**INDEX:**

[**Data Cleaning**](#_lhzdlb2ys41s) **2**

[**Data Selection**](#_l21yk9rgdizy) **2**

[**Data Modeling**](#_nger5zdctn1o) **3**

[**Business Questions**](#_d1deuxvlkcnv) **4**

[**Q1. What is the purpose of visit for the people coming to Hyatt Hotels ?**](#_45kvnetm7np) **4**

[**Q2. What is the Net Promoter Score for different states? Which State is doing the best and which state is doing the worst in terms of Net Promoter Score?**](#_38rjiby24p6g) **5**

[Linear Modelling on the Overall Data](#_134q3wig4ltq) 7

[Association rules mining](#_guxdnbnuuy64) 8

[Support Vector Machines(SVM)](#_ow2dkxsa5yf1) 10

[Naive Bayes](#_ve9qmzgcr4rl) 11

[**Q3. Does the purpose of travel for different customers reflect on the Net Promoter Score ?**](#_flh5fw3irke9) **12**

[**Business Customers:**](#_ub4p0c1cq4r) **12**

[**Q4. Is there a difference in criteria for recommending Hyatt hotels between East Coast and West Coast guests?**](#_keu6bbzhy8aa) **13**

[Linear Modelling](#_z2vu5nxf95ox) 14

[Arules analysis](#_tj3y9mqat56l) 16

[Support Vector Machines(SVM)](#_g2n28ze8wz7z) 17

[Naive Bayes](#_jlv3hjrjv321) 18

[**Leisure Customers:**](#_yxm050g9tdy1) **19**

[**Q4. Is there a difference in criteria for recommending Hyatt hotels between East Coast and West Coast guests?**](#_stvnr2u055dh) **21**

[Linear Modelling](#_abod2to53koc) 21

[Arules analysis](#_xgbarwvnja30) 23

[Support Vector Machines(SVM)](#_6dak8hb7ywys) 24

[Naive Bayes](#_k5bdvzb66lit) 25

[**Comparing the results from our analyses on all customers, business customers and leisure customers**](#_vyocnh24oafa) **26**

[Geographical visualization of the results](#_u7v2i2wnbahy) 27

# **Data Cleaning**

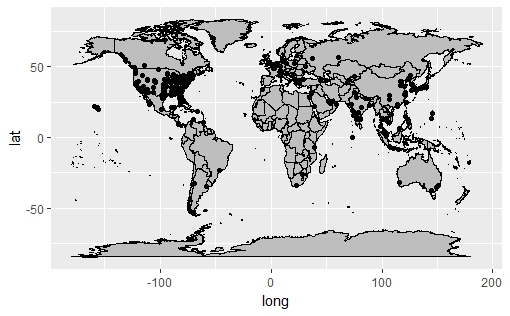
When it comes to doing any kind of analysis, one of the first steps to do is to collect the information necessary for the analysis and then store it in a manner that makes it easily accessible for use. For our analysis of Hyatt’s Net promoter Score, we have a whole year’s data from Hyatt. This dataset that we have is very big in size and contains a lot of information that we do not need for our analysis. So in order to increase speed and efficiency, we decided to remove some information that is irrelevant to our analysis. Moreover, the dataset that we have contains duplicate information as it appears to be a combination of information from different tables that could have overlapping information. Hence, we also removed this duplicate information such as the guest’s check-in-date which is mentioned in more than 2 different columns. Doing so leaves us with a dataset that only has information relevant to our analysis.

Another problem that we found when attempting our analysis was the abundance of missing or NA values that would make it difficult for us to understand how different factors affect the Net Promoter Score for Hyatt. We decided to remove factors(columns) that had over 80% of the data missing.

Our analysis will take ‘likelihood to recommend’ and ‘nps type’ as our dependent variables so it is necessary that we remove all instances with NA values in either, or both columns. The resulting subset left has relevant data for our analysis and is free of NA values based on the dependent variable which allows us to move forward with some modelling techniques.

# **Data Selection**

A close look at the data reveals that a significantly large proportion of Hyatt hotels are in the United States, hence this analysis will focus on Hyatt’s US locations.



We also wanted to examine whether Hyatt’s performance during regular season was different from Hyatt’s performance during the holiday season. For this purpose, the months of April and May were selected as regular season months. For holiday season, the months of December and January were selected (Christmas and New Years are two major holidays).

This analysis reveals very similar results, so we compared the data for these months. On a comparison across purpose of visit, it is revealed that more people visit Hyatt locations for leisure in holiday season while more people choose to stay at Hyatt for business purposes during the regular season. To overcome this hurdle, we concatenated some data from both seasons so the comparison would be more reasonable and across similar datasets.

It is very likely that a guest travelling for business purposes would value and prefer different amenities than a guest staying for leisure purposes. Hence, conducting separate analysis for guests based on purpose of visit and then comparing them is very useful.

This project report will answer some major business questions and provide valuable and actionable insight into Hyatt’s operations.

# **Data Modeling**

1. Linear Modelling
2. Association rules
3. SVM (Support Vector machines)
4. Naive Bayes

# **Business Questions**

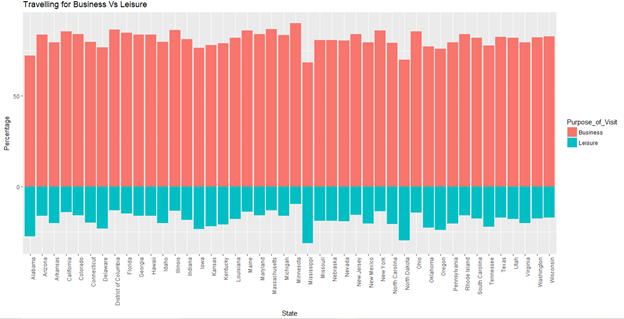
Q1. What is the purpose of visit for the people coming to Hyatt Hotels ?

Q2. What is the Net Promoter Score for different states? Which State is doing the best and which state is doing the worst in terms of Net Promoter Score?

Q3. Does the purpose of travel for different customers reflect on the Net Promoter Score ?

Q4. Is there a difference in criteria for recommending Hyatt hotels between East Coast and West Coast guests?

# Q1. What is the **purpose of visit** for the people coming to Hyatt Hotels ?



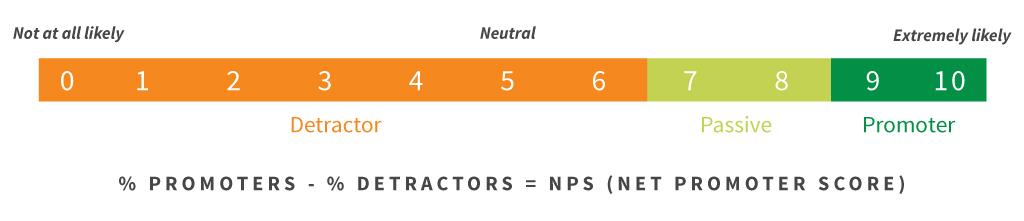
We realized from the above graph that the number of guests coming to Hyatt for business purposes are far greater than the number of guests coming to Hyatt for leisure. Its evident that Minnesota is state with the highest proportion of guests coming for business. On the other hand, Mississippi is clearly the state with the highest proportion of guests for leisure purpose(Despite that majority of guests in the state visit for business).

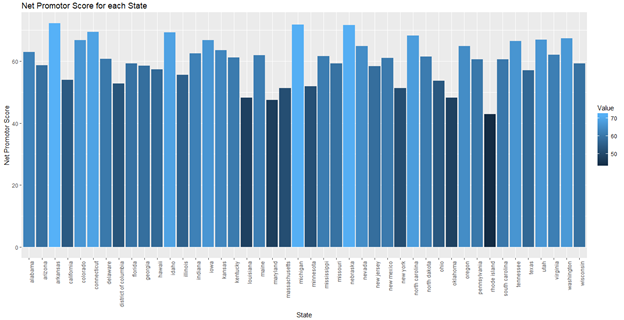
# Q2. What is the **Net Promoter Score** for different states? Which State is doing the best and which state is doing the worst in terms of Net Promoter Score?

Net Promoter Score is an index that can range from -100 to 100. It measures a customer's willingness to recommend a company or entity’s products to other people. A score of -100 means that all customers are detractors and would not recommend the company’s products whereas a score of 100 represents all customers being promoters(will definitely recommend to others).

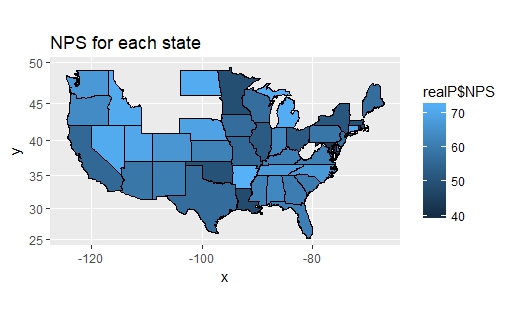
To calculate NPS, the following formula is used:

NPS = Percentage of promoters - Percentage of detractors

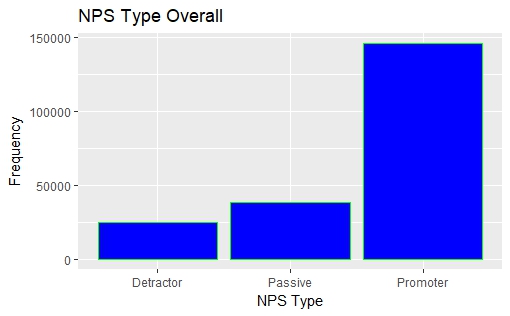




Doing so for all states reveals that Arkansas is the state with the highest NPS. The state with the lowest NPS was found out to be Maryland.



A summary of NPS Type for Hyatt



Our group also tried to map out the guest according to the NPS Types. As we can see that the majority of the customers rate the hotel as a promoter, but we will try understand which factor or purpose (people travelling for business or leisure) influence it more.

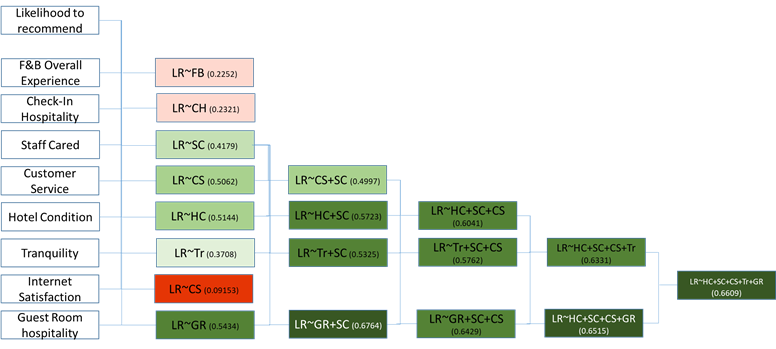
## Linear Modelling on the Overall Data

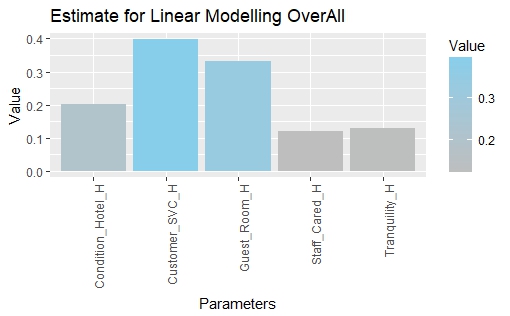
To understand how different factors affect the likelihood to recommend for a consumer, we used linear modelling.

Linear modelling is the process of describing a continuous variable as a factor of other variables. In other words, it is an attempt to explain the changes in a continuous variable as a result of changes in other (predictor) variables. R-square is a good determinant of how good a linear model is. The higher the r-square, the more accurately the change in a factor affects the dependent variable. In more mathematical, r-squared can be explained as how accurately can we predict the variability in Y with the change in value of parameter on X axis.

Linear modelling works best on numerical data, hence only the following variables were used for this part of the analysis.

The results of linear modelling with likelihood to recommend as the dependent variable are shown below.





We start with a single regression for all the above variables and try to build a more complex model based on our single regression results. The factors that have a very low ‘r-square’ (ie- they do not significantly affect the dependent variable) are dropped from our complex model in order to increase accuracy.

By doing so, we are left with the model where our dependent variable Likelihood to Recommend(LR) can be explained as a factor of Hotel Condition((HC) , Staff Cared(SC), Customer Service(CS), Tranquility(Tr) and Guest Room hospitality(GR).

This model, **LR~HC+SC+SC+Tr+GR** yields an r-square of 0.6609.

This means that about almost 66% of the variability in likelihood to recommend can be explained as a result of these variables.

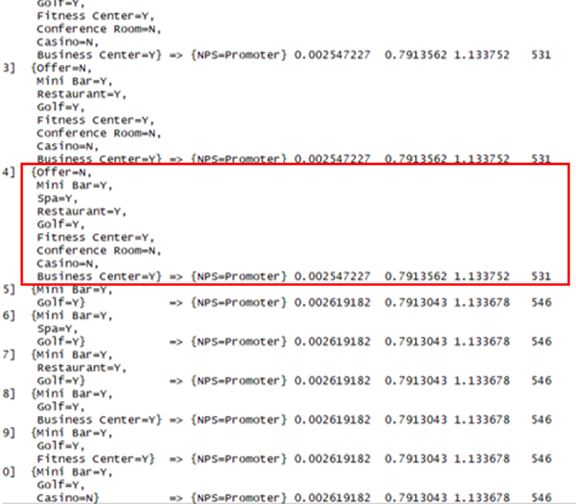
## Association rules mining

**Arules analysis**

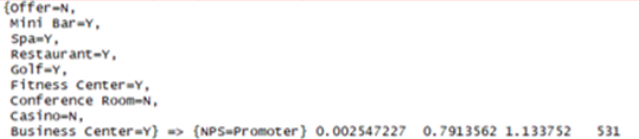
As mentioned above, linear modelling is useful for numerical data. To understand the effect of categorical variables on Hyatt’s NPS score, we need to use Arules(Association rules).

Association rules mining is a technique to understand the relationship and effect of a categorical factor on a dependent variable. Association rules or Arules are created by analyzing the data for frequent if/then patterns using a specified criteria support and confidence to identify the most important relationships. It is a very useful technique in analyzing and understanding customer behavior.

Our Arules analysis of the entire dataset revealed some interesting rules:



After sorting based on Confidence, the following association rule was found to be one of the best.



This means that the above combination, where Offer = N, Mini Bar = Y, Spa = Y, restaurant = Y, Golf = Y, Fitness center = Y, Conference Room = N, Casino = N, Business center = Y, is the one that results in most customers being Promoters. A 79% confidence also shows that this happens about 79% of the time these factors appear in this combination. Hence we can make the deduction that making an offer for the reservation or adding a conference room or casino in the hotel do not lead to an improved NPS score for Hyatt. Therefore, a conference room and casino might be unnecessary expenses for Hyatt if the goal is to just increase NPS.

On the other hand, this model offers strong evidence that the presence of a mini bar, spa, restaurant, a golfing facility and a fitness center can significantly increase a customer's likelihood to recommend the Hyatt hotel in question to other people.

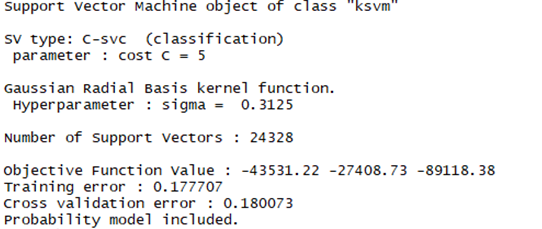
## Support Vector Machines(SVM)

Support Vector Machines(SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression purposes.

An SVM analysis is very useful to test how well a dependent variable can be explained as a factor of some ‘predictor’ variables.

In this report, an SVM analysis is conducted to see how well the NPS Type can be predicted based on our significant explanatory variables.

Results of SVM analysis on the data

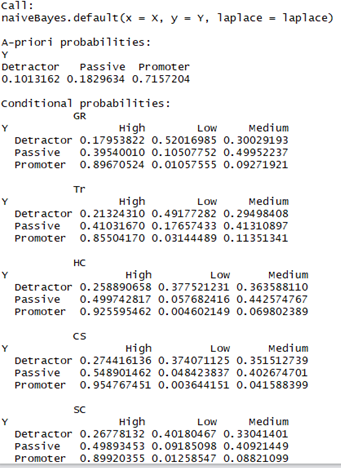


These results show that the model’s prediction is incorrect about 18% of the time. This means that these variables can successfully be used to predict the NPS Type about 82% of the time, which is a significantly high number.

This also means that if Hyatt made improvements in Hotel Conditions(HC), the level that the Staff Cares(SC) for the guests and increased the quality of Customer Service(CS), Tranquility(Tr) and Guest Room hospitality(GR), it’s overall people with NPS Type as promoter should increase. This would result in a higher Net Promoter Score.

## Naive Bayes

We also worked a little with naive bayes to get an overall understanding of the parameters that were had good r-squared value.



Another extremely useful comparison that is included in this report is the comparison of NPS across two different states for both purposes of visit-business or leisure.

For this purpose, New York and California are the two states selected. These states were selected because they offer a comparison across two sides of the country namely East Coast vs West Coast. Do different factors impact the likelihood for a customer to be a promoter or a detractor in these different areas? The data clearly shows that overall more guests visit California than New York and that California still has better NPS. So why does this phenomenon occur? Why do Hyatt’s customers like hotels in California more than they like the hotels in New York? What is Hyatt doing differently in its California hotels that is earning it better ratings in California?

# Q3. Does the purpose of travel for different **customers reflect on the Net Promoter Score** ?

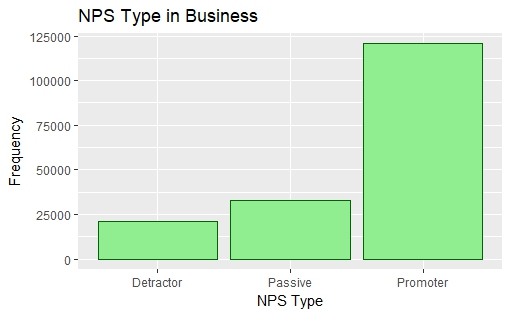
As mentioned previously in the report, conducting an analysis based on purpose of visit could be very useful. Hence, in this section we will use our various techniques to analyze Net Promoter Score as a function of other variables for both business and leisure customers.

This will show us if there is a difference in factors that affect a customer’s likelihood to recommend between business and leisure customers.

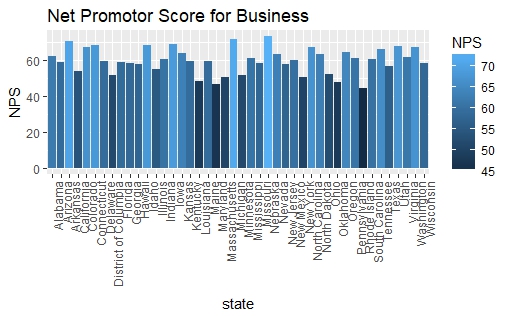
# **Business Customers**:

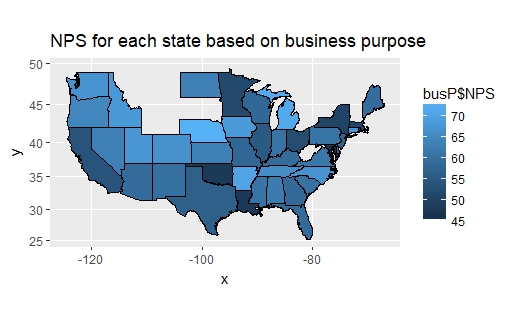
First we will conduct our analysis for the business customers in an attempt to understand their behavior.

We tries running the same tests on each business and Leisure, starting with calculating the numbers of guests under each NPS Type rating bracket (Promoter, Passive, Detractor)



Hyatt’s NPS Score for Business guests across different states



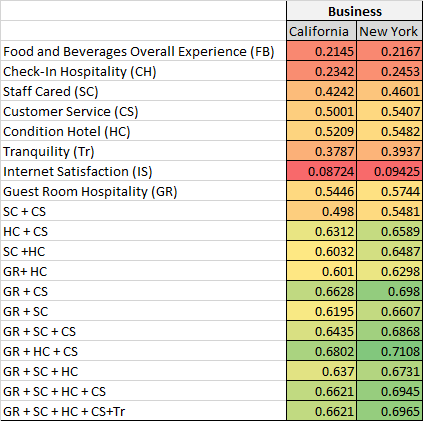


# Q4. Is there a difference in **criteria for recommending Hyatt hotels** between East Coast and West Coast guests?

This section will also answer this question by conducting the above analysis separately for New York(East Coast) and California(West Coast). The various modelling techniques are used here on data for both states so the models can be compared.

## Linear Modelling

The results of running various linear regressions with Likelihood to Recommend being modelled as a function of other variables for business guests in California vs business guests in New York are shown below. The column on the left lists the different linear regression models that were created and the variables used in them. The numerical values in the table represent the respective ‘R-square’ values for these various regression models. A higher R-square represents better relationship and means that the model significantly explains the resulting likelihood to recommend.



Based on these linear modelling results, similar factors impact a customer’s likelihood to recommend Hyatt in both New York and California. In California, the model that has the highest r-square and is most accurate is the one that only uses Guest Room Hospitality(GR), Hotel Condition(HC) and the quality of Customer Service(CS). This is our linear model for business guests in California. This model produces an adjusted r-square value of 0.6802 that means that almost 68% of the variation in the value of likelihood to recommend is a result of these variables.

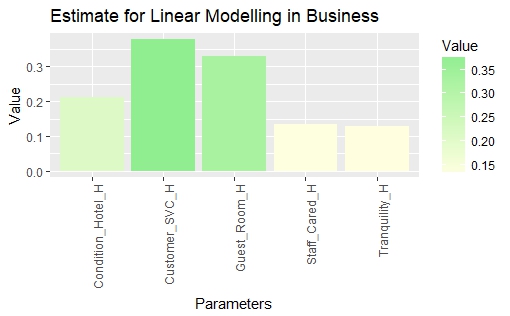
Similarly, in New York, the model with the best adjusted r-square value is the one that expresses likelihood to recommend as a result of Guest Room Hospitality(GR), Hotel Condition(HC) and the quality of Customer Service(CS). This is our linear model for business guests in New York. This model produces an adjusted r-square value of 0.7108 that means that almost 71% of the variation in the value of likelihood to recommend is a result of these variables.

However, it is evident that these 3 factors are slightly more significant for the New York model than they are for the California model as the model’s r-square for NY(0.7108)is higher than California (0.6802).

The models with Guest Room Hospitality(GR), Hotel Condition(HC), the quality of Customer Service(CS), Staff Cared(SC) and Tranquility(Tr) offer a slightly lower r-square but are still significant enough for comparison. This is the model that will be used in the comparison below.

Taking a closer look at the our models, we see that the coefficients of different variables in this models can tell which one of these variables has the most impact on the customers’ likelihood to recommend.

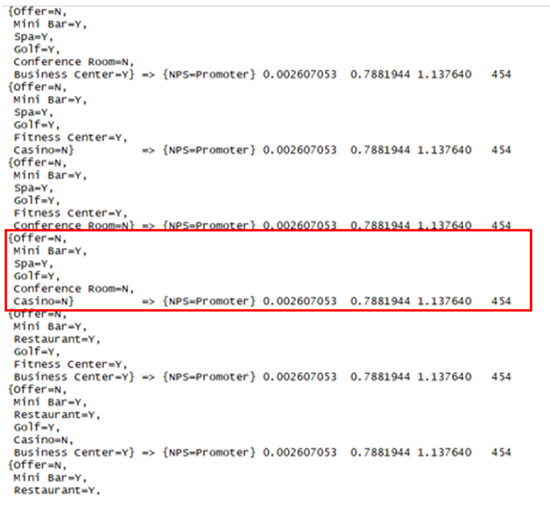
As we can see below, Customer Service(CS) has the most significant impact on the likelihood to recommend followed by Guest Room Hospitality. Hotel condition is next followed by Staff Cared and tranquility. We can also see clearly here that both Staff Cared and Tranquility have very little impact on the likelihood to recommend which explains why they were removed from our parsimonious model.



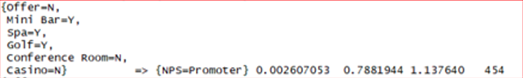
## Arules analysis

Since Linear Modelling only analyzes numerical data, we will use Arules again to understand the relationship and effect of categorical variables on a customer’s NPS type.

Following are the results from conducting an arules analysis on business customers only.



After sorting based on confidence, the following rule was found to be one of the best:

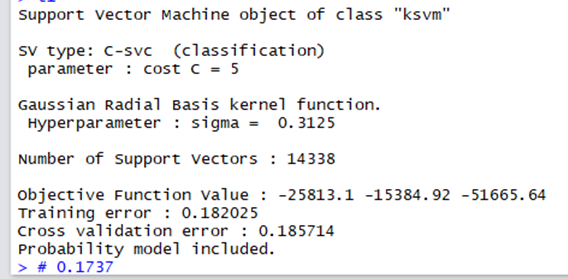


According to this, it appears that business customers frequently tend to be promoters when the hotel does not make a reservation offer, there is a minibar and a spa in the hotel and golf facility is available in the hotel. A 78.8% confidence means that of all the times these factors occur in this combination, about 79% of the times the customer ends up being a promoter. As per this analysis, the presence of a Conference room or casino does not significantly add to the likelihood of the customer being a promoter. Hence, spending a lot of money on adding a casino or a conference room to a hotel is unnecessary expenditure for Hyatt if the goal is to just boost NPS.

## Support Vector Machines(SVM)

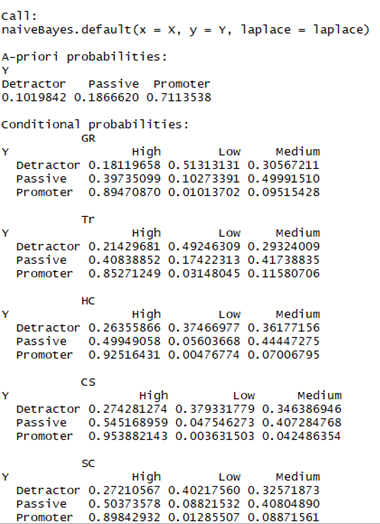
An SVM analysis is conducted here to see how well the NPS Type can be predicted for business customers based on our significant explanatory variables. The variables used here are the ones retrieved from the linear model on business customers. An SVM model is created using these variables and tested.

Results of SVM analysis on business customers



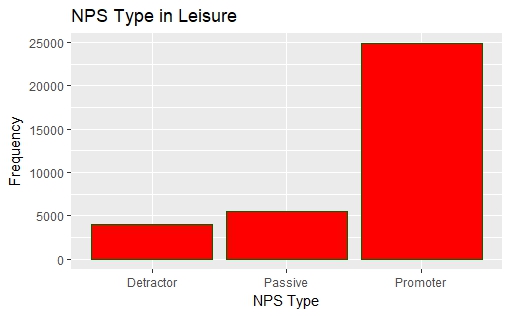
According to these results, the model is incorrect about 18.5% of the time. This means that over 81% of the times, this model can successfully predict whether a business customer will be a Promoter, Passive or Detractor just based on the hotel’s conditions, staff’s level of care, quality of customer service, guest room hospitality and tranquility.

## Naive Bayes

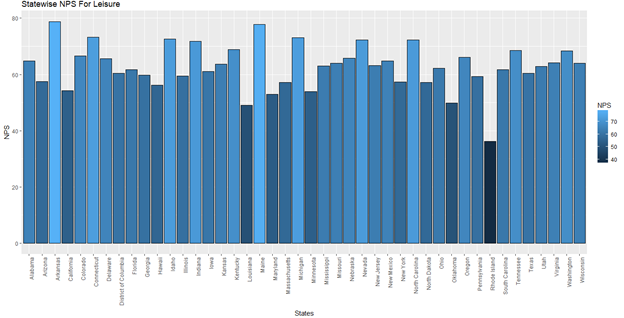


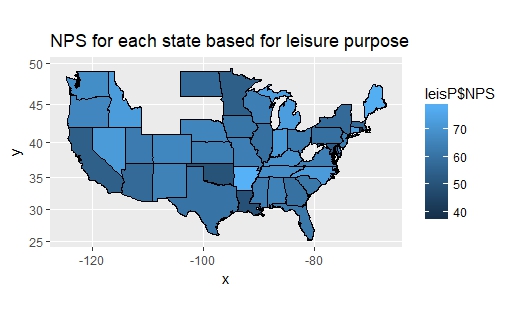
# **Leisure Customers:**

Having conducted our analysis for business customers, we will now move on to leisure customers and conduct the same analysis for them in order to understand their behavior.



Hyatt’s NPS Score for Leisure guests across different states

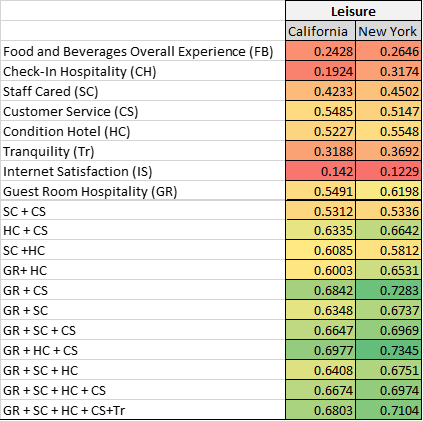




# Q4. Is there a difference in **criteria for recommending Hyatt hotels** between East Coast and West Coast guests?

## Linear Modelling

Shown below are the results of various single variable and multivariable linear regressions with likelihood to recommend as the dependent variable. These results are specific for leisure customers and are shown based on whether the customers are from California or New York. The column on the left lists the different linear regression models that were created and the variables used in them. The numerical values in the table represent the respective ‘R-square’ values for these various regression models. Higher the R-square, higher is the accuracy of the model and better is the relationship between the variables in it and the likelihood to recommend.



As was the case with business customers, this model suggests that similar factors affect a customer’s likelihood to recommend in both New York and California. In California, the model with the highest r-square and accuracy is the one that only uses Guest Room Hospitality(GR), Hotel Condition(HC) and the quality of Customer Service(CS). This model produces an adjusted r-square value of 0.6977 that means that almost 69% of the variation in the value of likelihood to recommend is a result of these variables. One notable thing here is that even after dropping the variable Hotel Condition from this model, the r-square value goes down by only a little bit. This suggests that Hotel Condition is not a very significant variable in this model and that leisure guests do not give very high importance to hotel condition.

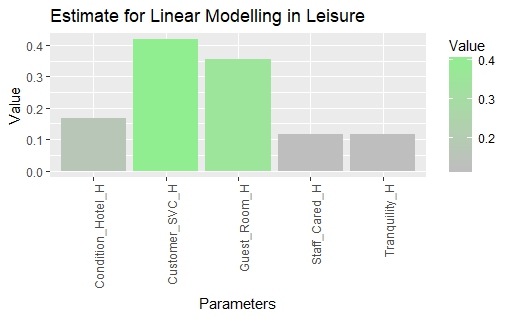
Similarly, in New York, the model with the best adjusted r-square value is the one that expresses likelihood to recommend as a function of Guest Room Hospitality(GR), Hotel Condition(HC) and the quality of Customer Service(CS). This model produces an adjusted r-square value of 0.7345 that means that almost 73.5% of the variation in the value of likelihood to recommend is a result of these variables.

Once again, it is tough to ignore that these 3 factors are slightly more significant for the New York model than they are for the California model as the model’s r-square for NY(0.7345)is higher than California (0.6977).

The model with Guest Room Hospitality(GR), Hotel Condition(HC), the quality of Customer Service(CS), Staff Cared(SC) and Tranquility(Tr) is not parsimonious but still offers a very high r-square. This is the model that will be used in the comparison below.

The following chart shows how significant the coefficients of our independent variables are when trying to predict likelihood to recommend.

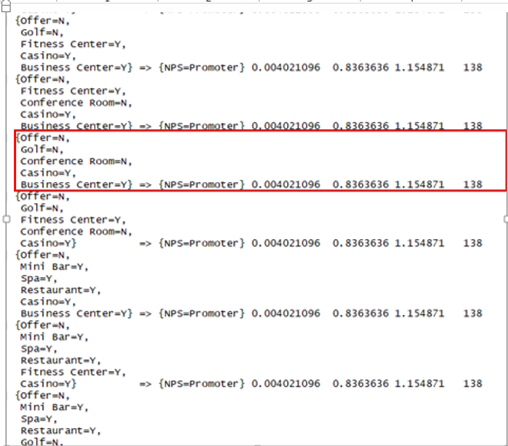
A closer look shows that Customer Service(CS) impacts likelihood to recommend more sharply that any other variable followed closely by Guest Room Hospitality. Hotel condition is next along with Staff Cared and tranquility. We can see clearly here that Hotel Condition, Staff Cared and Tranquility have very little impact on the likelihood to recommend.



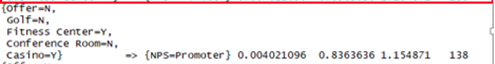
## Arules analysis

Since Linear Modelling only analyzes numerical data, we will use Arules again to understand the relationship and effect of categorical variables on a customer’s NPS type for our leisure customers

Following are the results from conducting an arules analysis on leisure customers only.



After sorting these rules based on their confidence level, the following rule was found to be one of the best:

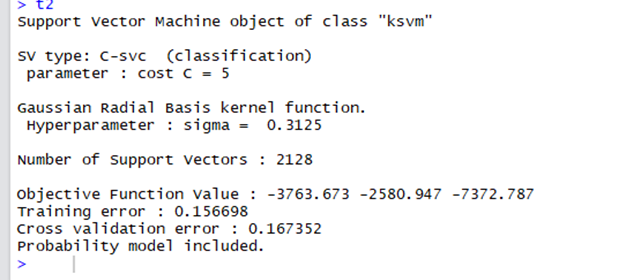


According to this rule, leisure customers tend to be promoters most often when the hotel does not make a reservation offer, there is no golf facility, no conference room but there is a casino and a fitness center. A 83.6% confidence means that of all instances when these factors occur simultaneously in this combination, about 83% of the times the customer ends up being a promoter, which is a very good ratio. This suggests that the presence of a golfing facility or a conference room do not significantly impact a leisure customer’s decision to recommend or not to recommend the hotel. Hence, spending money on adding a golf facility or a conference room to a hotel is unnecessary expenditure for Hyatt if the goal is to just boost NPS from its leisure customers.

## Support Vector Machines(SVM)

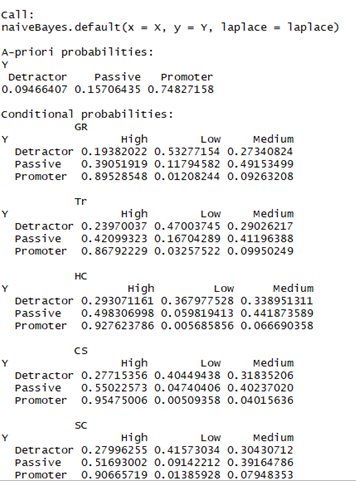
An SVM analysis is conducted here to see how well the NPS Type can be predicted for leisure customers based on the significant explanatory variables shortlisted from the linear model on leisure customers. An SVM model is thus created using these variables and tested.

Results of SVM analysis on leisure customers

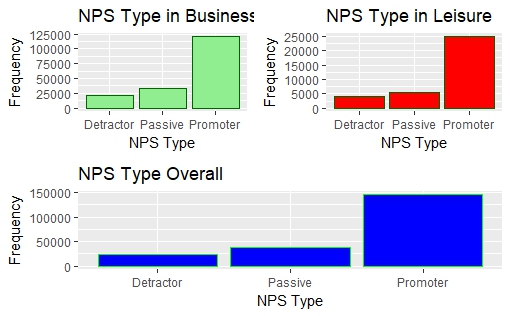


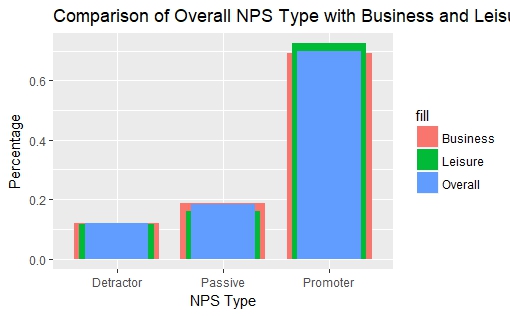
According to these results, the model is incorrect about 16.7% of the time. Over 83% times, this model can correctly predict whether a leisure customer will be a Promoter, Passive or Detractor for Hyatt just based on the hotel’s conditions, staff’s level of care, quality of customer service, guest room hospitality and tranquility.

## Naive Bayes

****

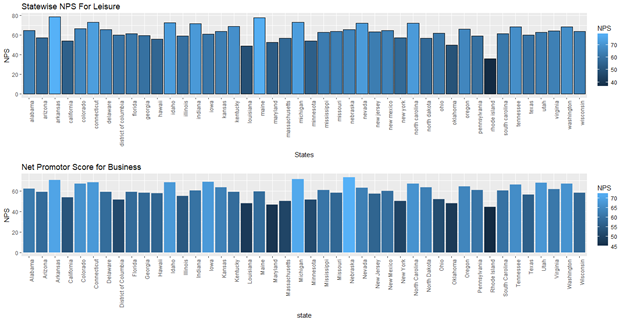
# **Comparing the results** from our analyses on all customers, business customers and leisure customers

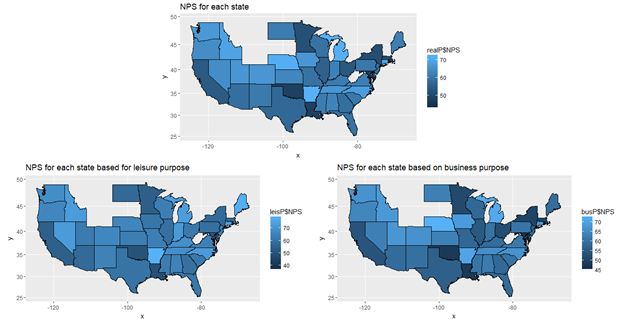




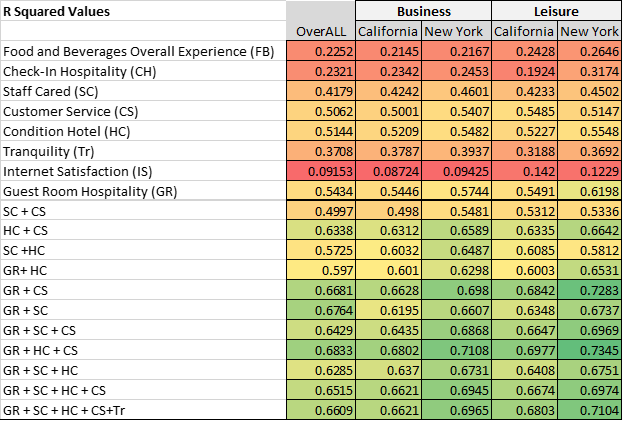
We can see from the charts above that the customers travelling for business are more than the number of guests coming in for leisure. So, it will not be wrong to say that the overall values or ratings are inclined towards how people travelling for business rate it.

## Geographical visualization of the results





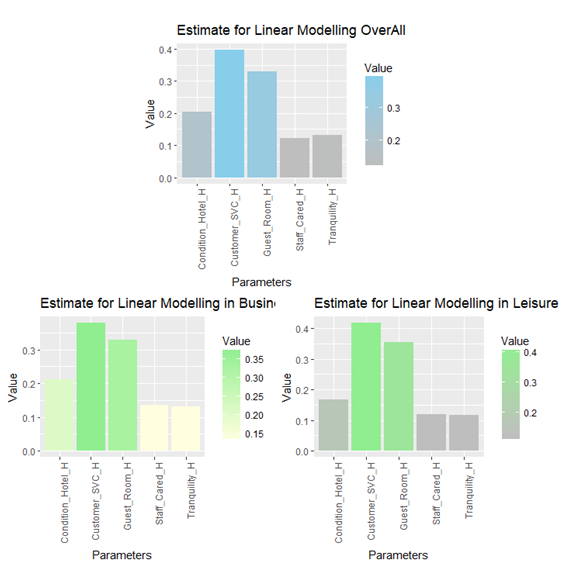
We realize from the charts above that Arkansas and Maine have substantially high Net Promoter Score from people travelling for Leisure, when compared with people travelling for business. Looking at from an unbiased perspective, we can see that in general people coming for leisure have given higher average net promoter score than that of the ones coming for business. Since all the other analysis did not show any path breaking results pointing out at why is this scenario as both the sets of customer have similar dependencies. We can not conclusively say what are exact reason behind it. However, it seems like Hyatt has a higher dependency on the parameters that were found above and they are really doing good in fulfilling the customers aspirations through the same.



For both Business and Leisure customers in New York, the same variables product the model with the best r-square. However, it is noteworthy that these three variables Guest Room Hospitality, Hotel Condition and Quality of Customer Service generate a better r-square for leisure customers in New York than business customers there. This signifies that leisure customers in New York give slightly more importance to these 3 factors than business customers do.

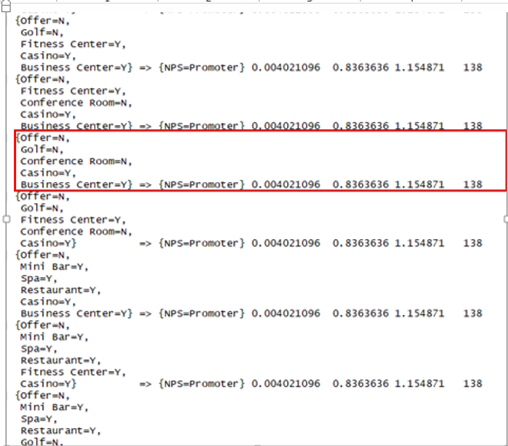
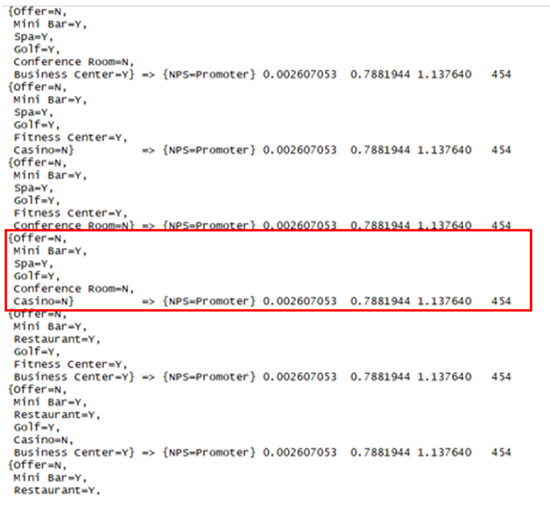
Similarly for California, the linear model with the three variables Guest Room Hospitality, Hotel Condition and Quality of Customer Service generates the best r-square for leisure customers. As with New York, leisure customers value these factors slightly more so than business customers. Moreover, it is observed that New Yorkers as a whole tend to give more importance to these factors than Californians as can be seen via the difference in their r-squared values.

When we analyze the coefficients of the 5 most significant variables for the linear models for both business and leisure customers, an interesting pattern emerges. The coefficient of Hotel condition is significantly higher for leisure customers than it is for business customers which clearly indicates that guests that stay at Hyatt locations in the US for leisure purposes denote more importance to the condition of the hotel. Hence, a leisure customers’ decision of whether or not the hotel should be recommended is significantly more dependent on the hotel’s condition that a business customer’s.



Some insight that could be key to Hyatt’s future performance here is that for its locations that receive a high percentage of leisure customers, Hyatt should work on improving hotel conditions. Based on the above analysis, locations with bad hotel conditions can deter NPS type significantly.

Business Leisure



Another interesting difference that we observed was that most business customers were promoters in instances when there was no casino in the hotel. This means they either do not prefer casinos or that they do not care enough about them to change their decision to recommend.

On the other hand, leisure customers react positively to the presence of a casino in the hotel in terms of their decision to recommend the hotel and their overall satisfaction from it.

This is a clear difference in consumer behavior for business and leisure customers when it comes to staying at Hyatt hotels. Hyatt could make great use of this inference by allocating its resources towards casinos in locations only where it receives a lot of leisure guests. Doing so can offer them a cost efficient way of boosting overall customer satisfaction and NPS.

Recommendations for Hyatt to increase their Net Promoter Score

Offer Casino facilities in areas where Hyatt receives more Leisure guests. The ARules analysis above showed how leisure customers prefer casinos more so than business customers.

c