**Cloud Final Project**

Team Members: Jaya Pavani Pathakota, Lakshmi Praneetha Singamreddy, Shruthi Asolkar

**Azure WebApp Url:** <https://cloudgroup44-fgdkbfc6cghzbpa6.centralus-01.azurewebsites.net>

**GitHub Link:** <https://github.com/jayapavanip/Data-Science-Project-on-Retail-Experience-in-Kroger>

Screenshots for Project Requirements:

1. **Write-Up on ML Models:**

* Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship and minimizes the error between predicted and actual values using the least squares method. While it's simple and interpretable, it struggles with non-linear relationships and complex datasets.

* Random Forest

Random Forest is an ensemble learning method based on decision trees. It creates multiple trees during training and combines their predictions (averaging for regression or majority voting for classification). It handles non-linear relationships, reduces overfitting, and performs well on complex data. However, it can be computationally intensive and less interpretable.

* Gradient Boosting

Gradient Boosting is another ensemble technique that builds models sequentially, with each new model correcting the errors of the previous ones. It uses weak learners (e.g., shallow decision trees) and combines them for a strong predictive model. Gradient Boosting excels in accuracy and handles non-linear relationships well but requires fine-tuning to avoid overfitting.

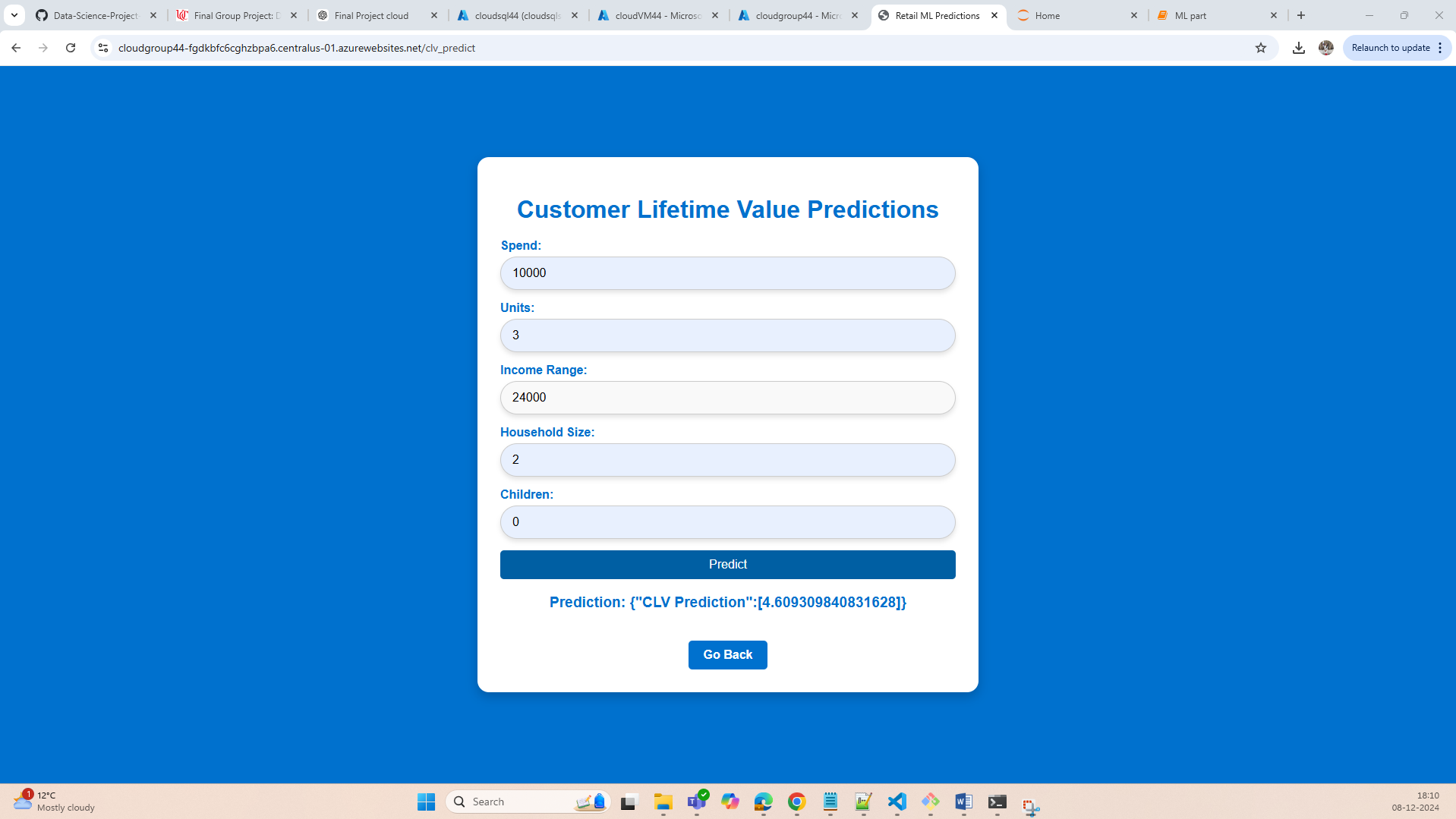
* Predictive Modelling for CLV

Gradient Boosting is the most suitable technique for predicting Customer Lifetime Value (CLV). CLV involves non-linear relationships and complex interactions between customer behaviour, spending patterns, and demographics. Gradient Boosting’s ability to capture such complexities makes it ideal for accurately forecasting long-term revenue potential, enabling businesses to prioritize high-value customers effectively.

How can we predict long-term revenue potential to prioritize high-value customers?

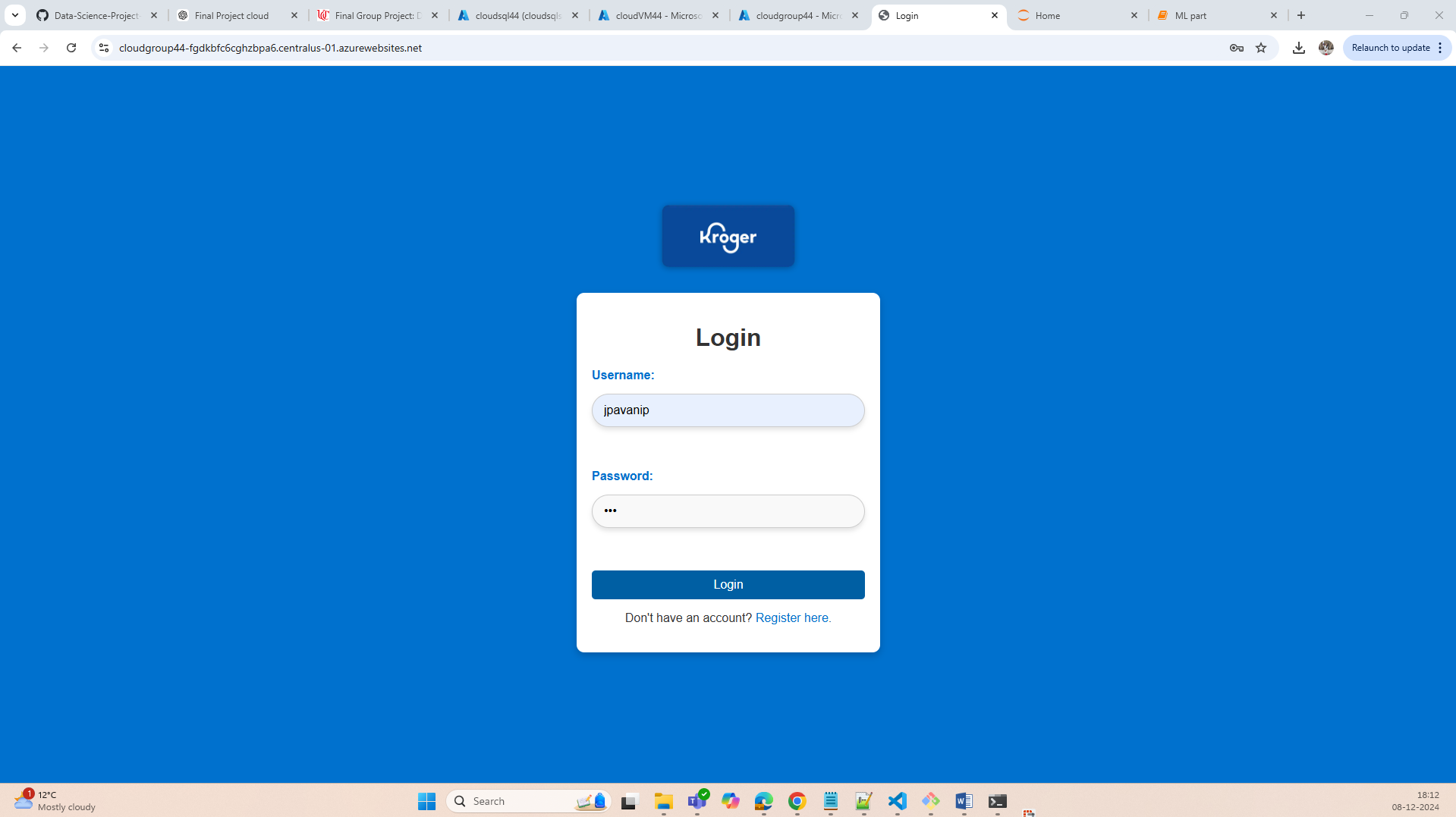
To predict long-term revenue potential, we can identify high-value customers by implementing a Customer Lifetime Value (CLV) prediction algorithm using Gradient Boosting, as demonstrated below:



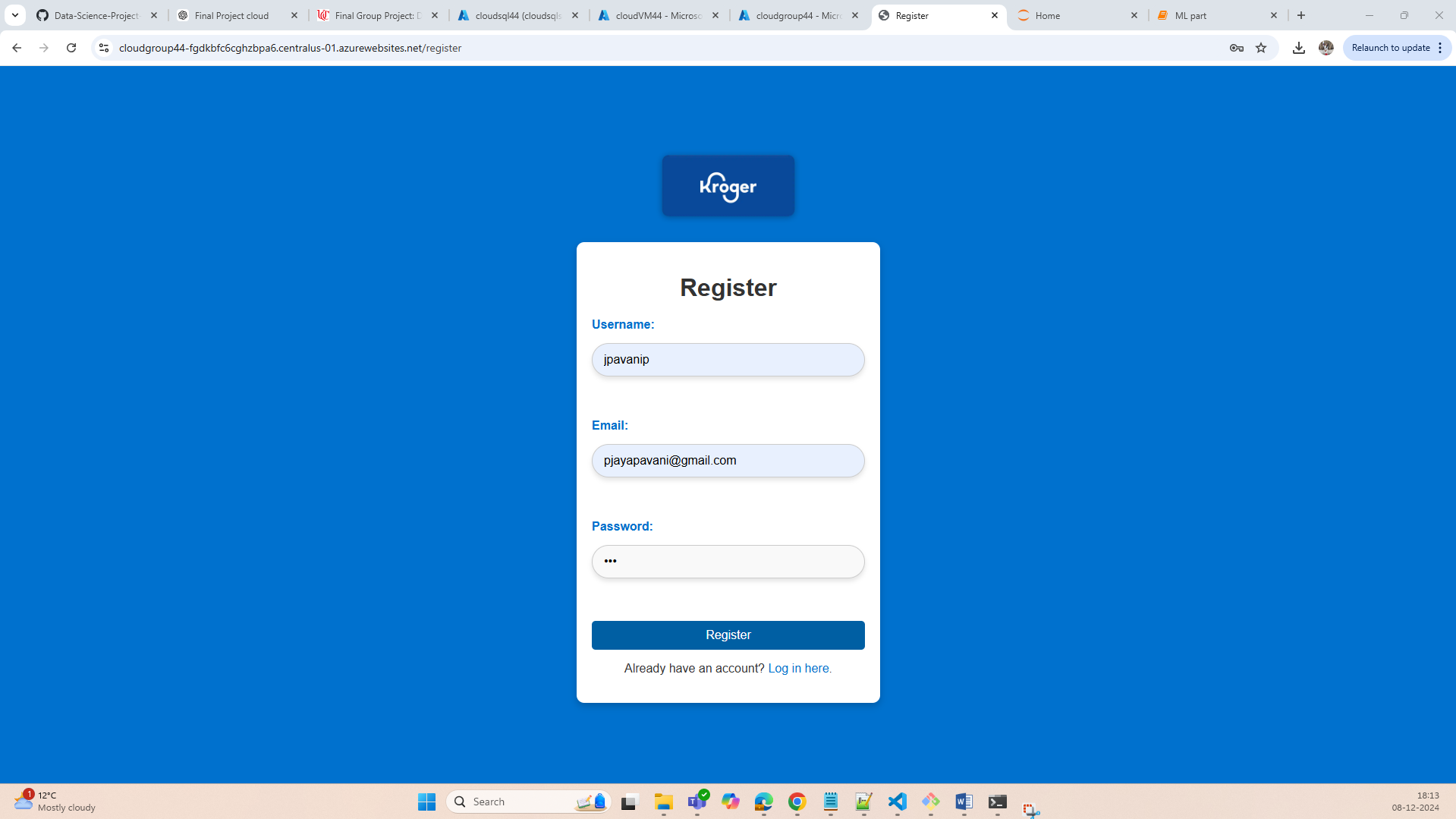


1. **Web Server Setup:**

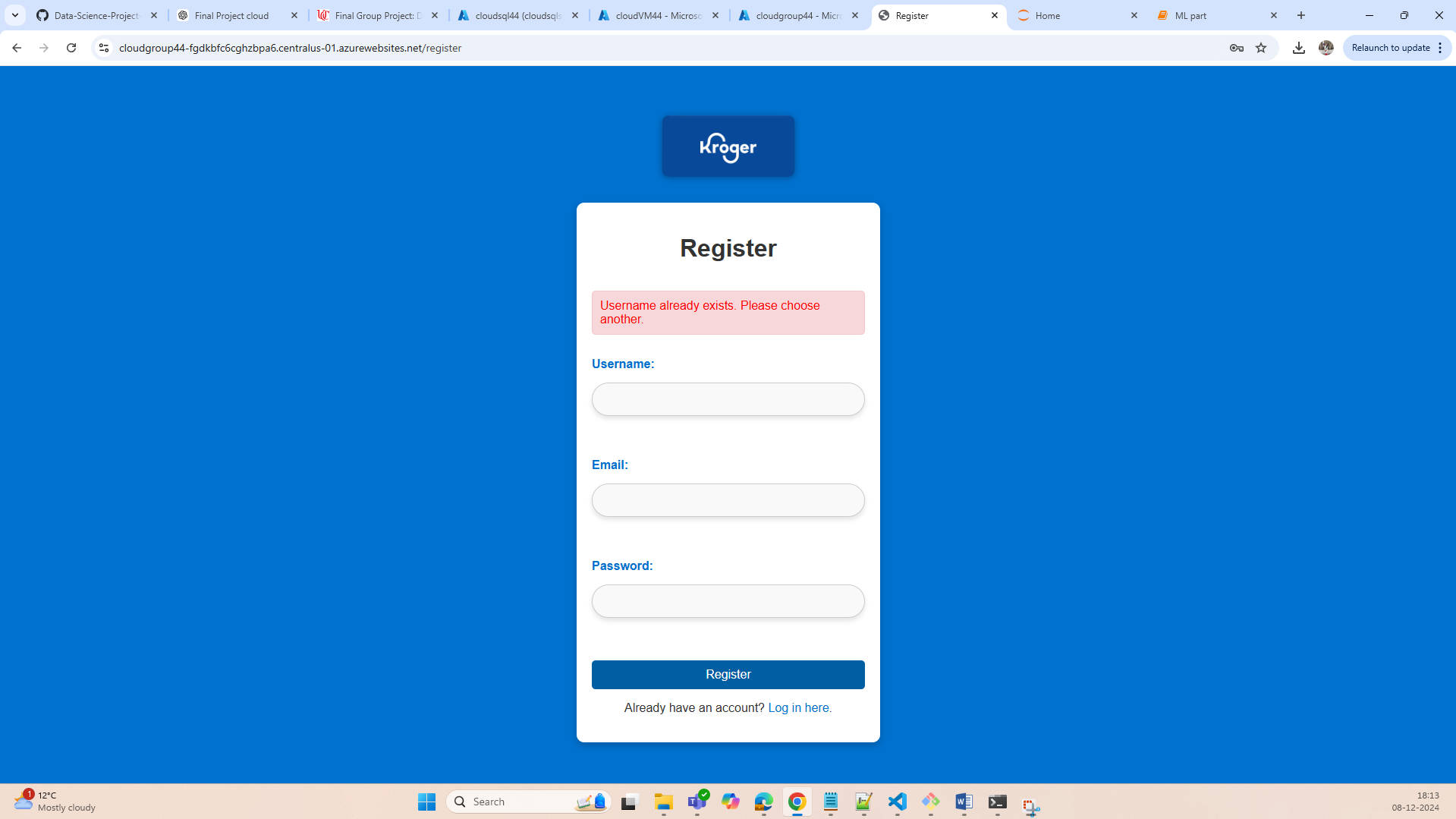
Login Page:



Registration page:



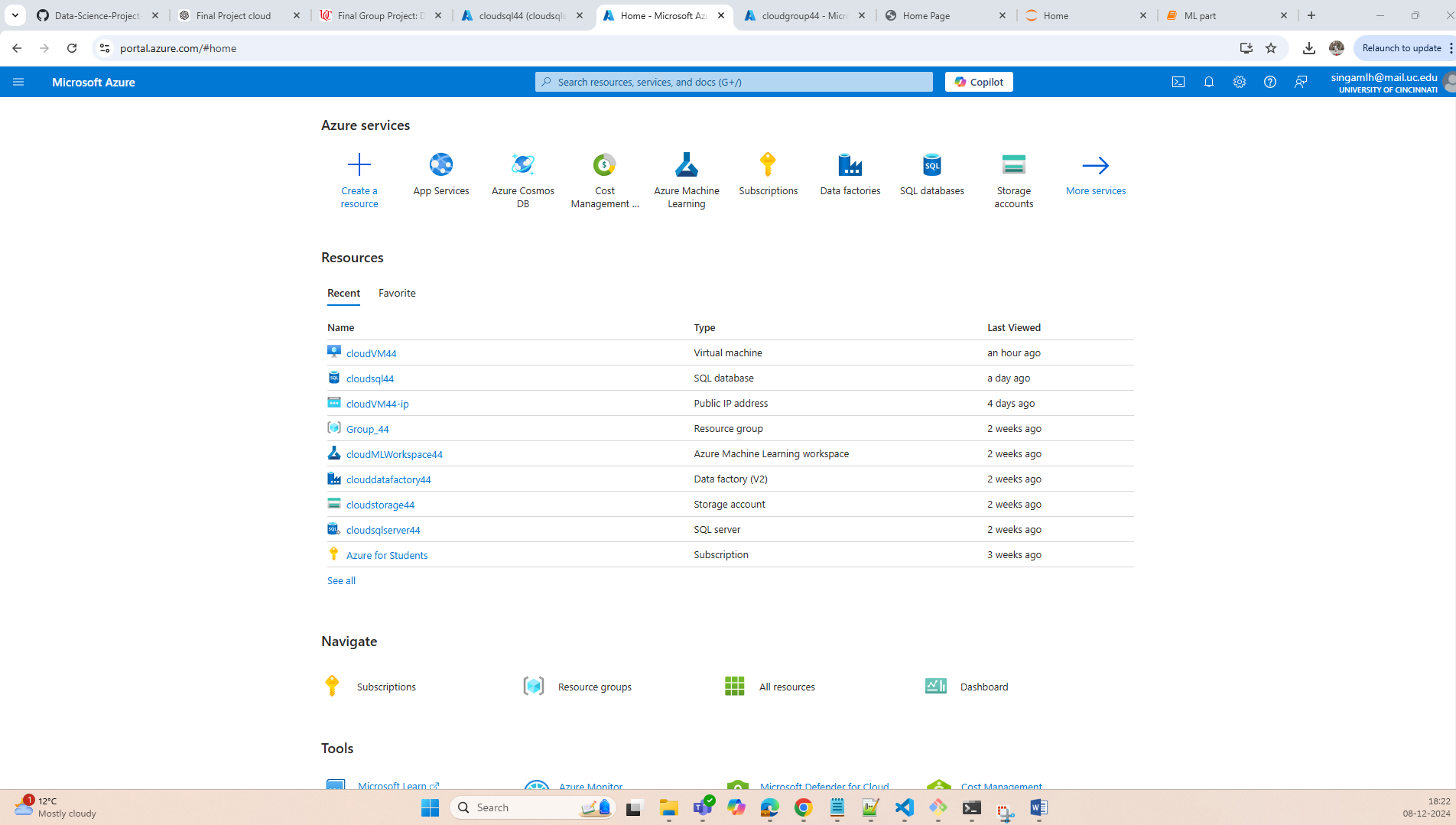
Added validations for register and login pages too, sample example:



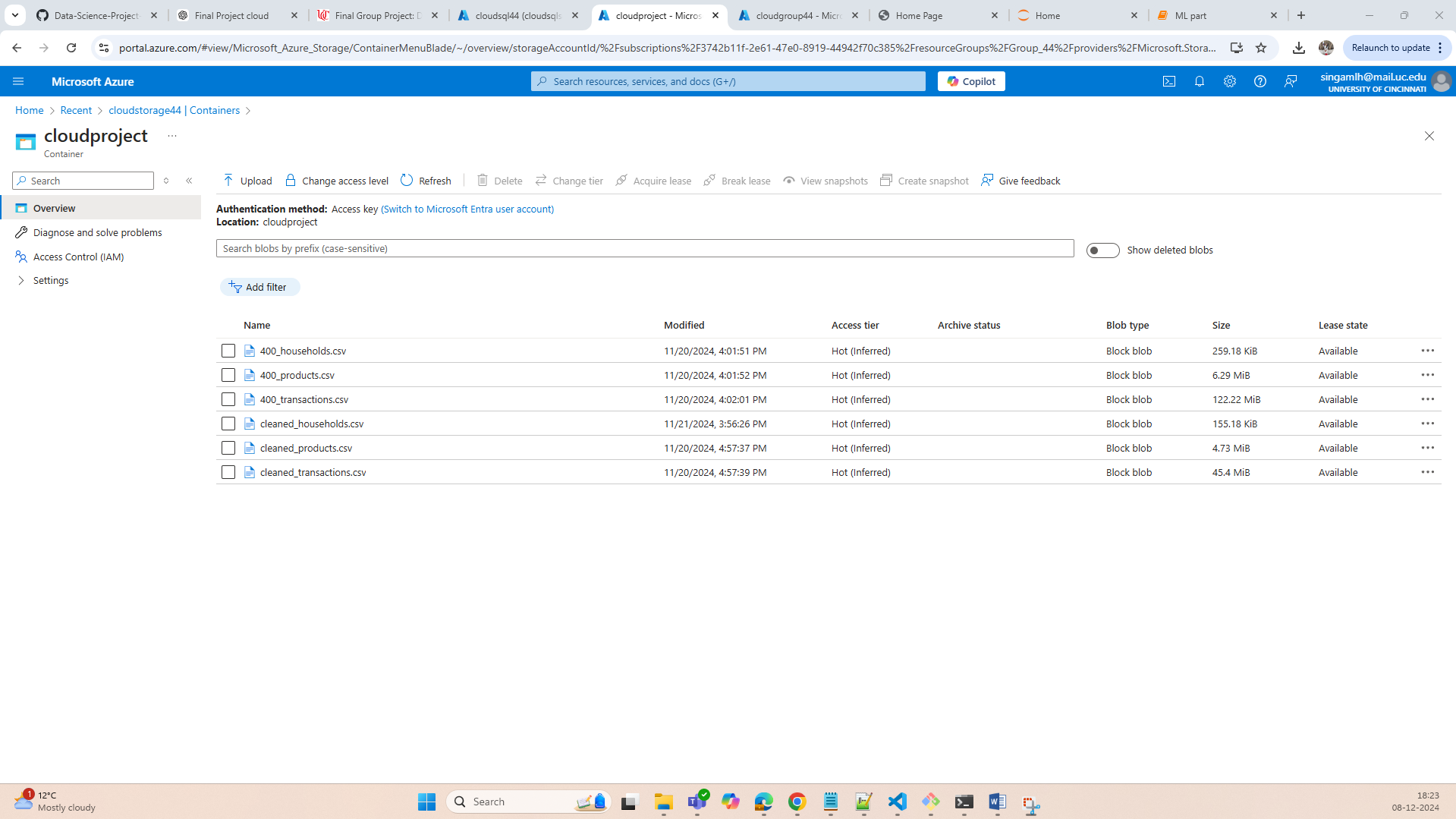
1. **Datastore and Data Loading:**

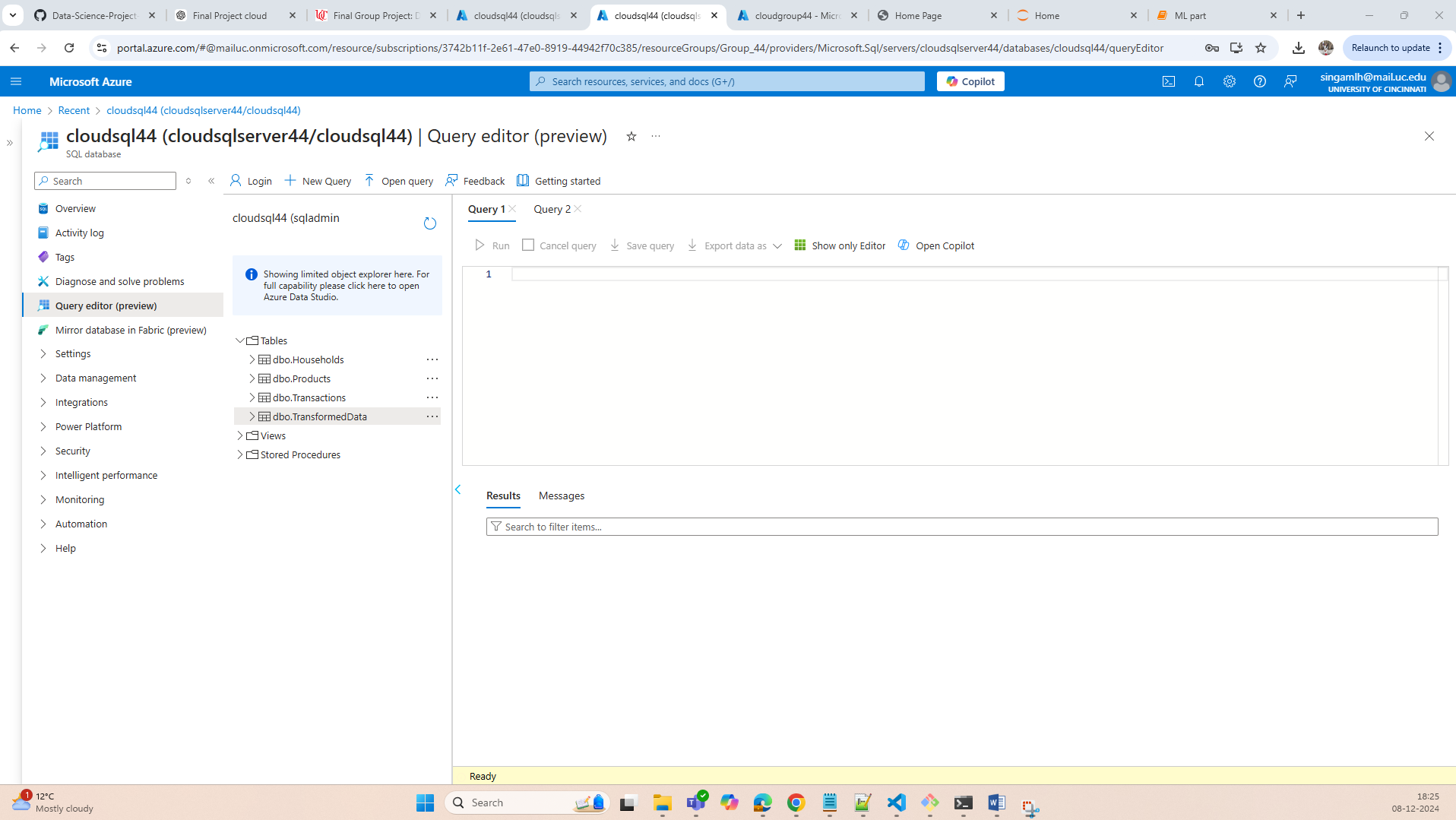
Procedure:

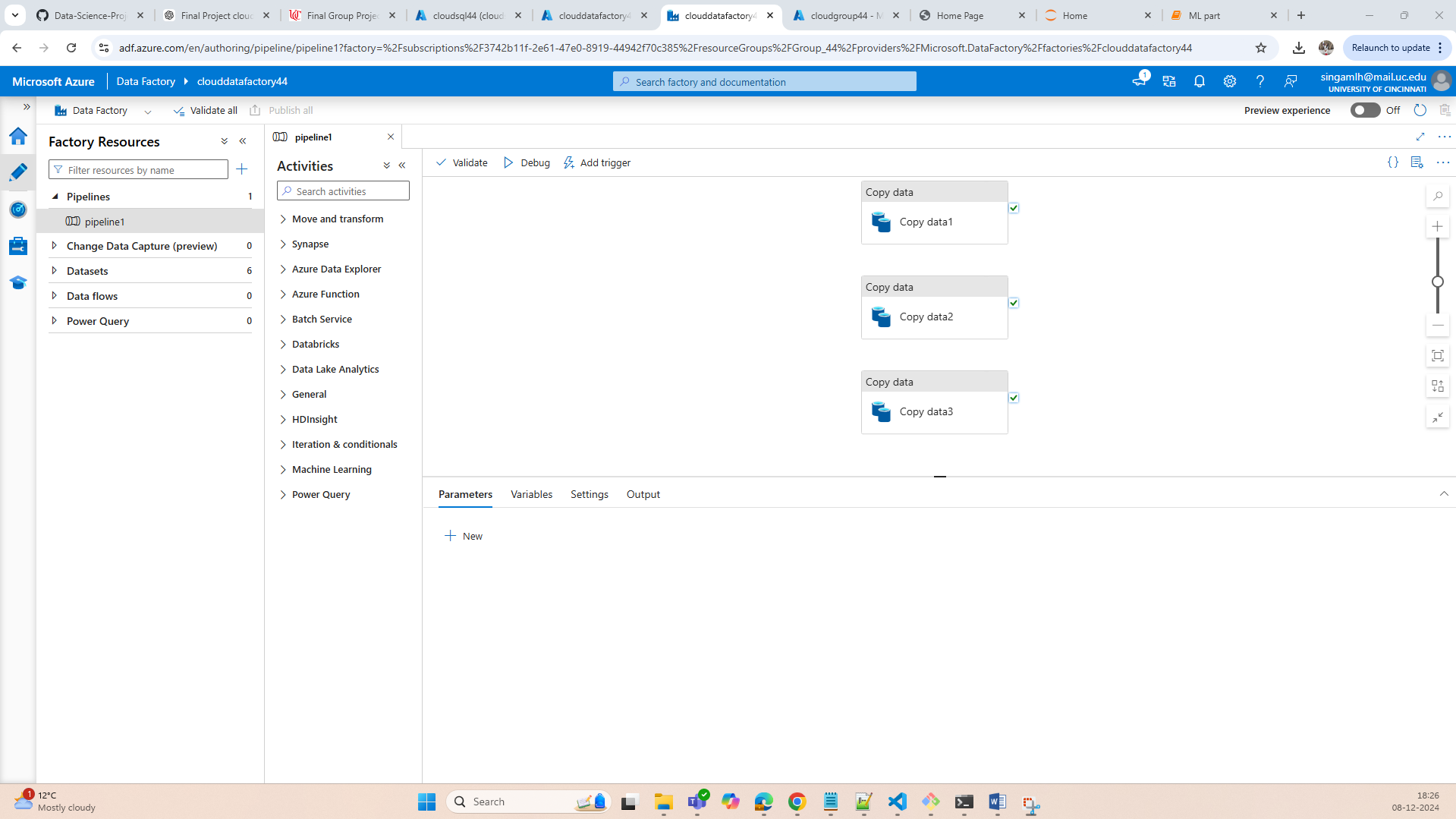
1. Created a Blob Storage and uploaded the 3 CSV flies. And then created an Azure SQL Database, followed by copying data through pipeline from Storage to SQL Database using Azure Data Factory, below are the screenshots:

Azure Home:  


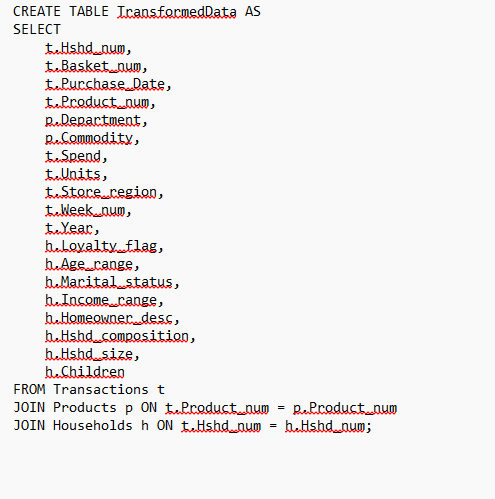
Azure Blob Storage:

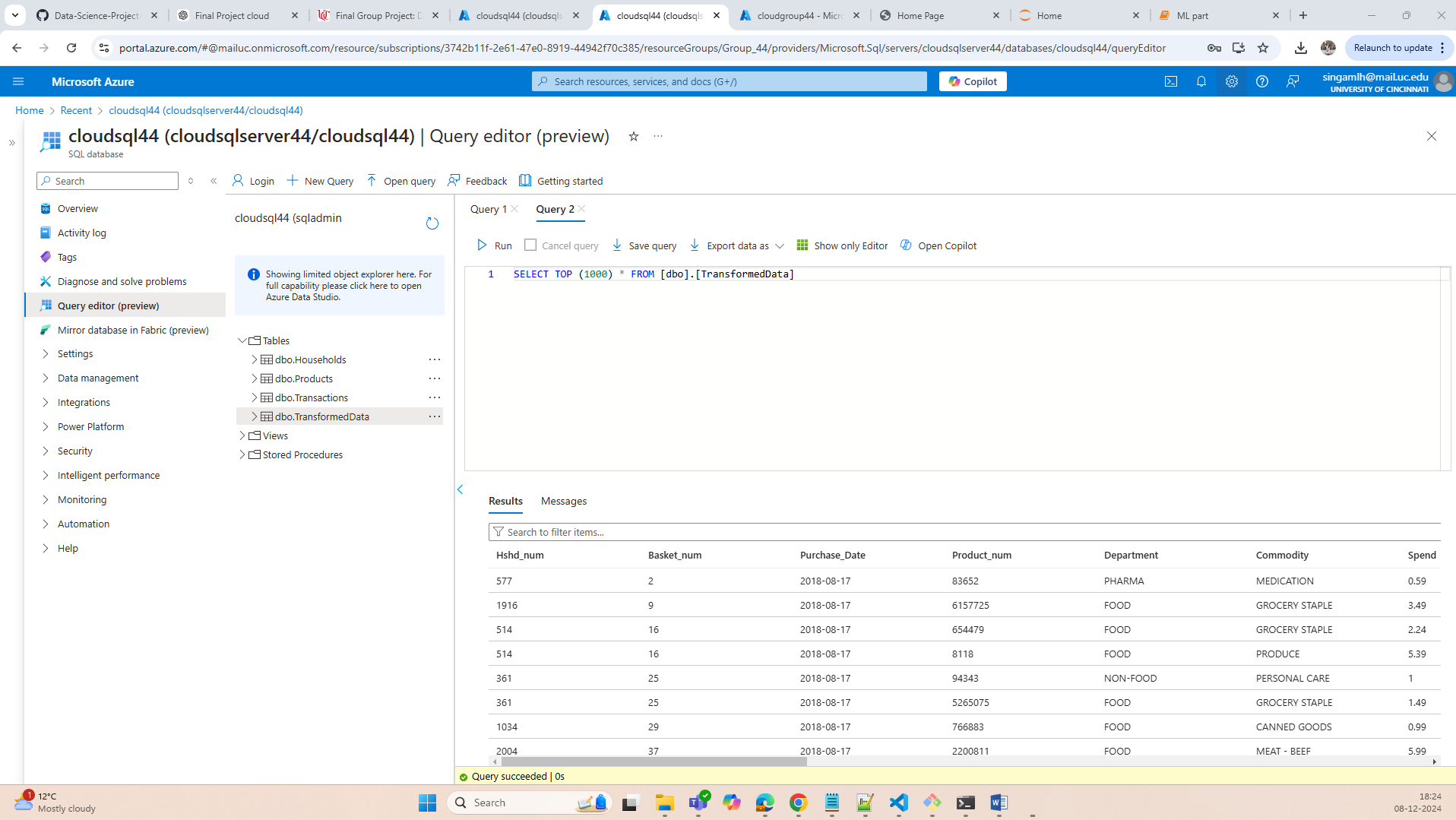


Azure SQL Database:  


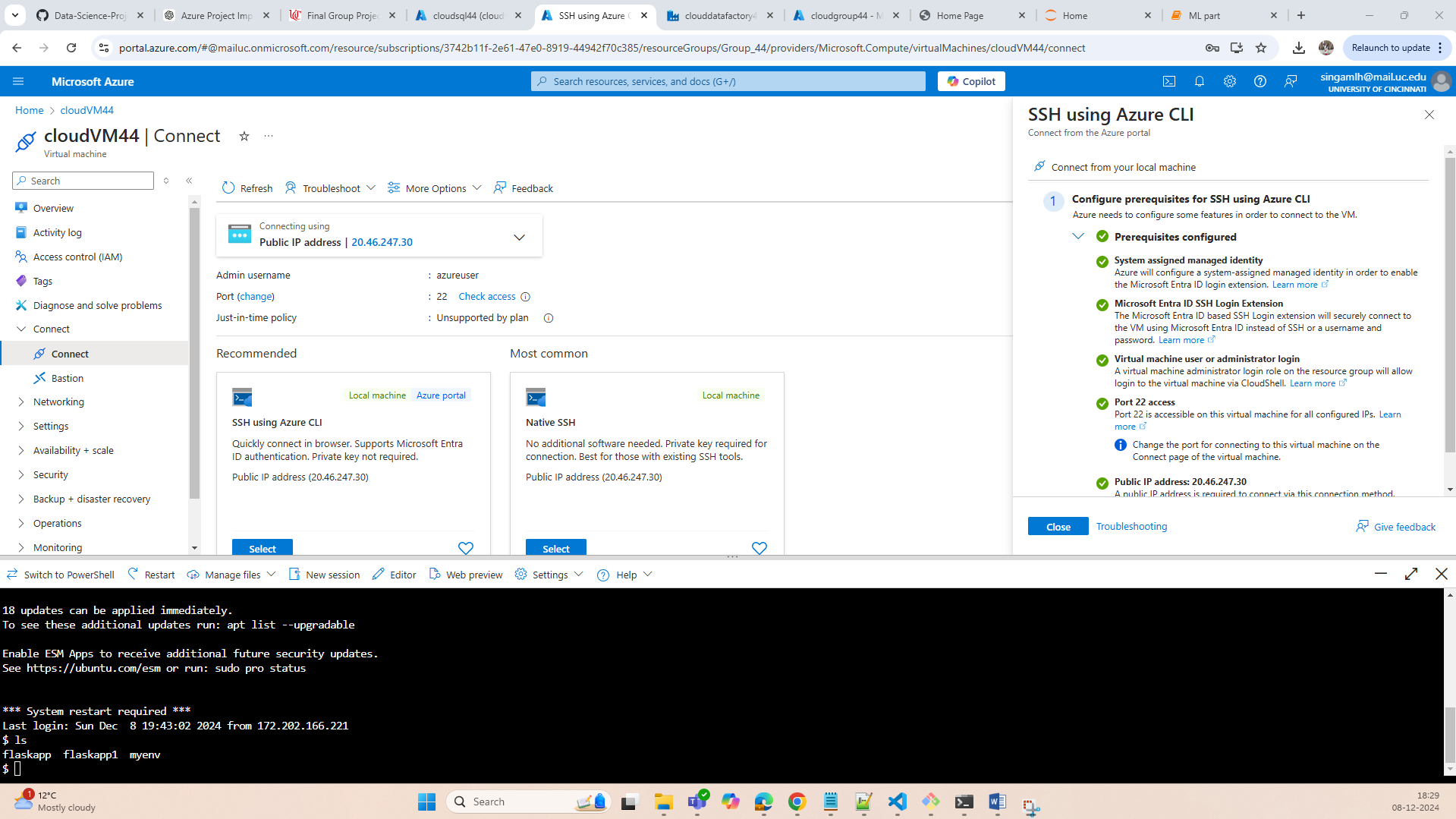
Azure Data Factory  


1. Created a TransformedData Table in Azure SQL by joining the 3 tables [Transactions, Products, Households].

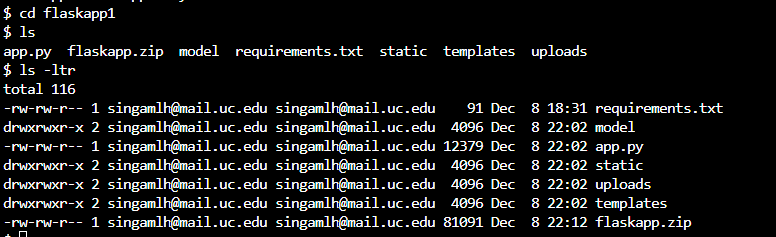
Join Query:  




1. Created an Azure VM and developed the flaskapp application, below is the folder structure:

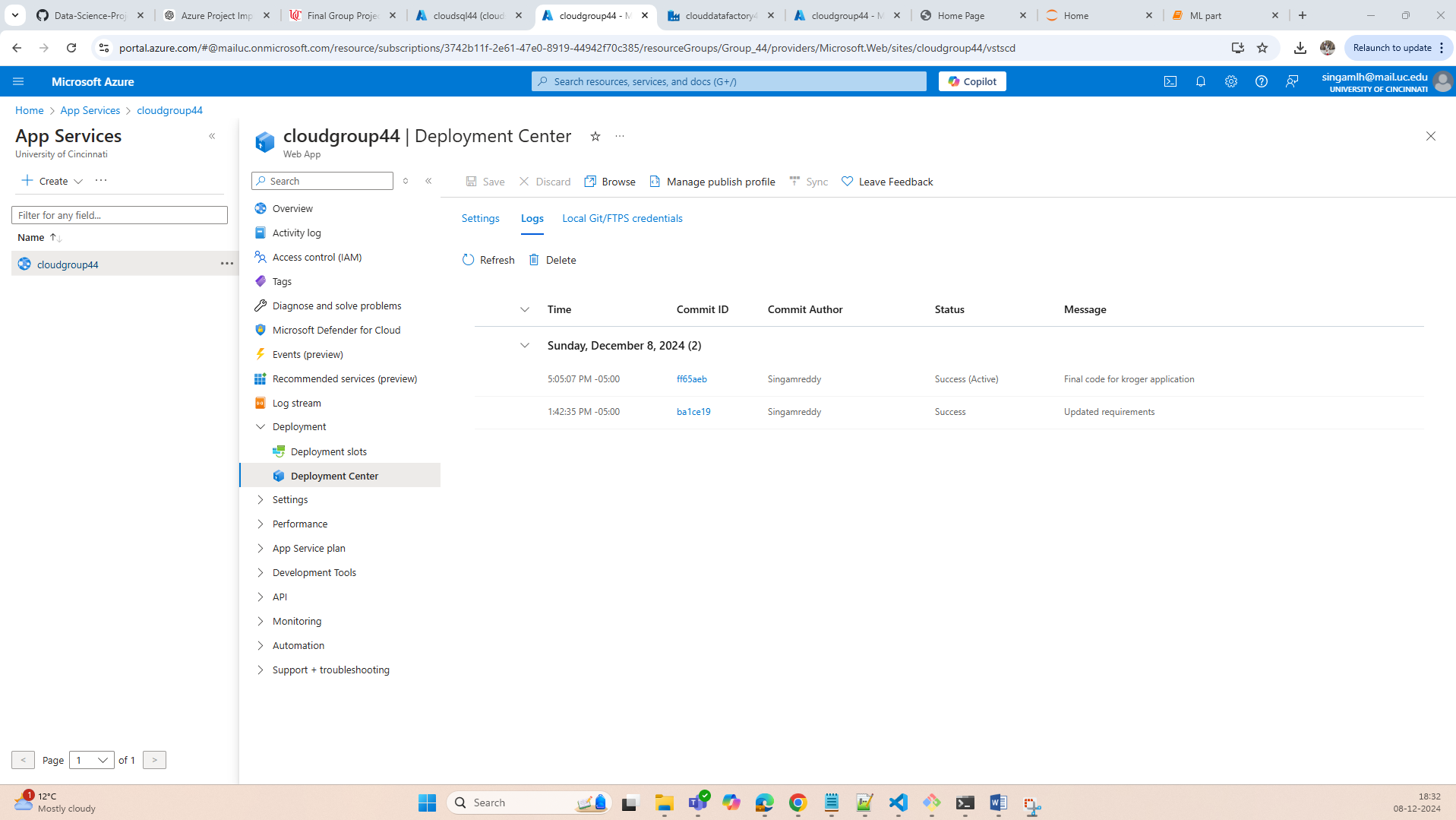


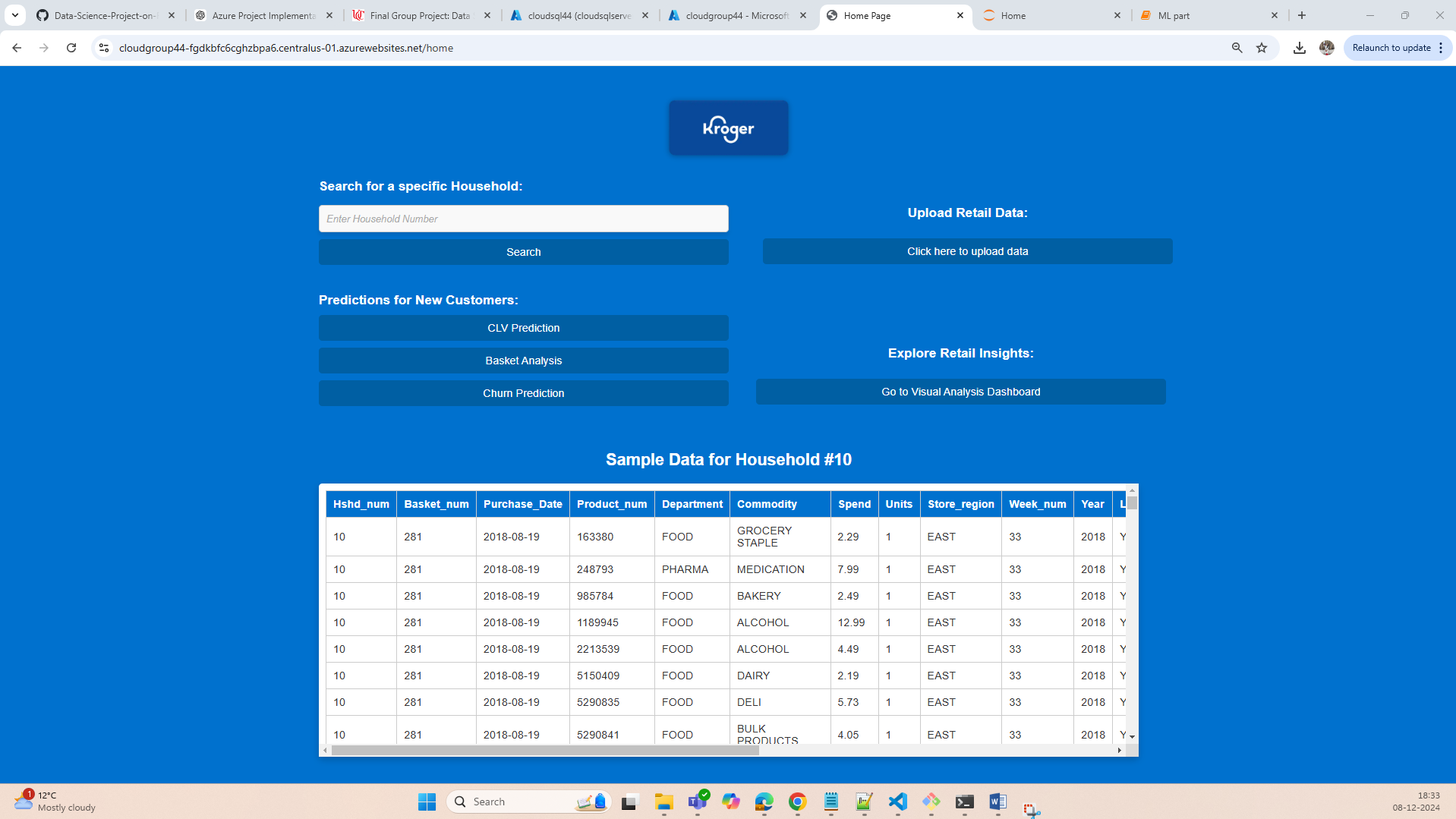
Folder Structure:



1. Created and pushed the flaskapp code to Azure WebApp using Local git and hosted the application in webapp.

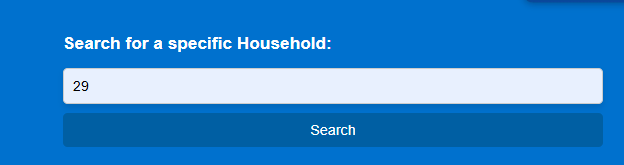
Azure WebApp and Deployment Logs:

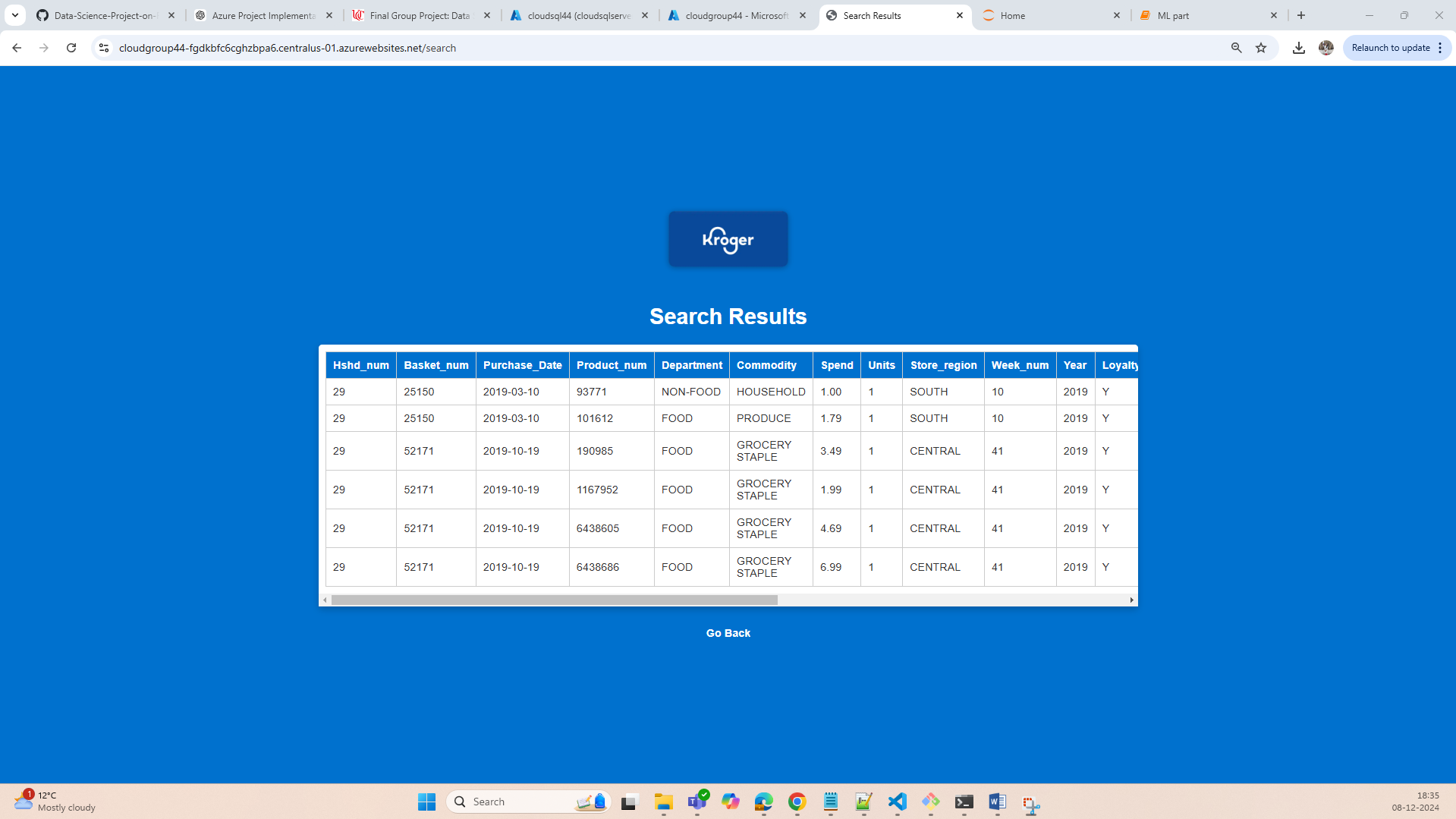


Azure Webapp Home Page and sample data pull for Household number = 10 and sorted by **Hshd\_num, Basket\_num, Date, Product\_num, Department, Commodity**:  


1. **Interactive Web Page:**

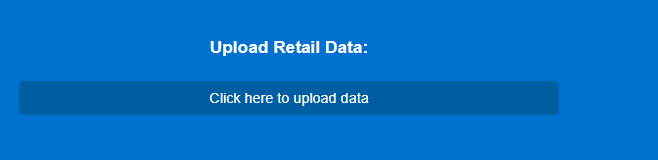
Data Pull for Specific Household 29:



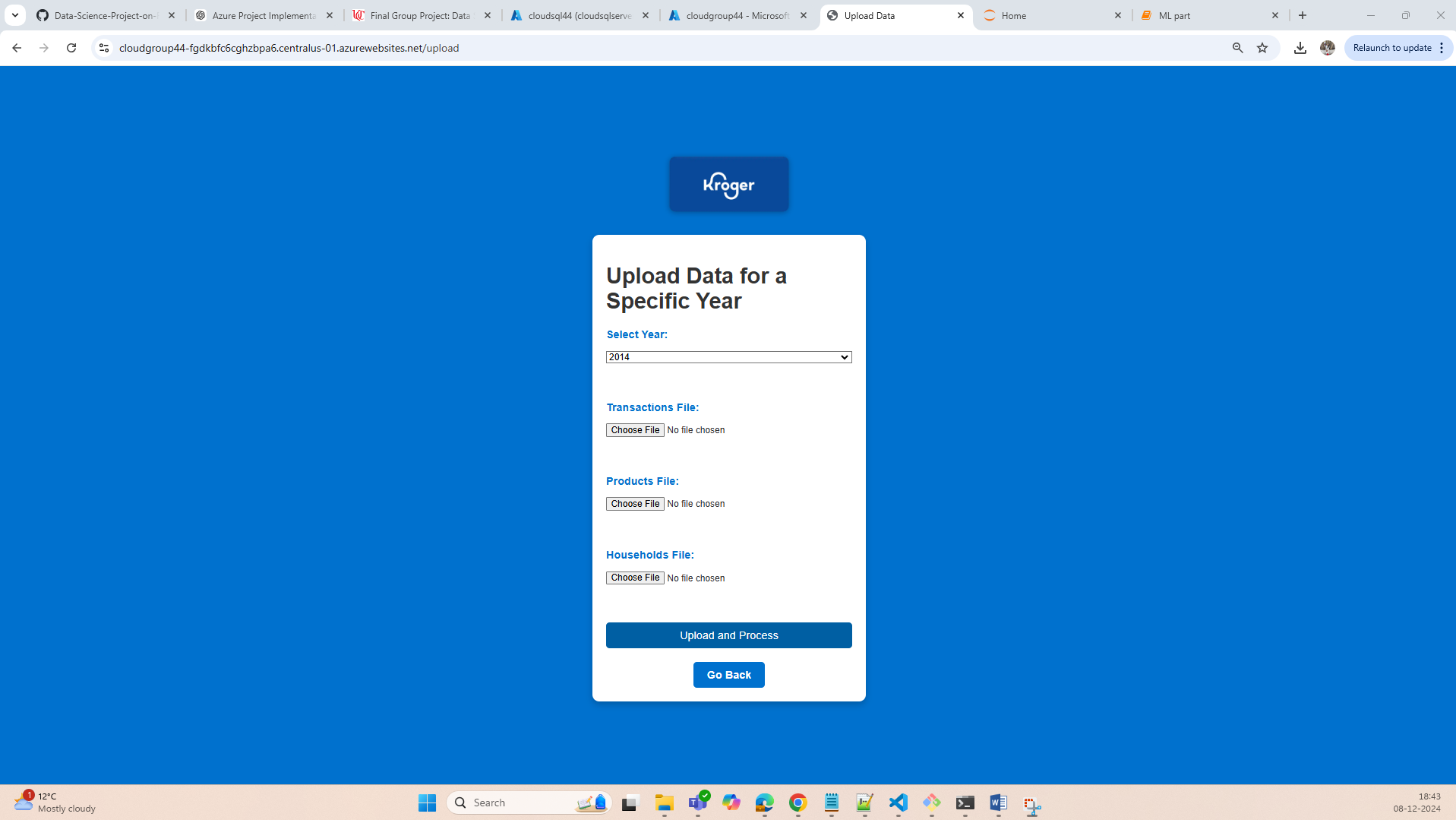


1. **Data Loading WebApp:**

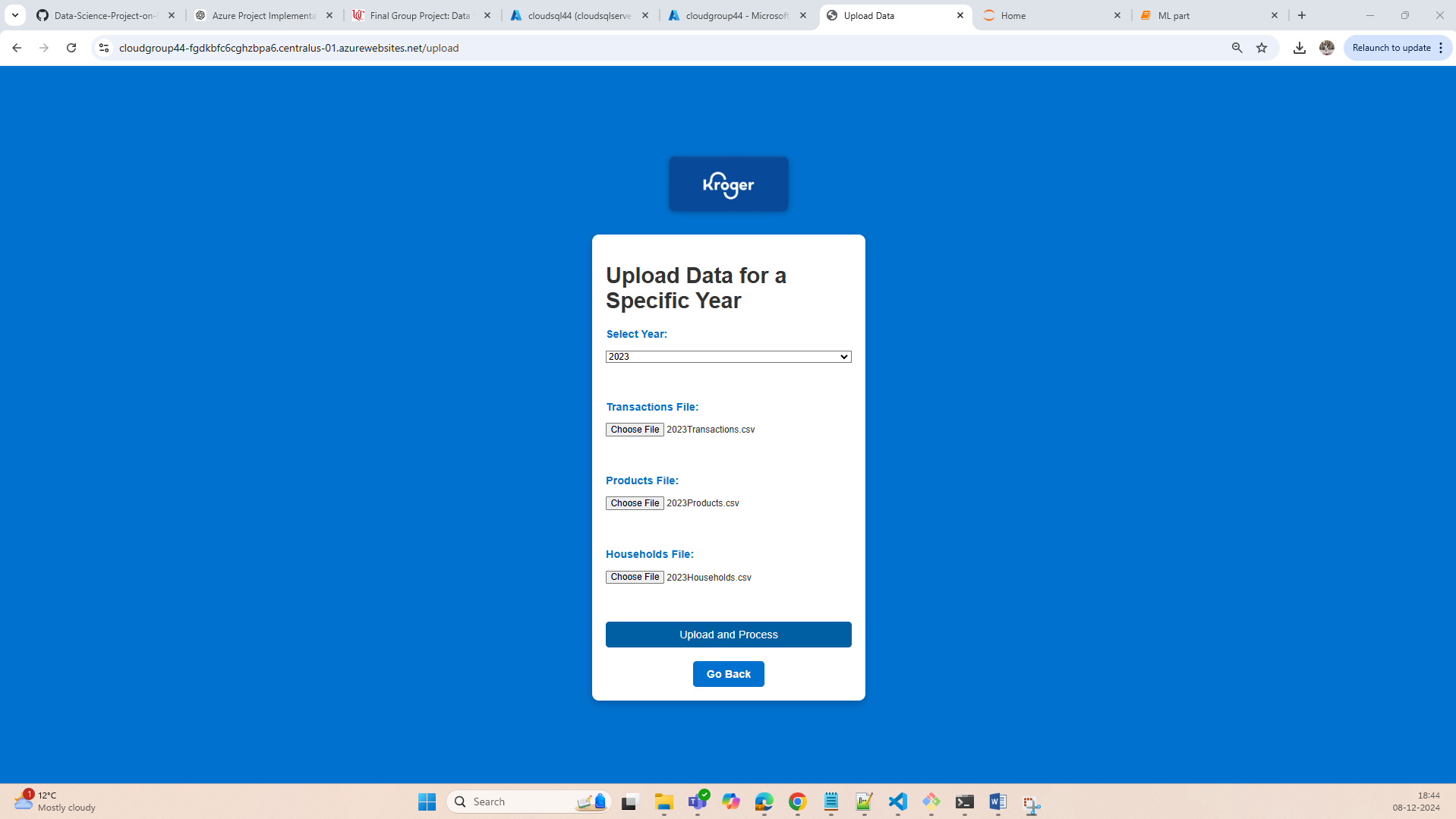
Below is the button from home page to navigate to upload data page:



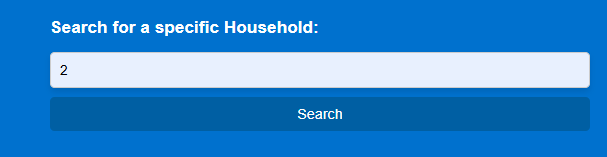
Upload data Page:

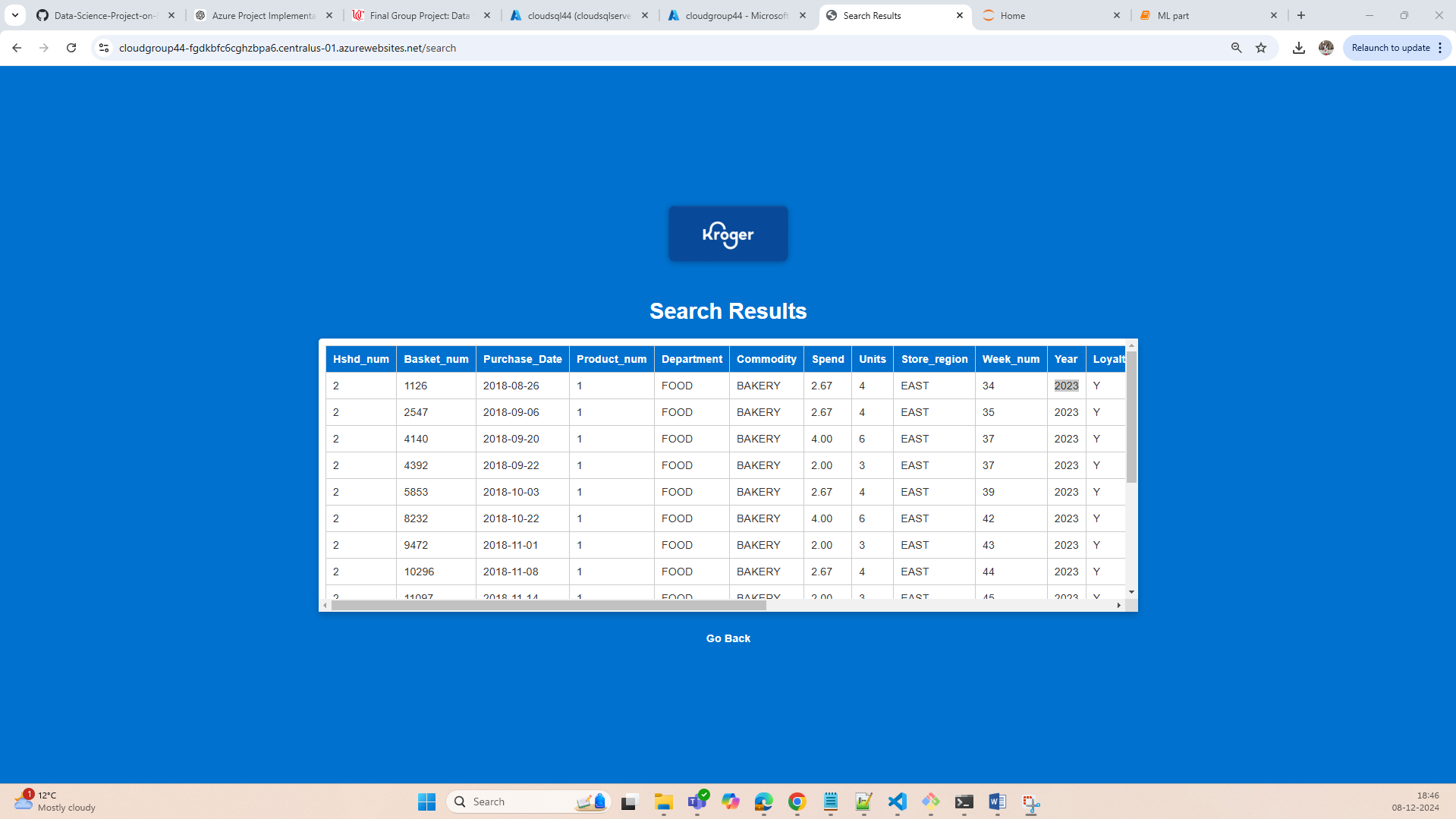


Uploaded sample data for Year 2023 for Household number 2:



Searching Uploaded data in Search bar:



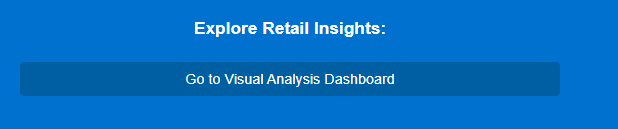


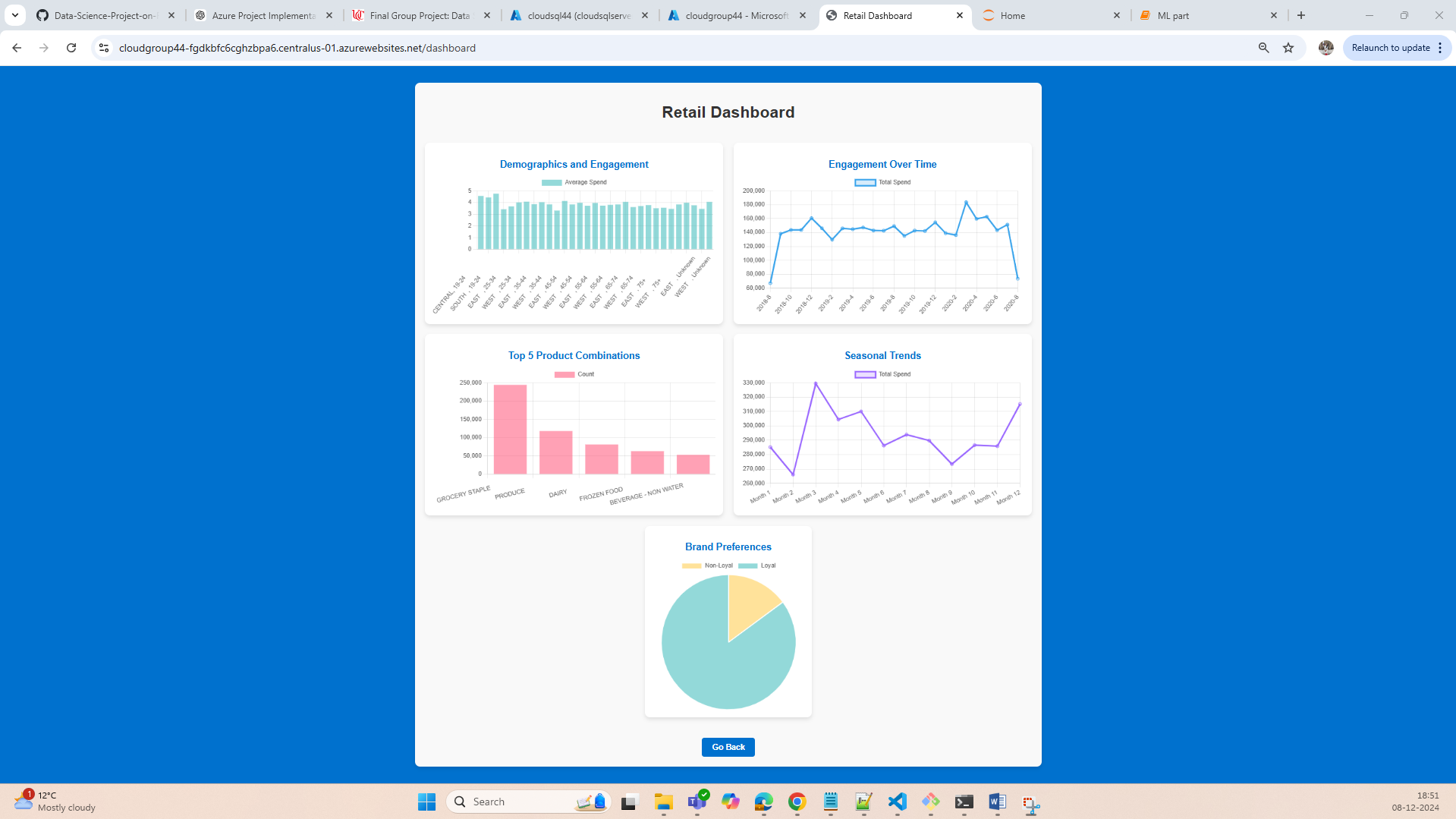
1. **Web Page with Dashboard:**

Retail Dashboard which answers possible Retail Questions like:

1. **Demographics and Engagement:**
2. **Engagement Over Time:**
3. **Basket Analysis:**
4. **Seasonal Trends:**
5. **Brand Preferences:**

Link from home page to navigate to Retail Dashboard:





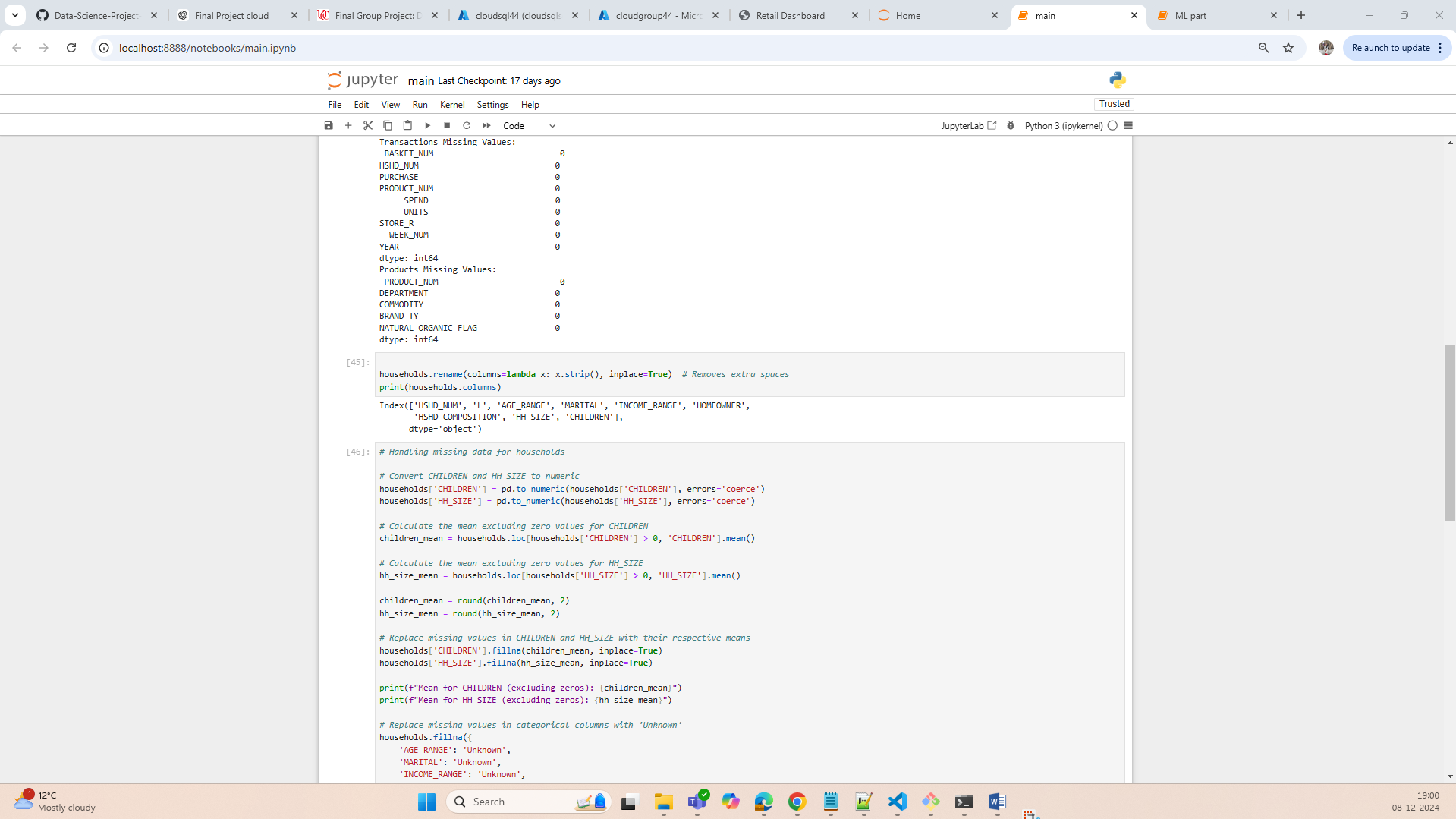
**Note: Since the flaskapp has to run 5 database queries with large data in database, it will take few seconds to load the visualization plots as above.**

**7. ML Model Application:**

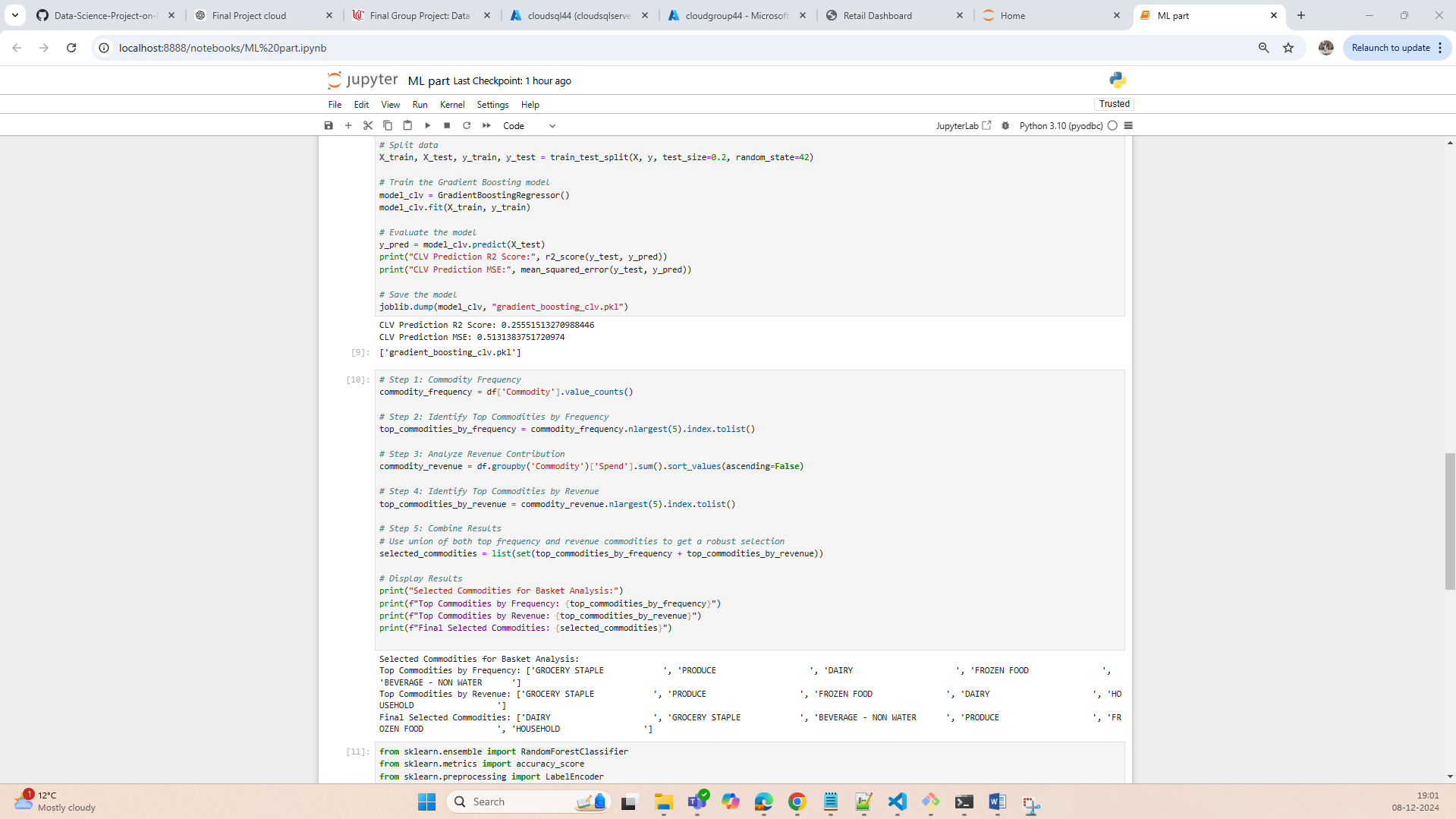
Procedure followed:

1. Cleaned the given 3 CSV files and saved the cleaned data.
2. Developed ML models for CLV Prediction, Basket Analysis and Churn Prediction and deployed the models in Flask application using .pkl files and predicted results for new customers.

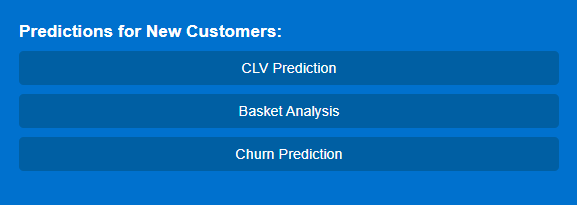
main.ipynb file to pre-processing the data:



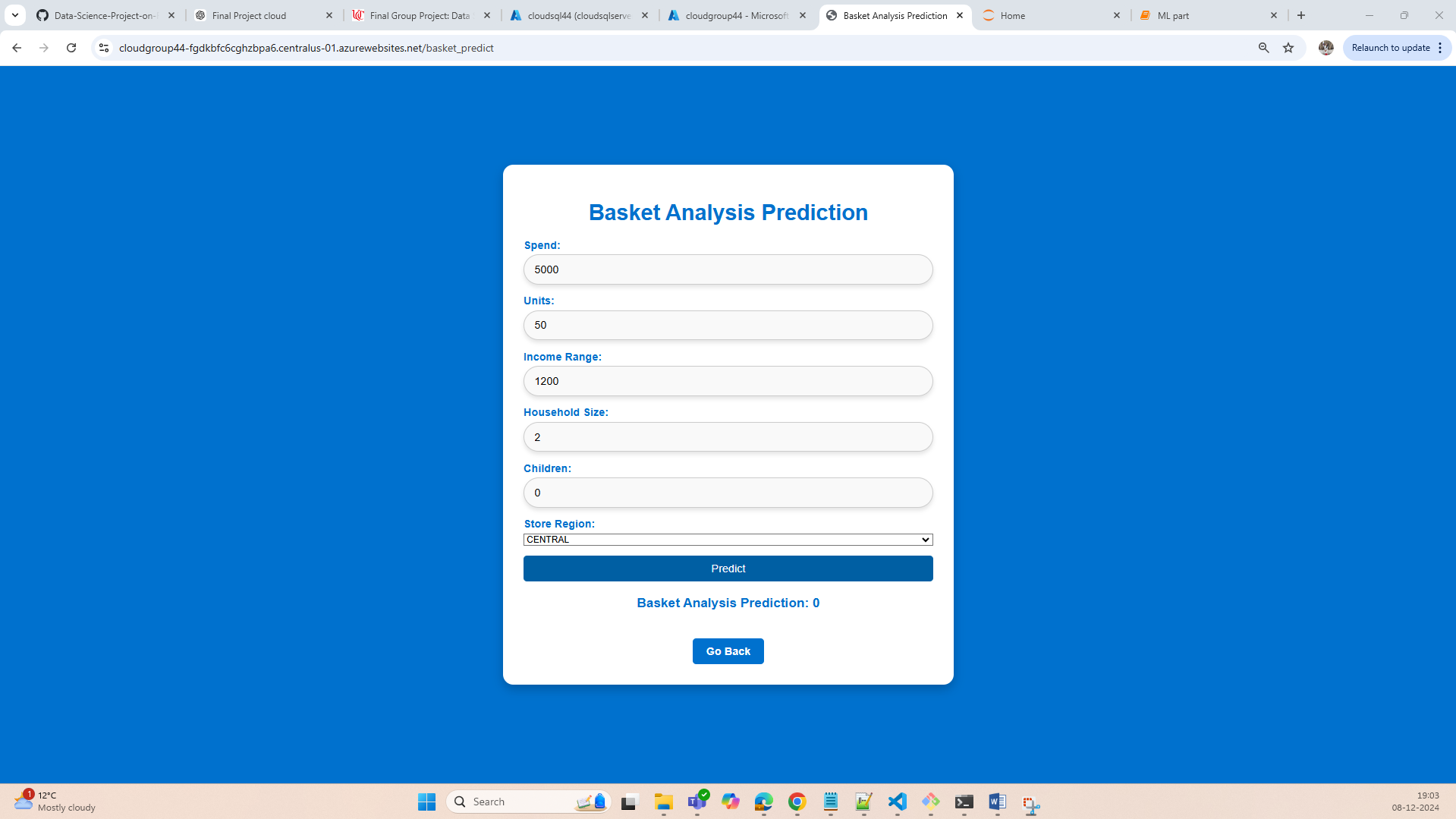
ML part.ipynb file to create and save ML models:



Link from Home page to navigate to respective prediction pages:



Basket Analysis prediction:



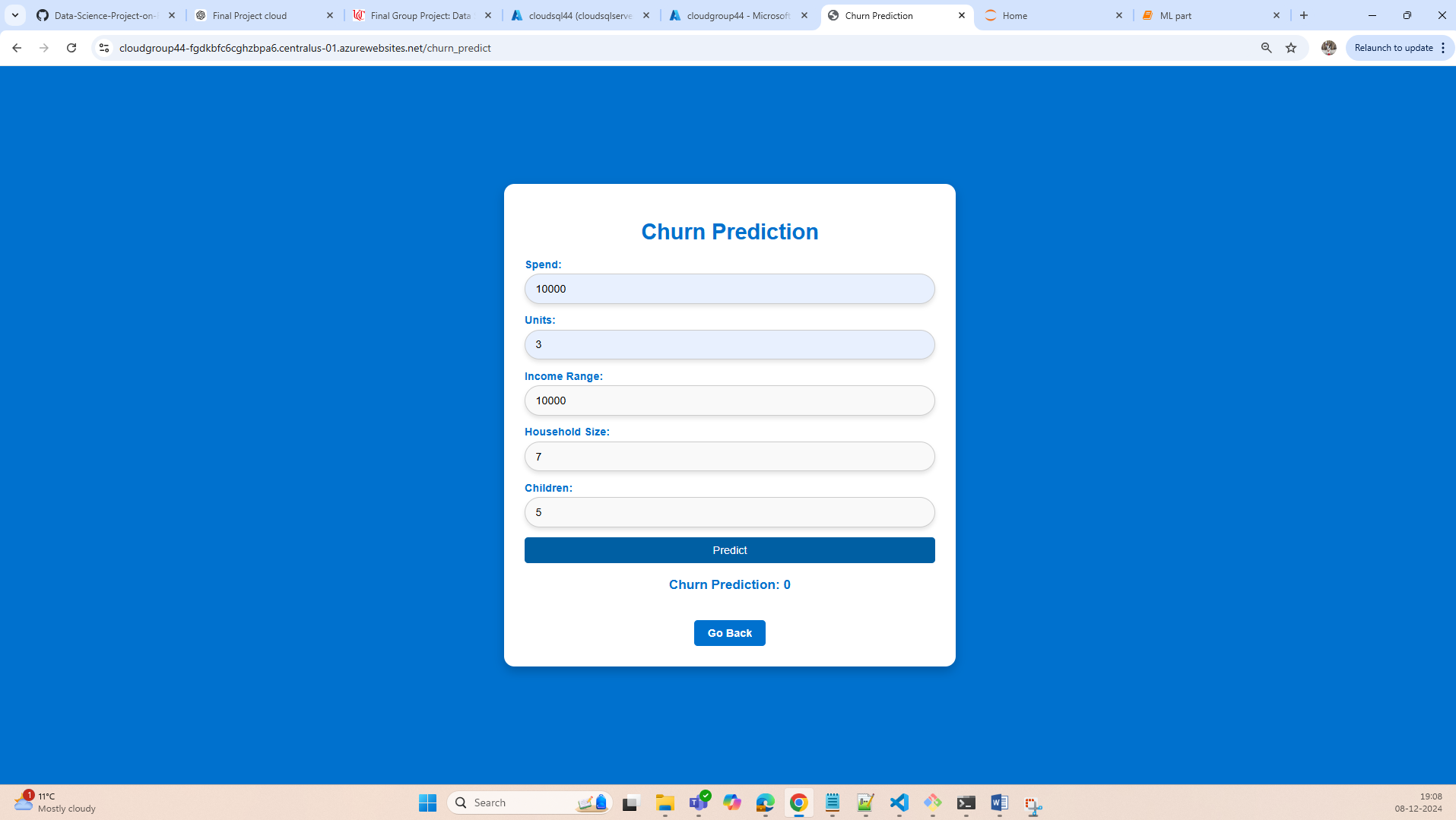
**Retail Question:** What are the commonly purchased product combinations, and how can they drive cross-selling opportunities?

To perform Basket Analysis, we utilized the **Random Forest** model due to its ability to handle high-dimensional data and capture complex relationships between features. Random Forest operates by constructing multiple decision trees and aggregating their results for classification or regression tasks, making it suitable for analysing customer purchasing behaviours.

Using this model, we identified key product combinations frequently purchased together, such as *FROZEN FOOD* and *DAIRY* or *GROCERY STAPLE* and *BEVERAGE - NON-WATER*. These insights can be leveraged to design targeted cross-selling strategies, such as offering discounts on complementary products or bundling items to encourage customers to increase their basket size.

By focusing on these combinations, retailers can drive revenue growth while enhancing customer satisfaction by meeting their needs effectively.

**8. Churn Prediction:**

****

Retail Question: Which customers are at risk of disengaging, and how can retention strategies address this?

To identify customers at risk of churn, we employed Logistic Regression, a robust and interpretable model for binary classification tasks. The model analysed key features such as spend, purchase frequency, household size, and loyalty flag to predict the likelihood of disengagement.

The analysis revealed patterns indicating high churn risk, such as reduced spending or infrequent purchases. Using correlation analysis, we observed that customers with lower loyalty scores or smaller basket sizes were more likely to disengage. Graphical representations of churn probabilities further supported these insights.

To address churn, targeted retention strategies can be implemented, such as personalized promotions, loyalty programs, and proactive communication to re-engage at-risk customers. These strategies can improve customer satisfaction, reduce churn rates, and enhance long-term revenue potential.