Hotel Review in Europe

Data Science in R

Spring 2019

Team MLB: Xiao Lu, Yange Bian, Luwen Mai, Yuyang Li

Abstract

Technological advancement makes online reviews the "digital word of mouth." Before their first purchases of certain products and services, consumers tend to rely on previous purchasing experiences of others to gather information and to reduce uncertainty. Hotel booking is no exception. Travelers, nowadays, have established the habits of looking up online reviews on the hotels before booking. This trend motivates our study. We would like to see if there is a correlation between keywords in online reviews and scores that hotels receive. The predictive power implied in this correlation could be a useful tool for hotels to cater to different needs of their customers, especially when cultural differences are at play.

Using multiple linear regressions, we are able to uncover factors that influence travelers' decisions of leaving positive or negative comments and scores, such as how many days it takes the travelers to leave their comments. Apart from that, we also see that travelers' nationalities, which indicate cultural differences, do have significant impacts on online review behaviors. What's more, we identified a few keywords that have strong predictive powers on review scores. Lastly, we analyzed the correlation by adding an interaction term and through spatial graph.

Table of Contents

Introduction

Online reviews are becoming increasingly important, especially for hotels. Before social media becomes a social trend, people have used to rely on word of mouth to gather information - not to mention that word of mouth "has the potential to influence customer purchase decisions" (Sparks and Browning, 2011). With technological advancements, consumers are paying more attention on online reviews so as to make informed decisions, including the time to book hotels. Studies have shown that online reviews have the power to influence traveler's booking decisions. In their study in 2004, Hennig-Thurau, Gwinner, Walsh, and Gremler have found that people nowadays have switched to gathering information from "electronic word of mouth (eWOM)," which enables people to view the advantages and disadvantages of a product or service and thus influences a firm's reputation (Hennig-Thurau et al., 2004). Supporting their claim, according to an experimental study by Sparks and Browning (2011), travelers are "more influenced by early negative information, especially when the overall set of reviews is negative." They also find that consumers "like to rely on easy-to-process information" when booking hotels (Sparks and Browning, 2011).

Other studies have shown that online reviews have significant impacts on consumers' booking intents and perceptions of trust. Sichtmann (2007) finds that consumers' willingness to book is heavily dependent on whether they trust a certain hotel, and hence marketers need to reduce uncertainty and build trust in their products and services. The component of consumers' trust, according to Johnson and Grayson (2005), could be consumer satisfaction in previous purchase experience, which serves as the building block for consumer confidence and future willingness to trust. Arguably, the same logic could be applied to reading online reviews before booking hotels, since there is no previous experience to rely on, travelers want to gather information of other people's purchase and interaction experiences, so as to reduce uncertainty and build up cognitive trust towards the hotels.

In terms of choosing hotels, cultural differences will affect travelers' opinions on what the important features of hotels are. Since cultural differences are often translated into different expectations to services, it is vital for "service providing companies such as hotels to realize that customer preferences are not identical all around the world" (Seo, 2012; Ueltschy et al., 2007). The study by Yeji Seo (2012) has shown that travelers from the United States, China and Japan have very different rankings in terms of satisfaction attributes of hotels. As the travelers from the United States have ranked "customization of services" and "reliability of services" as the two most important attributes. Chinese travelers have put more focus on "cleanliness of rooms" and "employee attentiveness to customers' needs." In the meantime, Japanese travelers pay more attention to "quick and efficient problem-solving skills" and "ability to efficiently cater to customer needs" (Seo, 2012). From this paper, we also find that Chinese and Japanese travelers have more similar rankings for satisfaction attributes, and this observation prompts us to categorize countries in our dataset into several geographic areas.

Cultural differences also influence consumer complaint behavior (CCB). Hofstede (1991) distinguishes power distance, long-term orientation, individualism–collectivism, masculinity–femininity and uncertainty avoidance as major factors influencing CCB. Specifically, power distance refers to “amount of respect and deference between those in superior and subordinate positions,” while masculinity-femininity refers to “relative emphasis on the achievement, interpersonal harmony which characterizes gender differences in some national cultures,” and individualism-collectivism refers to “whether one’s identity is defined by personal choices and achievements or by the character of the collective groups to which one is more or less permanently attached” (Yuksel, et al., 2006). Hofstede (1991) provides example of analysis based on these factors. He states that Asians tend to have high power distance and low uncertainty avoidance with masculinity and collectivism characteristics. This means that Asians have more respect on hierarchy and they have little concern over uncertainty. They view success as gaining social status and wealth and they pay more attention to the benefit of the group. However, Anglo-Saxon societies score low in power distance and have a high score with masculinity and individualism characteristics (Hofstede, 1991).

Data Description

The dataset we use is originally adapted from *kaggle.com*, and the data were collected from and owned by *booking.com*. This dataset has 17 variables, with 515,000 reviews and scores of 1,493 luxury hotels located within 6 countries in Europe. It also provides nationalities on travelers who write those reviews. To simplify our analysis, we choose to only look at the top 10 countries where most travelers are from. We also transform a few variables so that we can obtain the information we need. First, we extract the number of days for travelers to make reviews from “days_since_review” and transform it from character into numeric observations. In the original dataset, the location country of hotels are indicated by attached characters at the end of variable “Hotel_Address.” To have easier access to hotel locations, we separate this variable and set up a new variable “Hotel_Country” that only contains information of which country hotels are located at. Special attention is given to United Kingdom, we replace the original indication “Kingdom” by “UK.” For the purpose of controlling for cultural differences, we use variable “Reviewer_Nationality” to categorize travelers’ nationalities (the Top 10 tourist countries) by six geographic areas: “UK_IR” for United Kingdom and Ireland, “NE_SW_GE” for Netherlands, Switzerland and Germany, “Middle_East” for United Arab Emirates and Saudi Arabia, “US_CA” for United States of America and Canada, and “AUS” for Australia. Lastly, since we want to investigate review contents of both negative and positive reviews, we rank the frequency of occurrence of all words and pick out the top 5 keywords for negative and positive comments, respectively. To reduce noise in word frequencies, we de-capitalize words in all reviews and get rid of stopwords, numbers, punctuations, and whitespaces. We also exclude words that are meaningless to our analysis, such as “i,” “when,” and “one.”

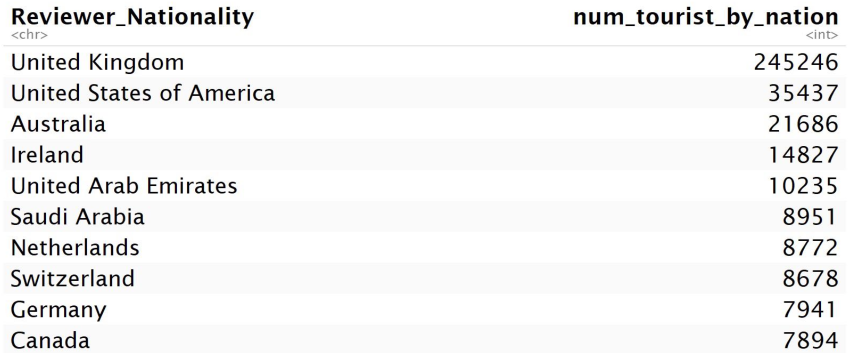
Methods

We use extensive multiple linear regressions in the analysis. In the first regression (Figure 7), the dependent variable is the review score to a certain hotel, given by a traveler. We regression this variable on the ratio between the number of negative reviews (score below 7) and total reviews (negative_ratio), the number of days the traveler takes to write the review (days_since_review), the word count difference between negative and positive comments (difference), and the total number of reviews a traveler gives. We hypothesize that there is a negative correlation between review scores to each of the variables mentioned above.

The second and third multiple linear regressions (Figure 16 and Figure 21) are about the influence of keywords on negative scores and positive scores, respectively. Apart from the dummy variables set up for keywords, we also have control variables that represent areas where the travelers come from and country dummies indicating which country the hotel is located at. Lastly, we include a control variable about how long it takes the travelers to write the reviews (days_since_review).

The fourth and fifth models are multiple linear regression model with interaction term (Figure 26) and spatial graph (Figure 28), respectively. In the fourth model, We add the interaction term Review_Total_Negative_Word_Counts:countrydummy. We also conduct the spatial analysis in order to analyze the topic from a different perspective.

Data Visualization



Reviewer_Nationality <chr>	num_tourist_by_nation <int>
United Kingdom	245246
United States of America	35437
Australia	21686
Ireland	14827
United Arab Emirates	10235
Saudi Arabia	8951
Netherlands	8772
Switzerland	8678
Germany	7941
Canada	7894

Figure 1. Top 10 Tourists Countries in Terms of Number of Travelers

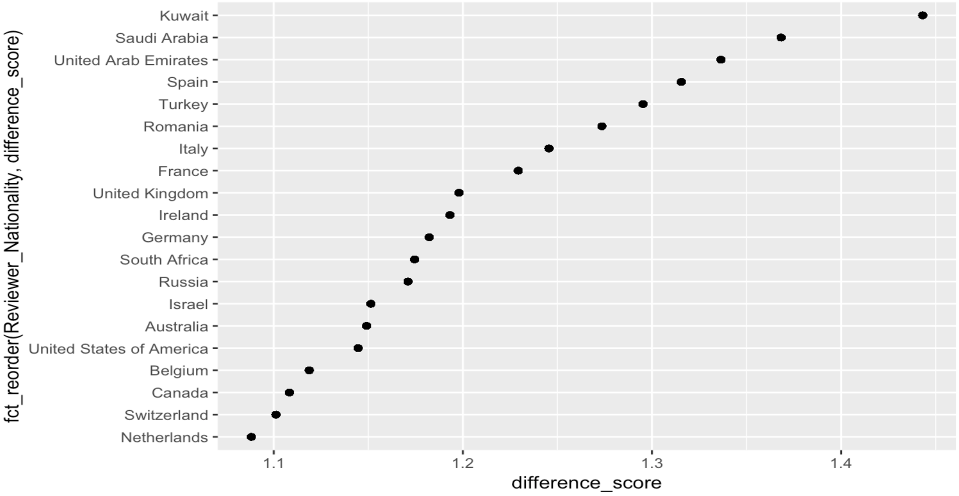


Figure 2. Scatterplot for the absolute difference between last year's and this year's average review scores among the Top 20 source-country of tourists

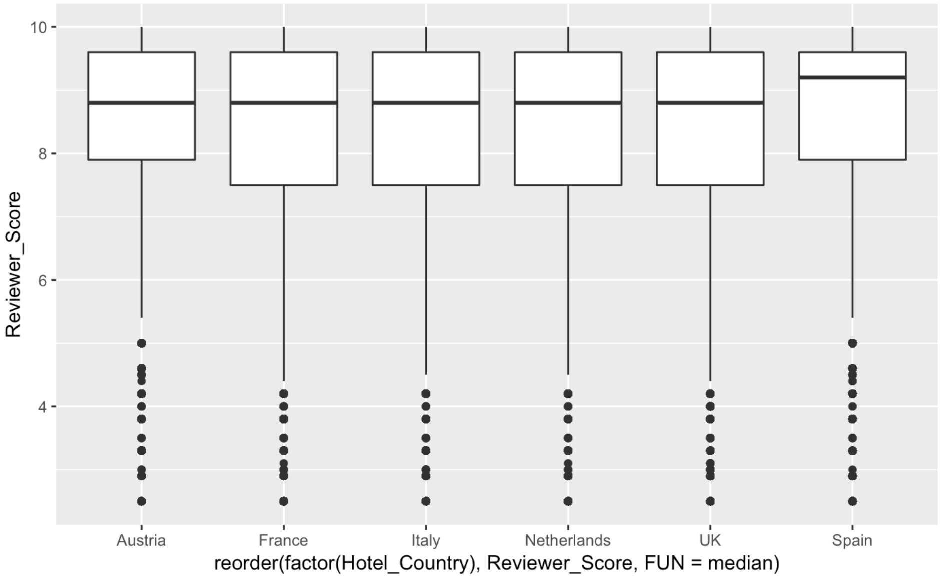


Figure 3. Boxplot of review scores for hotels in 6 countries, ordered by median value

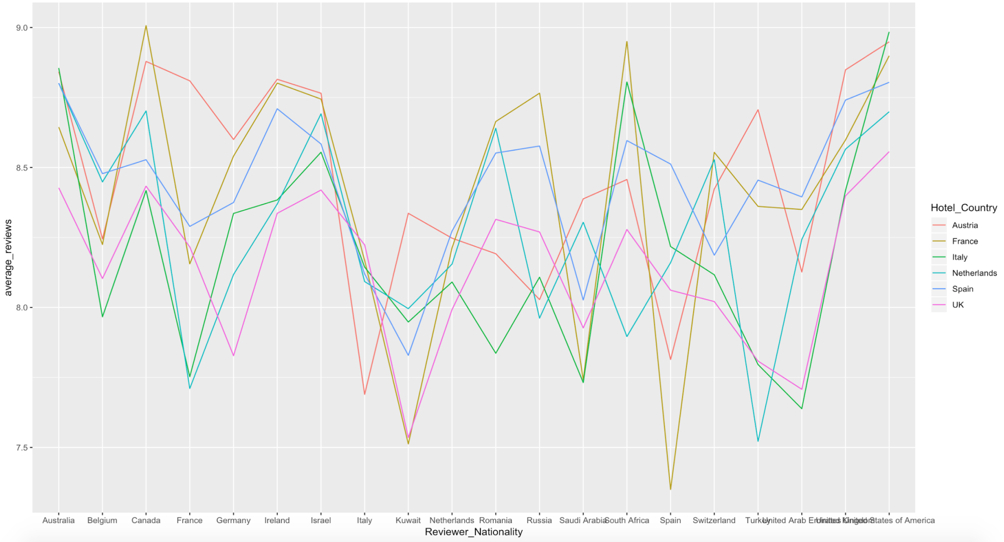


Figure 4. Line graph for the mean review score for Top 20 tourist countries, colored by the countries of hotels

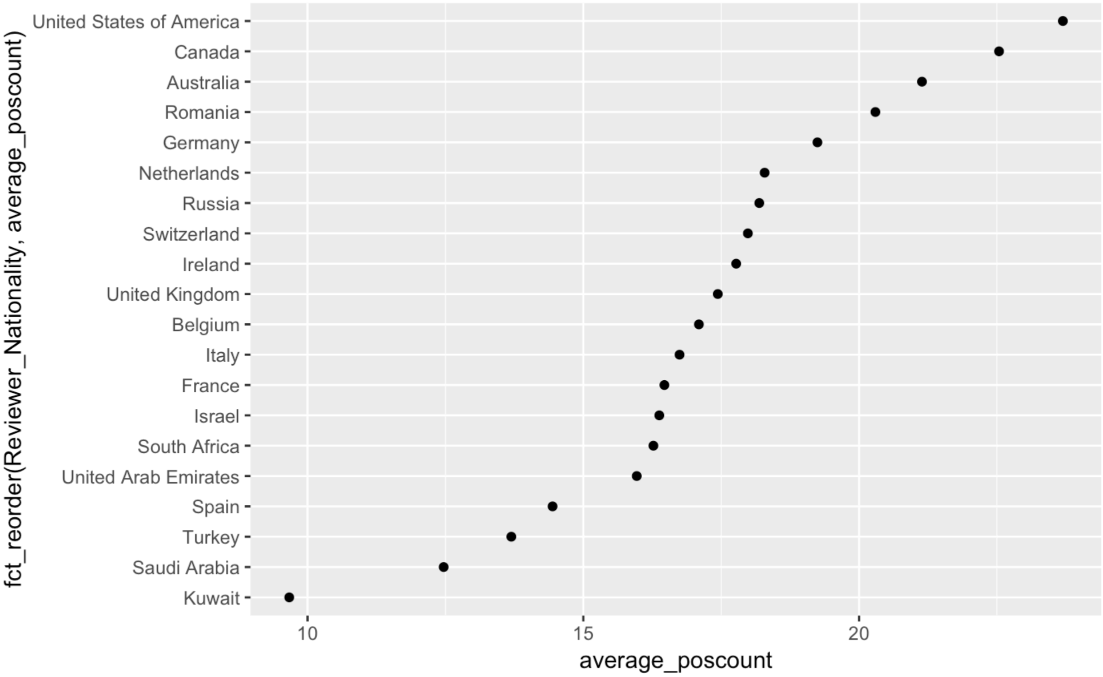


Figure 5. Scatterplot for the mean Total Positive Word Count for Top 20 tourist countries, ordered by the overall median

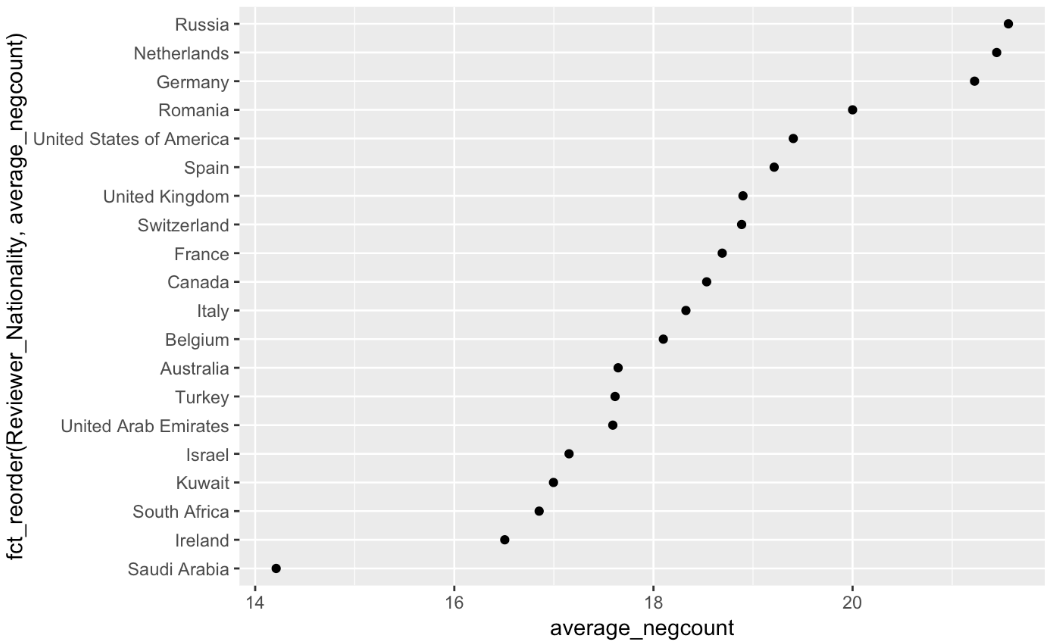
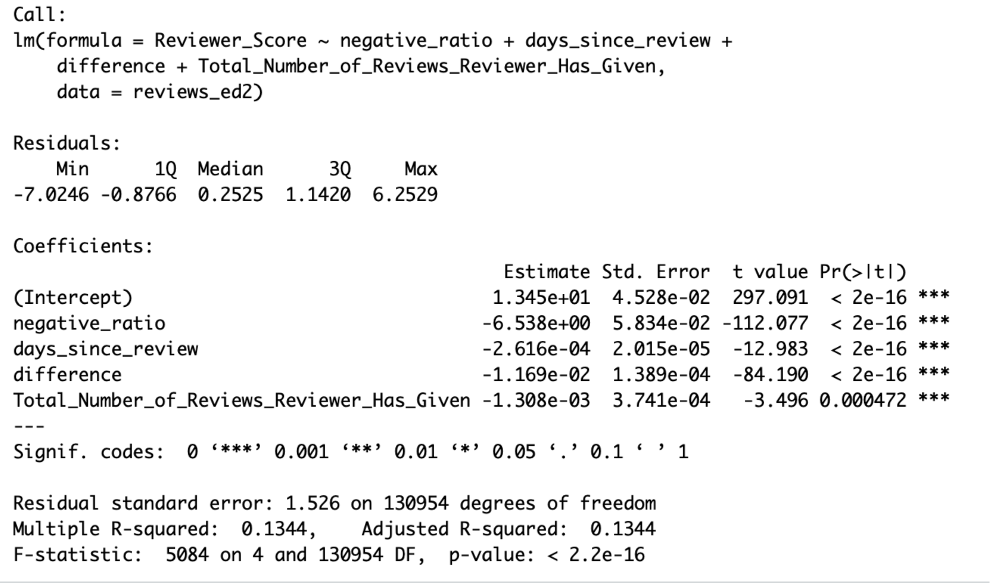


Figure 6. Scatterplot for the mean Total Negative Word Count for Top 20 tourist countries, ordered by the overall median

Results

Multiple Linear Regression #1

$$\widehat{Y} = 0.1345 - 6.538X_1 - 0.0002616X_2 - 0.01169X_3 - 0.001308X_4$$



```
Call:
lm(formula = Reviewer_Score ~ negative_ratio + days_since_review +
    difference + Total_Number_of_Reviews_Reviewer_Has_Given,
    data = reviews_ed2)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0246 -0.8766  0.2525  1.1420  6.2529

Coefficients:
                                             Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                 1.345e+01  4.528e-02  297.091  < 2e-16 ***
negative_ratio                             -6.538e+00  5.834e-02 -112.077  < 2e-16 ***
days_since_review                          -2.616e-04  2.015e-05  -12.983  < 2e-16 ***
difference                                 -1.169e-02  1.389e-04  -84.190  < 2e-16 ***
Total_Number_of_Reviews_Reviewer_Has_Given -1.308e-03  3.741e-04   -3.496 0.000472 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.526 on 130954 degrees of freedom
Multiple R-squared:  0.1344,    Adjusted R-squared:  0.1344
F-statistic:  5084 on 4 and 130954 DF,  p-value: < 2.2e-16
```

Figure 7: Summary of Parameter Estimates

Linear Model Assumptions:

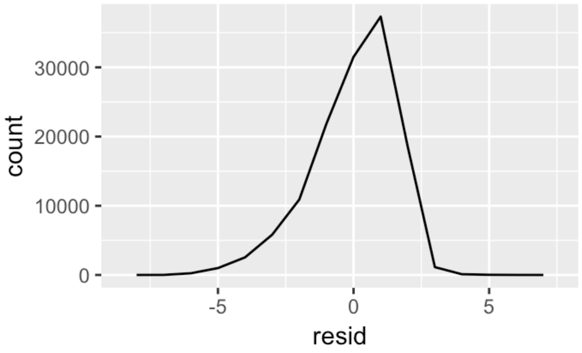


Figure 8: The Frequency Distribution of Residuals

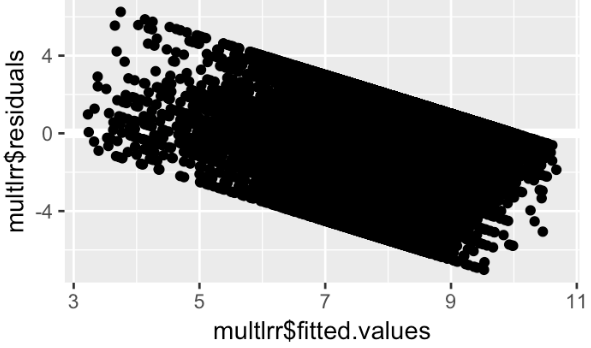


Figure 9: Residuals vs. Fitted Values

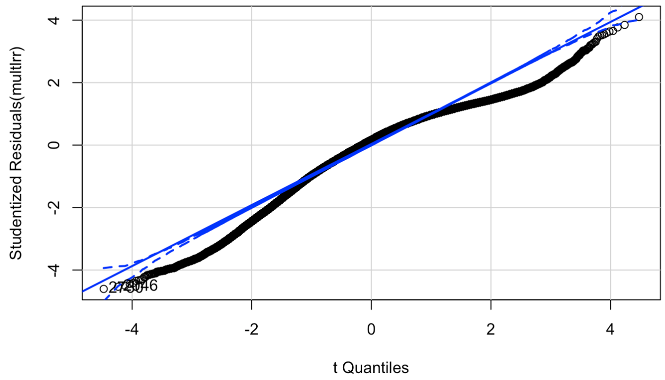


Figure 10: QQ Plot of Multiple Linear Regression #1

Regression Tree Modeling:

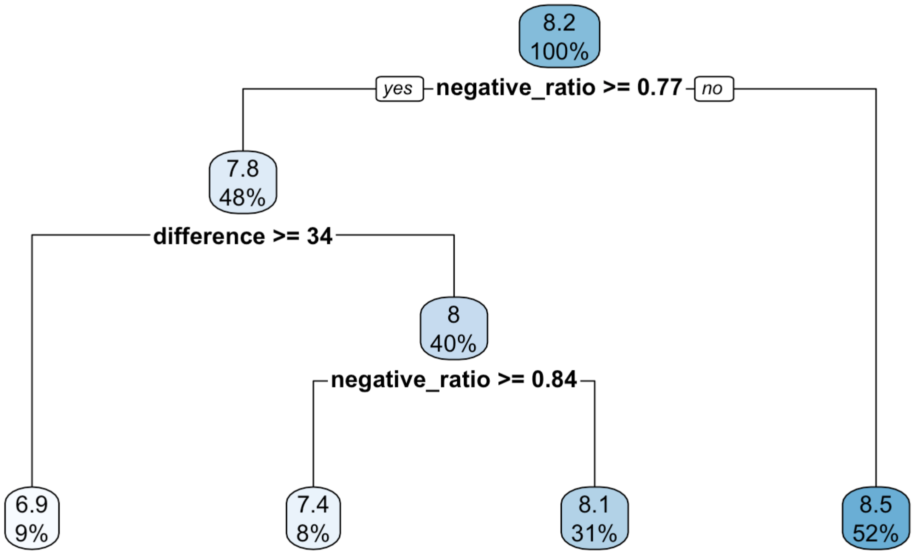


Figure 11: Regression Tree of Multiple Linear Regression #1

Multiple Linear Regression #2

Part 1: Negative Review Basic Info

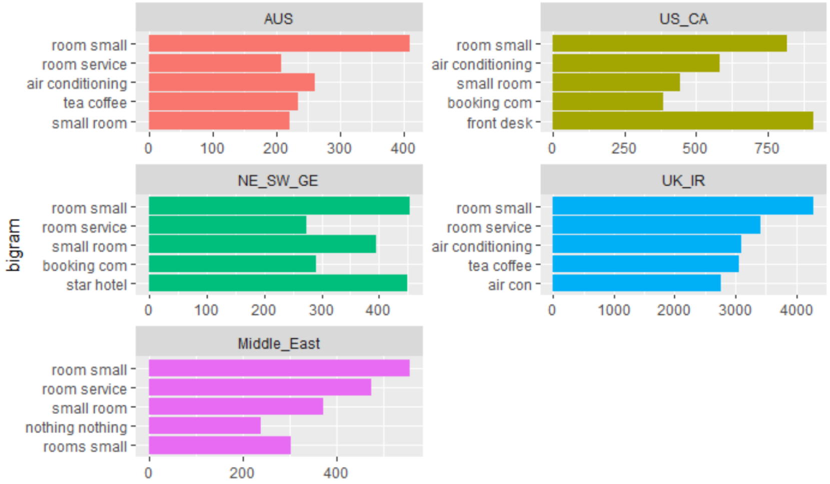


Figure 12. Group by Tourist Country

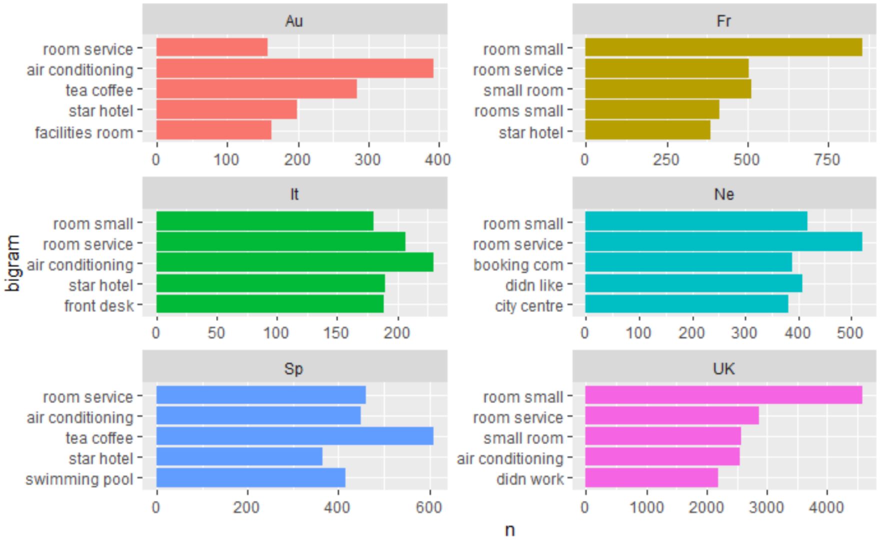
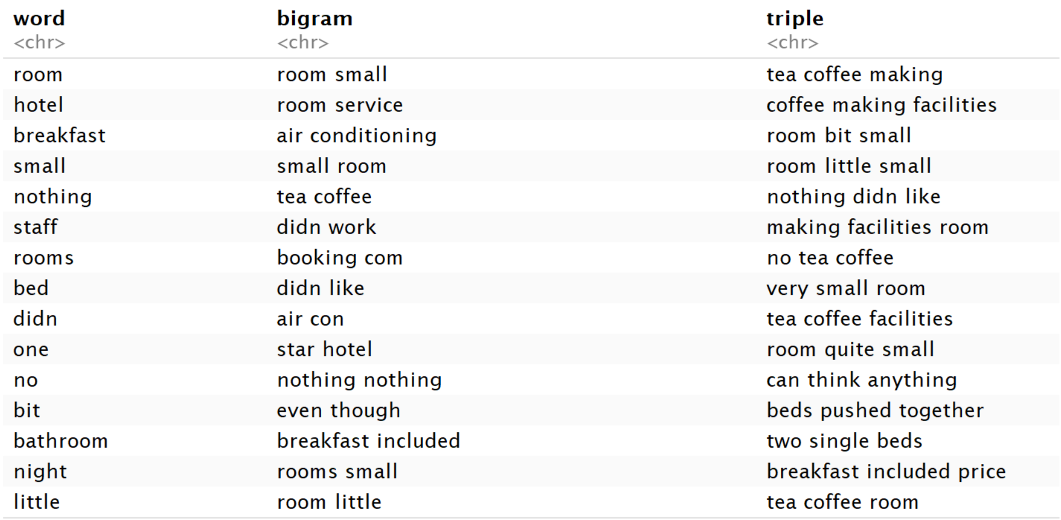


Figure 13. Group by Hotel Country



word <chr>	bigram <chr>	triple <chr>
room	room small	tea coffee making
hotel	room service	coffee making facilities
breakfast	air conditioning	room bit small
small	small room	room little small
nothing	tea coffee	nothing didn like
staff	didn work	making facilities room
rooms	booking com	no tea coffee
bed	didn like	very small room
didn	air con	tea coffee facilities
one	star hotel	room quite small
no	nothing nothing	can think anything
bit	even though	beds pushed together
bathroom	breakfast included	two single beds
night	rooms small	breakfast included price
little	room little	tea coffee room

Table 1. Comparison of Top 15 Highest Frequency Word and Bigram and Three-Words in Negative Review

Part 2: Positive Review Basic Info

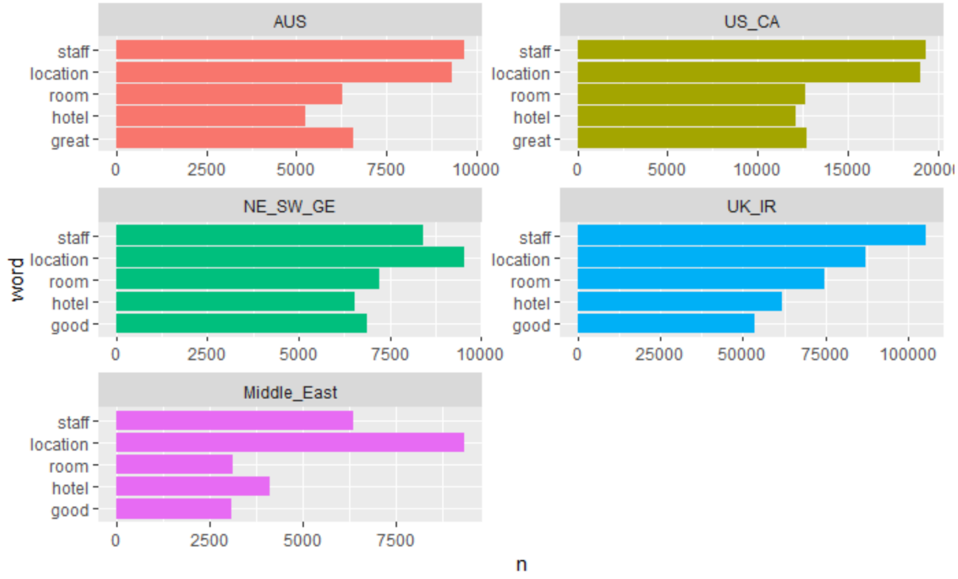


Figure 14. Group by Tourists Country

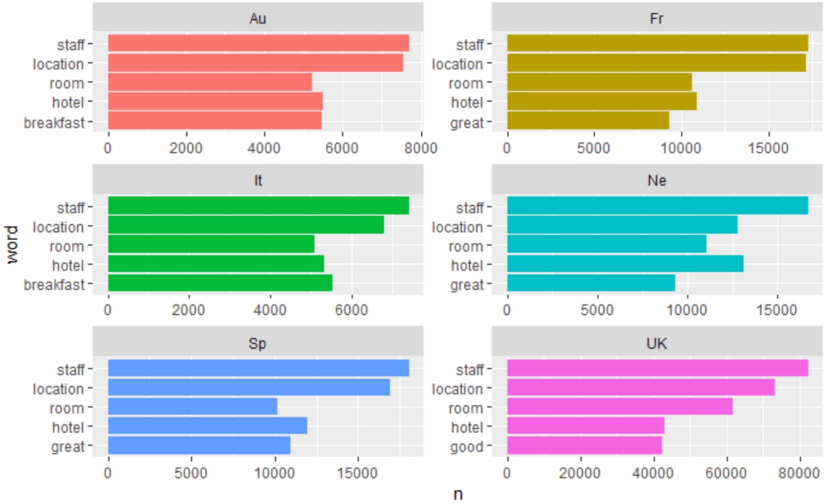
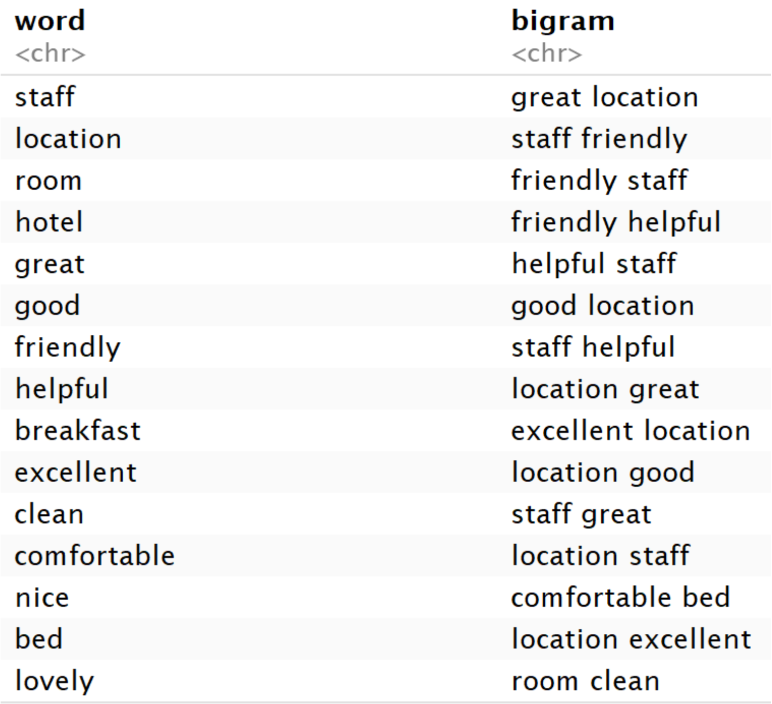


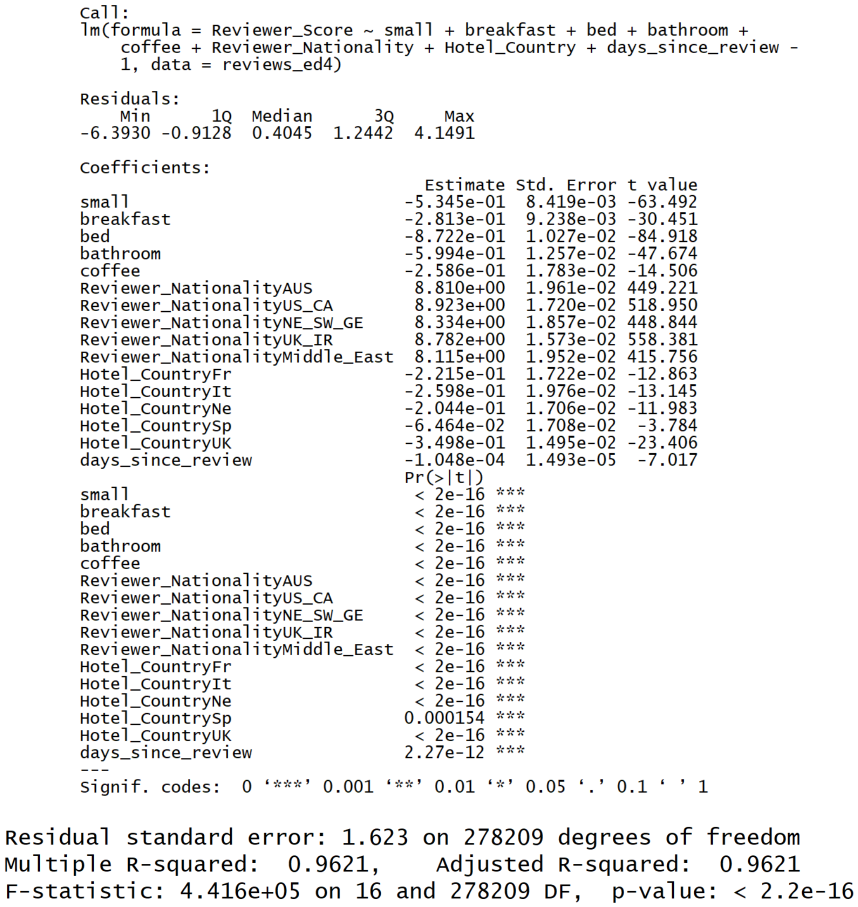
Figure 15. Group by the Hotel's Country



word <chr>	bigram <chr>
staff	great location
location	staff friendly
room	friendly staff
hotel	friendly helpful
great	helpful staff
good	good location
friendly	staff helpful
helpful	location great
breakfast	excellent location
excellent	location good
clean	staff great
comfortable	location staff
nice	comfortable bed
bed	location excellent
lovely	room clean

Table 2. Comparison of Top 15 Highest Frequency Word and Bigram in Positive Review

Part 3: Multiple Linear Regression for Negative Words



```
Call:
lm(formula = Reviewer_Score ~ small + breakfast + bed + bathroom + 
    coffee + Reviewer_Nationality + Hotel_Country + days_since_review - 
    1, data = reviews_ed4)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3930 -0.9128  0.4045  1.2442  4.1491 

Coefficients:
                                  Estimate Std. Error t value
small                           -5.345e-01  8.419e-03 -63.492
breakfast                       -2.813e-01  9.238e-03 -30.451
bed                             -8.722e-01  1.027e-02 -84.918
bathroom                        -5.994e-01  1.257e-02 -47.674
coffee                          -2.586e-01  1.783e-02 -14.506
Reviewer_NationalityAUS          8.810e+00  1.961e-02 449.221
Reviewer_NationalityUS_CA        8.923e+00  1.720e-02 518.950
Reviewer_NationalityNE_SW_GE     8.334e+00  1.857e-02 448.844
Reviewer_NationalityUK_IR        8.782e+00  1.573e-02 558.381
Reviewer_NationalityMiddle_East  8.115e+00  1.952e-02 415.756
Hotel_CountryFr                 -2.215e-01  1.722e-02 -12.863
Hotel_CountryIt                 -2.598e-01  1.976e-02 -13.145
Hotel_CountryNe                 -2.044e-01  1.706e-02 -11.983
Hotel_CountrySp                 -6.464e-02  1.708e-02  -3.784
Hotel_CountryUK                 -3.498e-01  1.495e-02 -23.406
days_since_review               -1.048e-04  1.493e-05  -7.017
                                Pr(>|t|)    
small                            < 2e-16 ***
breakfast                        < 2e-16 ***
bed                              < 2e-16 ***
bathroom                         < 2e-16 ***
coffee                           < 2e-16 ***
Reviewer_NationalityAUS          < 2e-16 ***
Reviewer_NationalityUS_CA        < 2e-16 ***
Reviewer_NationalityNE_SW_GE     < 2e-16 ***
Reviewer_NationalityUK_IR        < 2e-16 ***
Reviewer_NationalityMiddle_East  < 2e-16 ***
Hotel_CountryFr                  < 2e-16 ***
Hotel_CountryIt                  < 2e-16 ***
Hotel_CountryNe                  < 2e-16 ***
Hotel_CountrySp                 0.000154 ***
Hotel_CountryUK                  < 2e-16 ***
days_since_review               2.27e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.623 on 278209 degrees of freedom
Multiple R-squared:  0.9621,    Adjusted R-squared:  0.9621 
F-statistic: 4.416e+05 on 16 and 278209 DF,  p-value: < 2.2e-16
```

Figure 16. Regression Result for Negative Words

Linear Model Assumptions:

Based on the residual and QQ plots below, the linear modeling assumptions are not violated.

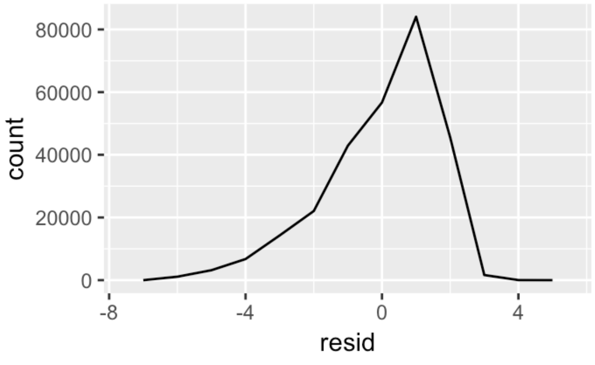


Figure 17. Frequency Distribution of Model Residuals

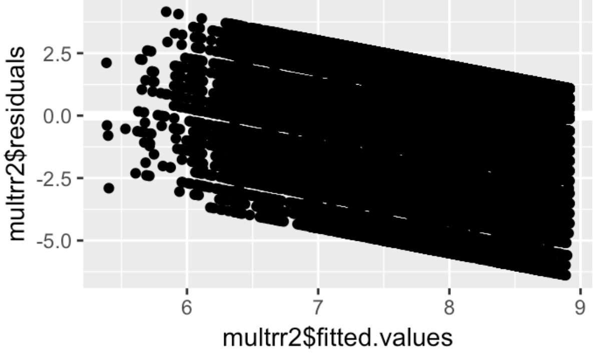


Figure 18. Residuals vs. Fitted Values

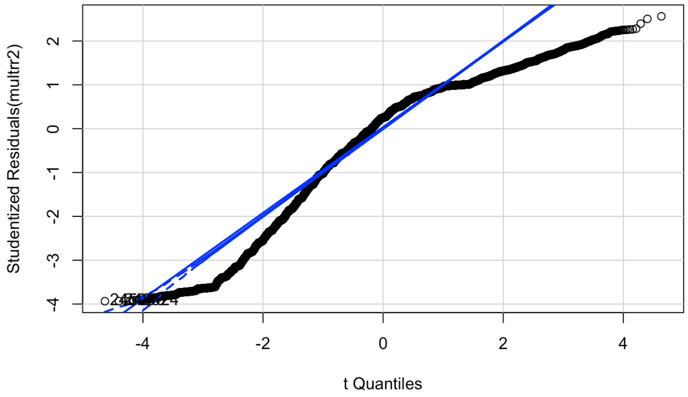
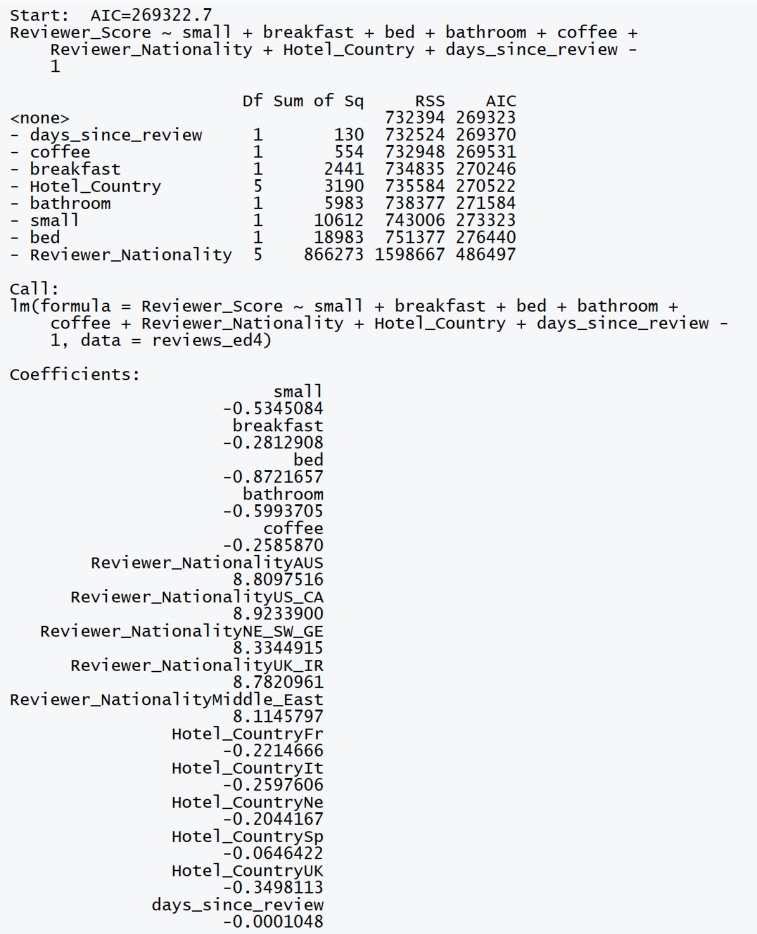


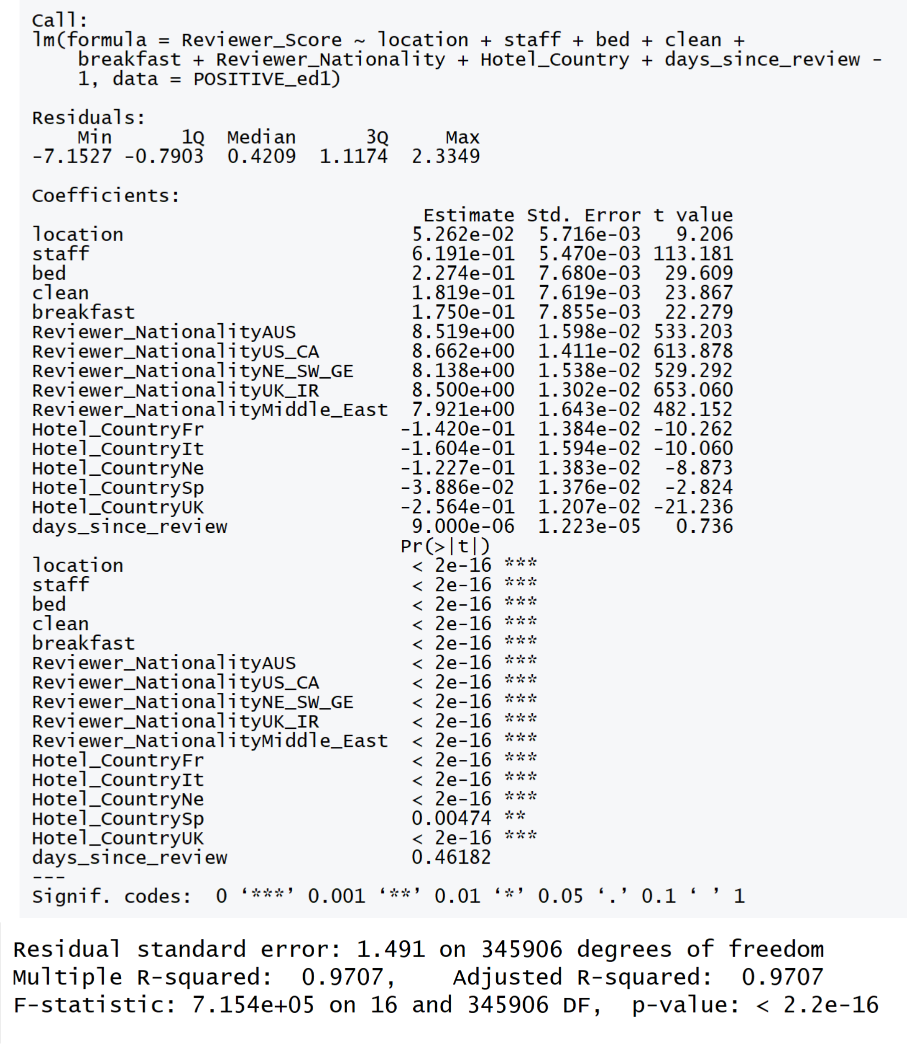
Figure 19. QQ plot of Regression Model



```
Start:  AIC=269322.7
Reviewer_Score ~ small + breakfast + bed + bathroom + coffee +
    Reviewer_Nationality + Hotel_Country + days_since_review -
    1

                       Df Sum of Sq     RSS    AIC
<none>                               732394 269323
- days_since_review     1       130  732524 269370
- coffee                1       554  732948 269531
- breakfast             1      2441  734835 270246
- Hotel_Country         5      3190  735584 270522
- bathroom              1      5983  738377 271584
- small                 1     10612  743006 273323
- bed                   1     18983  751377 276440
- Reviewer_Nationality  5    866273 1598667 486497

Call:
lm(formula = Reviewer_Score ~ small + breakfast + bed + bathroom +
    coffee + Reviewer_Nationality + Hotel_Country + days_since_review -
    1, data = reviews_ed4)

Coefficients:
                       small
                  -0.5345084
                   breakfast
                  -0.2812908
                         bed
                  -0.8721657
                    bathroom
                  -0.5993705
                      coffee
                  -0.2585870
        Reviewer_NationalityAUS
                   8.8097516
      Reviewer_NationalityUS_CA
                   8.9233900
   Reviewer_NationalityNE_SW_GE
                   8.3344915
      Reviewer_NationalityUK_IR
                   8.7820961
Reviewer_NationalityMiddle_East
                   8.1145797
               Hotel_CountryFr
                  -0.2214666
               Hotel_CountryIt
                  -0.2597606
               Hotel_CountryNe
                  -0.2044167
               Hotel_CountrySp
                  -0.0646422
               Hotel_CountryUK
                  -0.3498113
             days_since_review
                  -0.0001048
```

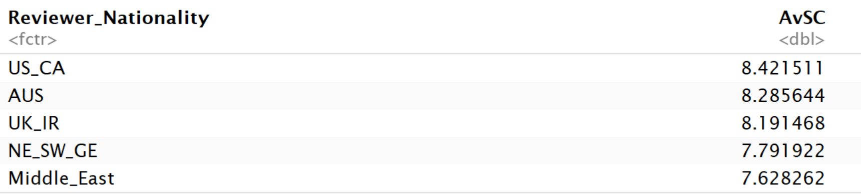
Figure 20. Stepwise Testing Procedure

Part 4: Multiple Linear Regression for Positive Words



```
Call:
lm(formula = Reviewer_Score ~ location + staff + bed + clean +
    breakfast + Reviewer_Nationality + Hotel_Country + days_since_review -
    1, data = POSITIVE_ed1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1527 -0.7903  0.4209  1.1174  2.3349

Coefficients:
                                  Estimate Std. Error t value
location                         5.262e-02  5.716e-03   9.206
staff                            6.191e-01  5.470e-03 113.181
bed                              2.274e-01  7.680e-03  29.609
clean                            1.819e-01  7.619e-03  23.867
breakfast                        1.750e-01  7.855e-03  22.279
Reviewer_NationalityAUS          8.519e+00  1.598e-02 533.203
Reviewer_NationalityUS_CA        8.662e+00  1.411e-02 613.878
Reviewer_NationalityNE_SW_GE     8.138e+00  1.538e-02 529.292
Reviewer_NationalityUK_IR        8.500e+00  1.302e-02 653.060
Reviewer_NationalityMiddle_East  7.921e+00  1.643e-02 482.152
Hotel_CountryFr                 -1.420e-01  1.384e-02 -10.262
Hotel_CountryIt                 -1.604e-01  1.594e-02 -10.060
Hotel_CountryNe                 -1.227e-01  1.383e-02  -8.873
Hotel_CountrySp                 -3.886e-02  1.376e-02  -2.824
Hotel_CountryUK                 -2.564e-01  1.207e-02 -21.236
days_since_review                9.000e-06  1.223e-05   0.736
                                Pr(>|t|)
location                         < 2e-16 ***
staff                            < 2e-16 ***
bed                              < 2e-16 ***
clean                            < 2e-16 ***
breakfast                        < 2e-16 ***
Reviewer_NationalityAUS          < 2e-16 ***
Reviewer_NationalityUS_CA        < 2e-16 ***
Reviewer_NationalityNE_SW_GE     < 2e-16 ***
Reviewer_NationalityUK_IR        < 2e-16 ***
Reviewer_NationalityMiddle_East  < 2e-16 ***
Hotel_CountryFr                  < 2e-16 ***
Hotel_CountryIt                  < 2e-16 ***
Hotel_CountryNe                  < 2e-16 ***
Hotel_CountrySp                  0.00474 **
Hotel_CountryUK                  < 2e-16 ***
days_since_review                0.46182
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.491 on 345906 degrees of freedom
Multiple R-squared:  0.9707,    Adjusted R-squared:  0.9707
F-statistic: 7.154e+05 on 16 and 345906 DF,  p-value: < 2.2e-16
```

Figure 21. Regression Result for Positive Words



Reviewer_Nationality <fctr>	AvSC <dbl>
US_CA	8.421511
AUS	8.285644
UK_IR	8.191468
NE_SW_GE	7.791922
Middle_East	7.628262

Table 3. Top 10 Countries’ Reviewer Average Score, Grouped by Region

Linear Model Assumptions:

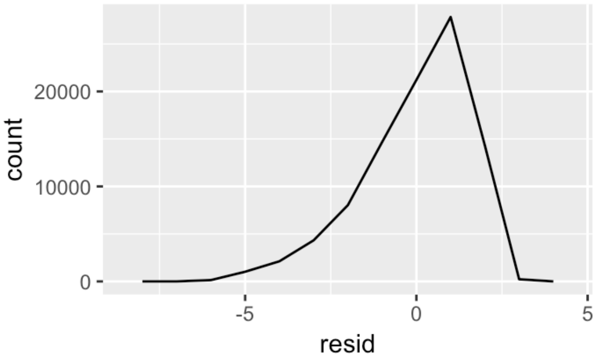


Figure 22. Frequency Distribution of Model Residuals

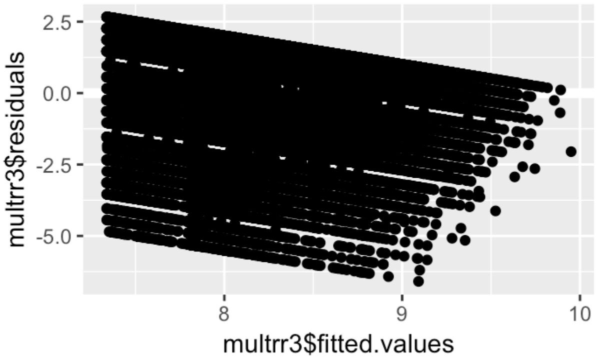


Figure 23. Residuals vs. Fitted Values

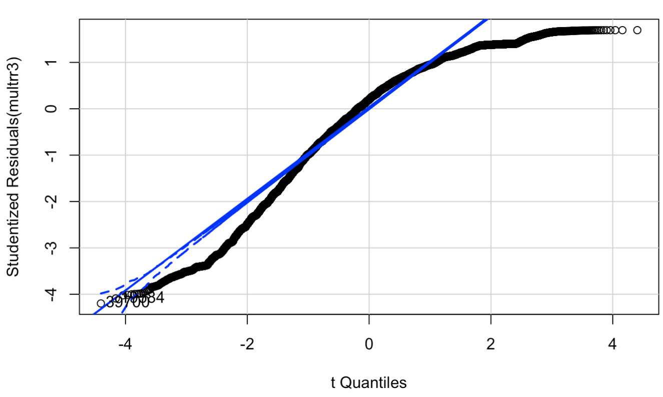
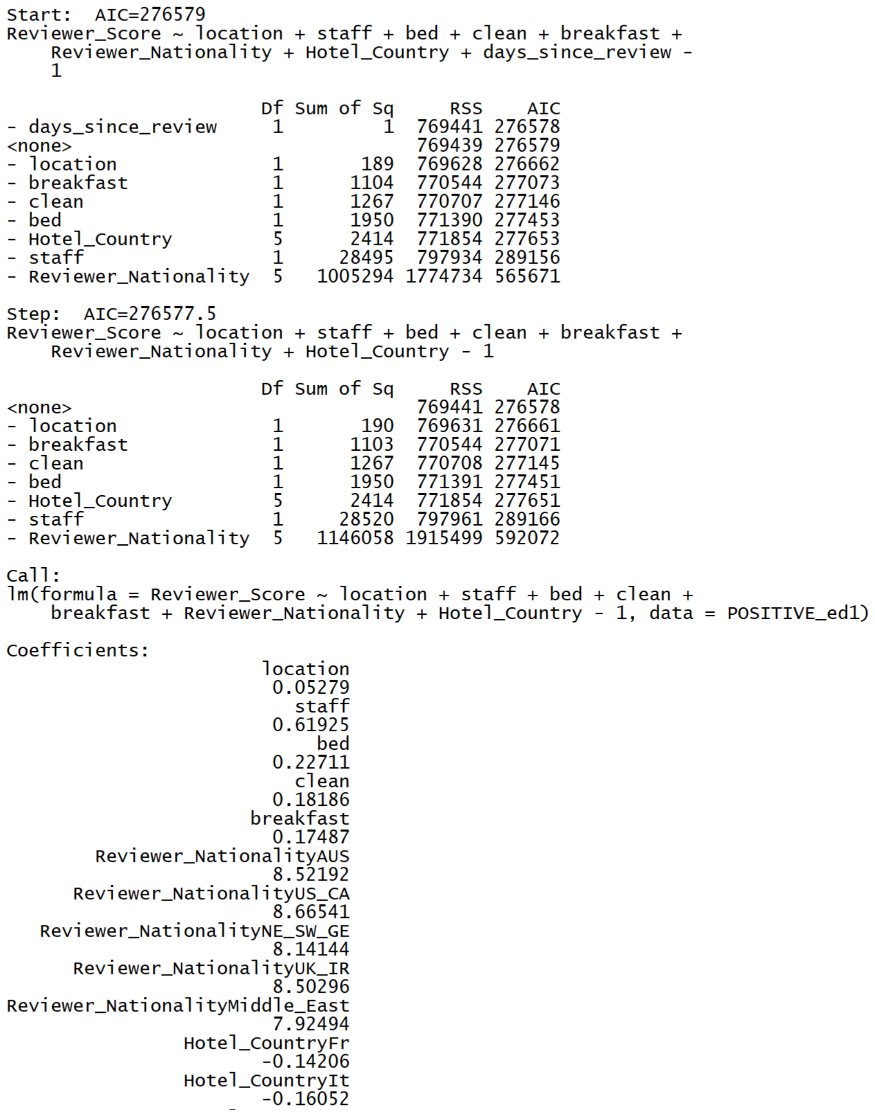


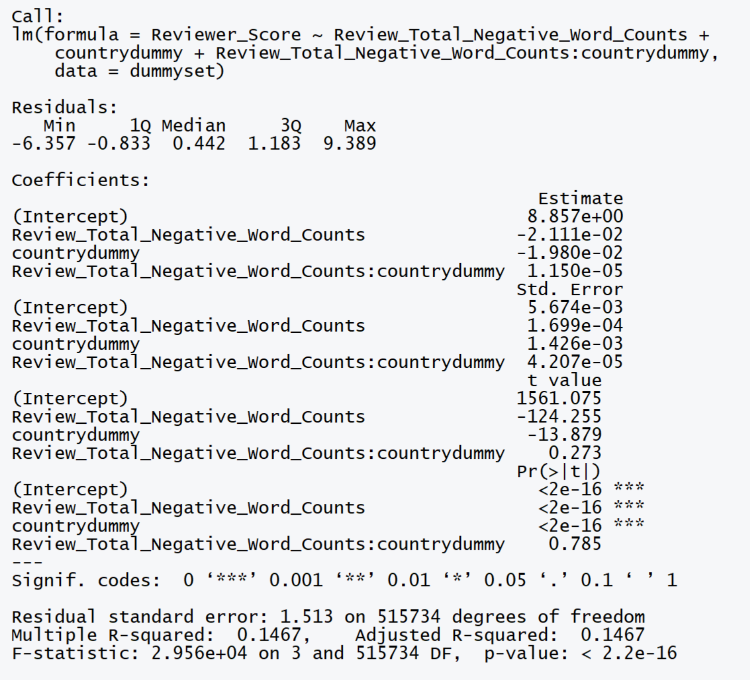
Figure 24. QQ plot of Regression Model



```
Start:  AIC=276579
Reviewer_Score ~ location + staff + bed + clean + breakfast + 
    Reviewer_Nationality + Hotel_Country + days_since_review - 
    1

                       Df Sum of Sq     RSS    AIC
- days_since_review     1         1  769441 276578
<none>                               769439 276579
- location              1       189  769628 276662
- breakfast             1      1104  770544 277073
- clean                 1      1267  770707 277146
- bed                   1      1950  771390 277453
- Hotel_Country         5      2414  771854 277653
- staff                 1     28495  797934 289156
- Reviewer_Nationality  5   1005294 1774734 565671

Step:  AIC=276577.5
Reviewer_Score ~ location + staff + bed + clean + breakfast + 
    Reviewer_Nationality + Hotel_Country - 1

                       Df Sum of Sq     RSS    AIC
<none>                               769441 276578
- location              1       190  769631 276661
- breakfast             1      1103  770544 277071
- clean                 1      1267  770708 277145
- bed                   1      1950  771391 277451
- Hotel_Country         5      2414  771854 277651
- staff                 1     28520  797961 289166
- Reviewer_Nationality  5   1146058 1915499 592072

Call:
lm(formula = Reviewer_Score ~ location + staff + bed + clean + 
    breakfast + Reviewer_Nationality + Hotel_Country - 1, data = POSITIVE_ed1)

Coefficients:
                       location  
                        0.05279  
                          staff  
                        0.61925  
                            bed  
                        0.22711  
                          clean  
                        0.18186  
                      breakfast  
                        0.17487  
        Reviewer_NationalityAUS  
                        8.52192  
      Reviewer_NationalityUS_CA  
                        8.66541  
   Reviewer_NationalityNE_SW_GE  
                        8.14144  
      Reviewer_NationalityUK_IR  
                        8.50296  
Reviewer_NationalityMiddle_East  
                        7.92494  
               Hotel_CountryFr  
                       -0.14206  
               Hotel_CountryIt  
                       -0.16052  
```

Figure 25. Stepwise Testing Procedure

Multiple Linear Regression with Interaction Term



```
Call:
lm(formula = Reviewer_Score ~ Review_Total_Negative_Word_Counts +
    countrydummy + Review_Total_Negative_Word_Counts:countrydummy,
    data = dummyset)

Residuals:
   Min     1Q Median     3Q    Max
-6.357 -0.833  0.442  1.183  9.389

Coefficients:
                                                Estimate
(Intercept)                                    8.857e+00
Review_Total_Negative_Word_Counts             -2.111e-02
countrydummy                                  -1.980e-02
Review_Total_Negative_Word_Counts:countrydummy  1.150e-05
                                              Std. Error
(Intercept)                                    5.674e-03
Review_Total_Negative_Word_Counts              1.699e-04
countrydummy                                   1.426e-03
Review_Total_Negative_Word_Counts:countrydummy  4.207e-05
                                               t value
(Intercept)                                   1561.075
Review_Total_Negative_Word_Counts             -124.255
countrydummy                                   -13.879
Review_Total_Negative_Word_Counts:countrydummy    0.273
                                              Pr(>|t|)
(Intercept)                                     <2e-16 ***
Review_Total_Negative_Word_Counts               <2e-16 ***
countrydummy                                    <2e-16 ***
Review_Total_Negative_Word_Counts:countrydummy   0.785
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.513 on 515734 degrees of freedom
Multiple R-squared:  0.1467,    Adjusted R-squared:  0.1467
F-statistic: 2.956e+04 on 3 and 515734 DF,  p-value: < 2.2e-16
```

Figure 26. Interaction Term Review_Total_Negative_Word_Counts:countrydummy

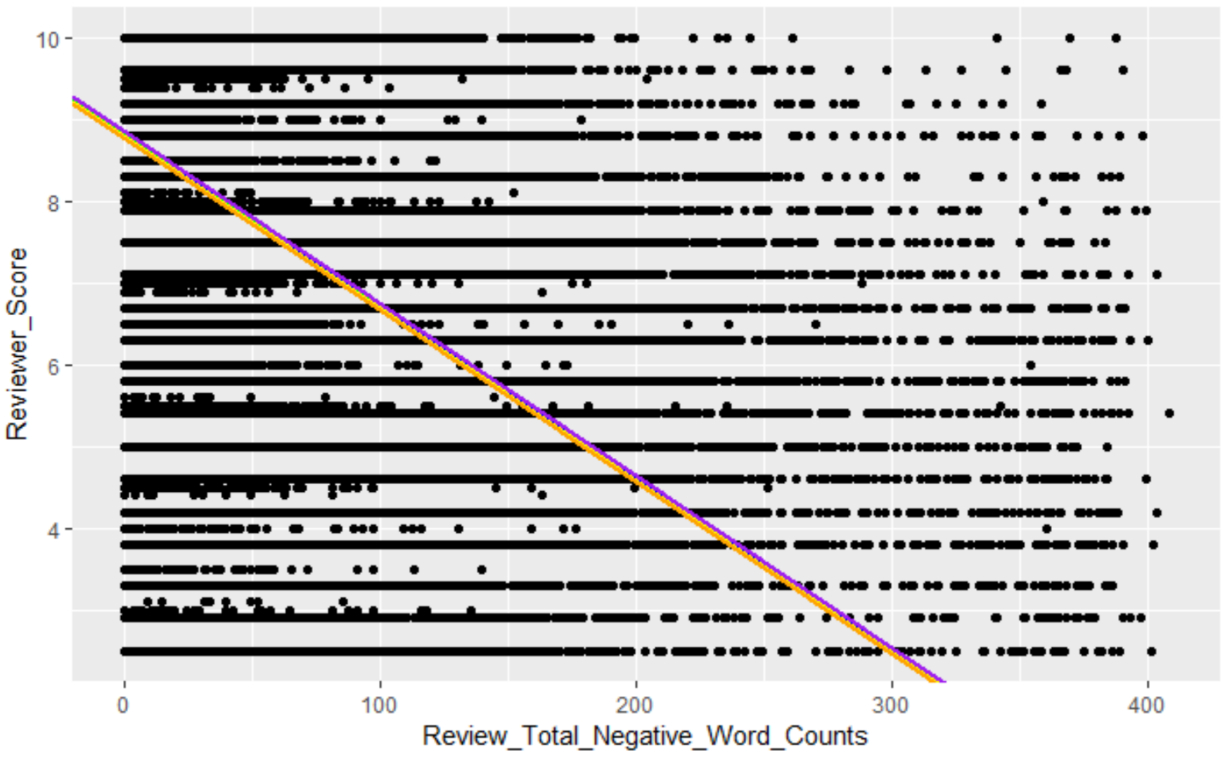


Figure 27. Plots of Regression with Interaction Term ‘Review_Total_Negative_Word_Counts:countrydummy’

Spatial Graph

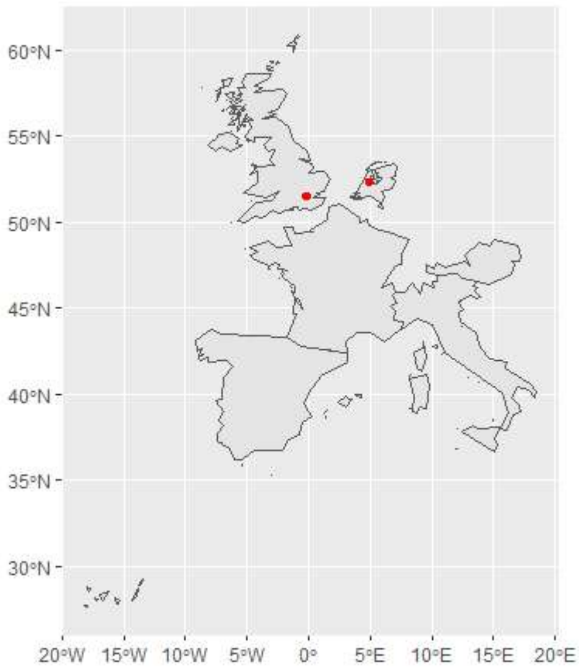


Figure 28. Spatial Graph of First 1000 Data Points

Discussion

The first multiple linear regression (Figure 7) is designed to see whether there is a general negative relationship between the reviewer's score and the following covariates: ratio of the negative reviews, time validity of the review (days_since_reivew), the discrepancy of reviews (the difference of word count between the positive or negative reviews) and the experience of the reviewers (total number of reviews given by each reviewer). Based on the result shown in Figure 7, one percent increase in the ratio of negative reviews would decrease the reviewer score by 6.538 after adjusting for other covariates. In addition, as the word count difference between positive and negative reviews increases, the reviewer score would decrease by 0.0169 after adjusting for other covariates. Even though all of the coefficients are statistically significant with infinitesimally small p-values, the magnitude of 'days_since_review' and 'Total_Number_of_Reviews_Reviewer_Has_Given' are very small and nearly negligible in the linear relationship with 'Reviewer_Score'. In this case, the time of review and experience of the reviewers have little, insignificant effect on the reviewer's score, and hence it would be plausible to conclude that the ratio of the negative reviews and the discrepancy of reviews would comparatively have more significant, negative impact on the reviewer score at almost every significance level.

In addition, the regression tree shown in Figure 11 (Regression Tree) partitions the continuous response variable, Reviewer_Score, into different pure percentages distribution of distinct values based on the two significant covariates, the ratio of the negative reviews and the discrepancy of reviews (negative_ratio and difference). Initially, the data is splitted into the subset of data with Reviewer_Score=8.2, further splitting into the subset of Reviewer_Score=8.5 and Reviewer_Score=7.8 based on negative_ratio >= 0.77. Different from binary response, the minimized impurity is computed based on the minimized sample variance, suggesting the nodes are the most likely homogenized subsets or groups of the data. Agreed with the findings from the previous multiple linear regression model, the regression tree model generally shows that as negative_ratio becomes larger such as negative_ratio>=0.77, the subsets of Reviewer_Score are comparatively lower (such as 7.8, 8, 6.9, 7.4, 8.1). Furthermore, if any of those subsets have word count difference >=34, the corresponding subsets would attribute to the lowest Reviewer_Score of 6.9. On the other hand, if negative_ratio is initially less than(<) 0.77, the subset, without further division, becomes rather stabilized attributing to 52% of homogenized groups with a higher Reviewer_Score of 8.5.

The robustness of the first linear regression could be shown by Figure 8 and Figure 9, in both cases the linear modeling assumptions are not violated. Figure 8 shows an approximate normal, bell-shaped distribution of the residuals. Figure 9 shows a randomly scattered yet a parallel linear pattern for the relationship between fitted and residual values, suggesting that fitted value of the response variable, Reviewer_Score, has many similar values and hence the residual appear parallel. More importantly, the QQ plot (Figure 10) suggests that slope of the residuals linear model is almost aligned with the slope of normal distribution in terms of theoretical quantiles.

For the multiple linear regression with respect to the keywords extracted from “Negative_Review” and “Positive_Review”, we also find some insights. First, in Figure 16, we find all coefficients (corresponding to different regressors) are statistically significant. Furthermore, though words pertinent to “room small” rank highest frequency among one word or two words (bigram), its coefficient is -0.5345, smaller in magnitude (absolute value) than the term “bathroom,” -0.5994, and “bed,” -0.8722. In our opinion, these results are reasonable: though tourists tend to complain most about the area of room, when they rate the hotel they lived in, they usually take into account factors related to the experience of living in the hotel. That is to say, even if the room is small, if the tourists had positive experience on other features of the hotel, the final score may not be that low. However, if travelers do not get to sleep well (corresponds to the term “bed” extracted from negative review) or have an awful experience taking a shower (corresponds to the term “bathroom”), the latter two are most important factors lowering the score given by the customer.

We can also find similar insights from the keywords drawn from positive review and the relationship between them and the score given by a customer (Table 2). We find that “great location” and “friendly staff” are two highest merits the customer would accredit to the hotel. In terms of how this good impression will influence the score they give (positive effect), we find that "friendly staff" outweigh that of "great location," in magnitude. This is also very intuitive. For instance, having a friendly staff may leave deeper impression of travelers than simply having a great location. The same information is also reflected by Figure 14. In addition, Figure 14 shows that tourists from Australia, US, and Canada prefer to use the word "great" when leaving good review, while tourists from UK, Ireland, Germany, and other countries prefer to use the word "good."

In terms of multicollinearity, the regression results from Figure 7 and Figure 16 show that the first two multiple regressions have all statistically significant coefficients aligned with significant F-statistics (p-value<2.2 e^-16), suggesting that none of the coefficients need to be removed and all coefficients are significantly contributed to the response variable. However, Figure 21 shows that the third multiple linear regression has one of the covariate, days_since_reviews, with comparative large P-value(0.46182), suggesting that days_since_reviews may not significantly contribute to the response variable in the linear model. As a result, a stepwise AIC procedure, as shown in Figure 25, is carried out to confirm the previous assumption and hence days_since_review is removed from the finalized, modified model.

We then add the interaction term Review_Total_Negative_Word_Counts:countrydummy in order to test whether the effect of Review_Total_Negative_Word on Reviewer_Score depends on the state of countrydummy, because we believe a reviewer's perspective and opinion on a country could affect the length of negative comments that reviewer gives. From Figure 26, the interaction term is insignificant since the p-value is 0.785, which is pretty large. Therefore, we remove the interaction term.

For the spatial graph, we try to plot the hotel location points of six European countries provided by the dataset. In the Figure 28, it shows the geographical map of the six countries in Europe. The two red points represent the hotel locations in UK and Netherlands. However, one limitation is that the whole dataset includes over 500000 data. It takes us over two hours to run the model. We finally decide to use first 1000 data as a try-out example to run our spatial model, which is more doable. It gives out the result as Figure 28 which only includes hotel locations in UK and Netherlands.

Reference

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word of mouth via consumer opinion platforms: what motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, 18(1), 38-52.

Hofstede, G. (1991). *Organisation and cultures: software of the mind*. NewYork: McGraw-Hill.

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business Research*, 58, 500-507.

Sichtmann, C. (2007). An analysis of antecedents and consequences of trust in a corporate brand. *European Journal of Marketing*, 41, 999-1015.

Seo, Yeji. (2012). Cultural Impact on Customer Satisfaction and Service Quality Evaluation in Hotels. *UNLV Theses, Dissertations, Professional Papers, and Capstones*. 1370. https://digitalscholarship.unlv.edu/thesesdissertations/1370

Sparks, B.A. & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, 32, 1310–1323.

Ueltschy, L. C., La roche, M., Eggert, A., & Bindl, U. (2007). Service quality and satisfaction: An international comparison of professional services perceptions. *The Journal of Services Marketing*, 21(6), 410-423.

Yuksel, A., Kilinc, U. K., Yuksel, F. (2006). Cross-national analysis of hotel customers' attitudes toward complaining and their complaining behaviours. *Tourism Management,* 27 (2006) 11–24.