CS246 Exam 2025 — Question 1: Frequent Itemsets (12 points)

Part 1 (6 points) — A-Priori Algorithm

Task: Find all frequent itemsets with support ⩾ 3 from 6 playlists.

We abbreviate each song by its first letter: **G**rape, **M**aple, **O**xytocin, **S**omeone, **N**umb, **D**ive Back in Time.

Playlist	Songs
1	G, M, O, S
2	G, M, O, N
3	D, G, M
4	S, N
5	G, M, N
6	D, G, M, O

Iteration 1 — Singletons

Item	Playlists	Support
G	1, 2, 3, 5, 6	**5** ✓
M	1, 2, 3, 5, 6	**5** ✓
O	1, 2, 6	**3** ✓
S	1, 4	2 ✗
N	2, 4, 5	**3** ✓
D	3, 6	2 ✗

S and D are pruned (support < 3).

Iteration 2 — Pairs

By the **monotonicity property**, a pair can only be frequent if *both* items are individually frequent. We only form pairs from {G, M, O, N} — 6 candidates instead of 15.

Pair	Playlists	Support
{G, M}	1, 2, 3, 5, 6	**5** ✓
{G, O}	1, 2, 6	**3** ✓
{G, N}	2, 5	2 ✗
{M, O}	1, 2, 6	**3** ✓
{M, N}	2, 5	2 ✗
{O, N}	2	1 ✗

Iteration 3 — Triplets

All C(4,3) = 4 triplets from the frequent items {G, M, O, N} are listed as candidates. However, by the monotonicity property, only {G,M,O} has all three subsets frequent ({G,M} ✓, {G,O} ✓, {M,O} ✓). The other three can be pruned since they contain infrequent pairs (e.g., {G,N}, {M,N}, {O,N}).

{G, M, O} appears in playlists 1, 2, 6 → **support = 3** ✓

Final Answer

Iteration	Candidate Itemsets	Frequent Itemsets
1	{G}, {M}, {O}, {S}, {N}, {D}	{G}, {M}, {O}, {N}
2	{G,M}, {G,O}, {G,N}, {M,O}, {M,N}, {O,N}	{G,M}, {G,O}, {M,O}
3	{G,M,O}, {G,M,N}, {G,O,N}, {M,O,N}	{G,M,O}

Part 2 (2 points) — Downside of Highest-Confidence Rules

Recall: conf({X,Y} → Z) = support({X,Y,Z}) / support({X,Y})

Answer: High confidence doesn't account for the **baseline popularity** of the recommended item. If Z appears in 90% of all playlists, then *any* rule predicting Z will have confidence ≈ 0.9 — but recommending it adds no value.

The fix is the **interest metric**: Interest(I → j) = |conf(I → j) − P(j)|, which filters out rules that are only confident because the consequent is universally popular.

Part 3a (2 points) — Frequent Buckets in PCY

Question: If a bucket meets the support threshold, are all pairs that hash to it frequent?

Answer: No. A bucket's count is the aggregate of *all* pairs that hash into it. Multiple infrequent pairs can inflate a bucket's total above the threshold even though no individual pair is frequent. A frequent bucket is a **necessary** condition for a pair to be frequent, not a **sufficient** one.

Part 3b (2 points) — Can PCY Produce False Negatives?

Answer: No. If a pair is truly frequent (count $\geqslant$ s), then that pair alone contributes at least s to its bucket, guaranteeing the bucket is marked frequent. The pair will pass the filter and be counted in Pass 2. PCY's pruning is conservative — it may keep non-frequent pairs (false positives) but never discards truly frequent ones.

	Can it happen?	**Why?**
False positive (bucket frequent, pair isn't)	Yes	Infrequent pairs can inflate a bucket
False negative (pair frequent, bucket isn't)	No	One frequent pair guarantees its bucket meets the threshold