## Mini-Abstract:

Online reviews are becoming increasingly important, especially for hotels. When booking accomodations, travelers are paying more attention on online reviews about the hotels. In particular, the quantity, length, time and attitude of the reviews are the main determinants of the overall review score of each hotel, thereby motivating our study.

This project deliverable performs a preliminary analysis of online reviews for hotels in Europe. Multiple relationships between the reviewer score and the key aspects of reviews, which include the length of reviews, time of reviews, the ratio of negative reviews and the number of reviews given by each reviewer, are analyzed separately using simple linear and logistic modeling methods.

Based on the data analysis results using dataset from booking.com, we find that review scores are negatively correlated with the length of reviews and the ratio of negative reviews, and positively associated with the number of reviews given by each reviewer. We find no significant relationship between review scores and how long it takes travelers to write reviews.

## Research Questions and Modeling Methods:

In this project, we will use simple linear regression and logistic regression to analyze how each key aspect of reviews, as defined earlier, quantitatively affect the overall review's score.

For the linear models, we are analyzing the question from different perspectives. In the first regression, the dependent variable is the review scores of each hotel and the main regressor is the ratio of negative reviews to total reviews. We will be able to see how one percent increase in the negative_ratio affects the average review score of a hotel. In the second regression, we are trying to find out whether how long it takes the travelers to write reviews has a negative impact on the scores they give. The third linear regression is based on the assumption that frequent travelers are less demanding in terms of hotel conditions and hence more likely to give higher scores. We use the number of comments that each person gives to represent travel frequencies. At this stage, we hope that these results will pave the way for our analysis in the final report: How can hotels use the online review to improve their future revenue. For instance, in the final report we would like to investigate whether travellers (booking the hotel via www.booking.com online) from certain region have certain preference? Since the hotels receive the online booking in advance, it can use the information to make some preparation and then better accommodate the travellers from certain region. Moreover, we intend to investigate

whether there is any pattern between certain word in the comment with online-booking traveller from certain region.

For the logistic models, the binary response, denoted by “yi” or “negative_reviews”, is a dummy variable that equals to 1 if the review score is less than 7, and 0 if greater or equal to 7. The mean response, $\pi i$=P(yi=1), is the probability of getting a review score less than 7, which is a boundary score separating positive and negative reviews. The predictor of this model, denoted by “x” or “pos_neg_diff”, is the absolute value of the difference between total word counts for positive and negative reviews. Hence, the dependent variable of this model is the natural logarithm of the ratio of probability of getting negative response to positive response, regressing on the predictor.

## Modeling Results:

### Part A: Linear Regression Result & Plot

```
lm(formula = score ~ negative_ratio, data = data_for_lm1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.01524 -0.21123  0.00706  0.25907  1.65241

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     11.9970     0.1126  106.53   <2e-16 ***
negative_ratio  -4.8565     0.1529  -31.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3955 on 909 degrees of freedom
Multiple R-squared:  0.5261,    Adjusted R-squared:  0.5256
F-statistic:  1009 on 1 and 909 DF,  p-value: < 2.2e-16
```

Table1: The Regression Result.

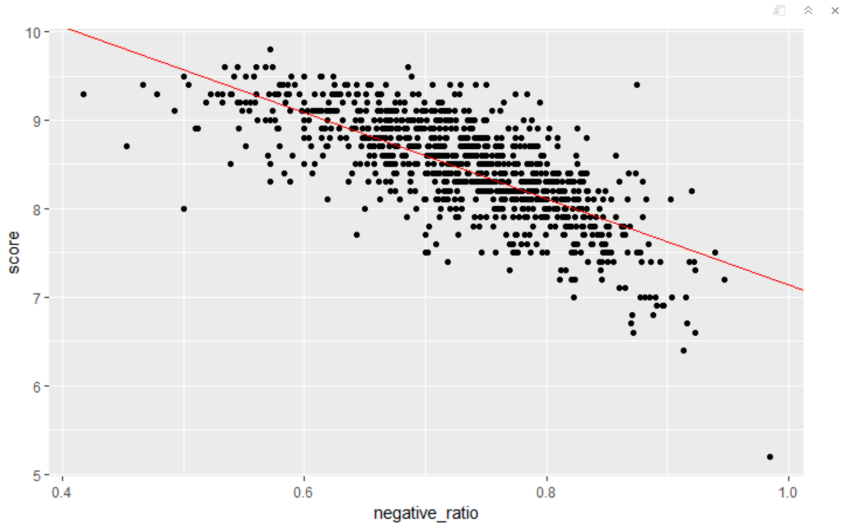


Figure1: hotels' score on y-axis against negative_ratio on x axis

```
lm(formula = Reviewer_Score ~ days_since_review, data =
reviews_ed2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9028 -0.8952  0.4087  1.2096  1.6122

Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)       8.388e+00  4.491e-03 1867.596   <2e-16 ***
days_since_review 2.046e-05  1.092e-05    1.874   0.0609 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.638 on 515736 degrees of freedom
Multiple R-squared:  6.81e-06,  Adjusted R-squared:  4.871e-06
F-statistic: 3.512 on 1 and 515736 DF,  p-value: 0.06092
```

Table 2: Regression Result of score and days_since_review

```
lm(formula = Reviewer_Score ~ num_of_reviews - 1, data =
reviews_ed6)

Residuals:
     Min       1Q   Median       3Q      Max
-110.175    4.429    6.764    8.255    9.664

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
num_of_reviews 0.3362663  0.0008797   382.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.027 on 350949 degrees of freedom
Multiple R-squared:  0.2939,    Adjusted R-squared:  0.2939
F-statistic: 1.461e+05 on 1 and 350949 DF,  p-value: < 2.2e-16
```

Table 3: Regression Results of Correlation Between Review Score and the total number of reviews the reviewer has left.

## Part B: Basic Checking for Linear Assumption

Whether the model is classical (small sample) OLS or large sample (asymptotic theory) OLS, the most important assumption is the normality of error term, ε. If this assumption is violated, the constructed t-statistics $\frac{b_j - c}{\sqrt{(X'X)_{jj}^{-1}\frac{e'e}{n-k}}}$ would not be t-distributed since the denominator, with our estimated standard error, is not chi-square distributed and the numerator is not standard normal distributed as well. Therefore, we conduct the following two basic plot checking for the error term.

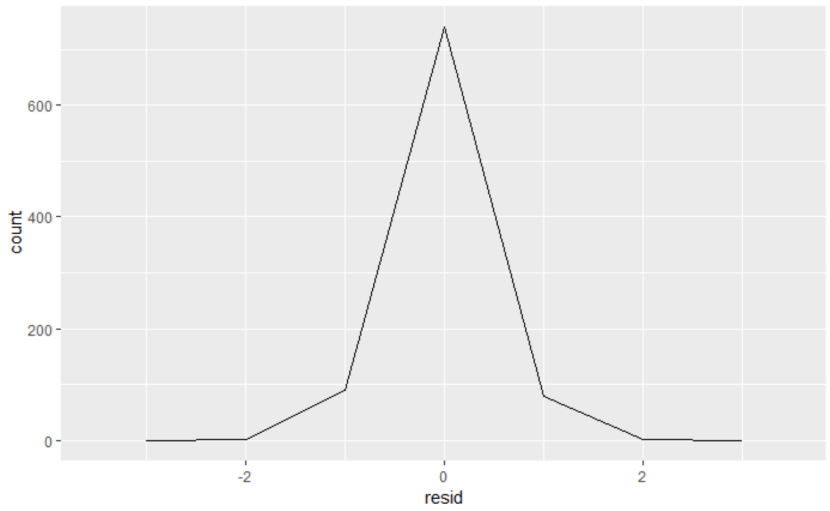

Table 4: The Probability Density Function of the Residual Term (as an unbiased estimation and ideal proxy for the unobserved error term)

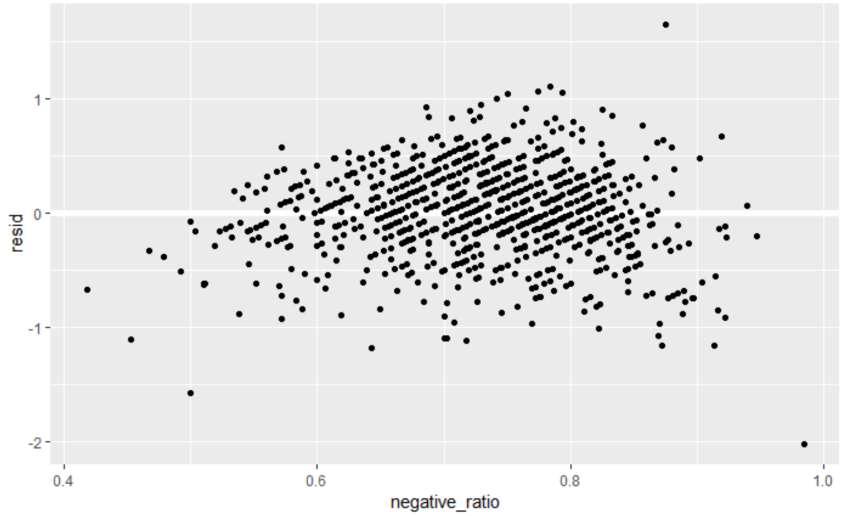


Table 5: The Conditional Distribution of Residual Term on negative ratio

## Part C: Logistic Regression Result & Plot

$$\log\left(\frac{\pi i}{1-\pi i}\right) = -1.974 + 0.016167x$$

```
 (Intercept) pos_neg_diff
 -1.97406883   0.01616707
Analysis of Deviance Table

Model: binomial, link: logit

Response: negative_reviews

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         515737     467612
pos_neg_diff  1    19072    515736     448539 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Chi-squared P-value Test

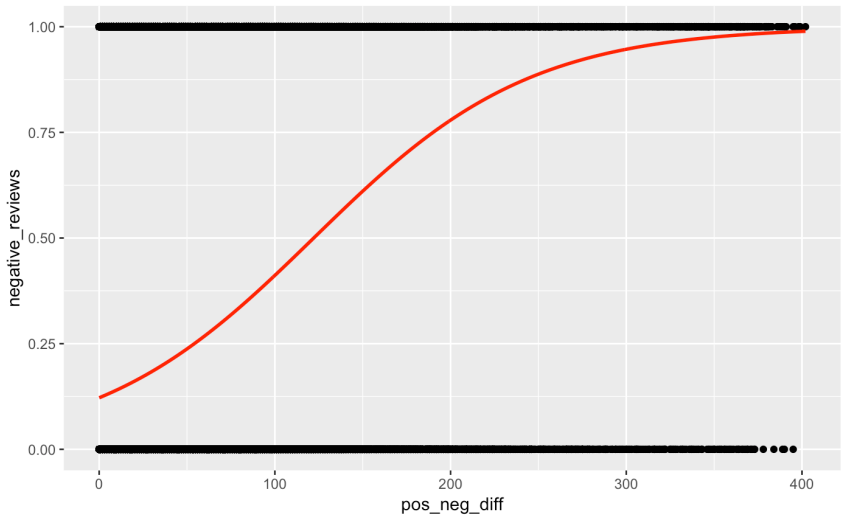


Figure 3: Logistic Curve

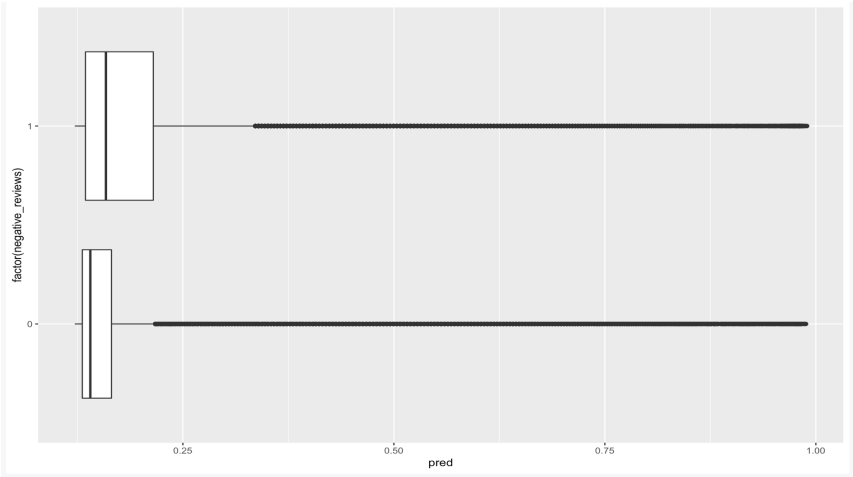


Figure 4: Logistic Boxplot

# EDA/Discussion:

## Linear Regression

Result of the linear model (Table 1) shows that one percent increase in the ratio of negative reviews among total reviews would decrease the average score by 0.05 points. This result is statistically significant at 1% level as the p-value is much smaller than the threshold value of 0.001. This negative relationship can also be seen on Figure 1.

The second linear regression (Table 2) is designed to see whether there is a negative relationship between how long it takes the travelers to write reviews (days_since_review) and the review scores. However, based on the results shown in Table 2, we find that even if we consider the coefficient of days_since_review is significant at 10% level, the magnitude is very small. This tells us that how long it takes travelers to write reviews has little to do with the scores they give to the hotels.

In the third linear regression (Table 3), our assumption is that if a reviewer has left more comments than others, he/she would have more experience in staying in hotels, and hence would be less demanding and are more prone to give a higher score. The result confirms our speculation. It shows that if a traveler has one more review written, the average score he/she gives would increase by 0.336.

## Logistic Regression

Result of the logistic model (Figure 2) shows that as the absolute difference of word count between negative and positive review increases, the log odds ratio(logit) of the probability of getting a low review score increases by about 1.162%. Also, the coefficient of x (pos_neg_diff) has extremely small p-value (<2.2e-16). Thus, there is sufficient evidence that the logistic relationship between the mean response and predictor is statistically significant.

The logistic curve (Figure 3) shows that as the absolute difference between the total word counts increases, the probability of getting a negative review increases. We also see that as the logistic curve changes from concaving up to concaving down, the rate of increase in probability decreases and the probability of getting negative review, a reviewer score less than 7, approaches to 1.

The logistic boxplot (Figure 4) shows that based on the logistic computation of the absolute difference between the total word counts, the interquartile range of probability of getting a negative review (denoted as 1), a reviewer score less than 7, is wider than the range of probability of getting a positive review (denoted as 0). Other than the outliers and similar minimum probability, the median probability of getting a negative review is greater than that of getting a positive review, and the maximum probability, as shown from the upper whisker, of getting a negative review is around 0.125 times more than that of getting a positive review.