**Project Report**

**On**

**Yahoo Finance Stock Prices ETL and Prediction**



*Submitted*

*In partial fulfilment*

*For the award of the Degree of*

**PG-Diploma in Big Data Technology**

**(PG-DBDA)**

**C-DAC, ACTS (Pune)**

|  |  |
| --- | --- |
| **Guided By:** | **Submitted By:** |
| **Prakash Sinha** | **Duy Khiem TRUONG – PRN 240840125063** |
|  | **Ayush Mishra - PRN 240840125009** |
|  | **Huynh Van Hoa – PRN 240840125062** |
|  | **Shubham Bhagwat – PRN 240840125051** |
|  | **Shubham Sawant – PRN 240840125049** |

**Centre for Development of Advanced Computing (C-DAC), ACTS**

**(Pune- 411008)**

## ACKNOWLEDGEMENT

I would like to express my sincere gratitude to **Prakash Sinha** for their guidance and support throughout this project. I would also like to thank my peers and family for their encouragement. Special appreciation goes to the providers of the dataset used in this study. Additionally, I extend my gratitude to CDAC Pune for their training and staff support, which significantly contributed to the completion of this project.

## ABSTRACT

This project aims to analyze stock price trends using the ARIMA model, specifically focusing on Apple Inc. (AAPL) stock prices. Time series forecasting is crucial for investors and financial analysts, and ARIMA (Auto-Regressive Integrated Moving Average) is a widely used technique for such predictions. The project involves data collection, preprocessing, model implementation, and performance evaluation. The results demonstrate ARIMA’s ability to capture patterns in AAPL stock price movements, providing insights into market behavior and potential investment strategies.

**Table of Contents**

Contents

[ACKNOWLEDGEMENT 2](#_Toc189819921)

[ABSTRACT 3](#_Toc189819922)

[Introduction 1](#_Toc189819923)

[Introduction 1](#_Toc189819924)

[Objective and Specifications 2](#_Toc189819925)

[LITERATURE REVIEW 3](#_Toc189819926)

[The ARIMA model 3](#_Toc189819927)

[ARIMA model for Stock Prices 3](#_Toc189819928)

[Methodology and Techniques 5](#_Toc189819929)

[Methodology: 5](#_Toc189819930)

[Dataset and Plot 6](#_Toc189819931)

[Implementation 7](#_Toc189819932)

[Check the API is available 8](#_Toc189819933)

[Get the stock prices from website 9](#_Toc189819934)

[Store the stock prices in to MinIO 9](#_Toc189819935)

[Format the stock prices 10](#_Toc189819936)

[Get the formatted csv 10](#_Toc189819937)

[Train the ARIMA model 11](#_Toc189819938)

[Results 15](#_Toc189819939)

[Airflow Dashboard 15](#_Toc189819940)

[MinIO Dashboard 20](#_Toc189819941)

[Conclusion 23](#_Toc189819942)

[Conclusion 23](#_Toc189819943)

[Final Thoughts 24](#_Toc189819944)

[References 25](#_Toc189819945)

**Chapter 1**

# Introduction

Detection of text regions either from handwritten or printed document images containing various non-textual information is a difficult task, and it can be more challenging to locate the position of the text regions when we deal with a doctor’s prescription.

## Introduction

The stock is exposed to different types of risk and uncertainties which have an impact on the price of the stock. It is difficult to predict the stock price. The stock price is influenced by various factors related to demand and supply. For predicting the price of a stock, we require dependent variables like stock market index, similar or identical company, sales, profits, earnings per share, etc. When the dependent variable does not have any impact and it is impossible to predict the stock price in such cases, the stock price is predicted by considering the past stock value on different time horizons like days, week, months, quarterly, or yearly by applying the autoregressive moving average method called ARIMA.

Stock prediction using time series analysis is an emerging area in predictive analytics. It has attracted many researchers because of the utility and accuracy of the model. The main objective of the ARIMA model is to study past observations based on which future models are generated to forecast a given variable. The success of the ARIMA model depends on appropriate model identification and evaluation.

It is essential to understand that the ARIMA model is applied under what circumstances.

* No dependent variable is available.
* Good sufficient historical data is available.
* Autocorrelation.

Time series models like ARIMA help in analyzing stock price trends based on historical data. This project applies ARIMA to predict future stock prices and evaluates its effectiveness using Apple Inc. (AAPL) as a case study.

## Objective and Specifications

Objectives:

* To demonstrate ETL pipeline using Airflow and other technology
* To build an ARIMA model for forecasting AAPL stock prices.
* To evaluate the model’s accuracy using performance metrics.
* To predict (forecast) future values of AAPL stock prices

Specifications:

* Programming Language: Python
* Technology Stack: Apache Airflow, Mlflow, MinIO, Apache Spark, Docker
* Libraries: Pandas, Statsmodels and others
* Dataset: Historical AAPL stock prices from Yahoo Finance through API

The project will put emphasis on the demonstration of ETL & prediction pipeline using Airflow and other related technology.

**Chapter 2**

# LITERATURE REVIEW

## The ARIMA model

The ARIMA (Auto-Regressive Integrated Moving Average) model has been widely used in time series forecasting due to its effectiveness in capturing linear dependencies in data. Box and Jenkins (1976) introduced the ARIMA model as a structured approach for analyzing and forecasting time series data. Since then, various studies have explored its applications in financial markets.

The ARIMA model forecasts time lags which are equally spaced in time horizon with univariate time series. In ARIMA, AR stands for autoregressive, which emphasizes on the relationship between the past values and the future values, Letter I stands for integrated, and MA stands for moving average.

It is represented by the equation:

yt = Ф0 + Ф1 it−1 + Ф2 it−2 +. . .+ Фpat-p + єt−θ1 є t−1−θ2 є t−2−. . .−θq є t−q

where actual data values are denoted as yt, coefficients are denoted as Фi and θj, Єi denotes the random errors, and integers p and q repre­sent the degrees of autoregressive and moving averages (Ayodele et al., 2014). The ARIMA model is a mixture of two equations: Autoregres­sive is the equation based on past lags and the moving average is based on error.

## ARIMA model for Stock Prices

Several studies have demonstrated the ARIMA model’s effectiveness in stock price forecasting. For instance, Fama (1970) discussed the Efficient Market Hypothesis (EMH), which suggests that stock prices follow a random walk, challenging the predictability of stock prices. However, studies by Pai and Lin (2005) and Atsalakis & Valavanis (2009) showed that ARIMA models could capture short-term market trends and patterns.

Comparative studies have been conducted to evaluate ARIMA’s forecasting performance against other models such as GARCH, LSTM, and hybrid models. Research by Zhang (2003) indicated that while ARIMA is robust for stationary time series, hybrid approaches combining ARIMA with machine learning techniques like artificial neural networks (ANN) improve forecasting accuracy. More recent studies highlight that ARIMA models, despite their simplicity, remain a strong baseline for financial time series forecasting (Hyndman & Athanasopoulos, 2018).

The ARIMA model is considered to be the most reliable model in such a situation; Box and Jenkins is the researchers and scientists who devel­oped the ARIMA model in 1970.

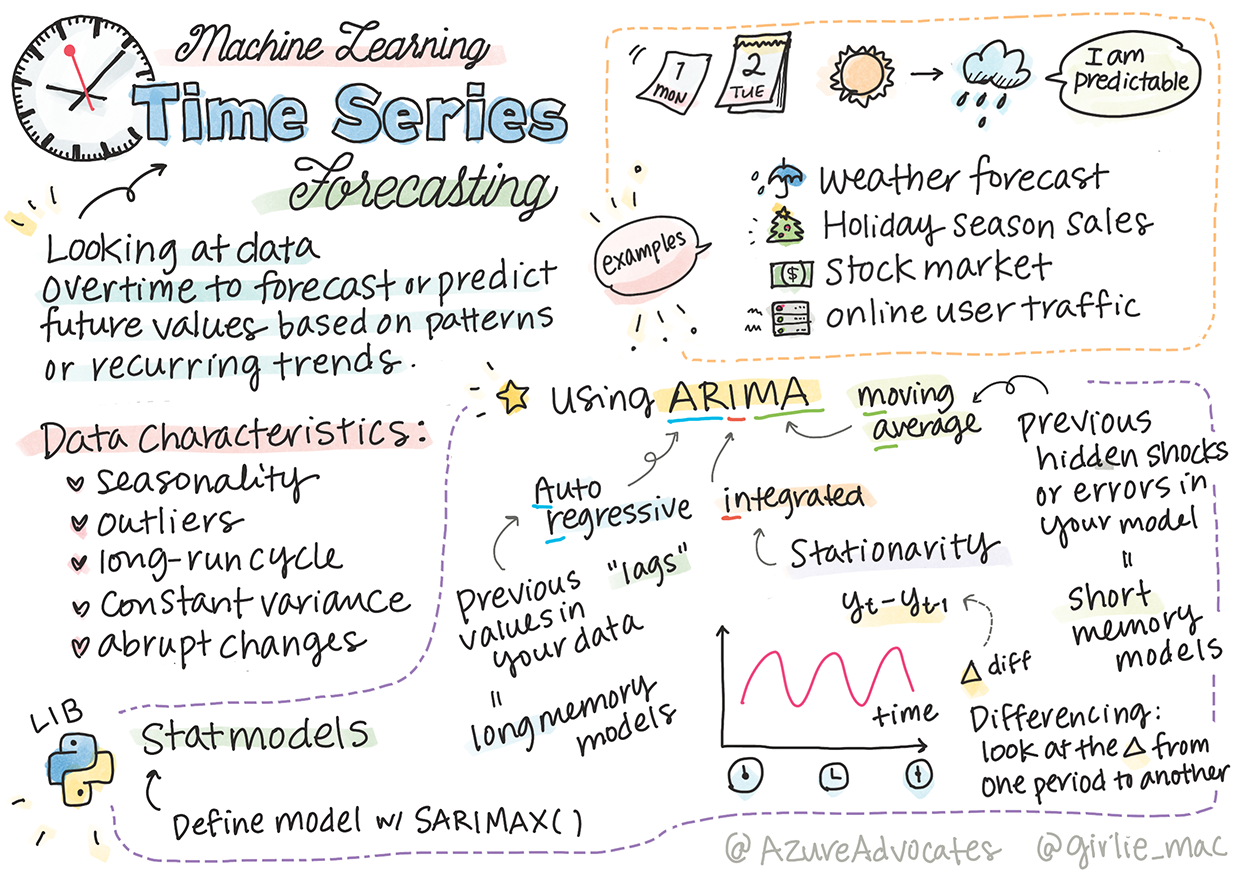
In the context of Apple Inc. (AAPL) stock price forecasting, researchers have applied ARIMA models to analyze historical trends and short-term fluctuations. Studies suggest that ARIMA can effectively capture seasonality and trend components in AAPL stock prices, making it useful for short-term investment strategies (Chen et al., 2013).

While ARIMA remains a widely used method, its limitations include the assumption of linearity and stationarity in data. Therefore, integrating ARIMA with other models, such as deep learning techniques, is suggested for enhanced predictive performance.

**Chapter 3**

# Methodology and Techniques

## Methodology:



* 1. *Data Source*

Yahoo Finance financial data is used to create the ARIMA model.

* 1. *Period of Study*

The interval for the selected data is the daily/hourly closing stock price from one year of AAPL for analysis.

* 1. *Software Used for Data Analysis*

Python Programming, Anaconda

* 1. *Model Applied*

For this study, we applied the ARIMA model.

* 1. *Limitations of the Study*

The study is restricted to the Stock Prices Index of AAPL only.

* 1. *Future Scope of the Study*

In the future, the study can be done on the macro level by applying it to a different stock at the same time.

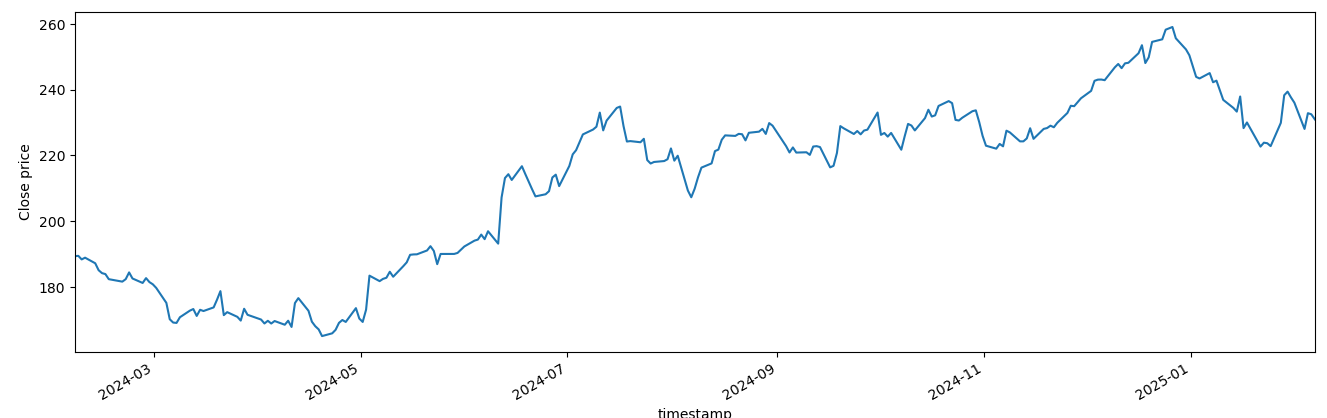
* 1. *Methodology*

The autoregressive integrated moving average (ARIMA) model pre­dicts the stock price by using past values. The ARIMA model is applied to understand the relationship between the past values of stock for pre­dicting its future predicted value. The ARIMA model is widely applied in the field of stock price prediction. The ARIMA model is imple­mented first by understanding the relationship between the past values of stock and its future value. Autocorrelation plays an important role in model development. The check for autocorrelation defines further steps of model evaluation and parameter estimation to select the best ARIMA model for stock prediction using Python. Research is carried out in three steps. First, we need to check the autocorrelation. Then, we need to evaluate different ARIMA models and compare the AIC of other models. The best model is selected with the lowest AIC.

## Dataset and Plot

The dataset is extracted from Yahoo Finance website





**Chapter 4**

# Implementation

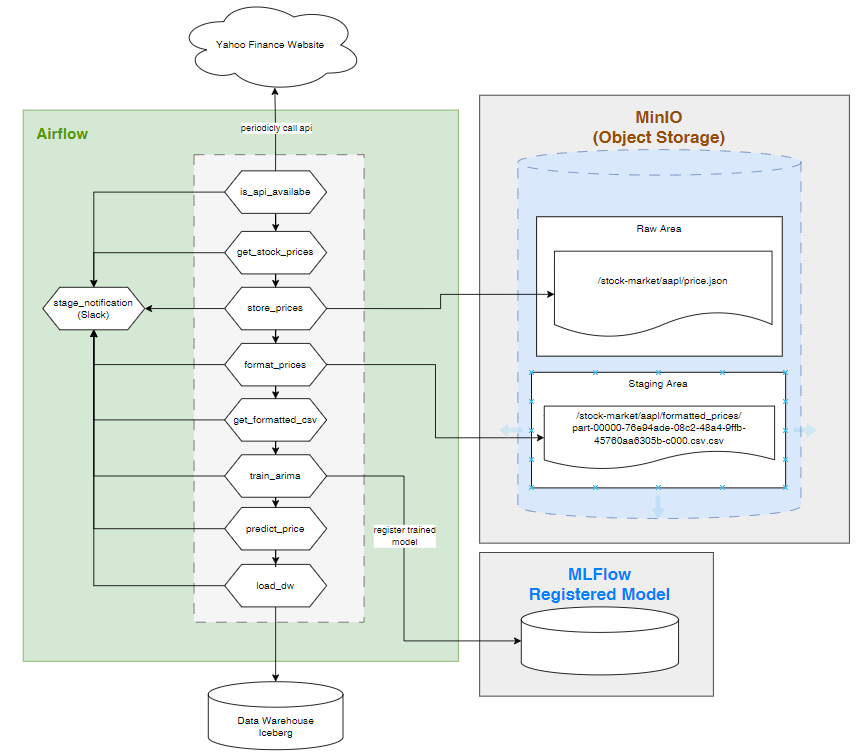
**Hardware Configuration**

* + CPU: 16 GB RAM, Quad core processor

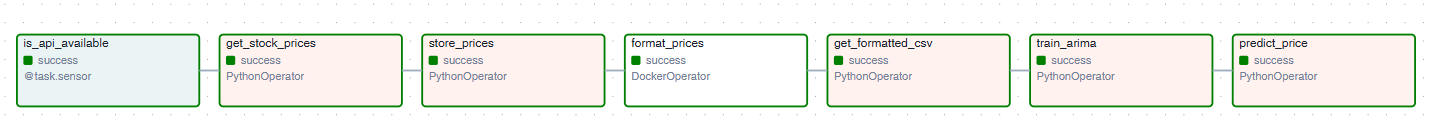
**Software Required**

* **Airflow**: Apache Airflow is a **workflow automation tool** that allows you to programmatically author, schedule, and monitor data pipelines. It is widely used for **ETL (Extract, Transform, Load) processes, machine learning workflows, and data engineering pipelines**.
* **Mlflow**: MLflow is an **open-source platform** for managing the machine learning (ML) lifecycle. It helps **track experiments, manage models, and deploy ML applications** efficiently.
* **MinIO**: MinIO is an **open-source, high-performance, S3-compatible object storage system**. It is designed for **cloud-native applications, big data, and AI/ML workloads**.
* **Spark**: Apache Spark is an **open-source, distributed computing framework** for processing large-scale data quickly. It is widely used for **big data analytics, machine learning, and real-time stream processing**.

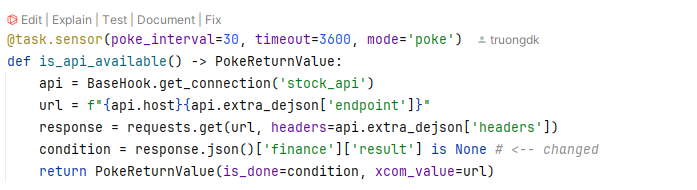
**Architecture:**



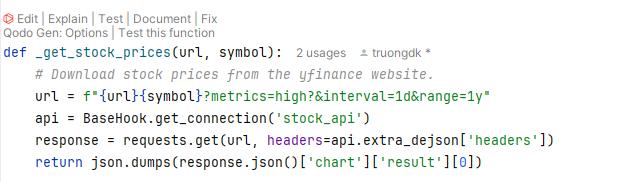
**Data Pipeline & Code base**



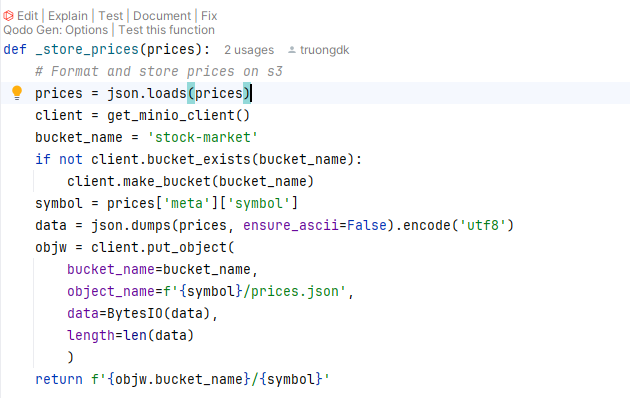
### Check the API is available



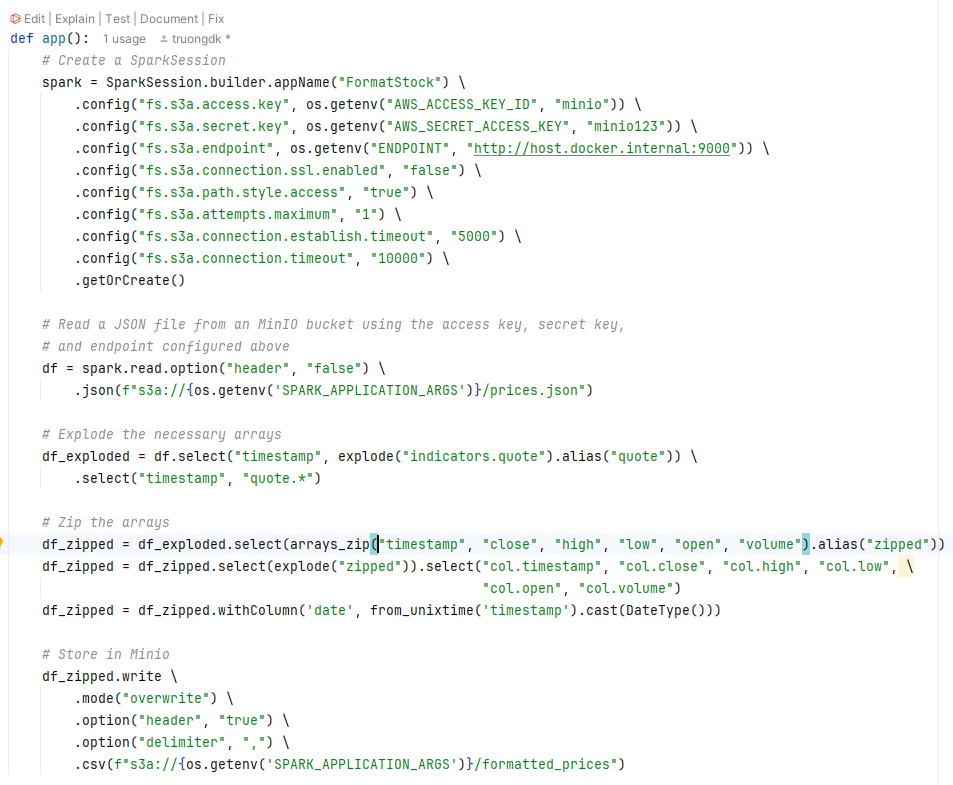
### Get the stock prices from website



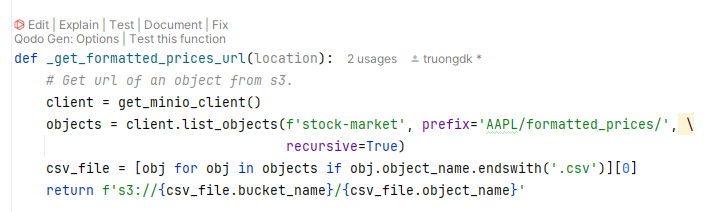
### Store the stock prices in to MinIO



### Format the stock prices



### Get the formatted csv

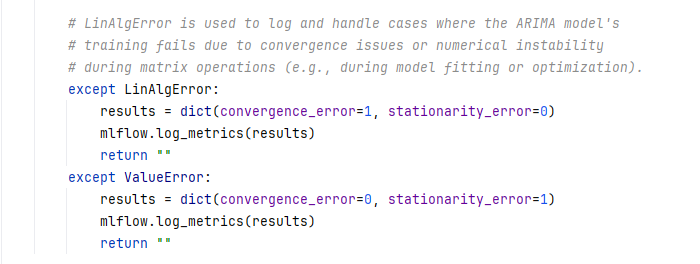


### Train the ARIMA model

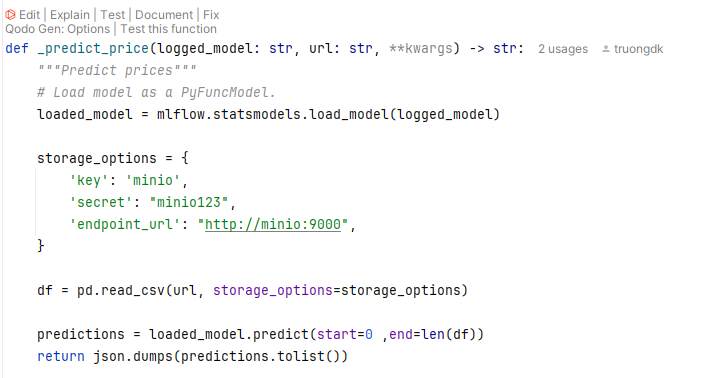




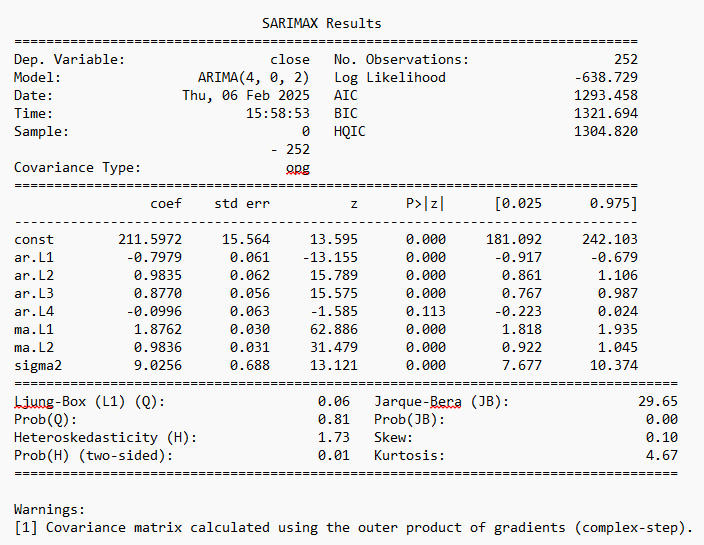




* 1. **Predict the stock prices**



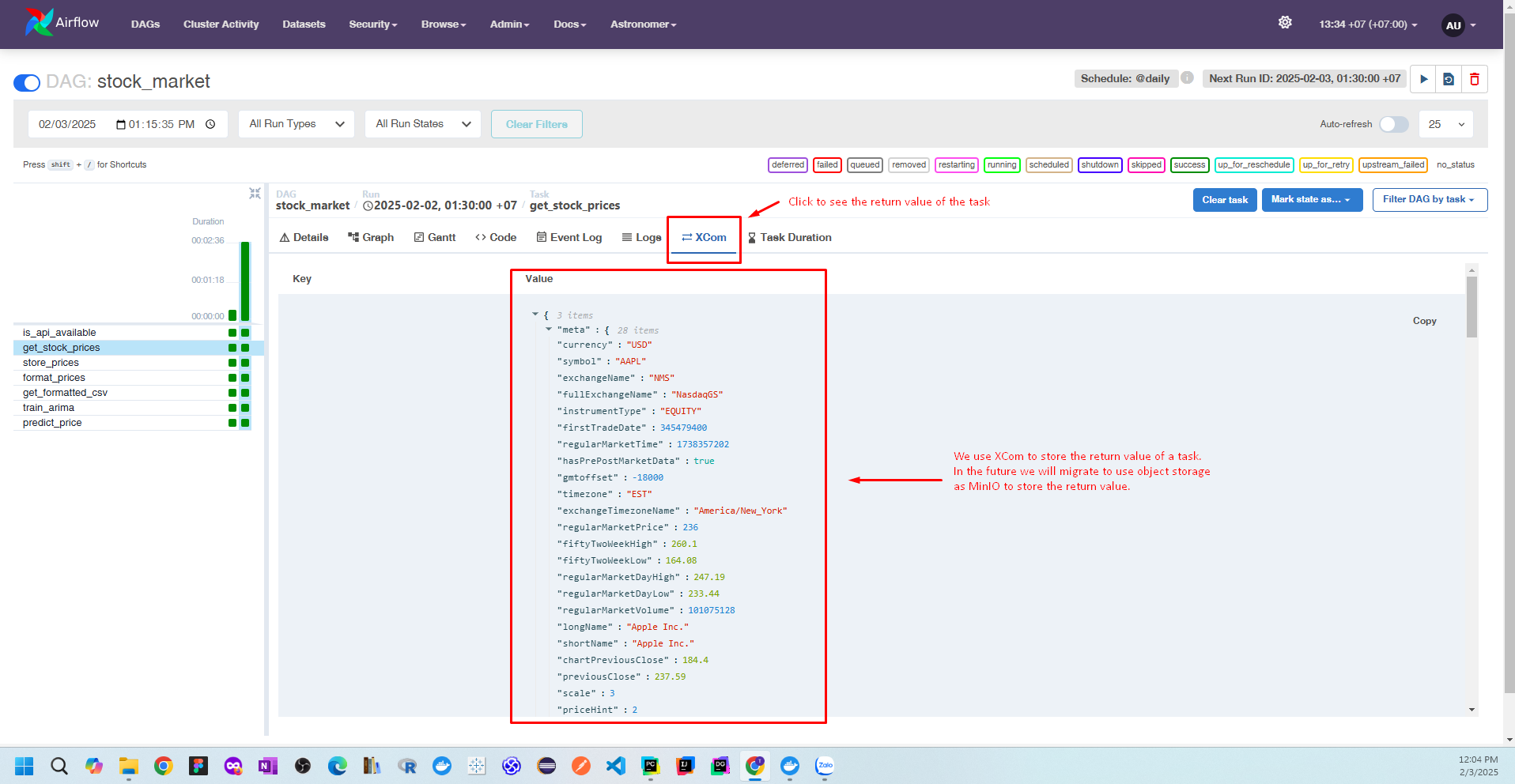
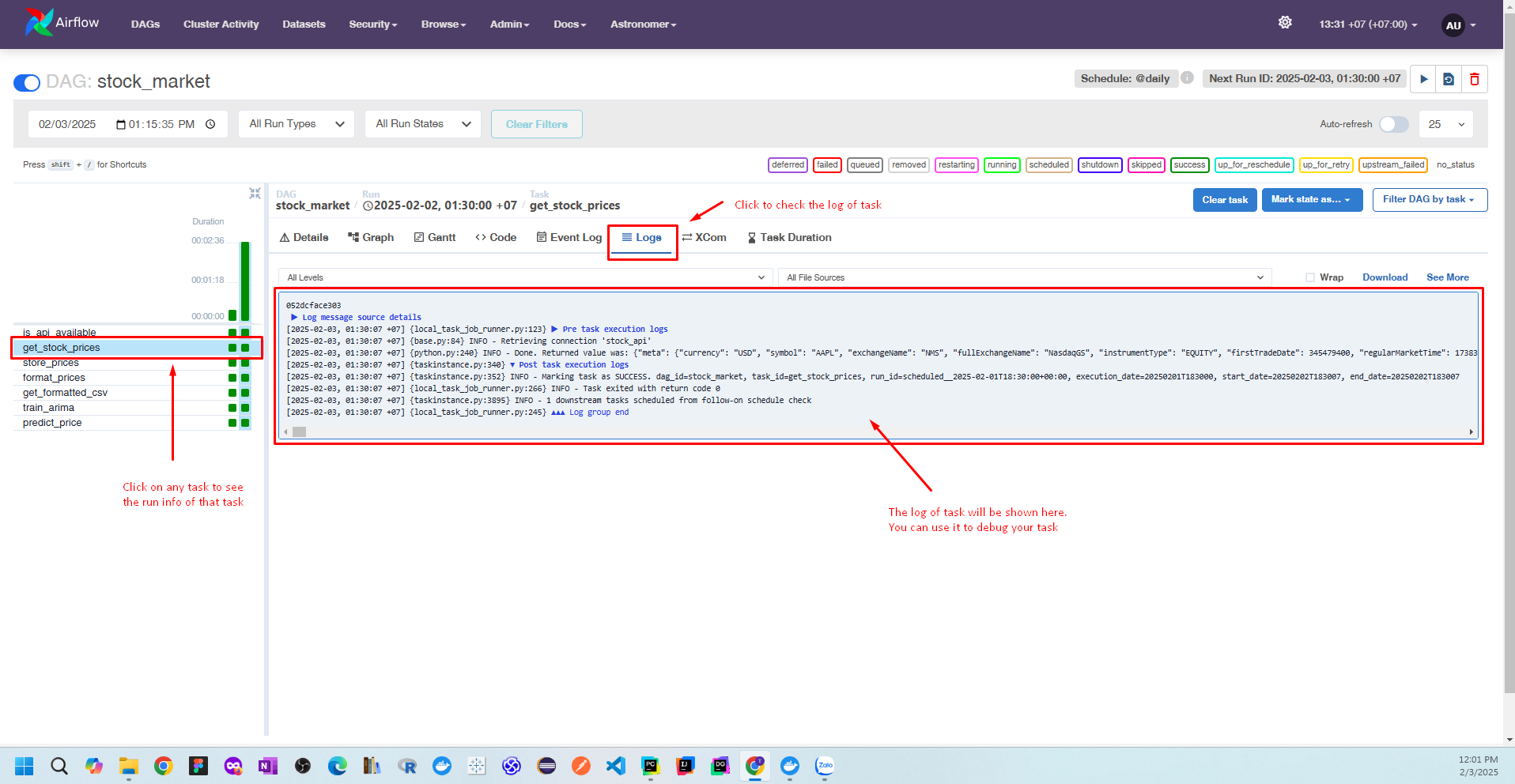
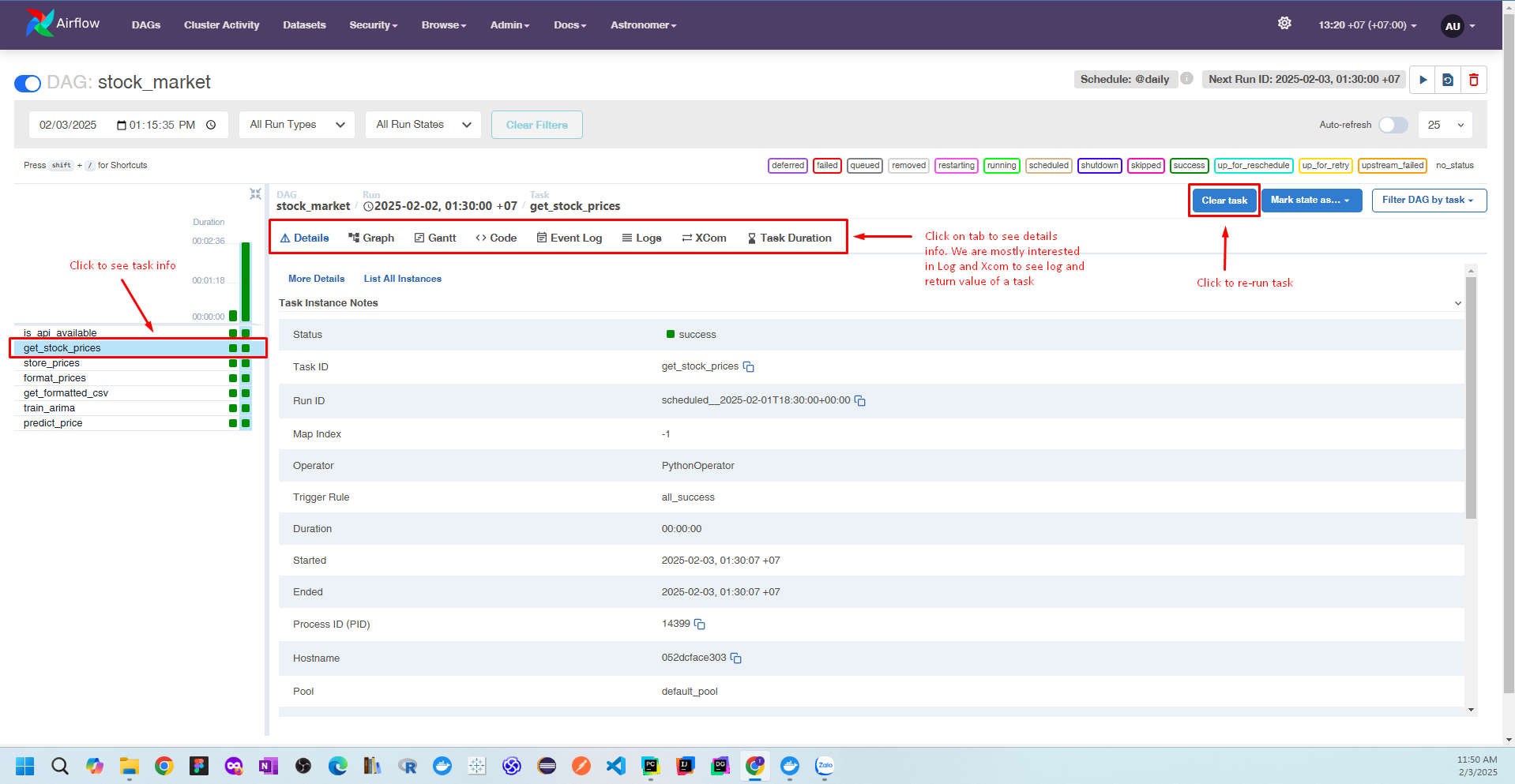
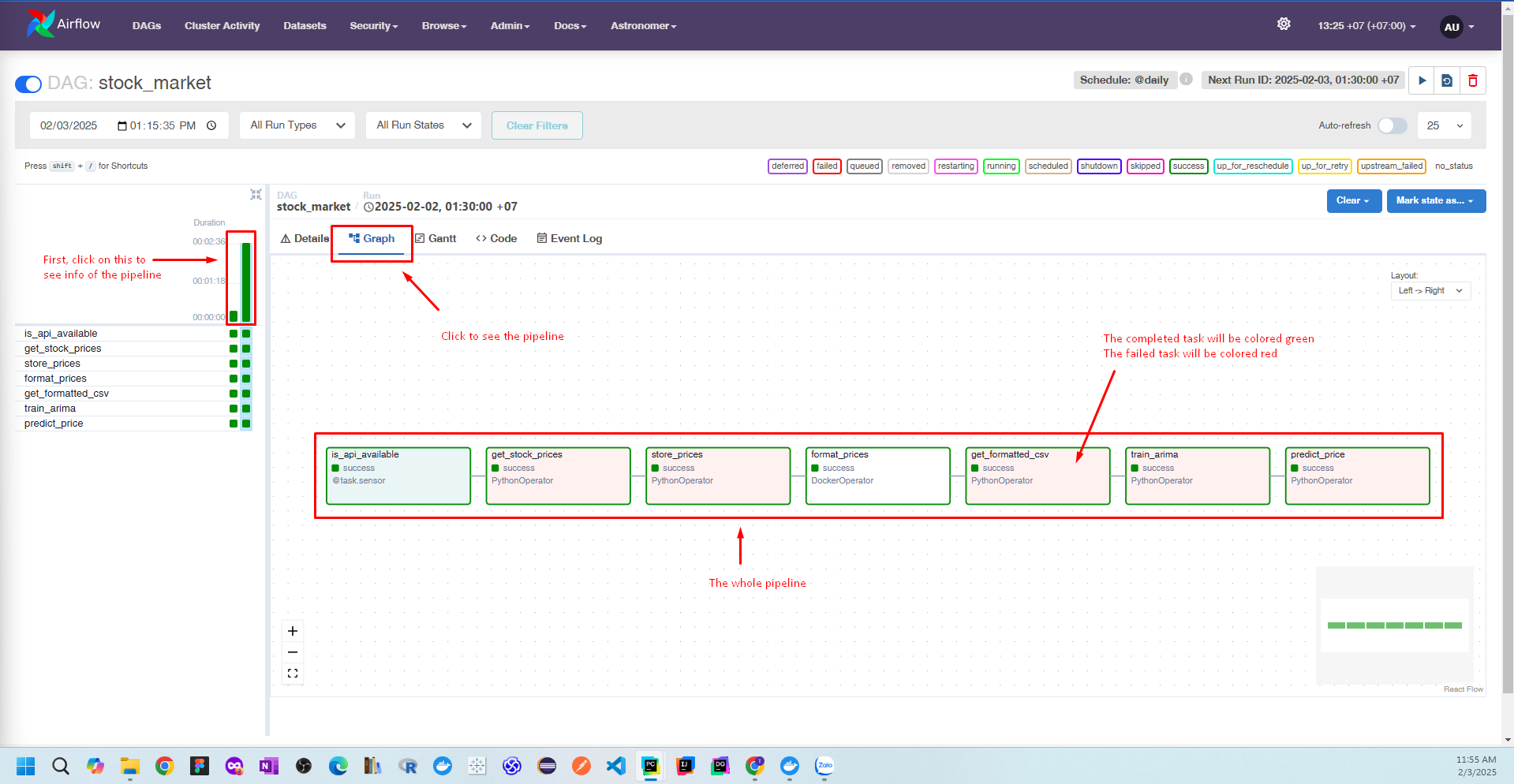
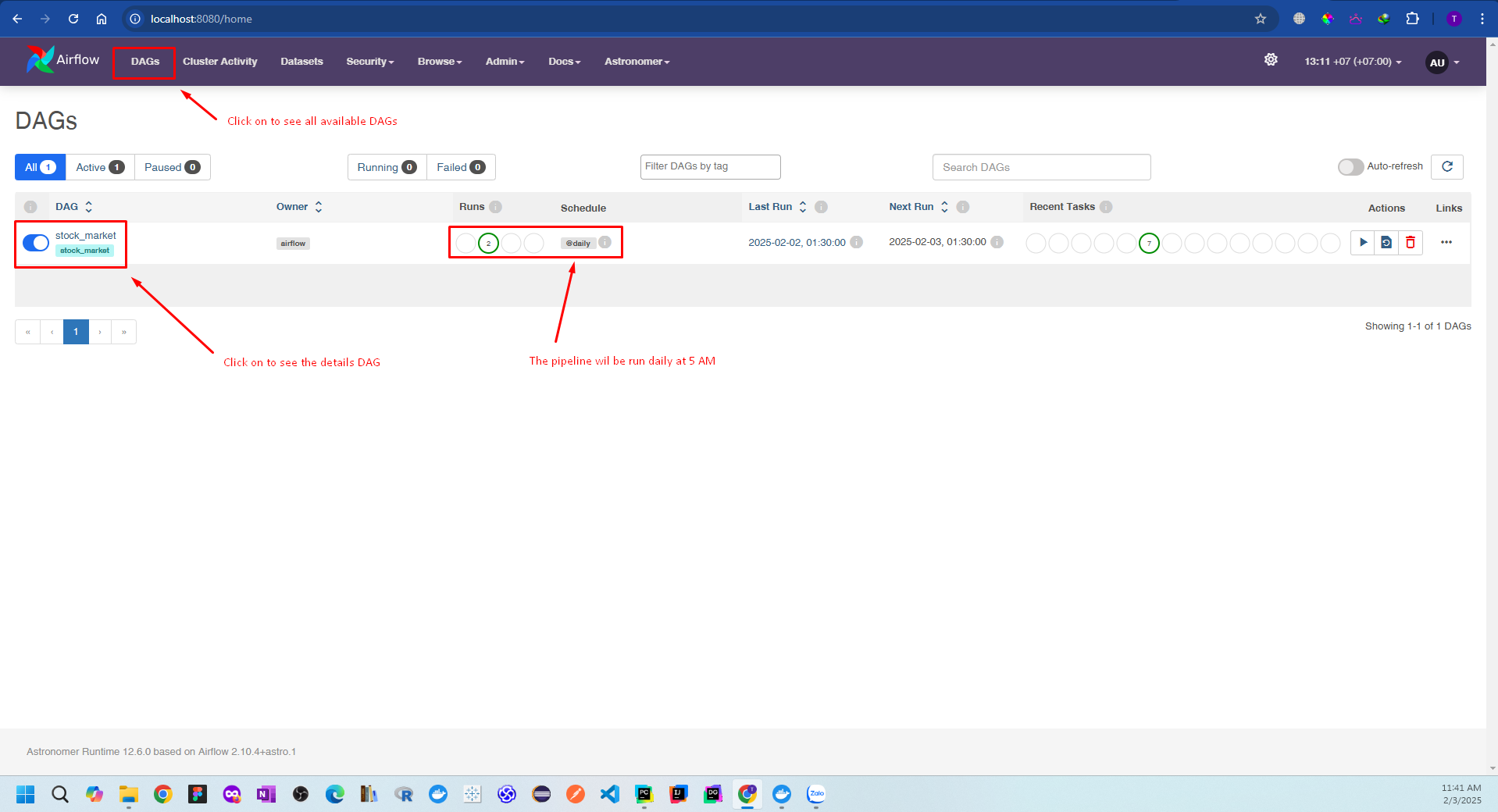
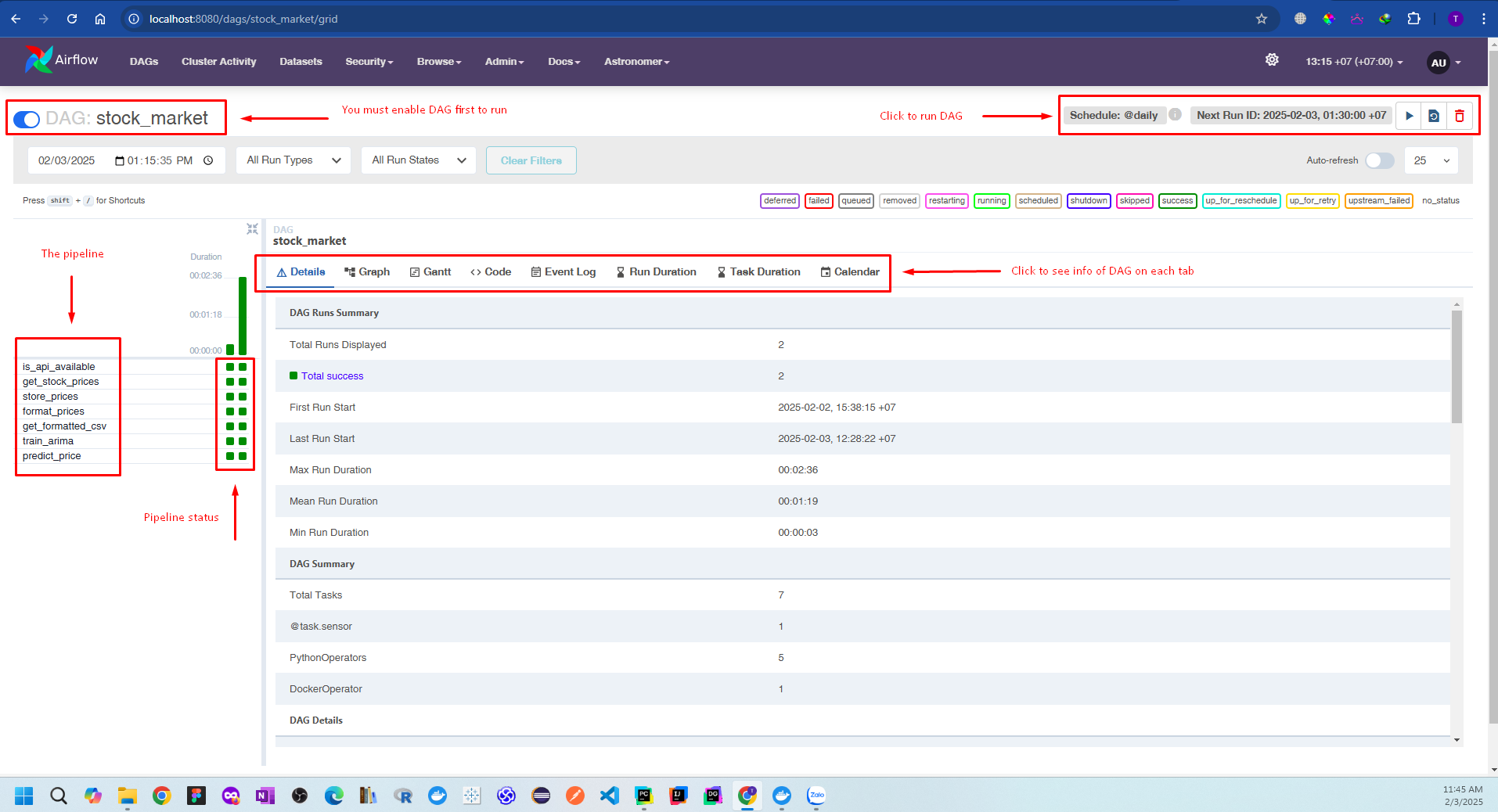
**ARIMA Model Summary**



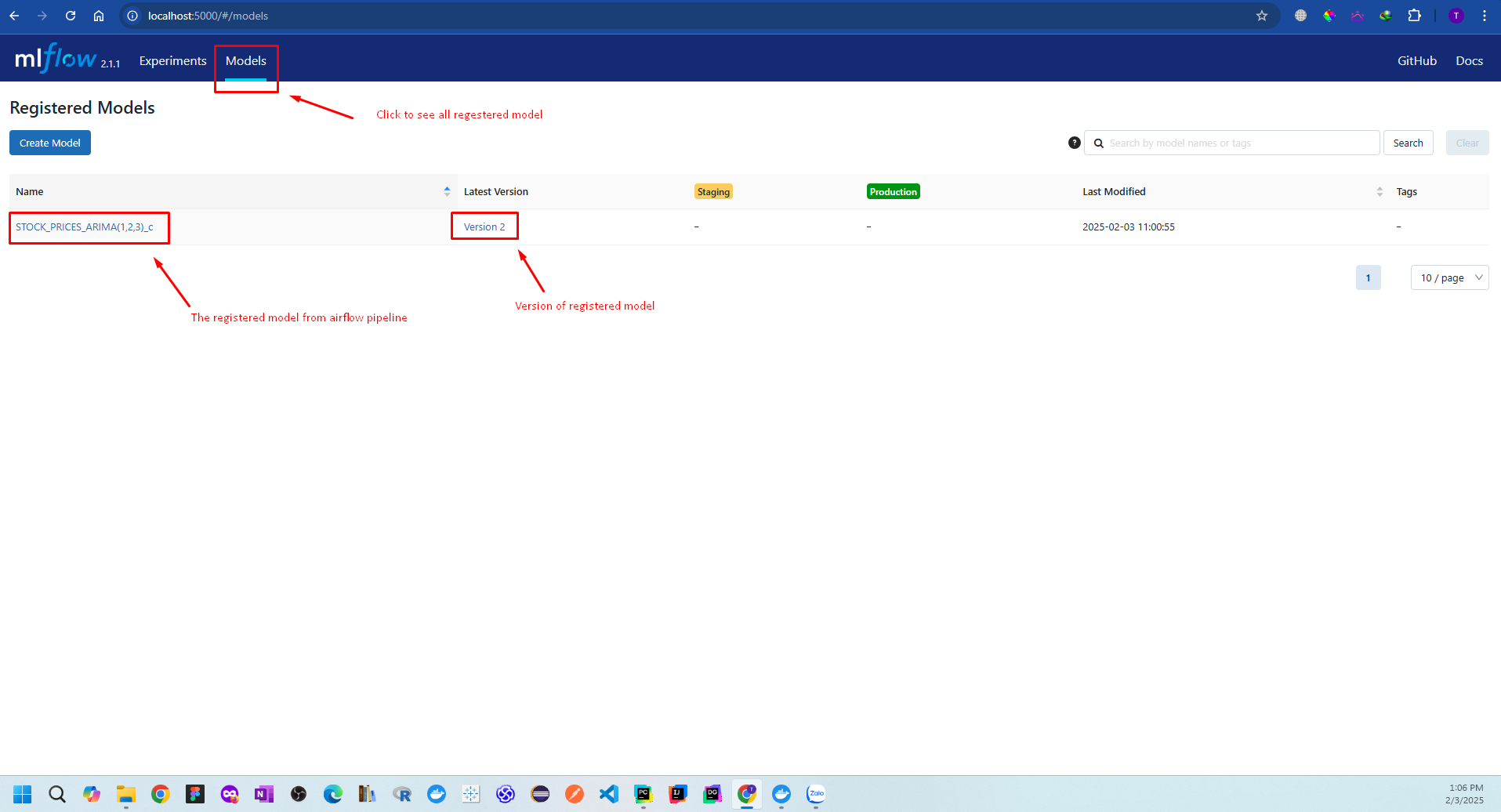
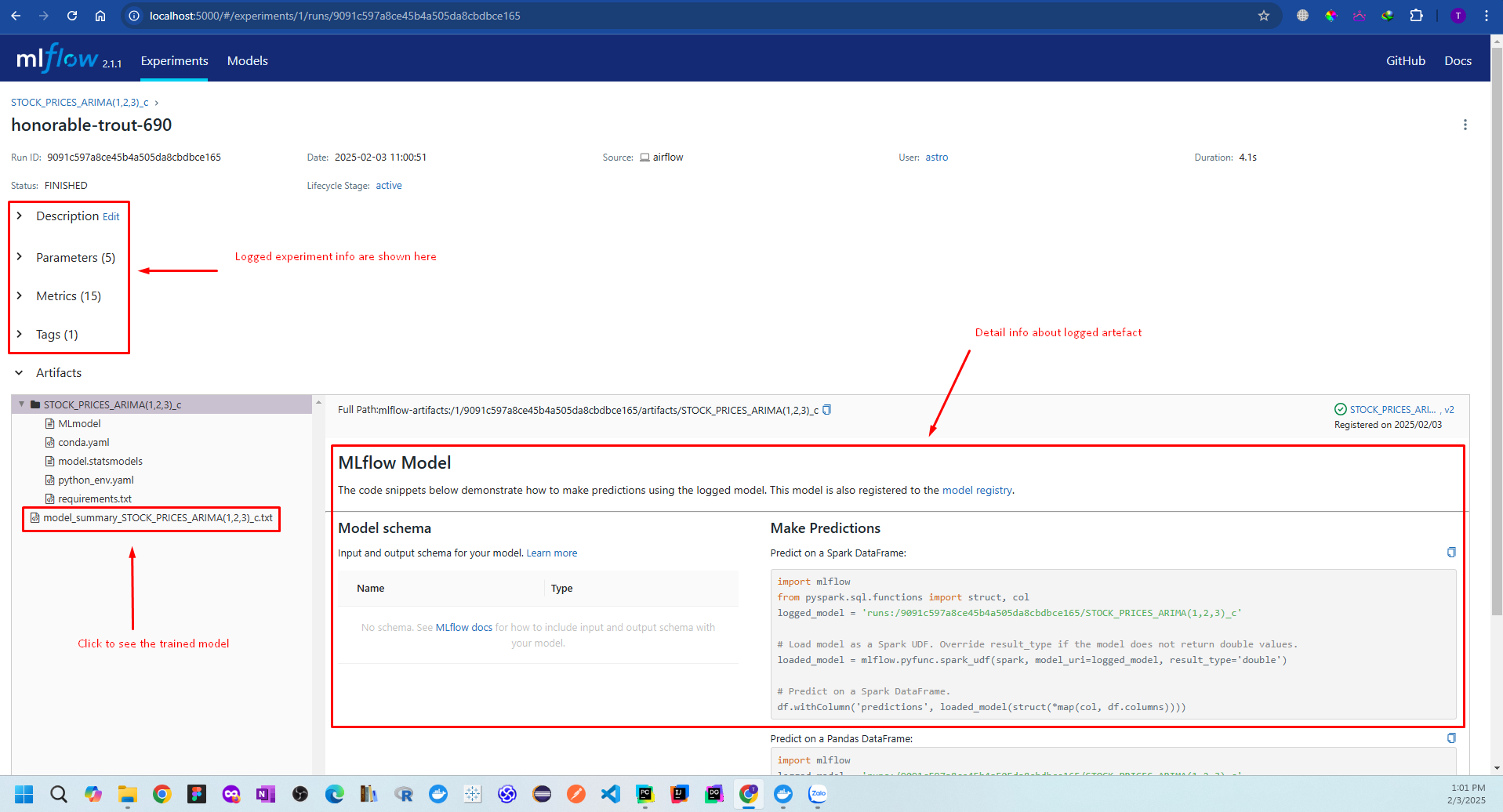
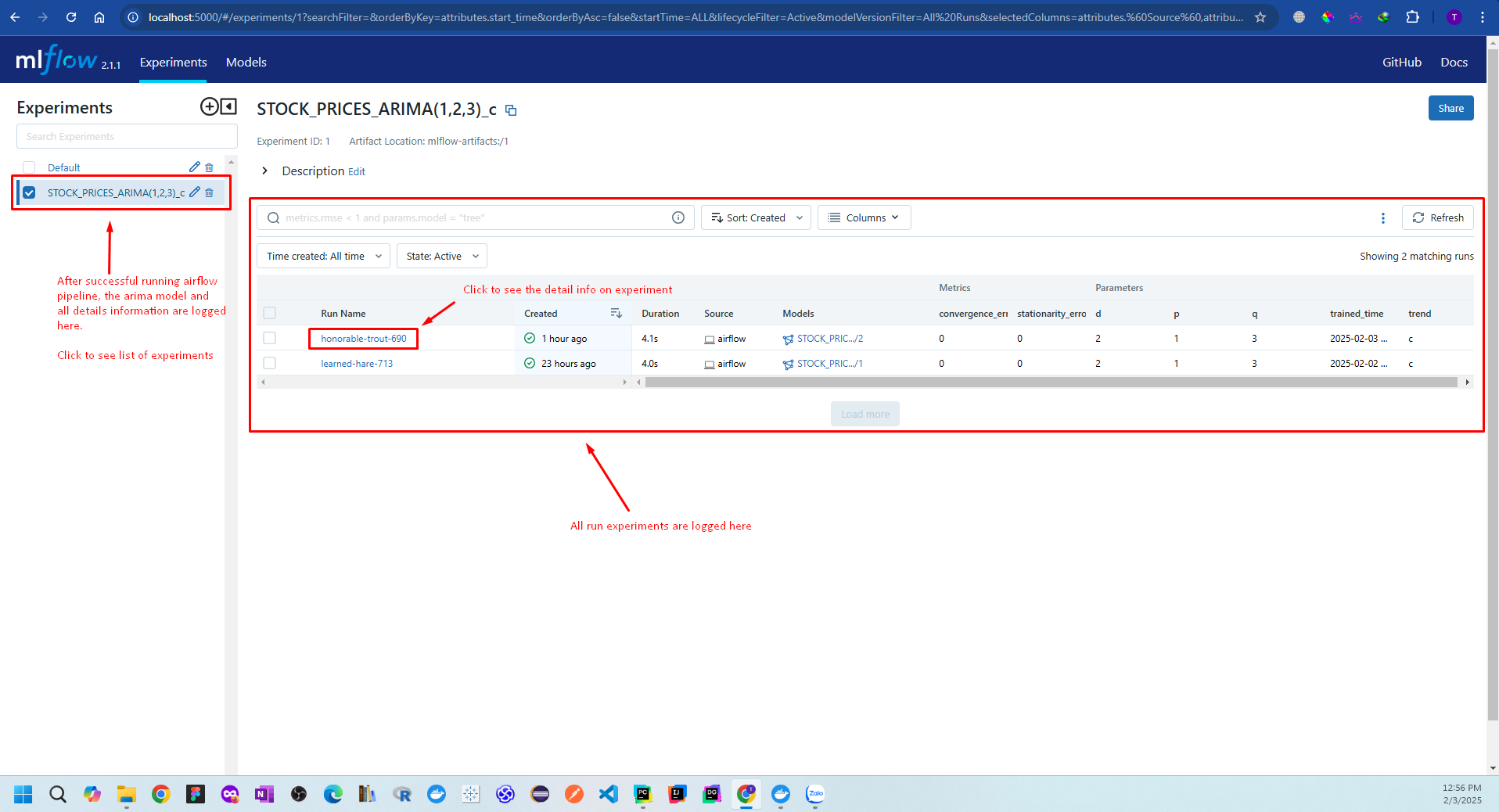
**Chapter 5**

# Results

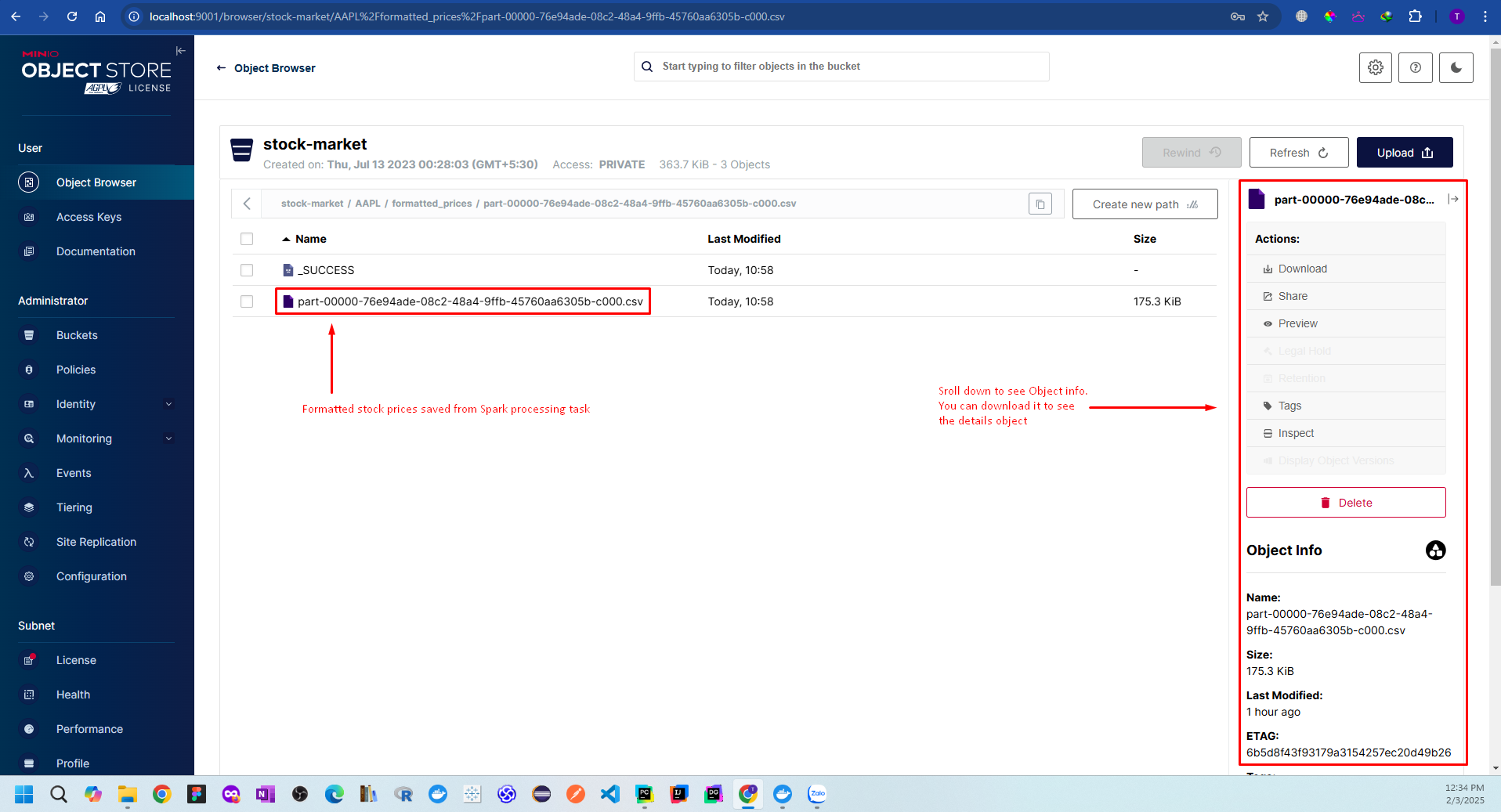
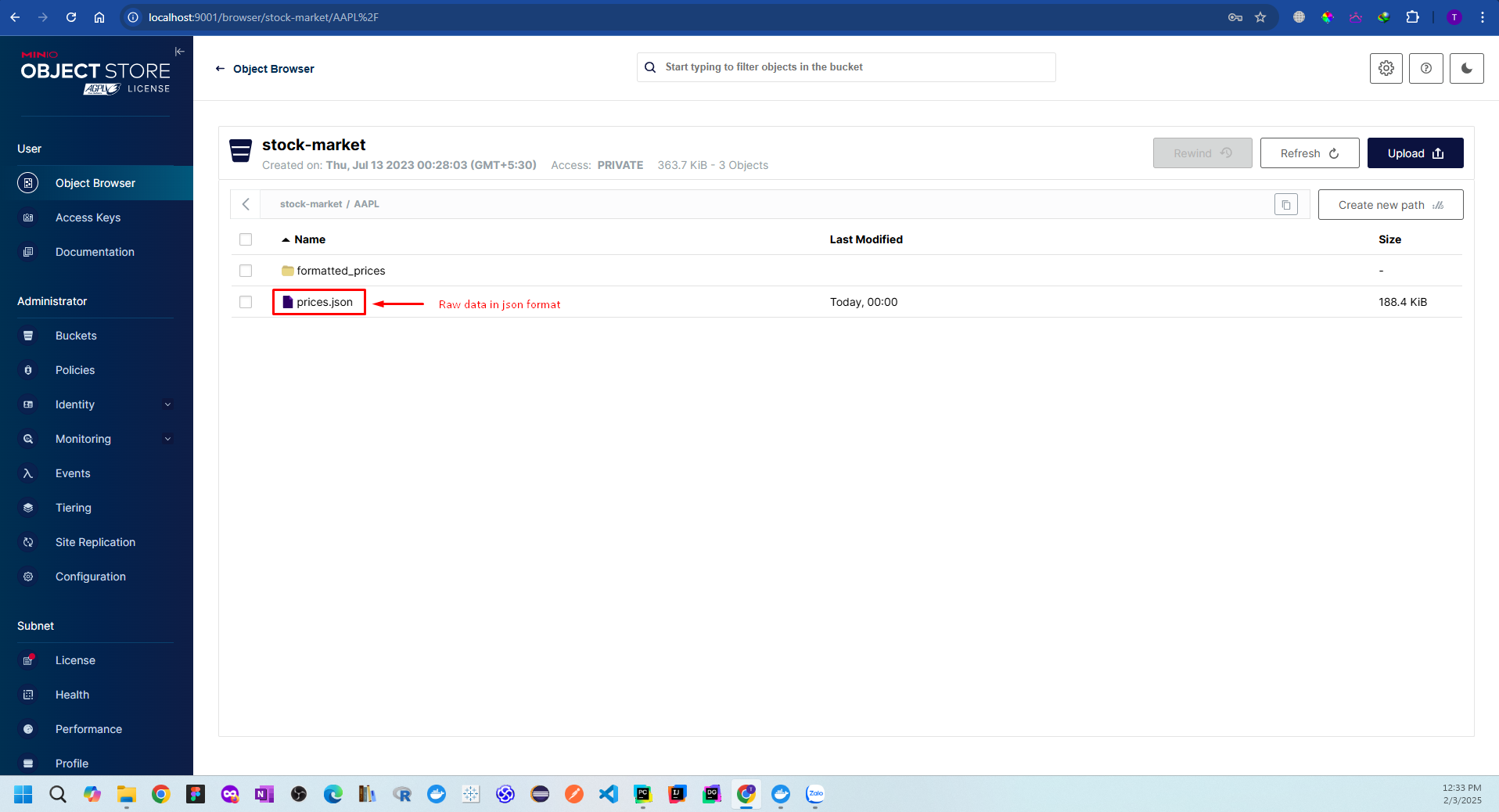
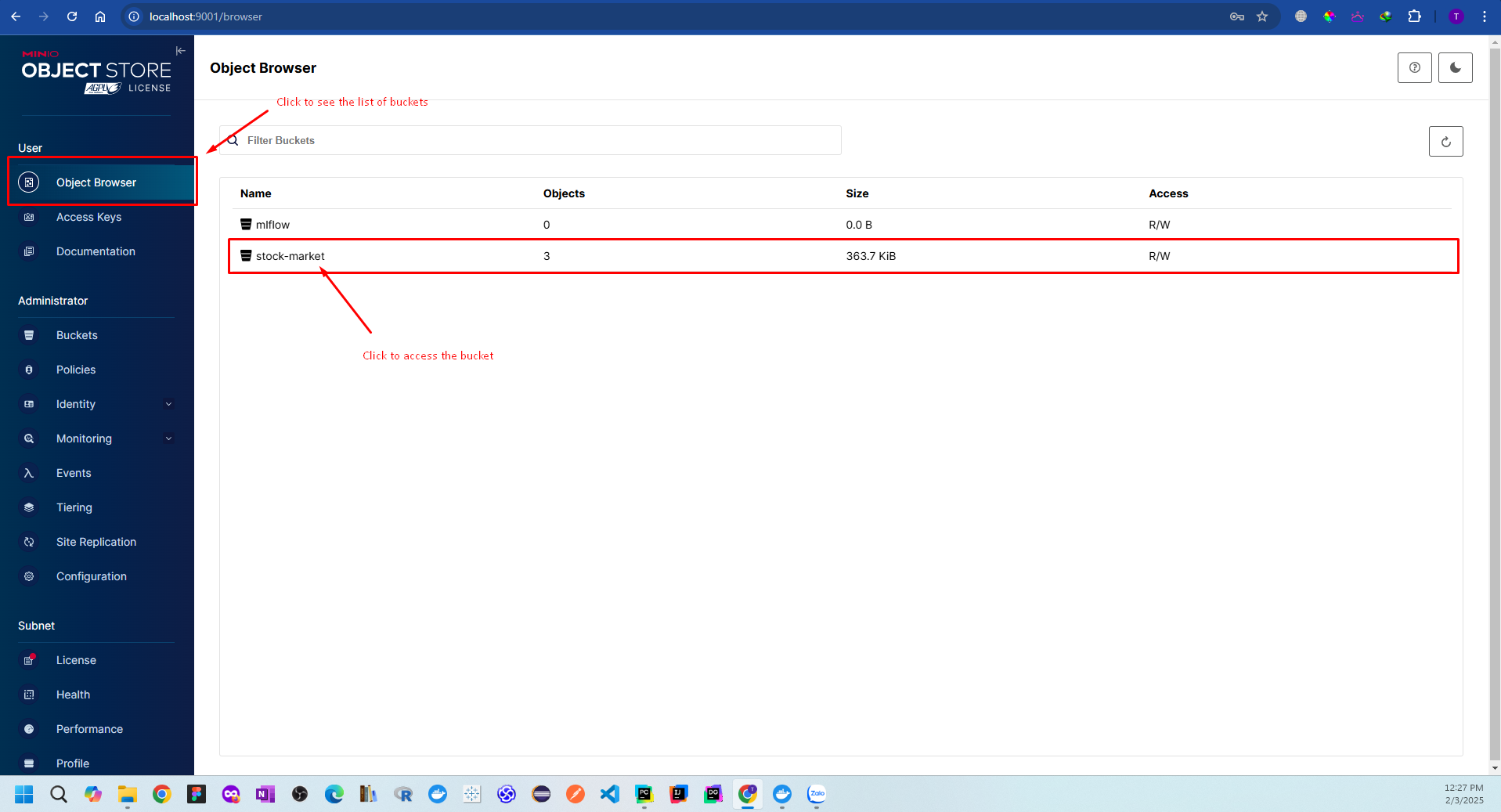
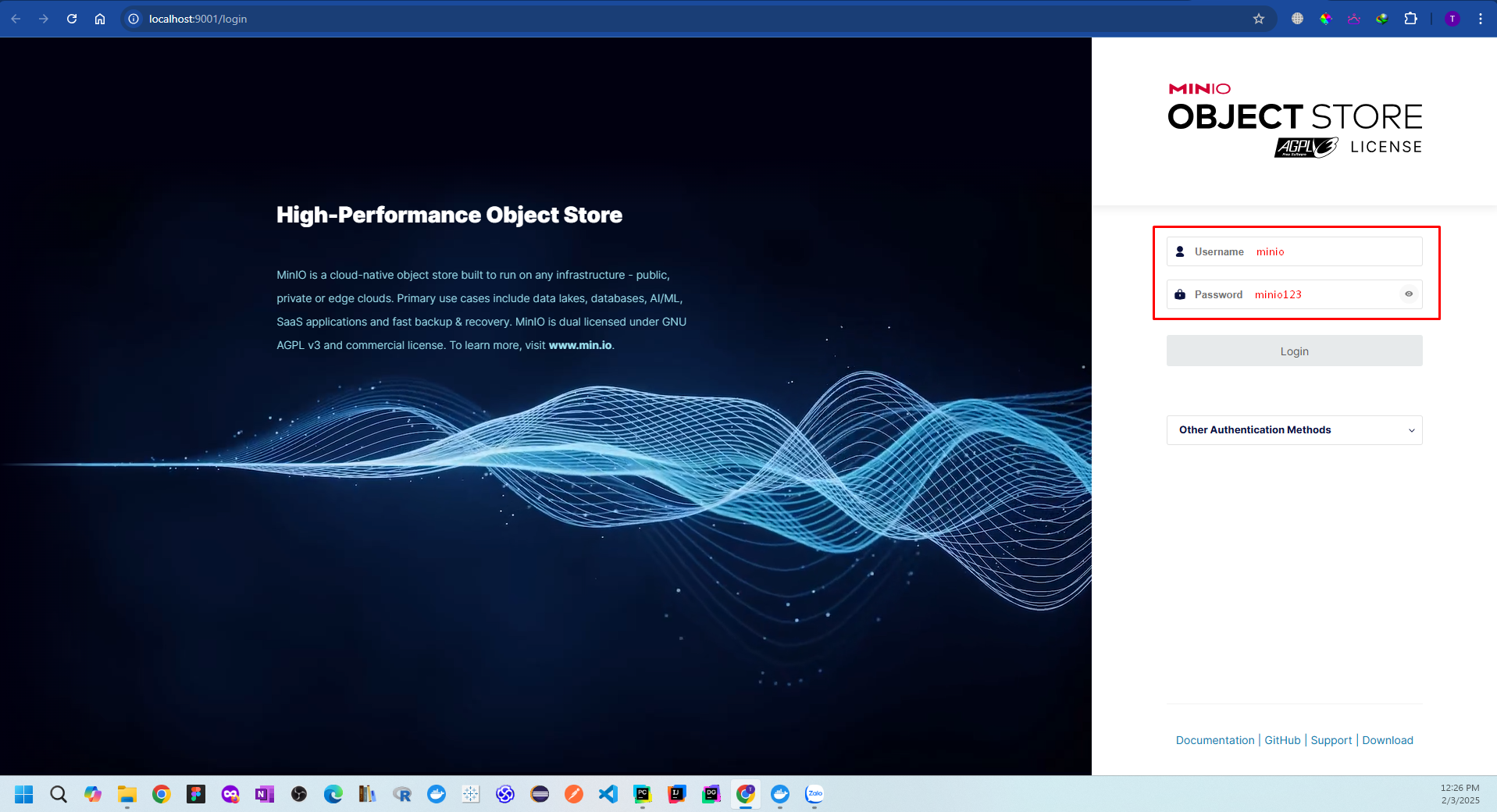
## Airflow Dashboard



**Mlflow**



## MinIO Dashboard



**Chapter 6**

# Conclusion

## Conclusion

In this project, we used the **ARIMA (Auto-Regressive Integrated Moving Average) model** to predict future stock prices using historical data obtained from **Yahoo Finance**. The key findings and insights from the project are as follows:

* **Data Preprocessing & Exploration**
  + We collected historical stock price data from Yahoo Finance and performed **time series analysis**.
  + The data was checked for **stationarity** using the **Augmented Dickey-Fuller (ADF) test**, and differencing was applied to make it stationary if needed.
* **Model Selection & Optimization**
  + We used **Auto ARIMA** and grid search techniques to find the best values for **p (Auto-Regressive), d (Differencing), and q (Moving Average)** parameters.
  + The model was trained using past stock prices and evaluated using metrics like **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and AIC/BIC scores**.
* **Prediction Results**
  + The ARIMA model successfully captured the trend and seasonality of stock prices.
  + However, **short-term predictions were more reliable** than long-term forecasts due to the high volatility of stock markets.
* **Limitations & Challenges**
  + **Stock prices are highly volatile** and influenced by external factors (news, macroeconomics, investor sentiment) that ARIMA cannot capture.
  + ARIMA is **best suited for univariate time series forecasting** and does not account for external variables like trading volume, interest rates, or market indicators.
  + The model assumes **linear relationships**, making it less effective for highly non-linear stock price movements.
* **Future Work & Improvements**
  + Incorporating **exogenous variables (ARIMAX model)** to improve prediction accuracy.
  + Exploring **LSTM (Long Short-Term Memory) or Transformer-based models** for better long-term forecasting.
  + Using **hybrid models** that combine statistical methods (ARIMA) with deep learning for enhanced performance.

## Final Thoughts

While ARIMA provides a solid baseline for stock price forecasting, it has limitations when dealing with highly volatile financial markets. For better predictions, integrating machine learning models, sentiment analysis, and macroeconomic factors would enhance forecasting accuracy.

**Chapter 7**

# References

[1] Box, G. E. P., & Jenkins, G. M. (1976). Time Series Analysis: Forecasting and Control.

[2] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice.

[3] Official documentation of Python libraries such as Statsmodels and Pandas.

[4] Historical AAPL stock price data from Yahoo Finance.