Tab 1

Project Title	**PatrolIQ - Smart Safety Analytics Platform**
Skills take away	**Python, Streamlit Cloud Deployment, Machine Learning, Data Analysis, Unsupervised Learning, Clustering Algorithms, Dimensionality Reduction, Geographic Data Analysis, Data Visualization, MLflow**
Domain	**Public Safety and Urban Analytics**

Problem Statement

Build a comprehensive urban safety intelligence platform that leverages unsupervised machine learning techniques to analyze crime patterns and optimize police resource allocation. Step into the role of a crime analyst working with law enforcement agencies to make cities safer through data-driven decisions.

Imagine you are a crime intelligence analyst at the Chicago Police Department. Every day, officers ask critical questions: "Where should we patrol tonight?", "Which neighborhoods need more resources?", "When do most crimes occur?". Your mission is to analyze 500,000 crime records and provide insights that could prevent crimes and save lives.

Urban areas face significant challenges in efficient police deployment and crime prevention due to lack of actionable insights from massive crime datasets. This project aims to solve critical public safety issues by identifying crime hotspots, understanding

temporal patterns, and providing intelligent analysis that law enforcement can act upon immediately.

The platform should deliver:

- Crime hotspot identification through geographic and temporal clustering
- Simplified visualization of complex crime patterns using dimensionality reduction
- Real-time analysis of 500,000 crime records from Chicago
- Advanced feature engineering from 22 crime and location variables
- MLflow integration for experiment tracking and model comparison
- Streamlit Cloud deployment for production-ready access.

Business Use Cases

Police Departments

- Optimize patrol route allocation and reduce response time by 60%
- Identify high-risk areas requiring increased police presence
- Predict crime patterns for proactive prevention strategies
- Evidence-based resource deployment for budget optimization

City Administration

- Data-driven urban planning for safer neighborhoods
- Strategic placement of surveillance systems and street lighting
- Justify public safety budget allocation with concrete insights
- Monitor crime trends across different districts and time periods

Law Enforcement Analytics Firms

- Provide crime intelligence services to multiple jurisdictions
- Develop predictive policing solutions for client cities
- Benchmark safety performance across different urban areas
- Generate comprehensive crime analysis reports for stakeholders

Emergency Response Systems

- Prioritize emergency calls based on area risk assessment
- Optimize ambulance and fire department deployment
- Coordinate multi-agency response in high-crime zones

- Real-time situational awareness for first responders

Approach

Step 1: Data Acquisition and Preprocessing

- Download full Chicago crime dataset (7.8 Million records, 1.7 GB)
- Sample 500,000 recent crime records for analysis
- Implement comprehensive data cleaning for missing values and inconsistencies
- Extract temporal features (hour, day of week, month) from datetime fields
- Apply data quality assessment and validation checks

Step 2: Exploratory Data Analysis

- Analyze crime distribution across 33 different crime types
- Study geographic patterns using latitude and longitude coordinates
- Investigate temporal trends (hourly, daily, monthly, seasonal patterns)
- Examine arrest rates and domestic incident correlations
- Generate comprehensive statistical summaries and crime insights

Step 3: Feature Engineering

- Create temporal features (hour of day, day of week, weekend flag, season)
- Generate geographic features (district clustering, coordinate binning)
- Develop crime severity scores based on crime type classification
- Apply categorical encoding for crime types and location descriptions
- Normalize geographic coordinates for distance-based calculations

Step 4: Unsupervised Learning - Clustering Analysis

Geographic Crime Hotspot Clustering - Minimum 3 Algorithms Required:

- K-Means Clustering for identifying distinct crime concentration zones
  - *Expected Result:* Identify 5-10 distinct geographic crime hotspots with clear cluster boundaries
- DBSCAN for density-based spatial clustering with noise detection
  - *Expected Result:* Detect high-density crime areas and filter out noise/outliers automatically
- Hierarchical Clustering for nested geographic area analysis

    - *Expected Result:* Create dendrogram showing hierarchical relationships between crime zones
- Model evaluation using silhouette score, Davies-Bouldin index, and elbow method
    - *Expected Result:* Achieve silhouette score above 0.5 for best performing algorithm
- Best performing clustering algorithm will be selected for deployment

Temporal Pattern Clustering:

- K-Means on temporal features (hour, day, month)
    - *Expected Result:* Identify 3-5 distinct time-based crime patterns (e.g., late night crimes, rush hour incidents)
- Identify peak crime time periods and seasonal patterns
    - *Expected Result:* Discover high-risk time slots and seasonal trends for resource planning
- Group similar time-based crime behaviors
    - *Expected Result:* Create temporal crime profiles for different types of incidents

Step 5: Unsupervised Learning - Dimensionality Reduction

Minimum 2 Techniques Required:

- PCA (Principal Component Analysis) for feature reduction and variance explanation
    - *Expected Result:* Reduce 22+ features to 2–3 principal components explaining 70%+ variance
    - *Expected Result:* Identify top 5 most important features driving crime patterns

- t-SNE or UMAP for 2D visualization of high-dimensional crime patterns
    - *Expected Result:* Generate clear 2D scatter plots showing distinct crime clusters
    - *Expected Result:* Visualize separation between different crime types and geographic zones

- **Reduce 22+ features to 2–3 principal components**

  - *Expected Result:* Maintain 70–80% of original data variance in reduced dimensions

- **Visualize crime clusters in lower-dimensional space**
  - *Expected Result:* Create interactive plots showing crime patterns in 2D/3D space

- **Identify most important features driving crime patterns**
  - *Expected Result:* Rank features by importance (e.g., time of day > location > crime type)

Step 7: MLflow Integration and Experiment Tracking

- Configure MLflow tracking server for organized experiment management
- Log clustering parameters (K values, distance metrics, algorithms)
- Track dimensionality reduction metrics (explained variance, reconstruction error)
- Compare different algorithm performances and select best models
- Implement model registry for version control

Step 8: Streamlit Application Development

- Multi-page web application with interactive crime visualizations
- Geographic crime heatmaps with cluster boundaries
- Temporal pattern analysis dashboards
- Interactive dimensionality reduction visualizations
- Model performance monitoring and MLflow integration

Step 9: Cloud Deployment and Production

- Deploy complete application on Streamlit Cloud platform
- Implement responsive design for cross-platform accessibility
- Configure automated deployment pipeline from GitHub repository
- Ensure proper error handling and user feedback mechanisms

Data Flow and Architecture:

Chicago Crime Dataset (7.8M Records)

↓

Download and Sample (500K Records)

↓

Data Quality Assessment & Preprocessing

↓

Feature Engineering & Exploratory Analysis

↓

Clustering Analysis (Geographic + Temporal)

↓

Dimensionality Reduction (PCA + t-SNE)

↓

MLflow Experiment Tracking

↓

Streamlit Application Development

↓

Cloud Deployment & Performance Testing

↓

Production-Ready Safety Intelligence Platform

Architecture Components:

- Data Layer: Chicago crime records with geographic and temporal information
- Processing Layer: Data cleaning, feature engineering, and sampling pipelines
- Model Layer: Clustering algorithms and dimensionality reduction with MLflow tracking
- Application Layer: Multi-page Streamlit web application with interactive visualizations
- Deployment Layer: Streamlit Cloud hosting with GitHub integration and CI/CD

Dataset: Chicago Crime Dataset

Dataset Scale:

- **Full Dataset:** 7.8 Million crime records (2001-2025)
- **Sample Used:** 500,000 recent crime records
- **Input Features:** 22 comprehensive variables
- **Crime Categories:** 33 distinct crime types
- **Geographic Coverage:** City of Chicago districts and wards

Data Source:

Chicago Data Portal - Crimes 2001 to Present (Public Dataset)

- **Official Link:** Chicago Live Dataset
- **Format:** CSV (Comma-separated values)
- **Update Frequency:** Daily updates from Chicago Police Department

How to Access the Dataset:

Download via Web Interface

1. Visit the official link:Chicago Live dataset
2. Click on **"Export"** button (top right corner)
3. Select **"CSV"** format
4. For large datasets, use filtering options before export or download in batches

Crime Type Distribution (33 Categories):

THEFT, BATTERY, CRIMINAL DAMAGE, NARCOTICS, ASSAULT, BURGLARY, MOTOR VEHICLE THEFT, ROBBERY, DECEPTIVE PRACTICE, CRIMINAL TRESPASS, WEAPONS VIOLATION, PUBLIC PEACE VIOLATION, OFFENSE INVOLVING CHILDREN, CRIM SEXUAL ASSAULT, SEX OFFENSE, GAMBLING, LIQUOR LAW VIOLATION, ARSON, INTERFERENCE WITH PUBLIC OFFICER, HOMICIDE, KIDNAPPING, INTIMIDATION, STALKING, OBSCENITY, and others

Data Set Explanation

Input Features (22 Variables):

Crime Identification:

- **ID:** Unique crime record identifier
- **Case Number:** Official police case reference number
- **IUCR:** Illinois Uniform Crime Reporting code
- **FBI Code:** FBI crime classification code

Crime Classification:

- **Primary Type:** Main crime category (33 types: THEFT, BATTERY, ASSAULT, etc.)
- **Description:** Detailed crime description and subcategory
- **Location Description:** Specific location type (STREET, RESIDENCE, APARTMENT, etc.)

Temporal Information:

- **Date:** Complete datetime when crime occurred
- **Year:** Extracted year of crime
- **Updated On:** Last update timestamp for the record

Geographic Features:

- **Block:** Anonymized street address block
- **Latitude:** Geographic latitude coordinate (41.6 to 42.0 range)
- **Longitude:** Geographic longitude coordinate (-87.9 to -87.5 range)
- **X Coordinate:** Illinois State Plane coordinate system X
- **Y Coordinate:** Illinois State Plane coordinate system Y
- **Location:** Combined latitude and longitude string

Administrative Boundaries:

- **Beat:** Police beat number (patrol area subdivision)
- **District:** Police district number (1-25)
- **Ward:** City council ward number (1-50)
- **Community Area:** Community area number (1-77)

Crime Status:

- **Arrest:** Boolean flag indicating if arrest was made (True/False)
- **Domestic:** Boolean flag for domestic violence incidents (True/False)

Engineered Features (Created during preprocessing):

- **Hour:** Extracted hour of day (0-23) from datetime
- **Day_of_Week:** Day name (Monday-Sunday)
- **Month:** Month number (1-12)
- **Season:** Season classification (Winter, Spring, Summer, Fall)
- **Is_Weekend:** Boolean flag for weekend crimes
- **Crime_Severity_Score:** Numerical score based on crime type

Expected Results

Technical Deliverables:

- Successfully process and analyze 500,000 crime records sampled from 7.8 million dataset
- Complete MLflow integration with organized experiment tracking for all models
- Deploy fully functional Streamlit Cloud application with interactive visualizations

Step 4: Clustering Analysis Results

Geographic Crime Hotspot Clustering:

- Identify 5-10 distinct crime zones on Chicago map
- K-Means: Creates circular hotspot zones with clear center points for patrol focus
- DBSCAN: Finds naturally formed high-crime areas and removes isolated incidents
- Hierarchical: Shows how zones relate (e.g., South Side breaks into smaller neighborhoods)
- Each cluster shows: total crimes, dominant crime types, and arrest rates
- Generate crime heatmap with color-coded risk zones (red = high risk, yellow = medium, green = low)
- Compare algorithms using silhouette score (target: above 0.5)
- Select best algorithm based on scores and practical usability

Temporal Pattern Clustering:

- Find 3-5 time-based crime patterns (late-night crimes, rush-hour incidents, weekend patterns)
- Identify peak danger times (e.g., 10 PM - 2 AM for violent crimes)
- Show seasonal trends (which months have more crimes)
- Compare weekday vs weekend crime patterns
- Create hourly heatmap showing when crimes happen most

Step 5: Dimensionality Reduction Results

PCA (Principal Component Analysis):

- Compress 22 features down to 2-3 main components capturing 70%+ information
- Identify which features matter most (likely: location, time, crime type)
- Create 2D scatter plot to visualize crime patterns
- Generate scree plot to show how much information each component captures
- Make data easier to understand and visualize

t-SNE Visualization:

- Create beautiful 2D plots where similar crimes cluster together
- Color-code by crime type to see natural groupings
- Color-code by time period to see day vs night patterns
- Compare with PCA results to validate findings
- Interactive plots allowing drill-down into specific crime groups

Project Evaluation Metrics

Technical Performance Evaluation (70%):

- Data preprocessing and sampling methodology (10%)
- Clustering analysis: minimum 3 algorithms with performance comparison (30%)
- Dimensionality reduction: PCA and one visualization technique (20%)
- MLflow integration: experiment tracking for all models (10%

Application Development and Deployment (30%):

- Streamlit application with interactive visualizations and user interface (20%)
- Cloud deployment stability, performance, and accessibility (10%)

Optional Bonus Deliverable (+10%):

- Docker containerization setup
- Complete Dockerfile and docker-compose.yml
- Deployment instructions for containerized environment
- Note: Docker is not mandatory.

Technical Tags

Python, Data Preprocessing, Feature Engineering, Unsupervised Learning, K-Means Clustering, DBSCAN, Hierarchical Clustering, PCA, t-SNE, UMAP, MLflow, Streamlit Cloud, Public Safety, Crime Analytics, Geographic Data Analysis, Temporal Analysis, Data Visualization, Big Data Processing

Timeline

The project should be completed and submitted **within 10 days** from the date it is assigned.

References:

TOPIC	LINK
Project Live Evaluation	Project Live Evaluation
EDA Guide	Exploratory Data Analysis (EDA) G…
Capstone Explanation Guideline	Capstone Explanation Guideline
GitHub Reference	How to Use GitHub.pptx
Project Orientation (English)	
Project Orientation (Tamil)	PatrolIQ_Project Session Recordin…
STREAMLIT RECORDING (English)	Special session for STREAMLIT(11/…
STREAMLIT DOCUMENTATION	Install Streamlit
ML FLOW Tutorial 1	ML FLOW 1

ML FLOW Tutorial 2	ML FLOW 2
MLflow DOCUMENTATION:	Getting Started with MLflow
Project Excellence Series [[Machine learning] (English)	**Project Excellence Series: Guided L..**
Project Excellence Series [[Machine learning] (Tamil)	**Project Excellence Series: Guided L..**
Project Excellence Series [Machine learning-Unsupervised learning] (English)	**Project Excellence Series: Guided L..**
Project Excellence Series [Machine learning-Unsupervised learning] (Tamil)	**Project Excellence Series: Guided L..**
Project Excellence Series [EDA] (English)	**Project Excellence Series: Guided L..**
Project Excellence Series [EDA] (Tamil)	**Project Excellence Series: Guided L..**

PROJECT DOUBT CLARIFICATION SESSION (PROJECT AND CLASS DOUBTS)

About Session: The Project Doubt Clarification Session is a helpful resource for resolving questions and concerns about projects and class topics. It provides support in understanding project requirements, addressing code issues, and clarifying class concepts. The session aims to enhance comprehension and provide guidance to overcome challenges effectively.

Note: Book the slot at least before 12:00 Pm on the same day

Timing: Monday-Saturday (3:30PM to 4:30PM)

Booking link :https://forms.gle/XC553oSbMJ2Gcfug9

LIVE EVALUATION SESSION (CAPSTONE AND FINAL PROJECT)

About Session: The Live Evaluation Session for Capstone and Final Projects allows participants to showcase their projects and receive real-time feedback for improvement. It assesses project quality and provides an opportunity for discussion and evaluation.

Note: This form will Open only on Saturday (after 2 PM) and Sunday on Every Week

Timing:

For DS and AIML
Monday-Saturday (05:30PM to 07:00PM)

Booking link : https://forms.gle/1m2Gsro41fLtZurRA

Created By:	Verified By:	Approved By:
Subhash Govindharaj	Nehlath Harmain	Nehlath Harmain