CS112 - Assignment 1

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

Question 1

(a) We can observe that the mean number of days for project duration is 643.6 days which is approximately 21.1 months. Looking at the median (592.0 days ≈ 19.5 months) and the interquartile range of durations, we can see that the most projects in the zone of fewer than 24 months. Figure 1 visualizes this claim. From the descriptive statistics discussed above, we can conclude that it is false to say that the project duration at approval is generally about two years (24 months); instead, it is likely to be around 592 days or 19.5 months (as the median).

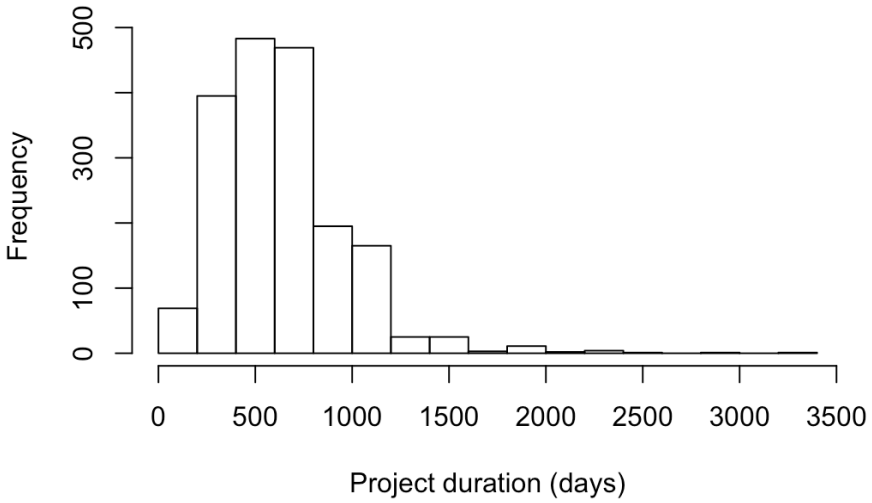

***Fig. 1*: Histogram of distribution of project durations (the number of days between original completion date and project approval date) with circulation date after 2008-01-01.**

Comparing project durations of earlier and later circulation dates has provided no observable change in general trend as seen in Fig. 2.

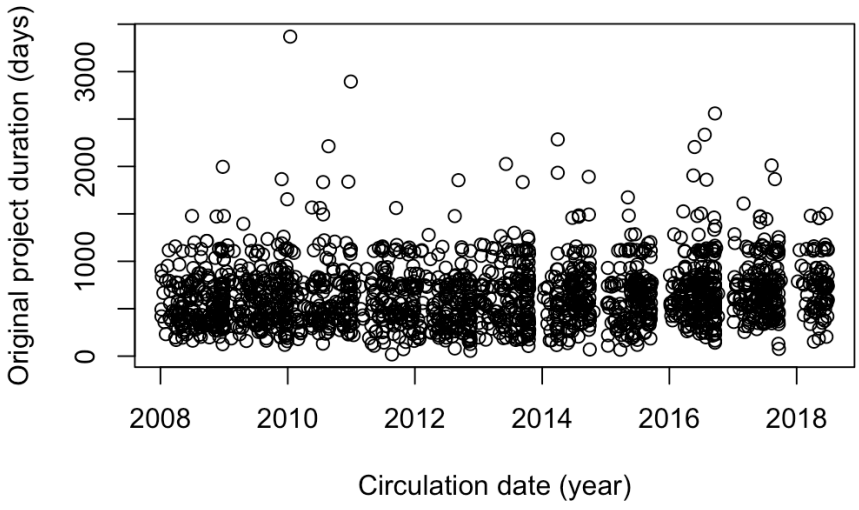

***Fig. 2*: Original project duration in projects with varying circulation date.**

(b) To calculate how the project duration has changed, I subtracted corresponding original completion dates from revised completion dates. The mean for this result is 573.6 days, whereas the median (50% quantile) is 485 days. This indicates that generally, projects are rescheduled for 485 days but there are rare exceptions (up to about 3837 days, 4th quantile) that “stretch” the mean larger. Also, we must note that there is no project with revised completion date earlier than original, so this disallows negative values to pull the mean to the median. These observations suggest that on average, projects are extended by 573.6 days (mean), but if the randomly selected project were to be extended, it would most likely be stretched by about 485 days (median).

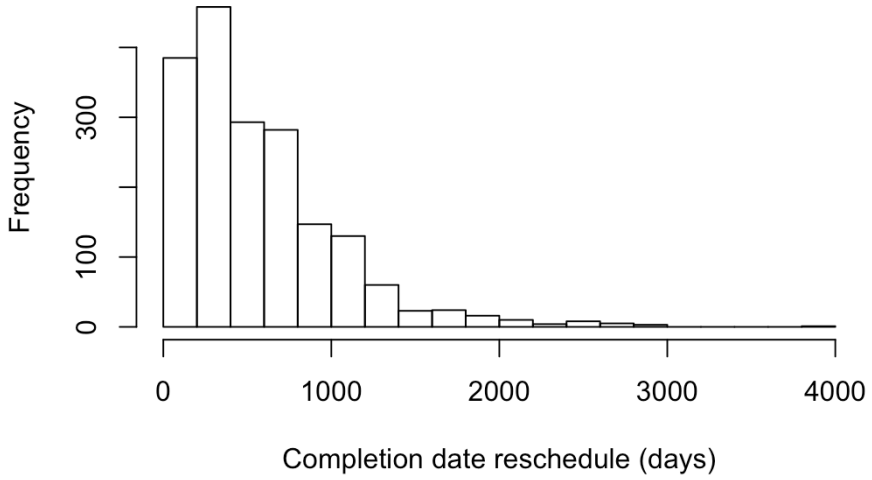

***Fig. 3*: Histogram of distribution of change in completion dates (the number of days between revised completion date and original completion date) in projects with circulation date after 2008-01-01.**

Question 2

Rating (0-3)	0	1	2	3
Fraction	0.02675227	0.1589085	0.682718	0.1316212

***Table 1*: Distribution of project ratings among all projects with circulation date after 2008-01-01.**

The most common rating for projects is “2” which is the rating for about 68% of projects. In descending order then come “1” (at 16%), “3” (at 13%), and “0” (at 3%). All the fractions in Table 1 should add up to 1 because all projects have a rating.

Question 3

Rating (0-3)	0	1	2	3
Fraction	0.01992032	0.1405805	0.7023335	0.1371656

***Table 2*: Distribution of project ratings among all projects that are not of type “PPTA” with circulation date after 2008-01-01.**

The most common rating for projects is “2” which is the rating for about 70% of projects. In descending order then come “1” (at 14%), “3” (at 14%), and “0” (at 2%). All the fractions in Table 2 should add up to 1 because all projects that are not of type “PPTA” have a rating.

Question 4

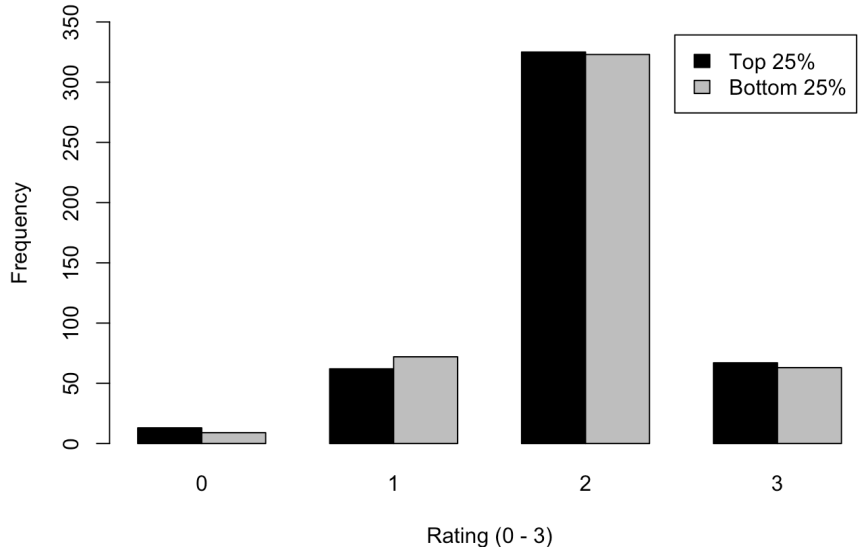

***Fig. 4*: Bar plot of rating frequencies in top 25% funded and bottom 25% funded projects**

First of all, looking at the distributions of ratings, we can see that there is no observable effect of the amount of funding on the ratings. To be specific, the mean rating of the top 25% funded projects is 0.6% bigger than that of the 25% bottom, while the mean funding is more than nine times larger. Secondly, even if we assume that there was an observable effect, the conclusion of causality is faulty because the fundings are correlated with different countries, departments, and other aspects of the projects. This means that we cannot fixate the control variables to establish a causal relationship between funding and ratings. Figure 5 shows such discrepancy in one of the control variables.

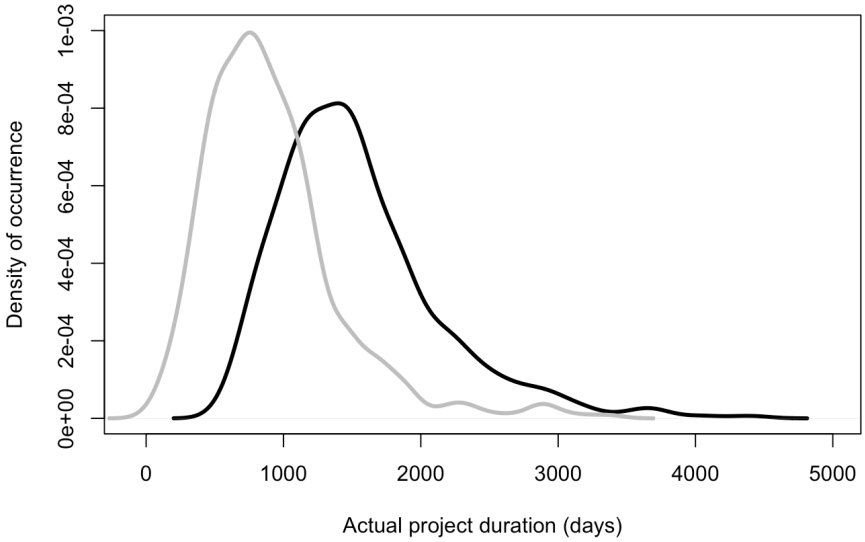

***Fig. 5*: Comparison of project durations in top 25% (black curve) and bottom 25% (grey curve) in terms of funding.**

Question 5

(a) Our objective would be to maximize the project success, measured with the rating.

(b) Imagining that we are assigned the country, department, type, and other characteristics of the projects (e.g., we're a Freelancer Data Scientist working being hired for a particular project). The levers we can tweak to affect the rating are the revised amount and the project duration. As we throw in more money, we would be able to hire better quality machinery and employees, but we would want to have the lowest possible optimal budget. When project duration is long, we would need to pay more, whereas if it's too short, we might not be able to meet the highest quality.

(c) If we were to have a lot of money, we could simultaneously start projects of the same type (e.g., we need a lot of parks!) but with revised amount and project duration pulled out from a joint uniform random distribution of, for example, [\$0; $\$10^8$] x [0 day; 5000 days]. When they finish building, we will get evaluations from park experts to see the caused rating.

(d) In this modeler, the dependent variable is the rating (the objective function), and the independent variables are the levers that we can control, like the revised amount and project duration.

(e) This is because foo data contains too many variables that can be correlated, and we are not able to look at those and determine the causation. Example: large revised amount and large project duration are correlated with a high rating, but from this, we cannot know which lever contributes more and by how much so that we can set an optimal balance of both.