## Dataset

## Methodology

We start by having an insight over the mentioned parameters by maintaining the accuracy of prediction models.

We have 31 parameters to consider. In order to eliminate some of the insignificant factors, we start by processing the correlation test. As set the significance threshold to α =0.05 = 5%. The parameters pass the numerical correlation are: lead\_time 0.293123, adults 0.060017, is\_repeated\_guest -0.084793, previous\_cancellations 0.110133, previous\_bookings\_not\_canceled -0.057358, booking\_changes -0.144381, agent -0.083114, days\_in\_waiting\_list 0.054186, required\_car\_parking\_spaces -0.195498, and total\_of\_special\_requests -0.234658.

However, when considering the categorial values, the result changes to the following: hotel 0.136531, lead\_time 0.241626, adults -0.055894, country -0.097097, market\_segment 0.238335, distribution\_channel 0.169727, is\_repeated\_guest -0.084793, previous\_cancellations 0.140984, previous\_bookings\_not\_canceled -0.057426, assigned\_room\_type -0.125211, booking\_changes -0.121746, deposit\_type 0.480434, agent 0.050608, days\_in\_waiting\_list 0.106141, customer\_type -0.135819, required\_car\_parking\_spaces -0.196524, total\_of\_special\_requests -0.215793, reservation\_status 0.980601, reservation\_status\_date-0.134554.

From these coefficient numbers, we can see the most significant relationship occurs between the reservation status which reflects the cancelation status. We eliminate this from factor list as it is a mere consequence. We can find that the deposit type has a high coefficient of 0.48. There are also lead\_time and total\_of\_special\_requests which have coefficient higher than 0.2. Finally, we eliminate the reservation\_status\_date as it is a metadata related to an eliminated field which indicate that its coefficient is a consequence of an eliminated coefficient too.

The obtained accuracy is about 0.

## Discussion