Project Background

Where2go is an online hotel booking website in China, which accounts for more than 50% market shares. Last year, they started a new project which aims to exploit the value of clickstream data.

They have generated two datasets based on the raw clickstream data: *CUSTOMER_LEVEL* data and *PURCHASE* data.

All the users in the dataset are mobile users. The marketing team wants to understand their customer base better. Through customer segmentation and profiling, they can improve targeting when launching a hotel price discount campaign.

Please note all outputs from this project are generated using Radiant (R package). The purpose of this project is to understand cluster analysis and interpretation of model outputs. Model accuracy and validation are not the focus of this project.

Datasets and Descriptions

Customer level data (*CUSTOMER_LEVEL.CSV*):

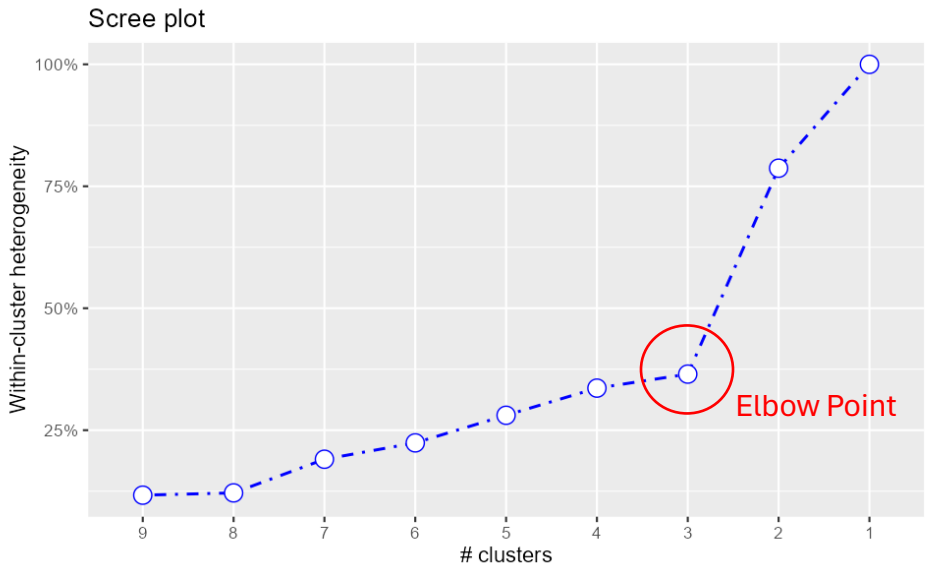
Uid: user id
Session_cnt: number of sessions
Hotel_num: number of hotels clicked in each session
Click_num: number of clicks in each session
Session_time: average time spent in one session (seconds)
Item_time: average time spent in each hotel (seconds)
Purchase_pct: percentage of session with purchase, i.e. conversion rate
Item_price: the average price of hotels viewed in each session.

Purchase data (*PURCHASE.CSV*):

Session: session id
Date: date of session
Hotelid: Hotel id
Num_click: number of click during the session
Tot_time: total time spent in the session (in minutes)
Price: price of hotel (in hundreds)
Buy: dummy variable indicating whether customer purchase in this session. 1 means purchase; 0 otherwise.

Part 1: Customer Segmentation & Profiling

Using *CUSTOMER_LEVEL* data, we will perform hierarchical cluster analysis to generate a scree plot. A scree plot shows the optimal number of clusters based on the elbow point.



According to our scree plot, the optimal number of clusters is 3. Keeping this in mind, we will now perform a K-means clustering using 3 clusters. We will analyze the output and understand how customer behavior differs across the 3 clusters.

```
K-means cluster analysis
Data         : CUSTOMER_LEVEL
Variables    : session_cnt, hotel_num, click_num, session_time, item_time, item_price, purchase_pct
Clustering by: K-means
Standardize  : TRUE
Observations : 8,998
Generated    : 3 clusters of sizes 1,441 | 3,319 | 4,238

Cluster means:
           session_cnt hotel_num click_num session_time item_time item_price purchase_pct
 Cluster 1       12.21     10.30     14.56     3,027.55    206.19     387.35         0.15
 Cluster 2       21.79      3.76      5.58     2,409.50    399.95     398.43         0.28
 Cluster 3        6.16      2.96      3.80       694.95    136.05     438.17         0.07

Percentage of within cluster heterogeneity accounted for by each cluster:

Cluster 1 21.98%
Cluster 2 47.40%
Cluster 3 30.62%

Between cluster heterogeneity accounts for 35.84% of the
total heterogeneity in the data (higher is better)
```

Cluster 1: Cheap-hotel-alternatives-seeking-customers

Based on the cluster analysis output, we see customers in **Cluster 1** have significantly higher means in # of hotels clicked per session (10.3), higher # of clicks per session (14.56) and higher average session time (3,027.55). These customers also have the lowest average hotel price viewed (387.35) with a moderate conversion rate (15%) compared to other clusters. Based on their behavior, we can profile **Cluster 1** as cheap-hotel-alternatives-seeking-customers. These customers tend to spend more time on the web looking through several hotel alternatives that have a low booking price before making a purchase.

Cluster 2: Purposeful-analytical-customers

Customers in **Cluster 2** spend the highest average time in each hotel (399.95) and have the highest conversion rate (28%). We see that they have the highest average number of sessions (21.79) and a relatively low number of hotels clicked (3.76). With this information, we can profile customers in **Cluster 2** as purposeful-and-analytical-customers. These customers know very well the location they are travelling to, the type of hotel they are looking for, hence they tend to spend more time digging into specific hotel listings within that region. They like to spend more time looking through hotel details like perks, amenities, reviews, room type and more so they can get a good grasp of what each hotel offers before making a purchase.

Cluster 3: Spontaneous-unplanned-customers

Customers in **Cluster 3** spend the least amount of session time (6.16) and has the least amount of session count (2.96). They do not view several hotel alternatives and the prices they view are higher (438.17) than that in Cluster 1 and 2. They also have the least conversion rate (7%). With this information, we can profile customers in **Cluster 3** as spontaneous-and-unplanned-customers. These types of customers do not plan head of time on hotel bookings, and they tend to start looking for hotels when it is near or during holidays. This explains why their conversion rate is low because the available bookings left are likely **1) more expensive hotels** and **2) less hotel options** as opposed to availabilities way before holidays.

Part 2: Targeting

The marketing team is interested in understanding the price sensitivity differences among the 3 clusters. Through understanding price sensitivities, the marketing team can more accurately launch a price discount campaign.

Merging our *CUSTOMER_LEVEL* data with *PURCHASE* data and cluster groups, we can generate a pivot table to show the relationship of mean hotel purchase price between clusters.

cluster	price
All	All
1	3.808
2	4.234
3	4.547
Total	**4.141**

According to our pivot table, the average hotel purchase price increases as we move from Cluster 1 to Cluster 3. This agrees with and validates our customer profiling in Part 1. Now, to understand the price sensitivity differences among clusters, we can build a logistics regression model to predict probability of hotel booking purchase and by including interaction terms of price*cluster.

```
Logistic regression (GLM)
Data                 : PURCHASE_cmb
Response variable    : buy
Level                : 1 in buy
Explanatory variables: num_click, tot_time, price, cluster
Null hyp.: there is no effect of x on buy
Alt. hyp.: there is an effect of x on buy

                    OR    OR% coefficient std.error  z.value p.value
 (Intercept)                        -4.579     0.029 -158.661  < .001 ***
 num_click       1.604  60.4%        0.472     0.005   93.613  < .001 ***
 tot_time        1.013   1.3%        0.013     0.000   52.249  < .001 ***
 price           0.930  -7.0%       -0.073     0.007  -10.380  < .001 ***
 cluster|2       5.224 422.4%        1.653     0.029   56.717  < .001 ***
 cluster|3       2.077 107.7%        0.731     0.041   17.972  < .001 ***
 price:cluster|2 0.976  -2.4%       -0.024     0.008   -3.215   0.001 **
 price:cluster|3 0.988  -1.2%       -0.013     0.010   -1.226   0.220

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Pseudo R-squared:0.178, Adjusted Pseudo R-squared:0.178
AUC: 0.815, Log-likelihood: -80425.824, AIC: 160867.649, BIC: 160955.037
Chi-squared: 34743.36 df(7), p.value < .001
Nr obs: 409,856
```

After adding the interaction term between price and clusters and using **Cluster 1** as our benchmark, we can now learn more about how price sensitivity differs across clusters. If we just look at the coefficients for clusters without interaction with price, we can infer that customers in **Cluster 1** tend to have lower conversion rate as depicted by a negative Odds Ratio Percentage of -7.0%. For each cluster's price coefficient, we can interpret them as follows:

Cluster 1: -0.073

A one-unit price reduction is associated with 7.3% sale increase.

Cluster 2: (-0.073 – 0.024) = -0.097

A one-unit price reduction is associated with 9.7% sale increase.

Cluster 3: (-0.073 – 0.013) = -0.086

A one-unit price reduction is associated with 8.6% sale increase.

From the marketing team's perspective when it comes to launching a price discount campaign, customers in **Cluster 2** will be the primary target as these customers are the most sensitive towards a price reduction.

Project Summary

One explanation for **Cluster 2** to be the most price sensitive is that these customers are experienced and analytical. They understand the hotels they are looking for and they have the highest conversion rate among all clusters. Their number of views for different hotels is very limited so if there is a price reduction in any of their target hotels, they are more likely to make a purchasing decision. **Cluster 1** is the least sensitive to a price reduction because these are the ones that also look out for a lower price alternative. They only have a moderate conversion rate compared to the rest of the clusters. If the price reduction isn't that significant, they are likely to search for hotels on other sites and seek out lower prices. Customers in **Cluster 3** are moderately sensitive to price changes because they tend to make spontaneous hotel booking decisions right before or during the holidays. This means that by the time they are checking for hotel bookings, most cheap options will have run out leaving them the expensive options left. Thus, a price reduction on the few availabilities left will somewhat incentivize them to make a purchase, but it won't be as impactful compared to customers in **Cluster 2** because those customers are experienced, up to date with hotel booking market and know exactly what they want. When a customer is experienced and has a good idea of what they want, a price reduction can really speed up the process of conversion.