**Notes about San Francisco crime dataset**

Initial Data Investigation

**sf\_data.csv**

Shape: 2,129,216 observations, 10 attributes

* Id: All are unique
* Category: 36 different categories
* Description:
  + 848 different categories 🡪 393 categories i.e., 46%, had less than 100 observations
    - I will not focus on description because of this
  + The description is uniquely within the category except for 9 incidences.
    - Especially *recovered vehicle* and *vehicle theft* sees overlapping descriptions
* Weekday: 7 days
* Date:
  + Min: 2003-01-01, Max: 2018-05-15
  + 2018-05-15 only observations in the hours 1 and 2 🡪 Is omitted from data set
  + will be partitioned into: day, year, month using DateTime type
* Time: I will create an ‘hour’ variable from this. Categorical with 12 different options.
* Label ['violent', 'other', 'prostitution', 'intoxication'], Unique for each category. The label *Prostitution* is only used in the category “prostitution”.
  + Should be excluded if trying to classify the category of the crime. Could be interesting for visualization purposes.

**sf\_districts.csv**

Shape: 592,854 observations, 2 attributes

* Three districts: ‘sunnydale’, ‘tenderloin’, ‘mission’.

**Combining the data sets**

* Only districts for a subset of the total amount of districts in San Francisco
* There exist 10 police districts in SF. Link: <https://sfgov.org/policecommission/police-district-maps>
* I choose to focus on the three given districts limiting my analysis by omitting the other districts.

**General comments**

* Only five observations of prostitution in Sunnydale across all years
* Sunnydale generally sees very little crime (~13000 observations across 5 years)
* Sunnydale is also located very differently from Mission and Tenderloin, which are more downtown