Food Demand Forecast

Business, Economic and Financial Data Project

Pierpaolo D'Odorico, Massimiliano Conte and Eddie Rossi

Food Demand Forecasting

The business problem:

A **meal delivery company** operates in multiple cities. They have various **fulfillment centers** in these cities for dispatching **meal orders** to their customers.
We need to **forecast** for upcoming weeks, so that these centers will **plan the stock** of raw materials accordingly.

Task:

Predict the demand for the next **10 weeks**!

Data sources

Data are collected in 3 different datasets, connected by keys.

Datasets

- Fulfilment centers data
- Meal info data
- Sales historical data

Fulfilment centers data

center_id	city_code	region_code	center_type	op_area
11	679	56	TYPE_A	3.7
13	590	56	TYPE_B	6.7
124	590	56	TYPE_C	4.0
66	648	34	TYPE_A	4.1
94	632	34	TYPE_C	3.6
64	553	77	TYPE_A	4.4

Fulfilment centers data features

Variables:

- **center_id**: Fulfilment identifier
- **city_code**: City id in which the center is located on
- **region_code**: Region id in which the center is located on
- **center_type**: Type of the center
- **op_area**: Size of the operational area

Unique values in dataset:

- **center_id**: 77
- **city_code**: 51
- **region_code**: 8
- **center_type**: 3

Meal info data

meal_id	category	cuisine
1885	Beverages	Thai
1993	Beverages	Thai
2539	Beverages	Thai
1248	Beverages	Indian
2631	Beverages	Indian
1311	Extras	Thai

Meal data features

Variables:

- **meal_id**: Meal identifier
- **category**: Category of food
- **cuisine**: Category of cuisine

Unique values in dataset:

- **category**: 14
- **cuisine**: 4

Sales info data

id	week	center_id	meal_id	checkout_price
1379560	1	55	1885	136.83
1466964	1	55	1993	136.83
1346989	1	55	2539	134.86
1338232	1	55	2139	339.50

id	base_price	email	homepage	num_orders
1379560	152.29	0	0	177
1466964	135.83	0	0	270
1346989	135.86	0	0	189
1338232	437.53	0	0	54

Sales data features

Variables:

- **id**: Id of the single transaction
- **week**: Temporal variable, we have 145 unique weeks
- **center_id**: Fulfilment identifier
- **meal_id**: Meal identifier
- **checkout_price**: Paid price for the product
- **base_price**: Full price of the product without promotion
- **emailer_for_promotion**: Binomial, promotion email or not
- **homepage_featured**: Binomial, product on web homepage
- **num_orders**: Number of orders for the meal and center

Sales data features

Unique values in dataset:

- **week**: 145
- **center_id**: 77
- **meal_id**: 51
- **emailer_for_promotion**: 2
- **homepage_featured**: 2

Create a unique dataset

We created a unique dataset **merging by keys**.

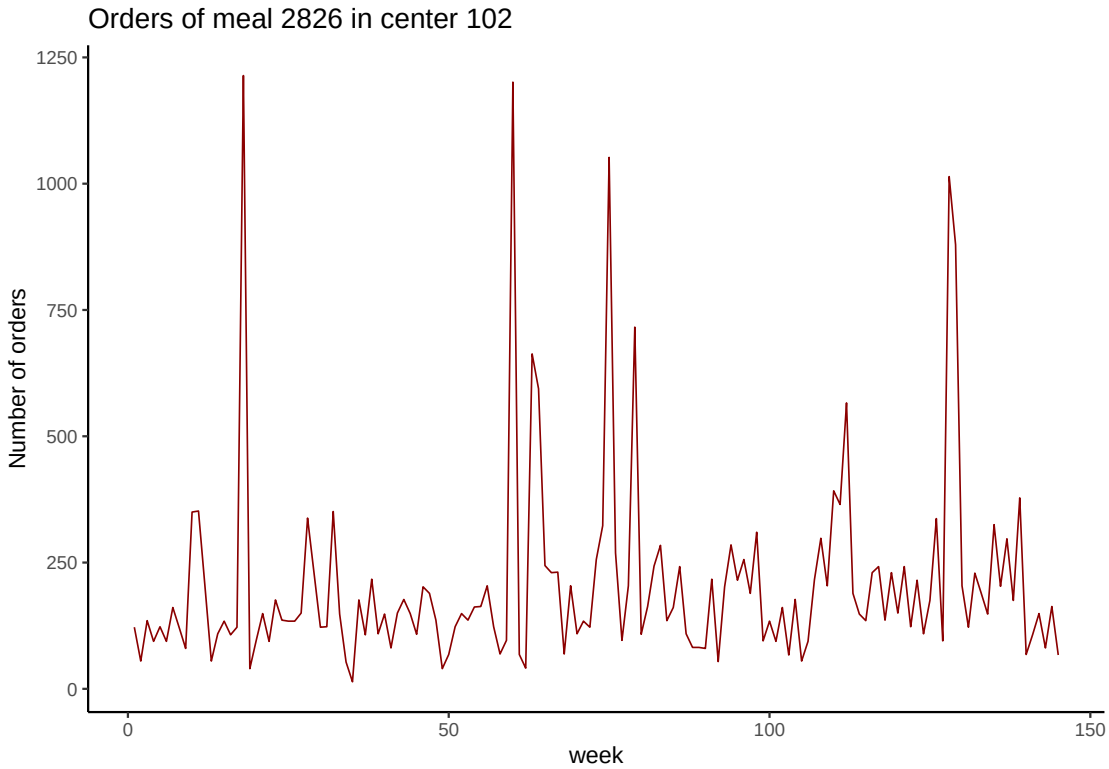
There are **0 NA's** in the complete dataset.

We will perform some exploratory data analysis:

- **Time series Exploration**: First look at time series behaviour
- **Univariate Analysis**: Looking at single variables behaviour
- **Multivariate Analysis**: Correlation between variables

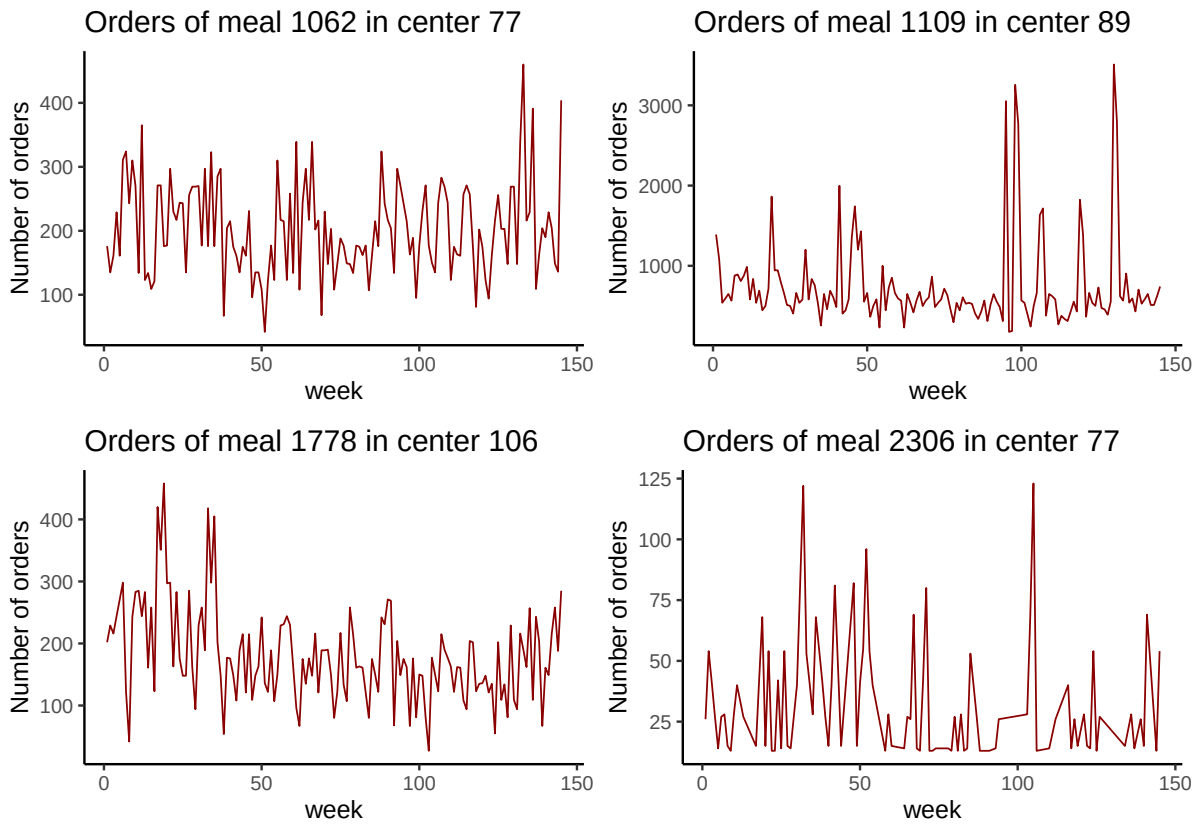
Time series Exploratory Analysis

In this business problem we deal with 77 **centers** and 51 **meals** for each center. We plot a random chosen series:



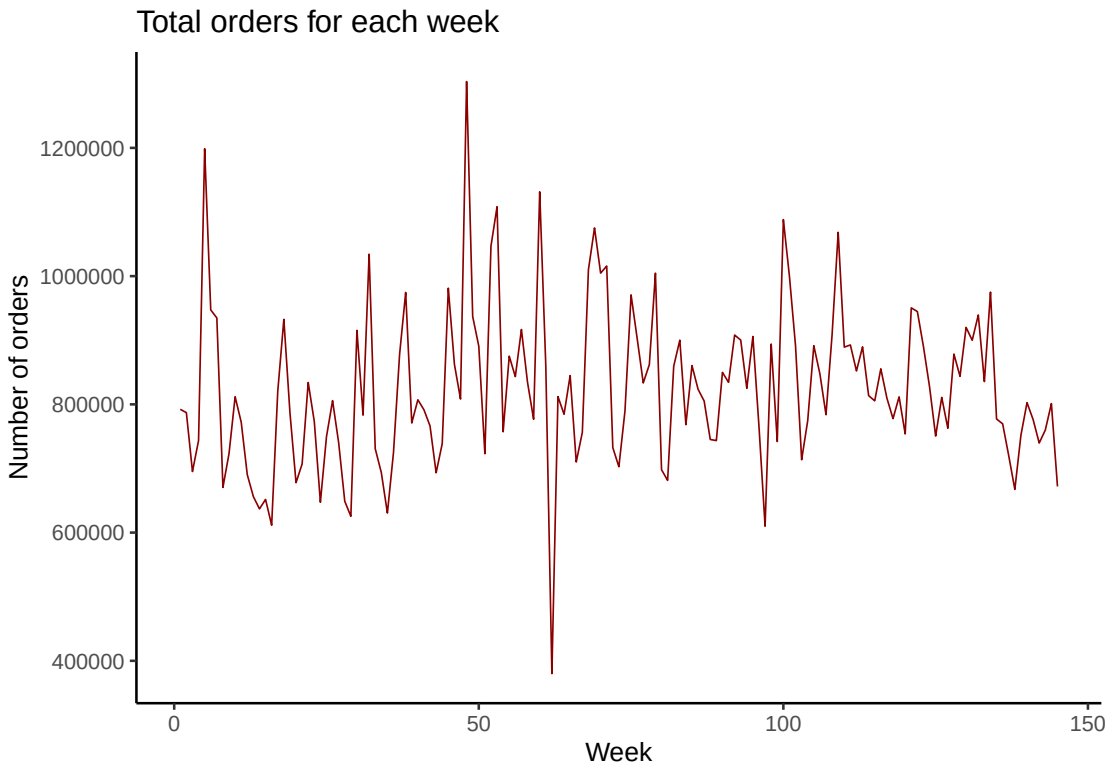
Time series Exploratory Analysis

Some other **examples** of time series:



Time series Exploratory Analysis

For looking to the **main trend** in the whole data we plot the **total number of orders** series.



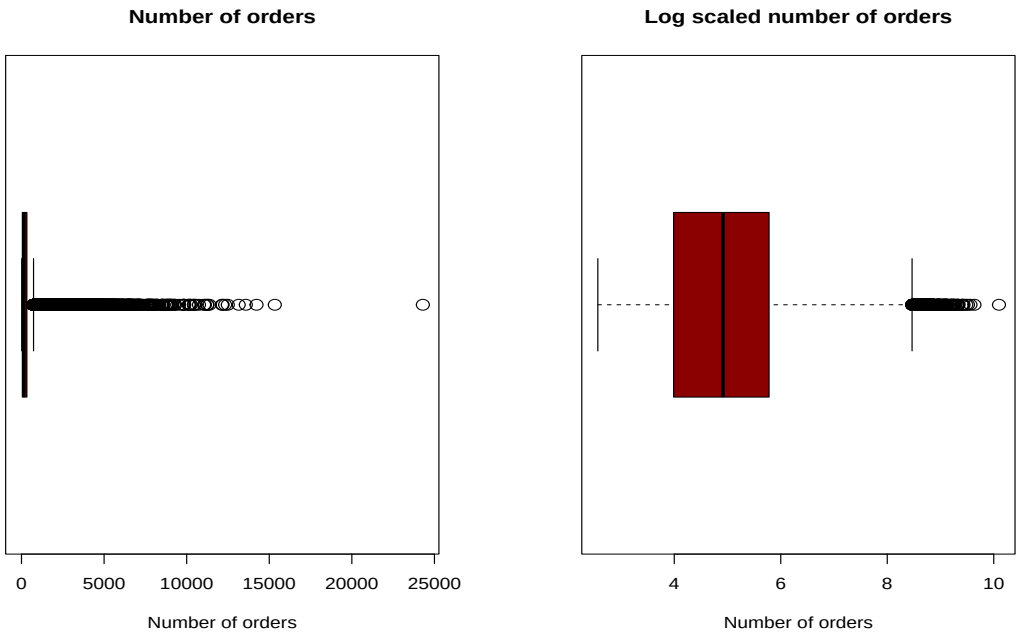
Time series Exploratory Analysis

Since we do **not** notice a **main trend** during weeks we will make some **analysis on data** that **doesn't involve** the **time dependence**.

With this assumption we can explore **variables behaviour** regardless of the week involved.

Univariate Analysis on numerical variables

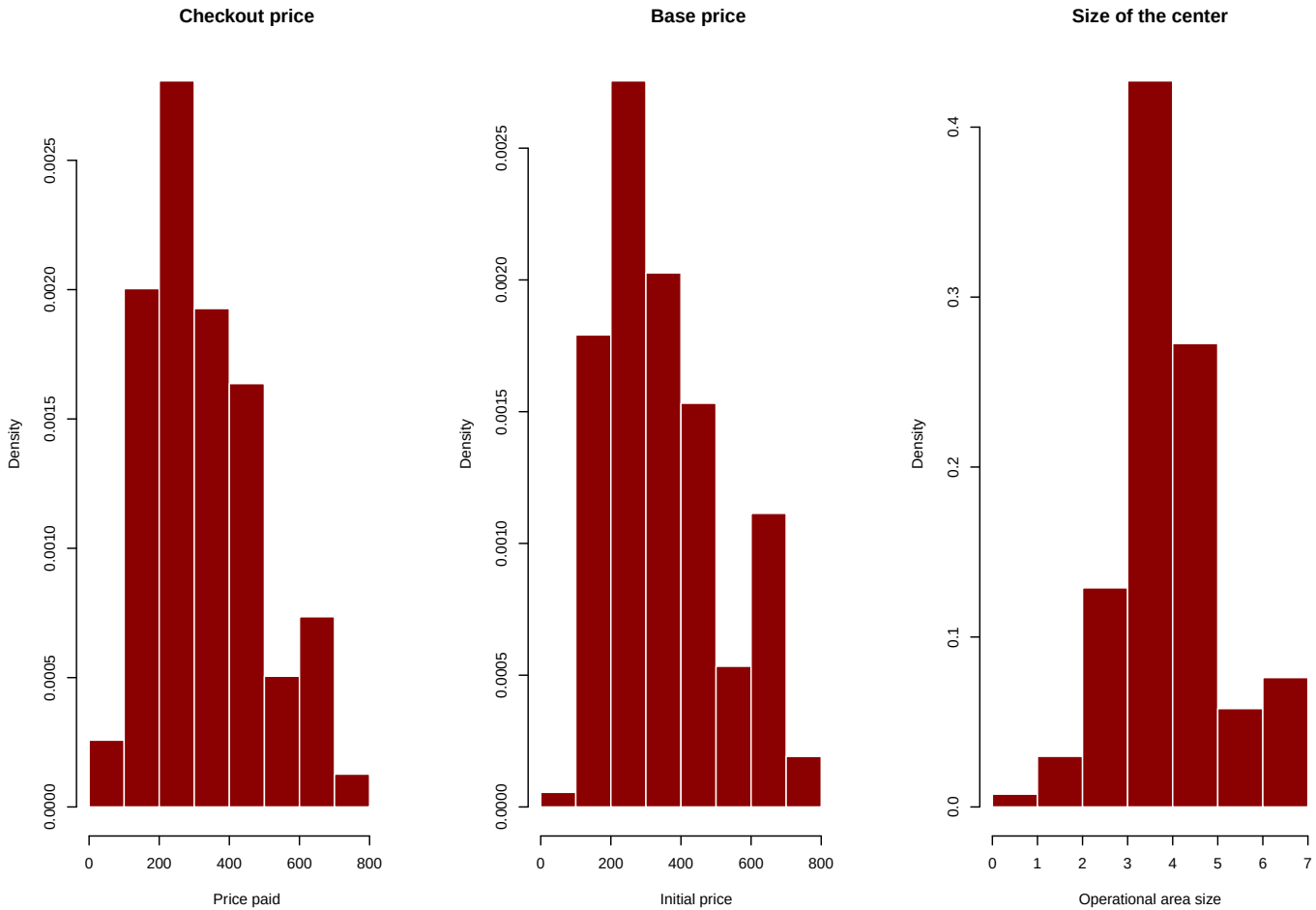
Number of orders boxplot vs log transformation for a better view:



High **number of orders** are related to some specific series from a specific center and meal with high demand. For this reason we don't consider them outliers.

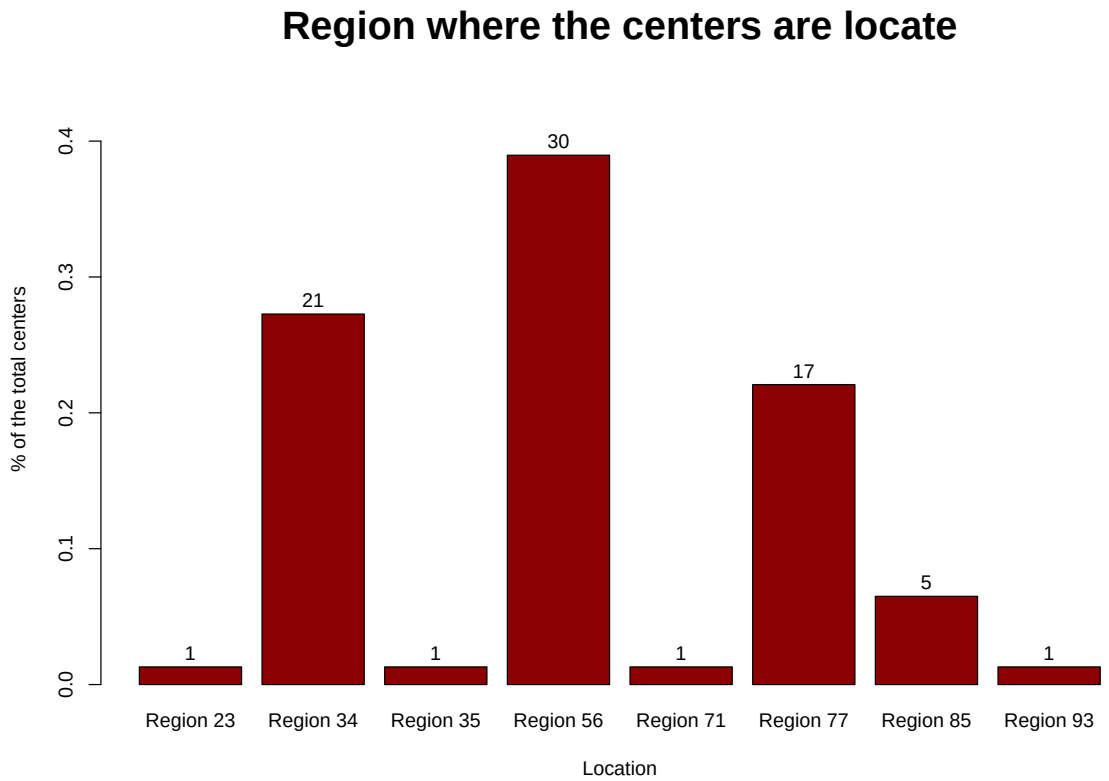
Univariate Analysis on numerical variables

We look at density of the numerical variables:



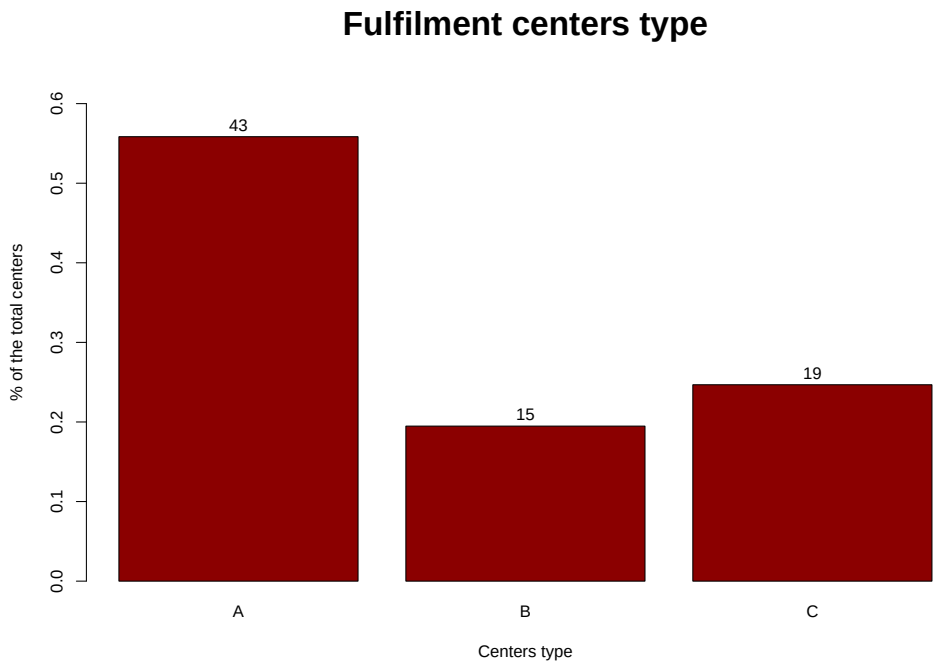
Univariate Analysis on categorical variables

We look at the region where centers are located

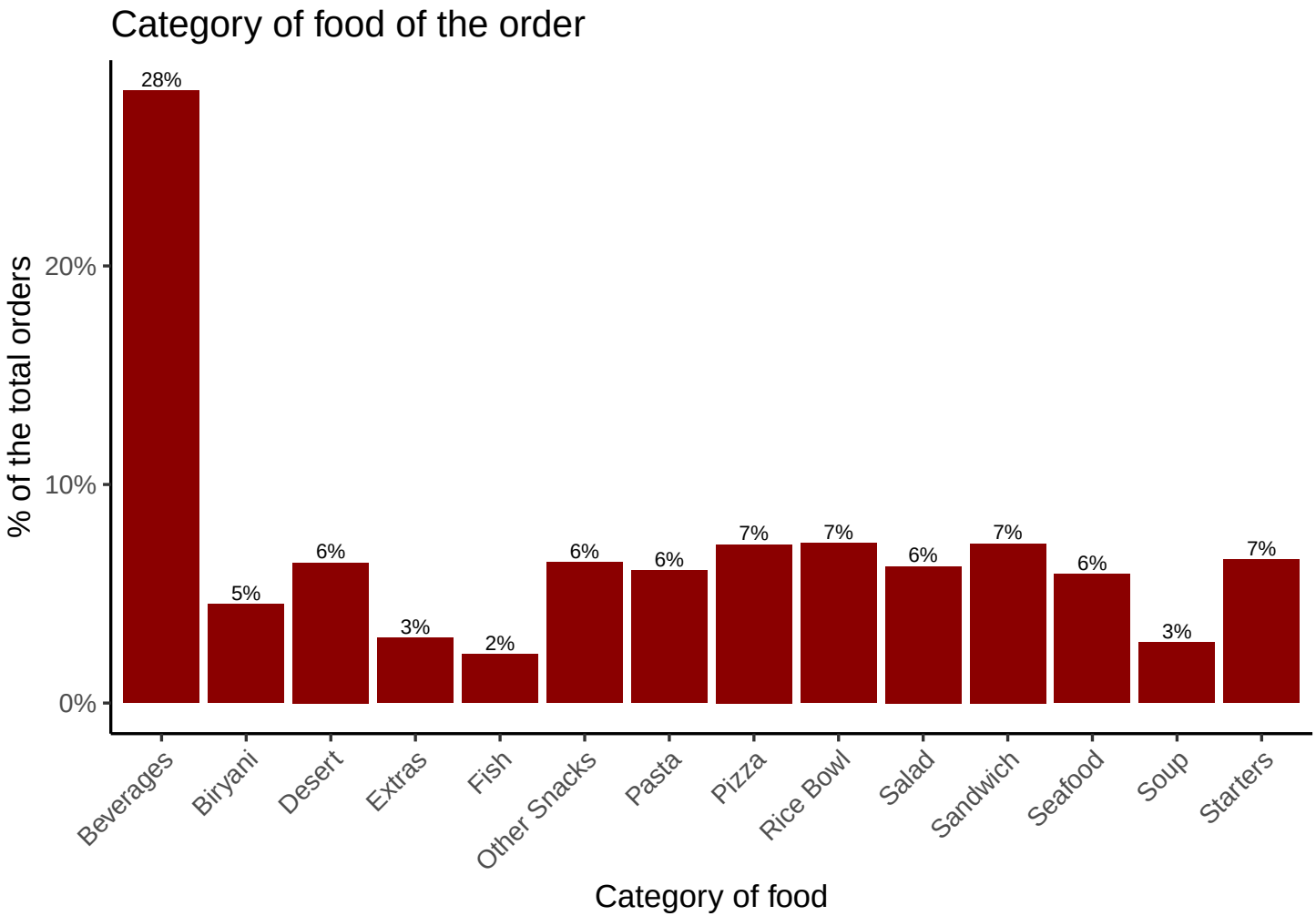


Univariate Analysis on categorical variables

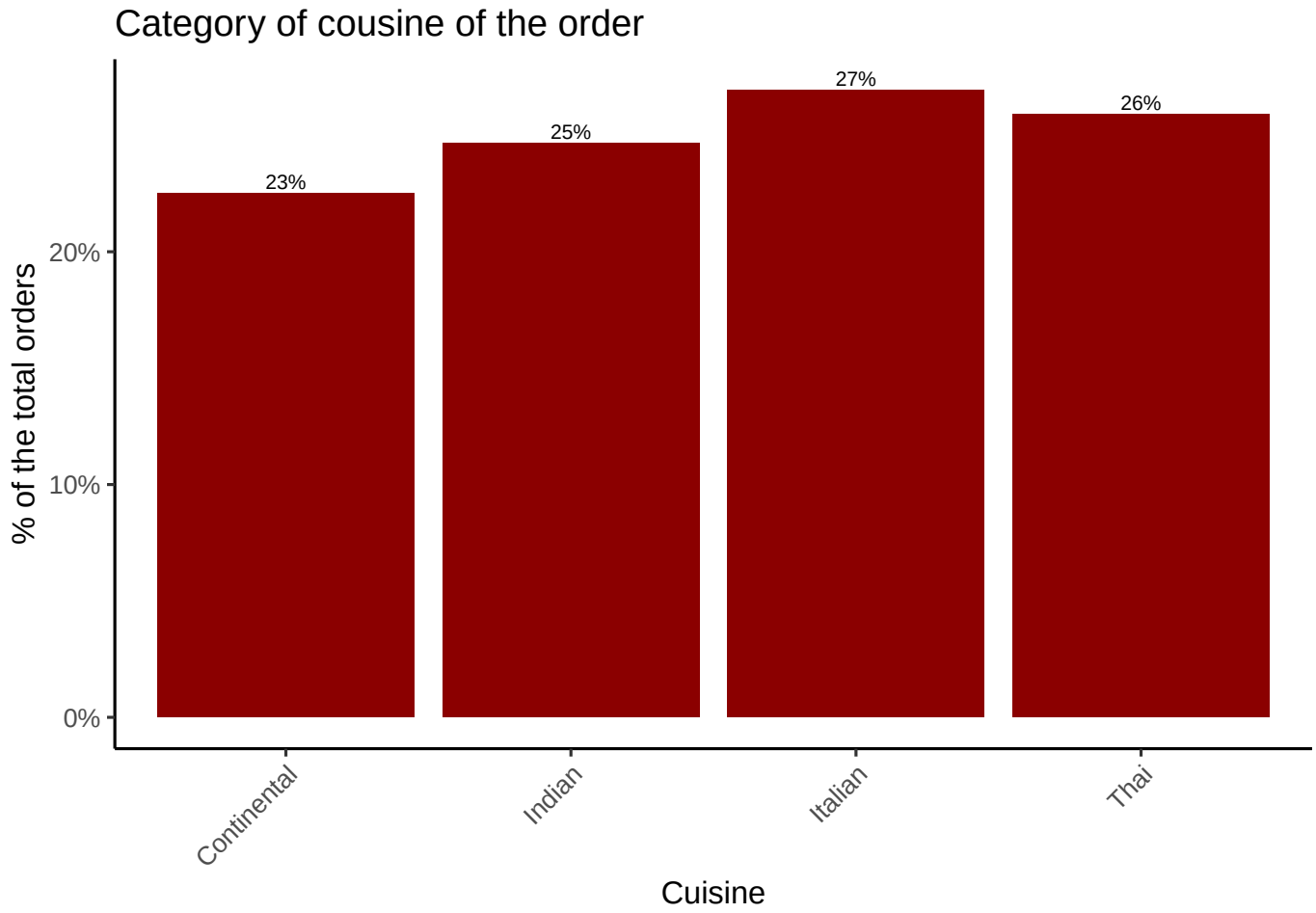
We have A, B and C center type.



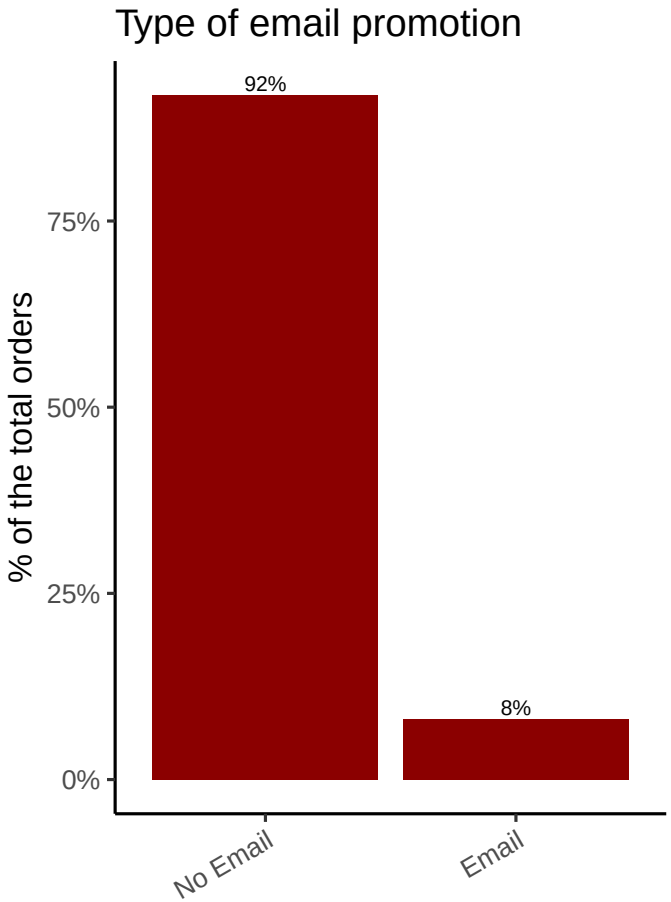
Univariate Analysis on categorical variables

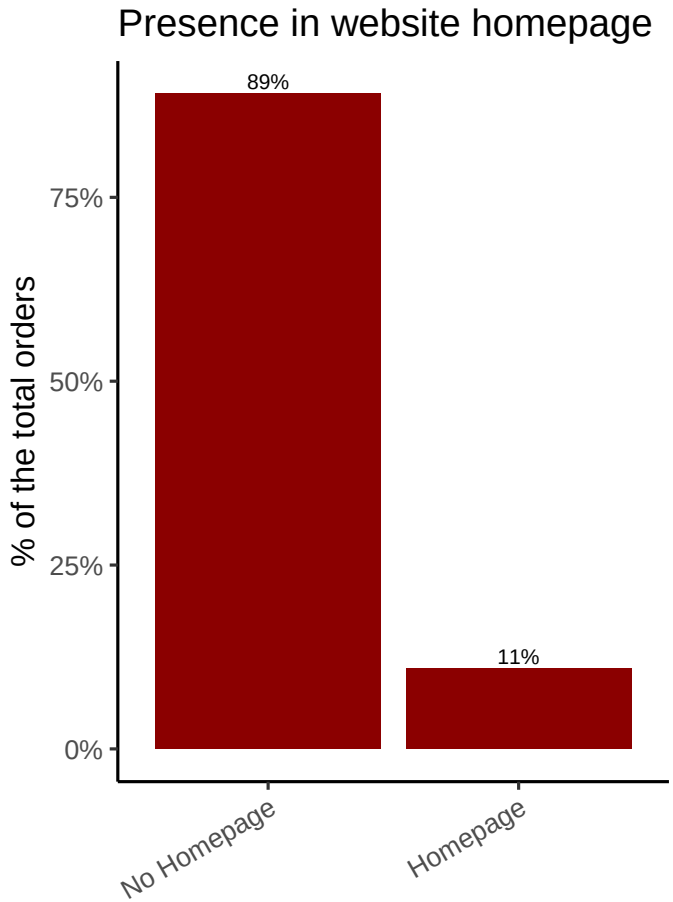


Univariate Analysis on categorical variables

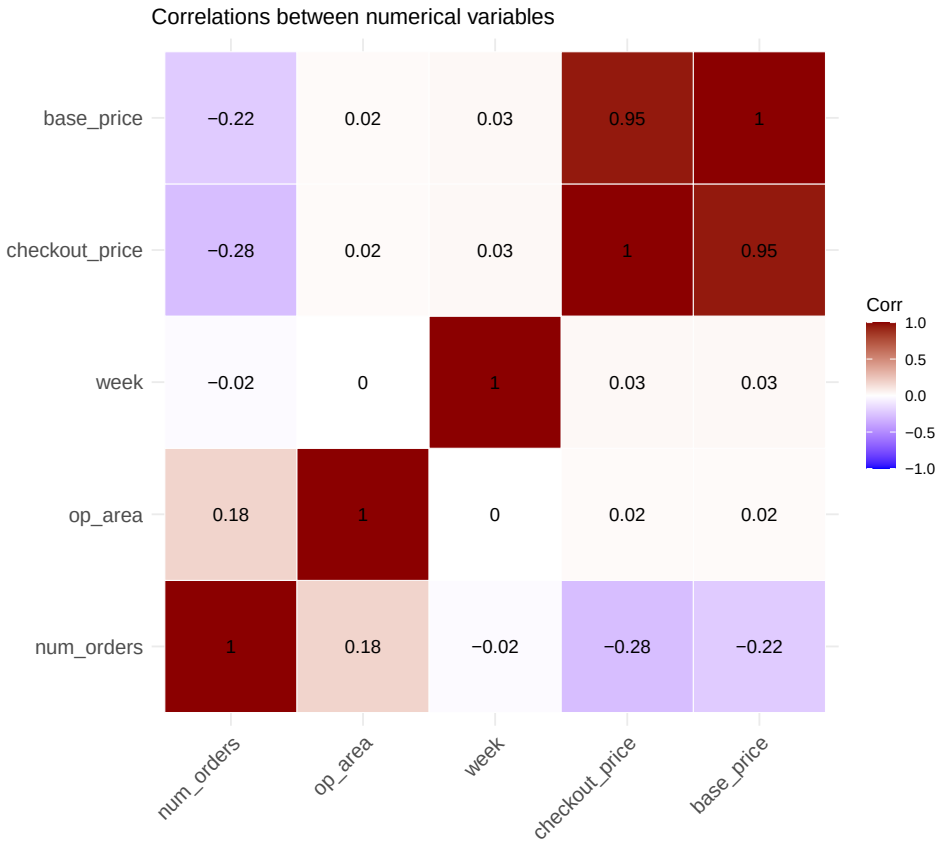


Univariate Analysis on categorical variables





Multivariate Analysis, correlation plot

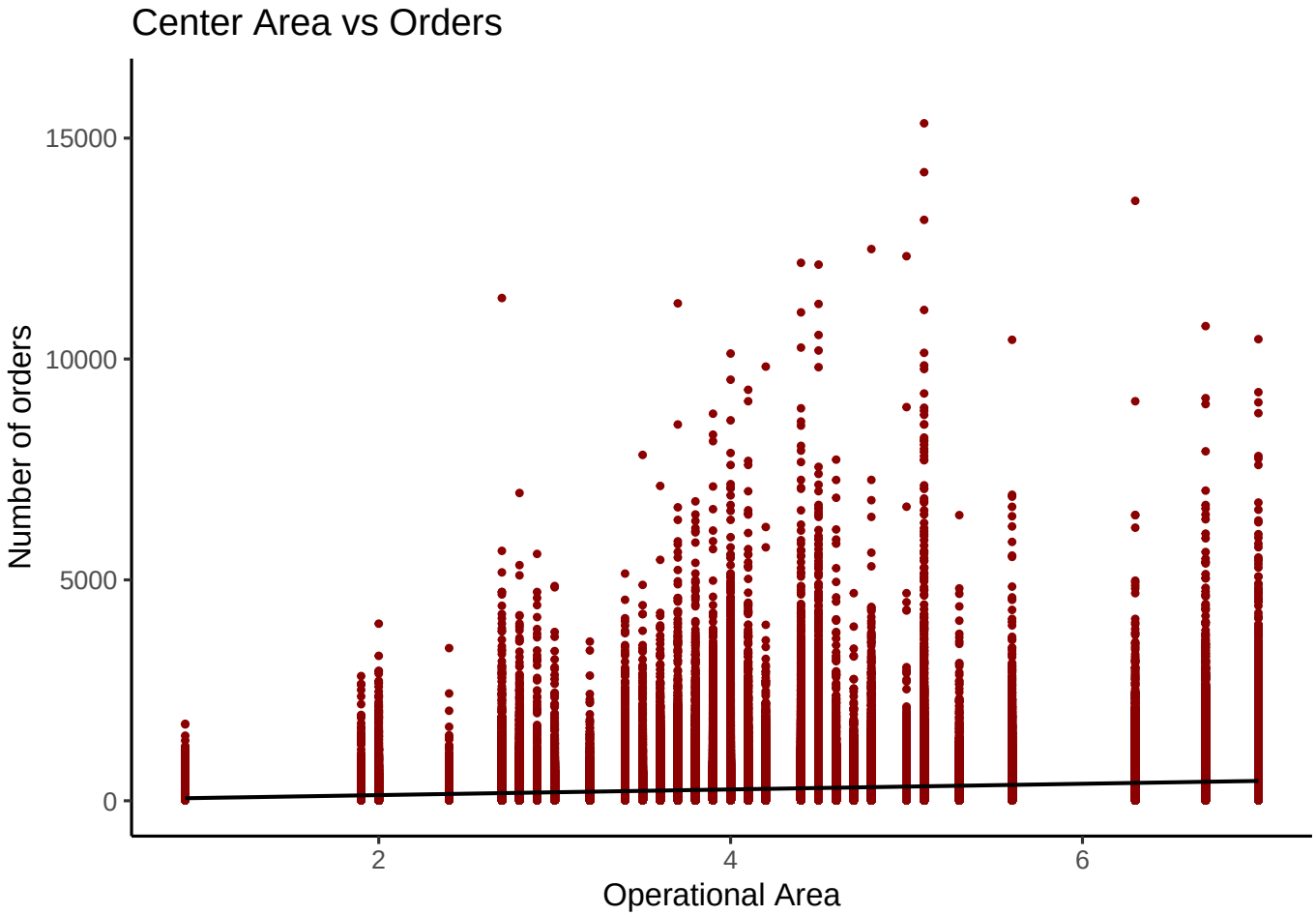


The **number of orders** is not highly correlated with other numerical variables.

Multivariate Analysis, numerical variables

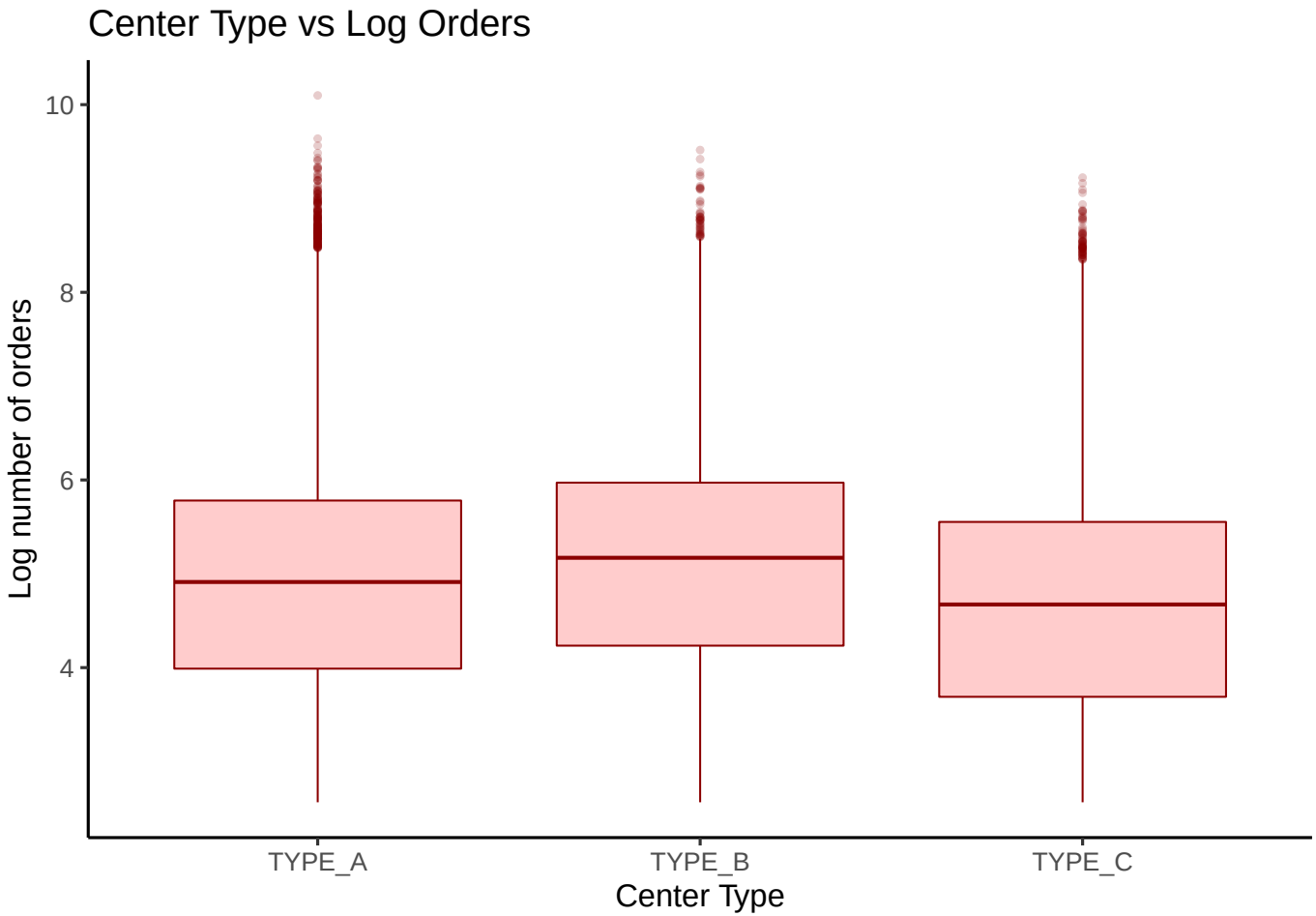
We decided to **remove Checkout Price** due to the nature of the variable, in a **real case scenario** we can't have a checkout price because checkout means that the order is confirmed.

Multivariate Analysis, numerical variables

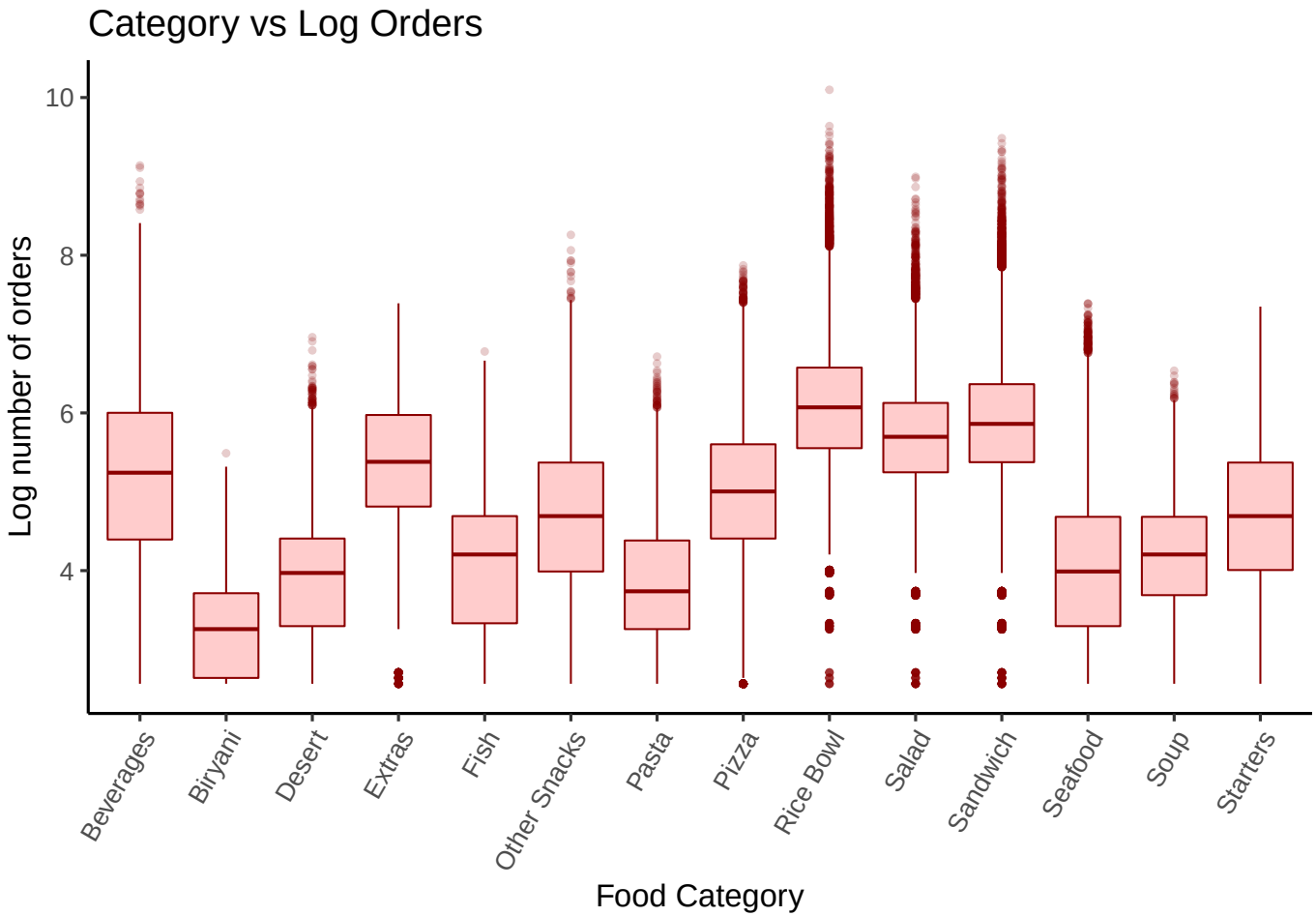


Number of orders seems to **lightly increase** for bigger area centers.

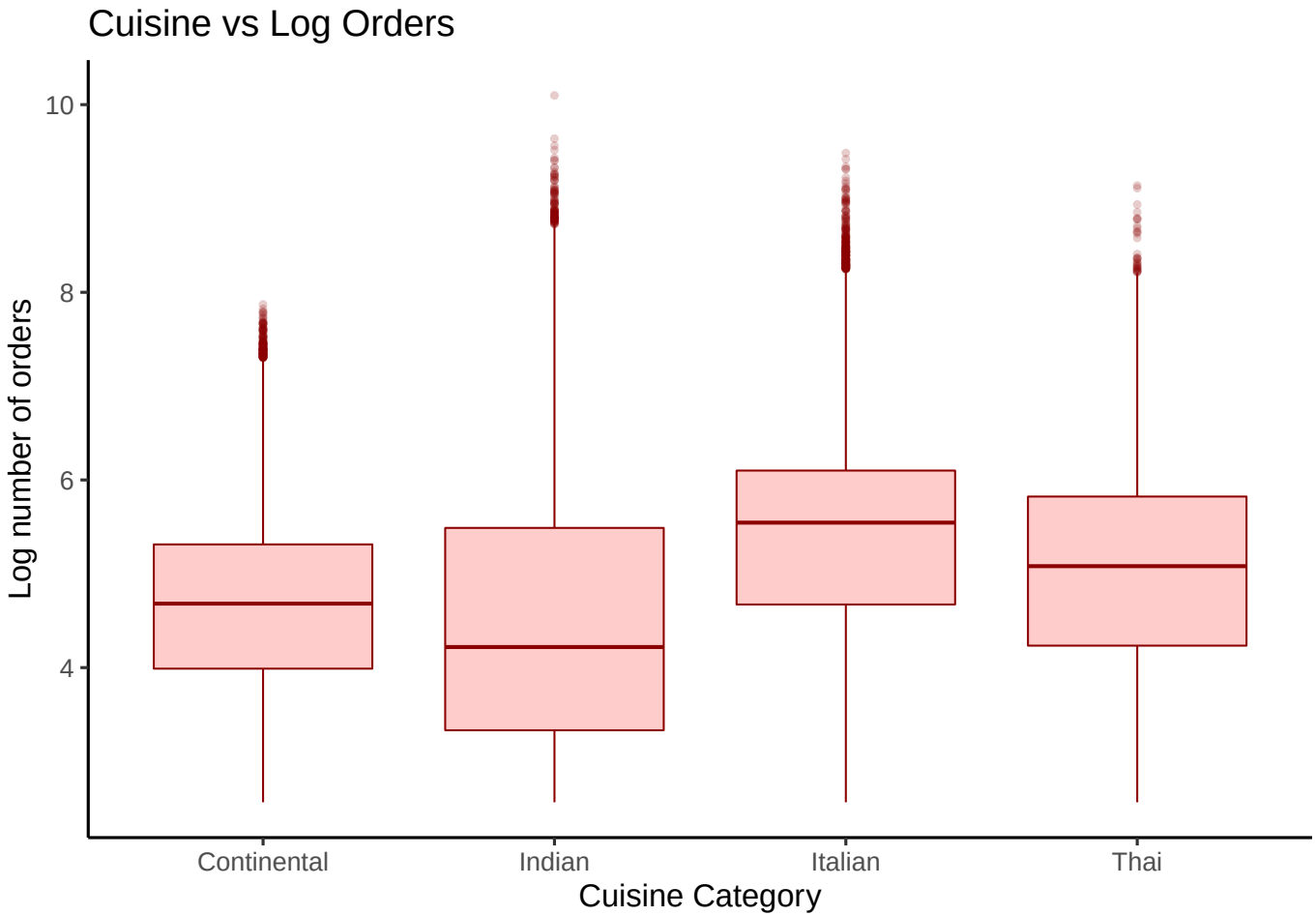
Multivariate Analysis, categorical variables



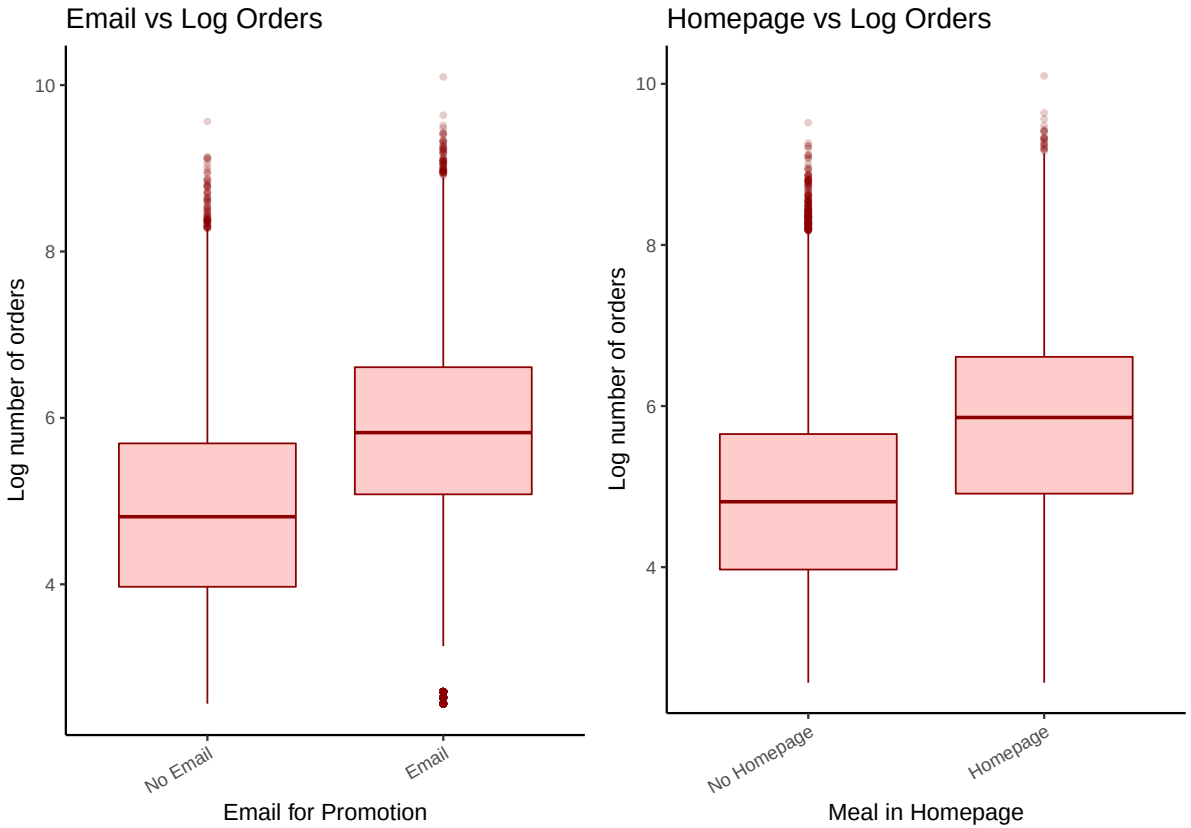
Multivariate Analysis, categorical variables



Multivariate Analysis, categorical variables



Multivariate Analysis, categorical variables



Promotions ad emails seem to **increase** the number of orders.

Modelling

Since we want to organize the goods for each specific fulfillment center, we need to forecast the demand for each specific center. Moreover, we also need to stratify for each unique meal, since each of them requires a different set of raw materials. We propose a two-stage approach:

- First we account for the temporal relationship using the linear model, obtaining (hopefully) i.i.d. residuals
- Then we model the obtained residuals, using some flexible method such as the gradient boosting

Linear model

We want to fit a straight line, between demand and time, for each combination of center and meal. This mean we should fit $N^o centers \cdot N^o meals$ $(77 \cdot 51 = 3927)$ linear models. But if we carefully craft some indicator variables we can specify all the simple linear models in to one single big linear model.

Linear model

$$Y_{ij} = \beta_{0ij} + \beta_{1ij} week + \mathcal{E}_{ij}$$
$$\forall i = 1, ..., 77; j = 1, ..., 51$$

Is equivalent to:

$$Y = \beta_0 + \beta_1 week + X_{ind}\beta_{level} + X_{ind}\beta_{slope} \cdot week + \mathcal{E}$$

Linear model

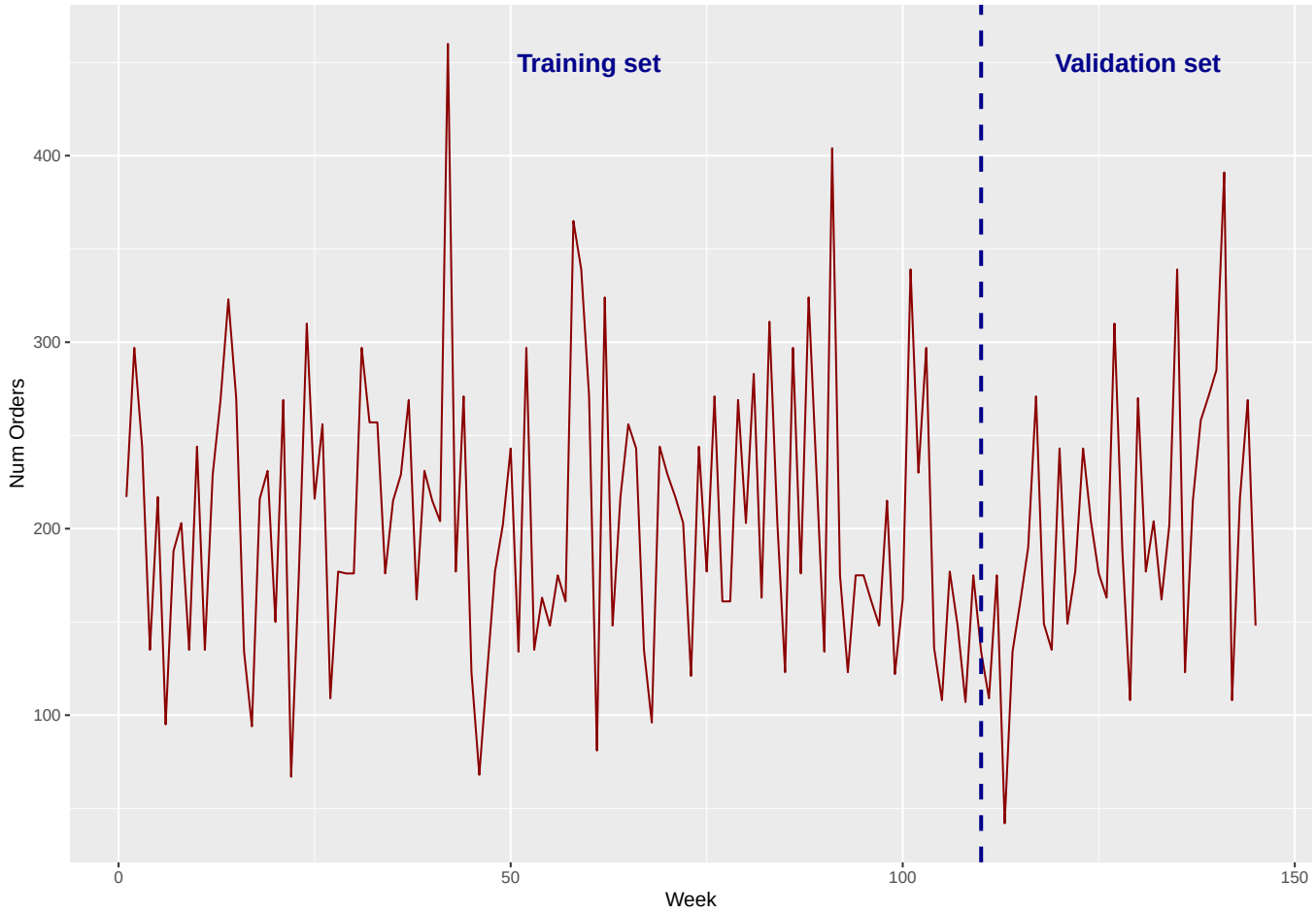
where X_{ind} is a vector with $51 \cdot 77 - 1 = 3926$ columns, and is obtained as the interaction between the dummy expansion of the categorical variables center_id and meal_id.

The model has $1 + 1 + 3926 + 3926 = 7854$ scalar parameters, that in the simple formulation there are 2 parameters for each model, so $2 \cdot 77 \cdot 51 = 7854$

Validation set and Test set

Dealing with time series data means that standard cross validation is not a viable option, since it breack the temporal dependency. We instead reserved a validation set, taking the last set of observations. The test set are the next 10 week, and the true number of orders stands on Kaggle.

Validation set and Test set

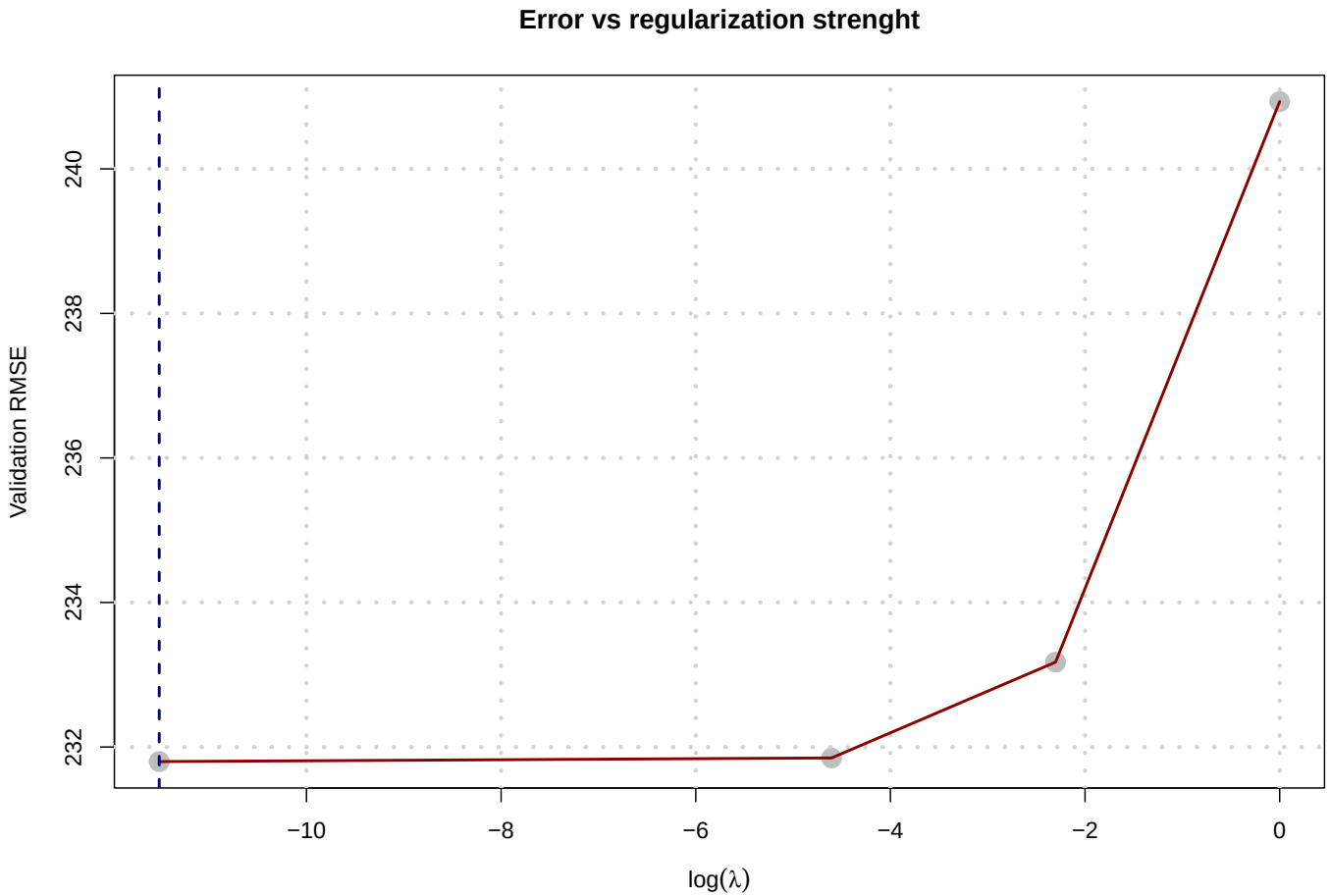


Regularization

We added elastic-net regularization in the estimation process:

$$\hat{\beta} = arg \min_{\beta \in \mathbb{R}^p} \left(\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right) \right)$$

No regularization is needed

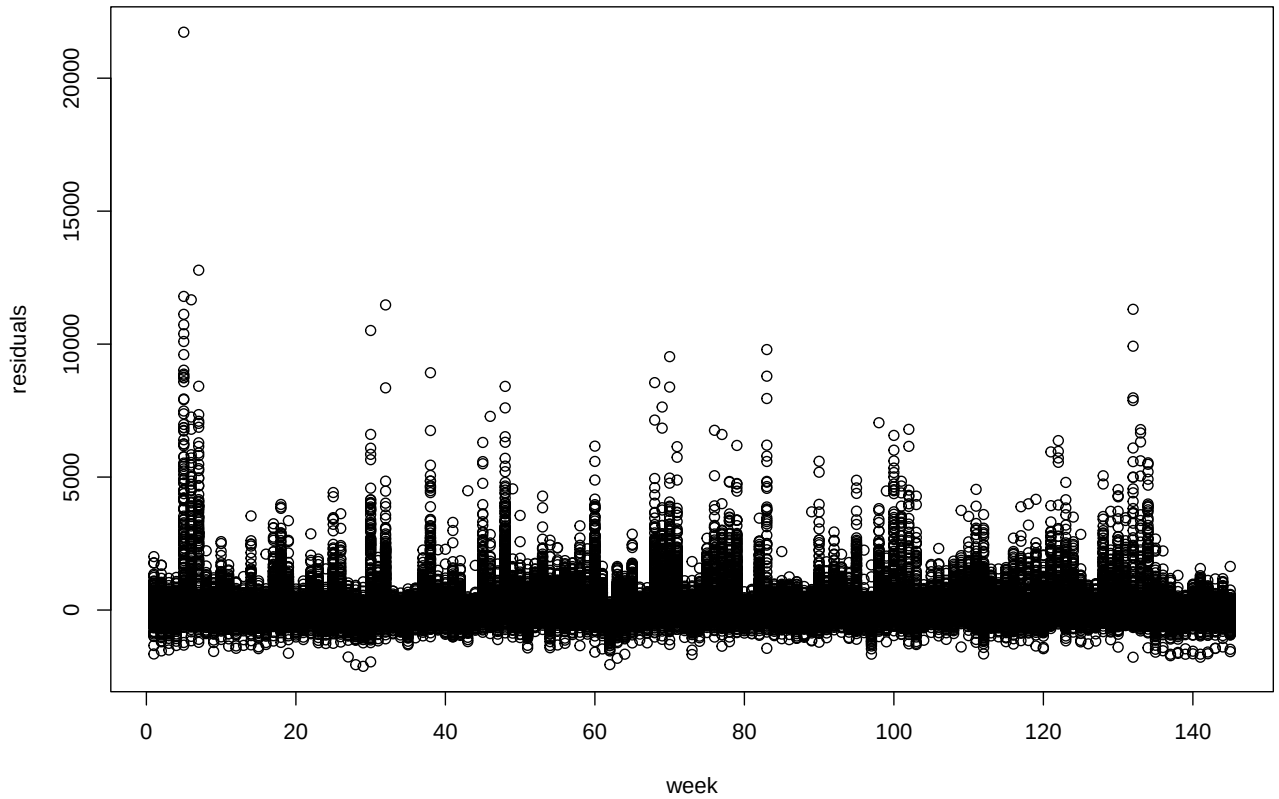


Results on validation set

Model	RMSE	MAE
Mean	350.9743	211.5107
LM	240.9311	106.4178
LM on ln(y)	367.1824	181.3026

Residuals linear regression

Let's compute and plot residuals both for the train and the validation set.



Gradient Boosting

Let us now try to further improve the accuracy in guessing the proper **number of orders** by using the prediction from the linear model as additional predictor.

For this purpose, **Gradient Boosting** is trained on the full training set, which includes the following 12 predictors:

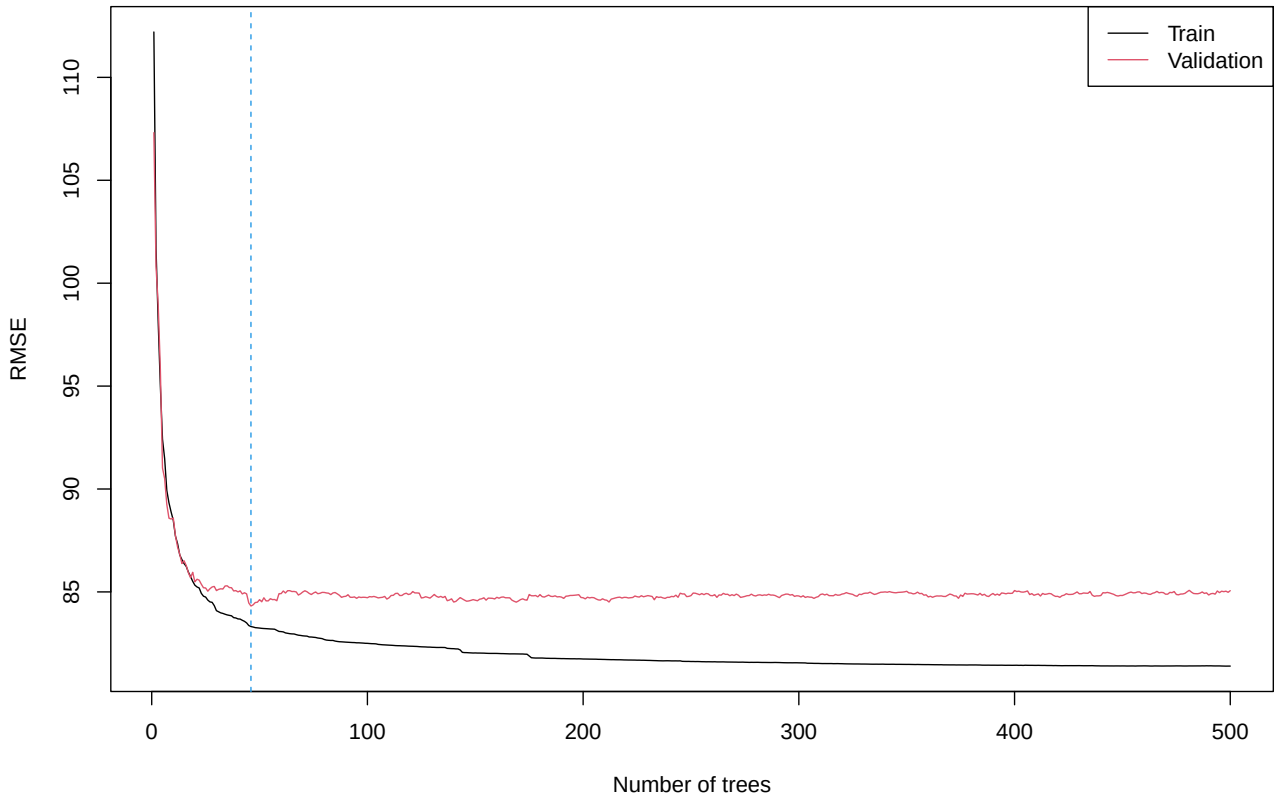
“meal_id”, “center_id”, “base_price”, “emailer_for_promotion”,

“homepage_featured”, “city_code,”region_code”, “center_type”,

“op_area”, “category”, “cuisine”, “predicted_lm”.

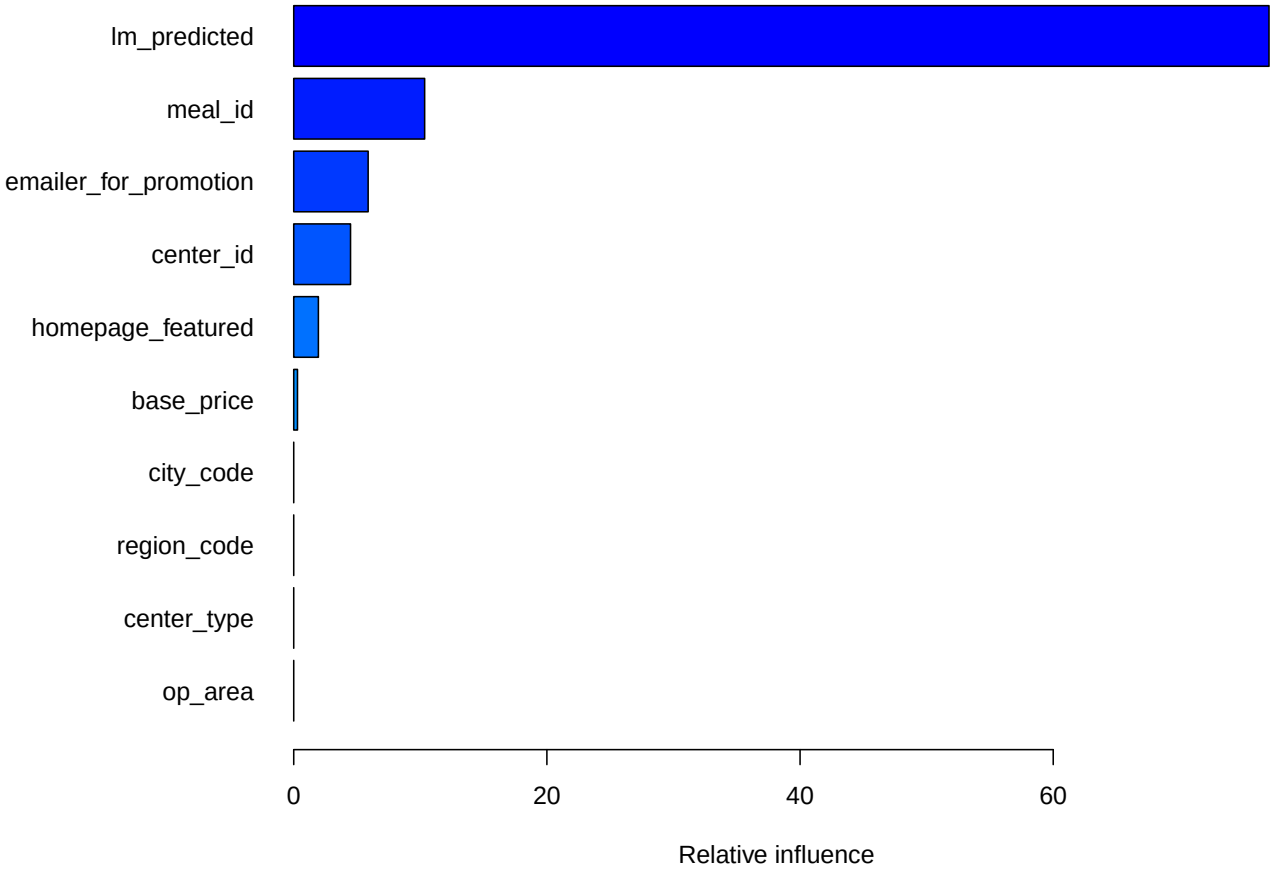
Training

The model is trained with **1000** trees, each of which has a depth of **2**.



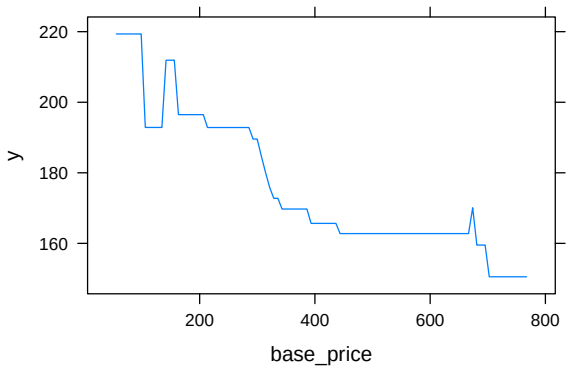
Relative influence plot

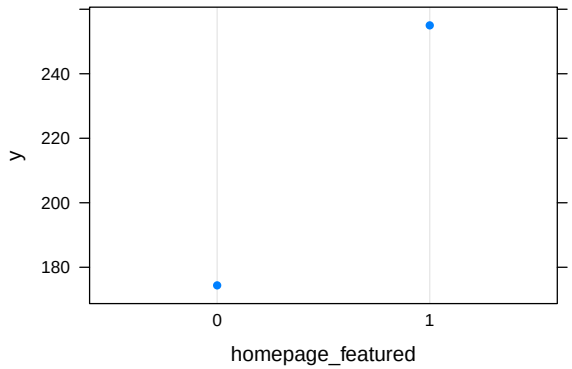
Let's take a look at which predictors are most important in discrimination.

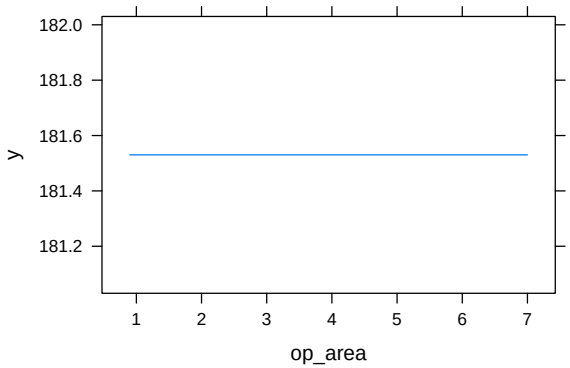


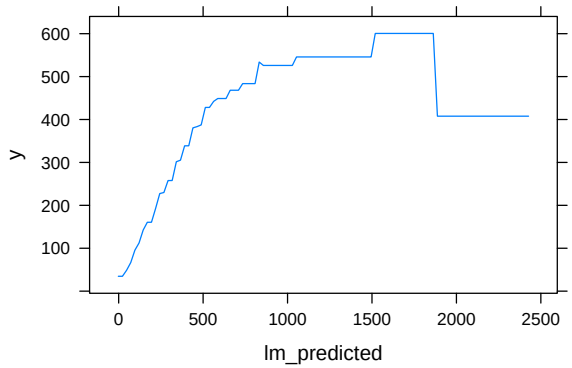
Marginal effects

Let's also plot the marginal effects that some predictors have on the response variable.









Results on validation set

Model	RMSE	MAE
Mean	350.97426	211.51070
LM	240.93105	106.41776
GB	84.31403	58.46229

Thanks for the attention

We hope that with this forecasting we are able to avoid a lot of food waste and to reduce transportation carbon emissions. Moreover, it will improve the company profit.

Pierpaolo D'Odorico, Massimiliano Conte and Eddie Rossi.