R Competency and The Drivetrain Approach to Decision Making

Viet Hoang Tran Duong

CS112 Spring 2019

(1) When projects are approved, they are approved for a certain period of time (until the time of "original completion date"). While projects are active, this "original" completion date is often pushed out (extended), and then there is a "revised" completion date.

<u>Claim 1</u>: You have been told that project duration at approval is generally about 2 years (24 months). In other words, (purportedly) when projects are approved, the difference between the original project completion date and the the approval date is (supposedly) approximately 24 months.

(a) Is this claim true? Explain.

We assume 1 month has 30 days. Hence, 24 months = 24*30 = 720 days

We calculated the time difference between the “Original Completion Date” and “Approval Date” for all qualified observations (after 2008-01-01 and not NA).

Median	Mean	Interquartile Range	Range
592.0	643.8	384.0	3351.0

Table 1.1: The table represents the parameters describe the distribution of the time difference between the “Original Completion Date” and “Approval Date” (unit: days)

As seen from table 1.1, the mean (643.8 days) and the median (592 days) difference between the "Original Completion Date" and "Approval Date" are smaller than 720 days. Also, the gaps between the mean, median compared to 720 days are 76.2 days and 128 days, respectively. These gaps are comparably big comparing to the Interquartile Range (384 days), meaning that the gaps are significant. Based on these comparisons, we can infer that, on average, the difference between the original project completion date and the approval date is less than 24 months.

Hence, this claim is not true.

Furthermore, the range and the standard deviation implies that the distribution of the time difference is very diverse.

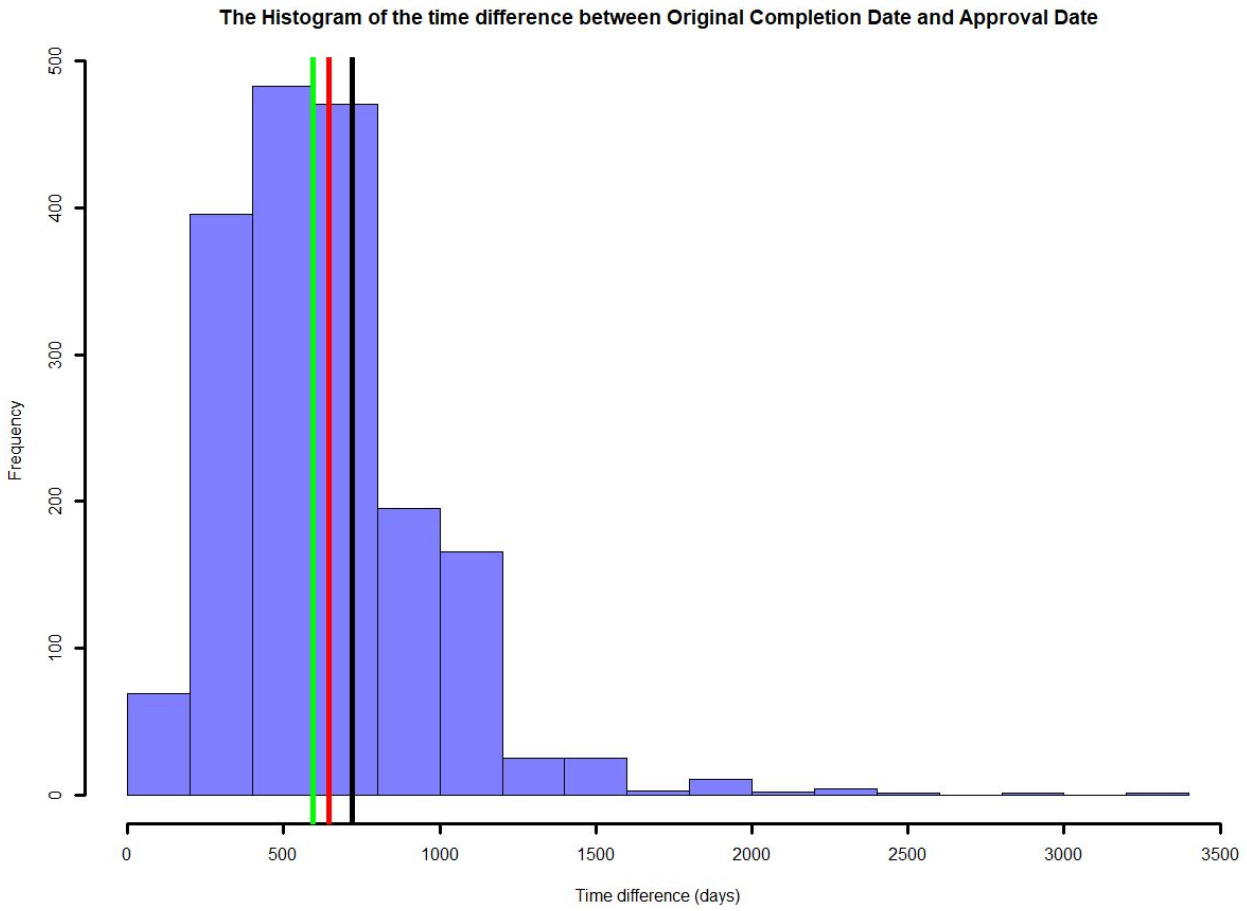

Figure 1.1: The histogram of the time difference between the Original Completion Date and Approval Date. The green line represents the median, the red line represents the mean, and the black line represents 720 days. As we can see, the distribution is very broad, positive skewed. Hence, the mean is susceptible to the influence of outliers on the right. The median will be the better measurement of the central tendency. Also, we can see the median is far from the 720 days (green line and black line). Hence, the given claim 1 is incorrect.

Has project duration at approval changed over time (consider projects circulated earlier and circulated later). Be sure to discuss mean durations, median durations, and the interquartile range of durations (using the "quantile" function). Approximate suggested length: 3-5 sentences.

The project duration is divided into two groups: Group CD_low contains 20% of the projects with the earliest Circulation Date, while Group CD_upper contains 20% of the projects with the latest Circulation Dat

Groups	Median	Mean	Interquartile Range	Range
Group CD_low	551.0	597.6	385.7	1877.0
Group CD_upper	711.5	726.1	363.7	2483.0
Difference (upper - lower)	160.5	128.5	-22.0	606.0

Table 1.2: The table represents the parameters describe the distribution of the time difference between the “Original Completion Date” and “Approval Date” of the 2 groups CD_low and CD_upper. (unit: days)

As seen from table 1.2, the Median, Mean of the CD_upper group are higher than of the CD_lower group. We can see that the projects with later Circulation Date have larger time differences than the projects with earlier Circulation Date. Hence, the project duration at approval has increased over time.

Also, for both groups, the interquartile range is much smaller than the range. This suggests that both groups contain many outlying data points but the time difference for the majority (50%) of projects has not deviated much from the median.

(b) How does original planned project duration differ from actual duration (if actual duration is measured as the duration between "ApprovalDate" and "RevisedCompletionDate"?) Once again, use # means, medians, and interquartile ranges to explain your results. Approximate suggested length: 3-5 sentences

First, we will calculate the actual duration of the projects.

For each project, we will then calculate the difference between the actual duration and the original planned project duration. Our calculations are reflected in the table below:

Min	Median	Mean	Interquartile Range	Range
0.0	485.0	573.5	517.0	3837.0

Table 1.3: The table represents the parameters describe the distribution of the difference between actual duration and the original planned project duration (unit: days).

From table 1.4, we can see that the actual duration is either more than or equal to the original planned project duration because the minimum of the difference between the two durations is 0 (days). Also, the median (485) and the mean (573.5) demonstrates the large gap between planned and actual duration. A large interquartile range and range implies a large diversity in the difference between the two durations.

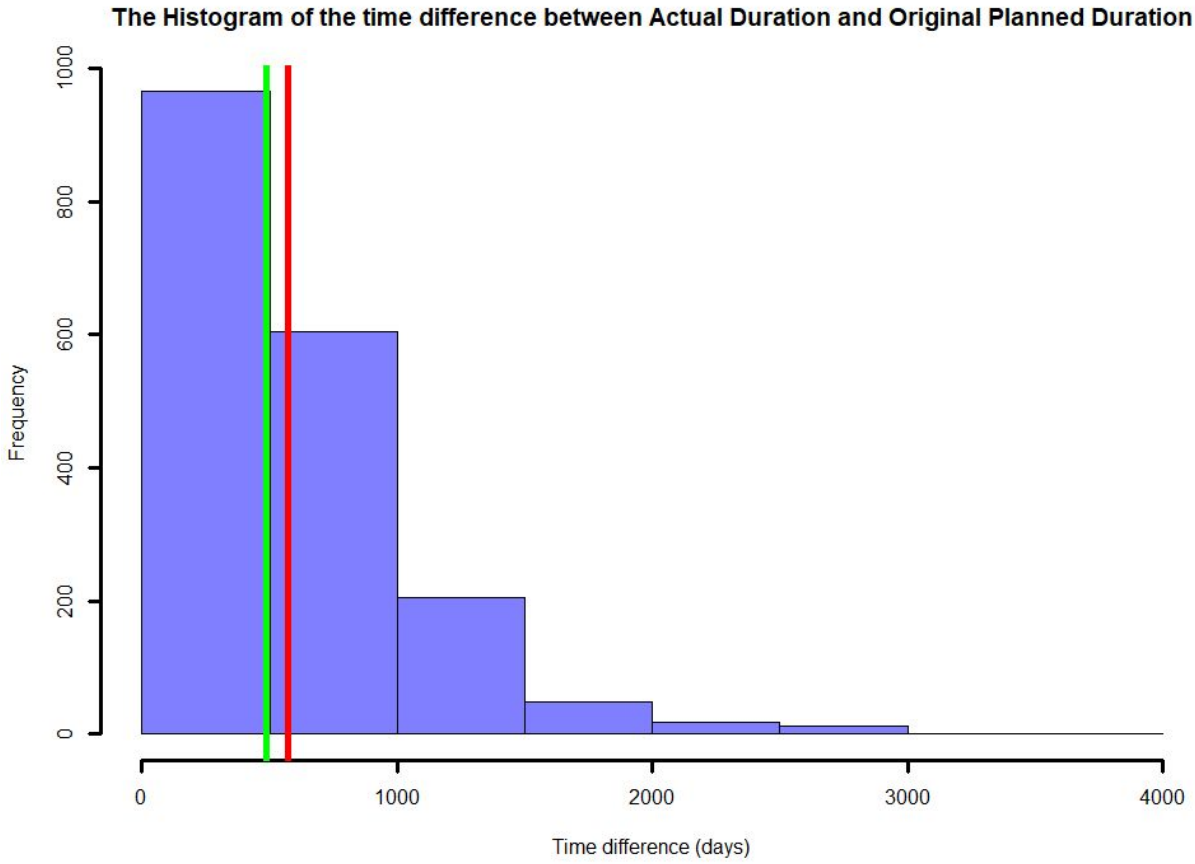

Figure 1.2: The histogram demonstrates the time difference between the Actual Duration and the Original Planned Duration. The green line represents the median, the red line represents the mean. As we can see, the distribution is very broad, positive skewed. Even though the majority of the projects has the gap less than 500 days (based on the median, 50% of the projects' gap was less than 485 days), the right tail makes the mean becomes bigger (573.5 days). As the mean is susceptible to the influence of outliers on the right, the median will be the better measurement of the central tendency.

(2) What % of projects that have ratings were rated 0? What % were rated 1? What % were rated 2? What % were rated 3? Answer these questions using a table or a figure. Provide a title AND an explanatory sentence or two that provides the numerical % results rounded to the nearest percentage-point.

Rating	Frequency (projects)	Percentage (%)
0	50	3
1	297	16
2	1276	68
3	246	13

Table 2.1: The table represents the frequency and the percentage of the projects rating.

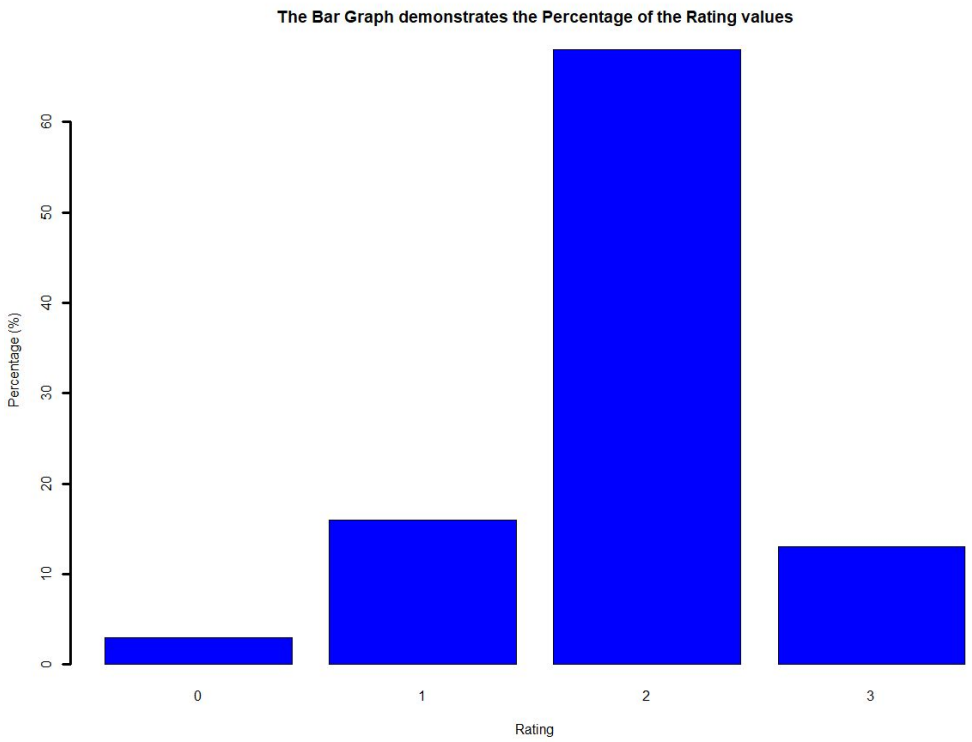

Figure 2.1: The bar plot represents the percentage of different project ratings

The majority of the projects (68%) receive the rating of 2. The lowest percentage is the projects with rating of 0, with 3% of the projects. The projects receive the rating of 1 and 3 are 16% and 13% of the total projects, respectively.

(3) Repeat problem 2, but this time exclude all PPTA projects. PPTA projects are more prone to negative ratings, because after a certain point in time only the low-rated PPTA projects required ratings. PPTA stands for "Project Preparatory Technical Assistance" and it is basically a project intended to set up a loan (often a very large multi-million-dollar loan). Only PPTAs that fail to "eventuate" to a loan are rated, which is why they are usually rated negatively.

Rating	Frequency (projects)	Percentage (%)
0	35	2
1	247	14
2	1234	70
3	241	14

Table 3.1: The table represents the frequency and the percentage of the projects rating of those excluding PPTA projects.

After excluding the PPTA projects, the majority of the remaining projects (70%) receive the rating of 2, while 2% of the projects receive a rating of 0 and 14% of the projects receive a rating of 1 and 3.

Compared to Table 2.1 and Figure 2.1, as we delete the PPTA projects, the percentage of the groups with high rating increase: the group with rating of 2 increases 2%, from 68% to 70%; the group with rating of 3 increases 1%, from 13% to 14%. The percentage of the groups with low rating decrease: the group with rating of 1 decreases 2%, from 16% to 14%; the group with rating of 1 decreases 1%, from 3% to 2%.

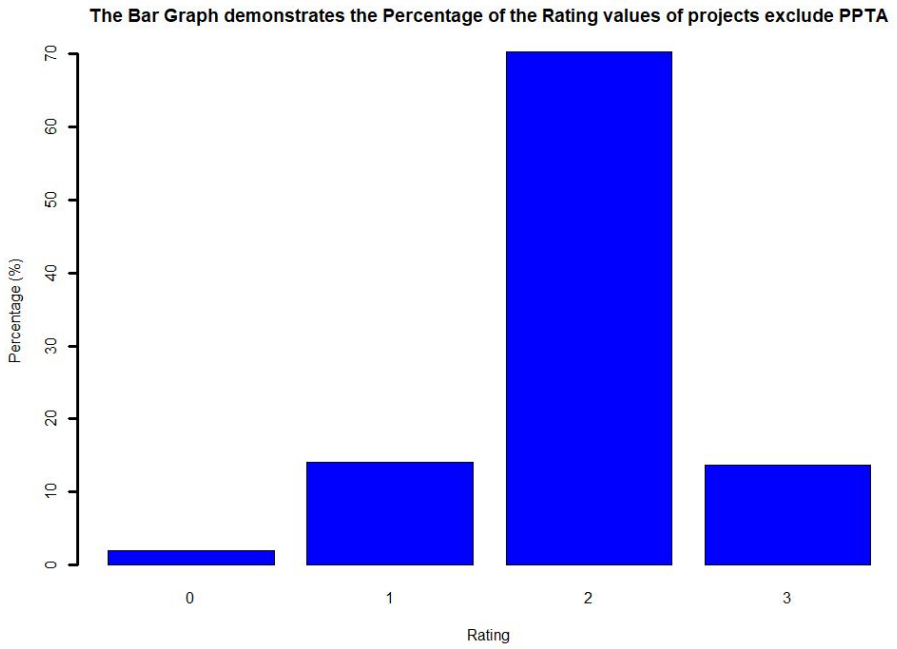

Figure 3.1: The bar plot represents the percentage of different project ratings of those excluding PPTA projects.

(4) Identify the top 25% of projects by "Revised.Amount" and the bottom 25% of projects by "RevisedAmount". ("RevisedAmount" shows the final project budget.)

Compare the ratings of these projects. Can you draw a causal conclusion about the effect of budget size on ratings? Why or why not?

Hint: Compare the characteristics of the two project groupings, e.g., "Dept", "Division", "Cluster", "Country"

Approximate suggested length: 3-5 sentences.

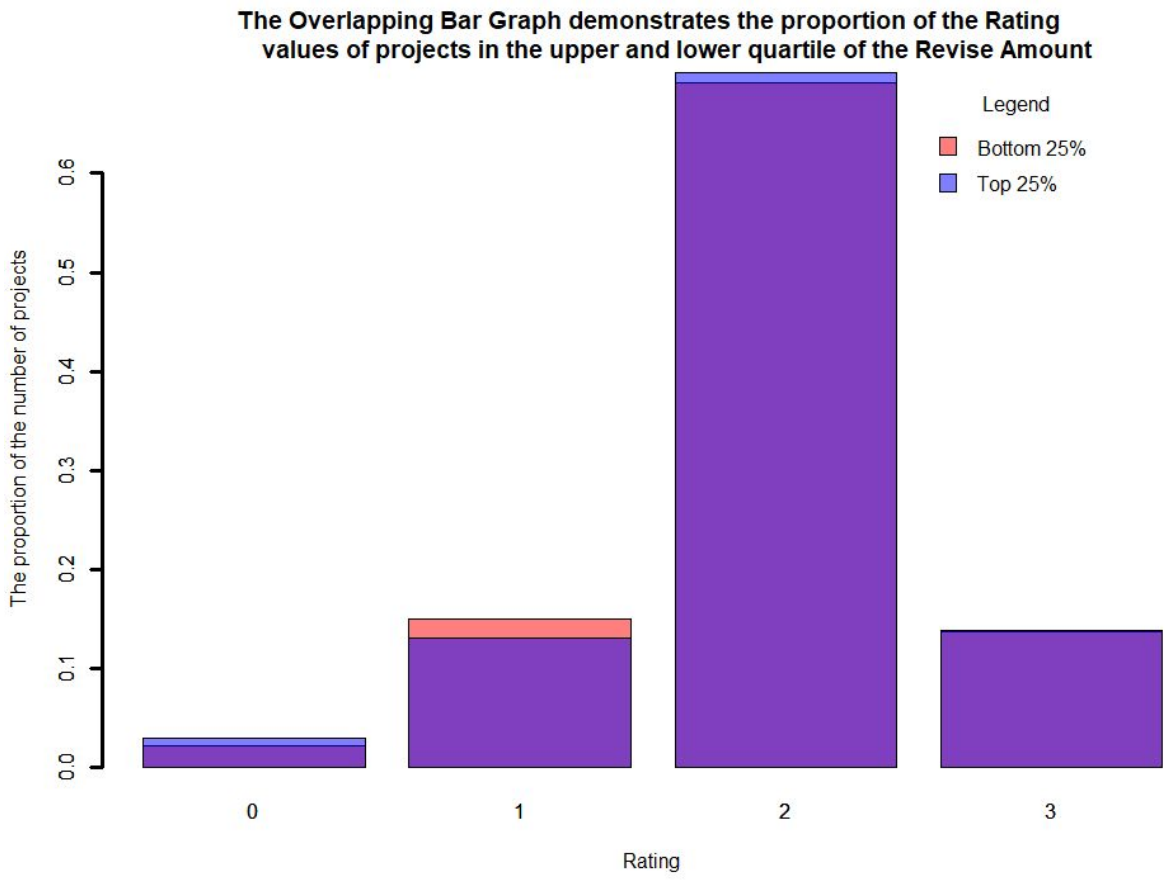

Figure 4.1: The Overlapping Bar graph demonstrates the proportion of the Rating values of projects in the upper and lower quartile of the final project budget. The red color represents the

projects in the bottom 25% of final project budget. The blue color represents the projects in the top 25% of final project budget. The purple represents the overlapping of those 2 groups.

From figure 4.1 we can see that the top 25% have a slightly higher proportion of projects that received a rating of 0 and 2, while the bottom 25% have a slightly higher proportion of projects that received a rating of 0. There is no significant difference in the proportion of the projects that received a rating of 3 in both the upper and lower quartile groups.

Rating	**Group**	**Percentage (%)**
0	Top 25%	2
	Bottom 25%	3
1	Top 25%	15
	Bottom 25%	13
2	Top 25%	69
	Bottom 25%	70
3	Top 25%	14
	Bottom 25%	14

Table 4.1: The table compares the percentage of the Rating values of projects in the top and bottom quartile of the final project budget.

From the table 5.1, the bottom quartile group has a higher percentage of projects with rating 0 and 2 by 1%, while the top quartile group has a higher percentage of projects with a rating 1 by 2%. The percentage of projects with a rating of 3 is the same.

From table 5.1 and figure 4.1, we can see little difference in ratings between projects in the top and bottom quartile by the "Revised Amount." Furthermore, we cannot draw any causal relations of the budget size on the rating because there are many confounding variables (i.e. "Dept.", "Country"). As this is observational studies, there are no control or treatment groups or random sampling. Therefore, we cannot and should not draw any causal inference from these observations.

(5) Imagine your manager asks you to apply Jeremy Howard's drivetrain model to the problem of optimal budget-setting to maximize project success (i.e., "Rating"). In such a situation, what would be the:

a. **Decision problem or objective?**

Our objective is to maximize the project success (by achieving a rating score of 3) while minimizing cost. Hence, we want to understand if there is a difference between the projects' rating if there is an increase or decrease in the projects' final budget.

b. **Lever or levers?**

Levers are characteristics those we can manipulate.

In this scenario, to achieve the objective, we can manipulate the amount of final budget assigned to a project - the "RevisedAmount" variable.

c. **Ideal RCT design?**

Our unit of analysis is projects that are identical in all variables (i.e., "Dept.", Country"). There will be one control group and two treatment groups, where each group should have an equal amount of projects and consist of at least 30 projects. The projects that are given more budget will be in group A, while projects that are given less budget will be in group B.

The control group will receive the median budget of the projects in the dataset: 0.6 (budget units). Group A whose treatment is receiving more budget amount will get the

additional 0.4 (budget units), making their budget 1.0 (budget unit), which is the value of the 3rd quartile of the budget distributions. Group B whose treatment is receiving less budget amount will get 0.2 (budget units) less, making their budget 0.4 (budget units), which is the value of the 1st quartile of the budget distributions. We will use the median, and quartile amount as our original data is right-skewed, which means that the mean will not fully reflect the central tendency.

By changing the assigning budget to different groups, we want to measure the change in the Rating with respect to the budget between the control and treatment groups.

Causal inference compares two sets of potential outcomes of the same units.

d. Dependent variable(s) and independent variable(s) in the modeler

Dependent variable: The project final budget ("RevisedAmount")

Independent variable: The project's Rating

e. And---Why would running RCTs and modeling/optimizing over RCT results be preferable to using (observational, non-RCT) "foo" data?

If we only conduct observational or non-RCT studies, there will be variables those we did not control or did not use randomized sampling to distribute. Those confounding variables could affect the results. Therefore, we cannot conclude the causal link between the budget size and the success (rating) of the projects.

In ideal case for RCT, at least, the projects should be identical in all variables those we can control and then get randomly assigned to the treatment and control groups to avoid selection bias. Also, randomization ensures that characteristics that might affect the relationship between the dependent and independent variables will be roughly equal across all arms of the study, minimizing allocation bias. From RCT, we can conclude more plausibly the causal

Link to the code:

https://github.com/viethoangtranduong/R-Competency-and-The-Drivetrain-Approach-to-Decision-Making/blob/master/R%20Competency%20and%20The%20Drivetrain%20Approach%20to%20Decision%20Making.R