IST-687 Final Project

ISHANI JARIWALA

MONIKA PAWAR

POOJA GODHWANI

SAI PRAVINDAR MOHAN

TUSHAR BADLANI

ARIJIT NANDY

Hotel Data Analysis

Table of Contents

INTRODUCTION

Dataset contains about 3M responses from a customer survey. The surveys are from the hospitality industry - Hyatt. There are about 120 columns for each observation (survey), some columns are about the person who responded to the survey (ex. a member of their rewards program, and if so, what level), some attributes are about the hotel (ex. location) and some are responses to the survey from the customer who stayed at the hotel (ex. would they recommend the hotel to a friend).

The focus was NPS (Net Promoter Score). The goal was to identify and then answer interesting questions, such as understanding how NPS varies across surveys (geography, different hotels, frequent vs non-frequent guests, etc).

DATA CLEANING AND PREPARATION

Dataset had many columns to be blank or NA. The complete dataset is about 100GB, but the analysis was done over the dataset of June and December. The dataset is first loaded into R from CSV format and the dataset is viewed before cleaning and preprocessing. Below is the str() and summary() of the dataset.

```
# Removing NA's from dataset
POV <- na.omit(POV)

nrow(POV)
#Total number of records in data set after removing NA's - 3643250

#Converting column NPS_Type from factor to character to remove blanks
POV$NPS_Type <- as.character(POV$NPS_Type)

#Removing blanks
POV <- POV[which(POV$NPS_Type != ""),]
```

INITIAL APPROACH

- Data Analysis on entire dataset
- Replacing NA's with mean
- Couldn't find a meaningful NPS
- We discover there were serious flaw with the approach

FINAL APPROACH

- Eliminating irrelevant columns

Certain columns were eliminated which were not relevant for our final analysis and prediction of our model.

R code to eliminate irrelevant attributes from the dataset.

```
# Retaining required columns in the dataset
AnalysisDF1 <- AnalysisDF[,which(names(AnalysisDF) %in% c("POV_CODE_C","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","F.B_FREQ_H","NPS_Type","State_PL","Country_PL"))]

# Business Question: Purpose of visit analysis
POV <- AnalysisDF1[,c("POV_CODE_C","NPS_Type","State_PL","Country_PL")]
```

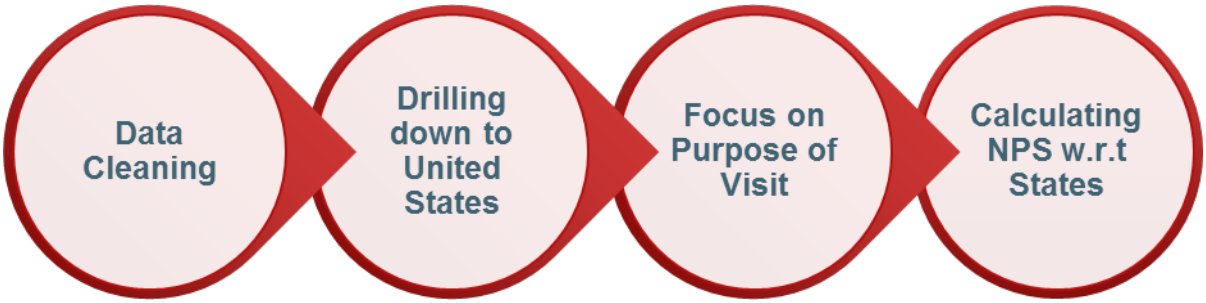
List of relevant columns that were considered in analysis:

- NPS_Type
- POV_CODE_C
- 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
- F.B_FREQ_H
- NPS_Type
- State_PL
- Country_PL"

Missing values

The dataset had many missing values, hence the concentration was only on the actual data which had values.

- **Concentrating on Purpose of visit as**



After data cleaning and drilling down to purpose of visit, we encountered that majority of the customer visits hotel for business purpose.

TARGET AUDIENCE

Data Science Expert - Erik Scott Anderson

Process Expert - Ivan Shamshurin

Client - Jeff Saltz

DATA ANALYSIS QUESTIONS, GRAPH ANALYSIS AND RECOMMENDATION

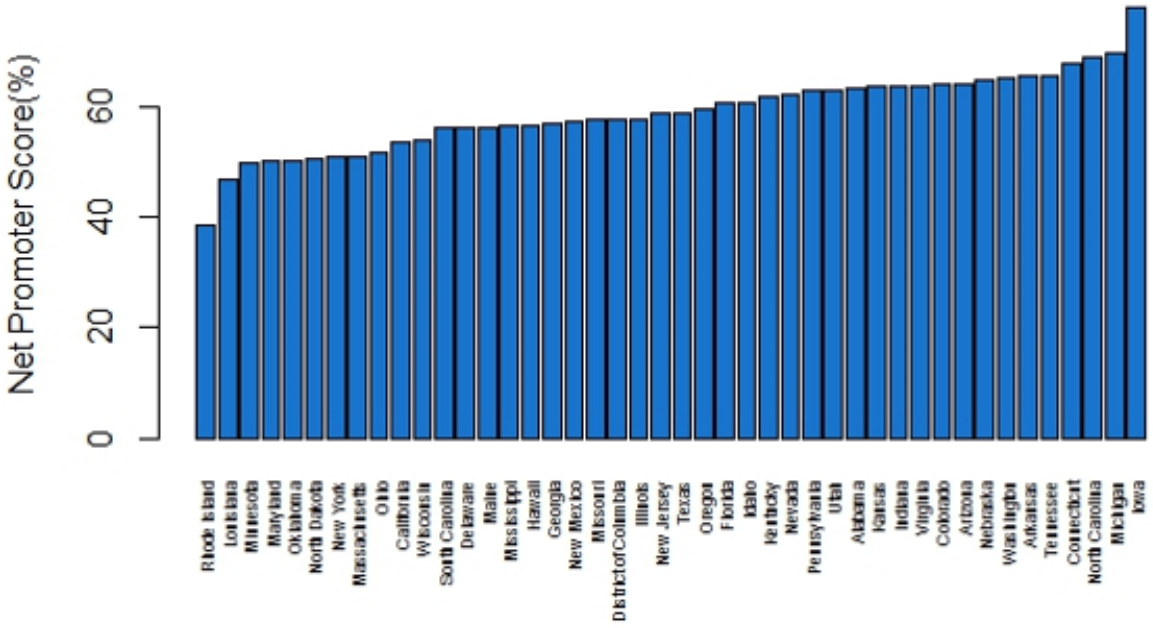
1. WHAT IS THE TREND/PATTERN OF CUSTOMERS IN USING HOTEL'S SERVICES AND THE DIFFERENCE, IF ANY, BASED ON CUSTOMER'S PURPOSE OF VISIT?

Method used: Heat Map

After drilling down the business process to purpose of visit, likelihood to recommend hotel based on major 9 factors that affected NPS. These factor were narrowed down to 44 states of USA. States which had no data were deleted from the dataset. Based on which the NPS were calculated.

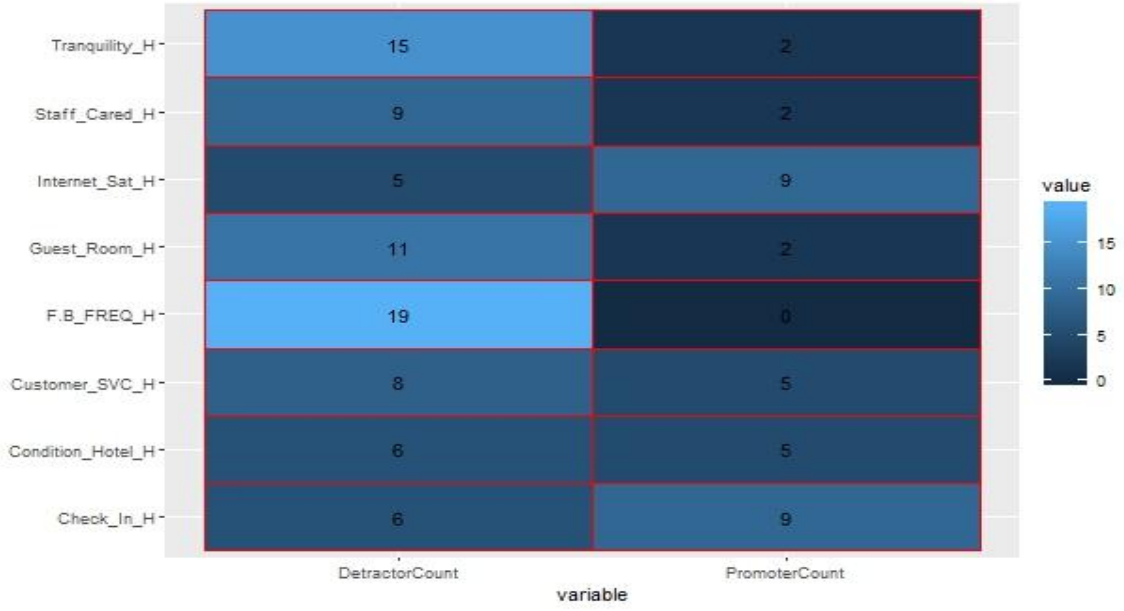
	Detractor	Passive	Promoter	Difference	Frequency	NPS
Rhode Island	120	126	349	229	595	38.48739
Louisiana	280	379	1101	821	1760	46.64773
Minnesota	181	310	845	664	1336	49.70060
Maryland	335	468	1479	1144	2282	50.13146
Oklahoma	109	174	505	396	788	50.25381
North Dakota	12	17	54	42	83	50.60241
New York	916	1200	4070	3154	6186	50.98610
Massachusetts	308	466	1437	1129	2211	51.06287
Ohio	408	484	1797	1389	2689	51.65489
California	2575	3536	12585	10010	18696	53.54086
Wisconsin	179	283	933	754	1395	54.05018
South Carolina	148	201	783	635	1132	56.09541
Delaware	41	55	217	176	313	56.23003
Maine	34	33	164	130	231	56.27706
Mississippi	28	32	143	115	203	56.65025
Hawaii	302	529	1789	1487	2620	56.75573
Georgia	669	875	3601	2932	5145	56.98737
New Mexico	168	242	945	777	1355	57.34317
Missouri	210	249	1124	914	1583	57.73847
District of Columbia	219	398	1363	1144	1980	57.77778
Illinois	870	1509	5325	4455	7704	57.82710
New Jersey	308	536	1944	1636	2788	58.68006
Texas	1560	2338	9377	7817	13275	58.88512

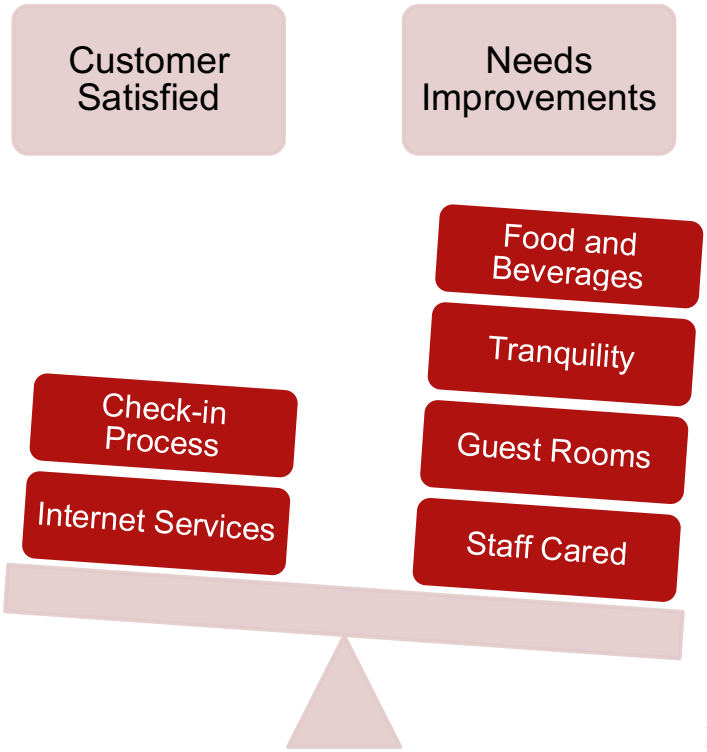
Important analysis from this data was Rhode Island had the least NPS.



Next focus of analysis was over the state Rhode Island which had the least NPS. Analysis now was done finding detractors for Rhode Island State.

Heat map below shows the factors affecting NPS for Rhode Island.





Recommendations

Rhode Island hotels should focus on the improving food and beverages services none of the detractors rated food and beverages services above 3. That clearly shows poor services being offered with respect to food and beverages.

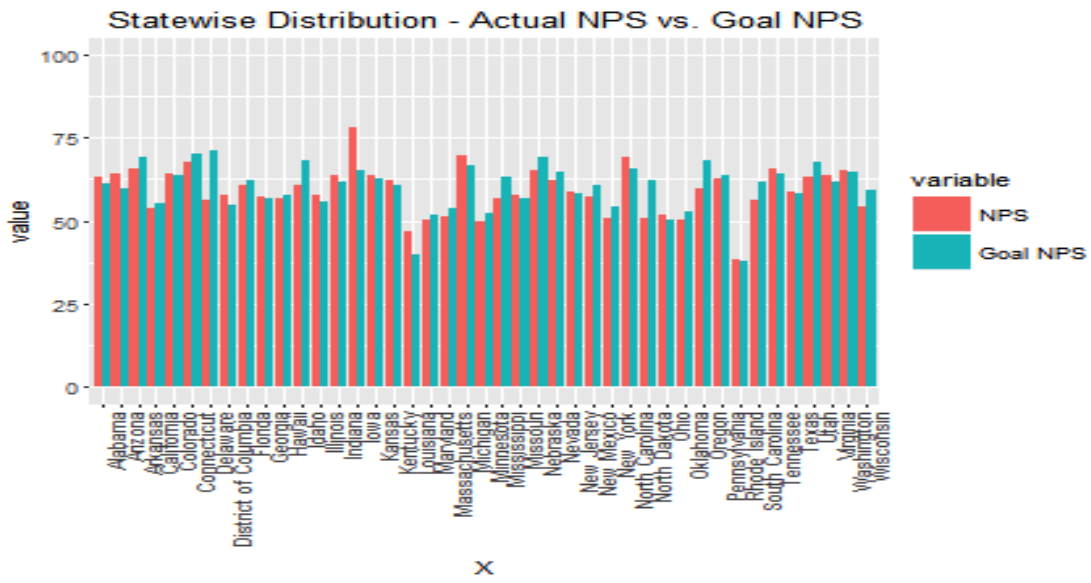
Also the tranquility, guest rooms and staff care needs significant improvements.

2. Comparing Average NPS value of each state with its Target NPS goal value (This value is set by the HyaTT group For each and every hotel)

Method Used: Visualization using Bar plots

After Analysis the dataset we came across a unique column called **Guest NPS Goal_PL** which contains Hotel's NPS goals value. This column helped us understand what NPS score the Hyatt expects its each and every individual hotel to have depending the location and business scenario. Instead of doing this analysis of each individual hotel, the data was grouped as per each state. The analysis help us understand which States are lacking in the achieving the set goal and what the reason for their poor performance.

The below graph show the comparison of goal NPS value and actual NPS value of each state.



Thus it's evident from the graph that though according to our previous analysis Rhode Island state has the lowest NPS value, its able to meet the set goal NPS value. This warns us to shift focus to other states.

Overall there are more than 22 states which are lacking behind to achieve the set Goal NPS target. Out of all these states **Delaware, Idaho and North Dakota** are states which have huge difference between their actual and goal NPS.

Thus this analysis will help Hyatt to concentrate on top performing states and properties This help them evaluate which services and amenities are working great in top performing states and implementing same strategies in low performing states can help them to increase the NPS value.

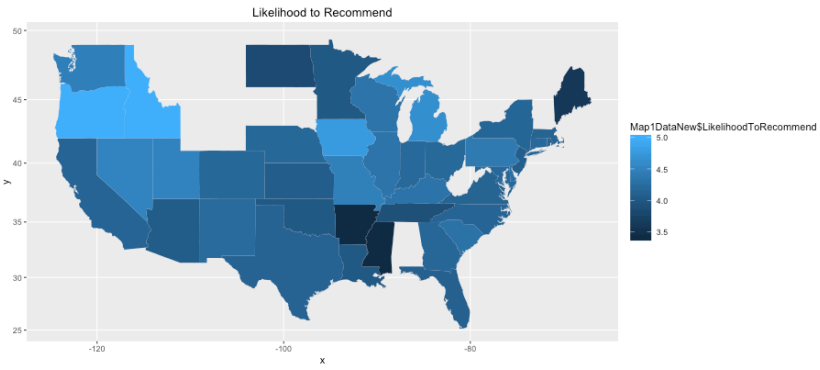
3. WHICH PROPERTY (GEOG. LOCATION) IS MOST PREFERRED DEPENDING ON STAY, FOOD, SERVICES, MEMBERSHIP DETAILS?

Method Used: Maps and Linear Modelling

For answering above question we took analyzed 4 different columns like likelihood to recommend, Tranquility, Hotel conditions and Customer service. For this part of the analysis we have only used data of detractors. We applied following filters to get the relevant dataset to work on.

Firstly, we removed all the NA values instead of replacing them with mean. As there were more than 70% values in the dataset.

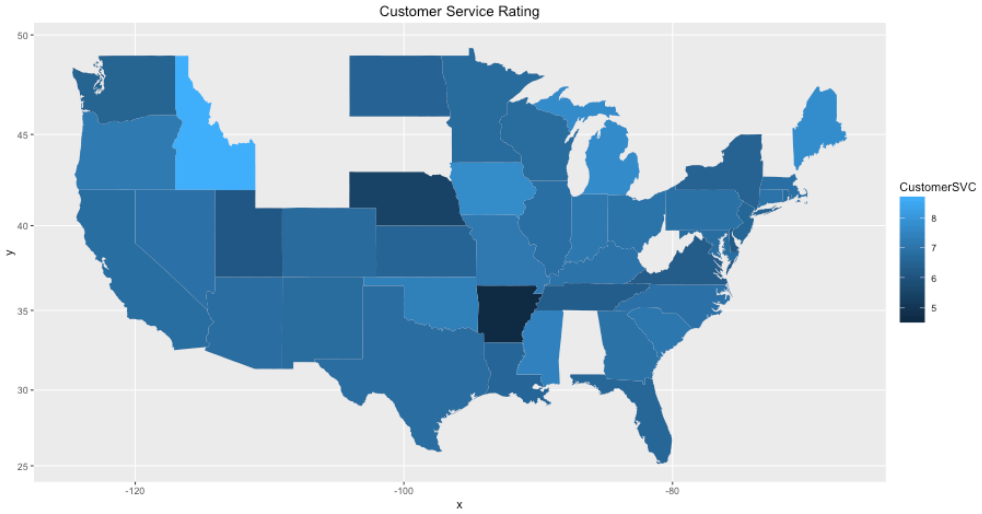
Secondly we only concentrated on the US based data as the majority of the detractors were in the country of US



Interpretation:

- The above map is made by taking an average of the likelihood to recommend and after looking at the results we can conclude that states such as Arkansas, Maine, and Mississippi have the least likelihood to recommend.

Our next analysis was based on the Customer rating data of the US and below is the map for detractors and the

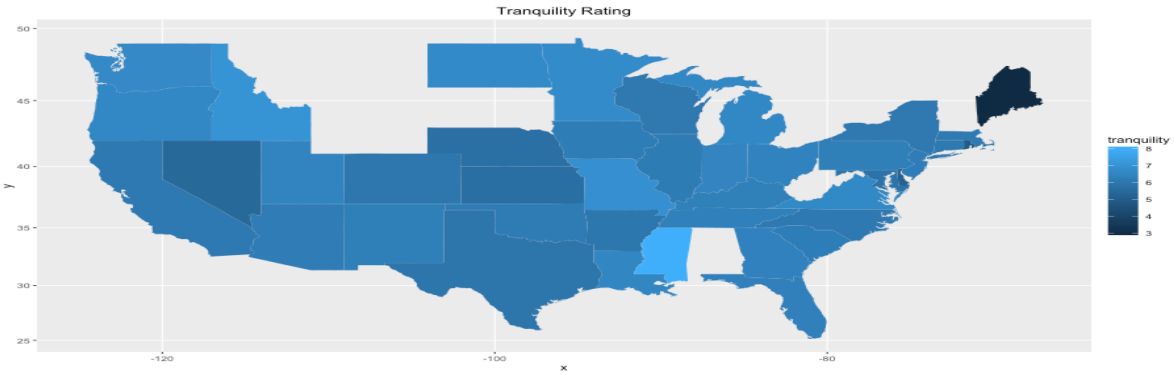


scale used is between 5-8.

Interpretation:

- The state of Arkansas has the least customer ratings.

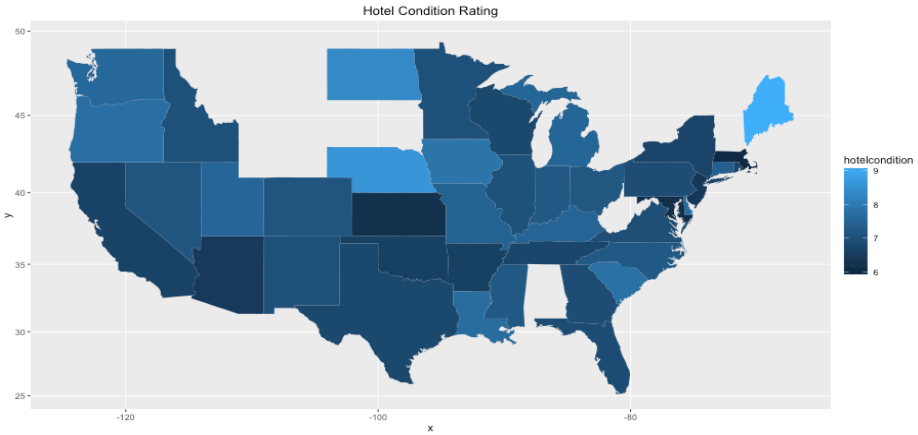
We applied the same analysis for Tranquility issues in United states and below is the map for the detractors.



Interpretation:

- Maine has the least Tranquility rating.

Similarly for the hotel condition ratings we found that,



Interpretation:

- There are no ratings below 6 in this map hotel condition is not the reason for the less rating for recommendations. Thus, discarding this analysis.

After looking at the maps and analyzing the results we did linear modelling for the same and following are the results:

We applied linear modelling for likelihood to recommend with respect to Customer Service, Tranquility, Hotel Condition and their combination to determine the dependency of these aspects on the likelihood to recommend.

49% of the time it was only dependent on the customer service

36% of the time it was only dependent to Tranquility and

45% of the time it was only dependent on the hotel conditions

After combining all these rating and then applying linear modelling it was 62% dependency.

Recommendation:

Recommendations on analysis for this question would be:

- Improvement in the customer service in the state of Arkansas
- Improvement in the Tranquility issues in the state of Maine

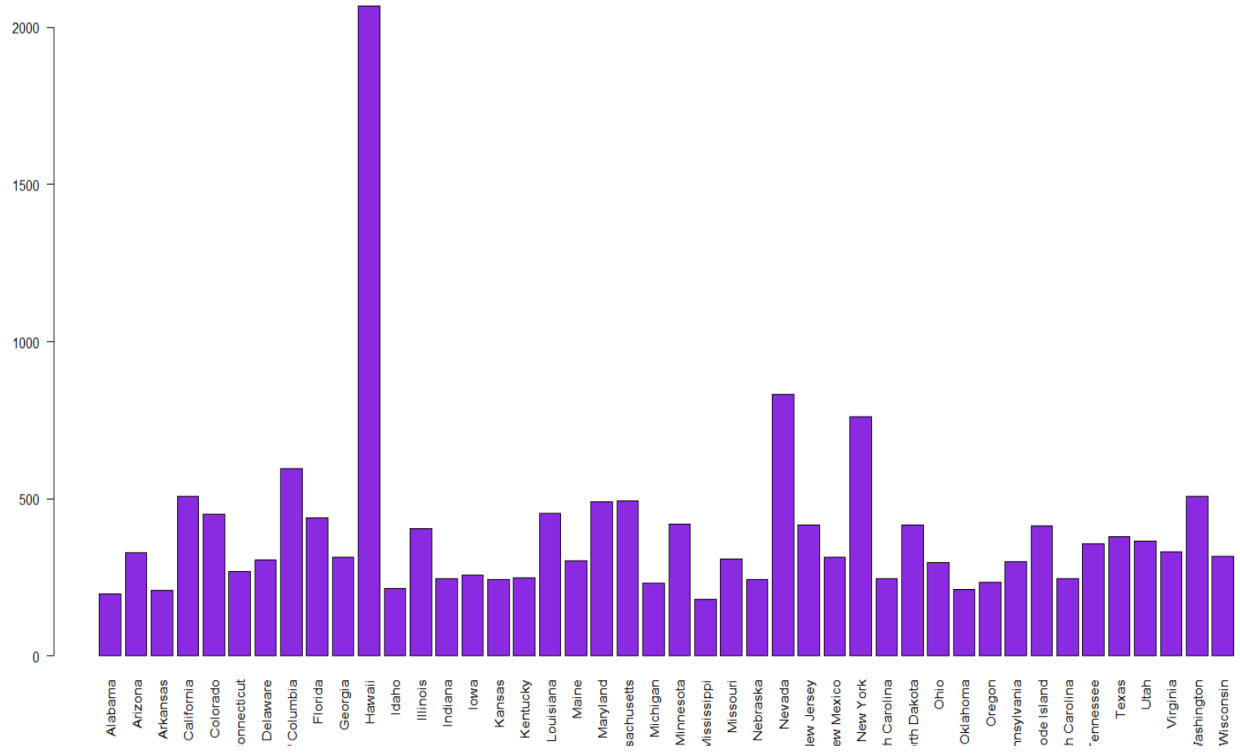
4. UNDERSTAND THE DIFFERENT REVENUE TRENDS AND PATTERNS ACROSS HOTELS IN UNITED STATES. WE FIRST ANALYZED GROSS, NETT AND ROOM REVENEUES, LENGTH OF STAY AND LIKLIHOOD TO RECOMMEND BY AGGREGATING ALL THE RECORDS AS PER EACH STATE AND THEN FINDING THE MEAN BY USIN COUNT OF EACH RECORD PER STATE.

Method used: Bar plots

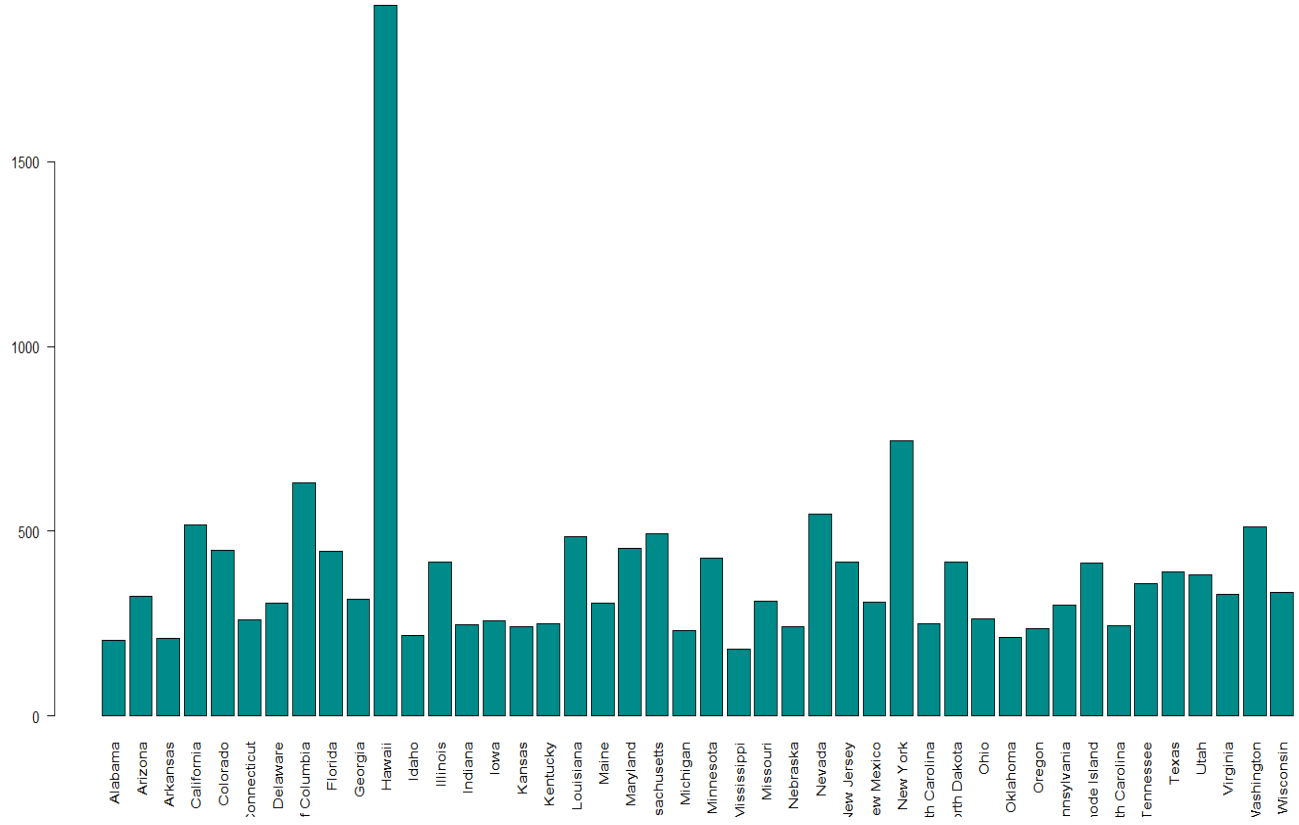
Used a smaller csv file 'USData.csv' that has the Columns City, State, US Region, Country, Number of Rooms, Length of Stay, Gross Revenue, Net Revenue and Room Revenue.

Thereafter, made smaller data frames. In a data table, calculated the count of instances for each state of ensure they can be used to perform aggregation functions with other parameter like Gross Revenue, Room Revenue and Net Revenue, as well as Length of stay.

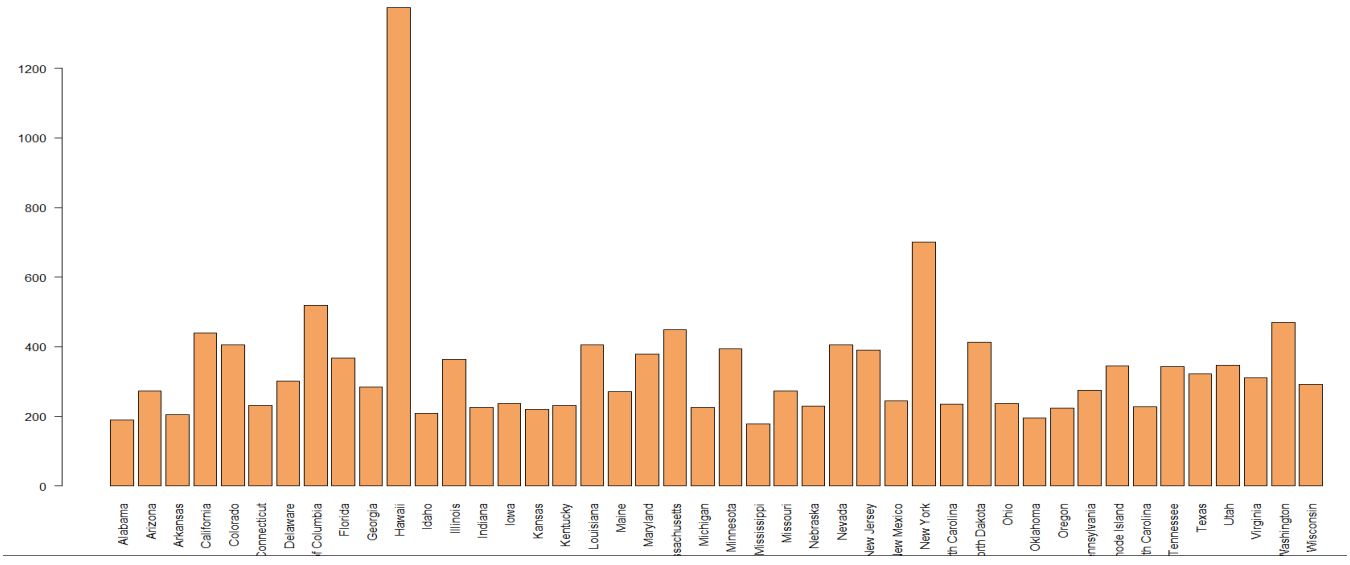
Plot of Average Gross Revenue per state:



Plot of Average Net Revenue per state:



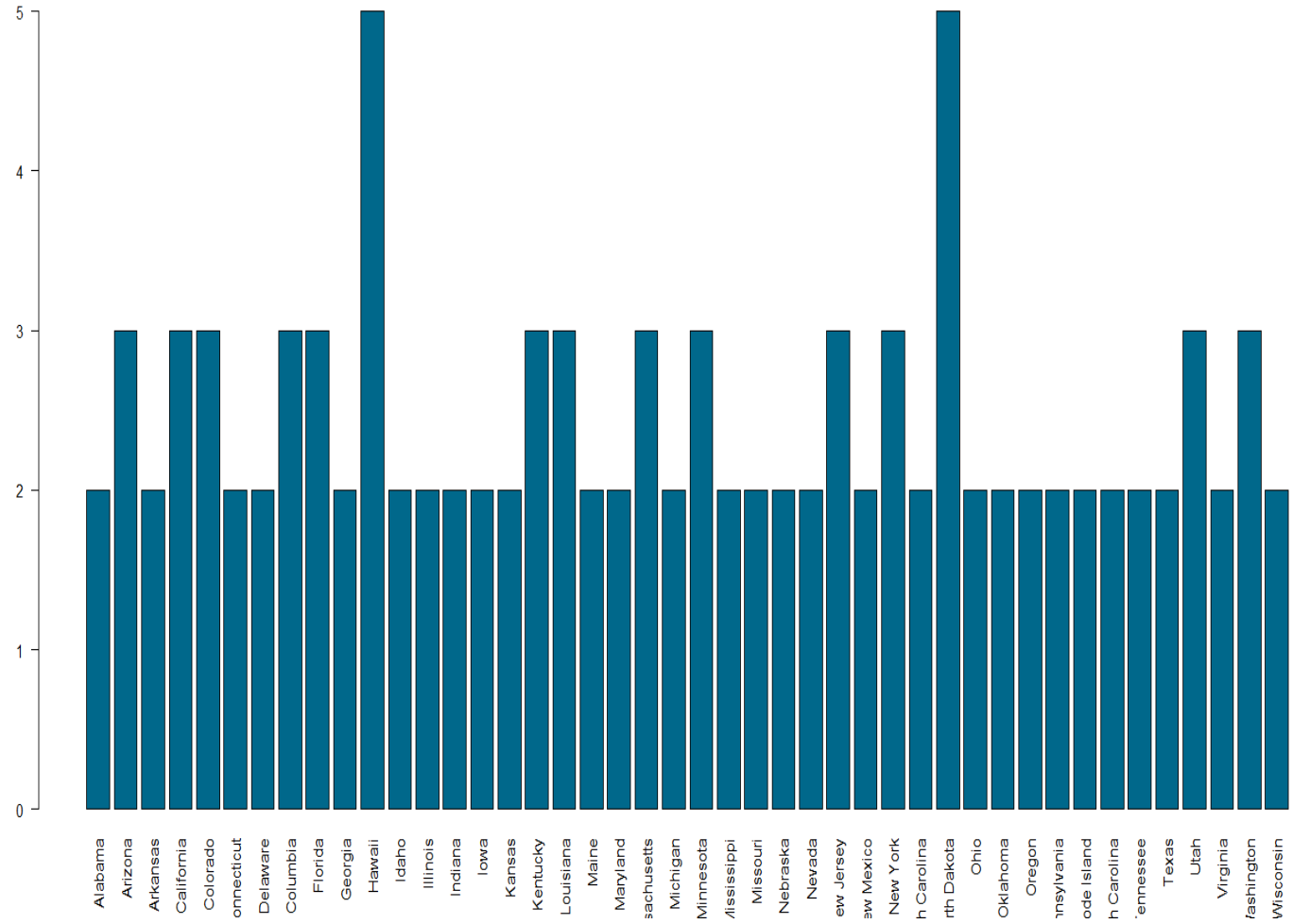
Plot of Average Room Revenue per state:



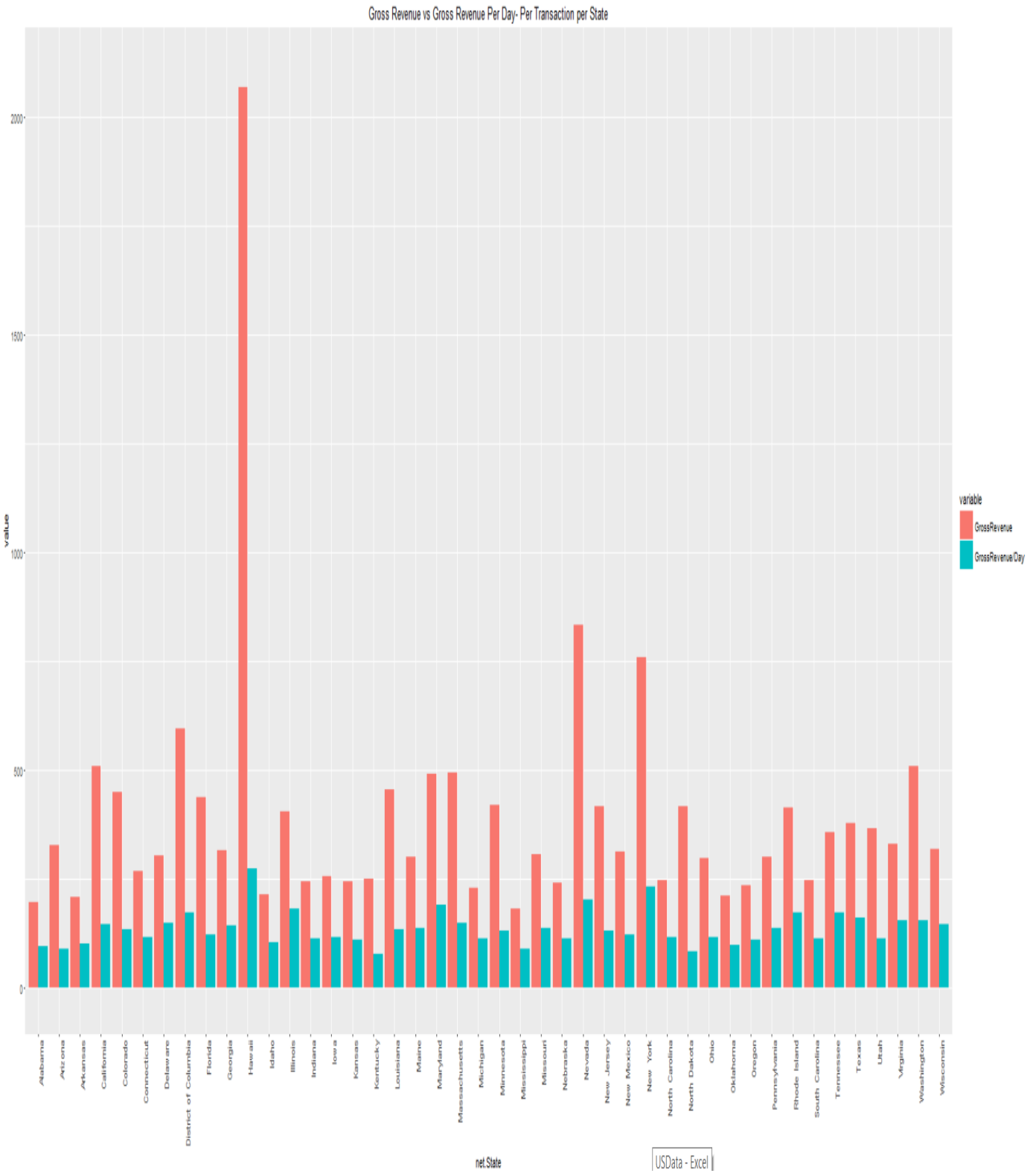
Conclusion: As we move from gross revenue to net revenue to room revenue, the general revenue values decrease. In the state of Nevada, this decline is a lot more due to which we can conclude that Nevada hotels have more revenue streams as compared to other hotels, and can be attributed to 335 casinos there. In Hawaii,

the difference in revenue is not very stark, hence most if not all the revenues are coming from hotel services itself.

Plot of Average Length of Stay per state:



This observation is interesting since there is nothing to suggest that the general revenue per state has anything to do with the length of stay for each state. In fact, quite a few hotels with longer length of stay had lower revenues generated per room. To best understand this situation, a plot of Gross Revenue and Gross Revenue per Day was created as shown below:


Gross Revenue vs Gross Revenue Per Day- Per Transaction per State
value
2000
1500
1000
500
0
variable
GrossRevenue
GrossRevenuePerDay
Alabama
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Nebraska
Nevada
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
Tennessee
Texas
Utah
Virginia
Washington
Wisconsin
net.State
USData - Excel

Observations:

1) Hawaii, the best performing hotel in revenues, has the highest change when compared to Revenue per Day. It is only slightly higher than other states which means customers stay there longer and that increases gross revenue.
2) States with maximum variance in both: Nevada, New York and Delaware
   States with minimum variance in both: Texas, Virginia, Texas, Tennessee, Michigan, Nebraska and Arkansas.
3) Best States to open hotel (as per Revenue): Hawaii, New York, Nevada, Maryland, Illinois
4) Not Best States to open hotel (as per Revenue): Kentucky, Mississippi, North Dakota, Arizona, Oklahoma

5. DISTRIBUTION OF CUSTOMERS TRAVELLING ON BUSINESS BASED ON THE NUMBER OF TIMES THEY USED THE FOOD AND BEVERAGES IN THE OUTLETS.

Method Used: Scatterplots

- The above data cluster will help understand the food and beverage preferences of customers who are travelling on business and why they remain a detractor.
- The output/refection of this question is to provide the customer detailed analysis of their food and beverage performance and recommendations as to where they can improve their focus.

Step 1:

We started to clean the huge dataset to remove NA values and then analyzed to it find columns which were relevant on this particular question.

Step 2:

We filtered the dataset to accommodate values only customers who were travelling on business and their food/beverage experiences.

Step 3:

After drilling down the dataset to accommodate only USA relevant data, we identified the following columns which were required for this respective question,

- Purpose of visit,
- Likelihood to recommend
- NPS type
- Food & Beverage frequency
- Food & Beverage experience
- Length of stay

Step 4:

After performing an in-depth data analysis, we found that there were different ranges of customer stay present in the data. There were 5 different ranges,

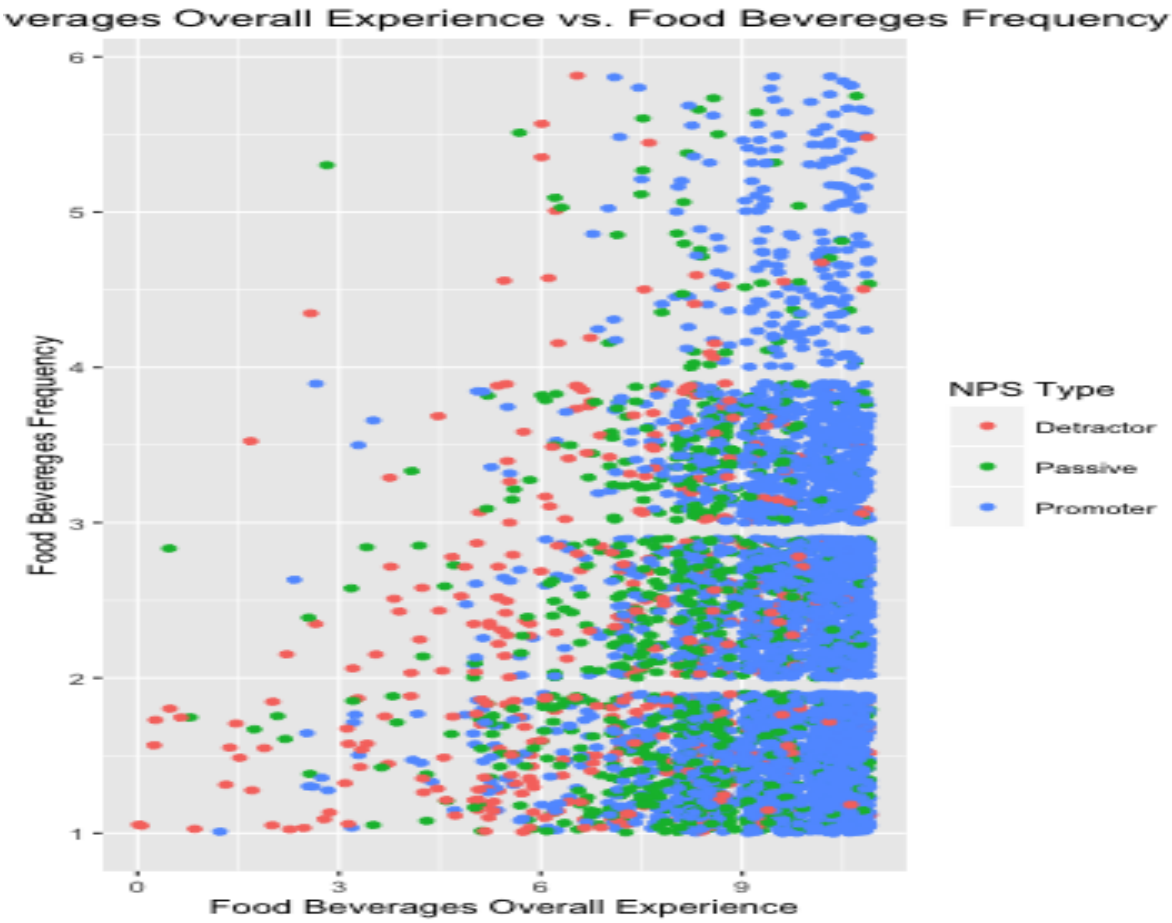
- 0-4
- 5-12
- 12-29
- 30-60
- 60-79
- 80-120

From the above ranges we found that there were only two of the ranges which were present through out 95% of the data. The two ranges were 0-4 and 5-11. We chose to proceed our analysis on 5-11 days of customer stay because if a customer is staying for 0-4 days then they are not going to visit/experience food and beverages for more number of times. **Therefore, we filtered the data to accommodate only customers who were traveling on business and staying for 5 -11 days.**

Step 5:

The data received after identifying the columns were plotted using ggplot function with Food and beverage experience on the x axis and food and beverage frequency (number of times the customer utilized the food and beverages) on the y axis. Also the points present on the map were with respect to the promoter, detractor and passives. This was done using ggpoint function.

After performing the steps mentioned above, the graph can be represented as follows,



Interpretation from the dataset & map

The interpretation/recommendations from the dataset and map mentioned above are,

- The customers who are travelling on business are more than the customers who are travelling on leisure.
- The customers travelling on business generally stay for 0-11 days.
- Most of the customers who are travelling on business tend to experience food and beverages less number of times.
- The customers who are experiencing food and beverages less number of times are more likely to be detractors like mentioned in the graph above.
- The customers who are experiencing food and beverages more number of times are more likely to be promoters like mentioned in the graph above.

Recommendations:

The client should focus on advertising/marketing strategies on social media or through partnering with corporate companies to increase the usage of food and beverage from the customers who are travelling on business. They should create awareness among the hotels throughout USA to focus on increasing customer experience when the business customers visit their outlets so that they feel welcomed and thereby increasing the possibility of revisiting the outlets.

Lessons Learnt:

- Identified the problem and also identified the data required to solve the problem.
- Understood basic of R programming and how data is organized, managed and used.
- Learnt basic principles and practices in data cleaning, screening and linking and applied on the dataset received from the client.
- Transformed data through processing, linking, aggregation, summarization, and searching
- We used different ways for visualization like maps and scatter plots to interpret the results within US.
- Taking decisions like Recommendations based on the results from the above analysis.
- We worked as a team, it helped in brainstorming sessions initially to identify the columns needed to solve the problem.
- After getting the data we analyzed data and generated different questions which will solve the purpose of the project and divided those questions among ourselves and the methods used to answer those questions.
- The contribution from all the team member was equal
- We met once in every week to discuss the progress of each other and discuss the results achieved by us.