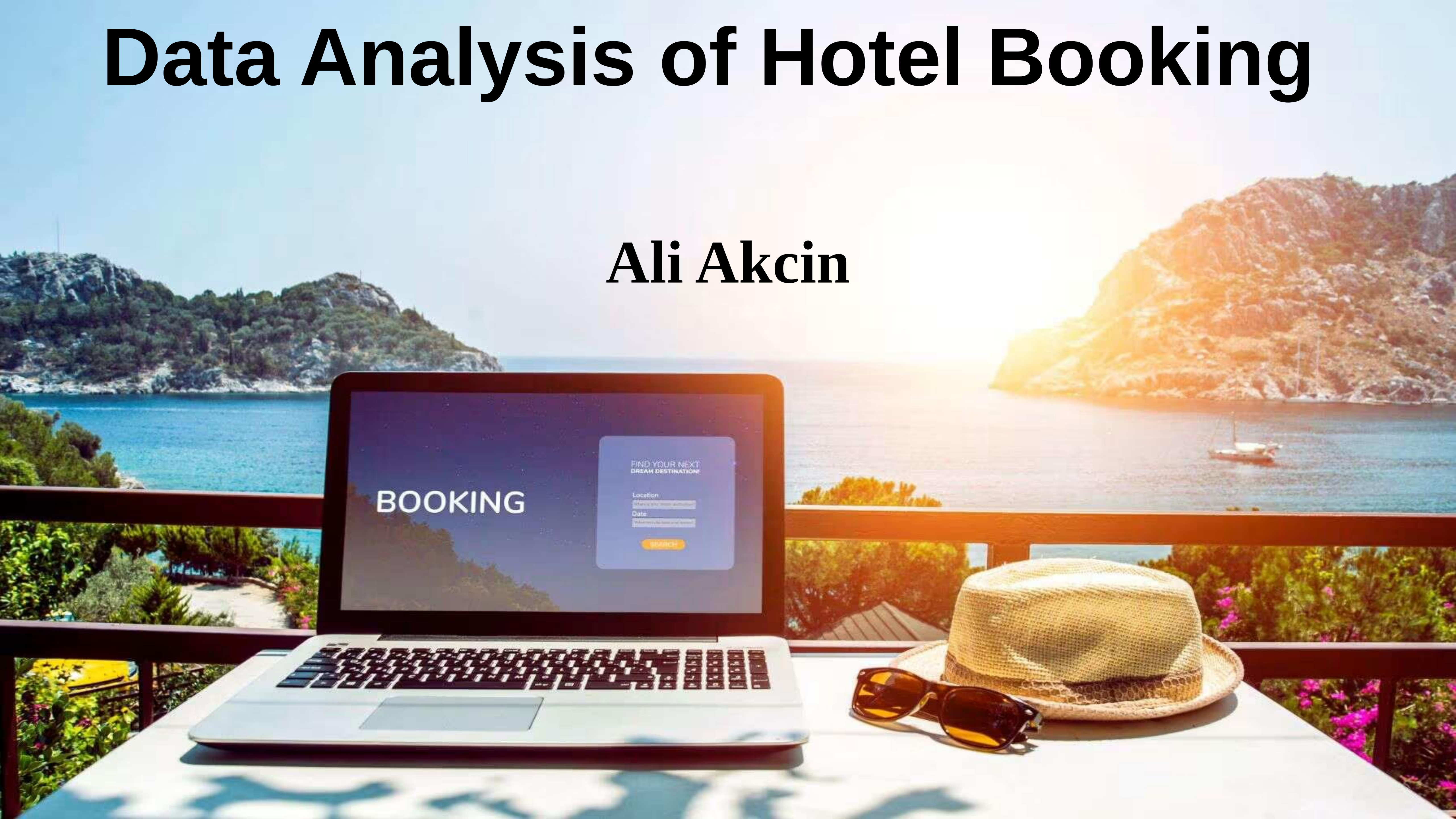
Data Analysis of Hotel Booking

Ali Akcin

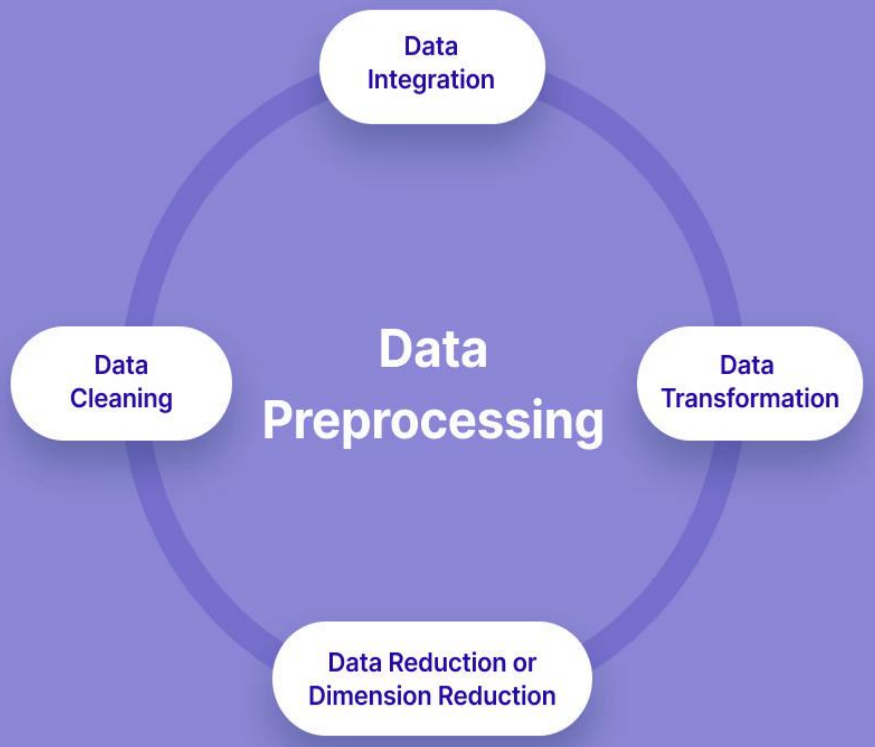


Problems Addressed

- Demand Forecasting
- Customer Segmentation
- Cancellation Analysis
- Optimal Room Allocation

Pre-Processing Data

- **Data**: Downloaded hotel dataset (119k records, 33 attributes).
- **Database**: Imported into PostgreSQL.
- **Cleaning**: Handled missing data, standardized values.
- **Transformation**: Normalized, filtered, and removed outliers.
- **Validation**: Verified with Weka.
- **Result**: Clean, ready dataset.



Data Preprocessing

Data Integration

Data Transformation

Data Reduction or Dimension Reduction

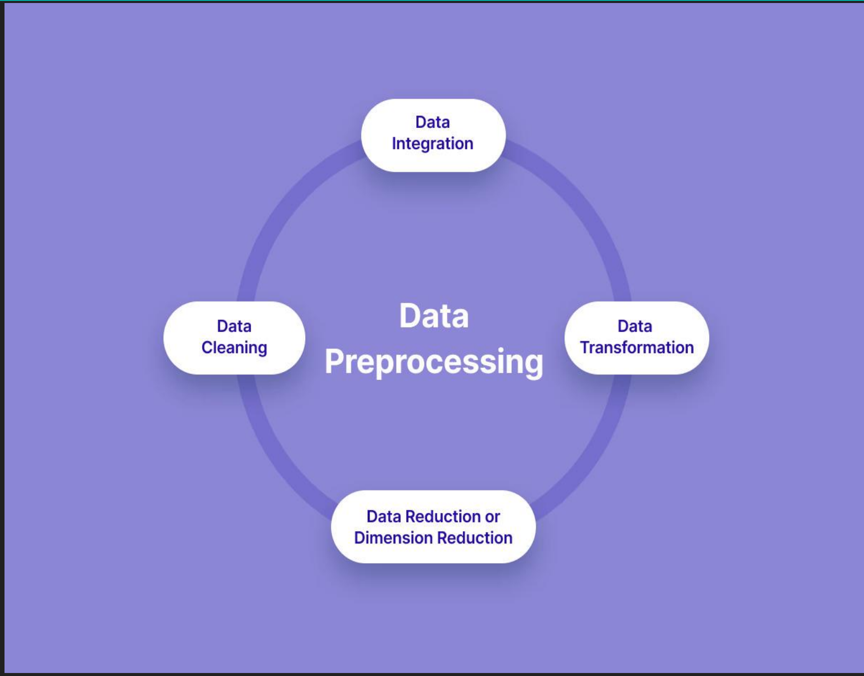
Data Cleaning

Pre-Processing Data

Approach:

- **"Company" Column**: Removed due to **112,593 (94%)** missing values, which could lead to biased or unreliable results.

- **"Children" Column**: Imputed missing values (4, 0%) using the **"ReplaceMissingValues"** filter in Weka, as the missing data was minimal, and imputation helped maintain the dataset's distribution.



Justification:

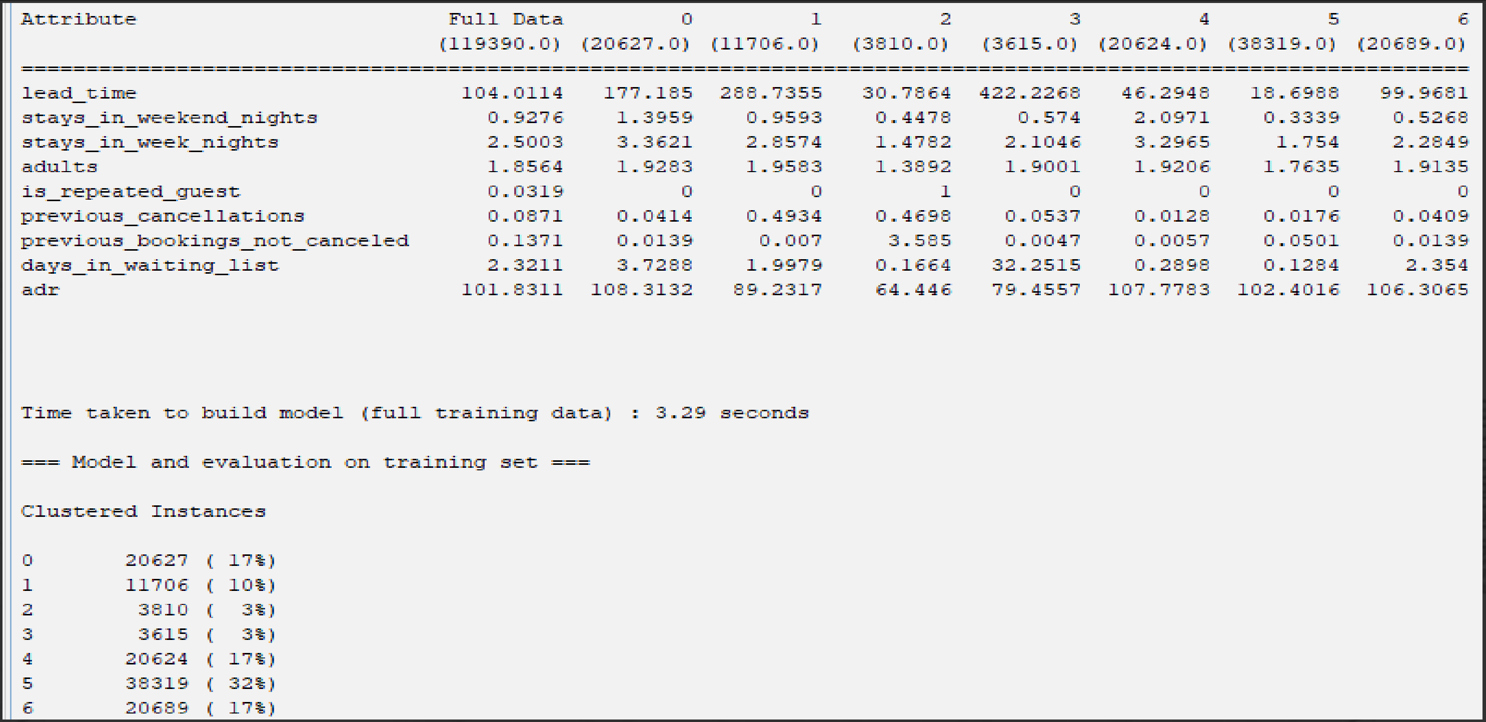
- Removing the **"company"** column eliminated unnecessary noise from the dataset, ensuring a more reliable analysis.

- Imputing the 'children' column with minimal missing values preserved the dataset's integrity, while the 'agent' and 'country' columns were processed similarly by imputing missing values with the mode, as they were categorical features with minimal missing data.

Data Analytics
K-Means Clustering

- Why K-Means Clustering ?

  - ✓ Efficient for large datasets
  - ✓ Scalable to multiple attributes
  - ✓ Clear and interpretable clusters
  - ✓ Simple and flexible algorithm



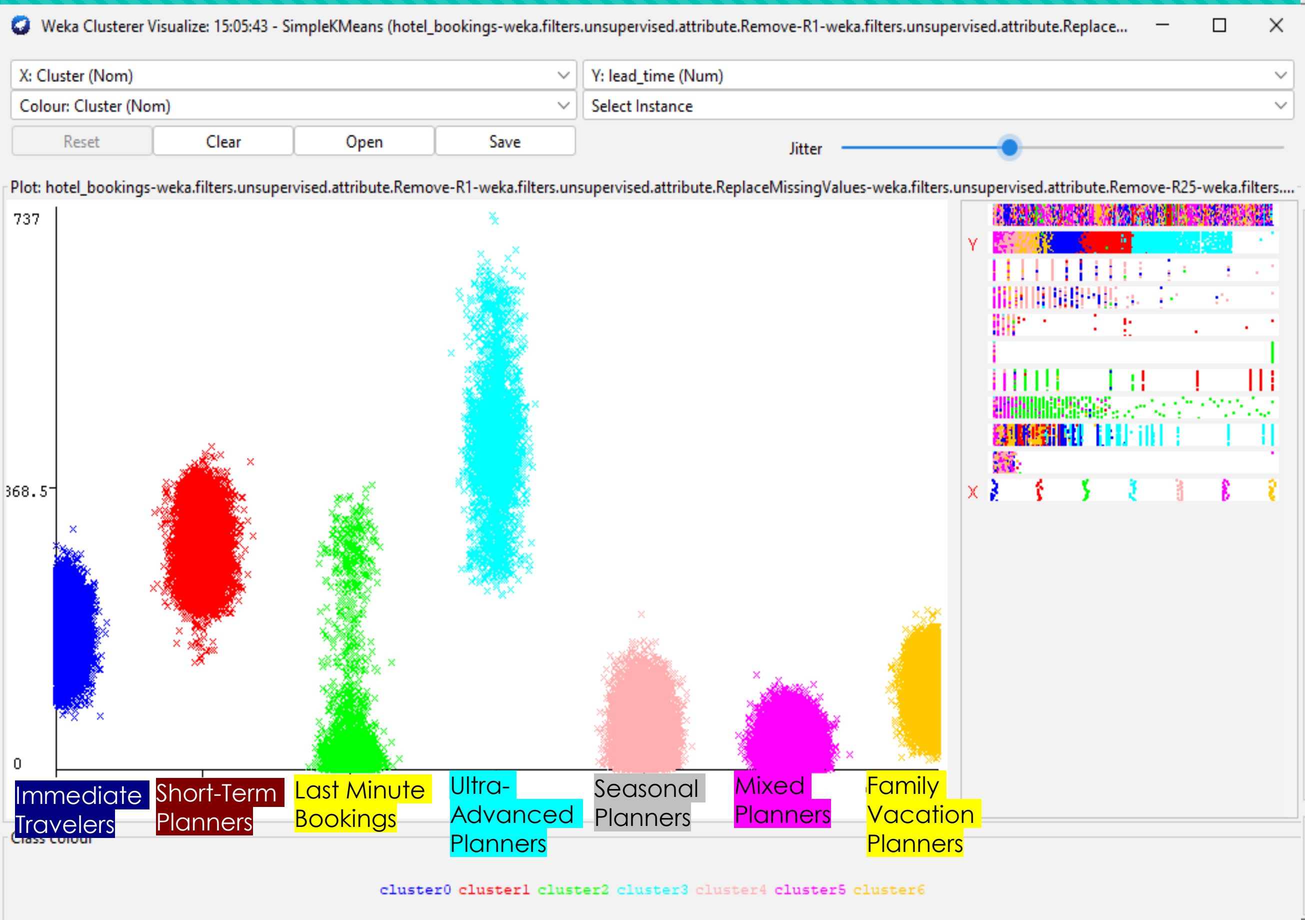
```
Attribute                          Full Data          0          1          2          3          4          5          6
                                   (119390.0)  (20627.0)  (11706.0)   (3810.0)   (3615.0)  (20624.0)  (38319.0)  (20689.0)
==========================================================================================================================
lead_time                            104.0114    177.185   288.7355    30.7864   422.2268    46.2948    18.6988    99.9681
stays_in_weekend_nights                0.9276     1.3959     0.9593     0.4478      0.574     2.0971     0.3339     0.5268
stays_in_week_nights                   2.5003     3.3621     2.8574     1.4782     2.1046     3.2965      1.754     2.2849
adults                                 1.8564     1.9283     1.9583     1.3892     1.9001     1.9206     1.7635     1.9135
is_repeated_guest                      0.0319          0          0          1          0          0          0          0
previous_cancellations                 0.0871     0.0414     0.4934     0.4698     0.0537     0.0128     0.0176     0.0409
previous_bookings_not_canceled         0.1371     0.0139      0.007      3.585     0.0047     0.0057     0.0501     0.0139
days_in_waiting_list                   2.3211     3.7288     1.9979     0.1664    32.2515     0.2898     0.1284      2.354
adr                                  101.8311   108.3132    89.2317     64.446    79.4557   107.7783   102.4016   106.3065




Time taken to build model (full training data) : 3.29 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       20627 ( 17%)
1       11706 ( 10%)
2        3810 (  3%)
3        3615 (  3%)
4       20624 ( 17%)
5       38319 ( 32%)
6       20689 ( 17%)
```

- Key patterns identified in **ADR (Average Daily Rate)**, **booking changes**, and **stay duration**.

- **Actionable insights**: Tailor pricing and promotions for each segment.

K-Means Clustering

Clustering of Hotel Bookings Based on Lead Time



Weka Clusterer Visualize: 15:05:43 - SimpleKMeans (hotel_bookings-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Replace...

X: Cluster (Nom)

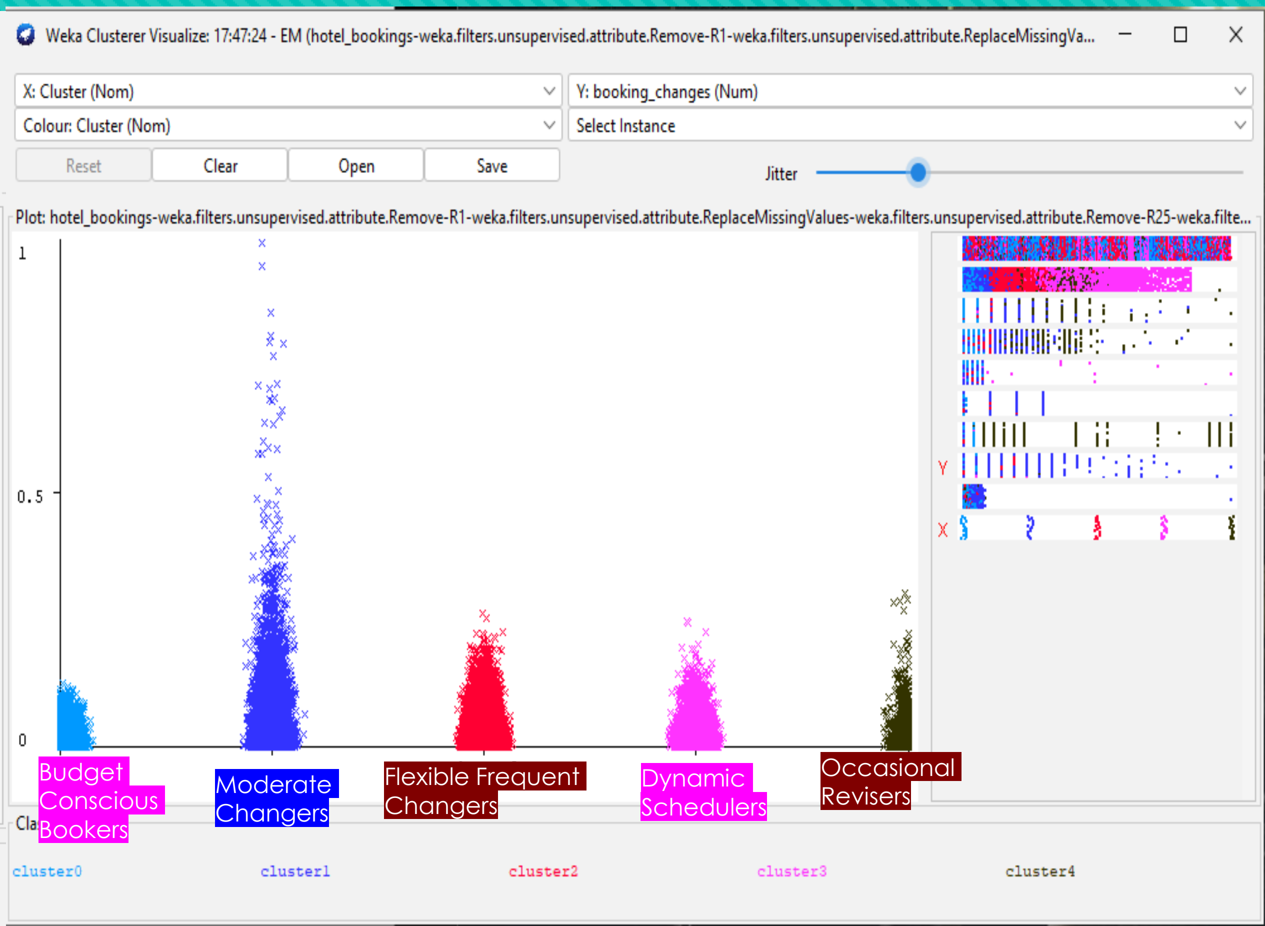
Y: lead_time (Num)

Colour: Cluster (Nom)

Select Instance

Plot: hotel_bookings-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Remove-R25-weka.filters....

Clustering of Hotel Bookings Based on Booking Changes



Weka Clusterer Visualize: 17:47:24 - EM (hotel_bookings-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.ReplaceMissingVa...

X: Cluster (Nom)

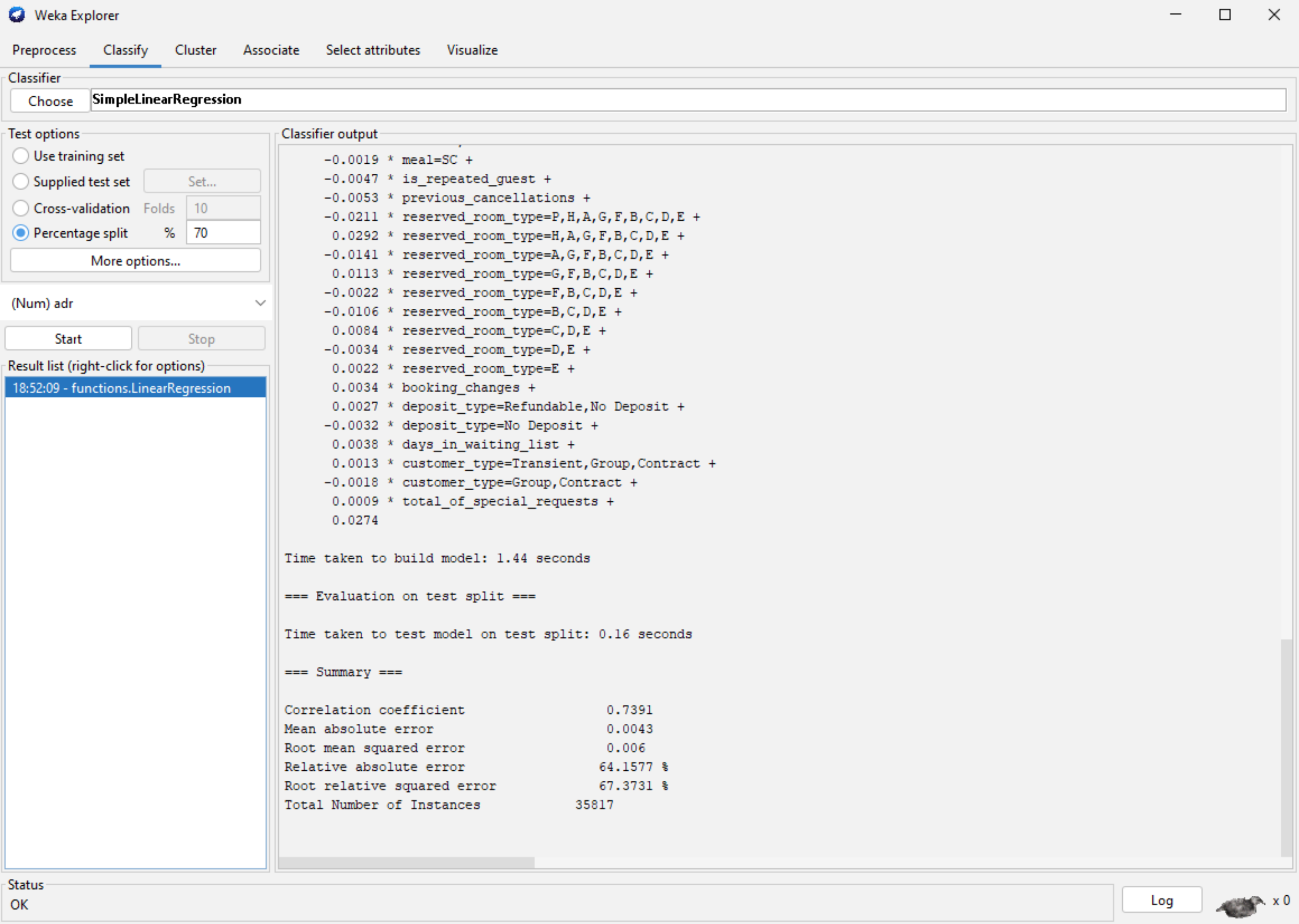
Y: booking_changes (Num)

Colour: Cluster (Nom)

Select Instance

Plot: hotel_bookings-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Remove-R25-weka.filte...

Data Analytics
Linear Regression



SimpleLinearRegression

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 70

(Num) adr

18:52:09 - functions.LinearRegression

Status: OK

```
      -0.0019 * meal=SC +
      -0.0047 * is_repeated_guest +
      -0.0053 * previous_cancellations +
      -0.0211 * reserved_room_type=P,H,A,G,F,B,C,D,E +
       0.0292 * reserved_room_type=H,A,G,F,B,C,D,E +
      -0.0141 * reserved_room_type=A,G,F,B,C,D,E +
       0.0113 * reserved_room_type=G,F,B,C,D,E +
      -0.0022 * reserved_room_type=F,B,C,D,E +
      -0.0106 * reserved_room_type=B,C,D,E +
       0.0084 * reserved_room_type=C,D,E +
      -0.0034 * reserved_room_type=D,E +
       0.0022 * reserved_room_type=E +
       0.0034 * booking_changes +
       0.0027 * deposit_type=Refundable,No Deposit +
      -0.0032 * deposit_type=No Deposit +
       0.0038 * days_in_waiting_list +
       0.0013 * customer_type=Transient,Group,Contract +
      -0.0018 * customer_type=Group,Contract +
       0.0009 * total_of_special_requests +
       0.0274

Time taken to build model: 1.44 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.16 seconds

=== Summary ===

Correlation coefficient                  0.7391
Mean absolute error                      0.0043
Root mean squared error                  0.006
Relative absolute error                 64.1577 %
Root relative squared error             67.3731 %
Total Number of Instances            35817
```

- **Predicted ADR** using hotel factors
- **Key trends** city hotels, lead time, seasonality
- **Moderate performance** 0.7391 correlation
- **Improvement needed** error of 0.0043 units

Data Analytics
J-48 Decision Tree

Why J48 Decision Tree?

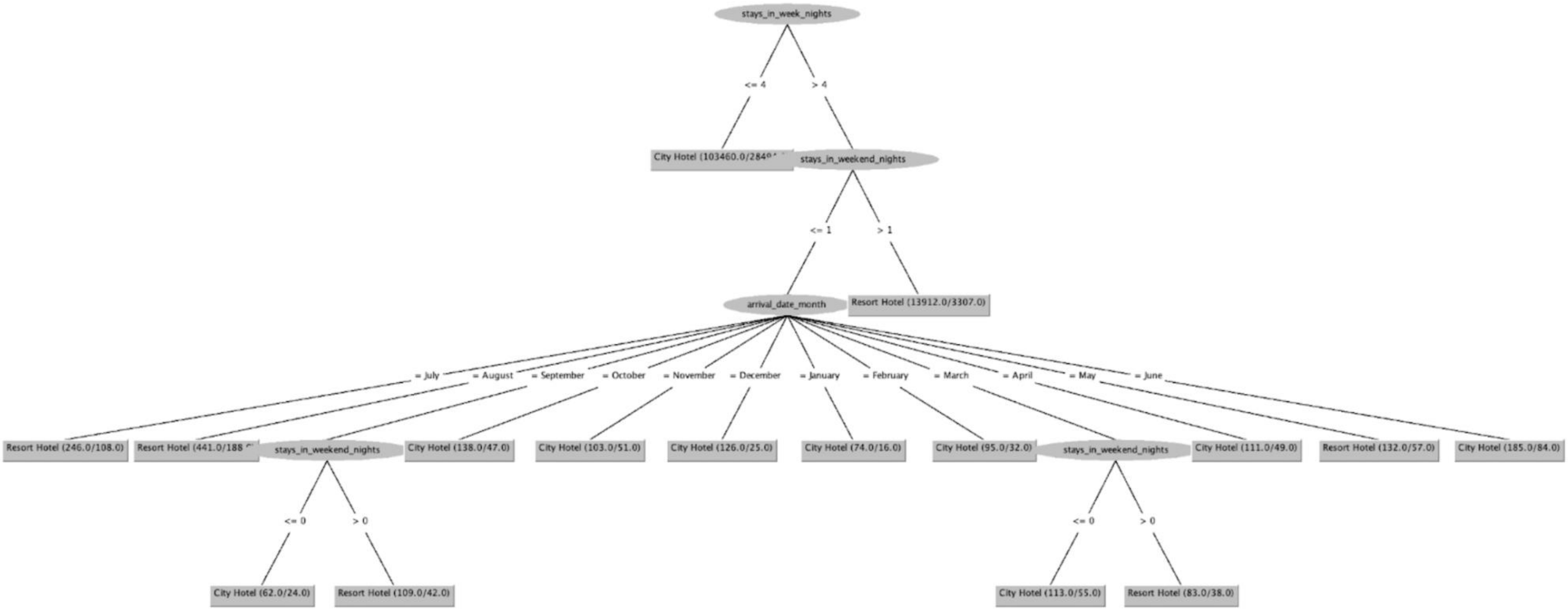
This decision tree will help:

- Predict the type of hotel booking based on stay patterns and seasonality.
- Understand patterns in guest behavior, such as which hotel is more popular during specific months or for specific durations of stay.
- Aid in marketing or operational decisions for hotels, like focusing offers on certain guest segments.

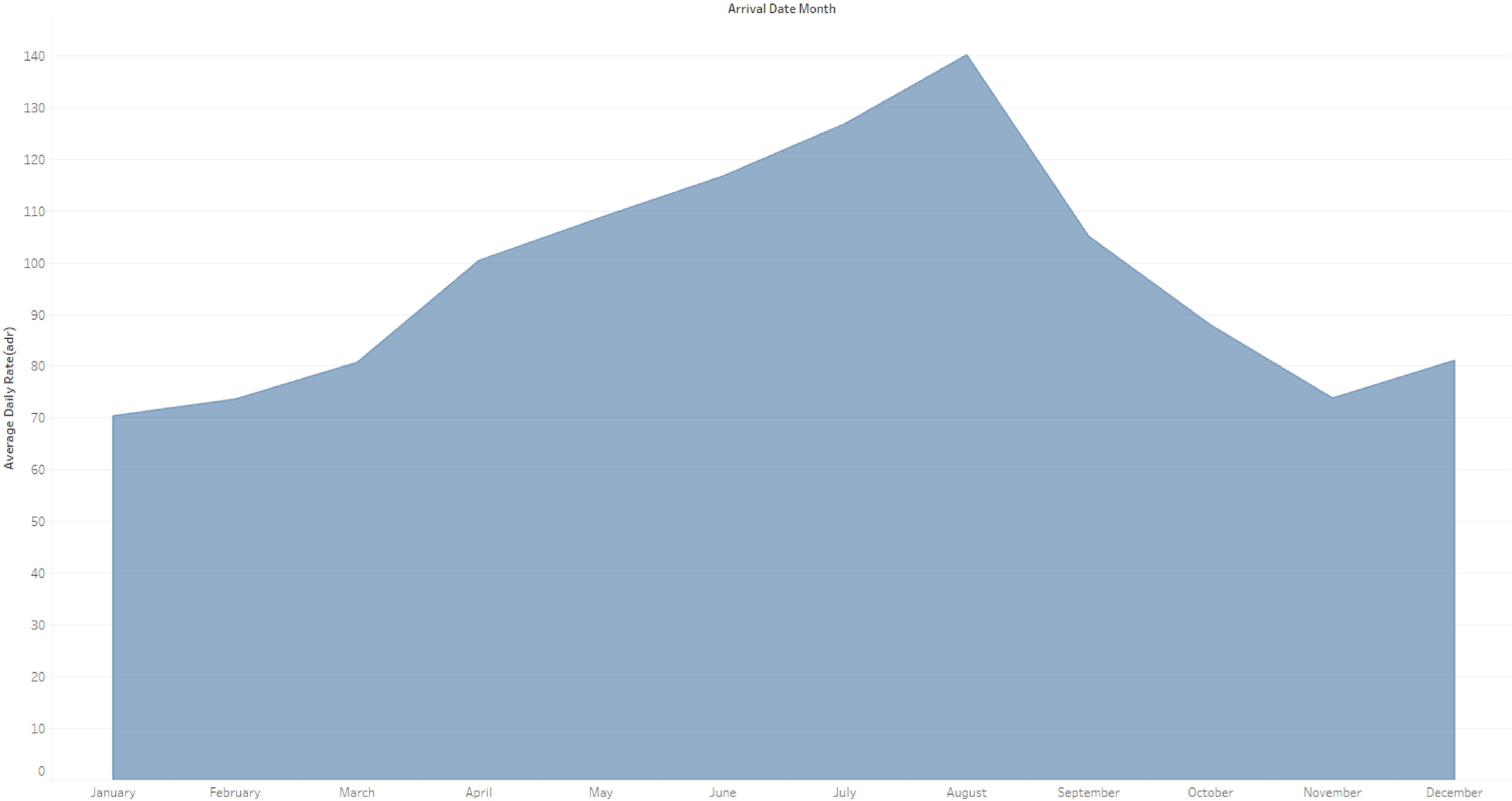
This decision tree focuses on predicting whether a hotel booking will be for a **City Hotel** or a **Resort Hotel** based on three main factors:

1.Length of Stay (Weekday Nights)
2.Length of Stay (Weekend Nights)
3.Arrival Month

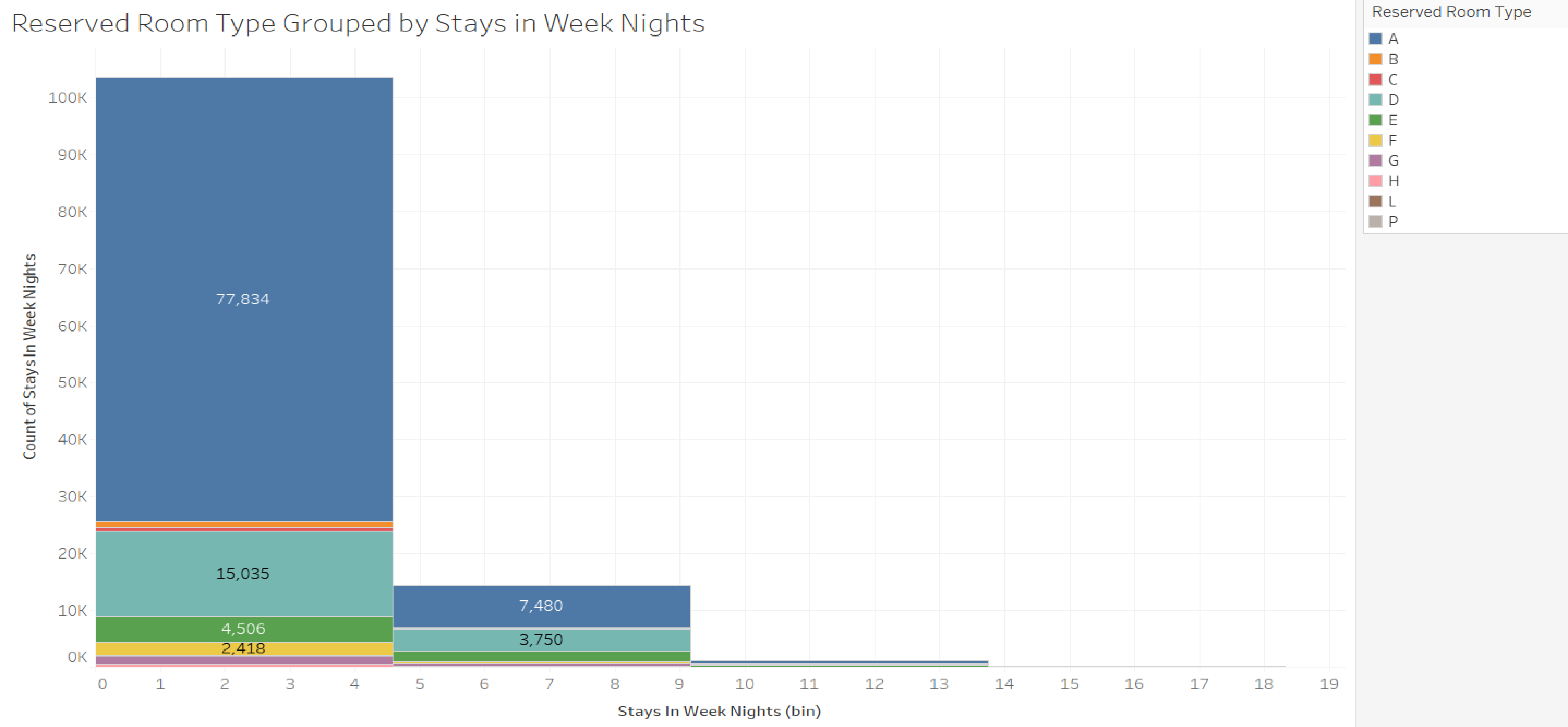
J-48 Decision Tree



Monthly Trends in Average Daily Rate (ADR)



Histogram

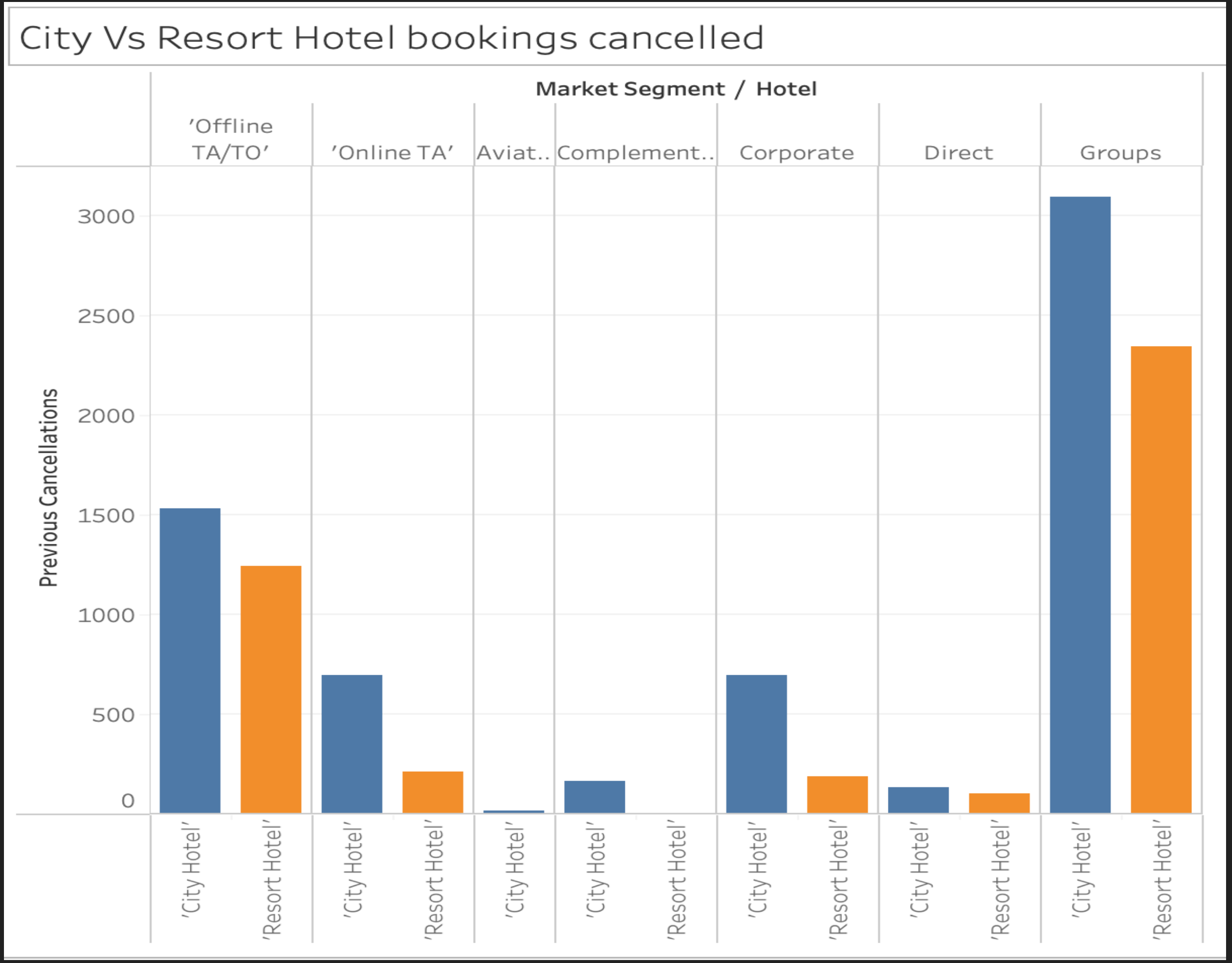


Reserved Room Type Grouped by Stays in Week Nights

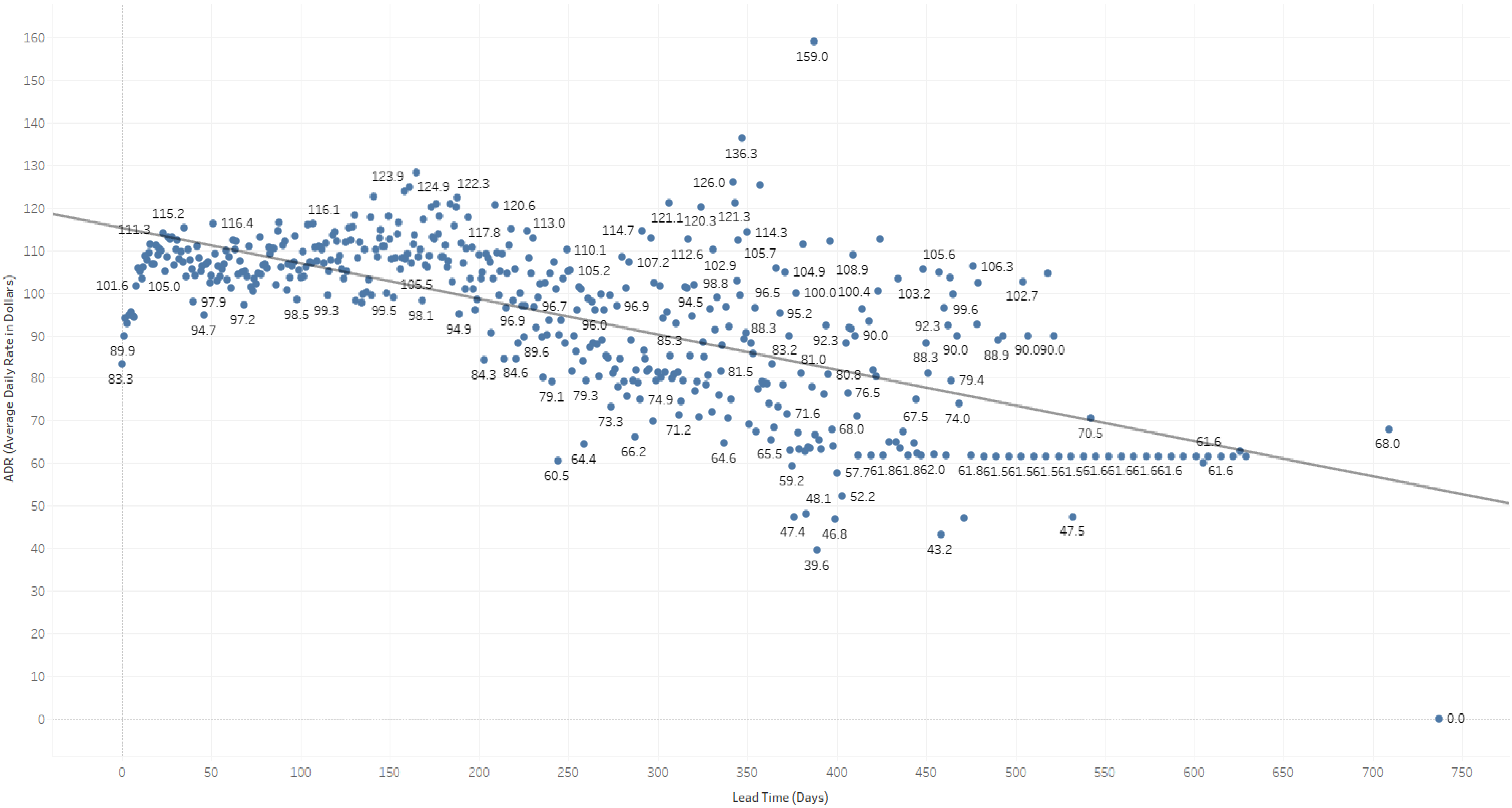
Bar Chart

- **City Hotels vs Resort Hotels:** City hotels consistently showed higher cancellation rates across all market segments. This could be due to factors like more frequent booking changes or higher business traveler volumes, where last-minute cancellations are more common.

- Bar charts allow us to effectively compare the cancellations of both city and resort hotels while also observing each market segment.



Impact of Lead Time on Average Daily Rate (ADR) Over Time


ADR (Average Daily Rate in Dollars)
Lead Time (Days)