Data-Based Commercial Strategy Report for Expedia

Free Cancellation, Traveling with Children, Range of Star Rating

Haochong Yang

March 31, 2022

Objective

- The **three** research questions my project focused on are:
  - "**Is the proportion of listing properties that can be cancelled freely in all searches 50%?**"
  - "**Is the proportion of searches that took children the same in June and in July?**"
  - "**Given a search including under age, what is the plausible range of star rating on average for the hotel that is suitable for this search?**" (Under age is defined as being either a child or infant)

Overall Introduction

- I focused on the sample data provided by Expedia and conducted this project.
- By focusing these topics, I can determine whether consumers are more inclined to properties providing free-cancellation policy. I can know whether the the pattern of traveling is affected by different schedule for schools. I can also see what range of star rating on average would parents choose if they bring their kids on trips.
- The sample provided has a great importance for Expedia because the sampling period is in summer and has the **peak amount of tourists**. I can give suggestions to Expedia on how to improve their work based on the given data and my analysis.

Data Summary

- The **data** that are used for analyzing in this project are a random sample consisted of 1000 searches made by consumers who made at least one click on the Expedia website during a period from 2021-06-01 to 2021-07-31.
- The variables being used in this project are the number of children, infants, cancellations and star ratings. For each research question, we adopt data wrangling methods for each variable above to be more suitable for our analysis. For each question, I first **select relevant variables**, then **created new variables** based on my goals.
- Variables will be **further specified** and illustrated in detail under each research question.

Research Q#1: Statement of the Question

Research Question 1

- Is the proportion of listing properties that can be cancelled freely in all searches 50%?

Variable Used

- The number of cancellations of three listings for each search.
- A new variable **"total cancellation"** is created in the data set by adding the values of three existing variables together. If there's no free cancellation allowed among three listed property, this variable shows "No Cancellation". If there is free cancellation allowed, this variable shows "Has Cancellation".
- Then, only **"total cancellation"** is selected because of its relevance.

Research Q#1: Set up of Statistical Model/Method

- I used one proportion hypothesis test, since it can help us determine whether the result we got from the sample is reliable.
- My **null hypothesis** is that the proportion of searches that contain at least one property with free-cancellation is 50% in all searches. Notice that $p_{has\ cancellation}$ is the proportion of searches that contain free cancellation in all searches.

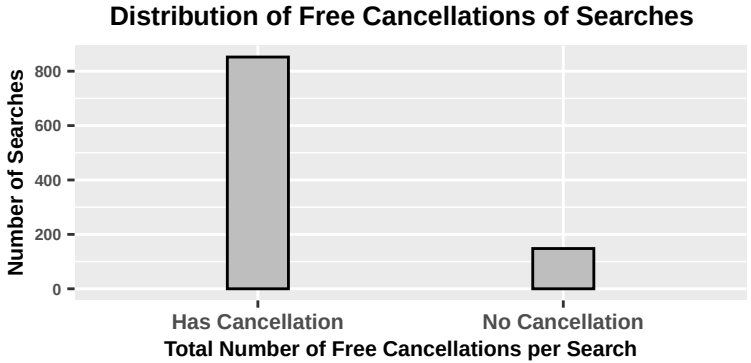
$$H_0 : p_{has\ cancellation} = 0.5$$

Similarly, the **alternate hypothesis** is that the proportion of searches that contain at least one property with free-cancellation is not 50% in all searches.

$$H_1 : p_{has\ cancellation} \neq 0.5$$

- Under the condition that the **null hypothesis** is true, I generated 10000 simulations.

Research Q#1: Relevant Visualization



- This graph displays the distribution of free cancellations of searches. Number of searches containing cancellation-free properties is larger than the searches that doesn't contain.

Research Q#1: Result

- The **test statistic** is 0.148, which means that the proportion of searches that don't contain free-cancellation at all in th entire sample is 0.148. The result from the above scenario is a **p-value equals 0**.
- The extreme **p-value** indicates that no value from the simulations is at least as extreme as the test statistic from the sample.
- This further illustrates that I have an **extremely strong evidence against** the null hypothesis, which suggests that the proportion of searches that contain free-cancellation in all searches that have been made on Expedia's website during the period from 2021-06-01 to 2021-07-31 isn't 50%.

Research Q#2: Statement of the Question

Research Question 2

- Is the proportion of searches that took children the same in June and in July?

Variable Used

- Variables used include number of children and check in date for each search.
- A new variable is created to determine whether the search contains children or not. The type of variable check in date is modified for easier extraction.
- Searches in June and July are filtered out specifically since we are comparing the proportion of bringing children out on trips. Then, only check in date and number of children are selected because they are relevant to our question.

Research Q#2: Set up of Statistical Model/Method

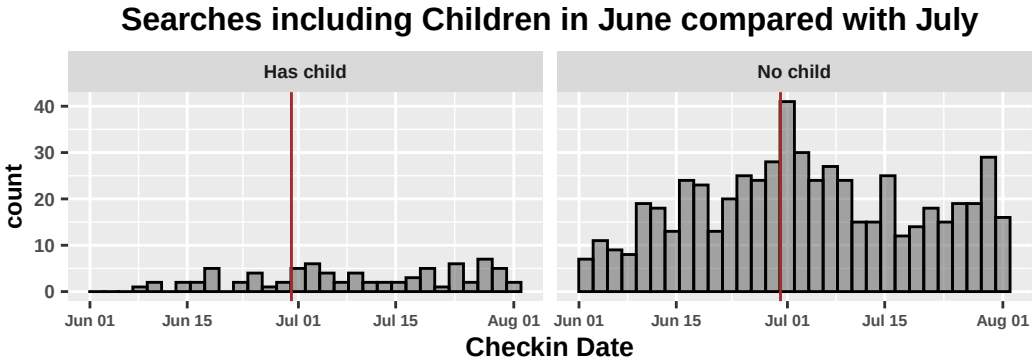
- I used a **two proportions hypothesis test** for this question. By comparing the proportion of searches that include children in all searches in June and the proportion of that in July, I can determine whether the proportion changed or not. I also generate simulations under our null hypothesis is true. If there are a certain number of proportions is as extreme as our real proportion from the sample, I don't have much evidence to reject the null hypothesis.
- My null hypothesis (H_0) is that the proportion of searches that include children are the same in June than July.

$$H_0 : p_{search\_has\_child} = p_{search\_no\_child}$$

So, the alternative hypothesis (H_1) should be that the proportion of searches that include children are different in June than July.

$$H_1 : p_{search\_has\_child} \neq p_{search\_no\_child}$$

Research Q#2: Relevant Visualization



- The red line above separates June and July. Comparison can be seen by viewing the same sides of two graphs above. The number of searches with children **increased slightly** from June to July, since more data is clustered at the right side of the brown line, which represents the end of June.

Research Q#2: Result and Interpretation

- The **test statistic** is -0.06586 approximately, which means that the difference in proportion of searches that contain children between June and July is -6.586%. In the simulation, **10000 differences** between the proportion mentioned above are generated and a **p-value of 0.0063**.
- This extremely small p-value suggests that I have strong evidence to reject our null hypothesis, hence showing that the proportion of searches that include children are different in June than July.

Research Q#3: Statement of the Question

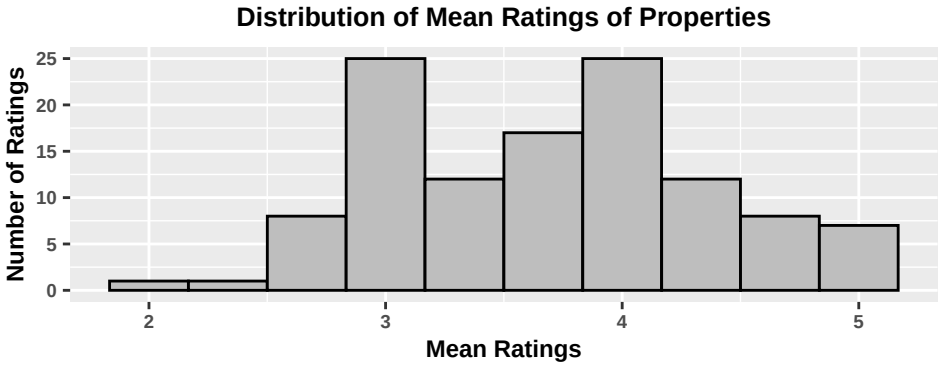
Research Question 3

- Given a search including under age, what is the plausible range of star rating on average for the hotel that is suitable for this search?

Variable Used & Processing

- Variables used include number of children, infants, and star ratings for three listings for each search.
- Number of under age is the sum of children and infant.
- Mean star rating is created by finding the average of the three ratings in each search.
- Filtered out searches that include children or infant.
- Created two new variables, one for mean ratings and one for number of under age.
- Removed variables that are not involved in the question.
- Calculated the mean value of star rating for each search.

Research Q#3: Relevant Visualization



- This bar graph shows the distribution of mean rating of each search that includes children and infants. We can see that the mean ratings which appeared the most are 3 and 4. There are more searches on high average ratings than on low average ratings.

Research Q#3: Set up of Statistical Model/Method

Method

- I used bootstrap sampling method to further investigate our question. By using the bootstrap sampling method, I can find a confidence interval, where the mean ratings fall in for searches that includes children or infant, from the sample I have.

Bootstrap

- Generated 10000 bootstrap samples from the original sample.
- Each bootstrap sample has a mean star rating value being documented.

Research Q#3: Result and Interpretation

- The 5th percentile among all the 10000 values generated is 3.58, which is the smallest value that is larger or equal to 5% of all values. Similarly, the 95th percentile is 3.79.
- So, the 90% confidence interval is from 3.58 stars to 3.79 stars on average. If I repeated this sampling procedure many times, 90% of the confidence intervals would include the true mean star ratings for the searches.
- Can say: I am 90% confident that the mean star rating for all searches that includes children or infant on Expedia during a period from 2021-06-01 to 2021-07-31 is between 3.58 stars and 3.79 stars, which is an interval higher than the middle between 0 stars and 5 stars.

Limitations

Seasonal Limitation

- Children are having summer vocation in June and July, so many searches made in that period of time include an abnormal number of children. This phenomenon is biased and is seasonal limited since there would be less number of children involved in searches made on school days. If Expedia wants to explore annual travel patterns, data with more variance on travel dates are important.

Data Type Problem!

- Most of the variables in the given data set are either discrete numerical data or nominal data, so these types of data can only provide a little variation when I'm analyzing it. The small differences among values may lead to a large deviance because I can't really see the trend changing, so the result predicted should be expected to have certain amount of error. Also, I simplified star rating as numerical data, but it actually requires hotel to pay more efforts to acquire a higher rating, so further quantification is needed for processing this variable.

Overall Conclusion

Conclusion for Question #1

- Consumers are finding properties with free-cancellation policy. As for the platform, I suggest that Expedia should recommend properties with free-cancellation policies, and subsidize those properties without this policy.

Conclusion for Question #2

- The result indicates that after the schools close, more children are expected to be on trips in July. So, I strongly suggest that Expedia should recommend more properties that provide family services to consumers more often in July than June.

Conclusion for Question #3

- Based on the result, I would like to encourage Expedia to recommend more properties with ratings from 3.58 stars to 3.79 stars for the searches that specified children or infants are included in the trip. Expedia shouldn't recommend lower star ratings properties because parents are seeking higher quality service for their kids.

References / Acknowledgement

- ggplot fonts and structure https://www.rstudio.com/resources/cheatsheets/
- ggplot color modification https://github.com/rstudio/cheatsheets/blob/main/data-visualization-2.1.pdf
- Date data type introduction https://blog.exploratory.io/filter-with-date-function-ce8e84be680
- Filtering a data type variable https://stackoverflow.com/questions/28335715/r-how-to-filter-subset-a-sequence-of-dates
- Modifying labels and axis https://ggplot2.tidyverse.org/reference/labs.html
- Changing output Theme https://hartwork.org/beamer-theme-matrix/

I would like to express our gratitude and thanks to Prof. Caetano, Prof. Schwartz, my TA Nick, as well as the TA that is reading our project right now! Thanks for your efforts!