

# R Tools for Automated Exploratory Data Analysis

Mateusz Staniak ([mtst@mstaniak.pl](mailto:mtst@mstaniak.pl))

Warszawa, 29 IX 2019



## Navigation

[Current Issue](#)

[Accepted articles](#)

[Archive](#)

[R News](#)

[News and Notes](#)

[Submissions](#)

[Reviews and Proofreading](#)

[Editorial Board](#)

## Subscribe

[RSS Feed](#) 

ISSN: 2073-4859

# The R Journal: accepted article

This article will be copy edited and may be changed before publication.

[The Landscape of R Packages for Automated Exploratory Data Analysis](#) 

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.

Received: ; online 2019-08-17, [supplementary material](#), (1.6 Kb)

CRAN packages: [cranlogs](#), [radiant](#), [visdat](#), [archivist](#), [xtable](#), [arsenal](#), [DataExplorer](#), [dataMaid](#), [dlookr](#), [ExPanDaR](#), [explore](#), [shiny](#), [exploreR](#), [funModeling](#), [inspectdf](#), [RtutoR](#), [SmartEDA](#), [data.table](#), [summarytools](#), [knitr](#), [ggplot2](#), [xray](#), [tableone](#), [describer](#), [skimr](#), [prettyR](#), [Hmisc](#), [ggfortify](#), [autoplotly](#), [gpairs](#), [GGally](#), [survminer](#), [cr17](#), [DALEX](#), [iml](#)

CRAN Task Views implied by cited CRAN packages: [ReproducibleResearch](#), [MissingData](#), [TeachingStatistics](#), [WebTechnologies](#), [Bayesian](#), [ClinicalTrials](#), [Econometrics](#), [Finance](#), [Graphics](#), [HighPerformanceComputing](#), [Multivariate](#), [OfficialStatistics](#), [Phylogenetics](#), [SocialSciences](#), [Survival](#)

---

# The why & what of autoEDA

- Crunchbase lists over 5,000 startups who are relying on machine learning for their main and ancillary applications, products and services today.

- 81% of machine learning startups Crunchbase tracks have had two funding rounds or less with seed, angel and early-stage rounds being the most common.

- According to [KPMG's Venture Pulse Report](#),

venture capital (VC) investment in artificial intelligence almost doubled in 2017, attracting \$12B compared to \$6B in 2016.

- Q2'18 was a second-straight record quarter for total Artificial Intelligence (AI) funding with total investments exceeding \$2.3B including eight mega-rounds over \$100M according to the latest [PwC/CB Insights MoneyTree Report from Q2 2018](#)

<https://www.forbes.com/sites/louiscolombus/2018/08/26/25-machine-learning-startups-to-watch-in-2018/>

## D/SRUPTION

[/ Articles](#) / [Resources](#) / [Magazine](#) / [Events](#) / [Partners](#) / [About Us](#)

### 4) Benevolent AI – pharmaceuticals

**Benevolent AI** is one of many startups disrupting the pharmaceutical industry. The company applies machine learning to improve the way that medicines are discovered, developed, tested and brought to market via several different steps. These include processing and modelling bioscience data, to give scientists hypotheses and ideas to explore; understanding the biology of a disease; finding the best responders for specific drug treatments in patients ahead of clinical trials; and designing molecules to ensure that drugs have the best chance of efficacy in patients. This not only speeds up the discovery and delivery of treatments, but also ensures they are more effective. Benevolent AI is based in London, with a research facility in Cambridge, and further offices in New York and Belgium. In 2018 the startup raised \$115m, bringing up its funding total to more than \$200m.

### 5) Hunters.AI – cybersecurity


**Hunters.AI** is an Israeli startup on a mission to protect the world from cybersecurity threats. Its AI-driven threat hunting technology constantly searches an organisation's systems for security breaches, and identifies attacks as soon as they are attempted. This gives companies the best chance to mitigate damage and isolate any serious risks. The technology also provides a full report of any incidents which do occur, detailing the timeline, location, risk level, target and any recommended actions. Hunters.AI was founded in 2018 and raised \$5.4m in its first seed round.

### 6) Pony.ai – autonomous vehicles

<https://disruptionhub.com/10-machine-learning-startups-transforming-industries/>

### Development of a Machine Learning Model Predicting an ICU Admission for Patients with Elective Surgery and Its Prospective Validation in Clinical Practice

Article Aug 2019

 Stefanie Jauk · Diether Kramer · Günther Stark · [...] · Johann Kainz

Frequent utilization of the Intensive Care Unit (ICU) is associated with higher costs and decreased availability for patients who urgently need it. Common risk assessment tool, like the ASA score, lack objectivity and do account only for some influencing parameters. The ai....

21 Reads

[Request full-text](#)

[Recommend](#) [Follow](#) [Share](#)

### Predicting Chemical Reaction Barriers with a Machine Learning Model

Article Mar 2019

Aayush R. Singh · Brian A. Rohr · Joseph A. Gauthier · Jens K. Nørskov

In the past few decades, tremendous advances have been made in the understanding of catalysis at solid surfaces. Despite this, most discoveries of materials for improved catalytic performance are made by a slow trial and error process in an experimental laboratory....

57 Reads · 2 Citations

[Request full-text](#)

[Recommend](#) [Follow](#) [Share](#)

### Part of the project: Optimal Machine Learning model for software defect prediction

Research from: [Tripti Lamba's Lab](#)

### Optimal Machine learning Model for Software Defect Prediction

Article [Full-text available](#) Feb 2019

 Tripti Lamba ·  Dr. kavita · A.K. Mishra

Machine Learning is a division of Artificial Intelligence which builds a system that learns from the data. Machine learning has the capability of taking the raw data from the repository which can do the computation and can predict....

1 Recommendation · 101 Reads

[Download](#)

[Recommend](#) [Follow](#) [Share](#)

### MACHINE LEARNING MODELS FOR PREDICTING FINANCIAL DISTRESS

<https://www.researchgate.net>



Welcome to H2O 3

Quick Start Videos

Cloud Integration

Downloading &amp; Installing H2O

Starting H2O

Getting Data into Your H2O Cluster

Data Manipulation

Algorithms

Cross-Validation

Variable Importance

Grid (Hyperparameter) Search

Checkpointing Models

Performance and Prediction

**AutoML: Automatic Machine Learning**

AutoML Interface

AutoML Output

Saving and Loading a Model

Productionizing H2O

Using Flow - H2O's Web UI

## AutoML: Automatic Machine Learning

In recent years, the demand for machine learning experts has outpaced the supply, despite the surge of people entering the field. To address this gap, there have been big strides in the development of user-friendly machine learning software that can be used by non-experts. The first steps toward simplifying machine learning involved developing simple, unified interfaces to a variety of machine learning algorithms (e.g. H2O).

Although H2O has made it easy for non-experts to experiment with machine learning, there is still a fair bit of knowledge and background in data science that is required to produce high-performing machine learning models. Deep Neural Networks in particular are notoriously difficult for a non-expert to tune properly. In order for machine learning software to truly be accessible to non-experts, we have designed an easy-to-use interface which automates the process of training a large selection of candidate models. H2O's AutoML can also be a helpful tool for the advanced user, by providing a simple wrapper function that performs a large number of modeling-related tasks that would typically require many lines of code, and by freeing up their time to focus on other aspects of the data science pipeline tasks such as data-preprocessing, feature engineering and model deployment.

H2O's AutoML can be used for automating the machine learning workflow, which includes automatic training and tuning of many models within a user-specified time-limit. [Stacked Ensembles](#) – one based on all previously trained models, another one on the best model of each family – will be automatically trained on collections of individual models to produce highly predictive ensemble models which, in most cases, will be the top performing models in the AutoML Leaderboard.

```
java -cp autoweka.jar weka.classifiers.meta.AutoWEKAClassifier
-timeLimit 5 -t iris.arff -no-cv
```

Figure 2: Command-line call for running Auto-WEKA with a time limit of 5 minutes on training dataset `iris.arff`. Auto-WEKA performs cross-validation internally, so we disable WEKA's cross-validation (`-no-cv`). Running with `-h` lists the available options.

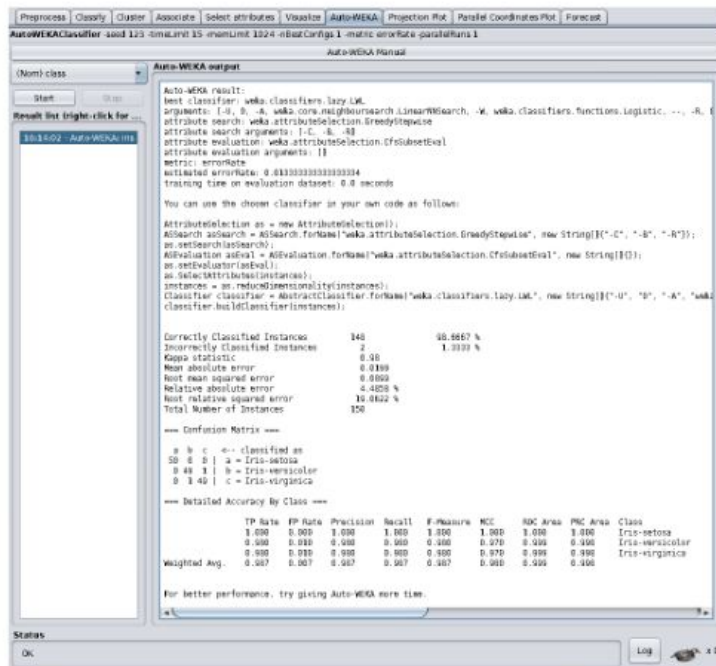


Figure 3: Example Auto-WEKA run on the iris dataset. The resulting best classifier along with its parameter settings is printed first, followed by its performance. While Auto-WEKA runs, it logs to the status bar how many configurations it has evaluated so far.

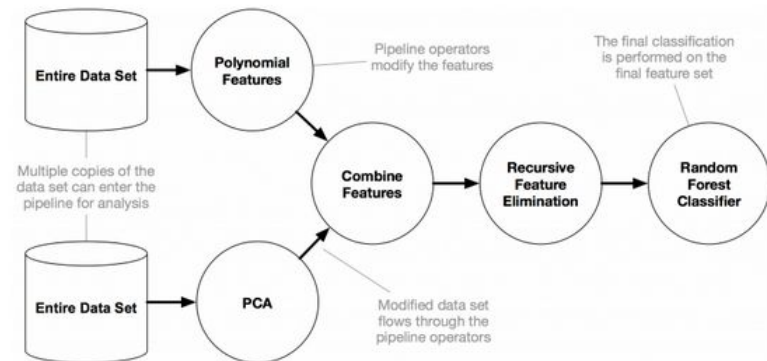
# TPOT

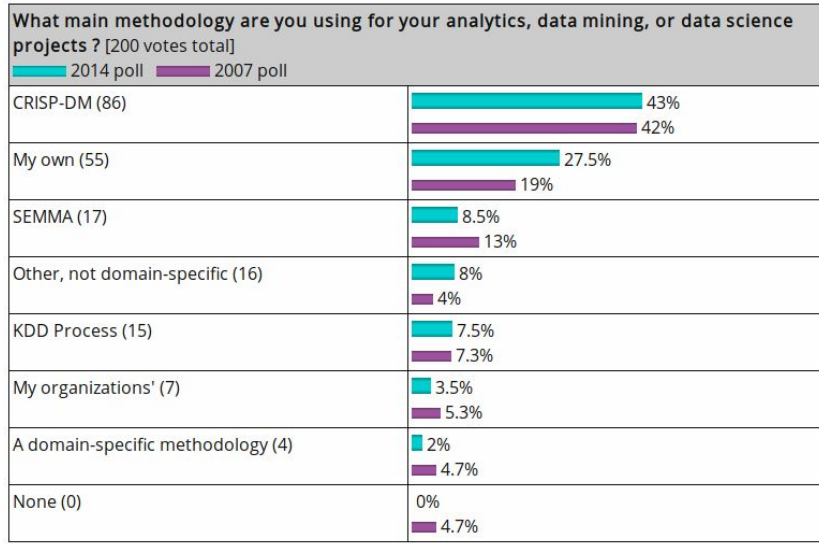
The Tree-Based Pipeline Optimization Tool (TPOT) was one of the very first AutoML methods and open-source software packages developed for the data science community. TPOT was developed by [Dr. Randal Olson](#) while a postdoctoral student with [Dr. Jason H. Moore](#) at the [Computational Genetics Laboratory](#) of the University of Pennsylvania and is still being extended and supported by this team.

The goal of TPOT is to automate the building of ML pipelines by combining a flexible [expression tree](#) representation of pipelines with stochastic search algorithms such as [genetic programming](#). TPOT makes use of the Python-based [scikit-learn](#) library as its ML menu.

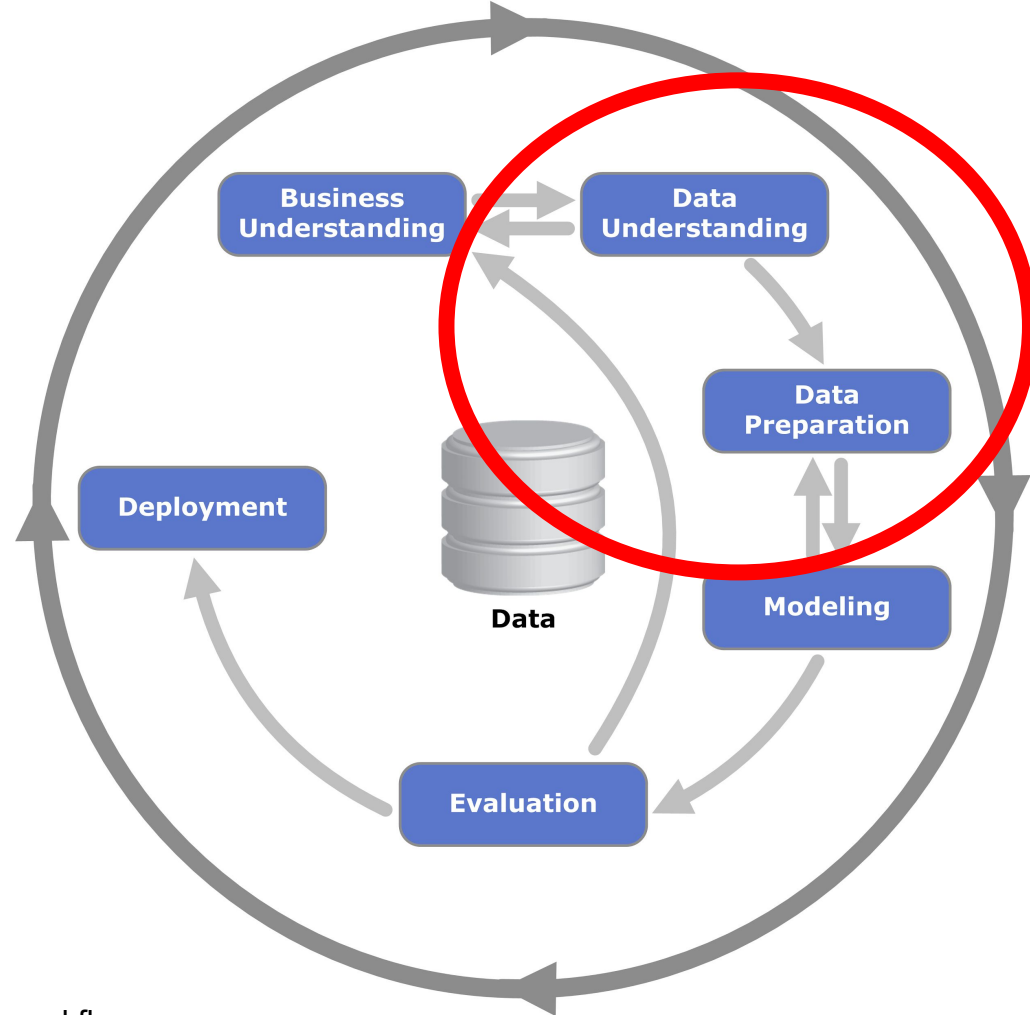
Several peer-reviewed papers have been published on TPOT. Our [first paper](#) in 2016 won a best paper award at the EvoStar computer science conference. Our [second paper](#) in 2016 won a best paper award at the GECCO computer science conference. We showed in a [2017 paper](#) presented at the GECCO conference how TPOT could be adapted to the analysis of big data from genetic studies of common human diseases. This paper was nominated for a best paper award. Here is our [latest paper](#) on some new operators to facilitate scaling TPOT to big data. Please contact us for reprints of these papers and others. These can also be found on [arXiv](#).

The TPOT software is open-source, programmed in Python, and available on [GitHub](#).





<https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>



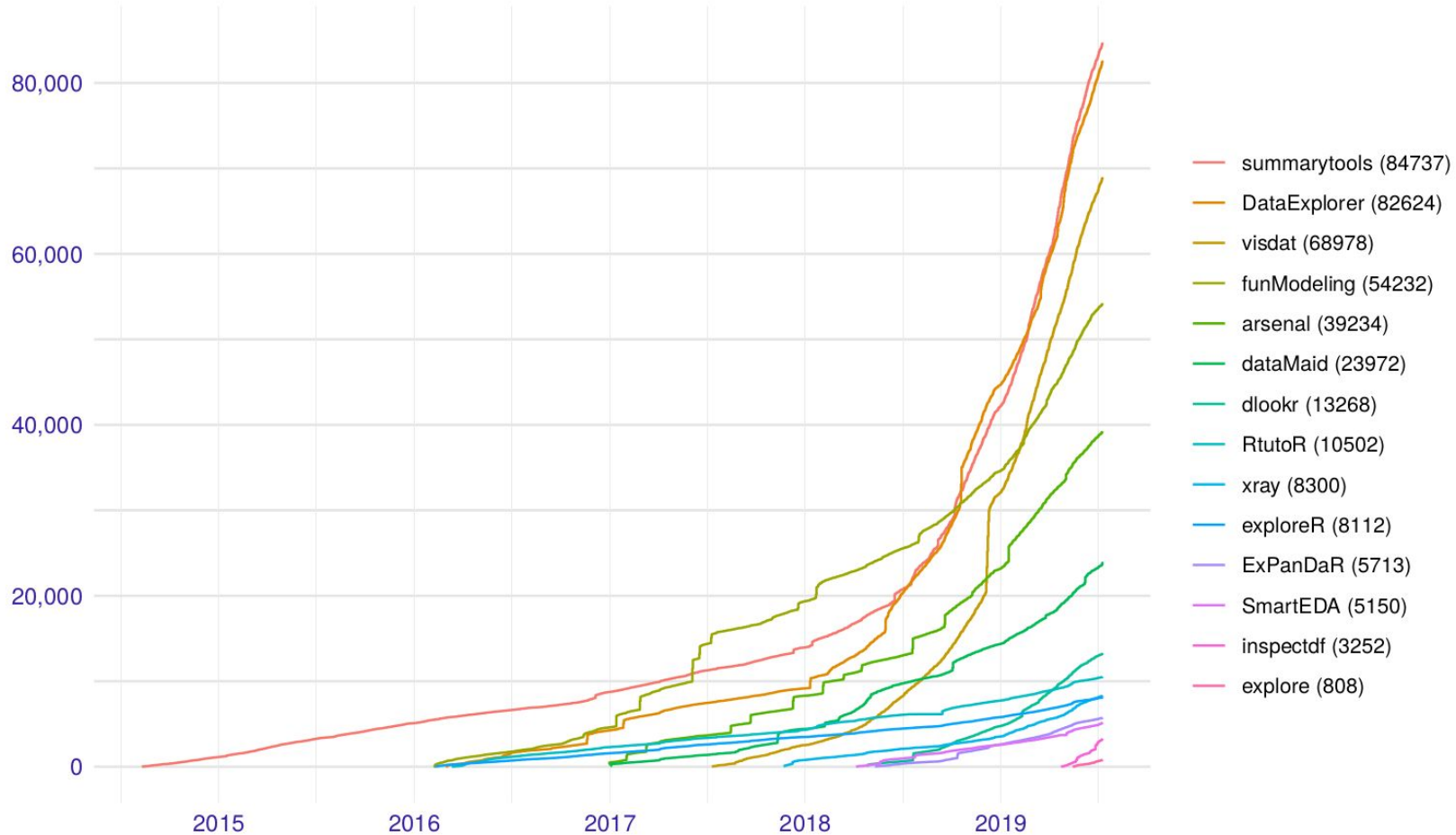
CRISP-DM workflow

<https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome>



## Total number of downloads

Based on CRAN statistics

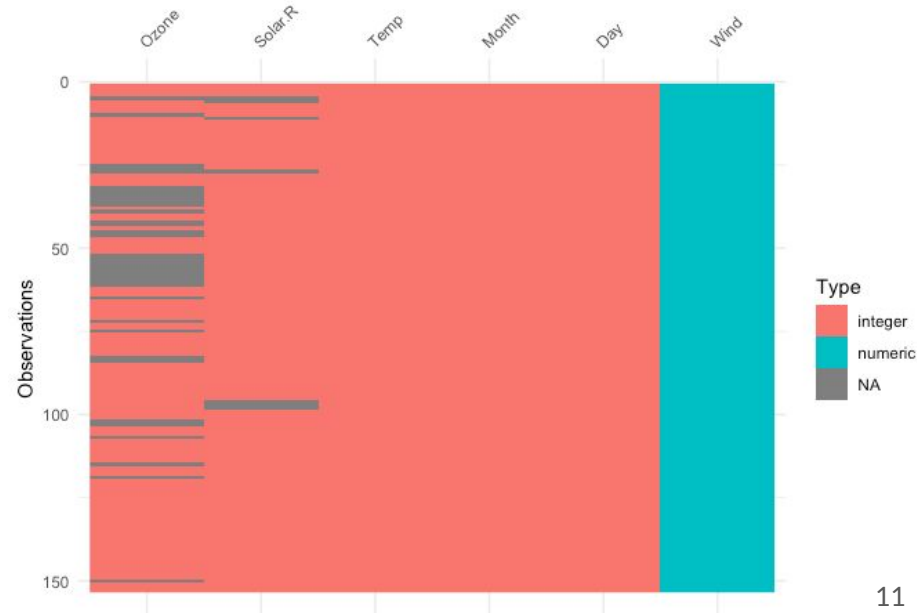


---

**What can you expect from  
autoEDA packages?**

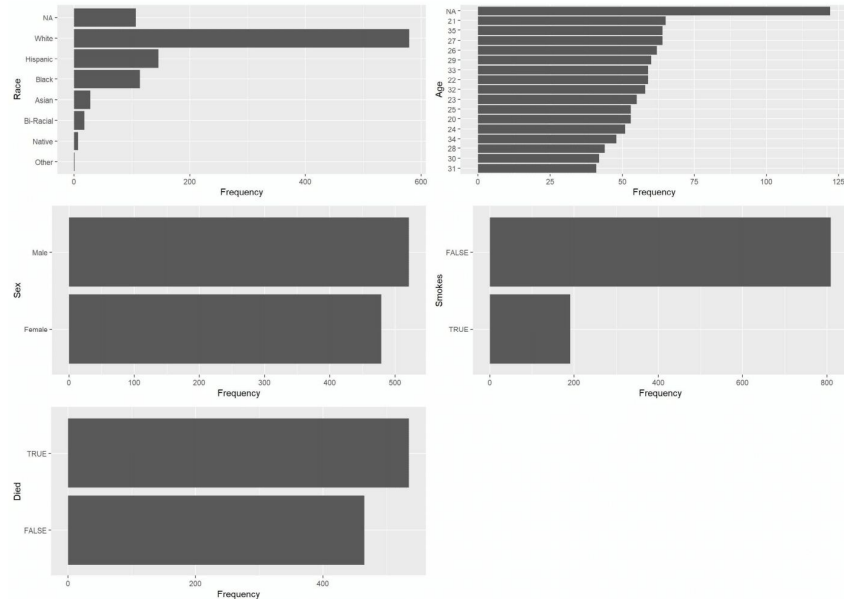
# Whole data summaries

- Missing value count
- Quantiles
- Scale & location
- Variable types
- Dataset size
- Datasets comparisons



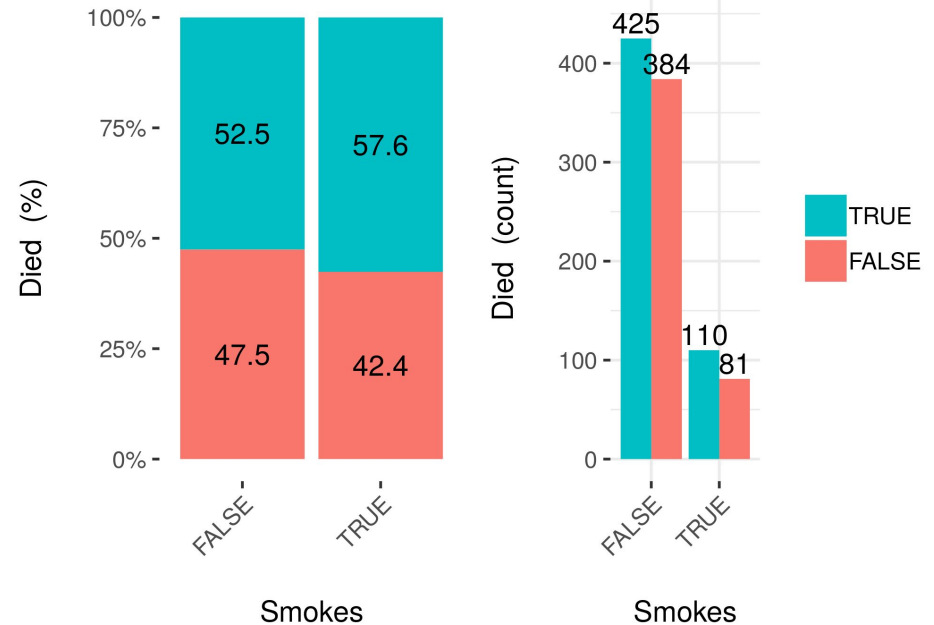
# Univariate distributions

- Descriptive statistics
- Histograms
- Bar plots
- QQ plots



# Bivariate relationships

- Scatter plots
- Grouped barplots
- Boxplots
- Contingency tables
- Correlation matrix







# More: transformations & multivariate analysis

## Transforming variables:

- skewness
- outliers
- arbitrary transformations

## Multivariate tools:

- regression models
- PCA
- visualization (Parallel Coordinate Plots)



## Reporting & interactivity

Typical approach:

All functions are run automatically  
and outputs are saved to pdf

Interactive approach:

The package assists in exploration -  
user chooses variables to explore  
interactively

---

# Top packages to look into

## Part 1

### Data report overview

The dataset examined has the following dimensions:

Feature	Result
Number of observations	1000
Number of variables	9

#### Checks performed

The following variable checks were performed, depending on the data type of each variable:

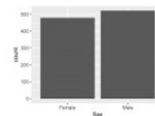
	character	factor	labelled	haven labelled	numeric	integer	logical	Date
Identify miscoded missing values	×	×	×	×	×	×		×
Identify prefixed and suffixed whitespace	×	×	×	×				
Identify levels with < 6 obs.	×	×	×	×				
Identify case issues	×	×	×	×				
Identify misclassified numeric or integer variables	×	×	×	×				
Identify outliers						×	×	×

Please note that all numerical values in the following have been rounded to 2 decimals.

1

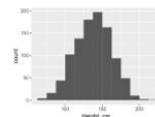
#### Sex

Feature	Result
Variable type	factor
Number of missing obs.	0 (0 %)
Number of unique values	2
Mode	"Male"
Reference category	Male



#### Height\_cm\_\_

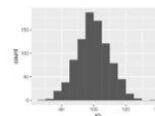
Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	365
Median	175.3
1st and 3rd quartiles	168.2; 182.03
Min. and max.	146.3; 207.2



- Note that the following possible outlier values were detected: '201.1', '207.2'.

#### IQ

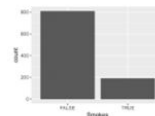
Feature	Result
Variable type	numeric
Number of missing obs.	102 (10.2 %)
Number of unique values	57
Median	100
1st and 3rd quartiles	93; 107
Min. and max.	68; 137



- Note that the following possible outlier values were detected: '68', '129', '137'.

#### Smokes

Feature	Result
Variable type	logical
Number of missing obs.	0 (0 %)
Number of unique values	2
Mode	"FALSE"



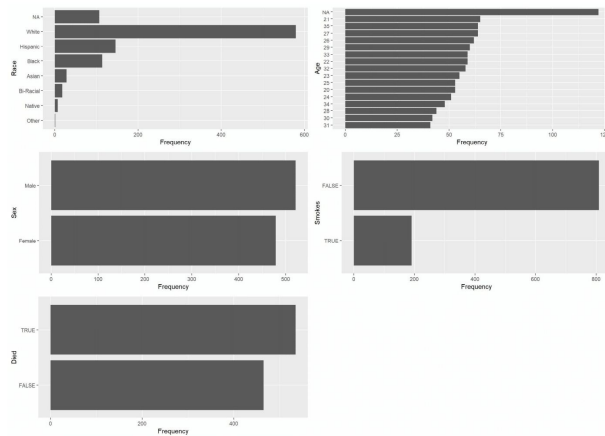
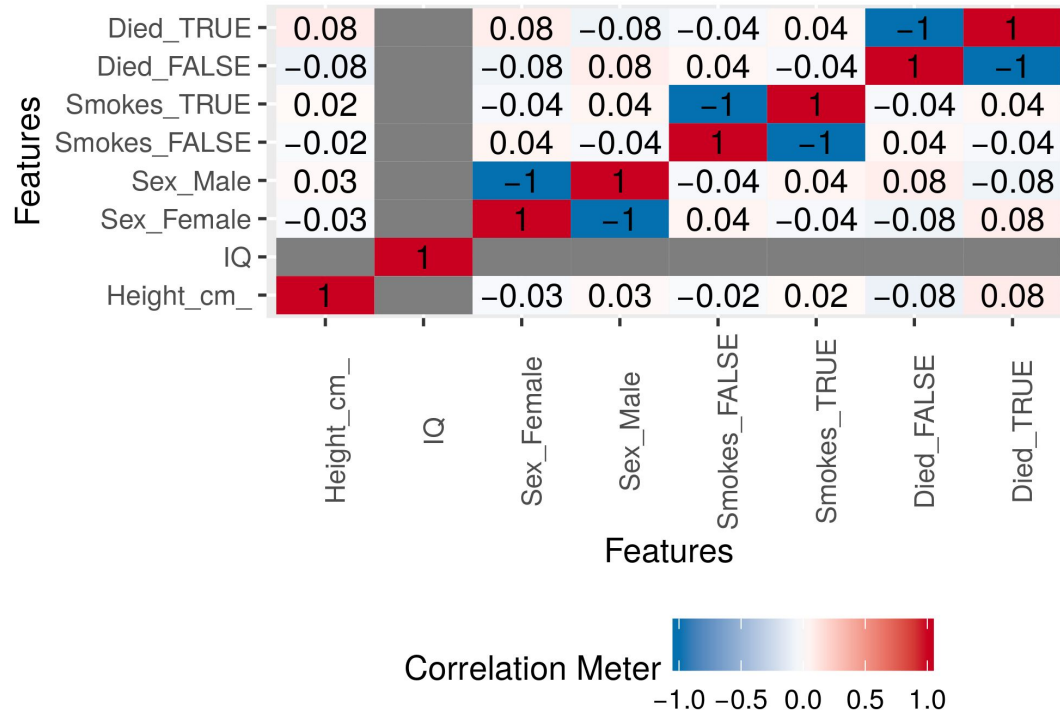
# dataMaid

Example report:

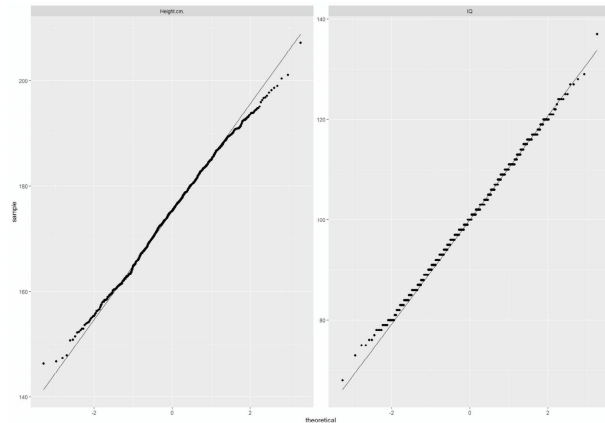
<https://bit.ly/2mMAIQJ>

# DataExplorer

Example report: <https://bit.ly/2mJLiT4>



QQ Plot



QQ Plot (by Died)



# Exploratory Data Analysis Report

2019-03-15

- Exploratory Data analysis (EDA)
  - 1. Overview of the data
  - 2. Summary of numerical variables
  - 3. Distributions of Numerical variables
    - Quantile-quantile plot for Numerical variables - Univariate
    - Density plots for Numerical variables - Univariate
    - Box plots for all numeric features vs categorical dependent variable - Bivariate comparison only with categories
  - 4. Summary of categorical variables
  - 5. Distributions of categorical variables

## Exploratory Data analysis (EDA)

Analyzing the data sets to summarize their main characteristics of variables, often with visual graphs, without using a statistical model.

### 1. Overview of the data

Understanding the dimensions of the dataset, variable names, overall missing summary and data types of each variables

```
# Overview of the data
ExpData(data=data,type=1)
# Structure of the data
ExpData(data=data,type=2)
```

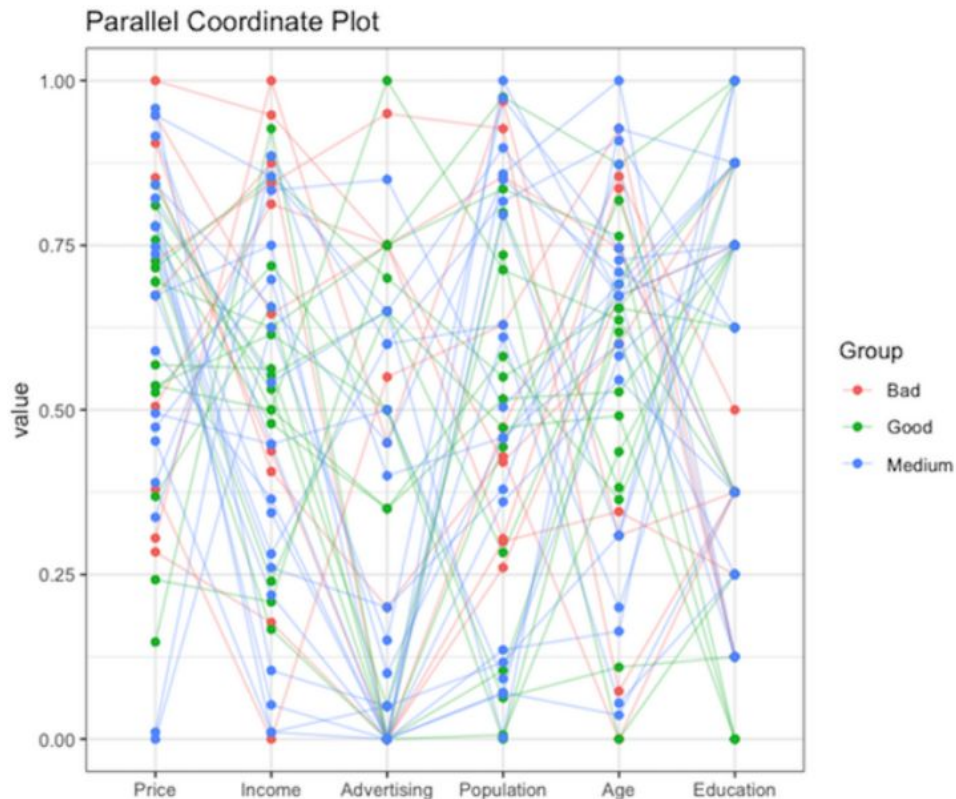
#### Overview of the data

Descriptions	Obs
<fctr>	<fctr>
Sample size (Nrow)	1000
No. of Variables (Ncol)	9
No. of Numeric Variables	2
No. of Factor Variables	3
No. of Text Variables	2
No. of Logical Variables	2
No. of Date Variables	0
No. of Zero variance Variables (Uniform)	0
% of Variables having complete cases	55.56% (5)
% of Variables having <50% missing cases	44.44% (4)
1-10 of 12 rows	Previous 1 2 Next

#### Structure of the data

## SmartEDA

Example report: <https://bit.ly/2mxvQJY>



IQ

normality test : Shapiro-Wilk normality test  
 statistic : 0.99821, p-value : 0.47445

type	skewness	kurtosis
original	0.0753	2.9651
log transformation	-0.2225	3.0516
sqrt transformation	-0.0725	2.9698

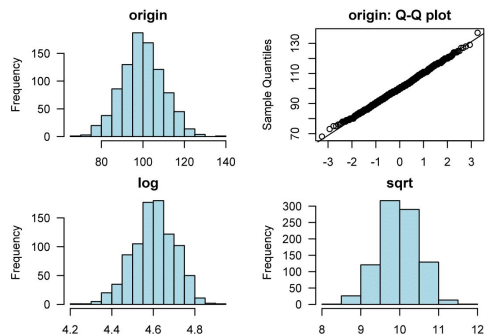


Figure 2.2: IQ

# dlookr

Example report: <https://bit.ly/2mKcFfZ>

## Chapter 1

## Introduction

The EDA Report provides exploratory data analysis information on objects that inherit `data.frame` and `data.frame`.

### 1.1 Information of Dataset

The dataset that generated the EDA Report is an `'data.frame'` object. It consists of 1,000 observations and 9 variables.

### 1.2 Information of Variables

Table 1.1: Information of Variables

variables	types	missing_count	missing_percent	unique_count	unique_rate
ID	character	0	0.0	1000	1.000
Race	factor	107	10.7	8	0.008
Age	character	122	12.2	17	0.017
Sex	factor	0	0.0	2	0.002
Height(cm)	numeric	0	0.0	365	0.365
IQ	numeric	102	10.2	58	0.058
Smokes	logical	0	0.0	2	0.002
Income	factor	100	10.0	901	0.901
Died	logical	0	0.0	2	0.002

The target variable of the data is 'Died', and the data type of the variable is logical.

### 1.3 About EDA Report

EDA reports provide information and visualization results that support the EDA process. In particular, it provides a variety of information to understand the relationship between the target variable and the rest of the variables of interest.

---

# Honorable mentions

### Group factor

None

Select a factor for subsetting specific analyses to.

### Outlier treatment

- ☒ No treatment  
☐ Winsorization 1%/99%  
☐ Winsorization 5%/95%  
☐ Truncation 1%/99%  
☐ Truncation 5%/95%

### By group factor

None

Indicate whether you want no outlier treatment or whether you want outliers to be winsorized to the given percentile or truncated if they exceed the given percentile. Give a by group if you want outlier treatment to be done independently by group.

### Bar Chart

#### Select factor to display

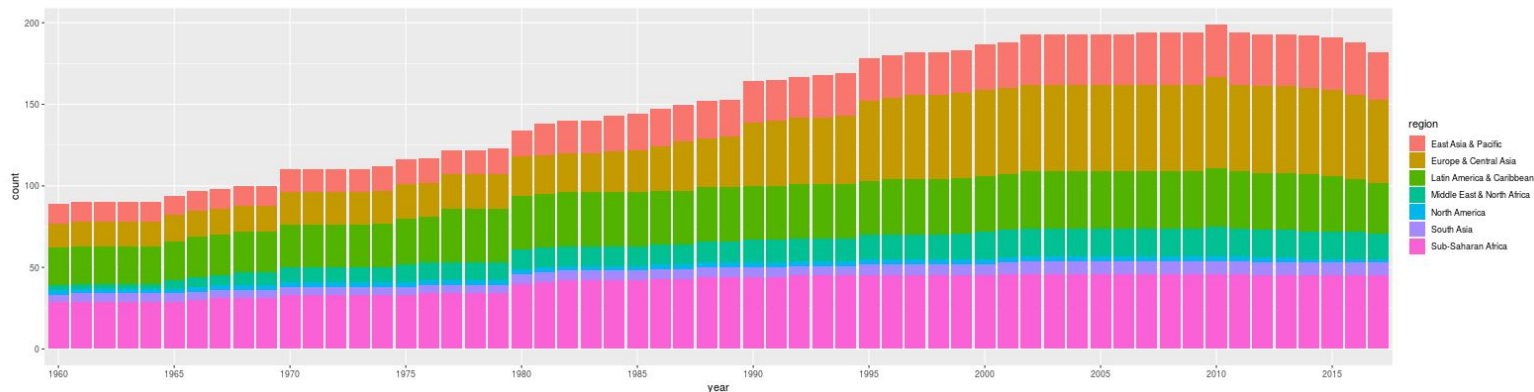
year

#### Select additional factor to display

region

☐ Relative display

Check if you want to see the additional factor relative to the first factor.



# ExPanDaR

## explore

### target

is\_versicolor

### variable

Sepal.Length

☒ auto scale

☐ split by target

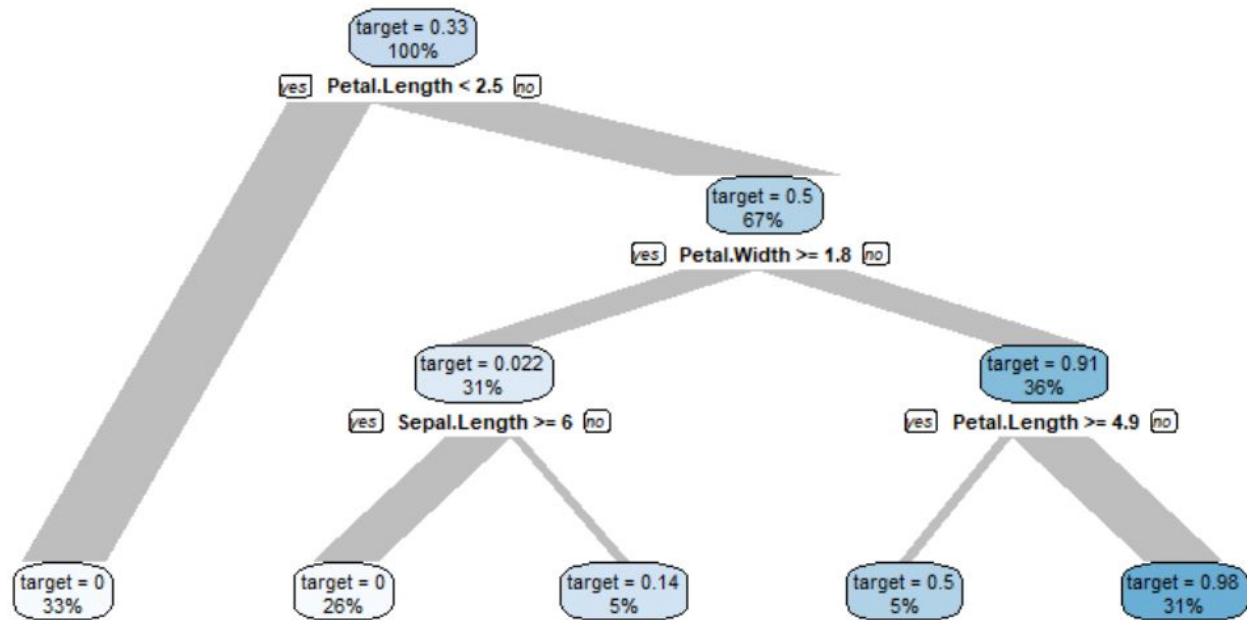
report all

variable

explain

overview

data



# explore



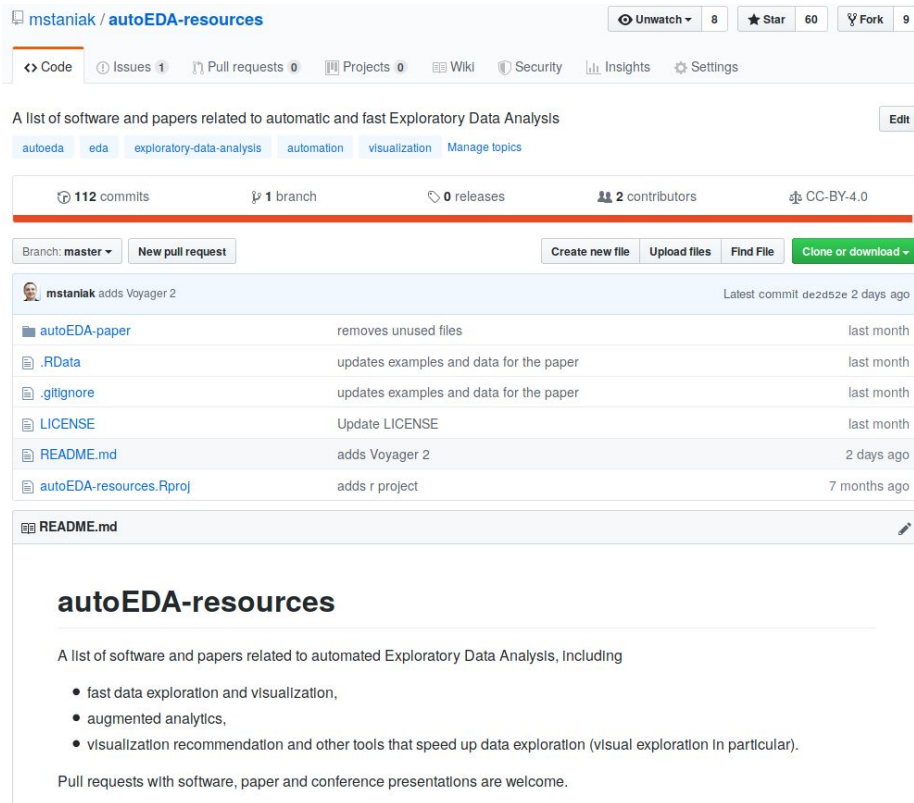
# More

- <https://journal.r-project.org/archive/2019/RJ-2019-033/>
- <https://github.com/mstaniak/autoEDA-resources>  
(includes packages from other languages)
- <http://blog.mstaniak.pl>
- 

Contact:

- [github.com/mstaniak](https://github.com/mstaniak)
- [mtst@mstaniak.pl](mailto:mtst@mstaniak.pl)

# Thank you for you attention



The screenshot shows the GitHub repository page for `mstaniak / autoEDA-resources`. At the top, there are navigation links: `<> Code`, `Issues 1`, `Pull requests 0`, `Projects 0`, `Wiki`, `Security`, `Insights`, and `Settings`. Below these, a description reads: "A list of software and papers related to automatic and fast Exploratory Data Analysis". There are tabs for `autoeda`, `eda`, `exploratory-data-analysis`, `automation`, `visualization`, and `Manage topics`. The repository statistics show `112 commits`, `1 branch`, `0 releases`, `2 contributors`, and `CC-BY-4.0` license. A red bar indicates the current branch is `master`. Below the statistics, there is a table of recent commits:

Commit	Message	Time
<code>mstaniak</code> adds Voyager 2	adds Voyager 2	2 days ago
<code>autoEDA-paper</code>	removes unused files	last month
<code>.RData</code>	updates examples and data for the paper	last month
<code>.gitignore</code>	updates examples and data for the paper	last month
<code>LICENSE</code>	Update LICENSE	last month
<code>README.md</code>	adds Voyager 2	2 days ago
<code>autoEDA-resources.Rproj</code>	adds r project	7 months ago

Below the commit table, there is a section for `README.md` with the title `autoEDA-resources`. The description reads: "A list of software and papers related to automated Exploratory Data Analysis, including". The list of items includes:

- fast data exploration and visualization,
- augmented analytics,
- visualization recommendation and other tools that speed up data exploration (visual exploration in particular).

Below the list, it says: "Pull requests with software, paper and conference presentations are welcome."