Testing and Investigating the Assumptions and Relationships From Couples' Expedia Search Results Recorded From 2021-06-01 to 2021-07-31

Umair Hussain and Jazli Muhammad Khairi Leong

March 31, 2022

Introduction

- The population we are looking to investigate are Expedia searches with adult counts equal to 2 with a child and infant count equal to 0 for the search data recorded from 2021-06-01 to 2021-07-31

- We presume these searches are likely to come from couples

- The goal of this project is to research and examine trends inferred from couples Expedia search queries and their respective search results

- Our audience are the collaborators at Expedia

Data Summary

Filtered out observations by their adult, child, and infant count. Specifically looking for search queries with an adult count of 2, child count of 0, and infant count of 0. This is done to inspect search results from potential couples.

- Created variable "stay_length" which takes the checkout_date and checkin_date of a search query and calculates their difference
- Created variable "avg_star" is created to measure the returns average star rating of the top three listings from the search query
- Created variable "avg_review_count" where it calculates the average number of reviews for the top 3 listings of a search query
- Created variable "avg_rating" also measured the average review rating for the top three listings
- Created variable "is_short" uses the newly created "stay_length" variable to distinguish if their length of stay was short or not. If stay_length was greater than 2 it would be FALSE, and if it was less than or equal to 2 it would be TRUE
- Created variable "is_reductions" checks if the top three listings of a search had a price reduction, if so the variable is TRUE otherwise it is FALSE

Research Question 1

Question
Is the proportion of search queries from the population that result in the top 3 listings all having price reductions equal to 50

This questions seemed interesting as we wanted to understand whether a significant number of the listings had price reductions for searches in the population. This analysis gives Expedia information on whether they're giving search queries with only 2 adults a significant number of price reduced listings.

The variable we used to solve this questions is is_reduction. To solve this question it would be most appropriate to use a **One Group Hypothesis Test**. The result from a One Group Hypothesis Test is the amount of evidence we have against our null hypothesis.

Hypotheses

Null hypothesis (H_0): The proportion of searches from the population that result in the top 3 listings all having price reductions equals 50

$$H_0 : p_{reduced} = 0.50$$

Alternative hypothesis (H_A or H_a or H_1): The proportion of searches from the population that result in the top 3 listings all having price reductions does not equals 50

$$H_1 : p_{reduced} \neq 0.50$$

where $p_{reduced}$ The proportion of the population that has search results where the top 3 listings have price reductions

Methods

- We solved this question with a Single Group Hypothesis Test
- To do this we filtered out the data as described by our population and grouped the observations by whether the search resulted in the top 3 listings all having a price reduction
- After wrangling the data we created a simulation using R assuming that the null hypothesis is true. After doing so we examine how extreme the proportion for our sample is compared to the simulated data.
- Doing all of this allows us to come to a conclusion and evaluate the evidence against the null hypothesis

Results

Figure 1

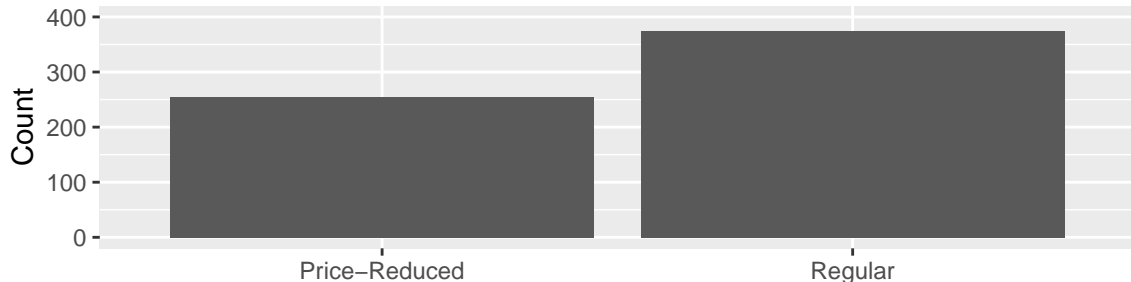Number of Searches With the Top 3 Listings Having a Reduction or Not

Figure 1 depicts $p_{reduced}$ for our sample of the population.

After performing the One Group Hypothesis Test we realized that our proportion seen in Figure 1 was very extreme for our null hypothesis. This led us to conclude that we have very strong evidence against our null hypothesis. This means that the proportion of search results that only contain price reduced listings for the top 3 is not equal to 50%.

Research Question 2

Question

Is there a difference between the mean average star rating for searches in the "short" and "long" group from the population?

Here we define a search to be in the short stay category as a search result where the stay_length is less than or equal 2 and stay_length being greater than 2 is a search in the long group.

This questions seemed interesting as we wanted to understand whether a people in the short group were recommended a different average star rating than people in the long group. This analysis gives Expedia information on whether they're showing one group of people different hotels depending on if their search specifies a short or long time.

The variable we used to solve this questions is is_short and avg_star. To solve this question it would be most appropriate to use a **Two Group Hypothesis Test (Randomization Test)**. The result from a Two Group Hypothesis Test is the amount of evidence we have against our null hypothesis.

Hypotheses

Null hypothesis (H_0): There is no difference in the mean average star rating for the top 3 results for searches in the short group and long group from the population

$$H_0 : \mu_{short} = \mu_{long}$$

Alternative hypothesis (H_A or H_a or H_1): There is a difference in the mean average star rating for the top 3 results for searches in the short group and long group from the population

$$H_1 : \mu_{short} \neq \mu_{long}$$

Where μ_{short} is the mean of the average star rating for searches in the short group and μ_{long} is the mean of the average star rating for searches in the long group

Methods

- We solved this question with a Two Group Hypothesis Test

- To do this we filtered out the data as described by our population and grouped the observations by whether the search was in the short group or the long group

- Next, we used the new edited tibble to create a simulation using R assuming that the null hypothesis is true. After doing so we examine how extreme the difference from μ_{short} and μ_{long} is for our sample and compare it against the simulated data.

- Doing this allows us to come to a conclusion and evaluate the evidence against the null hypothesis

Visualization

Figure 2

Average Star Rating For Searches
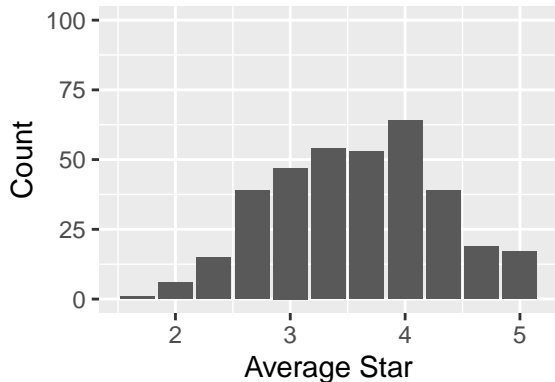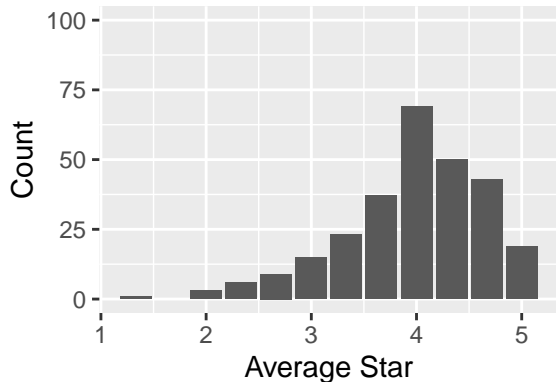in the Short Group

Figure 3

Average Star Rating For Searches
in the Long Group

Results

- Figure 2 and 3 make it evident that the searches in the long group were on average greater than the ones in the short group

- For our sample the difference between the mean of the average star rating for serches in the short group and long group was: about -0.39. This suggests that for our sample searches, the ones in the short group had a lower mean average star rating for the top 3 results than the searches in the long group.

- After performing the Two Group Hypothesis Test we came to the conclusion that the difference between μ_{short} and μ_{long} we saw from our sample was very extreme compared to what was observed in the simulation

- This led us to conclude that we have very strong evidence against the null hypothesis. This means depending on which group the search is in the average star of the top 3 listings would be different

Research Question 3

Question

Is there an association between the average star rating and the average number of reviews for the top 3 listings from the search?

This questions seemed interesting as we wanted to know whether the average star ratings for the top 3 listings resulted in a search correlates with the average number of reviews for the top 3 results. This all pertains to the searches containing only an adult count of 0. This analysis gives Expedia information on whether higher/lower average star rated searches have a higher/lower average review count.

The variables we used to solve this question are avg_star and avg_review. To solve this question it would be most appropriate to use a **Linear Regression**. We will construct 2 hypotheses to examine whether the relationship is linear.

Hypotheses

Null hypothesis (H_0): The slope equals 0 meaning there is no systematic linear effect of average star rating on the average review count for searches in our population

$$H_0 : \beta_1 = 0$$

Alternative hypothesis (H_A or H_a or H_1): The slope equals 0 meaning there is a systematic linear effect of average star rating on the average review count for searches in our population

$$H_1 : \beta_1 \neq 0$$

Where β_1 is the slope of the average star rating that helps predict the response variable: average review count

Methods

- We solved this question using Linear Regression

- To do this we filtered out the data as described by our population and plotted a scatter plot using the average star rating as the variable for the x axis and average review count as the variable for the y axis. Next we plotted the linear regression line to visually see the slope on the graph.

- After plotting the graph we examined the correlation and how much of the data was represented by the linear regression line

- After doing all this we created a linear model to examine the evidence against the null hypothesis and came to a conclusion

Results

Average Star Rating vs Average Review Count
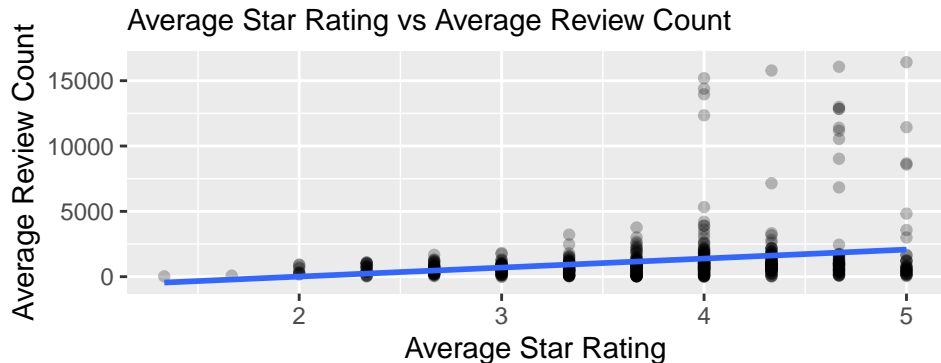
After investigating we came to the conclusion that there is strong evidence against the null hypothesis. This means that there is a linear association between average star rating and average review count for searches in our population. Additionally the association seems to be a weak positive linear association as the line does not explain most of the data.

Limitations

- Since we only had access to the search queries, the data set focuses on the search algorithm rather than consumer behaviour. This limits us because we cannot make any conclusions regarding certain groups of people rather we can only make inferences based on the search queries.

- It would have been helpful to have a larger data set so greater number of variables had values for them. As some variables had very little data within them making them impractical additions to our data set.

- The actual numerical price of the top 3 listings would be more insightful to answer question 2 rather than the average star rating

- The number of transactions would also be helpful to give insight for question 3 on whether searches, likely coming from couples, had more transactions on higher average star rated properties

Conclusions Part 1

All questions resulted in very strong evidence against their respective null hypotheses.

Takeaways

Question 1:

- For searches likely coming from couples the proportion of searches with all top 3 listings having a price reduction is not 50%
- To follow up this question it may be beneficial to check whether price reduced or non-price reduced searches generate more traffic
- Expedia could try and alter their algorithm to promote more listings from a certain group to appear more often to earn more money for searches in this population

Question 2:

- From this investigation Expedia can see that for searches likely coming from couples the average star rating for people in the short and long group are different
- They can further investigate whether which one gets a higher average star rated search

Conclusions Part 2

Takeaways

Question 2 (Continued)

- To maximize profits it would be beneficial to show higher star rated listings to people planning to searches in the long group as they may be more willing to spend the extra cash as they will be staying at the property longer.

Question 3:

- For searches likely coming from couples higher average star ratings seem to correlate with higher average review count
- The greater the number of reviews the listing has the more credible the listing appears to consumers
- A follow up question could be whether these higher star rated listings are more expensive
- To improve the platform we think Expedia should consider adding incentives for reviewing listings
- This way lower star rated listings could increase the number of reviews they have and in turn give those listings more credibility, thus giving the customer on a budget more options

Acknowledgements