FSAR Group Assignment

Fall 2024/2025

Deadline for Assignment 1: Friday, October 11th, 11:59PM.

Instructions

- Upload your assignment in **.Rmd format** to Moodle using the appropriate **submission link**. Ensure that **only one** member of your group submits the assignment.
- **R Markdown is mandatory**. The grader will compile your .Rmd file for grading. Ensure that your file compiles correctly to Word, HTML, or PDF with no errors.
- The generated report should have a **maximum length of 12 pages** (including all text, tables, and figures). Use **12-point** professional font for the main text (e.g., Times New Roman, Arial, or Calibri). Use **1-inch (2.54 cm) margins** on all sides and **1.5 line spacing** for the main text. Non-compliance with these specifications will result in grade penalties.
- Use comments (`#`) to organize, explain, and label your code clearly. **Good coding practices** and directory organization will be considered in grading.
- The **.Rmd document should include** all relevant R code but **R code should not appear** in the final report.
- **Presentation matters**: Your report should be styled **like a research paper** – professional, clear, and concise. Use **professional language** and structure, focusing on clarity and precision. **Key findings** should be highlighted, with brief interpretations focused on the most relevant insights, not an exhaustive presentation of all details.
- **Tables, figures, and results** should be concise and relevant to the analysis; they should be properly labeled, sized, and consistently formatted.
- **Citing your sources** is essential. If you use external resources, cite them properly. Use **APA citation style** for any references. Ensure that all sources are cited correctly, both in the text and in the reference list.

AI Tool Use Guidelines:

- You may use AI tools like ChatGPT to support your work (e.g., for brainstorming, clarifying concepts, or proofreading). However, **you are fully responsible** for all the content in your submission.
- Be critical and thoughtful in how you use AI-generated suggestions. Ensure that your analysis, interpretation, and discussion reflect **your own understanding** and work.
- Any misuse of AI tools that results in plagiarism will be treated as a violation of academic integrity.

Exploring the Impact of Airbnb Plus

Project Overview

In recent years, peer-to-peer (P2P) platforms like Airbnb have revolutionized the way people consume services. What began as a grassroots movement of sharing underutilized resources has now evolved into highly commercialized platforms catering to a wide array of consumer demands. As these markets shift from their early roots to more elite offerings, the question arises: can distinguishing high-quality services improve performance on these platforms?

Airbnb, the world's leading P2P accommodation platform, launched the Airbnb Plus program in 2018 to address the growing demand for premium accommodations. Listings under the Plus program are verified to meet higher quality standards, such as exceptional cleanliness, design, and comfort, offering guests an experience akin to a high-end hotel stay. This program aims to reduce search frictions for consumers by helping them easily identify premium offerings. But does quality differentiation through programs like Airbnb Plus lead to better matching between guests and hosts? Furthermore, does it impact all listings equally, or does it disproportionately favor higher-end listings?

In this project, you will analyze the Airbnb Plus program's impact on market-level performance, specifically examining whether the introduction of the program increased the number of booked nights in the affected areas. The focus will be on conducting an exploratory data analysis (EDA), visualizing trends, and testing hypotheses.

Dataset Description

The dataset for this project consists of Airbnb data from various U.S. cities **aggregated at the zip-code and month level**. This structure allows for the analysis of market-level trends and the impacts of the Airbnb Plus program on local markets before and after its introduction.

The variables include monthly measures of booking performance, pricing, and host characteristics, as well as socioeconomic and demographic indicators derived from the American Community Survey (ACS).

Below is the data dictionary for this dataset, containing the full variable list and descriptions:

Variable Name	Description
`listing_avg_review`	The average review rating of all listings
`listing_count_exit`	The number of leaving listings (i.e., exits)
`listing_count_entrant`	The number of new listings (i.e., entrants)
`listing_review_std`	The standard deviation of listing review ratings
`price_mean`	The average listing price
`price_mean_exit`	The average listing price of leaving listings
`price_mean_entrant`	The average listing price of new listings
`price_std`	The standard deviation of the average listing price
`timeperiod`	The year-month of an observation
`total_listing`	The number of total listings
`zipcode`	The ID of a zip code
`total_population`	Population
`med_household_income`	The median income of households
`housing_units`	Number of housing units
`rent`	The median rent of households
`plus_booking`	Average booked nights of Plus listings
`plus_price`	The average listing price of Plus listings
`close2plus_booking`	Average booked nights of close-to-Plus listings
`close2plus_price`	The average listing price of close-to-Plus listings

Variable Name	Description
regular_booking	Average booked nights of regular listings
regular_price	The average listing price of regular listings
close2plus_rating	The average review rating of close-to-Plus listings
regular_rating	The average review rating of regular listings
search_trend_airbnb	Airbnb Google search trend
search_trend_hotel	Hotel Google search trend
cancellation_flexible_rate	The ratio of listings that offer a flexible cancellation
treated	Treatment indicator for policy effect
policy_entry	Availability of the Plus program in a zip code in the current month (DID variable)
city_number	The ID of the city where a zip code belongs to
average_booked_nights	Average booked nights of all listings in a zip code in the current month
log_pop	The natural log format of total population
log_median_income	The natural log format of the median household income
log_median_rent	The natural log format of the median rent
log_housing_units	The natural log format of housing units
renter_occupied_rate	Percentage of renter-occupied housing units
above_college_rate	Percentage of the population with a high school degree or higher
year_built_to_now	The median age of all constructions
percent_income_spent_on_rent	Percentage of the median rent relative to the median household income
employment_rate	Percentage of the employed population
relative_time_m4	Relative time leads 4 months before the treatment period
relative_time_m3	Relative time leads 3 months before the treatment period
relative_time_m2	Relative time leads 2 months before the treatment period
relative_time_m1	Relative time lead 1 month before the treatment period
relative_time_0	The treatment period
relative_time_p1	Relative time lags 1 month after the treatment period
relative_time_p2	Relative time lags 2 months after the treatment period
relative_time_p3	Relative time lags 3 months after the treatment period
relative_time_p4	Relative time lags 4 months after the treatment period
log_pop_mean	Mean of the natural log of population over all time periods in a zip code
employment_rate_mean	Mean of the employment rate over all time periods in a zip code
log_median_income_mean	Mean of the natural log of median income over all time periods in a zip code
log_median_rent_mean	Mean of the natural log of median rent over all time periods in a zip code
log_housing_units_mean	Mean of the natural log of housing units over all time periods in a zip code
renter_occupied_rate_mean	Mean of the percentage of renter-occupied housing units over time
PISOR_mean	Mean percentage income spent on rent over all time periods
above_college_rate_mean	Mean percentage of population with higher education across all time periods
year_built_to_now_mean	Mean construction age over all time periods
average_booked_nights_201810	The number of booked nights in a zip code in October 2018
total_listing_201810	The number of listings in a zip code in October 2018
ln_total_listings	The natural log format of total listings
ln_total_hosts	The natural log format of total hosts
entire_ratio	The ratio of listings that provide an entire property
business_ratio	The ratio of business listings
instant_ratio	The ratio of listings that are instantly bookable
average_listing_price_201810	The average listing price in a zip code in October 2018

Variable Name	Description
market_thickness	Market thickness measured by the number of listings
high_thickness	Dummy variable indicating if market thickness is high
policy_entry_low_thickness	Interaction of policy entry and low market thickness
policy_entry_high_thickness	Interaction of policy entry and high market thickness
normalized_ln_market_thickness	Normalized market thickness
high_dispersion	Dummy variable indicating if rating dispersion is high
normalized_rating_dispersion	Normalized rating dispersion
policy_entry_low_dispersion	Interaction of policy entry and low rating dispersion
policy_entry_high_dispersion	Interaction of policy entry and high rating dispersion
plus_total_listing	The number of total Plus listings
close2plus_total_listing	The number of total close-to-Plus listings
regular_total_listing	The number of total regular listings
availability_future_two_months	Future available nights in the next two months
plus_rating	The average review rating of Plus listings
super_host_ratio	The ratio of super hosts
super_host_std	The standard deviation of being super hosts or not
total_bookings	The number of total booked nights
ln_total_bookings	The natural log format of total booked nights
newly_added_reviews	The average number of newly-added reviews
previous_listing_avg_review	The value of listing_avg_review in the previous time period
previous_price_mean	The value of price_mean in the previous time period
previous_host_avg_res	The value of host average response in the previous time period
previous_total_host	The value of total_host in the previous time period
previous_ln_total_listing	The natural log of total listings in the previous time period
previous_average_booked_nights	The average booked nights in the previous time period
ln_close2plus_booking	The natural log of close-to-Plus bookings
ln_regular_booking	The natural log of regular bookings
ln_plus_booking	The natural log of Plus bookings
ln_plus_price	The natural log of Plus listing prices
ln_close2plus_price	The natural log of close-to-Plus listing prices
ln_regular_price	The natural log of regular listing prices
ln_booked_nights	The natural log of booked nights
rating_dispersion	Rating dispersion measured by the standard deviation of review ratings
booking_rate	The average booking rate of all listings

PART 1

1. Dataset Overview

Provide an overview of the dataset structure. Include details such as the total number of observations, key outcome and explanatory variables, and the geographic and temporal scope of the dataset. Indicate which variables are central to your analysis and which are secondary.

Guidance: Focus on summarizing the key features of the dataset. Make sure to identify the most important variables for your analysis, as well as any initial observations related to the scope and structure of the data, such as the time period covered or how many zip codes and cities are represented. Complement with any information or insights you deem relevant for an outsider to understand the structure and key features of the data.

2. Data Quality Assessment

Evaluate the dataset for data quality issues such as missing values, outliers, or unusual patterns. Propose and explain how you plan to handle these issues, ensuring your approach is both justified and reproducible. Focus on addressing issues affecting the variables that are most relevant to your analysis.

Guidance: It's important to address any data quality issues in a thoughtful way. Not all missing values or outliers are problematic, so focus on the variables that matter most for your analysis and explain how you will handle any issues you find. Make sure your proposed solutions are reasonable, transparent and well documented.

3. Exploratory Data Analysis (EDA)

Conduct an Exploratory Data Analysis (EDA) that provides insights into the data. Use both summary statistics and visualizations to represent variable distributions and explore relationships between key variables. Structure your analysis in a way that tells a coherent story, rather than presenting disconnected graphs or tables.

Note that the dataset allows for analysis at different levels of aggregation, such as city, zip code, and time. You can explore the data from various perspectives, such as:

- **Zip Code/City-Levels**: Investigate how Airbnb performance varies across different zip codes or cities.
- **Time-Level**: Examine how Airbnb performance evolves over time, especially before and after the launch of the Airbnb Plus program.

Guidance: Focus your EDA on telling a clear, data-driven story that ties back to your key business questions. Use visuals and numeric summaries effectively, but avoid overloading your analysis with too many plots or tables. Be selective and thoughtful in your approach to ensure clarity and readability. You don't need to present and describe every single variable in the dataset, select the information that is more relevant.

4. Hypothesis Testing

Develop and test three research hypotheses about the impact of the Airbnb Plus on relevant business outcomes. Interpret your results, their statistical significance, and their potential business implications.

Guidance: When formulating your hypotheses, focus on questions that are directly relevant to the main business problems at hand. You are free to aggregate or structure the data in a way that supports your analysis. Be sure to interpret your findings discussing both the statistical outcomes and the practical implications of your results.