Final Data Challenge Report

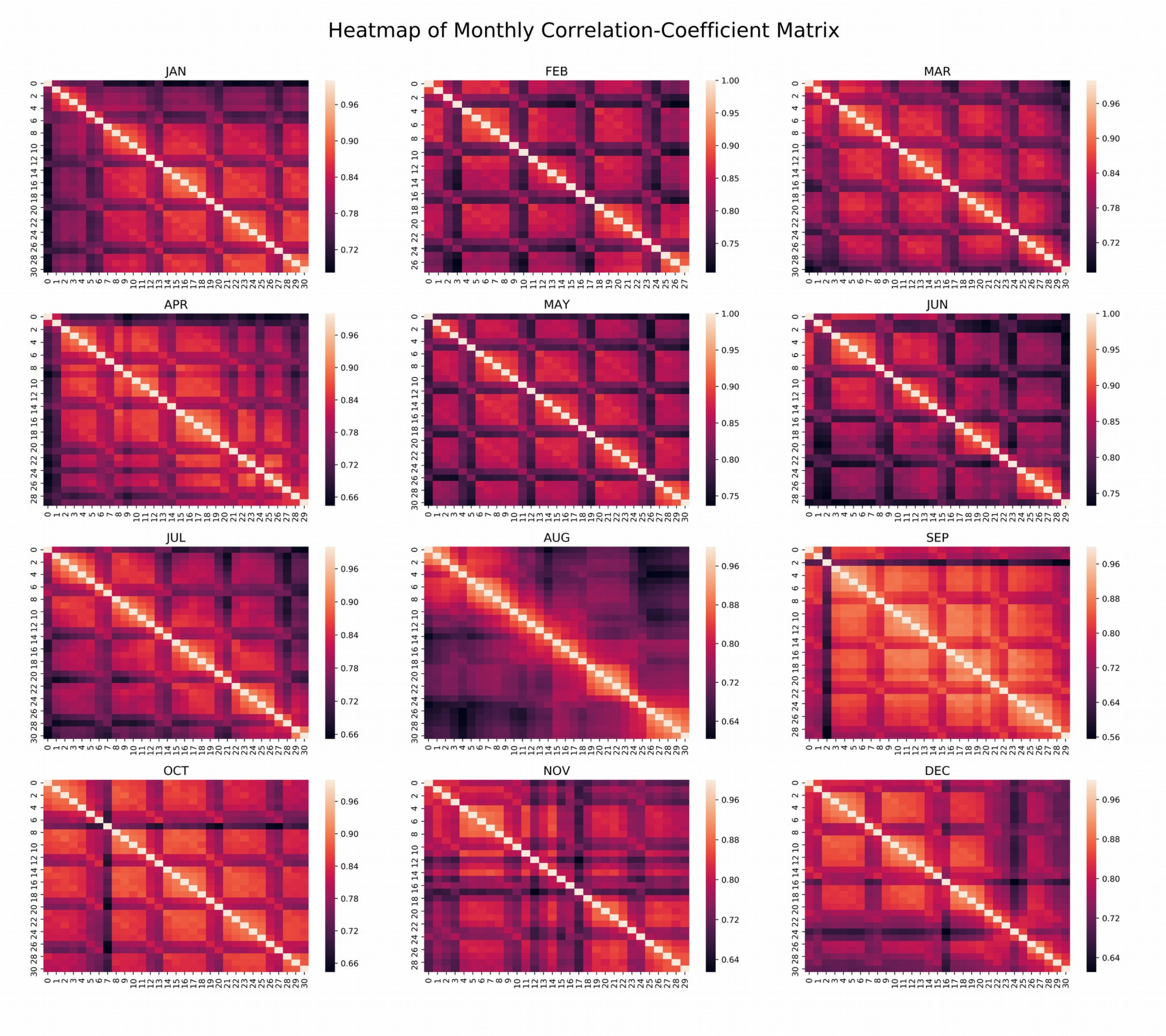
for Prof. Daniele Durante and Eni Gas e Luce

by MA Qitian 3068020

I. Exploratory Data Analysis

1) Missing Values in Training Set

Both numeric variables and categorical variables have missing data. If we drop training observations with at least one missing value, we will end up with 1/10 of the data, which is not acceptable, so we need to impute them.



The HeatMap above shows the correlation of daily energy consumption within a month. Except for August, a month for vacation, we clearly observe a weekly pattern: weekday consumptions are strongly correlated. Later plots will show this pattern more precisely. Hence, I replace missing consumption values with their corresponding weekdays / weekend mean. In rare occasions where the entire weekend data is missing, I replace them with the weekend mean of their corresponding month. Since the target is to predict the energy consumption of a single day, this pattern suggests that not all variables are useful for prediction and variable selection can come into play.

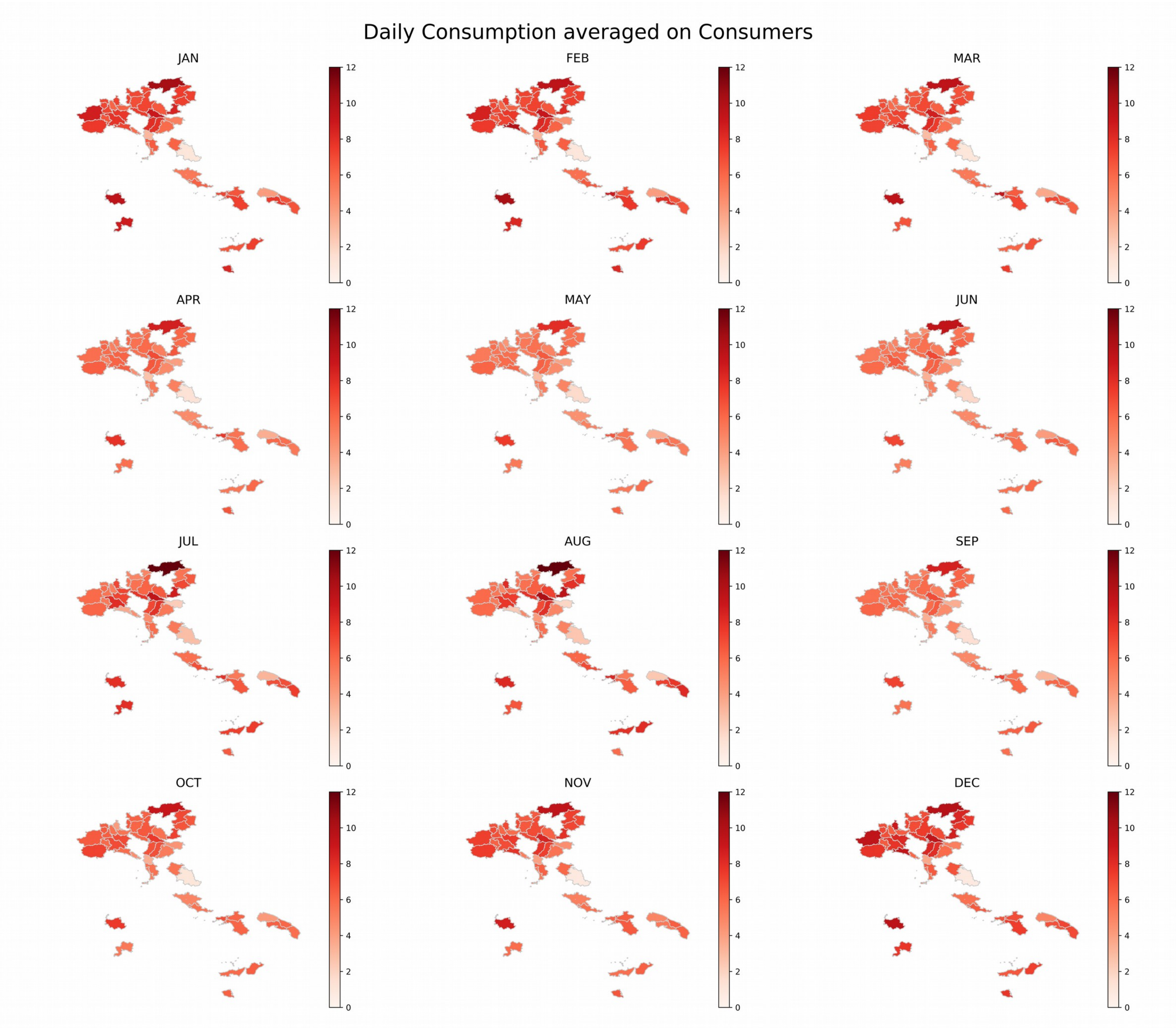
In terms of location, the data contains 224 CAPs, 212 CITYs, 44 PROVINCEs and 14 REGIONs. These geographical divisions are nested. Since the most sophisticated division CAPs contain no missing data, missing data from other more general divisions can be properly inferred.

About 20% of GENDERs is missing, the p-value for t-test is 0.0046, suggesting a strong difference between the two gender. Therefore I replaced them with a tag value.

About 28% of AGE_CUSTs is missing and there is no way we can infer that, so I dropped this variable. An alternative way could be to fill them with 0 and add another dummy variable.

The missing values in TYPE_OFFER and FAMILY_OFFER occur simultaneously, which suggests a special type of consumer so I replaced them with a tag value.

For other variables, the number of observations corrupted is very small, so I dropped the observations.



2) Missing Values in Test Set

For missing consumption data, I did the same as the training set, taking weekly pattern into consideration.

For missing categorical data in test set, things become harder since we cannot drop any observations. First, I defined similarity between an observation and a category by daily mean absolute difference in consumption between the observation and aggregated mean of the category and then I replaced the missing value with the closest category.

3) Categorical Difference between Training Set & Test Set

Some of the categorical inputs have rare categories that appear in the training set but not in the test set and vice versa. In the first scenario, I dropped the observations because it is not worthwhile to include extra dummy variables. In the second scenario, I inferred these category by the similarity defined in 2).

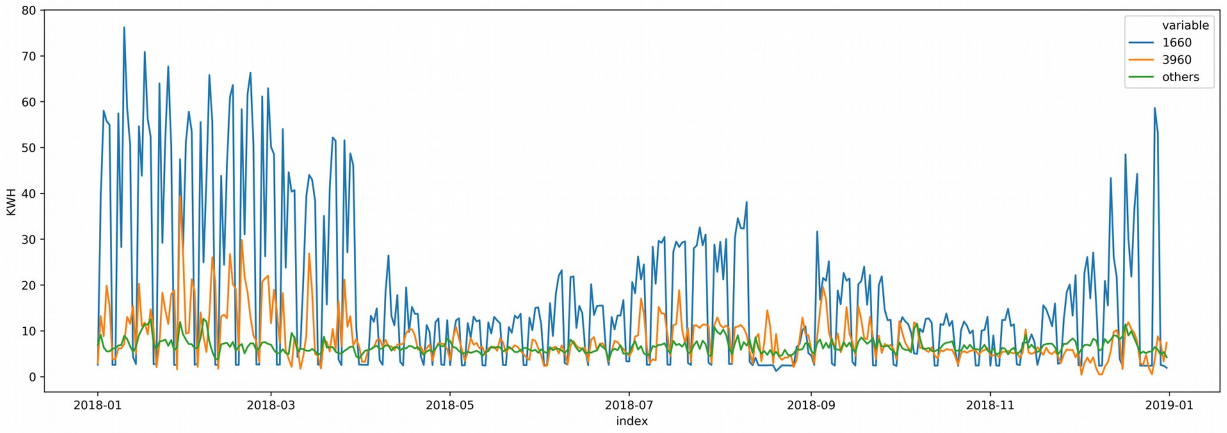
4) Variable Selection

Too many numerical variables and categorical variables with too many categories may cause overfitting. Hence we need to apply variable selection.

First we need to decide on which level do we differentiate the geographical locations, namely CAP, CITIES PROVINCE or REGION. Since the total number of categories are too big for a model to choose, the decision is better to be made by human. From the GeoMap in the previous page, we can easily see that even on the PROVINCE level, there are similarities between some neighboring provinces. On the other side, region may be too rough. Therefore I decided to keep PROVINCE and drop the others. I left the rest variable selection task to the model itself.

5) Outliers

From the Monthly Consumption plot below, we observe that IMPRESE behaves wildly. When we look into it (see the plot below), the wild behavior is driven by two irregular consumers 1660 and 3960 (indexed according to the order of the original training set). Therefore, I dropped these two observations.



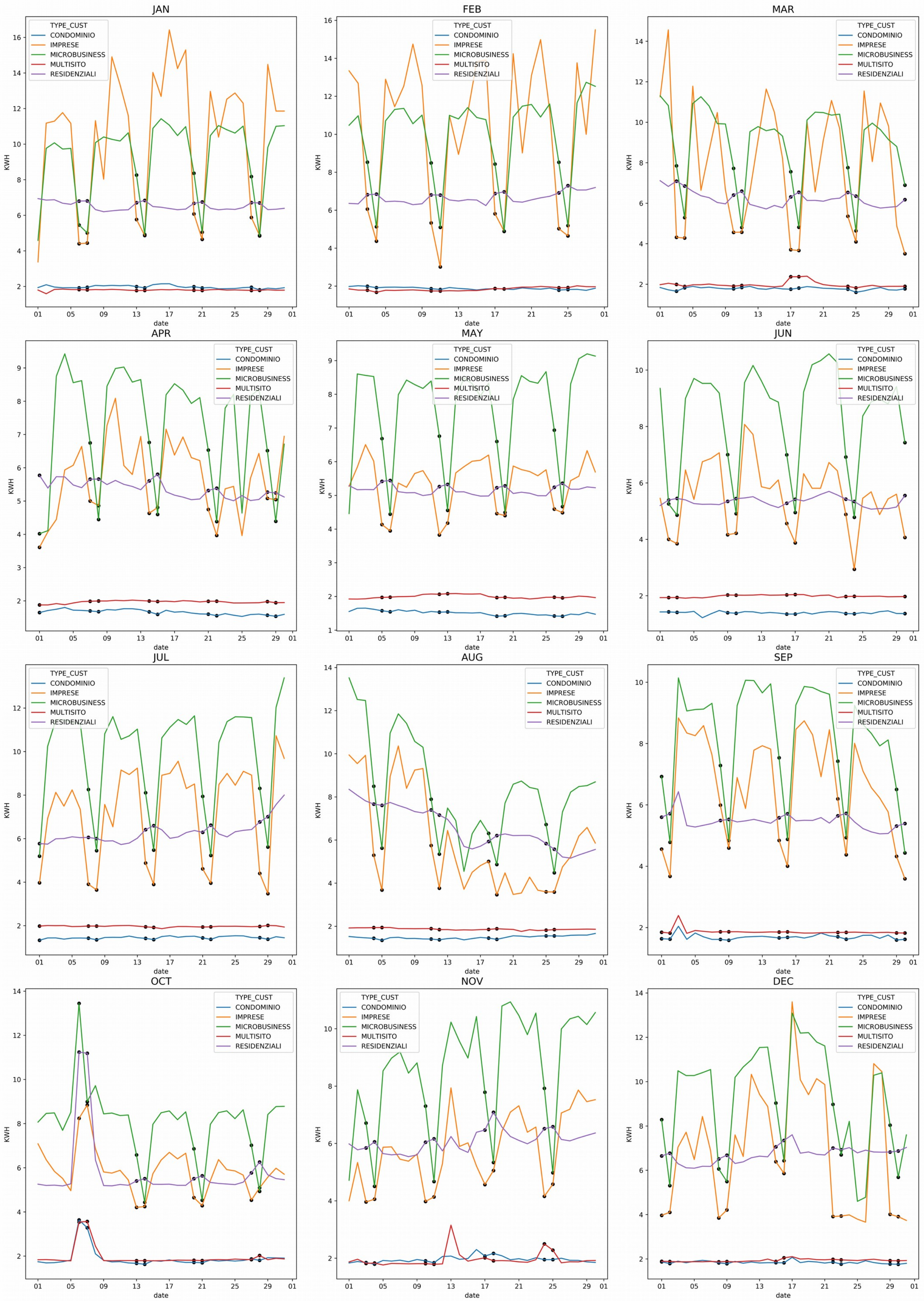
II Model

I applied LASSO with 10 cross-validation. The best alpha is 0.0002784 with 122 variables selected including 1 GENDER, 1 PAYMENT, 1 TYPE_OFFER, 4 FAMILY_OFFER, 13 PROVINCE, 3 ALTIMETRIC_ZONE, 1 CLIMATE ZONE. The consumption variables picked are as follows (counts / average coefficient)

Sun	Mon	Tue	Wed	Thu	Fri	Sat
13 / 0.139	12 / -0.0463	13 / -0.0435	14 / 0.0483	8 / 0.0179	13 / 0.0496	25 / 0.836

Since Jan. 12th is a Saturday, Saturday appears most frequently and has the biggest effect as expected. The category of GENDER selected is the missing tag (not M/F). This interesting phenomenon may be related to data recording method.

Monthly Consumption (dark points specify weekends)



III Future Improvement

Since many categorical variables are correlated, dropping some of them manually may improve the score.

Lasso does a very good job in variable selection but did not take the order of the data into consideration. Other model specially for time series data can improve the score.

After variable selection, other nonlinear model may improve the score further.

Iv Interesting Findings

From the HeatMap, we can see that for a particular consumer, North one consumes more energy than South one.

Consumption in Summer and Winter are more than that in Spring and Autumn, probably due to air-conditioning.

From the LinePlot, we can see that business consumption follows weekly pattern but residence consumption does not.

There is an abnormal consumption spike on October 6th and 7th. I wonder what happened on that day.