## 

Hyatt Corporation Survey Analysis & Recommendations

## horizontal line

## 

IST-687

June 25, 2018

Juri Boyar, Brynn Ehni, Randall Lueck, Joseph Rhyne, Rakesh Tuteja

## 

## 

## 

## 

## 

## 

## 

## 

# Table of Contents

[**Table of Contents**](#_up9r03vtyiv) **2**

[**Introduction**](#_m8iimuc11ozk) **3**

[Background](#_r04s314lny13) 3

[Scope](#_a7bgkjbtm1dy) 3

[Driving Business Questions](#_jok9tmvsg8wr) 4

[What are the demographics of our top customers?](#_g3xstyxri5yp) 4

[What amenities or groups of amenities drives a guest’s likelihood to recommend?](#_zhjqybsjkklt) 4

[What are the drivers of promoter likelihood to recommend scores?](#_42m3bsatjreg) 4

[What location has the highest NPS score per hotel type per region? What factors contribute to this score?](#_x3wcc9ajgnq2) 4

[**Data Architecture & Cleaning**](#_kgjuq140mwk5) **5**

[Data Population](#_3ltudapv4exf) 5

[Data Transformations](#_j3ivxwnuzus2) 5

[Data Cleaning](#_g1zfl8d01nwl) 5

[Data Dictionary](#_seybamtdjp4j) 6

[**Descriptive Statistics**](#_6fp8l7pogube) **10**

[**Business Question Analysis**](#_v4qs958s0711) **13**

[What are the demographics of our top customers?](#_qrd5e38h5kho) 13

[Methodology](#_t5t3cda6fcso) 13

[Statistical Analysis](#_b4fm2aym6w1m) 13

[What amenities or groups of amenities drives a guest’s likelihood to recommend?](#_t2x2aa3qcqhx) 19

[Methodology](#_3376o0eivti7) 19

[Statistical Analysis](#_z6kdlgh3nfum) 20

[Path Forward](#_vnw5daqkiwj8) 25

[What are the drivers of promoters to likelihood to recommend?](#_zha1mporggai) 25

[Methodology](#_oesv7lgr72f8) 25

[Statistical Analysis](#_fe8y256dqlt9) 26

[What location has the highest NPS score per hotel type per region? What factors contribute to this score?](#_hqnanlhcwj56) 30

[Methodology](#_r22k2ybazfy4) 30

[Statistical Analysis](#_wcqeqrtbgson) 31

[**Recommendations**](#_r5nf2e901fne) **32**

[**Appendix A: Reproducible R Code**](#_1rmhopi6czio) **34**

# Introduction

Hyatt Hotels collected data via customer surveys from 2011 to 2015. The year with the most relevant data for customers that stayed at Hyatt brands in the United States came from 2014. These customers answered a wide range of questions pertaining to the locations at which they stayed. However, in order to focus on the information that had the most substantial effect on the total NPS score for each location, Team 2 narrowed the data set to only a few select columns within the year 2014, across all brands, with the following questions in mind.

## Background

The Hyatt brand was created in 1957. Headquartered in Chicago, this corporation consists of 14 premier brands that each operate as a subsidiary under the Hyatt family. Each subsidiary owns, develops, manages, operates, licenses, or provides services to hotels, vacation rentals, residencies, or resorts. As of March 2018, Hyatt Hotels Corporation has 777 properties across 54 different countries. In 2017, the Hyatt Hotels Corporation generated 4.69 billion dollars in revenue. The Hyatt Hotels Corporation’s mission is to provide a distinctive experience for their guests.

## Scope

The aim of this analysis is to answer questions that will help drive more business to the various hotel brands through customer satisfaction. This document hopes to capture a number of categories that, when analyzed, prove to directly affect the overall NPS score of Hyatt hotel surveys and ultimately aid in providing recommendations to the Hyatt Hotels Corporation to further their mission. This analysis will focus on surveys completed by guests during 2014 at locations within the United States.

## Driving Business Questions

### What are the demographics of our top customers?

The first business question is to understand the demographics of the the guests who completed the survey. This will help the Team to further understand if there is a concentration of certain demographic of guests that stay at a particular location of a brand within the Hyatt Hotels Corporation’s portfolio. Understanding the strengths and guests of each brand will help specialize each brand to fit that particular audience in order to serve their needs in the best way possible.

## What amenities or groups of amenities drives a guest’s likelihood to recommend?

This question aims to determine how certain amenities (i.e. pool, conference center, etc.) may influence a guest’s stay and ultimately influence NPS score. Understanding if certain amenities are integral to a guest’s satisfaction will influence the types of properties the Hyatt Hotels Corporation may acquire in the future as they look to expand their portfolio and locations.

### What are the drivers of promoter likelihood to recommend scores?

This question aims to determine if the customer service, hotel condition or hotel atmosphere are strongly correlated to the overall likelihood to recommend score. If so, this could help determine areas of focus for the hotel chains, like customer service training, ensuring a tranquil atmosphere, or improving food and beverage quality.

### What location has the highest NPS score per hotel type per region? What factors contribute to this score?

This question aims to determine the best locations in each region based on NPS scores. Focusing on the five regions of the United States, we plan to discover which regions, business models, and locations drive NPS outcome. If successful, this data will provide the template for the qualities each hotel in that region should strive for.

# 

# Data Architecture & Cleaning

## Data Population

The data population consists of business stays at all Hyatt brands in the United States from the year 2014. This limits our population to 3,464,450 records across 38 variables since most of the data from the original data frame falls into these two categories. Focusing on this smaller population will still provide plenty of data to help target business travelers across hotel brands.

## Data Transformations

The Hyatt Survey data comes from four main sources with overlapping information: Enterprise Data Warehouse Checkout, Enterprise Data Warehouse Reservation, Medallia Invitation, and Medallia HySat. Because this information came from separate sources, with overlapping, null or potentially conflicting information, the Team created a hierarchy of sources for essential survey information. To do this, master columns were created for the following information:

* m\_check\_in\_date
* m\_check\_out\_date
* m\_length\_of\_stay
* m\_pov\_code
* m\_guest\_country
* m\_guest\_gender
* m\_survey\_status

## Data Cleaning

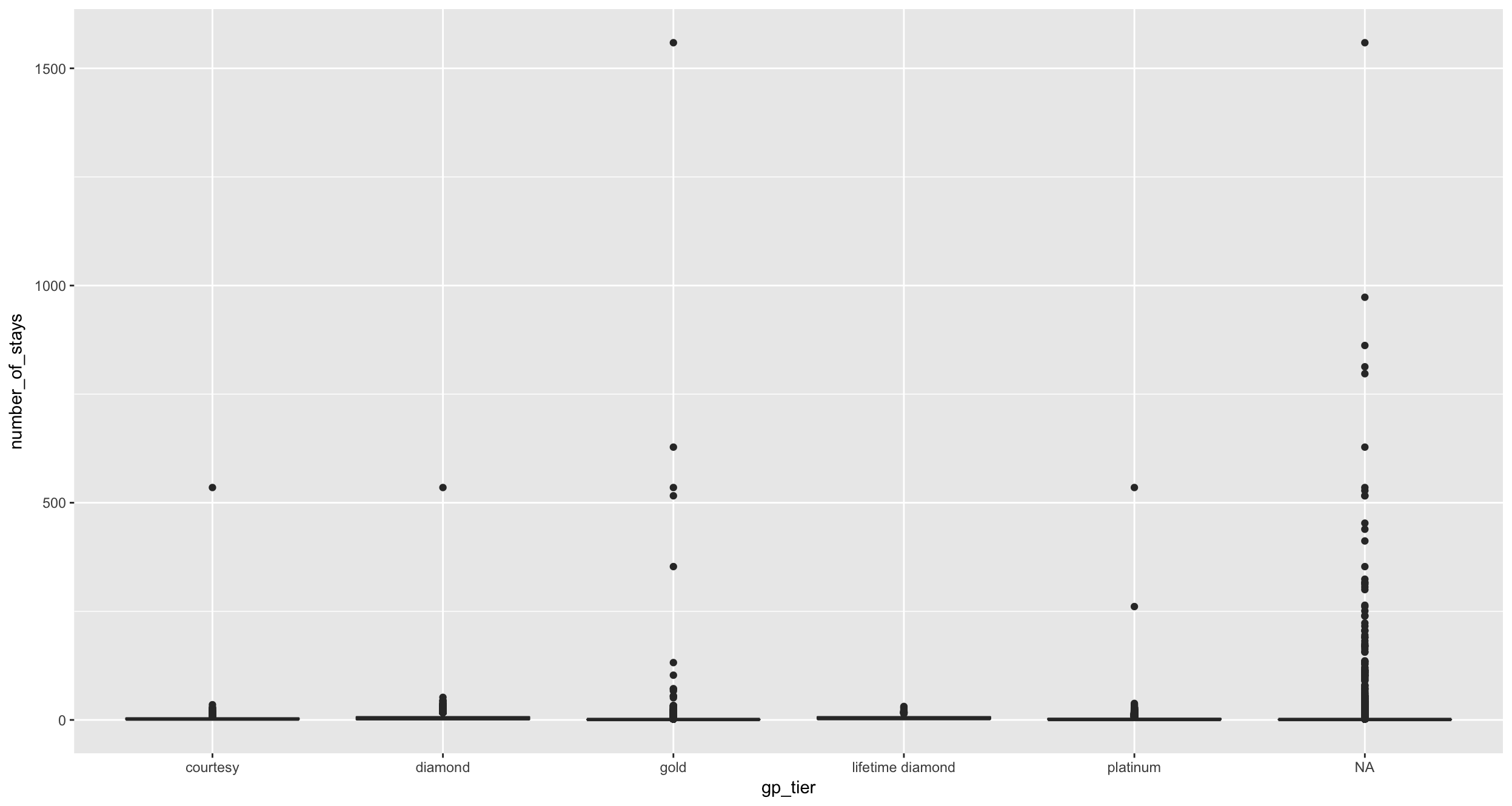
The data needed to be cleansed prior to analysis. All the satisfaction scores were converted to numeric types and all date values were stored as date types.The guest state information was cleaned to include only the 51 United States State Abbreviations (including the District of Columbia). The rest of the invalid values were set to null. The guest country values were cleaned to valid country names. If the guest state was populated, the guest country values were set to the United States. Finally, the GP tier information was conformed so the values were uniform. Equally important, we found repeat customers that had missing NPS scores. To deal with this, we placed their NPS score with a value of 9 because they chose to return to the Hyatt Corporation. All other missing values were addressed appropriately in each business question.

## Data Dictionary

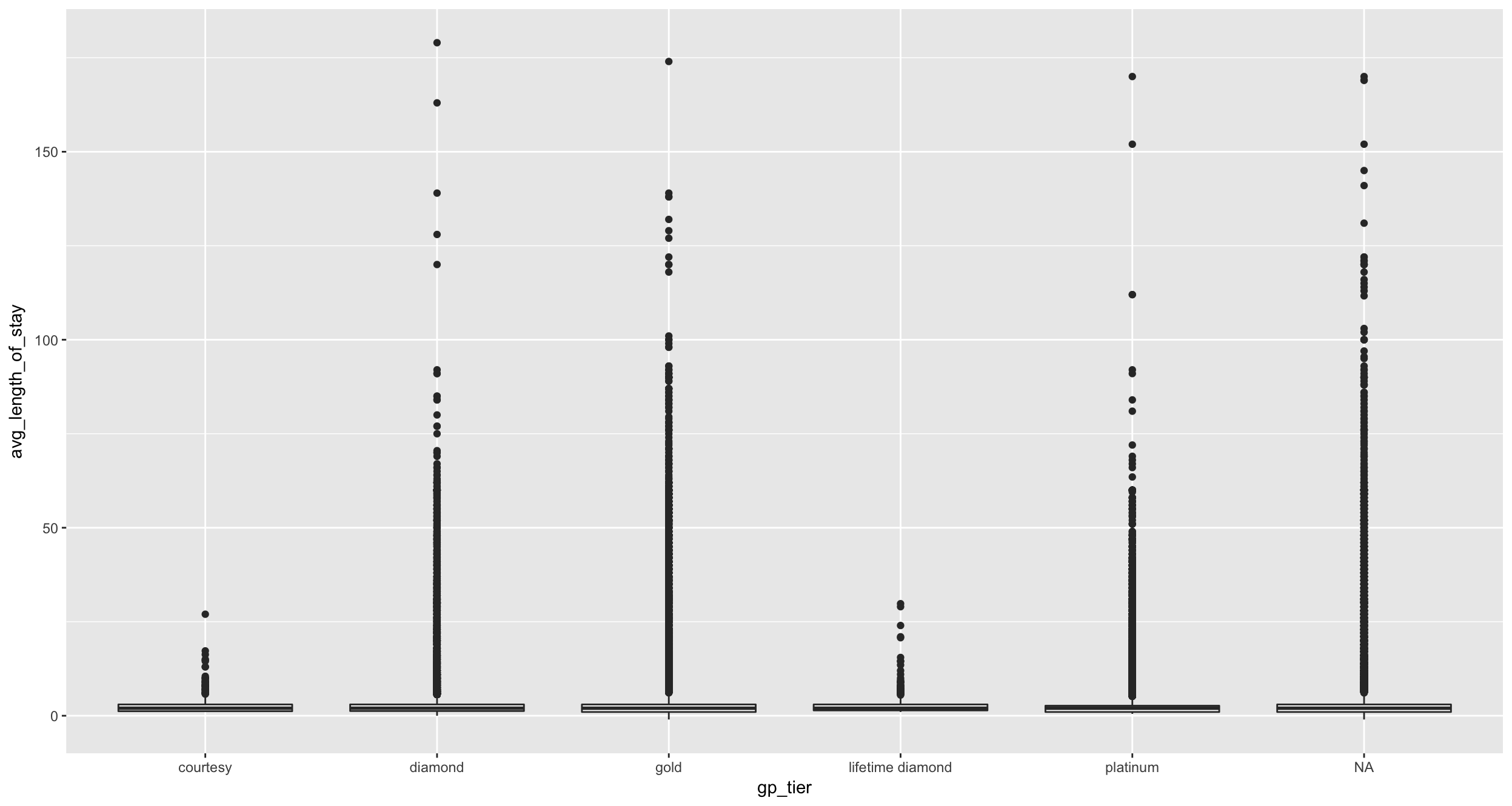
|  |  |  |  |
| --- | --- | --- | --- |
| **Column Number** | **Variable Name** | **Hyatt Definition** | **Group 2 Reasoning for Inclusion** |
| 1 | cons\_guest\_id\_c | System generated number that identifies a customer profile; this is generated in MDM/GEM | Uniquely identifies each guest. Important for gathering demographics for guests and determining if a guest has stayed multiple times. |
| 2 | m\_check\_in\_date | Check in date; for WALK status adjusted to the first in-house stay date | Date of check-in will be used to uniquely identify each stay of each guest, rather than check-out date. Cursory examination of the data shows that more complete records were included for 2014 and not for 2013 or 2015. |
| 3 | m\_pov\_code | Purpose of visit | Coalesce function used to combine data from myData$pov\_code\_c and myData$pov\_h. Purpose of visit, either business or pleasure, is important in determining the needs of a guest. Cursory examination of the data shows that the vast majority of guests stayed at Hyatt on a business trip. Group 2 intends to focus on business travelers. |
| 4 | pms\_total\_rev\_usd\_c | Total from the PMS in USD | Group 2 would like to see the total revenue from the guest's stay. Promoter guests should spend more money at the hotel. |
| 5 | adult\_num\_r | Number of adults for which the reservation was made | Number of individuals on the trip. Are rooms shared frequently? |
| 6 | children\_num\_r | Number of children associated with the reservation | Were any children along on the trip? This will have an impact on the desired hotel amenities. |
| 8 | m\_survey\_status | Status of the survey | Did the guest complete the survey? Completed surveys indicate a higher degree of effort on the guest's part.  Coalesced data from status\_h and e\_status\_i. |
| 9 | m\_guest\_country | Guest's country | Coalesced data contained in GUEST\_COUNTRY\_R and e\_country\_i. In some instances, GUEST\_COUNTRY\_R contains data when e\_country\_I does not and vice-versa. 'I' code variables are sourced differently than 'R' code variables. |
| 10 | guest\_state\_h | Guest's state | Important demographic. Most guests are domestic travelers. Home state information will be useful in gaining further insight about guests. |
| 12 | age\_range\_h | Guest's age range | Important demographic. Guests in different age ranges may prefer different amenities or have differing opinions. |
| 14 | net\_rev\_h | Net revenue | How much money was made off of the guest's stay. |
| 15 | room\_rev\_h | Room revenue | How much money was made off of the guest's room. Did the guest spend money elsewhere on the property? |
| 17 | likelihood\_recommend\_h | Likelihood to recommend metric; value on a 1 to 10 scale | Critical. This is a discrete numerical value used to determine whether a guest will be a detractor, neutral or a promoter. The NPS\_Type value is computed from this value. |
| 18 | overall\_sat\_h | Overall satisfaction metric; value on a 1 to 10 scale | Very important. In some instances, guests provided this score, but not the Likelihood\_Recommend\_H score. It would be possible to conclude that satisfied customers would provide a high recommendation score if it is missing. |
| 19 | award\_category\_pl | Gold Passport award redemption category | High value properties will require larger point redemptions. |
| 20 | brand\_pl | Hotel's brand | Do some customer have more loyalty to some brands than others? Are there more promoters for brands? |
| 21 | brand\_initial\_pl | Hotel's brand abbreviated | Abbreviation of Brand\_PL |
| 22 | gp\_tier | GP tier of the guest, coalesced from multiple sources (Reservation, Invitation, HySat) | What is the guest's loyalty status? Are some guests more loyal to certain Hyatt brands? Are loyalty members generally more satisfied when staying at Hyatt? |
| 23 | nps\_type | Indicates if the guest's HySat responses mark them as a promoter, a passive, or a detractor | Calculated from Likelihood\_Recommend\_H. Values of 9 and 10 are considered promoters. 7 and 8 are neutral. 1 to 6 are detractors. |
| 24 | spirit\_pl | Unique hotel identifier (5-letter code) | Hotel Attributes, for Visualization Map or Charts |
| 25 | property\_id\_pl | Unique hotel identifier (numeric) | Hotel Attributes, for Visualization Map or Charts |
| 26 | hotel\_name\_long\_pl | Full hotel name | Hotel Attributes, for Visualization Map or Charts |
| 27 | hotel\_name\_short\_pl | Abbreviated hotel name | Hotel Attributes, for Visualization Map or Charts |
| 28 | type\_pl | Type | Five categories of the business model for the hotel. Business, Convention, Franchise, Resort, and Select Service |
| 29 | location\_pl | Location | Type of community this hotel serves. Options are Airport, Resort, Suburban, and Urban. |
| 30 | status\_pl | Status of the hotel. Valid values: open, closed, pipeline, pre-opening | Hotel Attributes, for Visualization Map or Charts |
| 31 | city\_pl | City in which the hotel is located | Hotel Attributes, for Visualization Map or Charts |
| 32 | state\_pl | State in which the hotel is located | Hotel Attributes, for Visualization Map or Charts |
| 33 | us\_region\_pl | US region in which the hotel is located | Hotel Attributes, for Visualization Map or Charts |
| 34 | postal\_code\_pl | Zip code in which the hotel is located | Hotel Attributes, for Visualization Map or Charts |
| 35 | country\_pl | Country in which the hotel is located | Hotel Attributes, for Visualization Map or Charts |
| 36 | property\_latitude\_pl | Latitude of the hotel's location | Hotel Attributes, for Visualization Map or Charts |
| 37 | property\_longitude\_pl | Longitude of the hotel's location | Hotel Attributes, for Visualization Map or Charts |
| 38 | guest\_nps\_goal\_pl | Hotel's NPS goals | Hotel Attributes, for Visualization Map or Charts |
| 39 | hotel\_inventory\_pl | Size of hotel (number of rooms) | Hotel Attributes, for Visualization Map or Charts |
| 40 | all\_suites\_PL : valet\_parking\_pl | Flagged variables indicating whether a property offers the amenity | Used for examination of available amenities and their impact on guest NPS type. |
| 41 | nt\_rate\_r | Nightly rate | Used for examination of driving predictors of overall likelihood to recommend. |
| 42 | m\_length\_of\_stay | Length of guest’s stay | Used is descriptive statistics. |
| 43 | Guest\_room\_h : f&b\_overall\_experience\_h | NPS score variables indicating satisfaction with various hotel qualities | Used for examination of driving predictors of overall likelihood to recommend. |

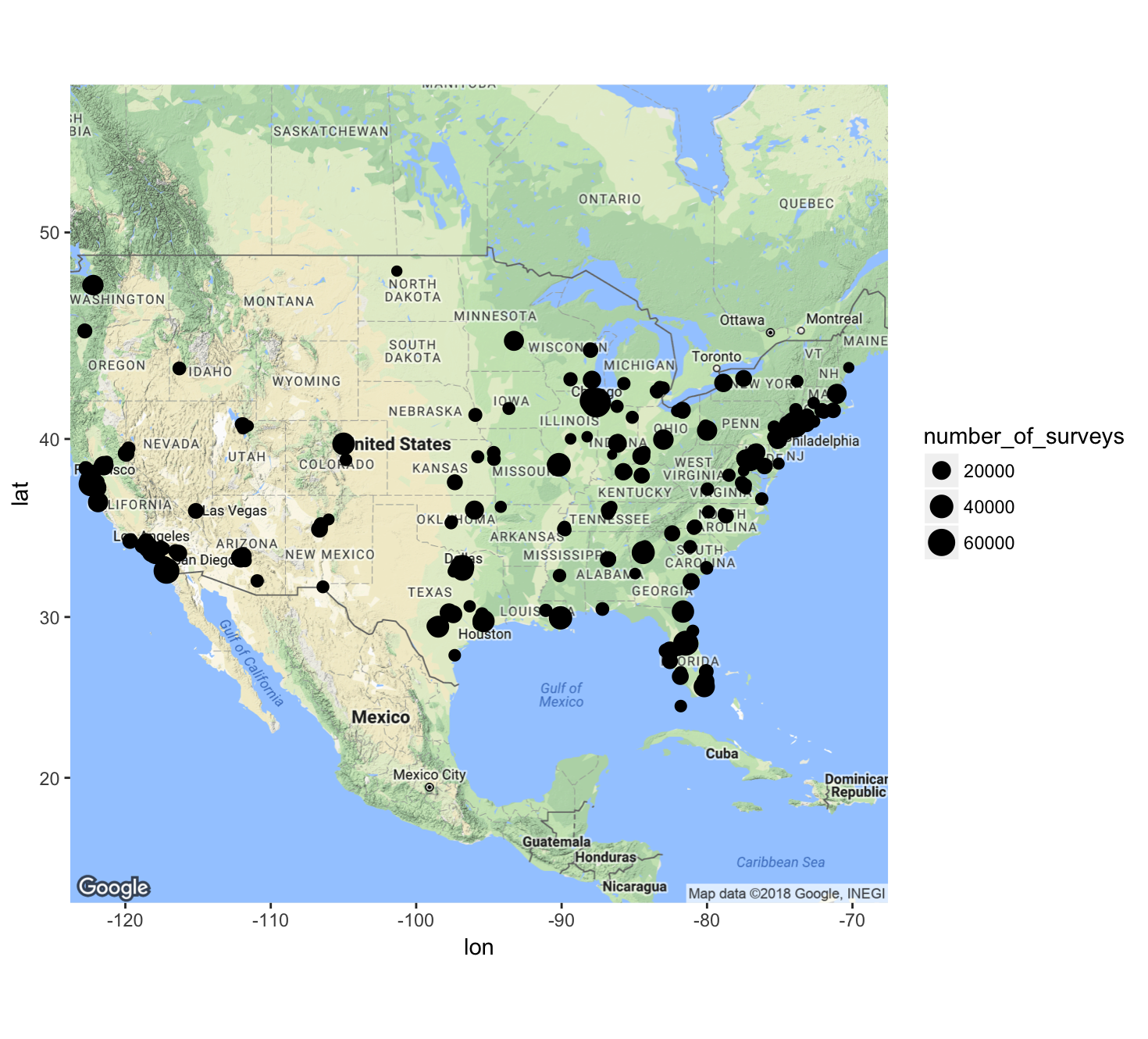
# Descriptive Statistics

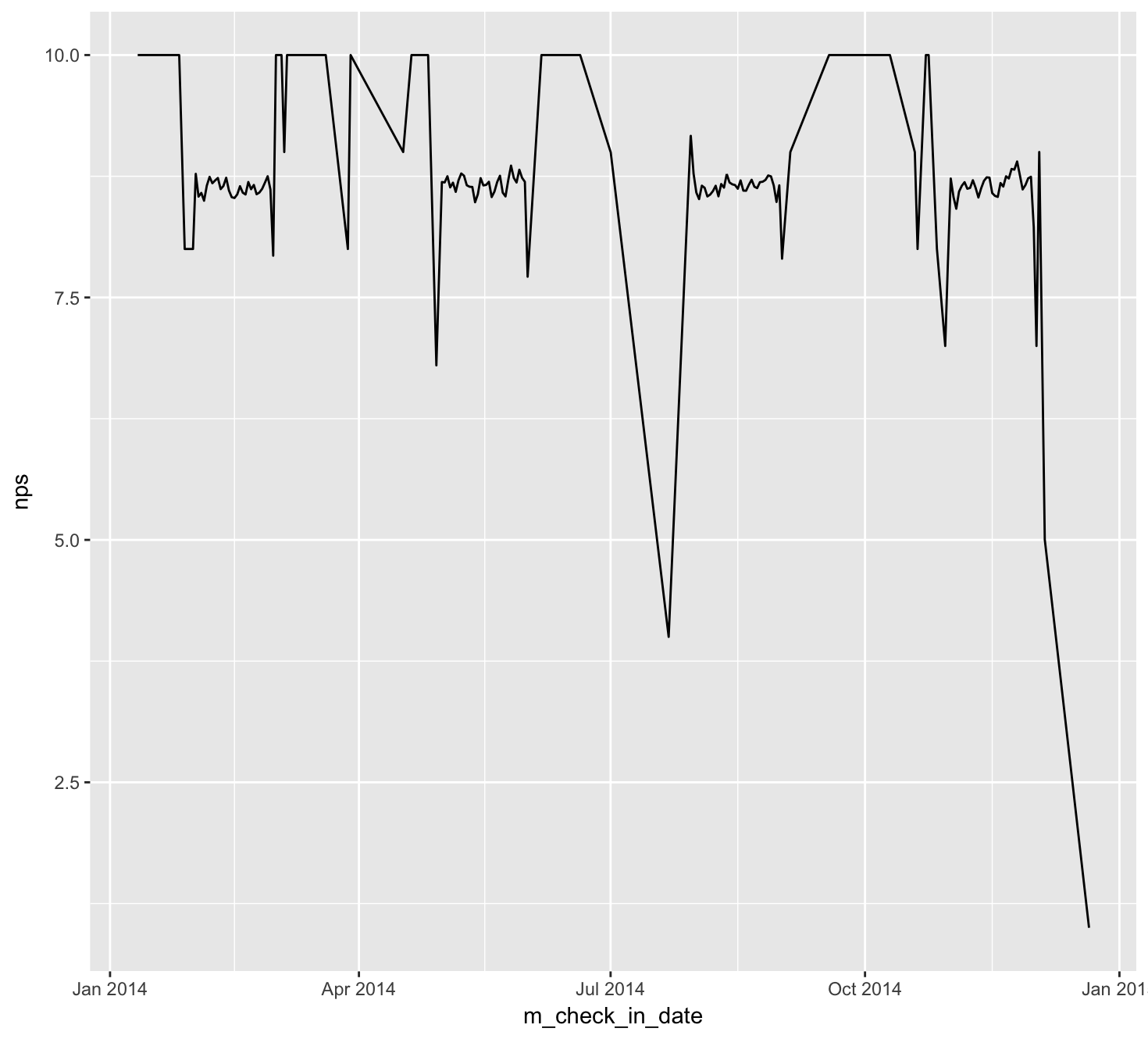
This boxplot shows the number of stays grouped by the gold passport tiers. Each cons\_guest\_id\_c is unique in this visual. As may be seen, there are outliers in this dataset, with repeat customers staying more than 1500 times in 2014.



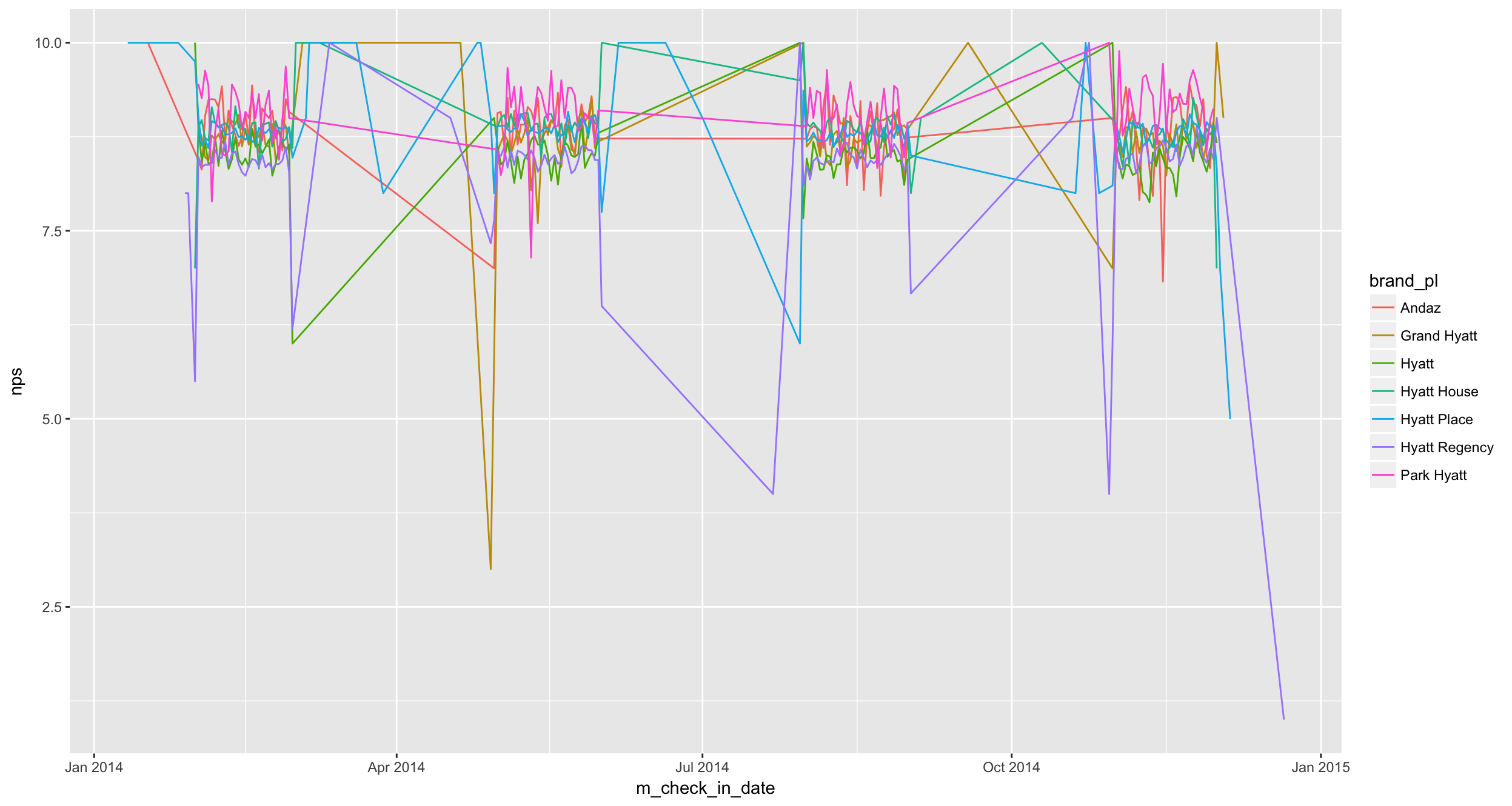
This boxplot shows the average length of stay per customer, grouped by gold passport tiers. It is clear that while each tier has approximately the same center around 1-5 business days, the gold, platinum and diamond members have a long tail compared to that of the courtesy and lifetime diamond members.



This map shows each of the US hotel locations. The size of the dot represents the number of surveys that came from that location. Thus, it is easy to see that the largest location represented in this dataset is Chicago, and that there are also large clusters in the New England, Florida, LA, and San Francisco area.

This graph shows the overall likelihood\_recommend\_h value over time. It is easy to tell in this graph that the majority of surveys are being entered in couple month intervals with a few months break in between. The lines stay relatively steady at 8 on the y-axis, then smooth out before jumping significantly. This indicates lower survey volume at these times.

Finally, this graph shows the overall likelihood to recommend broken out by brand. Here it may be seen that the Hyatt Regency brand had significant drops - more so than any other brand - in their NPS score during certain months (which is also when the survey count was lower).



# Business Question Analysis

## What are the demographics of our top customers?

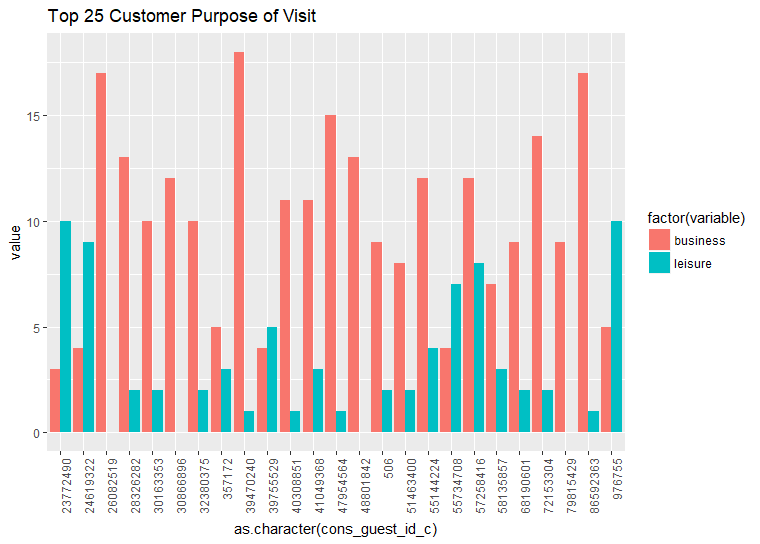
### Methodology

To study the demographics of the top customers, the team attempted to find where specific customers came from. We focused on the following information to extract the demographics of the top Hyatt customers across the country:

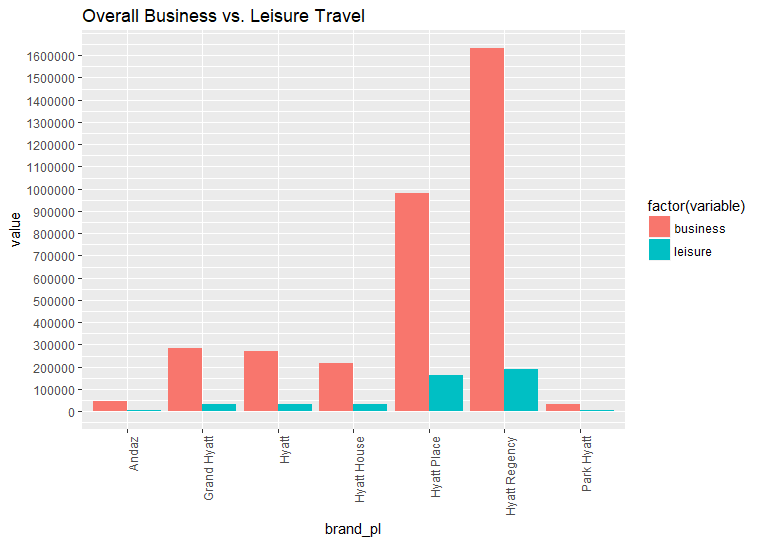
* Type of travel
* Hotels chosen by top customers
* Top spending guest countries
* Top spending guest states

### Statistical Analysis

To begin the analysis of top customer demographics, the team separated the data to show which type of travel, business or leisure, appeared most often for the top 25 customers.



The resulting graph shows that, while six of the top customers are more likely to travel for leisure when staying at one of the Hyatt brand hotels, most customers visit Hyatt hotels for business. To further investigate how often customers traveled for business instead of leisure at each brand of hotel, the team pulled the data for all stays within the data set for both categories:

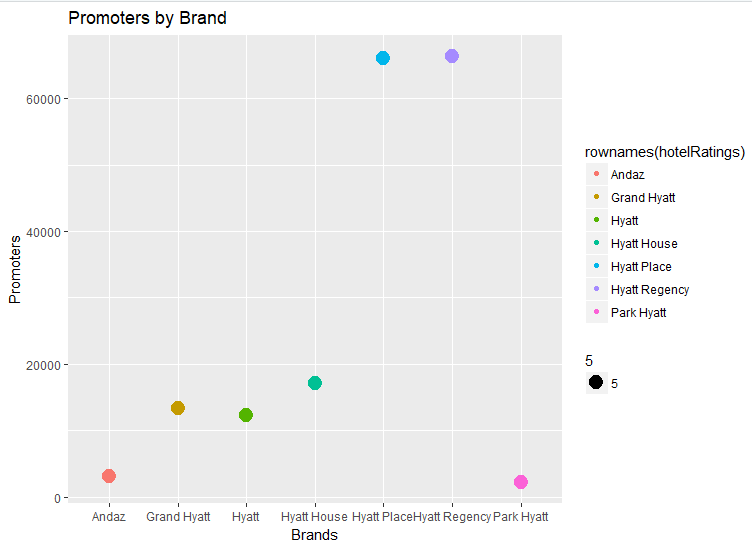


The above graph shows that customers overwhelmingly travel for business more frequently than they do for travel. This graph informs the decision to direct attention more toward business customers for most of the Hyatt brands. With focus on business customers over leisure customers, the team decided to see how different brands were impacted by their individual NPS score.

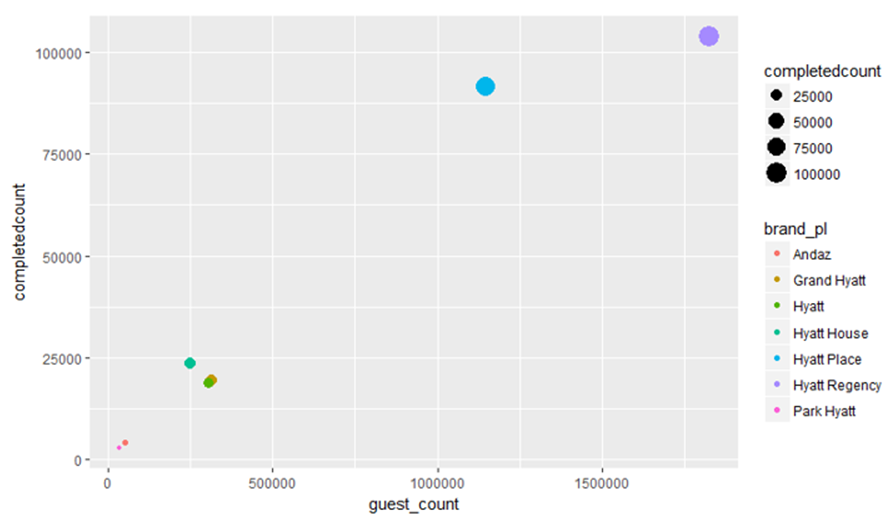
The below table identifies the top brands organized by the amount of customers who voted for an NPS score of 9-10 for the brand where they rented a room. This table identifies the Hyatt Regency and Hyatt Place as the locations where most customers stay. However, when likelihood to recommend is put into a ratio against not likely to recommend, Hyatt House emerges as the best brand.

|  |  |  |
| --- | --- | --- |
| **Brand** | **Not Likely to Recommend(<9)** | **Likelihood to Recommend (9+)** |
| Hyatt Place | 24128 | 66095 |
| Hyatt Regency | 35487 | 66363 |
| Hyatt House | 6018 | 17098 |
| Grand Hyatt | 5716 | 13389 |
| Hyatt | 6293 | 12266 |
| Andaz | 1104 | 3075 |
| Park Hyatt | 591 | 2207 |

## 

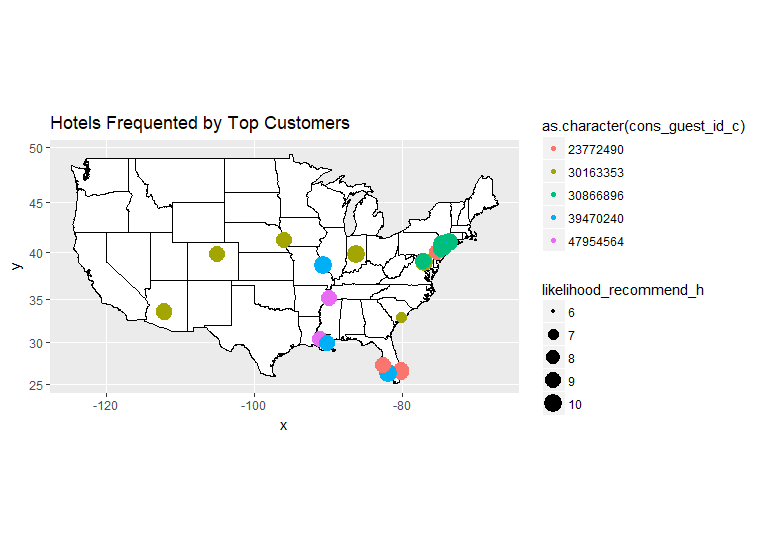


Hyatt House is the extended stay brand of Hyatt Hotels and accounts for the third largest United States footprint for Hyatt out of the seven brands found in the data set. In the above graph, Hyatt House has the third largest number of promoters. However, this brand has far fewer customers than either of the two top brands as seen in the graph below. Which shows the number of completed surveys per brand.



We’ll later explore some differences between hotel brands to try to find a driving factor for NPS scores between the business models associated with each of the hotel brands. For now, we’ll continue on to explore the locations most visited by our top customers.

The below map illustrates the locations most often booked by the five customers most likely to recommend staying at a Hyatt brand hotel:



The first customer listed on the map, ID# 23772490, was shown in an earlier chart to mostly stay at Hyatt hotels for leisure. This customer stays mainly on the east coast of the United States for both work and travel according to the graph. The other four top customers most often stay at Hyatt brand hotels for business, and also have their rooms booked by the companies for which they work. If this is true for most of the Hyatt’s customers, then a large portion of customers staying at Hyatt hotels for leisure may be doing so because they earn some sort of reward points from when they stay at Hyatt brand hotels for work.

There is a striking difference however, between these top customers, who all originally come from and work in the United States, and those customers who travel from other countries to stay in the United States for either business or leisure visits.

Visitors from other countries tend to spend up to six times as much as United States residents according to the below table.

Top 5 spending guest countries:

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Average Revenue** | **Median Revenue** | **Stays Per Country** |
| Tunisia | $2508.02 | $1136.91 | 25 |
| Greenland | $2308.86 | $932.35 | 3 |
| Bhutan | $2032.99 | $2032.99 | 1 |
| Libyan Arab Jamahiriya | $1945.82 | $1945.82 | 1 |
| Equatorial Guinea | $1880.05 | $1880.05 | 1 |

Top 5 spending guest states:

|  |  |
| --- | --- |
| **State** | **Average Revenue** |
| Hawaii | $745.29 |
| Alaska | $598.03 |
| Washington | $578.20 |
| Montana | $552.49 |
| Oregon | $544.33 |

The United States is in 182nd place compared to other countries, with only $407.72 per stay spent on average and a median price of $264.22. This seems very low but, three of the top five countries listed above only account for one stay, whereas United States residents account for 2.8 million stays out of the 3.9 million stays listed in the data set. This leads the team to believe that focus should be placed mostly on people traveling for business in the United States in any marketing campaign.

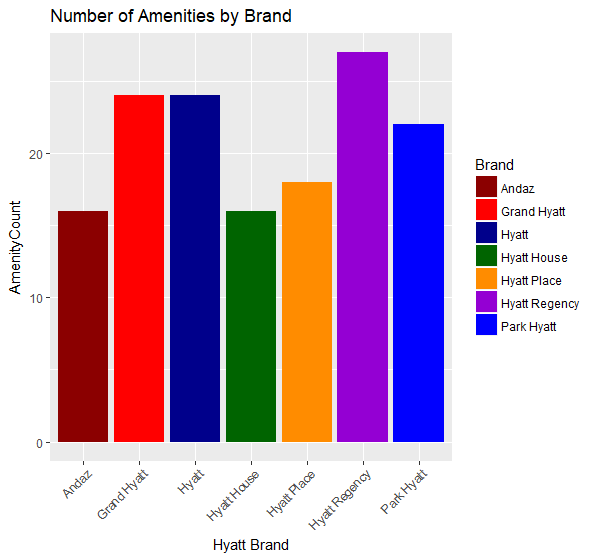
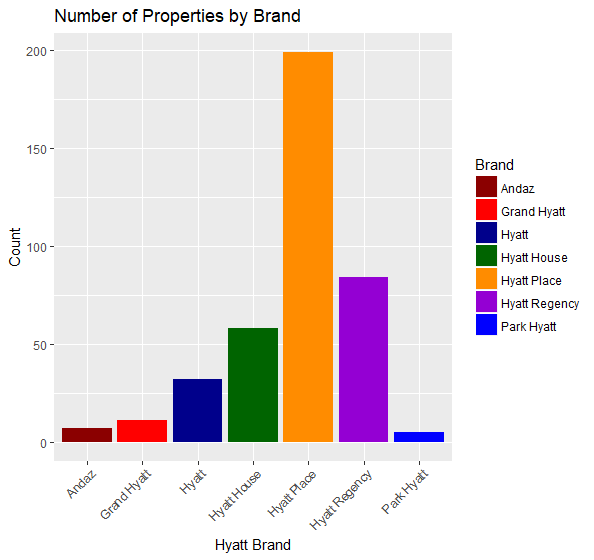
## What amenities or groups of amenities drives a guest’s likelihood to recommend?

### Methodology

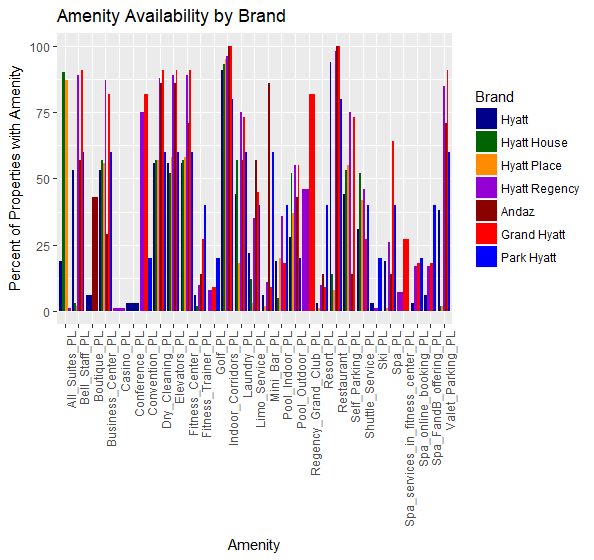
To answer this business question, the team needed to answer a number sub-questions related to amenities. These included: Which brands have certain amenities and which are missing? From business question 1, is there a correlation between the ratio of promoter to detractor NPS scores and the consistency of amenity availability? Overall, is there a relationship between certain amenities and NPS scores? Is it possible to predict the influence that an amenity will have on an NPS score?

### Statistical Analysis

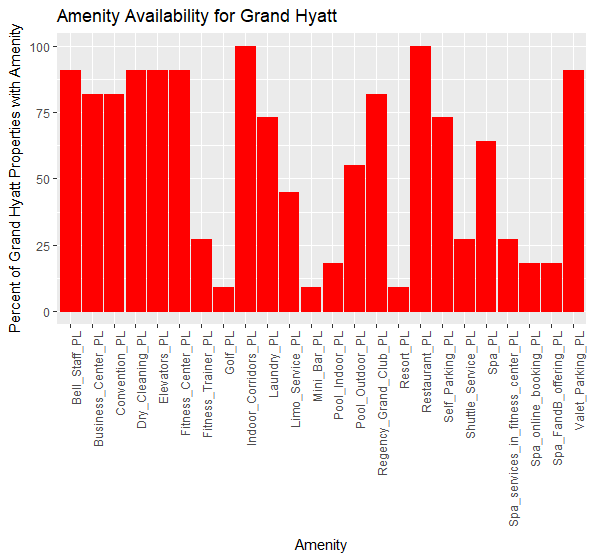
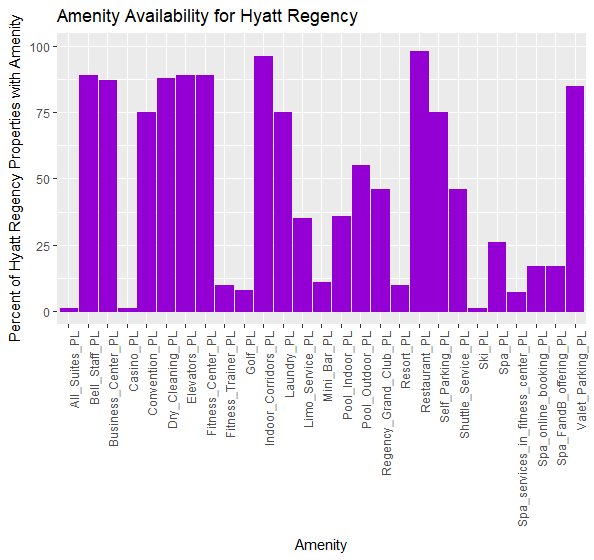
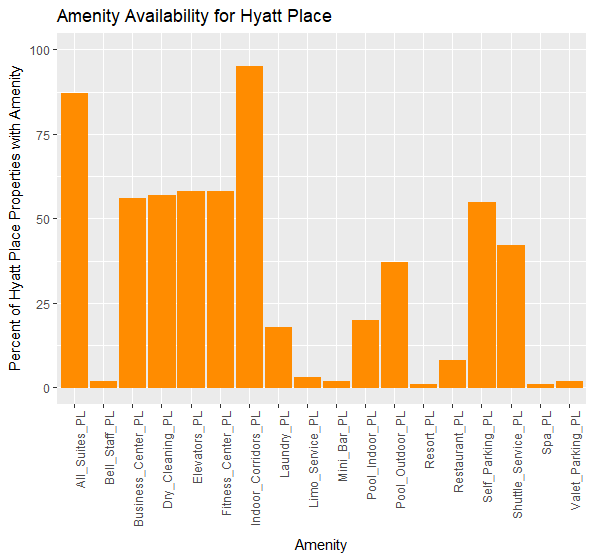
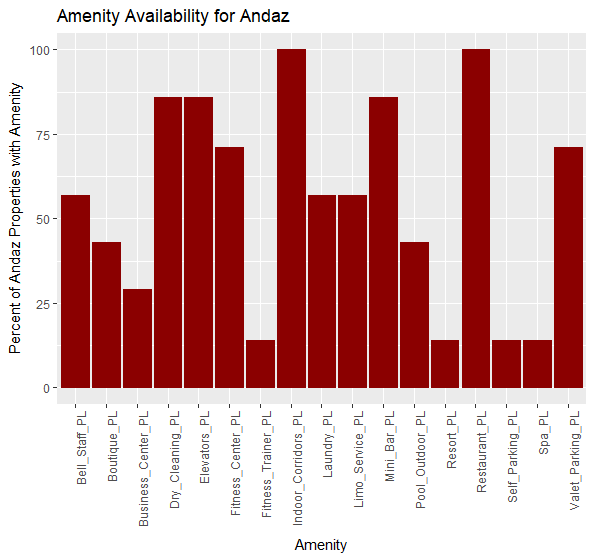
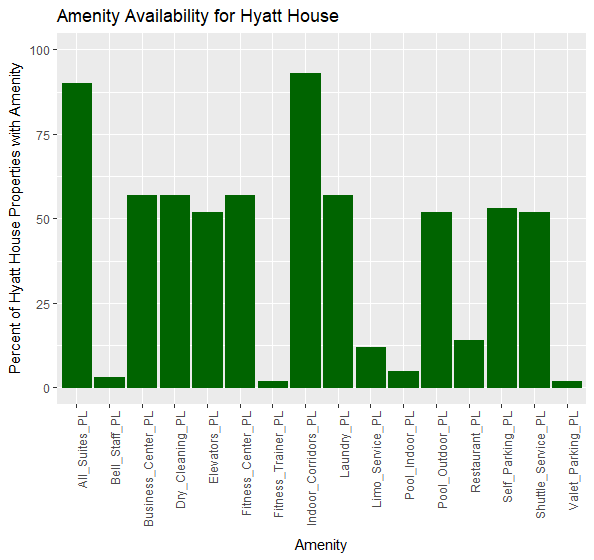
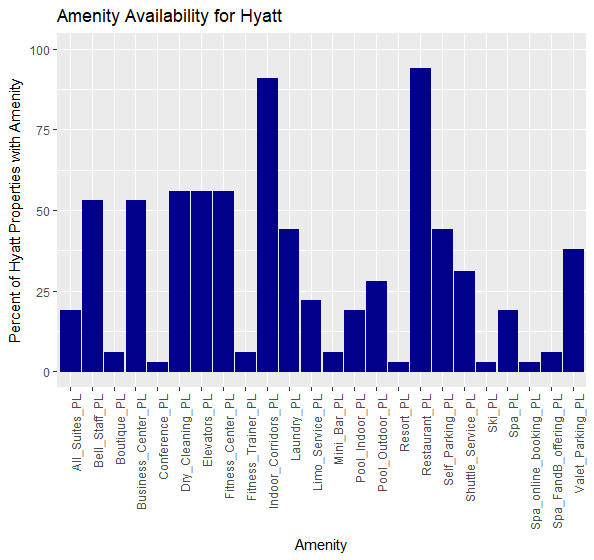
To begin, the team examined each brand and the total number of available amenities. Unique properties were extracted from the survey data and Hyatt Place was found to be the most abundant brand, with 199 properties. Hyatt Place was also towards the lower end of the brands in terms of amenities, offering only 18 total across all properties. The Hyatt Regency brand was the second most numerous with 84 properties, but offered 27 different amenities. Hyatt Regency had the worst likelihood to recommend ratio (promoters / detractors) of 1.87, while Hyatt Place had an above average ratio of 2.74. There does not appear to be relationship between the number of amenities offered and the NPS score. This was confirmed by creating a linear model where the number of amenities describes the NPS ratio. The model produced has an R-squared of only 0.12 and p-value of 0.23.

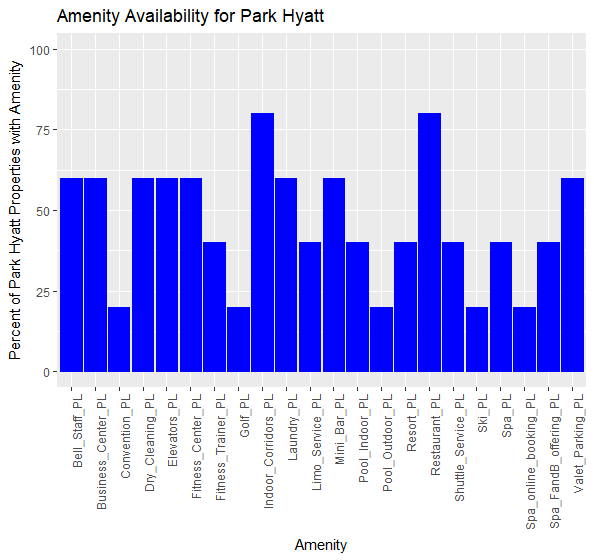


Next, the team investigated the mixture of amenities available at each of the brands. The plot below combines all the properties, showing the percentage of each brand’s properties with a particular amenity.



This plot is very busy, but it does give a sense of some of the gaps in available amenities. For example, very few properties have casinos, conference centers or skiing available. Certain brands are more consistent with the amenities they offer, while others, never offer a particular amenity at all. To get a better sense of this, each brand should be examined in turn.





The plots illustrate a number of statistics. With just 16 offered amenities among all of its properties, Hyatt House has the fewest, while Hyatt Regency has with most with 27. Of all the brands, Hyatt is the least consistent, with only 7 of the 24 possible amenities offered at 50% or more of the properties. Hyatt House is the most consistent, with 10 of the 16 possible amenities offered at 50% of more its properties.

The table below shows the likelihood to recommend ratio for each brand compared to consistency of amenities.

|  |  |  |
| --- | --- | --- |
| **Brand** | **Likely to Rec Ratio** | **% of Amenities Offered at 50% or More of Properties** |
| Hyatt | 1.95 | 29% |
| Hyatt Place | 2.74 | 39% |
| Hyatt Regency | 1.87 | 44% |
| Park Hyatt | 3.73 | 45% |
| Grand Hyatt | 2.34 | 58% |
| Andaz | 2.79 | 59% |
| Hyatt House | 2.84 | 63% |

There does not appear to be a correlation between consistency of amenities and the likelihood to recommend. This was confirmed by building a linear model where the percentage of amenities offered at 50% or more of properties describes the NPS ratio. The model has an R-squared of 0 and a p-value of 0.5.

To assist in narrowing the field of amenities to consider for further analysis, a Chi-Squared test for independence was completed. Each amenity was compared against the NPS type, either promoter or neutral/detractor. 22 of the 29 amenities were found to have p-values below an alpha of 1%. These 22 amenities were included in the Scalable Vector Machine model.

To help predict whether a guest will be a promoter or a neutral/detractor based on the combination of amenities, a Support Vector Machine model was built. The SVM was able to analyze a data set of 262,639 total surveys, of which 180,541 were promoters and 82,098 were detractors. This indicates a natural rate of 68% promoters. The SVM was trained on a random selection of one sixteenth of the data set (the size was constricted because of resource limitations). The Radial Basis kernel was used, along with a cost of 10 and cross validation setting of 10. When tested on the remainder of the data set, the SVM was able to correctly predict the NPS type 68% of the time. No permutation of amenities influenced the effectiveness of the model. This led the team to suspect that the model had simply learned that 68% of surveys would be promoters and that the amenities had very little or no impact on the prediction. The results from the prediction are shown below.

*Support Vector Machine object of class "ksvm"   
  
SV type: C-svc (classification)   
 parameter : cost C = 10   
  
Gaussian Radial Basis kernel function.   
 Hyperparameter : sigma = 0.2   
  
Number of Support Vectors : 10277   
  
Objective Function Value : -101622.3   
Training error : 0.309248   
Cross validation error : 0.311379   
[1] "Percent good: 68.780000"*

The second method employed to predict guest promoters was associative rule mining. All 29 of the amenities and the NPS\_Type variable were converted into a “shopping basket” to determine which amenities occurred together when a promoter NPS\_Type was on the right-hand side. The apriori function was used in R to conduct the mining with a support level of 0.25 and a confidence of 0.5. This yielded 17,724 rules, of which the top 5 by lift and confidence are shown below.

*lhs rhs support confidence lift count  
[1] {All\_Suites\_PL,   
 Indoor\_Corridors\_PL} => {NPS\_Type} 0.2776168 0.7246084 1.054112 72913  
[2] {All\_Suites\_PL} => {NPS\_Type} 0.2842495 0.7242925 1.053652 74655  
[3] {Business\_Center\_PL,   
 Indoor\_Corridors\_PL,   
 Shuttle\_Service\_PL} => {NPS\_Type} 0.2667426 0.6917365 1.006292 70057  
[4] {Business\_Center\_PL,   
 Fitness\_Center\_PL,   
 Indoor\_Corridors\_PL,   
 Shuttle\_Service\_PL} => {NPS\_Type} 0.2667426 0.6917365 1.006292 70057  
[5] {Business\_Center\_PL,   
 Elevators\_PL,   
 Indoor\_Corridors\_PL,   
 Shuttle\_Service\_PL} => {NPS\_Type} 0.2649683 0.6914158 1.005826 69591*

Only Hyatt House, Hyatt Place, Hyatt Regency and Hyatt offer property configurations with all suites. It is likely not the focus of other brands to offer only suites (e.g. extended stay). Only Andaz does not offer Shuttle Service. The remaining amenities are available at all Hyatt brands.

### Path Forward

It is the team’s conclusion that further analysis efforts would be better directed elsewhere. This is based on the inability of the team to find correlation between the number of amenities and promoters, the consistency of amenities and promoters, as well as the ineffectiveness of the SVM model. Based on the data contained in the survey results, hotel amenities do not appear to have an impact on the likelihood of guests to recommend Hyatt. Data regarding the quality or condition of the amenities might have an impact, but it was not included in the survey results.

The only recommendation which may be made is to offer a shuttle service at Andaz hotels.

## What are the drivers of promoters to likelihood to recommend?

### Methodology

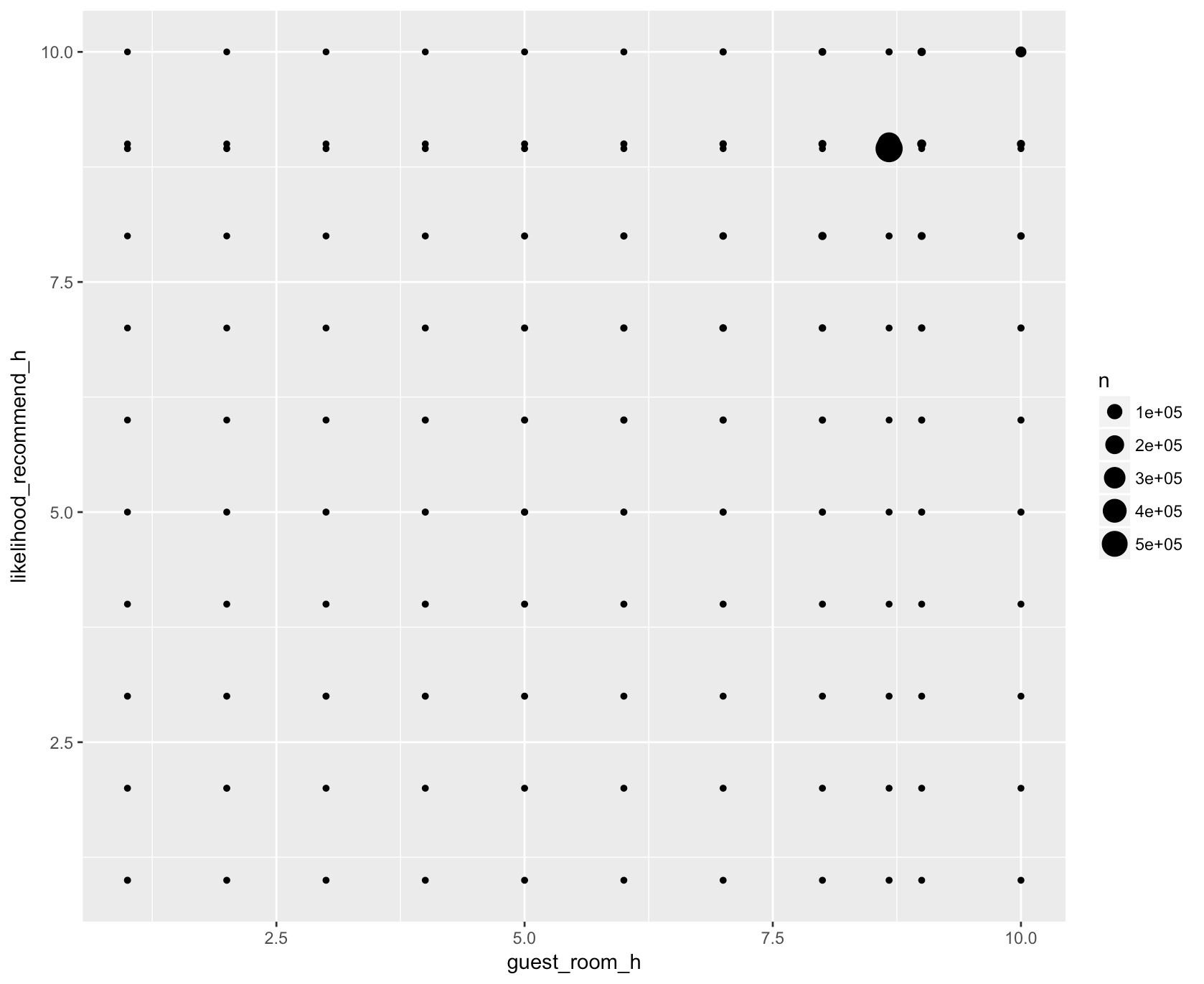
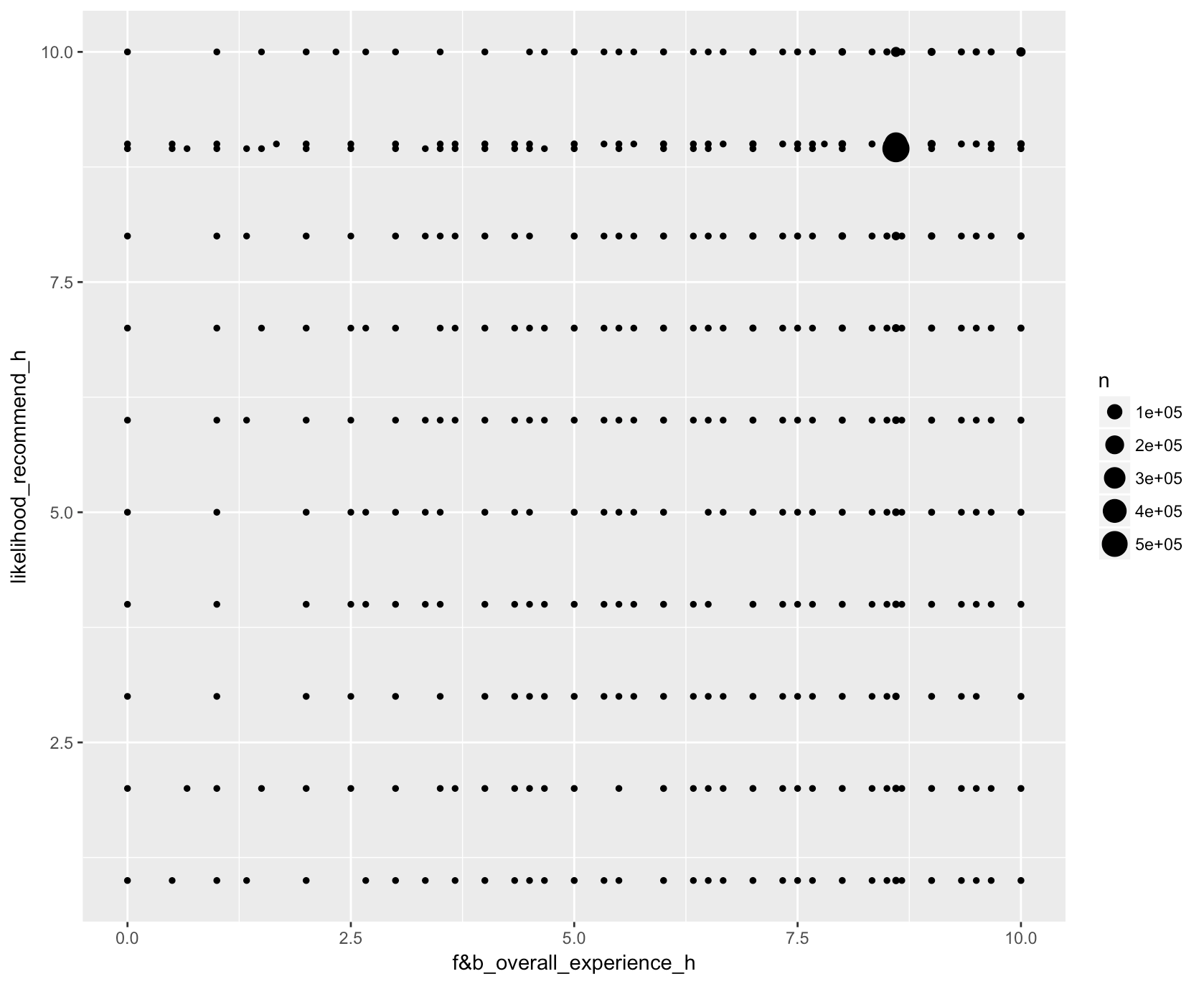
There are several NPS scores that are tracked in any given survey. These scores represent the quality of customer service, the check in process, food and beverage, guest room condition, and the hotel’s overall condition and feel. Each of these aspects are not captured in any other way in the survey regarding a particular stay.

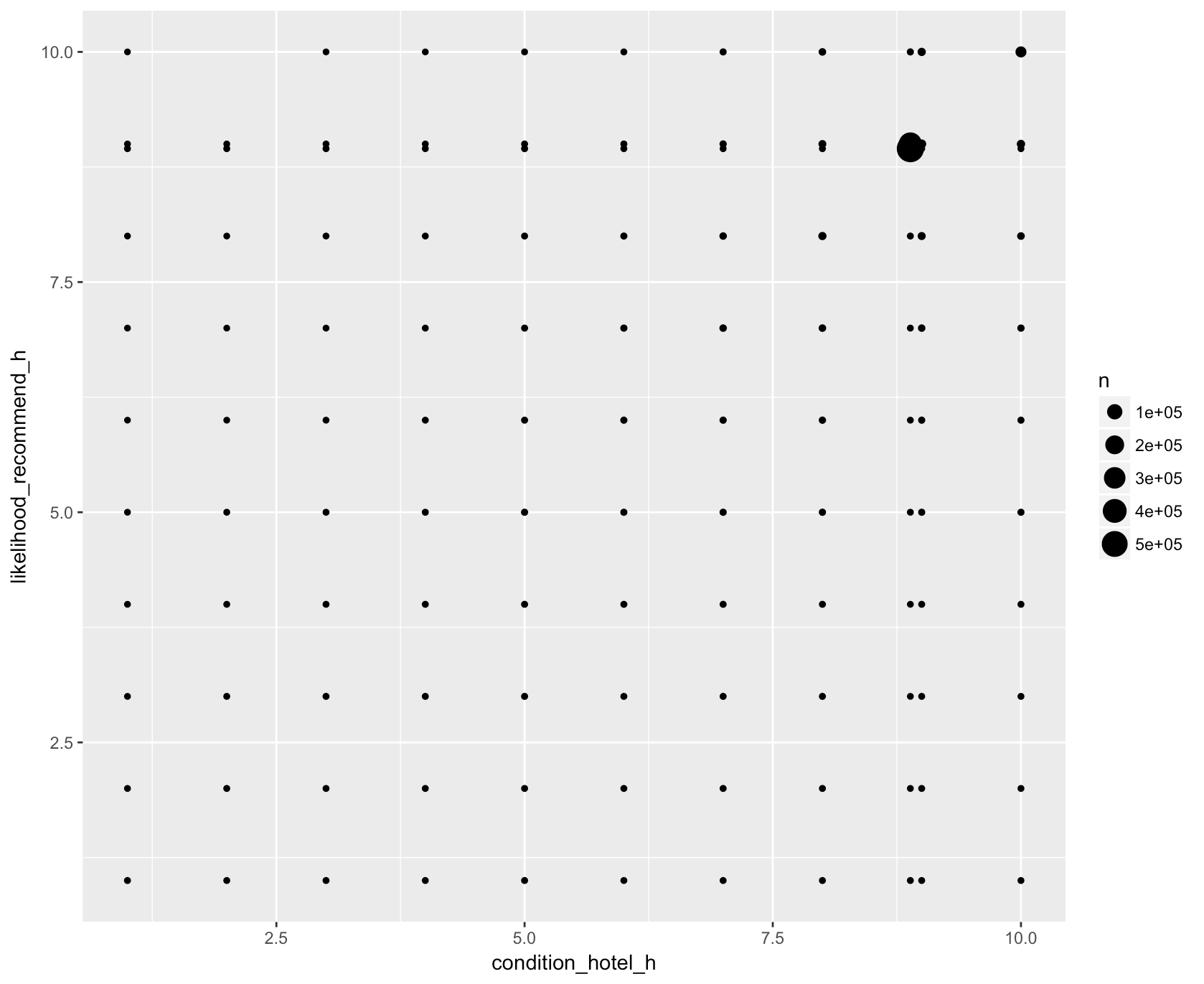
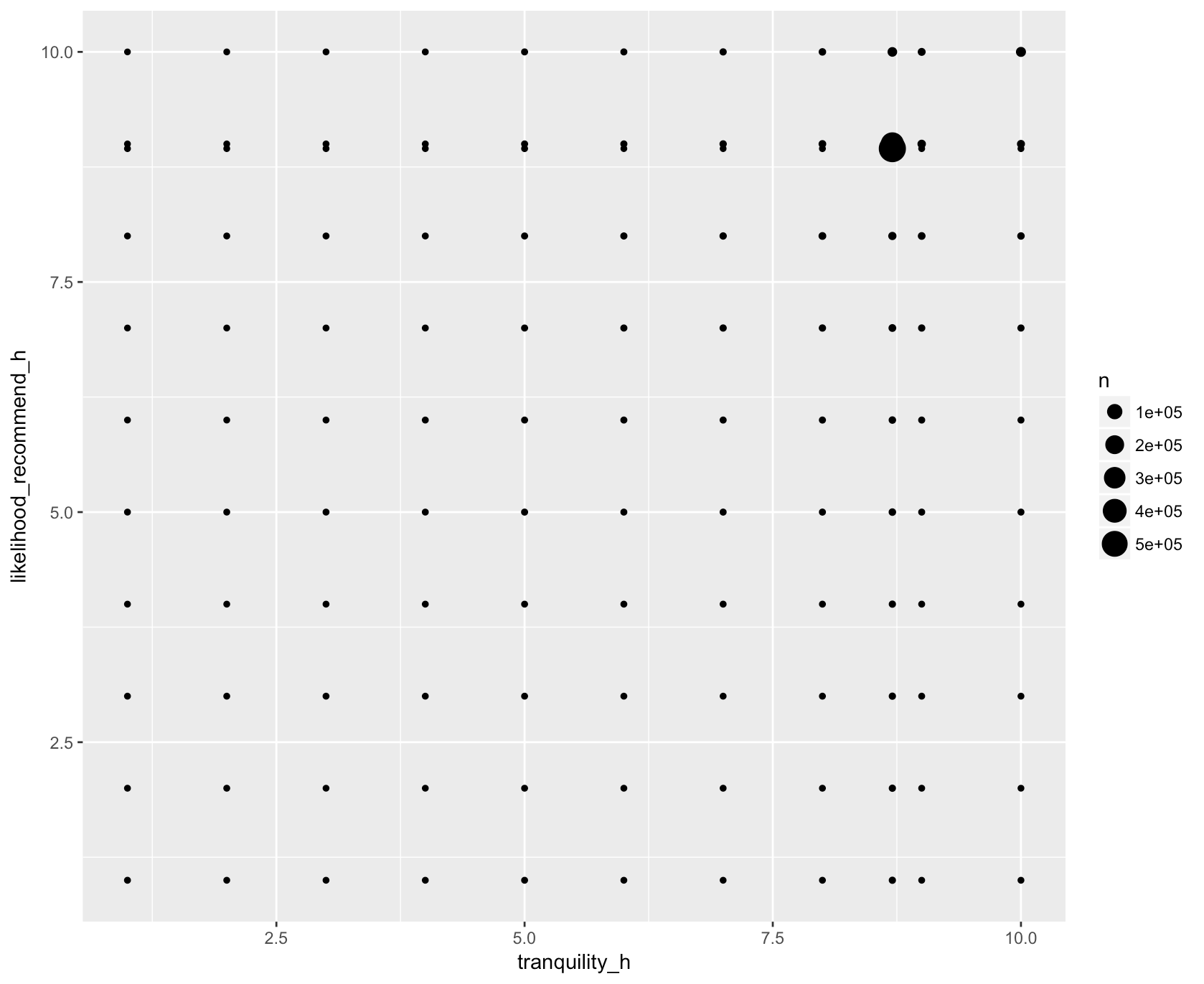
The team’s hypothesis is that one, or a combination, of these aspects largely contributes to the reason for an overall likelihood to recommend. For example, if a guest had particularly poor customer service, they are unlikely to recommend, etc. The team included the nightly rate as a way to control for expectations for each of these values, assuming that those who pay more for a room are more likely to expect higher levels of service and quality. The attributes will assess the following variables’ correlation to the likelihood\_recommend\_h score:

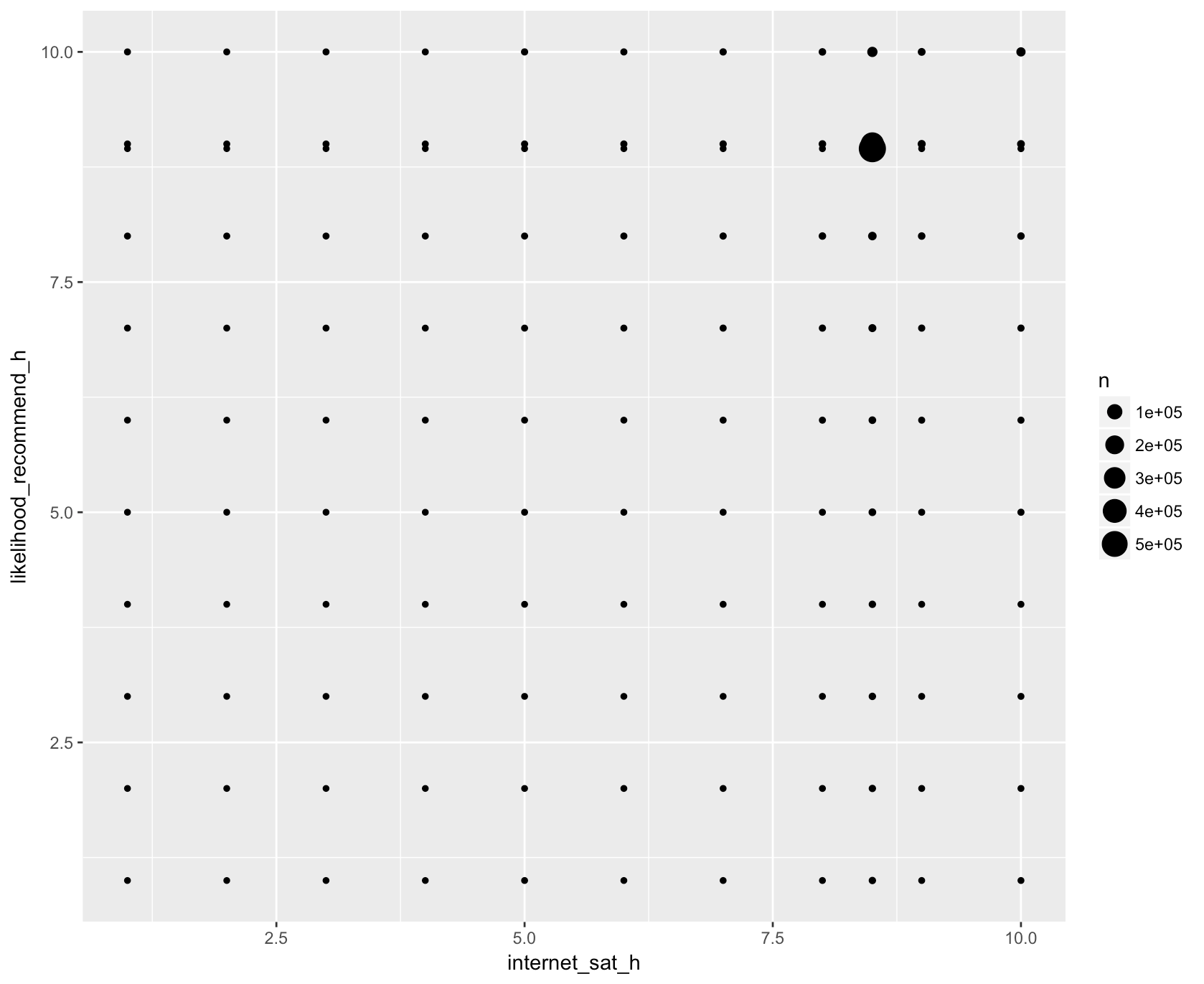
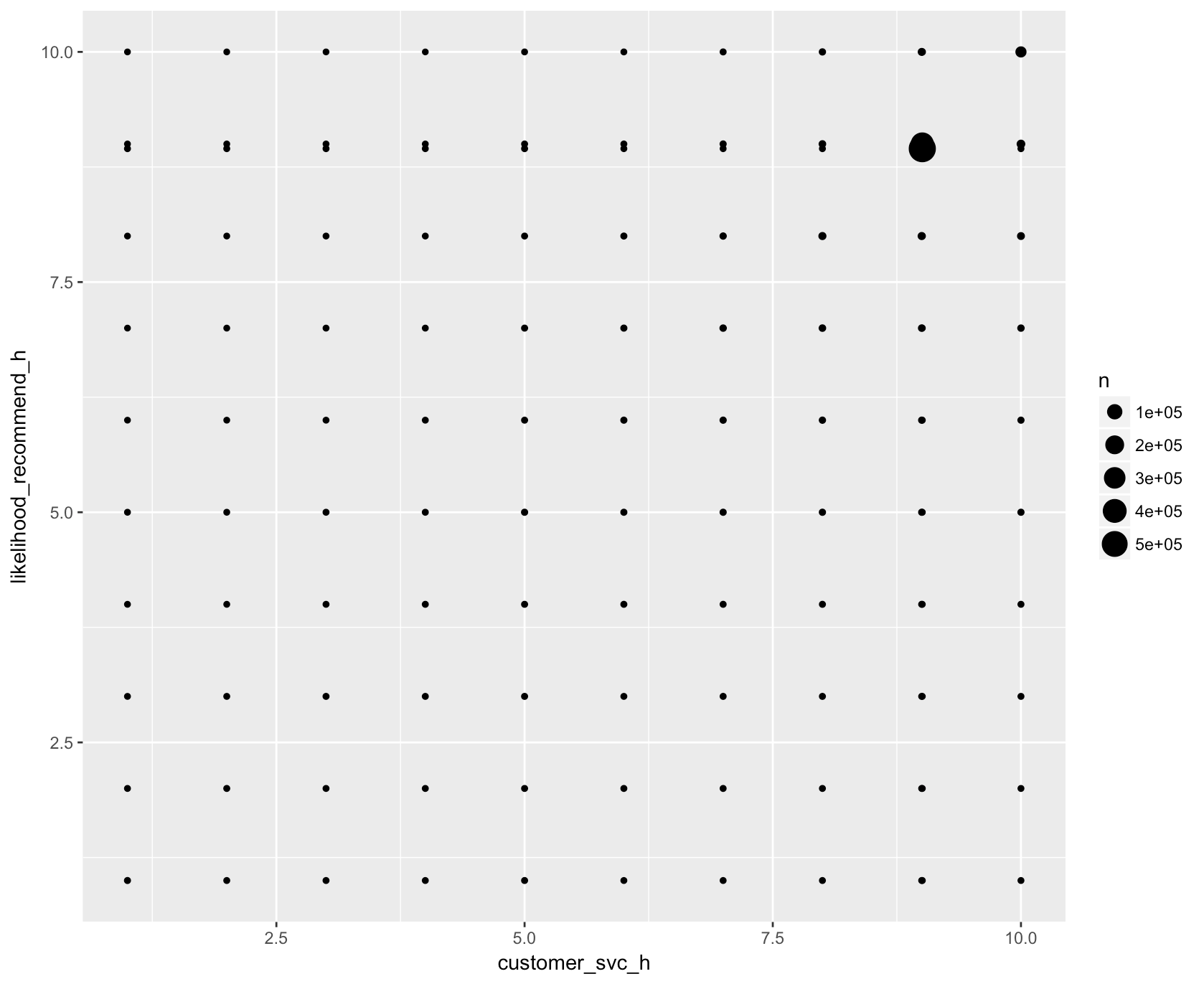
* guest\_room\_h
* tranquility\_h
* f&b\_overall\_experience\_h
* staff\_cared\_h
* condition\_hotel\_h
* internet\_sat\_h
* customer\_svc\_h
* nt\_rate\_r

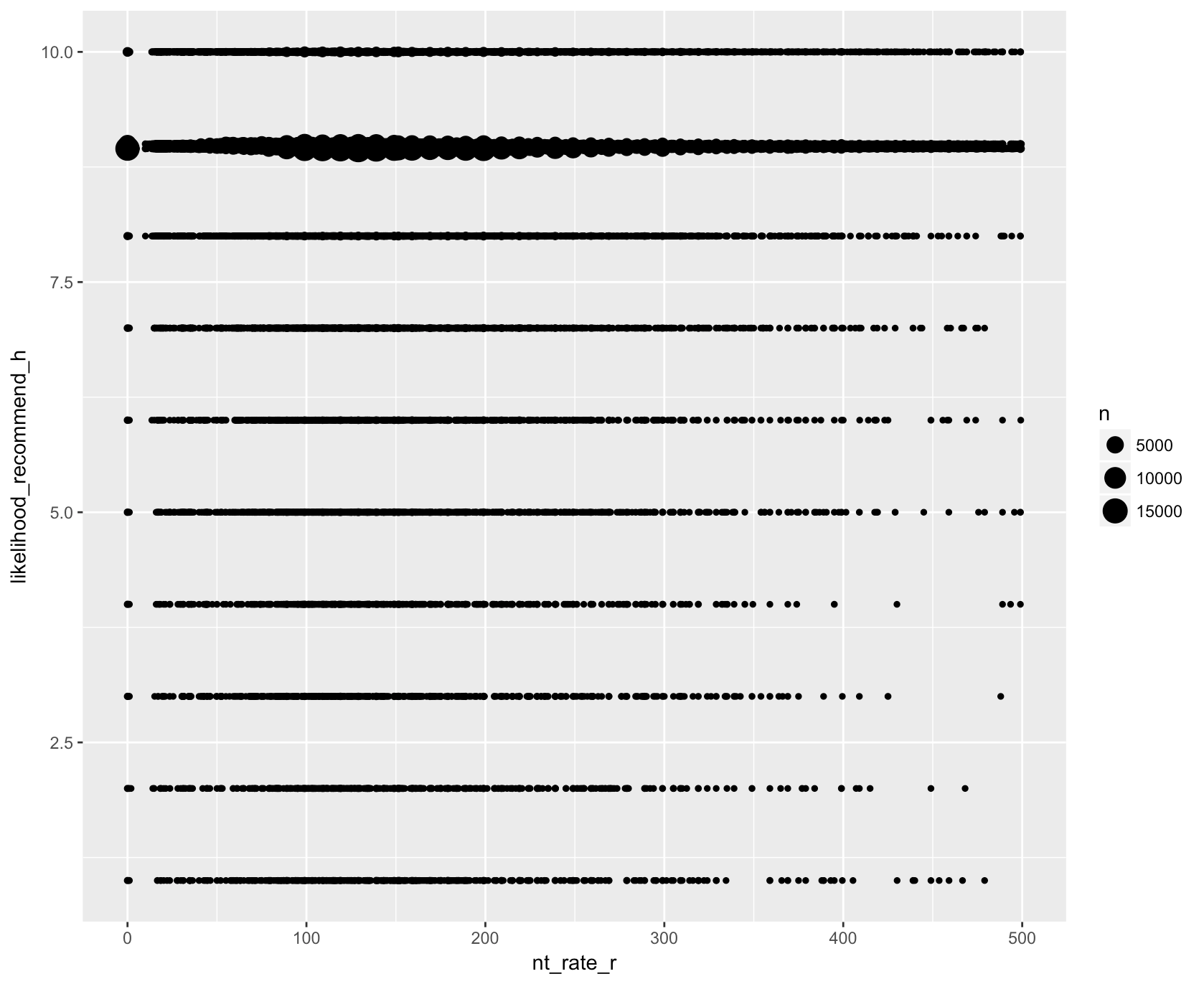
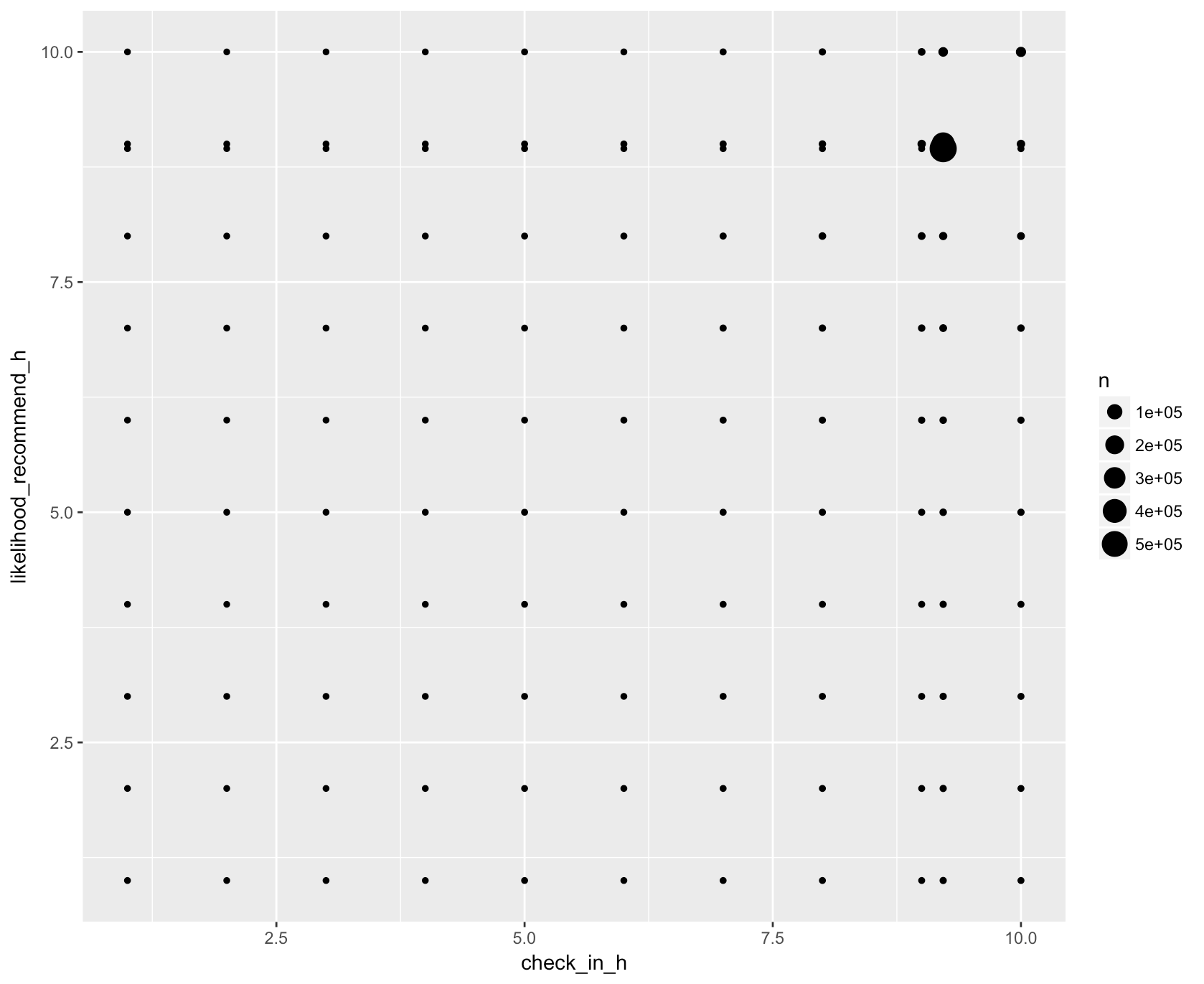
### Statistical Analysis

To deal with NAs within the data, all NAs of these variables were replaced with the mean of the variable. First each of the variables was graphed by the likelihood\_recommend\_h score to identify any apparent patterns in the data:









Looking at the data, there is no significant patterns to note, except that the mean of each of the scores lands relatively close to the same place regardless of the X variable.

Then, a linear model was fit to determine if there were any significant (or insignificant) predictors for the likelihood\_recommend\_h:

|  |  |
| --- | --- |
| **Predictor Name** | **p-value** |
| Intercept | <2e-16 \*\*\* |
| f&b\_overall\_experience\_h | <2e-16 \*\*\* |
| guest\_room\_h | <2e-16 \*\*\* |
| tranquility\_h | <2e-16 \*\*\* |
| condition\_hotel\_h | <2e-16 \*\*\* |
| customer\_svc\_h | <2e-16 \*\*\* |
| staff\_cared\_h | <2e-16 \*\*\* |
| internet\_sat\_h | <2e-16 \*\*\* |
| check\_in\_h | 0.481 |
| nt\_rate\_r | <2e-16 \*\*\* |

This model has Adjusted R-squared value of 0.5901. This means that 59% of the variance can be explained by this model. This full model clearly shows that the check in process is not a driver of the overall likelihood to recommend NPS score.

Finally, the team determined the most parsimonious model (the best fit model with the least amount of predictors) by using the step() function in R.

This final model had an AIC of -8685591 with the check\_in\_h predictor removed because it was not significant. The Adjusted R-Squared value did not change however. Here are the coefficients:

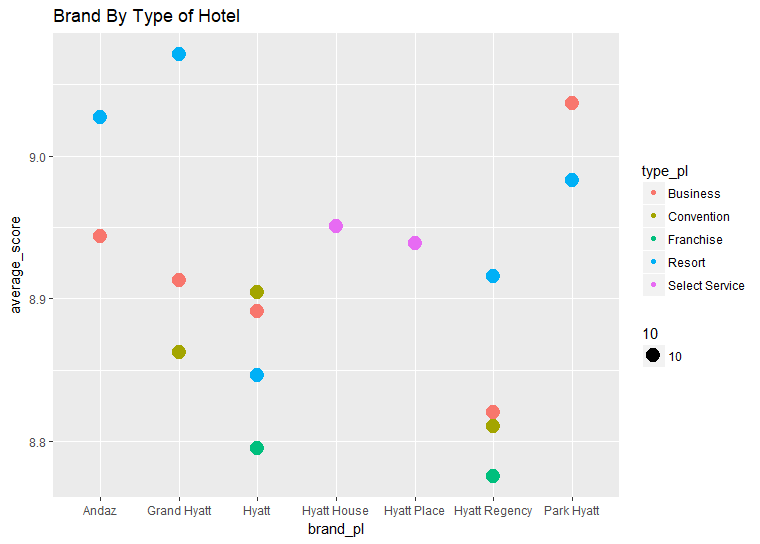
|  |  |
| --- | --- |
| **Predictor Name** | **Coefficient** |
| Intercept | -1.621e+00 |
| f&b\_overall\_experience\_h | 7.523e-02 |
| guest\_room\_h | 3.386e-01 |
| tranquility\_h | 8.318e-02 |
| condition\_hotel\_h | 2.357e-01 |
| customer\_svc\_h | 3.586e-01 |
| staff\_cared\_h | 7.992e-02 |
| internet\_sat\_h | 2.756e-02 |
| nt\_rate\_r | -6.085e-05 |

While this model does not have great predictive power, it was able to establish slight correlation between all of these aspects and the guest’s overall stay. Meanwhile, more surprisingly, the check in process had no significant correlation to the guest’s overall stay. This could warrant further analysis at a later date. Based on the low coefficients and the average adjusted R-Squared value, as well as a lack of pattern in the data, the team would not be confident in drawing any causation theories.

## What location has the highest NPS score per hotel type per region? What factors contribute to this score?

### Methodology

Across the Hyatt brands there exist five different business models which are labeled as “type\_pl” in the data. The team decided that separating the hotels by type would make for a better comparison among the hotels when deciding which businesses should copy each other to maximize their NPS scores. The survey lists these types as: Business, Convention, Franchise, Resort, and Select Service. Most hotels fall into multiple categories depending on their location. The below graph shows which brands fall into the five categories for hotel type along with the hotel’s average score by brand-type combination..



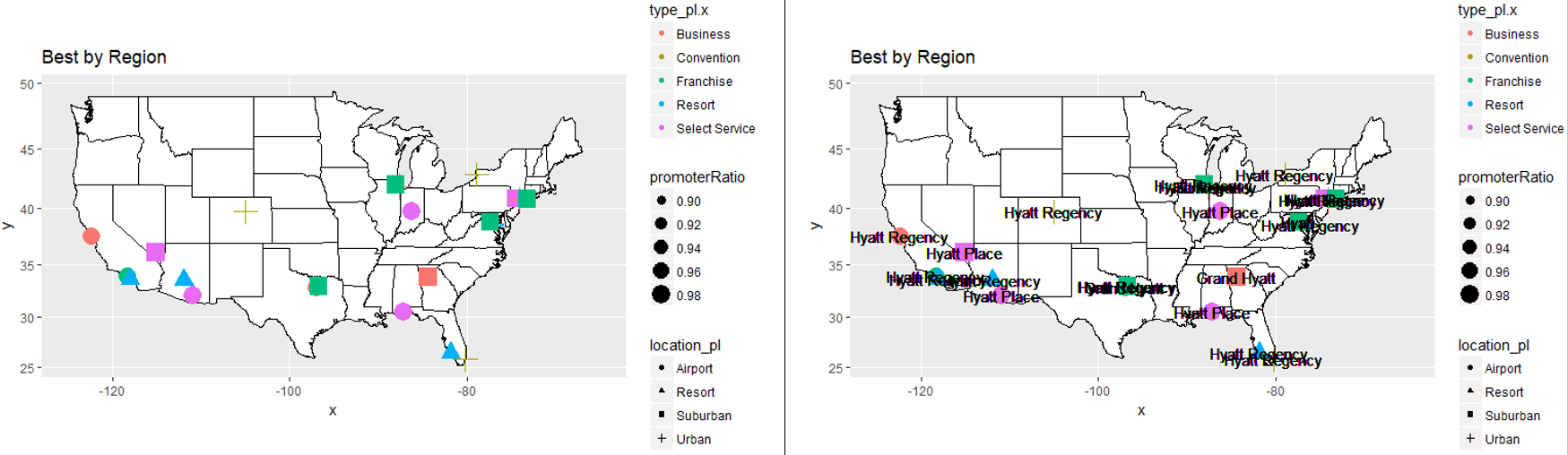
Furthermore, to account for location, the team looked at which of these hotels stood above their peers per business type within each US region. The regions are broken up into the five following categories: Northeast, Southeast, Midwest, Southwest, West.

To decide which hotels stood out numerically, a few factors common to most of the hotels were taken into account. These factors are: number of amenities per hotel, ratio of promoters to completed surveys per hotel, and the average NPS score for each hotel.

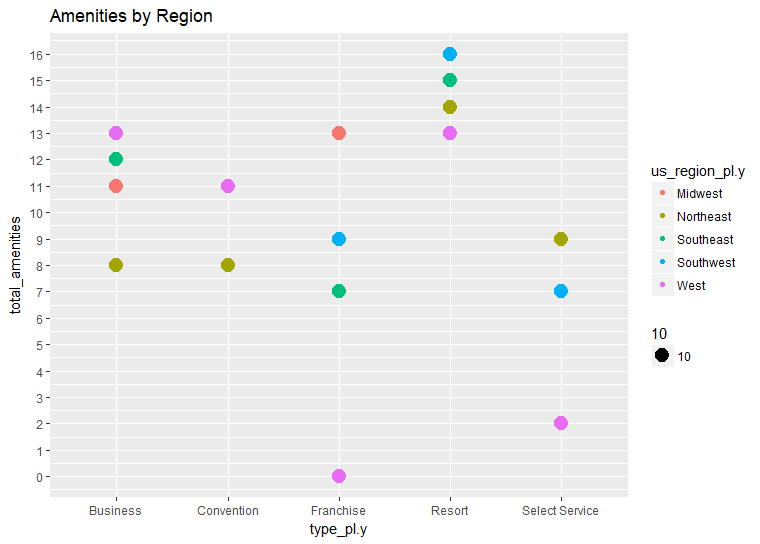
Once these three categories were figured out, and the correct hotels were collected into a data frame, the team began a statistical analysis of the data.

### Statistical Analysis

Of the 395 individual hotels surveyed from the United States, 24 stand out as the most highly promoted Hyatt hotels in their respective regions. This status was measured by dividing the amount of surveys completed at each location by the amount of promoters per location to get a ratio of promoters per location. Finding the locations with the highest ratio outcome provides a countrywide picture of the types, brands, and locations that work together the best to positively affect the NPS score at each hotel.



The above graphs represent the top hotels per type for the five regions of the United States. In the first map, we can see that all of the types of hotels have a varying promoter ratio, and were built in a wide range of communities that include airport, resort, suburban, and urban locations. In the second map, it becomes apparent that Hyatt Regency and Hyatt Place stand out as the highest ranked brand in every region regardless of type. Using this data, we decided to look at how many amenities the top hotels in each region have per type of hotel:



According to the above graph, all of the top hotels per region have more than one of the selected amenities from business question 2. The only option that didn’t have any of the expected amenities at the time this survey was conducted is located at the Los Angeles airport. This graph provides insight into what types of amenities are expected at each brand of the Hyatt’s hotels in each region. Although we discovered that amenities may not play a direct role in raising NPS score, it seems that high scoring luxury brand hotels such as those found in the resort and business categories tend to have more amenities for entertaining guests, while select service and franchise locations have a much wider spread depending on regional location. Highly rated convention hotels tend to have a moderate to high amount of amenities depending on location as well.

# Recommendations

1. Focus on marketing to businesses, not individuals.
2. There should be at least 3 amenities from this list at every hotel.
3. The customer service, hotel condition, guest room, and hotel atmosphere all appear to have similar scores as to the overall likelihood to recommend, with none being the ultimate driver.
4. To compete as a top hotel in each region according to hotel type, the different brands should try to follow the example of including a certain number of amenities depending on location and hotel type. In some cases they should include more of these amenities depending on brand, type, and region.

# Appendix A: Reproducible R Code

################### PUT ALL PRODUCTION CODE HERE #########################  
  
################################

############ DATA ARCHITECTURE  
################################

library(dplyr)

library(stringr)

library(logging)

library(ggplot2)

library(lubridate)

library(ggmap)

library(maps)

library(mapproj)

library(plotly)

library(reshape2)

library(sqldf)

### load data

load('~/Downloads/nps\_subset.Rdata')

myData <- DT\_exp\_final\_set

### make column names lowercase

names(myData) <- tolower(names(myData))

## replace spaces and hyphens in column names with underscores

names(myData) <- gsub('-| ', '\_', names(myData))

### transform columns to dates

myData$check\_in\_date\_c <- as.Date(myData$check\_in\_date\_c)

myData$check\_out\_date\_c <- as.Date(myData$check\_out\_date\_c)

myData$arrival\_date\_r <- as.Date(myData$arrival\_date\_r)

myData$departure\_date\_r <- as.Date(myData$departure\_date\_r)

myData$e\_delivereddate\_i <- as.Date(myData$e\_delivereddate\_i)

myData$e\_delivereddate\_adj\_i <- as.Date(myData$e\_delivereddate\_adj\_i)

myData$response\_date\_h[which(myData$response\_date\_h == '')] <- NA

myData$response\_date\_h <- as.Date(myData$response\_date\_h)

myData$guest\_checkin\_date\_h[which(myData$guest\_checkin\_date\_h == '')] <- NA

myData$guest\_checkin\_date\_h <- as.Date(myData$guest\_checkin\_date\_h)

myData$guest\_checkout\_date\_h[which(myData$guest\_checkout\_date\_h == '')] <- NA

myData$guest\_checkout\_date\_h <- as.Date(myData$guest\_checkout\_date\_h)

myData$eff\_date\_cc <- as.Date(myData$eff\_date\_cc)

### create master columns

myData$m\_check\_in\_date <- coalesce(myData$check\_in\_date\_c, myData$arrival\_date\_r, myData$guest\_checkin\_date\_h, myData$eff\_date\_cc)

myData$m\_check\_out\_date <- coalesce(myData$check\_out\_date\_c, myData$departure\_date\_r, myData$guest\_checkout\_date\_h)

myData$m\_length\_of\_stay <- coalesce(myData$length\_of\_stay\_c, myData$length\_stay\_h)

myData$m\_pov\_code <- coalesce(myData$pov\_code\_c, myData$pov\_h)

myData$m\_guest\_country <- coalesce(myData$guest\_country\_r, myData$e\_country\_i, myData$guest\_country\_h)

myData$m\_guest\_gender <- coalesce(myData$e\_hy\_gss\_gender\_i, myData$gender\_h)

myData$m\_survey\_status <- coalesce(myData$status\_h, myData$e\_status\_i)

## filter to just 2014 data

myData <- myData[year(myData$m\_check\_in\_date) == '2014',]

## remove columns included in master cols

myData <- myData %>%

select(-check\_in\_date\_c, -arrival\_date\_r, -guest\_checkin\_date\_h,

-check\_out\_date\_c, -departure\_date\_r, -guest\_checkout\_date\_h,

-length\_of\_stay\_c, -length\_stay\_h, -pov\_code\_c, -pov\_h,

-guest\_country\_r, -e\_country\_i, -guest\_country\_h,

-member\_status\_r, -e\_hy\_gss\_tier\_i, -gp\_tier\_h, - eff\_date\_cc,

-status\_h, -e\_status\_i, -e\_hy\_gss\_gender\_i, -gender\_h)

######################################################

############ DATA TRANSFORMATIONS/MUNGING/CLEANING  
#######################################################

## fix strings

myData$goldpassport\_flg\_r <- gsub('gp ', '', tolower(myData$goldpassport\_flg\_r))

### fix guest state format

myData$guest\_state\_h[grepl('[[:punct:]]| |[0-9]', toupper(myData$guest\_state\_h))] <- NA

myData$guest\_state\_h[which(myData$guest\_state\_h == '')] <- NA

### only keep US states

myData$guest\_state\_h[which(!(myData$guest\_state\_h %in% c('AL', 'AK', 'AZ', 'AR', 'CA', 'CO','CT', 'DE',

'FL', 'GA', 'HI', 'ID', 'IL','IN', 'IA','KS','KY',

'LA','ME','MD', 'MA','MI', 'MN', 'MS', 'MO', 'MT',

'NE', 'NV','NH','NJ','NM','NY','NC','ND','OH','OK',

'OR','PA','RI','SC','SD','TN','TX','UT','VT','VA',

'WA','WV','WI','WY','DC')))] <- NA

## set all entries with US states to US country

myData$m\_guest\_country[which(is.na(myData$guest\_state\_h) == FALSE)] <- 'UNITED STATES'

## set bad/unknown values to NA

myData$m\_guest\_country[grepl('[0-9]|^[A-Z]{2}$', myData$m\_guest\_country)] <- NA

myData$m\_guest\_country[which(myData$m\_guest\_country %in% c('','$CO','MIA','NEW','MSP','XZJ'))] <- NA

myData$gp\_tier <- tolower(myData$gp\_tier)

myData$gp\_tier[which(myData$gp\_tier == 'ldia')] <- 'lifetime diamond'

myData$gp\_tier[which(myData$gp\_tier == 'diam')] <- 'diamond'

myData$gp\_tier[which(myData$gp\_tier == 'plat')] <- 'platinum'

myData$gp\_tier[which(myData$gp\_tier == 'card')] <- 'courtesy'

myData$gp\_tier[which(myData$gp\_tier == '')] <- NA

#set state names to lowercase

myData$state\_pl <- tolower(myData$state\_pl)

#Change likelihood\_recommend\_h to numeric

myData$likelihood\_recommend\_h <- as.numeric(myData$likelihood\_recommend\_h)

myData$nps\_type[which(myData$nps\_type == '')] <- NA

#Repeat customers are de facto Promoters for NPS type and de facto 9 for blank likelihood\_recommend\_h. The following code converts any blanks for repeat customers to promoters.

#data frame for repeat customers with likelihood\_recommend-h and nps\_type

myData <- myData %>%

group\_by(cons\_guest\_id\_c) %>%

mutate(n = n(),

nps\_type = ifelse((n > 1) & is.na(nps\_type), "Promoter", nps\_type),

likelihood\_recommend\_h = ifelse((n > 1) & is.na(likelihood\_recommend\_h), 9, likelihood\_recommend\_h)) %>%

select(-n) %>%

ungroup()

####################################

############ DESCRIPTIVE STATISTICS  
####################################

myLocations <- myData %>%

select(property\_id\_pl, property\_longitude\_pl,

property\_latitude\_pl, hotel\_name\_long\_pl) %>%

group\_by(property\_id\_pl) %>%

mutate(number\_of\_surveys = n()) %>%

distinct()

# getting the map

us\_map2 <- get\_map("united states", zoom = 4)

# plotting the map with some points on it

ggmap(us\_map2) +

geom\_point(data = myLocations, aes(x = property\_longitude\_pl,

y = property\_latitude\_pl,

size = number\_of\_surveys))

myAvg <- myData %>%

select(m\_check\_in\_date, likelihood\_recommend\_h, brand\_pl) %>%

filter(is.na(likelihood\_recommend\_h) == FALSE) %>%

group\_by(m\_check\_in\_date, brand\_pl) %>%

mutate(nps = mean(likelihood\_recommend\_h)) %>%

select(-likelihood\_recommend\_h) %>%

distinct()

ggplot(data = myAvg, aes(x = m\_check\_in\_date, y = nps, colour = brand\_pl)) +

geom\_line()

myCustomers <- myData %>%

select(cons\_guest\_id\_c, gp\_tier, m\_length\_of\_stay) %>%

filter(is.na(cons\_guest\_id\_c) == FALSE) %>%

group\_by(cons\_guest\_id\_c) %>%

mutate(number\_of\_stays = n(),

avg\_length\_of\_stay = mean(m\_length\_of\_stay)) %>%

select(-m\_length\_of\_stay) %>%

distinct()

myAvgTier <- myCustomers %>%

group\_by(gp\_tier) %>%

mutate(avg\_stay = mean(number\_of\_stays)) %>%

select(-cons\_guest\_id\_c) %>%

distinct()

ggplot(myCustomers, aes(x = gp\_tier, y = avg\_length\_of\_stay)) +

geom\_boxplot()

ggplot(myCustomers, aes(x = gp\_tier, y = number\_of\_stays)) +

geom\_boxplot()

###################################

############ BUSINESS QUESTION 1  
###################################

#check total promoters before considering repeat customers

totalPromoters <- length(myData$cons\_guest\_id\_c[which(myData$likelihood\_recommend\_h >= 9)])

totalPromoters

#find repeat customers

myData$checkDupes <- (duplicated(myData$cons\_guest\_id\_c, incomparables = FALSE))

#enter the value of FALSE in the checkDupes column for all NA guest id values that show up as repeat customers.

myData$checkDupes <- ifelse(is.na(myData$cons\_guest\_id\_c) & myData$checkDupes == TRUE

,myData$checkDupes == FALSE

,myData$checkDupes)

#variable to identify where likelihood to recommend = NA

emptyLikelihood <- is.na(myData$likelihood\_recommend\_h)

#set repeat customer likelihood\_recommend\_h to 9 if blank

myData$likelihood\_recommend\_h <- ifelse(myData$checkDupes == TRUE & emptyLikelihood, 9, myData$likelihood\_recommend\_h)

#check total promoters after considering repeat customers

totalPromoters

#Brands graphed by amount of promoters

hotelByBrand <- ggplot(hotelRatings, aes(x = rownames(hotelRatings), y = hotelRatings$TRUE.))

hotelByBrand <- hotelByBrand + geom\_point(aes(color = rownames(hotelRatings), size = 5))

hotelByBrand <- hotelByBrand + labs(x = 'Brands', y = 'Promoters')

hotelByBrand <- hotelByBrand + ggtitle("Promoters by Brand")

hotelByBrand

# Find Top Customer

#--

fTopCustomer <- function (pimyDataSet, pinoOfCustomer) {

noOfCustomer <- pinoOfCustomer

MyCompletedSurvey <- pimyDataSet[pimyDataSet$m\_survey\_status=="COMPLETED",]

MyCompletedSurvey <- sqldf ("select cons\_guest\_id\_c , count(1) surveycount

from MyCompletedSurvey

where cons\_guest\_id\_c is not null

and cons\_guest\_id\_c > 0

group by cons\_guest\_id\_c

order by 2 desc" )

myTopCustomer <- subset(myData, cons\_guest\_id\_c %in% head(MyCompletedSurvey$cons\_guest\_id\_c,n=pinoOfCustomer))

return (myTopCustomer)

}

############ TOP CUSTOMER INFORMATION

################################

#-- Top Customer Call

myTopCustomer <- fTopCustomer ( myData , 25 )

#count business and leisure stays per top 25 customers

topCustomer.pov <- myTopCustomer %>%

group\_by(cons\_guest\_id\_c) %>%

summarise(business = length(m\_pov\_code[which(m\_pov\_code == 'BUSINESS')])

,leisure = length(m\_pov\_code[which(m\_pov\_code == 'LEISURE')]))

overall.pov <- myData %>%

group\_by(brand\_pl) %>%

summarise(business = length(m\_pov\_code[which(m\_pov\_code == 'BUSINESS')])

,leisure = length(m\_pov\_code[which(m\_pov\_code == 'LEISURE')]))

#Assign variable to get hotels listed by recommendation > 8

hotelRatings <- tapply(myData$likelihood\_recommend\_h,list(myData$brand\_pl,myData$likelihood\_recommend\_h > 8), length)

hotelRatings

#convert top customer and overall customer data to plot together on bar graphs

topCust.long<-melt(topCustomer.pov,id.vars="cons\_guest\_id\_c")

overall.long <- melt(overall.pov,id.vars="brand\_pl")

#graph top 25 customer leisure vs business

topCustomer.pov\_g <- ggplot(topCust.long, aes(x = as.character(cons\_guest\_id\_c), y = value, fill = factor(variable)))

topCustomer.pov\_g <- topCustomer.pov\_g + geom\_bar(stat = "identity", position = "dodge") +

scale\_y\_continuous(breaks = round(seq(min(topCust.long$value), max(topCust.long$value), by = 1)))

topCustomer.pov\_g <- topCustomer.pov\_g + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

topCustomer.pov\_g <- topCustomer.pov\_g + ggtitle("Top 25 Customer Purpose of Visit")

topCustomer.pov\_g

#graph top 25 customer leisure vs business

overall.pov\_g <- ggplot(overall.long, aes(x = brand\_pl, y = value, fill = factor(variable)))

overall.pov\_g <- overall.pov\_g + geom\_bar(stat = "identity", position = "dodge") +

scale\_y\_continuous(breaks = round(seq(min(0), max(overall.long$value), by = 100000)))

overall.pov\_g <- overall.pov\_g + theme(axis.text.x = element\_text(angle = 90, hjust = 1))

overall.pov\_g <- overall.pov\_g + ggtitle("Overall Business vs. Leisure Travel")

overall.pov\_g

#mean likelihood to recommend per top customer

topCustomer.avgRating <- tapply(myTopCustomer$likelihood\_recommend\_h, myTopCustomer$cons\_guest\_id\_c, mean, na.rm = TRUE)

topCustomer.avgRating <- data.frame(topCustomer.avgRating)

#state of the top customers

topCustomer.state <- tapply(myTopCustomer$state\_r, myTopCustomer$cons\_guest\_id\_c, unique)

# Scatter Plot

mySummary <- myData %>%

group\_by(brand\_pl) %>%

summarise(guest\_count=length(cons\_guest\_id\_c)

,completedcount=sum(ifelse(m\_survey\_status=="COMPLETED" , 1, 0)))

ggplot(mySummary, aes(x=guest\_count, y=completedcount)) + geom\_point(aes(size=completedcount, colour=brand\_pl))

rm(mySummary)

############ MAPPING TOP CUSTOMERS

################################

#focus only on top 5 customers

myTopFiveCustomer <- fTopCustomer ( myData , 5)

# Function to generate basic simple US map as starting point

fmyUSMap <- function () {

us <- map\_data("state")

g <- ggplot ( us , aes(map\_id=region) )

g <- g + geom\_map( map=us, fill="white", color="black")

g <- g + expand\_limits(x=us$long, y=us$lat)

g <- g + coord\_map() + ggtitle("Hotels Frequented by Top Customers")

return (g)

}

# Plot Top Customers

myTopFiveCustomer$region <- tolower(myTopFiveCustomer$state\_pl)

g <- fmyUSMap()

g <- g + geom\_point(data=myTopFiveCustomer

, aes(x=property\_longitude\_pl

,y=property\_latitude\_pl

,size=likelihood\_recommend\_h

,colour=as.character(cons\_guest\_id\_c)))

g

############ COMPARE COUNTRY REVENUE

################################

#Get mean revenue by country

medianRevenueByCountry <- sqldf("select m\_guest\_country,

mean(pms\_total\_rev\_c) as 'mean\_revenue'

median(pms\_total\_rev\_c) as 'median\_revenue'

from myData

group by m\_guest\_country

order by avg(pms\_total\_rev\_c) desc")

#stays per country

staysPerCountry <- myData %>%

group\_by(m\_guest\_country) %>%

summarise(stays\_by\_country = length(m\_guest\_country))

#Get mean revenue by state

sqldf("select guest\_state\_h, avg(pms\_total\_rev\_c) as 'mean\_revenue'

from myData

group by guest\_state\_h

order by avg(pms\_total\_rev\_c) desc")

#Get mean revenue by member status

sqldf("select gp\_tier, avg(pms\_total\_rev\_c) as 'mean\_revenue'

from myData

group by gp\_tier

order by avg(pms\_total\_rev\_c) desc")

###################################

############ BUSINESS QUESTION 2  
###################################

######################################

# amenitiesClean.R

# Clean and prepare the amenities portion

# of the Hyatt data set. This file should

# be sourced first!!!

# The tidyverse is a requirement.

######################################

library(tidyverse)

# Remove columns that are not needed for analyzing amenities

cleanAmenities <- function(amenitiesDF, dfType) {

# Remove spaces, dashes and ampersands in column names, they cause parsing errors

cols <- colnames(amenitiesDF)

cols <- gsub(" ","\_", cols)

cols <- gsub("&", "and", cols)

colnames(amenitiesDF) <- gsub("-","\_", cols)

# Add surveys where the guests gave an overall sat score

# but did not provide a likelihood to recommend score

amenities <- amenitiesDF %>%

mutate(NPS\_Type = replace(NPS\_Type, NPS\_Type == "" & Overall\_Sat\_H > 8, "Promoter"))

amenities <- amenities %>%

mutate(NPS\_Type = replace(NPS\_Type, NPS\_Type == "" & Overall\_Sat\_H < 9, "Detractor"))

# Required column lists

if(dfType == "full") {

# Include subset which will allow comparison with user feedback

amenities <- amenities %>%

select(CONS\_GUEST\_ID\_C, ARRIVAL\_DATE\_R, POV\_H, Likelihood\_Recommend\_H,

Overall\_Sat\_H, Guest\_Room\_H, Tranquility\_H, Condition\_Hotel\_H,

Customer\_SVC\_H, Staff\_Cared\_H, Internet\_Sat\_H, Check\_In\_H, FandB\_FREQ\_H,

FandB\_Overall\_Experience\_H, Brand\_PL, All\_Suites\_PL:Valet\_Parking\_PL, GP\_Tier, NPS\_Type)

} else if(dfType == "nps") {

# NPS type and amenities, remove surveys where an NPS type

# does not exist (no likelihood to recommend value)

amenities <- amenities %>%

select(NPS\_Type, All\_Suites\_PL:Valet\_Parking\_PL) %>%

filter(NPS\_Type != "")

# Change Y's to 1 and N's to 0

amenities <- amenities %>%

mutate\_at(.vars = vars(All\_Suites\_PL:Valet\_Parking\_PL),

.funs = funs(ifelse(. == "Y", 1, 0)))

# Change Promoter to 1 and Neutral/Detractor to 0

# Change NPS to promoter = 1, everything else 0

amenities <- amenities %>%

mutate\_at(.vars = vars(NPS\_Type),

.funs = funs(ifelse(. == "Promoter", 1, 0)))

# Do not convert to factors

} else if(dfType == "nps.factor") {

# NPS type and amenities, remove surveys where an NPS type

# does not exist (no likelihood to recommend value)

amenities <- amenities %>%

select(NPS\_Type, All\_Suites\_PL:Valet\_Parking\_PL) %>%

filter(NPS\_Type != "")

# Change Y's to 1 and N's to 0

amenities <- amenities %>%

mutate\_at(.vars = vars(All\_Suites\_PL:Valet\_Parking\_PL),

.funs = funs(ifelse(. == "Y", 1, 0)))

# Change Promoter to 1 and Neutral/Detractor to 0

# Change NPS to promoter = 1, everything else 0

amenities <- amenities %>%

mutate\_at(.vars = vars(NPS\_Type),

.funs = funs(ifelse(. == "Promoter", 1, 0)))

# Convert to factors

cols <- colnames(amenities)

amenities <- amenities %>%

mutate\_all(.funs = funs(factor(.)))

} else if(dfType == "brand") {

# Include the brand of the hotel and amenities

amenities <- amenities %>%

select(Property\_ID\_PL, Brand\_PL, All\_Suites\_PL:Valet\_Parking\_PL)

# Return only unique properties

amenities <- amenities %>%

unique()

} else {

# Otherwise, send back everything

amenities <- amenities %>%

select(CONS\_GUEST\_ID\_C:NPS\_Type)

}

return(amenities)

}

######################################

# amenitiesBrand.R

# Contains functions perform some descriptive

# statistics specifically for amenities.

######################################

# Property counts by brand

propCounts <- function(amenitiesArg) {

# First, clean up the data set

amenities <- cleanAmenities(amenitiesArg, "brand")

# Get a count of the number of properties for each brand,

# will be used to calculate percentages later

countAmen <- amenities %>%

select(Property\_ID\_PL, Brand\_PL) %>%

group\_by(Brand\_PL) %>%

summarise(Count = n())

# Convert to a data frame

countAmen <- countAmen %>%

as.data.frame()

print(countAmen)

return(countAmen)

}

# Amenities by Brand

brandAmenity <- function(amenitiesArg) {

# First, get property counts

countAmen <- propCounts(amenitiesArg)

# Next, clean up the working data set

amenities <- cleanAmenities(amenitiesArg, "brand")

# Get column names without Brand and Property ID

amenCols <- colnames(amenities)[-1]

amenCols <- amenCols[-1]

# Set up results data frame with a dummy entry

results <- data.frame("junk", "more junk", as.numeric(1001101))

names(results) <- c("Brand", "Amenity", "Prop\_With")

# Repeat for each amenity column

for(colName in amenCols) {

tibb <- amenities %>%

group\_by(Brand\_PL) %>%

filter\_(paste(colName, "==", "'Y'")) %>%

select(colName) %>%

summarise(count = n())

# Create a temp DF

tmp <- data.frame(tibb[[1]], colName, tibb[[2]])

names(tmp) <- c("Brand", "Amenity", "Prop\_With")

# And merge it with the results set

results <- rbind(results, tmp)

}

### Calculate Percentages

# Remove dummy value from results DF

results <- results[-1,]

# Initialize the holding vector with a dummy value

vec <- c(0)

# Cycle through and add the number of total properties for each brand

for(entry in results$Brand) {

vec <- c(vec, countAmen$Count[which(countAmen$Brand\_PL == entry)])

}

# Strip dummy value

vec <- vec[-1]

# Add to the results DF

results$Total\_Prop <- vec

# Calculate the percentage using the mutate function

results <- results %>%

mutate(Pct\_With = round((Prop\_With / Total\_Prop) \* 100, 0))

return(results)

}

######################################

# amenitiesChiSq.R

# Perform a Chi-Squared test on the amenities

# and NPS type to determine if the two are dependent

######################################

# Get the counts of promoters and detractors per amenity type

chiSqAmenity <- function(amenitiesArg) {

# First, clean up the data set

amenities <- cleanAmenities(amenitiesArg, "nps")

# Get column names without NPS\_Type

amenCols <- colnames(amenities)[-1]

# Set up results data frame

results <- data.frame("junk", as.numeric(1001101))

names(results) <- c("Amenity", "p.value")

# Loop through all amenity columns

for(colName in amenCols) {

tibb <- amenities %>%

select(c("NPS\_Type", colName))

#filter\_(paste(colName, "!=", 0))

print(table(tibb[[1]], tibb[[2]]))

# Perform chi squared analysis

chisq <- chisq.test(tibb[[1]], tibb[[2]], correct = FALSE)

print(chisq)

# Create a temp DF

tmp <- data.frame(colName, chisq$p.value)

names(tmp) <- c("Amenity", "p.value")

# And merge it with the results set

results <- rbind(results, tmp)

}

return(results[-1,])

}

######################################

# amenitiesKSVM.R

# Contains a function for creating a

# scalable vector machine model for the

# Hyatt amenities

######################################

library(kernlab)

# Create a KSVM for the amenities

svmAmenities <- function(amenitiesArg) {

# First, clean up the data set

amenities <- cleanAmenities(amenitiesArg, "nps.factor")

## Create training and testing data sets

# Random index for picking surveys

randIndex <- sample(1:dim(amenities)[1])

# Set a cutpoint at 1/16 of the data set, larger values

# create errors due to resource constraints

cutpoint1\_6 <- floor(dim(amenities)[1]/16)

# Portion of the data set for training

trainingAmenities <- amenities[randIndex[1:cutpoint1\_6],]

# Use the remainder of the data set for testing model out

testingAmenities <- amenities[randIndex[(cutpoint1\_6+1):length(randIndex)],]

# Create the model using the training portion of the data set

ksvmAmen <- ksvm(NPS\_Type ~ ., data = trainingAmenities, kernel = "rbfdot",

kpar = "automatic", C = 10, cross = 10,

prob.model = FALSE)

# What are the details of the model?

print(ksvmAmen)

# Run a prediction using the testing portion of the data set

ksvmAmenPred <- predict(ksvmAmen, testingAmenities)

# Determine if the model predicted the NPS type correctly

ksvmAmenGood <- data.frame(testingAmenities[,1], ksvmAmenPred,

ifelse(testingAmenities[,1] == ksvmAmenPred, TRUE, FALSE))

colnames(ksvmAmenGood) <- c("Test", "Pred", "Correct")

# Calculate the percent correct

ksvmAmenCorrect <- length(which(ksvmAmenGood$Test == ksvmAmenGood$Pred)) /

dim(ksvmAmenGood)[1]

# Present the results

ksvmAmenCorrect <- round(ksvmAmenCorrect \* 100, 2)

print(sprintf("Percent good: %f", ksvmAmenCorrect))

return(ksvmAmenGood)

}

######################################

# amenitiesAssoc.R

# Contains a function for running the

# Hyatt amenities through associative

# rule mining

######################################

library(arules)

library(arulesViz)

# Associative rule mining

assocAmenity <- function(amenitiesArg) {

# First, clean up the data set, do not return factors

amenities <- cleanAmenities(amenitiesArg, "nps")

# Make it a tibble for tidyverse functions

amenities <- as\_tibble(amenities)

# Run the amenities as a matrix through the apriori function

# High support and confidence values were selected because

# amenities tend to be clustered together by brand, making them

# artificially high

apAmen <- amenities %>%

as.matrix() %>%

apriori(parameter = list(supp = 0.25, conf = 0.5))

# Create a scatter plot showing support vs confidence and

# lift

plot(apAmen)

# Return the rules where NPS type is on the right hand side and

# lift is greater than 1. From inspection of the results, the best

# rules with NPS on rhs tend to have a lift of 1.00xxx

return(subset(apAmen, subset = rhs %in% "NPS\_Type" & lift > 1.00))

}

######################################

# amenitiesTest.R

# Run through descriptive analytics and

# call functions for models.

######################################

library(tidyverse)

library(ggplot2)

library(gplots)

workingSet <- DT\_exp\_final\_set

# Basic stats, how many promoters and detractors?

proDet <- cleanAmenities(workingSet, "nps.factor")

print(sprintf("Valid surveys considered: %d", dim(proDet)[1]))

# Promoters / detractors

pro <- sum(ifelse(proDet$NPS\_Type == 1, 1, 0))

print(sprintf("Promoters: %d", pro))

print(sprintf("Detractors: %d", ((dim(proDet)[1])-pro)))

print(sprintf("Promotor Percent: %f", 100 \* (pro / (dim(proDet)[1]))))

# Color coding for plots

# Hyatt House Place Regency Andaz Grand Park

hyatt\_colors <- c('Hyatt'=col2hex("darkblue"), 'Hyatt House'=col2hex("darkgreen"), 'Hyatt Place'=col2hex("darkorange"),

'Hyatt Regency'=col2hex("darkviolet"), Andaz=col2hex("darkred"), 'Grand Hyatt'=col2hex("red"),

'Park Hyatt'=col2hex("blue"))

# Property counts

propCount <- propCounts(workingSet)

print(sprintf("Total Number of Properties: %d", propCount))

# Plot the counts

ggplot(propCount, aes(x = Brand\_PL, y = Count, fill = Brand\_PL)) +

geom\_bar(stat = "identity") +

scale\_fill\_manual(values = hyatt\_colors) +

labs(x = "Hyatt Brand", fill = "Brand") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle("Number of Properties by Brand")

# Total number of amenities offered at each brand

AmenityCount <- c(16,24,16,18,27,24,22)

Brands <- c("Andaz", "Grand Hyatt", "Hyatt House", "Hyatt Place",

"Hyatt Regency", "Hyatt", "Park Hyatt")

offered <- data.frame(Brands, AmenityCount)

# Plot the total amenities offered

ggplot(offered, aes(x = Brands, y = AmenityCount, fill = Brands)) +

geom\_bar(stat = "identity") +

scale\_fill\_manual(values = hyatt\_colors) +

labs(x = "Hyatt Brand", fill = "Brand") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle("Number of Amenities by Brand")

# Add the NPS ratio (promoter / non-promoter)

npsRatio <- c(2.79, 2.34, 2.84, 2.74, 1.87, 1.95, 3.73)

offered <- data.frame(offered, npsRatio)

# Add the percent offered, this indicates consistency across

# properties in a brand

pctOffered <- c(0.59, 0.58, 0.63, 0.39, 0.44, 0.29, 0.45)

offered <- data.frame(offered, pctOffered)

# Run linear models to confirm observations

lmCount <- lm(formula = npsRatio ~ AmenityCount, data = offered)

lmConsistency <- lm(formula = npsRatio ~ pctOffered, data = offered)

# Get amenity statistics grouped by brand

brands <- brandAmenity(workingSet)

# Sort alphabetically by Brand

brands <- brands %>%

arrange(Brand)

print("Amenities by Brand: ")

print(brands)

# Pull individual Brands

andaz <- brands %>%

filter(Brand == "Andaz")

hyatt <- brands %>%

filter(Brand == "Hyatt")

hyatt\_house <- brands %>%

filter(Brand == "Hyatt House")

hyatt\_place <- brands %>%

filter(Brand == "Hyatt Place")

hyatt\_regency <- brands %>%

filter(Brand == "Hyatt Regency")

grand\_hyatt <- brands %>%

filter(Brand == "Grand Hyatt")

park\_hyatt <- brands %>%

filter(Brand == "Park Hyatt")

# Plot all

ggplot(brands, aes(x = Amenity, y = Pct\_With, group = Brand, fill = Brand)) +

geom\_bar(stat = "identity", position = "dodge") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability by Brand") +

ylab("Percent of Properties with Amenity") +

scale\_fill\_manual(values = hyatt\_colors) +

ylim(0, 100)

# Plot Hyatt

ggplot(hyatt, aes(x = Amenity, y = Pct\_With)) +

geom\_bar(stat = "identity", position = "dodge", color = "darkblue", fill = "darkblue") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability for Hyatt") +

ylab("Percent of Hyatt Properties with Amenity") +

theme(legend.position = "none") +

ylim(0, 100)

# Plot Hyatt House

ggplot(hyatt\_house, aes(x = Amenity, y = Pct\_With)) +

geom\_bar(stat = "identity", position = "dodge", color = "darkgreen", fill = "darkgreen") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability for Hyatt House") +

ylab("Percent of Hyatt House Properties with Amenity") +

theme(legend.position = "none") +

ylim(0, 100)

# Plot Hyatt Place

ggplot(hyatt\_place, aes(x = Amenity, y = Pct\_With)) +

geom\_bar(stat = "identity", position = "dodge", color = "darkorange", fill = "darkorange") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability for Hyatt Place") +

ylab("Percent of Hyatt Place Properties with Amenity") +

theme(legend.position = "none") +

ylim(0, 100)

# Plot Hyatt Regency

ggplot(hyatt\_regency, aes(x = Amenity, y = Pct\_With)) +

geom\_bar(stat = "identity", position = "dodge", color = "darkviolet", fill = "darkviolet") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability for Hyatt Regency") +

ylab("Percent of Hyatt Regency Properties with Amenity") +

theme(legend.position = "none") +

ylim(0, 100)

# Plot Andaz

ggplot(andaz, aes(x = Amenity, y = Pct\_With)) +

geom\_bar(stat = "identity", position = "dodge", color="darkred", fill = "darkred") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability for Andaz") +

ylab("Percent of Andaz Properties with Amenity") +

theme(legend.position = "none") +

ylim(0, 100)

# Plot Grand Hyatt

ggplot(grand\_hyatt, aes(x = Amenity, y = Pct\_With)) +

geom\_bar(stat = "identity", position = "dodge", color = "red", fill = "red") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability for Grand Hyatt") +

ylab("Percent of Grand Hyatt Properties with Amenity") +

theme(legend.position = "none") +

ylim(0, 100)

# Plot Park Hyatt

ggplot(park\_hyatt, aes(x = Amenity, y = Pct\_With)) +

geom\_bar(stat = "identity", position = "dodge", color = "blue", fill = "blue") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Amenity Availability for Park Hyatt") +

ylab("Percent of Park Hyatt Properties with Amenity") +

theme(legend.position = "none") +

ylim(0, 100)

# Run a chi-squared test for independence

x.sq <- chiSqAmenity(workingSet)

# Filter p-value for 1% of better

x.sq <- x.sq %>%

filter(p.value < 0.01)

# Associative Rules Mining

goodRules <- assocAmenity(workingSet)

# Print top rules

inspect(head(goodRules, n = 5, by = "confidence"))

# Create an SVM for amenities

svmResult <- svmAmenities(workingSet)

###################################

############ BUSINESS QUESTION 3  
###################################

q3\_dat <- myData %>%

select(likelihood\_recommend\_h, `f&b\_overall\_experience\_h`, guest\_room\_h,

tranquility\_h, condition\_hotel\_h, customer\_svc\_h, staff\_cared\_h, internet\_sat\_h,

check\_in\_h, nt\_rate\_r)

q3\_dat$nt\_rate\_r <- as.numeric(q3\_dat$nt\_rate\_r)

q3\_dat$likelihood\_recommend\_h[which(is.na(q3\_dat$likelihood\_recommend\_h))] <- mean(q3\_dat$likelihood\_recommend\_h[which(is.na(q3\_dat$likelihood\_recommend\_h) == FALSE)])

q3\_dat$`f&b\_overall\_experience\_h`[which(is.na(q3\_dat$`f&b\_overall\_experience\_h`))] <- mean(q3\_dat$`f&b\_overall\_experience\_h`[which(is.na(q3\_dat$`f&b\_overall\_experience\_h`) == FALSE)])

q3\_dat$guest\_room\_h[which(is.na(q3\_dat$guest\_room\_h))] <- mean(q3\_dat$guest\_room\_h[which(is.na(q3\_dat$guest\_room\_h) == FALSE)])

q3\_dat$tranquility\_h[which(is.na(q3\_dat$tranquility\_h))] <- mean(q3\_dat$tranquility\_h[which(is.na(q3\_dat$tranquility\_h) == FALSE)])

q3\_dat$condition\_hotel\_h[which(is.na(q3\_dat$condition\_hotel\_h))] <- mean(q3\_dat$condition\_hotel\_h[which(is.na(q3\_dat$condition\_hotel\_h) == FALSE)])

q3\_dat$customer\_svc\_h[which(is.na(q3\_dat$customer\_svc\_h))] <- mean(q3\_dat$customer\_svc\_h[which(is.na(q3\_dat$customer\_svc\_h) == FALSE)])

q3\_dat$staff\_cared\_h[which(is.na(q3\_dat$staff\_cared\_h))] <- mean(q3\_dat$staff\_cared\_h[which(is.na(q3\_dat$staff\_cared\_h) == FALSE)])

q3\_dat$internet\_sat\_h[which(is.na(q3\_dat$internet\_sat\_h))] <- mean(q3\_dat$internet\_sat\_h[which(is.na(q3\_dat$internet\_sat\_h) == FALSE)])

q3\_dat$check\_in\_h[which(is.na(q3\_dat$check\_in\_h))] <- mean(q3\_dat$check\_in\_h[which(is.na(q3\_dat$check\_in\_h) == FALSE)])

q3\_dat$nt\_rate\_r[which(is.na(q3\_dat$nt\_rate\_r))] <- mean(q3\_dat$nt\_rate\_r[which(is.na(q3\_dat$nt\_rate\_r) == FALSE)])

# ---------- visualize data -----------

q3\_dat\_sample <- q3\_dat[1:1000000,]

ggplot(q3\_dat\_sample, aes(x = `f&b\_overall\_experience\_h`, y = likelihood\_recommend\_h)) +

geom\_count()

ggplot(q3\_dat\_sample, aes(x = guest\_room\_h, y = likelihood\_recommend\_h)) +

geom\_count()

ggplot(q3\_dat\_sample, aes(x = tranquility\_h, y = likelihood\_recommend\_h)) +

geom\_count()

ggplot(q3\_dat\_sample, aes(x = condition\_hotel\_h, y = likelihood\_recommend\_h)) +

geom\_count()

ggplot(q3\_dat\_sample, aes(x = customer\_svc\_h, y = likelihood\_recommend\_h)) +

geom\_count()

ggplot(q3\_dat\_sample, aes(x = internet\_sat\_h, y = likelihood\_recommend\_h)) +

geom\_count()

ggplot(q3\_dat\_sample, aes(x = check\_in\_h, y = likelihood\_recommend\_h)) +

geom\_count()

ggplot(q3\_dat\_sample %>% filter(nt\_rate\_r < 500), aes(x = nt\_rate\_r, y = likelihood\_recommend\_h)) +

geom\_count()

q3\_model <- lm(formula = likelihood\_recommend\_h ~ ., data = q3\_dat)

summary(q3\_model)

step(q3\_model, data=q3\_dat, direction="backward")

best\_model <- lm(formula = likelihood\_recommend\_h ~ `f&b\_overall\_experience\_h` + guest\_room\_h +

tranquility\_h + condition\_hotel\_h + customer\_svc\_h + staff\_cared\_h +

internet\_sat\_h + nt\_rate\_r, data = q3\_dat)

summary(best\_model)

###################################

############ BUSINESS QUESTION 4  
###################################

#build a column to count for surveys that have a likelihood to recommend even if not completed

myData$has\_likelihood <- ifelse(myData$likelihood\_recommend\_h >= 1, 1, 0)

# Summary of specific columns having to do with regional information

mySummary <- myData %>%

group\_by(hotel\_name\_long\_pl,brand\_pl,us\_region\_pl, state\_pl, city\_pl, type\_pl,location\_pl, property\_longitude\_pl, property\_latitude\_pl) %>%

summarise( guest\_count=length(cons\_guest\_id\_c)

, completedcount=sum(ifelse(m\_survey\_status=="COMPLETED" , 1, 0))

, sumLikelihood=sum(ifelse(likelihood\_recommend\_h[which(likelihood\_recommend\_h >= 9)],1,0))

, sumHasLikelihood = sum(has\_likelihood, na.rm = TRUE)

, promoterRatio=sumLikelihood/(sumHasLikelihood)

, meanLikelihood=mean(likelihood\_recommend\_h,na.rm=TRUE)

, totalGoodAmenities = rowSums(myHotelsNew[,26:54] == "Y")

)

#-- PLOT INFORMATION FOR BEST HOTELS BY REGION PER TYPE --#

#subset of data that includes information for the best hotel per region

highestInRegion <- mySummary %>%

group\_by(us\_region\_pl, type\_pl) %>%

summarise(promoterRatio = max(promoterRatio))

#graph brand by type

brandByType <- myData %>%

group\_by(type\_pl, brand\_pl) %>%

summarise(average\_score = mean(likelihood\_recommend\_h, na.rm = TRUE))

brandByType.graph <- ggplot(brandByType, aes(x = brand\_pl, y = average\_score))

brandByType.graph <- brandByType.graph + geom\_point(data = brandByType, aes(color = type\_pl, size = 10))

brandByType.graph <- brandByType.graph + ggtitle("Brand By Type of Hotel")

brandByType.graph

#put all information from mySummary into highestInRegion to retain data for the subset of 24 hotels

highestInRegion <- merge(mySummary, highestInRegion, by = "promoterRatio")

highestInRegion$state\_pl <- tolower(highestInRegion$state\_pl)

#setting up information to plot to a map

us <- map\_data("state")

#merge all data from US and highestInRegion

newData <- merge(us, highestInRegion, by.x = 'region', by.y='state\_pl')

newData <- newData[,-21:-22]

#make a map that plots best hotels by type, location, and region without brand names

bestByRegion <- ggplot(us, aes(map\_id=region))

bestByRegion <- bestByRegion + xlim(-125, -65)

bestByRegion <- bestByRegion + geom\_map(map = us, fill = "white", color = "black")

bestByRegion <- bestByRegion + expand\_limits(x=us$long, y = us$lat)

bestByRegion <- bestByRegion + coord\_map() + ggtitle("Best by Region")

bestByRegion <- bestByRegion + geom\_point(data=newData

, aes(x = property\_longitude\_pl, y = property\_latitude\_pl, size=promoterRatio

,colour=type\_pl.x

,shape = location\_pl))

bestByRegion

#make a map that plots best hotels by type, location, and region with brand names

bestByRegion <- ggplot(us, aes(map\_id=region))

bestByRegion <- bestByRegion + xlim(-125, -65)

bestByRegion <- bestByRegion + geom\_map(map = us, fill = "white", color = "black")

bestByRegion <- bestByRegion + expand\_limits(x=us$long, y = us$lat)

bestByRegion <- bestByRegion + coord\_map() + ggtitle("Best by Region")

bestByRegion <- bestByRegion + geom\_point(data=newData

, aes(x = property\_longitude\_pl, y = property\_latitude\_pl, size=promoterRatio

,colour=type\_pl.x

,shape = location\_pl))

bestByRegion <- bestByRegion + geom\_text(data = newData, aes(x = property\_longitude\_pl, y = property\_latitude\_pl, label = brand\_pl))

bestByRegion

##FIND AMOUNT OF GOOD AMENITIES PER TOP HOTEL PER REGION

#extract good amenities for every hotel

myHotels <- unique(subset(myData, select=c(41:94)))

myHotels <- merge(myHotels,mySummary,by="hotel\_name\_long\_pl")

#remove insignificant amenities

myHotels <- myHotels%>%

select(-business\_center\_pl, -conference\_pl, -elevators\_pl, -pool\_indoor\_pl, -spa\_pl, -`spa\_f&b\_offering\_pl`)

#identify which columns represent amenities

colnames(myHotels[26:48])

#get a sum for good amenities per hotel

myHotels$totalAmenities <- rowSums(myHotels[,26:48] == "Y")

#fill in the amenity information on the newData data frame

newData <- merge(myHotels, newData, by = "hotel\_name\_long\_pl")

#get amount of amenities per top hotel

amenityTable <- newData %>%

group\_by(type\_pl.y, brand\_pl.y, us\_region\_pl.y) %>%

summarise(total\_amenities = mean(totalAmenities))

#graph amenities per region per type of hotel

businessGraph <- ggplot(amenityTable, aes(x = type\_pl.y, y = total\_amenities))

businessGraph <- businessGraph + geom\_point(data = amenityTable, aes(color = us\_region\_pl.y, size = 10))

businessGraph <- businessGraph + ggtitle("Amenities by Region")

businessGraph <- businessGraph + scale\_y\_continuous(breaks = round(seq(min(0), max(20), by = 1)))

businessGraph