University of Exeter

WORKING WITH DATA - MTHM501J

HOTEL GROUP SALES

Declaration

I want to acknowledge using ChatGPT as a tool to complete this project. Further, I acknowledge the use of GenAI tools in this assessment for the following:

[✓] For developing ideas.

[✓] To assist with research or gathering information.

[✓] To help me understand key theories and concepts.

[✓] To identify trends and themes as part of my data analysis

[✓] To suggest a plan or structure for my assessment.

[✓] To give me feedback on a draft.

[✓] To generate images, figures or diagrams.

[✓] To proofread and correct grammar or spelling errors.

I declare that I have referenced all use of GenAI outputs within my assessment in line with the University referencing guidelines.

Introduction

In the competitive hotel market, pricing is determined by a number of major factors such as Advance Booking, Special Requests, and Total Nights Stayed. Analysis of how these affect the average price per room and assist hotels in generating maximum revenue management while allowing customers to make wise booking choices.

Advance Booking the number of days between booking date and check-in has a price implication because of differential demand, early bookings having cheaper prices and last-minute bookings being subjected to premium pricing. Special Requests, including room category or value-added services, can increase costs because of customization and operational changes. Additionally, the Total Nights Stayed can influence the price per night, with longer stays sometimes receiving discounts or, in peak seasons, leading to higher rates due to demand trends.

This analysis explores the relationship between these factors and hotel room pricing using statistical and clustering techniques. The insights gained can assist hoteliers in refining their pricing strategies and provide valuable information for travellers looking to optimize their bookings.

Objective

The aim of this analysis is to enable customers to make booking decisions by understanding how different factors impact the average price per hotel room. More specifically, it is tested in this study whether variables like Advance Booking, Special Requests, and Total Nights Stayed bear any significant relationship with Room Price Avg.

This is accomplished by initially cleaning and transforming the data using mice and ggplot packages in a manner that missing values are handled and key variables are formatted for analysis. Multiple linear regression models will then be applied in ascertaining the statistical significance of these determinants. Statistical measures will also be applied in result interpretation so that it is understood how hotels modify their price strategies in accordance with booking patterns.

Though this we will know the Inference from evidence of data analysis to address the question and a conclusion based on result. A reproducible code is also provided at the end of the report in a bid to support transparency along with enabling further investigations.

Data

The data for this analysis was obtained from Kaggle, which is a platform that provides real-world datasets. From Kaggle, the hotel booking dataset was chosen, and this dataset provides in-depth customer booking information such as Advance Booking, Special Requests, Total Nights Stayed, and Room Price Avg.

To make the dataset clean and ready for analysis, multiple imputation techniques and visualization were done using the mice and ggplot packages. This included:

• Data cleaning by deleting unnecessary columns and fixing inconsistencies.

• Recoding categorical variables, i.e., Status, into meaningful labels.

• Missing values handling by multiple imputation methods using the mice package to increase the robustness of the analysis.

• Feature engineering, e.g., construction of a Total Nights Stayed variable to improve the identification of customer stay patterns.

Analysis and Results

The dataset used for this analysis includes key variables such as Advance Booking (days before check-in), Special Requests (additional services requested by the customer), Total Nights Stayed (duration of stay), and Room Price Avg (average price per night).

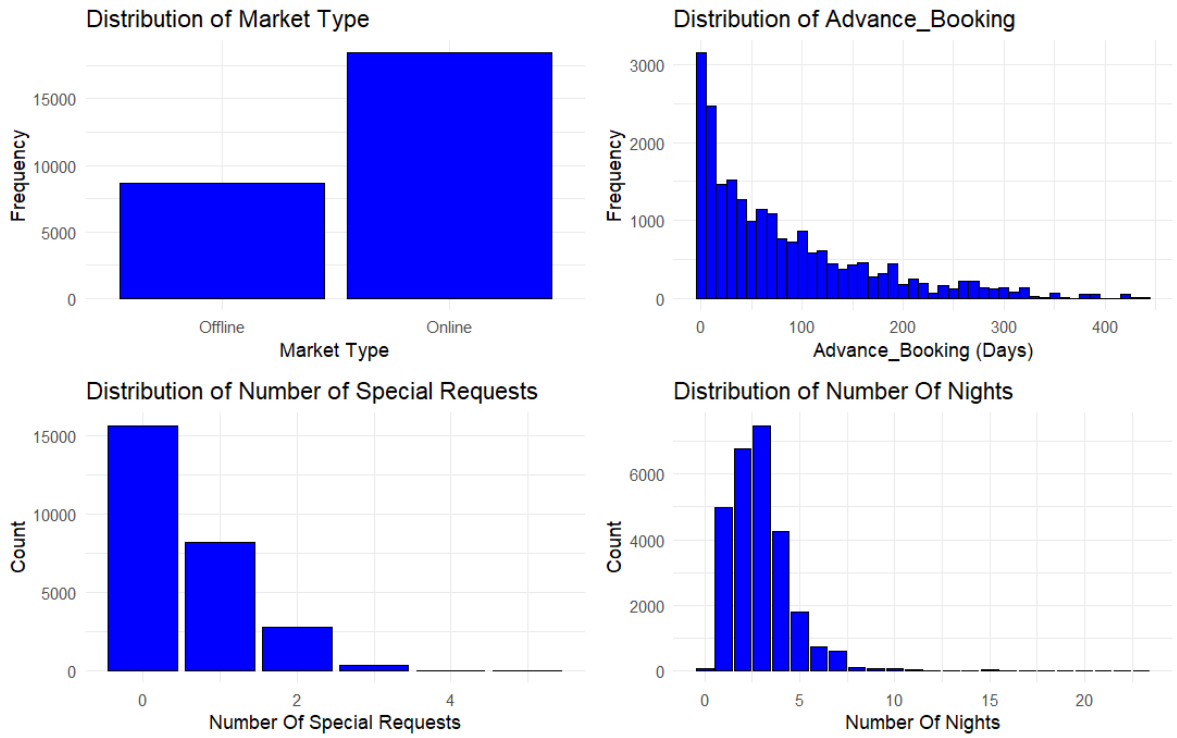
After performing data cleaning and preprocessing, we ensured that missing values were handled, and categorical variables were formatted for consistency. The data was then analysed using statistical techniques to identify relationships between booking behaviours and pricing trends.

Distribution Insights

Market Type: The data reveals that the number of online bookings is twice that of offline bookings. This suggests a growing preference for digital platforms, highlighting the need for competitive online pricing strategies and promotions.

Advance Booking Distribution: The bar graph indicates that most bookings occur either well in advance or at the last minute. We can say that the customers can demand for early discounts and hoteliers can fix a premium pricing for last-minute reservations.

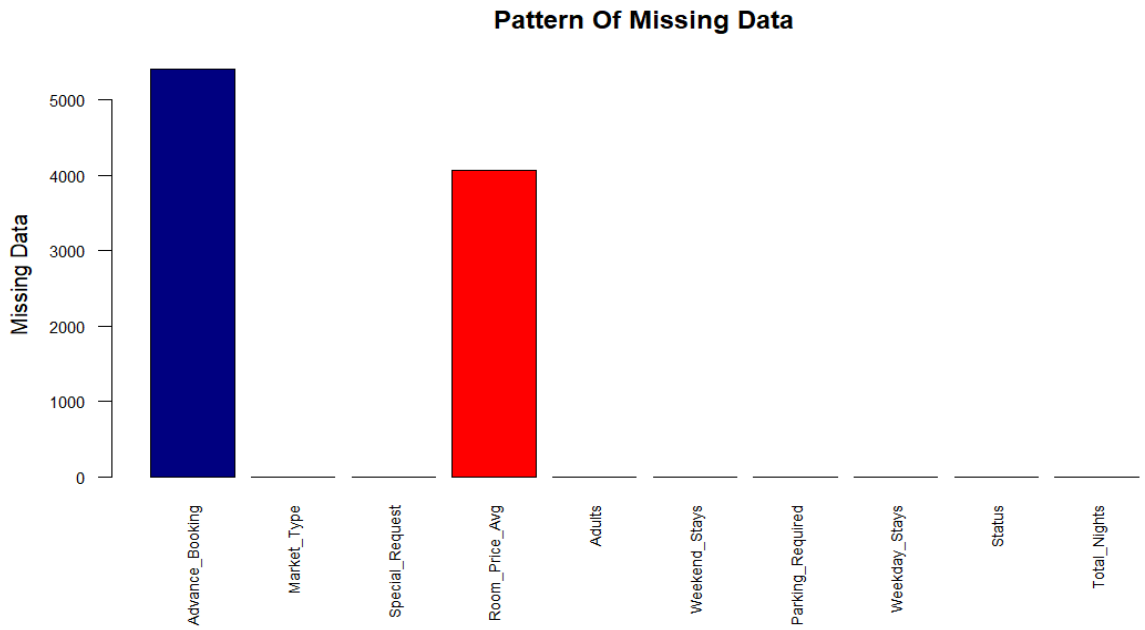
Special Requests: The distribution shows that a majority of customers don’t prefer special requests, while a smaller segment opts for premium services that contributes to the overall increase in room price based on the request made by the customers.



Total Nights Stayed: The plot reveals that most bookings are for short stays (1-3 nights), with very few guests opting for extending their stays in the hotels. This insight can help hoteliers tailor their pricing based on their guest behaviours.

Handling Missing Data & Regression Analysis

The dataset contained missing values in key columns, which were handled using multiple imputation via the mice package in R. The affected columns and their missing value percentages were:



- Advance_Booking: 20% missing
- Room_Price_Avg: 15% missing

To fill these gaps, the mice package in R was used with Predictive Mean Matching (PMM). This method replaces missing values with real observed ones, keeping the data reliable.

Term	Estimate	Std. Error	Statistic	p-Value
Intercept	96.1594	0.4768	201.70	<0.001
Advance_Booking	-0.0079	0.0030	-2.66	0.00897
Special_Request	7.8376	0.2970	26.39	<0.001
Total_Nights	0.3685	0.1311	2.81	0.00535

- The intercept of **96.16** suggests that when all independent variables are zero, the expected room price is approximately **96.16**.
- **Advance_Booking** has a small but statistically significant negative effect on room price (**β = -0.0079, p = 0.009**), indicating that booking further in advance slightly reduces the average room price.
- **Special_Request** has a strong positive effect on room price (**β = 7.84, p < 0.001**), meaning that customers making special requests tend to pay higher prices.
- **Total_Nights** also has a statistically significant positive effect (**β = 0.368, p = 0.005**), suggesting that longer stays are associated with slightly higher room prices.

Corelation Heatmap & Hypothesis Testing Analysis

A **heatmap** is a data visualization technique that uses colour to represent the magnitude of values in a matrix or dataset. It helps identify patterns, relationships, and trends within numerical data by encoding values with varying shades of colour.

- The heatmap suggests almost no correlation between Total_Nights and Room_Price_Avg (0.03), meaning that the length of stay does not significantly impact the average room price.
- There is a weak positive correlation between Room_Price_Avg and Special_Request (0.17) suggests that customers making special requests may pay slightly higher prices, but the effect is minimal.
- A weak negative correlation in Advance_Booking and Room_Price_Avg (-0.04) suggests that early bookings may be slightly associated with lower prices, but the effect is minor**.**

For each correlation test, the hypotheses are as follows:

- Null Hypothesis (H_0): There is no significant correlation between Room_Price_Avg and the given variable.
- Alternative Hypothesis (H_1): There is a significant correlation between Room_Price_Avg and the given variable.

A significance level of 0.05 (5%) was used to determine statistical significance.

Test	P-Value	Decision
Room_Price_Avg vs Advance_Booking	4.86×10^{-10}	Reject Null Hypothesis
Room_Price_Avg vs Special_Request	3.29×10^{-183}	Reject Null Hypothesis
Room_Price_Avg vs Total_Nights	1.97×10^{-6}	Reject Null Hypothesis

Since all p-values are **less than 0.05**, we reject the null hypothesis for all three tests, indicating that Room_Price_Avg has a statistically significant correlation with each of the tested variables.

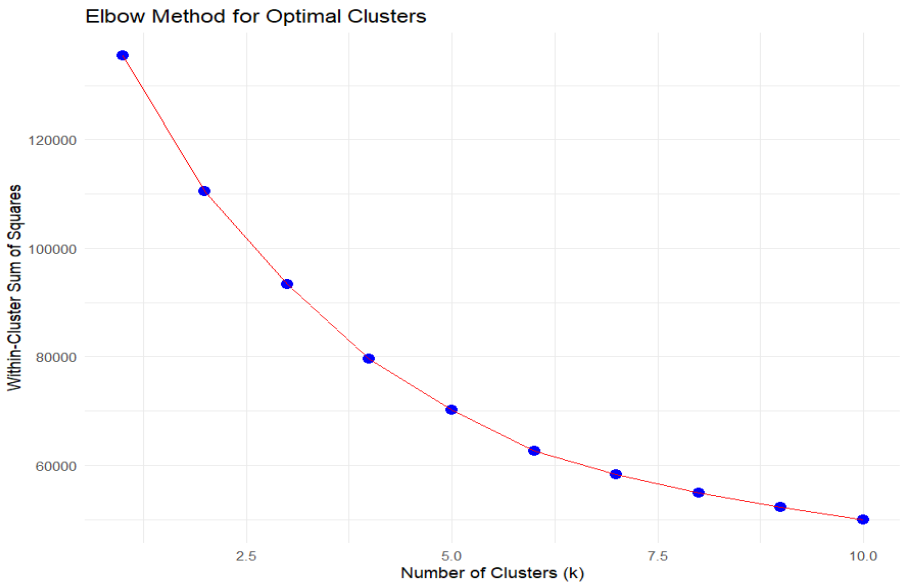
Although the correlation values are weak, the statistically significant p-values indicate that **the relationships exist**. This suggests that pricing strategies should still consider these factors, even if their direct impact is small.

Cluster Analysis

Cluster analysis is performed to group similar observations based on their characteristics. In this study, we use hierarchical clustering to segment the data into meaningful clusters. The data has features like Price of a room, Advance Booking days and number of nights. So here we are discovering hidden patterns i.e. different types of customers based on their booking behaviours.

Elbow Plot

The **Elbow Plot** is used to determine the optimal number of clusters by plotting the within-cluster sum of squares (WSS) for different cluster sizes. The point where the WSS curve starts to flatten is considered the ideal number of clusters.

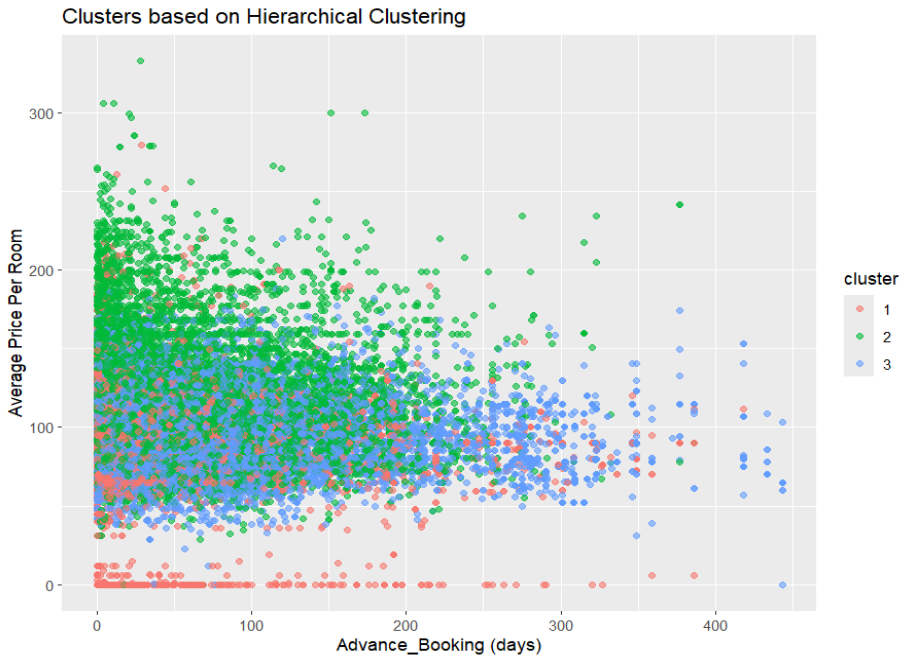


Selection of 3 Clusters

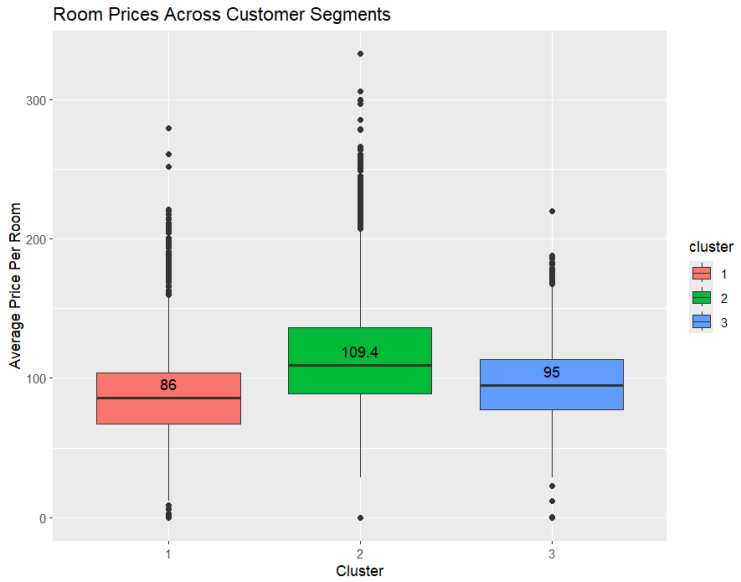
Based on the Elbow Plot and Dendrogram analysis, three clusters were chosen as the optimal number. This was to achieve a balance between intra-cluster variance minimization and meaningful segmentation. If there were too few clusters, intra-cluster variance would be high, and if there were too many, overfitting and poor generalization would occur. Three clusters allow for good differentiation without compromising interpretability.

Scatter Plot & Box Plot Visualization

A **scatter plot** is used to visualize the formed clusters in a two-dimensional space. This helps in interpreting how well the observations are separated and how distinct each cluster is.



The **box plots** is used to compare the distribution of key variables across the identified clusters. This provides insights into the differences among clusters and highlights key characteristics of each segment.



Cluster 1 (Red - Low Price, Early Booking) (Median Price: 86)

- Customers in this group book well in advance (ranging from 50 to 400+ days).
- They pay the lowest prices, benefiting from early booking discounts.
- The box plot confirms lower price variation, with a compact interquartile range (IQR).
- Likely consists of budget-conscious travellers or corporate clients who plan ahead to secure better deals.

Cluster 2 (Green - High Price, Short Notice) (Median Price: 109.4)

- This cluster consists of last-minute bookers (mostly within 0 to 150 days before stay).
- They pay the highest prices, reflecting dynamic pricing and urgent demand.
- The wide IQR and presence of many outliers suggest significant price fluctuations based on demand surges.
- Likely represents business travellers or urgent bookings, where convenience and availability outweigh price considerations.

Cluster 3 (Blue - Mid-range, Balanced Booking Time) (Median Price: 95)

- This cluster has a balanced distribution of booking times, from early to last-minute.
- The average price is moderate, sitting between clusters 1 and 2.
- The box plot reveals some price variation but not as extreme as Cluster 2.
- Represents a diverse group of leisure travellers and flexible planners who do not follow strict booking patterns but seek a balance between price and timing.

Limitations

1. **External Factors Not Accounted:** Seasonal trends, festive rates, and holidays play a significant role in room rates but were not covered in this study. Prices tend to skyrocket in peak travel periods, which has an impact on booking behaviour during cluster. Excluding these could lead to inaccurate customer segmentation.
2. **Potential Bias in Dataset:** The data might not represent all categories of hotels, for example, luxury, budget, or boutique hotels, equally. If the dataset is biased by some of the hotel types, the clusters discovered might not generalize to the entire market. A more representative dataset would render the clustering outcomes more stable.
3. **Imputed Missing Data and Uncertainty:** Missing values were addressed using statistical imputation, in which information is estimated based on current patterns. Although the process allows for completeness, it adds a degree of uncertainty. Certain booking and pricing behaviour patterns may be skewed as a result of such imputed values.
4. **Complexity of Dynamic Pricing Models:** Hotels tend to change room rates in real time based on demand, competition, and other market factors. Such dynamic fluctuations were not included in the clustering analysis as they might compromise the precision of price segmentation. Hence, last-minute price hikes or drops might not be reflected in the resultant clusters to their highest extent.

Conclusion

This analysis provides valuable insights into the key factors influencing hotel room rates, such as advance booking, special requests, and nights stayed. Based on statistical analysis and hierarchical clustering, we distinguished between three customer segments early bookers with discounted prices, last-minute bookers with high-end prices, and a balanced segment with medium price. These results show how dynamic pricing and booking times play influential roles in shaping hotel room rates, with practical implications for hoteliers and consumers.

The hypothesis testing validated the statistically significant relationship between room prices and the variables involved, highlighting the necessity for strategic price changes. Cluster analysis also highlighted that advance discounts bring in price-sensitive tourists, whereas last-minute bookings are more expensive because of demand spikes. These findings can assist hotels in streamlining revenue management strategies and tourists in making informed bookings.

However, some limitations have to be taken into consideration, such as seasonality in price variation, bias in the data sample, and the influence of real-time dynamic pricing. Follow-up research could include a bigger data sample, real-time price changes, and outside market conditions in order to increase the validity and generalizability of these results.

APPENDIX

```
# Install required packages if not installed
# install.packages("mice")
# install.packages("dplyr")
# install.packages("ggplot2")
# install.packages("reshape2")
# install.packages("patchwork")

# Load required libraries
library(mice)
library(dplyr)
library(ggplot2)
library(reshape2)
library(patchwork)

# Load the dataset
df <- read.csv("C:/Users/moham/OneDrive/Desktop/Rassignment/INNHotelsGroup.csv")

# --- EDA ---

# Combining 2 columns to get a single column
df <- df %>%
  mutate(Total_Nights = no_of_week_nights + no_of_weekend_nights)

# Remove unwanted columns
df <- df %>%
  select(-booking_id, no_of_week_nights, no_of_weekend_nights, -arrival_date, -rebooked)

# Repalcing 'Not Cancelled' to 'Confirmed' for better clarity
```

```
df$booking_status <- ifelse(df$booking_status == "Not Canceled", "Confirmed",
df$booking_status)

# Rename columns with meaningless names
df <- df %>% rename(
  Advance_Booking = lead_time,
  Market_Type = market_segment_type,
  Special_Request = no_of_special_requests,
  Room_Price_Avg = avg_price_per_room,
  Adults = no_of_adults,
  Parking_Required = required_car_parking_space,
  Status = booking_status)

# --- Visualization ---

# Visualize distributions
p1 <- ggplot(df, aes(x = Market_Type)) +
  geom_bar(fill = "blue", color = "black") +
  theme_minimal() +
  ggtitle("Distribution of Market Type") +
  xlab("Market Type") +
  ylab("Frequency")

p2 <- ggplot(df, aes(x=Advance_Booking)) +
  geom_histogram(binwidth=10, fill="blue", color="black") +
  theme_minimal() +
  ggtitle("Distribution of Advance_Booking") +
  xlab("Advance_Booking (Days)") +
  ylab("Frequency")
```

```
p3 <- ggplot(df, aes(x=Special_Request)) +
  geom_bar(fill="blue", color="black") +
  theme_minimal() +
  ggtitle("Distribution of Number of Special Requests") +
  xlab("Number Of Special Requests") +
  ylab("Count")

p4 <- ggplot(df, aes(x=Total_Nights)) +
  geom_bar(fill="blue", color="black") +
  theme_minimal() +
  ggtitle("Distribution of Number Of Nights") +
  xlab("Number Of Nights") +
  ylab("Count")

# Patchwork layout
(p1 | p2) / (p3 | p4)

# --- Missing Value Imputation ---

# Missing values
missing_values <- sapply(df, function(x) sum(is.na(x)))
missing_columns <- names(missing_values[missing_values > 0])
missing_values_per_column <- missing_values[missing_columns]
print("-------------------------")
print("Missing Values Per Column:")
print(missing_values_per_column)

# Visualize missing values
missing_data <- colSums(is.na(df))
```

```
par(mar=c(9, 4, 4, 2))
barplot(missing_data,
        col=c('navyblue', 'red'),
        names.arg=names(df),
        cex.axis=0.75,
        cex.names=0.7,
        las=2,
        ylab="Missing Data",
        main="Pattern Of Missing Data")

# Impute missing values using mice
imputed_data <- mice(df, m = 10, method = 'pmm', seed = 500, printFlag = FALSE)

# Extract the complete datasets
completed_data <- complete(imputed_data, "long", include = TRUE)

# Fit regression models on each imputed dataset
lm_models <- with(imputed_data, lm(Room_Price_Avg ~ Advance_Booking +
Special_Request + Total_Nights))

# Pool the results using Rubin's rules
pooled_results <- pool(lm_models)

# Print combined results
print("--------------------------")
print("Combined Results From Pooled Regression Models:")
summary(pooled_results)

# Extracting the complete data set
# Here we have used the 1st imputation for further analysis
```

```
complete_data <- complete(imputed_data, 1)
print("-------------------------")
print("Complete Dataset For Further Analysis:")
print(head(complete_data))

# --- Hypothesis Testing ---

# Perform correlation tests between Room_Price_Avg and other variables
test1 <- cor.test(complete_data$Room_Price_Avg, complete_data$Advance_Booking)
test2 <- cor.test(complete_data$Room_Price_Avg, complete_data$Special_Request)
test3 <- cor.test(complete_data$Room_Price_Avg, complete_data$Total_Nights)

# Frame hypotheses
hypothesis_results <- data.frame(
  Test = c("Room_Price_Avg vs Advance_Booking",
           "Room_Price_Avg vs Special_Request",
           "Room_Price_Avg vs Total_Nights"),
  P_Value = c(test1$p.value, test2$p.value, test3$p.value),
  Decision = ifelse(c(test1$p.value, test2$p.value, test3$p.value) < 0.05,
                    "Reject Null Hypothesis", "Fail to Reject Null Hypothesis")
)
print(hypothesis_results)

# --- Correlation Heatmap ---

# Calculate the correlation matrix
correlation_matrix <- cor(complete_data[, c("Room_Price_Avg", "Advance_Booking",
"Special_Request", "Total_Nights")], use = "complete.obs")
# Melt the correlation matrix into a long format
melted_correlation_matrix <- melt(correlation_matrix)
```

```
# Create the heatmap with correlation values and no x-axis labels
heatmap_plot <- ggplot(data = melted_correlation_matrix, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1, 1), space = "Lab",
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_blank(),  # Remove x-axis labels
        axis.text.y = element_text(size = 12),  # Adjust y-axis label size
        axis.title.x = element_blank(),  # Remove x-axis title
        axis.title.y = element_blank(),  # Remove y-axis title
        axis.ticks.x = element_blank(),  # Remove x-axis ticks
        axis.ticks.y = element_blank(),  # Remove y-axis ticks
        plot.title = element_text(hjust = 0.5)) +  # Center the plot title
  coord_fixed() +
  labs(title = "Correlation Heatmap") +
  geom_text(aes(label = round(value, 2)), color = "black", size = 4)  # Add correlation values
print(heatmap_plot)

# --- Hierarchical Clustering ---

# Use the first imputed dataset for clustering
clustering_data <- complete_data %>%
  select(Advance_Booking, Special_Request, Room_Price_Avg, Adults, Total_Nights) %>%
  scale()  # Normalize numerical data
# Compute distance matrix
distance_matrix <- dist(clustering_data, method = "euclidean")
```

```
# Perform hierarchical clustering
hc <- hclust(distance_matrix, method = "ward.D")

# Select relevant variables for clustering
clustering_data <- complete_data %>%
  select(Advance_Booking, Special_Request, Room_Price_Avg, Adults, Total_Nights) %>%
  scale()  # Normalize numerical data

# Compute the within-cluster sum of squares (WCSS) for different values of k
wcss <- sapply(1:10, function(k){
  kmeans(clustering_data, centers = k, nstart = 25)$tot.withinss
})

# Create elbow plot
elbow_plot <- ggplot(data.frame(k = 1:10, wcss = wcss), aes(x = k, y = wcss)) +
  geom_point(size = 3, color = "blue") +
  geom_line(color = "red") +
  labs(title = "Elbow Method for Optimal Clusters", x = "Number of Clusters (k)", y = "Within-Cluster Sum of Squares") +
  theme_minimal()
print(elbow_plot)

# Cut tree into k clusters
k <- 3
clusters <- cutree(hc, k)

# Add cluster labels to original data set
complete_data$cluster <- as.factor(clusters)

# Analyze clusters
```

```
print(summary(complete_data$cluster))


# Visualize clusters
ggplot(complete_data, aes(x = Advance_Booking, y = Room_Price_Avg, color = cluster)) +
  geom_point(alpha = 0.6) +
  labs(title = "Clusters based on Hierarchical Clustering", x = "Advance_Booking (days)", y = "Average Price Per Room")


# Visualize Price Distribution Across Clusters
ggplot(complete_data, aes(x = cluster, y = Room_Price_Avg, fill = cluster)) +
  geom_boxplot() +
  stat_summary(fun = median, geom = "text", aes(label = round(..y.., 2)),
               position = position_nudge(y = 10), color = "black") +  # Add median values as text
  labs(title = "Room Prices Across Customer Segments", x = "Cluster", y = "Average Price Per Room")
```