PROJECT REPORT

PROJECT NAME – COME VISIT AGAIN

TEAM NAME – CURVILICIOUS

AJAY CHHAJED

VISMAYA SOLANKI

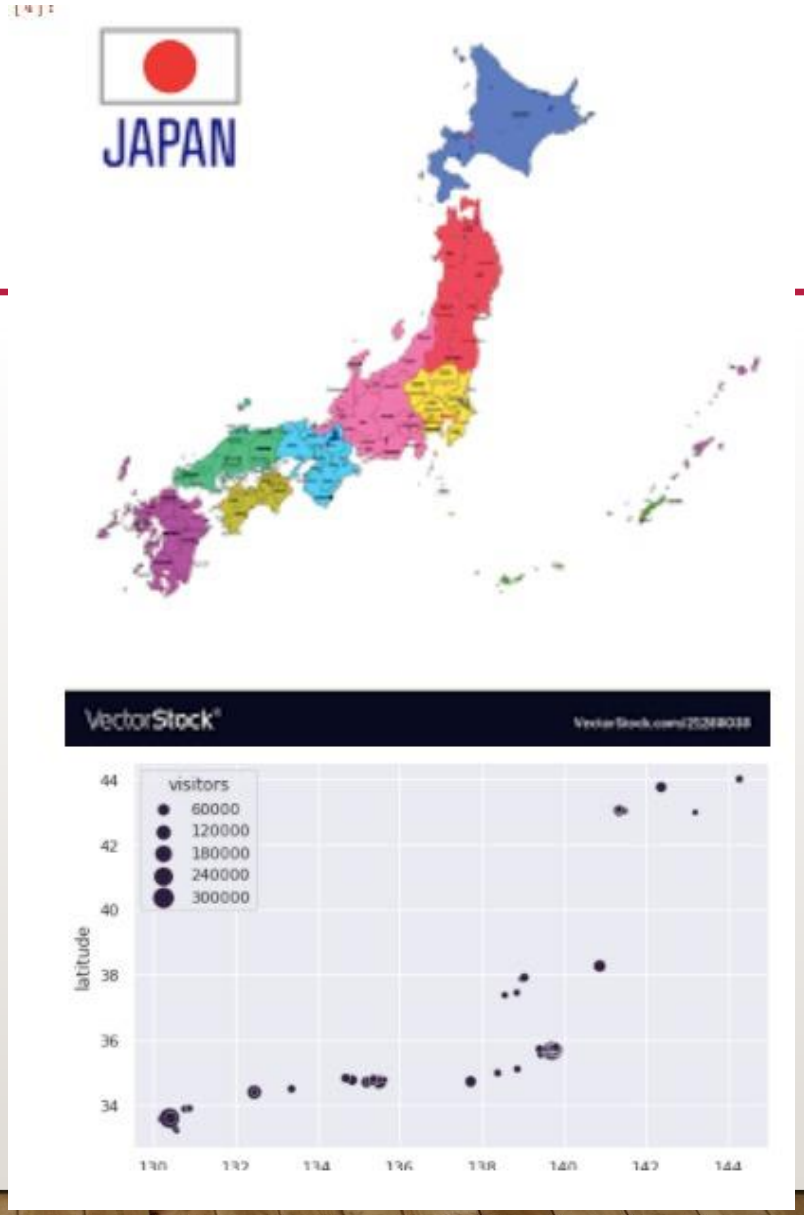
PROJECT DEFINITION

- In this project we were provided with time series data related to visitors in restaurants.
- Based on 2 streams of data (Chwiggy and Yomato) we had to predict the number of visitors for given future dates.
- It was a regression task.

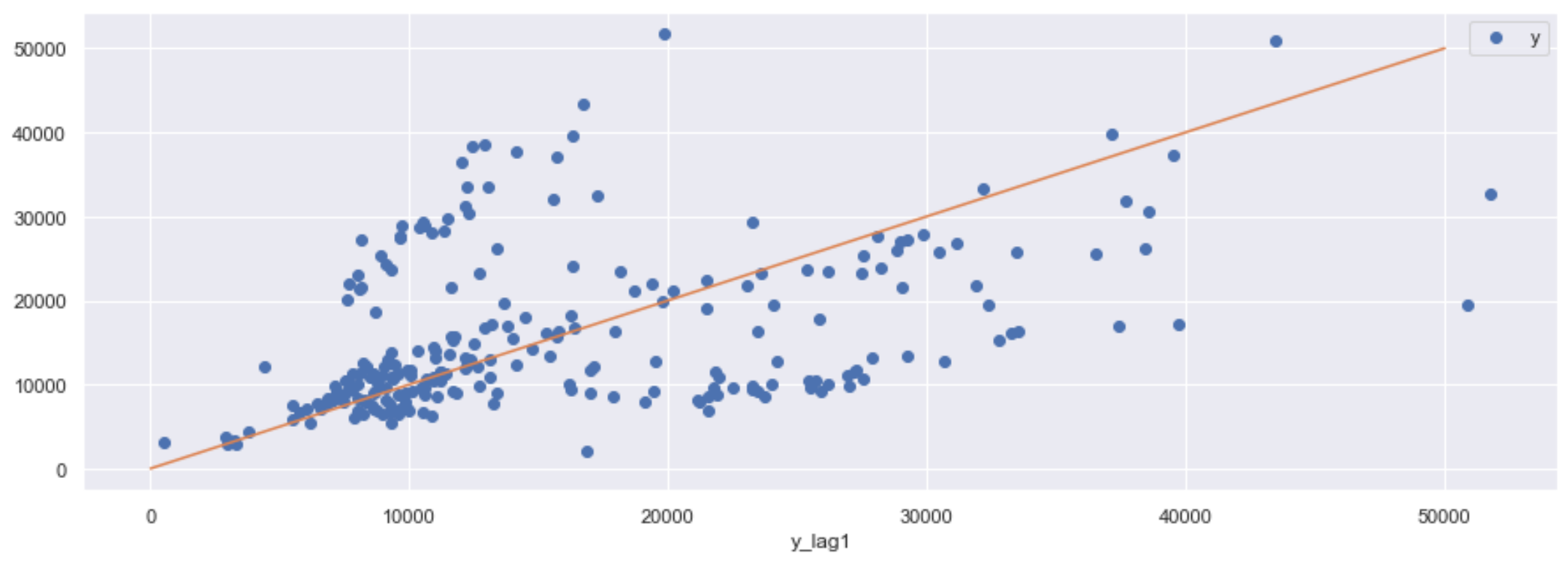
EDA

PLOTTING CO-ORDINATES ON GRAPH

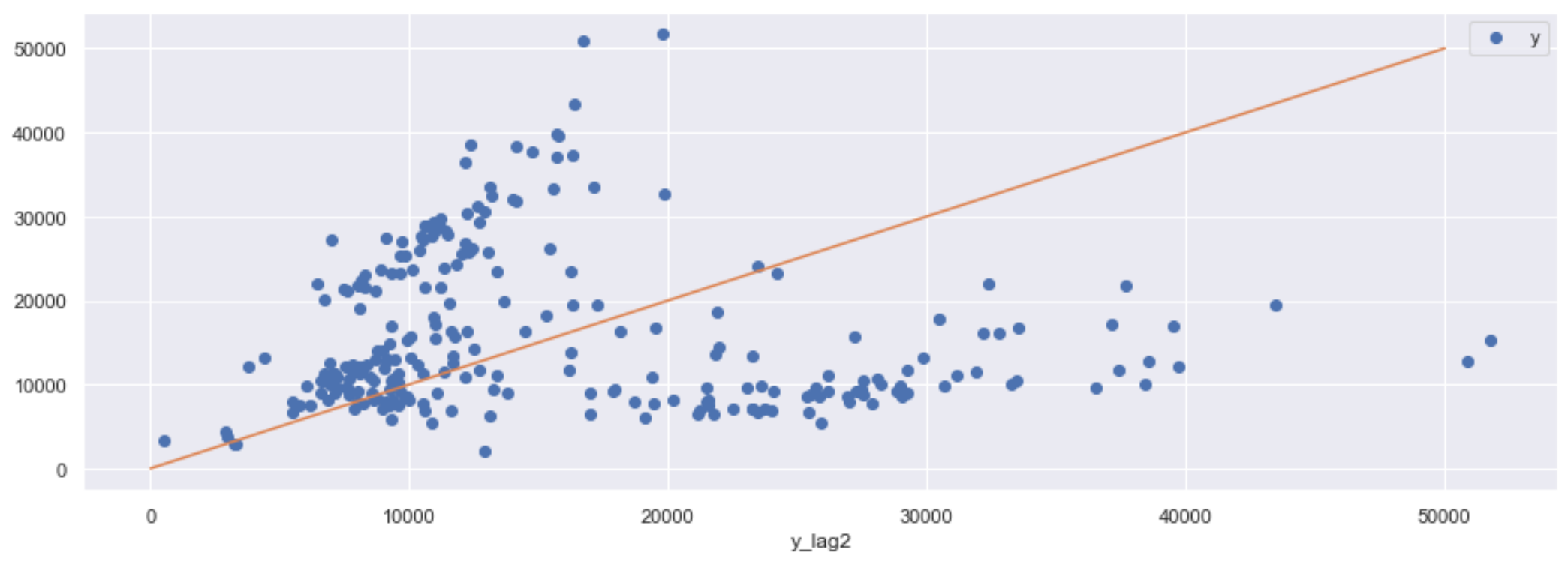
- Impression of graph of Japan can be seen in the plot.



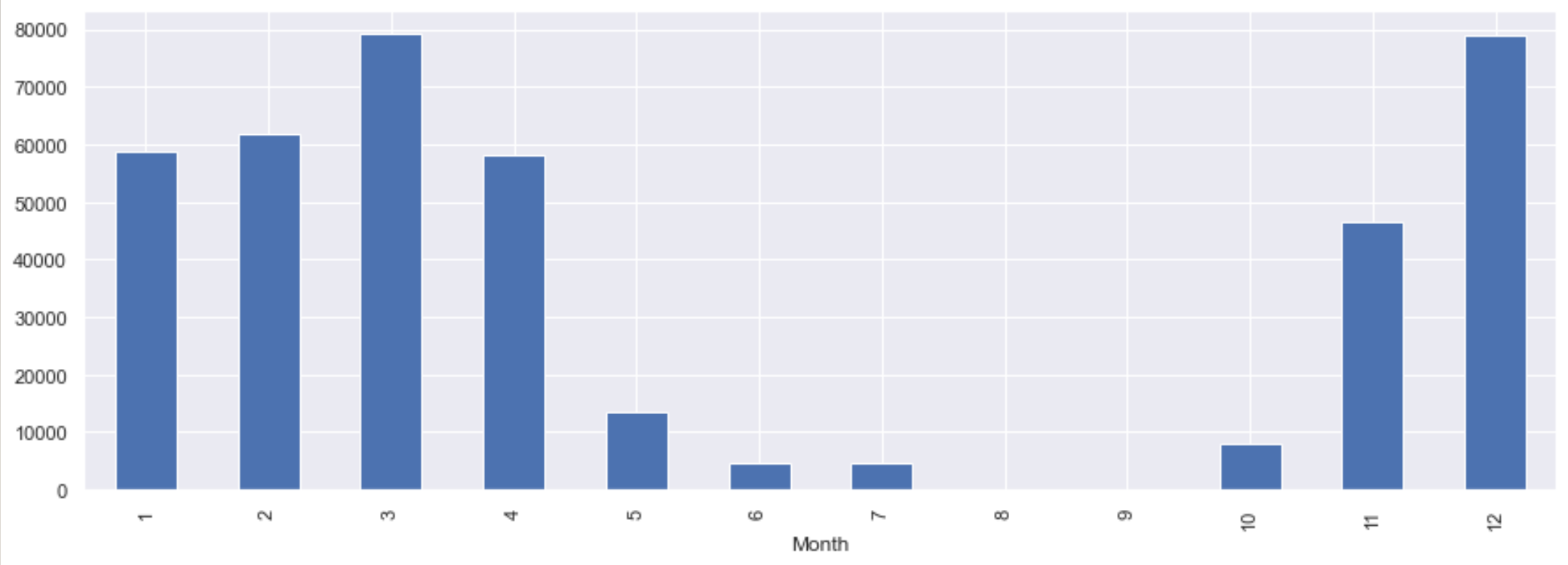
LAG GRAPH OF 1 DAY BACK FOR RESERVE VISITORS ON GIVEN DATE



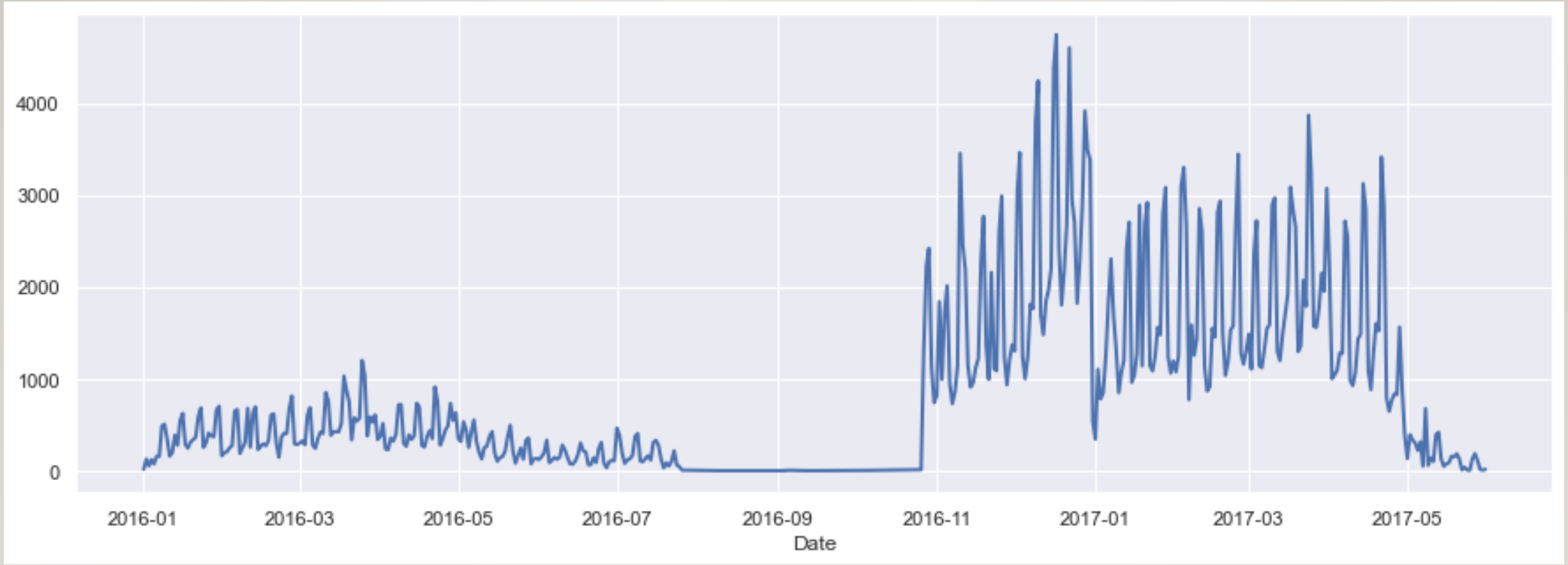
LAG GRAPH 2 DAYS BACK FOR RESERVE VISITORS ON GIVEN DATE



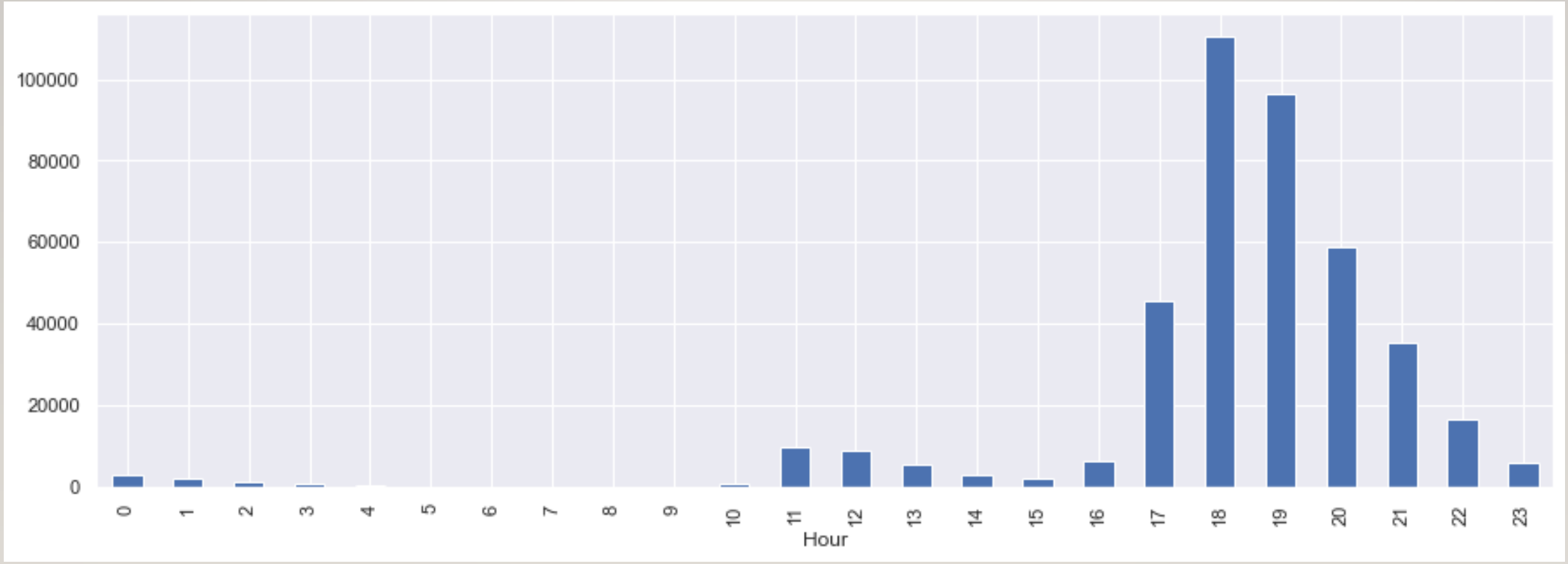
CHWIGGY RESERVE VISITORS VS MONTHS



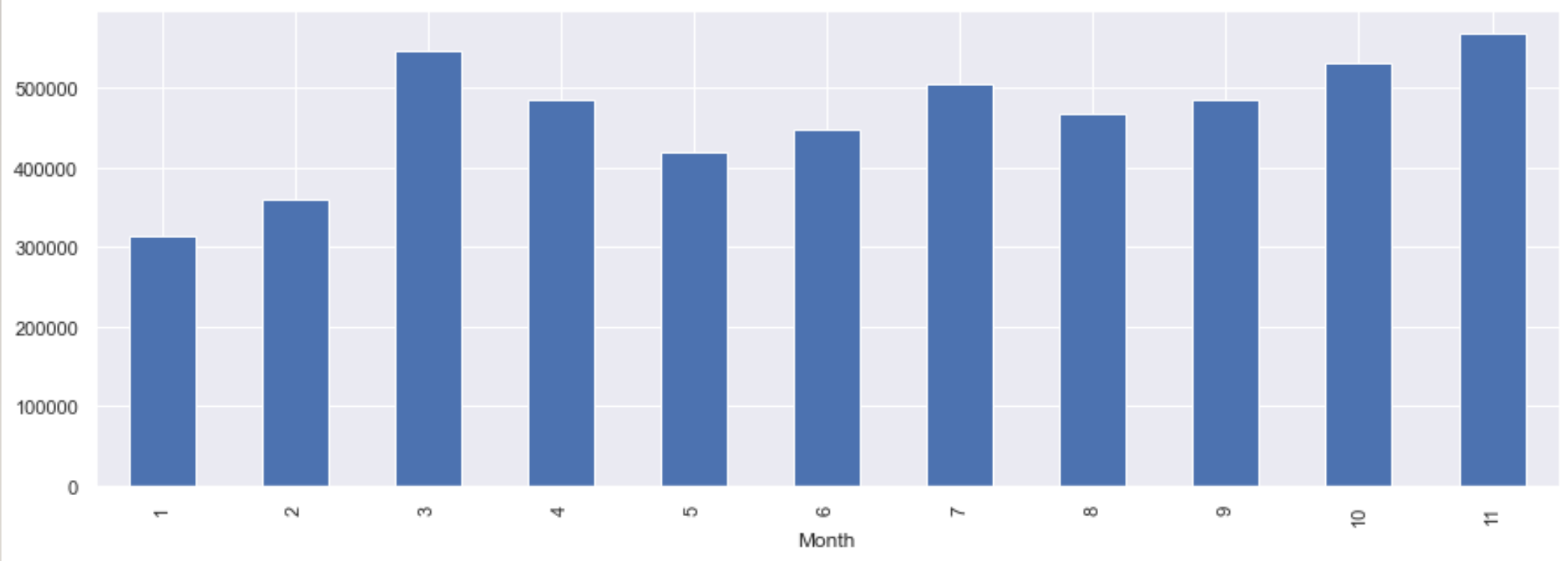
RESERVE VISITORS VS DATE FOR CHWIGGY



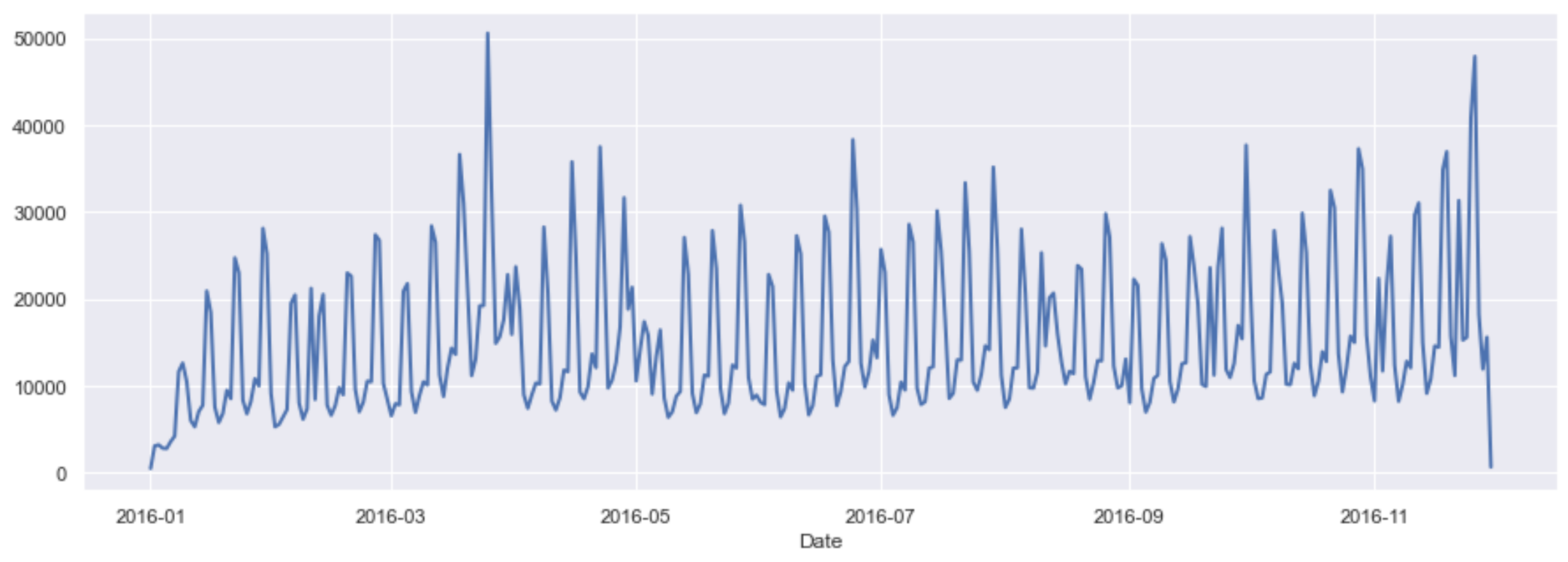
HOUR VS VISITORS FOR CHWIGGY



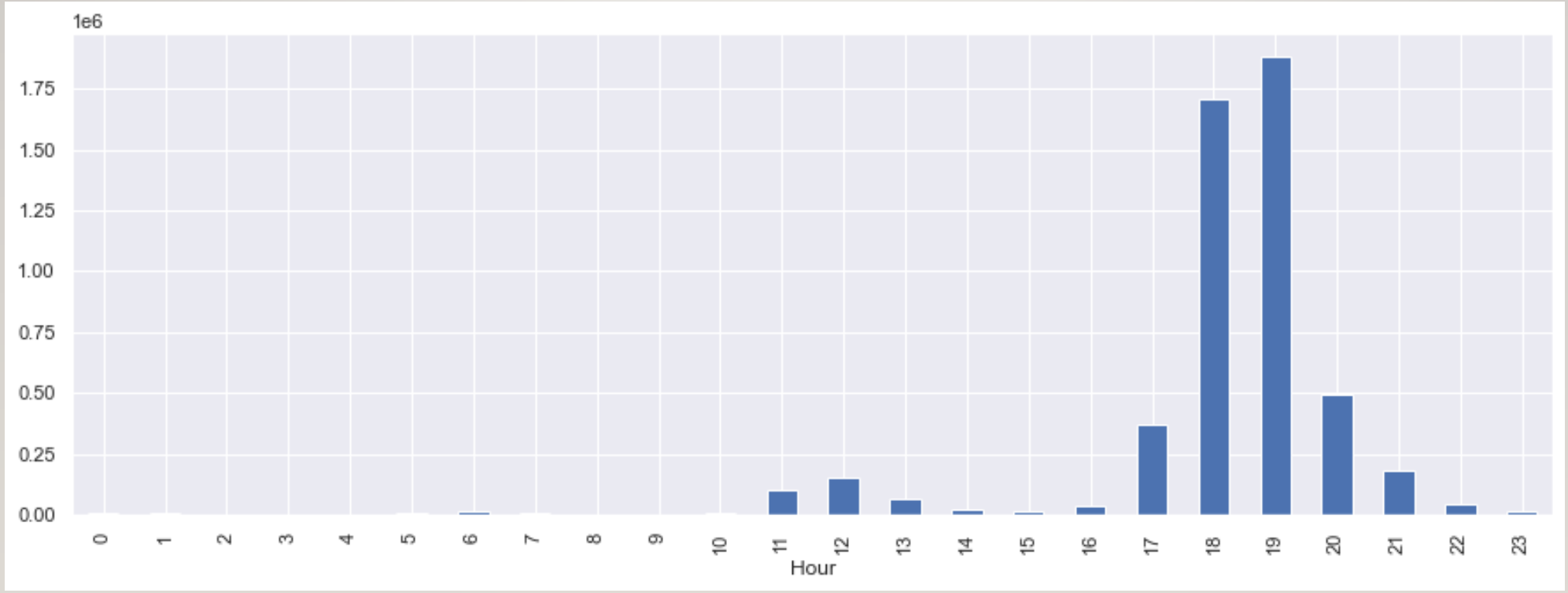
MONTH VS VISITORS FOR CHWIGGY



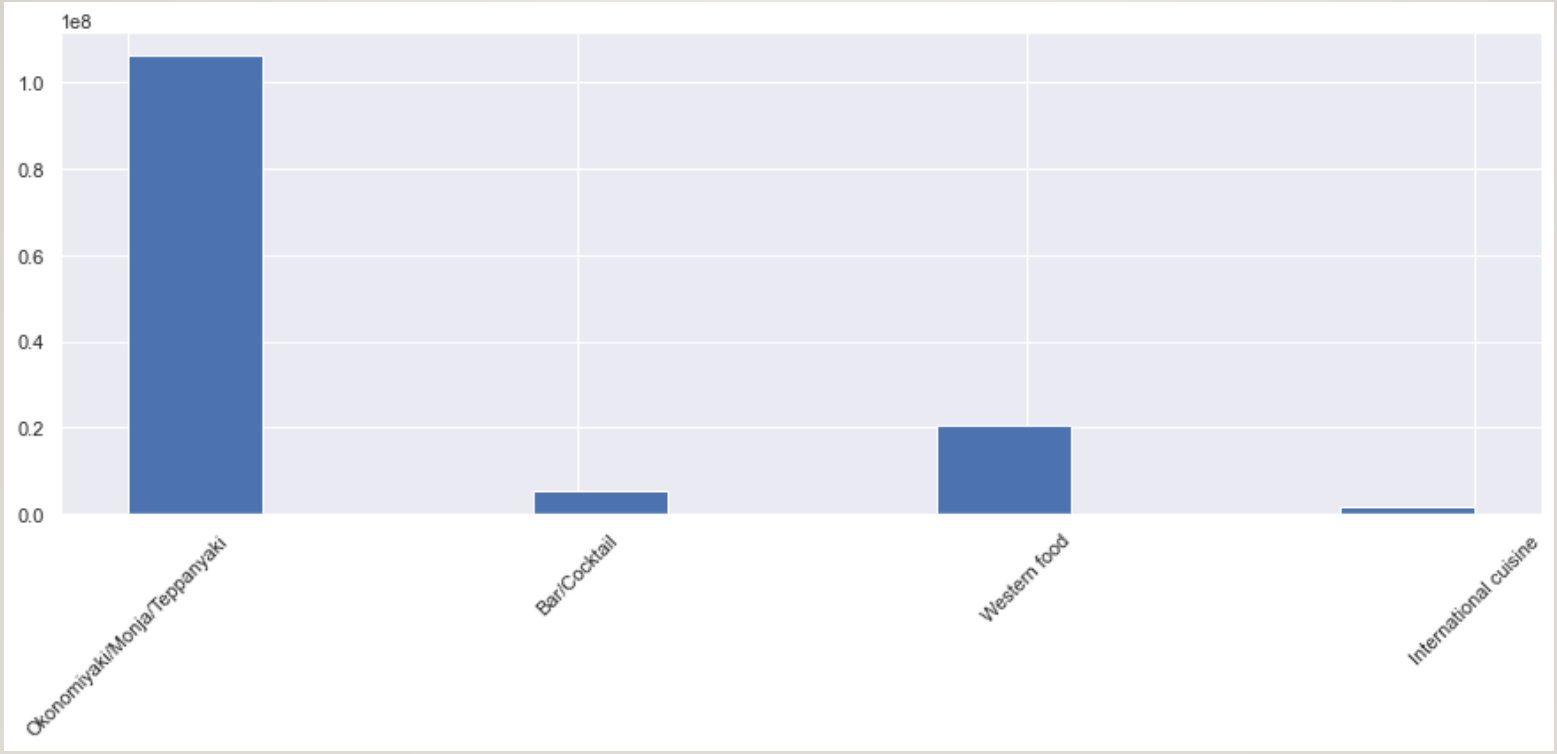
VISITORS VS DATE FOR YOMATO



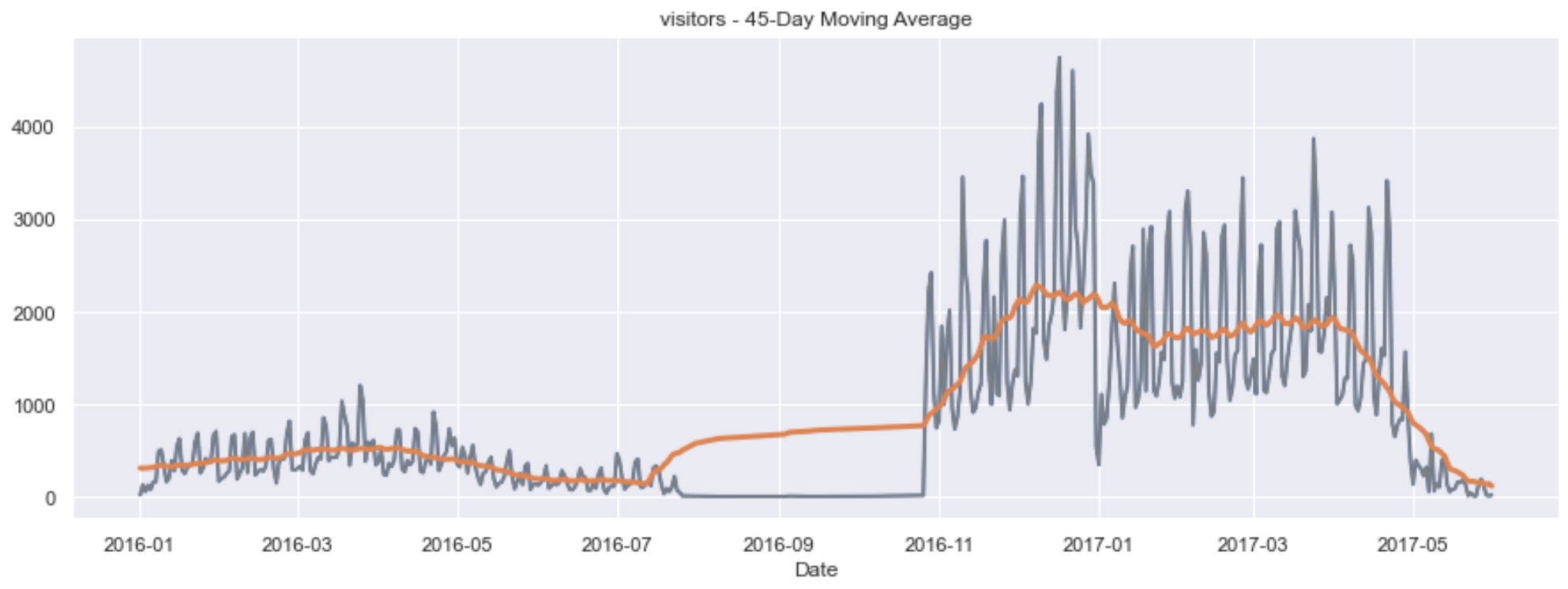
VISITORS VS TIME FOR YOMATO



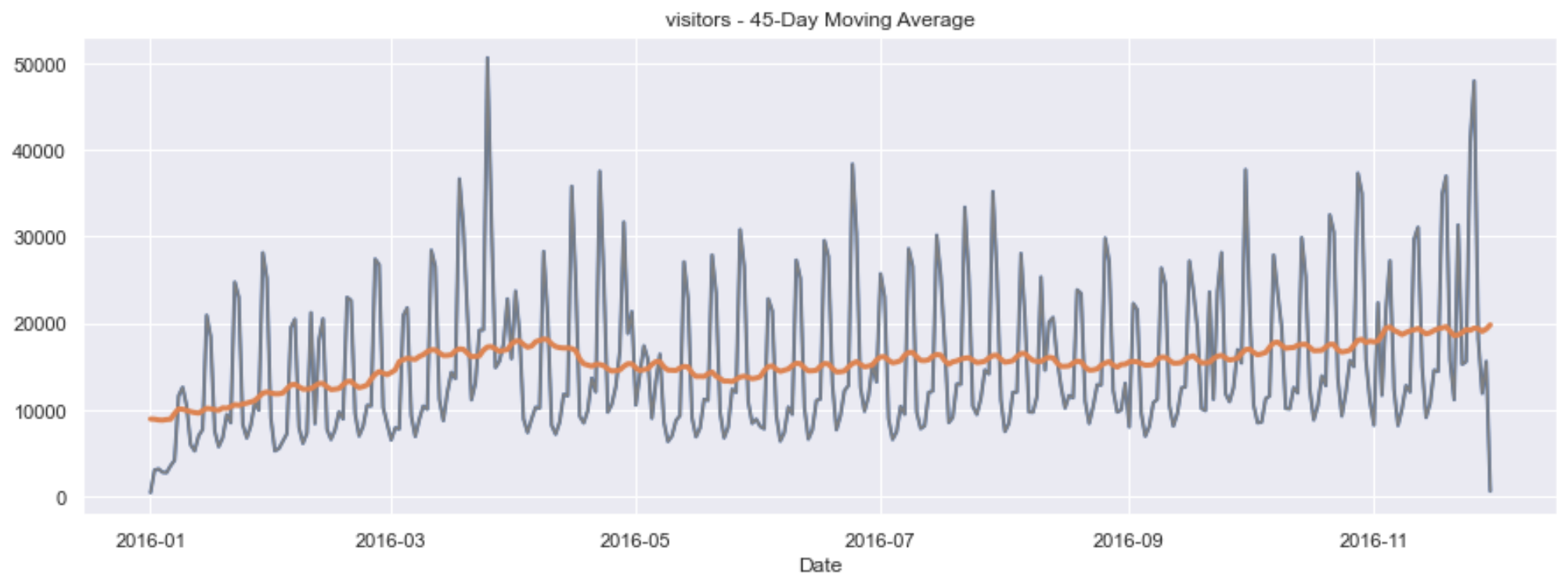
GENRE NAME VS RESERVE VISITORS, MERGED BY IDS AND THEN FOLLOWED BY GENRE NAMES FOR BOTH YOMATO AND CHWIGGY



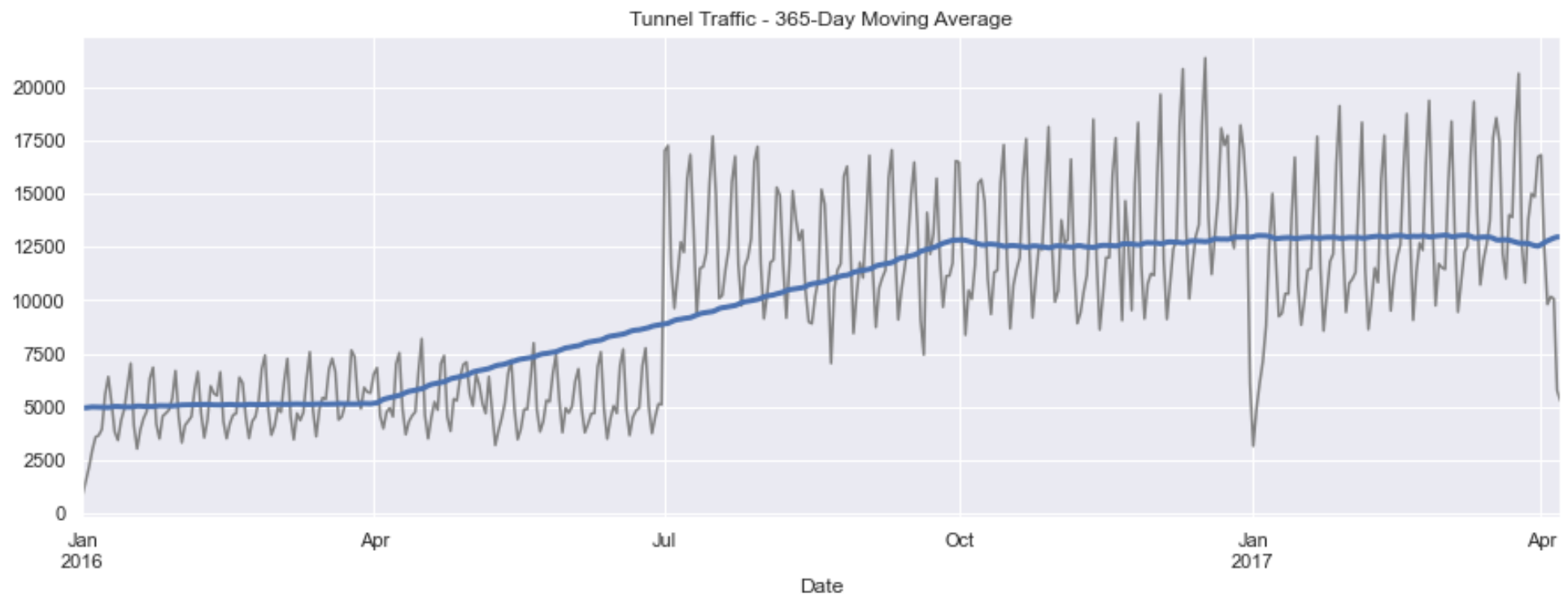
CHWIGGY MOVING AVERAGE OF RESERVE VISITORS(WINDOW = 45)



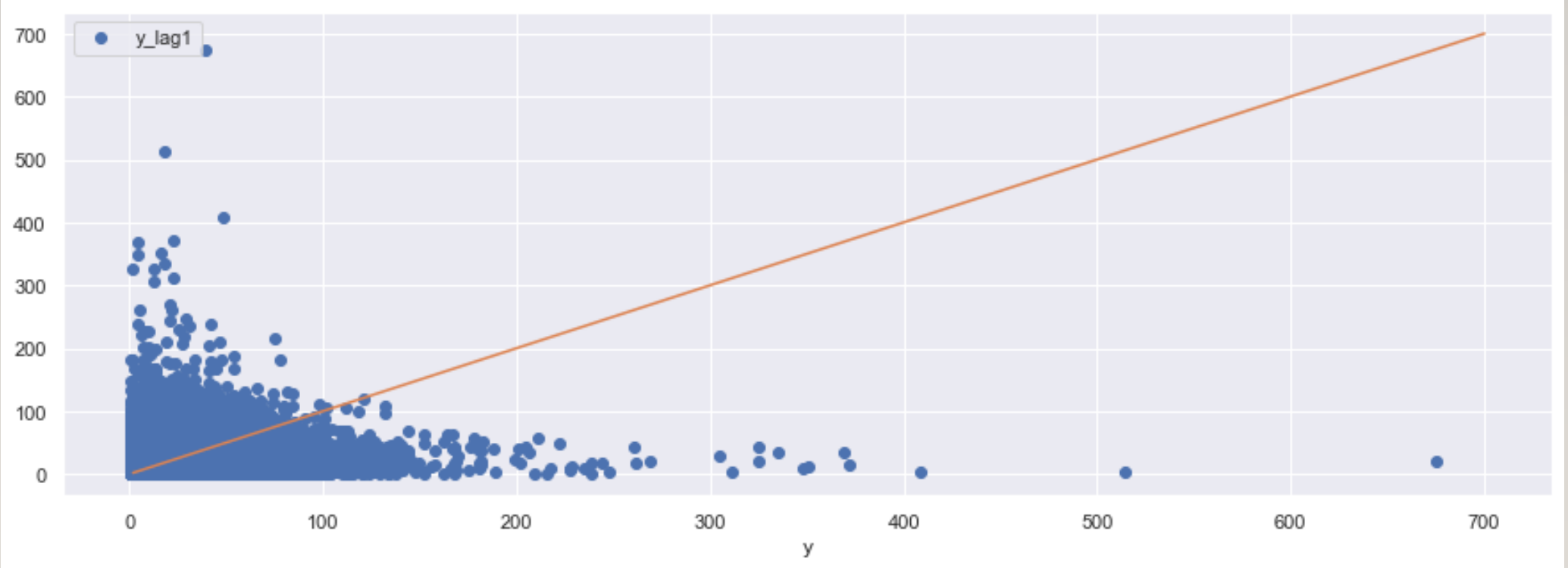
YOMATO MOVING AVERAGE OF RESERVE VISITORS(WINDOW = 45)



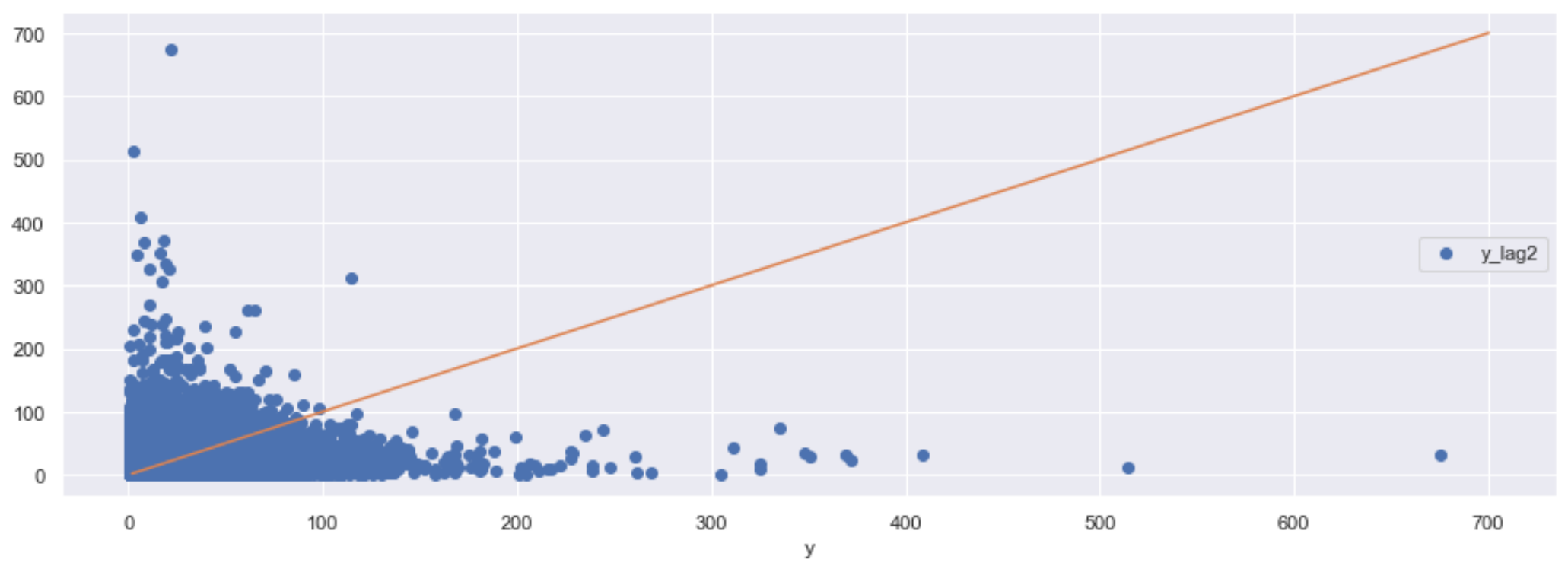
TRAINING FILE MOVING AVERAGE OF VISITORS(WINDOW = 45)



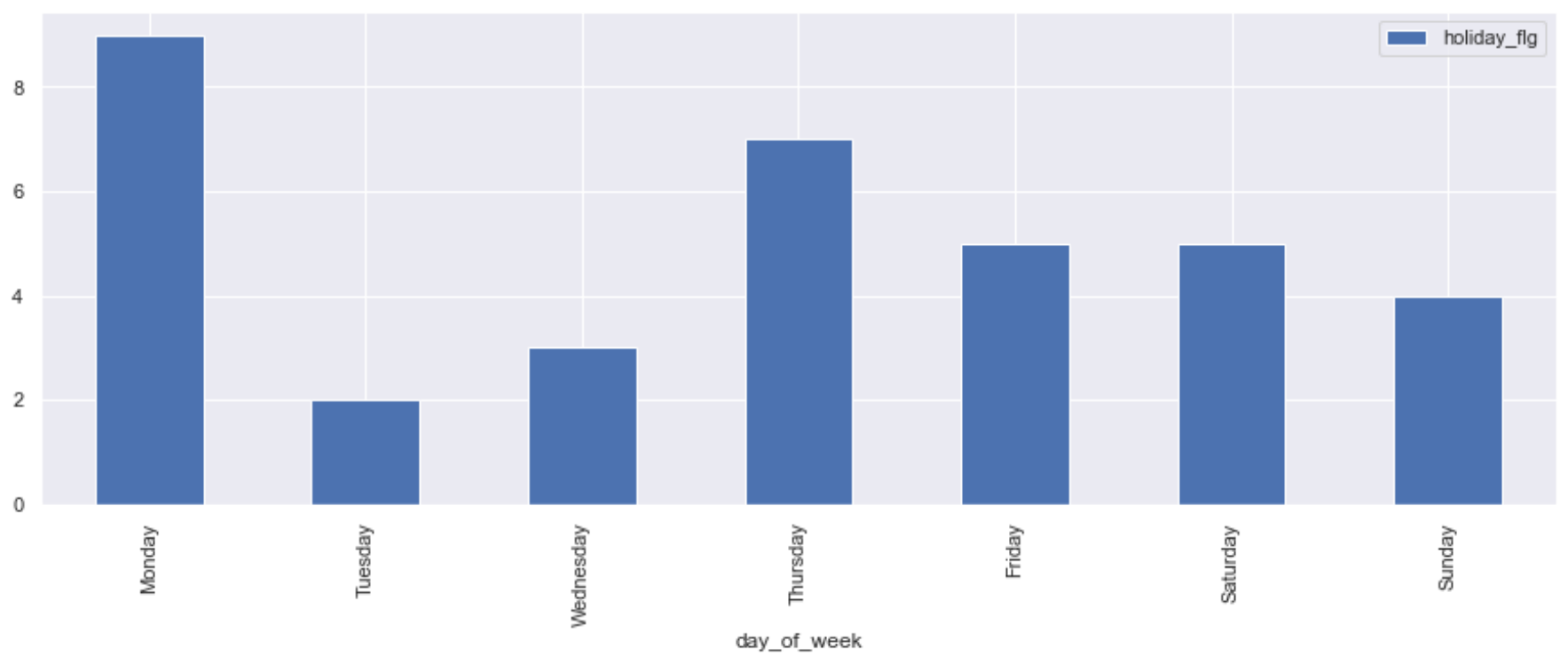
LAG GRAPH OF 1 DAY BACK FOR VISITORS ON GIVEN DATE FOR TRAINING FILE



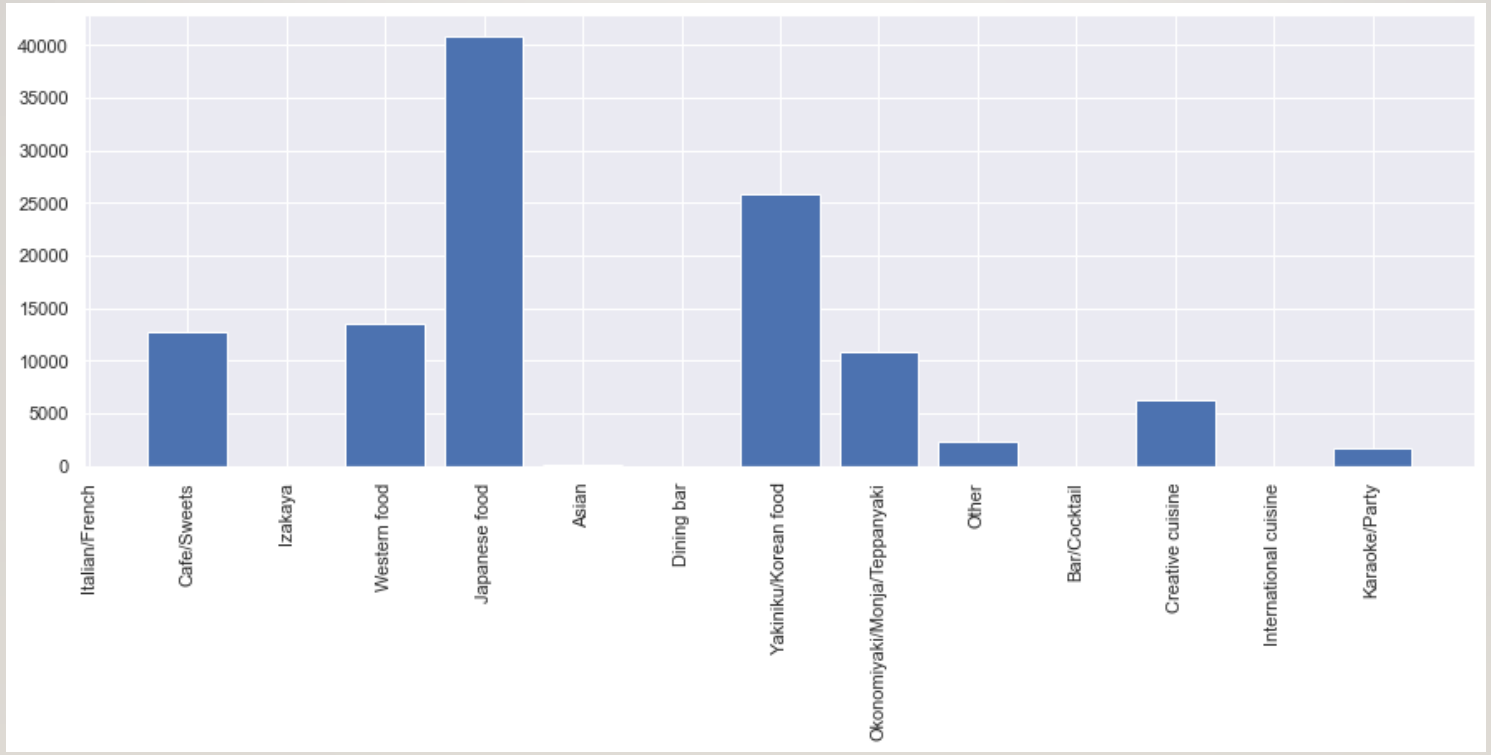
LAG GRAPH OF 2 DAYS BACK FOR VISITORS ON GIVEN DATE FOR TRAINING FILE



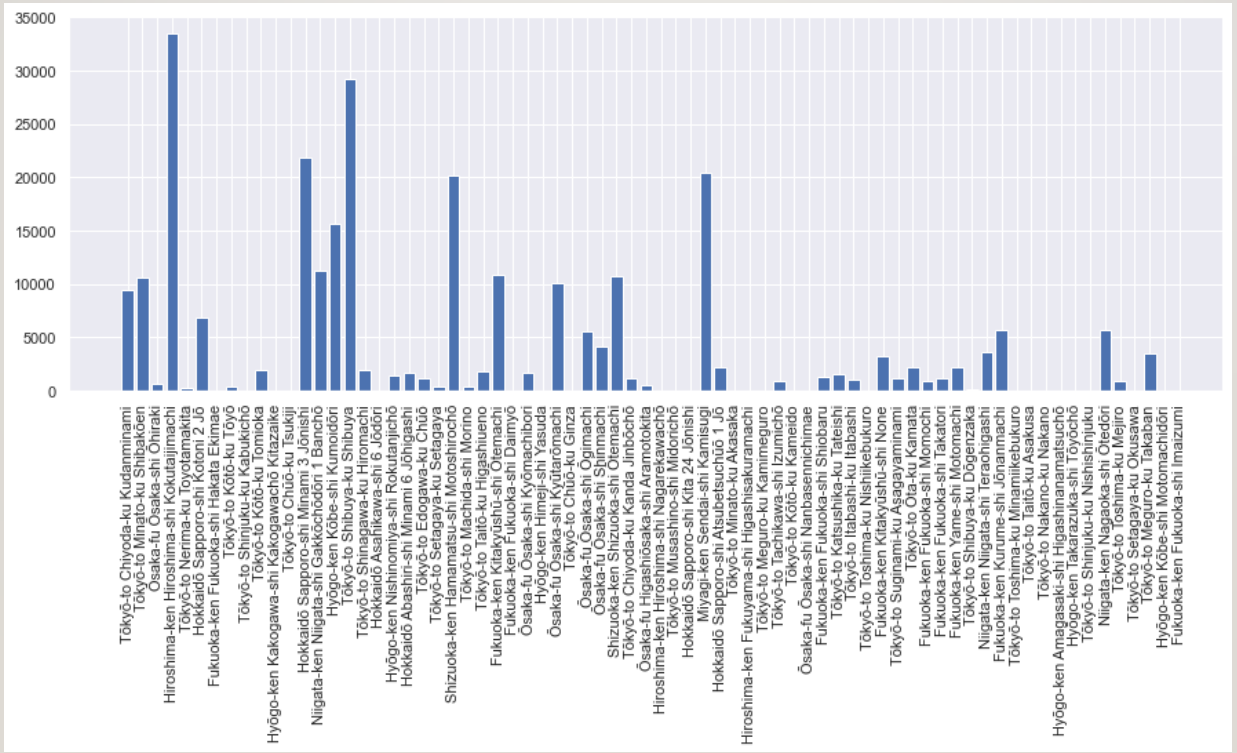
HOLIDAY COUNT VS WEEK DAY



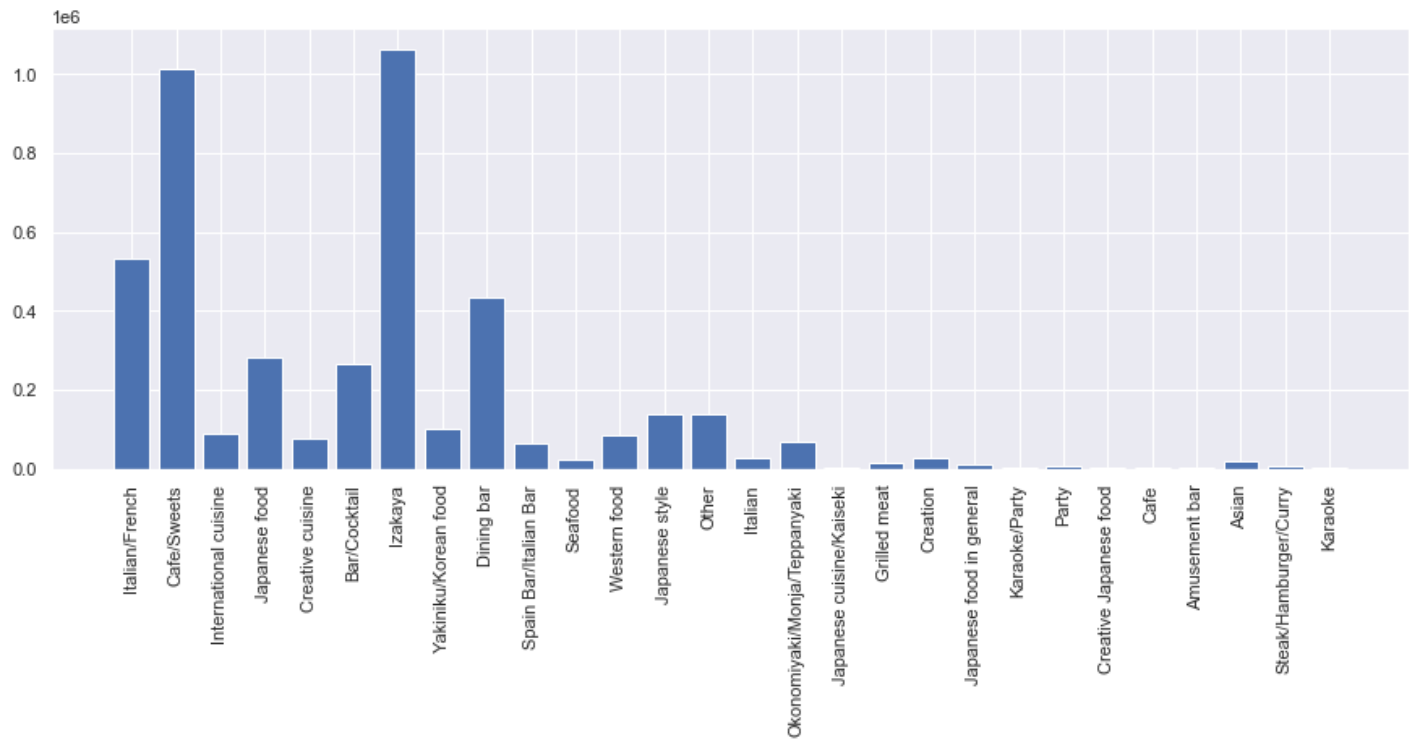
VISITORS VS GENRE NAME, MERGED BY IDS



VISITORS VS AREA NAME, MERGED BY IDS



VISITORS VS GENRE NAME, FOR TRAINING DATA



FEATURE SELECTION AND ENGINEERING

- We considered a range of features which helped us to improve the accuracy of prediction.
- In total we have 70 features including both one-hot encoded ones and self made ones.
- Following are some of the features we included :

FEATURES

1. Reserve Visitors
2. Lag features for past 6 days for reserve visitors.
3. Lead visitors features for upcoming 5 days for reserve visitors
4. Whether there is holiday or weekend on that day.
5. Whether there is holiday on the next day or not.
6. Mean, Median and Maximum visitors for particular restaurant.
7. Mean, Median and Maximum visitors for particular restaurant on given date.

8. Mean, Median and Maximum visitors for particular restaurant given month.
9. Latitude and Longitude of the restaurant.
10. Breakdown of date on day, date, month and year.
11. Week of the year
12. Features on the basis of time difference between visiting and reserving the restaurant.
13. One-Hot encoding for area name and genre name
14. Count frequency encoding for area name and genre name.

15. Arithmetic combinations on co-ordinates.

MODELS USED

- Linear Regression
- Support Vector Regressor
- Decision Tree
- KNeighbourRegressor
- Light Gradient Boosting Machine
- Extreme Gradient Boosting (Xgboost)
- Random Forest
- Stacking

BRIEF DESCRIPTION OF MODELS

We used Linear regression for the start on fewer number of features and obtained about 0.84 error with it.

Using Random Forest got 0.79 error on same number of features.

Then we tried SVR, Decision Tree, KNN. All of them gave error around 0.51.

After increasing number of features and using LGBM we got error 0.49 !!

Name of Model Used	Error
Linear Regression	0.84
Random Forest	0.79
SVR	0.51
Decision Tree	0.51
KNN	0.51
Xgboost	0.49868
Lgbm	0.49787
Stacking	0.52

CONTRIBUTIONS

Whole project was team effort and both contributed in their own way, coming up with ideas, models and code.

CONCLUSIONS

We understood the importance of feature engineering, after making real-life features which makes sense and are practical we observed drastic reduction in error.

For example using the next day holiday feature and taking sum of visitors for a particular restaurant on a particular date was a good improvement.

Using co-ordinates and genre-name as features our model improved.

REFERENCES

- All classroom material like slides, lectures.
- Discussion with TAs helped a lot
- Online video lectures from YouTube.
- Following are some websites which helped a lot.

https://stackoverflow.com/

http://www.geeksforgeeks.org/

https://www.kaggle.com/learn/time-series

http://medium.com/

https://towardsdatascience.com/

http://www.journaldev.com/

http://www.analyticssteps.com/