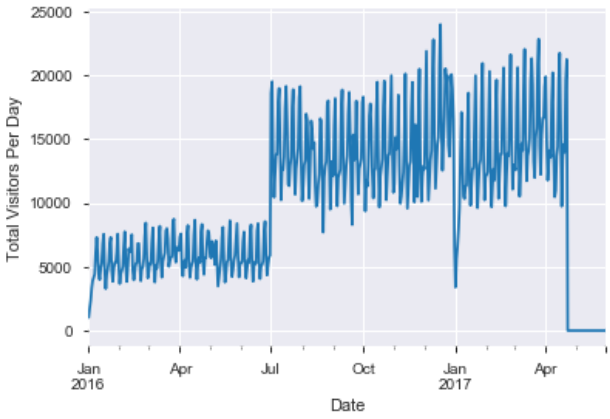
Restaurants face many problems; one of which is trying to predict how many customers will visit the restaurant on each day. This is helpful so that the appropriate number of staff can be scheduled and that the right amount of food can be ordered. If these can be optimized, costs can be reduced while still serving the customers needs.

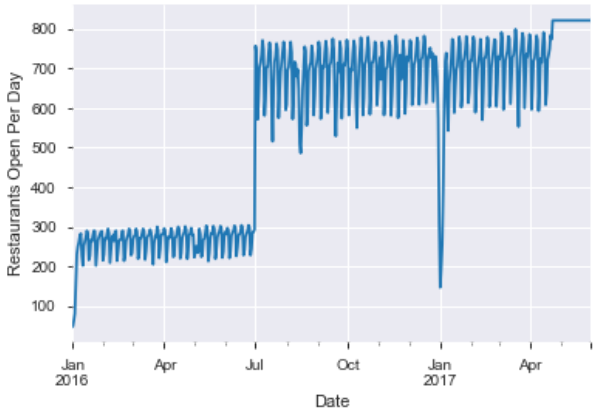
There were many files of data provided by the Kaggle competition. The first file contains a store ID, the date, and the number of visitors. I began by concatenating the test set to the training set so that I would have to merge the rest of the data only once. The file I merged to the visit data contained information about the store including genre and area name. The second file contained the day of the week for each date and whether it was a holiday.

When looking at what external data other people had used for this, I found that someone had gathered weather data from 1,600+ stations across Japan as well as how far each of these stations is from each of the 108 unique latitude/longitude pairs. For many of these locations, the closest station only measured precipitation or temperature. By averaging the 5 closest stations, I was able to get temperature and precipitation measurements for every location for every day. Some of the other measurements were still missing. I filled them first by the average value in the same city on that day, and I filled the rest by the average of all measurements on the day.

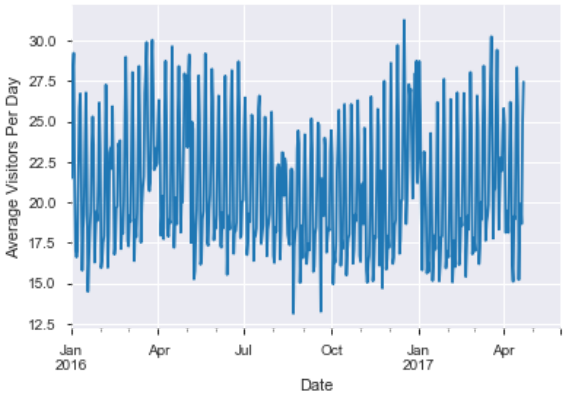
To begin my exploratory analysis, I used the pandas profiling module to create a report about each of the variables. It found that four of the weather measurements were highly correlated to others and should be excluded: average sea pressure, average vapor pressure, high temperature, and low temperature.

I found two interesting things when looking at the total visitors per day over the entire date range. There appears to be a weekly cycle and at the beginning of July the total visitors goes from ~50k to 150k. I thought this second change may be due to an increase in the number of restaurants, and I found that this also almost tripled in early July.



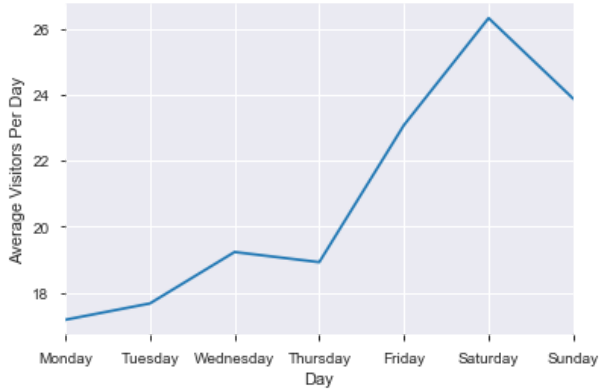


I decided to look at average visitors per day to remove the effect of a different number of restaurants being open each day. The daily data showed that there may be a seasonal variation in visitors, but the change is smaller than the day to day change. By averaging over a week and then a month, the seasonal change became more noticable.

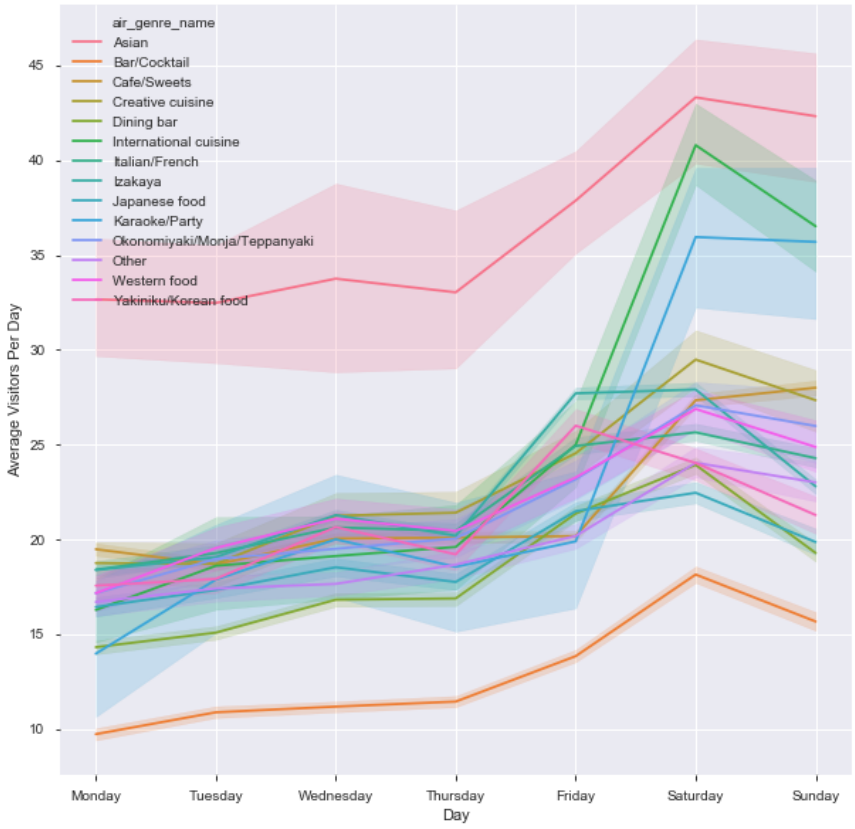


I next wanted to investigate whether total visitors per day graph was showing a weekly trend or if it was just noise. When looking at the average visitors per day based on the day of the week,

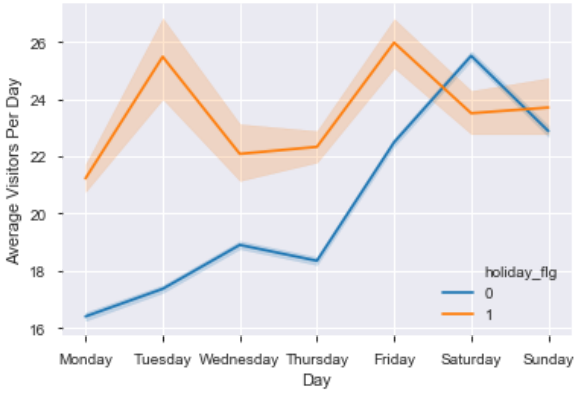
Monday through Thursday have a much smaller average than Friday through Sunday. Saturday is the maximum with an average 50% more than Monday.

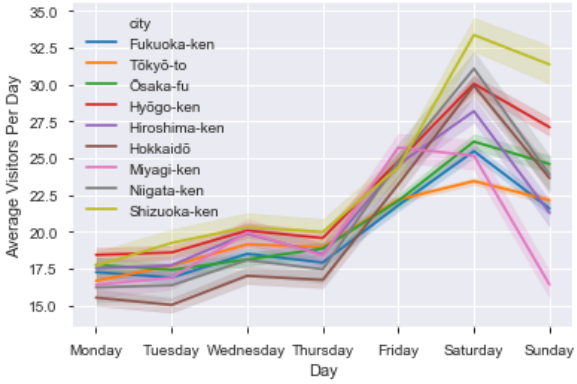


Looking at this weekly trend and separating by the other categorical features yielded some interesting results. Most of the 14 genres of restaurant are fairly tightly clustered with only 4 that stand out. The ‘Asian’ genre has a consistently much larger number of customers while the ‘BarCocktail’ genre is consistently much lower. The ‘International cuisine’ and ‘Karaoke/Party’



genres blend in with the rest on Monday through Friday, but have a much larger increase on Saturday and Sunday. If there is a holiday on a weekday, the number of visitors increases, but on Saturday and Sunday, there appears to be no effect. The city of the restaurant has no clear effect on the number of visitors.





I then ran an F-test on each of the categorical features and found that all have statistically significant statistics. This means that for each feature there is at least one pair of categories that have different means. The only feature where this is surprising is 'city.' This can be made sense of by realizing that even though they are clustered together, the highest mean may be significantly different from the lowest because there is so much data.