**Executive Summary and Implications**

A.  Develop an executive summary using your data and results from task 2. The summary should be written for a technical audience of your data-analytics peers or members of your team and should include each of the following:

•   a statement of the problem and the hypothesis

•   a summary of the data-analysis process

•   an outline of the findings

•   an explanation of the limitations of the techniques and tools used

•   a summary of proposed actions

•   expected benefits of the study (be as specific and quantitative as possible)

Our research objective for this project was to take our public dataset on hotel bookings retrieved from Kaggle and use it to test if we can create a model that can predict hotel booking cancellations. Since this dataset came with a binary classification of whether or not a booking was cancelled, it’s more or less ideal for a straightforward regression model. As such, we will be using multiple logistic regression to see if it can accurately predict booking cancellation using the following dependent variables:

lead\_time: Number of days that elapsed between the entering date of the booking into the PMS and the arrival date

adults: Number of adults

children: Number of children

babies: Number of babies

previous\_cancellations: Number of previous bookings that were cancelled by the customer prior to the current booking

booking\_changes: Number of changes/amendments made to the booking from the moment the booking was entered on the PMS

days\_in\_waiting\_list: Number of days the booking was in the waiting list before it was confirmed to the customer

adr: Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights

In order to test which variables are the most important, we will be running a few tests of significance initially. As such, our null (or original) hypothesis for this study will generally be that a predictive model cannot be created. Whereas our alternative hypothesis will be the inverse, that a predictive model can be successfully created with a reasonable degree of accuracy.

Moving on to our data analysis process, we first evaluated the data for any truncated or missing values. Once they were identified we removed them, along with any duplicates and standardized the data using z-scores. We do this in the event that any data is non-continuous and could potentially skew our tests of significance- which require continuous numerical values. As such, our study with be focusing on these specific types of variables. After observing our univariate features visually for any surface level insights, we ran our tests of significance (Principal Component Analysis and 1-way ANOVA) in order to sort our numerical variables by degree of significance. Our PCA helped identify that we should only be using around 5 features to avoid overfitting in our analysis. Combined with our 1-way ANOVA and a correlation matrix, we were able to narrow down that the most important numerical components of the dataset (from greatest to least) are:

adr, lead\_time, booking\_changes, days\_in\_waiting\_list, children, and adults

We then created bivariate histograms to compare our sorted features with our categorical response variable ‘is\_canceled’. Interestingly, this revealed a sharp distinction in booking cancellations among groups of 10 or more adults, and this finding was included in our final results.

Next, we were ready to create our predictive model using our sorted list of variables. We first created an initial model, and then a reduced model to compare the results. Since our former model returned a higher accuracy (lower AIC value), it was used on our test set to validate our results.

This leads us to our findings. We first ran a confusion matrix, which returned an accuracy of around 73% and then verified the results with an ROC Curve – which returned an AUC (area under the curve) value of around 0.671. Through this we can infer that the model has a high rate of sensitivity (true positive rate) but low specificity (true negative rate). Furthermore, we are able to reasonably predict whether or not a customer will cancel a booking based solely on a few numerical predictors with a relatively high degree of accuracy.

In terms of limitations to this study, a few come to mind. First, we would have compared different predictive models or advanced regression techniques, but that was outside the scope of this study. Techniques like lasso regression are particularly useful in isolating variables of significance quickly and efficiently. We could have also used categorical features when running our predictions, and that’s a viable approach for future study. We also didn’t include the element of time in our study, therefore seasonal and cyclical variations in bookings were not considered.

So as proposed actions for the future, one of the first things we could do is include categorical features as a method of approach for the future – as well as analyze the data as a time series for deeper insights. As far as propositions for the business, we could notify hotel administration of the risk associated with confirming bookings with 10 or more adults – as well as the influence that our most significant variables have on booking cancellation.

Lastly, the expected benefits of this study include the potential to reduce intentional overbookings by focusing on large parties involving 10 or more adults. Additionally, we could propose a substantial method of identifying ‘at risk’ bookings by considering the 3 most statistically significant features: number of booking changes, lead time, and average daily rate. Implementing both of these recommendations would effectively eliminate the need of having to double book as many rooms, and therefore frees up more room for potential revenue for the hotel.