Chapter 1

Date Cleaning and Preparation

1.1 New York Inspection Data

The main dataset for our project is retrieved from the official site of the State of New York. It contains a total of 28'300 A to C ratings from of food store inspections. When we downloaded the data on the XXX it has been the version of June 26, 2019 with observations from March 2018 to March 2019.

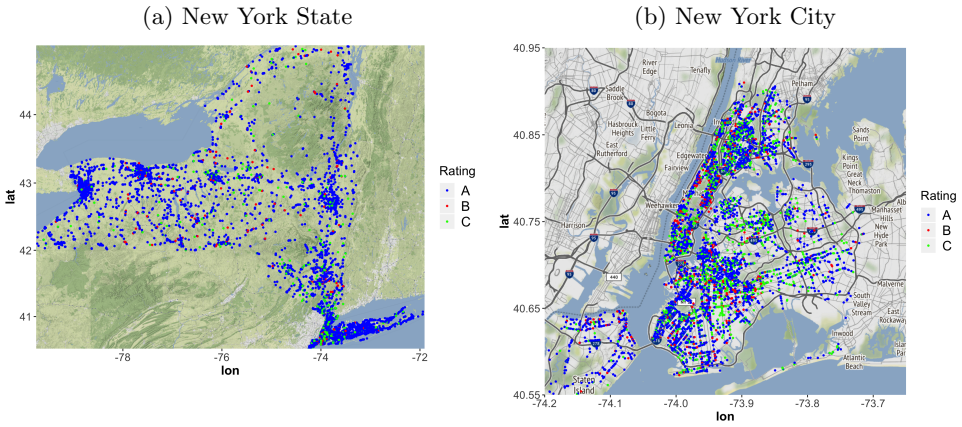
–¿ here Kaspar remove duplicate values
We will predict the food store inspection ratings from this dataset with different covariates whose derivation is outlined in the following sections.

1.2 Chain Information

1.3 Spatial Data

Some of the predictors rely on geolocation data. The cleaned inspection dataset contain spatial information in form of latitude and longitude data for most of the observations. The latitude and longitude information, however, is embedded in a larger address string. We create a function to extract the location data in two new columns. A check for `NA` values reveals that 748 values missing location information. The exact address is available for all shops though. Therefore, we combine, street, city and ZIP code to a single

Figure 1.1: Geographic Distribution of the inspected Shops



string that can be used with the Google Maps API[1] to obtain the missing latitude and longitude details. A Map of the completed data is illustrated in Figure 1.1 on state and city level.

With complete spatial information, we now compute the two variables "shops density in 1km radius" as well as "rating of the closest neighbor". To get the distances of coordinates in meters, we apply the Haversine Formula[2]

$$a = \sin^2(\frac{\Delta\alpha}{2}) + \cos(\alpha_1)\cos(\alpha_2)\sin^2(\frac{\Delta\lambda}{2}) \tag{1.1}$$

$$d = R\left(2\,\text{atan2}(\sqrt{a}, \sqrt{1-a})\right) \tag{1.2}$$

where α_i are latitudes, λ_i are longitudes, R is the earth's radius ($6371 \times 10^3 m$).

[1] In essence, the Google Maps API is a free service offered by Google. It requires only a one-time registration with a valid email address. Afterwards, it generates an API Key that must be included in the R-script with the command `register_google(key = "API KEY")`.

[2] The Haversine Formula has its roots in spherical trigonometry and calculates the geodesic distance – the shortest path between two points on a curved surface of a sphere, like the Earth. It needs to be mentioned that Haversine Formula does not take into account changes in altitude. However, it provides sufficient accuracy for the scope of this project.

1.4 Subway Data

In a next step, we amend our New York City inspection data with location information of subway stations. We use again an official dataset provided by the State of New York. The dataset contains the location of every subway entrance and its corresponding station. We are only interested in the station. Therefore, we remove all duplicates of stations with multiple entrances and then use again the Haversine Formula to find the distance of the closest station to every shop.

Chapter 2

Data Analysis

2.1 The Imbalance Issue

Closer scrutiny of Figure 1.1 sheds light on the imbalanced distribution of the original data. In fact, the rating A contributes to XX% on state level while B and C are only represented to XX% and XX% respectively.

Chapter 3

Conclusion

–¿ Wrong predictors