What is EDA?

Exploratory Data Analysis is the *first* systematic look at a new dataset.
You summarize, visualise, and sanity-check the data so you can:

- spot missing values, outliers, wrong data-types, duplicates.
- understand basic distributions and relationships (histograms, scatter plots, correlations)
- test early assumptions
- generate hypotheses and decide what features or models make sense next

Doing solid EDA greatly reduces bad surprises later in modelling.

Use cases for Data Preparation with AWS SageMaker AI

There are 3 primary use cases for *data preparation* with Amazon SageMaker AI. Choose the use case that aligns with your requirements, and then refer to the corresponding recommended feature.

The following are the primary uses cases when performing data preparation for Machine Learning.

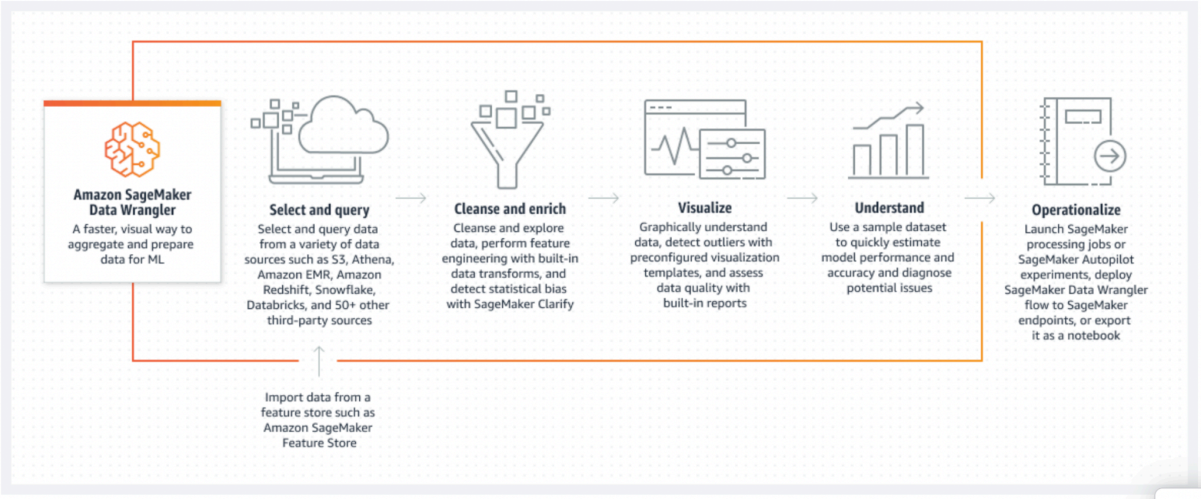
- **Use case 1**: For those who prefer a visual interface, SageMaker AI provides ways to explore, prepare, and engineer features for model training through a point-and-click environment. **For this we will use Data Wrangler within Amazon SageMaker Canvas**

Use Case 1- Exploratory Data Analysis with Amazon SageMaker Data Wrangler

Why SageMaker Data Wrangler?

Amazon SageMaker Data Wrangler reduces data prep time for tabular, image, and text data from weeks to minutes. With SageMaker Data Wrangler you can simplify data preparation and feature engineering through a visual and natural language interface. Quickly select,

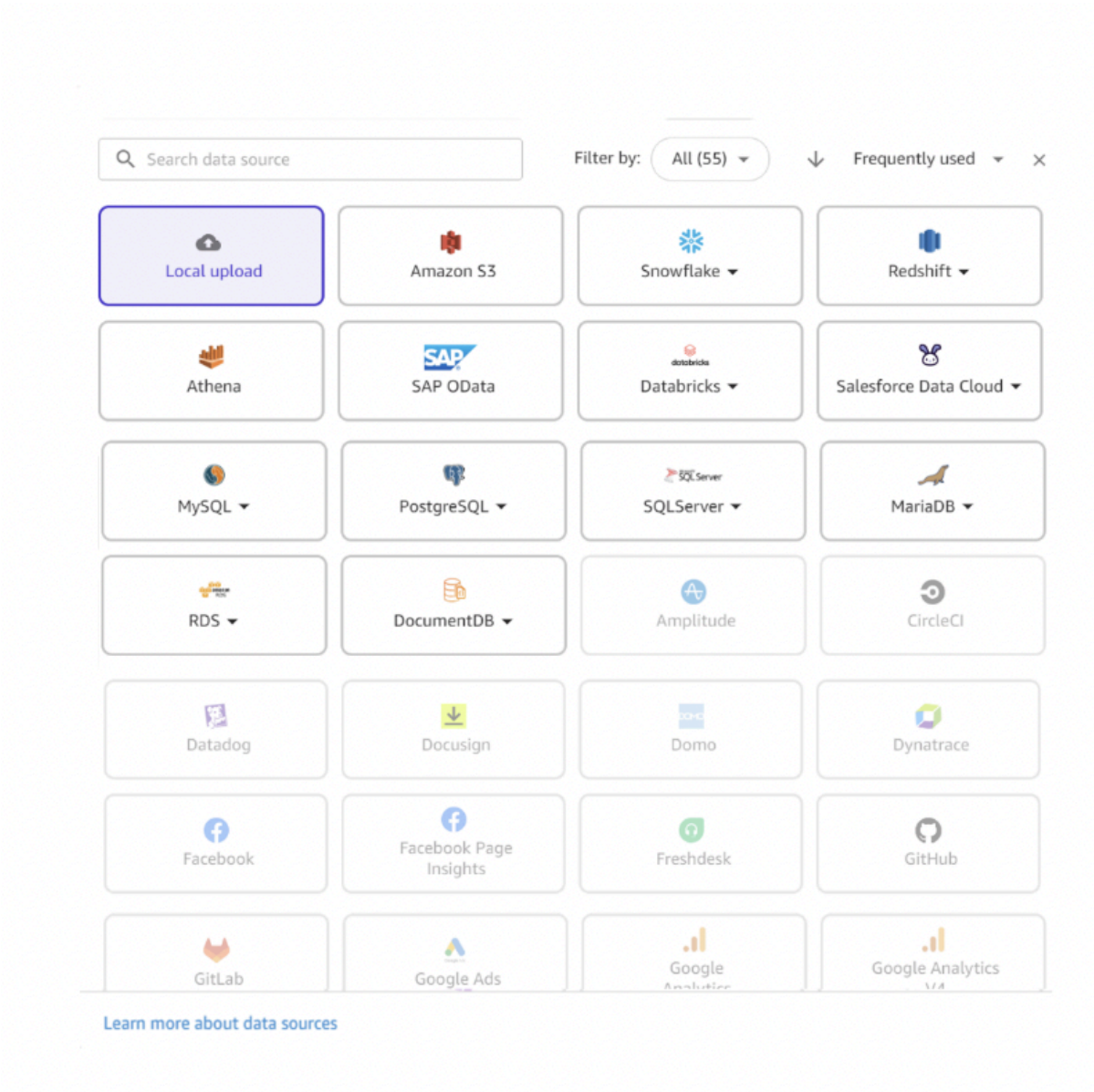
import, and transform data with SQL and over 300 built-in transformations without writing code. Generate intuitive data quality reports to detect anomalies across data types, and estimate model performance. Scale to process petabytes of data.



1. Access, select, and query data faster

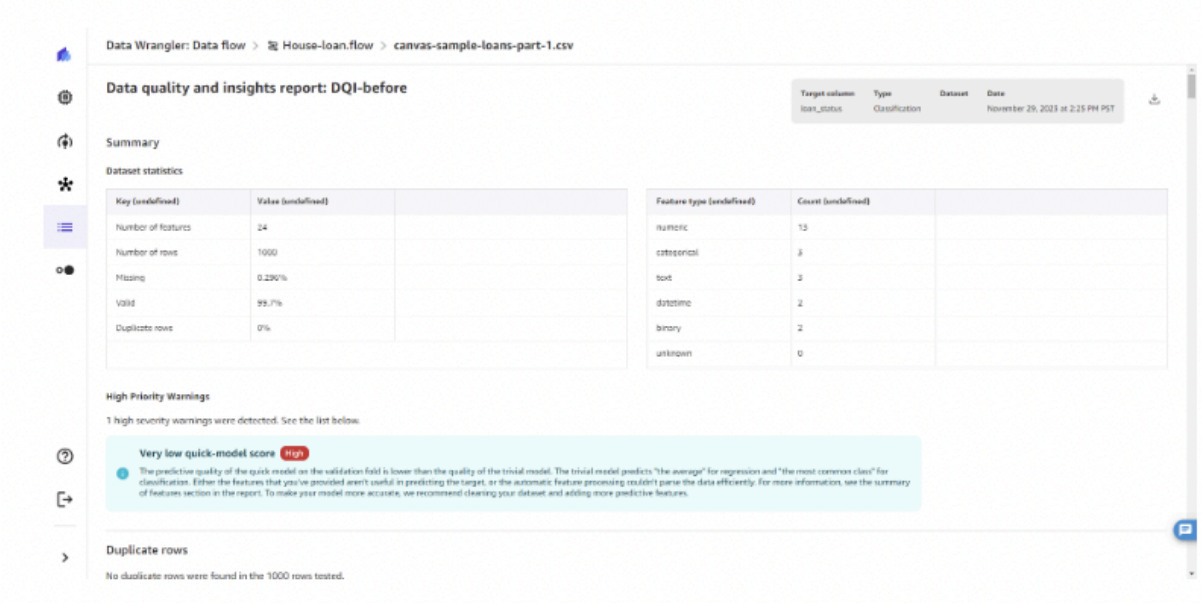
With SageMaker Data Wrangler, you can quickly access tabular, text, and image data from Amazon services such as S3, Athena, Redshift, and 50+ third-party sources. You can select data with visual query builder, write SQL queries, or import data directly in various formats

such as CSV, and Parquet.



2. Generate data insights and understand data quality

SageMaker Data Wrangler provides a data quality and insights report that automatically verifies data quality (such as missing values, duplicate rows, and data types) and helps detect anomalies (such as outliers, class imbalance, and data leakage) in your data. Once you can effectively verify data quality, you can quickly apply domain knowledge to process datasets for ML model training.



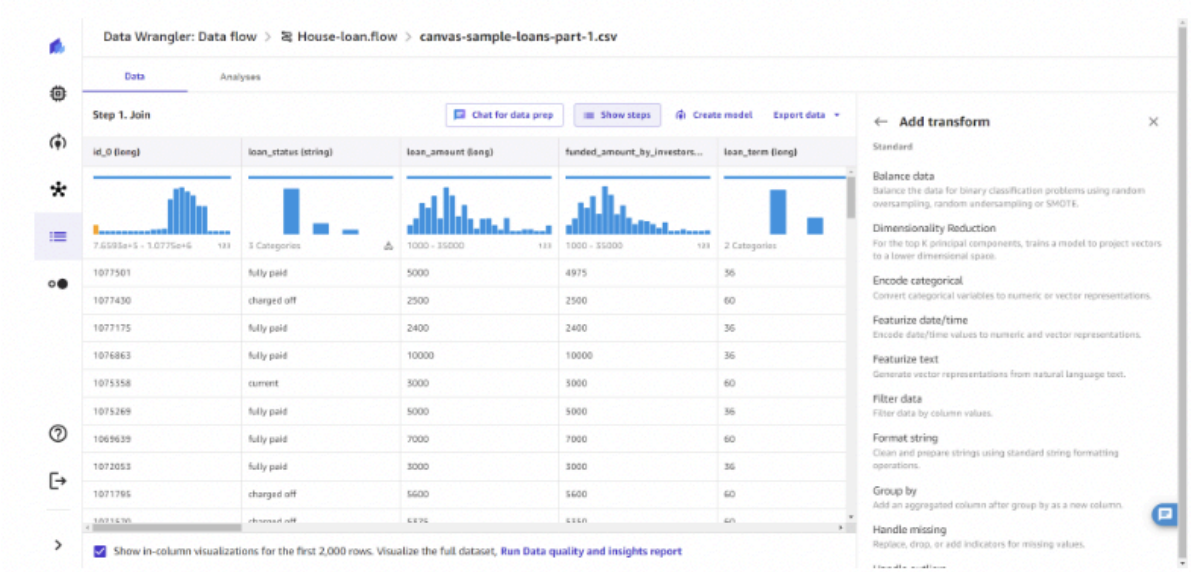
3. Understand your data with visualizations

SageMaker Data Wrangler helps you understand your data through robust built-in visualization templates such as histograms, scatter plots, feature importance, and correlations. Accelerate data exploratory with intuitive data quality reports that detect anomalies across data types and provide recommendations to improve data quality.

4. Transform data more efficiently

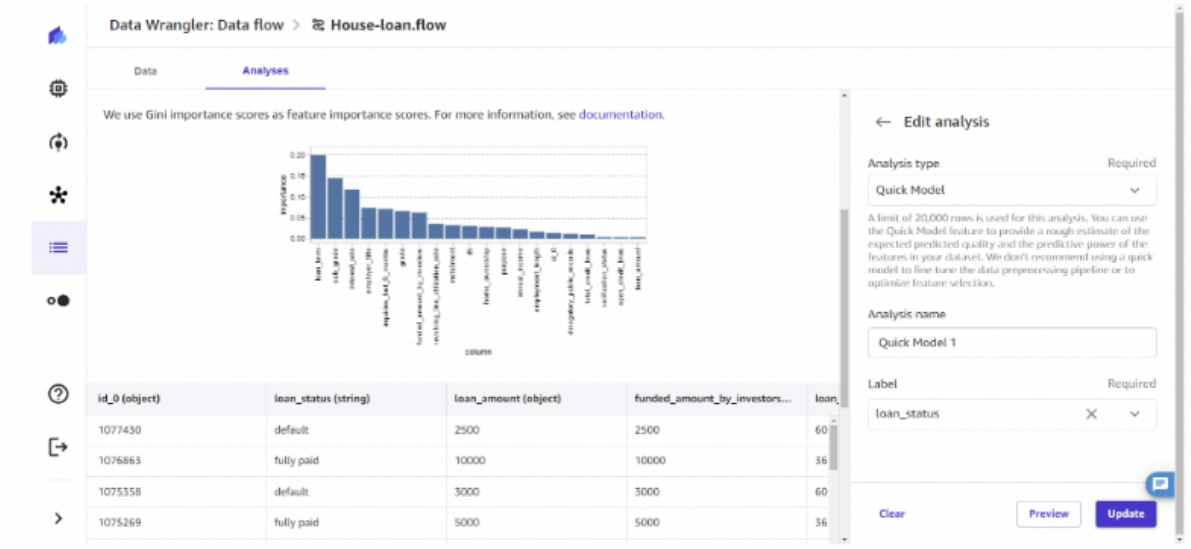
SageMaker Data Wrangler offers over 300 prebuilt PySpark transformations and a natural language interface to prepare tabular, timeseries, text and image data without coding.

Common use cases such vectorize text, featurize datetime, encoding, balancing data, or image augmentation are covered. You can also author custom transformations in PySpark, SQL, and Pandas or use natural language interface to generate code. A built-in library of code snippets simplifies writing custom transformations.



5. Understand the predictive power of your data

SageMaker Data Wrangler provides a Quick Model analysis to estimate your data's predictive power. You get estimated model accuracy, feature importance, and a confusion matrix to help you validate your data quality before training models.



- **Use case 2**: For users comfortable with coding who want more flexibility and control over data preparation, SageMaker AI integrates tools into its coding environments for exploration, transformations, and feature engineering. **For this we will use Data preparation with SQL in Studio**
- **Use case 3**: For users focused on scalable data preparation, SageMaker AI offers serverless capabilities that leverage the Hadoop/Spark ecosystem for distributed processing of big data. **For this we Prepare data using EMR Serverless applications in Studio**