Hotel Review in Europe

TEAM: MLB

Xiao Lu, Yange Bian, Luwen Mai, Yuyang Li

Agenda

Introduction

Research Question

Data and Method

EDA visualization

Modeling result

Conclusion

Introduction

- Digital “word of mouth”
- Online reviews have significant impacts on consumers’ booking intents and perceptions of trust
- Cultural differences will affect travelers’ opinions on what the important features of hotels are
- Cultural differences also influence consumer complaint behavior (CCB)

Research Question

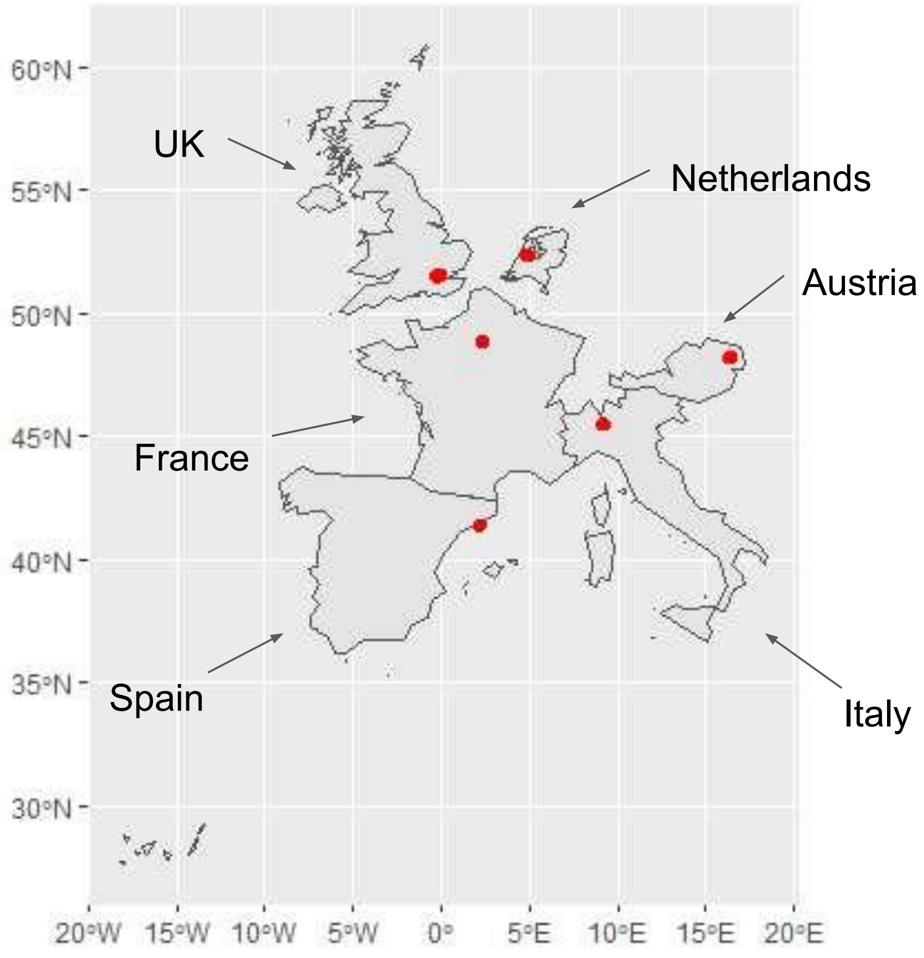
We would like to see if there is a correlation between keywords in online reviews and scores that hotels receive. The predictive power implied in this correlation could be a useful tool for hotels to cater to different needs of their customers, especially when cultural differences are at play.

Data and Method

- The dataset has 17 variables, with 515,000 reviews and scores of 1,493 luxury hotels located within 6 countries in Europe
- Choose the top 10 countries of travelers' origin
- Visualization using scatterplots, boxplot, bar charts and line graphs
- Multiple linear regressions for data analysis:
    1. review scores and influencing factors
    2. influence of keywords on positive/negative reviews
    3. regression with interaction term
- Spatial graph

Data Visualization-Spatial Analysis

1. Filter the original dataset by deleting all the duplicates from hotel_address and their associated longitudes and latitudes, and store the new data including three variable columns into a new dataset called unique.csv
2. each of the six red points on the right represents the geographical position of hotels in each of the six countries in Europe.



Limitation and Further Research of Spatial Analysis

Since the hotel address from the original dataset contains mostly only one city in each country. For example, in France, almost all hotel locations are in Paris. There is only one red point shown on the graph of each country.

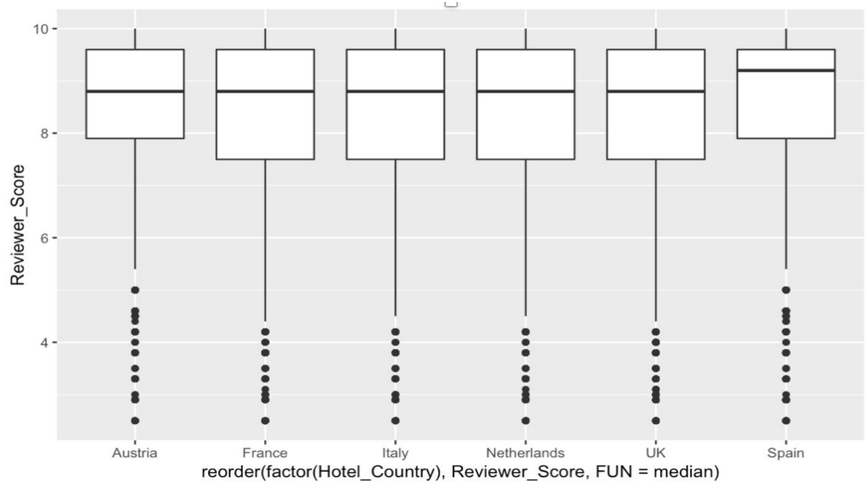
If we can find hotel addresses from more diverse cities in each country, for example, London, Manchester, Liverpool in UK, we may later implement clustering algorithm (where k=3 chosen) to do clustering on the points.

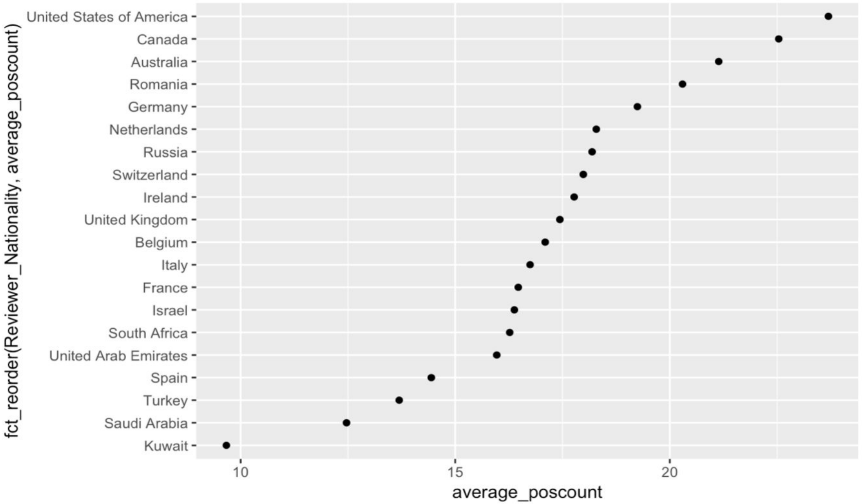
Example of pseudocode on the right.

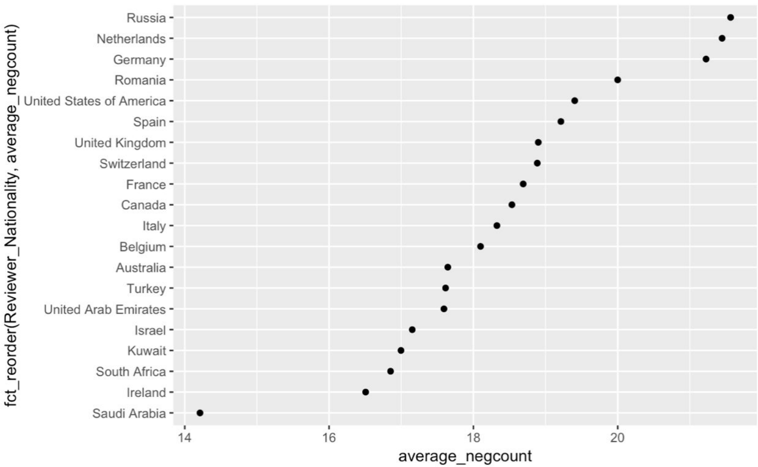
```
# assign three clusters
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3)
kmeans.fit(x)
y_kmeans = kmeans.predict(X)

# import library
from sklearn.metrics import pariwise_distances_argmin
def find_clusters(X, n_clusters, rseed=3):

# 1. randomly choose clusters
    rng = np.random.RandomState(rseed)
    i = rng.permutation(X.shape[0])[:n_clusters]
    centers = X[i]

    while True:

# 2. assign labels based on closest center
    labels = pairwise_distances_argmin(X, centers)

# 3. find new centers from means of points
    new_centers = np.array([X[labels == i].mean(0) for i in range(n_clusters)])

centers, labels = find_clusters(X, 3)
plt.scatter(X[:, 0],X[:, 1], c=labels, s=50, cmap='viridis')

    if np.all(centers == new_centers):
        break
    centers = new_centers
return centers, labels
```

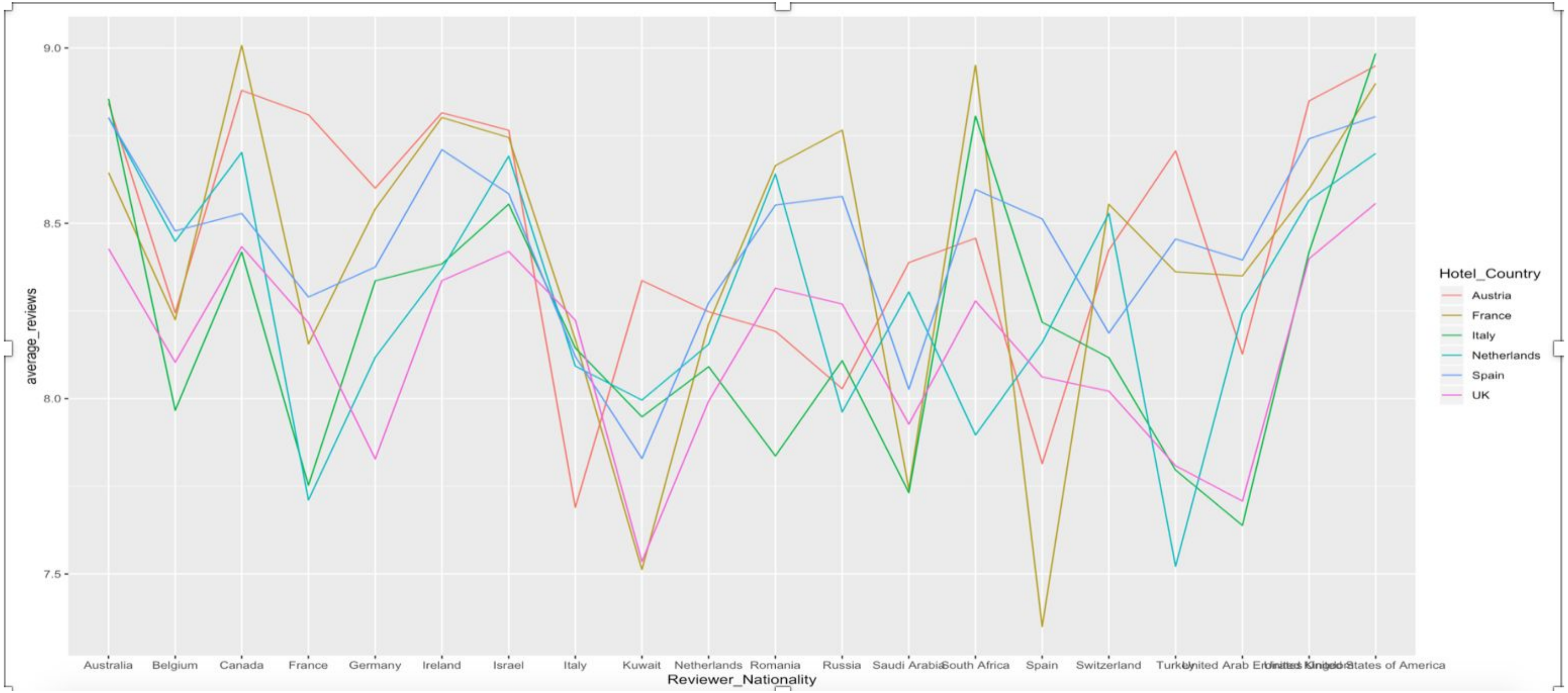
EDA Results/Data Visualization

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



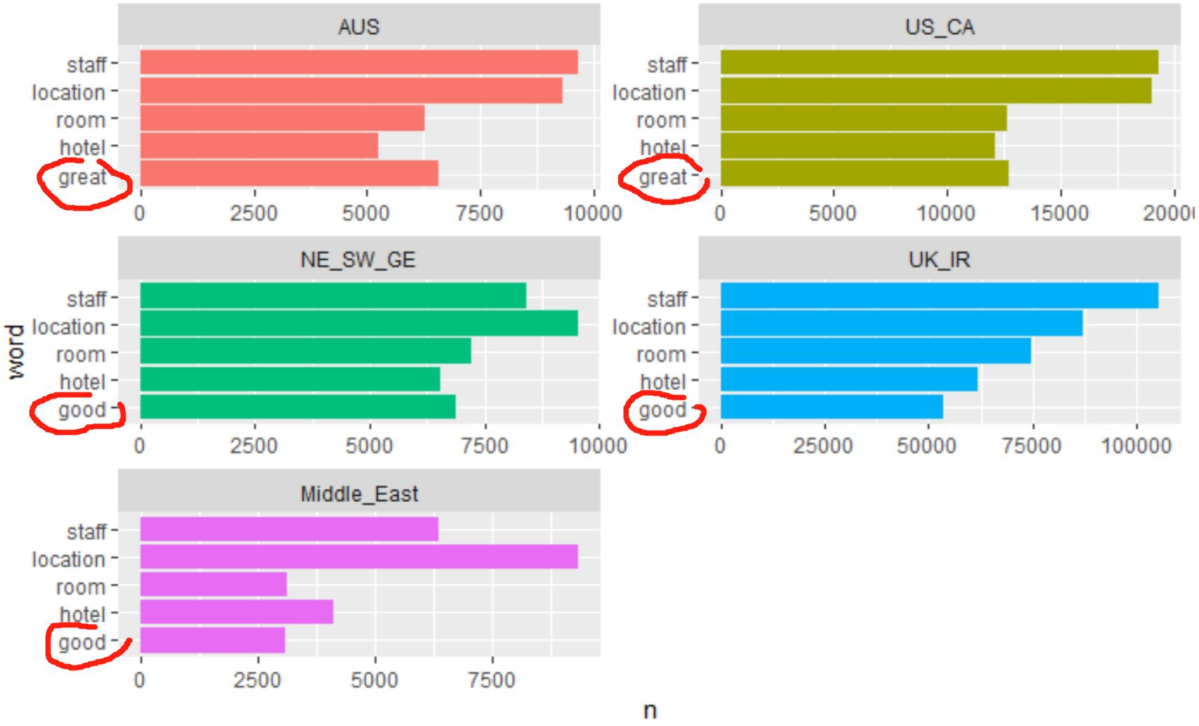






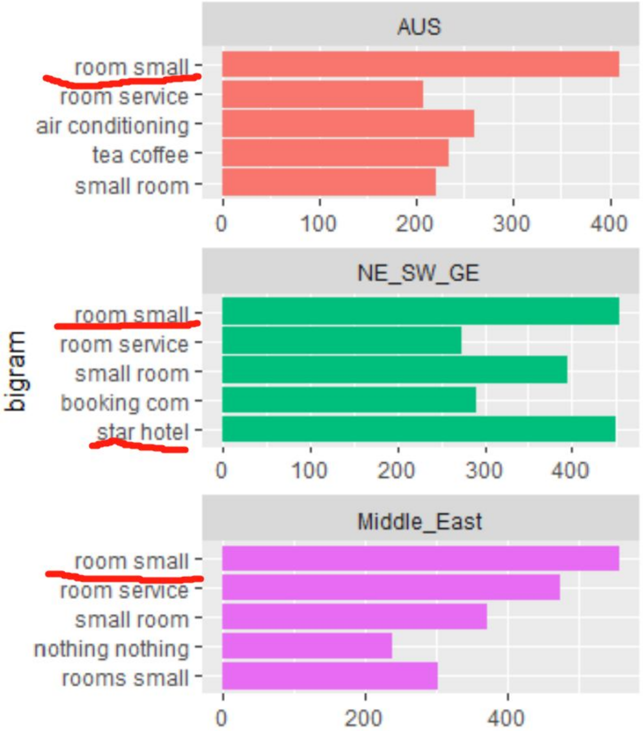
9.0
8.5
8.0
7.5
average_reviews
Australia
Belgium
Canada
France
Germany
Ireland
Israel
Italy
Kuwait
Netherlands
Romania
Russia
Saudi Arabia
South Africa
Spain
Switzerland
Turkey
United Arab Emirates
United Kingdom
United States of America
Reviewer_Nationality
Hotel_Country
Austria
France
Italy
Netherlands
Spain
UK

EDA Visualization: Positive Review



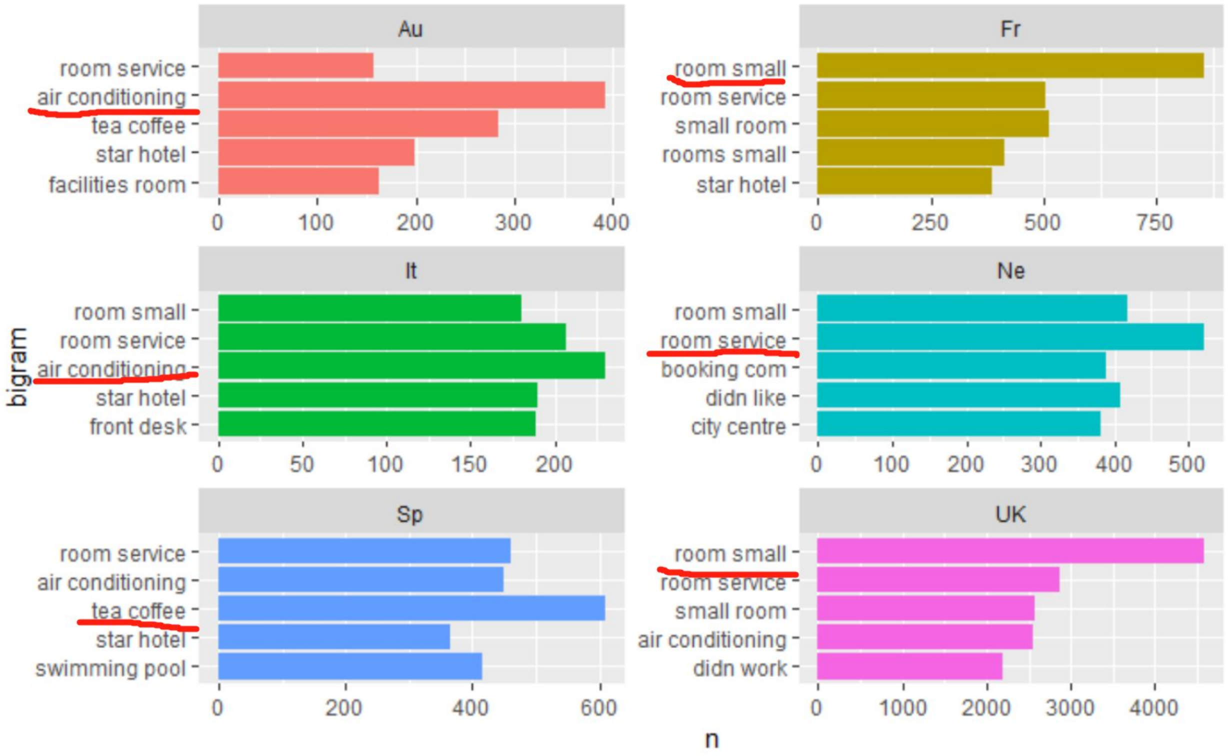
- Cultural Similarity
- Great VS Good
- Friendly Staff
- Good Location

EDA Visualization: Negative Review (1)



- Group by Tourists' Country
- Room Small
- Peculiarities

EDA Visualization: Negative Review (2)



- Group by Hotel Country
- Room Small
- Air Conditioner
- Peculiarity

EDA Visualization: Average Score

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

Is the above result consistent with people’s intuition?

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 1: Summary of Parameter Estimates

Comparatively Significant Covariates in terms of estimates and p-value:

- negative_ratio（ratio of negative reviews)
- difference（word count differences between positive/negative reviews)

Multiple Linear Regression #1: Modeling Assumptions

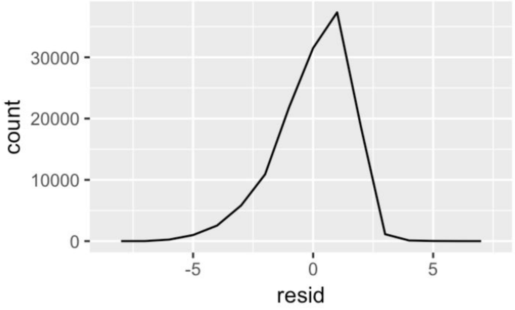


Figure 2: The Frequency Distribution of Residuals

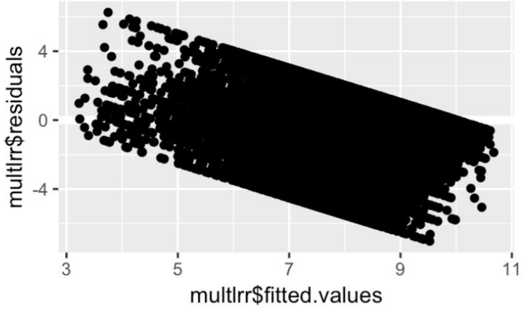


Figure 3: Residuals vs. Fitted Values

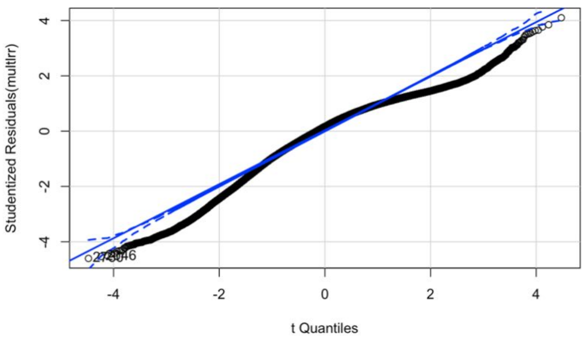


Figure 4: QQ Plot of Multiple Linear Regression #1

Regression Tree Modeling

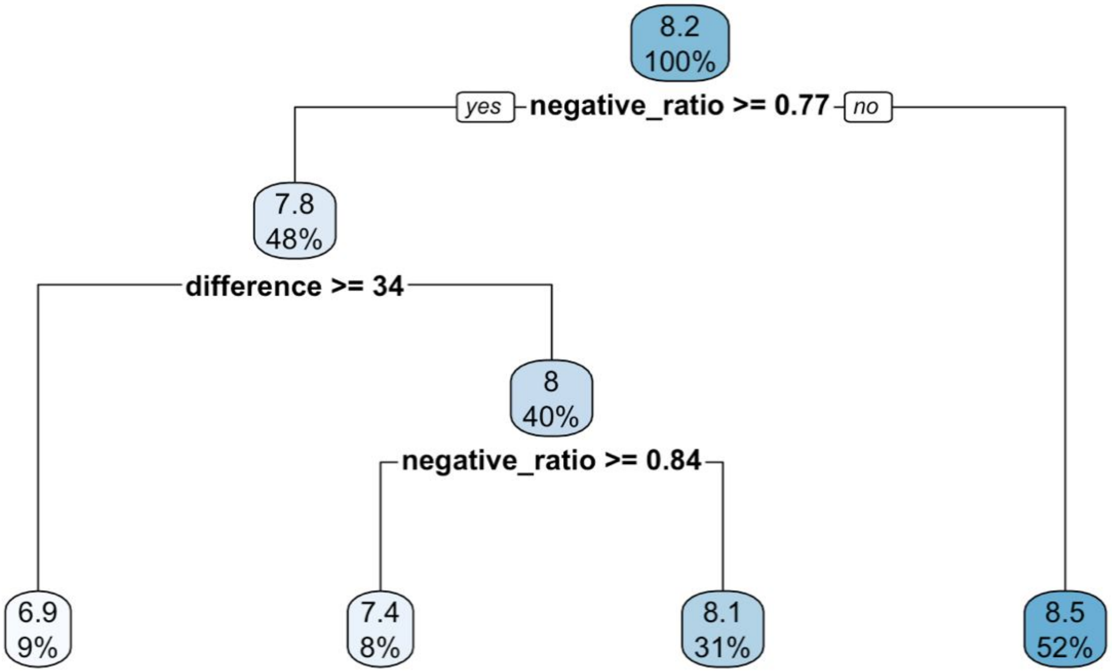
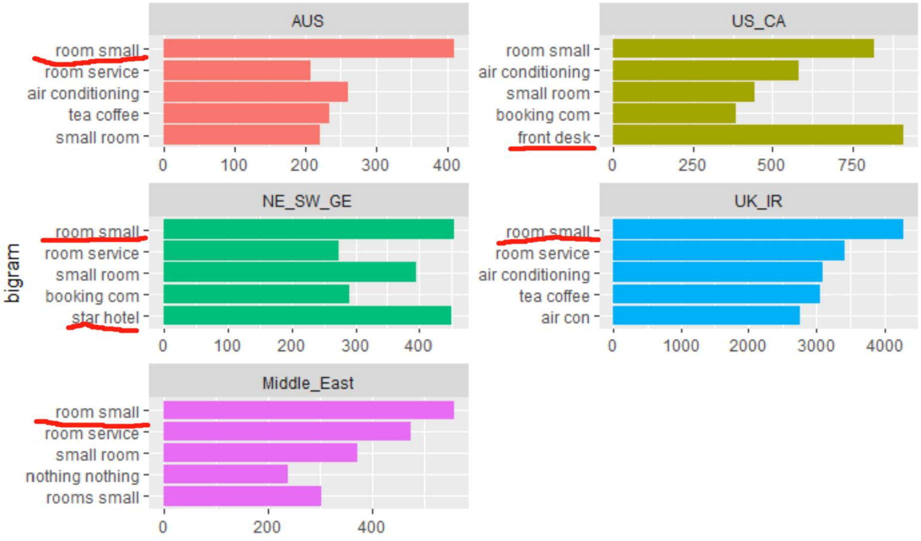


Figure 5: Regression Tree of Multiple Linear Regression #1

Modelling Results: Multiple Linear Regression for Negative Review

```
                              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 ***
```

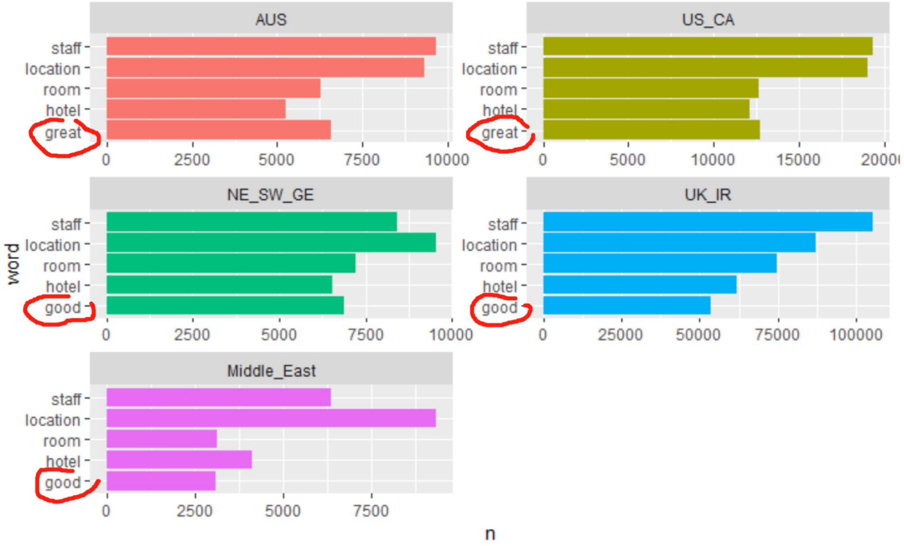
- Those closely related to living experience are ‘powerful’



Modelling Results: Multiple Linear Regression for Postive Review

	Estimate	Std. Error	t value	Pr(>|t|)	
location	5.262e-02	5.716e-03	9.206	< 2e-16	***
staff	6.191e-01	5.470e-03	113.181	< 2e-16	***
bed	2.274e-01	7.680e-03	29.609	< 2e-16	***
clean	1.819e-01	7.619e-03	23.867	< 2e-16	***
breakfast	1.750e-01	7.855e-03	22.279	< 2e-16	***
Reviewer_NationalityAUS	8.519e+00	1.598e-02	533.203	< 2e-16	***
Reviewer_NationalityUS_CA	8.662e+00	1.411e-02	613.878	< 2e-16	***
Reviewer_NationalityNE_SW_GE	8.138e+00	1.538e-02	529.292	< 2e-16	***
Reviewer_NationalityUK_IR	8.500e+00	1.302e-02	653.060	< 2e-16	***
Reviewer_NationalityMiddle_East	7.921e+00	1.643e-02	482.152	< 2e-16	***
Hotel_CountryFr	-1.420e-01	1.384e-02	-10.262	< 2e-16	***
Hotel_CountryIt	-1.604e-01	1.594e-02	-10.060	< 2e-16	***
Hotel_CountryNe	-1.227e-01	1.383e-02	-8.873	< 2e-16	***
Hotel_CountrySp	-3.886e-02	1.376e-02	-2.824	0.00474	**
Hotel_CountryUK	-2.564e-01	1.207e-02	-21.236	< 2e-16	***
days_since_review	9.000e-06	1.223e-05	0.736	0.46182	

- Friendly staffs do matters!



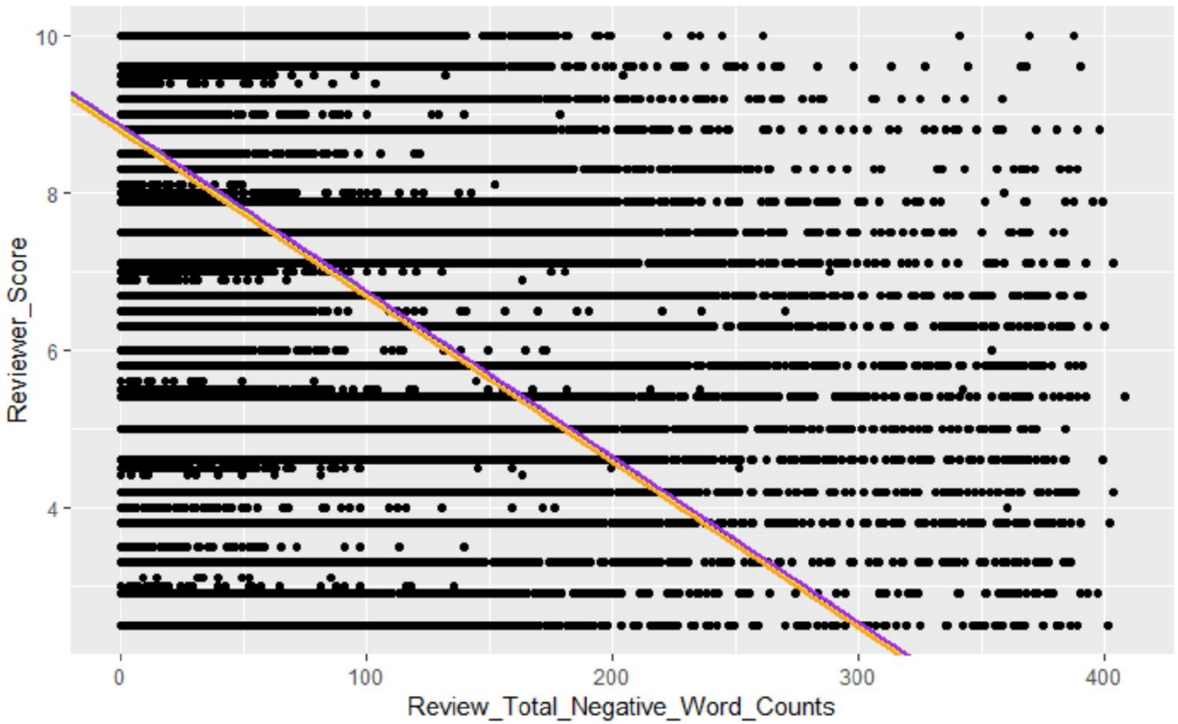
Interaction Term

Later, we add an interaction term Review_Total_Negative_Word_Counts:countrydummy to test whether the effect of Review_Total_Negative_Word_ Counts on Reviewer_Score depends on the state of countrydummy, where countrydummy is denoted shown at the bottom.

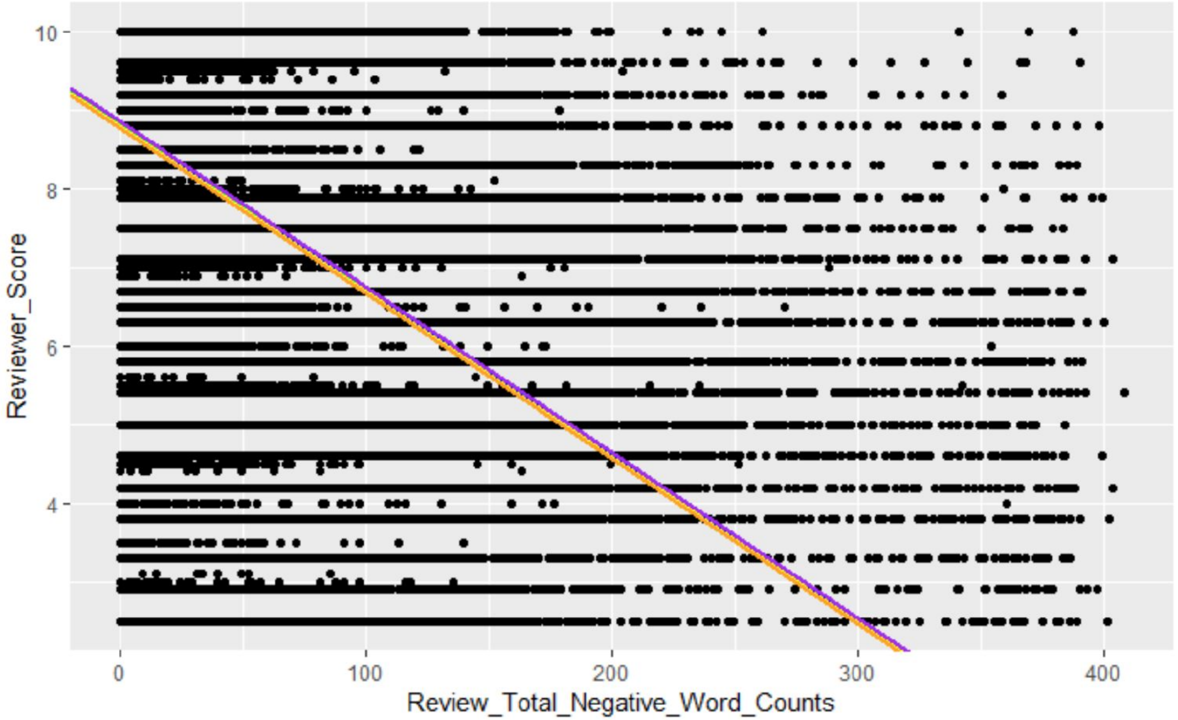
```
mutate(countrydummy = ifelse(Hotel_Country == 'France', 1,
                             ifelse(Hotel_Country == 'Italy', 2,
                                    ifelse(Hotel_Country == 'Netherlands', 3,
                                           ifelse(Hotel_Country == 'Spain', 4,
                                                  ifelse(Hotel_Country == 'UK', 5, 0))))))
```

From the table, the p-value is 0.785, which is insignificant. Therefore, we decide to remove the 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
```



We use six colors to represent six countries. From the graph on the left, we find that six lines are in high superposition.



Additionally, there is an interesting pattern that at lower reviewer scores, the negative word counts are more widely distributed whereas they are highly concentrated at lower values when the review scores are becoming larger.
We think the reasons could be travelers who are more dissatisfied with the hotel experience might write more negative comments to abreact their negative emotions.

Conclusion

```
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
```

- Joint Significance of Coefficients, and the Overall Goodness of Fit for Negative Words

```
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
```

- Joint Significance of Coefficients, and the Overall Goodness of Fit for Positive Words

- Finally, hotels/restaurants should give tourists a hospitable ambience, regardless of where the tourists come from!

Thank You！

- Q & A