Project 4

By: Jesus, Carl, Rekha, Kahan, Kuautli

Machine Learning

VIOLENT CRIME DETECTION

Introduction

- The "Violent Crime Detection" project utilizes data-driven analysis to enhance public safety by pinpointing areas prone to criminal activity. Our goal is to support effective security measures and strengthen community protection.
- Machine Learning, with its advanced predictive capabilities, provides a powerful framework for addressing the complexities of crime hotspot identification and prevention.
- This presentation highlights Machine Learning, which focuses on leveraging predictive modeling and data insights to analyze and forecast crime trends in Los Angeles.

Why Microsoft Azure?

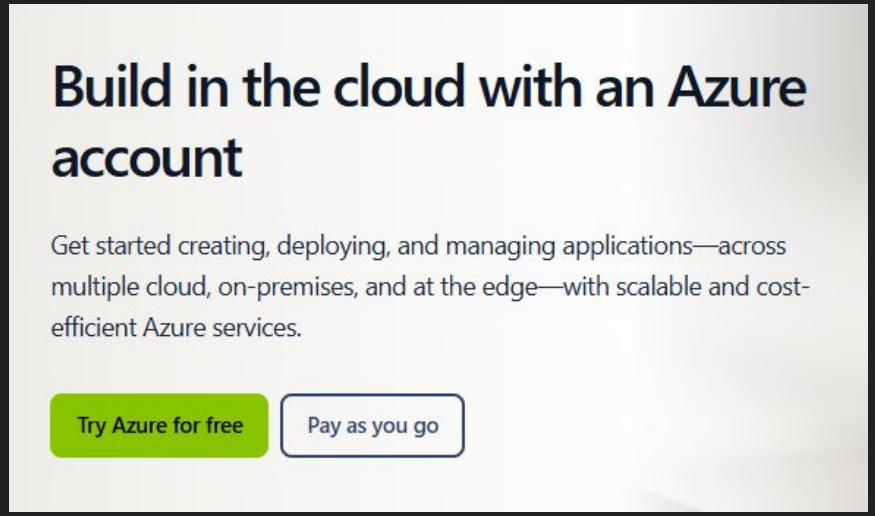
Using Microsoft Azure to host a PostgreSQL server provides several advantages that align with modern cloud-first strategies:

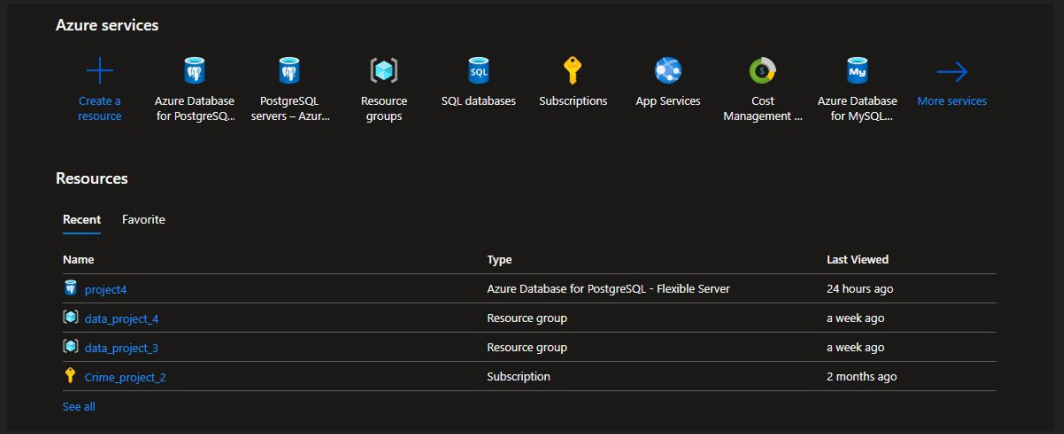
1. **Scalability and Performance:** Azure offers a fully managed PostgreSQL database service that automatically scales to handle high workloads. This ensures performance consistency even during peak usage, without manual intervention.
2. **High Availability and Disaster Recovery:** Azure provides built-in high availability through zone-redundant architecture and automated backups with point-in-time recovery. This reduces the risk of downtime and data loss.
3. **Cost Efficiency:** The flexible pricing model, including pay-as-you-go options, allows businesses to optimize costs based on their needs. Features like burstable performance and reserved capacity offer further savings.
4. **Security and Compliance:** Azure secures databases with features like advanced threat protection, encryption at rest and in transit, and virtual network integration. It also complies with industry standards such as GDPR, HIPAA, and SOC.
5. **Integration with Azure Ecosystem:** Hosting PostgreSQL on Azure allows seamless integration with other Azure services, such as Azure Data Factory for data migration, Azure AI for analytics, and Power BI for visualization.
6. **Managed Service Convenience:** Azure handles infrastructure management, updates, and patching, allowing teams to focus on application development rather than maintenance.

By leveraging these features, businesses can ensure a robust, secure, and scalable database environment while reducing operational overhead.

Setting up the SQL server

- First create an account with Azure if you don't already have one.
- After creating an account you will be taken to you Azure dashboard.

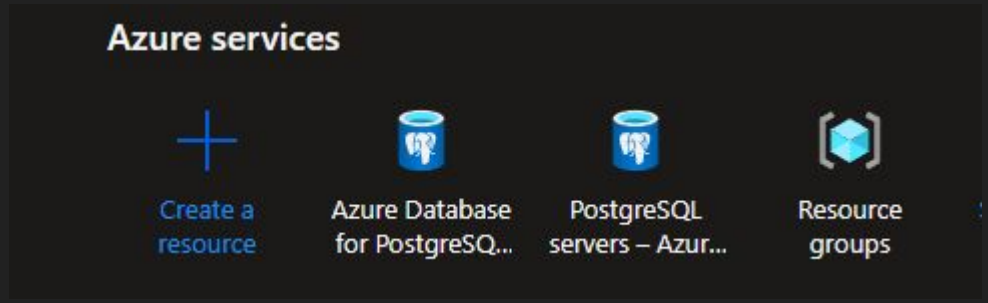


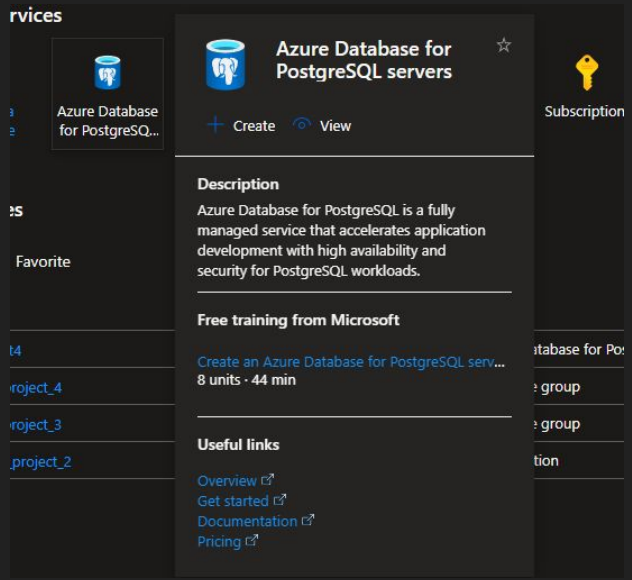


Setting up the SQL server

- Next you will need to create a new instance in the cloud (basically creating your subscription for the server).

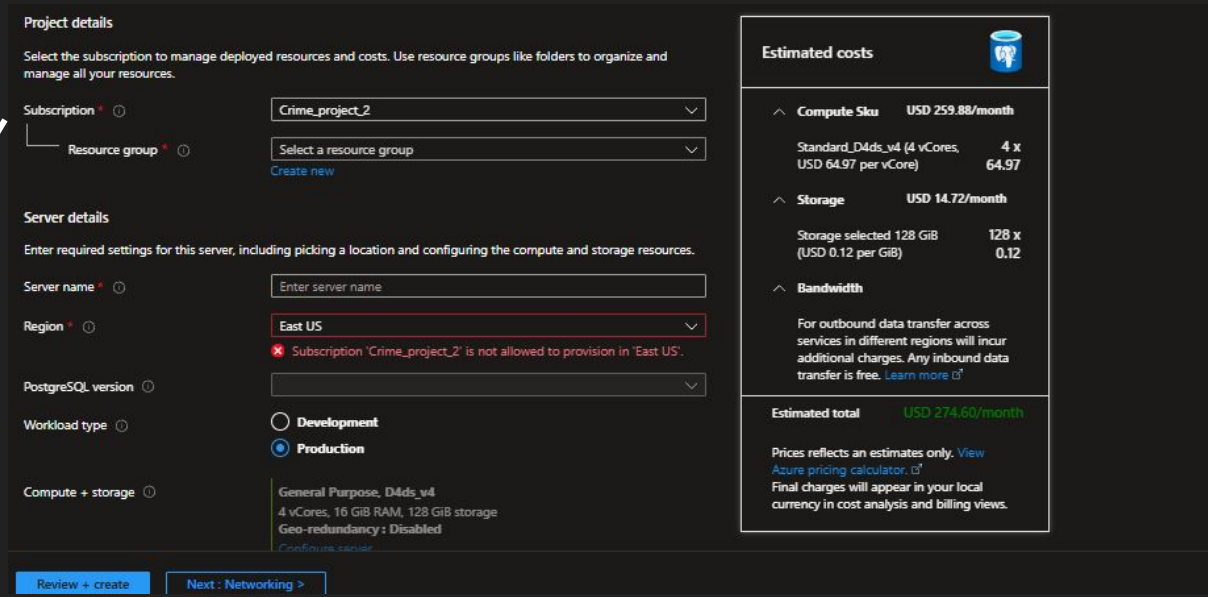
- Choose the type of SQL management service you wish to use.





Setting up the SQL server

- Lastly we just need to finalize set up of the server by selecting different usage options.



Using SQLAlchemy to create tables

```
########## Connecting to the Postgre server in Azure ##########

try:
    # Create the SQLAlchemy engine using the DATABASE_CONFIG from db_config.py
    connection_url = f"postgresql://{DATABASE_CONFIG['user']}:{DATABASE_CONFIG['password']}@{DATABASE_CONFIG['host']}:{DATABASE_CONFIG['port']}
    /{DATABASE_CONFIG['database']}"
    engine = create_engine(connection_url, connect_args={"sslmode": "require"})

    # Execute the query to create the table
    create_table_query = """
    CREATE TABLE IF NOT EXISTS crime_data_2020_to_present (
        DR_NO INT,
        Date_Rptd DATE NOT NULL,
        Date_Occ DATE NOT NULL,
        Time_Occ TIME NOT NULL,
        AREA INT,
        AREA_NAME VARCHAR(50),
        Rpt_Dist_No INT,
        Part_1_2 INT,
        Crm_Cd VARCHAR(5) PRIMARY KEY,
        Crm_Cd_Desc VARCHAR(75),
        VictAge INT,
        VictSex VARCHAR(1),
        Vict_Descent VARCHAR(1),
        Premis_Cd INT,
        Premis_Desc VARCHAR(75),
        Weapon_Use_Cd INT,
        Weapon_Desc VARCHAR(50),
        Status VARCHAR(50),
        Crm_Cd_1 INT,
        Crm_Cd_2 INT,
        Crm_Cd_3 INT,
        Crm_Cd_4 INT,
        LOCATION VARCHAR(50),
        Cross_Street VARCHAR(75),
        address_id NUMERIC
    );
    """

    with engine.connect() as conn:
        conn.execute(text(create_table_query))
        print("Table created successfully.")

    # Optionally, read data into a DataFrame
    query = "SELECT * FROM crime_data_2020_to_present;"  # Example query
    with engine.connect() as conn:
        date_time_df = pd.read_sql(query, conn)
        print(date_time_df)

except Exception as e:
    print("Error while connecting to PostgreSQL:", e)
```

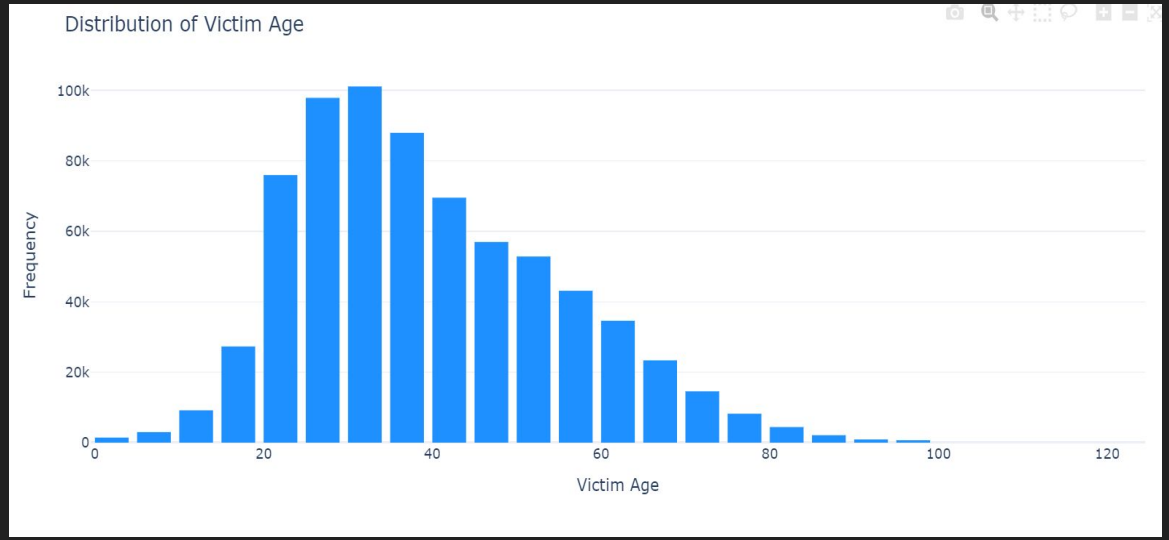
Objectives & Methodology

- **Objective:** The goal is to develop a machine learning model to identify violent crime in Los Angeles. Using LAPD data from 2020 to 2024, the model aims to offer insights that inform targeted security strategies for high-risk areas, promoting a safer community.
- **Data Collection & Methodology:**
  a. **Data Preprocessing:** Cleaning and preparing data for analysis.
  b. **Feature Engineering:** Extracting and selecting relevant attributes to improve model performance.
  c. **Model Development:** Training machine learning models to identify crime-prone areas.
  d. **Evaluation:** Testing the model for accuracy and reliability.

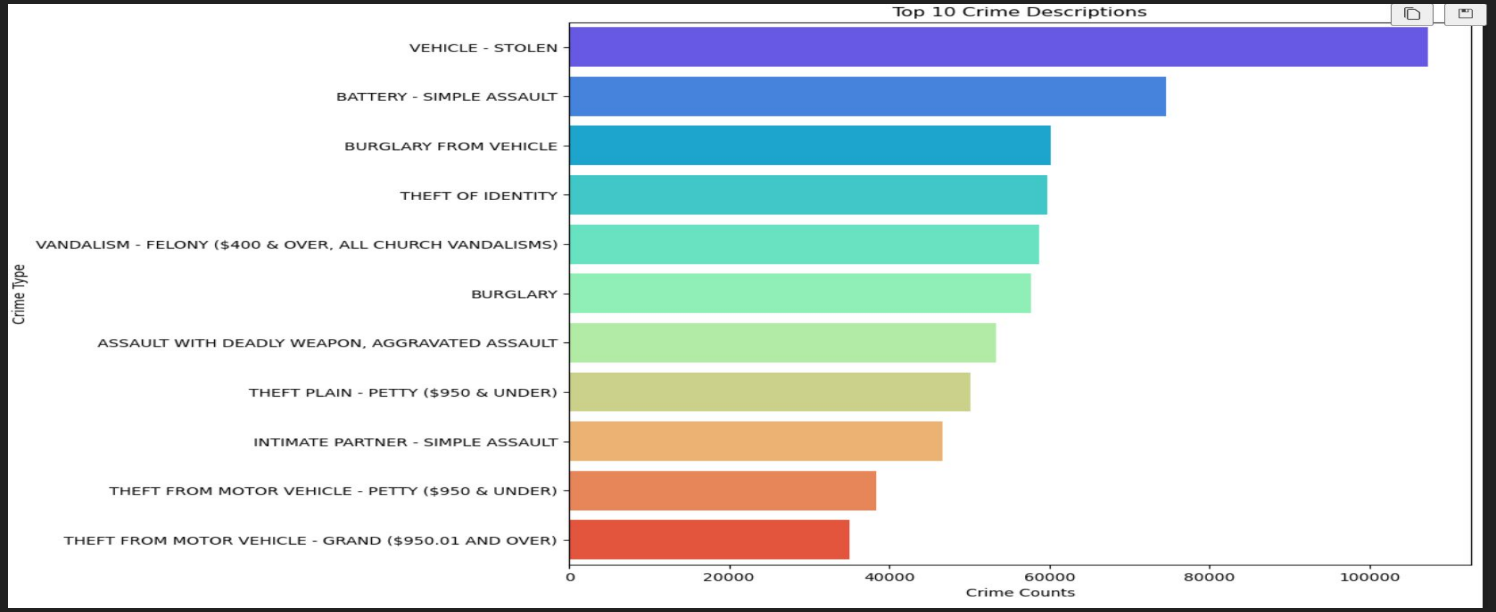
This structured workflow ensures the delivery of accurate, reliable insights to enhance community safety.

Distribution of counts by Victim Age

The histogram displayed illustrates the distribution of crime victims by age, a key component of our Exploratory Data Analysis (EDA). This insight indicates that the majority of crime victims fall within the 30 to 34 age range.

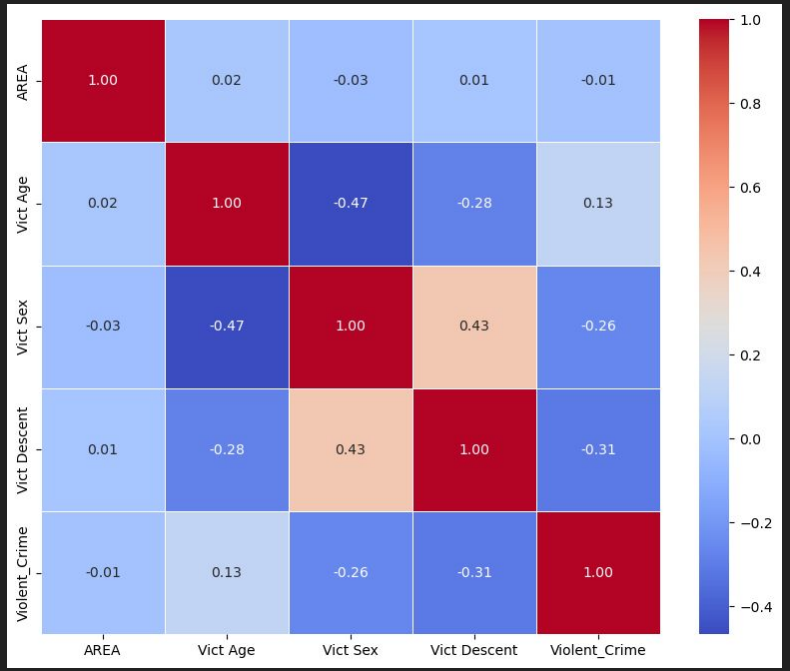


Distribution of Top 10 Crime Descriptions



Logistic Regression

- Dataset contains 900000 records
- 28 data columns captured per crime reported
  - AREA,Vict Age,Vict Sex,Vict Descent
  - Violent Crime (binary classification column)
- Violent Crime is what we are attempting to predict
- Multiple iterations in feature selection and data used
  - `X = ['Vict Age', 'Vict Sex', 'Vict Descent']`
  - `y = ['Violent_Crime']`



Model Prediction

All Years / Class Imbalance

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.71	0.93	0.80	69415
1	0.44	0.14	0.21	30585
accuracy			0.68	100000
macro avg	0.58	0.53	0.51	100000
weighted avg	0.63	0.68	0.62	100000

All Years / Balanced

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.69	0.63	0.66	53311
1	0.62	0.68	0.65	46689
accuracy			0.65	100000
macro avg	0.65	0.65	0.65	100000
weighted avg	0.66	0.65	0.65	100000

One Year / Balanced

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.68	0.66	0.67	10688
1	0.62	0.64	0.63	9312
accuracy			0.65	20000
macro avg	0.65	0.65	0.65	20000
weighted avg	0.65	0.65	0.65	20000

All Years One Area / Balanced

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.63	0.52	0.57	3985
1	0.59	0.69	0.64	4015
accuracy			0.61	8000
macro avg	0.61	0.61	0.61	8000
weighted avg	0.61	0.61	0.61	8000

Model Results

All Years / Balanced

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.69	0.63	0.66	53311
1	0.62	0.68	0.65	46689
accuracy			0.65	100000
macro avg	0.65	0.65	0.65	100000
weighted avg	0.66	0.65	0.65	100000

Precision (Positive Predictive Value):

- **Class 0 (0.69)**: 69% of those predictions are correct.
- **Class 1 (0.62)**: 62% of those predictions are correct.

Recall (Sensitivity/True Positive Rate):

- **Class 0 (0.63)**: The model correctly identifies 63% of all actual Class 0 instances.
- **Class 1 (0.68)**: The model correctly identifies 68% of all actual Class 1 instances.

Accuracy (0.65):

- The model correctly predicts 65% of all instances.

Conclusion

1. **Developed a machine learning mode**l to predict violent crime trends in Los Angeles using LAPD data from 2020–2024.
2. **Performed comprehensive exploratory data analysis (EDA)** to uncover patterns and relationships in the data, guiding feature selection and model design.
3. **Built and evaluated models** with balanced datasets, achieving 65% accuracy, and highlighted precision and recall metrics for crime prediction.
4. **Utilized Microsoft Azure's PostgreSQL** server for scalable, secure, and efficient data management.
5. **Visualized** key findings and predictions to provide actionable insights for improving public safety.


THANK YOU