# ## Tasks:

[A] analyze available events data to identify trends, seasonality impact and consistency of data across different US cities.

[B] build a machine learning model to predict the number of events of a given type for a particular date and location to better plan the allocation of city resources. The use of external datasets is likely necessary to improve model performance.

[C] make predictions more explainable by using model interpretability tools. You also need to ensure that the model won’t lead to undesired biases or discrimination when it is put into production.

[D] Evaluate the model performance during city-wide lockdowns during the COVID-19 pandemic, and apply corrections, if necessary.

# Techniques employed

Following are some of the techniques and tools employed in this project. You don’t need to be an expert in each of them, and material is provided to get you started.

--> Cleaning, filtering and pre-processing the data loaded from Parquet files:

pandas: a data analysis / manipulation library for Python

--> Analysis and visualization of location based and time-dependent data:

pandas

matplotlib: the reference Python plotting library

--> Interactive data visualization:

jupyter notebook: a web-based notebook environment for interactive computing

ipywidgets: interactive widgets for the Jupyter Notebook

--> Training and evaluation of machine learning models, model selection and hyper-parameter tuning, model interoperability

and validation:

scikit-learn: a machine learning library for Python

--> Creating python packages for data science projects.

--> Evaluating the risk of biases and discrimination when deploying models in production.

--> Dealing with unforeseen events (for example, COVID-19) in predictive models.

## Project outline

This project is organized into six milestones.

1. Load and Clean the Raw Data

2. Perform an Exploratory Analysis and Interactive Visualization of the Data

3. Build a machine learning model to predict the frequency (number) of different events in San-Francisco on a given date and in a particular neighborhood

4. Enhance the Model with External Data

5. Make the Model More Interpretable to Explain Predictions

We also need to ensure that there are no undesired biases and discrimination in the model predictions.

6. Predictions in 2020 and Generalization to Other Cities

Consider what would be necessary to generalize this model to other US cities (New-York, Chicago). Evaluate and, if possible, correct predictions made during the 2020 COVID-19 crisis.

The deliverable for each milestone is either a Jupyter notebook, a processed dataset, or an updated Python package.

Each milestone builds upon the results obtained in the previous ones.

Overall, this project is representative of a typical workflow in real-world Data Science applications.

## Books (Resources)

Data Science Bookcamp by Leonard Apeltsin

Exploring Data Science With chapters selected by John Mount and Nina Zumel

Introduction to Machine Learning with Python by Andreas C. Müller and Sarah Guido

Documentation for:

- pandas

- matplotlib

-scikit-learn

- ipywidgets

# - Manning Project FAQ and email mentor

**Milestone: 1 Load and Clean the Raw Data**

Objective:

--> Explore the data for different US cities to identify any data qualities issues.

--> Prepare a cleaned dataset with duplicate entries removed for each city.

Workflow

step-1 Load the data for San-Francisco and look at a few examples. For instance,

step-2 Identify potential data quality issues, including:

- missing values (or empty values for string columns)

- columns that contain no information

- columns that are of the wrong type (e.g. numbers encoded as strings).

- rows with identical content (duplicate rows)

- You can use pandas-profiling package to automate some of this analysis.

step-3 Store the cleaned dataset to disk in parquet format as safety-SanFrancisco-1.parquet with

- columns that contain constant values removed

- date columns converted to a datetime type

- categorical columns (that have less than a few 100 unique string values) converted to categorical data type, in order to use less memory.

- rows with identical values removed. This can be achieved with the DataFrame.drop\_duplicates method. It is recommended to write a python function with these steps which will be re-used later.

step-4 Examine the data for New-York city to ensure that it has the same data scheme. Apply the preprocessing from step 3 and store the results in the same way.

# Deliverable

The deliverable consists of:

- A Jupyter Notebook documenting your workflow, as you load and explore the data to identify potential data quality issues.

- A Python function (included in the notebook) allowing to clean and deduplicate the data as described in step 3 of the workflow.

- The resulting processed dataset for San-Francisco, and New-York stored as

safety-SanFrancisco-1.parquet, and safety-Nyc-1.parquet respectively.

#### Upload a link to your Jupyter Notebook (preferably hosted on GitHub) in the blank below and hit submit. After submitting, you can view an example solution in the next section.

### Importance to project

### -------------------------------

### In real word projects it is very important to check that the obtained data has no obvious issues or missing parts, before diving into detailed analysis or modeling. In particular, it should be sufficient to carry out the project objectives. If a new data extraction is necessary to address data quality issues, it is likely to take some time, particularly in large organizations, and may put the schedule of our project at risk. Therefore the earlier data quality issues are detected the better.

### Looking through the data manually on a few examples and typical scenarios is important to get more familiar with it, and check our initial assumptions. It may also lead to questions that could be worthwhile to investigate later.