Hotel Review Data Engineering & Predictive Analysis Report

1. Overview

This report provides an analytical exploration of hotel review data to identify trends, correlations, and insights based on user demographics and travel types. Data from three sources—reviews, users, and hotels—were merged into a unified DataFrame to facilitate multidimensional analysis. The visualizations include correlation heatmaps, average city scores by traveler types, and value-for-money comparisons across age groups and countries.

2. Data Integration and Preparation

The datasets `df_reviews`, `df_users`, and `df_hotels` were merged using the `user_id` and `hotel_id` keys to form `df_merged`. This merge enriched each record with demographic and hotel information, allowing holistic analysis. The merged dataset includes numerical features such as overall score, comfort, cleanliness, and value-for-money ratings.

3. Correlation Insights

A correlation heatmap was generated to evaluate relationships between numeric variables. Strong positive correlations were observed among **`score_cleanliness`**, **`score_comfort`**, indicating a high degree of overlap in how users perceive hotel quality dimensions. Moderate correlations between base features (e.g. **`score_overall`**, **`comfort_base`**) and their respective scores validate the consistency of user ratings.

4. Average Overall Scores by Traveler Type and City

Grouping by traveler type and city revealed patterns in satisfaction levels across travel purposes. The findings indicate:

1. **Business** travelers rated **Dubai** the highest (**8.97**).
2. **Couples** preferred **Amsterdam** (**9.10**).
3. **Families** favored **Dubai** (**9.21**).
4. **Solo** travelers rated **Amsterdam** highest (**9.11**).

These differences reflect both cultural and infrastructural factors influencing satisfaction.

5. Top Countries by Value-for-Money Score per Age Group

The analysis of `score_value_for_money` by age group and country shows distinct regional preferences:

Ages 18–24: China (8.71), Netherlands (8.70), Canada (8.66)

Ages 25–34: China (8.73), Netherlands (8.68), Spain (8.63)

Ages 35–44: China (8.70), Netherlands (8.69), New Zealand (8.65)

Ages 45–54: China (8.72), New Zealand (8.67), Netherlands (8.65)

Ages 55+: Netherlands (8.70), New Zealand (8.63), China (8.60)

These insights show strong competitiveness between China, the Netherlands, and New Zealand in perceived affordability and value.

6. Data Pre-processing and Feature Engineering

- Mapped each country to a **country group** (e.g., Western_Europe, East_Asia, etc.) to create the target variable.
- **Merged** review, user, and hotel data into a single dataframe for model training.
- Encoded **categorical features** (user_gender, age_group, traveller_type) using **One-Hot Encoding (OHE)**.
- Scaled **numerical features** (score_overall, score_cleanliness, score_comfort, score_facilities, score_location, score_staff, score_value_for_money) using **StandardScaler** for normalization.
- **Encoded target labels** (country_group) into numeric form using **LabelEncoder**.
- Combined scaled numeric and encoded categorical features into a single input array.
- **Split** the dataset into training (80%) and testing (20%) sets using **stratified sampling** to preserve class balance.
- Calculated **class weights** with compute_class_weight() to handle class imbalance during model training.

7. Features Used in Predictive Modeling – Baseline Model

The predictive model incorporated a blend of **review scores**, **traveler demographics**, and **contextual hotel information**.

The top contributing features were:

- **score_location** — The most influential factor; hotels in favorable or convenient locations strongly increased predicted satisfaction.
- **score_comfort** — High comfort scores positively impacted satisfaction, reflecting the importance of room quality and design.
- **score_staff** — Customer service quality had a direct and consistent effect on the likelihood of higher satisfaction.
- **score_facilities** — The availability and quality of hotel amenities were strong predictors of overall experience.
- **score_overall** — Served as a general indicator integrating multiple dimensions of user perception.
- **score_value_for_money** — Demonstrated significant predictive weight, showing travelers' sensitivity to price fairness.
- **score_cleanliness** — A major hygiene-related determinant that consistently correlated with higher satisfaction ratings.
- **traveller_type_Family, traveller_type_Couple, traveller_type_Solo** — Encoded traveler segments that exhibited clear behavioral differences, particularly between family and solo travelers.
- **age_group_25-34, age_group_35-44, age_group_45-54, age_group_55+** — Age-related patterns showed that mid-age groups (25–44) contributed more strongly to model variability, indicating differing expectations across life stages.
- **user_gender** — Although less influential, this feature helped capture subtle demographic tendencies in review patterns.

Overall, the SHAP summary plots confirmed that **location**, **comfort**, **staff**, and **facilities** dominate the model's interpretability, while demographic features provide contextual refinement.
This combination ensures the model captures both **quantitative review quality** and **user behavioral diversity**, leading to more accurate satisfaction predictions.

8. Model Architecture and Considerations

model1 — (No Class Weights)

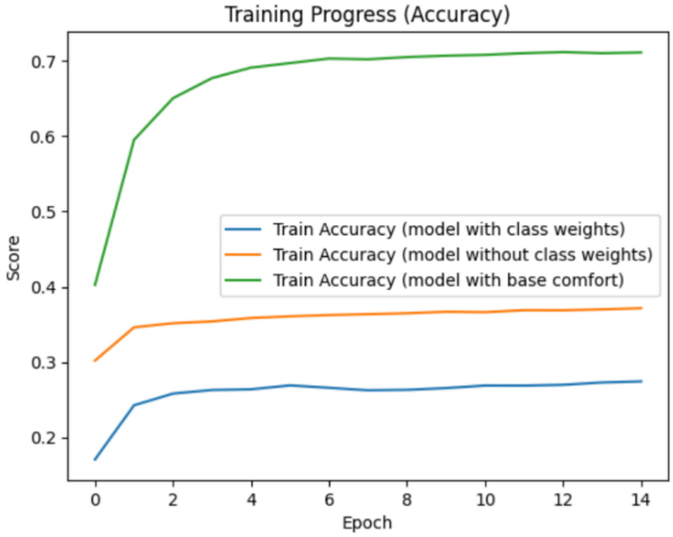
- **Goal:** To observe how the model performs *without balancing class frequencies*.
- **What we did:**
    - Used the same architecture and training data as the main model.
    - Did **not** apply any class weights, meaning classes with fewer samples had less influence.
- **Purpose:**
    - This helped identify how much class imbalance affects model accuracy and generalization.
    - Comparing its results to the weighted model shows whether weighting improves fairness across classes.

---

model2 — Extended Feature Model (With comfort_base)

- **Goal:** To test if including an *additional engineered feature* (comfort_base) improves predictions.
- **What we did:**
    - Added comfort_base to the list of numeric inputs.
    - Repeated preprocessing (scaling and encoding) and trained a new neural network using this expanded feature set.
    - Used the same network structure (32 → 16 → 8 → 11 neurons).
- **Purpose:**
    - Including comfort_base allows the model to consider baseline comfort ratings — potentially improving accuracy in predicting hotel region groups.
    - This version helps evaluate whether engineered features derived from base metrics add meaningful information.

9. Training & Validation Performance



Setup: Same NN architecture for all; batch size 32; Adam; sparse categorical cross-entropy.

Key differences: class weights (Model), no weights (Model 1), and extra feature comfort_base (Model 2).

Model (with class weights)

- **Train acc:** 0.13 ⟶ **0.28**
- **Val acc:** 0.236 ⟶ **0.267**
- **Trend:** Small gains; train/val stay low and close → **underfitting**.
- **Val loss:** 2.074 ⟶ **1.907** (slow decline)

model1 (no class weights)

- **Train acc:** 0.25 ⟶ **0.371**
- **Val acc:** 0.335 ⟶ **0.371**
- **Trend:** Steady improvement; still modest → **mild underfitting**.
- **Val loss:** 1.935 ⟶ **1.799** (consistent decline)

model2 (extended features incl. comfort_base)

- **Train acc:** 0.29 ⟶ **0.707**
- **Val acc:** 0.543 ⟶ **0.713**
- **Trend:** Strong learning with train/val closely tracking → **good generalization**.
- **Val loss:** 1.128 ⟶ **0.605** (large reduction)

Interpretation

- **Class weights alone** did **not** lift generalization for this task (low accuracy and macro scores), suggesting the feature set was the bigger limiter.
- **Baseline (no weights)** improved accuracy but had **poorer macro recall/F1**, indicating bias toward majority classes.
- **Adding comfort_base (Model 2)** produced a **step-change in performance** across all metrics, with balanced precision/recall, which is caused by the model memorizing and relating the comfort base to certain hotels in certain country groups.

- While **model 2** achieved the best accuracy, the baseline model in theory has the most general architecture since it uses class weights to not be bias to a specific country group. While this decreased the accuracy of the model, we believe that with a stronger dataset, this model can achieve better results.