Introduction:

In hospitality industry, customer reviews play important role in driving business and generating revenue. In order to succeed, hotels should focus on how to improve customer satisfaction that will drive overall increase in customer review ratings.

Descriptive Statistics:

The data being used in this study is Trip Advisor customer reviews for year 2012 obtained from DAIS at http://times.cs.uiuc.edu/~wang296/Data/. Original data was available in JSON format with following information:

- Hotel Information : Hotel detail as Hotel URL, Hotel ID and Hotel location specifics
- Customer Reviews for Each Hotel in the dataset as ratings, review date, and comments
- Each rating is based on scale of 1-5 with 1 as lowest rating and 5 as highest rating or very satisfied customer.
- Ratings are given as overall based on wholesome experience during the stay. This information is available for all the reviews
- Ratings are also given on specific service and accommodation provided. For example Room, Cleanliness, Service and Location. However, it depends on customers to provide ratings on each category, some categories and don't provide it all and just provide overall rating.

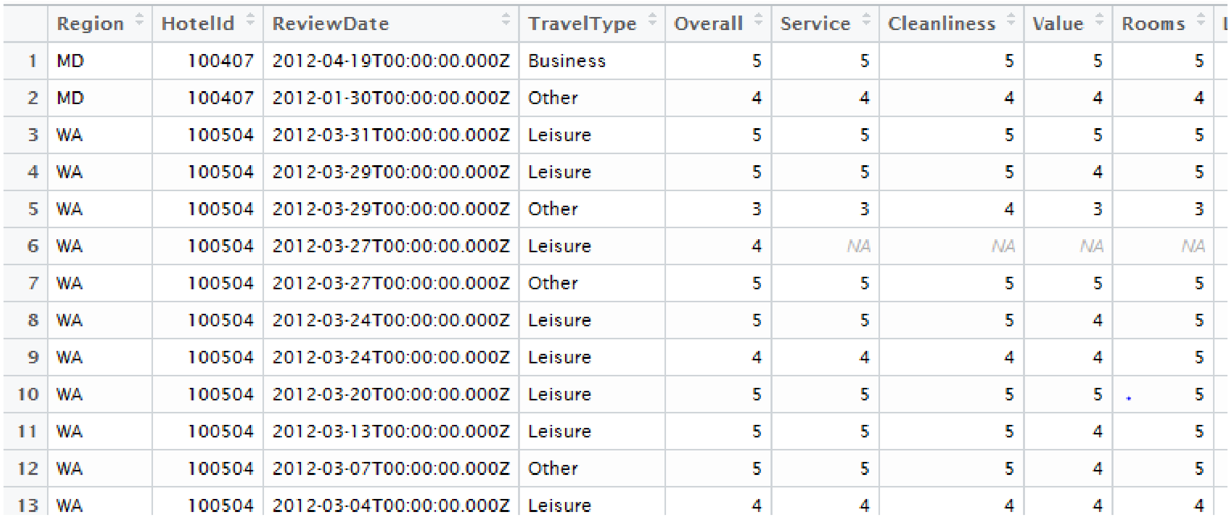
In order to use data for this study, I have added TravelType to each review based on customer comments.

- Leisure Travel: If there is anything mentioned in comments in regards to family members, pets, life events, concerts, games, holidays
- Business Travel: If there is anything mentioned in comments in regards to client, training, seminar
- Other Travel: If the comments does not fall into one of the two above mentioned categories

As Travel Type is based on customer comments, there is a chance that the review might have fallen into wrong category as compared to customers anticipated travel plan.

Data selection:

I have focused on data for year 2012. There are 115,049 reviews for Hotels mostly in USA and Europe. Data has been formatted to be used in tabular form as following:



	Region	HotelId	ReviewDate	TravelType	Overall	Service	Cleanliness	Value	Rooms
1	MD	100407	2012-04-19T00:00:00.000Z	Business	5	5	5	5	5
2	MD	100407	2012-01-30T00:00:00.000Z	Other	4	4	4	4	4
3	WA	100504	2012-03-31T00:00:00.000Z	Leisure	5	5	5	5	5
4	WA	100504	2012-03-29T00:00:00.000Z	Leisure	5	5	5	4	5
5	WA	100504	2012-03-29T00:00:00.000Z	Other	3	3	4	3	3
6	WA	100504	2012-03-27T00:00:00.000Z	Leisure	4	NA	NA	NA	NA
7	WA	100504	2012-03-27T00:00:00.000Z	Other	5	5	5	5	5
8	WA	100504	2012-03-24T00:00:00.000Z	Leisure	5	5	5	4	5
9	WA	100504	2012-03-24T00:00:00.000Z	Leisure	4	4	4	4	5
10	WA	100504	2012-03-20T00:00:00.000Z	Leisure	5	5	5	5	5
11	WA	100504	2012-03-13T00:00:00.000Z	Leisure	5	5	5	4	5
12	WA	100504	2012-03-07T00:00:00.000Z	Other	5	5	5	4	5
13	WA	100504	2012-03-04T00:00:00.000Z	Leisure	4	4	4	4	4

Goal:

The goal of this study is to analyze data using data using multinomial logistic regression model to analyze following:

- Are customers equally likely to give overall rating as 1, 2, 3, 4 or 5 based on travel type as Leisure or Business? I am using travel type of Other as reference.
- Does individual ratings for each category as Room, Service, Cleanliness and Location drive Overall rating?

Explanatory variables:

Variable	Category / Rating	Data type	Abbreviation
TravelType	Leisure, Business Other	Categorical	T
Rooms	1, 2, 3, 4, 5	Categorical	R
Service	1, 2, 3, 4, 5	Categorical	S
Cleanliness	1, 2, 3, 4, 5	Categorical	C
Location	1, 2, 3, 4, 5	Categorical	L

Table 1

Response variables:

Variable	Rating	Data Type	Abbreviation
Overall	1,2,3,4,5	Categorical	O

Table 2

Analysis:

I have decided to solve the stated problems separately.

Part 1:

H_0: The odds to give any Overall rating does not depend of Travel Type

H_A: The odds to give Overall rating depends on Travel Type for at least one rating

Following is the statistics of ratings falling into three categories as Leisure, Business and Other:

```
> with(data2012, table(Overall,TravelType))
       TravelType
Overall Business Leisure Other
      1      164    2862  2057
      2      237    3937  2096
      3      426   11114  4489
      4      904   26478 10577
      5     1110   33356 15242
```

Table 3

Given y represent Overall rating, and x represents reading for category as business or leisure, so given the formula for logits:

$$\log(y = i) = \log\left(\frac{p(y=i)}{1-(p=i)}\right) = \beta_{i0} + \beta_1 x_{\mathrm{i}2} + \beta_2 x_{\mathrm{i}3} \text{ for } i = 1..5$$

After obtaining the results from R for multinomial regression:

```
> data2012$TravelTypeL<-relevel(data2012$TravelType, ref="Other")
> testTravelType<-multinom(Overall~TravelTypeL, data=data2012)

Coefficients:
  (Intercept) TravelTypeLBusiness TravelTypeLLeisure
2  0.02007773          0.34999462          0.2989228
3  0.78166651          0.17216995          0.5752198
4  1.63824238          0.06917530          0.5865900
5  2.00365341         -0.09148068          0.4520932
```

Table 4

After plugging in these numbers to above equation:

$$\log(y = 2) = \log\left(\frac{p(y=2)}{1-(p=2)}\right) = 0.02 + 0.35x_{22} + 0.3x_{23}$$

$$\log(y = 3) = \log\left(\frac{p(y=3)}{1-(p=3)}\right) = 0.78 + 0.17x_{32} + 0.58x_{33}$$

$$\log(y = 4) = \log\left(\frac{p(y=4)}{1-(p=4)}\right) = 1.64 + 0.07x_{42} + 0.59x_{43}$$

$$\log(y = 5) = \log\left(\frac{p(y=5)}{1-(p=5)}\right) = 2 - 0.09x_{52} + 0.45x_{53}$$

Focusing on y = 4, one unit increase in business traveler will increase the odds of having overall rating = 4 by 0.07.

After calculating the predicted probability for odd of giving a rating of 1-5 by any type of customer:

```
> dTravelType<-data.frame(TravelTypeL=c("Other", "Business", "Leisure"))
> predict(testTravelType, newdata=dTravelType,"probs")
           1          2         3         4         5
1 0.05963880 0.06084831 0.1303174 0.3069082 0.4422873
```

```
2 0.05771683 0.08356470 0.1498126 0.3182912 0.3906147
3 0.03680951 0.05064075 0.1429711 0.3405590 0.4290196
```

Table 5

From above predicted probabilities, it is highly significant that customer with Travel Type as Business or Other will give any overall rating from 1-5 based on their experience during the stay. However for Leisure customer the odds of giving any overall rating is highly significant for rating 2 to 5 with the exception of overall rating 1.

Part 2

Now, I am analyzing if any overall rating is being derived from specific rating for Rooms, Service, Cleanliness and Location:

From initial analysis:

```
> with(data2012, table(Overall,Rooms))
       Rooms
Overall     1     2     3     4     5
      1  2894   744   717   162    60
      2  1098  2497  1753   529   123
      3   210  2136  9361  3009   734
      4    14   243  6599 21346  8591
      5    28    13   639  7977 39654
> with(data2012, table(Overall,Service))
       Service
Overall     1     2     3     4     5
      1  3372   677   560   107    48
      2  1449  1936  1842   604   211
      3   572  1815  6859  4596  1643
      4    72   384  4440 18073 13870
      5    25    20   444  5285 42944
> with(data2012, table(Overall,Cleanliness))
       Cleanliness
Overall     1     2     3     4     5
      1  2532   757   909   289   132
      2  1049  1589  2059  1011   335
      3   257  1497  6049  5821  1909
      4    17   231  3162 17565 15969
      5    20    12   307  4767 43535
> with(data2012, table(Overall,Location))
       Location
Overall     1     2     3     4     5
      1   900   399  1394  1144   764
      2   307   559  1594  2192  1393
      3   180   902  3453  5836  5159
      4    34   370  3705 11987 20839
      5    26    61  1238  6566 40757
```

Table 6

From the counts from above table, it seems that overall rating is correlated to specific rating.

```
>testRatings<-multinom(Overall~Rooms+Service+Cleanliness+Location, data=data2
012)
```

```
> summary(testRatings)
```

```
Coefficients:
  (Intercept)     Rooms   Service Cleanliness  Location
2   -4.196452 0.5557777 0.9261899   0.3579523 0.2332736
3  -10.581704 1.4525173 1.8592216   0.7683542 0.4723621
4  -23.444151 2.8536718 2.9624731   1.3640988 0.9874256
5  -44.280134 4.4816612 4.4608141   2.2247127 1.6031421
```

Table 7

From multinomial logistic regression equation, if y represents Overall rating:

$$\log(y = i) = \log\left(\frac{p(y=i)}{1-(p=i)}\right) = \beta_{i0} + \beta_1 x_{i2} + \beta_2 x_{i3} + \beta_2 x_{i3} + \beta_2 x_{i5} \text{ for } i = 1..5$$

After plugging in these numbers to above equation:

$$\log(y = 2) = \log\left(\frac{p(y=2)}{1-(p=2)}\right) = -4.2 + 0.56x_{22} + 0.92x_{23} + 0.36x_{24} + 0.24x_{25}$$

$$\log(y = 3) = \log\left(\frac{p(y=3)}{1-(p=3)}\right) = -10.58 + 1.45x_{32} + 1.85x_{33} + -0.78x_{24} + 0.47x_{25}$$

$$\log(y = 4) = \log\left(\frac{p(y=4)}{1-(p=4)}\right) = -23.44 + 2.85x_{42} + 2.96x_{43} + 1.36x_{24} + 0.99x_{25}$$

$$\log(y = 5) = \log\left(\frac{p(y=5)}{1-(p=5)}\right) = -44.28 + 4.48x_{52} + 4.46x_{53} + 2.22x_{24} + 1.6x_{25}$$

Focusing on Overall rating of 5, to achieve customer satisfaction up to level of 5, every unit increase in rating for Room ratings increase the odds of overall rating of 5 by 4.48. Also every increase in either Room or Service will increase Overall rating twice as more as Cleanliness and even much more than the increase in rating for Location.

Further, I have analyzed the data for Rooms ratings to check if Overall ratings are correlated.

Rooms ratings analysis using predicted probability :

```
> data2012$RoomsF<-factor(data2012$Rooms)
> data2012$RoomsFL<-relevel(data2012$RoomsF,ref="1")
> testRoomsRatings<-multinom(Overall~RoomsFL, data=data2012)
> summary(testRoomsRatings)
> dRooms<-data.frame(RoomsFL=c("1", "2", "3","4","5"))
> predict(testRoomsRatings, newdata=dRooms,"probs")
```

```
Rooms Rating (Horizontal/Rows) vs Overall Rating(Vertical/Columns)
            1           2          3           4           5
1 0.681895308 0.258704966 0.04950673 0.003282008 0.006610983
2 0.132074960 0.443291825 0.37921229 0.043143080 0.002277842
3 0.037602939 0.091935435 0.49091101 0.346041945 0.033508675
4 0.004902726 0.016017925 0.09112169 0.646394092 0.241563565
5 0.001216609 0.002497559 0.01492373 0.174753521 0.806608581
```

Table 8

From above table of predicted probablity for odds of getting Overall ratings based on Rooms rating shows that Overall ratings is almost correlated to Rooms ratings.

I have run the similar test for rest of the three categories as following:

<u>Service ratings analysis using predicted probability:</u>

```
Service Ratings (Horizontal/Rows) vs Overall Ratings (Vertical/Columns)

             1           2          3          4           5
1 0.6141544072 0.263901508 0.10415332 0.01313561 0.004655151
2 0.1400297165 0.400618103 0.37559855 0.07958023 0.004173398
3 0.0396092950 0.130206072 0.48488662 0.31388469 0.031413320
4 0.0037370170 0.021067915 0.16034780 0.63048809 0.184359180
5 0.0008190122 0.003594376 0.02798677 0.23624137 0.731358479
```

Table 9

<u>Cleanliness ratings analysis using predicted probability:</u>

```
Cleanliness Ratings (Horizontal/Rows) vs Overall Ratings (Vertical/Columns)

            1           2          3           4           5
1 0.653458708 0.270728471 0.06633285 0.004330126 0.005149843
2 0.185304616 0.388878742 0.36639248 0.056528088 0.002896079
3 0.072743569 0.164894458 0.48446408 0.253285863 0.024612033
4 0.009807355 0.034325873 0.19763939 0.596375021 0.161852363
5 0.002132528 0.005413589 0.03085300 0.258067722 0.703533165
```

Table 10

<u>Location ratings analysis using predicted probability:</u>

```
Location Rating (Horizontal/Rows) vs Overall Ratings (Vertical/Columns)

           1          2          3          4          5
1 0.62197631 0.21216578 0.12440523 0.02348106 0.01797162
2 0.17419059 0.24401835 0.39370339 0.16145226 0.02663541
3 0.12245942 0.14003507 0.30329811 0.32549467 0.10871274
4 0.04127400 0.07905654 0.21051345 0.43233962 0.23681639
5 0.01108378 0.02021359 0.07486039 0.30239979 0.59144244
```

Table 11

As I analyzed that predicted probability distributions for Overall ratings in regards to Rooms, Service, Cleanliness and Location ratings, it is evident that the odds of getting overall ratings is correlated to specific category ratings. Overall ratings for 1 and 5 are highly correlated to respective ratings in each category. Interestingly, the odds of getting Overall ratings of 3 or 4 is significant for each rating of Location.

Conclusion:

I have analyzed the dataset for customer review ratings for 2012 from Trip Advisor as collected by DAIS. From the analysis, I conclude following two problems:

- It is evident that customers with travel type of Business or Others will equally likely to give overall ratings from 1-5 based on their experience during the stay. Leisure customers will give a ratings from 2 to 5.
- Overall customer ratings is correlated to individual ratings for Rooms, Service and Cleanliness. In case of Location, odds of getting Overall ratings of 3 for any ratings of Location is high.

This dataset mostly include hotels from USA and Europe, so I will infer that given the facts above hotels in USA and Europe can achieve higher customer review rating by focusing more on Services like cleanliness, customer service and Rooms. Business Travelers are more independent in providing feedback and ratings based on their experience during the stay.

References:

Data: http://times.cs.uiuc.edu/~wang296/Data/ - TripAdvisor Dataset (JSON)

http://www.ats.ucla.edu/stat/r/dae/mlogit.htm - To understand and run multinomial regression analysis in R

https://www.youtube.com/watch?v=fDjKa7yWk1U – Tutorial to factor/relevel numerical categories

Python scripts to evaluate data and add TravelType as Leisure, Business or Other

Database to hold data and reformat for analysis: MongoDB.

Class Lectures – MSDS 6372