Nguyen, Trang

Professor: Myriam Quispe-Agnoli

BDA 620: Data Mining

Customer Behavior Prediction Using Multiple Linear Regression (MLR) and K-NN Regression

1. Introduction

Lisbon is the capital of Portugal, one of the most popular tourist destinations in Europe, attracting significant numbers of tourists each year. According to the National Statistics Institute, the tourism industry is one of the most significant contributions to the economic growth of Portugal. The industry brought millions of euros to Portugal's economy, and tourism in Lisbon contributed the largest share of revenues to the country's economy compared to other regions in Portugal [1]. Therefore, the tourism industry, particularly the hotel sector in Lisbon, Portugal, is a highly competitive market.

This study examines a dataset of a hotel's customer demographic, geography, and behaviors in Lisbon, Portugal, from 2015 to 2018. The purpose of the study is to find which factors contributed the most to the hotel's revenue increase. Does customer behavior impact to the increasing revenue of the hotel? Which distribution channel and market segment play the most significant roles in the hotel's revenue growth? To clarify the questions, data mining techniques are applied to profile profitable hotel customers and develop marketing strategies to increase hotel revenue.

2. Literature Review

Several articles are reviewed to ensure that the modeling approach is rigorous and aligns with contemporary research before conducting the study. Magnini, Honeycutt, and Hodge, authors of the article 'Data Mining for Hotel Firms: Use and Limitations,' stated that to increase revenue, maximize profits, and formulate effective marketing strategies, it is essential to understand customer behaviors in the hotel industry, including the origin of potential guests, their room preferences, and expenditure patterns. The authors highlighted the importance of data mining as well as predictive data mining techniques, including regression-type models, clustering, classification, decision trees, and association rules in discovering patterns, relationships and predicting customer behavior trends in complex datasets; by contrast, traditional statistical models still face some obstacles in these areas [2].

Another study conducted by Aslihan and Meltem applied RFM (Recency, Frequency, and Monetary) analysis to analyze customer segmentation as well as to discover profitable hotel customers of three five-star hotels in Antalya, Turkey. Different customer groups were categorized based on customer's ages, needs, and demands. In principle, customers with high RFM scores are the most valuable customers. The study found that most customers aged 35 to 44 stayed for shorter periods and selected standard rooms. The majority of the customers were Russians and Germans. In the study's results, the authors highlighted the effective RFM method in the cluster characteristics of hotel customers to help managers to generate new strategies to improve the hotel's financial performance [3].

3. Data Descriptions and Visualization Analysis

3.1 Data Gathering and Data Dictionary

The dataset obtained in this study is obtained from Mercer's Stetson-Hatcher School of Business. It is a real-world customer dataset from a hotel in Lisbon, Portugal. The raw dataset described 80,000 customers and consisted of 32 variables. Table 1 displays the data dictionary indicating the types of variables in the dataset, including customer personal, behavioral, demographic, and geographical information over three years. The dataset includes a total of 13 numerical variables describing customer age, the average number of days elapsed between the customer's booking date and arrival date, amounts of customer spending, the numbers of booking cancellations, no-shows, and check-ins, as well as the total number of persons and rooms per night. Additionally, it includes the number of days since the last stay and the number of days since the first stay. There are also categorical variables describing customer information, customer group types, and booking types in the dataset, such as nationality, name, identification document, distribution channel, and market segment. Furthermore, there are 13 Boolean variables describing customer room preferences.

3.2 Data Cleaning and Preprocessing

Cleaning and preprocessing dataset are important steps before conducting the study. 3,598 missing values were presented in the Age variable, and they were removed from the dataset. According to the data dictionary, a value of 1 in the variables DaysSinceLastStay and DaysSinceFirstStay indicates that the customer never stayed at the hotel. However, there were

18,249 semantic inconsistencies in the dataset due to typing errors, where a negative value (-1) was used instead of the expected value of 1. The typing errors were then corrected to the value of 1. Outliers in the Age variable were identified, as observed in Figure 1 and Table 2, and they were then removed due to their unrealistic values, which were considered semantic inconsistencies in the dataset. In Figure 1, outliers are also presented in the box plots of AverageLeadTime, LodgingRevenue, OtherRevenue, PersonsNights, and RoomNights. They were also removed due to the impact of outliers on the accuracy of predictions for the study. After cleaning and pre-processing, the variables were transformed into their correct types for the study, as indicated in Table 1, and the clean dataset includes information on 46,188 customers with 34 variables.

3.3 Data Dimension – Reducing and Expansion

NameHash and DocIDHash were removed from the dataset due to privacy concerns and their lack of relevance. New variables were introduced for the study's purpose, including TotalRevenue, RatePerNight, and AgeGroups. Table 3 displays Generalized Variance Inflation Factor (GVIF) values for different variables, revealing a high GVIF of 60.902 for MarketSegment, indicating the presence of multicollinearity. Therefore, a new variable named MarketSegmentCombined was created to group categories with insignificant differences in MarketSegment. To illustrate, two market segments, Other and Complementary, were combined into a new variable named Other-Complementary due to minimal differences between the two groups, as seen in Table 4 of Tukey's Multiple Comparisons of Means for TotalRevenue Across MarketSegment Levels.

3.4 **Data Mining Task Determination and Its Importance in Marketing Analytics**

The study aimed to identify the primary factors influencing the hotel's revenue growth. It examined whether customer behavior impacts the increase in hotel revenue and which distribution channel and market segment play the most significant roles in the hotel's revenue growth. To determine these impacts, exploratory data analysis was conducted to identify significant relationships between predictors and the target variable. The relationships between variables and patterns in the data are displayed in Figure 2. Data mining techniques were utilized in the study for marketing analytics purposes, examining factors that positively influence the hotel's revenue, including customer behaviors, booking types, and market segments. The objective is to enhance the hotel's profitability.

3.5 Partition Data

Before conducting the study, the cleaned dataset was partitioned into a 60% training set and a 40% validation set. The training dataset included 28,051 observations with 34 variables, while the validation set, also with 34 variables, described 18,701 observations. The purpose of using the training set in this study was to train models to learn patterns and relationships in the data. After the model was trained, it was tested on the validation set. The performance of the model on the validation set provided an estimate of how well the model might perform on new and unseen data.

4. Data Methodologies and Techniques

4.1 Multiple Linear Regression Model

In this study, a multiple linear regression model was employed to explore the relationship between customer preferences and the total revenues of the hotel. The dependent variable is the continuous outcome variable, TotalRevenue. The predictors in this study are Age, AverageLeadTime, OtherRevenue, BookingCanceled, BookingCheckedIn, RoomNights, DaysSinceLastStay, various variables related to customer room preferences, RatePerNight, DistributionChannel, and MarketSegmentCombined. According to Table 6 of the Pivot Table of Hotel Room Preferences, the percentages of SRAccessibleRoom, SRAwayFromElevator, SRNearElevator, and SRNoAlcoholInMiniBar are very low, indicating a low percentage of customer requirements. Therefore, these variables were omitted from the linear regression model.

The summary of the multiple linear regression model is displayed in Table 8. According to Table 8, OtherRevenue, RoomNights, SRHighFloor1, SRQuietRoom1, RatePerNight, and MarketSegment: Other_Complementary, Direct are statistically significant and positively contribute to the increase in the hotel's revenue (p-values < 0.05). If a customer stays at the hotel for an additional night, the model predicts an increase of 107.90 Euros in the hotel’s revenue. The Multiple R-squared is 0.896, indicating that the model explains 89.6% of the variance in TotalRevenue.

4.2 K-NN Regression Model

K-NN is a non-parametric as well as a supervised machine learning method utilized for both classification and regression tasks. K-NN Regression Model was employed in this study to predict the continuous outcome variable, TotalRevenue. The principle of the K-NN algorithm makes predictions by finding the K nearest data points to a given input. For numerical regression, it averages their target values [4]. The predictor variables used to predict the TotalRevenue are AverageLeadTime, PersonsNights, RoomNights, RatePerNight, OtherRevenue, DistributionChannel, and MarketSegmentCombined. As seen in Table 7, the optimal k value in this study is 3 based on the smallest RMSE, indicating that according to the cross-validated results, the model with k = 3 performs best in terms of RMSE among the tested values of k.

5. **Results**

Table 9 presents evaluation metrics comparing the Multiple Linear Regression model (MLR) and the K-NN Regression model. In comparison, MLR has a Mean Error (ME) of 0.257, indicating that the prediction of the hotel's total revenue is above the actual value. On the other hand, K-NN Regression has a lower mean error (-2.45), suggesting that the prediction of the hotel's total revenue is below the actual value. However, considering other metrics, K-NN Regression demonstrates better performance than MLR with lower errors. As seen in Figure 3, the scatter plot of K-NN shows the model fitting the data well. It is concluded that the K-NN model delivers more accurate predictions on the dataset.

To understand customer behavior, K-Means Clustering Analysis was applied in this study to cluster data based on similarities in customer preferences and customer spending behavior. In this analysis, the optimal k value determined by Elbow Method is 3, as indicated in Figure 4. With the optimal k value of 3, Table 10 shows that tourists traveling together in large groups, who stay for an extended period at the hotel, prefer to book in advance, typically a few months before arrival, as they can get a lower RatePerNight. In addition, as seen in Figures 5 and 6, most tourists prefer booking through a Travel Agent/Operator because the average RatePerNight for bookings through a Travel Agent/Operator is lower than for customers booked through the direct distribution channel.

Table 11 reveals that there were 26,799 customers in the market segment of Other – Complementary, and 6,169 customers in the market segment of Travel Agent/Operator who booked through a Travel Agent/Operator. To illustrate the accuracy of the evaluation, Table 12 demonstrates the implementation of the K-NN regression model to predict the price difference between different market segments and distribution channels. Customer A and Customer B booked the hotel at the same time, planning to stay for 1 night. However, Customer A, in the market segment as a group, goes with a group booking directly at the hotel and will pay 138 Euros, which is higher than Customer B, who is in the market segment of complementary booking at a Travel Agent/Operator. Therefore, the Travel Agent/Operator segment contributed the highest total revenue to the hotel from 2015 to 2018 compared to other booking distribution channels.

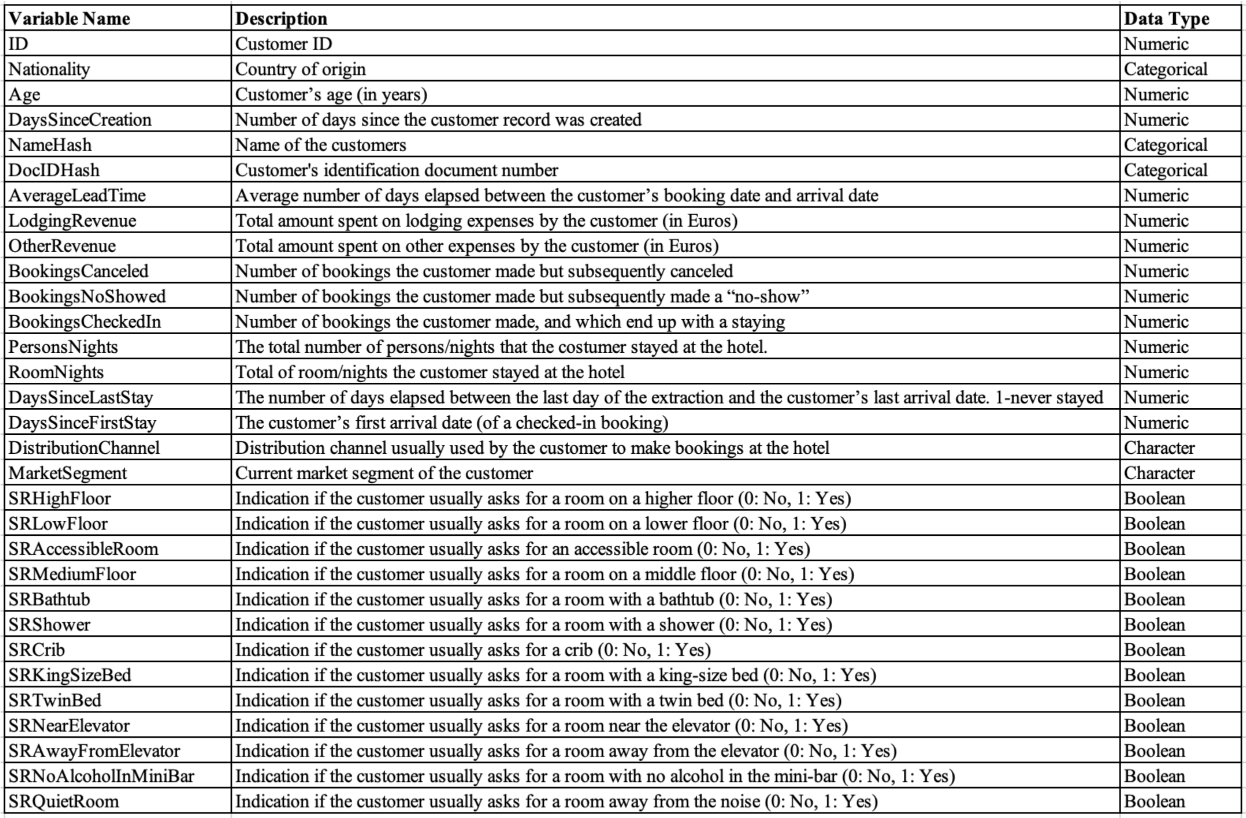
Figure 7, the Cumulative Total Revenue Gains Chart, and Figure 8, the Decile-wise Lift Chart, also reveal that the revenue of the top 10% of customers (1,870 customers) is predicted to be a total of 1,486,464 Euros, which is 2.06 times higher than the revenue of a random 1,870 customers, which is 694,569 Euros.

6. Conclusions and Recommendations

It is concluded that between Multiple Linear Regression (MLR) and K-NN Regression models, as seen in Table 9 and Figure 3, K-NN Regression outperforms MLR with superior performance in various metrics, while MLR overestimated the prediction of the hotel's revenue. The measurement of the average percentage difference between predicted and actual values in MLR is 12.4%, which is higher than that of K-NN (3.57%). Overall, the K-NN model provides more accurate predictions for the dataset. The model can be implemented for forecasting revenues as well as predicting customer behaviors and applied in marketing analytics.

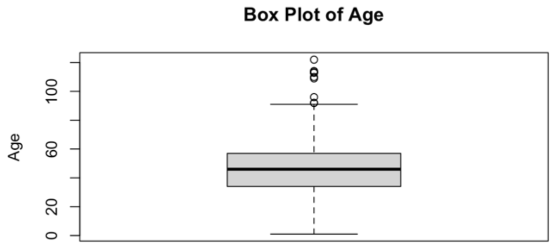
To increase the hotel's profit, it is recommended that the hotel should consider implementing strategies to attract customers seeking cheaper rates, such as special promotions, discounts, or package deals. The hotel should also consider offering exclusive deals or partnerships to encourage more bookings through Travel Agents/Operators and target the complementary market segment. In addition, the hotel should have customer loyalty programs to encourage customers to return by offering rewards, discounts, or exclusive packages to loyal customers.

Appendix

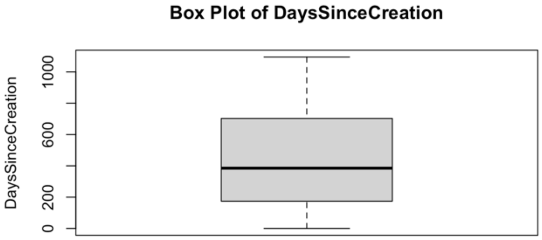


Variable Name	Description	Data Type
ID	Customer ID	Numeric
Nationality	Country of origin	Categorical
Age	Customer’s age (in years)	Numeric
DaysSinceCreation	Number of days since the customer record was created	Numeric
NameHash	Name of the customers	Categorical
DocIDHash	Customer's identification document number	Categorical
AverageLeadTime	Average number of days elapsed between the customer’s booking date and arrival date	Numeric
LodgingRevenue	Total amount spent on lodging expenses by the customer (in Euros)	Numeric
OtherRevenue	Total amount spent on other expenses by the customer (in Euros)	Numeric
BookingsCanceled	Number of bookings the customer made but subsequently canceled	Numeric
BookingsNoShowed	Number of bookings the customer made but subsequently made a “no-show”	Numeric
BookingsCheckedIn	Number of bookings the customer made, and which end up with a staying	Numeric
PersonsNights	The total number of persons/nights that the costumer stayed at the hotel.	Numeric
RoomNights	Total of room/nights the customer stayed at the hotel	Numeric
DaysSinceLastStay	The number of days elapsed between the last day of the extraction and the customer’s last arrival date. 1-never stayed	Numeric
DaysSinceFirstStay	The customer’s first arrival date (of a checked-in booking)	Numeric
DistributionChannel	Distribution channel usually used by the customer to make bookings at the hotel	Character
MarketSegment	Current market segment of the customer	Character
SRHighFloor	Indication if the customer usually asks for a room on a higher floor (0: No, 1: Yes)	Boolean
SRLowFloor	Indication if the customer usually asks for a room on a lower floor (0: No, 1: Yes)	Boolean
SRAccessibleRoom	Indication if the customer usually asks for an accessible room (0: No, 1: Yes)	Boolean
SRMediumFloor	Indication if the customer usually asks for a room on a middle floor (0: No, 1: Yes)	Boolean
SRBathtub	Indication if the customer usually asks for a room with a bathtub (0: No, 1: Yes)	Boolean
SRShower	Indication if the customer usually asks for a room with a shower (0: No, 1: Yes)	Boolean
SRCrib	Indication if the customer usually asks for a crib (0: No, 1: Yes)	Boolean
SRKingSizeBed	Indication if the customer usually asks for a room with a king-size bed (0: No, 1: Yes)	Boolean
SRTwinBed	Indication if the customer usually asks for a room with a twin bed (0: No, 1: Yes)	Boolean
SRNearElevator	Indication if the customer usually asks for a room near the elevator (0: No, 1: Yes)	Boolean
SRAwayFromElevator	Indication if the customer usually asks for a room away from the elevator (0: No, 1: Yes)	Boolean
SRNoAlcoholInMiniBar	Indication if the customer usually asks for a room with no alcohol in the mini-bar (0: No, 1: Yes)	Boolean
SRQuietRoom	Indication if the customer usually asks for a room away from the noise (0: No, 1: Yes)	Boolean

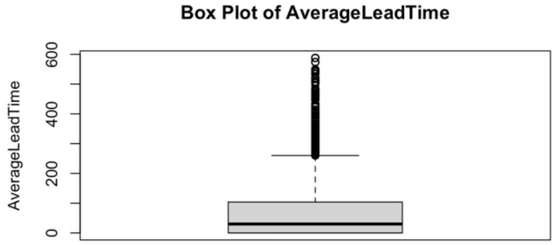
Table 1: Data Dictionary Table of Contributors in the Study



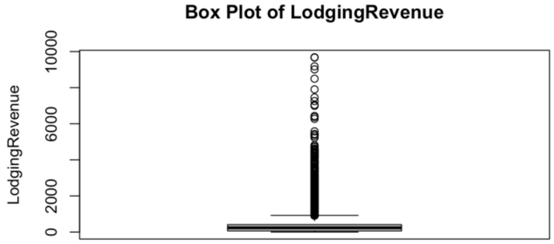
Box Plot of Age



Box Plot of DaysSinceCreation



Box Plot of AverageLeadTime



Box Plot of LodgingRevenue

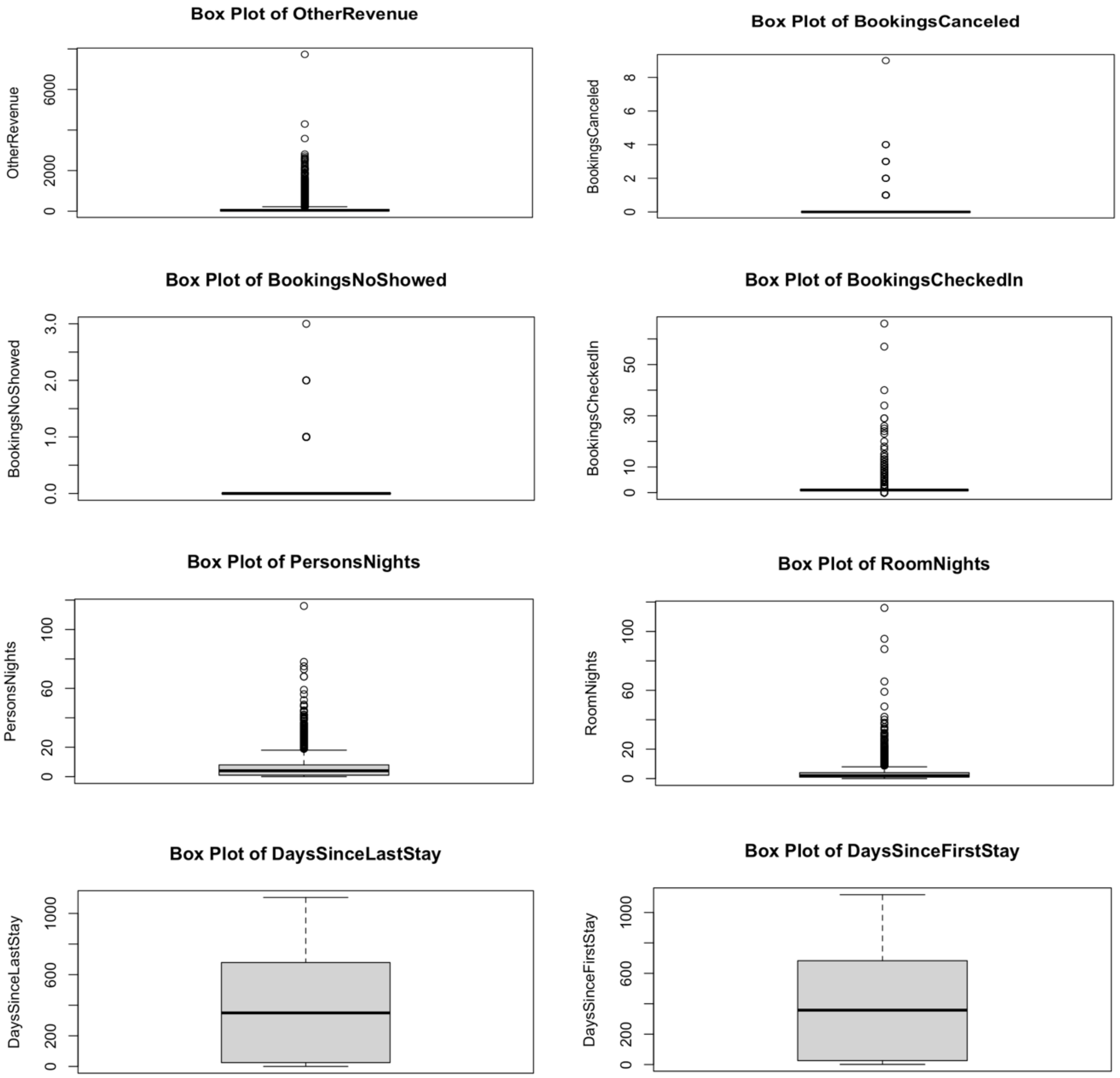
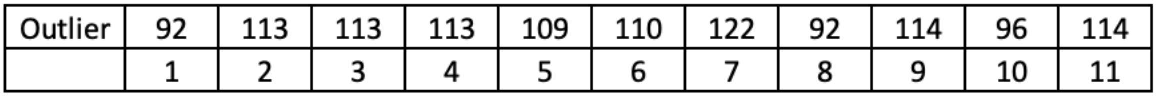
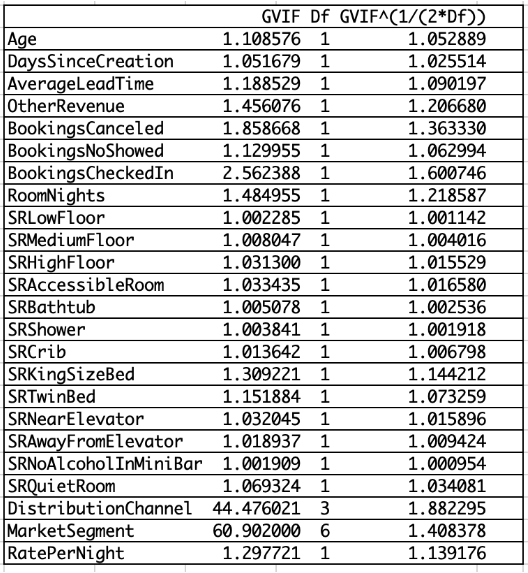


Figure 1. Box Plots of Numerical Variables in the Dataset



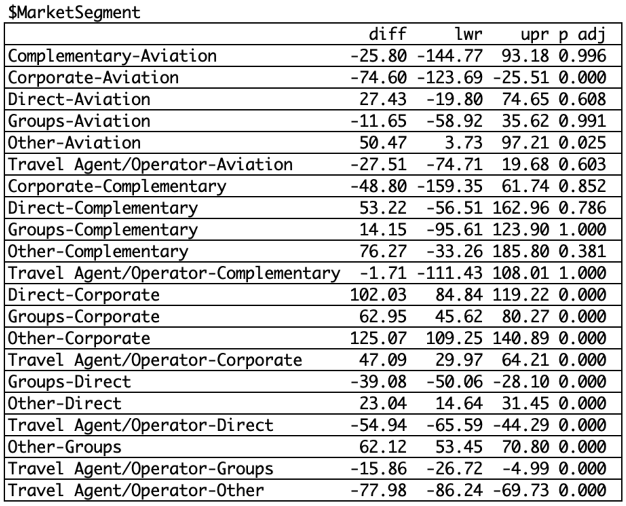
Outlier	92	113	113	113	109	110	122	92	114	96	114
	1	2	3	4	5	6	7	8	9	10	11

Table 2. Outliers Identified in The Variable `Age`



	GVIF	Df	GVIF^(1/(2*Df))
Age	1.108576	1	1.052889
DaysSinceCreation	1.051679	1	1.025514
AverageLeadTime	1.188529	1	1.090197
OtherRevenue	1.456076	1	1.206680
BookingsCanceled	1.858668	1	1.363330
BookingsNoShowed	1.129955	1	1.062994
BookingsCheckedIn	2.562388	1	1.600746
RoomNights	1.484955	1	1.218587
SRLowFloor	1.002285	1	1.001142
SRMediumFloor	1.008047	1	1.004016
SRHighFloor	1.031300	1	1.015529
SRAccessibleRoom	1.033435	1	1.016580
SRBathtub	1.005078	1	1.002536
SRShower	1.003841	1	1.001918
SRCrib	1.013642	1	1.006798
SRKingSizeBed	1.309221	1	1.144212
SRTwinBed	1.151884	1	1.073259
SRNearElevator	1.032045	1	1.015896
SRAwayFromElevator	1.018937	1	1.009424
SRNoAlcoholInMiniBar	1.001909	1	1.000954
SRQuietRoom	1.069324	1	1.034081
DistributionChannel	44.476021	3	1.882295
MarketSegment	60.902000	6	1.408378
RatePerNight	1.297721	1	1.139176

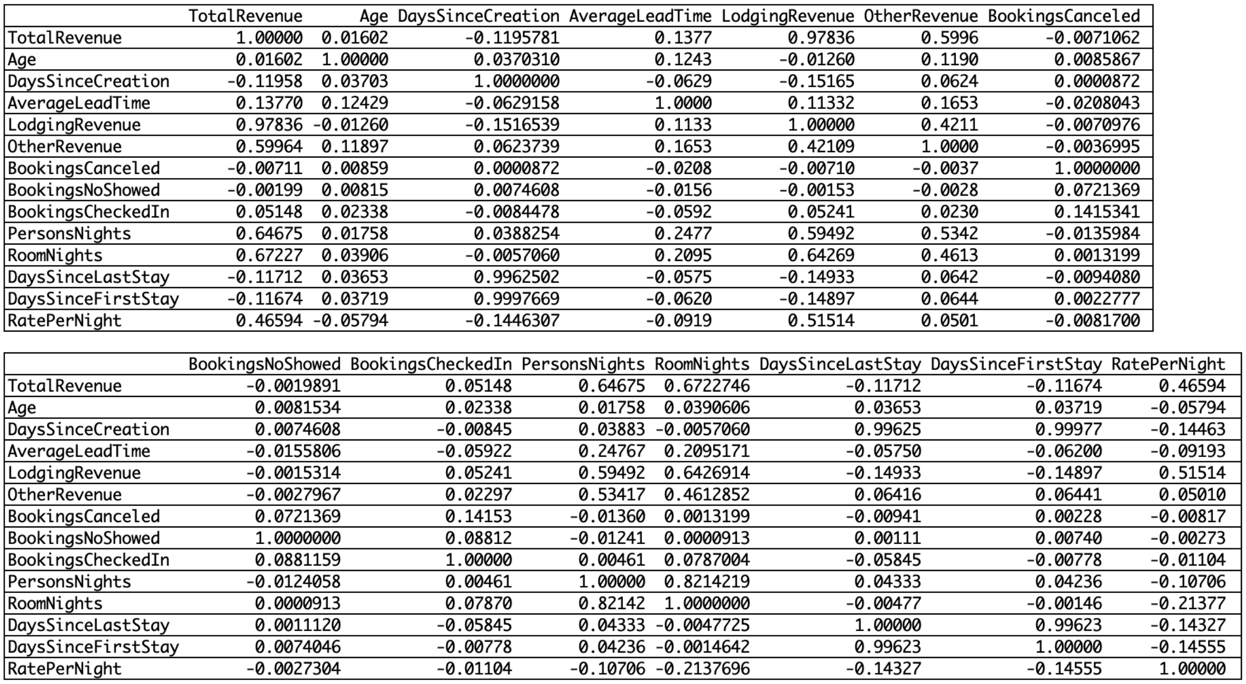
Table 3. VIF Evaluation: Identifying Multicollinearity



$MarketSegment

	diff	lwr	upr	p adj
Complementary-Aviation	-25.80	-144.77	93.18	0.996
Corporate-Aviation	-74.60	-123.69	-25.51	0.000
Direct-Aviation	27.43	-19.80	74.65	0.608
Groups-Aviation	-11.65	-58.92	35.62	0.991
Other-Aviation	50.47	3.73	97.21	0.025
Travel Agent/Operator-Aviation	-27.51	-74.71	19.68	0.603
Corporate-Complementary	-48.80	-159.35	61.74	0.852
Direct-Complementary	53.22	-56.51	162.96	0.786
Groups-Complementary	14.15	-95.61	123.90	1.000
Other-Complementary	76.27	-33.26	185.80	0.381
Travel Agent/Operator-Complementary	-1.71	-111.43	108.01	1.000
Direct-Corporate	102.03	84.84	119.22	0.000
Groups-Corporate	62.95	45.62	80.27	0.000
Other-Corporate	125.07	109.25	140.89	0.000
Travel Agent/Operator-Corporate	47.09	29.97	64.21	0.000
Groups-Direct	-39.08	-50.06	-28.10	0.000
Other-Direct	23.04	14.64	31.45	0.000
Travel Agent/Operator-Direct	-54.94	-65.59	-44.29	0.000
Other-Groups	62.12	53.45	70.80	0.000
Travel Agent/Operator-Groups	-15.86	-26.72	-4.99	0.000
Travel Agent/Operator-Other	-77.98	-86.24	-69.73	0.000

Table 4. Tukey's Multiple Comparisons of Means for TotalRevenue Across MarketSegment Levels



	TotalRevenue	Age	DaysSinceCreation	AverageLeadTime	LodgingRevenue	OtherRevenue	BookingsCanceled	BookingsNoShowed	BookingsCheckedIn	PersonsNights	RoomNights	DaysSinceLastStay	DaysSinceFirstStay	RatePerNight
TotalRevenue	1.00000	0.01602	-0.1195781	0.1377	0.97836	0.5996	-0.0071062	-0.0019891	0.05148	0.64675	0.6722746	-0.11712	-0.11674	0.46594
Age	0.01602	1.00000	0.0370310	0.1243	-0.01260	0.1190	0.0085867	0.0081534	0.02338	0.01758	0.0390606	0.03653	0.03719	-0.05794
DaysSinceCreation	-0.11958	0.03703	1.0000000	-0.0629	-0.15165	0.0624	0.0000872	0.0074608	-0.00845	0.03883	-0.0057060	0.99625	0.99977	-0.14463
AverageLeadTime	0.13770	0.12429	-0.0629158	1.0000	0.11332	0.1653	-0.0208043	-0.0155806	-0.05922	0.24767	0.2095171	-0.05750	-0.06200	-0.09193
LodgingRevenue	0.97836	-0.01260	-0.1516539	0.1133	1.00000	0.4211	-0.0070976	-0.0015314	0.05241	0.59492	0.6426914	-0.14933	-0.14897	0.51514
OtherRevenue	0.59964	0.11897	0.0623739	0.1653	0.42109	1.0000	-0.0036995	-0.0027967	0.02297	0.53417	0.4612852	0.06416	0.06441	0.05010
BookingsCanceled	-0.00711	0.00859	0.0000872	-0.0208	-0.00710	-0.0037	1.0000000	0.0721369	0.14153	-0.01360	0.0013199	-0.00941	0.00228	-0.00817
BookingsNoShowed	-0.00199	0.00815	0.0074608	-0.0156	-0.00153	-0.0028	0.0721369	1.0000000	0.08812	-0.01241	0.0000913	0.00111	0.00740	-0.00273
BookingsCheckedIn	0.05148	0.02338	-0.0084478	-0.0592	0.05241	0.0230	0.1415341	0.0881159	1.00000	0.00461	0.0787004	-0.05845	-0.00778	-0.01104
PersonsNights	0.64675	0.01758	0.0388254	0.2477	0.59492	0.5342	-0.0135984	-0.0124058	0.00461	1.00000	0.8214219	0.04333	0.04236	-0.10706
RoomNights	0.67227	0.03906	-0.0057060	0.2095	0.64269	0.4613	0.0013199	0.0000913	0.07870	0.82142	1.0000000	-0.00477	-0.00146	-0.21377
DaysSinceLastStay	-0.11712	0.03653	0.9962502	-0.0575	-0.14933	0.0642	-0.0094080	0.0011120	-0.05845	0.04333	-0.0047725	1.00000	0.99623	-0.14327
DaysSinceFirstStay	-0.11674	0.03719	0.9997669	-0.0620	-0.14897	0.0644	0.0022777	0.0074046	-0.00778	0.04236	-0.0014642	0.99623	1.00000	-0.14555
RatePerNight	0.46594	-0.05794	-0.1446307	-0.0919	0.51514	0.0501	-0.0081700	-0.0027304	-0.01104	-0.10706	-0.2137696	-0.14327	-0.14555	1.00000

Table 5. Correlation Matrix between Continuous Variables

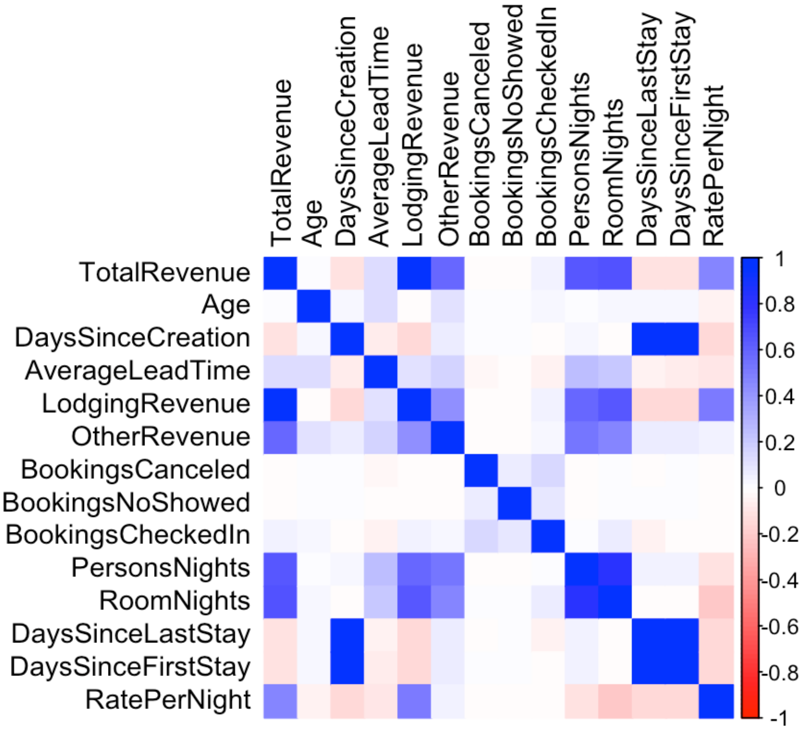
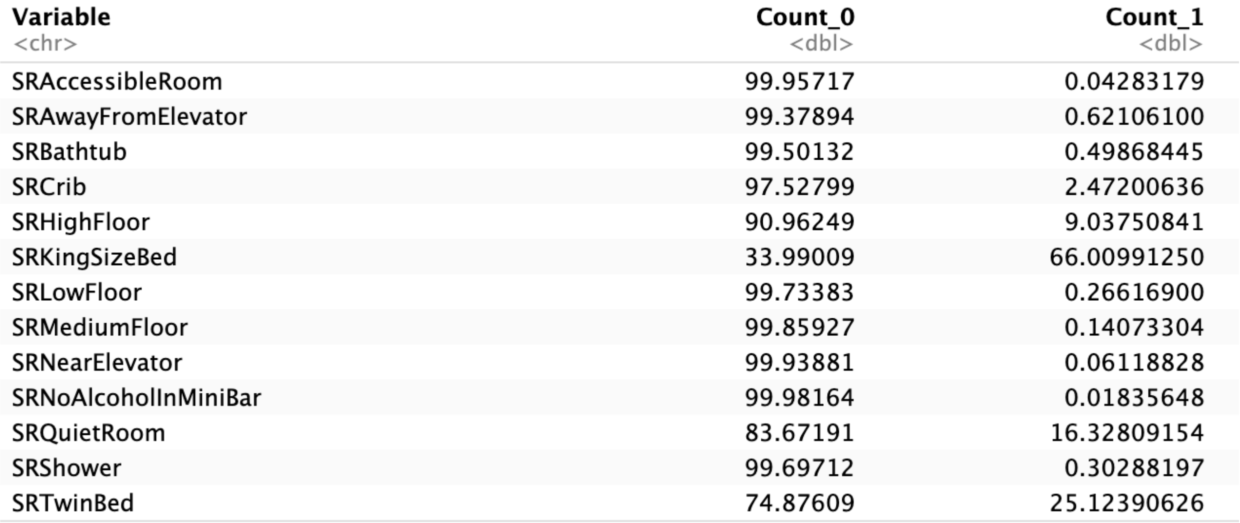
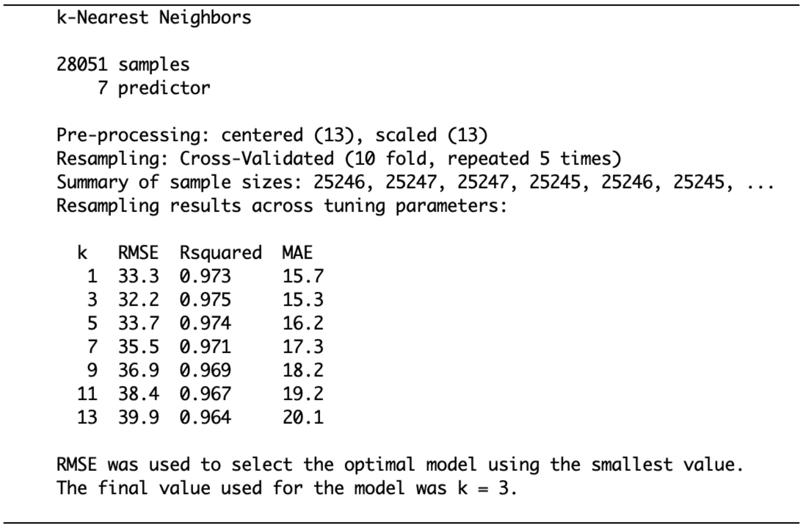


Figure 2. Correlation Matrix between Continuous Variables



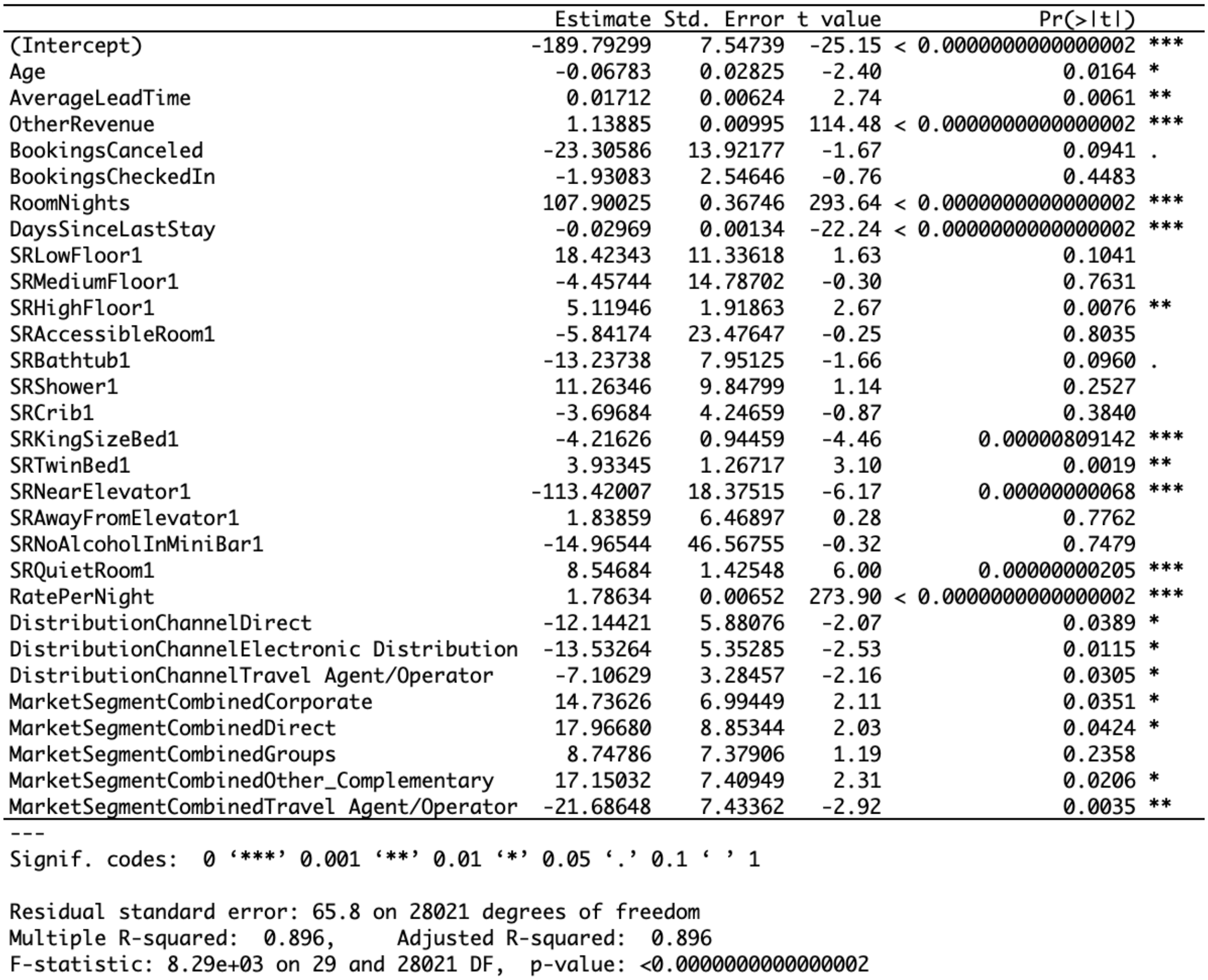
Variable	Count_0	Count_1
<chr>	<dbl>	<dbl>
SRAccessibleRoom	99.95717	0.04283179
SRAwayFromElevator	99.37894	0.62106100
SRBathtub	99.50132	0.49868445
SRCrib	97.52799	2.47200636
SRHighFloor	90.96249	9.03750841
SRKingSizeBed	33.99009	66.00991250
SRLowFloor	99.73383	0.26616900
SRMediumFloor	99.85927	0.14073304
SRNearElevator	99.93881	0.06118828
SRNoAlcoholInMiniBar	99.98164	0.01835648
SRQuietRoom	83.67191	16.32809154
SRShower	99.69712	0.30288197
SRTwinBed	74.87609	25.12390626

Table 6. Pivot Table of Hotel Room References



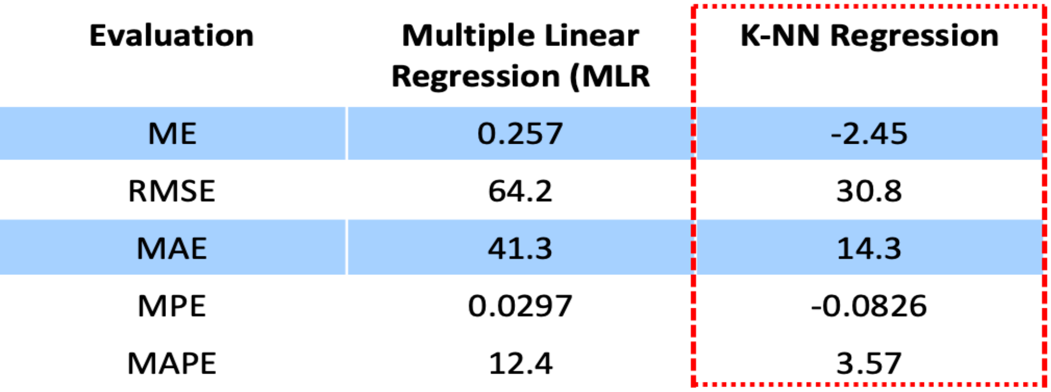
```
k-Nearest Neighbors

28051 samples
    7 predictor

Pre-processing: centered (13), scaled (13)
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 25246, 25247, 25247, 25245, 25246, 25245, ...
Resampling results across tuning parameters:

  k   RMSE  Rsquared  MAE 
   1  33.3  0.973     15.7
   3  32.2  0.975     15.3
   5  33.7  0.974     16.2
   7  35.5  0.971     17.3
   9  36.9  0.969     18.2
  11  38.4  0.967     19.2
  13  39.9  0.964     20.1

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 3.
```

Table 7. K-NN regression model



```
                                             Estimate Std. Error t value            Pr(>|t|)
(Intercept)                                -189.79299    7.54739  -25.15 < 0.0000000000000002 ***
Age                                          -0.06783    0.02825   -2.40              0.0164 *
AverageLeadTime                               0.01712    0.00624    2.74              0.0061 **
OtherRevenue                                  1.13885    0.00995  114.48 < 0.0000000000000002 ***
BookingsCanceled                            -23.30586   13.92177   -1.67              0.0941 .
BookingsCheckedIn                            -1.93083    2.54646   -0.76              0.4483
RoomNights                                  107.90025    0.36746  293.64 < 0.0000000000000002 ***
DaysSinceLastStay                            -0.02969    0.00134  -22.24 < 0.0000000000000002 ***
SRLowFloor1                                  18.42343   11.33618    1.63              0.1041
SRMediumFloor1                               -4.45744   14.78702   -0.30              0.7631
SRHighFloor1                                  5.11946    1.91863    2.67              0.0076 **
SRAccessibleRoom1                            -5.84174   23.47647   -0.25              0.8035
SRBathtub1                                  -13.23738    7.95125   -1.66              0.0960 .
SRShower1                                    11.26346    9.84799    1.14              0.2527
SRCrib1                                      -3.69684    4.24659   -0.87              0.3840
SRKingSizeBed1                               -4.21626    0.94459   -4.46          0.00000809142 ***
SRTwinBed1                                    3.93345    1.26717    3.10              0.0019 **
SRNearElevator1                            -113.42007   18.37515   -6.17          0.00000000068 ***
SRAwayFromElevator1                           1.83859    6.46897    0.28              0.7762
SRNoAlcoholInMiniBar1                       -14.96544   46.56755   -0.32              0.7479
SRQuietRoom1                                  8.54684    1.42548    6.00          0.00000000205 ***
RatePerNight                                  1.78634    0.00652  273.90 < 0.0000000000000002 ***
DistributionChannelDirect                   -12.14421    5.88076   -2.07              0.0389 *
DistributionChannelElectronic Distribution  -13.53264    5.35285   -2.53              0.0115 *
DistributionChannelTravel Agent/Operator     -7.10629    3.28457   -2.16              0.0305 *
MarketSegmentCombinedCorporate               14.73626    6.99449    2.11              0.0351 *
MarketSegmentCombinedDirect                  17.96680    8.85344    2.03              0.0424 *
MarketSegmentCombinedGroups                   8.74786    7.37906    1.19              0.2358
MarketSegmentCombinedOther_Complementary     17.15032    7.40949    2.31              0.0206 *
MarketSegmentCombinedTravel Agent/Operator  -21.68648    7.43362   -2.92              0.0035 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.8 on 28021 degrees of freedom
Multiple R-squared:  0.896,     Adjusted R-squared:  0.896
F-statistic: 8.29e+03 on 29 and 28021 DF,  p-value: <0.0000000000000002
```

Table 8. Multiple Linear Regression Model



Evaluation	Multiple Linear Regression (MLR	K-NN Regression
ME	0.257	-2.45
RMSE	64.2	30.8
MAE	41.3	14.3
MPE	0.0297	-0.0826
MAPE	12.4	3.57

Table 9. Evaluation Metrics between MLR model and K-NN Regression model

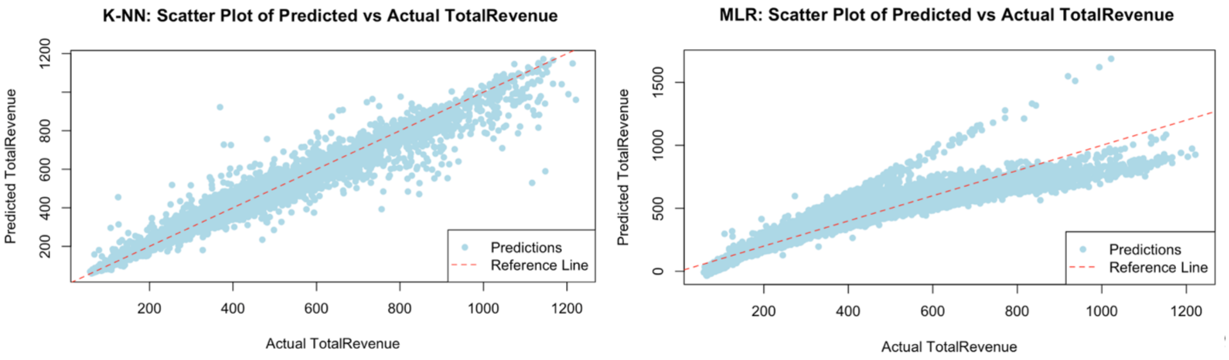


Figure 3. Scatter Plots of Predicted vs. Actual TotalRevenue in K-NN and MLR

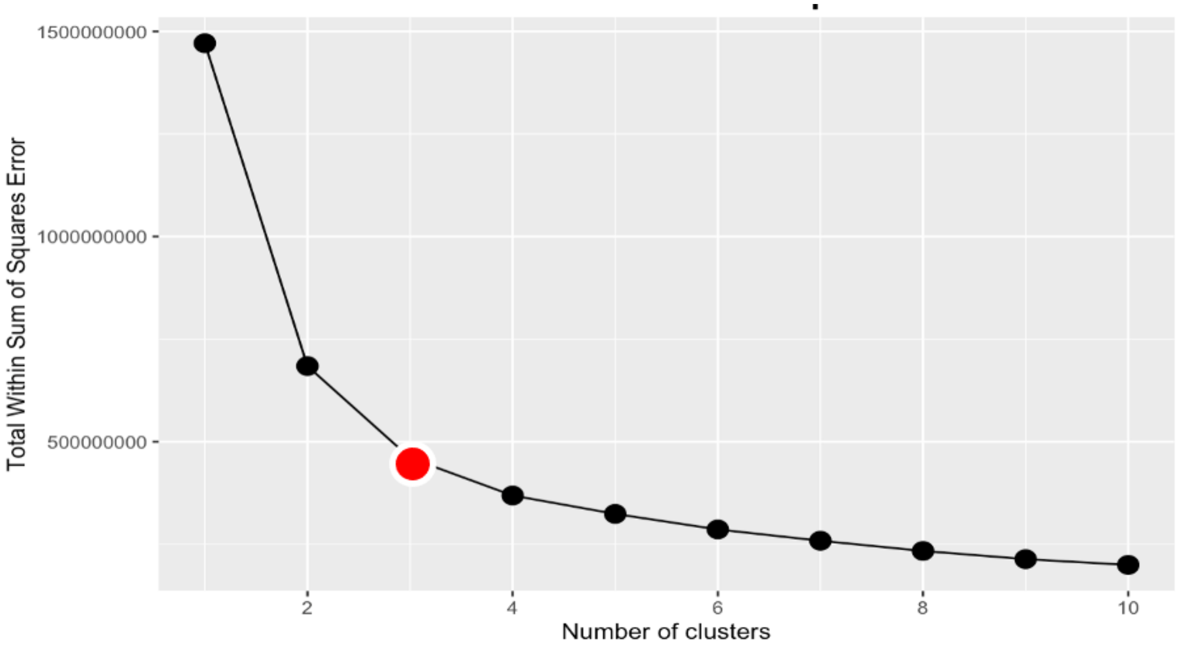
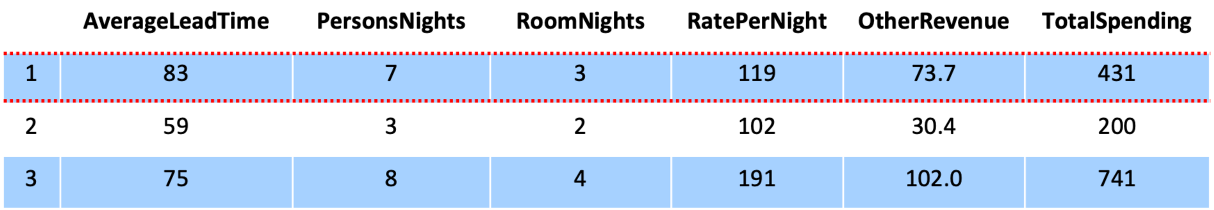


Figure 4. Elbow Method Determines Optimal k Value.



	AverageLeadTime	PersonsNights	RoomNights	RatePerNight	OtherRevenue	TotalSpending
1	83	7	3	119	73.7	431
2	59	3	2	102	30.4	200
3	75	8	4	191	102.0	741

Table 10. K-Means Clustering Analysis

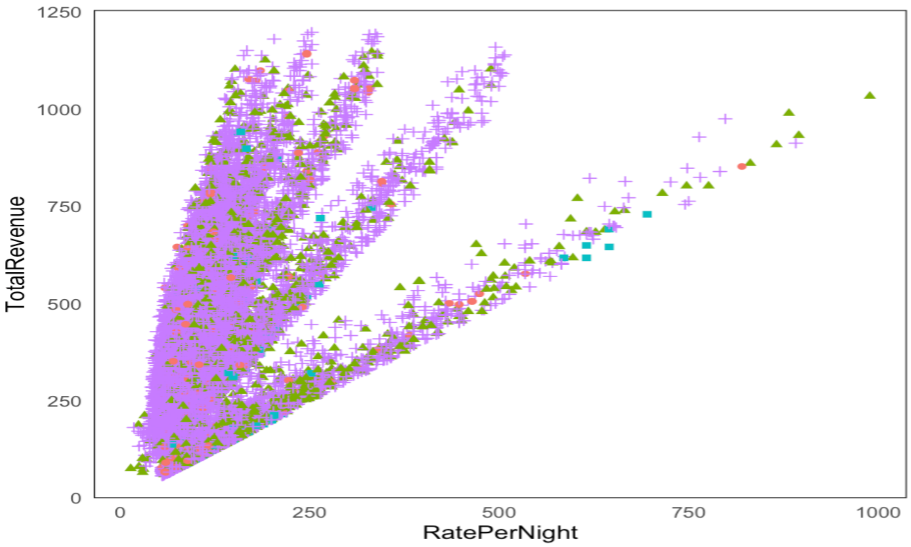


Figure 5. Total Revenue vs RatePerNight – Distribution Channel

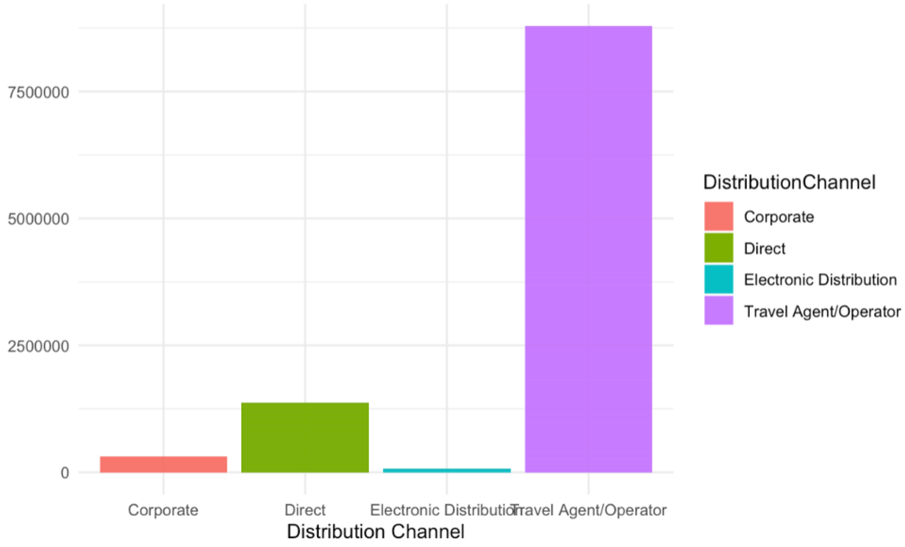
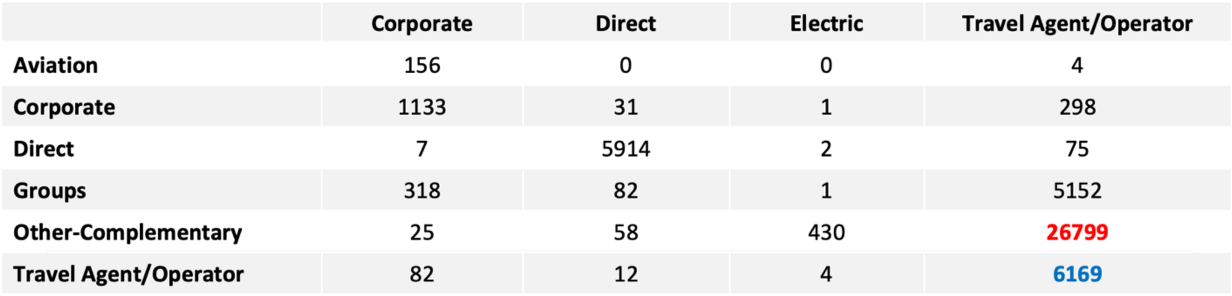
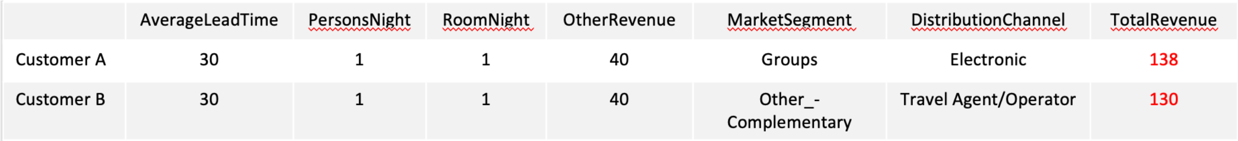


Figure 6. Total Revenue by Distribution Channel



	Corporate	Direct	Electric	Travel Agent/Operator
Aviation	156	0	0	4
Corporate	1133	31	1	298
Direct	7	5914	2	75
Groups	318	82	1	5152
Other-Complementary	25	58	430	**26799**
Travel Agent/Operator	82	12	4	**6169**

Table 11. Number of Customers in Different Market Segments Booking through Different Distribution Channels



	AverageLeadTime	PersonsNight	RoomNight	OtherRevenue	MarketSegment	DistributionChannel	TotalRevenue
Customer A	30	1	1	40	Groups	Electronic	138
Customer B	30	1	1	40	Other_-Complementary	Travel Agent/Operator	130

Table 12. Implement K-NN Regression Model to Predict Price

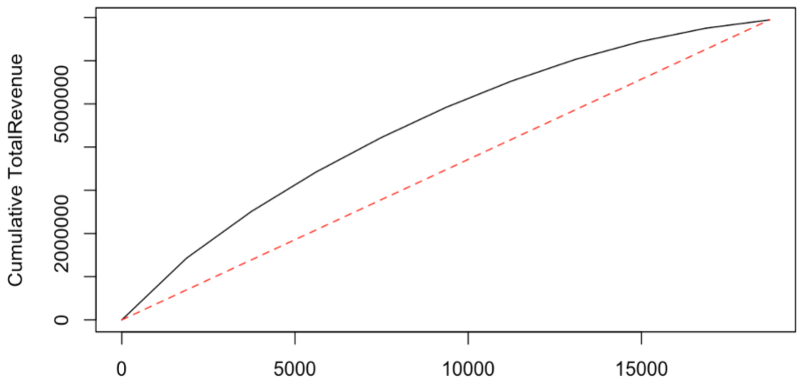


Figure 7. Cumulative Total Revenue Gains Chart

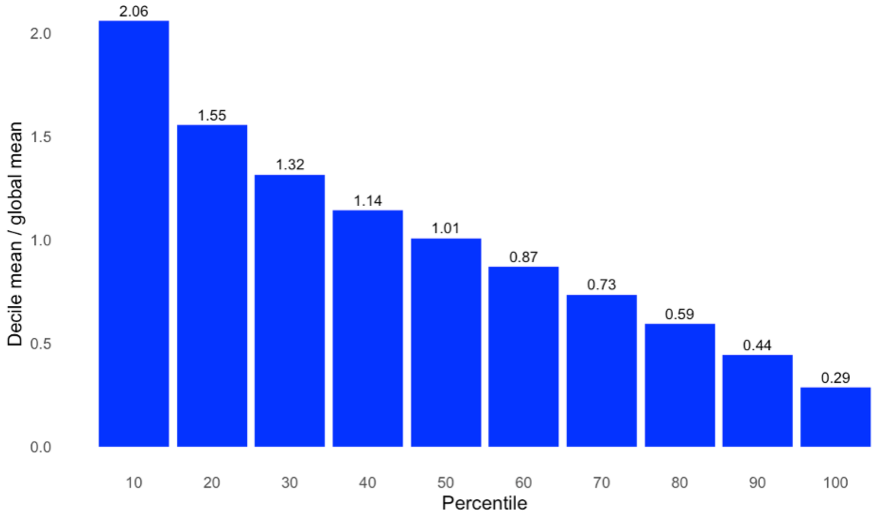


Figure 8. Decile-wise Lift Chart

References

1. Froyd, J. N. (2023, July 16). Tourism in Lisbon Reached a Record Revenue Per Room. Tourism Review. Retrieved October 29, 2023, from https://www.tourism-review.com/tourism-in-lisbon-reported-maximum-revenues-news13418.

2. Magnini, Vincent & Honeycutt, Earl & Hodge, Sharon. (2003). Data Mining for Hotel Firms: Use and Limitations. Cornell Hotel and Restaurant Administration Quarterly - CORNELL HOTEL RESTAUR ADMIN Q. 44. 94-105. 10.1177/0010880403442009.

3. Dursun-Cengizci, Aslıhan & Caber, Meltem. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tourism Management Perspectives. 18. 153-160. 10.1016/j.tmp.2016.03.001.

4. Verma, Nandini. (2023). Understanding K-Nearest Neighbors (KNN) Regression in Machine Learning. Retrieved December 13, 2023, from https://www.linkedin.com/pulse/understanding-k-nearest-neighbors-knn-regression-machine-verma-3jjqf/