Leonid Rempel

ADEC7460.02 Spring 2021 Predictive Analytics/Forecasting

Midterm Kaggle Project

**Problem:**

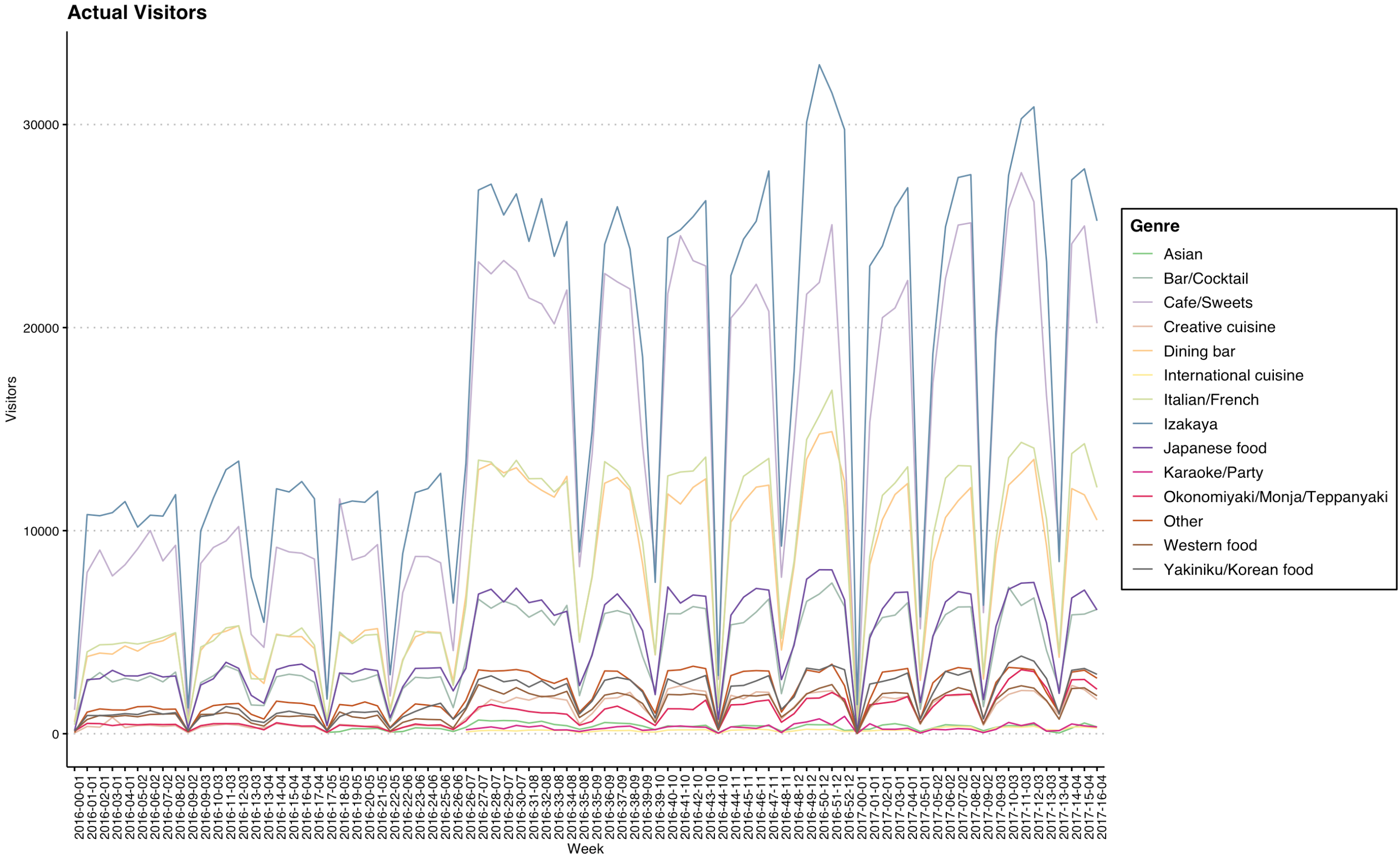
The problem posed by the challenge is to predict the total number of visitors to different restaurants in Japan - owned by Recruit Holdings - for future dates.

**Significance:**

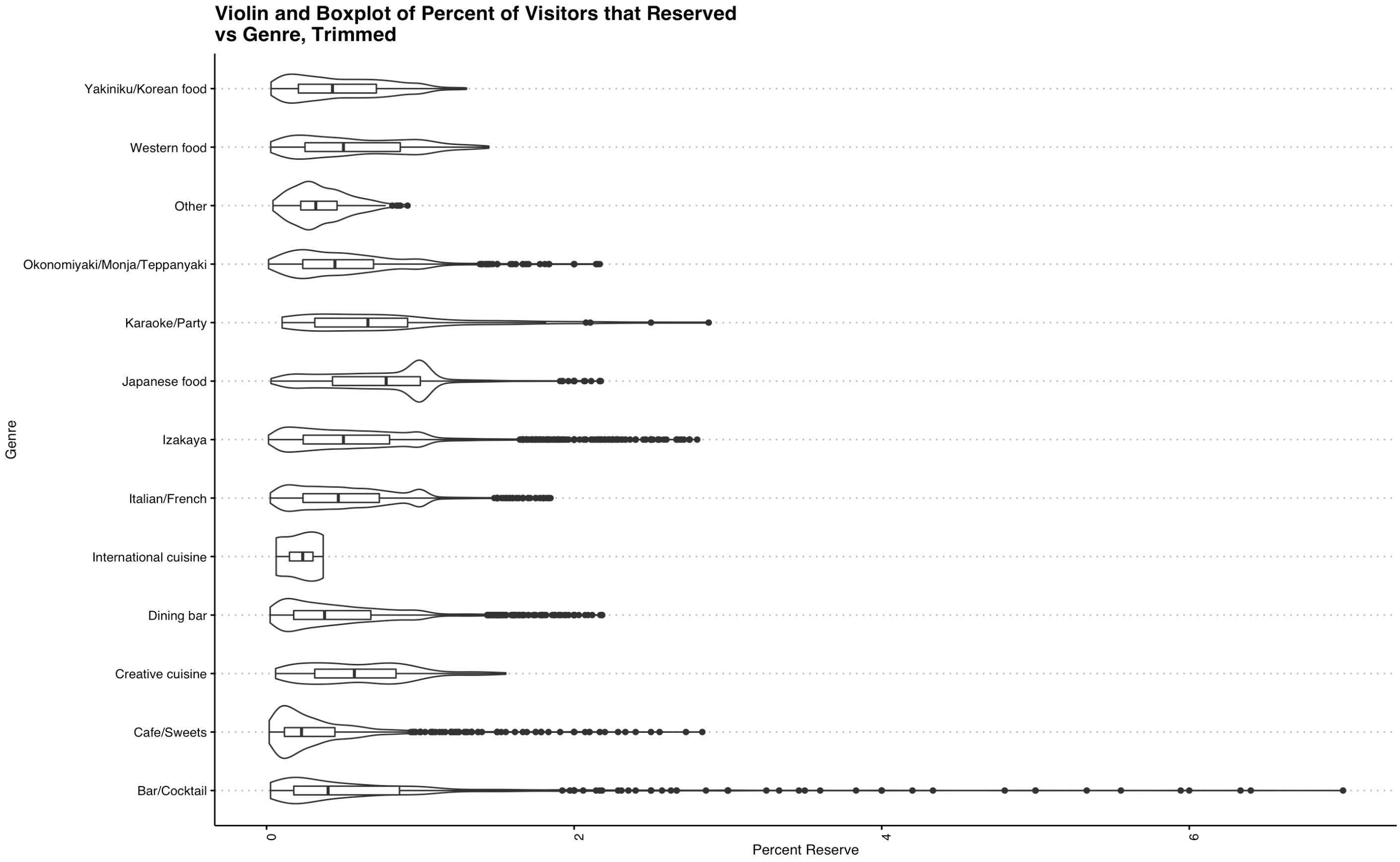
This challenge asks us to forecast panel data across a multitude of different genres, locations, and of course, time periods. For starters, if a restaurant knows how many customers to expect on a given day, then they will be able to properly stock supplies and schedule labor. There is a good amount of unpredictability and potentially a lot at stake for the company. If certain shops will not be able to meet expectations, then the company can know to adjust its strategy. Furthermore, this case study allows us to practice forecasting on a smaller scenario in preparation for forecasts of larger companies with more stores.

**Data:**

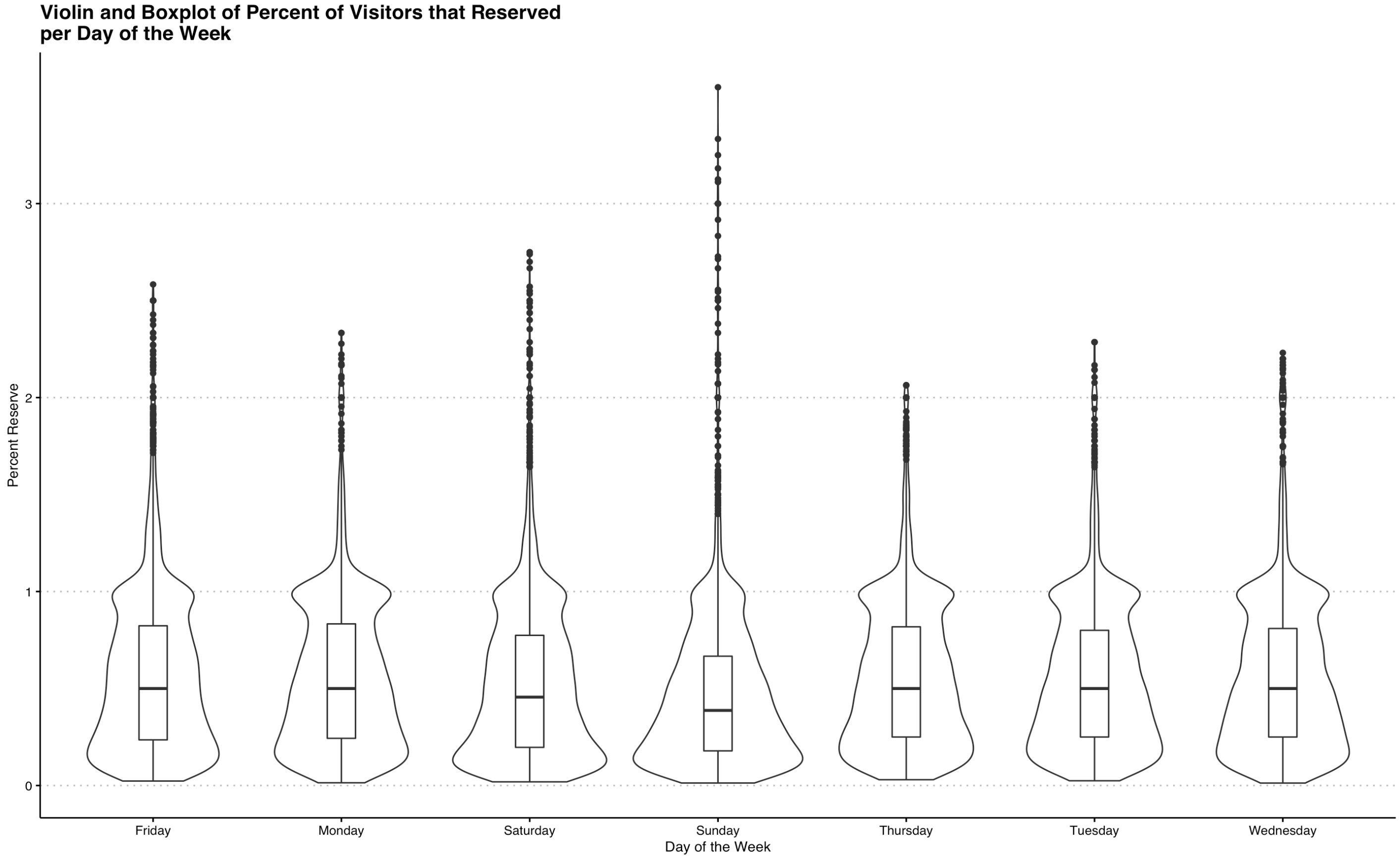
One useful breakdown of the data is by genre. Below is the number of visitors over time by genre.



Much of the numerical visitors comes from the Japanese bars, “Izakayas” followed by Café/Sweets, Dining bars, and Italian/French Cuisine. This is largely due to the fact that Izakayas are the most common types of restaurants that are associated with the company, in addition to Cafés, followed by Dining bars. For these three, reservations may occur, but they are not the bulk of the visits.

****

After removing extreme outliers, we can see from the graph above that there are much more reservations made than there are actual visitors – especially for the Bar/Cocktails, Izakayas, and Café/Sweets. For others, the amount of reservations made tracks better with the total number of customers that end up coming. It seems that people are somewhat fickle about their restaurant plans.

****

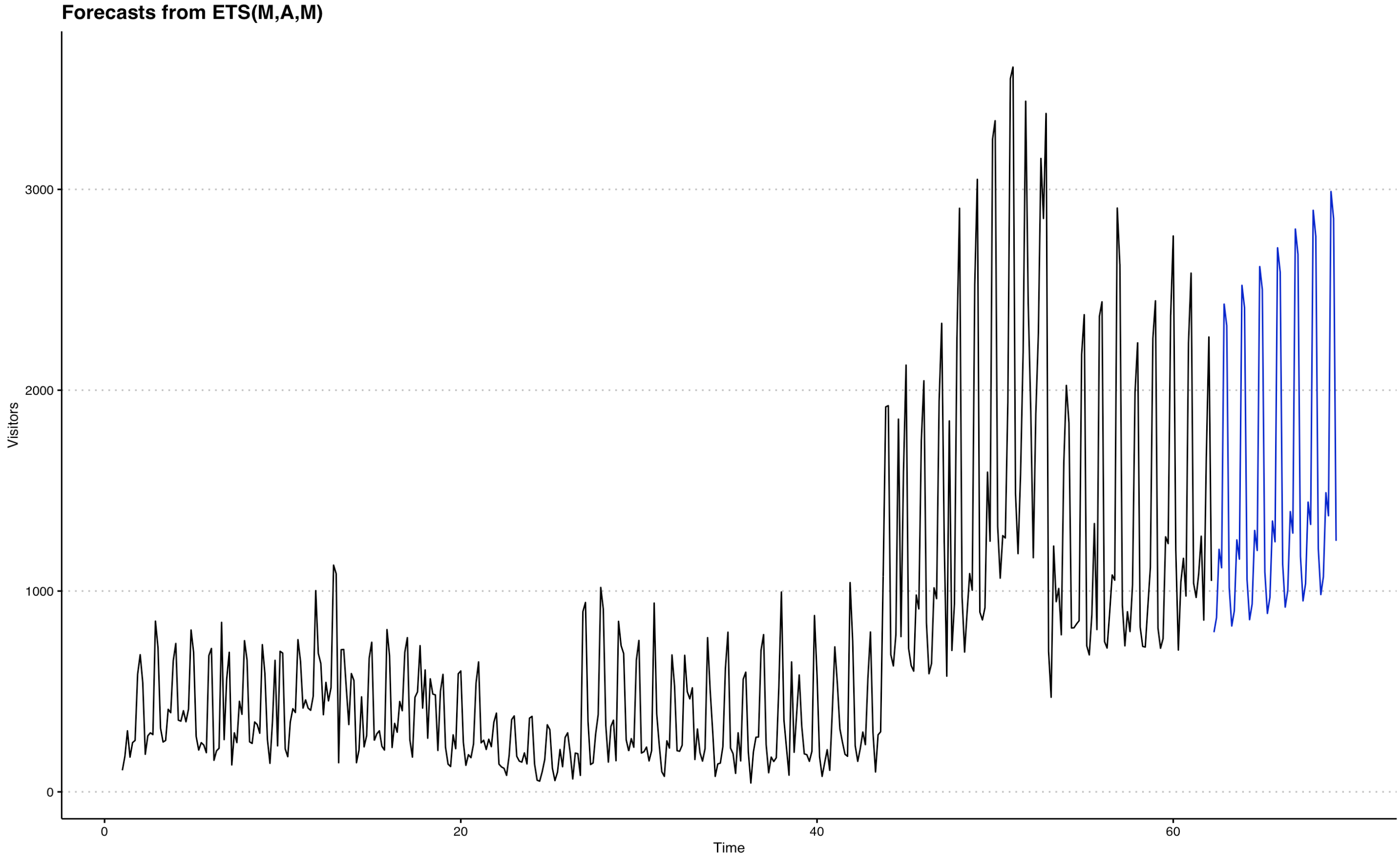
Lastly, it can be said that the ratio of reservations to visitors remains rather steady throughout the week, even though much less people end up coming on Saturdays and Sundays.

A lot of the cleaning to create these graphs came from merging together dataframes on reservations, information about those stores, and aggregating these so that for each store, there was one number that reported the total number of reservations made and the total number of visitors. After that, **the modeling that takes place is broken down by genre**, because it seems from the graph that it is the source of a lot of variation.

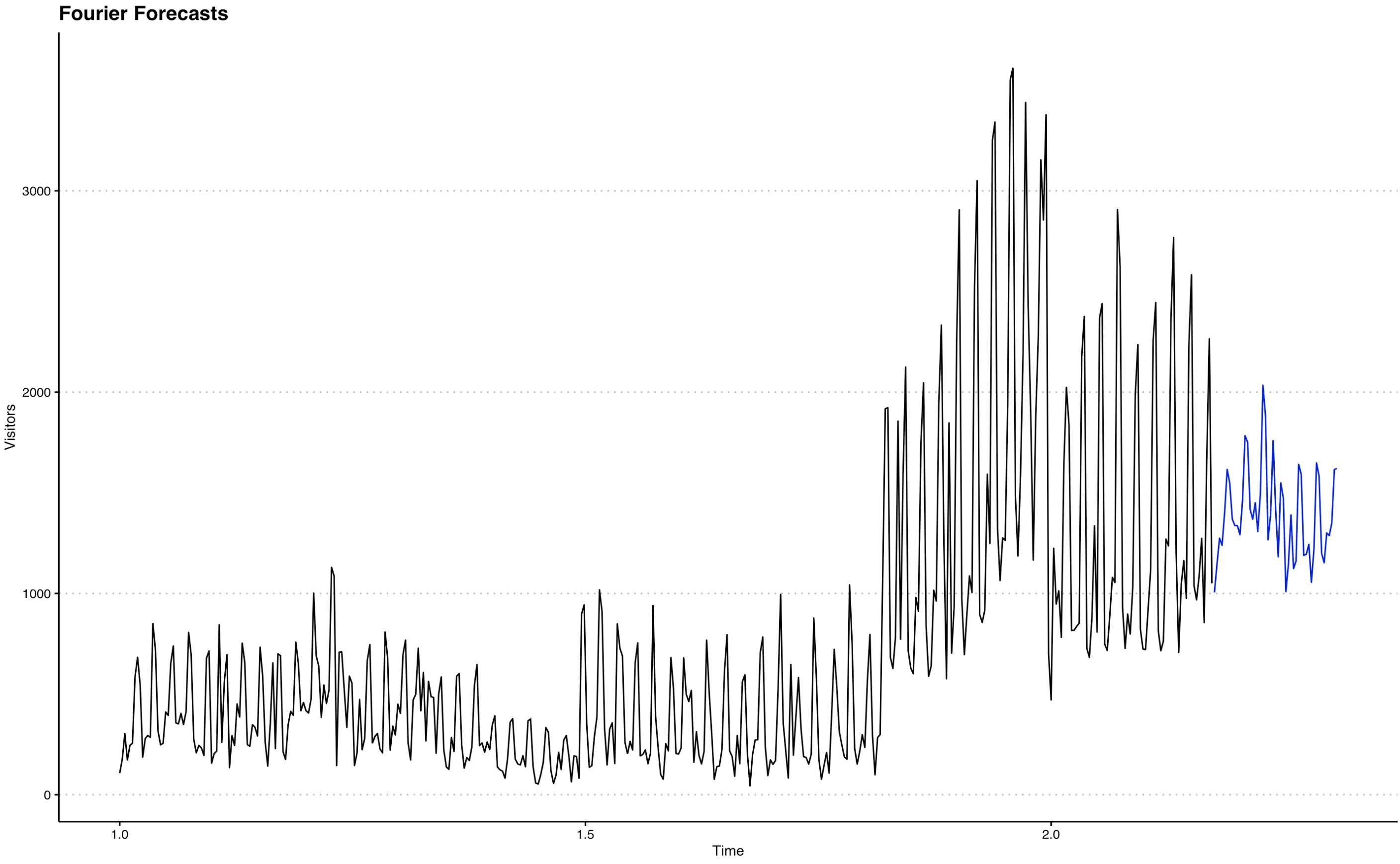
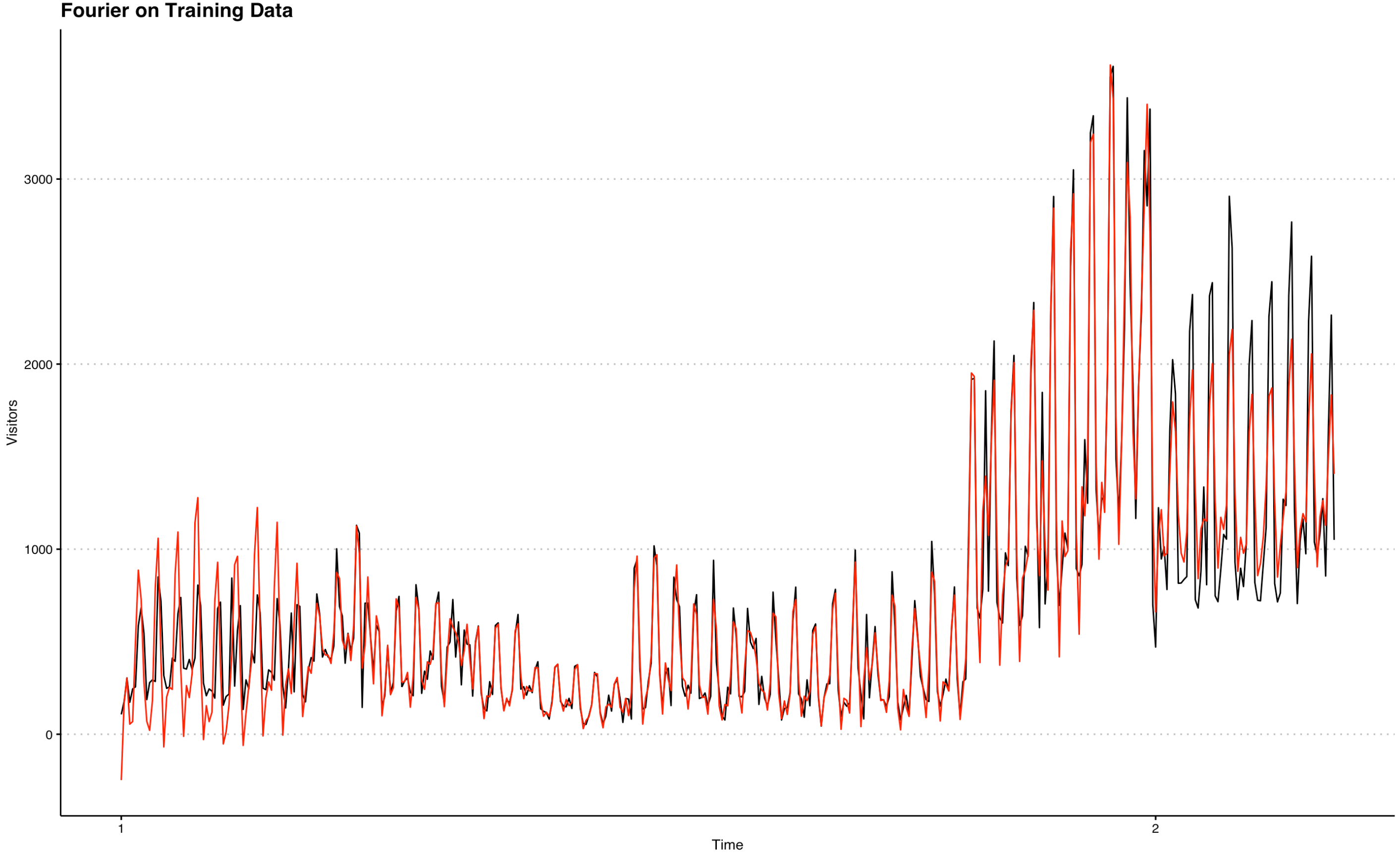
**Literature:**

One of the go to modeling techniques for this problem, a good starting point, would be an exponential smoothing, or ETS model. Such models are commonly used to forecast solar irradiance (Yang et al. 2015), but more applicable to the problem at hand, they have been used in short-term planning and management for a casino buffet restaurant (Hu et al. 2004). Another tool that is explored, and suggested by Hydman for daily forecasts, is the Fourier forecast. This has, more recently, been used in a comparison study of different modeling techniques for forecasting foot traffic (Abrishami and Kumar, 2018). Along the lines of detecting movement, the Fourier forecast has also been used to predict wider mobility patterns based off of social media (Yuan, Q et al. 2017). Lastly, even linear models get to play a role in forecasting as can be seen in a literature survey by Lasek, A. et al. (2015).

**Types of Models, Formulation, Performance:**

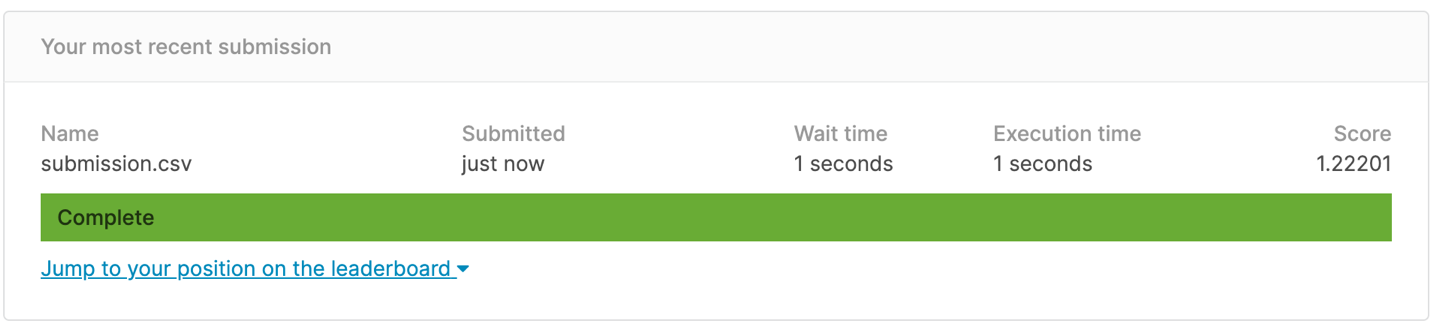


Featured above was my model of choice for forecasting, the ets, which usually came out to include a multiplicative error term, additive trend, and a multiplicative seasonality throughout the different genre breakdowns (above is the ETS forecast of the Izakaya). Coming closely was the Fourier model, which is featured below (from the Izakya) with the best model fit on the training data on the left and the forecasts on the right.

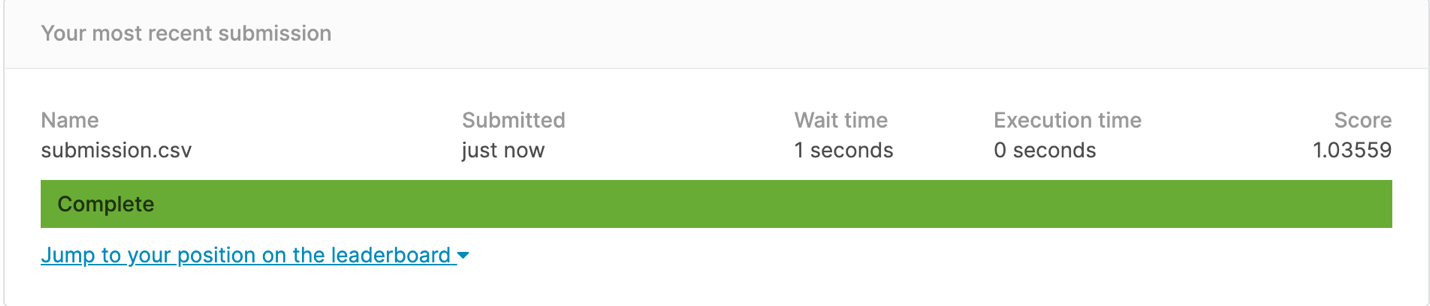


The reason for this is that the Fourier overfits the data. I used a 90/10 split, since this usually resulted in a similar number of forecasting periods as is asked in the challenge. The Fourier had a better MASE measure than the ETS on the validation set, but this is only good if we expect the out of sample data to be similar to the in sample data. The forecasting period ahead of us is Japan’s Golden Week, so I did not expect the future to look too similar to what the Fourier had already seen. Thus, the ETS was chosen because it had a much lower RMSE across the different genres, and this would be a better indicator of future performance.

After breaking down the visitor data by forecast and fitting an ets model, I ran into some issues which will be discussed in limitations. Owing to time and to some extent, knowledge constraints, I opted to assign the different genre stores the average value of visitors expected over the forecasting period. That is, once the ETS model forecasted the aggregate visitors by genre, I divided each of those forecasts for the 39 days by the amount of stores by genre. I also fitted an ETS model for each individual store and used it to forecast 39 periods out.



The above readout is for the forecasts that were done by dividing up the forecasts by genre.



The above readout is for the forecasts that were done by running an ETS model on each individual store. This latter method performed better, although still below benchmark.

**Limitations, Future Work, and Learning**

A major limitation was time. If there was more time, I would’ve explored how to forecast panel data more effectively, I know that the average of the aggregate forecast from an ETS would result in a less than ideal score by the way that the scoring metric is calculated. If we were more interested in how our stores will do by genre, it would’ve done much better. As a result, the average breakdown by genre was not enough and looking at individual stores did better. This is likely because I did not, and wasn’t quite sure how to, incorporate location into the mix. If I could go back, I might’ve even done a breakdown by genre and location to get the averages for places, although this would still have probably performed poorly.

I wanted to use linear models but I kept running into the issue of the data being a panel, and not a cross section or even a simple time series. The first thought with the exploration of the data by genre in aggregate was to somehow forecast reservations and then actual visitors, but I couldn’t quite wrap my head around how to do it. This is why I stuck to the more stochastic, and not relationship dependence ETS and explored Fourier models. They meant that I didn’t necessarily have to worry about the different interactions and the additional predictor creation, that I would otherwise do if I had more time to explore and read up on how others went about it.

I will be revisiting this challenge in the future when I figure out better techniques for predictor creation, a better way to code what I’m thinking about (lots of time goes into even phrasing a question correctly let alone exploring the answer), and a better way to incorporate meaningful relationships between the features into more advanced models.

Bibliography:

Abrishami, S. and Kumar, P., 2018, December. Using real-world store data for foot traffic forecasting. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 1885-1890). IEEE.

Hu, C., Chen, M. and McCain, S.L.C., 2004. Forecasting in short-term planning and management for a casino buffet restaurant. *Journal of Travel & Tourism Marketing*, *16*(2-3), pp.79-98.

Yang, D., Sharma, V., Ye, Z., Lim, L.I., Zhao, L. and Aryaputera, A.W., 2015. Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. *Energy*, *81*, pp.111-119.

Yuan, Q., Zhang, W., Zhang, C., Geng, X., Cong, G. and Han, J., 2017, February. PRED: Periodic region detection for mobility modeling of social media users. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 263-272).

Lasek, A., Cercone, N. and Saunders, J., 2016. Restaurant sales and customer demand forecasting: Literature survey and categorization of methods. *Smart City 360°*, pp.479-491.