## Chapter Three: Methodology

This project aims to create an API that can provide a function to predict if a reservation is most likely to cancel or not. To do this, we need to create a function that can evaluate a reservation according to its variables. The variables will be studied and tested to extract the factors or cancelation.

The study starts by defining the factors of cancelation based on previous reservation history. After that, we evaluate these factors to get the most effective factors. Then, we evaluate the factorizing process by finding its impact on the prediction of cancelation. To do that we implement a prediction model using logistic regression using all variables and another prediction model using the extracted factors. Consequently, we compare the precision of both models. Finally, we use the extracted factors to build a scoring function and a prediction function.

### Dataset and Variable Investigation

The dataset used for this project is downloaded from Kaggle website. it contains a variety of hotel reservations on formation collect it based on some potential factors.

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, and the number of available parking spaces, among other things. All personally identifying information has been removed from the data.

The dataset contains 32 column described in their original project as the following:

1. hotel: Hotel (H1 = Resort Hotel or H2 = City Hotel)
2. is\_canceled: Value indicating if the booking was canceled (1) or not (0)
3. lead\_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
4. arrival\_date\_year: Year of arrival date
5. arrival\_date\_month: Month of arrival date
6. arrival\_date\_week\_number: Week number of year for arrival date
7. arrival\_date\_day\_of\_month: Day of arrival date
8. stays\_in\_weekend\_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
9. stays\_in\_week\_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
10. adults: Number of adults
11. children: Number of children
12. babies: Number of babies
13. meal: Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner)
14. country: Country of origin. Categories are represented in the ISO 3155–3:2013 format
15. market\_segment: Market segment designation. In categories, the term “TA” means “Travel Agents” and “TO” means “Tour Operators”
16. distribution\_channel: Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”
17. is\_repeated\_guest: Value indicating if the booking name was from a repeated guest (1) or not (0)
18. previous\_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking
19. previous\_bookings\_not\_canceled: Number of previous bookings not cancelled by the customer prior to the current booking
20. reserved\_room\_type: Code of room type reserved. Code is presented instead of designation for anonymity reasons
21. assigned\_room\_type: Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
22. booking\_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
23. deposit\_type: Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; Non Refund – a deposit was made in the value of the total stay cost; Refundable – a deposit was made with a value under the total cost of stay.
24. agent: ID of the travel agency that made the booking
25. company: ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
26. days\_in\_waiting\_list: Number of days the booking was in the waiting list before it was confirmed to the customer
27. customer\_type: Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group – when the booking is associated to a group; Transient – when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party – when the booking is transient, but is associated to at least other transient booking
28. adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights
29. required\_car\_parking\_spaces: Number of car parking spaces required by the customer
30. total\_of\_special\_requests: Number of special requests made by the customer (e.g. twin bed or high floor)
31. reservation\_status: Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why
32. reservation\_status\_date: Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel

The second column is the dependent variable in this study, while the other columns represents the independent variables as potential factors.

### Factor Extraction

To extract the factors, we use chi-square test of independence. For nominated variables, we use correlation to find the effective variables. after that, we add another test based on the coefficient of the variables in the linear regression.

### Factors Evaluation

To evaluate the factors, we use two logistic regression models. One to predict the cancelation based on all variables in the dataset. The other one predicts the cancelation based on the nominated factors. The factorizing process is considered efficient if the precision of the factors model revealed better performance than the all variables model.

### Scouring Function

Assuming that the cancelation process follows a linear equation, we use the intercept and coefficients from the prediction model to create a scouring function that can calculate the sum of the factors multiplied by their coefficients in addition to the intercept. This scour is to be evaluated as a probability later using the logistic function:

The result will give the probability of cancelation. If the probability is more than 0.5, the reservation process is more likely to cancel.

### API Function

The API function is the last function that will use the scouring function to predict if a reservation would likely cancel or not. Its definition would be:

Boolean cancel (factorsArray)

It will return true if the factors indicate probability of cancelations and false otherwise.

## Chapter Four: Results Discussion

The question of this project is what are the main factors of hotel reservation cancelling? From this question we can hypothetically derive a sub question for each of the other parameters. However, we can disregard some parameters logically or practically during data wrangling.

We start with 31 proposed variables. These included timestamps, details, and status data. For the timestamps, we can decide what kind of time variables we should accept by understanding the logical scope of the problem. In other words, how would time affect the hotel reservation? and what timestamp are we talking about? If we considered that there might be some yearly routine that during some period of the year there would be many cancelations, then we do not need to know which year exactly we are talking about. In addition, this cannot be observed on daily bases. It is convenient to consider the period on weekly and monthly basis. For this, we can drop the other timestamps.

Now, when we talk about dates, we consider the arriving date in the reservation, not the reservation date, leaving date, nor status date. The other timestamps might be eliminated too. The details of the reservation include information about the hotel, agency, travelers, reservation type, average daily rate, and meals. Any variable considering this reservation is considered as a potential factor. The status is what we are trying to predict and analyze to find its factor. Namely, the calculation status. for this, we can consider all other status as not canceled. We do not need more details about the status.

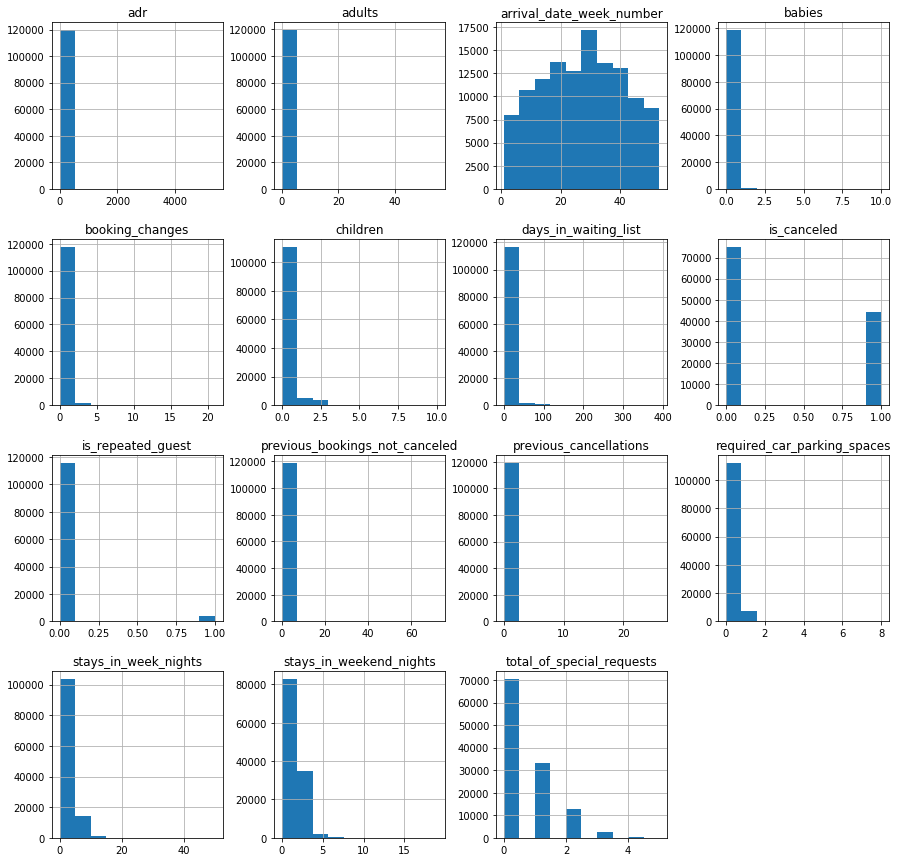
Based on the criteria above, we will drop the follwoing columns:

* lead\_time: Unneeded timestamp
* arrival\_date\_year: Unneeded timestamp
* arrival\_date\_day\_of\_month: Unneeded timestamp
* reservation\_status: Unneeded status details
* reservation\_status\_date: Unneeded timestamp

There are missing data in the children field, country field, agent field, and company field. From the data description in the introduction we can find that we may get the types we need from the market segment, distribution channel, and customer type and disregard the agent and company columns. So, we just drop these two. For the children field and country field, we may fill the missed data with 0 and 'Not' as mark that it was not inserted.

After cleaning, the dataset contains 33775 duplicates which is more 30% of the dataset. However, these duplicates are not real duplicates. They might be different in term of year, day, or other removed fields. For this reason, no duplicates might be eliminated.

The EDA includes having an insight of the statistics of the dataset. This insight will help in detecting if any bias occurs. After that, we evaluate each factor using chi-square test of independence. To create the list of the effective factors, we will calculate the correlation coefficient between all the variables to understand the relationship between the variables. This process might propose eliminating some unrelated variables. Finally, we use the remaining factors in creating prediction models, and compare the accuracy of these prediction models with the accuracy of the same models which included all the eliminated data. Starting with the numerical variables, we have the following histogram:

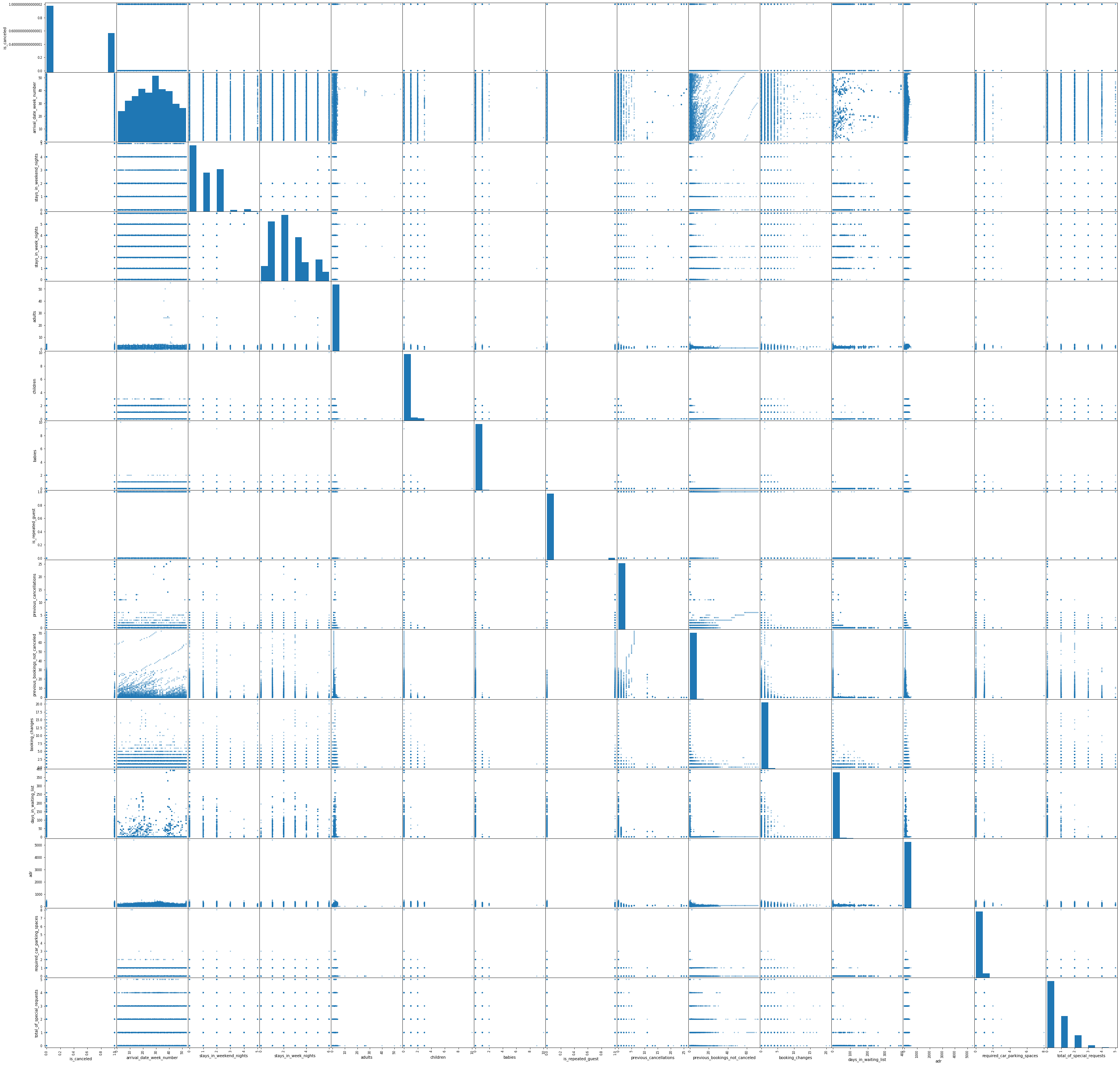
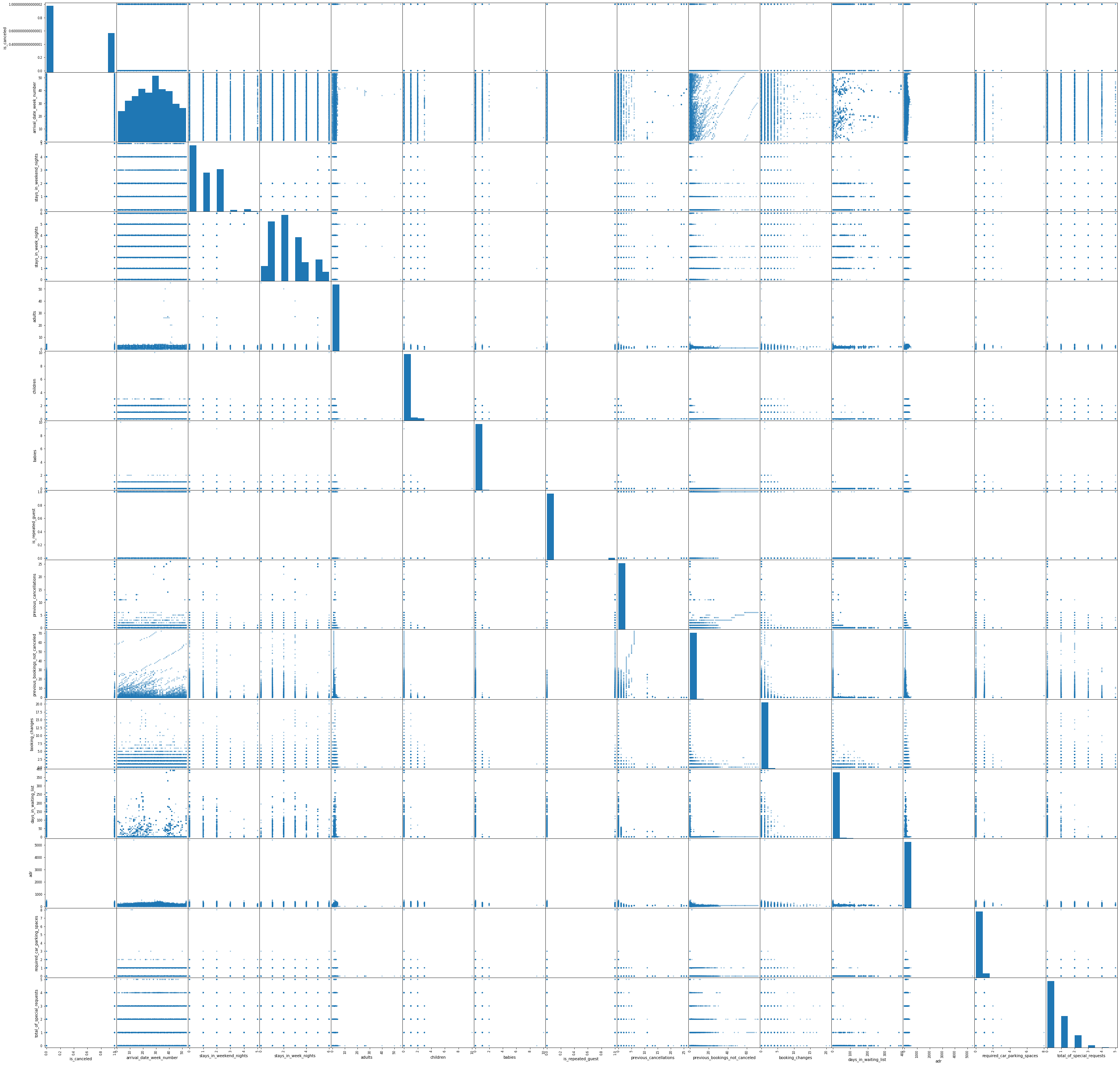


We start by initial evaluation for the factors using chi-square test of independence. We have 10 categorial variables. To deal with them, we encode them into numerical values. The test shows the following results:

|  |  |
| --- | --- |
| 0 | 5.178900e+04 |
| 1 | 0.000000e+00 |
| 2 | 3.886998e+06 |
| 3 | 8.805599e+07 |
| 4 | 1.588225e+05 |
| 5 | 9.070750e+05 |
| 6 | 7.945100e+04 |
| 7 | 3.408000e+04 |
| 8 | 2.284150e+04 |
| 9 | 1.414860e+05 |
| 10 | 2.881548e+07 |
| 11 | 6.288100e+05 |
| 12 | 2.601090e+05 |
| 13 | 2.509400e+04 |
| 14 | 3.495250e+04 |
| 15 | 2.830195e+05 |
| 16 | 3.245070e+05 |
| 17 | 4.587715e+05 |
| 18 | 4.296590e+05 |
| 19 | 2.259200e+04 |
| 20 | 1.033045e+07 |
| 21 | 1.039855e+05 |
| 22 | 5.767138e+11 |
| 23 | 2.967300e+04 |
| 24 | 1.537640e+05 |

The variable number 1 is the 'is\_canceled' field itself. In general, the test results indicates to a potential dependence between the dependent variable and the other nominated independent variables.

We have 15 numerical variables including two of them are actually Boolean but treated as numerical which are is\_canceled variable and is\_repeated\_guest variable. Numerically, it seems that there are a lot of outliers. Logically, it is understandable that most of the reservation would have adults less than ten for example. We would not consider a reservation for group of 50 is an outlier. This also applies to the average daily rate, babies, and car barking spaces. However, for fields like the bookiing\_changes, previous\_booking\_not\_canceled, previous\_cancellations, stays\_in\_week\_nights, and stays\_in\_weekend\_nights, the concept is the important but not the quantity. This means we might look on the impact of changing booking on the reservation cancellation. This impact would be effective from the first and second change. No need for to look at those of 20 changes. we may combine all of more than 5 changes as 5. in other words, we will decrease values more than the outlier threshold to the outlier threshold values in these fields. After that, we can try the scatter plot matrix to have an insight about the relationship between the variables:



The scatter plot matrix initially suggests that there might be a relationship between the cancelation and adults number, babies number. Unexpectedly, it seems that when there are previous cacelations, it is mor likely it will not cancel, while it might cancel with previous uncanceled resrervations. There is also relationship with car parking and booking changes. Numerically, we need to find the correlation between the numerical variables and the dependant variables:

|  |  |
| --- | --- |
| arrival\_date\_week\_number | 0.008148065395052901 |
| stays\_in\_weekend\_nights | -0.0017910780782611744 |
| stays\_in\_week\_nights | 0.024764629045872715 |
| adults | 0.06001721283956815 |
| children | 0.005036254836439323 |
| babies | -0.03249108920833264 |
| is\_repeated\_guest | -0.0847934183570878 |
| previous\_cancellations | 0.11013280822284255 |
| previous\_bookings\_not\_canceled | -0.057357723165947075 |
| booking\_changes | -0.14438099106132224 |
| days\_in\_waiting\_list | 0.054185824117780376 |
| adr | 0.047556597880386124 |
| required\_car\_parking\_spaces | -0.1954978174945085 |
| total\_of\_special\_requests | -0.2346577739690198 |

The correlation analysis reveals that there is significant correlation at alpha = 0.05 with the following columns: adults, is\_repeated\_guest, previous\_cancelations, previous\_bookings\_not\_canceled, booking\_changes, days\_in\_waiting\_list, requried\_car\_parking\_spaces, and total\_of\_special\_requests. At alpha = 0.01, stays\_in\_week\_nights, babies, and adr are added.

The correlation results for the categorial variables were the following:

|  |  |
| --- | --- |
| Hotel | 0.13653126949161642 |
| arrival\_date\_month | 0.011822120071305441 |
| meal | -0.01767760995132292 |
| country | -0.10044912870002404 |
| market\_segment | 0.23833549336078935 |
| distribution\_channel | 0.1697270301121236 |
| reserved\_room\_type | -0.04397743747112238 |
| assigned\_room\_type | -0.1252105041585338 |
| deposit\_type | 0.4804339866053031 |
| customer\_type | -0.13581931980513778 |

Based on these results, we will consider the following variables as factors of reservation cancelling: adults, is\_repeated\_guest, previous\_cancelations, previous\_bookings\_not\_canceled, booking\_changes, days\_in\_waiting\_list, requried\_car\_parking\_spaces, total\_of\_special\_requests, hotel, country, market\_segment, distribution\_channel, assigned\_room\_type, deposit\_type, customer\_type.

For evaluation, two prediction model are created. The one that uses all the variables shows an average precision of 0.82 while the model used the factors shows average precision of 0.85. By exploring the coefficients and eliminating the insignificant effective variables, we remain with the following five factors: is\_repeated\_guest, previous\_cancellations, previous\_bookings\_not\_canceled, required\_car\_parking\_spaces, andn deposit\_type. The model using these factors show precision of 0.86 with 0.98 precision in predicting cancelation.

The factors are used to create a function that can evaluate these factors and predict if the reservation is more likely to cancel or not. To create this function, we start with scoring function. The intercept and coefficients are rounded to 2 decimal level and used in an equation model. The coefficient values are -0.84, 2.21, -0.76, -6.3, and 2.66. The score is then processed in a logistic function. Based on the result of the logistic function, the cancel evaluation function returns true if the probability is more 0.5. Otherwise, it returns false. Finally, the cancel function can be interfaced in the requested form:

Boolean cancel (factorsArray)