Name: Nishit Shaileshbhai Rathod
Student Number: N01586439

Assignment

- Go to https://www.openml.org/
- Select a data set
- 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 has covers at least the suggested ToC (please see next slides)

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

Step 1: Selected the dataset from : OpenML

Step 2: Knowing the dataset.

The dataset is called "baskball" which has 96 instances and 5 features. The dataset was originally taken from –

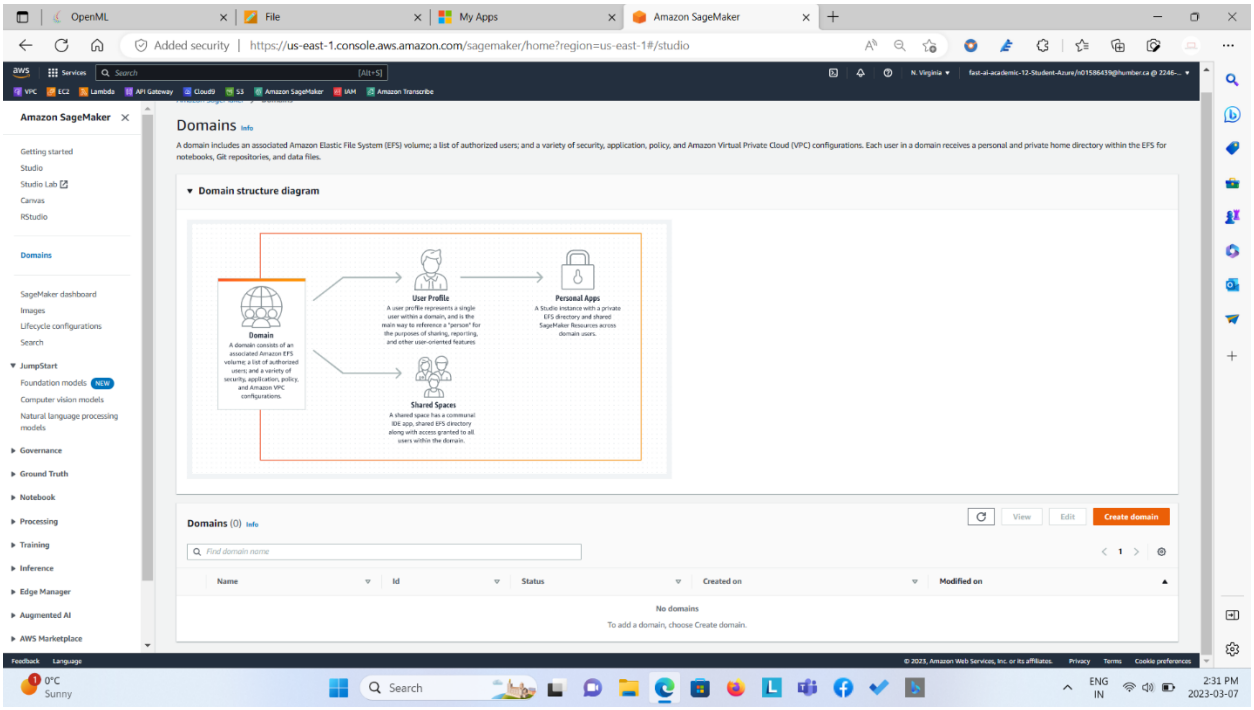
- Dataset from Smoothing Methods in Statistics (ftp stat.cmu.edu/datasets)
- Simonoff, J.S. (1996). Smoothing Methods in Statistics. New York: Springer-Verlag.

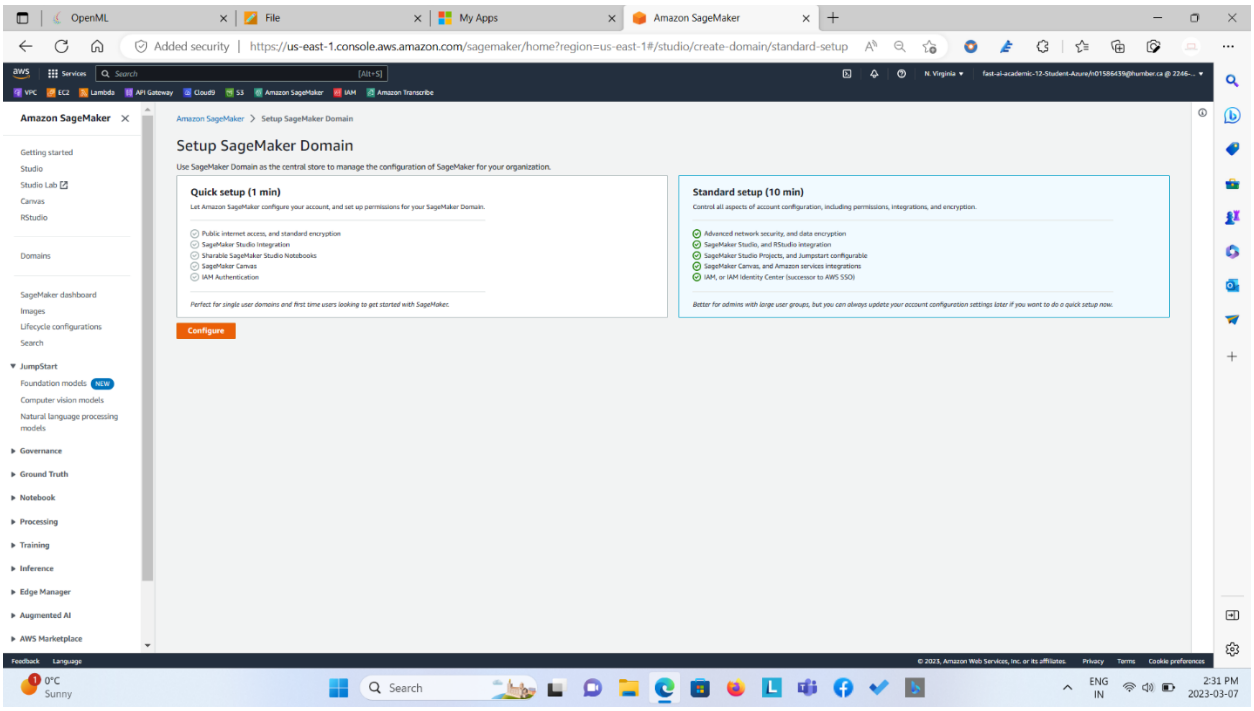
Data Features:

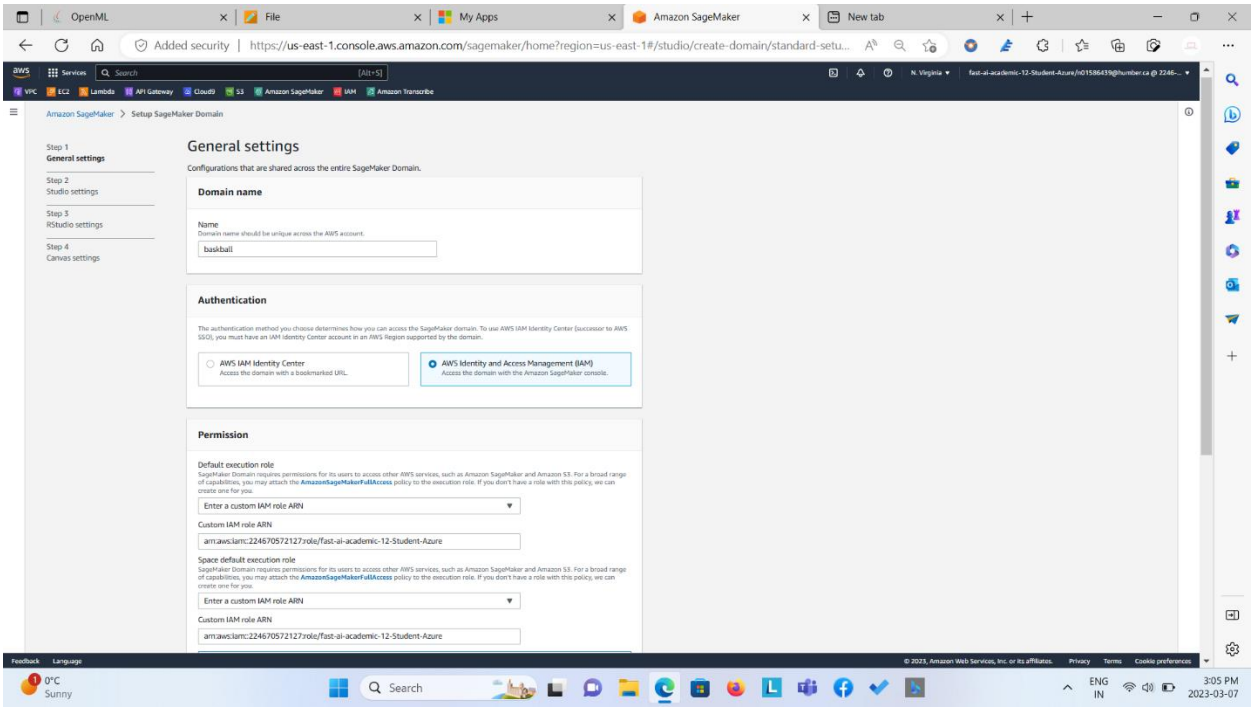
1. **points_per_minute (Target Variable) – Numeric type - 95 distinct values and 0 missing attributes.**
2. **assists_per_minute – Numeric Type - 96 distinct values and 0 missing attributes.**
3. **height – Numeric Type - 13 distinct values and 0 missing attributes.**
4. **time_played – Numeric Type - 94 distinct values and 0 missing attributes.**
5. **age – Numeric Type - 15 distinct values and 0 missing attributes.**

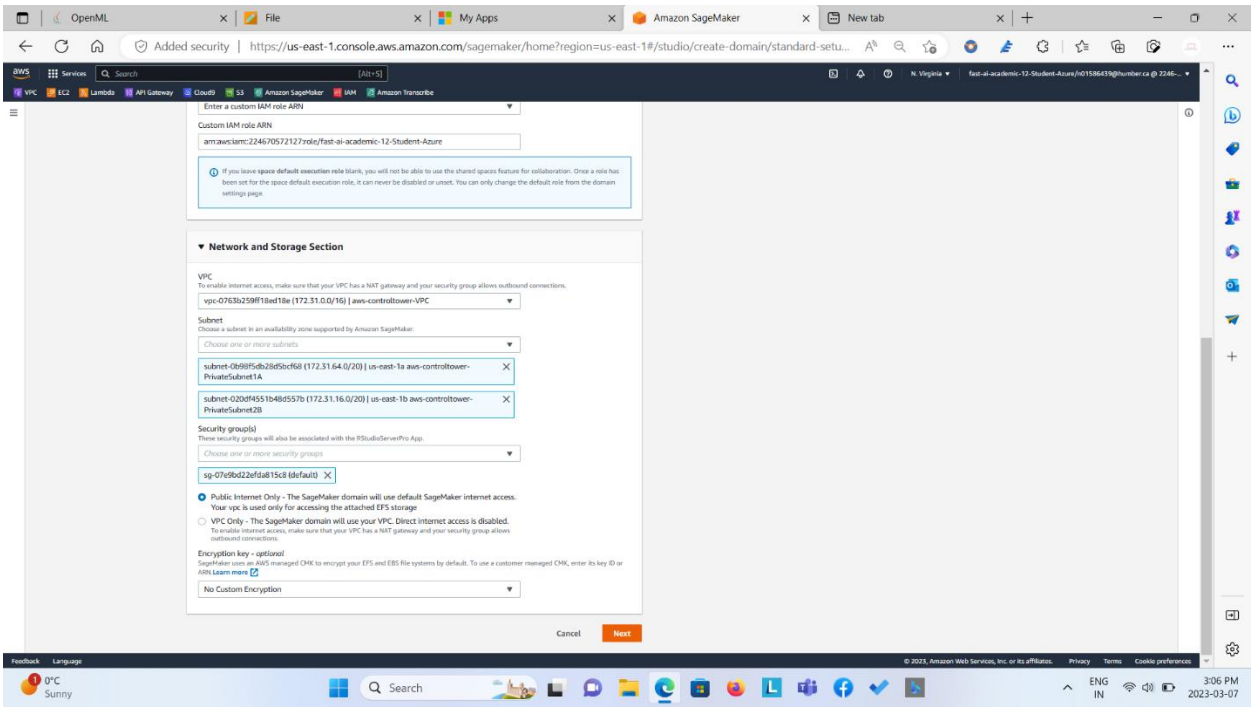
As the dataset didn't have Categorical feature as well as any missing values, I deliberately transformed the dataset and added "Sex" column as well as removed some of the values from the features. By doing this I can know have the hands-on practice in data pre-processing as well as in exploratory data analysis using AWS Sage maker Data Wrangler.

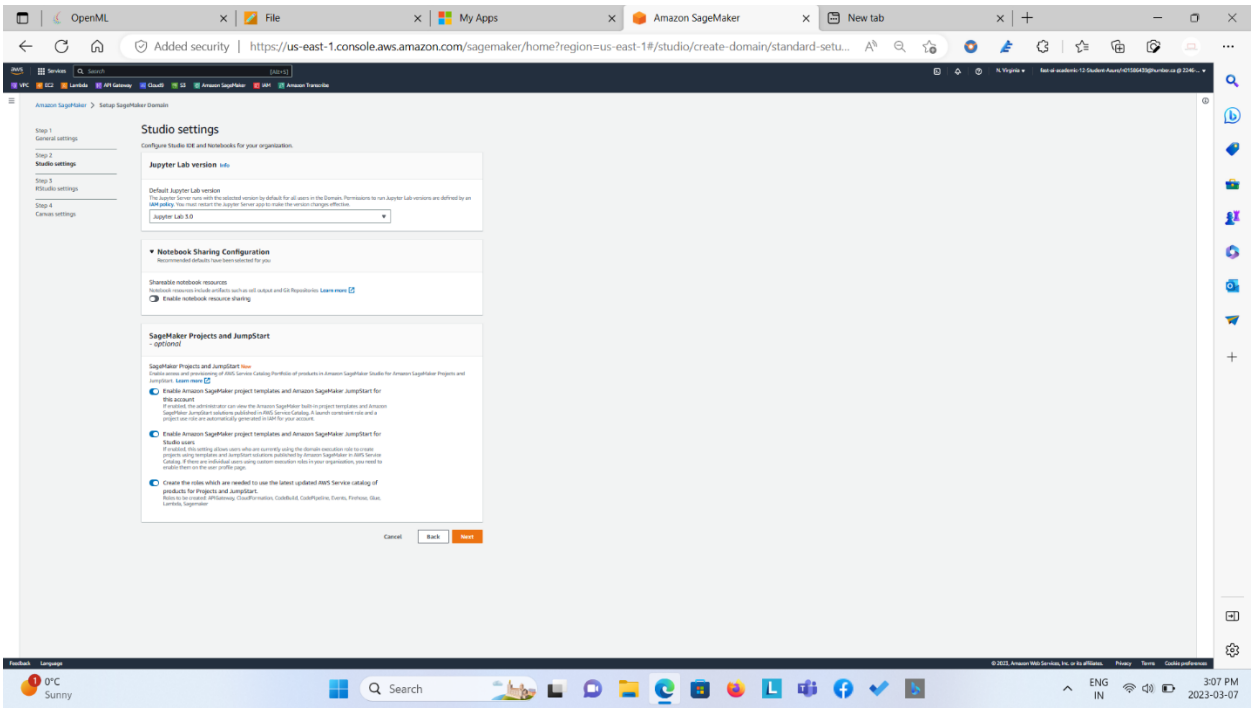
Step 3: Making the Domain, Creating the user in order to get excess to Studio.

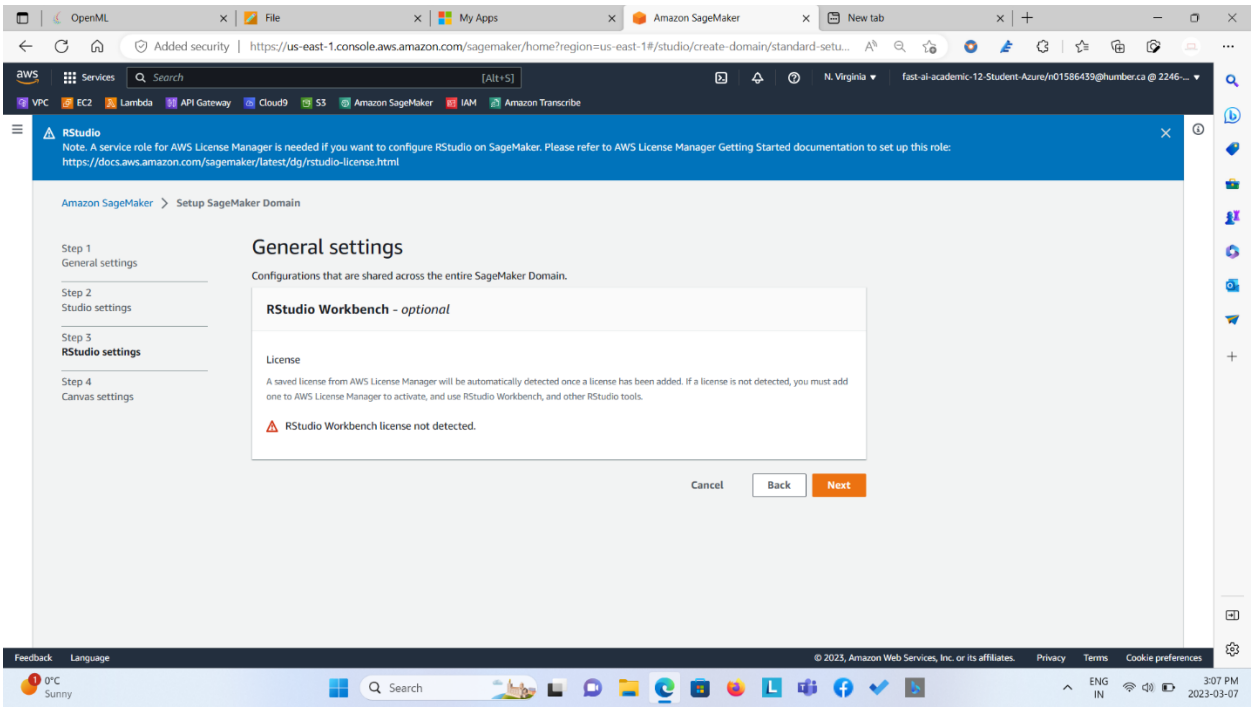


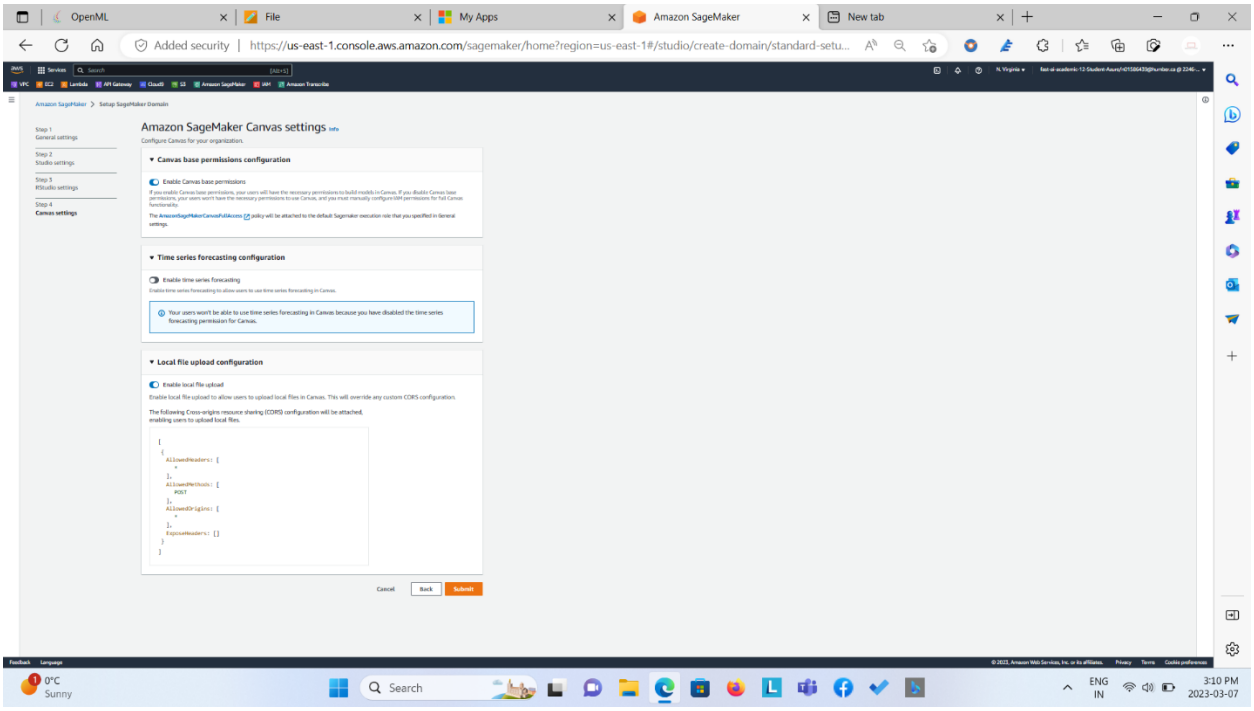

OpenML
File
My Apps
Amazon SageMaker
Added security | https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio/create-domain/standard-setup
Amazon SageMaker
Amazon SageMaker > Setup SageMaker Domain
Setup SageMaker Domain
Use SageMaker Domain as the central store to manage the 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
Getting started
Studio
Studio Lab
Canvas
RStudio
Domains
SageMaker dashboard
Images
Lifecycle configurations
Search
JumpStart
Foundation models
Computer vision models
Natural language processing models
Governance
Ground Truth
Notebook
Processing
Training
Inference
Edge Manager
Augmented AI
AWS Marketplace
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
0°C
Sunny
Search
ENG
IN
2:31 PM
2023-03-07

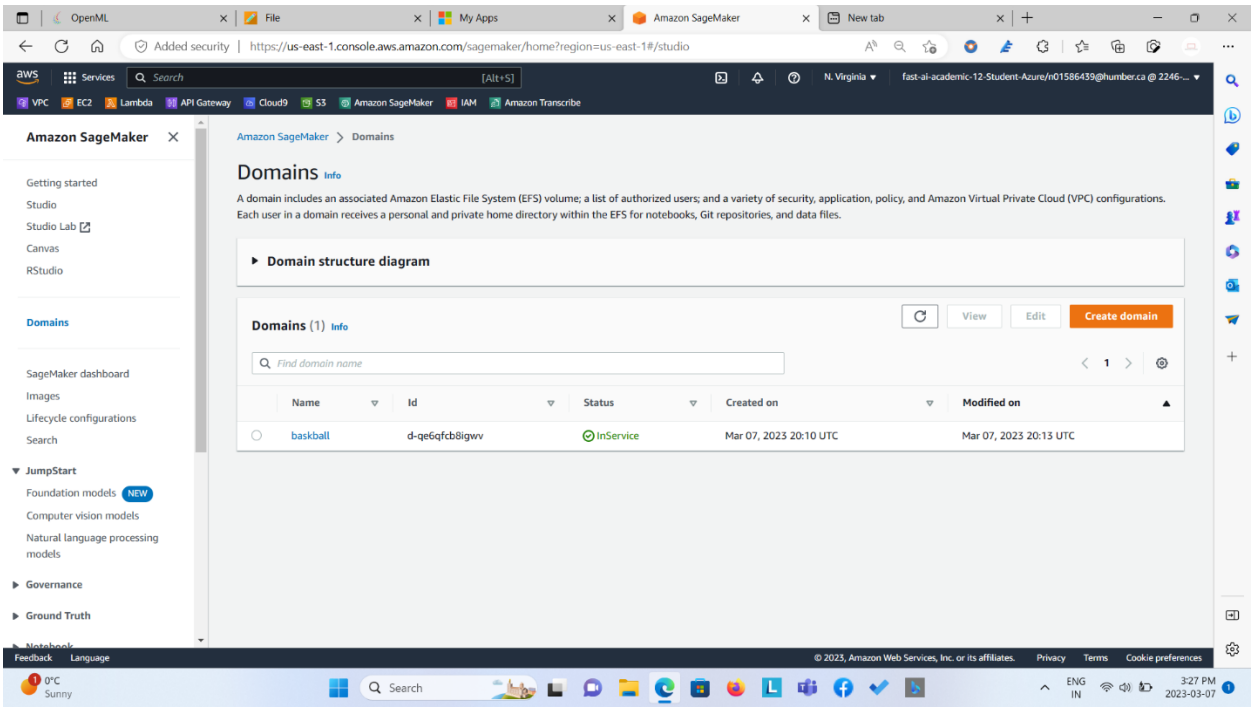

OpenML
File
My Apps
Amazon SageMaker
New tab
Added security | https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio/create-domain/standard-setu...
Amazon SageMaker > Setup SageMaker Domain
Step 1
General settings
Step 2
Studio settings
Step 3
RStudio settings
Step 4
Canvas settings
General settings
Configurations that are shared across the entire SageMaker Domain.
Domain name
Name
Domain name should be unique across the AWS account.
baskball
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
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.
Enter a custom IAM role ARN
Custom IAM role ARN
arn:aws:iam::224670572127:role/fast-ai-academic-12-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.
Enter a custom IAM role ARN
Custom IAM role ARN
arn:aws:iam::224670572127:role/fast-ai-academic-12-Student-Azure
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
0°C
Sunny
Search
ENG
IN
3:05 PM
2023-03-07

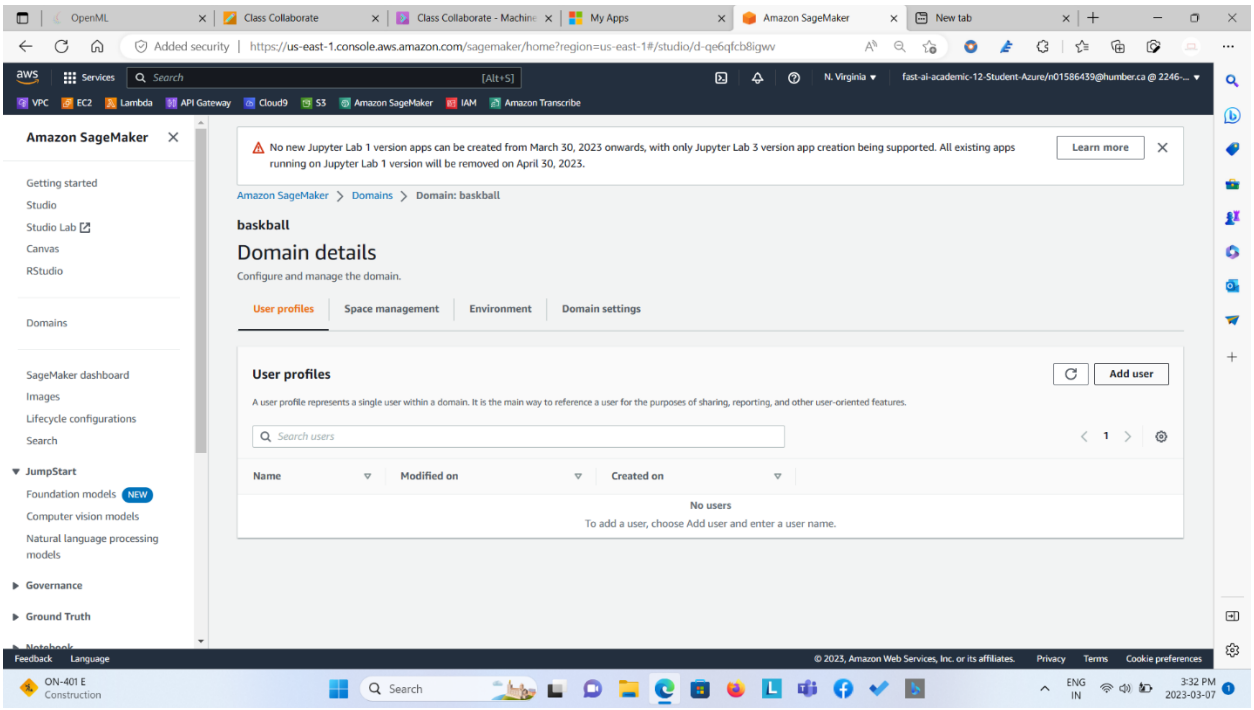

OpenML
File
My Apps
Amazon SageMaker
New tab
Added security | https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio/create-domain/standard-setu...
Enter a custom IAM role ARN
Custom IAM role ARN
arn:aws:iam::224670572127:role/fast-ai-academic-12-Student-Azure
Network and Storage Section
VPC
vpc-0763b259ff18ed18e (172.31.0.0/16) | aws-controltower-VPC
Subnet
Choose one or more subnets
subnet-0b98f5db28d5bcf68 (172.31.64.0/20) | us-east-1a aws-controltower-PrivateSubnet1A
subnet-020df4551b48d557b (172.31.16.0/20) | us-east-1b aws-controltower-PrivateSubnet2B
Security group(s)
Choose one or more security groups
sg-07e9bd22efda815c8 (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.
Encryption key - optional
No Custom Encryption
Cancel
Next
Feedback
Language
Privacy
Terms
Cookie preferences
0°C Sunny
Search
ENG IN
3:06 PM 2023-03-07

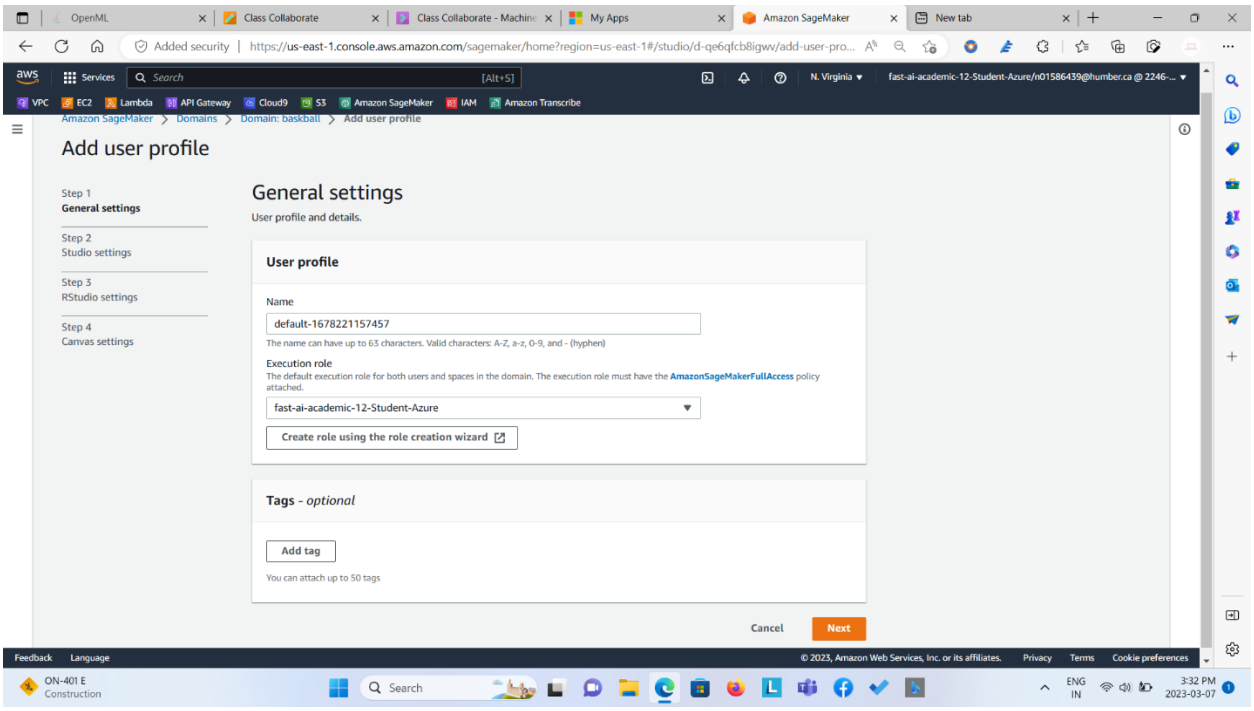

OpenML
File
My Apps
Amazon SageMaker
New tab
Added security | https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio/create-domain/standard-setu...
Amazon SageMaker > Setup SageMaker Domain
Step 1 General settings
Step 2 Studio settings
Step 3 RStudio settings
Step 4 Canvas settings
Studio settings
Configure Studio IDE and Notebooks for your organization.
Jupyter Lab version
Default Jupyter Lab version
Jupyter Lab 3.0
Notebook Sharing Configuration
Shareable notebook resources
Enable notebook resource sharing
SageMaker Projects and JumpStart
- optional
Enable Amazon SageMaker project templates and Amazon SageMaker JumpStart for this account
Enable Amazon SageMaker project templates and Amazon SageMaker JumpStart for Studio users
Create the roles which are needed to use the latest updated AWS Service catalog of products for Projects and JumpStart.
Cancel
Back
Next
Feedback
Language
Privacy
Terms
Cookie preferences
0°C Sunny
Search
ENG IN
3:07 PM 2023-03-07

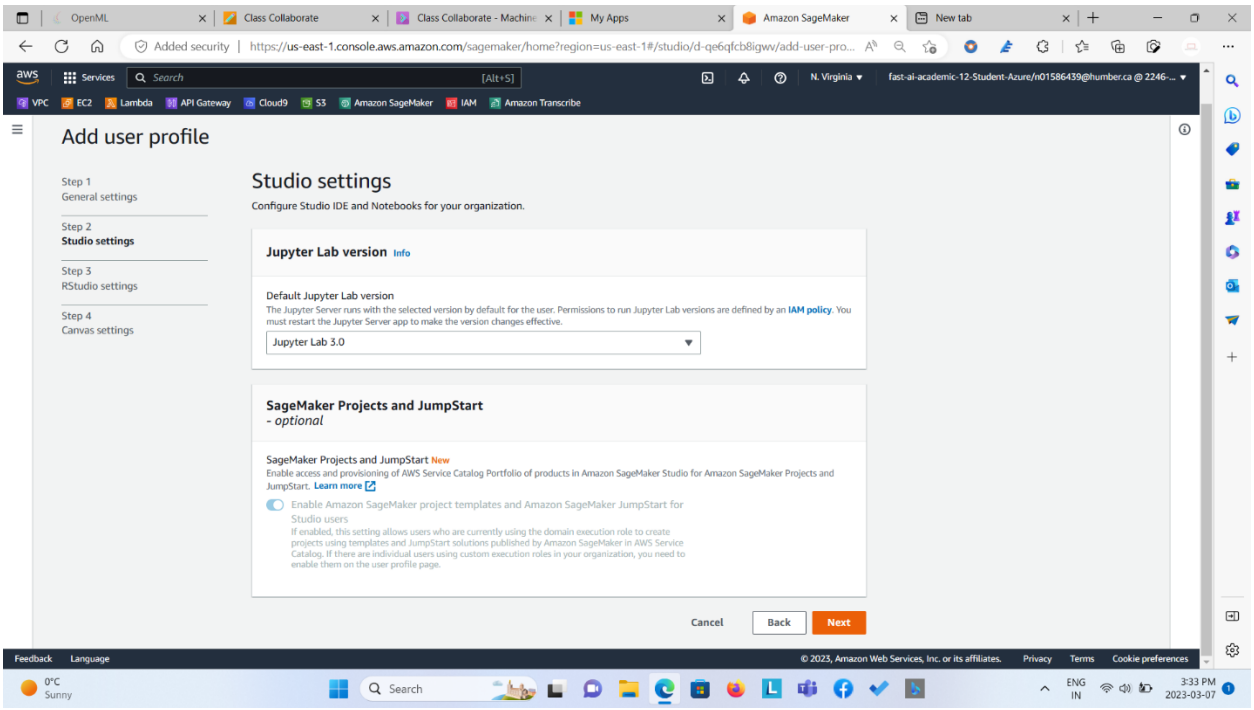

OpenML
File
My Apps
Amazon SageMaker
New tab
Added security | https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio/create-domain/standard-setu...
Services
Search
[Alt+S]
N. Virginia
VPC
EC2
Lambda
API Gateway
Cloud9
S3
Amazon SageMaker
IAM
Amazon Transcribe
RStudio
Note. A service role for AWS License Manager is needed if you want to configure RStudio on SageMaker. Please refer to AWS License Manager Getting Started documentation to set up this role:
https://docs.aws.amazon.com/sagemaker/latest/dg/rstudio-license.html
Amazon SageMaker > Setup SageMaker Domain
Step 1
General settings
Step 2
Studio settings
Step 3
RStudio settings
Step 4
Canvas settings
General settings
Configurations that are shared across the entire SageMaker Domain.
RStudio Workbench - optional
License
A saved license from AWS License Manager will be automatically detected once a license has been added. If a license is not detected, you must add one to AWS License Manager to activate, and use RStudio Workbench, and other RStudio tools.
RStudio Workbench license not detected.
Cancel
Back
Next
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
0°C
Sunny
Search
ENG
IN
3:07 PM
2023-03-07

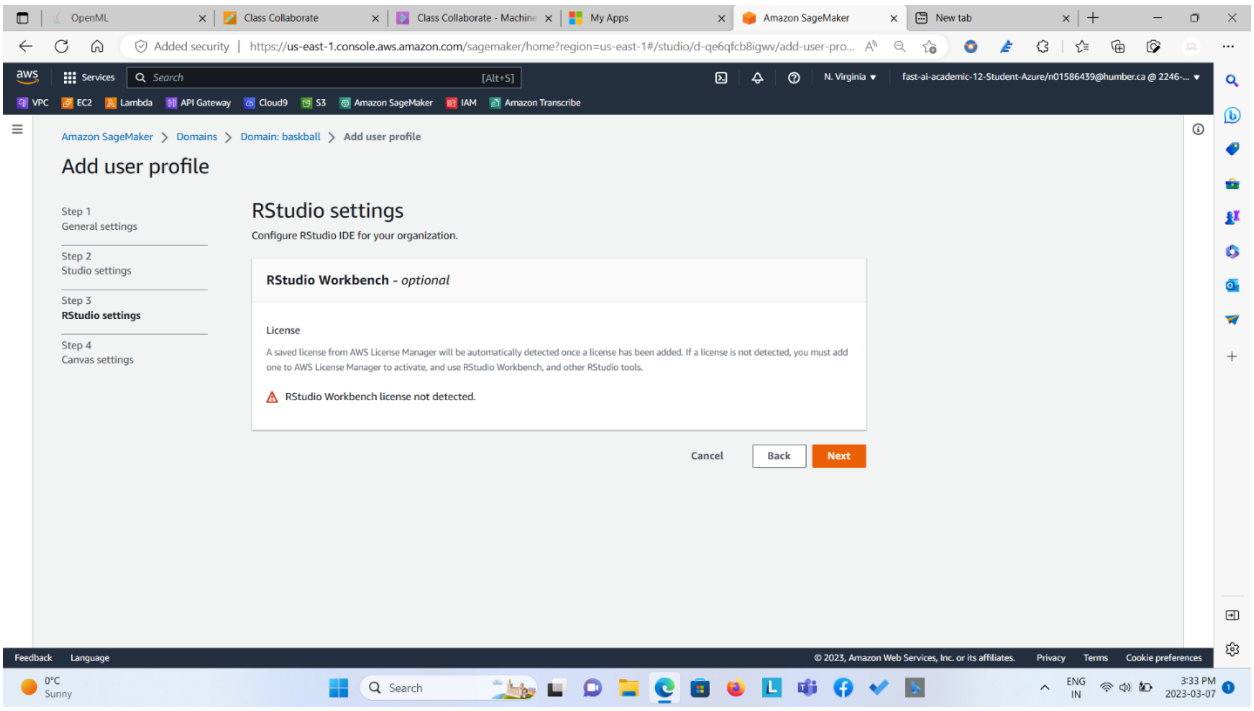

OpenML
File
My Apps
Amazon SageMaker
New tab
Added security | https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio/create-domain/standard-setu...
Amazon SageMaker Canvas settings
Canvas base permissions configuration
Time series forecasting configuration
Local file upload configuration
Cancel
Back
Submit
0°C
Sunny
Search
ENG
IN
3:10 PM
2023-03-07

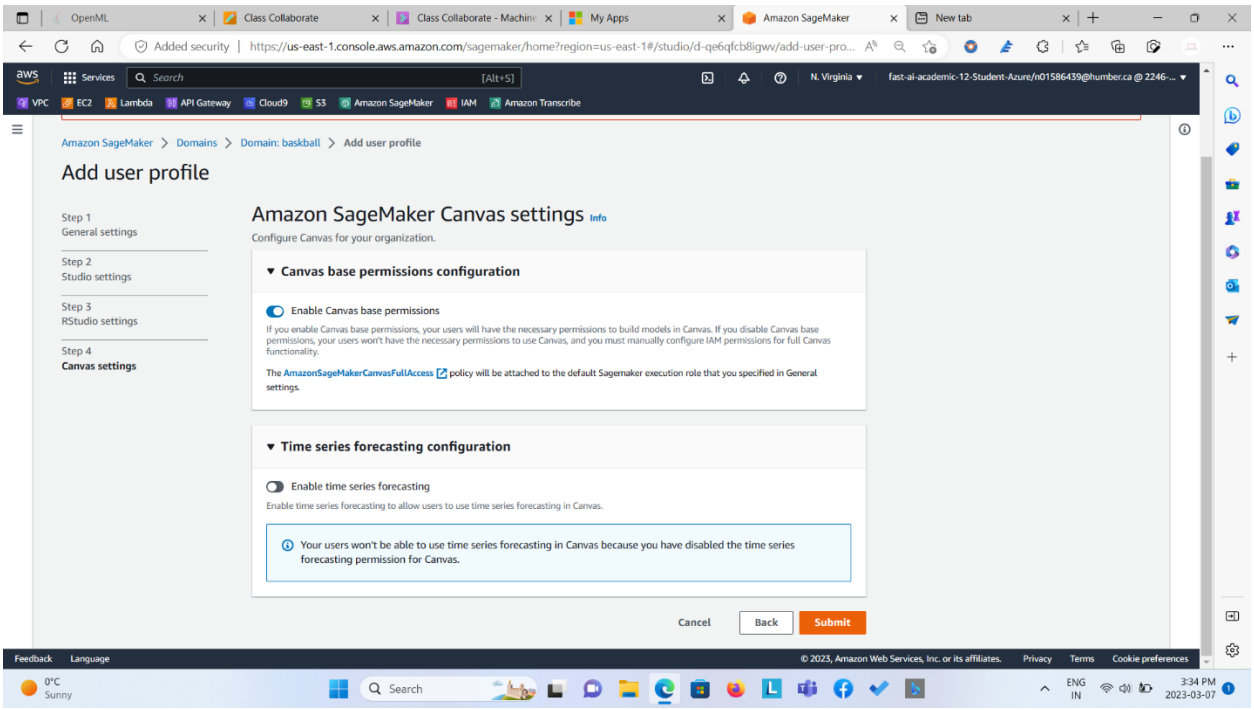

OpenML
File
My Apps
Amazon SageMaker
New tab
https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio
Services
Search
[Alt+S]
N. Virginia
VPC
EC2
Lambda
API Gateway
Cloud9
S3
Amazon SageMaker
IAM
Amazon Transcribe
Amazon SageMaker
Getting started
Studio
Studio Lab
Canvas
RStudio
Domains
SageMaker dashboard
Images
Lifecycle configurations
Search
JumpStart
Foundation models
NEW
Computer vision models
Natural language processing models
Governance
Ground Truth
Amazon SageMaker > Domains
Domains Info
A domain includes an associated Amazon Elastic File System (EFS) volume; a list of authorized users; and a variety of security, application, policy, and Amazon Virtual Private Cloud (VPC) configurations. Each user in a domain receives a personal and private home directory within the EFS for notebooks, Git repositories, and data files.
Domain structure diagram
Domains (1) Info
View
Edit
Create domain
Find domain name
Name
Id
Status
Created on
Modified on
baskball
d-qe6qfcb8igwv
InService
Mar 07, 2023 20:10 UTC
Mar 07, 2023 20:13 UTC
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
0°C
Sunny
Search
ENG
IN
3:27 PM
2023-03-07

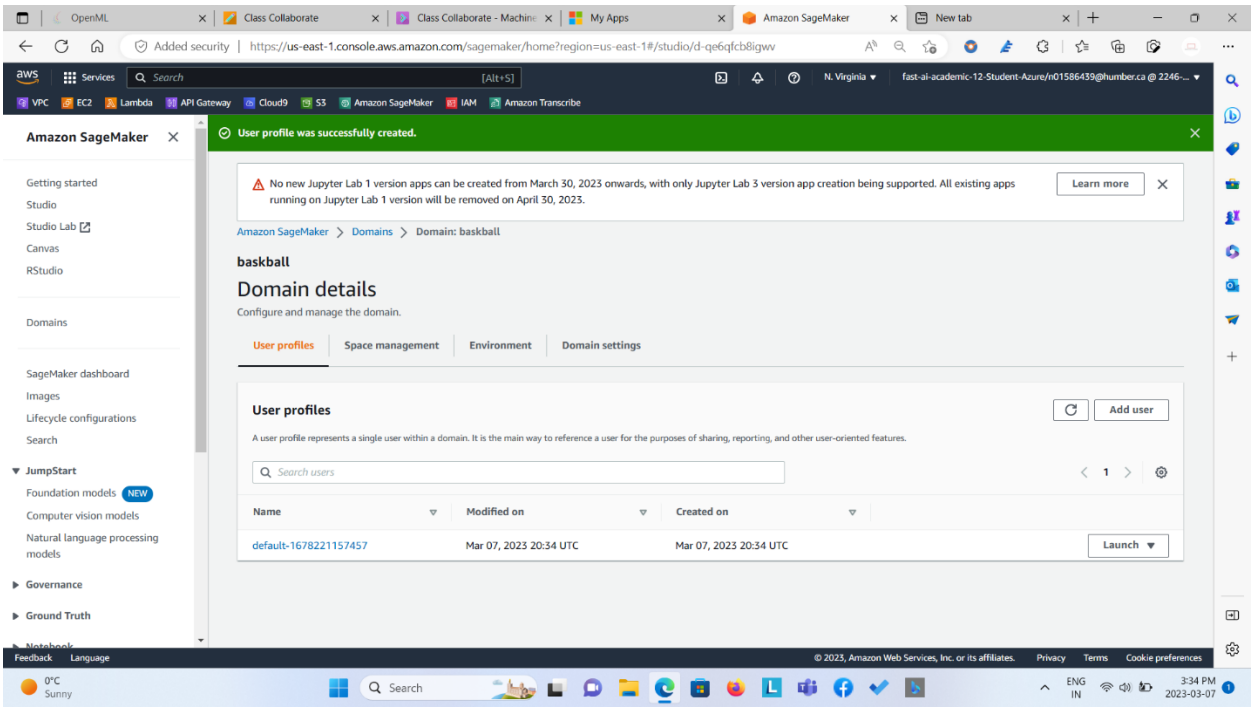

OpenML
Class Collaborate
Class Collaborate - Machine
My Apps
Amazon SageMaker
New tab
https://us-east-1.console.aws.amazon.com/sagemaker/home?region=us-east-1#/studio/d-qe6qfcb8igwv
Services
Search
[Alt+S]
N. Virginia
VPC
EC2
Lambda
API Gateway
Cloud9
S3
Amazon SageMaker
IAM
Amazon Transcribe
Amazon SageMaker
Getting started
Studio
Studio Lab
Canvas
RStudio
Domains
SageMaker dashboard
Images
Lifecycle configurations
Search
JumpStart
Foundation models
NEW
Computer vision models
Natural language processing models
Governance
Ground Truth
No new Jupyter Lab 1 version apps can be created from March 30, 2023 onwards, with only Jupyter Lab 3 version app creation being supported. All existing apps running on Jupyter Lab 1 version will be removed on April 30, 2023.
Learn more
Amazon SageMaker > Domains > Domain: baskball
baskball
Domain details
Configure and manage the domain.
User profiles
Space management
Environment
Domain settings
User profiles
Add user
A user profile represents a single user within a domain. It is the main way to reference a user for the purposes of sharing, reporting, and other user-oriented features.
Search users
Name
Modified on
Created on
No users
To add a user, choose Add user and enter a user name.
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
ON-401 E
Construction
Search
ENG
IN
3:32 PM
2023-03-07


Add user profile
Step 1
General settings
Step 2
Studio settings
Step 3
RStudio settings
Step 4
Canvas settings
General settings
User profile and details.
User profile
Name
default-1678221157457
The name can have up to 63 characters. Valid characters: A-Z, a-z, 0-9, and - (hyphen)
Execution role
The default execution role for both users and spaces in the domain. The execution role must have the AmazonSageMakerFullAccess policy attached.
fast-ai-academic-12-Student-Azure
Create role using the role creation wizard
Tags - optional
Add tag
You can attach up to 50 tags
Cancel
Next

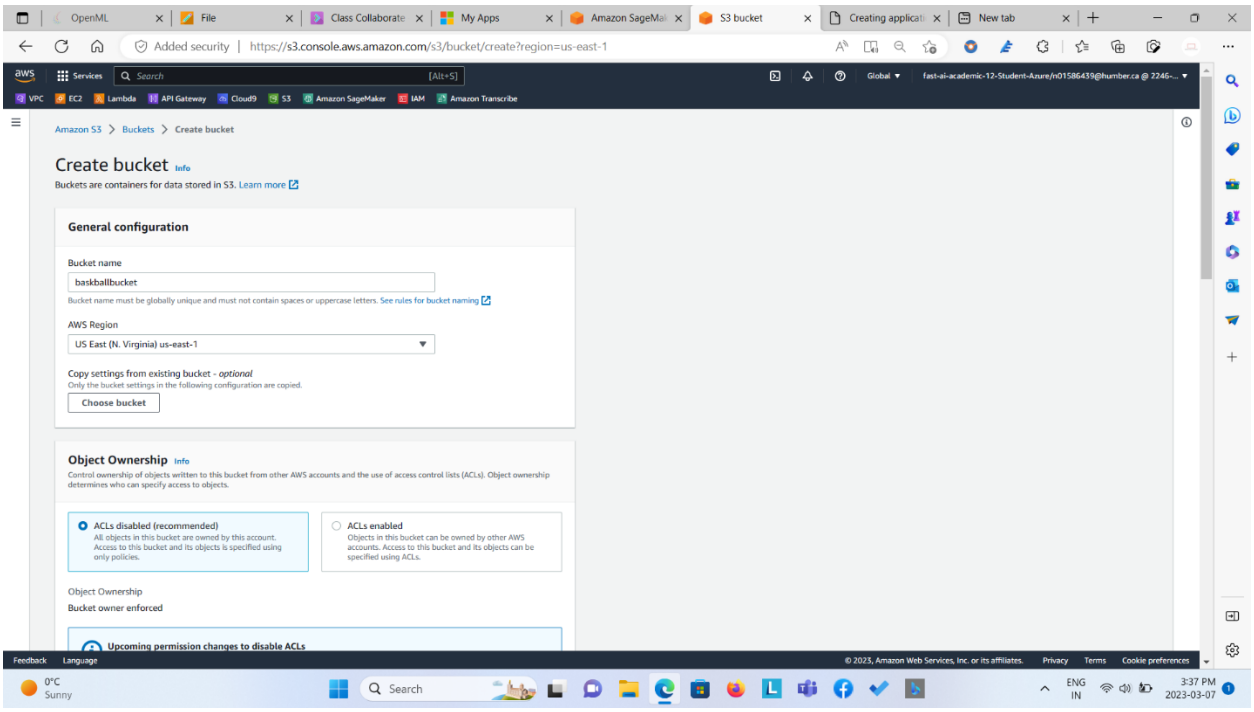

Add user profile
Step 1
General settings
Step 2
Studio settings
Step 3
RStudio settings
Step 4
Canvas settings
Studio settings
Configure Studio IDE and Notebooks for your organization.
Jupyter Lab version Info
Default Jupyter Lab version
The Jupyter Server runs with the selected version by default for the user. Permissions to run Jupyter Lab versions are defined by an IAM policy. You must restart the Jupyter Server app to make the version changes effective.
Jupyter Lab 3.0
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 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.
Cancel
Back
Next

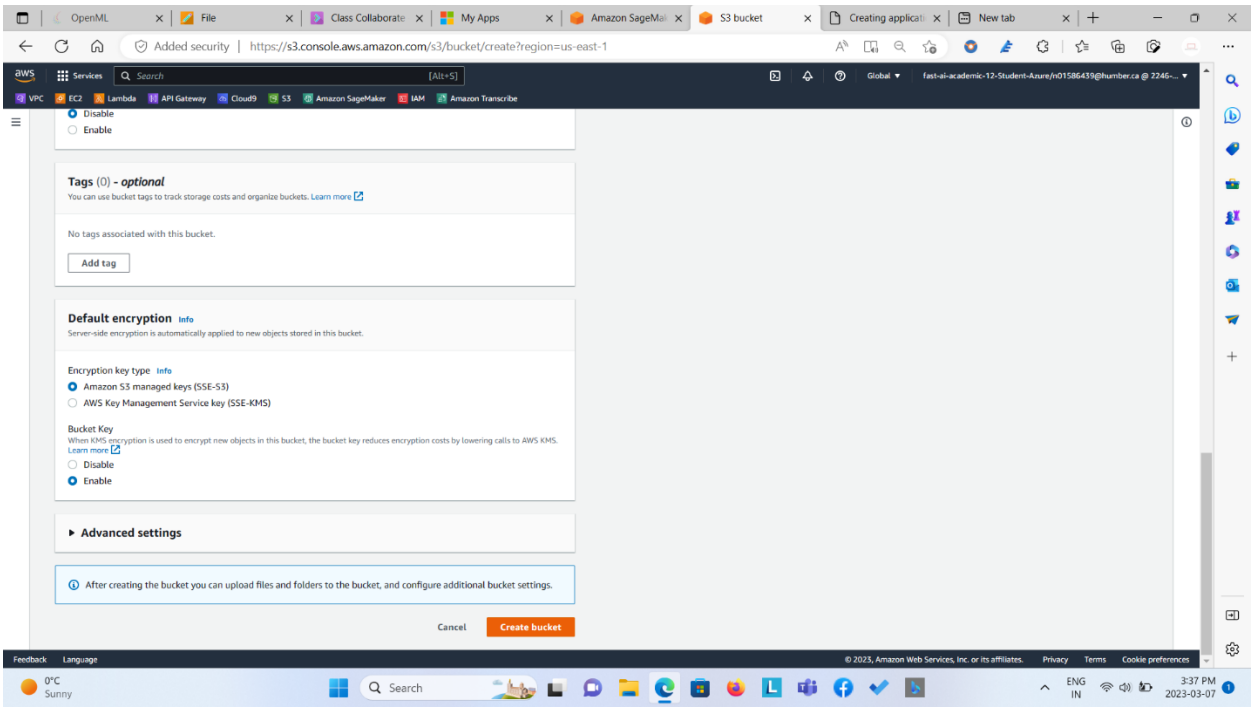

Amazon SageMaker > Domains > Domain: baskball > Add user profile
Add user profile
Step 1
General settings
Step 2
Studio settings
Step 3
RStudio settings
Step 4
Canvas settings
RStudio settings
Configure RStudio IDE for your organization.
RStudio Workbench - optional
License
A saved license from AWS License Manager will be automatically detected once a license has been added. If a license is not detected, you must add one to AWS License Manager to activate, and use RStudio Workbench, and other RStudio tools.
RStudio Workbench license not detected.
Cancel
Back
Next

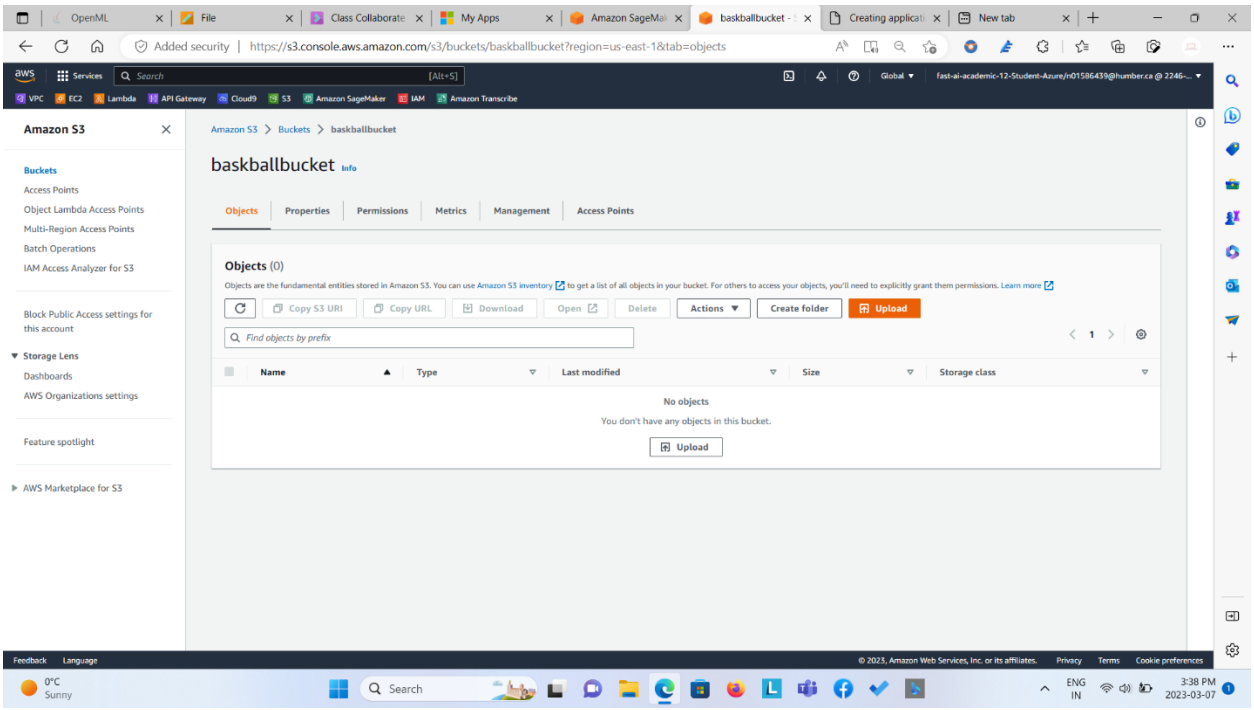

Amazon SageMaker > Domains > Domain: baskball > Add user profile
Add user profile
Step 1
General settings
Step 2
Studio settings
Step 3
RStudio settings
Step 4
Canvas settings
Amazon SageMaker Canvas settings Info
Configure Canvas for your organization.
Canvas base permissions configuration
Enable Canvas 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 AmazonSageMakerCanvasFullAccess policy will be attached to the default Sagemaker execution role that you 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.
Your users won't be able to use time series forecasting in Canvas because you have disabled the time series forecasting permission for Canvas.
Cancel
Back
Submit

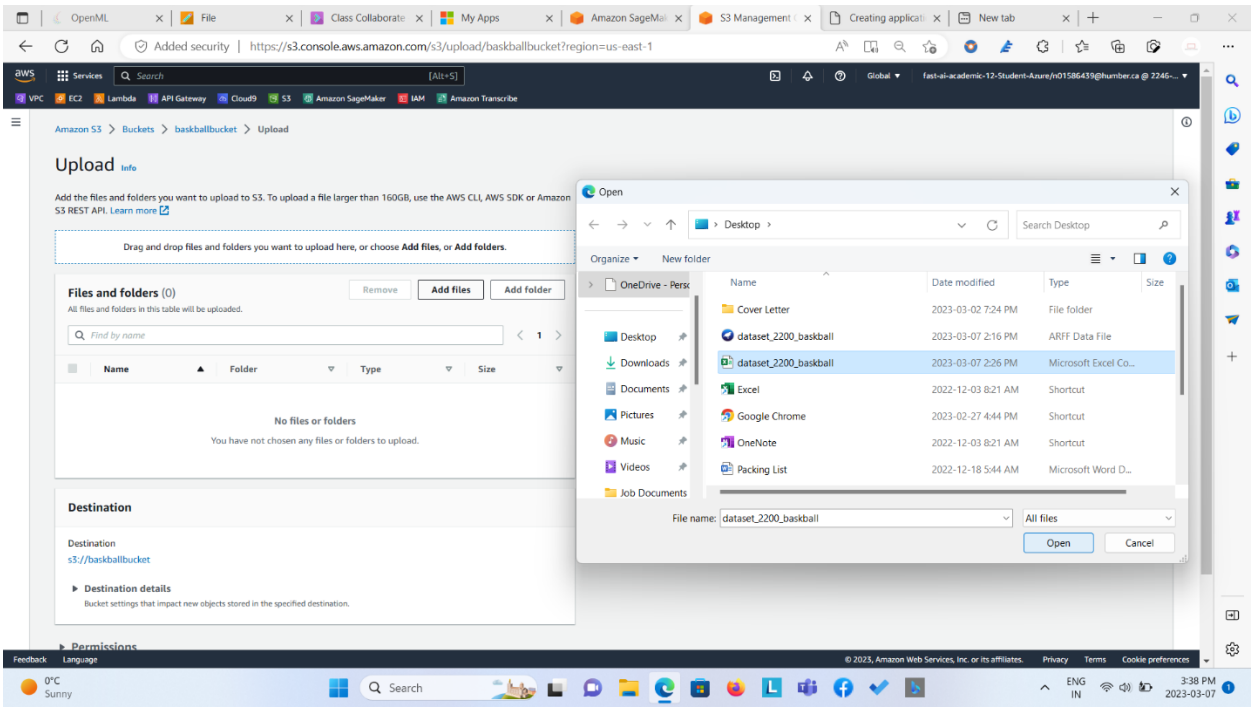


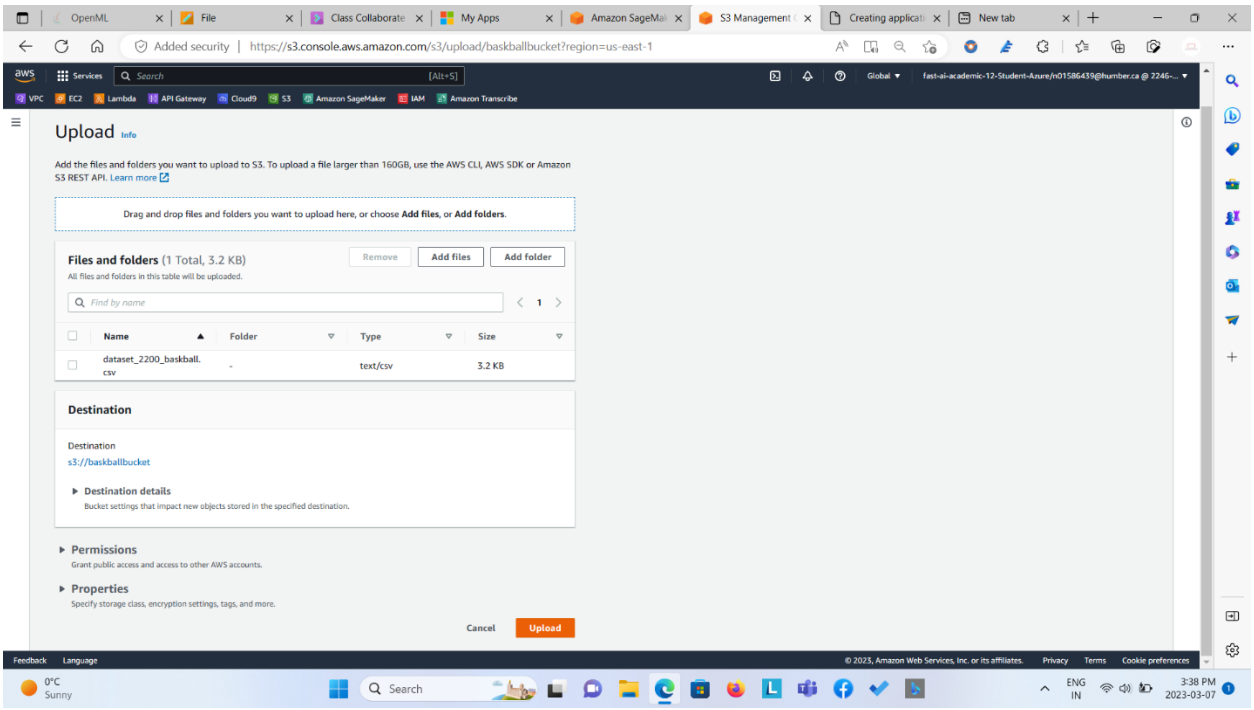
Step 4: Creating S3 Bucket and Uploading the dataset (.csv file) in it.

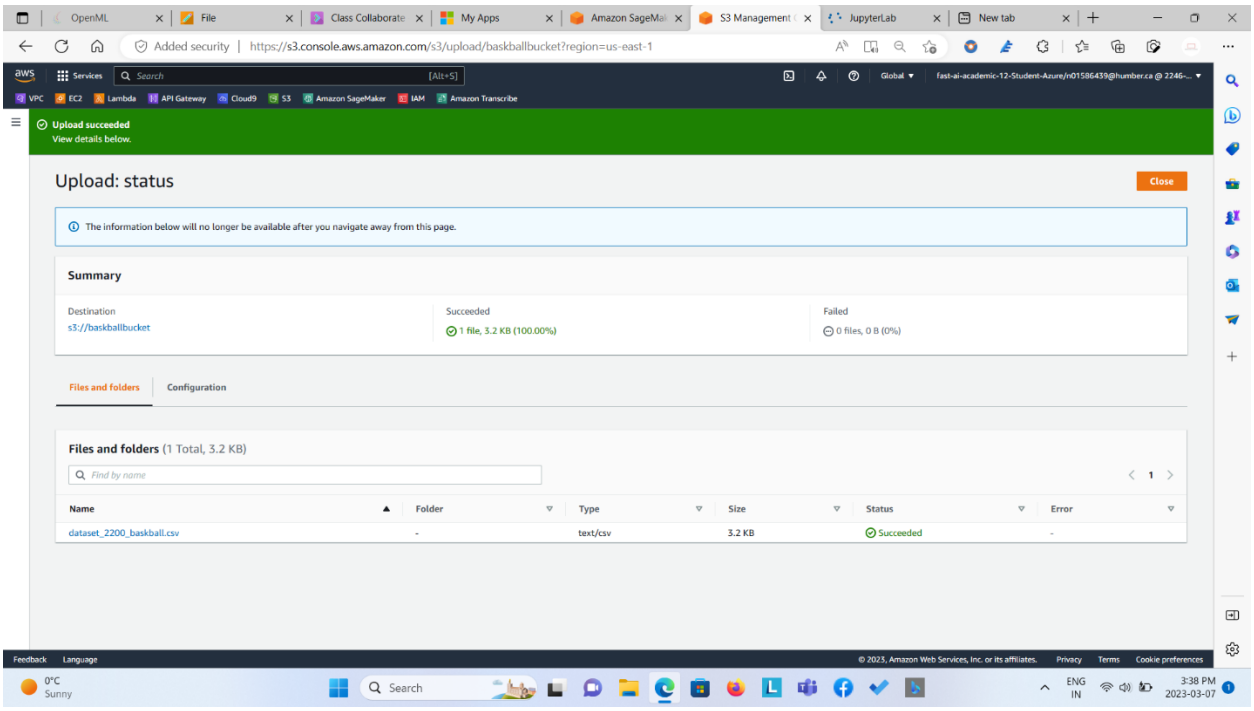



Disable
Enable
Tags (0) - optional
You can use bucket tags to track storage costs and organize buckets. Learn more
No tags associated with this bucket.
Add tag
Default encryption Info
Server-side encryption is automatically applied to new objects stored in this bucket.
Encryption key type Info
Amazon S3 managed keys (SSE-S3)
AWS Key Management Service key (SSE-KMS)
Bucket Key
When KMS encryption is used to encrypt new objects in this bucket, the bucket key reduces encryption costs by lowering calls to AWS KMS. Learn more
Disable
Enable
Advanced settings
After creating the bucket you can upload files and folders to the bucket, and configure additional bucket settings.
Cancel
Create bucket
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

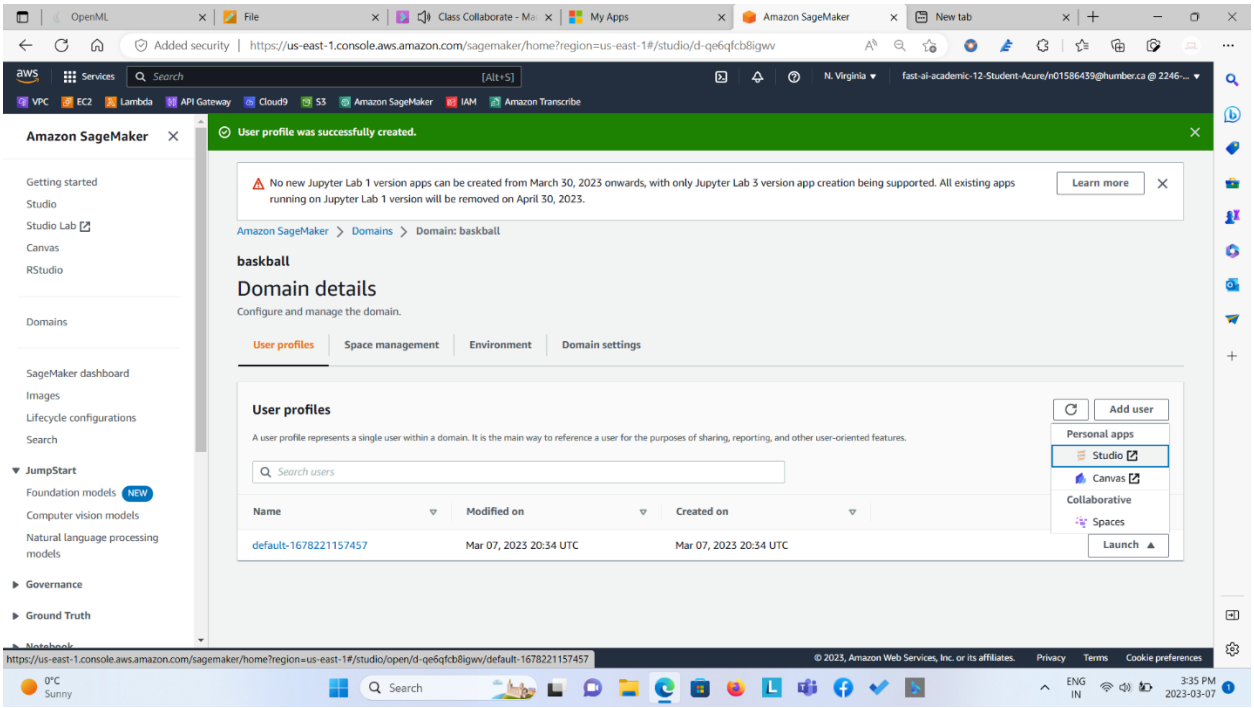

Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > baskballbucket
baskballbucket Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (0)
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
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
No objects
You don't have any objects in this bucket.
Upload
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

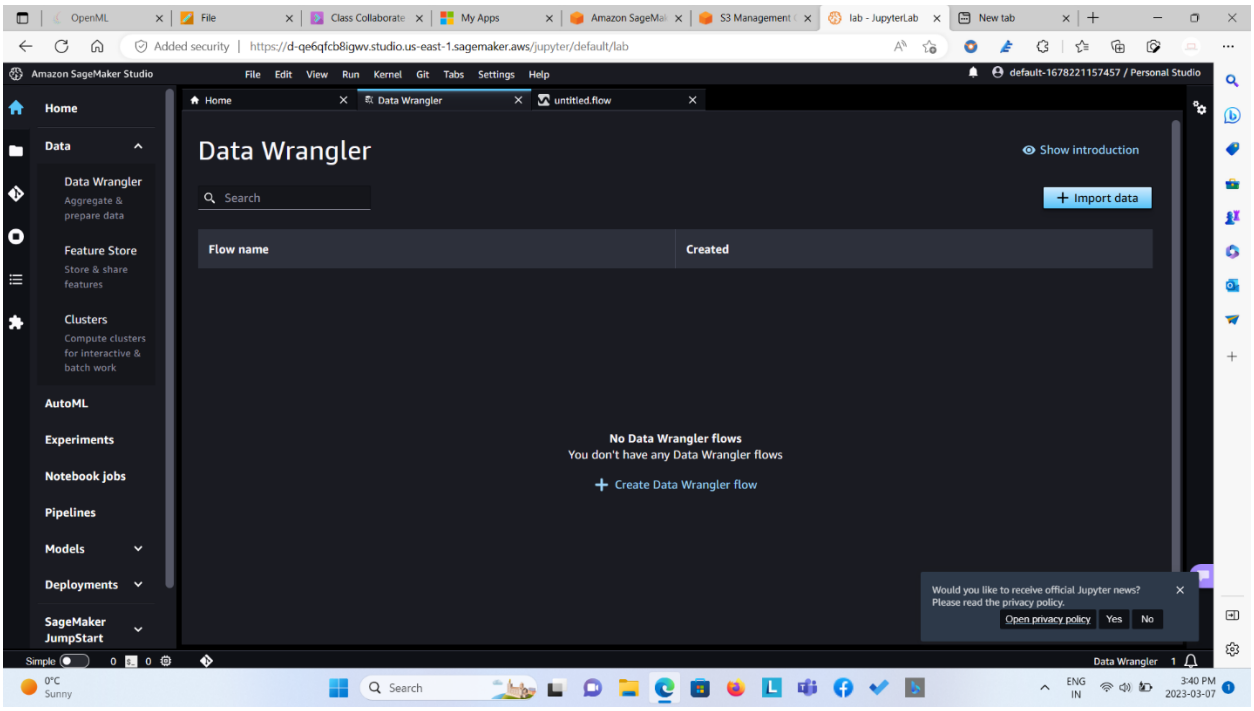

Upload
Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more
Drag and drop files and folders you want to upload here, or choose Add files, or Add folders.
Files and folders (0)
All files and folders in this table will be uploaded.
Remove
Add files
Add folder
Find by name
Name
Folder
Type
Size
No files or folders
You have not chosen any files or folders to upload.
Destination
Destination
s3://baskballbucket
Destination details
Bucket settings that impact new objects stored in the specified destination.
Permissions
Open
Desktop
Search Desktop
Organize
New folder
Name
Date modified
Type
Size
Cover Letter
2023-03-02 7:24 PM
File folder
dataset_2200_baskball
2023-03-07 2:16 PM
ARFF Data File
dataset_2200_baskball
2023-03-07 2:26 PM
Microsoft Excel Co...
Excel
2022-12-03 8:21 AM
Shortcut
Google Chrome
2023-02-27 4:44 PM
Shortcut
OneNote
2022-12-03 8:21 AM
Shortcut
Packing List
2022-12-18 5:44 AM
Microsoft Word D...
Desktop
Downloads
Documents
Pictures
Music
Videos
Job Documents
File name: dataset_2200_baskball
All files
Open
Cancel
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

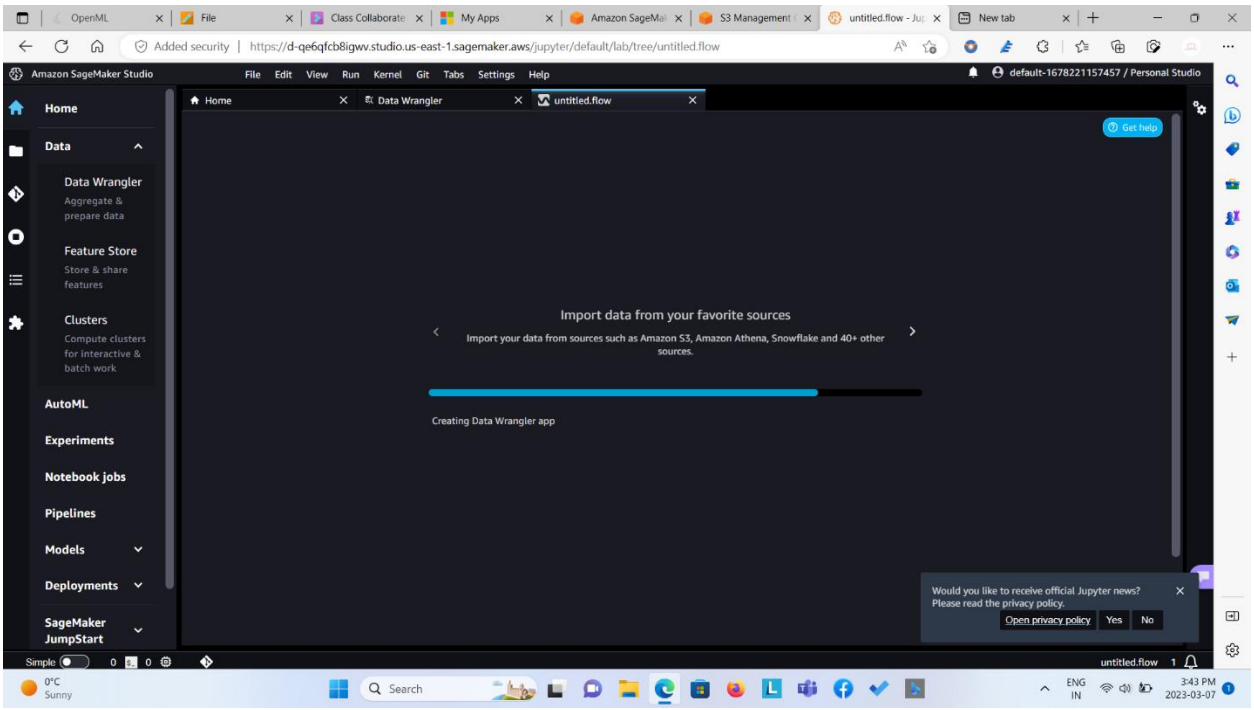

Upload
Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. Learn more
Drag and drop files and folders you want to upload here, or choose Add files, or Add folders.
Files and folders (1 Total, 3.2 KB)
All files and folders in this table will be uploaded.
Remove
Add files
Add folder
Find by name
Name
Folder
Type
Size
dataset_2200_baskball.csv
-
text/csv
3.2 KB
Destination
Destination
s3://baskballbucket
Destination details
Bucket settings that impact new objects stored in the specified destination.
Permissions
Grant public access and access to other AWS accounts.
Properties
Specify storage class, encryption settings, tags, and more.
Cancel
Upload
Feedback
Language
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

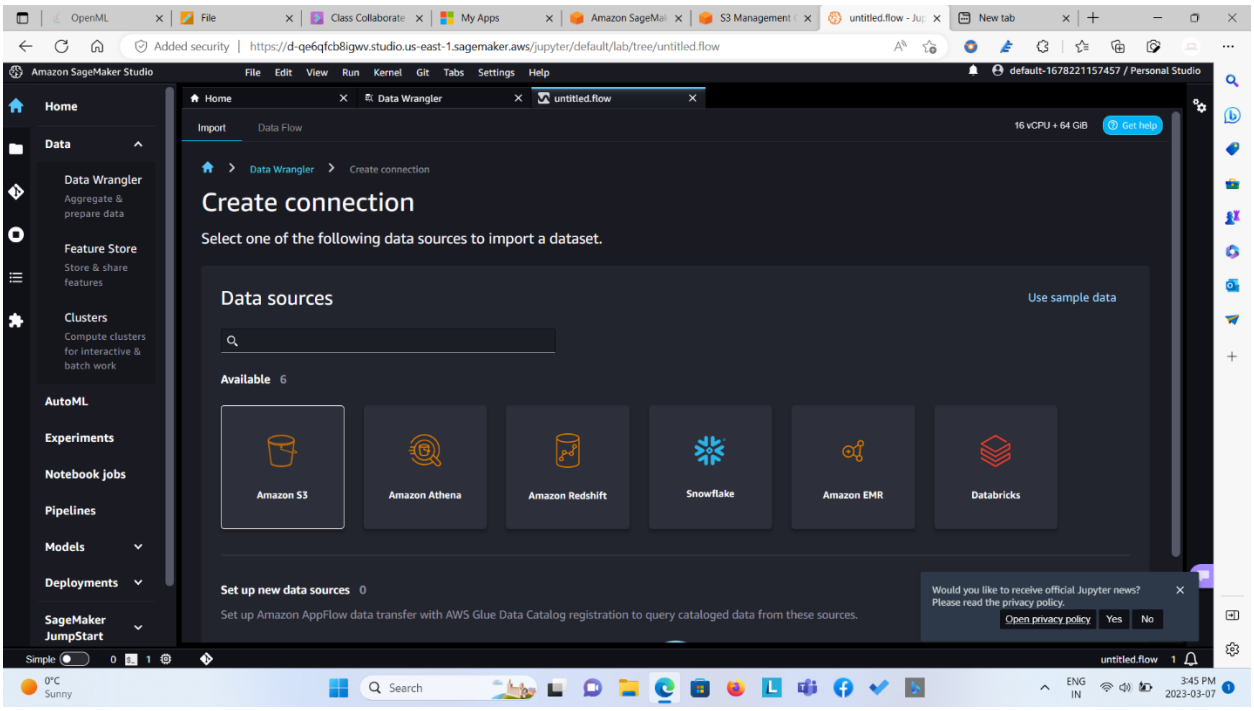


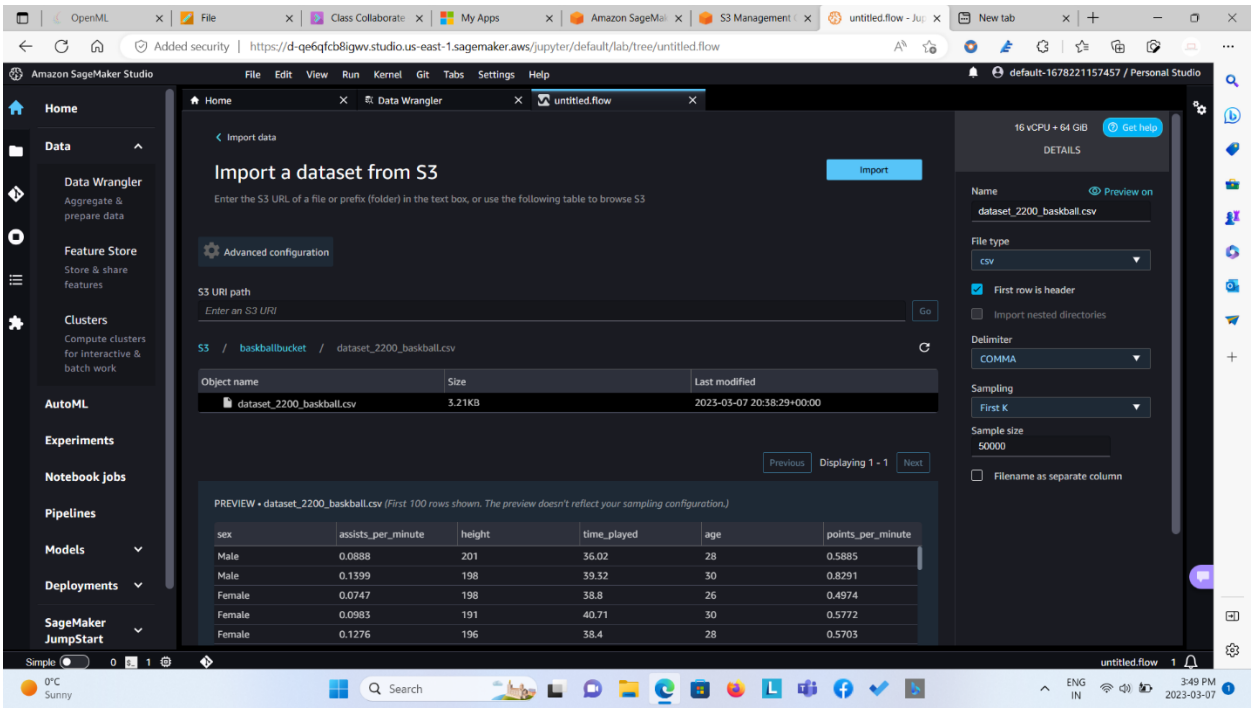
Step 5: Performing Data Cleaning/Data Pre-processing/Data Transformation/Data Analysis via. Data Wrangler from Studio by linking Dataset from S3 Bucket.

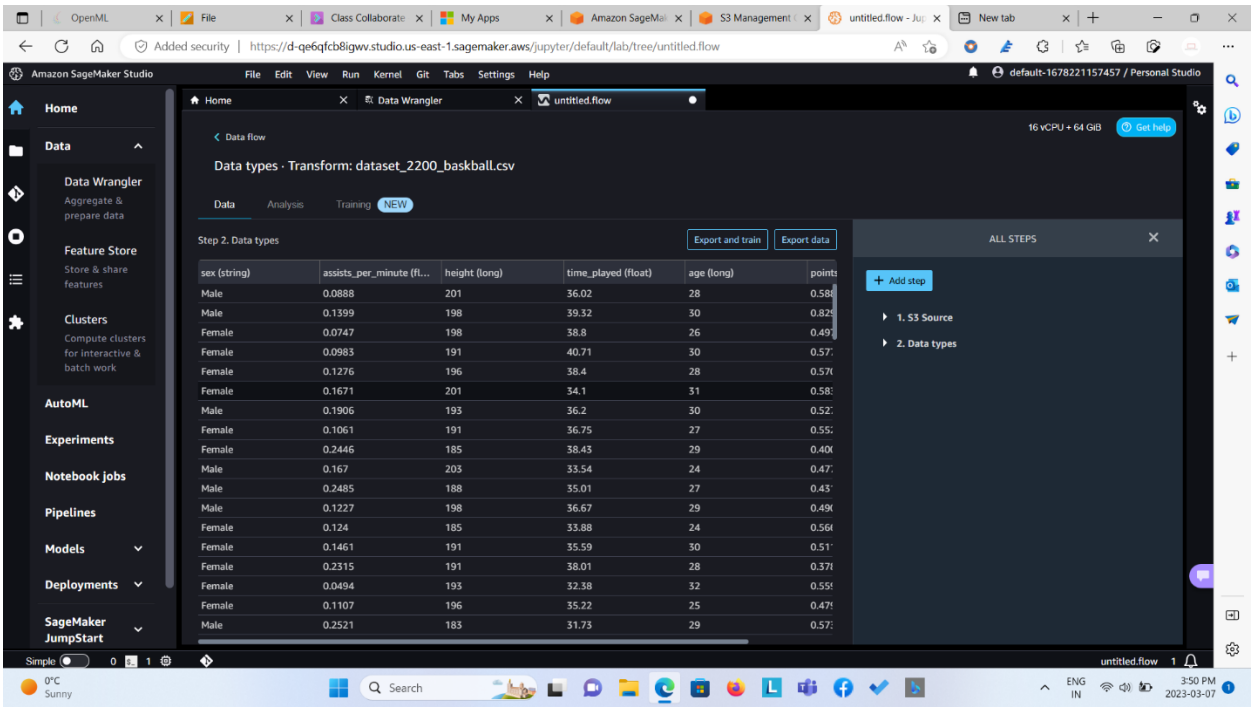


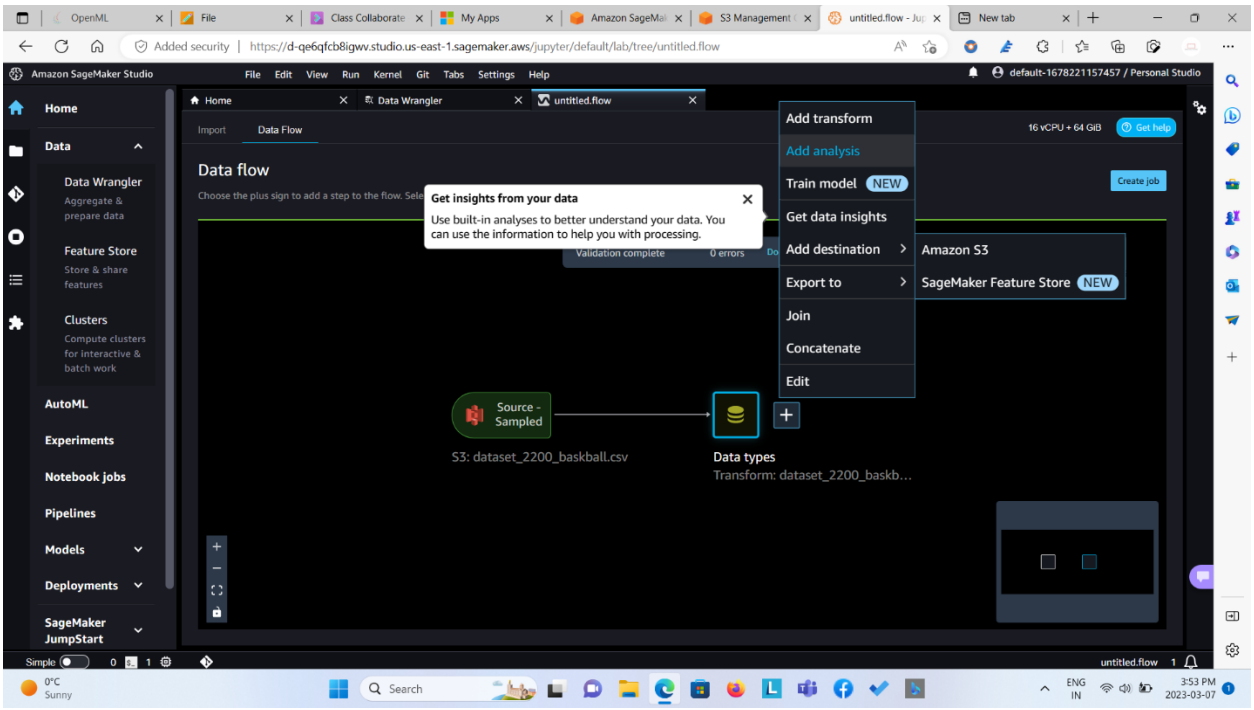

Amazon SageMaker Studio
File Edit View Run Kernel Git Tabs Settings Help
default-1678221157457 / Personal Studio
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Home
Data Wrangler
untitled.flow
Data Wrangler
Show introduction
Search
Import data
Flow name
Created
No Data Wrangler flows
You don't have any Data Wrangler flows
Create Data Wrangler flow
Would you like to receive official Jupyter news?
Please read the privacy policy.
Open privacy policy
Yes
No
Simple
Data Wrangler 1

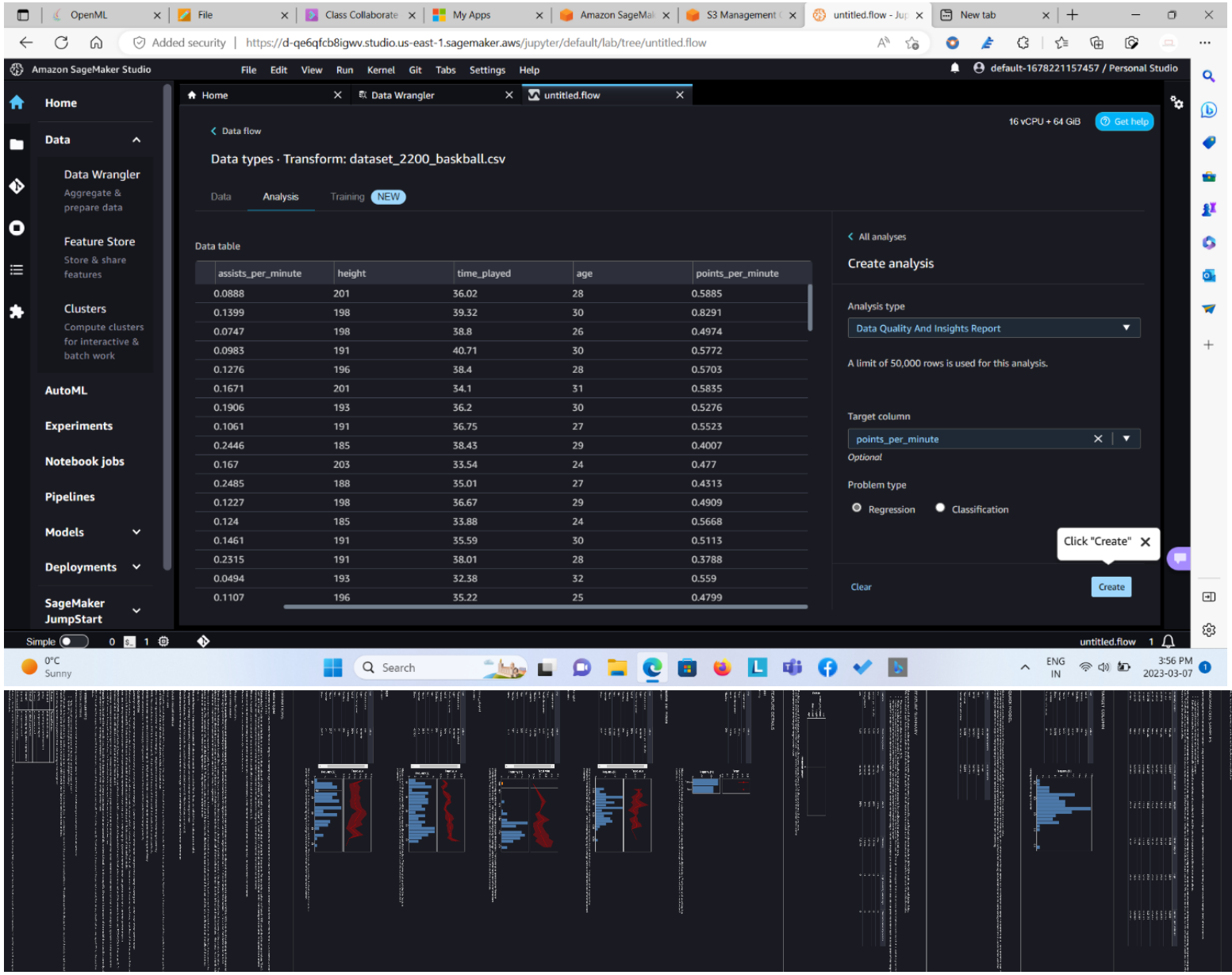

Amazon SageMaker Studio
File Edit View Run Kernel Git Tabs Settings Help
default-1678221157457 / Personal Studio
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Home
Data Wrangler
untitled.flow
Get Help
Import data from your favorite sources
Import your data from sources such as Amazon S3, Amazon Athena, Snowflake and 40+ other sources.
Creating Data Wrangler app
Would you like to receive official Jupyter news?
Please read the privacy policy.
Open privacy policy
Yes
No
Simple
untitled.flow 1

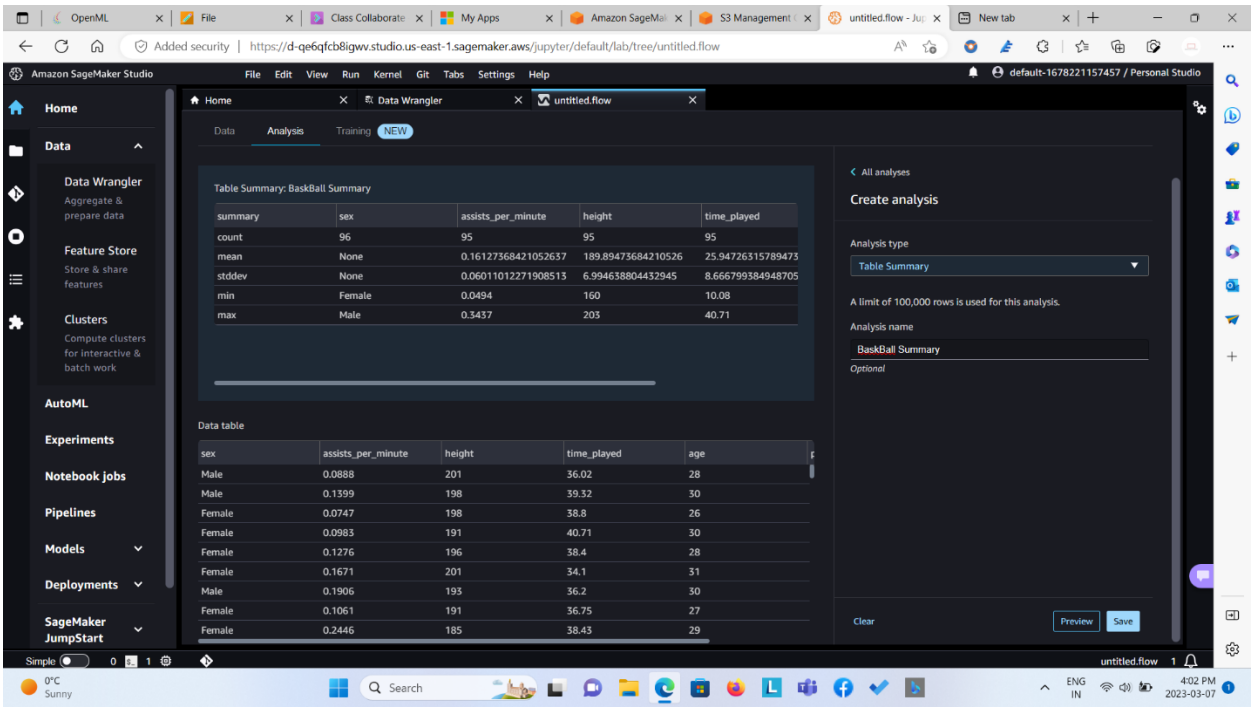

Amazon SageMaker Studio
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Import
Data Flow
16 vCPU + 64 GiB
Get help
Data Wrangler
Create connection
Create connection
Select one of the following data sources to import a dataset.
Data sources
Use sample data
Available 6
Amazon S3
Amazon Athena
Amazon Redshift
Snowflake
Amazon EMR
Databricks
Set up new data sources 0
Set up Amazon AppFlow data transfer with AWS Glue Data Catalog registration to query cataloged data from these sources.
Would you like to receive official Jupyter news?
Please read the privacy policy.
Open privacy policy
Yes
No


Amazon SageMaker Studio
Import data
Import a dataset from S3
Import
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
Enter an S3 URI
Go
S3 / baskballbucket / dataset_2200_baskball.csv
Object name
Size
Last modified
dataset_2200_baskball.csv
3.21KB
2023-03-07 20:38:29+00:00
Previous
Displaying 1 - 1
Next
PREVIEW • dataset_2200_baskball.csv (First 100 rows shown. The preview doesn't reflect your sampling configuration.)
sex
assists_per_minute
height
time_played
age
points_per_minute
Male 0.0888 201 36.02 28 0.5885
Male 0.1399 198 39.32 30 0.8291
Female 0.0747 198 38.8 26 0.4974
Female 0.0983 191 40.71 30 0.5772
Female 0.1276 196 38.4 28 0.5703
DETAILS
Name
Preview on
dataset_2200_baskball.csv
File type
csv
First row is header
Import nested directories
Delimiter
COMMA
Sampling
First K
Sample size
50000
Filename as separate column


Amazon SageMaker 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
Data flow
Data types · Transform: dataset_2200_baskball.csv
Data
Analysis
Training
NEW
Step 2. Data types
Export and train
Export data
ALL STEPS
Add step
1. S3 Source
2. Data types
16 vCPU + 64 GiB
Get help
sex (string)
assists_per_minute (fl...
height (long)
time_played (float)
age (long)
points
Male 0.0888 201 36.02 28
Male 0.1399 198 39.32 30
Female 0.0747 198 38.8 26
Female 0.0983 191 40.71 30
Female 0.1276 196 38.4 28
Female 0.1671 201 34.1 31
Male 0.1906 193 36.2 30
Female 0.1061 191 36.75 27
Female 0.2446 185 38.43 29
Male 0.167 203 33.54 24
Male 0.2485 188 35.01 27
Male 0.1227 198 36.67 29
Female 0.124 185 33.88 24
Female 0.1461 191 35.59 30
Female 0.2315 191 38.01 28
Female 0.0494 193 32.38 32
Female 0.1107 196 35.22 25
Male 0.2521 183 31.73 29
untitled.flow


Amazon SageMaker 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
Import
Data Flow
Data flow
Choose the plus sign to add a step to the flow.
Get insights from your data
Use built-in analyses to better understand your data. You can use the information to help you with processing.
Validation complete
0 errors
Add transform
Add analysis
Train model
NEW
Get data insights
Add destination
Amazon S3
Export to
SageMaker Feature Store
NEW
Join
Concatenate
Edit
Create job
16 vCPU + 64 GiB
Get help
Source - Sampled
S3: dataset_2200_baskball.csv
Data types
Transform: dataset_2200_baskb...
untitled.flow

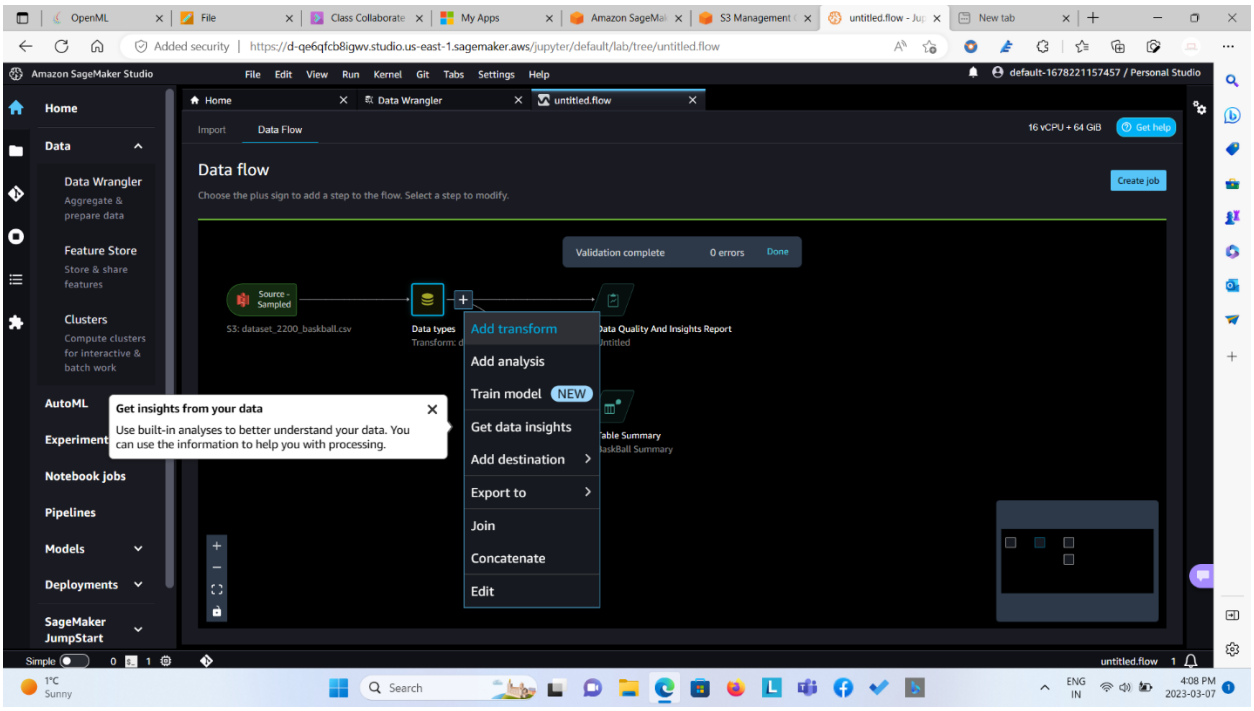

Amazon SageMaker Studio
File Edit View Run Kernel Git Tabs Settings Help
default-1678221157457 / Personal Studio
Home
Data
Data Wrangler
Aggregate & prepare data
Feature Store
Store & share features
Clusters
Compute clusters for interactive & batch work
AutoML
Experiments
Notebook jobs
Pipelines
Models
Deployments
SageMaker JumpStart
Home
Data Wrangler
untitled.flow
Data flow
16 vCPU + 64 GiB
Get help
Data types · Transform: dataset_2200_baskball.csv
Data
Analysis
Training
NEW
Data table
assists_per_minute height time_played age points_per_minute
0.0888 201 36.02 28 0.5885
0.1399 198 39.32 30 0.8291
0.0747 198 38.8 26 0.4974
0.0983 191 40.71 30 0.5772
0.1276 196 38.4 28 0.5703
0.1671 201 34.1 31 0.5835
0.1906 193 36.2 30 0.5276
0.1061 191 36.75 27 0.5523
0.2446 185 38.43 29 0.4007
0.167 203 33.54 24 0.477
0.2485 188 35.01 27 0.4313
0.1227 198 36.67 29 0.4909
0.124 185 33.88 24 0.5668
0.1461 191 35.59 30 0.5113
0.2315 191 38.01 28 0.3788
0.0494 193 32.38 32 0.559
0.1107 196 35.22 25 0.4799
All analyses
Create analysis
Analysis type
Data Quality And Insights Report
A limit of 50,000 rows is used for this analysis.
Target column
points_per_minute
Optional
Problem type
Regression
Classification
Click "Create"
Clear
Create
Simple
untitled.flow
0°C
Sunny
Search
ENG
IN
3:56 PM
2023-03-07

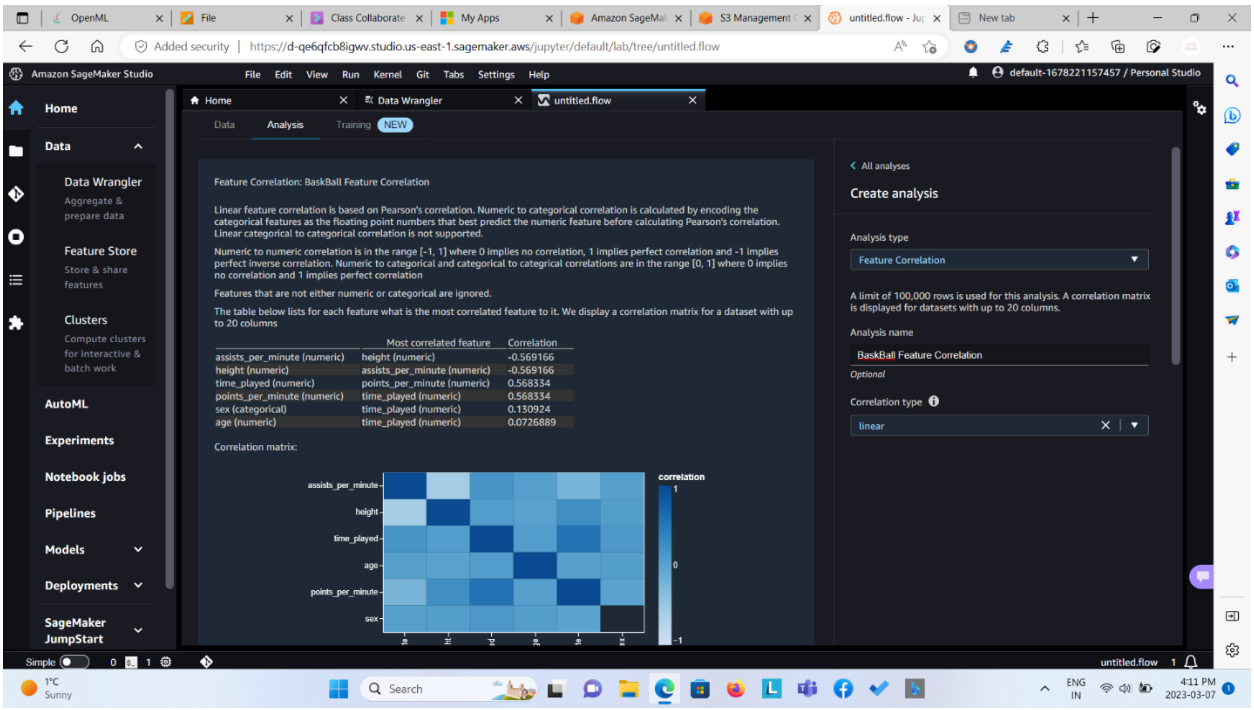


From the data analysis report I came to know about various things –

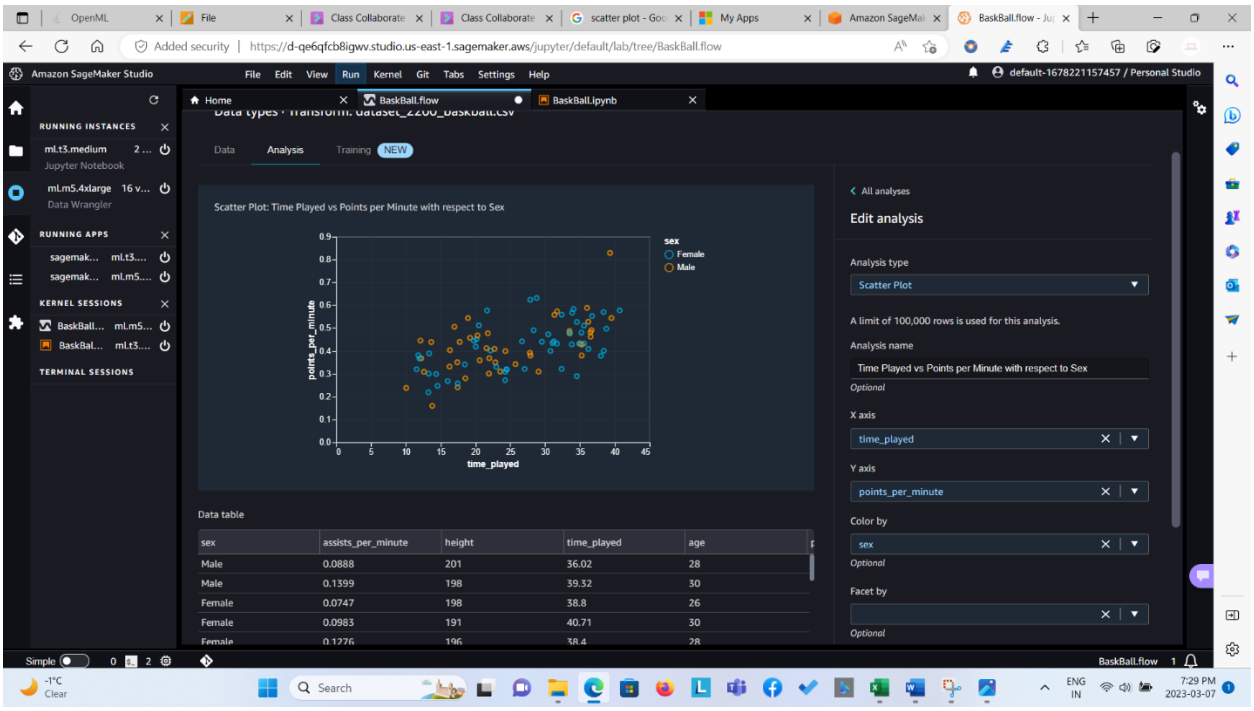
1. **The overview of the dataset in term for number of instances, number of features, missing data percentage, duplicate rows, data types in short all about the statistic of the dataset.**
2. **The descriptive summary of all the features in terms of min, max, mean, median, standard deviation, mode.**
3. **Summary and Distribution of the target variable as well as other features..**
4. **Quick overview about the regression model in terms of Validation scores and Test scores having R2, MSE, RMSE, MAE, Max Error, and Median Absolutely error as its metrics.**
5. **Machine Learning Terminologies**

This was report was necessary to generate as from it we could know about the dataset which we will be working upon as well as we would get clear idea about the features and domain we are working for.

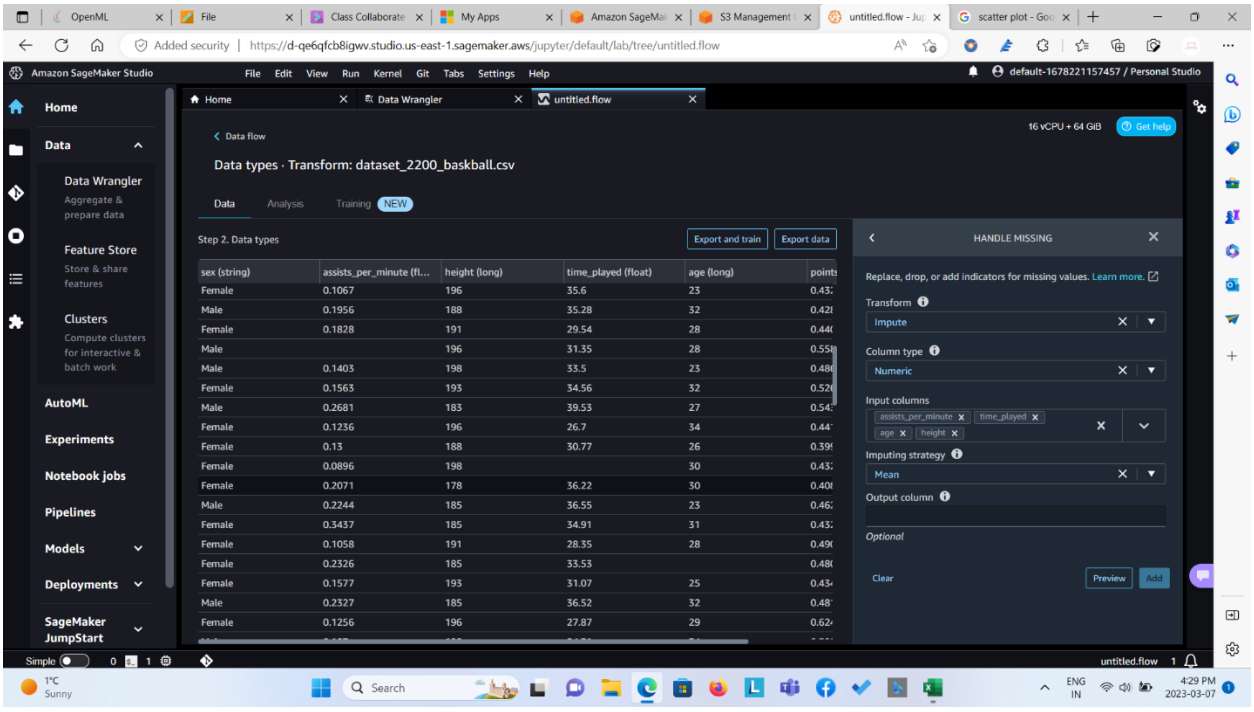


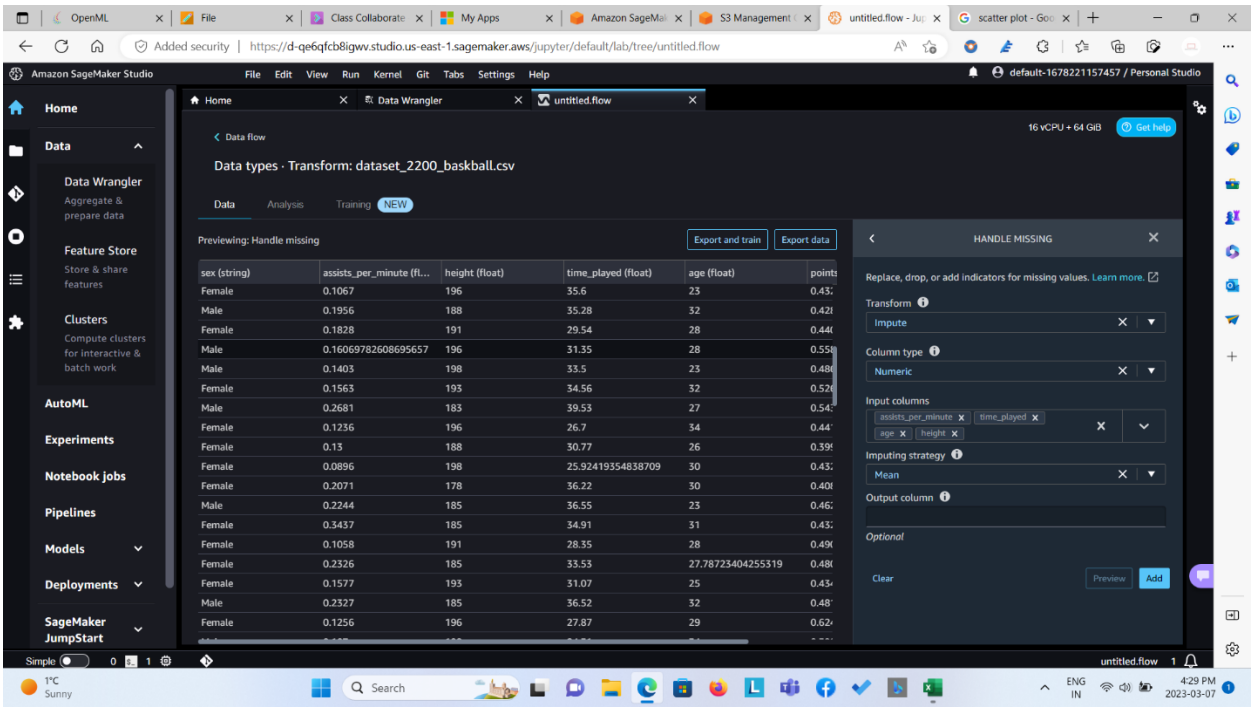


From the above heatmap, it is inferred that there is partial correlation between features which says that the features have partial significance over the target variable in order to predict the “points scored in a minute”.

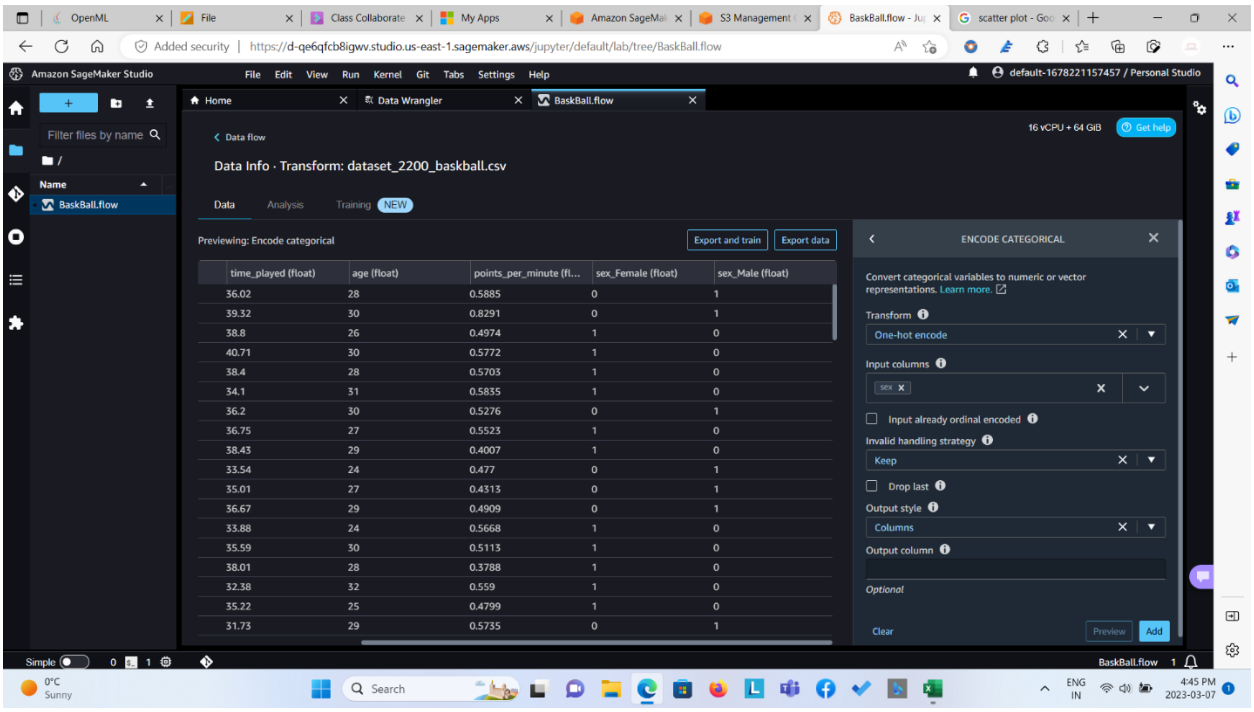


From the above scatter diagram, it is said the there is a relationship between time played and points per minute. It shows positive correlation and it plays vital role in predicting the points scored in a minute.





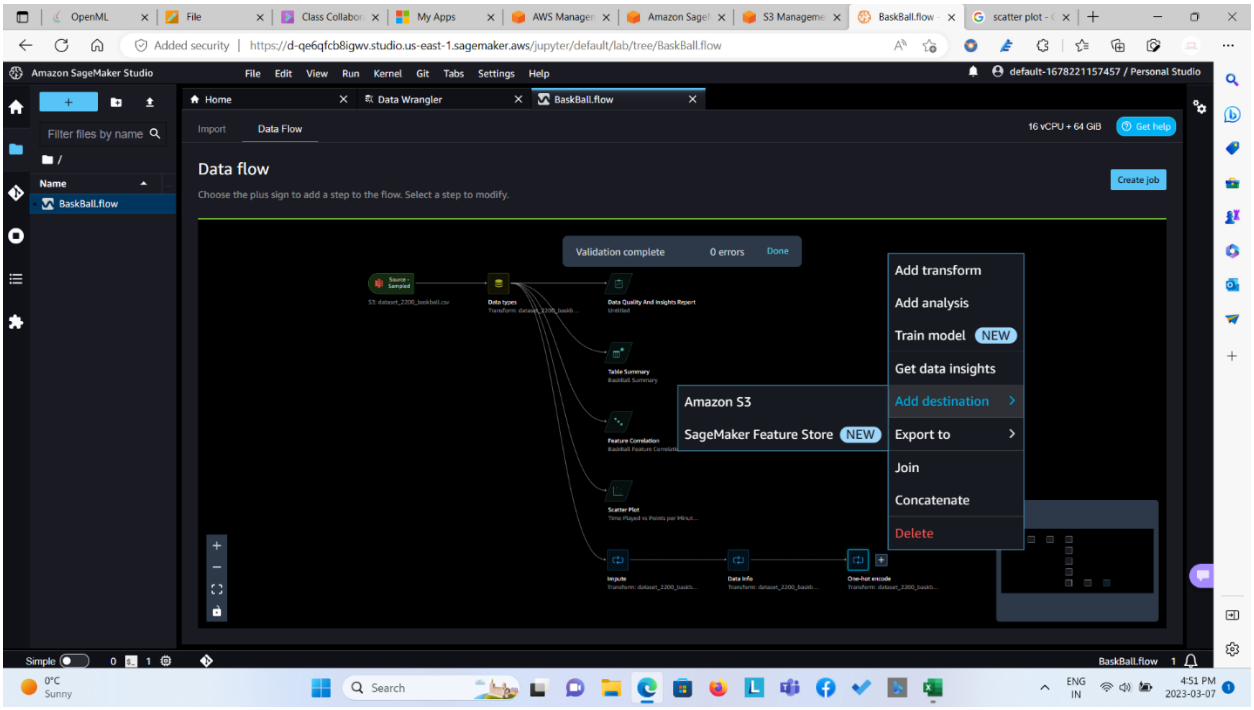
From the above data pre-processing step, we have replaced the missing values with the mean of the particular column. We did this so that we have a complete structured dataset in order to get the desire results in terms of accuracy and precision.



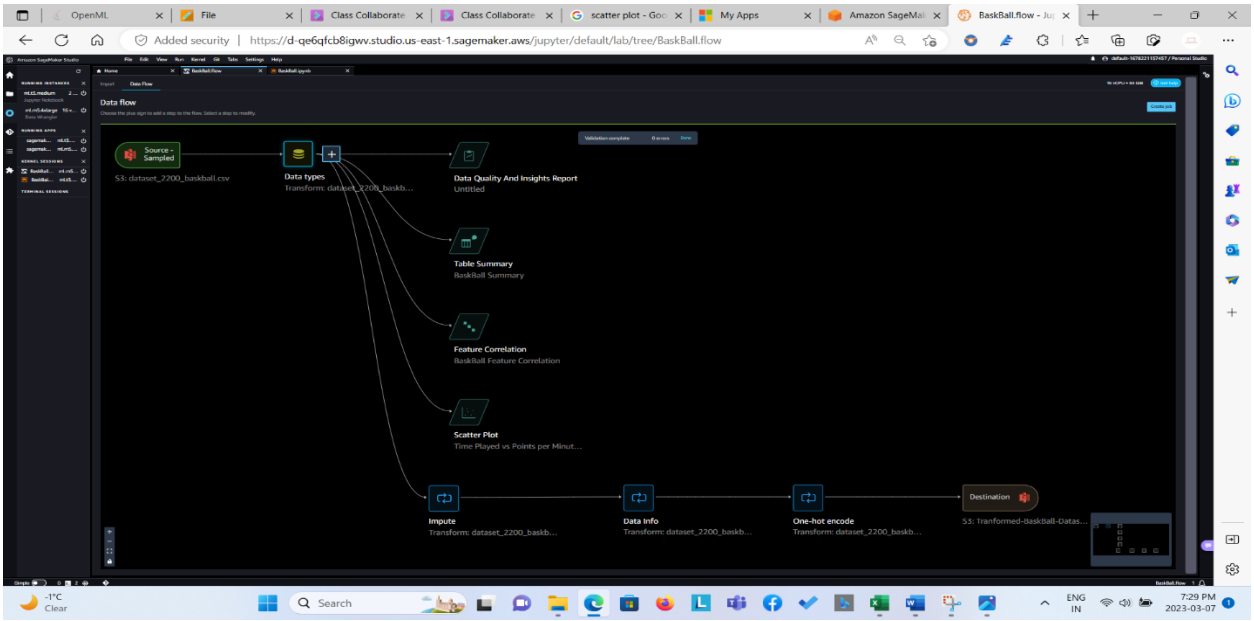
The step for one-hot-encoding is necessary for feature named “sex”, as we know that regression model can only be applied upon the features that are quantitative in nature. Here

sex is categorical variable, which also plays vital role in the prediction so to satisfy the regression condition we are converting it into numerical by encoding it as 0 and 1.

Step 6: Adding the Destination so that we can let the data flow to execute and the result to save into S3 bucket.



The Final Data-Flow



Later, created the job and the output was exported to the s3 bucket from data wrangler/data flow.