Big Data & Infrastructure Project

"Analyzing Crime and Education Data through a Data Lake Environment"

Alireza Foroughi
Ulster university's London Campus

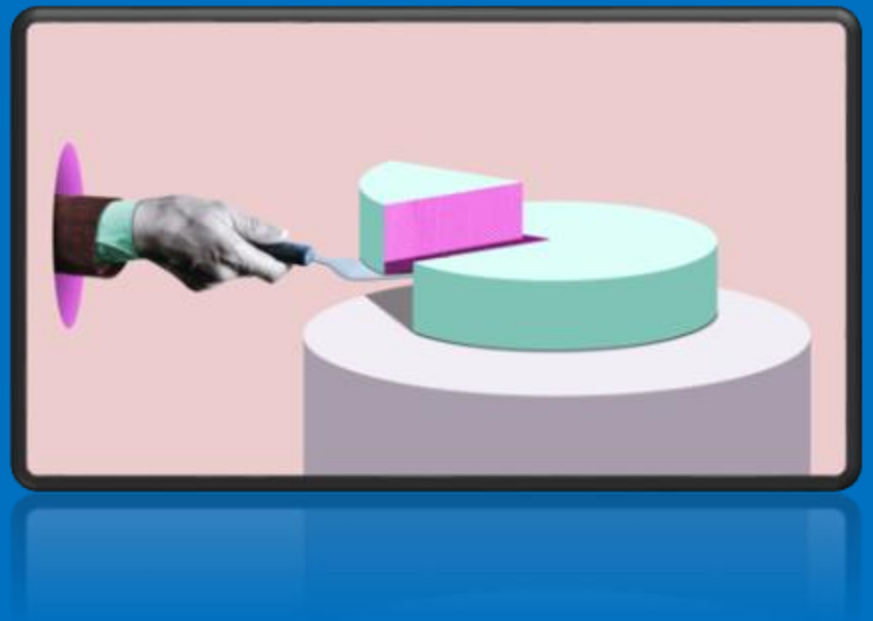
Discussion of the problem and justification of the dataset

1. Problem Statement:

"Two important measures of society development are **crime** and **education**.

In order to find patterns that could guide resource allocation and policy, this research investigates the ***relationship between income, education, and crime rates.***

2. Justification for Dataset Choice:



Crime Data:
Crucial for comprehending global governance and safety. aids in identifying high-risk locations that need attention.

Education and Income Data:
Economic growth and crime reduction are significantly influenced by education. The study can examine whether there is a relationship between lower crime rates and greater education levels by **merging income and education data**.

3. Expected Outcomes:

1. **Determine Correlations:**

This highlights the significance of education in lowering crime by illustrating the connection between greater crime rates and lower educational attainment in various geographical areas.

2. **Regional Insights:**

Emphasizing the distinctions between areas, such as Asia and Europe, it demonstrates that the discrepancies in income and educational achievement cause the crime rates in Asia to vary more.

3. **Policy Recommendations:**

Promoting balanced development and safer communities by offering insights for specific socioeconomic reforms and educational enhancements to combat crime in low-income, low-education areas.

Overview of the technical solution developed

1- Dataset Selection:

- Datasets on crime rates and education/income levels were sourced from **Kaggle**.

- *Reason:*

Kaggle provides a rich repository of well-maintained datasets that align with the project goals of analyzing educational attainment and crime rates.

2- Azure Storage Account and Container Creation:

- A Storage Account was created on **Azure** to store the datasets securely. A **container** within the storage account was set up to organize and manage files.

- *Reason:*
- Azure offers **scalable**, **cost-effective**, and **secure storage**, making it suitable for handling large datasets, with seamless integration into other Azure services like **Databricks**.

Choosing Databricks and Spark for Data Lake:

- **Apache Spark** was chosen as the processing engine, and **Databricks** was chosen as the **data lake environment**.

Why Spark and Databricks?

1. Scalability: PySpark's capacity to manage large datasets makes parallel processing and distributed computing possible.

2. Flexibility: Spark is perfect for preprocessing and analytics since it can handle both structured and unstructured data.

3. Integration: Databricks makes it easier to create and implement data pipelines **by directly integrating with Azure.**

Connecting Azure Storage to PySpark on Databricks:

- Azure Storage was linked to Databricks using **PySpark** to facilitate direct data access and analysis.

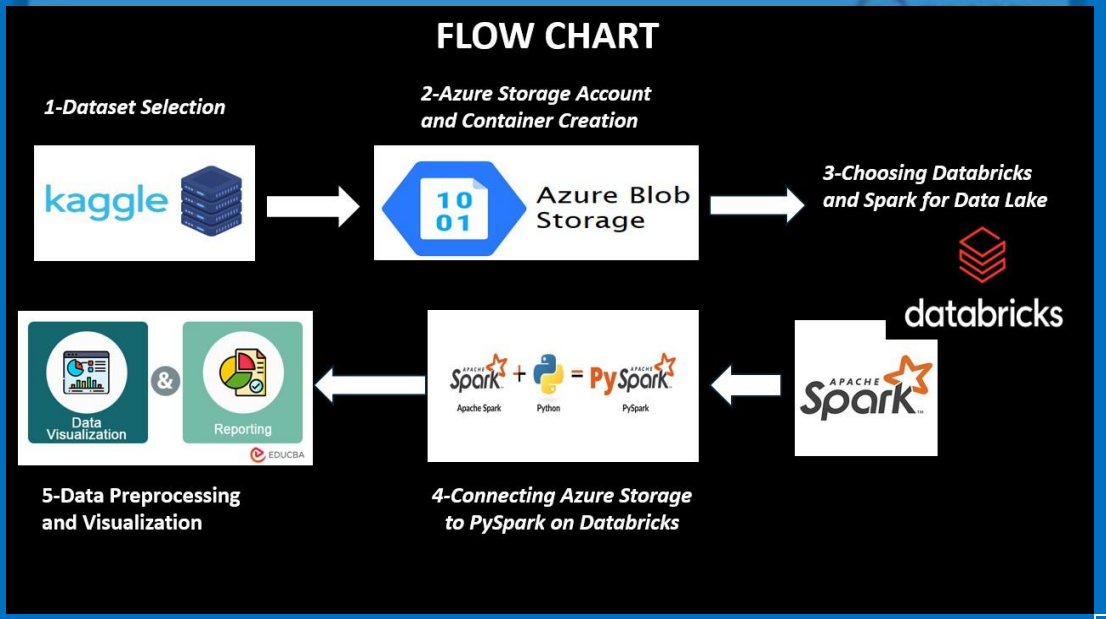
- *Reason:*

This setup allows for a smooth data pipeline from raw storage to transformation and visualization within the same ecosystem, reducing latency and improving workflow efficiency.

Data Preprocessing and Visualization:

•PySpark was used to **clean, combine, and analyze data** for preprocessing and visualization.

•Reason:
PySpark is an effective tool for processing massive amounts of data because it has strong **libraries** for data manipulation and can manage batch and streaming data effectively.



The analysis performed, and insight obtained

Large datasets were preprocessed, transformed, and shown using PySpark in Databricks. For effective data management and cleaning, Spark DataFrames were used to combine information on income, education, and crime rates.

1-Key Techniques Applied:

- **Filtering:** To guarantee data accuracy, null values and inconsistencies were eliminated.
- **Aggregation:** To compute average crime indices and identify trends, data was grouped by income and educational attainment.

2-Visualization Methods:

box plots, bar charts, and scatter plots to find relationships between crime, income, and education.

3- Why PySpark?

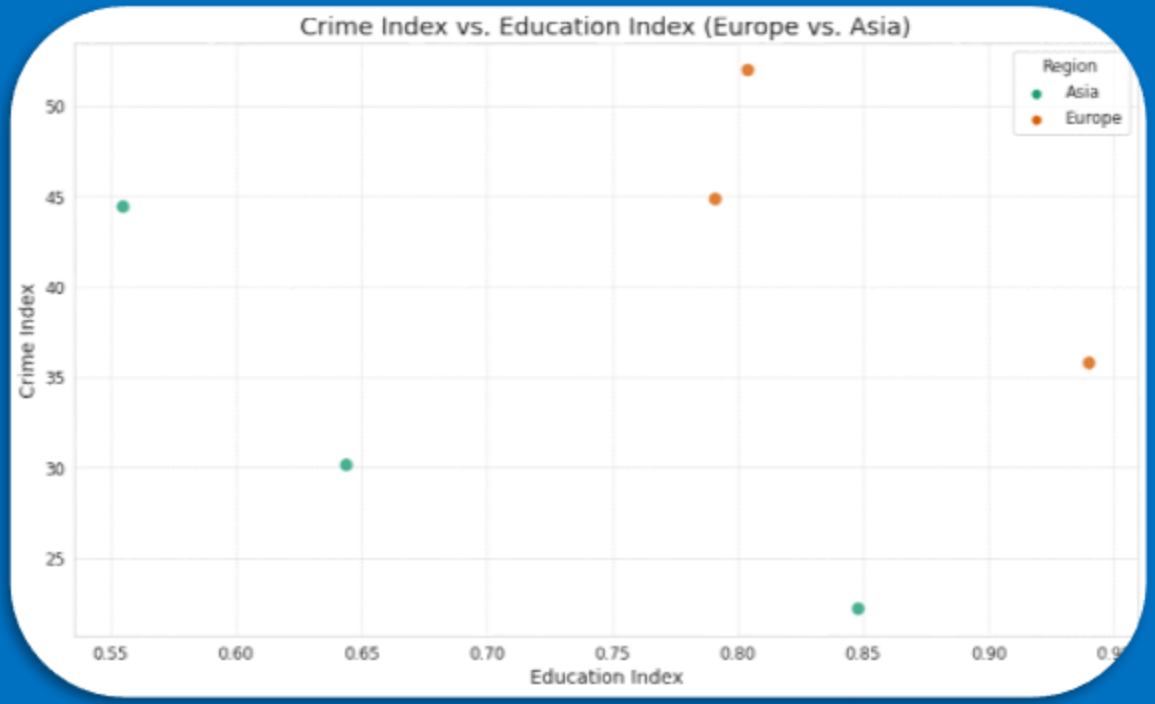
Scalable and Effective: Uses distributed processing to manage huge datasets.
Smooth Integration: Allows for easy data transmission to Databricks for analysis by connecting with Azure Blob Storage.

Exploratory Data Analysis

1. Scatter Analysis

Compare global crime and education indices to identify trends.

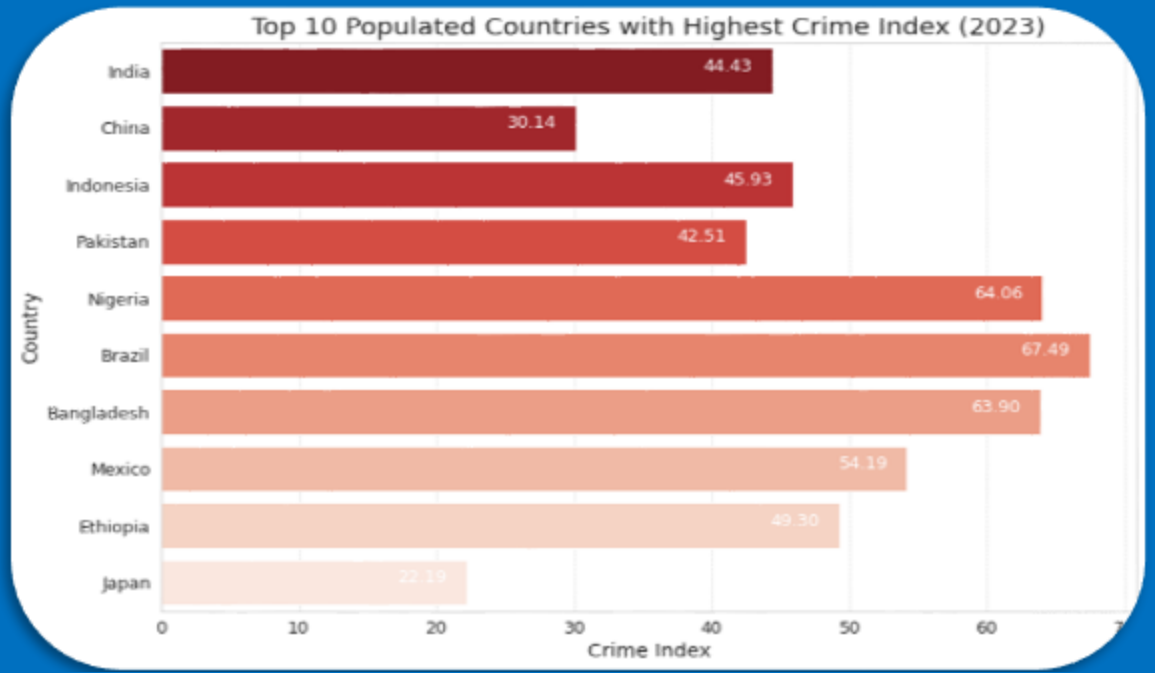
important regions (Asia vs. Europe) to examine regional differences in education and crime.



2.Important Findings:

In all regions, **higher crime rates are correlated with lower education indices.**

Crime rates vary more in Asia than in Europe, which may indicate unequal educational development.

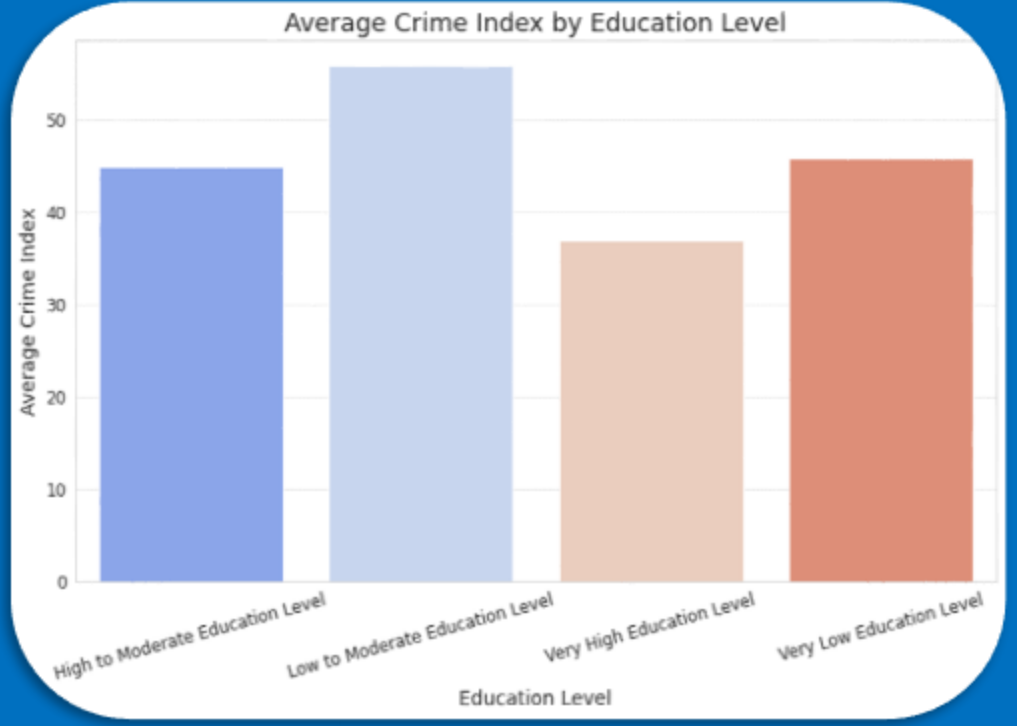


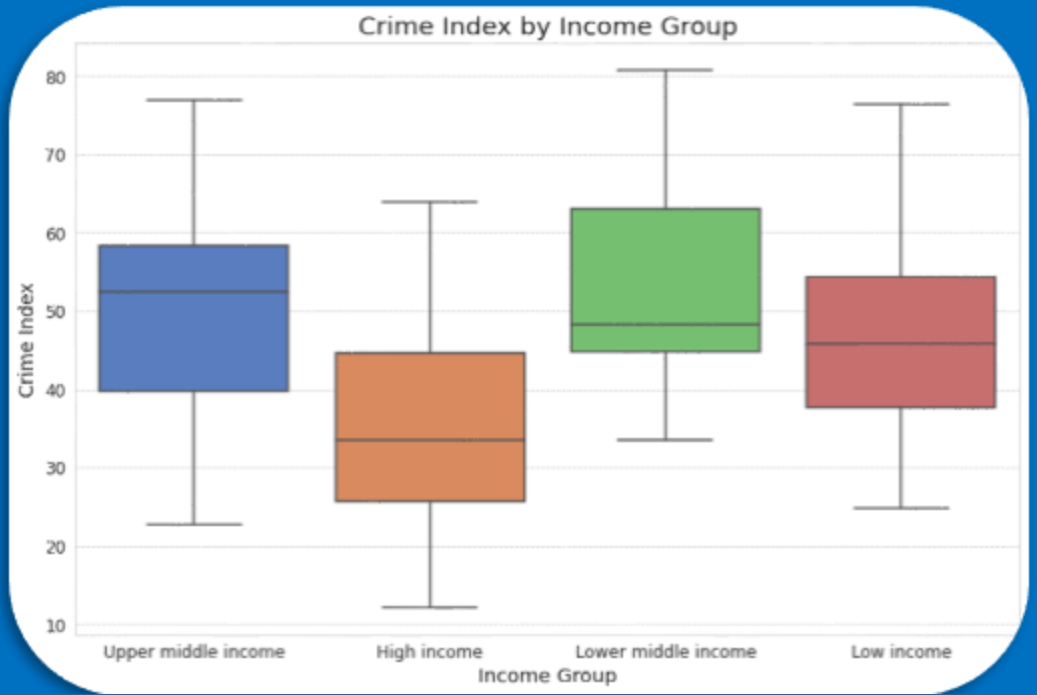
Comparative Analysis – Income and Education

Income Bracket Comparison:
Countries grouped according to revenue (Low, Middle, High) in order to examine the distribution of crimes.
There is less variation in crime rates across higher-income groups.

Crime rates were greater in lower-income nations with moderate levels of education.

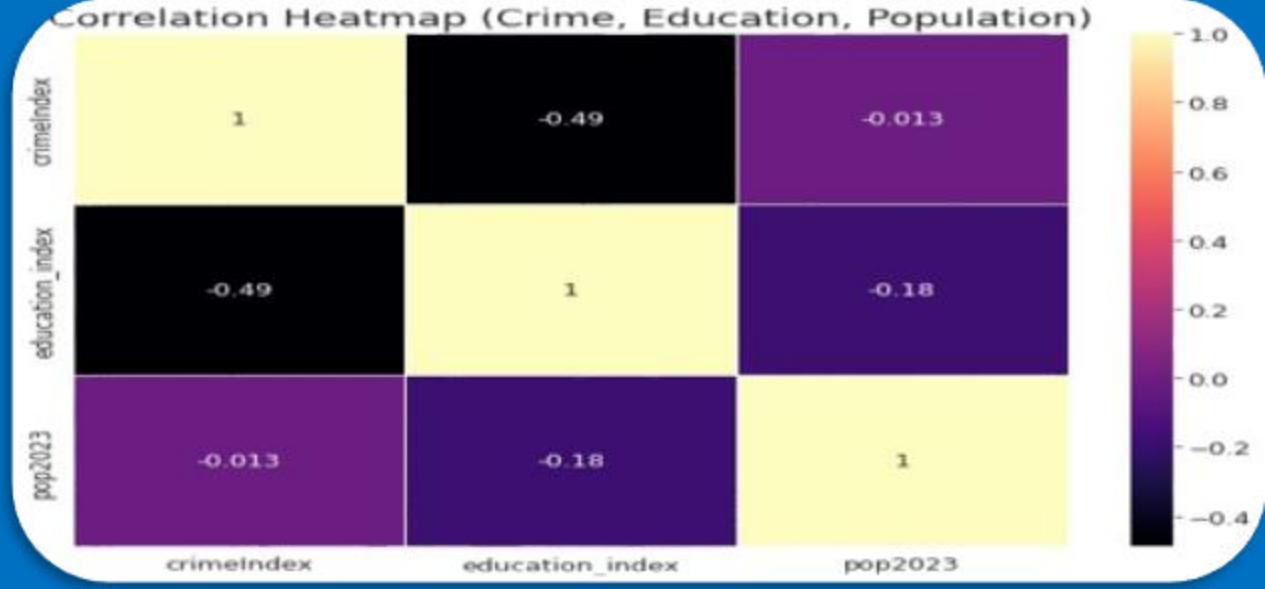
Result: Education is still a mitigating factor, **but income has a significant impact on crime.**





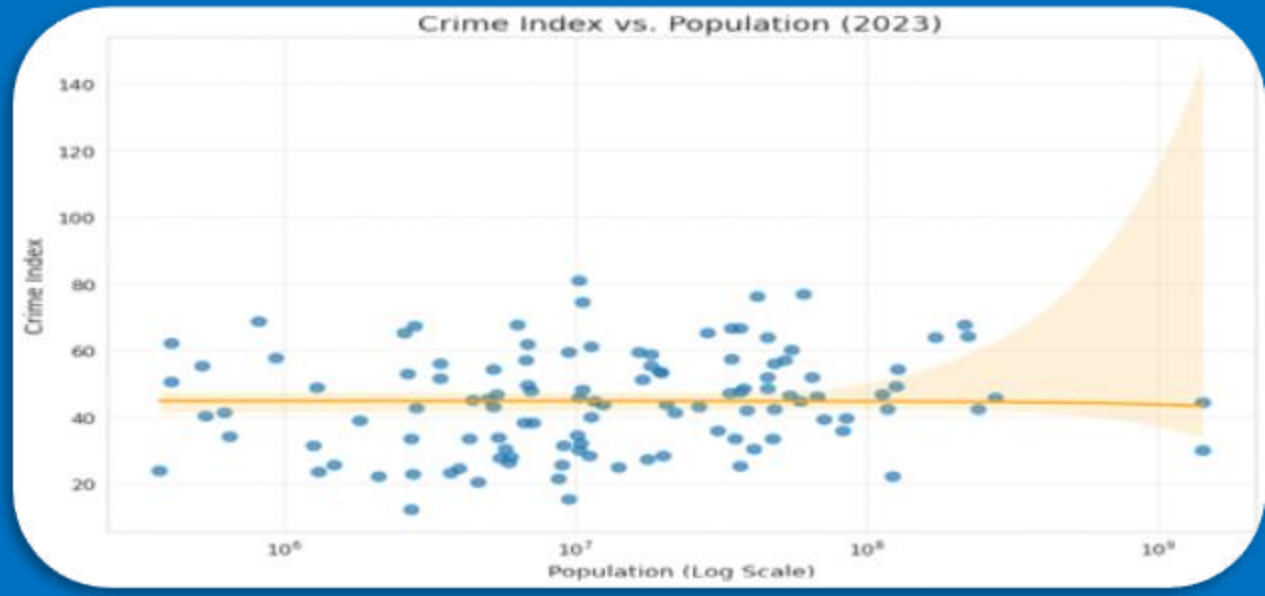
Correlation and Heatmap Analysis

The analysis of correlation:
crime rate and education index have a negative relationship.
Low income and increased crime are positively correlated.
Compared to income and education levels, population size had less of an impact.
Result: Income and education work together to greatly lower crime.



Population and Crime Trends

Impact on the Population:
There is little relationship between crime rate and population size.
Crime indices are higher in areas with a large population and low levels of education.



Conclusion:
The data shows that there is a direct correlation between crime, income, and education.
The necessity for focused educational changes in low-income, high-crime areas is highlighted by visualizations that shed light on regional discrepancies.

Concluding comments

This project successfully analyzed the correlation between income, education, and crime rates using PySpark and Databricks, offering insightful data visualization.

Reflection:

•The results of the investigation showed definite relationships between education and crime, backed up with bar charts and scatter plots.

Weaknesses:

•Deeper regional insights were limited by the granularity of the dataset.
•Analysis of crime trends across a number of years was not possible due to a lack of time-series data.

Improvements:

•bigger, more varied datasets to improve precision.
•Using geographic data to map crime scenes Socioeconomic aspects are analyzed to gain broader insights.

References

[1] Kaggle, "Crime Rate by Country 2023 Dataset," [Online]. Available: https://www.kaggle.com/datasets/xyz/crime-rate-by-country-2023.
[2] Kaggle, "Country Income and Education Level Dataset," [Online]. Available: https://www.kaggle.com/datasets/abc/country-income-and-education-level.
[3] Microsoft, "Azure Blob Storage Documentation," [Online]. Available: https://learn.microsoft.com/en-us/azure/storage/blobs/.
[4] Databricks, "Getting Started with Databricks and PySpark," [Online]. Available: https://www.databricks.com/.
[5] Apache Spark, "Apache Spark Documentation," [Online]. Available: https://spark.apache.org/docs/latest/.
[6] EDU CBA, "Data Visualization and Reporting," [Online]. Available: https://www.educba.com/data-visualization-and-reporting/.
[7] Matplotlib, "Matplotlib: Visualization with Python," [Online]. Available: https://matplotlib.org/stable/contents.html.