**Hotel Booking Prediction**

**Introduction-**

Every year, more than 140 million bookings made on the internet and many hotel bookings made through top-visited travel websites like Booking.com, Expedia.com, Hotels.com, etc. According to Google data, hotels are booked in advance of 12 weeks.

This dataset contains 31 features about booking information such as Average Daily Rate, Arrival Time, Room Type, Special Request, etc. between 2015 and 2017 years.

**Business Problem-**

This case study examines a dataset that contains information on hotel reservations and associated characteristics. We will look at what characteristics are accessible to hotel owners and how they may utilize them to forecast the result of a booking in this research. Hotel operators might better plan their company and understand how they can profit in these difficult times if they used the information acquired from this research. My algorithm will predict if a hotel booking will be canceled or not based on the features from each hotel booking. I aim to find best model to predict hotel booking cancellations with tree-based algorithms based on rest of the features found in the dataset. The goal of predictive analysis is to avoid overfitting and find the model that has the highest accuracy.

**Problem statements**

* What does the analysis/model building tell you?
* What are your recommendations?
* How would you pitch this business problem to a group of stakeholders to gain buy-in to proceed?
* Why should someone in the business care about this solution?
* What are some of the potential challenges or additional opportunities that need to be explored?

The scope of the project is based on dataset contains 31 features about booking information such as Average Daily Rate, Arrival Time, Room Type, Special Request, etc. between 2015 and 2017 years. Below is the scope of the effort:

• Data Selection

• Identify relevant attributes

• Data Preparation

• Feature Selection/Feature Engineering

• Exploratory Data Analysis

• Data Visualization

• Model Selection

• Model Evaluation

• Expected Results / Outcome

**Know the dataset**

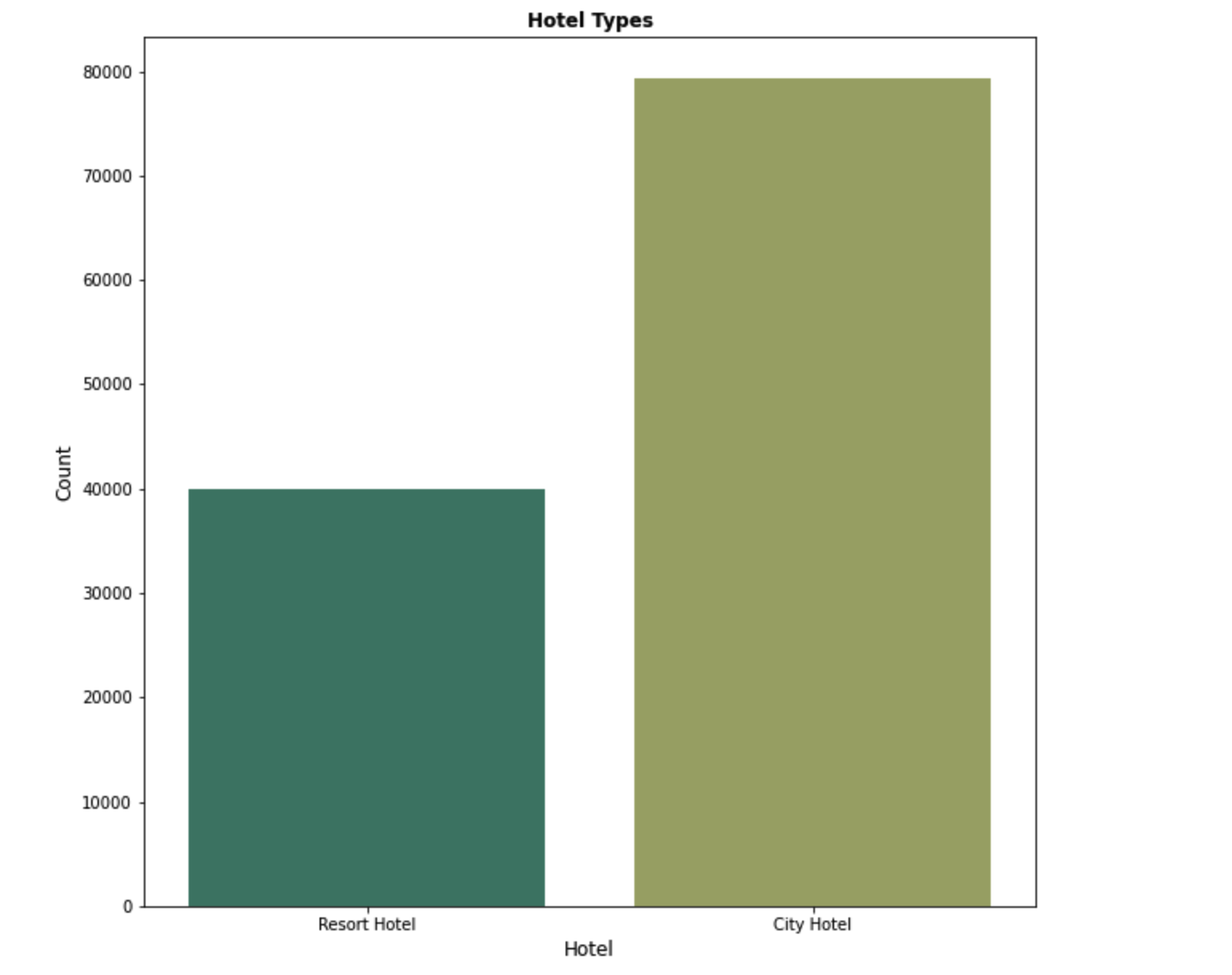
* hotel (H1 = Resort Hotel or H2 = City Hotel)
* is\_canceled Value indicating if the booking was canceled (1) or not (0)
* lead\_time Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
* arrival\_date\_year Year of arrival date
* arrival\_date\_month Month of arrival date
* arrival\_date\_week\_number Week number of year for arrival date
* arrival\_date\_day\_of\_month Day of arrival date
* stays\_in\_weekend\_nights Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
* stays\_in\_week\_nights Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
* adults Number of adults
* children Number of children
* babies Number of babies
* meal Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal
* country Country of origin. Categories are represented in the ISO 3155–3:2013 format
* market\_segment Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
* distribution\_channel Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
* is\_repeated\_guest Value indicating if the booking name was from a repeated guest (1) or not (0)
* previous\_cancellations Number of previous bookings that were cancelled by the customer prior to the current booking
* previous\_bookings\_not\_canceled Number of previous bookings not cancelled by the customer prior to the current booking
* reserved\_room\_type Code of room type reserved. Code is presented instead of designation for anonymity reasons.
* assigned\_room\_typeCode for the type of room assigned to the booking. Code is presented instead of designation for anonymity reasons.
* booking\_changes Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or out
* deposit\_type Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No
* agent ID of the travel agency that made the booking
* company ID of the company that made the booking or responsible for paying the booking.
* days\_in\_waiting\_list Number of days the booking was in the waiting list before it was confirmed to the customer
* customer\_type Type of booking, assuming one of four categories: Transient - Transient-Party - Contract - Group
* adr Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
* required\_car\_parking\_spaces Number of car parking spaces required by the customer
* total\_of\_special\_requestsNumber of special requests made by the customer (e.g., twin bed or high floor)
* reservation\_status Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out
* reservation\_status\_date Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to

**Exploratory Data Analysis (EDA)**

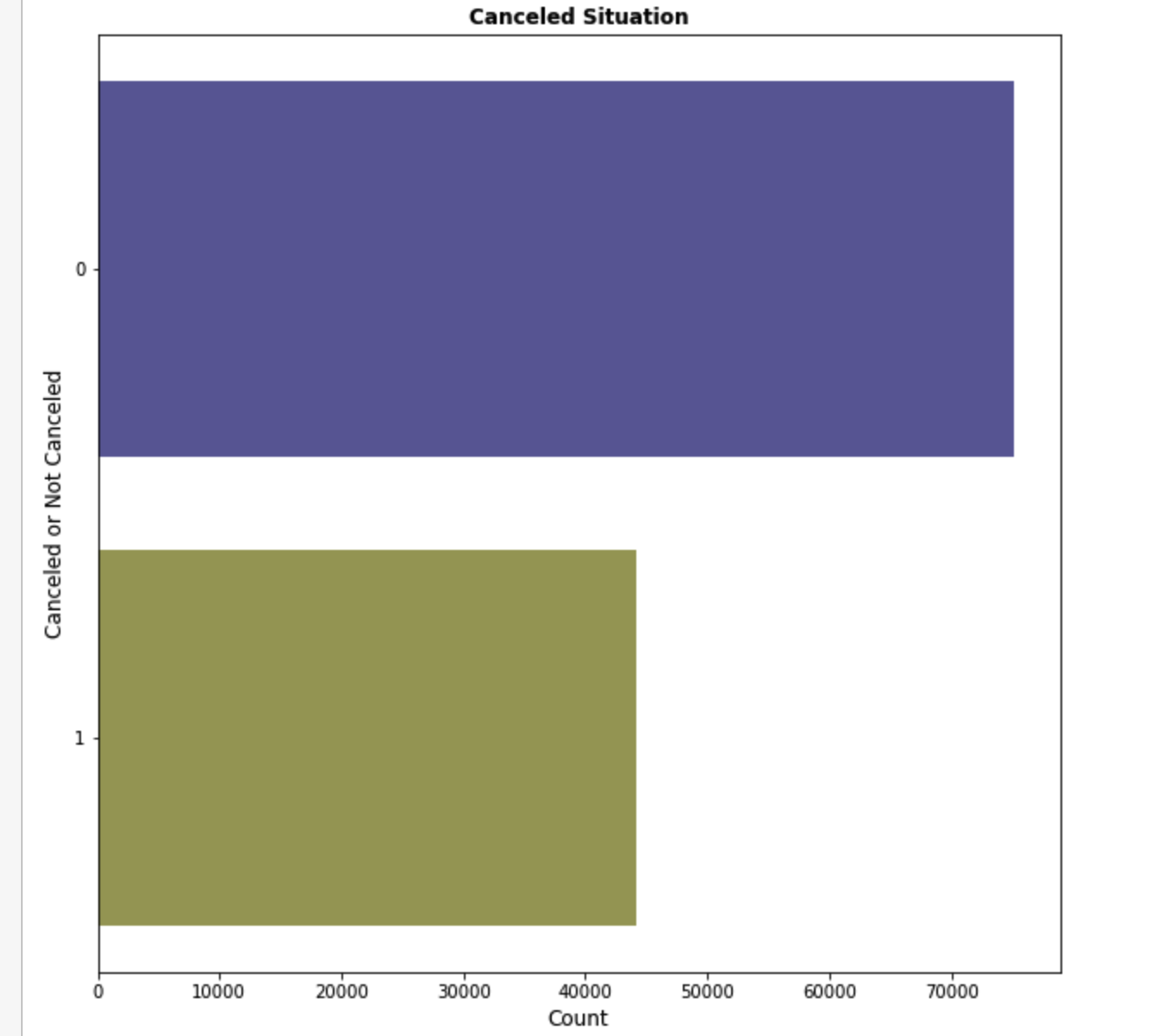
Topics covered and questions to answer from the data:

* Where do the guests come from?
* How much do guests pay for a room per night?
* How does the price per night vary over the year?
* Which is the busiest month?
* How long do people stay at the hotels?
* Bookings by market segment
* How many bookings were canceled?
* Which month have the highest number of cancelations?

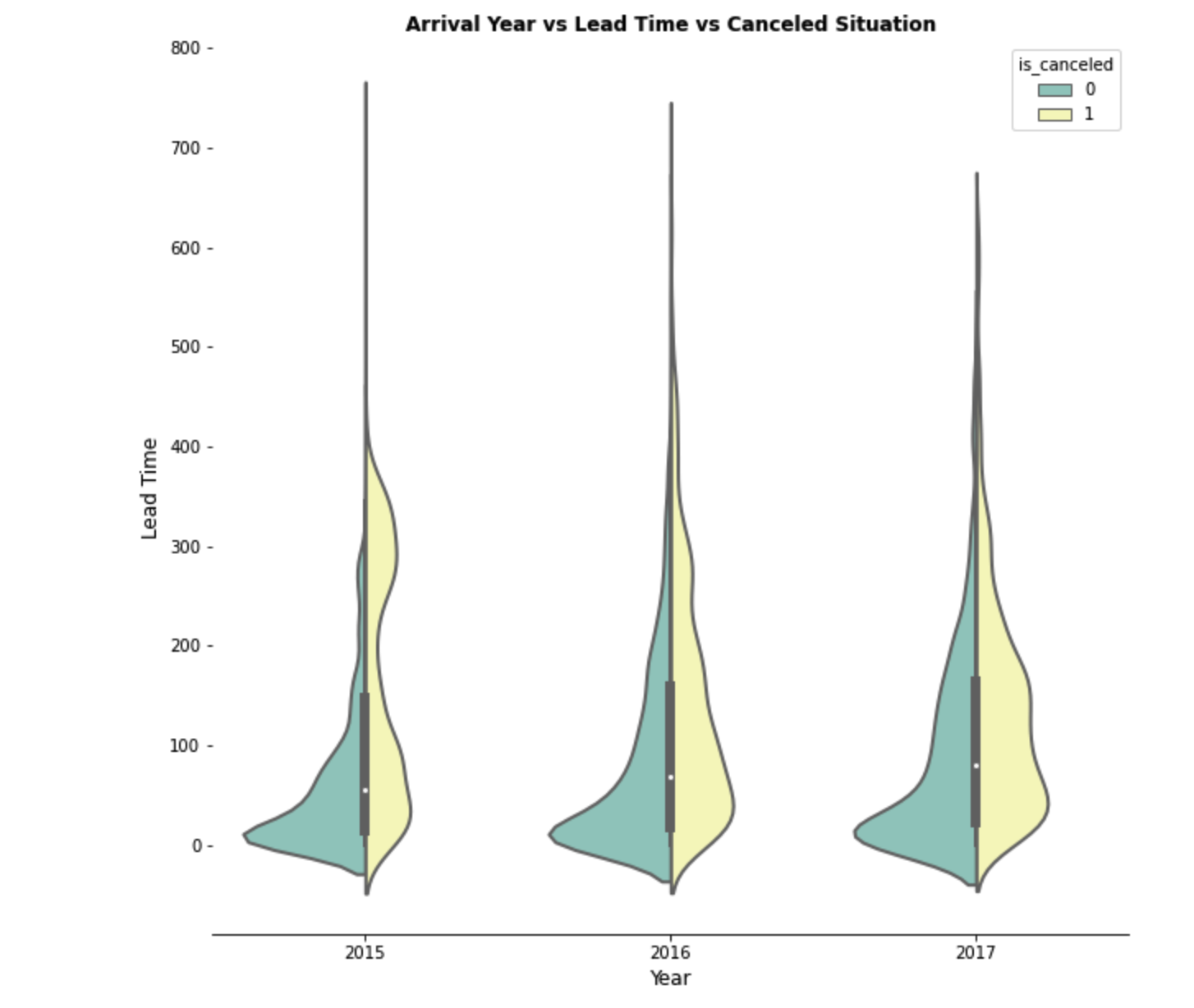
Hotel types



Now, let's dive into the target value of data. The numbers are similar with hotel features. While 37% of booking canceled, 63% of booking is not canceled. These numbers also show that there is no balanced problem on the target value.



Below graph shows the relationship of arrival\_date\_year to lead\_time with booking cancellation status. The graph created by violin plot. Violin plot is a hybrid of box plot and density plot. It shows the distribution of the data.



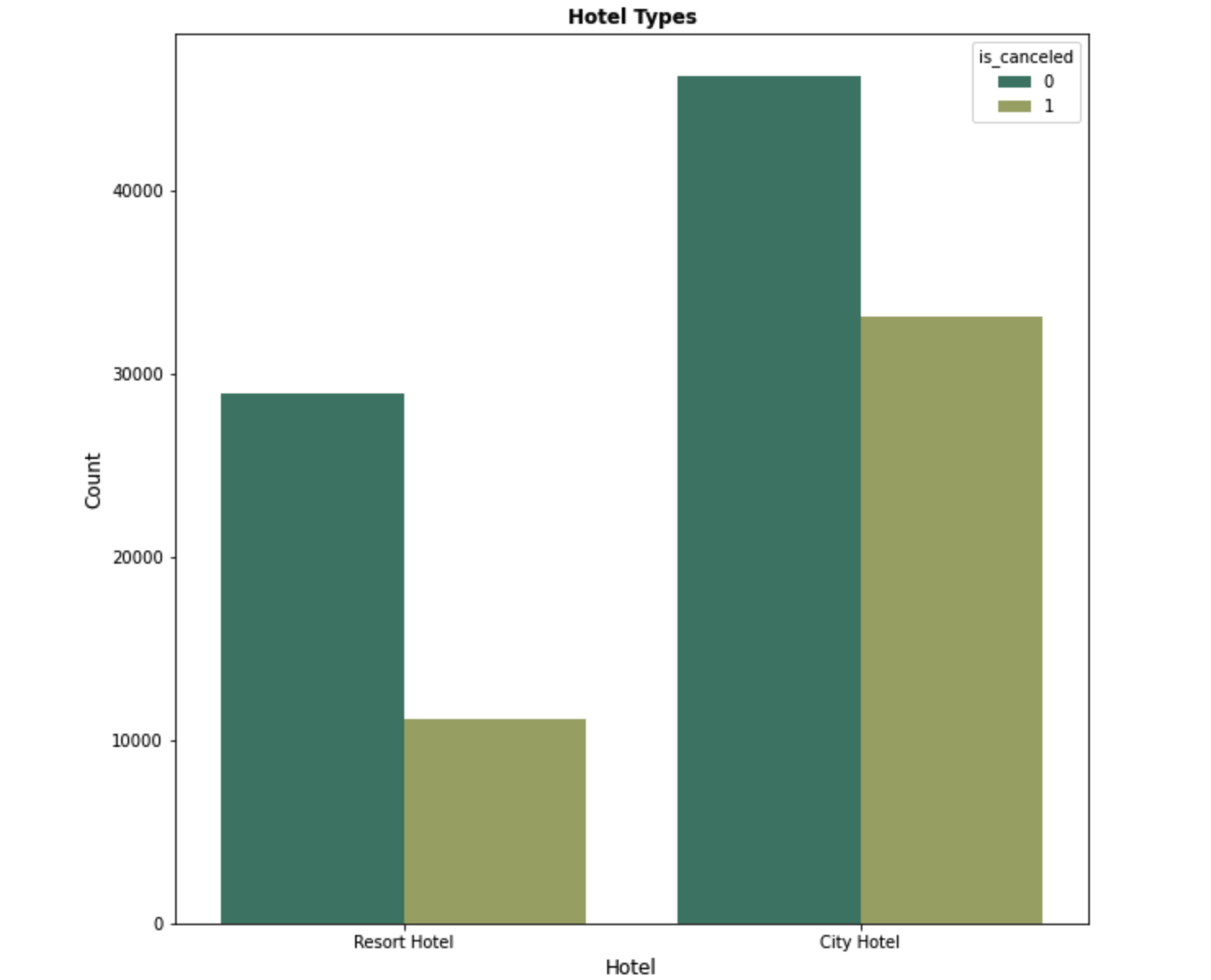
We need to some **data preprocessing** here as we can see

Children, Country, Agent and Company having missing values.

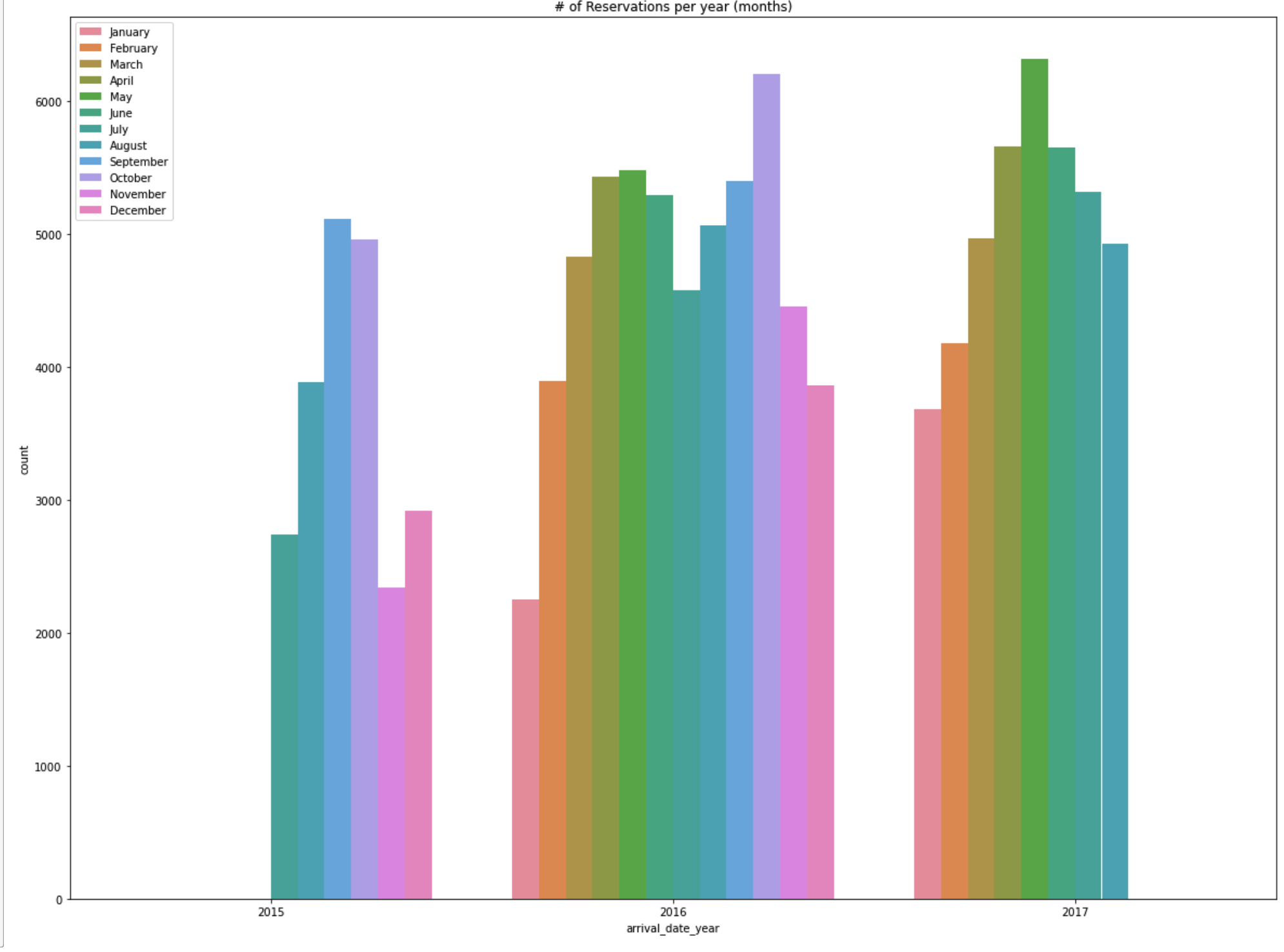
For Children we will default the missing values to 0

For country we will default the missing value to Unknown agent: If no agency is given, booking was most likely made without one. company: If none given, it was most likely private.

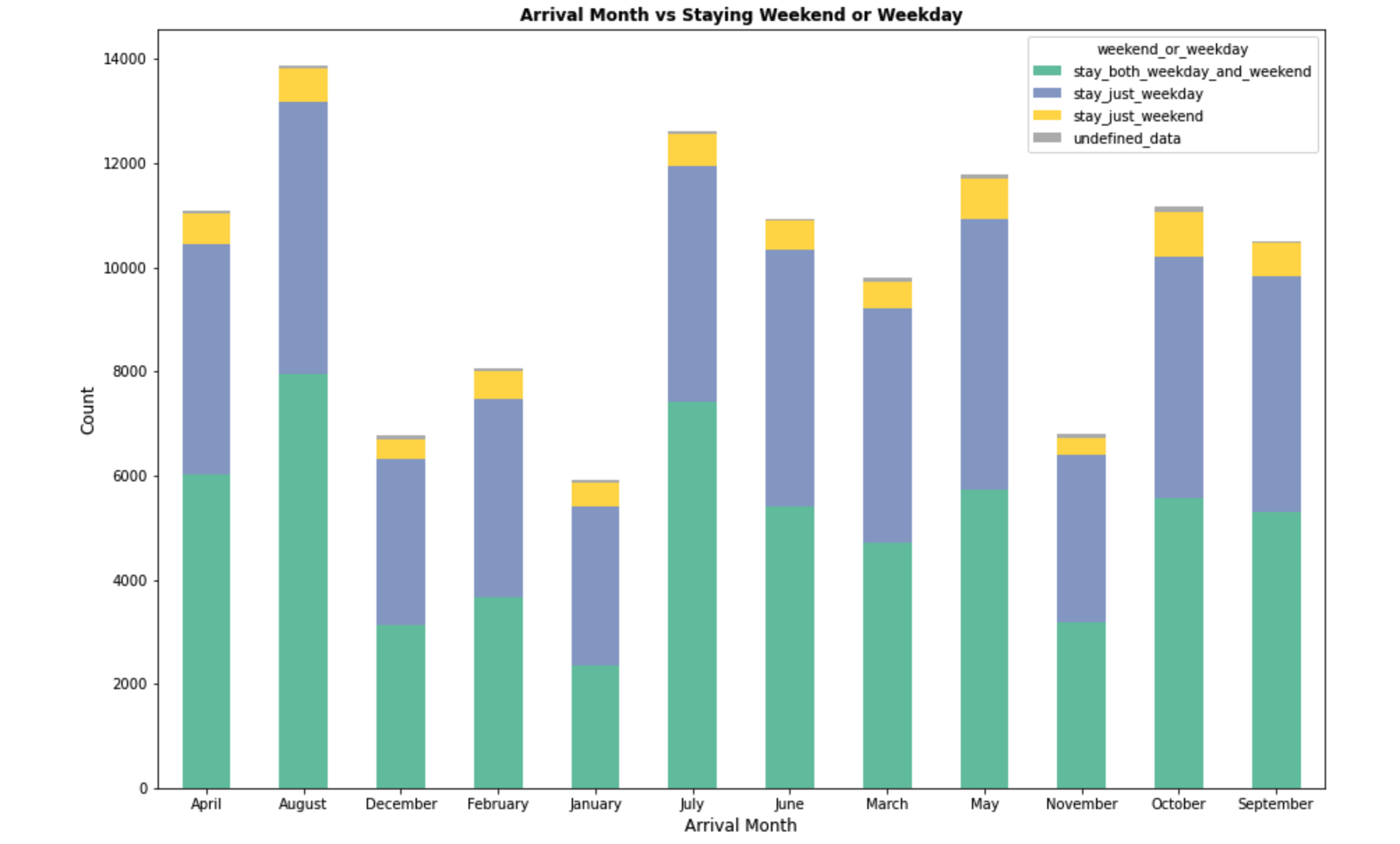
After we do the data preprocessing let us dig a little bit on type of hotels and its cancellation status. We can see the city hotels have majority of cancellation.



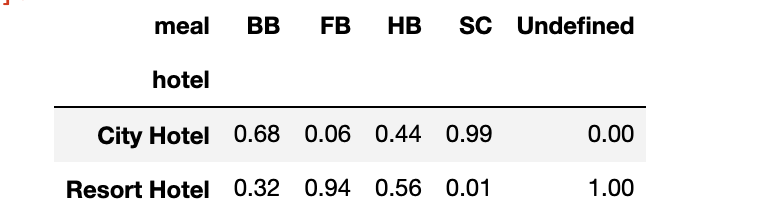
Now we need to take a deeper look which months the hotels were busy.



The next graph shows the link between the weekend or weekday characteristic and the arrival date month. The bar graph below illustrates that the majority of reservations were booked for simply weekdays or both weekdays and weekends. On the other hand, as compared to other categories, the number of people staying simply for the weekend is fairly low.



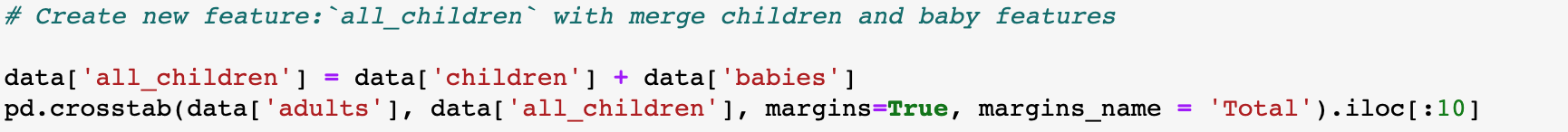
The below table shows frequency details about meal types according to the hotel types. Following the results, 68% of Bed&Breakfast booking made for City Hotel and almost every Full Board booking made in the Resort Hotel



Now we get to know about the data a bit let us dig into feature engineering and dimension reduction.

**Feature Engineering/Dimension Reduction-**

There is no obvious difference between children and babies these features gathered under the one feature which name is all\_children.



Another part is analyzing categorical features. Categorical labels converted into numerical form. This will help to be more understandable and implementable into machine learning algorithms. Some features are not ordinal such as country. In that case, One-Hot Encoding could be chosen. Due to the high number of categories, this method could incur higher computational cost. To help reducing that, Label Encoding method will be used.

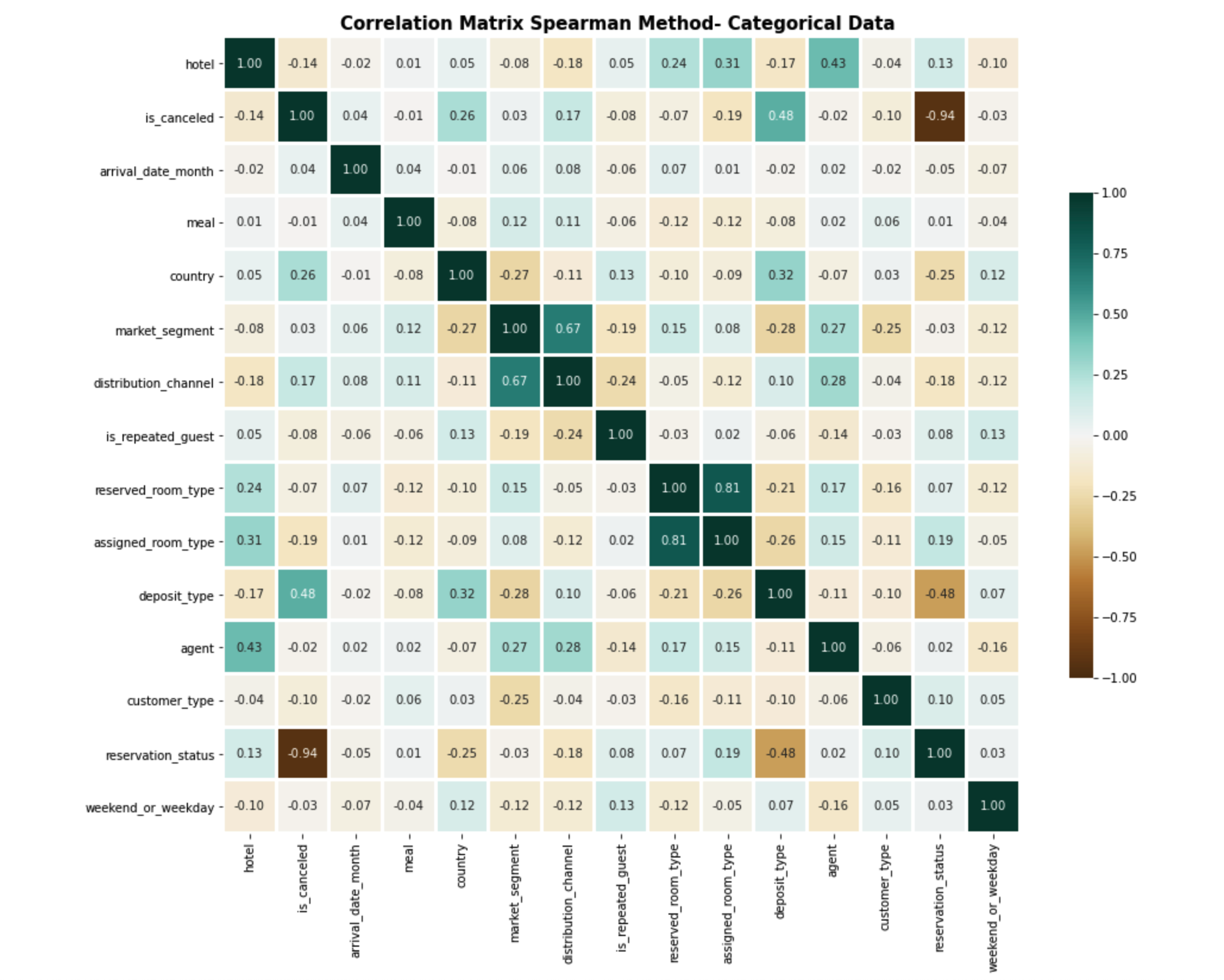
We are going to create two data frames here. One data frame has solely category information, whereas the other contains numerical information. A correlation matrix will be created using these two separate data sets.

For categorical data correlation matrices, the Spearman approach will be utilized, and for numerical data correlation matrices, the Pearson method will be used.

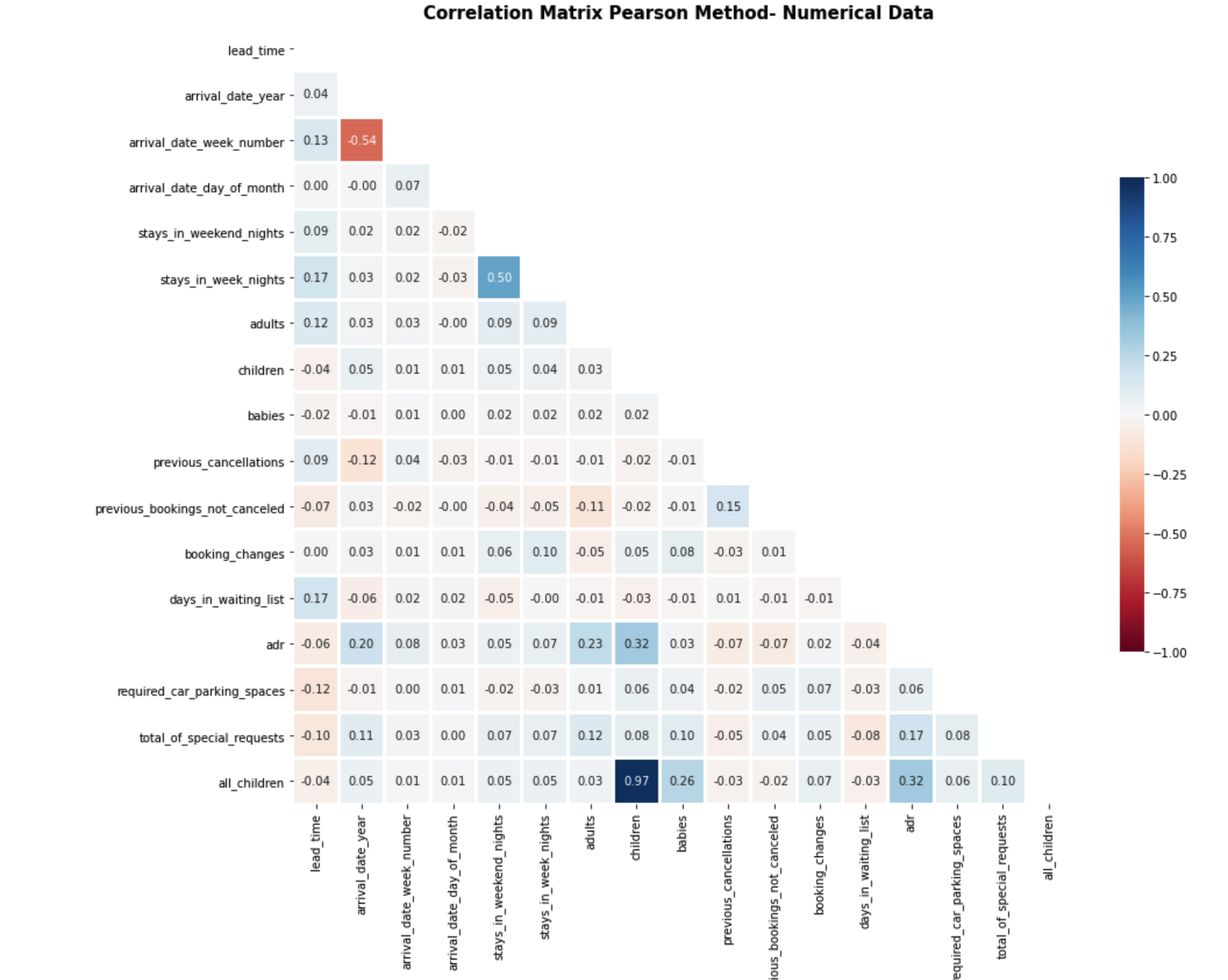
We are going to employ three different methods here to select features.

* Spearman Method
* Pearson coefficient
* Probability Density function

**Correlation Matrix with Spearman method**



**Correlation Matrix with Pearson Coefficient method**



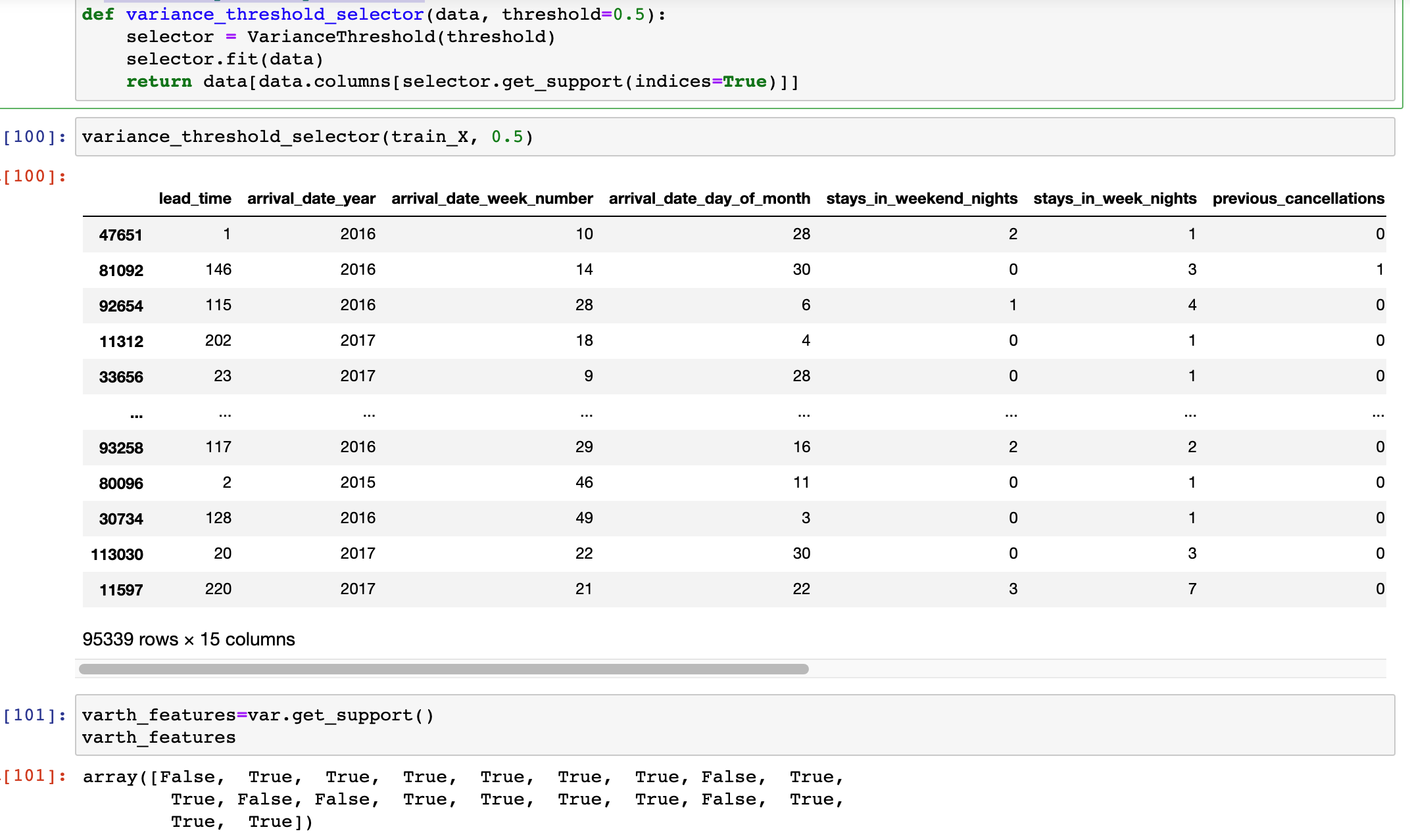
The correlation matrix above depicts if there are positive or negative links between them. Because of its negative connection with the is canceled feature, the reservation\_ status feature is given more prominence in those two heatmaps. The relationship with details is depicted in the table below. This high correlation may result in an incorrect prediction or overfitting. To avoid this, the reservation status feature will be removed.

On the other hand, because the all-children feature is made up of the children and the infants features, there is a strong link between the children and the all children features. As a result, the children function will be removed as well.

Last feature is reservation\_status\_date. Since this feature includes date type data and it could not convert another type, this feature will be eliminated.

**Probability density function**

Feature selection using Variance Threshold with threshold of 0.5



The probability density function of a given distribution is used to compute the variance threshold. If a characteristic has 95 percent or greater variability should be considered, it is extremely near to zero, therefore it may not be useful in model prediction and should be eliminated. The characteristics picked using the Variance threshold method are represented as True values

The columns hotel, arrival\_date\_year, is\_repeated\_guest, booking\_changes, deposit\_type & required\_car\_parking\_spaces should be removed because the PDF is close to Zero

**Model Creation**

Some tree-based methods were used to generate models in this section. Decision Tree, Random Forest, Extra Trees Classifier, and Extreme Gradient Boosting are some of the options. As bagging algorithms, Random Forest and Extra Tree Classification were chosen, and XGBoost was picked as one of the boosting techniques. As one tree algorithm, the Decision Tree algorithm was used.

Before model building, data will be split to train and test respectively 70% and 30% ratio. X\_train and X\_test data will be standardized with the Standard Scaler technique. After that, the Stratified K-Fold Cross Validation method will be used for resampling. Cross-validation is an important implementation to avoid overfitting. Stratified K-Fold Cross Validation method provides train/test indices to split data into train/test sets. Model parameters have been defined in the previous part.

The last part is comparison of classification reports of ML models.

First comparison the accuracy results.

Accuracy is a ratio of correct predictions to the total predictions. Its formula is (𝑇𝑃+𝑇𝑁)/(𝑇𝑃+𝐹𝑃+𝐹𝑁+𝑇𝑁) According to that, Random Forest have the highest correct prediction with 88%. Another performance metrics explained below:

Precision: It is the ratio of correctly predicted observation to the total positive predicted observation. Its formula is 𝑇𝑃/(𝑇𝑃+𝐹𝑃) Recall: It is the ratio of correctly predicted positive observations to the actual positive observations. Its formula is 𝑇𝑃/(𝑇𝑃+𝐹𝑁) Random Forest and the Extra Tree Classifier share the highest precision ratios. It means that both models predicted around 88% of all the positive labels correctly. On the other hand Random Forest has the highest recall ratio. It means that this model predicted 79% of actual positive observations correctly.

Yet to execute the above methods

**Model Evaluation/Selction**

Yet to execute. Confusion matrix will be shown here with model efficiency

**Conclusion/Recommendation**

Yet to evaluate

**Refererences**

<https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md>

<https://medium.com/analytics-vidhya/predicting-hotel-booking-cancellations-3fe40d4522f8>