Recruit Restaurant Visitor Forecasting

JULY 24

BUAN 6357
Authored by: Sai Rahul Dhulipalla

Contents

1.0 Summary....4

2.0 Introduction....4

3.0 About the Data....4

air_visit_data.csv: 4

air_reserve.csv / hpg_reserve.csv:....5

air_store_info.csv / hpg_store_info.csv:....6

store_id_relation.csv:....6

date_info.csv:....7

4.0 Exploratory Data Analysis 7

4.1 Air Visits Data:....7

4.2 Air Reserve Data:....9

4.3 HPG Reserve Data:....11

4.4 AIR Store Data: 12

4.5 HPG Store Data:....13

4.6 Date Info: 14

5.0 Time Series Analysis of Data for Forecasting Future Visitors number 16

5.1 ARIMA....16

5.1.1 Feature Engineering:....16

5.1.2 Example of running Auto ARIMA on few Restaurant time series:....16

5.1.3 Running Auto ARIMA on all the 829 Time series: 18

5.2 ETS Forecasting 18

5.2.1 Example of running ETS on few Restaurant time series:....18

5.2.2 Running ETS on all the 829 Time series: 20

5.3 Prophet Forecasting 20

5.3.1 Example of running Prophet on few Restaurant time series:....21

5.3.2 Running Prophet on all the 829 Time series: 23
5.4 ARIMA Vs ETS Vs Prophet.....24
6.0 Things that can be done to improve the performance of forecasting.....24
7.0 References 25

1.0 Summary

This is the report of a project done to build time series forecasting models that predicts the future visitors of restaurants in Recruit Holdings database. The dataset contains past visitor's data over 478 days for 829 different restaurants. The approach followed here is to build different models for different restaurants to do this task. Forecasting models were tried to be built using ARIMA, ETS and Prophet Forecasting methods. Forecasting method that produced models with low Root Mean Squared Error (RMSE) was chosen as the final method to build models. Here models built using the ETS methods gave the least RMSE value.

2.0 Introduction

Running a local restaurant isn't as easy as it looks. There are many unexpected troubles that pop up in running a restaurant.

One common difficulty is that restaurants need to know how many customers to expect each day to effectively purchase ingredients and schedule staff members to serve its customers well. This forecast isn't easy to make because many unpredictable factors affect restaurant attendance.

Recruit Holdings has unique access to key datasets that could make automated future customer prediction possible. So, it launched a competition on Kaggle to build models that can do these predictions. I included the link of competition in the reference section.

I choose to solve this prediction problem as part of my project for advanced business analytics course I am taking at The University of Texas at Dallas.

More details about the datasets, Exploratory data analysis I performed and the models I built are explained in further sections.

3.0 About the Data

The data was collected from Japanese restaurants. 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 data is based on the time range of Jan 2016 to Apr 2017.

Box folder Link to Data files: https://utdallas.box.com/s/je5xtbow4f5uznuq4qtu8z0c1kgsxw91
The individual data files and what they contain are mentioned below.

air_visit_data.csv:

- This is essentially the main training data set.
- It contains historical visit data for air restaurants.
- It contains 252,108 records for 829 different restaurants overtime period from January 2016 to April 2017.

The Glimpse of data can be found below:

	air_store_id	visit_date	visitors
1	air_ba937bf13d40fb24	2016-01-13	25
2	air_ba937bf13d40fb24	2016-01-14	32
3	air_ba937bf13d40fb24	2016-01-15	29
4	air_ba937bf13d40fb24	2016-01-16	22
5	air_ba937bf13d40fb24	2016-01-18	6
6	air_ba937bf13d40fb24	2016-01-19	9
7	air_ba937bf13d40fb24	2016-01-20	31
8	air_ba937bf13d40fb24	2016-01-21	21
9	air_ba937bf13d40fb24	2016-01-22	18
10	air_ba937bf13d40fb24	2016-01-23	26
11	air_ba937bf13d40fb24	2016-01-25	21
12	air_ba937bf13d40fb24	2016-01-26	11
13	air_ba937bf13d40fb24	2016-01-27	24
14	air_ba937bf13d40fb24	2016-01-28	21
15	air_ba937bf13d40fb24	2016-01-29	26

air_reserve.csv / hpg_reserve.csv:

- It contains the data on reservations made through the air / hpg systems.
- Air_reserve.csv contains 92,378 records for 314 different air stores.
- hpg_reserve.csv contains 2,000,320 records for 13325 different hpg stores.

The Glimpse of data can be found below:

	air_store_id	visit_datetime	reserve_datetime	reserve_visitors
1	air_877f79706adbfb06	2016-01-01 19:00:00	2016-01-01 16:00:00	1
2	air_db4b38ebe7a7ceff	2016-01-01 19:00:00	2016-01-01 19:00:00	3
3	air_db4b38ebe7a7ceff	2016-01-01 19:00:00	2016-01-01 19:00:00	6
4	air_877f79706adbfb06	2016-01-01 20:00:00	2016-01-01 16:00:00	2
5	air_db80363d35f10926	2016-01-01 20:00:00	2016-01-01 01:00:00	5
6	air_db80363d35f10926	2016-01-02 01:00:00	2016-01-01 16:00:00	2
7	air_db80363d35f10926	2016-01-02 01:00:00	2016-01-01 15:00:00	4
8	air_3bb99a1fe0583897	2016-01-02 16:00:00	2016-01-02 14:00:00	2

	hpg_store_id	visit_datetime	reserve_datetime	reserve_visitors
1	hpg_c63f6f42e088e50f	2016-01-01 11:00:00	2016-01-01 09:00:00	1
2	hpg_dac72789163a3f47	2016-01-01 13:00:00	2016-01-01 06:00:00	3
3	hpg_c8e24dcf51ca1eb5	2016-01-01 16:00:00	2016-01-01 14:00:00	2
4	hpg_24bb207e5fd49d4a	2016-01-01 17:00:00	2016-01-01 11:00:00	5
5	hpg_25291c542ebb3bc2	2016-01-01 17:00:00	2016-01-01 03:00:00	13
6	hpg_28bdf7a336ec6a7b	2016-01-01 17:00:00	2016-01-01 15:00:00	2
7	hpg_2a01a042bca04ad9	2016-01-01 17:00:00	2016-01-01 17:00:00	2

air_store_info.csv / hpg_store_info.csv:

- It contains details about the air / hpg restaurants including genre and location.
- Air_store_info.csv contains details about 829 stores.
- hpg_store_info.csv contains details about 4,690 stores.

The Glimpse of data can be found below:

	air_store_id	air_genre_name	air_area_name	latitude	longitude
1	air_0f0cdeee6c9bf3d7	Italian/French	HyÅ□go-ken KÅ□be-shi KumoidÅ□ri	34.69512	135.1979
2	air_7cc17a324ae5c7dc	Italian/French	HyÅ□go-ken KÅ□be-shi KumoidÅ□ri	34.69512	135.1979
3	air_fee8dcf4d619598e	Italian/French	HyÅ□go-ken KÅ□be-shi KumoidÅ□ri	34.69512	135.1979
4	air_a17f0778617c76e2	Italian/French	HyÅ□go-ken KÅ□be-shi KumoidÅ□ri	34.69512	135.1979
5	air_83db5aff8f50478e	Italian/French	TÅ□kyÅ□-to Minato-ku ShibakÅ□en	35.65807	139.7516
6	air_99c3eae84130c1cb	Italian/French	TÅ□kyÅ□-to Minato-ku ShibakÅ□en	35.65807	139.7516
7	air_f183a514cb8ff4fa	Italian/French	TÅ□kyÅ□-to Minato-ku ShibakÅ□en	35.65807	139.7516
8	air_6b9fa44a9cf504a1	Italian/French	TÅ□kyÅ□-to Minato-ku ShibakÅ□en	35.65807	139.7516
9	air_0919d54f0c9a24b8	Italian/French	TÅ□kyÅ□-to Minato-ku ShibakÅ□en	35.65807	139.7516

	hpg_store_id	hpg_genre_name	hpg_area_name	latitude	longitude
1	hpg_6622b62385aec8bf	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
2	hpg_e9e068dd49c5fa00	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
3	hpg_2976f7acb4b3a3bc	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
4	hpg_e51a522e098f024c	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
5	hpg_e3d0e1519894f275	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
6	hpg_530cd91db13b938e	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
7	hpg_02457b318e186fa4	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
8	hpg_0cb3c2c490020a29	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682
9	hpg_3efe9b08c887fe9a	Japanese style	TÅ□kyÅ□-to Setagaya-ku TaishidÅ□	35.64367	139.6682

store_id_relation.csv:

- It connects the air and hpg ids.
- There are 150 pairs connecting air and hpg ids.

The Glimpse of data can be found below:

Filter

	air_store_id	hpg_store_id
1	air_63b13c56b7201bd9	hpg_4bc649e72e2a239a
2	air_a24bf50c3e90d583	hpg_c34b496d0305a809
3	air_c7f78b4f3cba33ff	hpg_cd8ae0d9bbd58ff9
4	air_947eb2cae4f3e8f2	hpg_de24ea49dc25d6b8
5	air_965b2e0cf4119003	hpg_653238a84804d8e7
6	air_a38f25e3399d1b25	hpg_50378da9ffb9b6cd

date_info.csv:

- It contains Japanese holidays.
- It contains holiday flags over the period present in train and test datasets.

The Glimpse of data can be found below:

Filter

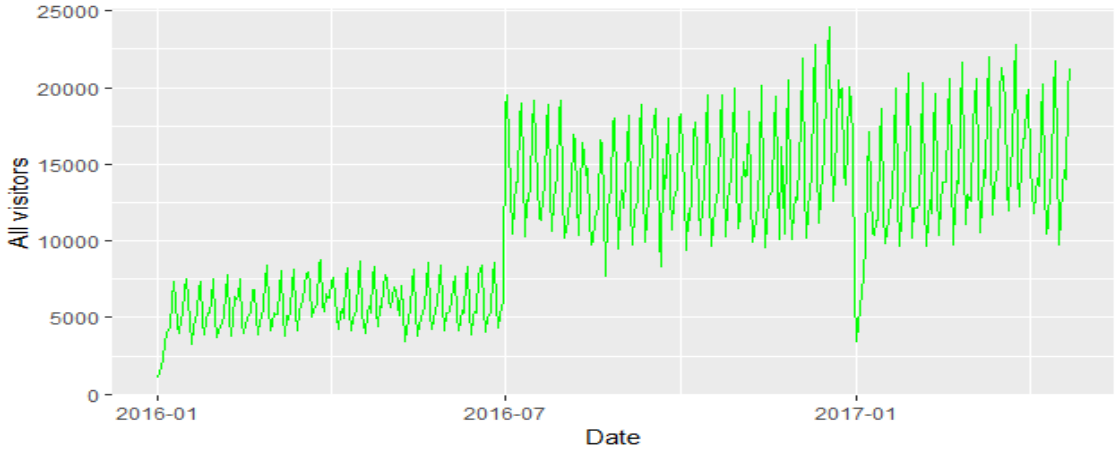
	calendar_date	day_of_week	holiday_flg
1	2016-01-01	Friday	1
2	2016-01-02	Saturday	1
3	2016-01-03	Sunday	1
4	2016-01-04	Monday	0
5	2016-01-05	Tuesday	0
6	2016-01-06	Wednesday	0

4.0 Exploratory Data Analysis

I tried exploring the data with different plots and came up with some conclusions. My exploration of data with their plots and the conclusions I could make are mentioned below.

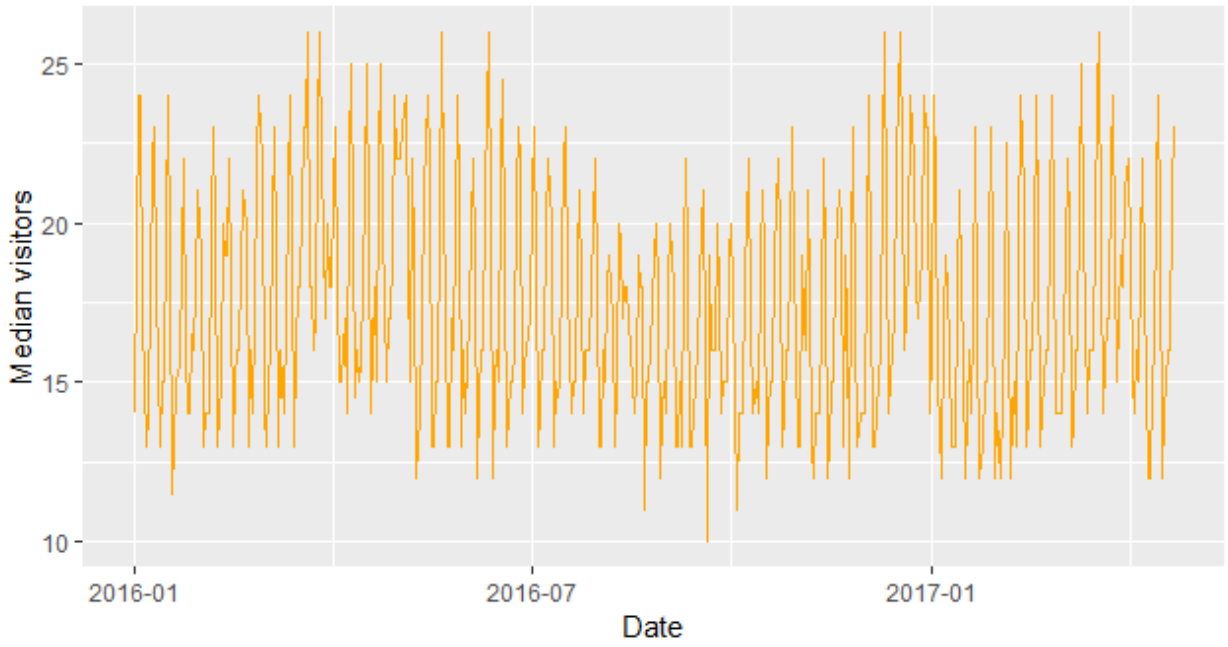
4.1 Air Visits Data:

A plot of total visitors for all restaurants on each day:



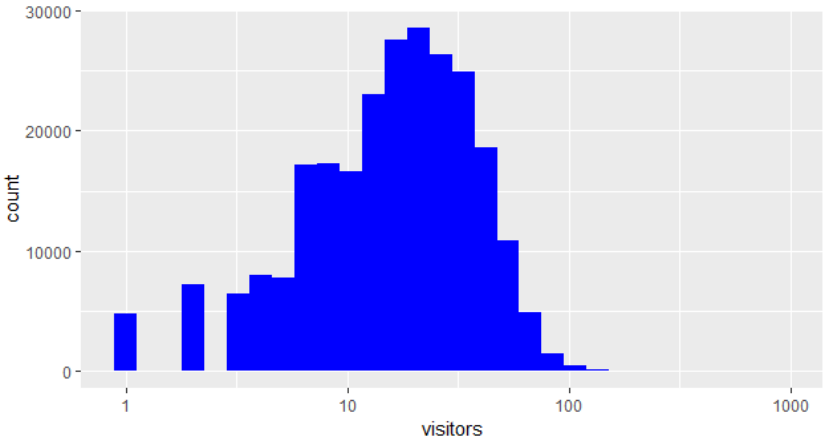
- We see a step in the total number of visitors from 2016-07, this might be because new restaurants might be added from this time period to the Air database.

A plot of the Median number of visitors for all restaurants on each day:



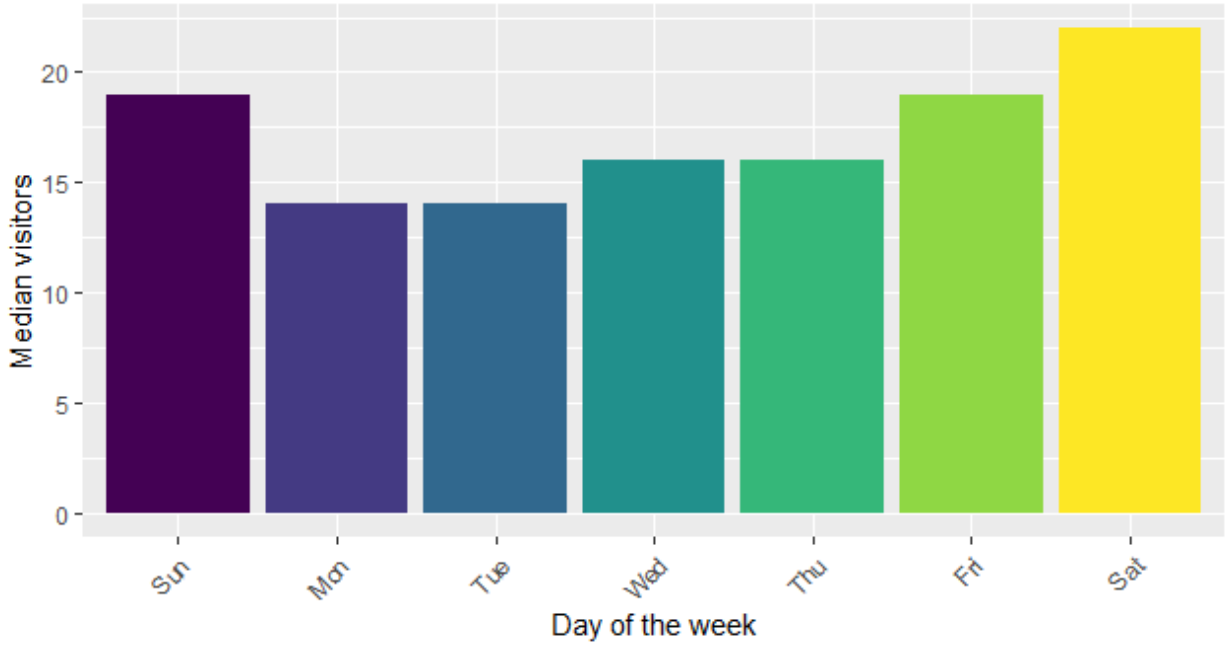
- Here we see both weekly and monthly seasonality.
- We see a flat trend here.

Histogram of visitors:



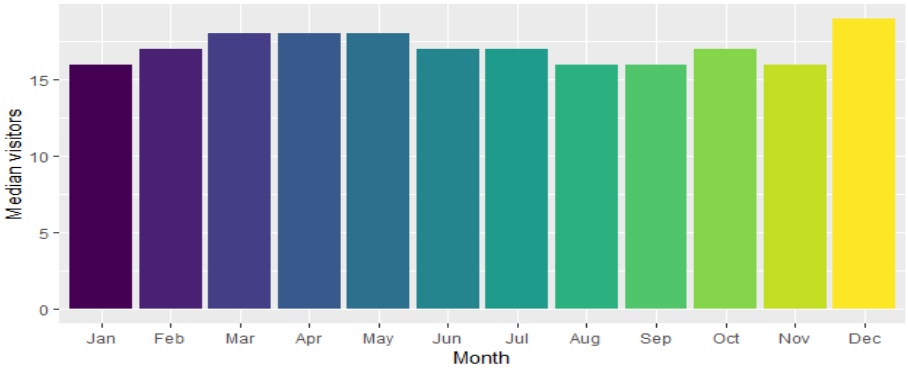
- From the above plot, we see that the median is somewhere around 20 visitors per day.

Bar graph of median visitors on each day of the week:



- We see a clear weekly seasonality.
- We find that the number of visitors is high on weekends and low on Monday and Tuesday which is expected.

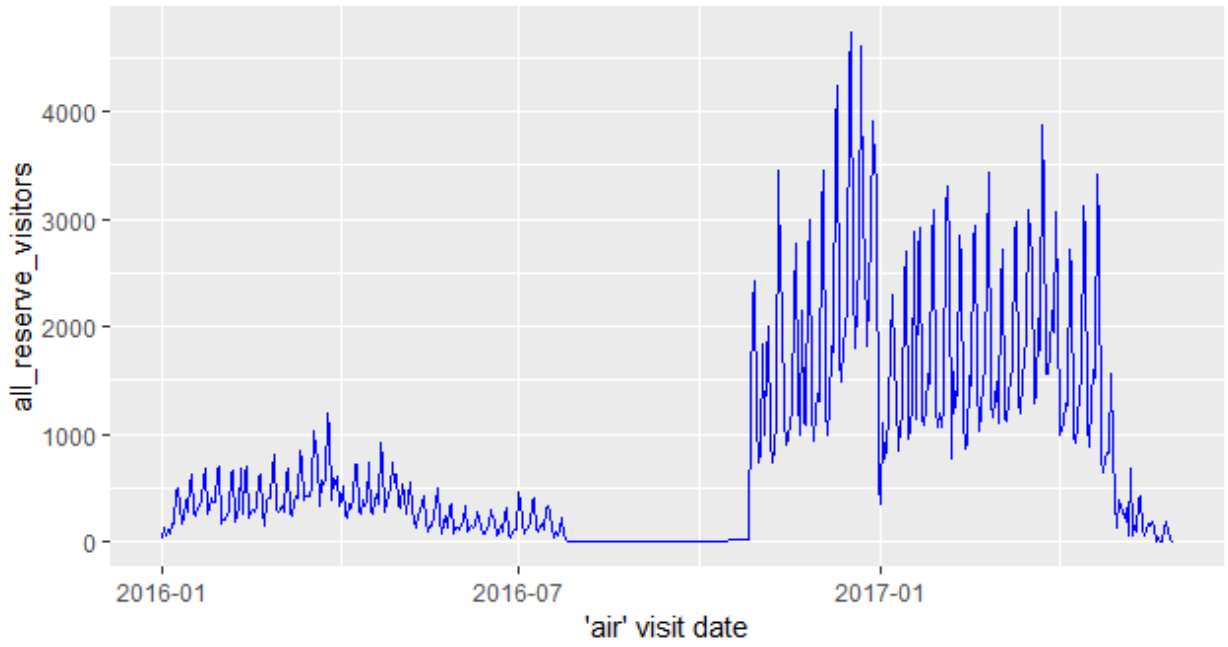
Bar graph of median visitors each month of the year:



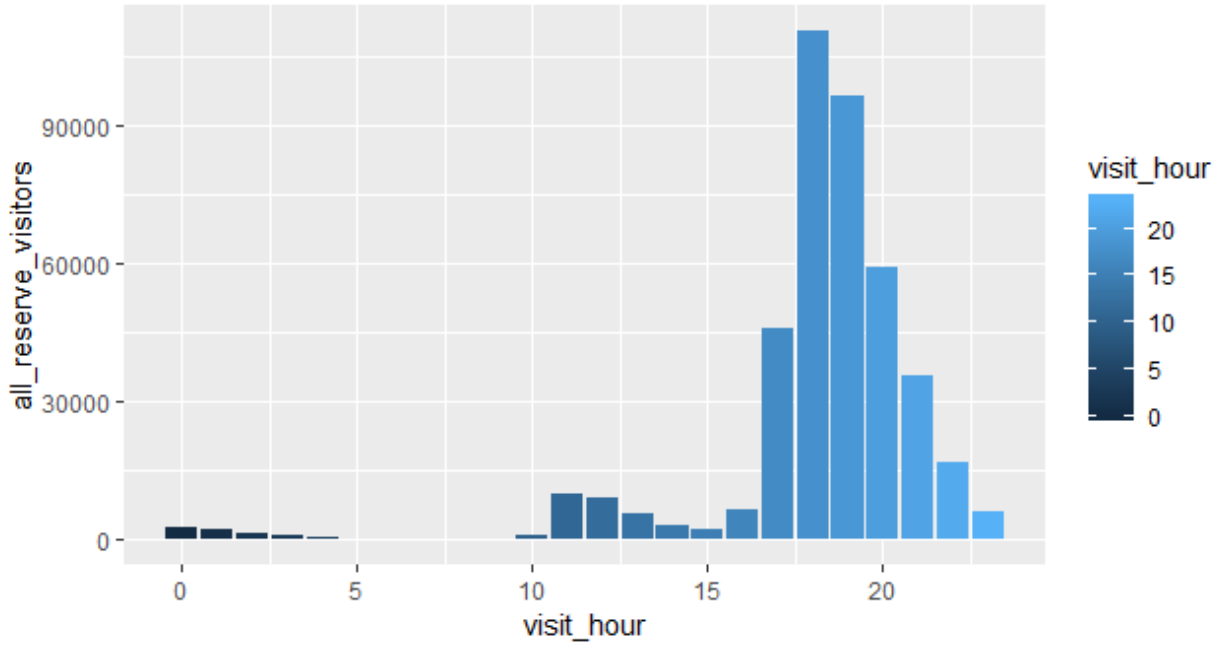
- We find that the number of visitors is high in December. We also see that visitors in March, April and May are also high.

4.2 Air Reserve Data:

The plot of total reservations for all restaurants on each day:

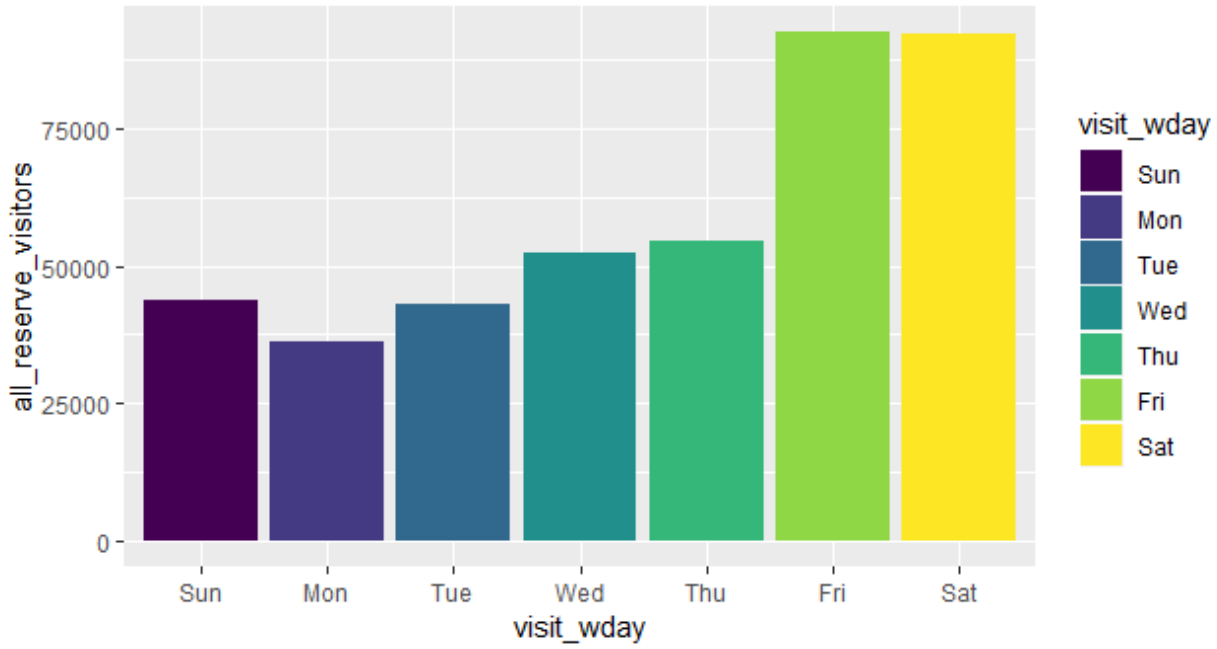


The plot of the number of reservations for each hour in a day:



- We see a greater number of reservations are in the evening times which is for dinner.

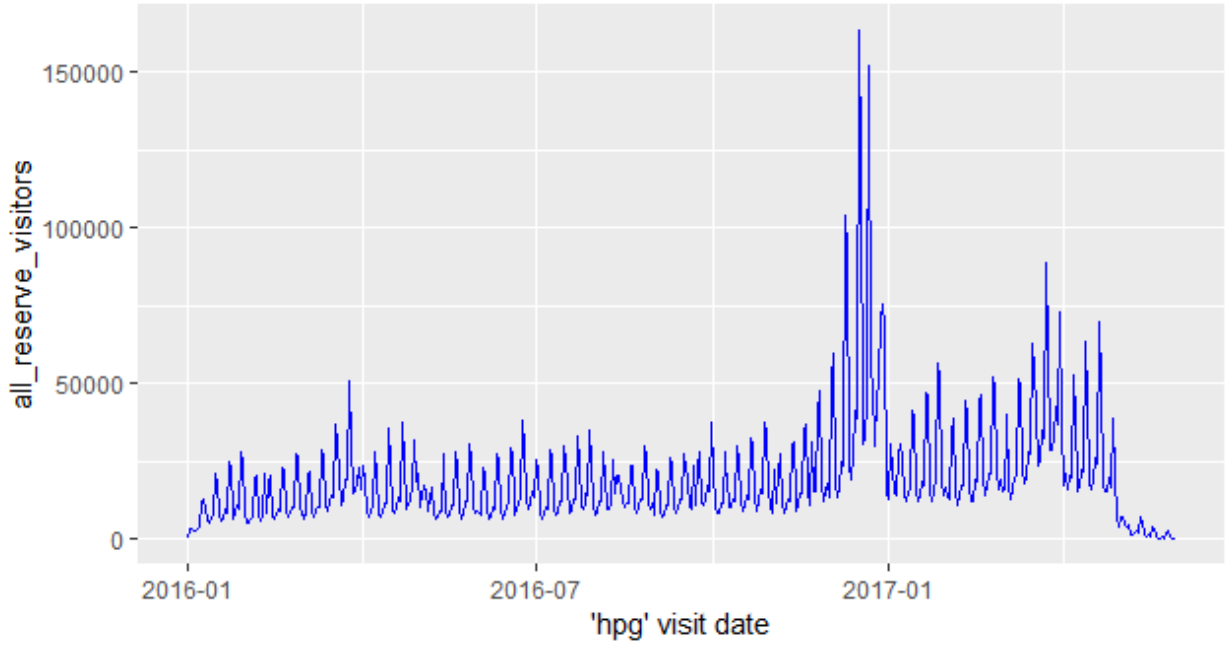
The plot of the number of reservations for each day of the week:



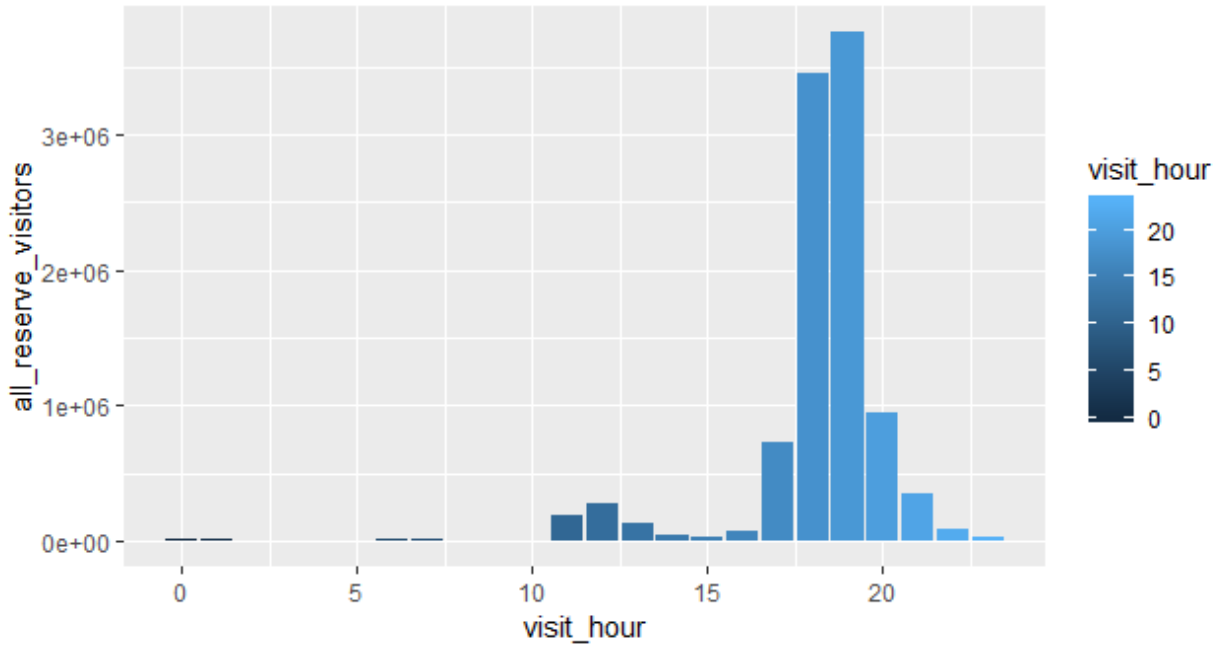
- We see a high number of reservations on Friday and Saturday. The number on Friday and Saturday visitors are almost the same.

4.3 HPG Reserve Data:

The plot of total reservations for all restaurants on each day:

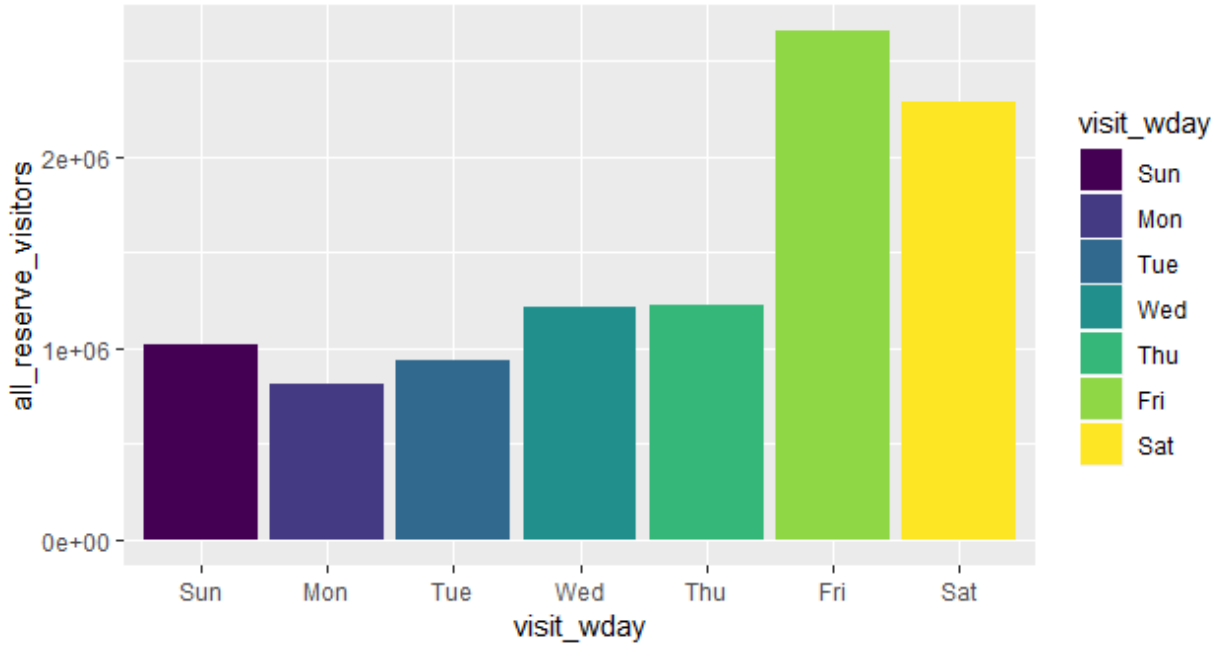


The plot of the number of reservations for each hour in a day:



- We see a greater number of reservations are in the evening times.

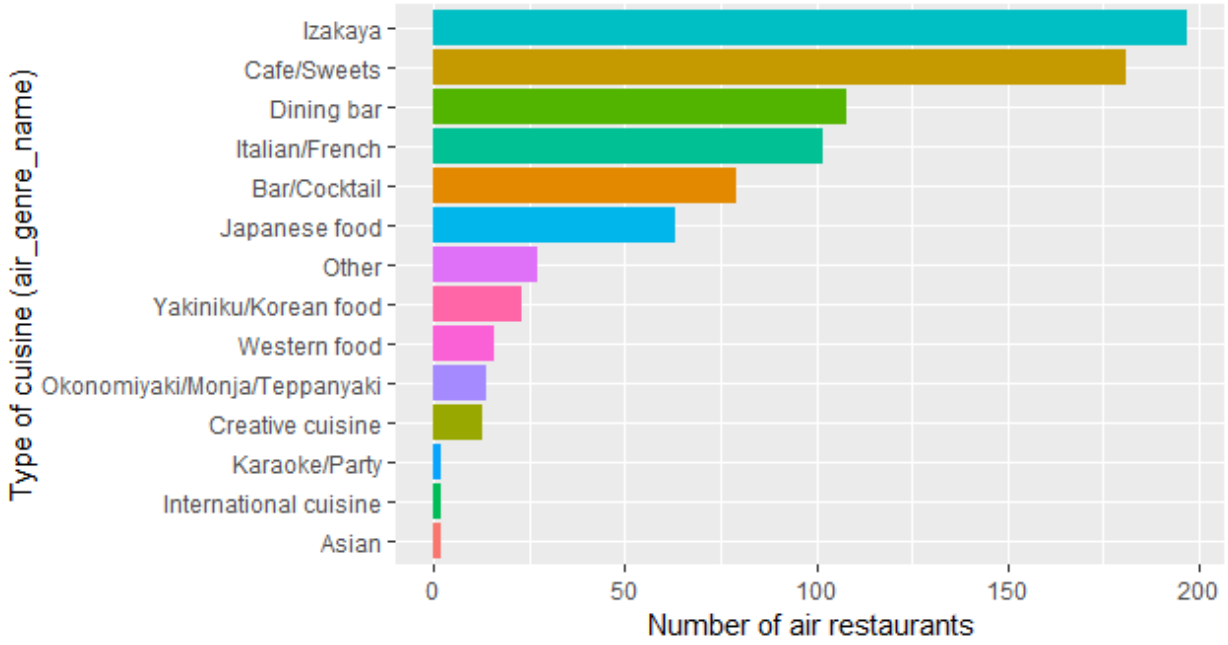
The plot of the number of reservations for each day of the week:



- We see a high number of reservations on Friday and next is Saturday. Unlike Air stores where number reservations are the same on Friday and Saturday, in hpg Friday has more reservations than Saturday.

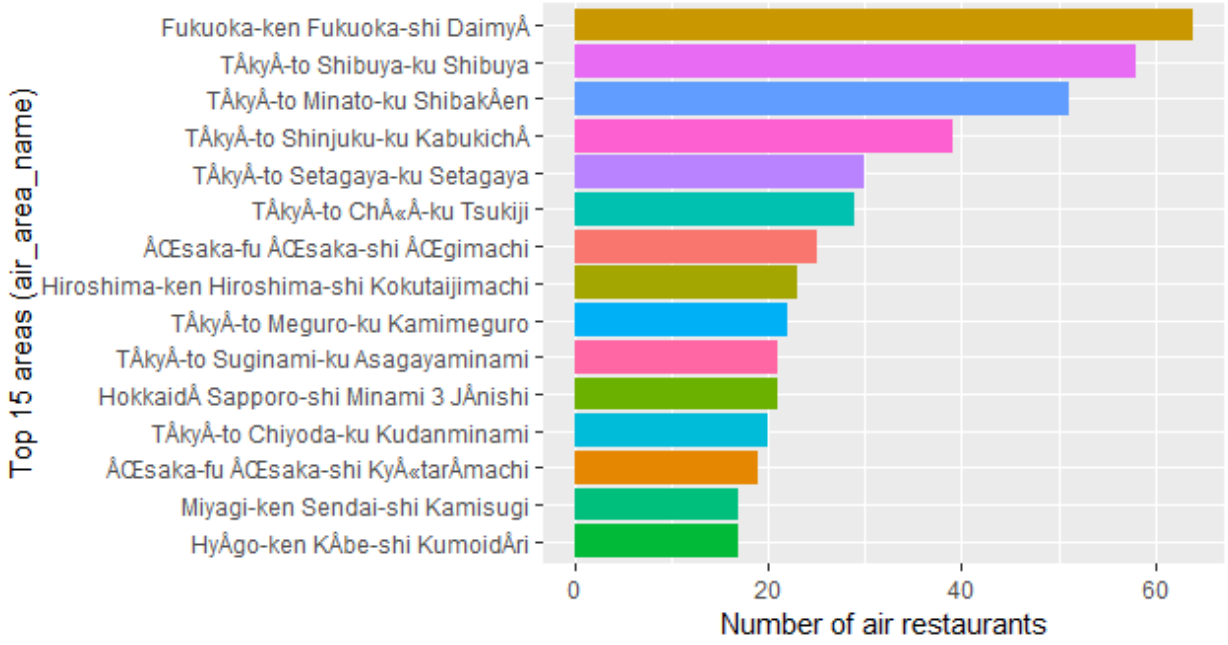
4.4 AIR Store Data:

The plot of Number of air restaurants for each genre:



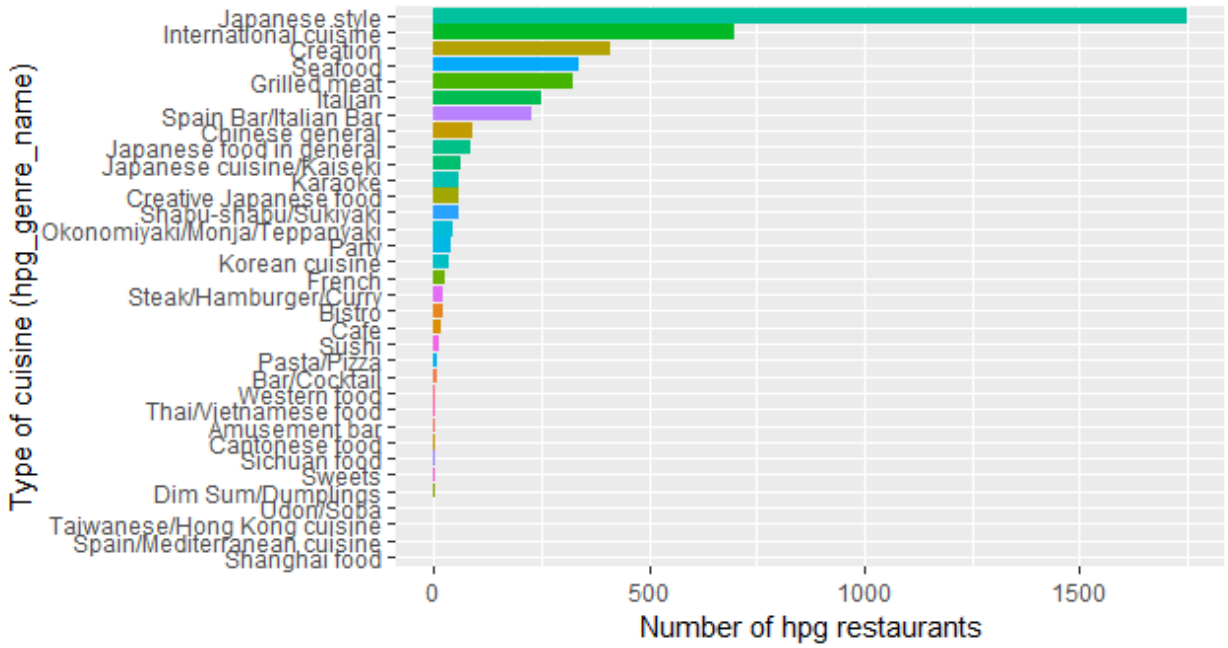
- We see that Izakaya & Café/Sweets type of restaurants are more in number in air data.

The plot of Top 15 areas with a high number of air restaurants:

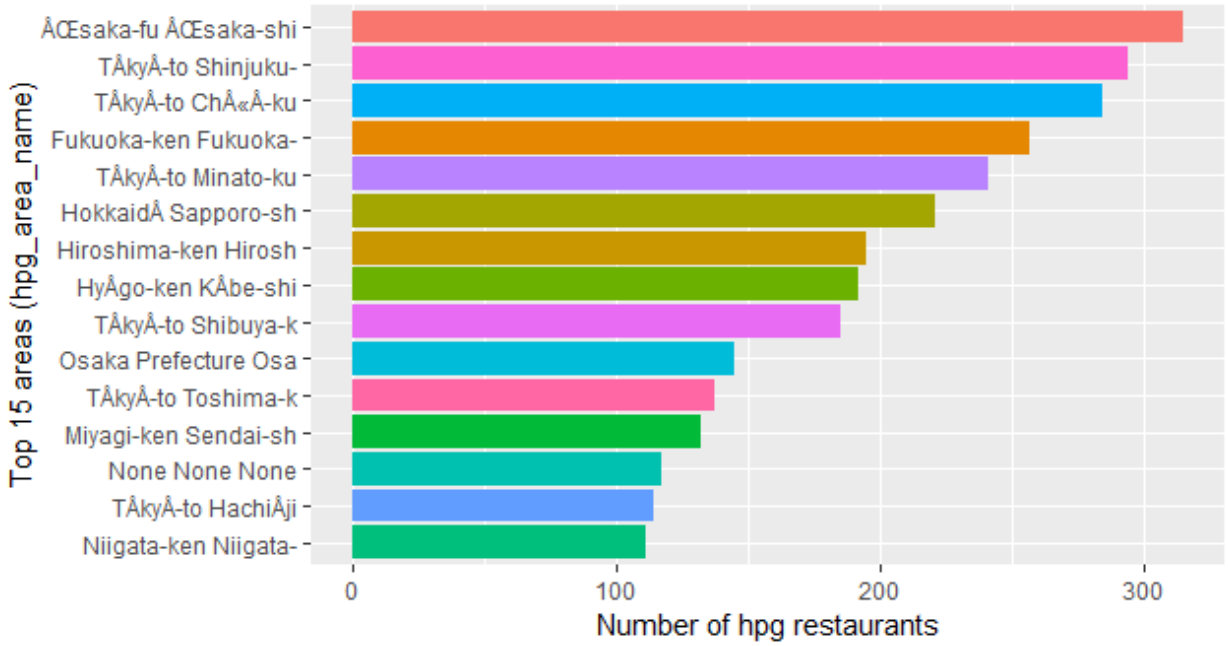


4.5 HPG Store Data:

The plot of Number of HPG restaurants for each genre:

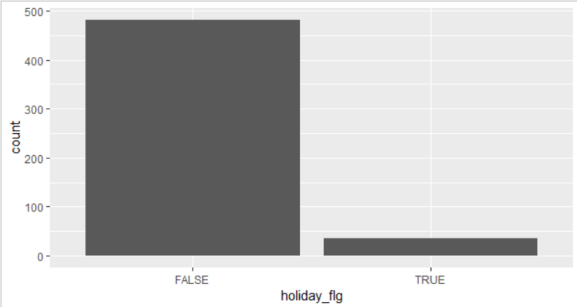


The plot of Top 15 areas with a high number of hpg restaurants:

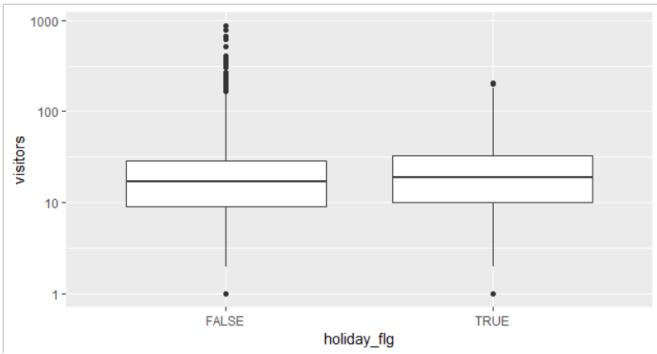


4.6 Date Info:

The plot of number holidays in a given period:

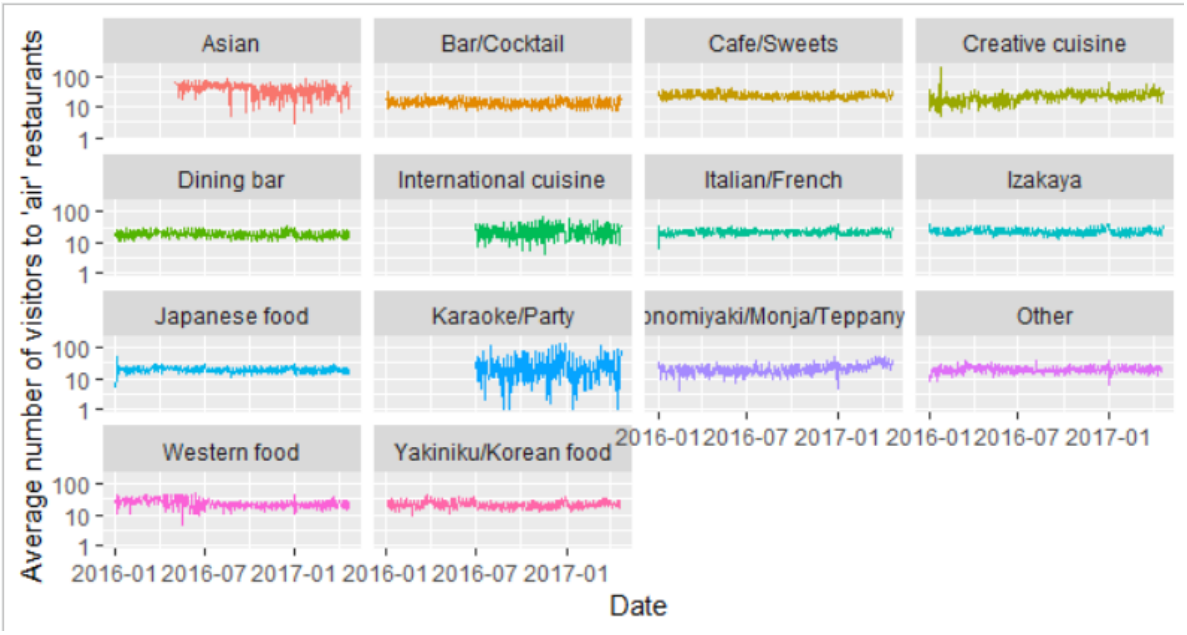


Boxplot showing the distribution of visitors on holidays and non-holidays:



- From the plot above we see that the median number of visitors on holidays and non-holidays is almost the same.
- The upper limit for the number of visitors on holidays is less than non-holidays which is an interesting point to investigate.

Time series plots of visitors for different restaurant genres:



- Almost all genres have a constant trend. Only Creative cuisine and Okonomiyaki have a slight growth trend.
- Asian cuisine & Karaoke/Party have noise in the data.

5.0 Time Series Analysis of Data for Forecasting Future Visitors number

From the past Air restaurants visit data provided, we know that we had data for 829 different restaurants for the period from 2016-01-01 to 2017-04-22, which means we have 829 different time series to analyze. The objective of this project is to predict the number of visitors these restaurants would have in the next 39 days. Since analyzing all the 829 restaurants individually would be hard, I thought to pick a method which can give me different models for all 829 restaurants automatically and have low RMSE for all the restaurants combinedly.

I tried building models using the below-mentioned methods and compared their performance to pick the method which gives least RMSE.

1) ARIMA
2) ETS
3) PROPHET forecasting Tool

5.1 ARIMA

ARIMA is a short form for Autoregressive integrated moving average. ARIMA has 3 parameters p, d, q as building blocks. To estimate these parameters individually for each time series is hard, so I used auto.arima model with stepwise = FALSE and approximation = FALSE to find out these parameters automatically.

5.1.1 Feature Engineering:

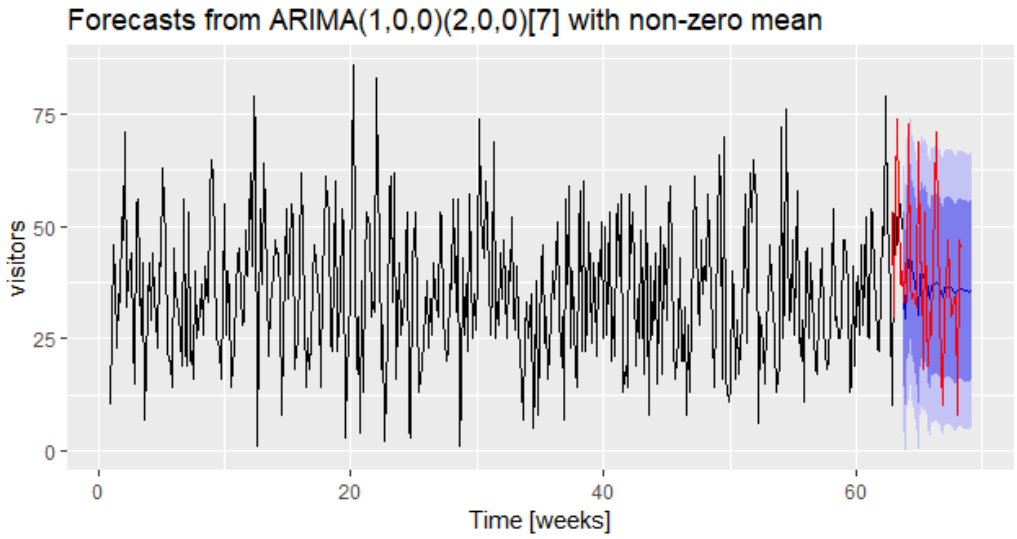
Before running Auto ARIMA model I did below mention feature engineering.

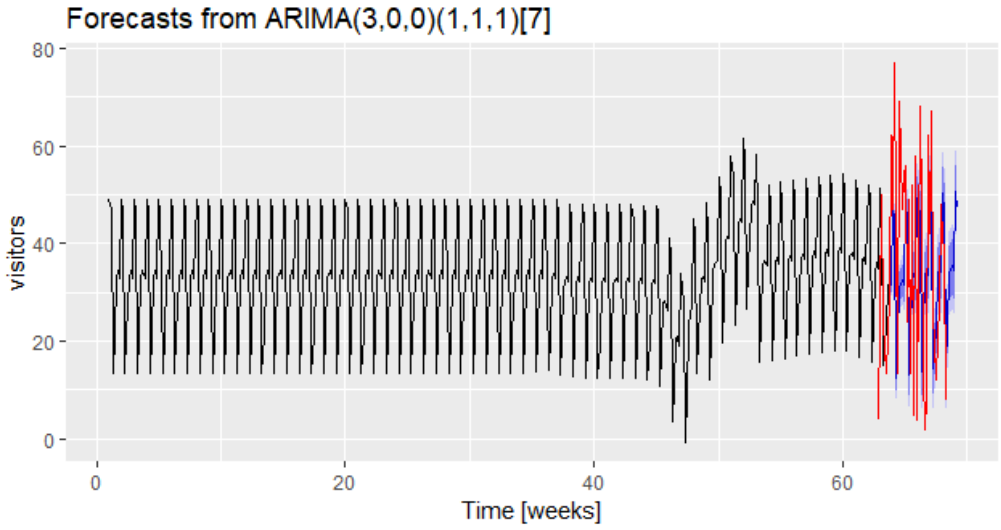
1) Since there are many NA values in time series of each restaurant, I imputed these NA values by taking the median number of visitors that restaurant had on those particular days of the week (i.e. Monday or Tuesday or Wednesday, etc.).
2) After imputing for missing values, I split the data into train and validation set with the length of the validation set same as the completion holder asked i.e. for 39 days.

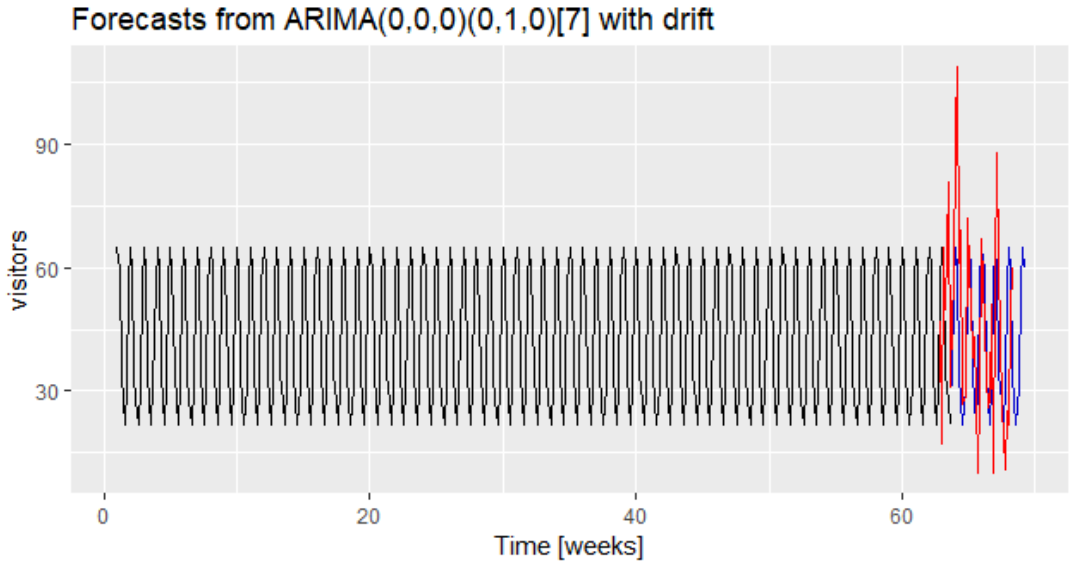
5.1.2 Example of running Auto ARIMA on few Restaurant time series:

Before running auto.arima on all the 829-time series I thought to run it on few time-series and check how they perform before running it on all. I ran ARIMA on 3 restaurants which have no NA values, medium number of NA values and a High number of NA values.

The plots of these 3-time series with predictions of the model over validation time in blue color and original values of validation in red are present below.


Forecasts from ARIMA(1,0,0)(2,0,0)[7] with non-zero mean
visitors
75
50
25
0
0
20
40
60
Time [weeks]


Forecasts from ARIMA(3,0,0)(1,1,1)[7]
visitors
80
60
40
20
0
0
20
40
60
Time [weeks]


Forecasts from ARIMA(0,0,0)(0,1,0)[7] with drift
visitors
90
60
30
0
20
40
60
Time [weeks]

From the above plots, we see that these models are predicting well for immediate time-periods and the predictions are not well for farther periods.

5.1.3 Running Auto ARIMA on all the 829 Time series:

I ran Auto Arima on all the 829-time series on my computer and it took me 4hours to complete. After running on all the Time series, I calculated the RMSE and found it to be 11.76298.

RMSE = 11.76298

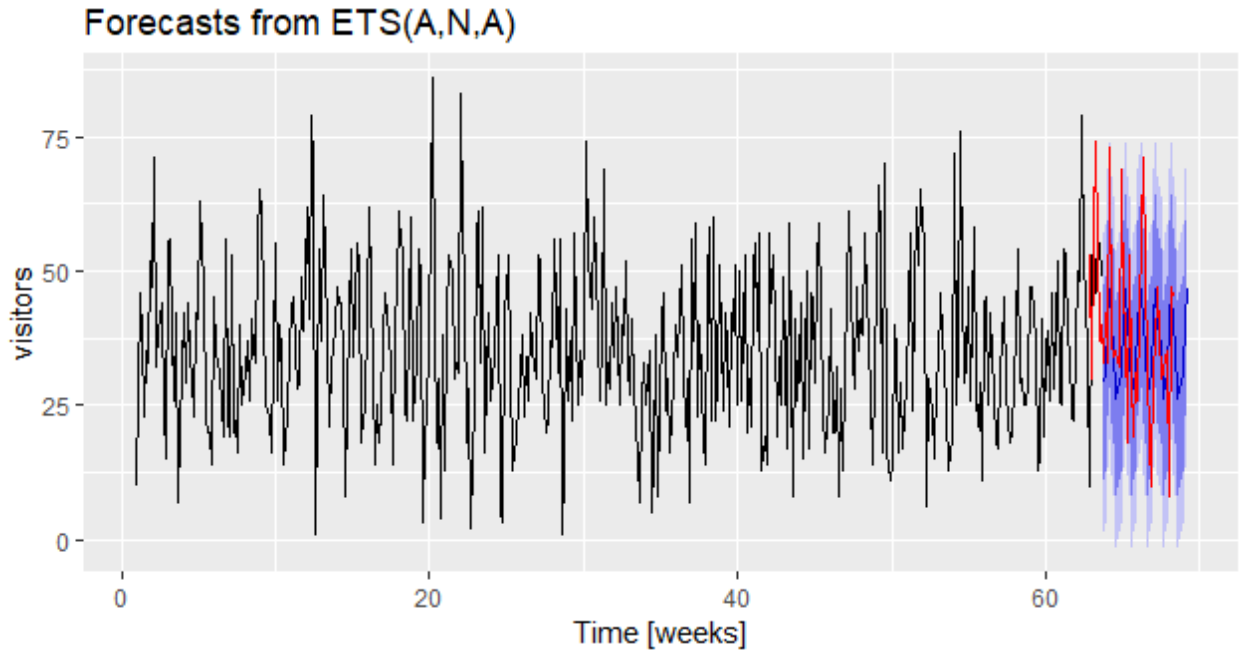
5.2 ETS Forecasting

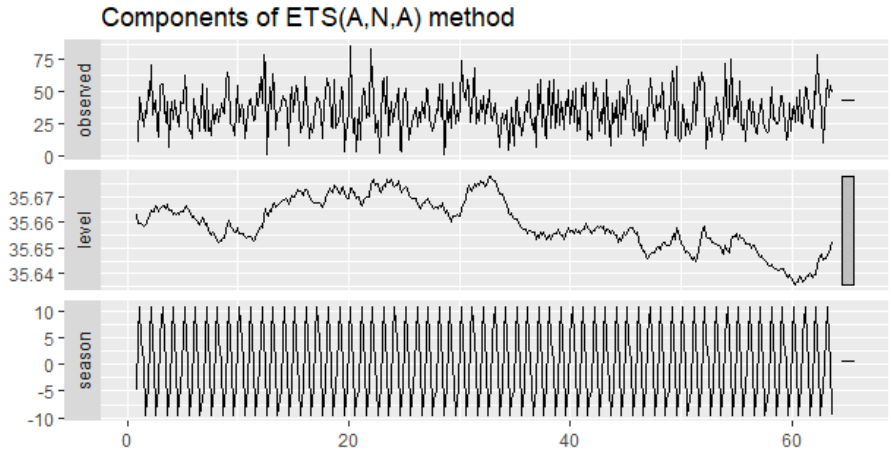
I tried Exponential Smoothing Methods using ETS in R as it automatically fixes the parameters. I Similarly implemented ETS models as ARIMA. I did the same feature engineering and test train split as in ARIMA.

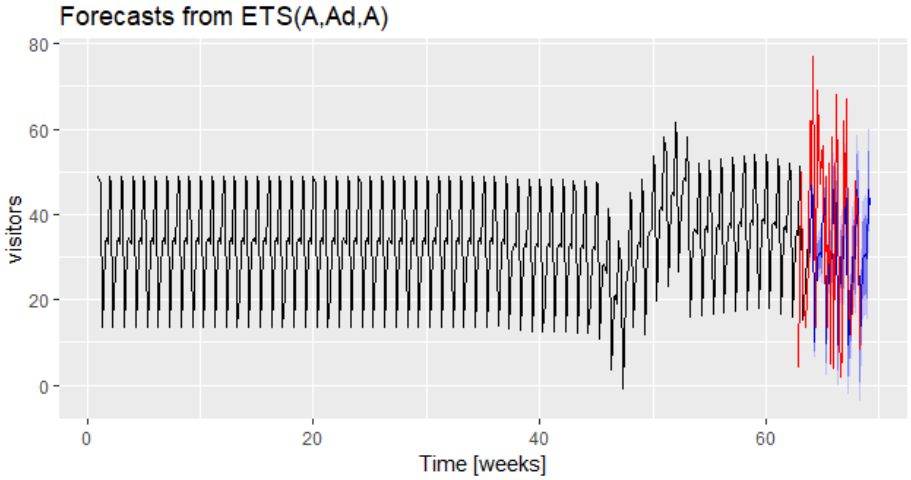
5.2.1 Example of running ETS on few Restaurant time series:

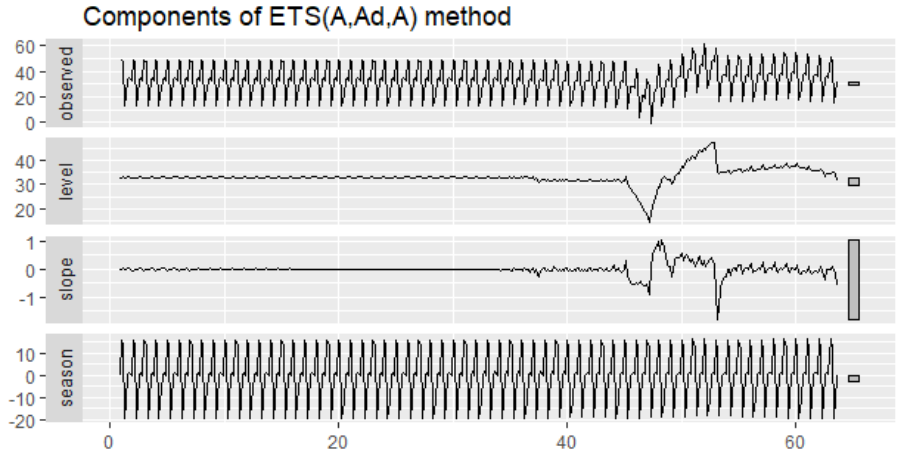
Before running ETS on all the 829-time series I thought to run it on few time-series and check how they perform before running it on all. I ran ETS on 3 restaurants which have no NA values, medium number of NA values and a High number of NA values.

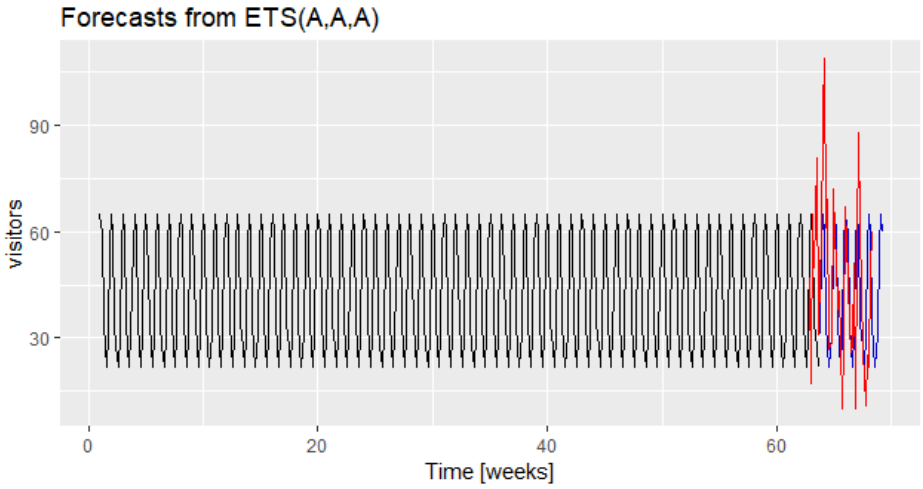
The plots of these 3-time series with predictions of the model over validation time in blue color and original values of validation in red are present below. I also included the decomposition of these time series below.

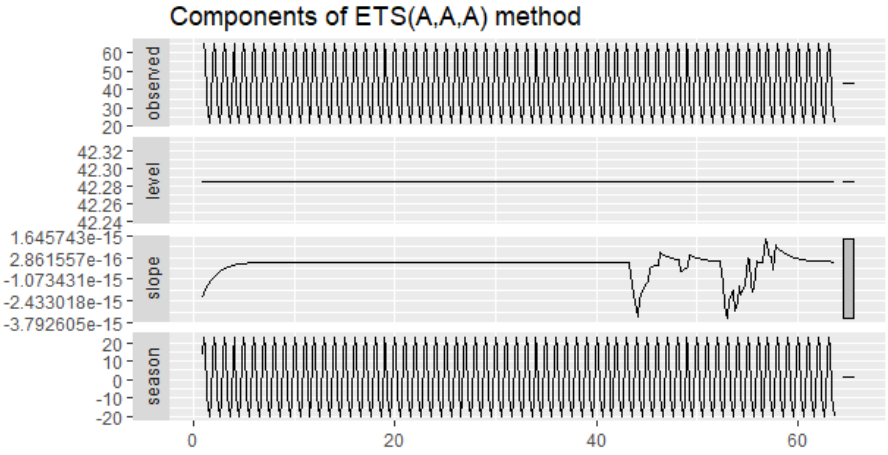



Components of ETS(A,N,A) method
observed
level
season
75
50
25
0
35.67
35.66
35.65
35.64
10
5
0
-5
-10
0
20
40
60


Forecasts from ETS(A,Ad,A)
visitors
80
60
40
20
0
0
20
40
60
Time [weeks]


Components of ETS(A,Ad,A) method
observed
level
slope
season
60
40
20
0
40
30
20
1
0
-1
10
0
-10
-20
0
20
40
60





From the above plots, we see that these models are a little better than ARIMA.

5.2.2 Running ETS on all the 829 Time series:

I ran ETS on all the 829-time series on my computer and it took me 15 minutes to complete. After running on all the Time series, I calculated the RMSE and found it to be 11.76298.

RMSE_ETS = 11.17782

5.3 Prophet Forecasting

The prophet is an Open source forecasting tool developed by Facebook. The advantages of it over ARIMA and ETS are

1) It handles NA values and outliers on its own automatically.
2) It gives us the possibility to include holidays and other special events in our forecast.

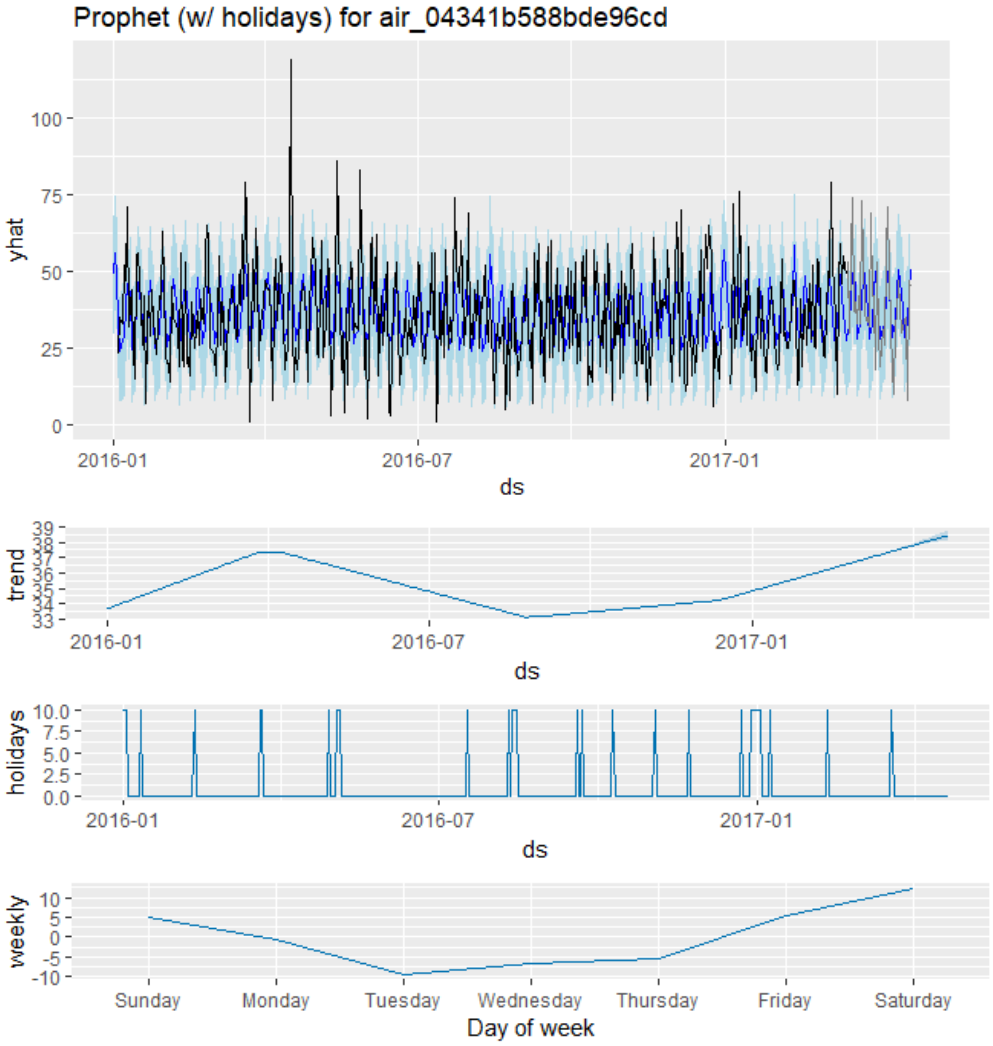
Since Prophet helps us include holidays in forecasting, and we have holidays in our data set, I included them in our modeling which was not possible with ARIMA and ETS. I didn't do any feature

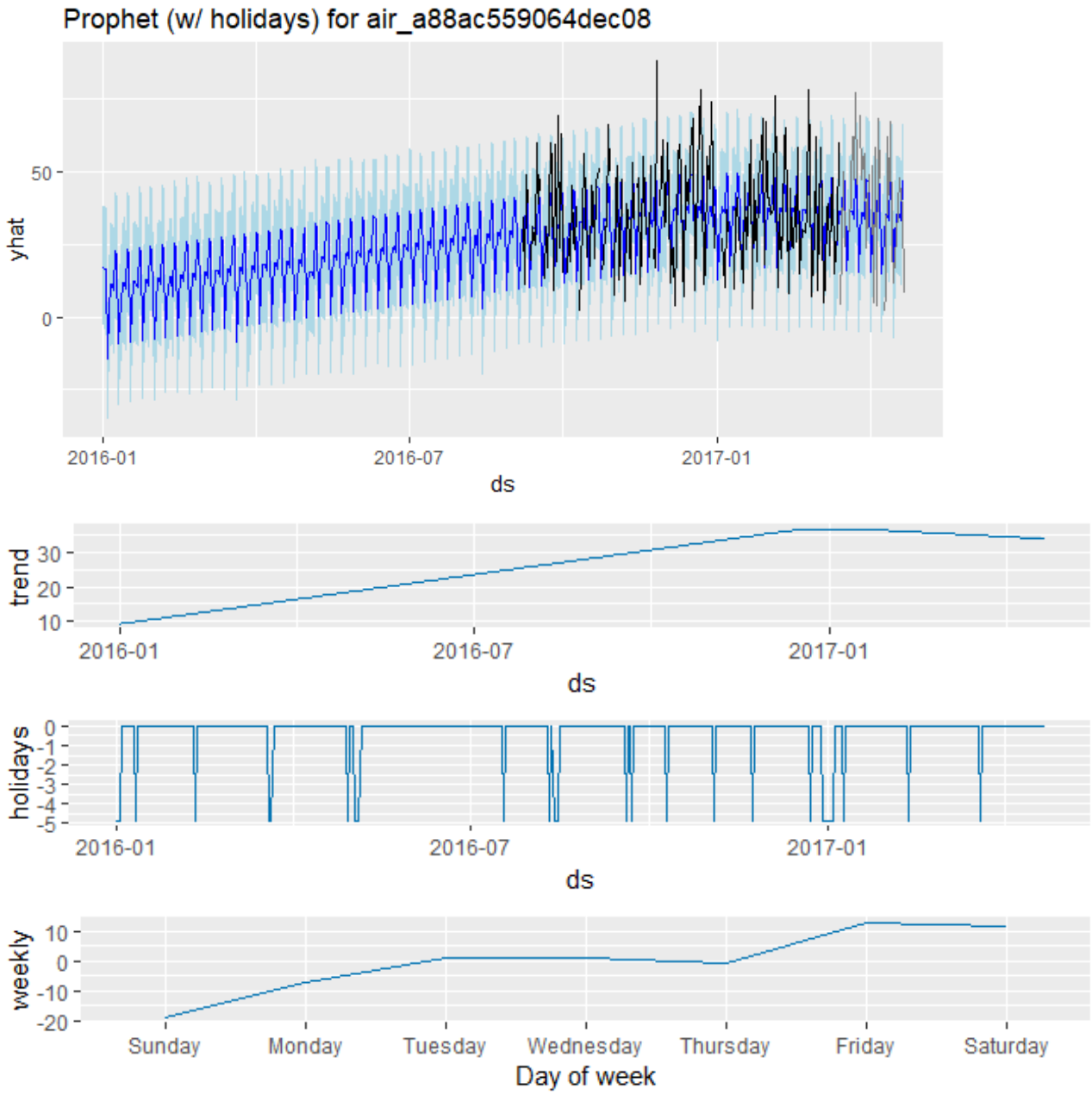
engineering to remove or replace NA's as Prophet does it. I converted the data set into the prophet input format and did test-train split.

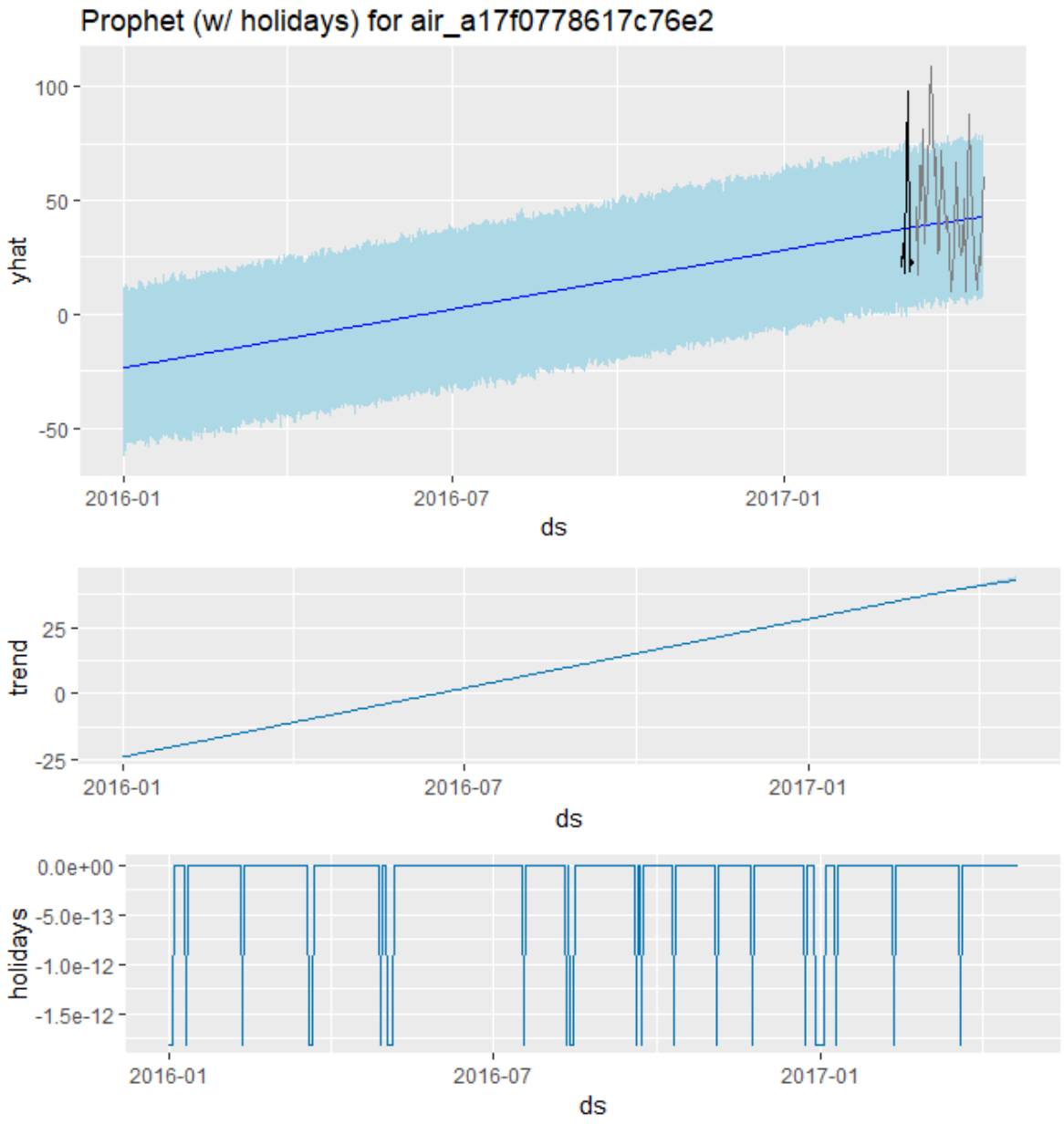
5.3.1 Example of running Prophet on few Restaurant time series:

Before running Prophet on all the 829-time series, I thought to run it on few time-series and check how they perform before running it on all. I ran Prophet on 3 restaurants which have no NA values, medium number of NA values and a High number of NA values.

The plots of these 3-time series with predictions of the model over validation time and original values of validation are present below. I also included the decomposition of these time series by Prophet are present below.




Prophet (w/ holidays) for air_a88ac559064dec08
yhat
50
0
2016-01
2016-07
2017-01
ds
trend
30
20
10
2016-01
2016-07
2017-01
ds
holidays
0
-1
-2
-3
-4
-5
2016-01
2016-07
2017-01
ds
weekly
10
0
-10
-20
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of week



From the above plots of the first two restaurants, we see that Prophet is doing a good job in recognizing the weekly seasonality. Prophet didn't recognize the Weekly seasonality for the 3rd restaurant because it has only 3 points in the training set.

5.3.2 Running Prophet on all the 829 Time series:

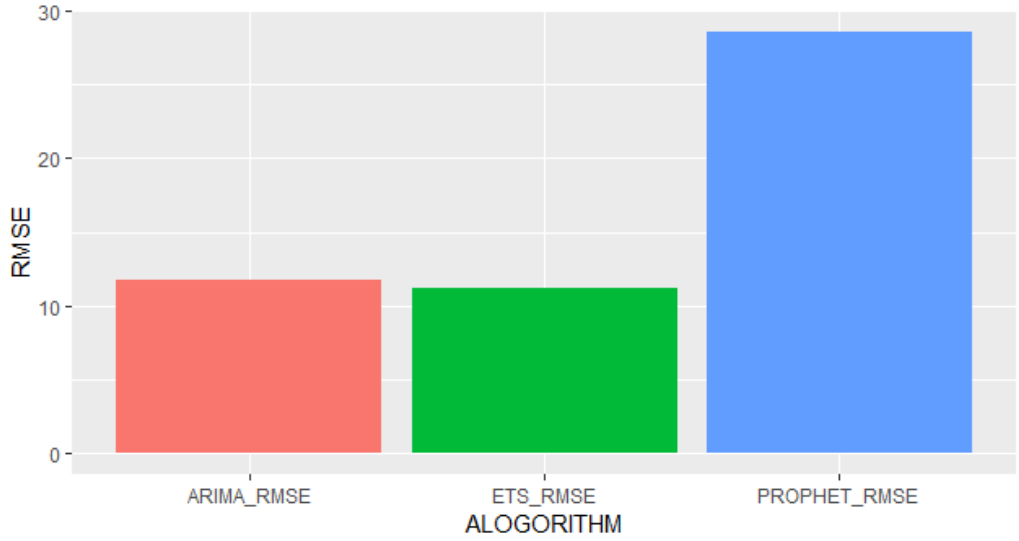
I ran Prophet on all the 829-time series on my computer and it took me 30 minutes to complete. After running on all the Time series, I calculated the RMSE and found it to be 28.5878.

RMSE_Prophet = 28.5878

5.4 ARIMA Vs ETS Vs Prophet

For comparing ARIMA, ETS & Prophet performance on all the time series I used RMSE as the measure.

Plot to compare them is present below.



From the plots above we find that RMSE of ARIMA and ETS are nearly similar but ETS is lesser when compared it exactly. So, we choose the models generated by **ETS** for forecasting the future number of visitor's restaurants would have.

6.0 Things that can be done to improve the performance of forecasting

Prediction can be improved by building a single global model instead of 829 different models for each restaurant. This model can be built by incorporating all the features of the restaurant, location, competition, and reservation data in it. We can build many engineered features which are specific to time series and put all these features in a booted regression Trees or neural nets and get a final global model which can predict the future customers.

This single model will have advantages like

1) Fast in prediction in real-time.
2) It will be able to predict even for new restaurants though it doesn't have historical data just by using the restaurant characteristics, over the other approach.

7.0 References

1. Link to Kaggle competition:
https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data
2. Other online References used:
https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/discussion
https://www.kaggle.com/headsortails/be-my-guest-recruit-restaurant-eda
https://otexts.com/fpp2/expsmooth.html
https://otexts.com/fpp2/arima.html
https://facebook.github.io/prophet/docs/quick_start.html#r-api