# Feedback for: What makes an R-package popular?

**Title abstract (5%)**

**Total marks:** 5%

I believe the title is adequate. The abstract discusses the aims, methods, data and findings.

**Background, motivation (10%)**

**Total marks:** 7%

The objectives of the project are fairly clear, and there is a brief discussion on previous studies that have explored a similar issue. However, the report does not explicitly address why it is important to study the factors that drive package download numbers.

For instance, all aims in section 5 are well explored and the results are interesting, but the implications (or benefit) that each of the results could have in a wider context are not discussed explicitly. Does knowing that a high number of GitHub commits is associated to high number of package downloads mean that users can make their packages more popular by submitting more commits?

**Objectives, significance (10%)**

**Total marks:** 8%.

The objectives of the project are clearly presented. To analyse the download patterns of R and R-packages, and to study how those patters vary across packages, release dates, names, update properties, etc. The contributions of the analysis are well articulated, although it’s hard to follow the exposition of tables and figures at times, as the student jumps straight into the interpretation before explaining the figure (Figures 5.9 and 5.10).

**Methodology, data, results, discussion (40%)**

**Total marks:** 37%.

The methodology used to analyse the data is well presented and discussed, and the visualization tools used to deal with different type of data challenges are adequate.

Most figures are very interesting. However, other figures such as that in Figure 5.4 are not clearly motivated. The student not only presents one type of figure, but when the data requires it, different type of visualizations that are more robust to the type of data are used (Figure 5.15). I would have appreciated more detail in the construction of complex figures such as the Lorenz curve figure.

**Conclusion (10%)**

**Total marks:** 7%.

The conclusions presented a brief description of the results and findings of the data analysis.

However, I think it requires a bit more discussion on the practical implications.

The caveats and drawbacks of the approach are not discussed. For instance, all the outputs produced in this report only measure association between number of downloads and other important characteristics of R and R-packages. However, it is hard to tell the direction of this causal relationship. For instance, do more GitHub commits lead to a larger number of downloads, or does a larger number of downloads implies that the package must be more reliable and thus more commits are necessary to implement corrections and improvement of bugs. Also, it is important to mention that the positive association of variables such as the total number of downloads and the total number of updates is probably mostly driven by the time variable, as both these variable increase over time.

**Technical details (10%)**

**Total marks:** 8%.

The data is described in section 3. The data is accessed through the package cranlogs.

However, some details of the data are still not clear. For instance, I do not fully understand yet why there are over 700000 CSV files collected. It would be useful to know what each of these files corresponds to.

**Writing and presentation (15%)**

**Total marks:** 12%.

The student has done a good job with the writing and the report is easy to read.

There is still some room for improvement. See below some illustrations of sentences that could be improved.

“Anyway, we could generally believed that a relatively popular R-package…”

“Alphabetical order played little roles in promoting the download volume.”

**Total mark: 84**