From Kaggle there is a bunch of data from the city of San Francisco. I took 2 datasets, one with building permit data and the other describing the patrons of the libraries. The goal was to take the well-structured CSV files and create a database in 3NF that allows mixed querying between the dataset. Additionally, there is enough geographic and demographic data in the dataset to make some exploratory judgements about some of the neighborhoods.

We downloaded the data from Kaggle and imported it into a jupyter notebook for exploration. We then used an online tool to design the database relationships. We settled on two main tables, one each for the permits and the patrons and 20 reference tables.

Cleaning the data was a straight forward process of renaming columns.

In the notebook we imported both of the flat table into POSTGres. We used a series of pandas functions to extract the unique keys for the reference tables put them into a dictionary and load the data. This method ensured that when we set the values on the main tables it’s constructed uniquely from the reference tables. From there we used a staging areas from the earlier imported flat tables to construct the main tables. Finally we designed views that recreated the original data tables.

The new normalized DB is about 88% the (141MB vs 160MB) than the two flat tables. Additionally for large searches (100k+ results) there is a measurable difference in the speed with which rows are retuned.