BUSINESS INTELLIGENCE

Project Report

June 1, 2025

By Qamar Raza 27140, Zain Sharjeel 26922, Abdullah Khalid 26969, Syed Muhammad Hussain 26992

https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

PRoject Overview

This project will perform a complete analysis of a dataset from the hotel industry. a dataset from the hotel industry, utilizing a dataset that details various aspects of hotel reservations such as booking timings, customer demographics, and deposit information.To approach this challenge, we will adopt the Design Thinking framework. The initial "Empathize" phase will involve gathering qualitative insights, ideally through interviews with hotel staff, and make an empathy map. Following this, the "Define" phase will focus on crystallizing the specific problems and analytical questions that our BI solution will target, ensuring our efforts are aligned with addressing real business problems.

Before developing visualizations, a critical phase of Exploratory Data Analysis (EDA) will be undertaken to thoroughly examine the dataset, identify initial patterns, spot any anomalies, and form early hypotheses regarding cancellation factors. This will be coupled with data wrangling, which includes cleaning the data, ensuring correct data types, and feature engineering.

We will start with a sprint brainstorming activity making the original framework of the project. We will then move onto the Prototype and Test stage, iteratively improving on the ideas and solutions. The final output will be a dashboard made to solve a problem identified in the “Define” stage of the design thinking framework.

Data Transformation

1. Data Overview

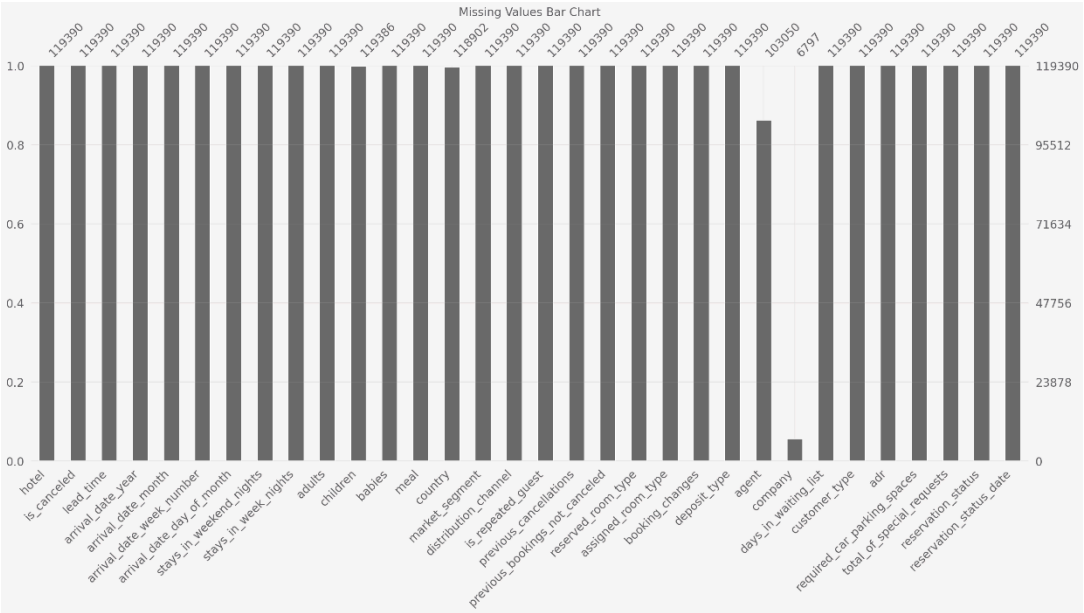
COLUMN	PURPOSE
hotel	Type of hotel (Resort Hotel or City Hotel).
is_canceled	Booking canceled (1) or not (0).
lead_time	Days between booking and arrival.
arrival_date_year	Year of arrival.
arrival_date_month	Month of arrival.
arrival_date_week_number	Week number of arrival year.
arrival_date_day_of_month	Day of the month of arrival.
stays_in_weekend_nights	Number of weekend nights (Sat/Sun) stayed.
stays_in_week_nights	Number of week nights (Mon-Fri) stayed.
adults	Number of adults.
children	Number of children.
babies	Number of babies.
meal	Type of meal booked (BB, HB, FB, SC).
country	Country of origin (ISO 3166-1 alpha-3 code).
market_segment	Market segment (e.g., TA for Travel Agents, TO for Tour Operators).
distribution_channel	Booking distribution channel (e.g., TA/TO).
is_repeated_guest	Booking by a repeated guest (1) or not (0).
previous_cancellations	Number of previous cancellations by the customer.
previous_bookings_not_canceled	Number of previous non-canceled bookings by the customer.
reserved_room_type	Code of room type reserved.
assigned_room_type	Code of room type assigned.
booking_changes	Number of changes made to the booking.

deposit_type	Type of deposit made for the booking.
agent	ID of the travel agency.
company	ID of the company/entity making the booking.
days_in_waiting_list	The number of days booking was on the waiting list.
customer_type	Type of booking (e.g., Transient, Contract).
adr	Average Daily Rate.
required_car_parking_spaces	Number of car parking spaces required.
total_of_special_requests	Number of special requests made.
reservation_status	Final booking status (Canceled, Check-Out, No-Show).
reservation_status_date	Date of the last reservation status update.
hotel	Type of hotel (Resort Hotel or City Hotel).
is_canceled	Booking canceled (1) or not (0).
lead_time	Days between booking and arrival.
arrival_date_year	Year of arrival.
arrival_date_month	Month of arrival.
arrival_date_week_number	Week number of arrival year.
arrival_date_day_of_month	Day of the month of arrival.
stays_in_weekend_nights	Number of weekend nights (Sat/Sun) stayed.
stays_in_week_nights	Number of weeknights (Mon-Fri) stayed.
adults	Number of adults.
children	Number of children.
babies	Number of babies.
meal	Type of meal booked (BB, HB, FB, SC).
country	Country of origin (ISO 3166-1 alpha-3 code).
market_segment	Market segment (e.g., TA for Travel Agents, TO for Tour Operators).
distribution_channel	Booking distribution channel (e.g., TA/TO).
is_repeated_guest	Booking by a repeated guest (1) or not (0).

previous_cancellations	Number of previous cancellations by the customer.
previous_bookings_not_canceled	Number of previous non-canceled bookings by the customer.
reserved_room_type	Code of room type reserved.
assigned_room_type	Code of room type assigned.
booking_changes	Number of changes made to the booking.
deposit_type	Type of deposit made for the booking.
agent	ID of the travel agency.
company	ID of the company/entity making the booking.
days_in_waiting_list	Number of days booking was on the waiting list.
customer_type	Type of booking (e.g., Transient, Contract).
adr	Average Daily Rate.
required_car_parking_spaces	Number of parking spaces required.
total_of_special_requests	Number of special requests made.
reservation_status	Final booking status (Canceled, Check-Out, No-Show).

2. Data Wrangling

We found inconsistent "NULL" string representations for missing data, potential data type mismatches for columns like agent, company, and children (initially floats), and the need to consolidate individual date components.



Handling missing values was a key step. "NULL" strings were first standardized to NaN. For the country column, missing values were imputed with the mode to preserve data for analysis. Missing agent and company IDs were filled with 0, indicating no intermediary, while missing children counts were set to 0, a common assumption for unspecified values. The reservation_status_date column also had missing entries (NaT after date conversion attempts), which were noted for potential future consideration based on specific analytical needs.

Data type corrections were then implemented. children, agent, and company were converted to integer types for accurate representation.

A significant transformation was the creation of a unified arrival_date column from its constituent year, month, and day parts, with month names mapped to numbers to ensure proper parsing. Any rows resulting in an invalid arrival_date (NaT) after this consolidation were removed due to the criticality of this field. Similarly, reservation_status_date was converted to a datetime object.

Categorical data inconsistencies were also addressed. 'Undefined' entries in the meal column were re-categorized as 'SC' (Self-Catering/No Meal) for clarity. Leading and trailing whitespaces were stripped from reserved_room_type and assigned_room_type to ensure consistent category representation. Bookings showing zero total guests (adults + children + babies) were identified and flagged, as these are atypical and may warrant further investigation or exclusion in specific analyses.

Missing data was addressed using targeted strategies for each affected column:

- **country**: Missing values imputed with the mode (most frequent country), as this is a common and reasonable assumption for categorical location data when no other information is available.
- **agent**: Missing agent IDs (originally 'NULL') filled with 0, chosen as a placeholder to numerically signify that no travel agent was involved in the booking, maintaining the column's numeric potential for ID-based analysis.
- **company**: Similar to agent, missing company IDs filled with 0, serving as a numerical indicator that no specific company was associated with or paid for the booking.
- **children**: Missing values for the number of children filled with 0, based on the common-sense assumption that an unspecified count typically means zero children for a booking.

To ensure data integrity, the dataset was checked for complete duplicate rows, and any found were removed to prevent skewed analytical results. Finally, after the successful creation and validation of the consolidated arrival_date datetime column, the original, now redundant, date component columns (arrival_date_year,

arrival_date_month, arrival_date_day_of_month, and arrival_date_week_number) were dropped.

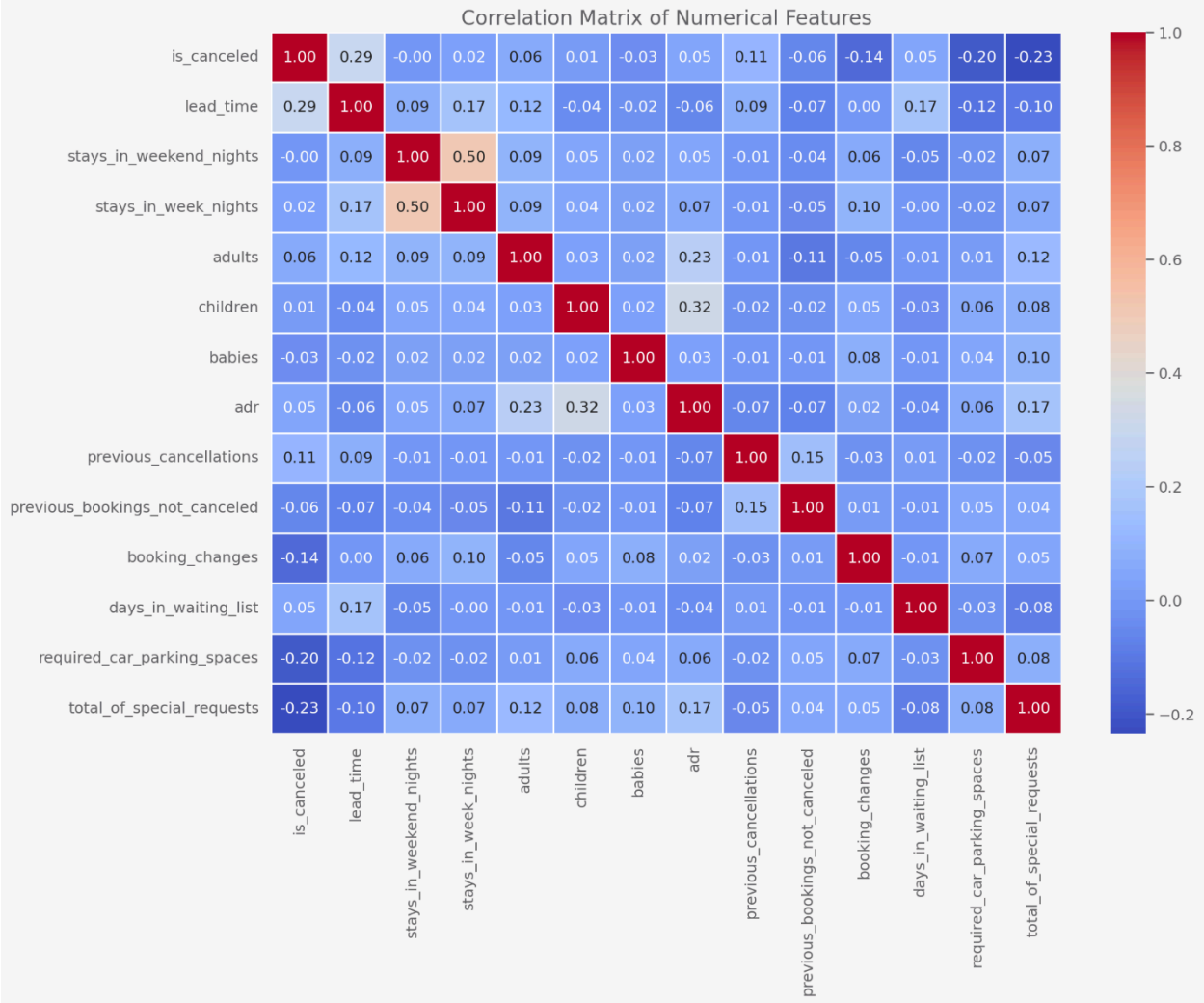
3. Univariate Analysis

Our exploration of the hotel booking data set aimed to understand its core characteristics and identify initial patterns, particularly those related to booking cancellations. The first step, univariate analysis, involved examining each feature individually. For numerical data, we observed that lead_time (the period between booking and arrival) is generally short, though some bookings are made significantly in advance. The adr (average daily rate) showed considerable variability, with instances of very high prices alongside some at or near zero, which might indicate complimentary stays or special promotions. Most bookings consist of one or two adults, with children and babies being less common. It was also noted that most guests do not have an extensive history of previous_cancellations or previous_bookings_not_canceled, and features like booking_changes or days_in_waiting_list frequently registered as zero. When looking at categorical features, the dataset includes both "Resort Hotel" and "City Hotel" types. A significant proportion of bookings were marked as is_canceled. "BB" (Bed & Breakfast) emerged as a popular meal choice, and "Online TA" (Travel Agents) was a dominant market_segment and distribution_channel. The majority of bookings were from new guests (not is_repeated_guest), "No Deposit" was the most prevalent deposit_type, and "Transient" was the leading customer_type. The final reservation_status of bookings primarily fell into "Check-Out," "Canceled," or "No-Show."

4. Bivariate Analysis

Bivariate analysis focused on the relationships between pairs of variables, with a particular emphasis on factors influencing is_canceled. The correlation matrix revealed a positive linear relationship between lead_time and is_canceled, suggesting that bookings made further in advance have a higher likelihood of cancellation. Similarly, previous_cancellations also showed a positive correlation with

current cancellations. Statistical tests, such as ANOVA, confirmed significant differences in the average values of numerical features like lead_time, adr, and previous_cancellations when comparing canceled and non-canceled bookings; for instance, canceled bookings generally had longer average lead times. For categorical variables, Chi-squared tests indicated significant associations between is_canceled and several features, including deposit_type, customer_type, market_segment, and is_repeated_guest. This means the cancellation rate varied notably based on the deposit policy in place and the type of customer making the booking.



5. Outlier Analysis

Outlier analysis was conducted to identify data points deviating substantially from the norm. Box plots visually highlighted potential outliers in lead_time (exceptionally long durations), adr (extremely high or very low/zero values), and previous_cancellations (a few customers with a high number of past cancellations). The Interquartile Range (IQR) method numerically corroborated these findings, particularly for adr and lead_time, and an initial look showed these outliers were present in both canceled and non-canceled bookings. Scatter plots, such as lead_time versus adr, provided a two-dimensional view, helping to spot unusual combinations and their cancellation status. The EDA indicates that factors like extended lead times, a history of prior cancellations, specific deposit policies, and distinct customer types appear to be statistically significant in predicting booking cancellations.

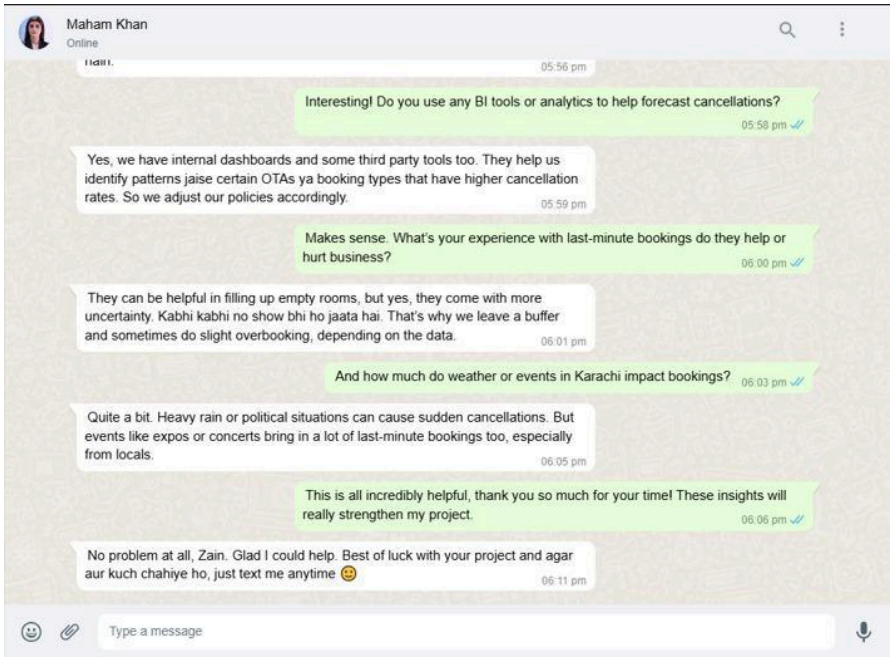
EMPHATIZE

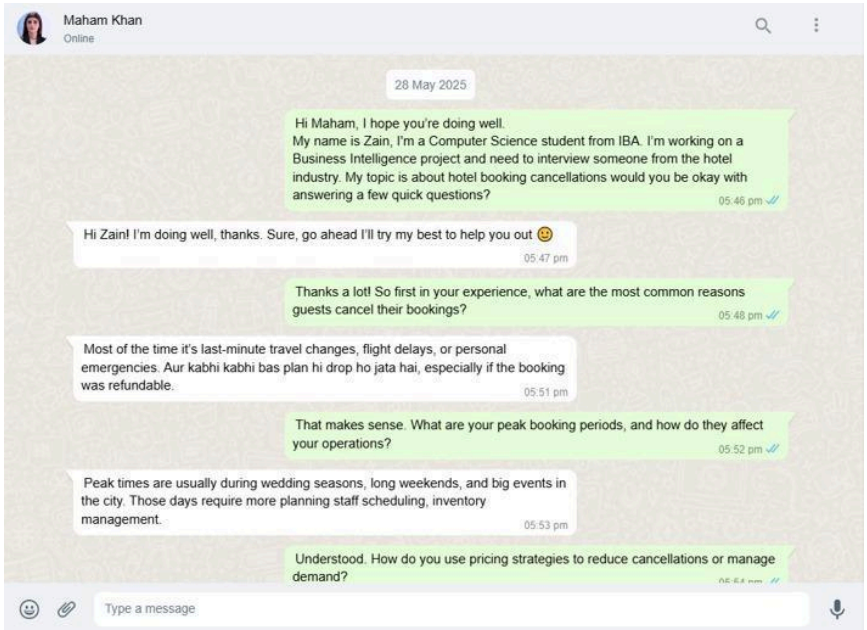
1. Interview Summary

https://www.linkedin.com/in/maham-khan-01403841/?originalSubdomain=pk

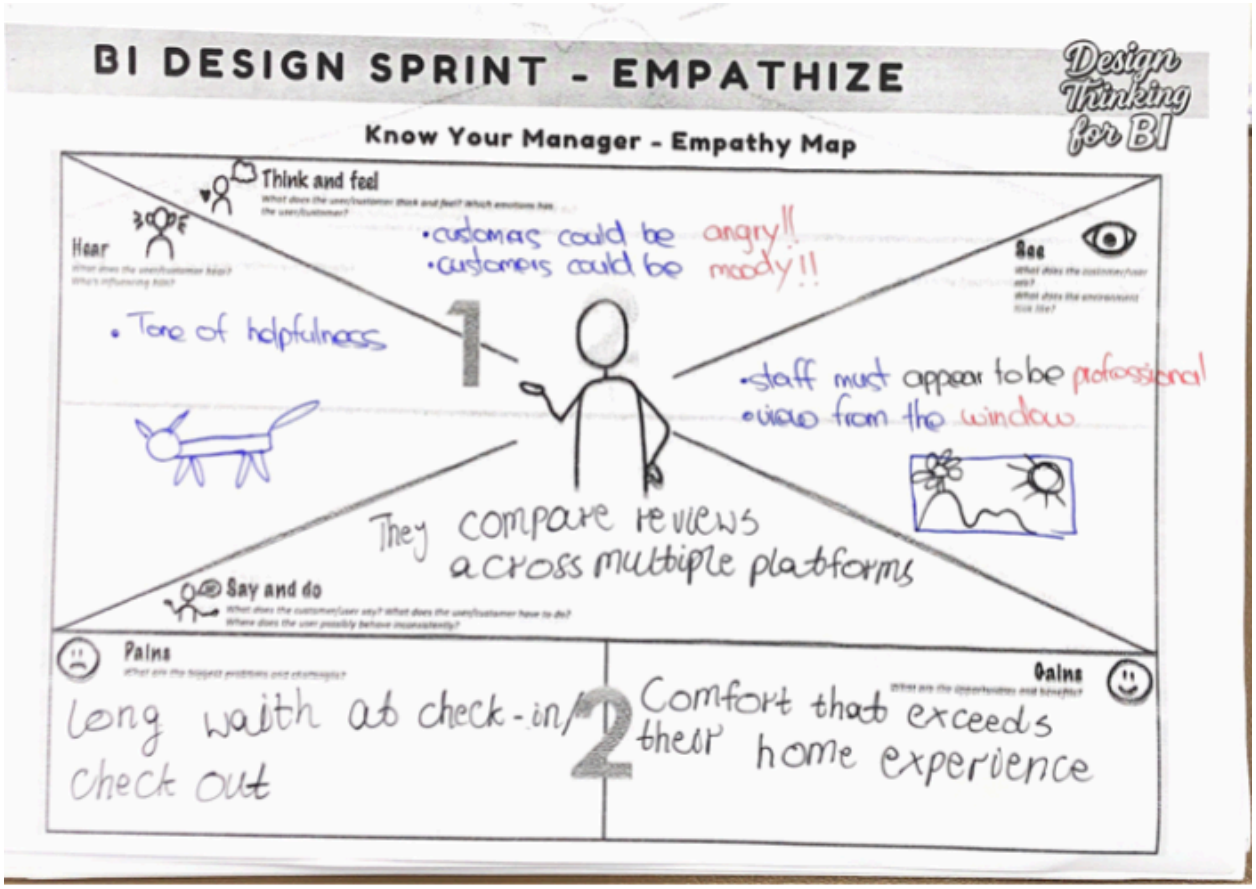
We was fortunate to interview Maham Khan, Cluster Asst. Director Marketing at Marriott Hotels, at the Marriott Hotel, Karachi, for our Business Intelligence project. Her revelations were quite insightful, especially on hotel booking cancellation practices and the role of BI tools in decision-making. She mentioned that last-minute changes to travel plans, flight issues, or attractive offers elsewhere are the primary reasons for cancellation. Also, she emphasized the use of dynamic pricing, non-refundable rates, and event-based forecasting as tools to reduce cancellations. Maham explained how BI dashboards allow them to analyze trends, adjust overbooking strategies, and

manage last-minute booking behavior. These practical experiences applied to what I studied and brought real-world applications to theoretical learning.

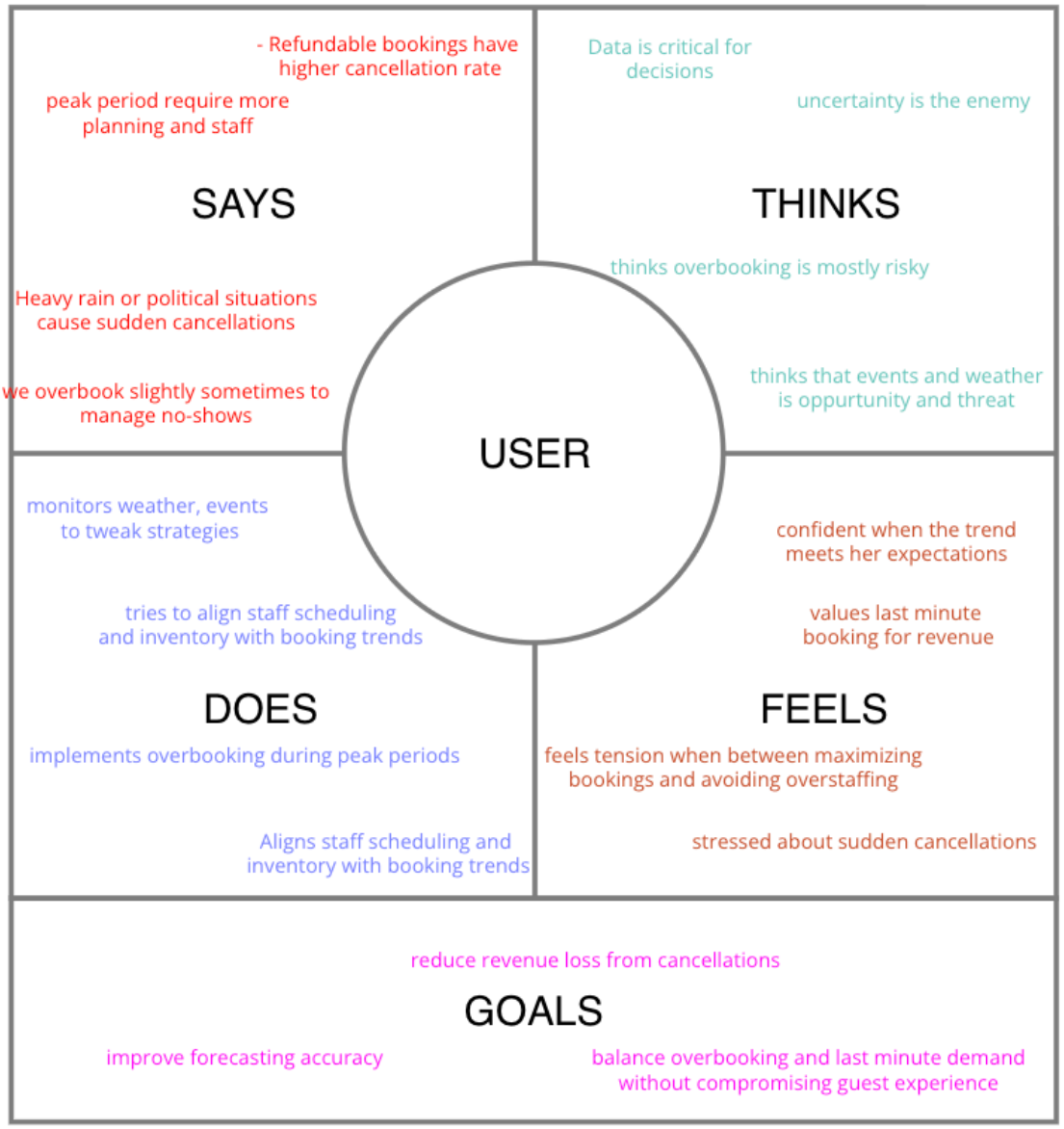




2. Empathy Map First Draft



3. Empathy Map After Interview



DEFINE

Problem: Analyzing booking cancellations.

Cancellations are significantly affecting our revenue, and we need to check which hotel policies are affecting cancellations.It's not just the number of cancellations, but when they happen and who is cancelling.

Why is this important?

We want to use these insights to develop targeted strategies maybe adjusting deposit policies for specific segments/lead times, offering non-refundable rates more strategically, optimizing our marketing spend on more reliable channels, or even implementing early intervention tactics for high-risk bookings.

How will solving this problem help?

Solving this problem will help us better undestand which hotel policies (deposit type, Channel), and customer type have the highest cancellation rate. We can then optimize our policies to reduce the cancellation rate or get a better idea of which types of customers are most likely to cancel. Being able to better predict cancellations will help the hotel make the most of its limited rooms during vacation season.

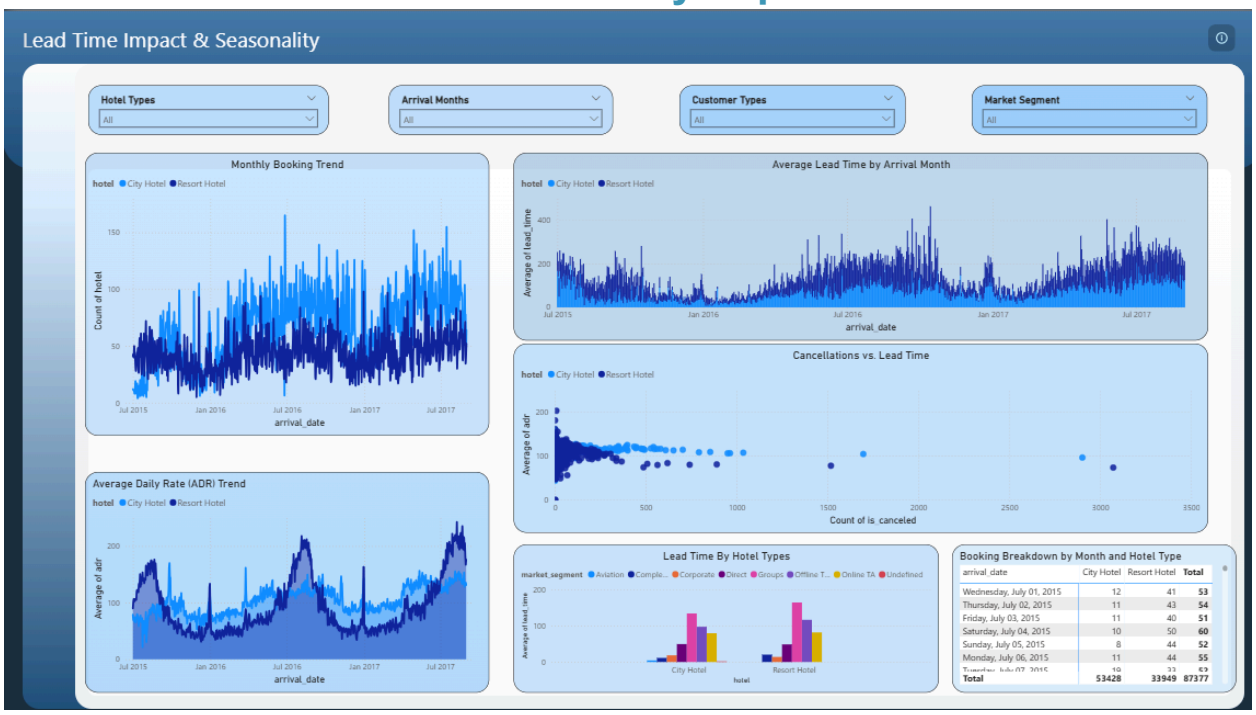
IDEATE – Business Quries

Business Query	Justification
High-Value Cancellations: Are we losing more revenue from cancellations of particular room types, longer stays, or bookings with higher ADRs? We need to quantify this.	Understanding which cancellations hurt the most financially (high ADR, specific valuable room types, or extended stays) helps prioritize retention efforts or consider targeted non-refundable policies for these segments. It quantifies the financial risk beyond just the number of cancellations.
Lead Time Impact & Seasonality: Do bookings made far in advance have a higher cancellation rate than last-minute ones? How does this interact with seasonality (e.g., long lead-time bookings for peak season are more or less likely to cancel than off-peak)? Is there a 'danger zone' in terms of lead time where cancellation risk peaks?	This explores the dynamic between booking horizon and cancellation. Identifying if a 'danger zone' in lead time exists allows for targeted interventions (e.g., reminder emails, deposit policies for certain lead times). Understanding seasonality's interaction helps refine forecasting and risk assessment for different periods.
Channel & Segment Risk: Are certain market segments (e.g., Online TAs vs. Direct bookings) or distribution channels more prone to cancellations? If so, is	This assesses the risk-reward profile of different acquisition channels and customer segments. If a high-cancellation channel also brings low ADR or high commission costs, its overall value to the business is questionable and might require renegotiation or strategic shifts.

the ADR from these high-cancellation channels high enough to justify the risk, or are we essentially paying high commission for unreliable bookings?	
Effectiveness of Deposits & Customer Behavior: How much does our 'No Deposit' policy contribute to cancellations compared to bookings where a deposit is made? Does this vary by customer type (e.g., Transient vs. Contract)? Are customers with a history of previous cancellations more likely to cancel again, even with a deposit?	This directly evaluates the impact of current deposit strategies and how different customer groups respond. It helps determine if deposit policies need adjustment for specific segments or if past behavior is a strong predictor irrespective of deposit, guiding customer relationship management.
Impact of Booking Changes: When customers make changes to their bookings, does this increase or decrease the likelihood of eventual cancellation? Is there a threshold of changes that signals a high-risk booking?	Analyzing the relationship between booking modifications and cancellations can help identify leading indicators of potential churn. If multiple changes significantly increase cancellation risk, this could trigger proactive engagement or flag bookings for closer monitoring.

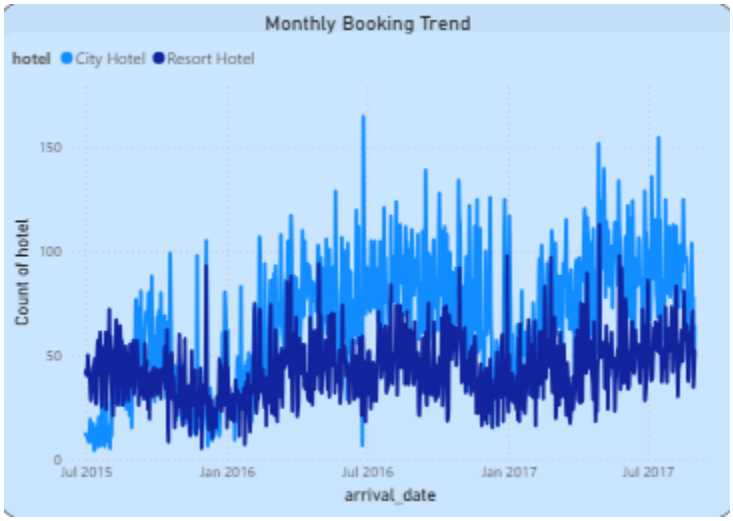
Prototype – BUilding the Story

1.How does time and Seasonality Impact the Cancellations?



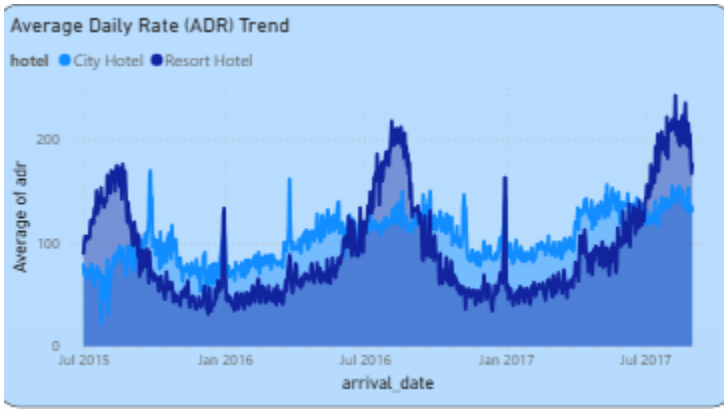
First we need to analyze general bookings and ADR trends before jumping into analysis of cancellation rates.

1)Monthly Booking Trend



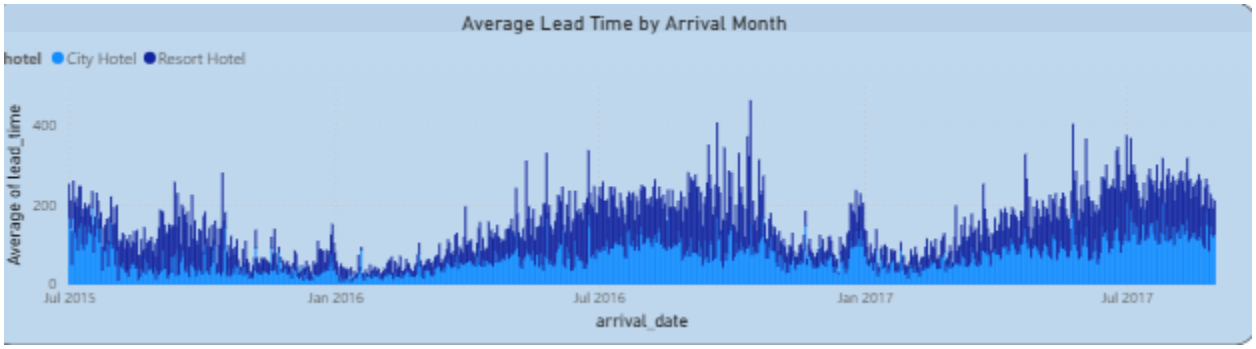
We see that on average; City hotels get more booking than Resort Hotels. An anomally here is that the hotels do not see a spike in bookings during the summer or winter vacations. No clear indication of spike in bookings during these times in our charts.

2)Average Daily Rate Trend



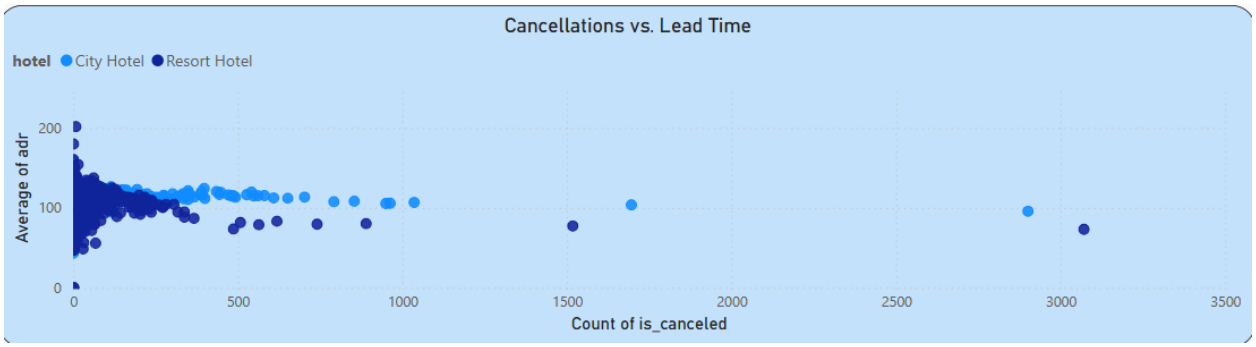
An interesting pattern to see here is that ADR spikes during the summer season(July) for Resort hotels, but this trend does not follow for the city hotels.

3)Average Lead Time by Month



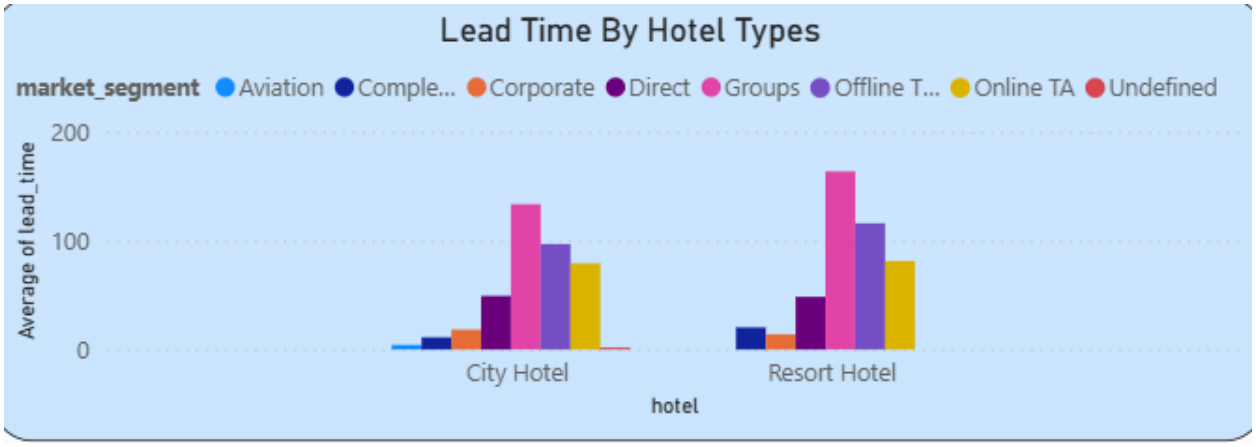
No clear conclusive pattern is observed for lead time by time. No concrete conclusion can be drawn about lead times.

4)Cancellations vs Lead Time



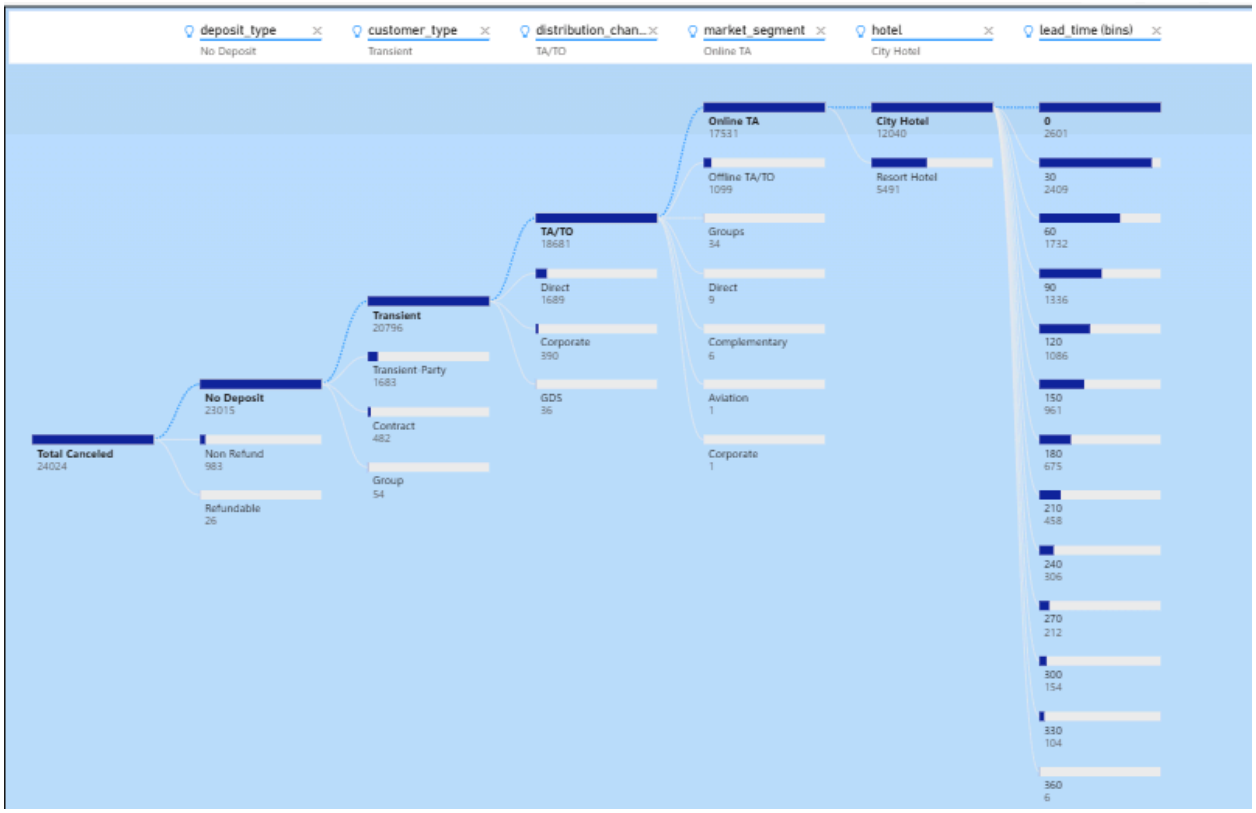
City Hotels experience significantly more cancellations than Resort Hotels, especially at higher average daily rates (adr), indicating a stronger correlation between price and cancellations in City Hotels.

5)Lead Time by Hotel Types



On average we see that market segment of type Group have higher lead time for both City and Resort Hotels.

2.Decomposition Tree

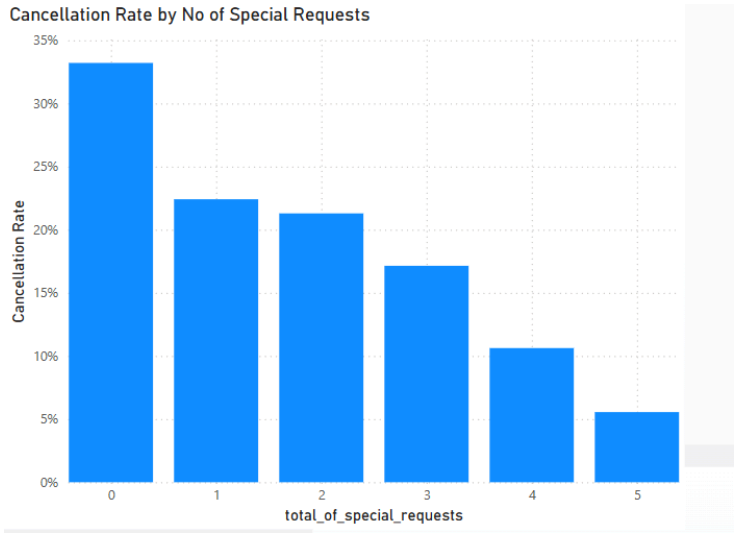


Before we dive into cancellation rates, we first analyze the decomposition tree for the total canceled measure. This allows us to get a quick overview of how Cancellations get distributed by the

Deposit type, customer type, distribution channel, market segment, hotel type and Lead time.

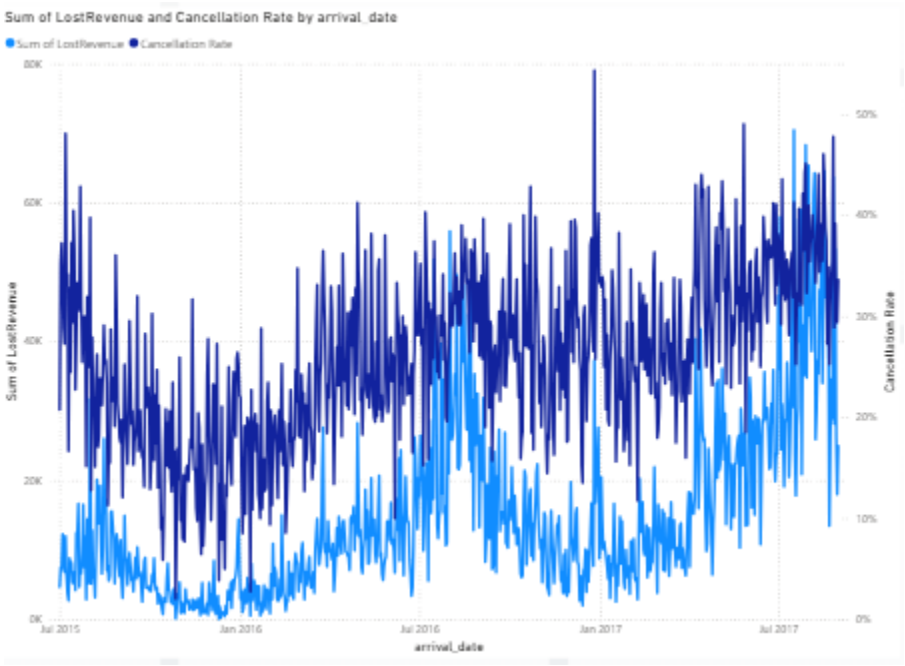
3. High Value Cancellations Analysis 1

1) Cancellation by Special Requests



Guests with more special requests are significantly less likely to cancel, suggesting higher commitment from these customers.

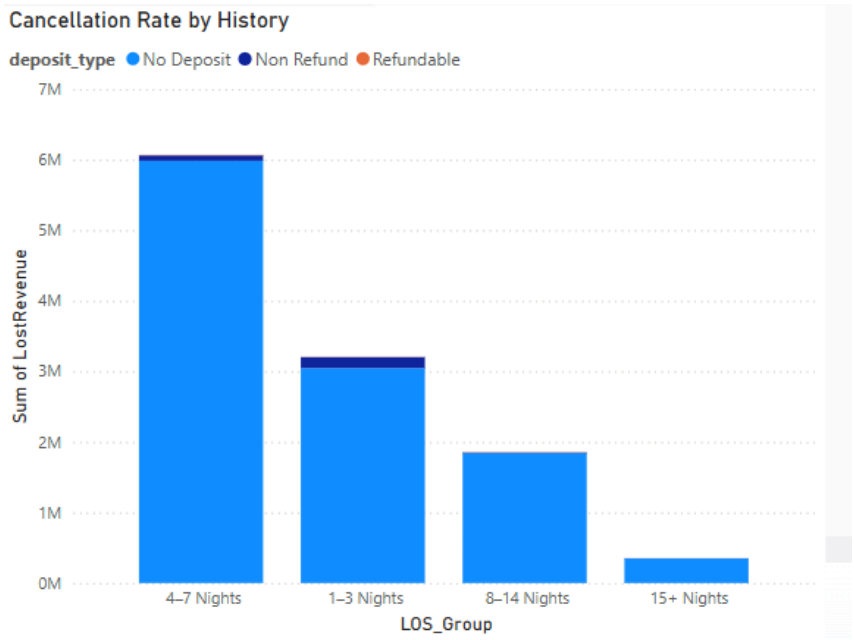
2)Lost Revenue By Cancellation Rate and Arrival Date



Both cancellation rate and lost revenue show a rising trend over time, with noticeable peaks during holiday and summer seasons, indicating seasonal impact on booking behavior.

4. High Value Cancellations Analysis 2

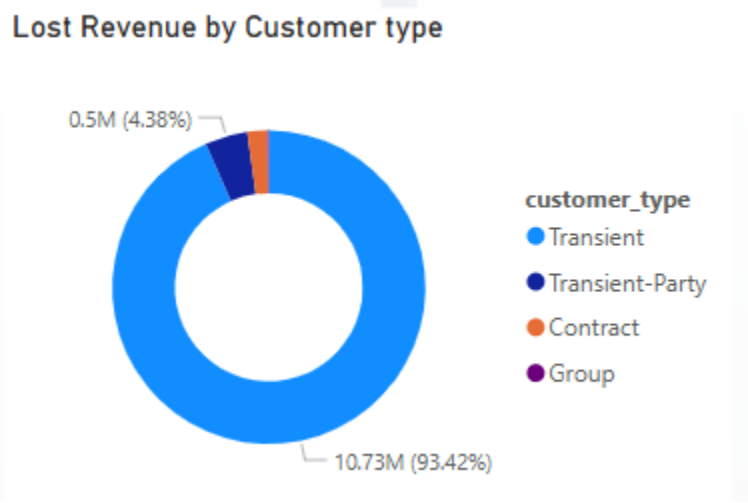
1)Cancellation Rate by History



The "4-7 Nights" length of the stay group generates the most lost revenue (around 6M), overwhelmingly from "No Deposit" cancellations. Shorter stays of "1-3 Nights" are the second largest source of lost revenue (over 3M), again predominantly due to "No

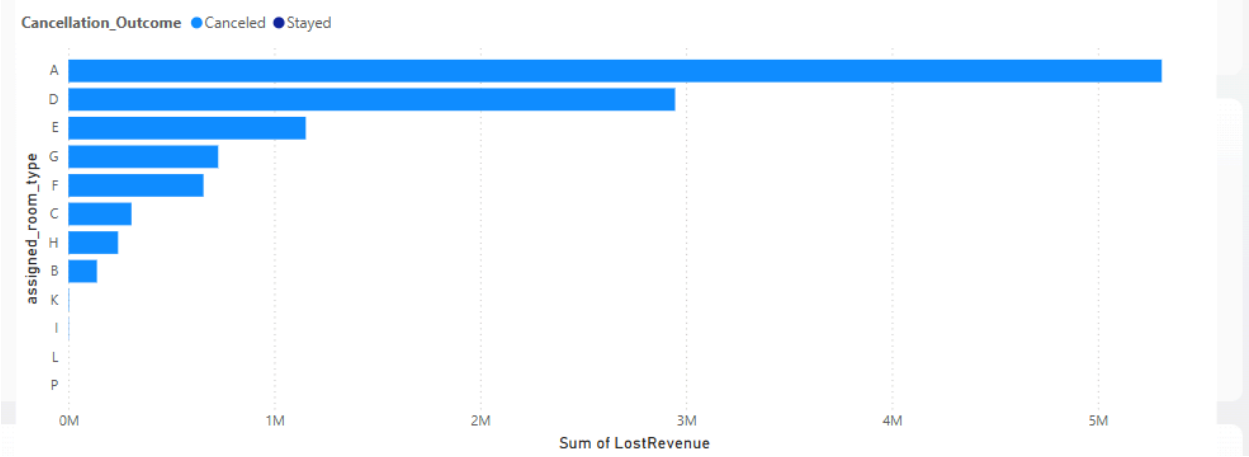
Deposit" bookings. "Non Refund" deposits contribute minimally to lost revenue across all stay lengths.

2)Lost Revenue By Customer Type



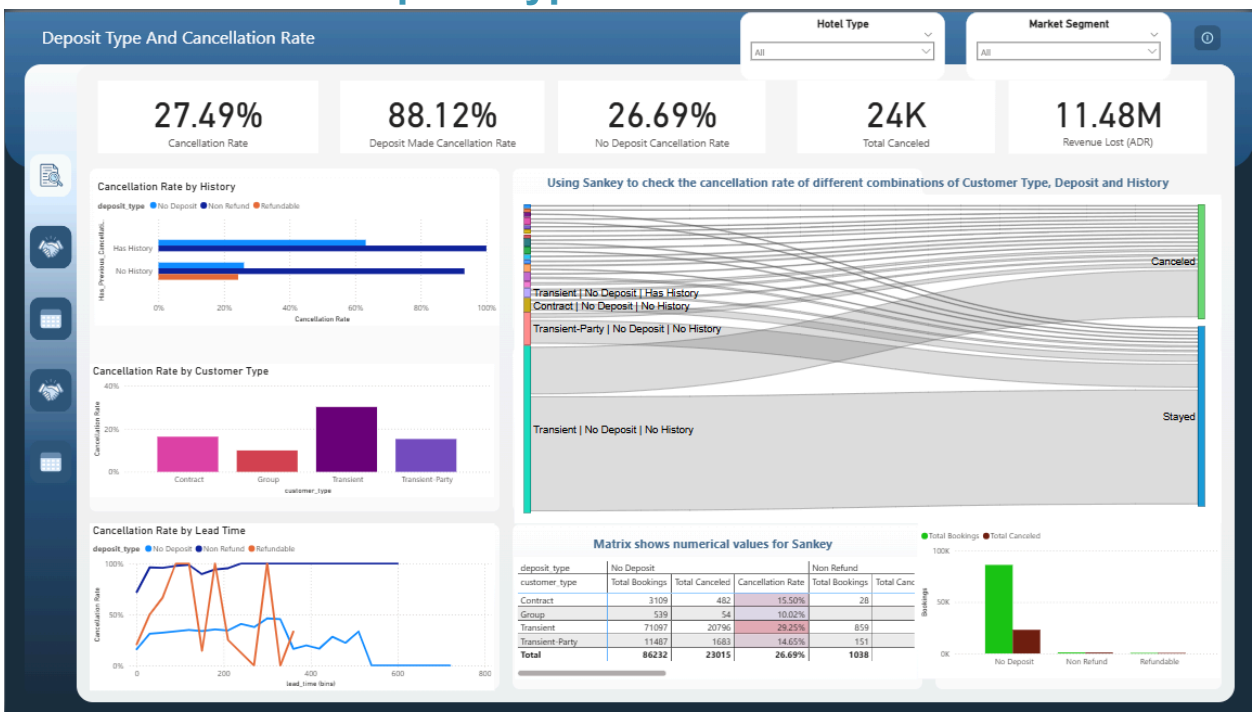
This donut chart shows "Lost Revenue by Customer type," with "Transient" customers overwhelmingly causing the most lost revenue at 10.73M (93.42%). "Transient-Party" customers contribute a much smaller 0.5M (4.38%) to lost revenue. "Contract" customers represent a negligible portion, and "Group" customers do not visibly contribute to lost revenue. It should be noted that this is likely due to a very high imbalance of Transient customers type.

3)Lost Revenue by Room Type

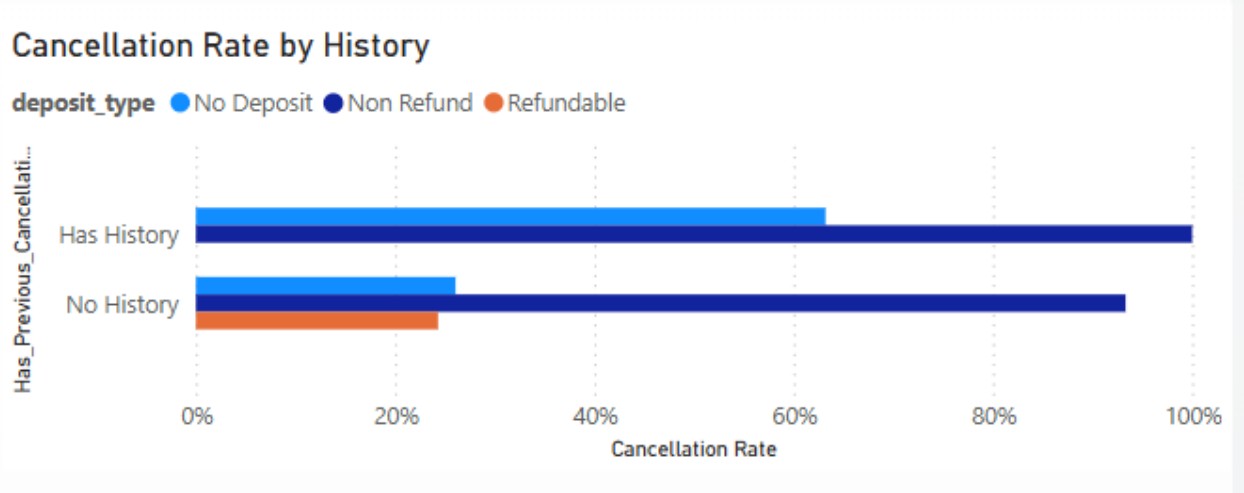


Room type "A" contributes the most to lost revenue, exceeding 5M. Room type "D" is the next highest contributor, with lost revenue around 3M. Subsequent room types (E, G, F, etc.) contribute progressively less lost revenue.

5.How does the deposit type effect Cancellations?

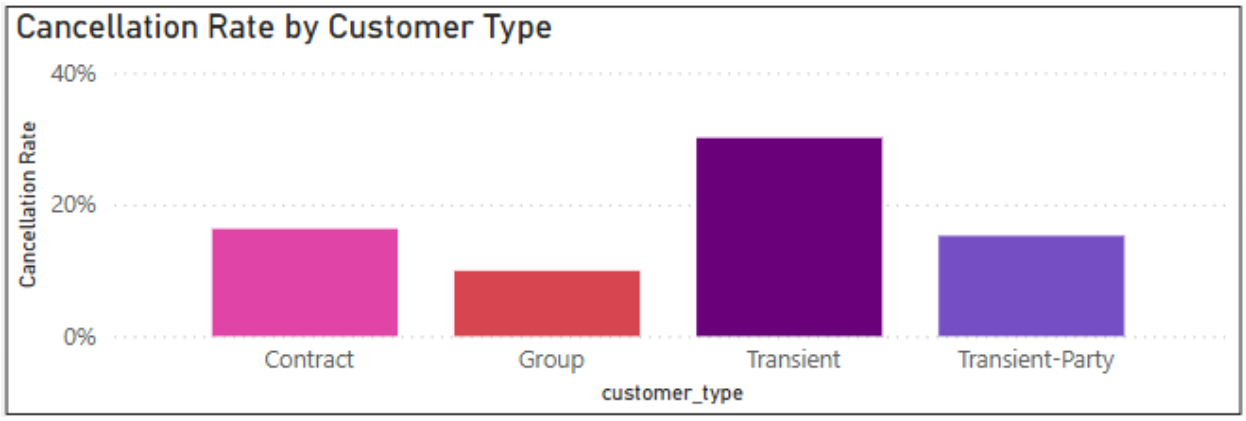


1)Cancellation Rate by History



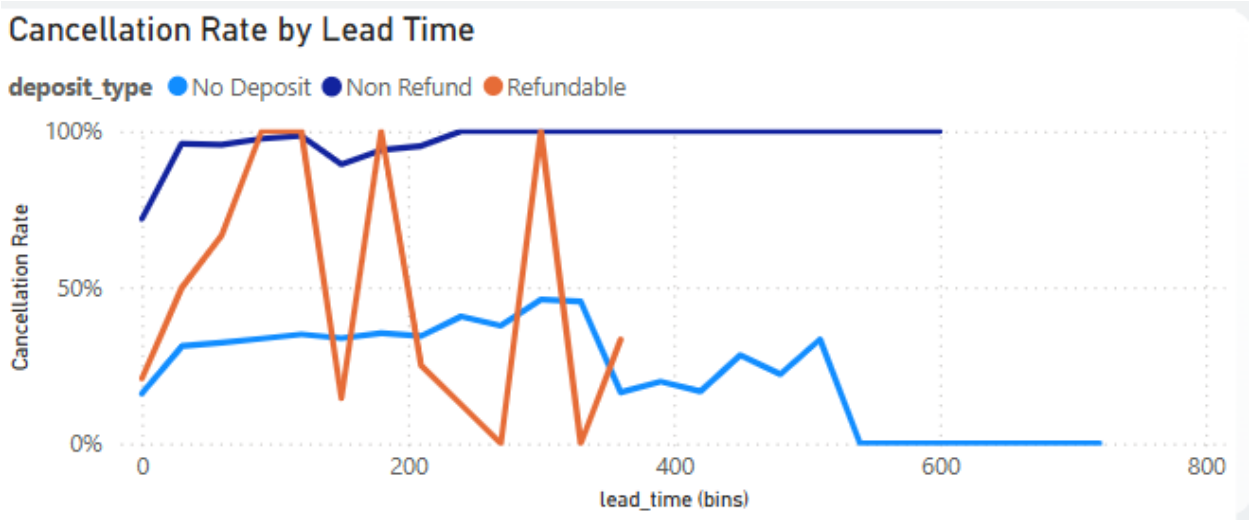
“Has History” indicated that the customer has previously cancelled registerations. This bar chart indicates that people with a history of cancellations have slightly higher cancellations than ones with “No History”.
An anomally that this chart indicates is that the “Non-Refund” deposit type has more cancellations than others. This is unconventional as you would except a non-refund policy to reduce cancellations.

2) Cancellation Rate by Customer Type



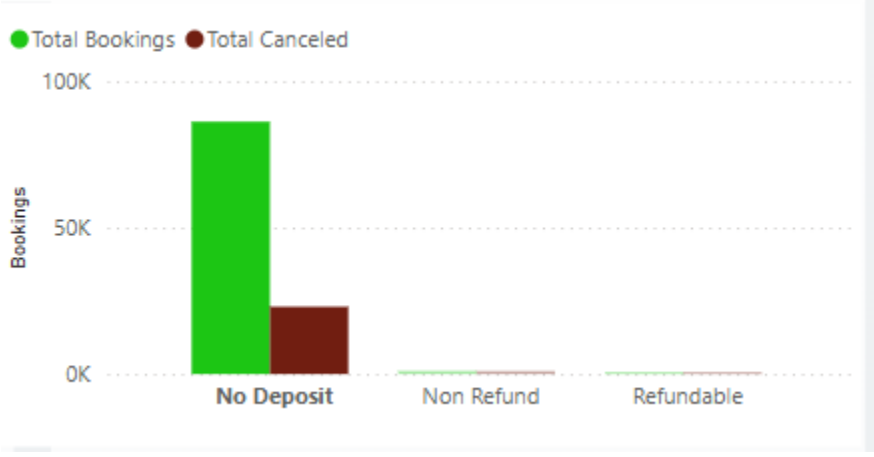
We see that Transient Customer type has the highest cancellation rate. We have also added a tooltip on this chart to indicates the deposit type distribution as a pie chart.

3) Lead Time by Cancellation Rate



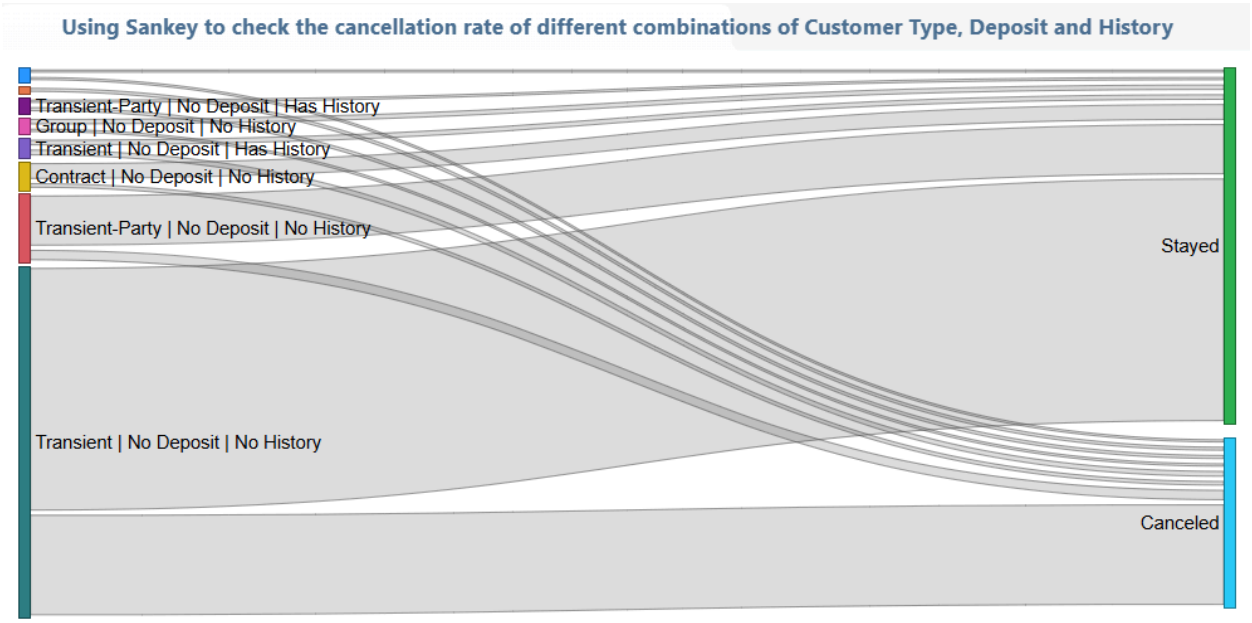
We see that there is no correlation between lead time and cancellation rate. You would expect higher lead time would have more cancellations but our dataset does not indicate this trend.

4)Distibution of Cancellation Rate by Deposit Policy



An anomaly here is that the booking with a deposit (Non-Refund, Refundable) actually have a higher cancellation rate than No deposit. However, we can also see that their records are very low in number so this could just be due to data imbalance.

5)Sankey to illustrate Cancellation Rate of All Combinations of Customer Type, Deposit Type and History



This is the most complex chart in this dashboard. Although a sankey here seems unconventional, it is actually the most suitable chart to illustrate get an overall picture of which combination of the data type have the highest cancellation rate. We have concatenated the Customer type | Deposit Type | History together to link together the cancellation rate. The size of the links indicate the number of booking records.

We see that Transient, No Deposit and No history have the lowest cancellation, and Group, No Deposit and Has history had the highest cancellation rate(the first link the the chart). We have also

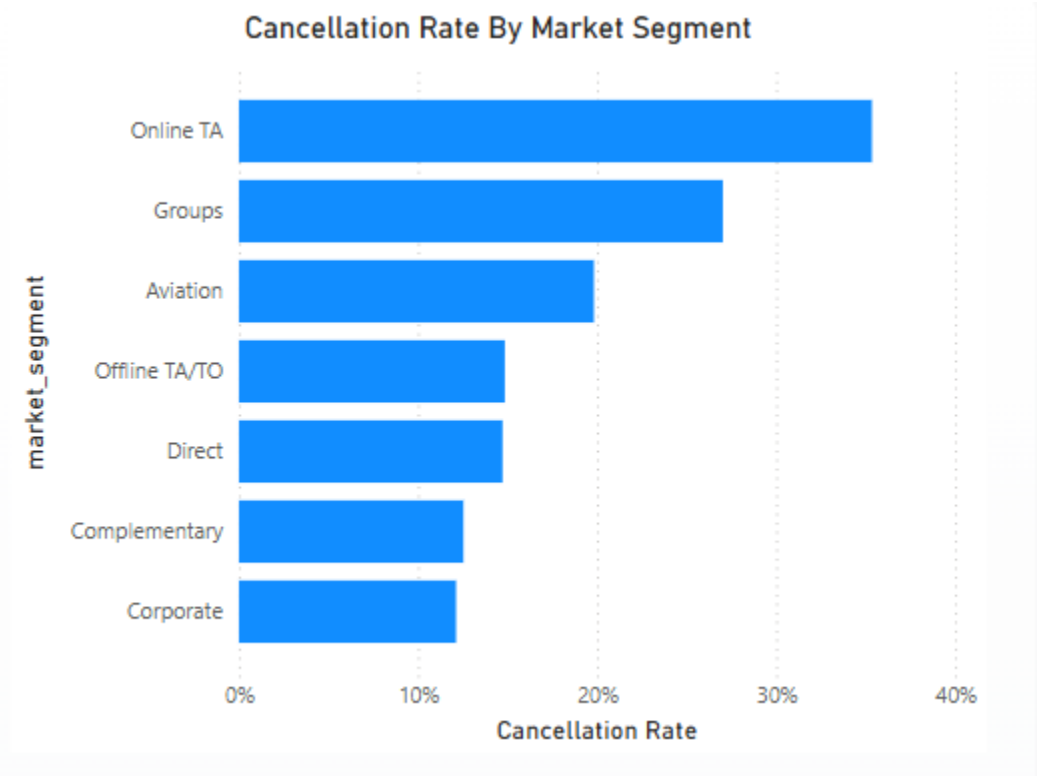
added a matrix view for the sankey so that we can get the numerical data for the above chart.

Matrix shows numerical values for Sankey

deposit_type	No Deposit		Total		
customer_type	eled	Cancellation Rate	Total Bookings	Total Canceled	Cancellation Rate
Contract	482	15.50%	3109	482	15.50%
Group	54	10.02%	539	54	10.02%
Transient	796	29.25%	71097	20796	29.25%
Transient-Party	683	14.65%	11487	1683	14.65%
Total	**015**	**26.69%**	**86232**	**23015**	**26.69%**

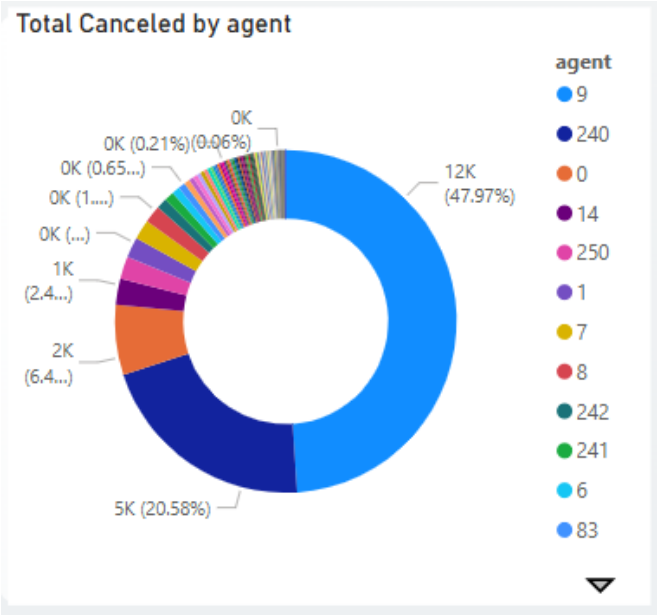
6. Cancellations By Market Segment.

1)Cancellations By Market Segment



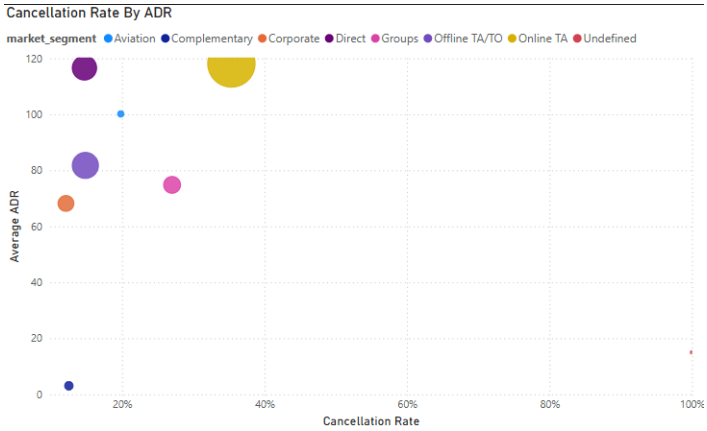
Online TA has the highest cancellation rate among all market segments. Groups and Aviation segments also exhibit significantly high cancellation rates."Corporate bookings show the lowest cancellation rate, with other segments like Offline TA/TO, Direct, and Complementary falling in the middle.

2)Cancellations By Agents



This donut chart displays "Total Canceled by agent," showing that agent "9" is responsible for the largest share of cancellations, nearly half of the total. Agent "240" accounts for the second highest number of cancellations, representing a significant portion. Agent "0" also contributes a noticeable amount, while a multitude of other agents each contribute much smaller percentages to the total.

3)Cancellation Rate by ADR



Online TA (large yellow bubble) shows a high Average ADR but also a high cancellation rate, representing a significant volume. In contrast, Direct bookings (dark purple) achieve a similarly high ADR but with a much lower cancellation rate, while the Undefined segment (tiny red) has an extremely high cancellation rate at a very low ADR.

Insights and Recommendations

1. Insights

- People that have given a deposit actually have a higher cancellation rate than people that did not. One would think that people that gave a deposit would have less cancellations, but our dataset does not follow this convention. This could be due to data imbalance.
- People with a history of cancellations have a higher cancellation rate.
- Higher lead time does not show any correlation with cancellation rate.

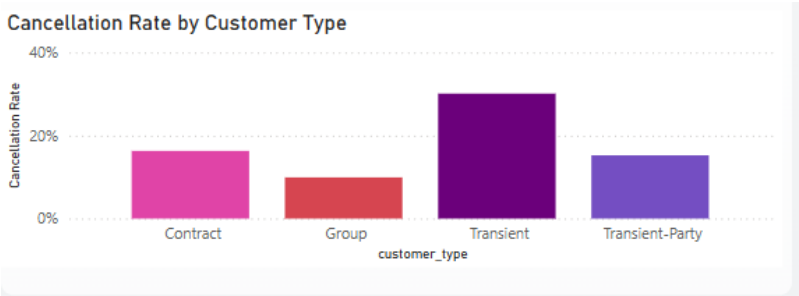
- Booking rates do not spike during summer or winter vacations, which is unexpected.
- However, ADR does increase during summer but only for City Hotels. This trend is not present for resort hotels.
- Group market segment have the smallest rate of cancellations. This makes sense as it is less likely that the entire group will cancel as compared to a single individual.
- We see the largest loss in revenue due to cancellations during the summer season (July)
- Room type A contributes the most to lost revenue. Room type D is the next highest contributor, with lost revenue. Subsequent room types (E, G, F, etc.) contribute progressively less lost revenue.

2. Business Recommendations

1)Optimizations Regarding Deposit Policy

Our analysis shows that despite the convention, customers that give a deposit actually have a higher cancellation rate. We recommend implementing the No deposit policy more often as it leads to less cancellations overall.

2)Which Customer to reject to preemptively avoid cancellations?



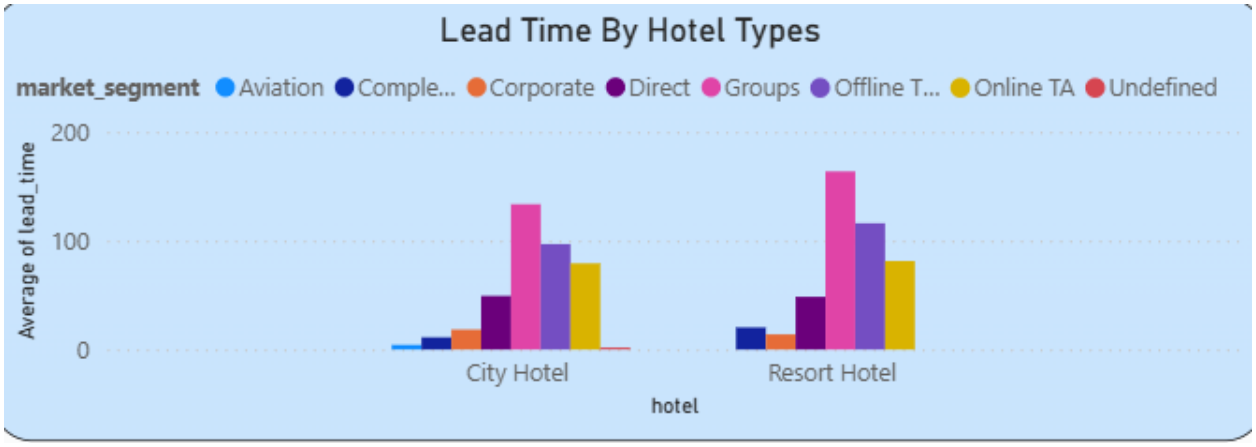
Our analysis shows that people that have previously cancelled the booking are more likely to cancel again. So, in seasons with the high number of bookings, customers that have previously made cancellations should be rejected from bookings in favour of people with no history of cancellations.

We also found out that Transient Customer type have the highest cancellation rate, so we should prioritise booking for non-Transient Customers as they are less likely to cancel. We can also use our insight that customers with previous cancellations are more likely to cancel again, so we should prioritise booking for non-Transient Customers with no prior history of cancellations as they are less likely to cancel.

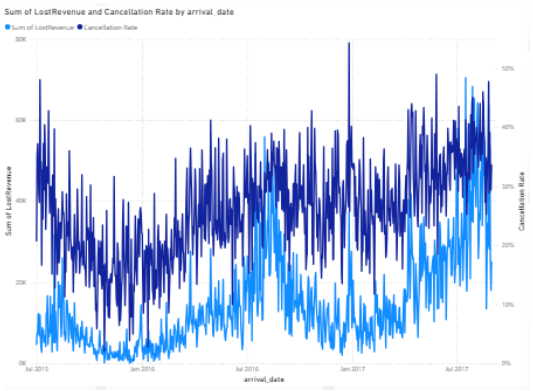
3)Priotise Group Market Segment.

Our analysis shows that the group market segment is the least likely to cancel. This makes sense as the entire group cancelling is less likely than a single individual cancelling. So we should try to attract more customers from this market segment to maximize room efficiency and minimize lost revenue due to cancellations.

However, it should be noted that Group segment has the highest lead time, so that the downside should be taken into account when room planning.

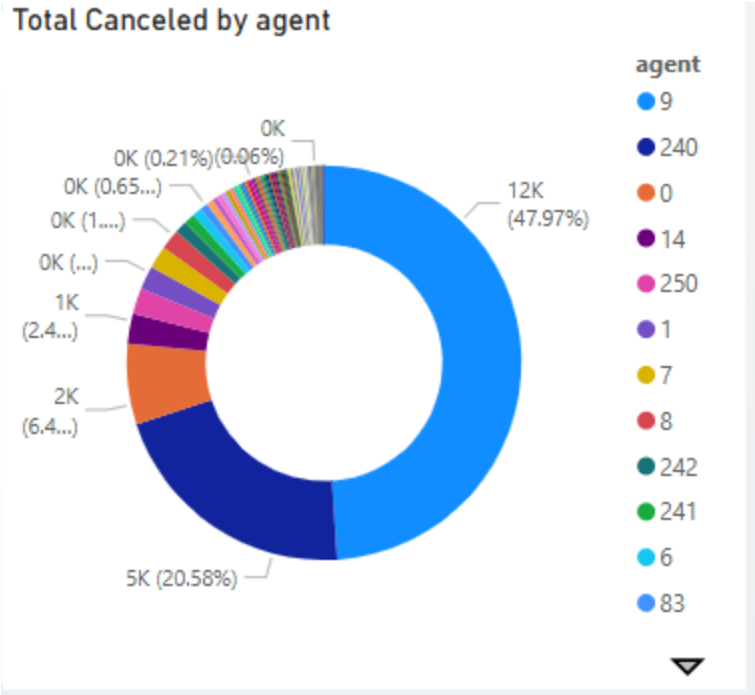


4)Target the Summer Seasons to minimize revenue lost by Cancellations.



Our analysis shows clear spikes in lost revenue during summer season(July) due to cancellations. I would be a good idea to target this month to minimize cancellations.

5)Investigate Agent 9



Almost half the cancellations are from bookings from agent 9. Do further investigation into why this is so. We could consider removing agent 9 to avoid cancellations.

Work Contribution

1)Qamar Raza : Deposit Type Dashbaord + Report + Presentation + EDA + Data Wrangling

2)Zain Sharjeel : Lead Time and Seasonality Dashboard + Interview

3)Abdullah Khalid : Market Segment Dashboad + Decompostion Tree Diagram

4)Syed Muhummand Hussain : HV1 + HV2 Dashboard + Background Data Analysis + Title Page

AI USE ACKNOWLEGEMENT

AI Tools were used in a responsible manner to supplement our analysis and report building 😊