Baltimore Crime Data Analytics

DECEMBER 17th 2024

Presented by:
Yatharth Kumar
Stuti Upadhyay
Ajay Kumar
Soumya Bhate
Arshdeep Singh
Mitaali Patel

Baltimore Crime Data Analytics

Project Background

Baltimore, the “charm city” and the biggest metropolitan area in the state of Maryland, a city filled with rich history and culture inherited from its past , has now been facing a persistent challenge of combating crime rate, resource allocation of law enforcement and public safety. Recent statistics indicate that Baltimore’s violent crime rate is significantly higher than the national average, with over 17,000 reported incidents of violent crimes annually. Property crimes such as burglary, larceny, and motor vehicle theft also present significant challenges, accounting for thousands of reported cases each year. These figures underscore the need for more strategic interventions and resource optimization.

With this project we aim to harness our skills in data management and analytics to delve deep into these issues. Our approach here is to analyze crime patterns, arrest trends, and the distribution of police stations and provide appropriate actionable insights. The ultimate goal that we plan on achieving with this project is not just to focus and discuss the current state of crime within the city but also to bridge the gap between law enforcement agencies and the communities they serve by promoting transparency and collaboration.

Data Overview

The dataset that was utilized for this project consists of;

1. Arrests:This dataset provides detailed records of arrests made within Baltimore. Key details include:
    a. Dates and timestamps of arrests.
    b. The type of offense or charge.
    c. The location of the arrest.
    d. Demographic information of the individuals involved, such as age, gender, and race.

2. Police Stations: This dataset contains information about police stations in Baltimore, including:
    a. Geographical locations (latitude, longitude).
    b. Operational details such as station names, addresses, and neighborhood associations.
    c. Information about the station commander and contact numbers.

3. Incidents: The incident dataset records various crime incidents across the city. Key attributes include:
    a. A unique identifier for each crime incident.
    b. The type of crime committed (e.g., theft, assault, burglary).
    c. Geographic details such as district, neighborhood, and coordinates (latitude, longitude).

d. Links to the related location where the incident occurred.

4. Locations: This dataset captures granular location details tied to both incidents and arrests. Key fields include:
    a. Unique location identifiers.
    b. Street tags, parcel identifiers (PRCL_PIN), and complete addresses.
    c. Geographic coordinates (X and Y) for spatial mapping.
    d. Neighborhood associations and other contextual information.

Key Project Objectives

1. Identification of Crime Hotspots – Understand and analyze our geographical data in order to pinpoint the specific areas with high crime rates that also result in a higher frequency of arrests.
2. Key Details for Resource Allocation – Evaluate the distribution of police stations in relation to our identified crime hotspot's locations.
3. Trends about the Arrest Record - Examine trends within the arrest data over the period of time as well as understand and examine any seasonal or demographic patterns that can be observed within it.
4. Insights about the Community - Provide certain insights that might be able to guide public awareness campaigns and community engagement initiatives.

Target Audience / Project Beneficiaries

1. Law Enforcement Agencies – Help the local law enforcement agencies in enhancing the operational efficiency within their department and aid into the strategic resource deployment.
2. Policy Makers – Provide viable insight that can be helpful in making more informed decisions as well as strategies on local crime prevention policies and could come in handy during the budget allocation.
3. Organizations based within the Community - To help local nonprofits in designing targeted intervention strategies and safety programs that are tailored to help and aid the affected communities.
4. Researchers and Analysts – This can greatly contribute towards the further exploration related to crime-related factors and can lead to a great amount of innovative solutions.
5. Private Investigation Organizations – This can certainly support independent investigation agencies and develop strategies where specific cases are being addressed. This can also lead to an increased awareness about crime trends and foster collaborative efforts to improve safety.
6. General Public – At last, this will help in increasing awareness among the wider population within the city about all the crime trends and foster into a lot of collaborative efforts that will aid greatly in improving the safety.

Methodology

The methodology for this project is structured in four major components with each one of them focusing on a certain specific aspect such as data processing, transformation and visualization. Our approach ensures a seamless flow from data collection to actionable insights, leveraging Azure services for efficient data handling and Power BI for visualizations.

Step 1: Data Retrieval from Data.gov API

1. API Integration : We begin by connecting to Data.gov API using an authenticated key to retrieve the relevant datasets related to crime incidents, arrests, location and police stations within Baltimore.
2. Data Selection : The datasets are carefully selected to ensure that they cover all necessary variables required for the analysis, including the geographical data, demographic information, incident details and arrest records.

Step 2: Azure Environment Setup

1. Uploading Flat Files : Once data is retrieved, flat files are uploaded to Azure Blob Storage using a Python script that automates the process and performs transformations, ensuring that any new data added to Data.gov is immediately available for processing.

2. ETL Pipeline Creation in Azure Data Factory : Azure Data Factory (ADF) is used to create an ETL pipeline. The pipeline is configured to extract data from Azure Blob Storage, and load it into an Azure SQL Database for further analysis.
3. Automated Data Refresh : The process is automated, allowing for seamless data updates and ensuring the pipeline remains operational as new datasets become available.

Step 3: Database Creation and Configuration on Azure SQL Database

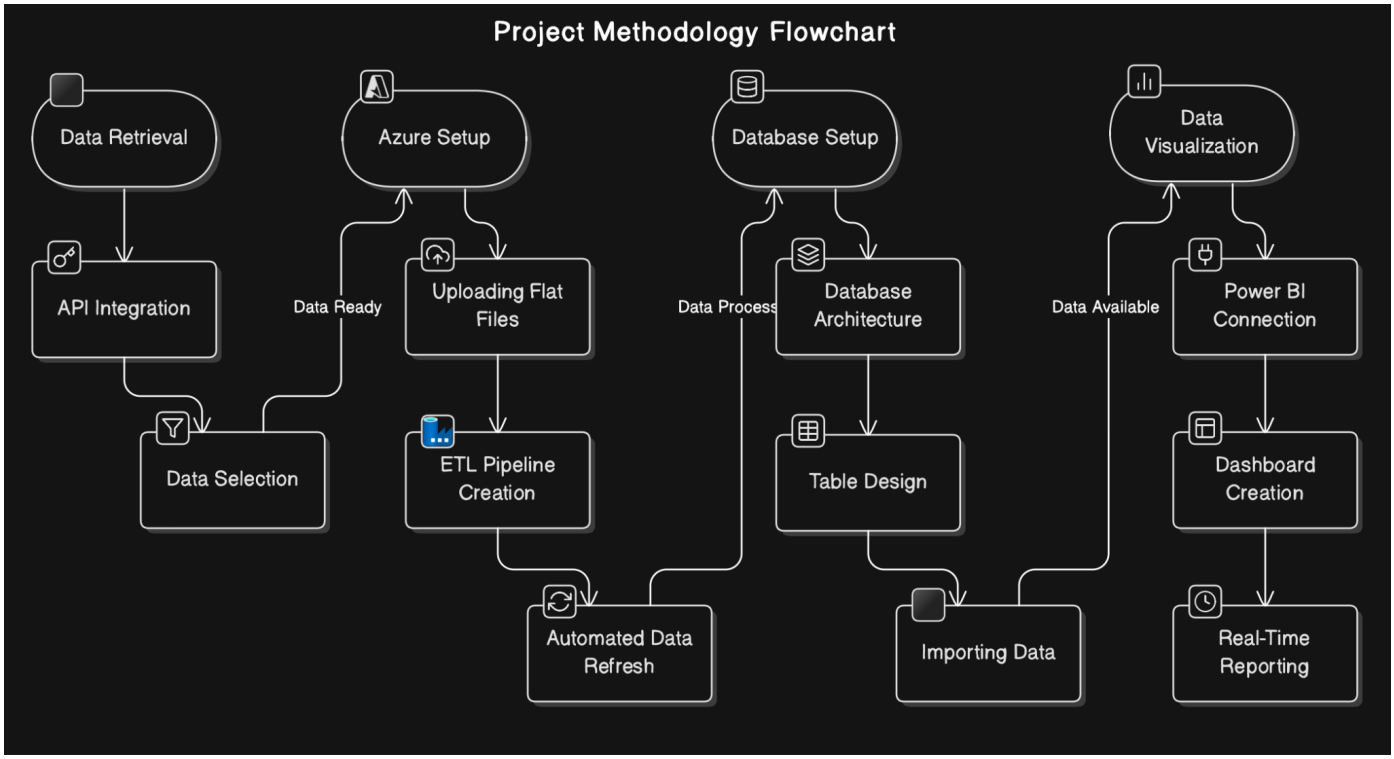
1. Database Architecture : The database architecture is designed to efficiently store and organize the various datasets. This includes the four major tables (PoliceStation, Location, Incident, and Arrest) that form the foundation for structured data storage.
2. Table Design : Each table is designed with key fields that link related information (e.g., linking incidents to locations and arrests to incidents), ensuring a relational structure for easy querying.
3. Importing Data : Using SQL Server Management Studio (SSMS) or Azure Data Studio, the cleaned data is imported from Azure Blob Storage to the Azure SQL Database. This step ensures that the data is ready for querying and analysis.

Step 4: Data Visualization in Power BI

1. Power BI Connection : Power BI is connected to Azure SQL Database to retrieve real-time data for visualization.
2. Dashboard Creation : Interactive dashboards are created in Power BI to visualize key insights, such as crime hotspots, arrest trends, and police station distribution.

Visualizations include maps, bar charts, time series, and heatmaps, which make the data easily interpretable for stakeholders.

3. Real-Time Reporting: The dashboards are designed to provide real-time updates based on the latest data available in the Azure SQL Database, allowing for dynamic reporting and decision-making.



By following this methodology, the project ensures that data is effectively collected, transformed, and visualized, providing actionable insights for law enforcement, policymakers, and the general public. The combination of Azure services and Power BI ensures a scalable, efficient, and user-friendly solution for crime data analytics.

Comprehensive View into the Database

There are four major tables for use in order to correctly structure our available data comprehensively:

	COLUMN_NAME	DATA_TYPE	CHARACTER_MAXIMUM_LENGTH	IS_NULLABLE
1	ArrestNumber	int	NULL	NO
2	Age	int	NULL	YES
3	Gender	char	1	YES
4	Race	char	1	YES
5	ArrestDateTime	datetime	NULL	YES
6	Charge	varchar	10	YES
7	ChargeDescription	varchar	255	YES

	COLUMN_NAME	DATA_TYPE	CHARACTER_MAXIMUM_LENGTH	IS_NULLABLE
1	IncidentNumber	varchar	20	YES
2	IncidentOffence	varchar	255	YES
3	IncidentLocation	varchar	100	YES
4	ArrestNumber	int	NULL	YES

	COLUMN_NAME	DATA_TYPE	CHARACTER_MAXIMUM_LENGTH	IS_NULLABLE
1	Location_Addr	varchar	50	YES
2	ArrestNumber	int	NULL	YES
3	Latitude	real	NULL	YES
4	Longitude	real	NULL	YES
5	District	varchar	20	YES

	COLUMN_NAME	DATA_TYPE	CHARACTER_MAXIMUM_LENGTH	IS_NULLABLE
1	GIS_ID	varchar	5	NO
2	SRCID_T	int	NULL	YES
3	subtype	varchar	30	YES
4	Address	varchar	255	YES
5	City	varchar	100	YES
6	State	varchar	2	YES
7	Zipcode	varchar	10	YES
8	X_Coord	real	NULL	YES
9	Y_Coord	real	NULL	YES
10	Name	varchar	100	YES
11	Neighborhood	varchar	100	YES
12	Commander	varchar	100	YES
13	Phone	varchar	15	YES
14	Edit_date	datetime	NULL	YES

1. **Police Station Table**

    - Represents information about police stations.
    - Key fields include:
        - GIS_ID: A unique identifier for the police station.
        - FType: The type of police station
        - Address: Station's physical address.
        - City, State, Zipcode: Location details.
        - X_Coord, Y_Coord: Geographic coordinates.
        - Name: Station name.
        - Neighborhood: Neighborhood location.
        - Commander: Name of the station commander.
        - Phone: Contact number.
        - Edit_Date: Timestamp of the last edit to the record.

2. **Location Table**

    - Stores detailed location information.
    - Key fields include:
        - LocationID: Unique identifier for each location.
        - Street_Tag: Street name or identifier.
        - PRCL_PIN: Potentially a parcel identification number.
        - City, State, Zipcode: Location details.

3. **Incident Table**

   - Tracks crime incidents.
   - Key fields include:
     - IncidentNumber: Unique identifier for each incident.
     - IncidentOffence: Type of crime.
     - IncidentLocation: Foreign key linking to the Location table.
     - District: Geographic district.
     - Neighborhood: Neighborhood location.
     - Latitude, Longitude: Precise geographic coordinates.
     - GeoLocation: Additional geographic details.

4. **Arrest Table**
   - Records details of arrests.
   - Key fields include:
     - ArrestNumber: Unique identifier for each arrest.
     - IncidentNumber: Links to the specific incident.
     - ArrestLocation: Foreign key linking to the Location table.
     - Age, Gender, Race: Demographic information of the arrested individual.
     - ArrestDateTime: Timestamp of the arrest.
     - Charge, ChargeDescription: Details of the legal charge.

Furthermore, following is our visualized schema in form of an Entity-Relationship Diagram (ERD);


ERD for Baltimore's Crime Database
police_station
id string pk
gis_id string
srcid_t string
subtype string
address string
city string
state string
zipcode string
x_coord float
y_coord float
name string
neighborhood string
commander string
phone string
edit_date timestamp
arrest
id string pk
age int
gender string
race string
arrestDateTime timestamp
charge string
chargeDescription string
arrestNumber string
incident
id string pk
incidentNumber string
incidentOffence string
incidentLocation string
arrestNumber string fk
1
location
id string pk
locationAddr string
arrestNumber string fk
1
latitude float
longitude float
district string

Cost Analysis

The cost analysis of this project provides a detailed view of the expenses incurred during the development and deployment of the project on the Azure platform. Below are the key findings:

1. Overall Costs :
    a. Total Cost Incurred: $1.68.
2. Cost Breakdown by Service : The costs are distributed across several Azure services, with the following contributions;
    a. Azure Data Factory v2: $1.17 (majority of the cost).
    b. SQL Database: $0.46.
    c. Storage: $0.03.
    d. Bandwidth: $0.02.
3. Cost Breakdown by Location : The expenses are split across two Azure locations;
    a. US East 2: $1.22 (majority of the cost).
    b. US East: $0.46.
4. Subscription and Resource Group Details :
    a. Subscription Used: Azure Subscription 1.
    b. Resource Group Name: "crimeanalytics".

Azure Subscription Insights :

1. Spending Rate: Currently at $1.68.
2. Top Products by Resource Usage:
    a. Azure Data Factory v2.
    b. SQL Databases.

c. Storage Accounts.
d. Bandwidth Usage.

3. Top Free Services Used:
    a. Networking.
    b. Data Transfer.
    c. Storage.
    d. Azure Cosmos DB.

Inferences

Incident and Arrest Trends:

1. Between 2010 and 2024, significant insights were observed across different age groups.
    a. Teenager Group (Age 13-19):
        i. Total arrests: 27.75k.
        ii. Gender Distribution: 88% were men, and 12% were women.
        iii. Race Distribution: Black (93%), White (5%), Unknown/Undisclosed (1.5%). Asian and Native American arrests were less than 1%.
        iv. Top Offenses: Narcotics (43%), Assault (6.6%), and Armed Person (4.3%).
    b. Young Adult Group (Age 20-30) - Year 2023:
        i. Total arrests: 5,006.
        ii. Gender Distribution: 77.5% were men, and 22.5% were women.
        iii. Race Distribution: Black (86.7%), White (6.75%), Unknown/Undisclosed (6%). Asian and Native American arrests were less than 1%.

iv. Top Offenses: Common Assault (21.7%), Assault Cut (17.3%), and Murder (13.04%).

Offense Types:

1. Major offenses include narcotics and assault for teenagers, whereas adults showed a significant number of violent offenses such as common assault and murder.
2. Drug-related offenses remain a key concern across all demographics.

Age and Gender Distribution:

1. Teenagers and young adults collectively make up a large share of arrests.
2. Gender disparities remain significant, with males accounting for over 77% of arrests in both groups.

Conclusions

The project provided valuable insights into the city's crime patterns and arrest trends, with a focus on optimizing law enforcement resource allocation. The key findings include:

1) Crime Trends:
    a) Teenagers aged 13-19 face higher rates of narcotics-related arrests, highlighting the need for preventive programs and drug rehabilitation efforts.
    b) Young adults aged 20-30 show a rising trend of violent crimes such as common assault and murder, which necessitates focused law enforcement strategies and targeted community interventions.
2) Racial Disparities:

a) Black individuals are disproportionately affected, comprising over 85% of arrests across both age groups. Policymakers must investigate systemic factors contributing to these disparities and create equitable prevention strategies.

3) Gender Distribution:
    a) Males consistently account for the majority of arrests, underlining the need for male-focused intervention programs, especially in youth demographics.
4) Crime Hotspots and Resource Allocation:
    a) Hotspots identified on maps point to high-crime neighborhoods requiring enhanced law enforcement, social services, and community engagement.
5) Key Offenses:
    a) Narcotics remain the top arrest category for teenagers, suggesting that addressing drug-related issues could have a significant impact on overall crime reduction.
    b) Violent crimes in the 20-30 age group, such as common assault and murder, demand immediate attention from law enforcement agencies.

Additionally, the use of Power BI to visualize these trends allowed for interactive, real-time data exploration, which can be instrumental for decision-makers in law enforcement, policy-making, and community engagement.

Furthermore, the cost analysis reflects that the majority of expenses stem from Azure Data Factory v2 and the SQL Database, primarily in the US East 2 region. The overall cost incurred so far is minimal, making the project highly cost-efficient for its scope and objectives. However, implementing budget tracking and enabling Azure Defender for added security could further optimize the resource management and safety of the project.

Project Risks :

1. One of the major technical risks with this happens in case if there are any new entries in the dataset after being updated by Data.gov it can lead to a probability of having one or more records that might not have the same data type and may not match our pipeline which can then lead to the crashing of the entire pipeline every time it is triggered. The data.gov API key should be secured and must be run by trusted personnel.
2. One of the major ethical issues might present itself in the form of systematic injustice. There is always a chance of ethically malicious individuals who can use the inferences derived from this to profile individuals within certain communities which can lead to more socially biased solutions by policy makers within the communities.