

# **Exploratory data analysis on a dataset of Crimes in India using Azure Databricks**

The examination of India's crime dataset provides officials with important information. My EDA explores patterns and trends, enabling researchers to contribute to formulating strategic choices. This blog summarizes my research and highlights important discoveries that advance knowledge of the nation's criminal activity.

## **Technologies used:**

Azure storage, Azure data factory, Azure data bricks, and key vault.

## **About dataset:**

The dataset consists of the total criminal activities in India on state wise, city wise, and total number of complaints.

## **Datasets:**

1. City.csv
2. State.csv
3. Complaints.csv

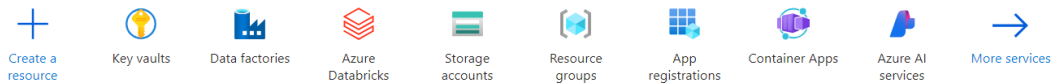
## **Overview of the project:**

The project leveraged Azure's cloud ecosystem to analyze a valuable dataset. First, the data was uploaded to GitHub for version control and then transferred to Azure Data Storage for secure, scalable storage. To unlock hidden insights, Azure Databricks seamlessly extracted the data for analysis. This tight integration with Azure services streamlined data processing and empowered informed decision-making.

This emphasizes the project's key points: cloud-based data management, analysis via Databricks, and efficient decision-making through integration.






Below are all the resources I have created:

## Azure services



## Resources

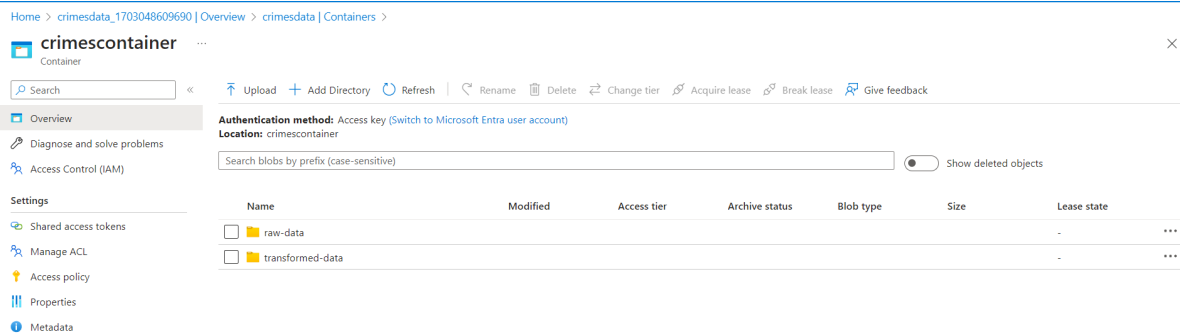
Recent Favorite

Name	Type	Last Viewed
 crimeskey	Key vault	a few seconds ago
 crimes-rg	Resource group	a few seconds ago
 crimesdf	Data factory (V2)	2 minutes ago
 crimesdbw	Azure Databricks Service	4 minutes ago
 crimesdata	Storage account	6 minutes ago

In the storage account, I created two folders - raw-data and transformed-data. The raw-data folder will contain the files that will be fetched from the GitHub account using Azure Data Factory.

GitHub account link:

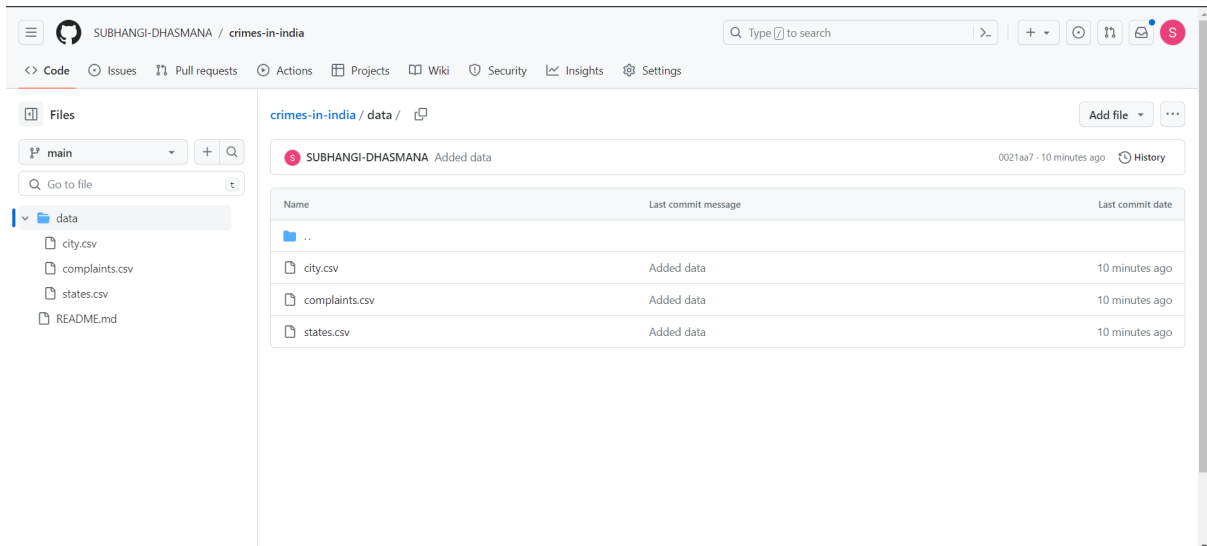
<https://github.com/SUBHANGI-DHASMANA/crimes-in-india/tree/main/data>



I have stored the datasets in CSV form which is downloaded from Kaggle:

<https://www.kaggle.com/dekomorisanae09/criminal-activities-in-india>

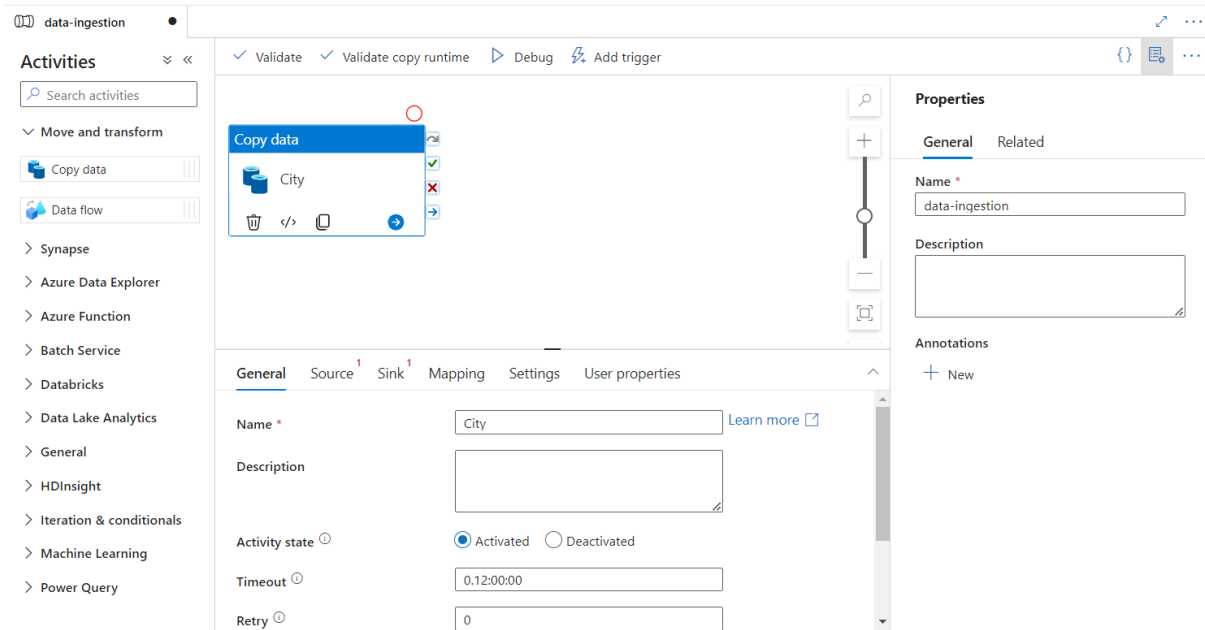
Azure Data Factory acts as a precision-engineered conduit for data, ensuring its smooth and efficient passage to Data Lake Gen 2. Acting as the heart of the system, it expertly directs raw data links to its source pipeline. From there, it orchestrates a seamless flow, guiding the data to its final destination within the "raw-data" folder of Data Lake Gen 2. This meticulous orchestration not only guarantees the integrity of the transfer but also establishes a well-structured foundation primed for subsequent analysis, setting the stage for meaningful insights to emerge.



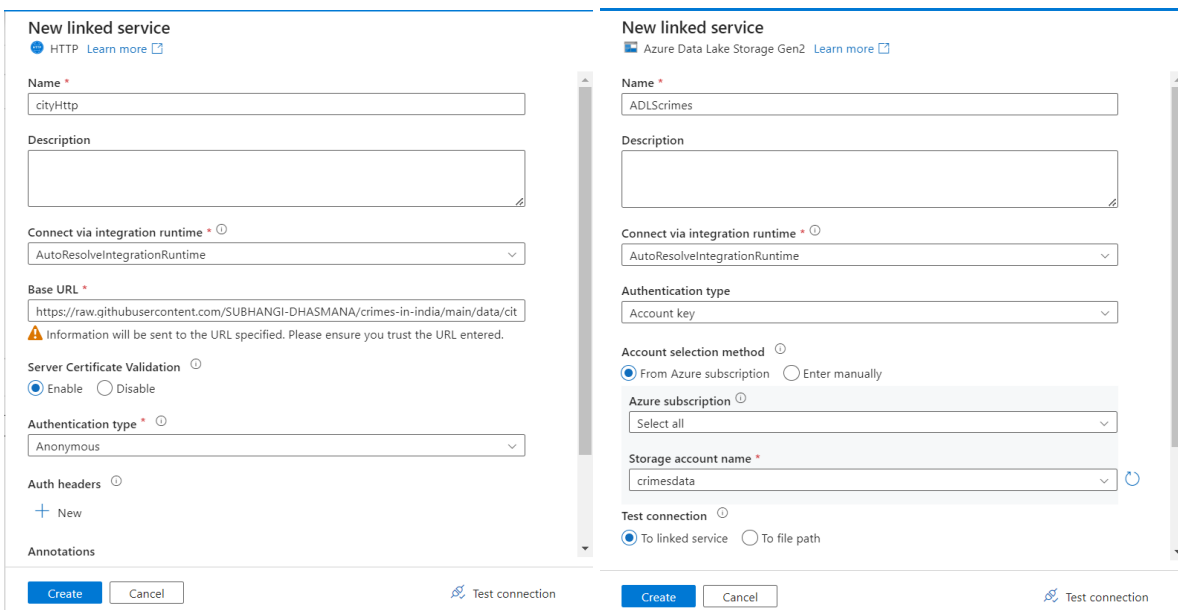
```
S.No, City, 2019, 2020, 2021, Population, RateofCognizableCrimes, ChargesSheetingRate
1, Agra, 6510, 6285, 5665, 17.5, 324.5, 58.0
2, Allahabad, 5621, 5455, 4130, 12.2, 339.4, 68.2
3, Amritsar, 2589, 3178, 3349, 11.8, 282.9, 54.8
4, Asansol, 4244, 4335, 4864, 12.4, 391.3, 95.8
5, Aurangabad, 5636, 6248, 7366, 11.9, 619.5, 82.6
6, Bhopal, 15367, 18329, 18831, 18.8, 1000.1, 88.3
7, Chandigarh City, 2819, 2583, 2401, 10.3, 234.0, 67.7
8, Dhanbad, 2102, 2801, 2508, 11.9, 209.9, 36.0
9, Durg-Bhilainagar, 5334, 5199, 5454, 10.6, 512.6, 85.1
```

```
raw.githubusercontent.com/SUBHANGI-DHASMANA/crimes-in-india/main/data/city.csv
```

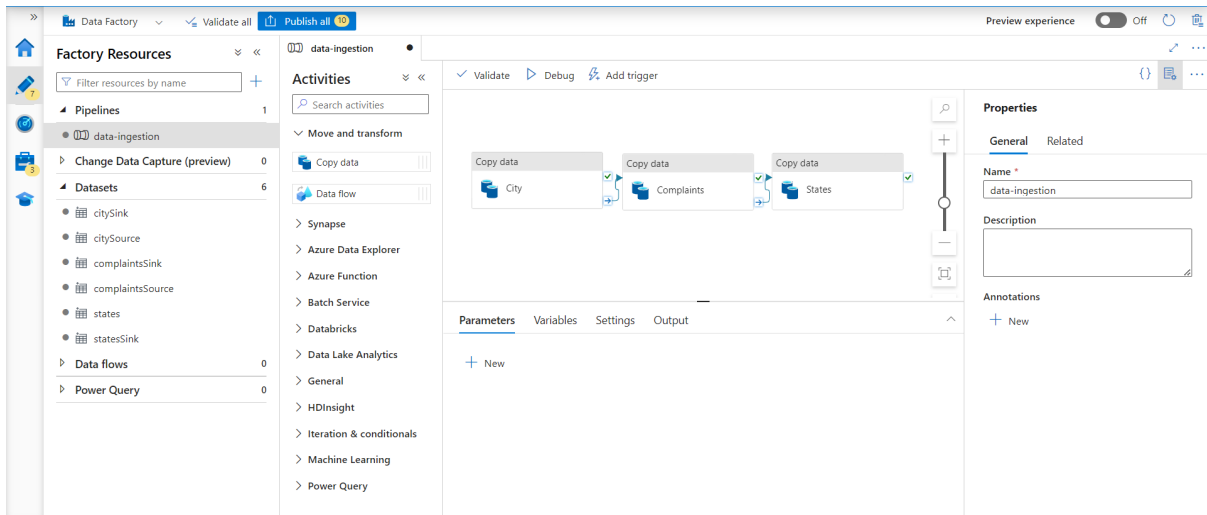
Within the Azure Data Factory, a dedicated "Data Ingestion" pipeline stands as a testament to efficiency and clarity. At its core, the "Copy Data" block, strategically positioned within the Move and Transform column, spearheads a streamlined transfer process. This meticulously designed pipeline unfolds as a visual symphony, its components harmoniously aligned to ensure a flawless influx of data. The intuitive visual representation serves as a guide, ensuring clarity and confidence in the pipeline's operations, fostering a seamless and efficient journey for the data it shepherds.



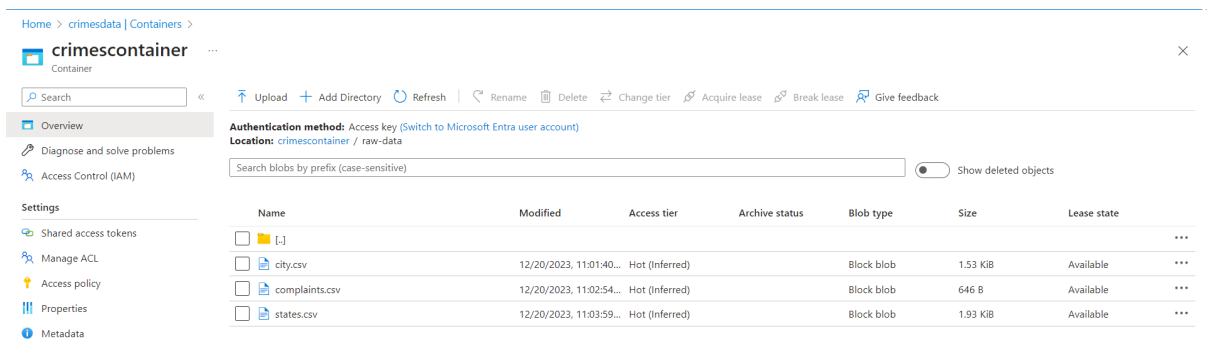
HTTP was selected as the source for the CSV dataset. A new link crafted with a base URL (<https://github.com/SUBHANGI-DHASMANA/crimes-in-india/blob/main/data/city.csv>) retrieves the city.csv dataset. In the sink, Azure Data Lake Gen 2 efficiently loads the data into Azure Data Storage, completing a seamless data transfer.



Similarly, we can create the “Copy Data” block for the rest of the dataset with sink location: “crimescontainer/raw-data/<name of the file>.csv”.



In the above image, the pipeline structure is highlighted by "Copy Data" blocks for the city, complaints, and state. Then we will click on validate all to validate our pipeline. Validation ensures that the configuration is error-free. Following successful validation, debugging and publishing are performed. After publishing, if we return to the raw-data folder it confirms the successful loading of all three datasets, completing a solid data pipeline.



The data ingestion part is completed, and now comes the main portion of the project which is launching the Azure Databricks to start our project.

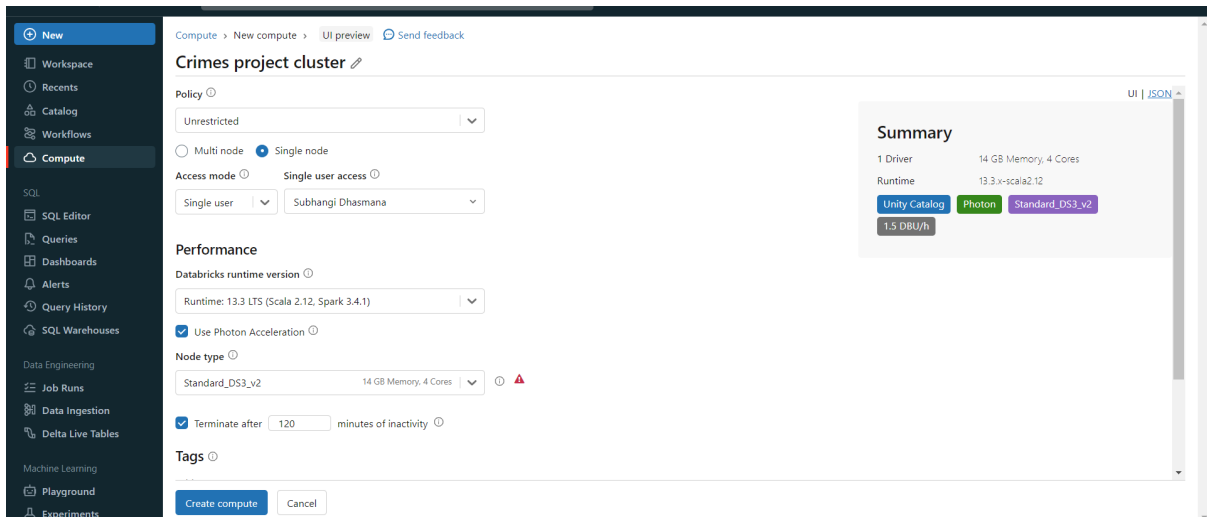
## What is Azure Databricks?

Azure Databricks is a big data analytics and machine learning platform hosted in the cloud. It combines the flexibility and scalability of Azure cloud services with Apache Spark, a powerful open-source data processing engine. It enables data engineers, data scientists, and analysts to work collaboratively and efficiently in a unified environment to perform data processing, data exploration, and machine learning tasks.

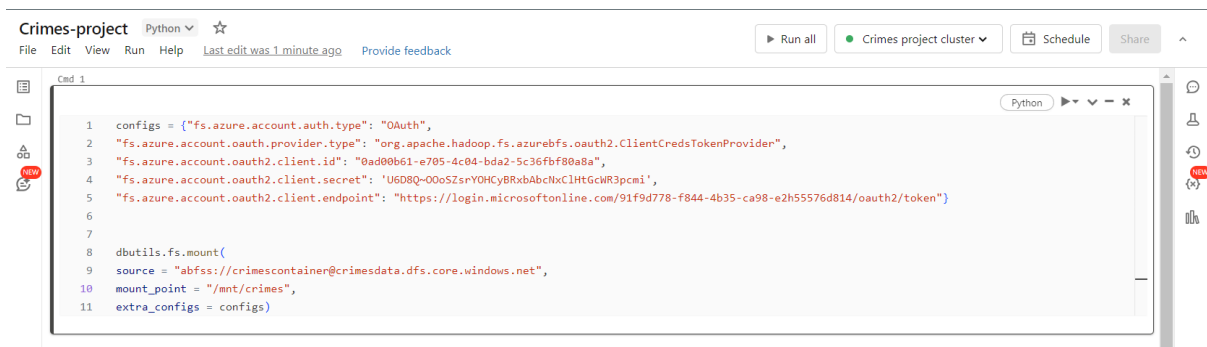
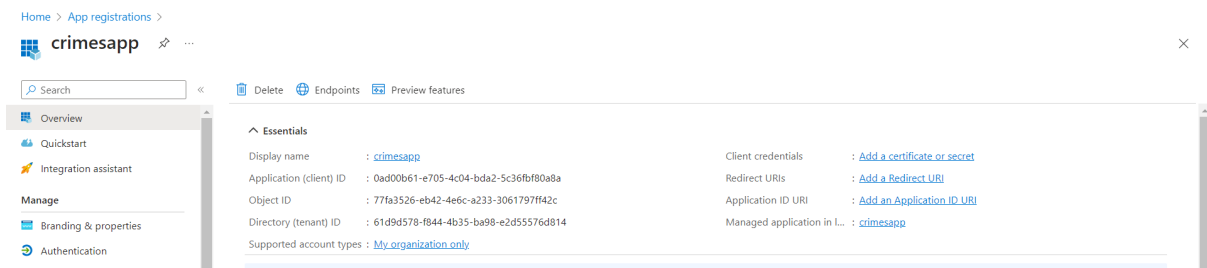
Creating a cluster in Azure Databricks is required to efficiently execute distributed data processing tasks. Clusters are made up of multiple computing nodes that collaborate to handle large amounts of data and complex computations. By forming a cluster, you can take advantage of the power of parallel processing, allowing for faster data analysis, exploration,

and machine learning tasks. Clusters can be scaled up or down based on workload demands, allowing resources and costs to be optimized.

First, we will click on compute to create a cluster in data bricks. I have used single-user access mode with runtime 13.3 LTS and Standard\_D53\_v2 node type.



The "crimesapp" app registration was created to allow Azure access to the dataset in the "crimesdata" storage account. Client and tenant IDs were copied and securely stored in a notepad for future use. A secret key, critical for data retrieval, is generated and discreetly saved. Protecting the client ID and secret key is critical for ensuring secure and controlled access to Azure data storage.



Azure Databricks' scope feature was used to secure credentials. Key Vault is important because secrets are generated and stored there for added security. By integrating with Key

Vault, sensitive information is kept secure, and Databricks notebooks can access credentials without directly exposing them.

```
Cmd 1
1 # dbutils.secrets.list(scope="dbscope")
Command took 0.07 seconds -- by subhangi788@gmail.com at 12/20/2023, 1:49:15 PM on Crimes project cluster

Cmd 2
1 import json
2
3 client_id_secret = dbutils.secrets.get(scope="dbscope", key="clients-id")
4 client_secret_secret = dbutils.secrets.get(scope="dbscope", key="clients-secret")
5
6 configs = {
7     "fs.azure.account.auth.type": "OAuth",
8     "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
9     "fs.azure.account.oauth2.client.id": client_id_secret,
10    "fs.azure.account.oauth2.client.secret": client_secret_secret,
11    "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/61d9d578-f844-4b35-ba98-e2d55576d814/oauth2/token"
12 }
13
14 dbutils.fs.mount(
15     source = "abfs://crimescontainer@crimesdata.dfs.core.windows.net",
16     mount_point = "/mnt/crimes",
17     extra_configs = configs)
Command took 0.12 seconds -- by subhangi788@gmail.com at 12/20/2023, 1:49:50 PM on Crimes project cluster
```

```
1 %fs
2 ls "/mnt/crimes"
```

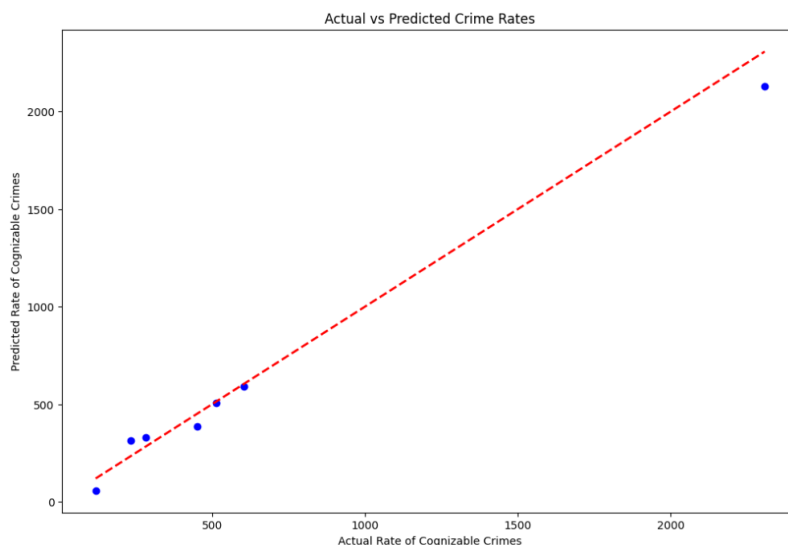
path	name	size	modificationTime
dbfs/mnt/crimes/raw-data/	raw-data/	0	1703048744000
dbfs/mnt/crimes/transformed-data/	transformed-data/	0	1703048756000

2 rows | 6.54 seconds runtime | Refreshed now

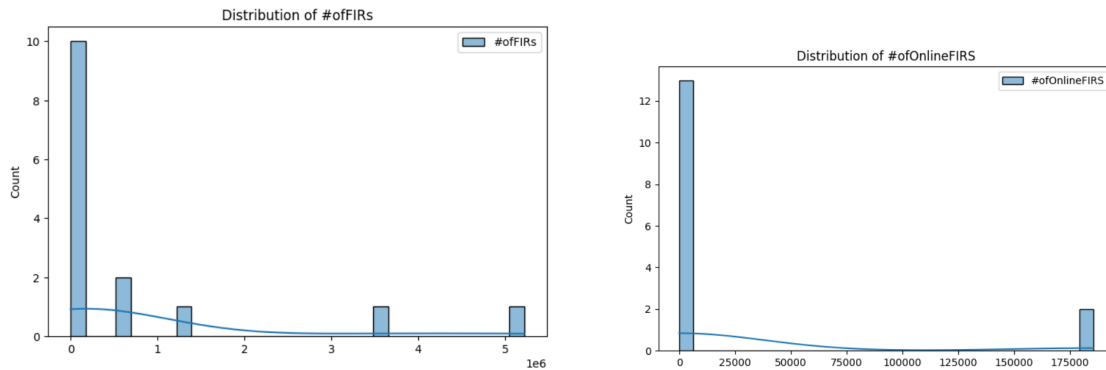
Command took 6.54 seconds -- by subhangi788@gmail.com at 12/20/2023, 11:32:13 AM on Crimes project cluster

## Crimes Rate Prediction Model:

Crime Rates Prediction Model was developed using linear regression, utilizing city-related features like population. Evaluation was performed with a Regression Evaluator, yielding a Mean Squared Error (MSE) of 6950.63 on the test data. This metric gauges the model's predictive accuracy, with lower MSE values indicating better performance.



## Exploratory Data Analysis On Crimes Complaint:

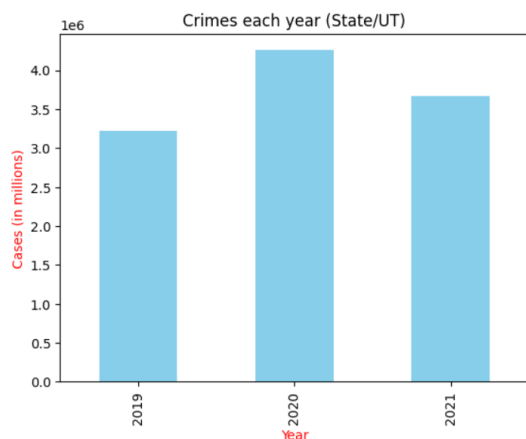


The first graph shows that there is a right skew in the distribution of the number of FIRs. This means that there are more FIRs with a lower number of FIRs than there are FIRs with a higher number of FIRs. The most common number of FIRs is 0.

The second graph suggests that the majority of online FIRs filed are in the lower range of the x-axis, with 0 being the most common number of FIRs. This can be interpreted in several ways:

- High abandonment rate: Many people may begin filing an online FIR but abandon the process before it is completed. This could be due to several factors, including a complex or confusing interface, technical issues, or a lack of trust in the online reporting system.
- Focus on minor incidents: The data could also indicate that a sizable proportion of online FIRs are filed for relatively minor incidents that do not necessitate police intervention.
- While the graph does not provide specific details about the types of FIRs filed, it's possible that online reporting is less common for serious crimes such as assault or robbery.

## Exploratory Data Analysis On States:



Result: In the year 2020, the highest number of crimes occurred.  
Total crimes in India: 11143313.0



## Challenges I faced during this project:

I was not able to access the datasets through Databricks stored in Azure data storage. Databricks authentication issues were resolved by granting access from an Azure Data Storage account. Overcoming a "permission denied" error required configuring access permissions for Databricks to interact with the Azure Data Storage account seamlessly. StackOverflow helped me successfully navigate and resolve the authentication issue.

Github link: <https://github.com/SUBHANGI-DHASMANA/crimes-in-india>

Create a free Azure account: <https://azure.microsoft.com/en-in/pricing/free-services>

## Business benefits:

### 1. Efficient Data Management:

- Utilized Azure Data Storage for secure, scalable storage.
- Leveraged GitHub for version control, ensuring data integrity.

### 2. Streamlined Data Processing:

- Azure Data Factory orchestrated seamless data transfer from GitHub to Data Lake Gen 2
- Well-designed pipelines ensured efficient ingestion and organization of raw data.

### 3. Integrated Cloud Ecosystem:

- Leveraged Azure services, including Azure Databricks, for comprehensive data analysis.
- Achieved smooth integration between Azure components for enhanced efficiency.

### 4. Enhanced Decision-Making:

- Azure Databricks facilitated exploratory data analysis, uncovering patterns and trends.
- Insights gained contribute to well-informed strategic decision-making.

### 5. Secure Data Handling:

- Implemented Azure Key Vault for secure storage of credentials, ensuring data access control.
- Secured sensitive information critical for data retrieval.

### 6. Predictive Modeling for Crime Rates:

- Developed a crime rates prediction model using linear regression.
- Evaluation through Regression Evaluator provided a metric (MSE) for predictive accuracy.

### 7. Data Visualization for Insights:

- Conducted exploratory data analysis on crimes complaints and states.
- Visualized findings, such as distribution of FIRs and crime occurrences over time.

### 8. Challenges Overcome:

- Resolved authentication issues in Azure Databricks for seamless data access.
- Successfully addressed "permission denied" error through StackOverflow assistance.

### 9. Open Collaboration and Documentation:

- Shared project resources on GitHub for transparency and collaboration.

- Documented the process, enabling others to understand and replicate the project.

**10. Business Impact:**

- Informed decision-making based on data insights.
- Improved efficiency in data processing and analysis.
- Enhanced data security measures for sensitive information.
- Potential for future predictive analytics to guide crime prevention strategies.