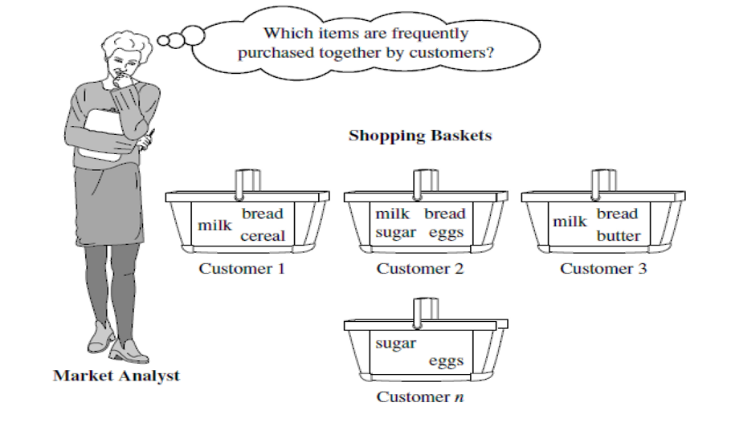
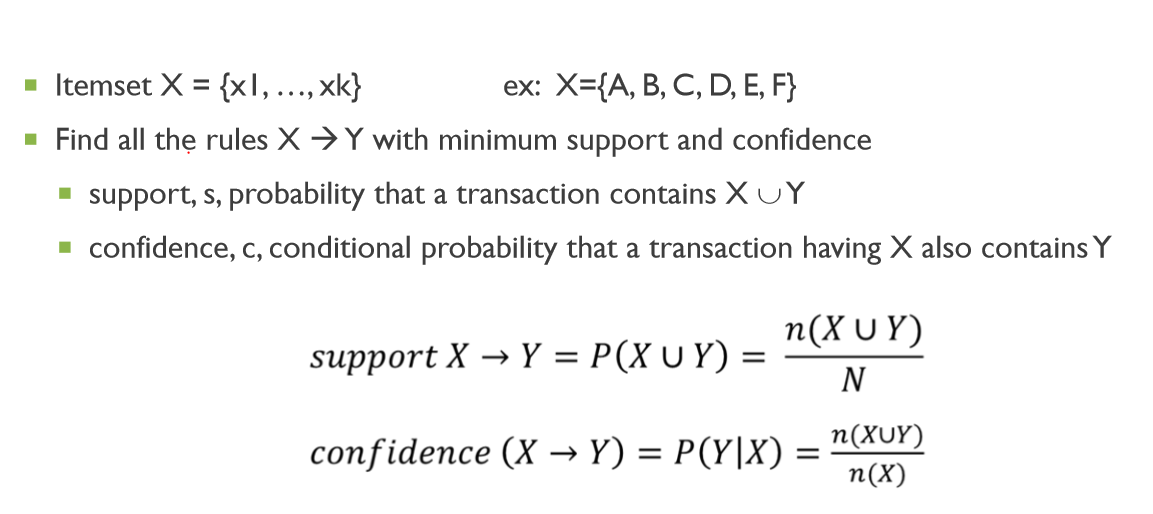


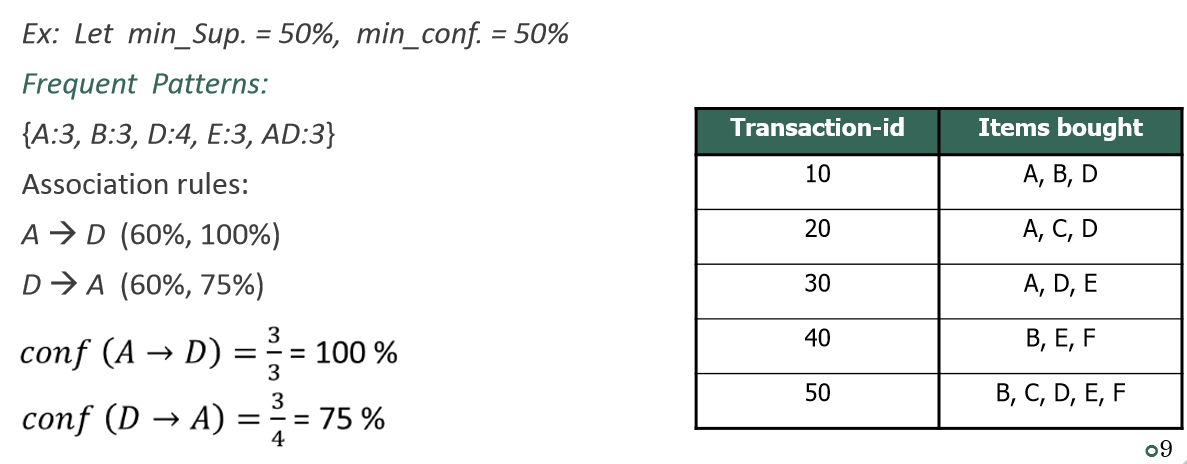
Mining Frequent Pattern, Association& Correlation

# Frequent Pattern

* **a** pattern **(a set of items, subsequences, substructures, etc.) that occurs** frequently **in a data set**
* **A Big role for mining association and correlation**
* **بس ليه بدور على الأشكال دى في الداتا أيه الحافز؟؟**
* Motivation**:** Finding inherent regularities in data
* **لما تلاقى ان السؤال بتاعك شبه الأسئلة الجاية يبقى فعلا محتاج الموضوع ده**
  + What products were often purchased together?— Beer and diapers?!
  + What are the subsequent purchases after buying a PC?
  + What kinds of DNA are sensitive to this new drug?
  + Can we automatically classify web documents?
* **و بيخش في مجموعة من التطبيقات اللى في الحياة**
* Applications: **Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis**
* Frequent Pattern are itemsets that appear frequently in a data set (e.g. Transaction record)
* Items that are frequently associated (e.g purchased) together can be represented as association rules**.Association Rule like =>**
* Computer → antivirus\_SW [Support = 2% , Confidence =60%]
* **معنى كده ان شراء الكمبيوتر بيأثر على شراء الانتى فيروس** باجمالى 2% من بياعة المنتجات **و** نسبة 60% من اللى اشتروا كمبيوترات اشتروا كمان انتى فيروس بس مش العكس ؛؛ احتمال شرطي أفتكر دكتور شريف
* **Support** and **Confidence** are measures of rule interestingness
* **و يعتبروا هم اللى بقارن بيهم الثريشولد بتوعى و بختبر بيهم مدى القرار على المنتجات و هو اللى بيعرفنى الباترن**
* **طبعا فيه ثريشولد أكتر من دول بس المحاضرة بنتكلم عن 2**
* 2% Support means 2% of Transactions Show that computers and antivirus\_SW are bought Together
* 60% Confidence means 60 % of customers who bought a computer also bought antivirus\_SW

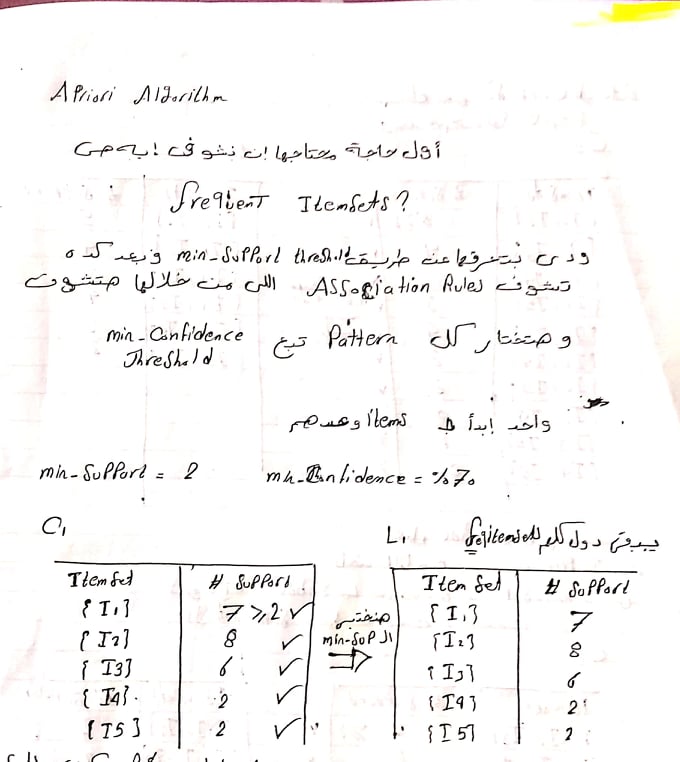
* Basics About **Association Rules**:
  + If frequency of itemset I satisfies min\_support count then I is a frequent itemset
  + Support(x) >= min\_support => x is frequent item set
  + If a rule satisfies min\_support and min\_confidence thresholds,it is said to be strong
    - problem of mining association rules reduced to mining frequent itemsets
  + Support(X)>= min\_support && Conf(X)>= min\_confidence => strong Association
  + Association rules mining becomes a two-step process:
    - Find all frequent itemsets that occur at least as frequently as a predetermined min\_support count
    - Generate strong association rules from the frequent itemsets that satisfy min\_support and min\_confidence

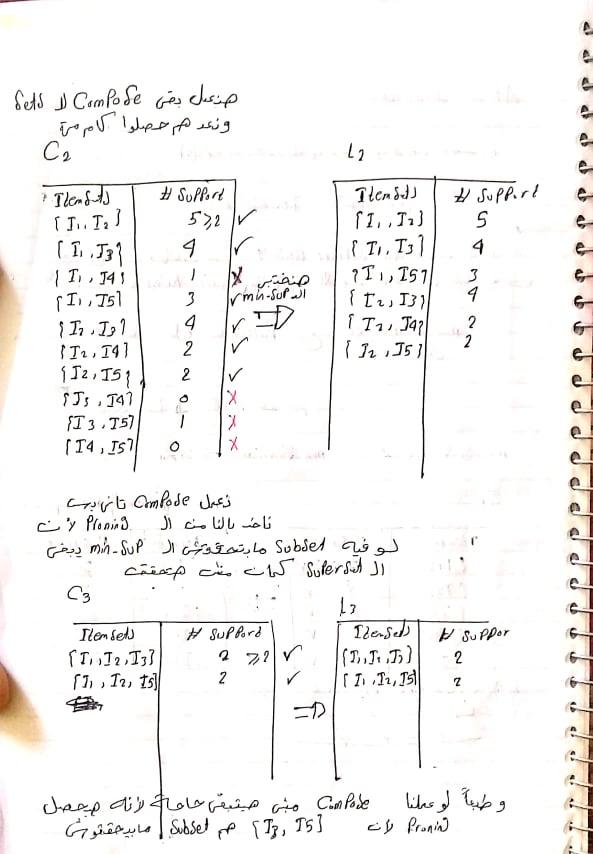




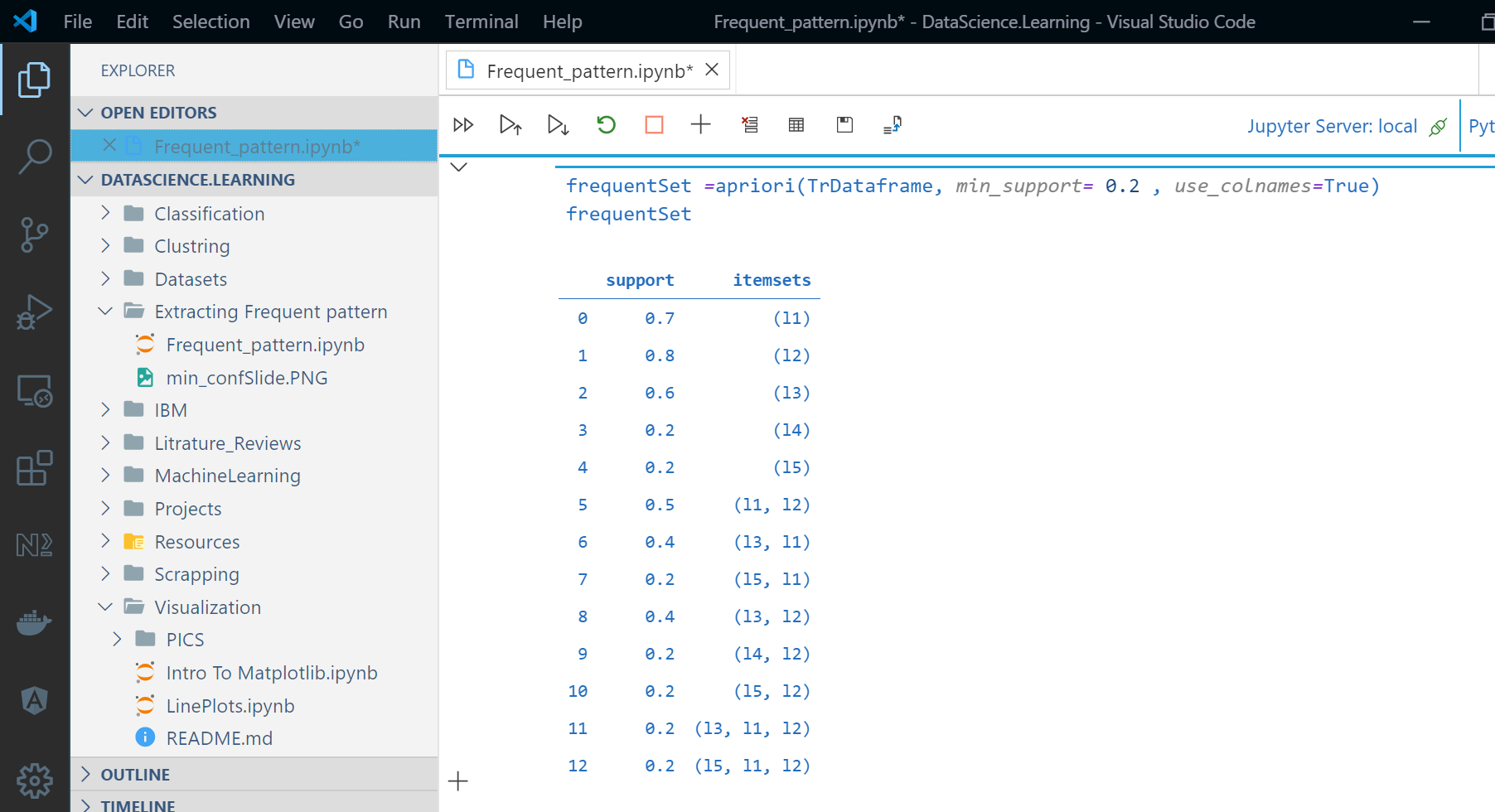
* نبدأ نحل المسألة دى ب 2 الجوريزم و هختبر الحلين بالبايثون كمان الله المستعان
  + Aprioiri
  + FP-Groth

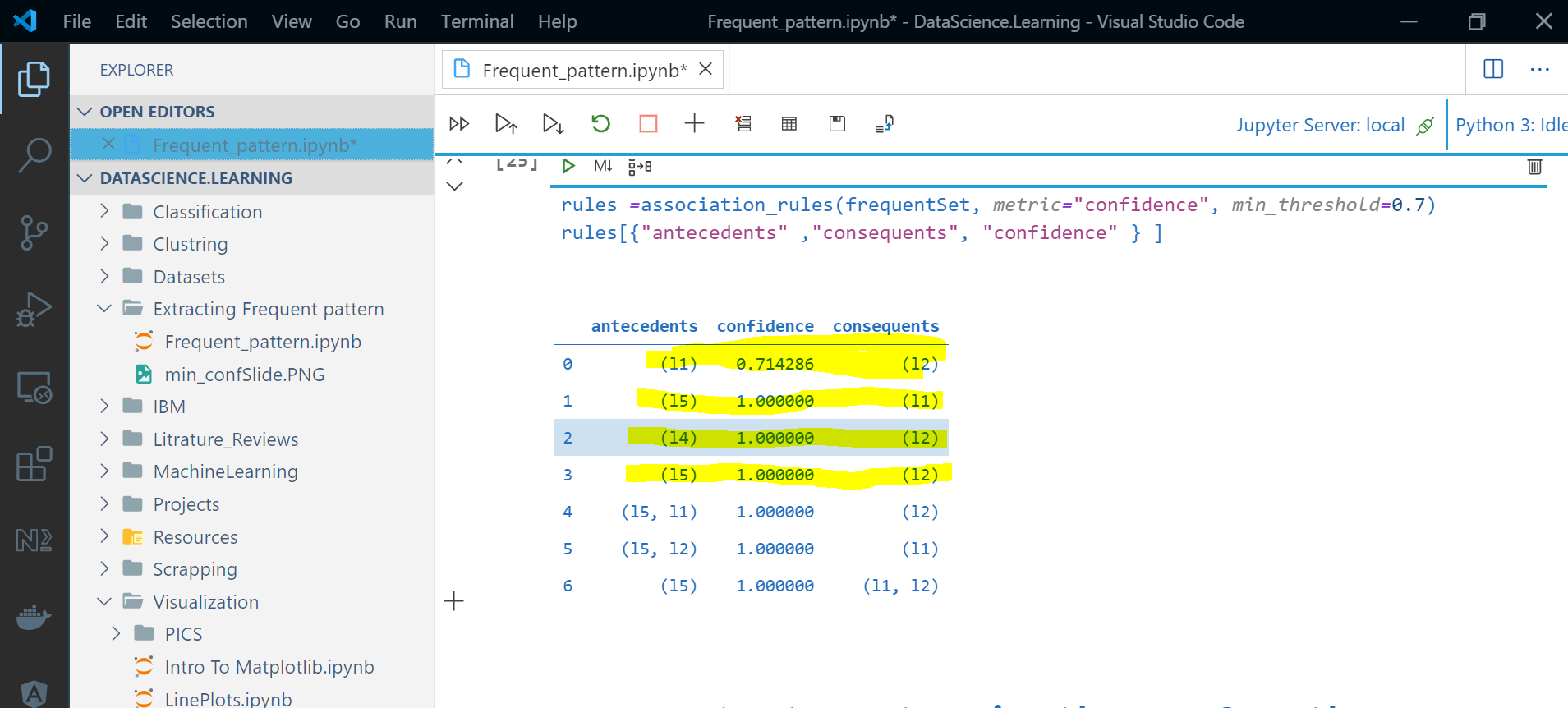
# Apriori Algorithm



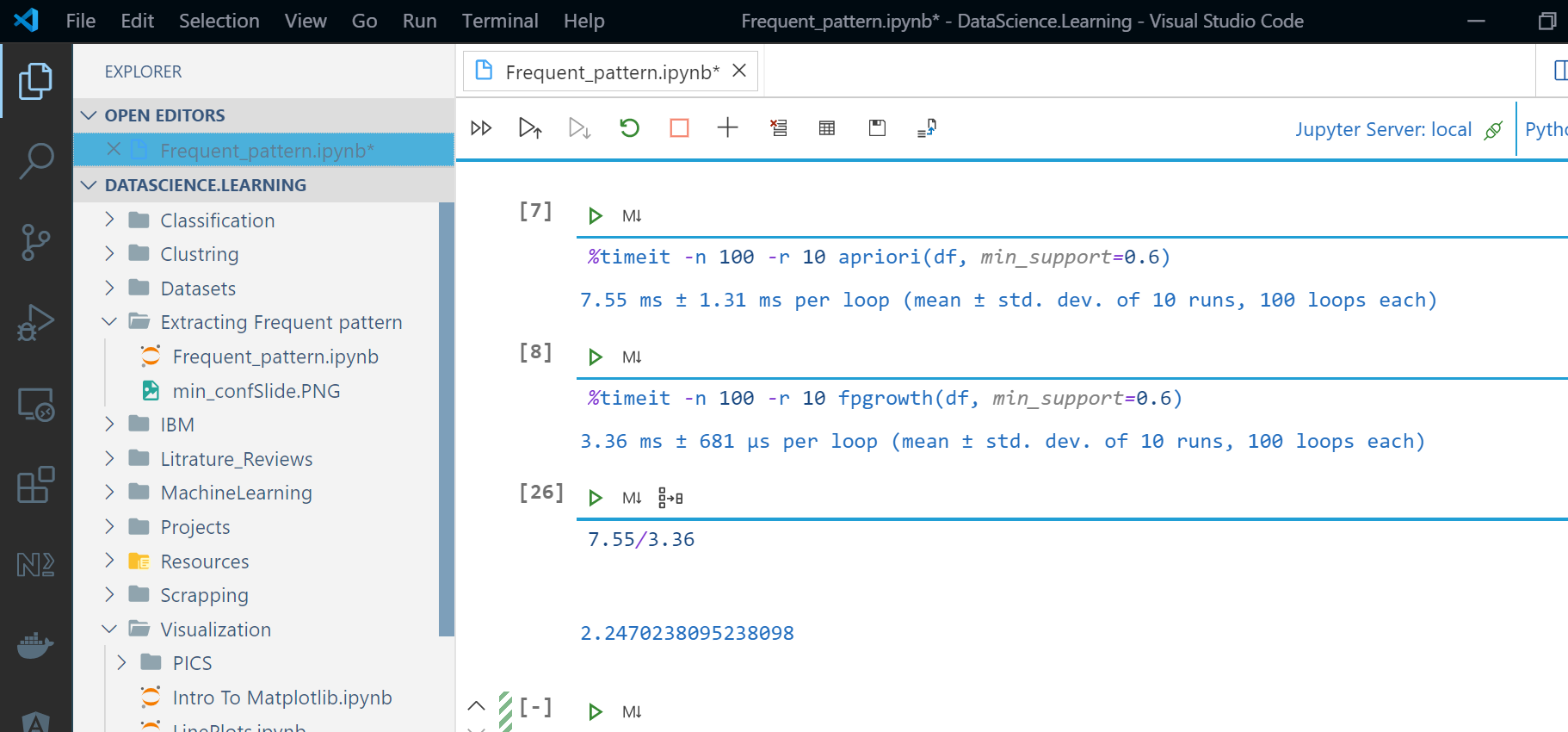


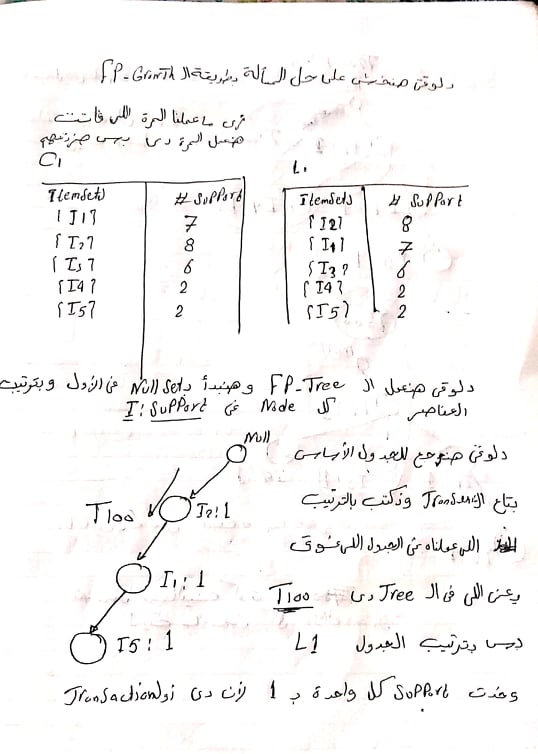
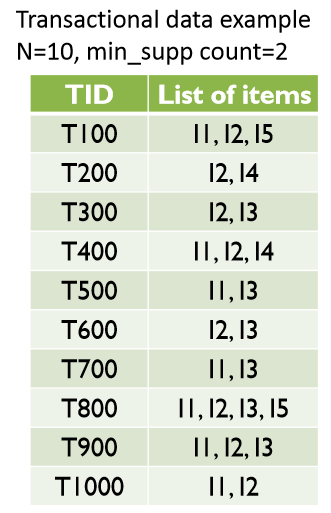
* أنا اسف خطى وحش جدا و حاسس بالموقف اللى انت فيه بس ربنا يستربس،، ده كله عشان ما بتقرأش المصادر يا عزيزى
* و بعد ما اتحلت منكم نحلها بايثون تمام و هنشوف ايه اللى فاضل من السترونج
* وده اللينك بتاع الكود اقرا المصادر بقى
* <https://github.com/AhmedKhalil777/DataScience.Learning/blob/master/Extracting%20Frequent%20pattern/Frequent_pattern.ipynb>

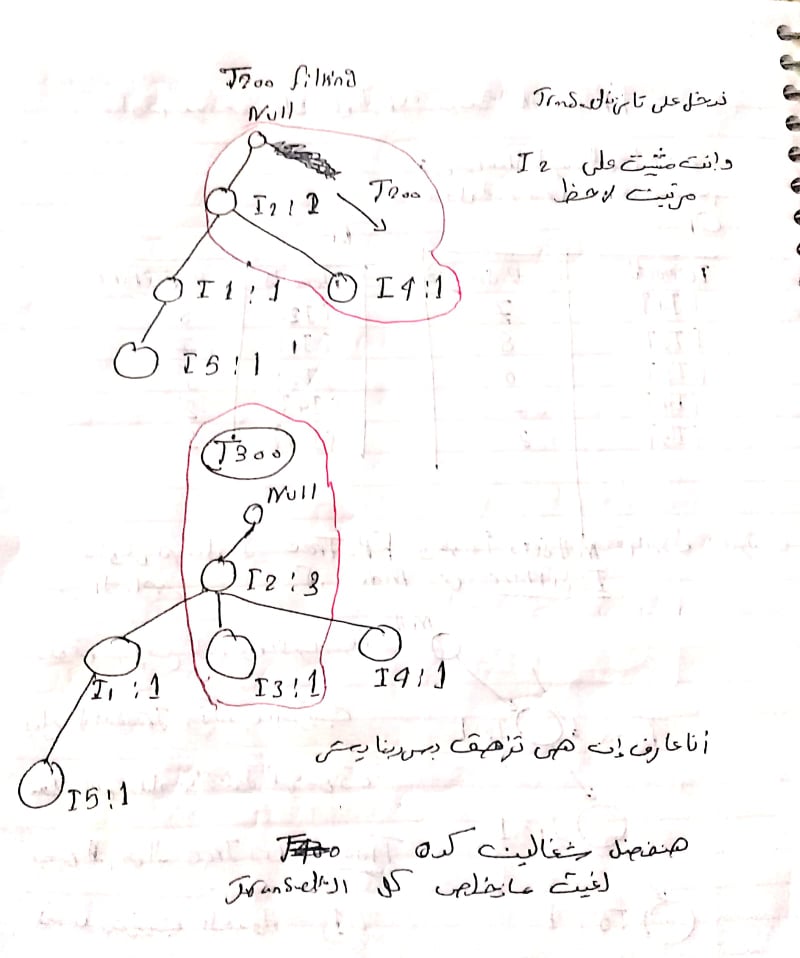
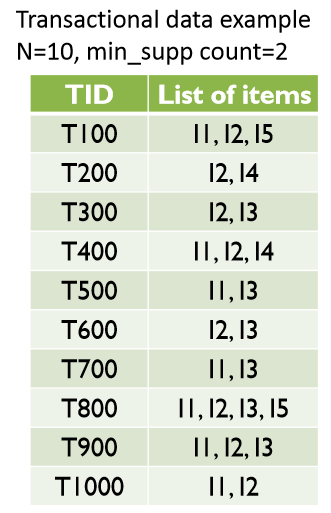


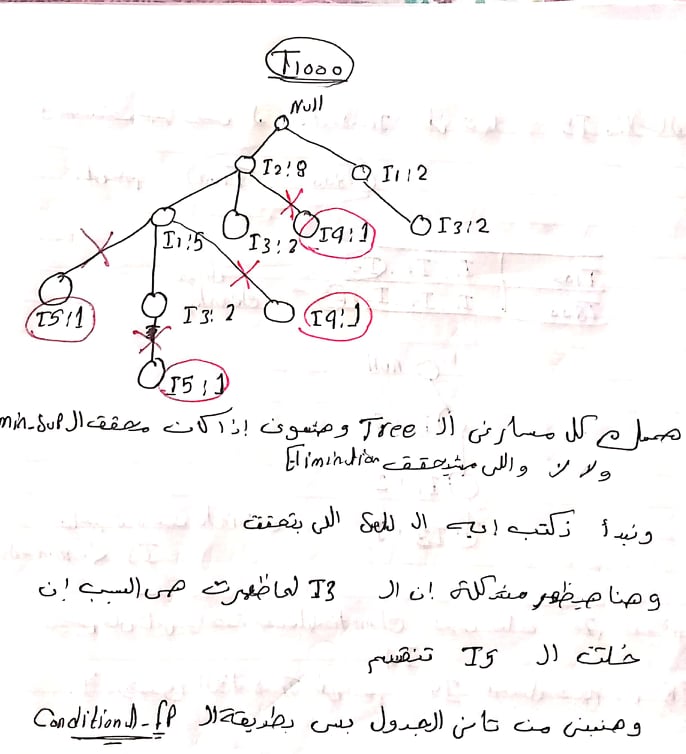


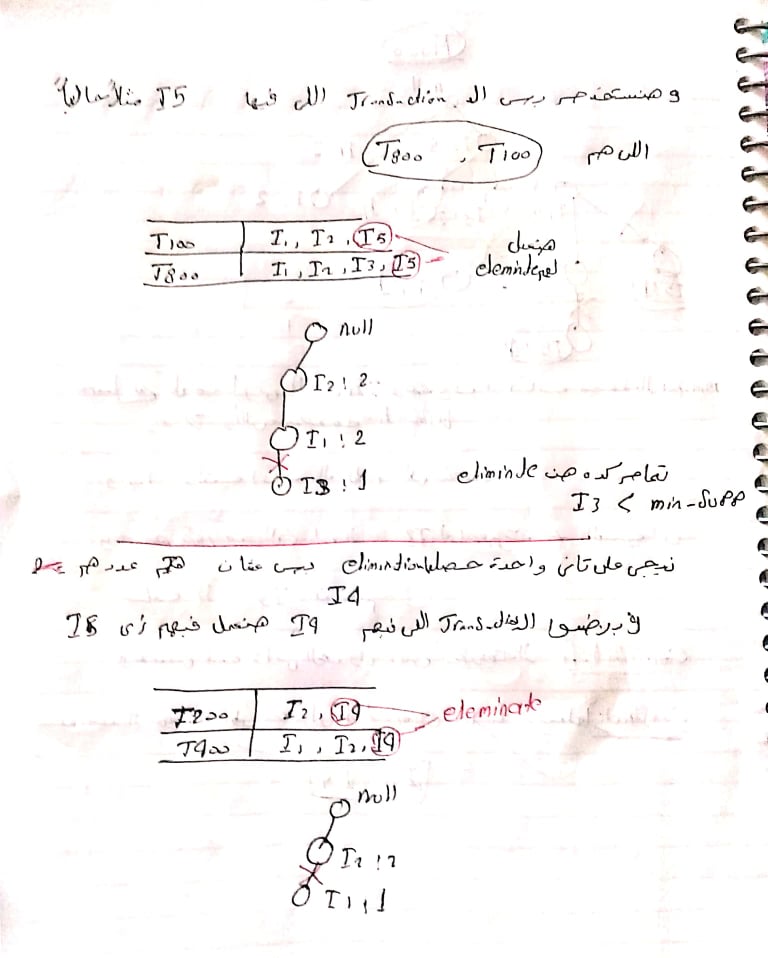
# FP-Growth

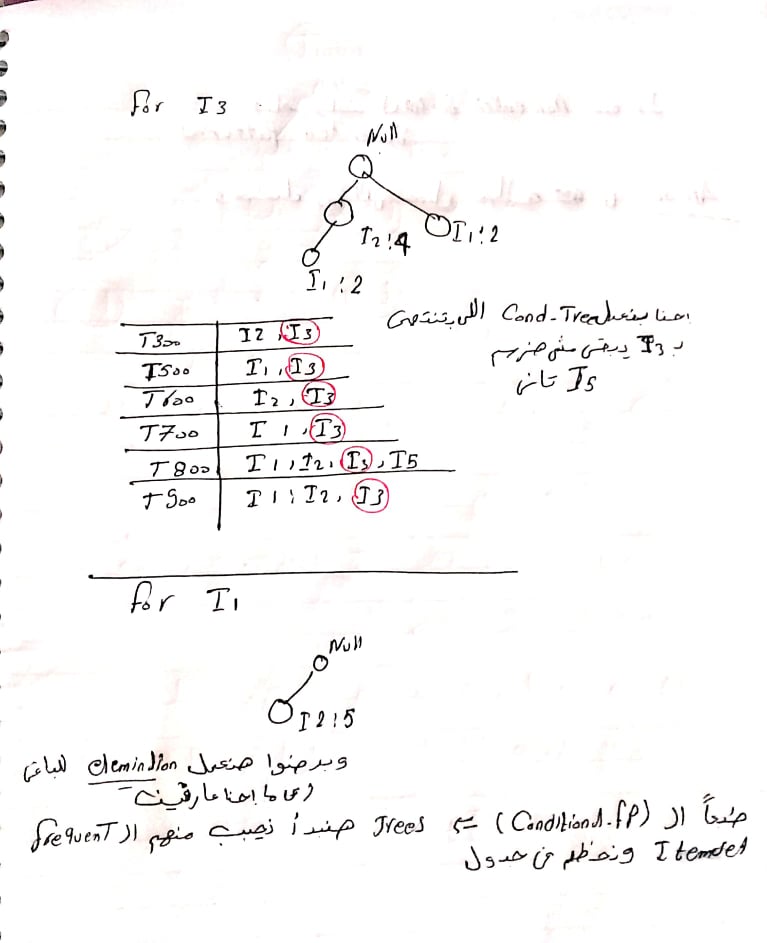
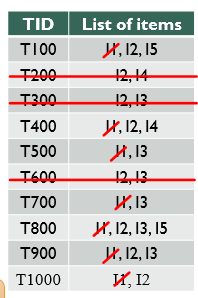
* أولا انا ما اتوقعش ان هي تيجى في امتحان بس للحظر بقول افهمها من المحاضرة
* اه هي تلخبط بس اللى فاهم ال
* Merge and Conquer
* هيفهمها كويس
* بس انا هعملها كود وهحسب ال
* Time cost
* بتاع الالجوريزمين و هتلاحظ فرق الوقت لو بتتعامل مع Big data
* بعد ال test Benchmarking
* اتضح ان في داتا بسيطة زى المسألة ان سرعة الالجوريزم ده أسرع مرتين وربع و طبعا هيختلف لو الداتا كبرت
* To avoid costly candidate generation
* Divide-and-conquer strategy:
  + Compress database representing frequent items into a frequent pattern tree (FPtree) – 2 passes over dataset
  + Divide compressed database (FP-tree) into conditional databases, then mine each for frequent itemsets – traverse through the FP-tree

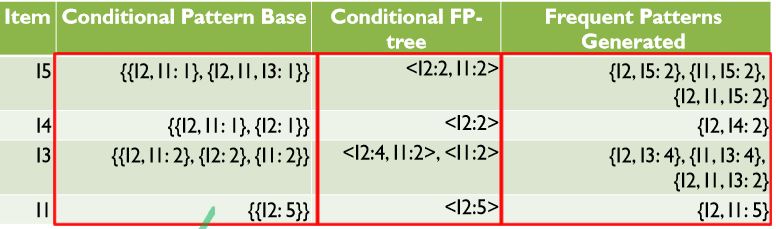


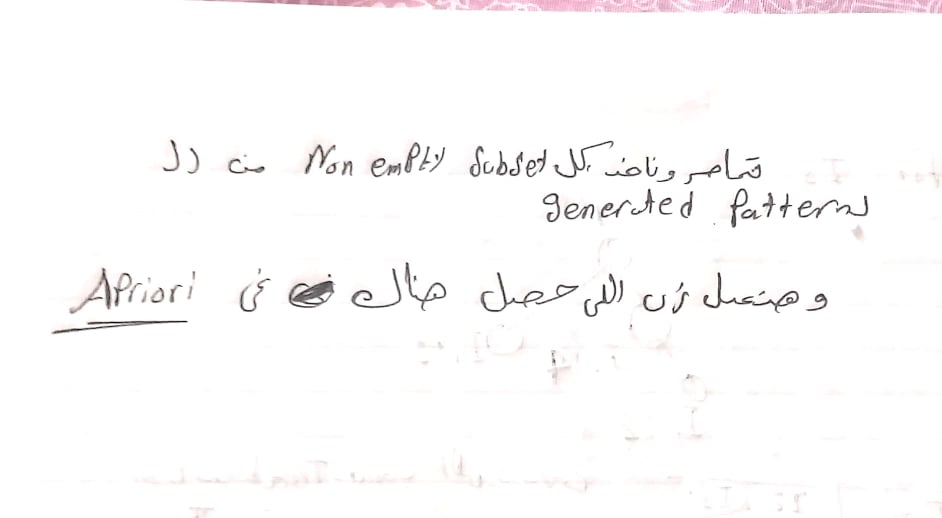












# Pattern Evaluation Method

* Not all association rules are interesting
  + Buys(X,”Computer games” → buys(X,”Videos”) [40%, 66%]
  + P(“videos”) is 75% > 66%
  + The two items are
    - negatively associated means buying one decreases the likelihood of buying the other
  + We need to measure “real strength” of rule
  + Correlation analysis
    - A→ B [support , confidence , correlation]
* يعنى نضيف قاعدة زيادة عشان نشوف اذا كان نشوف هو ايجابى او سلبى

