Finding the role of the free cancellation policy using Expedia data

Hi can I cancel my reservation? The government locked us up again.

Group 144: Robin Mao, Zheyu Zhou, Cowell Tang

March 31, 2022

Overall Introduction

Ever since the pandemic, global tourism experienced uncountable lost. As the tourism industry is recovering nowadays, people restarted to travel again. However, due to the uncertainity brought by the COVID-19 variants and government policies, the **cancellation policy** of the properties became relatively more critical. Therefore, we found it meaningful and educational to investigate research questions related to the **cancellation policy** of the properties using data from Expedia. Now, let us begin.

Research Question #01:

What is the range of plausible values for the stay duration for properties with and without first listing free cancellation?

1.Introduction

-This research question is given due to my personal experiences, which states that first listing free cancellation is necessary due to the uncertainty brought by the COVID-19 pandemic.

2.Objective

-In this research question, our goal is to create two **bootstrap confidence intervals** in order for the next research question to proceed.

3.Data Summary

-In this case, the population refers to all Expedia search data done between 2021-06-01 and 2021-07-31. The three variables we used are *checkout_date*,*checkin_date* and *is_free_cancellation1*.

Statistical Methods

Bootstrap Confidence Interval

1.Reason

-Bootstrap confidence interval can provide the **variation** of the population based on the samples, outputting a range of plausible values

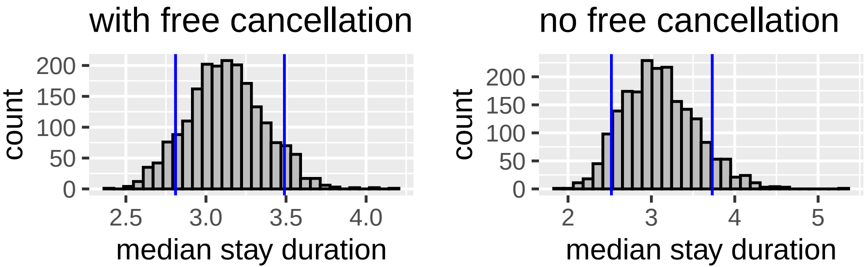
2.Advantages

-Requires only a small sample size

-More creditable, more intuitive

-Demonstrates the variation of the population

Visualisation & Conclusion



Each of these graph demonstrates a confidence interval on one sample statistics.

The exact x-values of the blue intervals are included in the next slide

Results & Conclusion: Research Question #01

Chart:exact x-value of the intervals

	Properties with first listing free cancellation	Properties without first listing free cancellation
First quantile	2.815	2.525
Third quantile	3.490	3.730

We can conclude that **we are 95% confident that the true means of the stay duration for properties with and without first listing free cancellation lie between the two intervals**

Limitation: In this case, the limitation refers to the fact that the model is true only using data during the pandemic. This means that if you use any pre-pandemic data onto the sampling distribution, the accuracy will be greatly reduced.

Research Question #02:

Will booking can be canceled without extra fees have a higher star rating?

1.Introduction

-When we travel, we like to use online booking to book hotels in advance. We always pay attention to the rating of accommodation when choosing.Therefore, I am interested in whether a free cancellation policy will increase the rating of a property.

2.Objective

-In this research question, our goal is to create three scatterplots in order to find out the **association** between the variables.

3.Data summary

-In this case, the population refers to all Expedia search data done between 2021-06-01 and 2021-07-31. The two variables we use are *is_free_cancellation1* and *star_rating*.

Statistical Methods

Simple Linear Regression

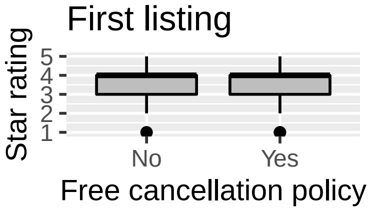
1.Reason

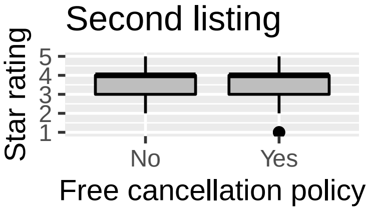
-A simple linear regression can demonsrate the association between the variables. With the correct model, we can also make prediction about future variables.

2.Advantages

-Easy to implement
-Interpret and effcient to train
-We can make predictions based on the model

Visualisation





In this case, a linear regression is performed between the *is_free_cancellation* and *star_rating*. The first graph demonstrates almost no difference between properties with different first listing free cancellation policy.
The second graph demonstrates that there is an outliner, representing a extreme value for samples with second listing free cancellation policy.
However, the differences are demonstrated in the chart in the next slide, showing the difference between the exact values.
We decided to show these graphs to provide a more intuitive visualisation of the data.

Results

Table: Coefficient-Level Estimates for a Model Fitted to Estimate Variation in Star Rating

The graph of third listing was ignored due to the limit on space, however the data is kept here.

	Predictor	B	SE	t	p
First Listed Booking	Intercept	3.63	0.045	80.68	<0.001
	Free Cancellation	0.17	0.057	2.93	0.004
Second Listed Booking	Intercept	3.68	0.045	81.14	<0.001
	Free Cancellation	0.07	0.057	1.33	0.185
Third Listed Booking	Intercept	3.57	0.044	80.57	<0.001
	Free Cancellation	0.19	0.056	3.43	<0.001

For general audiances, the p-value is what matters the most in this case.

Conclusion:Research Question #02

1.The chart shows an average difference of 0.17 in the response variable *star_rating1* between different free cancellation policies. The p-value is 0.004, which represents whether the observed relationship also exist in the greater population.

-Therefore, the chart demonstrates a significant linear relation between the free cancellation policy and star rating in the first listed booking.

2.The chart shows an average difference of 0.07 in the response variable *star_rating1* between different free cancellation policies. The p-value is 0.185, which represents whether the observed relationship also exist in the greater population.

-Therefore, the chart demonstrates no significant linear relation between the free cancellation policy and star rating in the second listed booking.

3.Using the same method above, we can say that the chart demonstrates a significant linear relation between the free cancellation policy and star rating in the third listed booking.

Limitation:

Limitation: The limitation is the assumption of linearity between the dependent and independent variables. However, in the real world, data are rarely linear separable. Thus, the existance of non-linear associations might decrease the accuracy.

Research Question #03:

1.Question

-Is the proportion of first listed booking with free cancellation 60%?

2.Motivation

-As the previous research questions address on the relation between free cancellation policies and other factors, we found it meaningful to find whether the proportion of first listed booking with free cancellation equals 60%. We choose 60% as on our first glimpse of the data set, there were about 60% of the properties that have free cancellation.

Statistical Methods

One Proportional Hypothesis Test

1.Variable Used

-The number of the first listed booking can be canceled freely and the number of the first listed booking cannot be canceled freely.

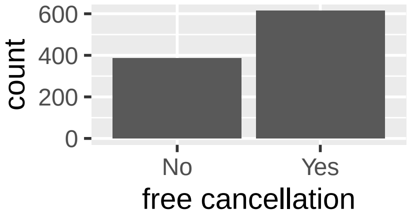
2.Data Wrangling

-Created a new variable that return “Yes” when first listed booking can be canceled freely and “No” when the first listed booking cannot be canceled freely.

3.Statistical Method

-The statistical method chose is one proportional hypothesis test. This inference creates simulations under the assumption that the *null hypothesis* is correct, and checks for extreme cases relaitve to the test statistics. This can be used to verify the strength of evidence against the null hypothesis. Because if the test statistic is very rare across all simulations, the null hypothesis has a lower chance being true. Vice versa, the null hypothesis is more likely to be true.

Visualisation



-The bar plot clearly indicates that the distribution of whether the first listed booking can be canceled freely. The number of the first listed booking having free cancellation, which is around 600, is 1.5 times to the first listed booking having not, which is around 400.

Hypotheses

-The **null hypothesis** for this question is that the proportion of the first listed booking that has free cancellation is 60%. The **alternative hypothesis** for this question is that the proportion of the first listed booking that has free cancellation is not 60%.

$$H_0 : p_{freecancellation} = 0.6$$

$$H_A : p_{freecancellation} \neq 0.6$$

-The **test statistic** which is the proportion of the first listed booking that has free cancellation is 0.614.

Result & Conclusion:Research Question #03

-We run 10000 times simulations by assuming the assumption is true and find the pvalue
This step is to find how rare the test statistic is

-Based on the p-value is equal to 0.1931, we can conclude that we have no evidence to against the null hypothesis that the proportion of booking of first listed that has free cancellation 60%

test statistic	pvalue
0.614	0.1931

Limitation: The concluding result from the hypothesis test are based on the p-value and therefore cannot be expressed with full certainty. We also cannot deny the probability of a type II error, which is fail to reject the *null hypothesis* when the *null hypothesis* is wrong.

Final Conclusion

-As this is a research question, it would be necessary for us to conclude our result relative to our initial question.

-Using the Expedia data, we can conclude that the first listing cancellation policy of the properties does affect the stay duration and star rating. And we can state that the proportion of booking with first listed free cancellation is **60%**. In advance, we can say that people do tend to care about the cancellation policy while booking properties during during the pandemic. We guess COVID really changed a lot of things, right?