Data Analysis of Factors that Influence the Listings on the Expedia Website

For all searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31

Kavesh Biersay, Vinayak Maharaj, Kowshik Mazumdar, Madhav Kanna Thenappan,
ProjectGroup 65

March 31, 2022

Introduction

- The objective for this project is to help collaborators at Expedia, Adam Woznica and Jan Krasnodebski, investigate "Recommendations and Search Patterns of Expedia Consumers".
- We are exploring what makes for a good recommendation and are looking for patterns in what consumers search for on Expedia booking sites.
- We are provided with the "expedia_data" dataset that consists of 1000 searches which accounts for of all searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31.
- Our target population for presenting the results is the general public therefore appropriate syntax and semantics will be employed to avoid ambuigity with use of technical jargon that non statiticans can't understand

Objectives

List of research questions

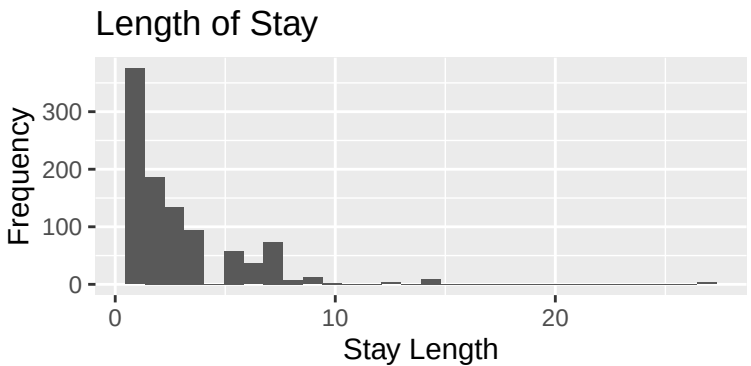
For this investigation we have chosen three research questions, the first being an exploratory question to determine the general information about the listings and the following questions were identifying the influence of specific factors on the search results

- What is the range of plausible values for the average stay length of each search consumer at 95% level confidence?
- Is the proportion of the first listed property a travel ad, equal to 50%?
- Is there a difference in the average rating of first listings with free cancellation and first listings with no free cancellation?

Question 1: Introduction

- Research question - **What is the range of plausible values for the average stay length of each search consumer at 95% level confidence?**
- For this question, the variables which will be utilized within this method would be the 'checkin_date' and 'checkout_date' variables within the data set. From the **checkin_date** and **checkout_date** variables, a new variable named **stay_length** will be created, which counts the number of days within consumers' stays.
- Our goal from this analysis will be to to determine an accurate range of values for the mean number of nights that each customer stayed in the listings provided
- In order to carry out this research question, we would use the method of Bootstrap Confidence Intervals specifically at the 95% level.

Question 1: Data Visualization



The histogram shows the distribution of stay length of all bookings in the sample data. It can be observed that the histogram is right skewed, unimodal and centred at stay length of 1 day.

Question 1: Statistical Analysis

- A bootstrap confidence interval test was used to narrow down the range of values which contain the mean amount of nights
- It was specifically done at the 95% level of confidence so that the results gathered would be more accurate in comparison to other levels of confidence which will result in more usable data for the Expedia researchers and greater application of results.
- We found the distribution of nights stayed in which contained the true average number of nights stayed
- We resampled the original sample with replacement 10,000 times which allowed for a simulation of sampling the population distribution however more feasible
- Lastly the new distribution formed was narrowed down further by excluding 2.5% of the extreme values arriving at our 95% confidence interval

Question 1: Results

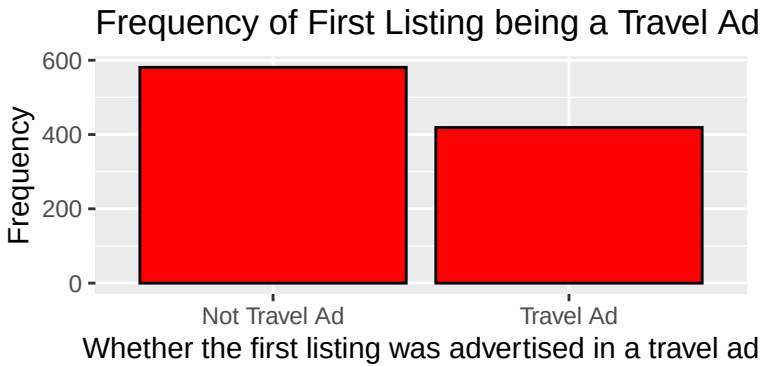
2.5%	97.5%
2.897	3.258

- From this bootstrapping investigation we are 95% confident that the mean stay length for listings on the Expedia website from 2021-06-01 to 2021-07-31 is between 2.897 and 3.258 nights
- This conclusion does not give the exact mean stay length however it gives a plausible range of values thats small enough that it will still be useful to the researchers
- In comparison to Expedia's competitor, Airbnb, this range contains a number that is lower (Airbnb's average stay length is 3.9 nights) which leads us to believe that there is room for improvement with the listings given to consumers

Question 2: Introduction

- Research question - **Is the proportion of the first listed property a travel ad, equal to 50%?**
- A one proportion hypothesis will be used with the research question taken as the assumption or null hypothesis in which we will determine the level of evidence against.
- The first listings were chosen as the variable being investigated because it shows the first preference from Expedia to the consumer criteria.
- The proportion being greater than 50% was chosen because my friend noticed that a majority of first listings have been in travel ads hence we would like to test this hypothesis.
- The variable chosen is "is_travel_ad1" which determines whether the first listing was advertised in a travel ad.

Question 2: Data Visualization



The above bar graph shows the frequency of the first listing being a Travel Ad or not. It can be observed that out of the sample data, 600 of the first listings were not Travel Ads, while the rest 400 were Travel Ads.

Question 2: Statistical Analysis

- Null Hypothesis (H_0): Among all the searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31, the proportion of the first listings which are advertised in travel ads is equal to 50%.
- Alternative Hypothesis (H_1): Among all the searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31, the proportion of the first listings which are advertised in travel ads is not equal to 50%.
- The test statistic, the p value, can be calculated from a graph made up of several repetitions of simulated data with proportions of advertised listings in each sample.
- The p value obtained can be used to come up with a valid conclusion on whether to reject or accept the null hypothesis.
- The test included 10,000 simulations to determine the proportion of 1st listings advertised as travel ads in 10,000 samples.

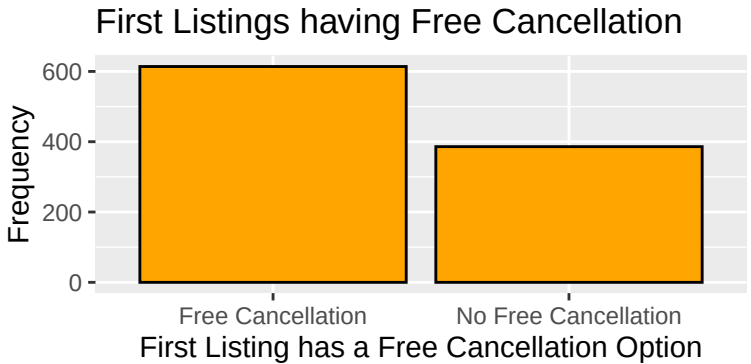
Question 2: Results

- The actual proportion of first listings advertised in a travel ad was found to be 0.419.
- The p-value was found to be 0.
- Since the p-value is 0 which is less than 0.001, we conclude that we have very strong evidence against the null hypothesis that among all the searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31, the proportion of the first listings which are advertised in travel ads is equal to 50%.
- Further research is requested to get a better idea on the influence of travel advertisements on the effectiveness of search results for consumers

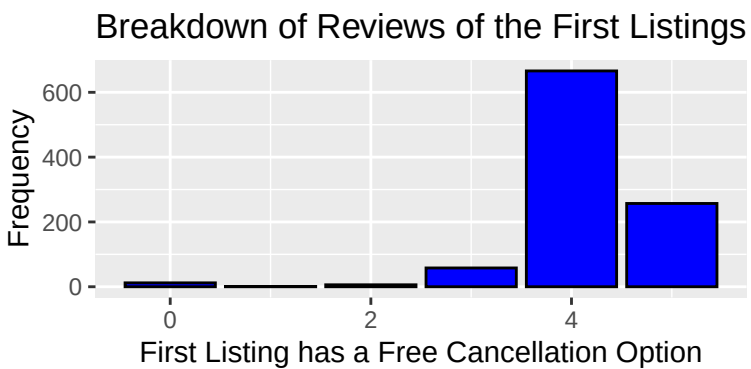
Question 3: Introduction

- Research question - **Is there a difference in the average rating of first listings with free cancellation and first listings with no free cancellation?**
- A two proportion hypothesis will be used with the research question taken as the assumption or null hypothesis in which we will determine the level of evidence against.
- The first listings were chosen as the variable being investigated because it shows the first preference from Expedia to the consumer criteria.
- The variables chosen are "is_free_cancellation1" which determines whether the first listing has a free cancellation option and "review_rating1" which is used to indicate the average rating of the 1st listing.

Question 3: Data Visualizations



The bar chart shows the frequency of First Listings in the sample dataset that had the Free Cancellation Option or not. It may be observed that 600 of the listings had the Free Cancellation option while 400 listings did not.



The bar chart shows the split for the reviews given for all first listings in the score range 1-5. The modal review score is 4 with over 600 listings whereas number of review scores were lower from the range 0-2 in comparison to 3-5. The data was also concentrated in this same 3-5 review score range.

Question 3: Statistical Analysis

- Null Hypothesis (H_0): Among all the searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31, there is no difference between the the average review rating for first listings between the groups which have free cancellation and don't have free cancellation.
- Alternative Hypothesis (H_1): Among all the searches on the Expedia website that span the period from 2021-06-01 to 2021-07-31, there is a difference between the the average review review rating for first listings between the groups which have free cancellation and don't have free cancellation.
- The test statistic, the p value, can be calculated from a graph made up of several repetitions of simulated data with differences in the means between first listings with and without free cancellation.
- The p value obtained can be used to come up with a valid conclusion on whether to reject or accept the null hypothesis.

Question 3: Results

- The actual difference between the average review first listings with and without free cancellation was found to be 0.1286729.
- The p-value was found to be 0.0085.
- We have a p value of 0.0085, this means we have strong evidence against the null hypothesis that there is no difference between the average review rating for first listings between the groups which have free cancellation and don't have free cancellation.
- This means that there is a clear difference in the review ratings for listings with and without free cancellation. Hence Expedia researches must have greater focus on including more listings with free cancellations.

Conclusion

- From the bootstrapping investigation we can state with 95% confidence that the mean stay length for listings on the Expedia website in the specified timeframe is between 2.897 and 3.258 nights.
  - In comparison to Expedia's competitor, Airbnb, it is a lower average stay length with Airbnb averaging 3.9 nights per customer, hence we have decided to analyze certain factors which may influence this.
- From the proportion based hypothesis testing we can conclude that we have very strong evidence against the fact that among all the searches on the Expedia website in the specified timeframe, the proportion of the first listings which are advertised in travel ads is equal to 50%.
- We can conclude that the option of first listings having free cancellation does lead to a difference in the average review rating according to the two proportion hypothesis which showed that there is strong evidence against there being no difference between the average review rating for first listings between the groups which have free cancellation and don't have free cancellation

Limitations

- For this investigation specifically questions 1 and 2, only thefirst listings we used so we can't necessarily generalize the findings for the whole dataset hence further analysis is needed.
- The data set was only for a small duration of time specifically during the covid-19 pandemic hence travel restrictions to countries would have had heavy influence on the data collected.
- The possible factors in the data dictionary which can influence listings chosen such as free wifi and breakfast were removed in the data set given hence possible other reasons for the indicator variable specifically in question 3 could not be recorded

References and Acknowledgements

- Airbnb Economic Impact. Retrieved March 20th, 2022, from https://blog.atairbnb.com/economic-impact-airbnb/#:~:text=Airbnb%20guests%20stay%20on%20average,%24713%20for%20the%20average%20visitor.

- The authors will like to thank Colin for their helpful suggestions on the visualizations used for the 2 proportion hypothesis tests and simplfying our bootstrapping interval approach for this project
- The authors will like to thank Uzair for their helpful suggestions in reevaluating our research questions including solidifying our approach for the porportion based hypothesis testing and overall selection of the 3 research questions