Assignment 7: Marketing Analytics

Preethi Abraham

26/04/2021

#By Preethi Susan Abraham  
  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data=read.csv("C:/Users/Preethi Abraham/Desktop/Brandeis Studies/Marketing Analytics/Assignment 7/hotelsat-data.csv")  
View(data)  
  
  
#Q1:Analyze the data to determine what aspects have the biggest impact on satisfaction  
  
  
#Creating a regression model with all 'satisfaction' parameters  
mod1=lm(satOverall~satCleanRoom +  
 satCleanBath +  
 satCleanCommon +   
 satFrontStaff +  
 satDiningStaff +  
 satHouseStaff +  
 satValetStaff +  
 satPerks +   
 satRoomPrice +  
 satDiningPrice +  
 satWifiPrice +  
 satParkingPrice +  
 satCity +  
 satCloseTransp +  
 satCloseEvents +  
 satPoints+  
 satRecognition, data=data)  
summary(mod1)

##   
## Call:  
## lm(formula = satOverall ~ satCleanRoom + satCleanBath + satCleanCommon +   
## satFrontStaff + satDiningStaff + satHouseStaff + satValetStaff +   
## satPerks + satRoomPrice + satDiningPrice + satWifiPrice +   
## satParkingPrice + satCity + satCloseTransp + satCloseEvents +   
## satPoints + satRecognition, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.92982 -0.62969 0.01264 0.61585 3.13811   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.018509 0.237568 -0.078 0.937910   
## satCleanRoom 0.097787 0.024528 3.987 7.03e-05 \*\*\*  
## satCleanBath 0.058437 0.029984 1.949 0.051496 .   
## satCleanCommon 0.030980 0.024791 1.250 0.211622   
## satFrontStaff 0.211916 0.035765 5.925 3.87e-09 \*\*\*  
## satDiningStaff 0.024947 0.028735 0.868 0.385445   
## satHouseStaff 0.058707 0.023772 2.470 0.013638 \*   
## satValetStaff 0.002993 0.026148 0.114 0.908888   
## satPerks 0.100832 0.030432 3.313 0.000944 \*\*\*  
## satRoomPrice -0.006211 0.023049 -0.269 0.787621   
## satDiningPrice 0.089935 0.026780 3.358 0.000804 \*\*\*  
## satWifiPrice 0.049088 0.025366 1.935 0.053154 .   
## satParkingPrice 0.037993 0.029249 1.299 0.194164   
## satCity 0.074920 0.028664 2.614 0.009046 \*\*   
## satCloseTransp 0.043335 0.022841 1.897 0.057989 .   
## satCloseEvents 0.014636 0.028389 0.516 0.606251   
## satPoints 0.065241 0.031336 2.082 0.037517 \*   
## satRecognition 0.033474 0.027352 1.224 0.221225   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9463 on 1482 degrees of freedom  
## Multiple R-squared: 0.3286, Adjusted R-squared: 0.3209   
## F-statistic: 42.66 on 17 and 1482 DF, p-value: < 2.2e-16

#satisfaction levels of cleanroom,frontstaff,perks,dining price are some of the biggest factors which contribute to the overall satisfaction level for customers. Apart from these factors, other satisfaction parameters such as clean bath, house staff,wifi price, city, close transportation and points also contribute to determining the overall satisfaction to a certain extent.  
  
#This model only contains the satisfaction parameters that are impacting the overall satisfaction level  
mod2=lm(satOverall~satCleanRoom +  
 satCleanBath +  
 satFrontStaff +  
 satHouseStaff +  
 satPerks +   
 satDiningPrice +  
 satWifiPrice +  
 satParkingPrice +  
 satCity +  
 satCloseTransp +  
 satPoints, data=data)  
summary(mod2)

##   
## Call:  
## lm(formula = satOverall ~ satCleanRoom + satCleanBath + satFrontStaff +   
## satHouseStaff + satPerks + satDiningPrice + satWifiPrice +   
## satParkingPrice + satCity + satCloseTransp + satPoints, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8813 -0.6322 0.0166 0.6152 3.1679   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.14862 0.18391 0.808 0.419145   
## satCleanRoom 0.10033 0.02437 4.116 4.06e-05 \*\*\*  
## satCleanBath 0.06555 0.02957 2.217 0.026774 \*   
## satFrontStaff 0.23149 0.03210 7.211 8.81e-13 \*\*\*  
## satHouseStaff 0.06130 0.02331 2.630 0.008629 \*\*   
## satPerks 0.11107 0.02944 3.772 0.000168 \*\*\*  
## satDiningPrice 0.08868 0.02661 3.332 0.000882 \*\*\*  
## satWifiPrice 0.04799 0.02432 1.973 0.048629 \*   
## satParkingPrice 0.03773 0.02611 1.445 0.148765   
## satCity 0.07973 0.02644 3.015 0.002613 \*\*   
## satCloseTransp 0.04339 0.02275 1.907 0.056686 .   
## satPoints 0.07598 0.02980 2.550 0.010871 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9458 on 1488 degrees of freedom  
## Multiple R-squared: 0.3266, Adjusted R-squared: 0.3216   
## F-statistic: 65.6 on 11 and 1488 DF, p-value: < 2.2e-16

#Satisfaction levels of clean room, front staff,perks and dining price impact the overall satisfaction the most. The other satisfaction levels contribute to the overall satisfaction to a certain extent  
  
  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

stargazer(mod1,mod2,type="text")#to compare both models

##   
## =======================================================================  
## Dependent variable:   
## ---------------------------------------------------  
## satOverall   
## (1) (2)   
## -----------------------------------------------------------------------  
## satCleanRoom 0.098\*\*\* 0.100\*\*\*   
## (0.025) (0.024)   
##   
## satCleanBath 0.058\* 0.066\*\*   
## (0.030) (0.030)   
##   
## satCleanCommon 0.031   
## (0.025)   
##   
## satFrontStaff 0.212\*\*\* 0.231\*\*\*   
## (0.036) (0.032)   
##   
## satDiningStaff 0.025   
## (0.029)   
##   
## satHouseStaff 0.059\*\* 0.061\*\*\*   
## (0.024) (0.023)   
##   
## satValetStaff 0.003   
## (0.026)   
##   
## satPerks 0.101\*\*\* 0.111\*\*\*   
## (0.030) (0.029)   
##   
## satRoomPrice -0.006   
## (0.023)   
##   
## satDiningPrice 0.090\*\*\* 0.089\*\*\*   
## (0.027) (0.027)   
##   
## satWifiPrice 0.049\* 0.048\*\*   
## (0.025) (0.024)   
##   
## satParkingPrice 0.038 0.038   
## (0.029) (0.026)   
##   
## satCity 0.075\*\*\* 0.080\*\*\*   
## (0.029) (0.026)   
##   
## satCloseTransp 0.043\* 0.043\*   
## (0.023) (0.023)   
##   
## satCloseEvents 0.015   
## (0.028)   
##   
## satPoints 0.065\*\* 0.076\*\*   
## (0.031) (0.030)   
##   
## satRecognition 0.033   
## (0.027)   
##   
## Constant -0.019 0.149   
## (0.238) (0.184)   
##   
## -----------------------------------------------------------------------  
## Observations 1,500 1,500   
## R2 0.329 0.327   
## Adjusted R2 0.321 0.322   
## Residual Std. Error 0.946 (df = 1482) 0.946 (df = 1488)   
## F Statistic 42.662\*\*\* (df = 17; 1482) 65.605\*\*\* (df = 11; 1488)  
## =======================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

#Q2 and Q3 have been done together below:  
  
  
#I created to a column called 'TotalMoneySpentOnStay to understand how much each customer is spending on average for everything during their entire stay(i.e room, food, wifi)  
data$TotalMoneySpentOnStay=(data$avgRoomSpendPerNight+data$avgFoodSpendPerNight+data$avgWifiSpendPerNight)\*data$nightsStayed  
  
mod3=lm(TotalMoneySpentOnStay~satCleanRoom +  
 satCleanBath +  
 satFrontStaff +  
 satHouseStaff +  
 satPerks +   
 satDiningPrice +  
 satWifiPrice +  
 satParkingPrice +  
 satCity +  
 satCloseTransp +  
 satPoints+distanceTraveled,data=data)  
summary(mod3)

##   
## Call:  
## lm(formula = TotalMoneySpentOnStay ~ satCleanRoom + satCleanBath +   
## satFrontStaff + satHouseStaff + satPerks + satDiningPrice +   
## satWifiPrice + satParkingPrice + satCity + satCloseTransp +   
## satPoints + distanceTraveled, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -627.3 -346.7 -115.8 250.2 9913.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 859.62219 104.76872 8.205 4.94e-16 \*\*\*  
## satCleanRoom -0.53762 13.78523 -0.039 0.9689   
## satCleanBath -9.23990 16.72853 -0.552 0.5808   
## satFrontStaff 19.85935 18.15703 1.094 0.2742   
## satHouseStaff -10.77816 13.18671 -0.817 0.4139   
## satPerks -21.44786 16.66371 -1.287 0.1983   
## satDiningPrice -11.93703 15.05080 -0.793 0.4278   
## satWifiPrice 6.07276 13.76100 0.441 0.6591   
## satParkingPrice 7.11816 14.77014 0.482 0.6299   
## satCity -33.21197 14.96303 -2.220 0.0266 \*   
## satCloseTransp 0.94137 12.87277 0.073 0.9417   
## satPoints 12.40096 16.86211 0.735 0.4622   
## distanceTraveled 0.00622 0.01586 0.392 0.6950   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 534.9 on 1487 degrees of freedom  
## Multiple R-squared: 0.00656, Adjusted R-squared: -0.001457   
## F-statistic: 0.8182 on 12 and 1487 DF, p-value: 0.6318

mod4=lm(avgRoomSpendPerNight~satOverall,data=data)  
summary(mod4)

##   
## Call:  
## lm(formula = avgRoomSpendPerNight ~ satOverall, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.841 -7.888 -0.253 7.112 50.414   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 160.1432 1.1345 141.152 < 2e-16 \*\*\*  
## satOverall -1.6511 0.2747 -6.011 2.31e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.21 on 1498 degrees of freedom  
## Multiple R-squared: 0.02355, Adjusted R-squared: 0.0229   
## F-statistic: 36.14 on 1 and 1498 DF, p-value: 2.308e-09

#It looks like overall satisfaction level has a negative impact on the avg room spend per night. This means that as the overall satisfaction of customers increase, the average room spend per night decreases, and this does not seem right.  
  
data$eliteStatus=as.factor(data$eliteStatus)  
data$visitPurpose=as.factor(data$visitPurpose)  
mod5=lm(avgRoomSpendPerNight~satOverall+distanceTraveled+eliteStatus+visitPurpose,data=data)  
summary(mod5)

##   
## Call:  
## lm(formula = avgRoomSpendPerNight ~ satOverall + distanceTraveled +   
## eliteStatus + visitPurpose, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -37.288 -8.064 -0.340 7.468 49.490   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.592e+02 1.578e+00 100.924 < 2e-16 \*\*\*  
## satOverall -1.350e+00 2.955e-01 -4.568 5.33e-06 \*\*\*  
## distanceTraveled 2.674e-04 3.616e-04 0.739 0.459731   
## eliteStatusNoStatus -3.798e-02 9.560e-01 -0.040 0.968318   
## eliteStatusPlatinum -4.515e+00 1.248e+00 -3.616 0.000309 \*\*\*  
## eliteStatusSilver -7.264e-01 9.025e-01 -0.805 0.421018   
## visitPurposeConcert 5.276e-02 1.130e+00 0.047 0.962757   
## visitPurposeConference 3.268e-01 1.087e+00 0.301 0.763653   
## visitPurposeOtherOrMixed 7.634e-01 1.178e+00 0.648 0.517053   
## visitPurposeSportsEvent 1.421e-01 1.019e+00 0.139 0.889180   
## visitPurposeVacation 1.728e-01 1.107e+00 0.156 0.875972   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.17 on 1489 degrees of freedom  
## Multiple R-squared: 0.03522, Adjusted R-squared: 0.02874   
## F-statistic: 5.436 on 10 and 1489 DF, p-value: 5.92e-08

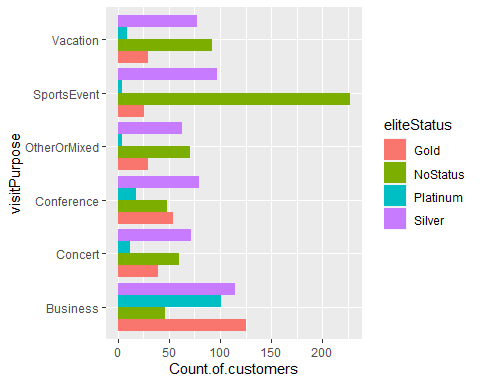
#In this model too, it can be seen that overall satisfcation will negatively affect the average room spend per night. This does not seem logical.  
  
data\_new <- data %>%  
 filter(eliteStatus=="NoStatus", visitPurpose=="Business")  
  
mod6 <- lm(TotalMoneySpentOnStay ~ satOverall, data=data\_new)  
summary(mod6)

##   
## Call:  
## lm(formula = TotalMoneySpentOnStay ~ satOverall, data = data\_new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -487.4 -319.5 -161.9 214.3 1218.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 96.32 334.12 0.288 0.7745   
## satOverall 150.80 81.03 1.861 0.0694 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 443.2 on 44 degrees of freedom  
## Multiple R-squared: 0.07297, Adjusted R-squared: 0.0519   
## F-statistic: 3.463 on 1 and 44 DF, p-value: 0.06944

#For customers whose elite status is 'No Status' and visit purpose is business related, overall satisfaction of those customers has some impact on the business's bottom line  
  
#Here, I have considered the 'bottom line' as the total money which our client makes from each customer(TotalMoneySpentOnStay)  
  
#Equation here is as follows:  
#TotalMoneySpentOnStay=96.32+150.80\*satOverall  
#Since 96.32 is not significant, our equation is   
  
#TotalMoneySpentOnStay= 150.80\*satOverall  
#If satOverall=7  
 #TotalMoneySpentOnStay=150.80\*7=1055.6 per customer from status of 'NoStatus' and business as their purpose of visit  
  
#There are 46 customers which belong to this category  
#So our client can expect a revenue of 1055.6\*46=$48557.6 from such customers alone  
  
customers\_based\_on\_status\_and\_visitpurpose=data%>%group\_by(visitPurpose,eliteStatus)%>%summarise(n())

## `summarise()` has grouped output by 'visitPurpose'. You can override using the `.groups` argument.

View(customers\_based\_on\_status\_and\_visitpurpose)  
names(customers\_based\_on\_status\_and\_visitpurpose)[names(customers\_based\_on\_status\_and\_visitpurpose) == "n()"] <- "Count.of.customers"  
  
ggplot(customers\_based\_on\_status\_and\_visitpurpose,aes(Count.of.customers,visitPurpose))+geom\_bar(aes(fill=eliteStatus),stat="identity", position=position\_dodge())



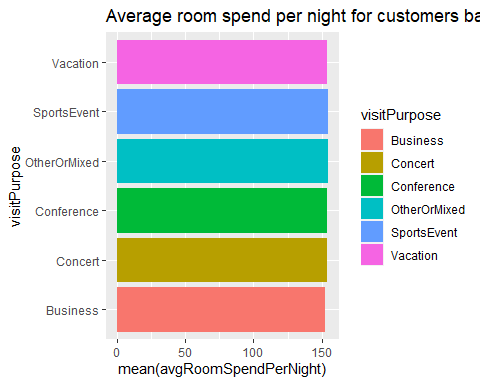
data\_new2 <- data %>%  
 filter(eliteStatus=="Gold", visitPurpose=="SportsEvent")  
  
mod7 <- lm( TotalMoneySpentOnStay~ satOverall, data=data\_new2)  
summary(mod7)

##   
## Call:  
## lm(formula = TotalMoneySpentOnStay ~ satOverall, data = data\_new2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -551.3 -377.1 -137.7 232.0 1451.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -146.0 485.3 -0.301 0.7662   
## satOverall 204.6 115.0 1.779 0.0879 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 547.3 on 24 degrees of freedom  
## Multiple R-squared: 0.1165, Adjusted R-squared: 0.07973   
## F-statistic: 3.166 on 1 and 24 DF, p-value: 0.08786

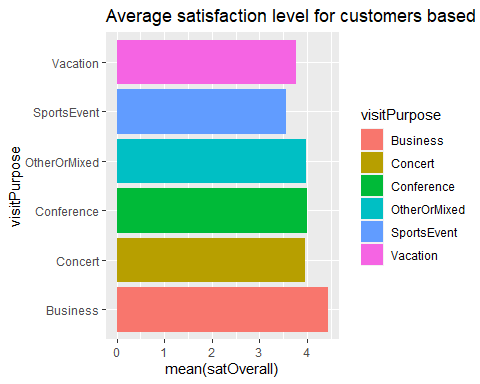
#It can be seen that overall satisfaction levels of customers whose elite status is 'Gold' and purpose of visit was due to a sports event seem to be having an impact on the total amount of money that they spend on their stay  
  
#TotalMoneySpentOnStay= 204.6\*satOverall  
##If satOverall=7  
 #TotalMoneySpentOnStay=204.6\*7= $1432.2 per customer from status of 'Gold' and 'Sports Event' as their purpose of visit  
  
#There are 26 customers from this category(Gold+SportsEvent), so our client can earn $1432.2\*26=37237.2  
  
  
data\_new3 <- data %>%  
 filter(eliteStatus=="NoStatus", visitPurpose=="Concert")  
  
mod8 <- lm( TotalMoneySpentOnStay~ satOverall, data=data\_new3)  
summary(mod8)

##   
## Call:  
## lm(formula = TotalMoneySpentOnStay ~ satOverall, data = data\_new3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -636.86 -333.64 -80.17 399.71 944.12   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 351.59 179.10 1.963 0.0544 .  
## satOverall 108.57 47.12 2.304 0.0248 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 427.1 on 58 degrees of freedom  
## Multiple R-squared: 0.08387, Adjusted R-squared: 0.06807   
## F-statistic: 5.31 on 1 and 58 DF, p-value: 0.02481

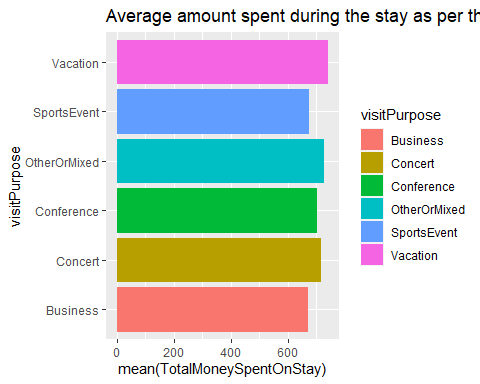
#It can be seen that overall satisfaction levels of customers who have 'NoStatus' and purpose of visit was due to a concert seem to be having an impact on the total amount of money that they spend on their stay  
  
#Equation:  
 #TotalMoneySpentOnStay=351.59+108.57\*satOverall  
#If satOverall=7  
 #TotalMoneySpentOnStay=351.59+108.57\*7=$1111.58 per customer  
  
#There are 60 customers from this category  
#Total money earned from customers belonging to 'No Status' and whose visit purpose is for a concert is 1111.58\*60=$66694.8  
  
  
x <- data %>%  
 group\_by(visitPurpose) %>%  
 summarise(mean(avgRoomSpendPerNight),mean(satOverall),mean(nightsStayed),mean(TotalMoneySpentOnStay))  
  
  
#Q4: Make 3 additional insights from the data that the client could use to develop their product or their marketing  
  
  
#Aditional Insights  
  
library(ggplot2)  
  
ggplot(x,aes(visitPurpose,`mean(avgRoomSpendPerNight)`,fill=visitPurpose))+geom\_bar(stat="identity", position=position\_dodge())+labs(title="Average room spend per night for customers based on their purpose of visit ")+coord\_flip()



#All customers seem to be spending similar amount of money for their room  
  
ggplot(x,aes(visitPurpose,`mean(satOverall)`,fill=visitPurpose))+geom\_bar(stat="identity", position=position\_dodge())+labs(title="Average satisfaction level for customers based on their purpose of visit ")+coord\_flip()



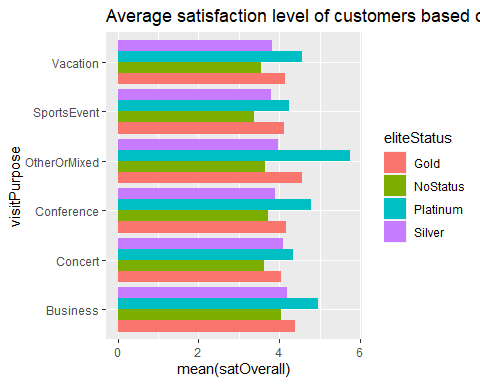
#Customers who visit our client's hotel for business purposes are the most satified whereas those who come for Sports Events are the least satisfied. From the previous graph on 'Average room spend per night for customers based on their purpose of visit', it can be seen that customers who stay at the hotel due to sports events spend the most on their room compared to other customers.   
  
#Our client should work on ensuring that customers are satisfied with their stay. Since the average satisfaction levels are in the range of 3 to 4, more work needs to be done to ensure that customers are satisfied with the stay at the hotel  
  
ggplot(x,aes(visitPurpose,`mean(TotalMoneySpentOnStay)`,fill=visitPurpose))+geom\_bar(stat="identity", position=position\_dodge())+coord\_flip()+labs(title="Average amount spent during the stay as per the customer's purpose of visit")



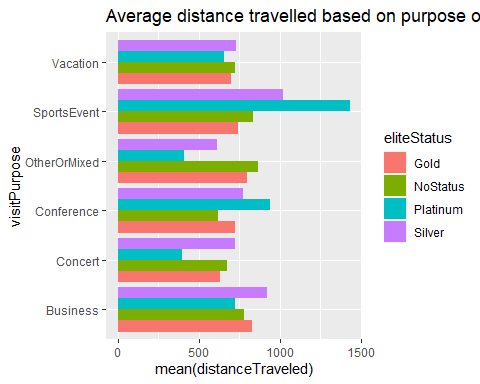
y=data %>%  
 group\_by(visitPurpose,eliteStatus) %>%  
 summarise(mean(distanceTraveled),mean(satOverall),mean(TotalMoneySpentOnStay))

## `summarise()` has grouped output by 'visitPurpose'. You can override using the `.groups` argument.

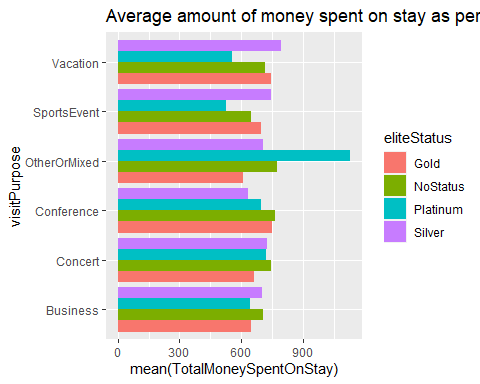
View(y)  
  
  
ggplot(y,aes(visitPurpose,`mean(satOverall)`))+geom\_bar(aes(fill=eliteStatus),stat="identity", position=position\_dodge())+coord\_flip()+labs(title="Average satisfaction level of customers based on purpose of visit and elite status")



#Customers from elite status of 'Platinum' and whose purpose of visit is 'Other or Mixed' seem to be the most satisfied. Our client should target such an audience.'Platinum' customers seem to be the most satisfied in every group. Customers who stay at the hotel for 'Sports Event' purpose seem to be the least satisfied. This shows that such customers('Sports Event as purpose of visit) are not the target audience for our client.  
  
  
ggplot(y,aes(visitPurpose,`mean(distanceTraveled)`))+geom\_bar(aes(fill=eliteStatus),stat="identity", position=position\_dodge())+coord\_flip()+labs(title="Average distance travelled based on purpose of visit and elite status")



#Customers whose elite status is platinum and purpose of visit is 'Sports Event' seem to be travelling the most to stay at the client's hotel  
  
ggplot(y,aes(visitPurpose,`mean(TotalMoneySpentOnStay)`))+geom\_bar(aes(fill=eliteStatus),stat="identity", position=position\_dodge())+coord\_flip()+labs(title="Average amount of money spent on stay as per visit pupose and elite status")



#Customers who visit the hotel for 'Other or Mixed' purposes from platinum status spend the most during their stay  
  
#Overall:  
 #1. Platinum customers are the most satisfied in each category of purpose of visit(Business, Sports Event etc)  
 #2. Client's target audience seems to be the customers who will visit the hotel for 'Other or Mixed' purposes and for 'Concerts'. These customers seem to be coming from nearby places since the average distance travelled by them seems to be the least. They are satisfied and also spend the most during their stay. Besides this, our client should also focus on 'vacationers' since such customers also spend the most during their stay   
#3.Our client should not focus too much on acquiring customers whose purpose of visit is due to a sports event. These customers are not from the local area(avg distance travelled is high), and they also spend the least amount of money during their stay at the hotel