2. Exercise

Data Science mit Python

Deadline: 09.12.25, Midnight

Wintersemester 2025/26

Patrick Schäfer, patrick.schaefer@hu-berlin.de

Agenda

- Questions?
- 2. Exercise

Question?

2.Exercise: Crime Statistics

Fighting Crime by Data Analysis

- Your aim is to support the police by developing preventive measures based on historical data from 2015-2018.
- A total of 319.073 minor to serious offences were identified during the period
- This exercise is about working exploratively on raw data

Dataset

- Look at the data
  - Use e.g. a text editor, Excel, Pandas

- The dataset contains the following columns:

```
'INCIDENT_NUMBER', 'OFFENSE_CODE',
'OFFENSE_CODE_GROUP', 'OFFENSE_DESCRIPTION',
'DISTRICT', 'REPORTING_AREA', 'SHOOTING',
'OCCURRED_ON_DATE', 'YEAR', 'MONTH', 'DAY_OF_WEEK',
'HOUR', 'UCR_PART', 'STREET', 'Lat', 'Long',
'Location‘
```

Dataset

- Familiarise yourself with the data set
- The data are stored in a flat CSV file in tabular form

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_
0	I182070945	619	Larceny	LARCENY ALL OTHERS	D14	808	NaN	2018-09-02 13:
1	I182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:
2	I182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:
3	I182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:
4	I182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:
...	...	...	...	...	...	...	...	

Data

- `'INCIDENT_NUMBER`:

  Running Number

- `'OFFENSE_CODE', 'OFFENSE_CODE_GROUP', 'OFFENSE_DESCRIPTION`:

  Description of the criminal offence

- `'DISTRICT', 'REPORTING_AREA`:

  Place where it took place and was reported

- `'SHOOTING`:

  Was a gun involved

- `'OCCURRED_ON_DATE`, `'YEAR', 'MONTH', 'DAY_OF_WEEK', 'HOUR`:

  Dates

- `'UCR_PART`:

  Seriousness of the offence

  (Part One: "Serious", Part Three: "Light")

- `'STREET', 'Lat', 'Long', 'Location`:

  Place/location of crime

Hierarchies

- 'OFFENSE_CODE_GROUP'
  =>'OFFENSE_CODE'
  =>'OFFENSE_DESCRIPTION'

619
=> Larceny
=> LARCENY ALL OTHERS

1402
=> Vandalism
=> VANDALISM

- 'YEAR'
  =>'MONTH'
  =>'DAY_OF_WEEK'

Exercise

- The exercise consists of three parts

- 1a) Classify each Column
  - Numerical or Categorical
- 1b) Exploratory Data Analysis
  - There are four key questions you must investigate
  - Key questions have follow-up questions to answer
- 2) Law Enforcement Dashboard (Vibe Coding)

Part 1a. Classify Columns

1. Classify the data types of each column
   a) Numerical
      - Continuous or Discrete
   b) Categorical
      - Nominal or Ordinal

2. If a column is discrete or ordinal:
   - State the reason for your decision

Part 1b. Key Questions

1. How has the total number of offences developed over the years?
   a) Which offences are the most frequent?
   b) How has the number of serious crimes `('Part One')` developed over the years?
   c) Why is the total number of offences (so) low in 2015 and 2018?

2. In which urban areas (`district`), broken down by year, were most crimes committed?
   a) In which urban areas (`district`) are most serious crimes `('Part One')` committed?
   b) Which types of serious crimes `('Part One')` occur most frequently in the urban area 'B2'?

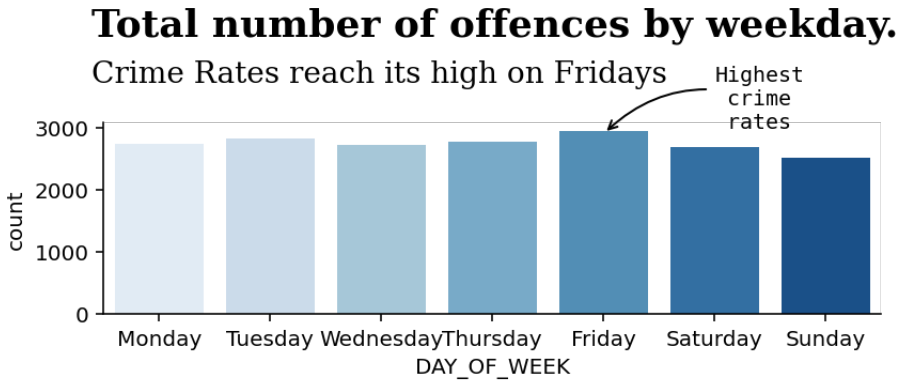
Part 1b. Key Questions (continued)

3. Are there (a) times, (b) days or (c) months when more serious crimes `('Part One')` occur?
   a) Do crimes tend to occur at night or during the day?
   b) When are the most police officers needed?

4. How has the number of shootings developed in recent years?
   a) In which district do most shootings take place?
   b) In which street do most shootings take place?
   c) At what times do most shootings take place?

Possible Exemplary Solutions

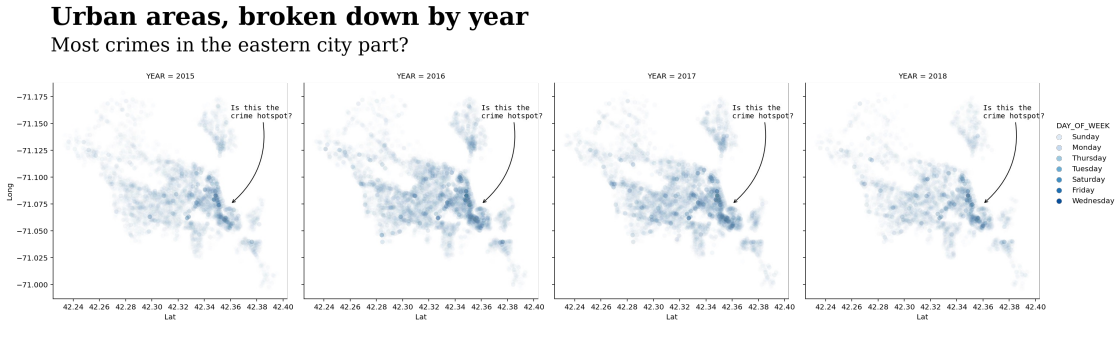
- Option 1: Plots



- Option 2: Tabular

Year	2015	2016	2017	2018
count	12256	19222	18316	11835

- Option 3: Maps (using Lat+Long)

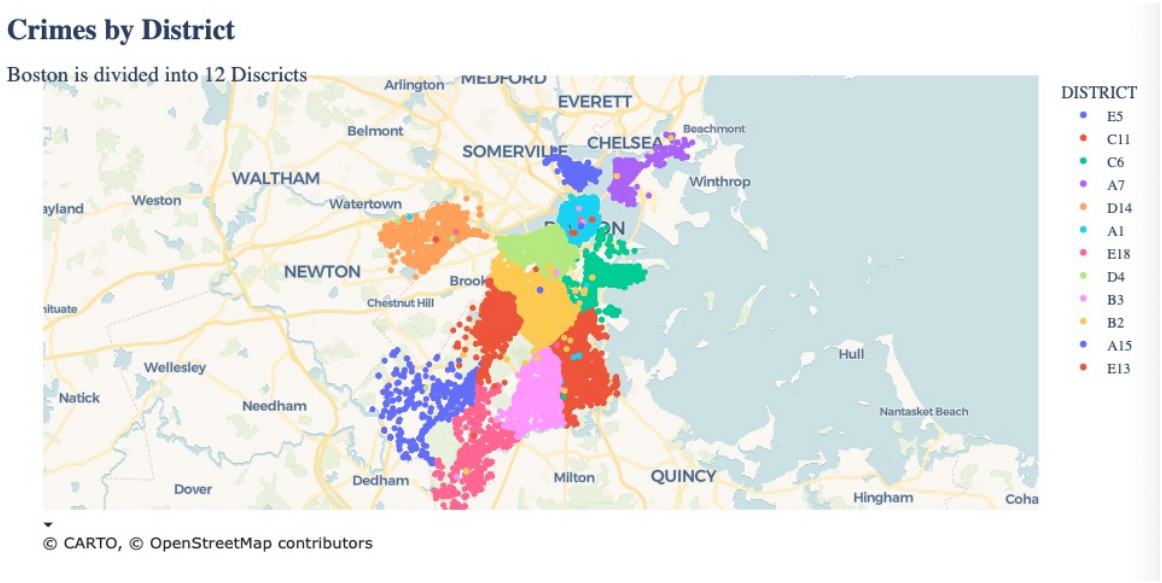


1. How has the total number of offences developed over the years?
   - **It has risen until … and decreased from …**
   - **The area has changed …**

Always add a description!

Possible Exemplary Solutions

- Option 3: Maps Continued



1. How has the total number of offences developed over the years?
   - **It has risen until ... and decreased from ...**
   - **The area has changed ...**

Always add a description!

Remarks

- Obviously, **selection, grouping** and **aggregation** are used in the key questions respectively
- Something known from databases:

```
SELECT count(incident_number)
  FROM crimes
 GROUP BY district
```

- Other than the previous exercise you are supposed to visualize and explain the results

Learning Goals

- You are given the freedom to use the following libraries
  - Seaborn,
  - Pandas,
  - Matplotlib,
  - (PlotLy for OpenStreetMap scatter_map)

- You will be given a starter Python: Exercise-2-primer.ipynp

Getting Started in Python

```
import pandas as pd
df = pd.read_csv('crime.csv.zip', compression='zip')
df
```

See Exercise-2-primer.ipynb in Moodle

	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_
0	I182070945	619	Larceny	LARCENY ALL OTHERS	D14	808	NaN	2018-09-02 13:
1	I182070943	1402	Vandalism	VANDALISM	C11	347	NaN	2018-08-21 00:
2	I182070941	3410	Towed	TOWED MOTOR VEHICLE	D4	151	NaN	2018-09-03 19:
3	I182070940	3114	Investigate Property	INVESTIGATE PROPERTY	D4	272	NaN	2018-09-03 21:
4	I182070938	3114	Investigate Property	INVESTIGATE PROPERTY	B3	421	NaN	2018-09-03 21:
...	...	...	...	...	...	...	...	

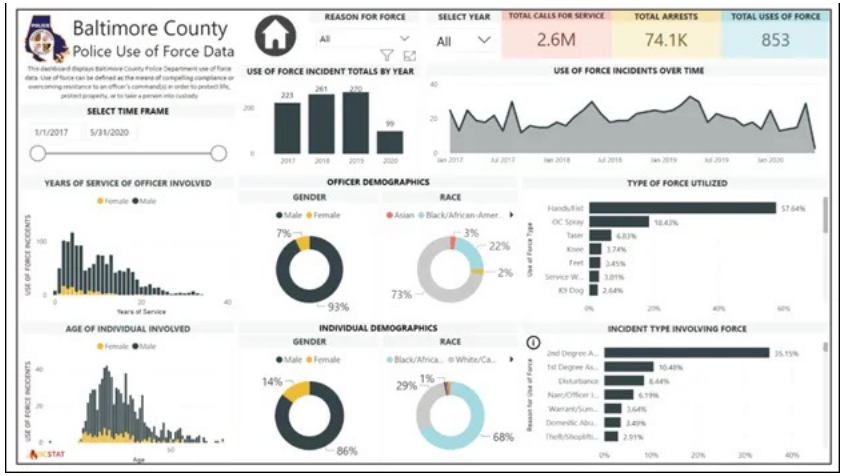
Jupyter Notebook

- Recommended way to get started
  - Download the starter **Notebook in Moodle**
  - Solve the exercise

- We will discuss the primers, next

Part 2. Law Enforcement Dashboard (Vibe Coding Task)

- Develop a **Streamlit-based** dashboard to assist law enforcement agencies in generating actionable insights from crime data through an intuitive interface
- **Required Features**
  - **Data:** Data must be uploaded – (read from Boston crimes.csv)
  - **Date Selection Mechanism**: Include a slider or date range selector allowing users to filter data dynamically based on specific time periods – can optionally show filters on discrict.
  - **Summary Statistics:** Display at least **three key aggregate metrics**, such as the total counts of UCR Part One offenses, UCR Part Two offenses, and shooting incidents,
  - **Data Visualizations:** Integrate **three distinct plots** to reveal patterns, such as a line chart for daily crime volumes with a 7-day rolling average, a bar chart for crimes distributed by district, or a pie chart of the top 10 offense codes.
- Use an LLM, document your prompts



Example

Part 2 – Continued

- Some hints:
  1. Do not put too much into one plot
  2. First ask the LLM to plan your app
  3. Use the LLM to fix errors
  4. Upload an image and as the LLM to reproduce this layout
  5. Ask for the LLM to simplify your code

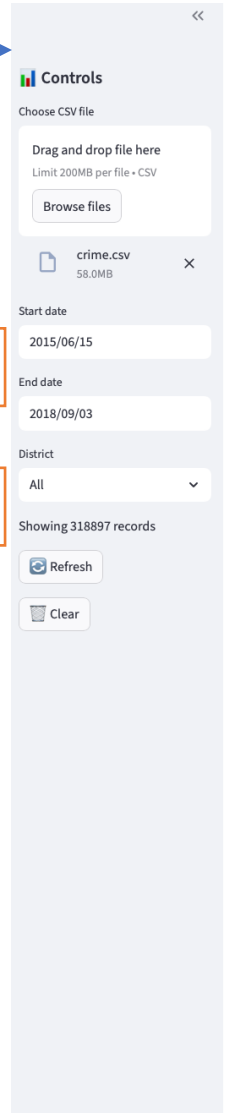
  6. Use a previous code base and past it into the chat to continue with this one
  7. It is ok to make final adjustments yourself (i.e. removing text)

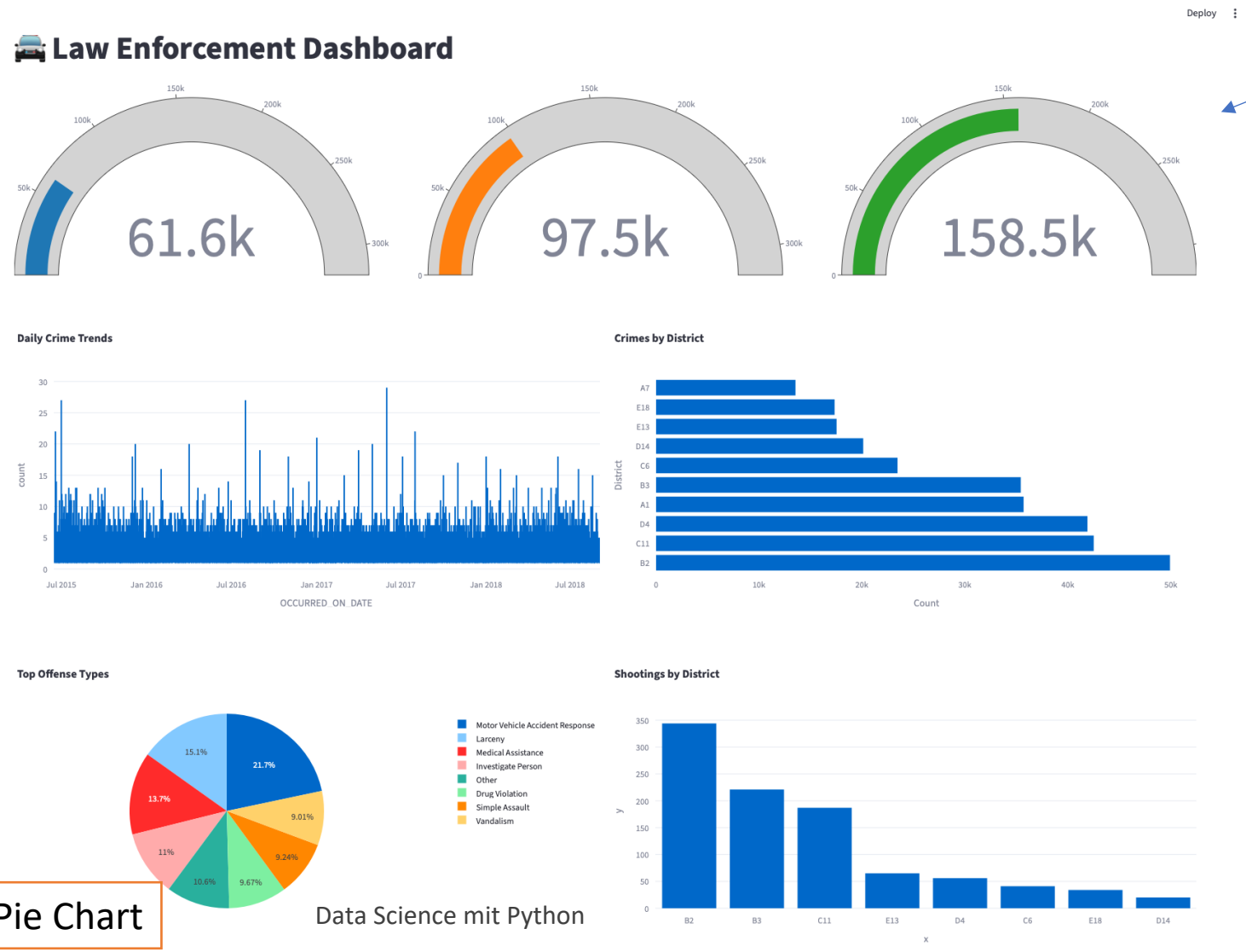
Dashboard Example (~30 Minutes, Claude)

Sidebar

st.date_input

st.selectbox





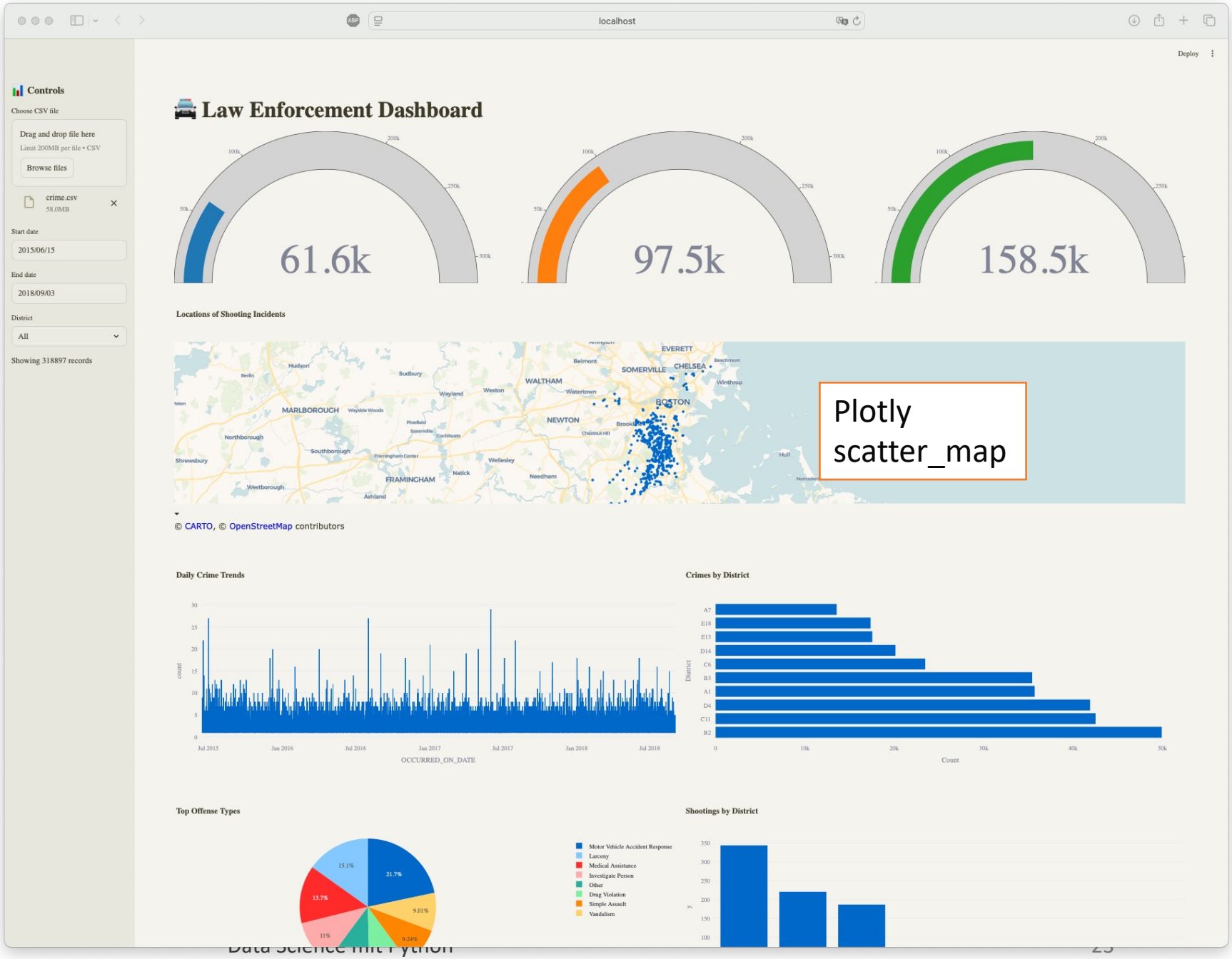
Plotly Gauge Plot

Pie Chart

More Inspiration

- Configure style with: .streamlit/config.toml

https://maxbraglia.substack.com/p/the-streamlit-theming-method-that



Submission Deadline: Tue, 09.12.25

- The task is considered to have been **successfully solved,** if
  - Part 1: An answer backed by evidence has been given for each of the **4 key questions**
  - Part 2: Your dashboard **can be executed** and contains **the requested features**
- **Deliverables:**
  - **Part 1: EDA (2 Files)**
    a) **Report - Exploratory Data Analysis:**
      - This contains only your **solutions and answers**:
        - These can be tabular or using plots
      - Format**: PPT, Word, PDF, HTML**

    b) **Your code**
  - **Part 2: Vibe Coding** (**>3 Files!**):
    - **The file** with **python code**
    - **One file** documenting the **used prompts** and the **used LLM**
    - **Screenshot(s)** or a small video

Questions?