Analysis on Engagement and Pricing Trends

Expedia Consumer Search Dataset

Helen Li & Alex Yu, Project Group 16

March 31, 2022

Introduction

Expedia is a searching and booking platform for travels. In this project to investigate "Recommendations and Search Patterns of Expedia Consumers," we aim to help collaborators at Expedia, Adam Woznica and Jan Krasnodebski, with understanding consumer search patterns and exploring ways to improve the existing recommendation algorithms.

The data is a random sample of property searches on the Expedia platform from 2021-06-01 to 2021-07-31, and only the search results were used.

Our research questions are:

1. Is the proportion of properties that get clicked on different for properties that are ads and properties that are not ads?
2. Is there an association between how early you try to book in advance and the price of the first listing?
3. What is the range of plausible price buckets that guests with children are interested in?

Is the proportion of properties that get clicked on different for properties that are ads versus not ads?

Rationale and Objective

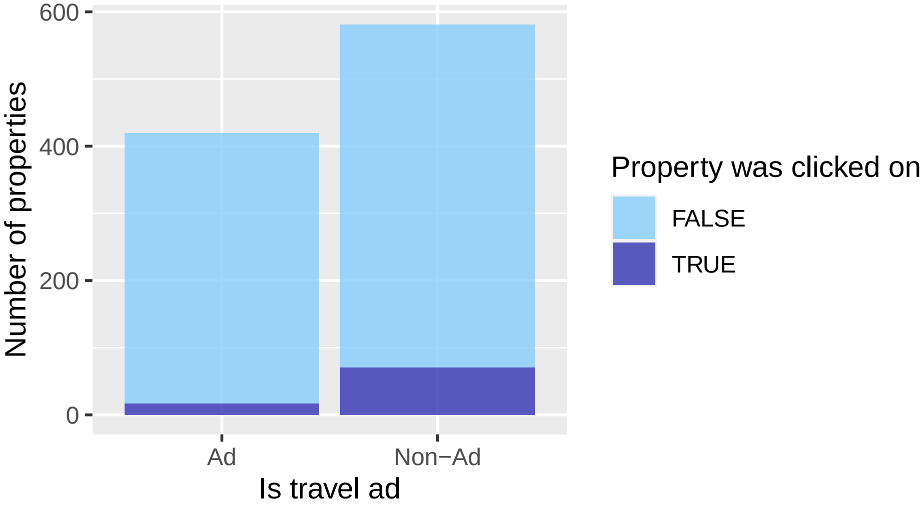
Ads are an important source of revenue for many websites. However, consumers may be less likely to click on properties that are ads because they are distrustful of advertisements.[1] This question seeks to determine whether the proportion of properties that consumers click on is different for properties that are labeled as ads versus not ads. Answering this question may help Expedia know whether sponsoring ads makes a difference to whether their listings get clicked on.

Data Exploring and Wrangling

- Figure 1 below shows the number of first listed properties that are ads and non-ads and also the number of first listed properties that received at least 1 click in each category.
- Created a new variable that records whether the first listing got clicked on, based on whether the given number of clicks on the first listing is greater than 0.

[1] Schomer, Audrey. "Consumer Attitudes Toward Digital Advertising 2021." *Insider Intelligence*, 22 July 2021, Link

Figure 1. Ads and Non−ads Listings



Statistical Method: Hypothesis Testing

Setting Up

- Computed the total number of ad-labeled first listings that got clicked on and divided it by the total number of ad-labeled first listings. Same process for non-ads.
- The difference between these two computed proportions is calculated as the test statistic.

Our Intended Analysis

- We want to know *how* unusual the computed test statistic is.
- We will run a simulation under the assumption that there is no difference between the two proportions. With this simulation's help, we can then analyze if we have enough evidence to reject this assumption.

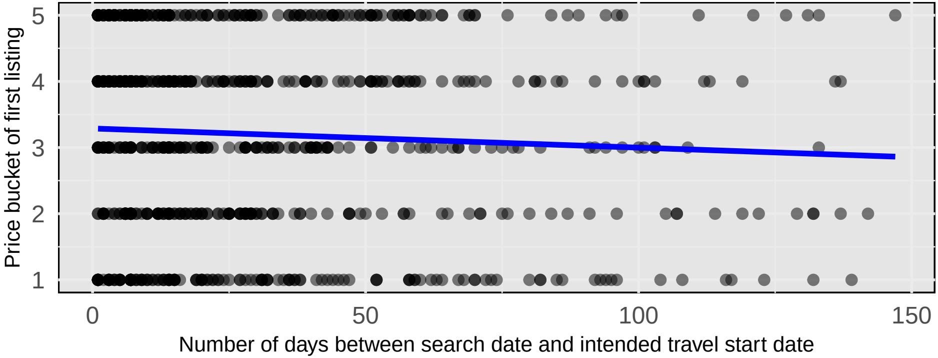
Results and Interpretation

- A p-value of 0 was produced from our test.
    - In other words, it is extremely unusual to obtain our test statistic when assuming that the proportion of first listed properties that get clicked on is the same for those that are ads and those that are not ads.
- Therefore, we have very strong evidence against this assumption.
    - This suggests that the proportion of first listed properties that get clicked on is different for ads and non-ads properties.
- Interestingly, from Figure 1, we also observe that although most of the first listed properties that got clicked on were non-ads, it seems like very few first listings got clicked on overall.

Is there an association between how early you book in advance and the price of the first listing?

Visually, there appears to be a weak negative linear association between earliness of a search and price of the top recommended listing. Earlier searches seem associated with lower prices.

Figure 2. Earliness of Search vs. First Price Bucket



Statistical Method: Simple Linear Regression

- The x-axis measures variable `day_diff`, or the number of days between the date of the search and the intended start travel date.
  - Filtered out 119 observations that searched too late or too early
  - "Too late" means `day_diff` ≤ 0; "too early" means `day_diff` ≥ 150 (5 months)[2]
- The y-axis measures variable `price_bucket1` (an integer in the range 1-5), and lower buckets correspond to lower prices.
  - Filtered out 57 observations that had missing values for `price_bucket1`
- The simple linear regression model $\mathit{price\_bucket1} = \beta_0 + \beta_1 \mathit{days\_diff}$ was used to assess the relationship between these two variables.
  - Estimates for intercept (β_0) and slope (β_1) reported below in Table 1
  - The fitted regression line in Figure 2 visualizes how these estimated coefficients are used to predict `price_bucket1` for a given `day_diff` value

[2] Kozicka, Patricia. "Best Time to Buy and When to Fly to Land Cheap Travel Deals." *Global News*, 13 Jan. 2017, Link

Results and Interpretation

Table 1. Linear Regression Coefficients

	Estimate	Standard Error	t value	P-value
(Intercept)	3.301316574	0.079314836	41.622939	2.294947e-186
day_diff	-0.003838271	0.001834283	-2.092519	3.677521e-02

- The intercept indicates that average price bucket of the first recommended listing is 3.301 when you search on the same day as your intended travel date.
- The slope suggests that when `days_diff` increases by 1 day (when you search a day earlier), `price_bucket1` decreases on average by 0.004.
- Intercept p-value is less than 0.001; slope p-value is between 0.01 and 0.05.
    - The smaller the p-value, the more unusual it is to obtain these estimated coefficient values under the assumptions that the true coefficient values are 0
    - Very strong evidence against the assumption that the true intercept is 0
    - Moderate evidence against the assumption that the true slope is 0
    - Further indicates that these two variables are associated

Results and Interpretation

Root Mean Squared Error (RMSE) & Coefficient of Determination (R^2)

- RMSE is a common measure of prediction error in linear regression models.
  - Trained model on 80% of the data and tested its prediction accuracy on the remaining 20%
- RMSE indicate that the average difference between our model's predicted and actual price bucket values for both training and testing data is about 1.43.
- Smaller RMSE suggests better accuracy, and RMSE for testing data is 1% lower (more accurate) than for training data.
  - Also indicates that the model can apply its learning to new data (data it was not trained on) and perform at around the same level of accuracy
- Only 0.66% (R^2) of the variability in `price_bucket1` is explained by our model.
  - Relative to `day_diff`, there may be other factors that are more related to `price_bucket1`

What is the range of plausible price buckets that guests with children search for?

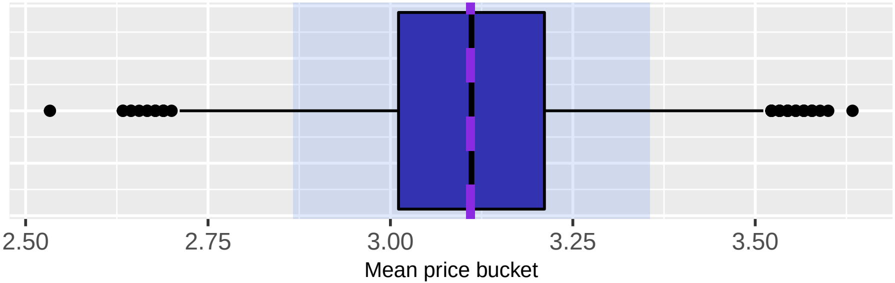
Rationale and Objectives: Understanding the price bucket of a specific population may allow for Expedia to better tailor their recommendations. Since guests with children may have different requirements for their properties, it may be helpful and useful to understand what price range they are interested in.

Statistical Method: Bootstrap Sampling

- Data
  - Restricted the sample to consumer searches that indicated the presence of at least 1 child
  - Filtered out 57 observations that had missing values for price bucket of the first listing
- Analysis
  - Created a new variable, `bucket_samp`, with 90 observations, that stores the price bucket of guests who searched with at least one child
  - We take a bootstrap sample that gives us a range of mean price buckets
  - Bootstrap sampling is a method to estimate a property of the data

Results and Interpretation

Figure 3. Mean Price Bucket of Searches from Guests with at least one Child



- The mean price bucket, indicated by the purple line in Figure 3, is approximately 3.089.
- Moreover, we are 95% confident that mean price bucket that Expedia users booking for at least one child are interested in is between 2.80 and 3.42.
  - If we repeated the same analysis procedure many times and got an interval each time, we expect 95% of these intervals to capture the true mean price bucket value
  - The confidence interval is shown by the blue shaded area in Figure 3

Conclusions and Recommendations

With the statistical analysis detailed in previous slides, we attempt to answer our research questions:

1. The proportion of clicks on the first listed properties that are ads and non-ads is different, and from Figure 1, we can see that the proportion is greater for non-ads properties. However, it is also important to keep in mind that very few properties were clicked on in both cases. Thus, avoiding placing ads properties as the first listing may increase the likelihood that more consumers click on the listing, but overall, we recommend that Expedia should explore more ways to increase interaction between consumers and listings across the board.

2. Searches done more in advance seems to be associated with a first recommended listing that's less expensive. If possible, consumers should aim at searching and booking earlier. Our model performed decently, but very little of the variability in the first listing's price is explained by how early the search is made. Thus, there seems to be other factors that are more related to the price of the top recommended listing. Potential factors include ratings, reviews, and availability and quality of services (food, facilities, etc.). Related variables are worth further exploring to better understand what role they play in Expedia's recommendation algorithm.
3. The mean price bucket of the first listing that guests with children see on Expedia is 3.098, and we are 95% confident that the mean price bucket falls between 2.87 to 3.36. Further studies could couple this information with a classification tree to predict which price bucket a guest is interested in depending on various factors (length of stay, number of children, country, etc.)

Limitations and Acknowledgements

- 1,000 consumer searches were included in the data, but very few consumers clicked on listed properties and/or made a transaction on one of the listings. This highly limited our ability to evaluate how “good” a recommendation is since the consumers usually had no response to the recommended listings from their searches.
- While we have a random sample of searches, the time frame is only from June to July of 2021. We do not know if we can extrapolate our results to different months or years.
- `price_bucket1` is a discrete variable that only takes on 1, 2, 3, 4, or 5 as its value. Instead, storing the price information as a continuous variable (be any number within a certain range) may provide more specific information about the prices and thus make our analyses more precise in questions 2 and 3.
- Our method to find the mean price bucket for guests with children works best on large samples that are roughly symmetric and continuous. Moreover, if the sample is *not* representative of the population, then our mean will be inaccurate.
- **We would like to thank the Expedia Group for the provided data and the STA130 teaching team for the support throughout the term**