San Francisco Crime Data

16-02-2022

Caroline Amalie Fuglsang-Damgaard

Præsentation af dataset

sf_crime_data.csv

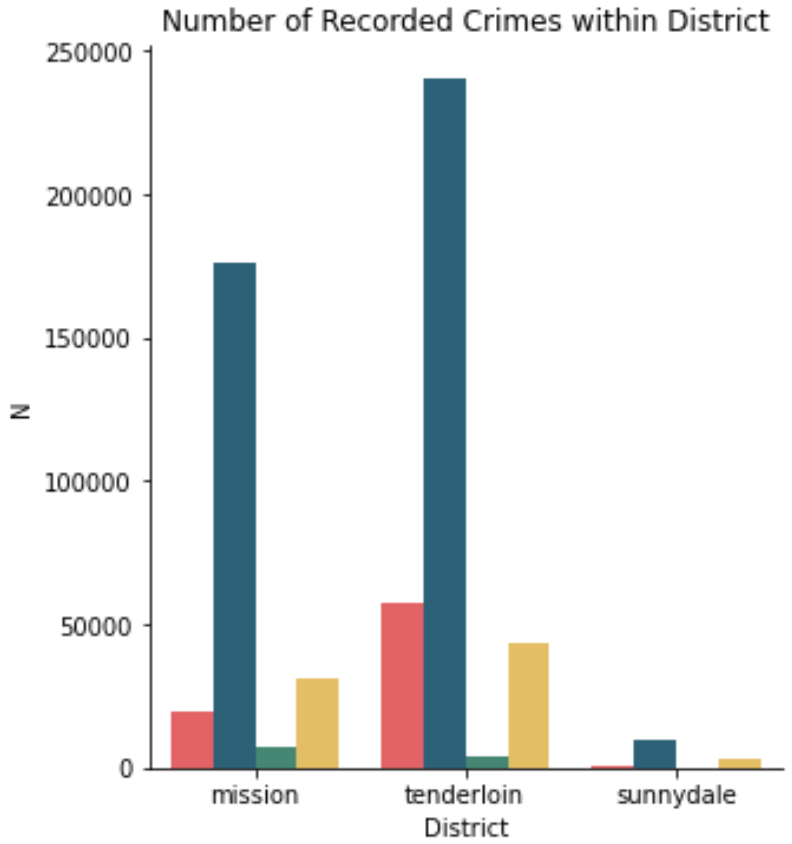
- 2.129.216 observationer, 10 attributter
- Id, category, description, weekday, date, time, resolution, longitude, latitude, label

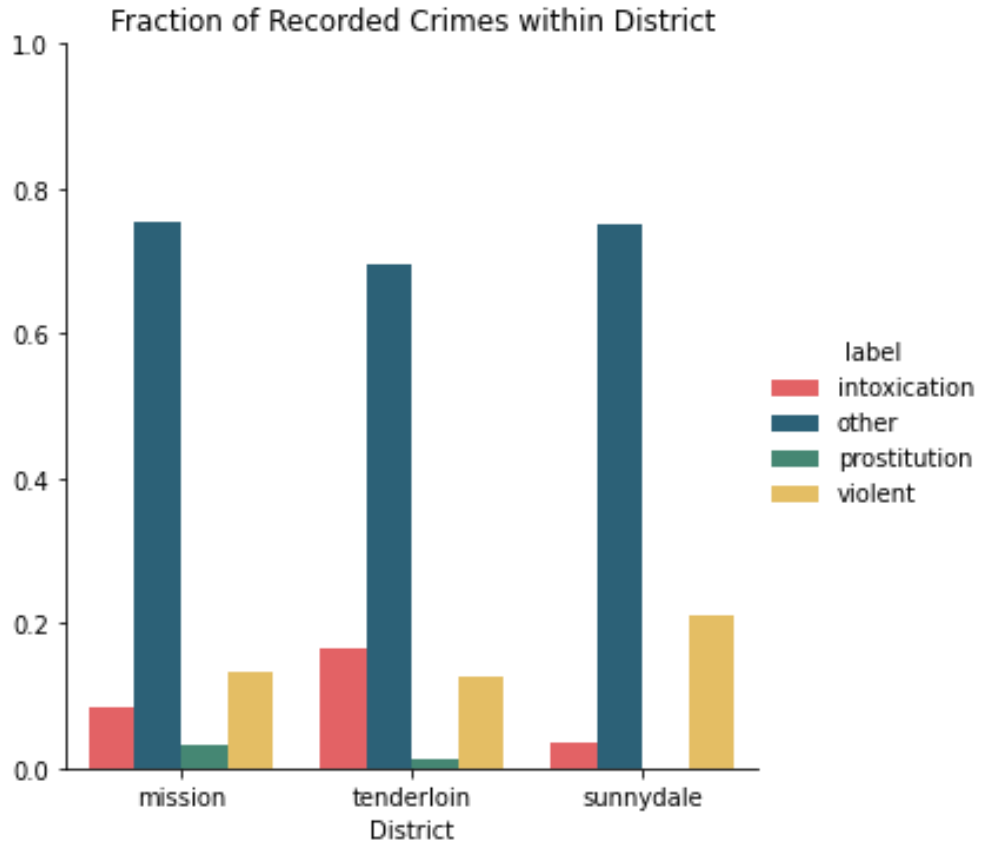
sf_districts.csv

- 592,854 observationer, 2 attributter
- Id, district
    - 3 unikke distrikter

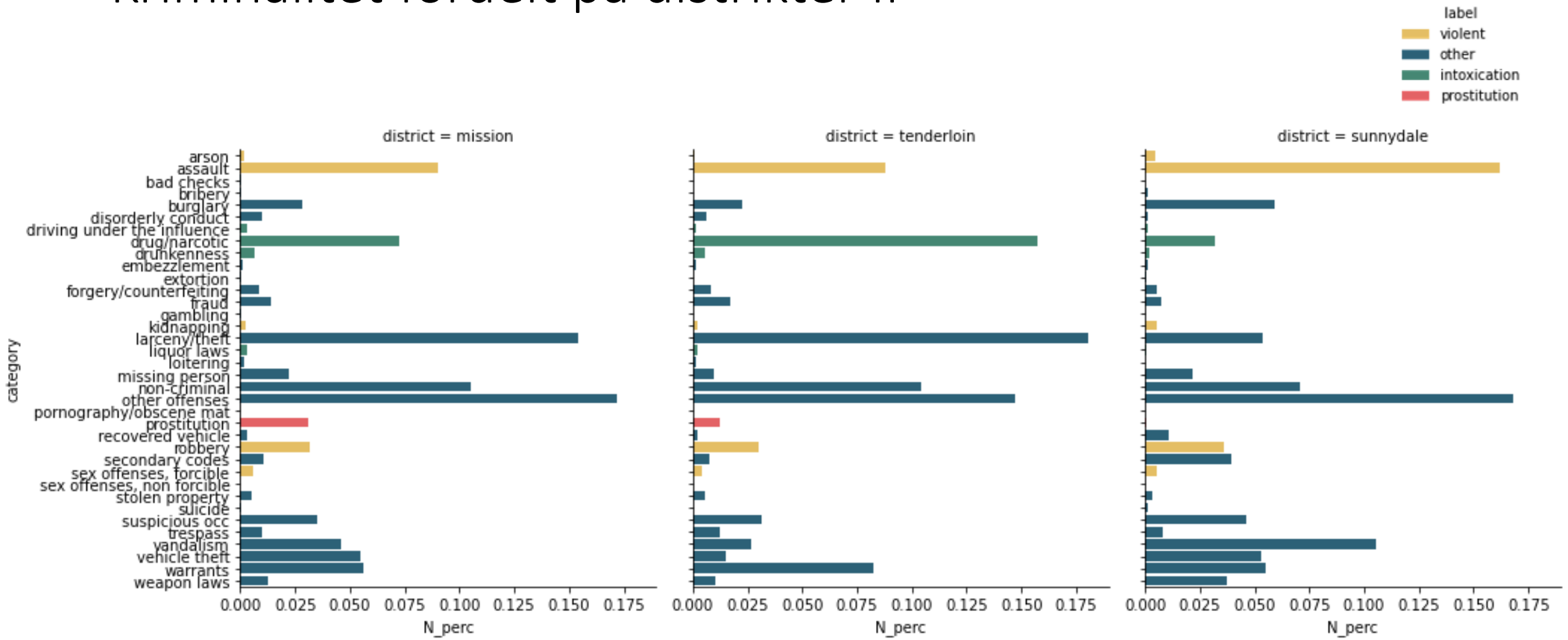
Kriminalitet fordelt på distrikter I

Distrikt	Mission	Tenderloin	Sunnydale
N	234.316	345.421	13.111

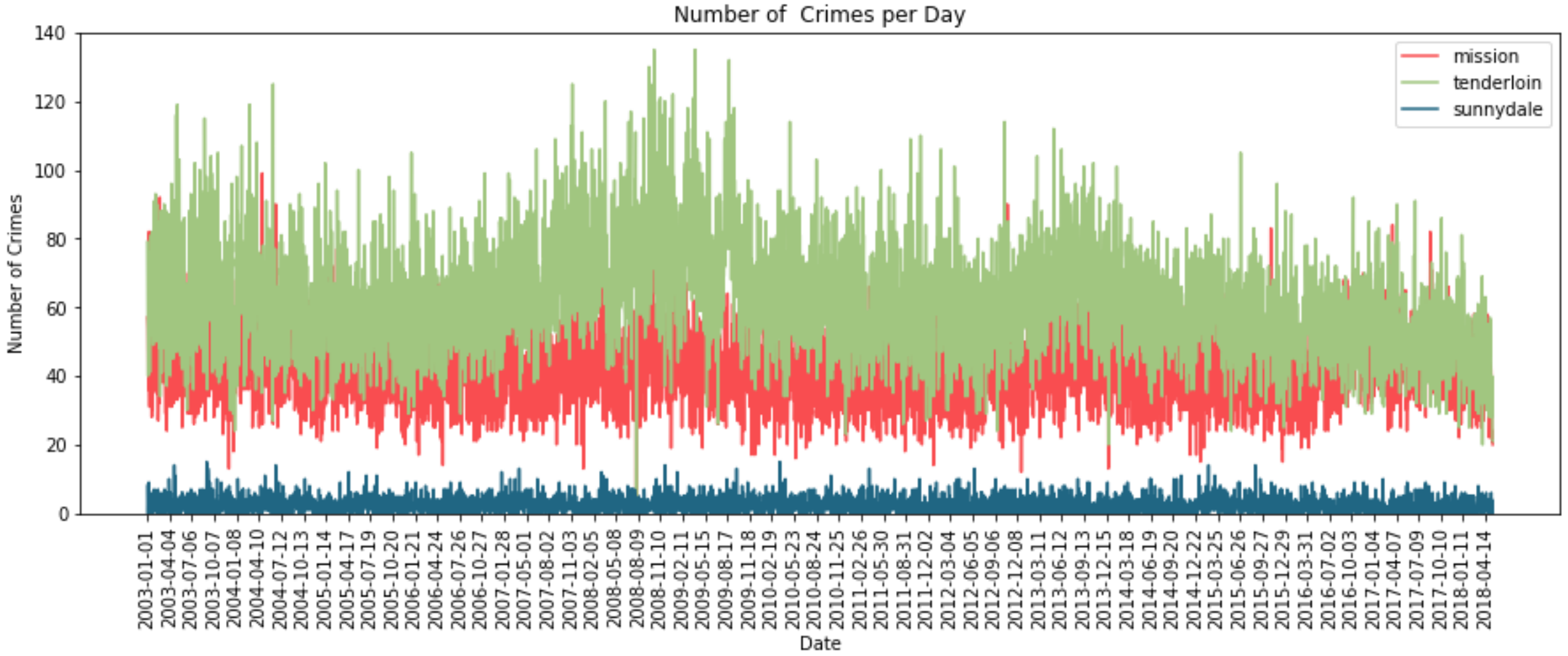




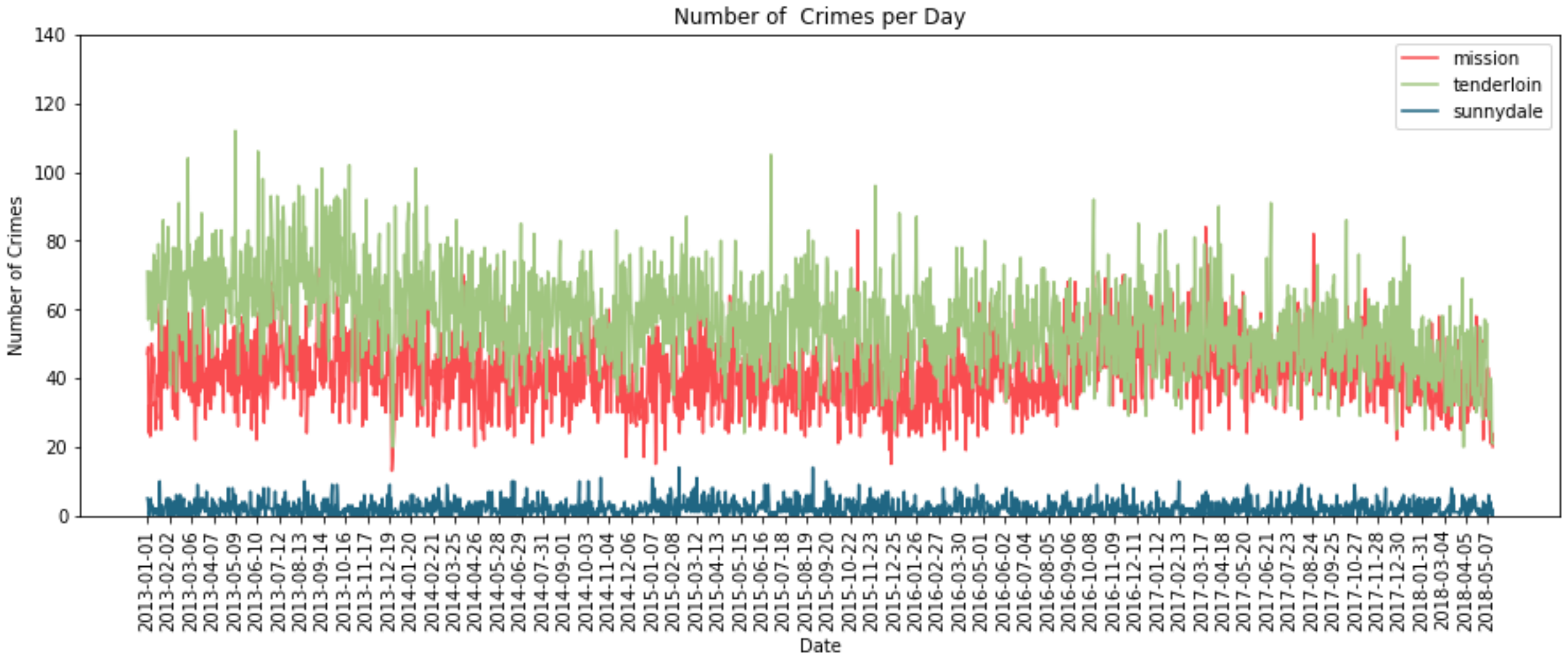
Kriminalitet fordelt på distrikter II



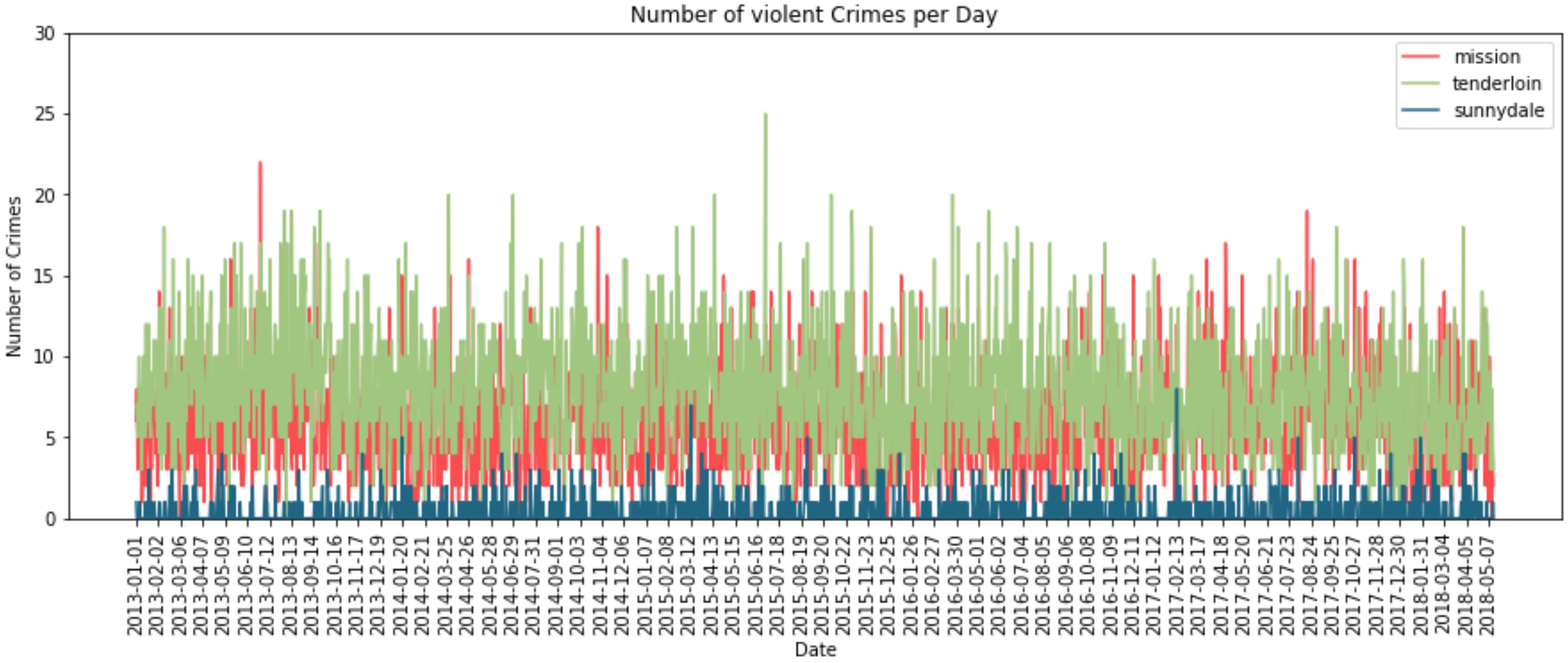
Kriminalitet over tid I



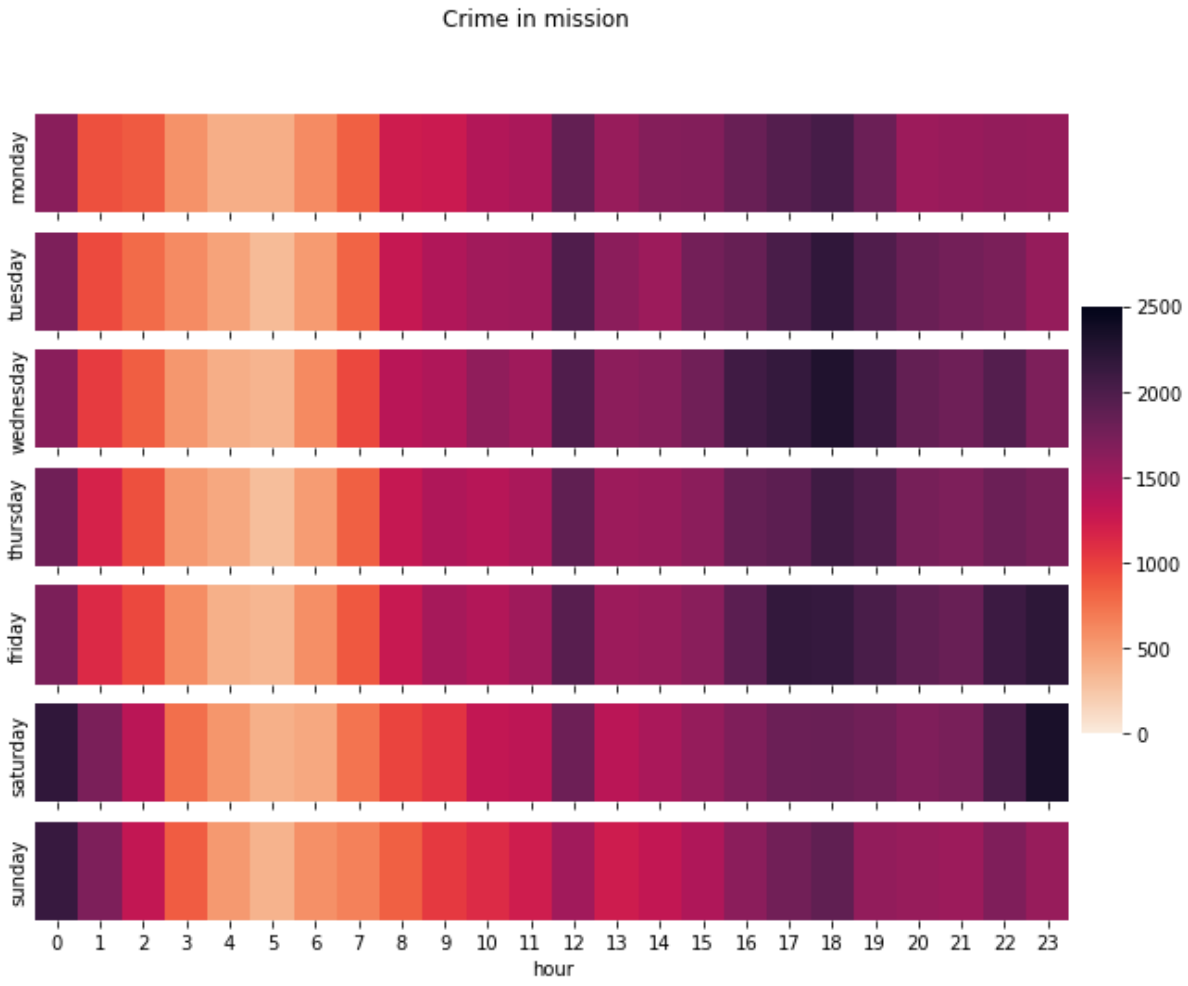
Kriminalitet over tid II

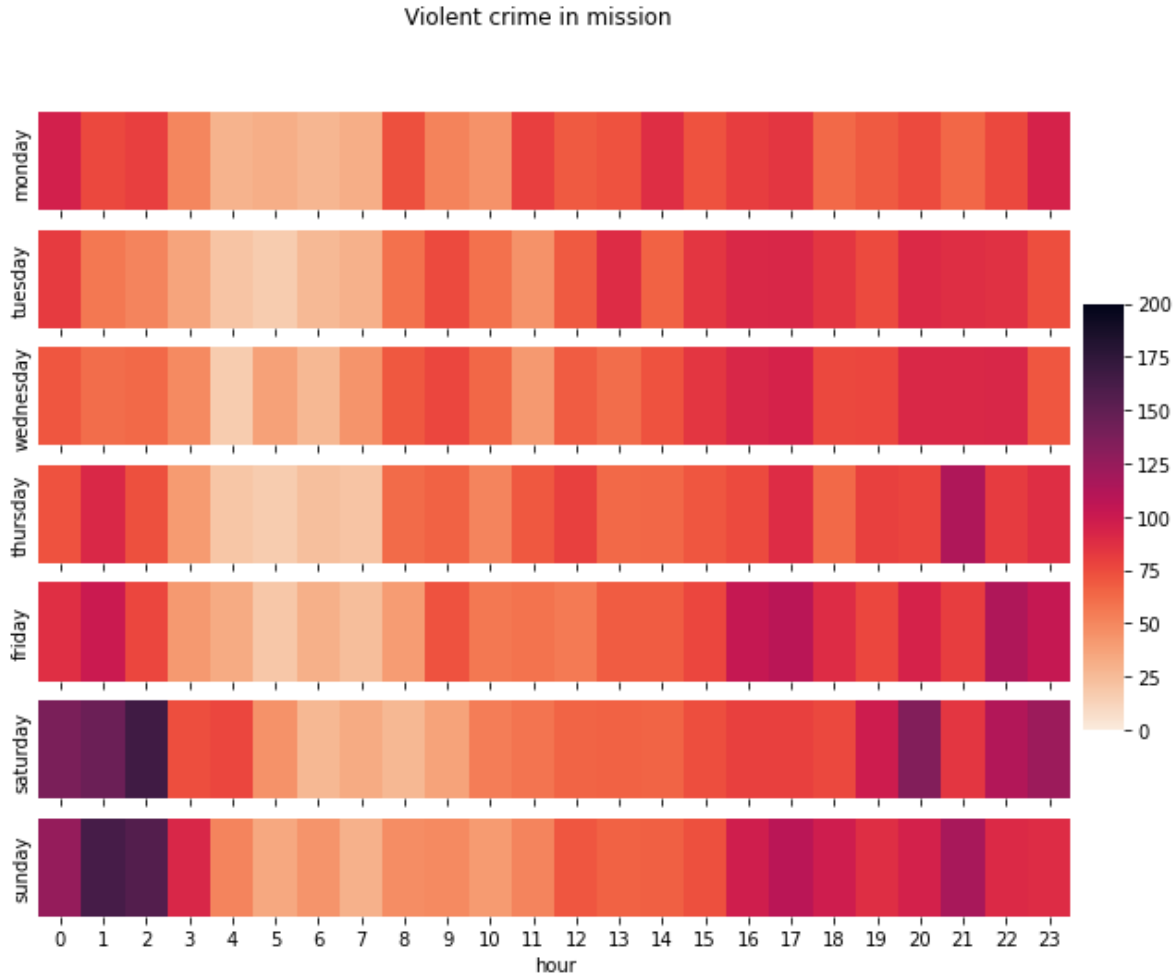


Kriminalitet over tid III



Voldelig kriminalitet fordelt på timer og uger





Case-idé I

Observationer

- Antallet af incidenser ser generelt nedadgående ud
- Men… mængden af voldelig kriminalitet ser stabil ud fra 2013-2018

Antagelser

- En voldelig situation kræver mere af politiet
- Voldelig kriminalitet har generelt større konsekvenser for ofrene og gør folk mere bange end anden kriminalitet
- Derfor har de tre distrikter har en interesse i at allokere deres ressourcer til bedre at kunne bekæmpe voldelig kriminalitet

Case-idé II

Spørgsmål

Kan vi bruge datasættet til at hjælpe de tre distrikter til at bekæmpe voldelig kriminalitet?

Ide til svar

Risikovurdering af om politiet møder en voldelig hændelse eller ej

Hvorfor

At kende risikoen for vold kan hjælpe politiet med at…

…Prioritere flere betjente til potentielt voldelige situationer

…Give den enkelte betjent det rette mindset inden han kører ud til en opgave

At være beredt på, at en situation kan være voldelig vil muligvis hjælpe politiet med at håndtere situationen bedre

→ Hvilket ultimativt ville kunne medføre et fald i voldelig kriminalitet

Modellering

Modellerer sandsynligheden for, at en given forbrydelse er voldelig baseret på attributterne

- *weekday*, *district*, *day, month, year, hour*
- Det er rimelig at tænke på hver forbrydelse som uafhængige hændelser
- Derfor anvendes der med et klassisk random train-val-test split

Logistisk Regression

Forbrydelse i er Bernoulli fordelt med sandsynligheden

$$p_i = P(Y = 1 \mid X = x_i)$$

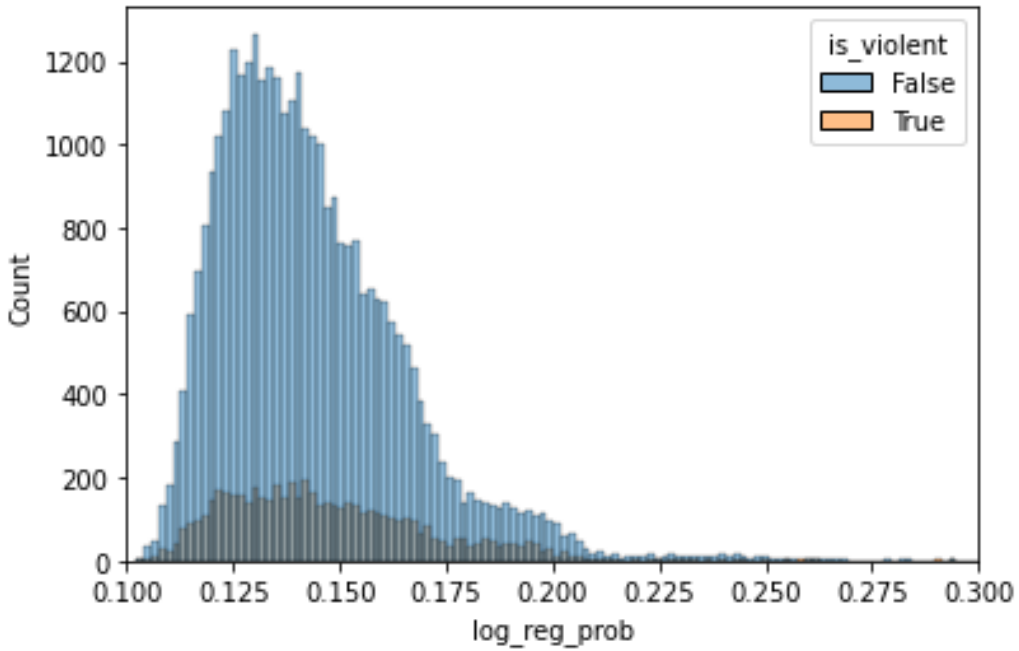
for at være voldelig

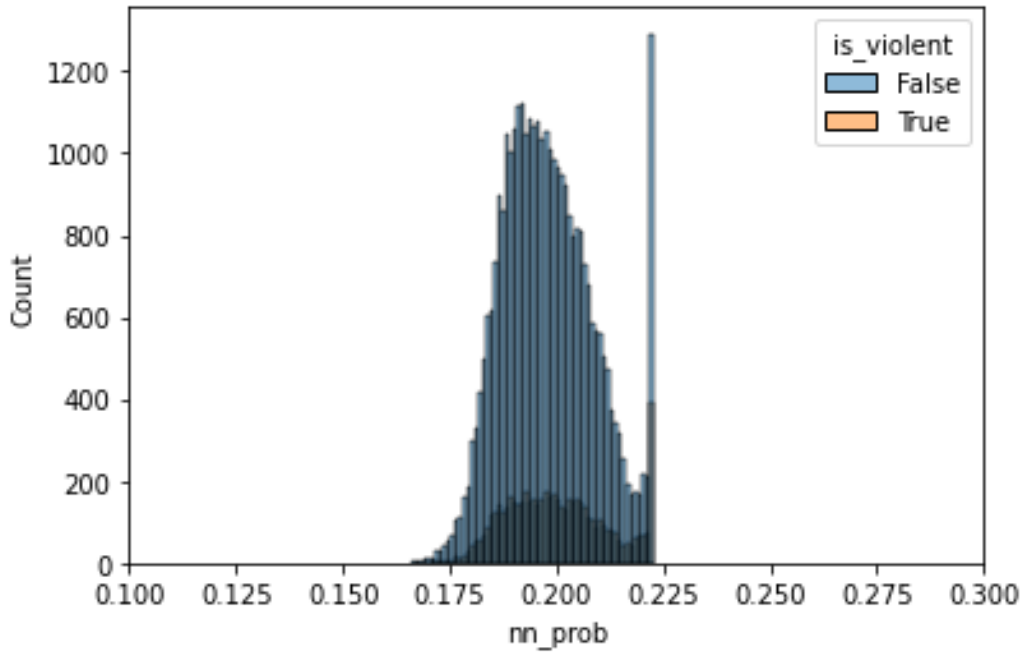
Neural Netværk

- Fleksibel ikke-linear model til at modellere sandsynligheden
- To hidden layers med 6 og 3 neuroner
- Learning rate på 1e-3
- Minibatches på 32, og early stopping samt max 10 epochs
- Sigmoid funktionen som output activation function

Resultater

Model	Accuracy test
Logistisk Regression	0.856
Neural Netværk	0.856





Opsummering

Konklusion

- De to modeller kan ikke skelne mellem voldelig og ikke-voldelig kriminalitet baseret på de givne attributter

Næste skridt

- Modellerne er ikke ret lovende → tilbage til det eksplorative

F.eks.

- Undersøg potentielle hot-spots for voldelig kriminalitet ved hjælp af koordinatdata
- Dyk ned i attributten "*category*"
  - Måske vil den større detaljegrad rumme forskelle, der er udvisket ved aggregeringen af voldelig/ikke-voldelig