# Cleaning

We used a combination of R and python to clean. We will first describe the cleaning process used in R for the datasets red light cameras, bike shops and bicycle parking spots. We used the tidyverse, and readr libraries in R to clean the data, and saved all ensuing datasets to CSV. Note we used the write\_csv function from the readr library, as it does not write the row names to the csv file. For the bike shop dataset, we first selected the relevant attributes and rename them to their corresponding names in our schema. We converted the number to integers instead of character, and the names all lower case, this ensures that if there were some naming choices by the different capitalization choices by the different datasets, we would still be able to join on these attributes. Some of the street numbers had a non-numeric character in it, and we removed these observations to ensure that the data conforms to our design. For the red-light camera data, we did a similar thing regarding renaming the attributes and forcing certain types. However, for this dataset we kept the id, which was in the original dataset to remove duplicates. We removed duplicates from this dataset through id, and then selected only the first street and second street. For the remaining datasets that relate to bike parking we treated them similarly regarding renaming and type casting. For these datasets we treated bike type of parking as a factor and forced “capacity” to be an integer. For the datasets that were not high-capacity parking, we did not have capacity information in those datasets and hence we treated each observation as a default capacity of one. Again, similar to the bike shop dataset we removed all observations that had a non-numeric character in their street number. Afterwards we joined the three parking datasets, by using row bind, and removed any duplicates. For some of the datasets the parking spot type had null values, so when saving we encoded NA as “”, to ensure that it is interpreted as null by psql when using \copy.