Cover Page

Name of the school: The George Washington University

Name of the author: Harshavardana Reddy Kolan

Name of the instructor: Reza Jafari

Date: 04/26/2024

Course Name: Data Visualization

Course Number: DATS 6401

App link:

https://dashapp-sk2ykdtmqq-ue.a.run.app/

Table of Contents:

Table of figures and tables:

1. **Abstract:**

This project aims to analyze a comprehensive dataset containing information on hotel bookings, encompassing various attributes related to guests, booking details, and reservation statuses. With a focus on understanding factors influencing booking cancellations, booking patterns, and guest preferences, the analysis employs statistical techniques and machine learning algorithms. Through exploratory data analysis and predictive modeling, the project seeks to uncover insights that can inform strategies for enhancing guest satisfaction, optimizing hotel operations, and minimizing revenue loss due to cancellations.

2.Introduction:

In the hospitality industry, managing hotel bookings effectively is paramount for ensuring guest satisfaction and maximizing revenue. Understanding the dynamics of booking patterns, guest preferences, and factors contributing to cancellations is crucial for hotel management to make informed decisions and streamline operations. This project delves into a detailed dataset encompassing various aspects of hotel bookings, including guest demographics, booking behavior, and reservation statuses.

The dataset offers a rich source of information for analyzing trends, identifying patterns, and extracting actionable insights. By leveraging statistical analysis and machine learning techniques, this project aims to explore the relationships between different attributes and booking outcomes. Specifically, the project focuses on investigating the following key areas:

1. Booking Cancellations: Understanding the factors contributing to booking cancellations is essential for minimizing revenue loss and optimizing resource allocation. By examining variables such as lead time, previous cancellations, and booking changes, we aim to identify patterns and predictors of cancellations.

2. Guest Preferences: Analyzing attributes such as meal preferences, special requests, and room type preferences can provide valuable insights into guest preferences and expectations. By understanding these preferences, hotels can tailor their services to enhance guest satisfaction and loyalty.

3. Booking Patterns: Examining trends in booking patterns, such as seasonality, market segments, and distribution channels, can help hotels optimize pricing strategies, marketing efforts, and inventory management.

Through this analysis, we aim to provide actionable recommendations for hotel management to improve operational efficiency, enhance guest experiences, and mitigate risks associated with booking cancellations. By leveraging data-driven insights, hotels can better anticipate guest needs, personalize services, and ultimately drive business success in a competitive hospitality landscape.

3.Dataset Description:

The dataset contains information about hotel bookings, including various attributes related to the guests, booking details, and reservation status.

Attributes:

1. **Hotel:** Type of hotel (Resort Hotel or City Hotel).
2. **is_canceled:** Binary variable indicating if the booking was canceled (1) or not (0).
3. **lead_time:** Number of days between booking date and arrival date.
4. **arrival_date_year:** Year of arrival date.
5. **arrival_date_month:** Month of arrival date.
6. **arrival_date_week_number:** Week number of arrival date.
7. **arrival_date_day_of_month:** Day of the month of arrival date.
8. **stays_in_weekend_nights:** Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay.
9. **stays_in_week_nights:** Number of week nights (Monday to Friday) the guest stayed or booked to stay.
10. **adults:** Number of adults.
11. **children:** Number of children.
12. **babies:** Number of babies.
13. **meal:** Type of meal booked.
14. **country:** Country of origin.
15. **market_segment:** Market segment designation.
16. **distribution_channel:** Booking distribution channel.
17. **is_repeated_guest:** Binary variable indicating if the guest is a repeated guest (1) or not (0).
18. **previous_cancellations:** Number of previous bookings that were canceled by the guest.
19. **previous_bookings_not_canceled:** Number of previous bookings that were not canceled by the guest.
20. **reserved_room_type:** Code of room type reserved.
21. **assigned_room_type:** Code for the type of room assigned to the booking.

22. **booking_changes:** Number of changes made to the booking.
23. **deposit_type:** Type of deposit made for the reservation.
24. **agent:** ID of the travel agency making the booking.
25. **company:** ID of the company/entity making the booking.
26. **days_in_waiting_list:** Number of days the booking was in the waiting list before it was confirmed.
27. **customer_type:** Type of booking (Transient, Contract, Group, or Transient-Party).
28. **adr:** Average daily rate (sum of transaction revenue divided by the number of staying nights).
29. **required_car_parking_spaces:** Number of car parking spaces required by the guest.
30. **total_of_special_requests:** Number of special requests made by the guest.
31. **reservation_status:** Last reservation status (Canceled, Check-Out, or No-Show).
32. **reservation_status_date:** Date at which the last status was set.

The dataset provides comprehensive information about hotel bookings, including guest demographics, booking behavior, and reservation status. Analysis of this dataset can offer insights into factors influencing booking cancellations, booking patterns, and guest preferences.

4.Data Preprocessing:

I performed several preprocessing steps on the dataset to handle missing values and engineer new features for enhanced analysis.

1. Identifying Missing Values:
    - Initially, I computed the percentage of missing values for each column in the dataset using the isna().sum() method.
    - Subsequently, I visualized the percentage of missing values for each column using a bar plot to understand the extent of missingness across different features.
2. Handling Missing Values:
    - For columns with missing values, I applied various strategies for imputation:
        - Replaced missing values in the "children," "country," "agent," and "company" columns with appropriate replacements such as 0 or "Unknown."
        - Imputed missing values in the "meal" column by replacing "Undefined" values with "SC" (Standard Charcoal).

        - Dropped rows with any remaining missing values after the initial imputation steps to ensure data completeness.

3. Feature Engineering:
    - Created new features to extract additional insights from the existing data:
        - Calculated the total number of guests ("total_people") by summing the counts of adults, children, and babies.
        - Computed the total number of nights stayed ("total_nights") by aggregating weekend nights and weeknights.
        - Derived the ratio of special requests per person ("special_requests_per_person") by dividing the total number of special requests by the total number of guests.
        - Calculated the lead time per night ("lead_time_per_night") by dividing the lead time by the total number of nights.
        - Obtained the average daily rate per person ("adr_per_person") by dividing the average daily rate by the total number of guests.
        - Generated the booking to arrival ratio ("booking_to_arrival_ratio") by dividing the lead time by the total number of nights.
4. Handling Remaining Missing Values:
    - After feature engineering, I checked for any remaining missing values and replaced them with the mean value of each respective column to ensure data completeness and maintain consistency.
5. Verification:
    - Finally, I validated the effectiveness of the preprocessing steps by rechecking for missing values using the isna().sum() method and confirmed that no missing values remained in the dataset.

By performing these preprocessing steps, I ensured that the dataset was ready for subsequent analysis, with missing values addressed and new informative features engineered to enrich the dataset's insights.

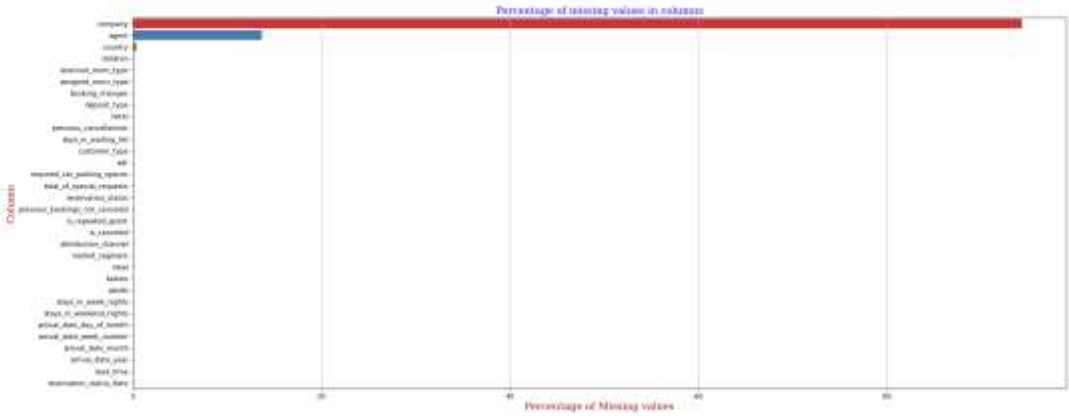


Fig 1- Bar Chart representing percentage of missing values in dataset

5. Outlier Detection and Removal:

I performed outlier detection and removal on the numerical columns of the dataset to ensure the robustness of subsequent analysis and modeling. Here's a summary of the outlier detection and removal process:

1. **Identifying Numerical Columns**:
    - Initially, I extracted the numerical columns relevant for outlier detection, including 'special_requests_per_person', 'lead_time_per_night', 'adr_per_person', 'booking_to_arrival_ratio', and 'adr'.
2. **Outlier Detection Using IQR Method**:
    - I applied the Interquartile Range (IQR) method to identify outliers in each numerical column.
    - For each column, I calculated the first quartile (Q1), third quartile (Q3), and the interquartile range (IQR).
    - Outliers were identified as values lying outside the range of 1.5 times the IQR below Q1 and above Q3.
3. **Outlier Removal**:
    - After identifying outliers in each column, I created a cleaned version of the dataset by filtering out the outliers.
    - Outliers were removed by retaining only the data points that fell within the acceptable range defined by the IQR method.
4. **Visualization**:

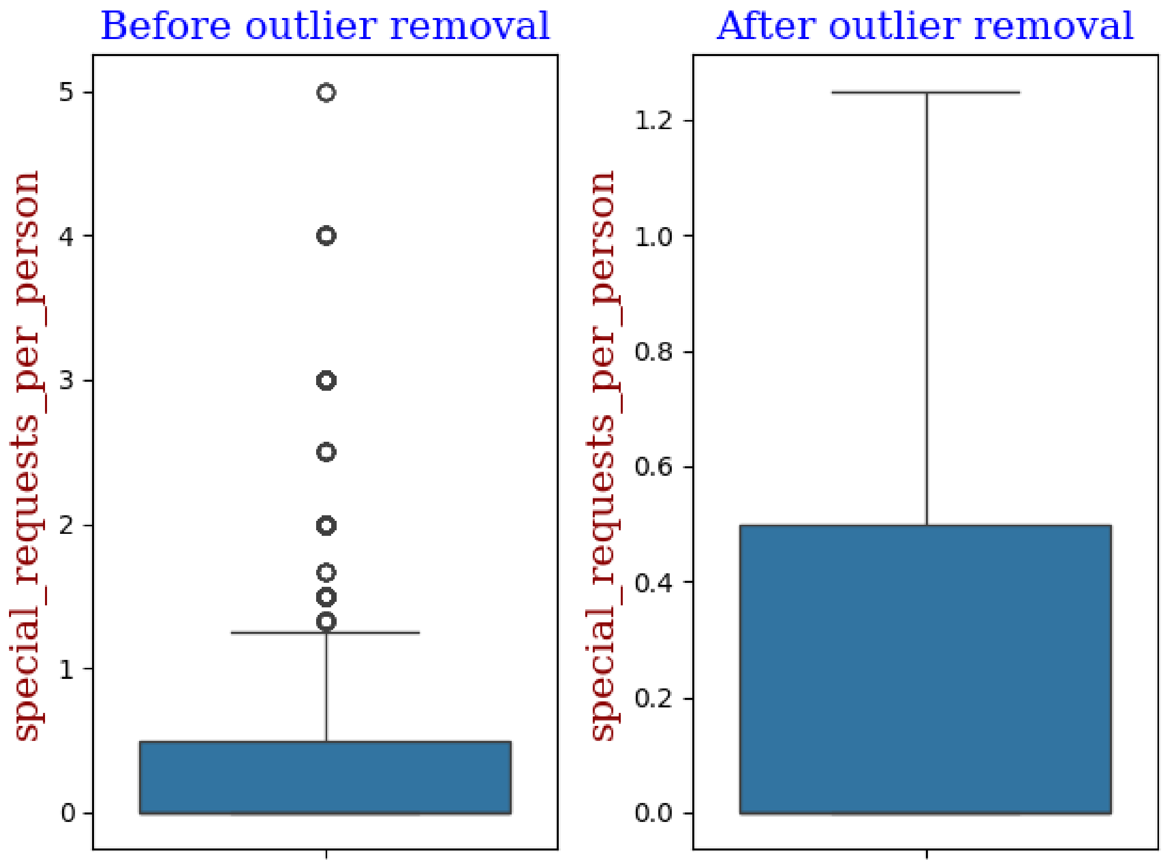


Fig2- Outlier box subplot showing before and after outlier removal for special requests per person

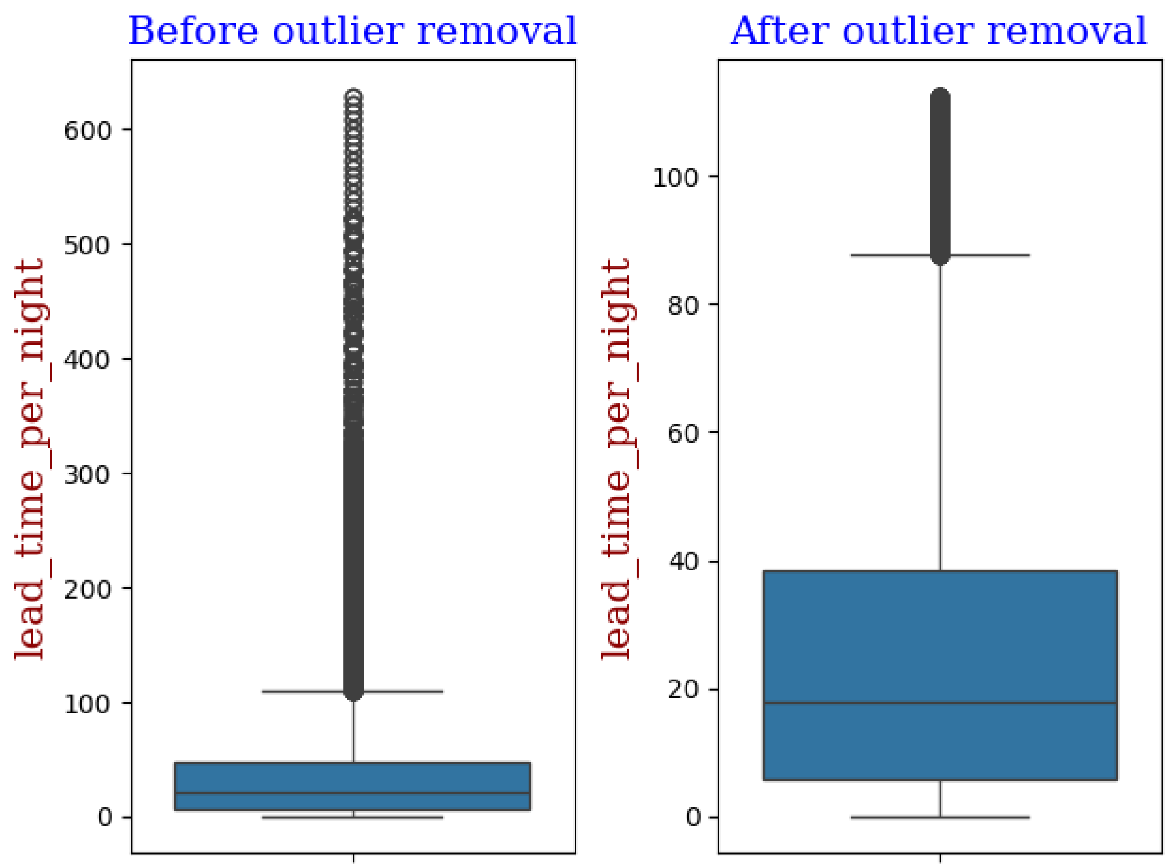


Fig3- Outlier box subplot showing before and after outlier removal for lead time per night

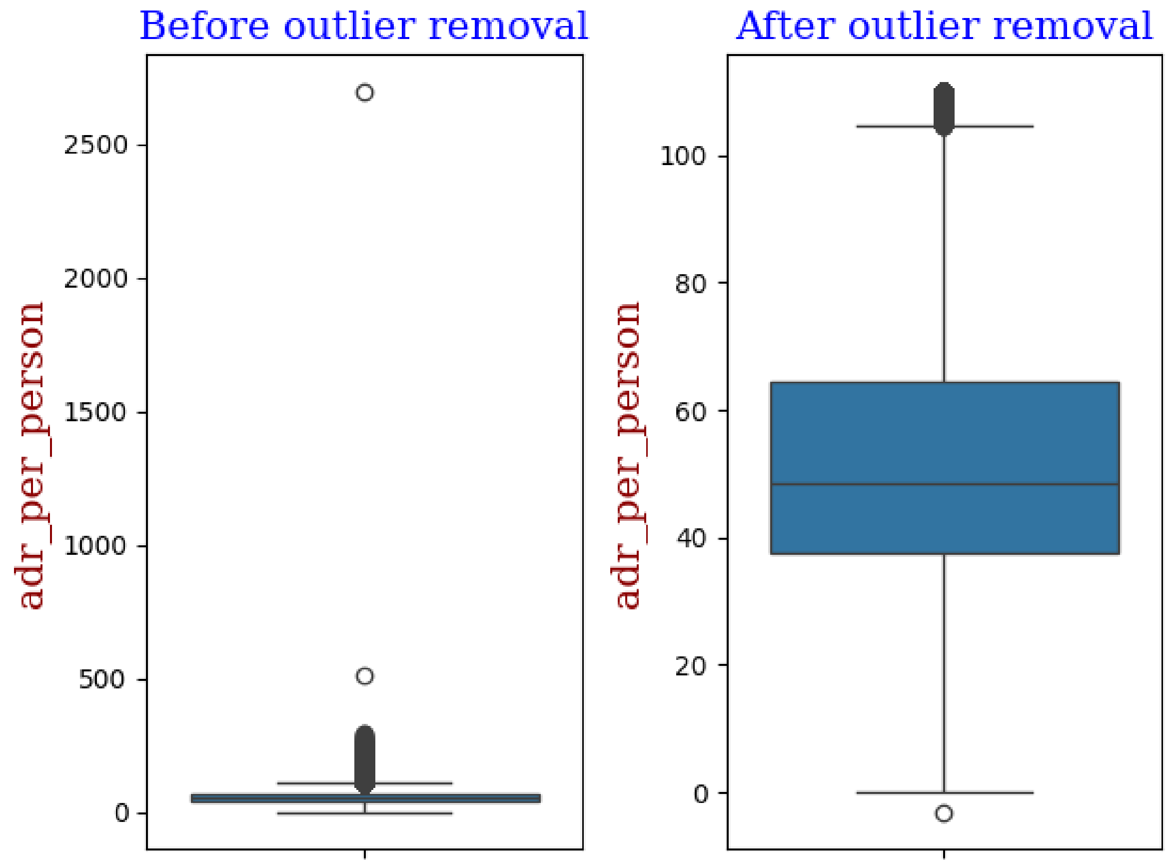


Fig4- Outlier box subplot showing before and after outlier removal for adr per person

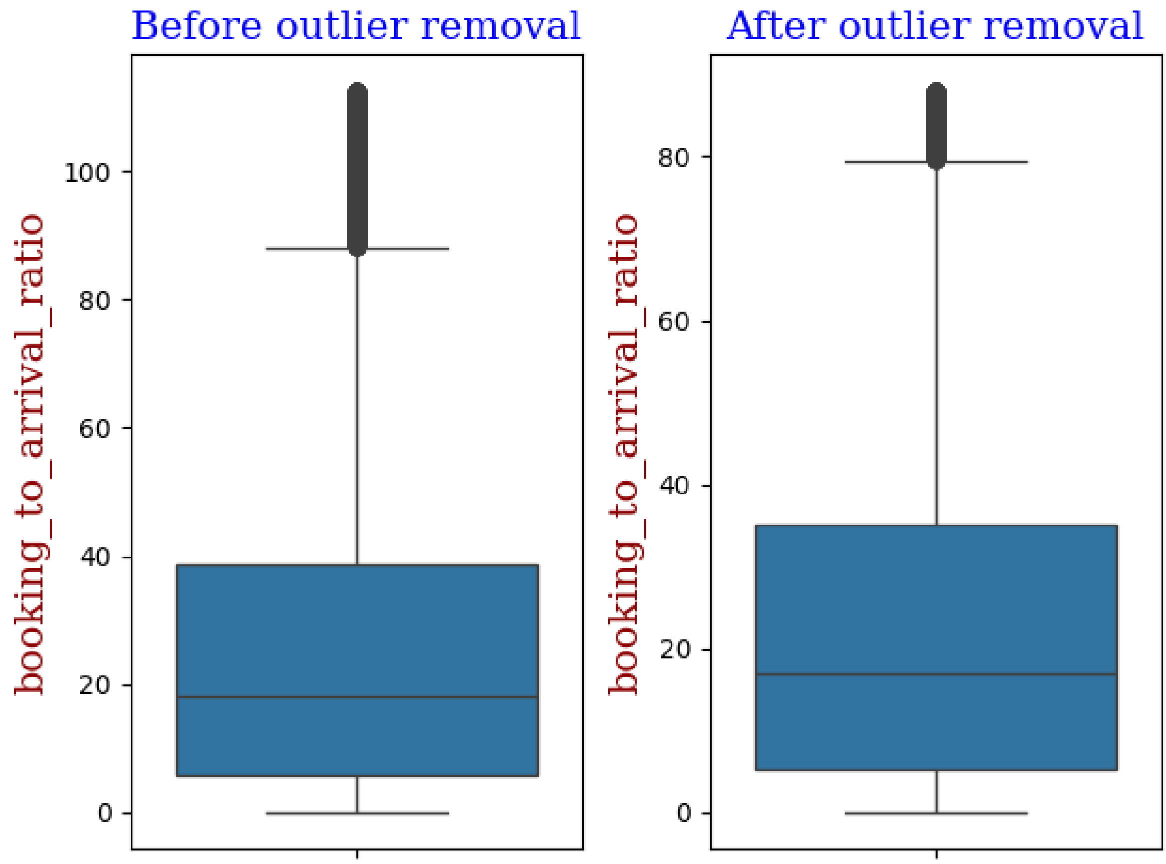


Fig5- Outlier box subplot showing before and after outlier removal for booking to arrival ratio

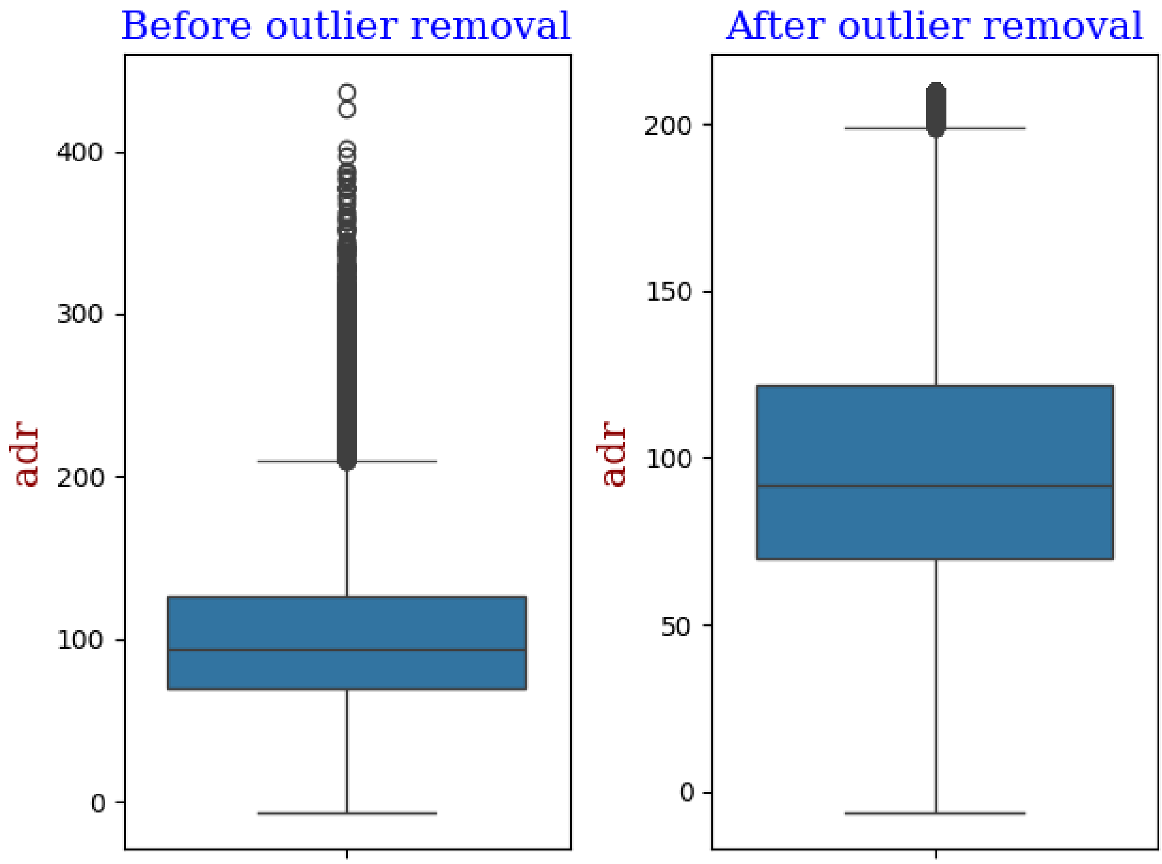


Fig6- Outlier box subplot showing before and after outlier removal for adr

- To visualize the effectiveness of outlier removal, I plotted boxplots before and after outlier removal for each numerical column.
- The boxplots allowed for a visual comparison of the distribution of data points before and after outlier removal, highlighting the impact of the outlier removal process.

5. **Verification**:

- Following outlier removal, I verified the completeness of the dataset by checking for any remaining missing values using the **isna().sum()** method.
- Any remaining missing values were dropped from the dataset to ensure data integrity and consistency.

By performing outlier detection and removal using the IQR method, I enhanced the quality of the dataset by mitigating the influence of outliers on subsequent analysis and modeling tasks. This preprocessing step contributes to more reliable and robust insights derived from the dataset.

6.PCA Analysis:

I conducted a Principal Component Analysis (PCA) on the numerical features of the dataset to reduce dimensionality and extract meaningful patterns from the data. Here's a summary of the PCA analysis:

1. Data Standardization:
    - Initially, I standardized the numerical features using the StandardScaler to ensure that all variables have a mean of 0 and a standard deviation of 1, thereby preventing features with larger scales from dominating the analysis.
2. Correlation Analysis:
    - I calculated the correlation coefficient matrix between the numerical features to identify correlations among them.
    - The correlation heatmap visualization provided insights into the relationships between different features, guiding the subsequent PCA analysis.
3. PCA Application:
    - PCA was applied to the standardized numerical features to transform the data into a new set of orthogonal variables called principal components.
    - I determined the number of principal components required to explain at least 90% of the variance in the data.
    - The condition number, which indicates the stability of the PCA solution, was calculated for both the original and reduced feature spaces.
4. Results and Interpretation:
    - The original feature space had a high condition number, suggesting potential multicollinearity issues due to high inter-feature correlations.
    - PCA revealed that two features could be removed while retaining 90% of the variance, resulting in a reduced feature space with a lower condition number.
    - The explained variance ratio showed that the first three principal components explained the majority of the variance in the data.
    - The cumulative explained variance plot depicted the trade-off between the number of principal components and the percentage of explained variance.

By performing PCA analysis, I effectively reduced the dimensionality of the dataset while preserving most of the variance, facilitating further analysis and interpretation of the data. This preprocessing step contributes to improved model performance and interpretability in subsequent modeling tasks.

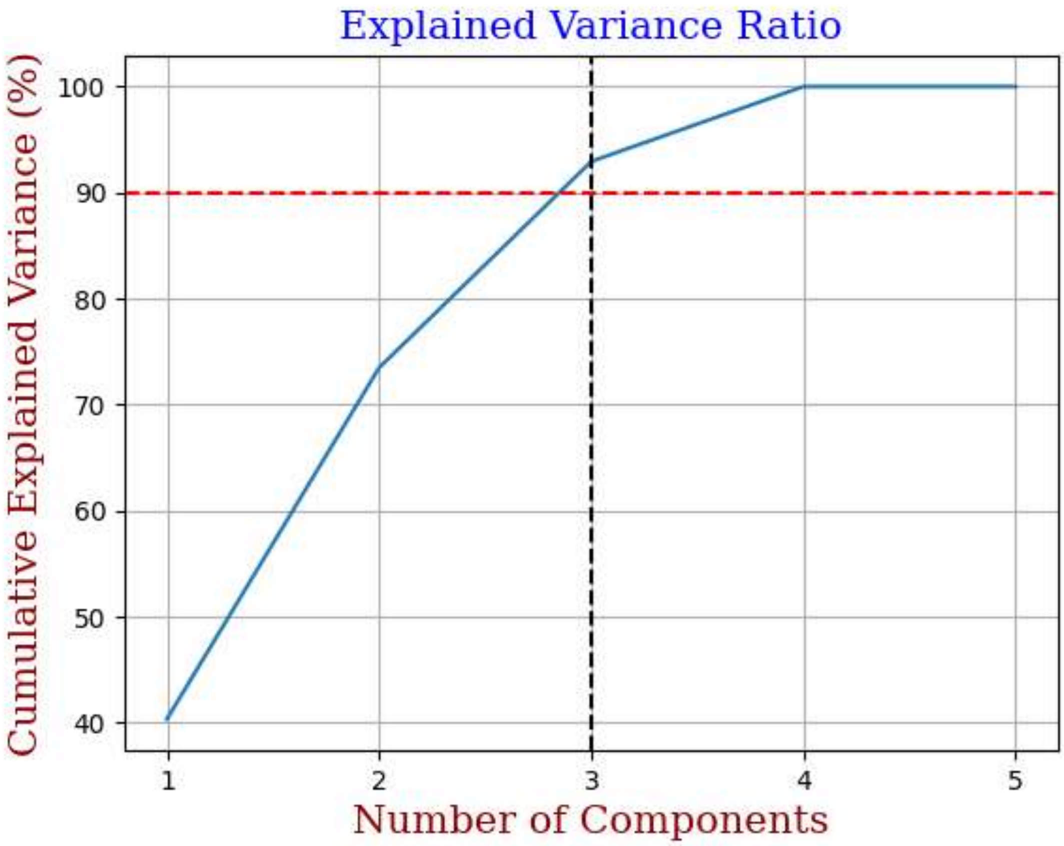


Fig 7 – Graph showing comenents for more than 90% cumulative explained variance

7.Normality Tests:

To assess the normality of the numerical features in the dataset, three different statistical tests were conducted: the D'Agostino and Pearson's omnibus test (Da_k_squared test), the Kolmogorov-Smirnov (K_S) test, and the Shapiro-Wilk test.

For the feature 'special_requests_per_person', all three tests rejected the null hypothesis of normality (p-values < 0.01), indicating that the data is not normally distributed. Similarly, for 'lead_time_per_night', 'adr_per_person', 'booking_to_arrival_ratio', and 'adr', all tests yielded p-values < 0.01, indicating non-normality of the data.

Specifically, the Da_k_squared test statistics were 128.48, 143.85, 58.93, and 70.70 for 'special_requests_per_person', 'lead_time_per_night', 'adr_per_person', and 'adr', respectively. The K_S test statistics ranged from 0.18 to 0.52 for these features, while the Shapiro-Wilk test statistics ranged from 0.37 to 0.85.

These results suggest that the assumption of normality does not hold for the examined features. Non-parametric methods or transformations may be more appropriate for further analysis to account for the non-normality of the data.

Column	Da_k_squared statistics	Da_k_squared p-value	K_S statistics	K_S p-valueShapiro statistics	Shapiro p-value	Normality
special_requests_per_person	11370.63	0.00	0.36	0.000.72	0.00	Not Normal
lead_time_per_night	11685.17	0.00	0.14	0.000.89	0.00	Not Normal
adr_per_person	3643.08	0.00	0.07	0.000.98	0.00	Not Normal
booking_to_arrival_ratio	11685.17	0.00	0.14	0.000.89	0.00	Not Normal
adr	2753.65	0.00	0.07	0.000.98	0.00	Not Normal

Table1- Normality Tests table.

8.Heatmap & Scatter Plot showing correlation:

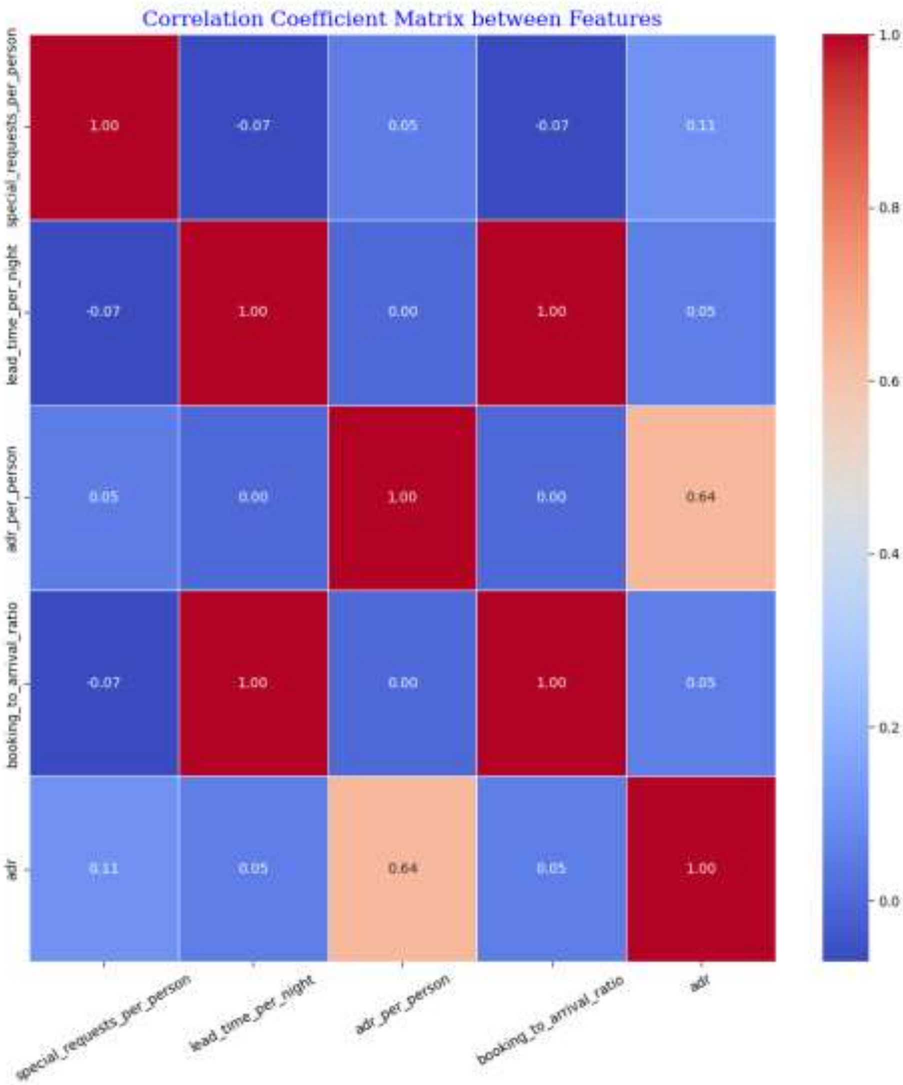


Fig8- Heatmap showing correlation between features

Observation:

The graph shows that the variables booking_to_arrival_ratio and lead_time_per night are highly correlated with the positive correlation of 1. This makes sense because the booking_to_arrival_ratio was calculated using lead time. Then, adr and adr_per_person are correlation with a high positive correlation of 0.64.

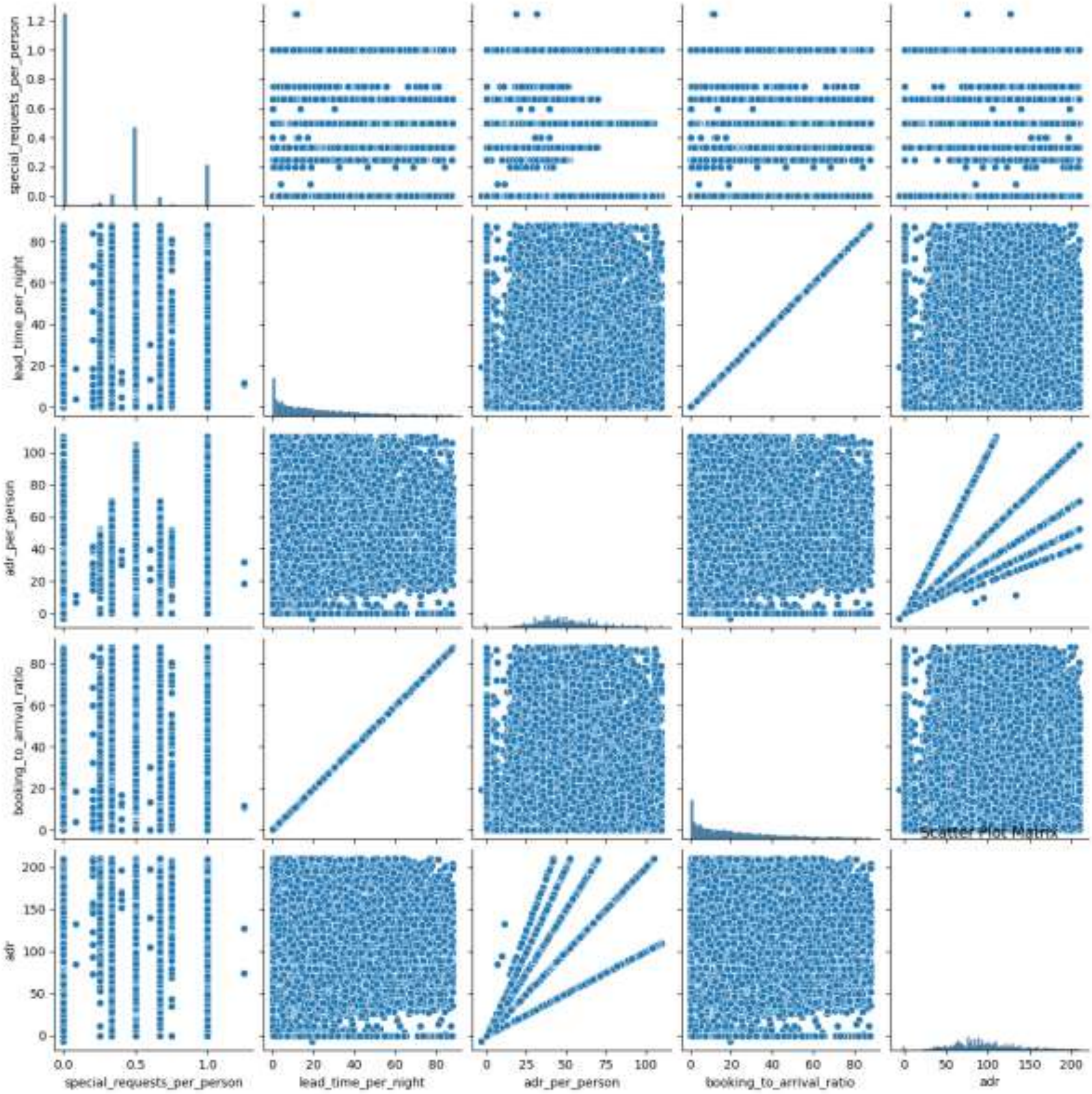


Fig9- Scatter matrix showing correlation between features

Observation:

The scatter matrix confirms the correlation plot as the variables booking_to_arrival_ratio and lead_time_per night are highly correlated with the positive correlation. Then, adr and adr_per_person are correlation with a high positive correlation.

9.Statistics(Multivariate KDE plot)

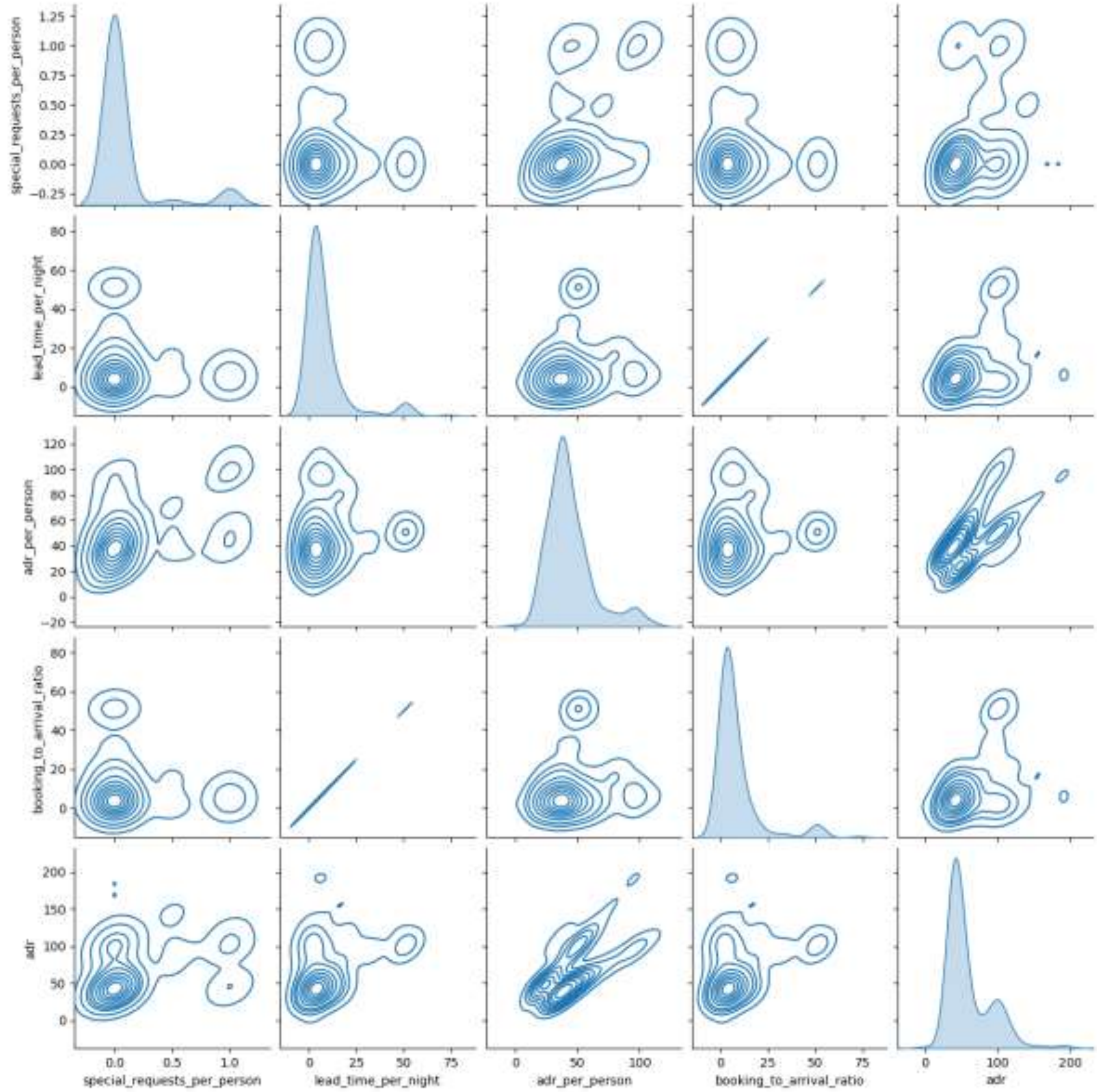


Fig10- KDE Multivariate plot showing correlation between features

Observation:

The multivariate KDE plot confirms the correlation plot and the scatter matrix as the variables booking_to_arrival_ratio and lead_time_per night are highly correlated with the positive correlation. Then, adr and adr_per_person are correlation with a high positive correlation.

10.Data Visualization

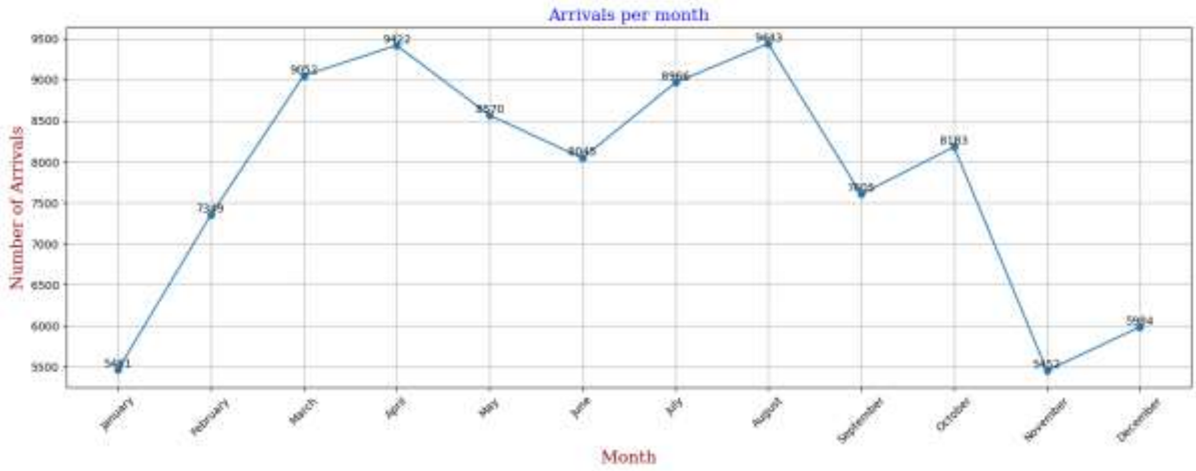


Fig11- Line plot of Arrivals per months

From the line plot, we can see that the Number of arrivals are high in August with 9443 arrivals. Followed by August, April has more arrivals. The reason might be because the people go to vacation in April and return back in August. The least number of arrivals are in November with only 5452 arrivals.

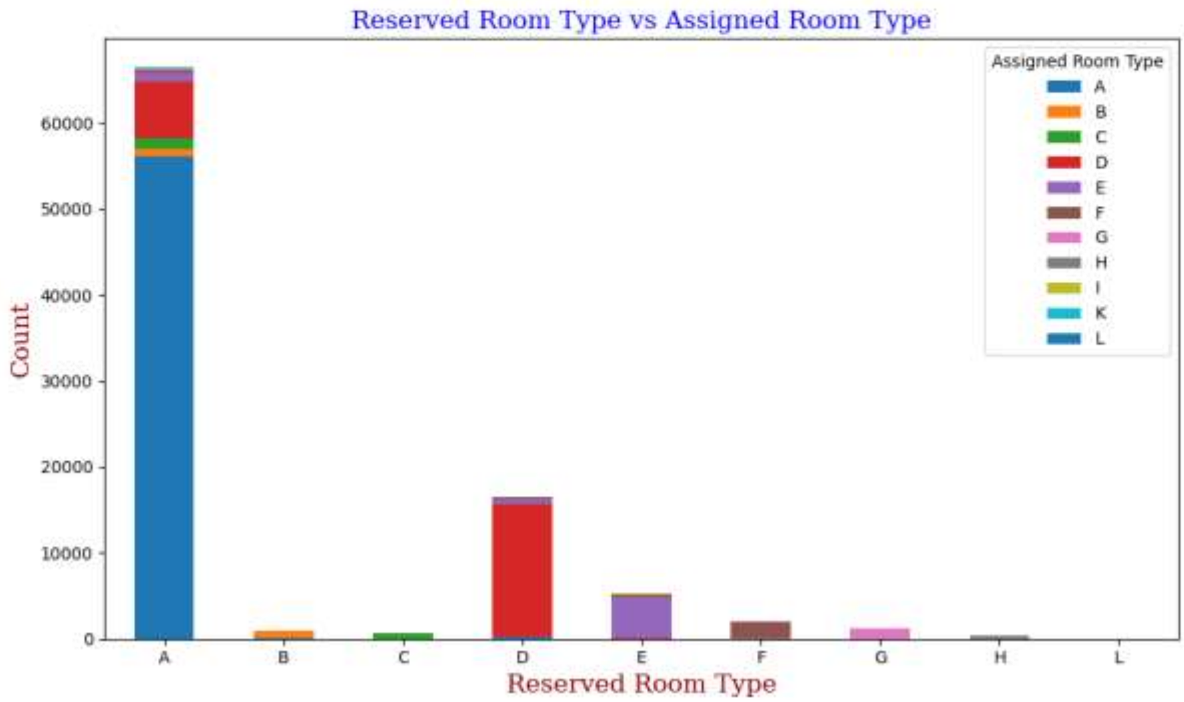


Fig12- Stacked Bar plot for Reserved Room Type vs Assigned Room Type

The graph shows that the room Type of A was reserved by the people but not assigned to some people. D was assigned to the people mostly when A was not available. This shows that room type A has more demand than any other room types. Rest of the room types were also assigned with other rooms sometimes.

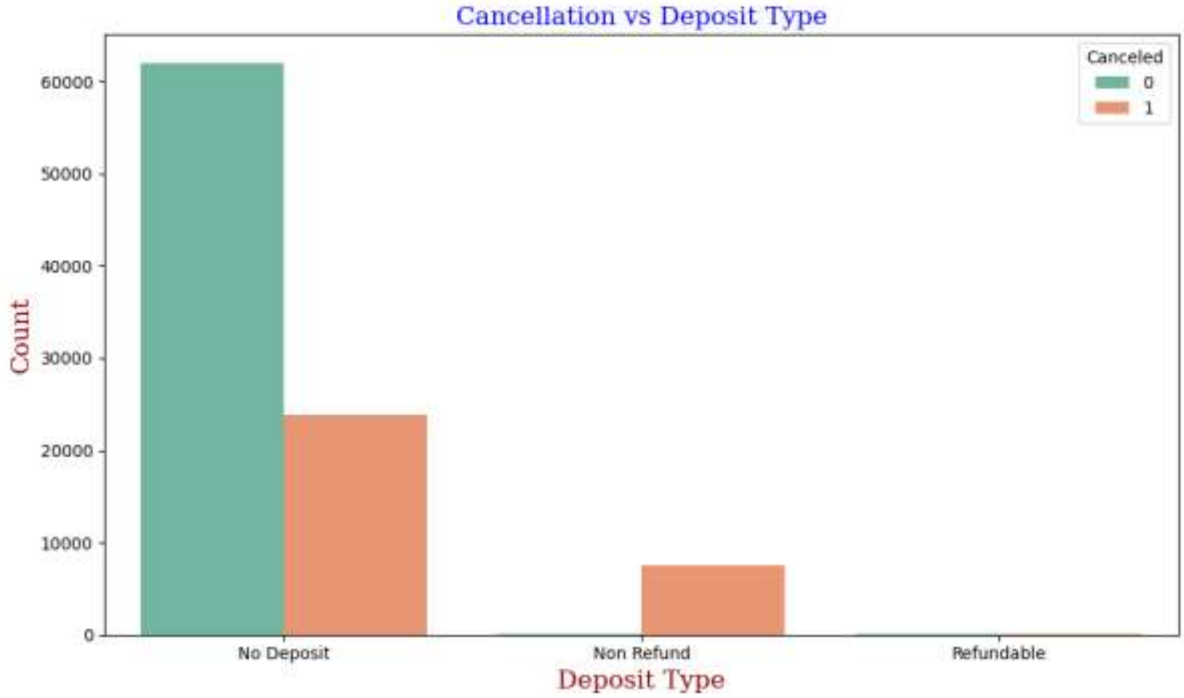


Fig13- Grouped bar plot cancellation vs Deposit Type

The grouped bar plot shows that the people with No Deposit are more than other categories. Many people cancelled as it was no Deposit. But there might be something wrong with the data as most people who booked the hotel with non_refundable deposit cancelled which doesn't make sense.

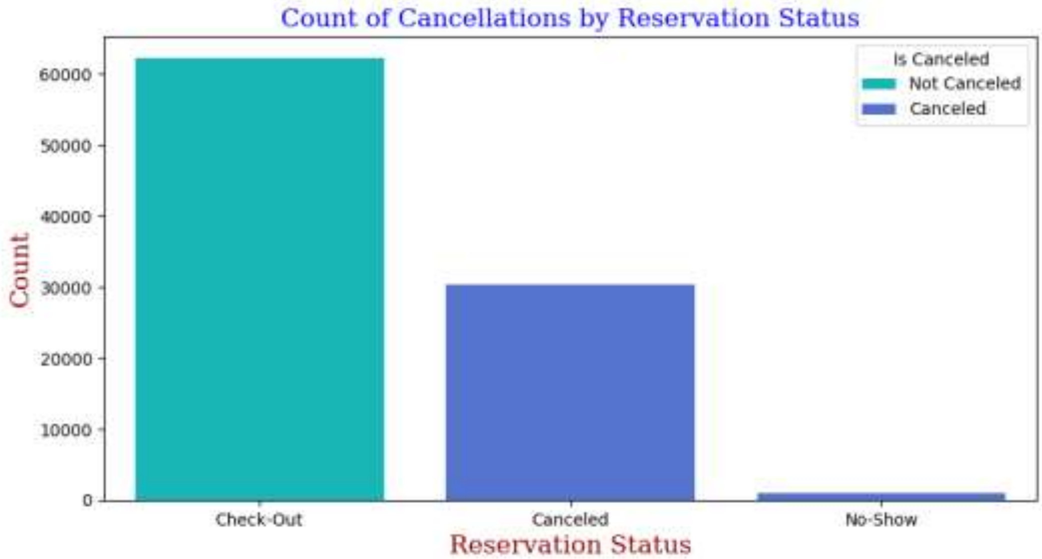


Fig14- Count plot of cancellations by Reservation Status

From the count plot we can see that people who Checked-out are more double than cancelled people. Finally, few people didn't show up and they even cancelled the ticket.

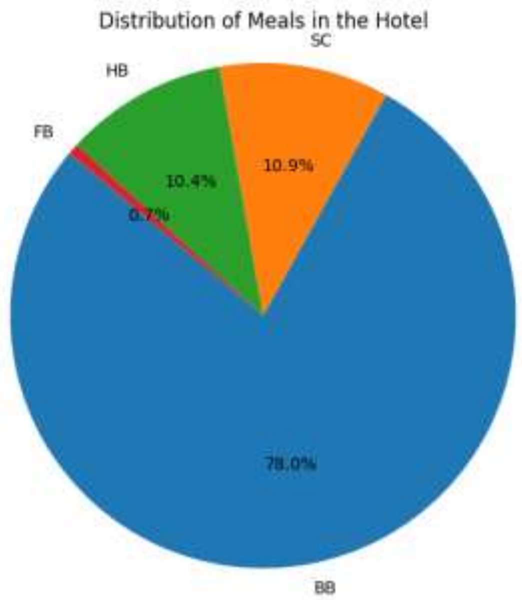


Fig15- Pie chart showing Distribution of Meals in the Hotel

From the pie chat we can see that the meal type of BB has most distribution than other meal types with 78% of the total distribution. Conversely, the lowest meal type being FB with only 0.7% of the total distribution.

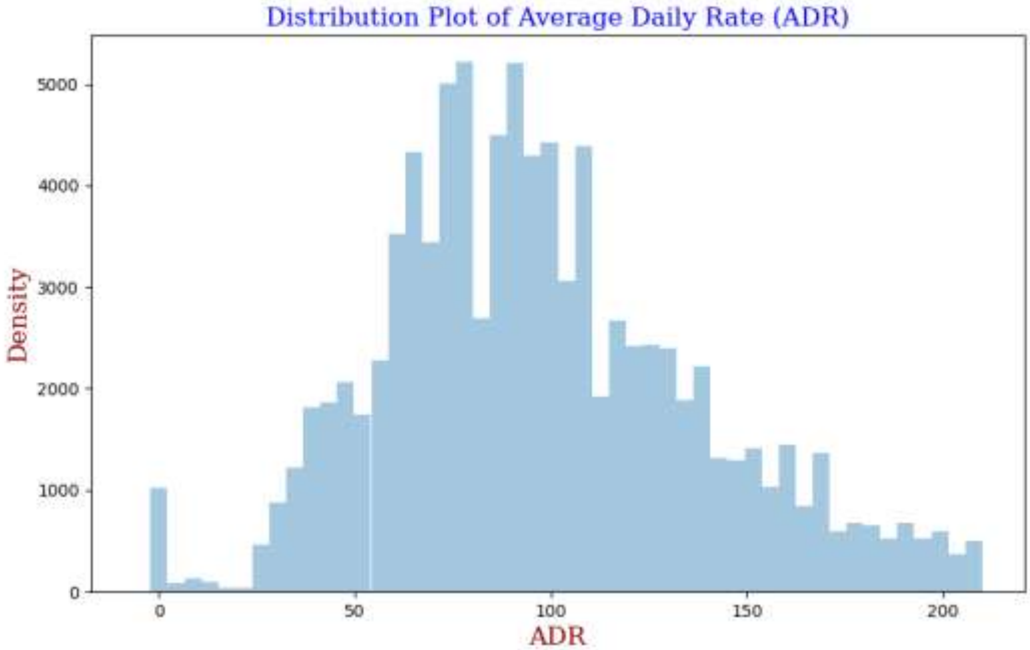


Fig16- KDE plot of Average Daily Rate(ADR)

From the distribution plot of Average Daily Rade(ADR), we can see that the distribution is close to normal but not normal. It is a bit left skewed.

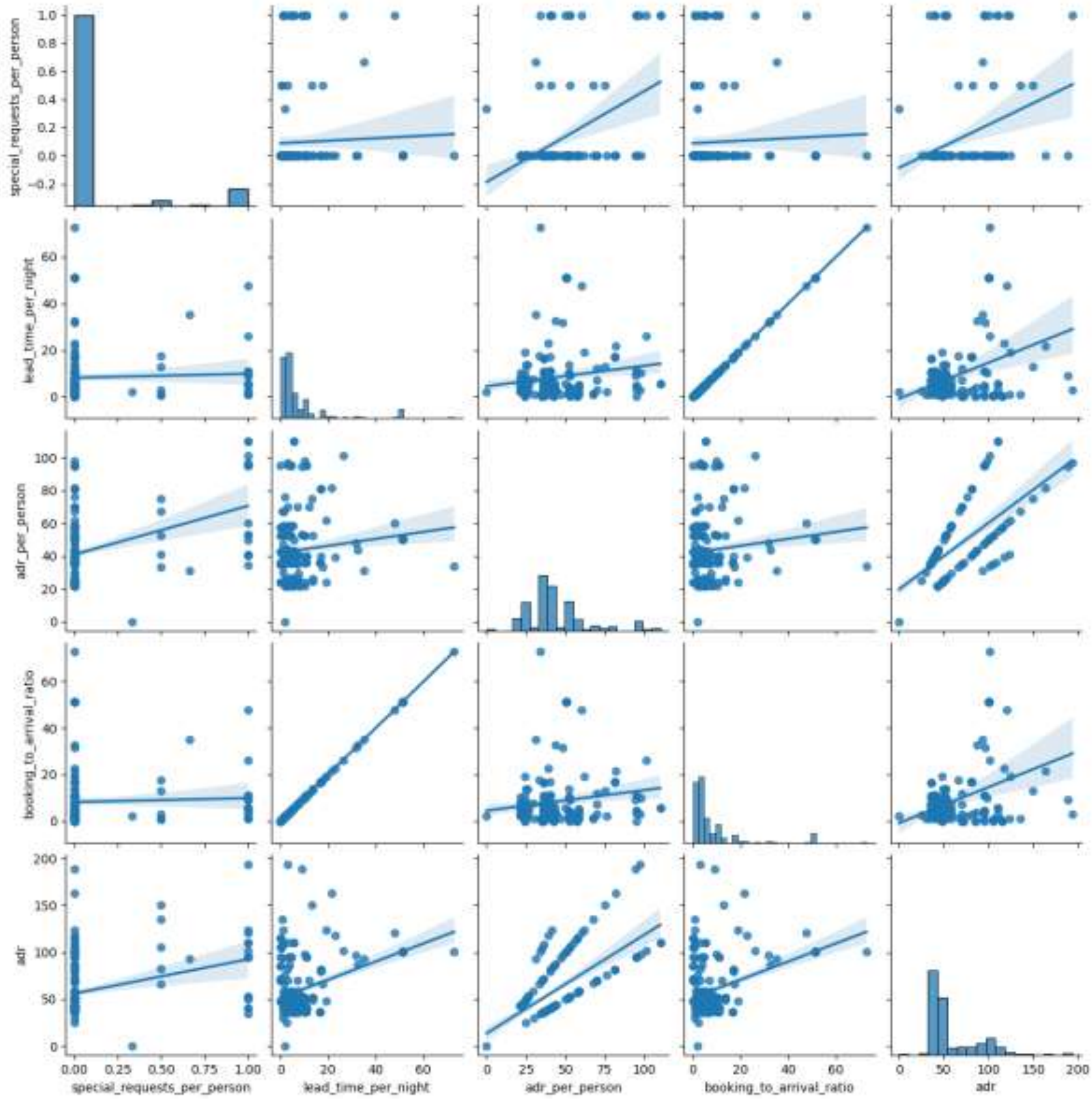


Fig17- Scatter plot with regression line showing correlation between features.

The scatter plot confirms the correlation plot as the variables booking_to_arrival_ratio and lead_time_per night are highly correlated with the positive correlation. Then, adr and adr_per_person are correlation with a high positive correlation.

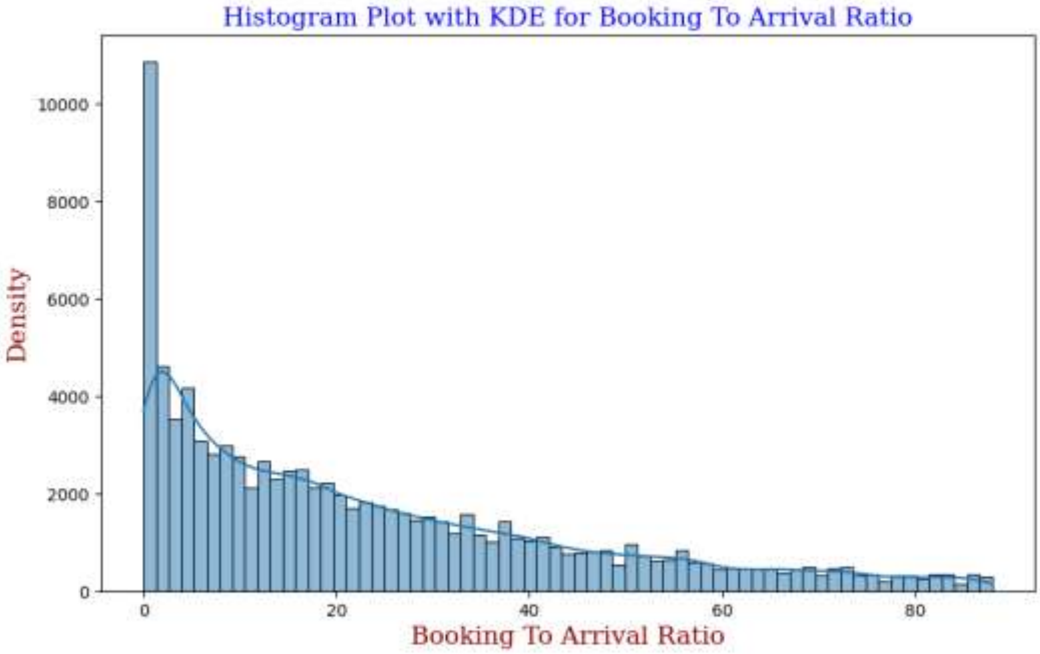


Fig18-Histogram plot with KDE for Booking To Arrival Ration.

From the histogram plow with KDE, we can see that the density of 0 Booking_to_Arrival ratio is very high and it goes down with the increase in value. This it the perfect example of a right skewed distribution.

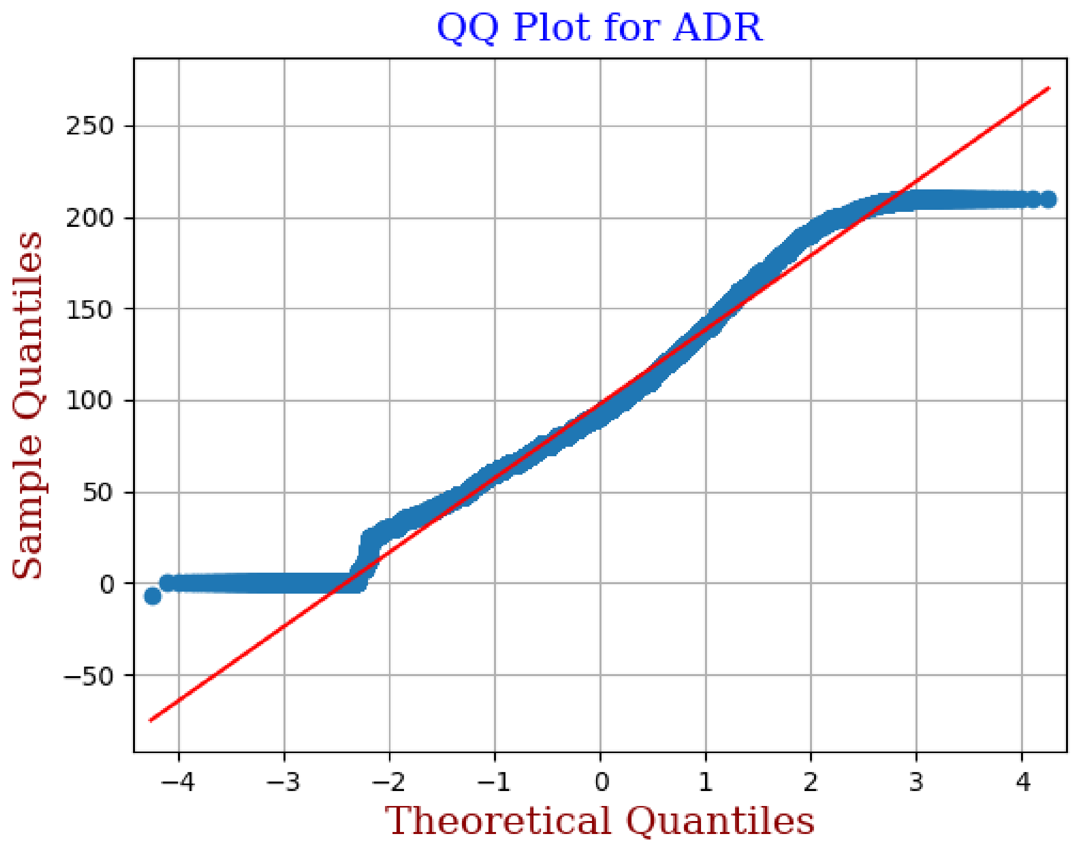


Fig19-QQ plot of ADR.

The qq plot for ADR shows that the ADR is near to normal distribution but it deviating from the normality at the extremes.

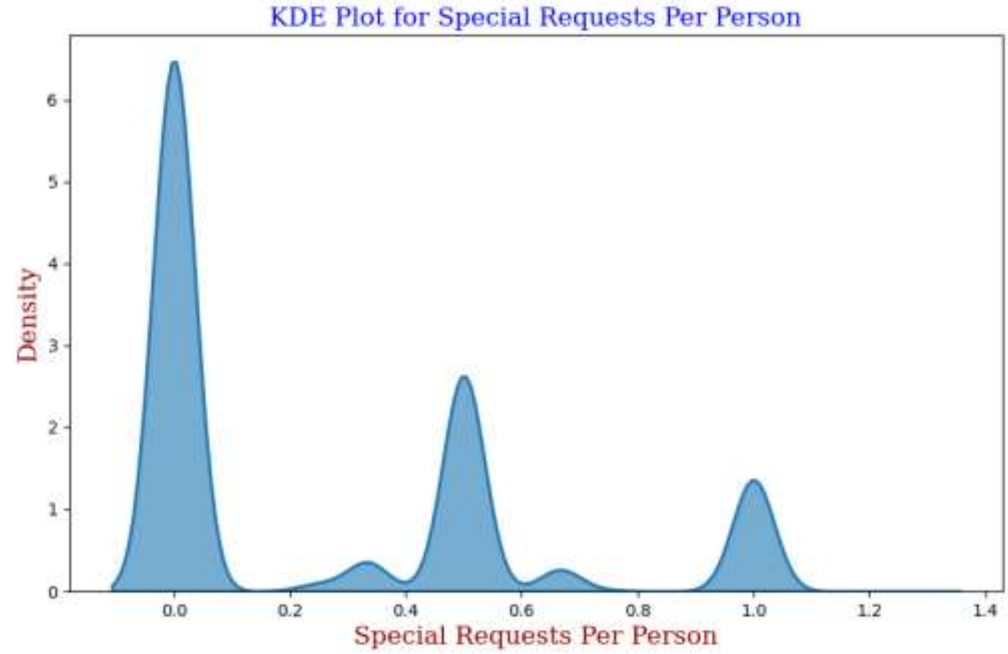


Fig20- KDE plot for Special Requests per Person.

The KDE plot for Special Requests per Person shows highest density at 0 and lowest density at 1. So, we can say that this plot is rightly skewed.

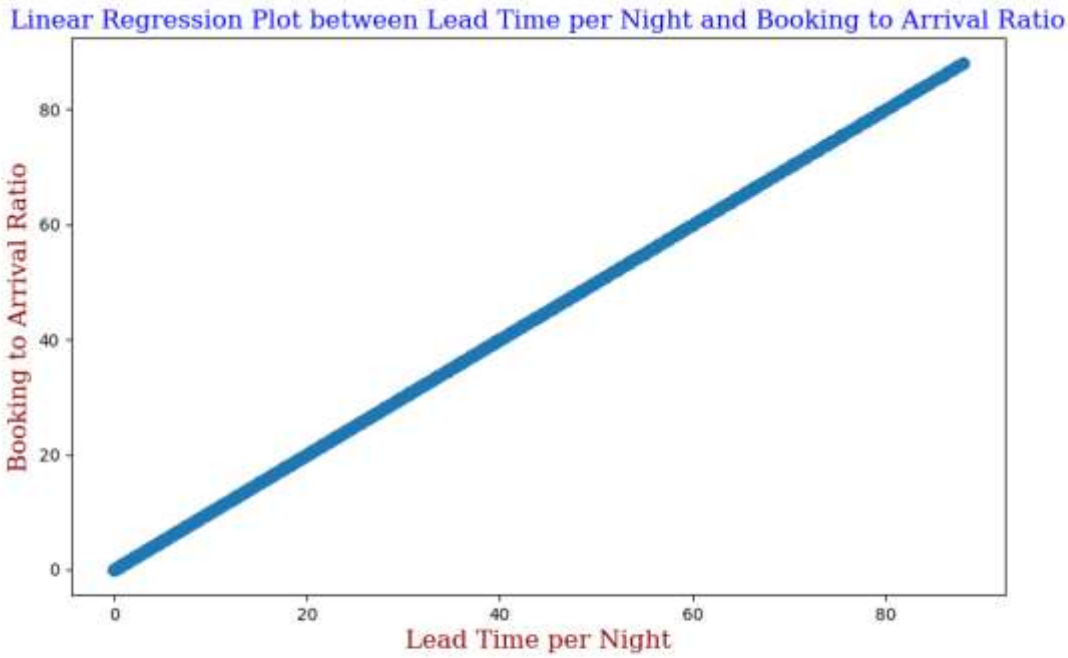


Fig21-Linear Regression Plot between Lead Time per Night and Booking to Arrival Ratio.

The Booking to Arrival Ratio and Lead Time per Night has the highest correlation as we observed from the previous graphs. So, there is a perfect linear line showing positive correlation.

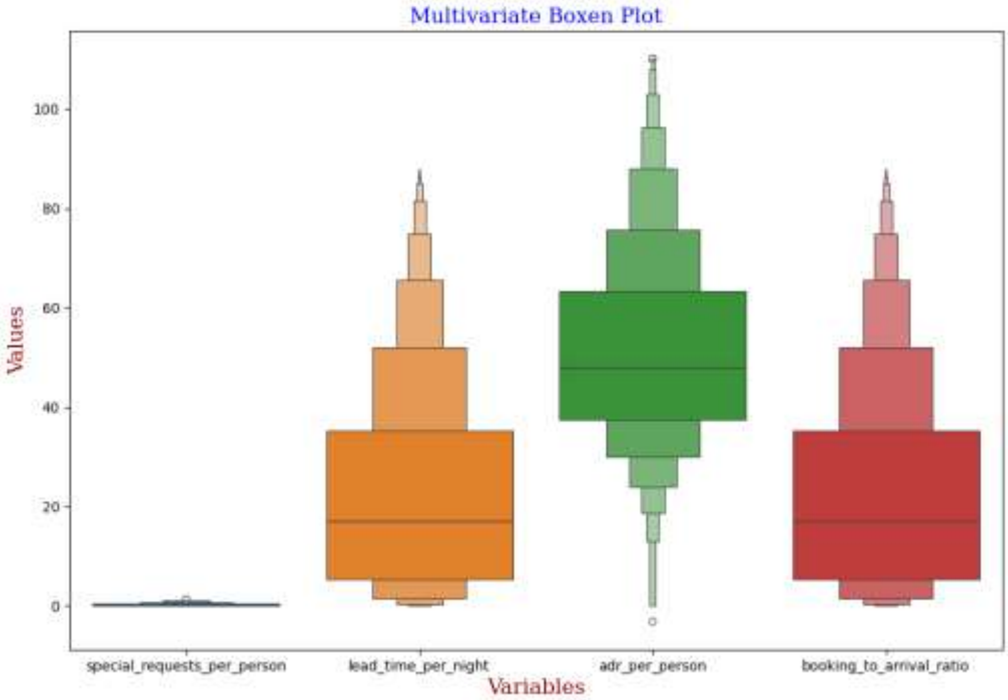


Fig21-Multivariate Boxen plot between features.

From the multivariate boxen plot, we can see that adr_per_person has the highest values compared to other variables. Conversely, special_requests_per_person has the lowest values compared to other variables.

Fig22-Area plot of Total people over Months.

From the area plot, we can see that total people over each month is highest in August because of the vacation. Followed by August, April is the second highest. Conversely, November has the lowest number of people and December and January has lower number of people followed by November.

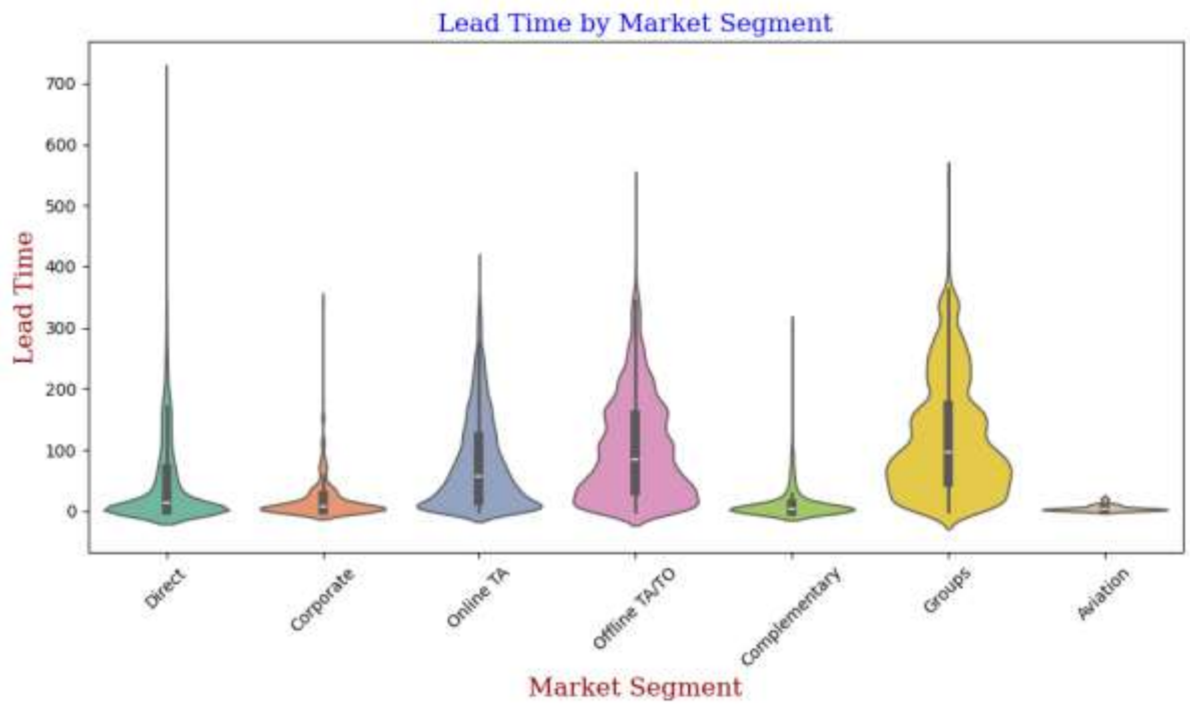


Fig23-Violin plot of lead time by Market segment.

From the violin plot, the lead time distribution for Groups market segement is more than other market segments. The lowest distribution of lead time is observed in Aviation.

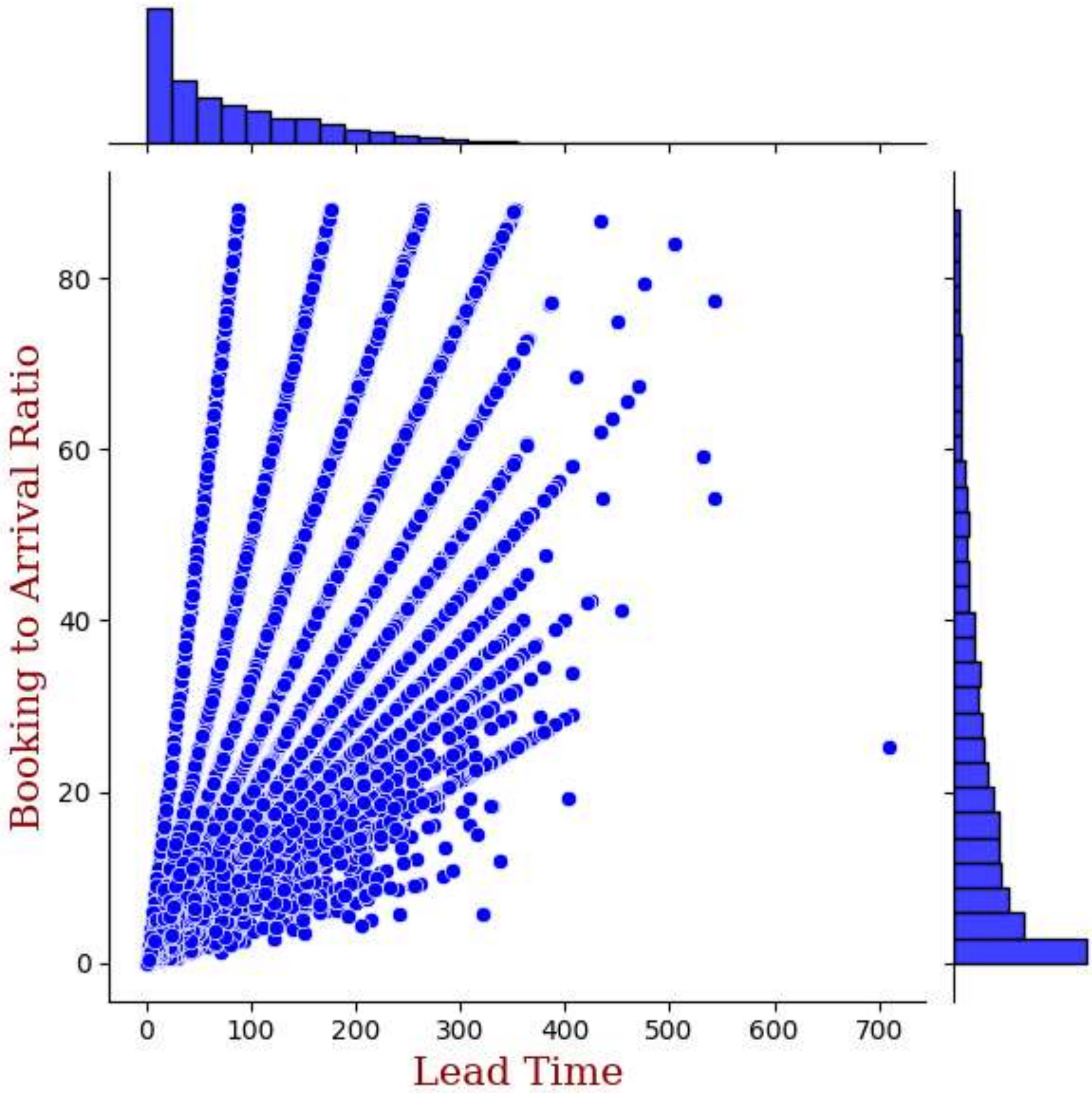


Fig24-Joint Scatter plot between Lead Time and Booking to Arrival Ratio.

The joint plot between the Booking to Arrival Ratio and Lead Time shows positive correlation by the scatter plot. The histograms shows the distrutions for lead time and booking to arrival ratio in the same graph

Fig25-Rug Plot for Booking to Arrival Ratio and ADR.

The rug plot shows the distribution of adr and booking arrival ratio in the edges of the plot. And the values are shown by the scatter.

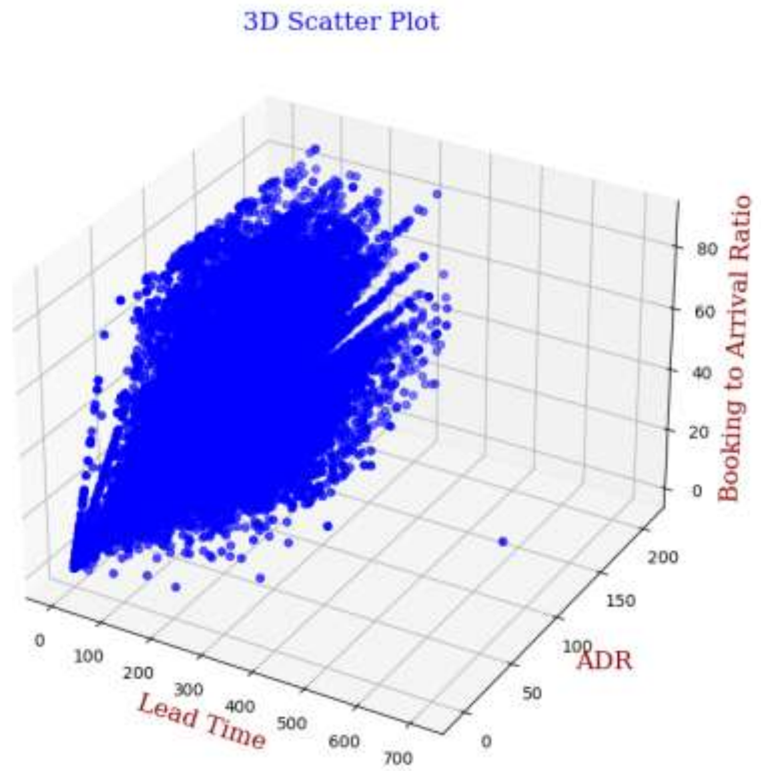


Fig26-3D scatter plot between lead Time, ADR and Booking to Arrival Ratio.

The 3d scatter plot shows the combined reperesentation of all the three variables- ADR, Booking to Arrival Ratio and Lead Time in a 3 dimensional projection.

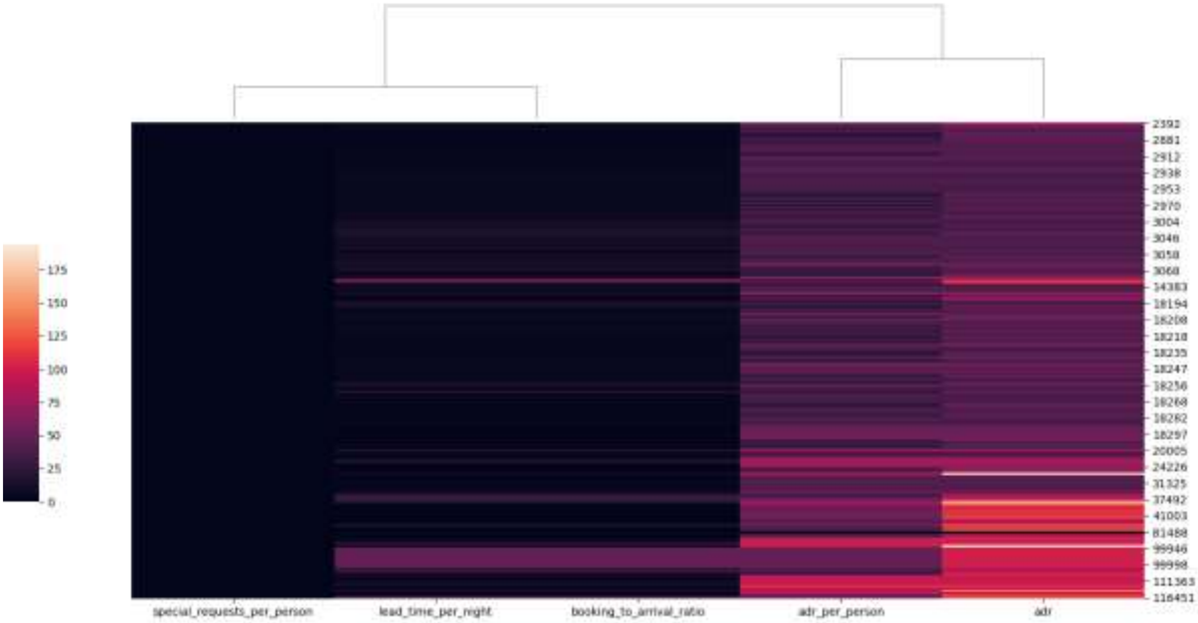


Fig27-Cluster Map between numerical features.

The cluster map shows that the distribution of values of special requests per person are very low as the color is darker and the distribution of values adr is higher as the color is lighter. This map forms as cluster with the given variables.

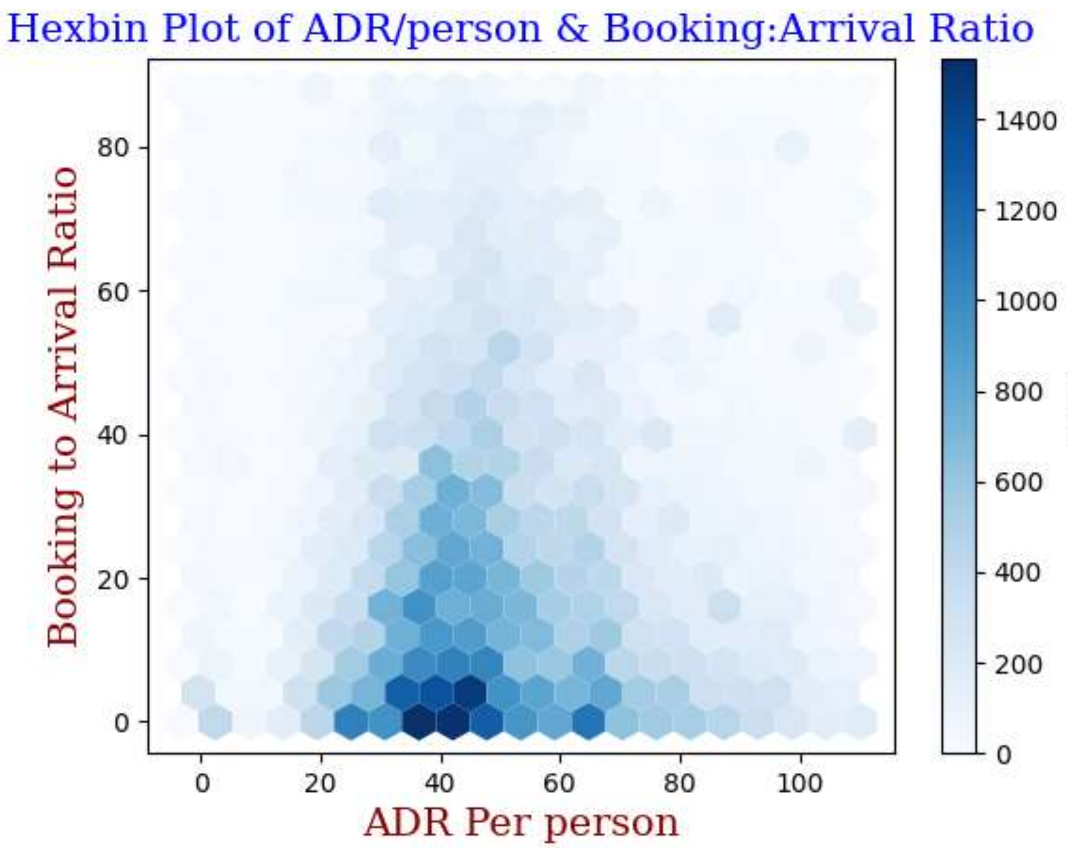


Fig28-Hexbin plot of ADR per person and Booking to Arrival Ratio.

The Hexbin plot shows that the count of Booking to Arrival Ratio value is higher around 0 and count of ADR per person values are higher at around 40.

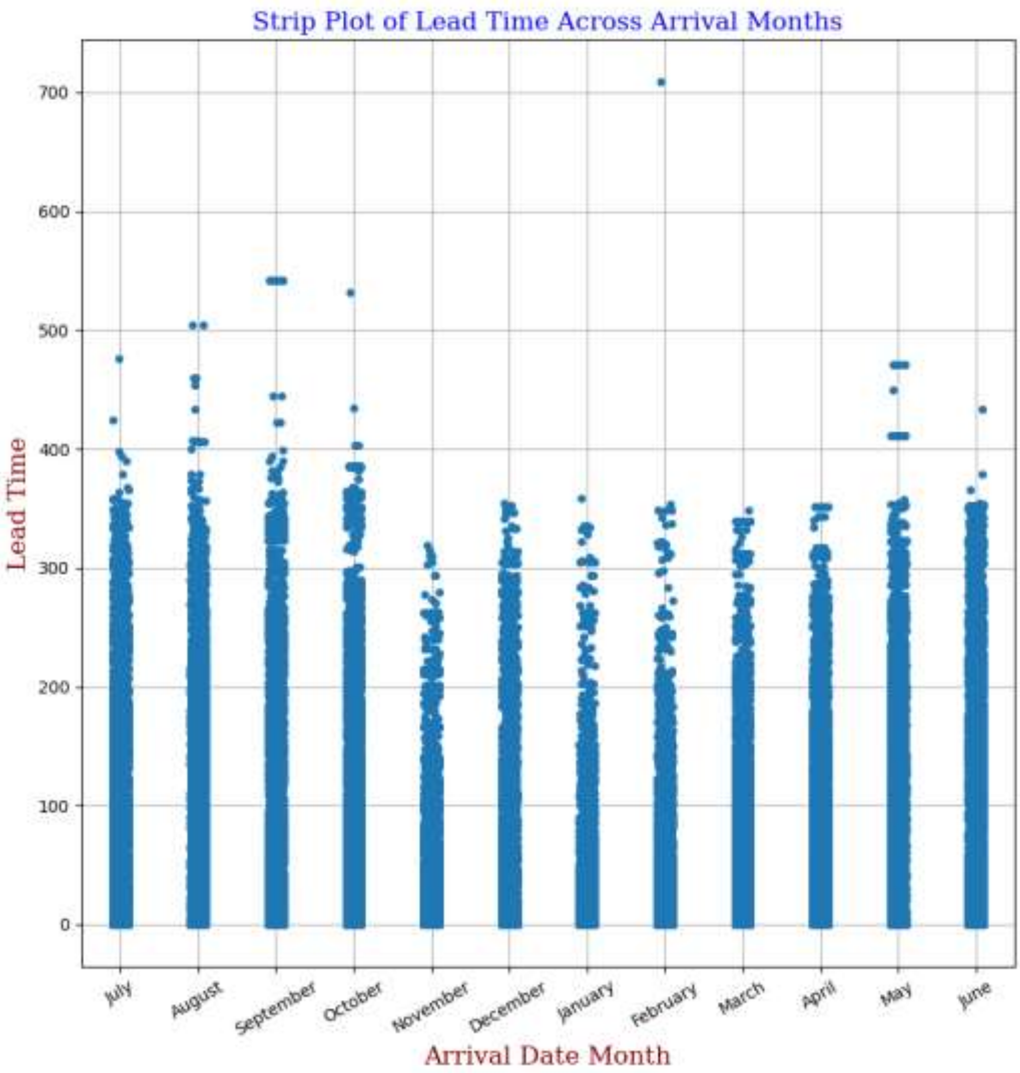
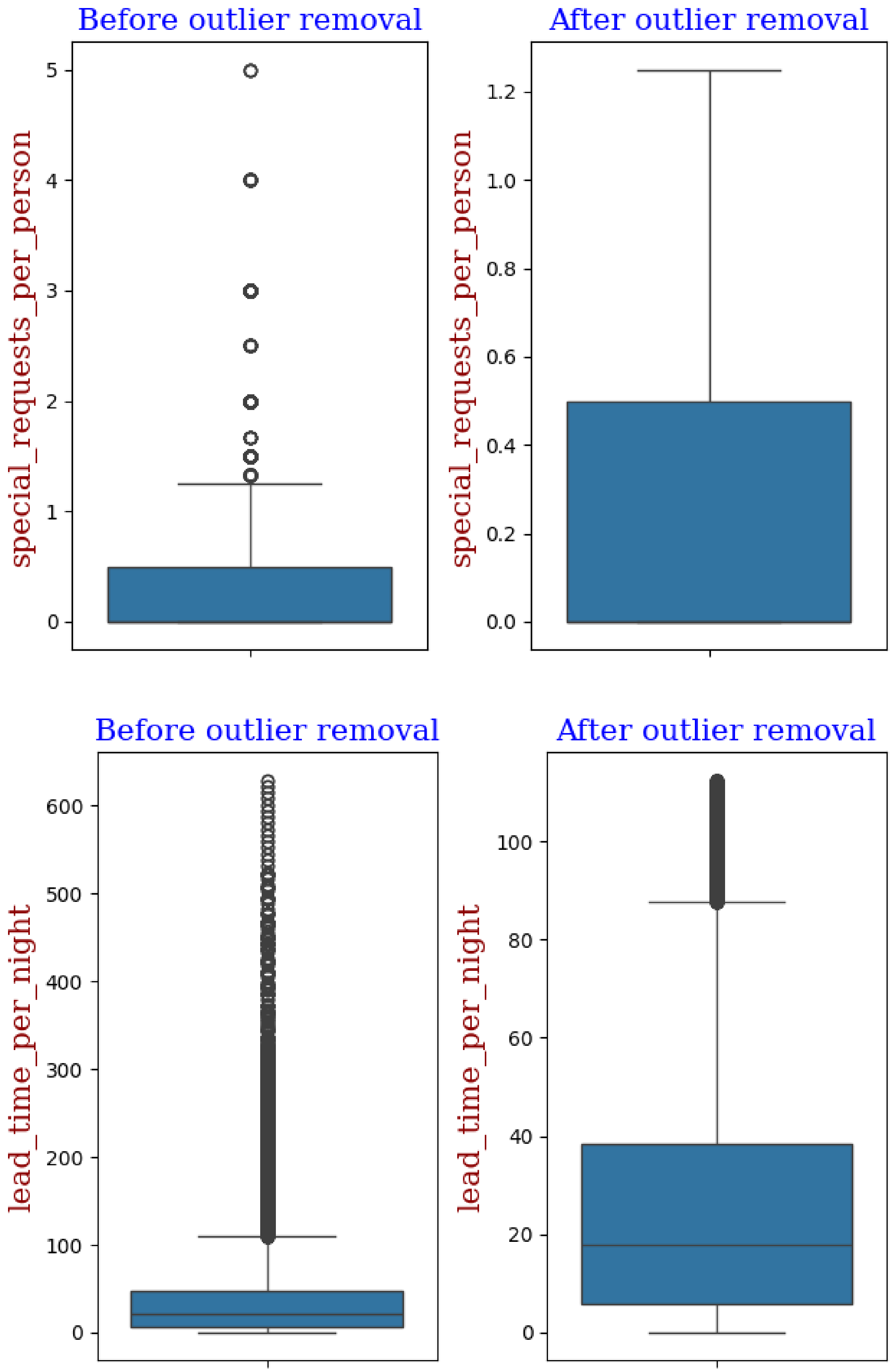


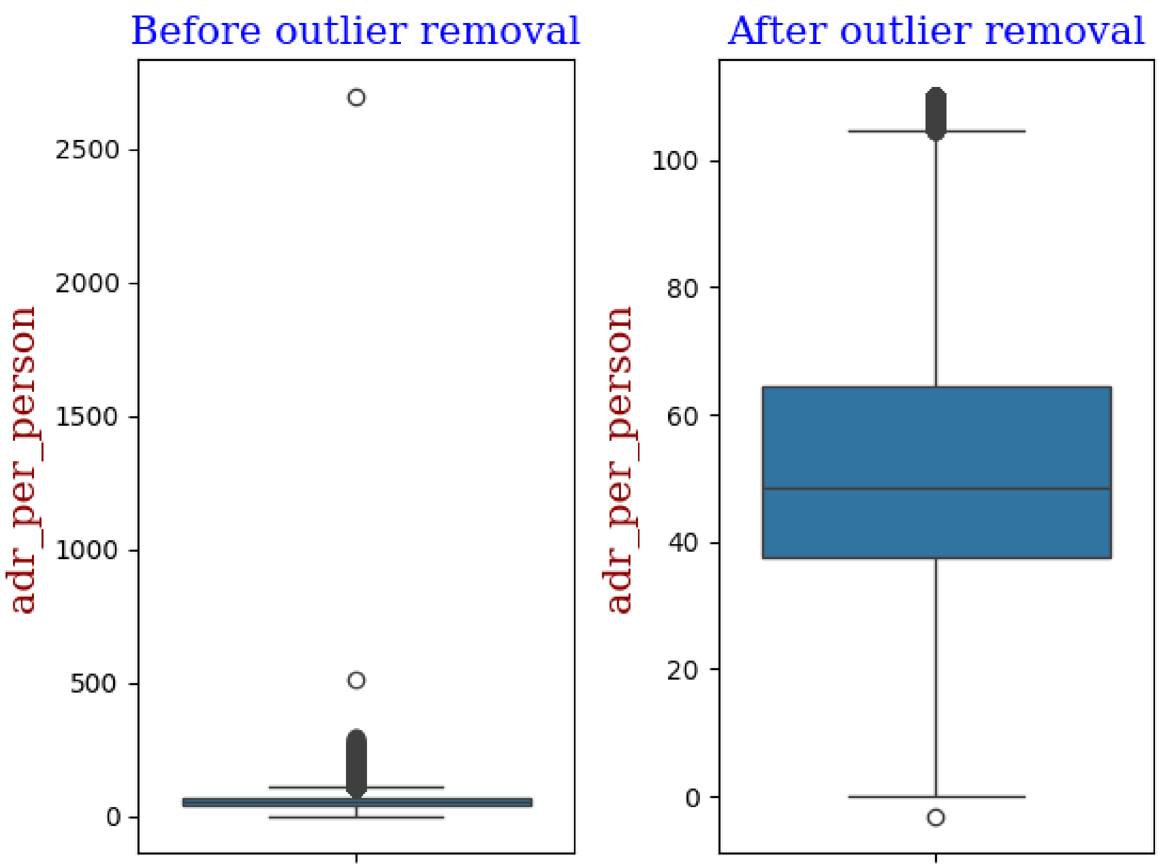
Fig29-Strip plot between lead time and arrival date month.

The strip plot shows the distribution of lead time across months. The lead times are high from June to October. The lead times are low but not much low for other months.

11.Subplots



Before outlier removal
After outlier removal
special_requests_per_person
5
4
3
2
1
0
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Before outlier removal
After outlier removal
lead_time_per_night
600
500
400
300
200
100
0
100
80
60
40
20
0



12.Tables

Column	Da_k_squared statistics	Da_k_squared p-value	K_S statistics	K_S p-valueShapiro statistics	Shapiro p-value	Normality
special_requests_per_person	11370.63	0.00	0.36	0.000.72	0.00	Not Normal
lead_time_per_night	11685.17	0.00	0.14	0.000.89	0.00	Not Normal
adr_per_person	3643.08	0.00	0.07	0.000.98	0.00	Not Normal
booking_to_arrival_ratio	11685.17	0.00	0.14	0.000.89	0.00	Not Normal
adr	2753.65	0.00	0.07	0.000.98	0.00	Not Normal

13.Dashboard

https://dashapp-sk2ykdtmqq-ue.a.run.app/

I designed and implemented a data visualization dashboard using Dash, a Python web framework for building analytical web applications. Here's what I did:

1. **Outlier Detection and Removal:**
    - I created a tab to visualize outliers in numerical columns using box plots and histograms.
    - Applied the Interquartile Range (IQR) method to detect and remove outliers.
    - Implemented callbacks to update the plots dynamically based on user selections.
2. **Normality Tests and Plots:**

- Developed a tab to perform normality tests on numerical columns and visualize the distribution using Q-Q plots and histograms.
- Utilized statistical tests such as D'Agostino & Pearson Test, Kolmogorov-Smirnov Test, and Shapiro-Wilk Test.
- Incorporated interactive elements to choose the test type and plot type, updating the results in real-time.

3. **PCA Analysis:**
    - Created a tab to perform Principal Component Analysis (PCA) on numerical features.
    - Implemented a slider to select the desired explained variance threshold, updating the PCA graph and information dynamically.
    - Calculated the condition number of the reduced feature space and displayed it as part of the analysis.
4. **Analysis Between Variables Using Plots:**
    - Designed a tab to explore the relationship between continuous and categorical variables using various plot types such as box plots, violin plots, and scatter plots.
    - Included an option to include hue as "hotel" for additional insights.
    - Implemented callbacks to update the plots based on user selections.
5. **Choropleth Map:**
    - Developed a tab to visualize the distribution of guests' home countries using a choropleth map.
    - Provided a range slider to adjust the color bar range dynamically.
6. **Other Plots:**
    - Implemented tabs for additional analyses, including animated bar plots of continuous variables over months and pie plots of categorical variables against revenue metrics.
    - Utilized dropdowns and checkboxes for user interaction and dynamic updates of the plots.
7. **Analysis by Selecting Specific Country:**
    - Designed a tab to analyze various variables by selecting a specific country.
    - Implemented dropdowns for country selection and displayed bar plots of selected variables for the chosen country.
    - Added a boolean switch to toggle between cancellation percentage and arrivals per month by hotel type.

14.Conclusion

The analysis of the dataset provided valuable insights into various aspects of hotel bookings, guest behavior, and operational dynamics. Here are the key conclusions drawn from the observations:

1. **Correlation Analysis:**
    - Variables such as booking_to_arrival_ratio and lead_time_per night exhibit a strong positive correlation, which is expected due to the calculation method.
    - Similarly, adr and adr_per_person are highly correlated, indicating a relationship between average daily rate and per person rate.
2. **Booking Patterns:**
    - The line plot reveals that August and April experience the highest number of arrivals, potentially due to vacation seasons.
    - Conversely, November sees the lowest number of arrivals, indicating off-peak periods for hotel bookings.
3. **Room Type Demand:**
    - Room type A shows the highest demand, as evidenced by reservations but not always being assigned.
    - This suggests a need for hotels to optimize inventory and allocation strategies for popular room types.
4. **Deposit Type and Cancellations:**
    - Contrary to expectations, a significant number of cancellations occur with bookings made without a deposit.
    - This discrepancy warrants further investigation into possible data anomalies or underlying reasons for cancellations.
5. **Reservation Status:**
    - The count plot highlights that the majority of guests check-out, indicating successful stays.
    - A smaller proportion of guests cancel, with even fewer being no-shows.
6. **Meal Preferences:**
    - BB meal type dominates the distribution, indicating a preference for bed and breakfast options among guests.
    - Conversely, FB meal type has the lowest distribution, suggesting less popularity or availability of full board options.
7. **Average Daily Rate Distribution:**
    - The distribution plot for ADR shows slight left skewness, indicating a tendency towards lower average daily rates.

- Further analysis may be needed to understand pricing strategies and their impact on booking behavior.

8. **Special Requests:**
    - The KDE plot for special requests per person suggests a majority of guests make no special requests.
    - This information can guide hotel staff in anticipating and fulfilling guest needs more effectively.
9. **Monthly Trends:**
    - August emerges as the busiest month, likely due to peak vacation periods.
    - April follows closely, indicating another peak period for hotel bookings.
    - November experiences the lowest number of guests, suggesting an off-peak season.
10. **Market Segment Analysis:**
    - Groups market segment shows a higher lead time distribution compared to other segments, indicating potentially longer planning periods for group bookings.
    - Conversely, Aviation segment exhibits the lowest lead time distribution.

15.References

- Data Visualization Society (datavisualizationsociety.com) - Provides resources, community forums, and events related to data visualization.
- Towards Data Science (towardsdatascience.com) - Offers articles, tutorials, and case studies on various data visualization techniques and tools.
- Tableau Public Gallery (public.tableau.com/en-us/gallery) - Offers a collection of interactive data visualizations created by the Tableau community, providing inspiration and examples for your own projects.