**Notes about San Fransisco crime dataset**

Initial Data Investigation

**sf\_data.csv**

* 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: 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**

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