Data Management Algorithm for Decision Making
02360003

Home Assignment 2 - Dry Part

Ido Tausi - 214008997
Afek Nahum - 214392706
Or Mutay - 206918633

Problem 1: Causal inference

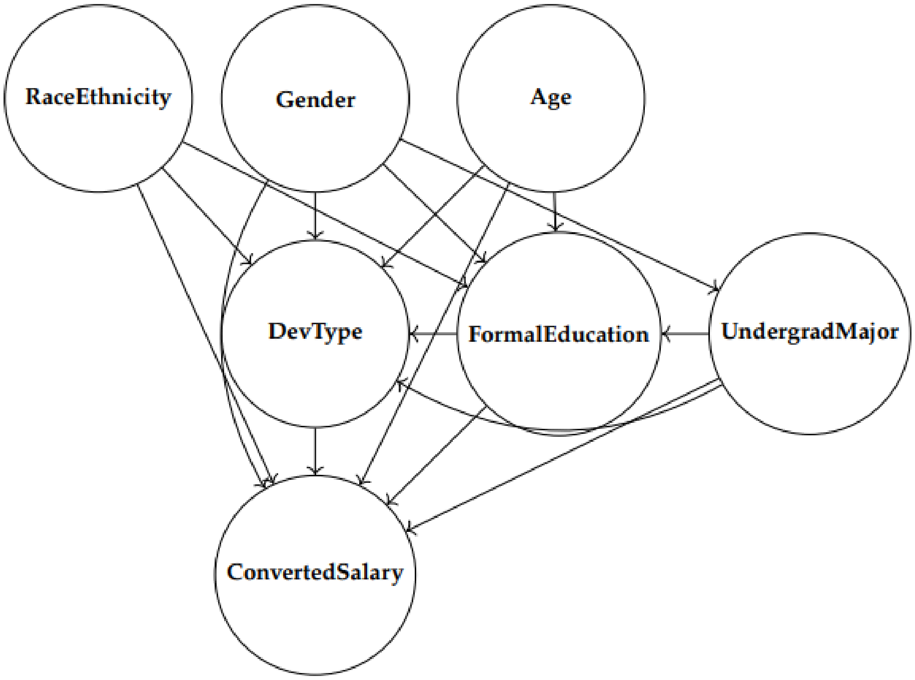


Figure 1: Partial causal DAG for the Stack Overflow dataset.

1.
(a) When $X = DevType,\ Y = ConvertedSalary$, the minimal set of variables that blocks all backdoor paths between X, Y is:
$\{RaceEthnicity, Gender, Age, FormalEducation, UndergradMajor\}$.
This set is minimal because for each variable in it, if we remove it then we get a backdoor path.

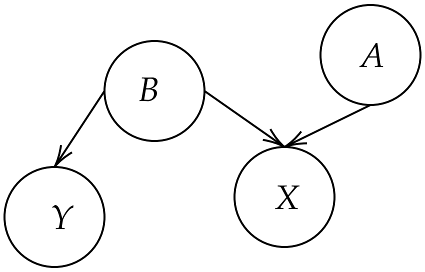
(b) When $X = Age,\ Y = ConvertedSalary$, the minimal set of variables that blocks all backdoor paths between X, Y is the empty set: $\{\}$.
This set is minimal because Age doesn't have any edges from other variables to him, therefore no backdoor paths.

(c) When $X = FormalEducation,\ Y = ConvertedSalary$, the minimal set of variables that blocks all backdoor paths between X, Y is: $\{RaceEthnicity, Gender, Age, UndergradMajor\}$.
This set is minimal because for each variable in it, if we remove it then we get a backdoor path.

2.
We are interested in a minimal adjustment set to block all backdoor paths between X and Y without including unnecessary variables. Adding extra variables can reduce accuracy, create bias, or complicate the analysis. A minimal set ensures we get an accurate and clear estimate of the true causal relationship, and its also desirable since it reduces the number of variables needed to control for confounding, and by that it reduces the model's complexity.

3.
By definition, parents of a treatment variable block all backdoor paths because they are not descendant of X (as they are his parents) and they will block all backdoor paths from X to Y as it blocks all incoming edges to X.

Parents do not always provide the minimal adjustment, for example:
A and B are the parents of X, but $\{B\}$ is the minimal adjustment set.



4.
In the Jupyter Notebook

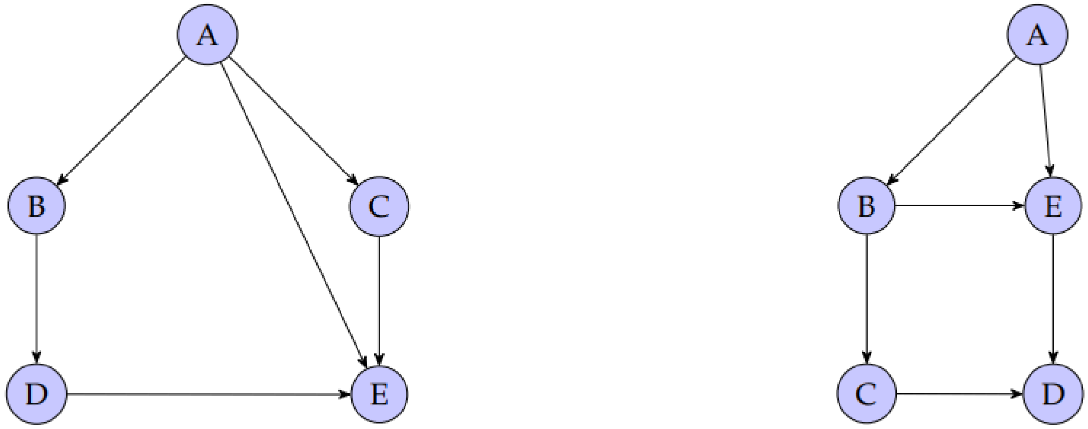
a) ATE (Master's degree): -1442.3064, P-value: [0.19083086]
b) ATE (Bachelor's degree): 2832.1884, P-value: [0.00170637]

5.
In the Jupyter Notebook

The group identified as potentially biased against is defined by the following attributes:
Age: 18 - 24 years old
Gender: Female
Race/Ethnicity: Black or African descent; Native American, Pacific Islander, or Indigenous Australian; White or of European descent

Average Treatment Effect (ATE): -90165.7738, P-value: 0.52122281

Problem 2: d-Separation



(a) DAG 1.

(b) DAG 2.

Figure 2: Causal DAGs for problem 2.

1.

Graph $DAG\ 1$:

1. For $X = D,\ Y = E$ the backdoor are: $D \leftarrow B \leftarrow A \rightarrow E,\ D \leftarrow B \leftarrow A \rightarrow C \rightarrow E$.
We can see that every backdoor path passes through B, therefore we can choose for our adjustment set $Z = \{B\}$, and every backdoor path would be blocked.

2. For $X = C,\ Y = E$ the backdoor are: $C \leftarrow A \rightarrow E,\ C \leftarrow A \rightarrow B \rightarrow D \rightarrow E$.
We can see that every backdoor path passes through A, therefore we can choose for our adjustment set $Z = \{A\}$, and every backdoor path would be blocked.

Graph $DAG\ 2$:

1. For $X = C,\ Y = D$ the backdoor are:$C \leftarrow B \rightarrow E \rightarrow D,\ C \leftarrow B \leftarrow A \rightarrow E \rightarrow D$.
We can see that every backdoor path passes through B, therefore we can choose for our adjustment set $Z = \{B\}$, and every backdoor path would be blocked.

2. For $X = E,\ Y = D$ the backdoor are:$E \leftarrow A \rightarrow B \rightarrow C \rightarrow D,\ E \leftarrow B \rightarrow C \rightarrow D$.
We can see that every backdoor path passes through B, therefore we can choose for our adjustment set $Z = \{B\}$, and every backdoor path would be blocked.

2.

a)

Proof:

given $X \perp_d (Y \cup A) | Z$, according to axiom (c) of weak union, it is implied that $X \perp_d Y | (Z \cup A)$ and also $X \perp_d A | (Z \cup Y)$ happen. Specifically, it is implied that $X \perp_d Y | (Z \cup A)$.

b)

Proof:

given $X \perp_d Y | Z$ and $X \perp_d A | (Z \cup Y)$, according to axiom (d) of contraction, it is implied that $X \perp_d (Y \cup A) | Z$.

c)

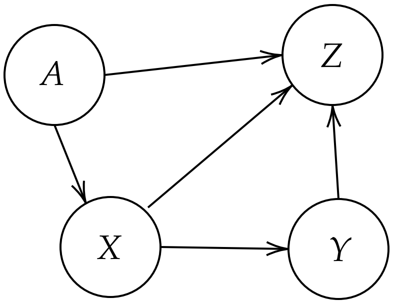
Proof:

given $X \perp_d Y | (Z \cup A)$ and $X \perp_d A | (Z \cup Y)$, according to axiom (e) of intersection, it is implied that $X \perp_d (Y \cup A) | Z$.

Now, according to axiom (b) of decomposition, on $X \perp_d (Y \cup A) | Z$, we get that $(X \perp_d Y | Z)$ and $(X \perp_d A | Z)$. Specifically, it is implied that $X \perp_d Y | Z$.

d)

Counter example:



Given this graph, all the backdoor paths from X to Y are blocked when Z is chosen for adjustment set. This is because each backdoor paths goes through Z, choosing it as the adjustment set will mean the all of them will be blocked. Therefore $X \perp_d Y | Z$.

Additionally, all the backdoor paths from Y to Z are blocked when X is chosen for adjustment set.

This is because each backdoor paths goes through X, choosing it as the adjustment set will mean the all of them will be blocked. Therefore $Y \perp_d Z | X$.

But the required condition $X \perp_d Z | Y$ doesn't happen, since there is a backdoor path $Z \leftarrow A \rightarrow X$ that Y doesn't block.

3.

We are interested in determining the number of ways to partition n nodes into three disjoint sets: X, Y and Z. Additionally, each node can be excluded from all sets, resulting in a fourth option. Therefore, each node can exist in one of four states: in set X, in set Y, in set Z, or in none of the sets.

Since each node has four independent choices, the total number of possible encoding for n nodes is 4^n. This gives us an upper bound for the total number of possible CI statements, although not all of these configurations will necessarily be valid CI statements.

4.

Sparse DAGs usually encode more CI statements. This is because they provide more opportunities for nodes to be $d - separated$, due to having fewer edges compared to the number of vertices. In contrast, a fully connected DAG would have very few conditional independence relationships, as it is harder to find nodes that are $d - separated$.

Problem 3: Frequent Itemsets

1.

In the notebook.

2.

To identify an over-represented subpopulation in the dataset, we used the Apriori algorithm to uncover frequent itemsets and generate association rules.

Step 1: Data Preparation

The dataset was preprocessed to ensure compatibility with the Apriori algorithm.
Missing values in numerical columns were handled by replacing them with the median of the respective column to reduce the impact of outliers. For categorical columns, missing values were replaced with the mode, ensuring minimal changes to the overall data patterns.
We discretized the `loan_amount` column into meaningful bins: `[300K-7.7M, 7.7M-14.5M, 14.5M-21.5M, 21.5M-39.5M]`, capturing the distribution of loan amounts. Additionally, categorical features such as `education`, `self_employed`, and `loan_status` were converted into binary formats.

Step 2: Transforming Data into Transactions

Each row of the dataset was transformed into a transaction containing feature-value pairs, as asked in q1.

Step 3: Executing the Apriori Algorithm

The Apriori algorithm was applied with a minimum support of 15% and a minimum confidence of 50%. To identify over-represented subpopulations, the resulting association rules were filtered by lift, with a threshold of `lift > 1.05`.

Results:

```
Over-represented Subpopulations:
{ loan_amount:21.5M-39.5M} -> { loan_status:1} (conf: 0.662, supp: 0.164,
lift: 1.064, conv: 1.118)
```

The results suggest that individuals with a loan amount in the range of `21.5M-39.5M` are over-represented among those with approved loan statuses (`loan_status:1`). We concluded that based on the confidence of 66.2%, that shows a high chance that people in this group would get their loans approved.

3.

To identify a loan term property associated with approved loan status, we used the Apriori algorithm on the dataset. First, we did the same data preparation as in q2. The Apriori algorithm was then applied with a minimum support of 5% and a minimum confidence of 60% to uncover meaningful associations.

The results showed that shorter loan terms, such as 2 and 4 years, are strongly associated with approved loan statuses. The rule $\{loan\_term: 2\} \rightarrow \{loan\_status: 1\}$ had a confidence of 78% and a lift of 1.253, indicating a higher-than-average likelihood of approval for this term.

Similarly, $\{loan\_term: 4\} \rightarrow \{loan\_status: 1\}$ exhibited a confidence of 81.9% and a lift of 1.316, highlighting an even stronger association. In contrast, longer terms like 12 and 18 years had lower confidence (~60%) and lift values below 1, suggesting weaker relationships with loan approvals.

These findings were derived by focusing the analysis on `loan_term` during preprocessing and filtering the resulting association rules. The insights suggest that shorter loan terms play a significant role in influencing loan approvals.

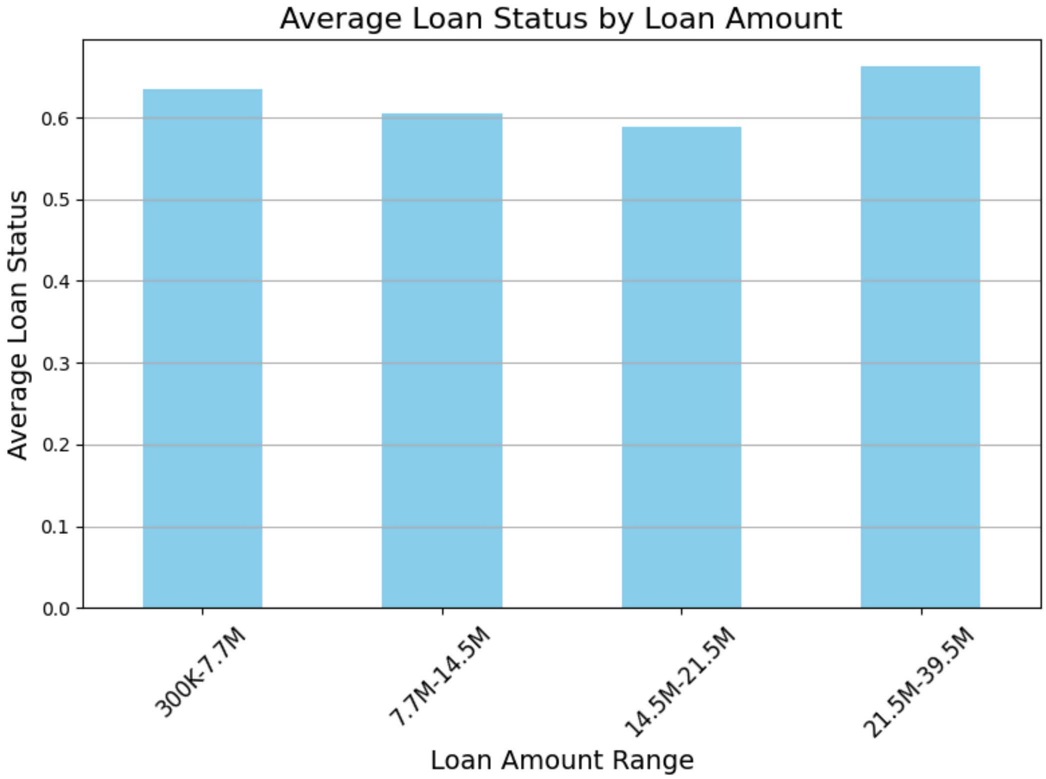
```
Loan Term Related Association Rules:
{ loan_term:12} -> { loan_status:1} (conf: 0.605, supp: 0.065, lift: 0.973,
conv: 0.957)
{ loan_term:18} -> { loan_status:1} (conf: 0.609, supp: 0.060, lift: 0.979,
conv: 0.966)
{ loan_term:2} -> { loan_status:1} (conf: 0.780, supp: 0.074, lift: 1.253,
conv: 1.715)
{ loan_term:4} -> { loan_status:1} (conf: 0.819, supp: 0.086, lift: 1.316,
conv: 2.085)
```

Problem 4: Interesting visualizations

1.

We choose `loan_amount`.

2.

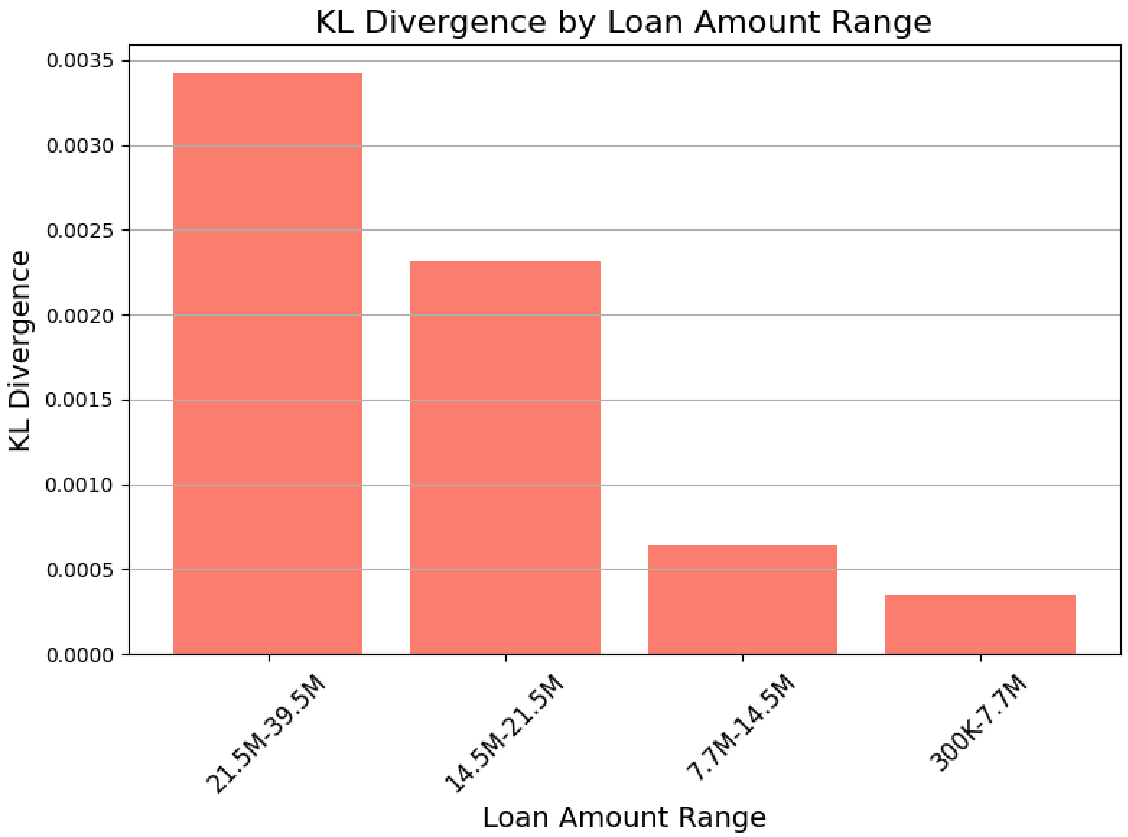


3.

Code:

In the notebook.

Results:



```
Most Divergent Subpopulation:
Loan Amount Range: 21.5M-39.5M, KL Divergence: 0.0034
```

Insights:

The plot reveals that the loan amount range `21.5M-39.5M` has the highest KL divergence, indicating that this subpopulation's loan approval rate deviates the most from the overall approval trend. Specifically, larger loans in this range have a significantly different approval pattern compared to other bins.

Subpopulations with smaller loan amounts, like `300K-7.7M` and `7.7M-14.5M`, show minimal divergence, suggesting their loan approval distributions closely follow the overall trend.