Microsoft | databricks

Exploratory Data Analysis

What is EDA?

Introduction to Exploratory Data Analysis (EDA) for Data Science

Exploratory Data Analysis (EDA) is a crucial first step in any data science project. It involves analyzing and visualizing datasets to uncover patterns, detect anomalies, test hypotheses, and check assumptions. The goal is to gain insights that can guide further analysis and decision-making.
EDA helps data scientists:

- Understand the overall structure and distribution of data.
- Identify missing values, outliers, and inconsistencies.
- Determine relationships between variables.
- Choose appropriate models for deeper analysis.

Introduction to Exploratory Data Analysis (EDA) for Data Science

Typical EDA techniques include:

- **Summary Statistics**: Using measures like mean, median, variance, and standard deviation to get a numerical snapshot of data.
- **Data Visualization**: Employing histograms, box plots, scatter plots, and heatmaps to visually explore distributions and relationships.
- **Correlation Analysis**: Examining how different variables interact using correlation matrices and statistical tests.
- **Handling Missing Data**: Detecting gaps and deciding whether to impute, ignore, or remove missing values.
- **Feature Engineering**: Creating new variables to improve predictive power or simplify analysis.

By applying EDA effectively, data scientists ensure their datasets are clean, well-understood, and ready for further modelling.

How does Azure Databricks Help?

What are the EDA tools in Azure Databricks?

Azure Databricks has built-in analysis and visualization tools in both Databricks SQL and in Databricks Runtime. For an illustrated list of the types of visualizations available in Azure Databricks, see Visualization types.

EDA in Databricks SQL

Azure Databricks has built-in support for charts and visualizations in Databricks SQL.

You can create and run Exploratory Data Analysis SQL queries on Unity Catalog managed Tables living on the DELTA Lake. Queries can be visualized as

- AI/BI Dashboards based on Databricks SQL.
- Visualisations in Databricks SQL.

EDA in Databricks Runtime

Databricks Runtime provides a pre-built environment that has popular data exploration libraries already installed. You can see the list of the built-in libraries in the release notes.

In a Databricks Python notebook, you can combine SQL and Python to explore data. When you run code in a SQL language cell in a Python notebook, the table results are automatically made available as a Python DataFrame. For details, see Explore SQL cell results in Python notebooks.

Azure Databricks Exploratory Data Analysis (EDA)

With Azure Databricks notebooks, data scientists can perform EDA using familiar tools. For example, the upcoming demo uses some common Python libraries to handle and plot data, including:

- **Numpy**: a fundamental library for numerical computing, providing support for arrays, matrices, and a wide range of mathematical functions to operate on these data structures.

- **pandas**: a powerful data manipulation and analysis library, built on top of NumPy, that offers data structures like DataFrames to handle structured data efficiently.

- **Plotly**: an interactive graphing library that enables the creation of high-quality, interactive visualizations for data analysis and presentation.

- **Matplotlib**: a comprehensive library for creating static, animated, and interactive visualizations in Python.

Azure Databricks also provides built-in features to help you explore your data within the notebook output, such as filtering and searching data within tables, and zooming in on visualizations. You can also use Databricks Assistant to help you write code for EDA.

Demo