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, op_area : 30

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:

meal_id : 51, category : 14, cuisine : 4

Meal info data*

id	week	center_id	meal_id	base_price	num_orders
1379560	1	55	1885	152.29	177
1466964	1	55	1993	135.83	270
1346989	1	55	2539	135.86	189
1338232	1	55	2139	437.53	54
1448490	1	55	2631	242.50	40
1270037	1	55	1248	252.23	28

*binomial variables and “checkout_price” are removed for a better dataset view.

Meal 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

Meal data features

Unique values in dataset:

id : 456548, week : 145, center_id : 77, meal_id : 51, checkout_price : 1992, base_price : 1907, emailer_for_promotion : 2, homepage_featured : 2, num_orders : 1250

Create a unique dataset

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

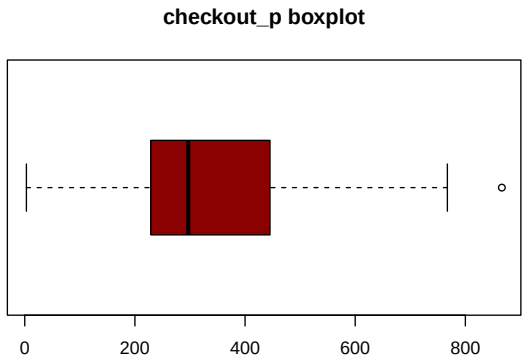
There are **0 NA's** in the complete dataset.

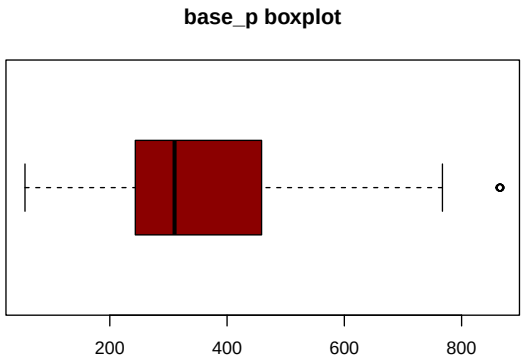
We will perform some exploratory data analysis:

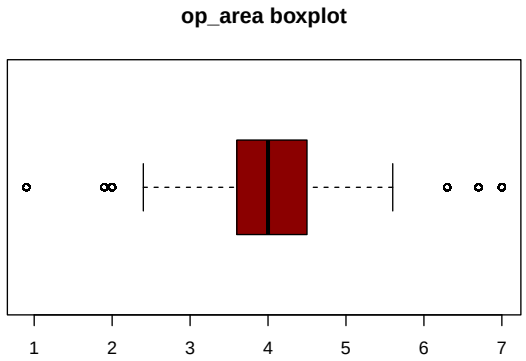
- **Univariate Analysis**: Looking at single variables behaviour
- **Multivariate Analysis**: Correlation between variables

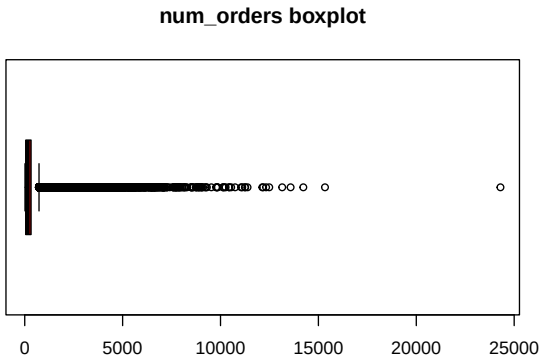
Univariate Analysis on numerical variables

We plot boxplots of **numerical variables**:



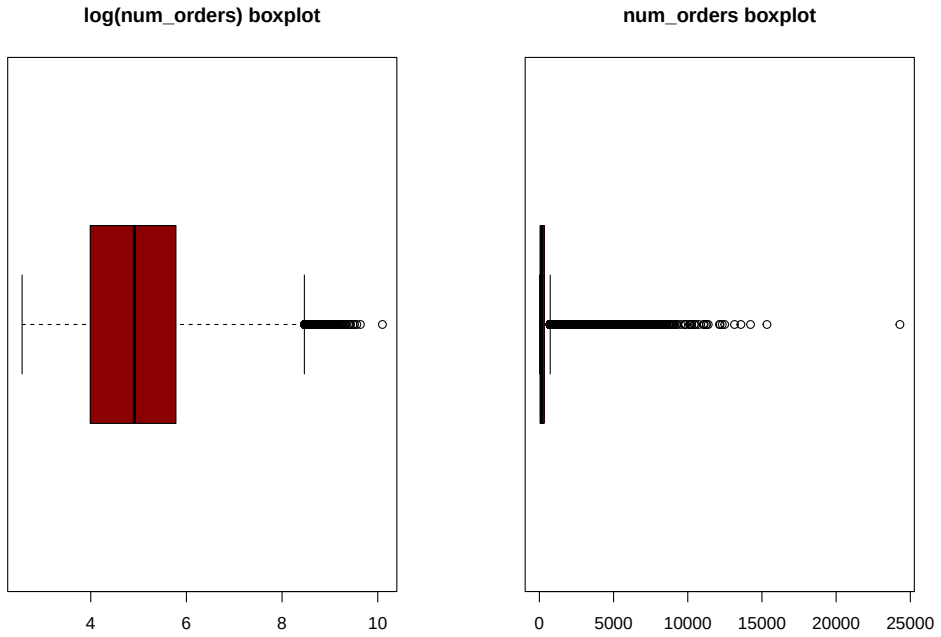






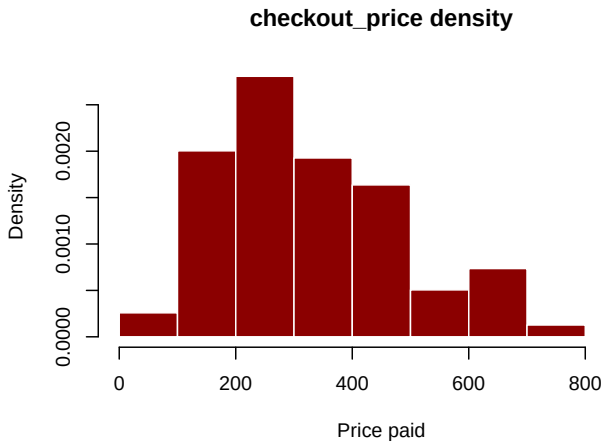
Univariate Analysis on numerical variables

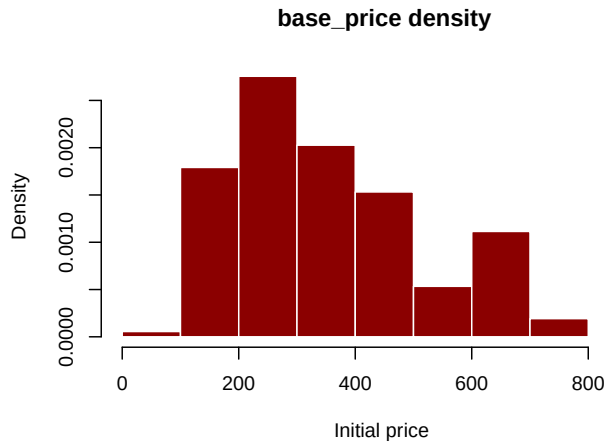
num_orders boxplot vs log transformation for a better view:

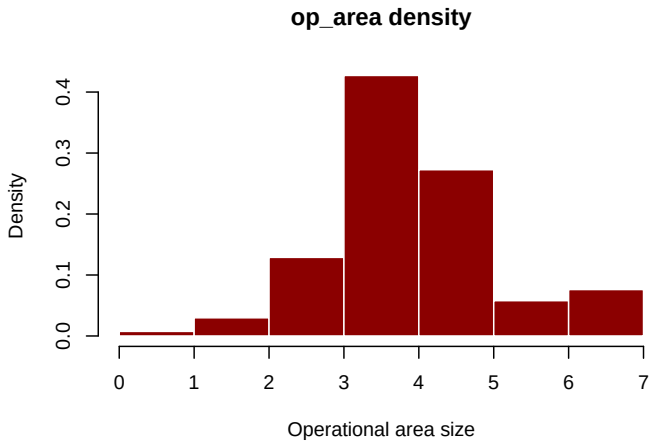


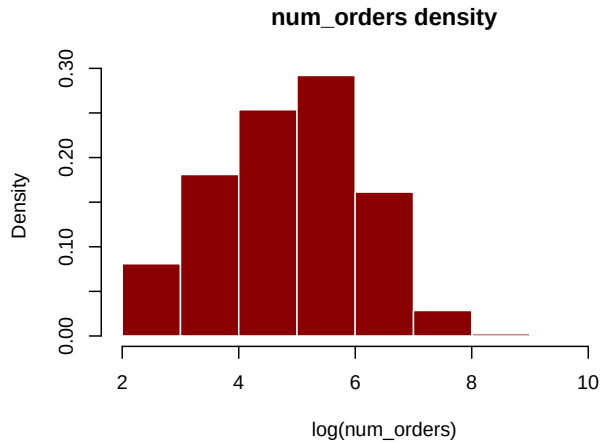
High **num_orders** are related to some specific series from a specific center and meal with high demand. For this reason we don't consider them outliers.

Univariate Analysis on numerical variables

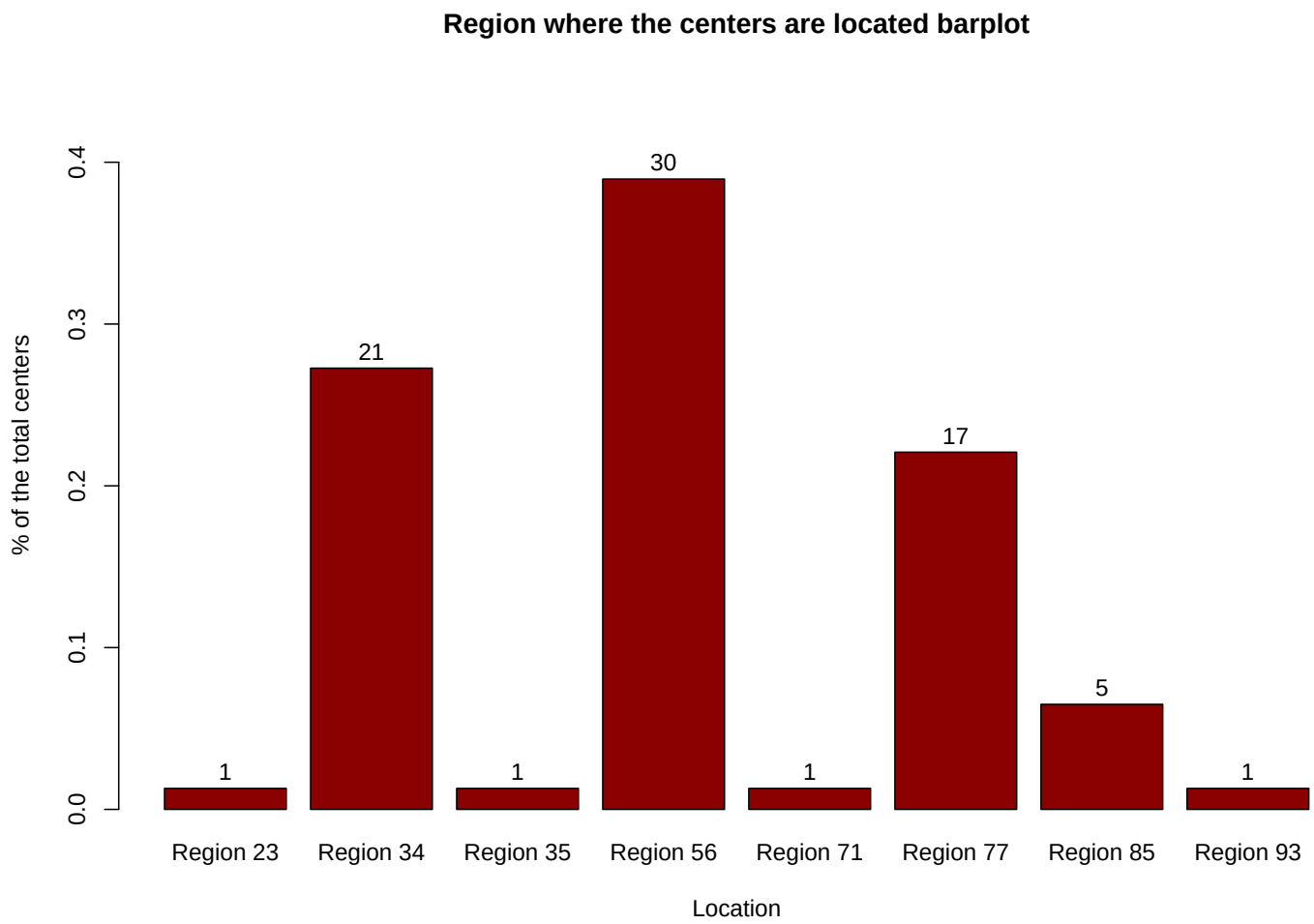






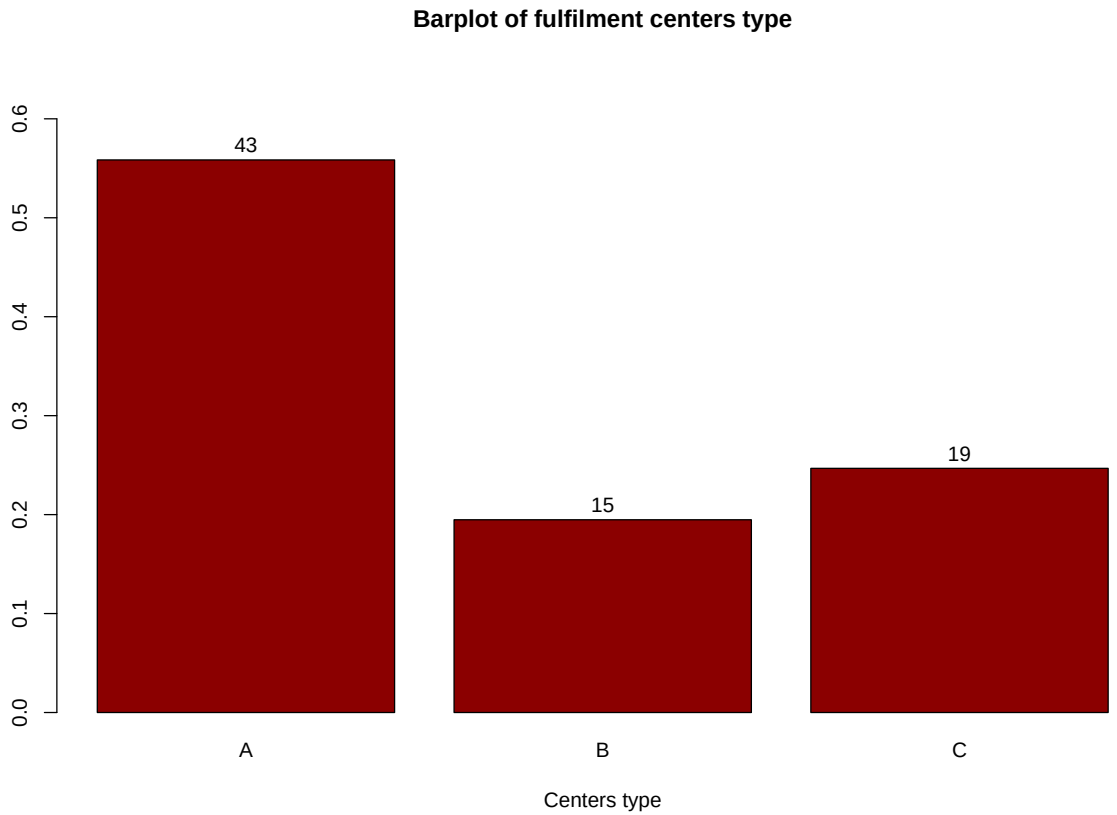


Univariate Analysis on categorical variables

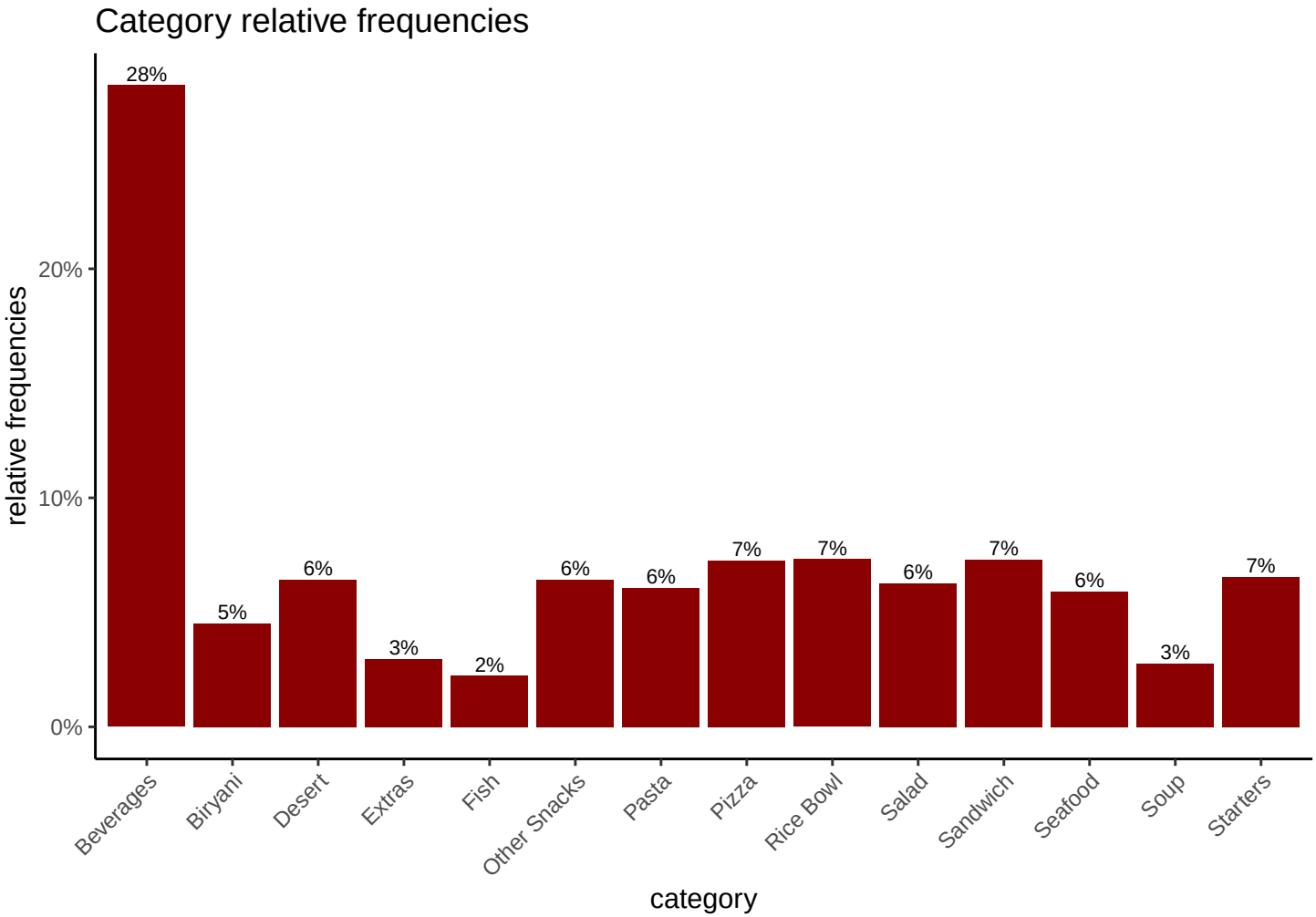


Univariate Analysis on categorical variables

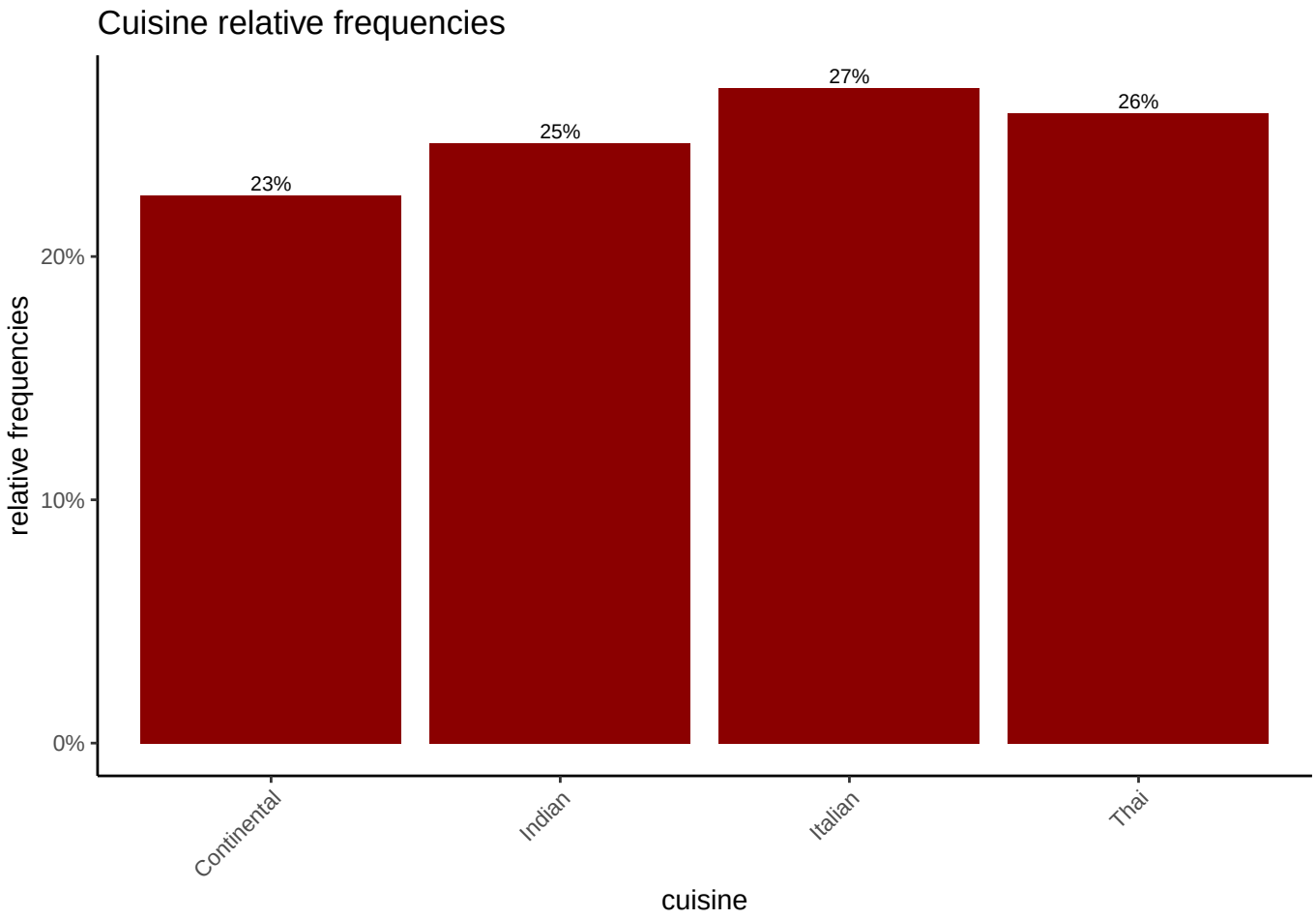
center_type: There are 3 centers types: A, B and C



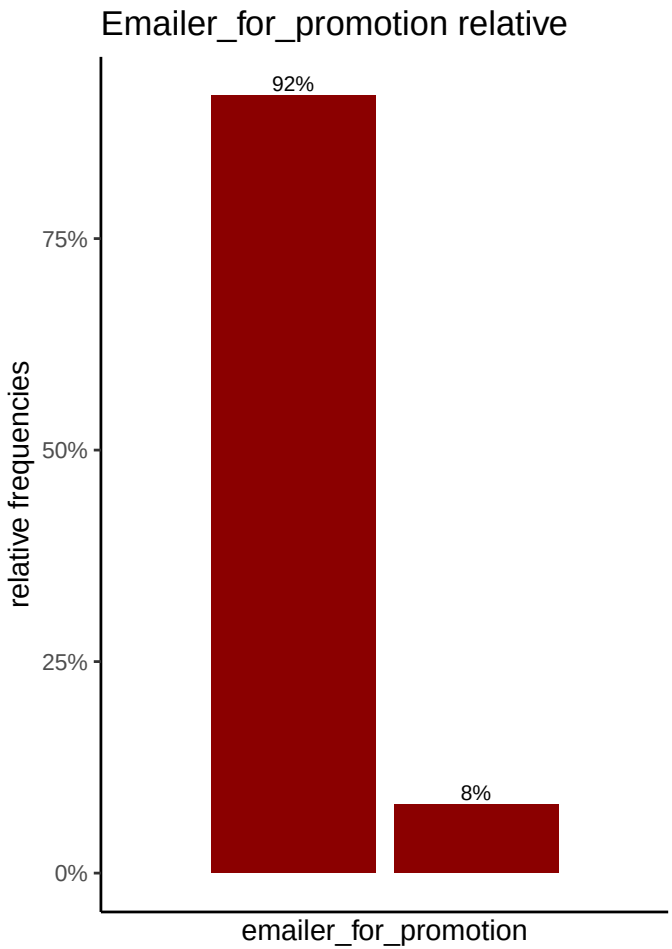
Univariate Analysis on categorical variables

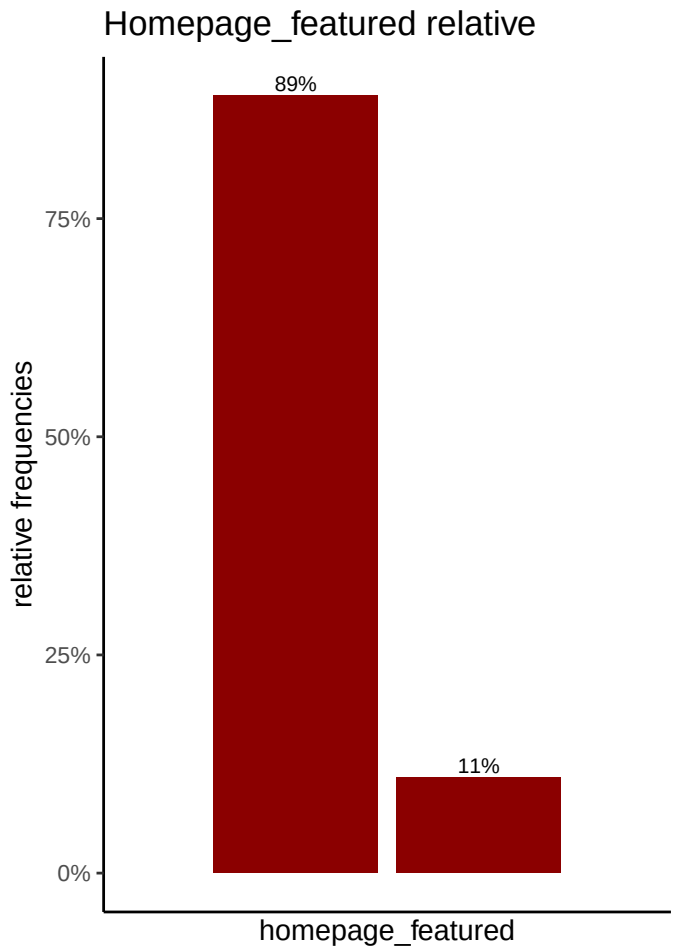


Univariate Analysis on categorical variables

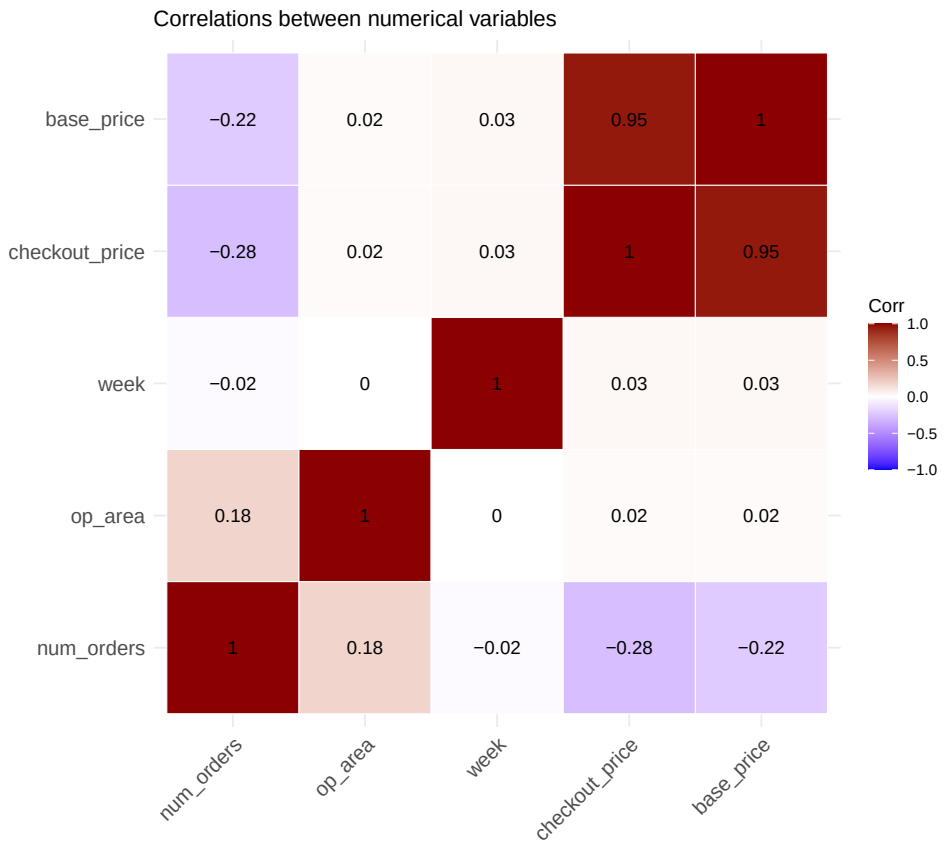


Univariate Analysis on categorical variables



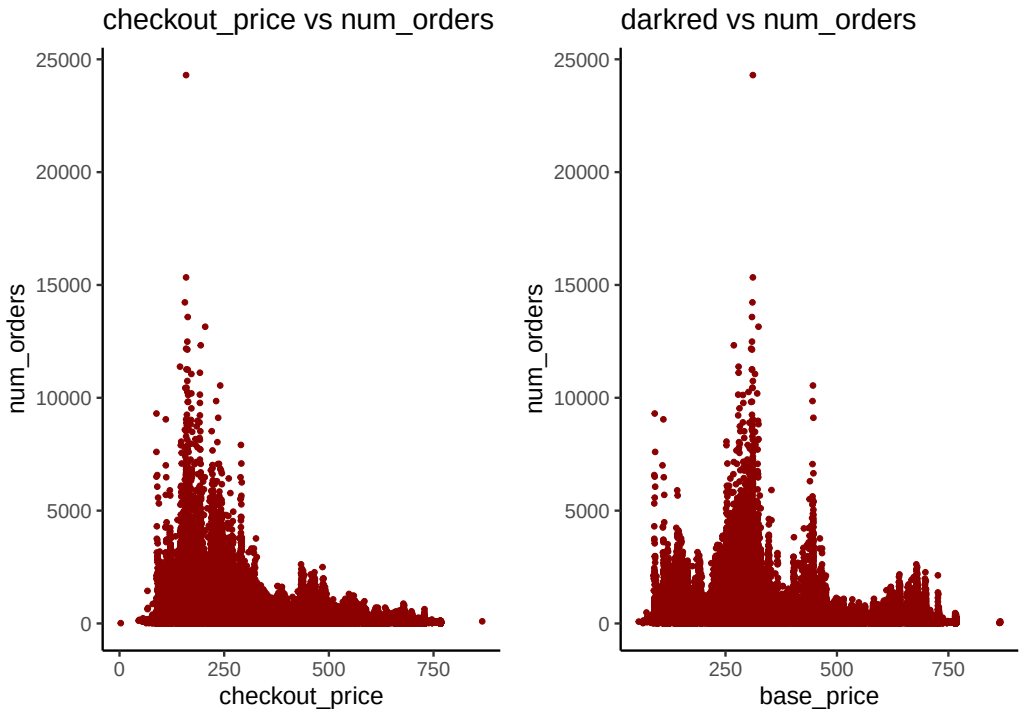


Multivariate Analysis, correlation plot



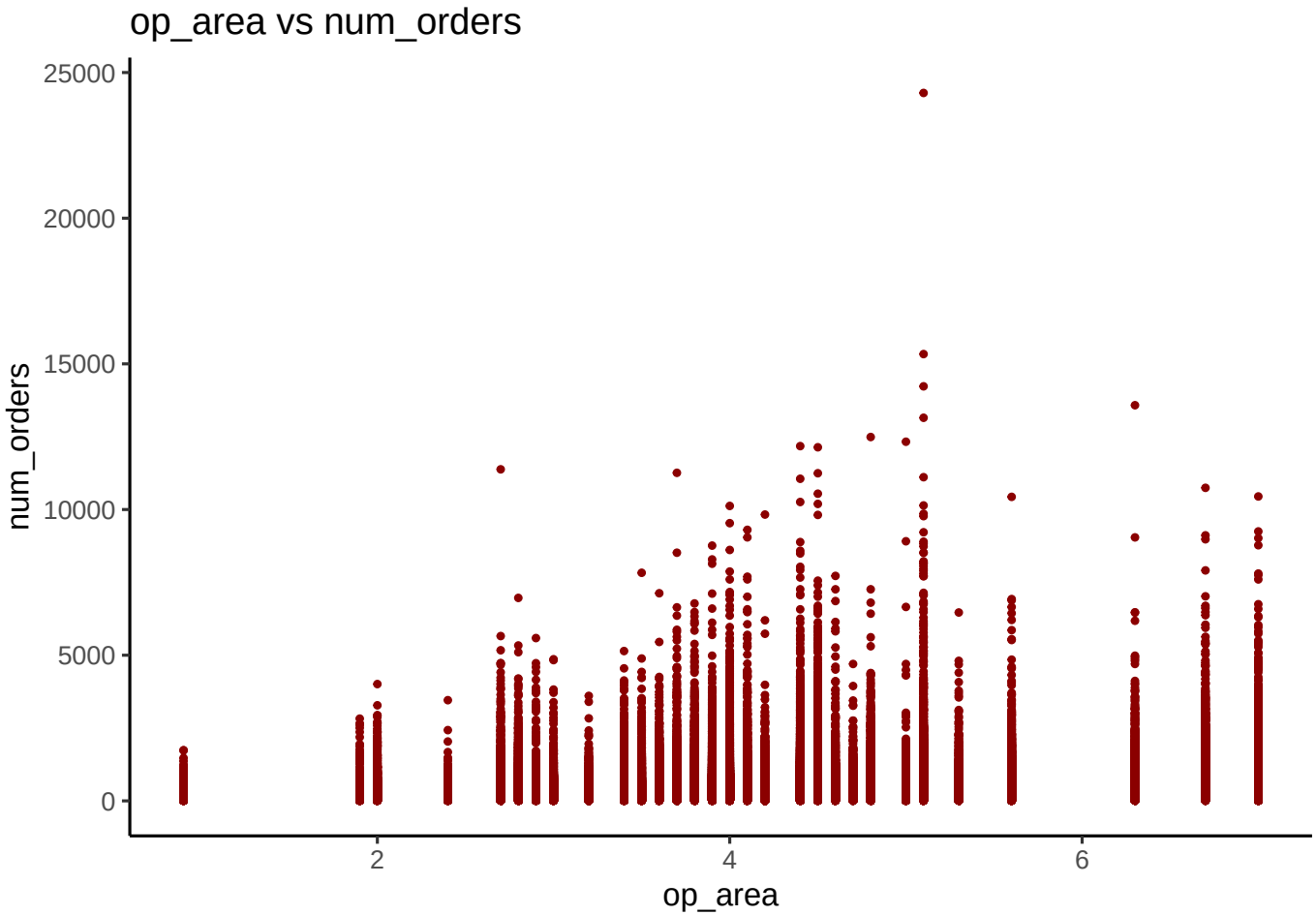
num_orders is not highly correlated with other numerical variables.

Multivariate Analysis, numerical variables



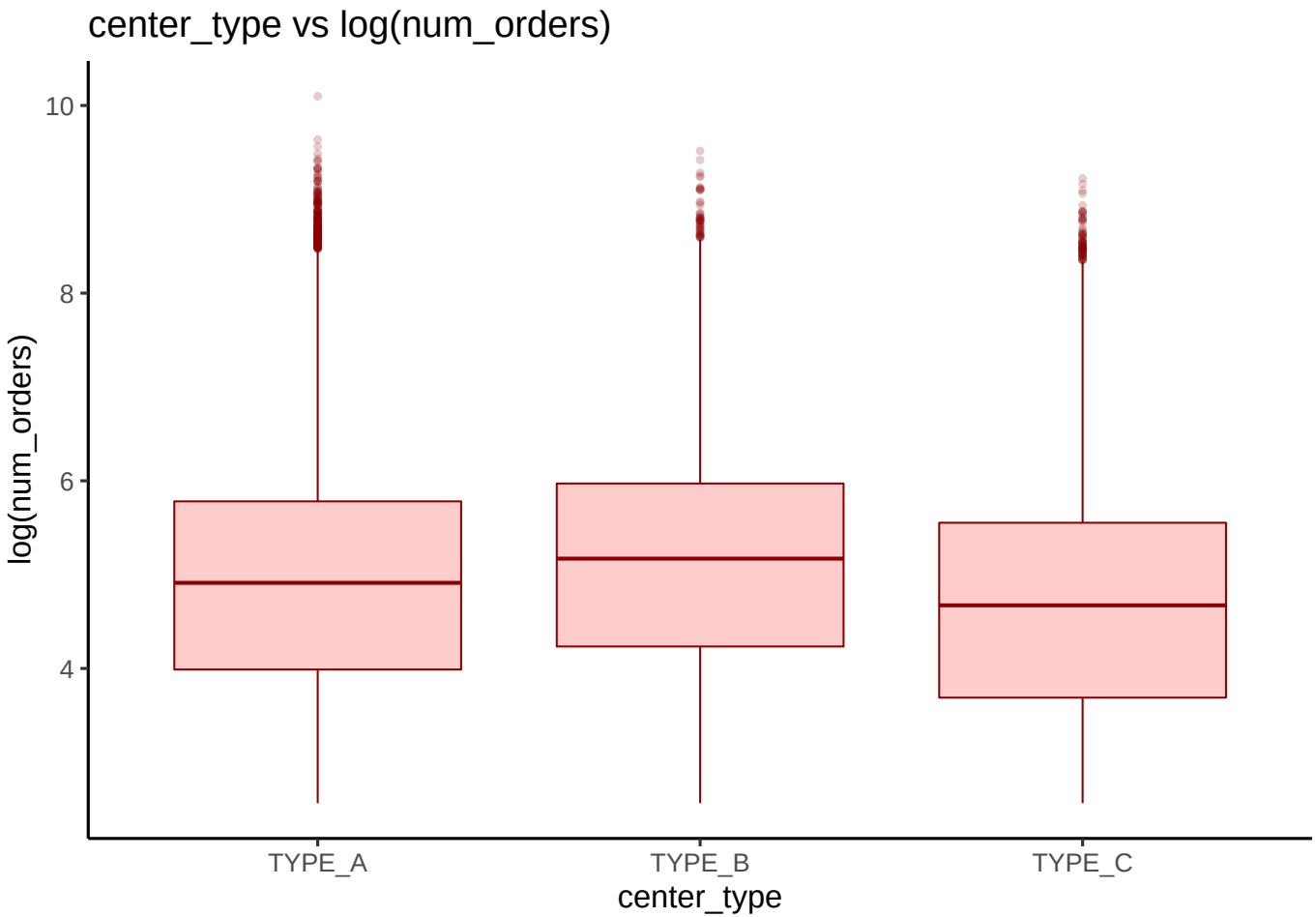
We decided to remove checkout_price due to the nature of the variable, in a real case scenario we can't have a checkout price because checkout means that the order is confirmed.

Multivariate Analysis, numerical variables

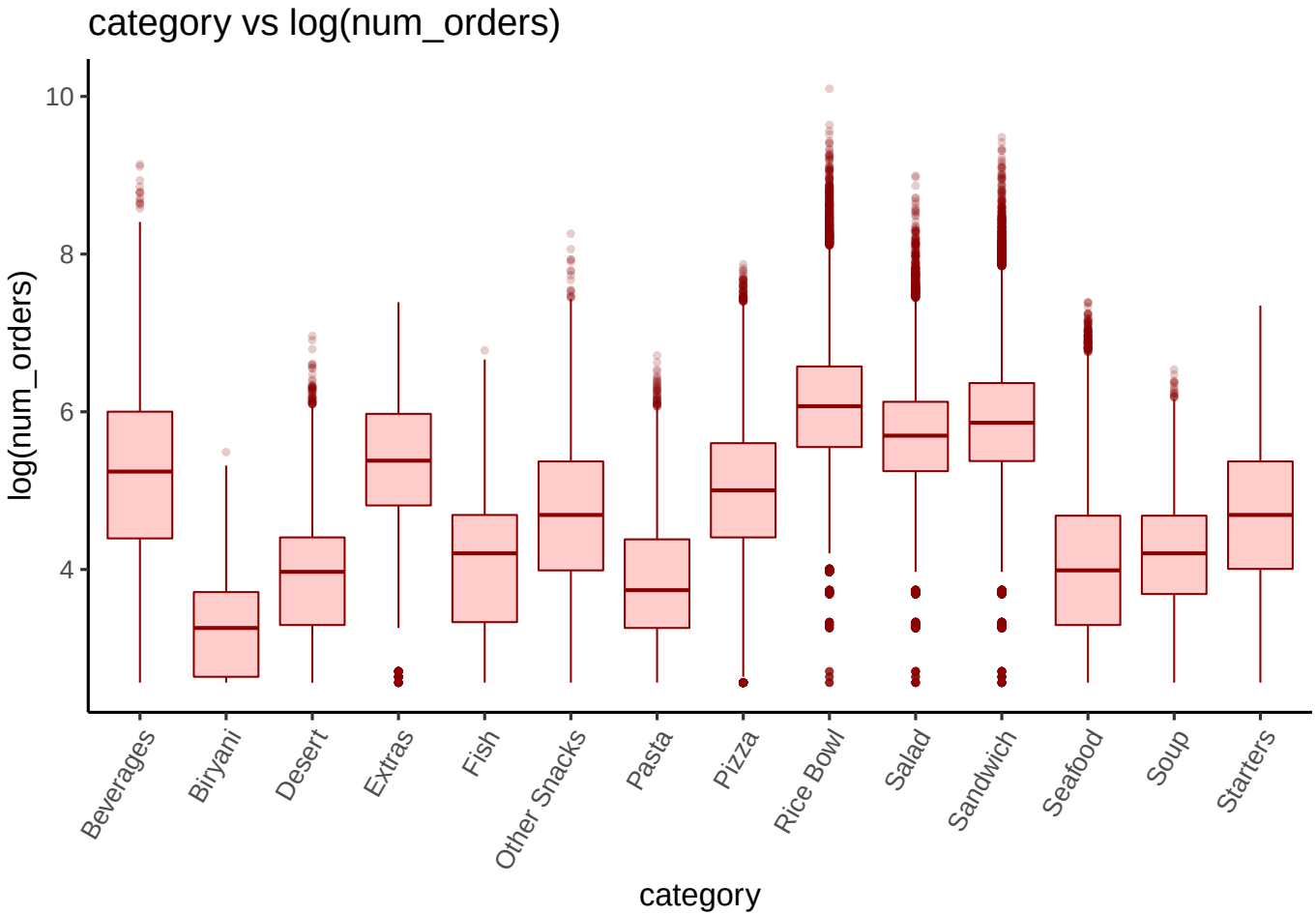


num_orders seems to increase for bigger op_area centers

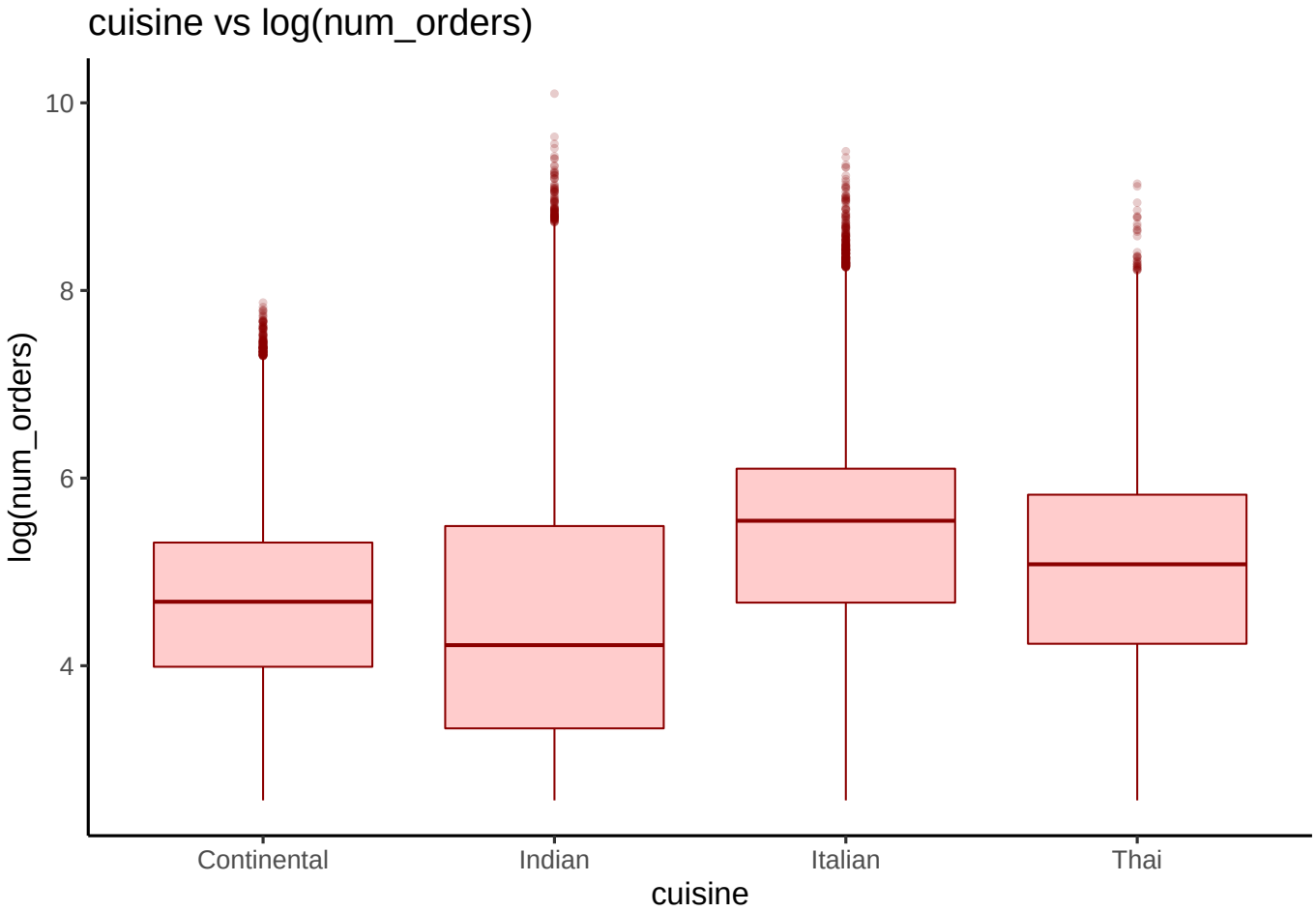
Multivariate Analysis, categorical variables



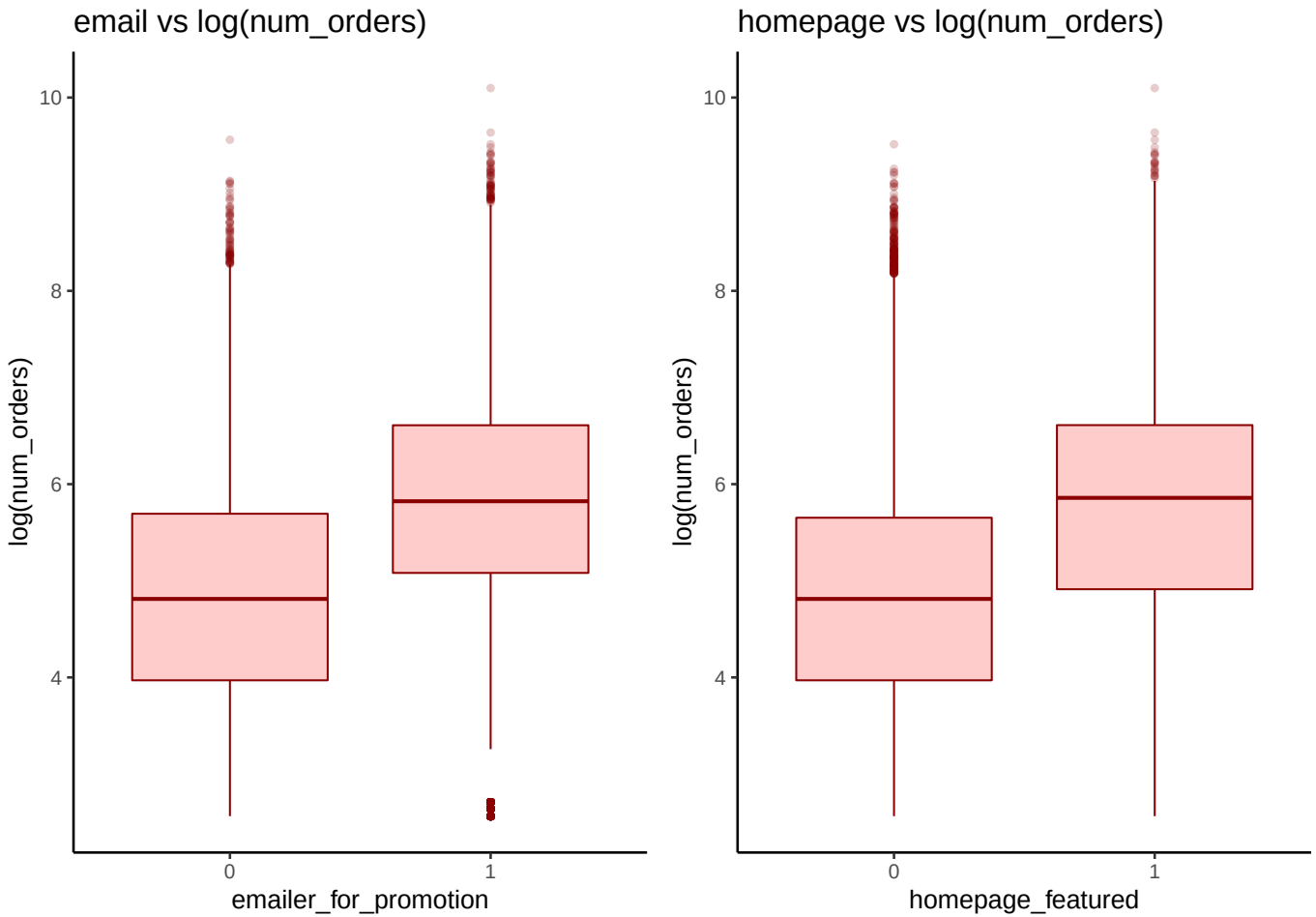
Multivariate Analysis, categorical variables



Multivariate Analysis, categorical variables



Multivariate Analysis, categorical variables



Promotions ad email seem to increase the orders

Modelling

Since we want to organize the goods for each specific fulflment 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

$$Y_{ij} = \beta_{0ij} + \beta_{1ij} week$$

$$\forall i = 1, ..., 77; j = 1, ..., 51$$

Is equivalent to:

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

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 $2 \cdot 77 \cdot 51 = 7854$

Results