Browsing and Travel Patterns of North-American Expedia Customers

Chin Chin Jim and Lina Jin, Group 102

March 31, 2022

Introduction

- In this research project, we will explore data on Expedia search pages in the interests of Expedia developers and stakeholders.
- Our data analyses will help assess the efficiency of certain features on Expedia search pages, and investigate the searching patterns of Expedia users. Any correlations or associations that we conclude can be used to make changes to Expedia that will make the searching process more streamlined.
- The population of interest is North American Expedia users in the year 2021. We will use a sample of North American Expedia searches from June to July 2021.

Question 1: Is the proportion of Expedia consumers looking for hotels in the summer 0.25?

- Summer (around May to August), is known as "High Season" for American travelers, making it the busiest season for hotels.
- With our data, we can see if the most popular season (over 25%) of the Expedia consumers follow the pattern and look to travel during the summer.
- We can use single-sample hypothesis testing to find evidence for the association.

Hypotheses

Null hypothesis (H_0): The proportion of Expedia consumers looking for hotels in the summer is 0.25.

$$H_0 : p_{summer} = 0.25$$

Alternative hypothesis (H_1): The proportion of Expedia consumers looking for hotels in the summer is not 0.25.

$$H_1 : p_{summer} \neq 0.25$$

Data visualizations

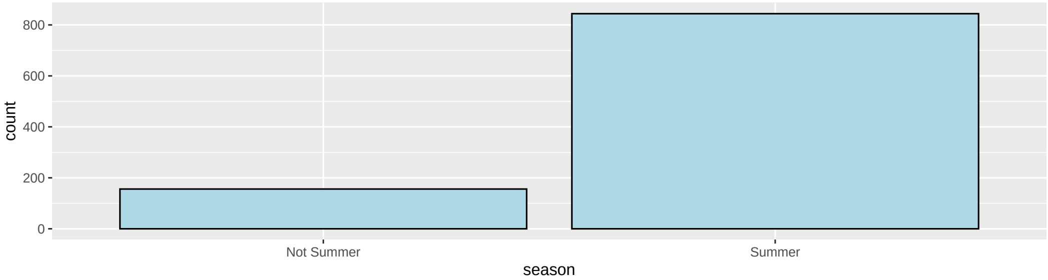


Figure 1

Figure 1 shows that our data tells us that .844 of the Expedia customers are looking to stay during the summer. We can easily see that the number of Expedia customers that plan to stay during the summer is significantly more than those who plan to stay in other months.

Setup of the Statistical Method

By creating a simulation that visualizes a distribution of data with the assumption that our null hypothesis is true, we can see if the proportion of values from our simulation that are at least as extreme as the sample data (p-value).

Data Summary

- We created a new variable called "season" based off of "checkin_date", where customers who are starting their stay in June, July, or August take on the value "summer" and for any other month they take on the value "not summer" (see Fig. 1).

Conclusion

- We observe that none (or a very trivial amount) of the values from our simulation are at least as extreme as the original statistic (.844), then we have strong evidence against our null hypothesis.
- Expedia could increase advertisements or hotel property availability during summer time to increase hotel bookings on their website.

Limitations

There are limitations to single hypothesis tests because we can only conclude there are differences between those who look to book in the summer versus those who do not. There could be other factors that contribute to why the proportion of travelers during the summer is not 25%.

Question 2: Is the average number of clicks the same for listings that are travel ads versus not travel ads?

- Travel ads can be sponsored by travel companies that wish to promote their services on Expedia. We can use data on the "success" of these travel ad listings, measured in clicks, to examine their performance compared to non-travel ad listings
- We will analyze if there is an influence on the number of clicks on a listing based on whether or not the listing is a travel ad.

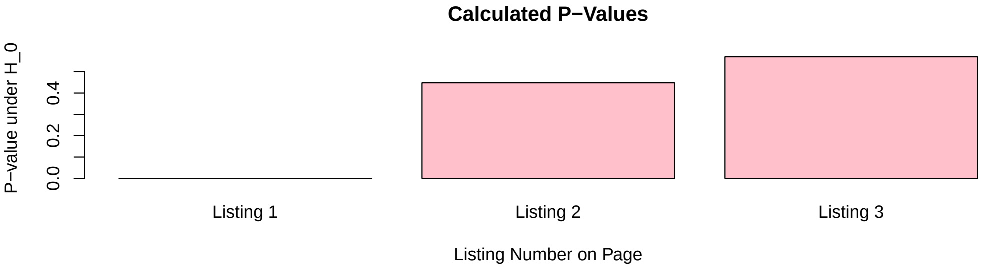
Hypotheses

Null hypothesis (H_0): There is no difference in the mean number of clicks to a property listing for listings that are travel ads and listings that are not travel ads.

$$H_0 : \mu_{travel\_ad} = \mu_{not\_travel\_ad}$$

Alternative hypothesis (H_1): The mean number of clicks are different for listings that are travel ads and listings that are not travel ads.

Data visualizations



Visualization (cont.)

- We chose to visualize this data using a barplot (Figure 2) which can be used to easily compare the values of the three different p-values, and determine which are $<$ 0.10.
- Figure 2 shows us that only listing 1 has a p-value < 0.1 after running our two-group hypotheses tests.
- Listing 1 has a p-value of 0, Listing 2 has a p-value of 0.448, and Listing 3 has a p-value of 0.570.
- The p-values also seem to increase with each listing further down the page, suggesting a decreasing relevance between travel ads and number of clicks past the first listing.

Statistical Methods

- By using whether or not a listing was a travel ad or not, and the number of clicks it received, we created a two-sided hypothesis test simulation to test if there was any difference in travel ads vs non travel ads for number of clicks.
- While simulating, we assumed that the number of clicks would be equal among the two groups.
- We then compared the results of the simulation to the actual data collected, to see if our data was unusual under the case that the two groups did receive equal clicks.

Data Summary

We used is_travel_ad1, is_travel_ad2, and is_travel_ad3 (independent variables) with num_clicks1, num_clicks2, and num_clicks3 (dependent variables) respectively.

Conclusion

- For Listing 1, there is very strong evidence against the Null Hypothesis that there is no difference in the mean number of clicks for travel ads vs non travel ads.
- As for Listings 2 & 3, there is no evidence against the Null Hypothesis. Whether a listing is a travel ad or not is only relevant to the number of clicks it receives if it is the first listing on the page.
- Customers may be less inclined to click past the first listing in general, decreasing total click count for the second and third listings
- In the interest of Expedia developers, they may want to charge more for travel companies listing ads in the first listing spot, since these get more click traffic

Limitations

- Our data only gives the number of clicks within the first 180 minutes of the search, so it may help our hypothesis if we could have access to the total number of clicks within a given period.
- We could make a better conclusion if we had access to data of more properties from Expedia, not just the first three listings

Question 3: Is there an association between the number of people going on a trip and the month of the check-in date?

The time of year can influence the travel patterns of different customers. By looking at the trend of the times that people search to check in on Expedia and the number of people they plan to travel with, we can see if there is a correlation between the two variables.

Hypotheses

Null hypothesis (H_0): There is no association between the number of people going on a trip and the month of the check-in date.

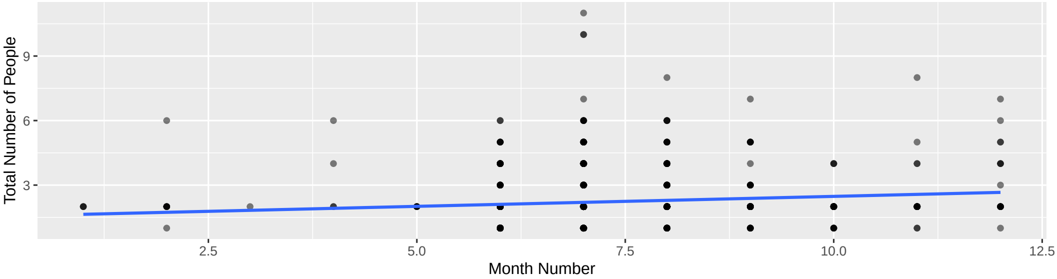
$$\hat{\beta}_1 = 0$$

Null hypothesis (H_1): There is an association between the number of people going on a trip and the month of the check-in date.

$$\hat{\beta}_1 \neq 0$$

Data visualizations

Fig 3: Total Number of People Attending Based on Month of Expedia Search



Setup of the Statistical Method

Variables	Values
$\hat{\beta}_0$	1.54927521
$\hat{\beta}_1$	0.09263662
R^2	0.01328409
p-value	2.598475e-04

- We fit an estimated regression line based (the “best” line) from our plot to demonstrate the relationship between the month number and the total number of people.
- The estimate of β_0 is about 1.55, which means that when the month is 0, the average group of people is 1.55.
- The estimate of β_1 is about .09, which means that with every month, we can account for a .09 difference in the number of people.
- Since the p-value is small, we know that we have strong evidence against the null hypothesis.

Conclusion

- There is a positive weak correlation between check-in date and total number of people.
- This visualization allows us to see that the months of the highest number of people are the summer months (around-June, July, August), and late-Winter months (late-November, December).
- This suggests that the number of people attending trips is higher during western holiday seasons (Christmas, summer vacation, etc.).
- Expedia may look to charge more or increase ad presence during these times.
- Also they may prioritize properties with a high quantity of rooms (to accommodate for increased total guest count) and place them at the top of each search page to make it easier for the customer during peak seasons.

Conclusion (cont.)

Limitations

- Since our R^2 is small, very little of our variability is explained by our regression model.
- Data from more consumers and/or from a larger sample of dates can help solidify if there is an association between the month and the number of people planning to stay within our population, to see if the trend follows.
- There may be other factors that explain the increased number of people depending on the month, so access to more variables can help us see if there are other associations between the number of people and another variable.

Summary of Findings

Question 1: The majority of Expedia customers book their trips during the summer.

Question 2: The number of clicks is affected by whether a listing is a travel ad or not, but only for the first listing on the page.

Question 3: There is a weak positive association between the month of Expedia searching and the number of people attending a trip, with a higher total number of people in the summer and late winter months. This is possibly due to summer vacation and winter holiday breaks.

Summary (cont.)

Next Steps

We can:

- Promote Expedia more during the summer months and charge more for ad space during that time.
- Charge more for travel ads in the first listing space.
- Prioritize hotels that have enough space for larger groups of people during the peak months.
- Include data on actual purchases rather than searches in future analyses, to further explore the spending habits of Expedia customers