OPIM326 Project

G2 Group 1

10 November, 2019

Contents

1 Problem Overview

We decided to embark on an Exploratory Data Analysis project using the dataset and background info provided by Recruit Restaurant Visitor Forecasting. We scoped the project according to the following problem statements:

1. To understand the significant factors that affect the number of customers/day using visualization techniques (**Descriptive Analytics**)
2. To estimate the number of customers/day by accounting for factors such as location, cuisine, reservations, time period and weather data (**Predictive Analytics**)
3. To discover heuristic guidelines to aid new restaurant owners in decision making

This report will be useful for individuals who seek to understand more about the analytics methodology used to analyse the Japan F&B industry. Specifically, we have narrowed down two group of user profiles who might extract significant value from this report:

1) Operations Manager operating a large restaurant chain with many outlets across Japan, who wants to know more about the demand factors that affect daily visitor flow, so much so that he is well informed to plan ahead in terms of supply planning or new outlet planning
2) Aspiring Entrepreneur who wish to understand more about the Japan F&B industry landscape so that he can make better decisions in terms of the key consideration factors (cuisine genre, location, store size, reservation policy etc.)

1.1 Data Set

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): similar to Yelp" (search and reserve) and "AirREGI / Restaurant Board (air): similar to Square" (reservation control and cash register). The training data is based on the time range of Jan 2016 - most of Apr 2017, while the test set includes the last week of Apr plus May 2017.

These are the individual files:

- **air_visit_data.csv**: historical visit data for the *air* restaurants. This is essentially the main training data set.
- **air_reserve.csv** / **hpg_reserve.csv**: reservations made through the *air* / *hpg* systems.
- **air_store_info.csv** / **hpg_store_info.csv**: details about the *air* / *hpg* restaurants including genre and location.
- **store_id_relation.csv**: connects the *air* and *hpg* ids
- **date_info.csv**: essentially flags the Japanese holidays.

To better illustrate the effectiveness of our forecasting techniques, we will be using **air_visit_data.csv** data from Jan 2016 - Mar 2017 as training set and Mar - Apr 2017 as test set. The rest of the csv files will be used to for feature engineering.

2 Data Preprocessing

2.1 Loading Data

```
air_visits <- read.csv('air_visit_data.csv')
air_reserve <- read.csv('air_reserve.csv')
hpg_reserve <- read.csv('hpg_reserve.csv')
air_store <- read.csv('air_store_info.csv')
hpg_store <- read.csv('hpg_store_info.csv')
holidays <- read.csv('date_info.csv')
store_ids <- read.csv('store_id_relation.csv')
```

2.2 Summary of air visitor data

```
##                air_store_id         visit_date        visitors
##  air_5c817ef28f236bdf:   477   2017-03-17:   799   Min.   :  1.00
##  air_36bcf77d3382d36e:   476   2017-03-03:   791   1st Qu.:  9.00
##  air_a083834e7ffe187e:   476   2017-04-14:   791   Median : 17.00
##  air_d97dabf7aae60da5:   476   2017-03-24:   789   Mean   : 20.97
##  air_232dcee6f7c51d37:   475   2017-03-31:   786   3rd Qu.: 29.00
##  air_60a7057184ec7ec7:   475   2017-04-21:   784   Max.   :877.00
##  (Other)             :249253   (Other)   :247368
```

```
## Observations: 252,108
## Variables: 3
## $ air_store_id <fct> air_ba937bf13d40fb24, air_ba937bf13d40fb24, air_b...
## $ visit_date   <fct> 2016-01-13, 2016-01-14, 2016-01-15, 2016-01-16, 2...
## $ visitors     <int> 25, 32, 29, 22, 6, 9, 31, 21, 18, 26, 21, 11, 24,...
```

Number of unique *air* restaurants: 829

Mean number of visitors/day in *air* restaurants: 20.9737612

3 Data Visualization

3.1 Visitor Time Series

We can start by first looking at the visitor data in terms of overall daily visitor count, daily visitor distribution, monthly distribution, and daily distribution. .

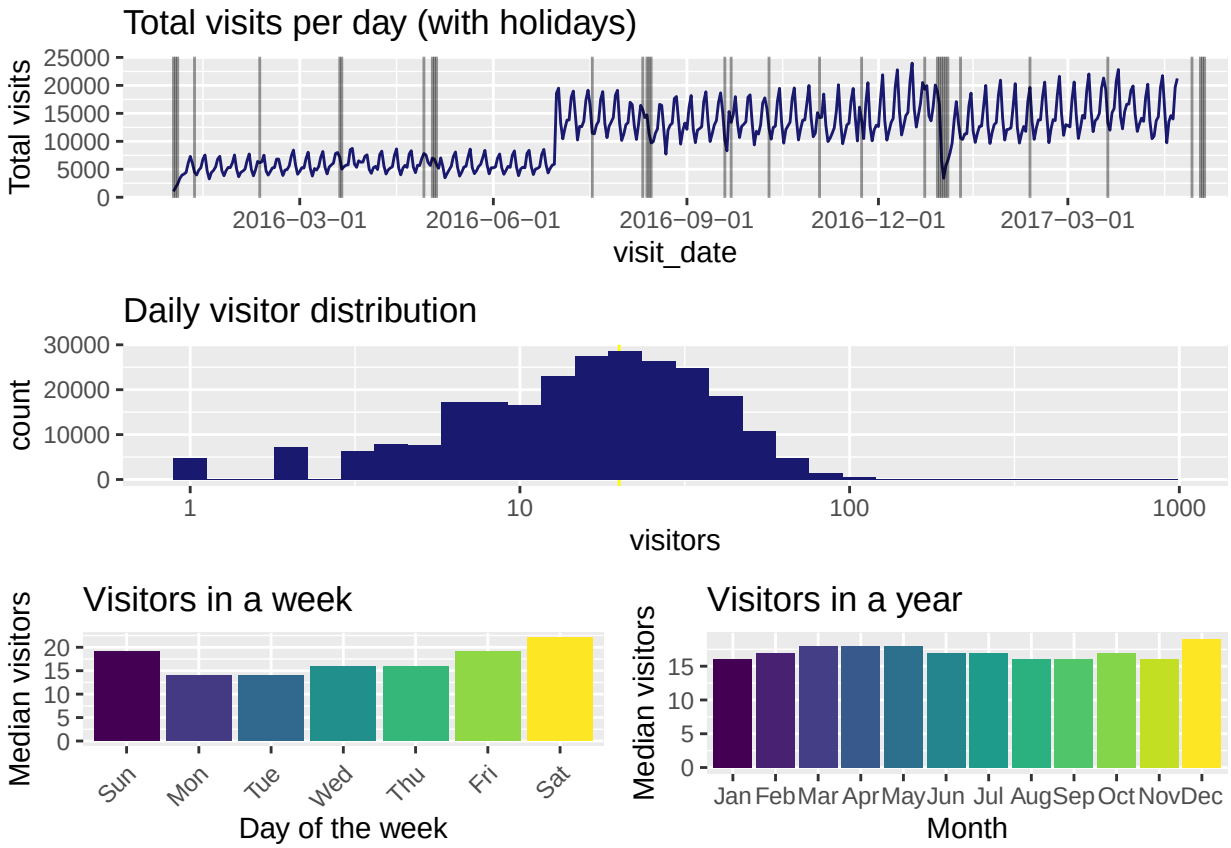


Figure 1: Air Restaurants Visitor Information

From the figures above, we can make the following observations:

- Daily number of visitors assumes a lognormal distribution, which is good for linear regression methods (requiring a simple transformation of log(n)).
- There seem to be a huge spike in aggregate visitor count on July 2016, which we assume to be due to an update in the air website database to include more restaurants.
- There is a large drop in aggretate visitor count at the start of 2017, so perhaps there are some missing data in the data set. We will explore in the later section if it affects our prediction accuracy.
- The visitor count is higher on Fridays, Saturdays and Sundays, which is not surprising for the F&B industry.
- There seem to be a drop in visitor count from July till November, before reaching a peak in December. We can hypothesize that it has something to do with the weather conditions, or simply due to the addition of smaller stores into the website database.

Now, let's look at the forecast period (22nd March to 22nd April) in 2016 to see if we have any anomolies in the data.

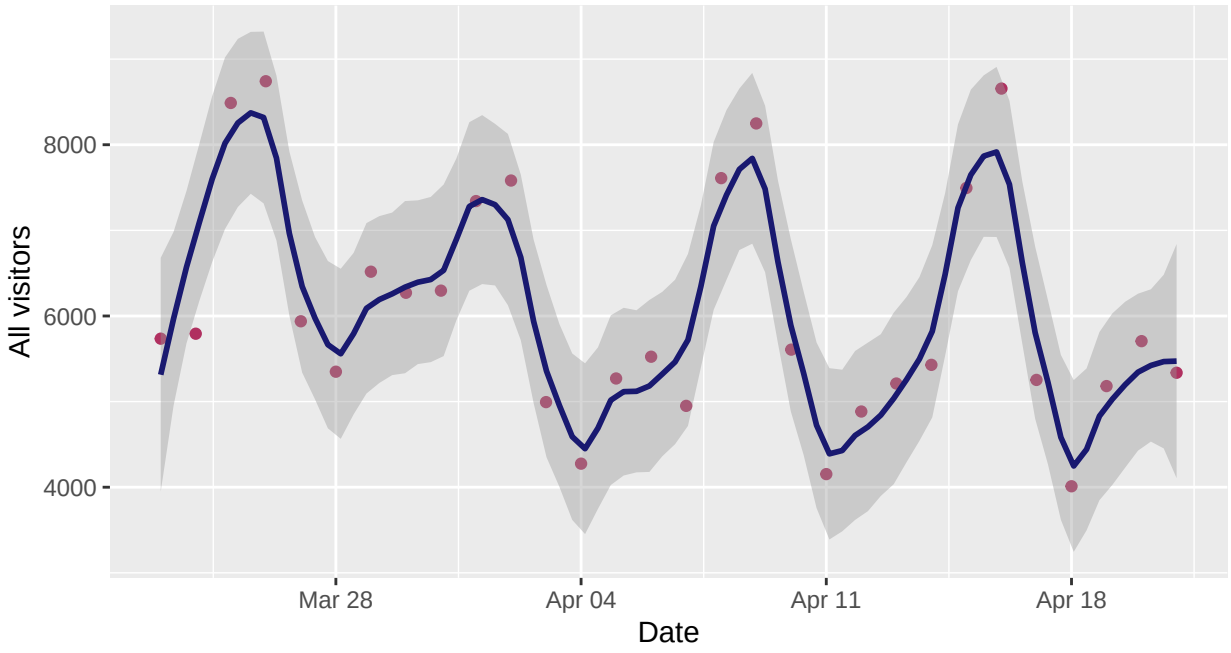


Figure 2: 22 Mar - 22 Apr Visitor Data

Here, we see that the actual data (red dots) is quite well behaved with a regular cyclical trend(blue line) and a slight decreasing trend.

3.2 Restaurant Overview

We can first look at the restaurant count split by genre and area to see if the data set is skewed towards a particular catagory.

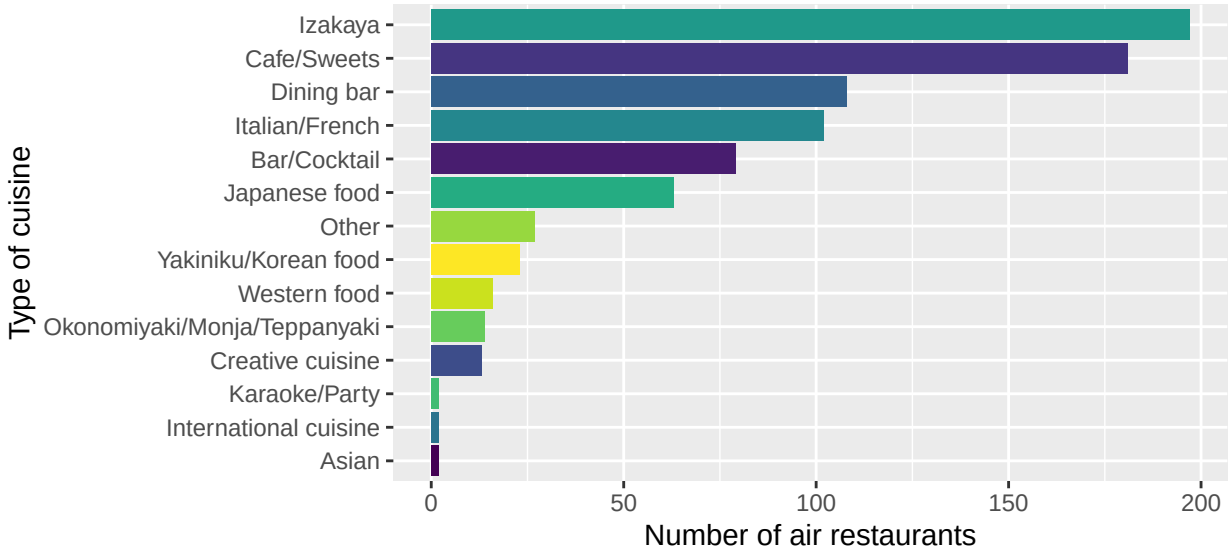


Figure 3: Number of air restaurants by cuisine genre

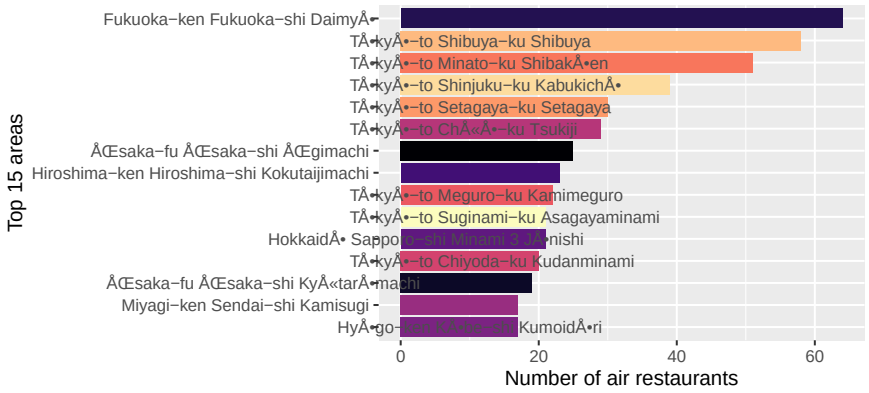


Figure 4: Number of air restaurants by area

From the figures above, we can make the following observations:

- Izakaya and Cafe seems to be the two most popular restaurants in Japan, followed by dining bar, Italian/French and Japanese food(assuming no selection bias on AirREGI website).
- Most of the *air* restaurants are located in the Tokyo prefecture, followed by Fukuoka, Osaka, Hiroshima and Hokkaido.

4 Time-Series Forecasting

Now let's get our hands dirty and do some forecasting. Our general methodology will be as such:

1) Decompose date column into useful date-time features. You can see these features like individual variables.
2) Extract additional useful features such as holidays, average historical visits, and number of reservations.
3) Standard regression/classification methods for daily visitor prediction
4) Select a few store IDs for validation test. We selected *air_id = "air_ba937bf13d40fb24"*, *air_id = "air_07bb665f9cdfbdfb"* and *air_id = air_1c0b150f9e696a5f"* to demonstrate the effectiveness of our model under different scenarios.

4.1 Feature Engineering

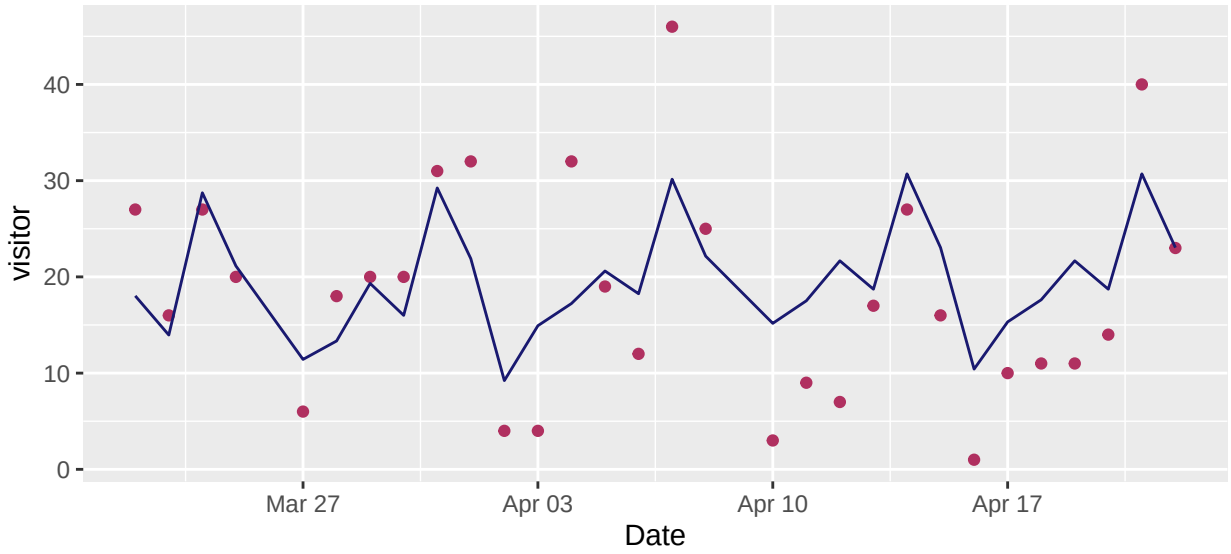
Given that we are dealing with time series data, we broke down the date column into day of the week, week of the month, and month of the year, so that we can do prediction using standard linear regression. In addition, we created two new features:

1. **vis_mean_dow** - mean no. of visitors to the restaurant on that particular day of the week
2. **res_for_date** - no. of reservation scheduled

visit_date	air_store_id	dow	visitors	month	woy	vis_mean_dow	res_for_date	holiday_flg
2016-01-01	air_b2d8bc9c88b85f96	5	16	1	53	24.32308	0	1
2016-01-01	air_d0a1e69685259c92	5	7	1	53	23.31884	0	1
2016-01-01	air_506fe758114df773	5	4	1	53	20.17391	0	1

4.2 Linear Regression & Classification Trees

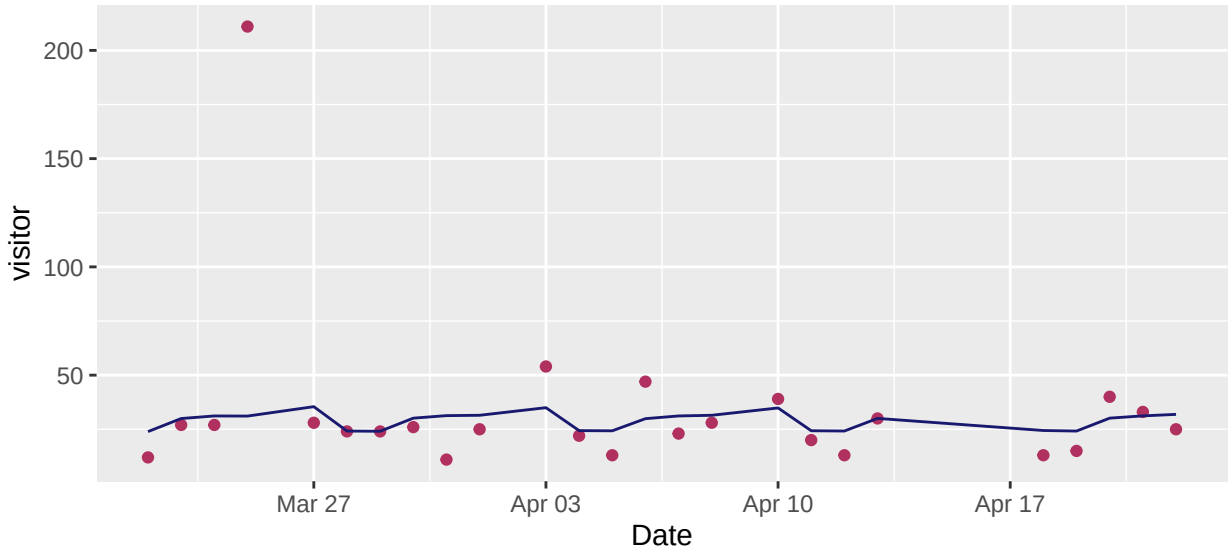
4.2.1 Case 1: store *air_ba937bf13d40fb24*



- Mean Absolute Error: 6.3959602
- Average predicted daily visitors: 19.6648248
- Average actual daily visitors: 18.2666667

In this case, we can see that time series forecasting of daily visitors is reasonably accurate for store owners to plan for supplies, especially for F&B stores with prerishable raw materials.

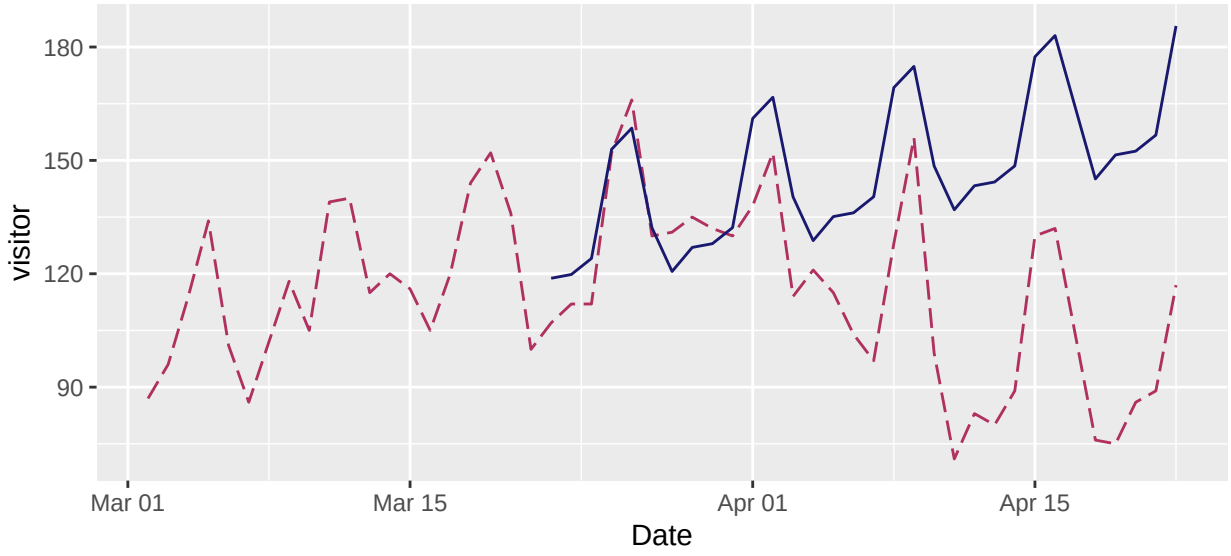
4.2.2 Case 2: store *air_07bb665f9cdfbdfb*



- Mean Absolute Error: 14.3104168
- Average predicted daily visitors: 28.9605873
- Average actual daily visitors: 33.2

We can see that while the overall trend is reasonably accurate, the forecasted values poorly accounts for anomalies. In cases like these, it would be advisable to exercise sound business judgement to preempt days with sudden surge in visitor count.

4.2.3 Case 3: store air_1c0b150f9e696a5f



- Mean Absolute Error: 33.5787441
- Average predicted daily visitors: 146.4558269
- Average actual daily visitors: 114.8064516

In this case, while the general peaks and troughs appear to be accurate, our model failed to account for the overall decreasing trend. Upon further inspection it appears that this is because the earliests recorded visit date is 2017-03-02, merely a month before the dates that we are forecasting. This could be due to the restaurant being new in the system, thus the model failed to capture a trends that could have been caught with more data. Alternatively, it could also be due to the restaurant being recently opened, such that the increasing trend in visitors in the first month was due to initial interest in the new restaurant. Our model then captured and wrongly extrapolated the trend, as the initial interest began to wane as time passes.

4.3 Technical Remark

Based on the cases presented above, we can see that while general prediction methods work well for stores with stable seasonality and trend factors, they are nonetheless inadequate for unique cases like case 2 & 3. This small exercise aptly ellucidates the importance of well-engineered features, which have tremendous impact on the predictive power of a model.

5 Heuristic Guidelines

5.1 Supply Planning

From the cases above, we can see that while time series forecasting works very well for stores with stable trend and seasonality, they fit worse for stores with very high variance. Nevertheless, for stores with visitor data similar to case 1, time series forecasting is a very useful tool to better manage store supplies, especially in the F&B industry where a large portion of raw ingredients are perishables.

5.2 Staff Scheduling

When it comes to staff scheduling, a key factor to analyze on top of visitor data would be worker productivity. Assuming that worker productivity is homogeneous across all stores, we can make good decisions based on the results shared in this report. However, to truly optimize the number of part time and full time workers to hire for the entire year, we would recommend the decision makers to also consider the average work rate of their workers to achieve better service levels and customer experience.

6 Model Limitations and Future Plans

6.1 Diagnostic & Prescriptive Analytics

In this report, our analysis has been restricted mostly to data visualization (**Descriptive Analytics**) and forecasting (**Predictive Analytics**), with some preliminary discussion on decision making. However, to fully uncover the potential of analytics, there remain much work to be done in terms of validifying hypothesis(**Disgonistic Analytics**) and providing actionable recommendations(**Prescriptive Analytics**). With this in mind, we highly encourage further exploration and analysis through refining problem statements, collecting more quality data, and running randomized controlled trials (RCTs) to gain a deeper understanding of the F&B industry in Japan and the critical success factors required to run a fully automated and optimized Operations Model.

6.2 Alternative Time-Series Forecasting Methods

Note that our method for analysing time-series/panel data isn't the most sophisticated as we are essentially pooling the data to do analysis, ignoring factors like time lags(autocorrelation), heterogeneity, endogeneity and other issues related to time-series and panel data. Hence, we have list out some tools that we would like to explore in the future as our knowledge of data science and analytics grow.

- **ARIMA**: Known as "Auto-Regressive Integrated Moving Average" Model, ARIMA addresses some of the inherent problems with time series data by including a set of parameters (p,d,q), which accounts for the autocorrelation and non-stationarity within a time series data set. To further account for seasonality factors, Seasonal ARIMA (SARIMA) can be used instead to increase model fit.
- **Prophet**: Prophet is a forecasting model designed to deal with multiple seasonalities, developed by Facebook's core Data Science team. The model assumes that the time series data can be decomposed into trend factor, seasonality factor and holiday factor. THe model fitting is framed as a curve-fitting exercise, so it does not explicitly take into account for temporal dependence structure in the data.

In this project, we think that Prophet would work quite well, given that there are clear seasonality and holiday factors within the data (hence non-stationary).

6.3 Additional Features

We would also like to note some features that we can include into our model to increase its predictive power.

- **Transportation Data** - We hypothesize that restaurants nearer to public modes of transportation (buses & trains) will be more assessible and therefore have more daily visitors. We believe that analysing the marginal effect of such variables will be useful for restaurants seeking to open new chains and would like to assess the relative benefits between two locations.
- **Similarity Index** - We hypothesize that restaurants with similar cuisines (e.g. Singaporean & Malaysian cuisine) would have lower average daily visitor count due to the substitution effect. It would be interesting to proof or disproof this hypothesis as it then allows aspiring entrepreneurs to make better decisions on whether to open their store nearer or further away from restaurants with similar offerings.
- **Area "Crowdedness"** - We hypothesize that there is a correlation between the density of restaurants within an area(q) and the average daily number of customers(n), assuming that other confounding factors like food quality, cuisine genre, assessibility etc. are kept constant. We think that such an analysis would definitely benefit aspiring entrepreneurs who want to make a decision on where to open a store, balancing the tradeoff between high traffic and high competition(also high fixed cost).
- **Weather Forecast** - We hypothesize that extreme weather conditions are likely to impede people from visiting restaurants(instead opting for home-cook meal). Hence, we think that adding quality weather data as additional features will boost our time series prediction model in a meaningful way.

7 Credit

Lastly, we would like to credit the following sources for inspiring our research methods in this report:

- Be My Guest - Recruit Restaurant EDA by Heads or Tails
- A Very Extensive Recruit Exploratory Analysis by Troy Walters