First of all, it’s difficult to handle that big amount of data. There are nearly over 200.000 different building permits and 43 columns. The meaning of the columns was not always clear. Even after looking at the dataset dictionary, you didn’t really know what information the column contains. For example, what is meant with „Exisiting Units“? Therefore, exploring those columns is quite difficult (you sometimes don’t get what the plot shows you). Deciding which columns to drop because there are not necessary for the exploration is also a challenge if you don’t exactly know the meaning of all variables. One reason for this could be that building permits are not really within one’s grasp. In comparison to the dataset during our lessons which was easy to understand and something you can imagine, the San Francisco dataset is far apart from our everyday life and nothing you are familiar with. Because you cannot change the data set, the only way to handle this difficulty was to search for explanations in the internet and to get somehow familiar with the dataset by looking at the data.

A big problem was that the „Current Status“ column consists not only of Strings but has 14 different variations. Predicting and calculating correlations with other columns is not possible without constructing another column or replacing the values. We decided to sort the statuses into two different categories: permitted and not permitted, so that in the column you have 0 or 1 (false or true). This makes the exploration a lot easier. Another method would be to do hot encoding and creating a new column for each status. But this would make the exploration more complex and therefore, effortful.

Last but not least the differentiation between the exercises was quite difficult because there is no clear line between for example A and B. Does e.g. factorizing and categorizing belong to exercise A or exercise B? That’s why we decided to do everything in one notebook (also because the exercises are somehow connected with each other because you needed the cleaned and imputed version of the dataset from A in the other two exercises).