Exploring the Impact of the Plus Program in Airbnb Performance

Ana Ferreira, Claudius Kroflin, Edet Bassey, Federico Scandizzo

Introduction

In 2018, Airbnb introduced the Plus program, an initiative designed to improve match-making for users seeking premium accommodations. The core objective of this study is to evaluate the impact of the program and determine whether differentiating high-quality services contributes to improved platform performance.

The research will be structured by first providing a dataset overview, outlining the relevant performance metrics, explanatory and other secondary variables. Next, a data quality assessment is given by explaining how missing values, inconsistencies and outliers were dealt with. The exploratory data analysis comes next, starting with the evolution of performance metrics overtime, highlighting the two implementation dates of the Airbnb Plus program to emphasize before and after changes.

The primary research question guiding this analysis is: "What is the impact of Airbnb Plus? Specifically, does distinguishing high-quality services lead to enhanced platform performance?"

1. Dataset Overview

The Dataset has a total of 14604 observations across 99 variables. It covers 773 zipcodes within 11 US cities from the beginning of August 2017 (`timperiod` = 8) to the end of October 2019 (`timperiod` = 34).

The Airbnb Plus Program was only implemented in 5 cities, such as *San Francisco*, *Denver*, *Nashville*, *New Orleans* and *Washington DC*. San Francisco got the program implemented in February 2018 (`timeperiod = 14`) while the others had it starting in October 2018 (`timeperiod = 22`).

For the scope of the research, the relevant key variables are the following:

- **Performance Metrics:** `total_bookings`, `booking_rate`, `listing_avg_review`, `average_booked_nights`
- **Secondary variables:** `zipcode`, `timeperiod`, `city_number`, `policy_entry`

Other than the above, two new variables were created and added to the data set:

- `city_name`: transforms the `city_number` to the name of the city according to the `zipcodes` included in it for relevant cities and to *Others* for those who did not had the plus program introduced at any point in time
- `date`: transforms the `timeperiod` into the correspondent *year-month* format

2. Data Quality Assessment

To assess the quality of the dataset, it is important to look at the distribution of the key identified variables.
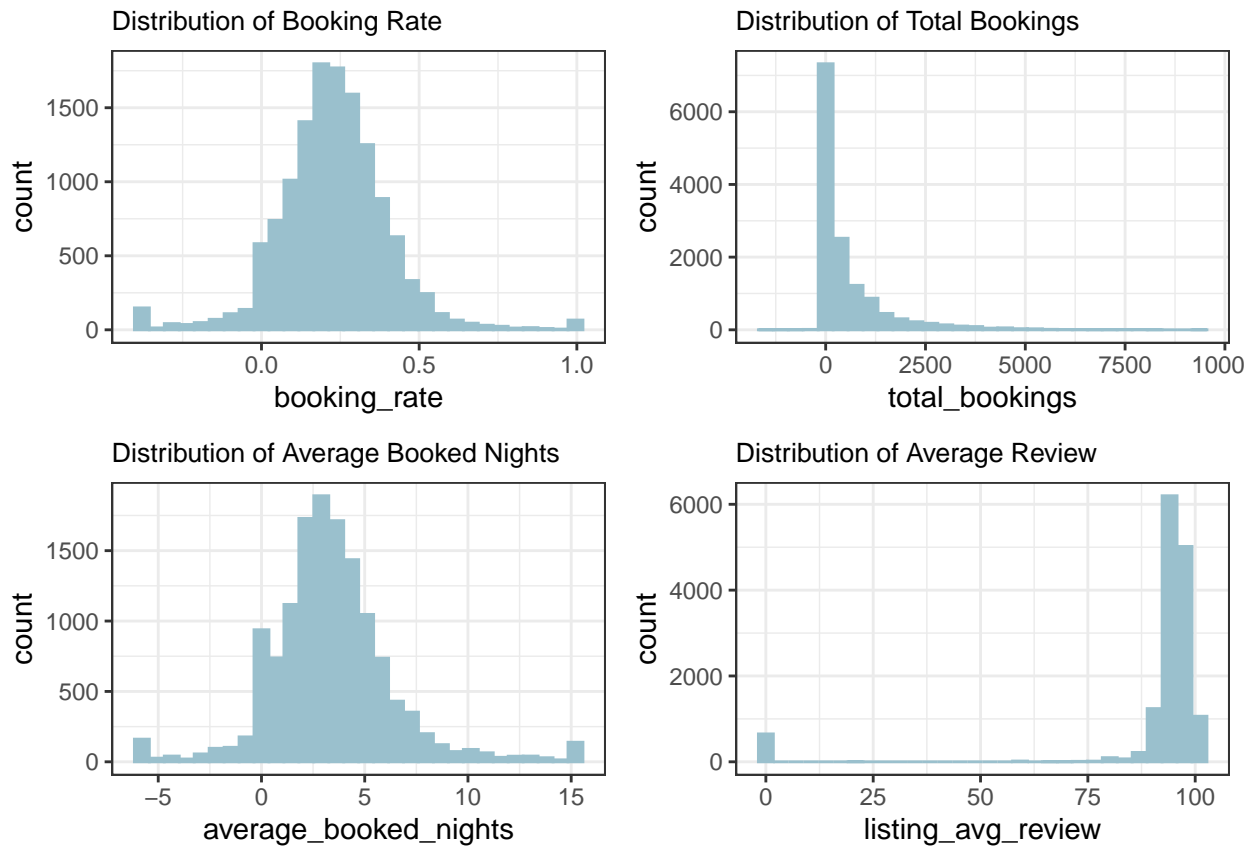
The variables `booking_rate` and `average_booked_nights` have both negative to positive values. When cancellations outweigh the bookings, the observation is a negative value, and vice-versa. In the `listing_avg_review` there are some "0" values, which affect the mean and distribution of this variable. Given that Airbnb review ranges on a scale from 1 to 5 stars, the 0 values will be treated the same as missing values (described later in the section). Other than the above, it seems like there is no other significant presence of outliers. To further investigate the variables, a box plot distribution is shown for each one.
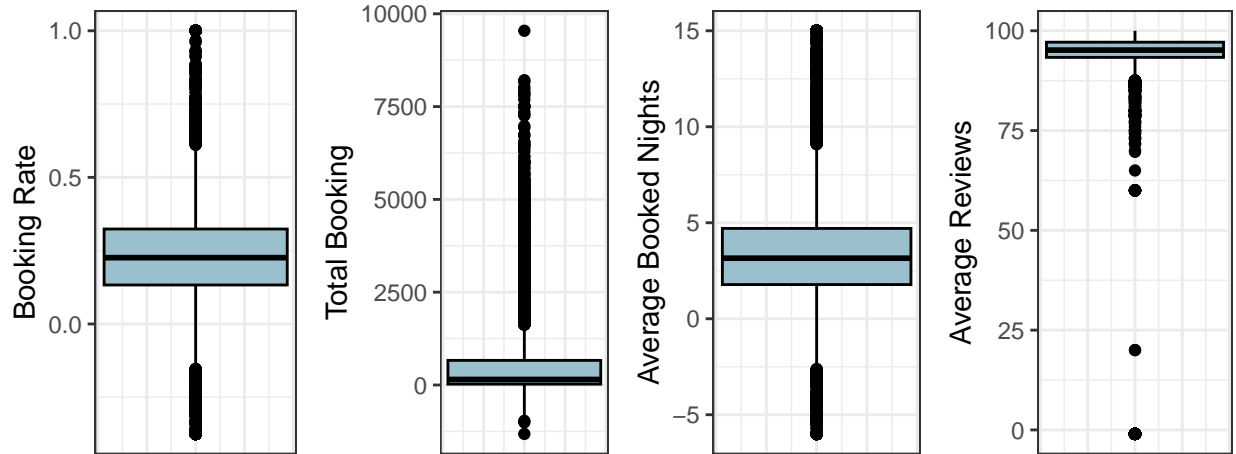
Most of the box plots show that the minimum and maximum values fit within the distribution. The `listing_avg_review` is the only with more of a significant outlier, apart from the 0's that were tackled previously. This outlier was left unchanged, since it is most likely just a very bad review.

All the missing values were counted and identified between the variables. Removing these values would could create gaps in time for some zipcodes and, thus, for some cities. For this reason, it was decided to input the mean value of each variable in the correspondent combination of `city_name` and `timeperiod`. After creating a table with those means, there were some missing values which indicated that there is no information for some of these combinations. As a solution, each missing value was filled with the one of the next time period (i.e. `booking_rate` for Denver, `timeperiod` = 18 was missing so it was assumed it was equal to Denver, `timeperiod` = 19 that was complete). Next, missing values in the dataset were replaced with the respective value of the means table. This strategy allowed the analysis to introduce as less bias as possible by not removing potentially relevant data and without significant changes in the overall mean of cities and time period – these are two crucial variables for this analysis.

	n_na	Mean Before	Mean After	% Change
total_bookings	953	542.27	552.25	1.84
booking_rate	1374	0.23	0.23	1.04
listing_avg_review	650	95.12	95.11	0.00
average_booked_nights	953	3.35	3.38	0.82

After doing the adjustments above, it was checked whether there were any duplicate rows – there were not. Next, `timeperiod` and `policy_entry` were investigated further.

To make sure that there are no gaps in the timeline for each one of the relevant cities, a table was made that shows 0 if no information is present in the data for that city and time and 1 otherwise (see table below). It can be noticed that data for Denver, Washington DC and Nashville only start from April 2018 instead of August 2017. So, the conclusions for those cities will be based on that time period. Also, Denver had missing data in May 2018 which created a gap in time. To manage this an artificial row was created for that time period, which is simply the mean between April and June 2018 for that city.

City	Count Present Timeperiods
San Francisco	27
Denver	18
New Orleans	27
Washington DC	19
Nashville	19
Others	27

`policy_entry` = 0 means no Plus program and `policy_entry` = 1 means Plus program was present. Given this, the sum of all policy entries for cities before the implementation of the Plus program should be 0. The sum of all policy entries for cities after the implementation of the Plus program should be equal to the number of rows. This simple analysis was confirmed to be correct.

Given the above, the final dataset is now ready and is the only one used throughout this research. The final dataset has 14605 observations across 10 variables.

3. Exploratory Data Analysis

This section will present a comprehensive exploration of the performance metrics over time. A focus will be placed on the two implementation dates of the Airbnb Plus program (February 2018 in San Francisco and October 2018 in other cities), enabling a visual comparison of platform performance before and after the program's introduction.
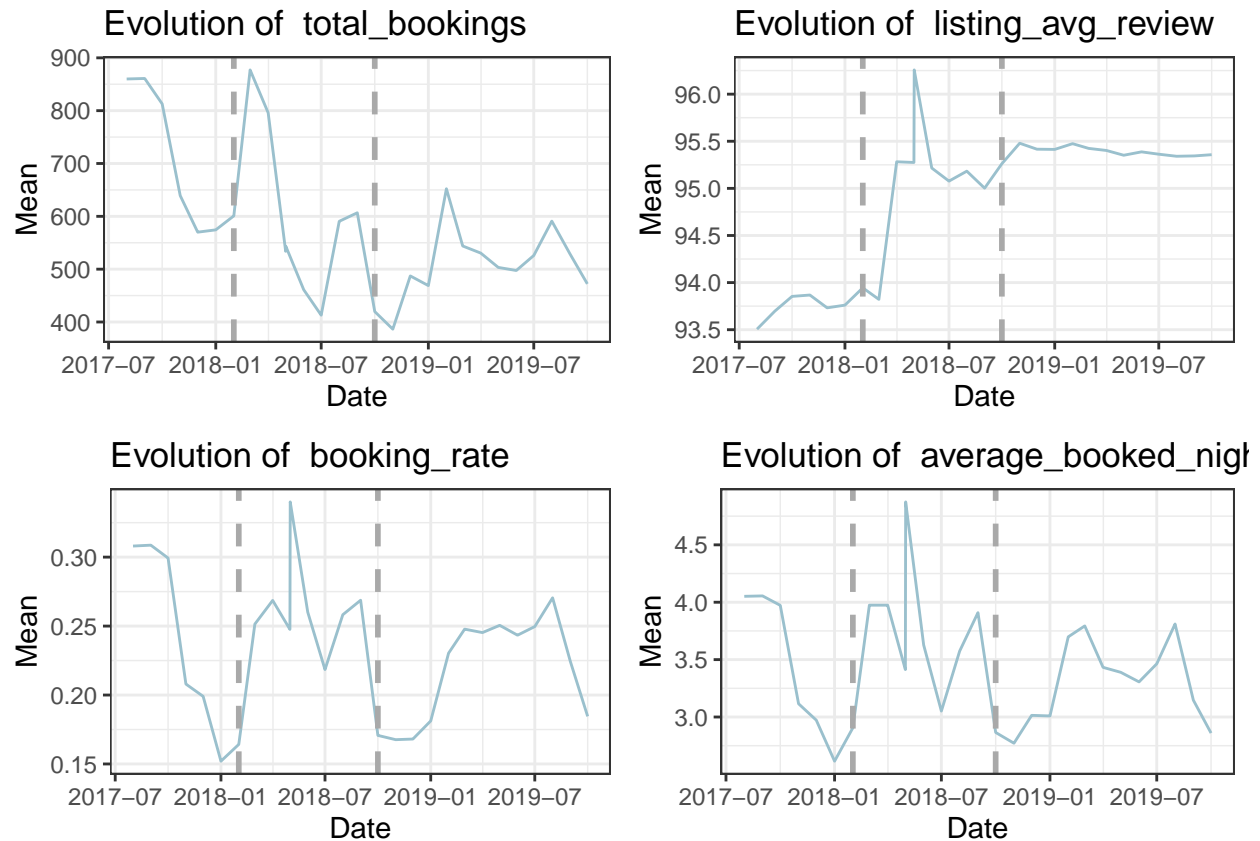

Booking rate and average booked nights have a very similar performance since they measure the same metric. The most notable insight is how the implementation of the Plus program in San Francisco alone was able to significantly increase these metrics. In the second implementation, said metrics did see some performance increase but eventually dropped, suggesting the Plus program was more of a success in San Francisco than in the other cities. Listing average reviews had quite a volatile performance, suggesting mixed results in the Plus program meeting customer expectations. Later in other cities, reviews seemed to have consistently increased showing an improvement in customer satisfaction. Total bookings severely decreased after the Plus program in San Francisco,

indicating mixed results when compared to the positive performance of booking rates. Perhaps, users booked more frequently but also cancelled more often which can explain how total bookings dropped. Later in other cities, total bookings also saw a consistent increase, like booking rates and average reviews.

Analysis of Booking Rate increase by city

How much of the booking rate did the Plus program contribute to? Was it more effective in San Francisco or in the later cities? In order to answer these questions, we will compare this metric to those cities that did not receive the Plus program at all.
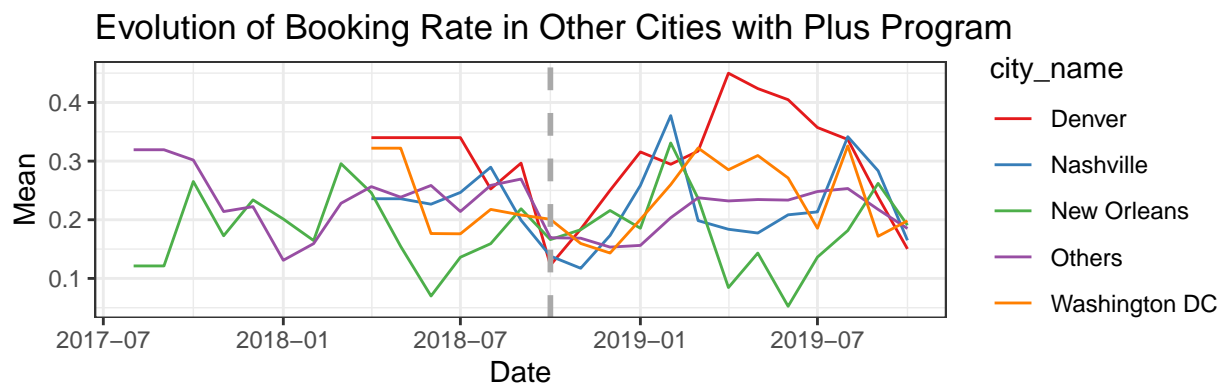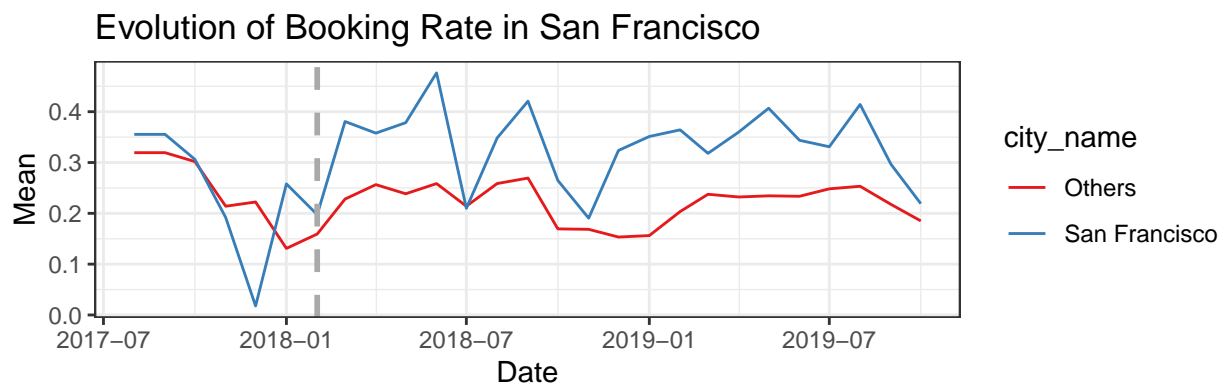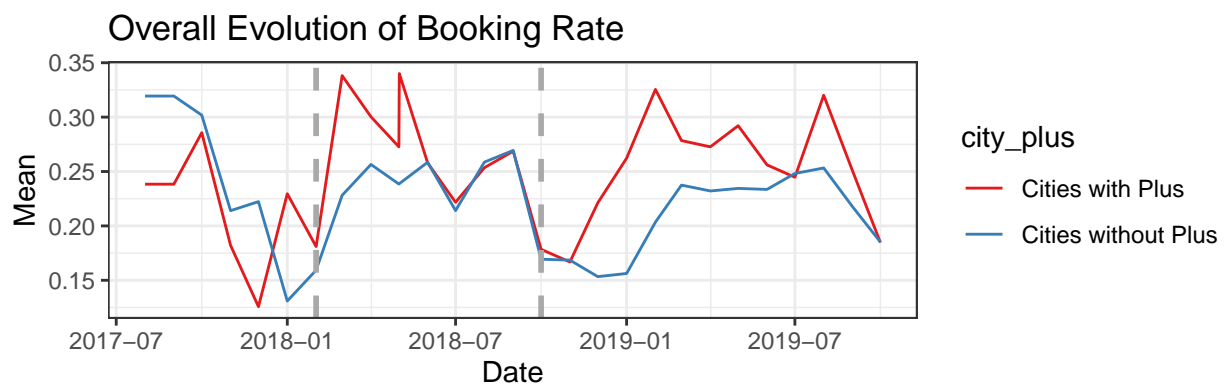
The Plus program in San Francisco definitely increased booking rate against cities that did not receive it. The program later implemented in other cities (Denver, Nashville, New Orleans and Washington DC) also had increased booking rates compared to all other cities. However, as it was suggested in previous plots, the program in San Francisco seems to have been more successful. It is possible that the Plus program in San Francisco was more of a success because it was an unannounced pilot program. The program in other cities could have been less of a success for a large variety of reasons.

Analysis of Average Rating in second stage cities

If booking rates in the cities that got the Plus program (excluding San Francisco) didn't see much of an increase, did they at least have better average ratings, resulting in improved customer satisfaction?
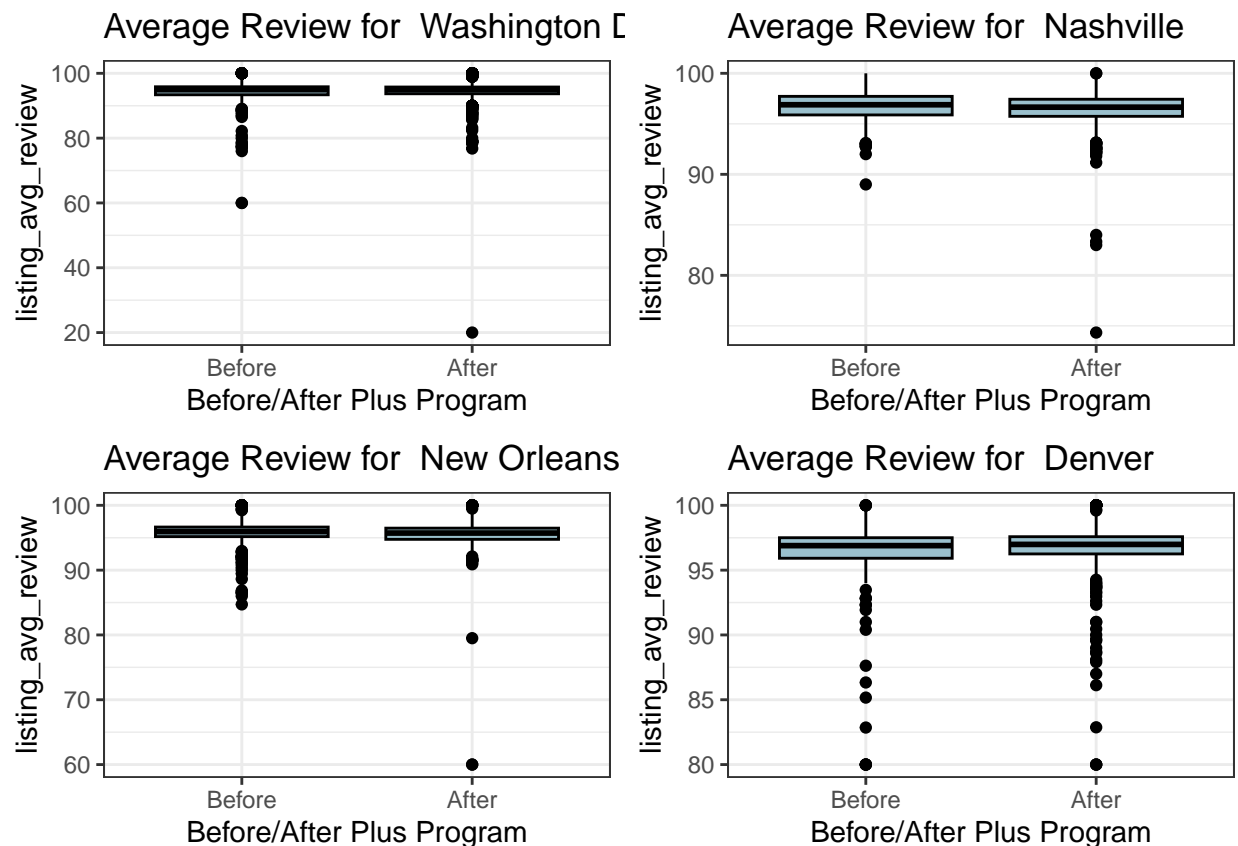
Interestingly, the average rating in Washington DC, Nashville, New Orleans and Denver didn't change that much. The average rating was already very high. Perhaps, the Plus program in these cities did meet customers' expectations which allowed the rating to stay consistently high.

To recap, the assumptions we can extrapolate from the previous visual inspections are the following: **overall booking rates increased in cities after receiving the Plus program**; however, this same metric **does not seem to have significantly improved in Denver, Nashville, New Orleans, and Washington DC**; finally, in these cities the **average review also seems to remain unchanged**.

4. Hypothesis Testing

To test the assumptions above, a short series of hypothesis testing will be conducted. However, because there have been two implementations of the Plus program, comparing the booking rate means of the Plus cities before and after at different time periods introduces bias, since there is mixed data between the two implementations. For this reason, the means of all the booking rates will be **grouped by policy entry** (whether the Plus program was present or not), instead of by time period, to still determine before and after effects.

The first hypothesis is, therefore, the following:

- H_0 : The booking rate mean before Plus implementation is equal or larger than that after implementation.

- H_a : The booking rate mean before Plus implementation is smaller than that after implementation.

In order to test this hypothesis a left-tailed t-test will be conducted.

policy_entry	Mean	SD
0	0.24	0.16
1	0.26	0.18

statistic	t_df	p_value	alternative	estimate	lower_ci	upper_ci
-4.004542	2398.474	3.2e-05	less	-0.0247803	-Inf	-0.0145979

The p-value of the test is extremely small, close to 0. Thus, the observed difference between the two group means is **statistically significant** at a 95% confidence level, leading to reject the null hypothesis of equality of means (p<0.05).

The second hypothesis will test the **strength of the relationship** between booking rates and the implementation of the Plus program in Denver, Nashville, New Orleans and Washington DC.

- H_0 : There is no significant relationship between the booking rate and the implementation of the Plus program. *($\beta1 = 0$)*

- H_a : There is a significant relationship between the booking rate and the implementation of the Plus program. *($\beta1 \neq 0$)*

In order to test this hypothesis a linear regression model will be applied.

plus	Mean	SD
After	0.24	0.19
Before	0.24	0.15

Table 6: Relation Between Booking Rate and Plus Program

	Dependent variable:
	booking_rate
plusBefore	−0.002
	(0.007)
Constant	0.240***
	(0.004)
Observations	2,538
R^2	0.00004
Residual Std. Error	0.177 (df = 2536)
Note:	*p<0.1; **p<0.05; ***p<0.01

Even though $\beta1 \neq 0$ (-0.0023), the p-value for this coefficient is greater than 0.05. At a 95% confidence level, the coefficient is **not statistically significant**, indicating not enough statistical evidence that there is a significant relationship between the two variables.

Finally, the third hypothesis states the following:

- H_0 : The average review mean for the second stage cities is equal before and after the plus implementation.

- H_1 : The average review mean for the second stage cities is different before and after the plus implementation.

In order to test this hypothesis a two-tailed t-test will be conducted.

policy_entry	Mean	SD
0	95.55	3.51
1	95.66	3.59

statistic	t_df	p_value	alternative	estimate	lower_ci	upper_ci
-0.7300316	1878.441	0.4654619	two.sided	-0.1072067	-0.3952171	0.1808038

The p-value of the test is greater than 0.05. So at a 95% confidence level, there is **no evidence of significant difference** between the means, leading to fail to reject the null hypothesis of equality of means.

Conclusion

In conclusion, it can be deduced that overall **booking rates significantly improved** after implementing the Plus program. It was found that the difference between the booking rate means was statistically significant.

But, the change in **booking rate means was different** between the first implementation of the Plus program in San Francisco and the second in Nashville, New Orleans, Washington DC and Denver.

The first program in San Francisco seems much more successful, perhaps because it was essentially an unannounced pilot program. Therefore, it was worth investigating the effect of the program

11

specifically in Nashville, New Orleans, Washington DC and Denver. Interestingly, **no evidence of a significant relationship was found between the implementation of Plus program and increase in booking rates**.

Since the Plus program seemed to be less of a success in these cities when compared to San Francisco, it was worth exploring whether customer satisfaction changed instead. After testing for it, it was shown that there is **not enough evidence to conclude that there is a difference between means of listing average reviews**.

In other words, the Plus program increased overall booking rates. The Plus program in San Francisco was a success, although it was a pilot one. In the cities where the program was later implemented, there was no relationship between the booking rates and the implementation of the program. Customer satisfaction might have played a role, except that no significant difference was found. This analysis was insightful to suggest several demographic topics for further research. Why was the pilot program such a success? Did employment rate play a role? Perhaps, San Francisco is a larger CBD than other cities, leading to better match-making for premium accommodations. Did seasonality matter during the second implementation? It could be that the program was less of a success since it was introduced after Summer, a time where people generally come back from holidays and book less vacations.

Disclaimer

This research paper had its code deliberately debugged with the help of AI, making sure that the inevitable human errors were swiftly corrected by our digital assistant. Rest assured, no robots were harmed in the process, and the human touch remained essential. Other than debugging, AI was also used for coding suggestions and alternatives.

Fun fact: this disclaimer was written by AI