Methodology

1. Data Preprocessing and Transformation

- **Extract Offerings and Destinations from Tags:**
  - Utilize a mapping file to parse the tags column and extract information about offerings and destinations.
  - Implement custom functions to process the tags and create new columns representing these extracted features. This step ensures that the data is categorized and structured in a manner that facilitates further analysis.
- **Generate New Columns from Ratings:**
  - Extract two new columns from the ratings column:
    - norm_user_rating: Represents normalized user ratings, scaled between 0 and 100 for consistency.
    - user_rating: Represents the original user ratings. This dual representation helps in various analytical scenarios, providing both normalized and raw data views.
- **Derive Sentiment Column:**
  - Sentiment analysis categorizes reviews into positive, neutral, or negative sentiments, which is crucial for understanding customer feedback.
  - **Approach 1: Using Pre-trained Sentiment Analysis Pipeline**
    - Use the pre-trained sentiment analysis model akhooli/xlm-r-large-arabic-sent, which is based on XLM-R (Cross-lingual Roberta), a transformer model designed for cross-lingual understanding.
    - XLM-R is a variant of the Roberta model, which itself is an optimized version of BERT (Bidirectional Encoder Representations from Transformers). Roberta improves on BERT by training with larger mini-batches, longer sequences, and removing the next sentence prediction objective.
    - The model processes review text by tokenizing it, feeding it through the transformer network, and interpreting the output logits to assign a sentiment label.
  - **Approach 2: Custom Model (LLM - Llama 3)**
    - Using LLM model to do the sentiment analysis, a 4bit quantized Llama was chosen. Llama 3 is a state-of-the-art transformer-based model designed to handle large-scale language tasks with high accuracy and efficiency. It builds on the architecture of previous LLMs, incorporating improvements in model scaling, training data, and fine-tuning techniques.
    - Llama3 proves to be a powerful tool for sentiment analysis, especially when paired with the right prompts.
    - LLMs like Llama3 excel at grasping the nuances of language, including sarcasm, humor, and double meanings. This is crucial for accurate sentiment analysis, which can be easily misled by literal interpretations.
    -

2. Text Cleaning and NLP Analysis

- **Text Cleaning:**
  - Implement comprehensive NLP preprocessing techniques:
    - Remove stop words to eliminate common but uninformative words from the text.
    - Apply stemming to reduce words to their base forms, ensuring consistency in text analysis.
    - Remove punctuation, special characters, and digits to clean the text and focus on meaningful words.
    - Convert text to lowercase for uniformity, aiding in accurate analysis.
- **NLP Analysis:**
  - **Term Frequency-Inverse Document Frequency (TF-IDF):**
    - TF-IDF is a numerical statistic that reflects the importance of a word in a document relative to a collection of documents (corpus). It is calculated as the product of two terms: Term Frequency (TF), which measures how frequently a term appears in a document, and Inverse Document Frequency (IDF), which measures how important a term is within the corpus.
    - This transformation helps in highlighting important words in the reviews while downplaying common words.
  - **Non-negative Matrix Factorization (NMF):**
    - NMF is a dimensionality reduction technique used to decompose the TF-IDF matrix into two lower-dimensional matrices: a basis matrix and a coefficient matrix. This decomposition helps in identifying latent topics in the text data.
    - By applying NMF to the TF-IDF matrix, we can uncover underlying themes or topics present in the reviews, providing insights into common discussion points and trends.

3. Exploratory Data Analysis (EDA)

- **Sentiment Distribution:**
  - Analyze and visualize the distribution of sentiments across the dataset by categorizing reviews into positive, neutral, and negative sentiments.
  - Visualize this distribution using bar charts or pie charts to understand the overall sentiment trends within the data.
- **Offerings and Destinations Analysis:**
  - Examine the frequency and distribution of different offerings and destinations extracted from the tags.
  - Explore correlations between these features and other variables, such as user ratings and sentiments, to identify patterns and relationships.
- **Ratings Analysis:**
  - Investigate the patterns and trends in user ratings, analyzing how they relate to other variables like sentiment.

    - Explore the relationship between user ratings and sentiment to identify any notable correlations or trends.
- **Variable Correlations:**
    - Use statistical methods to examine correlations between various variables in the dataset.
    - Visualize these relationships using heatmaps or scatter plots to uncover significant patterns or insights, helping to understand the interplay between different features.

Findings

Sentiment Analysis:

1. **Positive Sentiment Dominance**:
    - Most of the customer feedback (69.8%) is positive.
    - Negative and neutral sentiments are nearly equally distributed at around 15% each.

Keyword/Category Identification:

2. **Tourism and Accommodation Lead**:
    - Tourism attractions and sites are the most frequently mentioned offerings, followed by accommodation.
    - Other categories like food & beverage, retail, and religious offerings have lower frequencies.
3. **Destination Popularity**:
    - Riyadh, Madinah, and Jeddah are the top destinations with the highest number of mentions.
    - Smaller cities like Al Diriyah and Al Ula have significantly fewer mentions.
4. **User Ratings Distribution**:
    - There is a concentration of high ratings, with a significant spike at 100, indicating many users rate their experiences very highly.
5. **Offerings Comparison by Destination**:
    - Each top destination has a distinct pattern of offerings. For example, Riyadh has high mentions in tourism and accommodation, while Madinah excels in tourism and religious offerings.

6. **Normalized User Ratings by Offering**:

- Accommodation: Generally high ratings across most destinations, with notable peaks in Yanbu and Jizan.
- Food & Beverage: High ratings in destinations like Jizan and Al Ahsa, indicating good satisfaction levels in these areas.
- Religious: Exceptional ratings in cities like Makkah, reflecting strong positive feedback for religious offerings.
- Retail: Riyadh, Khobar, and Jeddah stand out with very high ratings in this category.
- Tourism Attractions/Sites: High ratings are consistent in destinations such as Abha, Riyadh, and Yanbu, showing strong user satisfaction.

7. **Regional Differences**:
    - Variations in user ratings across different destinations highlight regional preferences and satisfaction levels. For instance, the high ratings in Makkah for religious offerings and Riyadh for retail suggest differing strengths that cater to unique visitor interests.

Recommendations

Enhancing Customer Experience:

1. **Focus on Positive Feedback**:
    - Leverage the dominant positive sentiment by highlighting positive reviews in marketing campaigns to attract new customers.
    - Encourage satisfied customers to share their positive experiences on social media and review platforms.
2. **Address Negative Feedback**:
    - Conduct a deeper analysis of negative and neutral feedback to identify common issues or areas of improvement.
    - Implement a feedback loop where negative comments are addressed promptly and effectively to enhance customer satisfaction.

Service and Offering Improvements:

3. **Improve Lesser-mentioned Offerings**:
    - Invest in enhancing food & beverage, retail, and religious offerings, as these areas have lower frequencies but potential for growth.
    - Consider customer preferences and trends to diversify and improve these services.

4. **Tailored Services by Destination**:
    - Develop destination-specific strategies. For instance, enhance tourism experiences in Riyadh and Madinah to maintain their leading positions.
    - Promote lesser-known destinations like Al Diriyah and Al Ula with targeted marketing campaigns highlighting unique attractions.
5. **Improving Specific Offerings:**
    - Enhance offerings in categories with more mixed ratings, such as "Retail" and "Food & Beverage," by introducing new and diverse options that cater to different customer preferences.
    - Partner with local businesses to improve the variety and quality of retail and dining experiences, ensuring they meet the high standards expected by customers.

User Engagement and Rating Management:

5. **Encourage Honest Feedback**:
    - Encourage customers to leave detailed feedback and ratings, providing insights for continuous improvement.
    - Implement loyalty programs or incentives for customers who provide feedback, especially in less frequented categories.
6. **Maintain High Standards**:
    - Given the high user ratings across various offerings, maintain and further improve service standards to ensure continued customer satisfaction.
    - Regularly train staff and update facilities to meet and exceed customer expectations.

- All the codes and files can be found at: https://github.com/alikhater75/Artifact-DS_A-L2