MILESTONE 2 - INITIAL FINDINGS

PROJECT: TABLE FOR FOUR

TEAM: WILDCATTS ANALYTICS

CATHERINE TOLLEY
TAMARA WILLIAMS
TOM ALIG
SHEELA RAO

May 13, 2018

TABLE OF CONTENTS

PROBLEM STATEMENT

With the world's ninth-largest population (127 million) (Wikipedia, n.d.) (McGushion, n.d.)Japan has approximately 107 million mobile phone users (Statista, n.d.). Adding in roughly 24 million foreign tourists with their smartphones, these mobile phone users represent a marketing opportunity for the travel, restaurant, and entertainment business. Recruit Holdings Co., Ltd. is a global holding company, headquartered in Japan. For this engagement, we will be targeting the Media and Solutions vertical of Recruit Holdings' (RH) business, focusing on the company's offerings in the restaurant sector.

Recruit Holdings' AirREGI point-of-sale (POS) product currently offers cash register and table management services via iOS tablet and mobile devices for restaurants in Japan. A related Recruit Holdings business, Hot Pepper Gourmet (HPG), a restaurant marketing website and mobile app, provides Japanese diners with a searchable interface to locate restaurants and obtain promotional coupons. The company has targeted these two products for continued growth in 2018. Recruit Holdings has hired WildCATTS Analytics to further its vision of connecting customers with businesses, specifically in the Japan restaurant market. The proposed project will provide new business insights and drive shareholder value.

By tapping into Recruit Holdings breadth of data collection in the Japan restaurant market, WildCATTS Analytics will add predictive modeling to the capabilities of the AirREGI system. Achieving success in the project goals will enable Recruit Holdings to adopt the same capabilities across its multiple product channels worldwide.

PROJECT GOALS

1. Develop a robust predictive model to predict daily restaurant volume up to 30 days in advance.
2. Develop a dashboard for the AirREGI POS solution to deliver predictions in context, enabling restaurant managers to take appropriate actions, including pushing customized promotions to the HPG marketing product.
3. Develop a prototype of a mobile experience for restaurant clients to interact with and potentially respond to volume predictions.
4. Maximize insights from Recruit Holdings' valuable restaurant sector data with meaningful data visualizations.
5. Recommend strategy to monetize prediction-enhanced AirREGI Point of Sale premium-subscription solution.

DESCRIPTION OF THE DATA

The primary source of data for this project is the Kaggle Recruit Restaurant Visitor Forecasting data set. The dataset is comprised of seven files at differing levels of granularity, forming a relatively complex data ecosystem. The files and their relationships are shown below in Figure 1.

Table	Records	Variables
Total Visit Data	252,108	3 plus modeling target visit count
Date Data	517	3
AirREGI Restaurant Info	829	5
HPG Restaurant Info	4,690	5
AirREGI Reservation Data	92,378	4
HPG Reservation Data	2,000,320	4
Store Relation Data	150	2

Table 1

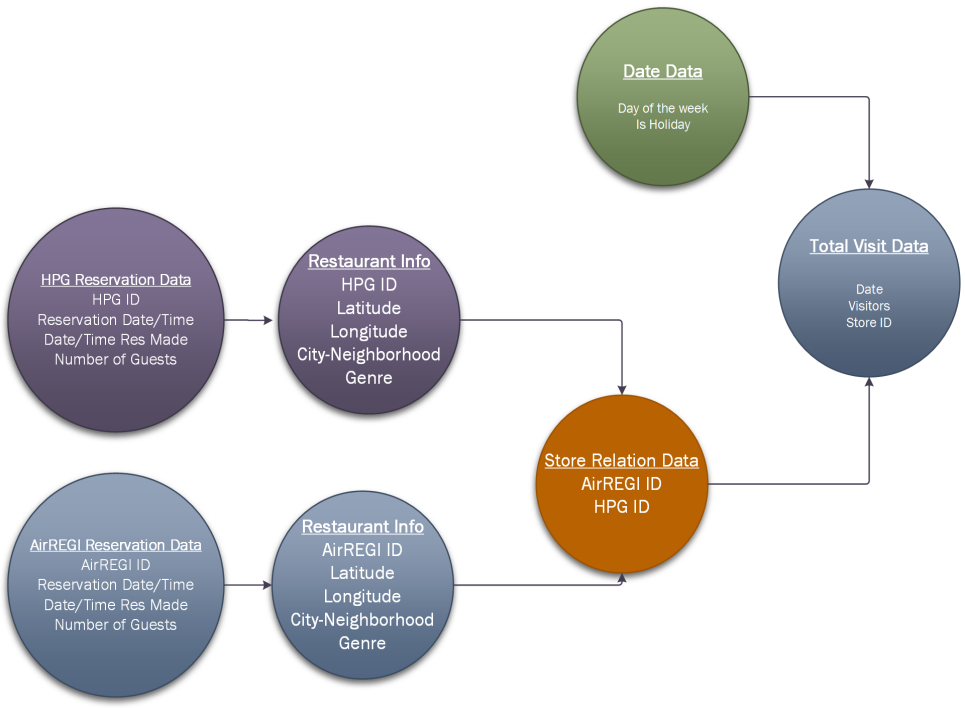

Figure 1 - Visual Representation of Kaggle Data Files

OVERVIEW OF THE DATA

- Approximately 16 months of historical data for the AirREGI restaurants will be used to predict the next 39 days of visitor volume
- Daily visitor counts demonstrate periodic patterns but are less complete for the first 6 months of historical data provided. The predominant period is weekly with the highest counts on Saturdays.
- Many missing daily visitor counts can be attributed to regular weekly open and closed days for each restaurant and to holiday closures.
- Holidays impact restaurant visitor counts, when open.
- Restaurant geographic coordinates are grouped approximately to the labeled geographic area, which has high cardinality.
- Reservations are typically made 24 hours prior to dining and are most often made for the evening meal.
- Reservation data from the AirREGI system are inconsistent during the first year of the data. The usefulness of AirREGI reservations for predictive modeling is questionable.
- Reservation data from the Hot Pepper Gourmet (HPG) web and mobile reservation system are more consistent but contains reservations for only 150 of the 829 AirREGI restaurants targeted. The usefulness of HPG reservation data for predictive modeling is questionable.
- There are outliers in both the daily visitor counts and the reservation data.

Initial investigation into the training data was primarily carried out using R. Some visual exploration was also done in Microsoft PowerQuery/PowerBI and Tableau. The model training data has 252,108 observations of 829 restaurants utilizing the AirREGI POS product, spanning 517 calendar days from January 1, 2016 to April 22, 2017.

AIRREGI RESTAURANT DATA

The AirREGI restaurants are labeled with the restaurant genre, the geographic area, and the latitude and longitude. There are 14 distinct restaurant genre categories (Figure 2), with extremely unequal distribution. The most common genres are Izayaka and cafe/sweets.

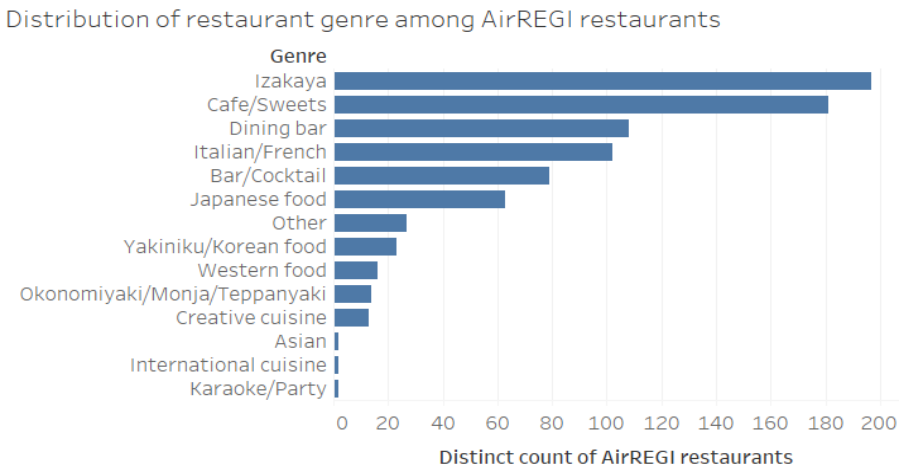

Figure 2 - Restaurant genres

Several restaurants share the same latitude and longitude pairs as a result of intentional obfuscation of individual restaurant locations and identities, resulting in 108 unique coordinate pairs among the 829 restaurants. Restaurants are located across Japan (Figure 3). The area variable demonstrates high cardinality with 103 distinct geographic areas.

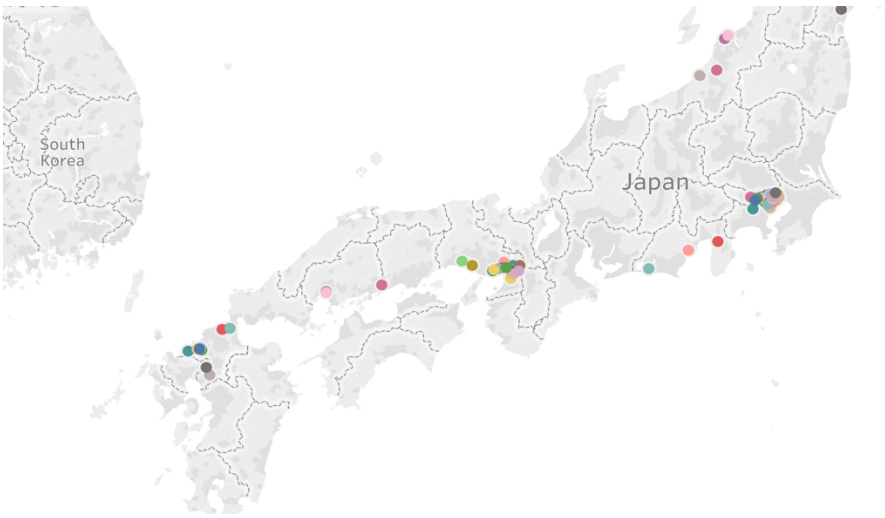

Figure 3 - Map of AirREGI restaurant locations

The distribution of the prediction target, visits per day per restaurant per day, is centered around 20 visits, shown in orange. (Figure 3). There are a few rare outlier days with more than 100 visitors per restaurant. It is difficult to explore outlier daily visits in the absence of hourly visits and restaurant capacity data.

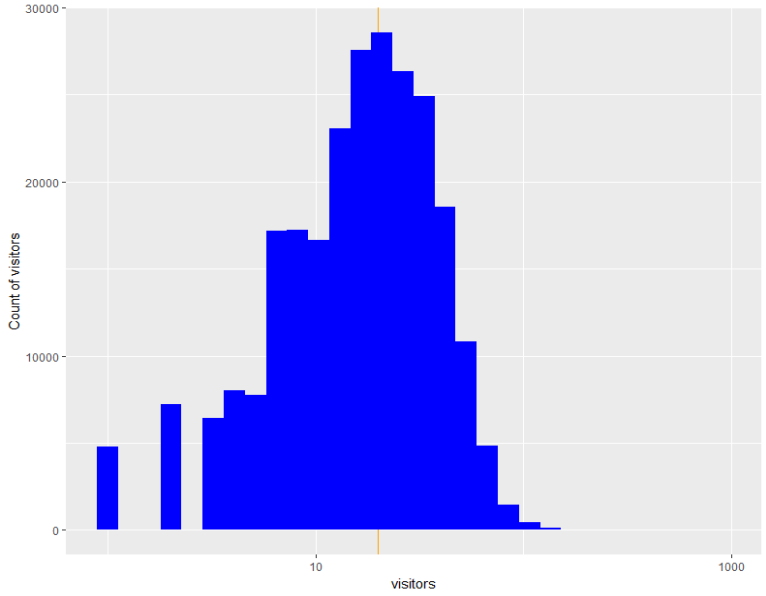

Figure 4 - Distribution of visitors per day

While a weekend holiday has little impact on the visitor numbers, and even decreases them slightly, there is a much more pronounced effect for the weekdays, especially Monday and Tuesday (Figure 5).

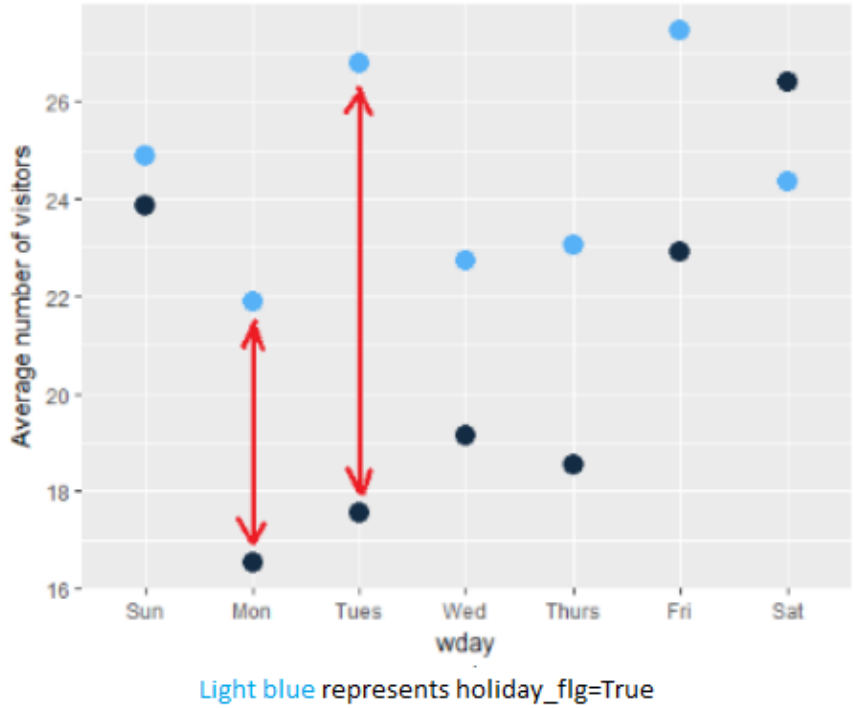

Figure 5 - Impact of holidays on visitor count

Historical visitor totals across all restaurants are presented in Figure 6. There is a periodic pattern that most likely corresponds to a weekly cycle. An outlier date with very few visits is noted on New Year's Day, likely due to a large number of restaurants closed for the holiday. Notice that there is an increase in volume starting July 2016. The plot also demonstrates seasonality and a slight upward trend, with more visits in December and in the most recent months of the data.

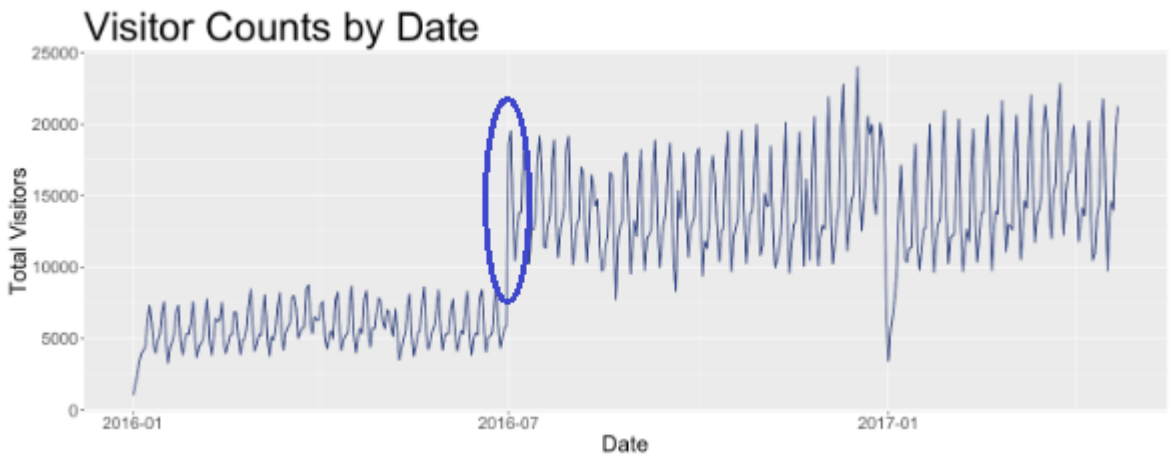

Figure 6 - Plot of Total Number of Visitors Per Date

The sudden increase in visitors in July 2016 is explained by the addition of approximately 400 restaurants to the data. In the upper panel of Figure 7, the dark blue trend shows the number of restaurants with visitor counts captured by date. The grey trend shows the number of restaurants with no visitor count captured by date. The sudden increase in restaurants with visitor count may be explained by the onboarding of new restaurants or increased data capture. Most of the restaurants have visitor counts in the last 10 months of the training data.

The variation in visitor count by date also demonstrates a periodic weekly cycle. Furthermore, there are dips in the number of restaurants with visitor counts associated with holidays, indicated by the holiday flags in the lower panel of Figure 7. This suggests the periodic absence of visitor counts may be due to restaurant closures.

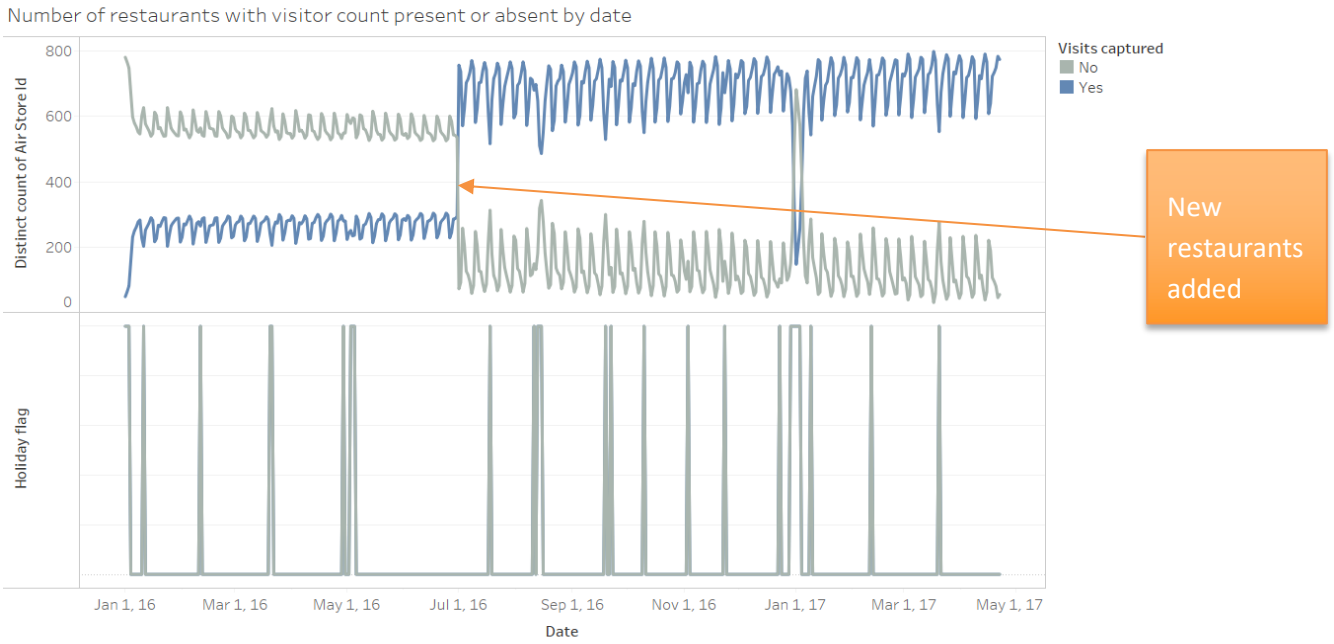

Figure 7 - Number of restaurants with visitor count captured by date, with holiday flags

Upon further investigation, individual restaurants had data for 4 days per week on average, with a bimodal distribution of restaurants with 3-4 days per week and restaurants with 5-6 days per week (Figure 8). This pattern was similar across restaurant genres and prefectures (not shown).

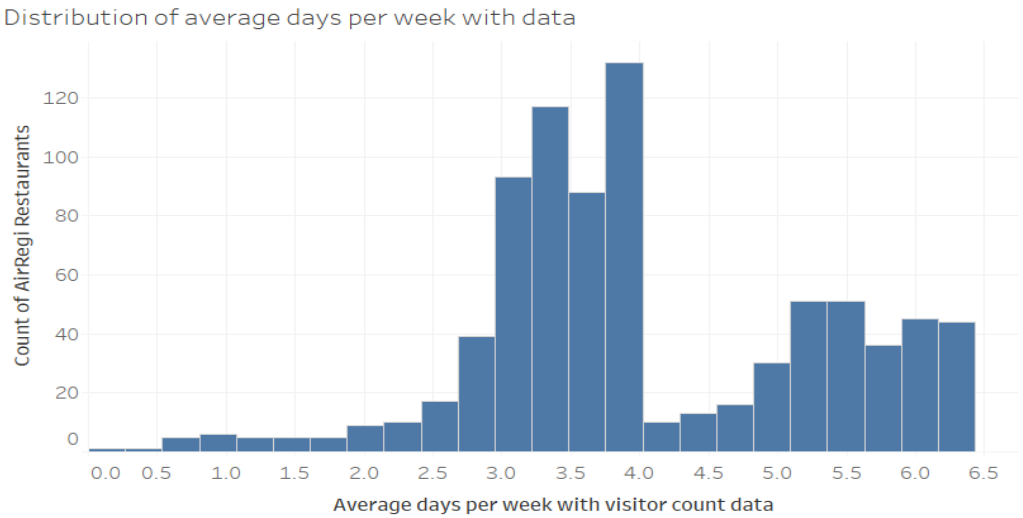

Figure 8 - Restaurant average days per week with visitors

These observations will be taken into account when imputing missing historical visitor counts for individual restaurant-date combinations prior to modeling. The majority of missing dates are likely the result of weekly planned restaurant closures as part of their normal operating schedule.

Holidays account for both missing visitor counts due to restaurant closures and impacts to visitor counts due to changes in diner patterns. The holiday period known as Golden Week: April 29 - May 5 is an important example. The plot in Figure 9 shows visitor counts around this date range as well as a smoothing fit in blue. Notice the blue line trending downward around the end of the first quarter in 2016. This is the negative impact of Golden Week on the restaurant business. Holiday information can be used effectively for modeling special cases.

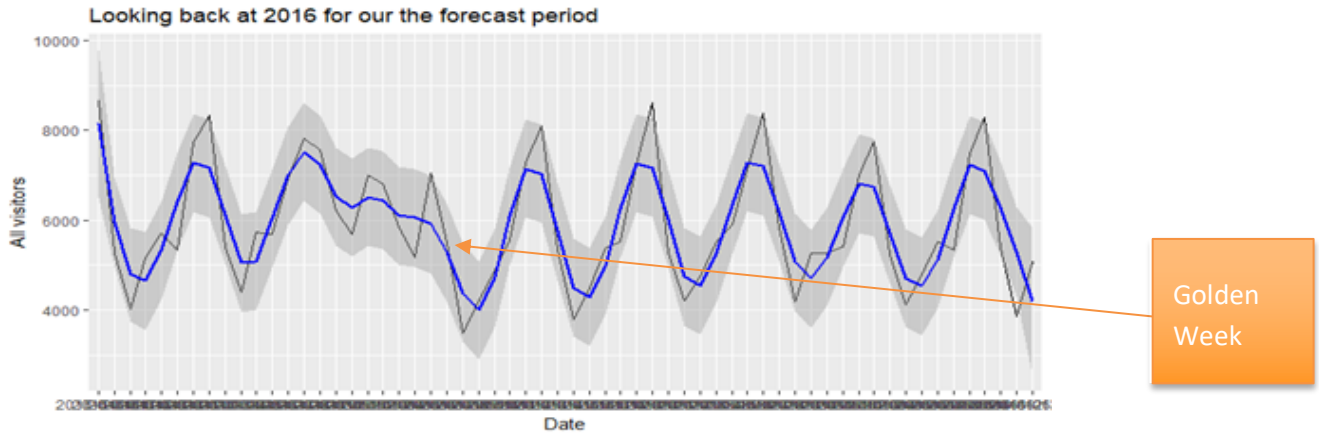

Figure 9 - Impact of Golden Week

The remaining unexplained missing dates are attributed to abnormal restaurant closure, failure to capture visit counts within AirREGI for the date, dates prior to the restaurant's use of AirREGI, or other technical issues leading to missing data.

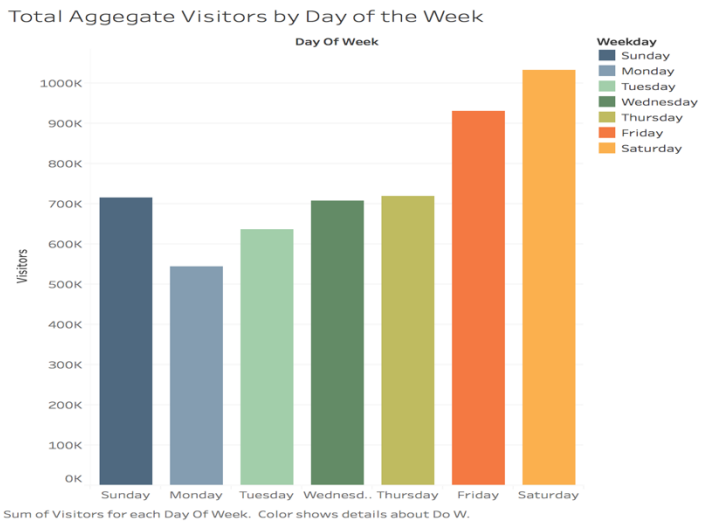

Figure 10 - Total Visitors by the Day of the Week

Drilling down on the periodicity of restaurant visits, a weekly pattern emerges that is intuitive for the restaurant sector: the highest volume is on Saturdays, followed by Fridays, and the lowest volume is on Monday (Figure 10). Visits also vary by month, with the highest volume in March and the lowest volume in May and June.

An important consideration for modeling time series data is stationarity. The training series was tested for stationarity using the Augmented Dickey-Fuller test and confirmed to be stationary. This indicates the data can be modeled using time series forecasting-class models.

In addition to general patterns for all restaurants, individual restaurants demonstrated widely varying patterns of visit volume. Figure 11 shows the time series and autocorrelation of two restaurants. The plot of visitor count against date shows a unique pattern for each. The autocorrelation patterns shown below each time series plot demonstrate the periodicity of the data. Restaurant 101 has a random pattern of visit counts, while restaurant 201 has a clear weekly pattern of visit counts. Predictive models should account for individual restaurant characteristics, and Recruit Holdings will benefit from analysis of both general patterns and specific restaurant patterns in tailoring its AirREGI feature offerings.

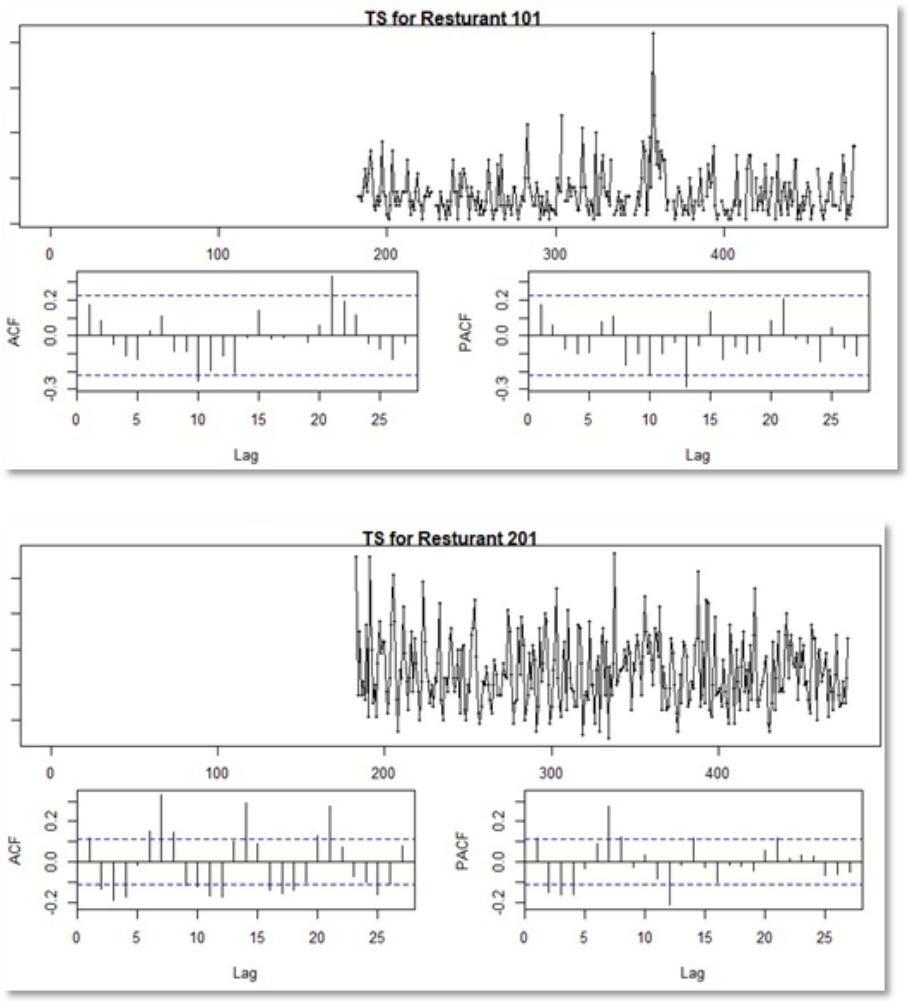

Figure 11 - Examples of individual restaurant visitor patterns

RESERVATIONS DATA FROM AIRREGI AND HPG SYSTEMS

Recruit Holdings provided reservation data from the AirREGI POS system as well as the HPG mobile app. Figure 12 shows the number of visitors reserved through each system over time. Fewer reservations were made in 2016 through the AirREGI system and none at all for a long stretch of time between July 2016 and early October 2016. The volume only increased during the end of that year. Reserved visits were more consistent in 2017. The artificial decline we see after the first quarter of 2017 is most likely related to these reservations being at the end of the training time frame, which means that long-term reservations would not be part of this data set. For HPG reservations, reserved visits follow a more consistent pattern, with a clear spike in December 2016.

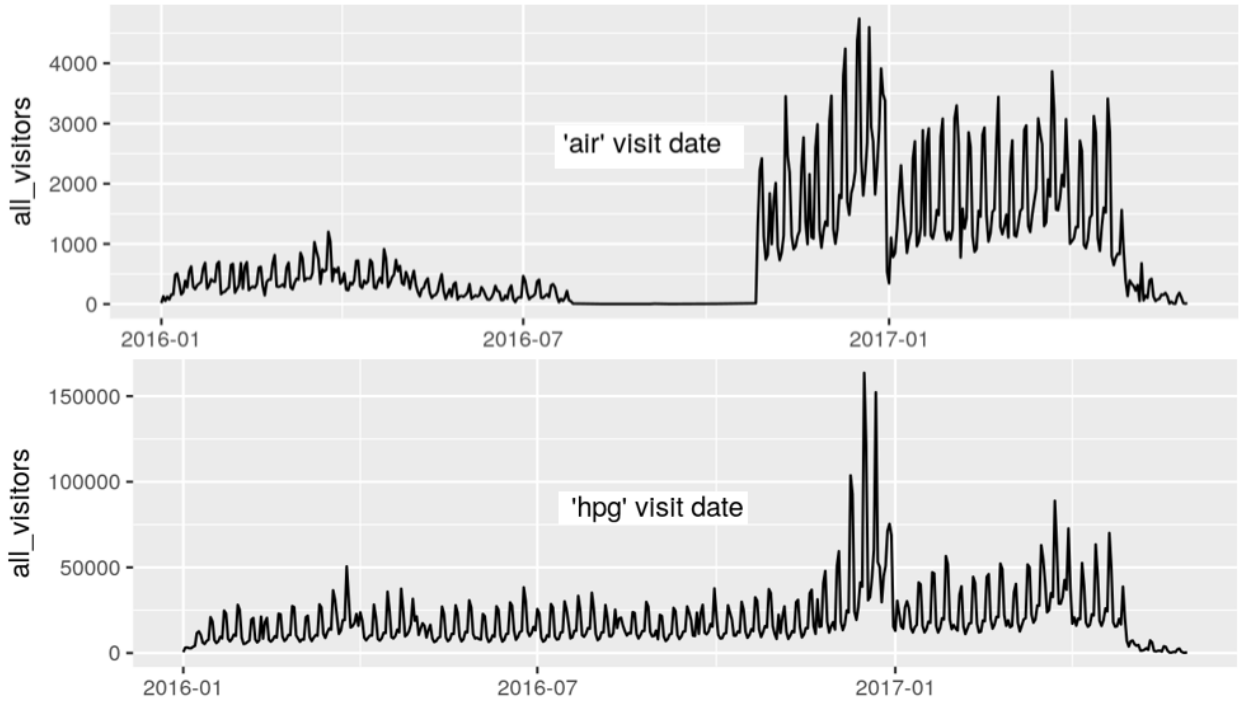

Figure 12 - Time distribution of reservation data

ADDITIONAL FINDINGS

Outliers. The view of total visitors per month (Figure 13) shows evidence of outliers in red, which may impact the models.

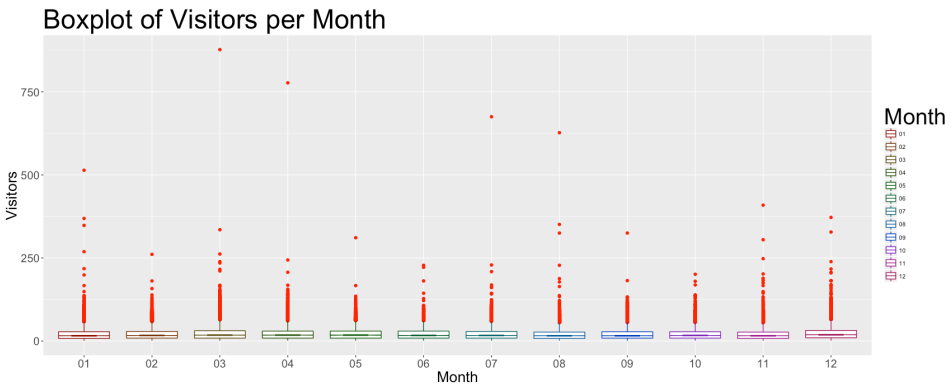

Figure 13 - Boxplot of Total Visitors Showing Outliers by Month

There was a significant outlier in the reservation data. A single restaurant had an abnormally high number of reservations for a single date. Upon examination, the reservations were made from February 2016 up until 5 days prior to the visit date. The source of this anomaly is unclear.

Number of Reservations	Number of Visitors	Earliest Reserved Date	Latest Reserved Date
362	2,241	2/4/16	11/5/16

Table 2 - Reservations for AirREGI store air_a17f0778617c76e2 for 11/10/2016

Potential Clusters. The visitors by city and genre view of the data in Figure 14 reveals there are natural clusters in the data.

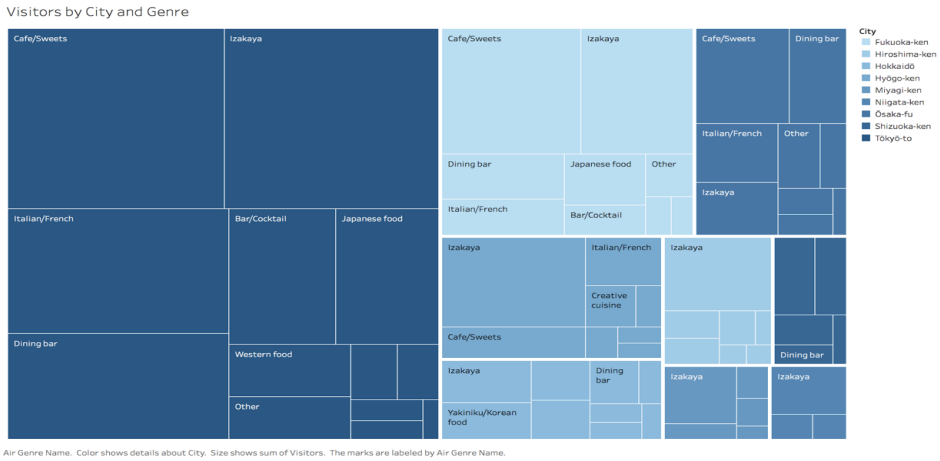

Figure 14 - Potential Clusters, Visits Aggregated by Prefecture and Genre. Size = Visits

Prefectures. The first portion of the the high-cardinality restaurant area name text variable was common across many restaurants. The text was parsed resulting in 9 distinct regions. These labels and the coordinates were referenced against Japanese geographic names and determined to be prefectures, the Japanese equivalent of counties. As shown in Figure 15, the majority of AirREGI restaurants fall in the Tokyo Prefecture. Prefectures will be used as a lower-cardinality descriptor of restaurant location for clustering, modeling and dashboards.

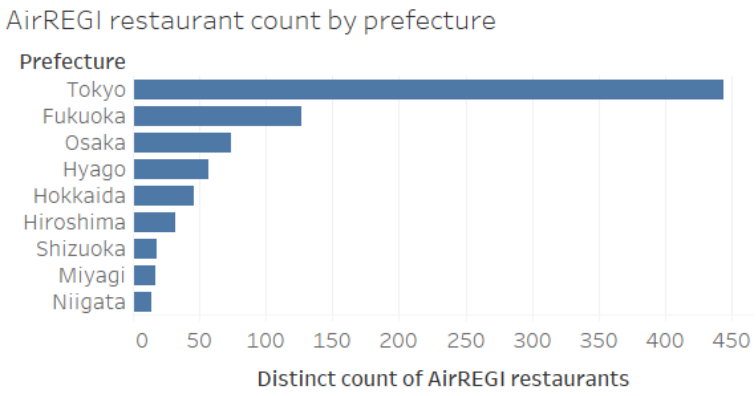

Figure 15 - Restaurant prefectures

DATA MANIPULATION

AUGMENTATION AND FEATURE ENGINEERING

Our handling of the data was done as a series of steps, beginning with the base Kaggle data. First, a dataset comprised of all the reservation data was created by joining HPG and AirREGI reservations into a single set and adding the restaurant information such as latitude and longitude. Next, a dataset showing the visits per store and all related restaurant information was created. These two sets, reservations and visits were used as the base files for augmenting with external data.

In the belief that weather and proximity to various geographical features might impact restaurant visits, we augmented our data with both weather, and “places” data. Weather data was obtained from a weather data EDA shared on the Kaggle site by Hunter McGushion (McGushion, n.d.). Using the weather data .csv files he provided, along with his mapping between restaurants and the nearest weather collection station, we mapped 14 weather variables onto the base dataset. In addition to weather, we also added geographical data collected from Google Places (Google, Inc., n.d.) via a Python script written by a team member. Using the latitude and longitude of each restaurant, HTTP requests were formed to query:

1. subway/train stations within a 500-meter radius of the restaurant
2. museums and zoo within a 500-meter radius
3. bars, malls, and movie theatres within a 500-meter radius

From the results returned by the Google Places query, the distinct number of place names were counted, and appended to dataset to form the new, doubly enhanced dataset with both weather and nearby places.

Predictions will be made at the restaurant level. Historical volume, reservations and the geographic features described above were available at differing levels of granularity. Consolidation of summary

features at the restaurant level provides insights into AirREGI POS customers and summary variables for predictive modeling. Table 3 summarizes features derived across five intuitive categories.

Category	Derived Features
Availability	Number/percent of dates with visitor counts, Number/percent holidays with visitor counts, Mean days per week open
Location	Subway stops, tourist attractions, nightlife, prefecture, zone
Reservations	Number/percent of dates with reservations, mean/median daily reservations, Mean/median daily reserved visitors
	Mean/median reservation lead hours, number of AirREGI and HPG reservations
Volume	Mean/median daily visitors, busiest day of week, mean visits on busiest day, slowest day of week, mean visits on slowest day
Other	Genre, restaurant is also on HPG

Table 3

The final data asset after all new feature additions is depicted in Figure 16.

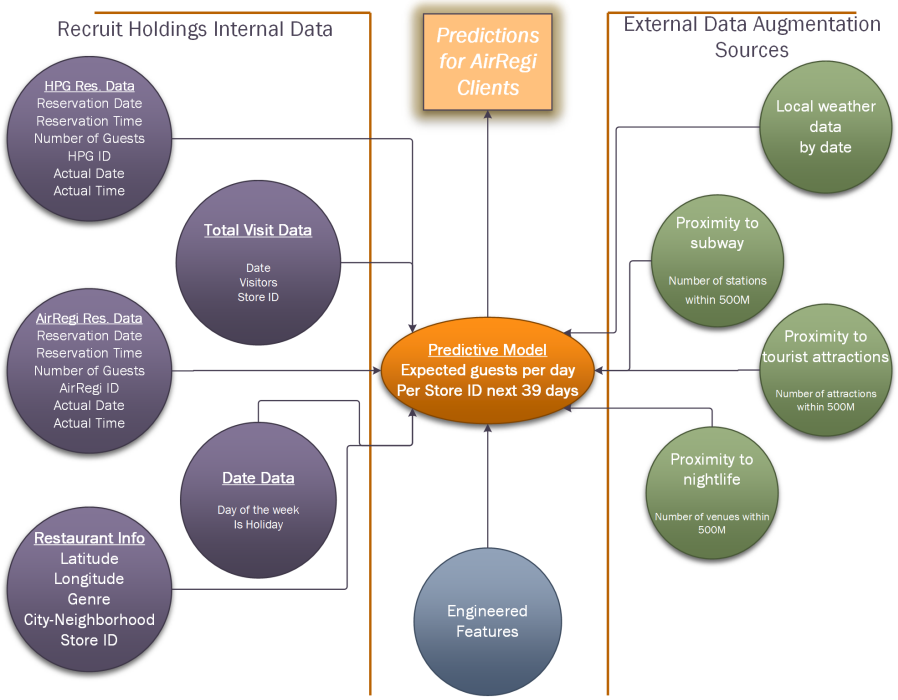

Figure 16 - Enhanced project data

MISSING VALUE TREATMENT

Two modelling paths are under development. One path uses ARIMA modelling, the other uses boosted trees. In the case of boosted trees, initial models were built without any imputation; missing values were simply converted to 0. Subsequent models will be built using a more refined imputation strategy informed by the patterns discovered in the data.

For time series models like exponential smoothing and auto ARIMA, a cascading conditional approach was used for imputation.

1. In order to reflect the first day of a restaurant's visitor counts, all NAs before the first visits row were ignored.
2. Missing data every 7 days were interpreted as a regular weekly closure: Impute 0.
3. Missing data at irregular weekdays where the date was a holiday were interpreted as restaurant closure: Impute 0.
4. Remaining unexplained missing values: Impute median visitor count for the restaurant.

The example below illustrates the imputation process for a few restaurants in the month of March 2017.

A	B	C	D	E	F	G
Date	Day	Holiday Flg	Day	air_97c	air_97e	air_9828
3/1/2017	Wednesday	0	1	NA	21	21
3/2/2017	Thursday	0	2	10	NA	13
3/3/2017	Friday	0	3	18	16	18
3/4/2017	Saturday	0	4	27	35	23
3/5/2017	Sunday	0	5	15	10	18
3/6/2017	Monday	0	6	5	1	7
3/7/2017	Tuesday	0	7	6	21	NA
3/8/2017	Wednesday	0	8	NA	13	16
3/9/2017	Thursday	0	9	9	NA	8
3/10/2017	Friday	0	10	16	17	20
3/11/2017	Saturday	0	11	25	22	26
3/12/2017	Sunday	0	12	18	37	NA
3/13/2017	Monday	0	13	12	13	24
3/14/2017	Tuesday	0	14	11	11	NA
3/15/2017	Wednesday	0	15	NA	20	15
3/16/2017	Thursday	0	16	10	NA	15
3/17/2017	Friday	0	17	9	22	17
3/18/2017	Saturday	0	18	28	39	18
3/19/2017	Sunday	0	19	19	54	12
3/20/2017	Monday	1	20	5	26	NA
3/21/2017	Tuesday	0	21	5	25	NA
3/22/2017	Wednesday	0	22	NA	30	17
3/23/2017	Thursday	0	23	20	NA	19
3/24/2017	Friday	0	24	6	34	5
3/25/2017	Saturday	0	25	15	33	7
3/26/2017	Sunday	0	26	11	47	22
3/27/2017	Monday	0	27	NA	7	6
3/28/2017	Tuesday	0	28	8	21	NA
3/29/2017	Wednesday	0	29	16	34	22
3/30/2017	Thursday	0	30	9	NA	17
3/31/2017	Friday	0	31	11	19	29

NA's marked in blue occur once a week - imply a weekly resturant closure

NA's on days with holiday_flg= 1 - imply holiday closure

Other NAs - interpreted as true missing data

Figure 17 - Illustration of imputation inferences

Outliers were also addressed prior to modeling. For simplicity we employed the 'tsclean' function from the *forecast* package for R, which is known to work well on both seasonal and non-seasonal time series. Finally, in order to predict volume at the restaurant level for future dates, the final modeling dataset was converted to a wide format so that models could be run in a loop.

ANALYSIS OF DATA

CORRELATIONS

The consolidated data are shown in a correlation plot in Figure 18. Variables are intuitively correlated with respect to weather, geographic location, and nearby places. Note there is little to no univariate correlation to the target, visitors. Historical visits for each individual restaurant are the most powerful predictor; however, earlier findings suggest clusters of variables may demonstrate a relationship to the target.

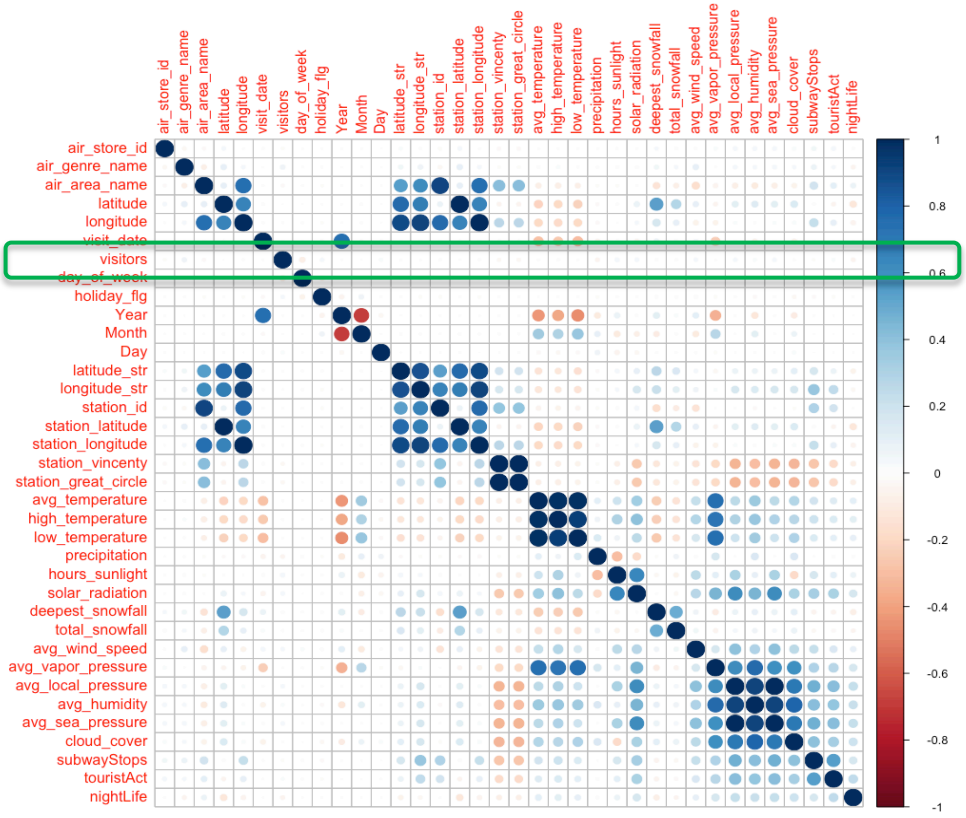

Figure 18 - Correlations among potential restaurant predictors

The correlations in the aggregated reservation data are shown in Figure 19. The number of reserved visitors has a moderate little correlation to visitor count. However, the inconsistencies in the reservation data previously noted limit their use in predictive models.

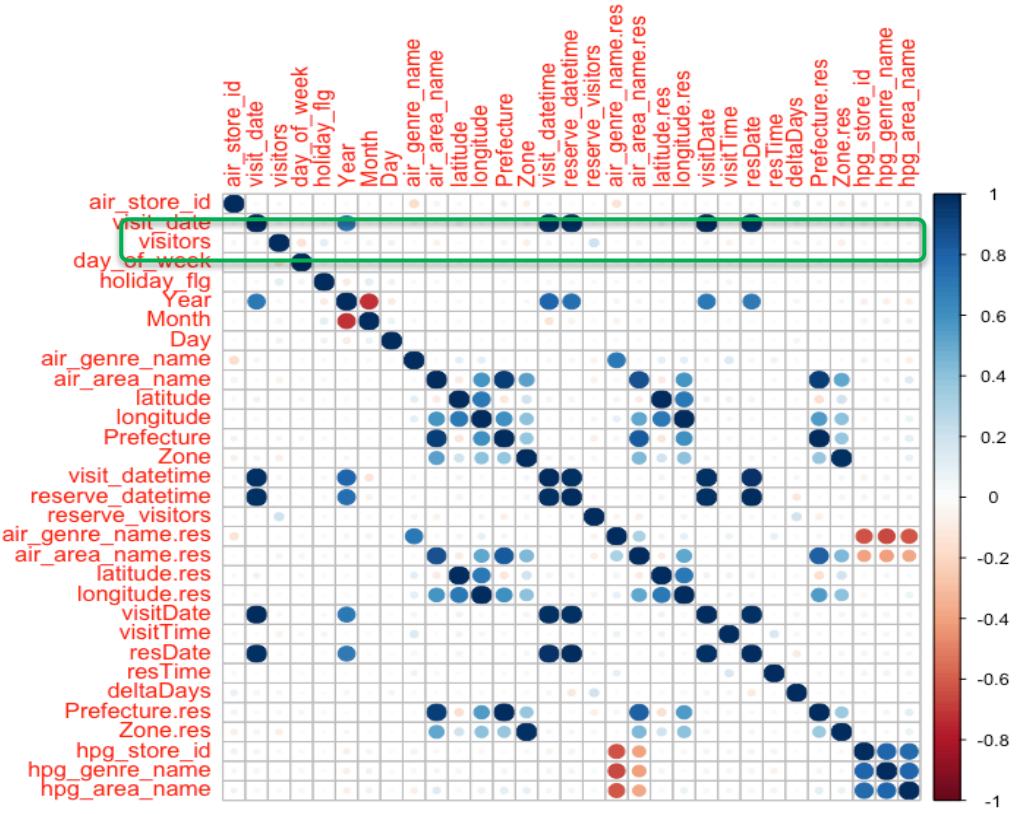

Figure 19 - Correlations among potential reservation predictors

CLUSTERING

The availability of many restaurant characteristics in varying combinations suggests clustering will provide valuable insights. Cluster membership will also be tested in models for predictive value. Cluster profiles may provide insights for Recruit Holdings into their AirREGI customer base and provide a basis for customer segmentation and targeting the prediction-enhanced AirREGI POS product.

An initial effort at clustering the restaurant data yielded promising results in preliminary models. After consultation with our colleague, Dr. Donald Wedding, of Northwestern University, the clustering approach has been refined. Refinements are currently in process, and a brief overview of the approach is described here.

The process is to first group the enhanced restaurant data according to the intuitive categories described in Table 3 in the previous section. For instance, all variables in the “Availability” category are

grouped together. Similar groups were created with other variables for “Location”, “Reservation”, and “Volume.”

The variables in each group are standardized with mean of zero and standard deviation of one. All 829 restaurants were then clustered with a k-means clustering function. Using the example of the Availability group, a cluster size of six optimizes the trade-off between having clusters of meaningful size and ensures that the individual data points are “close” to their assigned cluster.

The 829 restaurants are clustered on the availability variables as shown in Figure 20. Further review of these clusters shows that restaurants in Cluster 6 are significantly busier during the weekdays than are restaurants in other clusters. Cluster 6 also has a much higher proportion of restaurants located in Tokyo than the typical restaurant in the data set and tends to be closed for holidays far more often than restaurants in other clusters.

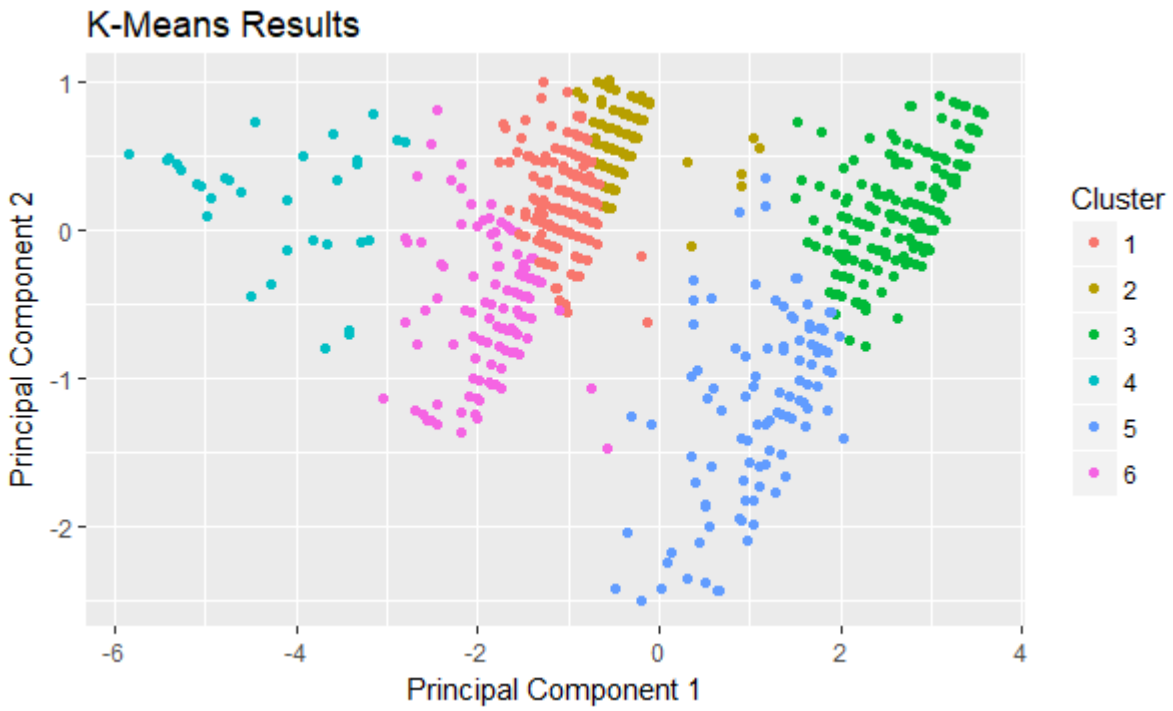

Figure 20 - Clustering of Availability variables

Clusters will be built for the remaining variable groupings and then profiled as in the example above. Interesting segments will be used to frame insights for the Recruit Holdings dashboard, and cluster memberships will be trialed in subsequent predictive models.

PRELIMINARY MODELS

For this data, the two modeling methods most likely to be feasible solutions are boosted trees and classic time series models. This dataset is from a Kaggle Competition. The root mean squared logarithmic error (RMSE) was used to evaluate and score submissions. The RMSE may be thought of as the "distance" between the actual values and the predictions; a lower error value indicates a more accurate model.

BOOSTED TREES CLASS OF MODELS

Boosted trees are attractive for handling complex patterns and relationships in the data, and time series will address the periodicity of visitor volume. Initial approaches to modeling include two different boosted tree algorithms and two different traditional forecasting algorithms. Initial modeling results are promising, particularly when using the augmentation data.

The first boosted tree approach used the R package 'ForecastXGB' which is under development by Peter Ellis (Ellis, n.d.) Ellis describes his package as follows:

> *The forecastxgb package aims to provide time series modelling and forecasting functions that combine the machine learning approach of Chen, He and Benesty's xgboost with the convenient handling of time series and familiar API of Rob Hyndman's forecast.* (Ellis, Timeseries forecasting using extreme gradient boosting, n.d.)

The second approach used XGBoost (Chen, He, Benesty, Khotilovich, & Tang, n.d.)to model the data without attempting to model the time series explicitly. Incrementally complex models were built using the raw data and adding additional features: Kaggle data only, Kaggle + weather, and so on. The results of submitting the predictions produced by each model to the Kaggle scoring system are summarized below:

Submission Description	Private Score	Public Score
XGBoost using Kaggle+weather+clustersv1, cluster column moved to earlier in the dataframe	0.709	0.704
XGBoost using Kaggle + weather + clustersv1 data	0.714	0.704
XGBoost using Kaggle + weather + metro data	0.798	0.807
XGBoost using Kaggle + a subset of weather data	0.783	0.768
XGBoost using Kaggle + weather	0.752	0.734
XGBoost using Kaggle data Only	0.834	0.852
ForecastXGB using Kaggle data only, lag = 1	1.295	1.016
ForecastXGB using Kaggle data only, lag = 7	0.874	0.848

The boosted tree models show promising results. Adding the weather data improved results, however adding the geographical data did not. This may have been caused by the obfuscation of the exact restaurant location in the Kaggle dataset yielding undifferentiated geographically position data. Adding the first version of the cluster membership variable to the dataset and moving it earlier in the dataset resulted in significant test error improvement. The first model using cluster data, built by appending the cluster membership information to the end of the data frame, did not perform as well. The implication of this is that the trees being built by the algorithm branch in a more meaningful and better differentiated way when the cluster variable occurs earlier in the branch-decision process.

TIME SERIES CLASS OF MODELS

Forecasting methods were selected to model the time series. Exponential smoothing was chosen for its simplicity and auto ARIMA for its ability to handle seasonality, differencing and auto-regression, which is a linear combination of past values of visits. Both ETS and Auto ARIMA models were implemented using Forecast package from Rob J Hyndman (Hyndman, n.d.).

Time series models require complete historical training data. The dataset was imputed as explained in the 'Missing values and imputation' section.

GENERAL APPROACH FOR TIME SERIES MODELING

The imputed air_visit dataset in conjunction with holiday information was reshaped into the wide form with a column representing each restaurant while the rows represented visits for each day. A loop was run for all 829 restaurants and individual models were built.

It was challenging to determine how to explore the performance of these sets of models. There was no single time series modeling class that worked well for all restaurants. So, reviewing accuracy for randomly selected restaurants served best. AIC and RMSE were used to check performance.

ETS

An Exponential Smoothing (ETS) model was implemented without any transformations in the data. The package helps pick type of smoothing automatically based on minimizing AIC (Akaike's Information Criterion). For the best ETS set of models, best AIC was 1191.18. Here's an example restaurant 10's residuals. Both ACF and PACF values are within acceptable limits.

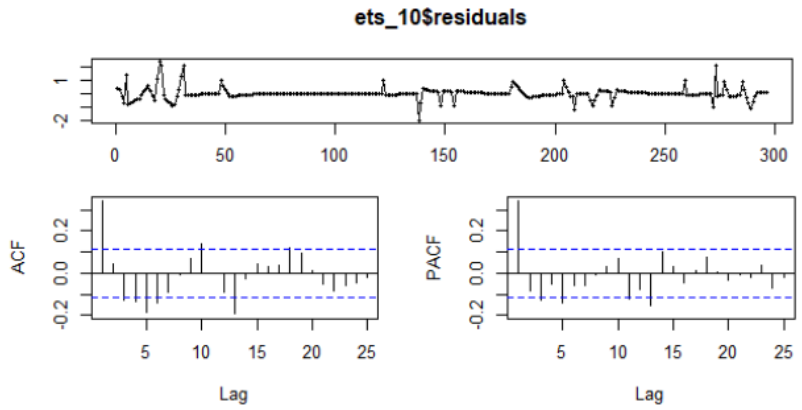

Figure 21 - Residuals Plot ETS model for restaurant 10

AUTO ARIMA

The second modelling class used was auto ARIMA, which stands for Auto Regressive Integrated Moving Average (Hyndman, n.d.) Although one can run ARIMA models and select the autoregressive (p), differencing (d) & moving average (q) parameters, auto-ARIMA handles it and picks appropriate values. For the best auto ARIMA set of models, the package selected ARIMA(1,0,0) for (p,d,q). The AIC value was way better than the ETS set of models, at 292.33. Here's an example restaurant 10's residuals plot, which are within normal limits.

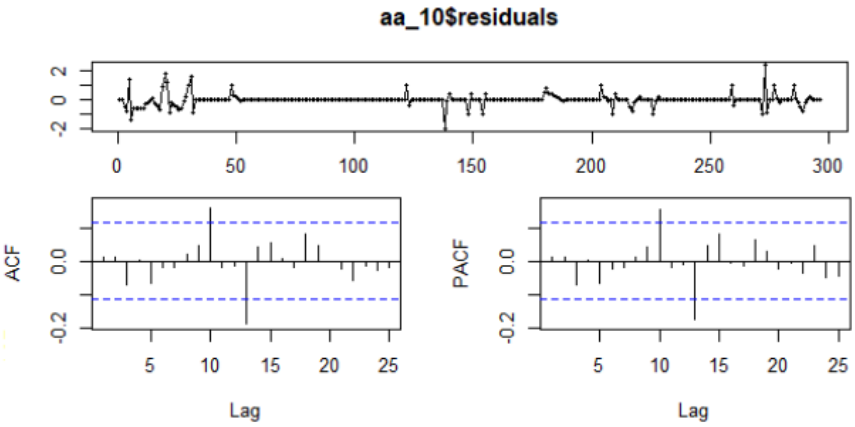

Figure 22 - Residuals Plot auto-ARIMA model for restaurant 10

Here are the scores on Kaggle. Both ETS and Auto ARIMA models worked very well. The winning public score in the Kaggle competition was 0.501.

Submission and Description	Private Score	Public Score
Auto ARIMA	0.613	0.587
ETS	0.616	0.601

DASHBOARD PROTOTYPES

WildCATTS Analytics is developing two Tableau dashboards to address the project goals. The first is a prototype dashboard for incorporating visitor predictions in the AirREGI POS product suite, aimed at enhancing value for restaurant managers. The second is a comprehensive dashboard for Recruit Holdings to assess and monitor their AirREGI business. Both dashboards are interactive and will leverage the enriched data developed during the project engagement.

AIRREGI STORE MANAGER DASHBOARD

SCENARIO: Yuichi Sato is a restaurant manager in the Osaka prefecture. He manages two stores which specialize in serving Teppanyaki. As part of the value add for the Table for Four project, Recruit Holdings can now offer Yuichi a new way to view data on his restaurants. After going through an RH setup experience to select his restaurants, Yuichi can see a map showing his stores, Yuichi's Original Teppanyaki, and Yuichi's Store Two

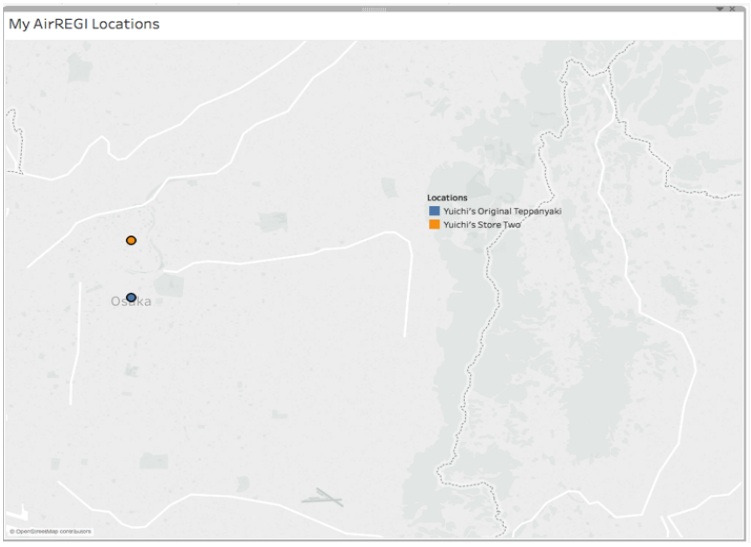

Figure 23 - AirREGI store manager map

Clicking on either location will show a view of that individual restaurant or performing a multi-select will show both. In the mocked-up example below, the restaurant owner is being presented with a view showing both restaurants' recent counts of actual visitors as the yellow line, in conjunction with the forecasted visitors for the next several days as the bars. At this moment the team does not have test predictions for dates which are part of the training data, therefore we do not currently have data for actuals and predictions which overlap. The mock-up was created using synthetic data for the "actual' values in the part of the chart where the series overlap.

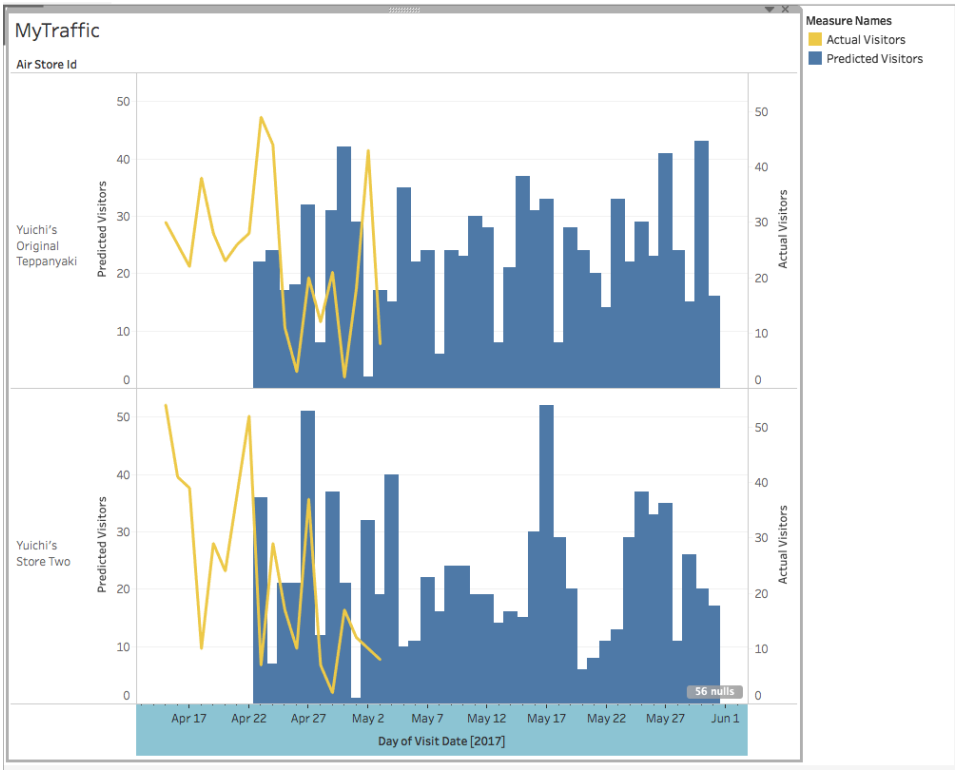

Figure 24 - Actual vs. predicted volume for restaurants selected from the map above.

RECRUIT HOLDINGS DASHBOARD

The Recruit Holdings dashboard will display a map of all AirREGI restaurants. This will serve as a graphic interface point of entry for drilling down to individual regions and restaurants (Figure 24). A second view will provide insights into restaurant characteristics, such as genre and volumes, incorporating insights from clustering analysis. A third view will track the restaurant volume predictions across the AirREGI customer base. Finally, we will explore adding a fourth view to examine reservation patterns from both AirREGI and HPG.

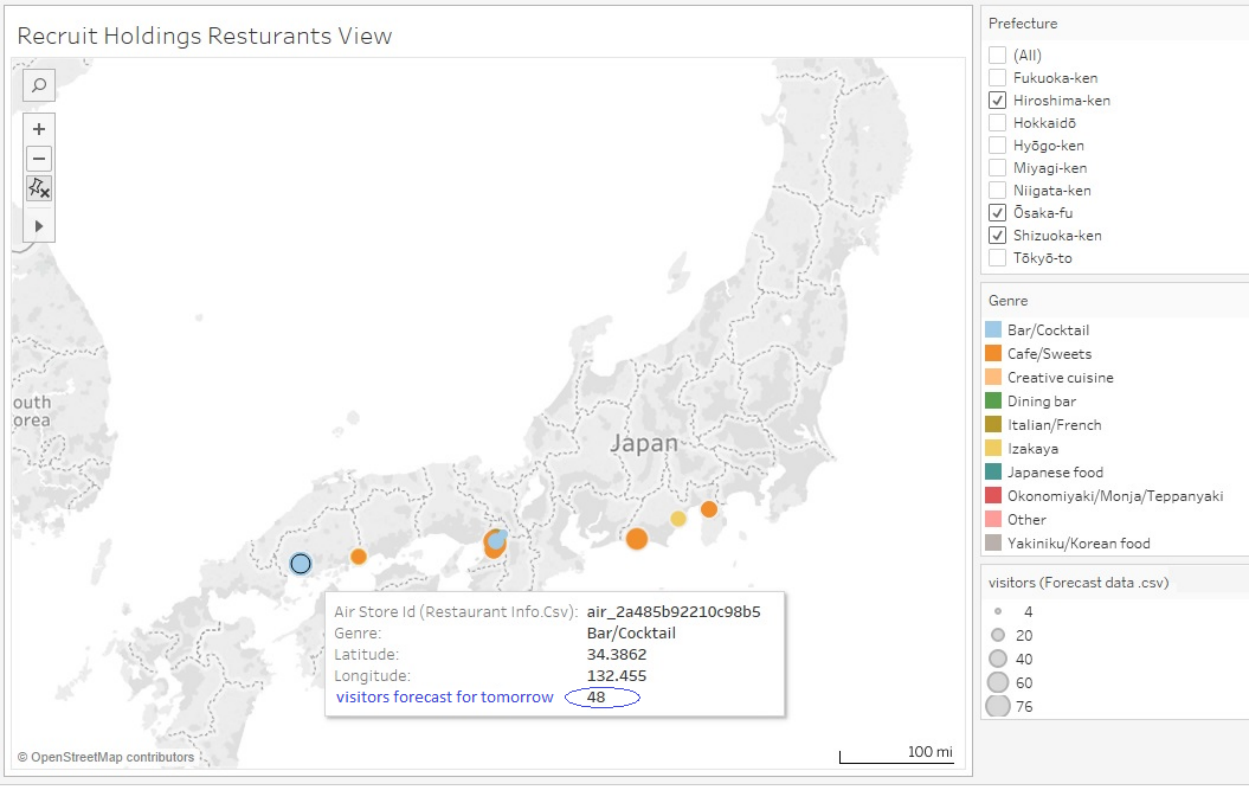

Figure 25 - Recruit Holdings dashboard map view

MOBILE EXPERIENCE

WildCATTS analytics is exploring an enhancement to the AirREGI mobile experience based on the restaurant manager dashboard described in the previous section. After evaluating several potential platforms, we have established the best point of entry is to extend dashboard capabilities to a mobile experience using Tableau Server. This work will be incorporated into our next phase of development.

BUSINESS IMPACT

The ability to accurately predict restaurant volume up to 30 days in advance is a small but important part of RH's overall data strategy. WildCATTS Analytics assistance in solving this piece of the puzzle will help RH achieve its goal: *"The more options you have, the richer your life may be. That's why here at Recruit, we are hard at work to give you as many options as possible."*

RH's Business Support Pack for restaurants, according to its annual report, "comes with six core functions: website creation, online advertising, website reservation tools, table management and reservation books, customer management, and messaging. We offer a variety of rates and plans to customers, who select the one that suits their business best." (Recruit Holdings) RH began charging restaurants for these services in January 2017.

The ability to predict restaurant volume will aid these core functions by optimizing a restaurant's cost of goods, staffing, and use of incentives such as coupons to drive traffic.

FINANCIAL IMPACT

RH does not provide details on the cost of its Business Support Pack. According to the New York Times, Yelp charges $250 per month for a reservations-only system. The additional data tools, advertising, website creation, and other services would certainly cost a premium on top of the $250 monthly fee.

From a conservative estimating standpoint, it is assumed that RH will charge $250 per month per restaurant, which equates to $3,000 per year per restaurant. Per RH's annual report, in 2016 it had sales of its Air Platform to restaurants of approximately $342 million USD. Dividing the $342 million by $3,000, it can be estimated that RH's Business Support Pack is in 114,000 restaurants in Japan. Add in an estimated 25% growth, as it is a new product, and first year revenues for the product can be estimated to be over $384 million USD. See Table 2 for calculations. Table 2 further estimates various costs associated with the system. The bottom line comes to over $278 million USD in profit for the first year for RH.

Conservatively estimate that improved ability to predict restaurant volume is providing 1% of value to the entire Business Support Pack, as mentioned above in the form of optimizing a restaurant's cost of goods, staffing, and use of incentives such as coupons to drive traffic. This project would therefore provide approximately $2.78 million USD in value to RH in its first year.

Further recall that RH has many other business lines, including the housing, bridal, educational, automobile, travel, dining, and human resources industries. RH also operates in over 60 countries. The ability to leverage the predictive algorithms and lessons learned from this project and apply elsewhere in RH can yield an exponential amount of value to RH's bottom line.

Recruit Holdings, Premium Data Services		
All financial numbers in USD		
REVENUE	**CALCULATION**	**NOTES**
annual fee per restaurant customer	$3,000	$250 per month per restaurant
estimated number of subscribers at start of year	114,000	Per annual report, $342 million USD in rev. $342mm/$3000 per year = 114,000
first year growth of new customers	28,500	25% growth in premium subscribers per YEAR
estimated revenue year one	**$384,750,000**	(starting subscription * annual fee) + (0.5 * new customers * annual fee) [0.5 new customers because half will be with RH, on average, for full year]
COSTS		
One time fixed cost of software	$10,000,000	Estimate
Ongoing software support per year	$32,062,500	$250 per restaurant per YEAR (also factors in 25% growth)
WildCATTS Analytics Fee for Current Engagement	$234,000	$325 per hour, 4 people, 20 hours per week, 9 weeks
Sales, Promotions, Entertainment, etc.	$64,125,000	$500 per restaurant, per YEAR
estimated total costs year one	**$106,421,500**	
PROFIT		
estimated margin year one	$278,328,500	For entire RH Business Support Pack
	$2,783,285	**Margin attributable to improved restaurant predictive modeling**

Table 4

PRELIMINARY CONCLUSIONS

Recruit Holdings' data assets for its AirREGI product provide a rich source of business insights and a strong foundation for predictive modeling. Data can be easily augmented with readily available weather and location-specific information. Derived restaurant characteristics show promise for clustering and segmentation as well as predictive modeling of restaurant visitors. Preliminary models show promising results. Additional feature selection, model tuning, and exploration of other modeling classes, such as Bayesian and random forest, are expected to improve predictive accuracy further.

The augmented data lends itself to providing ongoing business insights and embedding predictions within the business context using dashboards for both Recruit Holdings and its AirREGI restaurant

managers. Leveraging HPG reservation data for the modeling objective is challenging due to the poor overlap between AirREGI and HPG datasets and questionable data findings. We will continue to explore the utility of the reservation data and make recommendations for improving its value. Finally, our financial analysis demonstrates incorporation of visitor predictions into the AirREGI POS product offering is profitable for Recruit Holdings.

PROJECT STATUS

The project is divided into 4 phases, outlined below. Phase 1 was completed on 4/22/2018, and Phase 2 is being submitted on time on 5/13/2018. There are no known schedule risks or impacts at this time.

- **Completed**
  - Phase 1 - Team formation, and initial engagement with customer and customer problem.
    - ✓ Deliverable – Initial statement of work (this document)
  - Phase 2 - Deep dive into the business problem and build first round of models.
    - ✓ Deliverable - Initial Findings Report, project status, and Executive Summary of progress to date.
- **Scheduled future work**
  - ➢ Phase 3 - Due 5/29/2018 - Final model selection and testing. Finalize dashboard build. Develop mobile experience. Generate final reporting artifacts and recommendations.
    - ● Deliverable - Executive Summary Report, Final Report with all models, code, data analysis, dashboard(s), mobile experience, and recommendations.
  - ➢ Phase 4 - Due 6/10/2018 - Presentation of results.
    - ● Deliverable – Live, or pre-recorded presentation for executive staff of the results and recommendations along and associated slide deck.
- **Known Risks or schedule impacts**
  - ➢ None at this time.

WORKS CITED

Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (n.d.). *Package 'xgboost'*. Retrieved from CRAN: ftp://cran.r-project.org/pub/R/web/packages/xgboost/xgboost.pdf

Ellis, P. (n.d.). *forecastxgb-r-package*. Retrieved from https://github.com/: https://github.com/ellisp/forecastxgb-r-package

Ellis, P. (n.d.). *Timeseries forecasting using extreme gradient boosting*. Retrieved from Free Range Statistics: http://freerangestats.info/blog/2016/11/06/forecastxgb

Google, Inc. (n.d.). *Place Search*. Retrieved from Google Places API for Web: https://developers.google.com/places/web-service/search

Hyndman, R. (n.d.). *Package 'forecast'*. Retrieved from CRAN: https://cran.r-project.org/web/packages/forecast/forecast.pdf

McGushion, H. (n.d.). *Exhaustive Weather EDA/File Overview* . Retrieved from Kaggle: https://www.kaggle.com/huntermcgushion/exhaustive-weather-eda-file-overview

Recruit Holdings . (n.d.). *Recruit Holdings 2017 Annual Report.* Retrieved from https://recruit-holdings.com: https://recruit-holdings.com/assets/pdf/annual/2017/annual_2017_en_all.pdf

Statista. (n.d.). *Number of mobile phone users in Japan from 2013 to 2020 (in millions)*. Retrieved from Statista The Statistics Portal: https://www.statista.com/statistics/274672/forecast-of-mobile-phone-users-in-japan/

Wikipedia. (n.d.). *Japan*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Japan