

Assignment – 05

Make sure we are in central and everything Is deleted or shutdown and no instances running.

The screenshot shows the Amazon SageMaker console interface. The top navigation bar includes the AWS logo, a search bar, and the region set to 'Central'. The left sidebar contains navigation links for 'Getting started', 'Studio', 'Studio Lab', 'Canvas', 'RStudio', 'Admin configurations', and 'JumpStart'. The main content area displays the 'User Details' for a user named 'sreemanasa' within the 'Domain: mydomain5'. The 'Apps' table lists two applications: 'sagemaker-data-wrang-ml-m5-4xlarge-2ba8da4813c5fe7300bede5b3023' and 'default', both with a 'Delete' button. The 'Details' panel on the right shows the user's name, execution role, creation time, and status as 'InService'.

App name	Status
sagemaker-data-wrang-ml-m5-4xlarge-2ba8da4813c5fe7300bede5b3023	Delete
default	Delete

Details
Name sreemanasa
Execution role arn:aws:iam::085812980140:role/fast-ai-academic-43-Student-Azure
Created On Thu Mar 07 2024 00:34:27 GMT-0500 (Eastern Standard Time)
Status InService

The screenshot shows the Amazon SageMaker console interface, specifically the 'Domains' page. The top navigation bar includes the AWS logo, a search bar, and the region set to 'Central'. The left sidebar contains navigation links for 'Getting started', 'Studio', 'Studio Lab', 'Canvas', 'RStudio', 'Admin configurations', and 'JumpStart'. The main content area displays the 'Domains' page, which includes a description of domains and a table showing no domains are currently listed. The 'Create domain' button is visible in the top right corner of the table area.

Name	Id	Status	Created on	Modified on
No domains To add a domain, choose Create domain.				

Introduction:

I choose a dataset named salaries.csv which is uncleaned and uploaded it to S3 bucket in MyApps.

Then we create a domain in Amazon SageMaker then we land into a page seen as SageMaker Studio classic.

By opening the data wrangler, we create a data flow shown below where we get the insights of each step we have gone through while performing data analysis.

We import data from S3 bucket. Always when creating a new domain or creating the bucket we have to ensure that we are in central region.

Data Flow:

DI(Data Quality and Insights report)

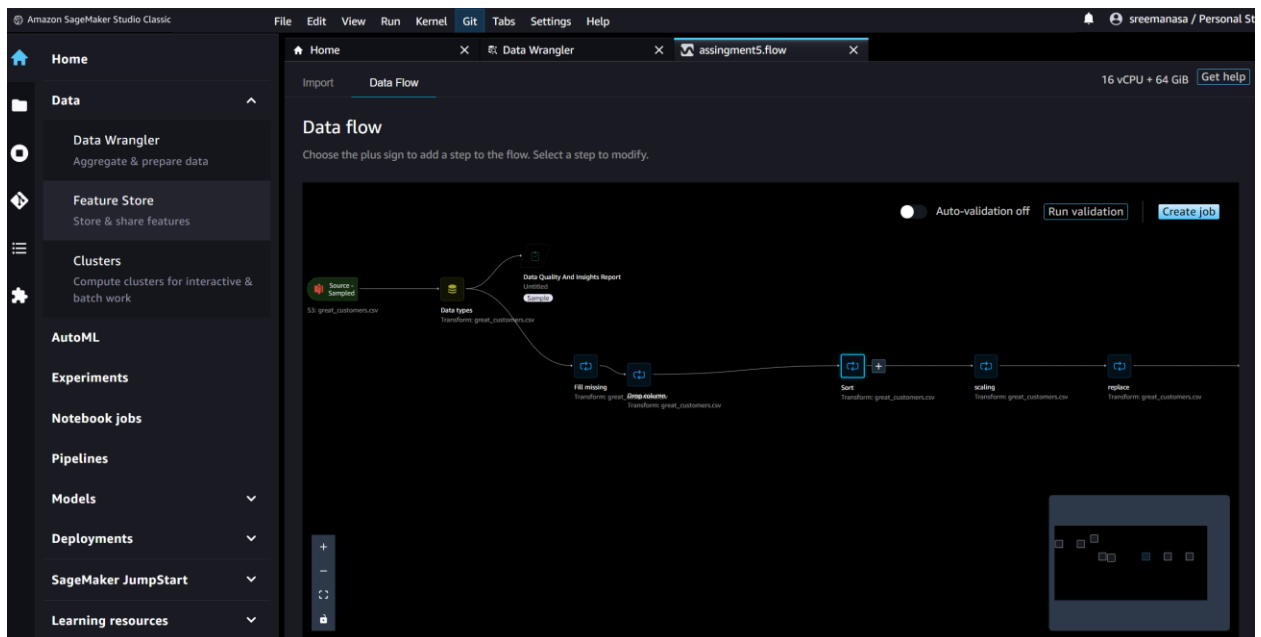
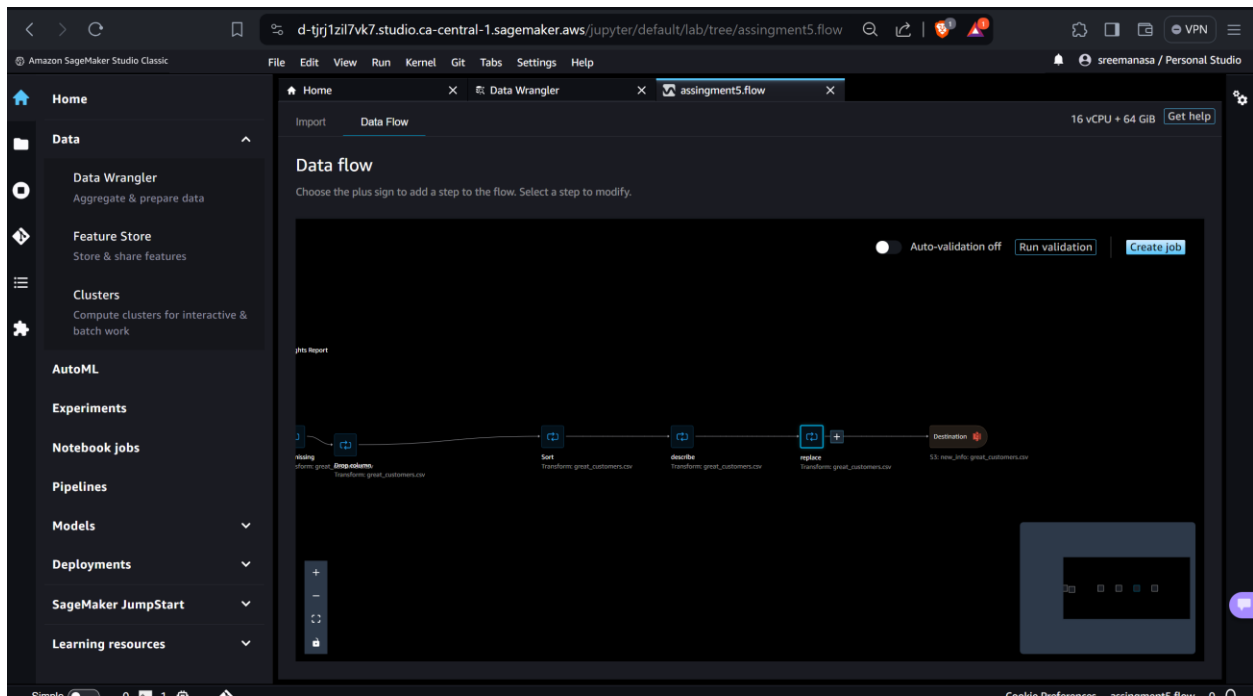
Data Analysis Performed:

1. Fill Missing values
2. Drop Columns
3. Sort the Dataset according to userId in ascending.
4. Describe the dataset using describe().
5. Replace(custom): Replaced the column Name from salary to Payment.

```
import pandas as pd
```

```
df.rename(columns={'salary': 'Payment'}, inplace=True)
```

After completing the data analysis, we save the flow in destination in s3 bucket.



Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

Home Data Data Wrangler Feature Store Clusters AutoML Experiments Notebook jobs Pipelines Models Deployments SageMaker JumpStart Learning resources

Home Data Wrangler assingment5.flow

16 vCPU + 64 GiB Get help

< Data flow

Sort · Transform: great_customers.csv

Data Analysis Training **NEW**

Step 1. S3 Source Visualizations off Export and train Export data

The following is a visualization of the first 2,000 rows of your dataset. To analyze the entire dataset, run a Data Quality and Insights Report.

user_id (string)	age (string)	workclass (string)
1.0001e+6 - 1.0488e+6	14 - 90	3 Categories
1004889	14.0	private
1012811	25.0	private
1006870	21.0	private
1022149	23.0	private
1029558	26.0	private
1022394	26.0	private
1026358	58.0	private
1026126	23.0	private

ALL STEPS

+ Add step

1. S3 Source

Name: great_customers.csv

Type: S3

Optional

S3 URI: s3://as5bucket/great_customers.csv

Optional

Content Type: csv

Optional

Has Header: ☒

2. Data types

2. Data Types

Amazon SageMaker Studio Classic

File Edit View Run Kernel Git Tabs Settings Help

Home Data Data Wrangler Feature Store Clusters AutoML Experiments Notebook jobs Pipelines Models Deployments SageMaker JumpStart Learning resources

Home Data Wrangler assingment5.flow

16 vCPU + 64 GiB Get help

< Data flow

replace · Transform: great_customers.csv

Data Analysis Training **NEW**

Step 2. Data types Visualizations off Export and train Export data

The following is a visualization of the first 2,000 rows of your dataset. To analyze the entire dataset, run a Data Quality and Insights Report.

user_id (long)	age (float)	workclass (string)
1.0001e+6 - 1.0488e+6	14 - 90	3 Categories
1004889	14	private
1012811	25	private
1006870	21	private
1022149	23	private
1029558	26	private
1022394	26	private
1026358	58	private
1026126	23	private

ALL STEPS

+ Add step

1. S3 Source

2. Data types

Column name	Type
user_id	Long
age	Float
workclass	String
salary	Float
education_rank	Long
marital-status	String
occupation	String
race	String
sex	String
mins_beerdrinking_year	Float
mins_exercising_year	Float

3. Drop Columns

The screenshot shows the Amazon SageMaker Studio Classic interface. The left sidebar contains navigation options: Home, Data, Feature Store, Clusters, AutoML, Experiments, Notebook jobs, Pipelines, Models, Deployments, SageMaker JumpStart, and Learning resources. The main panel displays a Data Wrangler workflow for 'great_customers.csv'. The current step is 'Step 4. Drop column'. A visualization of the first 2,000 rows of the dataset is shown, with columns 'user_id (long)', 'age (float)', and 'workclass (string)'. The 'workclass' column is selected for dropping. The right sidebar shows the 'ALL STEPS' list, including '1. S3 Source', '2. Data types', '3. Fill missing', '4. Drop column', '5. Sort', and '6. describe'.

4. Sort Data by userId

The screenshot shows the Amazon SageMaker Studio Classic interface. The left sidebar contains navigation options: Home, Data, Feature Store, Clusters, AutoML, Experiments, Notebook jobs, Pipelines, Models, Deployments, SageMaker JumpStart, and Learning resources. The main panel displays a Data Wrangler workflow for 'great_customers.csv'. The current step is 'Step 5. Sort'. A visualization of the first 2,000 rows of the dataset is shown, with columns 'user_id (long)', 'age (float)', and 'workclass (string)'. The 'user_id' column is selected for sorting. The right sidebar shows the 'ALL STEPS' list, including '1. S3 Source', '2. Data types', '3. Fill missing', '4. Drop column', '5. Sort', and '6. describe'. The 'Sort' step is configured to sort by 'user_id' in 'Ascending' order.

5. Describe

Amazon SageMaker Studio Classic interface showing the 'replace' transform step in the 'Data Wrangler' workflow.

Step 6. describe

The following is a visualization of the first 2,000 rows of your dataset. To analyze the entire dataset, run a [Data Quality and Insights Report](#).

user_id (long)	age (float)	workclass (string)
1000006	29	0
1000015	43	private
1000017	33	private
1000029	22	self_employed
1000030	46	government
1000031	56	self_employed
1000039	20	private
1000041	54	self_employed

ALL STEPS

- 4. Drop column
- 5. Sort
- 6. describe
 - Name: describe
 - 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.describe()

```

Replace

Amazon SageMaker Studio Classic interface showing the 'replace' transform step in the 'Data Wrangler' workflow.

Step 7. replace

The following is a visualization of the first 2,000 rows of your dataset. To analyze the entire dataset, run a [Data Quality and Insights Report](#).

user_id (long)	age (float)	workclass (string)
1000006	29	0
1000015	43	private
1000017	33	private
1000029	22	self_employed
1000030	46	government
1000031	56	self_employed
1000039	20	private
1000041	54	self_employed

ALL STEPS

- 6. describe
- 7. replace
 - Name: replace
 - 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.

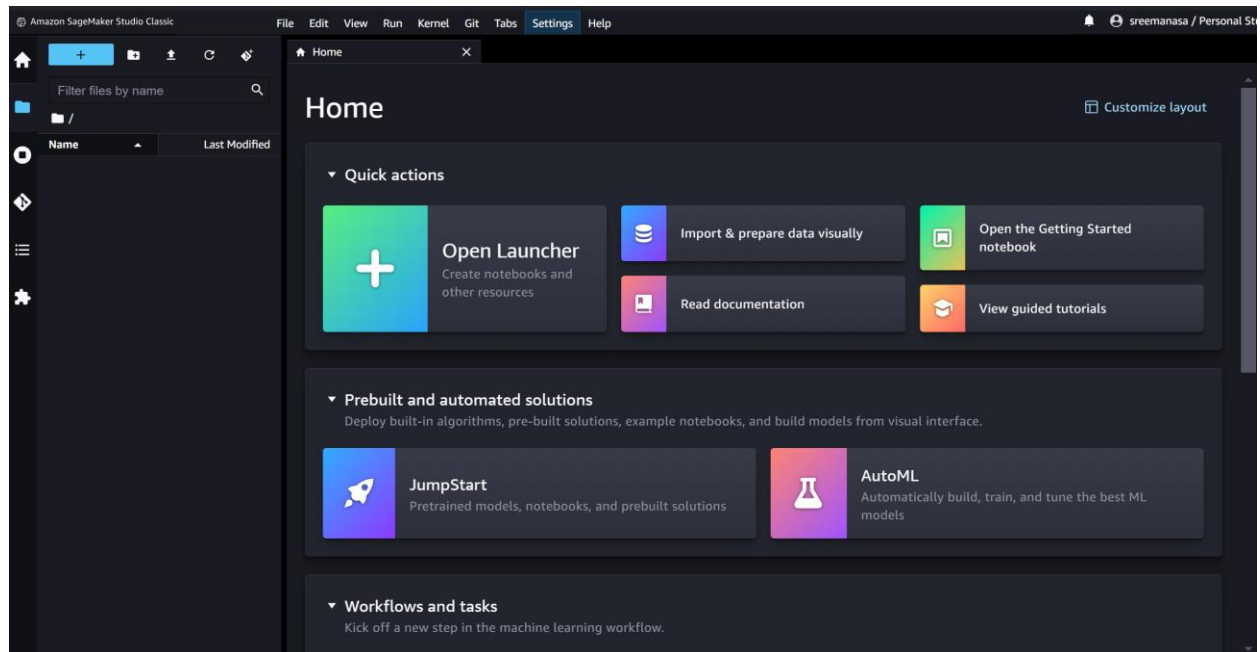
```

1 # Table is available as variable 'df'
2 import pandas as pd
3 df.rename(columns={'salary': 'Payment'},
4           inplace=True)

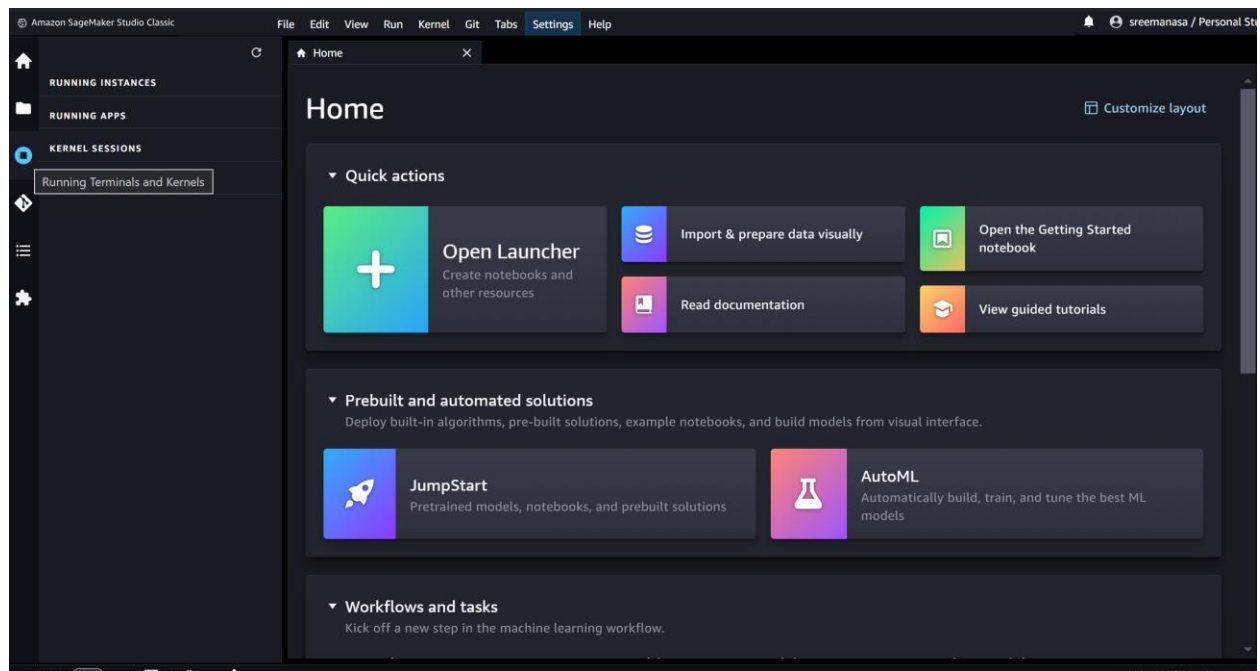
```

Clear Preview Update

Deleted the Wrangler flow: As we have created a flow, once we use it, we have to make sure to delete the files from the folder and shut down all the kernels.



No running Terminals:



As we have used the domain, We need to delete them in order to consume less credits.