# Hotel Booking Cancellation Prediction

Data Preprocessing Pipeline

***The dataset initially shows no missing values in any field—a rare but ideal scenario for real-world data! This suggests strong data collection practices or prior cleaning. However, we still implemented safeguards in our pipeline to handle potential future inconsistencies."***

```
Missing Values After Handling:
Booking_ID                  0
number of adults            0
number of children          0
number of weekend nights    0
number of week nights       0
type of meal                0
car parking space           0
room type                   0
lead time                   0
market segment type         0
repeated                    0
P-C                         0
P-not-C                     0
average price               0
special requests            0
date of reservation         0
booking status              0
dtype: int64
```

```
Missing Values Before Handling:
Booking_ID                  0
number of adults            0
number of children          0
number of weekend nights    0
number of week nights       0
type of meal                0
car parking space           0
room type                   0
lead time                   0
market segment type         0
repeated                    0
P-C                         0
P-not-C                     0
average price               0
special requests            0
date of reservation         0
booking status              0
dtype: int64
```

"The dataset contains **zero duplicate entries**—both before and after preprocessing—indicating highly clean and unique booking records. With **36,285 entries** and **17 features**, we have a robust, non-redundant dataset for training our cancellation prediction model."

```
Number of duplicates before: 0
Number of duplicates after: 0
New shape: (36285, 17)
```

***"Normalization standardizes features for fair model training.***

Data after outlier treatment:

| | number of adults | number of children | number of weekend nights |
|---|---|---|---|
| count | 36285.0 | 36285.0 | 36285.000000 |
| mean | 2.0 | 0.0 | 0.810087 |
| std | 0.0 | 0.0 | 0.867286 |
| min | 2.0 | 0.0 | 0.000000 |
| 25% | 2.0 | 0.0 | 0.000000 |
| 50% | 2.0 | 0.0 | 1.000000 |
| 75% | 2.0 | 0.0 | 2.000000 |
| max | 2.0 | 0.0 | 5.000000 |

| | number of week nights | lead time | average price | special requests |
|---|---|---|---|---|
| count | 36285.000000 | 36285.000000 | 36285.000000 | 36285.000000 |
| mean | 2.178145 | 83.767893 | 102.968399 | 0.606642 |
| std | 1.290708 | 81.662186 | 31.678904 | 0.746950 |
| min | 0.000000 | 0.000000 | 20.750000 | 0.000000 |
| 25% | 1.000000 | 17.000000 | 80.300000 | 0.000000 |
| 50% | 2.000000 | 57.000000 | 99.450000 | 0.000000 |
| 75% | 3.000000 | 126.000000 | 120.000000 | 1.000000 |
| max | 6.000000 | 289.500000 | 179.550000 | 2.500000 |

## *"Label encoding preserves ordinal relationships for meal plans."*

```
Categorical columns before encoding:
ut actions['Booking_ID', 'type of meal', 'room type', 'market segment type',
        'date of reservation', 'booking status'],
      dtype='object')

After encoding:
   type of meal  room type  market segment type  booking status
0             0          0                    3               1
1             3          0                    4               1
2             0          0                    4               0
3             0          0                    4               0
4             3          0                    4               0
```

Table showing type of meal → encoded values (0, 1, 2)

*"Min-Max scaling ensures equal feature weighting in models."*

After normalization:

| | number of adults | number of children | number of weekend nights |
|---|---|---|---|
| count | 36285.0 | 36285.0 | 36285.000000 |
| mean | 0.0 | 0.0 | 0.162017 |
| std | 0.0 | 0.0 | 0.173457 |
| min | 0.0 | 0.0 | 0.000000 |
| 25% | 0.0 | 0.0 | 0.000000 |
| 50% | 0.0 | 0.0 | 0.200000 |
| 75% | 0.0 | 0.0 | 0.400000 |
| max | 0.0 | 0.0 | 1.000000 |

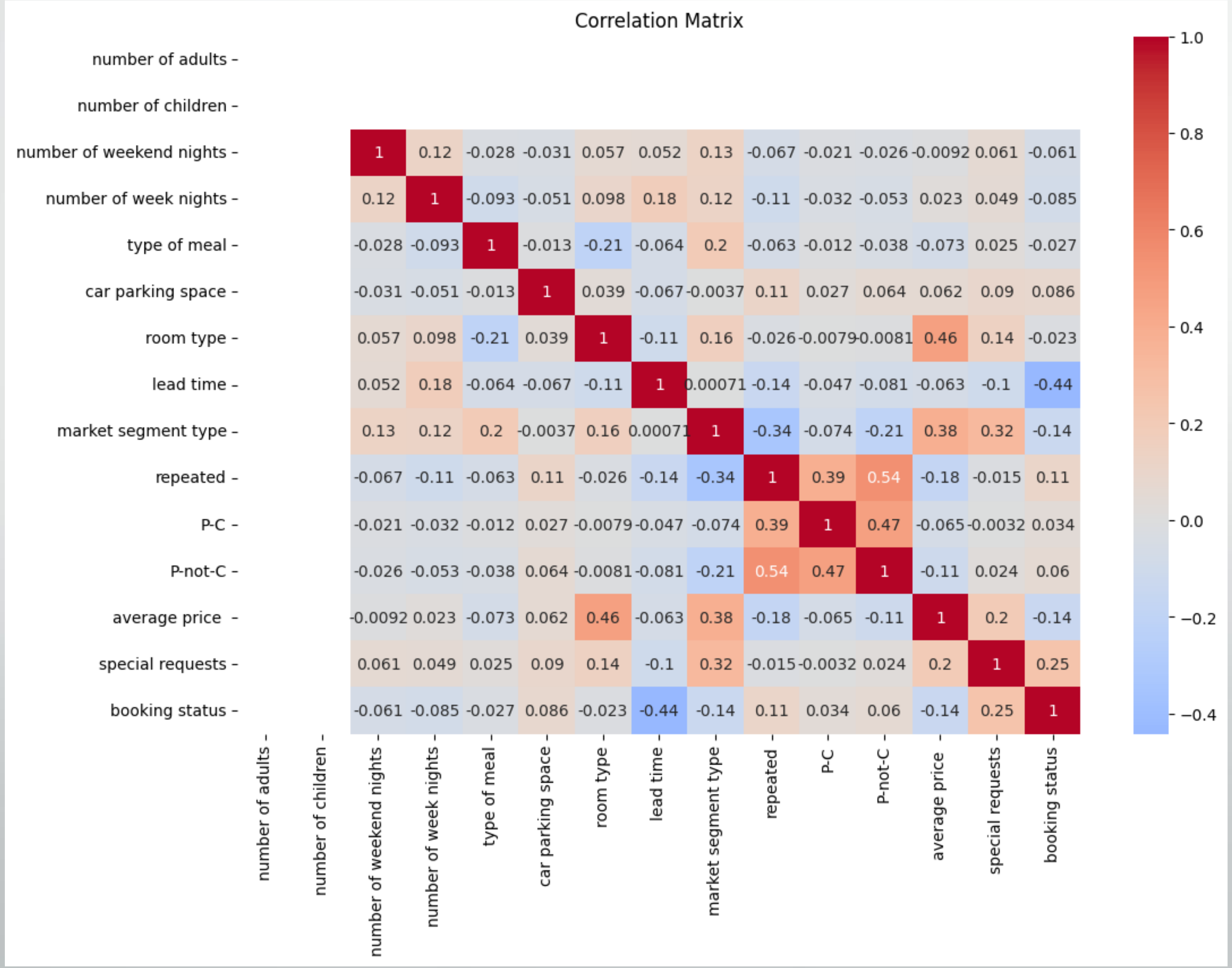
| | number of week nights | lead time | average price | special requests |
|---|---|---|---|---|
| count | 36285.000000 | 36285.000000 | 36285.000000 | 36285.000000 |
| mean | 0.363024 | 0.289354 | 0.517748 | 0.242657 |
| std | 0.215118 | 0.282080 | 0.199489 | 0.298780 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 0.166667 | 0.058722 | 0.375000 | 0.000000 |
| 50% | 0.333333 | 0.196891 | 0.495592 | 0.000000 |
| 75% | 0.500000 | 0.435233 | 0.625000 | 0.400000 |
| max | 1.000000 | 1.000000 | 1.000000 | 1.000000 |

*"Upsampled minority class to prevent model bias toward majority."*

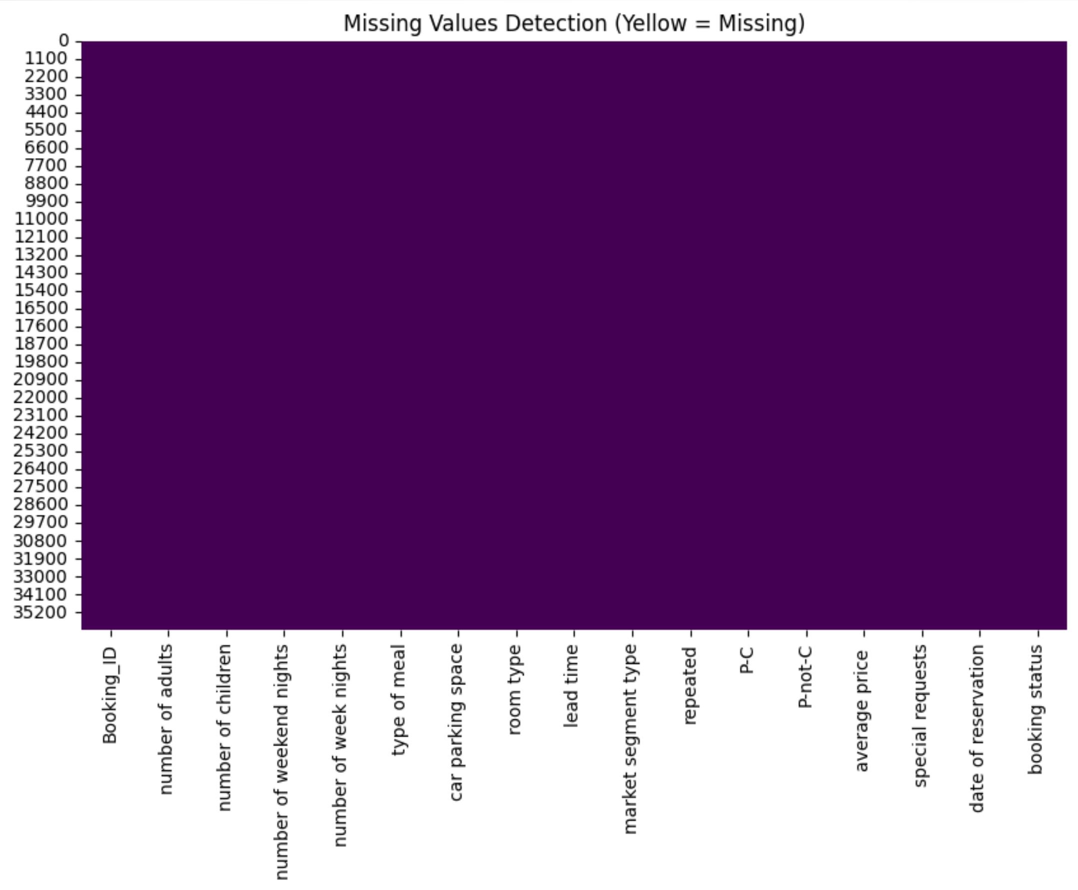
```
Class distribution before balancing:
booking status
1    24396
0    11889
Name: count, dtype: int64

Class distribution after balancing:
booking status
0    11889
1    11889
Name: count, dtype: int64
```

# "No features dropped – all correlations < 0.8."



## *"No missing values detected – rare for real-world data! Our pipeline includes safeguards anyway."*



1."Feature Distributions Post-Normalization"
2."Standardized Features: Adults, Lead Time & Price"
3."Normalized Data for Model Readiness"
4."From Raw to Scaled: Key Features"