NG CHEN TING [31861148]

FIT3003 Business intelligence and data warehousing

MStay Analysis for Homestay and Rental Services

Data Cleaning

Duplicated Records

Duplicates may lead inaccurate calculations. To identify them, datas are grouped by their Primary Key (PK) and the occurrences of each value is counted. By applying a filter to display rows where the PK count exceeded one, records that were duplicated can be easily identified. Duplicated records were found in **Booking and Host Table** which were removed to ensure only one original record is retained.

Null Values

Any NULL value in a primary key field disrupts the ability to uniquely identify records causing significant issues in database operations. **A NULL value** was found in the **AMENITY table**, where the description was marked as "unknown." This record was removed to maintain data consistency and ensure the integrity of the primary key.

Referential Integrity

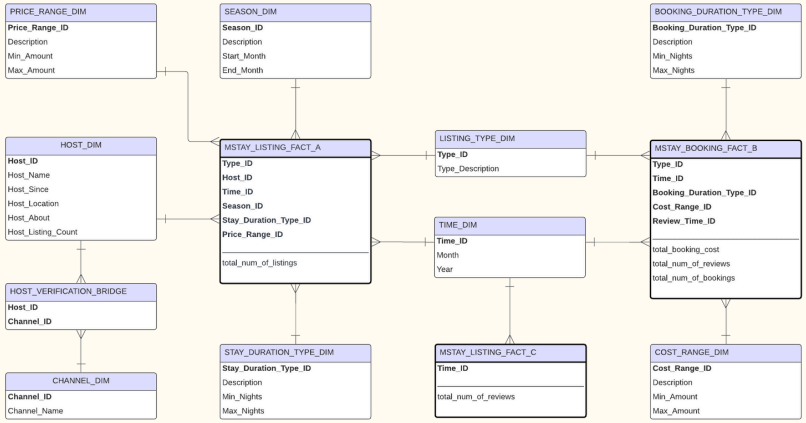
Review, Listing, and Host Verification tables contain foreign key values that do not correspond to valid primary keys in the referenced tables. **3 rows were affected** by this problem. To resolve, the affected foreign key values were first replaced with NULL values then removed in a single operation to maintain the integrity of the database after all values are checked.

Invalid Dates

Invalid dates were identified as a source of inconsistency in the dataset. **66 records** were found where the Review date was earlier than the Booking date, which is logically incorrect. These records are removed to ensure the reliability of the data.

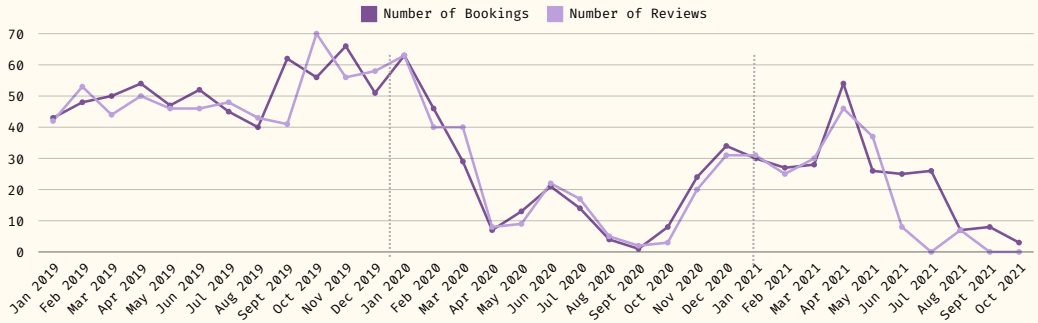
Star Schema ☆

Multi-fact table is made to handle different subjects, such as listings, bookings, and reviews. All three fact tables share one similar dimension which is the **Time Dimension** to enable time-based analysis across the 3 subjects. A **bridge table** also exists between Host and Channel Dimension because each Host may have used different Channels and vice versa. This star schema facilitates easy calculations of aggregated values, such as the **number of listings**, **average booking cost**, and **total number of reviews**.



Descriptive Analysis

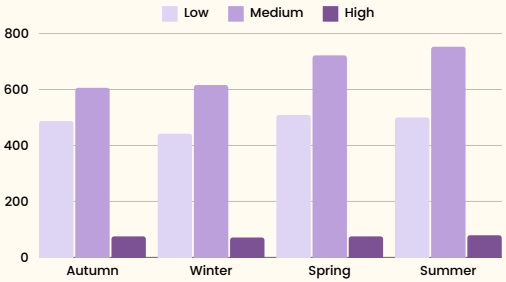
Number of Booking and Review Trends [2019 to 2021]



In 2019, booking counts showed a steady increase, peaking in November 2019 with 66 bookings. There's a **sharp decline in bookings for 2020**, particularly in April where only 7 bookings were made, likely due to the global pandemic. Although bookings **gradually recover in the first half of 2021**, they remained below pre-pandemic levels.

One notable trend is a **one-month buffer period** where high booking volumes were often followed by a spike in reviews the following month. In October 2019, bookings decreased to 56, but reviews surged to 70, indicating that many reviews from September bookings were submitted in October. However, the **buffer pattern became inconsistent post-pandemic**. From July to October 2021, very few to no reviews were submitted, despite some ongoing bookings.

Listings Distribution by Season and Price Range

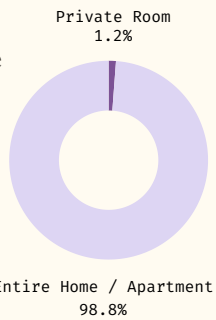


Listings are **mostly in the medium price range**, with the fewest in the high range. Although summer is the peak season overall, spring has slightly more low-price listings than summer, with a difference of 9 listings. Despite this, **Summer** shows the highest activity among the four seasons, making it the peak season for listings.

Total and Average Booking Cost by Listing Type

Only two types of listings are being utilised, despite four available options. Entire homes and apartments significantly outperform private rooms in bookings. Although entire **homes and apartments have a higher average booking cost**, they attract a much **larger number of bookings**. This suggests that customers prioritise space and amenities over price, demonstrating a strong preference for these accomodations, even at a higher cost.

Percentage of Bookings



Average Booking Cost