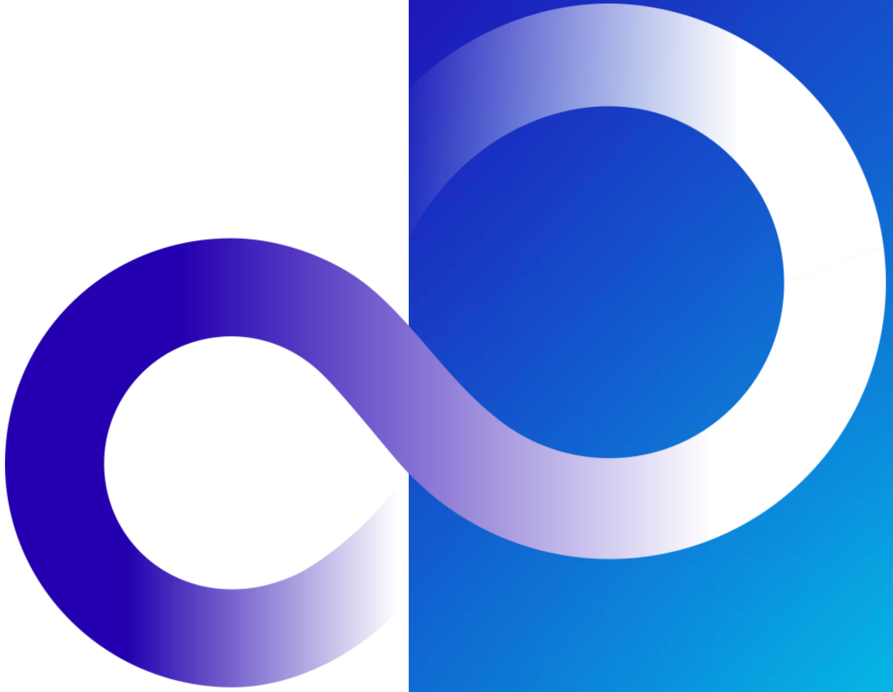
Fujitsu Kozuchi Auto Data Wrangling User Manual



Dec 20, 2023

FUJITSU Reserch
FUJITSU Limited

Fujitsu Auto Data Wrangling

Automatic pre-processing for tabular data using generative AI

Reduce efforts of data preperation in AI application by automatic pre-processing for tabular data via data cleaning and data enrichment using generative AI.

Challenges

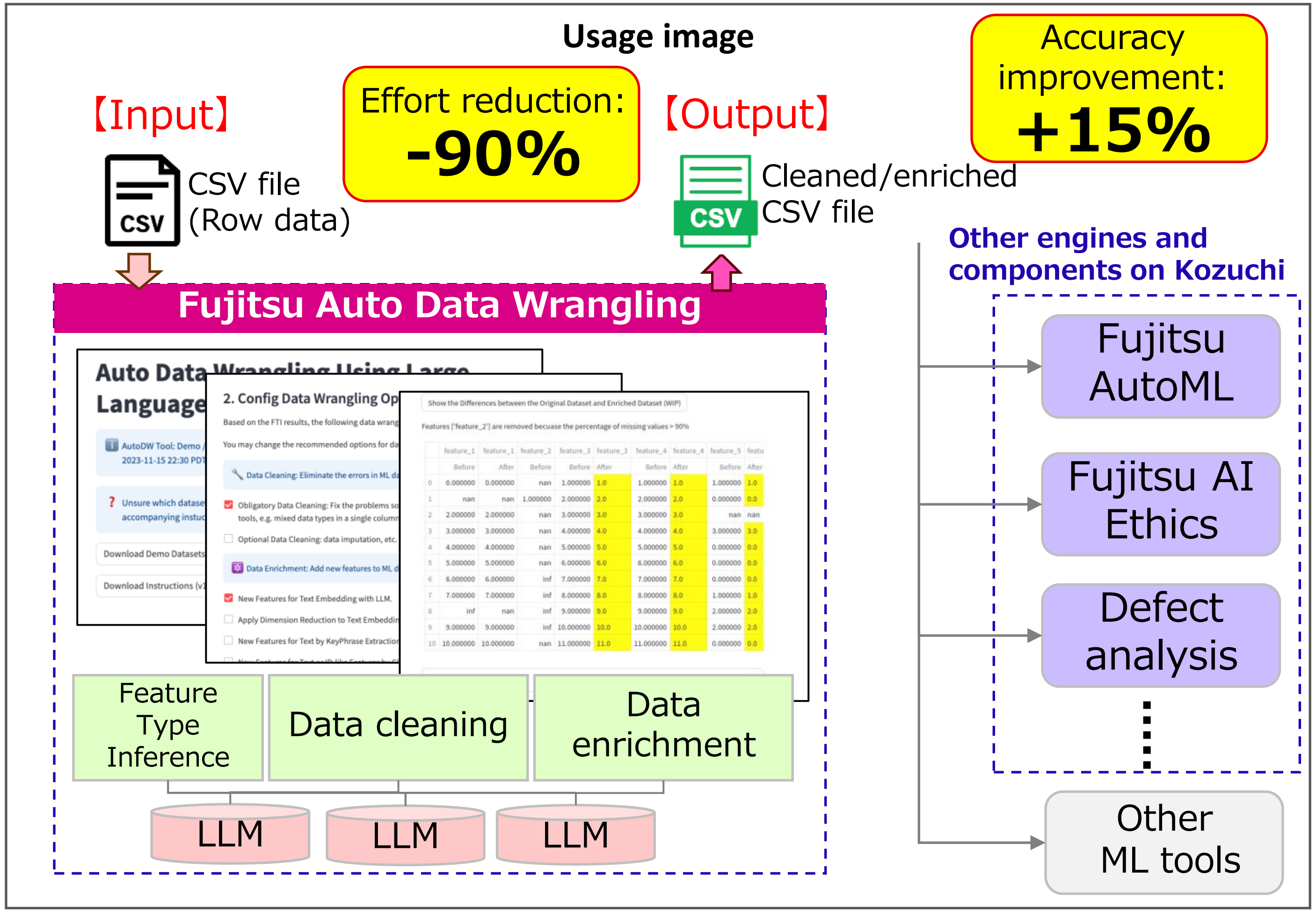
- Applying machine learning to tabular data in the field requires a lot of effort of pre-processing, such as data cleaning and data enrichment
- Conventional machine learning cannot handle a variety of text features as they are, which limits their accuracy

Solutions

- **Automatic data cleaning** based on **Feature Type Inference** using LLMs
- **Automatic data enrichment** that creates new features by analyzing existing text features using LLMs
- **Achieving both automation and scalability** by using different LLMs depending on data wrangling processes

Strengths

- Effort reduction of data preparation (90% reduction compared with manual coding of data wrangling)
- Automatic data enrichment for text features not addressed by other companies' data wrangling tools
- Improving accuracy of ML 15%+ (in case of using Fujitsu AutoML)



Usage image

【Input】 CSV file (Row data)

Effort reduction: **-90%**

【Output】 Cleaned/enriched CSV file

Accuracy improvement: **+15%**

Auto Data Wrangling Using Large Language

AutoDW Tool: Demo / 2023-11-15 22:30 PDT

? Unsure which datase accompanying instuc

Download Demo Datasets

Download Instructions (v1

2. Config Data Wrangling Op

Based on the FTI results, the following data wrang

You may change the recommended options for da

Data Cleaning: Eliminate the errors in ML da

☑ Obligatory Data Cleaning: Fix the problems so tools, e.g. mixed data types in a single column

☐ Optional Data Cleaning: data imputation, etc.

Data Enrichment: Add new features to ML d

☑ New Features for Text Embedding with LLM.

☐ Apply Dimension Reduction to Text Embeddin

☐ New Features for Text by KeyPhrase Extraction

Show the Differences between the Original Dataset and Enriched Dataset (WIP)

Features ['feature_2'] are removed becuase the percentage of missing values > 90%

	feature_1	feature_1	feature_2	feature_3	feature_3	feature_4	feature_4	feature_5	featu
	Before	After	Before	Before	After	Before	After	Before	After
0	0.000000	0.000000	nan	1.000000	1.0	1.000000	1.0	1.000000	1.0
1	nan	nan	1.000000	2.000000	2.0	2.000000	2.0	0.000000	0.0
2	2.000000	2.000000	nan	3.000000	3.0	3.000000	3.0	nan	nan
3	3.000000	3.000000	nan	4.000000	4.0	4.000000	4.0	3.000000	3.0
4	4.000000	4.000000	nan	5.000000	5.0	5.000000	5.0	0.000000	0.0
5	5.000000	5.000000	nan	6.000000	6.0	6.000000	6.0	0.000000	0.0
6	6.000000	6.000000	inf	7.000000	7.0	7.000000	7.0	0.000000	0.0
7	7.000000	7.000000	inf	8.000000	8.0	8.000000	8.0	1.000000	1.0
8	inf	nan	inf	9.000000	9.0	9.000000	9.0	2.000000	2.0
9	9.000000	9.000000	inf	10.000000	10.0	10.000000	10.0	2.000000	2.0
10	10.000000	10.000000	nan	11.000000	11.0	11.000000	11.0	0.000000	0.0

What you can and cannot do?

- What you can do : Preprocessing Before Machine Learning
  - Data Cleaning
    - Format the input data to avoid errors when applying it to machine learning tools such as Fujitsu AutoML
  - Data Enrichment
    - Analyze the feature columns of input data and add new feature columns that contribute to improving analysis efficiency
- What it can’t do: Machine learning
  - For machine learning processing such as data learning, classification, and prediction, please use existing machine learning tools such as Fujitsu AutoML.

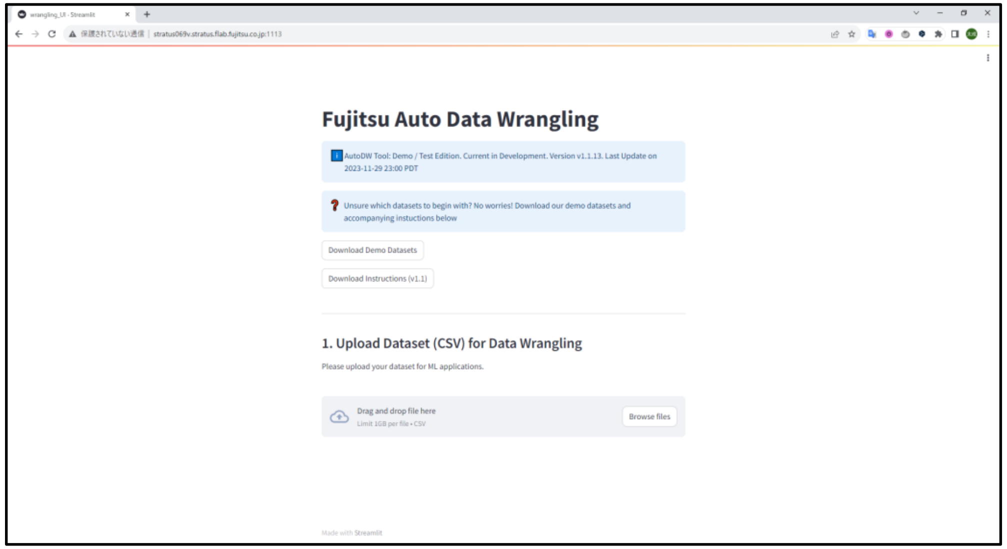
Two ways to use

- Interactive Demo
  - You can experience the operation of the Fujitsu Auto Data Wrangling using sample data provided by us.

- Proof of Concept (PoC)
  - You can use the Fujitsu Auto Data Wrangling with your own data.

Access to the environment

- Preparation
  - Download “Environment_Access_Manual” from the access manual
  - **PoC**
    - Refer to “How to Connect with Azure VPN Gateway” section
  - **Interactive Demo**
    - Refer to “Bastion connection with Remote Desktop” section
- Access
  - http://10.0.0.139:8550/ from a web browser
  - Confirmation is successful if the start screen, as displayed on the right, appears



Fujitsu Auto Data Wrangling

AutoDW Tool: Demo / Test Edition. Current in Development. Version v1.1.13. Last Update on 2023-11-29 23:00 PDT

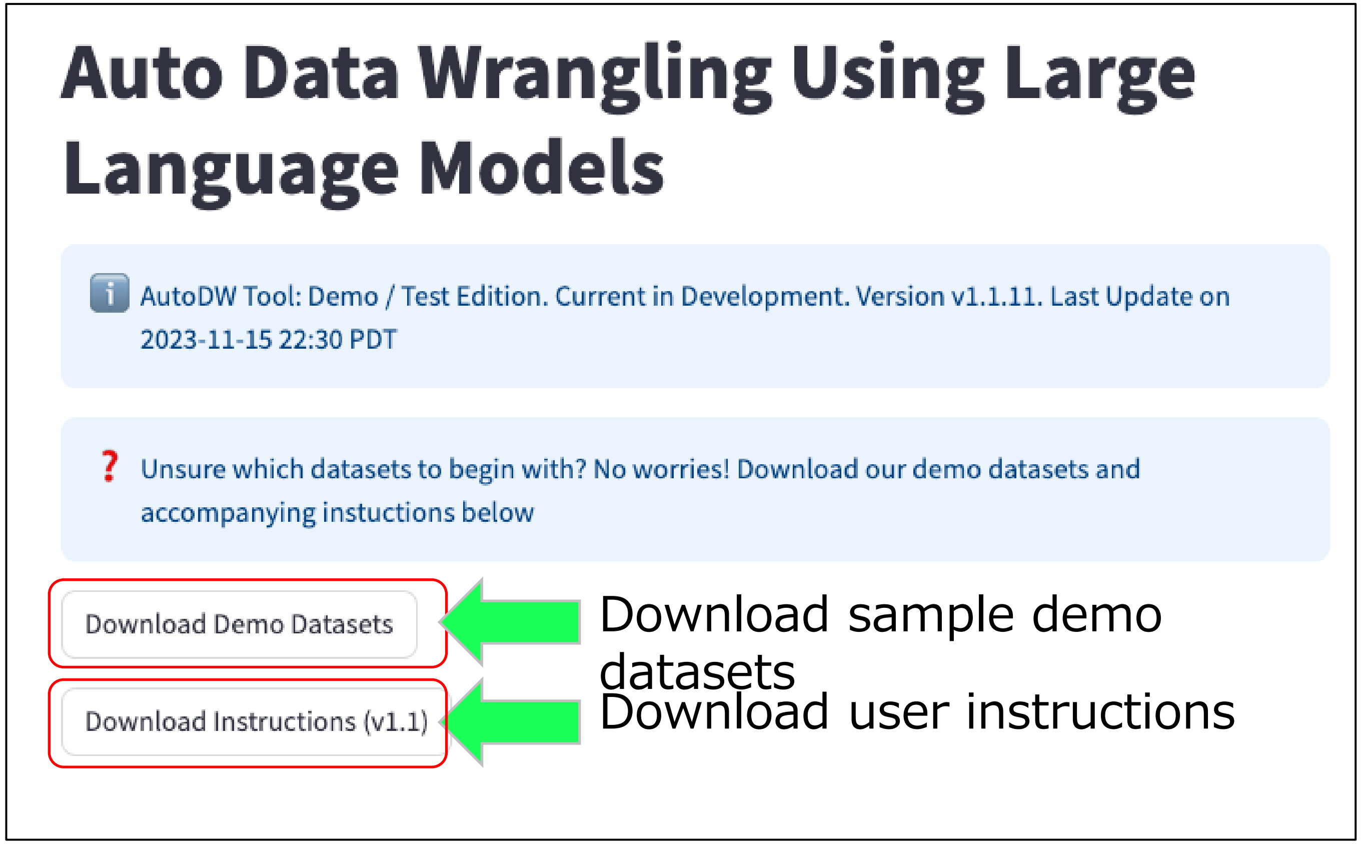
Unsure which datasets to begin with? No worries! Download our demo datasets and accompanying instuctions below

1. Upload Dataset (CSV) for Data Wrangling

Please upload your dataset for ML applications.

2. How to use Fujitsu Auto Data Wrangling webapp

2.1 Getting Started with Sample Datasets (Optional)



Auto Data Wrangling Using Large Language Models

AutoDW Tool: Demo / Test Edition. Current in Development. Version v1.1.11. Last Update on 2023-11-15 22:30 PDT

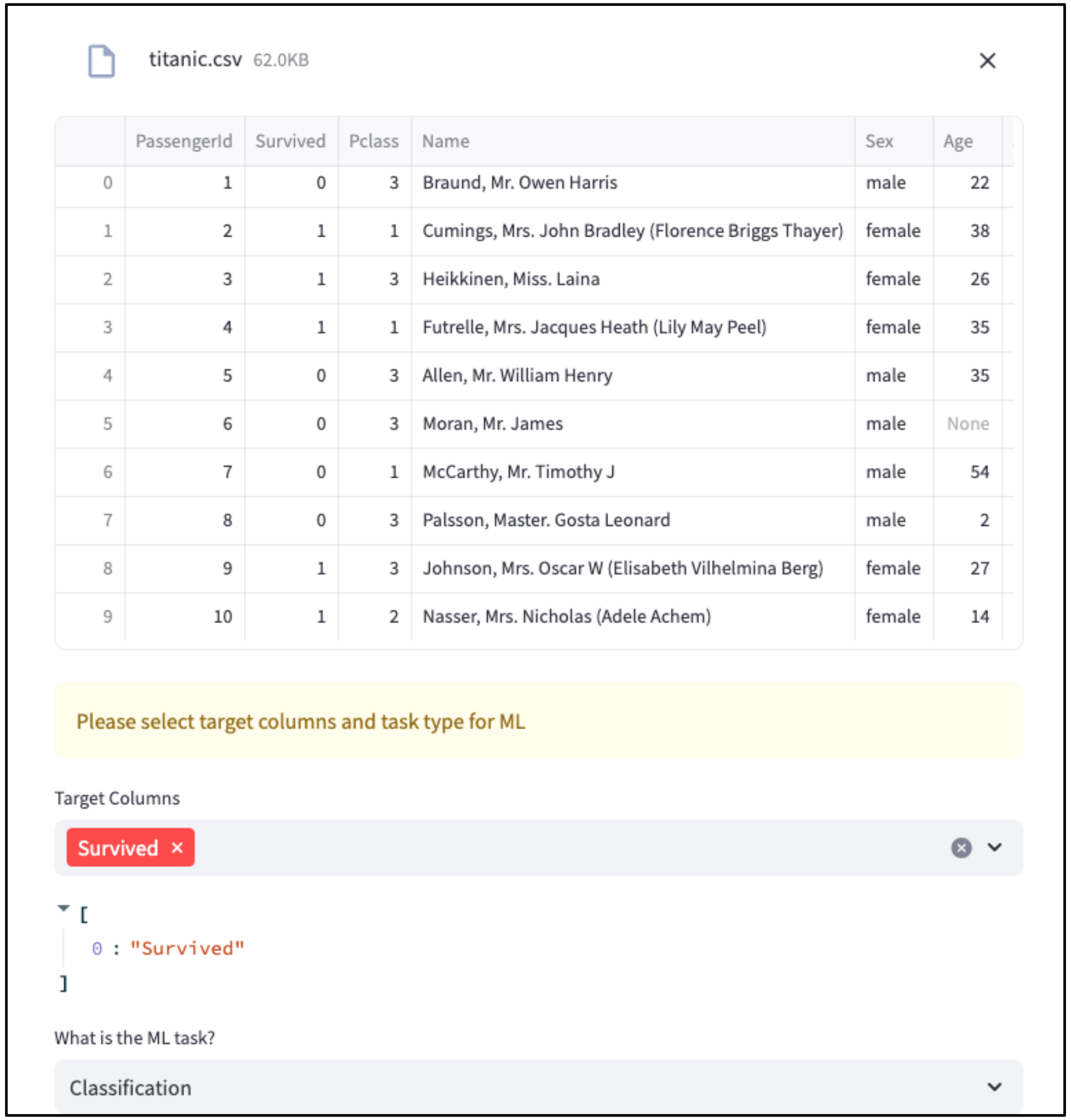
❓ Unsure which datasets to begin with? No worries! Download our demo datasets and accompanying instuctions below

Download sample demo datasets

Download user instructions

- If you are unsure about which datasets to begin with, you may download some sample demo datasets and check the effects of auto data wrangling first
- You may also download the user manual / instructions

Sample Demo Dataset #1



titanic.csv 62.0KB

	PassengerId	Survived	Pclass	Name	Sex	Age
0	1	0	3	Braund, Mr. Owen Harris	male	22
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38
2	3	1	3	Heikkinen, Miss. Laina	female	26
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35
4	5	0	3	Allen, Mr. William Henry	male	35
5	6	0	3	Moran, Mr. James	male	None
6	7	0	1	McCarthy, Mr. Timothy J	male	54
7	8	0	3	Palsson, Master. Gosta Leonard	male	2
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14

Please select target columns and task type for ML

Target Columns

Survived ×

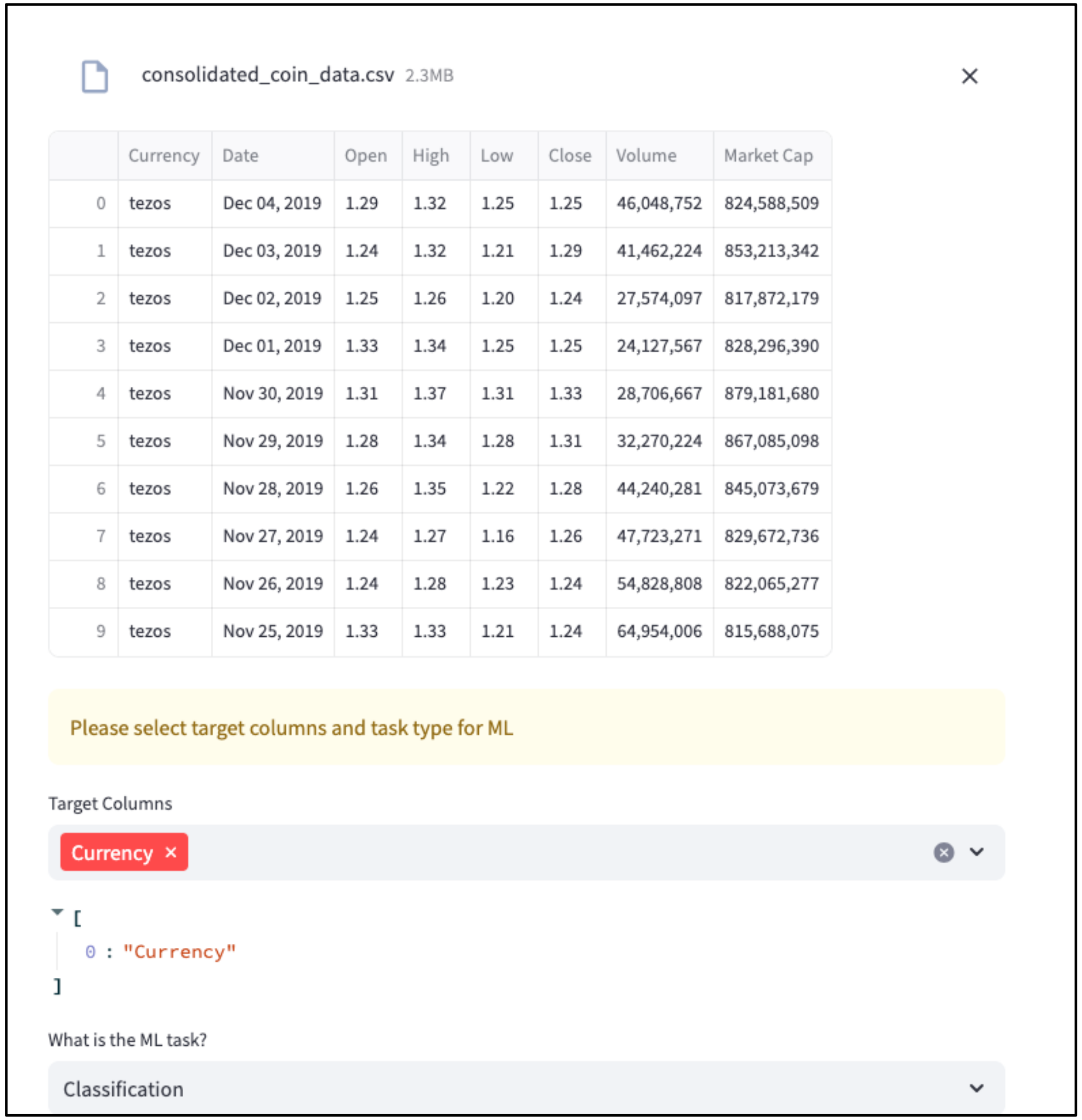
```
[
  0 : "Survived"
]
```

What is the ML task?

Classification

- titanic.csv
  - Common dataset for demo purpose
  - Target column: Survived
  - ML task: Classification

Sample Demo Dataset #2



consolidated_coin_data.csv 2.3MB

	Currency	Date	Open	High	Low	Close	Volume	Market Cap
0	tezos	Dec 04, 2019	1.29	1.32	1.25	1.25	46,048,752	824,588,509
1	tezos	Dec 03, 2019	1.24	1.32	1.21	1.29	41,462,224	853,213,342
2	tezos	Dec 02, 2019	1.25	1.26	1.20	1.24	27,574,097	817,872,179
3	tezos	Dec 01, 2019	1.33	1.34	1.25	1.25	24,127,567	828,296,390
4	tezos	Nov 30, 2019	1.31	1.37	1.31	1.33	28,706,667	879,181,680
5	tezos	Nov 29, 2019	1.28	1.34	1.28	1.31	32,270,224	867,085,098
6	tezos	Nov 28, 2019	1.26	1.35	1.22	1.28	44,240,281	845,073,679
7	tezos	Nov 27, 2019	1.24	1.27	1.16	1.26	47,723,271	829,672,736
8	tezos	Nov 26, 2019	1.24	1.28	1.23	1.24	54,828,808	822,065,277
9	tezos	Nov 25, 2019	1.33	1.33	1.21	1.24	64,954,006	815,688,075

Please select target columns and task type for ML

Target Columns

Currency ×

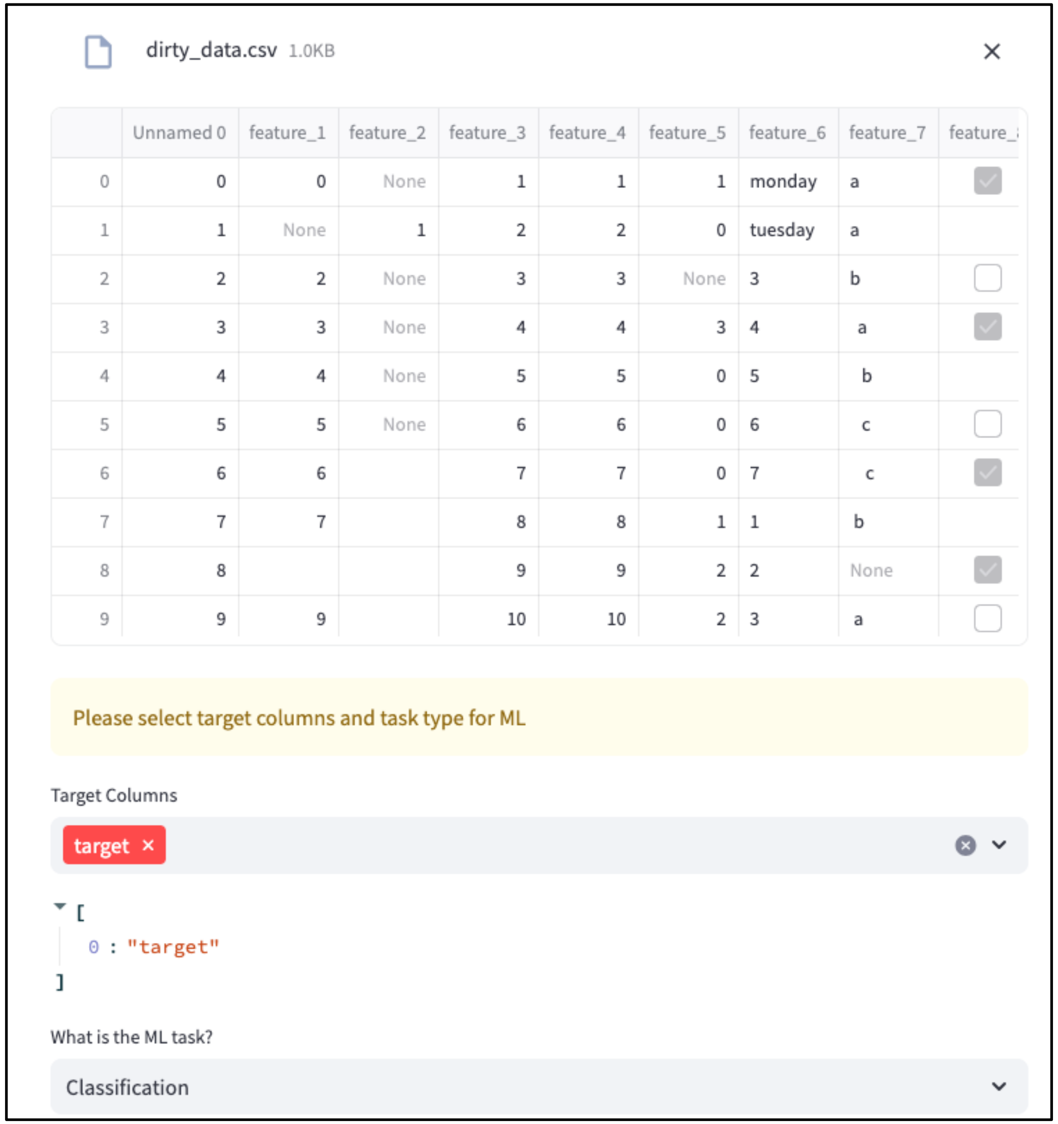
```
[
  0 : "Currency"
]
```

What is the ML task?

Classification

- consolidated_coin_data.csv
  - A representative data to show the effect of FTI
    - The numbers look numerical values, but they are strings (or numerical numbers embedded in strings)
  - Target column: Currency
  - ML task: Classification

Sample Demo Dataset #3



dirty_data.csv 1.0KB

	Unnamed 0	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_
0	0	0	None	1	1	1	monday	a	☑
1	1	None	1	2	2	0	tuesday	a	
2	2	2	None	3	3	None	3	b	☐
3	3	3	None	4	4	3	4	a	☑
4	4	4	None	5	5	0	5	b	
5	5	5	None	6	6	0	6	c	☐
6	6	6		7	7	0	7	c	☑
7	7	7		8	8	1	1	b	
8	8			9	9	2	2	None	☑
9	9	9		10	10	2	3	a	☐

Please select target columns and task type for ML

Target Columns

target ×

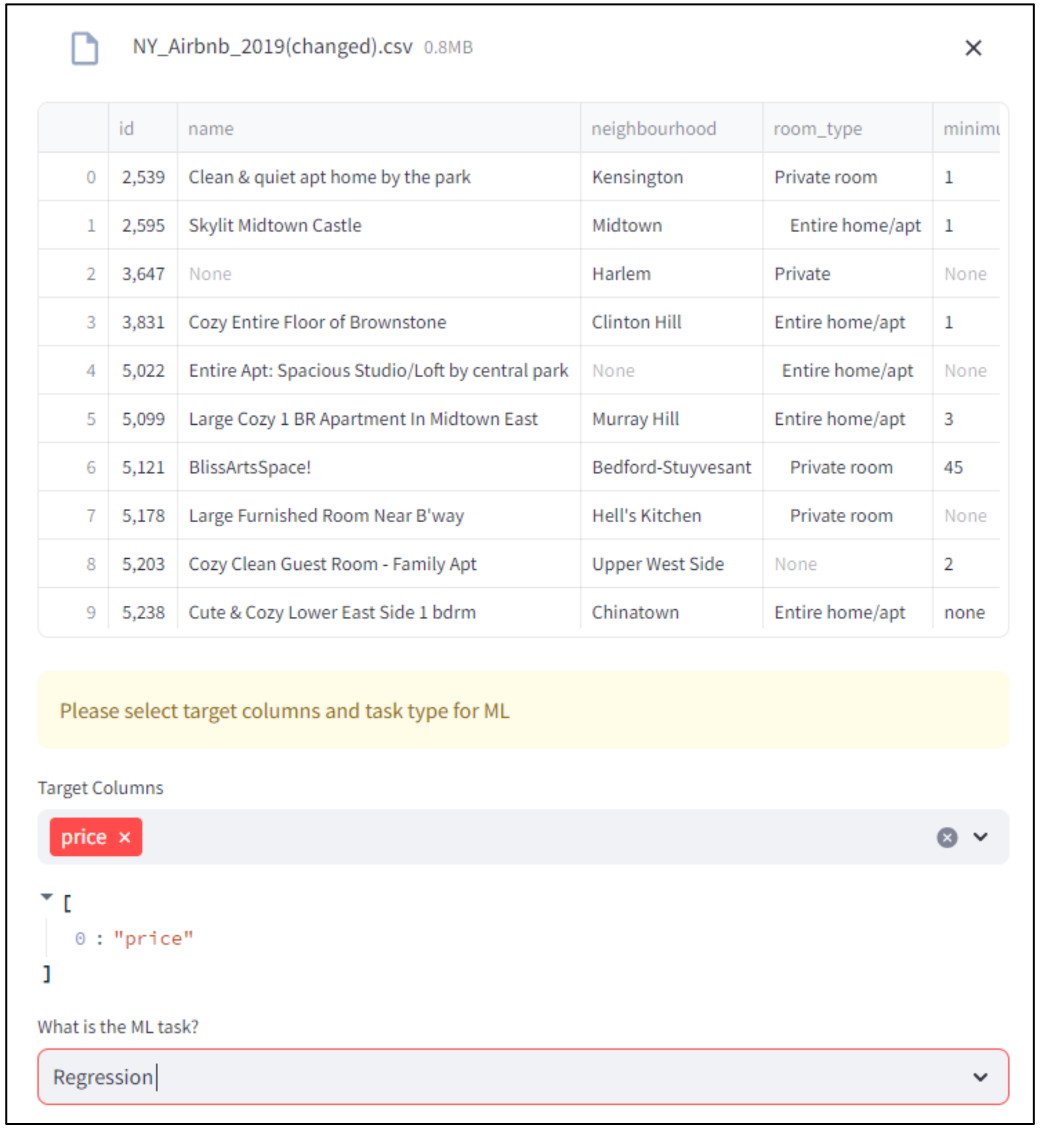
```
[
  0 : "target"
]
```

What is the ML task?

Classification

- dirty_data.csv
  - A representative data to show the effect of data cleaning
    - A small dataset full of different type of errors (e.g. mixed data type in a single column, NaN value decoded as “?”, etc.)
  - Target column: target
  - ML task: Classification

Sample Demo Dataset #4



NY_Airbnb_2019(changed).csv 0.8MB

	id	name	neighbourhood	room_type	minimu
0	2,539	Clean & quiet apt home by the park	Kensington	Private room	1
1	2,595	Skylit Midtown Castle	Midtown	Entire home/apt	1
2	3,647	None	Harlem	Private	None
3	3,831	Cozy Entire Floor of Brownstone	Clinton Hill	Entire home/apt	1
4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

Please select target columns and task type for ML

Target Columns

price ×

```
[
  0 : "price"
]
```

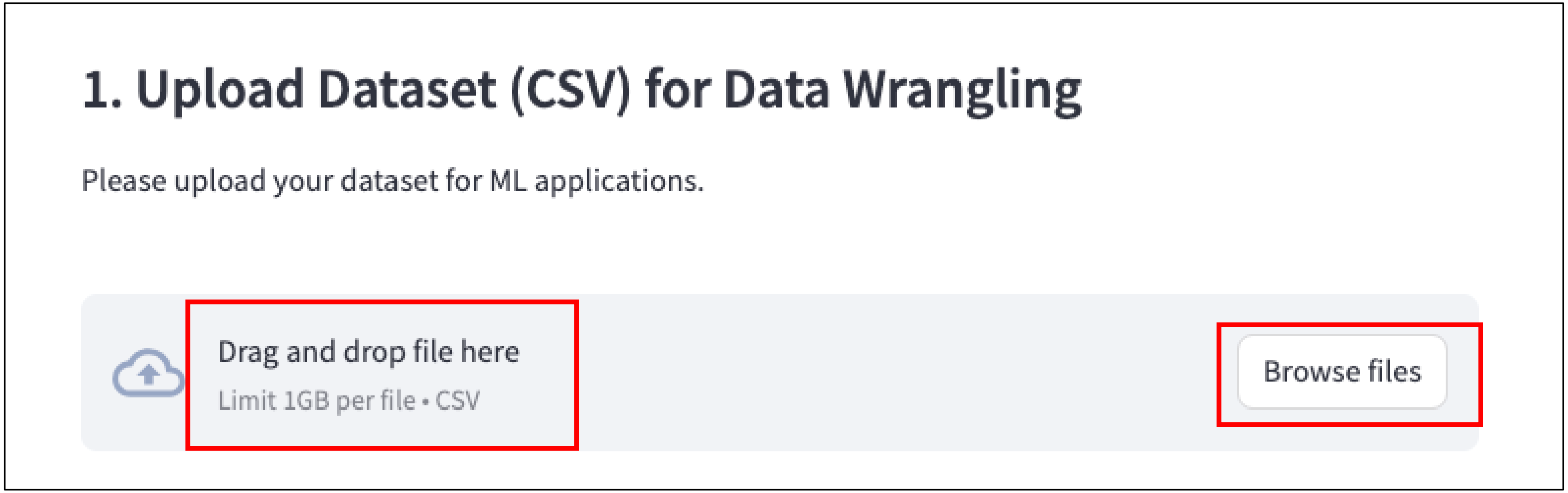
What is the ML task?

Regression

- NYC_Airbnb_2019(changed).csv
  - A representative data to show the effect of FTI, data cleaning, data enrichment
    - Include feature columns of text type, datetime type, and numeric type
    - Embedded by editing dirty_data.csv similar error
    - Created by modifying Airbnb listings and metrics in NYC, NY, USA (2019)©Airbnb Licensed under CC BY 4.0
  - Target column: Price
  - ML task:Regression

The following pages are explained using this data.
Use this file in the demo video.

2.2 Specify Dataset & Problem Setting (1)



1. Upload Dataset (CSV) for Data Wrangling

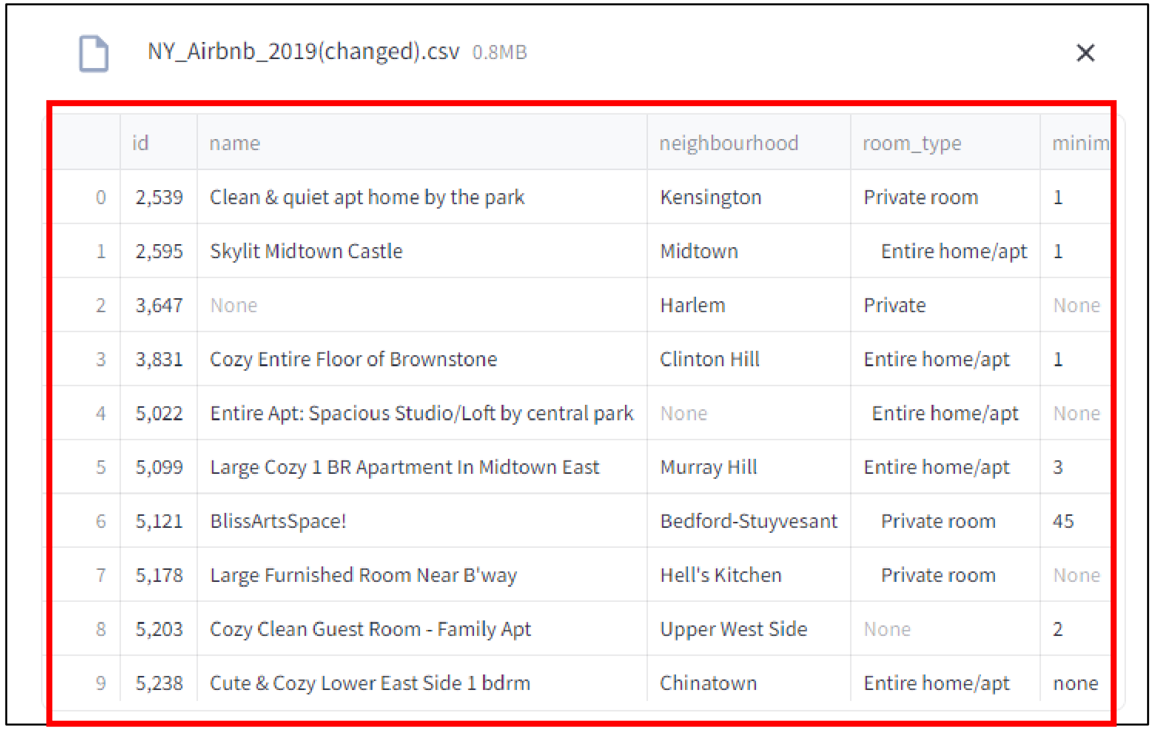
Please upload your dataset for ML applications.

Drag and drop file here

Limit 1GB per file • CSV

- Two approaches to upload a dataset (CSV) for data wrangling
  - Drag and drop a csv file
  - Click “Brower files” button and select a csv file

2.2 Specify Dataset & Problem Setting (2)

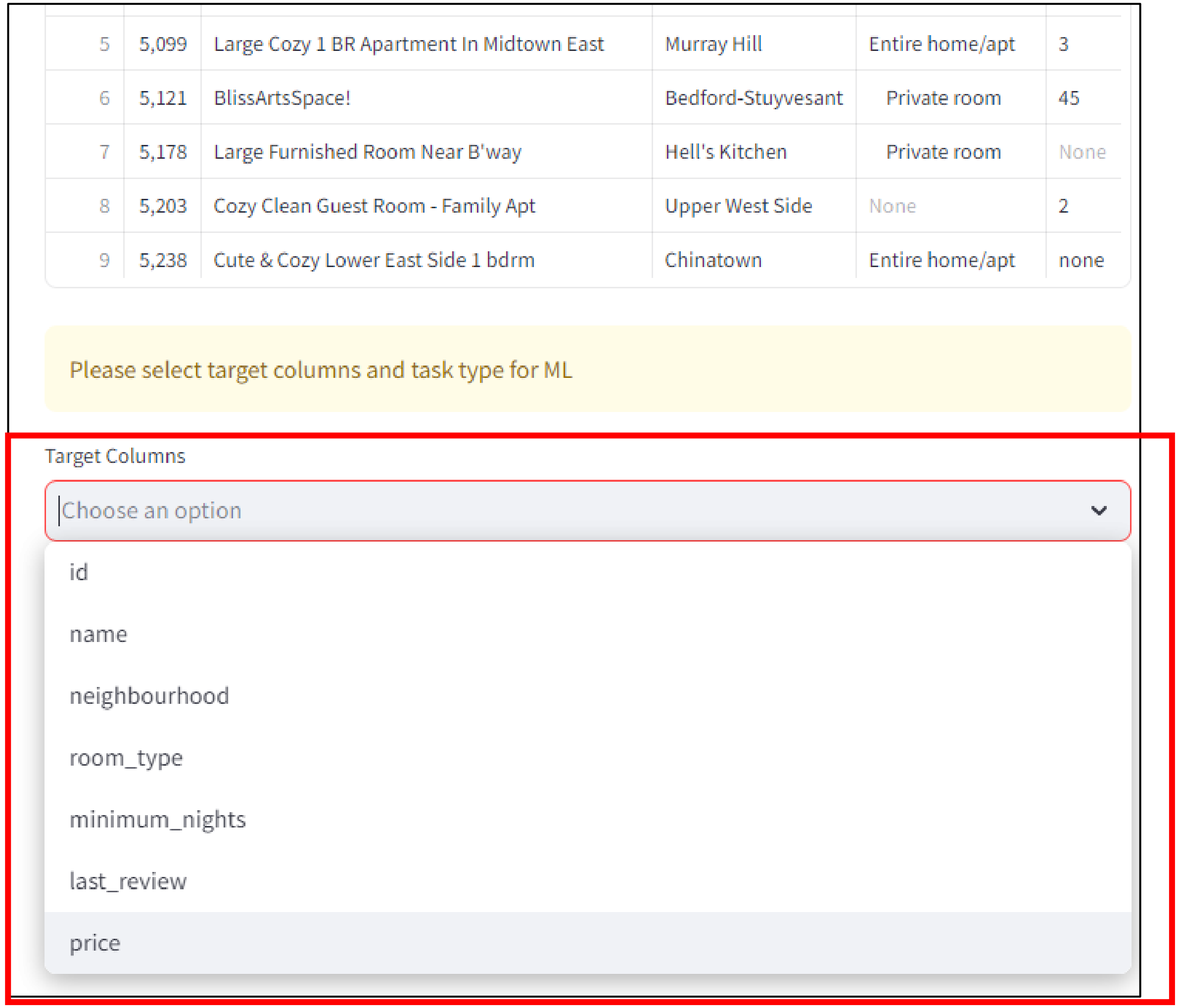


NY_Airbnb_2019(changed).csv 0.8MB

	id	name	neighbourhood	room_type	minim
0	2,539	Clean & quiet apt home by the park	Kensington	Private room	1
1	2,595	Skylit Midtown Castle	Midtown	Entire home/apt	1
2	3,647	None	Harlem	Private	None
3	3,831	Cozy Entire Floor of Brownstone	Clinton Hill	Entire home/apt	1
4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

- A preview of the uploaded dataset is displayed.

2.2 Specify Dataset & Problem Setting (3)



5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

Please select target columns and task type for ML

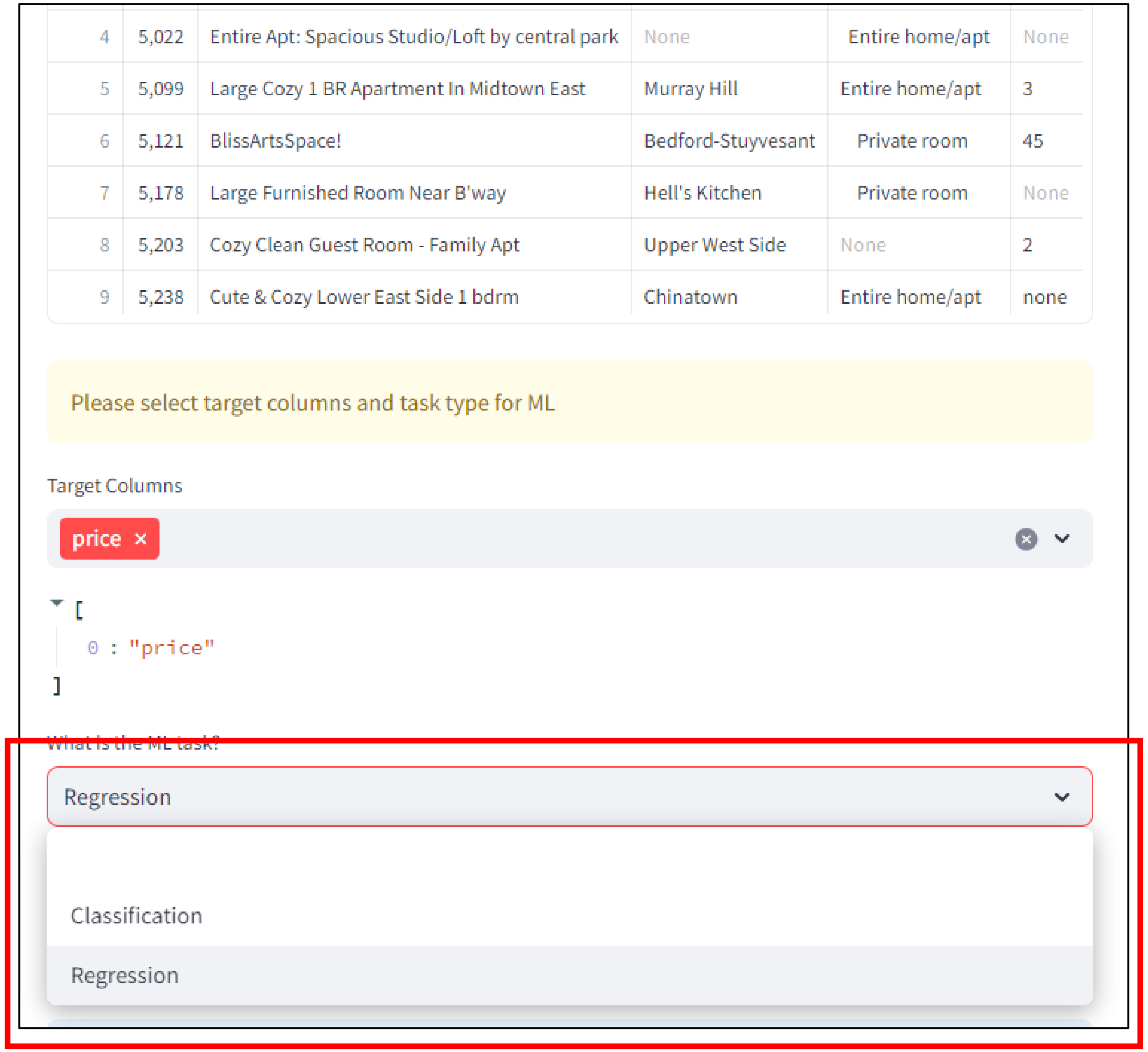
Target Columns

Choose an option

- id
- name
- neighbourhood
- room_type
- minimum_nights
- last_review
- price

- Select target variables (columns you want to predict) in “Target Columns”

2.2 Specify Dataset & Problem Setting (4)



4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

Please select target columns and task type for ML

Target Columns

price ×

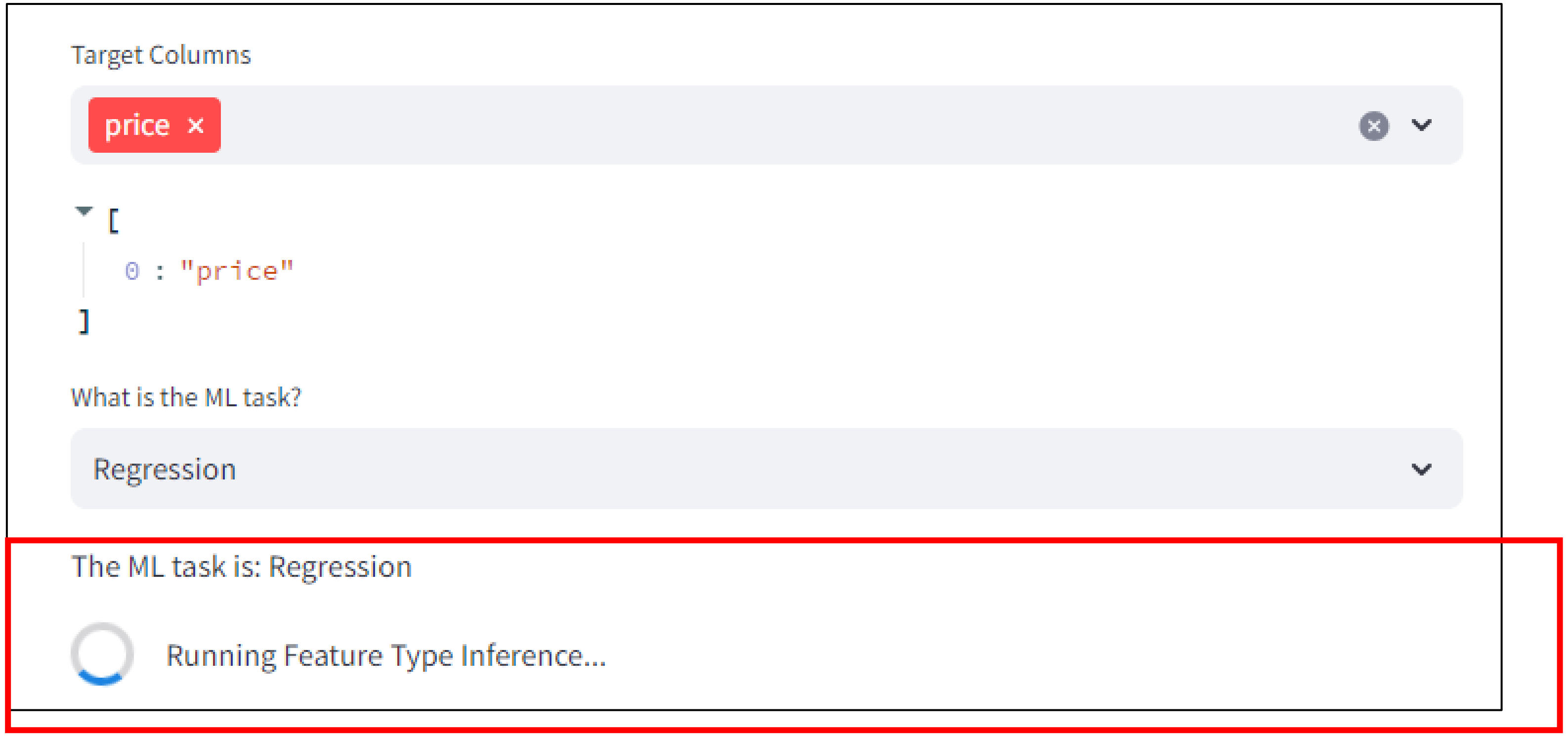
```
[
  0 : "price"
]
```

What is the ML task?

Regression

- Select ML tasks, the two options are:
  - Classification: for ML tasks that predict discrete class labels
  - Regression: for ML tasks that predict a continuous quantity

2.3 Feature Type Inference (1)



Target Columns

price ×

```
[
  0 : "price"
]
```

What is the ML task?

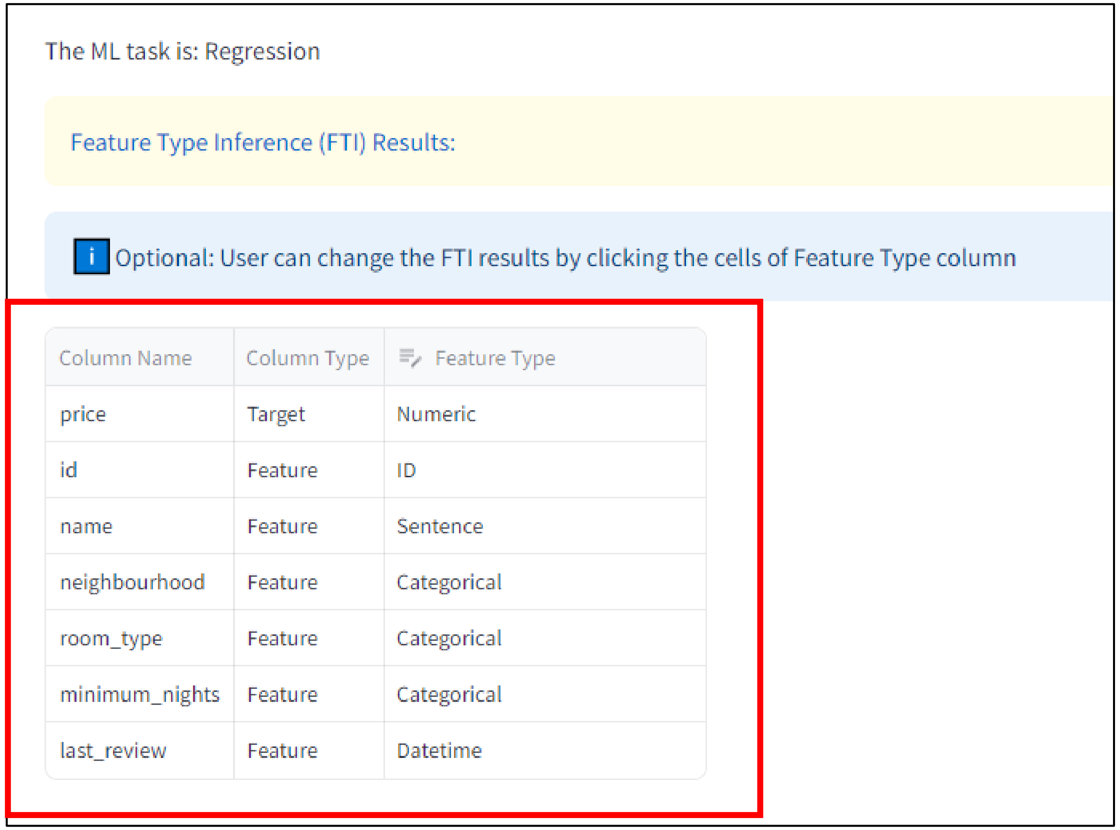
Regression

The ML task is: Regression

Running Feature Type Inference...

- After the Target columns and ML task are specified, feature type inference (FTI) is automatically executed for the uploaded dataset, which predicts the feature type for each columns of the dataset.

2.3 Feature Type Inference (2)



The ML task is: Regression

Feature Type Inference (FTI) Results:

i Optional: User can change the FTI results by clicking the cells of Feature Type column

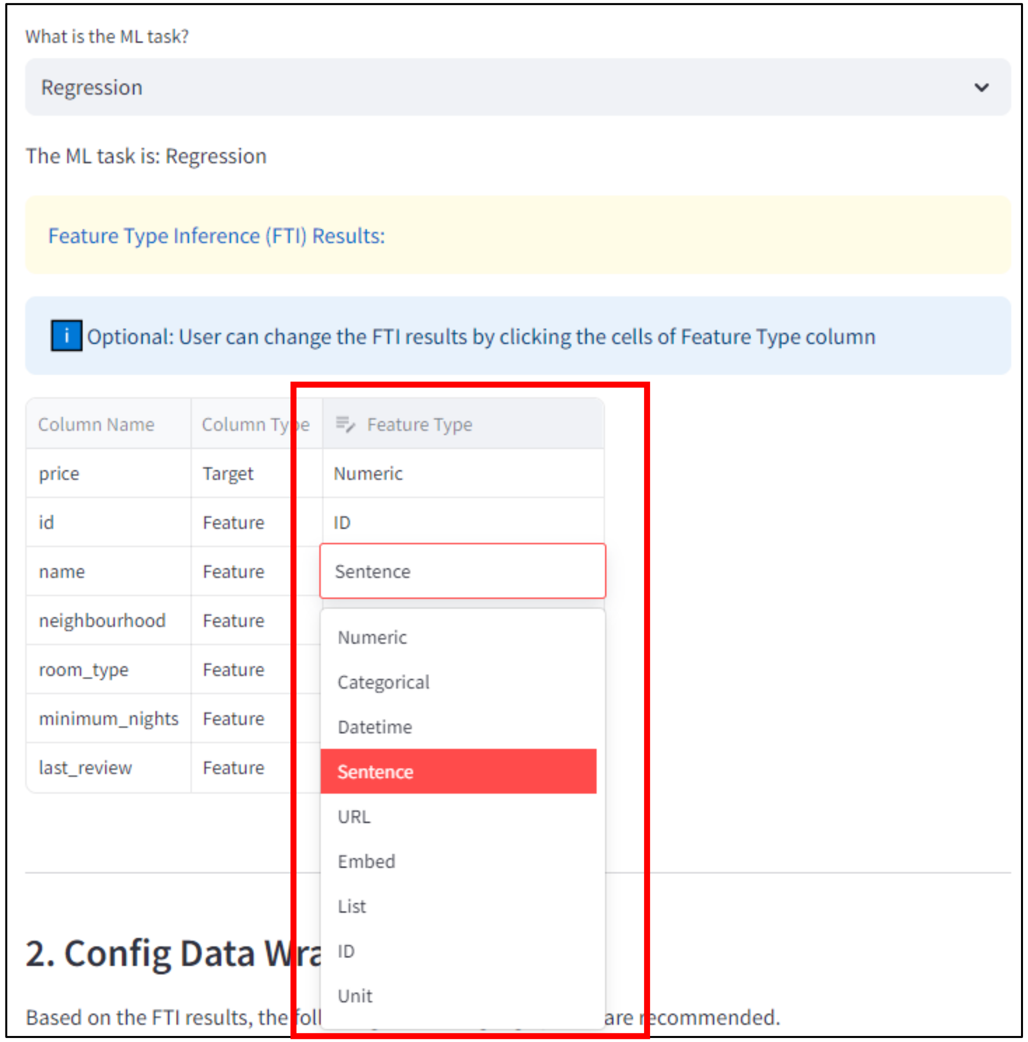
Column Name	Column Type	Feature Type
price	Target	Numeric
id	Feature	ID
name	Feature	Sentence
neighbourhood	Feature	Categorical
room_type	Feature	Categorical
minimum_nights	Feature	Categorical
last_review	Feature	Datetime

- After the FTI prediction is completed, a table with three columns is displayed:
  - Column Name: shows the name of each column in the uploaded dataset
  - Column Type: shows a column is a target or a feature
  - Feature Type: shows the predicted feature type for each column

2.3 Feature Type Inference (3)

- Feature Types
  - **Numeric**: quantitative data, such as 1, 2, 3, ⋯
  - **Categorical**: a variable with a set number of groups, such as male, female
  - **Datetime**: dates and times, such as 11-23-2022, 15:20PM, etc.
  - **Sentence**: a set of words, such as “auto data wrangling tools are useful”, etc.
  - **URL**: Uniform Resource Locator, such as https://www.fujitsu.com/global/about/research/
  - **Embed**: A string with numerical value embedded, such as $1,000
  - **List**: A sequence of several variables, grouped together, such as [A, B, C], [1, 2, 3], etc.
  - **ID**: An identity column, such as the index
  - **Unit**: a determinate quantity as a standard of measurement, such as 100m, 60kg, etc.
  - **Sign**: A string with symbols such as >=, <= with numerical values, examples are >50, <=100, etc.
  - **Range**: A string shows a range of numerical values, such as 60-100

Edit FTI Results (Optional)



What is the ML task?

Regression

The ML task is: Regression

Feature Type Inference (FTI) Results:

i Optional: User can change the FTI results by clicking the cells of Feature Type column

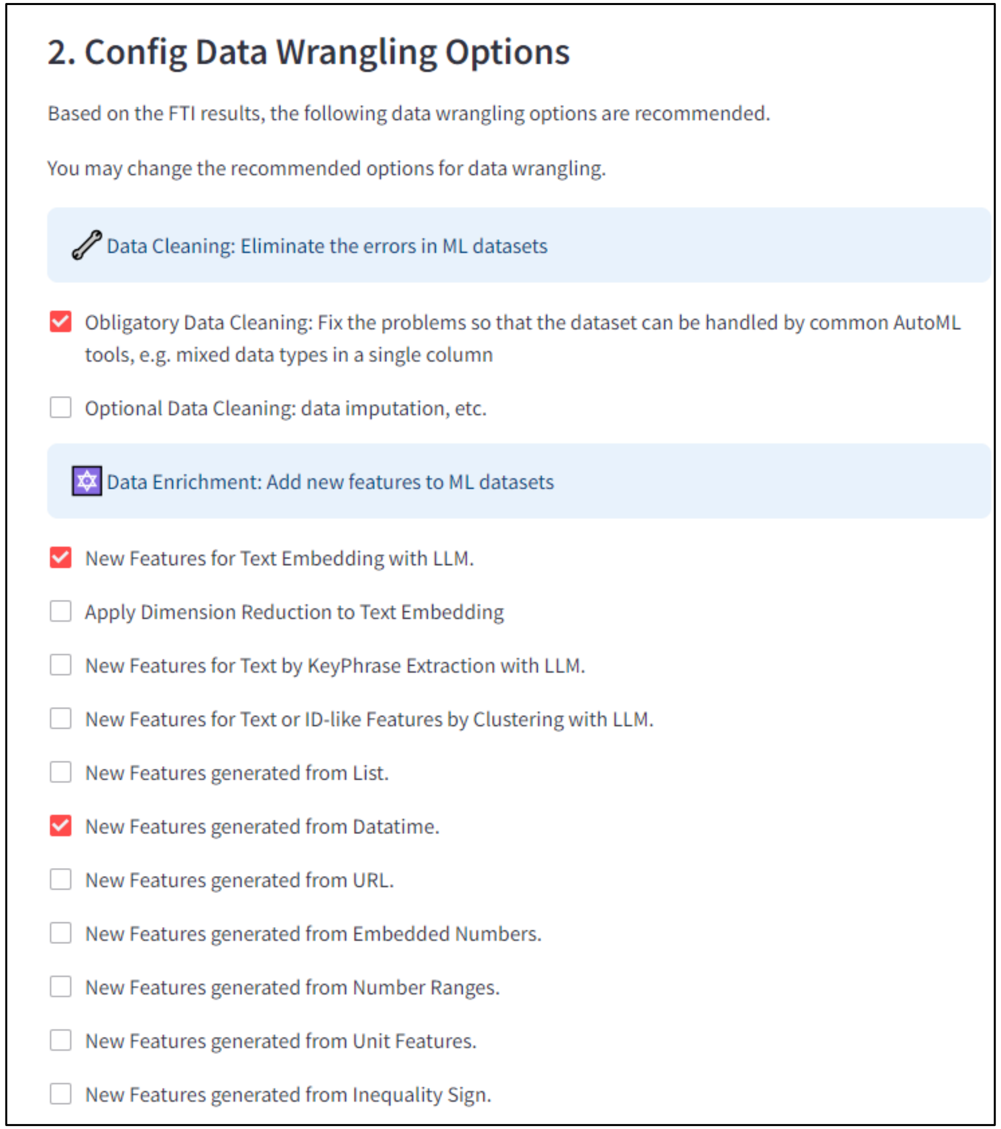
Column Name	Column Type	Feature Type
price	Target	Numeric
id	Feature	ID
name	Feature	Sentence
neighbourhood	Feature	
room_type	Feature	
minimum_nights	Feature	
last_review	Feature	

2. Config Data Wra

Based on the FTI results, the foll are recommended.

- The "Feature Type" column in the FTI results is editable
- If user feels some of the FTI results is not satisfied, you can click the cell to change the predicted FTI type
- The following processing will be based on the feature type that user selected

2.4 Config data wrangling options (1)



2. Config Data Wrangling Options

Based on the FTI results, the following data wrangling options are recommended.

You may change the recommended options for data wrangling.

Data Cleaning: Eliminate the errors in ML datasets

☑ Obligatory Data Cleaning: Fix the problems so that the dataset can be handled by common AutoML tools, e.g. mixed data types in a single column

□ Optional Data Cleaning: data imputation, etc.

Data Enrichment: Add new features to ML datasets

☑ New Features for Text Embedding with LLM.

□ Apply Dimension Reduction to Text Embedding

□ New Features for Text by KeyPhrase Extraction with LLM.

□ New Features for Text or ID-like Features by Clustering with LLM.

□ New Features generated from List.

☑ New Features generated from Datatime.

□ New Features generated from URL.

□ New Features generated from Embedded Numbers.

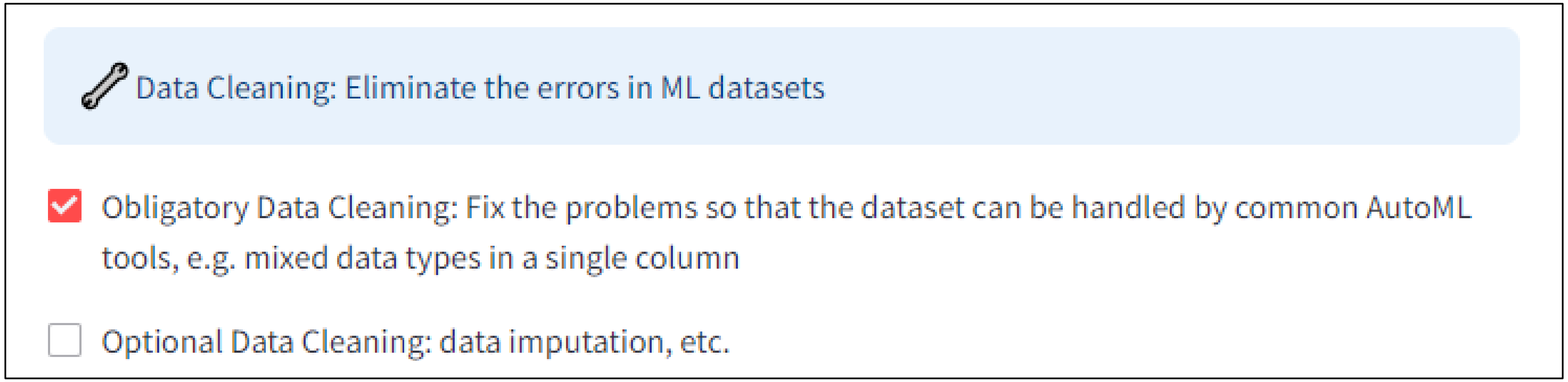
□ New Features generated from Number Ranges.

□ New Features generated from Unit Features.

□ New Features generated from Inequality Sign.

- Based on the FTI results, certain data wrangling options are suggested automatically.
- Nonetheless, users have the flexibility to modify these options according to their requirements.
- Two major data wrangling options:
  - Data Cleaning: eliminate the errors in the dataset
  - Data enrichment: add new features to the dataset

2.4 Config data wrangling options (2)



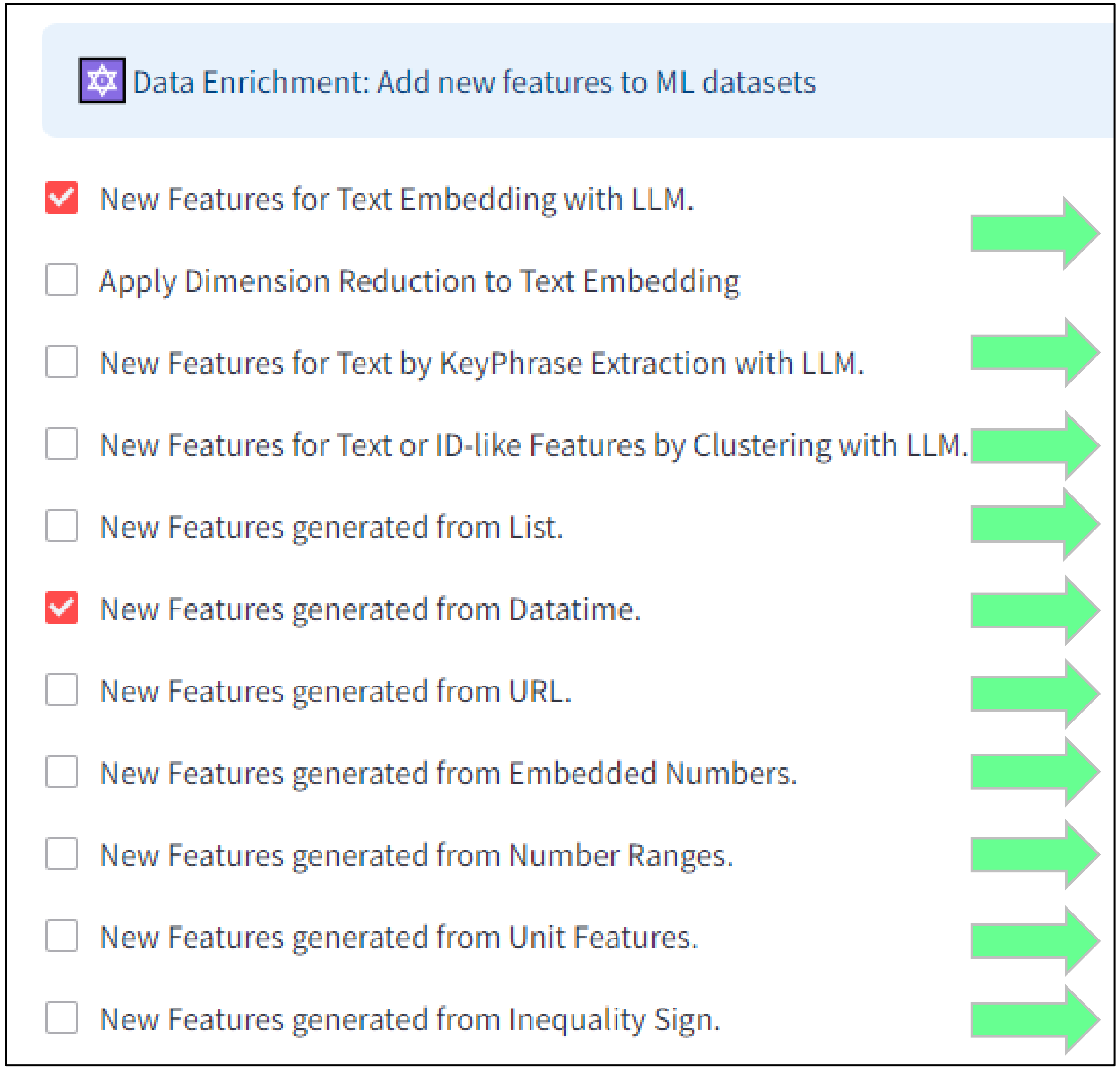
Data Cleaning: Eliminate the errors in ML datasets

☑ Obligatory Data Cleaning: Fix the problems so that the dataset can be handled by common AutoML tools, e.g. mixed data types in a single column

☐ Optional Data Cleaning: data imputation, etc.

- Data cleaning has two options:
  - **Obligatory data cleaning** is essential to address issues in the dataset, enabling seamless compatibility with popular AutoML tools. This process involves checking for potential errors, such as decoding datasets when necessary, refining headers, eliminating irrelevant features, addressing NaN cells in the target, handling infinite values, cleaning columns with mixed data types, and cleaning text columns, etc.
  - **Optional data cleaning** serves as an additional layer on the obligatory data cleaning results to ensure the dataset is fully prepared for AutoML tools. This step may involve operations such as data imputation and encoding the target column. These operations are considered optional, as many common AutoML tools are capable of performing them as well.

2.4 Config data wrangling options (3)



Data Enrichment: Add new features to ML datasets

- ☑ New Features for Text Embedding with LLM.
- □ Apply Dimension Reduction to Text Embedding
- □ New Features for Text by KeyPhrase Extraction with LLM.
- □ New Features for Text or ID-like Features by Clustering with LLM.
- □ New Features generated from List.
- ☑ New Features generated from Datatime.
- □ New Features generated from URL.
- □ New Features generated from Embedded Numbers.
- □ New Features generated from Number Ranges.
- □ New Features generated from Unit Features.
- □ New Features generated from Inequality Sign.

- Data enrichment based on FTI

Add new text embedding features using LLM for the ”Sentence” features (more details in the next page)

Add new keyphrase features for the ”Sentence” features using LLM

Add new clustering features for the ”Sentence” features using LLM

Add new features for the ”List” features

Add new features for the ”Datetime” features

Add new features for the ”URL” features

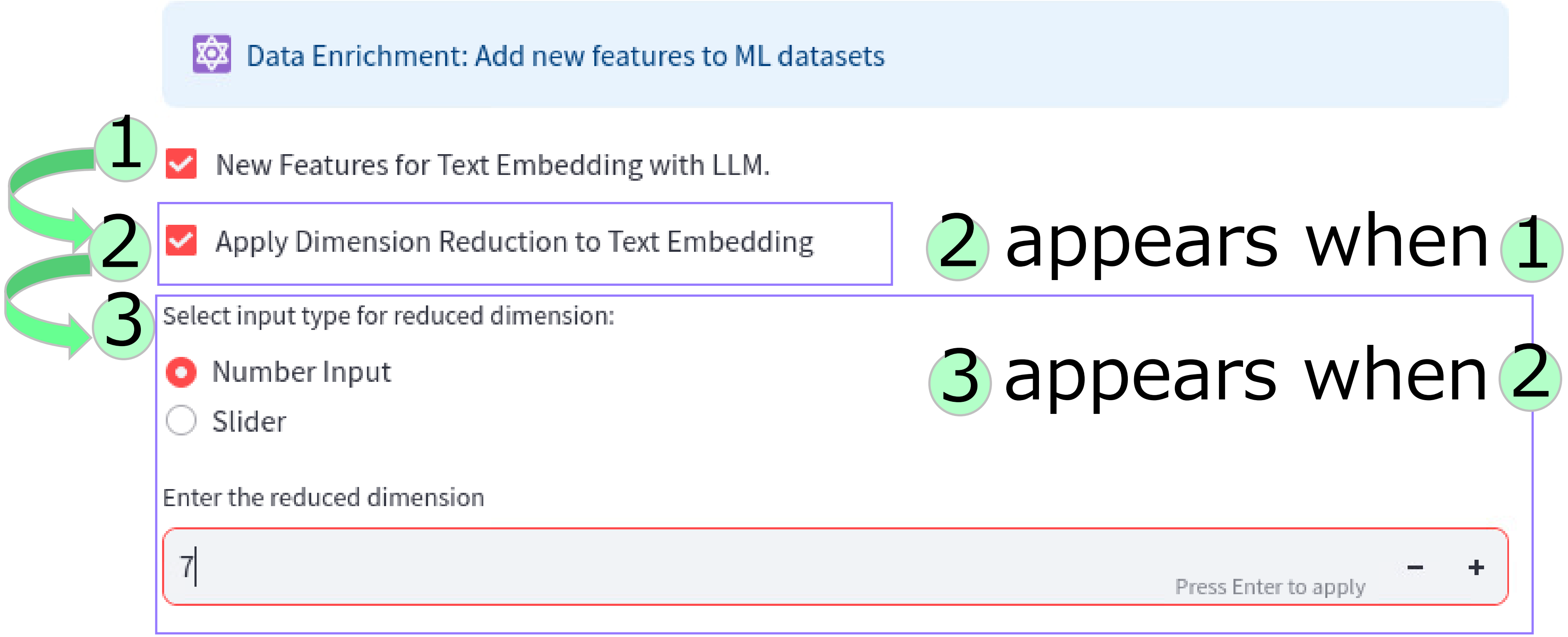
Add new features for the ”Embed” features

Add new features for the ”Range” features

Add new features for the ”Unit” features

Add new features for the ”Sign” features

2.4 Config data wrangling options (4)



Data Enrichment: Add new features to ML datasets

1 ☑ New Features for Text Embedding with LLM.

2 ☑ Apply Dimension Reduction to Text Embedding

3 Select input type for reduced dimension:

● Number Input

○ Slider

Enter the reduced dimension

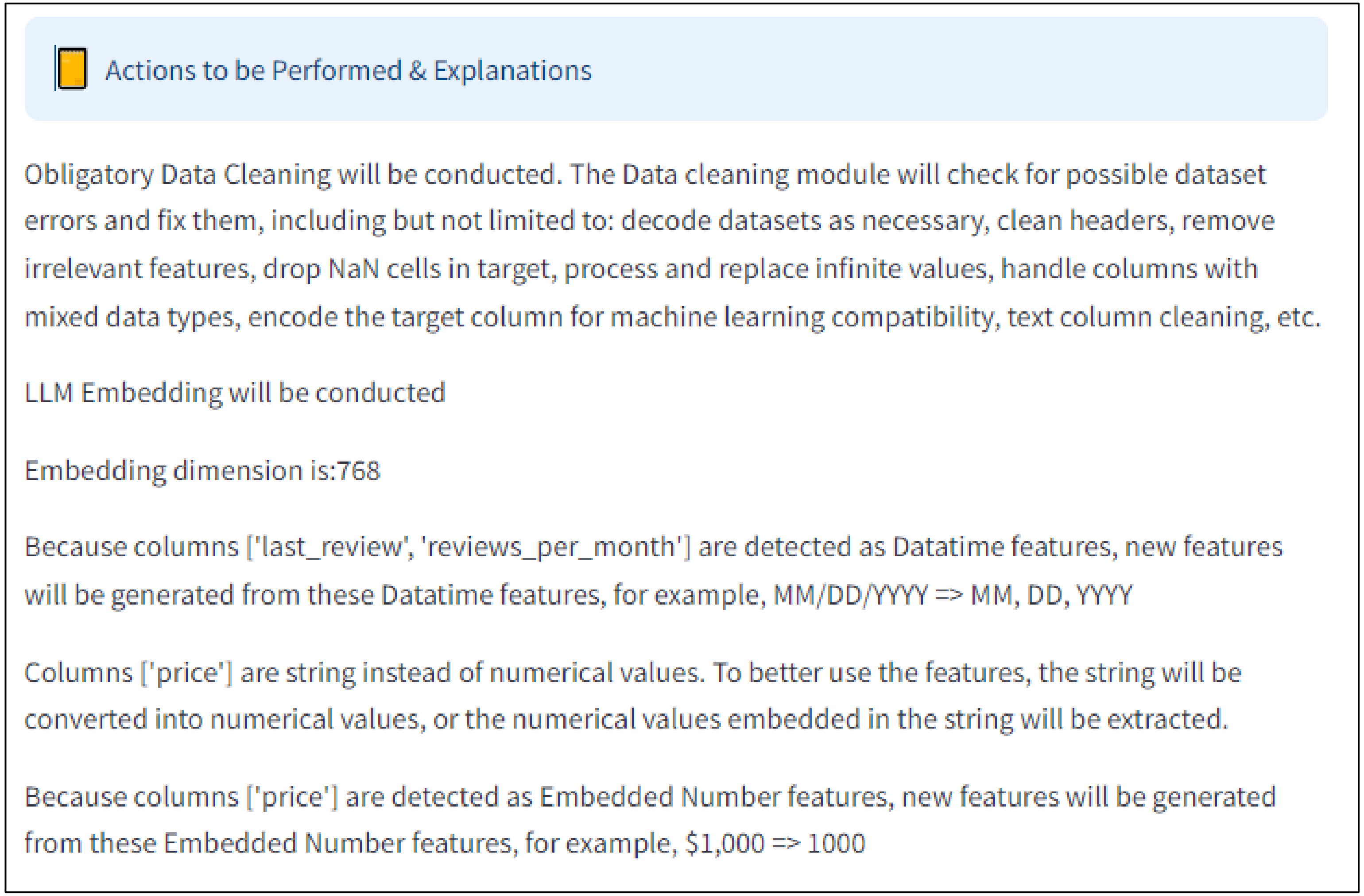
7 Press Enter to apply − +

2 appears when 1 is checked

3 appears when 2 is checked

- Text embedding configuration
  - You can apply dimension reduction to text embedding by checking 2
  - You can configure the number of dimensions of text embedding (from 1 to 768) by using either a number input box or slider, as shown in 3

2.4 Config data wrangling options (5)



Actions to be Performed & Explanations

Obligatory Data Cleaning will be conducted. The Data cleaning module will check for possible dataset errors and fix them, including but not limited to: decode datasets as necessary, clean headers, remove irrelevant features, drop NaN cells in target, process and replace infinite values, handle columns with mixed data types, encode the target column for machine learning compatibility, text column cleaning, etc.

LLM Embedding will be conducted

Embedding dimension is:768

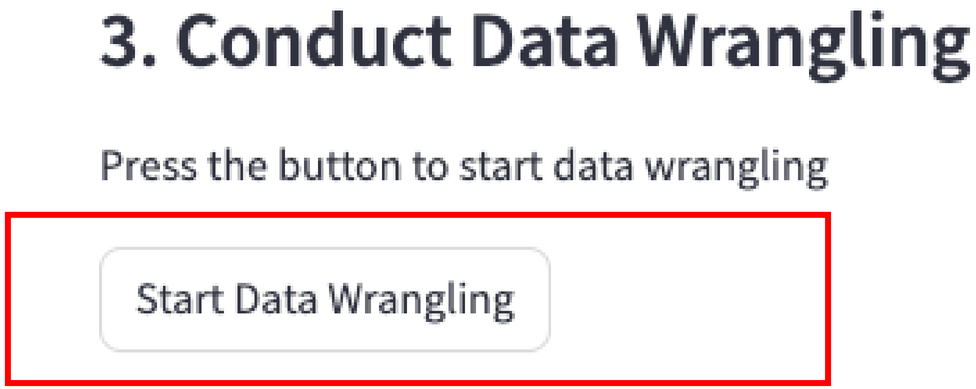
Because columns ['last_review', 'reviews_per_month'] are detected as Datatime features, new features will be generated from these Datatime features, for example, MM/DD/YYYY => MM, DD, YYYY

Columns ['price'] are string instead of numerical values. To better use the features, the string will be converted into numerical values, or the numerical values embedded in the string will be extracted.

Because columns ['price'] are detected as Embedded Number features, new features will be generated from these Embedded Number features, for example, $1,000 => 1000

- The data wrangling actions are summarized, explained, and are displayed in UI.

2.5 Data Wrangling Execution & Results (1)

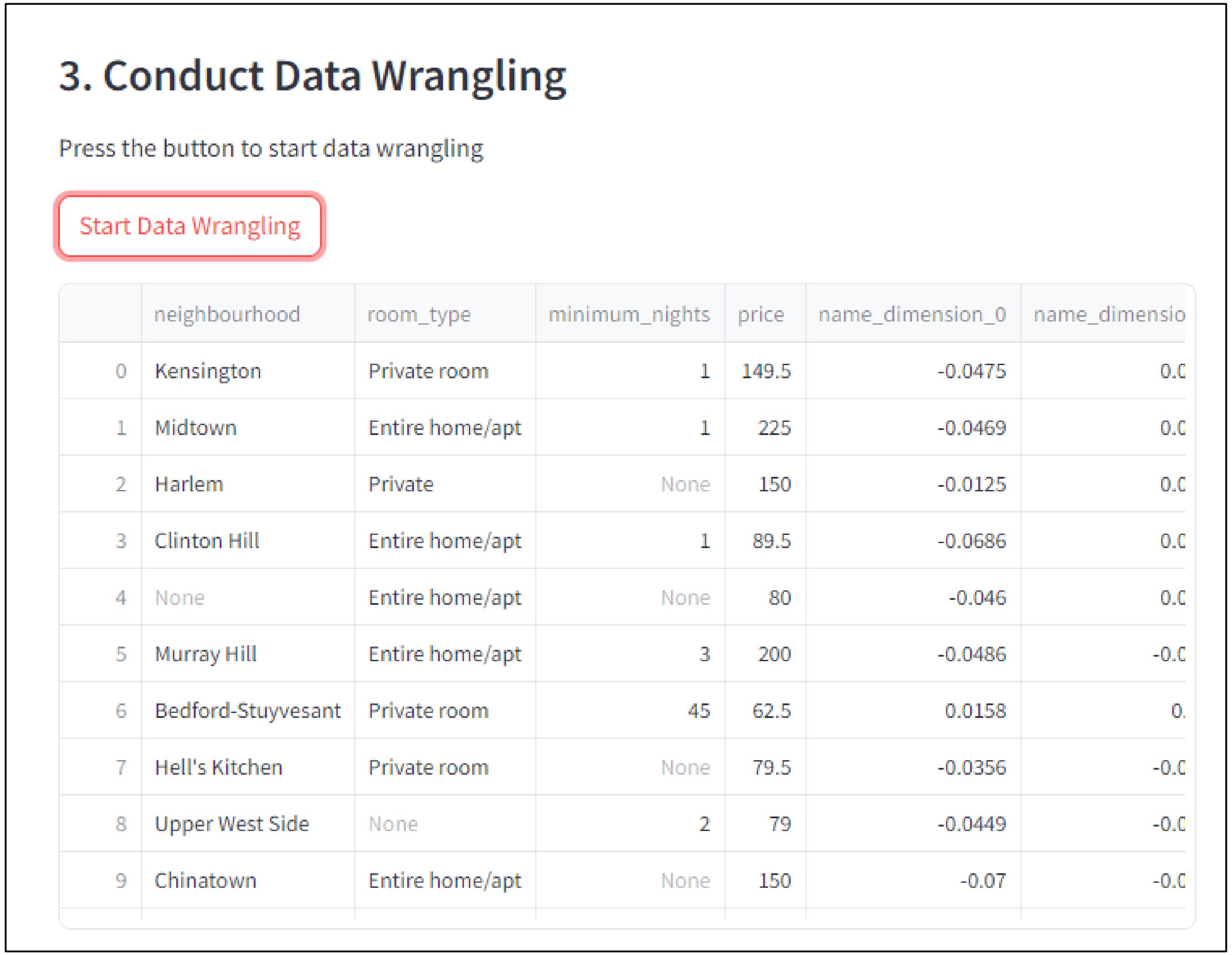


3. Conduct Data Wrangling

Press the button to start data wrangling

- After the data wrangling options are configured, click the button “Start Data Wrangling” to start the data cleaning and enrichment processing based on the configurations

2.5 Data Wrangling Execution & Results (2)



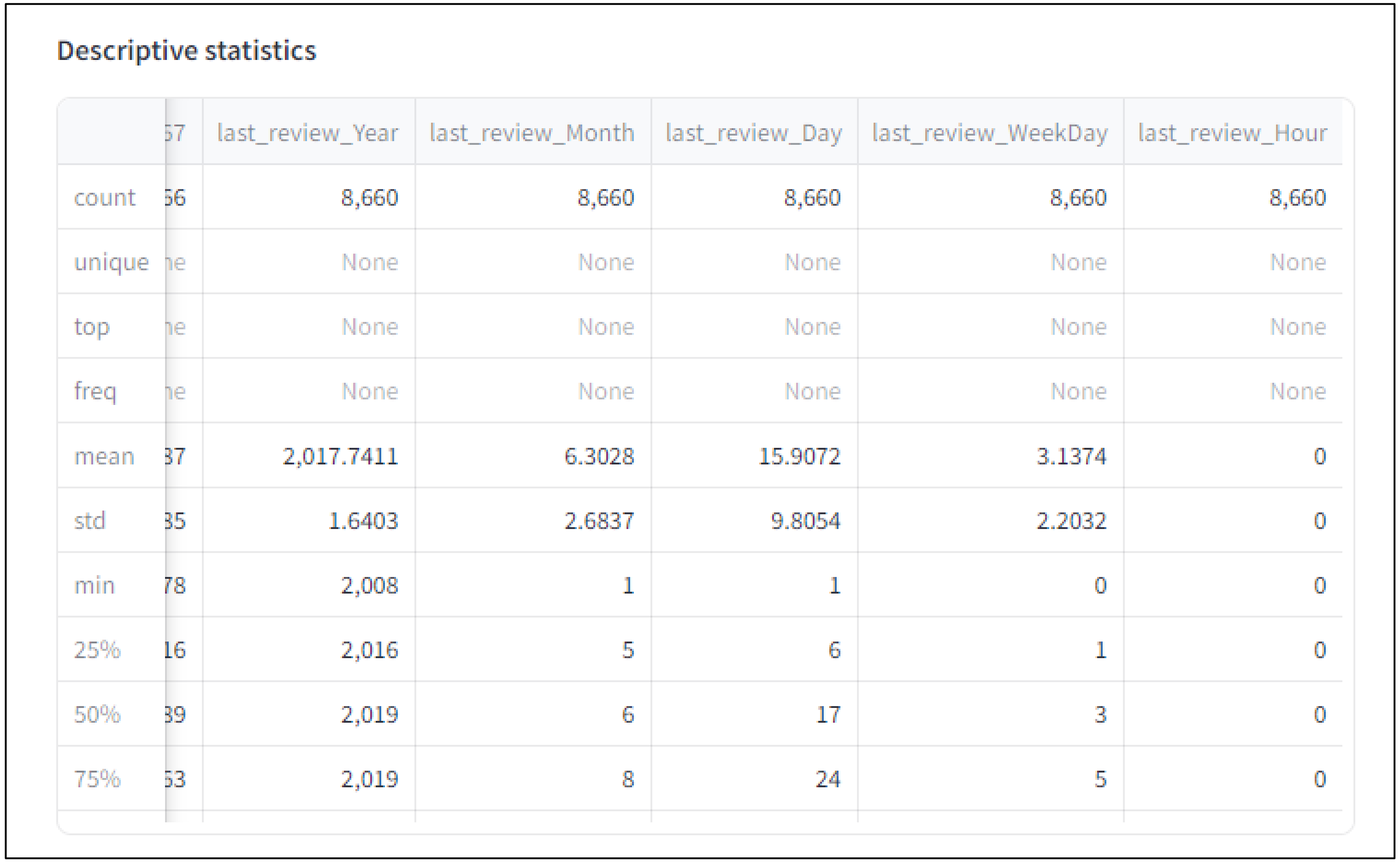
3. Conduct Data Wrangling

Press the button to start data wrangling

	neighbourhood	room_type	minimum_nights	price	name_dimension_0	name_dimensio
0	Kensington	Private room	1	149.5	-0.0475	0.0
1	Midtown	Entire home/apt	1	225	-0.0469	0.0
2	Harlem	Private	None	150	-0.0125	0.0
3	Clinton Hill	Entire home/apt	1	89.5	-0.0686	0.0
4	None	Entire home/apt	None	80	-0.046	0.0
5	Murray Hill	Entire home/apt	3	200	-0.0486	-0.0
6	Bedford-Stuyvesant	Private room	45	62.5	0.0158	0.
7	Hell's Kitchen	Private room	None	79.5	-0.0356	-0.0
8	Upper West Side	None	2	79	-0.0449	-0.0
9	Chinatown	Entire home/apt	None	150	-0.07	-0.0

- After the data wrangling processing is done, the cleaned & enriched dataset is displayed

2.5 Data Wrangling Execution & Results (3)

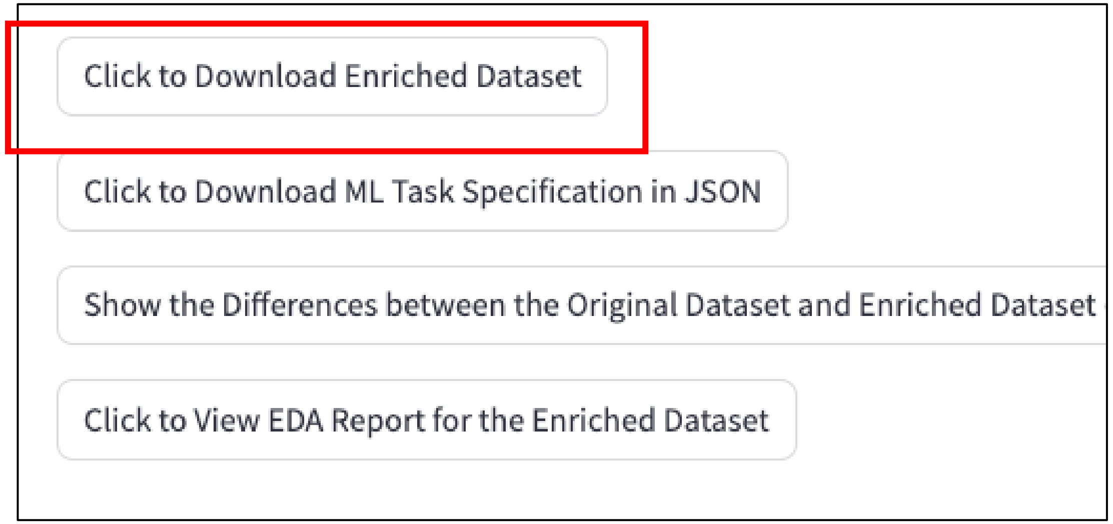


Descriptive statistics

	57	last_review_Year	last_review_Month	last_review_Day	last_review_WeekDay	last_review_Hour
count	56	8,660	8,660	8,660	8,660	8,660
unique	ne	None	None	None	None	None
top	ne	None	None	None	None	None
freq	ne	None	None	None	None	None
mean	37	2,017.7411	6.3028	15.9072	3.1374	0
std	35	1.6403	2.6837	9.8054	2.2032	0
min	78	2,008	1	1	0	0
25%	16	2,016	5	6	1	0
50%	39	2,019	6	17	3	0
75%	53	2,019	8	24	5	0

- The descriptive statistics for the cleaned & enriched dataset is also displayed

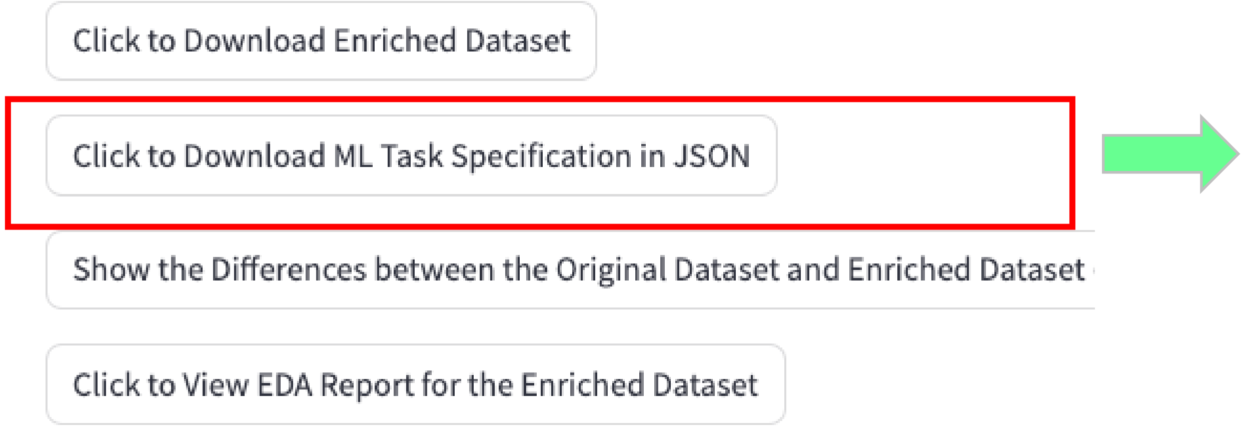
2.5 Data Wrangling Execution & Results (4)

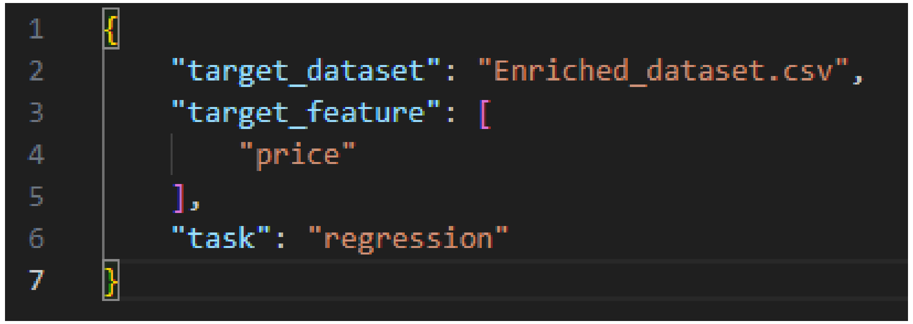


Click to Download Enriched Dataset
Click to Download ML Task Specification in JSON
Show the Differences between the Original Dataset and Enriched Dataset
Click to View EDA Report for the Enriched Dataset

- Click the button “Click to Download Enriched Dataset” to download the cleaned & enriched dataset to your local machine. The downloaded dataset is in CSV format and can be further used for other applications such as AutoML.

2.6 Data Wrangling Execution & Results (5)

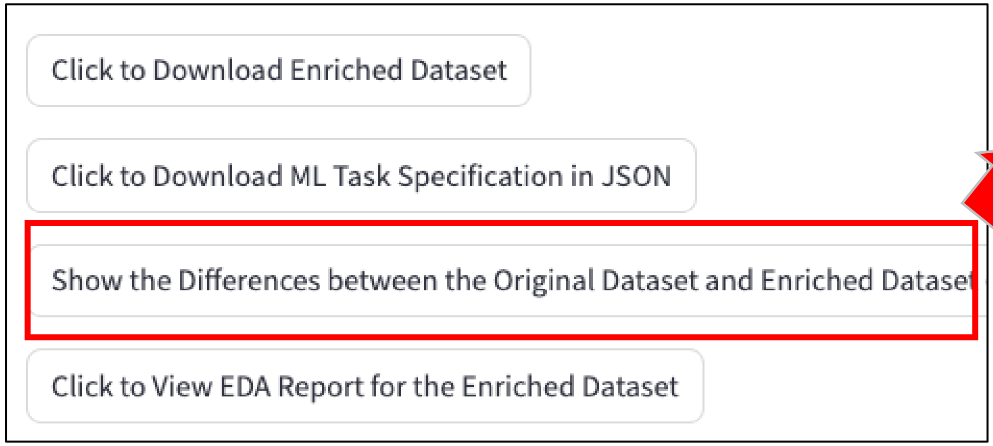

Click to Download Enriched Dataset
Click to Download ML Task Specification in JSON
Show the Differences between the Original Dataset and Enriched Dataset
Click to View EDA Report for the Enriched Dataset



```
{
    "target_dataset": "Enriched_dataset.csv",
    "target_feature": [
        "price"
    ],
    "task": "regression"
}
```

○Click the button “Click to Download ML Task Specification” to download the ML task specification for the cleaned & enriched dataset to your local machine. The ML task specification is in JSON format and can be further used for other applications such as AutoML.

2.6 Data Wrangling Execution & Results (6)

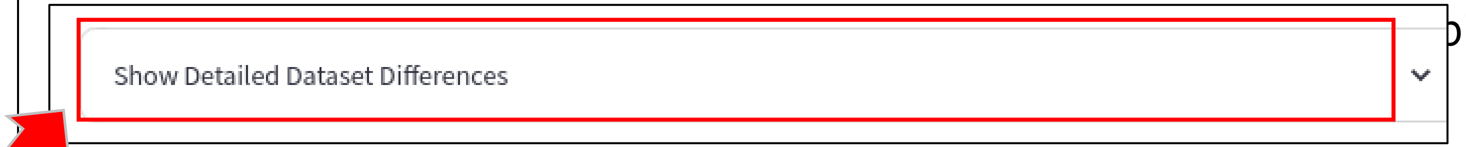

Click to Download Enriched Dataset

Click to Download ML Task Specification in JSON

Show the Differences between the Original Dataset and Enriched Dataset

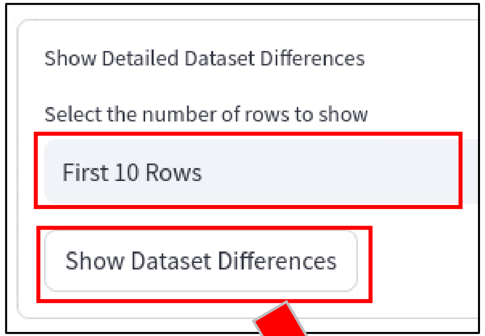
Click to View EDA Report for the Enriched Dataset

- Click the above button to
  - Show the differences (highlighted in yellow color) between the original dataset and enrich dataset
    - By default, only 10 rows are displayed, but you can configure it by clicking “select the number of rows to show”
  - Explain the reason if columns are dropped


Show Detailed Dataset Differences

bility

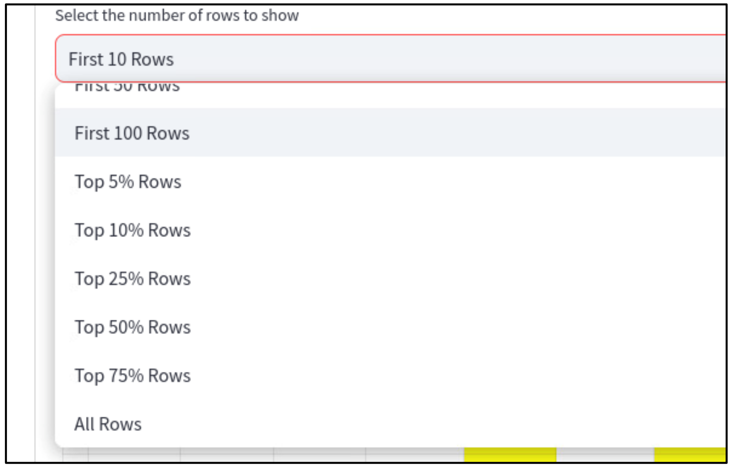
展開


Show Detailed Dataset Differences

Select the number of rows to show

First 10 Rows

Show Dataset Differences


Select the number of rows to show

First 10 Rows

First 50 Rows

First 100 Rows

Top 5% Rows

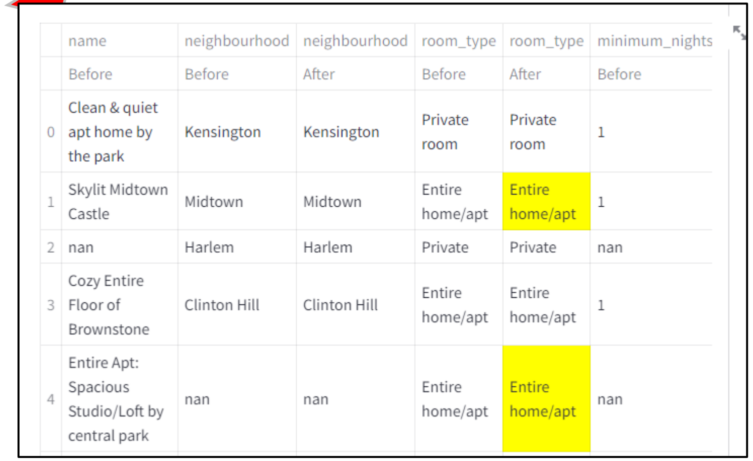
Top 10% Rows

Top 25% Rows

Top 50% Rows

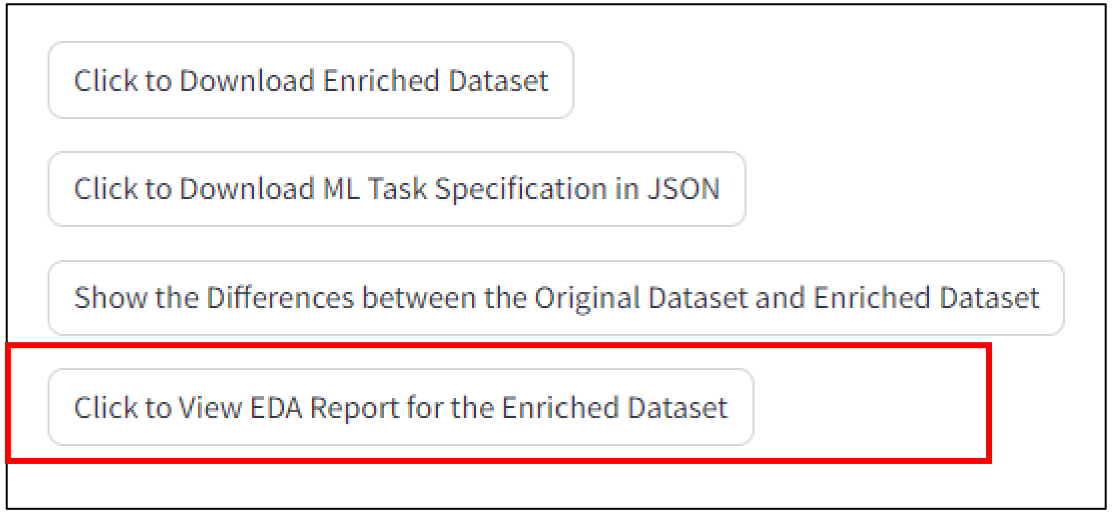
Top 75% Rows

All Rows

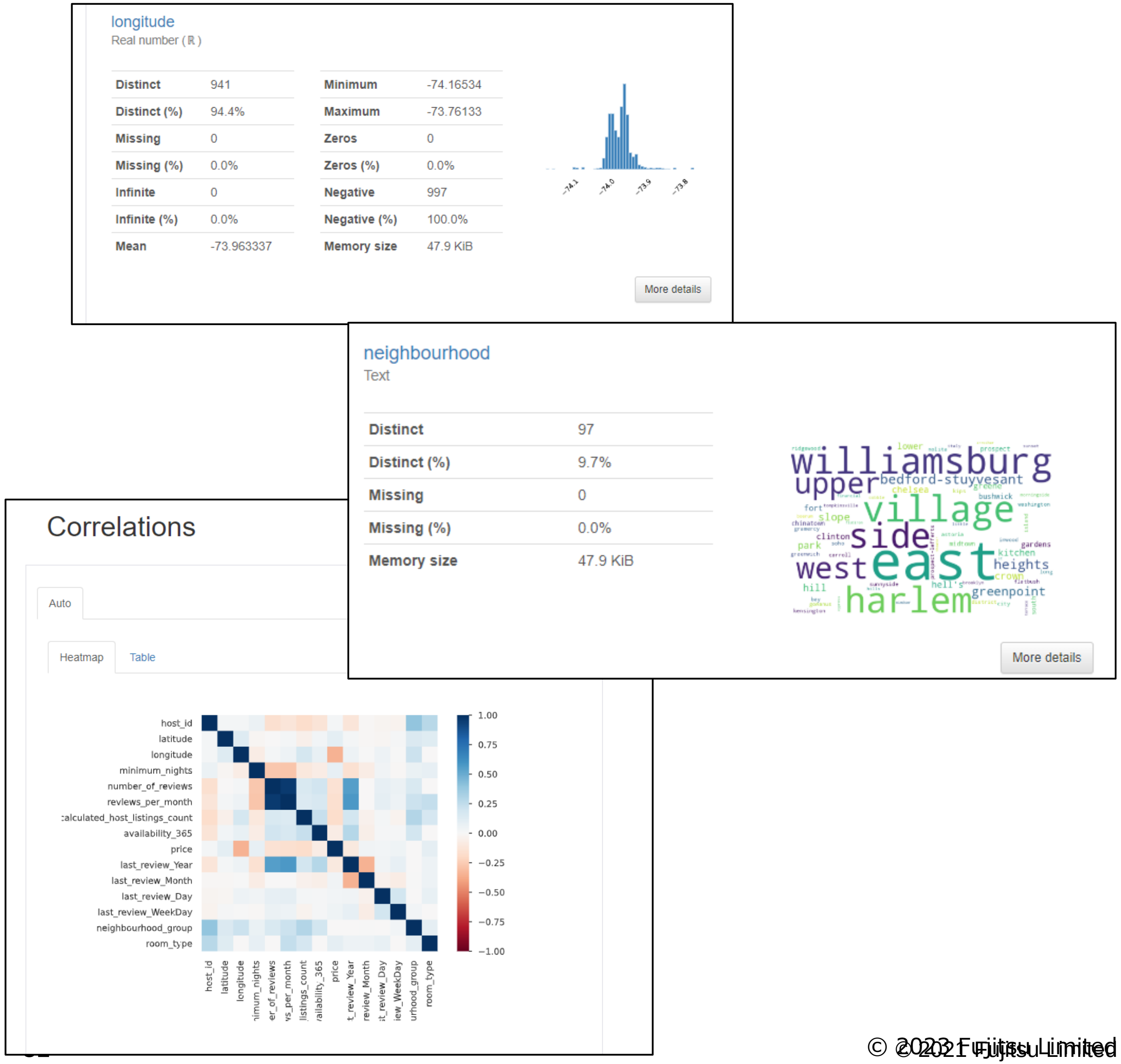


	name	neighbourhood	neighbourhood	room_type	room_type	minimum_nights
	Before	Before	After	Before	After	Before
0	Clean & quiet apt home by the park	Kensington	Kensington	Private room	Private room	1
1	Skylit Midtown Castle	Midtown	Midtown	Entire home/apt	Entire home/apt	1
2	nan	Harlem	Harlem	Private	Private	nan
3	Cozy Entire Floor of Brownstone	Clinton Hill	Clinton Hill	Entire home/apt	Entire home/apt	1
4	Entire Apt: Spacious Studio/Loft by central park	nan	nan	Entire home/apt	Entire home/apt	nan

2.6 Data Wrangling Execution & Results (7)


Click to Download Enriched Dataset
Click to Download ML Task Specification in JSON
Show the Differences between the Original Dataset and Enriched Dataset
Click to View EDA Report for the Enriched Dataset

- Click the button to generate the EDA report for the cleaned & enriched dataset
  - Note: When there are too many columns in the enriched dataset (e.g. if the LLM text embedding is applied), the generation of EDA report could be slow due to the interaction visualizations between every column pairs



longitude

Real number (ℝ)

Distinct	941	Minimum	-74.16534
Distinct (%)	94.4%	Maximum	-73.76133
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	997
Infinite (%)	0.0%	Negative (%)	100.0%
Mean	-73.963337	Memory size	47.9 KiB

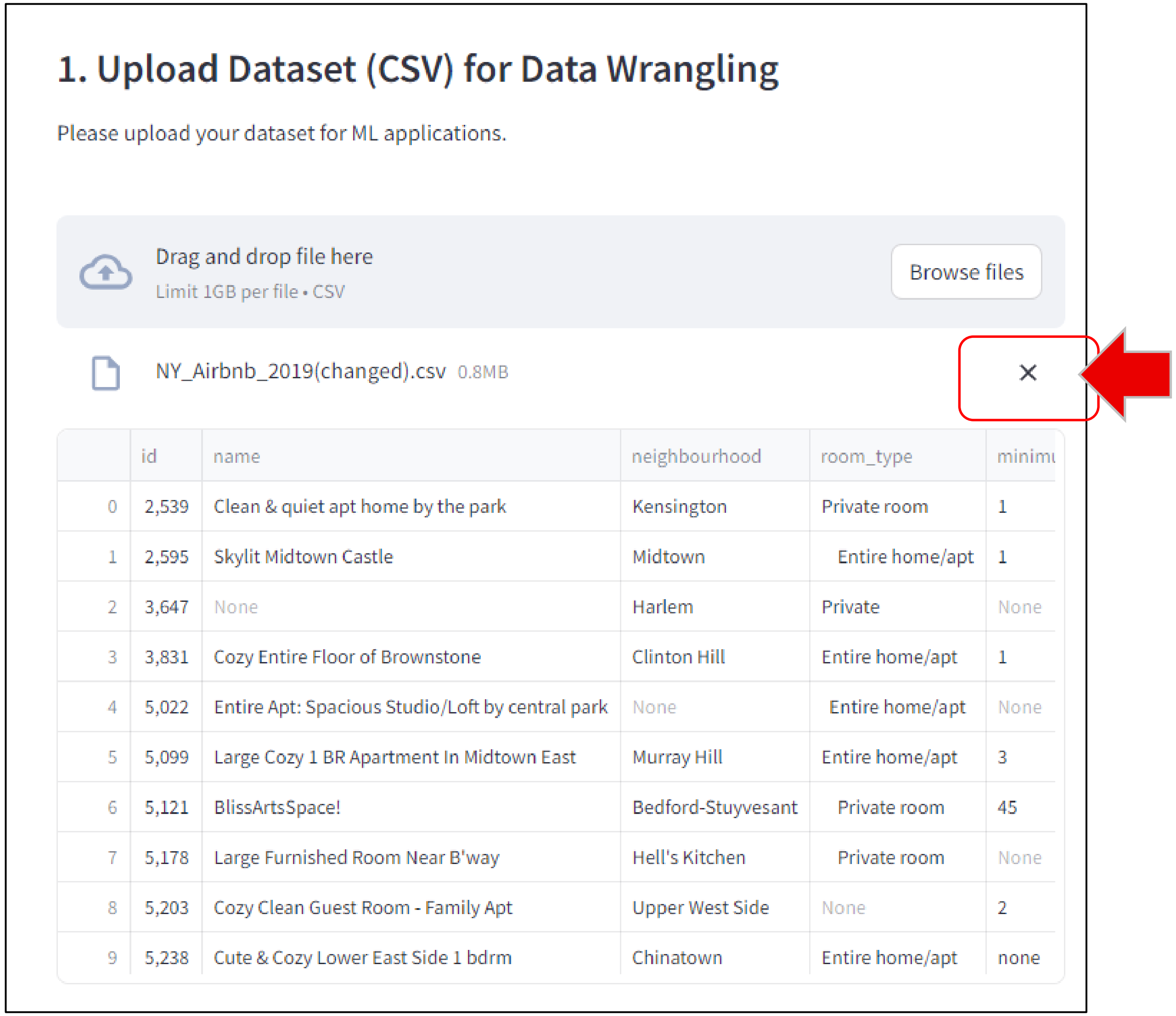
neighbourhood

Text

Distinct	97
Distinct (%)	9.7%
Missing	0
Missing (%)	0.0%
Memory size	47.9 KiB

Correlations

Try a Different Dataset



1. Upload Dataset (CSV) for Data Wrangling

Please upload your dataset for ML applications.

	id	name	neighbourhood	room_type	minimu
0	2,539	Clean & quiet apt home by the park	Kensington	Private room	1
1	2,595	Skylit Midtown Castle	Midtown	Entire home/apt	1
2	3,647	None	Harlem	Private	None
3	3,831	Cozy Entire Floor of Brownstone	Clinton Hill	Entire home/apt	1
4	5,022	Entire Apt: Spacious Studio/Loft by central park	None	Entire home/apt	None
5	5,099	Large Cozy 1 BR Apartment In Midtown East	Murray Hill	Entire home/apt	3
6	5,121	BlissArtsSpace!	Bedford-Stuyvesant	Private room	45
7	5,178	Large Furnished Room Near B'way	Hell's Kitchen	Private room	None
8	5,203	Cozy Clean Guest Room - Family Apt	Upper West Side	None	2
9	5,238	Cute & Cozy Lower East Side 1 bdrm	Chinatown	Entire home/apt	none

- If you complete the data wrangling for a dataset and want to try a different dataset, please click here to start the data wrangling for a new dataset.

Contact us (for customers)

- Please contact your Fujitsu representative.

Thank you