CS112 - R Competency and The Drivetrain Approach to Decision Making

Link to R code: https://github.com/anggunberlian/cs112/blob/master/CS112%20Assignment%201.R

(1)

(a) For the claim that project approval is generally about 24 months, it is not true because the mean or average project duration (between the original project completion date and the approval date) is 21.7 months, and the median is 20 months. Having the median with a similar result with the mean, we can infer that the dataset is roughly equally distributed from the lowest and highest values. Hence, it is more appropriate to say that the project duration is generally 20-22 months.

For the claim that the project can be extended, it is true because 1) the average project duration between revised completion date and the approval date is 40.6 months which is certainly larger than the average original project duration, 21.7 months.

To the question: if the length of the project delay has changed over time, the answer is no. From F*igure* 1., we can see that time to time, the length of the delay have mostly been similar to the data points cluster around the same area.

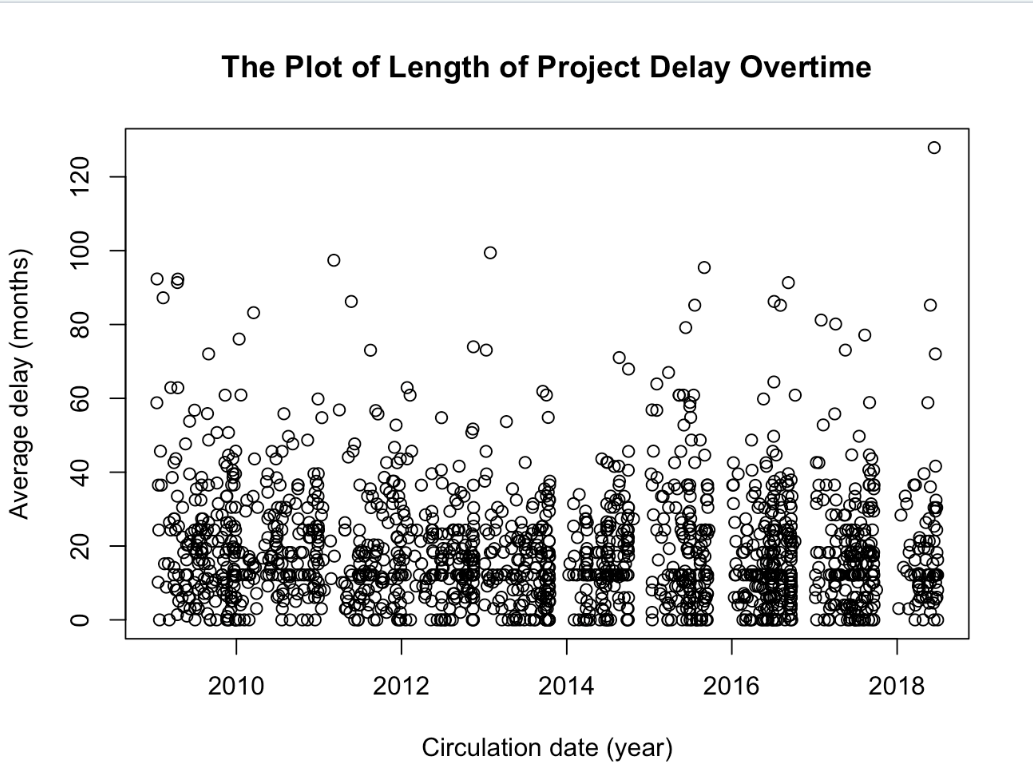
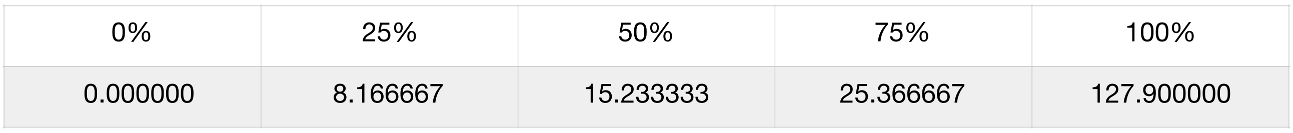


Figure 1.

Additionally, the mean and median of length project delay is 18.96 or 19 months and 15.2 months, while the interquartile ranges are as shown in *Figure 2*. To find those values in R, I need to exclude the NAs in which I used na.rm argument (na.rm = TRUE) that could remove the missing values and allow me to perform other functions to calculate non-missing values.



0%	25%	50%	75%	100%
0.000000	8.166667	15.233333	25.366667	127.900000

Figure 2.

c. The difference between planned project duration and actual project duration is illustrated in *Figure 3,* showing that the two graphs do not overlap*.* The planned duration (red line) has its peaks around the 20-month duration, while the actual duration (blue line) has its peaks around the 40-month duration. In addition to the visualization, here is the means, medians, and interquartile ranges are shown in *Figure 4*. The interquartile ranges for actual duration are always higher (e.g., 25th percentile for planned duration is 13.5 months while it is 28 months for the actual duration), supporting the insight that projects are completed much later than planned.

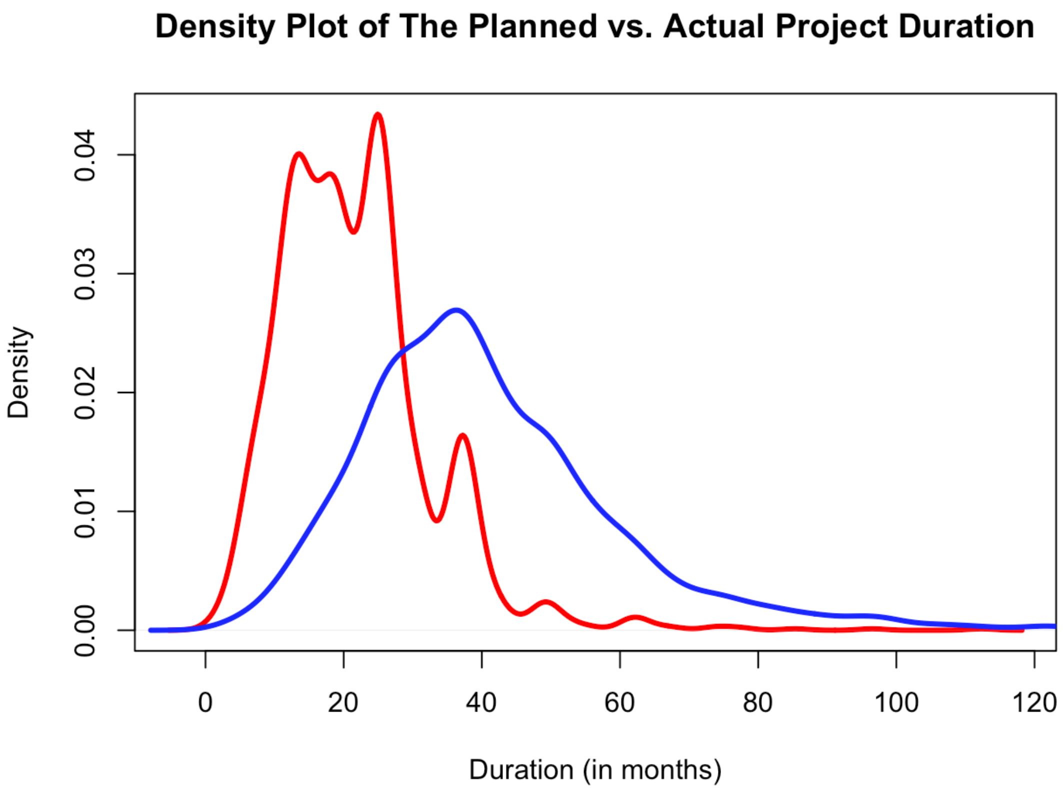


Figure 3.

	Planned Duration (in months)	**Actual Duration (in months)**
Mean	21.7	40.6
Median	20	37.3
Interquartile 1 (0%)	0.6	1.9
Interquartile 2 (25%)	13.5	28
Interquartile 3 (50%)	20	37.3
Interquartile 4 (75%)	26.4	49.3
Interquartile 5 (100%)	112.3	146.5

Figure 4. Table

(2) Projects that are completed from 2010 to the present time generally get rating 2 or the second-highest rating. Projects in that category also get relatively low percentage low rating (e.g., 2% for 0 ratings).

	Percentage ratings for projects completed from 2010 - now
Rating 0	2%
Rating 1	11%
Rating 2	72%
Rating 3	14%

Figure 5. Table of Percentage Ratings for Project Completed between 2010 to Present Time.

(3) From *Figure 6.* below, PATA projects mostly did not get any ratings, as the highest percentage rating is 12% for rating 2. In other words, for PATA projects, they have many missing values for their rating.

	Percentage ratings for PATA Projects
Rating 0	0%
Rating 1	1%
Rating 2	12%
Rating 3	3%

Figure 6. Table of Percentage Ratings for Policy and Advisory Technical Assistance ("PATA") Projects.

(4) After identifying the projects that are on top and bottom 10% of the final budget (revised amount), I wanted to see if there is any effect on budgets to rating. Hence, I plotted the ratings for both categories. The idea is that, if the top 10% got a higher rating and the bottom 10% got a lower rating in general or vice versa — it might indicate a relationship between budget and rating. However, as shown in *Figure 7.,* there is an overlap between the red line, which represents the top 10% with the blue line that represents the bottom 10%. Hence, it indicates that how much budget a project gets does not affect the ratings.

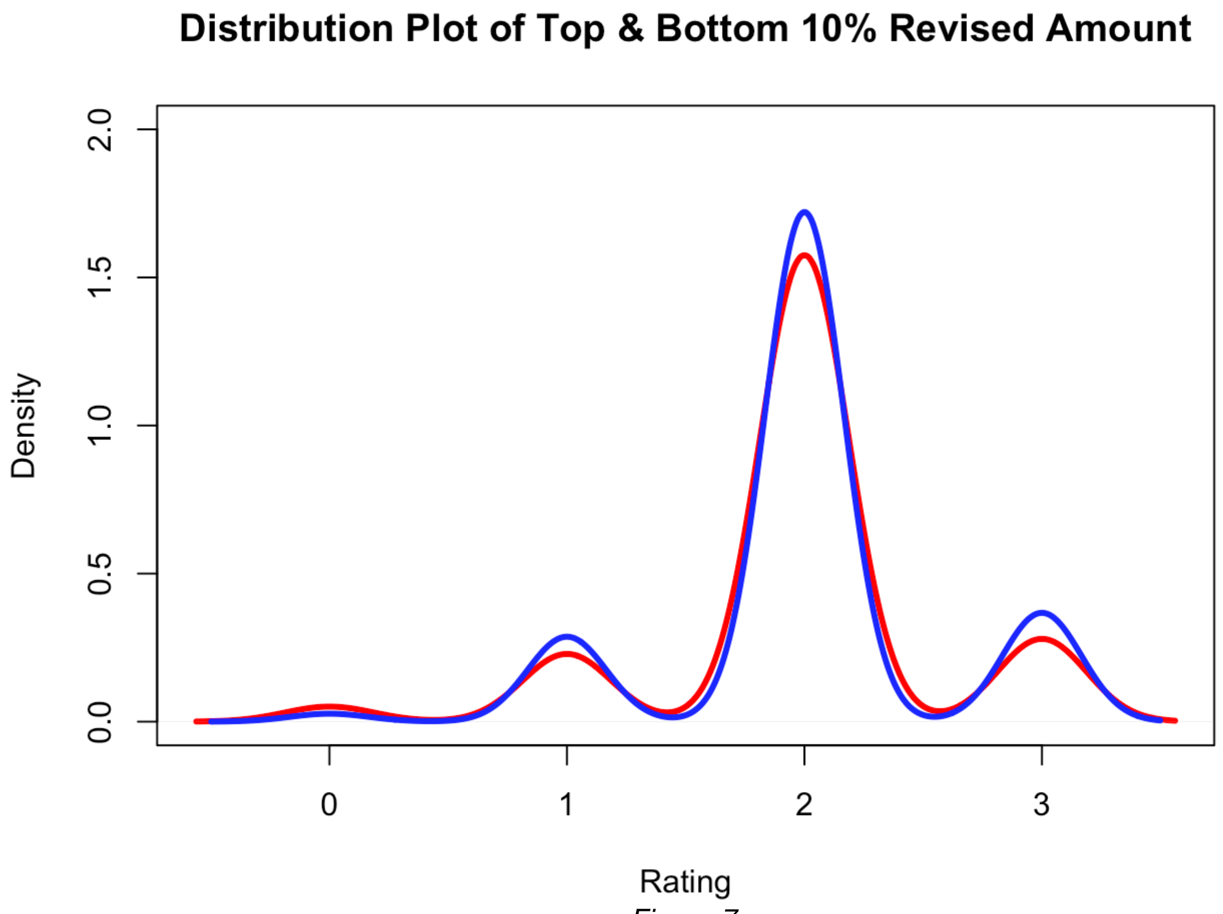


Figure 7.

Additionally, I created a bar plot comparing the two (top and bottom) according to their "Dept" characteristic. The idea behind it is that if the two groups have similar representation on their characteristic(s), then comparing the two is plausible. However, through *Figure 8.,* we can see that the two groups are mostly coming from different departments. Therefore, the two are not comparable.

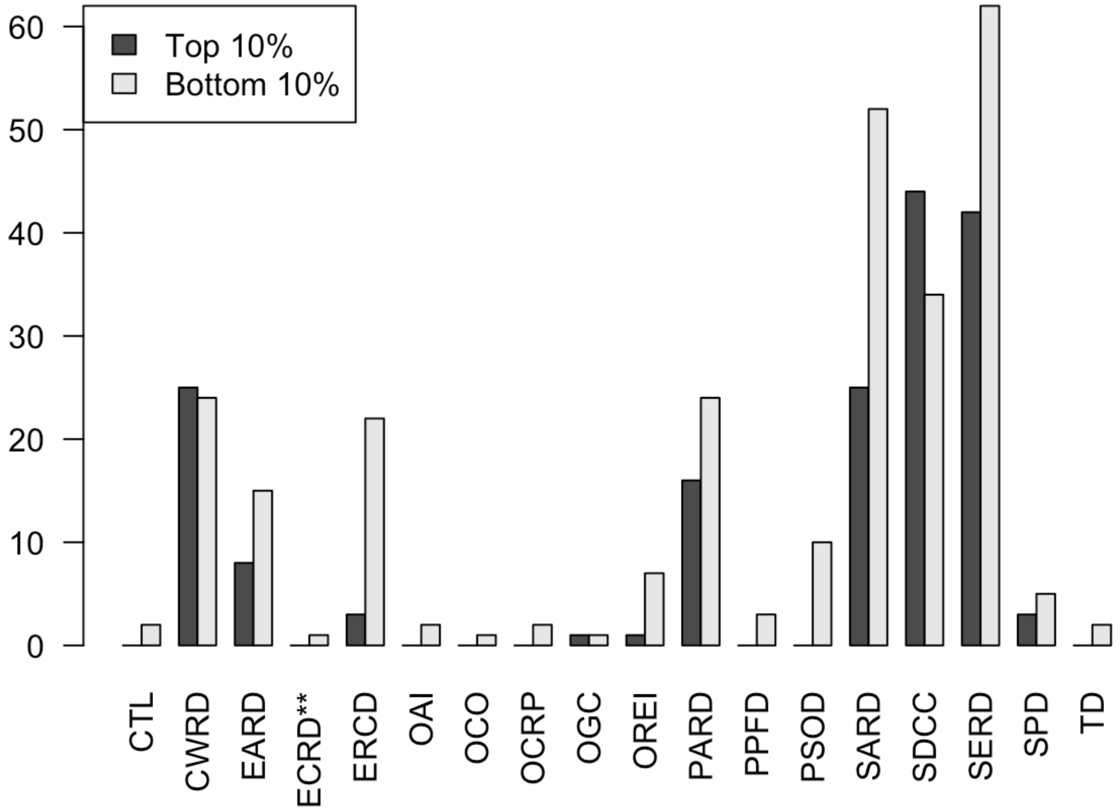


Figure 8. A bar plot comparing the type of "Dept" (department) for top and bottom 10% projects.

(5)

(a) Decision problem or objective?

The objective is to set the *right budget* in which that it is enough to run the project but it does not allow the project to delay.

(b) Lever or levers?

The levers are aspects/variables that we could change to achieve our objectives. In this case, the lever would be the project budget. We need to find the project budget range that would allow the project to run (e.g., not too short on cash) and be completed on time.

(c) Ideal RCT design?

The ideal RCT design would be random assignment of project budget to high number of projects. This would allow us to collect data in the end and analyze the pattern, e.g., how much project budget needed to complete a project on time? or is it possible at all to find it? could it be based on other characteristics, e.g., projects with X conditions should get X amount, while projects with Y conditions needs to get Y amount.

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

A modeler is a model that represent the underlying things that we want to know. In this case the dependent variable would be the project budget while other things would be the independent, e.g., project type, region, department, etc.

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

It is preferable as RCT allows us to randomize the assignment, and it requires many data points. Therefore, with RCT, we have better chance of modeling how things would work in real world, compared to non-RCT that could be biased (e.g., observational) or that they are too small of pool of data and might not represent the real world.