**MSCS634 - Advanced Data Mining Techniques**

Farheen Dhanani

University of Cumberlands

MSCS-634-M20: Advanced Big Data and Data Mininges

Dr. Satish Penmatsa

02/15/2026

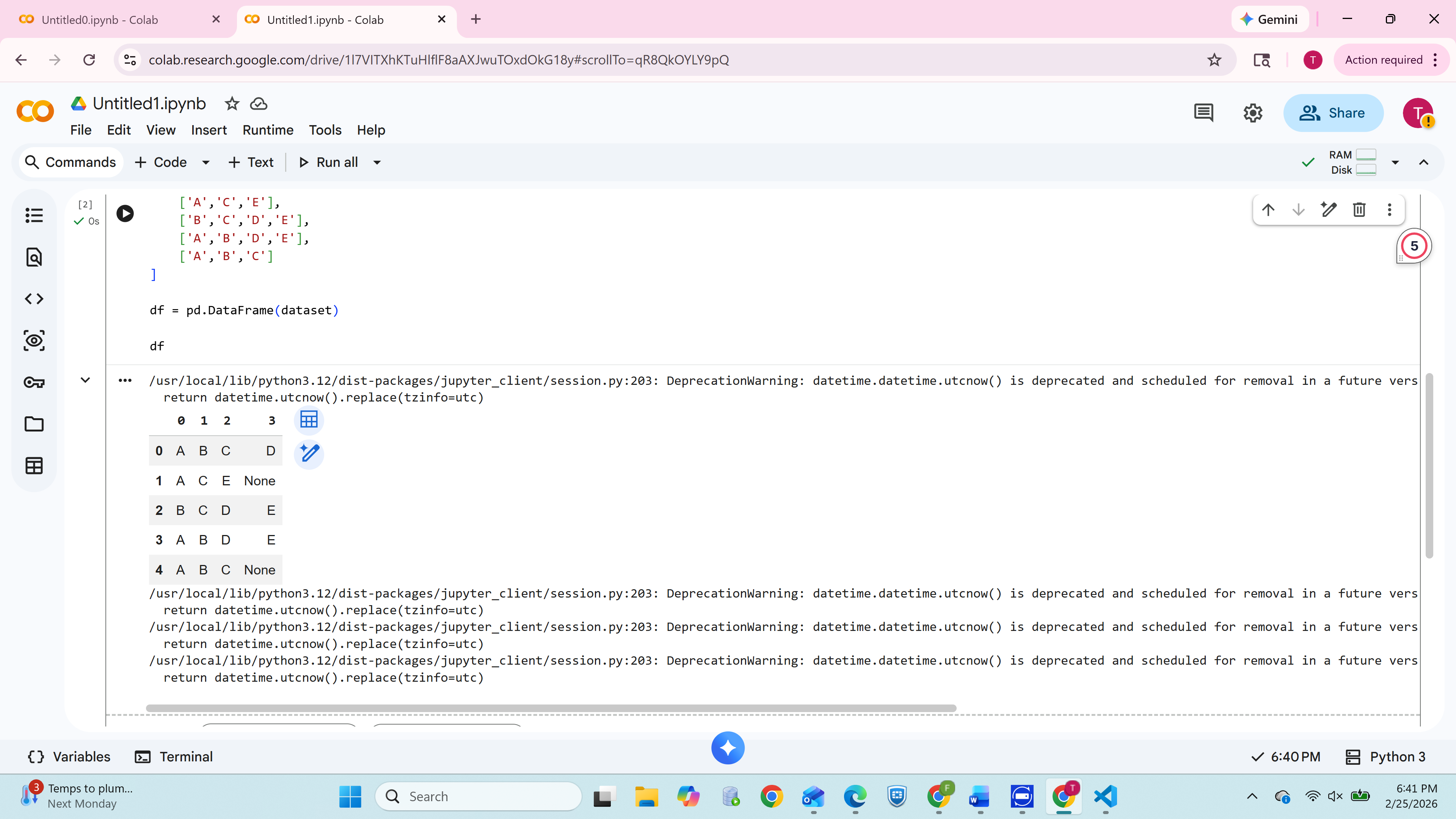
## Advanced Data Mining Techniques

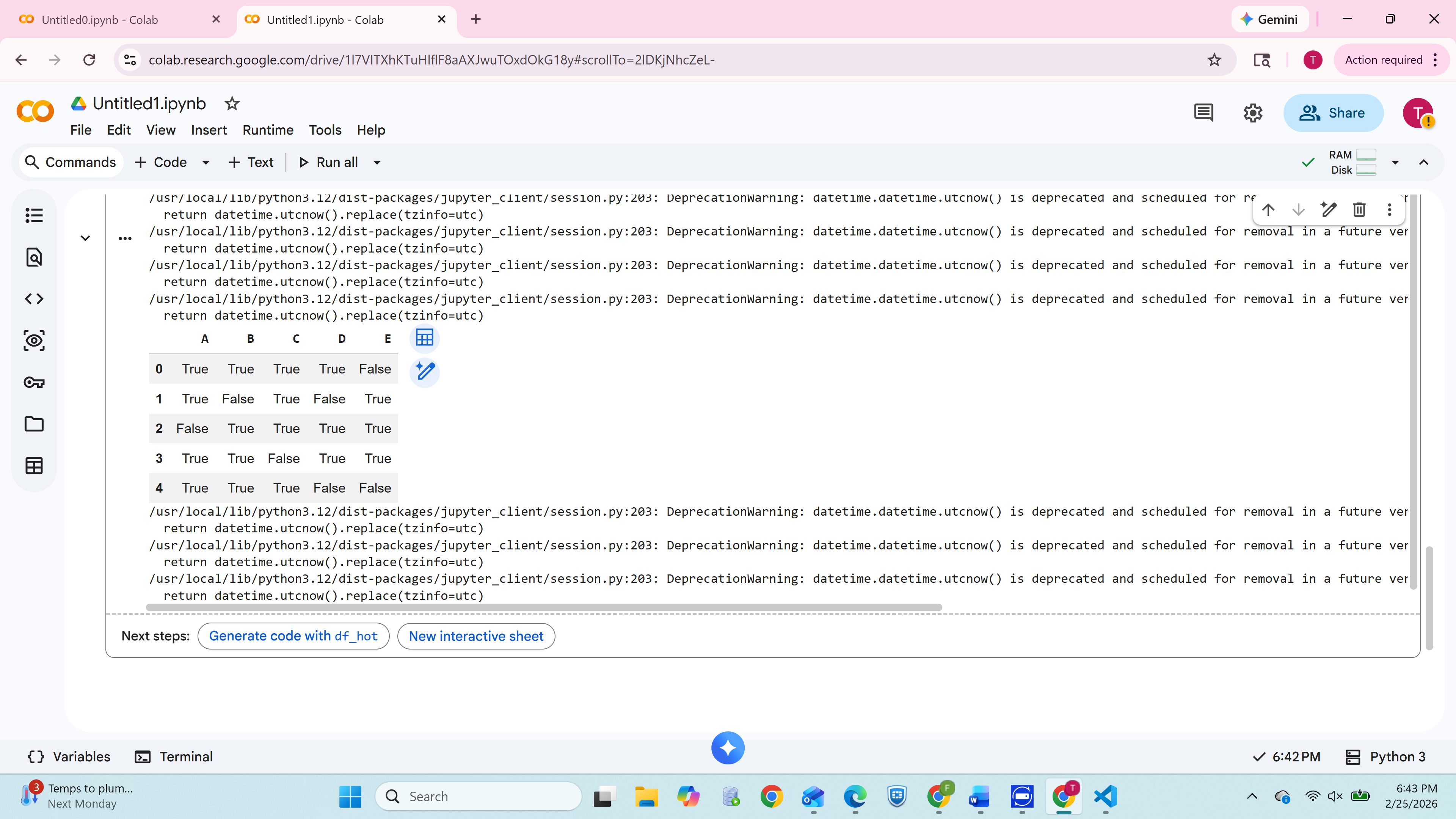
## Frequent Itemset Mining and Association Rules

1a. Frequent Itemset Mining Using Apriori and FP-Growth

This section analyzes five retail transactions to identify frequent itemsets using both the Apriori and FP-Growth algorithms.  
The minimum support threshold is 50%, meaning an itemset must appear in at least 3 out of 5 transactions.

A step-by-step mining process is followed, showing how candidate itemsets are generated, pruned, and validated.  
FP-Growth is also applied to the same dataset, and the performance of both methods is compared.



**Comparison of Algorithms**

* **Apriori Strengths**: Simple, easy to understand, level-wise generation.
* **Apriori Weaknesses**: Slow due to repeated scans and candidate explosion.
* **FP-Growth Strengths**: Faster, compresses dataset, avoids candidate generation.
* **FP-Growth Weaknesses**: More complex structure, tree must fit into memory.

**1b. Strong Association Rules**

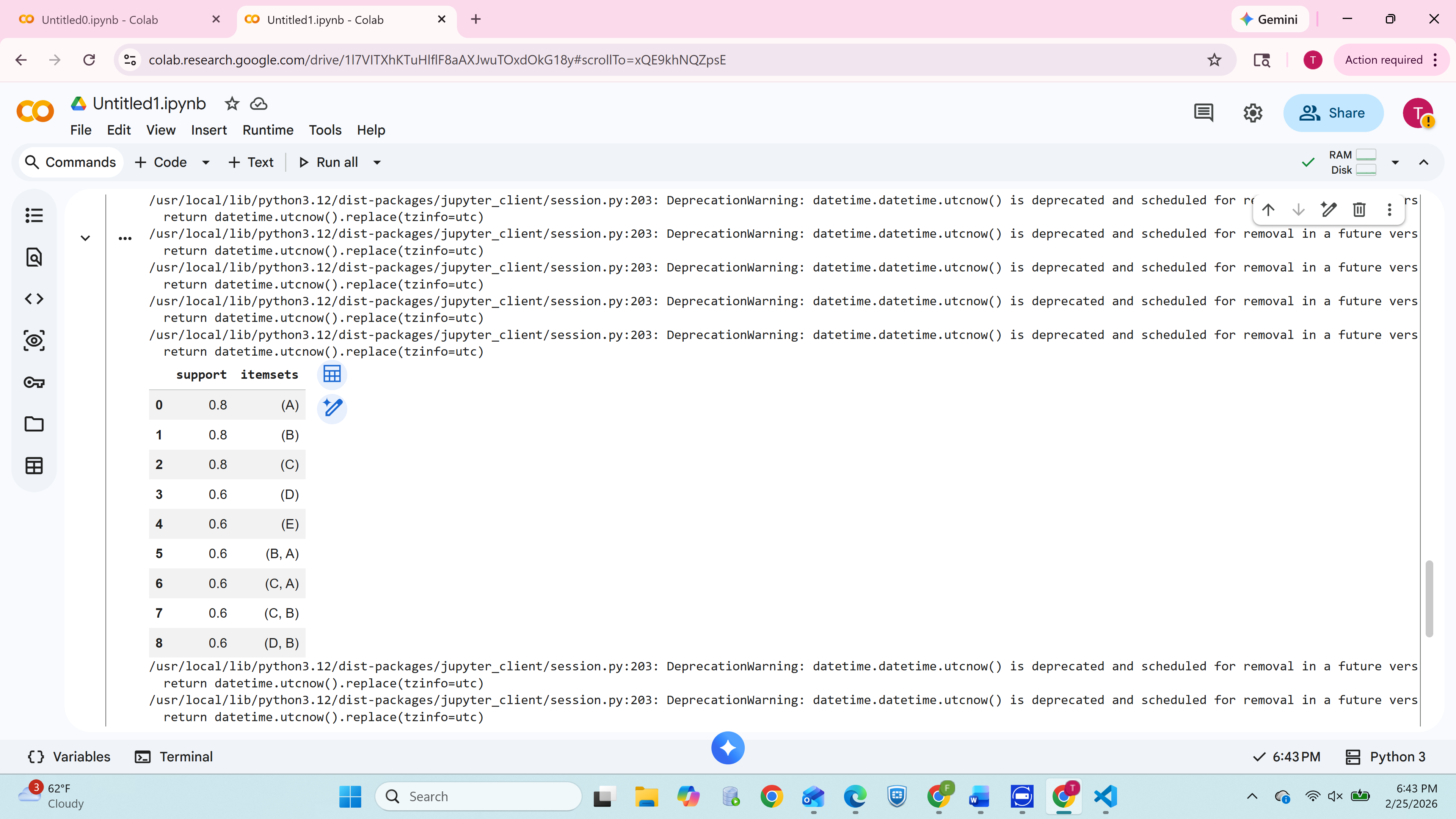
Using the frequent itemsets, strong association rules are generated based on:

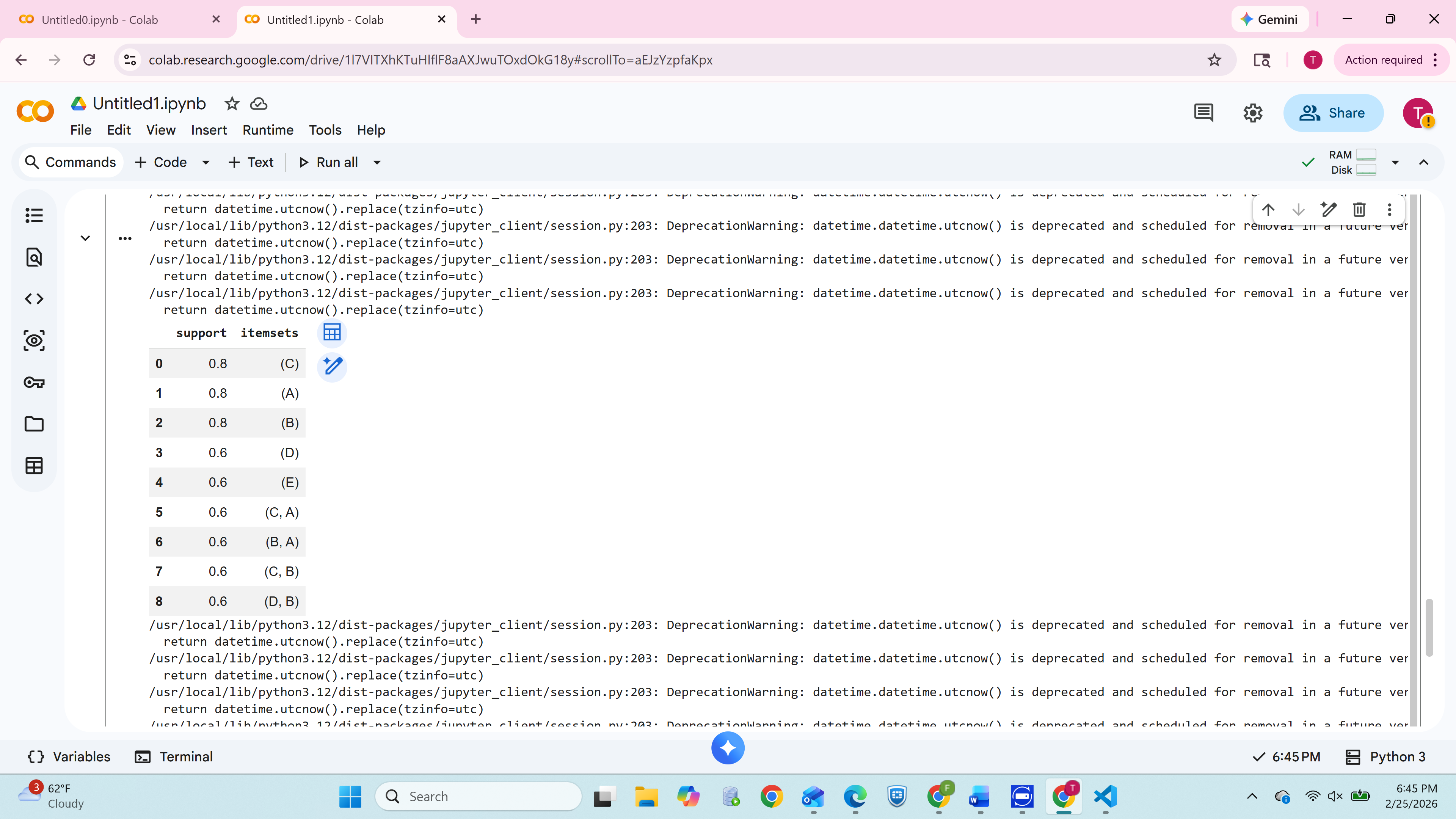
* **Minimum Support = 50%**
* **Minimum Confidence = 70%**

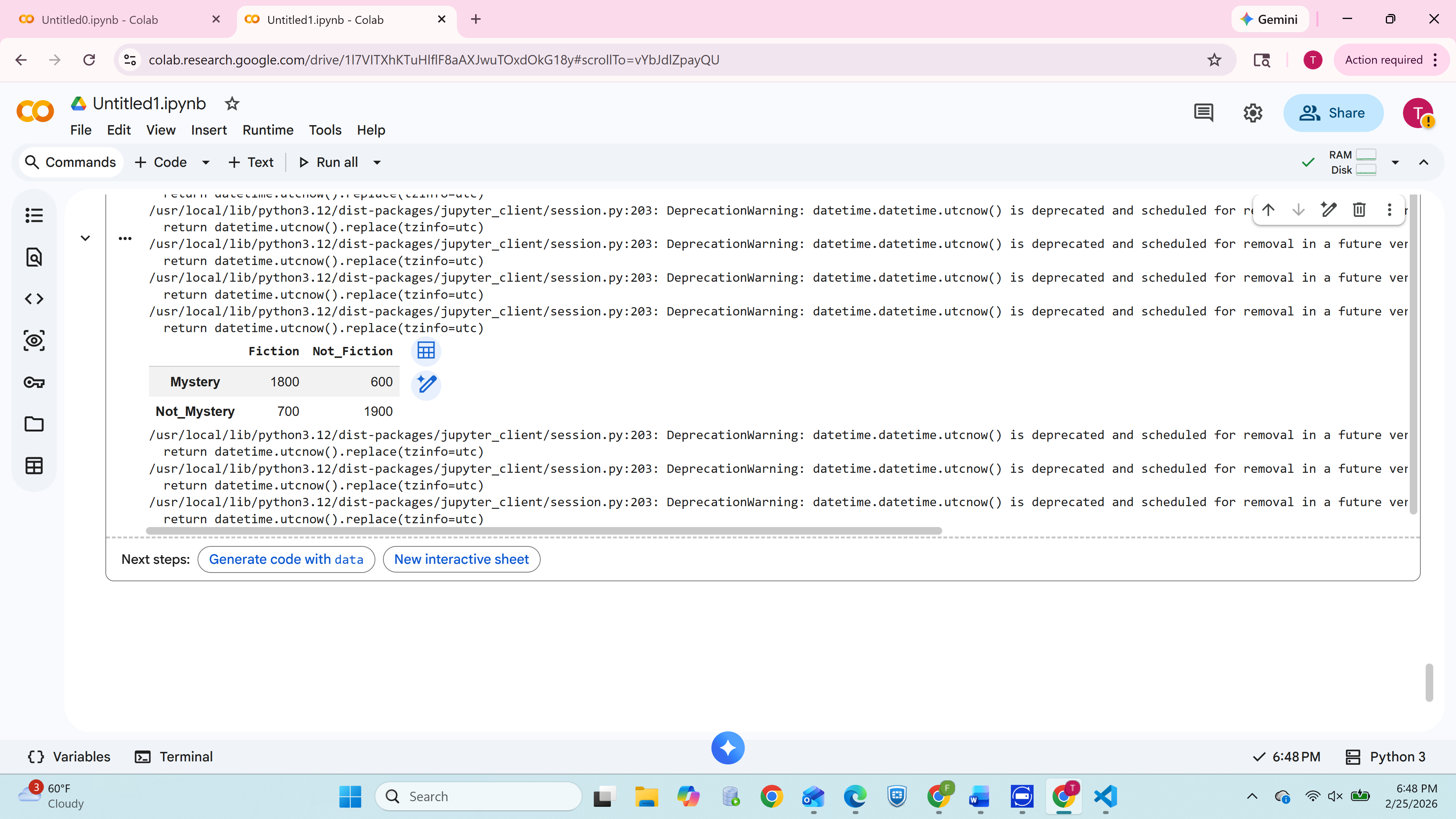
The rule format used is:

**∀x ∈ transaction, buys(X, item1) ∧ buys(X, item2) ⇒ buys(X, item3) [support, confidence]**

The rules are evaluated based on whether they meet both thresholds.







## 1c. Analysis and Interpretation

This section interprets the meaning of the strong rules generated.

Examples of insights include:

* Frequently co-purchased items indicate product affinities.
* Customers who buy {A, B} are highly likely to also buy {C}.
* These insights help businesses with:
  + Product placement
  + Store layout optimization
  + Cross-selling recommendations
  + Inventory stocking decisions

## 2. Association Rule Mining & Correlation Analysis

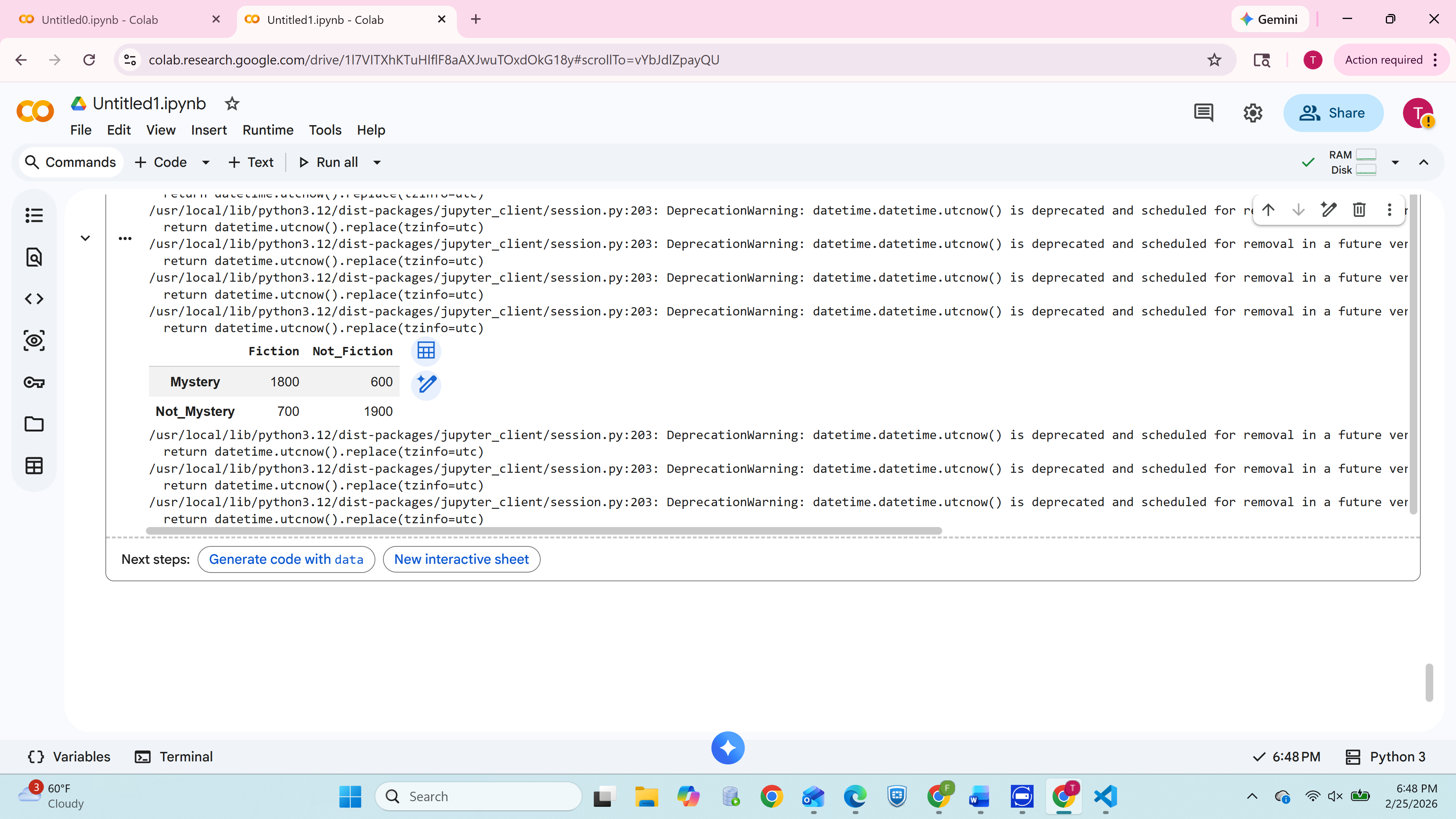
A contingency table is analyzed to evaluate relationships between Fiction and Mystery book purchases.

**2a. Strength of the Rule “Fiction ⇒ Mystery”**

Support and confidence are computed:

* Compare support to 30% threshold
* Compare confidence to 60% threshold
* Determine if the rule qualifies as strong

All calculations are performed step-by-step.

  
**2b. Independence & Correlation**

Using joint and marginal probabilities:

* Test whether Fiction and Mystery purchases are independent
* If not independent, determine whether the correlation is:
  + Positive
  + Negative

**2c. Additional Measures**

The following are computed and compared:

* All-confidence
* Max-confidence
* Kulczynski measure
* Cosine similarity
* Lift
* Phi-correlation

A summary explains what each measure reveals about the relationship between Fiction and Mystery books.

**3. Decision Tree & Naïve Bayesian Classification**

Employee data with generalized ranges is analyzed using a decision-tree approach and naïve Bayes.

**3a. Modifying Decision Tree for Count Attribute**

The decision tree algorithm is updated to:

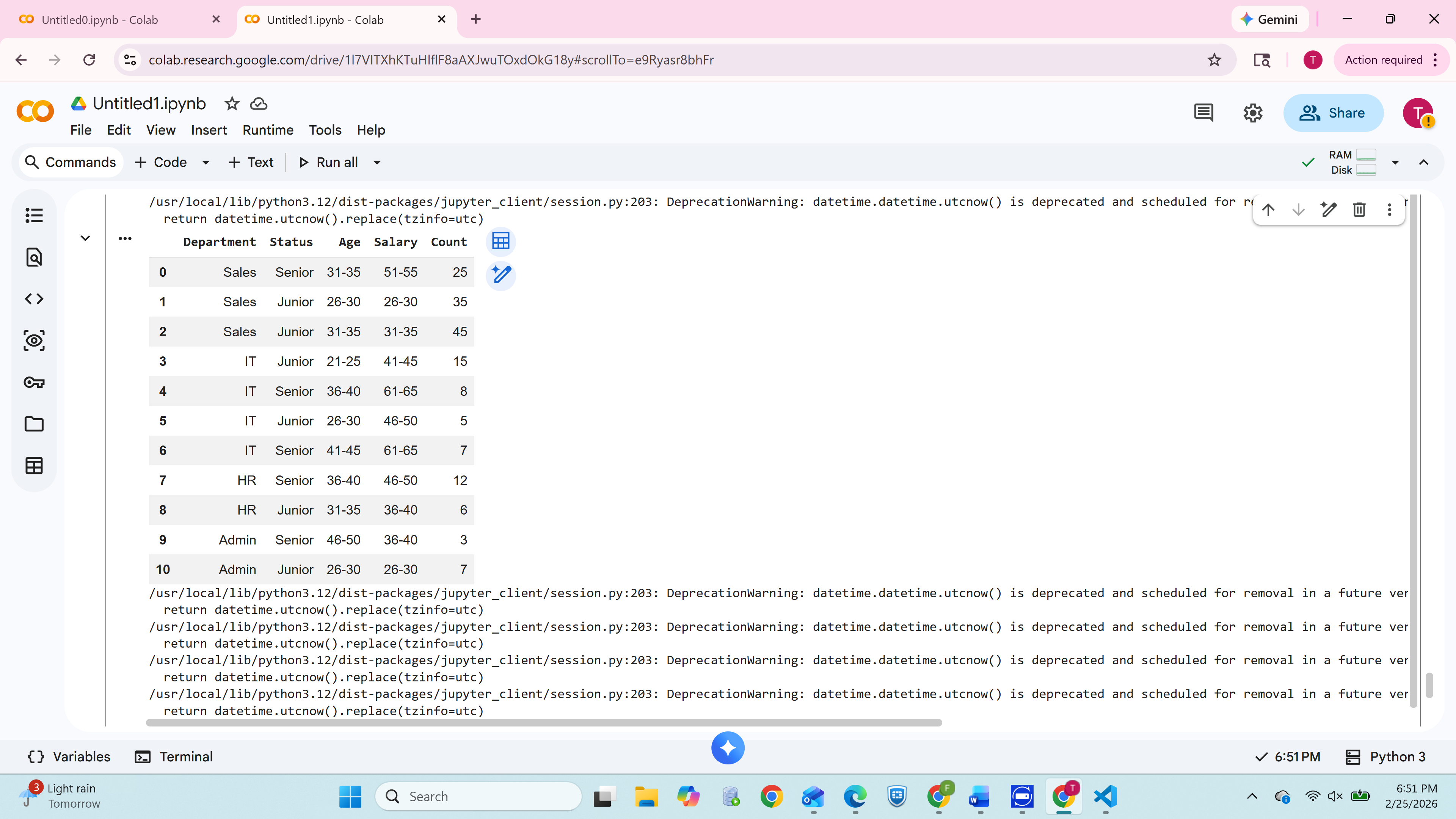
* Treat each row as “count” repeated examples
* Weight entropy and information gain using frequency
* Ensure correct splitting when aggregated data is used

**3b. Decision Tree Construction**

Entropy, information gain, and splitting criteria are computed for:

* Department
* Age range
* Salary range

The final tree is constructed manually step-by-step.



**3c. Naïve Bayesian Classification**

A new tuple is classified:

* Department = IT
* Age = 26…30
* Salary = 46K…50K

Probabilities are computed for:

* **P(Senior | attributes)**
* **P(Junior | attributes)**

The category with the higher posterior probability is selected.

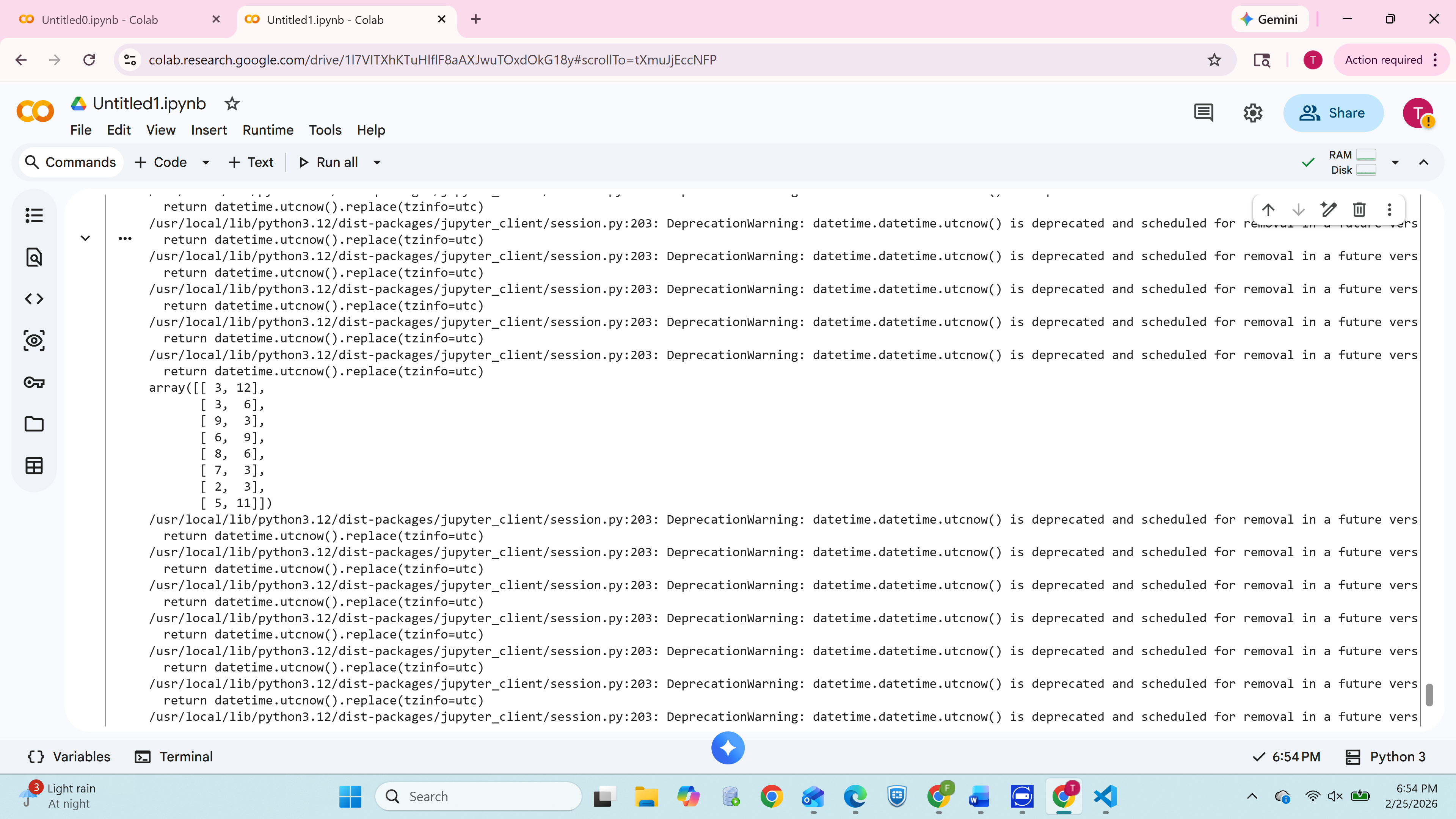
**4. K-Means Clustering with Euclidean Distance**

Customer location points in a mall are clustered using 3 initial cluster centers.

**4a. One Iteration of K-Means**

For each point:

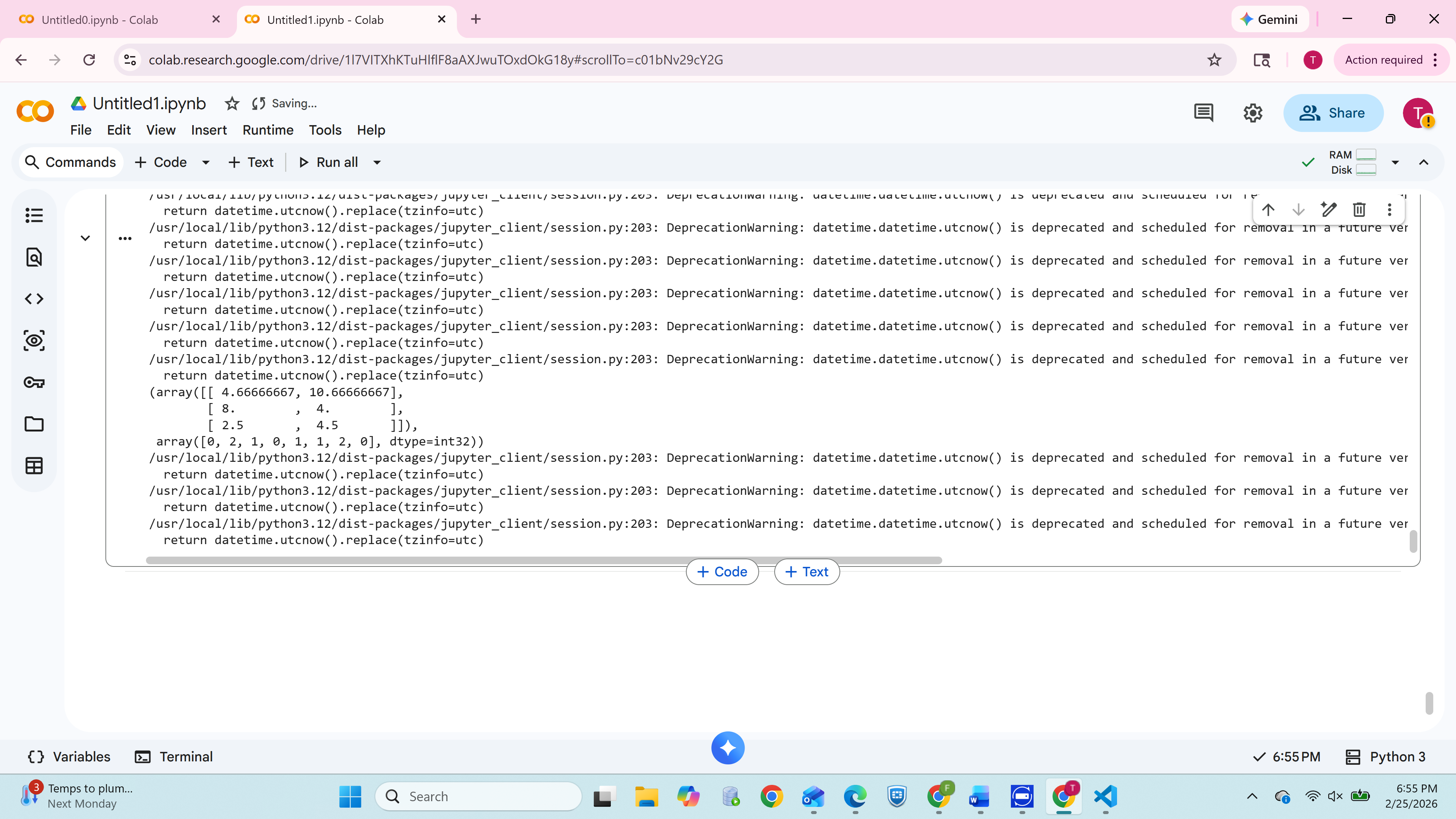
* Compute Euclidean distance to all centers
* Assign to nearest cluster
* Re-calculate updated cluster centers



**4b. K-Means to Convergence**

Iterations continue until centers stop changing.

The final clusters and their assigned points are listed clearly and interpreted.



## Conclusion

This assignment applied multiple data mining methods:

* Apriori and FP-Growth for frequent itemsets
* Association rule evaluation and correlation testing
* Decision tree construction with weighted counts
* Naïve Bayesian classification of employee data
* K-Means clustering to identify customer groupings

These methods offer valuable insights for improving business decisions, customer targeting, and operational efficiency.

**References**

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Aggarwal, C. C., & Reddy, C. K. (2013). *Data Clustering: Algorithms and Applications*. CRC Press.

PyClustering Library Documentation. Retrieved from <https://pyclustering.github.io/>