Report

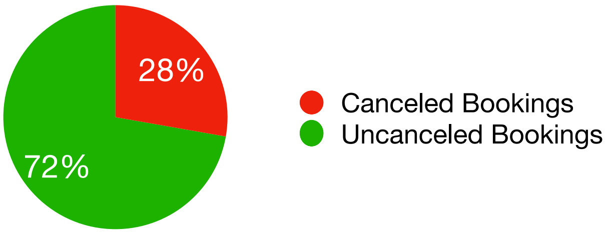
Note: All values and statistics shown here are from the H1.csv dataset.

Initial Preprocessing:

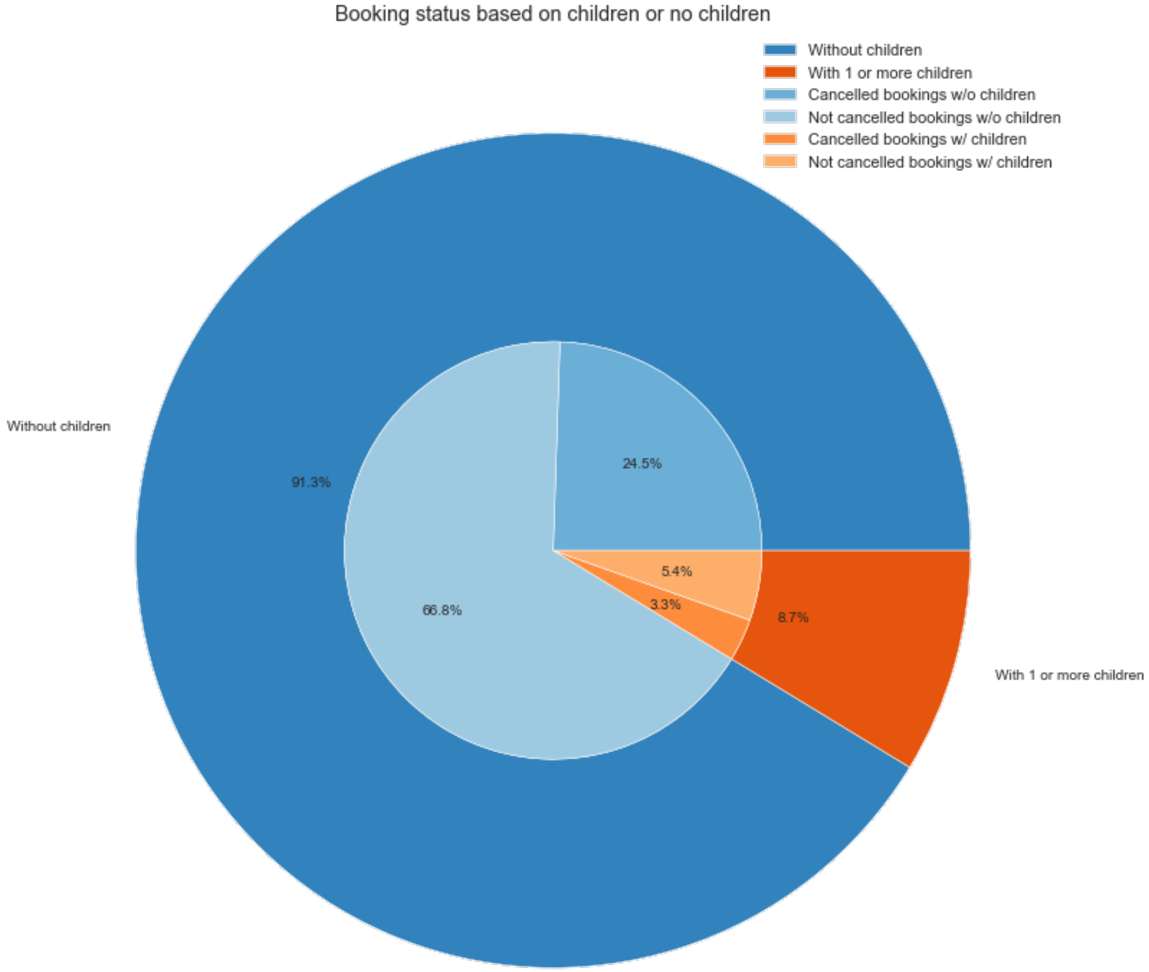
To get started, I replaced all the string “NULL” values in the dataset with bumpy nan values so that they are identified as missing values by pandas.

Statistical Analysis:

An initial analysis on the number of canceled bookings revealed that only about 28% of the bookings were canceled, while 72% of the bookings remained not canceled. This simple statistic suggests that a large majority of bookings were not canceled.

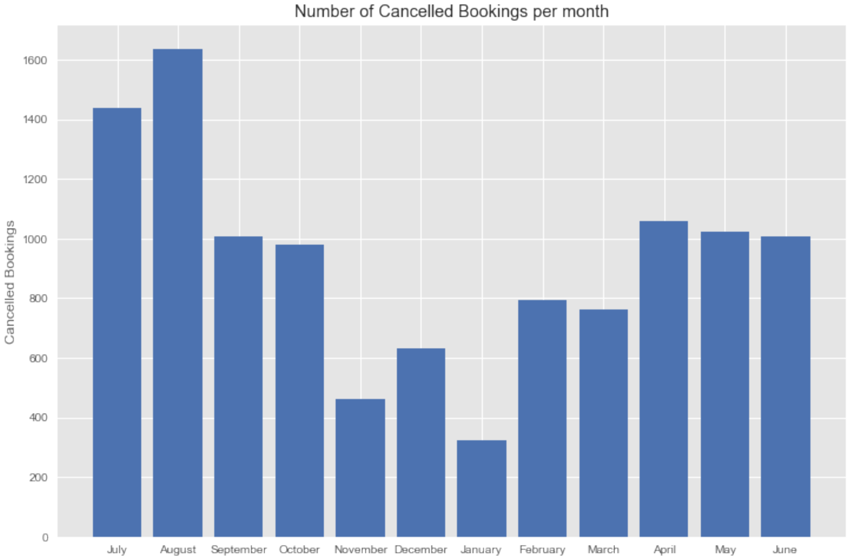


After this, I calculated the booking cancellation rate of people with or without children. I found that Here are my findings represented visually:



A large majority of bookings created are by guests without children. About 25% of people without children and ~37% of people with one or more children canceled their reservations.

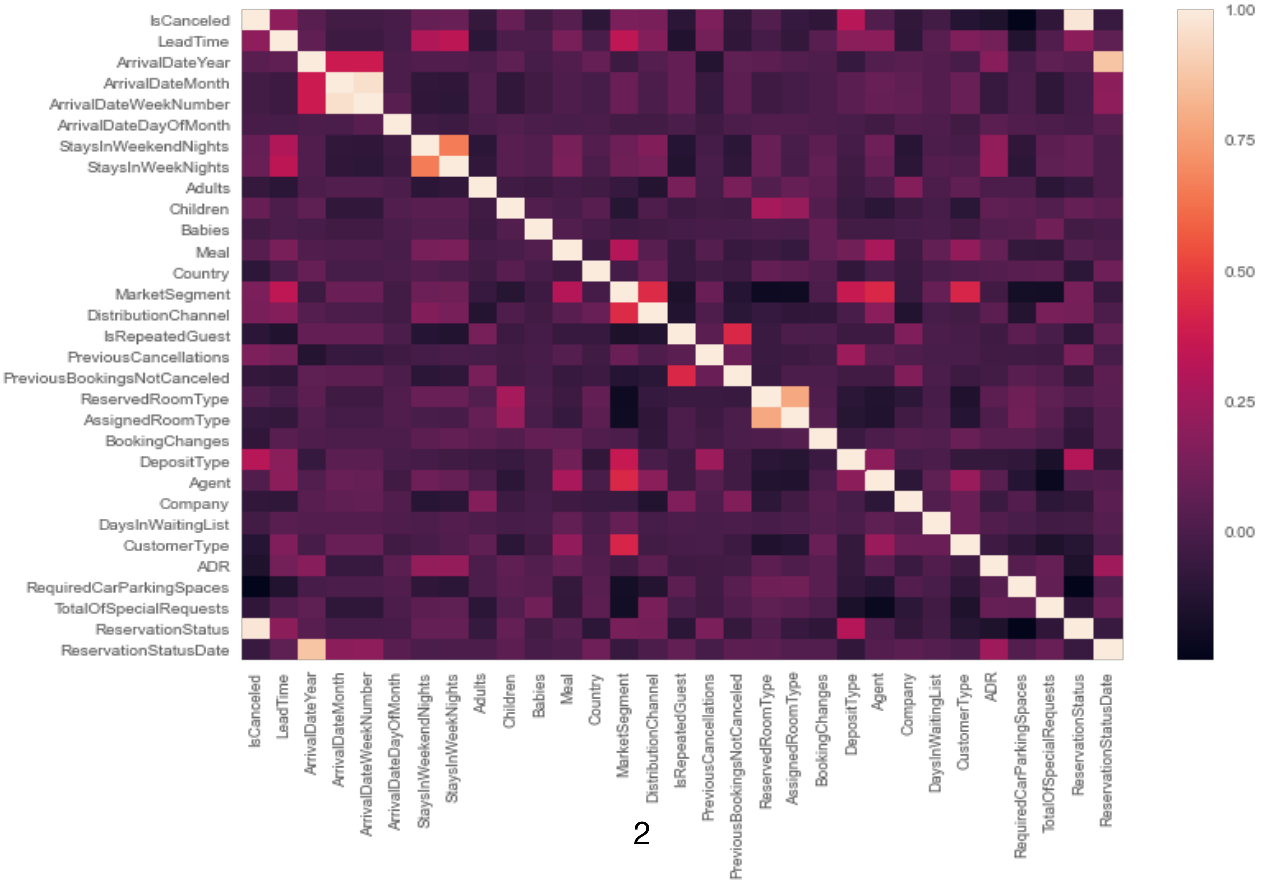
Finally, I calculated the number of cancellations during each month. Here are my findings visualized:



As visible, August and July had the most cancellations over the years. These months also had the most number of reservations. This follows the trend that one of the best times to visit Portugal is during August and July according to this website. However, the high cancellation rates could be due to two possibilities: guests could be rethinking their travel plans due to the high summer heat during these months, or they could have found better deals at resorts closer to the beach.

Data preprocessing for model building/Classification:

First, I created a confusion matrix displaying the correlation between features (shown below). From the correlation, I decided to delete the features with a correlation above 95%.



The feature *ReservationStatus* is highly correlated with the target value *IsCanceled* and *ArrivalDateWeek* is highly correlated with *ArrivalDateMonth*. I removed these features. I then encoded the categorical features except for the date features with a Binary Encoder. For the features containing a date, I performed a sine and cosine transformation so that months and days are represented in a cyclical manner. I dropped the *Company* feature since it had a significant amount of missing values (~92% of the data). Since the *Country* column did not have many missing values (~1% of the data), I removed the rows with missing values. Finally, I decided to impute the missing values in the *Agent* feature using a KNN imputer with 5 neighbors. Before doing this, I split the data into training and test sets so that the test set can be imputed with the KNN imputer that is trained on the training set to avoid an information leak.

Classification models:

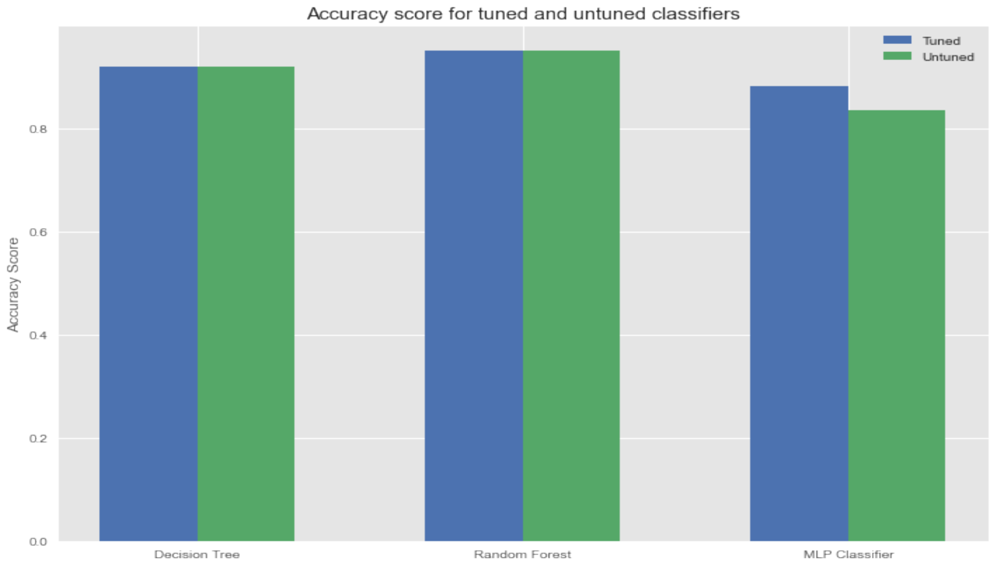
I chose to apply the Decision Tree, Random Forest, and Multilayer Perceptron classifiers.

First, I tested these classifiers without tuning any parameters. The Random Forest Classifier resulted in the highest accuracy score, while the Multilayer Perceptron Classifier resulted in the worst.

I then ran grid search to determine the best parameters for each classifier. The grid search for Decision Tree built 1280 models, Random Forest built 90 models, and Multilayer Perceptron built 216 models.

The tuned and untuned accuracy scores for each classifier are presented in the table and chart below:

	Untuned Parameters accuracy score	**Tuned Parameters accuracy score**	**Difference (Tuned - Untuned)**
Decision Tree Classifier	0.9192	0.9210	0.0018
Random Forest Classifier	0.9515	0.9516	0.0001
Multilayer Perceptron Classifier	0.8491	0.8832	0.0472



Conclusion:

The Random Forest Classifier prevailed as the most effective model after tuning, while the Multilayer Perceptron Classifier still had an accuracy score below 90%. However, the MLP Classifier had the highest accuracy gain from the tuning process as visible from the chart. The accuracy gains for Decision Tree and Random Forest are negligible. In conclusion, the tuned Random Forest Classifier is the best model to determine if a user is going to cancel their reservation.