Automatic Exploratory Data Analysis Tools in R - an Overview

Mateusz Staniak

11 - 03 - 2019

Contents

Introduction

With autoML tools like h2o Driverless AI or autoKeras, building predictive models is becoming easier and faster. But the first step in every Data Science project is understanding the particular dataset and the patterns that can be found inside it. It usually reffered to as Exploratory Data Analysis. This report is a summary of the existing R tools for automatic (or fast) EDA.

Automatic EDA Tasks

EDA tools have multiple possible goals and most of the tools only try to address only some of them.

- Whole dataset summaries: provide information about sample size, number of variable and their types, possibly relationships between several datasets and meta-data such amount of disk space used.
- Data validity: perform checks related to missing data and atypical values.

- Univariate summaries: depends on variable type. For numerical variables, usually typical descriptive statistics such as centrality and dispersion measures, for categorical data, unique levels and associated counts.
- Bivariate summaries: present simple relationships, either between one variable of interest and all other variables (contigency tables, scatter plots, survival curves, plots of distribution (boxplots, histograms, bar plots) by values of a variable), all pairs of variables (correlation matrices and plots) or chosen pairs of variables.
- Feature engineering: make variables more suitable for modeling. This includes PCA and other dimension reduction techniques, merging levels of categorical variables, transforming numerical variables (for example Box-Cox transformation).
- Data visualization: finding visual insight in the data. This task is particularly challenging when the dataset is high dimensional.
- Quick reporting: create a report based on the above points.

Overview of R Tools for autoEDA

In this section, eight R packages are shortly summarized. One of them is only available on GitHub (`autoEDA`), the rest is on CRAN. The list is not exhaustive, but these are the most matured general-purpose packages. Other libraries exist, for example `RBioPlot` which was written specifically for molecular biology data.

dataMaid

The `dataMaid` package has two central functions: the `check` function, which performs checks of data consistency and validity, and `makeDataReport`, which automatically creates a report (in PDF, DOCX or HTML format) . The goal is to detect unsual (outliers, anomalies, incorretly encoded) and missing values. The report contains whole dataset summary (variables and their types, number of missing values and if a problem was detected) and summaries (plot of the distribution and some descriptive statistic) for each variable separately. It is possible to manually define new checks and summaries.

xray

The `xray` package offers three functions for the analysis of data prior to statistical modeling:

- detecting anomalies (missing data, zero values, blank strings and infinite numbers),
- drawing (through histograms and barplots) and printing (through quantile tables) univariate distributions of each variable,
- drawing plots of variables over time (for a specified time variable). Plots can be saved to png images.

visdat

The package `visdat` is maintained by rOpenSci. It consist of 6 functions that help visualize:

- variables types and missing data,
- types of each value in each column,
- clusters of missing values,
- differences between two datasets,
- where given conditions are satisfied in the data,
- correlation matrix for the numerical variables. Each of these functions returns a single `ggplot2` plot that shows a rectangular representation of the dataset.

dlookr

The `dlookr` package provides tools for 3 types of analysis: data diagnosis (correctness, missing values and outliers detection), exploratory data analysis, feature engineering (imputation, dychotomization, transformation of continuous features). It can also automatically generate a pdf report for all these analyses.

For data diagnosis, types of variables are reported along with counts of missing values and and unique values. Variables with low proportion of unique values are described separately. All the typical descriptive statistics are provided for each variable. Outliers are detected and distributions of variables before and after outlier removal are plotted.

In EDA report, descriptive statistics are presented along with normality tests and histogram of variables and their transformation that reduce skewness (logarithm and root square). Correlation plots are shown for numerical variables. If target variable is specified, plots that show relationship between the target and each predictor are also included.

Transformation reports compare descriptive statistics and plots for each variable before and after imputation, skewness-removing transformation and binning.

DataExplorer

`DataExplorer` is a new package that helps automatize EDA and simple feature engineering. It can also generate a report with data summaries. It has functions for:

- whole dataset summary (dimensions, types of variables, missing values etc),
- missing values visualization (percentage of missigness in each column),
- plotting distributions of variables (numerical and categorical variables separately),
- QQ Plots,
- plotting correlation matrices,
- visualizing PCA results,
- plotting relationships between target variable and predictors (scatterplots and boxplots),
- replacing missing values by a constant,
- grouping sparse categories,
- creating dummy variables and dropping features. The automatic report can be customized. By default, it consists of all the above points except feature engineering.

funModeling

The package `funModeling` is a reach set of tools for EDA connected to the book Data Science Live Book -Open Source- (2017). These tools include

- dataset summary,
- plots and descriptive statistics for categorical and numerical variables,
- correlation analysis (classical and based on information theory),
- plots of distribution of target variables vs predictors (bar plots, boxplots, histograms),
- quantitative analysis for binary target variables,
- different methods of discretization,
- variable scaling,
- outlier treatment,
- gain and lift curves.

autoEDA

`autoEDA` package is a GitHub-based tool for univariate and bivariate visualizations. It can also generate a pdf report with the plots of distributions of predictors grouped by outcome variable or distribution of outcome by predictors.

arsenal

The `arsenal` package is a set of 5 tools for data exploration:

- descriptive statistics by levels of a target variable (like Table 1), also for paired observation (for example longitudinal data),

- comparing data frames,
- frequency tables for multiple categorical variables,
- fitting and summarizing simple statistical models (linear regression, Cox model etc). Results of each function can be saved to a short report.

Feature Comparison

In this section, I will compare how different packages address autoEDA tasks. For a quick overview of package features, see the table below.

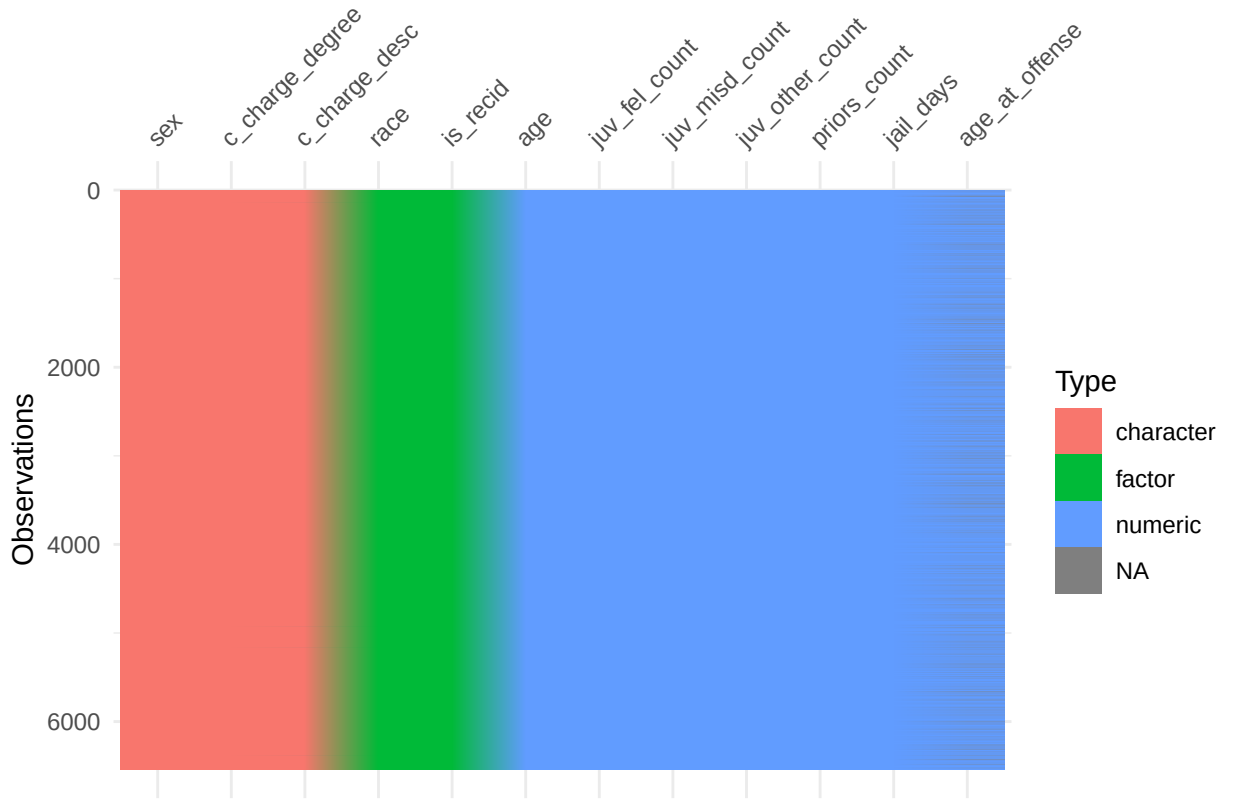
```
comparison_table %>%
  rename(task = X1, feature = X2) %>%
  mutate_if(is.numeric, function(x) ifelse(x == 1, "x", "")) %>%
knitr::kable() %>%
  kable_styling(full_width = F)
```

task	feature	DataExplorer	dataMaid	funModeling	visdat	arsenal	xray	au
Whole dataset	Variable types	x	x	x	x			x
Whole dataset	Dataset size	x	x	x	x			x
Whole dataset	Other info	x			x			
Whole dataset	Compare two datasets				x	x		
Data validity	Missing values	x	x	x	x		x	x
Data validity	Redundant columns		x	x	x			x
Data validity	Outliers		x	x				x
Data validity	Atypical values		x				x	
Data validity	Level encoding		x					
Univariate analysis	Descriptive statistics		x	x			x	x
Univariate analysis	Histograms	x	x	x				x
Univariate analysis	Boxplots	x						
Univariate analysis	Bar plots	x	x	x				x
Univariate analysis	QQ plots	x						
Bivariate analysis	Descriptive statistics							
Bivariate analysis	Correlation matrix	x						
Bivariate analysis	1 vs each correlation			x				x
Bivariate analysis	Time-dependency					x	x	
Bivariate analysis	Bar plots by target	x		x				x
Bivariate analysis	Histograms by target			x				x
Bivariate analysis	Scatter plots	x						
Bivariate analysis	Contigency tables					x		
Bivariate analysis	Other stats. for factors			x				
Multivariate analysis	PCA	x						
Multivariate analysis	Stat. Models					x		
Feature engineering	Imputation							
Feature engineering	Scaling			x				
Feature engineering	Skewness reduction							
Feature engineering	Outlier treatment			x				
Feature engineering	Binning (cont. vars)	x		x				
Feature engineering	Merging factor levels	x						
Quick reporting	PDF/html reports	x	x					x
Quick reporting	Saving plots/outputs			x		x	x	

Whole dataset summaries

`visdat` package offers most original summaries of full dataset.

```
visdat::vis_dat(recid)
```



The drawback of this approach is that it is not well suited for high dimensional data. But for smaller number of variables, it gives a good overview of the dataset.

Most packages that provide a whole dataset summary take a similar approach and present names and types of variables, number of missing values and sometimes unique values or other statistics. This is true for `autoEDA` (`dataOverview` function), `dataMaid` (`makeDataReport` result), `funModeling` (`df_status` function) and `DataExplorer` (`introduce` function), which provides the information separately on two plots - one for dataset structure, one for missing data. The `dlookr` package provides summaries for numerical variables and categorical variables are presented only in the report, separately (`describe` function).

Data validity

Some package offer automated checks for the data. This include at least outlier detection. The `dataMaid` package's main purpose is to find inconsistencies and errors in the data. It finds possible outliers, missing values, low-frequency and possibly mis-encoded levels of factors. All these information can be summarized in a quality report. The `dlookr` package offers similar functionality. There are two main differences:

- the report does not contain possible mis-encoded factors,
- outlier analysis is supplemented with plots showing variable distribution before and after removing the outliers.

The analysis is rather simple, for example in zero-inflated variables non-zero values are treated as outliers (`dlookr`).

Univariate summaries

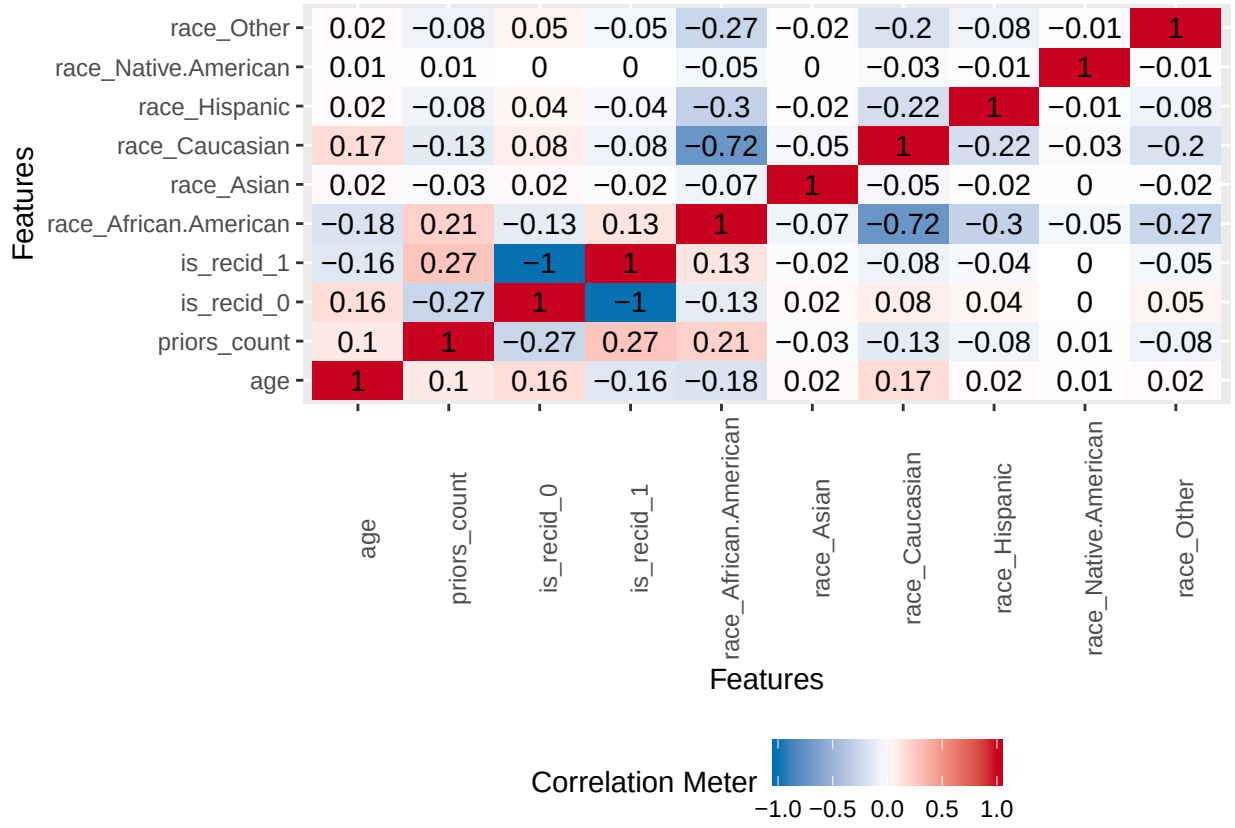
All the tools that support univariate analysis take a similar approach. For categorical variables, counts are reported and bar plots are presented, while histogram or boxplots and typical descriptive statistics (including quantiles, sometimes skewness) are used for continous variables.

In `dataMaid` and `dlookr` packagas, these plots are presented variable-by-variable in the report. In other packages (`dataExplorer`, `funModeling`) groups of plots are shown together (wall of histograms, wall of barplots). Notably, the `dlookr` reports skewness of variables and in case a skewed variables is found,it shows the distribution after some candidate transformations to reduce the skewness. This package also reports normality.

Bivariate summaries

The `funModeling` package only support calculating correlations between variables and a specified target. `DataExplorer`, `vis_dat` and `DataExplorer` packages can plot correlation matrices. They differ in categorical variables treatment. Some packages require only numerical features (`vis_dat`). Interestingly, in `DataExplorer`, low-cardinality categorical features are converted to 0-1 variables and plotted alongside numerical variables.

```
DataExplorer::plot_correlation(recid_small)
```



The report from `autoEDA` package consists of a limited number of bar plots / boxplots with target variable as one of the dimensions. The `arsenal` package only presents variable summaries by levels of a chosen categorical variable. Similarly, in `DataExplorer`, `dlookr` and `funModeling`, scatter plots and boxplots with a specified target variable on one of the axis can be plotted. Additionally, `funModeling` and `dlookr` draw histograms/densities of continuous features by target. The `funModeling` package also has unique options of drawing barplots of discretized variable by target and quantitative analysis for binary outcome based on representativeness and accuracy.

```
knitr::kable(funModeling::categ_analysis(recid, "race", "is_recid")) %>%
  kable_styling(full_width = F)
```

race	mean_target	sum_target	perc_target	q_rows	perc_rows
African-American	0.399	1301	0.586	3257	0.498
Native American	0.333	5	0.002	15	0.002
Caucasian	0.288	650	0.293	2256	0.345
Hispanic	0.279	149	0.067	535	0.082
Other	0.247	109	0.049	441	0.067
Asian	0.182	6	0.003	33	0.005

Feature engineering

The `dataMaid` package assumes that every decision regarding the data should be made by the analyst and does not provide any tools for data manipulation after diagnosis. Most of the packages only provide exploration tools. Exceptions are `dlookr`, `funModeling` and `DataExplorer`. `DataExplorer` tools are limited to imputation by a constant, merging levels of factors and creating dummy variables.

The `dlookr` package can create a report that presents different possible transformations of features. Missing values can be imputed by mean/median/mode and distributions of variables before and after the procedure compared. The same is done for imputation of outliers. Logarithmic and root square transforms are proposed for skewed variables. Different methods of binning continuous variables are also presented, including Weight of the Evidence.

The `funModeling` package can perform discretization of variable using equal frequency criterion or gain ratio maximization. It can also scale variables to the interval [0, 1]. Outliers can be treated using Tukey or Hampel method.

Data visualization

Data visualization is mostly limited to univariate and bivariate plots described above. Some original visualizations for whole dataset are provided by `visdat` package. Currently, none of the packages perform visualization recommendation. Available plots are limited to standard uni- and bivariate graphics.

Quick reporting

`DataExplorer`, `dlookr` and `dataMaid` packages are capable of generating good quality reports. They consists of all (or most) possible outputs of the package which are organized either by variable (`dataMaid`, `dlookr`) or by type of variable (`DataExplorer`). `autoEDA` package generates a minimal report with bivariate plots. Example reports from these packages can be found in the `usecase/` directory. Packages `arsenal`, `funModeling` and `xray` have an option of saving outputs to files.

Summary

Undoubtely, the presented have many advantages. Still, there are many open problems related to automated data exploration that could vastly improve them.

Strengths

- The packages `dlookr`, `dataMaid` and `DataExplorer` are capable of creating good quality reports.
- `DataExplorer` has very good visualizations for PCA.
- `DataExplorer` handles categorical variables on correlation plots by creating dummy features, which is a unique idea compared to other packages.

- The `visdat` package, while probably not the best choice for high dimensional data, features interesting take on initial whole dataset exploration.
- The `dlookr` package is capable of selecting skewed variable and proposing transformations.
- `dataMaid` is a good tool for finding problems in the data.
- For datasets with a moderate number of features, `DataExplorer`, `funModeling` and `dlookr` give a reasonable insight into variables stributions and simple relationships.

Weaknesses

- All the presented tools are likely to fail in typical situations with imperfect data. In particular, they are usually not robust to issues like zero-variance/constant variables (`dataExplorer` can't generate a report in this case). Error messages in some cases not uninformative.
- In some situations, they lack flexibility. For example, in `DataExplorer` arguments can be passed to `cor` function, but not to `corrplot` function.
- In case of *walls* of histograms or barplots, no selection is being done and no specific order is chosen to promote most interesting distributions. Moreover, for high-dimensional data or high-cardinality factors the plots are unreadable or impractical. This is especially true for `DataExplorer` and `funModeling` functions (e.g. `cross_plot`), even though `DataExplorer` removes too large factors from the panels.
- Plots are limited to bivariate relationships, when exploring higher dimensional dependencies would be interesting (for example through colors and sizes).
- Support for time-varying variables and non-classical (not IID) problems such as survival analysis is limited or non-existent.
- Automated reports could be enriched by textual annotations and descriptions, either built from a simple template or from a generative model.
- Only one of the packages addresses the issue of skewed variables. Proposing transformations of continuous features (other than binning) could be also useful.
- Exploration based on simple statistical models (for example scatter plot smoothing) is not an option in any of the packages.
- None of the packages addresses issues such a multicolinearity.
- Missing data imputation more advanced than imputing a constant is delegated to other variables.
- Some of the above issues (lack of recommendations, not performing imputation or other transformation) result in the packages being not suitable for iterative work.