Restaurant Visitor Prediction

Team 9
Tianhao Guo, Jiani Gao, Tingting Ma, Mo Cheng, Jinyan Lyu

Business Target

- To avoid letting customers waiting for a long time.
- To avoid the waste of food.
- Help to provide a win-win situation: predict how many visitors a restaurant is expected to have in a given time.

Data Description

- 7 subsets that contains reservation, visit, together with other information
- Collect from two websites:

  a restaurant reservation website AirREGI

  a restaurant rating website Hot Pepper Gourmet.
- Already split into training and testing sets:

  training data covers the dates from 2016 until April 2017,

  test set covers the last week of April and May of 2017.

Data Exploration

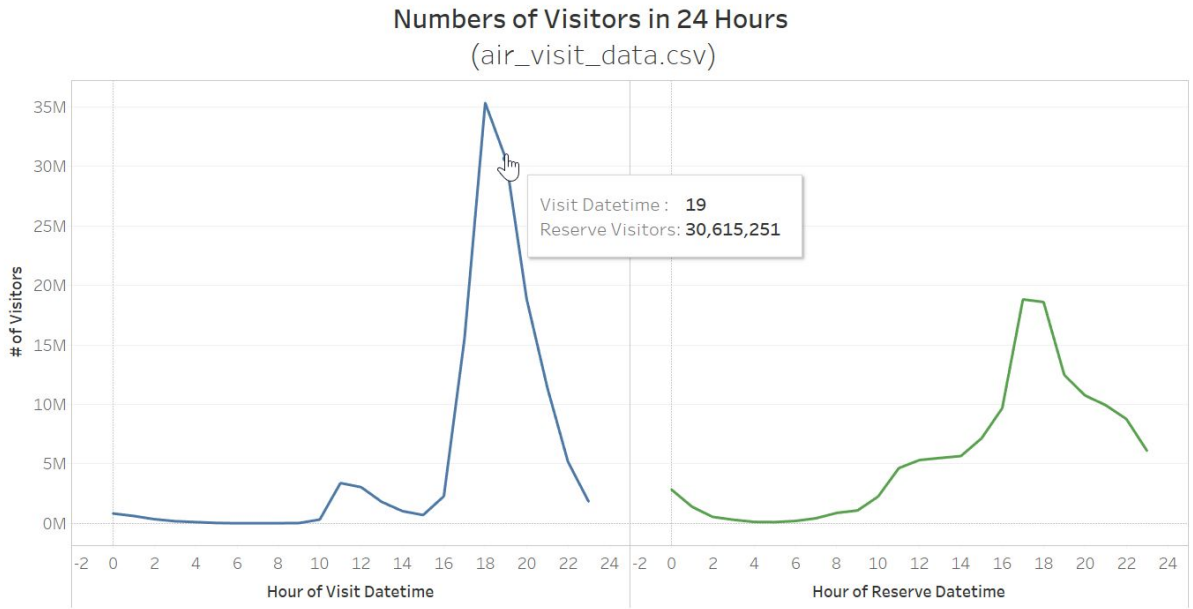


Figure 1: Total numbers of visitors of all resturants in 24 hours from AirREGI dataset.

Data Exploration

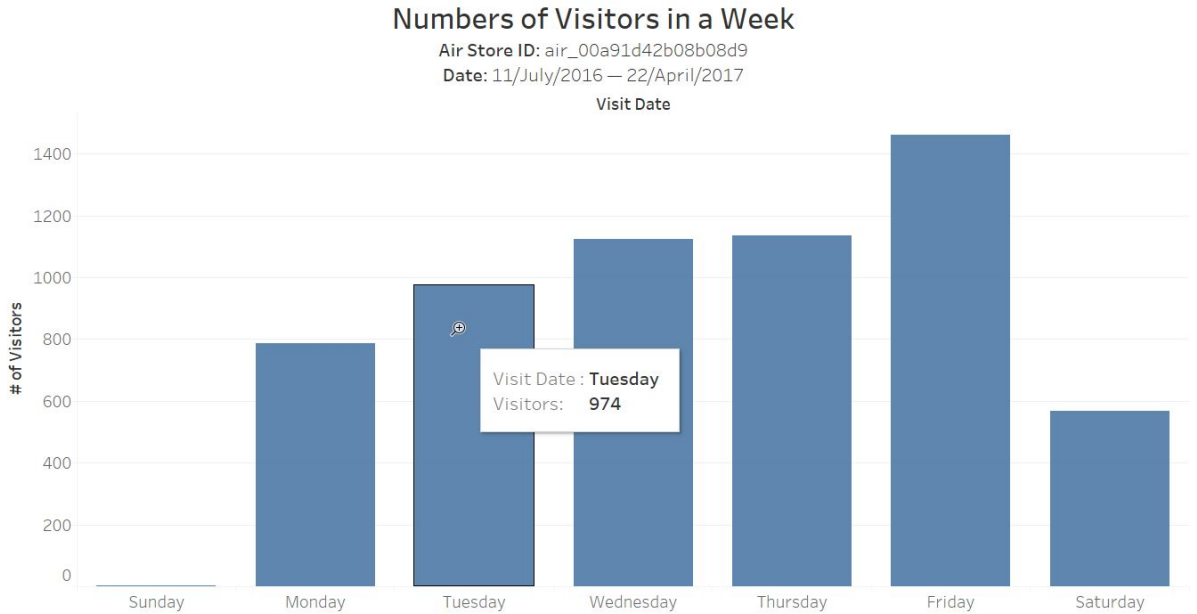


Figure 2: Number of visitors in a week of a resturant (ID: air_00a91d42b08b08d9) from 11/July/2016 to 22/April/2017.

Data Exploration

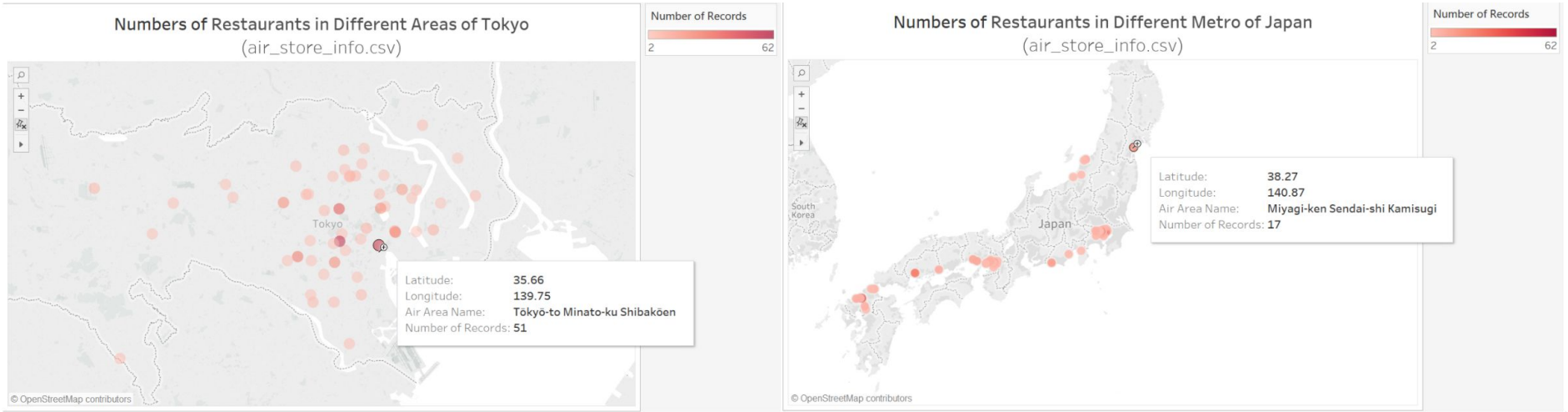


Figure 3: Distribution of restaurants in Tokyo and Japan.

Data Exploration

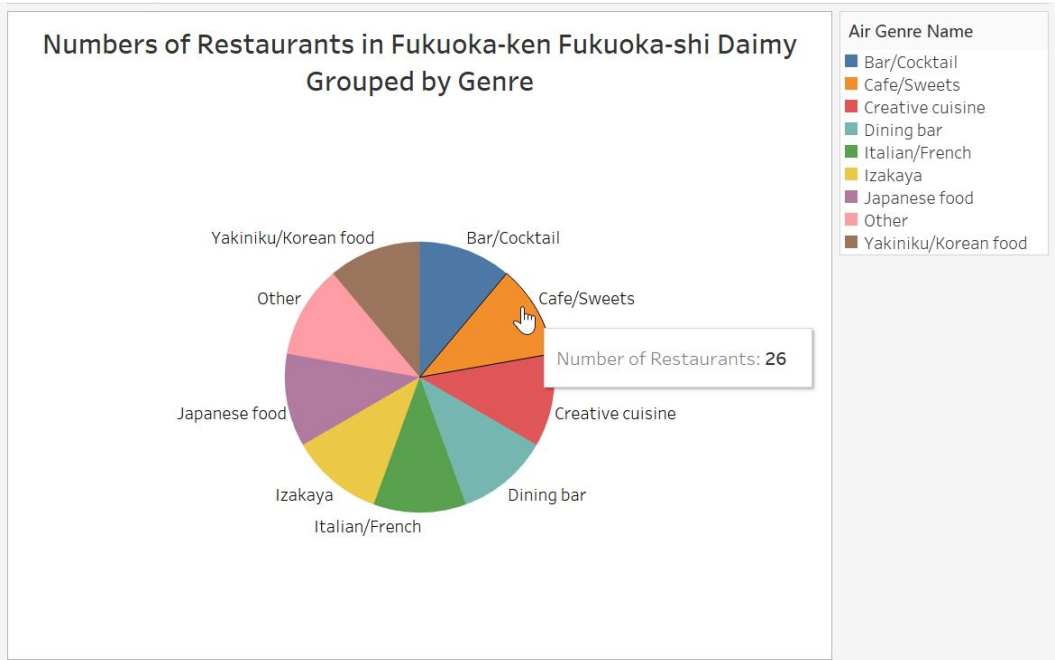
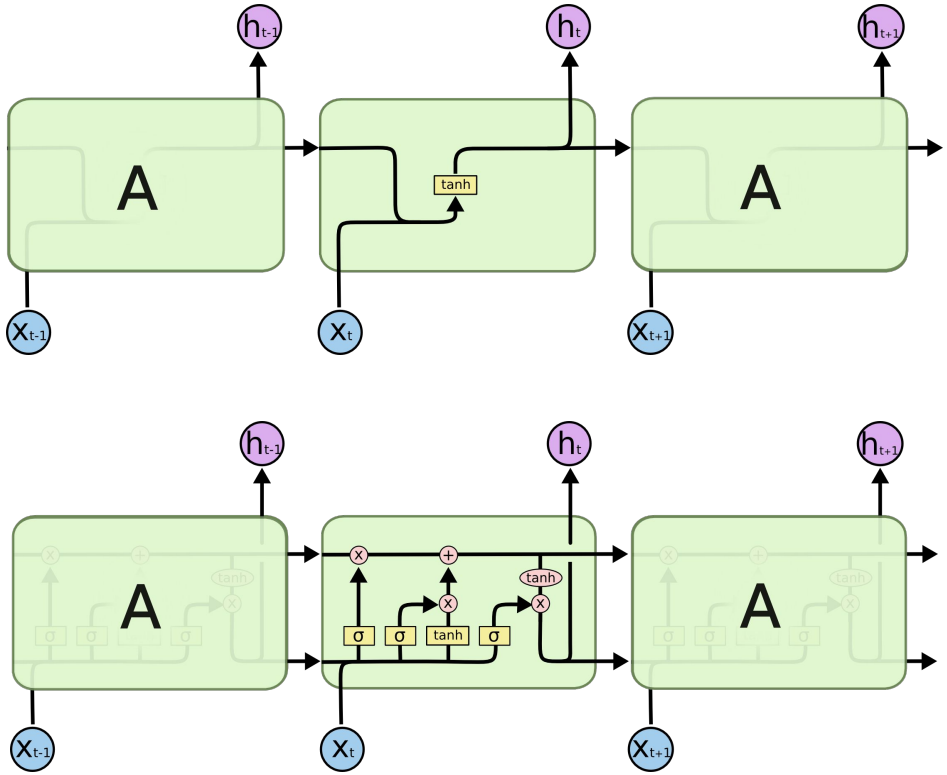


Figure 4: numbers of different genres of restaurants in Fukuoka-ken Fukuoka-shi Daimyo

The Math Part

Long Short-Term Memory (LSTM) recurrent neural network

- A special kind of RNN which is capable of learning long-term dependencies;
- Designed to avoid the long-term dependency problem;
- Well-suited to classify, process and predict time series given time lags of unknown size and duration between important events.



The Hacking Part

- Process the raw data
- Merge two ralational datasets to a single dataframe
- Convert genes to numerical type

New data looks like:

	air_store_id	visit_date	visitors	dow	year	month	min_visitors	mean_visitors	median_visitors	max_visitors	count_observations	air_genre_name
0	air_ba937bf13d40fb24	2016-01-13	25	2	2016	1	7.0	23.843750	25.0	57.0	64.0	4.0
1	air_ba937bf13d40fb24	2016-01-14	32	3	2016	1	2.0	20.292308	21.0	54.0	65.0	4.0
2	air_ba937bf13d40fb24	2016-01-15	29	4	2016	1	4.0	34.738462	35.0	61.0	65.0	4.0
3	air_ba937bf13d40fb24	2016-01-16	22	5	2016	1	6.0	27.651515	27.0	53.0	66.0	4.0
4	air_ba937bf13d40fb24	2016-01-18	6	0	2016	1	2.0	13.754386	12.0	34.0	57.0	4.0

The Hacking Part

- Use Pandas shift() function to transform time series data into a supervised learning problem
- Use MinMaxScaler from scikit-learn to normalize all features to range(0,1)
- Defined our own series_to_supervised() function to shift columns of variables

New data looks like:

	var1(t-1)	var2(t-1)	var3(t-1)	var4(t-1)	var5(t-1)	var6(t-1)	var7(t-1)	var8(t-1)	var9(t-1)	var10(t-1)	var11(t-1)	var12(t-1)	var13(t-1)	var14(t-1)	var15(t-1)	var16(t-1)	
1	0.666667	0.0	0.0	0.022556	0.076686	0.071895	0.023918	0.985714	0.214286	0.038835	0.768135	0.904147	0.0	0.0	0.0	0.0	0
2	0.666667	0.0	0.0	0.030075	0.132432	0.124183	0.055809	0.971429	0.214286	0.757282	0.815045	0.968543	0.0	0.0	0.0	0.0	0
3	0.666667	0.0	0.0	0.082707	0.288725	0.274510	0.095672	1.000000	0.571429	0.650485	0.815973	0.968186	0.0	0.0	0.0	0.0	0
4	0.666667	0.0	0.0	0.015038	0.035856	0.026144	0.015945	0.857143	0.142857	0.902913	0.794438	0.940490	0.0	0.0	0.0	0.0	0
5	0.666667	0.0	0.0	0.120301	0.228969	0.209150	0.082005	1.000000	0.500000	0.300971	0.793805	0.938517	0.0	0.0	0.0	0.0	0

The Hacking Part

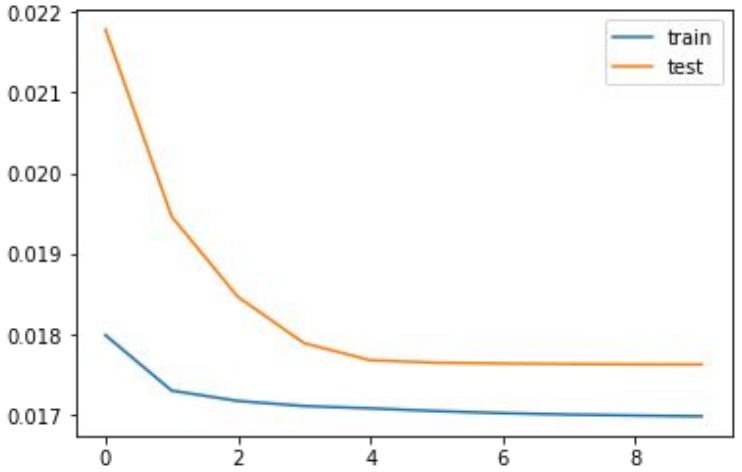
Define and fit the model

- 75% as training data, 25% test data.
- Input data are reshaped into a 3D format [samples, timesteps, features] for LSTM
- Defined the LSTM model with 64 neurons in first hidden layer, 1 neuron in output layer

The Hacking Part

- Get the model after 10 training epochs with batch size 100
- plot the loss of training and test
- Used this model to predict the test data
- root-mean-square error (RMSE)=0.797.

*RMSE is a frequently used measure of the differences between values predicted by a model and the values actually observed.



Loss of training and test in the model

Conclusion

- By analyzing the result of training model and predicted test data, we can see the difference between prediction and actual data is small for enough to business product.
- Thus, we believe that our model can help restaurants to reasonably allocate time and prepare food according to the predicted customer traffic we provide.

Questions?