

## Step 0- Start Creating Flow

The screenshot shows the SageMaker Studio Home page. On the left, there's a sidebar with sections for Applications (5), Home, Running instances, Compute, and Data. Under Data, the Data Wrangler and Feature Store options are visible. The main content area features an "Onboarding plan" box with three cards: "Take the tour", "Access your EFS data in JupyterLab and CodeEditor", and "Access your Studio Classic apps". Below this is a "Recent spaces" section.

The screenshot shows the Data Wrangler page within SageMaker Studio. The sidebar remains the same. The main content area has a banner about Data Wrangler being available in SageMaker Canvas. It then displays the "Data Wrangler" title and a brief description: "Aggregate, explore and prepare data to build Generative AI or ML solutions in SageMaker Canvas, no code required." It shows a "Status" button indicating the service is "Running". Below this, there's a "How Data Wrangler works" section with four icons: Import data, Prepare data, Export, and Train. The "Import data" section describes aggregating data from sources like Amazon S3, Athena, Snowflake, and others. The "Prepare data" section describes exploring and transforming data with natural language or a visual interface. The "Export" section describes creating pipelines using SageMaker Processing jobs and SageMaker Pipelines. The "Train" section describes using data to build custom models with AutoML. At the bottom, there's an "Examples and tutorials" section with four cards: Tabular data, Image data, Text data, and Time series data.

## Step 1- Import Data

**Data Wrangler**

Search data flows Import data flows Import and prepare

Dataset type Tabular CSV or Parquet files Image PNG and JPG files

**Data flows** Jobs

Data Wrangler data flows are recipes that capture all the data preparation steps you performed on the data. With data flows, you can import data from Amazon S3, Amazon Athena, Snowflake, and over 50 data sources to join, transform and analyze data using over 300 built-in operators and a natural language interface.

Flow name	Created	Last updated
New data flow 2025-4-28 4:57:00 PM.flow	04/28/2025 4:57 PM	04/28/2025 5:28 PM

**Resources**

**Get started**

- What is Amazon SageMaker Data Wrangler
- Get started with Data Wrangler in Canvas

**Documentation**

- Pricing in Canvas
- AWS technical guide

**What's new**

- Accelerate data preparation for ML with comprehensive data preparation capabilities and a natural language interface

**Import tabular data**

Select a data source: **Amazon S3**

Input S3 endpoint

Provide the ARN, URI, Aliases should have the form

**Amazon S3**

Name
sagemake
sagemake
sagemake

Learn more about data sources

Cancel Next

Search data source Filter by: All (56) Frequently used

Local upload	Canvas Datasets	Amazon S3	Snowflake
Redshift	Athena	Databricks	Salesforce Data Cloud
SQL Server	MySQL	Oracle	Apache Hadoop

## Import tabular data

← Previewing 1 file Showing the first 100 rows

Import settings

If your data has special character delimiters, use the advanced import settings to specify a custom delimiter. [Learn More](#)

bankloan (1).csv ▾

Delete

ID	Age	Experience	Income	Zi
1	25	1	49	91
2	45	19	34	90
3	39	15	11	94
4	35	9	100	94
5	35	8	45	91
6	37	13	29	92

### Import settings

Settings apply to all imported files. [Learn more](#)

Dataset name \*

bankloan (1).csv

### Sampling

Sample your dataset for faster exploration. Your full dataset will be used for data export or model build.

[Learn more](#)

Sampling method \* ⓘ

Random

Random sampling ensures that each row has an equal probability of being chosen.

Sample size ⓘ

50000

1 50k 100k 150k 200k  
(Recommended)

### Advanced

Cancel

Back

Import

## Step 2- Generate Data Quality Insights

Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow

Add data ▾

Step 2: Data types

Data flow Data Analyses

Run validation

Source

Local File Upload: bankloan.csv

Data types

Transform: ban

Columns: -- Rows: --

Help Log out

Home Amazon Q Data Wrangler Datasets My Models ML Ops Ready-to-use Gen AI Help Log out

Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow > bankloan.csv

**Data Quality And Insights Report** +

Data Quality And Insights Report: 30-04-2025\_10:38:09

No Preview available

Create analysis

Data size \* ⓘ

Sampled dataset

Full dataset

Instance type \* ml.m5.4xlarge

Number of instances \* ⓘ 2

It will take around 10 minutes to generate a report on the full dataset. For a quicker report, choose sampled dataset.

**Clear** **Create**

Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow > bankloan.csv

**Step 2: Data types** Data flow Data Analyses Create model

**Data Quality And Insights Report** +

**Data quality and insights report: 30-04-2025\_10:38:09**

Target column	Type	Dataset	Date
Loanapproved	Classification	bankloan.csv	30 April 2025 at 10:52 GMT+5:30

**Summary**

**Dataset statistics**

Key	Value
Number of features	14
Number of rows	5000
Missing	0%
Valid	100%
Duplicate rows	0%

Feature type	Count
numeric	9
categorical	0
text	0
datetime	0
binary	4
unknown	0

### Step 3- Add Cleaning Transformations- click on data flow and then add transform

**Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow**

**Step 2: Data types**

Validation complete    0 errors    Done    Run validation

Source → Data types

Local File Upload: bankloan.csv

Data types

Transform: ban

Columns: 14    Rows: 5,000

**Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow > bankloan.csv**

**Step 2. Data types**

Chat for data prep    Data    Analyses    Create model

Build a data insights report to identify data quality issues, recommends fixes, and estimates feature importance and model accuracy.

Get data insights

ID (long)	Age (long)	Experience (long)	Income (long)
1 - 5000	23 - 67	-3 - 43	8 - 224
1	25	1	49
2	45	19	34
3	39	15	11
4	35	9	100
5	35	8	45
6	37	13	29
7	53	27	72
8	50	24	22

Sampling: 50,000    Columns: 14    Rows: 5,000    Show visualizations

← Filter or drop rows    X

Filter or drop rows that don't match a pattern that you specify.

Filters

Column name \*    Age

Condition \*    <

Value \*    18

Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow > bankloan.csv

**Step 3. Filter or drop rows** Chat for data prep Data flow Data Analyses Create model

Build a data insights report to identify data quality issues, recommends fixes, and estimates feature importance and model accuracy.

Get data insights X

ID (long) Age (long) Experience (long) Income (long)

1 - 5000	23 - 29	-3 - 5	8 - 205
1	25	1	49
12	29	5	45
23	29	5	62
55	29	5	44
59	28	2	93
75	28	3	135
86	27	2	109
90	25	-1	113

Sampling: 50,000 Columns: 14 Rows: 488 Show visualizations

← Manage columns X

Move, drop, duplicate or rename columns in the dataset. [Learn more](#).

Transform \* ⓘ

Rename column X

Input column Loanapproved X

New name ⓘ

Loan\_approved

Additional columns to rename

Log out

## Step 4&5 - Univariate EDA, Bivariate and Correlation

Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow > bankloan.csv

**Step 4: Rename column** Data flow Data Analyses Create model

Data Quality And Insights Report +

Data Quality And Insights Report: 30-04-2025\_11:07:23

No Preview available

Create analysis

Analysis type \*

Data Quality And Insights Report

Histogram

Table Summary

Bias Report

Feature Correlation

Target Leakage

Quick Model

Scatter Plot

Multicollinearity

Full dataset

Clear Create

Log out

## Step 6- Feature Engineering

Data Wrangler: Data flow > New data flow 2025-4-30 10:37:36 AM.flow > bankloan.csv

Step 4. Rename column Chat for data prep Data flow Data Analyses Create model

Build a data insights report to identify data quality issues, recommends fixes, and estimates feature importance and model accuracy. Get data insights X

ID (long)	Age (long)	Experience (long)	Income (long)
1 - 5000	23 - 29	123	123
1	25	1	49
12	29	5	45
23	29	5	62
55	29	5	44
59	28	2	93
75	28	3	135
86	27	2	109
90	25	-1	113

Custom transform Python (Pandas) X v

Using Python (Pandas) requires your dataset to fit in memory and only uses a single instance in batch computation. It is ideal for smaller datasets less than 2GB and experimentation but we recommend Python (PySpark) or Python (User-Defined Function) for production use-cases

```
1 # Table is available as variable
2 df['debt_to_income'] = df['Mortga
```

Clear Preview Add

Sampling: 50,000 Columns: 14 Rows: 488 Show visualizations

Home Amazon Q Data Wrangler Datasets My Models ML Ops Ready-to-use Gen AI Help Log out