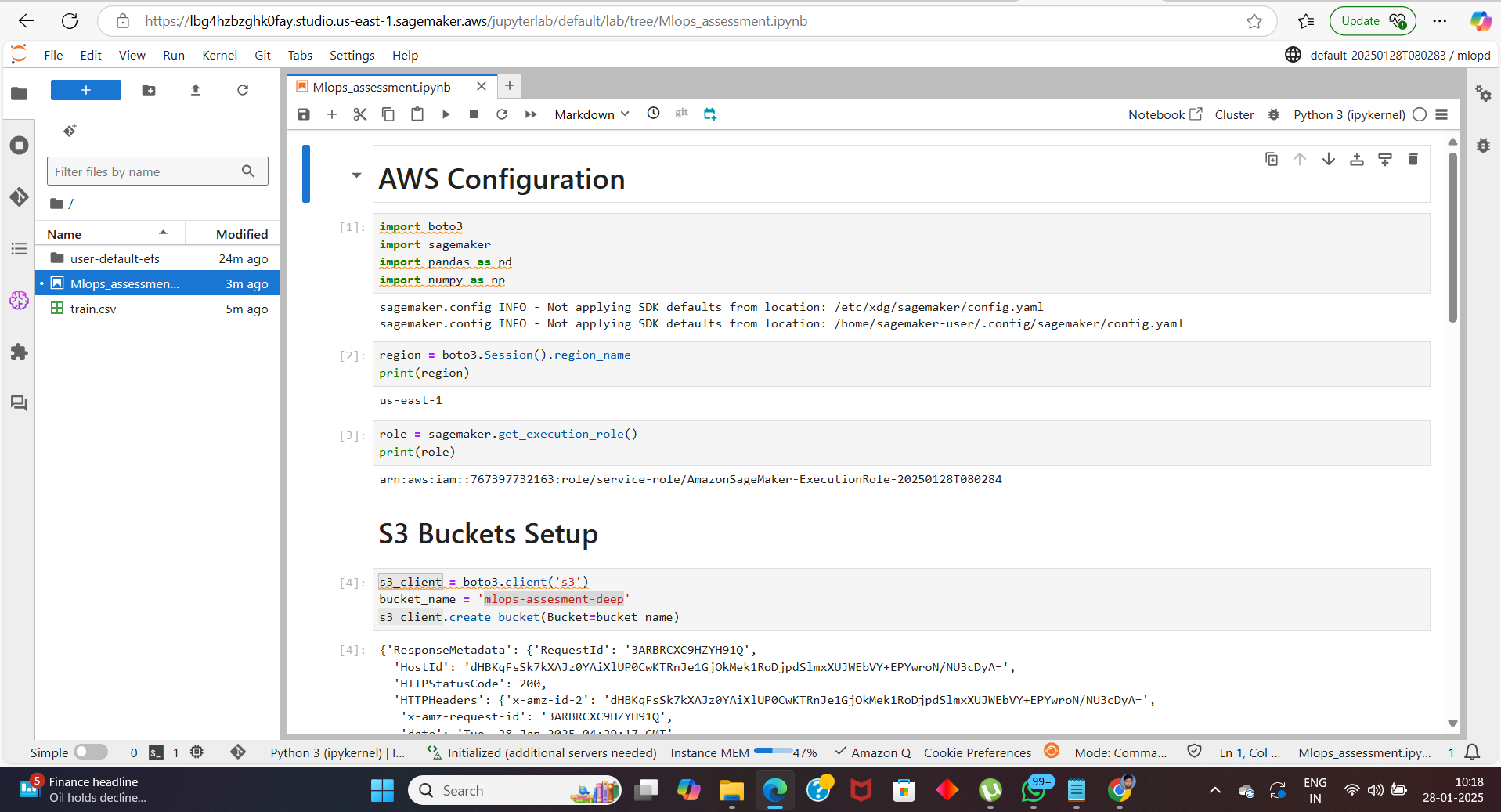
Technical Documentation

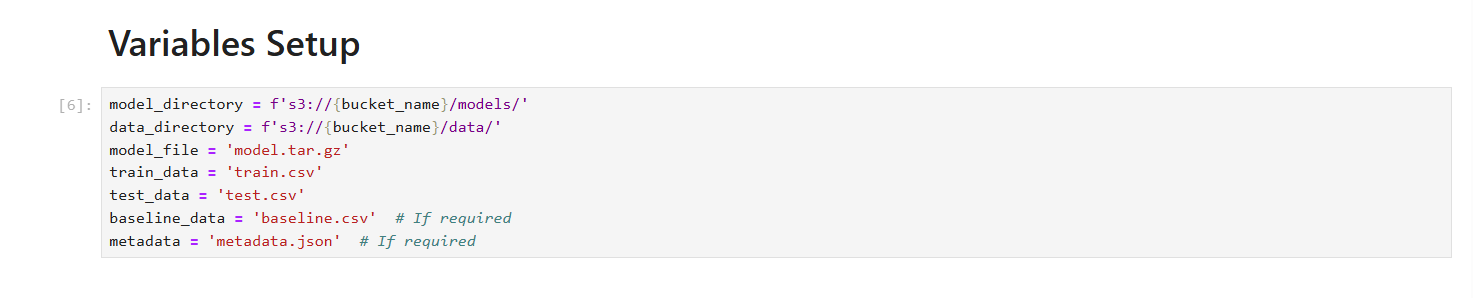


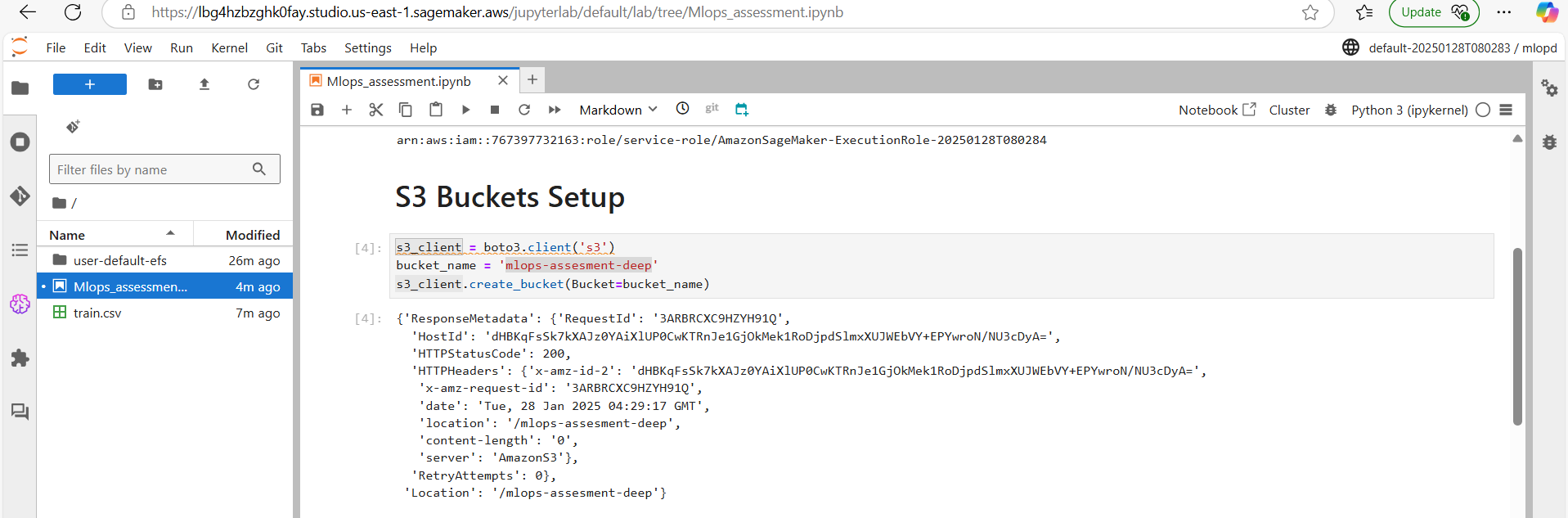
Proposed by Deep Prasad

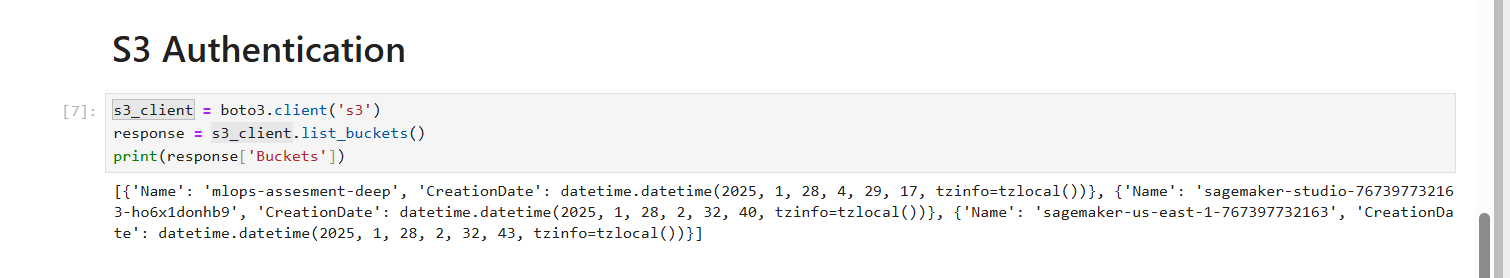
2121507

[Deep.Prasad@congnizant.com](mailto:Deep.Prasad@congnizant.com)









Challenges Faced: None

**Data Eyeballing**

Wrote several and performed detailed EDA to find the following observations:  
  
**observations**

* recordID can be ignored beacuse its basically a no. that identitifies each row uniquely
* is\_feartured\_sku can be ignored beacuse its a category and for training it is necessary it remains a number
* is\_display\_sku can be ignored beacuse its a category and for training it is necessary it remains a number

**total\_price**

* mean > median a slight right skewness will be there in the data and the standard deviation is high so data is not well distributed and mean > standard deviation the normal curve will be a bit flattened and range of the data is 41 to 562
* outliers are present

**base\_price**

* mean > median a slight right skewness will be there in the data and the standard deviation is high so data is not well distributed and mean > standard deviation the normal curve will be a bit flattened and range of the data is 61 to 562.
* comparing the ranges of the total\_price and base price it is clear that there total\_price < base\_price which means some items are being sold below base price hence reason can be investigated for this loss in total\_price
* outliers are present

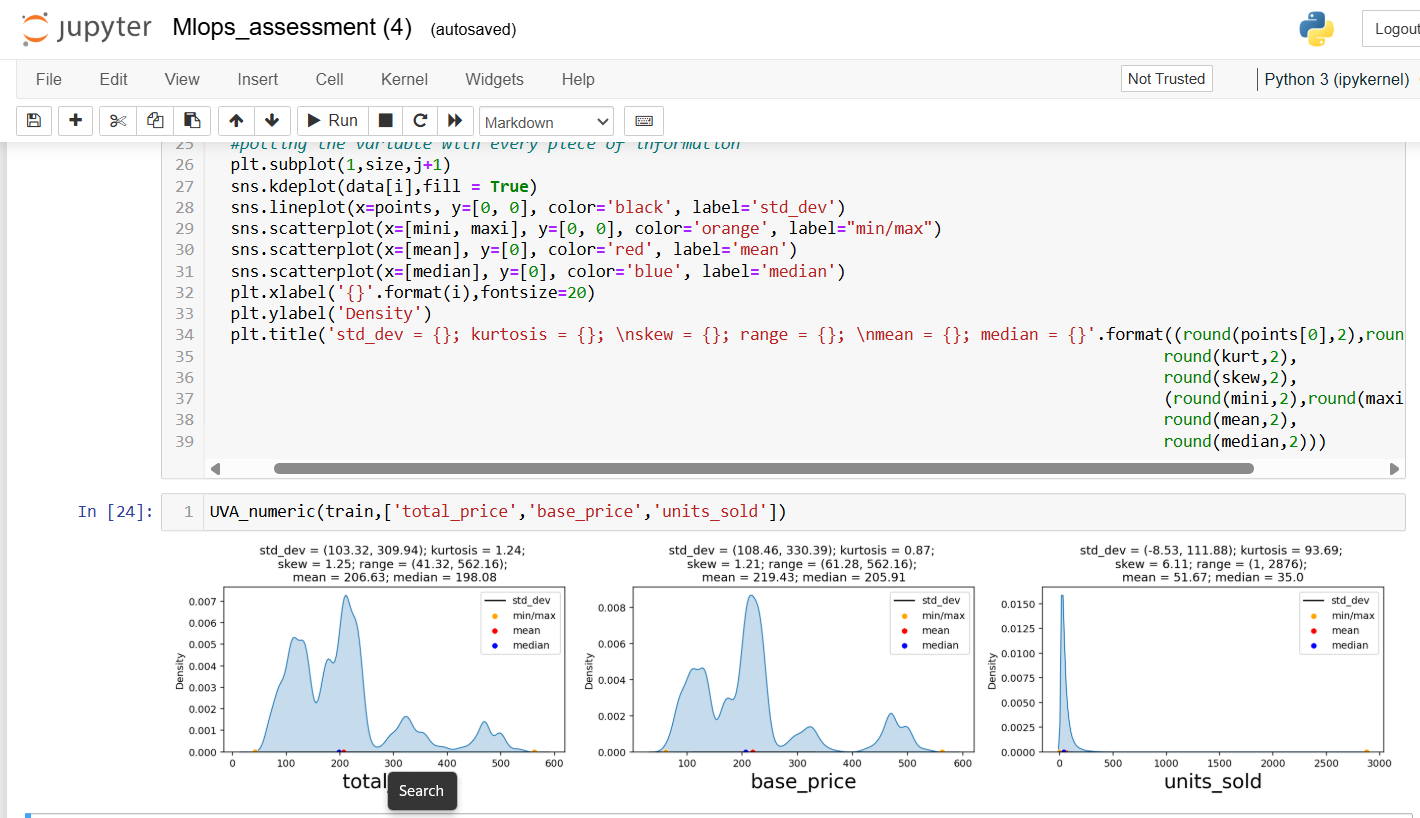
**units\_sold**

* mean > median a right skewness will be there in the data and the standard deviation is high so data is not well distributed and mean < standard deviation the normal curve will be a sharp peak and range of the data is 1 to 2876
* outliers are present

Challenges Faced: Outliers

Handing techinique: removed using boxplot and whiskers i.e any outlier 1.5 times the upper lower whiskers are removed

Decision: it wont affect our model we have had enough data dropping miniscule amount of data affect much



**observations**

**total price**

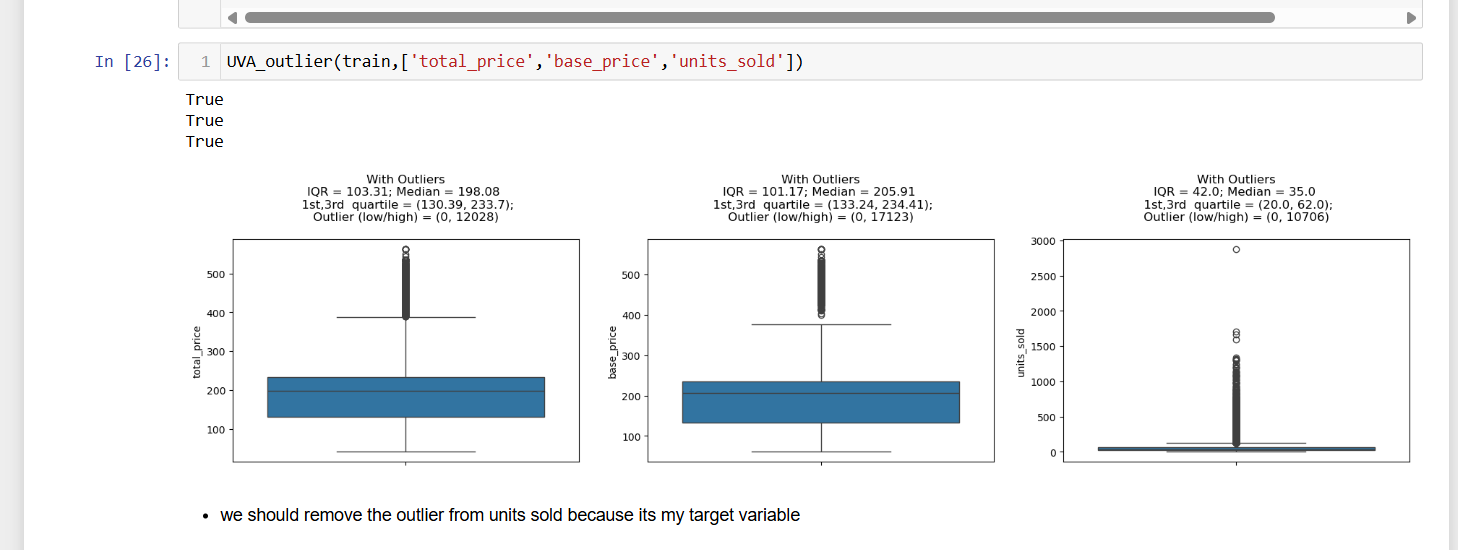
* clearly it has outliers but these outliers are important can be point of interest cane help further analysis
* skew >0.5: positively/right skewed
* kurtosis > 0: High extreme values are there and fatter tail we can visually verified

**base price**

* clearly it has outliers but these outliers are important can be point of interest cane help further analysis
* skew >0.5: positively/right skewed
* kurtosis > 0: High extreme values are there and fatter tail we can visually verified

**units sold**

* clearly outliers are there but visually not able to verify
* skew >>> 0.5 -> positively/right skewed
* kurtosis >>> 0 High extreme values are there and fatter tail we can visually verified

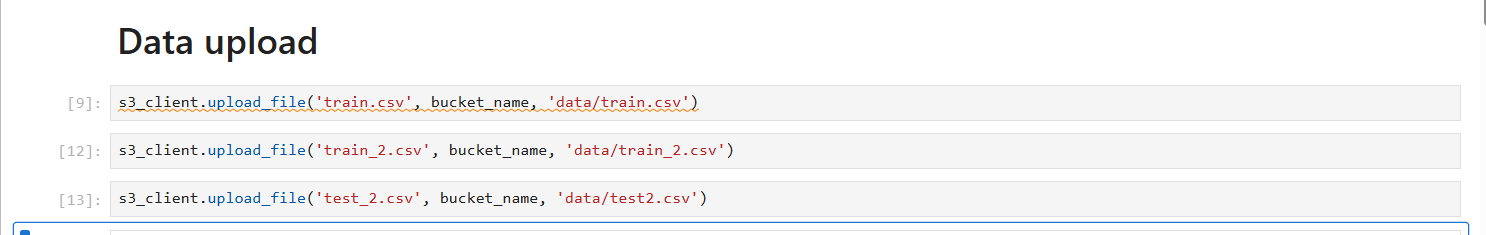


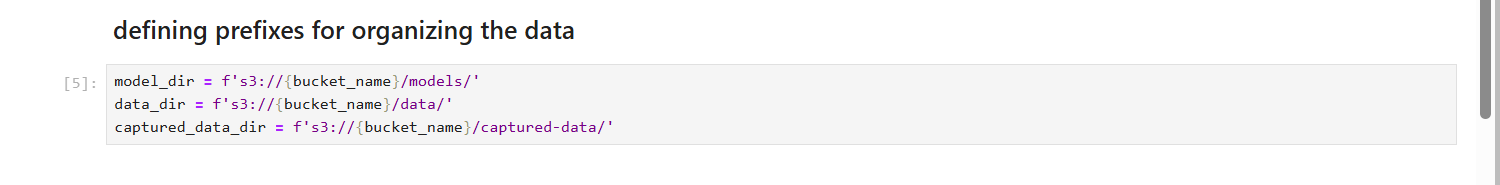
Challenge: Week is string here it needs to be done feature extraction because it will not allow us to train our model



Scaling and Encoding

* One-Hot Encoding turns each unique category (like store\_id and sku\_id) into a separate column with binary values (0 or 1).
* We dropped the first category of both columns (store\_id and sku\_id) to avoid multicollinearity. This means one category from each will not have a corresponding column, which is a standard approach to prevent redundancy.
* At the end it prevents model to give importance to those Features which have higher values.

UPLOADING THE FILES IN S3

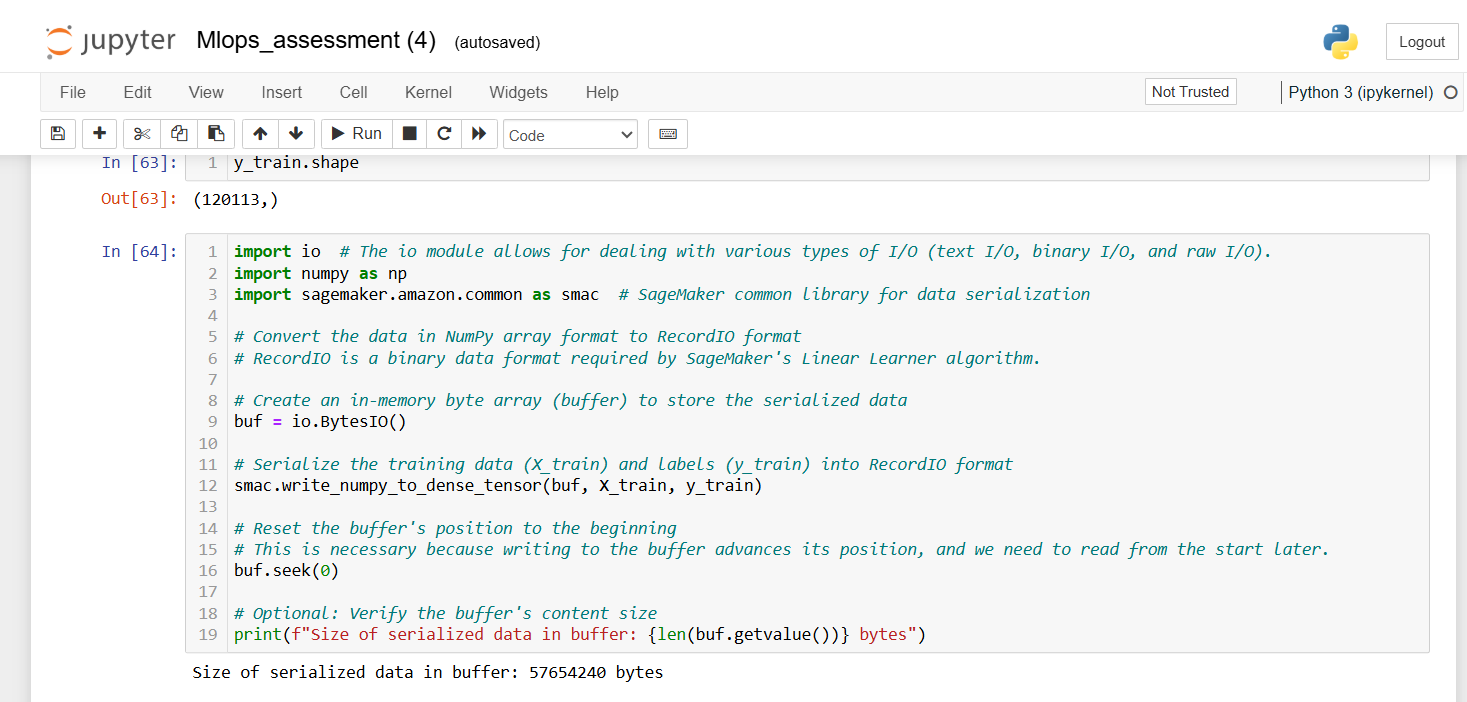


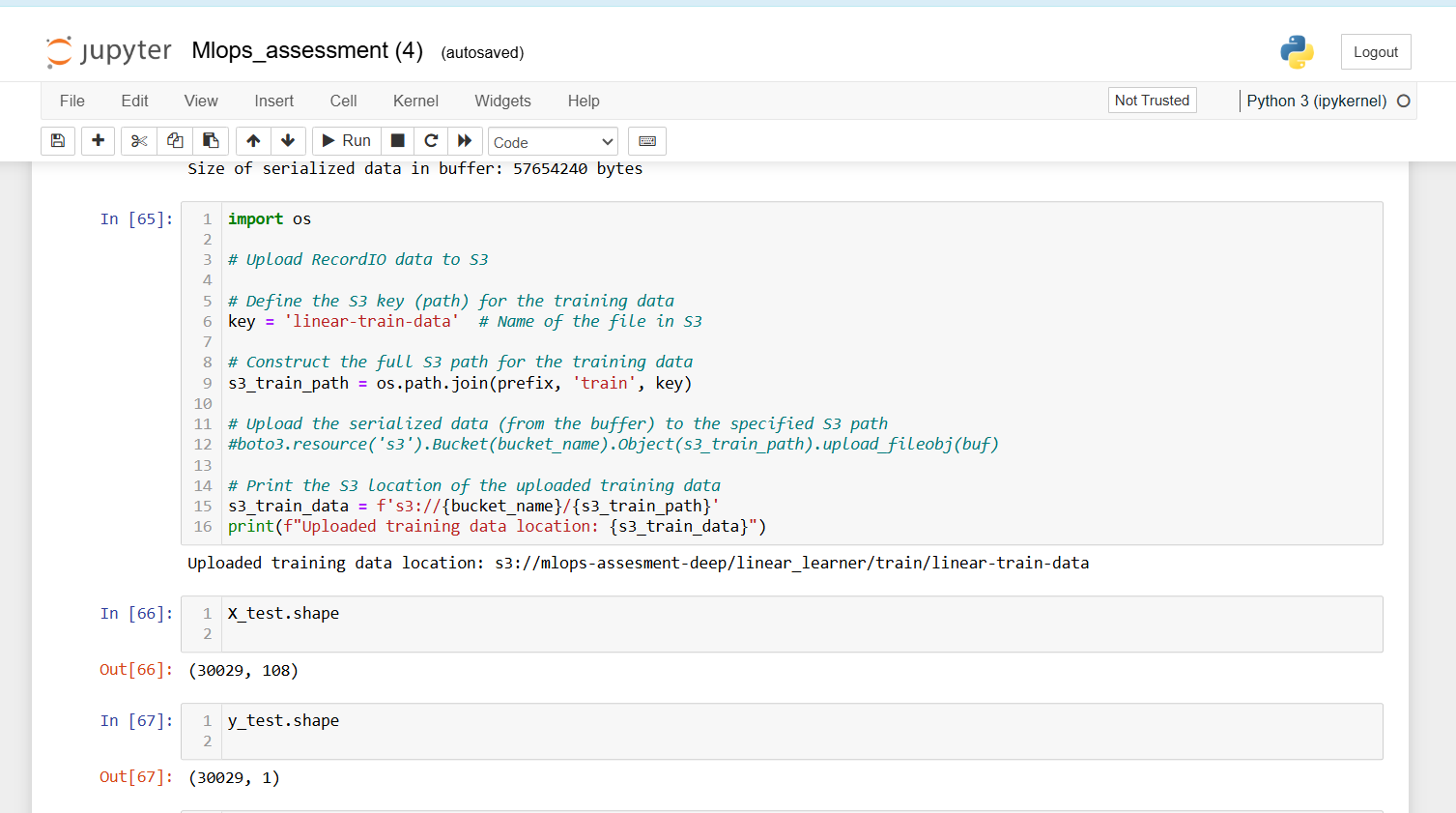


Challenges: Naming Conventions

Handling: decided to keep the names based clarity and when the reads it he would make an idea why the name was given

PRE-TRAINING THE MODEL

Taking care of Input and Output formats of the Sagemakers linear learner algorithm that expects in Record IO format using SMAC.  


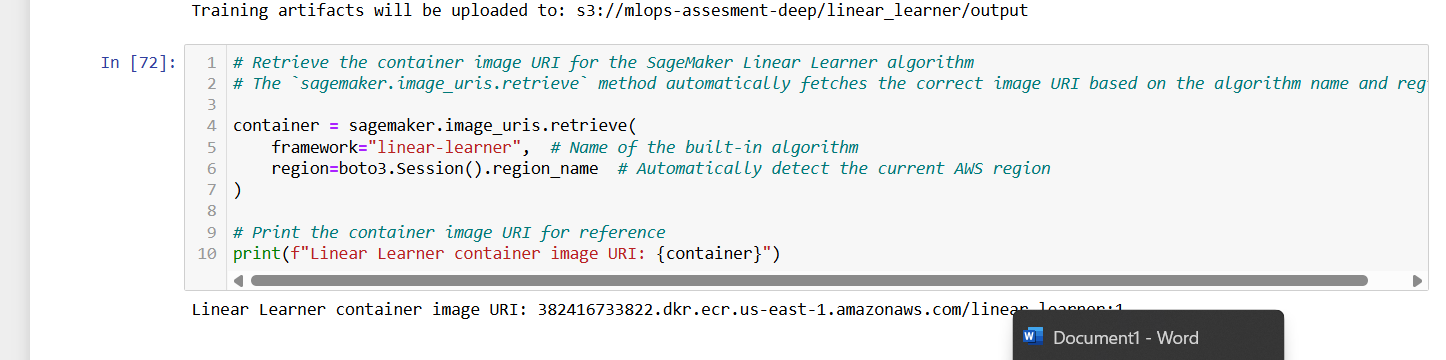


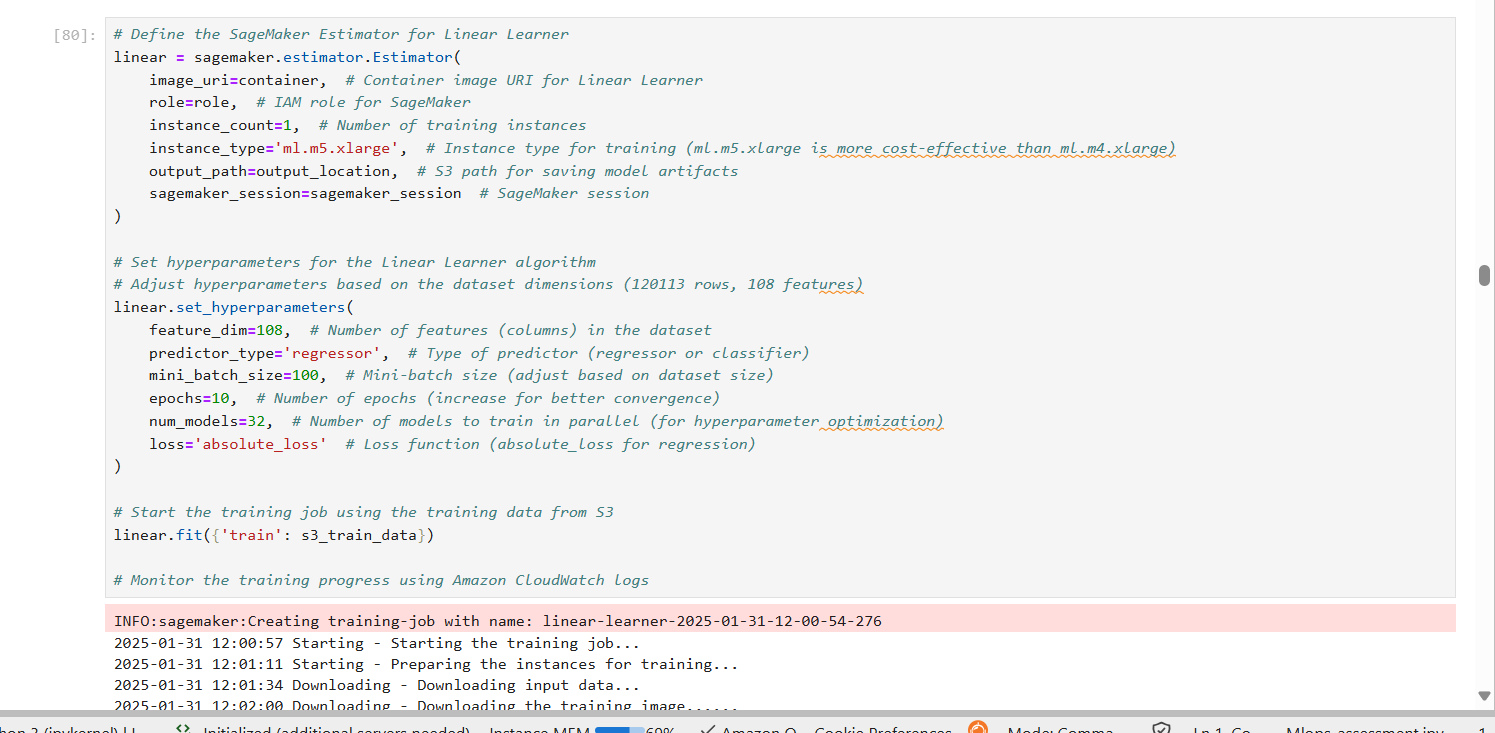
CREATING PLACEHOLDERS FOR MODEL OUTPUT



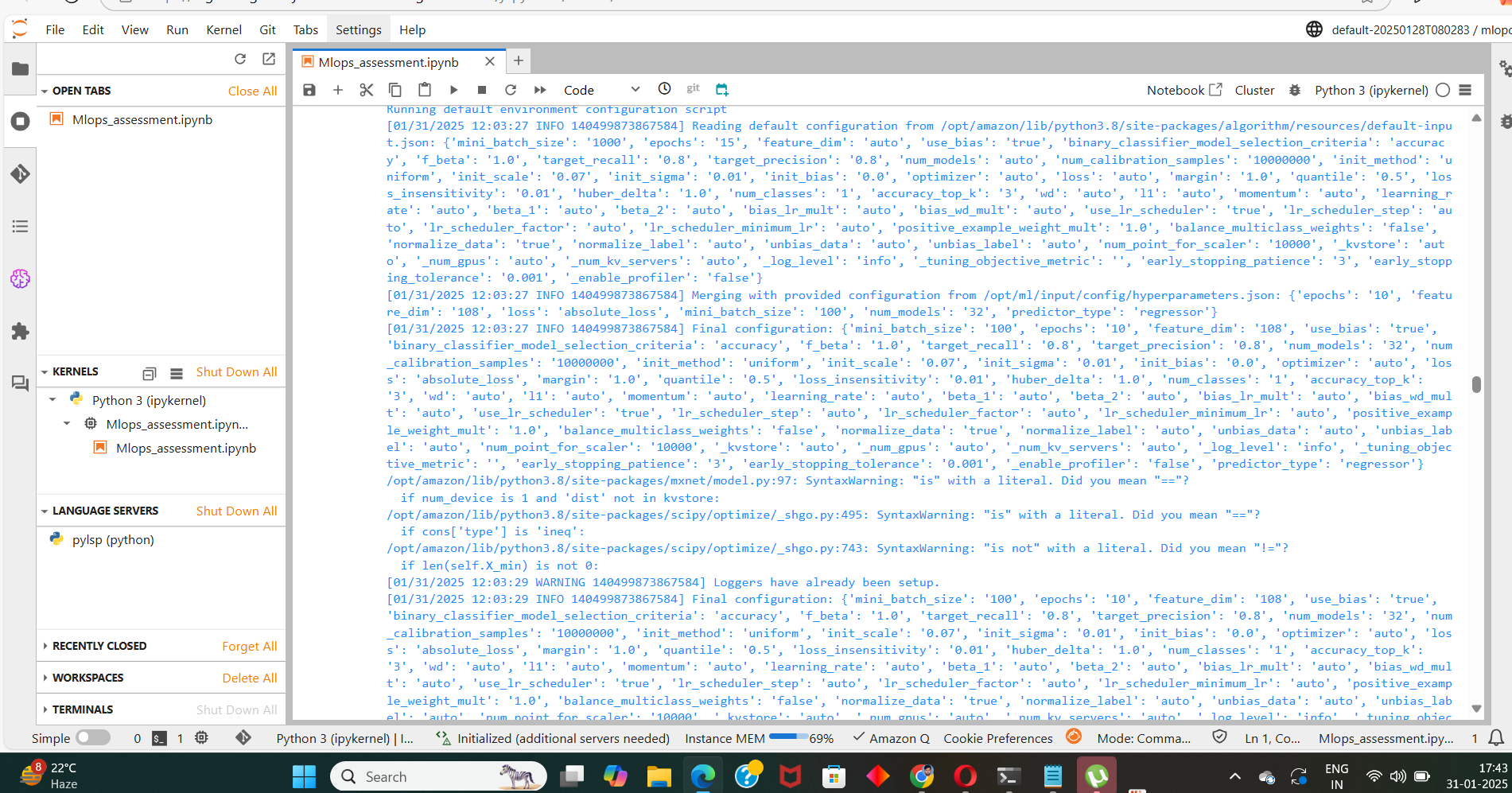
RETRIEVING THE MODEL IMAGE

Because no need to whole training process again and again we can use standard linear learner algorithm from aws for our TRAINING.

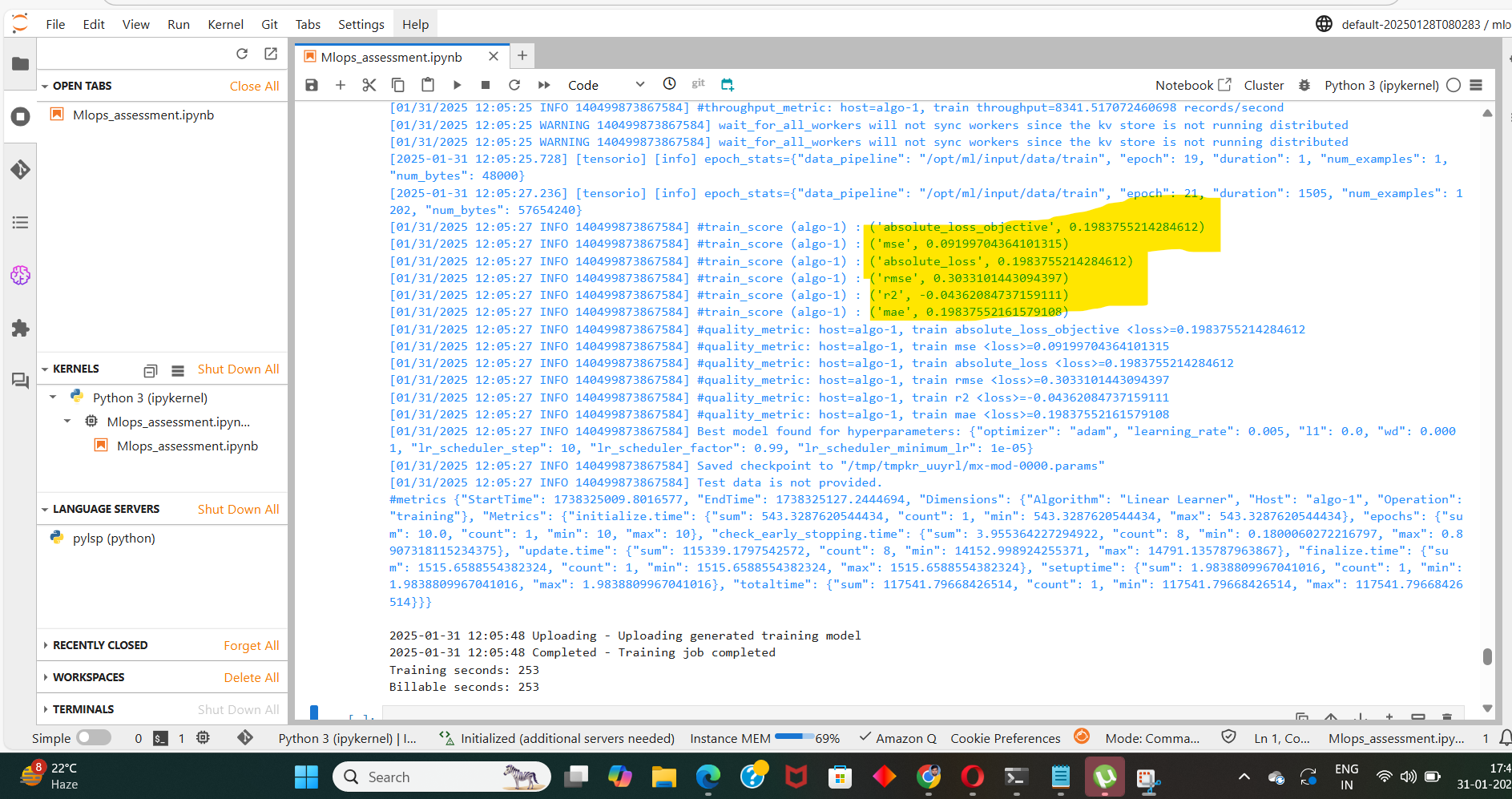


TRAINING THE MODEL 

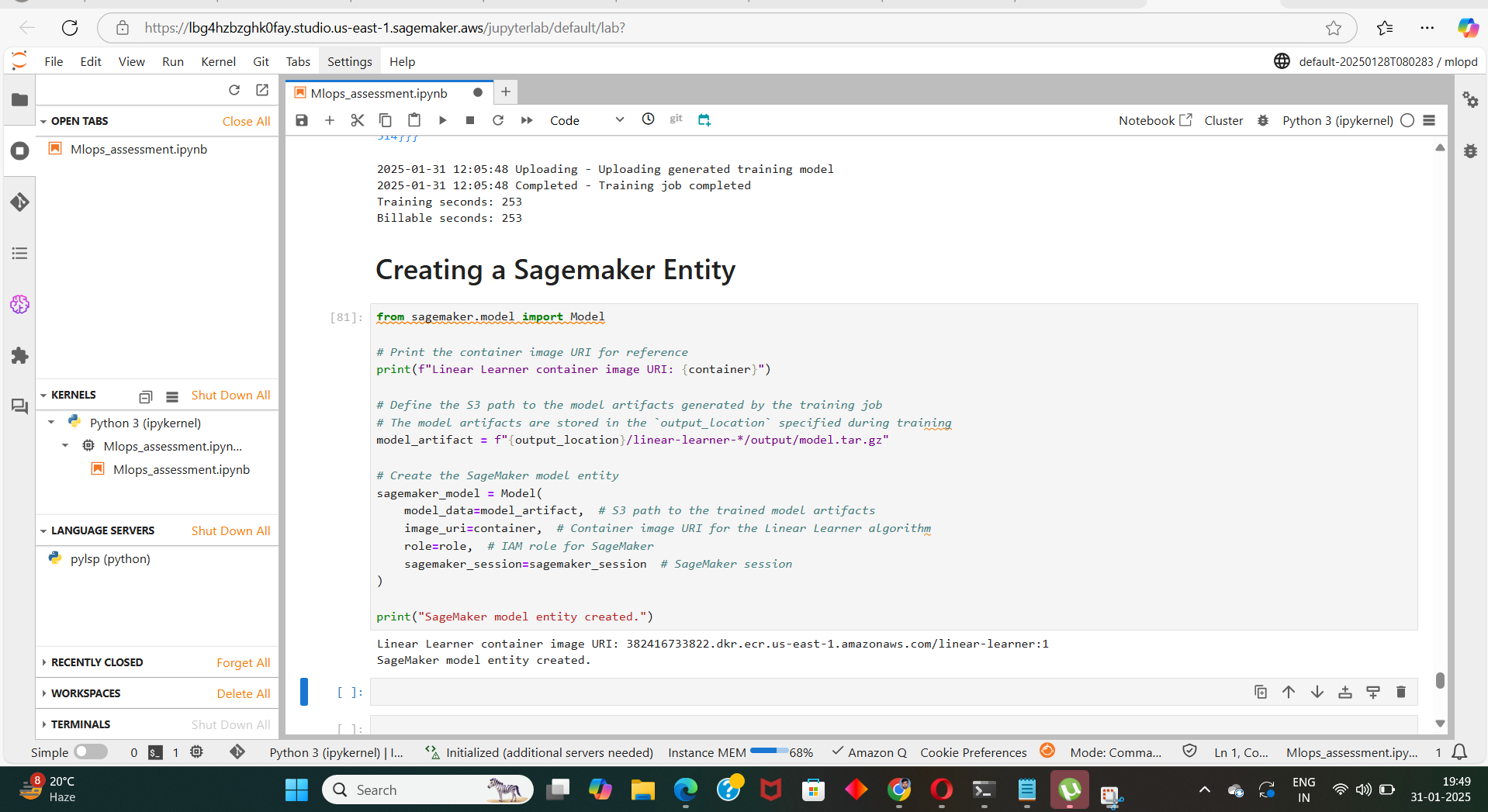
INPROGRESSING – TRAINING THE MODEL



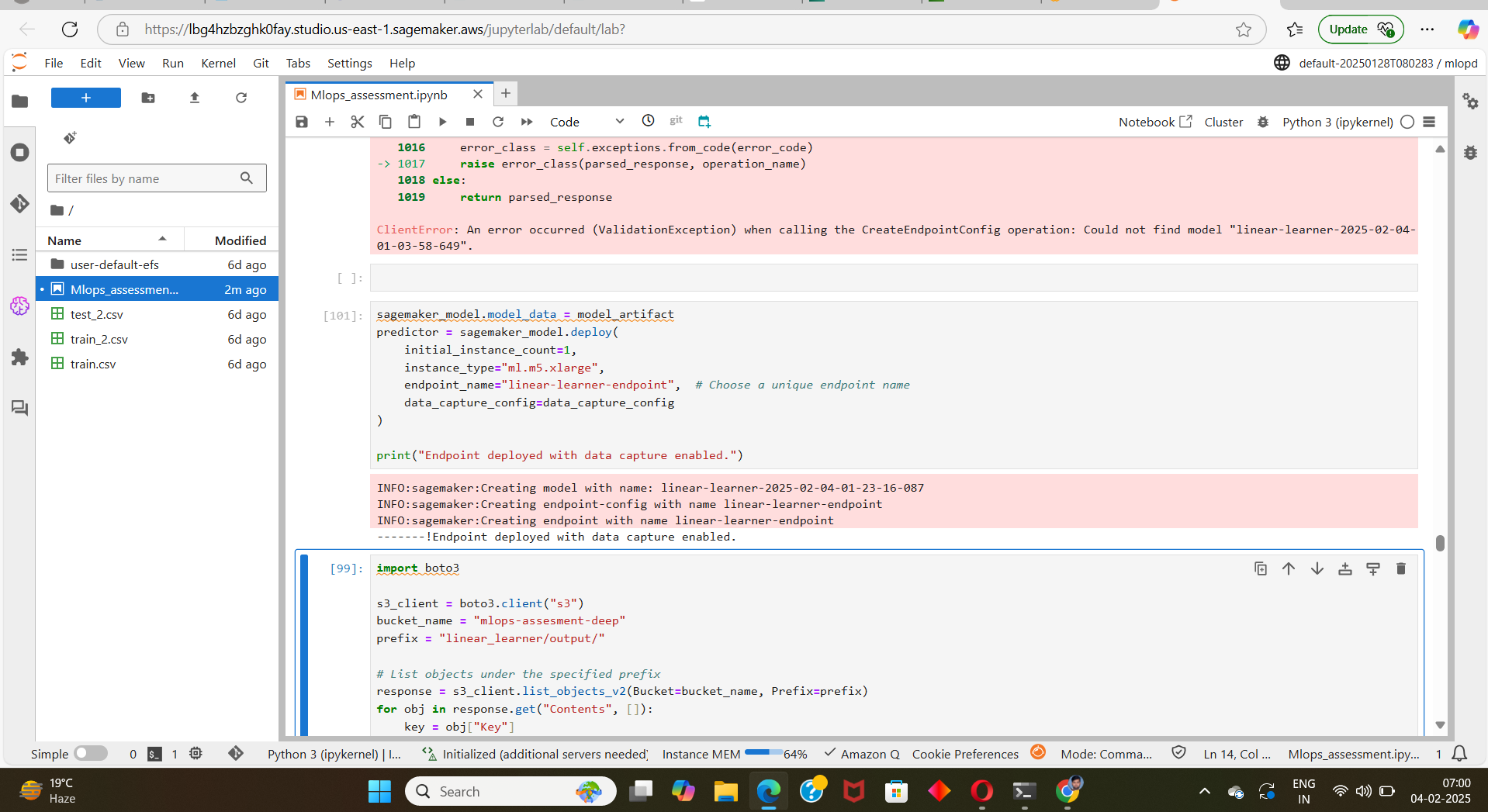
TRAIN END



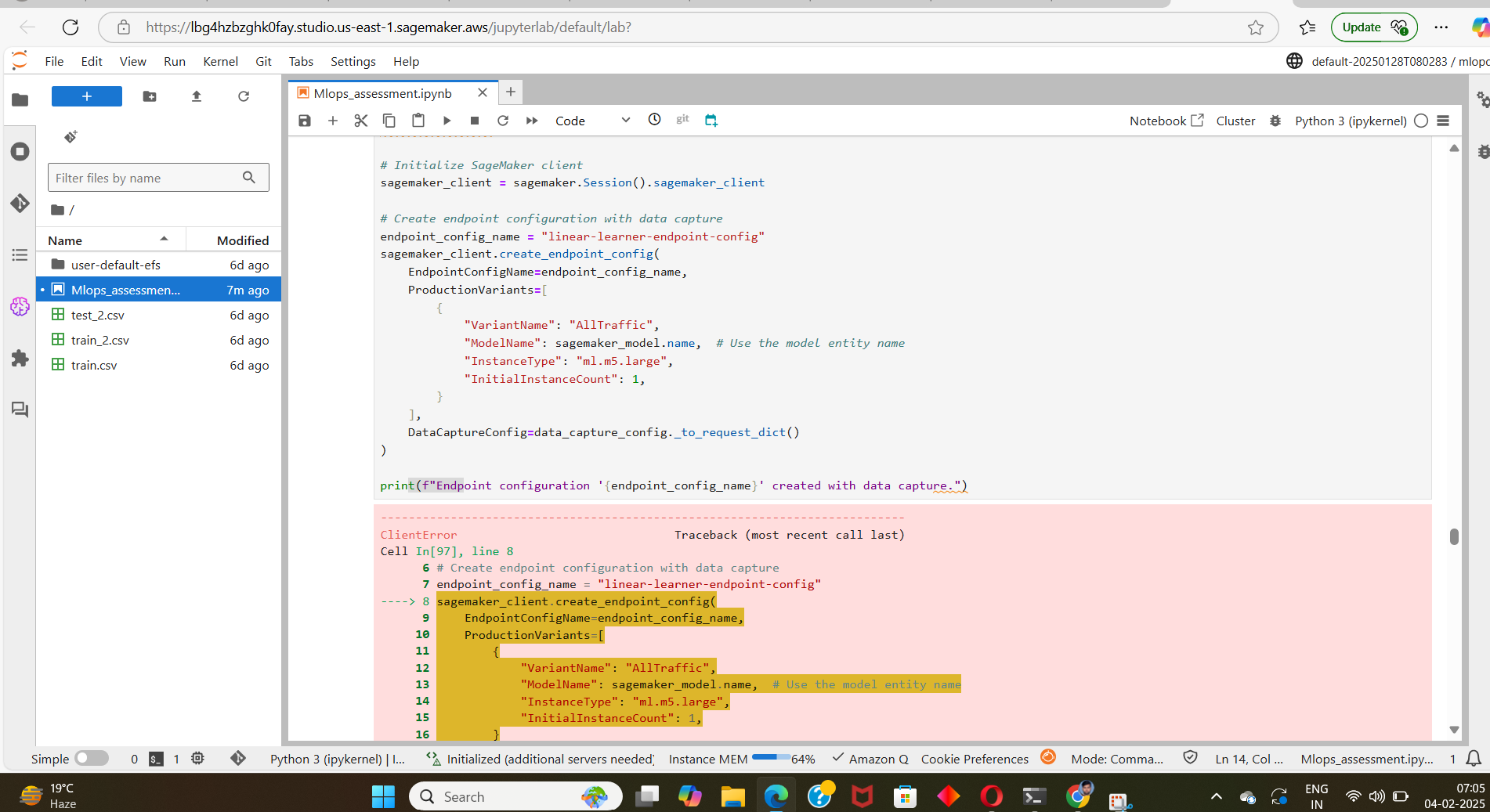
ENTITY CREATION



ENDPOINT DEPLOYMENT WITH DATA CAPTURE ENABLED

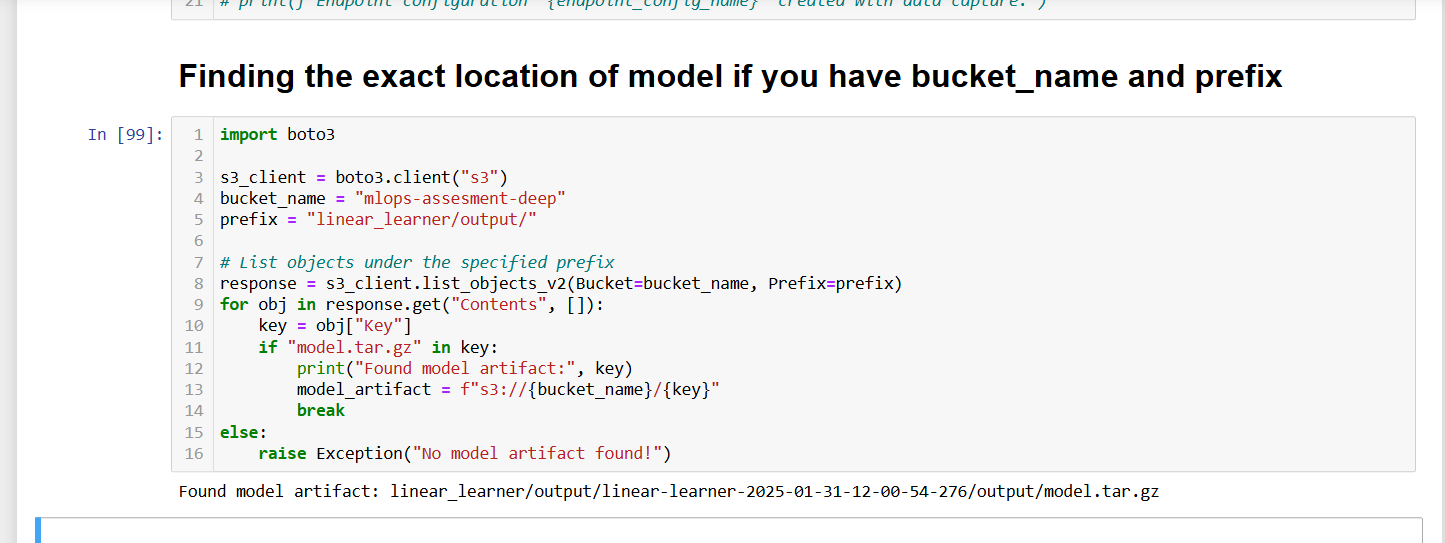


ENDPOINT ERROR

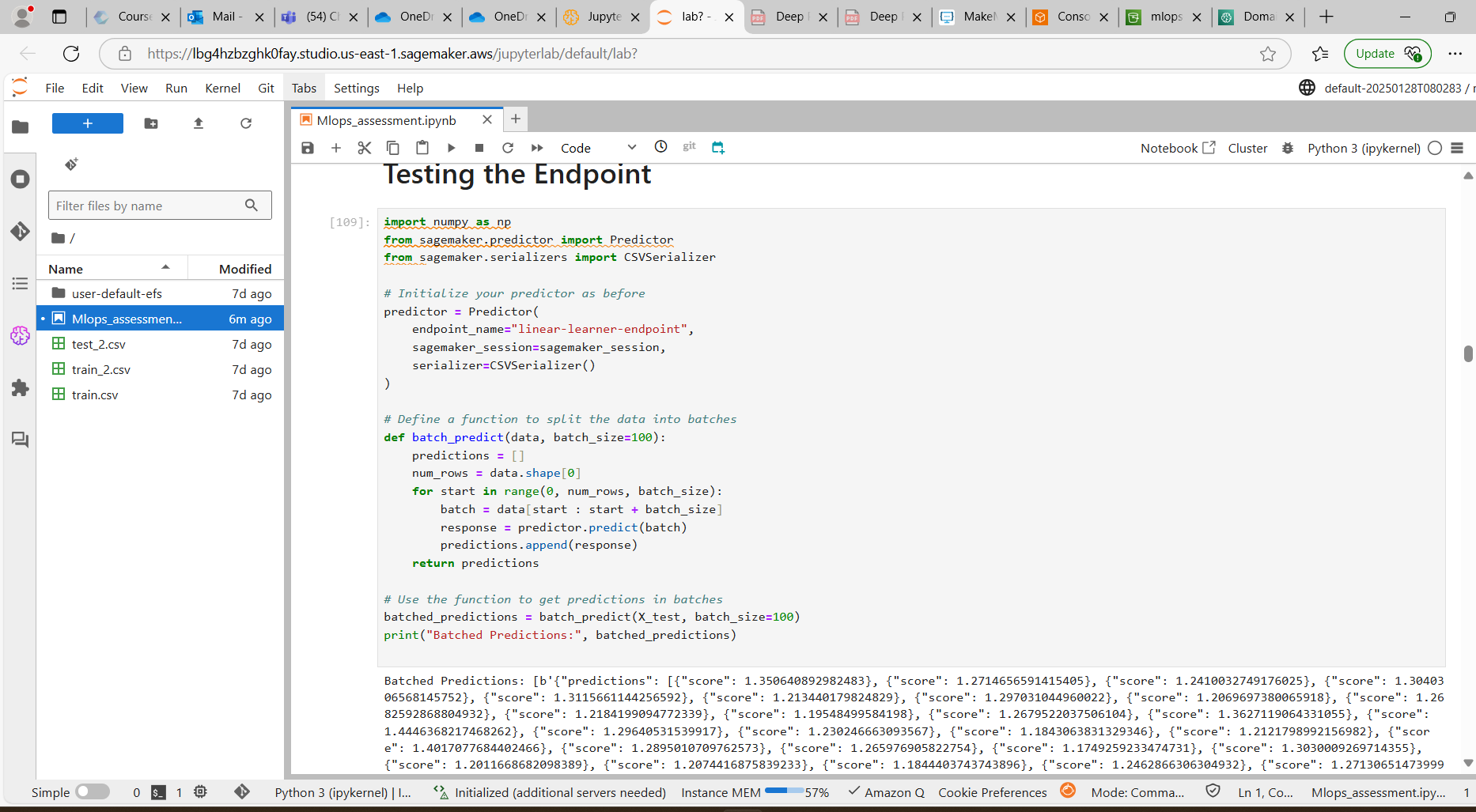


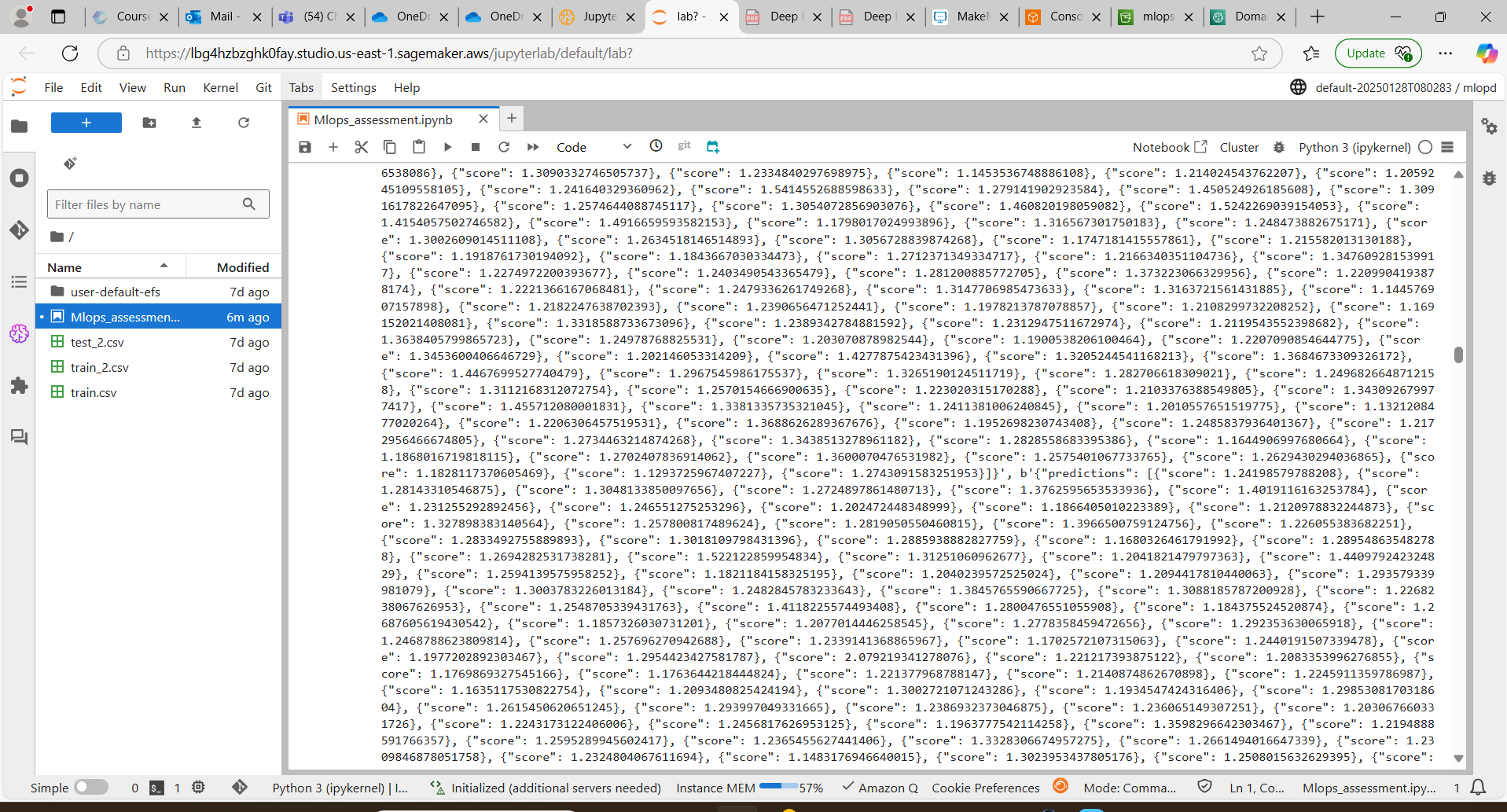
Challenges faced: client error it was not able to find the model location I wrote a code exact location of model.

Handling: I wrote a code to find the exact location of model.

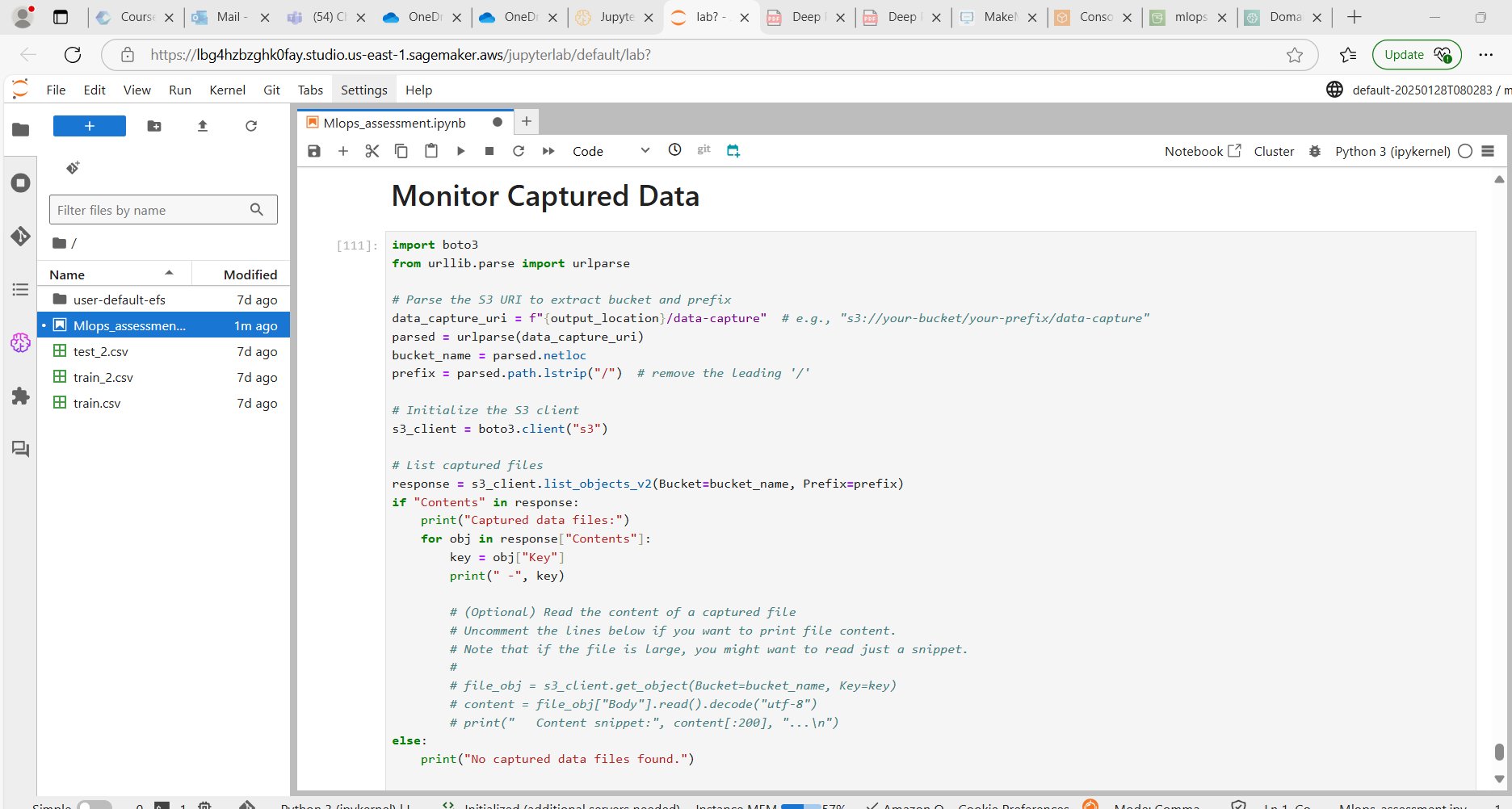


TESTING THE ENDPOINT

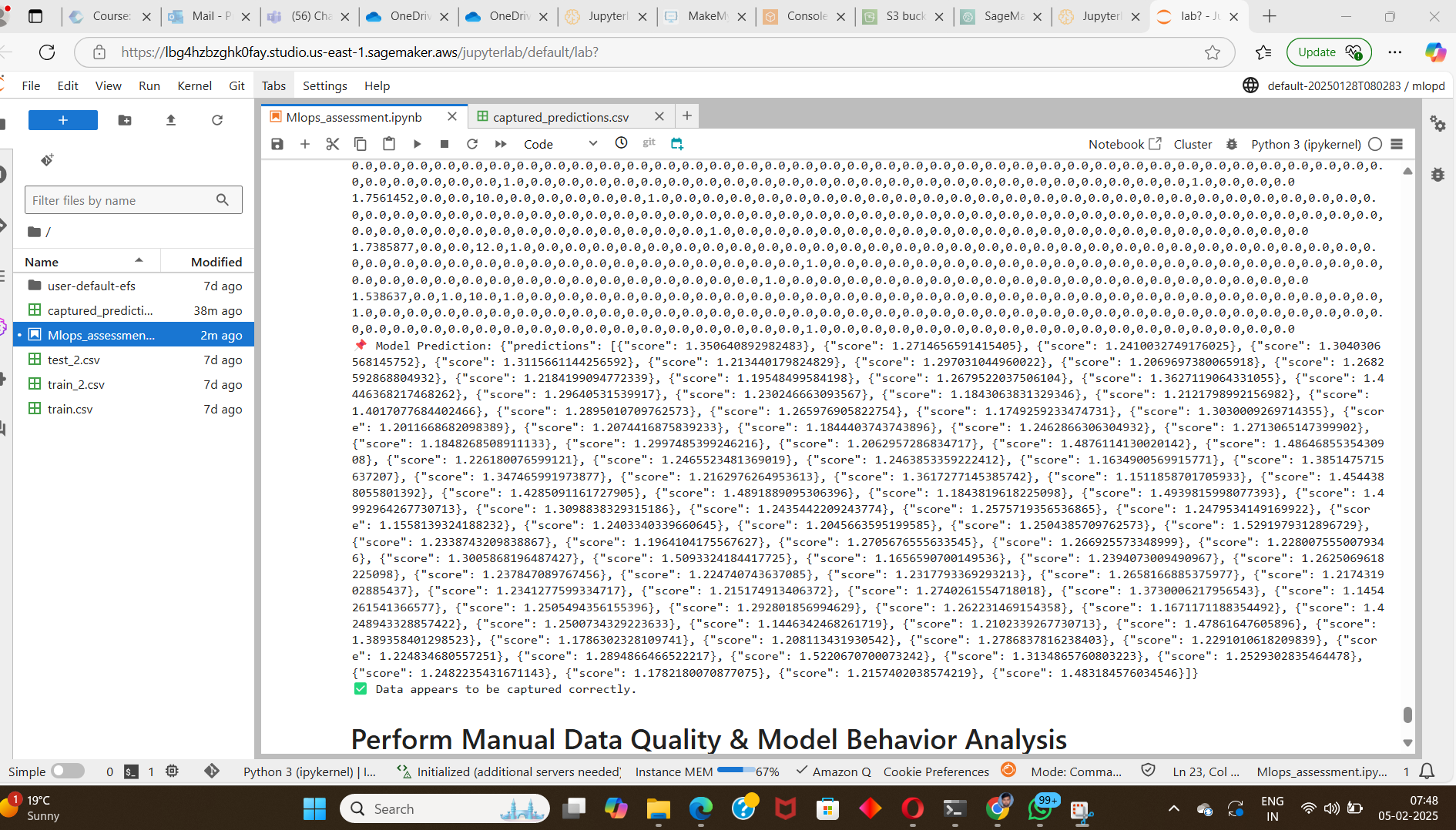
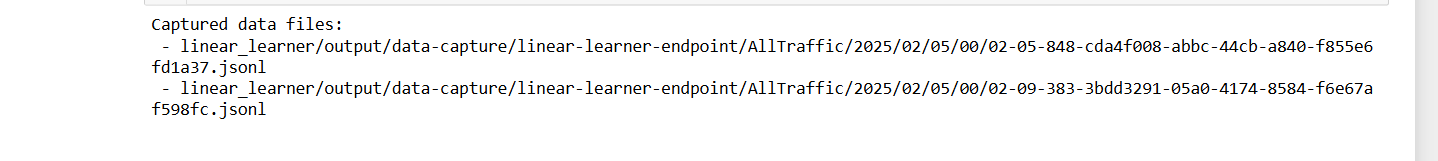




VERIFY THE CAPTURED DATA

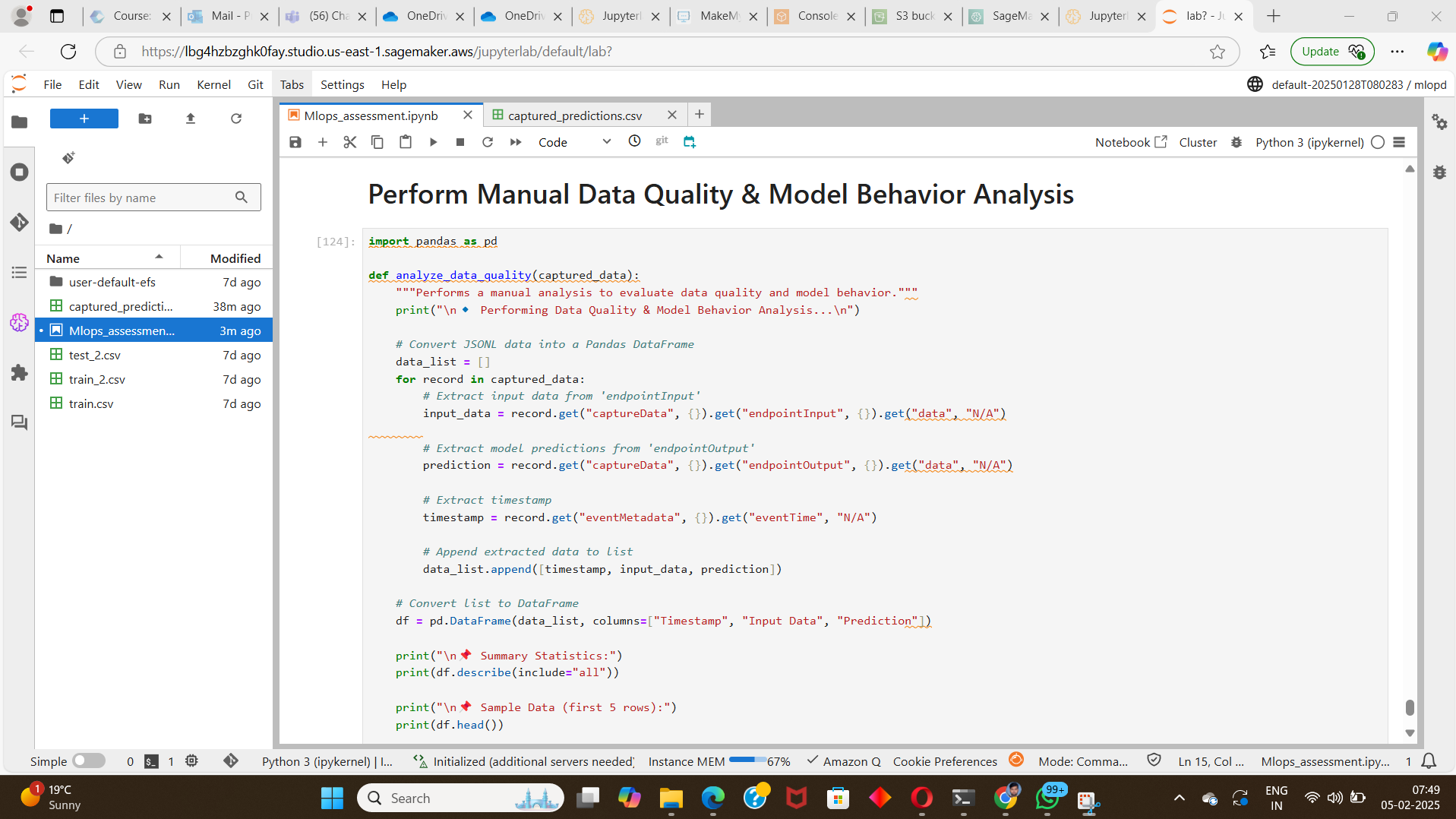


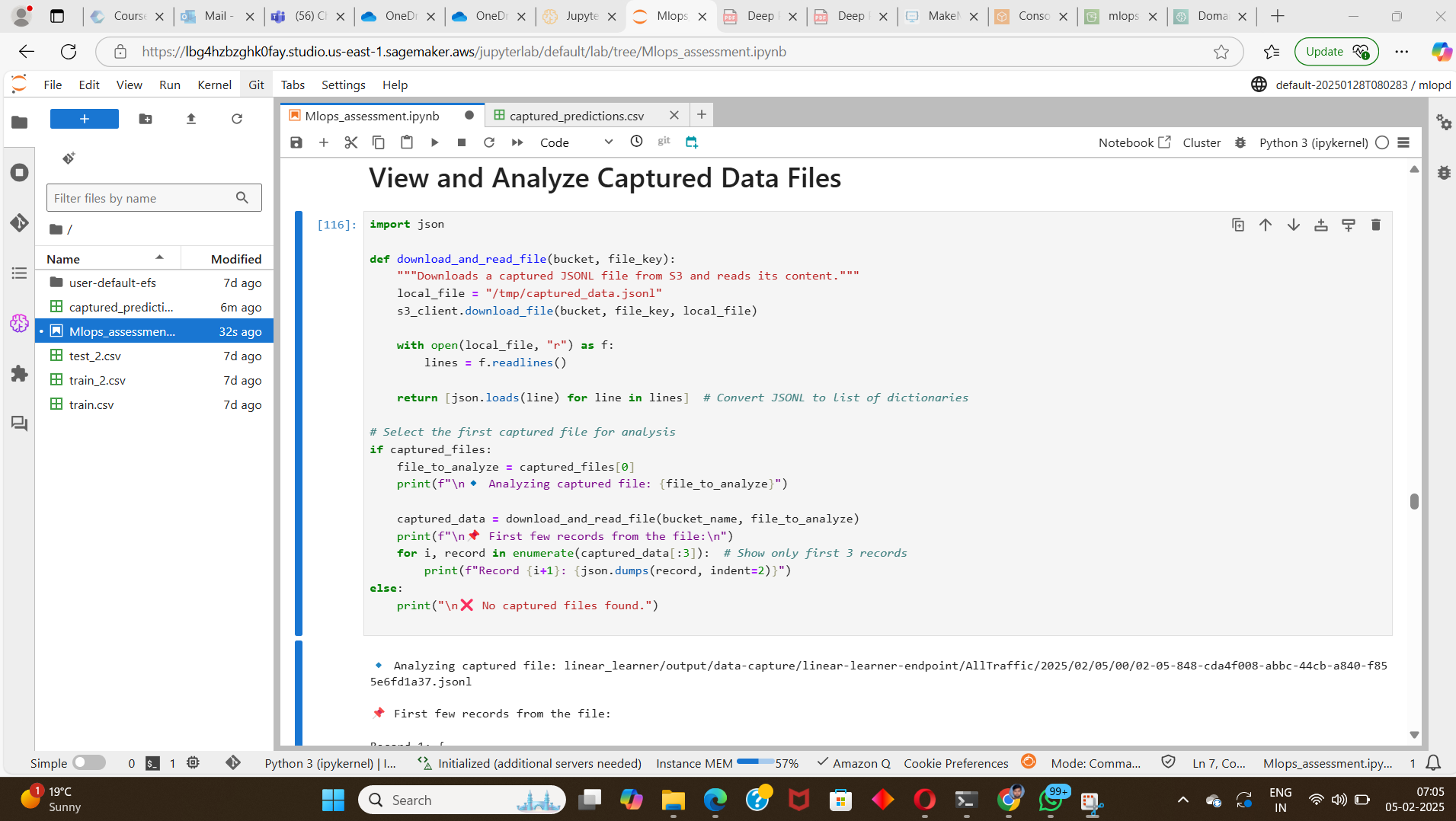
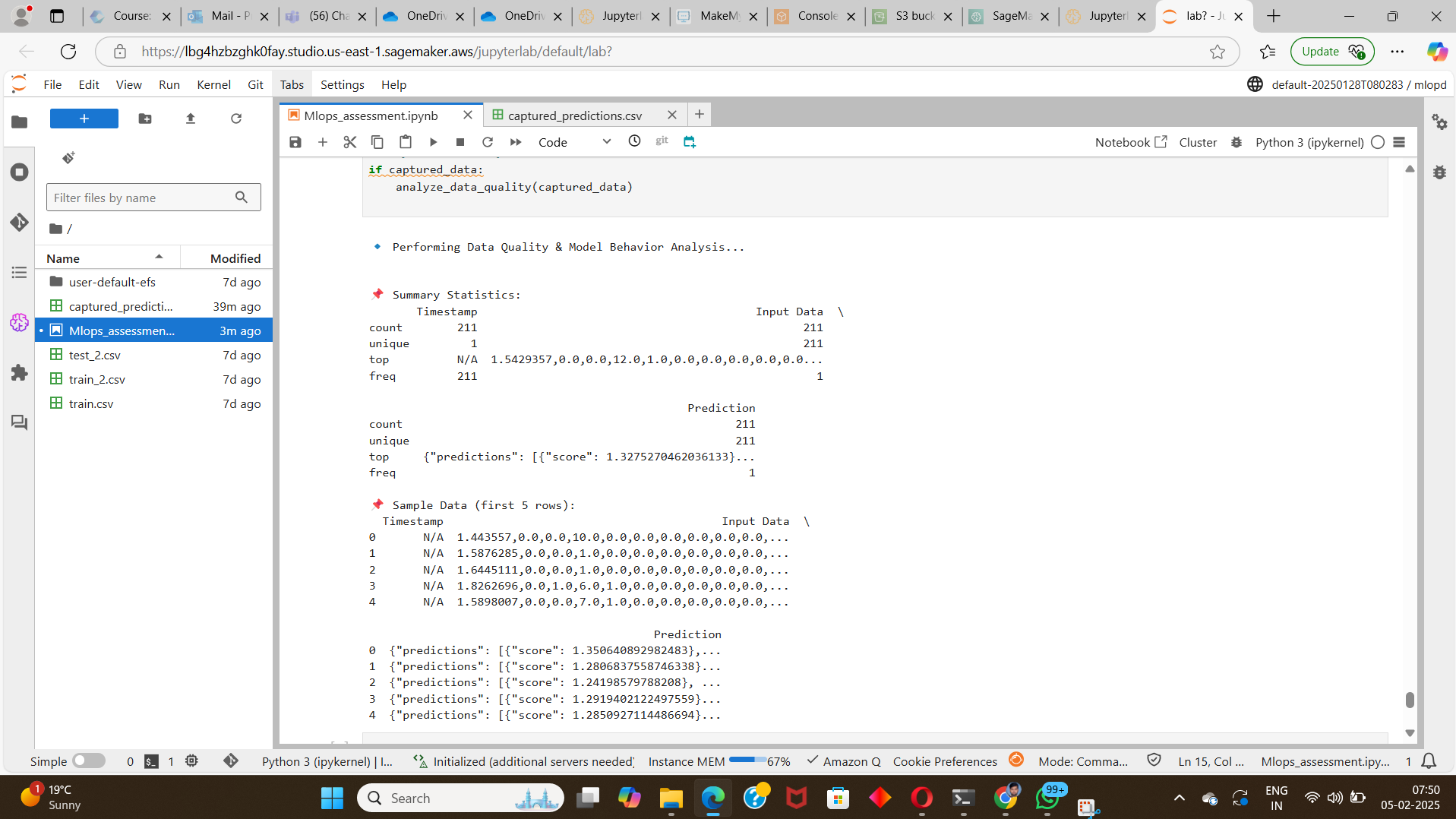
While Monitoring the Data I was not able to see because large file size so I passed the data in batches of 100 rows and monitored it



[C:\Users\Deep Kumar Prasad\Pictures\Screenshots\Mlops assesment\capture in csv format.mp4](file:///C:\Users\Deep%20Kumar%20Prasad\Pictures\Screenshots\Mlops%20assesment\capture%20in%20csv%20format.mp4)

**Behavioural Analysis**





COULD NOT COMPLETE THE BELOW CODE BECAUSE PROJECT TRANSIOTION AND INADEQUATE KNOWLEDGE FOR CICD.  
  
# .github/workflows/mlops-pipeline.yml

name: MLOps CI/CD Pipeline

on:

push:

branches: [ main ]

paths:

- 'scripts/\*\*'

- 'models/\*\*'

- 'data/\*\*'

jobs:

build-test-deploy:

strategy:

matrix:

os: [ubuntu-latest, windows-latest]

runs-on: ${{ matrix.os }}

container:

image: amazonlinux:2023

steps:

- name: Checkout code

uses: actions/checkout@v3

- name: Set up Python 3.8

uses: actions/setup-python@v4

with:

python-version: 3.8

- name: Install dependencies

run: |

python -m pip install --upgrade pip

pip install -r requirements.txt

pip install black flake8 pytest

- name: Lint with flake8

run: flake8 scripts/ --count --show-source --statistics

- name: Format with Black

run: black --check scripts/

- name: Run unit tests

run: pytest tests/

- name: Configure AWS credentials

uses: aws-actions/configure-aws-credentials@v2

with:

aws-access-key-id: ${{ secrets.AWS\_ACCESS\_KEY\_ID }}

aws-secret-access-key: ${{ secrets.AWS\_SECRET\_ACCESS\_KEY }}

aws-region: us-east-1

- name: Train model

run: |

python scripts/train.py \

--bucket-name ${{ secrets.S3\_BUCKET }} \

--role-arn ${{ secrets.SAGEMAKER\_ROLE }}

- name: Deploy model

run: |

python scripts/deploy.py \

--model-s3-path s3://${{ secrets.S3\_BUCKET }}/models/model.tar.gz \

--endpoint-name sales-forecast-endpoint \

--instance-type ml.m5.large