1 Introduction

The aim of this challenge is to predict the future numbers of restaurant visitors. This makes it a Time Series Forecasting problem. 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.

- **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.
- **sample_submission.csv**: serves as the *test* set. The *id* is formed by combining the *air* id with the visit date.

2 Overview: File structure and content

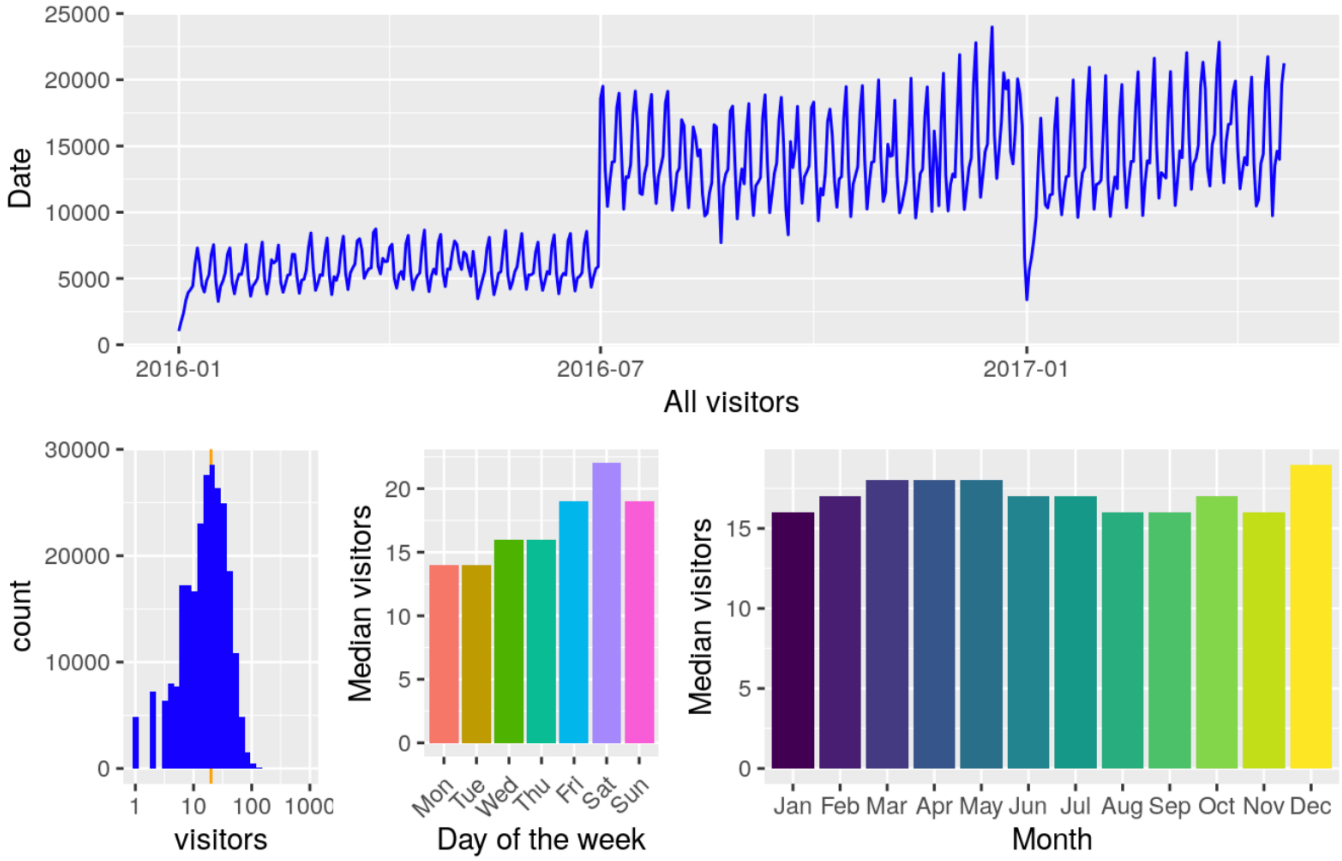
```
> air_visits
# A tibble: 252,108 x 5
   air_store_id        visit_date visitors wday  month
   <chr>               <date>        <dbl> <ord> <ord>
 1 air_ba937bf13d40fb24 2016-01-13       25 三    " 1"
 2 air_ba937bf13d40fb24 2016-01-14       32 四    " 1"
 3 air_ba937bf13d40fb24 2016-01-15       29 五    " 1"
 4 air_ba937bf13d40fb24 2016-01-16       22 六    " 1"
 5 air_ba937bf13d40fb24 2016-01-18        6 一    " 1"
 6 air_ba937bf13d40fb24 2016-01-19        9 二    " 1"
 7 air_ba937bf13d40fb24 2016-01-20       31 三    " 1"
 8 air_ba937bf13d40fb24 2016-01-21       21 四    " 1"
 9 air_ba937bf13d40fb24 2016-01-22       18 五    " 1"
10 air_ba937bf13d40fb24 2016-01-23       26 六    " 1"
# … with 252,098 more rows
```

```
> air_reserve
# A tibble: 92,378 x 12
   air_store_id visit_datetime      reserve_datetime
   <chr>        <dttm>              <dttm>
 1 air_877f797… 2016-01-01 19:00:00 2016-01-01 16:00:00
 2 air_db4b38e… 2016-01-01 19:00:00 2016-01-01 19:00:00
 3 air_db4b38e… 2016-01-01 19:00:00 2016-01-01 19:00:00
 4 air_877f797… 2016-01-01 20:00:00 2016-01-01 16:00:00
 5 air_db80363… 2016-01-01 20:00:00 2016-01-01 01:00:00
 6 air_db80363… 2016-01-02 01:00:00 2016-01-01 16:00:00
 7 air_db80363… 2016-01-02 01:00:00 2016-01-01 15:00:00
 8 air_3bb99a1… 2016-01-02 16:00:00 2016-01-02 14:00:00
 9 air_3bb99a1… 2016-01-02 16:00:00 2016-01-01 20:00:00
10 air_2b8b29d… 2016-01-02 17:00:00 2016-01-02 17:00:00
# … with 92,368 more rows, and 9 more variables:
```

```
> air_store
# A tibble: 829 x 9
   air_store_id air_genre_name air_area_name prefecture latitude
   <chr>        <fct>          <fct>         <chr>         <dbl>
 1 air_0f0cdee… Italian/French Hyōgo-ken Kō… Hyōgo-ken      34.7
 2 air_7cc17a3… Italian/French Hyōgo-ken Kō… Hyōgo-ken      34.7
 3 air_fee8dcf… Italian/French Hyōgo-ken Kō… Hyōgo-ken      34.7
 4 air_a17f077… Italian/French Hyōgo-ken Kō… Hyōgo-ken      34.7
 5 air_83db5af… Italian/French Tōkyō-to Min… Tōkyō-to       35.7
 6 air_99c3eae… Italian/French Tōkyō-to Min… Tōkyō-to       35.7
 7 air_f183a51… Italian/French Tōkyō-to Min… Tōkyō-to       35.7
 8 air_6b9fa44… Italian/French Tōkyō-to Min… Tōkyō-to       35.7
 9 air_0919d54… Italian/French Tōkyō-to Min… Tōkyō-to       35.7
10 air_2c6c79d… Italian/French Tōkyō-to Min… Tōkyō-to       35.7
# … with 819 more rows, and 4 more variables: longitude <dbl>,
```

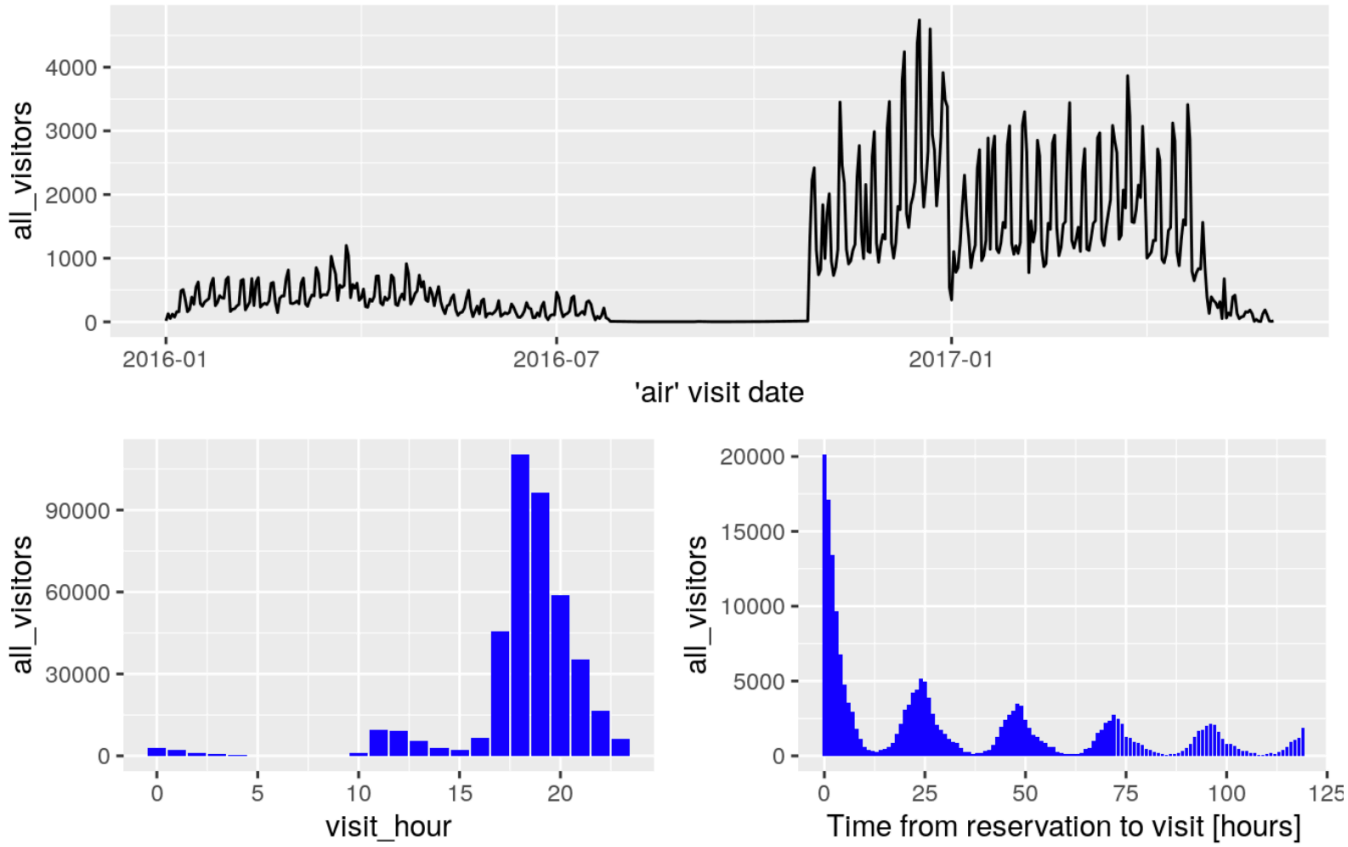
3 Individual feature visualisations

3.1 Air Visits



- There is an interesting long-term step structure in the overall time series. This might be related to new restaurants being added to the data base. In addition, we already see a periodic pattern that most likely corresponds to a weekly cycle.
- The number of guests per visit per restaurant per day peaks at around 20 (the orange line). The distribution extends up to 100 and, in rare cases, beyond.
- Friday and the weekend appear to be the most popular days; which is to be expected. Monday and Tuesday have the lowest numbers of average visitors.
- Also during the year there is a certain amount of variation. Dec appears to be the most popular month for restaurant visits. The period of Mar - May is consistently busy.

3.2 Air Reservations

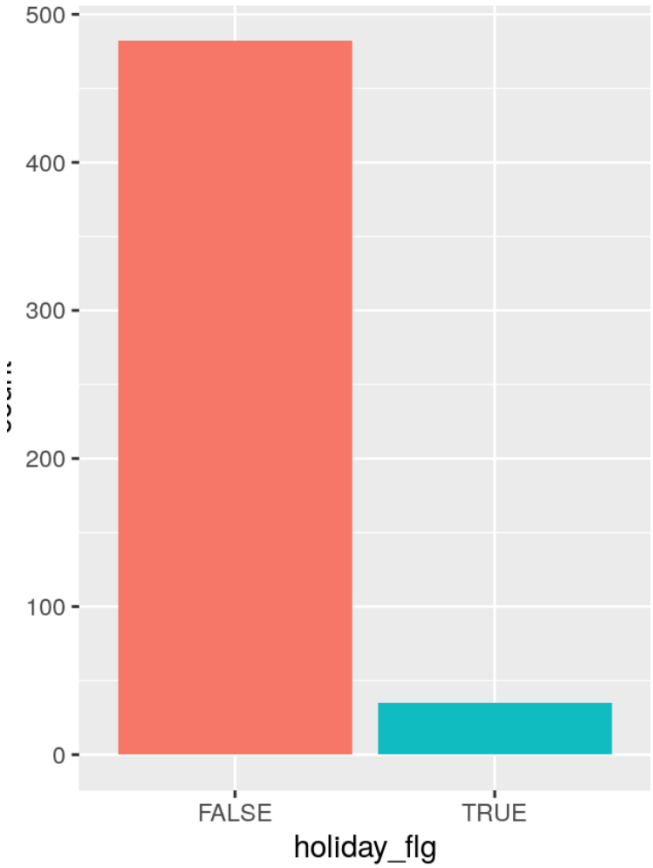


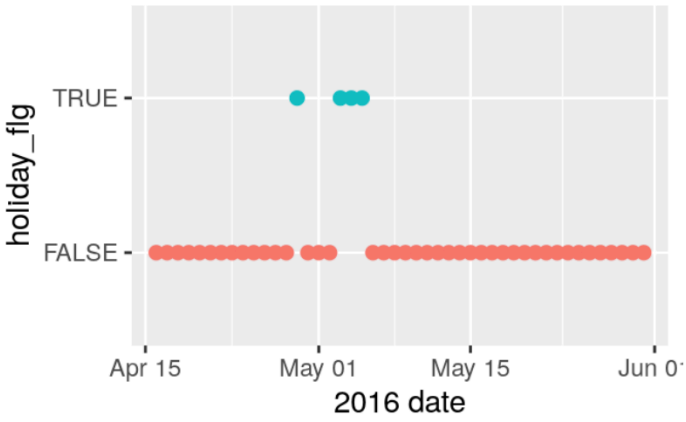
- There were much fewer reservations made in 2016 through the *air* system; even none at all for a long stretch of time. The volume only increased during the end of that year. In 2017 the visitor numbers stayed strong. The artifical decline we see after the first quarter is most likely related to these reservations being at the end of the *training* time frame, which means that long-term reservations would not be part of this data set.
- Reservations are made typically for the dinner *hours* in the evening.
- The time, here shown in hours, between making a reservation and visiting the restaurant follow a nice 24-hour pattern. The most popular strategy is to reserve a couple of hours before the visit, but if the reservation is made more in advance then it seems to be common to book a table in the evening for one of the next evenings. This plot is truncated to show this pattern, which continues towards longer time scales. Very long time gaps between reservation and visit are not uncommon. Those are the most extreme values for the *air* data, up to more than a year in advance:

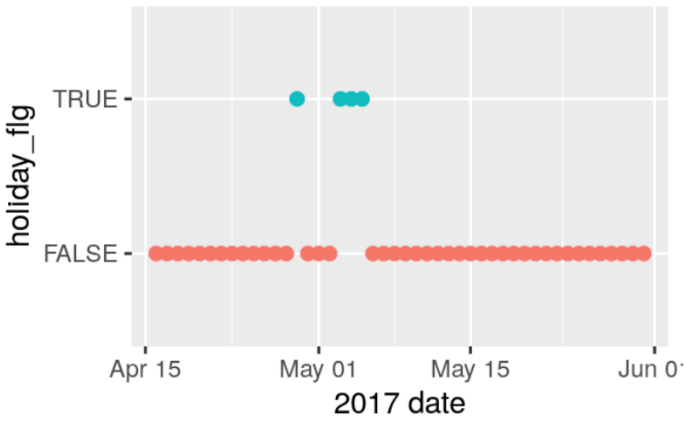
3.3 Air Store

- There are lots of Izakaya gastropubs in our data, followed by Cafe's. We don't have many Karaoke places in the *air* data set and also only a few that describe themselves as generically "International" or "Asian".
- Fukuoka has the largest number of *air* restaurants per area, followed by many Tokyo areas.

3.4 Holidays

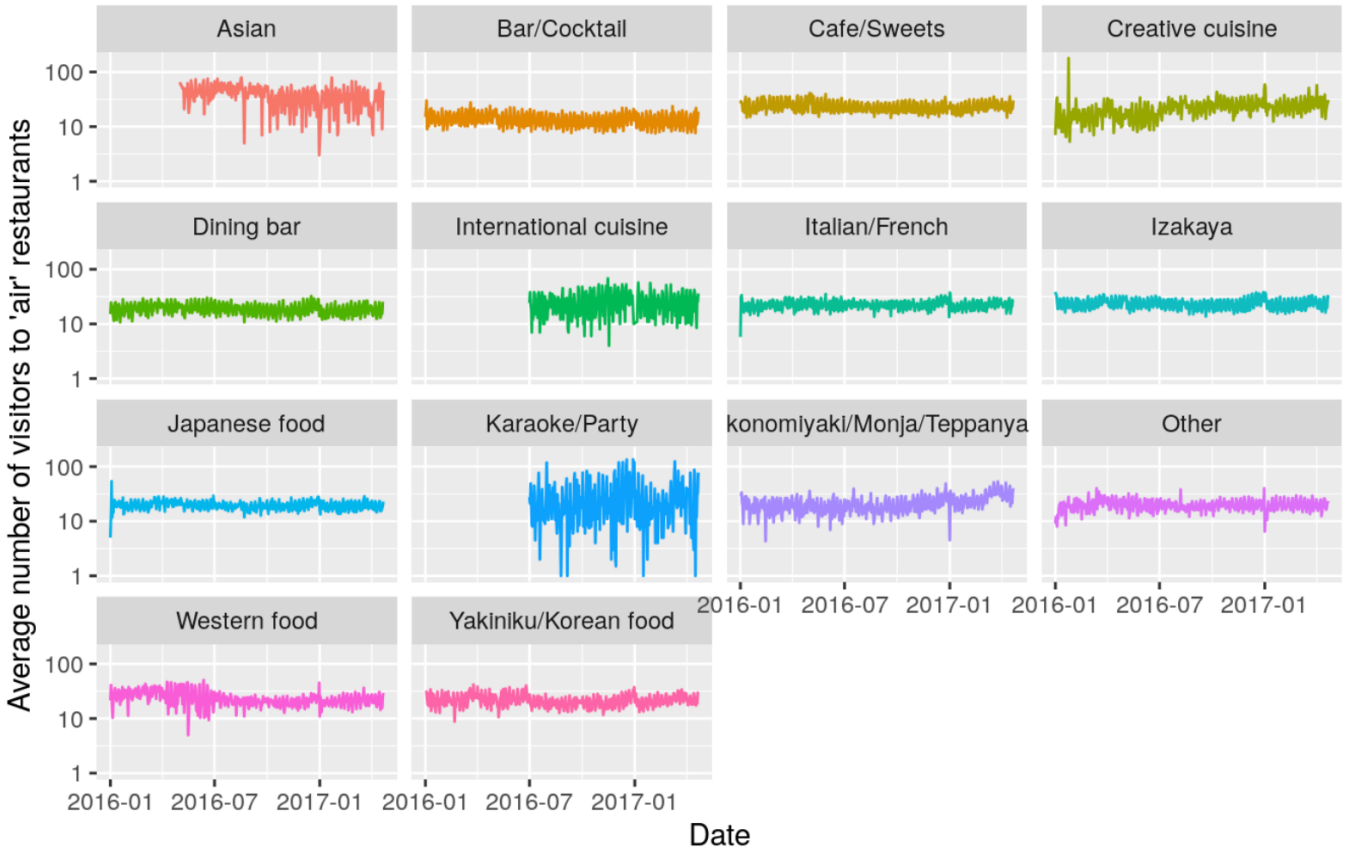




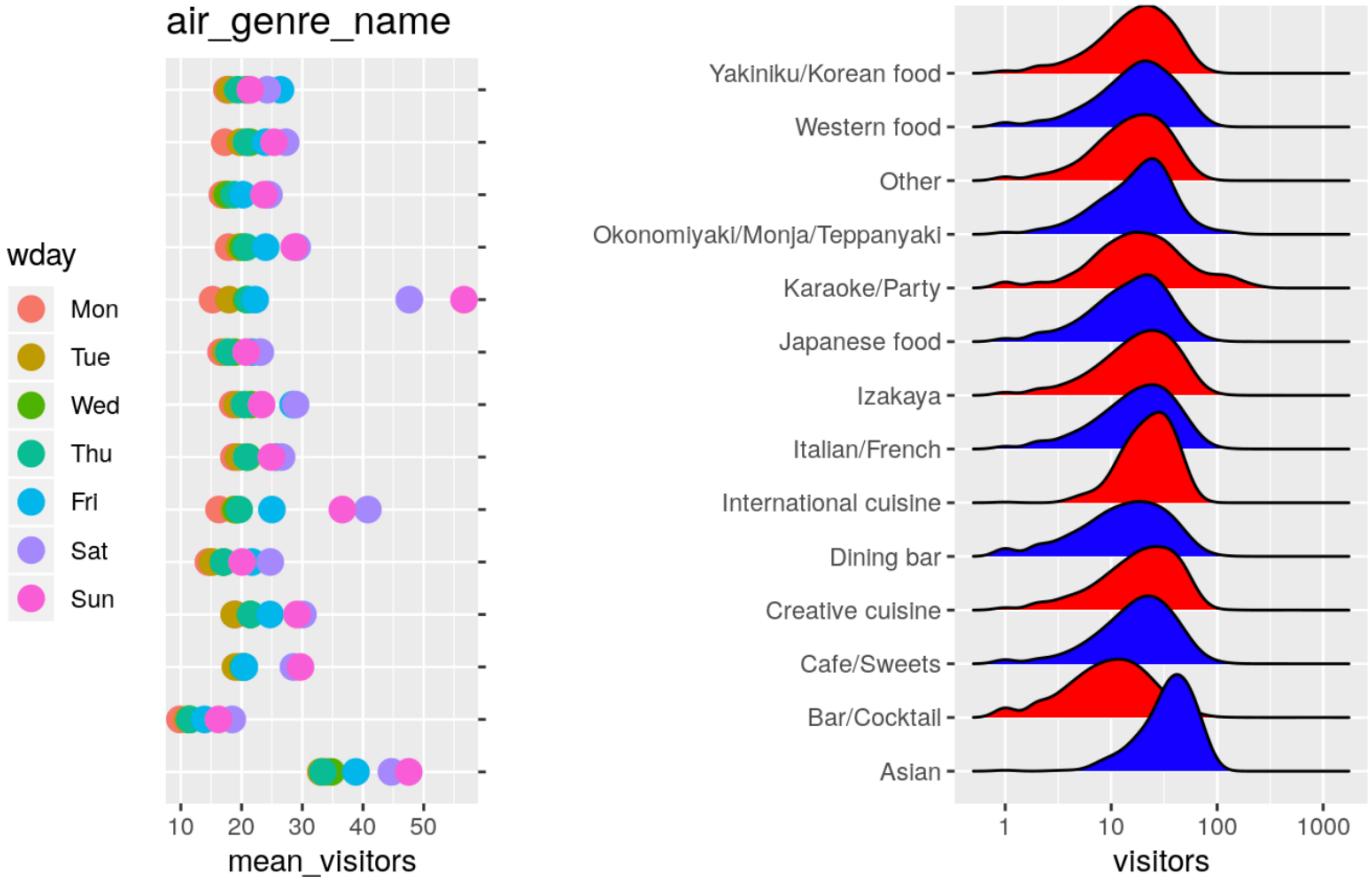


4 Feature relations

4.1 Visitors per genre

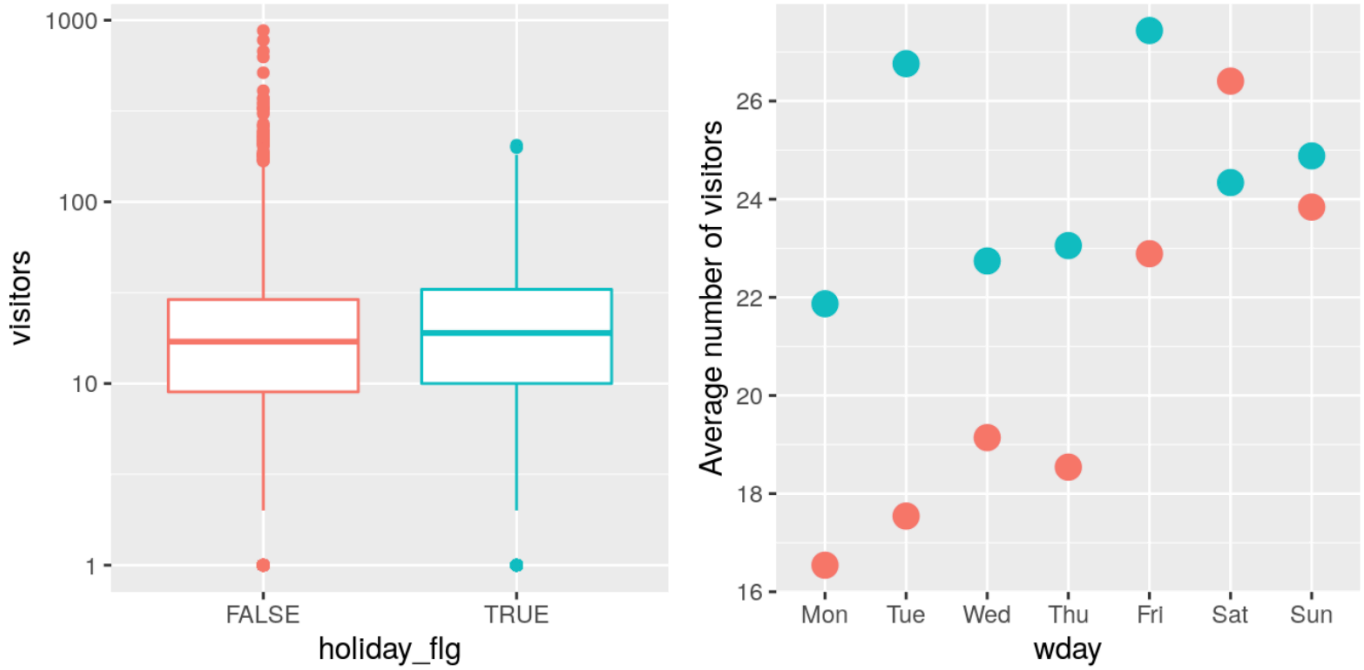


- The mean values range between 10 and 100 visitors per genre per day. Within each category, the long-term trend looks reasonably stable. There is an upward trend for “Creative Cuisine” and “Okonomiyaki” et al., while the popularity of “Asian” food has been declining since late 2016.
- The low-count time series like “Karaoke” or “Asian” (see Fig. 6) are understandably more noisy than the genres with higher numbers of visitors.



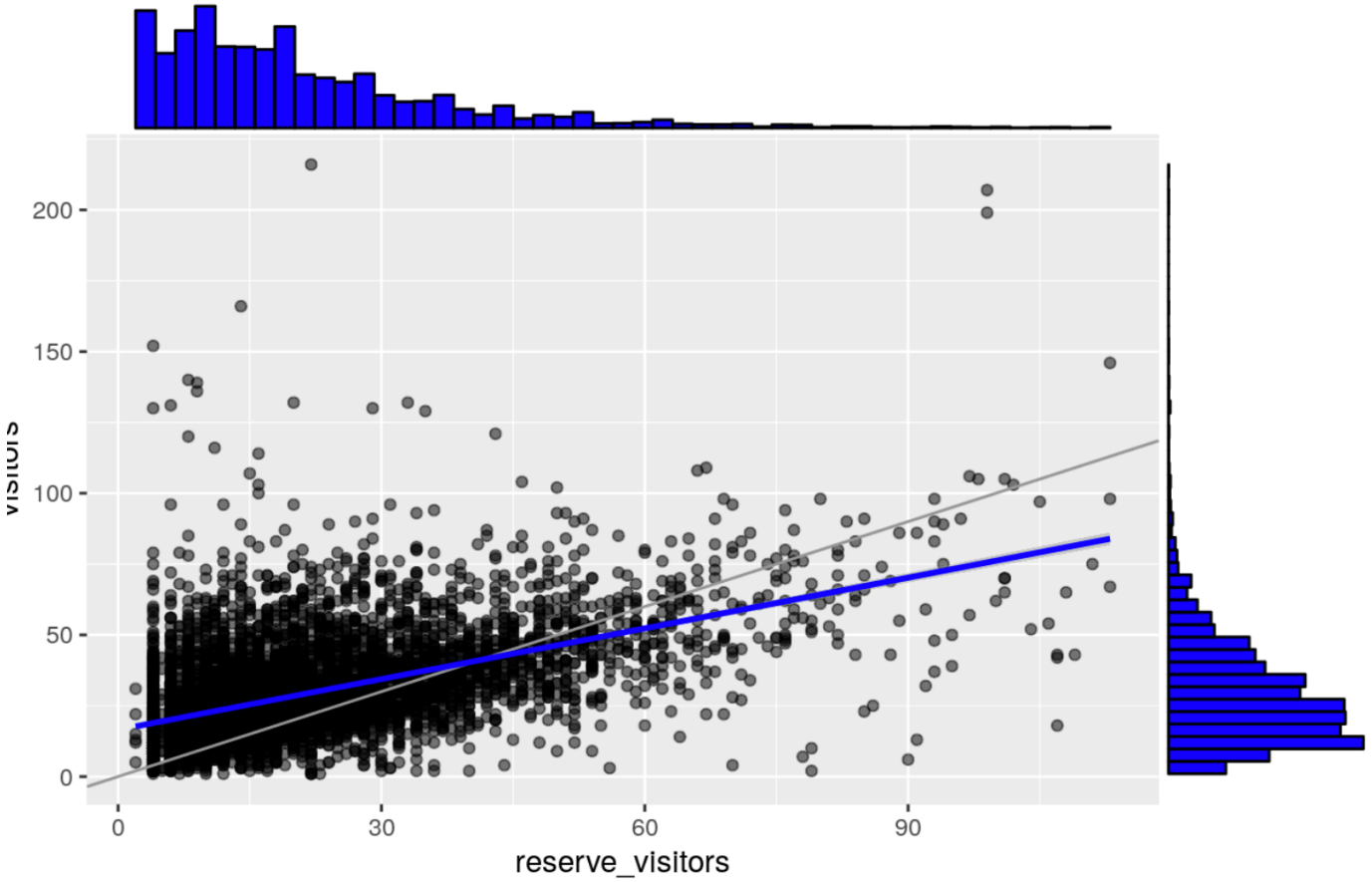
- The biggest difference between weekend and weekdays exists for the “Karaoke” bars, which rule the weekend. A similar trend, although with a considerably smaller gap, can be seen for the “International” cuisine.
- No *genre* really goes against the trend of busier weekends. The smallest variations are in the generic “Other” category, the “Japanese” food, and also the “Korean” cuisine which is the only category where Fridays are the busiest days. General “Bars/Cocktail” are notably unpopular overall.
- The density curves confirm the impression we got from the week-day distribution: the “Asian” restaurants have rarely less than 10 visitors per date and the “Karaoke” places show a very broad distribution due to the strong impact of the weekends. Note the logarithmic x-axis.

4.2 The impact of holidays

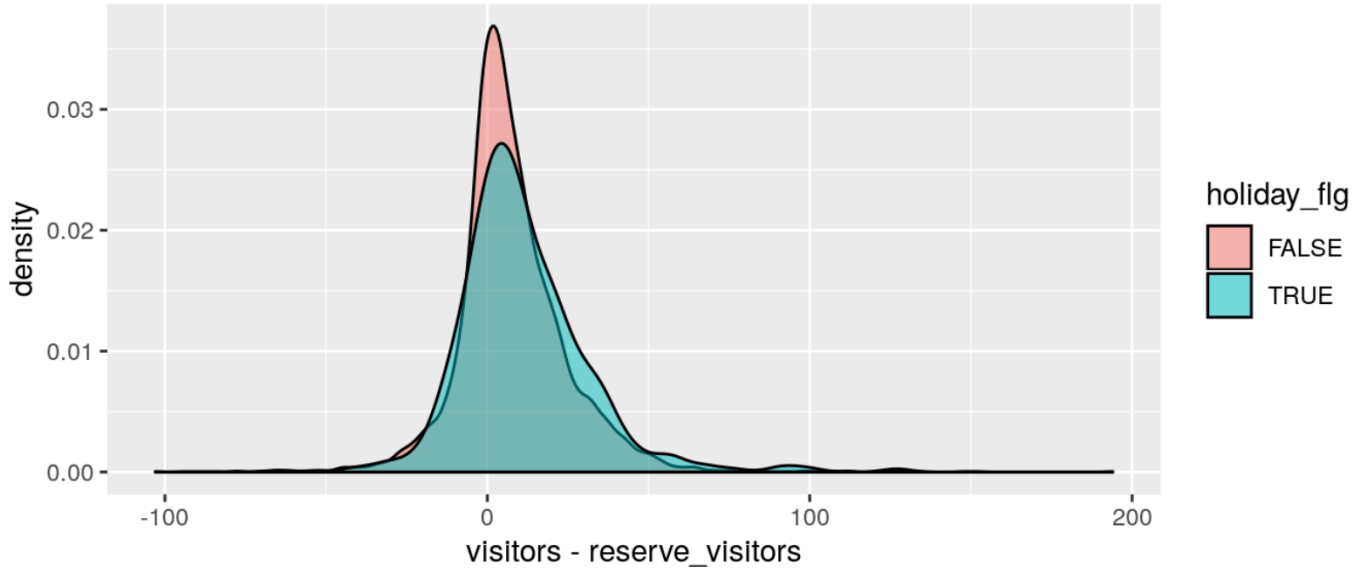


- Overall, holidays don’t have any impact on the average visitor numbers (left panel). As so often, more information is hidden in the details.
- While a weekend holiday has little impact on the visitor numbers, and even decreases them slightly, there is a much more pronounced effect for the weekdays; especially Monday and Tuesday (right panel).

4.3 Reservations vs Visits



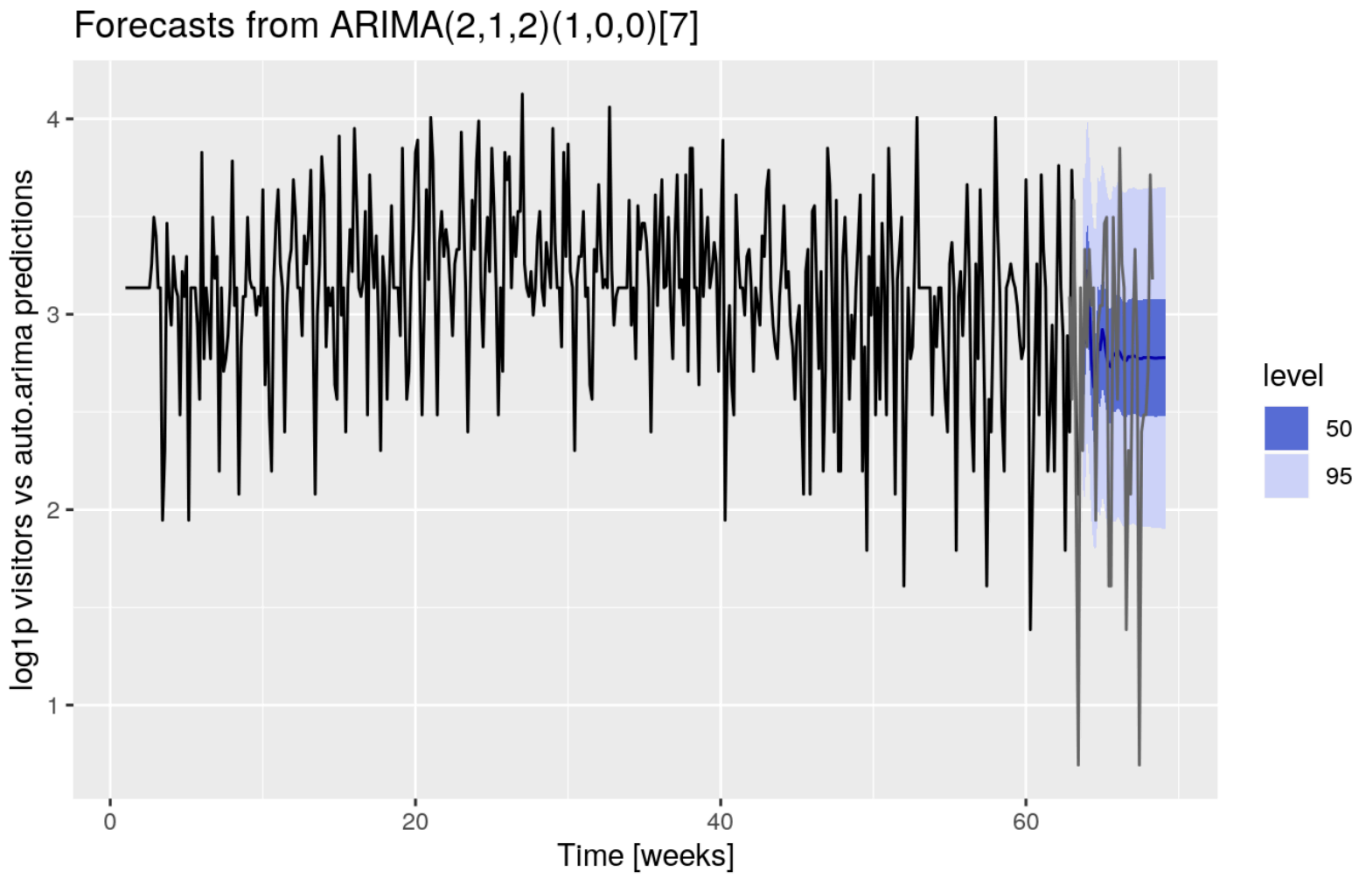
- The scatter points fall largely above the line of identity, indicating that there were more *visitors* that day than had reserved a table. This is not surprising, since a certain number of people will always be walk-in customers.
- A notable fraction of the points is below the line, which probably indicates that some people made a reservation but changed their mind and didn't go.
- The linear fit suggests a trend in which larger numbers of *reserve_visitors* are more likely to underestimate the eventual *visitor* numbers. This is not surprising either, since I can imagine that it is more likely that a large group of people walk in a restaurant without reservation.



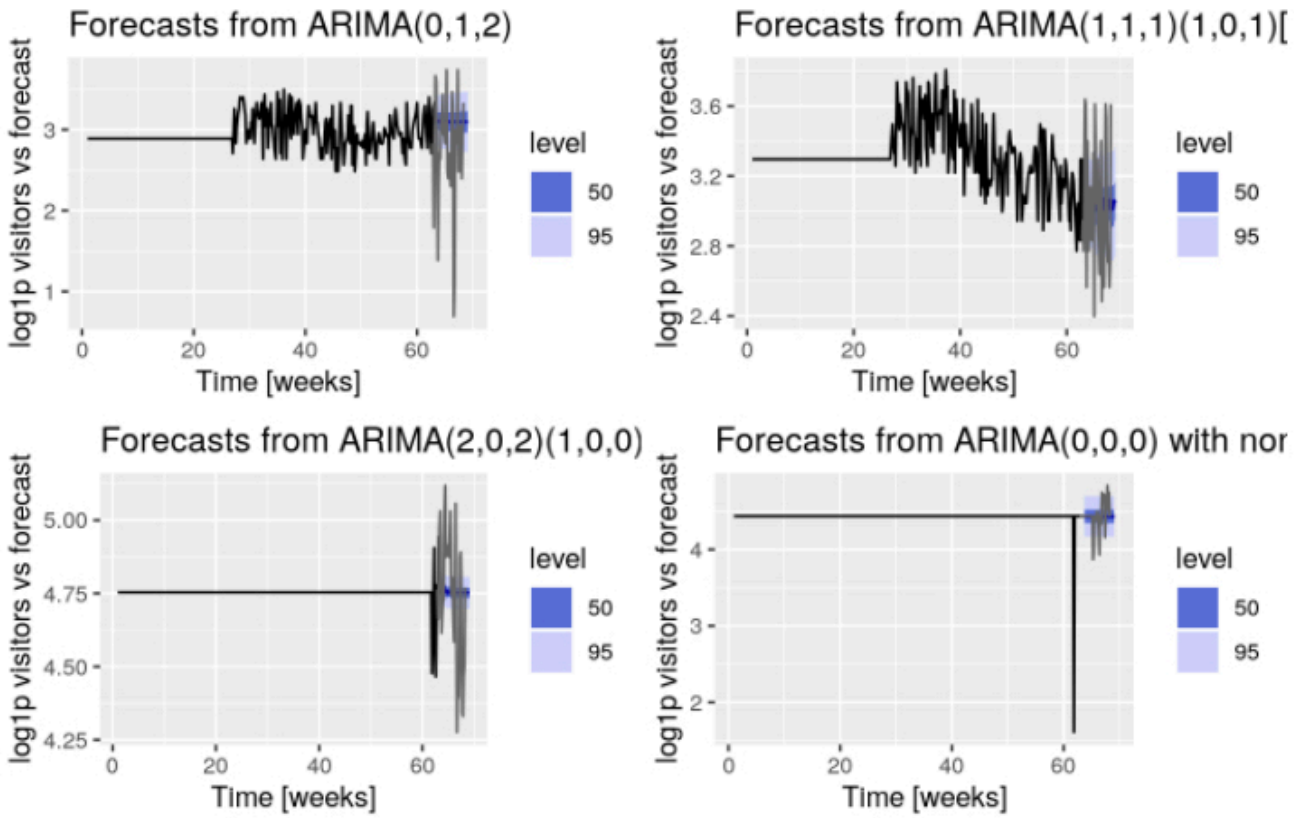
- There are somewhat higher numbers of *visitors* compared to *reservations* on a holiday. The peaks are almost identical, but we see small yet clear differences towards larger numbers.

5 Forecasting methods

5.1 ARIMA / auto.arima

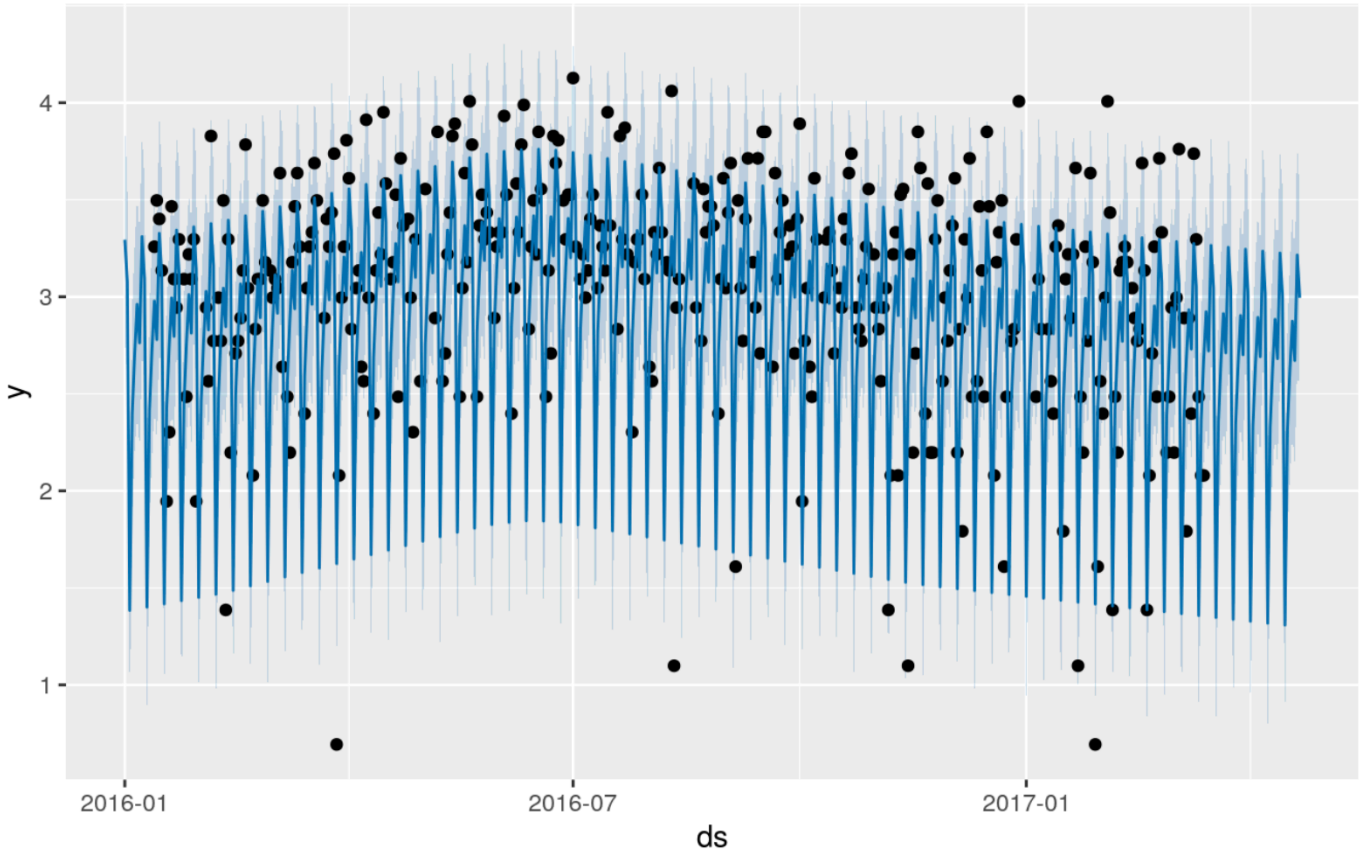


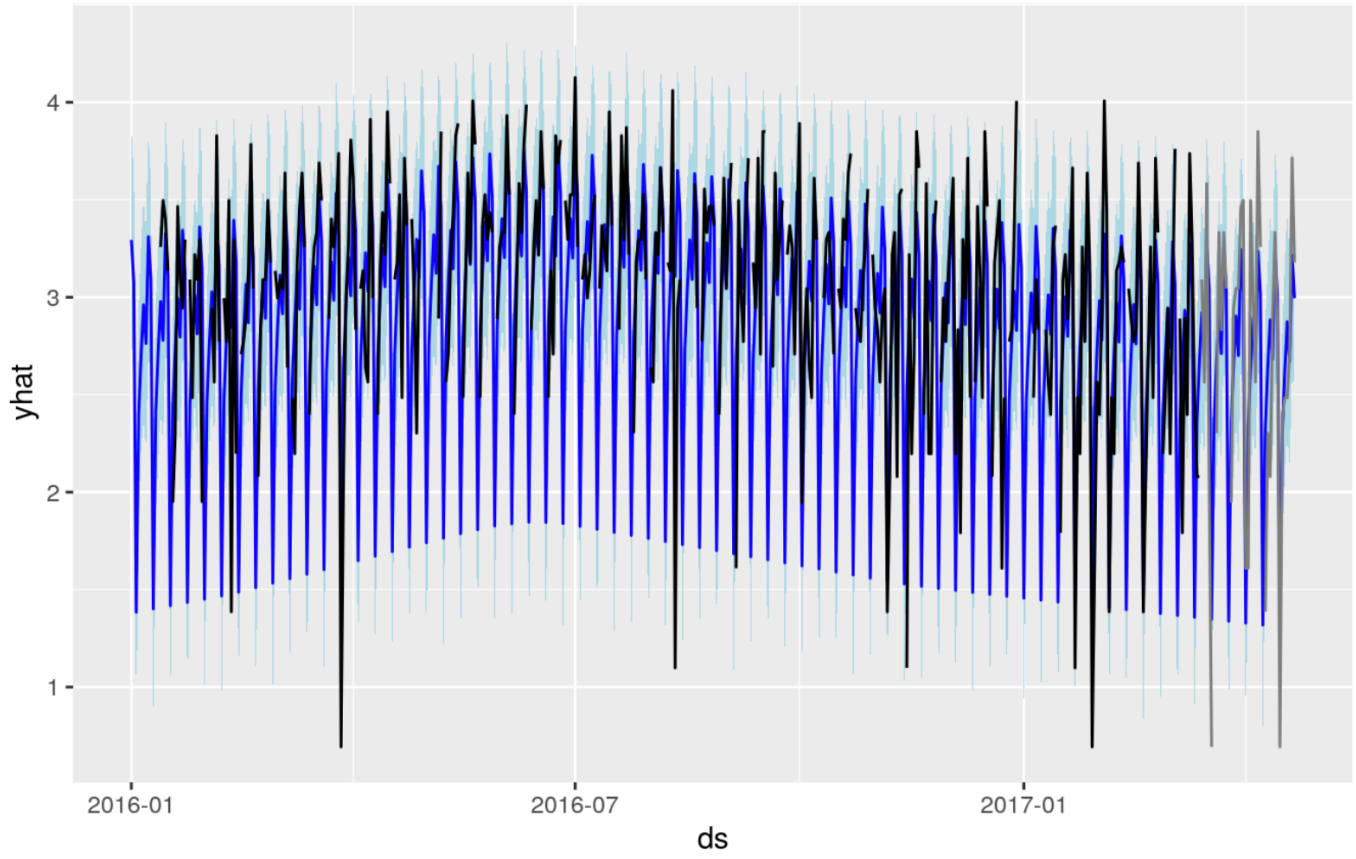
- We find that the first days of the forecast fit quite well, but then our prediction is not able to capture the larger spikes. Still, it's a useful starting point to compare other methods to.

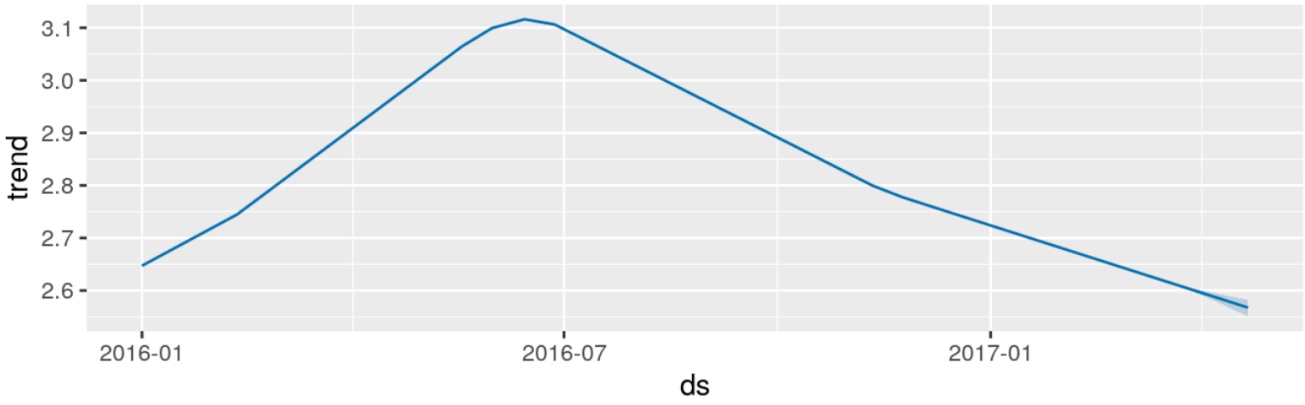


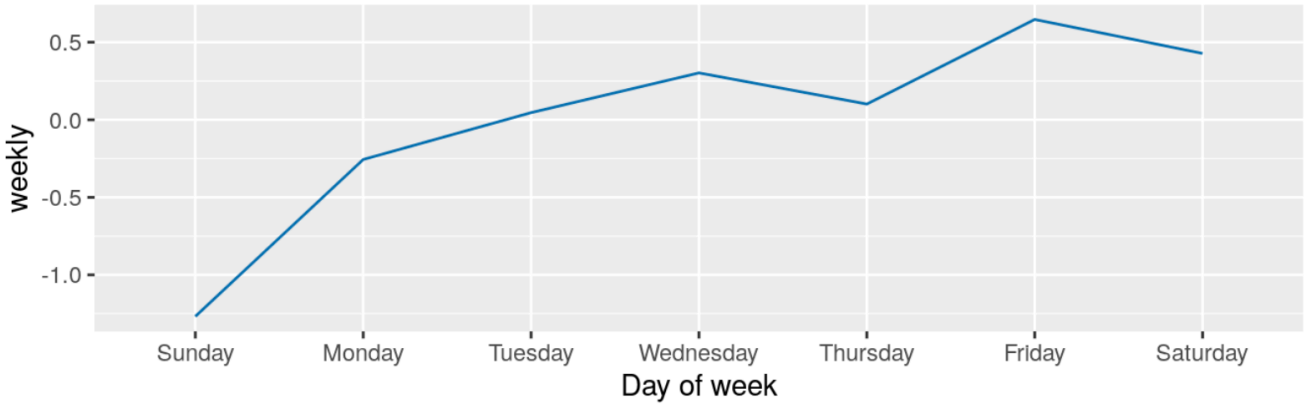
- The two time series' in the upper panels are reasonable complete, but we see that the long gaps (and our median filling) lead to problems in the predictions in the upper left panel where we loose the weekly periodicity. The upper right panel retains this periodicity and the predictions for the first days are relatively decent, but then we quickly under-predict the amplitude of the variations.
- The lower panels include two of the outliers from our time-series parameter space above; and here we see cases where things go really wrong. These kind of peculiar time series could lead to a bad performance for any otherwise decent forecasting algorithm if they contain a large enough fraction of visits in the test data set.

5.2 Prophet

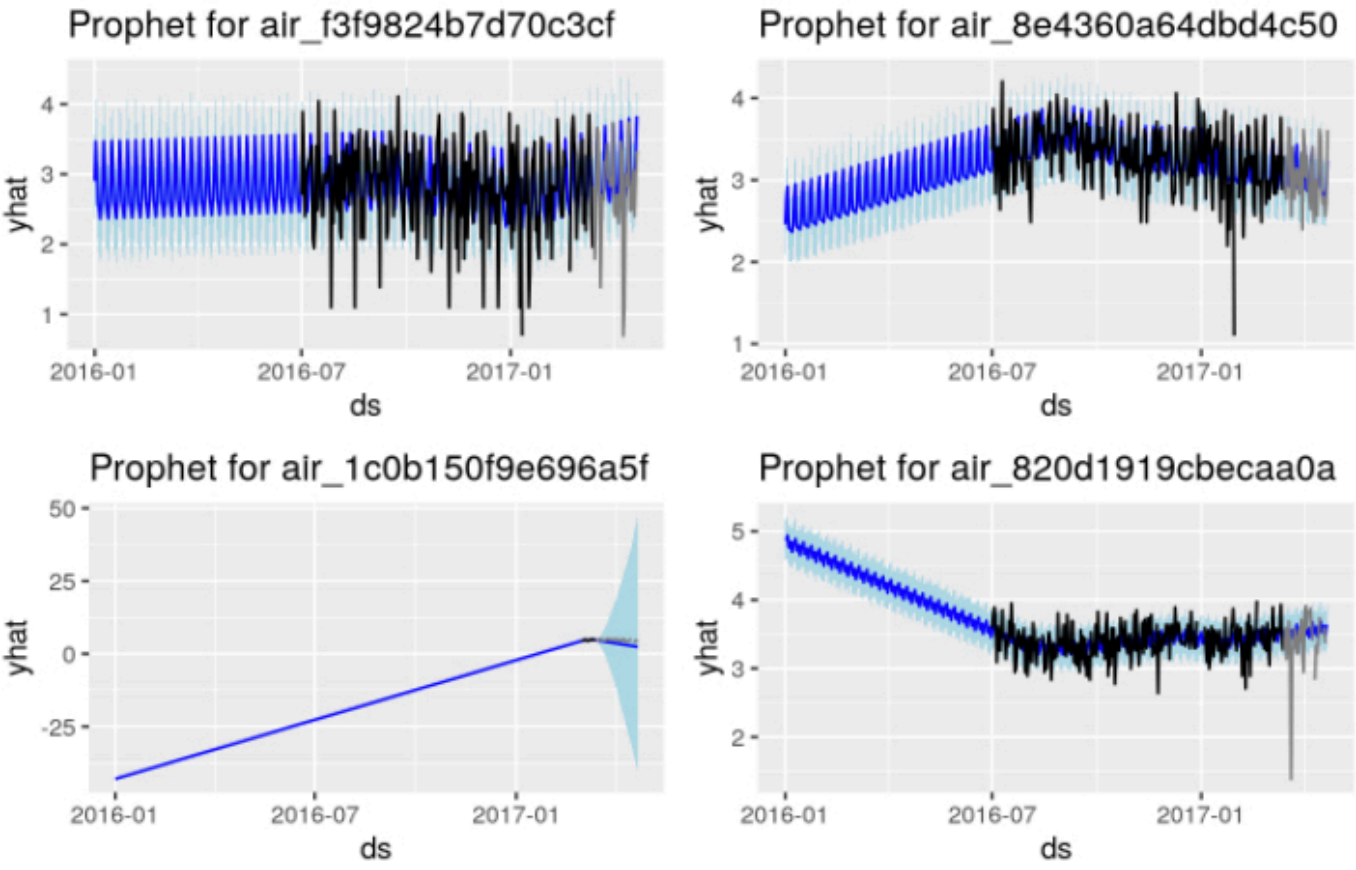

y
4
3
2
1
2016-01
2016-07
2017-01
ds


yhat
4
3
2
1
2016-01
2016-07
2017-01
ds

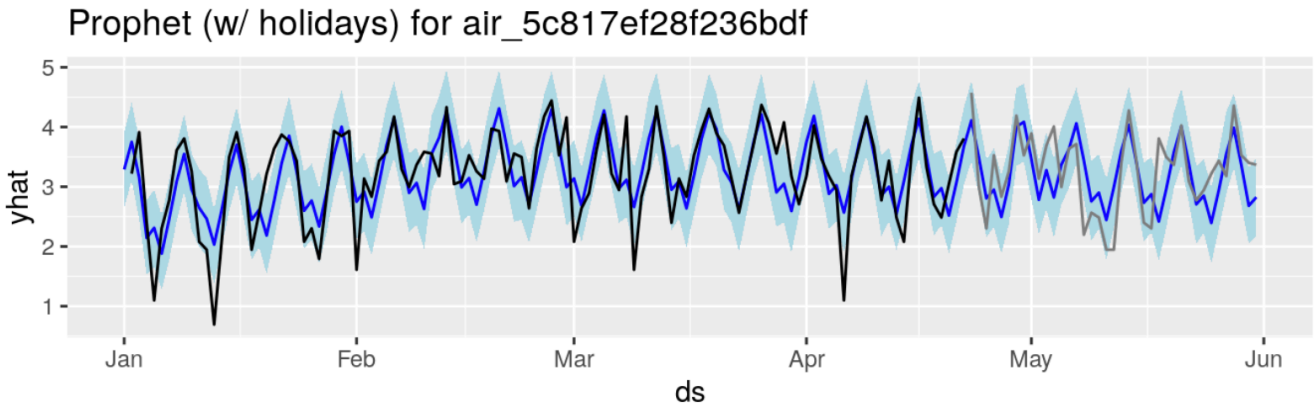


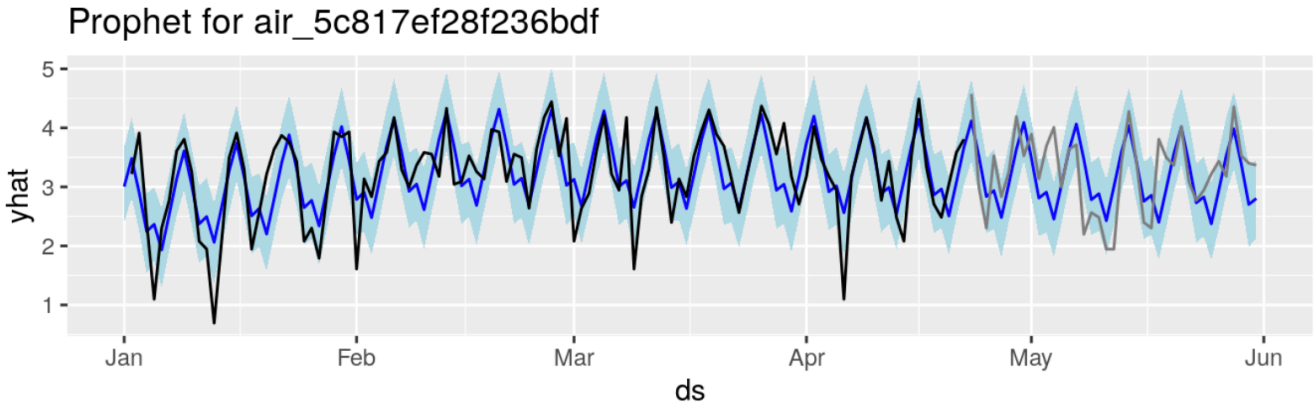


- Prophet detects a weekly variation pattern which is similar to what we had found before, in that Fri/Sat are more popular than the rest of the week.



- Looks like we're overestimating the trend component, which could probably use less flexibility, for at least three of these. The forth time series has too little data for prophet to be able to do much.





- There is a subtle improvement in fitting the Golden Week *visitors* when including holidays. The performance of this component might improve if there are more holidays included in the *training* set.