***(1). **https://doi.org/10.1186/s40537-023-00781-w**

Name: Punit Rajesh Shah
Student ID: 14373530

APPENDIX

Bert Cosine similarity.ipynb

```
import pandas as pd
import torch
from transformers import BertTokenizer, BertModel
from sklearn.metrics.pairwise import cosine_similarity
import re
import matplotlib.pyplot as plt
import warnings

warnings.filterwarnings("ignore")

# Function for text preprocessing
def preprocess_text(text):
    if pd.isna(text):
        return ""
    # Convert to lowercase
    text = str(text).lower()
    # Remove punctuation
    text = re.sub(r'[^\w\s]', '', text)
    return text


# Load the Excel files into pandas DataFrames
df_booking = pd.read_excel("V1BookingHotel_-_TH_SG.xlsx")
df_ai_reviews = pd.read_excel("reviewsAIV1.xlsx")

# Preprocess text data in both datasets
df_booking["Text"] = df_booking["Text"].apply(preprocess_text)
df_ai_reviews["Reviews"] = df_ai_reviews["Reviews"].apply(preprocess_text)

# Initialize BERT tokenizer and model
tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")
model = BertModel.from_pretrained("bert-base-uncased")

# Preprocess and tokenize AI-generated reviews
ai_reviews = df_ai_reviews["Reviews"].tolist()
ai_reviews_tokenized = tokenizer(ai_reviews, padding=True, truncation=True, return_tensors="pt")

# Obtain BERT embeddings for AI-generated reviews
with torch.no_grad():
    ai_reviews_outputs = model(**ai_reviews_tokenized)
    ai_reviews_embeddings = ai_reviews_outputs.last_hidden_state.mean(dim=1)

# Function to calculate contextual similarity scores
def calculate_similarity_scores(text):
    if not isinstance(text, str):
        text = str(text)
    # Tokenize and encode the input text
    text_tokenized = tokenizer(text, padding=True, truncation=True, return_tensors="pt")
```

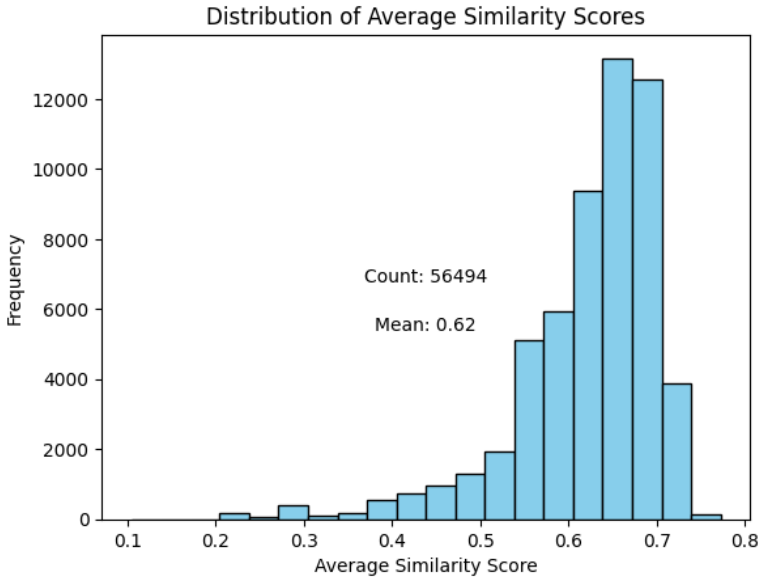
```
    # Obtain BERT embeddings for the input text
    with torch.no_grad():
        text_outputs = model(**text_tokenized)
        text_embedding = text_outputs.last_hidden_state.mean(dim=1)

    # Calculate cosine similarity between the input text and AI-generated reviews
    similarity_scores = cosine_similarity(text_embedding, ai_reviews_embeddings)

    return similarity_scores


# Iterate through each review in df_booking and calculate similarity scores
similarity_scores_list = []
average_similarity_scores = []
for idx, row in df_booking.iterrows():
    text = row["Text"]
    similarity_scores = calculate_similarity_scores(text)
    similarity_scores_list.append(similarity_scores)
    average_similarity_score = similarity_scores.mean()
    average_similarity_scores.append(average_similarity_score)

# Add similarity scores to df_booking
#df_booking["Similarity Scores"] = similarity_scores_list

# Add average similarity scores to df_booking
df_booking["Average Similarity Score"] = average_similarity_scores

# Save the updated DataFrame with similarity scores
df_booking.to_excel("V1BookingHotel_with_AvgSimilarity_Scores.xlsx", index=False)
```

Hotel Data Visualization.ipynb

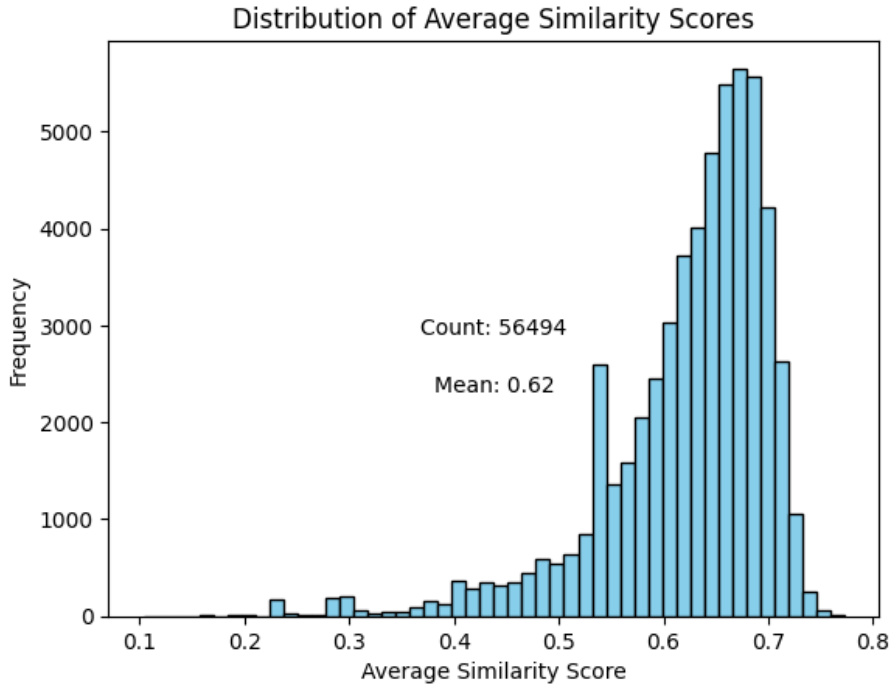
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
import warnings

warnings.filterwarnings("ignore")
# Load the Excel files into pandas DataFrames
df_booking = pd.read_excel("V1BookingHotel_with_AvgSimilarity_Scores.xlsx")
df_ai_reviews = pd.read_excel("reviewsAIV1.xlsx")
# Visualize distribution of average similarity scores
plt.hist(df_booking["Average Similarity Score"], bins=20, color='skyblue', edgecolor='black')
plt.title('Distribution of Average Similarity Scores')
plt.xlabel('Average Similarity Score')
plt.ylabel('Frequency')
# Add labels for count and mean
plt.text(0.5, 0.5, f'Count: {len(df_booking["Average Similarity Score"])}', horizontalalignment='center',
verticalalignment='center', transform=plt.gca().transAxes)
```

plt.text(0.5, 0.4, f'Mean: {np.mean(df_booking["Average Similarity Score"]):.2f}',
horizontalalignment='center', verticalalignment='center', transform=plt.gca().transAxes)

plt.show()

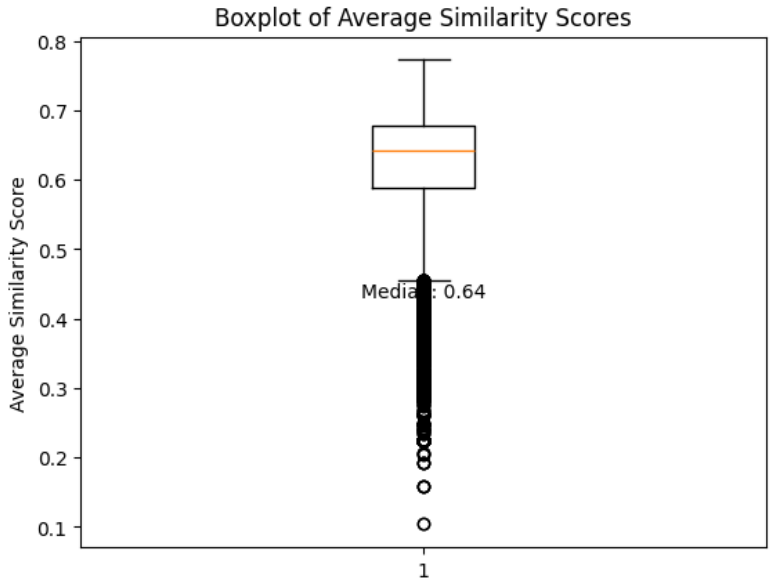



```
# Visualize distribution of average similarity scores
plt.hist(df_booking["Average Similarity Score"], bins=50, color='skyblue', edgecolor='black')
plt.title('Distribution of Average Similarity Scores')
plt.xlabel('Average Similarity Score')
plt.ylabel('Frequency')
# Add labels for count and mean
plt.text(0.5, 0.5, f'Count: {len(df_booking["Average Similarity Score"])}', horizontalalignment='center',
verticalalignment='center', transform=plt.gca().transAxes)
plt.text(0.5, 0.4, f'Mean: {np.mean(df_booking["Average Similarity Score"]):.2f}',
horizontalalignment='center', verticalalignment='center', transform=plt.gca().transAxes)

plt.show()
```

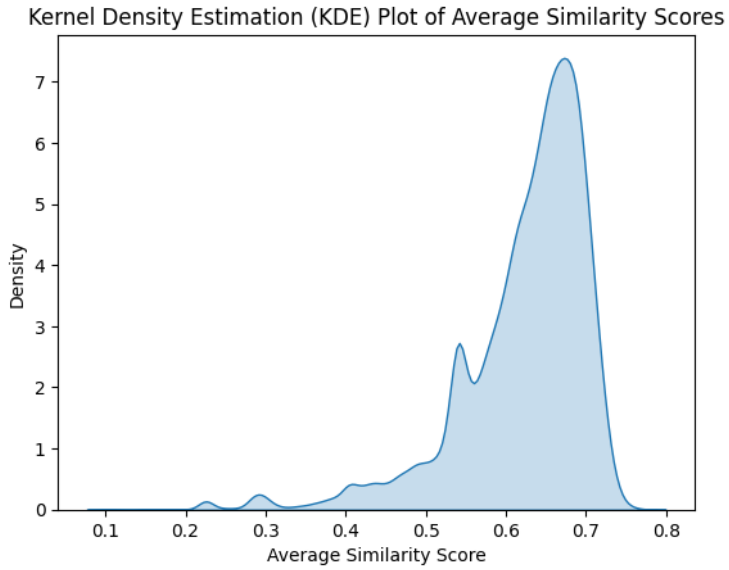



```
# Boxplot of Average Similarity Scores
plt.boxplot(df_booking["Average Similarity Score"])
```

```
plt.title('Boxplot of Average Similarity Scores')
plt.ylabel('Average Similarity Score')
# Add label for median
plt.text(0.5, 0.5, f'Median: {np.median(df_booking["Average Similarity Score"]):.2f}',
horizontalalignment='center', verticalalignment='center', transform=plt.gca().transAxes)

plt.show()
```

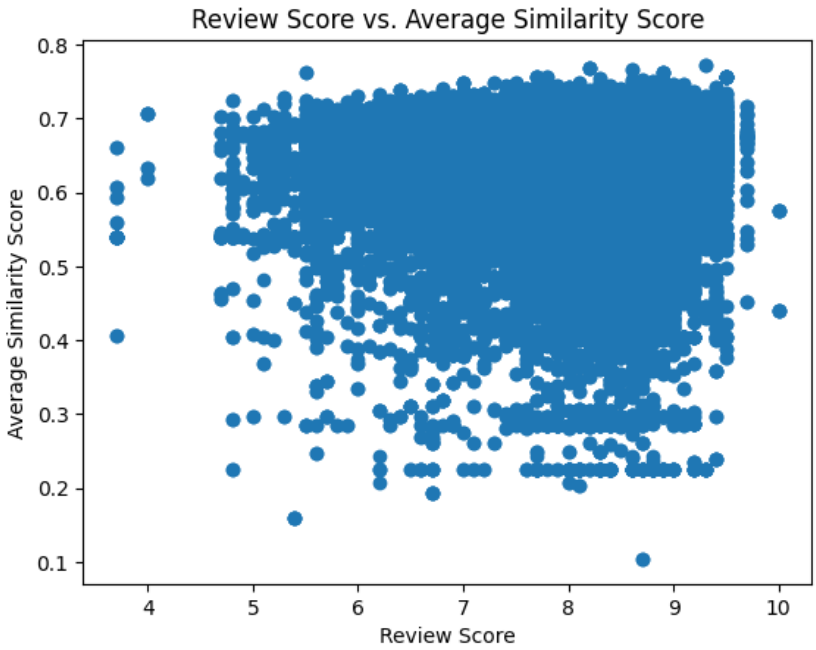



```
# Kernel Density Estimation (KDE) Plot of Average Similarity Scores
sns.kdeplot(df_booking["Average Similarity Score"], shade=True)
plt.title('Kernel Density Estimation (KDE) Plot of Average Similarity Scores')
plt.xlabel('Average Similarity Score')
plt.ylabel('Density')
plt.show()
```

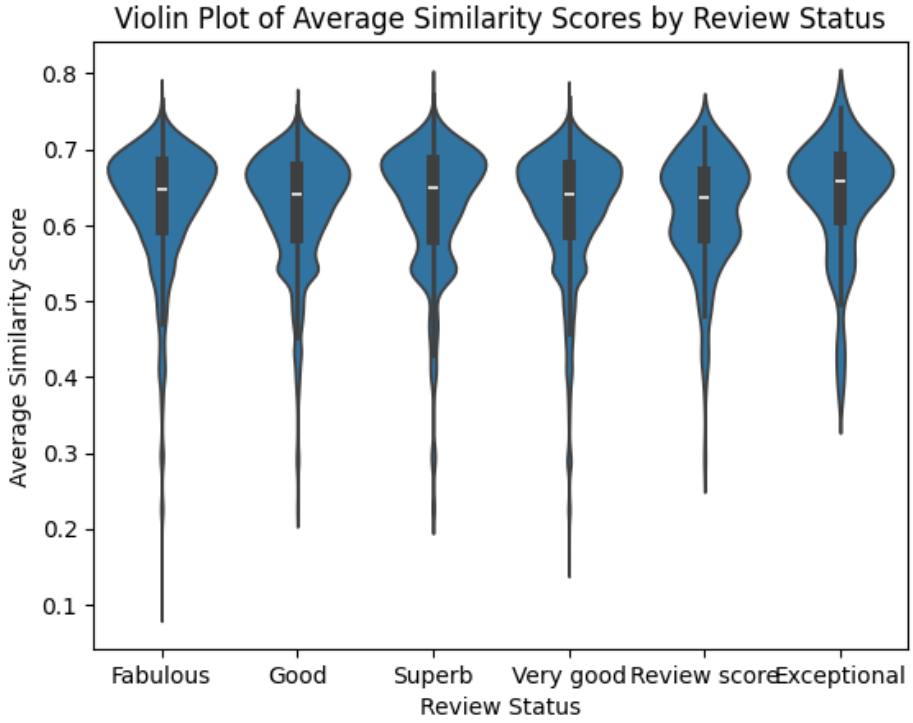



```
# Scatter plot of Review Score vs. Average Similarity Score
plt.scatter(df_booking['Review Score'], df_booking['Average Similarity Score'])
plt.title('Review Score vs. Average Similarity Score')
plt.xlabel('Review Score')
plt.ylabel('Average Similarity Score')
plt.show()
```

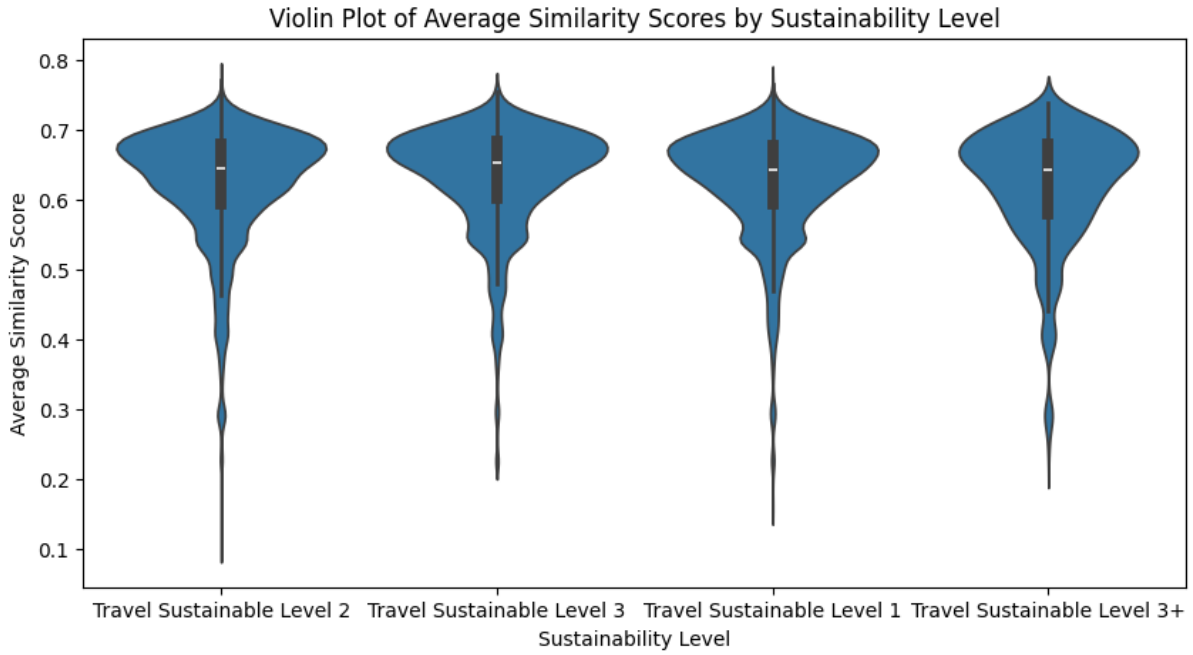
Name: Punit Rajesh Shah
Student ID: 14373530



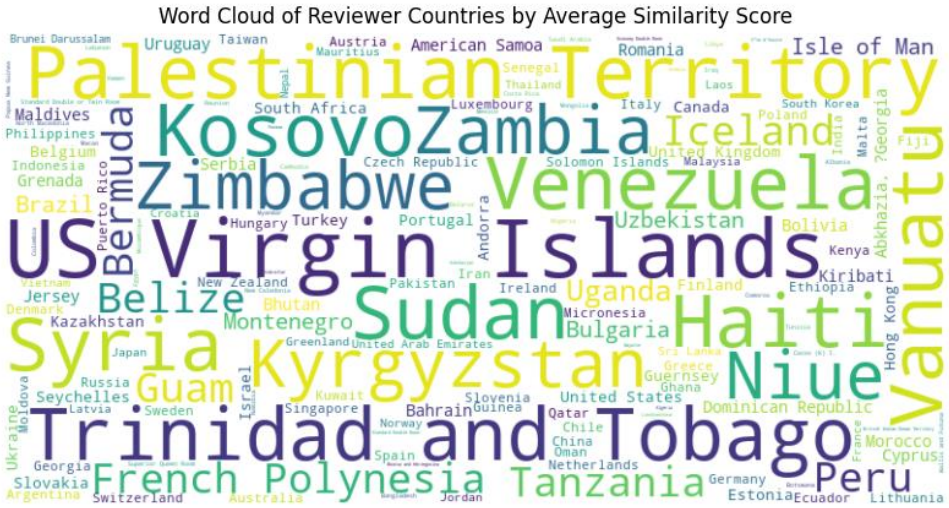

```
# Violin plot of Average Similarity Scores by Review Status
sns.violinplot(x='Review Status', y='Average Similarity Score', data=df_booking)
plt.title('Violin Plot of Average Similarity Scores by Review Status')
plt.xlabel('Review Status')
plt.ylabel('Average Similarity Score')
plt.show()
```




```
# Violin plot of Average Similarity Scores by Review Status
plt.figure(figsize=(10, 5))

sns.violinplot(x='Sustainability Level', y='Average Similarity Score', data=df_booking)
plt.title('Violin Plot of Average Similarity Scores by Sustainability Level')
plt.xlabel('Sustainability Level')
plt.ylabel('Average Similarity Score')
plt.show()
```



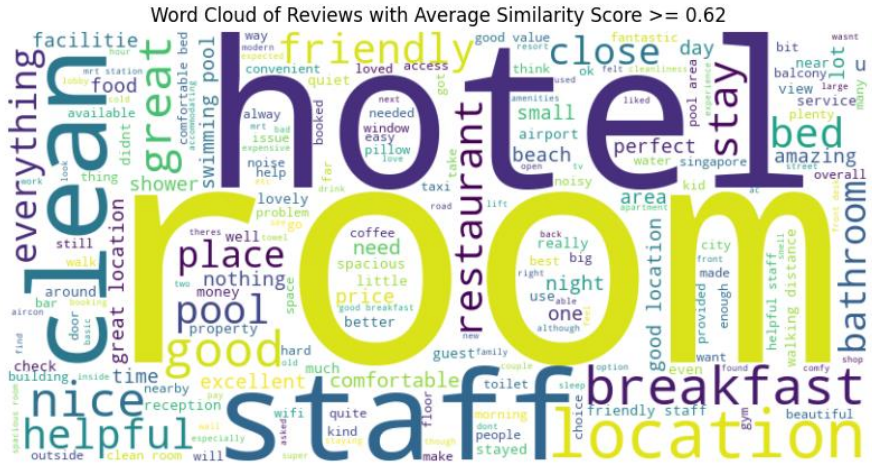

```
# Aggregate data by Reviewer Country and calculate mean of Average Similarity Score
country_similarity_mean = df_booking.groupby('Reviewer Country')['Average Similarity Score'].mean()

# Generate word cloud with country names weighted by mean similarity score
wordcloud = WordCloud(width=800, height=400,
background_color='white').generate_from_frequencies(country_similarity_mean)

# Display the word cloud
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.title('Word Cloud of Reviewer Countries by Average Similarity Score')
plt.axis('off')
plt.show()
```




```
# Word cloud of all reviews
all_reviews_text = ' '.join(df_ai_reviews['Reviews'])
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(all_reviews_text)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.title('Word Cloud of Reviews')
```

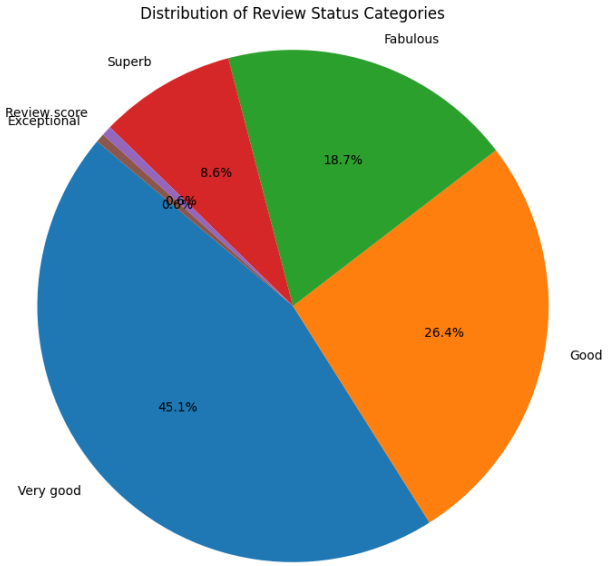
```
plt.axis('off')
plt.show()
```


```
# Filter rows with Average Similarity Score >= 0.62
high_similarity_df = df_booking[df_booking['Average Similarity Score'] >= 0.62]
high_similarity_reviews = ' '.join(high_similarity_df['Text'])
wordcloud = WordCloud(width=800, height=400,
background_color='white').generate(high_similarity_reviews)
plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.title('Word Cloud of Reviews with Average Similarity Score >= 0.62')
plt.axis('off')
plt.show()
```

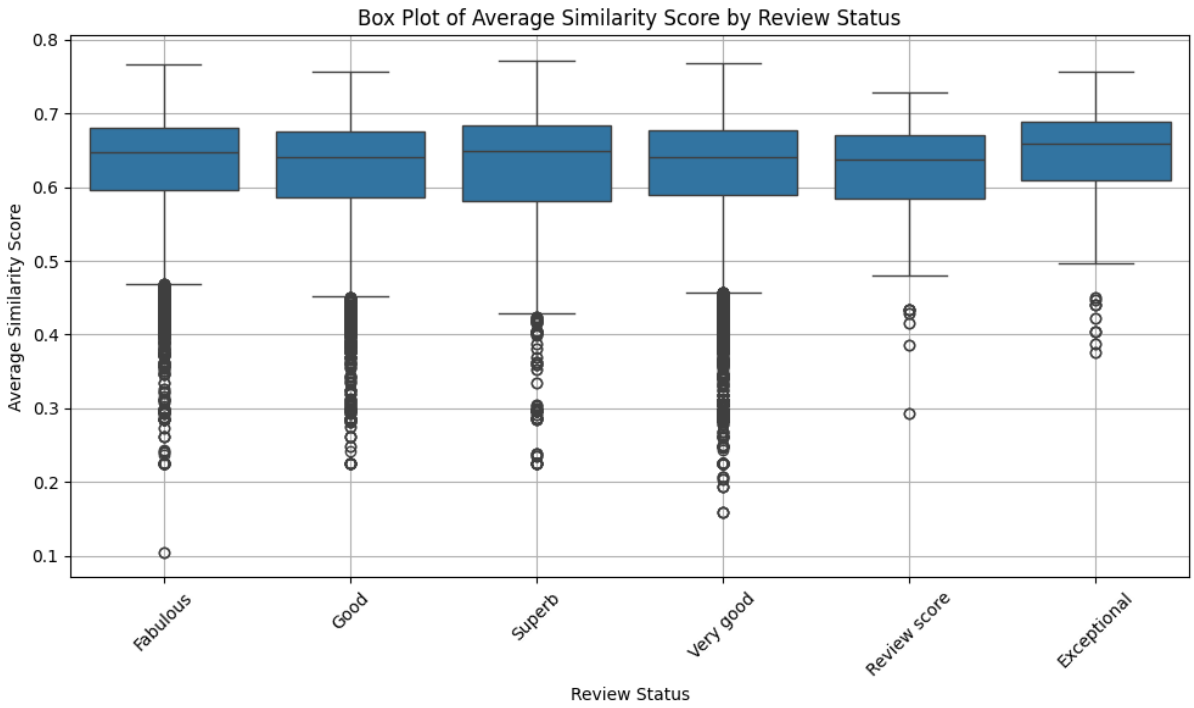



```
# Count the occurrences of each "Review Status" category
review_status_counts = df_booking["Review Status"].value_counts()

# Create a pie chart
plt.figure(figsize=(8, 8))
plt.pie(review_status_counts, labels=review_status_counts.index, autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Review Status Categories')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```

Name: Punit Rajesh Shah
Student ID: 14373530

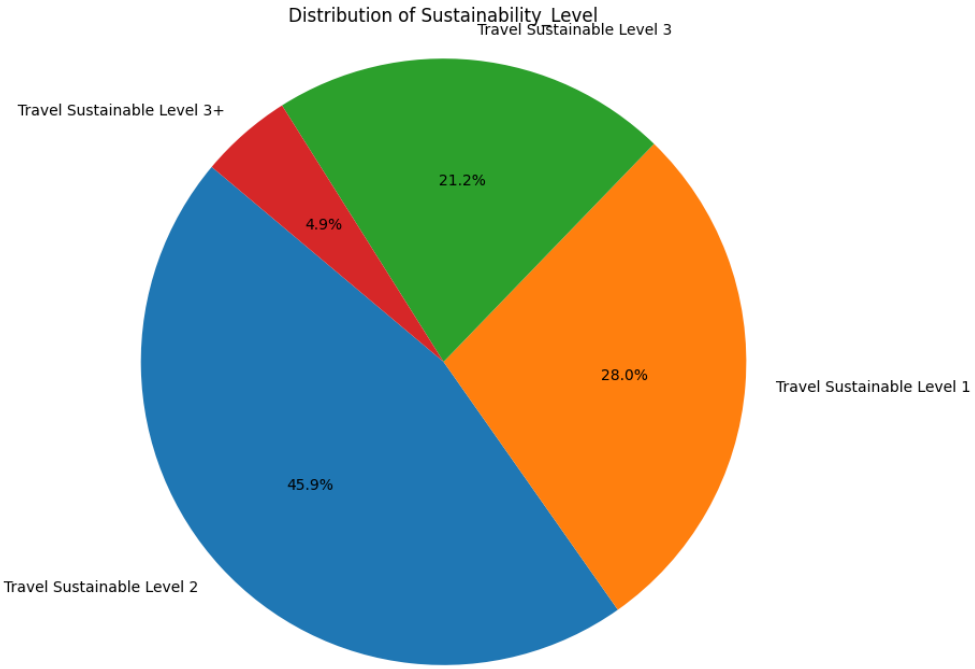



```
# Box plot of Average Similarity Score by Review Status
plt.figure(figsize=(10, 6))
sns.boxplot(x='Review Status', y='Average Similarity Score', data=df_booking)
plt.title('Box Plot of Average Similarity Score by Review Status')
plt.xlabel('Review Status')
plt.ylabel('Average Similarity Score')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```

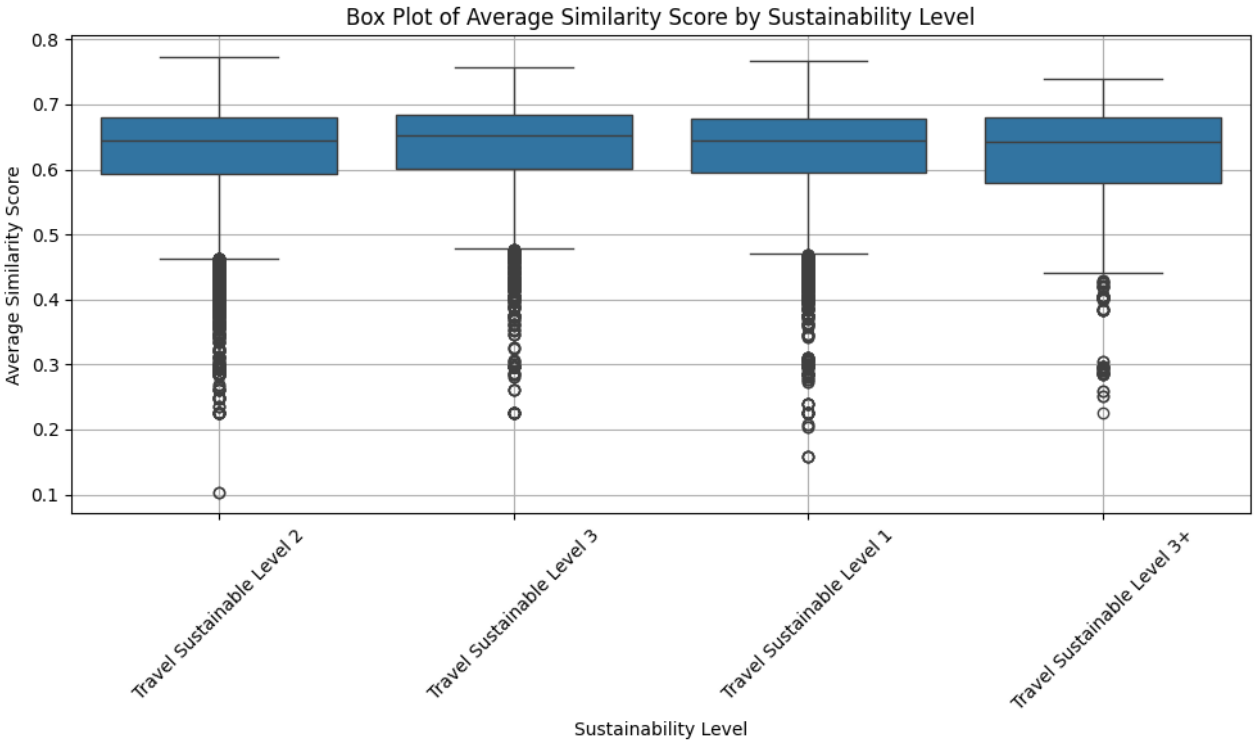



```
# Count the occurrences of each "Review Status" category
Sustainability_Level_counts = df_booking["Sustainability Level"].value_counts()

# Create a pie chart
plt.figure(figsize=(8, 8))
plt.pie(Sustainability_Level_counts, labels=Sustainability_Level_counts.index, autopct='%1.1f%%',
startangle=140)
```

```
plt.title('Distribution of Sustainability_Level')
plt.axis('equal')  # Equal aspect ratio ensures that pie is drawn as a circle.
plt.show()
```

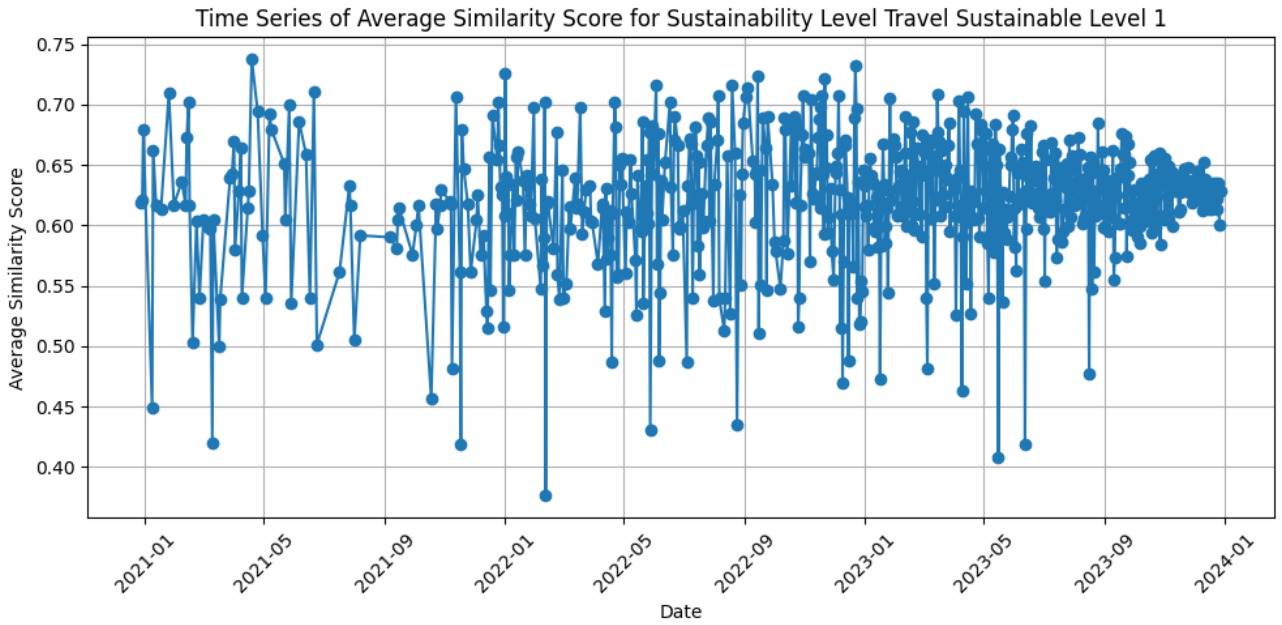


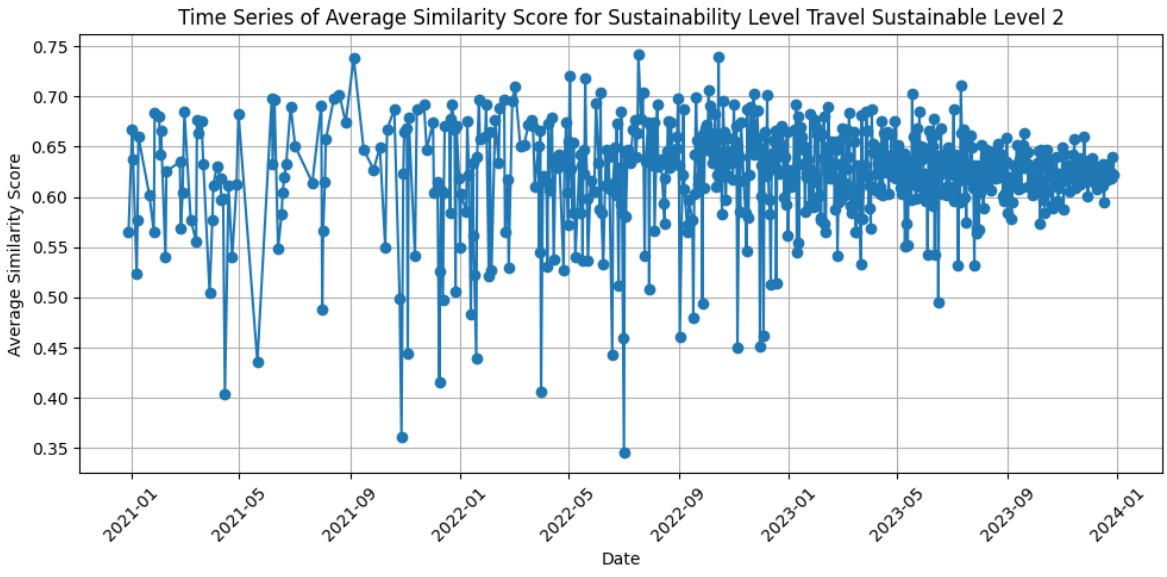

```
# Box plot of Average Similarity Score by Review Status
plt.figure(figsize=(10, 6))
sns.boxplot(x='Sustainability Level', y='Average Similarity Score', data=df_booking)
plt.title('Box Plot of Average Similarity Score by Sustainability Level')
plt.xlabel('Sustainability Level')
plt.ylabel('Average Similarity Score')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



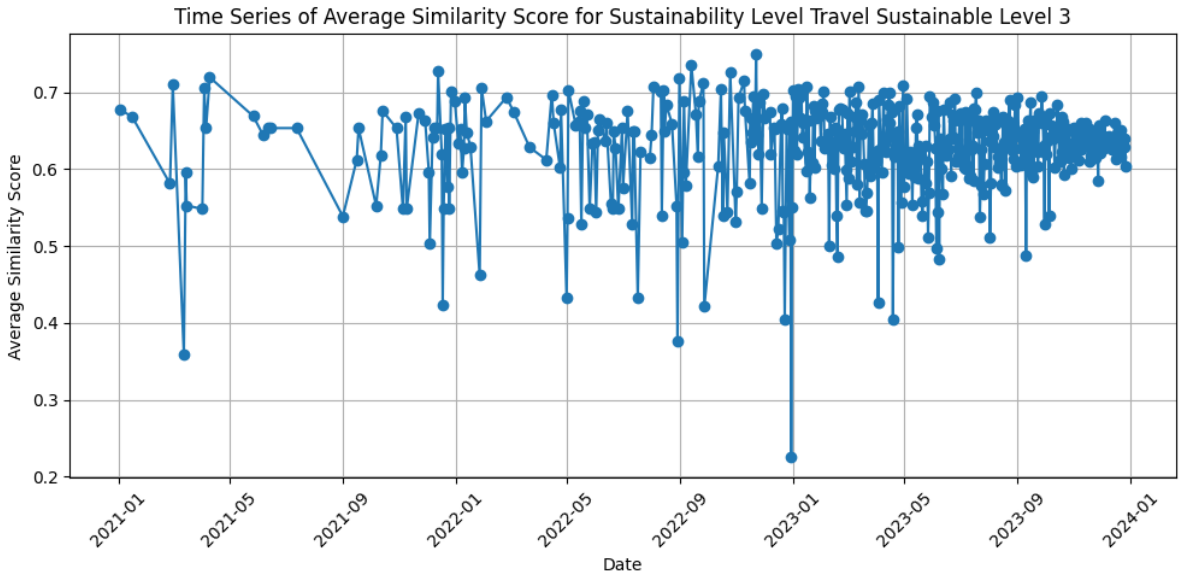
Name: Punit Rajesh Shah
Student ID: 14373530

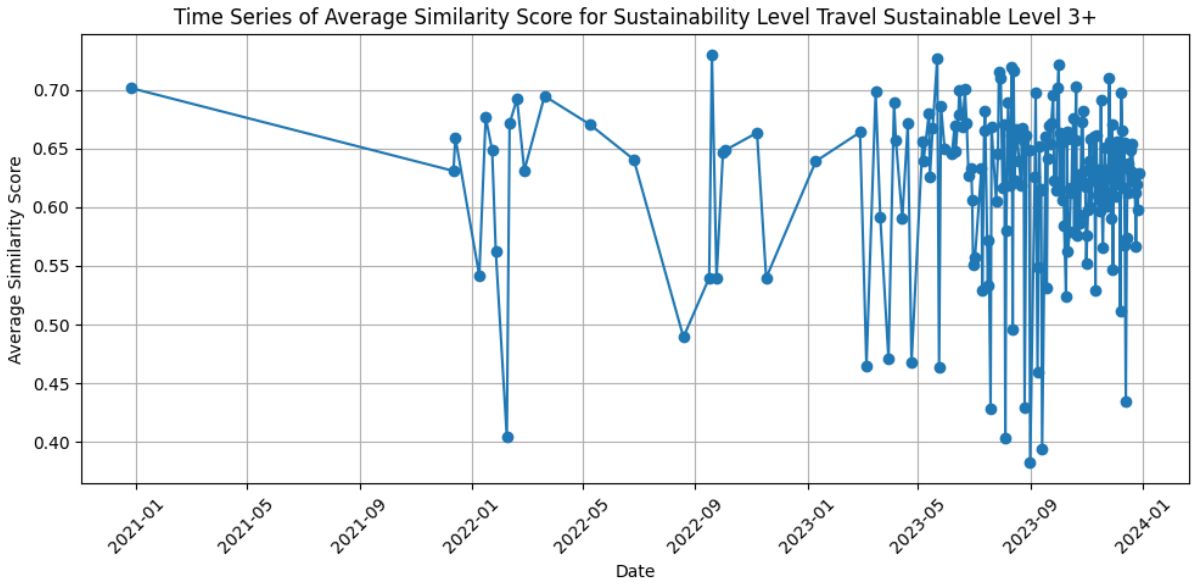
```
# Time series plot of Average Similarity Scores for each Sustainability Level with dates arranged in ascending order
df_booking['Date'] = pd.to_datetime(df_booking['Date'], errors='coerce')
df_booking3 = df_booking.dropna(subset=['Date'])
grouped_by_sustainability_date = df_booking3.groupby(['Sustainability Level', 'Date'])['Average Similarity Score'].mean().reset_index()

for sustainability_level, group in grouped_by_sustainability_date.groupby('Sustainability Level'):
    group_sorted = group.sort_values('Date')  # Sort by date in ascending order
    plt.figure(figsize=(10, 5))
    plt.plot(group_sorted['Date'], group_sorted['Average Similarity Score'], marker='o', linestyle='-')
    plt.title(f"Time Series of Average Similarity Score for Sustainability Level {sustainability_level}")
    plt.xlabel("Date")
    plt.ylabel("Average Similarity Score")
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.tight_layout()
    plt.show()
```





Name: Punit Rajesh Shah
Student ID: 14373530






```
# Time series plot of Average Similarity Scores for each hotel with dates arranged in ascending order
df_booking['Date'] = pd.to_datetime(df_booking['Date'], errors='coerce')
df_booking2 = df_booking.dropna(subset=['Date'])
grouped_by_hotel_date = df_booking2.groupby(['Name', 'Date'])['Average Similarity
Score'].mean().reset_index()

for name, group in grouped_by_hotel_date.groupby('Name'):
    group_sorted = group.sort_values('Date')  # Sort by date in ascending order
    plt.figure(figsize=(10, 5))
    plt.plot(group_sorted['Date'], group_sorted['Average Similarity Score'], marker='o', linestyle='-')
    plt.title(f"Time Series of Average Similarity Score for Hotel {name}")
    plt.xlabel("Date")
    plt.ylabel("Average Similarity Score")
    plt.xticks(rotation=45)
    plt.grid(True)
    plt.tight_layout()
    plt.show()

# all the graphs can be found in the code file
```