Predicting Hotel Reservation Cancellation

01

Introduction

Background

- Hospitality relies on reservations for revenue
- Cancellations cause revenue loss and operational issues
- Unpredictable cancellations impact planning
- Need for a predictive model for accurate forecasts

Problem Statement

“Derive a predictive model to minimize cancellations”

Objectives

1. Develop predictive models
2. Identify key factors influencing cancellation
3. Provide actionable insights and recommendations

Methodology

- **Dataset description:** H1 (resort hotel) subset
- **Data preprocessing and EDA:** Cleaning, handling missing values, and encoding categorical variables (one-hot encoder, imputer)
- **Model Fitting:** Model development using logistic regression, decision trees, KNN
- **Integration Technique:** SMOTE
- **Model evaluation metrics:** Accuracy, precision, Recall

02

Data Collection

Original
Variables: 31
Observations: ~40,000

After Cleaning
Variables: 21
Observations: ~37,000

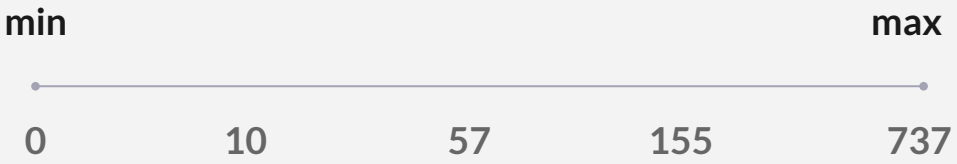
Bookings between July 1, 2015, and August 31, 2017

03
Data Preprocessing

Is Canceled- Target

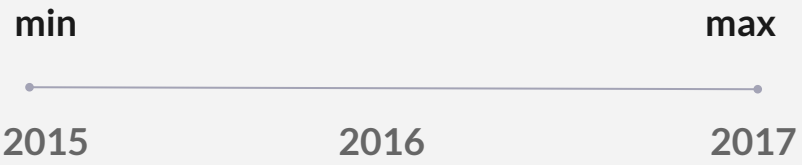
Lead Time

Numerical



Arrival Date Year

Categorical



Arrival Date Month

Categorical

January

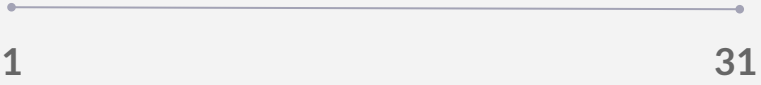
December

Arrival Date Week Number

Numerical

Arrival Date Day Of Month

Numerical



Stays In Weekend Nights

Stayed 0 Weekend night

Stayed more than 1 Weekend night

Stays in Week Nights

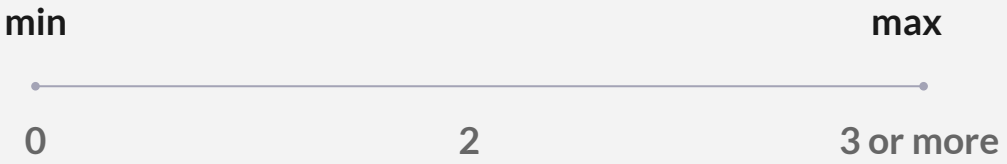
No prior stays

1-5 weeknight stays

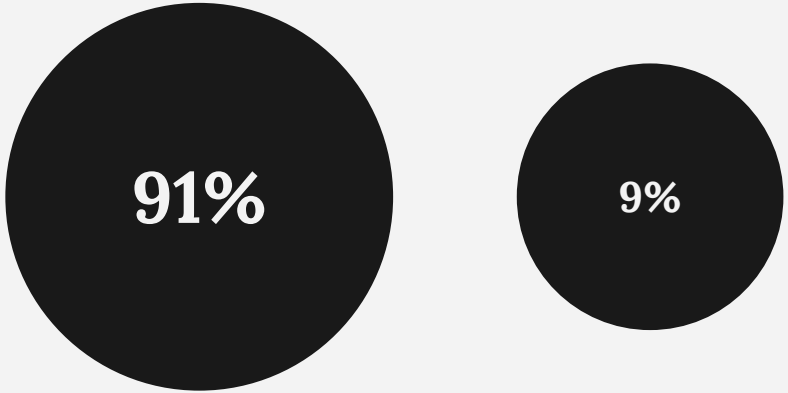
More than 5 weeknight stays

Adults

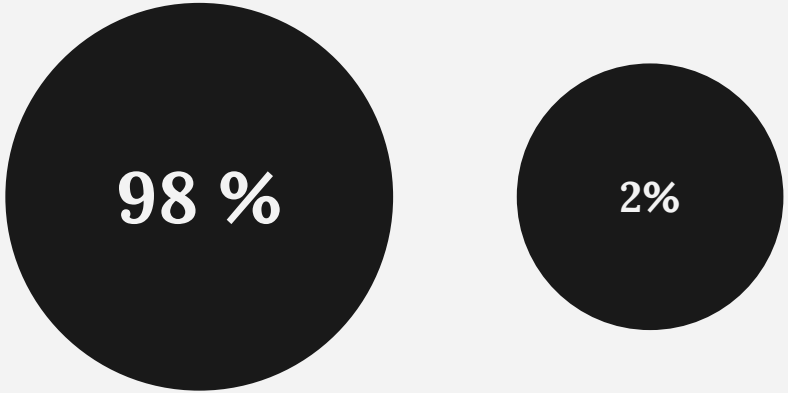
Numerical



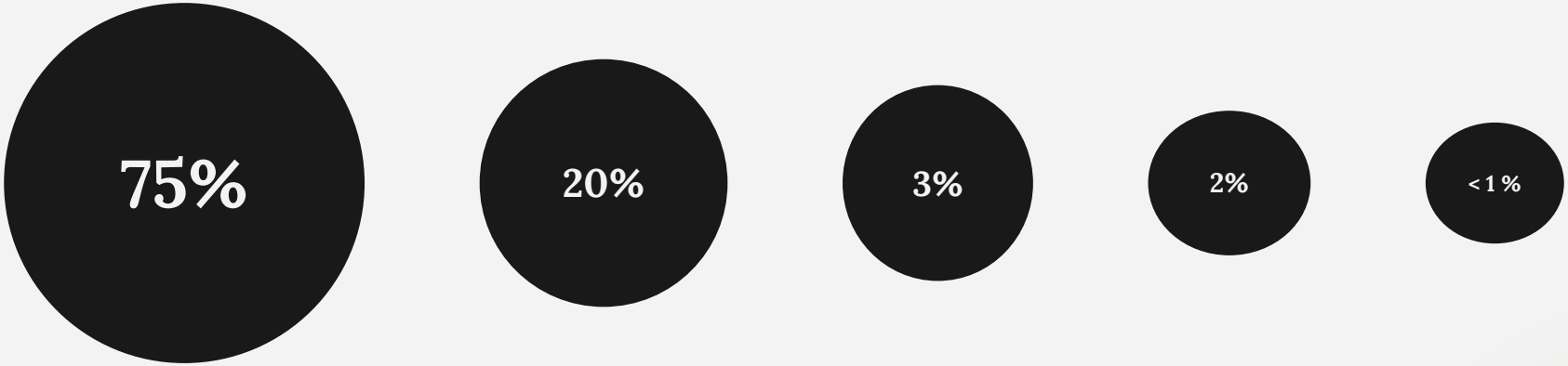
Children



Babies



Meals



Bed & Breakfast

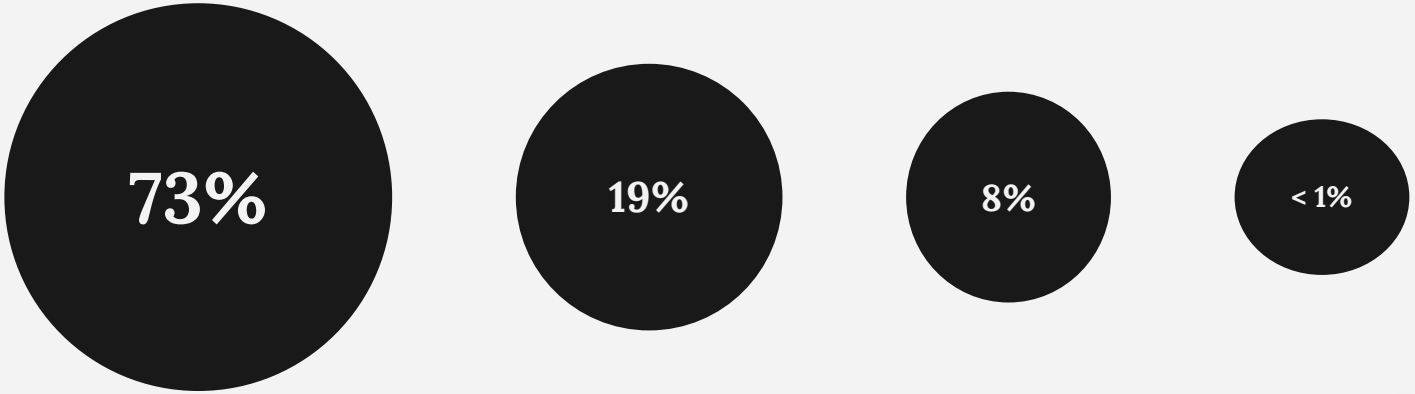
Half board

Undefined

Full board

SC

Distribution Channel



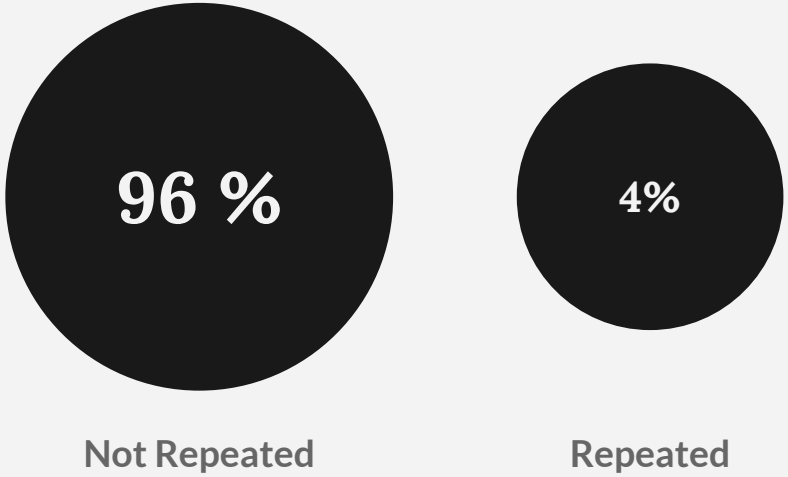
Travel Agents/
Tour Operators

Direct

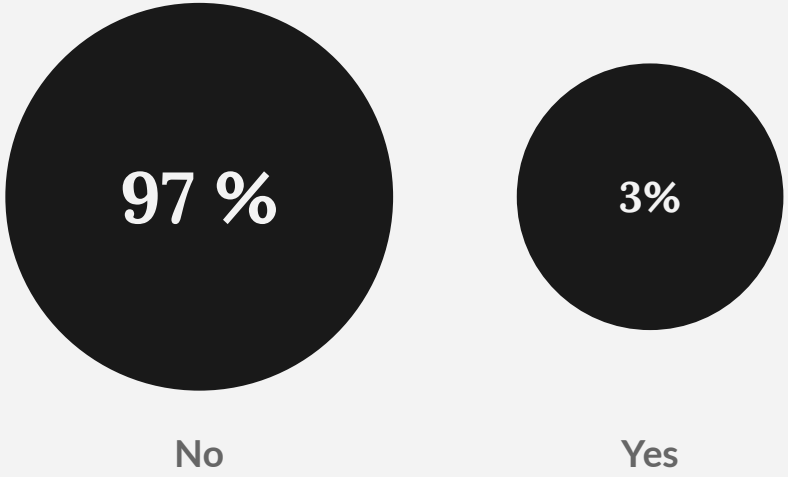
Corporate

Undefined

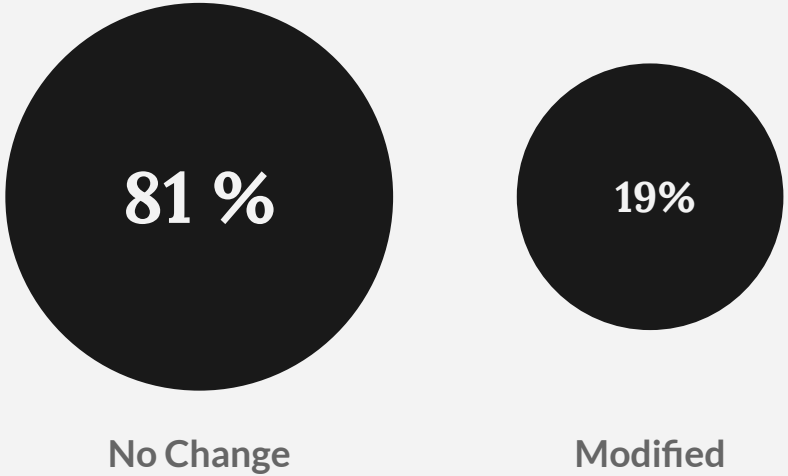
Repeated Guest



Previous Cancellation



Booking Changes



Deposit Type

Categorical

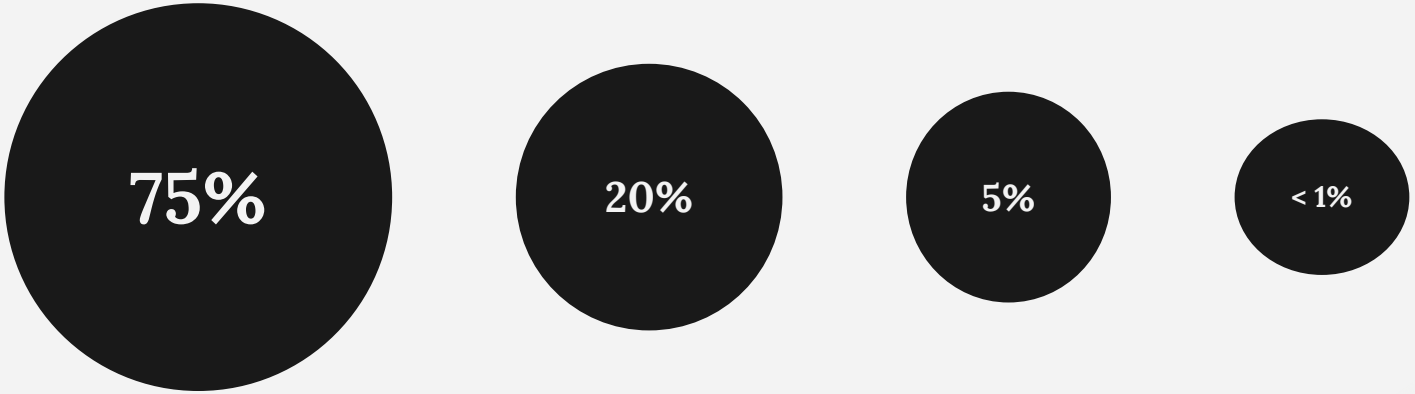
No Deposit

Non - Refund

Refundable

Customer Type

Categorical



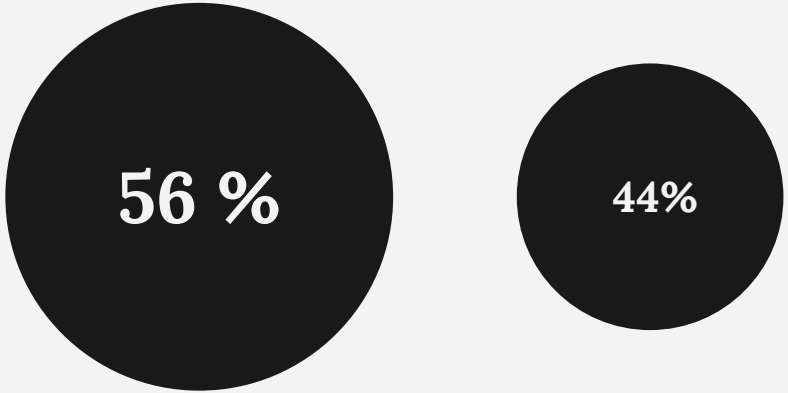
Transient

Transient-Party

Contract

Group

Special Requests



Without any Special Requests

At least one Special Request

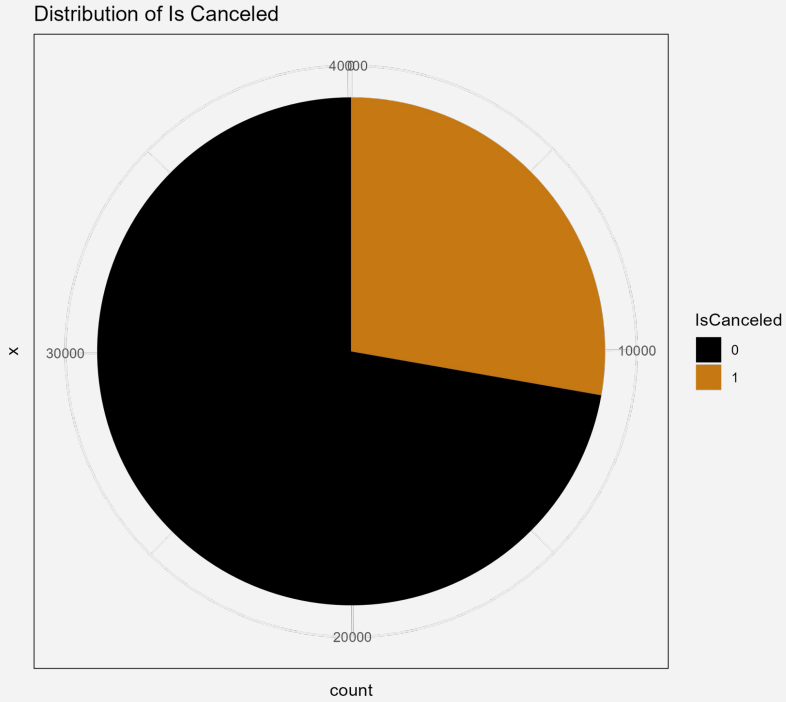
04

Exploratory Analysis

IsCanceled - Target

A smaller fraction of cancellations

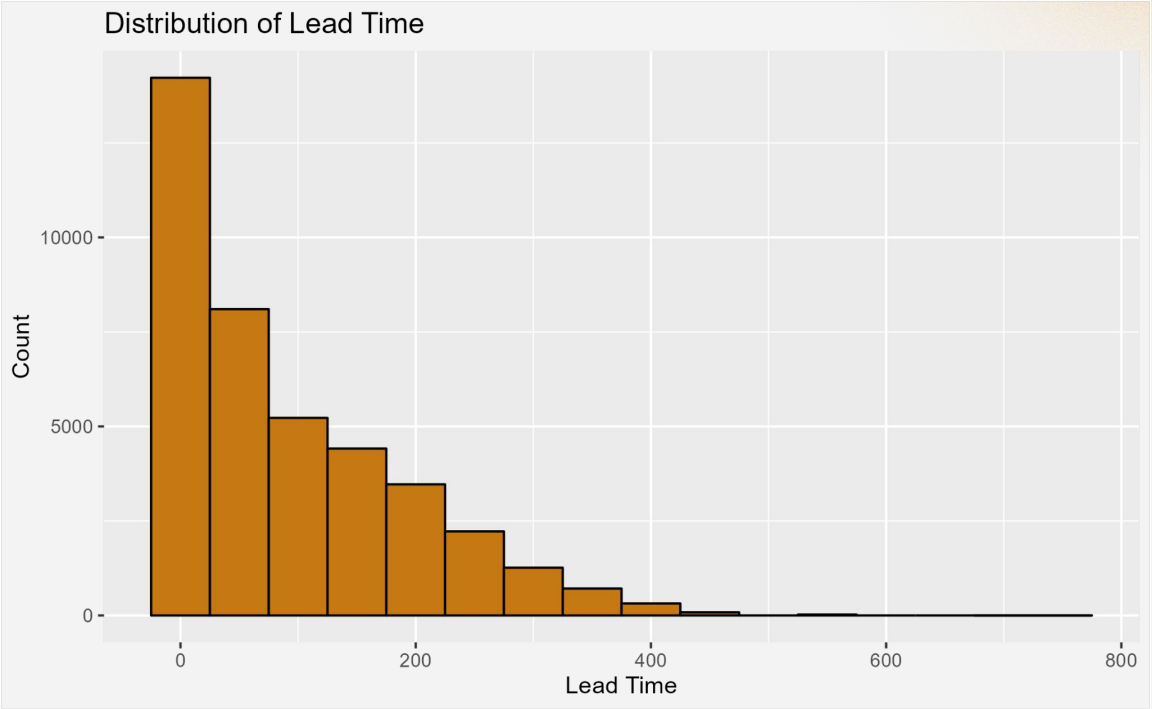
Transformation- ***Factorize***



LeadTime

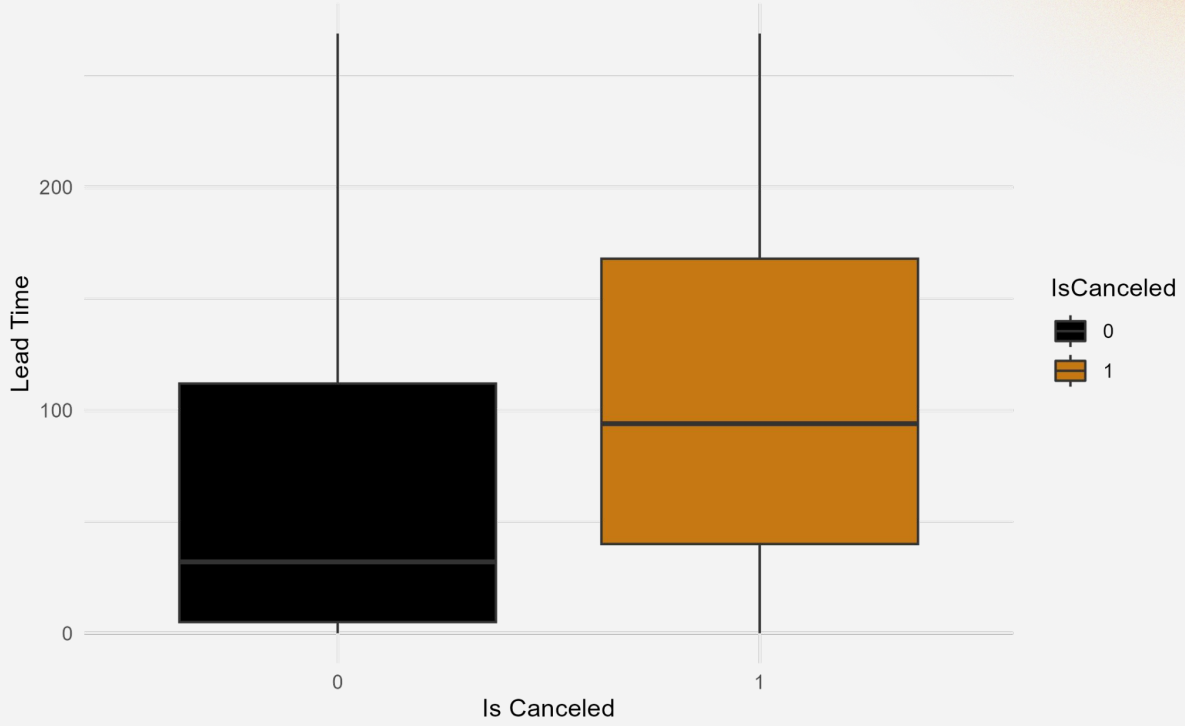
High frequency of shorter lead time

Remove the outliers



LeadTime

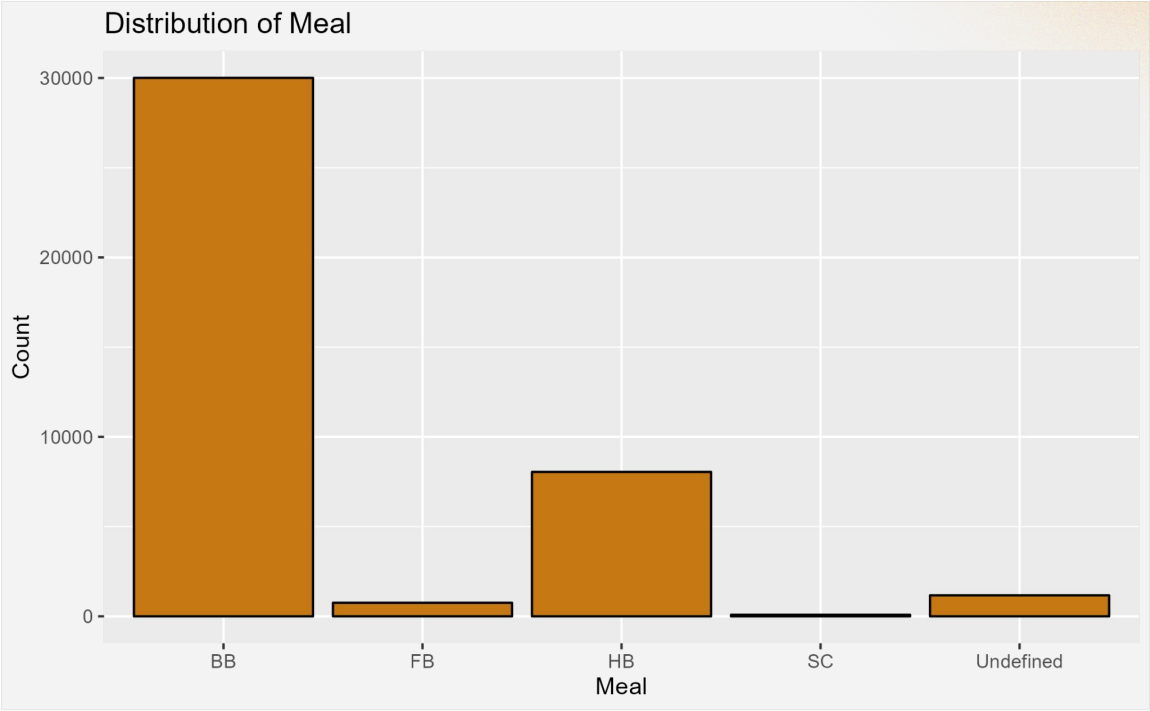
Bookings canceled have longer lead time, with more variation



Meals

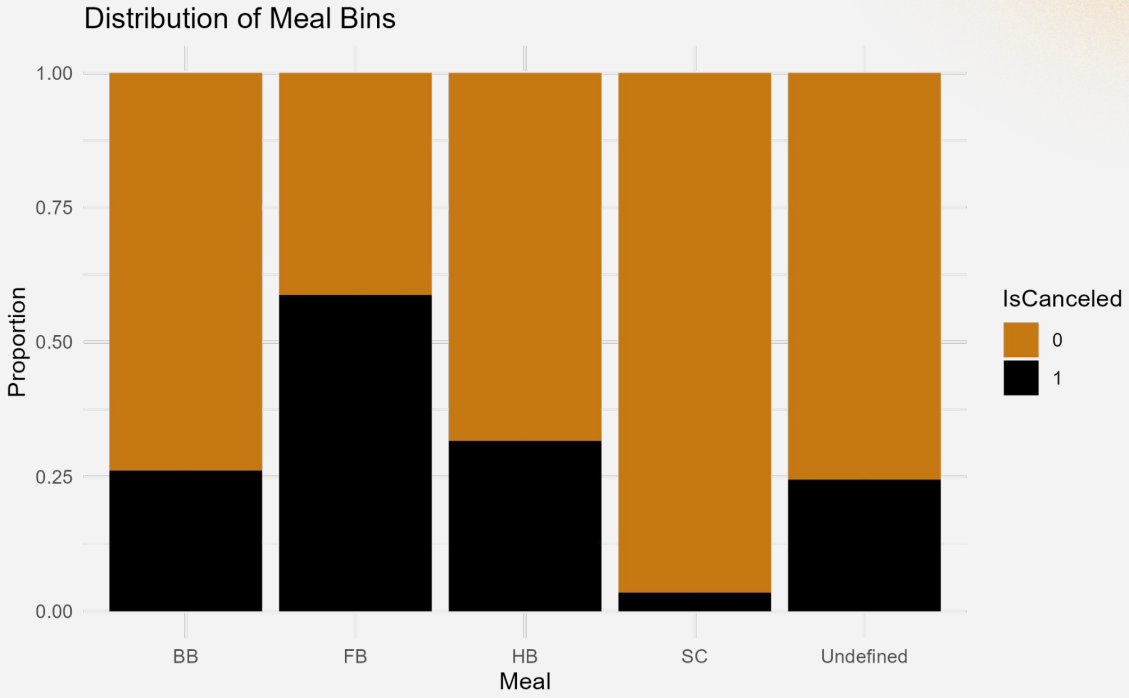
Bed & Breakfast is the most popular meal package

Transformation- ***Factorize***



Meals

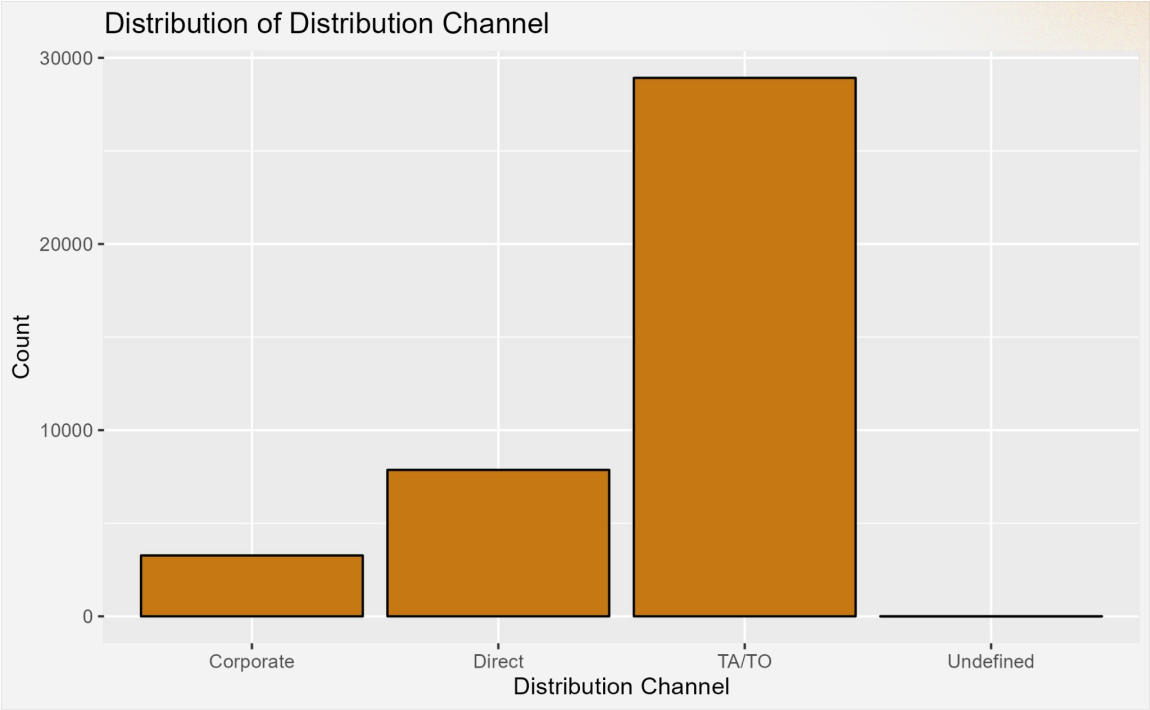
Variation in cancellation rates



DistributionChannel

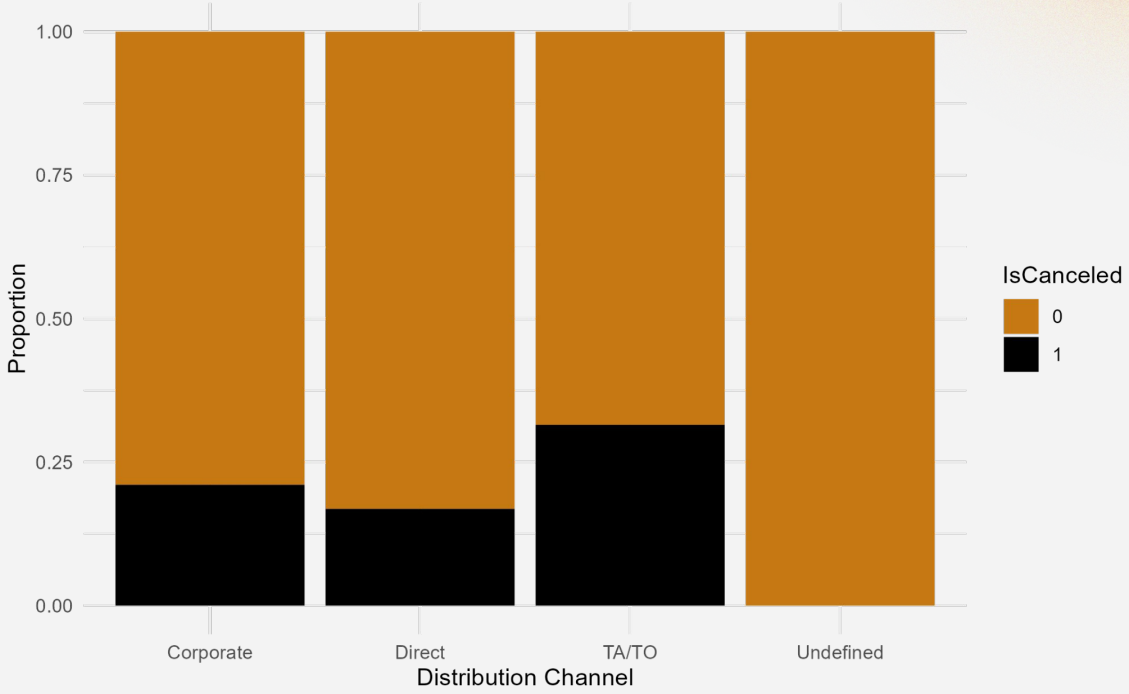
Travel agencies and tour operators are dominant

Transformation- ***Factorize***



DistributionChannel - Bins

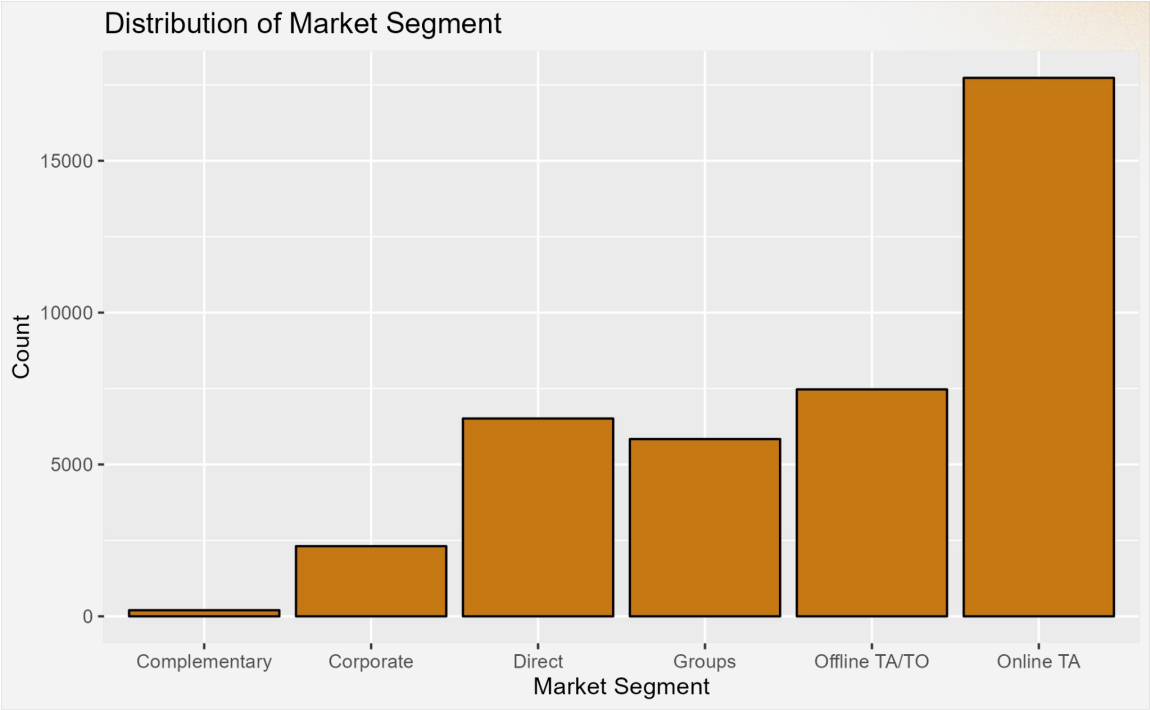
Variation across categories



MarketSegment

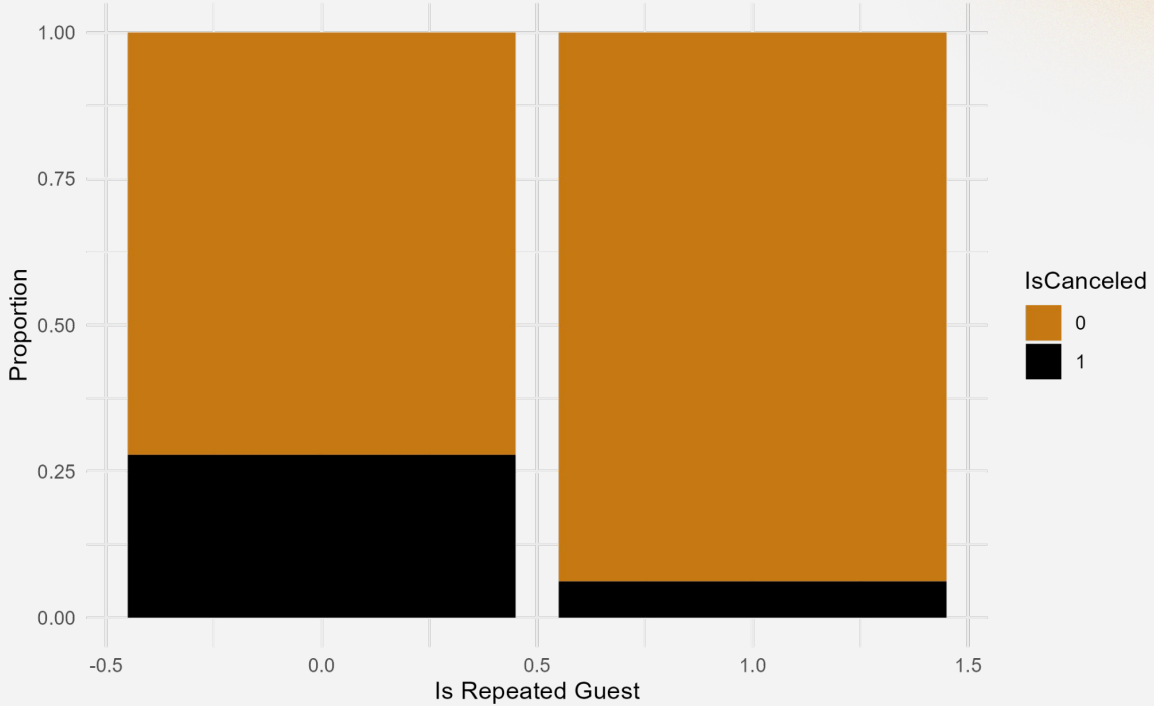
Behaves similarly as Distribution Channel

Dropped



IsRepeatedGuest - Bins

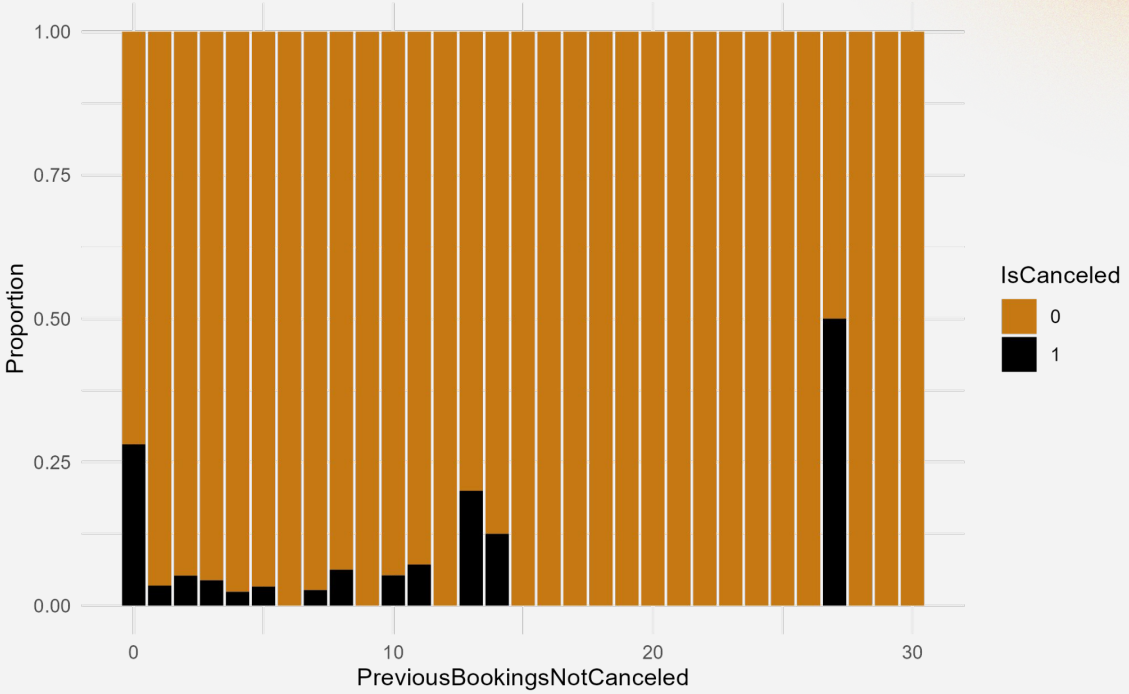
History of low previous bookings



PreviousBookingsNotCanceled

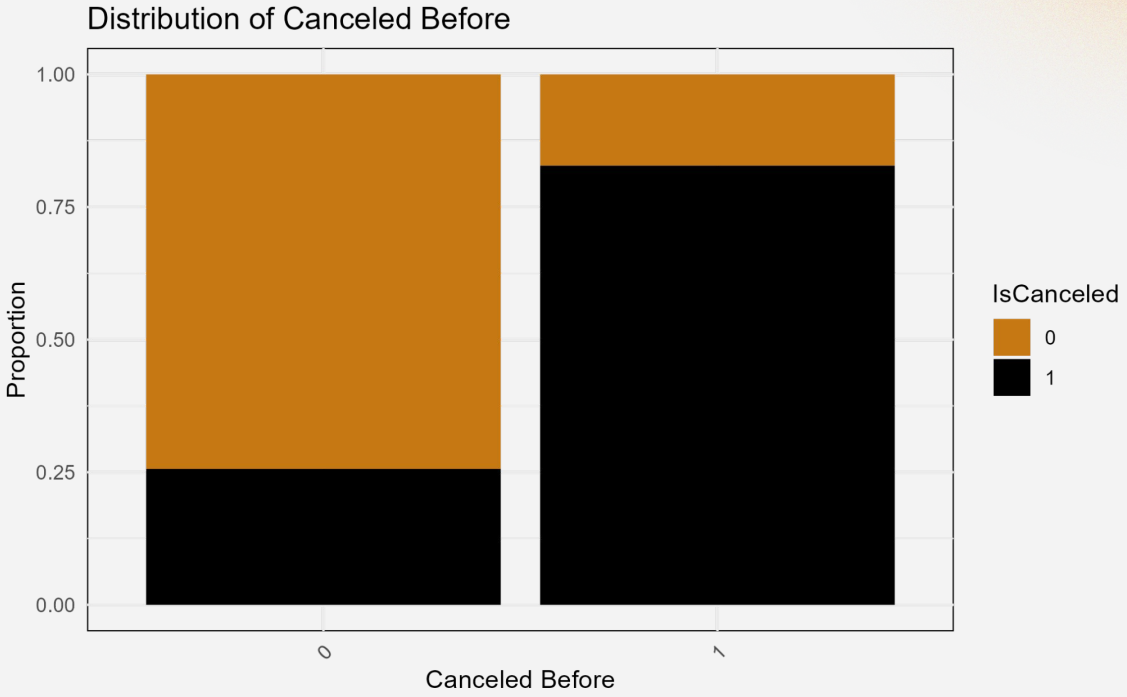
History of low previous cancellations

Transformation- ***Bins- 0,1***



CanceledBefore - Bins

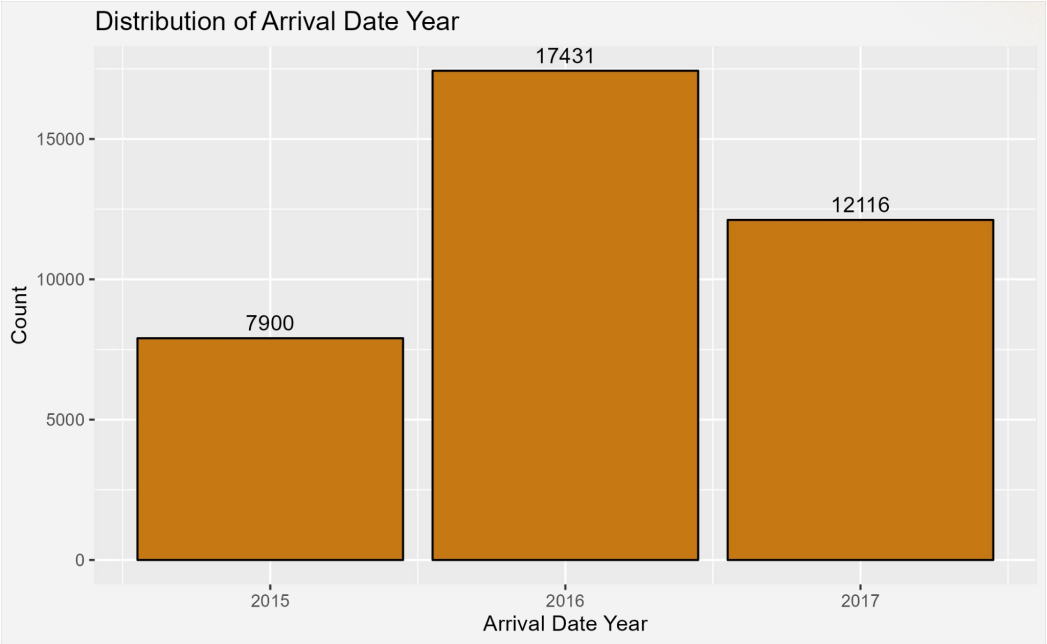
Variations across categories



ArrivalDateYear

Fluctuation across years

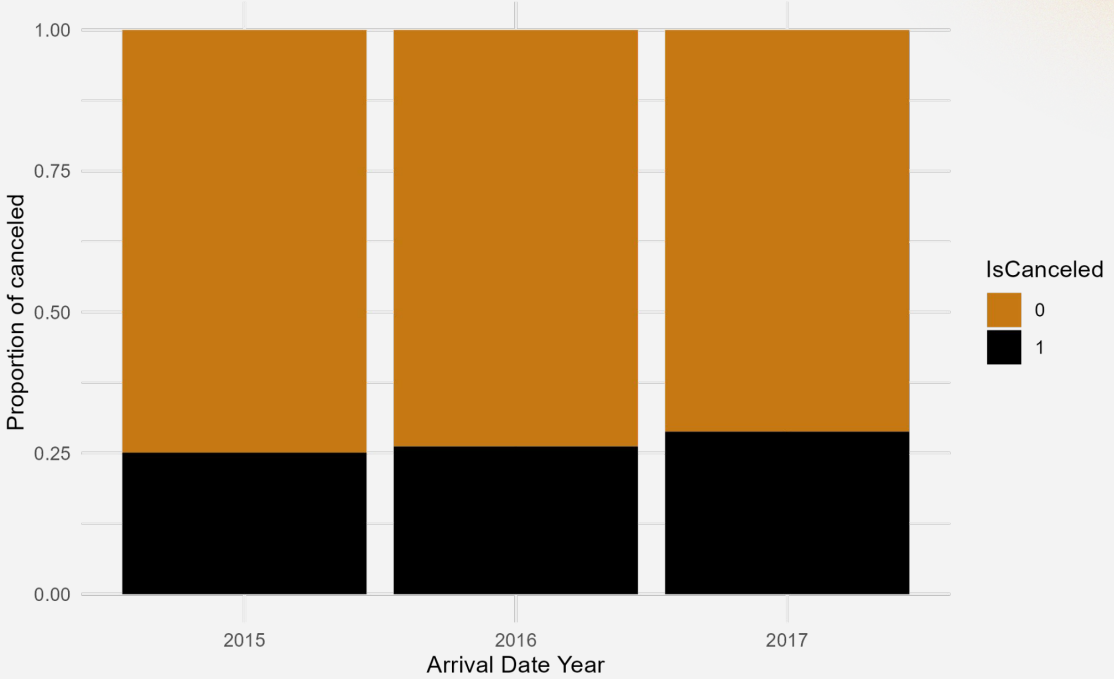
Transformation- ***Factorize***



ArrivalDateYear - Bins

Proportion is consistent across the years

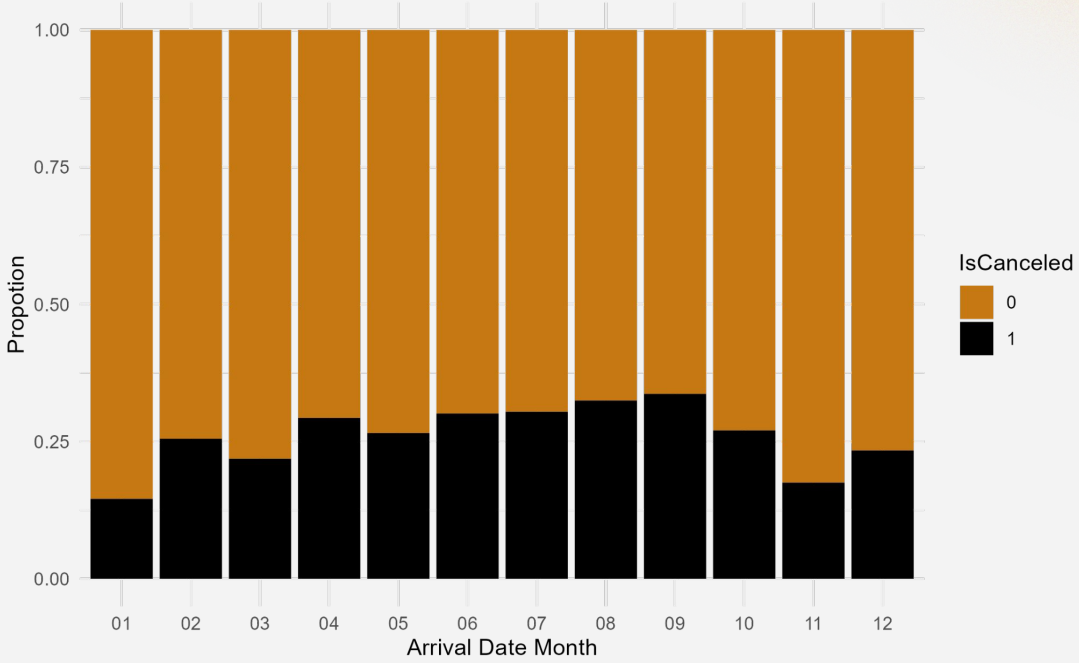
Dropped



ArrivalDateMonth

Seasonal variation

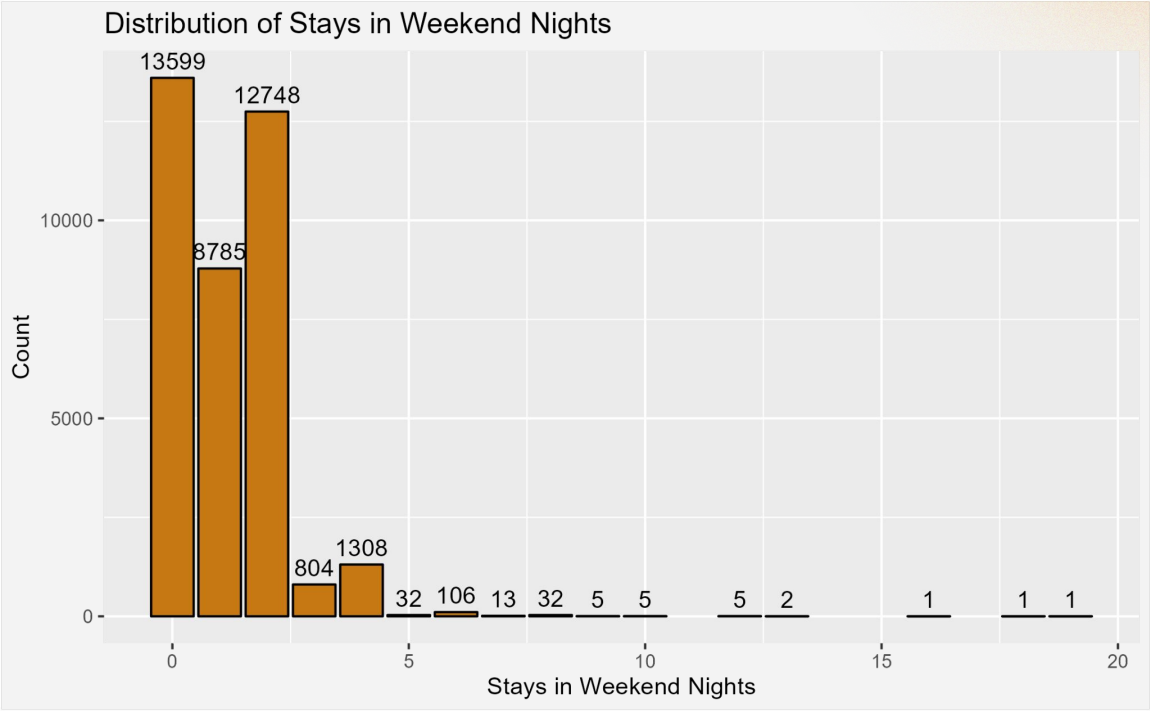
Transformation- ***Factorize***



StaysInWeekendNights

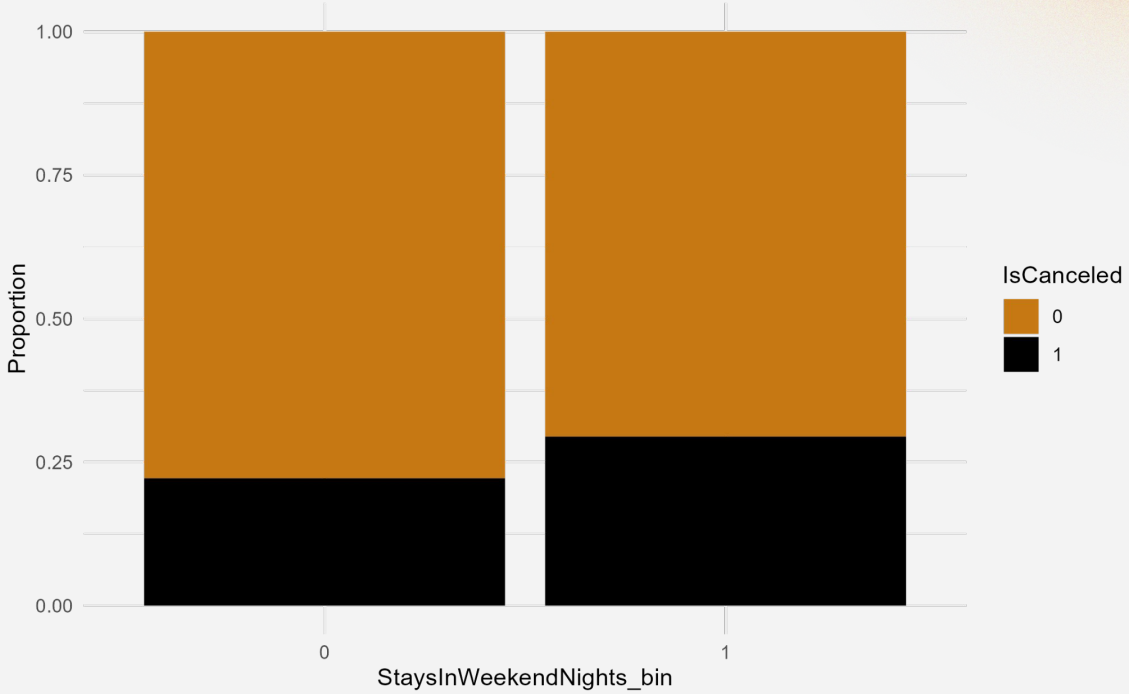
Few or no weekend nights, with rare long stays

Transformation- ***Bins- 0,1***



StaysInWeekendNights - Bins

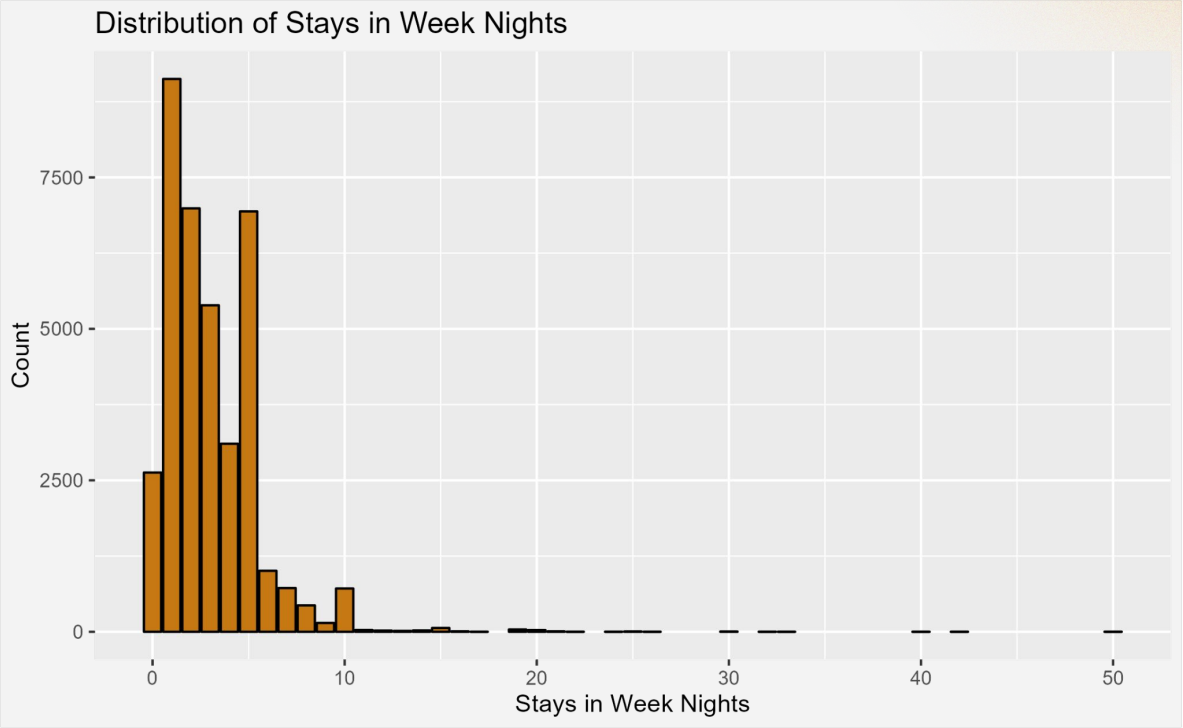
Variation across categories



StaysInWeekNights

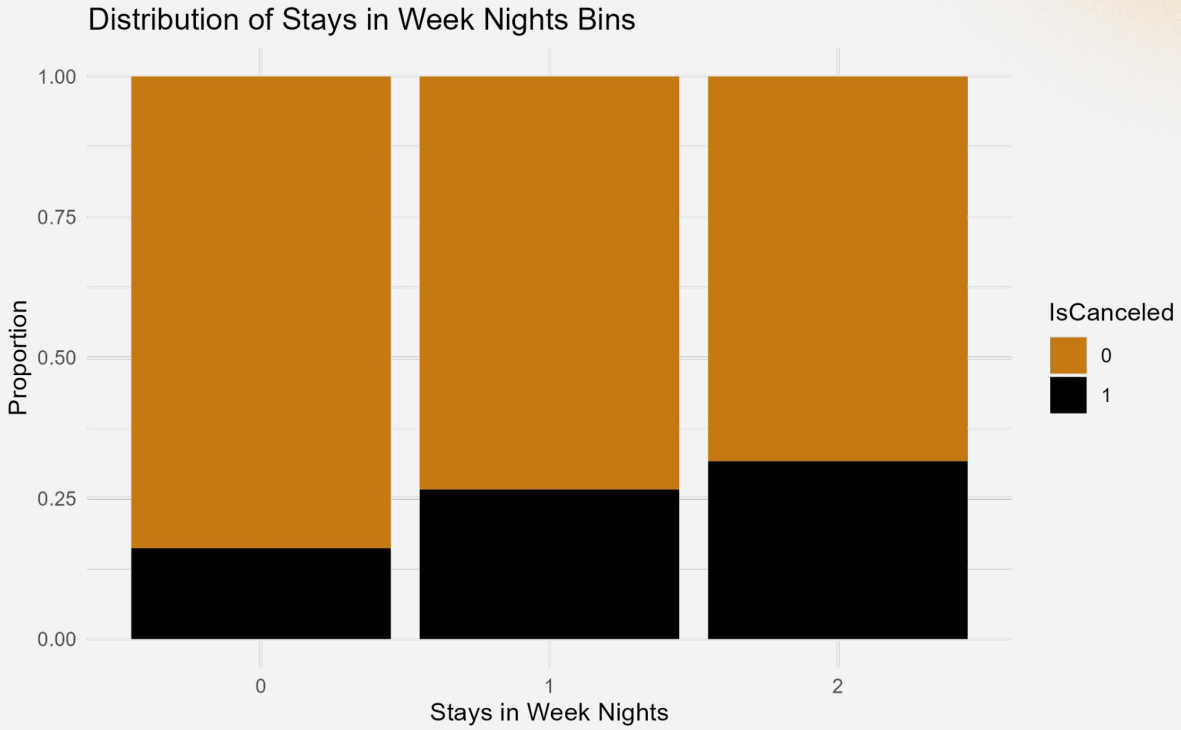
Right-skewed with few weeknights

Transformation- ***Bins- 0,1,2***



StaysInWeekNights - Bins

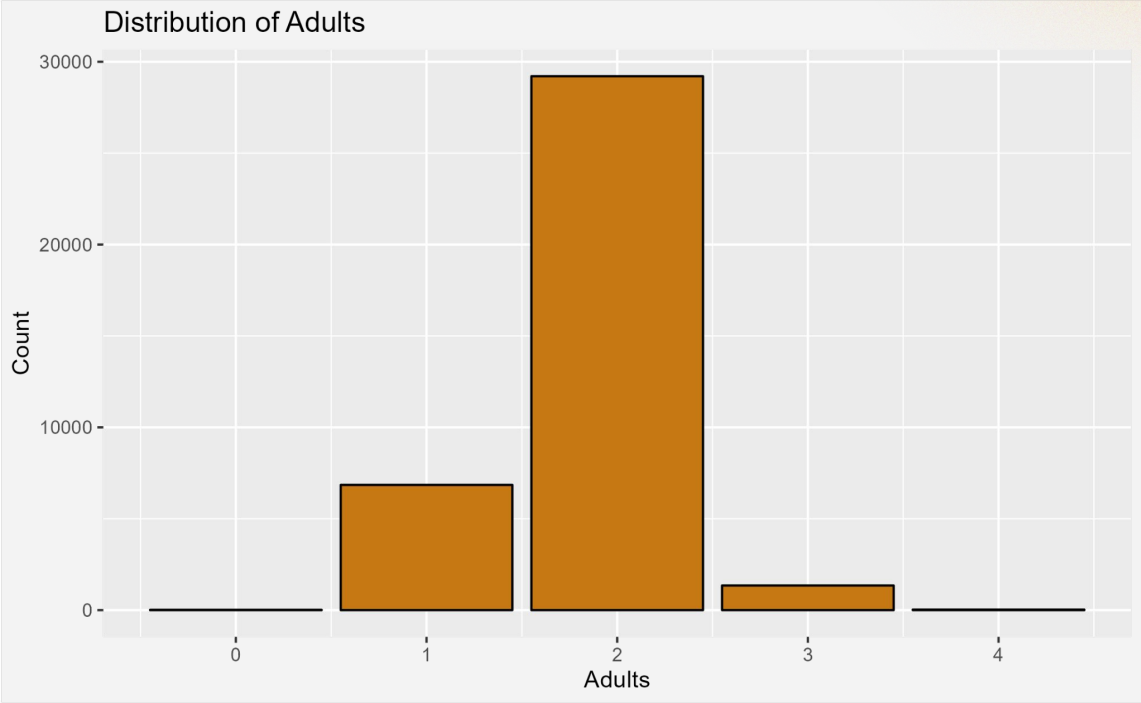
Variation across categories



Adults

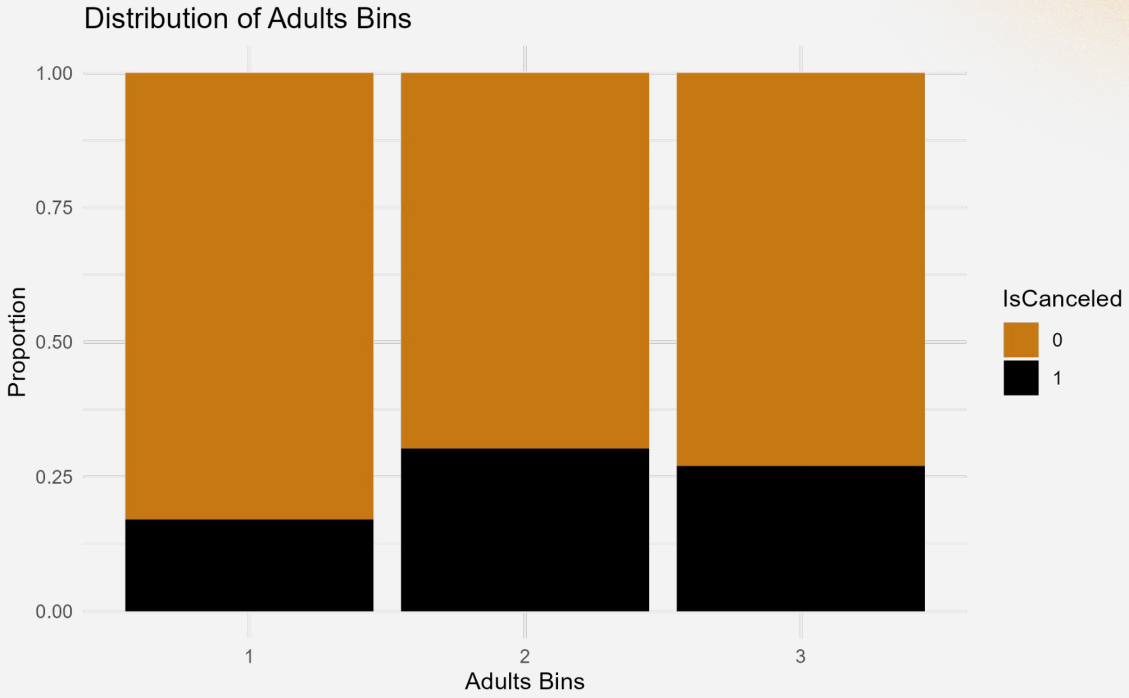
Most bookings are made for two adults

Transformation- ***Bins- 1,2,3***



Adults - Bins

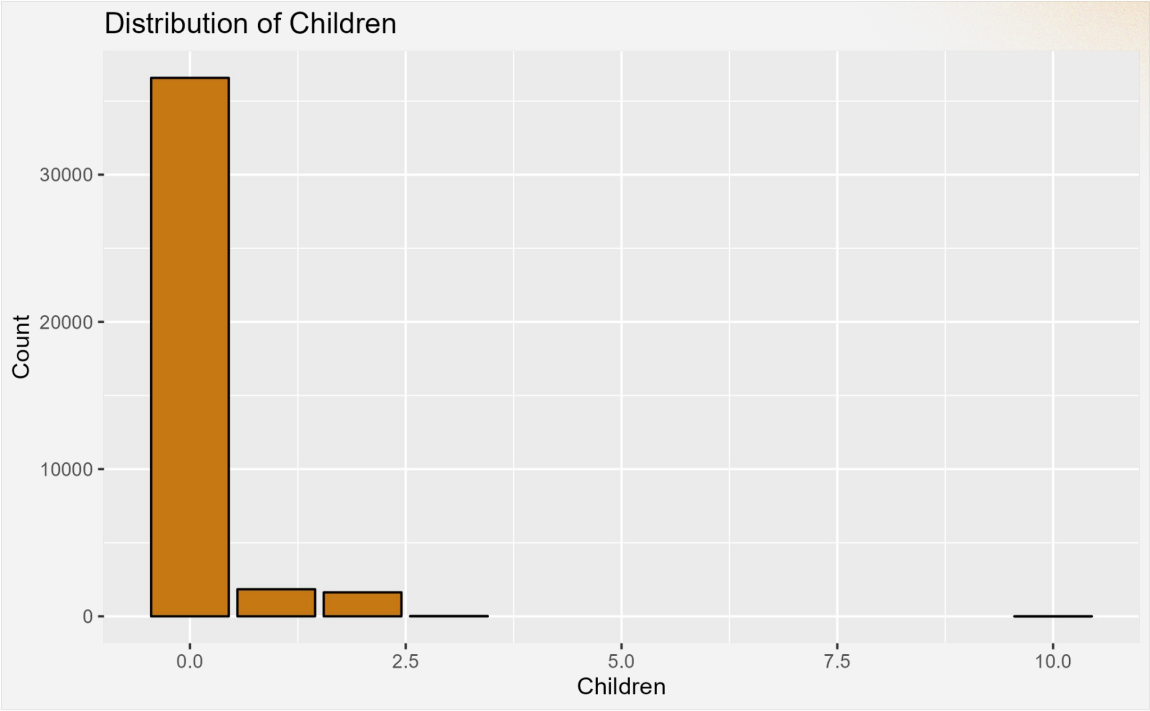
Variation across categories



Children

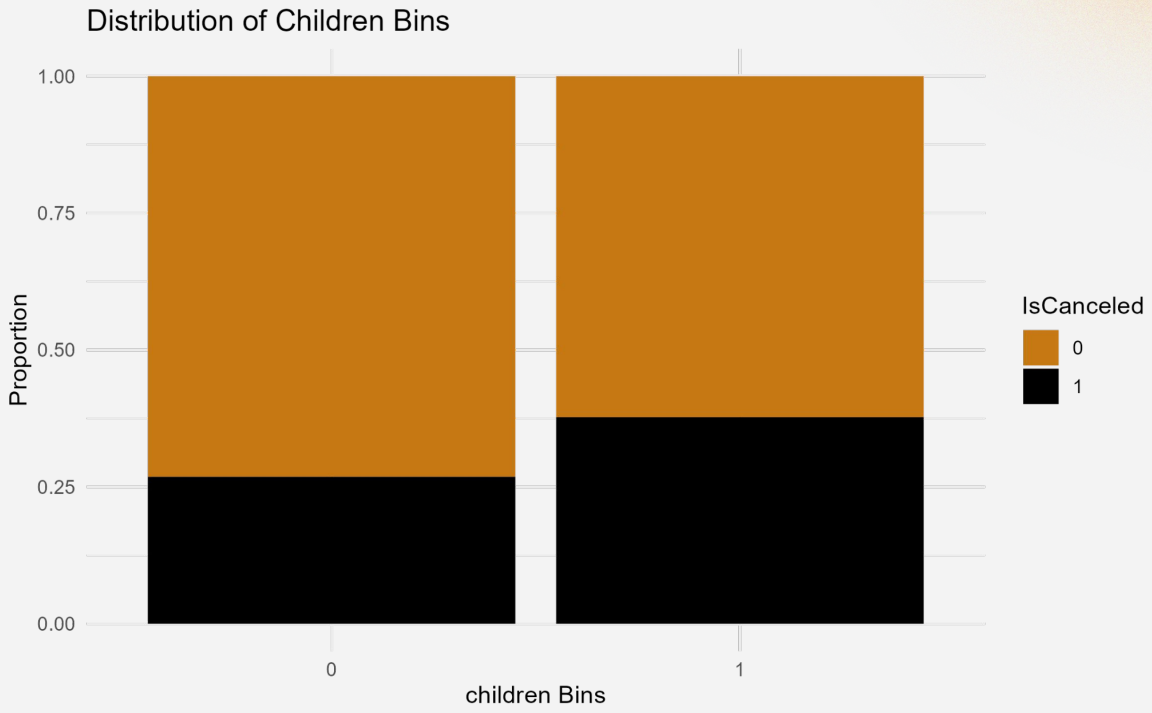
Right skewed, with very few for one or more children

Transformation- ***Bins- 0,1***



Children - Bins

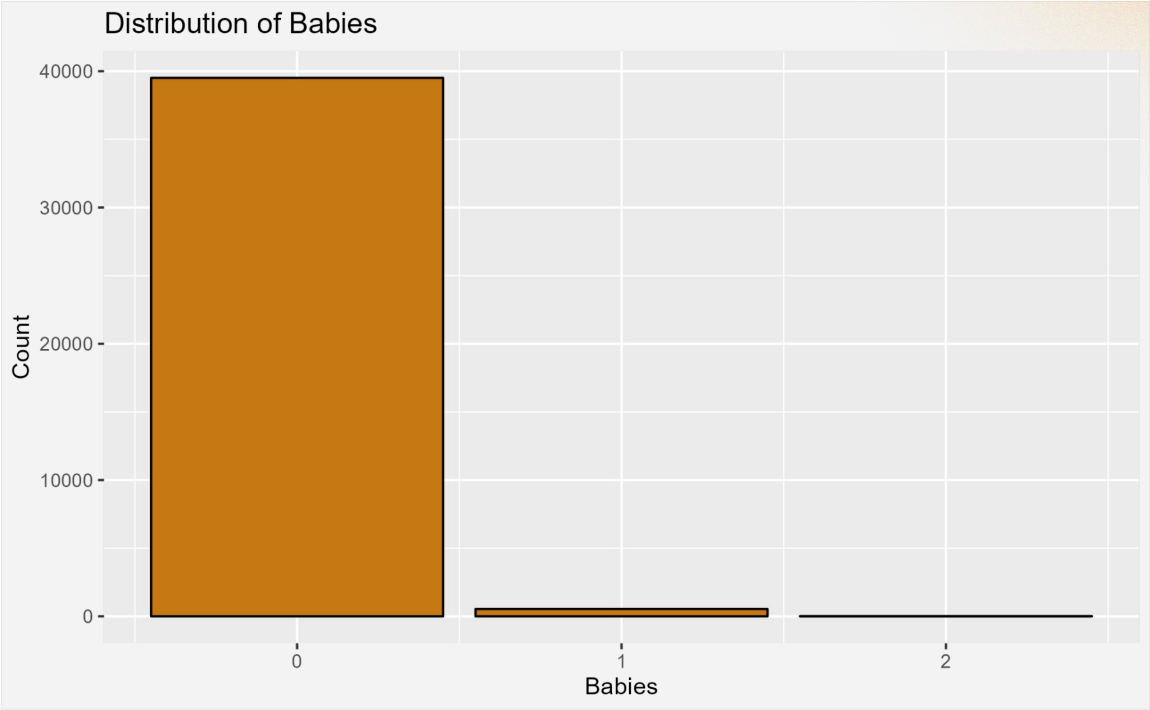
Variation across categories



Babies

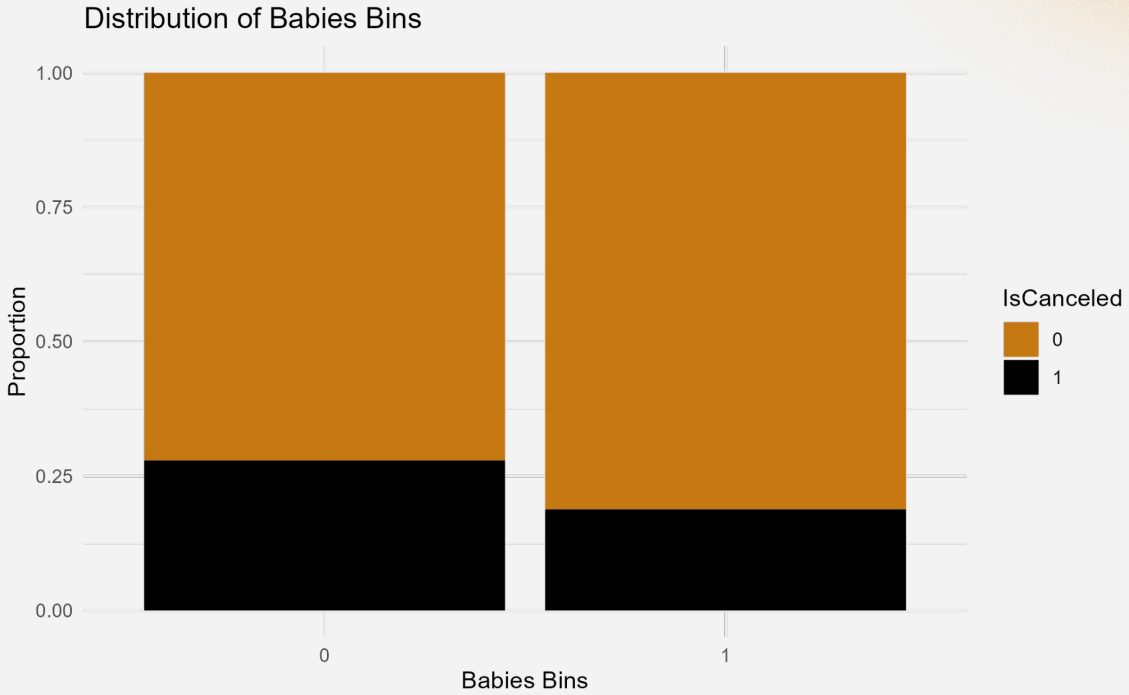
Right skewed, with very few for one or more babies

Transformation- ***Bins- 0,1***



Babies - Bins

Variation across Categories



Country

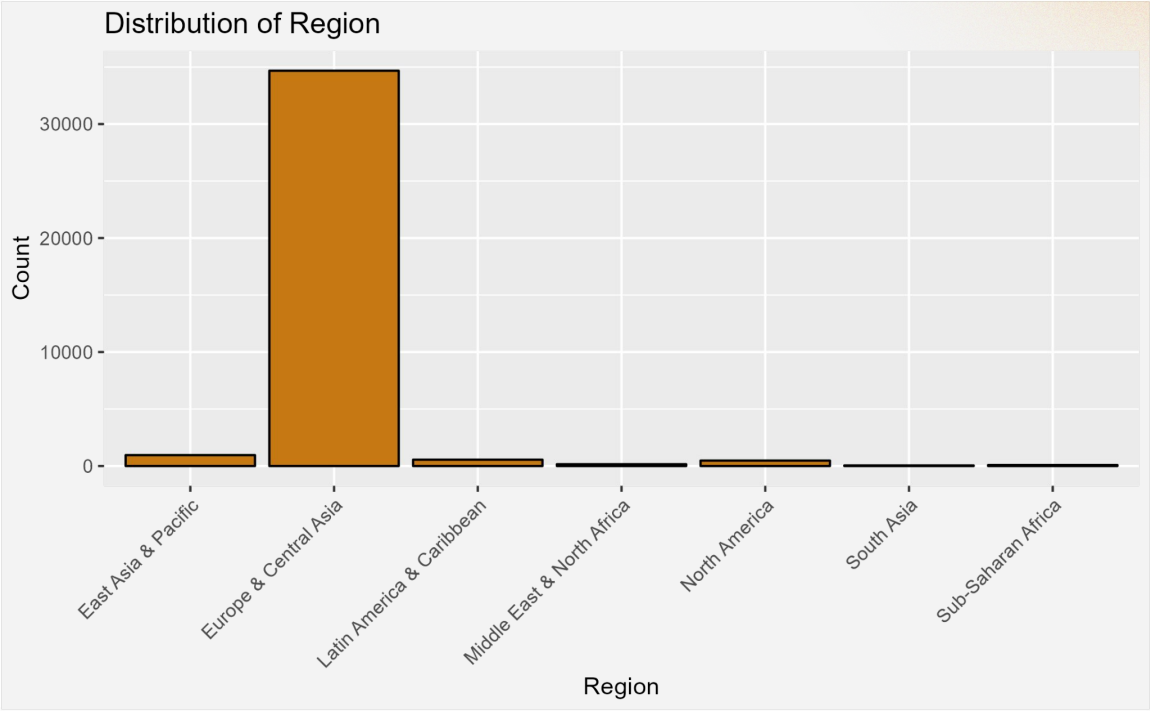
Unique country codes: 126

Transformation- Merged with World Bank data and Adjusted country codes

Region

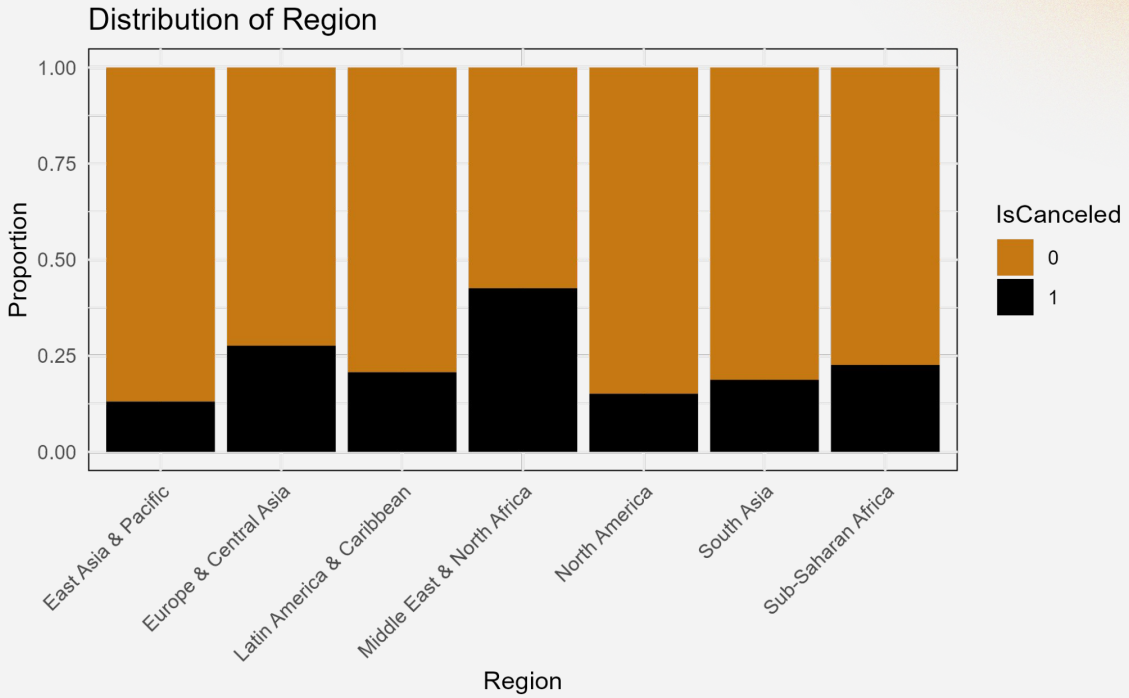
Most bookings come from Europe & Central Asia.

Transformation- ***Factorize***



Region - Bins

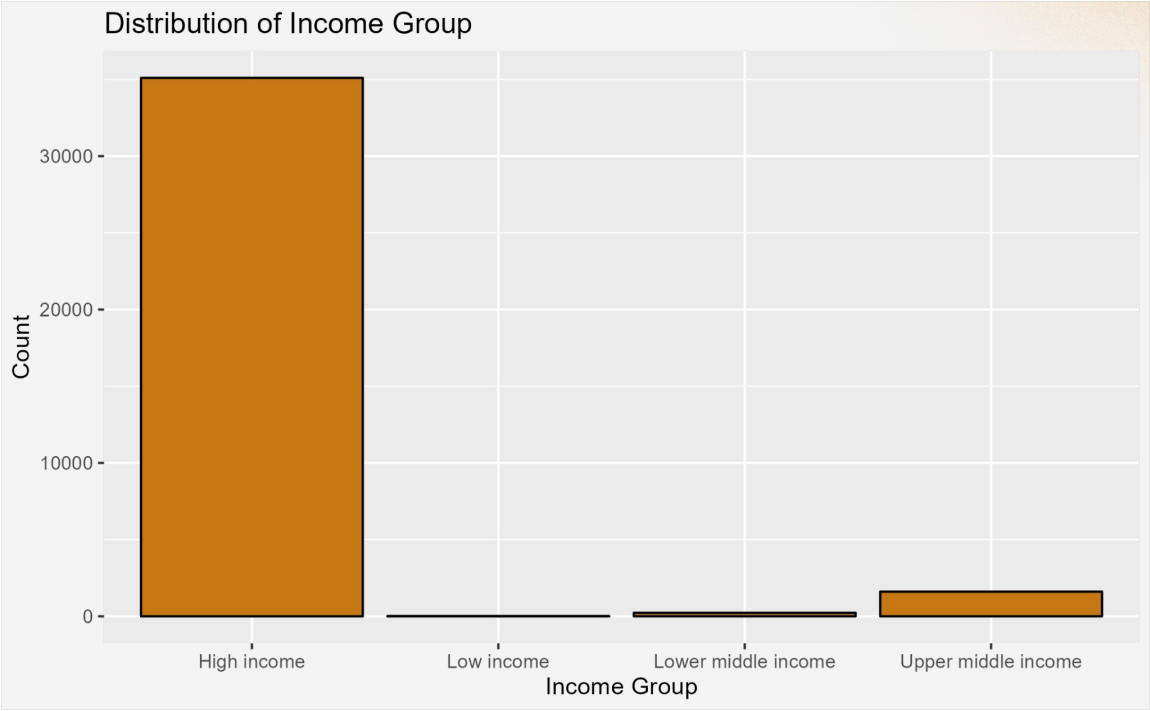
Variations across regions



IncomeGroup

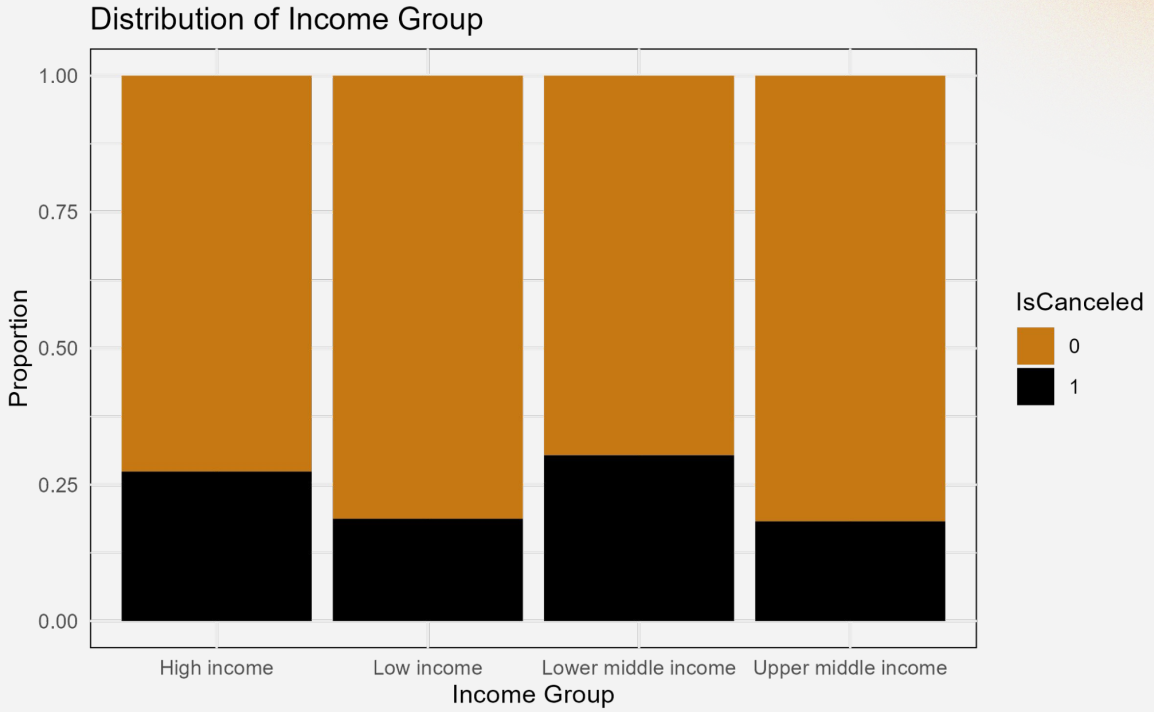
Most bookings come from high-income regions

Transformation- ***Factorize***



IncomeGroup - Bins

Variation across income groups



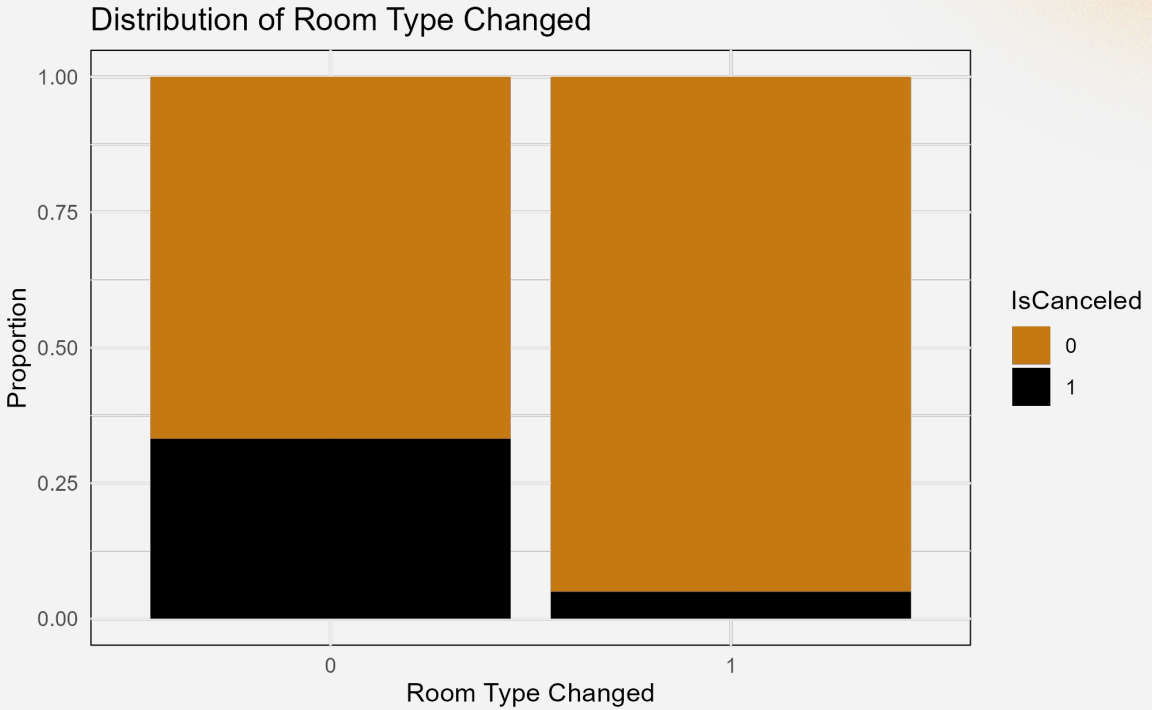
ReservedRoomType + AssignedRoomType

Transformation- Computed difference between the columns,
Create a new variable 'RoomTypeChanged'

RoomTypeChanged - Bins

Variation across categories

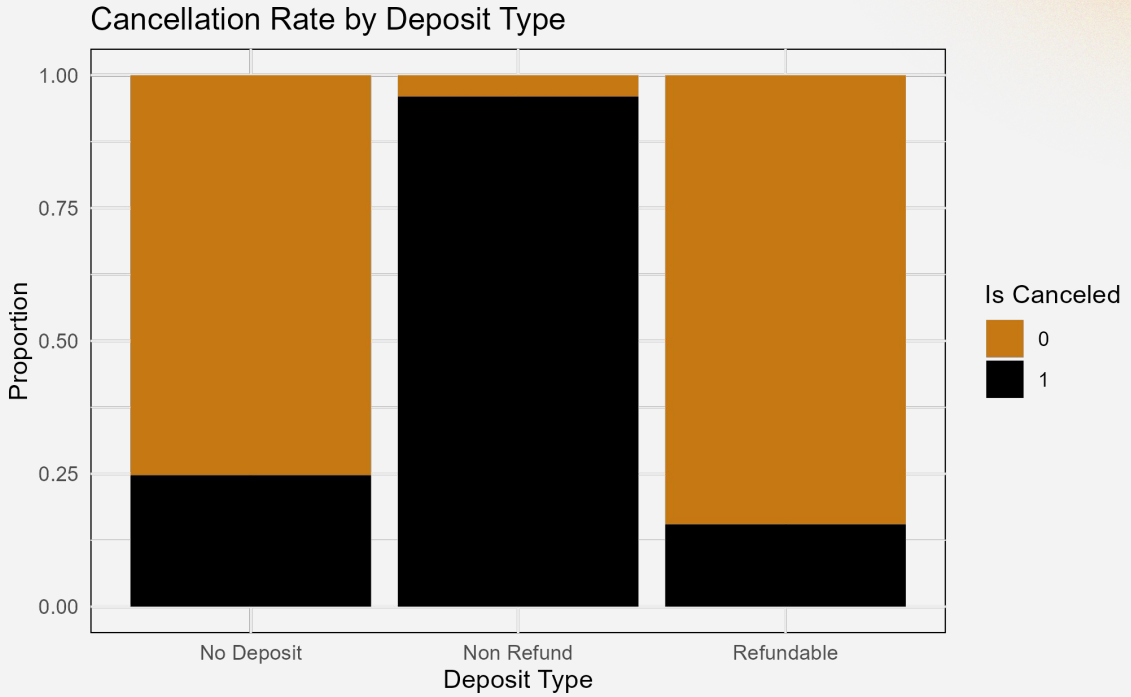
Transformation- ***Factorize***



DepositType - Bins

Bookings with no refund have the highest cancellation rates

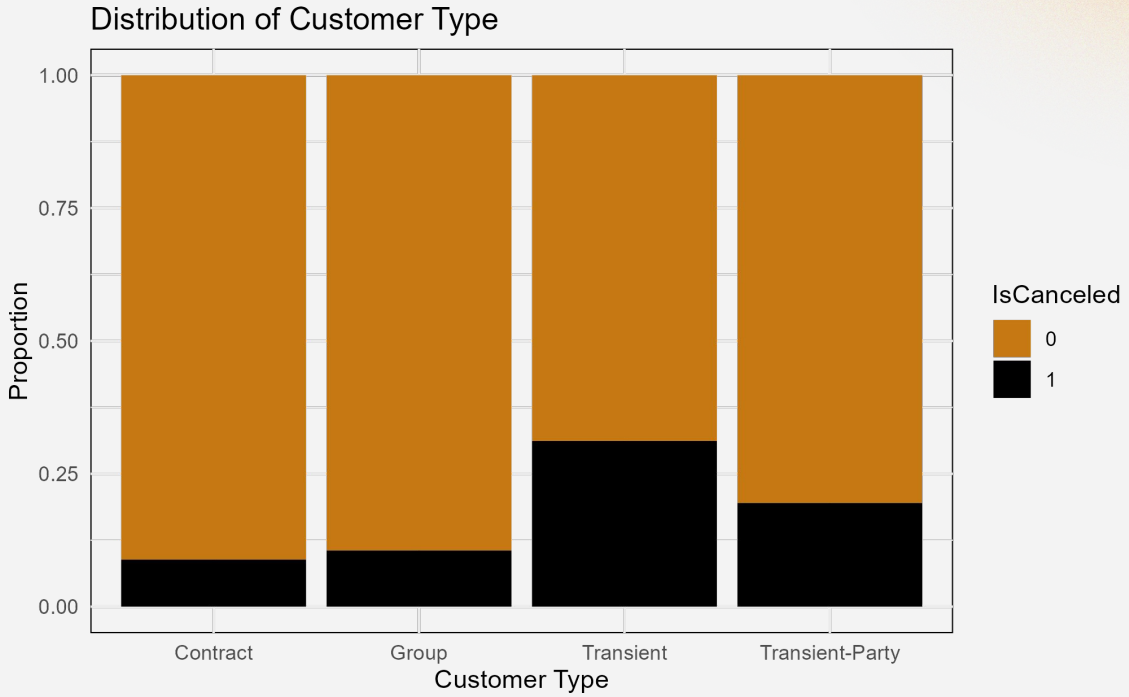
Transformation- ***Factorize***



CustomerType - Bins

Variations across categories

Transformation- ***Factorize***



05
Model Fitting

Logistic Regression

Logistic Regression

Binary Classification

Probability Estimation

GLM: Pipeline

Numerical Transformer

Categorical Transformer

GLM: Pipeline

Numerical Transformer

Imputer (median)

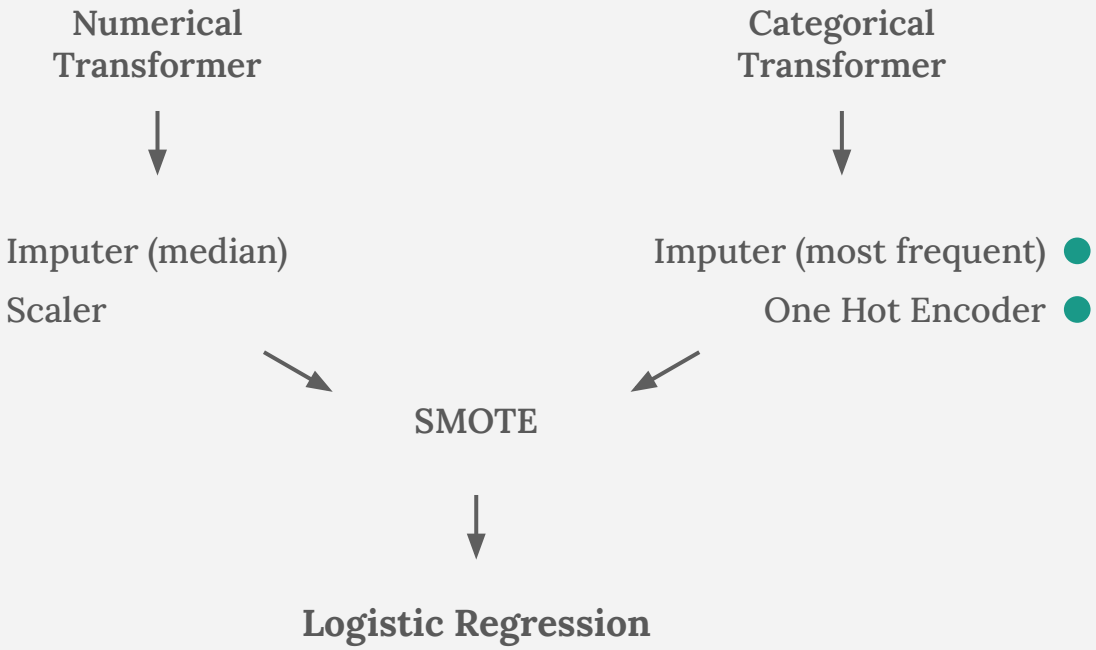
Scaler

Categorical Transformer

Imputer (most frequent)

One Hot Encoder

GLM: Pipeline



Dataset splits

Training & Validation Set

Testing Set

Model 1: Baseline

IsCanceled

~

All Variables

Model 1: Performance

IsCanceled
~
All Variables

66%
Accuracy

41%
Precision

75%
Recall

Model 2: Step AIC

IsCanceled

~

Backward Selection

Model 2: Performance

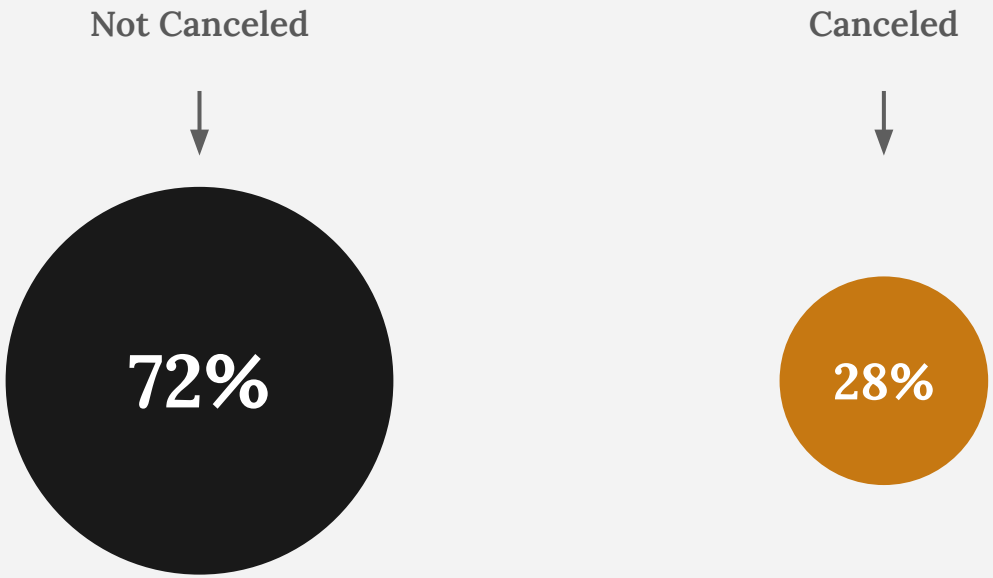
IsCanceled
~
Backward Selection

68%
Accuracy

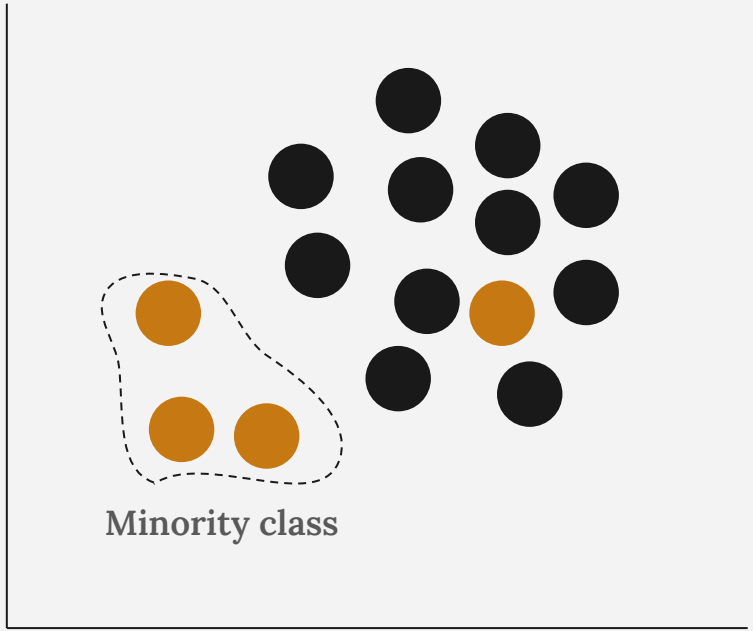
43%
Precision

77%
Recall

Addressing Class Imbalance



SMOTE: Identify Minority Cases



SMOTE: Synthetic Generation

Model 3: SMOTE + AIC

Model 3: SMOTE

IsCanceled
~
Forward Selection

80%
Accuracy

67%
Precision

42%
Recall

GLM: Model Evaluation

Baseline vs. **SMOTE + AIC**

GLM: Model Evaluation

Accuracy

66% → **80%**

GLM: Model Evaluation

Precision

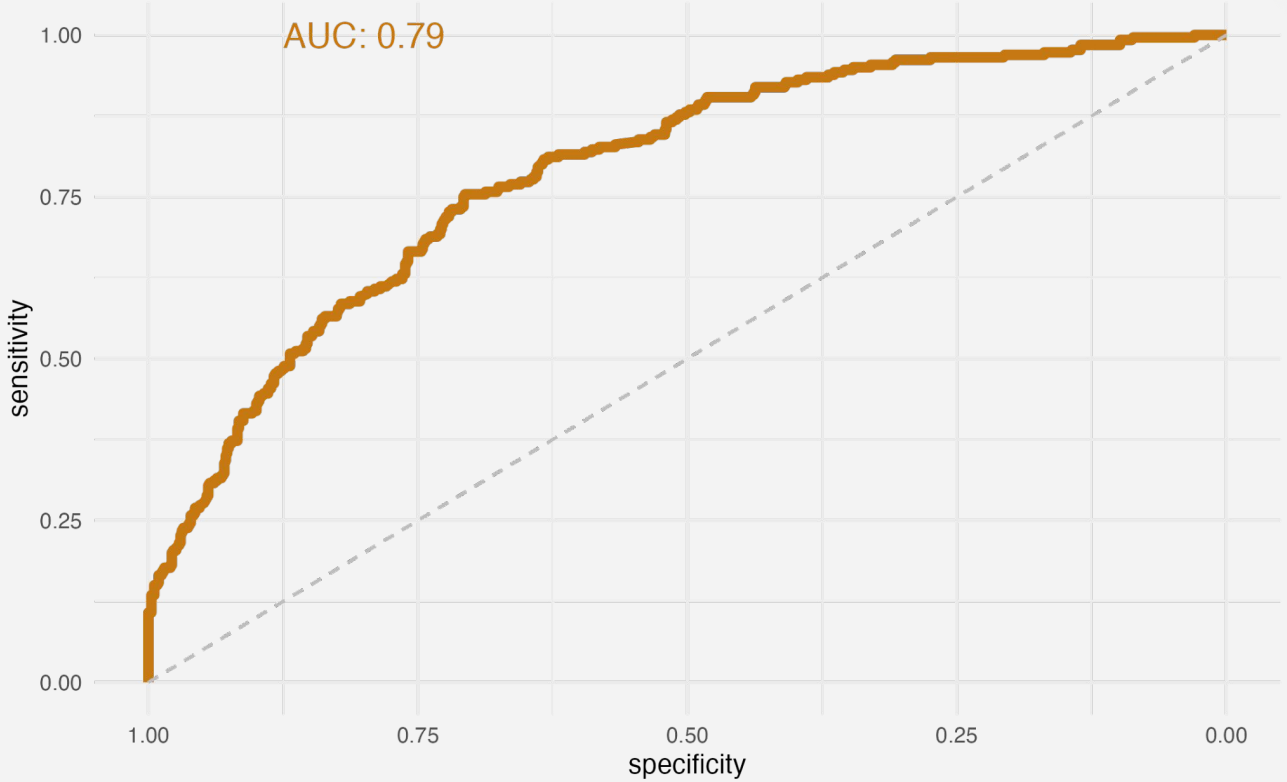
41% → 67%

GLM: Model Evaluation

Recall

75% → **42%**

GLM: Model Evaluation



K-Nearest Neighbors

K-Nearest Neighbors

Binary Classification

Probability Estimation

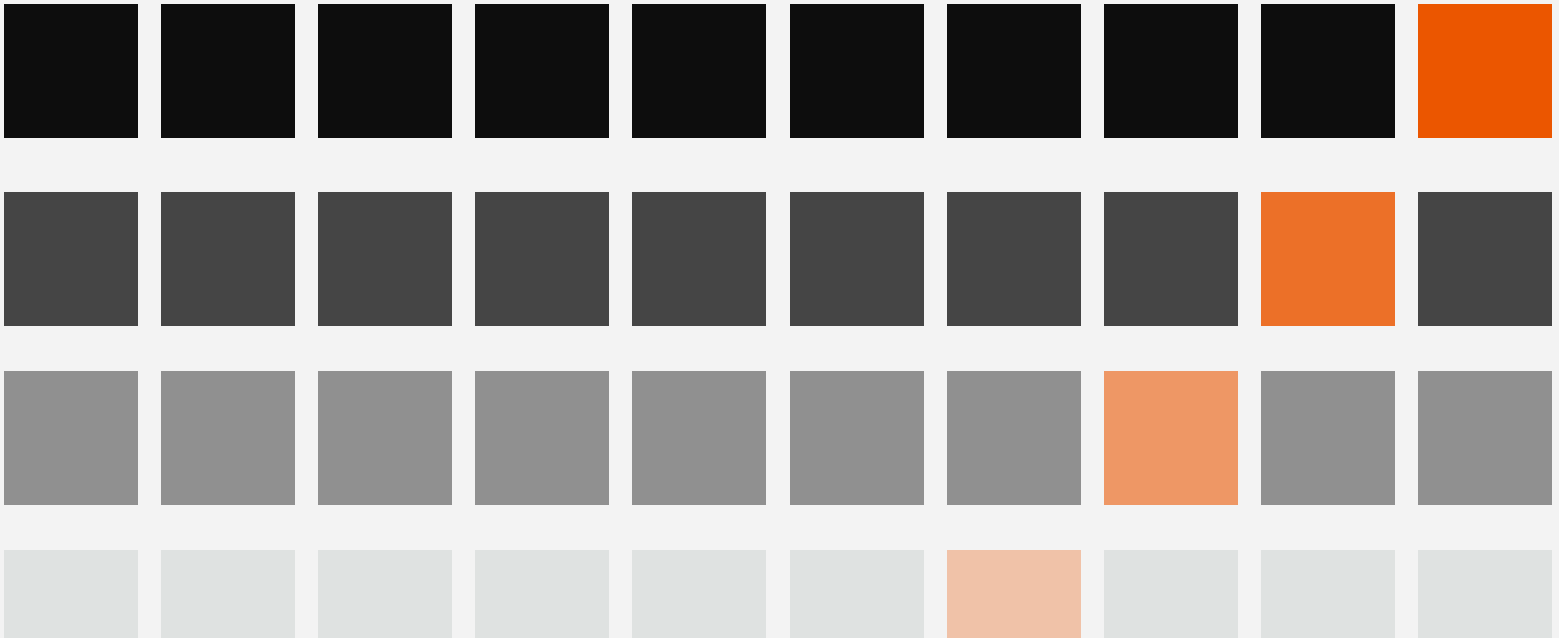
Dataset splits

Training & Validation Set

Testing Set

Cross-Validation

K = 1 : 15



Model 1: Baseline

IsCanceled

~

All Variables

Model 1: Performance

IsCanceled
~
All Variables

73%
Accuracy

47%
Precision

22%
Recall

Model 2: SMOTE

Model 2: Performance

IsCanceled
~
All Variables

66%
Accuracy

40%
Precision

65%
Recall

KNN: Model Evaluation

Baseline vs. **SMOTE**

KNN: Model Evaluation

Accuracy

73% → **66%**

KNN: Model Evaluation

Precision

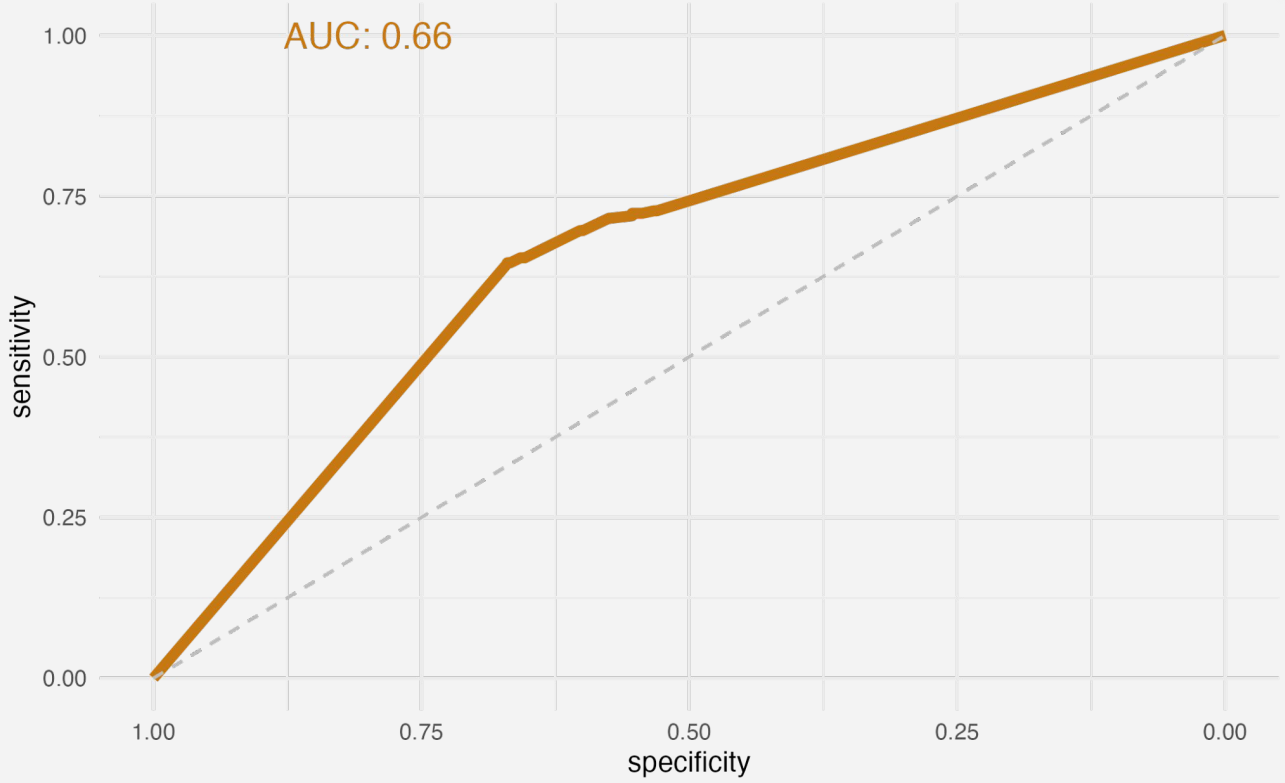
47% → **40%**

KNN: Model Evaluation

Recall

22% → **65%**

KNN: Model Evaluation



KNN: Curse of dimensionality

20 Features

Random Forest Classifier

Random Forest Classifiers

Ensemble Learning

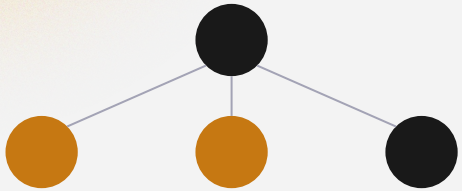
Immune to Overfitting

Random Forest Classifiers

Sample

Random Forest Classifiers

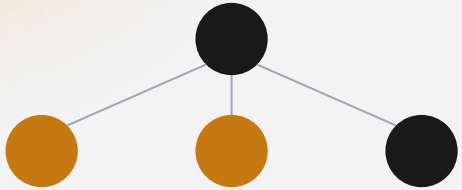
Sample



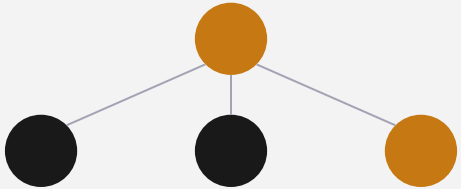
Tree #1

Random Forest Classifiers

Sample



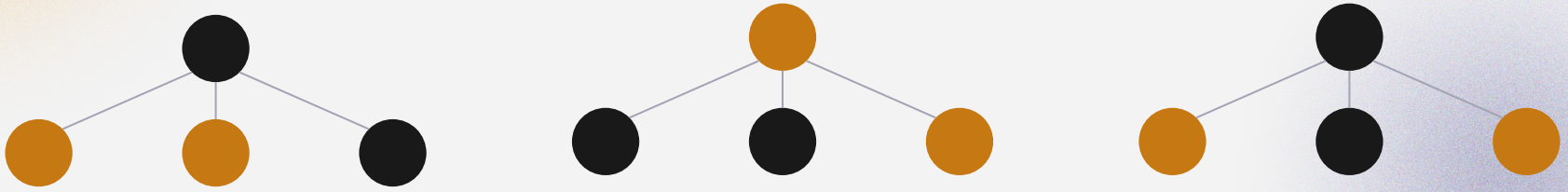
Tree #1



Tree #2

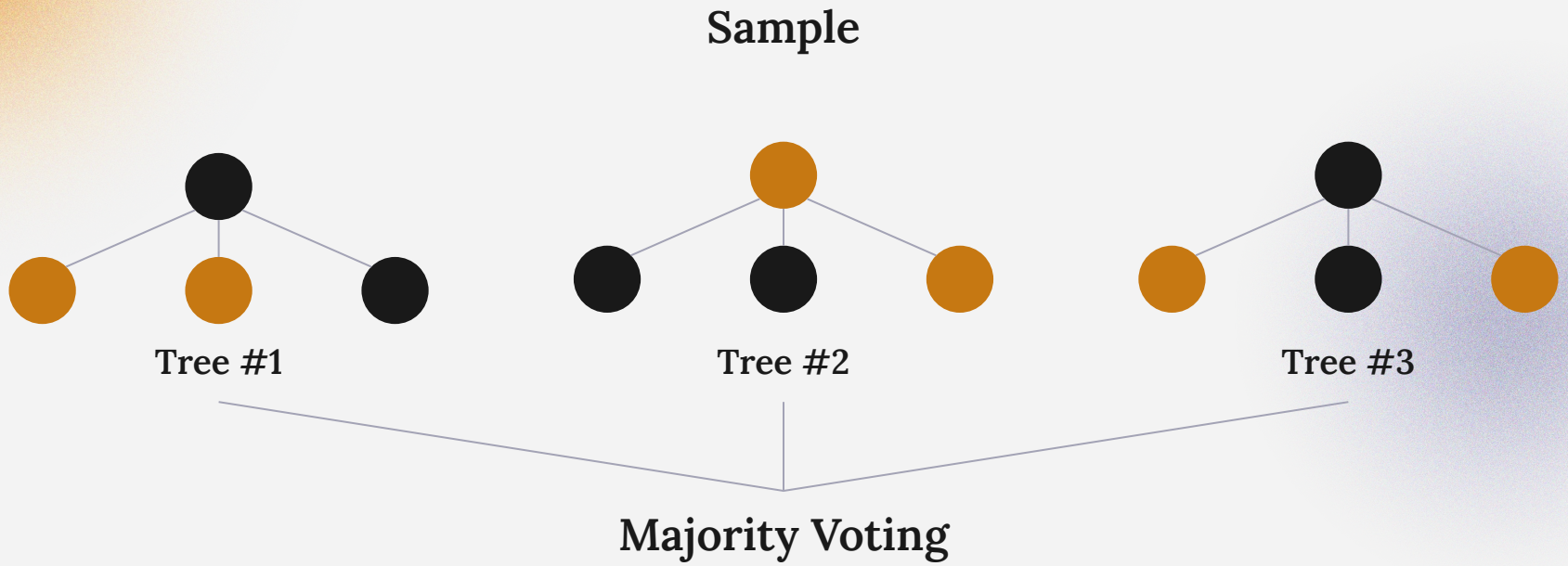
Random Forest Classifiers

Sample



Tree #1 Tree #2 Tree #3

Random Forest Classifiers



Model 1: Baseline

IsCanceled

~

All Variables

Model 1: Performance

IsCanceled
~
All Variables

78%
Accuracy

56%
Precision

57%
Recall

Model 2: Hypertuning

Grid CV

~

Min Node Size
Split Rule
Mtry

Model 2: Performance

IsCanceled
~
All Variables

80%
Accuracy

77%
Precision

33%
Recall

Model 3: SMOTE + Tuning

Model 3: Performance

73%
Accuracy

48%
Precision

83%
Recall

RF: Model Evaluation

Accuracy

78% → 73%

RF: Model Evaluation

Precision

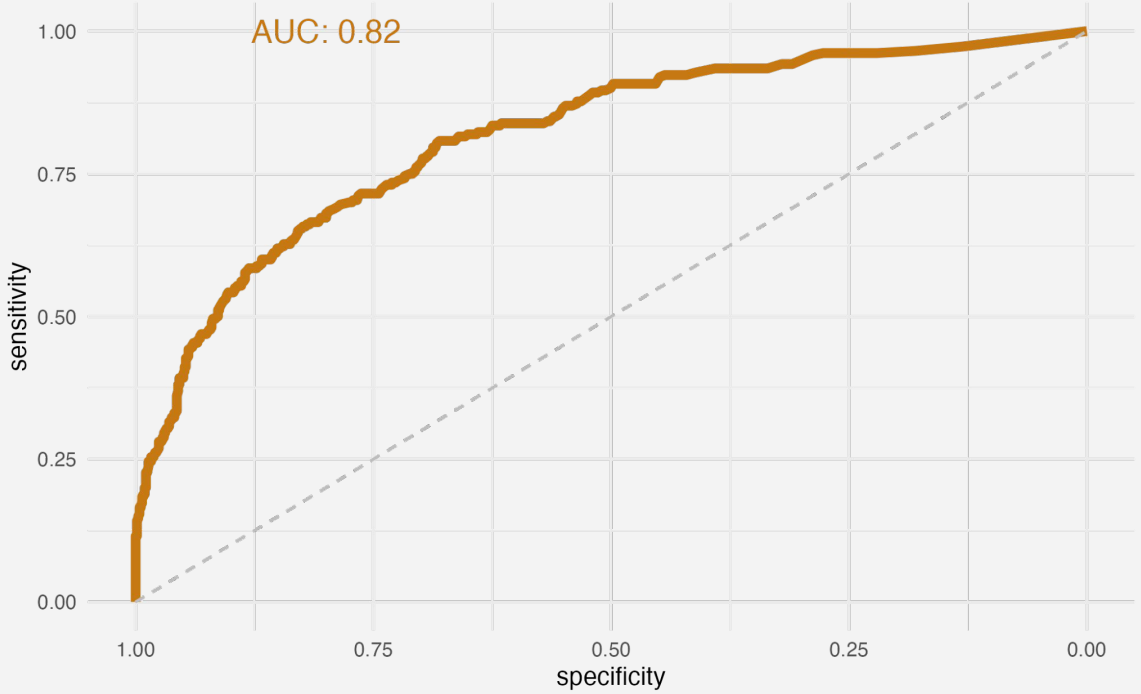
56% → 48%

RF: Model Evaluation

Recall

57% → 83%

RF: Model Evaluation



06

Model Evaluation

Evaluation Criteria

Accuracy

Recall

Precision

Accuracy

Random Forest Classifier	81%
Logistic Regression	80%
K-Nearest Neighbors	62%

Recall

Random Forest Classifier	83%
Logistic Regression	67%
K-Nearest Neighbors	38%

Precision

Logistic Regression	66%
Random Forest Classifier	48%
K-Nearest Neighbors	47%

7
Model Selection

Trade-offs

Precision vs. Sensitivity

Trade-offs

Lowering Threshold - Higher Sensitivity

Trade-offs

Raising Threshold - **Lower** Sensitivity

Trade-offs

Cancellation Cost vs. Overbooking Cost

Trade-offs

If cancellations are costly

•

Optimize for Sensitivity

Trade-offs

If cancellations are costly

•

Use **Random Forest Model**

Trade-offs

If overbooking is costly

•

Optimize for Precision

Trade-offs

If overbooking is costly

•

Use Logistic Regression Model

8 Conclusions

Model Utility

Better Inventory Management

Model Utility

More Confidence in **Overbooking**

Model Utility

Inference on **Cancellation Patterns**

Future Scope

- Building a neural network
- Make inference on the reasons behind cancellations based on logistic regression coefficients

Thank you