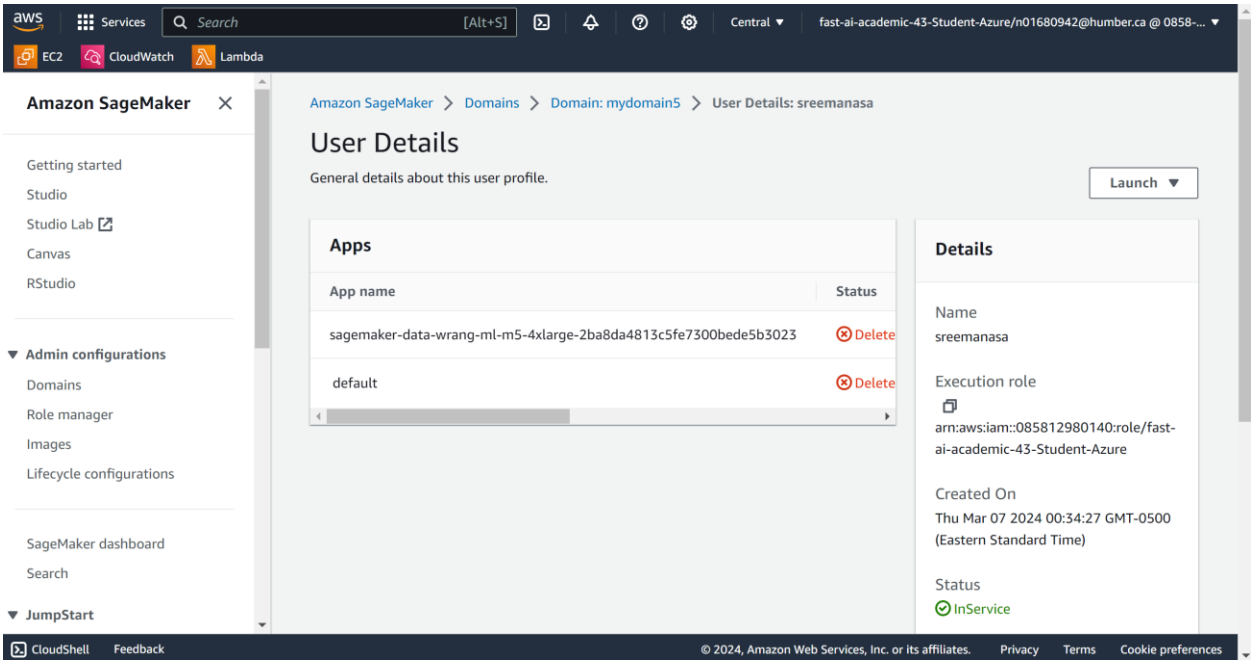
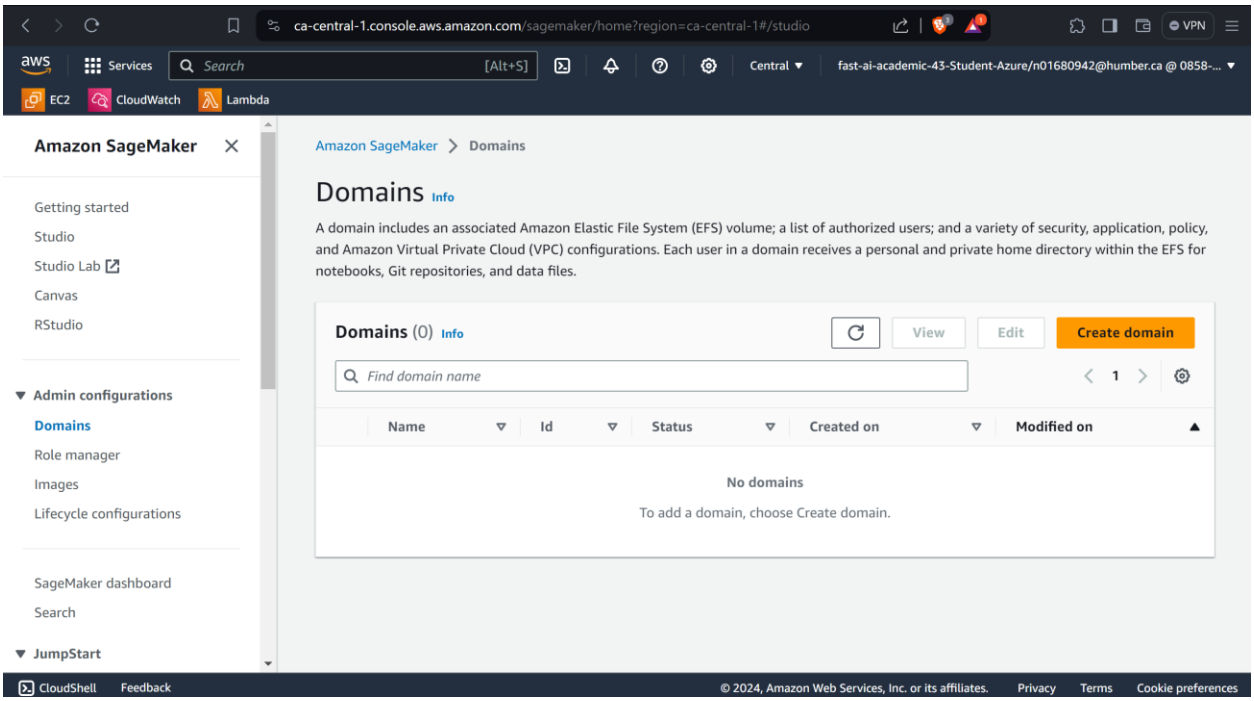
Assignment – 05

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





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 creaking the bucket we have to ensure that we are in central region.

Data Flow:

DI(Data Quality and Insights report)

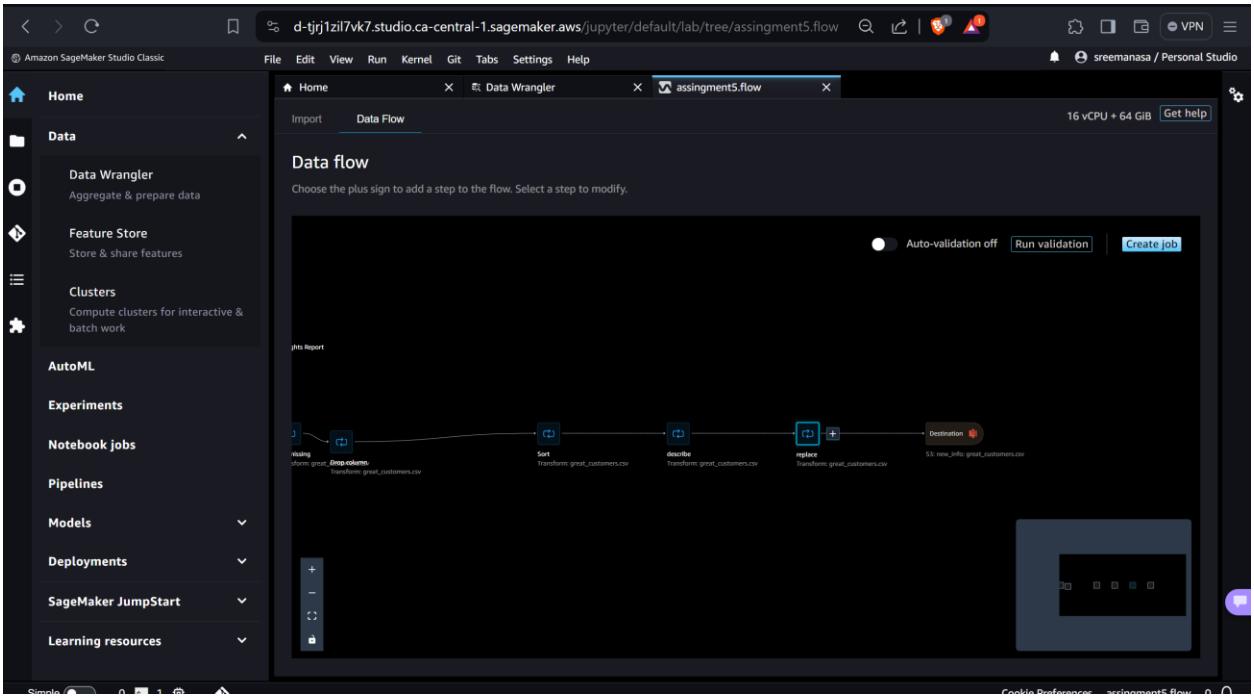
Data Analysis Performed:

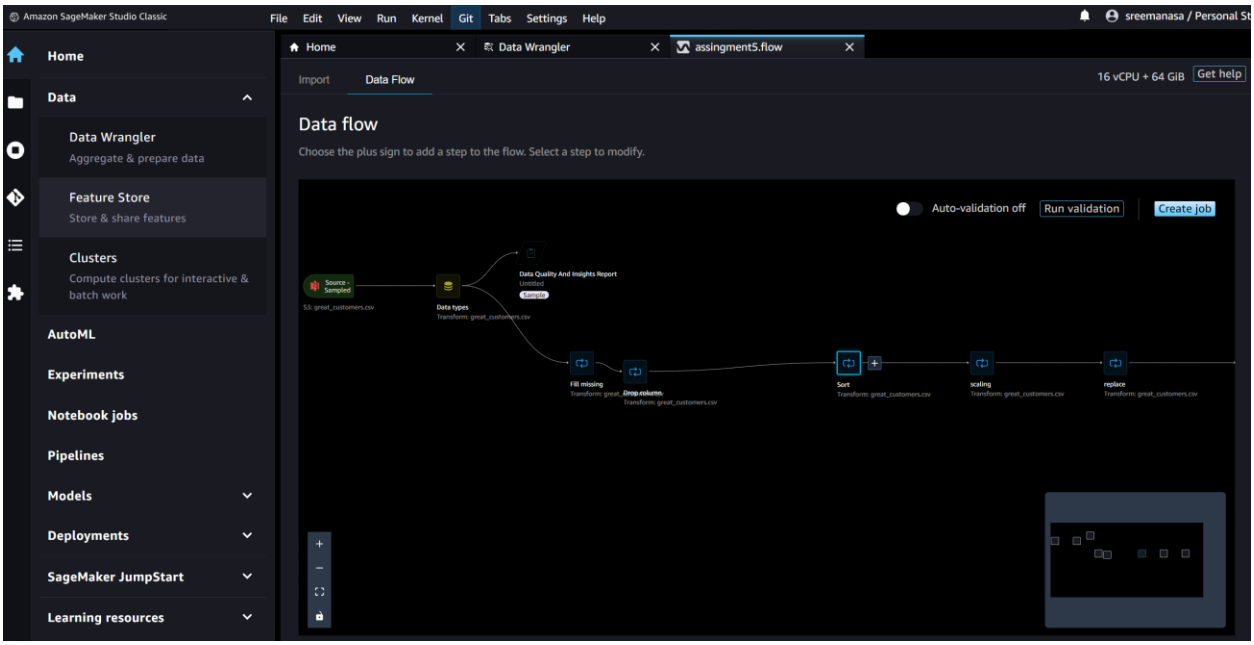
1. Fill Missing values
2. Drop Columns
3. Sort the Dataset according to userId in ascending.
4. Describe the dataset using describe().
5. Replace(custom): Replaced the column Name from salary to Payment.

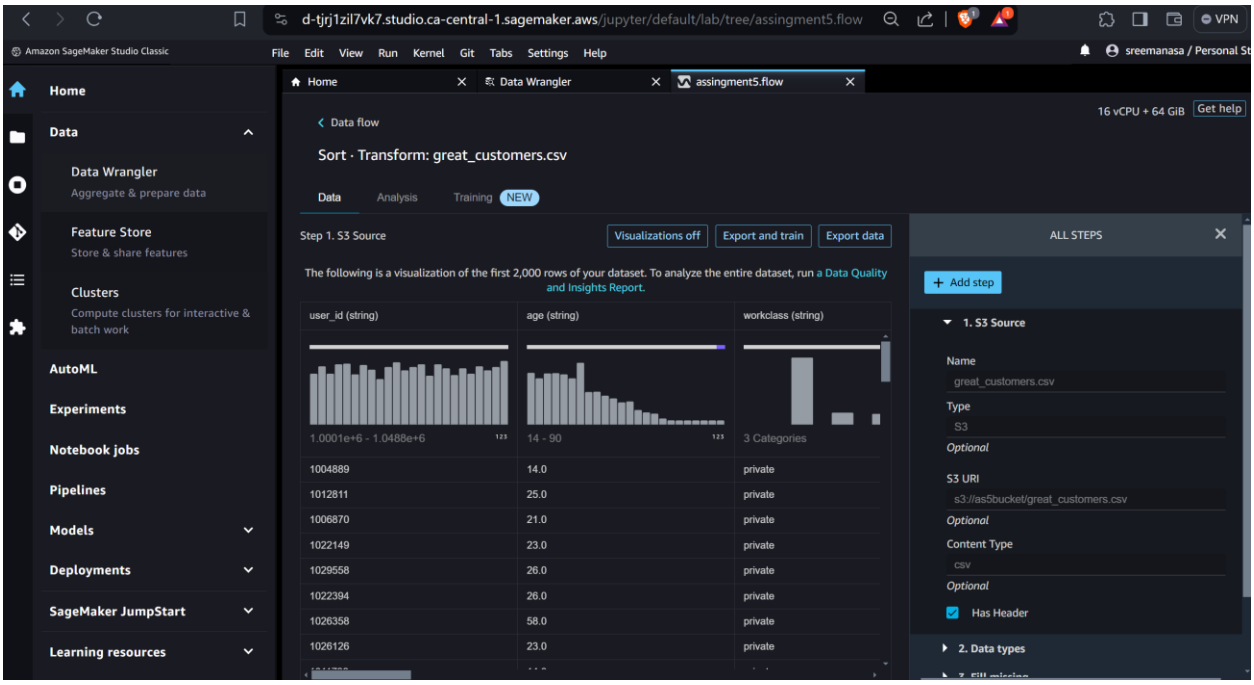
```
import pandas as pd

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

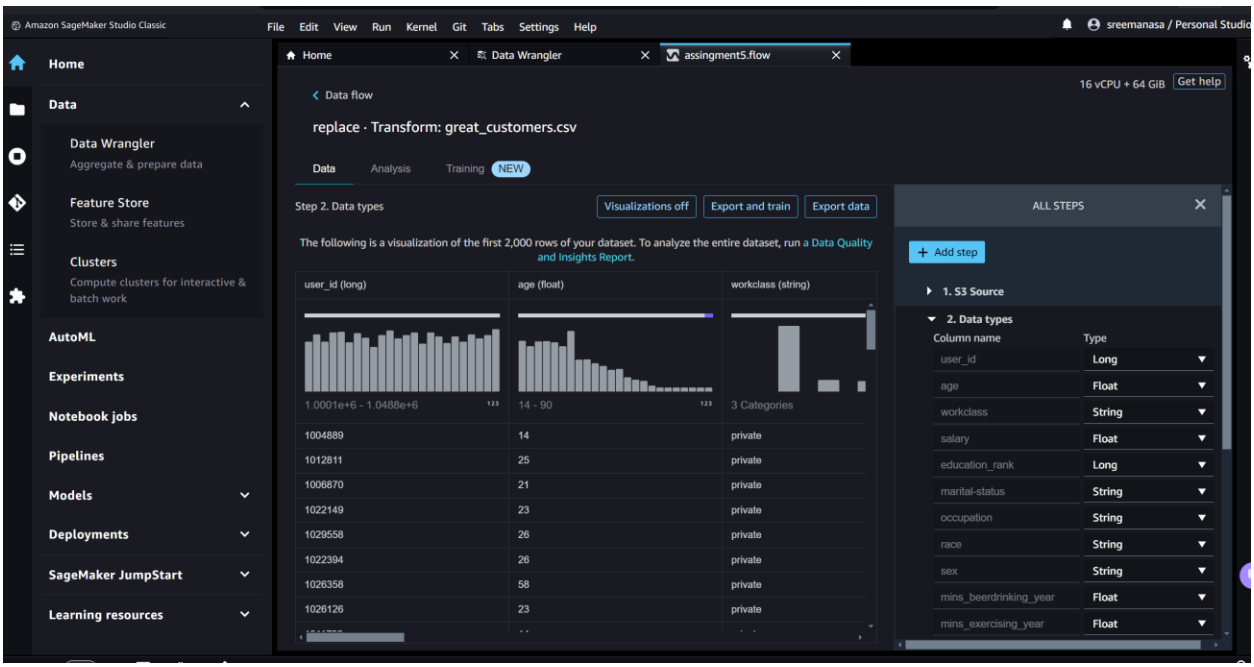
After completing the data analysis, we save the flow in destination in s3 bucket.



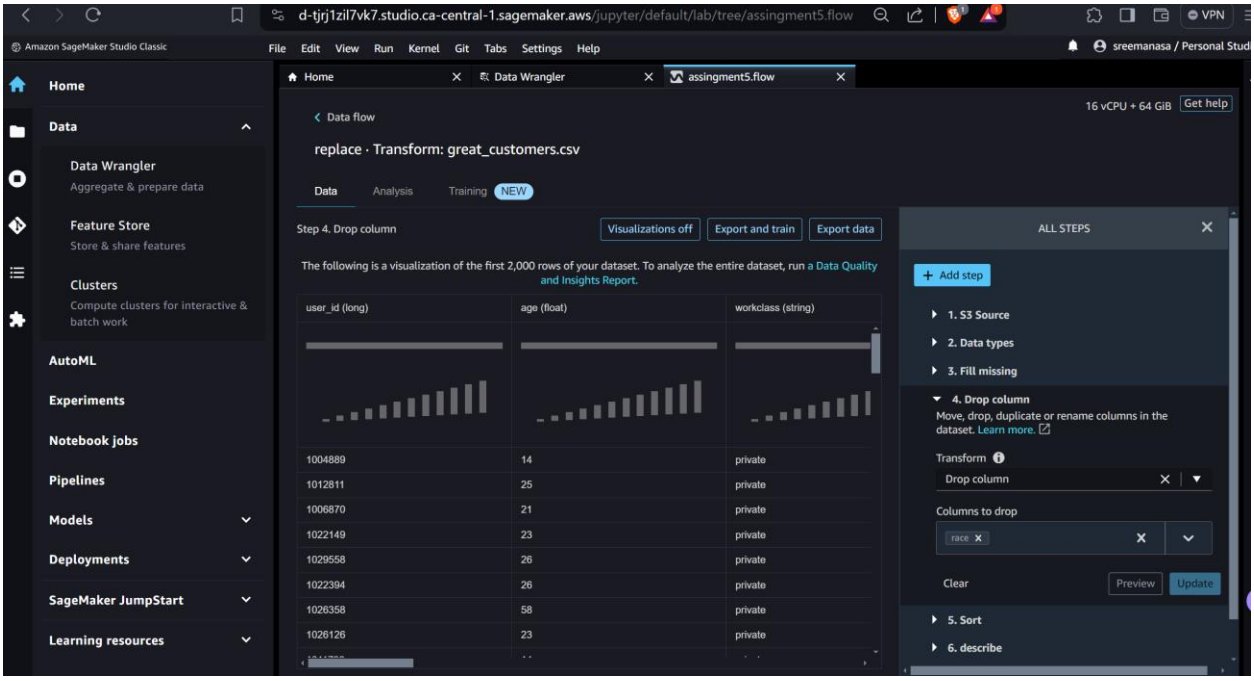



d-tjrj1zil7vk7.studio.ca-central-1.sagemaker.aws/jupyter/default/lab/tree/assingment5.flow
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Learning resources
Home
Data Wrangler
assingment5.flow
Data flow
16 vCPU + 64 GiB
Get help
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

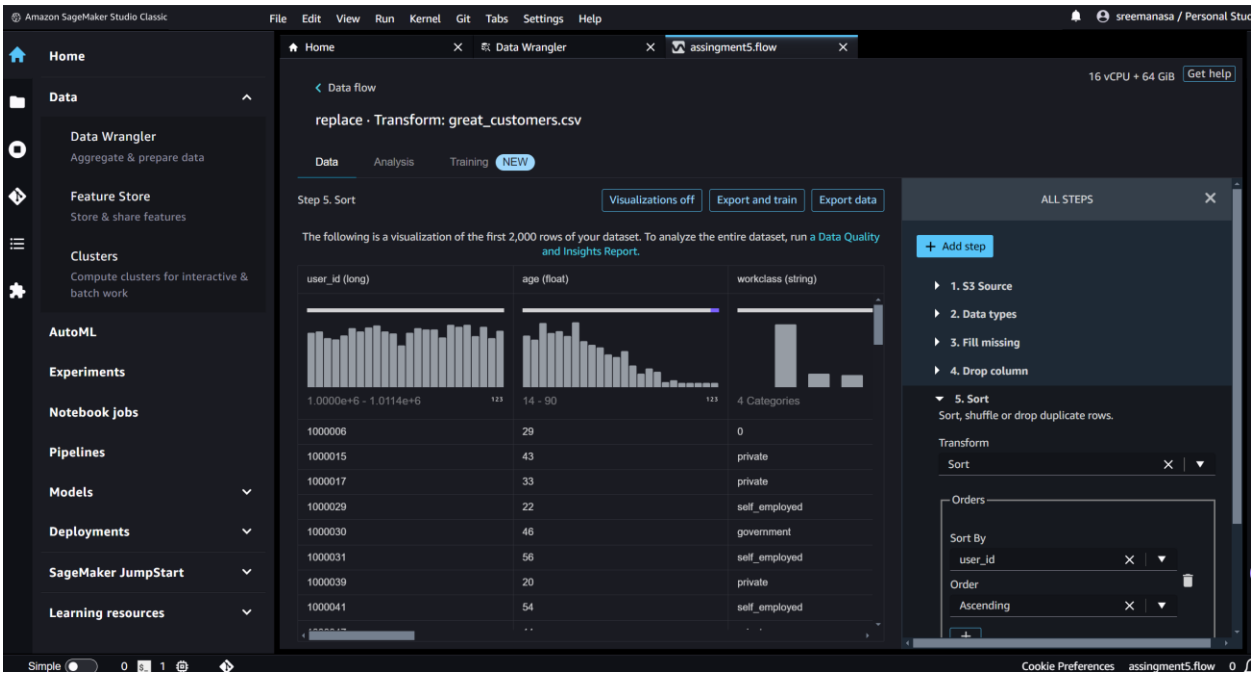
2. Data Types


Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
sreemanasa / Personal Studio
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Learning resources
Home
Data Wrangler
assingment5.flow
Data flow
16 vCPU + 64 GiB
Get help
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

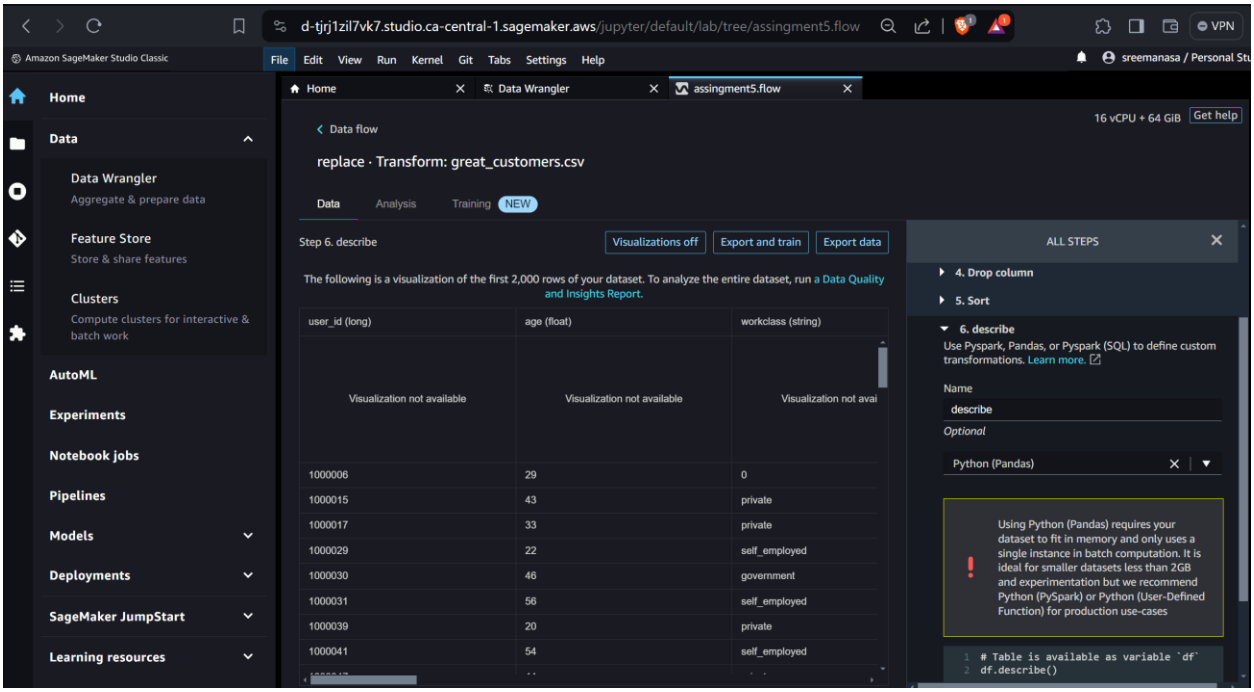
3. Drop Columns



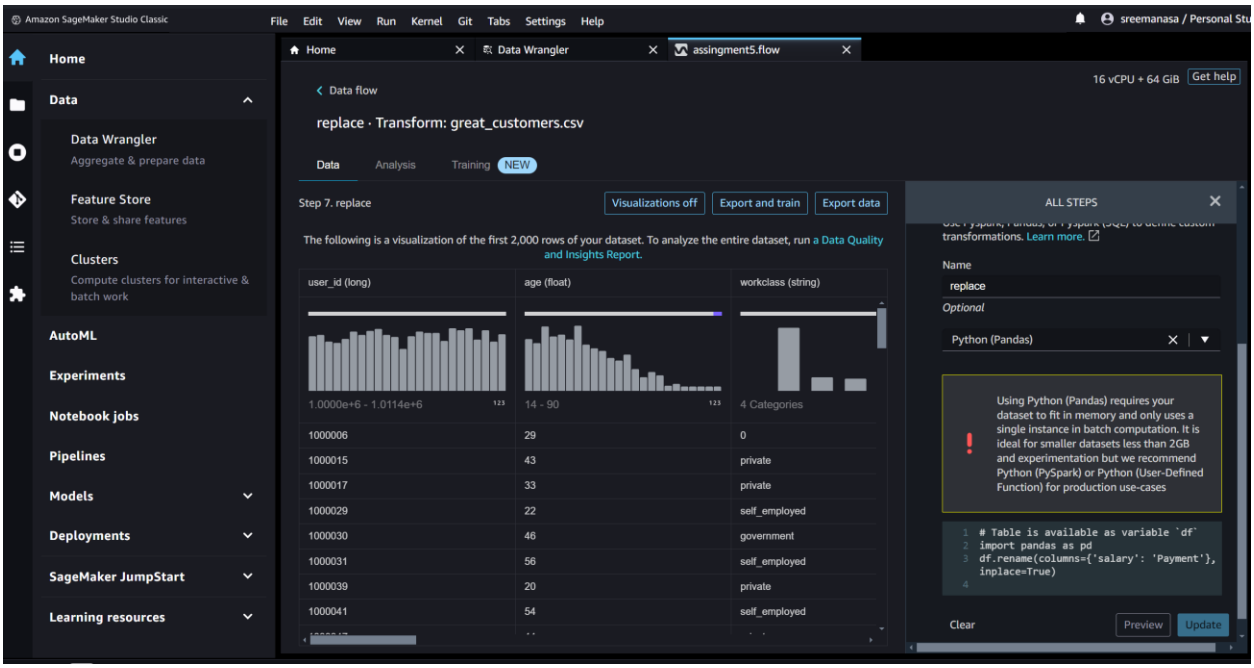
4. Sort Data by userId



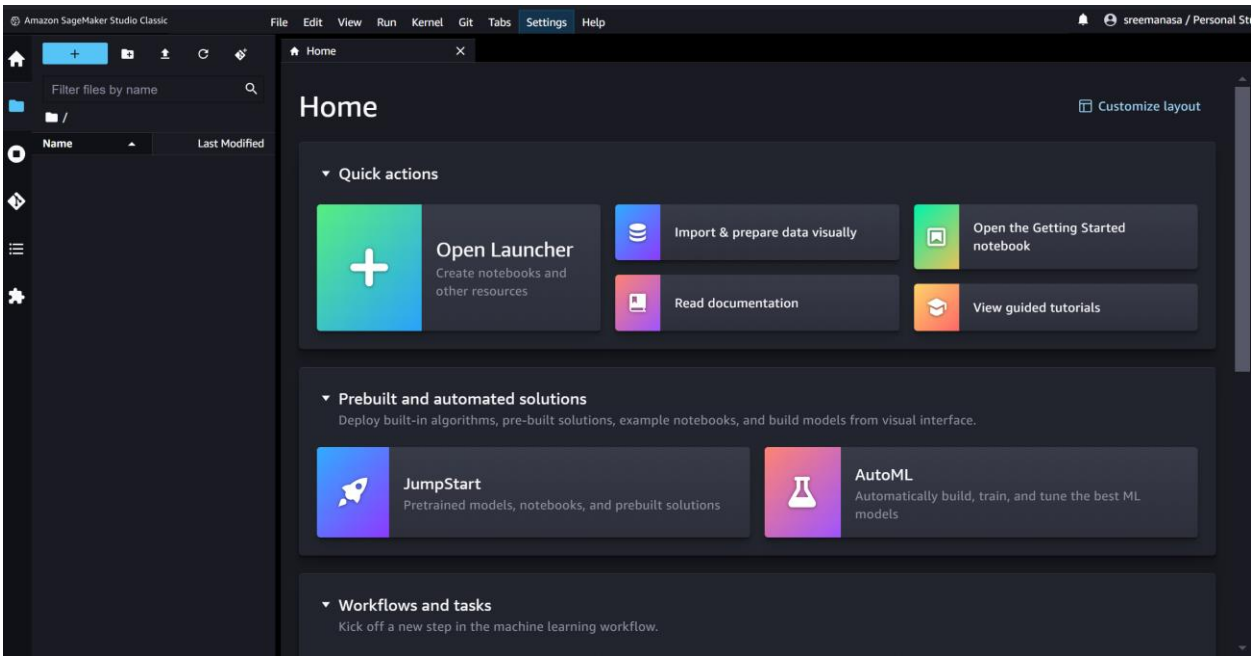
5. Describe


d-tjrj1zil7vk7.studio.ca-central-1.sagemaker.aws/jupyter/default/lab/tree/assingment5.flow
Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Learning resources
Home
Data Wrangler
assingment5.flow
Data flow
16 vCPU + 64 GiB
Get help
replace · Transform: great_customers.csv
Data
Analysis
Training
NEW
Step 6. describe
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)
Visualization not available
Visualization not available
Visualization not avai
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
Use Pyspark, Pandas, or Pyspark (SQL) to define custom transformations. Learn more.
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
# Table is available as variable `df`
df.describe()

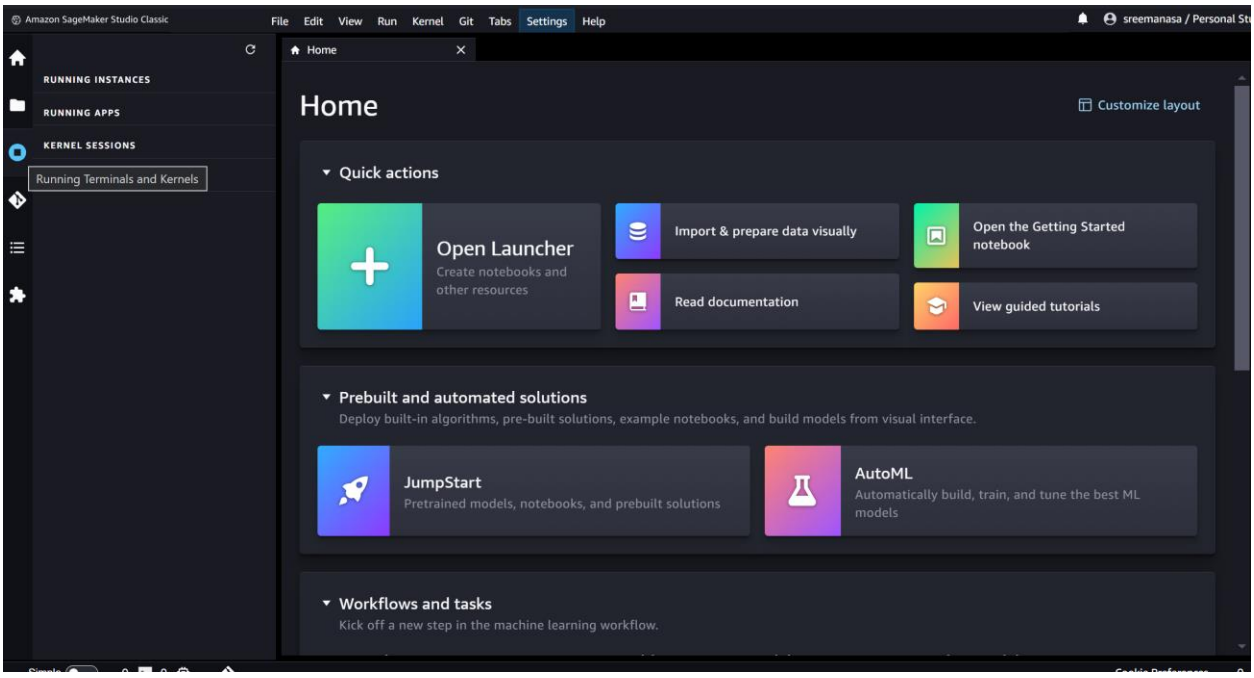
Replace


Amazon SageMaker Studio Classic
File Edit View Run Kernel Git Tabs Settings Help
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Learning resources
Home
Data Wrangler
assingment5.flow
Data flow
16 vCPU + 64 GiB
Get help
replace · Transform: great_customers.csv
Data
Analysis
Training
NEW
Step 7. replace
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)
1.0000e+6 - 1.0114e+6
14 - 90
4 Categories
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
transformations. Learn more.
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
# Table is available as variable `df`
import pandas as pd
df.rename(columns={'salary': 'Payment'}, 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.