Project Report

Course Name	**Data Science essentials**
Product Name (Marketing & Sales)	PDDS
Module Name	**Data Science Essentials**
Product Name (Marketing & Sales)	Data Science Essentials

Student name	**Assessor name**
Gangeswaran Rajavalarmathi	RyanSuryanto

Date issued	**Completion date**	**Submitted on**
12-09-2023	18-10-2023	15-10-2023

Project title	**Student Registration Form Development**

Learner declaration
I certify that the work submitted for this assignment is my own and research sources are fully acknowledged. Student signature: Date:

Content

Project Overview: Describe the Project with summary of analytical processes and project outcomes
(Explain the Project in your own words in 15 – 20 lines)

Data Preparation: The project starts with data preparation, involving tasks like data cleansing to ensure data quality and data exploration to understand the dataset's structure and features. This phase sets the foundation for subsequent analytical tasks.

Feature Identification: The project aims to identify which features within the dataset may serve as predictors of flight delays or early arrivals. This step involves statistical analysis to select relevant variables.

Model Development: Machine learning techniques are employed to build predictive models. These models are trained on the prepared data, using algorithms and feature engineering to predict flight delays and early arrivals.

Model Evaluation: The performance of the trained models is evaluated using metrics such as accuracy, precision, recall, and mean absolute error. This step ensures that the models are effective in their predictions.

Model Deployment: The most successful model is deployed as a web service, enabling real-time predictions for flight delays. This allows the model to be integrated into applications for practical use.

1. **Project Technical Environment:** (Describe the Microsoft Azure Machine Learning Platform)

Microsoft Azure Machine Learning (Azure ML) platform is a cloud-based service provided by Microsoft that offers a comprehensive set of tools and services for building, training, deploying, and managing machine learning models. It is designed to streamline the entire machine learning lifecycle, making it easier for data scientists and developers to work on projects ranging from simple data analysis to complex machine learning tasks. Here's a description of the key features and components of the Azure ML platform

1. **Workspace:** Azure ML provides a dedicated workspace where you can manage all aspects of your machine learning projects. It serves as a central hub for organizing data, experiments, models, and resources.
2. **Notebooks:** You can create and run Jupyter notebooks directly within the Azure ML environment. This allows for interactive data exploration and model development.
3. **Automated Machine Learning (AutoML):** AutoML is a powerful feature that automates much of the machine learning process, including feature engineering, algorithm selection, and hyperparameter tuning. It makes machine learning more accessible to those without deep expertise in the field.
4. **Data Preparation:** Azure ML includes tools for data preparation, cleansing, and transformation. You can explore data, handle missing values, and create data pipelines for seamless integration into your machine learning workflow.
5. **Model Training and Experimentation:** You can build and train machine learning models using various algorithms and libraries. Azure ML offers version control for models and code, making it easy to track changes and collaborate with team members.
6. **Model Management:** The platform allows you to package and deploy machine learning models as web services, making them accessible for real-time predictions and integration into applications.
7. **Scalability:** Azure ML provides scalable computing resources, enabling you to train models on large datasets and handle high computational loads.
8. **Integration:** Azure ML seamlessly integrates with other Microsoft Azure services, such as Azure Databricks, Azure Synapse Analytics, and Azure DevOps, allowing for end-to-end data and AI solutions.

9. **Collaboration:** Teams can collaborate on projects in a collaborative, version-controlled environment, with the ability to share code, data, and experiments.
10. **Monitoring and Model Management:** Azure ML offers tools for monitoring model performance and managing deployed models. You can track model drift and retrain models as needed.
11. **Security and Compliance:** The platform includes security features such as data encryption, identity and access management, and compliance with industry standards and regulations.

Benefits of Azure ML:

1. **Scalability:** Azure ML can handle projects of any size, from small-scale experiments to large-scale machine learning initiatives.
2. **Ease of Use:** The platform offers a user-friendly interface and supports various programming languages, making it accessible to data scientists and developers with different skill levels.
3. **Integration:** Integration with other Azure services simplifies the creation of end-to-end data and AI solutions.
4. **AutoML:** Automated machine learning accelerates model development and makes it accessible to a broader audience.
5. **Collaboration:** Azure ML promotes collaboration and version control within a team, streamlining the development process.
6. **Cost-Effective:** You pay for the resources you use, making it cost-effective and suitable for projects with varying resource demands.

2. **Analytical Technique & Tools used :** Describe the Analytical Technique used in the Project

1. **Descriptive Statistics:**
    - Description: Descriptive statistics are used to summarize and describe the main features of a dataset. This includes measures like mean, median, standard deviation, and quartiles.
    - Application: Descriptive statistics are often used for initial data exploration and to understand the distribution of data.
2. **Inferential Statistics:**
    - Description: Inferential statistics involve making predictions or inferences about a population based on a sample of data. This includes hypothesis testing, regression analysis, and confidence intervals.
    - Application: Inferential statistics help in making predictions and drawing conclusions from data, such as testing hypotheses or estimating population parameters.
3. **Machine Learning:**
    - Description: Machine learning encompasses a wide range of algorithms and techniques that allow computers to learn from data and make predictions or decisions without being explicitly programmed.
    - Application: Machine learning is used for tasks such as classification, regression, clustering, and recommendation systems. Common algorithms include decision trees, linear regression, support vector machines, and deep learning neural networks.
4. **Clustering:**
    - Description: Clustering techniques group similar data points into clusters based on similarity or distance measures.
    - Application: Clustering is used for segmentation, customer profiling, and anomaly detection. K-Means, DBSCAN, and hierarchical clustering are commonly used algorithms.
5. **Natural Language Processing (NLP):**
    - Description: NLP techniques are used to analyze and process text data, including sentiment analysis, topic modeling, and language understanding.
    - Application: NLP is applied in tasks like text classification, chatbots, and language translation. Tools like NLTK, spaCy, and GPT models can be used.

Tools:

1. **Python:**
    - Description: Python is a popular programming language with a wide range of libraries and frameworks for data analysis and machine learning, such as NumPy, Pandas, Scikit-Learn, and TensorFlow.
2. **R:**
    - Description: R is a programming language and environment specifically designed for statistical analysis and data visualization. It has a rich ecosystem of packages like ggplot2 and dplyr.
3. **Jupyter Notebook:**
    - Description: Jupyter Notebook is an interactive environment for writing and executing code. It's commonly used for data exploration and sharing results.
4. **Tableau:**
    - Description: Tableau is a data visualization tool that allows for the creation of interactive and shareable dashboards.
5. **Microsoft Azure Machine Learning:**
    - Description: Azure ML is a cloud-based platform for building, training, and deploying machine learning models.
6. **SQL (Structured Query Language):**
    - Description: SQL is used for data manipulation and querying in relational databases.
7. **Excel:**
    - Description: Excel is a widely used tool for data analysis, especially for small-scale and basic data processing tasks.

3. Data Science Project Team – Roles and Responsibilities Table

Role	Responsibilities
Data Scientist	- Data analysis and modeling
	- Building and training machine learning models
	- Feature selection and engineering
	- Data visualization and interpretation
Data Engineer	- Data collection and storage setup
	- Designing and maintaining data pipelines
	- Data preprocessing and cleansing
	- Ensuring data accessibility and availability
Data Analyst	- Data exploration and visualization

Role	Responsibilities
	- Creating reports and dashboards for stakeholders
	- Identifying data patterns and trends
	- Collaborating with domain experts for context
Machine Learning Engineer	- Model deployment and integration
	- Model optimization and scaling
	- Collaborating with data scientists on model development
Domain Expert	- Providing subject matter expertise
	- Offering domain-specific insights for analysis
	- Collaborating with data scientists for context
Business Analyst	- Defining project goals and success criteria

Role	Responsibilities
	- Requirements gathering and documentation
	- Ensuring alignment with business objectives
Data Quality Analyst	- Data validation and quality assurance
	- Data cleansing and error detection
	- Ensuring data accuracy, completeness, and consistency
Project Manager	- Project planning and resource management
	- Coordination of the team's efforts
	- Communicating with stakeholders and clients
	- Ensuring project timelines and budgets are met

4. **Activity 1: Activity Summary**

5. **Data Scientist:**
    a. Data scientists are responsible for data analysis, model building, and predictive analytics. They use statistical techniques and machine learning algorithms to extract insights from data and build predictive models.

6. **Data Engineer:**
    a. Data engineers are responsible for data collection, storage, and retrieval. They design and maintain data pipelines, ensuring that data is accessible and in the right format for analysis.

7. **Data Analyst:**
    a. Data analysts focus on exploring and visualizing data. They provide descriptive insights, create reports, and support decision-making by presenting data in a comprehensible format.

8. **Machine Learning Engineer:**
    a. Machine learning engineers specialize in deploying machine learning models into production. They work on model optimization, scaling, and integration with software systems.

9. **Domain Expert:**
    a. A domain expert has subject matter knowledge relevant to the organization's industry. They collaborate with data scientists to provide context and domain-specific insights for more accurate analysis.

10. **Business Analyst:**
    a. Business analysts bridge the gap between the data team and the business stakeholders. They define project goals, requirements, and success criteria, ensuring that data science projects align with business objectives.

11. **Data Quality Analyst:**
    a. Data quality analysts focus on data cleansing and validation, ensuring that the data used for analysis is accurate, complete, and consistent.

12. Project Manager:

a. Project managers oversee data science projects, ensuring they are delivered on time and within budget. They coordinate the efforts of the team, manage resources, and communicate with stakeholders.

Task 2: Review of Data Cleansing, Exploration, and Machine Learning Activities

Data Cleansing: In the data cleansing activity, the team identified and corrected errors, inconsistencies, and inaccuracies in the dataset. This process involved removing duplicate records, filling in missing values, and addressing outliers, resulting in clean and reliable data for analysis.

Data Exploration: During data exploration, the team visualized and summarized the data to gain a better understanding of its characteristics. This step included generating descriptive statistics, creating data visualizations, and identifying patterns or trends in the data.

Machine Learning: In the machine learning activity, the team developed and trained predictive models using the cleaned data. The models were evaluated for their accuracy and performance, and the best-performing model was selected for deployment.

Task 3: Report on Output of Each Activity

Data Cleansing: The data cleansing activity successfully cleaned the dataset, resulting in improved data quality. This step enhanced the accuracy and reliability of the data, which is crucial for downstream analysis and model development.

Data Exploration: Data exploration provided valuable insights into the dataset. Key patterns and trends were identified, helping the team make informed decisions about the variables and features to include in the machine learning models.

Machine Learning: The machine learning activity resulted in the development of predictive models that can make data-driven predictions. The best-performing model, based on evaluation metrics, was selected for deployment, potentially providing actionable insights for the organization.

13. Activity 2: Flight DataSet- An Overview

The flight Data set consist of below parameters:

Year

Month

Carrier

OriginAirportID

DestAirportID

CRSDepTime

CRSArrTime

Month

DayofMonth

DayOfWeek

Carrier

OriginAirportID

DestAirportID

CRSDepTime

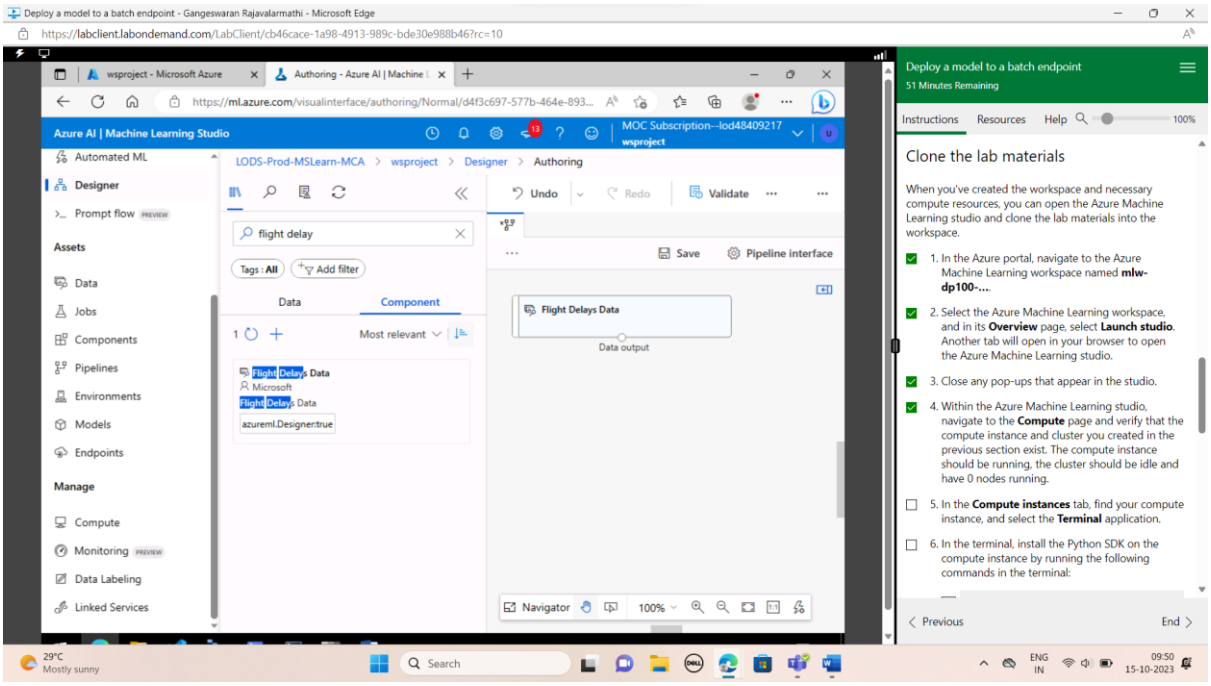
DepDelay

CRSArrTime

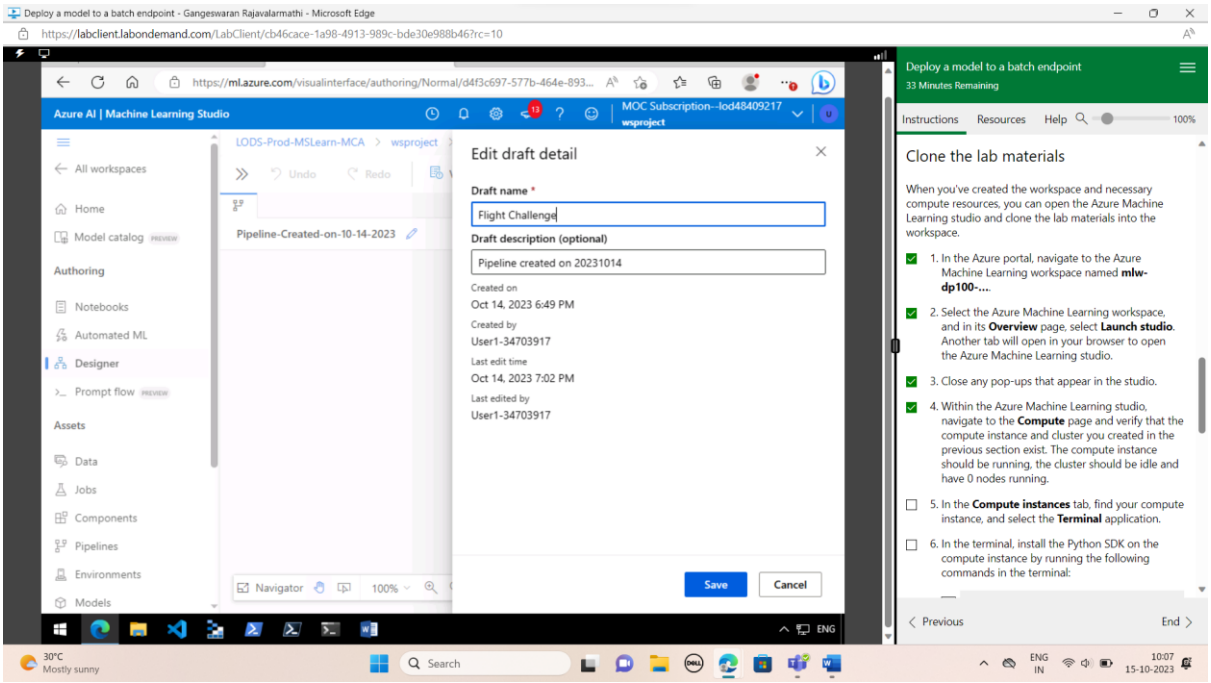
ArrDelar

14. Screen-shots of each task of Activity 3

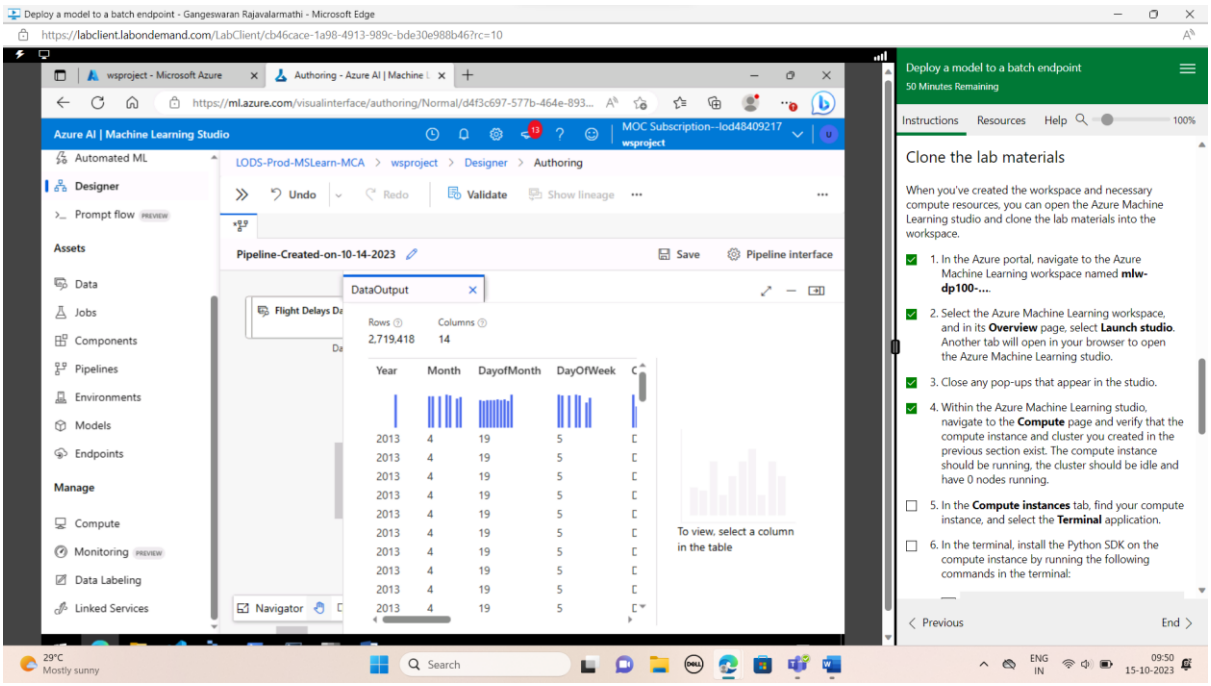
Task 1: Sign in to Azure Machine Learning Workspace



Task 2: Create a new experiment, with an appropriate name like "Flights Challenge



Task 3: Add the Flights Delay Data sample dataset to the experiment, and then visualize its contents

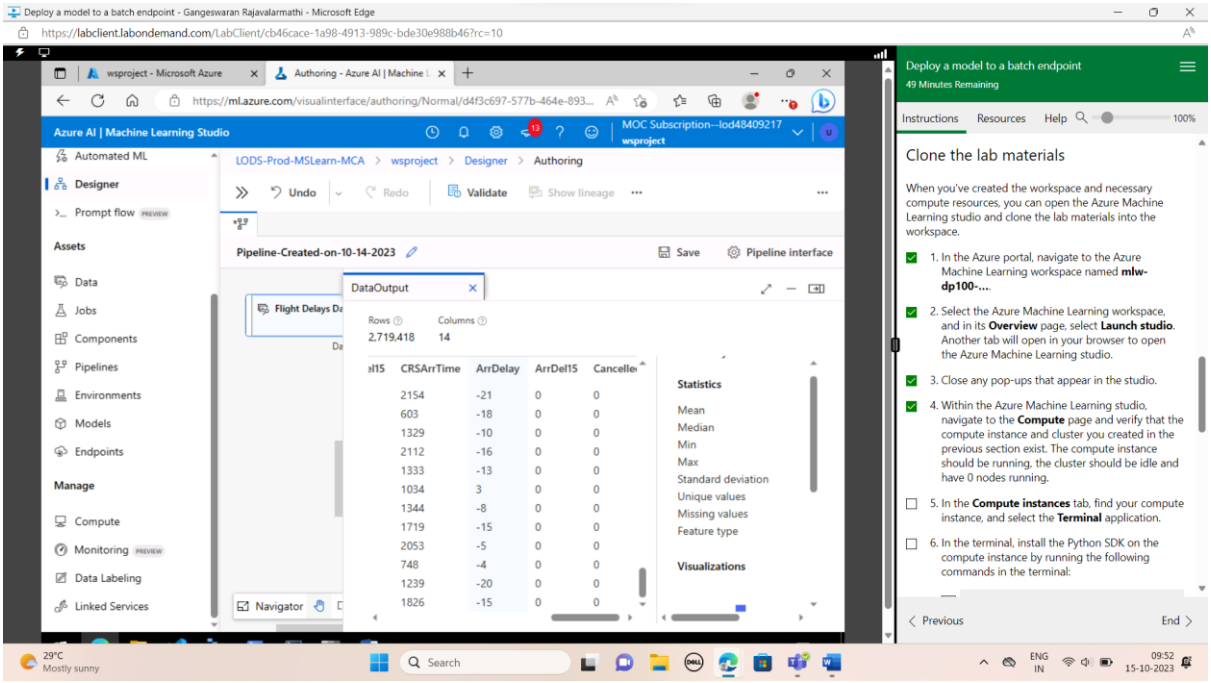


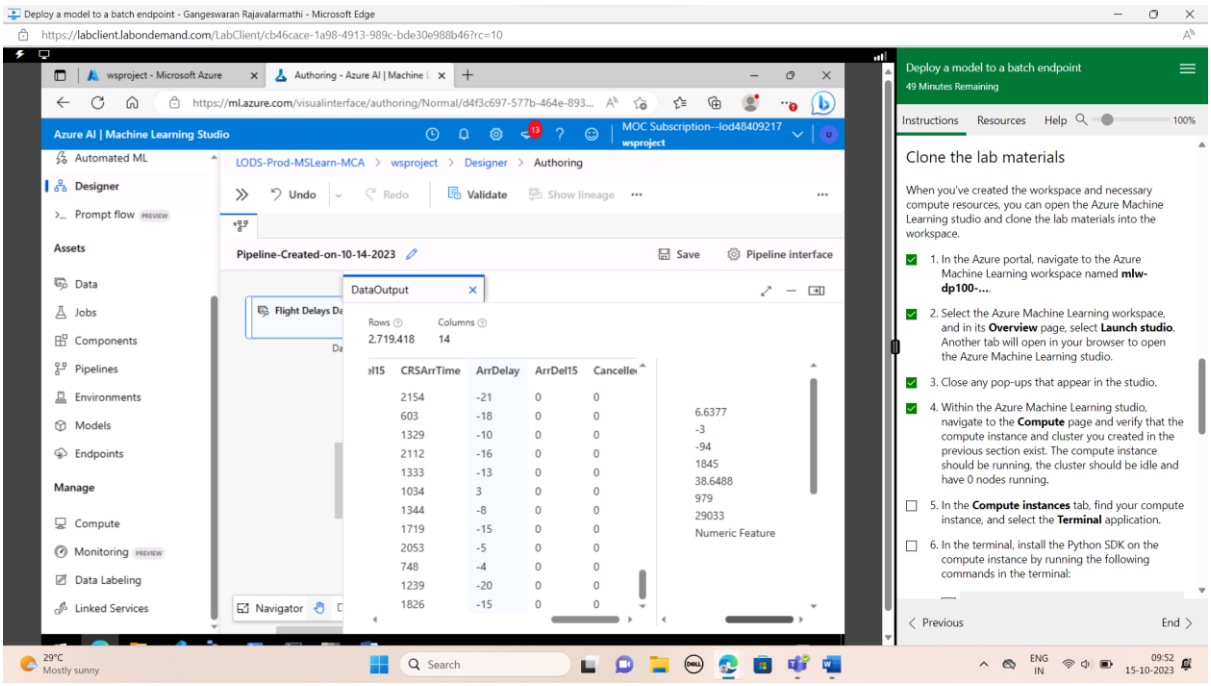
Task 4: Answer the following questions: -

How many rows are in the dataset?2,719,418

- Whatis the mean value of the ArrDelay column? 6.6377

Include screen shots of each task of Activity 3 in the Project Report

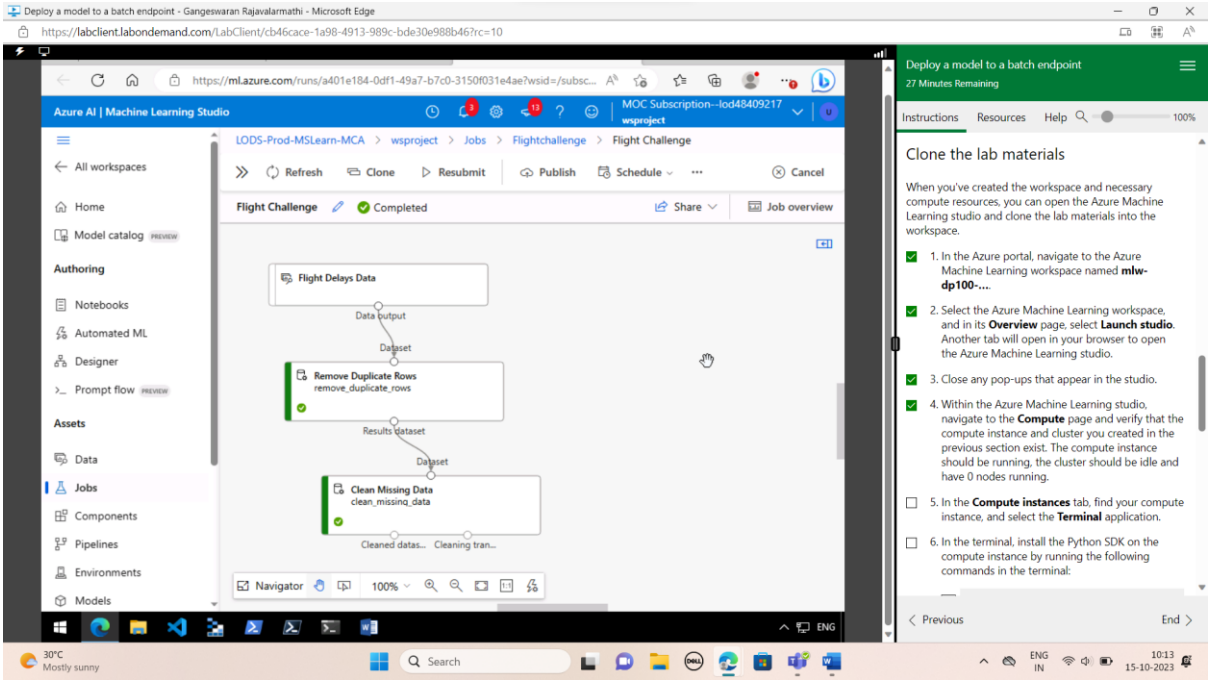




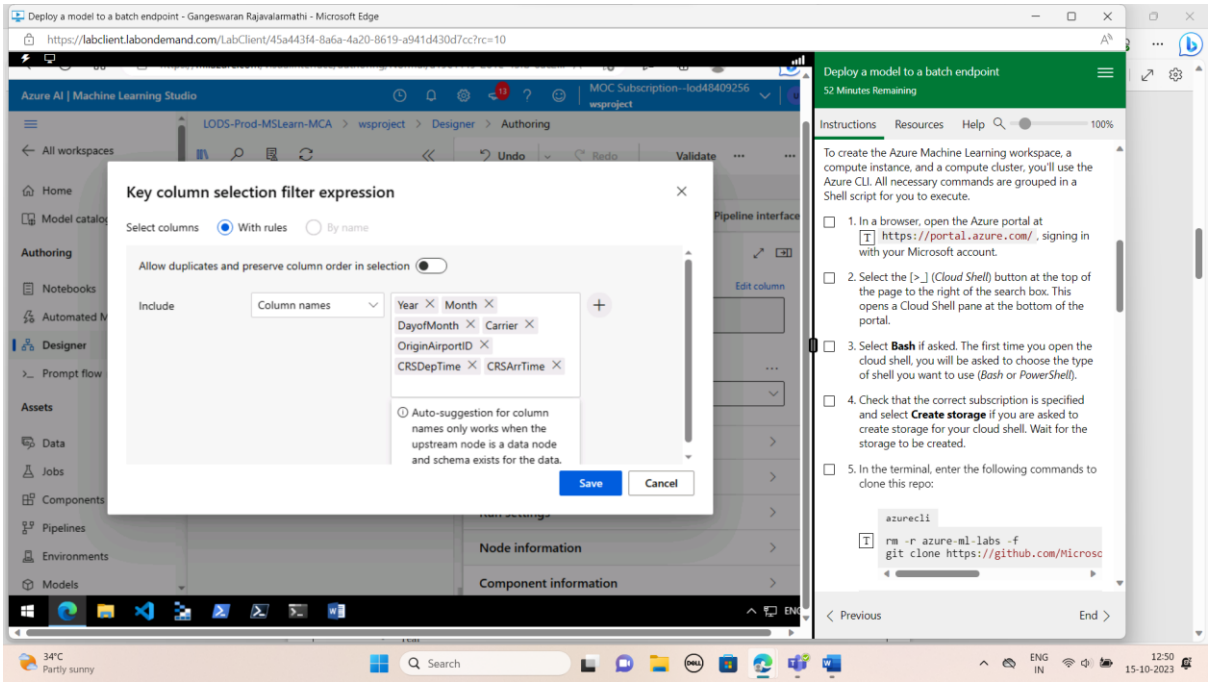
15. Screen-shots of each task of Activity 4

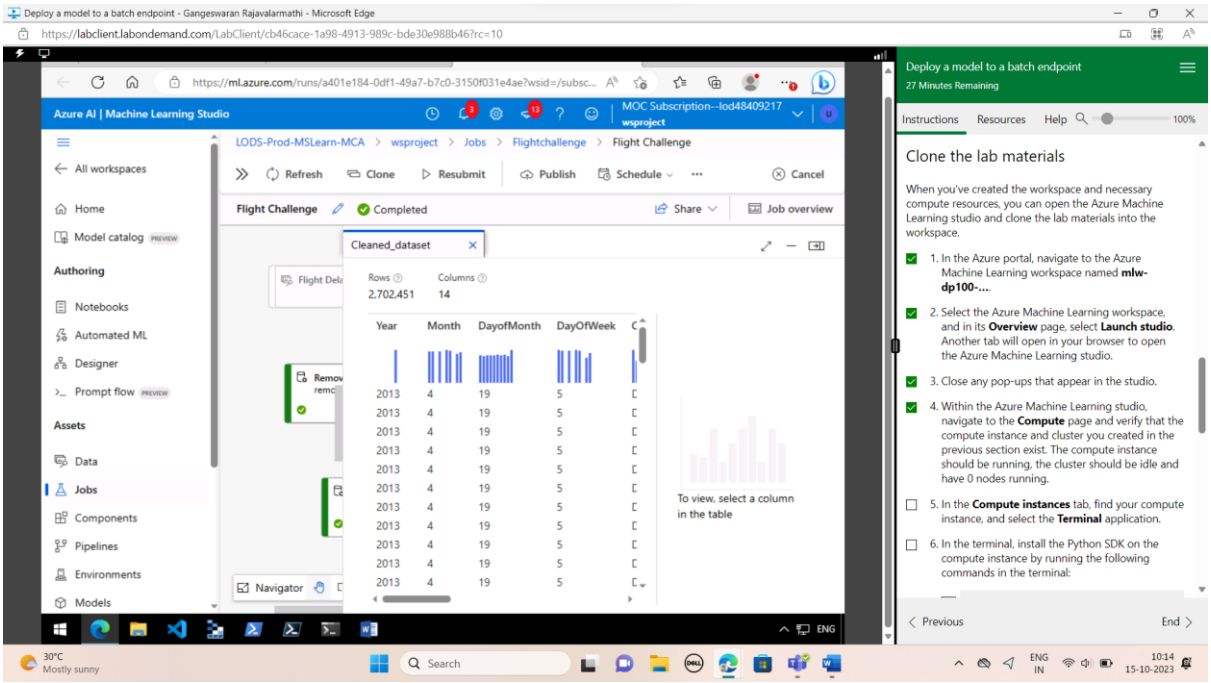
Task 1: Remove duplicate rows (retaining the first instance of each row).

Rows are considered duplicates in this dataset if they have matching values for all the following fields: - Year - Month - DayofMonth - Carrier - Origin Airport ID - Dest Airport ID - CRS Dep Time - CRS Arr Time Use the built-in Azure Machine Learning module



Task 2: After removing the duplicate rows, replace missing values in the DepDelay and ArrDelay columns with the value 0 (zero). Use the built-in Azure Machine Learning Remove Duplicate Rows module,

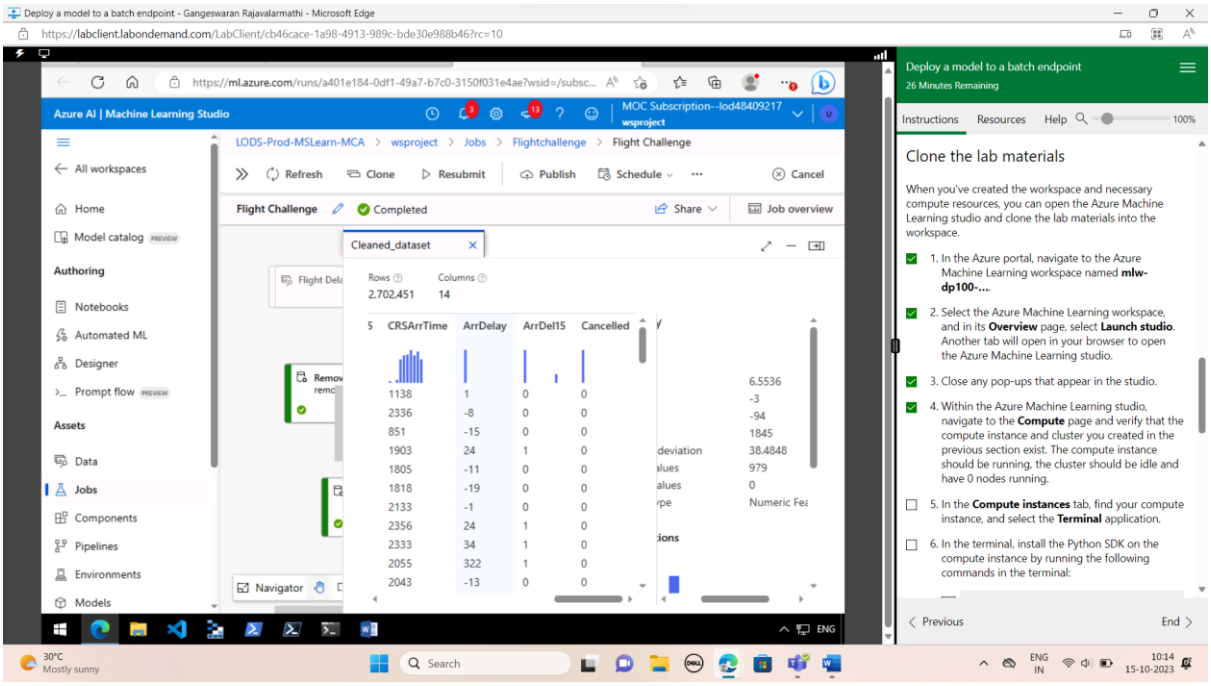




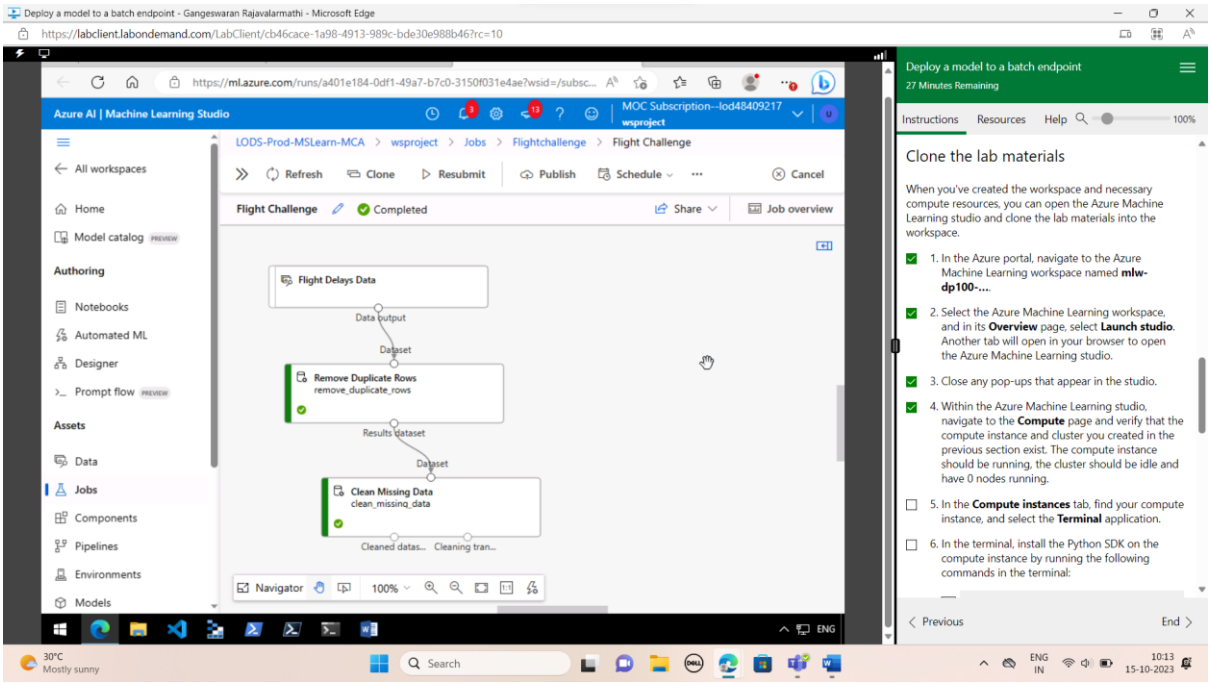
Task 3: Answer the following questions, after you have removed duplicate rows and replaced missing values, -

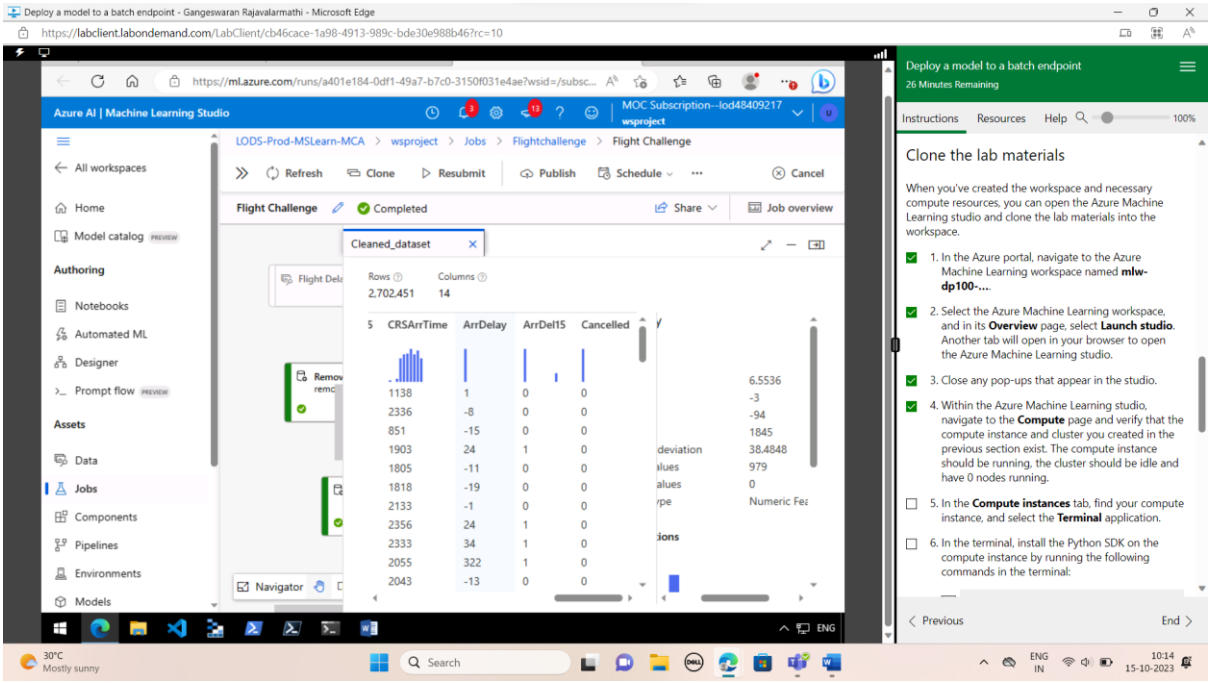
How many rows remain in the dataset? 2,702.451

- What is the mean value of the ArrDelay column? 6.5536



Include screen shots of each task of Activity 4 in the Project Report. After completing the Data Cleanup, perform activities and tasks related to Data Exploration.

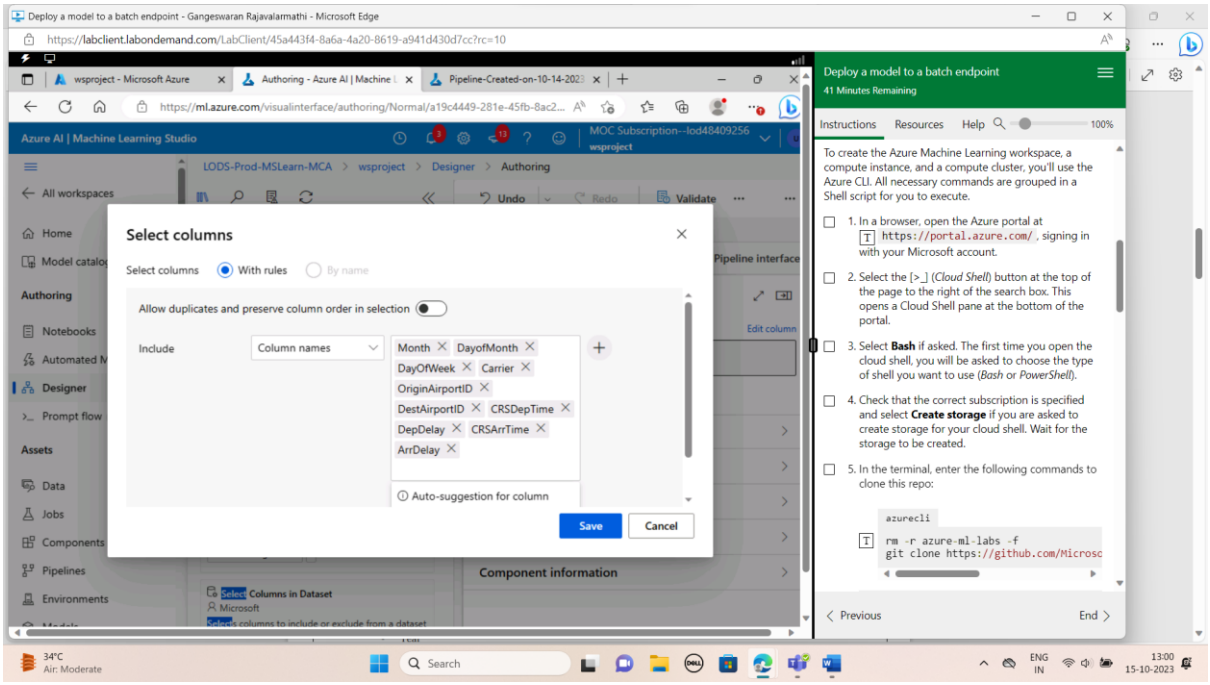




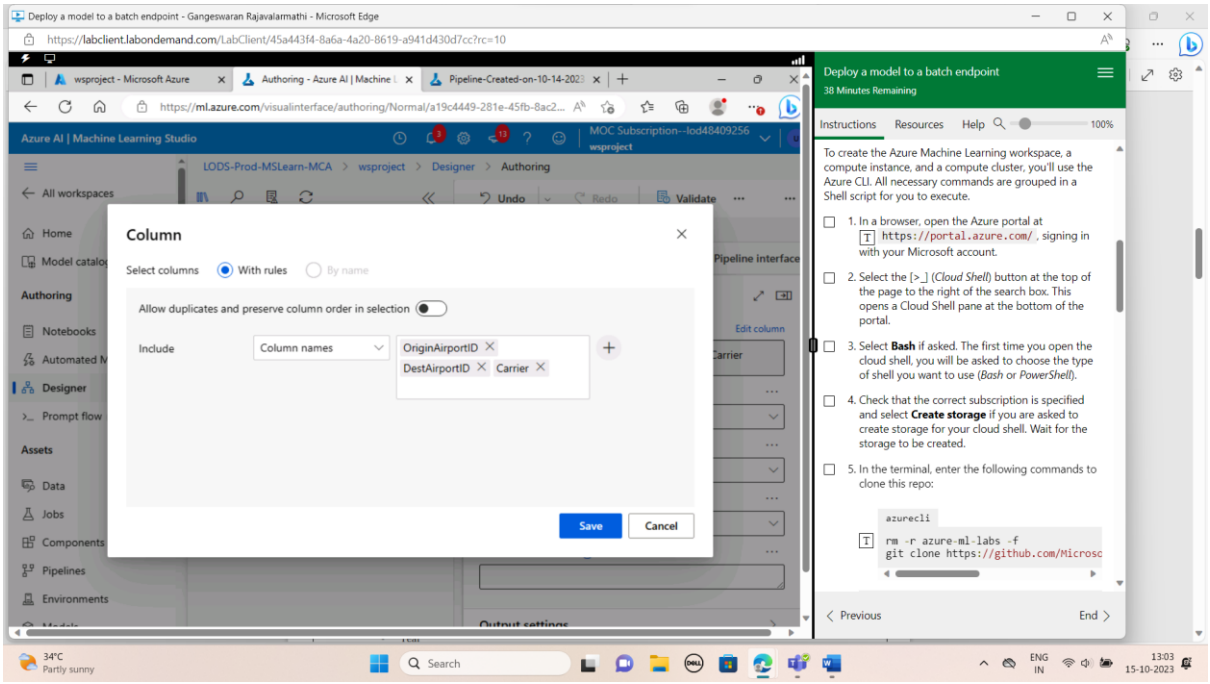
Activity 6: Train a Regression Model To predict a numeric value, such as the number of minutes delayed or early a flight arrives, train a regression model. Perform the following tasks to train a regression model: -

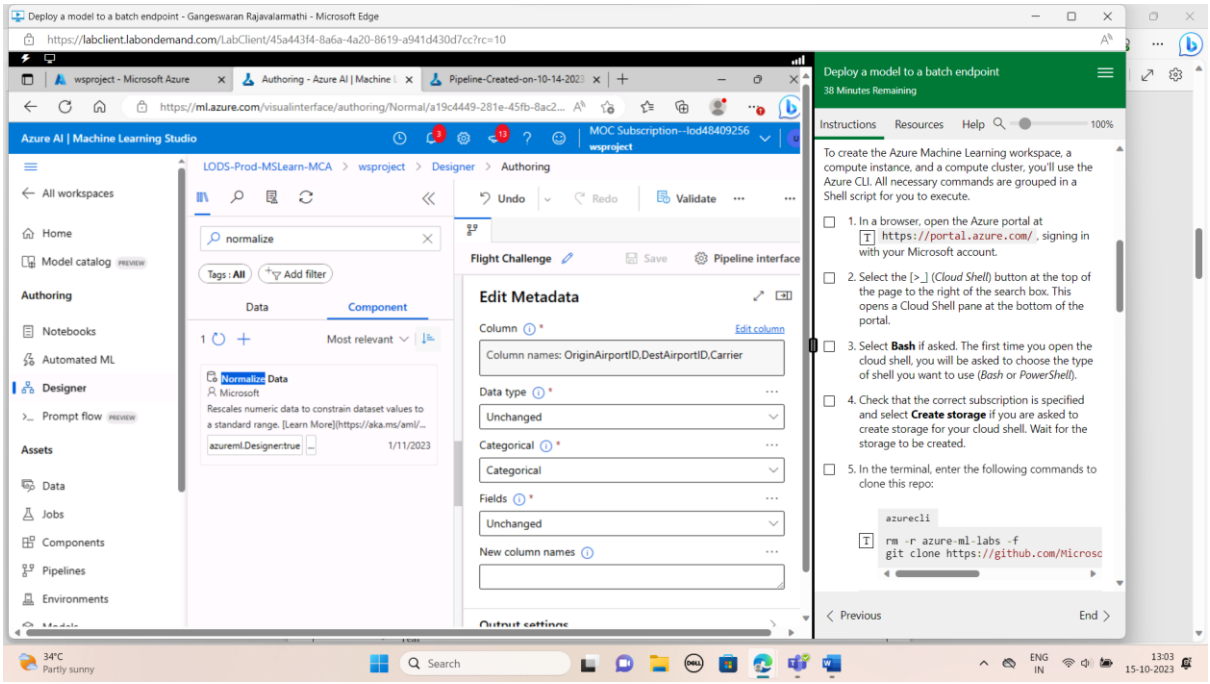
Task 1: Return to the Azure Machine Learning experiment you created in Part 1. -

Task 2: Add a Select Columns in Dataset module, and use it to select only the Month, DayofMonth, DayOfWeek, Carrier, OriginAirportID, DestAirportID, CRSDepTime, DepDelay, CRSArrTime, andArrDelay columns.

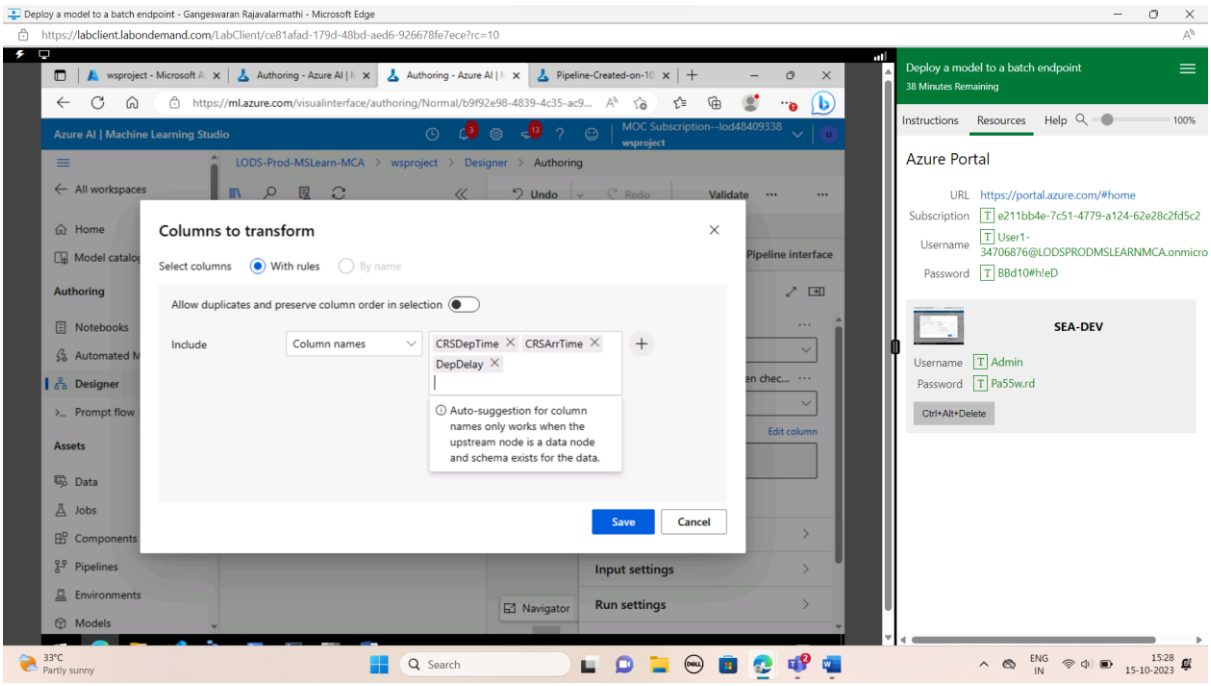


- **Task 3: Add an Edit Metadata module and use it to make the OriginAirportID, DestAirportID, and Carrier columns Categorical.**

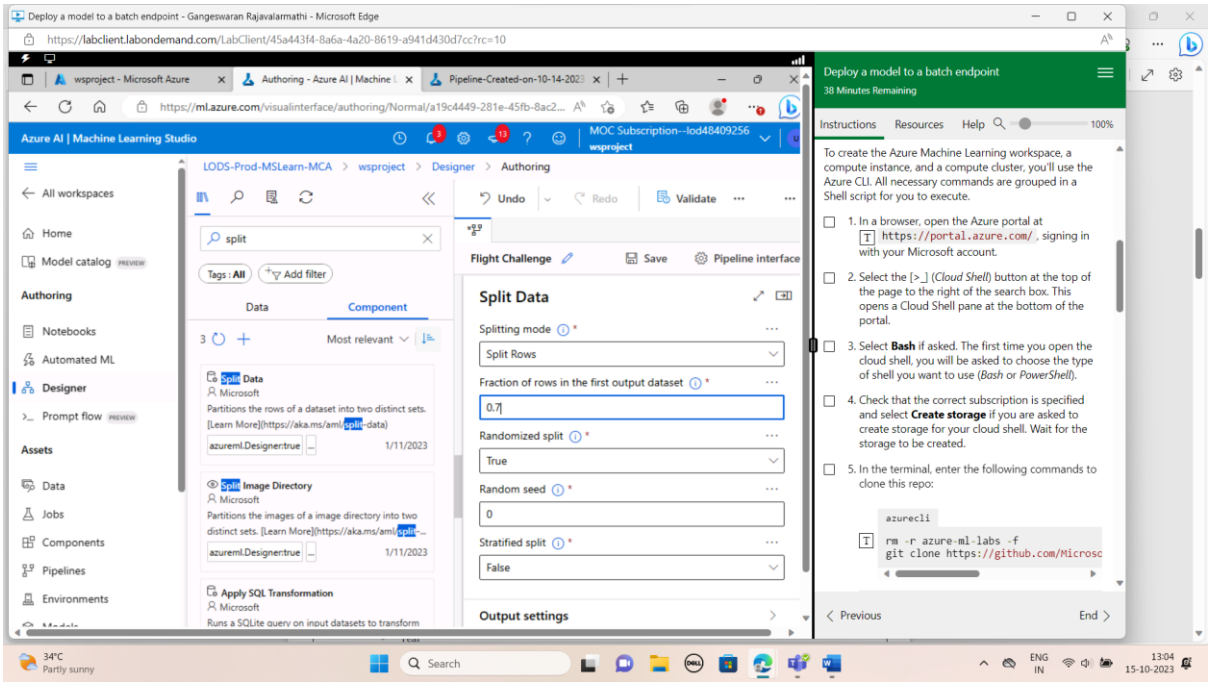




Task 4: Add a Normalize Data module and use it to standardize the CRSDepTime, CRSArrTime, and DepDelay columns using the ZScoretransformation method.

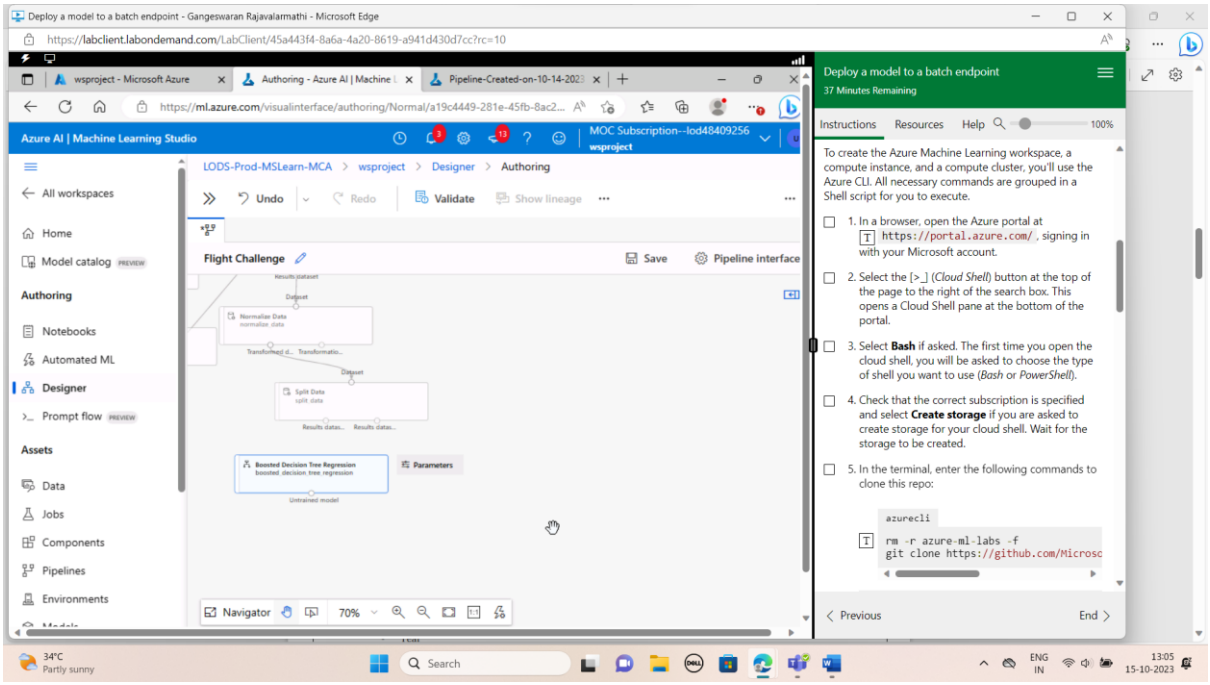


- Task 5: Add a Split Data module and use it to split the rows into 70% / 30% subsets. Use a random seed value of 0.

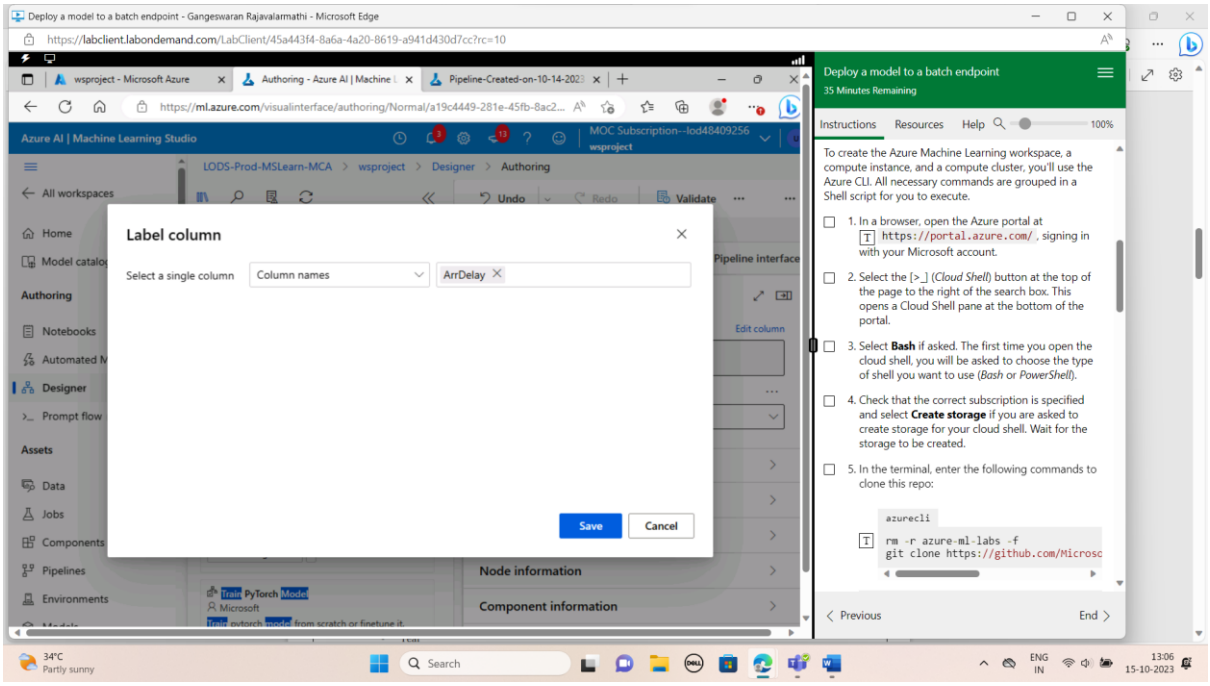


-

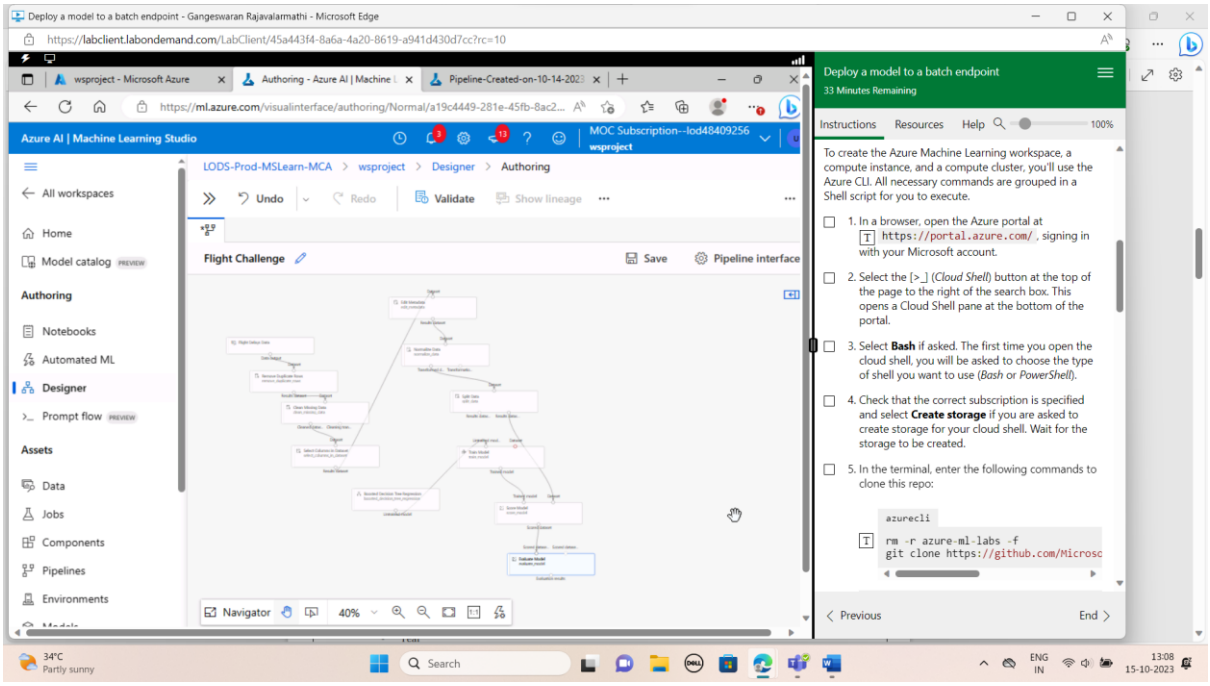
Task 6: Add a Boosted Decision Tree Regression module and a Train Model module. Then use the default settings to train the model with the 70% data split to predict the ArrDelay label column.



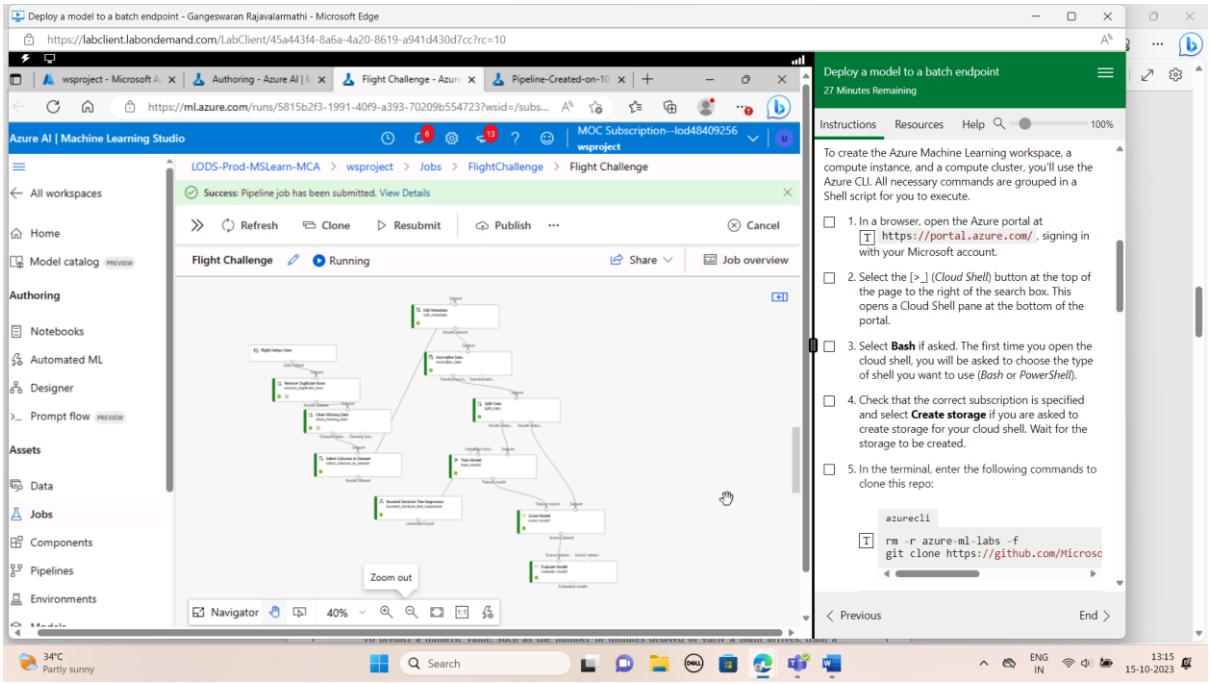
- Task 7: Add a Score Model module, and use it to score the trained model using the 30% split of data.

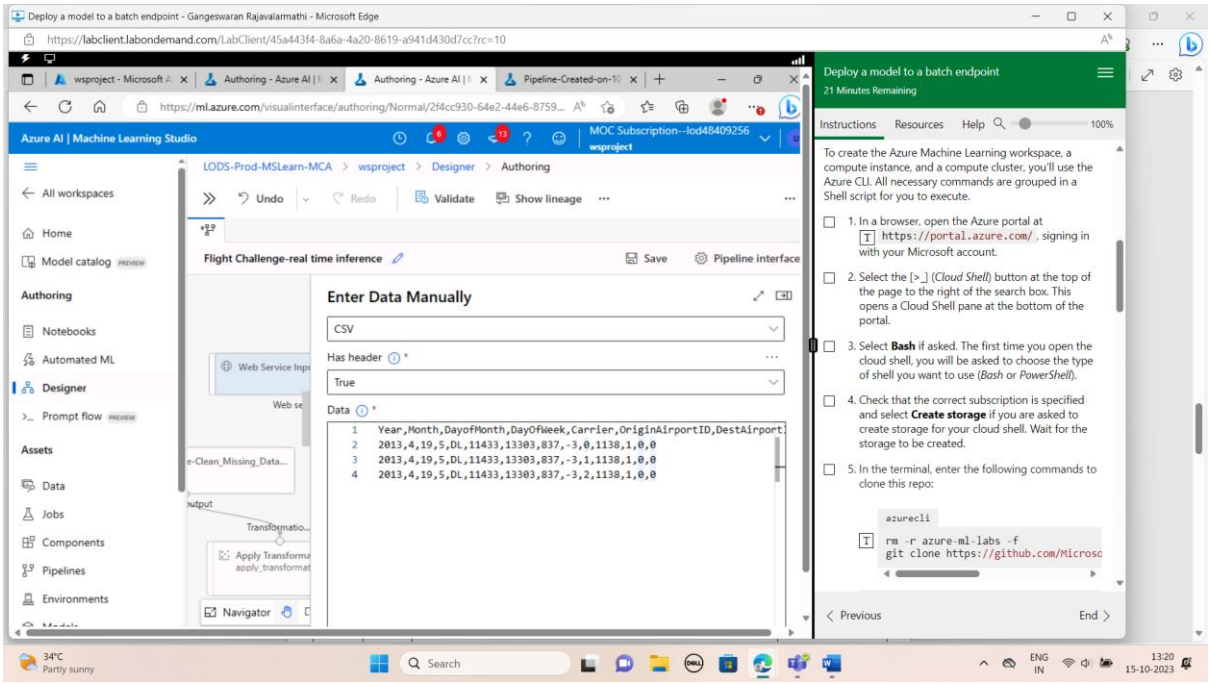


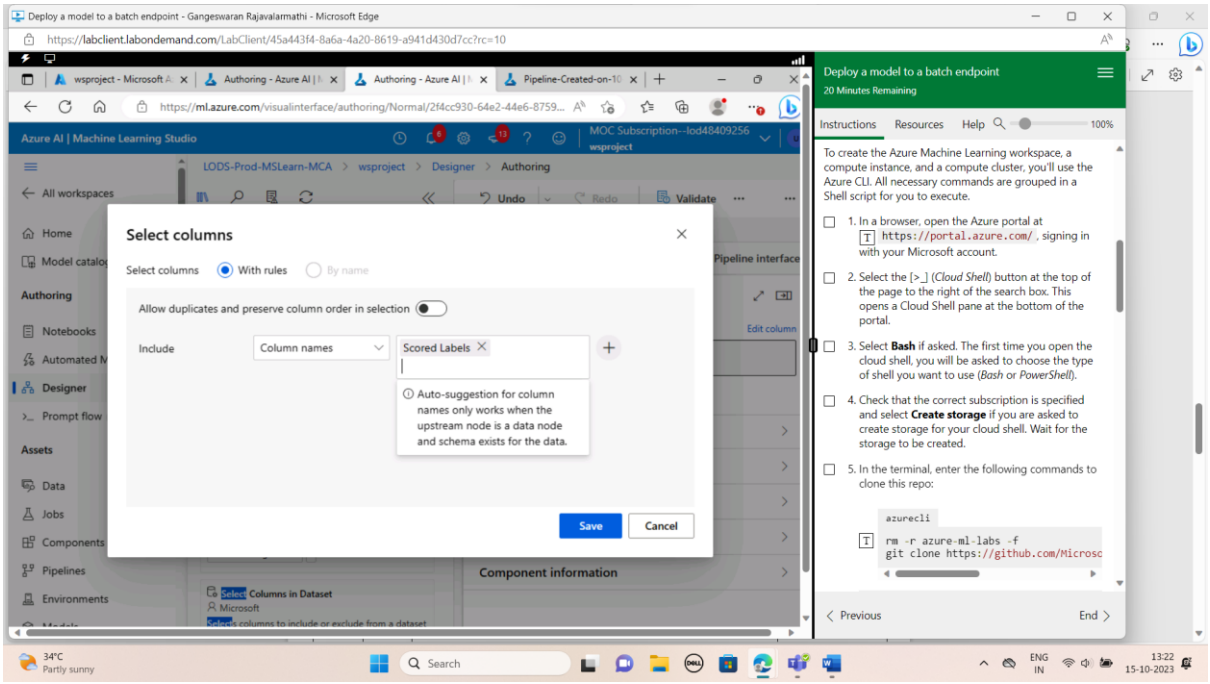
Task 8: Add an Evaluate Model module and use it to evaluate the results from the Score Model module



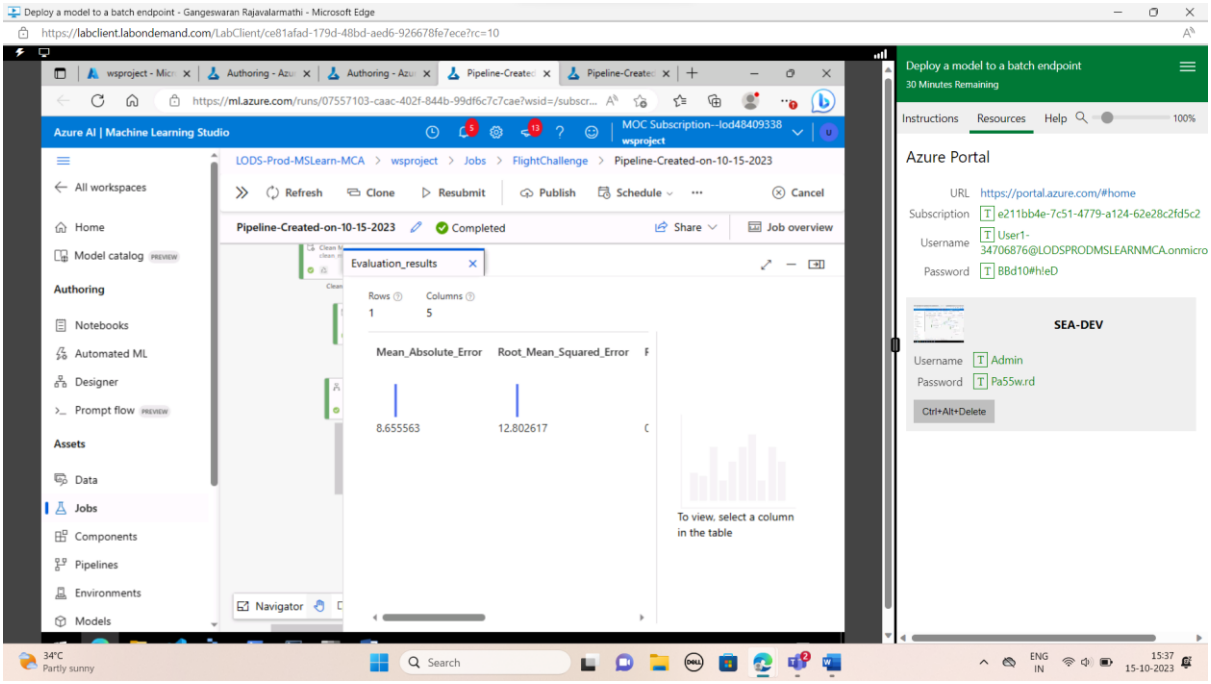
16. Screen-shots of each task of Activity 5

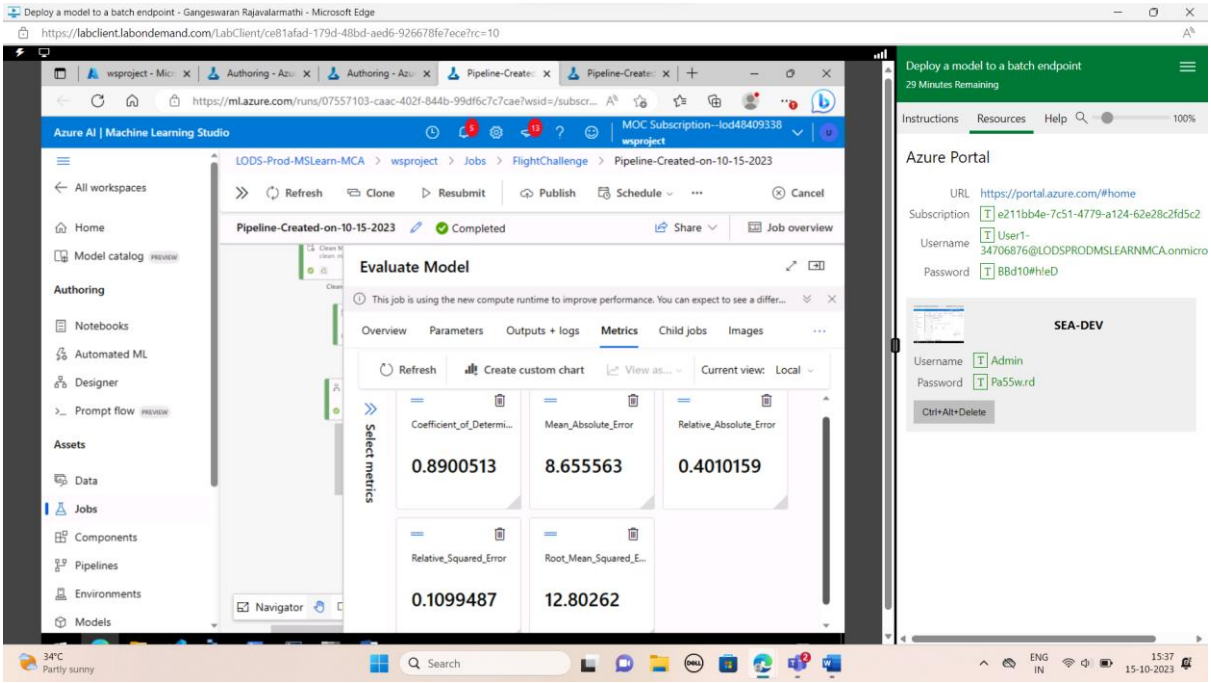



Deploy a model to a batch endpoint - Gangeswaran Rajavalarmathi - Microsoft Edge
https://labclient.labondemand.com/LabClient/45a443f4-8a6a-4a20-8619-a941d430d7cc?rc=10
Azure AI | Machine Learning Studio
Flight Challenge-real time inference
Enter Data Manually
CSV
Has header
True
Data
Year,Month,DayofMonth,DayOfWeek,Carrier,OriginAirportID,DestAirport
2013,4,19,5,DL,11433,13303,837,-3,0,1138,1,0,0
2013,4,19,5,DL,11433,13303,837,-3,1,1138,1,0,0
2013,4,19,5,DL,11433,13303,837,-3,2,1138,1,0,0
Deploy a model to a batch endpoint
21 Minutes Remaining
Instructions
Resources
Help
To create the Azure Machine Learning workspace, a compute instance, and a compute cluster, you'll use the Azure CLI. All necessary commands are grouped in a Shell script for you to execute.
1. In a browser, open the Azure portal at https://portal.azure.com/, signing in with your Microsoft account.
2. Select the [>_] (Cloud Shell) button at the top of the page to the right of the search box. This opens a Cloud Shell pane at the bottom of the portal.
3. Select Bash if asked. The first time you open the cloud shell, you will be asked to choose the type of shell you want to use (Bash or PowerShell).
4. Check that the correct subscription is specified and select Create storage if you are asked to create storage for your cloud shell. Wait for the storage to be created.
5. In the terminal, enter the following commands to clone this repo:
azurecli
rm -r azure-ml-labs -f
git clone https://github.com/Microsc
Previous
End

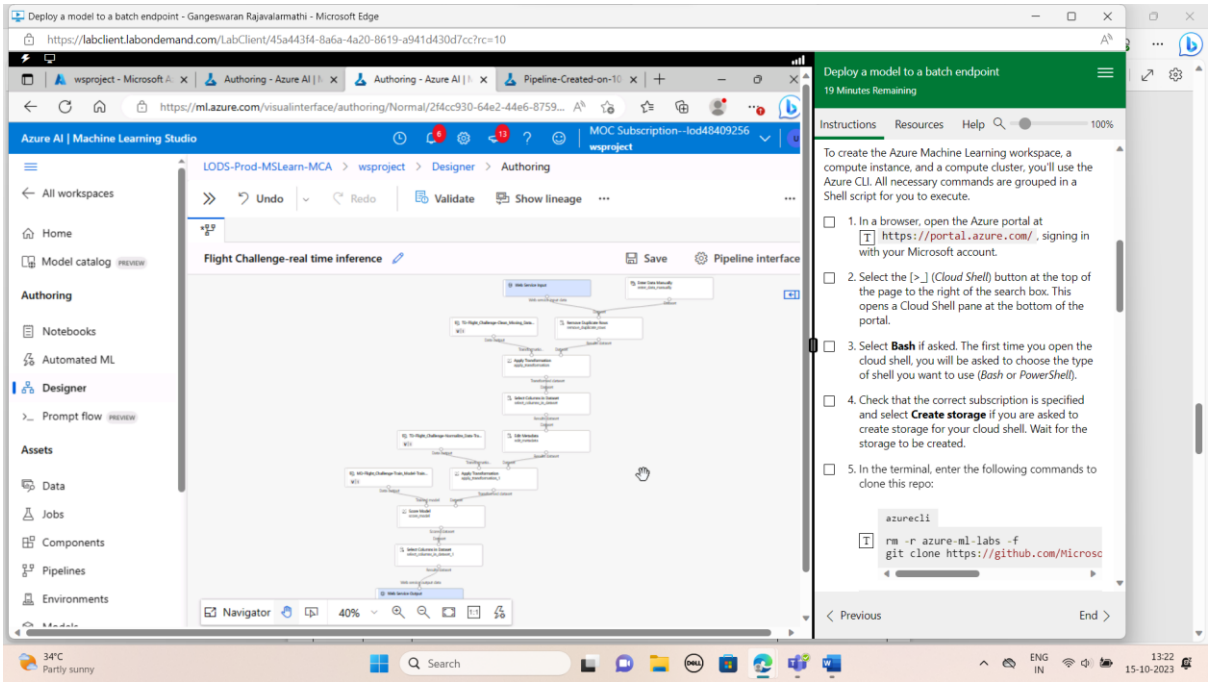

Deploy a model to a batch endpoint - Gangeswaran Rajavalarmathi - Microsoft Edge
https://labclient.labondemand.com/LabClient/45a443f4-8a6a-4a20-8619-a941d430d7cc?rc=10
Azure AI | Machine Learning Studio
Select columns
Select columns
With rules
By name
Allow duplicates and preserve column order in selection
Include
Column names
Scored Labels
Auto-suggestion for column names only works when the upstream node is a data node and schema exists for the data.
Save
Cancel
Component information
Deploy a model to a batch endpoint
20 Minutes Remaining
Instructions
Resources
Help
To create the Azure Machine Learning workspace, a compute instance, and a compute cluster, you'll use the Azure CLI. All necessary commands are grouped in a Shell script for you to execute.
1. In a browser, open the Azure portal at https://portal.azure.com/, signing in with your Microsoft account.
2. Select the [>_] (Cloud Shell) button at the top of the page to the right of the search box. This opens a Cloud Shell pane at the bottom of the portal.
3. Select Bash if asked. The first time you open the cloud shell, you will be asked to choose the type of shell you want to use (Bash or PowerShell).
4. Check that the correct subscription is specified and select Create storage if you are asked to create storage for your cloud shell. Wait for the storage to be created.
5. In the terminal, enter the following commands to clone this repo:
azurecli
rm -r azure-ml-labs -f
git clone https://github.com/Microsc
Previous
End

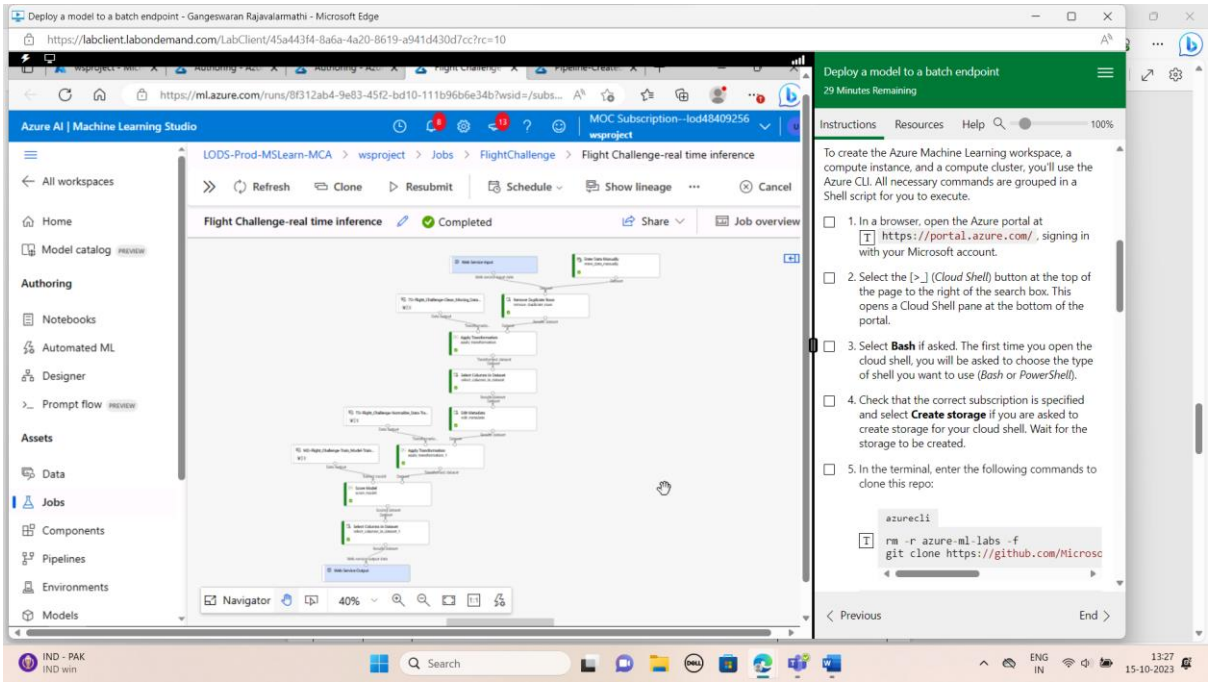
17. Screen-shots of each task of Activity 6





18. Screen-shots of each task of Activity 7





19. Screen-shots of each task of Activity 8