**MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE**

**NATIONAL TECHNICAL UNIVERSITY OF UKRAINE**

**" IHORY SIKORSKY KYIV POLYTECHNIC INSTITUTE"**

**Liubov Oleshchenko**

# **Statisical Methods Of ML**

# Laboratory Work 1

**K-Means clustering algorithm in Python**

**Kulubecioglu Mehmet**

**IM-14 FIOT**

**Class Number: 12**

**Kyiv**

**IHORY SIKORSKY KYIV POLYTECHNIC INSTITUTE**

**2024**

**Using K-Means Clustering Algorithm in Python**

## **1. Introduction**

This report presents an implementation of the K-Means clustering algorithm in Python to analyze and segment datasets. The analysis is performed on two datasets:

1. **Loan Application Dataset (clustering.csv)** – Clustering applicants based on income and loan amount.
2. **Wholesale Customers Dataset (Wholesale customers data.csv)** – Segmenting wholesale customers based on annual spending across different product categories.

## **2. Required Libraries and Tools**

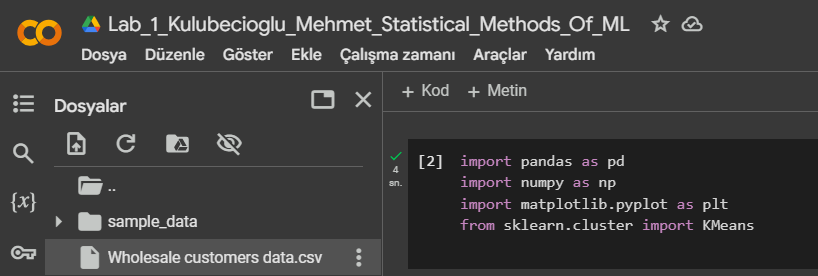
To implement the clustering algorithm, the following libraries were used:

* **pandas** for data manipulation.
* **numpy** for numerical computations.
* **matplotlib.pyplot** for data visualization.
* sklearn.cluster.KMeans for implementing the K-Means algorithm.
* **sklearn.preprocessing.StandardScaler** for data standardization.

Google Colab was used as the development environment.

## **3. Clustering Loan Applicants**

### **Step 1: Importing Required Libraries**

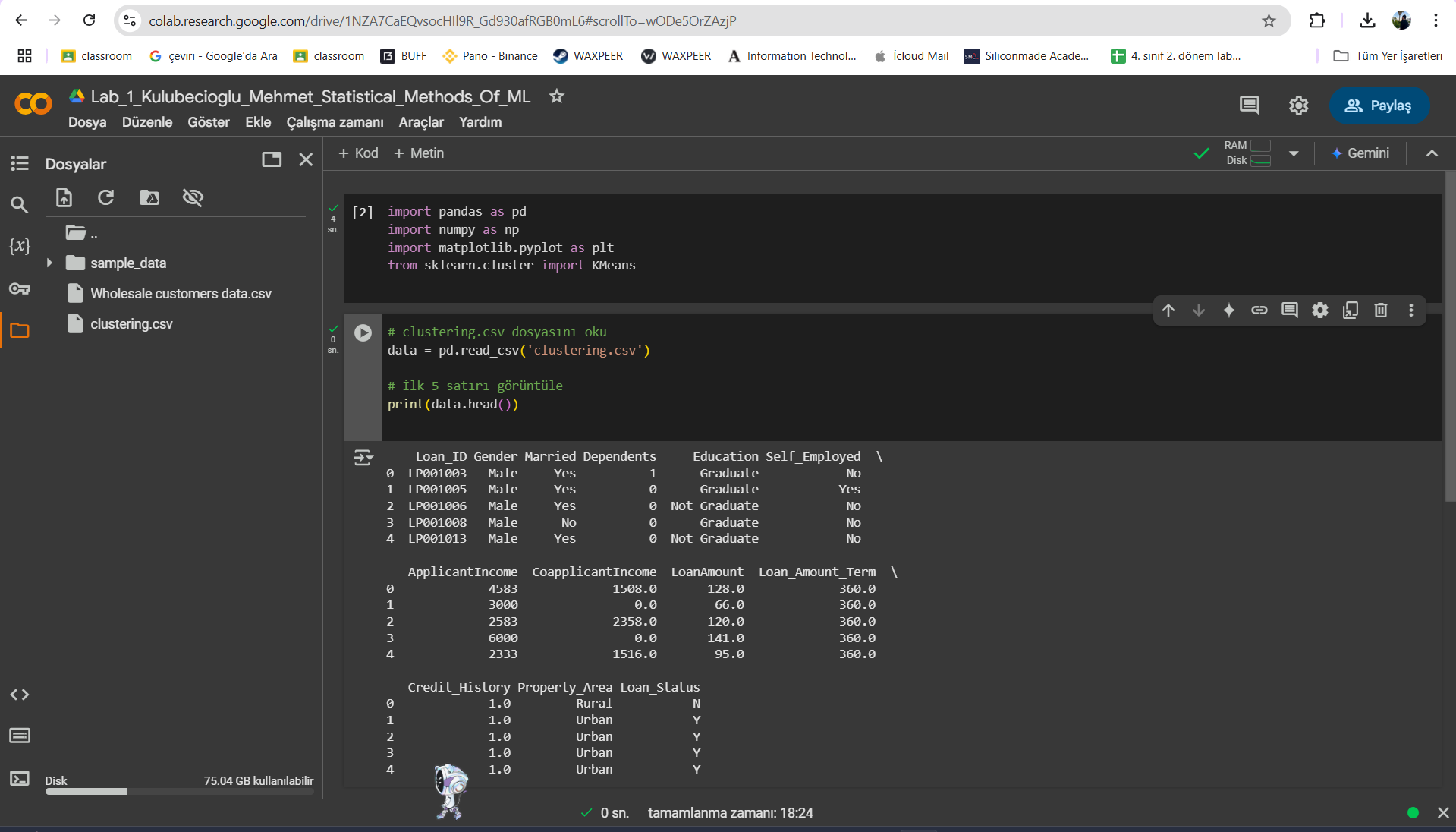


### 

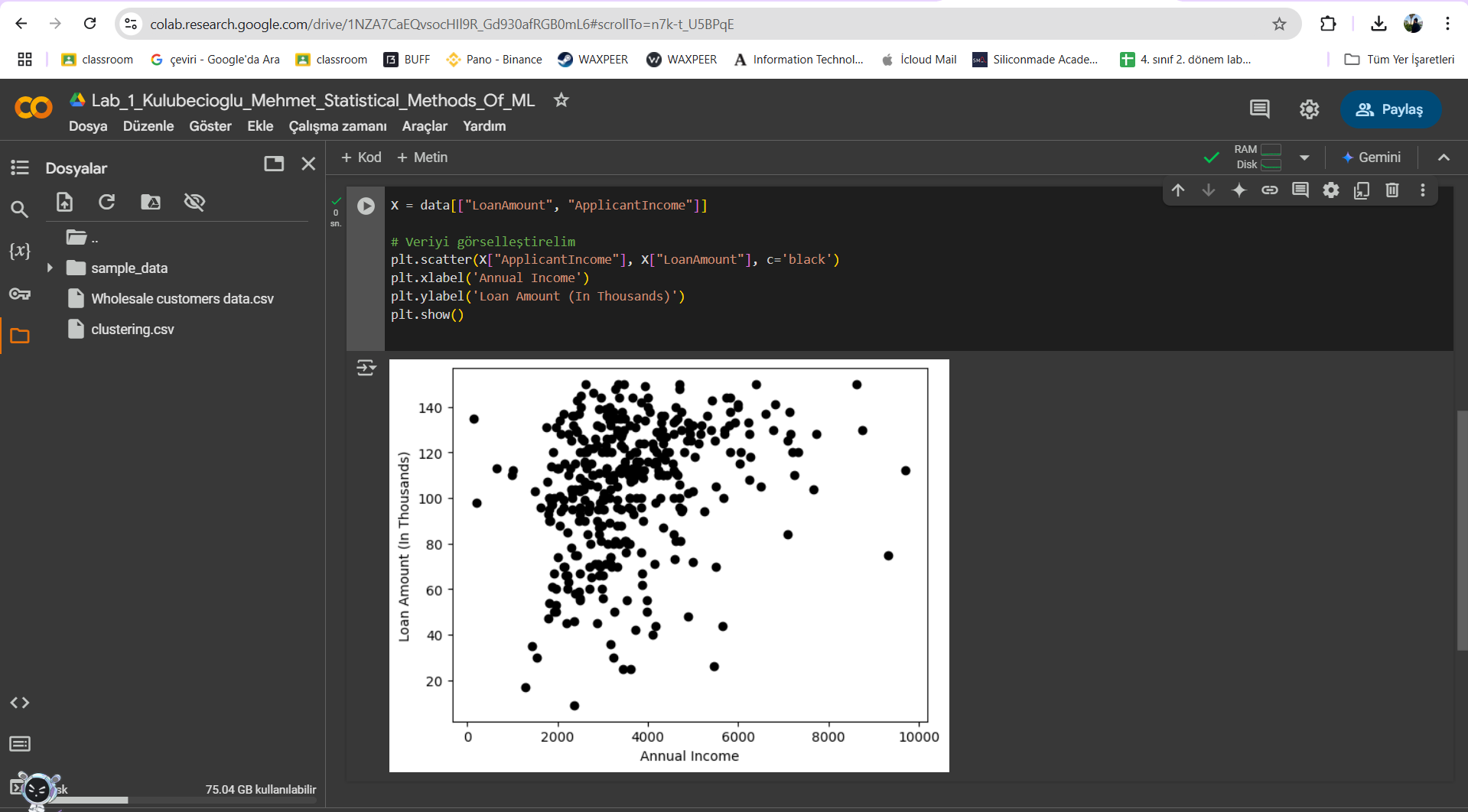
### 

### 

### **Step 2: Loading and Exploring the Dataset**



### **Step 3: Selecting Features and Visualizing Data**

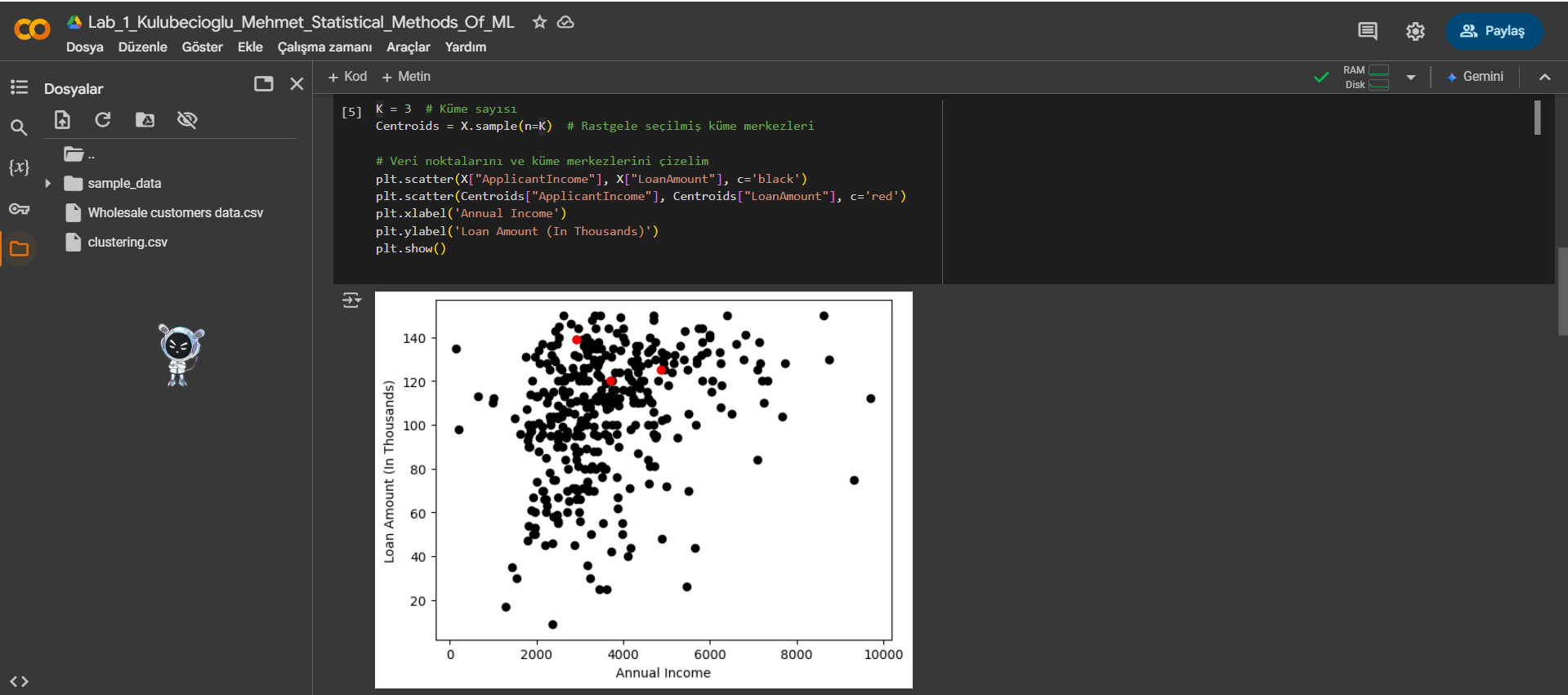


### 

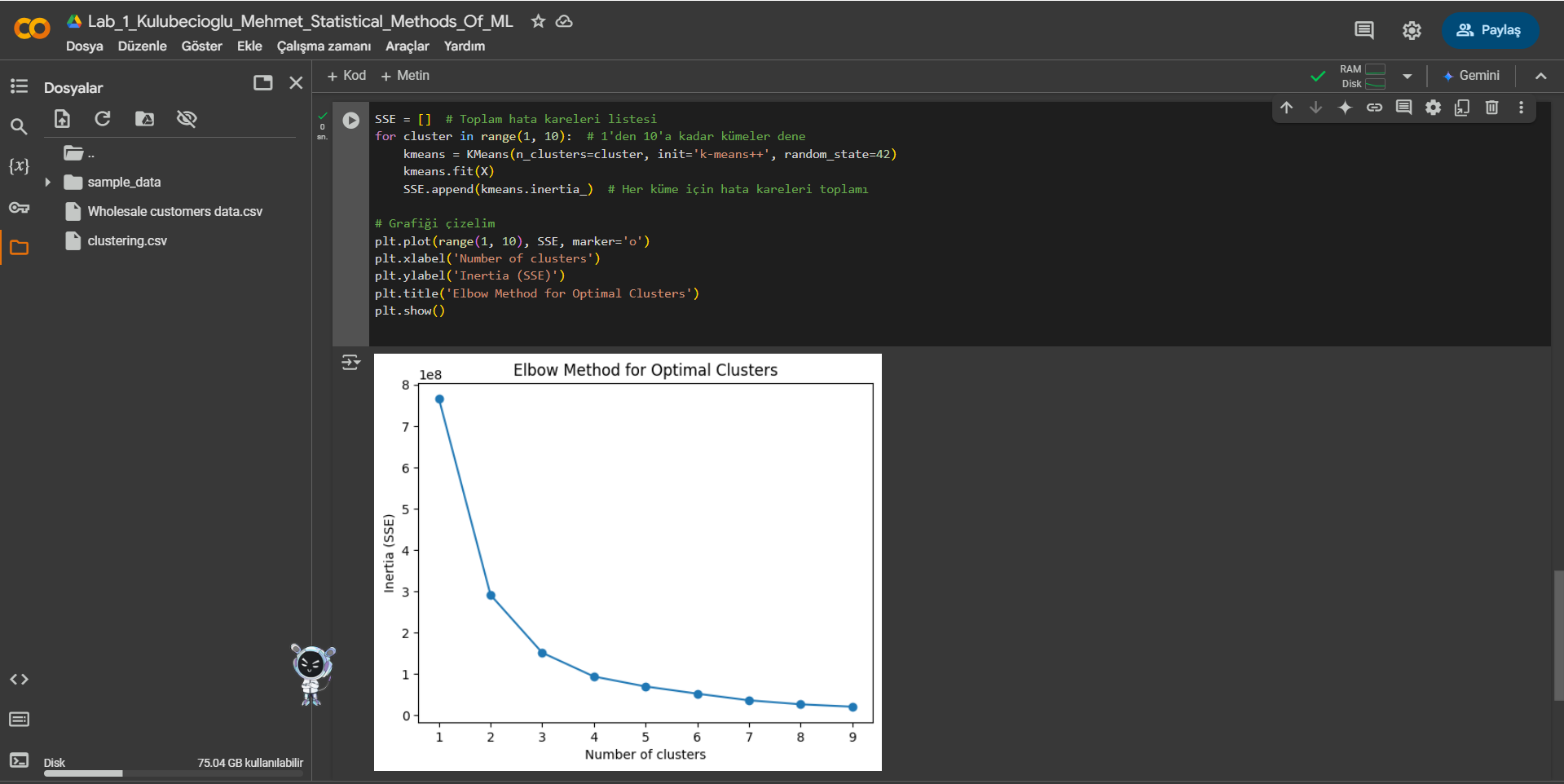
### 

### 

### **Step 4: Applying K-Means Clustering**

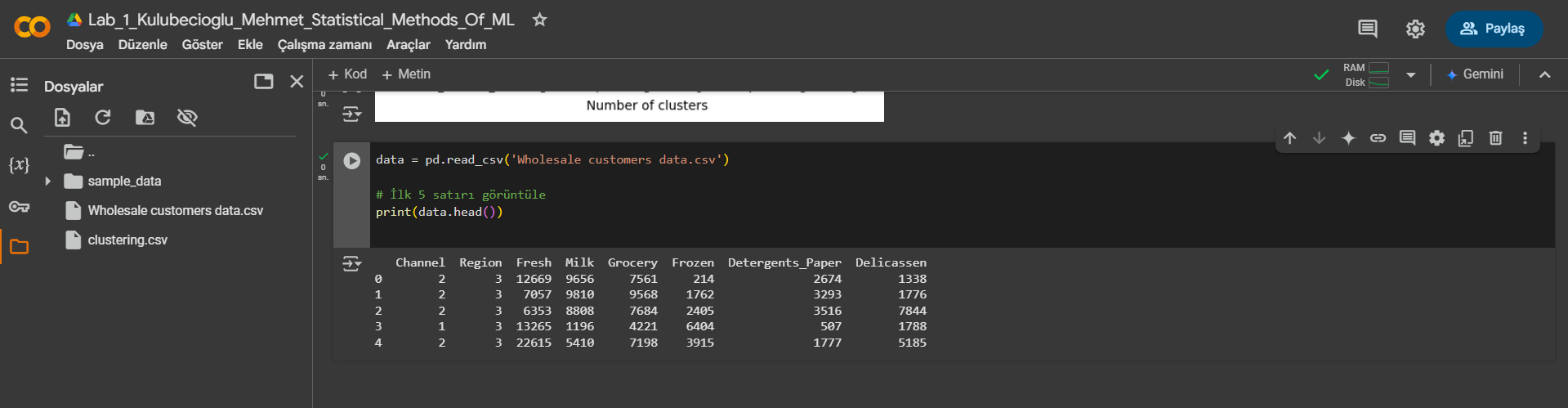


### **Step 5: Determining the Optimal Number of Clusters (Elbow Method)**

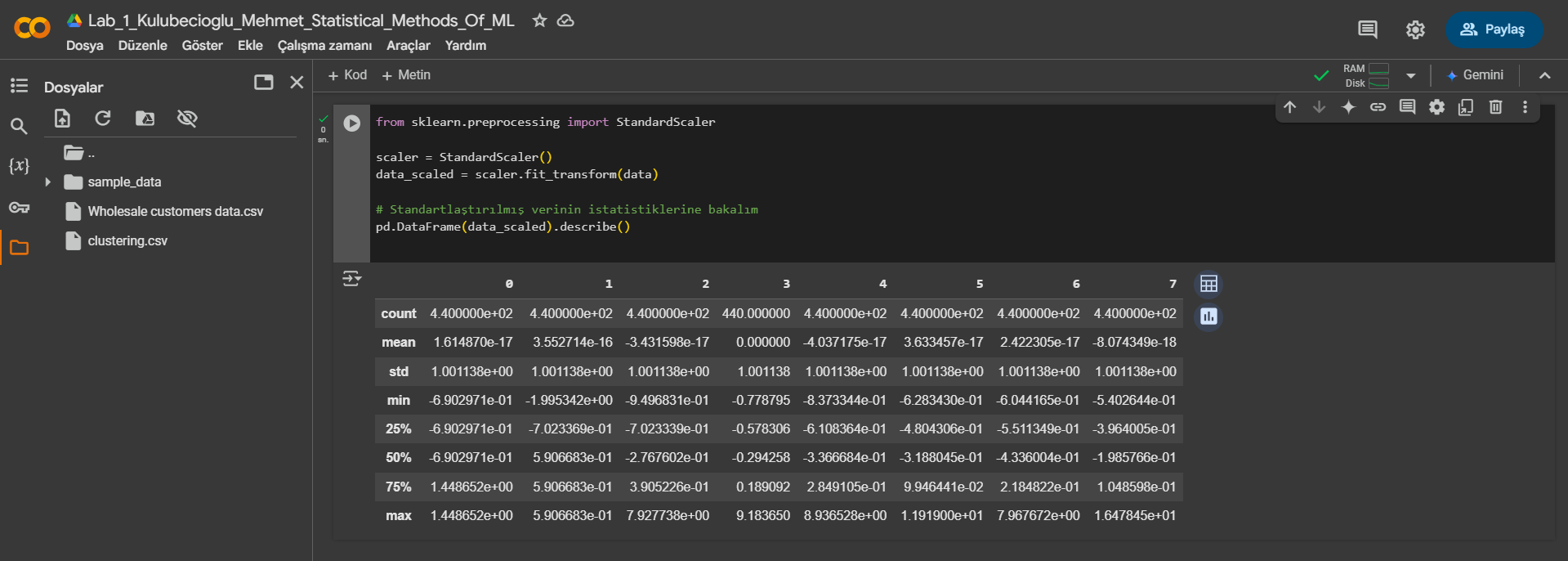


## **4. Clustering Wholesale Customers**

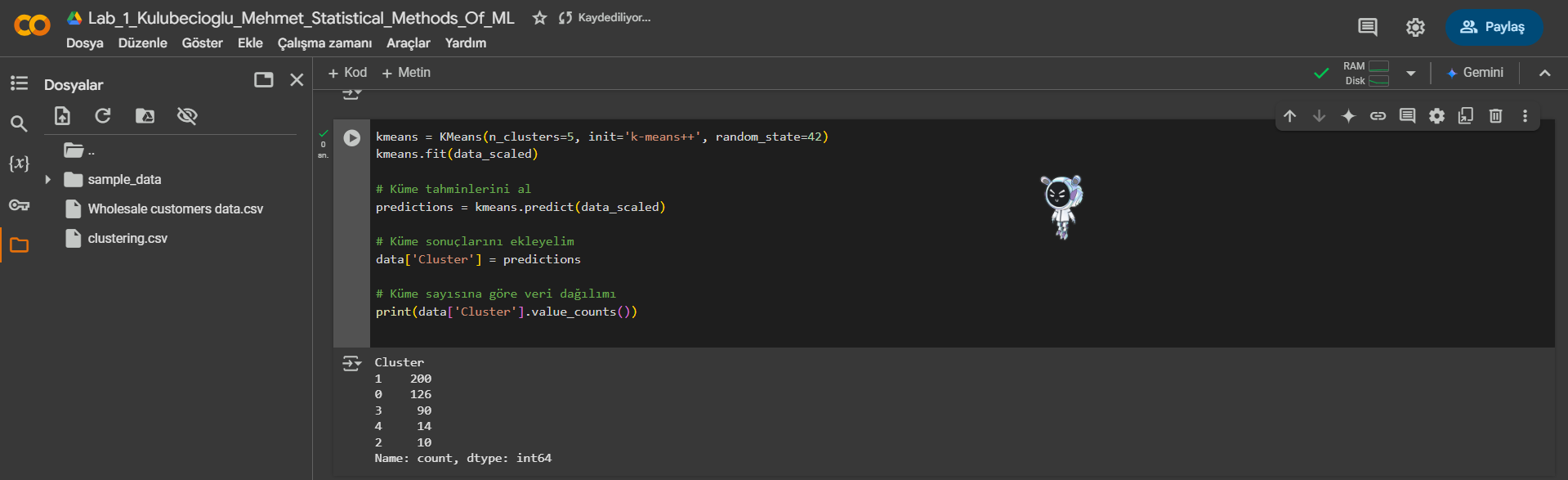
### **Step 1: Loading the Dataset**



### **Step 2: Standardizing the Data**

****

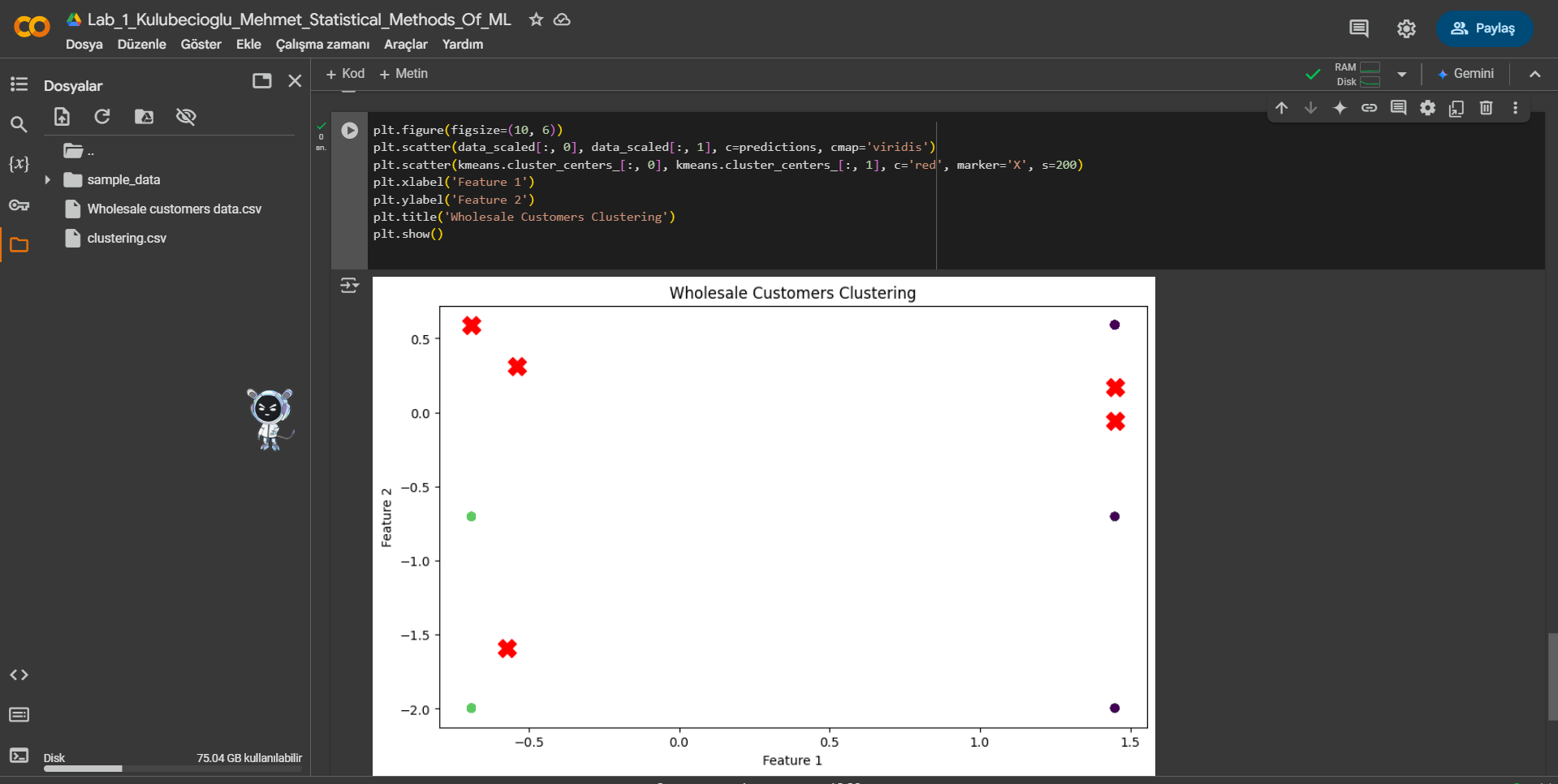
### **Step 3: Implementing K-Means Clustering**



### 

### 

### **Step 4: Visualizing the Clusters**



## 

## 

## 

## **5. Answers to Questions**

### **1. What is clustering?**

Clustering is an unsupervised learning technique that groups similar data points into clusters based on patterns in the data.

### 

### **2. What properties of clusters do you know?**

* **Homogeneity within clusters:** Data points within a cluster should be similar.
* **Heterogeneity between clusters:** Data points in different clusters should be as different as possible.

### **3. What applications of clustering in real scenarios do you know?**

* Customer segmentation
* Document clustering
* Image segmentation
* Recommendation systems

### **4. What clustering evaluation metrics do you know?**

* **Inertia (Sum of Squared Errors - SSE)**
* **Dunn Index**
* **Silhouette Score**

### **5. What is K-Means Clustering?**

K-Means is a clustering algorithm that partitions data into k clusters, minimizing the variance within each cluster.

### **6. How to choose the right number of clusters in K-Means?**

The **Elbow Method** is commonly used, where we plot the inertia for different k values and select the optimal number where the curve bends.

### **7. What is the K-Means++ algorithm used for?**

K-Means++ improves the initialization of centroids to avoid poor clustering results.

### **8. How to implement K-Means clustering algorithm and K-Means++ algorithm for centroid initialization in Python?**

Using **sklearn.cluster.KMeans**, we specify **init='k-means++'** while creating the model.

### **9. What is data standardization used for?**

Standardization scales data to have a mean of 0 and a standard deviation of 1, ensuring that features contribute equally to clustering.

### **10. What clustering algorithms do you know?**

* **K-Means**
* **Hierarchical Clustering**
* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**
* **Gaussian Mixture Models (GMM)**