Data Mining
Constructing test sets for individual project

and

bettering your grade
Select approximate FDs that vary in domain cardinality and in degree of support $X \rightarrow Y$, where

- $|X| = 0$, $|X| = 1$, $|X| = 2$, $|X| = 3$, … (Depth = 0, Depth = 1, Depth = 2, Depth = 3, …)
- Min-sup = 1.0, Min-sup $\geq$ 0.95, Min-sup $\geq$ 0.90, Min-sup = 0.85, …
B→C 0.91 There are

- 86 rows with B=b₁ and C=c₁,
- 6 rows with B=b₁ and C=c₃,
- 2 rows with B=b₁ and C=c₂,
- 3 rows with B=b₃ and C=c₂,
- 2 rows with B=b₂ and C=c₁,
- 1 row with B=b₂ and C=c₂

\[
\frac{(86+3+2)}{100} = 0.91
\]
Select approximate FDs that vary in domain cardinality and in degree of support $X \Rightarrow Y$, where

- $|X| = 0, |X| = 1, |X| = 2, |X| = 3, \ldots$ (Depth = 0, Depth = 1, Depth = 2, Depth = 3, \ldots)
- Min-sup = 1.0, Min-sup $\geq 0.95$, Min-sup $\geq 0.90$, Min-sup = 0.85, \ldots
If you
• correctly implement find-approximate-functional-dependencies and all your auxiliary functions, for both on data sets you will be given ahead of time and those we use to grade;
• nicely format and comment your code with comprehensible and informative function header comments;
then you will receive an A- score (90%).

If, in addition,
• you implement some efficiency enhancement (such as pruning), and explain it clearly in comments at the top of the submission file (perhaps comparing runtime before and after enhancement); or
• simply instrument the code (top level function find_fds) and report runtime results as depth-limit varies for a fixed data set and minimal-support, and as minimal-support varies for a fixed data set and depth-limit; or
• Give a short write-up (e.g., one page) on results on an additional “real-world” data set, such as the “happiness” data set (already formatted for you +5%) or translate, test, and write up results with another real-world data set, such as Congressional Voting Records https://www.congress.gov/roll-call-votes, which has not been translated for you +10%);
then you can receive up to 100%.