Bar Plot Analysis

Initial exploratory analysis included creating bar plots of the variables related to the bookings cancellation status. Cancellations and non-cancellations were the dependent variables (y) against the independent variables (x). Further analysis was conducted by Data was then transforming the dependent variable (canceled or not cancelled) into a ratio of the number of cancellations against the total number of booking’s.

Results:

The results indicate that there are clear trends for several variables that increase the likelihood of a booking being cancelled.

The first is the lead time analysis. This showed that as lead time increased a party was more likely to cancel. This relationship is not obvious from the bar plot due to the large quantity of data (<40,000 data points). Most bookings were days in advance, while fewer were more than 6 months out. The lower number of data point with greater lead times could also mask actual relationships.

Chart

Description automatically generated

The analysis of number of babies shows that a booking without a baby is more likely to cancel that one with. However, this is not a strong coloration. Analysis as a ratio (number of cancellations by total number of bookings) makes the relationship more visible. A person traveling without a baby is approximately 28% likely to cancel verses 19% and 23% for a party booking with 1 and 2 babies respectively.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Analysis of deposit type shows that a large majority of booking parties did not put a deposit down. Still, roughly a quarter (25%) canceled their reservation. The most surprising revelation from deposit type as ratio analysis is the high percentage of non-refundable deposits parties that cancelled. The data shows that nearly 95% of parties that submitted a non-refundable deposit canceled. This is counterintuitive to what would be expected with anything non-refundable.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

When analyzing the number of parking spaces requested at booking, it is clear that customers who do not request a parking space are more likely to cancel. Parking space analysis as a ratio revealed a more interesting relationship. Of all the bookings that requests one or more parking space. None of the cancelled their reservation. Because there were no cancellations for parties that requested parking spaces, this insight was only visible after plotting the dependent variable as *is not cancelled*.

Chart

Description automatically generated Chart, shape, square

Description automatically generated

Chart, bar chart

Description automatically generated

Customers were grouped by type contract, group, transient, or non-transient party and plotted against the total number of cancellations and non-cancellations for each group. Analysis for number of cancellations shows that a larger for all customer types a customer is more likely to keep the reservation than cancel. Plotting customer type against cancellations as a ratio reveals that Transient parties are more likely to cancel than non-transient group or contract parties.

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Previous cancellations also revealed an interesting and useful trend in cancelations. Analysis shows that the vast majority of parties do not have previous cancelations. This makes analyzing raw data difficult. When plotted as ratio of cancellations to total bookings a pattern emerged. While customers with more than one cancelation were more likely to cancel, customers with greater than 14 previous cancellations, canceled their booking 100% of the time.

Chart

Description automatically generated Chart, bar chart

Description automatically generated

Code

#Subset Data based on IsCanceled

canceled<- subset(hotelData, IsCanceled == 1)

notCanceled<-subset(hotelData, IsCanceled == 0)

#Bar plot from raw data

#LeadTime

ggplot()+ geom\_bar(data=hotelData, aes(x=LeadTime, group=as.factor(IsCanceled), fill=as.factor(IsCanceled)), position="dodge")+ labs(x="Lead Time", y= "Count of Lead Time", title= "Lead Time Analysis")

#Babies

ggplot()+ geom\_bar(data=hotelData, aes(x=Babies, group=as.factor(IsCanceled), fill=as.factor(IsCanceled)), position="dodge")+ labs(x="# of Babies", y= "Count", title= "# of Babies Analysis")

#DepositType

ggplot()+ geom\_bar(data=hotelData, aes(x=as.factor(DepositType), group=as.factor(IsCanceled), fill=as.factor(IsCanceled)), position="dodge")+ labs(x="Deposit Type", y= "Count", title= "Deposit Type Analysis")

#CustomerType

ggplot()+ geom\_bar(data=hotelData, aes(x=CustomerType, group=as.factor(IsCanceled), fill=as.factor(IsCanceled)), position="dodge")+ labs(x="Customer Type", y= "Count", title= "Customer Type Analysis")

#RequiredCarParkingSpaces

ggplot()+ geom\_bar(data=hotelData, aes(x=RequiredCarParkingSpaces, group=as.factor(IsCanceled), fill=as.factor(IsCanceled)), position="dodge")+ labs(x="Required Car Parking Spaces", y= "Count of Lead Time", title= "Parking Analysis")

#Previous Cancellations

ggplot()+ geom\_bar(data=hotelData, aes(x=PreviousCancellations, group=as.factor(IsCanceled), fill=as.factor(IsCanceled)), position="dodge")+ labs(x="Previous Cancellations", y= "Count of Lead Time", title= "Previous Cancellation Analysis")

#As ratio analysis

#Subset Data based on IsCanceled

canceled<- subset(hotelData, IsCanceled == 1)

notCanceled<-subset (hotelData, IsCanceled == 0)

#Count number of different variables for Meals

a<-data.frame(table(canceled$Meal))

b<-data.frame(table(hotelData$Meal))

#Merge a & b on Var 1

c<- merge(a,b,by="Var1")

#Add new Col with cancel rate

c$Meal\_Cancel\_Rate<- c$Freq.x/c$Freq.y

c

#ggplot meal cancellation rate

ggplot(c, aes(y=Meal\_Cancel\_Rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Meal Type", y="Cancellation Rate by Meal", title="Meal type analysis")

#Deposit Analysis as Ratio

DepositType\_Canceled<- data.frame(table(canceled$DepositType))

DepositType\_Total<- data.frame(table(hotelData$DepositType))

DepositType\_asRatio<- merge(DepositType\_Canceled, DepositType\_Total, by="Var1")

DepositType\_asRatio$DepositType\_cancel\_rate<-DepositType\_asRatio$Freq.x/DepositType\_asRatio$Freq.y

ggplot(DepositType\_asRatio, aes(y=DepositType\_cancel\_rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Deposit Type", y="Cancellation Rate by Deposit Type", title="Deposit analysis")

#CustomerType as Ratio

d<-data.frame(table(canceled$CustomerType))

e<-data.frame(table(hotelData$CustomerType))

f<- merge(d,e,by="Var1")

f$CustomerType\_cancel\_rate<- f$Freq.x/f$Freq.y

ggplot(f, aes(y=CustomerType\_cancel\_rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Customer type", y="Cancellation Rate by Customer Type", title="Customer type analysis")

#Lead Time

g<-data.frame(table(canceled$LeadTime))

h<-data.frame(table(hotelData$LeadTime))

I<- merge(g,h,by="Var1")

I$LeadTime\_cancel\_rate<- I$Freq.x/I$Freq.y

ggplot(I, aes(y=LeadTime\_cancel\_rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Booking Lead Time", y="Cancellation Rate by Lead Time", title="Booking Lead Time Analysis")

#Country of Booking

j<-data.frame(table(canceled$Country))

k<-data.frame(table(hotelData$Country))

l<- merge(j,k,by="Var1")

l$Country\_cancel\_rate<- l$Freq.x/l$Freq.y

ggplot(l, aes(y=Country\_cancel\_rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Country", y="Cancellation Rate by Country", title="Booking Country Analysis")+ theme(axis.text.x=element\_text(angle=90))

#Previous Cancellations

m<-data.frame(table(canceled$PreviousCancellations))

n<-data.frame(table(hotelData$PreviousCancellations))

o<- merge(m,n,by="Var1")

o$PreviousCancellation\_cancel\_rate<- o$Freq.x/o$Freq.y

ggplot(o, aes(y=PreviousCancellation\_cancel\_rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Previous Cancellations", y="Cancellation Rate by Previous Cancellations", title="Previous Cancelation Analysis")

#Required Parking Spaces

p<-data.frame(table(canceled$RequiredCarParkingSpaces))

q<-data.frame(table(hotelData$RequiredCarParkingSpaces))

r<- merge(p,q,by="Var1")

r$RequiredCarParkingSpaces\_Cancellation\_rate<- r$Freq.x/r$Freq.y

ggplot(r, aes(y=RequiredCarParkingSpaces, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Parking Spaces Required", y="Cancellation Rate by Parking Spaces", title="Parking Space Analysis")

#Required Parking Spaces (Not Cancelled)

s<-data.frame(table(notCanceled$RequiredCarParkingSpaces))

t<- merge(s,q,by="Var1")

t$RequiredCarParkingSpaces\_nonCancellation\_rate<- t$Freq.x/t$Freq.y

ggplot(t, aes(y=RequiredCarParkingSpaces, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Parking Spaces Required", y="nonCancellation Rate by Parking Spaces", title="Parking Space Analysis (Not Canceled)")

#Babies

u<-data.frame(table(canceled$Babies))

v<-data.frame(table(hotelData$Babies))

w<- merge(u,v,by="Var1")

w$Babies\_cancellations\_rate<- w$Freq.x/w$Freq.y

ggplot(w, aes(y=Babies\_cancellations\_rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="# of Babies", y="Cancellation Rate by Babies", title="# Babies Analysis")

#LeadTime (does not work)

LeadTime.df<- canceled%>%

select(LeadTime)%>%

mutate(LeadTime\_bins= ifelse(LeadTime <=30, "30 days",

ifelse(LeadTime<=180, "6 mo",

ifelse (LeadTime<=365, "1 yr",

"long rng"))))

#canceled[,LeadTime=cut(LeadTime, breaks=c(0,30,180,365,Inf),labels=c("<30","30-180","180-365",">365"))]

levels(canceled$LeadTime)=c("<30","30-180","180-365",">365")

table(canceled$LeadTime)

g<-data.frame(table(canceled$LeadTime))

h<-data.frame(table(hotelData$LeadTime))

I<- merge(g,h,by="Var1")

I$LeadTime\_cancel\_rate<- I$Freq.x/I$Freq.y

ggplot(I, aes(y=LeadTime\_cancel\_rate, (x=Var1), fill=Var1))+ geom\_bar(stat="identity")+ labs(x="Booking Lead Time", y="Cancellation Rate by Lead Time", title="Booking Lead Time Analysis")