Team Restaurant EDA Final Project

Balajee Devesha Srinivasan (basrini@iu.edu) Ranvir Singh Virk (rsvirk@iu.edu)

Introduction

Aim

- The aim of this project is to conduct a thorough analysis and modeling of the various factors influencing customer visitation trends to Japanese restaurants. By identifying and understanding the key determinants that drive these trends—such as seasonal variations, location and genre of restaurants, and others—we intend to develop a robust predictive model.

Why Japanese restaurants ?

- Well, who doesn't like Japanese food and culture? Jokes aside we care about this project because it teaches us how to develop and interpret:-
  - Business Strategy: Essential for resource allocation and operational efficiency.
  - Market Dynamics: Understanding trends helps us adjust to dynamic market conditions.

Why should anyone else care ?

- It helps us learn how to model and explore:-
  - Industry Insights: Valuable for stakeholders in the restaurant and hospitality sector.
  - Consumer Behavior: Reflects broader trends in customer dining preferences and habits.

Data Description

- Our data is "Recruit Restaurant Visitor Forecasting dataset" from Kaggle
- The data comes in the shape of 8 relational files which are derived from two separate Japanese websites that collect user information:

  Hot Pepper Gourmet (HPG): like Yelp, here users can search restaurants and make a reservation for their chosen restaurant.

  Air REGI/Restaurant Board (AIR): like Square, a reservation control and cash register system for users to book an appointment of their liking.

- The dataset represents a ***sample of visitor trends*** **from Japanese restaurants**, primarily sourced from two restaurant reservation websites: HPG and AIR. It ***offers valuable insights into consumer behavior and patterns in the restaurant industry*** within Japan.
- And the dataset is significant as it ***reflects real-world trends*** and can be instrumental in forecasting demand, aiding restaurant management, and ***understanding the impact of external factors like holidays on visitor trends***.
- The Individuals in this dataset are the **restaurant visitors** which may include tourists, locals or anyone in general who make a reservation using the aforementioned websites.

Key variables in this dataset include:

- ***Visitors***: The count of daily visitors to the restaurants. It's the primary variable of interest for forecasting demand.
- ***Reservations***: Data from both HPG and Air systems, capturing the number of reservations made and the number of visitors per reservation.
- ***Restaurant Information***: This includes "air_store_info" and "hpg_store_info", providing details about the restaurant's genre and geographical location.
- ***Date Information***: Includes data on Japanese holidays, which is crucial for understanding seasonal patterns and special occasions affecting visitor numbers.

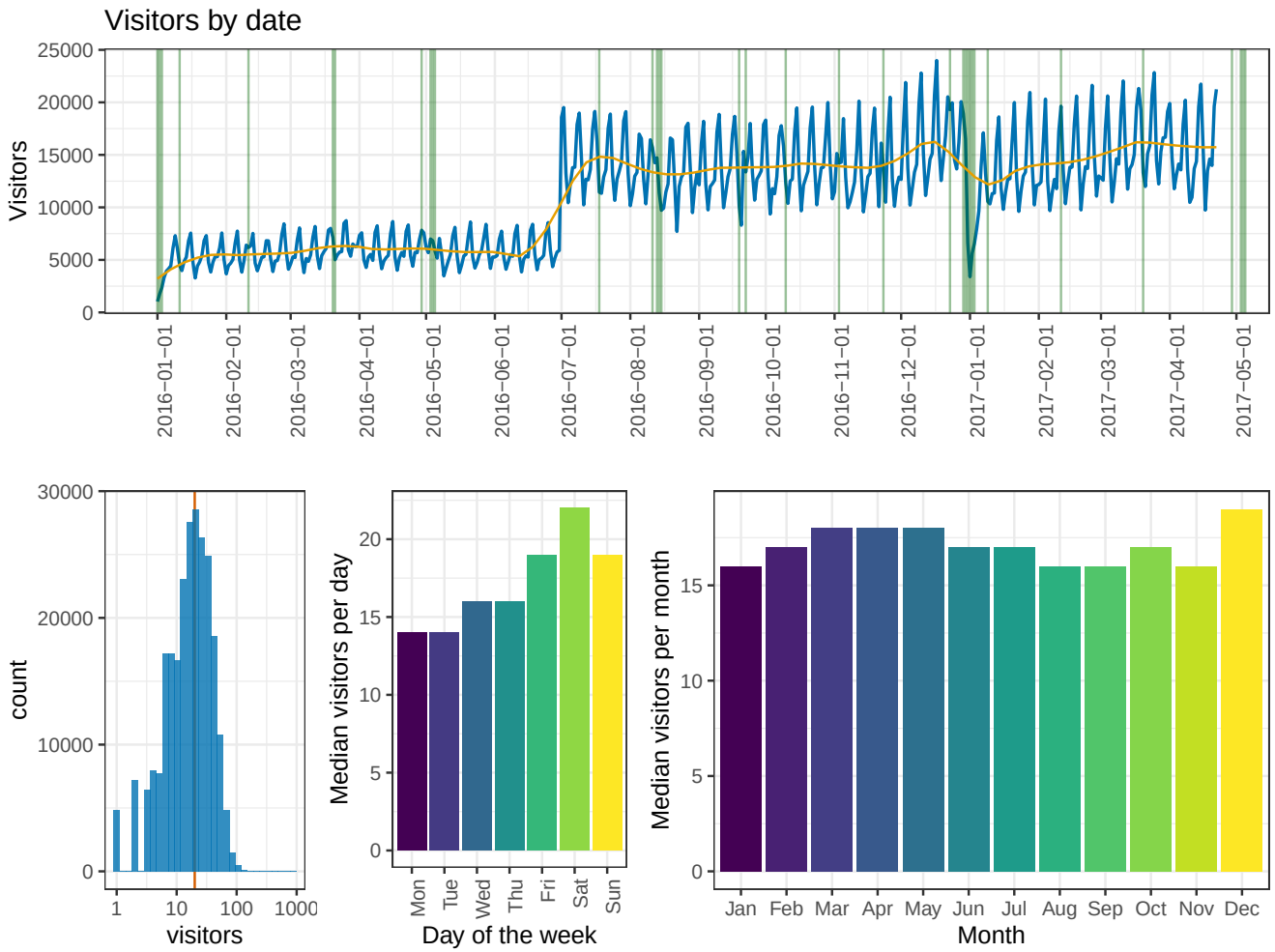
These variables are measured using actual transactional data and reservations made through the two websites, offering a reliable representation of visitor patterns.

Data Exploration

- Air visits: The air_visits.csv has numbers for each visit_date and air_store_id for 829 different stores.
- Air Reserve: This file includes the date and time of the reservation, as well as those of the visit. We have reservation numbers for 314 air stores.
- HPG Reserve: Similar to Air reserves we have reservation numbers for 13325 hpg stores.
- Air Store: This includes the name of the particular cuisine along with the name of the area. There are 829 different stores.
- HPG Store: Similar to Air store, includes the name of the particular cuisine along with the name of the area but with the addition of "style" at the end of cuisine name. There are 4690 different stores.
- Holidays: The data info file is stored as holidays has the binary flags for the holidays in dtae format.
- Store IDs: This is a relational file that connects the air and hpg ids. There are only 150 pairs, which is less than 20% of all air stores.
- Overall, There are no missing data in the provided files.

Now that the basic description is completed lets try to visualize our Key target variables and its associations.

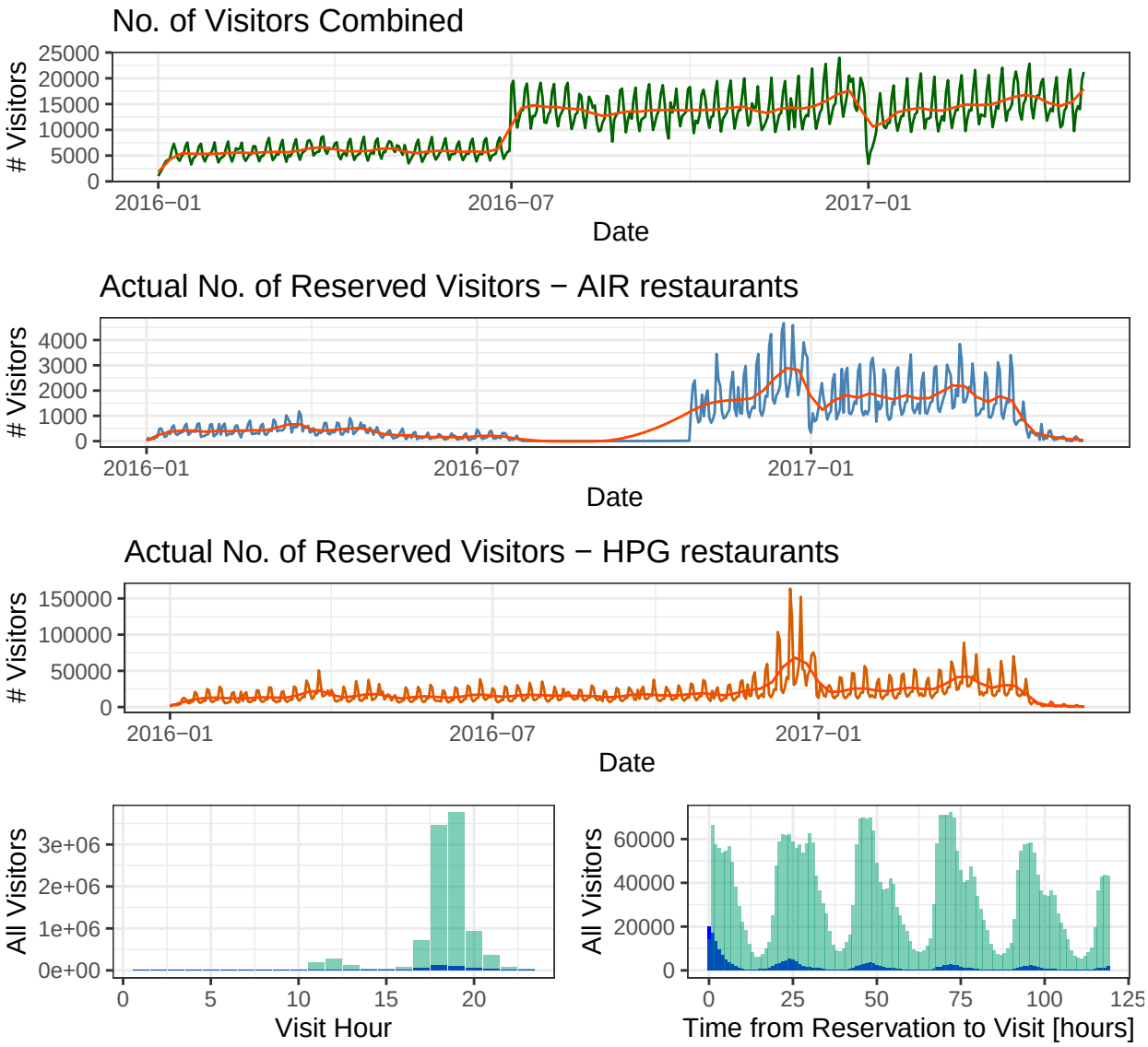
Target Variable - Visits to the Restaurants



Observations

- The above plots visualizes the number of visitors to the restaurants as a time series and also tries to visualize the long-term structure present in the data
- We see a few periodic and cyclical trends corresponding to the weekly and monthly/yearly trends as shown above.
  - On average the visitors peak to 20 per day with max being around 100.
  - The weekends seem to be the most popular time to visit, which is to be expected. Conversly, early week is the least popular time for visiting.
  - As far as months go, December and End of Year sees the largest crowd however we also observe steady business from March to May.
- One peculiar observation can be seen as an near 125 - 150% increase in the number of visits to restaurants during mid of 2016. The reason behind the hike is addition of ~500 new restaurants to the AIR database.
- Additionally, The sharp fall in visits on 1st of Jan is due to the new-year's eve, as most of the restaurants stay closed on new-years day.

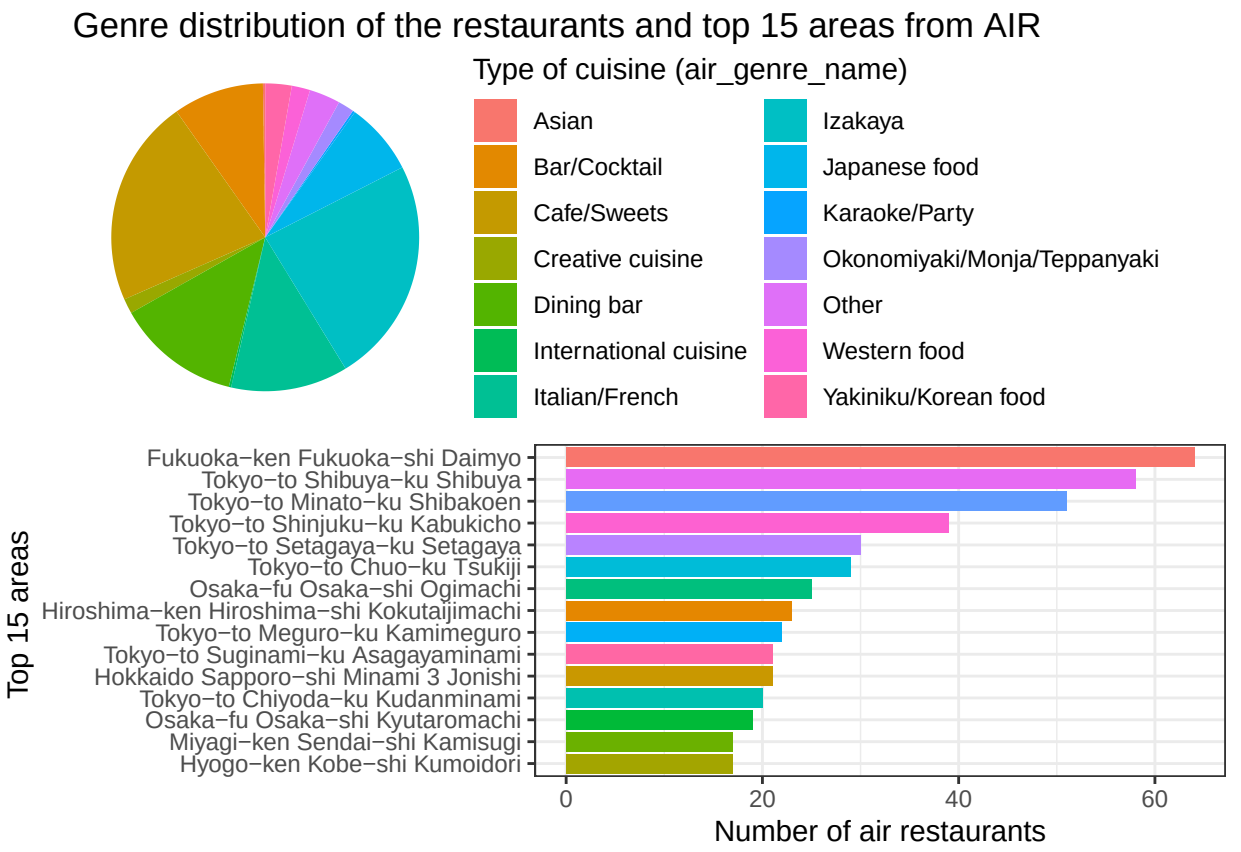
Reservations V/S Visitors

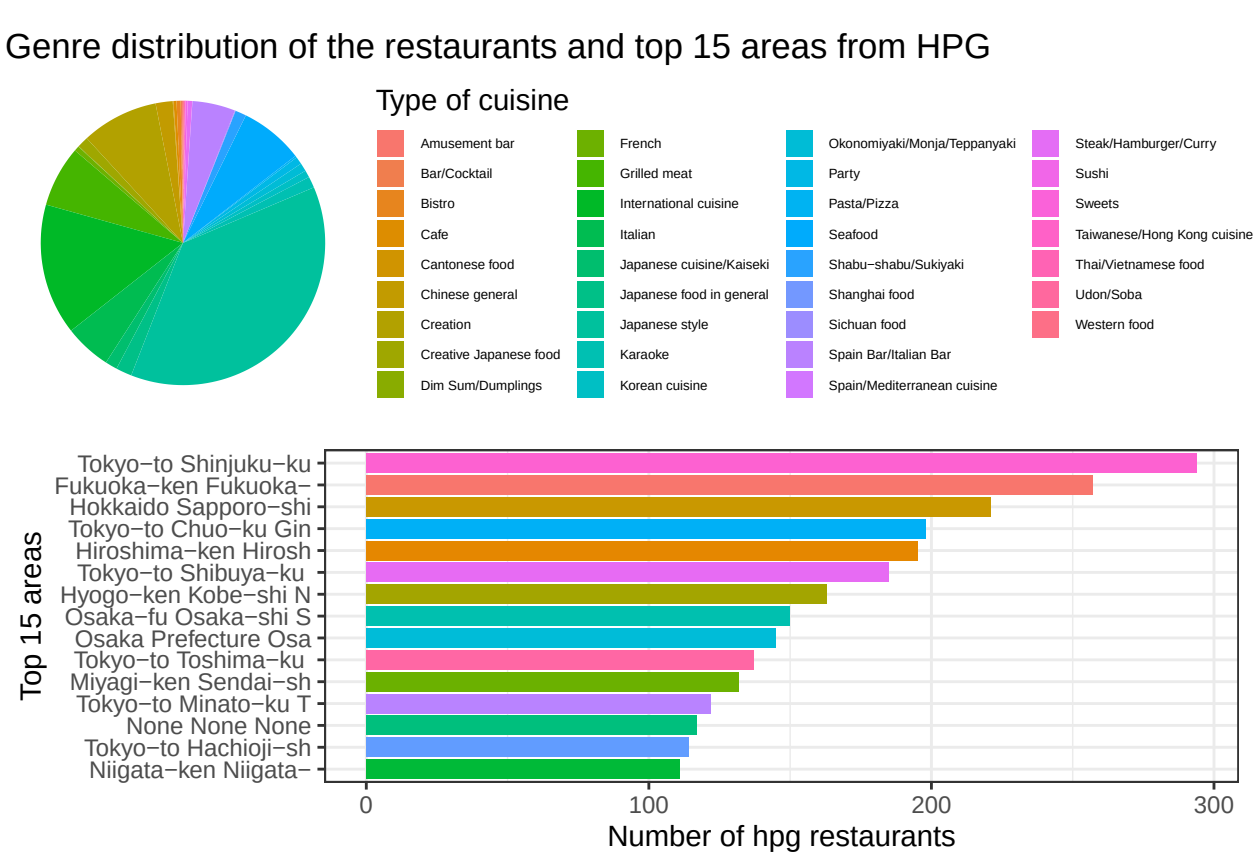


Observations:

- The combined total visitors plot over span of around 2 years shows several pronounced spikes in the data, which could indicate special events, promotions, or data collection anomalies.
- There is a clear concentration of visits around the typical dining hours, peaking in the evening. Showing the highs frequency of visits around dinner time, which is consistent with typical dining patterns. There's also a smaller peak during lunch hours, indicating a significant, albeit lower, number of visits.
- The peaks at regular intervals could represent common booking behaviors, such as making reservations one day (24 hours), two days (48 hours), or a week (around 168 hours) in advance.

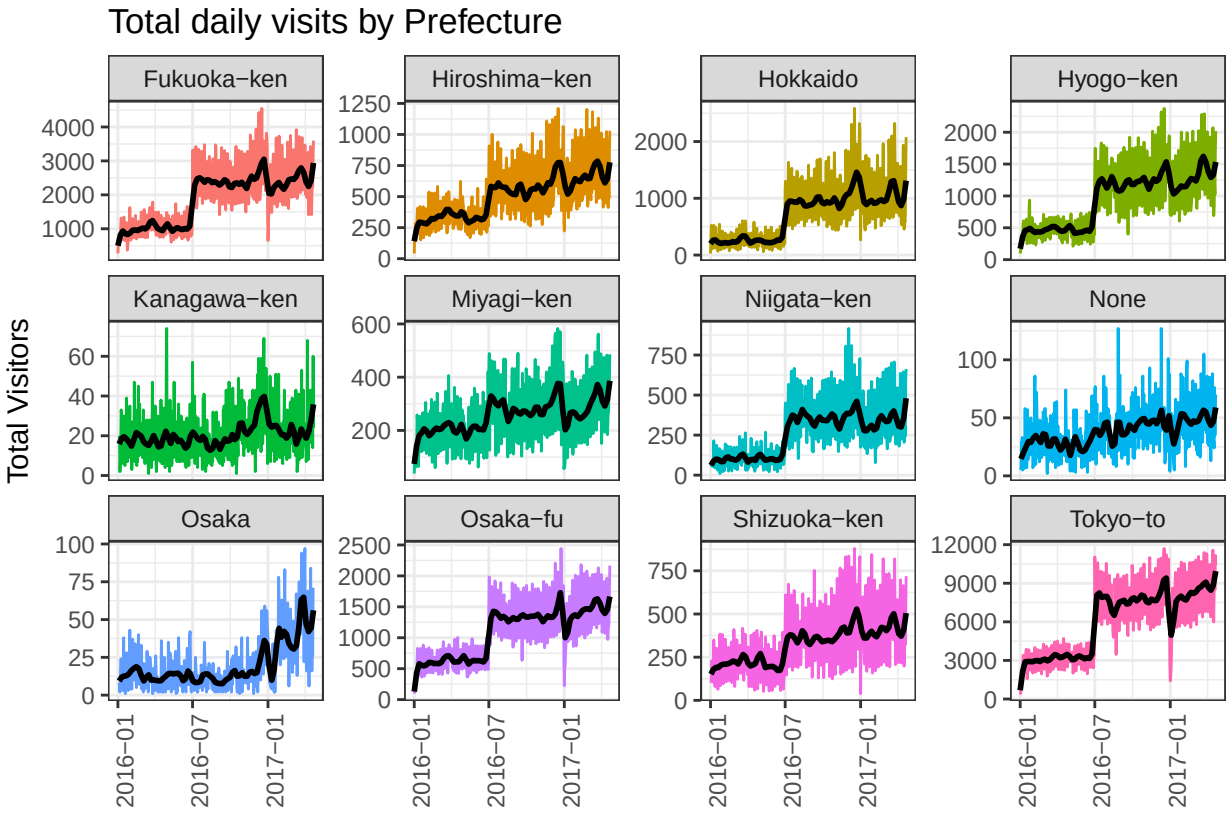
Genre and Top areas



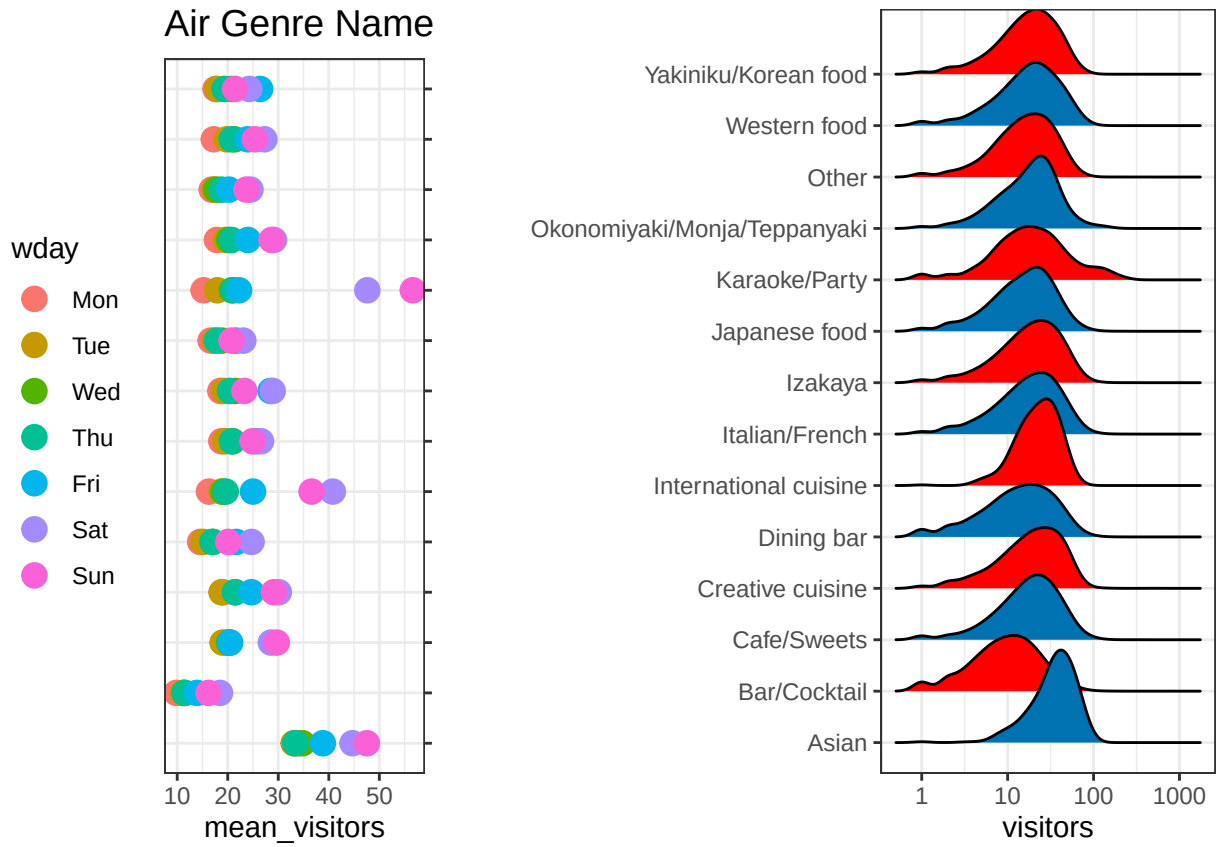


Observations:

- Traditional Japanese and Japanese style food seems to hold the largest market share, followed by international cuisine and creative cuisine. This indicates a strong local preference but also a substantial interest in diverse culinary experiences.
- Niche cuisines like Spanish and French have smaller representations, potentially indicating less competition and opportunity for growth.
- The variety of cuisines, including a notable segment for Asian and creative cuisines, suggests that AIR caters to a market with eclectic tastes, balancing traditional Japanese and broader Asian flavors with innovative dining concepts.
- For the chart demonstrating the distribution of AIR restaurants by cuisine type, Izakaya is the most prevalent genre, likely due to its cultural significance and popularity in Japan. The presence of various international cuisines, such as Italian/French and Western food, indicates a cosmopolitan dining scene.
- Beyond weekly cycles, the loess smoothing lines could also be capturing seasonal trends in dining, such as holiday periods where certain prefectures might see a surge in visitors due to local events or as seasonal destinations.
- The patterns vary by region, with some prefectures like Tokyo-to, reflecting its large population and status as a culinary hotspot. In contrast, prefectures like Miyagi-ken and Shizuoka-ken show more modest visit totals.
- The black loess smoothing lines reveal the underlying trends and weekly cycles. Some prefectures, like Osaka, display more pronounced weekly cycles, potentially indicative of weekend surges in dining out.
- The plots suggests that while some areas are culinary hotspots with a high density of restaurants, other regions might offer market opportunities for new establishments, especially for cuisines that are currently underrepresented



Genre V/S Weekday

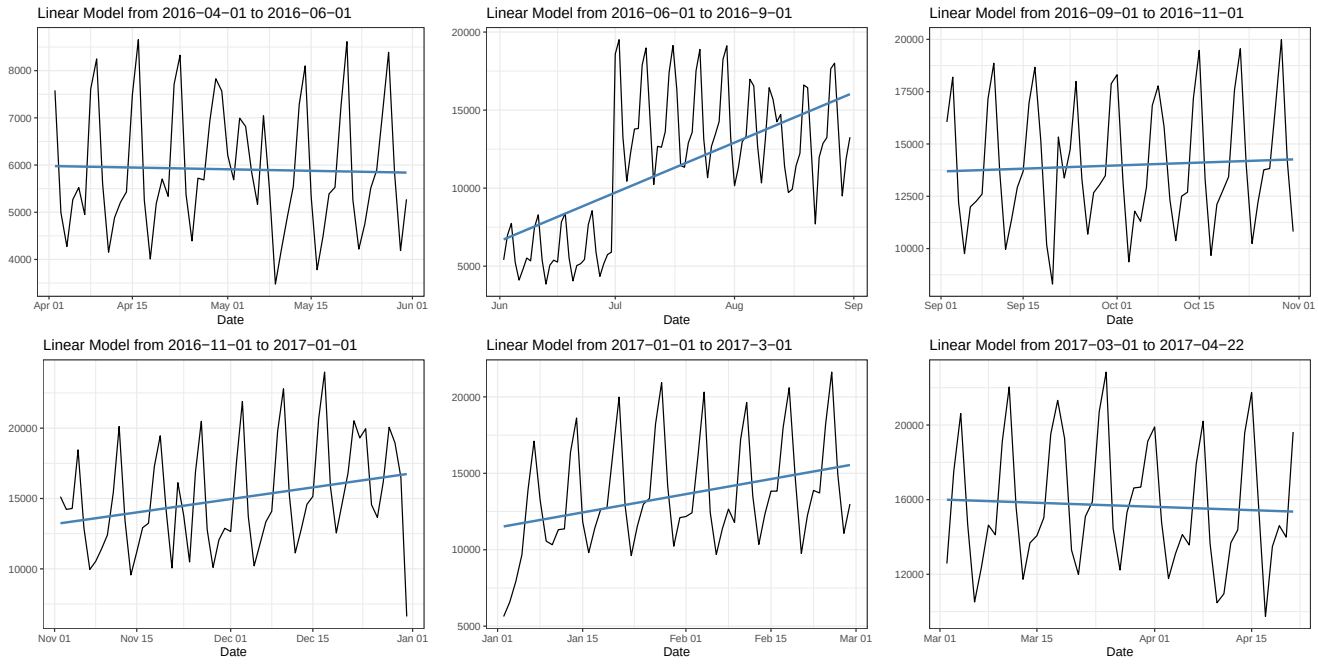


- The analysis of these plots could inform restaurant owners about the average expected customer flow throughout the week, helping to manage staffing and inventory.
- The consistent mean visitors across the week suggest that there may not be significant spikes or drops on any particular day for any genre, but there are 3 cuisines that do have an significant increase towards the weekends those being karaoke, Asian, and International. On top of that, Bar/Coctails tend to stay low.
- The wider distribution in the density plots for certain genres indicates variability in visitor numbers, which could be due to special events, promotions, or other factors that occasionally draw in more or fewer visitors.
- Restaurant genres with narrower peaks in their density distributions may have more predictable visitor numbers, allowing for more precise planning and operations.

Overall, these plots provide a multifaceted view of the dining landscape, highlighting both the average trends and the variability in restaurant patronage, which is crucial for making informed business decisions.

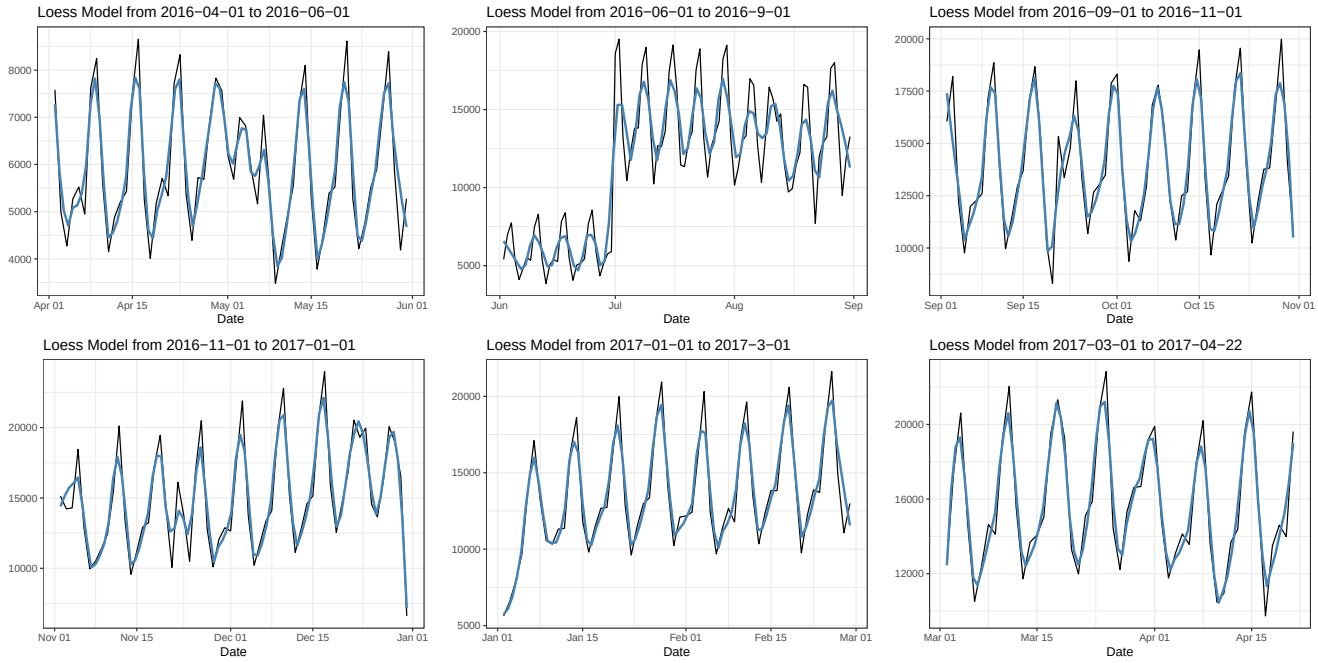
Preliminary Modelling

Linear model



The series of linear model fits depicted in the graphs illustrate that while linear trends can capture the broad direction of visitor counts over time, they do not account for the pronounced weekly fluctuations evident in the actual data. These models are beneficial in establishing a general trend, such as the significant upward trend from June to September, which can be useful for long-term capacity planning. However, the consistent underfitting of peak times and overfitting of off-peak times indicates that additional models, perhaps with higher complexity, would better capture the nuanced patterns for more precise operational and strategic decision-making, such as weekly staffing and inventory management.

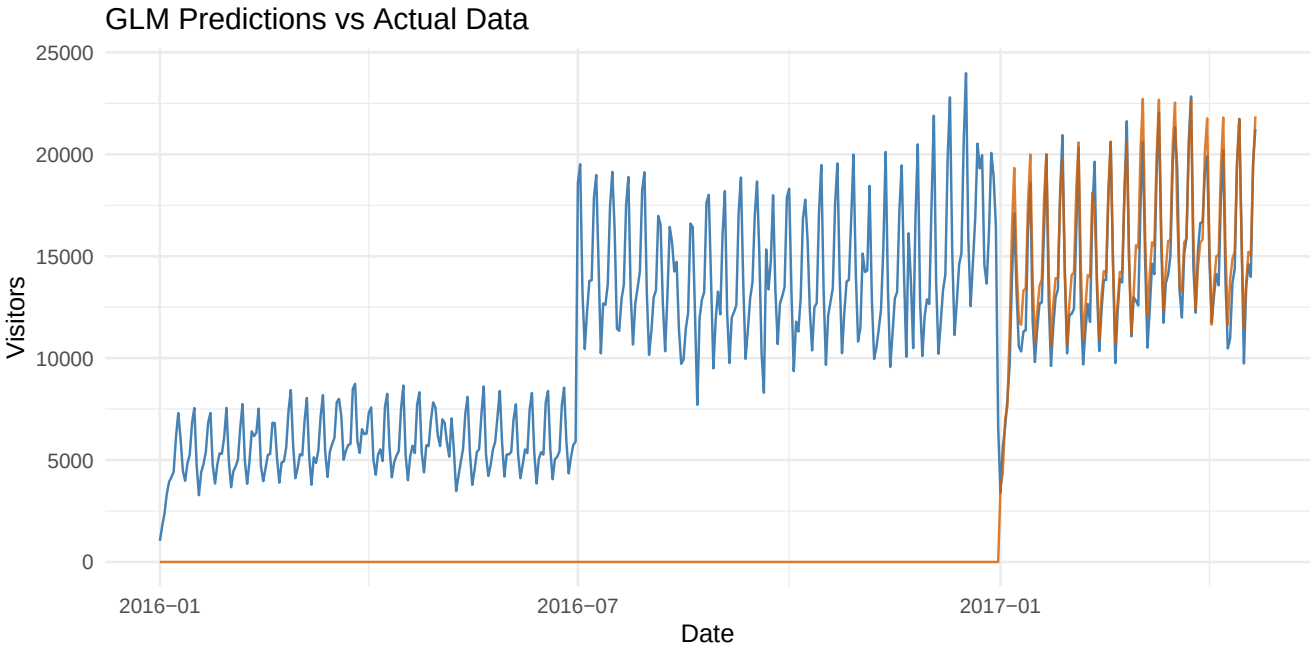
Loess model



The LOESS models accurately reflect the weekly visitor trends. These insights are critical for planning workforce

and managing supplies. With a robust R-squared value of 0.8271, the models prove their merit in closely following visitor behavior. However, despite their precision, LOESS models may not be the best fit for scaling up due to their intensive computations. Therefore, we might consider Generalized Linear Models (GLM) for their broader applicability and ability to integrate various influencing factors, thus enhancing our strategic planning capabilities.

GLM with interactions

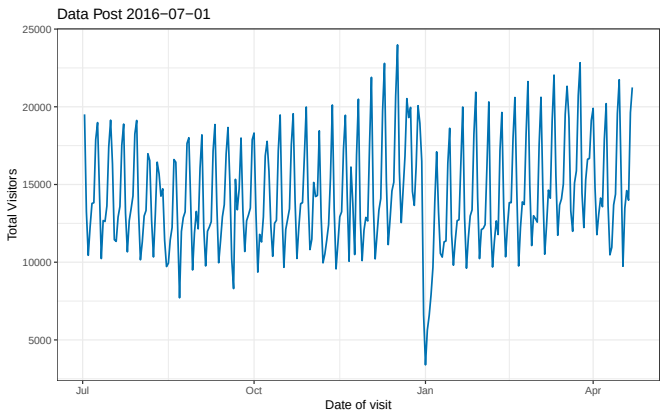


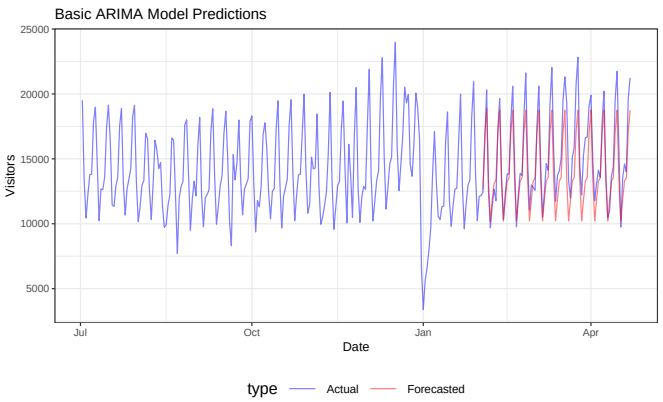
The GLM model's use of a Poisson distribution for predicting the number of visitors at Japanese restaurants, starting from 2017, reveals its strengths and limitations. The orange line, representing the model's predictions, closely traces the actual visitor count shown in steel blue, indicating a general understanding of visitor trends. However, there are discrepancies where the model fails to match the actual numbers perfectly, highlighting areas for improvement in capturing daily fluctuations.

The key interactions that proved most insightful in this model are weekdays, holidays, and location, closely followed by genre. These factors are crucial because they directly influence customer flow patterns. Weekdays and holidays represent different customer behaviors, with weekdays typically showing regular, predictable patterns and holidays introducing variability due to special events or seasonal changes. Location is also a significant factor, as it dictates the potential customer base's size and characteristics, varying from busy urban centers to quieter suburban areas. Lastly, the genre of the restaurant can attract different demographics and influence peak times.

Advanced Modelling

The ARIMA model, tailored to our data after mid-2016, serves as a solid foundation for predicting the number of visitors at Japanese restaurants. This time series forecasting model, adjusted with data following the July 2016 update, demonstrates its effectiveness in forecasting future trends. The left side of the plot shows the actual data with its natural variability, while the right side presents the model's forecasts, which start to diverge slightly from the actual data. This divergence highlights the model's capability for short-term prediction but also signals the need for ongoing adjustments to maintain long-term accuracy. By using this model, we're better equipped to predict and prepare for the ebb and flow of customer visits, ensuring that our restaurant can meet demand efficiently.





Conclusions

The extensive data analysis of Japanese restaurants reveals that location is a pivotal factor affecting visitor trends. Through some data exploration and looking at individual features used for modeling, we can see that urban areas, especially Tokyo, emerge as hotspots with a high density of visitors, likely due to their larger populations and status as economic and cultural hubs. Tokyo has the largest crowd of both local and foreign audience, weather it be from a different part of Japan or the World Tokyo has established itself as one of the most touristy places. The preference for certain cuisines, such as *Izakaya*, *Café/Sweets*, and *Japanese-style options*, suggests a strong market for these genres, particularly in bustling city centers. Conversely, regions like *Miyagi-ken* and *Shizuoka-ken* display more modest visitor numbers, potentially offering untapped opportunities for new ventures.

If one was trying to open a restaurant in Japan, it is essential to focus on:

- **Location Selection**: Targeting areas with high foot traffic, such as Tokyo's *Shinjuku* and *Shibuya* districts, could capitalize on the established influx of potential customers. Being tourist spot and a world's biggest city, Tokyo has an unfathomable number of restaurants with varying cuisine options and style.
- **Cuisine Differentiation**: While traditional Japanese and Asian cuisines are prevalent, there is room for growth in less-saturated markets such as international and niche cuisines, as evidenced by the diversity of restaurant types. These dishes give a variety to the native Japanese audience but also provide a familiarity to the international tourists.
- **Understanding Customer Patterns**: Observing the clear weekly and seasonal patterns in restaurant visitation can inform operational decisions like staffing and inventory management. It is very clear that most busy days for your restaurant are going to be Fridays and Saturdays, this will require a greater number of staff and preaping for the meals to be cooked.

In essence, success in Japan's competitive restaurant industry hinges on a strategic location that aligns with the type of cuisine offered, understanding the local market dynamics, and adapting to the rhythmic flow of customer visitation to optimize service and maximize business potential.

Adding insights from various statistical models we used elevates our analysis of Japanese restaurant visitation trends. These models affirm that while location is crucial, understanding the rhythm of customer visits underscored by our findings of peak times weekly is equally vital for operational efficiency. Linear and LOESS models pinpoint trends and nuances in urban demand, suggesting that areas like Tokyo's Shinjuku and Shibuya could be lucrative due to their bustling activity. The GLM and ARIMA models enhance our ability to forecast and plan for fluctuating visitor numbers, which is indispensable for managing the ebb and flow in busy city centers and for positioning new restaurants.

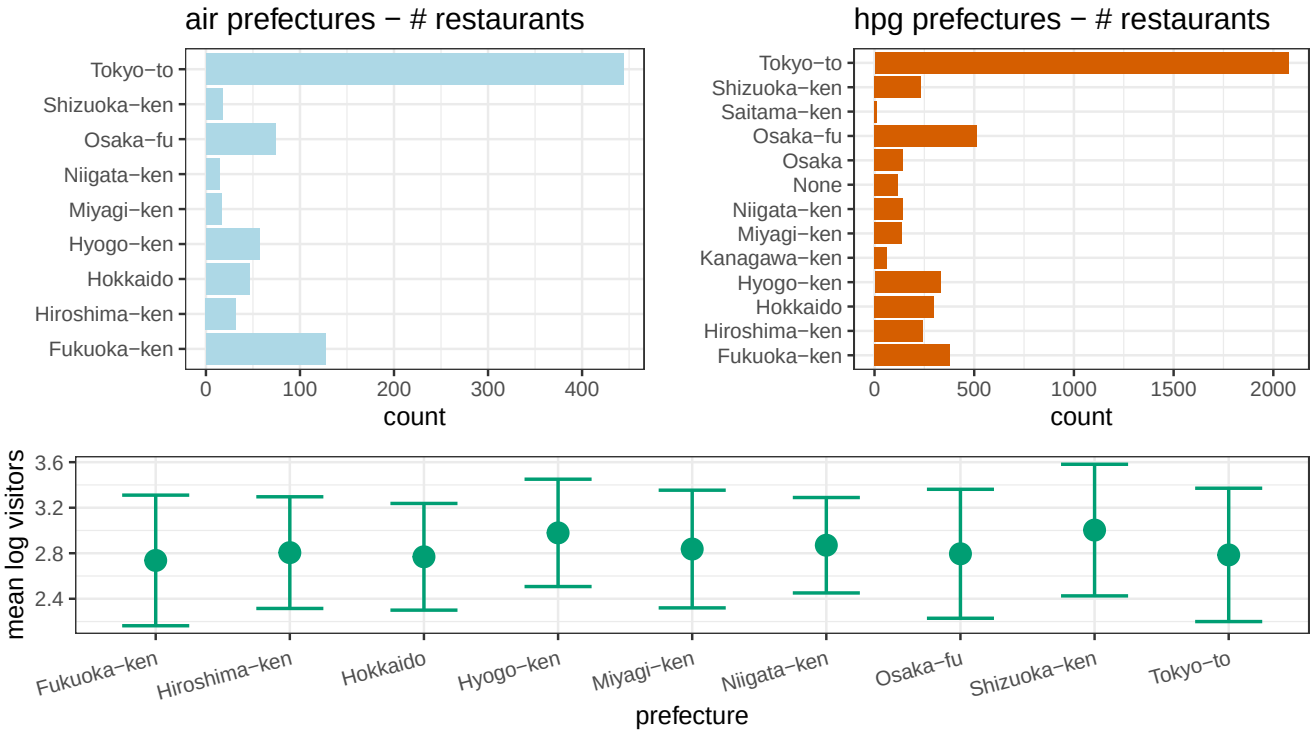
Limitations and Improvements

The analysis of Japanese restaurant visitation trends has been comprehensive, yet there are inherent limitations to the process. While urban areas like Tokyo are recognized as prime locations, over-saturation in these markets could lead to increased competition, which may not be adequately captured by broader data trends. Additionally, factors

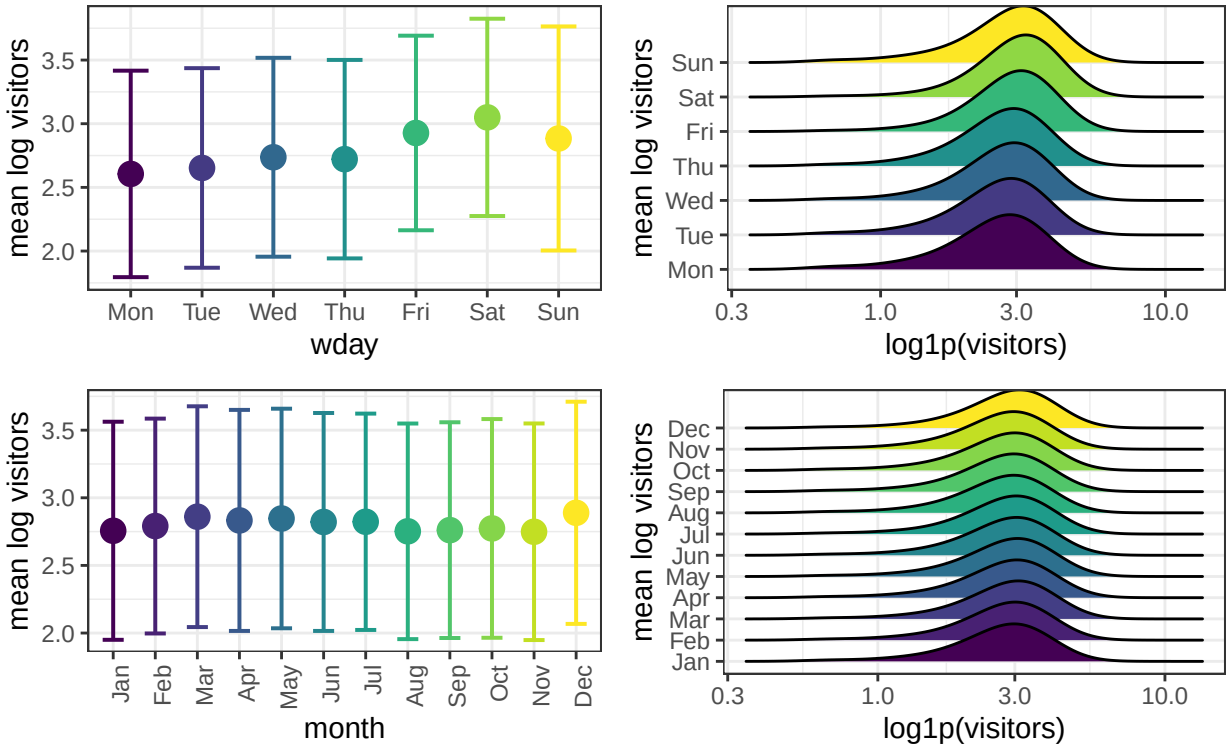
such as changing consumer preferences, economic fluctuations, and unforeseen events like public health concerns can alter visitation patterns in ways that static models may not predict accurately.

Appendix

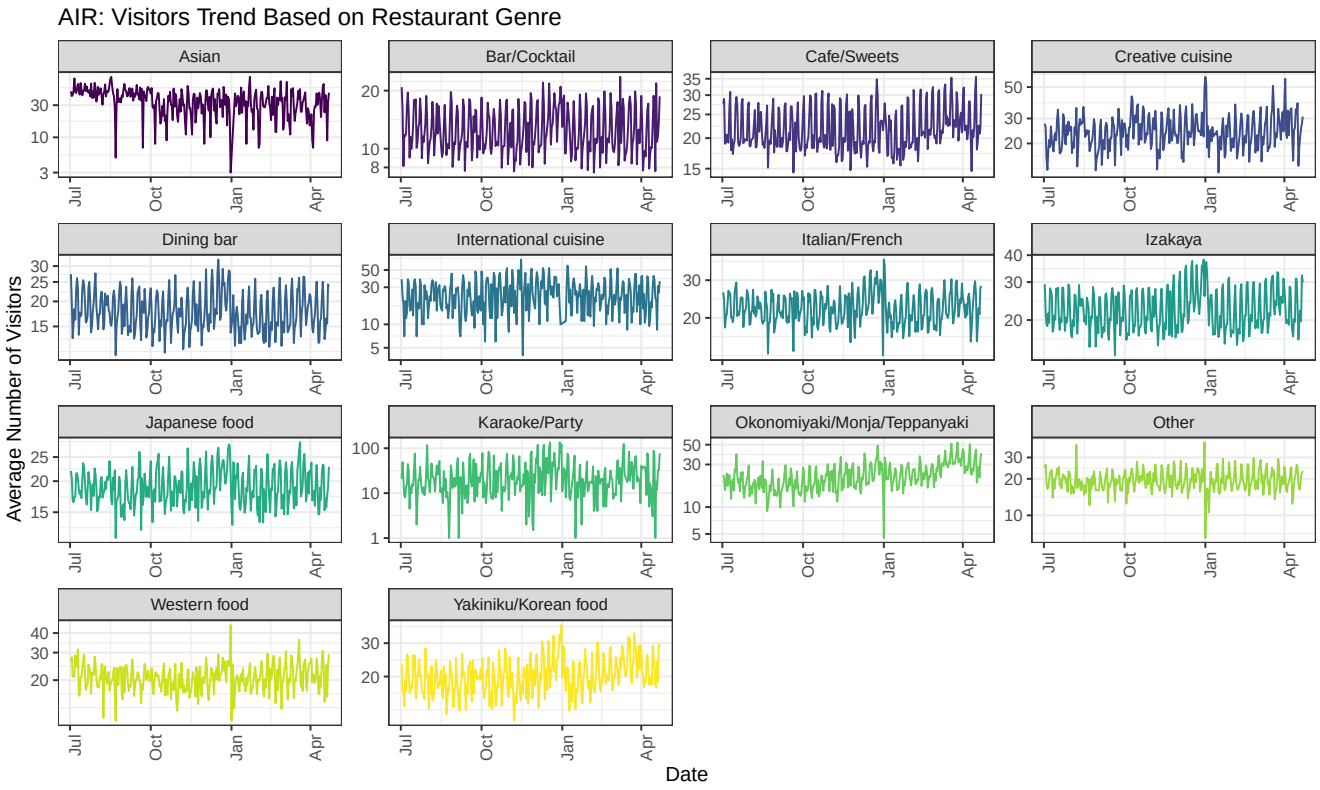
More Feature Interactions and transformation



Why to use logq transformation for the visitors for advanced modelling



Visitors per genre and prefecture

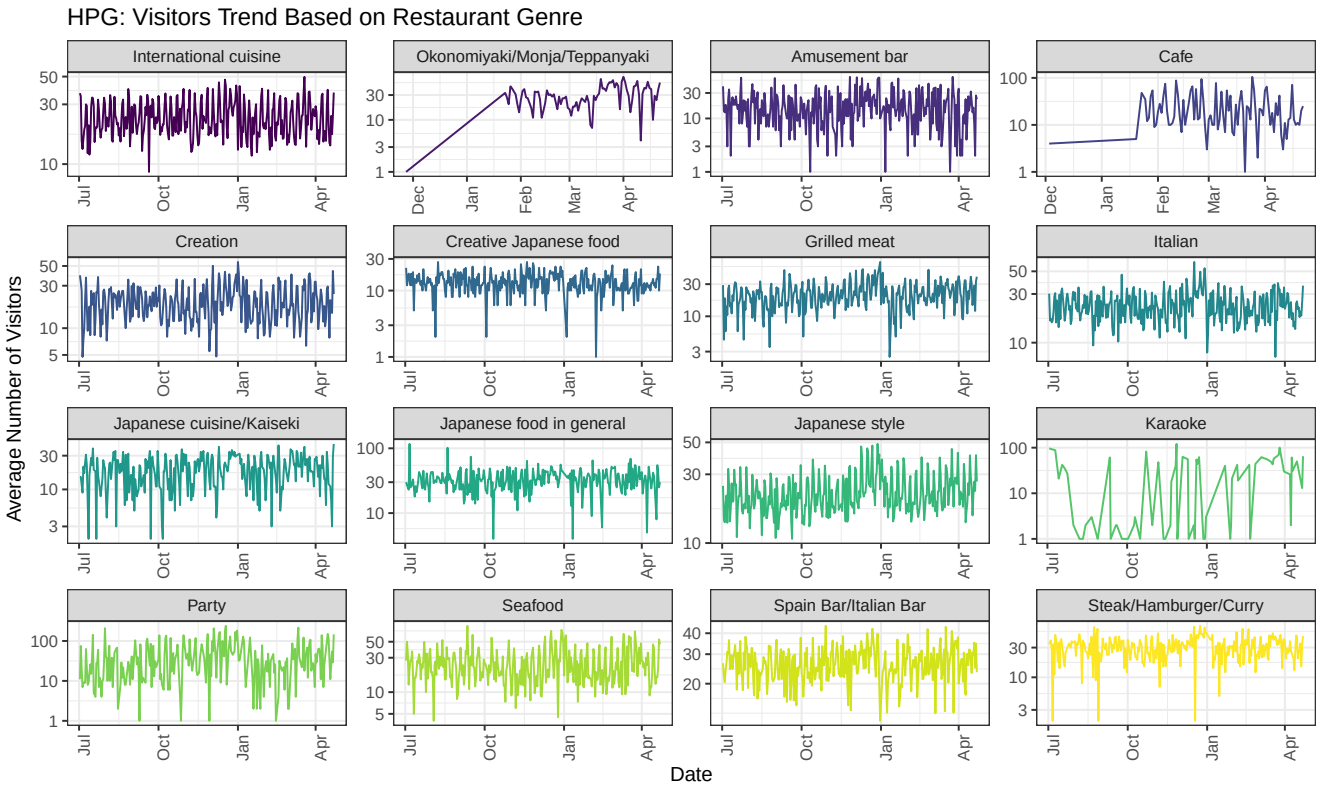


Observations:

- All cuisine genres demonstrate a relatively stable number of visitors throughout the year, with no significant

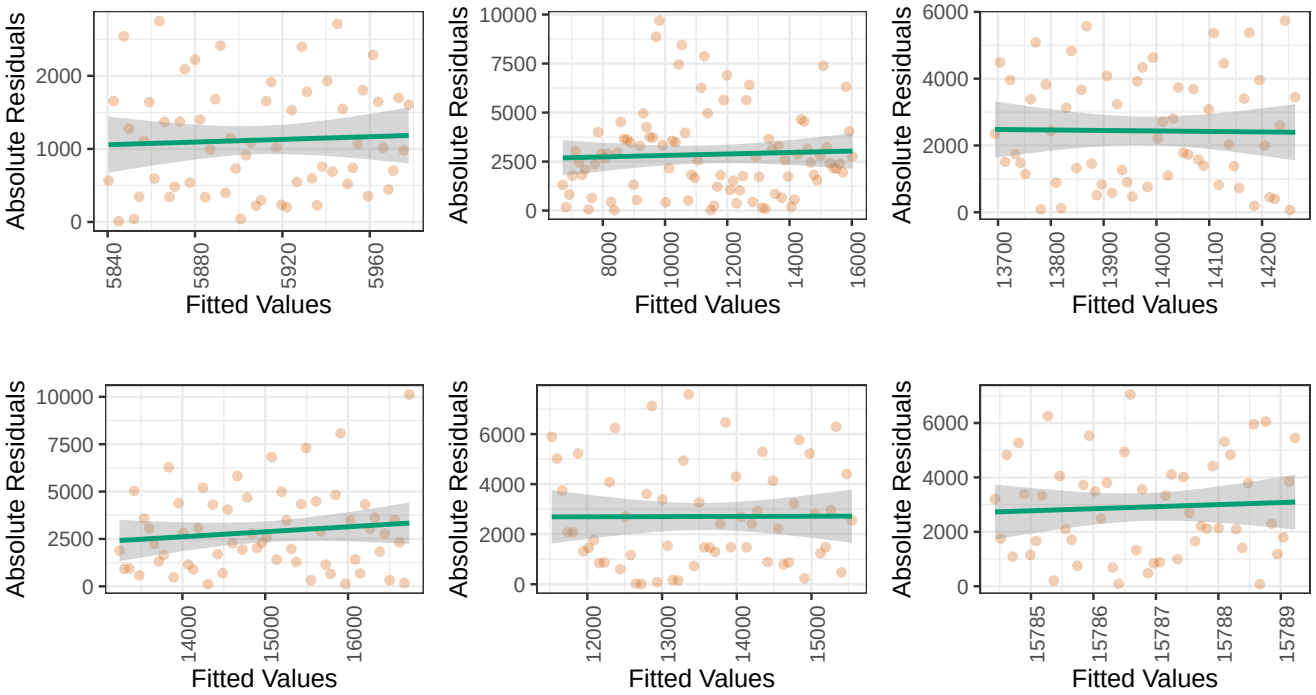
long-term upward or downward trends. This suggests a consistent demand for various dining experiences

- Genres such as Izakaya, Café/Sweets, and Bar/Cocktail maintain a steady flow of visitors, which is indicative of the popularity of casual and social dining settings in Japanese culture.
- Some variability in visitor numbers can be seen, possibly reflecting seasonal fluctuations, holidays, or special events that temporarily affect the popularity of certain cuisines
- Cuisines such as Creative cuisine, Karaoke/Party, and Okonomiyaki/Monja/Teppanyaki show lower average visitors but maintain a consistent base, likely serving niche markets that cater to specific customer preferences.



Modelling

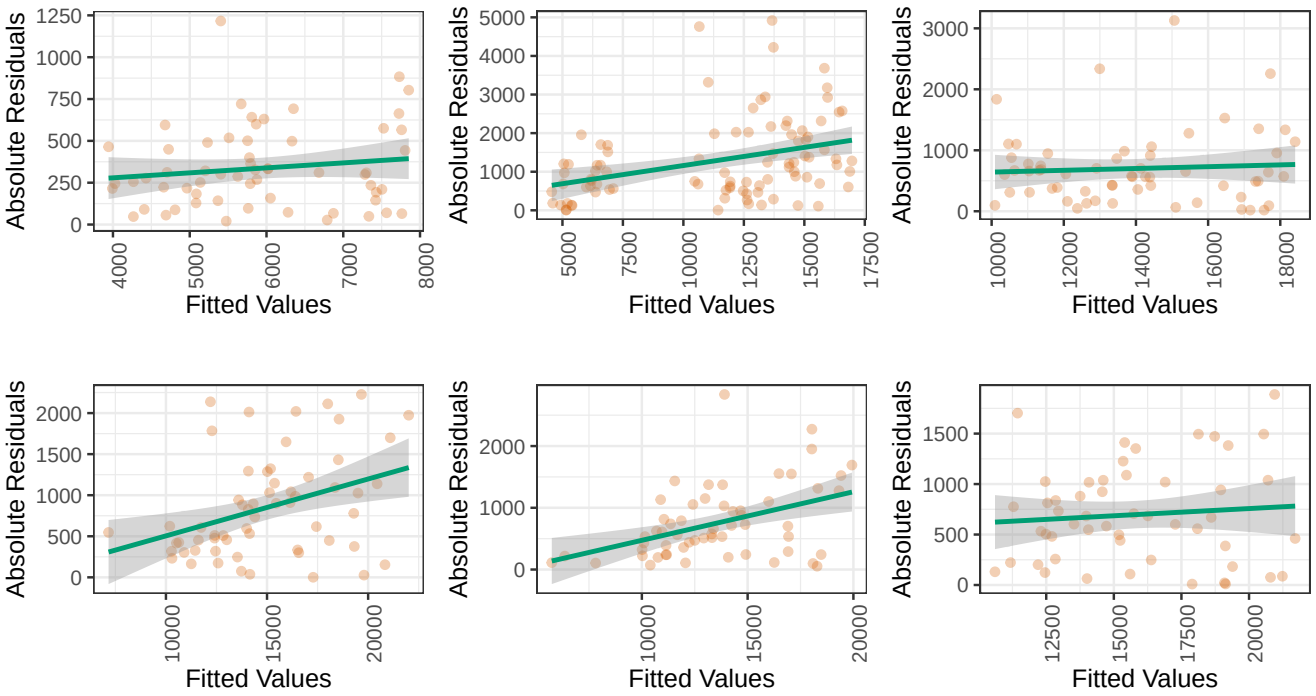
Linear Model Residuals



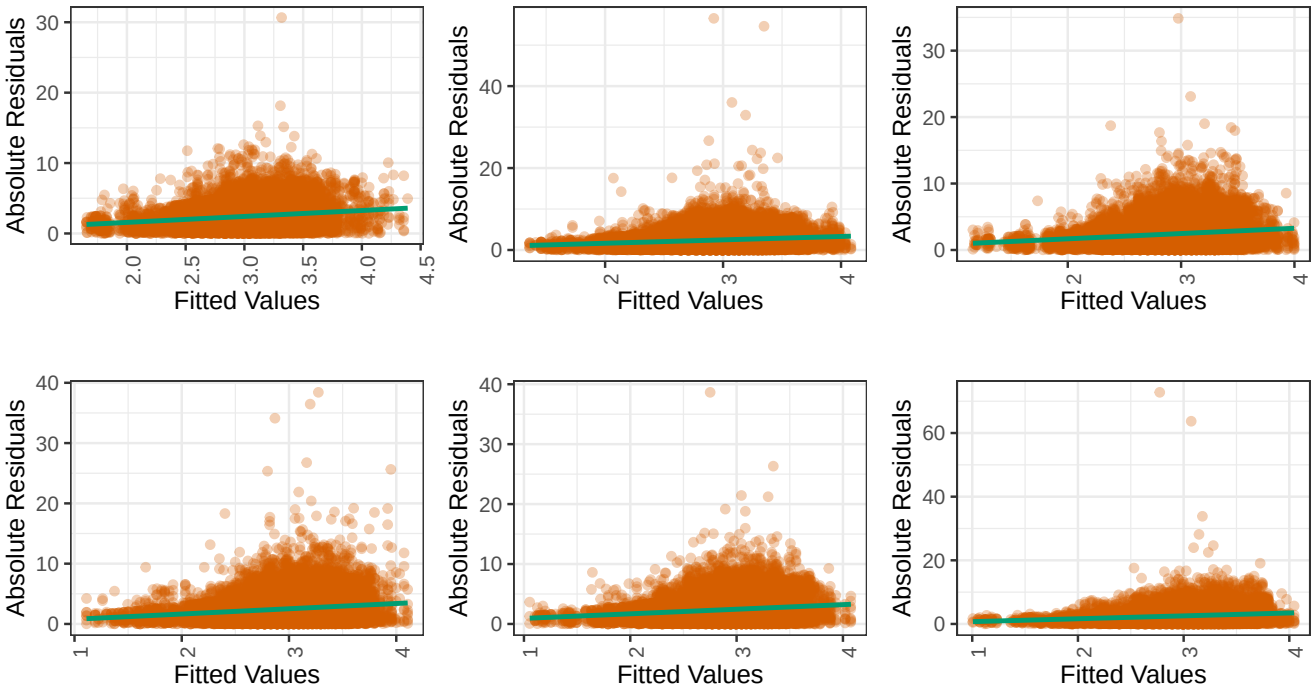
- The residual plot shows the differences between the observed and predicted values by the linear model. A good fit would result in residuals being randomly scattered around zero, without any discernible pattern.
- The clustering of residuals around zero for most fitted values suggests that the linear model is reasonably well-fitted, though some outliers are present, indicating potential over- or underestimations by the model at certain points.

  Low R-squared values are obtained for shorter time periods when linear models are applied to restaurant visitation data, showing a poor fit and implying that a straightforward linear method would not be appropriate for capturing the intricate visitor behavior patterns. The model's fit gets slightly better over a longer time frame, but it still can't account for a sizable amount of the variability. This indicates the need for a more complex model that can take seasonal impacts, cyclic trends, and other exogenous factors impacting restaurant visits into account.

Loess Model Residuals



GLM



- glm(formula = visitors ~ wday * holiday_flg + month + air_genre_name + air_area_name + latitude + longitude, family = poisson(), data = training_data)
- Interpretation
  - wday.Q has a high z-value and low p-value, indicating a strong linear effect of weekdays on visitor counts.

- holiday_flg1 has a positive coefficient with high significance, implying holidays typically see more visitors.
- latitude and longitude: The negative coefficients may suggest a specific geographic trend, such as more visitors in certain areas.

ARIMA