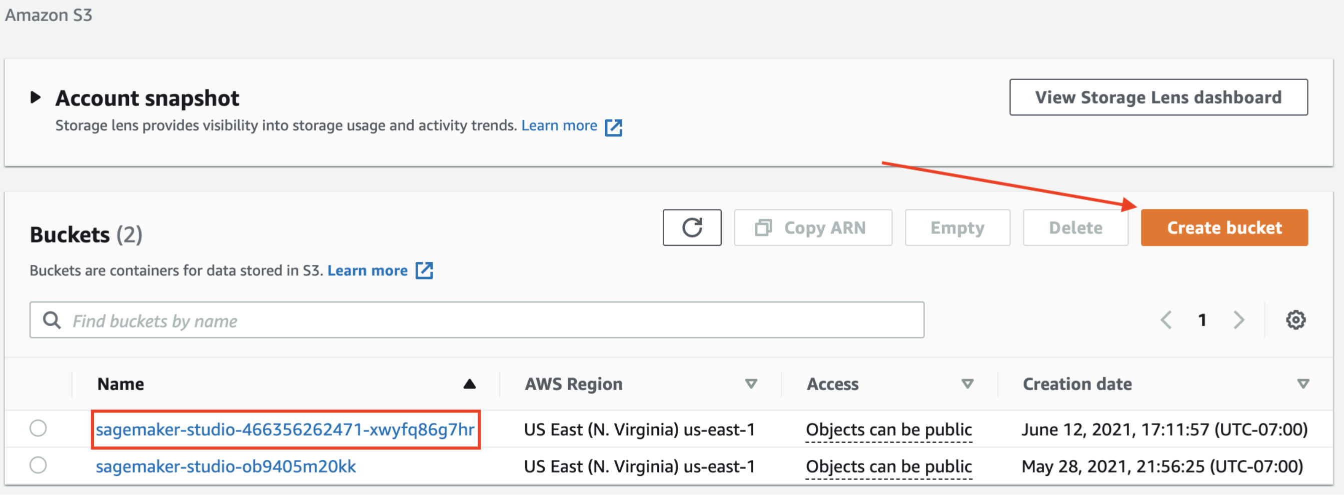
Solution: Data Wrangler

If you haven't done so already, you'll create an S3 bucket to upload files to and use throughout the course. Normally one will be created when you start using Sagemaker Studio, as pictured below.



Storage lens provides visibility into storage usage and activity trends. Learn more

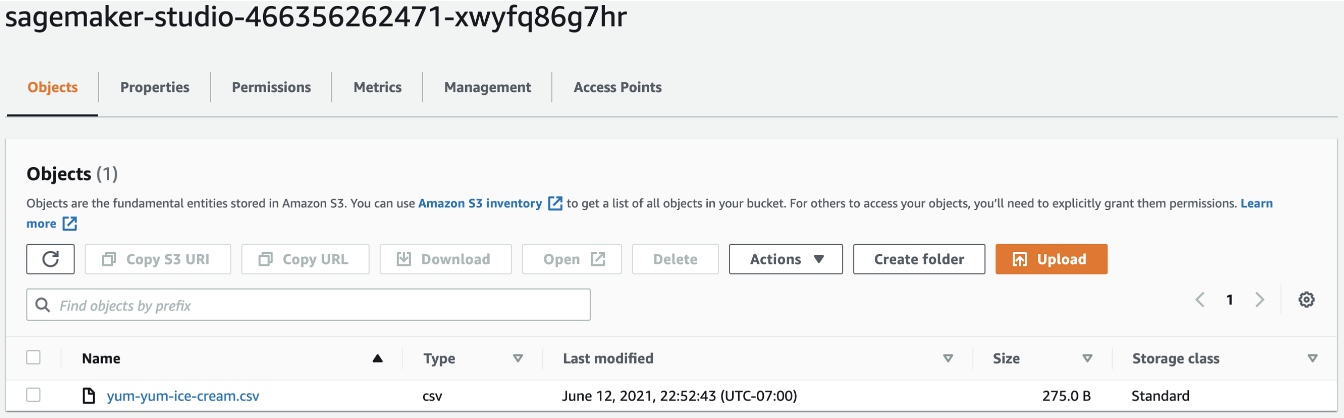
Buckets (2)

Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	sagemaker-studio-466356262471-xwyfq86g7hr	US East (N. Virginia) us-east-1	Objects can be public	June 12, 2021, 17:11:57 (UTC-07:00)
○	sagemaker-studio-ob9405m20kk	US East (N. Virginia) us-east-1	Objects can be public	May 28, 2021, 21:56:25 (UTC-07:00)

Choose or create a new bucket

Upload the `yum-yum-ice-cream.csv` file to your S3 bucket. The default settings that prompt you while uploading is acceptable.



sagemaker-studio-466356262471-xwyfq86g7hr

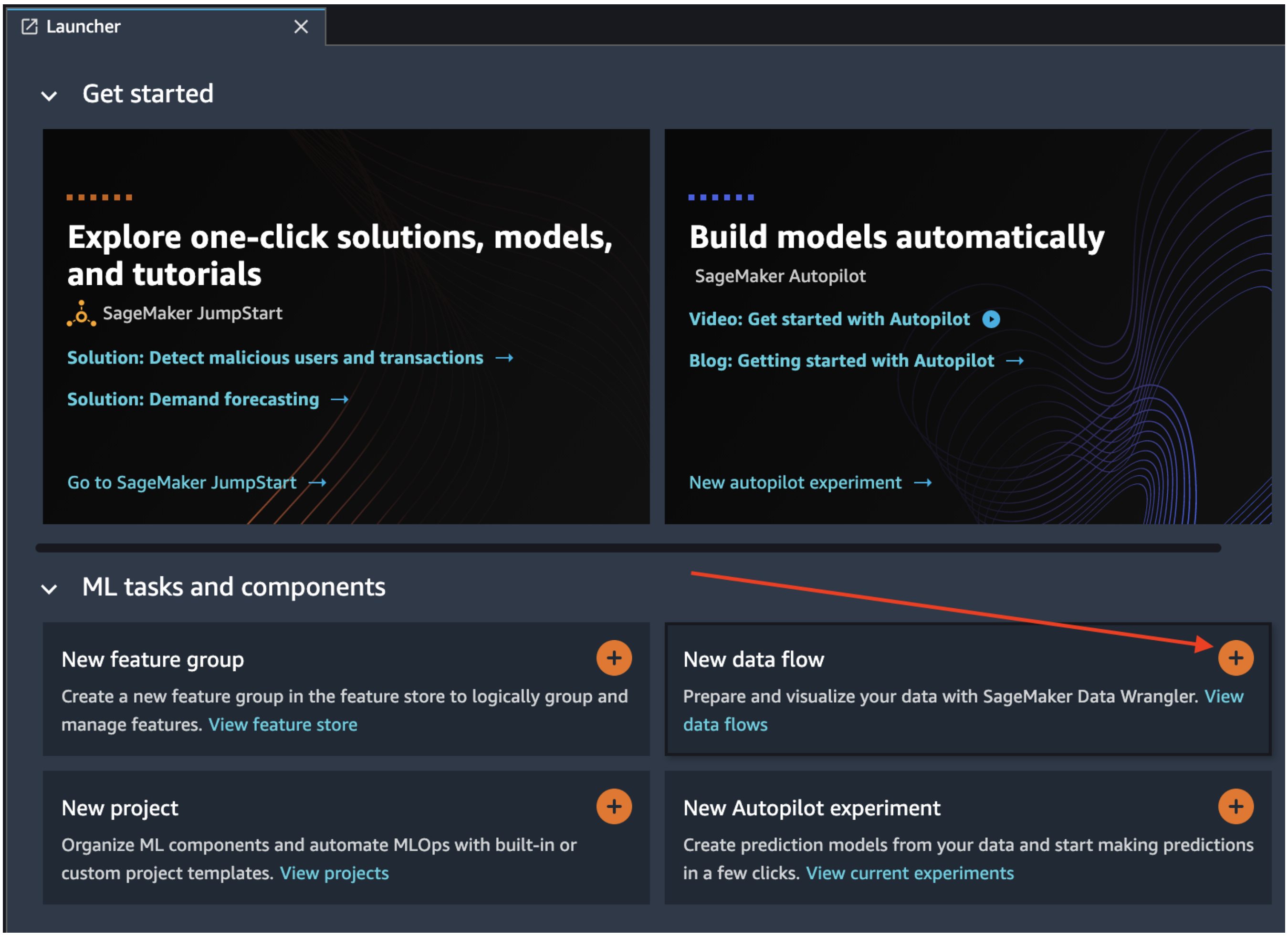
Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	yum-yum-ice-cream.csv	csv	June 12, 2021, 22:52:43 (UTC-07:00)	275.0 B	Standard

Upload File to Bucket

When in the Sagemaker Studio IDE, you'll see a display for creating a `New data flow` , clicking on the `+` start your new Data Wrangler data flow.



Launcher
Get started
Explore one-click solutions, models, and tutorials
SageMaker JumpStart
Solution: Detect malicious users and transactions →
Solution: Demand forecasting →
Go to SageMaker JumpStart →
Build models automatically
SageMaker Autopilot
Video: Get started with Autopilot
Blog: Getting started with Autopilot →
New autopilot experiment →
ML tasks and components

New feature group +

Create a new feature group in the feature store to logically group and manage features. View feature store

New data flow +

Prepare and visualize your data with SageMaker Data Wrangler. View data flows

New project +

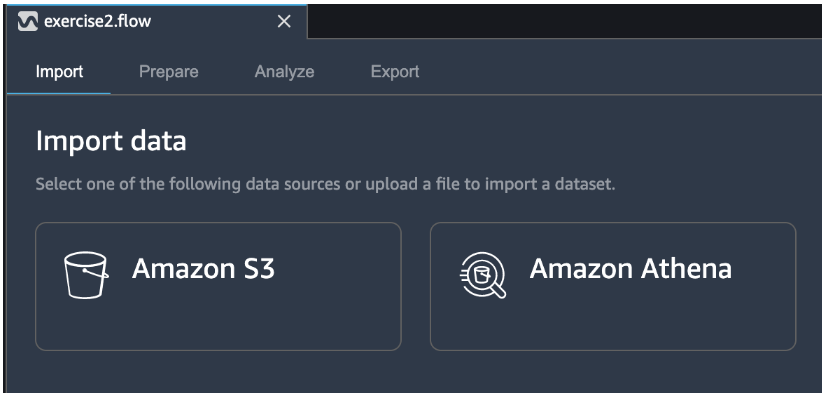
Organize ML components and automate MLOps with built-in or custom project templates. View projects

New Autopilot experiment +

Create prediction models from your data and start making predictions in a few clicks. View current experiments

Sagemaker Studio Data Wrangler

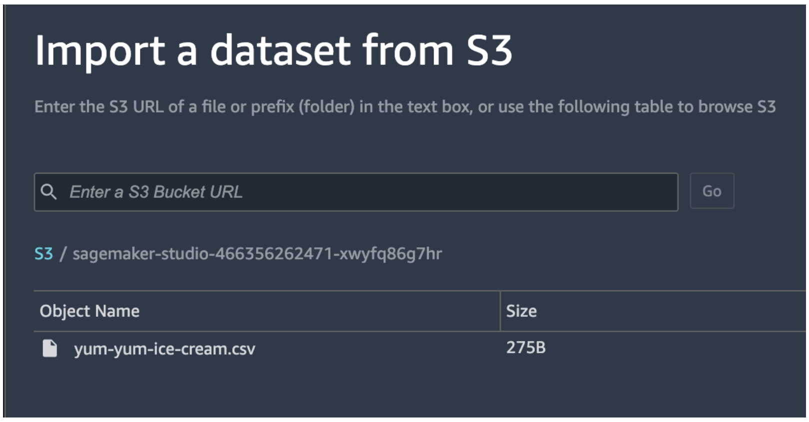
You already imported your file into S3, but in order to use it in Data Wrangler, you'll need to import it further from S3.



Import data

Select one of the following data sources or upload a file to import a dataset.

Data Wrangler S3


Import a dataset from S3

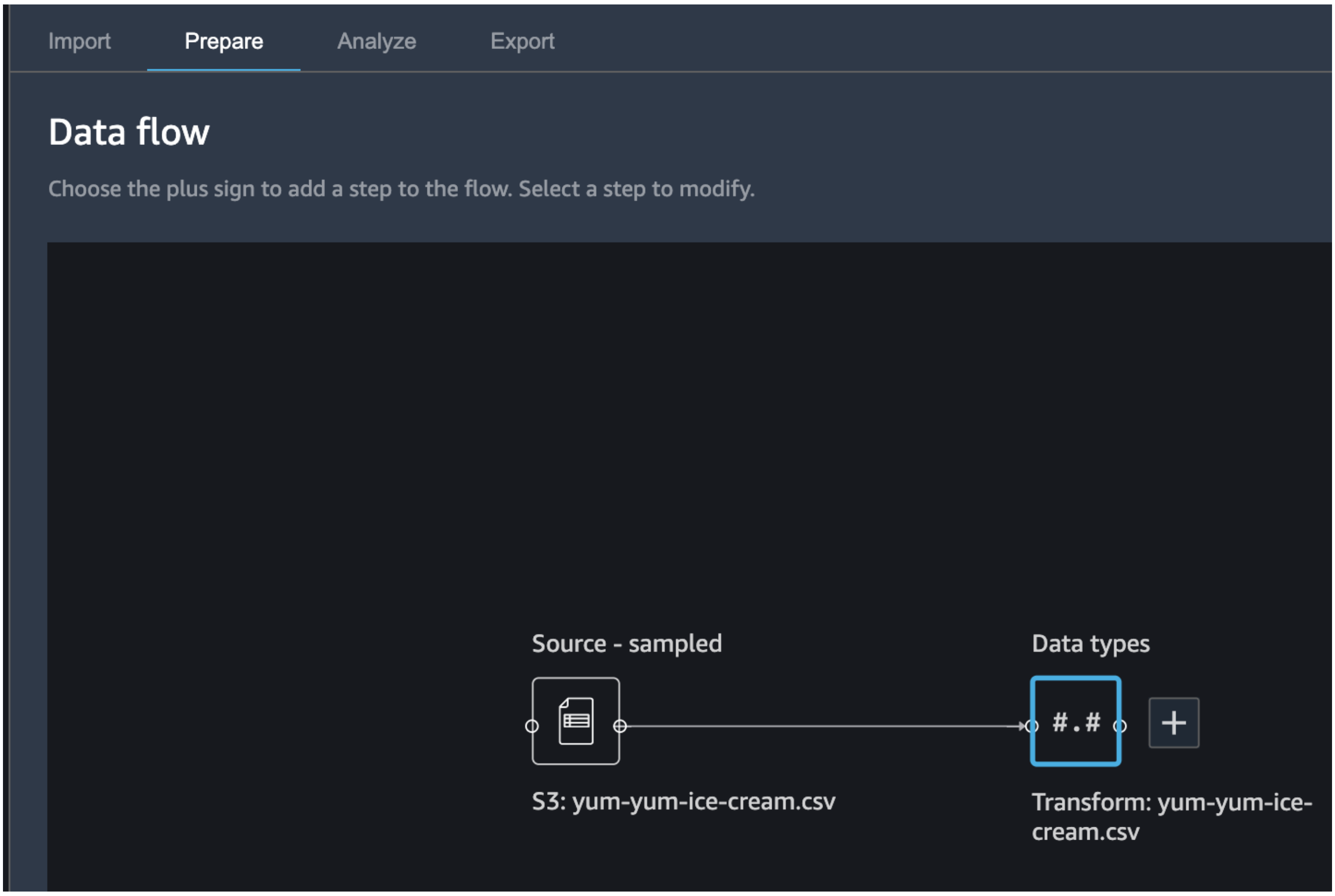
Enter the S3 URL of a file or prefix (folder) in the text box, or use the following table to browse S3

S3 / sagemaker-studio-466356262471-xwyfq86g7hr

Object Name	Size
yum-yum-ice-cream.csv	275B

Import data file

After the file has been imported, the Data flow will be presented to you. Start transforming the data by clicking the `+` next to the `Data types` section. You'll be adding a transformation.

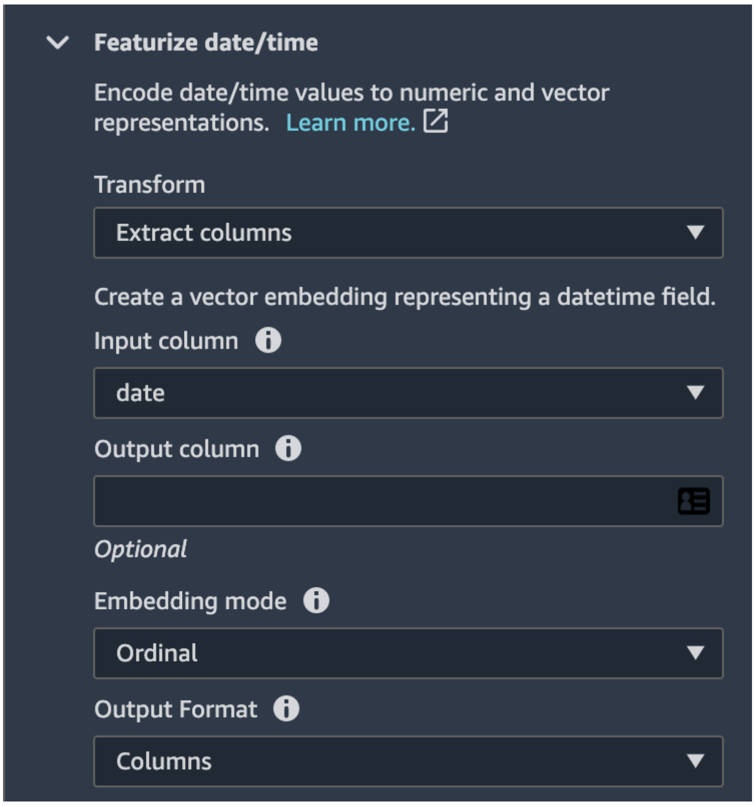


Data flow

Choose the plus sign to add a step to the flow. Select a step to modify.

Data Flow

Transform the `date` field to have `year`, `month`, and `day` as their own features.



Featurize date/time

Encode date/time values to numeric and vector representations. Learn more.

Transform: Extract columns

Create a vector embedding representing a datetime field.

Input column: date

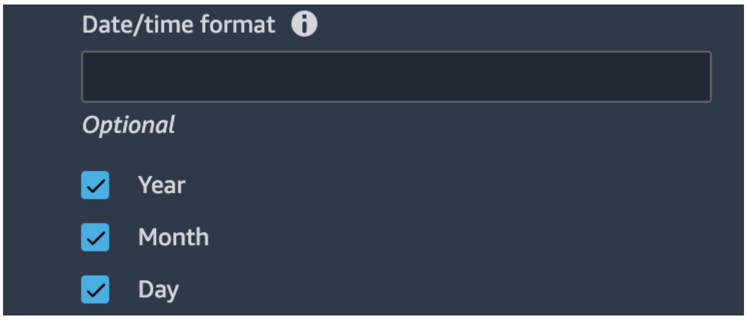
Output column:

Optional

Embedding mode: Ordinal

Output Format: Columns

Extract Data Features



Date/time format:

Optional

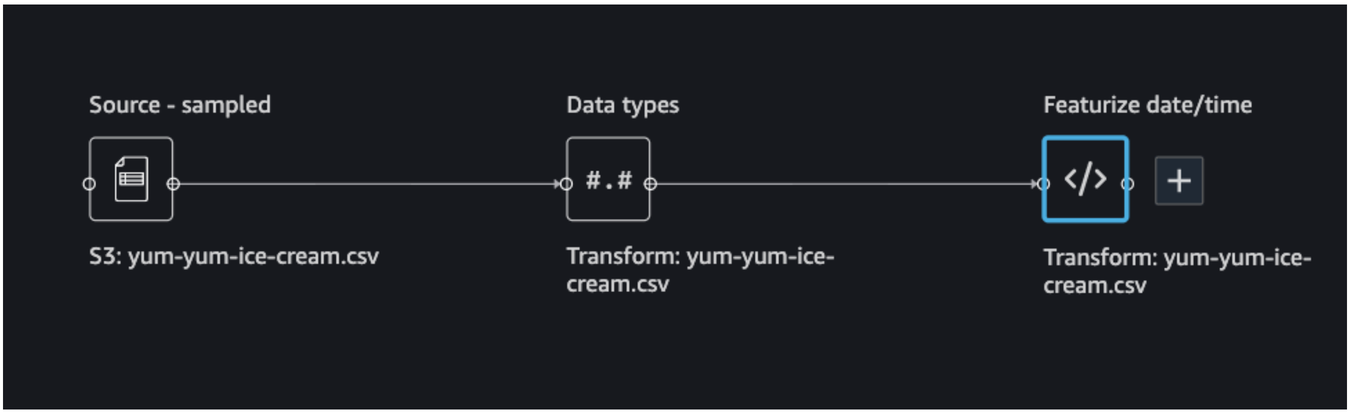
☑ Year

☑ Month

☑ Day

Extract Year, Month, Day

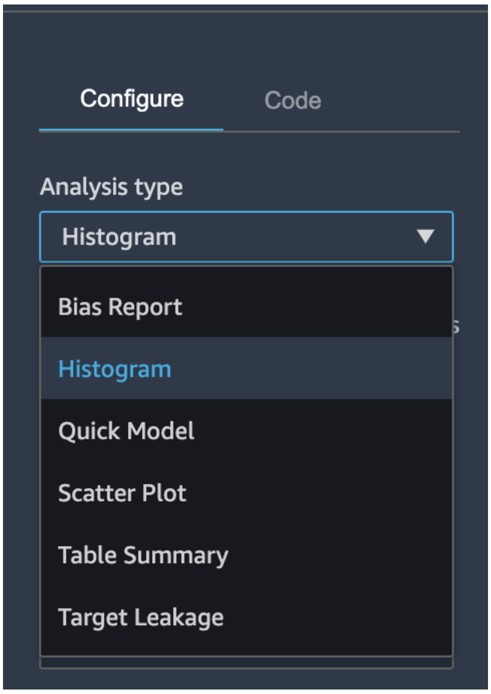
After transforming the date field, your data flow will be extended to visualize your changes.



Source - sampled
S3: yum-yum-ice-cream.csv
Data types
#.#
Transform: yum-yum-ice-cream.csv
Featurize date/time
</>
+
Transform: yum-yum-ice-cream.csv

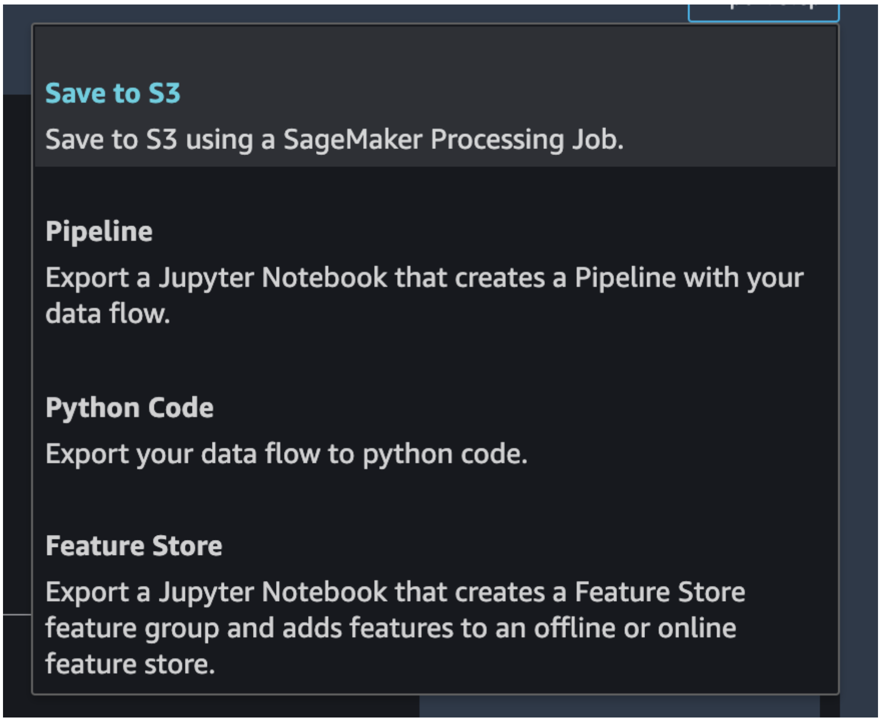
Data Flow With Features

Create the two visualizations, `Table Summary` and `Histogram` .



Histogram and Table Summary

Now complete the exercise by exporting your data flow to S3. Save the results to S3.


Save to S3
Save to S3 using a SageMaker Processing Job.

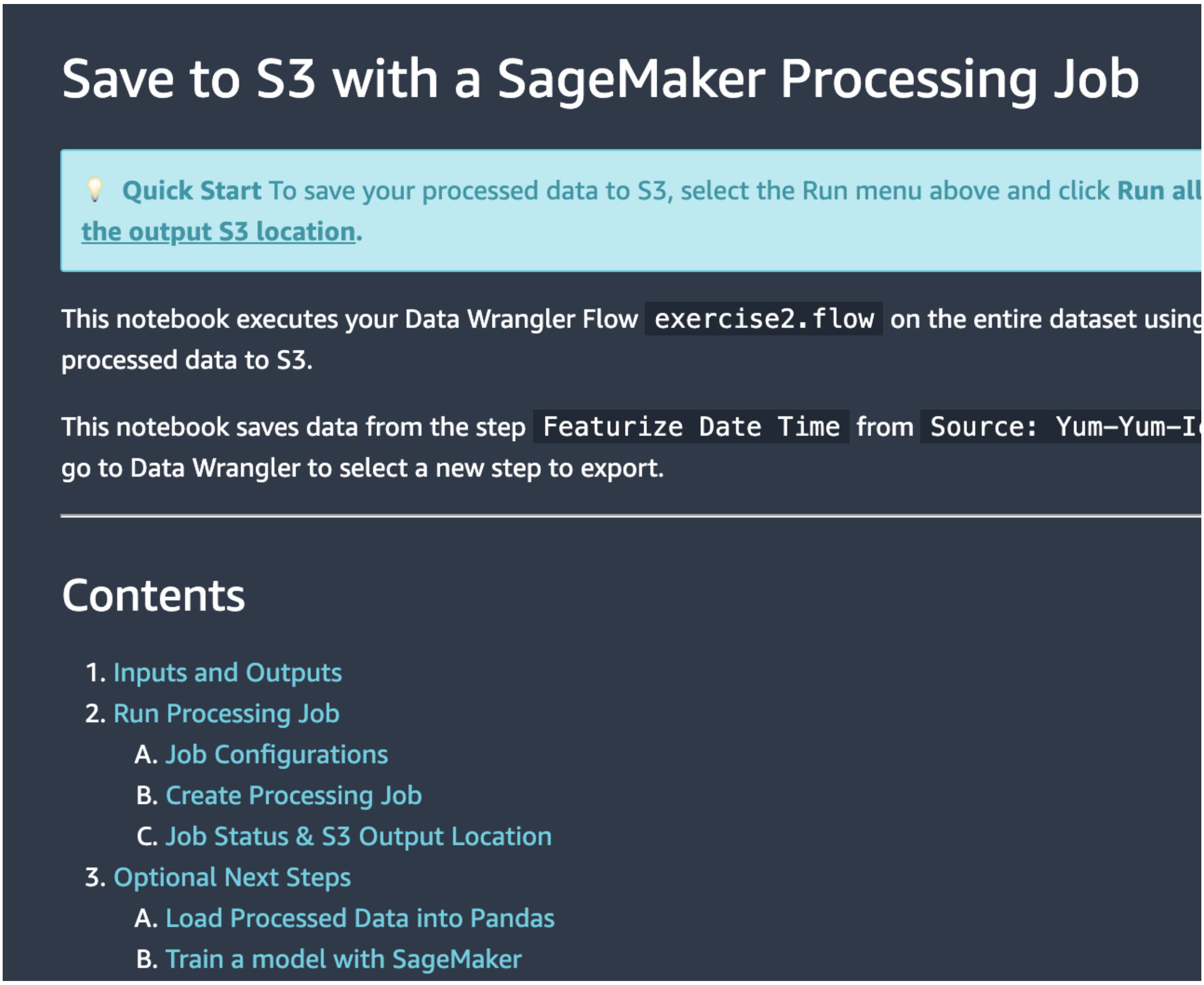
Pipeline
Export a Jupyter Notebook that creates a Pipeline with your data flow.

Python Code
Export your data flow to python code.

Feature Store
Export a Jupyter Notebook that creates a Feature Store feature group and adds features to an offline or online feature store.

Export to S3

When exporting your data flow to S3, it will create a Jupyter notebook called `Save to S3 with a SageMaker Processing Job` . The notebook will have everything configured for you to execute. To export your data, execute the cells up to the `Optional Next Steps`. The optional steps are not required for this exercise.



Save to S3 with a SageMaker Processing Job

💡 **Quick Start** To save your processed data to S3, select the Run menu above and click **Run all** the output S3 location.

This notebook executes your Data Wrangler Flow `exercise2.flow` on the entire dataset using processed data to S3.

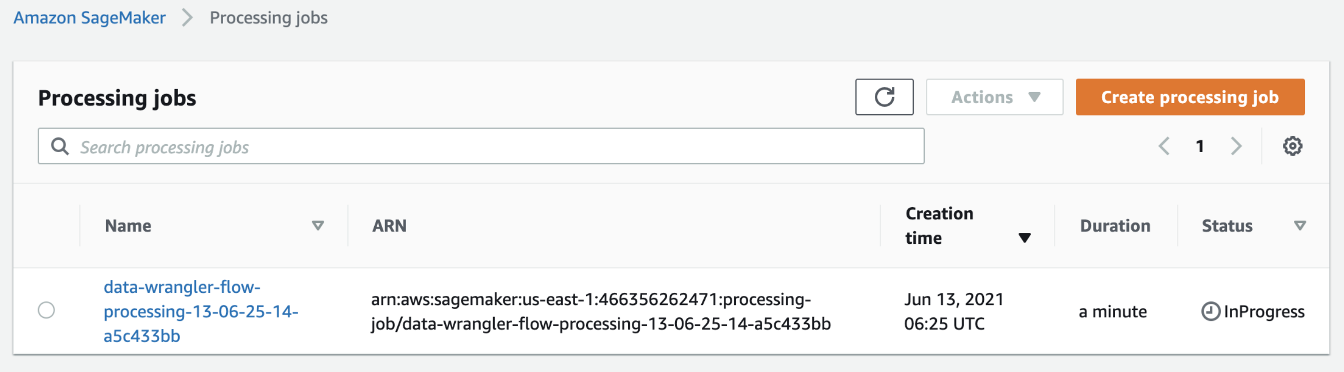
This notebook saves data from the step `Featurize Date Time` from `Source: Yum-Yum-I` go to Data Wrangler to select a new step to export.

Contents

1. Inputs and Outputs
2. Run Processing Job
   A. Job Configurations
   B. Create Processing Job
   C. Job Status & S3 Output Location
3. Optional Next Steps
   A. Load Processed Data into Pandas
   B. Train a model with SageMaker

Processing Job

You can view the status of your data flow export by navigating to `Processing Jobs` in the Sagemaker console.

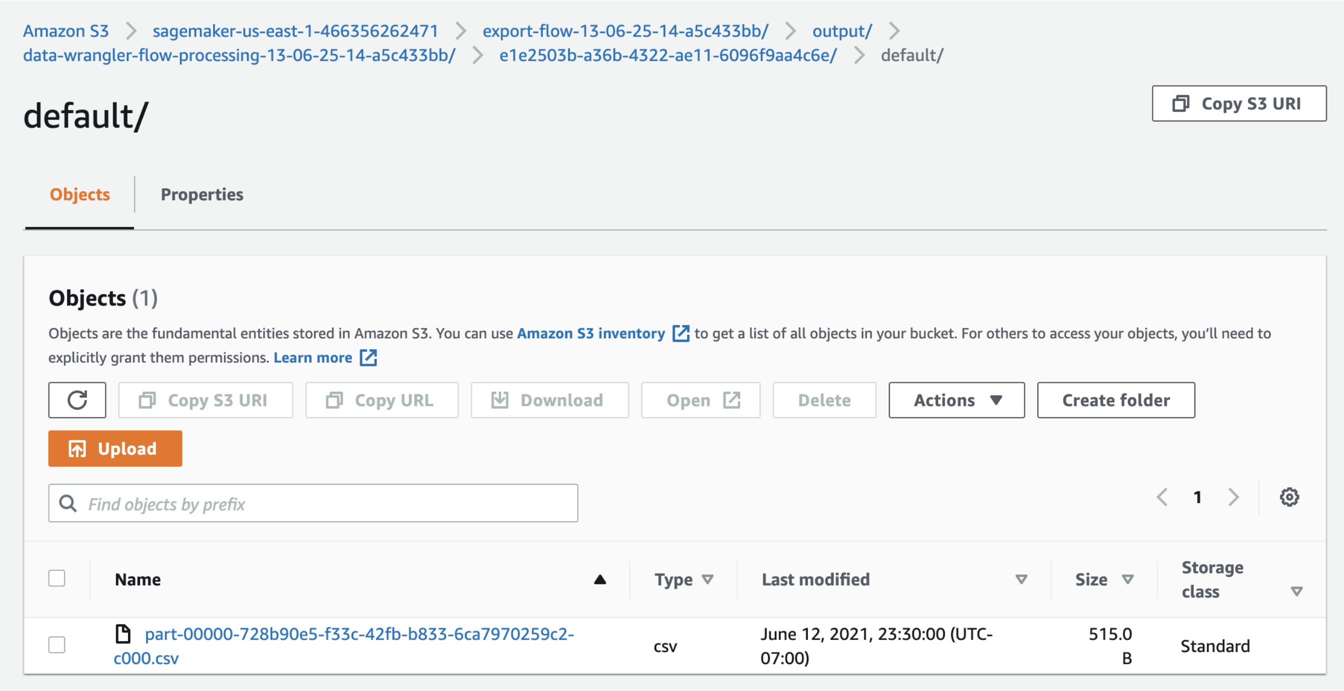


Processing jobs

	Name	ARN	Creation time ▼	Duration	Status
○	data-wrangler-flow-processing-13-06-25-14-a5c433bb	arn:aws:sagemaker:us-east-1:466356262471:processing-job/data-wrangler-flow-processing-13-06-25-14-a5c433bb	Jun 13, 2021 06:25 UTC	a minute	InProgress

Processing Job in Progress

After the job is complete, the data will be available in S3 under a named `export-flow` directory. Below is an example of such a path.



Amazon S3 > sagemaker-us-east-1-466356262471 > export-flow-13-06-25-14-a5c433bb/ > output/ > data-wrangler-flow-processing-13-06-25-14-a5c433bb/ > e1e2503b-a36b-4322-ae11-6096f9aa4c6e/ > default/

default/

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	part-00000-728b90e5-f33c-42fb-b833-6ca7970259c2-c000.csv	csv	June 12, 2021, 23:30:00 (UTC-07:00)	515.0 B	Standard

Processed Job Results

The file that will be exported will look something like below. The naming convention is automatically created during the data job, but the data inside will reflect the date transformation including `date_year` , `date_month` , and `date_day` .

The exercise is a simple one, but the underlying principles are what make Data Wrangler so powerful. Because everything is managed through AWS, we can easily create and manage workflows creating complex processes that power our machine learning applications.

part-00000-728b90e5-f33c-42fb-b833-6ca7970259c2-c000.csv

```
date,ice_cream_type,topping,location,date_year,date_month,date_day
2021-01-01,1,1,1,2021.0,0.0,0.0
2021-01-01,2,1,2,2021.0,0.0,0.0
2021-01-01,1,2,2,2021.0,0.0,0.0
2021-01-01,3,1,1,2021.0,0.0,0.0
2021-01-01,1,2,2,2021.0,0.0,0.0
2021-01-01,1,2,2,2021.0,0.0,0.0
2021-01-01,1,1,1,2021.0,0.0,0.0
2021-01-02,1,1,1,2021.0,0.0,1.0
2021-01-02,3,3,1,2021.0,0.0,1.0
2021-01-02,3,2,2,2021.0,0.0,1.0
2021-01-02,2,3,2,2021.0,0.0,1.0
2021-01-02,2,3,2,2021.0,0.0,1.0
2021-01-02,3,1,1,2021.0,0.0,1.0
2021-01-02,1,2,2,2021.0,0.0,1.0
```