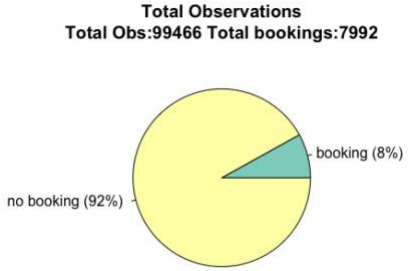
# Customer Segmentation

The goal is to use data analytics to create a profile that captures characteristics for different customer groups. Our data is from a hotel booking company that was supplied for a Kaggle competition and can be found in the data folder of this project. This data will be used to make recommendations to explore ways to increase the booking rate by efficiently using marketing channels and customizing campaigns to meet the needs of different customer segments.

First, the source.R script was opened and the working directory was set. The packages needed for the analysis were then initialized. The data was loaded and explored using the explore.R script. The data has 100,000 observations and 24 columns. The important columns in this data set include information about booking date, mobile bookings, marketing packages, marketing channels, check-in/out dates, number of adults/children, number of rooms, destination distance, user origin city and other relevant information.

The data was tested to make sure there were no obvious logical errors. This would include issues such as the check-in date being after the booking date, booking dates after the check-out date, check-in dates after check-out dates, a room booked for a child without an adult, etc. There are observations in the data where there are no check-in or check-out dates. After confirming no bookings were actually made for those entries, they were removed from the data set. There were check-out dates that were scheduled before the check-in dates which is not possible, therefore these entries and were also removed. After looking at the frequency of the duration of the stay, outliers that had excessive stays (longer than 43 days) were also removed from the data. After checking if there were any bookings made with negative days-notice, I saw that -1 days-notice had bookings, the others observations with larger days of negative notice did not have bookings made. There is the potential that the booking date may be 1 day prior to check-in when there's a 1am check-in. In this case the check-in time is considered to be the late afternoon the previous day even if it was booked at 1am. Any days-notice of more than one day prior was deleted from the data. Any rooms booked for a child without an adult was also deleted from the data as this is not possible under most hotel rules.

After removing the data that did not pass the logic check, the data sample sizes were explored. Here is a graph exploring these values:

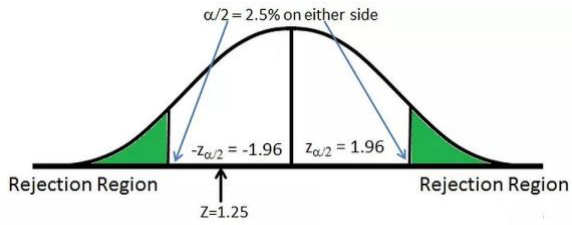



The first area explored were the underperforming and overperforming marketing channels. We can run a 2-sample 2-tailed t-test to identify these channels. First, we aggregate the data so that we have one row for each marketing channel and a column for the average booking rate and number of bookings. Then, we calculate the average booking rate and number of bookings for all channels combined. We can

look at the average booking rate for a specific channel and compare it to the overall booking rate, but this will not tell us if it's a under or over performer because there is random sampling error or if there's a real statistical underlying difference between the two groups. We will perform a 2-sampled t-test, the first group is one individual marketing channel and the second group contains the remaining channels, therefore we calculate the booking rate and the number of bookings for the second group, and use these values to calculate the z-score and p-values. We use these p-values to find which channels are statistically significant at the 95% confidence level. Our hypothesis is:

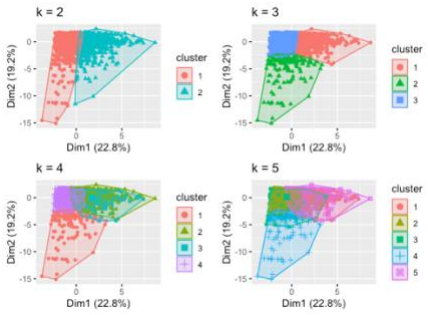
$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$

The null hypothesis is that the channel has no statistical difference between the two groups and the alternative hypothesis is that the two groups are in fact statistically different. If the p-value is less than (1-.95)/2=0.025, we can reject the null hypothesis.

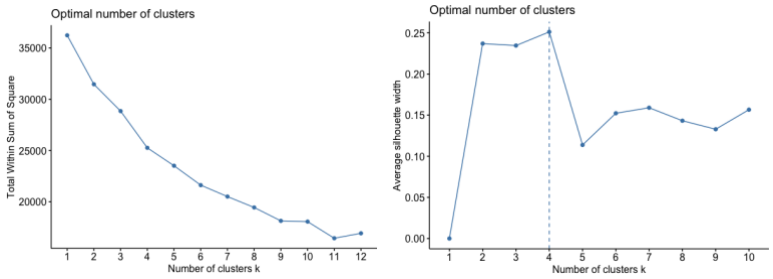



Result: The marketing channels where we cannot reject the null, are marketing channels: 6, 8, and 10, meaning they are not statistically significant at a .95 confidence level. The statistically significant under performers are marketing channels: 0,1,2,3, and 7. The statistically significant over performers are marketing channels: 4,5, and 9. Knowing the overperforming and underperforming channels will allow us to better utilize our marketing budget.

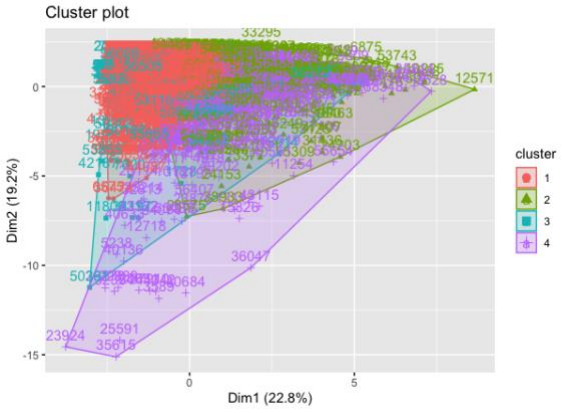
Next, we want to identify which cities have similar customers in order to customize different marketing campaigns to different groups of cities. Looking at the list of variables, I selected eight features that I think are important in distinguishing different customer characteristics between cities and would help us understand how to target these cities. I chose duration, notice, destination distance, mobile, package, adult count, children count, and room count. We then aggregated the data so each row was a unique city and the columns are the averages of the eight variables per city. We standardized the data using the scale function and perform a K-means analysis on this data to get the clusters of cities that are similar. The number of clusters should not be so large that we don't have enough samples in each group, or that there's too small a difference between the customers, and at a certain point more groups does not necessarily help us fit the model better and takes more computational time. Also, the main goal is to target these groups with different campaigns and we would not be able to run a large number of campaigns at one time due to budget and resource limitations. We can explore with different numbers of clusters.



After exploring and looking at the elbow and silhouette plot, the optimal number of groups was 4.




We used the *fviz_cluster* function to visualize the groups. It does this by performing principle component analysis on the data and reduces the dimensions so that we can have the best visual representation in a two-dimensional plot.




,

We aggregate the data so each row is a specific cluster and look at the averages of all of the variables to get a general idea of the customers in each cluster.

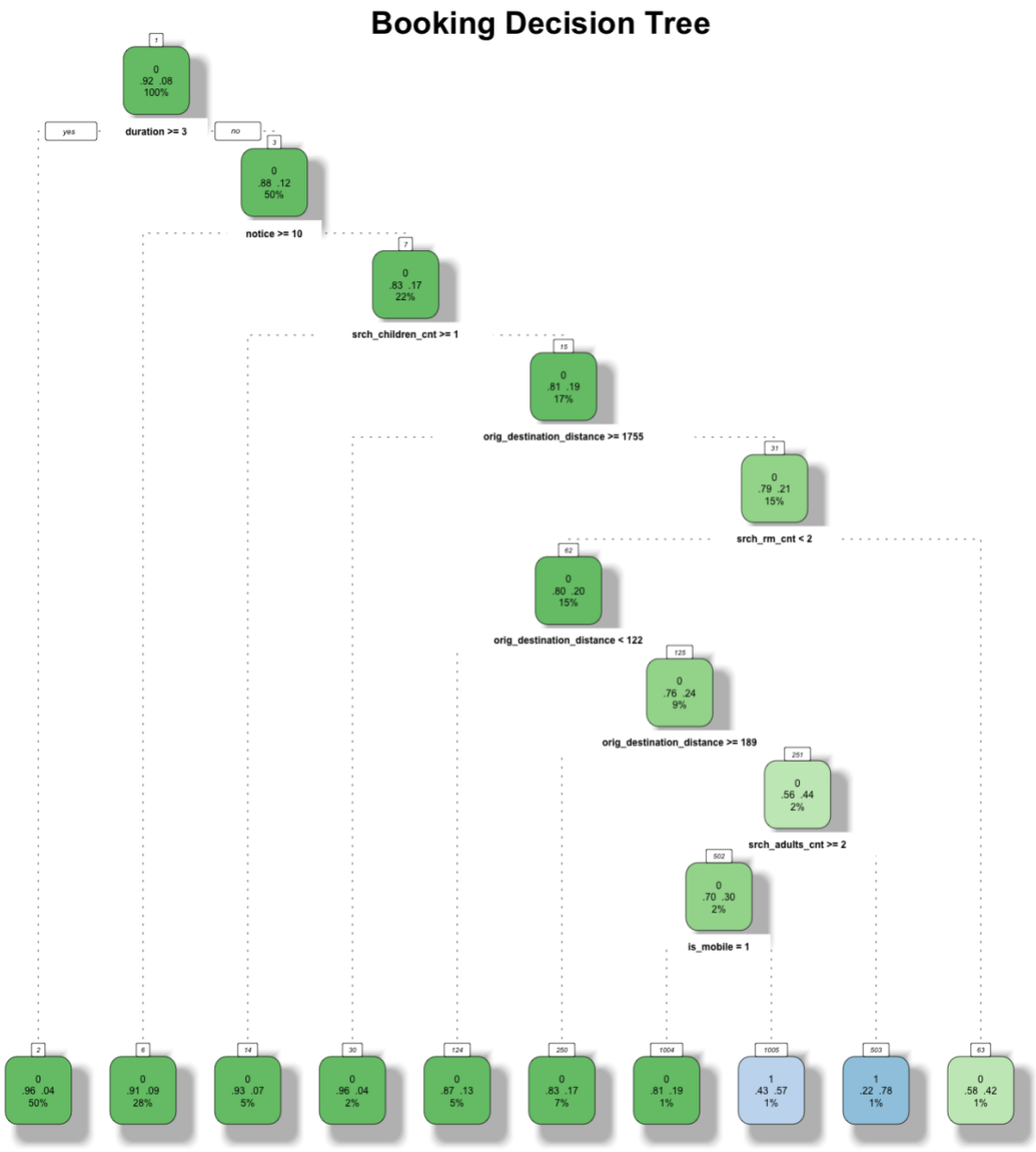
| duration | notice | orig_destination_distance | is_mobile | is_package | srch_adults_cnt | srch_children_cnt | srch_rm_cnt | cluster |
|---|---|---|---|---|---|---|---|---|
| 3.153452 | 51.18560 | 1828.967 | 0.12848065 | 0.2262066 | 2.027774 | 0.3529023 | 1.097762 | 1 |
| 4.605429 | 76.54055 | 1940.651 | 0.11022840 | 0.5723270 | 2.034757 | 0.3333333 | 1.099305 | 2 |
| 3.081081 | 52.90837 | 1475.341 | 0.54845089 | 0.2715887 | 2.030323 | 0.3559657 | 1.086355 | 3 |
| 4.334547 | 93.01158 | 4644.312 | 0.09033752 | 0.1700860 | 2.178028 | 0.3759100 | 1.187624 | 4 |

From this table we see that on average cluster 4 is not likely to use their phone and gives the most days-notice. Cluster 3 is most likely to search bookings from their phone and gives less days-notice, cluster 2 is most likely to search for bookings with a package. I created a mapping data frame of which cities belong to which cluster and merge this data back into the original data frame.

Now we can look at a decision tree to understand what causes people to choose to book. I decided to look at cluster number 3 so that we can learn how to target this group. The data is split into two data frames, a training set (75% of the total data set) and a testing data set (the remaining 25%). The training data set will be used to build a classifying algorithm that can accurately predict the testing data set and then compare the predicted results with the actual results of the test set. Next, we look at the sample sizes of the two data sets to make sure there are enough in each category (either booked or not booked).

| | | |
|---|---|---|
| Train | No Booking | 1267 |
| Train | Booking | 112 |
| Test | No Booking | 428 |
| Test | Booking | 32 |

The sample sizes are not as large as I'd hoped, so we adjusted the complexity parameter in the *rpart* function in order to grow the nodes on the tree. This may cause overfitting and poor results, but may give us a general idea of how a customer behaves.




Cluster 3

Each node shows the predicted class (0=no booking, 1=booking), the predicted probability of booking, and the percentage of observations in the node. This decision tree model gives a 92.61% accuracy rate on predicting bookings on the testing dataset for cluster 3. Looking at this tree if someone in a city from cluster 3 searches for hotels and plans for a stay longer than 3 days we know they're automatically less likely to make the booking. In this case we can target them with a mobile ad or booking deal since this group is more likely to book via mobile device.