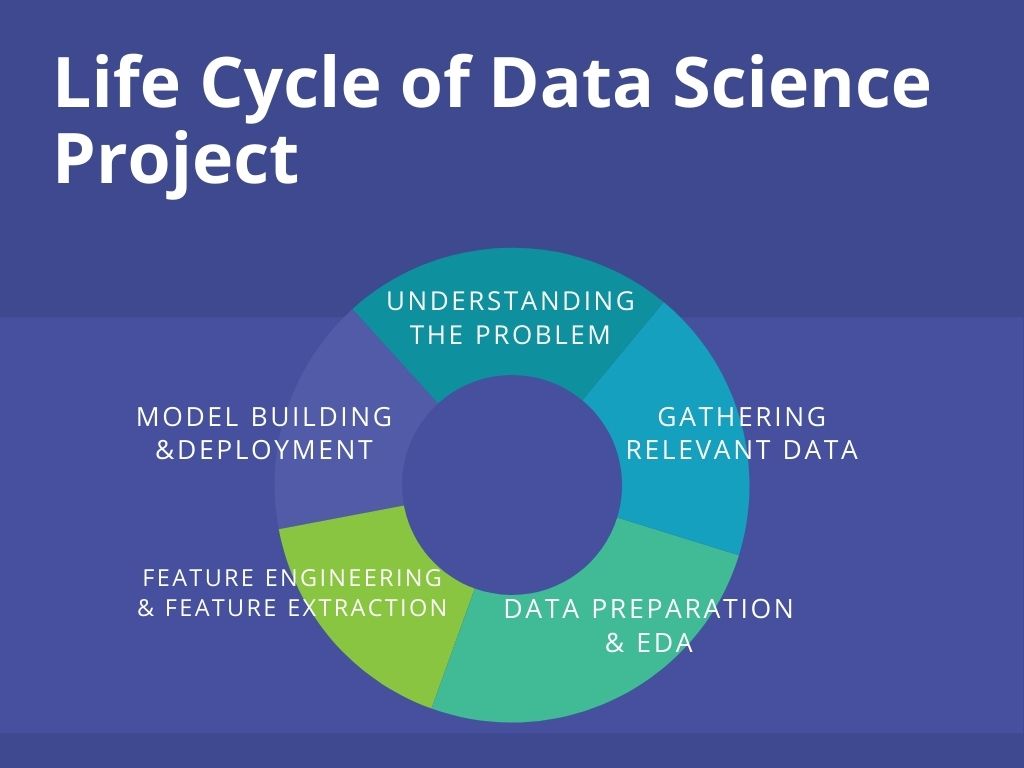
**IST - 687**

**FINAL PROJECT REPORT**

Introduction**:**

As consultants for this hotel, we are analyzing hotel cancellations and the reasons behind them to try to predict them and help reduce them. To do so, we create visualizations and models to explore the data and try to create actionable insights for the hotel management to use in response to these cancellations to try to reduce their occurrences.

We are following the Data Science Life Cycle Model:



**Step 1: Understanding the problem :**

At first, we understood the problem and noted down all the points. We decided on our approach to the problem. We did brainstorming and understood what business questions have to be solved.

**Step 2: Gathering the relevant data :**

We obtained the data from the link below:

https://intro-datascience.s3.us-east-2.amazonaws.com/Resort01.csv

Apart from this, we also referred to a few resources on the internet.

**Step 3: Data Preparation and EDA :**

The most crucial step is data cleaning and data preparation when we gather the data. The data might be having missing values; the dataset might also have some outliers. Therefore, we have to clean the data and select only relevant attributes.We also performed the Exploratory Data Analysis where used the basic functions such asstr(),summary(),table() and some advanced plots to understand the data

**Step 4: Feature Engineering and Feature Extraction :**

Feature engineering is a process where we decide which attributes in the hotel data set can highly contribute to cancellations and at the beginning worked only on those attributes and later explored other attributes which might have influenced the cancellations.

**Step 5: Model Building and Deployment :**

In our project we have tried and implemented SVM, Decision Tree, Association Rule Mining and Logistic Regression Model.

1. **Business Questions**

After analyzing and understanding the problem statement we have come up with the following business questions and presented it one by one.

1. Are repeated guests more likely to cancel their bookings?

2. Are cancellations more on week nights or weekend nights?

3. Which countries have the most and least amount of repeated customers?

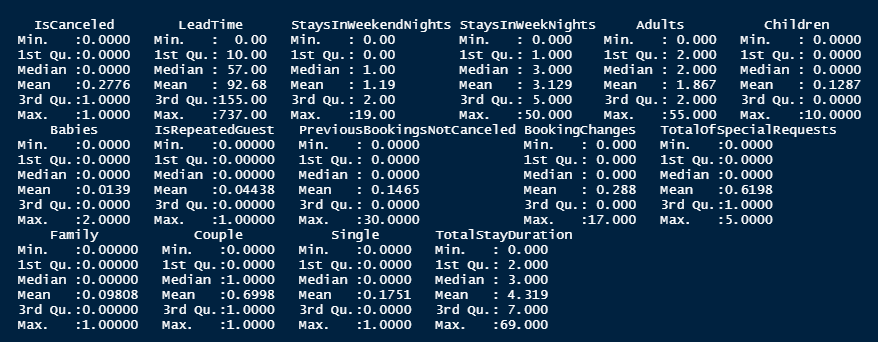
4. Does weekend or weekday stay have an impact on customer cancellations?

5. Are there any booking changes/cancellations if a customer is not assigned their reserved room type for their stay?

6. Are there any booking cancellations if a repeating customer doesn’t get assigned a parking spot?

7. From which country people are traveling the most?

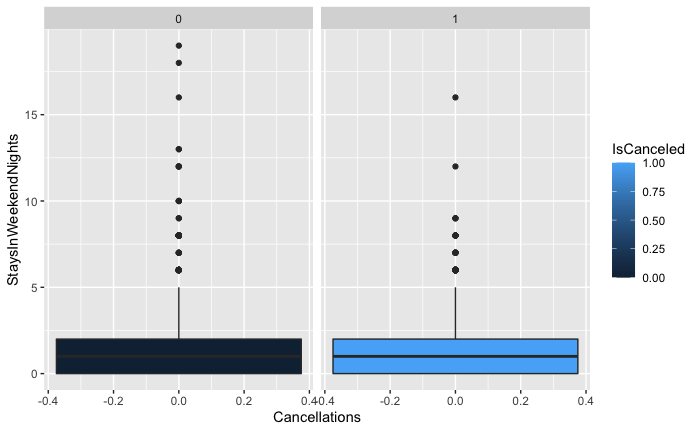
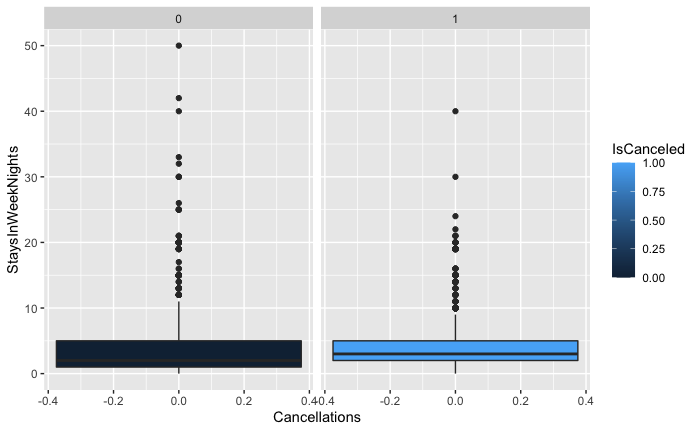
1. **Initial Visuals(Maps/Boxplots)**
2. **Summary Statistics:**

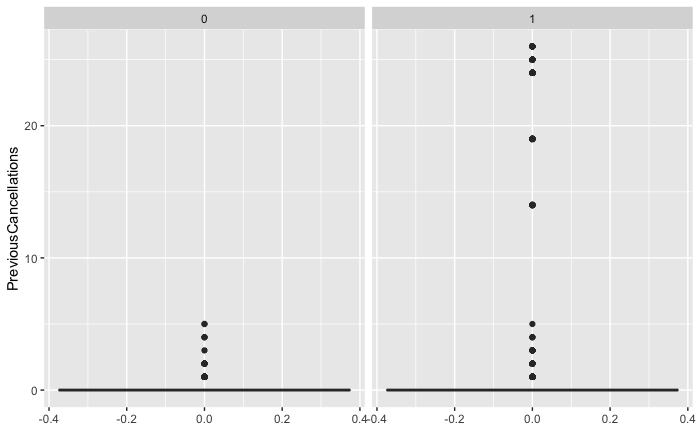
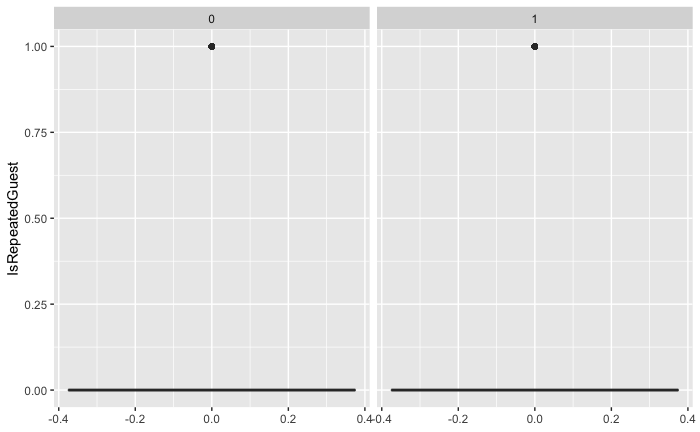


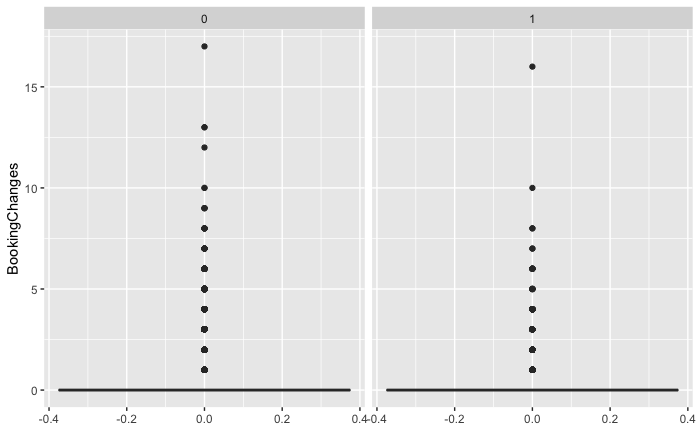
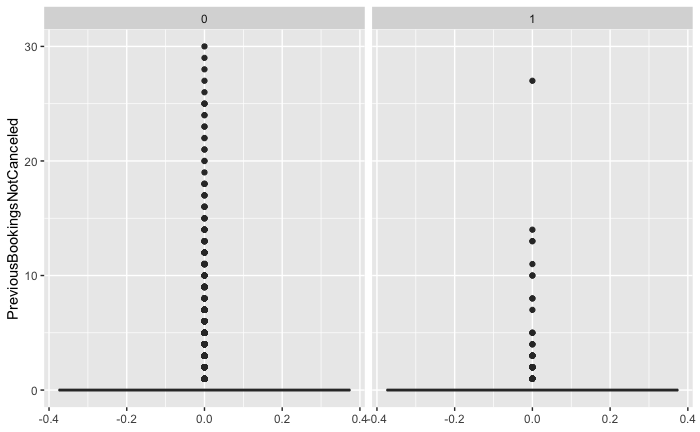
1. **Box Plots, Violin Plots and Scattered for Numeric Values:**

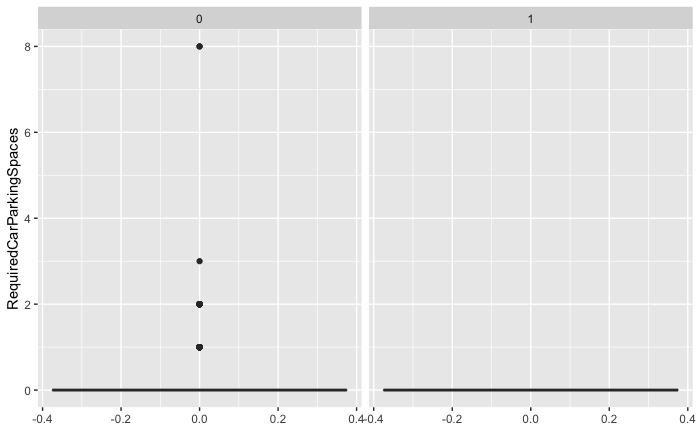
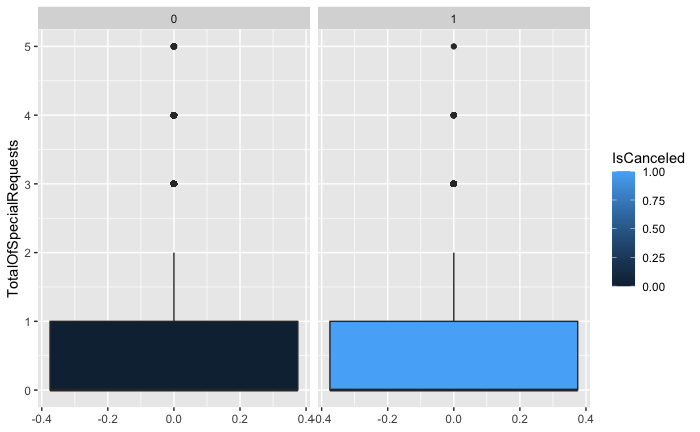
We created numerous box plots with single characteristics to see if we could glean any information. As a result, after constructing box plots for numeric data, LeadTIme provided us with only one insight, which is described below.

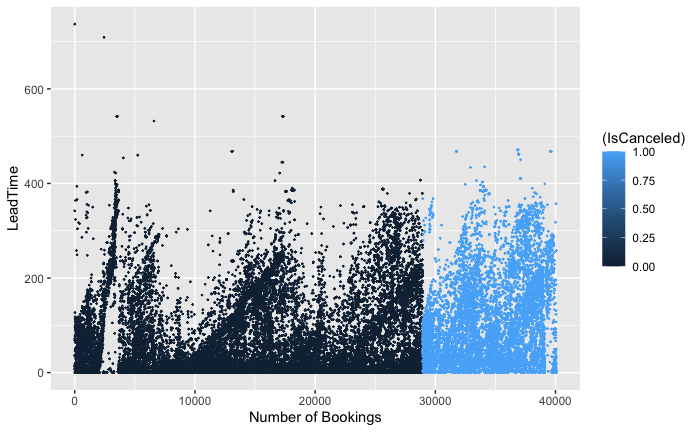
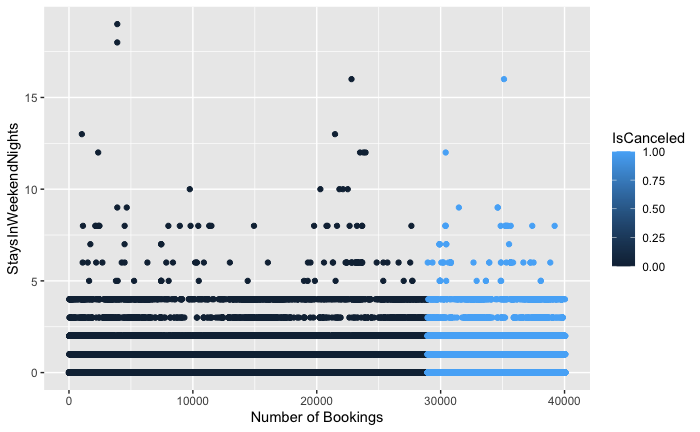
**Plots with no insights:**

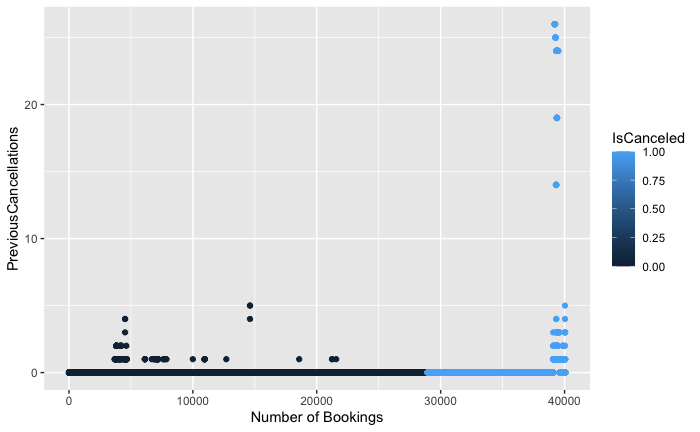
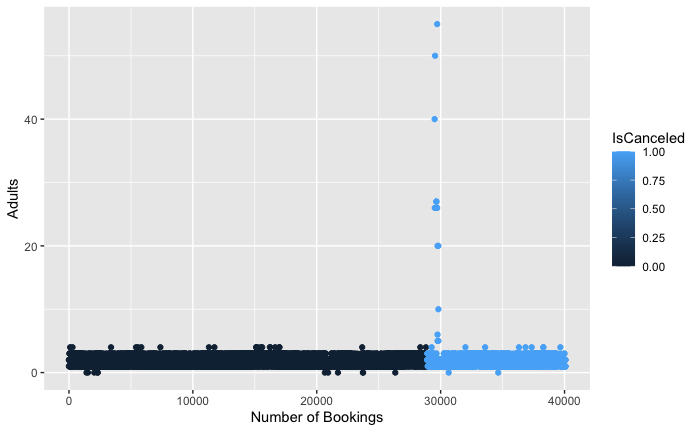




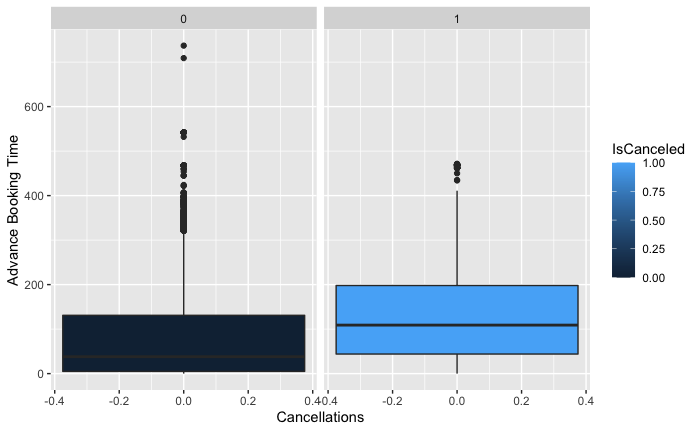




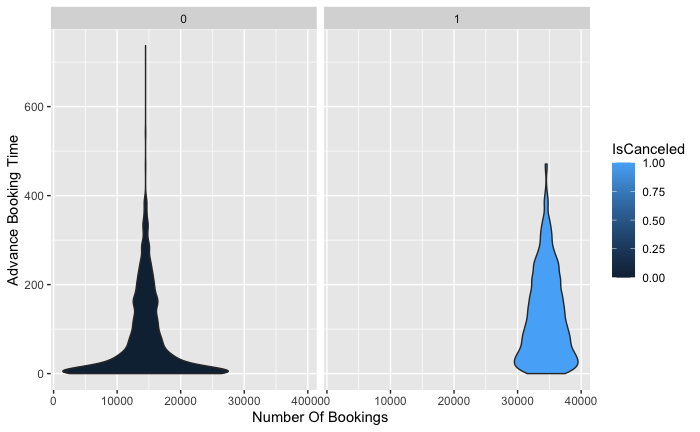


**Plots with Insights:**

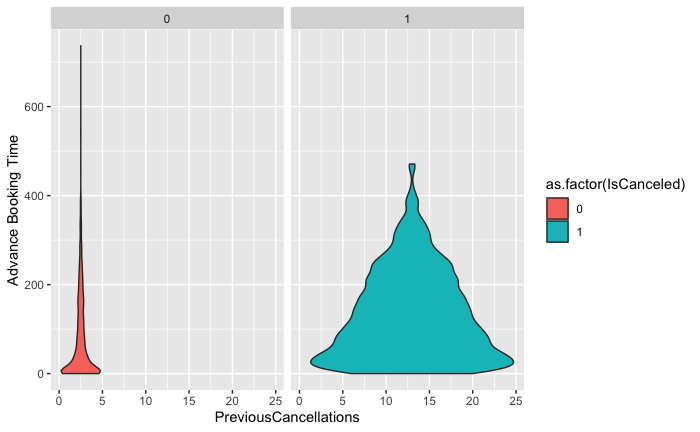


The lead time is the number of days between booking a hotel and the actual arrival date.

As the lead time increases, the percentage of cancellations rises. The percentage of cancellations climbs as the lead time increases. We may deduce from the box plot that cancellations occur if the booking is made more than 100 days in advance. If it is booked within the next 20 days, it is less likely to be canceled.

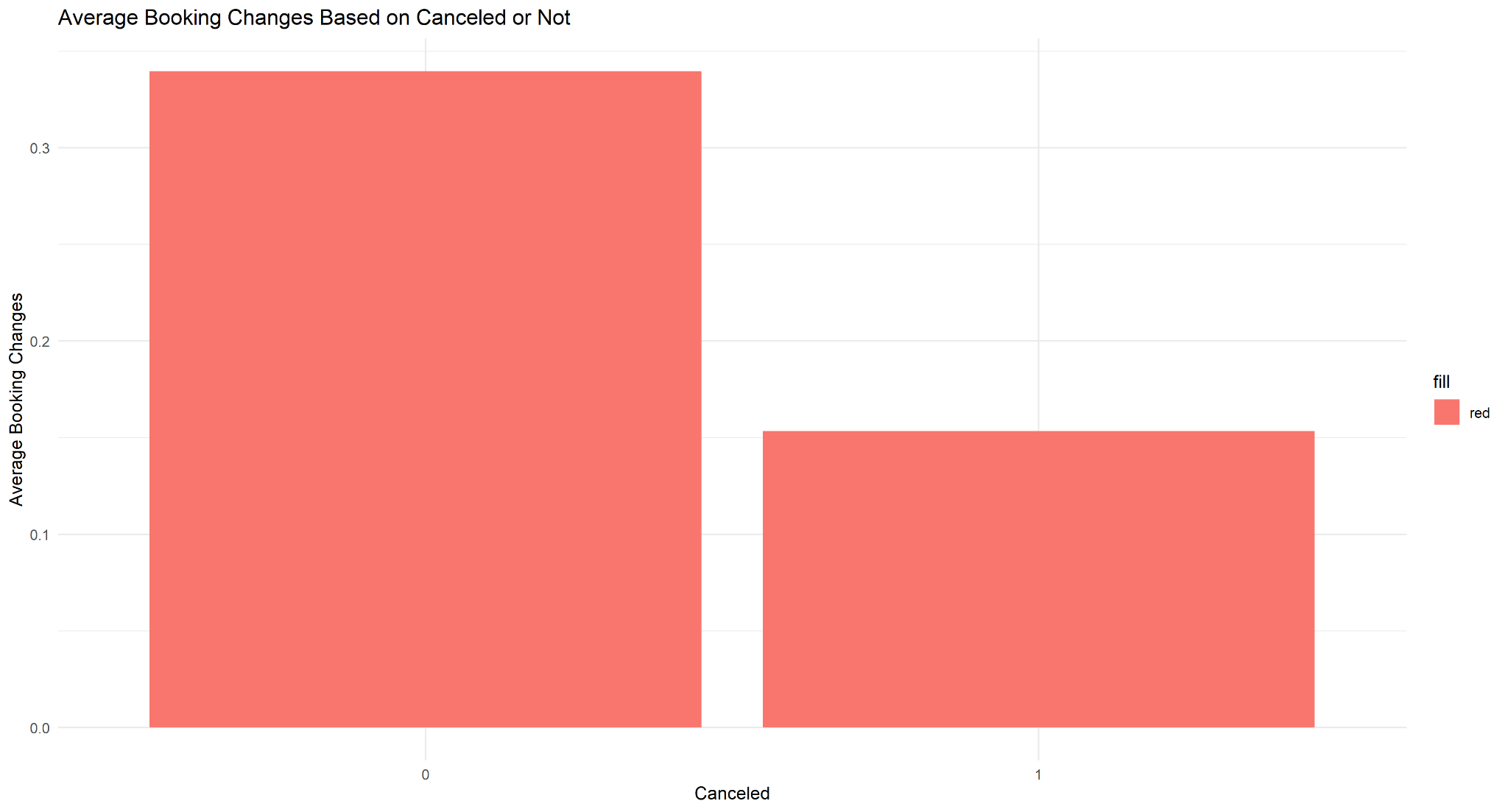


This is a Violin Plot which depicts the same thing in the above graph with a different approach where we can see that as the lead time increases, the percentage of cancellations rises.

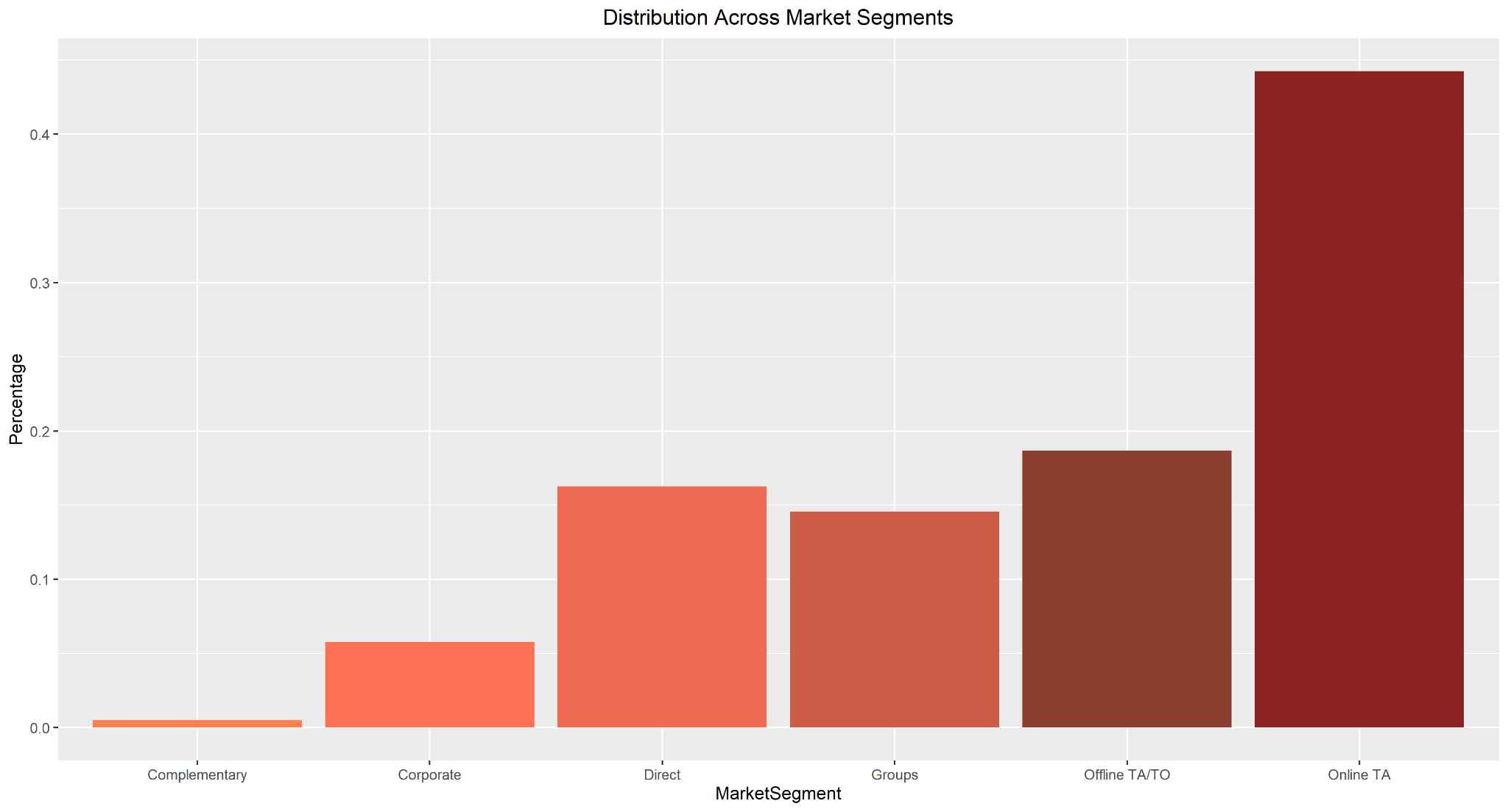


People who have had more than two cancellations with appointments made less than 20 days ago are more likely to cancel the booking, as shown by this violin plot.

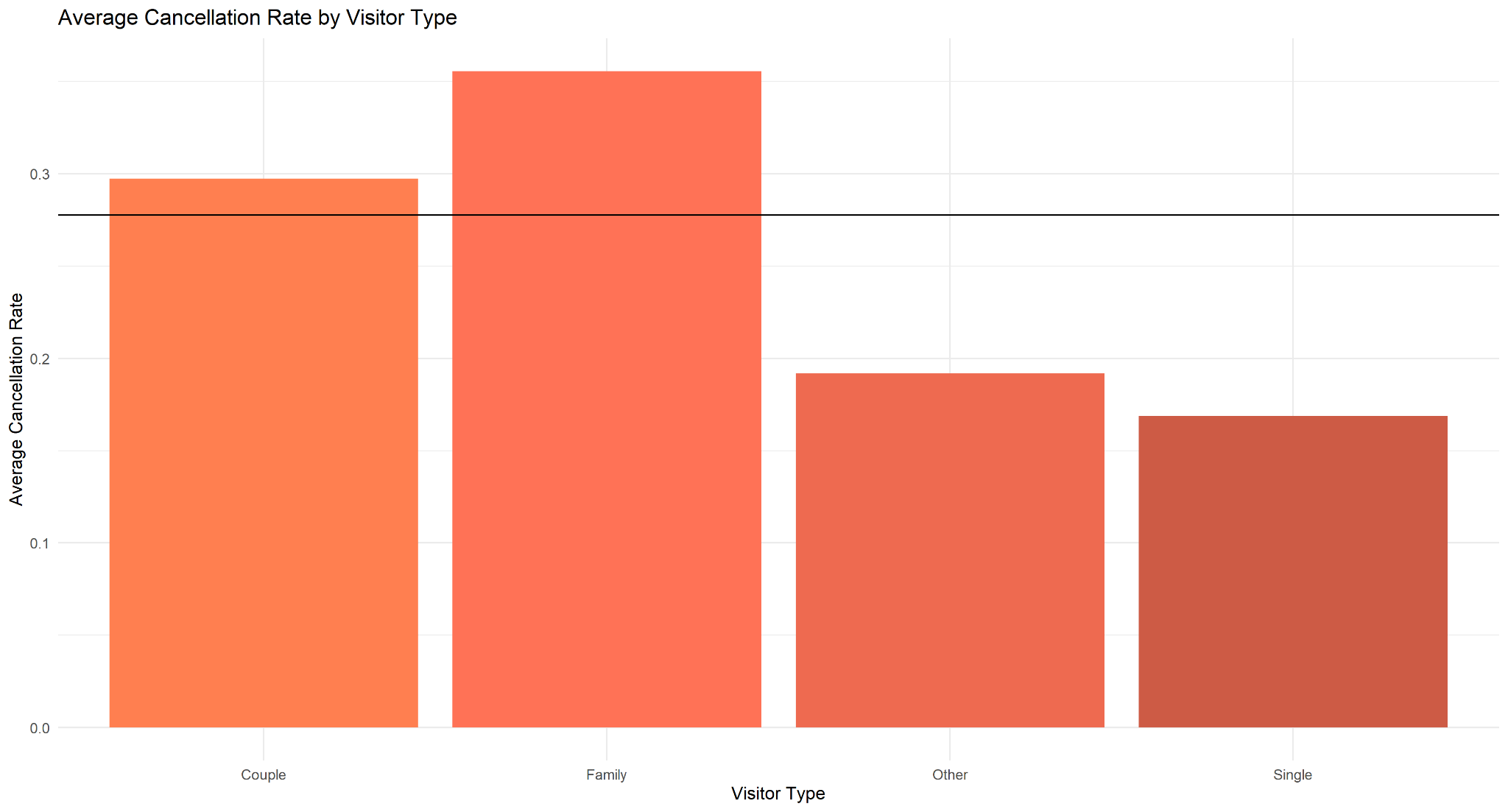
1. **Bar Plots for Categorical Value and derived attributes: (Single Attribute)**



This visual shows the average number of booking changes, broken down by whether the reservation got cancelled or not. This shows that customers that make more booking changes are less likely to end up cancelling their reservation.



This shows the proportion of customers that falls under each market segment. Online travel agents make up almost half of the total market, while complimentary makes up almost none.



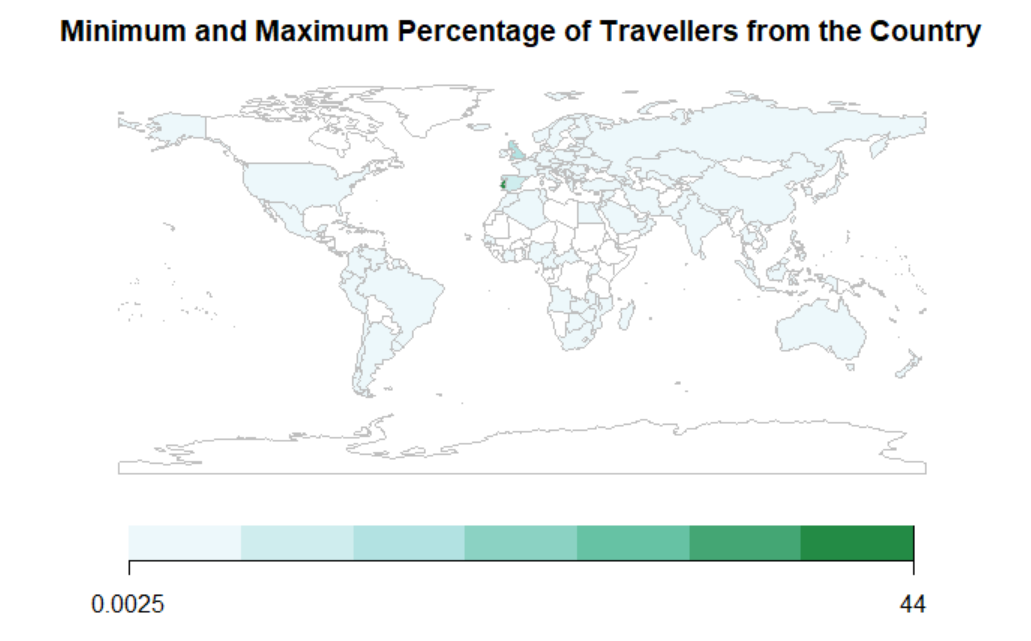
This visualization illustrates the average cancellation rate for each visitor type in the dataset, as well as including a comparison to the overall average cancellation rate for the dataset as a whole. Families and couples both cancel at a higher rate than the overall average, while the Single and Other groups cancel at a lower rate than the average.

1. **Maps:**

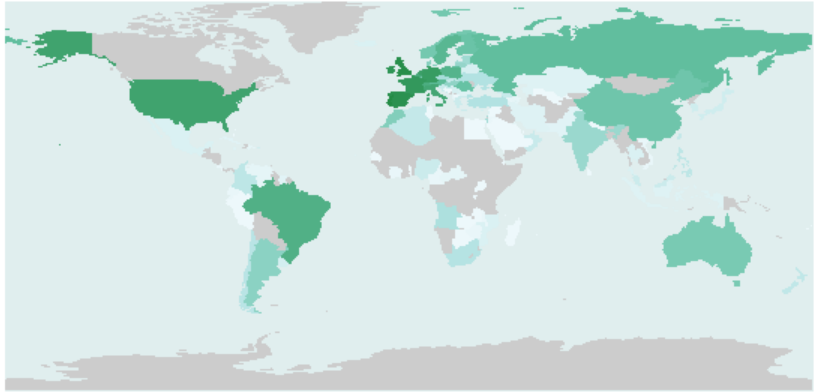
Cartography, or the process of designing maps, is an ancient skill that requires communication, intuition, and some imagination. In R we can create the maps using r functions and packages such as library(rworldmap),library(dplyr),joinCountryData2Map and many other functions and plots. Maps are best to represent critical analysis and we have represented three maps as below.

We have plotted a World Map which presents two countries.

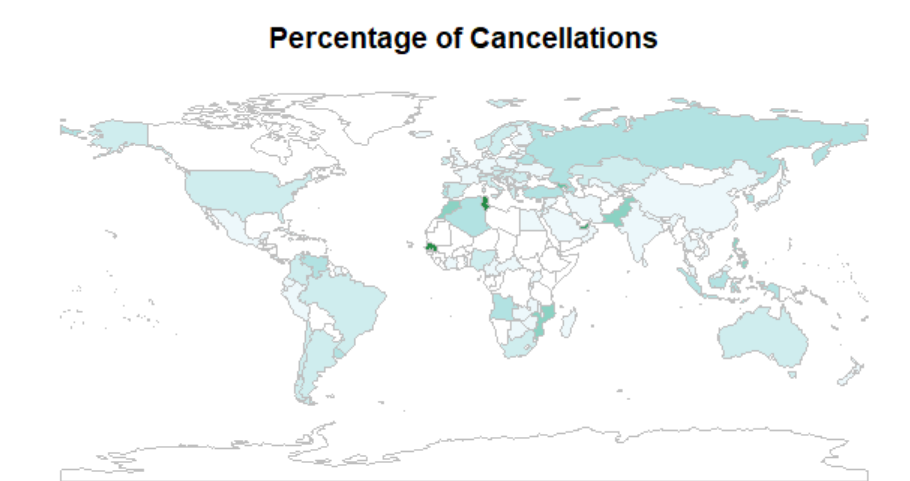
1. The country from which we have minimum number of travelers
2. The country from which we have maximum number of travelers



To dive deeper and focus on other countries we plotted a map which represents the number of travellers from all the countries present in the dataset.



The last graph is the most important graph and this graph represents the cancellation rates of each country



1. **Data Acquisition, Cleansing, Transformation, Munging**

CODE USED TO CLEAN AND TRANSFORM THE DATA AS NEEDED (ProjectCSV is the df that the CSV was read into initially):

ProjectCSV2 <- ProjectCSV[,-c(9,14,15)]

ProjectCSV2$Family <- ifelse(ProjectCSV2$Children > 0 | ProjectCSV2$Babies > 0, 1,0)

ProjectCSV2$Couple <- ifelse(ProjectCSV2$Adults == 2 & ProjectCSV2$Family == 0, 1,0)

ProjectCSV2$Single <- ifelse(ProjectCSV2$Adults == 1 & ProjectCSV2$Family == 0, 1,0)

ProjectCSV2$TotalStayDuration <- ProjectCSV2$StaysInWeekendNights + ProjectCSV2$StaysInWeekNights

ProjectCSV2$VisitorType <- ifelse(ProjectCSV2$Family == 1, "Family", ifelse(ProjectCSV2$Couple == 1,"Couple",ifelse(ProjectCSV2$Single == 1,"Single", "Other")))

ProjectCSV2$Booking <- 1

ProjectCSV2$RequiredCarParkingSpaces[ProjectCSV2$RequiredCarParkingSpaces>1] <- "2+"

THIS PART IS FOR LOOKING AT ROOM CHOICES BECAUSE THAT PART WAS REMOVED FOR A SUBSET TO ALLOW THE SVM MODEL TO RUN

ProjectCSV$Family <- ifelse(ProjectCSV$Children > 0 | ProjectCSV$Babies > 0, 1,0)

ProjectCSV$Couple <- ifelse(ProjectCSV$Adults == 2 & ProjectCSV$Family == 0, 1,0)

ProjectCSV$Single <- ifelse(ProjectCSV$Adults == 1 & ProjectCSV$Family == 0, 1,0)

ProjectCSV$TotalStayDuration <- ProjectCSV$StaysInWeekendNights + ProjectCSV$StaysInWeekNights

ProjectCSV$VisitorType <- ifelse(ProjectCSV$Family == 1, "Family", ifelse(ProjectCSV$Couple == 1,"Couple",ifelse(ProjectCSV$Single == 1,"Single", "None")))

ProjectCSV$DoNotGetRoomChoice <- ifelse(ProjectCSV$ReservedRoomType == ProjectCSV$AssignedRoomType,1,0)

CREATES A SUBSET JUST INCLUDING REPEATED GUESTS

ProjectCSV2 %>%

filter(IsRepeatedGuest == 1) -> RepeatGuestSet

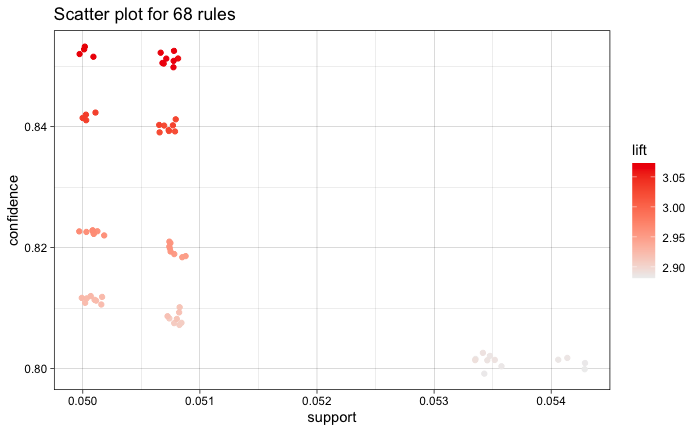
CREATES A SUBSET JUST INCLUDING COUPLES

ProjectCSV2 %>%

filter(Couple == 1) -> CouplesSet

1. **Models**
2. Association Rule Mining :

We had 68 rules but the most significant were 24 rules with a lift more than 3.



**ASM Model Visualization:**

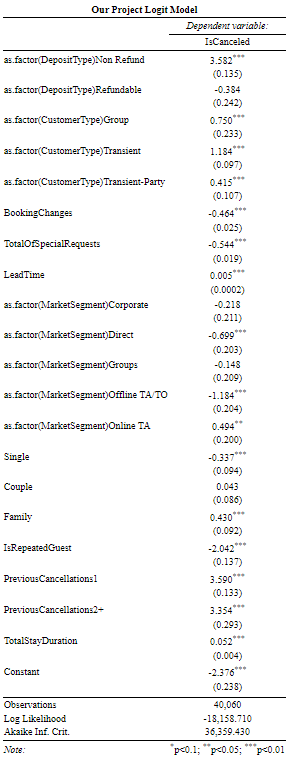
We made a graph of the rules so that we could see which ones have the most lift. Following the achievement of the maximum lift, We devised a set of rules with a lift greater than 3.00. Then, gathered a total of 24 rules with a lift greater than 3.00 and visualized the following points from them.

* Country Portugal has the highest number of cancellations, as well as the highest number of bookings.
* The majority of cancellations are made by individuals or groups, as the bookings had no children or newborns, according to the ruleset.
* The Group segment is the market segment with the highest cancellations.
* Before canceling a booking, most people do not make any changes to it. As the rule showed that the bookings had 0 changes.
* In general, guests who cancel their bookings are not a repeated guest
* The number of special requests made by customers who cancel their bookings is 0.
* People have canceled the bookings due to no parking space.
* Previous Booking Not Canceled = 0

1. Logit Model:

ProjectLogitModel <- glm(IsCanceled ~ as.factor(DepositType) + as.factor(CustomerType) + BookingChanges + TotalOfSpecialRequests + LeadTime + as.factor(MarketSegment) + Single + Couple + Family + IsRepeatedGuest + PreviousCancellations + TotalStayDuration + I(LeadTime^2), data=ProjectCSV2, family="binomial")

summary(ProjectLogitModel)



**There are a lot of significant variables in this model:**

Variables that have significant positive effects on the likelihood of cancellation (make it more likely they cancel) are: The total stay duration, having previously canceled a booking, the reserving party being a family (as opposed to Other), booking through an online travel agent instead of Complimentary, booking more time in advance of the day of reservation, being a group, transient, or transient party (all as compared to contract), or making a non-refundable deposit instead of no deposit at all.

Variables that have significant negative effects on the likelihood of cancellation (make it less likely they cancel) are: Being a repeat guest, being a single traveler as compared to the Other group, booking through direct or an offline travel agent instead of complimentary, or having more special requests or booking changes.

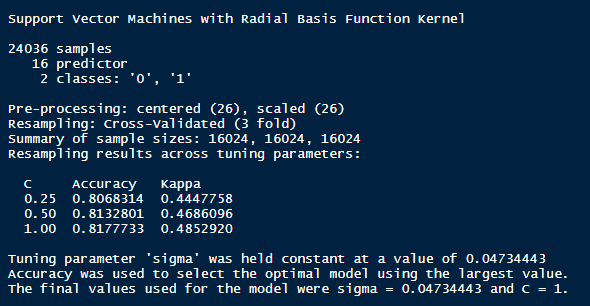
Overall, the logit model shows a solid amount of significant variables that have a clearly measured effect on customers’ cancellation decisions.

SVM:

tr\_control <- trainControl(method = "cv", number = 3)

ProjectSVMModel2 <- train(as.factor(IsCanceled) ~ ., data=ProjectTrainingSet, method="svmRadial", trControl = tr\_control, preProc=c("center","scale"))

ProjectSVMPred <- predict(ProjectSVMModel2, newdata=ProjectTestSet, type = "raw")

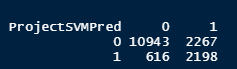


* Confusion Matrix

ProjectConfMatrix <- table(ProjectSVMPred,ProjectTestSet$IsCanceled)

ProjectConfMatrix

sum(diag(ProjectConfMatrix))/sum(ProjectConfMatrix)

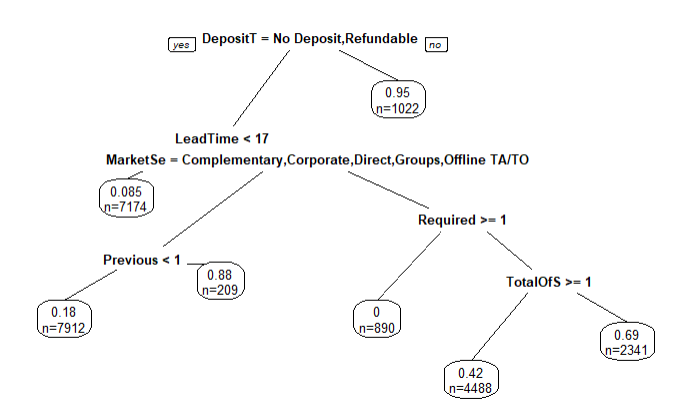


* Calculated Accuracy = 0.8200824
  + Accurately predicts the true cancellation result 82% of the time

Decision Tree:

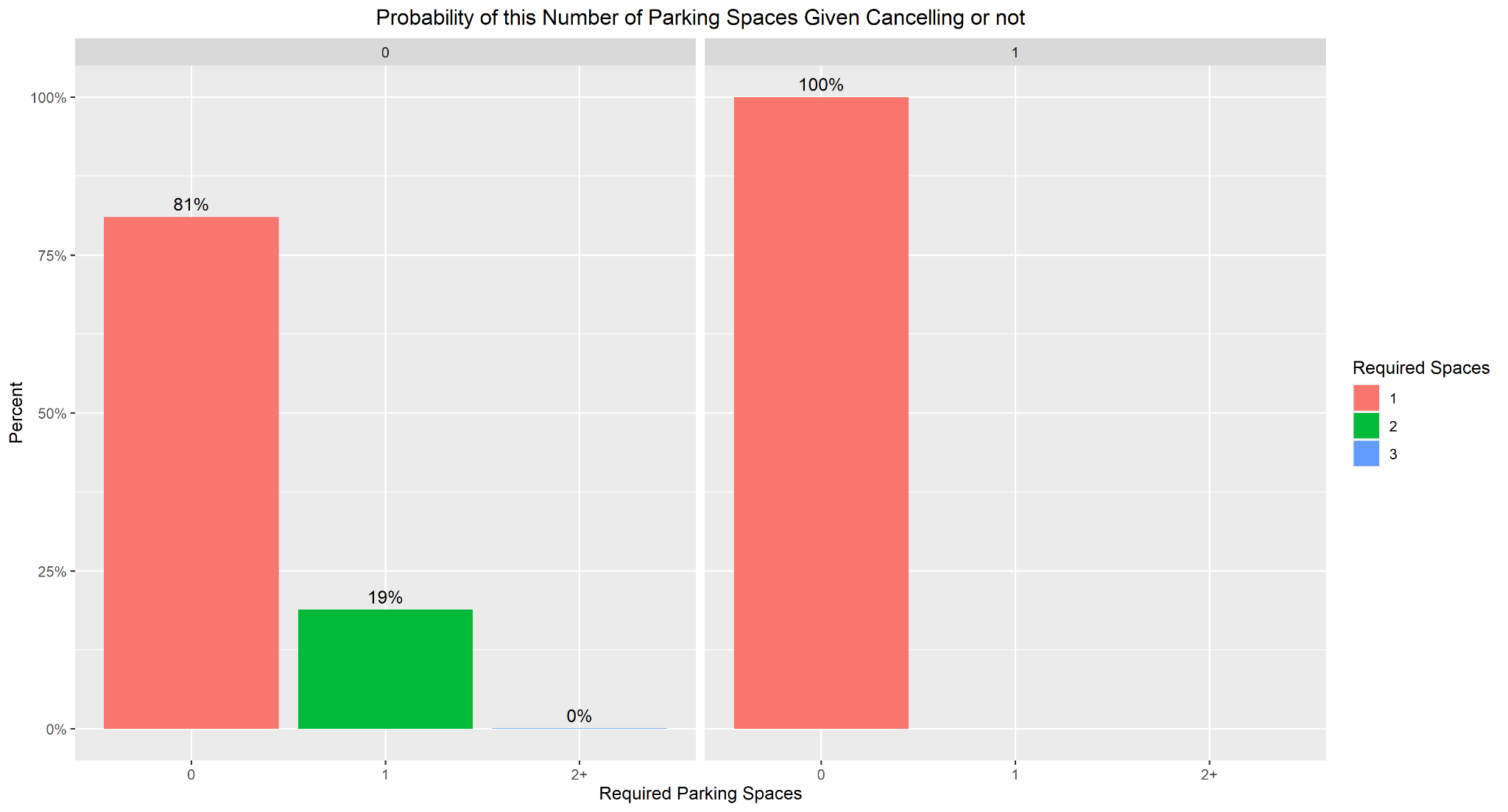
ProjectTree <- rpart(IsCanceled ~ ., data=ProjectTrainingSet)

prp(ProjectTree, faclen = 0, cex = 0.8, extra = 1)

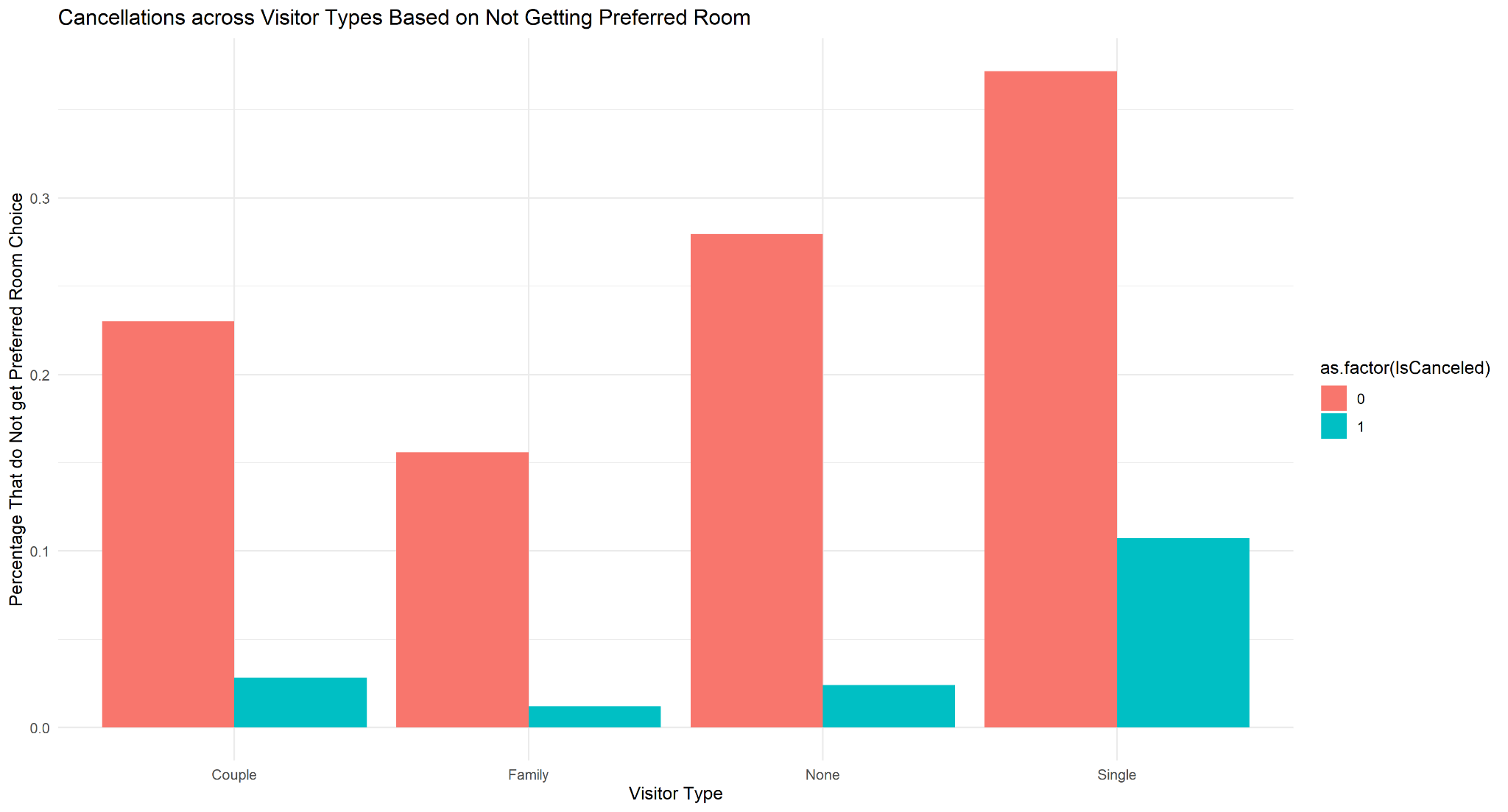


The first branch of the decision tree is the deposit type: if a non-refundable deposit is made, the customer cancels 95% of the time. From there, if the customer makes the reservation less than 17 days in advance, they cancel just 8.5% of the time. If not, it looks at whether they made the reservation through an online travel agent or not. If not, then if the customer requires multiple parking spaces, they literally never cancel the reservation. If they require one or fewer but make multiple special requests, they cancel 42% of the time as opposed to 69% if they make one or fewer special requests. If they did make the reservation through an online travel agent, they have an 88% likelihood of cancelling if they have previously cancelled a reservation versus just an 18% chance if not.

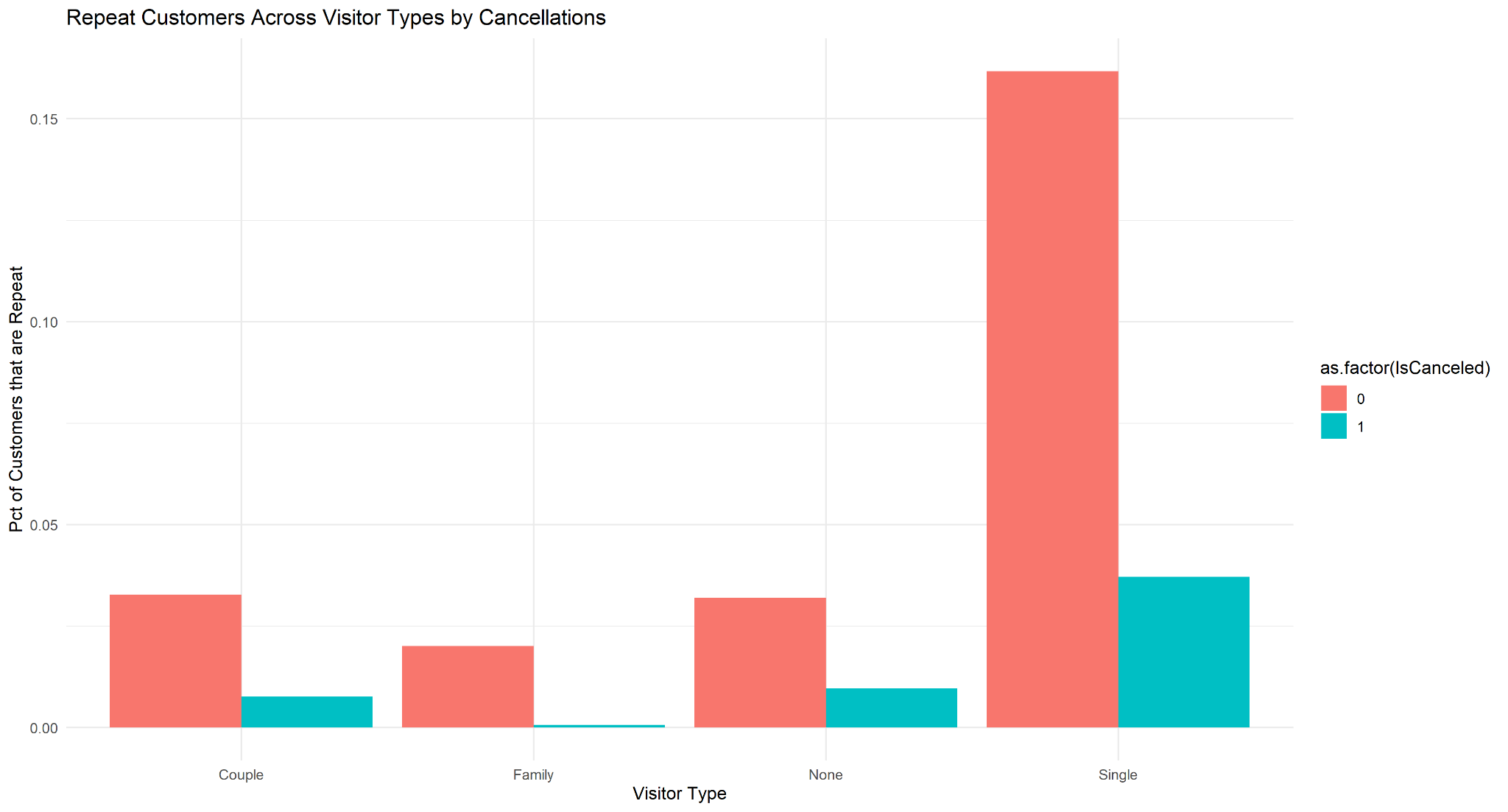
1. **Modeling Visuals**



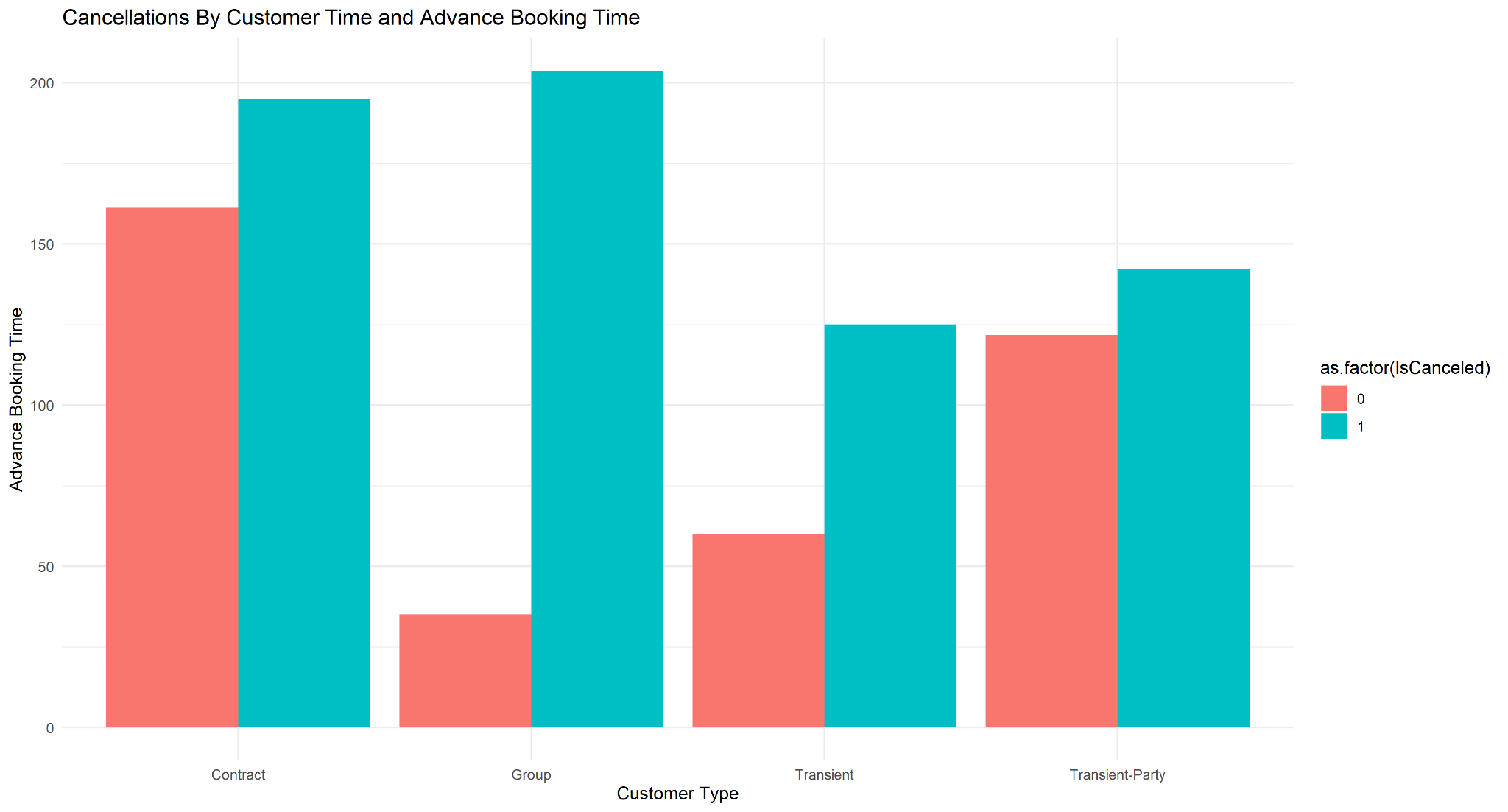
This visualization shows the percentage of each number of parking spaces given if the reservation was ultimately cancelled or not. The main takeaway from this is that every cancelled reservation did not require a parking space, showing that visitors that put in the effort to get their parking spot will follow through on their reservations.



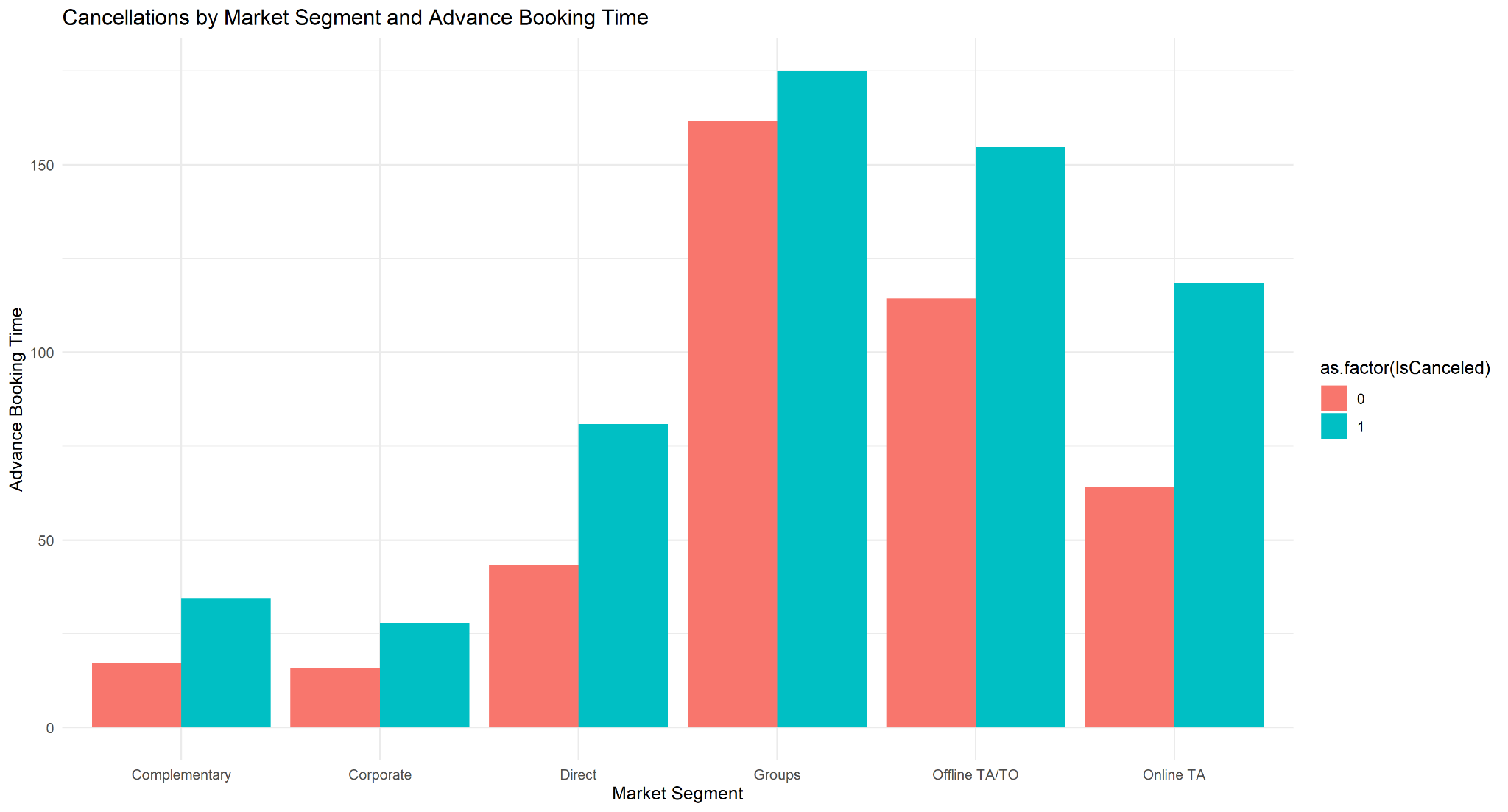
This visual shows the proportion of customers that didn’t receive their room of choice across the different visitor type groups, separated by canceled or not canceled. This visual seems counter-intuitive because it shows people that don’t get their preferred room to be less likely to cancel their reservations, which is odd because we would’ve expected the exact opposite.



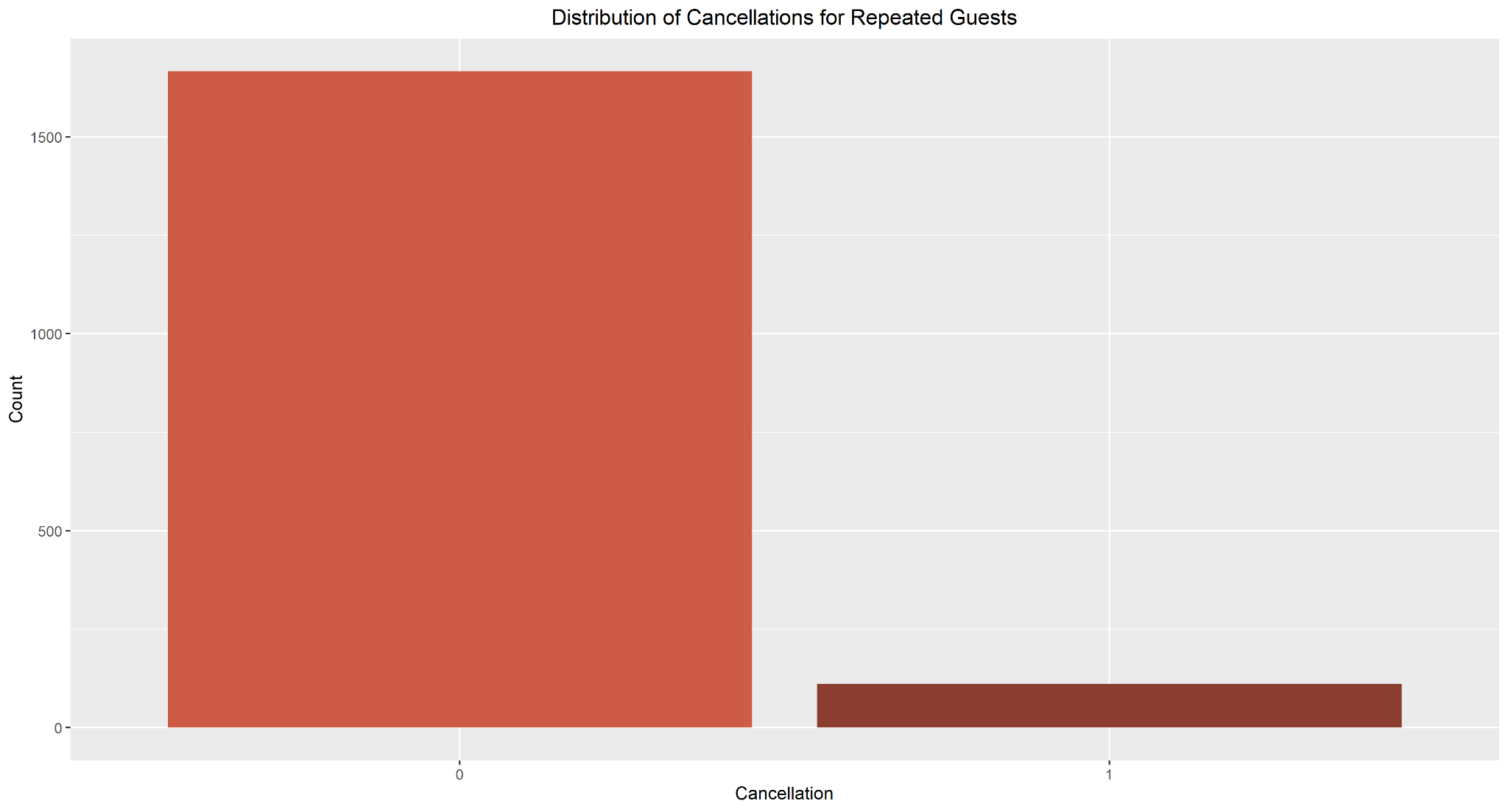
This shows a breakdown across visitor types of the proportion of total customers that are repeat customers, again split between canceled and uncanceled reservations. Single customers are significantly less likely to cancel their reservations if they are repeat customers, although all types of repeat customers are less likely to cancel to varying degrees as well.



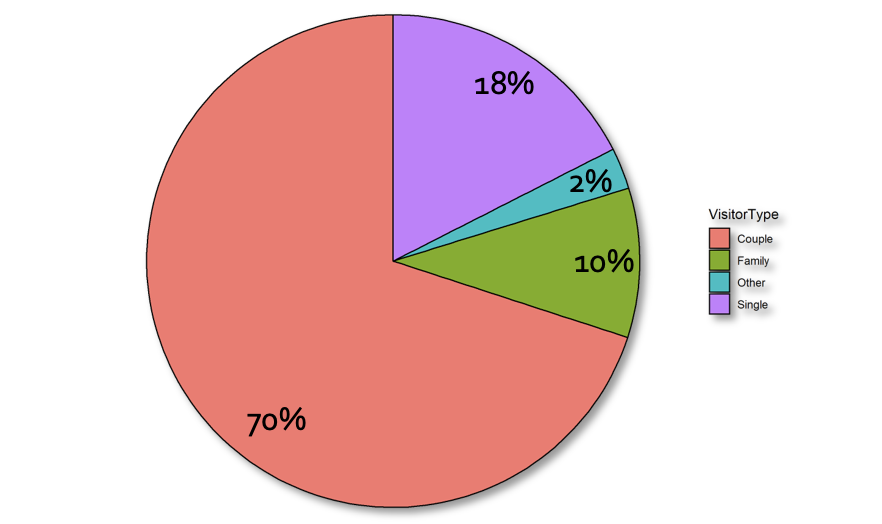
This visualization shows the average number of days the customers book in advance of their stay, broken down by both customer type and cancellation status. For each customer type, the customer cancels more for higher advance booking times, but the most valuable insight is for groups, where the average advance booking time for non-cancellations is roughly 30-35 days, whereas it is just above 200 days for cancelled reservations.



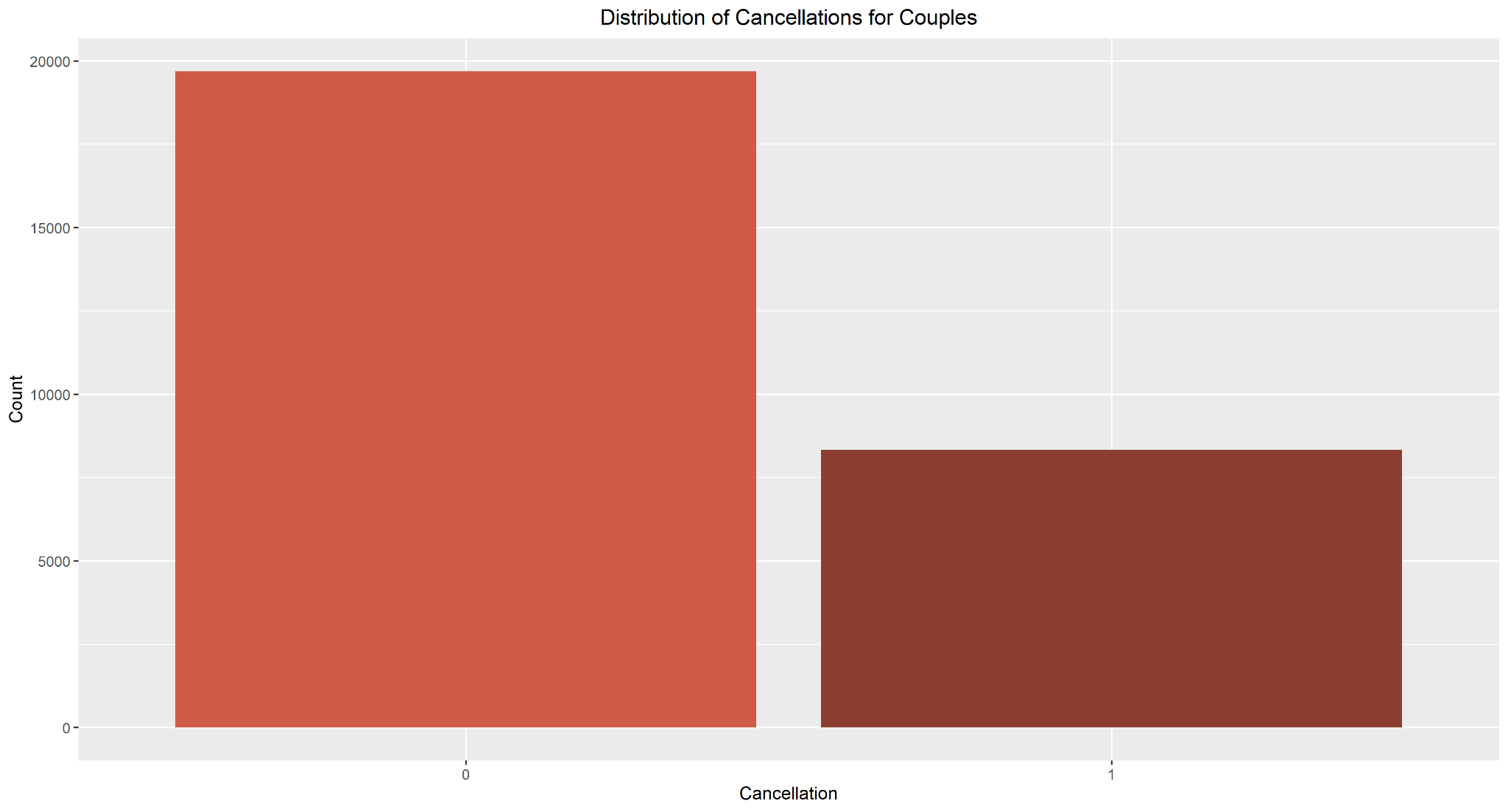
This visualization again shows the average number of days the customers book in advance of their stay, but this time broken down by both market segment and cancellation status. For each market segment, the customer still cancels more for higher advance booking times, but the most valuable insight is for groups, where the average advance booking time for non-cancellations is roughly 30-35 days, whereas it is just above 200 days for cancelled reservations.



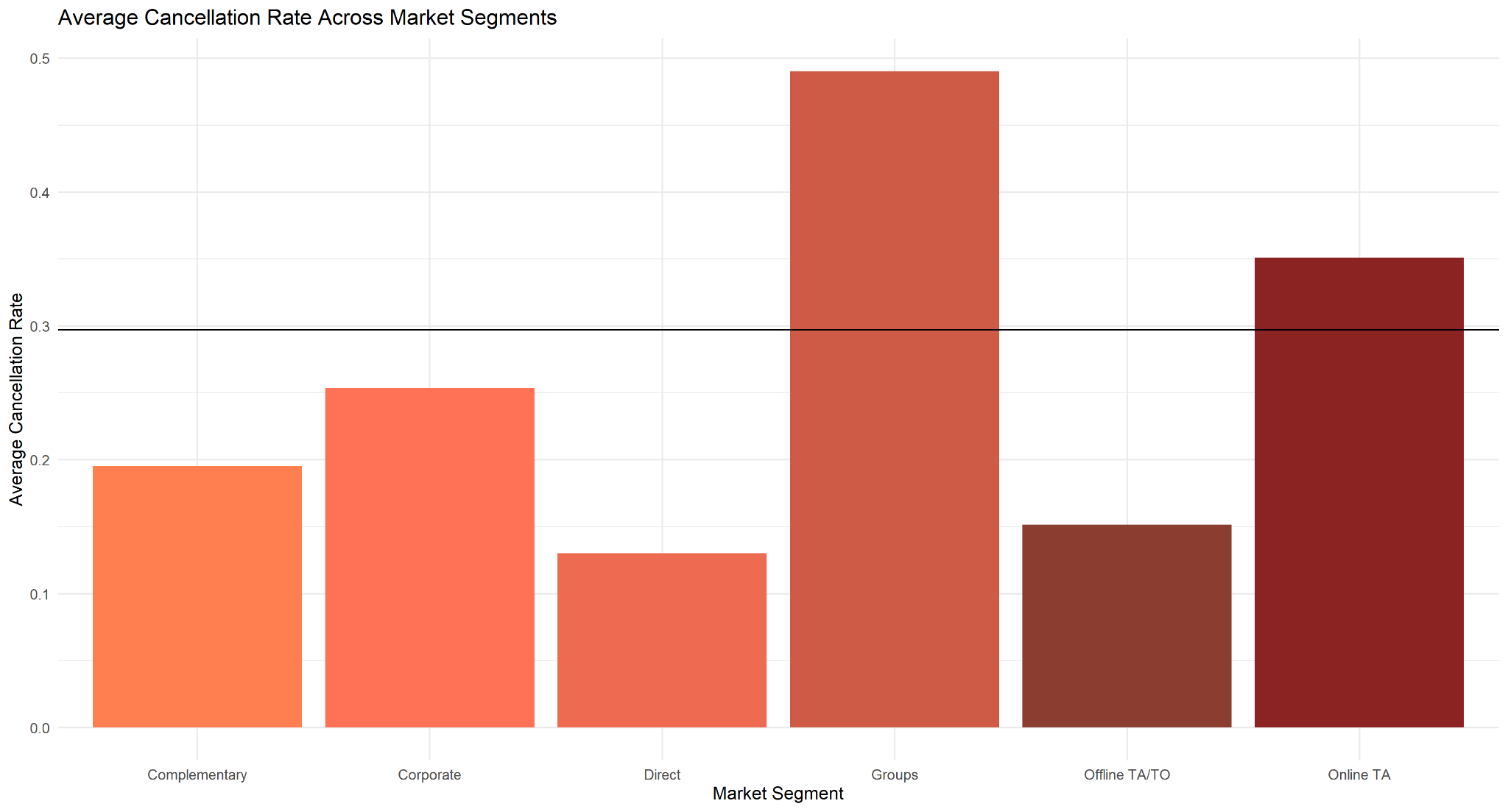
This graphic comes from a subsetted dataset that only includes guests that are returnees and looks at how many of those guests cancel and don’t cancel. Overwhelmingly, returning guests are highly unlikely to cancel their reservations.



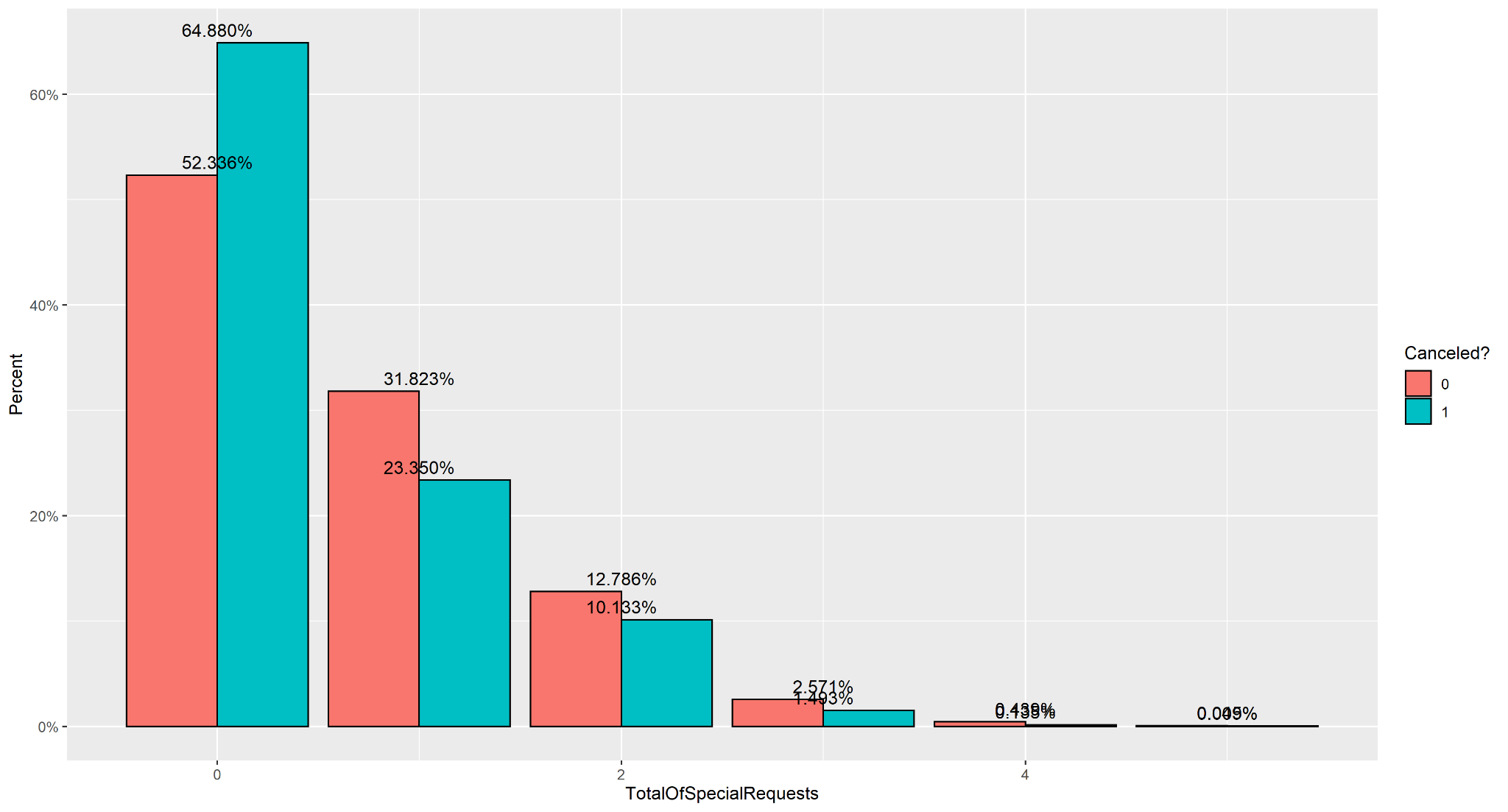
This is a pie chart investigating the breakdown of visitor types of guests than are returnees. Couples make up over 50%, while Single visitors make up the vast majority of the rest.



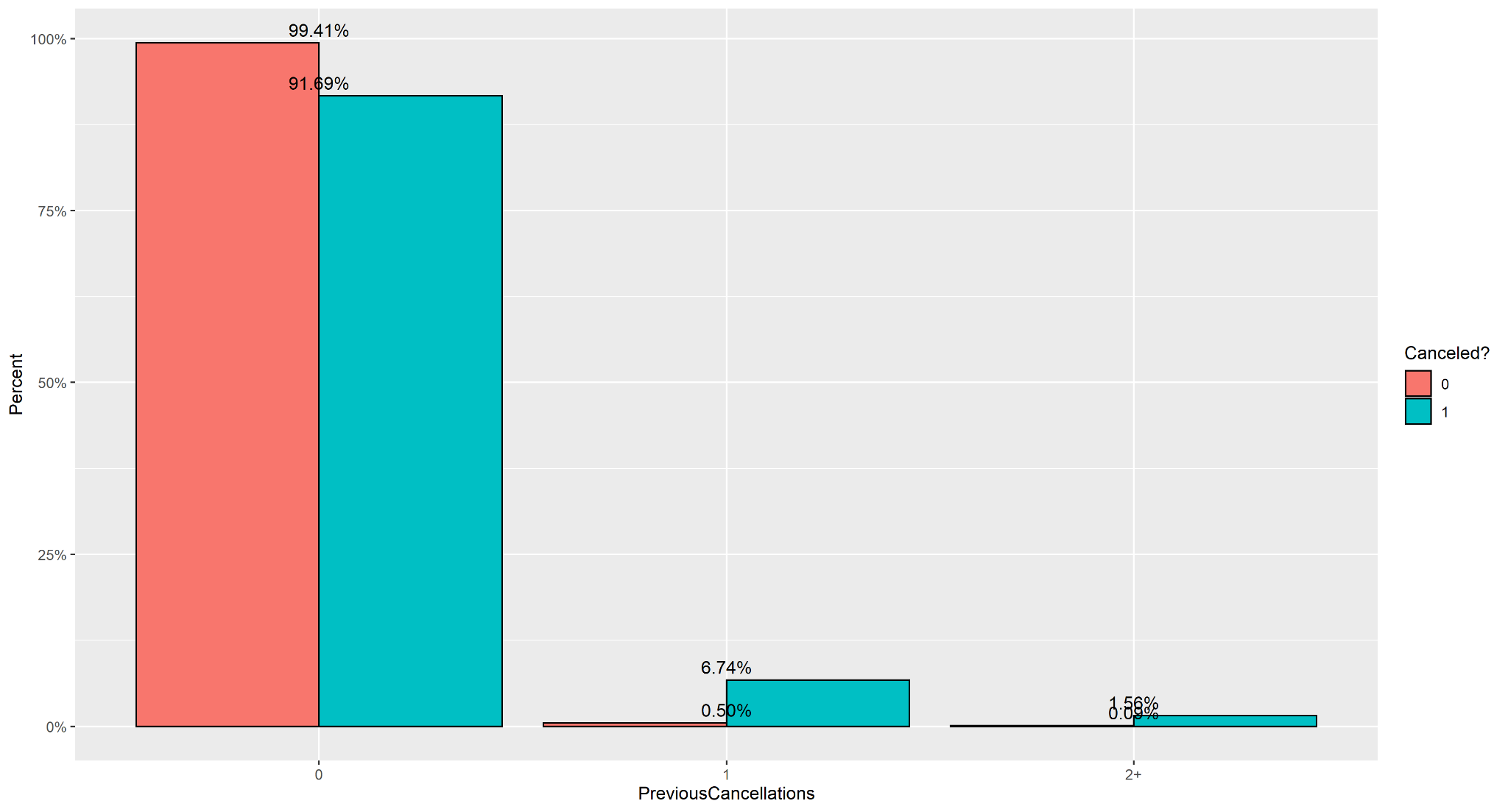
This shows the distribution of cancellations for the subset that only contains customers that are couples, which is exactly 2 adults and no children or babies. There are more non-cancelled reservations than canceled ones, which is very much in line with the overall trend.



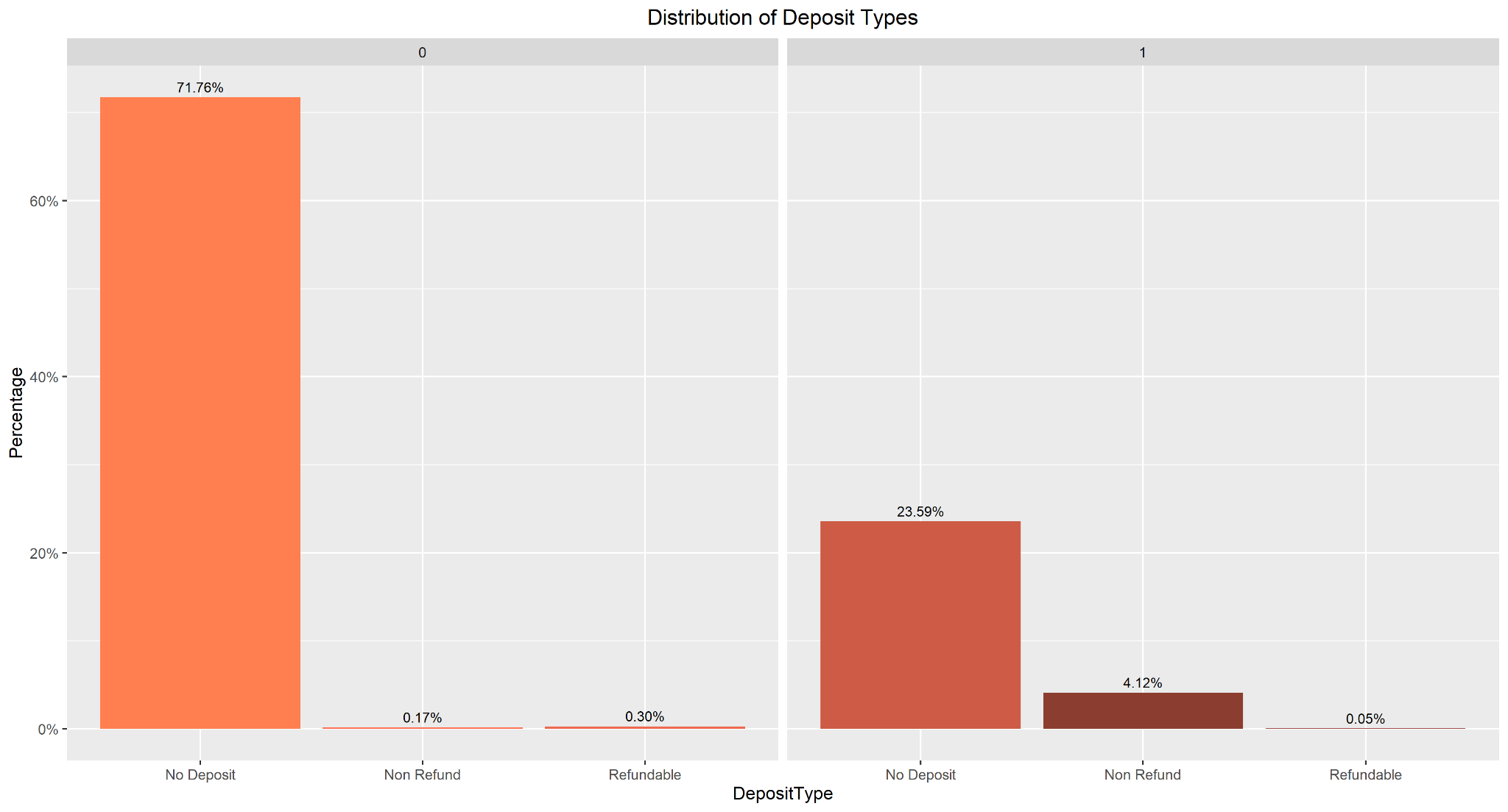
This graphic depicts the average cancellation rate for each separate segment of the market, compared to the overall average as a horizontal line. Groups have the highest average cancellation rate,with Online travel agents being the only other segment above the overall average. Direct and Offline travel agents have the lowest average cancellation rates.



As you can see, the more special requests a client makes of the hotel, the less likely he is to cancel his stay!



Customers with previous cancellations are extremely likely to cancel their next reservation.



Contrary to what we would expect, customers appear much more likely to cancel reservations with a non-refundable deposit than actually keep them.

1. **Actionable Insights and Conclusions**

* There has never been a cancellation by a customer that listed at least one required parking space
* Customers that don’t cancel their booking make an average of roughly 0.2 more booking changes
* The proportion of customers who canceled their bookings that made 0 special requests is 12.5% higher than for customers who didn’t cancel their bookings
* Couples make up 70% of the returning guests
* 17% of solo customers who didn’t cancel their bookings are returnees as opposed to 3.5% who do cancel
  + **Recommendations:**
    - Offer incentives to customers to return to the hotel for another stay, with a special focus on solo travelers
      * Cancellations are less likely for repeat customers no matter what, but for solo travelers, the difference is significantly larger
      * Returning guests are more likely to build loyalty towards the hotel and return yet again in the future
      * Creating and maintaining customer loyalty and returns must be a top priority
* Nearly 45% of customers book through an Online travel agent
* More lead time makes Customers more likely to cancel their bookings
* Families cancel 35% of the time, Couples 30% of the time, Customers who fall into “Other” 19% of the time, and Solo travelers 17% of the time
  + **Recommendations:**
    - Work on ways to make the hotel a better fit for families given that they are the group with the highest cancellation rates
      * Advertise family discounts and family-friendly amenities
      * Highlight nearby attractions that will appeal to families if there are any
      * Train staff on how to anticipate the needs of families, especially those with younger children who could be more of a challenge
      * Get feedback from families who stay at the hotel for what they liked and what needs to be improved
* Over 93% of the customers who have previously canceled a booking cancel their bookings this time
  + **Recommendations:**
    - Create a survey for guests who cancel their reservations
      * Listen to their responses and insights
      * Make adjustments if they have complaints or suggestions of things that are fixable
    - The goal is to make the guests feel that their feedback matters and is being listened to
    - We also want to improve the service of the hotel for before their arrival but after the reservation is made if that is needed
* For the Group market segment, there is a difference of roughly 180 days for the average advance booking time for canceled and non-cancelled reservations, with canceled ones being booked more in advance
* 96% of customers who place a non-refundable deposit then cancel their booking
* A longer stay duration makes cancellation more likely across every customer type and visitor type
* Takeaways from our best Model which is ASM:
  + Portugal is the country with the most cancellations, as well as the most bookings
  + The majority of cancellations are made by the group market segment
  + Customers that don’t have babies or children are more likely to cancel
  + Customers who don’t make changes to their booking are more likely to cancel it
  + Repeated guests are generally less likely to cancel
  + Guests who cancel don’t make special requests.
  + Not having a parking space can cause a customer to cancel their reservation
  + Not having stayed at the hotel before makes a customer more likely to cancel