WIE3007 Data Mining & Warehousing

Semester 1, Session 2023/2024

Lecturer

Prof. Dr. Teh Ying Wah

Group Assignment

Leveraging Data Mining for Enhanced Hotel Booking Strategies

Group Members	Student ID	Occ
Thayananth A/L Kumaresan	U2105590	1
Kalkazbek Alikhan	S2043519	2
Cher Jia Wen	U2005415	1
Visalini a/p Vijayan	U2102811	2
Ahmed Ishag	S2124564	1

Table of Contents

1.0 Introduction

1.1 Background

The dataset under consideration revolves around hotel bookings, a domain crucial to the functioning of the hospitality industry. In the hospitality sector, understanding the intricacies of booking patterns and discerning customer preferences is paramount.

The dataset provides an opportunity to delve into the nuances of guest behavior, enabling hoteliers to gain insights into the factors contributing to successful bookings or cancellations. By unraveling the patterns within the dataset, the hospitality industry can make informed decisions to improve service delivery, tailor marketing strategies, and elevate the guest experience.

In the context of this dataset, the hospitality industry is presented with a tool for retrospective analysis and predictive modeling. Identifying trends in booking behavior allows for anticipatory measures, aiding hotels in managing room availability, optimizing pricing strategies, and proactively addressing customer needs.

1.2 Objectives

- To uncover recurring patterns in booking behaviour for adapting to seasonal trends and optimizing resource allocation.
- To identify factors influencing cancellations by analyzing the dataset for the development of strategies to mitigate cancellations and enhance revenue management.
- To develop predictive models to forecast the likelihood of booking cancellations so that stakeholders can implement preventive measures and improve overall booking success rates.

2.0 Dataset Selection

2.1 Data Source

The dataset for this project was obtained from Kaggle source, a well-known platform for data science and machine learning resources. The reference for the dataset at the following URL: Kaggle Hotel Reservations Classification Dataset

2.2 Dataset Information

Features: no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space, room_type_reserved, lead_time, arrival_year, arrival_month, arrival_date, market_segment_type, repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests.

Target Variable: booking_status (Flag indicating if the booking was canceled or not).

2.3 Justification

The reason for choosing this hotel booking dataset is grounded in its inherent potential to provide valuable insights into booking behavior and the factors that influence cancellations. Several key factors contribute to the justification for choosing this dataset:

1. Relevance to Hospitality Industry:

The dataset revolves around hotel bookings, a critical aspect of the hospitality industry. Dataset analysis can offer insights directly applicable to the challenges and dynamics faced by hotels and accommodation providers.

2. Real-world Application of Data Mining:

The dataset aligns with the real-world application of data mining in the hospitality domain. By applying the SAS SEMMA methodology, we aim to extract meaningful patterns, insights, and models that can inform decision-making processes in the hotel industry.

3. Potential for Predictive Analysis:

With the inclusion of the 'booking_status' variable indicating whether a booking was canceled or not, the dataset offers a platform for predictive analysis. Understanding the patterns and variables influencing cancellations can be pivotal for hotels in optimizing their booking strategies.

In summary, the dataset's richness in relevant features, its alignment with real-world hospitality scenarios, and the potential to uncover actionable insights make it a compelling choice for this data mining project. The findings from this analysis are expected to contribute significantly to the enhancement of booking strategies and customer satisfaction within the hotel industry.

3.0 Understanding the Dataset

3.1 Structure and Features

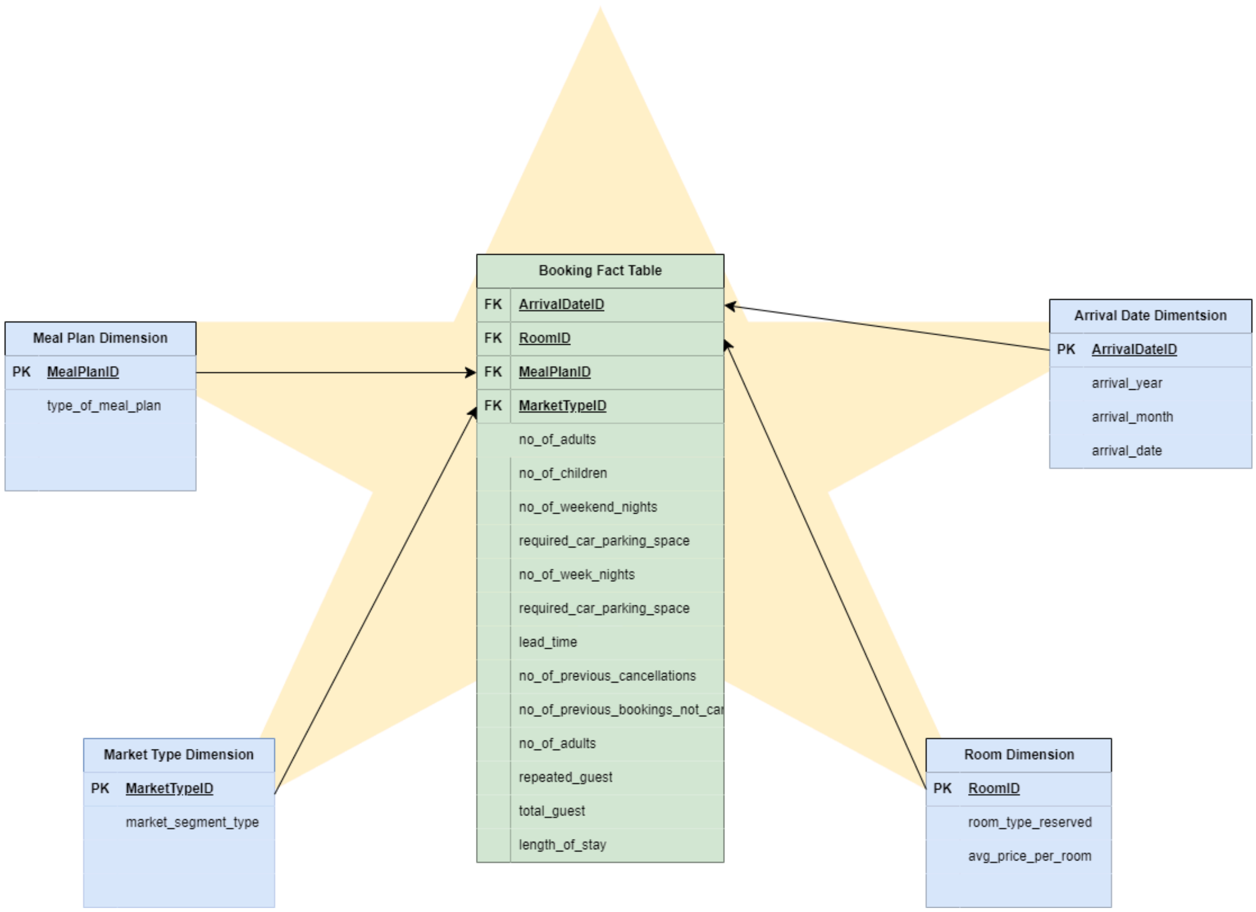
Number of Entries (Rows): 36275

Number of Features (Columns): 19

Field Name	Description	Data Type
Booking_ID	Unique identifier of each booking	String or Integer
no_of_adults	Number of adults	Integer
no_of_children	Number of children	Integer
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	Integer
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	Integer
type_of_meal_plan	Type of meal plan booked by the customer	String
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1 - Yes)	Integer (0 or 1)
room_type_reserved	Type of room reserved by the customer; values are ciphered (encoded) by INN Hotels	String
lead_time	Number of days between the date of booking and the arrival date	Integer
arrival_year	Year of arrival date	Integer
arrival_month	Month of arrival date	Integer
arrival_date	Date of the month	Integer
market_segment_type	Market segment designation	String
repeated_guest	Is the customer a repeated guest? (0 - No, 1 - Yes)	Integer (0 or 1)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking	Integer
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking	Integer

avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic (in euros)	Float or Decimal
no_of_special_requests	Total number of special requests made by the customer (e.g., high floor, view from the room)	Integer
booking_status	Flag indicating if the booking was canceled or not	String or Integer

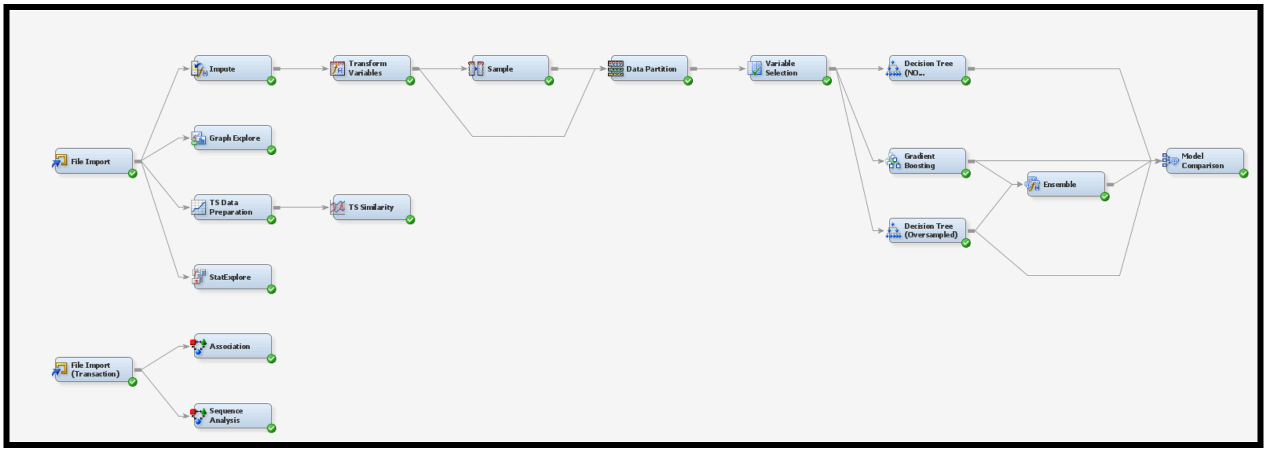
3.2 Star Schema



Clearer Image can be found here:
https://drive.google.com/file/d/1qy7oVVOp1J4DoTZR8GsQ8duR2lgUoGdn/view?usp=sharing

4.0 Applying SAS SEMMA Methodology

Final Model

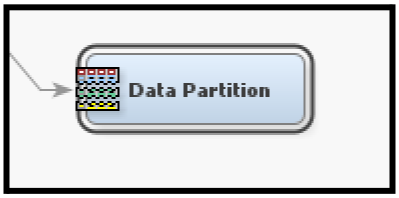


This documentation outlines a comprehensive predictive modeling analysis conducted on a dataset related to hotel bookings. The objective of this analysis is to develop accurate and robust models for predicting booking outcomes. The process encompasses various stages, including data exploration, preprocessing, feature engineering, and the evaluation of multiple machine learning algorithms.

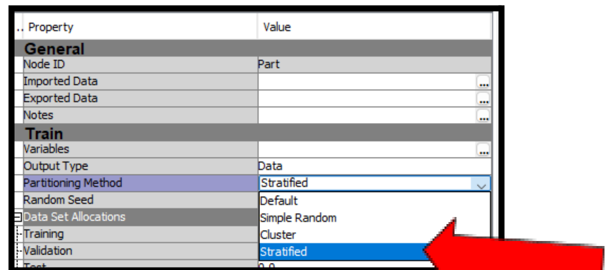
4.1 Sample

This step entails choosing a subset of the appropriate volume dataset from a vast dataset that has been given for the model's construction. The goal of this initial stage of the process is to identify variables or factors (both dependent and independent) influencing the process. The collected information is then sorted into preparation and validation categories.

Node Name :	Data Partition	**Configuration**



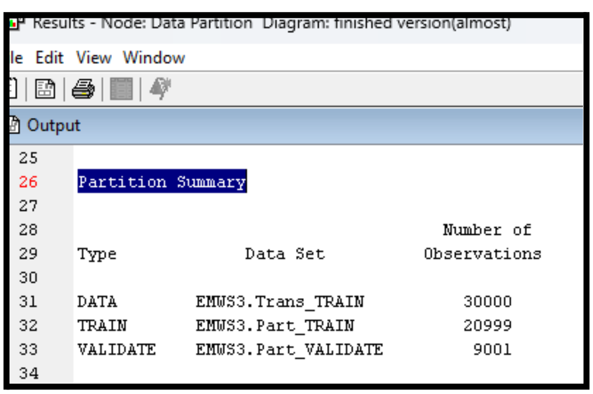
Data Partition nodes are used to generate a partition field that splits the data into separate subsets or samples for the training, testing, and validation stages of model building. By using one sample to generate the model and a separate sample to test it, we can get a good indication of how well the model will generalize to larger datasets that are similar to the current data.



We chose Stratified as our partitioning method.

For the dataset allocation we allocated 70% for training and 30% for validation.

Result

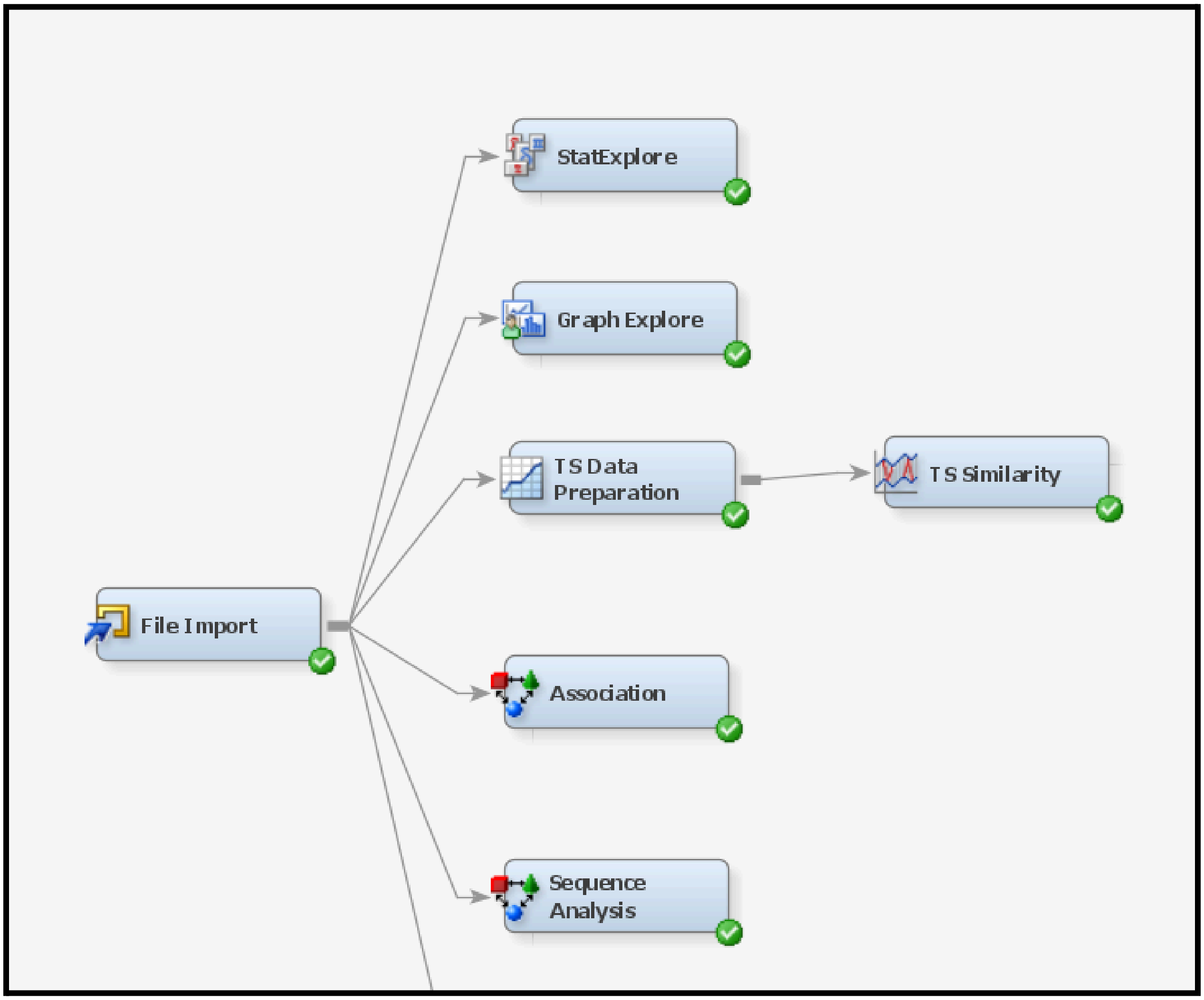


These findings indicate the successful division of the initial dataset into distinct training and validation sets, facilitating the training and evaluation of predictive models. The training set, containing 20,999 observations, is used to train the models, while the validation set, with 9,001 observations, is employed to assess their performance on unseen data.

4.2 Explore

The xxploration stage allowed us to look into the data by searching for relationships, trends, and anomalies to gain understanding and ideas. There were multiple nodes used for multiple exploratory analyses as in the diagram below.

1. Explotoray Data Analysis
2. Time Series Analysis
3. Association & Sequence Analysis
4. DBSCAN Clustering(Knime)

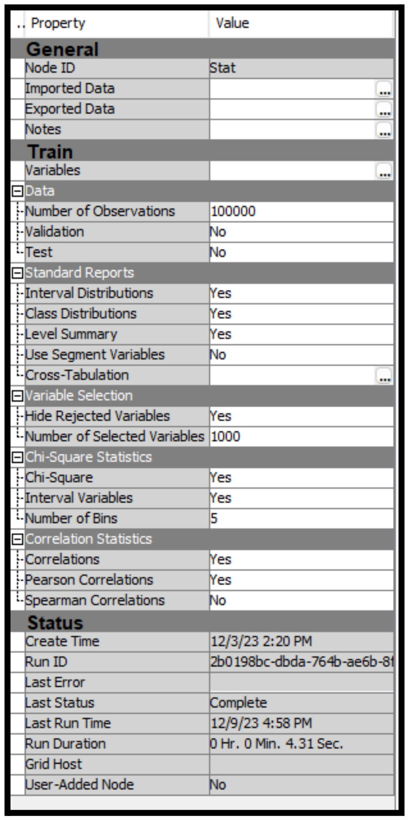


4.2.1 Exploratory Data Analysis

Exploratory Data Analysis

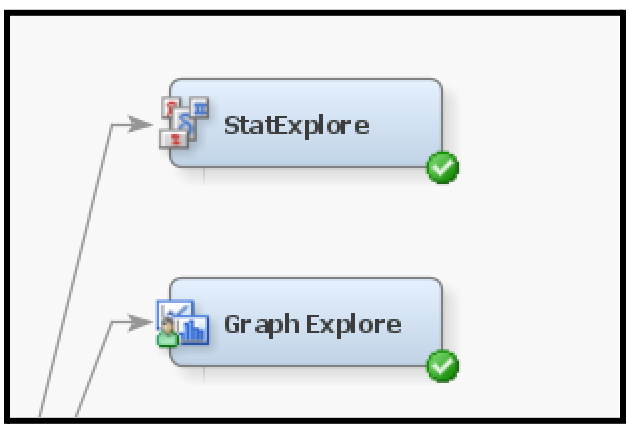
Node Name :	StatExplore	**Configuration**

StatExplore node is a multipurpose tool that can be used to examine variable distributions and statistics in the data sets. The StatExplore node generates summarization statistics.



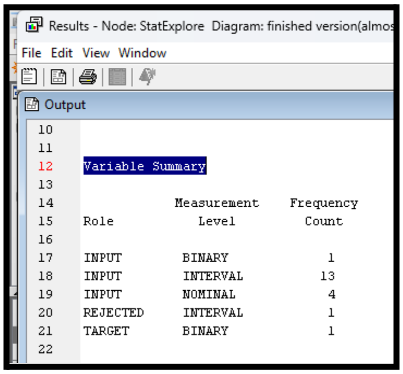
Property	Value
General	
Node ID	Stat
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Data	
Number of Observations	100000
Validation	No
Test	No
Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	...
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variables	1000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	Yes
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No
Status	
Create Time	12/3/23 2:20 PM
Run ID	2b0198bc-dbda-764b-ae6b-8f
Last Error	
Last Status	Complete
Last Run Time	12/9/23 4:58 PM
Run Duration	0 Hr. 0 Min. 4.31 Sec.
Grid Host	
User-Added Node	No

No specific configuration was sele



cted.

Result



Results - Node: StatExplore Diagram: finished version(almos

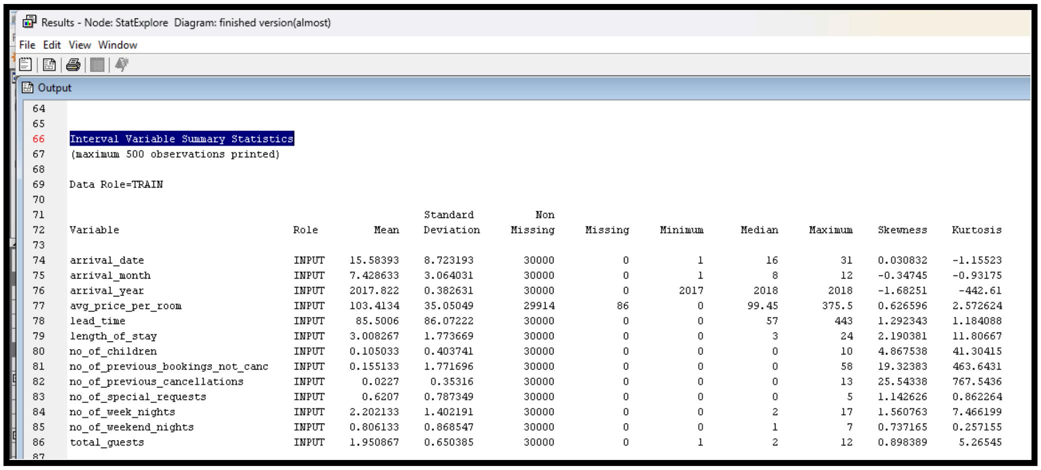
File Edit View Window

Output

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	BINARY	1
INPUT	INTERVAL	13
INPUT	NOMINAL	4
REJECTED	INTERVAL	1
TARGET	BINARY	1

In this output we can see the frequency counts of each field type that we will be dealing with further. There are 4 nominal values and 13 interval ones.



Results - Node: StatExplore Diagram: finished version(almost)

File Edit View Window

Output

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
arrival_date	INPUT	15.58393	8.723193	30000	0	1	16	31	0.030832	-1.15523
arrival_month	INPUT	7.428633	3.064031	30000	0	1	8	12	-0.34745	-0.93175
arrival_year	INPUT	2017.822	0.382631	30000	0	2017	2018	2018	-1.68251	-442.61
avg_price_per_room	INPUT	103.4134	35.05049	29914	86	0	99.45	375.5	0.626596	2.572624
lead_time	INPUT	85.5006	86.07222	30000	0	0	57	443	1.292343	1.184088
length_of_stay	INPUT	3.008267	1.773669	30000	0	0	3	24	2.190381	11.80667
no_of_children	INPUT	0.105033	0.403741	30000	0	0	0	10	4.867538	41.30415
no_of_previous_bookings_not_canc	INPUT	0.155133	1.771696	30000	0	0	0	58	19.32383	463.6431
no_of_previous_cancellations	INPUT	0.0227	0.35316	30000	0	0	0	13	25.54338	767.5436
no_of_special_requests	INPUT	0.6207	0.787349	30000	0	0	0	5	1.142626	0.862264
no_of_week_nights	INPUT	2.202133	1.402191	30000	0	0	2	17	1.560763	7.466199
no_of_weekend_nights	INPUT	0.806133	0.868547	30000	0	0	1	7	0.737165	0.257155
total_guests	INPUT	1.950867	0.650385	30000	0	1	2	12	0.898389	5.26545

Next as a result of running this node, it also presents summary statistics for all the interval variables as above.

Key findings are as below :

- **Skewness and Kurtosis:**

no_of_children	INPUT	0.105033	0.403741	30000	0	0	0	10	4.867538	41.30415
no_of_previous_bookings_not_canc	INPUT	0.155133	1.771696	30000	0	0	0	58	19.32383	463.6431
no_of_previous_cancellations	INPUT	0.0227	0.35316	30000	0	0	0	13	25.54338	767.5436

  - Above 3 variables have abnormally high skewness and kurtosis.
  - The variables are :
    - no_of_children
    - no_of_previous_bookings_not_canc
    - no_of_previous_cancellations
  - These problem will be treated in the later stage.
- **Presence of Missing Values:**

Variable	Role	Mean	Standard Deviation	Non Missing	Missing
arrival_date	INPUT	15.58393	8.723193	30000	0
arrival_month	INPUT	7.428633	3.064031	30000	0
arrival_year	INPUT	2017.822	0.382631	30000	0
avg_price_per_room					86

- As indicated in the picture above, there are 86 missing values in the avg_price_per_room variable. This will be then imputed in the later stage.

Results - Node: StatExplore Diagram: finished version(almost)

File Edit View Window

Output

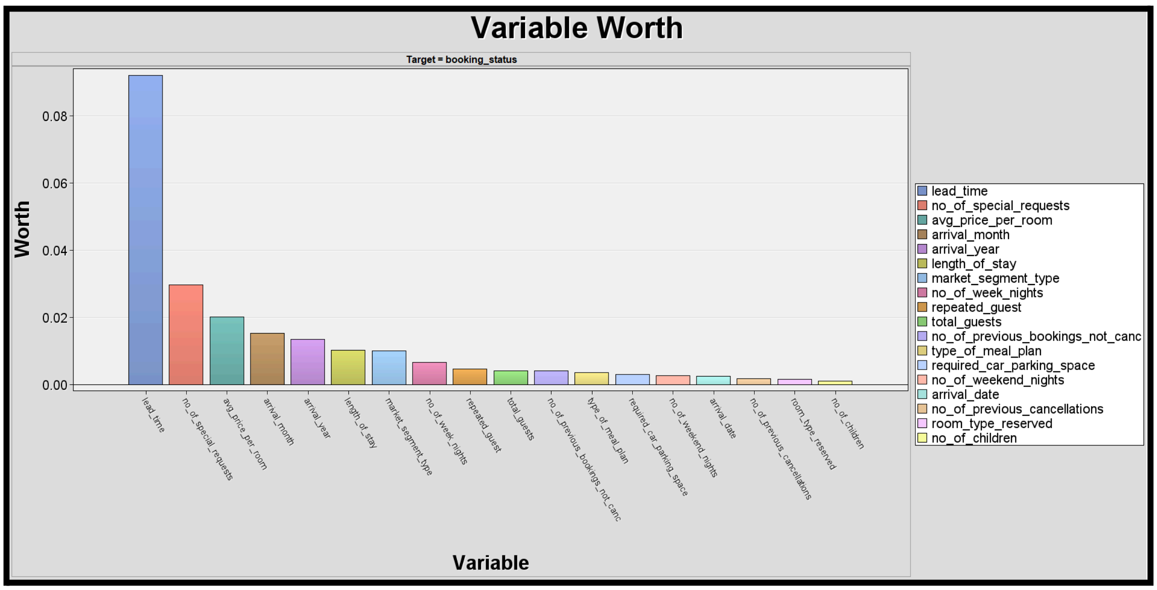
```
booking_status       1              2           0          9815



Chi-Square Statistics
(maximum 500 observations printed)

Data Role=TRAIN Target=booking_status

Input                                Chi-Square    Df      Prob

lead_time                             5502.5163     4    <.0001
no_of_special_requests                 995.0204     4    <.0001
arrival_year                           934.2196     1    <.0001
market_segment_type                    715.4852     4    <.0001
arrival_month                          545.4652     4    <.0001
avg_price_per_room                     534.2270     5    <.0001
repeated_guest                         339.8776     1    <.0001
required_car_parking_space             222.7899     1    <.0001
type_of_meal_plan                      219.1374     3    <.0001
no_of_week_nights                      184.8766     4    <.0001
length_of_stay                         154.2180     4    <.0001
no_of_weekend_nights                   121.9469     4    <.0001
no_of_previous_bookings_not_canc        51.2550     4    <.0001
room_type_reserved                      43.9391     6    <.0001
total_guests                            38.2518     2    <.0001
no_of_previous_cancellations            27.1623     3    <.0001
arrival_date                            10.8054     4    0.0288
no_of_children                           0.4336     2    0.8051
```

The Chi-Square statistics table provides valuable insights into the relationship between various features and the target variable (booking_status). Here are key insights and findings based on the provided Chi-Square statistics:

- **Highly Significant Predictors:**

- The features "lead_time," "no_of_special_requests," "arrival_year," "market_segment_type," and "arrival_month" have exceptionally high Chi-Square values, indicating a strong association with booking_status. The p-values for these features are all below 0.0001, suggesting a very low probability that the observed associations are due to chance.

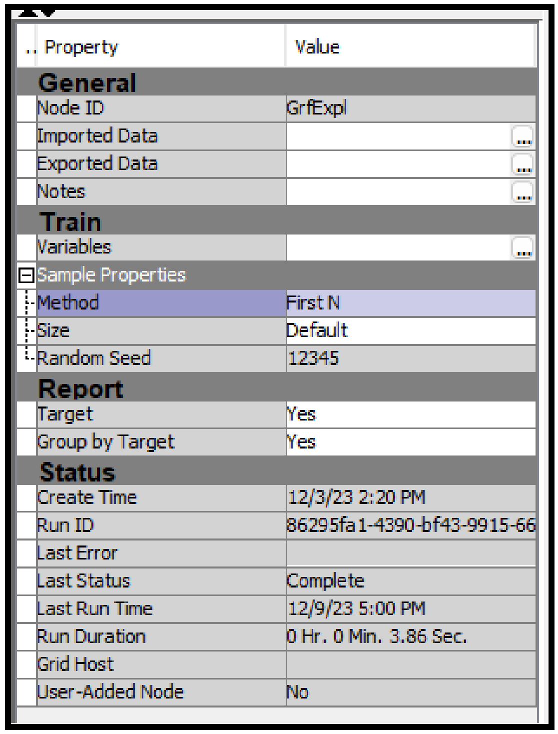
- **Insignificant Predictor:**
    - "no_of_children" has a Chi-Square value of 0.4336 and a p-value of 0.8051, indicating that it is not a significant predictor of booking_status. The high p-value suggests that the observed association may be due to chance.



The node also variable worthiness graph as above indicates the best variable for the prediction problem. As we can see the lead_time shows the highest worthiness while the no_of_children shows the least.

Node Name:	GraphExplore	**Configuration**

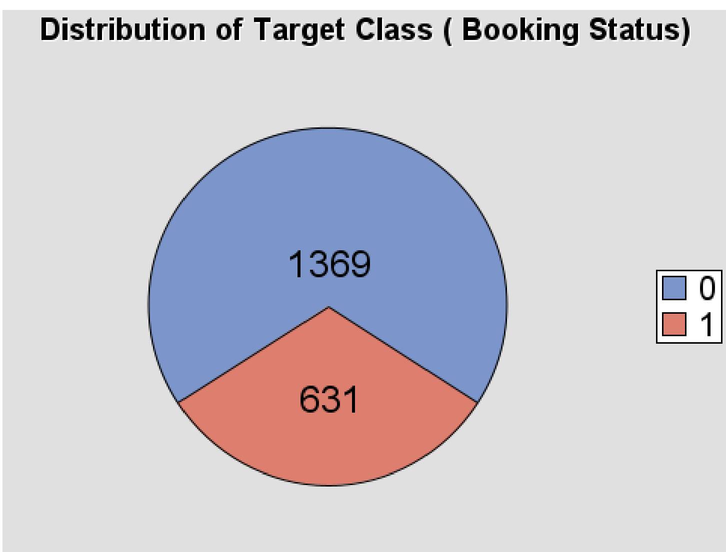
The Graph Explore node is on the Explore tab of the Enterprise Miner tools bar. The Graph Explore node is an advanced visualization tool that enables us to explore large volumes of data graphically to uncover patterns and trends and reveal extreme values in the database.



Property	Value
General	
Node ID	GrfExpl
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Sample Properties	
Method	First N
Size	Default
Random Seed	12345
Report	
Target	Yes
Group by Target	Yes
Status	
Create Time	12/3/23 2:20 PM
Run ID	86295fa1-4390-bf43-9915-66
Last Error	
Last Status	Complete
Last Run Time	12/9/23 5:00 PM
Run Duration	0 Hr. 0 Min. 3.86 Sec.
Grid Host	
User-Added Node	No

No specific configuration was selected.

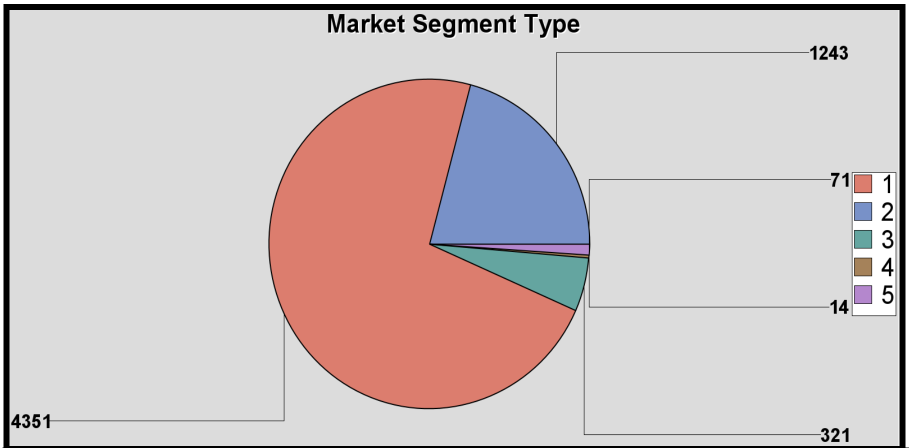
Result & Analysis



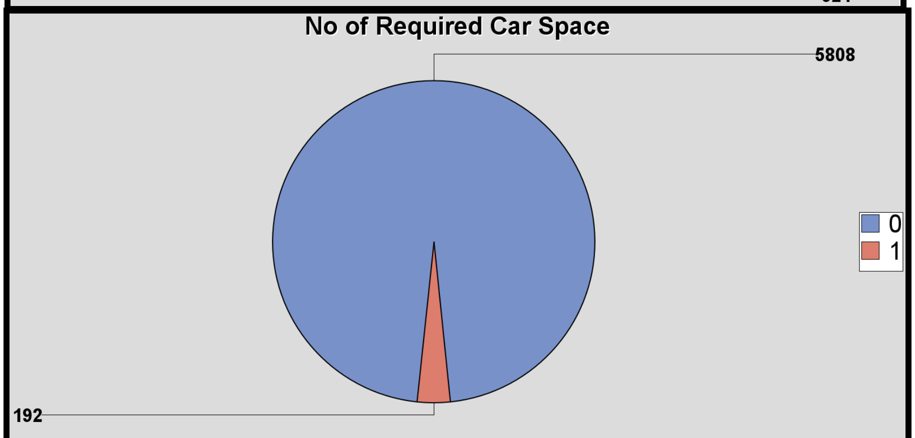
This diagram represents the distribution of booking status values. “1” refers to canceled, while “0” is not canceled. The proportion of distribution is close to 2:1 as shown in a graph.

As we can see, the class is imbalanced thus, a class balancing technique needs to be done and it is addressed here

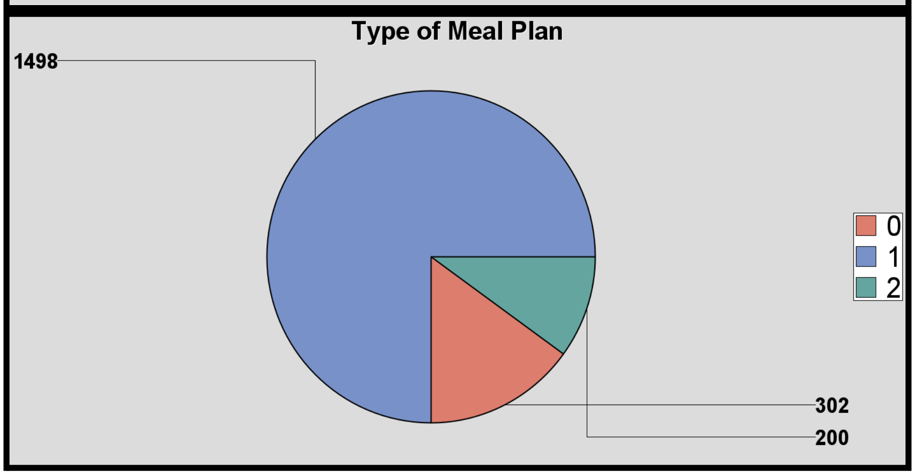
Each of these graphs represents distributions in terms of frequency for fields in the dataset.



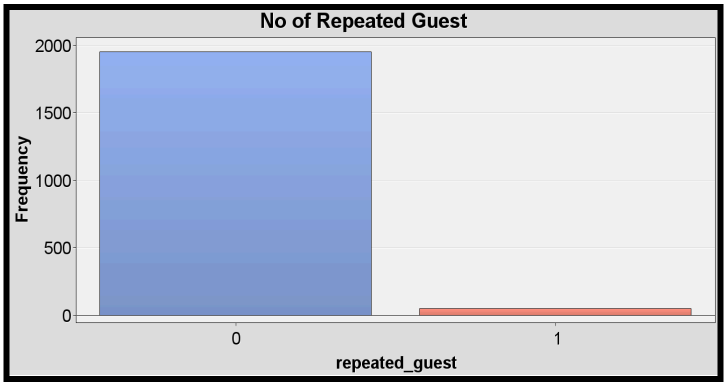
In this graph "Market segment type" has five values with each respective frequency of their appearance in the dataset.

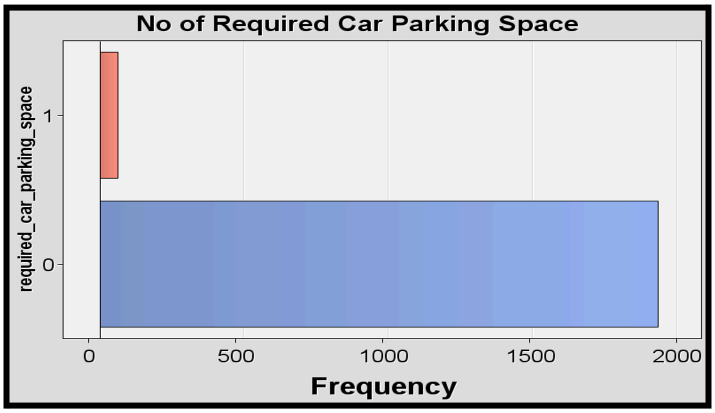


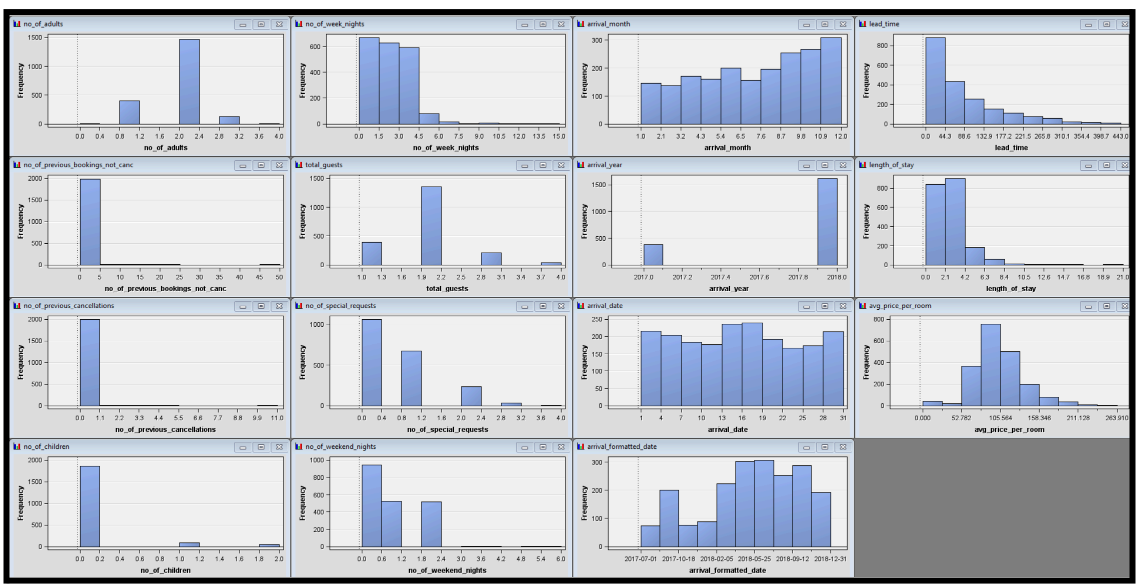
In this graph "Number of required car space" has two values with each respective frequency of their appearance in the dataset.



The graph above represents a distribution of meal types in a dataset among data records. There are 3 data values that are overwhelming whole dataset.



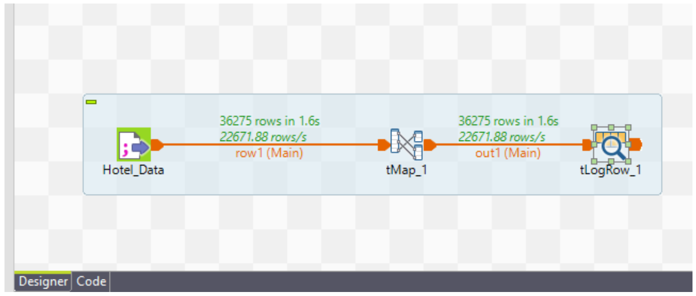


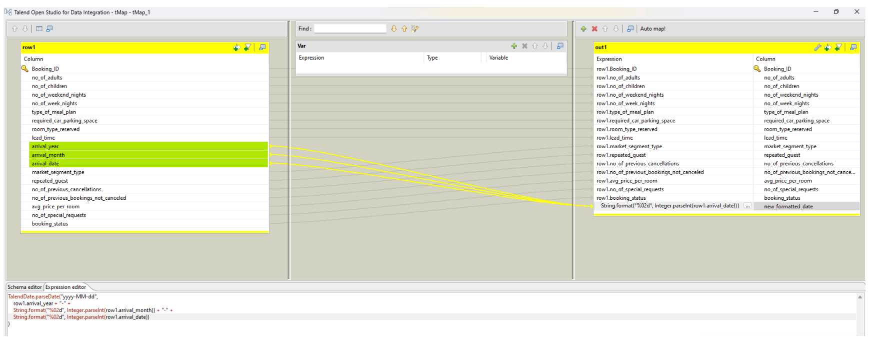


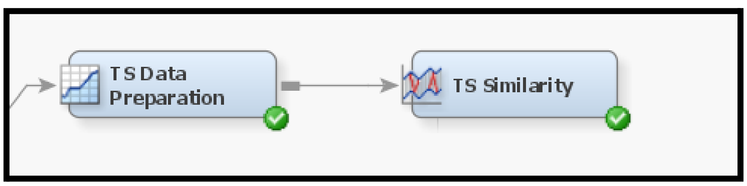
Above is the histogram graph on the distribution of continuous variables. Again we can witness how some variables are highly skewed.

4.2.2 Time series Analysis

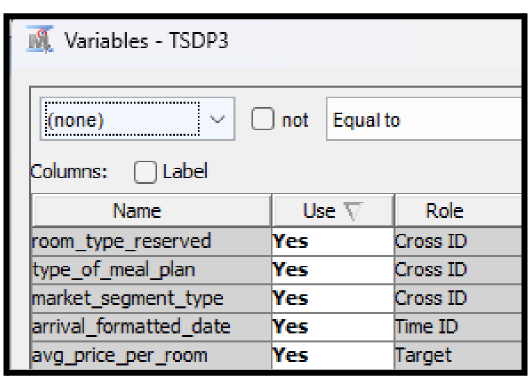
To prepare our data for Time Series analysis some preparation needs to be done. We used Talend Open Studio for Data Integration as our tool. Below are the screenshots of the work done on the platform.



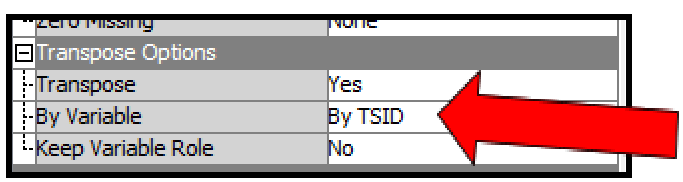




In Time Series Similarity, the **TS Data Preparation** was configured to set Transpose by variable **"By TSID"** as shown in the picture below and then the variables selected for the analysis are shown in the right picture.


Variables - TSDP3

Name	Use	Role
room_type_reserved	Yes	Cross ID
type_of_meal_plan	Yes	Cross ID
market_segment_type	Yes	Cross ID
arrival_formatted_date	Yes	Time ID
avg_price_per_room	Yes	Target



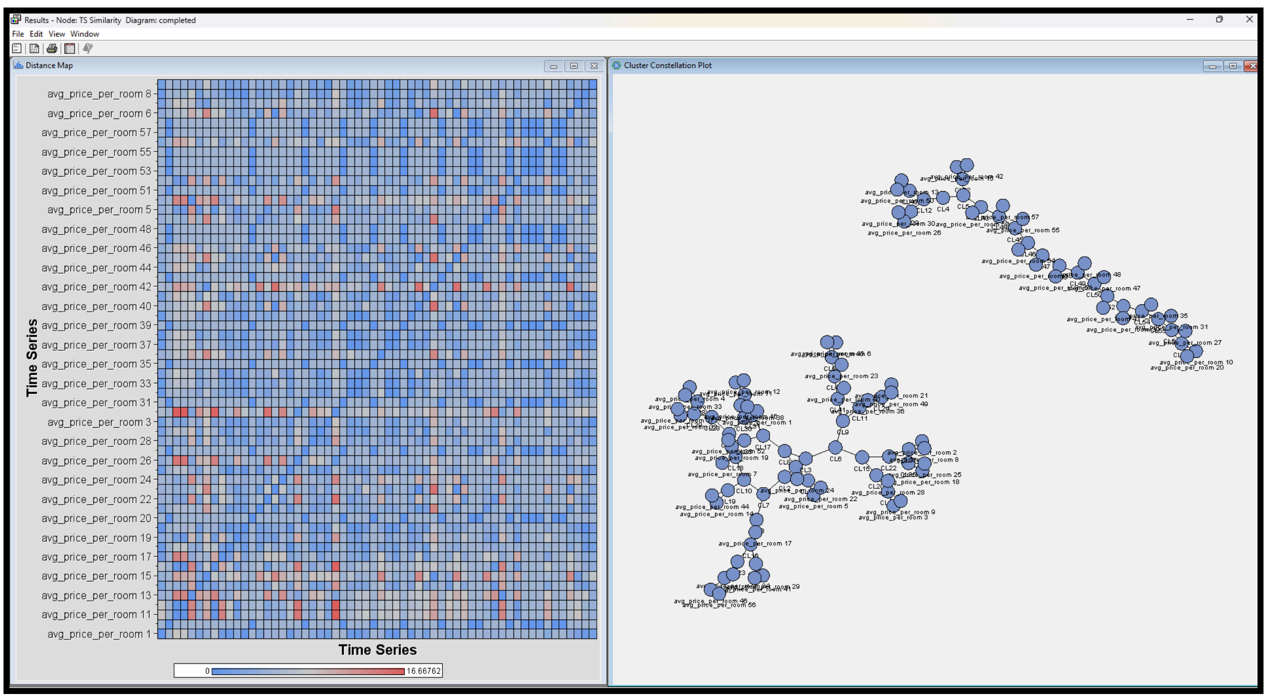
Transpose Options	
Transpose	Yes
By Variable	By TSID
Keep Variable Role	No

After running the TS Data Preparation node, it gave the following results. This result shows that there are 87 time series were created.

TSID Map Summary Table

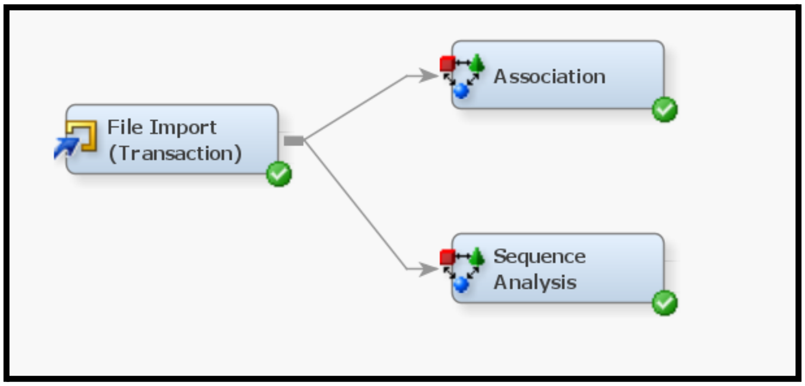
Name	Level	Count	Percent
booking_status	0	55	63.21839
booking_status	1	32	36.78161
market_segment_type	1	28	32.18391
market_segment_type	3	12	13.7931
market_segment_type	4	4	4.597701
market_segment_type	5	16	18.3908
market_segment_type	2	27	31.03448
room_type_reserved	7	9	10.34483
room_type_reserved	6	12	13.7931
room_type_reserved	5	12	13.7931
room_type_reserved	4	18	20.68966
room_type_reserved	3	3	3.448276
room_type_reserved	2	10	11.49425
room_type_reserved	1	23	26.43678
type_of_meal_plan	0	18	20.68966
type_of_meal_plan	3	3	3.448276
type_of_meal_plan	2	21	24.13793
type_of_meal_plan	1	45	51.72414
TSID		87	100

Then the TS Similairty nodes were executed and yielded the results below.

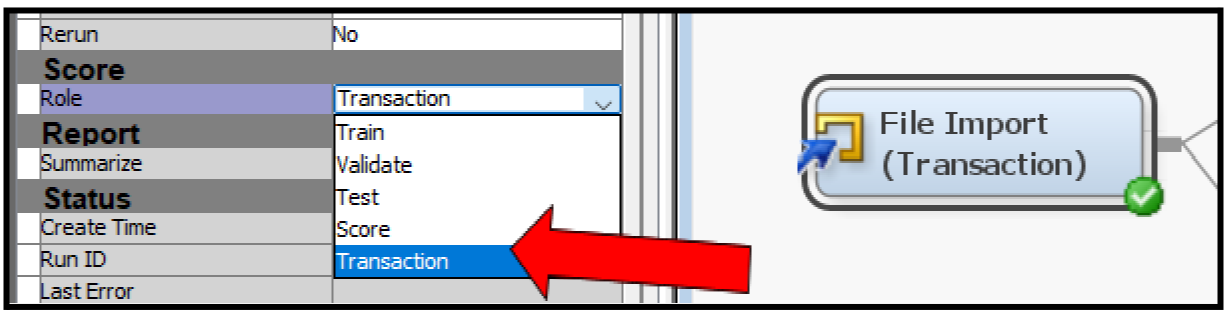


On the left, it shows the distance map. The darker the blue the closer the distance is and the red is the exact opposite of it. On the right are the formed clusters showing that the node is able to successfully cluster the timer series.

4.2.3 Association and Sequence Analysis



The above diagram shows the flow and nodes used for the Association and Sequence Analysis. As for this analysis, we needed to use another import node as this time the data plays a different role.

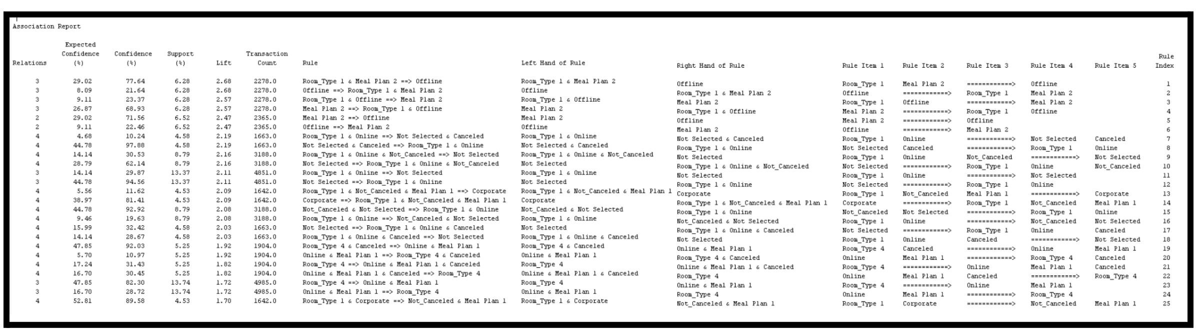


Thus to prepare for the analysis, in the File Import node, we have to set the role to "Transaction" as seen in the image above.

4.2.3.1 Association Analysis

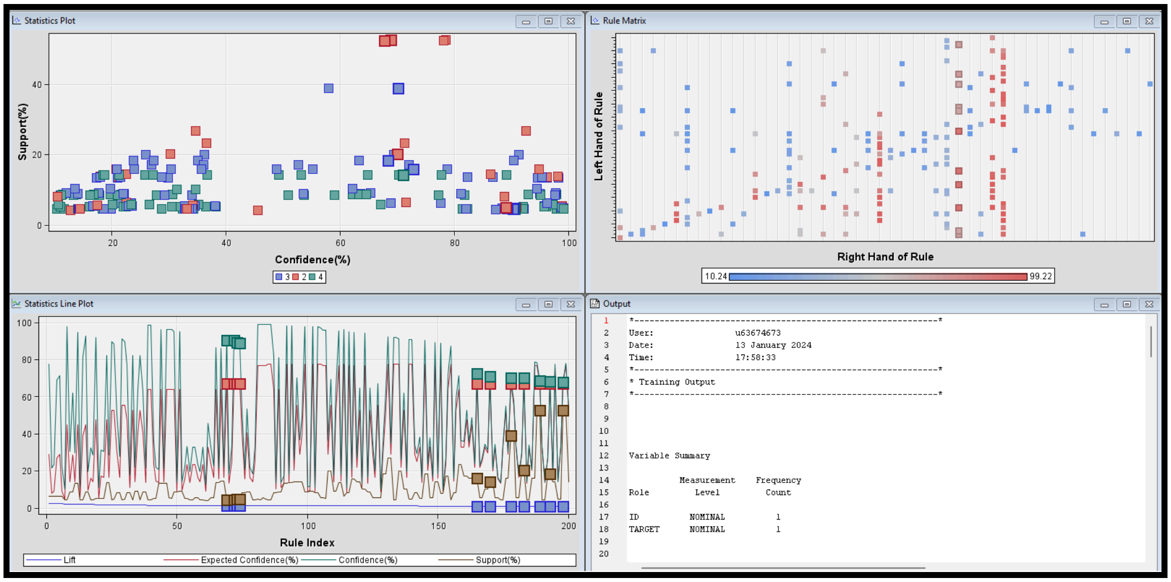
Firstly, we performed the Association Analysis. We set the roles for each variable as seen in the image above.

With that, the Association analysis was executed and yielded the following rules.



Association Report

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule	Left Hand of Rule	Right Hand of Rule	Rule Item 1	Rule Item 2	Rule Item 3	Rule Item 4	Rule Item 5	Rule Index
3	29.02	77.64	6.28	2.68	2278.0	Room_Type 1 & Meal Plan 2 ==> Offline	Room_Type 1 & Meal Plan 2	Offline	Room_Type 1	Meal Plan 2	============>	Offline		1
3	8.09	21.64	6.28	2.68	2278.0	Offline ==> Room_Type 1 & Meal Plan 2	Offline	Room_Type 1 & Meal Plan 2	Offline	============>	Room_Type 1	Meal Plan 2		2
3	9.11	23.37	6.28	2.57	2278.0	Room_Type 1 & Offline ==> Meal Plan 2	Room_Type 1 & Offline	Meal Plan 2	Room_Type 1	Offline	============>	Meal Plan 2		3
3	26.87	68.93	6.28	2.57	2278.0	Meal Plan 2 ==> Room_Type 1 & Offline	Meal Plan 2	Room_Type 1 & Offline	Meal Plan 2	============>	Room_Type 1	Offline		4
2	29.02	71.56	6.52	2.47	2365.0	Meal Plan 2 ==> Offline	Meal Plan 2	Offline	Meal Plan 2	============>	Offline			5
2	9.11	22.46	6.52	2.47	2365.0	Offline ==> Meal Plan 2	Offline	Meal Plan 2	Offline	============>	Meal Plan 2			6
4	4.68	10.24	4.58	2.19	1663.0	Room_Type 1 & Online ==> Not Selected & Canceled	Room_Type 1 & Online	Not Selected & Canceled	Room_Type 1	Online	============>	Not Selected	Canceled	7
4	44.78	97.88	4.58	2.19	1663.0	Not Selected & Canceled ==> Room_Type 1 & Online	Not Selected & Canceled	Room_Type 1 & Online	Not Selected	Canceled	============>	Room_Type 1	Online	8
4	14.14	30.53	8.79	2.16	3188.0	Room_Type 1 & Online & Not_Canceled ==> Not Selected	Room_Type 1 & Online & Not_Canceled	Not Selected	Room_Type 1	Online	Not_Canceled	============>	Not Selected	9
4	28.79	62.14	8.79	2.16	3188.0	Not Selected ==> Room_Type 1 & Online & Not_Canceled	Not Selected	Room_Type 1 & Online & Not_Canceled	Not Selected	============>	Room_Type 1	Online	Not_Canceled	10
3	14.14	29.87	13.37	2.11	4851.0	Room_Type 1 & Online ==> Not Selected	Room_Type 1 & Online	Not Selected	Room_Type 1	Online	============>	Not Selected		11
3	44.78	94.56	13.37	2.11	4851.0	Not Selected ==> Room_Type 1 & Online	Not Selected	Room_Type 1 & Online	Not Selected	============>	Room_Type 1	Online		12
4	5.56	11.62	4.53	2.09	1642.0	Room_Type 1 & Not_Canceled & Meal Plan 1 ==> Corporate	Room_Type 1 & Not_Canceled & Meal Plan 1	Corporate	Room_Type 1	Not_Canceled	Meal Plan 1	============>	Corporate	13
4	38.97	81.41	4.53	2.09	1642.0	Corporate ==> Room_Type 1 & Not_Canceled & Meal Plan 1	Corporate	Room_Type 1 & Not_Canceled & Meal Plan 1	Corporate	============>	Room_Type 1	Not_Canceled	Meal Plan 1	14
4	44.78	92.92	8.79	2.08	3188.0	Not_Canceled & Not Selected ==> Room_Type 1 & Online	Not_Canceled & Not Selected	Room_Type 1 & Online	Not_Canceled	Not Selected	============>	Room_Type 1	Online	15
4	9.46	19.63	8.79	2.08	3188.0	Room_Type 1 & Online ==> Not_Canceled & Not Selected	Room_Type 1 & Online	Not_Canceled & Not Selected	Room_Type 1	Online	============>	Not_Canceled	Not Selected	16
4	15.99	32.42	4.58	2.03	1663.0	Not Selected ==> Room_Type 1 & Online & Canceled	Not Selected	Room_Type 1 & Online & Canceled	Not Selected	============>	Room_Type 1	Online	Canceled	17
4	14.14	28.67	4.58	2.03	1663.0	Room_Type 1 & Online & Canceled ==> Not Selected	Room_Type 1 & Online & Canceled	Not Selected	Room_Type 1	Online	Canceled	============>	Not Selected	18
4	47.85	92.03	5.25	1.92	1904.0	Room_Type 4 & Canceled ==> Online & Meal Plan 1	Room_Type 4 & Canceled	Online & Meal Plan 1	Room_Type 4	Canceled	============>	Online	Meal Plan 1	19
4	5.70	10.97	5.25	1.92	1904.0	Online & Meal Plan 1 ==> Room_Type 4 & Canceled	Online & Meal Plan 1	Room_Type 4 & Canceled	Online	Meal Plan 1	============>	Room_Type 4	Canceled	20
4	17.24	31.43	5.25	1.82	1904.0	Room_Type 4 ==> Online & Meal Plan 1 & Canceled	Room_Type 4	Online & Meal Plan 1 & Canceled	Room_Type 4	============>	Online	Meal Plan 1	Canceled	21
4	16.70	30.45	5.25	1.82	1904.0	Online & Meal Plan 1 & Canceled ==> Room_Type 4	Online & Meal Plan 1 & Canceled	Room_Type 4	Online	Meal Plan 1	Canceled	============>	Room_Type 4	22
3	47.85	82.30	13.74	1.72	4985.0	Room_Type 4 ==> Online & Meal Plan 1	Room_Type 4	Online & Meal Plan 1	Room_Type 4	============>	Online	Meal Plan 1		23
3	16.70	28.72	13.74	1.72	4985.0	Online & Meal Plan 1 ==> Room_Type 4	Online & Meal Plan 1	Room_Type 4	Online	Meal Plan 1	============>	Room_Type 4		24
4	52.81	89.58	4.53	1.70	1642.0	Room_Type 1 & Corporate ==> Not_Canceled & Meal Plan 1	Room_Type 1 & Corporate	Not_Canceled & Meal Plan 1	Room_Type 1	Corporate	============>	Not_Canceled	Meal Plan 1	25



We filtered our rules to focus on where the right-hand rule was Canceled and Not Canceled to understand the patterns of when the booking is being canceled and not canceled.

Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Rule
2	32.76	45.57	4.15	1.39	Meal Plan 2 ==> Canceled
4	32.76	38.19	5.25	1.17	Room_Type 4 & Online & Meal Plan 1 ==> Canceled
3	32.76	38.05	5.48	1.16	Room_Type 4 & Online ==> Canceled
2	32.76	36.51	23.36	1.11	Online ==> Canceled
3	32.76	36.03	17.24	1.10	Online & Meal Plan 1 ==> Canceled
3	32.76	35.71	15.99	1.09	Room_Type 1 & Online ==> Canceled
4	32.76	34.61	10.25	1.06	Room_Type 1 & Online & Meal Plan 1 ==> Canceled
3	32.76	34.46	5.46	1.05	Room_Type 4 & Meal Plan 1 ==> Canceled
4	32.76	34.28	4.58	1.05	Room_Type 1 & Online & Not Selected ==> Canceled
2	32.76	34.16	5.70	1.04	Room_Type 4 ==> Canceled
3	32.76	33.98	4.62	1.04	Online & Not Selected ==> Canceled
3	32.76	33.45	4.65	1.02	Room_Type 1 & Not Selected ==> Canceled
2	32.76	33.12	4.68	1.01	Not Selected ==> Canceled

From the picture above, we can see that expected confidence, those who selected Meal Plan 2, Room Type 4, Online will cancel their booking with a confidence level of 32.76% while the highest confidence level is when customers select Meal Plan 2 at 45.57%

Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Right Hand of Rule ▲
4	67.24	90.62	4.53	1.35	1642.0	Room_Type 1 & Meal Plan 1 & Corporate ==> Not_Canceled	Not_Canceled
3	67.24	90.34	4.57	1.34	1656.0	Room_Type 1 & Corporate ==> Not_Canceled	Not_Canceled
3	67.24	89.33	4.92	1.33	1783.0	Meal Plan 1 & Corporate ==> Not_Canceled	Not_Canceled
2	67.24	89.09	4.95	1.33	1797.0	Corporate ==> Not_Canceled	Not_Canceled
3	67.24	72.76	16.02	1.08	5812.0	Offline & Meal Plan 1 ==> Not_Canceled	Not_Canceled
4	67.24	71.01	14.29	1.06	5182.0	Room_Type 1 & Offline & Meal Plan 1 ==> Not_Canceled	Not_Canceled
3	67.24	70.13	38.97	1.04	14136	Room_Type 1 & Meal Plan 1 ==> Not_Canceled	Not_Canceled
2	67.24	70.05	20.33	1.04	7375.0	Offline ==> Not_Canceled	Not_Canceled
2	67.24	68.82	52.81	1.02	19156	Meal Plan 1 ==> Not_Canceled	Not_Canceled
3	67.24	68.40	18.38	1.02	6667.0	Room_Type 1 & Offline ==> Not_Canceled	Not_Canceled
2	67.24	67.75	52.54	1.01	19058	Room_Type 1 ==> Not_Canceled	Not_Canceled

For Not Cancel booking, customers who choose Room Type 1, Meal Plan 1, Corporate, or Offline group will not cancel their booking with an expected confidence level of 67.24% while the highest confidence level is when customers choose Room Type 1, Meal Plan 1 and Corporate at 90.62%.

4.2.3.2 Sequence Analysis

Secondly, we performed the Sequence analysis, by choosing the same Association node but was renamed to Sequence Analysis. Then in the configuration panel, we had to configure the rules where the Export Rule by ID is set to “Yes” as on the left picture and on the right are the configurations for the Sequence panel for the analysis.

Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes

Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction	0.0
Support Type	Percent
Support Count	1
Support Percentage	2.0

Then the sequence analysis node was executed and yielded the results below :

Rule Description

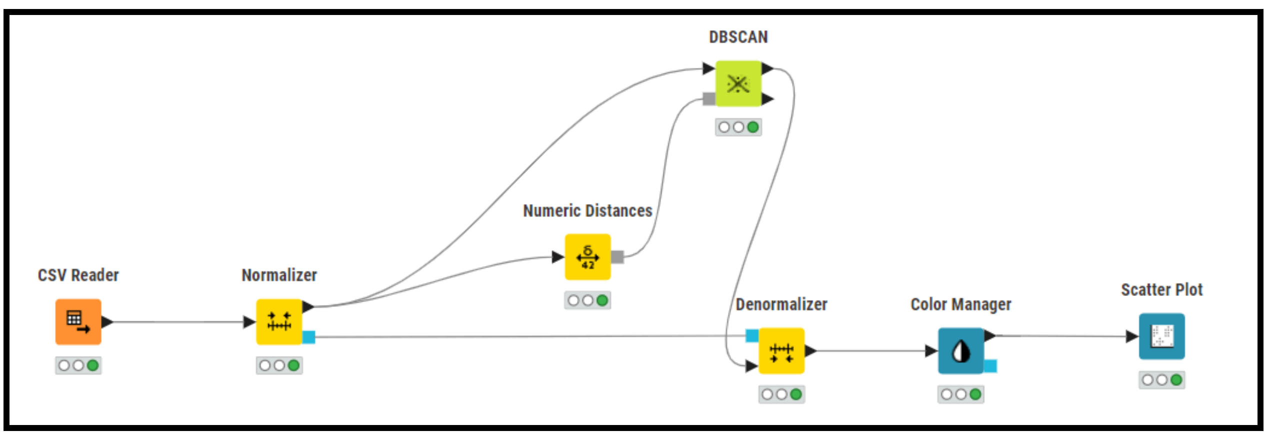
Map	Rule
RULE1	Room_Type 1 & Meal Plan 2 ==> Offline
RULE2	Offline ==> Room_Type 1 & Meal Plan 2
RULE3	Room_Type 1 & Offline ==> Meal Plan 2
RULE4	Meal Plan 2 ==> Room_Type 1 & Offline
RULE5	Meal Plan 2 ==> Offline
RULE6	Offline ==> Meal Plan 2
RULE7	Room_Type 1 & Online ==> Not Selected & Canceled
RULE8	Not Selected & Canceled ==> Room_Type 1 & Online
RULE9	Room_Type 1 & Online & Not_Canceled ==> Not Selected
RULE10	Not Selected ==> Room_Type 1 & Online & Not_Canceled
RULE11	Room_Type 1 & Online ==> Not Selected
RULE12	Not Selected ==> Room_Type 1 & Online
RULE13	Room_Type 1 & Not_Canceled & Meal Plan 1 ==> Corporate
RULE14	Corporate ==> Room_Type 1 & Not_Canceled & Meal Plan 1
RULE15	Not_Canceled & Not Selected ==> Room_Type 1 & Online
RULE16	Room_Type 1 & Online ==> Not_Canceled & Not Selected
RULE17	Not Selected ==> Room_Type 1 & Online & Canceled
RULE18	Room_Type 1 & Online & Canceled ==> Not Selected
RULE19	Room_Type 4 & Canceled ==> Online & Meal Plan 1
RULE20	Online & Meal Plan 1 ==> Room_Type 4 & Canceled
RULE21	Room_Type 4 ==> Online & Meal Plan 1 & Canceled
RULE22	Online & Meal Plan 1 & Canceled ==> Room_Type 4
RULE23	Room_Type 4 ==> Online & Meal Plan 1
RULE24	Online & Meal Plan 1 ==> Room_Type 4
RULE25	Room_Type 1 & Corporate ==> Not_Canceled & Meal Plan 1
RULE26	Corporate ==> Not_Canceled & Meal Plan 1
RULE27	Room_Type 4 ==> Online & Not_Canceled & Meal Plan 1
RULE28	Online & Not_Canceled & Meal Plan 1 ==> Room_Type 4
RULE29	Not_Canceled & Corporate ==> Room_Type 1 & Meal Plan 1
RULE30	Corporate ==> Room_Type 1 & Meal Plan 1
RULE31	Room_Type 4 & Not_Canceled ==> Online & Meal Plan 1
RULE32	Online & Meal Plan 1 ==> Room_Type 4 & Not_Canceled
RULE33	Meal Plan 1 & Corporate ==> Room_Type 1 & Not_Canceled

From this we can derive the insights that ,

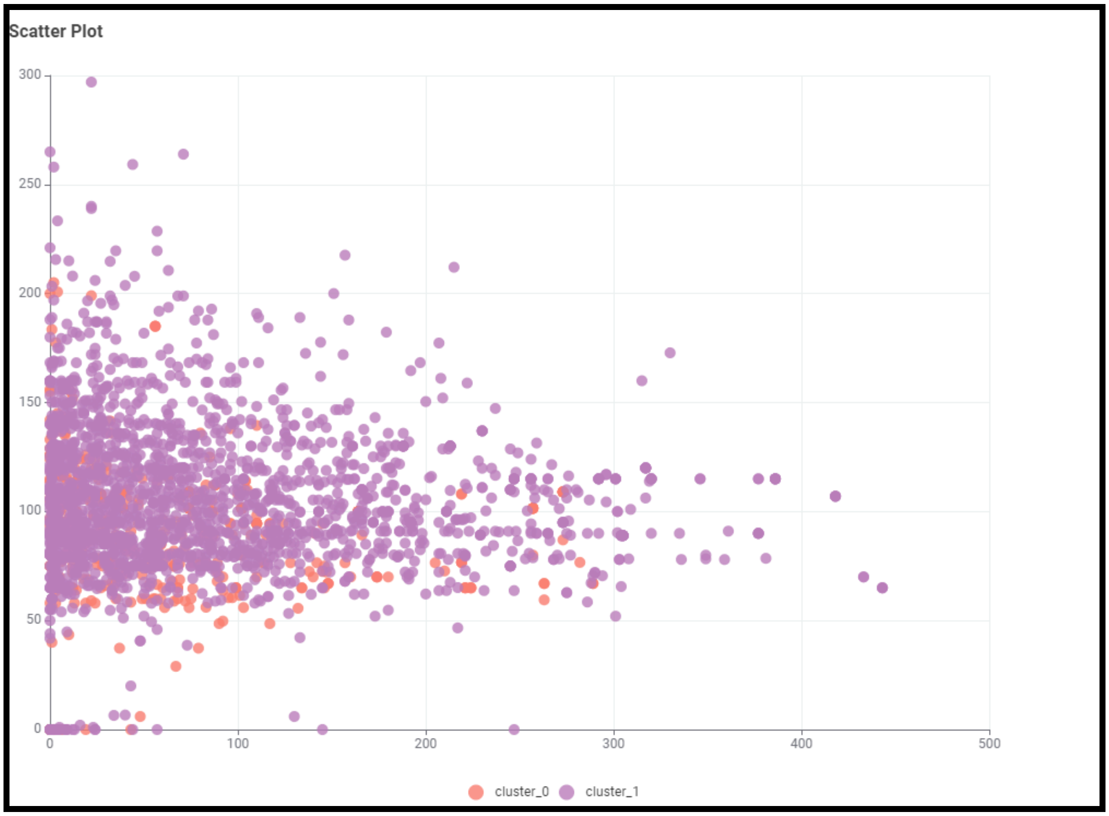
- **Frequent Transitions:**
    - Certain transitions between states are frequent and may represent common patterns in the data.
    - Examples include transitions involving 'Room_Type 1' and 'Meal Plan 2,' both leading to and from the 'Offline' state.
- **Cancellation Patterns:**
    - Several rules involve the 'Canceled' state, indicating specific sequences leading to or from cancellations.
    - For instance, sequences like 'Room_Type 1 & Online ==> Not Selected & Canceled' suggest that certain online bookings for 'Room_Type 1' lead to cancellations.
- **Meal Plan Preferences:**
    - Patterns related to meal plans are evident, such as transitions involving different meal plan types ('Meal Plan 1,' 'Meal Plan 2').
    - For example, 'Room_Type 1 & Not_Canceled & Meal Plan 1 ==> Corporate' suggests that specific bookings with 'Room_Type 1' and 'Meal Plan 1' lead to the 'Corporate' state.

4.2.4 DBSCAN Clustering

To perform the **DBSCAN** clustering, we used Knime as our tool. Below are the flow and nodes used to perform the clustering.



After setting the configuration as above, it yielded the results as below.



From the analysis, the insights discovered that two clusters have been formed which makes sense as we have data on those canceled and not cancelled indicating that there are clear clusters and patterns which we can dive into.

4.3 Modify

In this step, lessons learned in the exploration phase from the data collected in the sample phase are derived with the application of business logic. In other words, the data is parsed and cleaned, then passed onto the modeling stage, and explored if the data requires refinement and transformation.

Overall we performed 7 tasks under this stage. These are

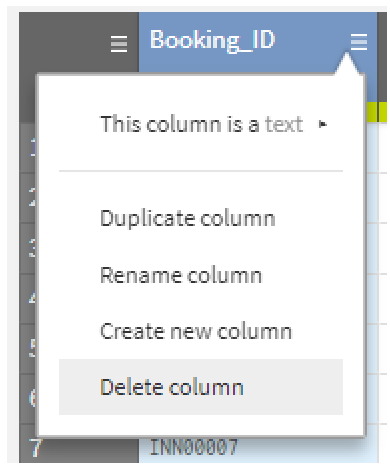
- **Task 1: Remove unwanted variable**
- **Task 2: Encode Categorical Variables**
- **Task 3: Feature Engineering**
- **Task 4: Variable Preparation**
- **Task 5: Impute Missing Values**

- **Task 6: Data Transformation**
- **Task 7: Variable Selection**

All those tasks are explained below.

4.3.1 Task 1: Remove unwanted variable

- One variable which is Booking ID was identified as an irrelevant variable for the prediction problem. Thus it was removed from the data using the Talend Data Preparation tool.



4.3.2 Task 2: Encode Categorical Variables

There were a few variables that were in the text that needed to be converted to numerical format as we will be feeding our data to a machine-learning model. To perform this task, we used the Talend Data Preparation tool.

Variables that were chosen are:

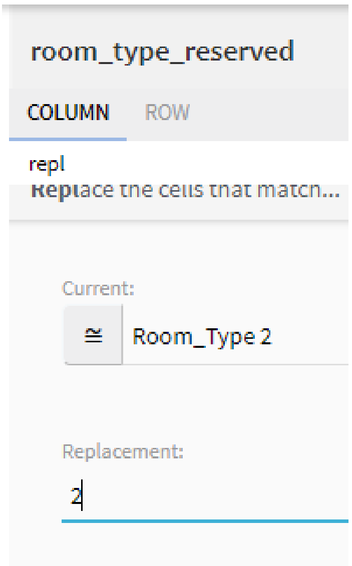
- room_type_reserved
- type_of_meal_plan
- market_segment_type
- booking_status

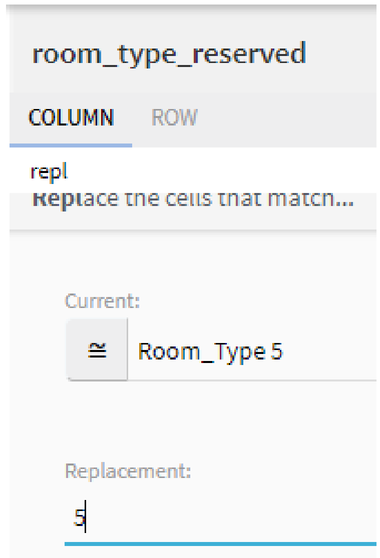
The encoding of variables are as below:

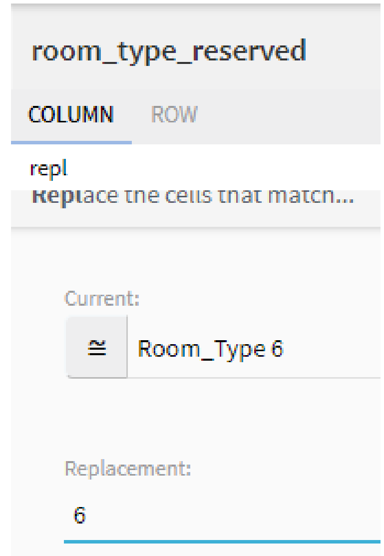
- **room_type_reserved**
  - Room_Type 1 → 1
  - Room_Type 2 → 2
  - Room_Type 3 → 3
  - Room_Type 4 → 4
  - Room_Type 5 → 5
  - Room_Type 6 → 6
- **type_of_meal_plan**
  - Not_selected → 0
  - Meal Plan 1 → 1
  - Meal Plan 2 → 2
  - Meal Plan 3 → 3
- **Market_segment_type**
  - Online → 1
  - Offiline → 2
  - Corporate → 3
  - Aviation → 4
  - Complementary → 5
- **booking_status**
  - Cancelled →1
  - Not_cancelled → 0

The tables below are some screenshots attached for the variables above as an evidence of the task.

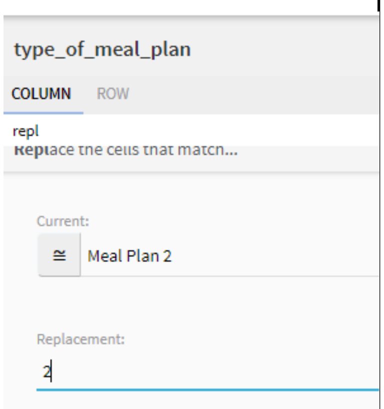
room_type_reserved

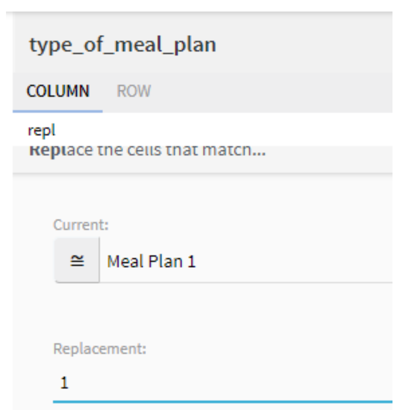


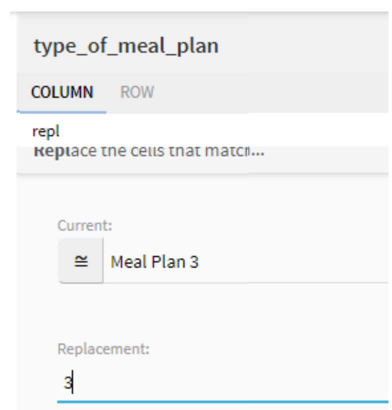




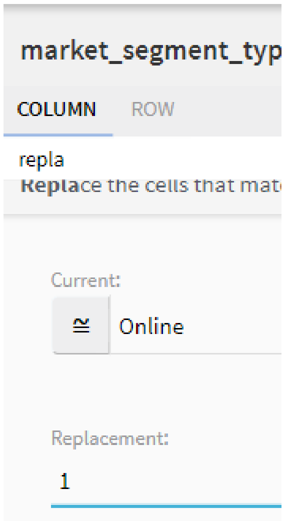
type_of_meal_plan

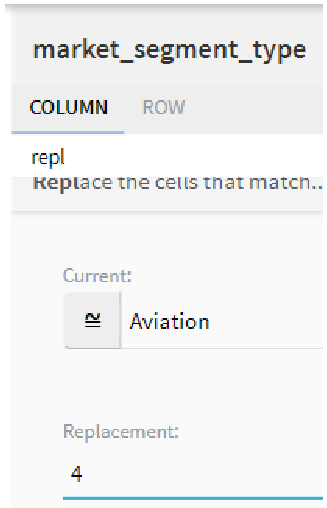


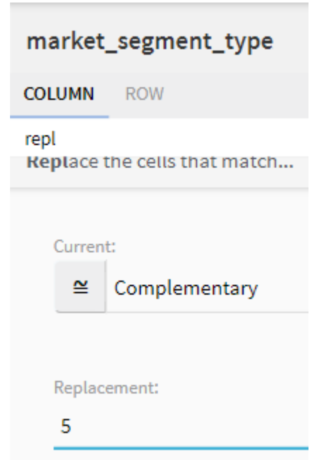




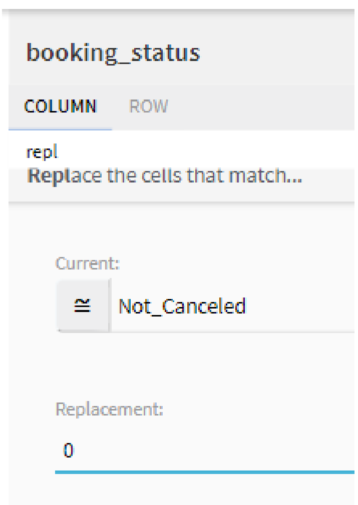
market_segment_type

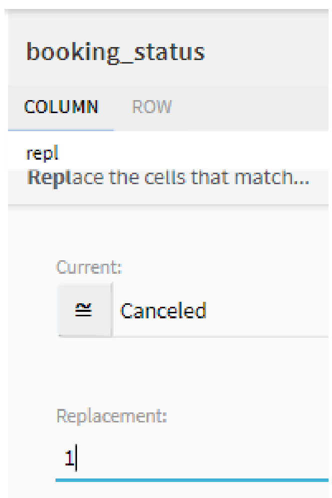






booking_status





4.3.3 Task 3 : Feature Engineering

Feature engineering feature extraction or feature discovery is the process of extracting features from raw data to support training a downstream statistical model. We also adopted this method to our datasets. To complete this task we used the Talend Data Preparation tool as well.

We have successfully created two new features such as below :

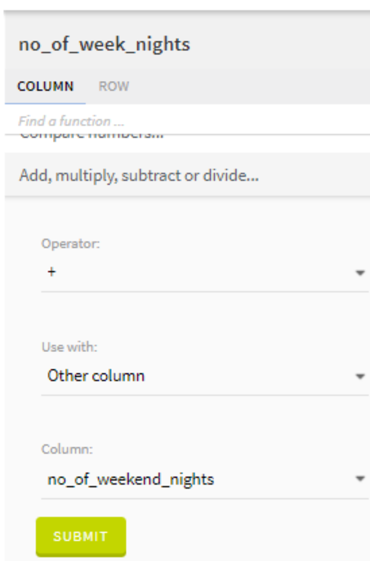
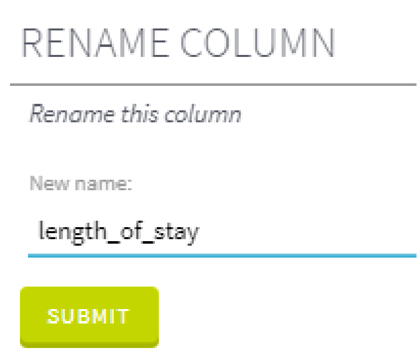
length_of_stay = no_of_weekend_nights + no_of_weekday_nights

- This variable indicates the duration of the stay by totaling up the number of weekday and weekend nights.

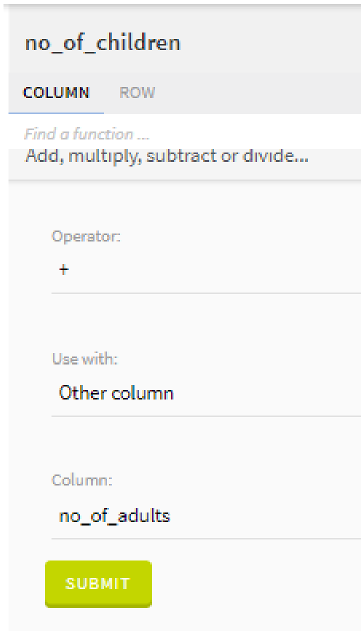
total_guests = no_of_adults + no_of_child

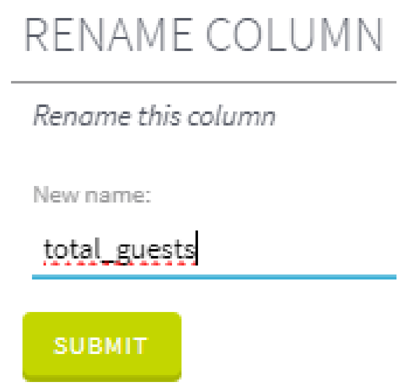
- This variable indicates the total number of guests by totaling up the number of adults and children.

The tables below are some screenshots attached for the variables above as evidence of the task.

<table>
<tr><th colspan="2">length_of_stay</th></tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
</table>

total_guests





4.3.4 Task 4: Variable Preparation

After performing the tasks above, the data was loaded into SAS Enterprise Miner. After uploading the data, we have to prepare the variables by assigning roles and levels to the variables.

Below are the changes done.

<u>**Role**</u>
booking_status: Input → Target

<u>**Level**</u>
booking_status: Interval → Binary
market_segment_type: Interval → Nominal
repeated_guest: Interval → Binary
required_car_parking_space: Interval → Binary
room_type_reserved: Interval → Nominal
type_of_meal_plan: Interval → Nominal

Below is the before and after of the task.

Before:

Variables - FIMPORT

(none) | not | Equal to | ...

Columns: Label | Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
arrival_date	Input	Interval	No		No	.	.
arrival_month	Input	Interval	No		No	.	.
arrival_year	Input	Interval	No		No	.	.
avg_price_per_r	Input	Interval	No		No	.	.
booking_status	Input	Interval	No		No	.	.
lead_time	Input	Interval	No		No	.	.
length_of_stay	Input	Interval	No		No	.	.
market_segment	Input	Interval	No		No	.	.
no_of_children	Input	Interval	No		No	.	.
no_of_previous_	Input	Interval	No		No	.	.
no_of_previous_	Input	Interval	No		No	.	.
no_of_special_r	Input	Interval	No		No	.	.
no_of_week_nig	Input	Interval	No		No	.	.
no_of_weekend	Input	Interval	No		No	.	.
repeated_guest	Input	Interval	No		No	.	.
required_car_pa	Input	Interval	No		No	.	.
room_type_rese	Input	Interval	No		No	.	.
total_guests	Input	Interval	No		No	.	.
type_of_meal_p	Input	Interval	No		No	.	.

After:

The highlighted rows are the ones that had changes.

Variables - FIMPORT

(none) | not | Equal to | ...

Columns: Label | Mining | Basic

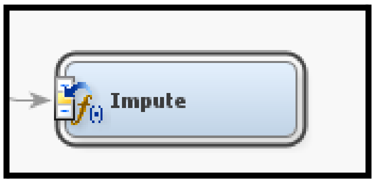
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
arrival_date	Input	Interval	No		No	.	.
arrival_month	Input	Interval	No		No	.	.
arrival_year	Input	Interval	No		No	.	.
avg_price_per_room	Input	Interval	No		No	.	.
booking_status	Target	Binary	No		No	.	.
lead_time	Input	Interval	No		No	.	.
length_of_stay	Input	Interval	No		No	.	.
market_segment_type	Input	Nominal	No		No	.	.
no_of_children	Input	Interval	No		No	.	.
no_of_previous_bookings_not_canc	Input	Interval	No		No	.	.
no_of_previous_cancellations	Input	Interval	No		No	.	.
no_of_special_requests	Input	Interval	No		No	.	.
no_of_week_nights	Input	Interval	No		No	.	.
no_of_weekend_nights	Input	Interval	No		No	.	.
repeated_guest	Input	Nominal	No		No	.	.
required_car_parking_space	Input	Binary	No		No	.	.
room_type_reserved	Input	Binary	No		No	.	.
total_guests	Input	Interval	No		No	.	.
type_of_meal_plan	Input	Nominal	No		No	.	.
VAR1	Rejected	Interval	No		No	.	.

4.3.5 Task 5: Impute Missing Values

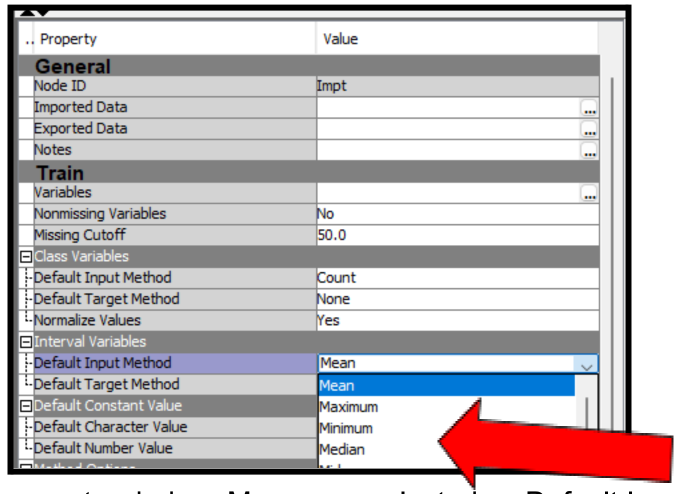
There were some missing values found in our dataset which have been mentioned in the Explore stage. Thus, this task was crucial to address this problem

Impute node was selected and steps were done and the results are explained below.

Node Name:	Impute	**Configuration**



The Impute Node is more versatile and provides multiple imputation methods to handle missing values. It allows users to choose from a variety of imputation techniques, such as regression imputation, k-nearest neighbor imputation, and more.



In the property window, Mean was selected as Default Input Method. The mean is the most common measure of a variable's central tendency; it is an unbiased estimate of the population mean. The mean is the preferred statistic to use to replace missing values if the variable values are at least roughly symmetric

Result

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP_avg_price_per_room	MEAN	IMP_REP_avg_price_per_room	M_REP_avg_price_per_room	105.011	INPUT	INTERVAL	Replacement: avg_price_per_room	541

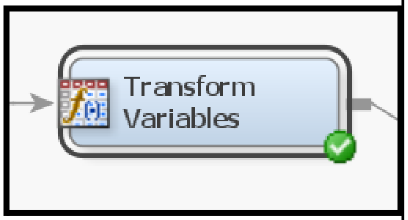
The variable "REP_avg_price_per_room" underwent imputation using the mean imputation method. These findings summarize the imputation process for the specified variable

- **Imputed Variable:**
    - The imputed values were stored in a new variable called "IMP_REP_avg_price_per_room."
- **Imputed Value:**
    - The mean imputed value for "REP_avg_price_per_room" is 105.011.

4.3.6 Task 6: Data Transformation

As mentioned in the Explore stage, some variables undergo skewness.

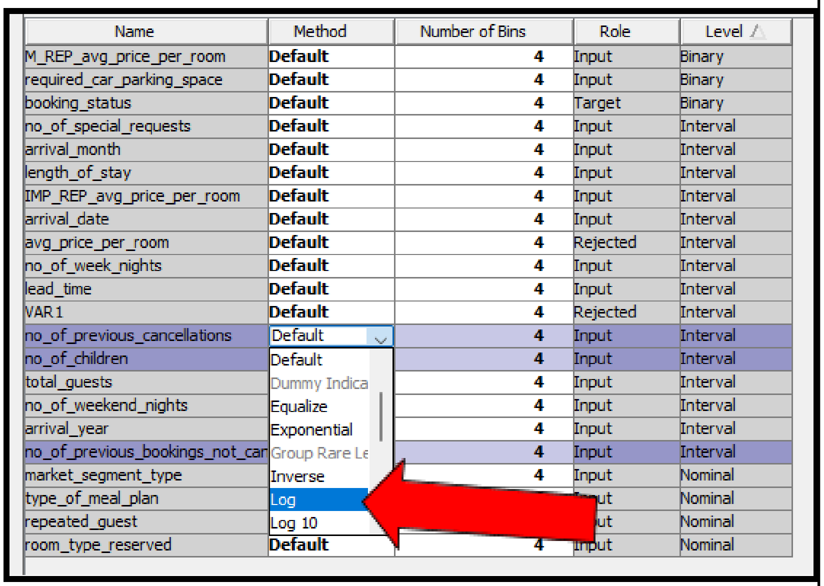
Node Name:	Transform Variables	**Configuration**



The Transform Variables node allows users to create new variables from existing variables. The original variables used in the calculations will remain unchanged, however new column(s) will be added to the dataset.

As highlighted in Explore stage on the high skewness and kurtosis problem, there are 3 variables selected for the transformation which are again:

- no_of_children
- no_of_previous_bookings_not_canc
- no_of_previous_cancellations



Name	Method	Number of Bins	Role	Level
M_REP_avg_price_per_room	Default	4	Input	Binary
required_car_parking_space	Default	4	Input	Binary
booking_status	Default	4	Target	Binary
no_of_special_requests	Default	4	Input	Interval
arrival_month	Default	4	Input	Interval
length_of_stay	Default	4	Input	Interval
IMP_REP_avg_price_per_room	Default	4	Input	Interval
arrival_date	Default	4	Input	Interval
avg_price_per_room	Default	4	Rejected	Interval
no_of_week_nights	Default	4	Input	Interval
lead_time	Default	4	Input	Interval
VAR1	Default	4	Rejected	Interval
no_of_previous_cancellations	Default	4	Input	Interval
no_of_children	Default	4	Input	Interval
total_guests	Dummy Indica	4	Input	Interval
no_of_weekend_nights	Equalize	4	Input	Interval
arrival_year	Exponential	4	Input	Interval
no_of_previous_bookings_not_car	Group Rare Le	4	Input	Interval
market_segment_type	Inverse	4	Input	Nominal
type_of_meal_plan	Log		ut	Nominal
repeated_guest	Log 10		ut	Nominal
room_type_reserved	Default	4	Input	Nominal

For the method Log was selected as transformation method for these variables

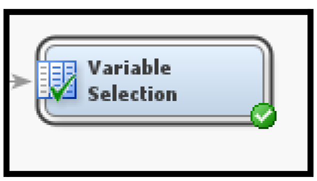
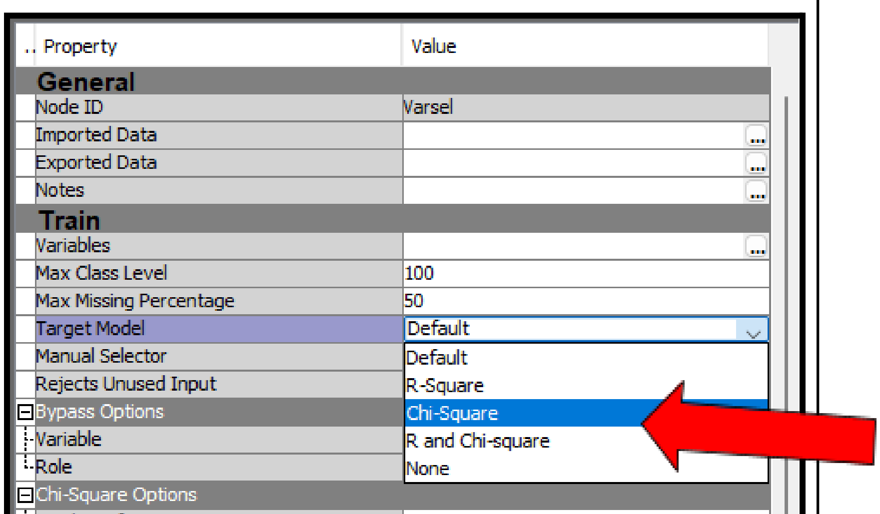
Results

Transformations Statistics

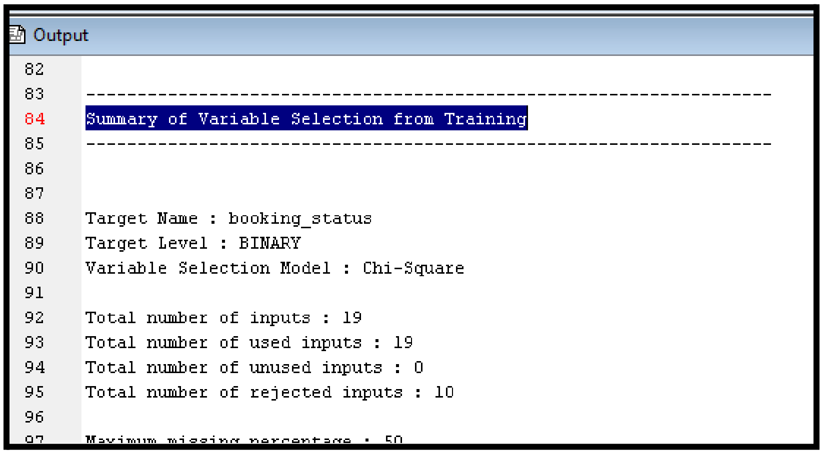
Source	Method	Variable Name	Formula	Skewness
Output	Computed	LOG_no_of_children	log(no_of_children + 1)	3.629558
Output	Computed	LOG_no_of_previous_bookings_...	log(no_of_previous_bookings_not_canc + 1)	8.916943
Output	Computed	LOG_no_of_previous_cancellatio...	log(no_of_previous_cancellations + 1)	14.35091

The above table indicates the newly created variables as the output of running the node. As we can see the transformed variables have reduced skewness

4.3.7 Task 7: Variable Selection

Node Name:	Variable selection	**Configuration**
 The Variable Selection node quickly identifies input variables that are useful for predicting the target variable or variables. The input status is assigned to these variables.		 Chi Square is chosen since it is a classification problem.

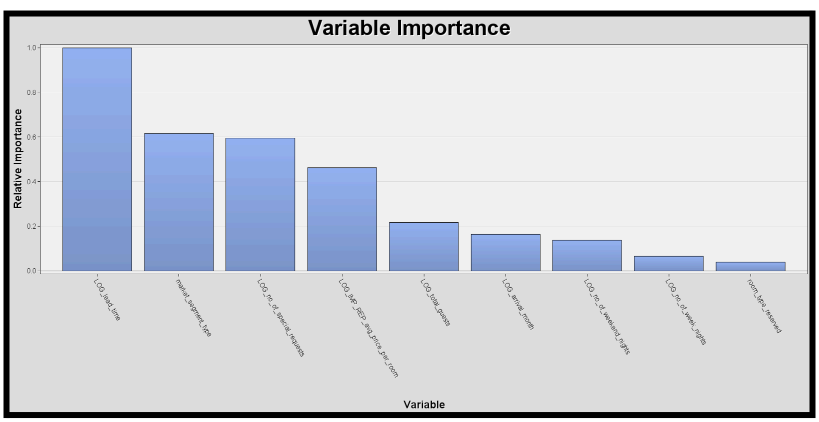
Result



```
Output

--------------------------------------------------------------------
Summary of Variable Selection from Training
--------------------------------------------------------------------


Target Name : booking_status
Target Level : BINARY
Variable Selection Model : Chi-Square

Total number of inputs : 19
Total number of used inputs : 19
Total number of unused inputs : 0
Total number of rejected inputs : 10

Maximum missing percentage : 50
```

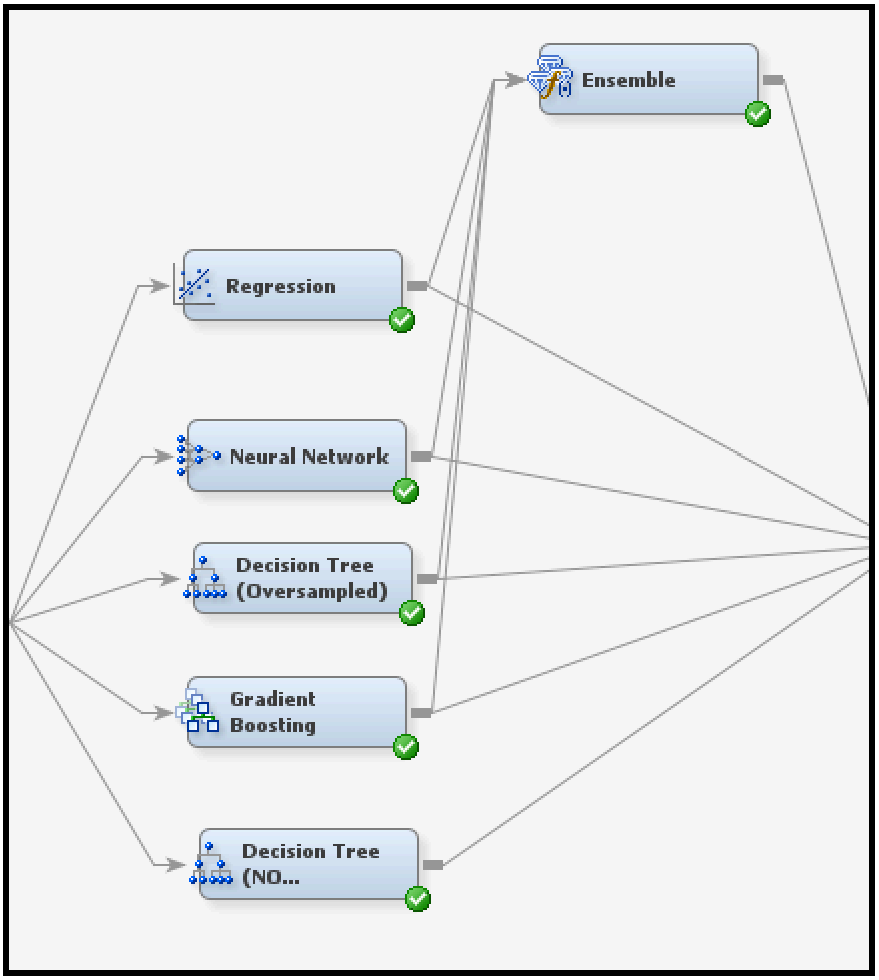
The summary indicates that Chi-Square was used for variable selection with the target variable being "booking_status" (a binary outcome). Here are key points from the summary:

- **Target Variable:**
  - **Name:** booking_status
  - **Type:** Binary (two possible outcomes)
- **Variable Selection Method:**
  - **Model**: Chi-Square
  - The Chi-Square test was employed to select relevant variables for predicting the binary outcome of booking_status.
- **Number of Inputs:**
  - **Total Inputs**: 19
  - **Used Inputs**: 19
  - **Unused Inputs**: 0
  - All 19 input variables were considered in the variable selection process, and none of them were excluded.



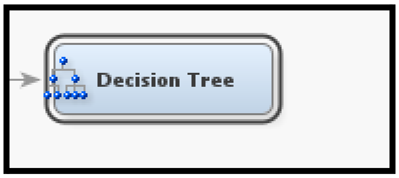
The graph indicates the order of relative importance of the variables/ As we can see LOG_lead_time has the highest importance while room_type_reserved has the least importance.

4.4 Model



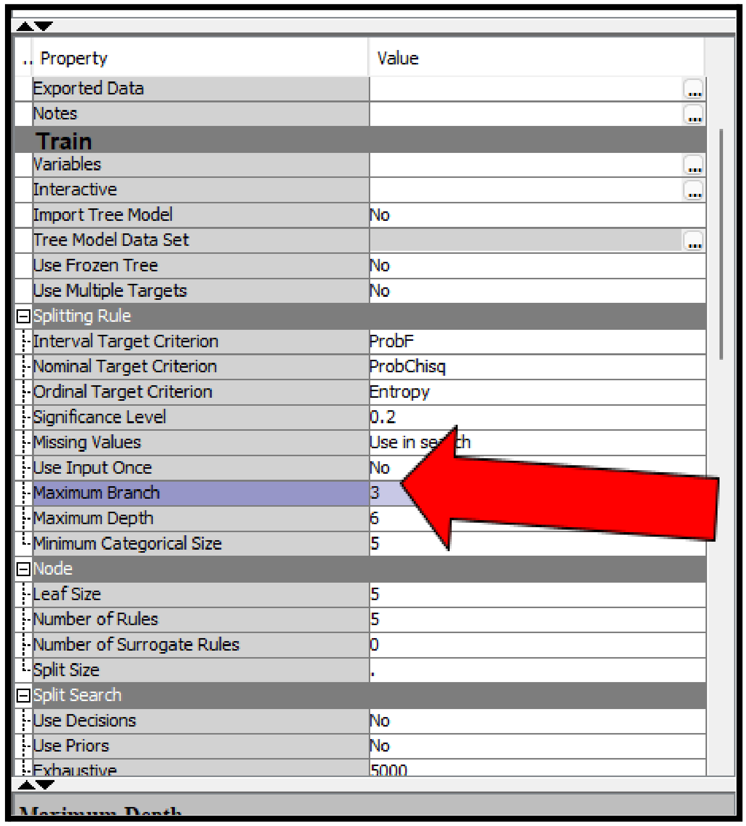
Ensemble
Regression
Neural Network
Decision Tree
(Oversampled)
Gradient
Boosting
Decision Tree
(NO...

<table>
<tr><td>Node Name:</td><td>Decision Tree</td><td>Configuration</td></tr>
<tr><td colspan="2">

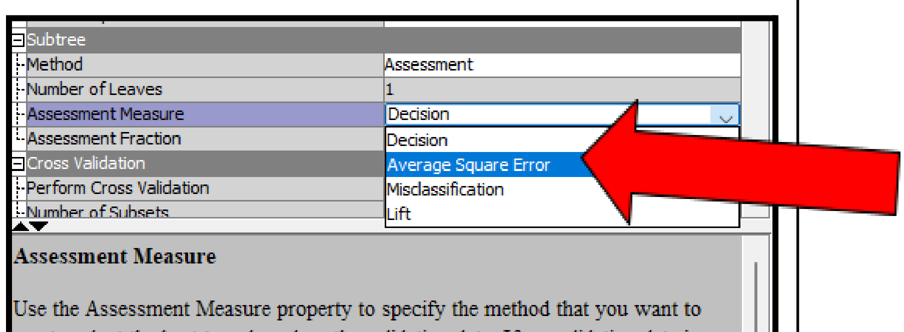


An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. The original segment contains the entire data set and is called the root node of the tree. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. The type of decision depends on the context. In predictive modelling, the decision is the predicted value.

</td><td>



From the properties table, the highlighted indicates the number of branches for the tree was selected as 3.



For the assessment measure, the Average Square Error was selected.

</td></tr>
</table>

Result

Fit Statistics

Target=booking_status Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	20999.00	9001.00
MISC	Misclassification Rate	0.16	0.16
MAX	Maximum Absolute Error	0.99	0.99
SSE	Sum of Squared Errors	4882.01	2082.67
ASE	Average Squared Error	0.12	0.12
RASE	Root Average Squared Error	0.34	0.34
DIV	Divisor for ASE	41998.00	18002.00
DFT	Total Degrees of Freedom	20999.00	.

From the fit statistics table, we can see that the analysis was conducted on a dataset with 20,999 observations for training and 9,001 observations for validation. Some key things to take note of were :

- The decision tree model achieved a misclassification rate of 16% on both the training and validation datasets.
- The maximum absolute error was 0.99 for both datasets.
- The sum of squared errors (SSE) was 4,725.49 for training and 2,016.10 for validation.
- The average squared error (ASE) was 0.11 for both datasets.
- The root average squared error (RASE) was 0.34 for training and 0.33 for validation.

Event Classification Table

Data Role=TRAIN Target=booking_status Target Label=' '

False Negative	True Negative	False Positive	True Positive
1995	12797	1332	4875

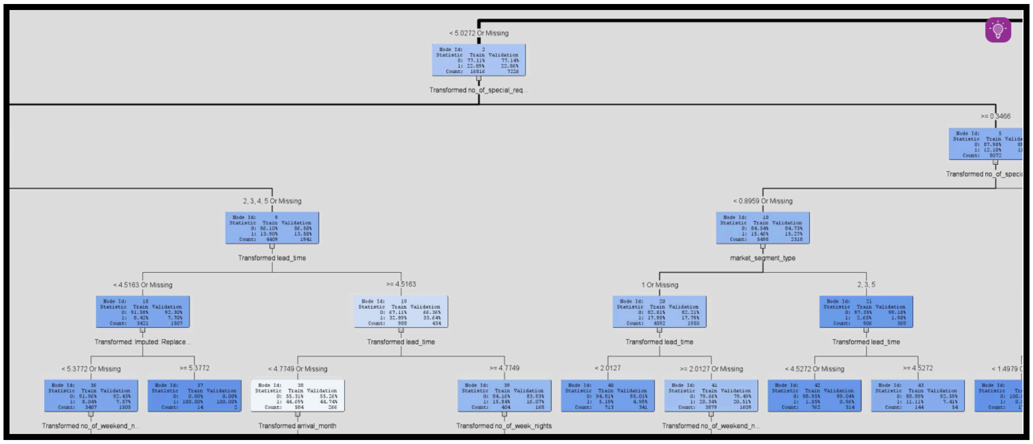
Data Role=VALIDATE Target=booking_status Target Label=' '

False Negative	True Negative	False Positive	True Positive
851	5479	577	2094

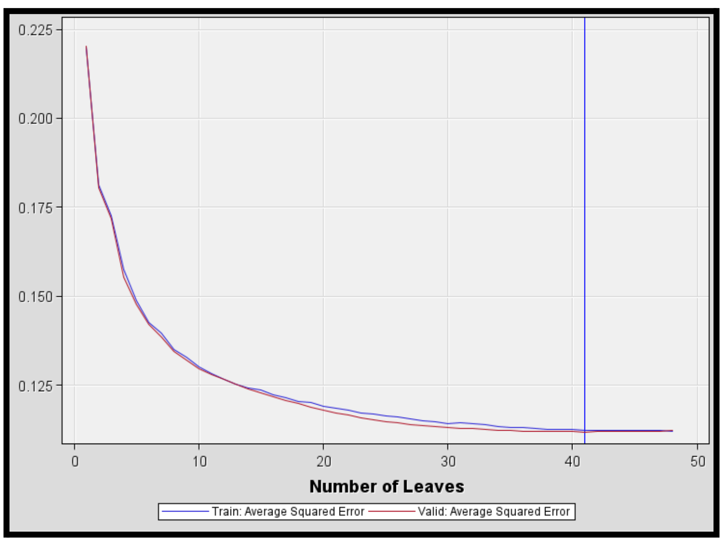
The classification tables reveal the model's performance on the binary target variable "booking_status" for both the training and validation datasets.

Key findings from the table are:

- The model achieved high accuracy (85.26% in training, 85.43% in validation) for correctly predicting instances where bookings were not cancelled (Target 0).
- The accuracy for correctly identifying instances where bookings were cancelled (Target 1) was also high, reaching 81.42% in training and 81.13% in validation.
- Challenges were observed in distinguishing between non-cancelled bookings (Outcome 0) and cancelled bookings (Outcome 1), as indicated by lower accuracy percentages (32.91% in training, 32.39% in validation) for Target 1 in the non-cancelled category.
- The classification tables highlight potential areas for model refinement, particularly in improving the model's ability to differentiate between non-cancelled and cancelled bookings.



Above is the part of the decision tree that was modelled. (The full picture was too huge to be included thus why a small portion was selected to be shown here.)

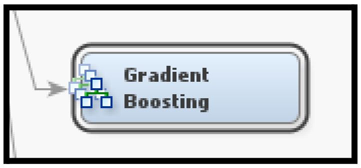


Training Set: The MSE on the training set measures how well the model fits the training data. It quantifies the average squared difference between the actual and predicted values.

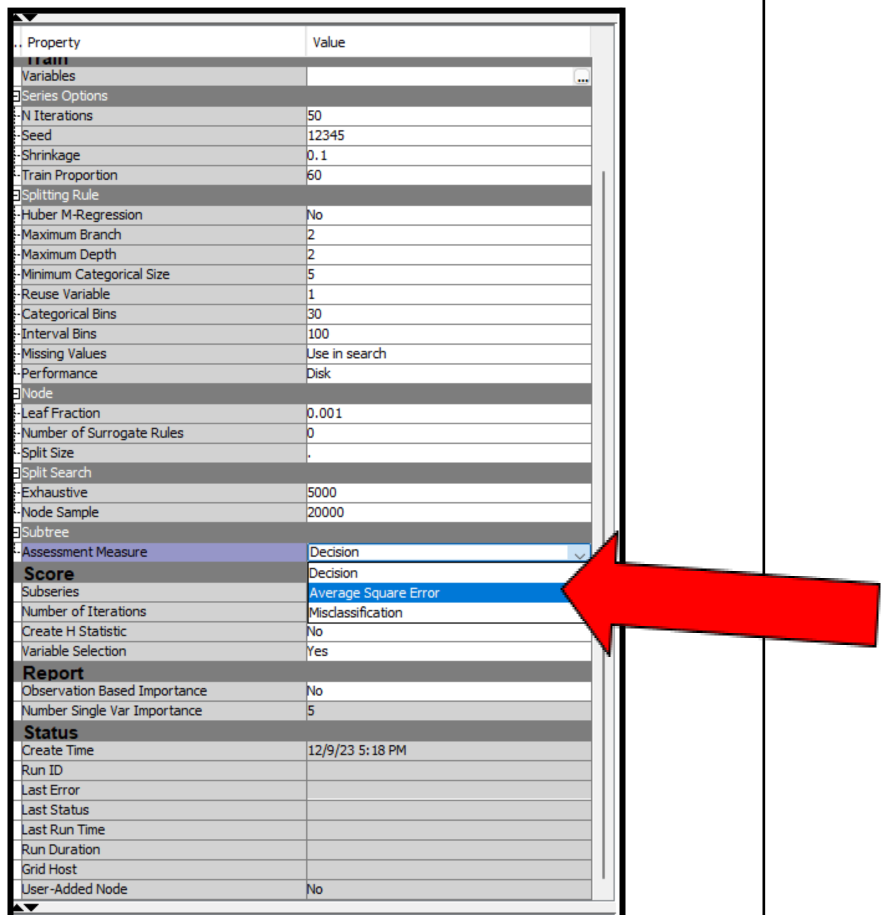
Validation Set: The MSE on the validation set evaluates how well the model generalizes to new, unseen data. It indicates how the model is expected to perform on data it has not seen during training.

As we can see both the sets are reducing the error as the number of leaves increases.

Node Name:	Gradient Boosting	**Configuration**

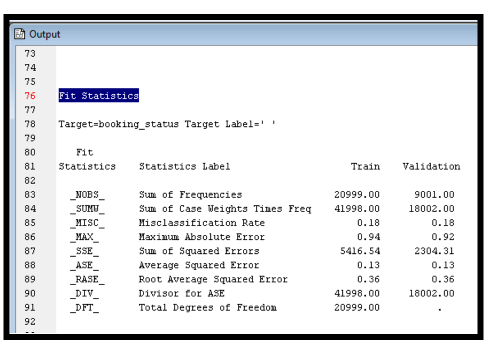


This node uses a partitioning algorithm described in "Greedy Function Approximation: A Gradient Boosting Machine," and "Stochastic Gradient Boosting" by Jerome Friedman. A partitioning algorithm searches for an optimal partition of the data defined in terms of the values of a single variable.



As in the previous model, the assessment measure Average Square Error was selected.

Result



Output

Fit Statistics

Target=booking_status Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	20999.00	9001.00
SUMW	Sum of Case Weights Times Freq	41998.00	18002.00
MISC	Misclassification Rate	0.18	0.18
MAX	Maximum Absolute Error	0.94	0.92
SSE	Sum of Squared Errors	5416.54	2304.31
ASE	Average Squared Error	0.13	0.13
RASE	Root Average Squared Error	0.36	0.36
DIV	Divisor for ASE	41998.00	18002.00
DFT	Total Degrees of Freedom	20999.00	.

From the fit statistics table, we can see that the analysis was conducted on a dataset with 20,999 observations for training and 9,001 observations for validation. Some key things to take note of were:

- **Misclassification Rate:**
  - Both the training and validation datasets exhibit a misclassification rate of 18%.
- **Maximum Absolute Error:**
  - The model's maximum absolute error is 0.94 for training and 0.92 for validation.
- **Sum of Squared Errors (SSE):**
  - The SSE is 5,416.54 for training and 2,304.31 for validation, indicating a lower error in the validation dataset.
- **Average Squared Error (ASE):**
  - The ASE is 0.13 for both training and validation, providing an average measure of prediction error.
- **Root Average Squared Error (RASE):**
  - RASE is 0.36 for both training and validation, indicating similar average prediction errors.
- **Effective Sample Size:**
  - The effective sample size, represented by the sum of case weights times frequency, is 41,998.00 for training and 18,002.00 for validation.

Classification Table

Data Role=TRAIN Target Variable=booking_status Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	83.6635	90.8698	12839	61.1410
1	0	16.3365	36.4920	2507	11.9387
0	1	22.8197	9.1302	1290	6.1431
1	1	77.1803	63.5080	4363	20.7772

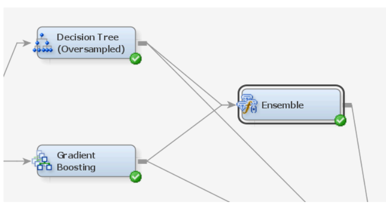
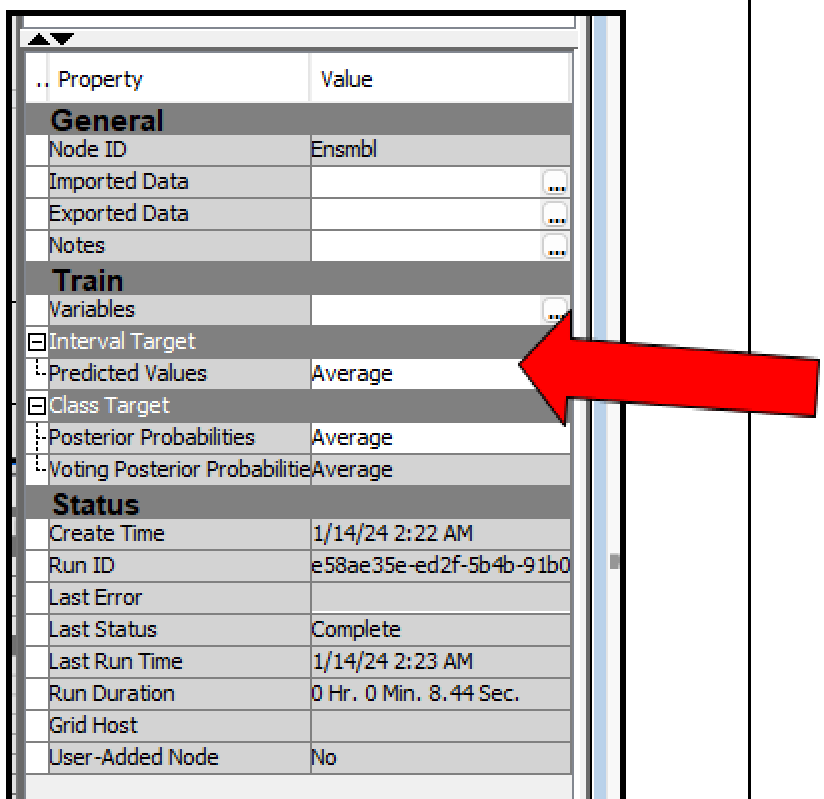
Data Role=VALIDATE Target Variable=booking_status Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	83.9082	91.1823	5522	61.3487
1	0	16.0918	35.9593	1059	11.7654
0	1	22.0661	8.8177	534	5.9327
1	1	77.9339	64.0407	1886	20.9532

Some of the key findings are:

- **Training Dataset:**
  - The Gradient Boosting model achieved an 83.66% accuracy in correctly predicting non-cancelled bookings (Target 0) and a 77.18% accuracy in predicting cancelled bookings (Target 1).
  - The model struggled to differentiate between non-cancelled (Outcome 0) and cancelled bookings (Outcome 1) in both training and validation datasets, resulting in lower accuracies for Target 1 in the non-cancelled category (36.49% in training, 35.96% in validation).
- **Validation Dataset:**

- Similar to the training dataset, the model demonstrated an 83.91% accuracy for correctly predicting non-cancelled bookings (Target 0) and a 77.93% accuracy for predicting cancelled bookings (Target 1).
- The misclassification rates remained consistent at 16.09% for non-cancelled bookings and 22.07% for cancelled bookings in the validation dataset.

Node Name:	Ensemble	**Configuration**
		

Result

Fit Statistics

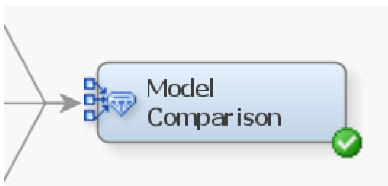
Target=booking_status Target Label=' '

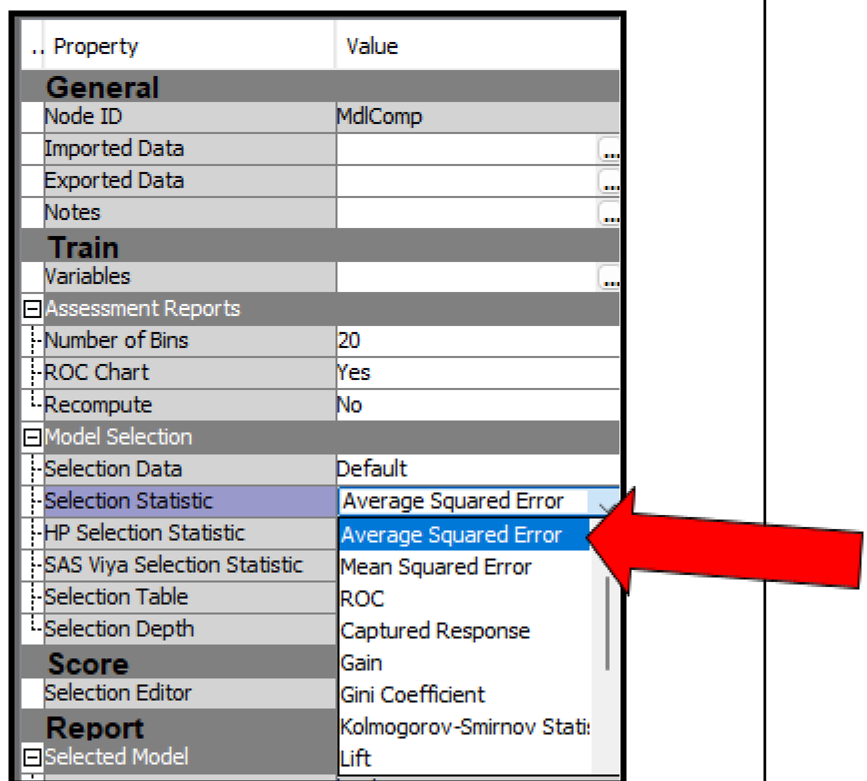
Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.13	0.13
DIV	Divisor for ASE	41998.00	18002.00
MAX	Maximum Absolute Error	0.89	0.88
NOBS	Sum of Frequencies	20999.00	9001.00
RASE	Root Average Squared Error	0.36	0.36
SSE	Sum of Squared Errors	5401.23	2293.98
DISF	Frequency of Classified Cases	20999.00	9001.00
MISC	Misclassification Rate	0.16	0.16
WRONG	Number of Wrong Classifications	3394.00	1446.00

Some of the key findings are that the model achieves a score of 0.13 for Average Squared error which is still higher compared to other models.

4.5 Assess

In this final SEMMA stage, the model is evaluated for how useful and reliable it is for the studied topic. The data can now be tested and used to estimate the efficacy of its performance.

<table>
<tr><td>Node Name:</td><td>Model Comparison</td><td>Configuration</td></tr>
<tr><td colspan="2">


The Model Comparison node belongs to the Assess category in the SAS data mining process of Sample, Explore, Modify, Model, and Assess (SEMMA). The Model Comparison node enables you to compare the performance of competing models using various benchmarking criteria. The combined criteria enable the analyst to make cross-model comparisons and assessments.
</td><td>


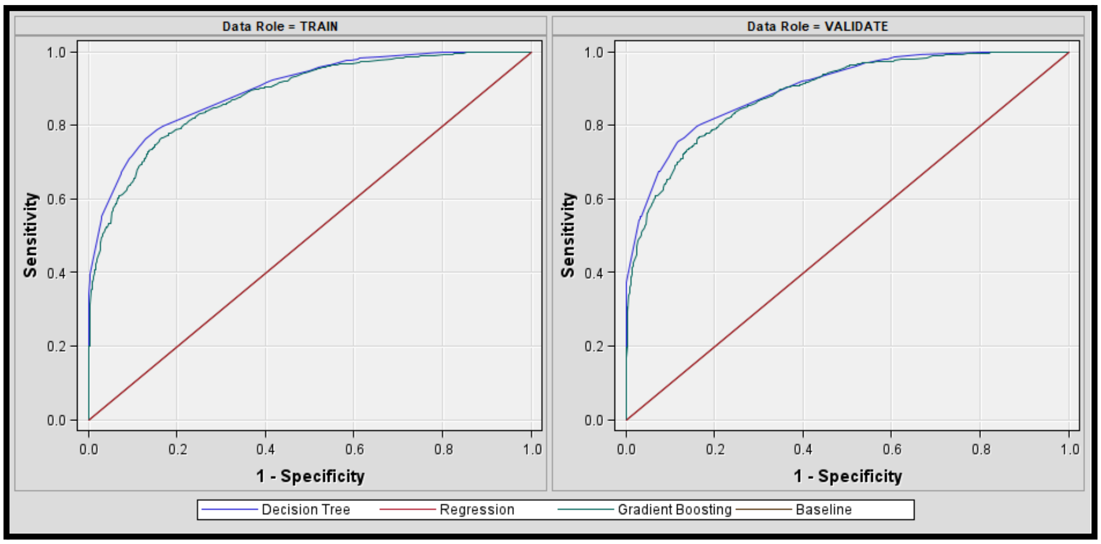
For the selection statistics criteria Average Squared Error was selected. It chooses the model with the smallest average squared error value.
</td></tr>
</table>

Results

Fit Statistics

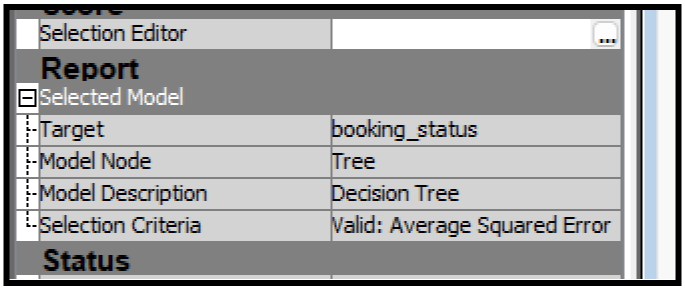
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Tree	Decision Tree	0.11199	0.11252	0.15777	0.15743
	Boost	Gradient Boosting	0.12800	0.12897	0.18082	0.17698
	Reg	Regression	0.22014	0.22013	0.32716	0.32719



Below are the key findings from the fit statistics table and ROC Chart:

- **Decision Tree Model:**
  - Selected model based on the Valid Average Squared Error (VASE).
  - Achieved an average squared error of 0.11199 on the validation dataset.
  - Demonstrated a misclassification rate of 15.74% on the validation dataset.
- **Neural Network Model:**
  - Had a higher average squared error (0.12675) compared to the Decision Tree on the validation dataset.
  - The associated misclassification rate was 17.47% on the validation dataset.
- **Gradient Boosting Model:**
  - Showed an average squared error of 0.12800 on the validation dataset.
  - Had a misclassification rate of 17.70% on the validation dataset.

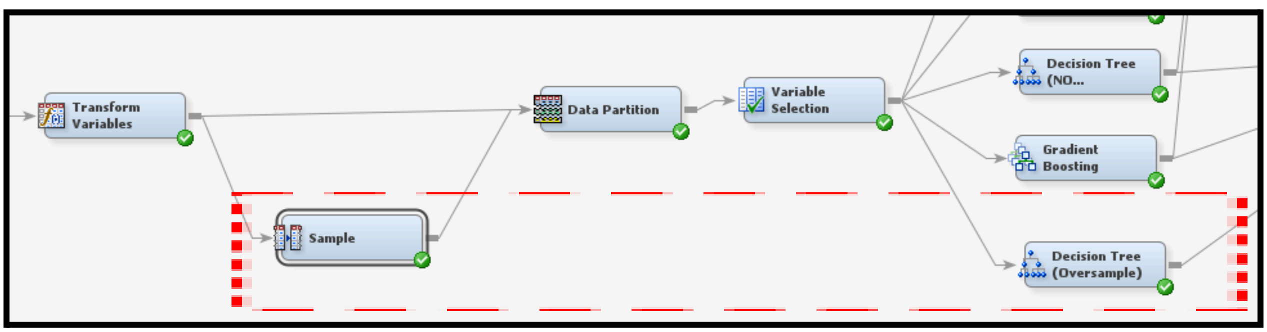


The reports in the property table indicate that the Decision tree was chosen as the final model for the prediction problem.

From the above analysis, we can conclude that Decision Tree was the best model. However, we did witness the model suffer from overfitting due to the class imbalance. It is addressed in the next steps.

4.5.1 Addressing Overfitting & Class Imbalance

To address the overfitting & class imbalance, we added a Sample node and configured to oversampling as below.



Modified Flow

Train	
Variables	...
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Computed
Observations	.
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	Yes
Based on Count	Yes
Exclude Missing Levels	No

Configuration for Oversampling in Sample Node

After executing the new flow, it yielded the results below.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Average Squared Error
Y	Tree	Tree	Decision Tree (Oversampled)	booking_st...	0.114707
	Tree2	Tree2	Decision Tree (NO Oversample)	booking_st...	0.116244
	Neural	Neural	Neural Network	booking_st...	0.129993
	Boost	Boost	Gradient Boosting	booking_st...	0.12863
	Ensmbl	Ensmbl	Ensemble	booking_st...	0.128607
	Reg	Reg	Regression	booking_st...	0.220126

As we can see over here, the oversampled data has performed better than the data without oversampled and it has proven that we successfully addressed the overfitting and class imbalance.

5.0 Conclusion

In conclusion, the analysis of the hotel booking dataset has provided valuable insights into the intricate dynamics of the hospitality industry. By applying the SAS SEMMA methodology to explore, preprocess, model, and assess the data, our group has successfully unravelled patterns and trends that hold significance for hotel management and strategic decision-making.

5.1 Key Findings

- **Booking Patterns and Customer Preferences:**
  - We observed distinct patterns in booking behaviour, shedding light on the factors influencing successful reservations. Understanding customer preferences, such as room type reservations and meal plans, equips hoteliers with the knowledge to tailor services to meet guest expectations.
- **Factors Influencing Cancellations:**
  - The identification of factors contributing to booking cancellations is critical for devising strategies to minimize such occurrences. Insights gained from the dataset offer a foundation for proactive measures and improved revenue management.
- **Optimizing Customer Engagement:**
  - By analyzing features like special requests and historical booking behavior, we have uncovered opportunities to enhance customer engagement. This knowledge allows hotels to provide personalized experiences that resonate with guest preferences.
- **Predictive Modeling for Booking Status:**
  - The development of predictive models has empowered us to forecast the likelihood of booking cancellations. This proactive approach enables hotels to implement preventive measures and improve overall booking success rates.
- **Contributions to the Hospitality Industry:**
  - Our findings contribute to the broader understanding of data-driven practices within the hospitality sector. The actionable insights derived from this analysis can serve as a guide for hotel management in optimizing operational strategies, marketing campaigns, and customer experiences.

5.2 Recommendations and Future Work

As we conclude this project, we recommend the following areas for further exploration:

- **Dynamic Pricing Strategies:** Investigate the implementation of dynamic pricing strategies based on historical booking patterns and customer behaviour.
- **Personalization in Marketing:** Explore avenues for further personalization in marketing campaigns to target specific customer segments effectively.
- **Real-time Monitoring:** Implement real-time monitoring systems to promptly identify potential cancellations and take preventive actions.

5.3 Acknowledgment

We express our gratitude to Kaggle for providing the dataset and fostering a collaborative environment for data science enthusiasts. This project would not have been possible without the wealth of resources and opportunities offered by the Kaggle community.

In summary, this data mining project has not only deepened our understanding of hotel booking dynamics but also laid the groundwork for practical applications that can positively impact the hospitality industry. We look forward to the continued evolution of data-driven strategies and their role in shaping the future of hotel management.

6.0 Appendix

1. Github
2. Youtube Video Link