Uncovering the Secrets of the Maven Repository

Contact

Mail: N.H.C.Tomassen@student.tudelft.nl
Supervisor: Mehdi Keshani
Responsible Professor: Sebastian Proksch
Delft University of Technology

Introduction

Maven Central is repository that contains a large number of common libraries used in Java. It serves as a place where developers can store and share projects made in Java. This research will analyze Maven Central, specifically the space requirement of the libraries. This analysis provides insights into the average size of libraries and how to reduce size for future library maintainers. This is beneficial as smaller artifacts take less space in the repository and requires users to download smaller files.

Research Questions

1. How big is an average library on Maven?
2. How is the space requirement of the libraries distributed in the ecosystem?
3. What are the reasons for the larger sizes?

Key Terms

- .m2 folder: the local directory where maven caches libraries downloaded from the central repository.
- JAR: Archive files similar to zip used to package Java classes and related metadata and resources.
- POM: XML file that contains information about a project and is used by Maven to build the project.

Methods

- Figure 1 shows a high level overview of the architecture used in this research.
- First the Maven Index is read and stored in a table
- Then, we use a data selection strategy to select libraries from this table. The distribution of our sample can be seen in Figure 3, and the distribution of packages on Maven Central in Figure 2.
- After selecting which packages to analyze, their jars and pom files are downloaded into the local .m2 folder by the resolver.
- The individual jars and pom files are analyzed by extractors and metadata is stored in a database. This metadata includes the size of the package, the number of files, the number of direct dependencies and the number of transitive dependencies.

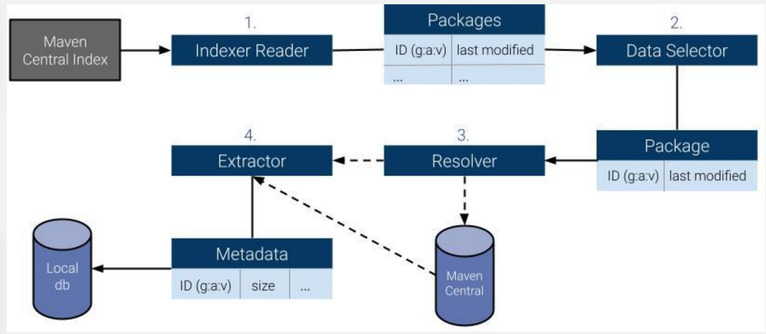


Figure 1: High level component diagram of main architecture

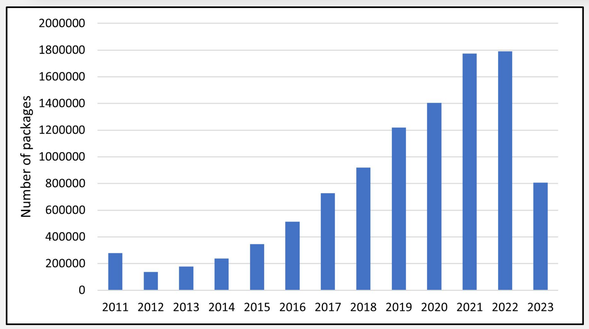


Figure 2: Distribution of packages released on Maven Central per year

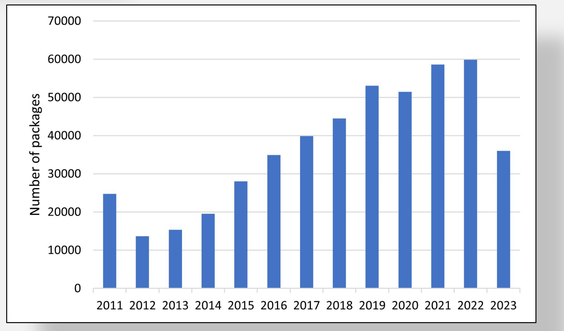


Figure 3: Selected distribution of packages in our sample

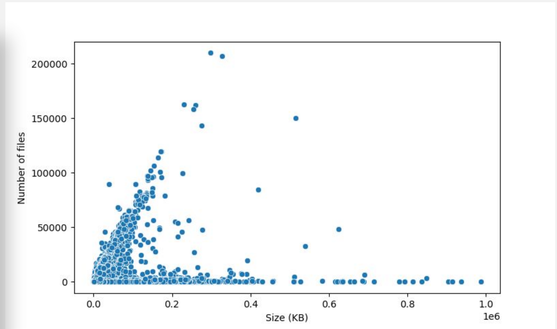


Figure 4: Correlation between size and number of files

Results & Discussion

- Average space requirement of a library is 1447 KB, the median is 25.9 KB and the largest package is 986,867 KB.
- Figure 5 shows the distribution of size in Maven Central.
- There are two distinct categories of artifacts in Figure 4: Those that slightly increase in size when they add more files and those that don't have that many files to begin with and are usually small but can become very large too. We manually analyzed the cases where libraries are massive but the number of files is low.
- The analysis revealed that almost all these packages are machine learning or big data projects with massive data files.
- We recommend not including files like this in the package and instead uploading them somewhere else and document how to connect to them.
- The threats to validity in this study are that not every package in our sample could be analyzed because in some cases, the parent pom was hosted on a different repository. Furthermore our data selection strategy does not analyze every version of each package but only one version per package.

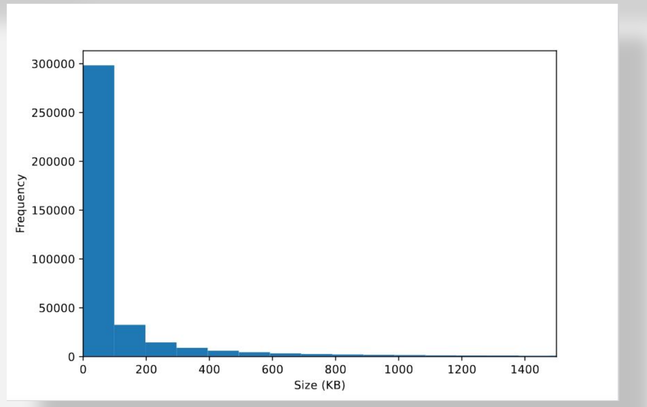
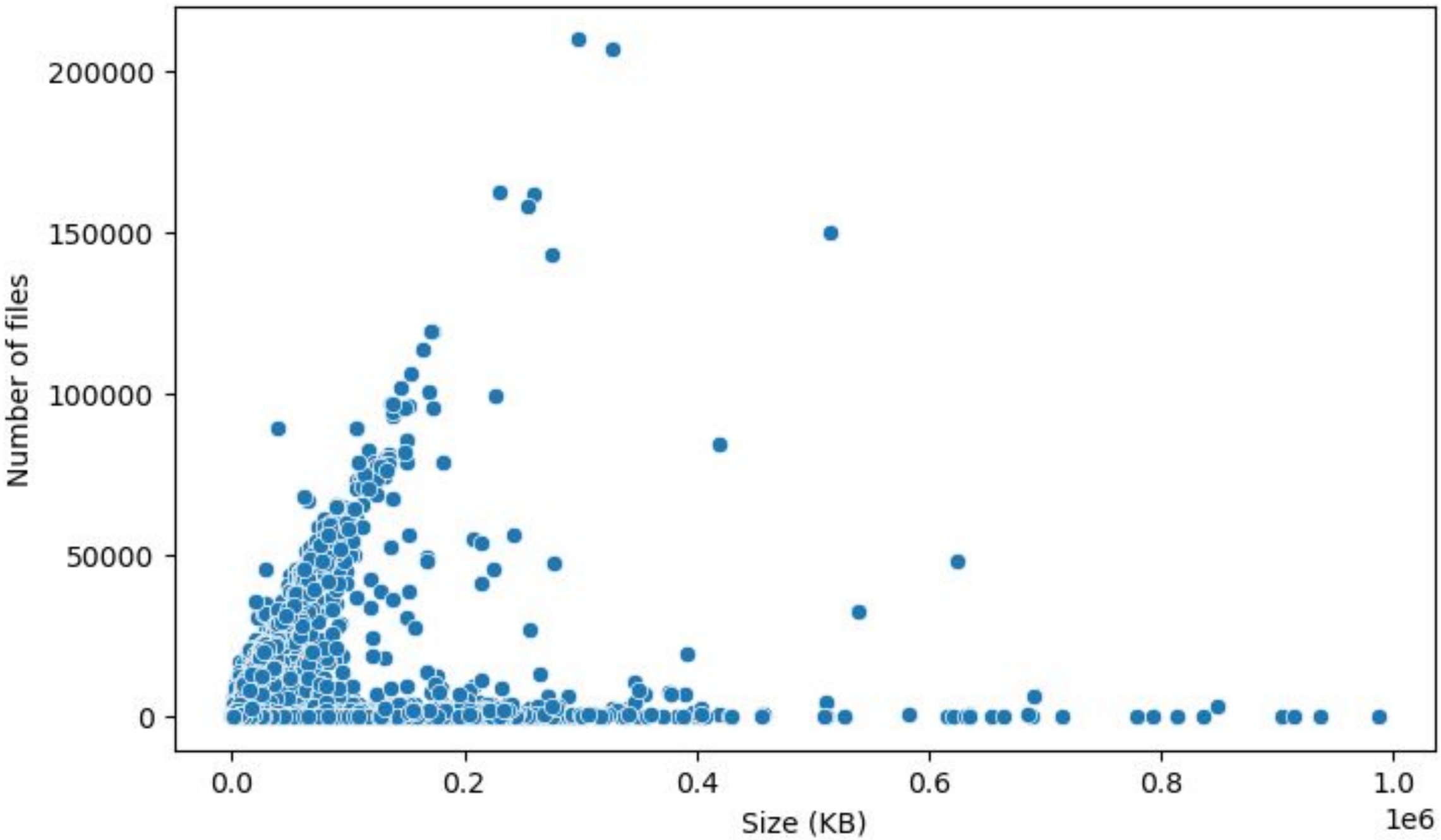
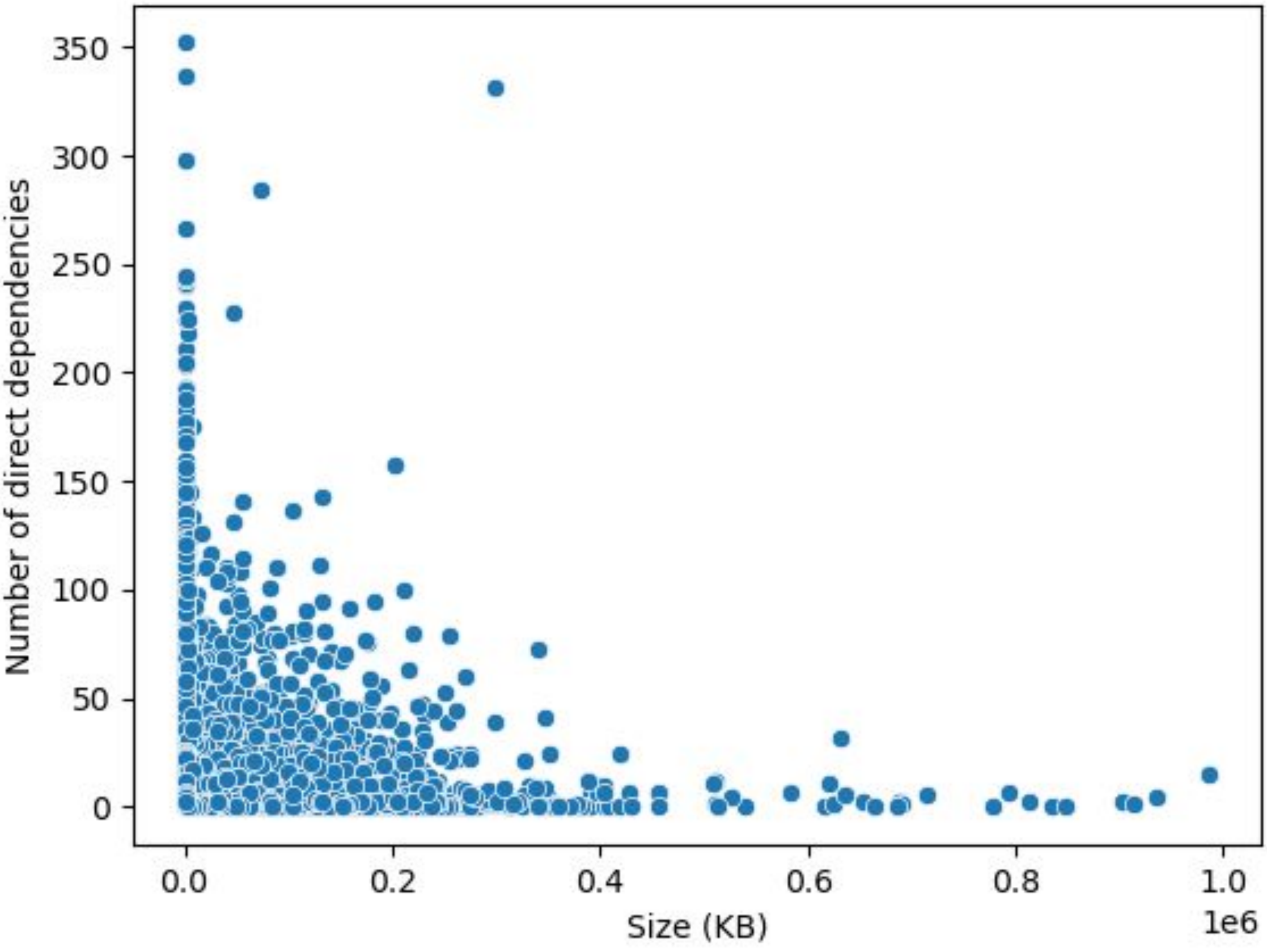


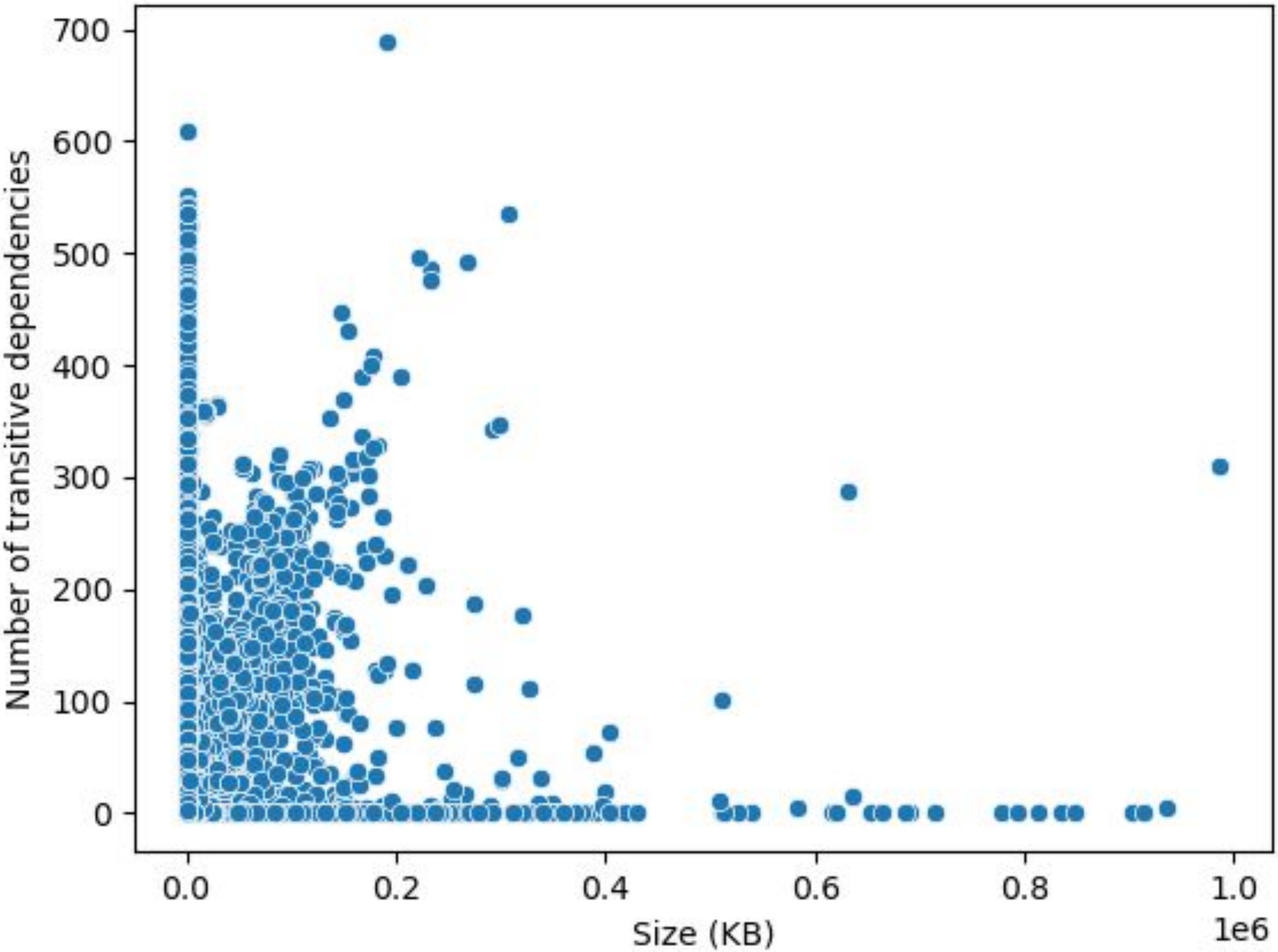
Figure 5: Distribution of size in Maven Central

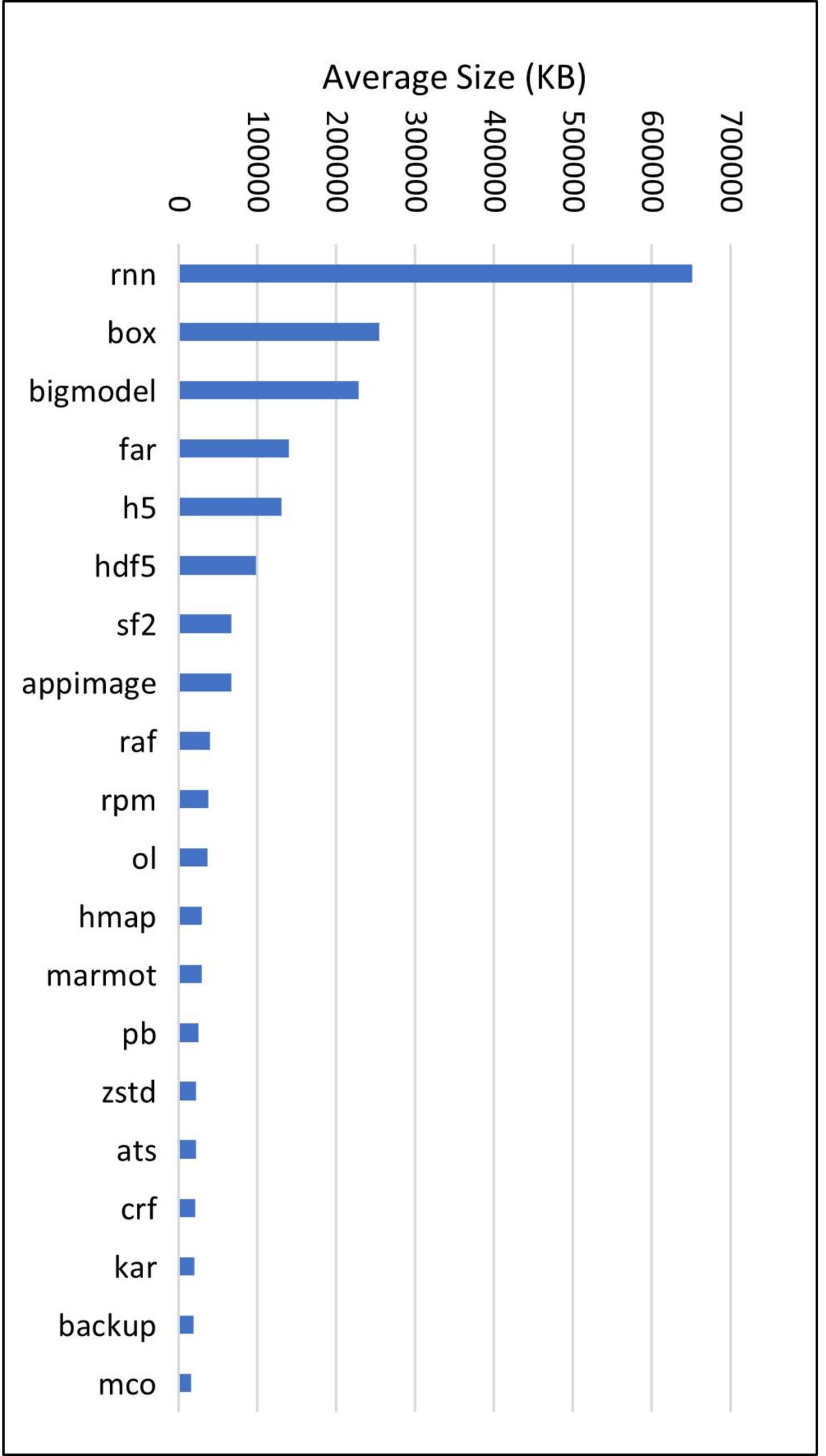
References

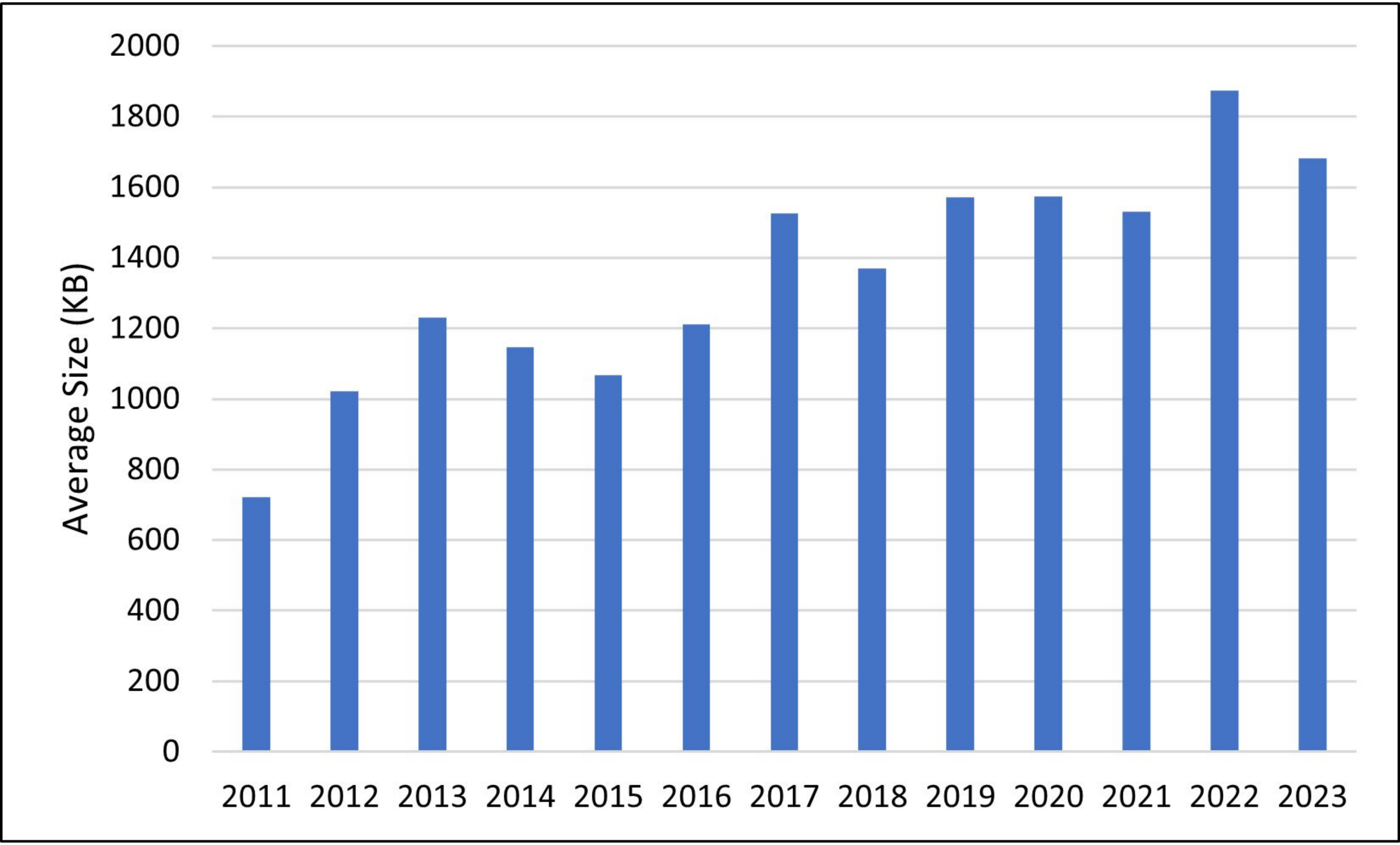
[1] Cesar Soto-Valero, Nicolas Harrand, Martin Monperrus, and Benoit Baudry. A comprehensive study of bloated dependencies in the maven ecosystem. 2021.
[2] Dimitris Mitropoulos, Vassilios Karakoidas, Panos Louridas, Georgios Gousios, and Diomidis Spinellis. The bug catalog of the maven ecosystem. 2014.


200000
150000
100000
50000
0
Number of files
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Size (KB)

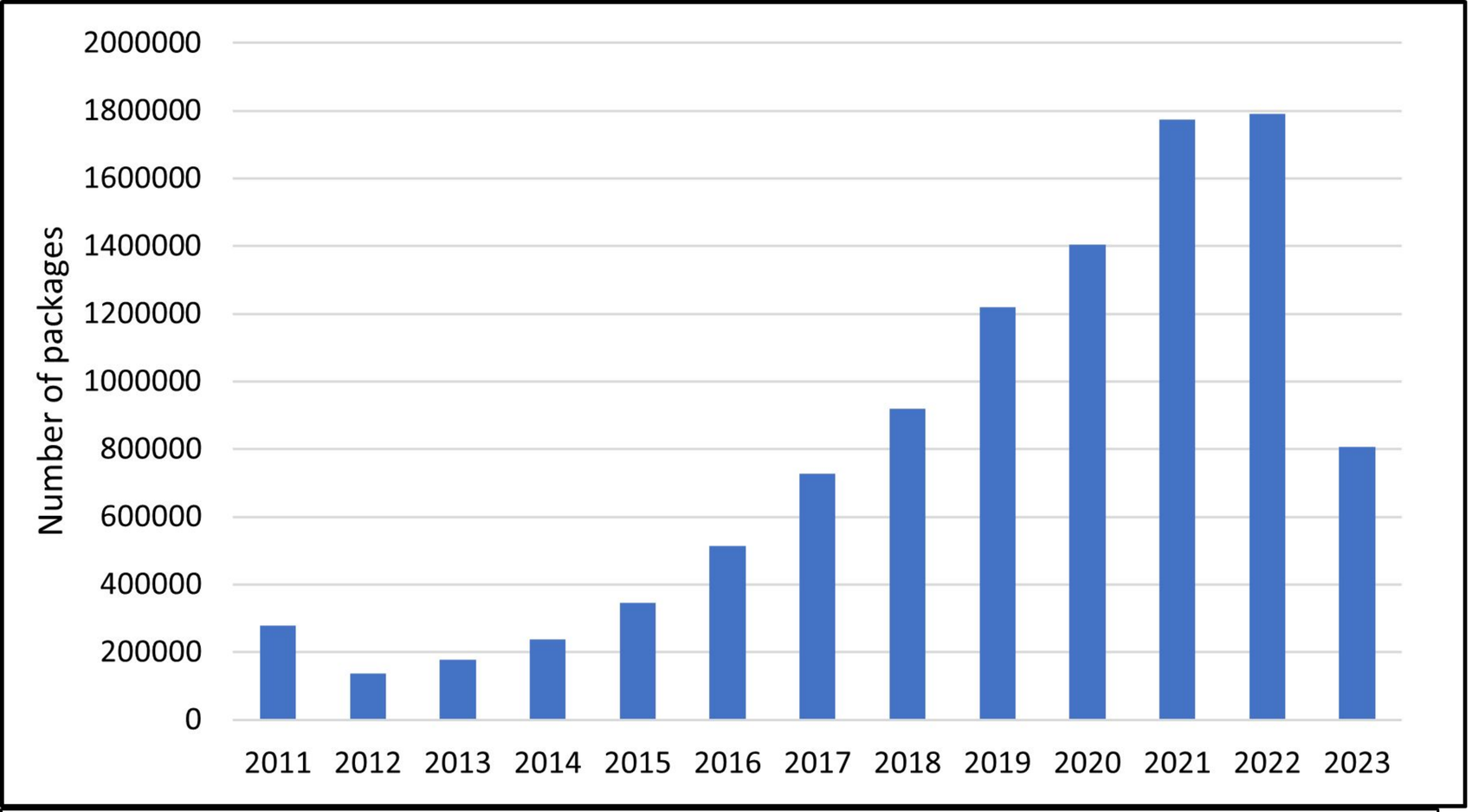

Number of direct dependencies
350
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Size (KB)

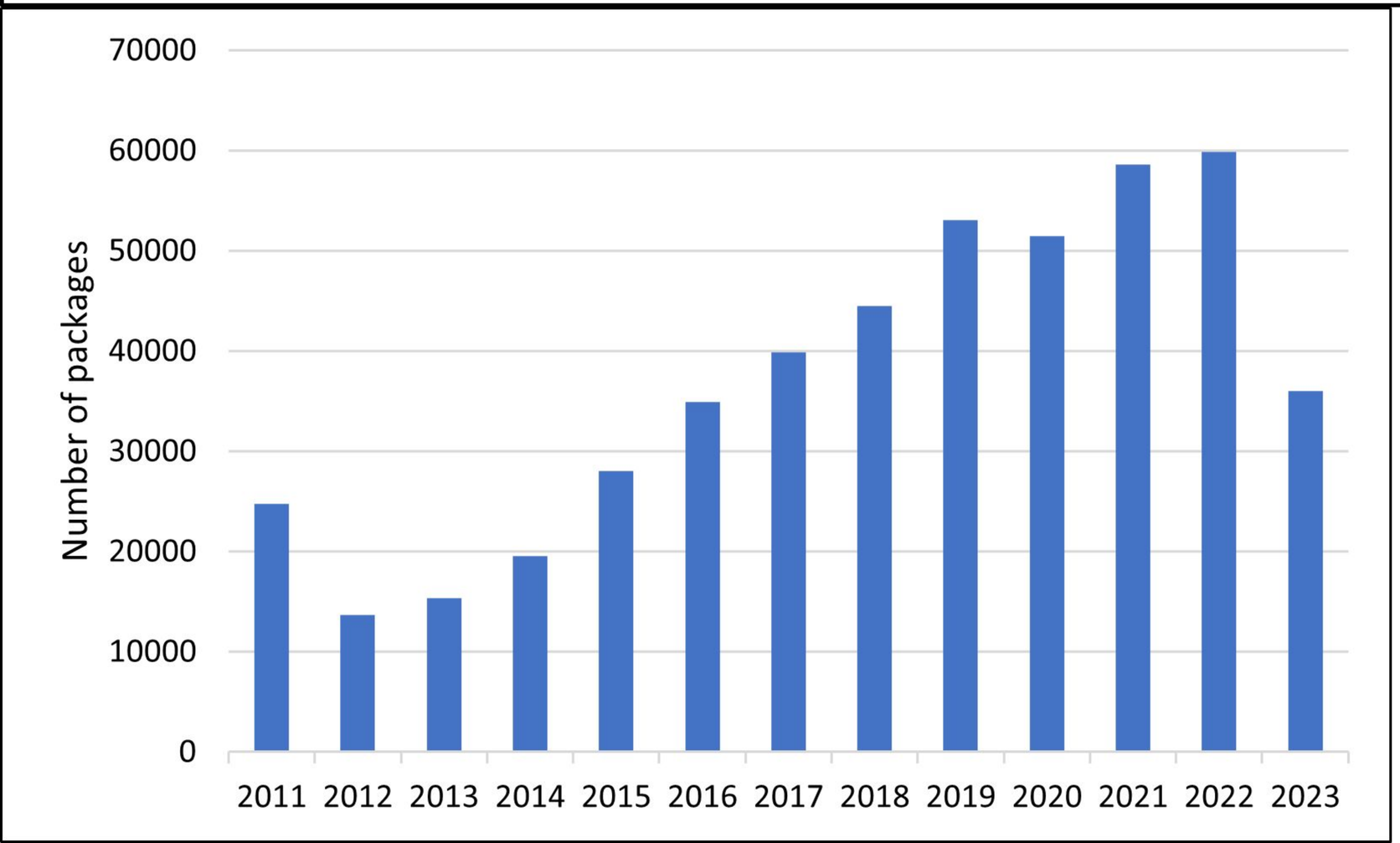

Number of transitive dependencies
700
600
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Size (KB)


Average Size (KB)
0
100000
200000
300000
400000
500000
600000
700000
rnn
box
bigmodel
far
h5
hdf5
sf2
appimage
raf
rpm
ol
hmap
marmot
pb
zstd
ats
crf
kar
backup
mco


Average Size (KB)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023

class
sjsir
js
png
xml
nir
properties
noextension
jar
html
tasty
java
res
txt
xsb
mf
json
svg
gif
css


Number of packages
2000000
1800000
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023


Number of packages
70000
60000
50000
40000
30000
20000
10000
0
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023