Data Warehousing for MStay: From Cleaning to Insightful Analysis with Star Schema

OBJECTIVE

The aim of this project was to clean and optimize the MStay data quality issues, including duplication, constraint violations, and missing values. By ensuring data integrity and accuracy, a star schema design was implemented to enable efficient data analysis. This allowed for detailed insights into booking patterns, pricing trends, guest behavior, and listing performance, including identifying peak booking periods, analyzing the impact of listing types, and

evaluating the distribution of high-priced listings across seasons. STARSCHEMA DESIGN DATA CLEANING **Duplication Issues** ListingTypeDIM ListingPriceDIM Duplicate records in the Booking and Host tables were resolved using the SELECT DISTINCT method to retain only unique entries, improving data accuracy and consistency. BookingDurationDIM ListingDurationDIM ookingDurationID **Constraint Violations** BookingDurationDescription CategoryDescription ListingFACT Foreign key violations, such as invalid host_id or booking_id BookingDurationID references, were identified and removed, restoring the ListingTimeDIM relational integrity of the database for stable, accurate Number_Of_Bookings SeasonID analysis. Total Booking Cost HostID BookingPriceDIM PriceRangeID **Null Values** BookPriceRangeID BookPriceRangeDescription Number Of Listings Null values were found in key columns like AMM_ID BookingTime and REVIEW_COMMENT. These were replaced with BookingTimeID defaults such as "Unknown" and "No Comment" to ListingSeasonDIM preserve data completeness and minimize Year information loss. SeasonDescription Numeric Validation ReviewFACT ReviewTimeDIM Inconsistent values, like negative prices or booking ReviewTimeID BookingPriceID HostDIM costs, were corrected (e.g., -150 to 150) to ensure BookingDurationIDnumeric fields adhered to logical expectations, **HostName** Num Of Reviews improving data reliability. ChannelDIM HostChannelBRIDGE ChannellD DATA ANALYTICS ChannelName ChannellD Trends in Average

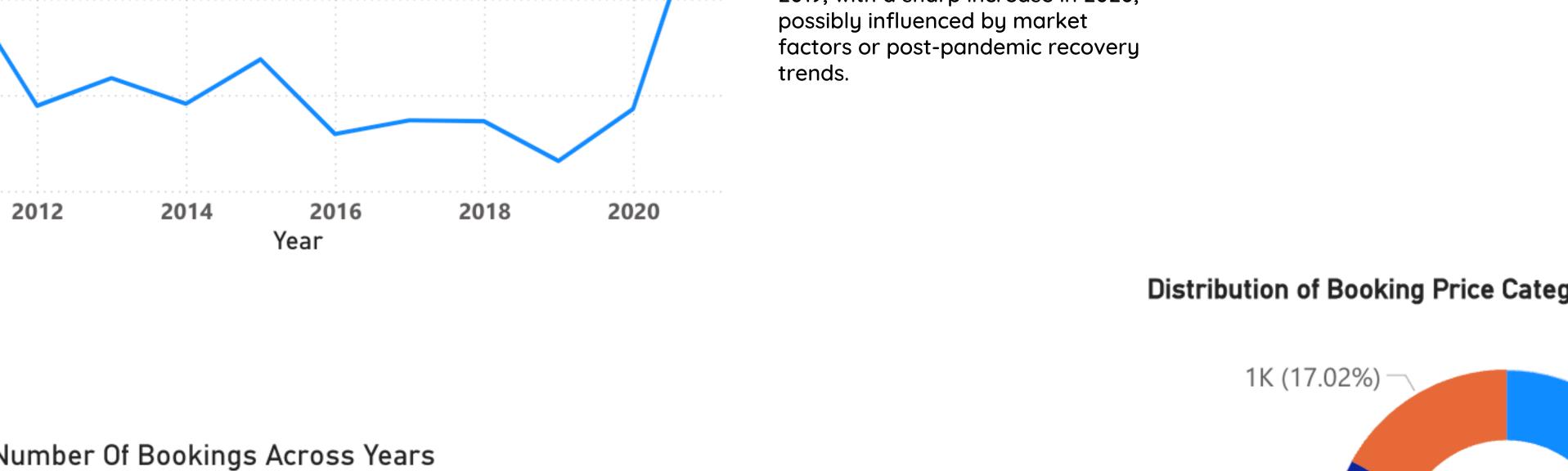
Average Booking Across Years 8,000 Cost 7,500 300king 2,000

Average Bo

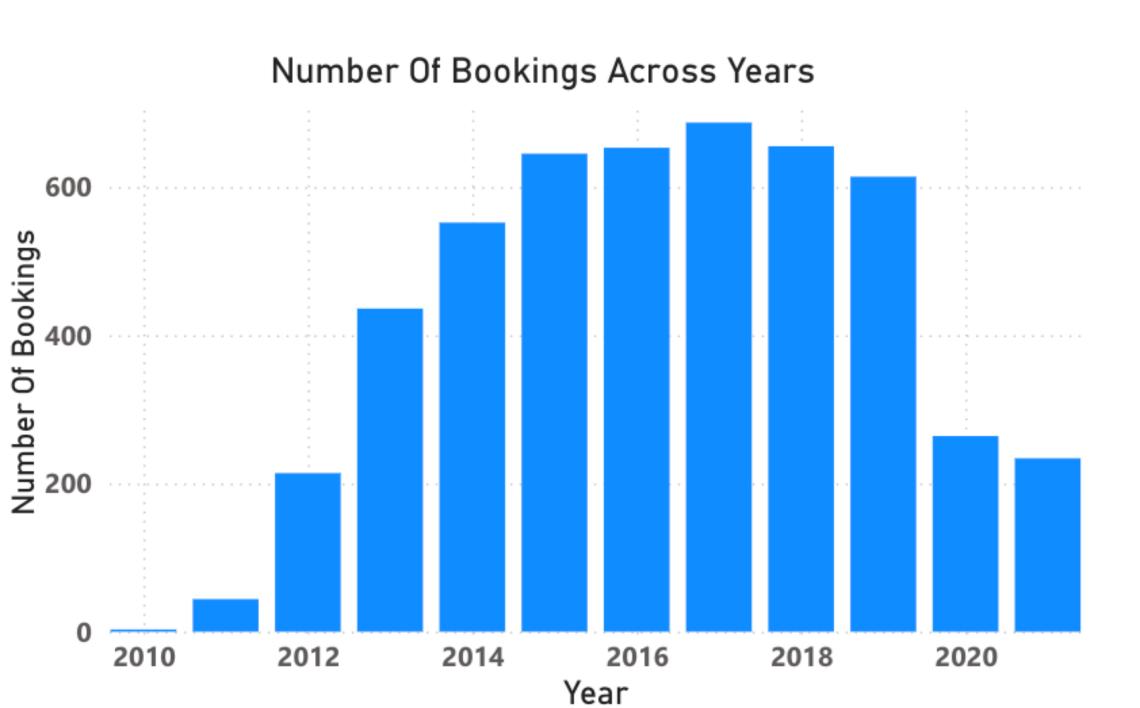
6,000

2010





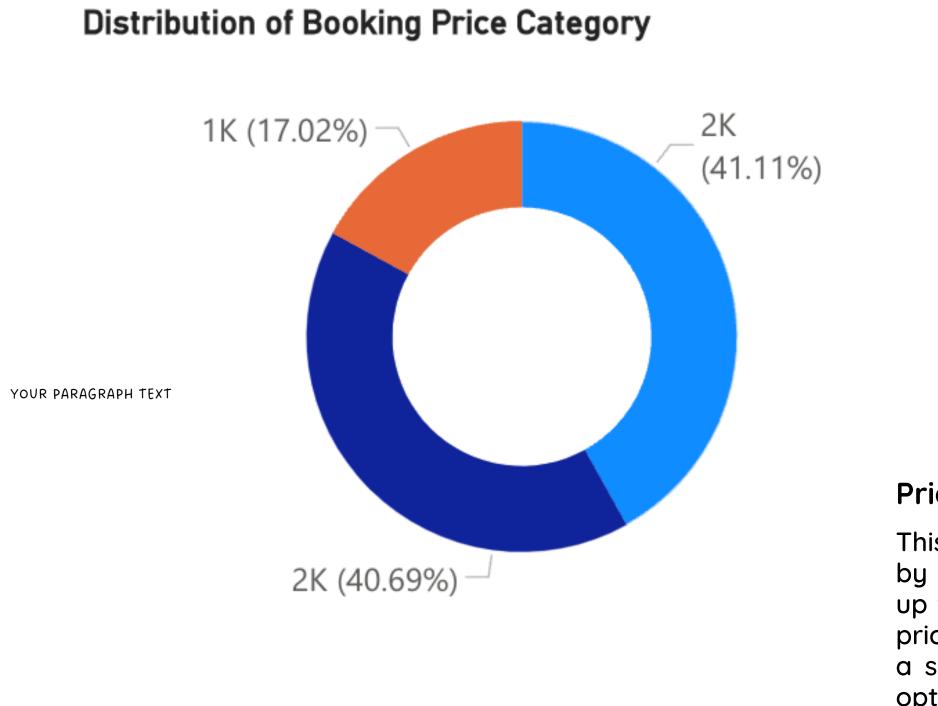
Booking Costs Over



This bar chart visualizes the total number of bookings year, highlighting a peak between 2014 and 2018 followed by a decline recent years. The decrease in 2020 could be attributed external to factors such as travel restrictions.

Yearly Booking

Volume



booking volume. PRICE_CATEGORY Low

Number of Reviews Per

follows a similar pattern to

bookings, peaking between

correlation between reviews

and bookings suggests that

guest feedback trends are

The number of reviews

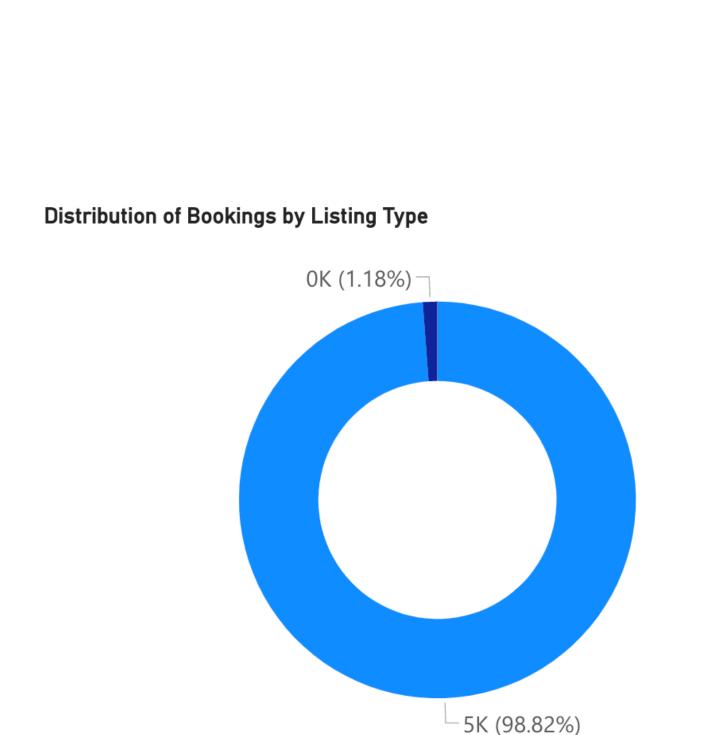
2014 and 2018 before

dropping in 2020. The

directly influenced by

Year

Price Category Distribution This pie chart displays the distribution of bookings by price category. Medium-priced bookings make up the majority (41.11%), followed closely by lowerpriced bookings. High-priced bookings account for a smaller portion, indicating that most customers opt for affordable or moderately priced listings.



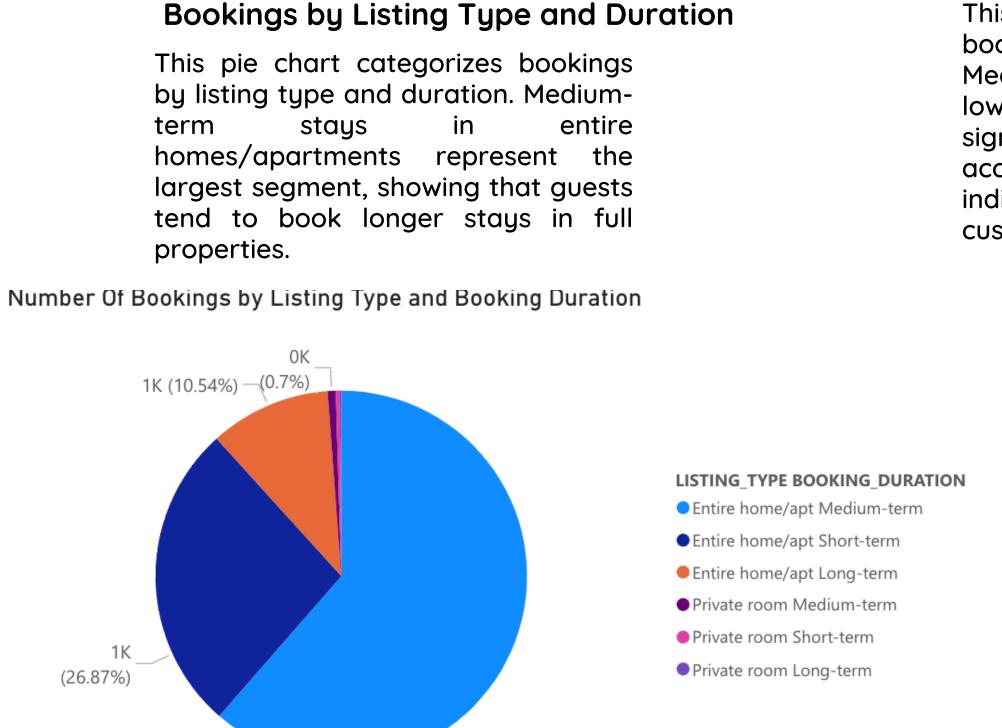
Bookings by Listing Type

Bookings for entire homes/apartments dominate the listing type distribution, accounting AGRAPTTEXT total bookings, private rooms constitute a much smaller share. This suggests that guests prefer property rentals over shared accommodations.

LISTING TYPE

Private room

Entire home/apt



─ 3K (61.41%)

Price Category Trends in Bookings

Medium

High

This bar chart shows the number of bookings in each price category. Medium-priced bookings dominate, with low-priced listings also attracting significant bookings. High-priced listings account for the fewest bookings, indicating price sensitivity among customers. 2000



Reviews Across Years