Further Investigation on Impact of Plus Program in Airbnb Performance

Ana Ferreira, Anna-Maria Fiederling, Jasper Sänger, Paulo Cheng

Introduction

Previous exploratory data analysis and hypothesis testing, to evaluate the impact of the program and determine whether differentiating high-quality services contributes to improved platform performance, indicated that the implementation of Airbnb Plus Program yielded a general positive impact on the booking rates in San Francisco, the pilot market subjected to Airbnb Plus Program roll-out. This result was not replicated for cities in the second implementation period, where hypothesis tests returned no significant improvement in market outcome variables.

In this study, to further scrutinize the effect of the Airbnb Plus Program and build on previously established hypotheses, regression modelling is applied using the Difference in Difference framework, identifying differences before and after treatment for treated and control groups.

Hypotheses

In the previous investigation, a short series of hypothesis testing was conducted in order to test the explored assumptions related to the question of interest.

The alternative for the first hypothesis stated the following:

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

The results for this test showed a p-value 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 tested the strength of the relationship between booking rates and the implementation of the Plus program in Denver, Nashville, New Orleans and Washington DC, the cities of the second implementation of the program.

H_a : **There is a significant relationship between the booking rate and the implementation of the Plus program.** *(β1 ≠ 0)*

Even thought β1 ≠ 0 (-0.0023), the p-value for this coefficient was 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 alternative for the third hypothesis stated the following:

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

The p-value of the test was 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.

In summary, the Plus program seem to have significantly improved overall booking rates. However, no significant relationship was found between the program and increased booking rates in Nashville, New Orleans, Washington DC, and Denver *(second treatment group)*, leading to the conclusion that the previous result was influenced by a positive result in the first implementation

in San Francisco. Additionally, there was no evidence to suggest a difference in customer satisfaction, as measured by average listing reviews, in the second group cities. These findings establish the baseline scope and motivation of the preceding report.

Model Specification

A fixed effects model was chosen in order to account for differing time periods of policy implementation, unique zip code characteristics and the unbalance in the data across these two indexes *(time period & zip code)*.

Given that, the model for `booking_rate` is specified as follows:

$$\text{Booking Rate}_{it} = \beta_0 + \beta_1 \text{Policy}_{it} + \beta_2 \text{Plus}_{it} * \text{Policy}_{it} + \beta_3 \text{Big City}_{it} * \text{Policy}_{it} + \beta_{4,...,k} \text{control}_{it} + \gamma_i + \delta_t + \epsilon_{it}$$

The model for `listing_avg_review` is specified as follows:

$$\text{Average Review}_{it} = \beta_0 + \beta_1 \text{Policy}_{it} + \beta_2 \text{Plus}_{it} * \text{Policy}_{it} + \beta_3 \text{Big City}_{it} * \text{Policy}_{it} + \beta_{4,...,k} \text{control}_{it} + \gamma_i + \delta_t + \epsilon_{it}$$

Equations valid for k-4 control variables, that in this case corresponds to the same three for both models: `search_trend_airbnb`, `entire_ratio` and `super_host_ratio`. γ_i represents the `timeperiod` dummies and δ_t the id dummies, in this case `zipcode`. Two variables were created: `big_city`, a binary variable *(1 for cities with populations over 1 million, 0 otherwise)*, and `plus`, a categorical variable with three levels *(1st impl, 2nd impl, No impl)* indicating the program implementation phase or lack thereof.

Dataset Overview & Model-Free Investigation

The final cleaned data set has a total of 14643 observations across 18 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 data cleaning process involved a deep dive into the data, looking at distributions, missing

values, presence of outliers and inconsistencies for the relevant variables. For the variable `listing_avg_review`, an unlikely large number of value "0" was present. These were assumed to be NA values and treated as such. For the same variable, one outlier was treated by imputing the next closest minimum value.

Inconsistencies in the indexes *(zip codes across time periods)* were identified and showcased an unbalanced panel data set. 507 zip codes have entries past April 2018 only and 98 zip codes left the data set before the last time period. Some zip codes have entry gaps in between first and last observed time periods. For one collection of zip codes which belong to the city Denver, an entire gap in entries for the time period equivalent to May 2018 was present. As this seem to be different from the other observed gaps, it was treated by creating an artificial row for each zip code with imputed values derived from the mean between the prior and preceding month across all variables.

For the scope of this research, the data analysis that will follow will be mainly focused in the outcome variables of interest: `booking_rate` and `listing_avg_review`, their evolution over time, trend and relations with other important features.

Table 1: Summary Statistics

Variable	N	Mean	St.Dev.	Min	25%	Median	75%	Max
booking_rate	14643	0.23	0.17	-0.38	0.14	0.23	0.32	1.00
listing_avg_review	14643	95.14	3.25	70.00	93.66	95.30	97.14	100.00
policy_entry	14643	0.15	0.36	0.00	0.00	0.00	0.00	1.00
search_trend_airbnb	14643	6.47	3.76	1.00	3.25	6.75	8.00	19.75
entire_ratio	14643	0.59	0.26	0.00	0.40	0.60	0.78	1.00
super_host_ratio	14643	0.18	0.18	0.00	0.07	0.14	0.24	1.00

In the summary statistics table above, `booking_rate` across all time periods and zip codes in the cleaned data set displays a mean of 0.23. The `listing_avg_review` key variable indicates good reviews, with a high mean score of 95 and values ranging only from 70 to 100. The variable `booking_rate` displays a negative minimum value. This could suggest data entry errors and will be addressed in the assumptions sections in more detail. The variable `search_trend_airbnb` ranges

from 3.76 to 19.75 with the mean value being 6.47. It is also possible to conclude that 15% of the zipcodes present had the Plus program introduced, 59% of listings offer entire apartments and 18% have a super host.
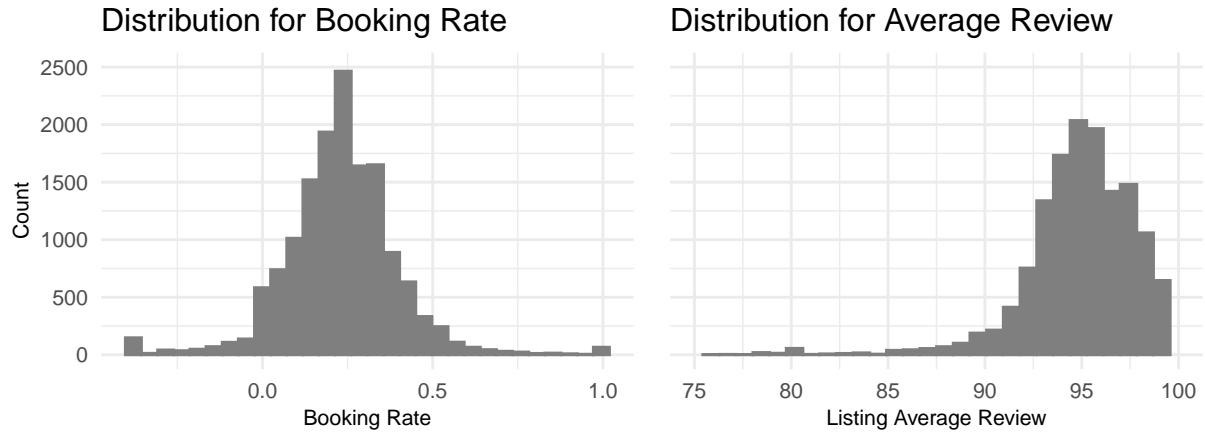

The `booking_rate` distribution is approximately normal, with few values near the extremes 0 or 1, indicating most rates are moderate. The `listing_avg_review` distribution is left-skewed, mainly concentrated above 90, with a near-normal shape within that range.
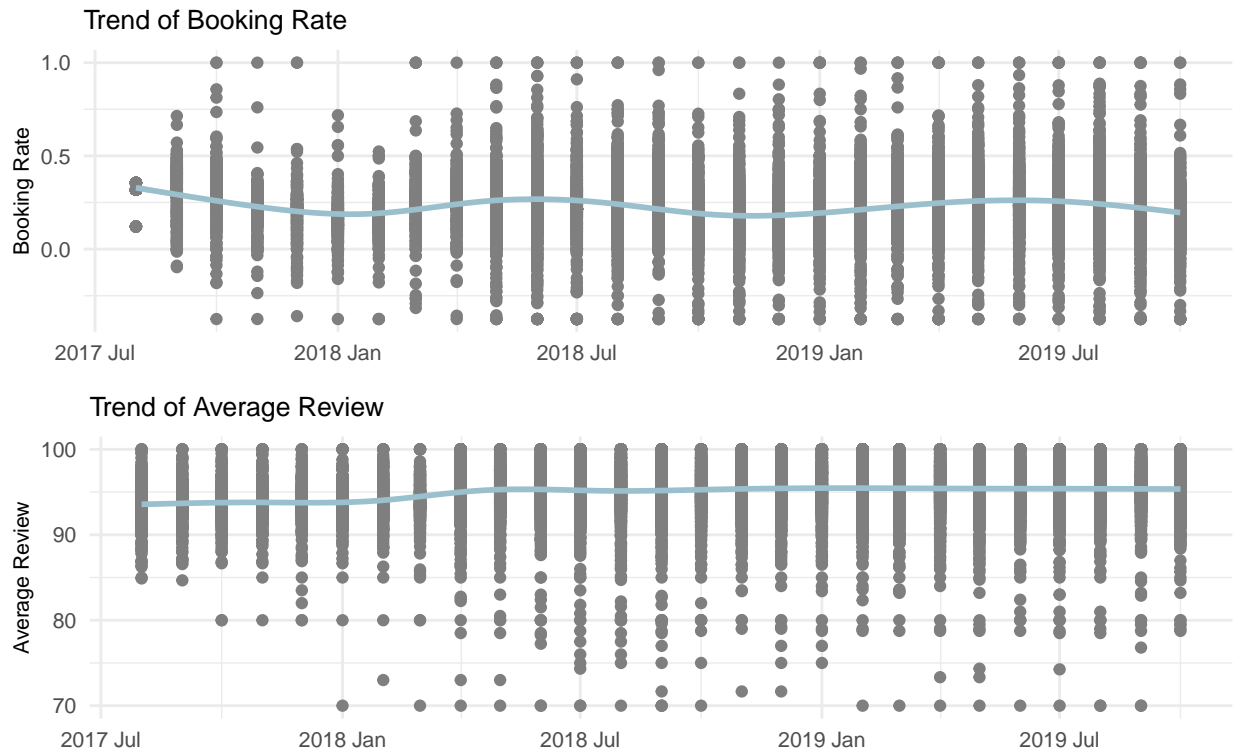
The trend line for booking rates displays minor fluctuations over time. `booking_rate` remain rel-

atively stable with slight dips in February 2018 and November 2018, periods that coincide approximately with program implementation. `listing_avg_review` are consistently high with a slight upward trend over time. There is some presence of outliers but those do not seem to influence the trend line drastically.

Evolution of Booking Rate

Evolution of Average Review
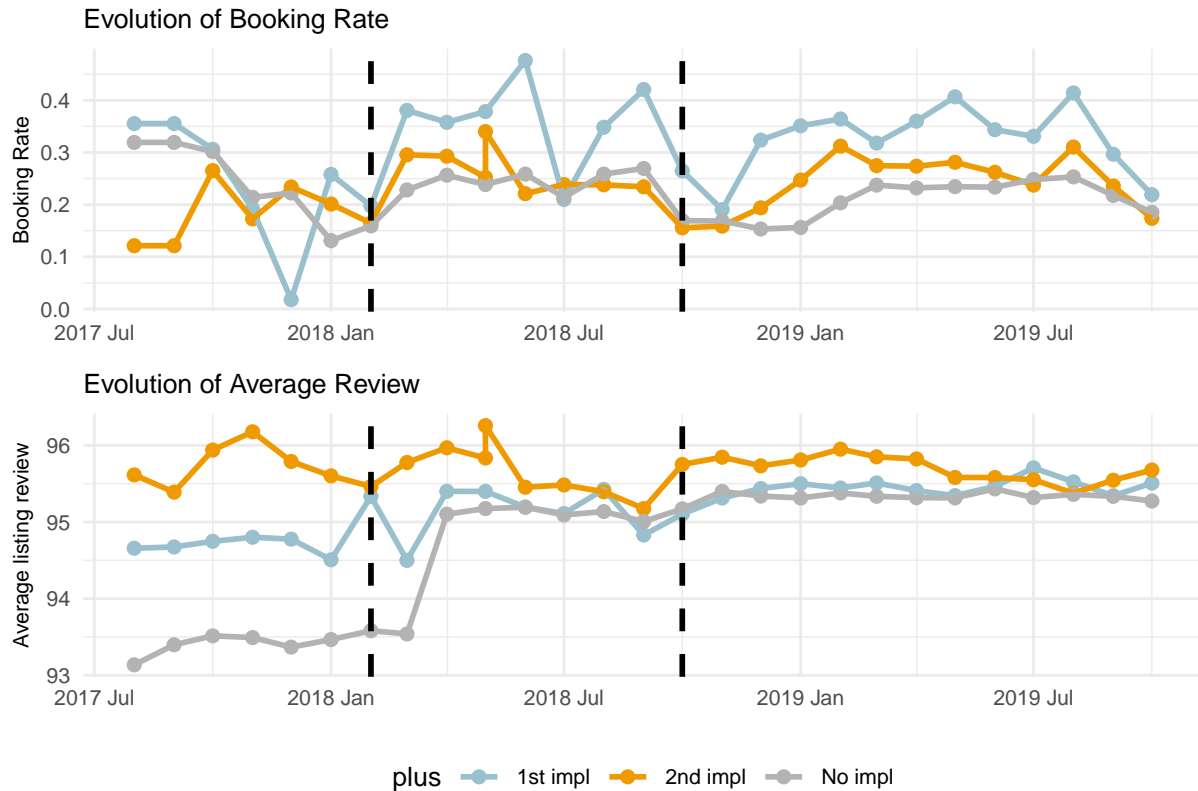
The top plot shows `booking_rate` over time, with the *1st impl* group consistently higher, followed by *2nd impl*. The bottom plot shows stable average listing reviews, with *2nd impl* scoring highest, followed by *1st impl*, and *No impl* consistently lowest. The significant dip to a vlue of 0% in `booking_rate` in December 2019 is likely a data collection error or an extraordinary unknown event.

Regression Analysis: Estimating the Impact of Airbnb Plus

After the previous analysis, the above specified models were conducted and after modelling a test for serial correlation was performed with the results indicating a p-value smaller than 0.05 leading to reject the null of no serial correlation, at a 95% confidence level. For that reason, the standard errors showed in the following models summary were corrected to be the robust one's.

Table 2: Fixed Effects Models

	Dependent variable:	
	Booking Rate with robust SEs	Average Review with robust SEs
	(1)	(2)
policy_entry1	0.051***	−0.721*
	(0.018)	(0.419)
search_trend_airbnb	0.015***	−0.028
	(0.002)	(0.027)
entire_ratio	0.076***	0.196
	(0.023)	(0.523)
super_host_ratio	0.015	1.723***
	(0.022)	(0.291)
policy_entry1:plus2nd impl	−0.060***	0.185
	(0.016)	(0.372)
policy_entry1:big_cityYes	0.017	0.779**
	(0.012)	(0.324)
Time Period	Yes	Yes
Observations	14,643	14,643
Note:		*p<0.1; **p<0.05; ***p<0.01

Note: For the interpretations of the models a 0.05% significance level will be considered.

According to the `booking_rate` model results, Airbnb Plus had a positive effect on the `booking_rate` during its first implementation. Holding all other factors constant, booking rate increased by an average of 0.051 units after the program started. However for the second implementation, holding all else constant, the booking rate decreased by an average of 0.009 units. Furthermore, the interaction between `policy_entry` and `big_city` *(added to examine differences in effects based on city characteristics)* yielded a statistically insignificant result. This indicates that there is insufficient evidence to suggest differences in the program's impact between

small and big cities.

The `listing_avg_review` model shows no statistical evidence of changes in customer expectations before and after the program's introduction. Notably however, post-implementation, average reviews in big cities are higher by 0.779 units on average, controlling for other factors.

The control variables in the booking rate model behaved as expected, showing a positive effect overall, although the `super_host_ratio` variable did not yield statistically significant results. In the average review model, only `super_host_ratio` showed a significant result, indicating the anticipated positive effect.

In summary, these results align with the hypotheses previously tested, with the models reinforcing the conclusions drawn: the program had a **significant impact in its first implementation** in San Francisco. However, there is **no evidence of a similar effect during the second implementation**. Additionally, there is **no evidence to suggest that the program significantly improved customer expectations**, as reflected in higher average reviews.

Assumptions and Model Diagnostics

**MLR.1**: *The population model is linear in the parameters.*
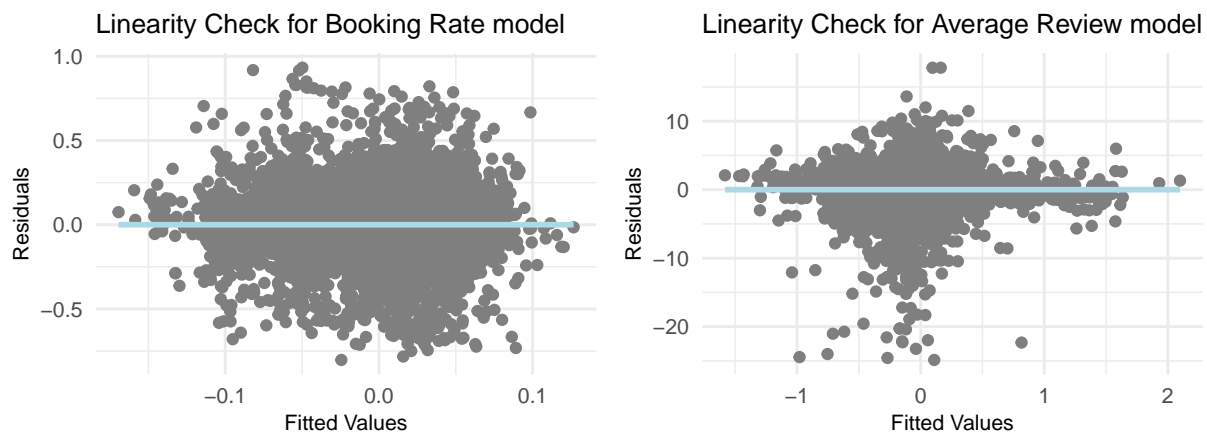
The diagnostic plots suggest that both models meet this first assumption of linearity as the Residuals vs Fitted plots depict straight lines.

MLR.2: *Random sample of n observations.*

The assumption of random sampling requires that the data used in the analysis is representative of the population of interest, with each observation independently and randomly drawn. We do lack explicit information about the data collection process, however, the data set is composed of aggregated Airbnb metrics across various U.S. zip codes and time periods, which suggests that the data is not systematically biased toward any particular subset of the population. In addition, no explicit evidence suggests that the data was collected in a way that would violate the independence or randomness of observations, making this assumption reasonable in our analysis.

MLR.3: *None of the independent variables is constant, and there in no exact linear relationships among the independent variables (no perfect collinearity).*
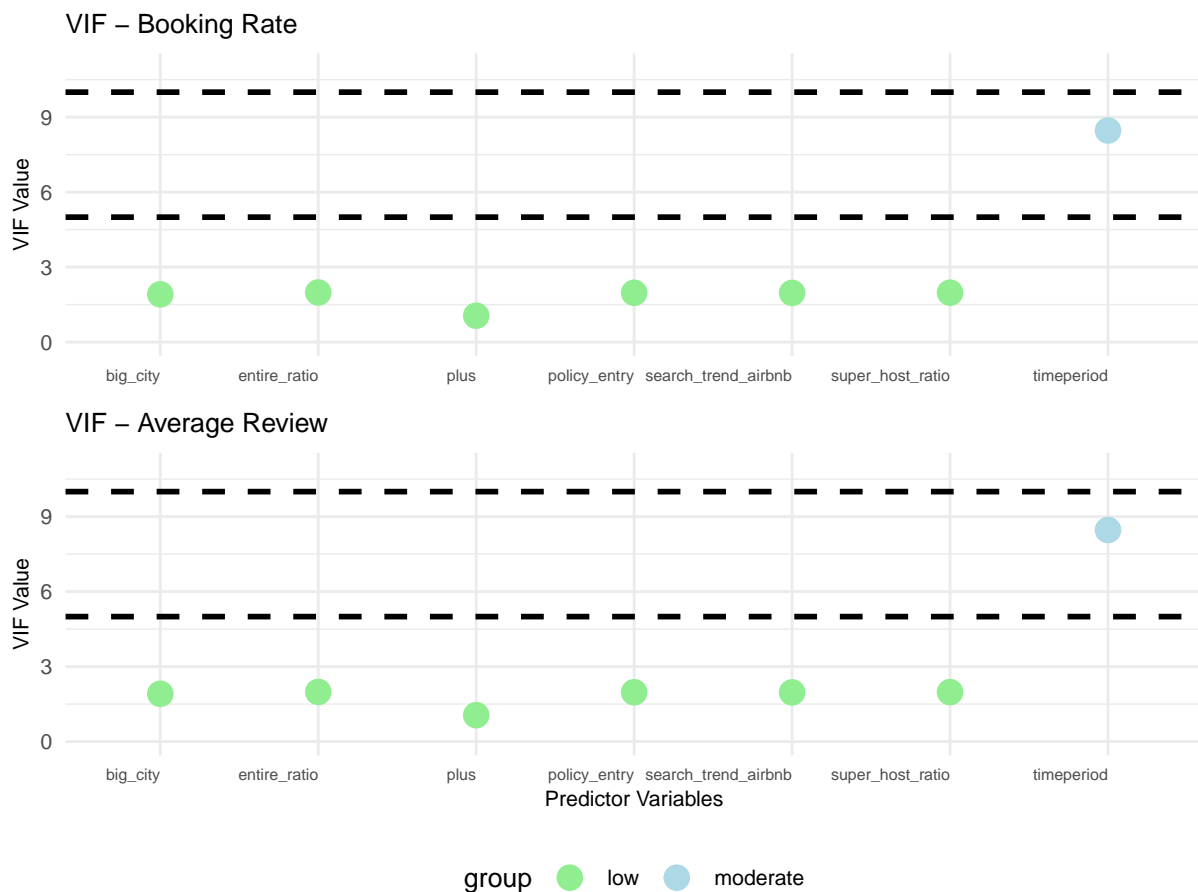
The VIF plot for both models show exactly the same values as all independent variables are common for both models. These plots show that most independent variables have VIF values below 5, indicating no significant multicollinearity issues. Only `time_period` displays a moderate but

acceptable VIF value, as it is still below 10. This satisfies the MLR.3 assumption of no perfect collinearity, ensuring the model's stability. The predictors are moderately correlated, and standard errors are unlikely to be inflated.

MLR.4: *The error ϵ has an expected value of zero given any values of the independent variables (zero conditional mean).*

The `booking_rate` model could be influenced by time-varying factors *(hidden in the error term)* such as natural catastrophes, cultural events, or market characteristics *(e.g., competitors, new state regulations)*, which may correlate with explanatory variables and introduce bias. Negative booking rates in the data could suggest potential measurement errors in the explained variable, and measurement error in the explanatory variables are assumed to not exist. Reverse causality is a concern, as selection for policy implementation could depend on pre-existing booking rates, creating potential endogeneity. This selection bias could distort the estimated effect of `policy_entry`, reflecting correlation rather than causation.

The `listing_avg_review` model may be affected by omitted time-varying factors like local service quality or platform-level policy changes, which could introduce bias. No evidence suggests measurement errors or reverse causality in this model.

MLR.5: *The error ϵ has the same variance given any values of the explanatory variables (homoskedasticity).*

To diagnose heteroskedasticity, a Breusch-Pagan test was conducted for each model. At a 95% confidence level, both tests yielded very small p-values (<0.05), leading us to reject the null hypothesis of homoskedasticity. This indicates that the variance of the errors is not constant. However, the interpretations of the results remain valid, as the use of robust standard errors accounts for both heteroskedasticity and serial correlation.

MLR.6: *The population error ϵ is independent of the explanatory variables and is normally distributed with zero mean and variance σ^2.*

Since the sample used for modeling is large, the reliability of inference is supported by the Central Limit Theorem, which ensures that this assumption holds.

Conclusions and Business Implications

The central finding of the analysis challenges conventional assumptions about plus service upgrades. Contrary to expectations, the Airbnb Plus program did not lead to a significant improvement in customer satisfaction, measured by average reviews. However, the plus program showed evidence of a positive impact on booking rates in San Francisco, with an increase of 0.051 units post-implementation. In contrast, for the second launch of the plus program, cities did not experience significant changes, and on average a slight decline in booking rates. The results suggest that the success of the program in San Francisco may be related to the strong brand presence of Airbnb. The company was founded there and still locates its headquarters there *(Airbnb, 2024)*. These factors likely contributed to increased demand for plus listings. In other cities, the discrepancy of demographic and economic conditions may have limited the program's impact, emphasizing the importance of market-specific strategies for future roll-outs.

Future launches should focus on leveraging existing brand strength and aligning service upgrades with market-specific demand to ensure both higher booking rates and improved customer satisfaction. Markets with strong brand presence, high tourism demand, and affluent customer bases are more likely to benefit from premium programs like Plus. Findings additionally challenge the assumption about customer satisfaction. The plus program failed to improve average listing reviews, reflecting the complexity of consumer profiles and behavior. Programs should be tailored to both customer expectations, such as improved service quality, as well as the regional market characteristics when launching future service upgrade initiatives.

The analysis of the Airbnb Plus program has significant implications for the hospitality and platform economy. It highlights the importance of more data-driven strategies, in order to understand regional market dynamics and optimizing premium service roll-outs. The findings have lead to an increased interest in exploring how local economic conditions, brand presence, and customer behavior influence the success of service upgrades as such.

Limitations

Data Quality

- The presence of missing observations introduces potential biases. Future studies would benefit from ensuring high-quality data collection from reliable sources as well as more explanation on how data got conducted and is to be interpreted.

- The reliance on aggregated zip-code-level data may obscure individual-level trends or insights.

Causal Inference

- Although fixed-effects models help control for unobserved heterogeneity, potential endogeneity issues remain. For example, pre-existing booking trends could influence the selection of treated zip codes. More detailed data on how treated and non-treated zip codes were selected would strengthen causal interpretations.

Generalizability

- The findings may not generalize to all cities or time periods, particularly given the variability in program success across different markets.

AI use note: during this research the help of AI was used to debugging, layouts and text improvements.