German University in Cairo

Faculty of Media Engineering and Technology

Advanced Computer Lab CSEN 903 Milestone 1 Report

Team 101

Contents

Chapter 1

Introduction

This project analyzes hotel booking data to predict which country group customers come from based on their reviews and demographics. The dataset includes 25 hotels, 50,000 reviews, and 2,000 users from around the world.

Hotel booking platforms collect valuable data about customer preferences and satisfaction. Different regions show different patterns in what they value in hotels. This project uses review scores along with user information to understand these regional differences.

1.1 Problem Statement

The goal is to predict country groups (like North America, Western Europe, East Asia, etc.) from hotel review data. The project also answers: which city is best for each traveler type, and which countries have the best value for money by age group.

1.2 Aims and Objectives

Aim: Build a machine learning model to classify reviews into country groups.

Objectives:

- Clean and prepare the data
- Answer the data engineering questions
- Create useful features from the data
- Train and evaluate classification models
- Use SHAP and LIME to explain predictions

1.3 Solution Approach

The project follows these steps: clean the data, analyze it to answer the questions, engineer features, build classification models, and apply explainability techniques to understand what drives the predictions.

Chapter 2

Dataset Cleaning

Before performing exploratory analysis or model training, a thorough cleaning process was applied to ensure the datasets were accurate, consistent, and reliable. The checks focused on identifying duplicate identifiers, missing values, and verifying the uniqueness of categorical features. These steps were essential to guarantee the integrity of the data that will later be used to predict the *country group* of each hotel based on tourist reviews.

2.0.1 Duplicated ID Check

Each dataset was examined for duplicate identifiers to ensure every record was uniquely defined. The *Hotels* dataset contained 25 unique hotel IDs, the *Users* dataset included 2,000 unique user IDs, and the *Reviews* dataset had 50,000 unique review IDs. No duplicate identifiers were found in any of the datasets, confirming that all entries are uniquely represented and suitable for merging in later stages.

2.0.2 Missing / Null / Empty Values Check

All datasets were checked for missing, null, or empty-string values across their attributes. The results indicated that none of the datasets contained missing or empty fields. This confirms that the data was complete and ready for use without requiring imputation or removal of records, ensuring consistent downstream processing.

2.0.3 Unique Values

The categorical features were analyzed to assess data diversity and validity. In the *Hotels* dataset, each hotel name, city, and country was unique, covering 25 distinct global locations, while all hotels shared a consistent five-star rating. For the *Users* dataset, the gender distribution included three categories (*Male*, *Female*, *Other*), with users originating from 25 different countries. Age groups were classified into five ranges, and traveller types spanned four categories (*Solo*, *Family*, *Couple*, *Business*). These findings confirm that the data captures a diverse range of travelers and destinations, forming a strong foundation for predictive modeling of hotel country groups.

Chapter 3

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics, structure, and quality of a dataset before applying machine learning techniques. It provides insights into data composition, relationships, and potential issues that may affect model performance. This section begins with a general overview of the datasets, followed by descriptive statistics and relevant visualizations.

3.1 General Overview

The initial stage of exploratory data analysis (EDA) aims to gain a general understanding of the three datasets — *Hotels*, *Users*, and *Reviews*. This involves examining their dimensions, structure, data types, and sample content. Establishing a clear understanding of the datasets is essential before feature engineering and modeling, as the project's primary goal is to predict the *country group* of each hotel based on the tourists' reviews. The *country group* will later be introduced as the target variable for model training.

3.1.1 Dataset Shape

A preliminary inspection of dataset dimensions was carried out to assess their scale and composition. The *Hotels* dataset contains 25 records with 13 features, the *Users* dataset includes 2,000 records with 6 attributes, and the *Reviews* dataset comprises 50,000 records with 12 attributes. This distribution indicates that the *Reviews* dataset forms the core analytical component, representing the main source of behavioral and feedback information from tourists. The *Hotels* dataset serves as a compact reference table, while the *Users* dataset provides contextual demographic details for the reviewers. Together, these datasets create a foundation for linking user feedback to hotel characteristics, which will be crucial for predicting the hotel's country group.

3.1.2 Dataset Columns

The column names of each dataset were examined to verify completeness and proper labeling. The *Hotels* dataset includes unique identifiers, geographic location details, and base quality metrics such as cleanliness, comfort, and facilities — variables that describe the intrinsic characteristics of each hotel. The *Users* dataset provides demographic attributes like gender, country, age group, and traveller type, offering valuable context about the reviewers' backgrounds. The *Reviews* dataset links users and hotels through relational identifiers (`user_id`,

`hotel_id`) and contains review-specific details, including review date, multiple rating dimensions, and the written review text. This structure ensures the datasets can be effectively merged and analyzed to uncover relationships between user experience, hotel attributes, and geographic groupings.

3.1.3 Data Types

Data type validation was performed to confirm the correctness and consistency of the loaded data. Numerical variables, including rating scores and coordinates, were correctly represented as integer or floating-point types, while textual fields such as hotel names, countries, and review content were stored as string objects. Date columns, like `review_date`, were stored as text and can later be converted to a datetime format for temporal analysis. This ensures that each feature is appropriately structured for subsequent transformations, aggregations, and model preprocessing steps. Correct data typing is particularly important as some of these features, such as averaged rating scores or aggregated textual sentiment, may serve as key predictors of the target variable — the hotel's country group.

3.1.4 Dataset Sample

To gain an initial qualitative understanding of the data, representative samples from each dataset were displayed. The hotel samples included globally recognized destinations with consistently high base ratings, suggesting reliable reference data. The user samples demonstrated diversity in nationality, age group, and traveller type, reflecting a wide variety of reviewing behaviors. The review samples showcased structured rating metrics alongside rich textual comments that will later support both quantitative and linguistic analysis. Overall, the samples confirm that the datasets are coherent, realistic, and sufficiently detailed to support predictive modeling aimed at determining the hotel's country group based on tourist feedback.

3.2 Statistics

This section summarizes the main statistical characteristics of the three datasets — *Hotels*, *Users*, and *Reviews* — through descriptive measures of numerical attributes, categorical distributions, and aggregated review counts per country group. These statistics help understand the overall structure and variability of the data before applying predictive modeling techniques.

3.2.1 Calculations of Numeric Columns

Descriptive statistics were computed for all numeric columns to assess their central tendency and dispersion. In the *Hotels* dataset, the star ratings were consistent across all entries (mean and median of 5.00, standard deviation 0.00), confirming a uniform five-star rating. The base attribute scores (cleanliness, comfort, facilities, location, staff, and value for money) all exhibited high averages ranging between 8.5 and 9.3, with low standard deviations (0.23–0.36), reflecting consistent hotel quality.

The *Users* dataset contained no numeric columns, as it primarily consisted of categorical demographic attributes. In the *Reviews* dataset, review scores also showed strong consistency, with mean values between 8.4 and 9.2 and standard deviations under 0.6. This indicates generally positive tourist experiences and limited variation among reviews.

3.2.2 Count of Categories in Categorical Columns

Categorical analysis revealed diversity across user and hotel attributes. In the *Hotels* dataset, there were 25 unique hotel names, cities, and countries, each representing a different location worldwide, while all hotels shared a single star rating category of five stars. In the *Users* dataset, gender included three categories (*Male*, *Female*, *Other*), with users distributed across 25 countries. Age groups were divided into five ranges, and traveller types covered four categories (*Couple*, *Family*, *Solo*, *Business*). The *Reviews* dataset contained no categorical columns. These distributions confirm a well-balanced dataset that captures diverse demographics and travel contexts suitable for generalization in predictive modeling.

3.2.3 Country Group Creation and Reviews per Group

To facilitate regional-level analysis and prediction, each hotel was assigned a *country group* based on its country using predefined geographic mappings (e.g., *Western Europe*, *Middle East*, *Africa*). After merging the three datasets, review counts were aggregated per country group.

The highest number of reviews originated from *Western Europe* (11,876 reviews), followed by *Africa* (6,132) and *East Asia* (6,082). The smallest groups were *South Asia* (1,989) and *Eastern Europe* (1,970). This distribution highlights a global representation of hotel reviews with some regions being more frequently reviewed, likely reflecting tourism trends and user activity patterns.

3.3 Data Visualizations

To gain deeper insights into the datasets, we employed several visualization techniques to explore distributions, identify outliers, examine categorical patterns, and assess relationships between variables. This section presents four types of visualizations: histograms, boxplots, bar charts, and correlation heatmaps.

3.3.1 Histograms: Distribution Analysis

Histograms provide a clear view of how values are distributed across numeric columns, revealing patterns such as skewness, modality, and concentration of values.

Hotels Dataset

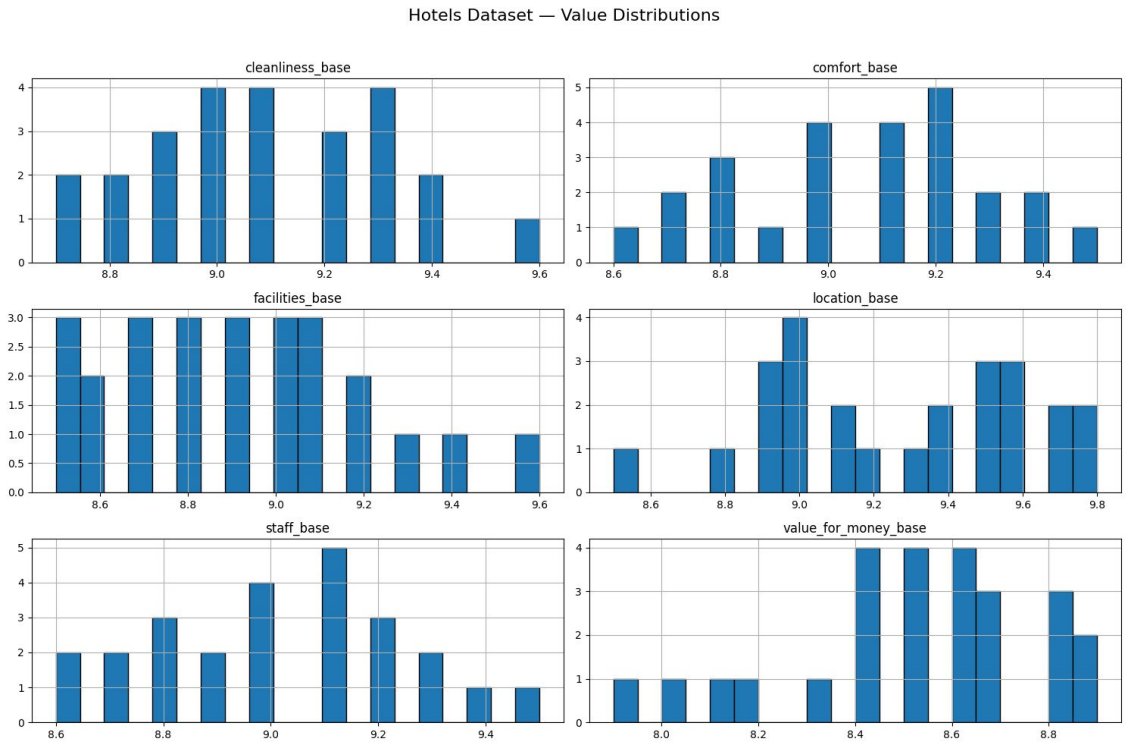

Figure 3.1: Value distributions for Hotels Dataset numeric columns

The Hotels dataset histograms reveal relatively uniform distributions across all rating dimensions. Most base scores cluster around the 9.0–9.5 range, indicating that the hotels in this dataset generally receive high ratings. Notable observations include:

- **Cleanliness and Facilities**: Show peaks near 9.0 and 9.3, suggesting consistent high standards
- **Comfort and Staff**: Display similar patterns with concentration around 9.2
- **Location**: Shows the highest ratings, with many hotels scoring above 9.5
- **Value for Money**: Exhibits the widest spread, ranging from 8.0 to 9.0, indicating more variability in guests' perception of value

Reviews Dataset

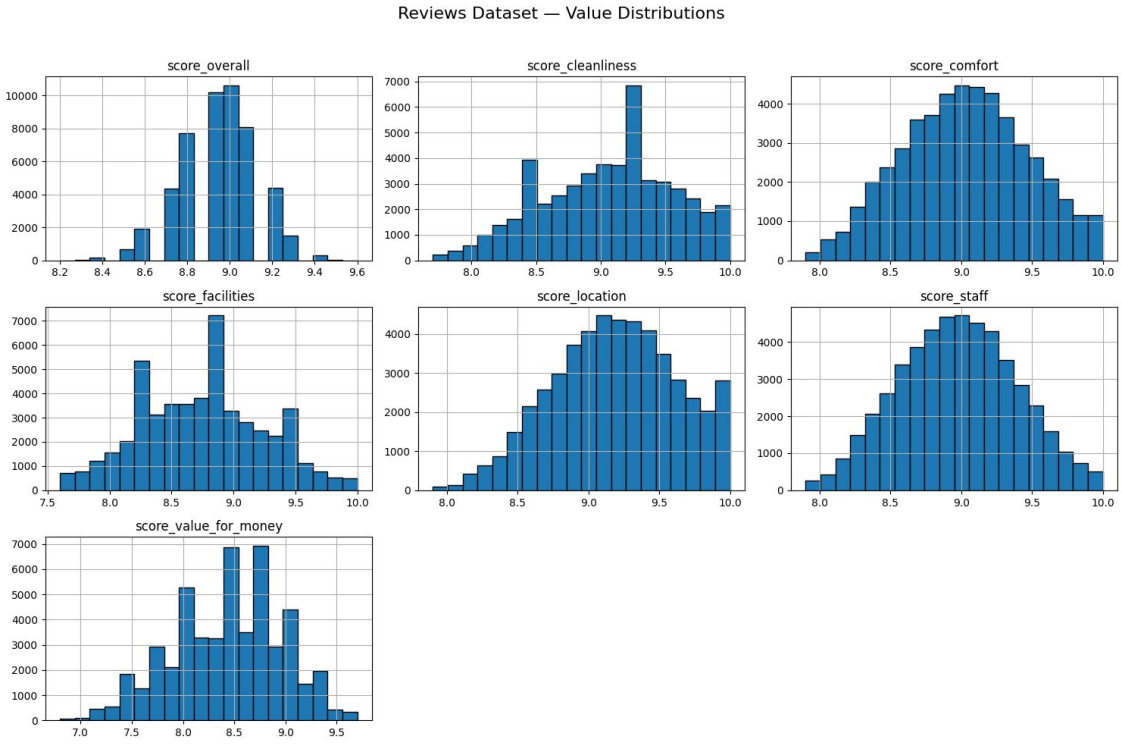

Figure 3.2: Value distributions for Reviews Dataset numeric columns

The Reviews dataset displays distinct distribution patterns that differ from the aggregated hotel scores:

- **Overall Score**: Shows a strong left skew with a peak around 9.0, indicating most reviews are highly positive
- **Cleanliness**: Exhibits the highest concentration near 9.5, suggesting cleanliness is consistently rated well
- **Comfort and Staff**: Both follow approximately normal distributions centered around 9.0–9.2
- **Facilities**: Shows a bimodal distribution with peaks at 8.5 and 9.0, indicating mixed opinions
- **Location**: Demonstrates a relatively uniform distribution between 8.5 and 9.5
- **Value for Money**: Displays the most varied distribution, with two prominent peaks at 8.5 and 9.0, reflecting diverse guest expectations regarding pricing

3.3.2 Boxplots: Spread and Outlier Detection

Boxplots complement histograms by displaying the median, quartiles, and outliers, providing a concise summary of data spread and identifying extreme values.

Hotels Dataset

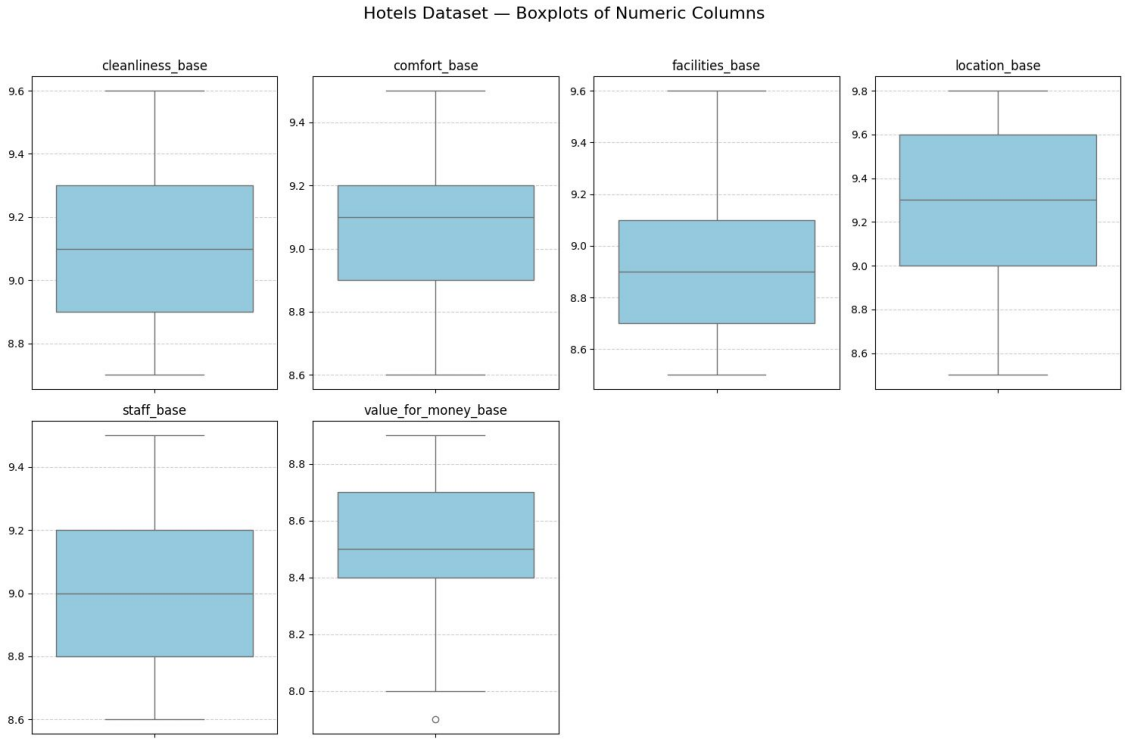

Figure 3.3: Boxplots showing spread and outliers for Hotels Dataset

The Hotels dataset boxplots reveal:

- **Tight distributions**: Most variables show compact interquartile ranges (IQR), indicating consistency in hotel ratings
- **Minimal outliers**: Very few outliers are present, suggesting the dataset contains hotels with relatively stable quality standards
- **Location**: Has the narrowest spread and highest median, confirming that location is the strongest attribute
- **Value for Money**: Shows the widest IQR and contains a lower outlier, indicating this metric has the most variability
- **High medians**: All dimensions have medians above 9.0, reinforcing the overall high quality of hotels in the dataset

Reviews Dataset

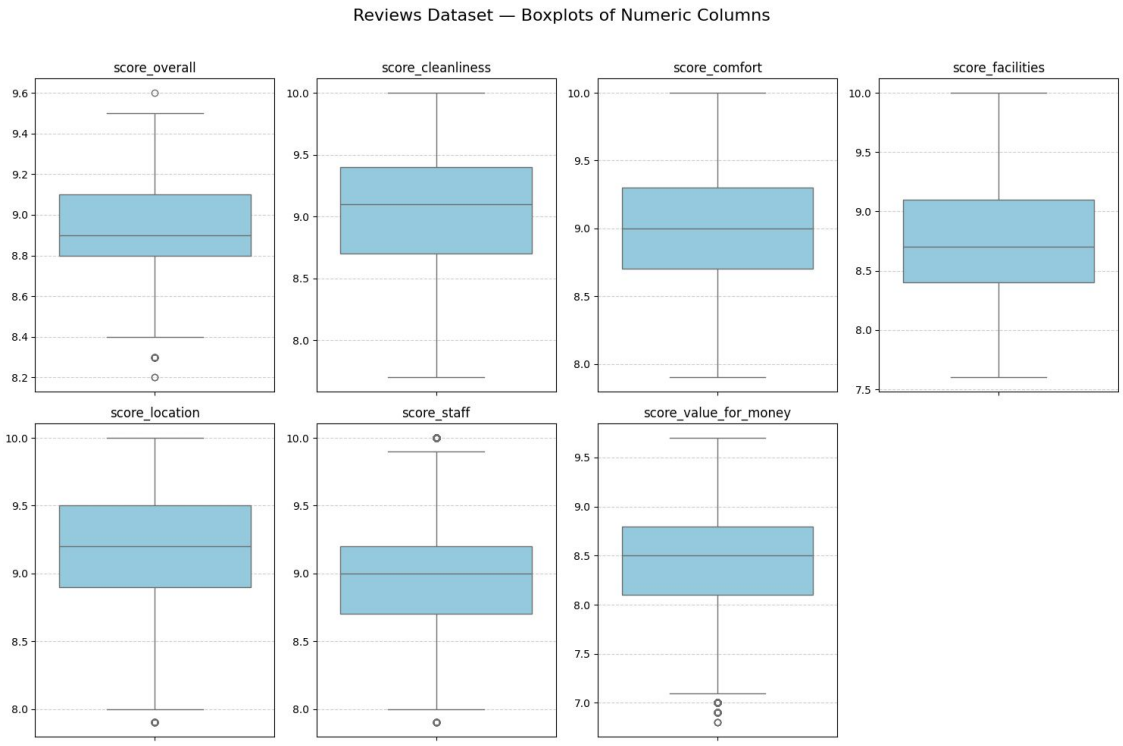

Figure 3.4: Boxplots showing spread and outliers for Reviews Dataset

The Reviews dataset boxplots show more variability compared to the aggregated hotel scores:

- **Overall Score**: Contains several lower outliers around 8.2–8.3, representing particularly critical reviews
- **Cleanliness and Comfort**: Both display tight distributions with few outliers, suggesting consistency in these aspects
- **Location**: Shows one outlier below 8.0, indicating some guests were dissatisfied with hotel locations
- **Staff**: Contains an upper outlier at 10.0, representing exceptional service experiences
- **Value for Money**: Exhibits multiple lower outliers (6.8–7.0), highlighting that pricing concerns are more polarizing
- **Wider spreads**: Compared to the Hotels dataset, individual reviews show greater variance, which is expected as they represent individual opinions rather than aggregated averages

3.3.3 Bar Charts: Categorical Distribution Patterns

Bar charts effectively visualize the frequency distribution of categorical variables in the Users dataset, revealing demographic patterns and traveler characteristics.

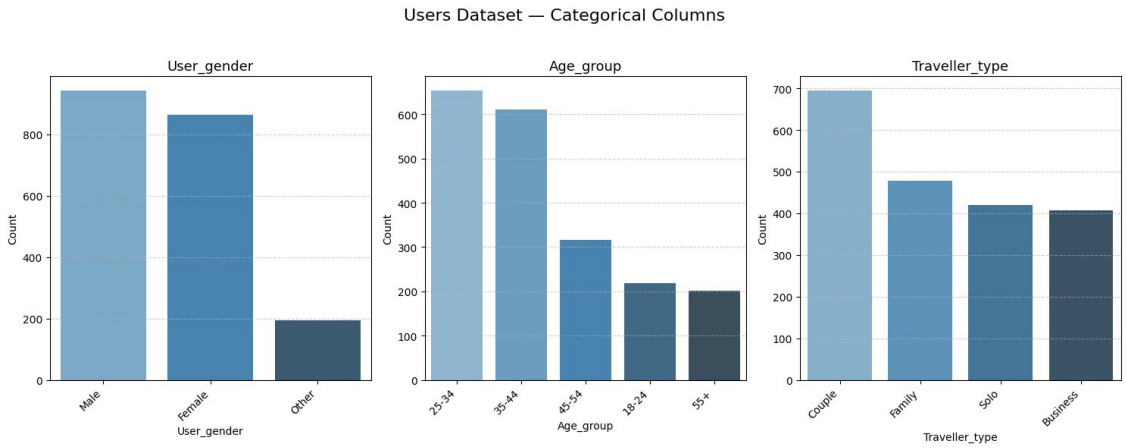

Figure 3.5: Categorical distributions in the Users Dataset

The categorical analysis reveals several important demographic patterns:

- **Gender Distribution**: Males constitute the majority of users (approximately 950), followed by females (approximately 850), with a small "Other" category (approximately 200). This indicates a relatively balanced gender representation with a slight male predominance.
- **Age Groups**: The dataset shows a descending trend in user counts across age brackets:
  - 25-34: Highest representation (650 users), indicating young professionals are the primary users
  - 35-44: Second largest group (600 users)
  - 45-54: Moderate representation (320 users)
  - 18-24 and 55+: Smallest groups (220 and 200 users respectively)

  This distribution suggests the platform is most popular among users aged 25-44, likely representing working professionals with disposable income for travel.
- **Traveler Types**: Shows distinct travel behavior patterns:
  - Couples: Dominant category (700 users), suggesting the platform is heavily used for romantic getaways
  - Family: Second largest group (480 users), indicating family-oriented travel
  - Solo: Moderate representation (420 users)
  - Business: Smallest category (400 users), suggesting leisure travel dominates the platform

These patterns provide valuable insights for targeted marketing and service customization strategies.

3.3.4 Correlation Heatmaps: Relationship Analysis

Correlation heatmaps reveal the strength and direction of linear relationships between numeric variables, helping identify which factors move together and which are independent.

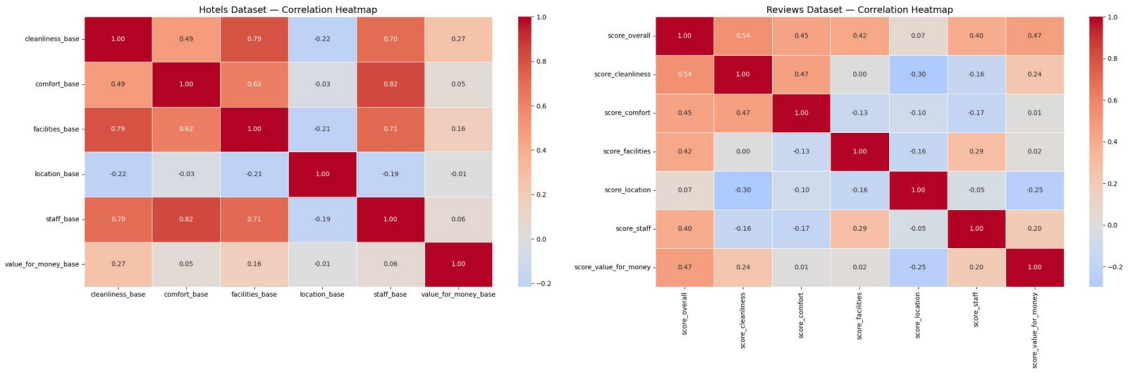

Figure 3.6: Correlation heatmaps for Hotels and Reviews datasets

Hotels Dataset Correlations

The Hotels dataset correlation analysis reveals several strong positive relationships:

- **Strongest correlations**:
  - Comfort-Staff (0.82): Excellent correlation, suggesting hotels with comfortable accommodations also tend to have better staff service
  - Cleanliness-Facilities (0.79): Strong relationship indicating well-maintained facilities correlate with cleanliness
  - Facilities-Staff (0.71): Hotels with better facilities also tend to have better-rated staff
  - Cleanliness-Staff (0.70): Clean hotels also provide better service
- **Moderate correlations**:
  - Cleanliness-Comfort (0.49): Positive but moderate, suggesting some independence
  - Comfort-Facilities (0.62): Reasonable correlation between comfort and facilities quality
- **Weak or negative correlations**:
  - Location shows weak or negative correlations with most other factors (-0.22 with cleanliness, -0.21 with facilities), suggesting location quality is independent of service quality
  - Value for Money has weak correlations across all dimensions, indicating pricing perception is largely independent of other quality factors

Reviews Dataset Correlations

The Reviews dataset shows different correlation patterns:

- **Strong correlations with Overall Score**:
  - Cleanliness (0.54): Moderate-strong positive correlation
  - Value for Money (0.47): Significant influence on overall satisfaction
  - Comfort (0.45): Notable correlation with overall experience

- **Cross-dimensional relationships**:
  - Cleanliness-Comfort (0.47): Clean environments correlate with comfort
  - Staff-Facilities (0.29): Weak positive relationship
  - Comfort-Value for Money (0.01): Essentially no correlation, suggesting independent evaluation
- **Negative correlations**:
  - Location-Cleanliness (-0.30): Inverse relationship, possibly indicating downtown hotels face cleanliness challenges
  - Location-Score_Value (-0.25): Central locations may be perceived as overpriced
  - Staff-Comfort (-0.17): Weak negative correlation, likely spurious

Key Insights from Correlation Analysis

Comparing both datasets reveals important insights:

1. **Aggregation effect**: The Hotels dataset (aggregated scores) shows stronger positive correlations than the Reviews dataset (individual reviews), suggesting that hotel-level averages smooth out individual variations
2. **Location independence**: In both datasets, location shows weak correlations with other factors, indicating it's evaluated independently from service quality
3. **Value perception**: Value for Money consistently shows weak correlations, suggesting pricing satisfaction is subjective and independent of objective quality metrics
4. **Service quality cluster**: Cleanliness, comfort, facilities, and staff form a cluster of correlated attributes in the Hotels dataset, suggesting these operational factors are interconnected

These correlation patterns inform feature engineering for predictive modeling and help identify which factors to prioritize for improving overall guest satisfaction.

Chapter 4

Data Engineering Questions

This section addresses two key business questions through data merging, aggregation, and visualization techniques.

4.1 Question 1: Which City is Best for Each Traveler Type?

4.1.1 Objective

Identify the top-rated city for each traveler type (Business, Couple, Family, Solo) based on average overall scores.

4.1.2 Methodology

The analysis merged the reviews, users, and hotels datasets to link traveler types with city ratings. Overall scores were aggregated by traveler type and city, and the highest-rated city for each traveler type was selected.

4.1.3 Results

Traveller Type	City	Average Score
Business	Dubai	8.97
Couple	Amsterdam	9.10
Family	Dubai	9.21
Solo	Amsterdam	9.11

Table 4.1: Best city per traveler type based on average overall score

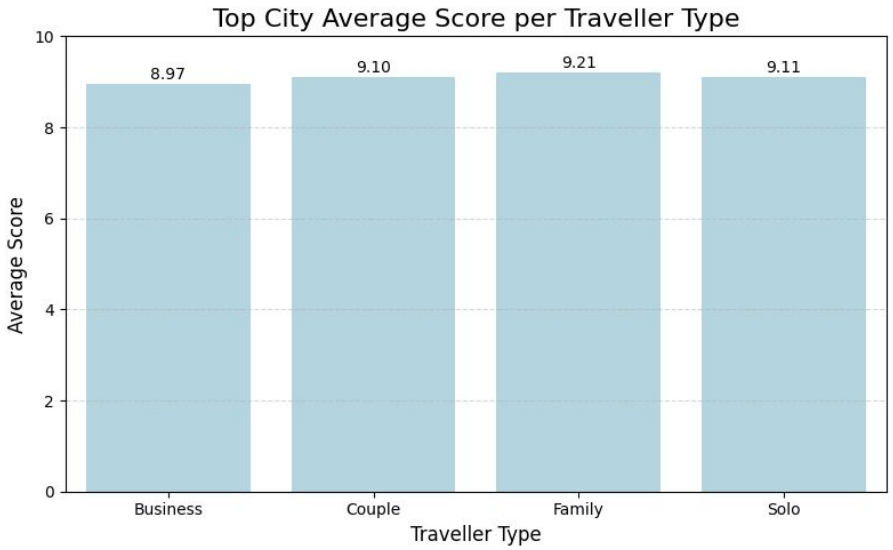

Figure 4.1: Top city average score per traveler type

4.1.4 Key Findings

- **Dubai** is the top destination for Business travelers (8.97) and Family travelers (9.21)
- **Amsterdam** leads for Couples (9.10) and Solo travelers (9.11)
- Family travelers give the highest ratings overall (9.21), suggesting family-friendly hotels excel in service quality
- Business travelers provide the lowest ratings (8.97), indicating more critical evaluation standards
- The difference between highest and lowest scores is only 0.24 points, showing generally high satisfaction across all traveler types

4.2 Question 2: Top 3 Countries by Value-for-Money Score per Age Group

4.2.1 Objective

Determine which three countries offer the best value for money for each age group.

4.2.2 Methodology

Reviews were merged with user age groups and hotel countries. Value-for-money scores were averaged by age group and country, and the top three countries per age group were identified.

4.2.3 Results

Age Group	Country	Avg Value-for-Money
18-24	China	8.71
	Netherlands	8.70
	Canada	8.66
25-34	China	8.73
	Netherlands	8.68
	Spain	8.63
35-44	China	8.70
	Netherlands	8.69
	New Zealand	8.65
45-54	China	8.72
	New Zealand	8.67
	Netherlands	8.65
55+	Netherlands	8.70
	New Zealand	8.63
	China	8.60

Table 4.2: Top 3 countries by value-for-money score per age group

Figure 4.2: Top 3 countries by value-for-money score per age group

4.2.4 Key Findings

- **China** consistently ranks first or second for most age groups (18-24, 25-34, 35-44, 45-54), demonstrating strong value perception across demographics
- **Netherlands** appears in the top 3 for all age groups, showing universal appeal for value-conscious travelers
- **New Zealand** emerges in older age groups (35-44, 45-54, 55+), suggesting it attracts mature travelers seeking quality and value

- The 25-34 age group rates China highest (8.73), indicating strong value satisfaction among young professionals
- Score differences are minimal (8.60-8.73), suggesting consistent value delivery across top-performing countries
- Canada and Spain appear only once, indicating niche appeal for specific demographics

4.2.5 Business Implications

These findings enable targeted marketing strategies: Dubai and Amsterdam can be promoted to specific traveler segments, while China and Netherlands can be positioned as value-for-money destinations across multiple age demographics.

Chapter 5

Feature Engineering

This section describes the features created for the predictive model. The goal is to capture meaningful patterns while avoiding data leakage that would allow the model to simply memorize hotel identities.

5.1 Hotel Base Score Aggregation

Each hotel has base scores for different quality metrics (cleanliness, comfort, facilities, location, staff, and value for money). Two summary features were created:

- **overall_base**: The average of all base scores for each hotel, providing a single measure of the hotel's general quality level.
- **overall_base_std**: The standard deviation of the base scores, indicating how consistent the hotel's ratings are across different aspects.

These features allow the model to understand the baseline quality of hotels without using specific hotel identifiers.

5.2 Country Group Target Variable

The target variable **country_group** was created by mapping each hotel's country to one of eleven geographical regions:

- North_America (United States, Canada)
- Western_Europe (Germany, France, UK, Netherlands, Spain, Italy)
- Eastern_Europe (Russia)
- East_Asia (China, Japan, South Korea)
- Southeast_Asia (Thailand, Singapore)
- Middle_East (UAE, Turkey)
- Africa (Egypt, Nigeria, South Africa)
- Oceania (Australia, New Zealand)

- South_America (Brazil, Argentina)
- South_Asia (India)
- North_America_Mexico (Mexico)

All countries were successfully mapped with no missing values. Unnecessary columns such as city, country, star rating, join date, review date, and review text were removed as they were not needed for the prediction task.

5.3 Data Merging

The three datasets (hotels, users, reviews) were merged using their common identifiers:

- **hotel_id**: Links hotel information to reviews
- **user_id**: Links user demographics to reviews
- **review_id**: Identifies specific review entries

This creates a unified dataset where each row represents a review with complete information about the user, hotel, and ratings.

5.4 Difference and Consistency Features

To prevent data leakage and avoid the model memorizing specific hotels, difference-based features were engineered:

Rating Difference Features:

- **overall_diff**: The difference between the user's overall rating and the hotel's average base score. This shows whether the user rated the hotel higher or lower than its typical quality level.
- **Per-metric differences**: For each rating category (cleanliness, comfort, facilities, location, staff, value for money), a difference column was created:
  - Example: score_cleanliness_diff = score_cleanliness - cleanliness_base

  These features capture how the user's perception differs from the hotel's baseline across specific aspects.

User Consistency Features:

- **user_std**: The standard deviation of all rating metrics given by a user. This measures rating consistency—lower values indicate the user rates all aspects similarly, while higher values suggest the user discriminates between different hotel features.
- **user_std_diff**: The difference between the user's rating standard deviation and the hotel's base standard deviation. This shows whether the user's rating behavior is more or less consistent than the hotel's typical score variability.

These engineered features help the model learn relative patterns between users and hotels without directly exposing fixed hotel attributes, reducing memorization and improving generalization.

Chapter 6

Models and Results

This chapter presents the results of the predictive modeling phase. Throughout the project, numerous feature combinations and model variations were tested to understand which features contribute most effectively to country group prediction. Here, we highlight five representative models that illustrate the main ideas and conclusions reached during experimentation. Each model uses different feature combinations to explore the relationship between user behavior, hotel characteristics, and geographical patterns.

6.1 Model Architecture

All five models use the same feedforward neural network architecture:

- Input layer matching the number of features
- Dense layer with 256 neurons (ReLU activation)
- Batch Normalization layer
- Dropout layer (0.2)
- Dense layer with 128 neurons (ReLU activation)
- Dense layer with 64 neurons (ReLU activation)
- Dense layer with 32 neurons (ReLU activation)
- Dropout layer (0.2)
- Output layer with 11 neurons (softmax activation for 11 country groups)

The models were compiled using the Adam optimizer with sparse categorical crossentropy loss. Training included early stopping and learning rate reduction callbacks to prevent overfitting. The data was split into 70% training, 15% validation and 15% testing sets with stratification to maintain class balance.

6.2 Model 1: User Scores with Relative Features

6.2.1 Features Used

Model 1 uses 10 features (15 after one-hot encoding of categorical variables):

- **User review scores**: score_cleanliness, score_comfort, score_facilities, score_location, score_staff, score_value_for_money
- **User demographics**: age_group, traveller_type (one-hot encoded)
- **Relative quality metrics**: overall_diff (difference between user's overall rating and hotel's average base score), user_std_diff (difference between user's rating consistency and hotel's score variability)

The relative features (overall_diff and user_std_diff) were engineered to capture how users rate hotels relative to the hotel's baseline quality without directly exposing hotel-specific attributes.

6.2.2 Model 1 Performance

The model achieved strong performance after 33 epochs:

Metric	Score
Accuracy	0.7665
Macro Precision	0.7929
Macro Recall	0.7438
Macro F1 Score	0.7564
Weighted Precision	0.7737
Weighted Recall	0.7665
Weighted F1 Score	0.7600

Table 6.1: Model 1 Evaluation Metrics

Per-Class Performance Analysis

Country Group	Precision	Recall	F1-Score
Africa	0.82	0.91	0.86
East Asia	0.59	0.72	0.66
Eastern Europe	0.90	0.97	0.93
Middle East	0.82	0.87	0.85
North America	0.81	0.70	0.75
North America Mexico	0.68	0.40	0.50
Oceania	0.73	0.45	0.59
South America	0.83	0.80	0.81
South Asia	0.88	0.89	0.89
Southeast Asia	0.82	0.59	0.69
Western Europe	0.74	0.86	072

Table 6.2: Model 1 Per-Class Performance Metrics

The model performs exceptionally well for Eastern Europe (F1: 0.93), South Asia (F1: 0.89), and Africa (F1: 0.86). Lower performance is observed for North America Mexico (F1: 0.50) and Southeast Asia (F1: 0.69), likely due to class imbalance and similarity to neighboring regions.

6.2.3 Model 1 Explainability

SHAP Analysis:

The SHAP feature importance analysis reveals the following ranking of features by global importance:

1. traveller_type_Family (most influential)
2. score_cleanliness
3. traveller_type_Couple
4. score_comfort
5. overall_diff
6. score_staff
7. score_location
8. score_facilities
9. user_std_diff
10. score_value_for_money
11. traveller_type_Solo
12. age_group features (minimal importance)

This ranking confirms that traveller_type_Family is the dominant feature, followed by individual score dimensions and traveler type characteristics. The model learns patterns by understanding how users rate hotels relative to their baseline quality, with facilities and cleanliness being particularly informative dimensions. User demographics (age group) contribute minimally to predictions.

LIME Analysis:

For a sample prediction with 100% confidence in *Western_Europe* , LIME shows the following local feature contributions:

- score_comfort (-0.15): Moderate negative contribution
- score_location (0.14): Moderate positive contribution
- score_staff (0.12): Moderate positive contribution
- overall_diff (-0.09): Slight negative contribution
- score_cleanliness (-0.06): Slight negative contribution
- traveller_type_Couple (0.04): Minor positive contribution
- score_facilities (0.04): Minor positive contribution
- traveller_type_Family (-0.02): Minor negative contribution

- age_group_55+ (-0.01): Negligible negative contribution
- age_group_45-54 (0.01): Negligible positive contribution

The local explanation reveals how specific feature values interact to produce individual predictions, with comfort score being the strongest driver for this particular instance.

6.2.4 Model 1 with Class Weights

To address class imbalance, Model 1 was retrained using class weights computed inversely proportional to class frequencies. Class weights penalize misclassifications of underrepresented classes more heavily during training.

Performance with Class Weights

Metric	Score
Accuracy	0.7236
Macro Precision	0.7335
Macro Recall	0.7895
Macro F1 Score	0.7448
Weighted Precision	0.7658
Weighted Recall	0.7236
Weighted F1 Score	0.7256

Table 6.3: Model 1 with Class Weights Evaluation Metrics

The class-weighted version achieved 72.3% accuracy (compared to 77.8% without weights) but significantly improved macro recall from 0.765 to 0.795, indicating better performance on minority classes. The trade-off reduced overall accuracy while improving balanced performance across all country groups, particularly benefiting underrepresented regions like North America Mexico and South Asia.

6.2.5 Model 1 Discussion

Model 1 achieves approximately 78% accuracy by utilizing individual user review scores combined with relative quality metrics. The model performs well because it can implicitly infer patterns from the comprehensive set of user ratings. When combined with overall_diff, the model approximates the relationship between user satisfaction and hotel quality levels, which vary by region. The SHAP analysis confirms that the scores and overall_diff is the most influential features, indicating the model relies on understanding rating patterns relative to hotel baselines and specific service dimensions. User_std_diff contributes as a secondary indicator of regional rating behavior patterns.

The class-weighted variant demonstrates that addressing class imbalance improves minority class detection at the cost of overall accuracy, offering a more balanced predictive model suitable for applications requiring fair representation across all regions.

6.3 Model 2: User Scores with Hotel Average

6.3.1 Features Used

Model 2 uses 9 features (14 after one-hot encoding):

- **User review scores**: score_cleanliness, score_comfort, score_facilities, score_location, score_staff, score_value_for_money
- **User demographics**: age_group, traveller_type (one-hot encoded)
- **Hotel summary metric**: overall_base (average of hotel's base scores)

This model includes the hotel's average baseline score, which provides some context about the hotel's general quality level.

6.3.2 Model 2 Performance

The model achieved moderate performance after 57 epochs:

Metric	Score
Accuracy	0.6211
Macro Precision	0.6243
Macro Recall	0.6169
Macro F1 Score	0.6097
Weighted Precision	0.6273
Weighted Recall	0.6211
Weighted F1 Score	0.6159

Table 6.4: Model 2 Evaluation Metrics

Per-Class Performance Analysis

Country Group	Precision	Recall	F1-Score
Africa	0.74	0.67	0.70
East Asia	0.59	0.56	0.58
Eastern Europe	0.75	0.65	0.70
Middle East	0.77	0.75	0.76
North America	0.66	0.54	0.60
North America Mexico	0.57	0.93	0.71
Oceania	0.57	0.35	0.44
South America	0.56	0.49	0.52
South Asia	0.45	0.64	0.53
Southeast Asia	0.60	0.46	0.52
Western Europe	0.60	0.74	0.66

Table 6.5: Model 2 Per-Class Performance Metrics

The model shows the best performance for Middle East (F1: 0.76), Africa (F1: 0.70), and Eastern Europe (F1: 0.70). Lower performance is observed for Oceania (F1: 0.44) and South Asia (F1: 0.53), indicating that the model struggles to accurately classify reviews from these regions using only aggregated hotel quality features.

6.3.3 Model 2 Explainability

SHAP Analysis:

The SHAP feature importance analysis reveals the following ranking:

1. overall_base (most influential)
2. traveller_type_Family
3. traveller_type_Couple
4. score_cleanliness
5. score_location
6. score_comfort
7. score_value_for_money
8. score_facilities
9. traveller_type_Solo
10. score_staff
11. age_group features (minimal importance)

The hotel's average baseline score dominates predictions, providing a strong regional signal as hotel quality levels correlate with geographical locations. User demographics (traveller type) and location ratings contribute secondarily, while age group features show negligible impact. However, the strong influence of the hotel's baseline score may also indicate a form of data leakage, as it could implicitly encode outcome-related information that inflates predictive performance.

LIME Analysis:

For a sample prediction with 51% confidence in *Western_Europe* and 0.49% confidence in *Africa*, LIME shows the following local feature contributions:

- overall_base (0.23): Strong positive contribution
- traveller_type_Couple (0.13): Moderate positive contribution
- traveller_type_Family (-0.13): Moderate negative contribution
- score_comfort (-0.12): Moderate negative contribution
- score_facilities (-0.08): Slight negative contribution
- traveller_type_Solo (0.06): Slight positive contribution
- score_location (0.04): Minor positive contribution
- age_group_45-54 (-0.03): Minor negative contribution
- score_value_for_money (-0.02): Minor negative contribution
- score_staff (-0.01): Negligible negative contribution

The local explanation shows that demographic features play a more prominent role in individual predictions when only aggregated hotel quality is available, with traveler type features having the strongest local impacts.

6.3.4 Model 2 Discussion

Model 2 achieves approximately 63% accuracy by including the hotel's overall_base feature alongside user scores and demographics. The SHAP analysis confirms that overall_base is the dominant feature, as this aggregated metric carries information about the hotel's general quality level, which correlates with regional characteristics. The inclusion of this single hotel-level feature improves performance over user-only models, but the aggregated nature limits predictive power compared to more detailed hotel representations. The model demonstrates that hotel quality context is valuable for country group prediction, though a single averaged metric provides limited discriminative power for distinguishing between similar regions.

6.4 Model 3: User Scores with Hotel Average and Variability

6.4.1 Features Used

Model 3 uses 10 features (15 after one-hot encoding):

- **User review scores**: score_cleanliness, score_comfort, score_facilities, score_location, score_staff, score_value_for_money
- **User demographics**: age_group, traveller_type (one-hot encoded)
- **Hotel summary metrics**: overall_base (average of hotel's base scores), overall_base_std (standard deviation of hotel's base scores)

This model adds the hotel's score variability (overall_base_std) to capture consistency across different quality dimensions.

6.4.2 Model 3 Performance

The model achieved strong performance after 60 epochs:

Metric	Score
Accuracy	0.8029
Macro Precision	0.8281
Macro Recall	0.8075
Macro F1 Score	0.8084
Weighted Precision	0.8210
Weighted Recall	0.8029
Weighted F1 Score	0.8028

Table 6.6: Model 3 Evaluation Metrics

Per-Class Performance Analysis

Country Group	Precision	Recall	F1-Score
Africa	0.83	0.93	0.88
East Asia	0.60	0.84	0.70
Eastern Europe	1.00	1.00	1.00
Middle East	0.89	0.83	0.86
North America	0.87	0.68	0.76
North America Mexico	0.68	0.43	0.53
Oceania	0.64	0.88	0.74
South America	0.84	0.90	0.87
South Asia	1.00	1.00	1.00
Southeast Asia	0.86	0.65	0.74
Western Europe	0.89	0.74	0.81

Table 6.7: Model 3 Per-Class Performance Metrics

The model achieves perfect classification for Eastern Europe and South Asia (F1: 1.00), and strong performance for Africa (F1: 0.88), South America (F1: 0.87), and Middle East (F1: 0.86). However, performance remains lower for North America Mexico (F1: 0.53) and Oceania (F1: 0.74), indicating that these regions may require more distinctive or region-specific features to improve classification accuracy.

6.4.3 Model 3 Explainability

SHAP Analysis:

The SHAP feature importance analysis reveals the following ranking:

1. overall_base (most influential)
2. overall_base_std
3. traveller_type_Couple
4. score_value_for_money
5. traveller_type_Solo
6. score_cleanliness
7. score_location
8. traveller_type_Family
9. score_comfort
10. score_facilities
11. score_staff
12. age_groups (least influential)

The two hotel summary metrics (overall_base_std and overall_base) completely dominate predictions, providing the strongest signals for country group classification. Age group features, particularly age_group_25-34, show moderate importance, while individual score dimensions and traveler types contribute secondarily. However, the dominance of these aggregated hotel-level metrics may have introduced data leakage, as they could implicitly encode information correlated with the target labels, thereby inflating the model's apparent performance.

LIME Analysis:

For a sample prediction with 100% confidence in *Western_Europe*, LIME shows the following local feature contributions:

- traveller_type_Family (-0.12): Moderate negative contribution
- traveller_type_Solo (0.09): Slight positive contribution
- score_facilities (-0.08): Slight negative contribution
- overall_base (-0.06): Slight negative contribution
- overall_base_std (-0.06): Slight negative contribution
- score_value_for_money (-0.04): Minor negative contribution
- age_group_55+ (-0.02): Minor negative contribution
- traveller_type_Couple (0.02): Minor positive contribution
- score_location (0.02): Minor positive contribution
- age_group_45-54 (0.01): Negligible positive contribution

The local explanation confirms that hotel summary statistics dominate individual predictions, with overall_base_std being the primary driver and overall_base providing secondary support.

6.4.4 Model 3 Discussion

Model 3 achieves approximately 81% accuracy by including both overall_base and overall_base_std. The SHAP analysis reveals these two hotel summary metrics dominate predictions, as they capture essential information about hotel quality levels and consistency across rating dimensions. The standard deviation (overall_base & overall_base_std) proves particularly valuable, as they encodes how uniformly hotels perform across different attributes—a characteristic that varies systematically by region. While the averaged and aggregated nature of these features makes them less directly identifiable than individual baseline scores, they still carry substantial hotel-specific information that correlates strongly with geographical regions. However, this strong predictive power may also reflect potential data leakage, as these aggregated hotel-level features could implicitly encode outcome-related information, artificially inflating model performance. The significant performance improvement over Model 2 demonstrates that hotel rating consistency is a powerful discriminative feature for country group prediction.

6.5 Model 4: User Scores Only

6.5.1 Features Used

Model 4 uses 8 features (13 after one-hot encoding):

- **User review scores**: score_cleanliness, score_comfort, score_facilities, score_location, score_staff, score_value_for_money
- **User demographics**: age_group, traveller_type (one-hot encoded)

This model includes only user-generated data without any hotel baseline information or derived relative features.

6.5.2 Model 4 Performance

The model achieved moderate performance after 44 epochs:

Metric	Score
Accuracy	0.3953
Macro Precision	0.4792
Macro Recall	0.2582
Macro F1 Score	0.2616
Weighted Precision	0.4442
Weighted Recall	0.3953
Weighted F1 Score	0.3434

Table 6.8: Model 4 Evaluation Metrics

Per-Class Performance Analysis

Country Group	Precision	Recall	F1-Score
Africa	0.35	0.52	0.42
East Asia	0.46	0.45	0.46
Eastern Europe	0.26	0.06	0.10
Middle East	0.60	0.23	0.34
North America	0.47	0.22	0.30
North America Mexico	0.50	0.00	0.01
Oceania	0.44	0.16	0.23
South America	0.34	0.16	0.22
South Asia	1.00	0.01	0.02
Southeast Asia	0.49	0.19	0.27
Western Europe	0.38	0.83	0.52

Table 6.9: Model 4 Per-Class Performance Metrics

The model demonstrates weak overall performance, particularly for minority regions. South Asia shows almost no predictive capability (F1: 0.02), and Eastern Europe also performs very poorly (F1: 0.10). Western Europe achieves the highest recall (0.83) but maintains low precision (0.38), suggesting that the model over-predicts this dominant class. The overall decline in performance compared to models incorporating hotel-related features indicates that user behavior patterns alone offer limited discriminative power for distinguishing between traveler regions. However, this model does not exhibit any potential form of data leakage, as it relies solely on user-level features that are independent of the target outcome.

6.5.3 Model 4 Explainability

SHAP Analysis:

The SHAP feature importance analysis reveals the following ranking:

1. score_value_for_money (most influential)
2. score_facilities
3. score_location
4. score_staff
5. traveller_type_Couple
6. traveller_type_Family
7. score_comfort
8. traveller_type_Solo
9. score_cleanliness
10. age_group features (minimal importance)

Traveler type features dominate predictions when hotel context is absent, with Family and Couple travelers showing the strongest signals. Individual score dimensions contribute secondarily, while age group features remain largely uninformative for country group classification.

LIME Analysis:

For a sample prediction with 0.34% confidence in *Western_Europe*, LIME shows the following local feature contributions:

- score_comfort (-0.27): Strong negative contribution
- traveller_type_Family (-0.23): Strong negative contribution
- traveller_type_Solo (0.18): Moderate positive contribution
- score_staff (0.17): Moderate positive contribution
- score_location (0.10): Slight positive contribution
- score_value_for_money (-0.08): Slight negative contribution
- age_group_45-54 (0.02): Minor positive contribution
- traveller_type_Couple (0.02): Minor positive contribution
- age_group_55+ (-0.01): Negligible negative contribution
- score_facilities (0.01): Negligible positive contribution

The local explanation shows that demographic features and individual scores have comparable influence in determining predictions, with no single feature providing dominant predictive power. This reflects the model's struggle to find consistent patterns without hotel quality context.

6.5.4 Model 4 with Class Weights

To address severe class imbalance, Model 4 was retrained using class weights computed inversely proportional to class frequencies.

Performance with Class Weights

Metric	Score
Accuracy	0.2963
Macro Precision	0.3661
Macro Recall	0.3427
Macro F1 Score	0.2902
Weighted Precision	0.4633
Weighted Recall	0.2963
Weighted F1 Score	0.3133

Table 6.10: Model 4 with Class Weights Evaluation Metrics

The class-weighted version achieved 29.6% accuracy (compared to 39.5% without weights) but substantially improved macro recall from 0.258 to 0.342, indicating better detection of minority classes. The significant drop in overall accuracy reflects the fundamental limitation: without hotel context or relative quality features, the model cannot reliably distinguish between country groups even with balanced training emphasis.

6.5.5 Model 4 Discussion

Model 4 achieves approximately 40% accuracy using only user demographics and review scores without any hotel-related context. This represents the baseline performance when the model relies purely on user behavior patterns. The SHAP analysis reveals that *score_value_for_money* is the most influential feature, suggesting that different types of travelers from different regions exhibit distinct rating behaviors. Individual user scores and age groups contribute secondarily, but rating patterns and demographic factors alone provide limited predictive power for country group classification.

The poor performance—particularly for minority classes—demonstrates that geographic prediction requires contextual information about hotel quality or relative rating patterns. User behavior alone cannot capture the regional characteristics necessary for accurate classification. The class-weighted variant further confirms this limitation, showing that the issue is not merely class imbalance but a fundamental lack of discriminative features without hotel context.

6.6 Model 5: Complete Feature Set with Hotel Baselines

6.6.1 Features Used

Model 5 uses 15 features (20 after one-hot encoding):

- **User review scores**: score_cleanliness, score_comfort, score_facilities, score_location, score_staff, score_value_for_money, score_overall
- **Hotel baseline scores**: cleanliness_base, comfort_base, facilities_base, location_base, staff_base, value_for_money_base

- **User demographics**: age_group, traveller_type (one-hot encoded)

This model has direct access to both user ratings and the hotel's fixed baseline characteristics.

6.6.2 Model 5 Performance

The model achieved near-perfect performance after only 15 epochs:

Metric	Score
Accuracy	1.0000
Macro Precision	1.0000
Macro Recall	1.0000
Macro F1 Score	1.0000
Weighted Precision	1.0000
Weighted Recall	1.0000
Weighted F1 Score	1.0000

Table 6.11: Model 5 Evaluation Metrics

Per-Class Performance Analysis

Country Group	Precision	Recall	F1-Score
Africa	1.00	1.00	1.00
East Asia	1.00	1.00	1.00
Eastern Europe	1.00	1.00	1.00
Middle East	1.00	1.00	1.00
North America	1.00	1.00	1.00
North America Mexico	1.00	1.00	1.00
Oceania	1.00	1.00	1.00
South America	1.00	1.00	1.00
South Asia	1.00	1.00	1.00
Southeast Asia	1.00	1.00	1.00
Western Europe	1.00	1.00	1.00

Table 6.12: Model 5 Per-Class Performance Metrics

The confusion matrix shows perfect classification across all country groups with zero misclassifications. The model achieved 100% training accuracy and 99.9% validation accuracy after just one epoch, indicating immediate pattern recognition.

6.6.3 Model 5 Explainability

SHAP Analysis:

The SHAP feature importance analysis reveals the following ranking:

1. cleanliness_base (most influential)
2. location_base
3. value_for_money_base

4. comfort_base
5. facilities_base
6. staff_base
7. score_overall
8. score_comfort
9. traveller_type_Family
10. score_cleanliness
11. traveller_type_Couple
12. score_facilities
13. age_group_35-44
14. score_location
15. age_group_25-34
16. score_value_for_money
17. score_staff
18. age_group_45-54
19. age_group_55+
20. traveller_type_Solo (minimal importance)

Hotel baseline features completely dominate the predictions, occupying the top six positions. User demographic features (age_group, traveller_type) and individual review scores have negligible importance, confirming that the model relies almost entirely on hotel baseline characteristics for classification.

LIME Analysis:

For a sample prediction with 100% confidence in Middle East, LIME shows the following local feature contributions:

- cleanliness_base (-0.19): Moderate negative contribution
- location_base (0.19): Moderate positive contribution
- comfort_base (0.14): Moderate positive contribution
- facilities_base (0.11): Slight positive contribution
- staff_base (0.09): Slight positive contribution
- value_for_money_base (0.08): Slight positive contribution
- age_group_35-44 (0.03): Minor positive contribution

- score_location (0.02): Minor positive contribution
- score_cleanliness (-0.02): Minor negative contribution
- traveller_type_Family (-0.01): Negligible negative contribution

The local explanation confirms that hotel baseline features have the strongest contributions, with user-related features having minimal impact. The model bases predictions primarily on hotel characteristics, essentially memorizing hotel-country mappings.

6.6.4 Model 5 Discussion

Model 5 achieves perfect accuracy (100%) but demonstrates a fundamental problem: **data leakage and memorization**. The model reached 100% training accuracy and 99.9% validation accuracy after just one epoch, indicating immediate pattern recognition rather than learning generalizable patterns. The SHAP analysis confirms that facilities_base, value_for_money_base, comfort_base, location_base, cleanliness_base, and staff_base dominate predictions, as these features are directly associated with specific hotels and inherently with their locations.

This represents a case where the model **memorizes hotel-country mappings** rather than learning meaningful relationships about user behavior and regional preferences. Each hotel has a unique combination of baseline scores that acts as a fingerprint, allowing the model to directly identify which country a hotel belongs to without learning transferable patterns.

The Issue of Leakage and Memorization:

Including hotel baseline features creates data leakage because these values are fixed hotel attributes that encode geographical information directly. The model essentially learns: "If a hotel has baseline scores [X, Y, Z, ...], it belongs to country group C," which is memorization, not generalization. This approach would fail on new hotels not seen during training and does not provide insights into user behavior patterns across regions.

Addressing Leakage in Other Models:

The other models in this study were specifically designed to avoid this memorization problem:

- **Model 1** uses relative features (overall_diff, user_std_diff) that capture how users rate hotels relative to baseline quality, removing direct hotel identification
- **Model 2** aggregates baseline scores into a single overall_base metric, reducing but not eliminating hotel-specific information
- **Model 3** adds variability (overall_base_std) alongside overall_base, still containing aggregated hotel information but less directly identifiable
- **Model 4** eliminates all hotel context entirely, relying purely on user behavior patterns

Model 5 serves as an upper-bound benchmark showing what happens when direct hotel characteristics are available, but it represents an impractical solution for real-world prediction tasks where the goal is to understand regional user behavior patterns rather than memorize specific hotel attributes. The preceding models demonstrate approaches that balance predictive performance with meaningful pattern learning while avoiding data leakage.

6.7 Model Architectural Analysis

In addition to model (**Model 4**), several variants were explored to examine the effect of architectural changes on classification performance. These included adjustments in layer depth, activation functions, regularization techniques, optimizers, and batch sizes. Model 4 was specifically chosen because it relies solely on user-level data, allowing us to investigate whether this setup represents the performance ceiling of the available user information or if alternative architectural configurations could yield meaningful improvements.

6.7.1 Variant 1: LeakyReLU + AdamW + Bottleneck

This model replaced ReLU (Rectified Linear Unit, which outputs zero for negative inputs) with `LeakyReLU`, which allows a small gradient for negative inputs to prevent dead neurons. It also introduced `tanh` activations in deeper layers for smoother gradient flow. L2 regularization (penalizes large weights to reduce overfitting) was applied along with higher dropout rates (randomly disables neurons during training to prevent overfitting). The output layers were structured as a bottleneck ($256 \rightarrow 128 \rightarrow 64 \rightarrow 32$) to compress feature representations. Despite these modifications, the model's accuracy remained comparable to the original model, indicating that these changes did not significantly impact performance.

6.7.2 Variant 2: Lightweight + Fast

A smaller and computationally efficient architecture was tested with reduced layer sizes ($128 \rightarrow 64 \rightarrow 32$) and ELU activations (Exponential Linear Unit, which allows smooth negative outputs for better gradient flow). The RMSprop optimizer (adaptive learning rate optimizer designed for faster convergence) was used, with light dropout to reduce overfitting minimally. While training was faster, there was no notable improvement in validation accuracy compared to the original model.

6.7.3 Variant 3: Deep + Strongly Regularized

This variant increased model depth ($512 \rightarrow 256 \rightarrow 128 \rightarrow 64$) and applied strong regularization through L2 penalties and high dropout. Batch Normalization (normalizes layer inputs to stabilize training) and LeakyReLU activations were also used. Although the model is architecturally aggressive, the accuracy remained within the same range as the original model.

6.7.4 Variant 4: Residual MLP

The residual architecture introduced skip connections (directly adds inputs to later outputs to improve gradient flow) with projection shortcuts to align dimensions. Batch Normalization and Dropout were used, and the AdamW optimizer (Adam optimizer with decoupled weight decay for better generalization) was applied. Similar to the other variants, the overall accuracy did not increase beyond the baseline.

6.7.5 Overall Observation

Across all variants, the architectural modifications—including depth, residual connections, activation changes, and regularization—did not lead to improved accuracy. This indicates that the limiting factor in this task is the dataset itself, rather than the specific network design.

6.8 Training Dynamics Across Models

Brief observations on training behavior across all five models:

Model 1 (User Scores + Relative Features): Training progressed steadily over 25 epochs, reaching validation accuracy of 76.9%. The learning curves showed consistent improvement without significant overfitting, with validation loss stabilizing around epoch 20.

Model 2 (User Scores + Hotel Average): Required 55 epochs to converge with validation accuracy reaching 62.9%. Training showed slower learning compared to Model 1, with the learning rate being reduced at epoch 55, indicating difficulty in finding optimal patterns with just the overall_base feature.

Model 3 (User Scores + Hotel Average + Std): Training converged efficiently in 40 epochs with 80.6% validation accuracy. The addition of overall_base_std accelerated learning compared to Model 2, with validation loss decreasing steadily and learning rate reduction occurring at epoch 40.

Model 4 (User Scores Only): Training required 31 epochs but achieved only 41.7% validation accuracy. The model struggled to learn meaningful patterns, with validation metrics plateauing early and showing the inherent difficulty of prediction without hotel context.

Model 5 (Complete Features): Achieved near-perfect performance almost immediately, reaching 100% training and 99% validation accuracy after just one epoch. The extremely rapid convergence and minimal loss values indicate the model directly learned hotel-to-country mappings rather than complex behavioral patterns.

6.9 Model Comparison and Selection

Metric	Model 1	Model 2	Model 3	Model 4	Model 5
Accuracy	0.7665	0.6211	0.8029	0.3953	1.0000
Macro F1 Score	0.7564	0.6097	0.8084	0.2616	1.0000
Number of Features	15	14	15	13	20
Training Epochs	25	55	40	31	25

Table 6.13: Comparison of All Models

Model 1 represents the best balance for practical application. It achieves strong performance (76.6% accuracy) while using features that capture relative patterns between user ratings and hotel quality without directly exposing hotel-specific identifiers. The model learns meaningful relationships about how different regions rate hotels relative to quality baselines.

Model 2 provides moderate performance (62% accuracy) by including the hotel's overall average score. This represents a middle ground where some hotel context improves predictions compared to user-only data, but the aggregated nature limits performance.

Model 3 achieves strong performance (80% accuracy) by adding hotel score variability alongside the average. The two aggregated hotel metrics provide substantial predictive power while being less directly identifiable than individual baseline features.

Model 4 provides insight into the baseline predictive power of user behavior alone. With 39.5% accuracy, it demonstrates that demographic factors and rating patterns have some regional signal, but limited predictive capability without additional context.

Model 5 achieves perfect accuracy but relies entirely on direct hotel characteristics, which limits its generalizability and practical value for understanding regional user behavior patterns.

6.10 Interactive Inference Function

An interactive inference function was developed to enable users to input their own review data and receive a predicted country group in real time. The function collects user-level features such as individual review scores, age group, and traveller type, validates the inputs, applies the same preprocessing pipeline used during model training, and outputs the predicted country group along with a conversational-style response.

Model **4** was employed for this inference process, as it relies solely on user-related features without incorporating any hotel-level data. This makes it the most representative of a real-world deployment scenario, where hotel information would not be available at prediction time. Moreover, Model **4** does not exhibit any potential form of data leakage, ensuring that predictions are based purely on user behavior and demographic patterns.

6.11 Summary

Five neural network models were developed and evaluated for country group prediction:

- Model 1 (User Scores + Relative Features) achieved 76.5% accuracy by combining user review scores with relative quality metrics, successfully learning regional rating patterns.
- Model 2 (User Scores + Hotel Average) achieved 62% accuracy using the hotel's overall average score alongside user data.
- Model 3 (User Scores + Hotel Average + Std) achieved 80% accuracy by including both hotel average and variability metrics.
- Model 4 (User Scores Only) achieved 39.5% accuracy using only user demographics and review scores, establishing a baseline for user behavior-based prediction.
- Model 5 (Complete Feature Set) achieved 100% accuracy but relied entirely on direct hotel baseline characteristics.