A Splash of Expedia Search Patterns!

Kindly Fasten Your Seatbelts

Malaikah Hussain, Hannah Pan & Julie Tang - Group 5

March 31, 2022

Research Question 1: Introduction

TripAdvisor claims 45% of travelers with a smartphone use their mobiles to book their travel itinerary. Does **Expedia** have the same proportion of mobile users?

With the rise of mobile usage, it is important to understand the **percentage of Expedia users** that actually use their mobile platform. This can give an indication of a potential preference between platforms and reveals how Expedia's mobile platform compares to competitors such as TripAdvisor.

RQ1: Objective & Data Summary

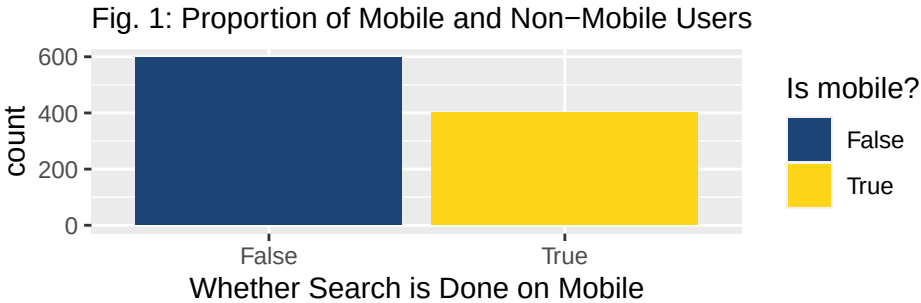
Objective

Using a sample of searches on Expedia, we would like to determine whether or not the true proportion of mobile searches is 45% of the overall searches. In particular, we are exploring the question: **Is the proportion of mobile searches performed on Expedia's mobile platform 45% of the total searches?**

Data Summary

Data Wrangling: Since we will only be using the variable `is_mobile`, to make the data frame simpler, all other variables have been removed.

RQ1: Visualization



Bar Graph: Displays the proportion of searches done on mobile in comparison to other platforms for our sample data. From this, we can estimate that mobile searches comprise around 40% of the total searches in our data. This is likely possible under the null hypothesis and means that the true proportion of overall mobile searches could be 45%, but must be tested further.

RQ1: Statistical Methods

Hypothesis Testing: A statistical method which involves evaluating evidence against a value for a **parameter** (number that describes the population). In our case we are assessing the **hypothesis** (called the null hypothesis) that the proportion of mobile searches **(parameter)** is equal to 45% **(value)**.

Hypothesis Testing Steps

1. Find the proportion of mobile searches for the Expedia sample data.
2. Simulate more samples under the null hypothesis and find the proportion of mobile searches for each of the samples.
3. Evaluate evidence against the null hypothesis.

RQ1: Results

- The sample data of overall Expedia searches gives us a proportion 40.2% mobile users.
- After conducting hypothesis testing, we find that the probability of more extreme values than the test statistic under the null hypothesis is 0.358.
- Since it is over 0.10, we have no evidence to reject the hypothesis that the proportion of mobile users is 45%.
- From this we can conclude that Expedia may have a similar proportion of mobile users to its competitors like TripAdvisor.

Research Question 2: Introduction and Objective

- **Expedia** users undoubtedly consider both **review score** and **number of reviews** when picking a listing.
- Important to grasp a pattern/relationship so that users can get recommended the most suitable listings.

Research Question 2

What is the association between the **number of reviews** and the mean customer **review score**?

- **Number of Reviews**: The quantity of reviews per listing, to the nearest 25.
- **Review Score**: The mean customer review score for the listed property on a scale from 0 to 5, rounded to nearest integer. 0 means that there are no reviews.

RQ2: Data Summary

Data Wrangling: Since we are interested in a dataset containing **unique** listing review scores and number of reviews, the given dataset was cleaned to only include observations of this precondition.

Out of the **3000** customer searches (1000 per 3 listings), **2618** listings were **unique**.

Data Wrangling Steps

1. We create a new dataset containing the original column **destination of listing**, and combine the **star rating**, **review score**, and **review count** of the 1st, 2nd, and 3rd listings into 3 columns. Instead of 1+9 columns of 1000 observations, there are now 1+3 columns of 3000 observations.
2. We remove the duplicate rows. The dataset now has 2618 observations.

RQ2: Statistical Methods

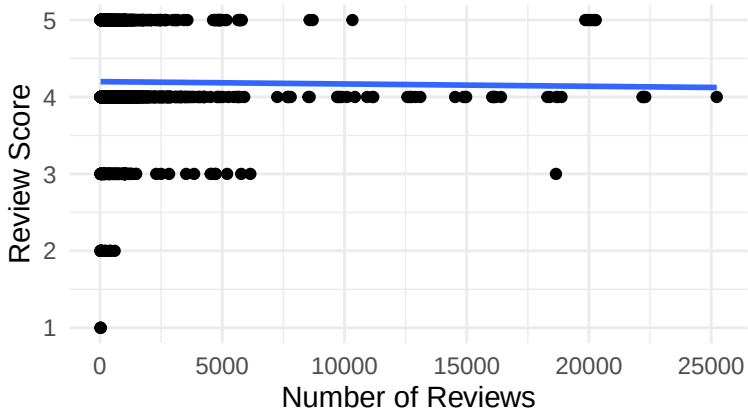
Simple Linear Regression: A model which uses values from a line of best fit to explain the relationship between the predictor and the response variable. If there is no association, the slope of the fitted line would be 0.

Simple Linear Regression Steps:

1. Create the scatter plot, where the x-axis (**predictor**) is the number of reviews, and the y-axis (**response variable**) is the review score.
2. Plot and calculate the best fit line, and calculate strength of the correlation between our variables (the **r value**).
3. Create the summary table, the equation of the fitted line, and evaluate the **null hypothesis**. The result is determined by the **p-value**, which measures the strength of evidence against the null hypothesis.

RQ2: Results

Fig. 2: Number of Reviews vs. Review Score



The equation of the fitted line is $y = 4.20 - 3.02 \cdot 10^{-6}x$, the strength of the linear association is -0.012, and the p-value is 0.55.

RQ2: Results

- Null hypothesis: the slope of the fitted line is equal to 0.
- P-value indicates that there is no evidence against our null hypothesis, aligns with our slope of approximately 0.
- Strength of correlation is close to 0, showing that there is little to no correlation between the predictor and response variable.
- This mirrors our conclusion drawn from the summary table, as a slope of 0 means that the predictor does not affect the response variable in any way.

Implications:

- A bit surprising, results show that those two variables may not have been affecting one another as much as anticipated.
- For the future, weights could be added in order to improve the algorithm: an 'overall rating' calculated from a sample equation (0.5 x number of reviews) + (0.5 x review score).

Research Question 3: Introduction and Objective

Pay-Per-Click

Expedia uses a PCP marketing model whereby **Expedia** receives monetization from the advertiser per customer's click on the listing ad. Determining the possible **average number of reviews** that travel ads receive **reveals** the **effectiveness** of these ads on the quantity of reviews **Expedia** listings receive.

Research Question 3

What is the **range of plausible values** for the average **number of reviews** for **Expedia** property listings that are **travel advertisements**?

- **Plausible Values**: An interval of the form (x, y) where the original mean number of review value may lie
- **Number of Reviews**: The quantity of reviews per travel ad, rounded to the nearest 25
- **Travel Ads**: Listings labeled as "Ad"

RQ3: Data Summary

Data Wrangling: Since rows in the dataset are **customer searches** and the RQ is interested in the **listings**, the data was cleaned to identify unique **listings** from the 1000 searches.

Out of the **2618** Expedia listings **902** were travel ads.

Data Wrangling Steps

1. The dataset is cleaned to obtain **listings** from the given 1000 customer **searches**.
2. Similar values across **star rating**, **review score**, **review count** and **destination address** imply duplicate listings and are eliminated from the dataset.
3. The dataset is filtered to only include listings that are ads (`travel_ad == 1`) to generate a dataset with 902 rows named `review_rating_902`.

RQ3: Statistical Methods

Bootstrapping: A statistical method which involves the random re-sampling of data values from the original sample to obtain a **range of plausible** values of the population statistic.

Bootrapping Steps

1. A *sample* of our sample data `travel_reviews_902` is taken with replacement
2. The mean of the bootstrap sample data values/`review counts` is calculated
3. Step 1. and 2. is repeated 5000 times to get a *distribution* of bootstrapped means
4. The 2.5th to 97.5th percentile of the mean values is calculated, revealing the interval

RQ3: Results

Fig. 3: Distribution of Review Counts for Travel Ads

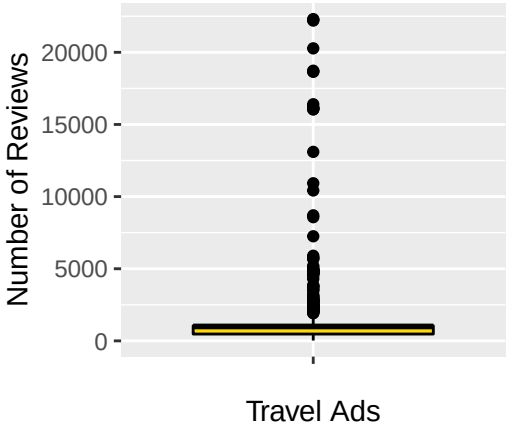
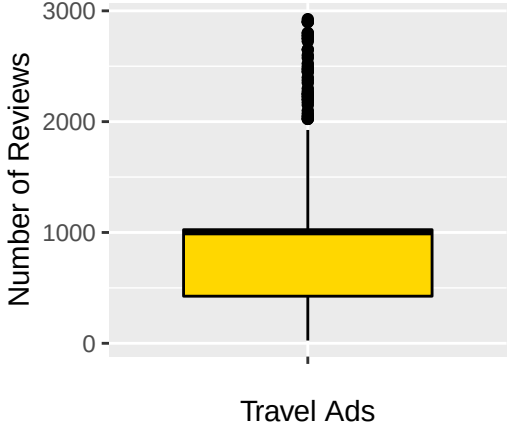


Fig. 4: Distribution of Review Counts for Travel Ads (count < 3000)



RQ3: Results

The range of plausible values for the average number of reviews for Expedia property listings that are travel advertisements is (1168, 1479)

Boxplot Description:

1. While 50% of the distribution lie near the 400 to 1000 range, the extreme outliers display the variability of the data. This means that in the process of bootstrapping, some samples may contain a higher proportion of outliers, being unrepresentative of the overall population.
2. All listings that have reviews greater than 2000 are outliers.

Confidence

Compared to the range of the values in the dataset, the 95% **Confidence Interval** is narrow (small), implying that if we got a different sample from the population, we could expect to get a similar value for the estimated mean.

Limitations of Approach / Model

1. **RQ 1**: Hypothesis testing only allows for evaluating evidence against the null hypothesis but cannot tell the true value of a proportion. Thus, for the conclusion, since there was not enough evidence to conclude the proportion of mobile users is not 45% it could very well be but it also could be a range of other values.
2. **RQ 2**: Non-continuous parameters were used for both the predictor and the response variable of the linear regression.
3. **RQ 2/3**: The method to remove possible duplicates is error-prone, and does not take into account that the review scores and the numbers of reviews of listings can change over time.

Conclusion

Summary of Findings:

1. Expedia may have a similar proportion of mobile users to its competitors like TripAdvisor.
2. There is little to no association between the number of reviews and the review score of a listing.
3. We are 95% confident that the mean review count for listings that are advertisements is between 1168 and 1479.

References and Acknowledgements

The authors would like to thank **Maya**, **Ziming**, **Nayan**, **Chris**, and **Nick M.** for their helpful suggestions and comments that improved the presentation of this poster.

Ar, R., Scoa. (2016, December 1). Install.packages fails in knitr document: "trying to use cran without setting a mirror". Stack Overflow. Retrieved March 31, 2022, from https://stackoverflow.com/questions/33969024/install-packages-fails-in-knitr-document-trying-to-use-cran-without-setting-a

TripAdvisor study reveals 42% of travelers worldwide use smartphones to plan or book their trips. Tripadvisor. (2015, June 30). Retrieved March 29, 2022, from https://ir.tripadvisor.com/news-releases/news-release-details/tripadvisor-study-reveals-42-travelers-worldwide-use-smartphones