

Feature Engineering in Amazon SageMaker

- AWS Academy Virginia

Using SageMaker instance for feature engineering

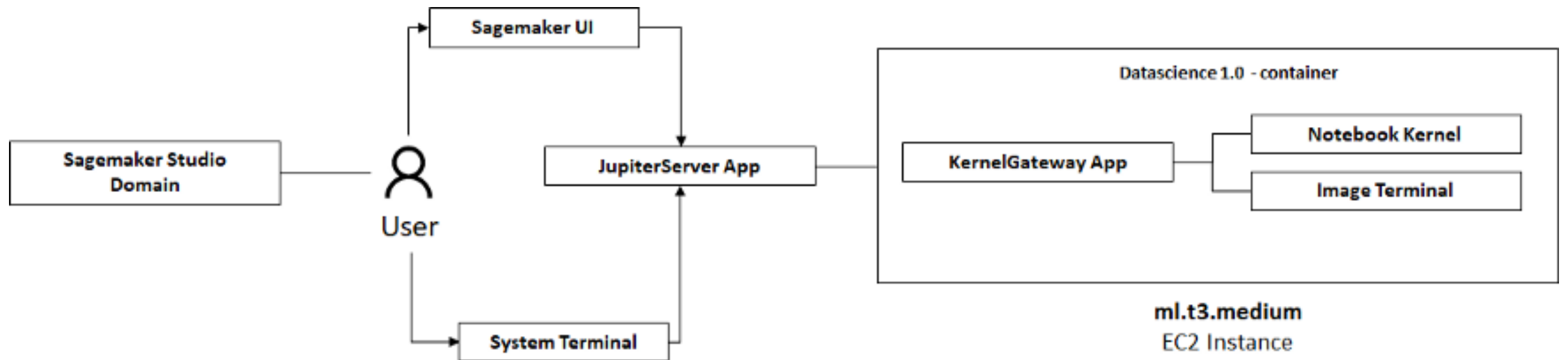
- Upload the `car_prediction_data.csv` file in S3
- Open `Day 05.ipnyb` and read the content to learn how
 - To analyze and visualize the data in Amazon SageMaker notebook instance
 - How to clean data in Amazon SageMaker notebook instance
 - Save cleaned data locally
 - Push the cleaned data into S3 Bucket



SageMaker Studio Data Wrangler

- 1) Use myapps for this part of lecture
- 2) Make sure you are in Virginia region

Studio Architecture



Launch SageMaker Studio

- Create SageMaker Domain
- Select Standard Setup

Amazon SageMaker > Setup SageMaker Domain

Setup SageMaker Domain

Use SageMaker Domain as the central store to manage the configuration of SageMaker for your organization.

Quick setup (1 min)

Let Amazon SageMaker configure your account, and set up permissions for your SageMaker Domain.

- ✓ Public internet access, and standard encryption
- ✓ SageMaker Studio Integration
- ✓ Sharable SageMaker Studio Notebooks
- ✓ SageMaker Canvas
- ✓ IAM Authentication

Perfect for single user domains and first time users looking to get started with SageMaker.

Standard setup (10 min)

Control all aspects of account configuration, including permissions, integrations, and encryption.

- ✓ Advanced network security, and data encryption
- ✓ SageMaker Studio, and RStudio integration
- ✓ SageMaker Studio Projects, and Jumpstart configurable
- ✓ SageMaker Canvas, and Amazon services integrations
- ✓ IAM, or IAM Identity Center (successor to AWS SSO)

Better for admins with large user groups, but you can always update your account configuration settings later if you want to do a quick setup now.

Configure

Configure Standard Setup

- Select the domain name and select IAM

Domain name

Name

Domain name should be unique across the AWS account.

Authentication

The authentication method you choose determines how you can access the SageMaker domain. To use AWS IAM Identity Center (successor to AWS SSO), you must have an IAM Identity Center account in an AWS Region supported by the domain.

☐ **AWS IAM Identity Center**
Access the domain with a bookmarked URL.

☒ **AWS Identity and Access Management (IAM)**
Access the domain with the Amazon SageMaker console.

Permission

See the following picture first and then read the next two slides that has instructions about item 1-4 in the following picture

Permission

Default execution role
SageMaker Domain requires permissions for its users to access other AWS services, such as Amazon SageMaker and Amazon S3. For a broad range of capabilities, you may attach the [AmazonSageMakerFullAccess](#) policy to the execution role. If you don't have a role with this policy, we can create one for you.

1 Enter a custom IAM role ARN ▼

2 Custom IAM role ARN
arn:aws:iam::272861155995:role/fast-ai-academic-16-Student-Azure

Space default execution role
SageMaker Domain requires permissions for its users to access other AWS services, such as Amazon SageMaker and Amazon S3. For a broad range of capabilities, you may attach the [AmazonSageMakerFullAccess](#) policy to the execution role. If you don't have a role with this policy, we can create one for you.

3 Enter a custom IAM role ARN ▼

4 Custom IAM role ARN
arn:aws:iam::272861155995:role/fast-ai-academic-16-Student-Azure

Permission

- You need to assemble the role as instructed below:

1) Select **custom IAM Role** when you get to the permissions step. You select **custom** in fields 1 and 3 shown in the picture.

2) Assemble the role like this:

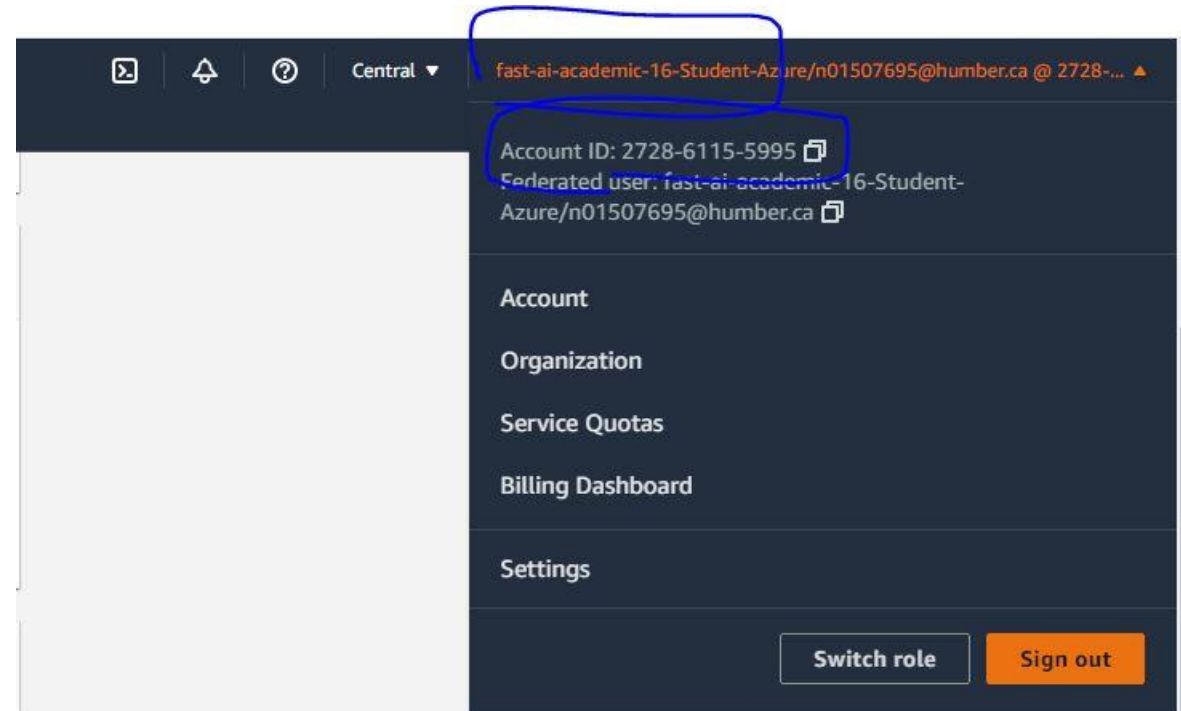
- `arn:aws:iam::AccountName:role/fast-ai-academic-nn-Student-Azure`

3) For example, if your AWS account number is 272861155995 and I assigned you the myapps account #16 this is the resulting text:

`arn:aws:iam::272861155995:role/fast-ai-academic-16-Student-Azure`

Permission

- You can get the AWS account number by clicking on the top right side, as shown below. Make sure you remove (-) between digits when you assemble the role text I mentioned above.



Network Setting

VPC

To enable internet access, make sure that your VPC has a NAT gateway and your security group allows outbound connections.

vpc-06e729a2a08af20c4 (172.31.0.0/16) | aws-controltower-VPC ▼

Subnet

Choose a subnet in an availability zone supported by Amazon SageMaker.

Choose one or more subnets ▼

subnet-063e600d62c572f34 (172.31.64.0/20) | ca-central-1a aws-controltower-PrivateSubnet1A ✕

subnet-05e6b3d2248c207de (172.31.16.0/20) | ca-central-1b aws-controltower-PrivateSubnet2B ✕

Security group(s)

These security groups will also be associated with the RStudioServerPro App.

Choose one or more security groups ▼

sg-0aff74529d6b71f04 (default) ✕

- ☒ Public Internet Only - The SageMaker domain will use default SageMaker internet access. Your vpc is used only for accessing the attached EFS storage
- ☐ VPC Only - The SageMaker domain will use your VPC. Direct internet access is disabled.
To enable internet access, make sure that your VPC has a NAT gateway and your security group allows outbound connections.


We do not need to share the notebooks

- This for scenarios that more than one person wants to work on the same notebook

▼ Notebook Sharing Configuration

Recommended defaults have been selected for you

Shareable notebook resources


Notebook resources include artifacts such as cell output and Git Repositories [Learn more](#) 

☐ Enable notebook resource sharing

All Jumpstart features must be enabled

SageMaker Projects and JumpStart - *optional*

SageMaker Projects and JumpStart **New**

Enable access and provisioning of AWS Service Catalog Portfolio of products in Amazon SageMaker Studio for Amazon SageMaker Projects and JumpStart. [Learn more](#) 

- ☒ **Enable Amazon SageMaker project templates and Amazon SageMaker JumpStart for this account**
If enabled, the administrator can view the Amazon SageMaker built-in project templates and Amazon SageMaker JumpStart solutions published in AWS Service Catalog. A launch constraint role and a project use role are automatically generated in IAM for your account.
- ☒ **Enable Amazon SageMaker project templates and Amazon SageMaker JumpStart for Studio users**
If enabled, this setting allows users who are currently using the domain execution role to create projects using templates and JumpStart solutions published by Amazon SageMaker in AWS Service Catalog. If there are individual users using custom execution roles in your organization, you need to enable them on the user profile page.

After clicking on **Next**, and **Submit**

Canvas Settings

- Ignore R Studio error and continue to Canvas configuration

Disable time series forecast

Amazon SageMaker Canvas settings [Info](#)

Configure Canvas for your organization.

▼ Canvas base permissions configuration

☒ Enable Canvas base permissions

If you enable Canvas base permissions, your users will have the necessary permissions to build models in Canvas. If you disable Canvas base permissions, your users won't have the necessary permissions to use Canvas, and you must manually configure IAM permissions for full Canvas functionality.

The [AmazonSageMakerCanvasFullAccessPolicy](#) will be attached to the default SageMaker execution role that you have specified in General settings.

▼ Time series forecasting configuration

☒ Enable time series forecasting

Enable time series forecasting to allow users to use time series forecasting in Canvas.

Amazon Forecast role

Canvas needs permission to connect to Amazon Forecast on your behalf to enable time series forecasting in Canvas.

☒ Create and use a new execution role

☐ Use an existing execution role

New IAM role suffix

Your role will be prefixed with "AmazonSagemakerCanvasForecastRole-" and includes the policy named

[AmazonSagemakerCanvasForecastRolePolicy](#)

20221112T185826

The name can have up to 63 characters. Valid characters: A-Z, a-z, 0-9, and - (hyphen)

Before clicking on Create

- After creating domain, make sure the **default** user is there or create it if it is not there

The SageMaker Domain is ready
Choose your user name, then choose Launch app to get started.

Jupyter Lab 3 will be the officially supported version in SageMaker Studio, while Jupyter Lab 1 will only receive security fixes from August 31, 2022.

Amazon SageMaker > Domains > Domain: default-1668308714784

Control Panel

Configure and manage SageMaker domain, users, and apps.

Users

Search users

1

Name ▾	Modified on ▾	Created on ▾	
default-1668308506700	Nov 13, 2022 03:10 UTC	Nov 13, 2022 03:10 UTC	<div>Launch app ▾</div>

Launch the Studio

Control Panel

Configure and manage SageMaker domain, users, and apps.

Users

Add user

Q Search users

< 1 > ⚙

Name ▾	Modified on ▾	Created on ▾
default-1668308506700	Nov 13, 2022 03:10 UTC	Nov 13, 2022 03:10 UTC

Launch app ▲

Studio

Canvas

Apps

Configure apps to manage your ML workflow using SageMaker.



Upload Titanic dataset to S3

[Amazon S3](#) > [Buckets](#) > [day-05-mk](#)



day-05-mk [Info](#)

[Objects](#) | [Properties](#) | [Permissions](#) | [Metrics](#) | [Management](#) | [A](#)

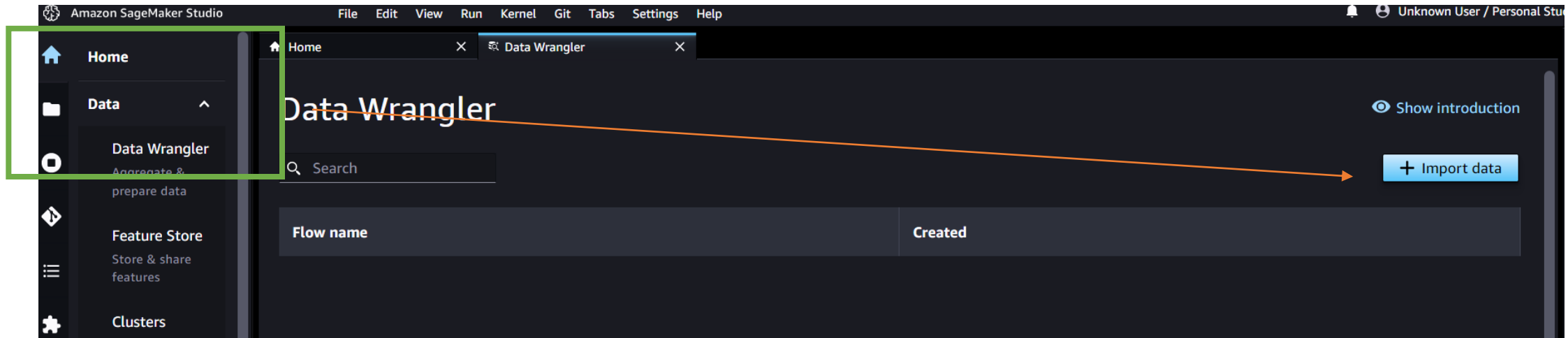
Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#)

<input type="checkbox"/>	Name	Type
<input type="checkbox"/>	 car_prediction_data.csv	csv
<input type="checkbox"/>	 titanic.csv	csv

Start DW flow and import **titanic** data



NOTE: when you click on Import, the studio creates a new server for DW. If you do not terminate the server after you are done with DW, your credit will be used up (see next slide). Make sure you terminate the DW server. I show you how to do that.

A new big server for DW

- The server shown in pending mode will be available to DW

Amazon SageMaker > Domains > Domain: day05 > User Details: default-1692368204257

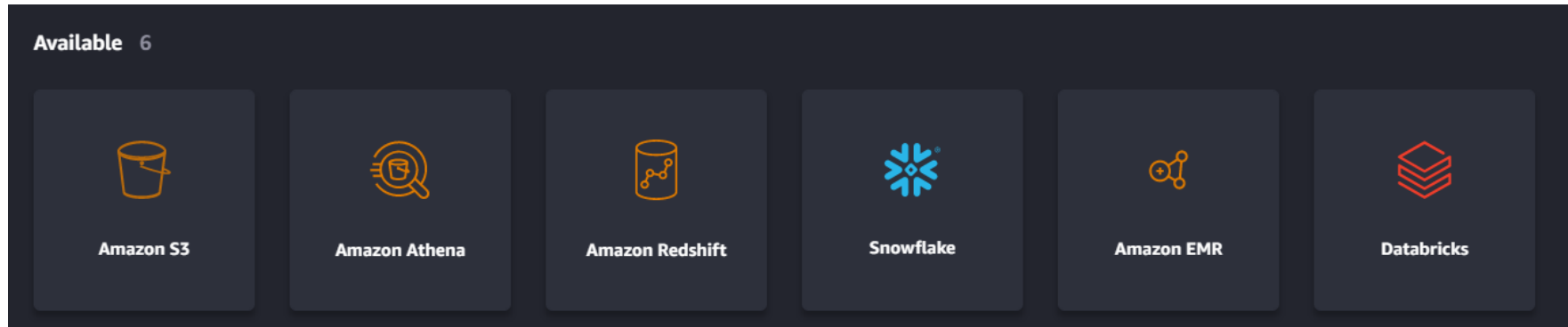
User Details

General details about this user profile.

Launch ▼

Apps				
App name	Status	App type	Created	Action
sagemaker-data-wrang-ml-m5-4xlarge-b741c1a025d542c78bb538373f2d	⋮ Pending	KernelGateway	Sun Oct 01 2023 07:58:25 GMT-0700 (Pacific Daylight Time)	Delete app
default	✔ Ready	JupyterServer	Fri Aug 18 2023 07:17:32 GMT-0700 (Pacific Daylight Time)	Action ▼

Select the bucket and data set



1. PassengerId: Unique Id of a passenger
2. Survived: If the passenger survived(0-No, 1-Yes)
3. Pclass: Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
4. Name: Name of the passenger
5. Sex: Male/Female
6. Age: Passenger age in years
7. SibSp: No of siblings/spouses aboard
8. Parch: No of parents/children aboard
9. Ticket: Ticket Number
10. Fare: Passenger Fare
11. Cabin: Cabin number
12. Embarked: Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

< Import data

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

⚙️ Advanced configuration

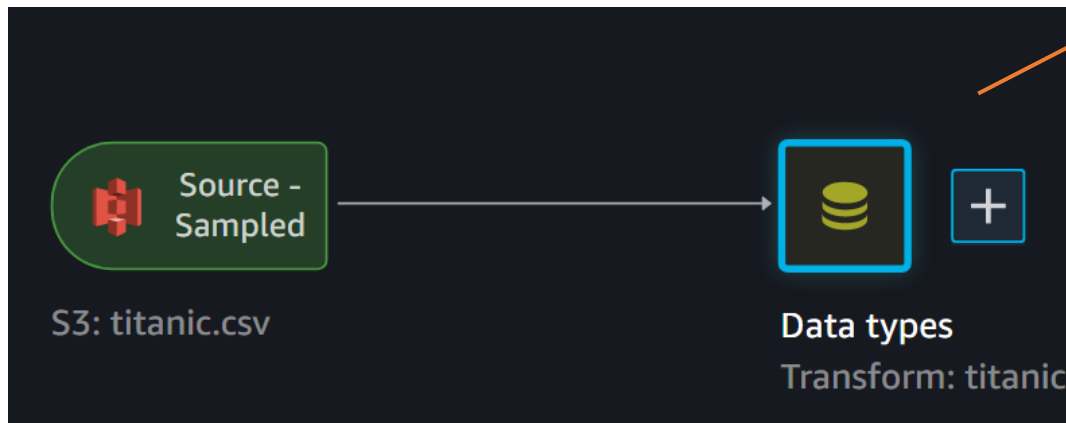
S3 URI path

S3 / day-05-mk

Object name	Size
car_prediction_data.csv	16.81KB
titanic.csv	114.96KB

Generating a quick report

- Click on the **+** sign and click on **Get data insight**
- Select **Data Quality** report and **Survived** in Target column
- Select **Classification** in prediction type



The screenshot shows the "Create analysis" panel in a data tool. At the top, there is a back arrow and the text "All analyses". Below this is the heading "Create analysis". The panel contains the following fields:

- Analysis type:** A dropdown menu with "Data Quality And Insights Report" selected.
- Target column:** A dropdown menu with "survived" selected, accompanied by a close button (X) and a dropdown arrow.
- Optional:** A section containing the "Problem type" field.
- Problem type:** Two radio buttons are present: "Regression" (unselected) and "Classification" (selected).

At the bottom of the panel, there is a "Clear" button on the left and a "Create" button on the right. A tooltip with the text "Click 'Create'" and a close button (X) is visible near the "Create" button.

Data quality and insights report

- A report to get information that might help you with data exploration and feature engineering
- It gives you information such as the % of missing values and the number of outliers.
- If you have issues with your data, such as **target leakage** or **imbalance**, the insights report can bring those issues to your attention.

Observe the following reports

- SUMMARY → MISSING VALUE

- DUPLICATE ROWS

- QUICK MODEL

- CONFUSION Matrix

- Prediction Power

- Feature details and their
Missing %

SUMMARY			
Dataset statistics			
Key		Feature type	
Value		Count	
Number of features		numeric	6
Number of rows		categorical	2
Missing		text	4
Valid		datetime	0
Duplicate rows		binary	1
		unknown	0

Data Exploration

- Choose the + next to the **Data type** step in your data flow and select **Add analysis**
- In the **Analysis** area, select **Table summary** from the dropdown list.
- Give the table summary a **Name**.
- Select **Run** to preview the table that will be created.

Observations

- Fare average (mean) is around **\$33**, while the max is over **\$500**. This column likely has outliers.
- This dataset uses **? Or empty spaces** to indicate missing values. A few number of columns have missing values: *cabin*, *embarked*, and *home.dest*

cabin	embarked
1309	1309
None	None
None	None
?	?
T	S

- The age category is missing over 250 values (different between 1309 and 1046).
- Go back to the data flow. Next, clean your data using the insights gained from these stats.

Drop Unused Columns

- Choose + next to the **Data type** step in your data flow and choose **Add transform**.
- Choose **Manage columns**
- Under **Transform**, make sure **Drop column** is selected
- Under **Columns to drop**, specify the following column names: cabin, ticket, name, sibsp, parch, home.dest, boat, body
- Choose **Preview**
- Choose **Add**

Using Pandas to drop the columns

- Alternatively, you could drop the columns by using the following code as well

```
cols = ['name', 'ticket', 'cabin', 'sibsp', 'parch', 'home.dest', 'boat', 'body']  
df = df.drop(cols, axis=1)
```

Clean up Missing Values

- Before we start fixing the missing values let's see how many missing value we have. Run `df.info()` as shown.
- You **do not** need to add it to the flow, that is just for you to get a better idea

As you see **Age** and **Fare** have different counts

```
6  0  pclass    1309 non-null  int64
7  1  survived  1309 non-null  int64
8  2  sex       1309 non-null  object
9  3  age       1046 non-null  float64
10 4  fare      1308 non-null  float64
11 5  embarked  1309 non-null  object
dtypes: float64(2), int64(2), object(2)
memory usage: 61.5+ KB
```

CUSTOM TRANSFORM

Name

Optional

Python (Pandas)

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 'df'

2 df.info()

Clear

Preview

Add

Output

```
1 <class 'pandas.core.frame.DataFrame'>
2 RangeIndex: 1309 entries, 0 to 1308
3 Data columns (total 6 columns):
4 #   Column      Non-Null Count  Dtype
5 ---  ---
6 0  pclass      1309 non-null  int64
7 1  survived    1309 non-null  int64
8 2  sex         1309 non-null  object
9 3  age         1046 non-null  float64
10 4  fare        1308 non-null  float64
11 5  embarked    1309 non-null  object
```

Handling missing value

- Click on **+** after Drop Column → Data Transform
- Choose **Handling missing**.
- Choose **Drop missing** for the **Transformer**.
- Choose *age* for the **Input column**.
- Choose **Preview** to see the new data frame, and then choose **Add** to add the transform to your flow.
- Repeat the same process for *fare and embark*.

Post report

- Use `df.info()` again to see the result
- All the counts are the same now

```
1 <class 'pandas.core.frame.DataFrame'>
2 RangeIndex: 1045 entries, 0 to 1044
3 Data columns (total 6 columns):
4 #   Column      Non-Null Count  Dtype
5 ---  ---
6 0    pclass      1045 non-null   int64
7 1    survived    1045 non-null   int64
8 2    sex         1045 non-null   object
9 3    age         1045 non-null   int64
10 4    fare        1045 non-null   float64
11 5    embarked    1045 non-null   object
12 dtypes: float64(1), int64(3), object(2)
13 memory usage: 49.1+ KB
14
```

CUSTOM TRANSFORM

Use Pyspark, Pandas, or Pyspark (SQL) to define custom transformations. [Learn more.](#)

Name

Optional

Python (Pandas)

!

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 `df`

2 `df.info()`

Clear

Preview

Add

Output

```
1 <class 'pandas.core.frame.DataFrame'>
2 RangeIndex: 1045 entries, 0 to 1044
3 Data columns (total 6 columns):
4 #   Column      Non-Null Count  Dtype
5 ---  ---
6 0    pclass      1045 non-null   int64
7 1    survived    1045 non-null   int64
8 2    sex         1045 non-null   object
9 3    age         1045 non-null   int64
10 4    fare        1045 non-null   float64
11 5    embarked    1045 non-null   object
```

One Hot Encoding

- **Custom Transform** → **Pandas** → **Name: Encode**
- In the **Custom Transform** section, choose **Python (Pandas)** from the dropdown list and add the code
- Choose **Preview** and **Add**

```
import pandas as pd

dummies = []
cols = ['pclass','sex','embarked']
for col in cols:
    dummies.append(pd.get_dummies(df[col]))

encoded = pd.concat(dummies, axis=1)

df = pd.concat((df, encoded),axis=1)
```

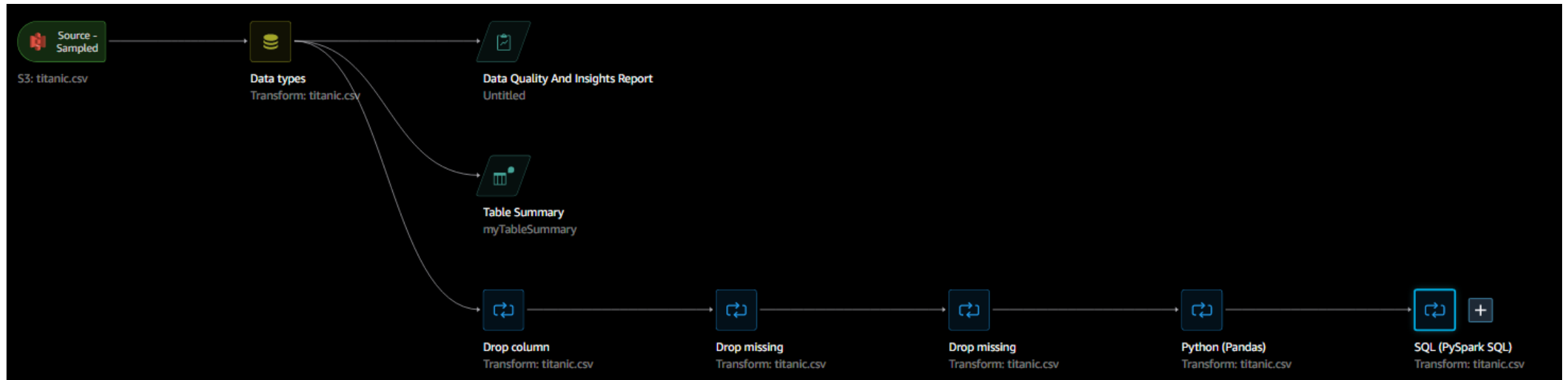
Custom SQL: SELECT Columns

- You can select the columns you want to keep using SQL.
- In the **Custom Transform** section, select **SQL (PySpark SQL)** from the dropdown list
- Enter the following in the code box.

```
SELECT survived, age, fare, 1, 2, 3, female, male, C, Q, S FROM df;
```

- Choose **Preview** and then **Add**
- **The columns listed in your SELECT statement** are the only remaining **columns**

This is what you should have now



Add a destination

- You can let the data flow to be Executed and the result to be saved into S3 bucket

Add a destination

Amazon S3

Dataset name

titanitoutput

File type

CSV (*.csv)

Delimiter

Comma (,)

Compression

None

Amazon S3 location ⓘ

s3://day05-mk-studio/

Browse

Your data is exported to the following S3 location: s3://day05-mk-studio/{processing-job-name}

PARTITIONING

Number of partitions ⓘ

Partition by column ⓘ

Select...

Optional

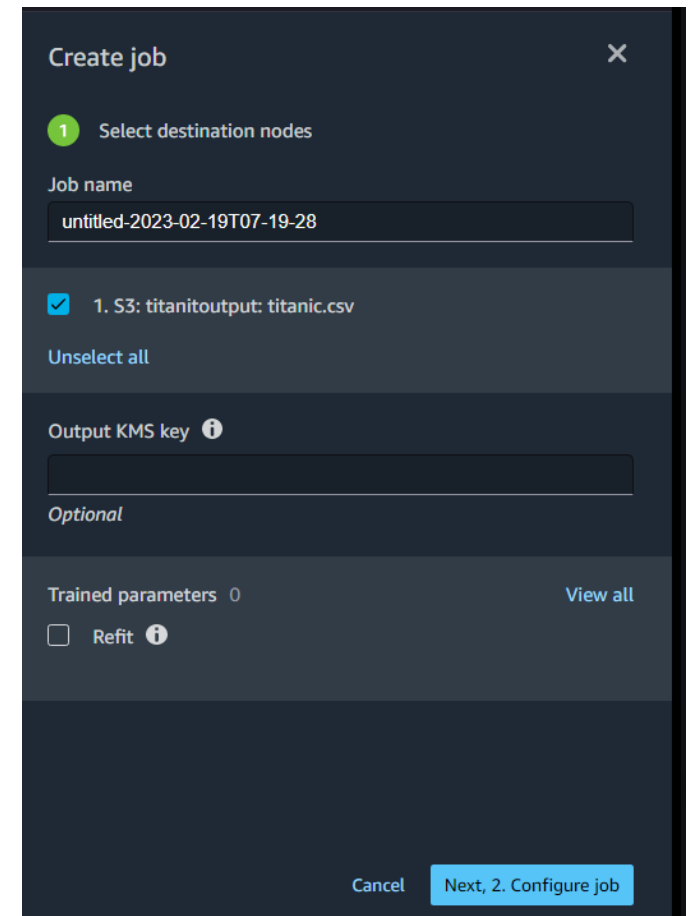
Optional

Cancel

Add destination

After you create a flow with S3 as destination

- You need to create a **job** to run that flow now or schedules



The screenshot shows a 'Create job' dialog box with a dark theme. At the top right is a close button (X). Below the title, there is a green circle with the number '1' followed by the text 'Select destination nodes'. A 'Job name' field contains the text 'untitled-2023-02-19T07-19-28'. Below this, a list of destinations is shown with a checked checkbox and the text '1. S3: titanitoutput: titanic.csv'. An 'Unselect all' link is below the list. The 'Output KMS key' field is empty, with an information icon (i) to its right. Below this field is the word 'Optional'. A 'Trained parameters' section shows '0' parameters and a 'View all' link. A 'Refit' checkbox is unchecked, with an information icon (i) to its right. At the bottom right are two buttons: 'Cancel' and 'Next, 2. Configure job'.

Create job

1 Select destination nodes

Job name

untitled-2023-02-19T07-19-28

☒ 1. S3: titanitoutput: titanic.csv

Unselect all

Output KMS key ⓘ

Optional

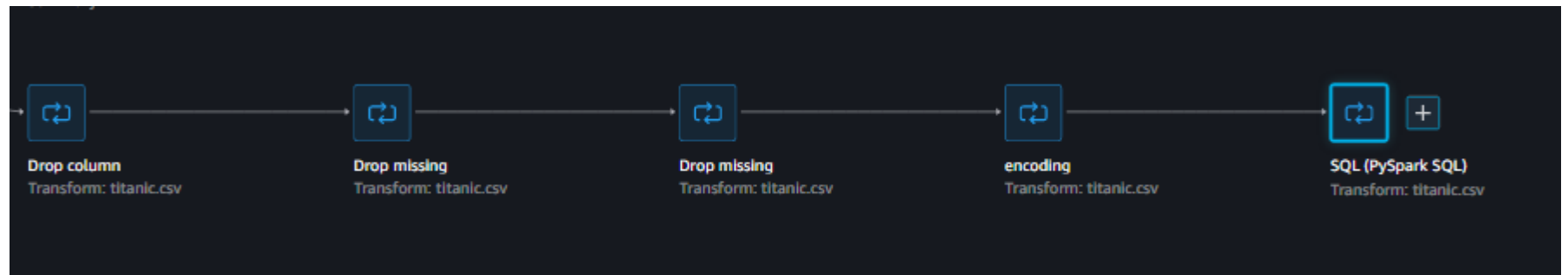
Trained parameters 0 View all

☐ Refit ⓘ

Cancel Next, 2. Configure job

Export the DW flow to a Notebook

- If you export your data flow using a **Data Wrangler**, the process automatically **creates** a Jupyter Notebook
- This notebook automatically opens in your Studio instance and is **configured to run a SageMaker processing job** to run your Data Wrangler data flow, which is referred to as a Data Wrangler job.
- Save your **data flow** → Select **File** and then select **Save Data Wrangler Flow**.
- Click on + in the last step and Select **Amazon S3 (via Jupyter notebook)**



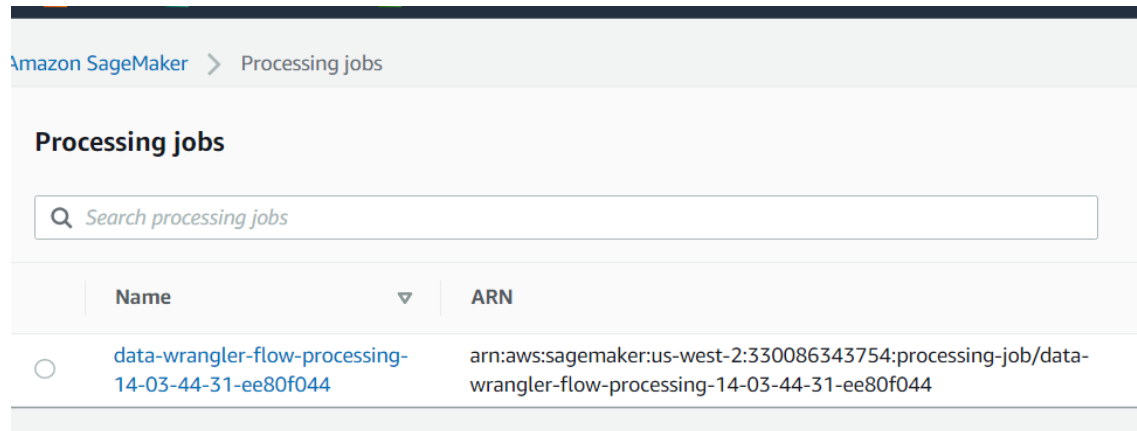
Using the notebook

- Choose any **Python 3 (Data Science)** kernel for the **Kernel**.
- **Before start running the cells read the following instructions:**
 - Do not run optional parts related to training **but** run the “**(Optional) Configure Spark Cluster Driver Memory**”
 - **Do not start training**
 - Change the output bucket to your own bucket where you downloaded the titanic data

```
# You can configure this with your own bucket name, e.g.  
bucket = "day-05-mk"
```

See the results

- Follow the execution flow in SageMaker console processing job and in the S3



The screenshot shows the Amazon SageMaker console interface. At the top, there is a breadcrumb navigation bar with 'Amazon SageMaker' and 'Processing jobs'. Below this, the title 'Processing jobs' is displayed. A search bar with the placeholder text 'Search processing jobs' is present. Below the search bar is a table with two columns: 'Name' and 'ARN'. The 'Name' column has a dropdown arrow. One job is listed in the table.

	Name	ARN
<input type="radio"/>	data-wrangler-flow-processing-14-03-44-31-ee80f044	arn:aws:sagemaker:us-west-2:330086343754:processing-job/data-wrangler-flow-processing-14-03-44-31-ee80f044

Clean up

- **Very important:**

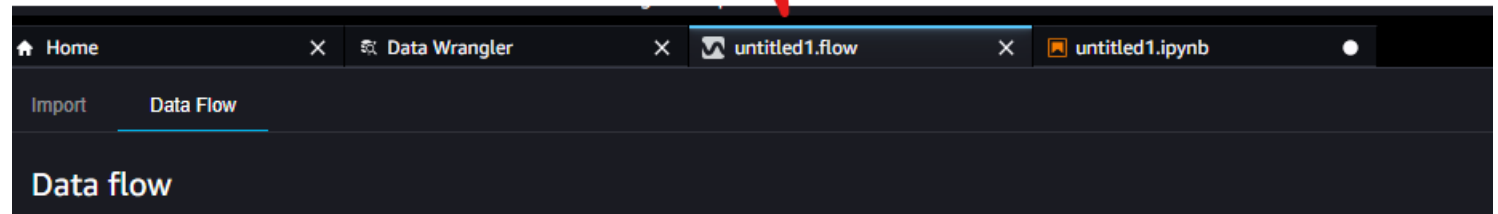
- Make sure you terminate the DW server and other servers otherwise it will use up your credit.
- Close all the tabs in the Studio that have the file flow files. Every time you starts the SM studio, if it has an open tab for .flow, it will create that huge DW server.

Amazon SageMaker > Domains > Domain: day05 > User Details: default-1692368204257

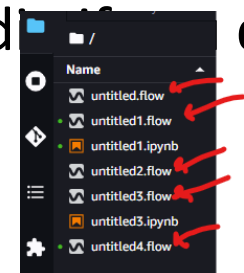
User Details

General details about this user profile.

Apps		
App name	Status	App type
sagemaker-data-scienc-ml-t3-medium-ccb588b5efaf671be41927273f0c	Deleted	KernelGateway
sagemaker-data-wrang-ml-m5-4xlarge-b741c1a025d542c78bb538373f2d	Deleted	KernelGateway
default	Ready	JupyterServer



- To avoid mistakes, it is better you delete all .flow files in the Studio. do not want to use credit for no reason.

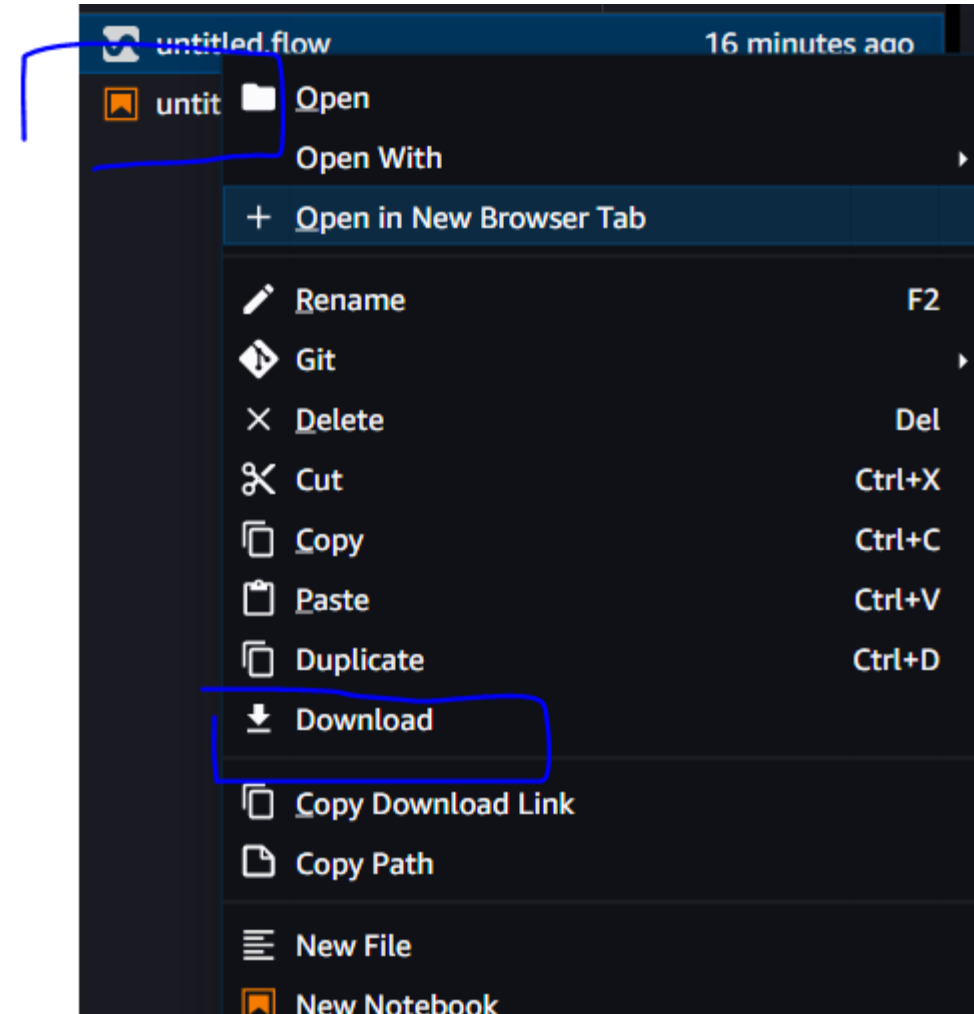


Assignment

- Go to <https://www.openml.org/> or <https://www.kaggle.com/>
- Select a data set. You may need to change the data format
- Analyze the data in DW and find some opportunities to improve the quality of data (like removing missing value or scaling data or one hot encoding, etc.)
- Clean/transform data in Studio DW
- Upload the following items in BB
 - Your selected data set
 - DW .flow file (to learn how to download the .flow file, see next slide)
 - A report that covers at least the suggested ToC (please see next slides)

Download the .flow file

- Right click on the file
- Select Download



ToC for the report

- Data set fields descriptions
- The visualization and analysis that you have done in DW and what you learn out of those visualizations. Each analysis comes with a picture and your description beneath
- The transformations you have done, again you have to explain why you have chosen that transformation, include the picture and explain the results after transformation (with pictures)
- If you have added a code, include that in the report in the right spot