### Dataset description

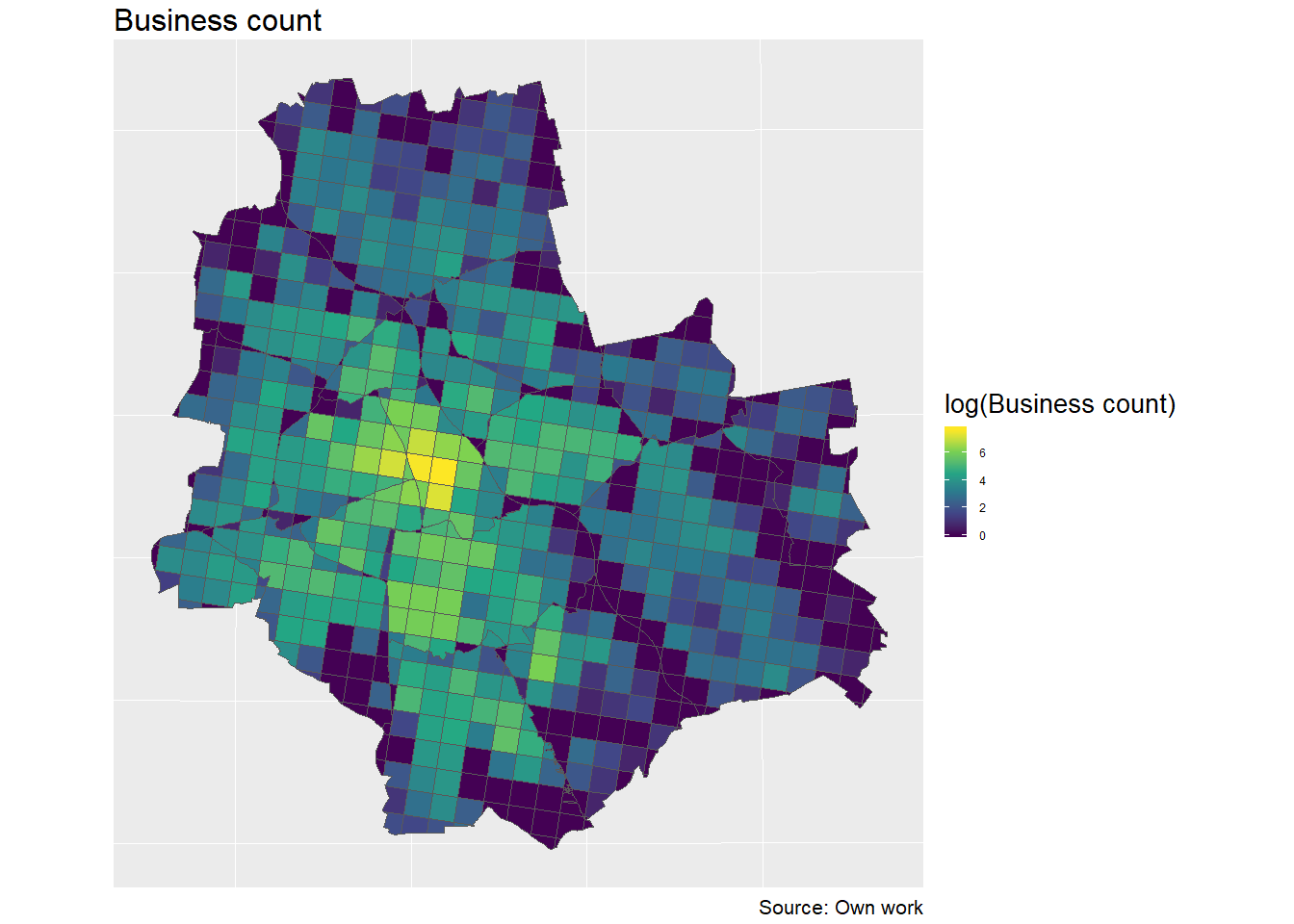
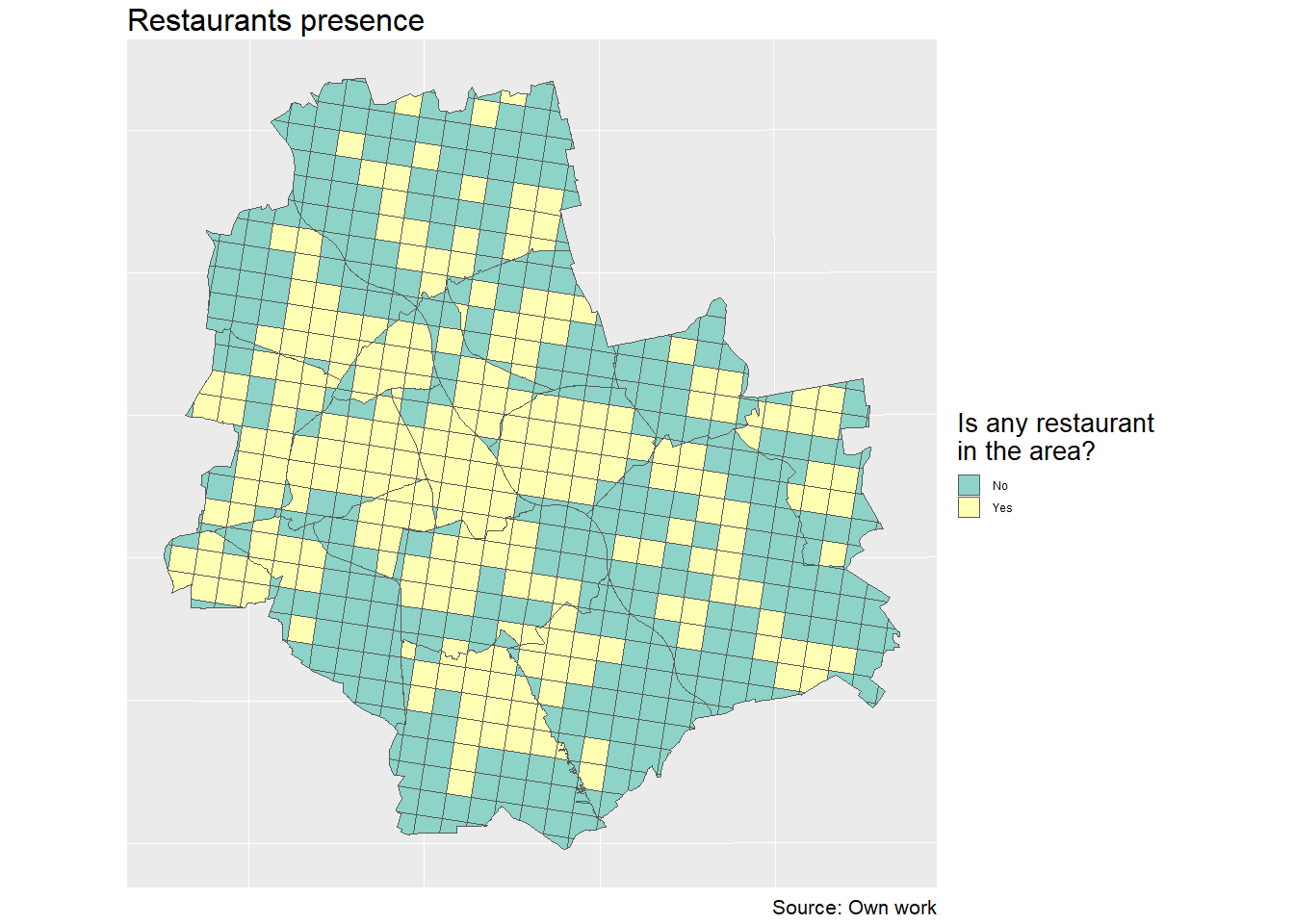
In my study I have restricted the analysis to the Warsaw metropolis. The variables included in the dataset are:

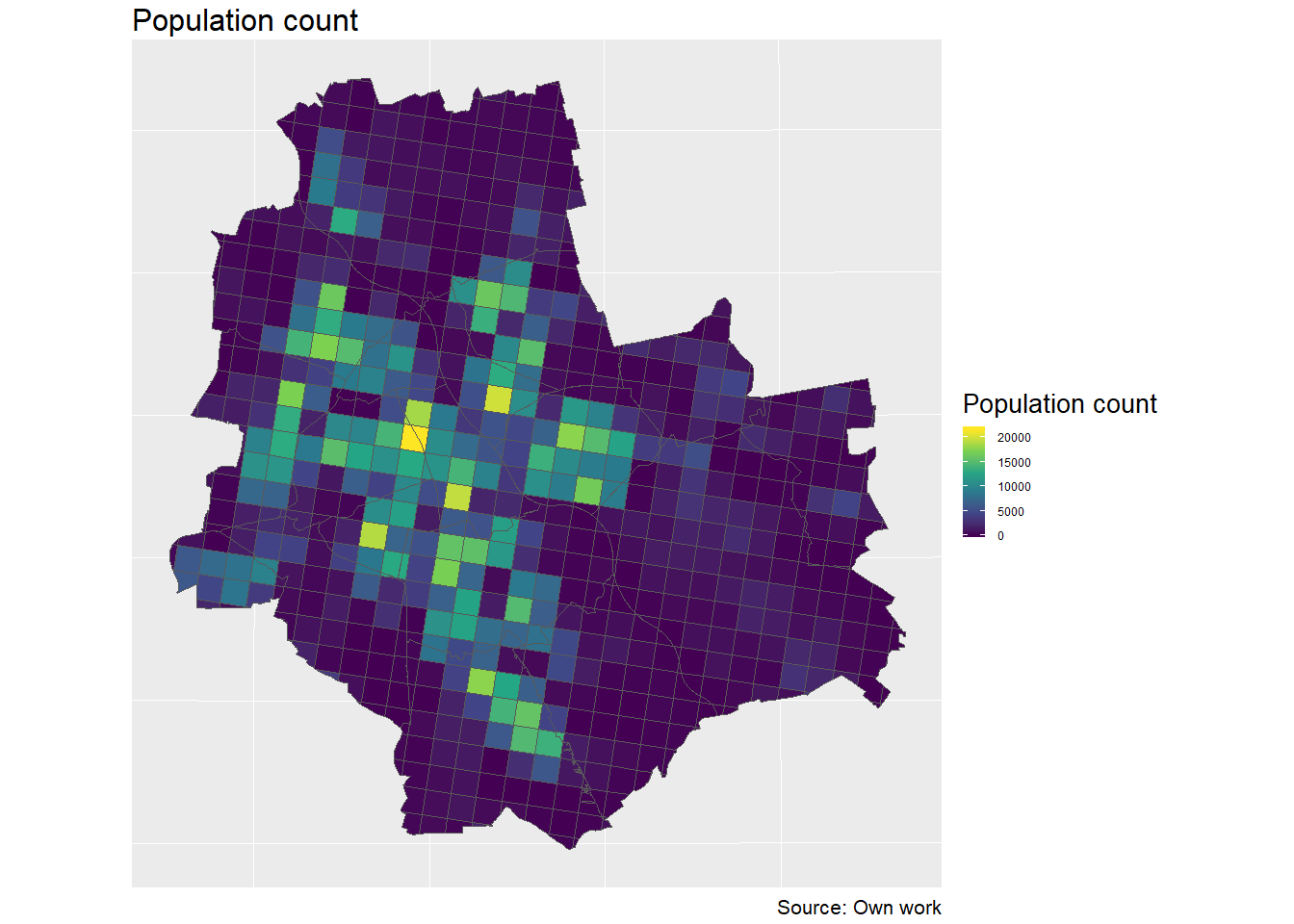
* Restaurants locations
* Businesses locations
* Population density
* Bus stops locations
* Roads locations.

These features come from various sources. Restaurants’ locations were obtained from Zomato website. There were 2341 observations total, but due to incorrect addresses, 72 restaurants were excluded. Population density comes from 2011 GUS National Census (<https://geo.stat.gov.pl/nsp-2011>). The data about businesses was gathered from (…). Location of bus stops and roads was obtained using Open Street Map service.

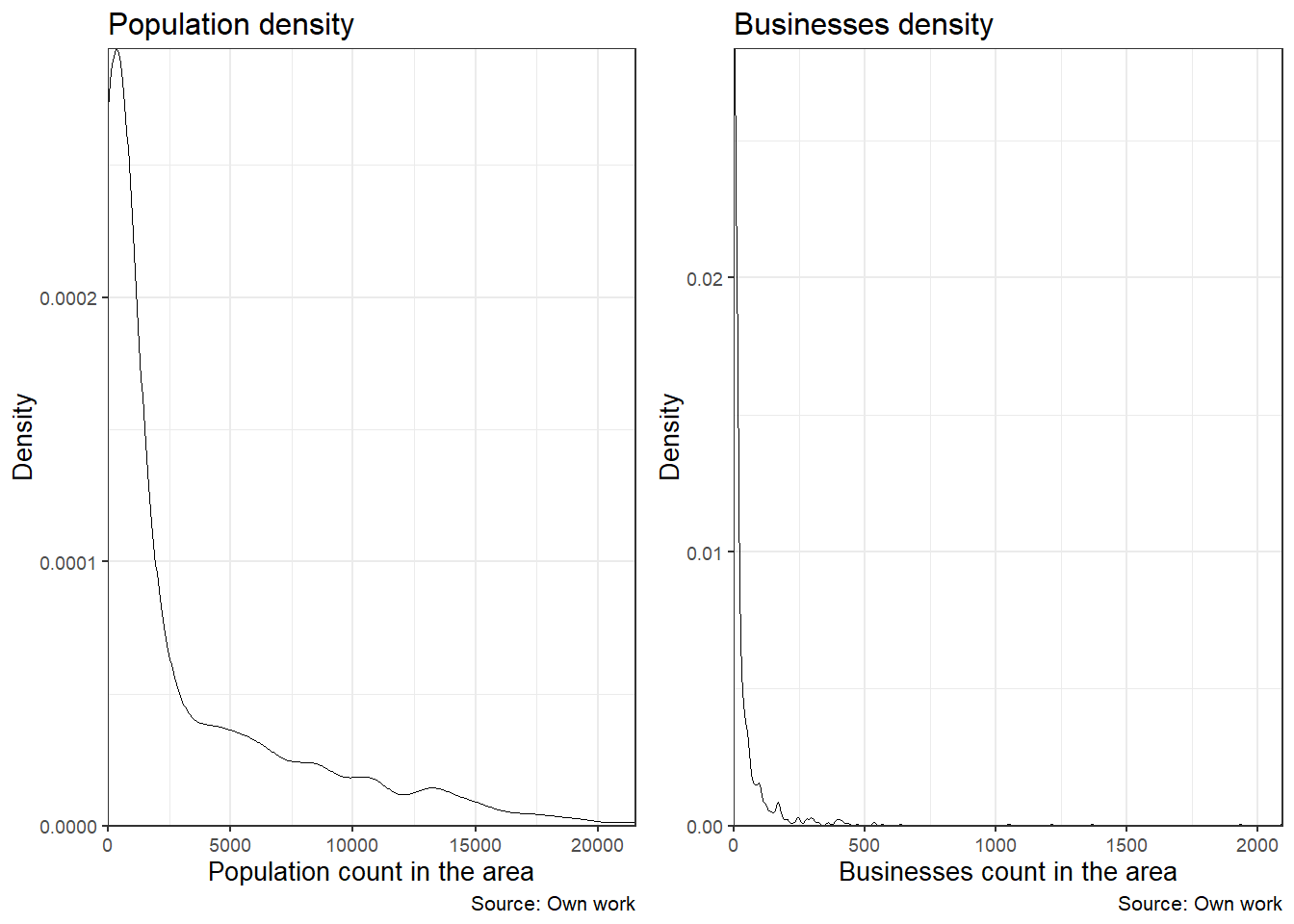
Although restaurants, businesses and infrastructural features (bus stops and roads) are points data, population density is in a form of an 1km x 1km aggregated grid. Thus, to asses population influence on the presence of restaurants, it was necessary to convert all variables to the same format. To do this, all variables were binned to a grid in the same resolution as population density data.

The map of restaurants locations shows that there exists high centrality. Also, in regions far from city center it is visible that restaurants are located in proximity to the largest streets, some of which are exit roads. Population and business presence are also highly concentrated in the city center.

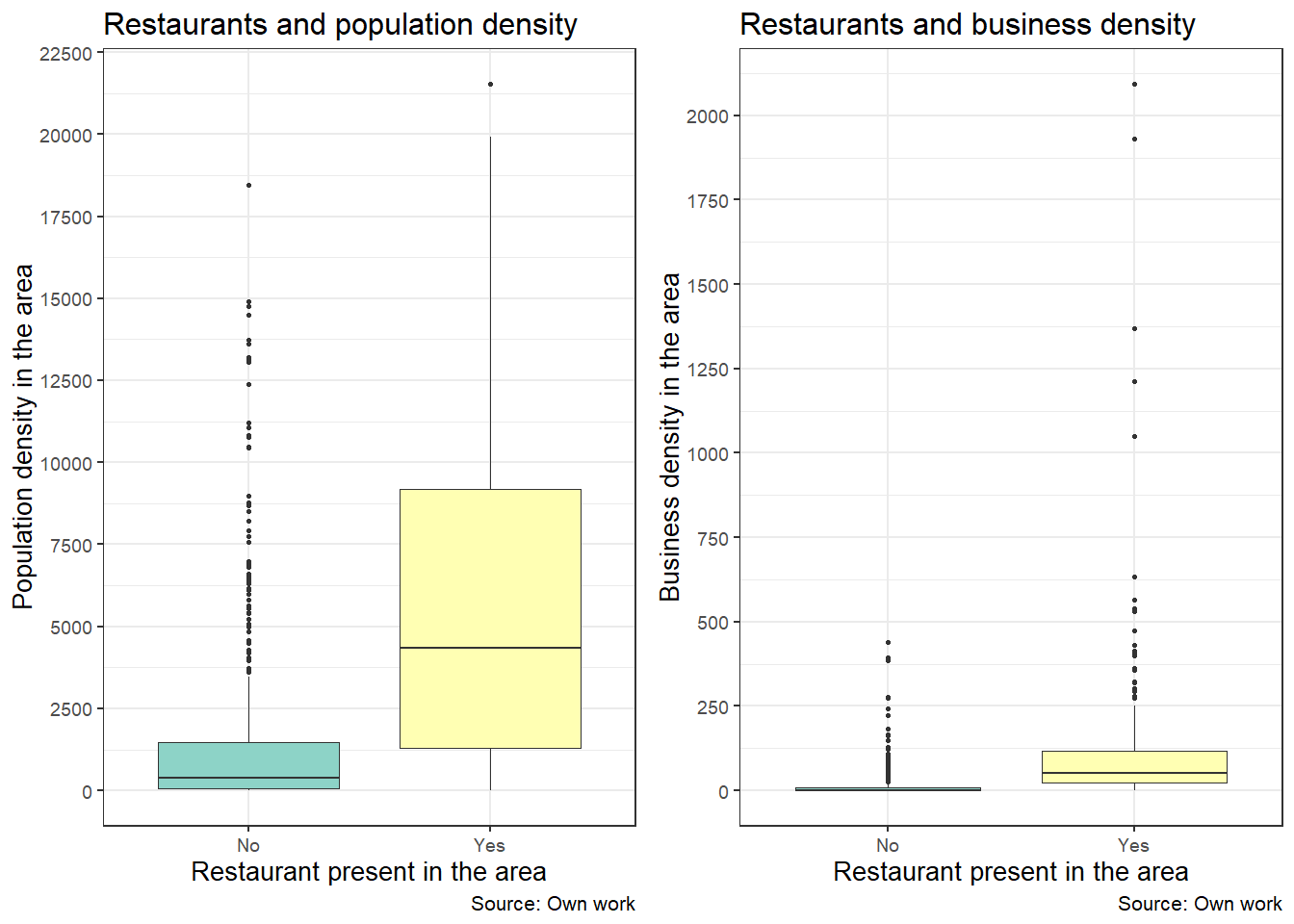




After binning the points data to a grid it can be seen that both restaurants and businesses locations distributions are highly skewed. Typical power law distribution is observed, with majority of values close to 0 and few observations with extreme values. The population density data is also highly right-skewed, but to a way lower extent than the other two variables.



As shown on the boxplots below, the subsamples containing and not containing any restaurants are significantly different in terms of business and population locations. Average business count in a grid cell in which the restaurants was present was 116.62, while in regions without restaurants was only 14.53. Similarly, average population density in restaurants’ regions was 5784.74, compared to 1613.98 in regions without restaurants presence.



The join-count statistic was performed on restaurants presence data. With p-value< 0.0001, there is evidence that spatial autocorrelation in target variable exists. This means that estimates using non-spatial modeling will be biased, and there is necessity to take spatial dimension into account.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | min | Q1 | median | mean | Q3 | max |
| Population count | 0,00 | 142,75 | 738,00 | 2976,70 | 4295,00 | 21531,00 |
| Restaurants count | 0,00 | 0,00 | 0,00 | 2,81 | 1,00 | 228,00 |
| Business count | 0,00 | 0,00 | 7,00 | 47,89 | 39,25 | 2093,00 |
| Bus stops count | 0,00 | 0,00 | 3,00 | 5,66 | 8,00 | 54,00 |
| Total roads length | 0,00 | 711,47 | 3594,25 | 4887,42 | 8151,69 | 20111,20 |

Table 1- basic descriptive statistics of used variables