

# Hotel Data Analysis - Midterm Report

Zhihan Wang, Junrui Ye

## 1 Introduction

- We are using hotel-booking data from Kaggle website to analyze the impact of different factors in hotel booking. The final goal of this project is to predict whether a hotel will be booked and how much the gross booking cost in USD (GBU) will be. The midterm report focuses on analysis of which features will affect the prediction the most and fit preliminary models for predicting GBU value of a new booking which is the first milestone before reaching the final goal.

## 2 Data Cleaning

- We select factors that we believe are somehow related to our prediction. Missing values were filtered out except those in GBU column. This is done using Python since the data is too big to process with Julia(all RAM will be used and computer could not handle). So the new data frame contains no NA values in any columns other than gross bookings usd. This gives us a data frame without unnecessary NA values which is small enough for Julia to process yet large enough to build models with without under-fitting problems.
- To check data corruptions, the columns with Boolean values were checked to see whether the values are all 0s and 1s. For the columns with specific IDs or codes from the website, it is not clear whether the values are corrupted as no information of the detailed codes were given.

## 3 Data Description

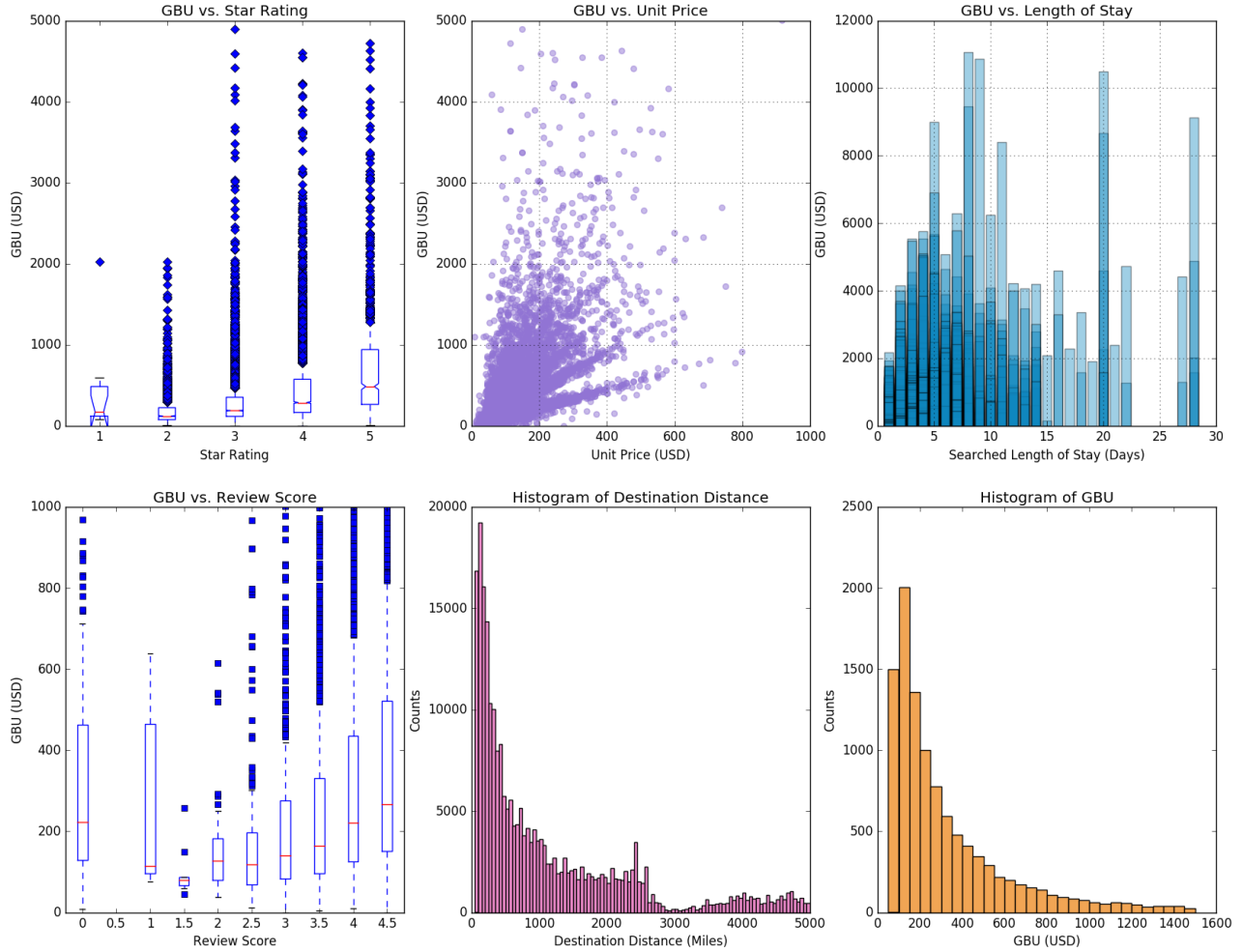
- Data Descriptions

Column name	Description	Data type
date_time	Timestamp	string
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	int
posa_continent	ID of continent associated with site_name	int
user_location_country	The ID of the country the customer is located	int
user_location_region	The ID of the region the customer is located	int
user_location_city	The ID of the city the customer is located	int
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be	double
user_id	ID of user	int
is_mobile	1 when a user connected from a mobile device, 0 otherwise	tinyint
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	int
channel	ID of a marketing channel	int
srch_ci	Checkin date	string
srch_co	Checkout date	string
srch_adults_cnt	The number of adults specified in the hotel room	int
srch_children_cnt	The number of (extra occupancy) children specified in the	int
srch_rm_cnt	The number of hotel rooms specified in the search	int
srch_destination_id	ID of the destination where the hotel search was performed	int
srch_destination_type_id	Type of destination	int
hotel_continent	Hotel continent	int
hotel_country	Hotel country	int
hotel_market	Hotel market	int
is_booking	1 if a booking, 0 if a click	tinyint
cnt	Numer of similar events in the context of the same user	bigint
hotel_cluster	ID of a hotel cluster	int
destinations.csv		
Column name	Description	Data type
srch_destination_id	ID of the destination where the hotel search was performed	int
d1-d149	latent description of search regions	double

- Main Features to Focus on
  - The main features to focus on include hotel features like star rating, hotel unit price, review score, customer features like historical booking price of the customer, and search features like length of stay, number of rooms. We choose these features because intuitively they are thought to affect the gross booking values. We perform some preliminary analysis on these data. The relationship between GBU and each of these features were plotted (shown below) and general patterns were discovered as expected.

## 4 Descriptive Statistics

- Plots and Feature Exploration

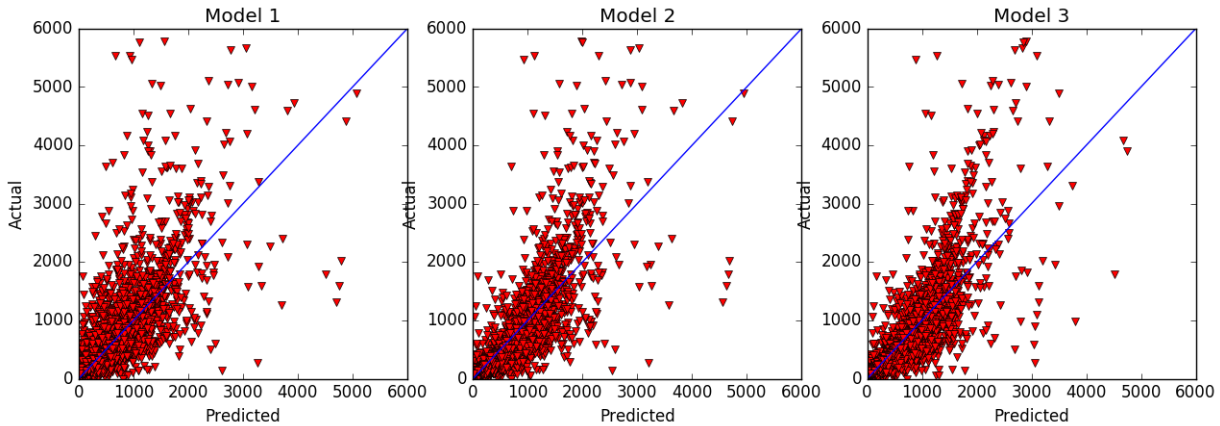


- The graphs above show the general correlations of gross booking cost in USD(GBU) with other features specified.
- We have selected star rating, hotel unit price, searched length of stay, review scores to be the main features of the model(discussed in detail in the Analysis Strategies section). This is because it is clear that the first four plots of those features show a great correlation between GBU and the feature specified on the x axis.
- More plots and analysis were also done to evaluate other features which are not shown here. Some histograms of features are presented here to show how the distribution of the features are.

## 5 Analysis Strategies

- Models and Graphs:
  - We perform regressions on selected data to predict whether the customer will book the hotel and the GBU of the bookings.

- For preliminary analysis, we fit three different linear regression models:  
 Model1: Regression on four features: price, days, review rating, star rating  
 Model2: Regression on nine features: price, days, number of rooms, star rating, promotion boolean, brand boolean, location score, booking window, visitor's average booked hotel price  
 Model3: Multiply days and number of room together. Other features remain the same as Model2.



- How to Test Effectiveness of the Models
  - Mean Square Error  
 Compute the mean square error for each model, choose the model with the least error.
  - Test Models on Test Set  
 After picking the best model, it was tested on the test set in order to check its error and effectiveness.
  - Cross-validation  
 Cross-validation was used in order to check the least squared errors of each model thus pick the best one for future analysis and prediction.
- How to Avoid Under-fitting and Over-fitting
  - To prevent under-fitting, large amount of data was used to perform the analysis on. Also, various features were used to fit the initial models and validated using Cross-validation to measure their significance on the prediction.
  - To prevent over-fitting, we used regularization to reduce the variance of the model. Lasso regression and l1 regularization were used in order to test the significance of each feature on the prediction and perform feature elimination when needed.

## 6 Future Plan

- Future Data Modeling and Prediction
  - Regularization will be used later to discover details of the features included since they would be penalized less compared to l1 regularization.
  - Linear or polynomial regression(depends on the model testing results) will be used to determine a detailed price of hotel booking. Thus when having certain information of a hotel search in the future, the potential total price of the booking can be predicted which would also be helpful for prediction of whether the search would lead to a booking.
  - Since the result is a Boolean value, logistic regression may be considered as a decent model to study whether a specific hotel search will lead to a booking.
- Data Visualization
  - As further analysing goes, more data visualization will be produced in order to give a better understanding and explanation to audience of what stories the data is conveying and what result our analysis is providing.