DSO 530 Final Project

Expedia Hotel Recommendations

Which hotel type will an Expedia customer book?

Group 6

Tong Xie
Yidan Xue
Yifeng Wang
Ningdong Wang
Weichen Zhang

USC Marshall
School of Business

Contents

USC Marshall
School of Business

PART 1

Business Goal

What big-data-driven business can we build from users' behavior?

Business Goal

- Predict the hotel type that a user would like to book
- Provide 10 personalized hotel recommendations to travelers

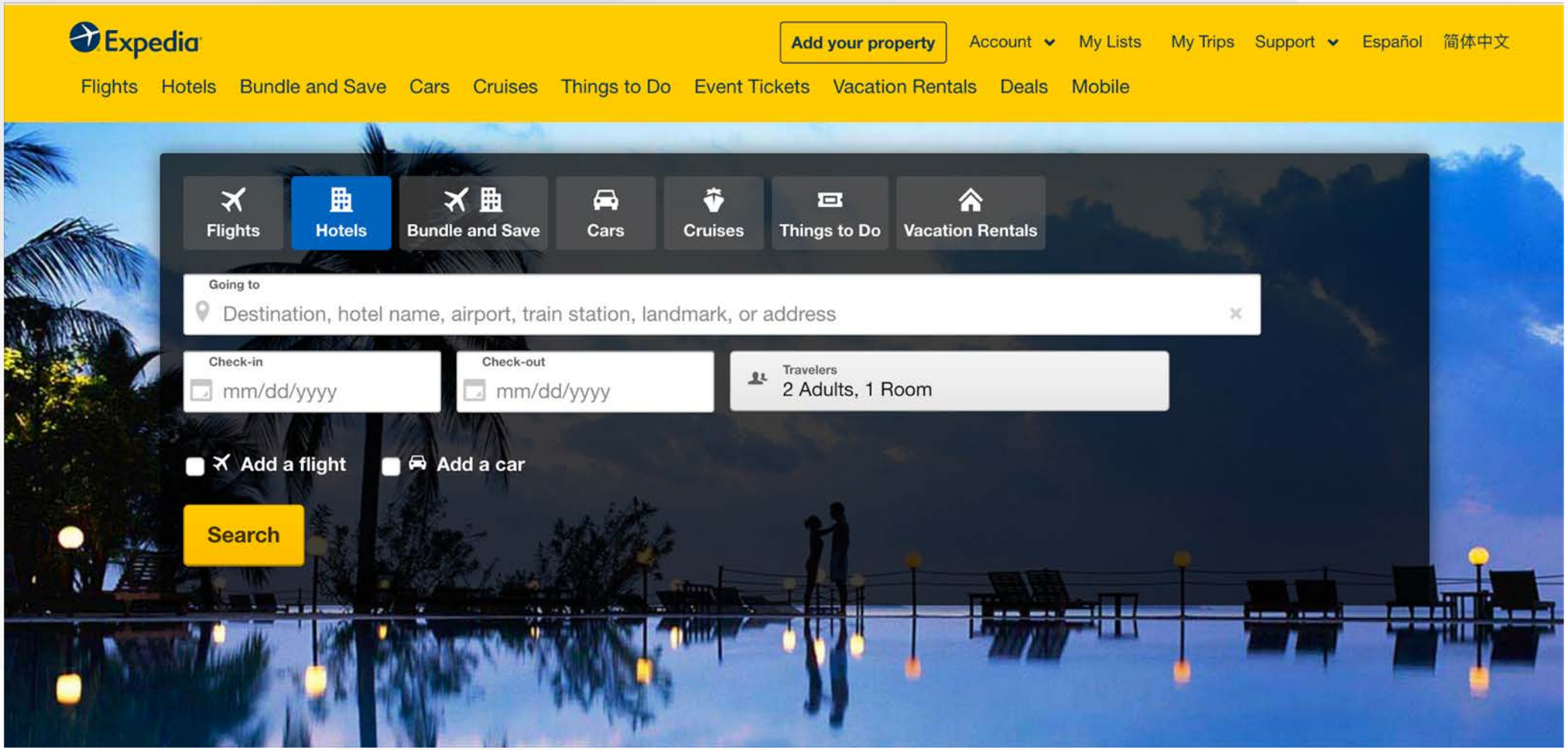
USC Marshall
School of Business

PART 2

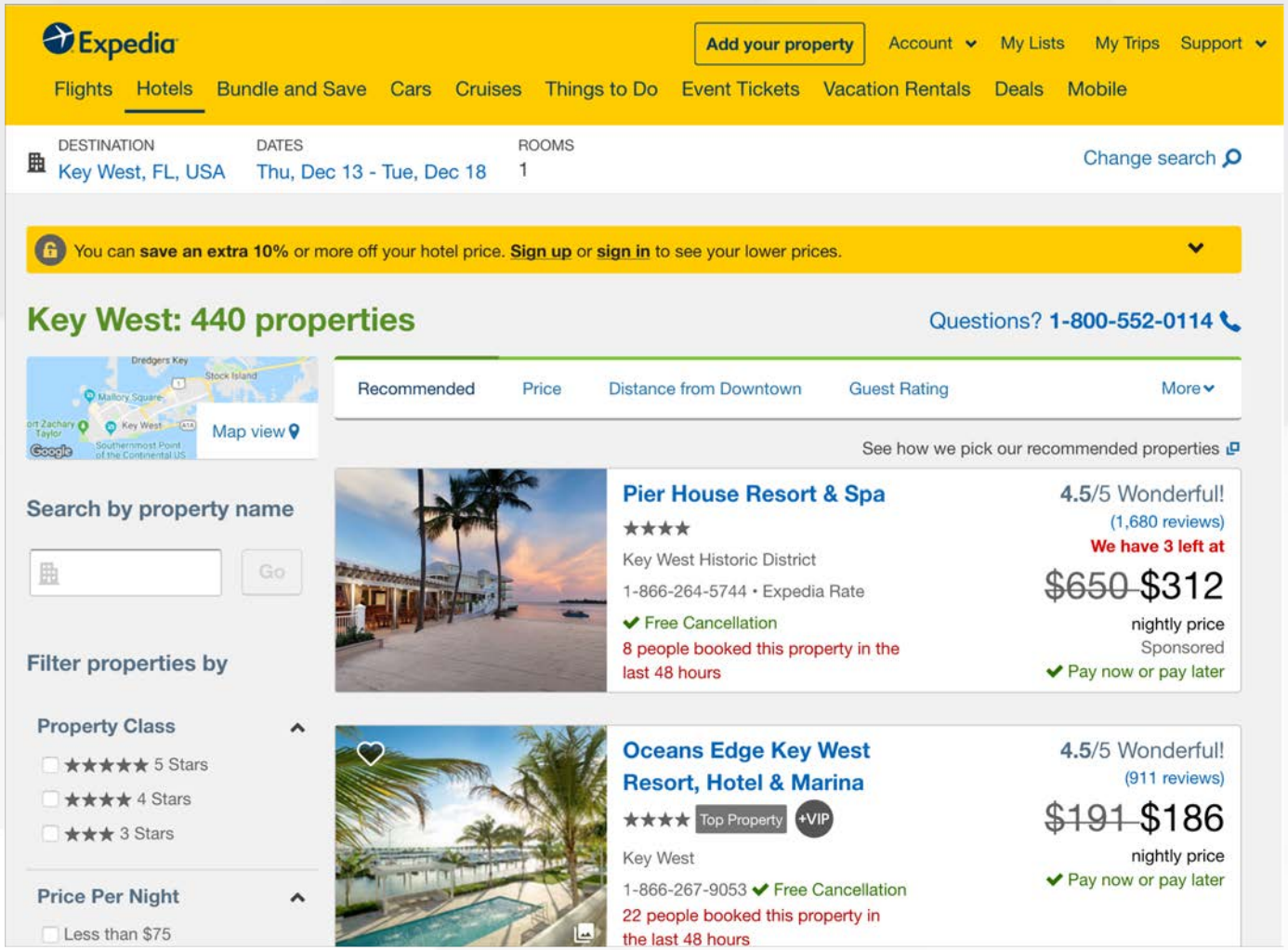
Data understanding

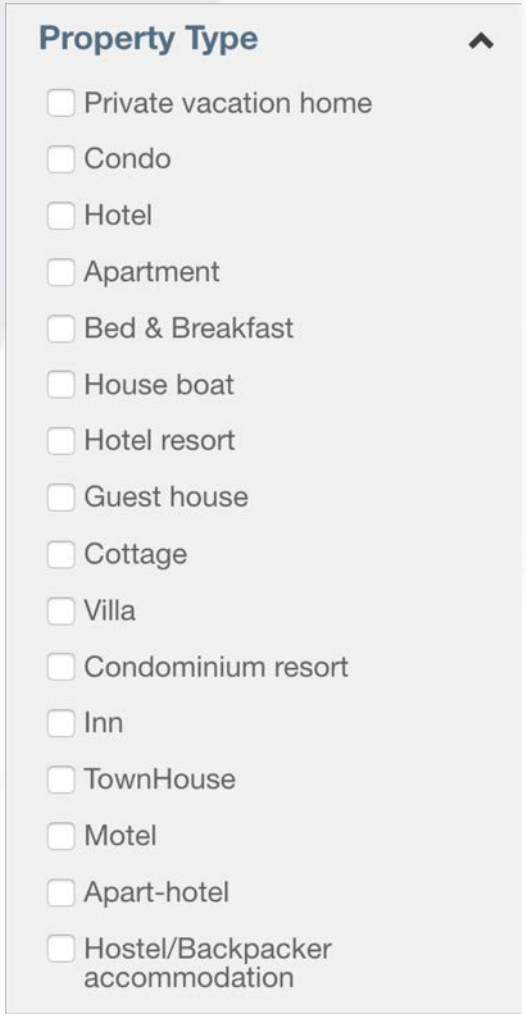
How are data generated?

Expedia Website



Search Result





USC Marshall
School of Business

PART 3

Exploratory Data Analysis

How does data look like? What expect variables can be created?

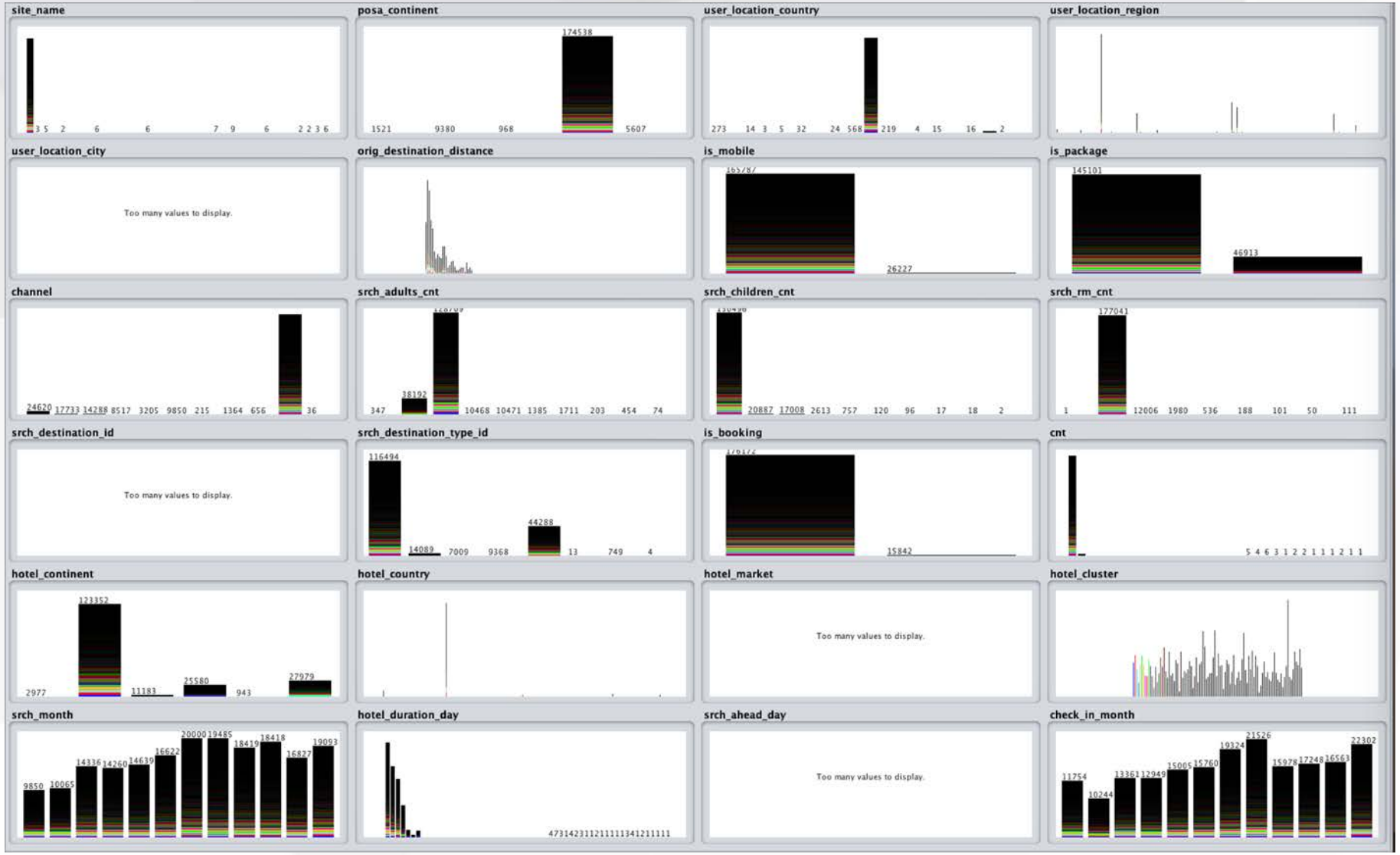
Challenges in Handling the Dataset

USC Marshall School of Business

	5 numerical variables	15 categorical variables	3 date/time variables	Target variable
192,014 rows	very different scales	at most over 10,000 unique values	noise, outliers	100 levels

Masked data — difficult to group

Visualization



Variable Creation

- **date_time** Timestamp
- **srch_ci** Check in date
- **srch_co** Check out date

Expert Variable	Definition
hotel_duration_day	Days that users plan to stay in the hotel
srch_ahead_day	Days in advance that users plan a trip
srch_month	The month that users perform the search
check_in_month	The month that users plan to check in

USC Marshall
School of Business

PART 4

Model Selection and Evaluation

Challenges and progress in the modeling process

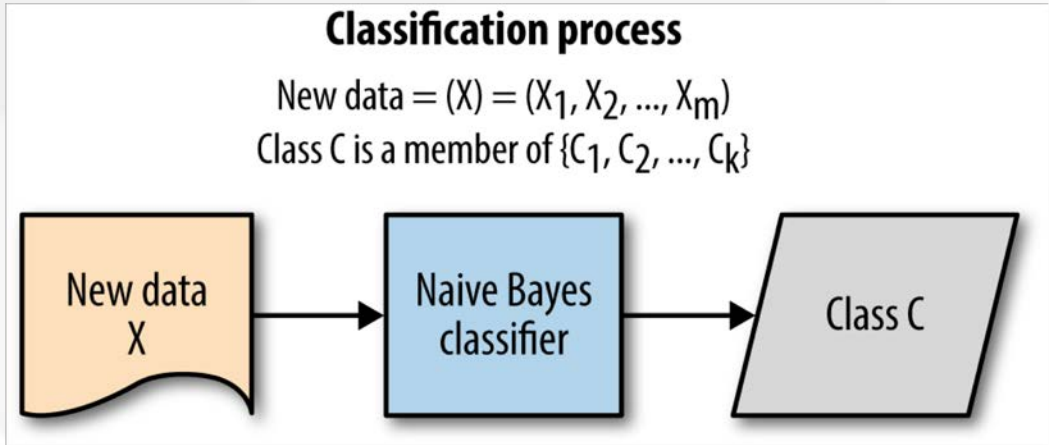
Approaches we took

kNN	Hard to determine the type of distance for classification
Random Forest	Computationally intensive
XGBoost	First used grouped levels, 50% accuracy Need to create a dummy variable for each categorical level,computationally intensive
Neural Network	Hard to choose a proper activate function for output layers to output more than one prediction with proper accuracy

Improved Method

Naïve Bayes Model

- Easy and fast to predict the class of testing data set
- When assumption of independence holds, Naive Bayes classifier performs better
- Performs well with categorical input variables compared to numerical variables.



Improved Method

Naïve Bayes Model

Suitable for datasets with many categorical variables

For 10 predictions,

1st Trial Accuracy: 30.0%

Aggregate numerical to categorical

2nd Trial Accuracy: **42.9%**

Improved Method

Naïve Bayes Model

Suitable for datasets with many categorical variables

For 10 predictions,

2nd Trial Accuracy: 42.9%

Laplacian Correction

3rd Trial Accuracy: **59.6%**

To be continued...

Improved Method

USC Marshall School of Business

Naïve Bayes Model

Recap Assumption: **independency** among variables

- hotel_continent and hotel_country
- user_location_country and user_location_region

have high correlation.

So, remove the former ones.

For 10 predictions,

3rd Trial Accuracy: 59.6%

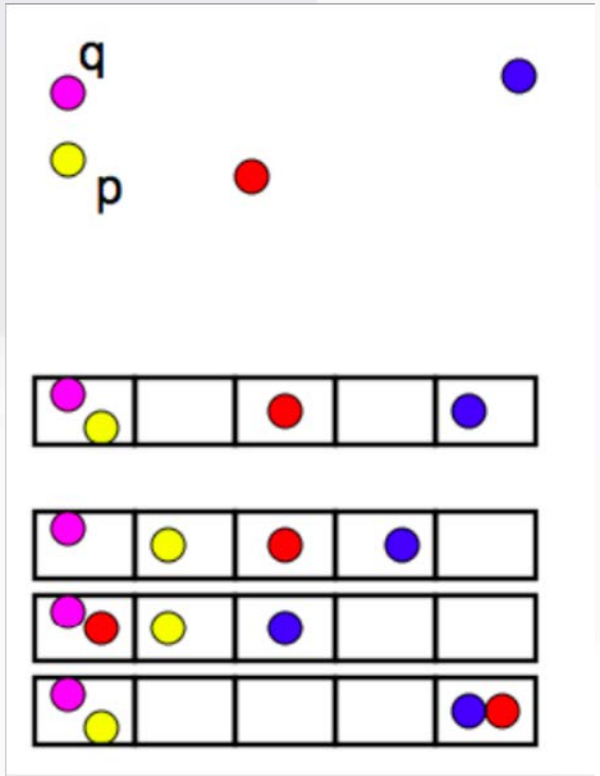
Remove highly correlated attributes

4th Trial Accuracy: **60.3%**

Further Steps

Locality Sensitive Hashing (LSH)

- Commonly used to deduplicate large quantities of documents, webpages, and other files
- Data points are hashed into buckets
- Suitable for categorical variables with a large number of levels



USC Marshall
School of Business

PART 5

Business Insights

To which business scenario can we apply our model?

Business Insights

To Customers

- **Make personalized booking recommendations based on user locations and destinations**
  - Attract new customers
  - Develop existing customers' loyalty

To Business

- **Discover popular hotel types based on user search habits**
  - Partner with hotels with a high booking frequency before competitors do
  - 25% commission in one booking

Q&A

THANKS!

Appendix

- How do hotel clusters look like?
- Final Variables
- Data Cleaning
- Chi-Square Test of Independence

What is the “hotel cluster”?

Final Variable - Numeric

Numerical Variable	Description
orig_destination_distance	Physical distance between a hotel and a customer at the time of search
srch_adults_cnt	Number of adults specified in the hotel room
srch_children_cnt	Number of (extra occupancy) children specified in the hotel room
srch_rm_cnt	Number of hotel rooms specified in the search
cnt	Number of similar events in the context of the same user session
hotel_duration_day	(Check out date - Check in date) in days
srch_ahead_day	(Check in date - User search date) in days

Final Variable - Categorical (1)

Categorical Variable	Description	Number of levels
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)	36
posa_continent	ID of continent associated with site_name	5
user_location_country	The ID of the country the customer is located	19
user_location_region	The ID of the region the customer is located	211
user_location_city	The ID of the city the customer is located	6062
hotel_continent	Hotel continent	6
hotel_country	Hotel country	182
hotel_market	Hotel market	1916
srch_month	The month that a user conducts the booking search	12

Final Variable - Categorical (2)

Categorical Variable	Description	Number of levels
is_mobile	1 when a user connected from a mobile device, 0 otherwise	2
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise	2
channel	ID of a marketing channel	11
check_in_month	The month that the check in date falls in	12
is_booking	1 if a booking, 0 if a click	2

Final Variable - Categorical (3)

Categorical Variable	Description	Number of levels
srch_destination_id	ID of the destination where the hotel search was performed	11365
srch_destination_type_id	Type of destination	8
is_booking	1 if a booking, 0 if a click	2

Data Cleaning

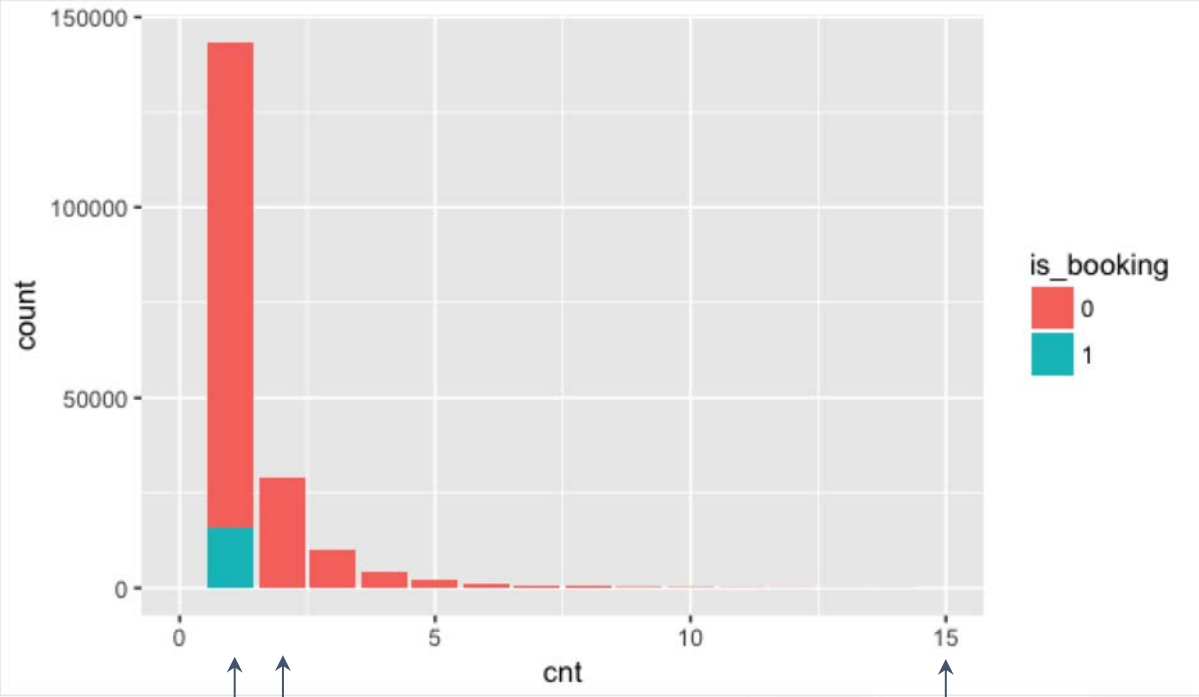
- Delete rows with null values in check in or check out dates
- Delete variables that do not describe the user behavior: user ID
- Delete categorical variables with over 10,000 levels
  - some categorical variables are overlapping (city, region, country)
  - keep those that make sense and with moderate and manageable levels

Chi-Square Test of Independence

- Conducted Chi-square test between each categorical variable and the target variable.
- All the p values are smaller than 0.05, therefore we reject the hypothesis of independence.
- All the categorical variables are correlated to the target variable.

Categorical Variable	p-value
site_name	0
posa_continent	0
user_location_country	0
user_location_region	0
user_location_city	0
is_mobile	9.090527e-69
is_package	0
channel	0
srch_destination_id	0
srch_destination_type_id	0
is_booking	0
hotel_continent	0
hotel_country	0
hotel_market	0
srch_month	1.563823e-190
check_in_month	0

cnt filled by is_booking



Specific destination
Well-prepared travel plan
More valuable users

Just looking around...

cnt

Number of similar events in the context of the same user session

is_booking

1 if a booking, 0 if a click