PROJEKT 1

Programowanie i analiza danych dla zaawansowanych w R

Zainteresowanie w ujęciu czasowym ⑦   ⬇ ⟨⟩ ↗

100

75

50

25

Średnia   1 sty 2004   1 cze 2010   Uwaga   1 lis 2016   Uwaga
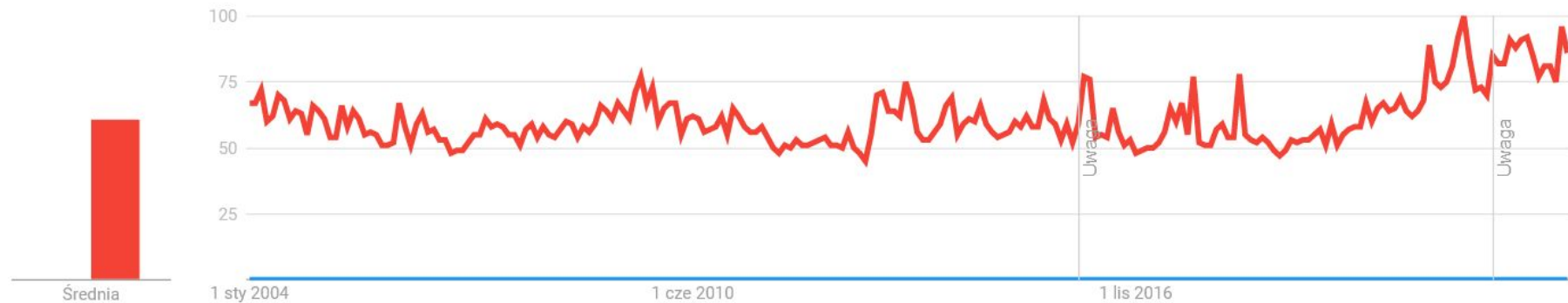
The Landscape of R Packages for Automated Exploratory Data Analysis
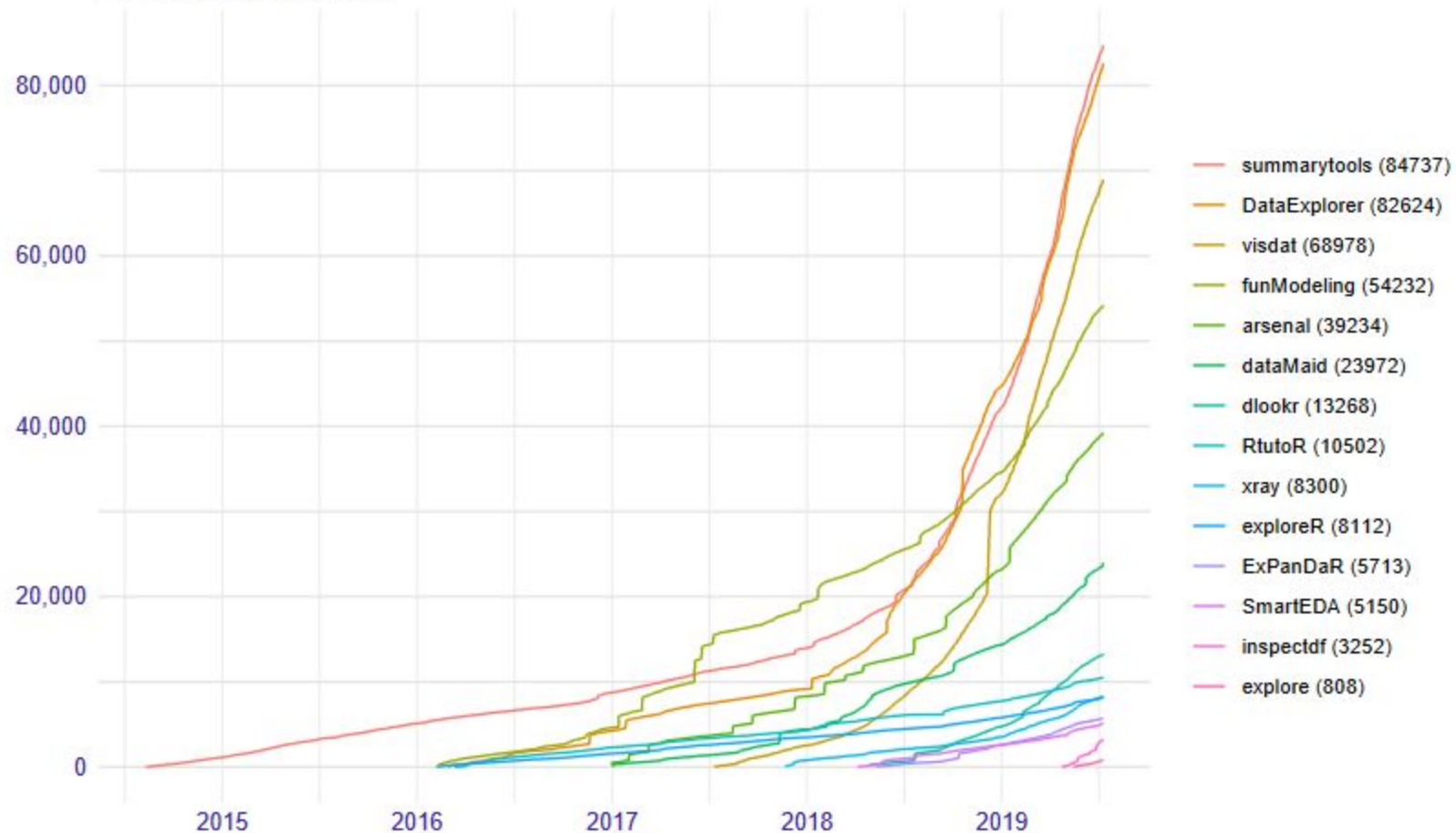
by Mateusz Staniak and Przemysław Biecek

Abstract The increasing availability of large but noisy data sets with a large number of heterogeneous variables leads to the increasing interest in the automation of common tasks for data analysis. The most time-consuming part of this process is the Exploratory Data Analysis, crucial for better domain understanding, data cleaning, data validation, and feature engineering.

There is a growing number of libraries that attempt to automate some of the typical Exploratory Data Analysis tasks to make the search for new insights easier and faster. In this paper, we present a systematic review of existing tools for Automated Exploratory Data Analysis (autoEDA). We explore the features of fifteen popular R packages to identify the parts of analysis that can be effectively automated with the current tools and to point out new directions for further autoEDA development.

1) charakterystyka istniejących pakietów R służących do zautomatyzowania analizy danych i porównanie ich możliwości

2) identyfikacja obszarów, w których AutoEDA mogłaby zostać ulepszona, ocena potencjału

Total number of downloads

Based on CRAN statistics

- summarytools (84737)
- DataExplorer (82624)
- visdat (68978)
- funModeling (54232)
- arsenal (39234)
- dataMaid (23972)
- dlookr (13268)
- RtutoR (10502)
- xray (8300)
- exploreR (8112)
- ExPanDaR (5713)
- SmartEDA (5150)
- inspectdf (3252)
- explore (808)

Total number of downloads
Based on CRAN statistics, 27.02.2023

arsenal (280930)	dlookr (156991)	exploreR (28880)	RtutoR (26876)	visdat (1077517)
DataExplorer (626895)	ExPanDaR (70532)	funModeling (224318)	SmartEDA (486034)	xray (47843)
dataMaid (82441)	explore (64283)	inspectdf (71918)	summarytools (667126)	

900000

600000

300000

0

2014   2016   2018   2020   2022
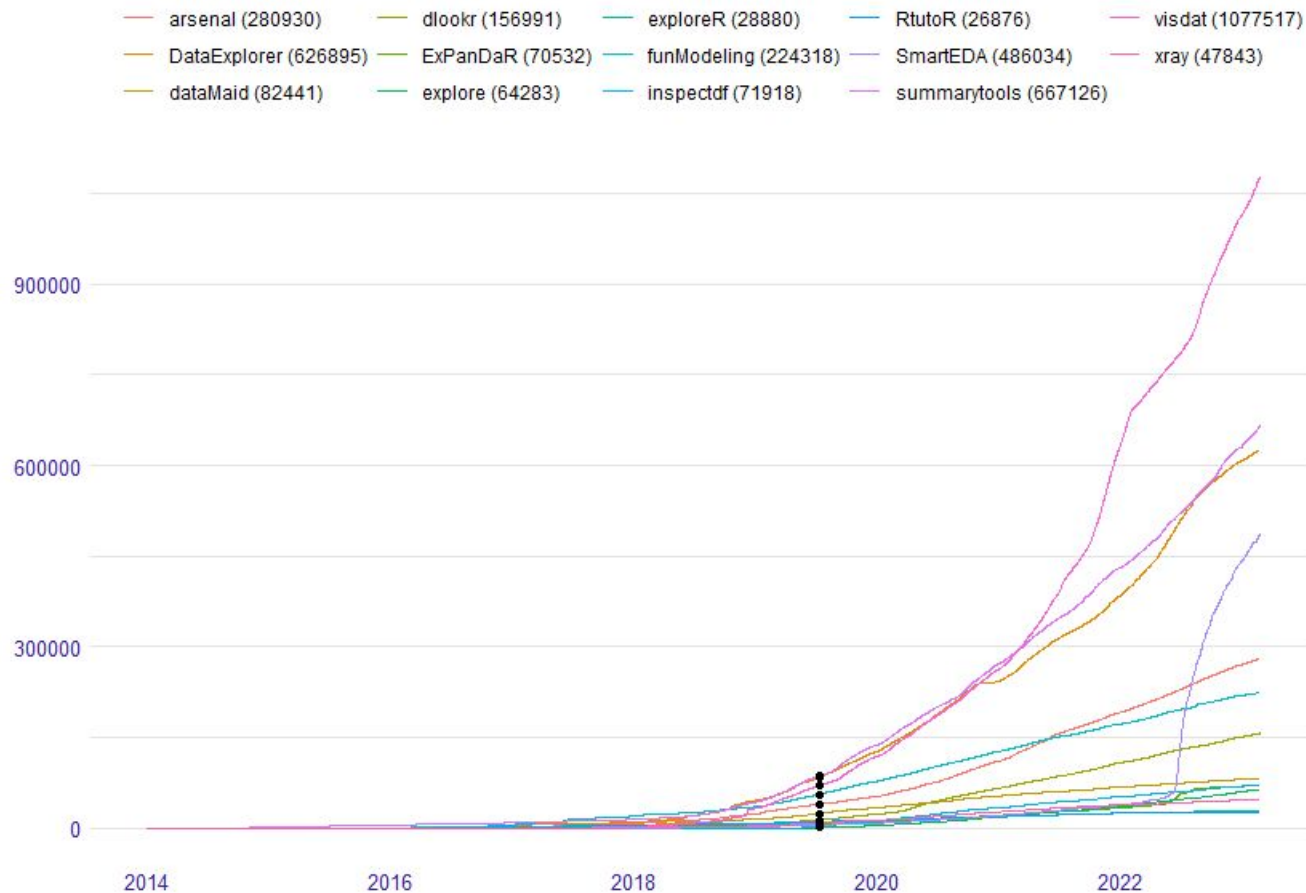
	CRAN			GitHub				
package	downl.	debut	age	stars	commits	contrib.	issues	forks
arsenal	39234	2016-12-30	2y 6m	59	637	3	200	4
autoEDA	-	-	-	41	20	1	4	12
DataExplorer	82624	2016-03-01	3y 4m	235	187	2	121	44
dataMaid	23972	2017-01-02	2y 6m	68	473	2	45	18
dlookr	13268	2018-04-27	1y 2m	35	54	3	9	12
ExPanDaR	5713	2018-05-11	1y 2m	32	197	2	3	14
explore	808	2019-05-16	0y 1m	15	114	1	1	0
exploreR	8112	2016-02-10	3y 5m	1	1	1	0	0
funModeling	54232	2016-02-07	3y 5m	58	126	2	13	18
inspectdf	3252	2019-04-24	0y 2m	117	200	2	12	11
RtutoR	10502	2016-03-12	3y 3m	13	7	1	4	8
SmartEDA	5150	2018-04-06	1y 3m	4	4	1	1	2
summarytools	84737	2014-08-11	4y 11m	255	981	6	76	33
visdat	68978	2017-07-11	2y 0m	313	426	12	122	39
xray	8300	2017-11-22	1y 7m	63	33	4	10	5

Table 1: Popularity of R packages for autoEDA among users and package developers. First two columns summarise CRAN statistics, last five columns summarise package development at GitHub. When a repository owned by the author is not available, the data were collected from a CRAN mirror repository. Data was gathered on 12.07.2019.

package	CRAN downl.	debut	age	GitHub stars	commits	contrib.	issues	forks
arsenal	39234	2016-12-30	2y 6m	59	637	3	200	4
autoEDA	-	-	-	41	20	1	4	12
DataExplorer	82624	2016-03-01	3y 4m	235	187	2	121	44
dataMaid	23972	2017-01-02	2y 6m	68	473	2	45	18
dlookr	13268	2018-04-27	1y 2m	35	54	3	9	12
ExPanDaR	5713	2018-05-11	1y 2m	32	197	2	3	14
explore	808	2019-05-16	0y 1m	15	114	1	1	0
~~exploreR~~	~~8112~~	~~2016-02-10~~	~~3y 5m~~	~~1~~	~~1~~	~~1~~	~~0~~	~~0~~
funModeling	54232	2016-02-07	3y 5m	58	126	2	13	18
inspectdf	3252	2019-04-24	0y 2m	117	200	2	12	11
~~RtutoR~~	~~10502~~	~~2016-03-12~~	~~3y 3m~~	~~13~~	~~7~~	~~1~~	~~4~~	~~8~~
SmartEDA	5150	2018-04-06	1y 3m	4	4	1	1	2
summarytools	84737	2014-08-11	4y 11m	255	981	6	76	33
visdat	68978	2017-07-11	2y 0m	313	426	12	122	39
xray	8300	2017-11-22	1y 7m	63	33	4	10	5

dataReporter

Table 1: Popularity of R packages for autoEDA among users and package developers. First two columns summarise CRAN statistics, last five columns summarise package development at GitHub. When a repository owned by the author is not available, the data were collected from a CRAN mirror repository. Data was gathered on 12.07.2019.

package	downl. prev	downl. now	downl. diff	%	stars prev	stars now
arsenal	39234	272575	233341	6.947418	59	207
DataExplorer	82624	617658	535034	7.475528	235	462
dataMaid	23972	77987	54015	3.253254	68	135
dlookr	13268	156991	143723	11.832303	35	171
ExPanDaR	5712	70532	64820	12.348039	32	134
explore	808	64283	63475	79.558168	15	137
exploreR	8112	25353	17241	3.125370	1	0
funModeling	54535	204760	150225	3.754653	58	88
inspectdf	3252	71918	68666	22.115006	117	237
RtutoR	10502	22509	12007	2.143306	13	0
SmartEDA	5150	486034	480884	94.375534	4	34
summarytools	84737	652983	568246	7.705996	255	466
visdat	68978	1074964	1005986	15.584157	313	428
xray	8300	47054	38754	5.669157	63	73

27-02-2023