Unified Approach to the Exploratory Data Analysis Phase

Vincent Wolowski

April 2021

Automatic EDA

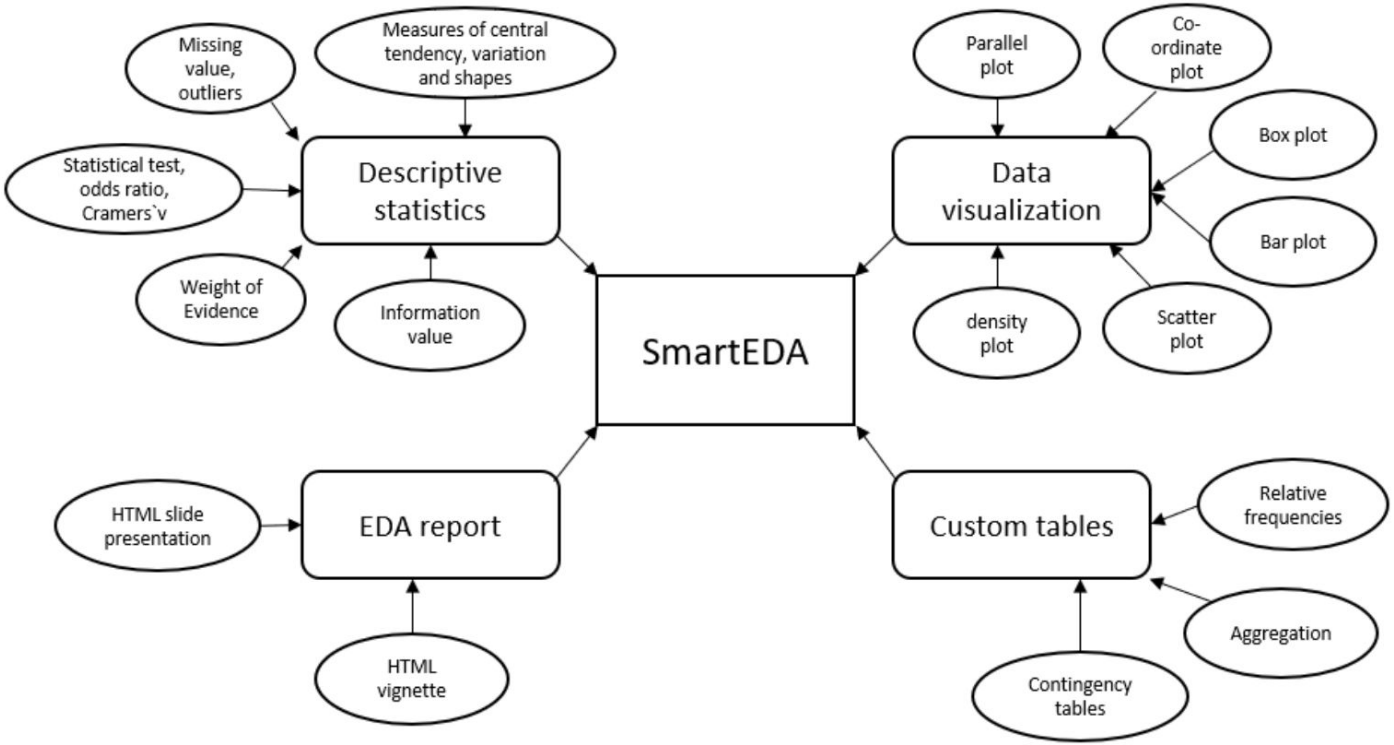
R packages for Automatic EDA

	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.

Source: *M.Staniak & P.Biecek: The Landscape of R Packages for Automated Exploratory Data Analysis, 2019*

Functionalities of R package SmartEDA



Source: *https://github.com/daya6489/SmartEDA*

Comparison of SmartEDA with other packages

	CRAN packages						
Exploratory analysis features	SmartEDA	dlookr	DataExplorer	Hmisc	exploreR	RtutoR	summarytools
Describe basic information for input data	✓	✓	✓	✓			✓
Function to provide summary statistics for all numerical variable (automatically scans through each variable and select only numeric/integer variables)	✓	✓		✓			✓
Function to provide plots for all numerical variable (automatically scans through each variable and select only numeric/integer variables)	✓		✓		✓		
Function to provide summary statistics and plots for all character or categorical (automatically scans through each variable and select only character/categorical variables)	✓	✓	✓	✓			✓
Function to provide plots for all character or categorical (automatically scans through each variable and select only character/categorical variables)	✓		✓				
Customized summary statistics - extension of data.table package	✓						
Normality / Co-ordinate plots	✓	✓	✓				
Feature binarization / Binning		✓	✓				
Standardize /missing imputation / diagnose outliers		✓	✓		✓		
HTML report using rmarkdown / Shiny	✓	✓	✓			✓	

Source: *https://github.com/daya6489/SmartEDA*

R packages for Automatic EDA

Package	Method	Description
DataExplorer	create_report()	Creates a detailed summary HTML report *(comment: show example)*
SmartEDA	ExpReport()	Creates a detailed summary HTML report (*comment: show example)*
AEDA	fastReport()	Creates a detailed Rmd report (*comment: show example)*
dataMaid	makeDataReport()	Creates a detailed report in Word, PDF or HTML (*comment: show example)*
dlookr	diagnose_web_report() diagnose_paged_report()	Reports the information for diagnosing the quality of the data
ExPanDaR	ExPanD()	Creates a shiny based app for interactive exploratory data analysis (*comment: show example)*

Descriptive Summary

Descriptive Summary for Continuous Variables

Metrics	Description
obs_count	The counts of the observations
obs_pct	The percentage of the observations
missing_count	The counts of missing observations
missing_pct	The percentage of the missing observations
mean/median/mode	Main measures of central tendency
min/max	Minimum and maximum for each observation
range	Maximum - minimum
quartiles	First, second, third, fourth quartile
IQR	Interquartile range
MAD	Median absolute deviation as a robust measure of the variability (amount of variation)
variance	Variance as a measure of the variability
std. dev.	Standard deviation as a measure of the variability
CV	Coefficient of variation (relative standard deviation) as a standardized measure of dispersion of a probability distribution or frequency distribution (as percentage).
Skewness	As a measures the deviation of a variable's given distribution from the normal distribution, of the asymmetry of a distribution. As a guideline, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.
Kurtosis	As a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. High kurtosis implies the variable’s distribution has a heavy tail, or outliers, low kurtosis implies the variable’s distribution has a light tail, or lack of outliers.

Descriptive Summary for Categorical Variables

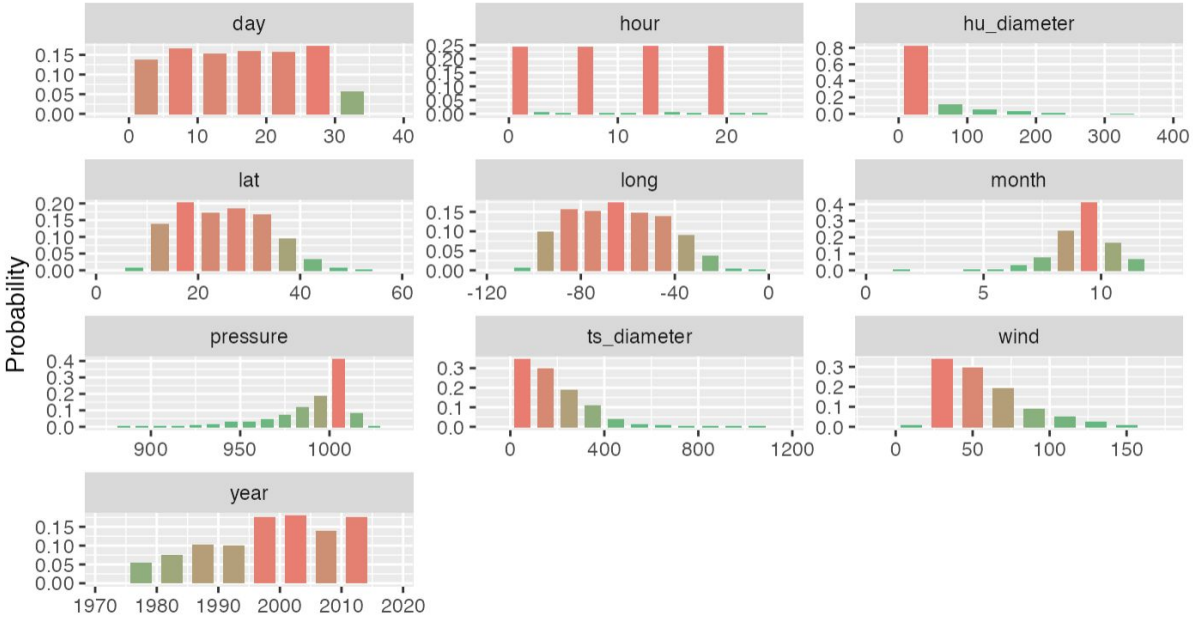
Metrics	Description
is_factor	Boolean if it is an R factor
is_ordered_factor	Boolean if it is an ordered R factor
is_date	Boolean if it is a R date type.
unique_values_count	The number of unique values
unique_values_pct	The number of unique values as percentage
missing_count	The counts of missing observations
missing_pct	The percentage of the missing observations
mode	The most observed value
per category: count	For each value category the number of values
per category: pct	For each value category the number of values expressed as percentage
Shannon entropy	A measure of the “information content” of a variable, the amount of information required to describe a variable.
UAC	Unalikeability coefficient, measure of the variability for categorical variables, represents the proportion of possible comparisons (pairings) which are unalike. Implemented in ragree::unalike().

Descriptive Summary for Dataframes

Metrics	Description
rows	Number of rows
columns	Number of columns
discrete_columns	Number of categorical variables
continuous_columns	Number of continuous variables
all_missing_values	Total number of missing values
complete_rows	Number of rows without any missing values
total_observation	Number of columns without any missing values
memory_usage	Memory size of dataframe

Exploring Numeric Features

inspectdf::inspect_num(data, breaks = 10) %>% inspectdf::show_plot()
A histogram is generated for each numeric feature



Exploring Numeric Features

skimr::skim(data)

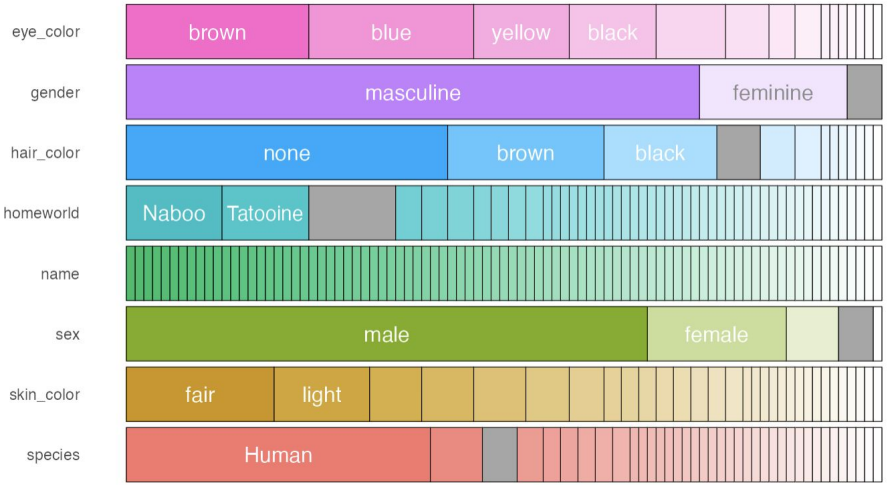
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
displ	0	1	3.47	1.29	1.6	2.4	3.3	4.6	7	
year	0	1	2003.50	4.51	1999.0	1999.0	2003.5	2008.0	2008	
cyl	0	1	5.89	1.61	4.0	4.0	6.0	8.0	8	
cty	0	1	16.86	4.26	9.0	14.0	17.0	19.0	35	
hwy	0	1	23.44	5.95	12.0	18.0	24.0	27.0	44	

modelsummary::datasummary_skim(data)

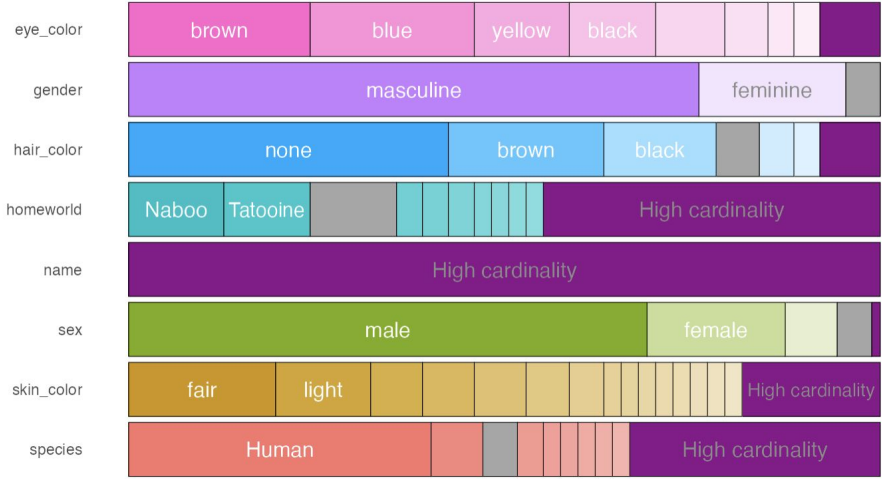
	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
bill_length_mm	165	1	43.9	5.5	32.1	44.5	59.6	
bill_depth_mm	81	1	17.2	2.0	13.1	17.3	21.5	
flipper_length_mm	56	1	200.9	14.1	172.0	197.0	231.0	
body_mass_g	95	1	4201.8	802.0	2700.0	4050.0	6300.0	

Exploring Categorical Features

```
inspectdf::inspect_cat(data) %>%
inspectdf::show_plot()
```



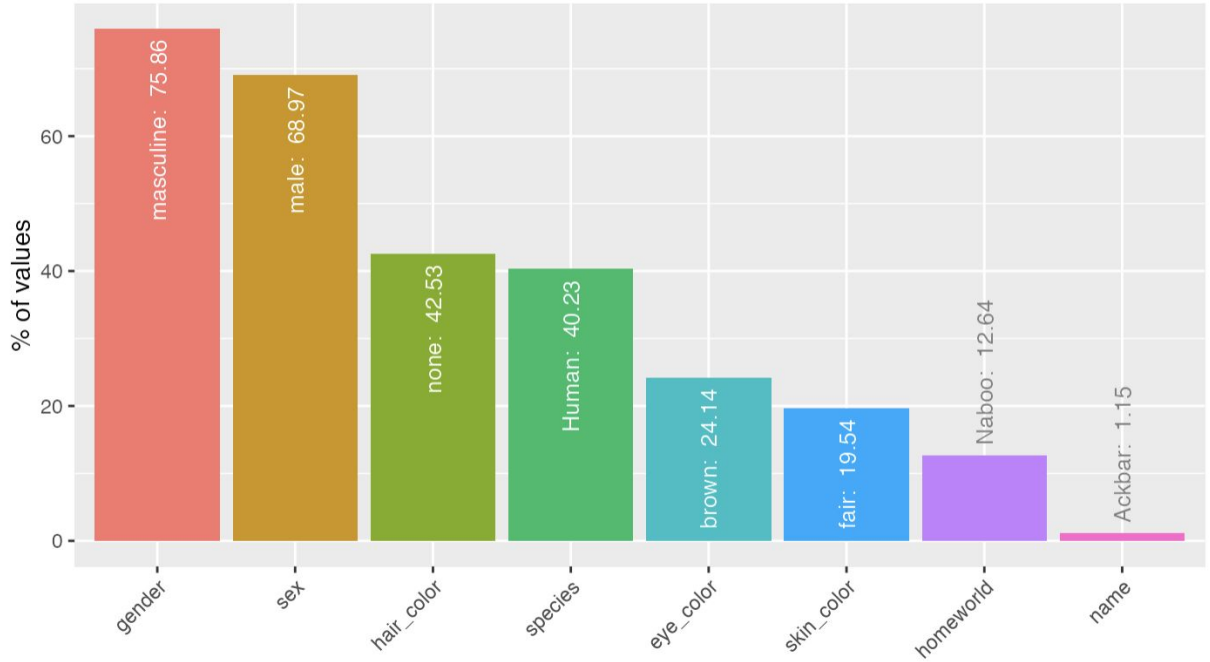

```
inspectdf::inspect_cat(data) %>%
inspectdf::show_plot(high_cardinality = 1)
```



Frequency of categorical levels.
Gray segments are missing values.
In purple bundled together are categories shown that occur only a small number of times.

Exploring Categorical Features

inspectdf::inspect_imb(data) %>% inspectdf::show_plot()
Understanding if categorical columns are dominated by a single level,
the most frequently occurring categorical level in each column.



Descriptive Summary: Univariate Visualizations

Histogram	Shows the frequency distribution for a continuous variable
Bar Chart	Presents a categorical variable with rectangular bars with heights or lengths proportional to the values that they represent
Box Plot	Standardized way of displaying the dataset based on a five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles
Density Plot	Represents the distribution of a continuous variable by using a kernel density estimate to show the probability density function
QQ Plot	A Normal Quantile-Quantile Plot compares two probability distributions by plotting their quantiles against each other. The observed quantiles (depicted as dots) with the quantiles to be expected if the data were normally distributed (depicted as a solid line). These plots can reveal outliers, differences in location and scale, and other differences between the distributions.

Descriptive Summary: Bivariate Visualizations

Scatterplot	Used to observe relationships between continuous variables
Density Plot	Using hexagonal grids (package hexbin, ggplot2 with geom_hex())

Descriptive Summary: Multivariate Visualizations

Heatmap	two-dimensional representation of data in which values are represented by colors
Igloo-Plot	Visualization of multidimensional datasets

R packages for Summary Statistics

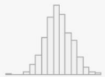
Package	Method	Description
dlookr	describe()	Computes descriptive statistics of numeric variables for exploratory data analysis
dlookr	diagnose() diagnose_report()	Computes descriptive statistics for continuous and categorical variables
summarytools	descr()	Provides univariate statistics for numerical data. Calculates mean, sd, min, Q1*, median, Q3*, max, MAD, IQR*, CV, skewness*, SE.skewness*, and kurtosis* on numerical vectors. (*) Not available when using sampling weights.
summarytools	dfSummary()	Summary of a data frame consisting of: variable names and types, labels if any, factor levels, frequencies and/or numerical summary statistics, barplots/histograms, and valid/missing observation counts and proportions.
skimr	skim()	Provides a broad overview of a data frame. It handles data of all types, dispatching a different set of summary functions based on the types of columns in the data frame.
psych	describe()	Provides summary statistics, the ones most useful for scale construction and item analysis in classic psychometrics
DataExplorer	introduce()	Describe basic information for input data

R packages for numeric and categorical column summaries

Package	Method	Description
inspectdf	inspect_num()	A histogram is generated for each numeric feature
inspectdf	inspect_cat()	Frequency of categorical levels
inspectdf	inspect_imb()	The most frequently occurring categorical level in each column is shown

Data Frame Summary by summarytools::dfSummary()

tobacco
Dimensions: 1000 x 9
Duplicates: 2

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Missing
1	gender [factor]	1. F 2. M	489 (50.0%) 489 (50.0%)		22 (2.2%)
2	age [numeric]	Mean (sd) : 49.6 (18.3) min < med < max: 18 < 50 < 80 IQR (CV) : 32 (0.4)	63 distinct values		25 (2.5%)
3	age.gr [factor]	1. 18-34 2. 35-50 3. 51-70 4. 71 +	258 (26.5%) 241 (24.7%) 317 (32.5%) 159 (16.3%)		25 (2.5%)
4	BMI [numeric]	Mean (sd) : 25.7 (4.5) min < med < max: 8.8 < 25.6 < 39.4 IQR (CV) : 5.7 (0.2)	974 distinct values		26 (2.6%)
5	smoker [factor]	1. Yes 2. No	298 (29.8%) 702 (70.2%)		0 (0%)
6	cigs.per.day [numeric]	Mean (sd) : 6.8 (11.9) min < med < max: 0 < 0 < 40 IQR (CV) : 11 (1.8)	37 distinct values		35 (3.5%)
7	diseased [factor]	1. Yes 2. No	224 (22.4%) 776 (77.6%)		0 (0%)
8	disease [character]	1. Hypertension 2. Cancer 3. Cholesterol 4. Heart 5. Pulmonary 6. Musculoskeletal 7. Diabetes 8. Hearing 9. Digestive 10. Hypotension [3 others]	36 (16.2%) 34 (15.3%) 21 (9.5%) 20 (9.0%) 20 (9.0%) 19 (8.6%) 14 (6.3%) 14 (6.3%) 12 (5.4%) 11 (5.0%) 21 (9.5%)		778 (77.8%)

Source: *https://cran.r-project.org/web/packages/summarytools/vignettes/introduction.html*

Core functions of R package summarytools

Function	Description
`freq()`	**Frequency Tables** featuring counts, proportions, cumulative statistics as well as missing data reporting
`ctable()`	**Cross-Tabulations** (joint frequencies) between pairs of discrete/categorical variables, featuring marginal sums as well as row, column or total proportions
`descr()`	**Descriptive (Univariate) Statistics** for numerical data, featuring common measures of central tendency and dispersion
`dfSummary()`	**Data Frame Summaries** featuring type-specific information for all variables: univariate statistics and/or frequency distributions, bar charts or histograms, as well as missing data counts and proportions. Very useful to quickly, detect anomalies and identify trends at a glance

Source: *https://cran.r-project.org/web/packages/summarytools/vignettes/introduction.html*

Descriptive Summary

EDAWB::get_summary_stats(data)

Quantitative variables
variable, n_count, n_pct missing_count, missing_pct, unique_count, unique_rate, var, sd, cv, mean, median, mad, min, max, range, iqr, skewness, kurtosis

Qualitative variables
variable, n_count, n_pct, missing_count, missing_pct, unique_count, unique_rate, total_num_levels, levels, counts_per_levels, ratio_per_levels, levels_rank, mode, entropy, uac, uac_perry

DataExplorer::introduce(data)

Summary of dataframe
rows, columns, discrete_columns, continuous_columns, all_missing_columns, total_missing_values, complete_rows, total_observations, memory_usage

Outlier Detection

Methods for Outlier Detection

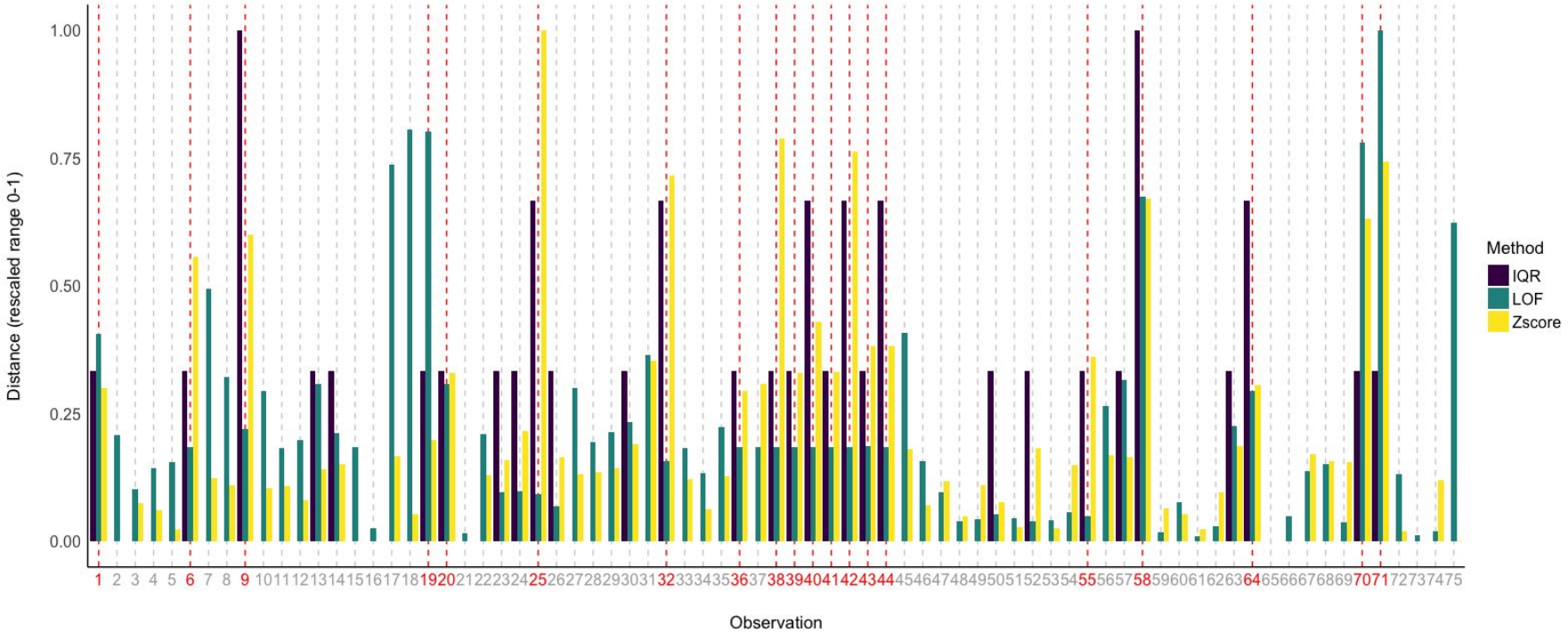
Package	Method	Description
EDAWB	find_outliers_by_iqr()	IQR based outlier detection.
EDAWB	find_outliers_by_quantile()	Quantile based outlier detection.
EDAWB	find_outliers_by_zscore()	Z-score based outlier detection.
outliers	outlier()	Find value with largest difference from the mean.
outliers	chisq.out.test()	Chi-squared test for outlier.
outliers	cochran.test()	Cochran’s C test to check if largest variance in several groups of data is "outlying" and this group should be rejected. Alternatively, if one group has very small variance, it can be tested for "inlying" variance.
car	outlierTest()	Reports the Bonferroni p-values for testing each observation in turn to be a mean-shift outlier, based Studentized residuals in linear (t-tests), generalized linear models (normal tests), and linear mixed models.

Methods for Outlier Detection

Package	Method	Description
EDAWB	find_outliers_by_hampel_filter()	Hampel filter.
outliers	grubbs.test()	Grubbs’s test.
outliers	dixon.test()	Dixon’s test.
EnvStats	rosnerTest()	Rosner’s test.
DescTools	LOF()	A function that finds the local outlier factor (Breunig et al.,2000) of the input data matrix using k neighbours.
performance	check_outliers()	Can be "all" or some of c("cook", "pareto", "zscore", "iqr", "mahalanobis", "robust", "mcd", "ics", "optics", "lof").
dlookr	diagnose_outlier()	Produces outlier information for diagnosing the quality of the numerical data.
robustbase	adjOutlyingness()	Computes Skewness-adjusted multivariate "outlyingness" of all observations. Outlyingness here is a generalization of the Donoho-Stahel outlyingness measure, where skewness is taken into account via the medcouple, i.e. a robust concept and estimator of skewness. The medcouple is defined as a scaled median difference of the left and right half of distribution, and hence not based on the third moment as the classical skewness.

Visualizing Outliers with the R package *performance*

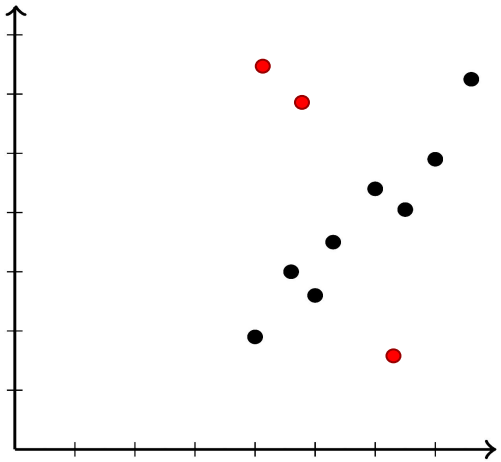
```
res <- performance::check_outliers(data, method = c("zscore", "iqr","lof"))
plot(res, type = "bars")
```



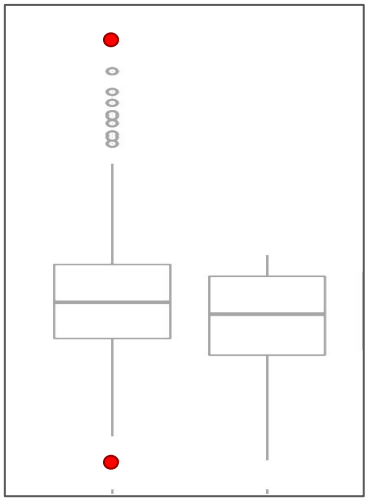
Source : https://easystats.github.io/see/articles/performance.html

Visualizing Outliers

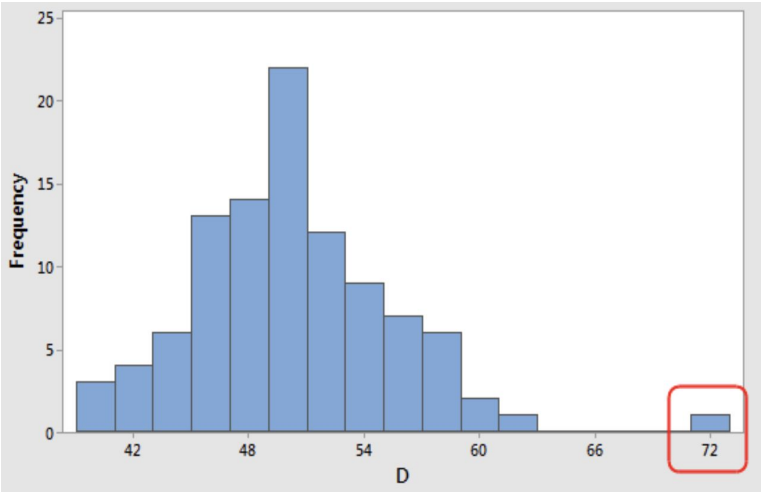
Scatterplot



Boxplot

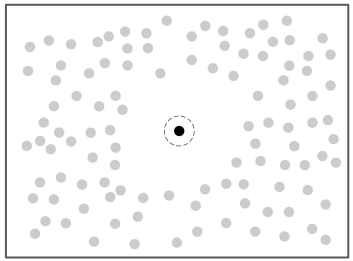


Histogram

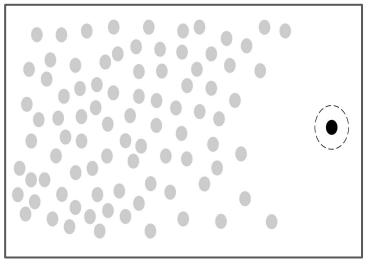


Visualizing Outliers

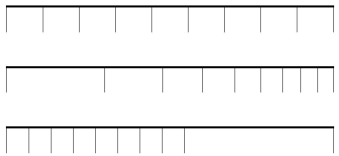
Point of Focus



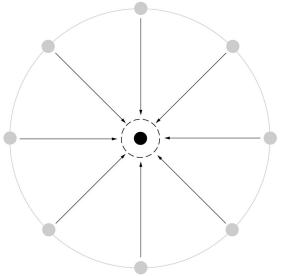
Breakout



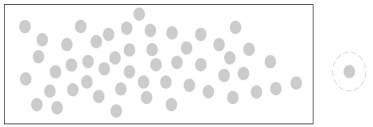
Scale Adjustment

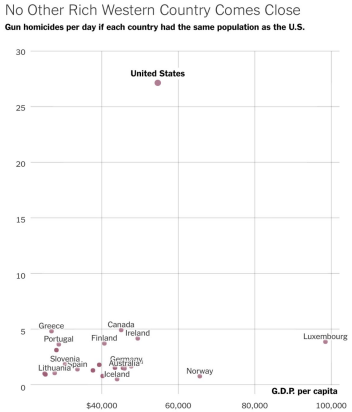


Reference Point

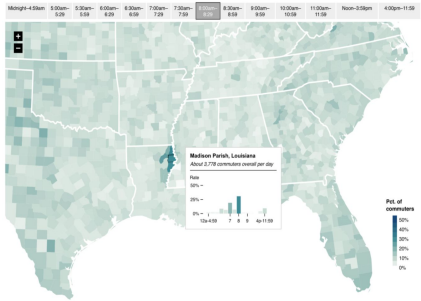


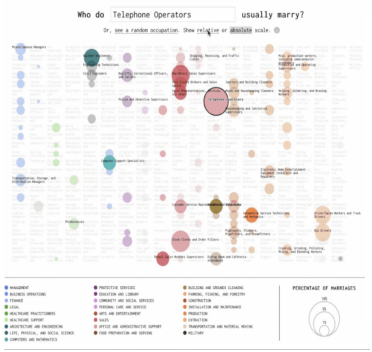
Providing Context



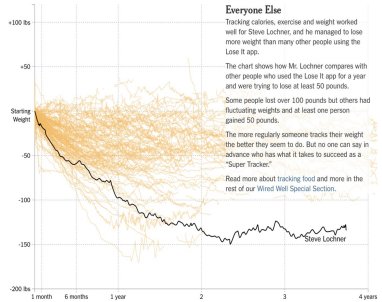


Commute Time by Country

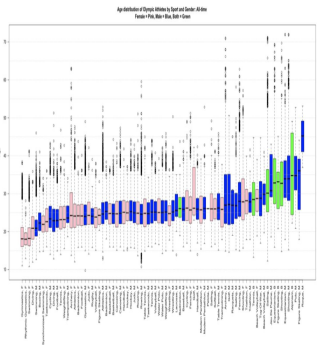




Diary of Food Tracker



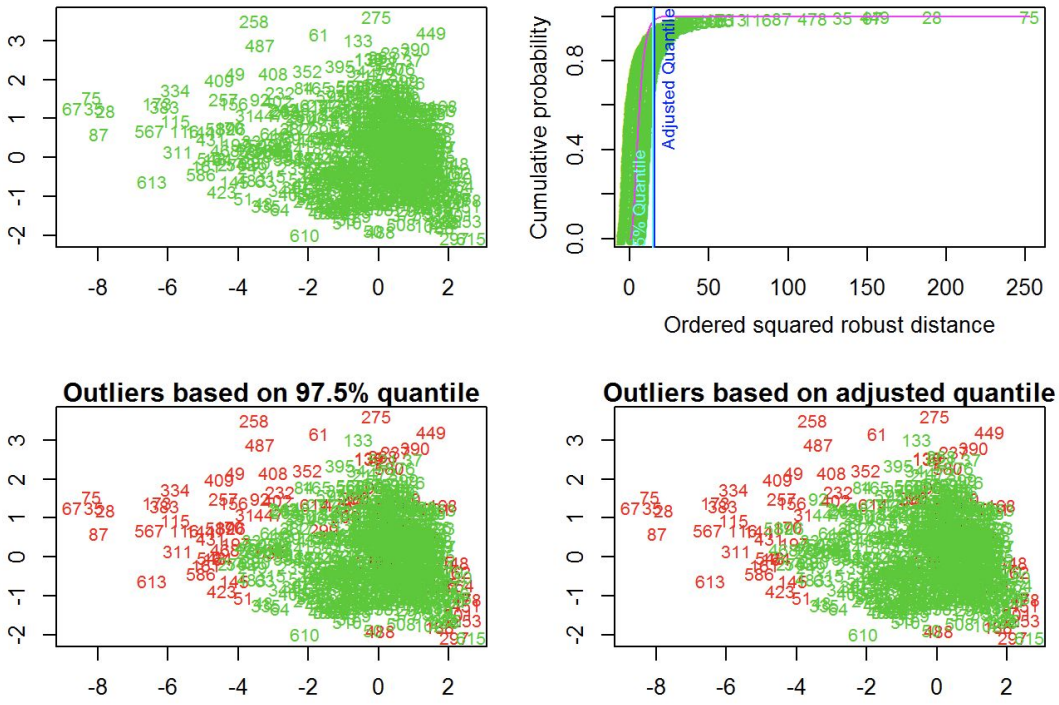
Age distribution Olympics Sports



Source: https://flowingdata.com/2018/03/07/visualizing-outliers/

Visualizing Outliers using R Package *mvoutlier*

mvoutlier::aq.plot(data)

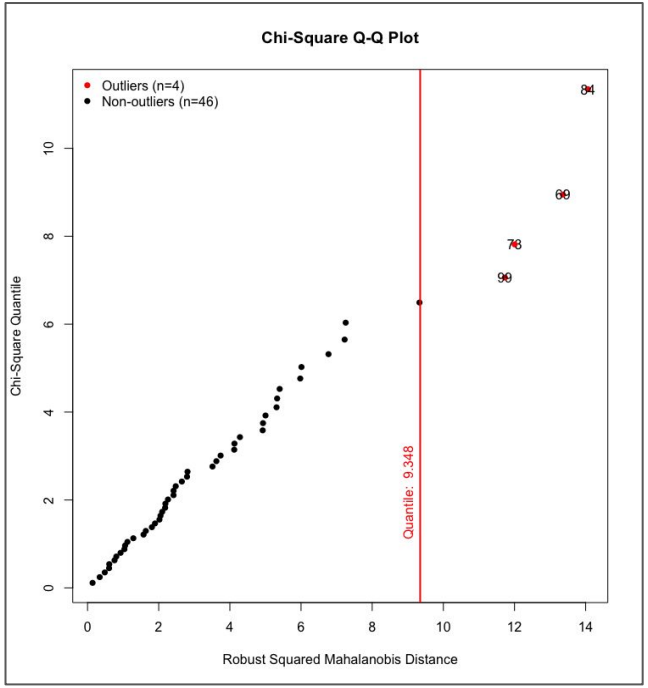


Source: https://rpubs.com/Treegonaut/301942

Outlier Detection: Visualizing Outliers using R Package *MVN*

Performs multivariate normality tests, univariate normality of marginal distributions through plots and tests, multivariate Box-Cox transformation, and multivariate outlier detection, i.e. (adjust) quantile method based on Mahalanobis distance.

```
library(MVN)
data(iris)
versicolor <- iris[51:100, 1:3]
MVN::mvn(versicolor, multivariateOutlierMethod = "quan", showOutliers = TRUE)
```




```
$multivariateNormality
           Test        HZ   p value MVN
1 Henze-Zirkler 0.7137223 0.3255775 YES

$univariateNormality
             Test     Variable Statistic   p value Normality
1 Anderson-Darling Sepal.Length    0.3608    0.4333    YES
2 Anderson-Darling Sepal.Width     0.5598    0.1406    YES
3 Anderson-Darling Petal.Length    0.5551    0.1446    YES

$Descriptives
              n  Mean   Std.Dev Median Min Max  25th 75th        Skew   Kurtosis
Sepal.Length 50 5.936 0.5161711   5.90 4.9 7.0 5.600  6.3  0.09913926 -0.6939138
Sepal.Width  50 2.770 0.3137983   2.80 2.0 3.4 2.525  3.0 -0.34136443 -0.5493203
Petal.Length 50 4.260 0.4699110   4.35 3.0 5.1 4.000  4.6 -0.57060243 -0.1902555

$multivariateOutliers
   Observation Mahalanobis Distance Outlier
84          84               14.065    TRUE
69          69               13.360    TRUE
73          73               12.005    TRUE
99          99               11.729    TRUE
```

Outlier Detection

EDAWB::get_outlier_summary(continuous_data)
Outlier tests for continuous data: IQR, z-score, quantile (0.025,0.975), Hampel Filter

EDAWB::get_outliers_by_performance_check_outliers(continuous_data)
Wrapper for *performance::check_outliers()*.
Outlier tests for continuous data:

- Z-score: standard score
- IQR: Interquartile Range
- Univariate methods is to compute for each variable some sort of "confidence" interval and consider as outliers values lying beyond the edges of that interval.
  - Equal-Tailed Interval ("eti"),
  - Highest Density Interval ("hdi"),
  - Bias Corrected and Accelerated Interval ("bci").

  The default threshold is 0.95, considering as outliers all observations that are outside the 95% CI on any of the variable.
- Mahalanobis distance for multivariate outliers detection
- Minimum Covariance Determinant (MCD)
- Invariant Coordinate Selection (ICS)
- The Ordering Points To Identify the Clustering Structure (OPTICS) algorithm (Ankerst et al., 1999)
- Local Outlier Factor

Source: https://rdrr.io/cran/performance/man/check_outliers.html

Missingness

Patterns of Missingness

Rubin (1976) classification: every data point has some likelihood of being missing. The process that governs these probabilities is called the missing data mechanism or response mechanism. The model for the process is called the missing data model or response model.

Missing Completely at Random (MCAR)
A certain value is missing has nothing to do with its hypothetical value and with the values of other variables. It occurs entirely at random.

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = 0|\psi)$$

Missing at Random (MAR)
The probability of being missing is the same only within groups defined by the *observed* data.

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi) = \Pr(R = 0|Y_{\text{obs}}, \psi)$$

Missing not at Random (MNAR)
Two possible reasons are that the missing value depends on the hypothetical value or is dependent on some other variable's value.

$$\Pr(R = 0|Y_{\text{obs}}, Y_{\text{mis}}, \psi)$$

Sources:
https://stefvanbuuren.name/fimd/sec-MCAR.html
https://en.wikipedia.org/wiki/Missing_data
https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Patterns of Missingness

Let the data be represented by the $n \times p$ matrix Y. In the presence of missing data Y is partially observed. Notation Y_j is the jth column in Y, and Y_{-j} indicates the complement of Y_j, that is, all columns in Y except Y_j. The missing data pattern of Y is the $n \times p$ binary response matrix R.

Univariate and multivariate

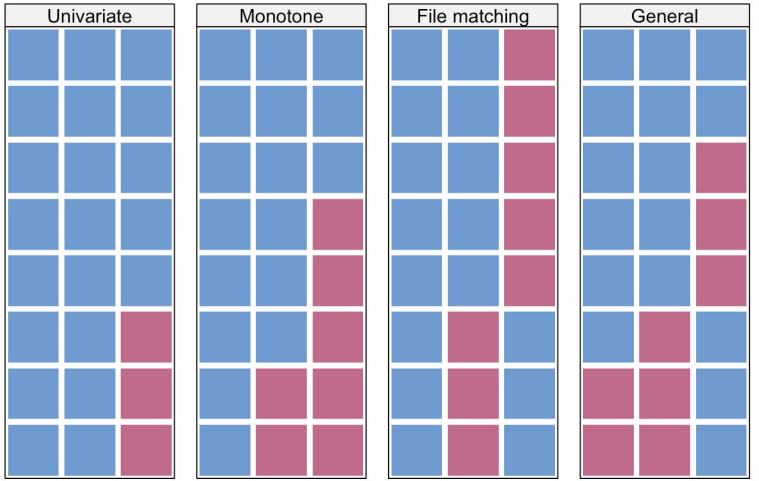
A missing data pattern is said to be univariate if there is only one variable with missing data. Otherwise it is multivariate.

Monotone and non-monotone (or general)

A missing data pattern is said to be monotone if the variables Y_j can be ordered such that if Y_j with $k > j$ are also missing. This occurs, for example, in longitudinal studies with drop-out. If the pattern is not monotone, it is called non-monotone or general.

Connected and unconnected

A missing data pattern is said to be connected if any observed data point can be reached from any other observed data point through a sequence of horizontal or vertical moves.



Examples for missing data patterns. Blue is observed, red is missing.

Source: https://stefvanbuuren.name/fimd/missing-data-pattern.html

Tabular Information on Missingness

EDAWB::get_pct_missing_column_wise(data, 2)

	var	pct_missing
1	CD8IMMPH	43
2	CD8LOC	43
3	BMISI	1
4	BWGHTSI	1
5	USUBJID	0
6	STUDYID	0
7	STUDYNAME	0
8	STUDY_PHASE	0
9	AGE	0
10	SEX	0
11	RACE	0
12	ETHNIC	0
13	COUNTRY	0
14	BECOG	0
15	SMOKHX	0
16	TOBHX	0
17	APCPI	0
18	BHGHTCM	0
19	BSDIABP	0
20	BSSYSBP	0

EDAWB::get_pct_missing_row_wise(data, 2)

	row_index	pct_missing
10	19	10
11	20	10
12	22	10
13	23	10
14	24	10
15	26	10
16	32	10
17	36	10

EDAWB::get_freq_missing_column_wise(data)

	variable	freq_missing (#rows:100)
1	CD8IMMPH	43
2	CD8LOC	43
3	BMISI	1
4	BWGHTSI	1
5	USUBJID	0
6	STUDYID	0
7	STUDYNAME	0
8	STUDY_PHASE	0
9	AGE	0
10	SEX	0
11	RACE	0
12	ETHNIC	0
13	COUNTRY	0
14	BECOG	0
15	SMOKHX	0
16	TOBHX	0
17	APCPI	0
18	BHGHTCM	0
19	BSDIABP	0
20	BSSYSBP	0

EDAWB::get_freq_missing_row_wise(data)

	row	freq_missing (#cols:20)
1	1	2
2	2	2
3	3	2
4	4	2
5	5	2
6	8	2
7	13	2
8	14	2
9	15	2
10	19	2

Tabular Information on Missingness

naniar::miss_var_summary(data, order=TRUE, add_cumsum=TRUE)

	variable	n_miss	pct_miss	n_miss_cumsum
1	CD8IMMPH	43	43	43
2	CD8LOC	43	43	86
3	BMISI	1	1	87
4	BWGHTSI	1	1	88
5	USUBJID	0	0	0
6	STUDYID	0	0	0
7	STUDYNAME	0	0	0
8	STUDY_PHASE	0	0	0
9	AGE	0	0	0
10	SEX	0	0	0
11	RACE	0	0	0
12	ETHNIC	0	0	0
13	COUNTRY	0	0	0
14	BECOG	0	0	0
15	SMOKHX	0	0	0
16	TOBHX	0	0	0
17	APCPI	0	0	86
18	BHGHTCM	0	0	86
19	BSDIABP	0	0	87
20	BSSYSBP	0	0	87

Tabular Information on Missingness

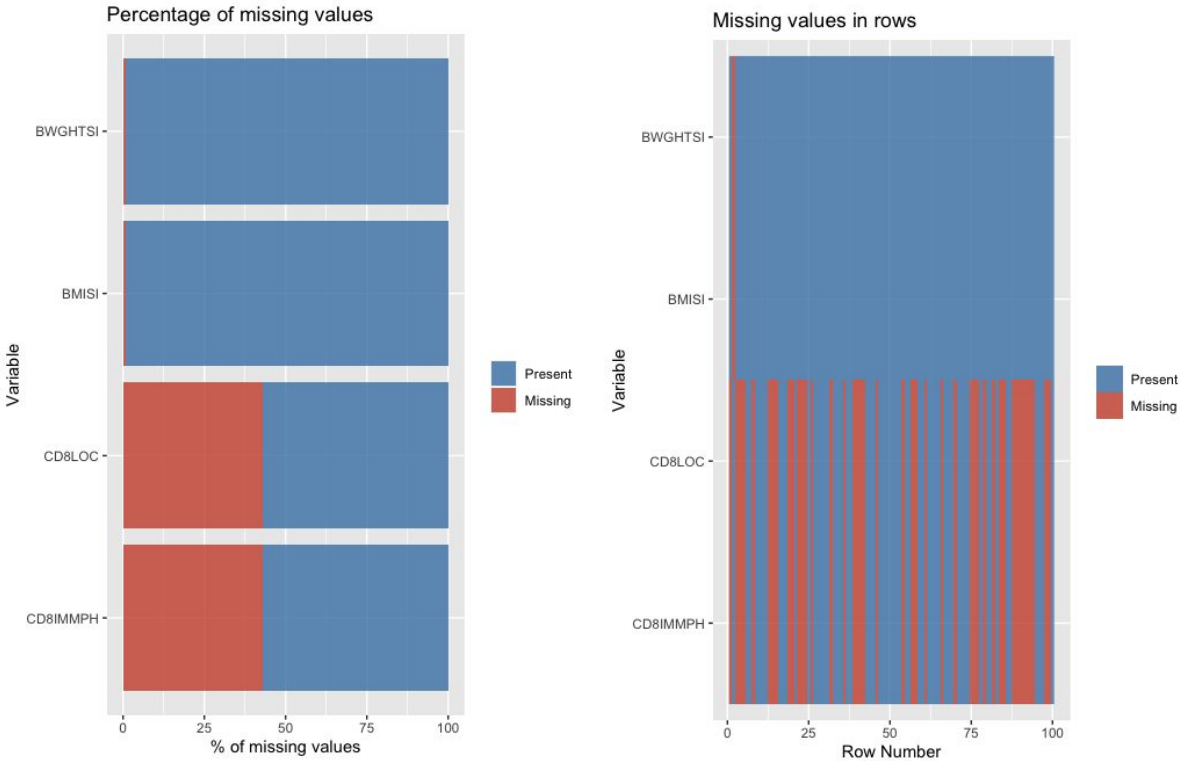
EDAWB::get_pct_bins_of_nas_for_rows_and_columns(data, get_standard_pct_missing_bins())

	pct_missing	num_rows (#100)	pct_rows
1	0%	56	56
2	(0-10]%	44	44
3	(10-20]%	0	0
4	(20-30]%	0	0
5	(30-40]%	0	0
6	(40-50]%	0	0
7	(50-60]%	0	0
8	(60-70]%	0	0
9	(70-80]%	0	0
10	(80-90]%	0	0
11	(90-100)%	0	0
12	100%	0	0

	pct_missing	num_columns (#20)	pct_columns
1	0%	16	80
2	(0-10]%	2	10
3	(10-20]%	0	0
4	(20-30]%	0	0
5	(30-40]%	0	0
6	(40-50]%	2	10
7	(50-60]%	0	0
8	(60-70]%	0	0
9	(70-80]%	0	0
10	(80-90]%	0	0
11	(90-100)%	0	0
12	100%	0	0

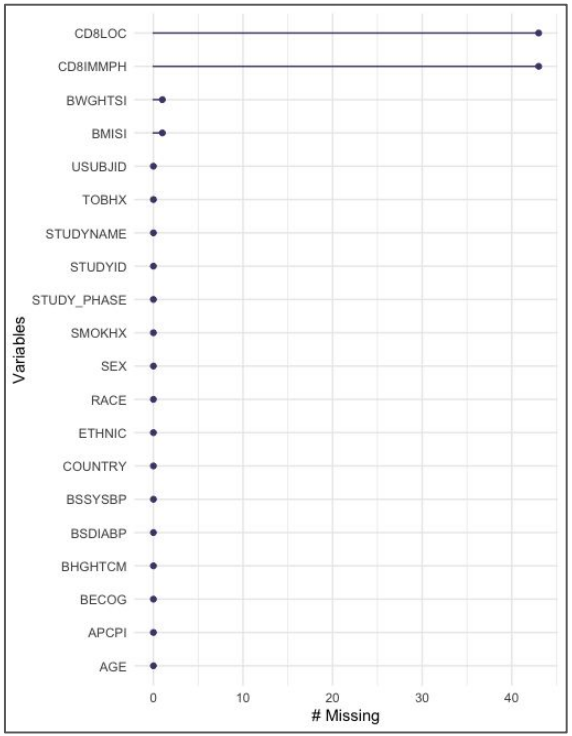
Visualization of Missingness

EDAWB::get_pct_missing_plots_for_columns_and_rows(data)

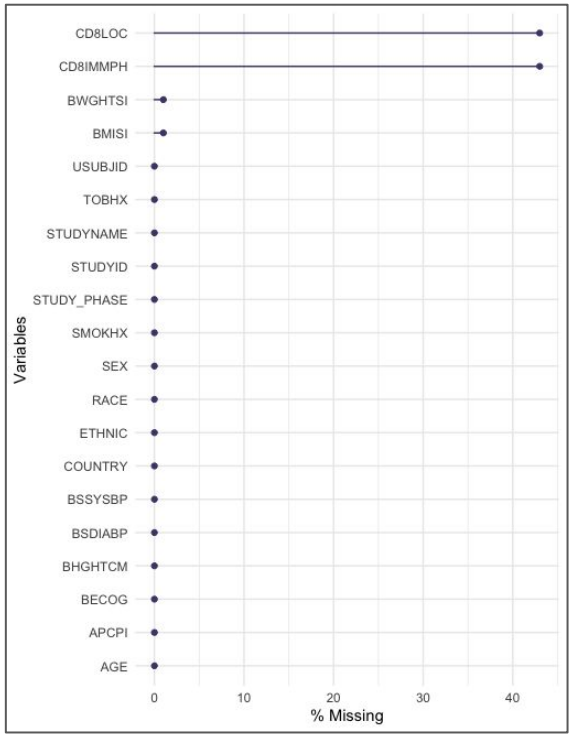


Visualization of Missingness

naniar::gg_miss_var(data, show_pct = FALSE)

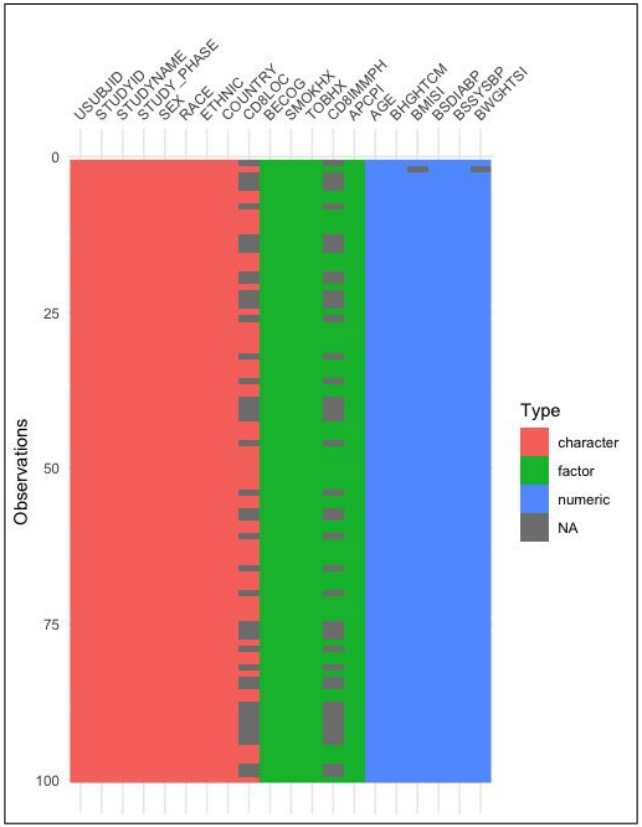


naniar::gg_miss_var(data, show_pct = TRUE)



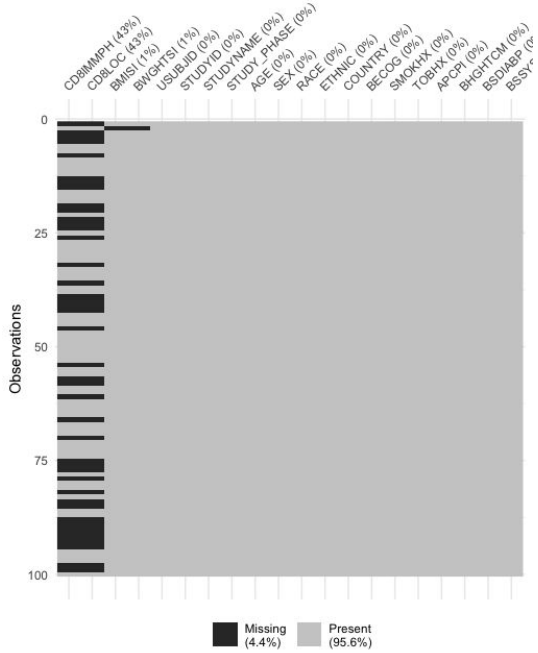
Visualization of Missingness

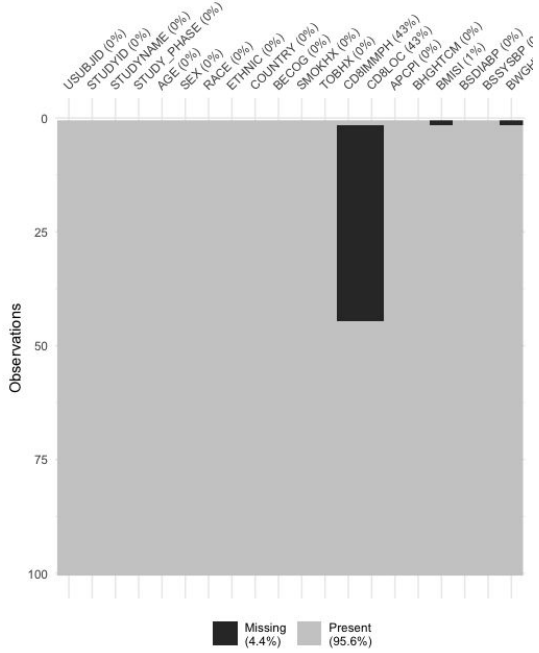
visdat::vis_dat(data, warn_large_data=FALSE)

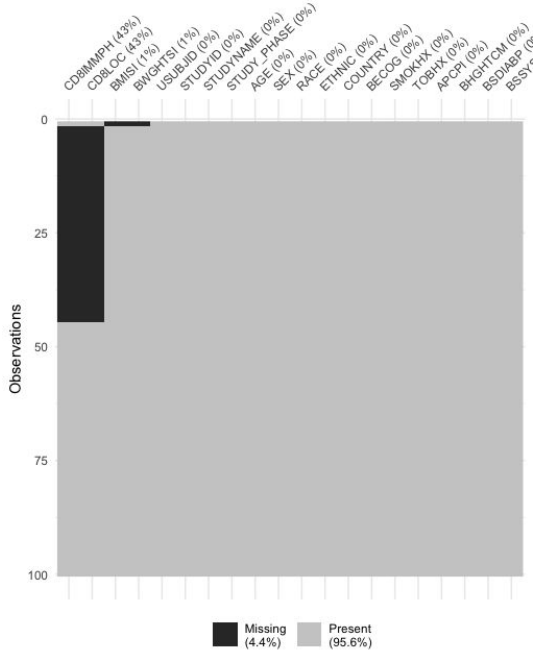


Visualization of Missingness

visdat::vis_miss(data, warn_large_data=FALSE, sort_miss = TRUE)
visdat::vis_miss(data, warn_large_data=FALSE, cluster = TRUE)
visdat::vis_miss(data, warn_large_data=FALSE, sort_miss = TRUE, cluster = TRUE)

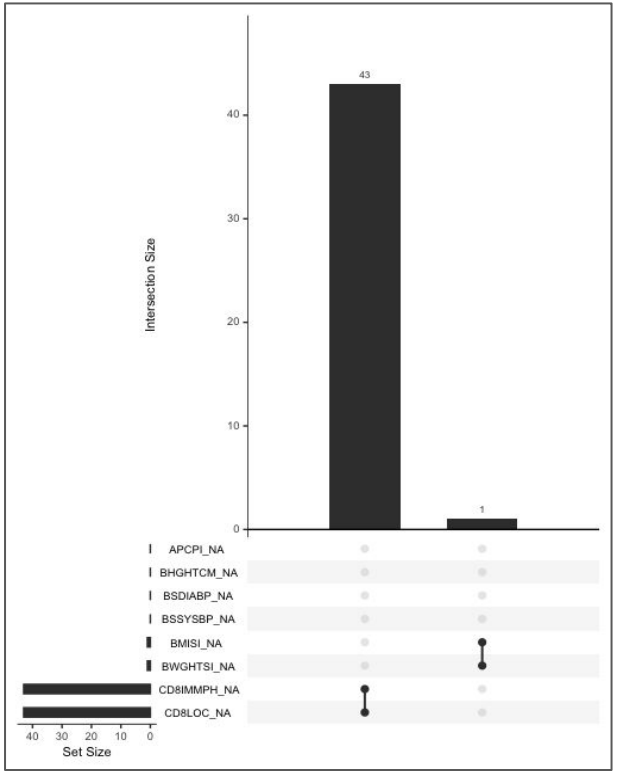






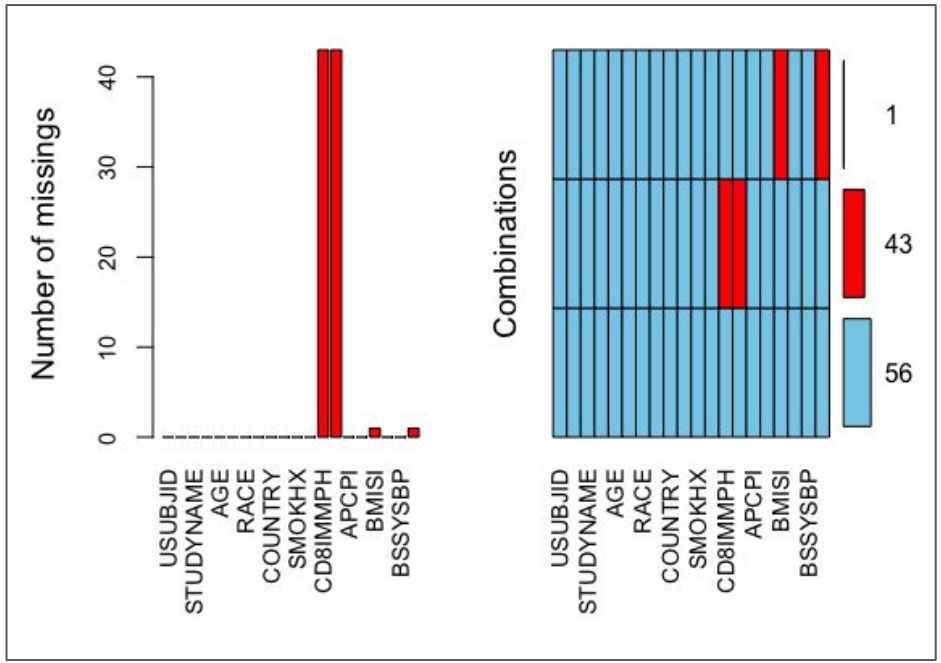
Visualization of Missingness

naniar::gg_miss_upset(data, nsets = 20, nintersects = 30)



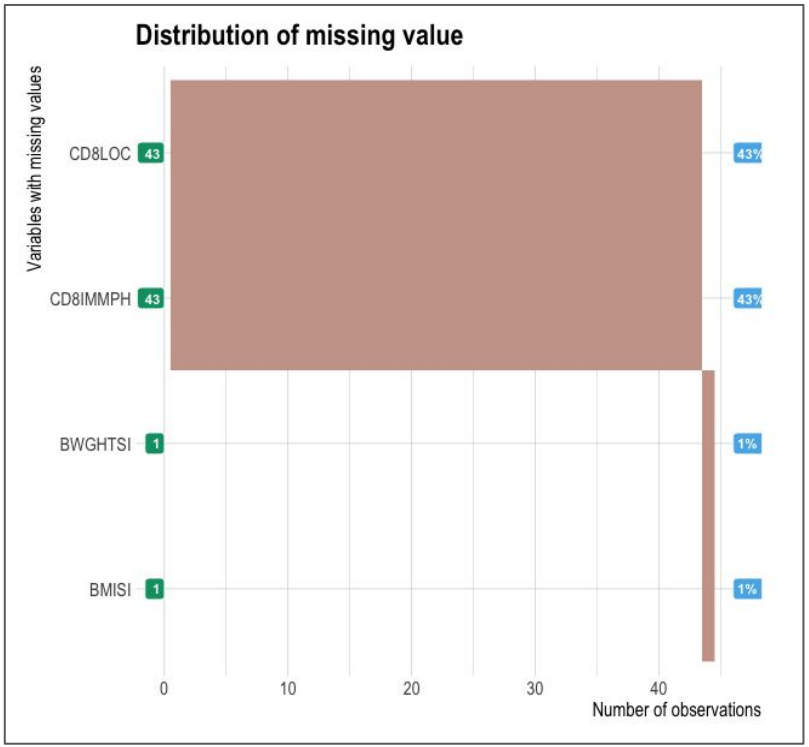
Visualization of Missingness

VIM::aggr(data, prop = F, numbers = T, combined = F, labels = names(data), cex.axis = .9, oma = c(10,5,5,3))

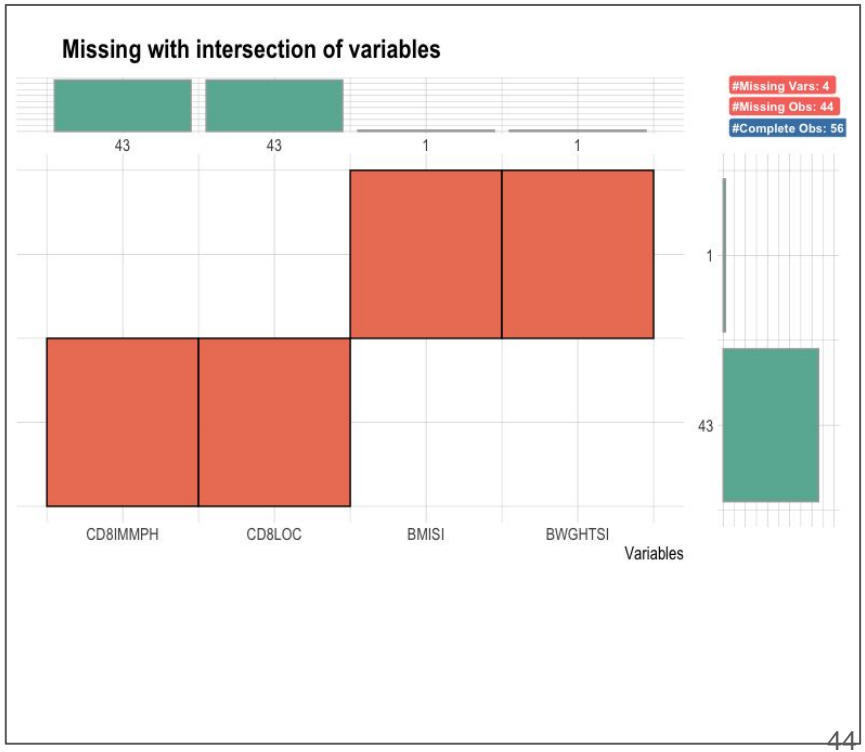


Visualization of Missingness

dlookr::plot_na_hclust(data, main = "Distribution of missing value")

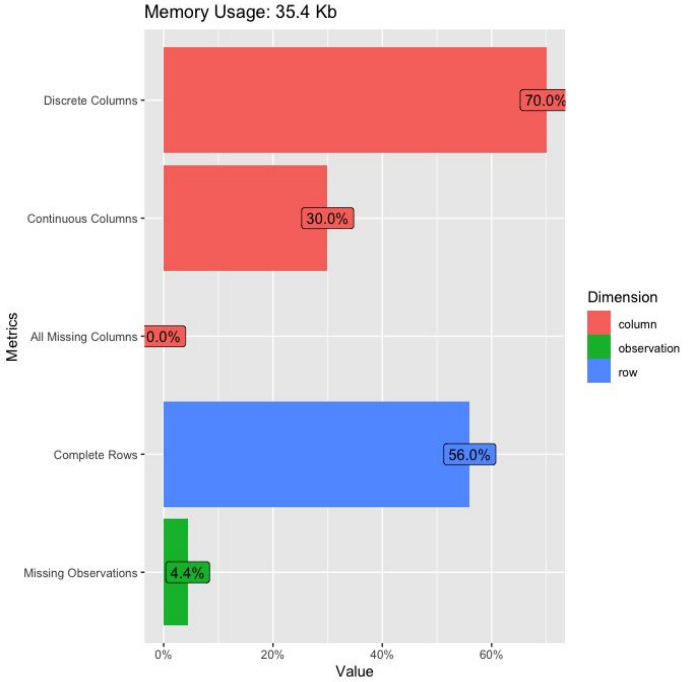


dlookr::plot_na_intersect(data)

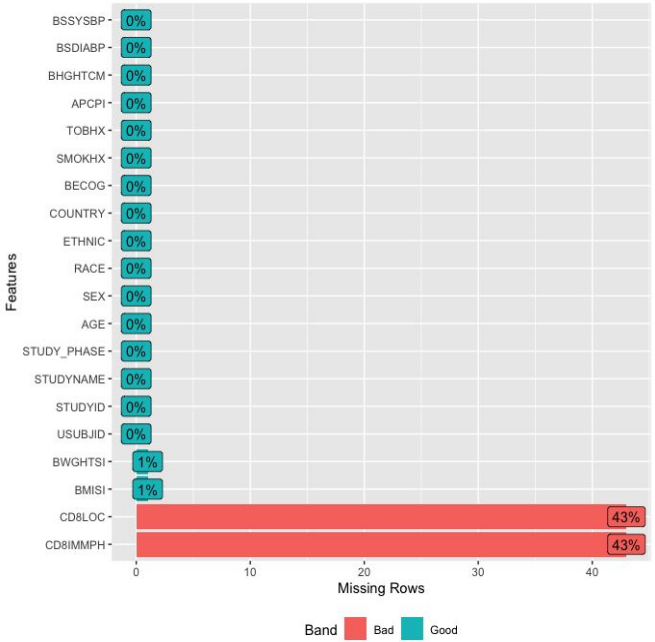


Visualization of Missingness

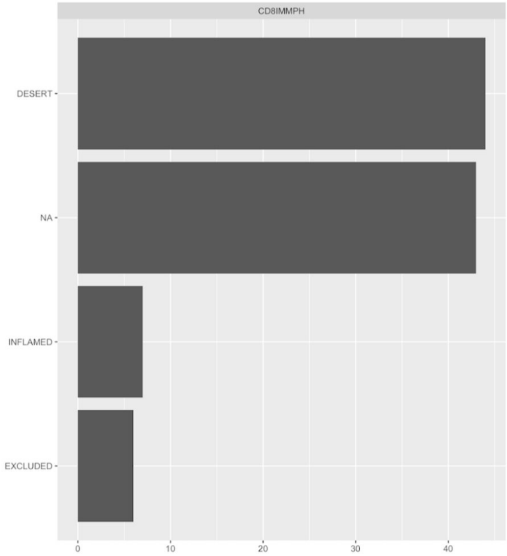
DataExplorer::plot_intro(data)



DataExplorer::plot_missing(data)



DataExplorer::plot_bar(data)
Frequency distribution of all discrete variables



Missingness

Selection of useful methods:

- *naniar::miss_var_summary(data, order=TRUE, add_cumsum=TRUE)*
- *EDAWB::get_pct_bins_of_nas_for_rows_and_columns(data, get_standard_pct_missing_bins())*
- *visdat::vis_miss(data, warn_large_data=FALSE, sort_miss = TRUE, cluster = TRUE)*
- *DataExplorer::plot_missing(data)*

Removal of NAs

Automatic Removal of NAs by rows or by columns

EDAWB::remove_nas_by_using_only_rows(data, threshold)

Input: dataframe df, threshold th between [0-100]

Output: modified dataframe with rows removed where row-wise %NA >= threshold

Procedure:

```
for each row r in df do
    r_pct_na ← get %NA
    if  r_pct_na ≥ th then
        remove r
return df
```

EDAWB::remove_nas_by_using_only_columns(data, threshold)

Input: dataframe df, threshold th between [0-100]

Output: modified dataframe with columns removed where column-wise %NA >= threshold

Procedure:

```
for each column c in df do
    c_pct_na ← get %NA
    if  c_pct_na ≥ th then
        remove c
return df
```

Recursive Removal of NAs across rows and columns

EDAWB::run_recursive_na_removal_using_columns_and_rows()

Input: dataframe df, threshold th between [0-100]

Output: modified dataframe with rows and columns removed such that %NA < threshold

Procedure:

```
def rec_na_removal_spread(df, th):
  r_pct_na ← get max %NA per rows
  c_pct_na ← get max %NA per columns
  if c_pct_na ≥ r_pct_na AND c_pct_na > th then
    remove column c from df
  if r_pct_na > c_pct_na AND r_pct_na > th then
    remove row r from df
  if th is reached then
    return df
  else
    rec_na_removal_spread(df, th)
```

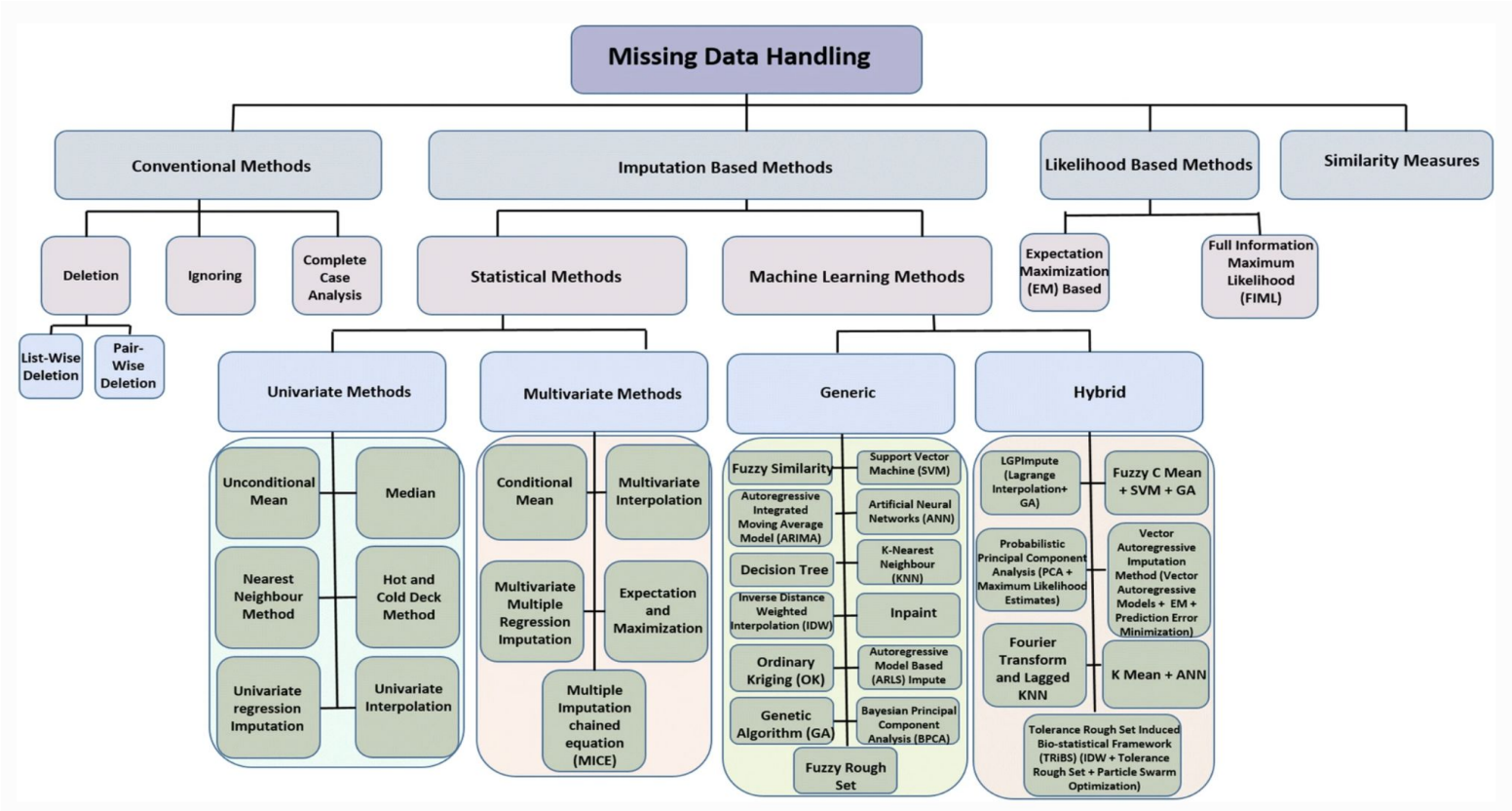
Removal of NAs

Selection of useful methods:

- *EDAWB::remove_nas_by_using_only_rows(data, threshold)*
- *EDAWB::remove_nas_by_using_only_columns(data, threshold)*
- *EDAWB::run_recursive_na_removal_using_columns_and_rows()*

Imputation

Imputation Techniques



Source: N.Gangadharan et al.Metaheuristic approaches in biopharmaceutical process development data analysis, 2019

R packages for Data Imputation

An incomplete list...

- **mice:** Multivariate Imputation by Chained Equations
- **VIM:** Visualization and Imputation of Missing Values
- **Amelia:** multiple imputation of multivariate incomplete data. It uses an algorithm that combines bootstrapping and the EM algorithm to take draws from the posterior of the missing data
- **missForest:** Nonparametric missing value imputation using Random Forest, particularly in the case of mixed-type data
- **mi:** provides functions for data manipulation, imputing missing values in an approximate Bayesian framework, diagnostics of the models used to generate the imputations, confidence building mechanisms to validate some of the assumptions of the imputation algorithm, and functions to analyze multiply imputed data sets with the appropriate degree of sampling uncertainty.
- **Hmisc:** various imputation functions
- **miceRanger:** Multiple Imputation by Chained Equations with Random Forests
- **simputation:** interface to different imputation methods

mice Package

The mice package contains functions to

- Inspect the missing data pattern
- Impute the missing data m times, resulting in m completed data sets
- Diagnose the quality of the imputed values
- Analyze each completed data set
- Pool the results of the repeated analyses
- Store and export the imputed data in various formats
- Generate simulated incomplete data
- Incorporate custom imputation methods

univariate imputation methods

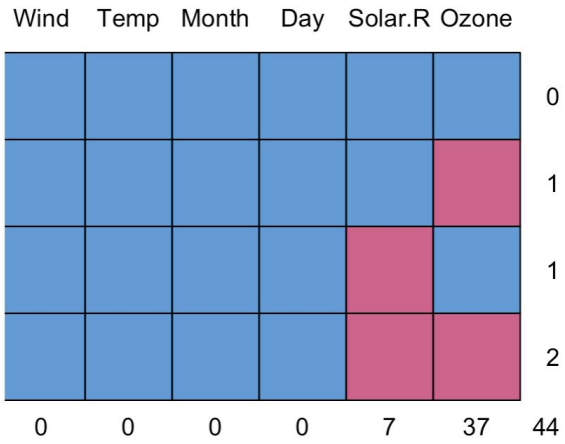
Method Name	Data Type	Description
pmm	any	Predictive mean matching
midastouch	any	Weighted predictive mean matching
sample	any	Random sample from observed values
cart	any	Classification and regression trees
rf	any	Random forest imputations
mean	numeric	Unconditional mean imputation
norm	numeric	Bayesian linear regression
norm.nob	numeric	Linear regression ignoring model error
norm.boot	numeric	Linear regression using bootstrap
norm.predict	numeric	Linear regression, predicted values
quadratic	numeric	Imputation of quadratic terms
ri	numeric	Random indicator for nonignorable data
logreg	binary	Logistic regression
logreg.boot	binary	Logistic regression with bootstrap
polr	ordered	Proportional odds model
polyreg	unordered	Polytomous logistic regression
lda	unordered	Linear discriminant analysis
2l.norm	numeric	Level-1 normal heteroscedastic
2l.lmer	numeric	Level-1 normal homoscedastic, lmer
2l.pan	numeric	Level-1 normal homoscedastic, pan
2l.bin	binary	Level-1 logistic, glmer
2lonly.mean	numeric	Level-2 class mean
2lonly.norm	numeric	Level-2 class normal
2lonly.pmm	any	Level-2 class predictive mean matching

Sources:
Stef van Buuren, Karin Groothuis-Oudshoorn: mice: Multivariate Imputation by Chained Equations in R
https://www.rdocumentation.org/packages/mice/versions/3.13.0/topics/mice
https://stefvanbuuren.name/fimd/how-to-generate-multiple-imputations.html

mice Package

Displaying missing-data patterns for investigating any structure of missing observations in the data.
In specific case, the missing data pattern could be (nearly) monotone. Monotonicity can be used to simplify the imputation model.

```
library(mice)
mice::md.pattern(airquality)
```




```
    Wind Temp Month Day Solar.R Ozone
111    1    1     1   1       1     1  0
35     1    1     1   1       1     0  1
5      1    1     1   1       0     1  1
2      1    1     1   1       0     0  2
       0    0     0   0       7    37 44
```

Source: N.Gangadharan et al.Metaheuristic approaches in biopharmaceutical process development data analysis, 2019

mice Package

```
library(mice)
imp <- mice::mice(airquality,
                  m=5,          # Number of multiple imputations
                  maxit=50,     # A scalar giving the number of iterations.
                  meth='pmm', # Single string or a vector of strings, specifying the method to be used for each column in data.
                  printFlag=FALSE,
                  seed=500)
summary(imp)
```

```
Class: mids
Number of multiple imputations:  5
Imputation methods:
  Ozone Solar.R    Wind    Temp   Month     Day
  "pmm"   "pmm"      ""      ""      ""      ""
PredictorMatrix:
        Ozone Solar.R Wind Temp Month Day
Ozone       0       1    1    1     1   1
Solar.R     1       0    1    1     1   1
Wind        1       1    0    1     1   1
Temp        1       1    1    0     1   1
Month       1       1    1    1     0   1
Day         1       1    1    1     1   0
```

```
mice::complete(imp, include=TRUE, action="broad")
```

	Ozone.0	Solar.R.0	Wind.0	Temp.0	Month.0	Day.0	Ozone.1	Solar.R.1	Wind.1	Temp.1	...
1	41	190	7.4	67	5	1	41	190	7.4	67	
2	36	118	8.0	72	5	2	36	118	8.0	72	
3	12	149	12.6	74	5	3	12	149	12.6	74	
4	18	313	11.5	62	5	4	18	313	11.5	62	
5	NA	NA	14.3	56	5	5	14	237	14.3	56	
6	28	NA	14.9	66	5	6	28	82	14.9	66	

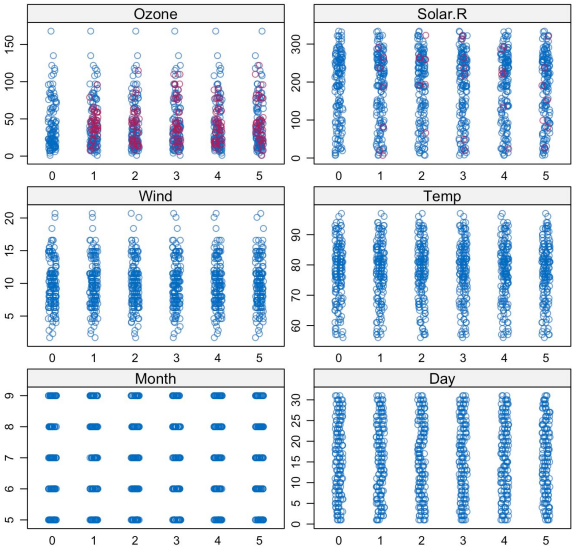
...
...

```
mice::complete(imp, action=1)
```

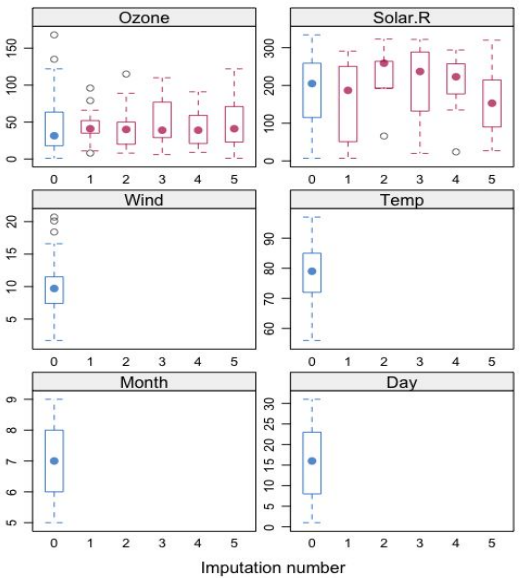
	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	14	237	14.3	56	5	5
6	28	82	14.9	66	5	6

mice Package

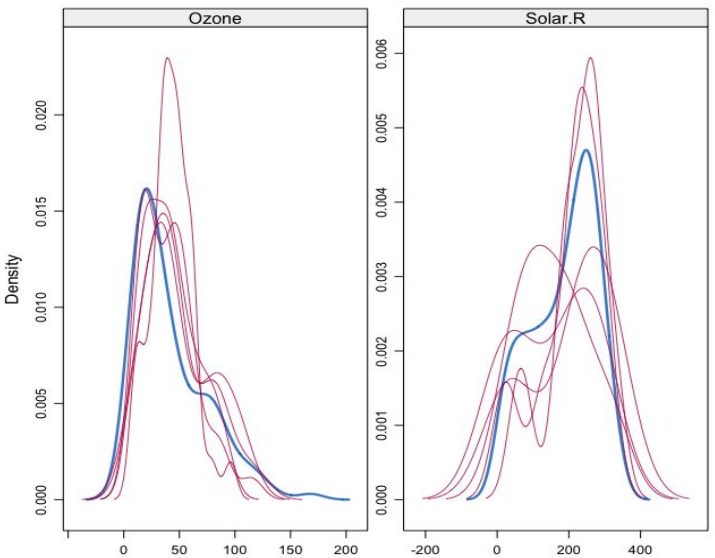
mice::stripplot(imp)



mice::bwplot(imp)



mice::densityplot(imp)



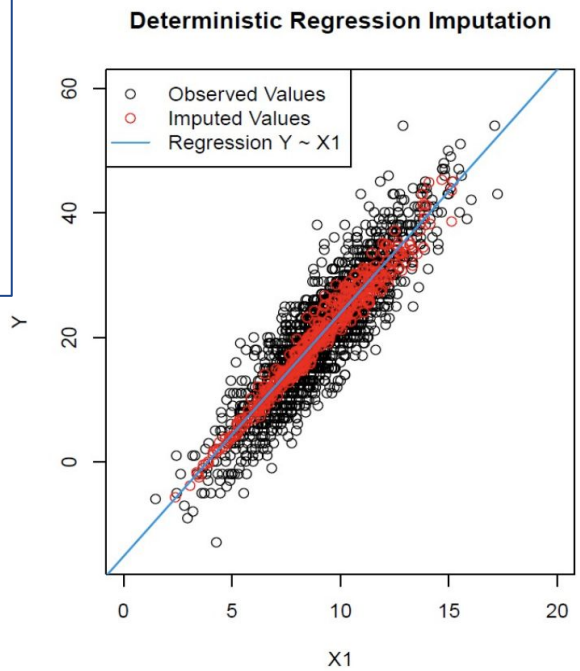
Regression Imputation using *mice*

Regression imputation fits a statistical model on a variable with missing values. Predictions of this regression model are used to substitute the missing values in this variable.

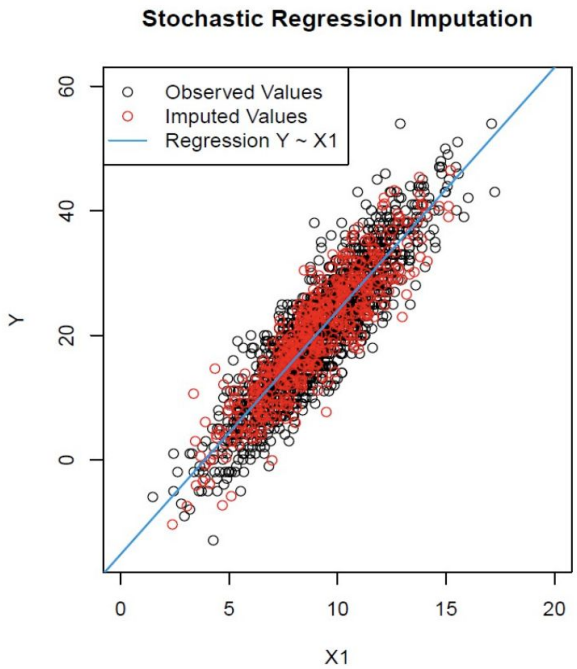
Deterministic regression imputation replaces missing values with the exact prediction of the regression model.

Stochastic regression imputation adds a random error term to the predicted value.

```
imp <- mice::mice(data, method = "norm.predict", m = 1)
data_det <- mice::complete(imp)
```



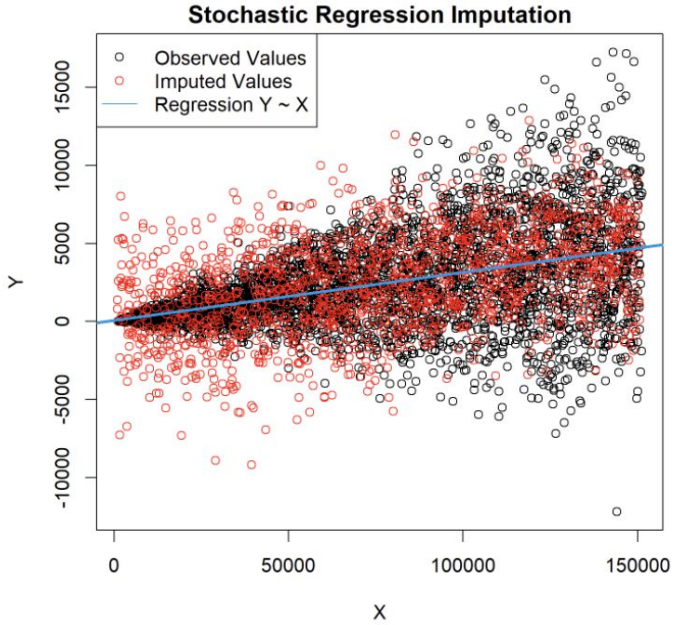

```
imp <- mice::mice(data, method = "norm.nob", m = 1)
data_sto <- mice::complete(imp)
```



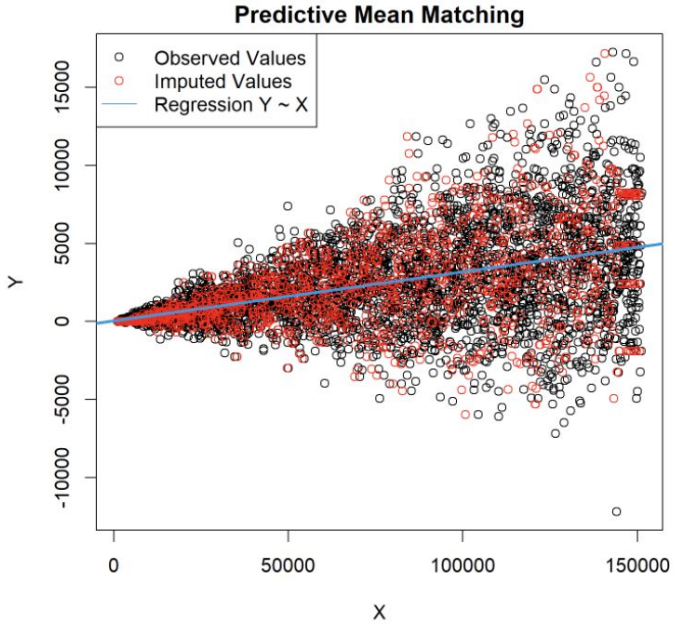
Source: https://statisticsglobe.com/regression-imputation-stochastic-vs-deterministic/

Imputation: Predictive Mean vs Stochastic Regression Imputation using *mice*

```
imp <- mice::mice(data, method = "norm.nob", m = 1)
data_sto <- mice::complete(imp)
```




```
imp <- mice::mice(data, method = "pmm", m = 1)
data_pmm <- mice::complete(imp)
```



Source: https://statisticsglobe.com/predictive-mean-matching-imputation-method/

VIM Package: Imputation Methods

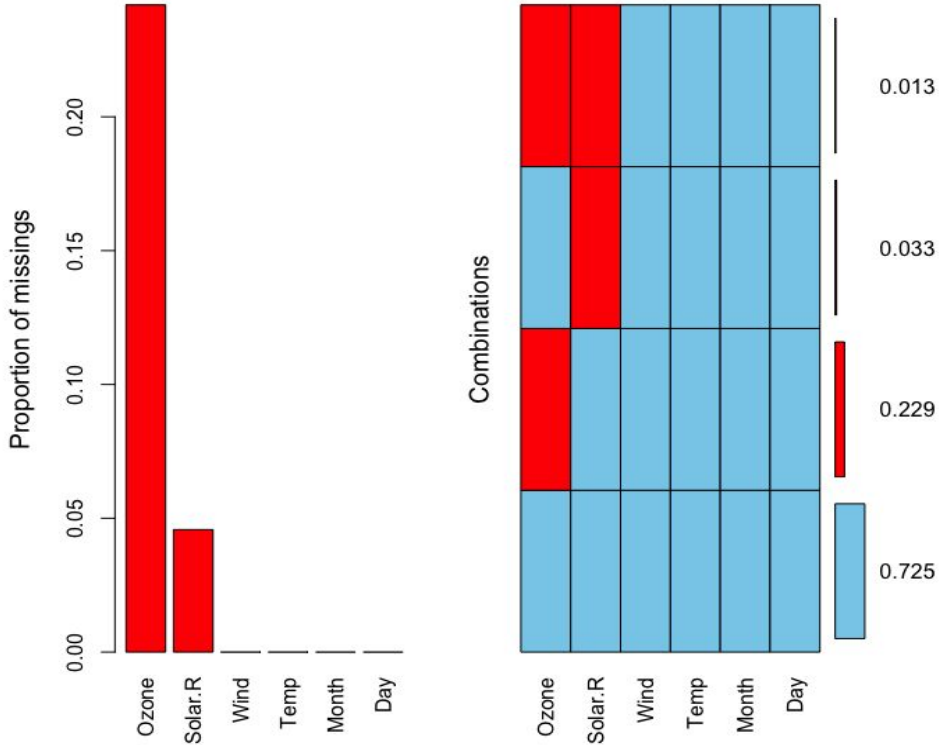
Methods for imputation

- *hotdeck():* Hot-deck imputation
- *kNN():* k nearest neighbor imputation
- *irmi():* Iterative robust model-based imputation
- *regressionImp():* Impute missing values based on a regression model
- *rangerImpute():* Impute missing values based on a random forest model
- *matchImpute():* Suitable donors are searched based on matching of the categorical variables

VIM: Visualizations of Missingness Patterns

Plotting the amount of missing/imputed values in each variable and the amount of missing/imputed values in certain combinations of variables.

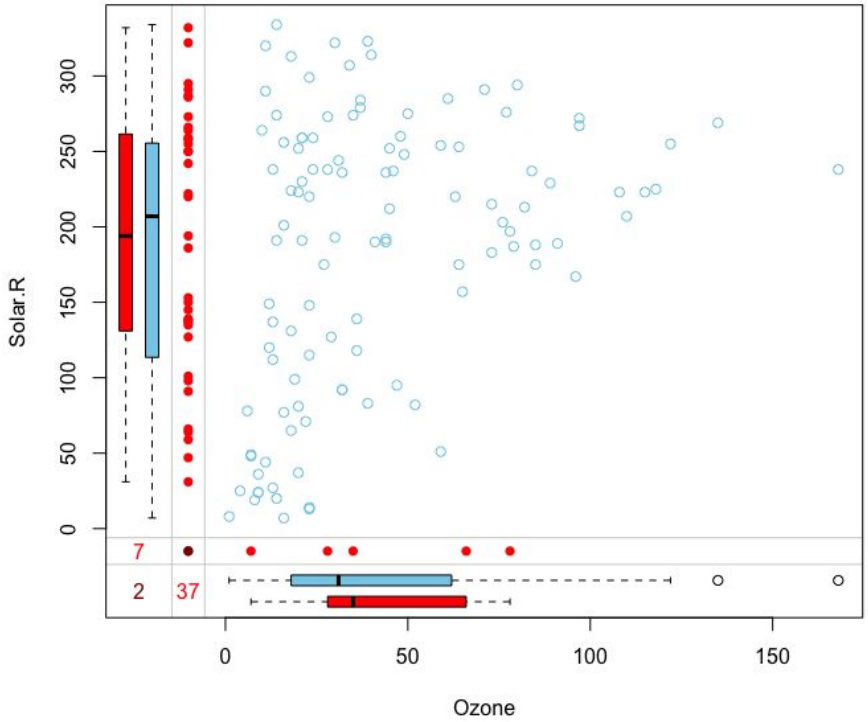
```
library(VIM)
VIM::aggr(airquality,
          bars = TRUE,
          numbers = TRUE,
          prop = TRUE,
          combined = FALSE,
          varheight = FALSE,
          only.miss = FALSE,
          sortVars = TRUE,
          sortCombs = TRUE)
```



VIM: Visualizations of Missingness Patterns

Scatterplot with information about missing/imputed values shown in the plot margins.
Imputed values are highlighted in the scatterplot.

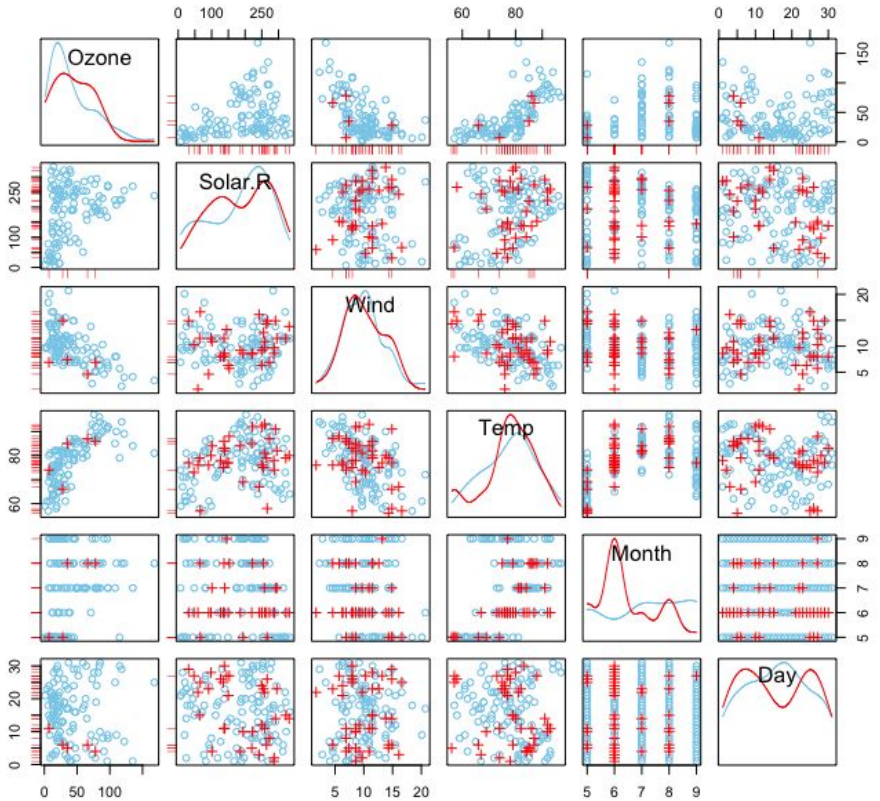
```
library(VIM)
VIM::marginplot(airquality[c(1,2)])
```



VIM: Visualizations of Missingness Patterns

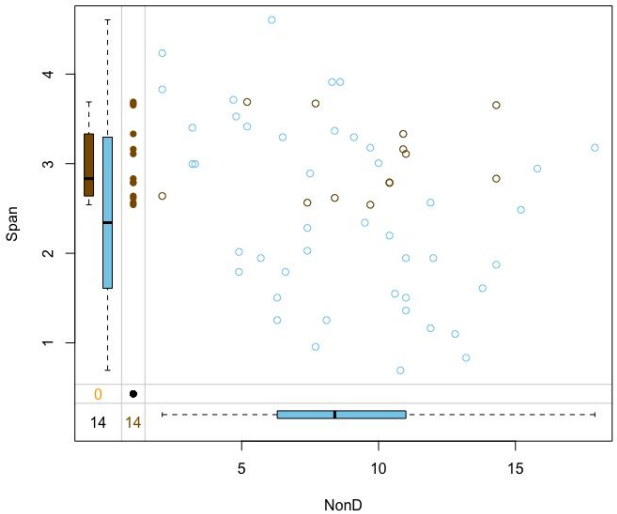
Scatterplot matrix in which observations with missing/imputed values in certain variables are highlighted.

```
library(VIM)
VIM::scattmatrixMiss(airquality)
```

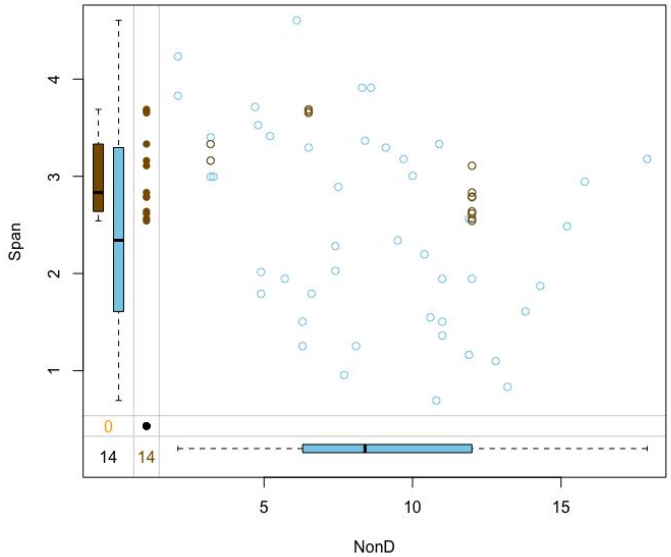


VIM: Assessment of Imputation

```
imp_hotdeck <- VIM::hotdeck(dataset, variable = "NonD")
imp_hotdeck[, c("NonD", "Span", "NonD_imp")] %>%
    VIM::marginplot(delimiter = "_imp")
```




```
imp_knn <- VIM::kNN(dataset, variable = "NonD")
imp_knn[, c("NonD", "Span", "NonD_imp")] %>%
    VIM::marginplot(delimiter = "_imp")
```



Source: https://cran.r-project.org/web/packages/VIM/vignettes/donorImp.html

R Package *simputation*

Imputation methods

- Model based (optionally add [non-]parametric random residual)
  - linear regression
  - robust linear regression
  - ridge/elasticnet/lasso regression
  - CART models (decision trees)
  - Random forest
- Multivariate imputation
  - Imputation based on the expectation-maximization algorithm
  - missForest (=iterative random forest imputation)
- Donor imputation (including various donor pool specifications)
  - k-nearest neigbour (based on gower's distance)
  - sequential hotdeck (LOCF, NOCB)
  - random hotdeck
  - Predictive mean matching
- Other
  - (groupwise) median imputation (optional random residual)
  - Proxy imputation: copy another variable or use a simple transformation to compute imputed values.
  - Apply trained models for imputation purposes.

simputation functions

function.	model	package	R.recommended
impute_rlm	M-estimation	MASS	yes
impute_en	ridge/elasticnet/lasso	glmnet	no
impute_cart	CART	rpart	yes
impute_rf	random forest	randomForest	no
impute_rhd	random hot deck	VIM (optional)	no
impute_shd	sequential hot deck	VIM (optional)	no
impute_knn	k nearest neighbours	VIM (optional)	no
impute_mf	missForest	missForest	no
impute_em	mv-normal	norm	no

Source: https://cran.r-project.org/web/packages/simputation/vignettes/intro.html

Imputation

Selection of useful methods:

- *EDAWB::impute_by_column_mean()*
- *EDAWB::impute_by_column_median()*
- *EDAWB::impute_by_column_mode()*
- *mice::mice()*
- *VIM::hotdeck()*
- *VIM::kNN()*

Feature Transformation

Transformation

Package	Method	Description
EDAWB	do_log2_transformation()	Produces a log2 transformation
EDAWB	do_zscore_transformation()	z-score (standard score) provides a measure of how far from the mean a data point is by giving the number of standard deviations by which the value of a raw score is above or below the mean value of what is being observed or measured.
EDAWB	do_minmax_transformation()	All features are transformed to be in the range [0,1]
EDAWB	do_softmax_transformation()	Produces a softmax transformation to map the input values between 0 and 1 so that they can be interpreted as probabilities. It is a probability distribution consisting of probabilities proportional to the exponentials of the input numbers.

Discretization of Continuous Data

Package	Method	Description
EDAWB	discretize_by_median()	Discretization by median
EDAWB	discretize_by_quartiles()	Discretization by quartiles
EDAWB	discretize_by_deciles()	Discretization by deciles
arules	discretize(): method="interval"	Discretization by equal interval width
arules	discretize(): method="frequency"	Discretization by equal frequency
arules	discretize(): method="cluster"	Discretization by k-means clustering
arules	discretize(): method="fixed"	Discretization by specified interval boundaries

Source: https://www.rdocumentation.org/packages/arules/versions/1.6-8/topics/discretize

Discretization of Continuous Data using varrank::discretization()

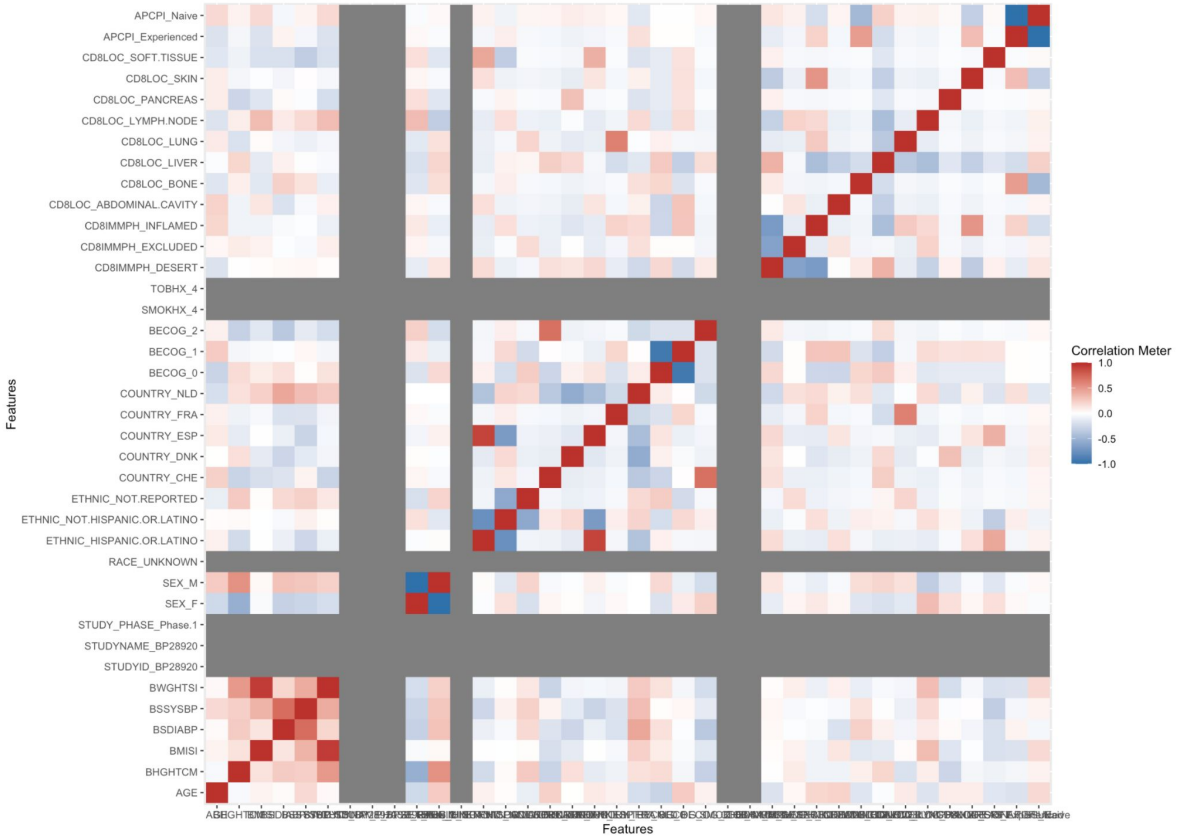
Parameter	Rule	Number of Bins
fd	Freedman–Diaconis rule	$\frac{range(x) * n^{1/3}}{2 * IQR(x)}$
doane	doane's rule	$1 + \log_2 n + \log_2 1 + \frac{\lvert g \rvert}{\sigma_g}$
cencov	Cencov's rule	$n^{1/3}$
rice	Rice' rule	$2n^{1/3}$
terrell-scott	Terrell-Scott's rule	$(2n)^{1/3}$
sturges	Sturges's rule	$1 + \log_2(n)$
scott	Scott's rule	$range(x)/\sigma(x)n^{-1/3}$
kmeans	applies the classical k-means clustering to one-dimensional continuous data	

Source: https://www.math.uzh.ch/pages/varrank/reference/discretization.html

Correlation Analysis

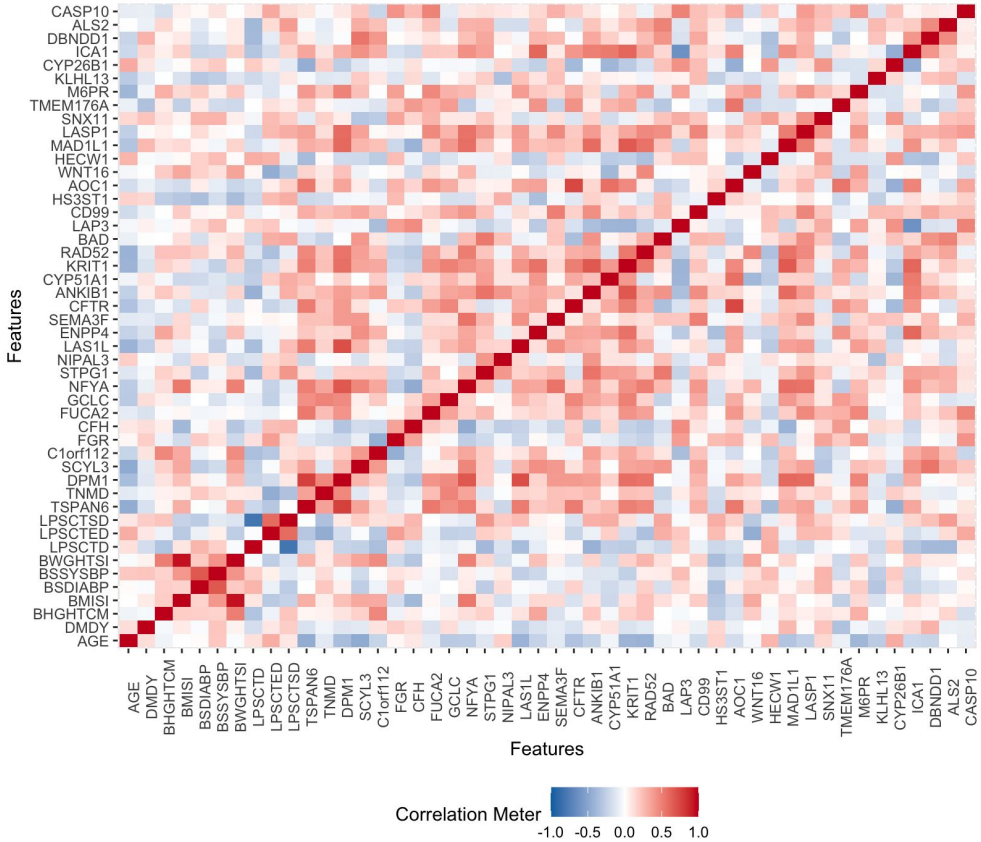
Correlation Analysis

DataExplorer::create_report(data, output_file='dataexplorer_report.html')



Correlation Analysis

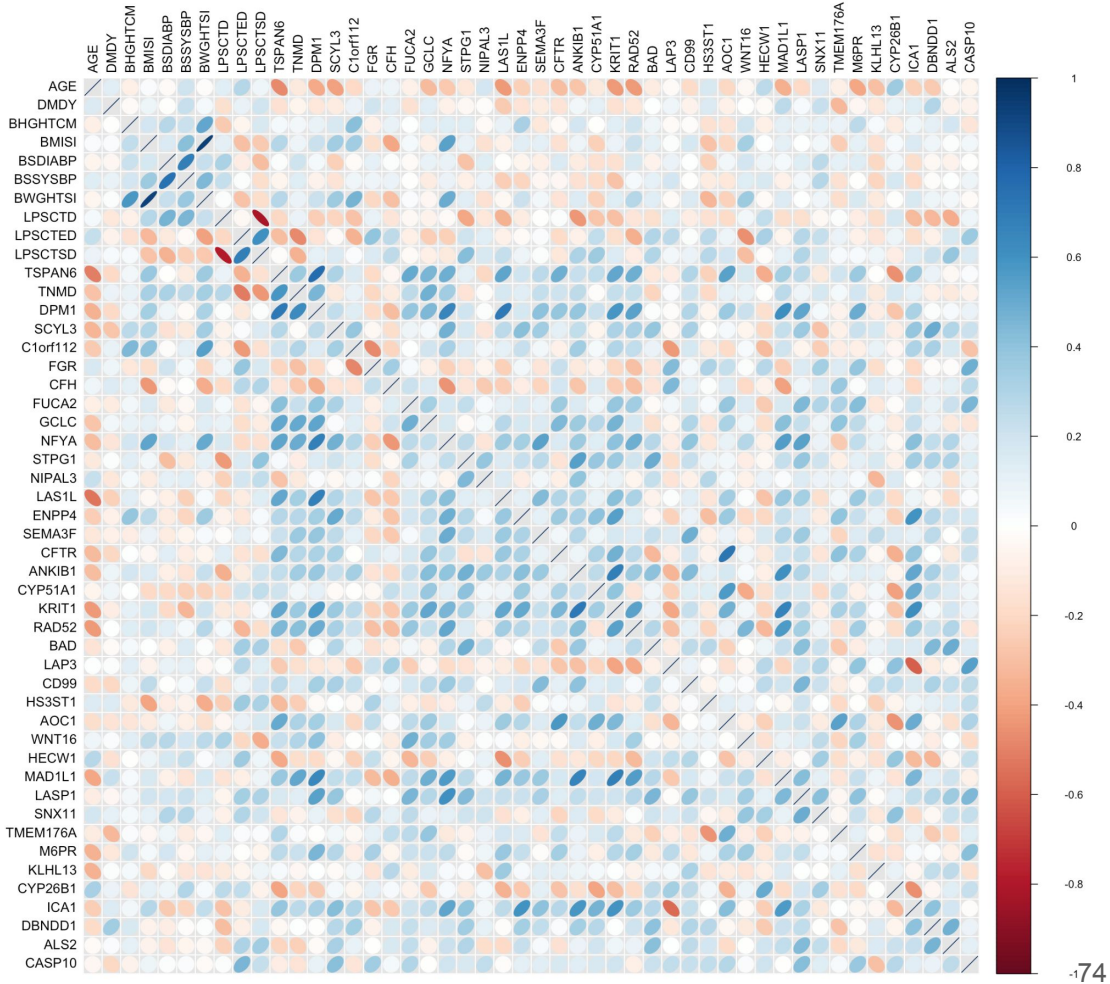
DataExplorer::plot_correlation(continuous_data)



Correlation Analysis

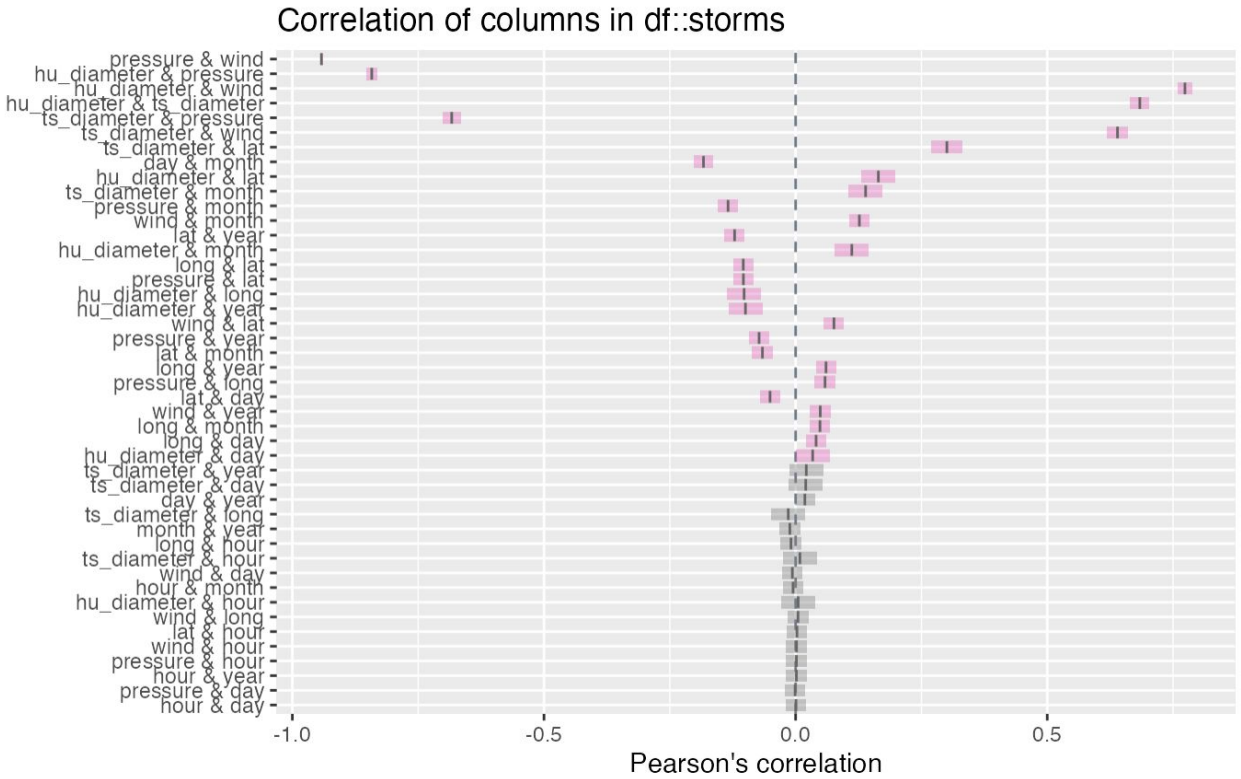
ExPanDaR::ExPanD(data)

This plot visualizes sample correlations (Pearson above, Spearman below diagonal). Reports correlations for all continuous variables.



Correlation Analysis

inspectdf::inspect_cor(data, method = "pearson") %>% inspectdf::show_plot()
Pearson, Kendall and Spearman correlations for numeric columns in one, two or grouped dataframes.



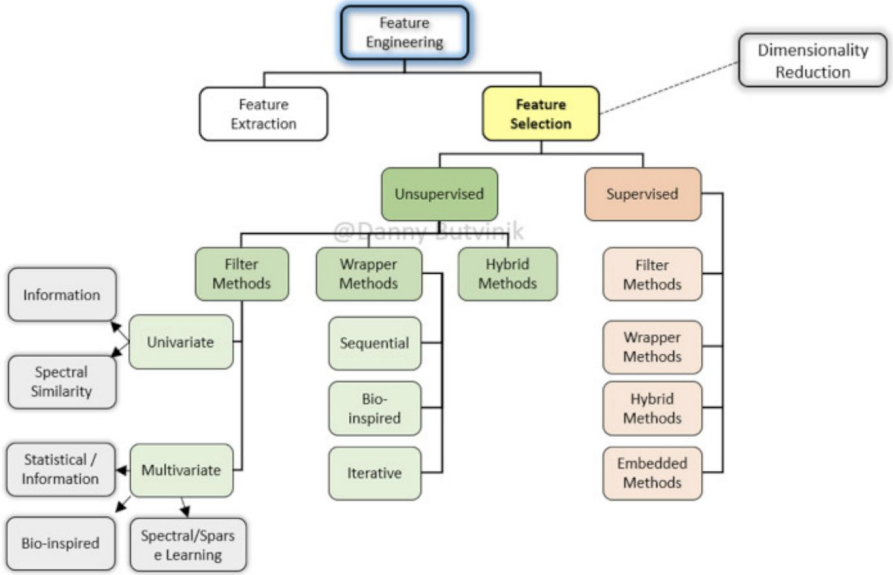
R packages for Correlation Analysis

Package	Method	Description
PerformanceAnalytics	chart.Correlation()	Visualization of a Correlation Matrix. On top the (absolute) value of the correlation plus the result of the cor.test as stars. On bottom, the bivariate scatterplots, with a fitted line
DataExplorer	create_report()	Creates a data profiling report with a correlation plot included
DataExplorer	plot_correlation()	Creates a correlation heatmap for all discrete categories
GGally	ggcorr()	Creates a correlation matrix plot using ggplot2
GGally	ggpairs()	Creates a ggplot2 based generalized pairs plot to show scatterplot matrix
ExPanDaR	ExPanD()	A shiny based web app that uses ExPanDaR functionality for interactive data exploration including correlation plots
inspectdf	inspect_cor()	Pearson, Kendall and Spearman correlations for numeric columns in one, two or grouped dataframes

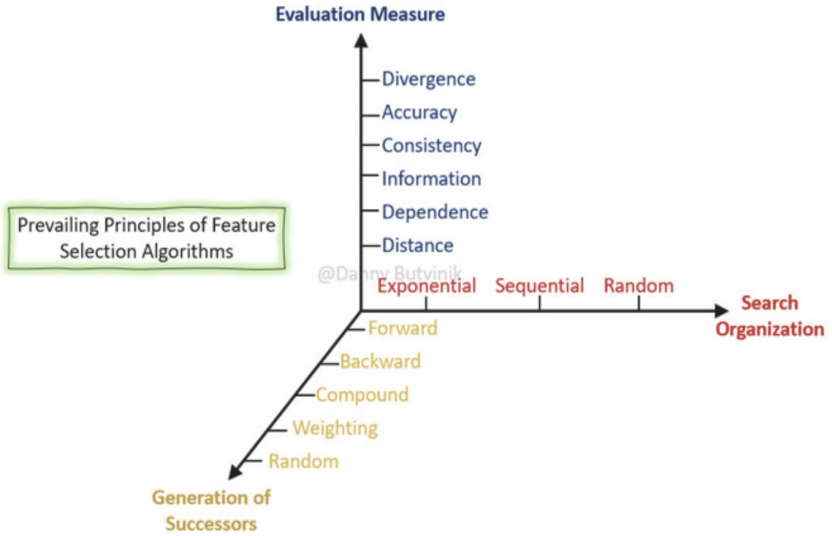
Feature Selection

Feature Selection

High-level taxonomy for feature selection



Characterization of a feature selection algorithm



Source: https://www.kdnuggets.com/2021/06/feature-selection-overview.html

Feature Selection

data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$

target variable y (discrete or continuous)

feature selection algorithm:
selection of subset of $k << p$ columns, $\mathbf{X}_S \in \mathbb{R}^{n \times k}$
that are most relevant to the target variable y

R packages for feature selection

- FSelector
- FSelectorRcpp
- Boruta
- Varrank
- varSelRF
- FSinR
- caret

Feature Filtering

What is feature filtering?

- Filter methods select features from a dataset without a target variable and independently for any machine learning algorithm.
- These methods rely only on the characteristics of these variables.
- The goal is to eliminate irrelevant, redundant, constant, duplicated, and correlated features.
- Univariate filter methods evaluate and rank a single feature according to certain criteria.
- Multivariate filter methods evaluate the entire feature space. They take into account features in relation to other ones in the dataset.

Feature Filter Methods

Basic Filter Methods

- Constant Features
- Quasi-Constant Features
- Duplicated Features

Correlation Filter Methods

- Pearson correlation coefficient
- Spearman's rank correlation coefficient
- Kendall's rank correlation coefficient

Statistical & Ranking Filter Methods

- Mutual Information: measure of the mutual dependence of two variables. It measures the amount of information obtained about one variable through observing the other variable.
- Chi-squared Score: testing relationships between categorical variables. Suited for categorical variables and binary targets only, and the variables should be non-negative and typically boolean, frequencies, or counts.
- ANOVA Univariate Test: measures the dependence of two variables. Assumes a linear relationship between the variables and the target, and also that the variables are normally distributed. It's well-suited for continuous variables and requires a binary target.
- Univariate ROC-AUC /RMSE: uses machine learning models to measure the dependence of two variables. It's suitable for all variables, and also makes no assumptions about their distribution.

Source: https://www.kdnuggets.com/2021/06/feature-selection-overview.html

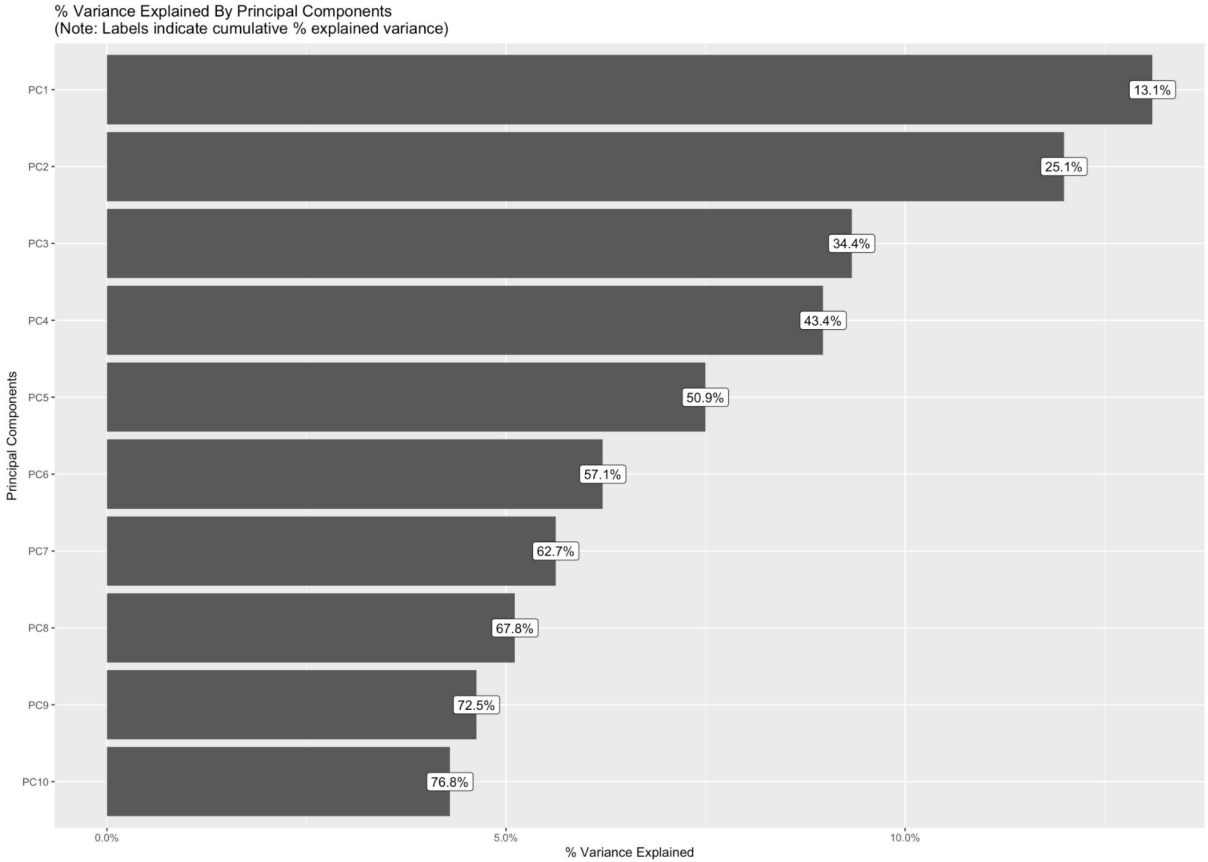
Feature Filtering

Package	Method	Description
EDAWB	select_top_features_by_variance()	Order features by variance and select top n features
EDAWB	select_top_features_by_mad()	Order features by MAD and select top n features
EDAWB	select_top_features_by_entropy()	Order features by entropy and select top n features
EDAWB	select_top_features_by_mutual_information()	Order features by mutual information and select top n features
EDAWB	select_top_features_by_correlation()	Order features by correlation coefficient starting with the least correlated ones and select top n features
EDAWB	select_top_features_by_laplacian_score()	Order features by Laplacian Score and select top n features

PCA Analysis

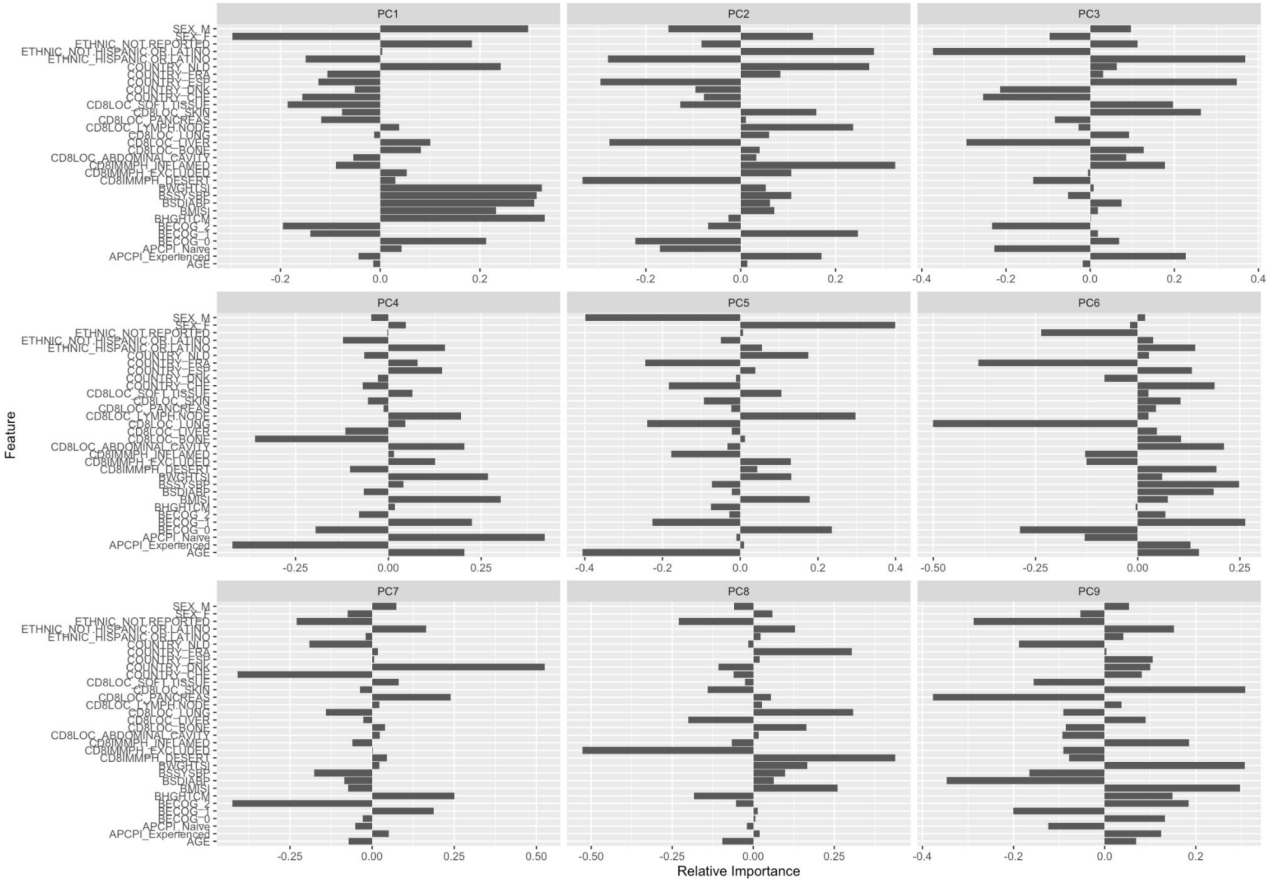
Principal Component Analysis

DataExplorer::create_report(data, output_file='dataexplorer_report.html')



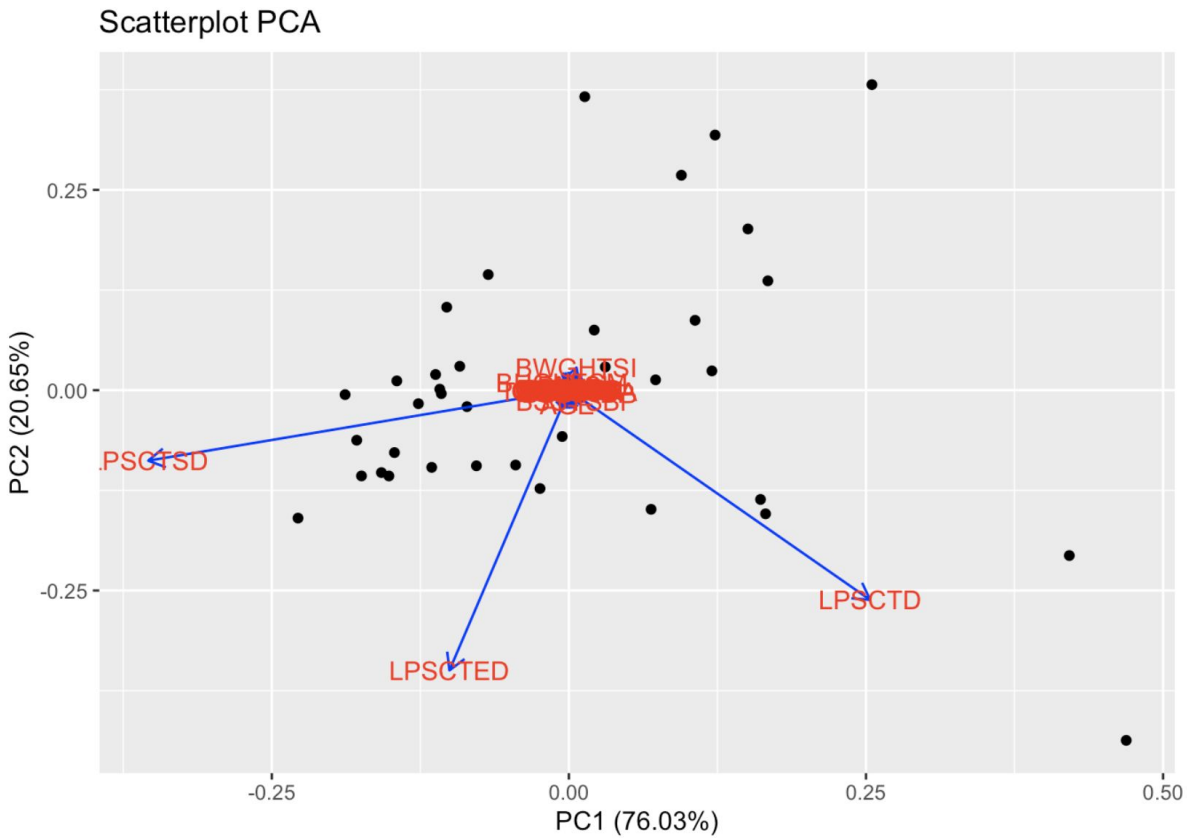
Principal Component Analysis

DataExplorer::create_report(data, output_file='dataexplorer_report.html')



Principal Component Analysis

AEDA::fastReport(data=data)



R Package *explor* by Julien Barnier

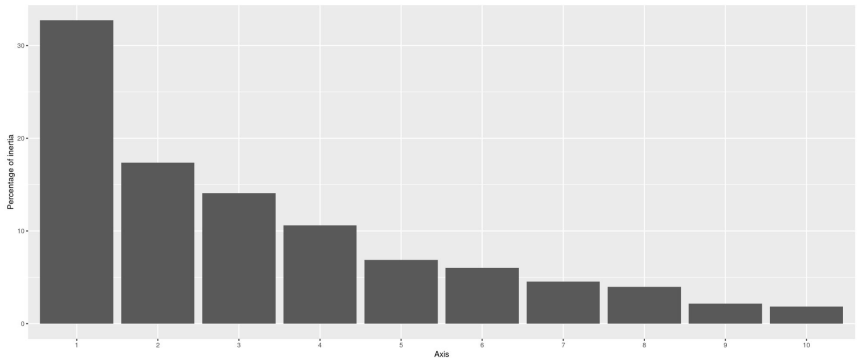

```
library(FactoMineR)
library(explor)

data(decathlon)
pca <- FactoMineR::PCA(decathlon[,1:12], quanti.sup = 11:12, graph = FALSE)
explor::explor(pca)
```

PCA | Eigenvalues | Variables plot | Variables data | Individuals plot | Individuals data

Dimensions to plot

10

Eigenvalues histogram



Eigenvalues table

Axis	%	Cum. %
1	32.7	32.7
2	17.4	50.1
3	14.0	64.1
4	10.6	74.7
5	6.8	81.6
6	6.0	87.5
7	4.5	92.1
8	4.0	96.0
9	2.1	98.2
10	1.8	100.0

Links: *https://github.com/juba/scatterD3*, *https://juba.github.io/explor/articles/introduction_en.html*

R Package *explor* by Julien Barnier

Analysis	Function	Package	Notes
Principal Component Analysis	PCA	FactoMineR	-
Correspondance Analysis	CA	FactoMineR	-
Multiple Correspondence Analysis	MCA	FactoMineR	-
Principal Component Analysis	dudi.pca	ade4	Qualitative supplementary variables are ignored
Correspondance Analysis	dudi.coa	ade4	-
Multiple Correspondence Analysis	dudi.acm	ade4	Quantitative supplementary variables are ignored
Specific Multiple Correspondance Analysis	speMCA	GDAtools	-
Multiple Correspondance Analysis	mca	MASS	Quantitative supplementary variables are not supported
Principal Component Analysis	princomp	stats	Supplementary variables are ignored
Principal Component Analysis	prcomp	stats	Supplementary variables are ignored
Correspondance Analysis	textmodel_ca	quanteda.textmodels	Only coordinates are available

Links: *https://github.com/juba/scatterD3*, *https://juba.github.io/explor/articles/introduction_en.html*

R Package *explor* by Julien Barnier

PCA | Eigenvalues | Variables plot | Variables data | Individuals plot | Individuals data

X axis
Axis 1 (32.72%)

Y axis
Axis 2 (17.37%)

Labels size
4 10 20
4 6 8 10 12 14 16 18 20

Minimum contribution to show label
0

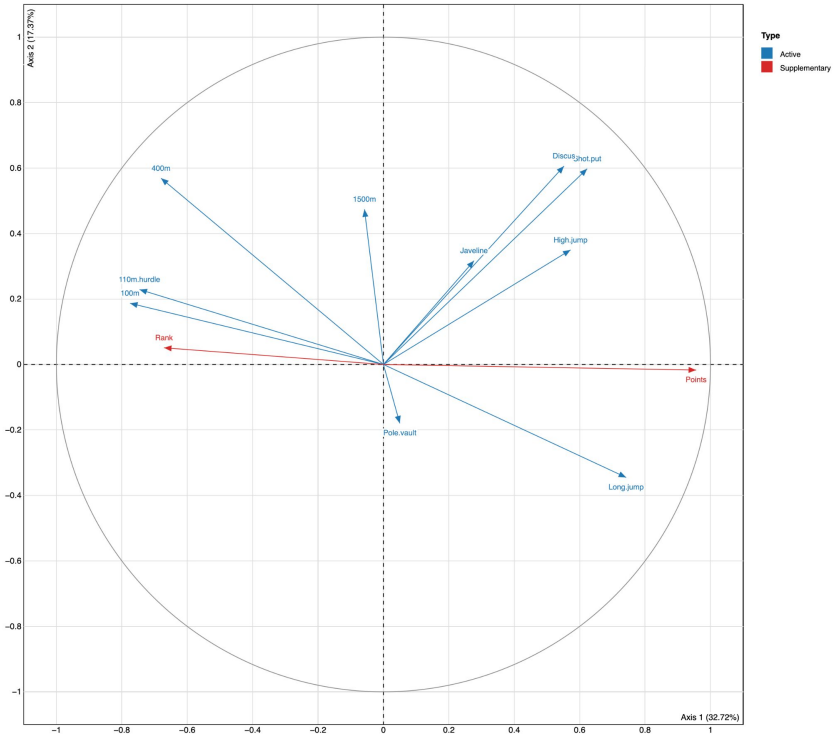
Points color :
Variable type

Supplementary variables

Supplementary variables to display
Rank
Points

Animations

Lasso selection
Get R code
Export as SVG



Links: https://github.com/juba/scatterD3, https://juba.github.io/explor/articles/introduction_en.html

R Package *explor* by Julien Barnier

PCA | Eigenvalues | Variables plot | **Variables data** | Individuals plot | Individuals data

Dimension

Axis 1 (32.72%)

Active variables

Show 10 entries — Search:

Variable	Coord	Contrib	Cos2	Cor
100m	-0.775	18.34	0.600	-0.775
110m.hurdle	-0.746	17.02	0.557	-0.746
Long.jump	0.742	16.82	0.550	0.742
400m	-0.680	14.12	0.462	-0.68
Shot.put	0.623	11.84	0.388	0.623
High.jump	0.572	10.00	0.327	0.572
Discus	0.552	9.33	0.305	0.552
Javeline	0.277	2.35	0.077	0.277
1500m	-0.058	0.10	0.003	-0.058
Pole.vault	0.050	0.08	0.003	0.05

Showing 1 to 10 of 10 entries — Previous 1 Next

Supplementary variables

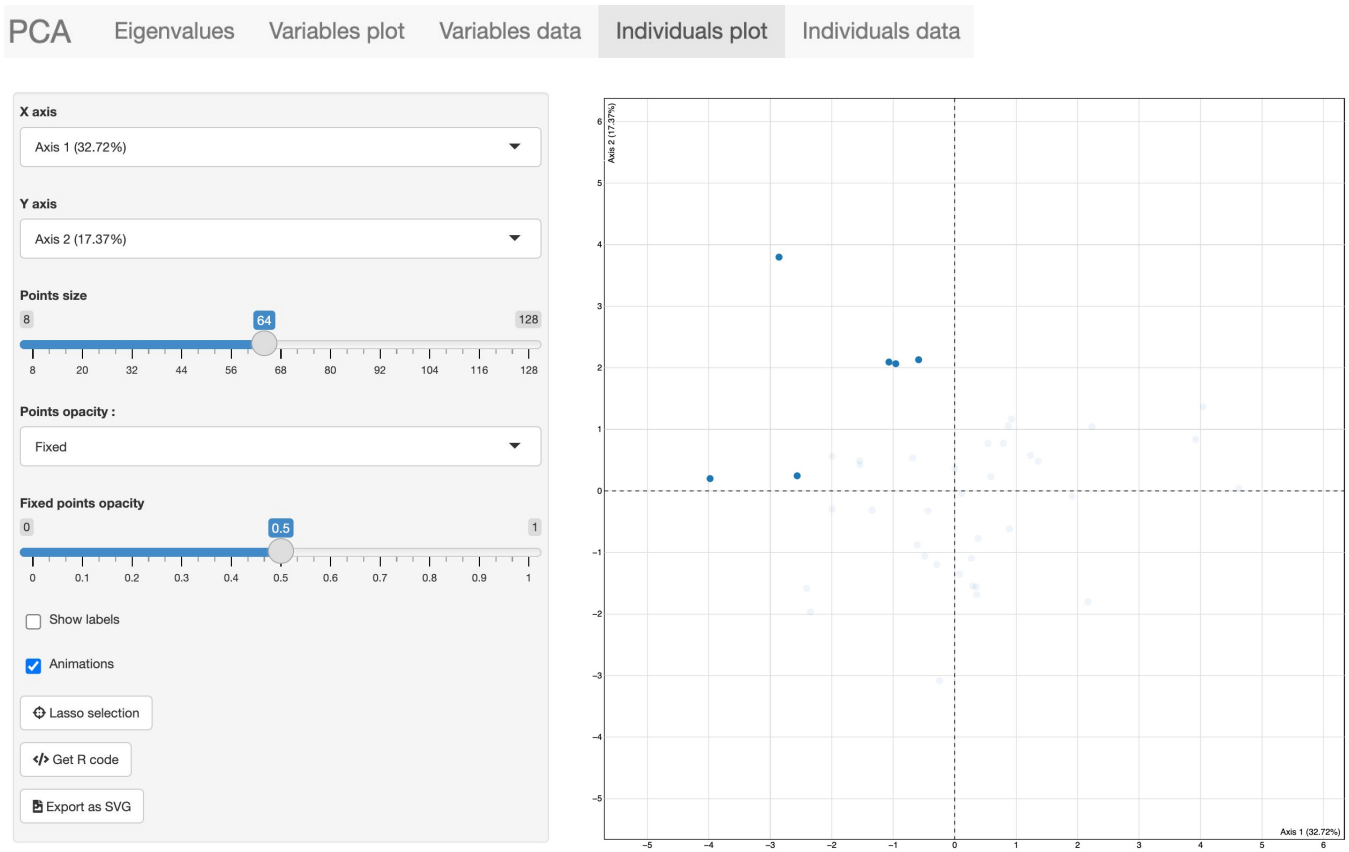
Show 10 entries — Search:

Variable	Coord	Cos2	Cor
Points	0.956	0.914	0.956
Rank	-0.671	0.450	-0.671

Showing 1 to 2 of 2 entries — Previous 1 Next

Links: https://github.com/juba/scatterD3, https://juba.github.io/explor/articles/introduction_en.html

R Package *explor* by Julien Barnier



Links: https://github.com/juba/scatterD3, https://juba.github.io/explor/articles/introduction_en.html

R Package *explor* by Julien Barnier

PCA | Eigenvalues | Variables plot | Variables data | Individuals plot | Individuals data

Dimension

Axis 1 (32.72%)

Active individuals

Show 20 entries Search:

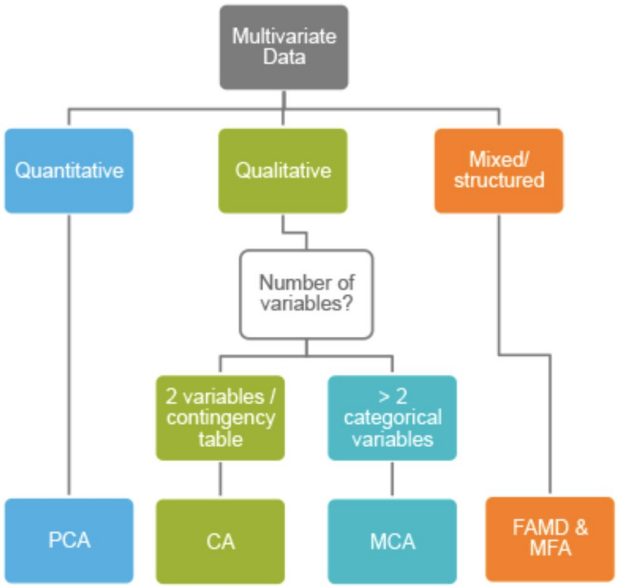
Name	Coord	Contrib	Cos2
Karpov	4.620	15.91	0.852
Sebrle	4.038	12.16	0.695
Clay	3.919	11.45	0.711
Macey	2.233	3.72	0.423
Warners	2.168	3.50	0.530
Bernard	1.906	2.71	0.455
KARPOV	1.358	1.38	0.160
CLAY	1.235	1.14	0.124
Zsivoczky	0.925	0.64	0.130
Hernu	0.889	0.59	0.238
Smith	0.870	0.56	0.061
SEBRLE	0.792	0.47	0.112
McMULLEN	0.588	0.26	0.053
Pogorelov	0.540	0.22	0.051
Ojaniemi	0.380	0.11	0.026
WARNERS	0.357	0.10	0.022
Averyanov	0.349	0.09	0.019
Nool	0.295	0.07	0.009
ZSIVOCZKY	0.272	0.06	0.011
Schoenbeck	0.114	0.01	0.004

Showing 1 to 20 of 41 entries Previous 1 2 3 Next

Links: https://github.com/juba/scatterD3, https://juba.github.io/explor/articles/introduction_en.html

Influence Analysis

Principal Component Methods



PCA: Principal Component Analysis
(M)CA: (Multiple) Correspondence Analysis
FAMD: Factor Analysis Mixed Data
MFA: Multiple Factor Analysis

Visualizing Multivariate Data Analysis Results

Sources:
http://www.sthda.com/english/articles/22-principal-component-methods-videos/
http://www.sthda.com/english/wiki/factoextra-r-package-easy-multivariate-data-analyses-and-elegant-visualization

Dimensionality reduction methods

Source: Lan Huong Nguyen, Susan Holmes: Ten quick tips for effective dimensionality reduction, PLOS 2019

Method	Input Data	Method Class	Nonlinear	Complexity
PCA	continuous data	unsupervised		$\mathbb{O}(max(n^2p, np^2))$
CA	categorical data	unsupervised		$\mathbb{O}(max(n^2p, np^2))$
MCA	categorical data	unsupervised		$\mathbb{O}(max(n^2p, np^2))$
PCoA (cMDS)	distance matrix	unsupervised		$\mathbb{O}(n^2p)$
NMDS	distance matrix	unsupervised		$\mathbb{O}(n^2h)$
Isomap	continuous*	unsupervised	✓	$\mathbb{O}(n^2(p + log\ n))$
Diffusion Map	continuous*	unsupervised	✓	$\mathbb{O}(n^2p)$
Kernel PCA	continuous*	unsupervised	✓	$\mathbb{O}(n^2p)$
t-SNE	continuous/distance	unsupervised	✓	$\mathbb{O}(n^2p + n^2h)$
Barnes–Hut t-SNE	continuous/distance	unsupervised	✓	$\mathbb{O}(nh\ log\ n)$
LDA	continuous (X and Y)	supervised		$\mathbb{O}(np^2 + p^3)$
PLS (NIPALS)	continuous (X and Y)	supervised		$\mathbb{O}(npd)$
NCA	distance matrix	supervised	✓	$\mathbb{O}(n^2h)$
Bottleneck NN	continuous/categorical	supervised	✓	$\mathbb{O}(nph)$
STATIS	continuous	multidomain		$\mathbb{O}(n^2P, nP^2)$
DiSTATIS	distance matrix	multidomain		$\mathbb{O}(n^2P, nP^2)$

Basic properties: input data required, method class, linear or nonlinear, and runtime complexity in terms of: n—the number of observations, p—the number of features in the original data, k—the selected number of nearest neighbors, h—the number of iterations, and P—the total number of variables in all available datasets collected on n samples in the case of multidomain data.

*Commonly, Isomap estimates geodesic distances between data points from Euclidean distances, and Diffusion Map and Kernel PCA compute Gaussian kernels and thus require continuous data input. However, it is possible to use categorical data if other dissimilarities or kernels are used.

Abbreviations: CA, correspondence analysis; cMDS, classical multidimensional scaling; LDA, linear discriminant analysis; MCA, multiple CA; NCA, neighborhood component analysis; NIPALS, nonlinear iterative partial least squares; NMDS, nonmetiric multidimensional scaling; NN, neural network; PCA, principal component analysis; PCoA, principal CA; t-SNE, t-Stochastic Neighbor Embedding; PLS, partial least squares

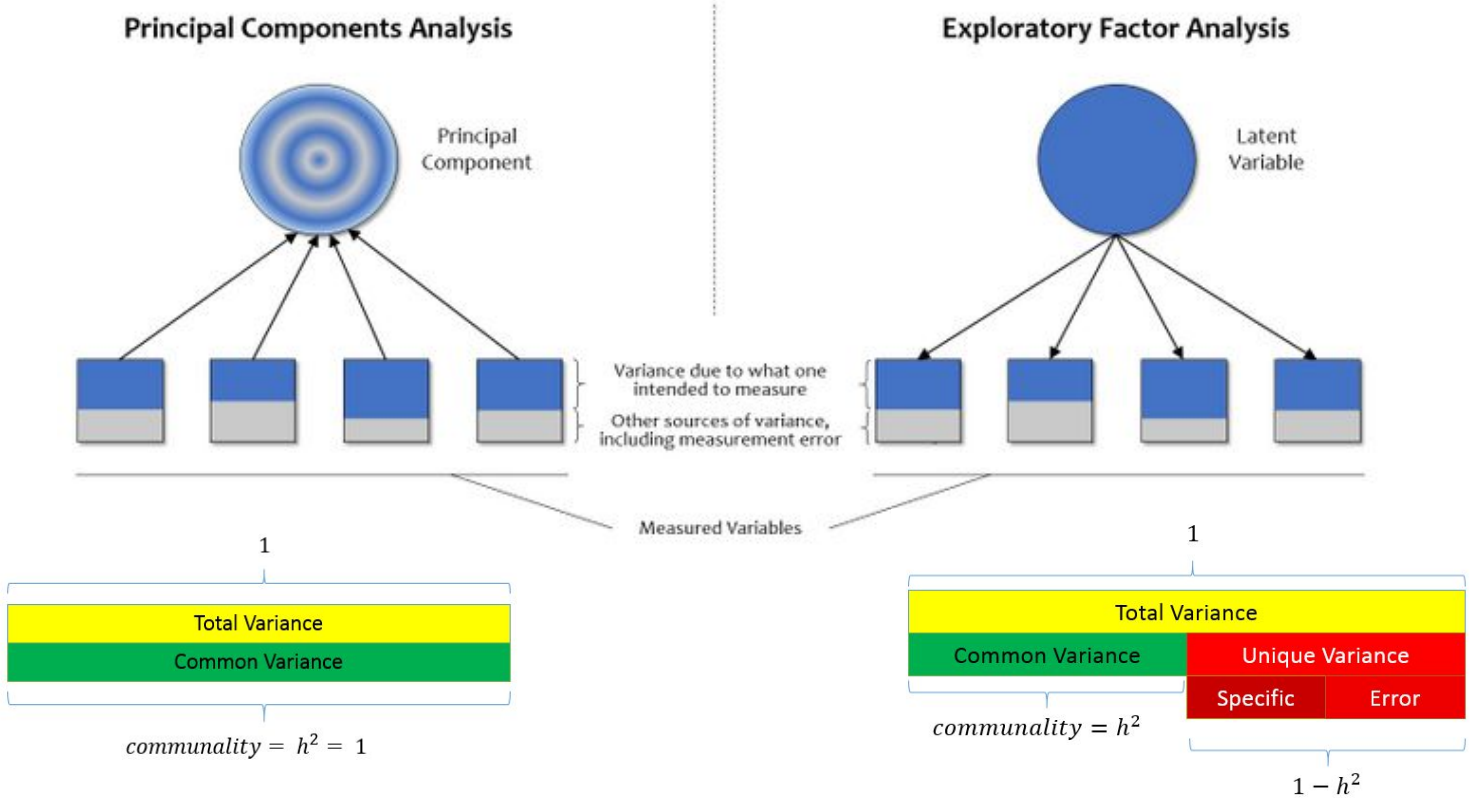
R and Python Implementations

Source: Lan Huong Nguyen, Susan Holmes: Ten quick tips for effective dimensionality reduction, PLOS 2019

Method	R function	Python function
PCA	`stats::prcomp`	`sklearn.decomposition.PCA`
CATPCA	`gifi::princals`	
CA	`FactoMineR::CA`	
MCA	`FactoMineR::MCA`	
PCoA (cMDS)	`stats::cmdscale`	`sklearn.manifold.MDS`
NMDS	`ecodist::nmds`	`sklearn.manifold.MDS`
Isomap	`vegan::isomap`	`sklearn.manifold.Isomap`
Diffusion Map	`diffusionMap::diffuse`	
(Barnes–Hut) t-SNE	`Rtsne::Rtsne`	`sklearn.manifold.TSNE`
LDA	`MASS::lda`	`sklearn.discriminant_analysis.LinearDiscriminantAnalysis`
PLS (NIPALS)	`mixOmics::pls`	`sklearn.cross_decomposition.PLSRegression`
DiSTATIS	`DistatisR::distatis`	
Procrustes	`vegan::procrustes`	`scipy.spatial.procrustes`

Software packages and function performing specified DR techniques available in R and python. R implementations are given as `package_name::function_name`; listed python functions come from `sklearn` and `scipy` libraries. The outputs of most linear DR methods can be visualized in R with `factoextra` package [25], used to generate a number of the plots in this article. Abbreviations: CA, correspondence analysis; CATPCA, categorical PCA; cMDS, classical multidimensional scaling; DR, dimensionality reduction; LDA, linear discriminant analysis; MCA, multiple CA; NIPALS, nonlinear iterative partial least squares; NMDS, nonmetiric multidimensional scaling; PCA, principal component analysis; PCoA, principal CA; t-SNE, t-Stochastic Neighbor Embedding; PLS, partial least squares

Principal Components or Factor Analysis?



Sources:
https://towardsdatascience.com/principal-components-or-factor-analysis-fcc98225b932
https://stats.idre.ucla.edu/spss/seminars/efa-spss/

Independent Component Analysis

As in PCA, we are looking for N different vectors onto which we can project our observations to give a set of N maximally independent signals (sources)

Instead of using variance as our independence measure (i.e. decorrelating) as we do in PCA, we use a measure of how statistically independent the sources are.

In large biological data sets, the loading vectors should only assign large weights to important variables.
That means the distribution of any loading vector should be **super-Gaussian**: most of the weights are very close to zero while only a few have large (absolute) values.

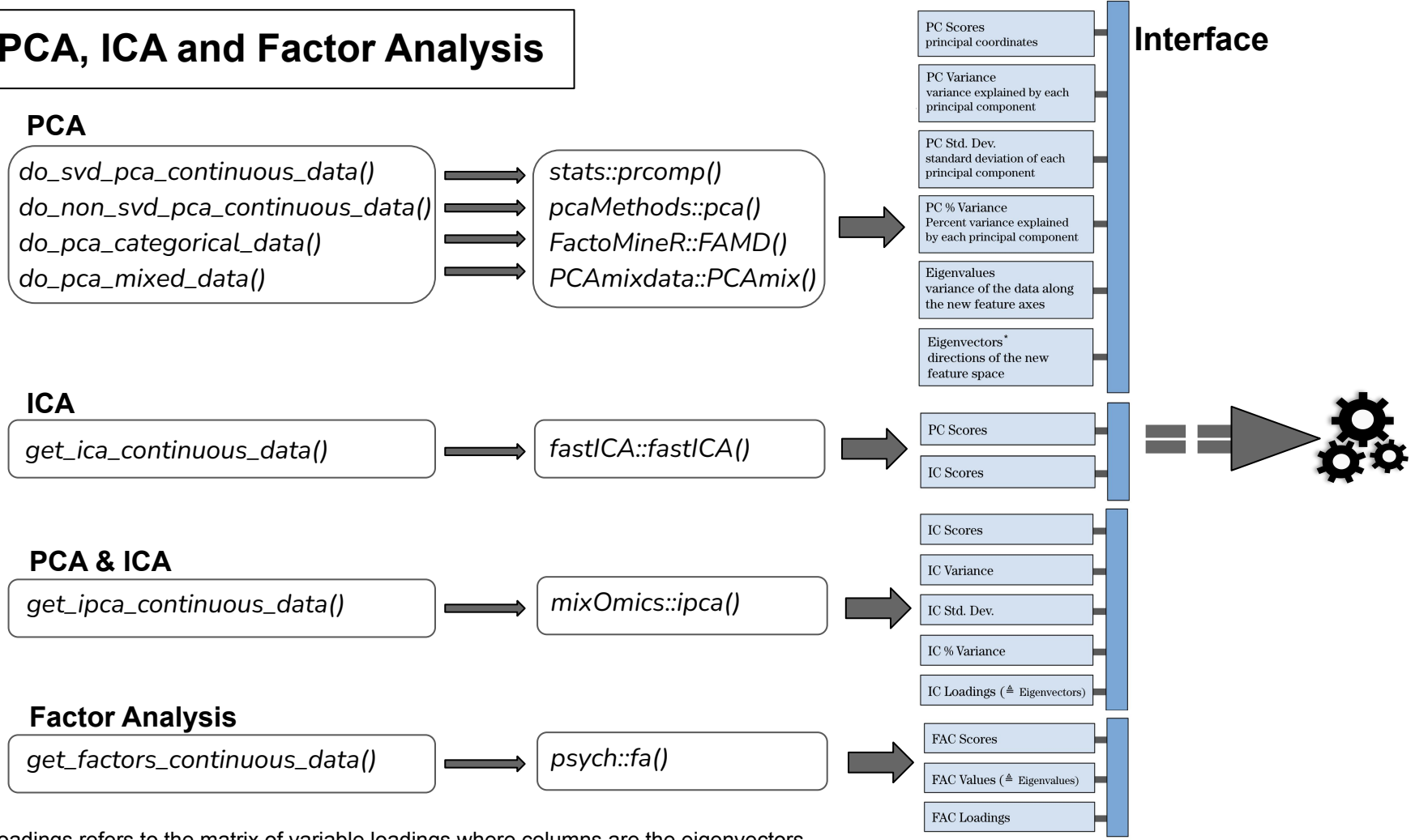
Due to the existence of noise, the distribution of any loading vector is distorted and tends toward a Gaussian distribution according to the Central Limit Theorem.

By maximizing the non-Gaussianity of the loading vectors using FastICA, we obtain more noiseless loading vectors.

We then project the original data matrix on these noiseless loading vectors, to obtain independent principal components, which should be also more noiseless and be able to better cluster the samples according to the biological treatment.

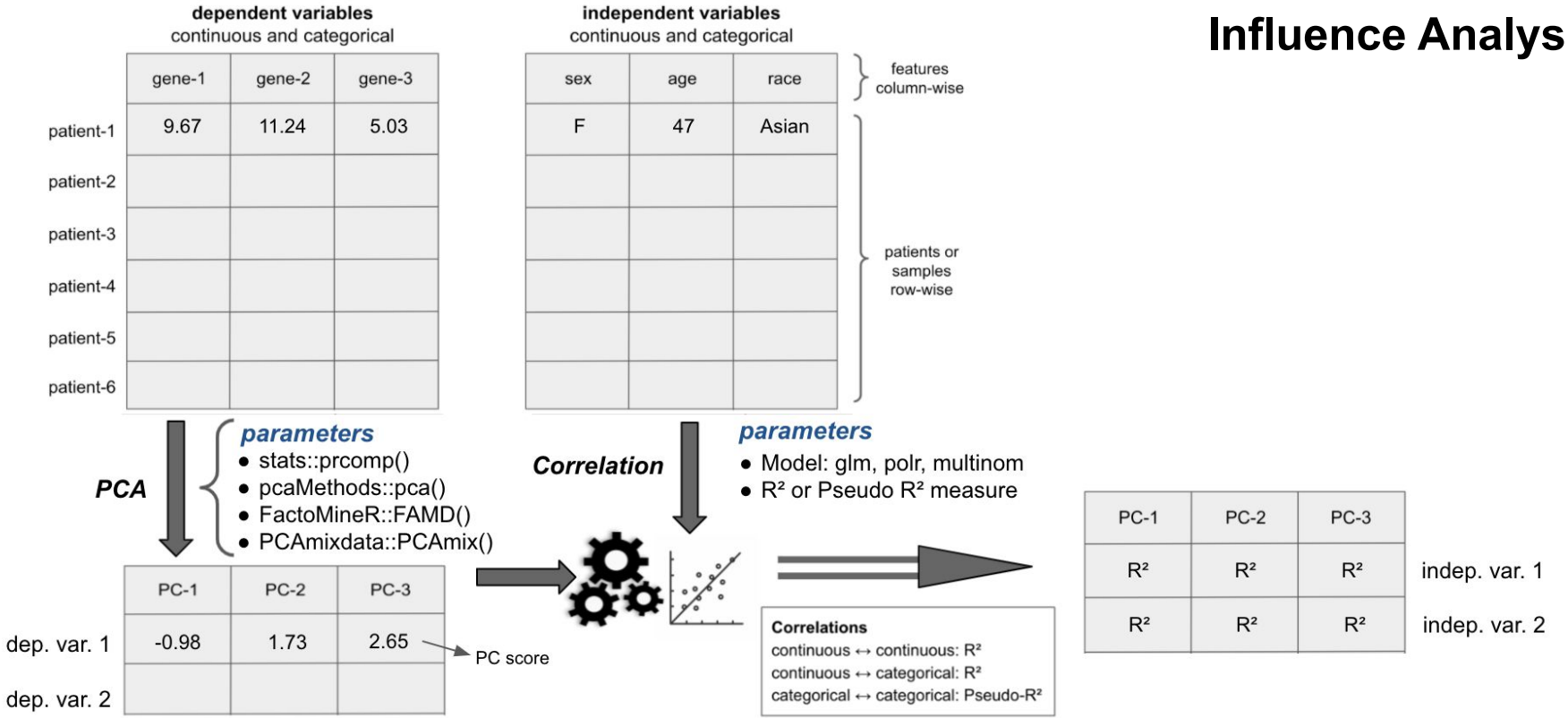
Sources:
http://www.mit.edu/~gari/teaching/6.555/SLIDES/BSShandouts.pdf
https://rdrr.io/cran/mixOmics/man/ipca.html

PCA, ICA and Factor Analysis



*Loadings refers to the matrix of variable loadings where columns are the eigenvectors.

Influence Analysis




```
do_variance_analysis(
  dependent_vars = dependent_vars,
  independent_vars = independent_vars,
  pc_scores = pc_scores,
  rsq_threshold_pct = 25)
```

Pseudo R² Measures implemented in the R Package *DescTools*

DescTools::PseudoR2(x, which = NULL)

x: the *glm*, *polr* or *multinom* model object to be evaluated.

which: one of the pseudo R2 methods listed below as character.

`McFadden`	McFadden pseudo-*R^2*
`McFaddenAdj`	McFadden adjusted pseudo-*R^2*
`CoxSnell`	Cox and Snell pseudo-*R^2* (also known as ML pseudo-*R^2*)
`Nagelkerke`	Nagelkerke pseudo*R^2* (also known as CraggUhler *R^2*)
`AldrichNelson`	AldrichNelson pseudo-*R^2*
`VeallZimmermann`	VeallZimmermann pseudo-*R^2*
`McKelveyZavoina`	McKelvey and Zavoina pseudo-*R^2*
`Efron`	Efron pseudo-*R^2*
`Tjur`	Tjur's pseudo-*R^2*
`AIC`	Akaike's information criterion
`LogLik`	log-Likelihood for the fitted model (by maximum likelihood)
`LogLikNull`	log-Likelihood for the null model. The null model will include the offset, and an intercept if there is one in the model.
`G2`	differenz of the null deviance - model deviance

Source: https://rdrr.io/cran/DescTools/man/PseudoR2.html

Influence Analysis: Example Workflow in R

```
pca_data <- get_pca_mixed_data(data = dependent_vars, npc = 10)

pc_scores <- get_principle_components(pca_data = pca_data, npc = 10)

cf <- do_variance_analysis(dependent_vars = dependent_vars,
                           independent_vars = independent_vars,
                           pc_scores = pc_scores,
                           rsq_threshold_pct = 25)
rsq_df = cf$rsq_df   # Dataframe: rows: independent vars, columns: PCs from dependent vars, values: R²
top_explanatory_vars_df = cf$top_explanatory_vars_df  # Sorted variables according to R²

create_pc_heatmap(rsq_df, 10, 'PC') # Heatmap of R² matrix

create_paired_pcs_plot(...) # Scatterplot matrix
create_vars_pcs_plot(...) # Violin plot

cf_rsq_df <- do_corr_analysis(independent_vars = independent_vars,
                              cor_vars = rownames(top_explanatory_vars_df),
                              model = 'glm',
                              method = 'McFadden')

create_feature_corr_matrix(as.matrix(cf_rsq_df)) # Pairwise correlations of independent variables
```

PCA Methods implemented in the R Package *pcaMethods*

- **SVD:** a fast method which is also the standard method in R but which is not applicable for data with missing values.
- **NIPALS:** an iterative fast method which is applicable also to data with missing values.
- **RNIPALS:** PCA by non-linear iterative partial least squares.
- **PPCA:** Probabilistic PCA which is applicable also on data with missing values. Missing value estimation is typically better than NIPALS but also slower to compute and uses more memory.
- **BPCA:** Bayesian PCA which performs very well in the presence of missing values but is slower than PPCA. A port of the matlab implementation by Shigeyuki Oba.
- **robustPca:** PCA implementation based on robustSVD and robust to outliers in a data set. It can also handle missing values, but not intended to be used for missing value estimation.
- **NLPCA:** Non-linear PCA which can find curves in data and in presence of such can perform accurate missing value estimation.

Source: https://www.bioconductor.org/packages/release/bioc/manuals/pcaMethods/man/pcaMethods.pdf

FactoMineR::FAMD()

- Factor analysis of mixed data (FAMD) is a principal component method dedicated to analyze a data set containing both quantitative and qualitative variables (Pagès 2004).
- The FAMD algorithm acts as PCA (principal component analysis) quantitative variables and as MCA (multiple correspondence analysis) for qualitative variables.
- MCA represents data as points in a low-dimensional Euclidean space and can be seen as the counterpart of principal component analysis for categorical data.
- Quantitative and qualitative variables are normalized during the analysis in order to balance the influence of each set of variables to determine the dimensions of variability.
- The continuous variables are scaled to unit variance and the categorical variables are transformed into a disjunctive data table (crisp coding) and then scaled using the specific scaling of MCA.

Sources:

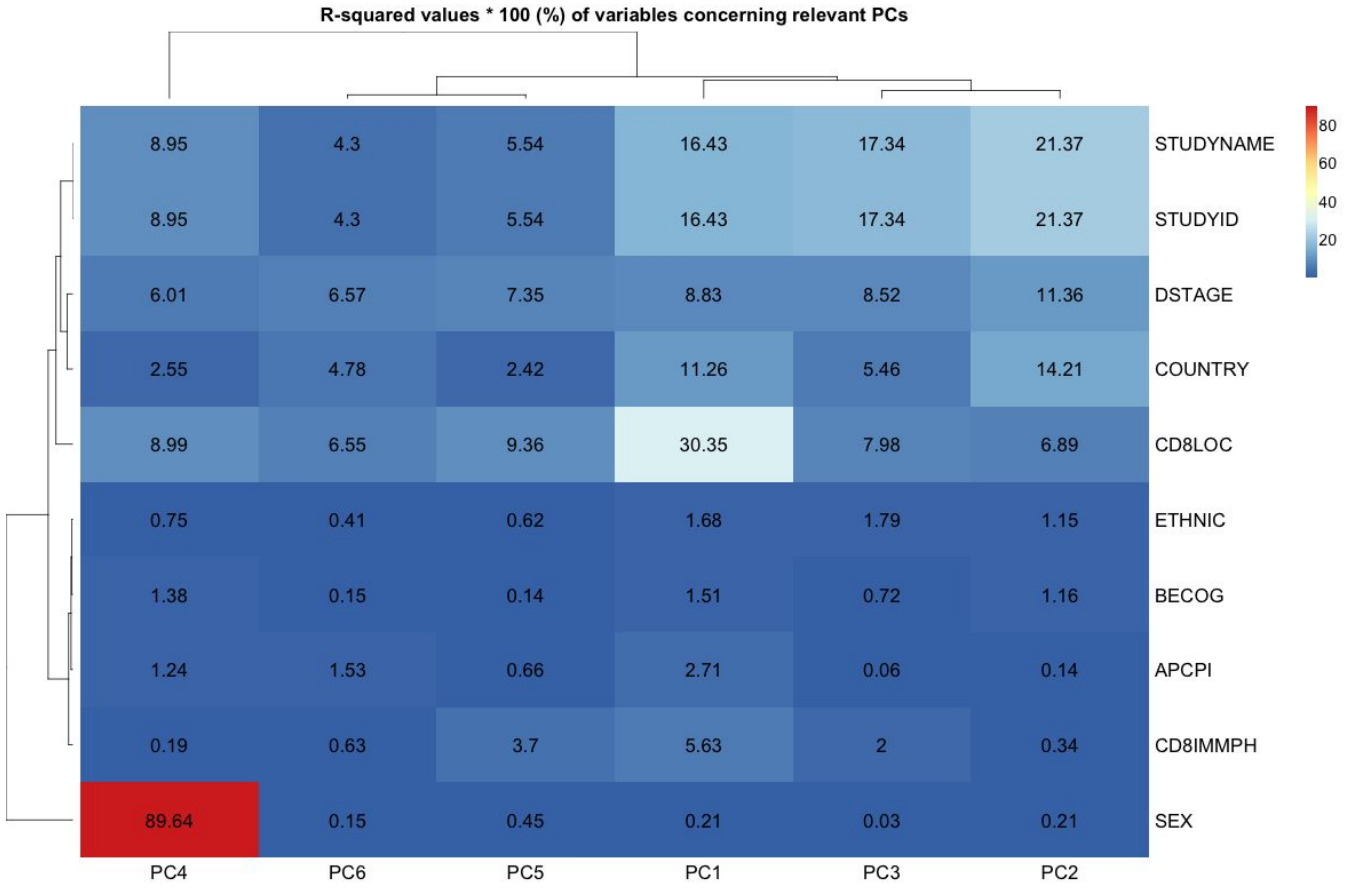
Pagès, J. 2004. “Analyse Factorielle de Donnees Mixtes.” Revue Statistique Appliquee 4: 93–111.
F. Bertrand et al.Using Factor Analyses to explore data generated by the National Grapevine Wood Diseases Survey, 2007

PCAmixdata::PCAmix()

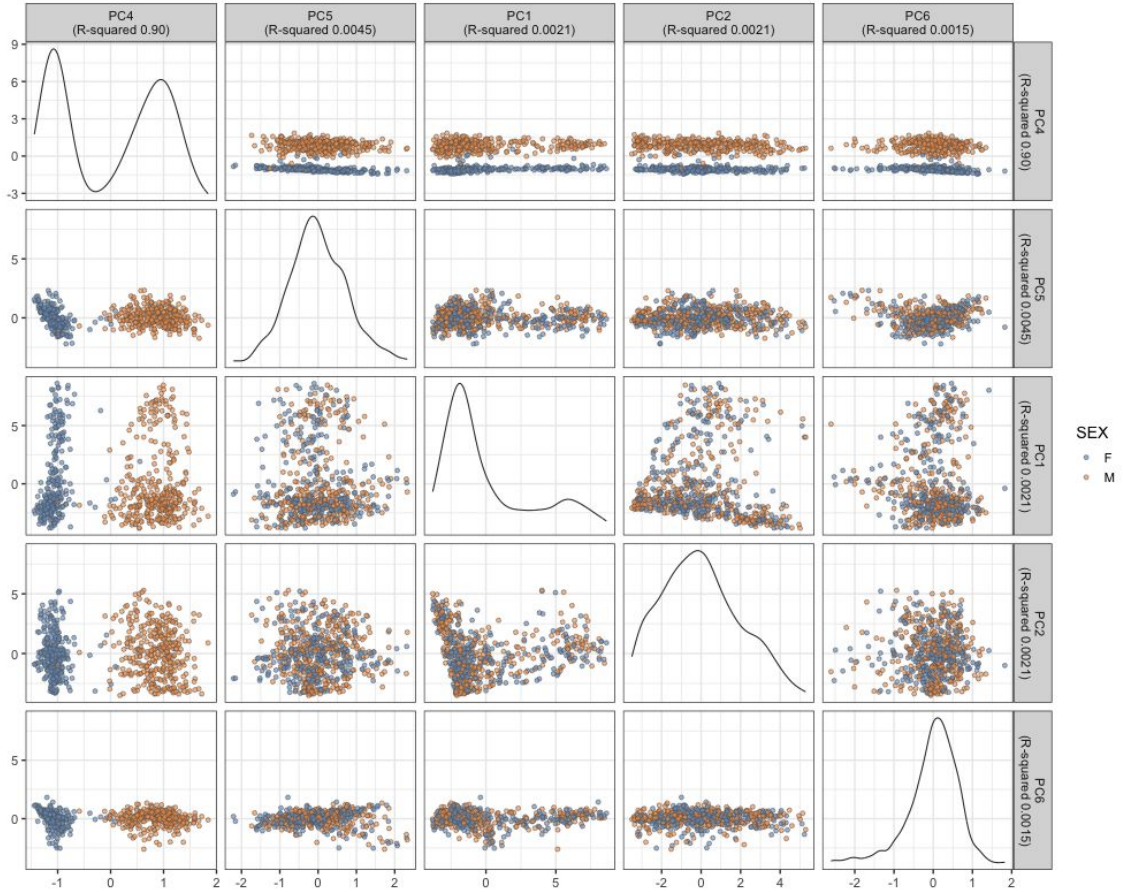
- Performs principal component analysis of a set of observations described by a mixture of qualitative and quantitative variables.
- For quantitative variables a standard PCA is performed.
- For qualitative variables s standard MCA is performed.
- Missing values are replaced by means for quantitative variables and by zeros in the indicator matrix for qualitative variables.
- PCAmix performs squared loadings.
    - Squared loadings for a qualitative variable are correlation ratios between the variable and the principal components.
    - For a quantitative variable, squared loadings are the squared correlations between the variable and the principal components.

Sources:
https://cran.r-project.org/web/packages/PCAmixdata/vignettes/PCAmixdata.html
https://rdrr.io/cran/PCAmixdata/man/PCAmix.html

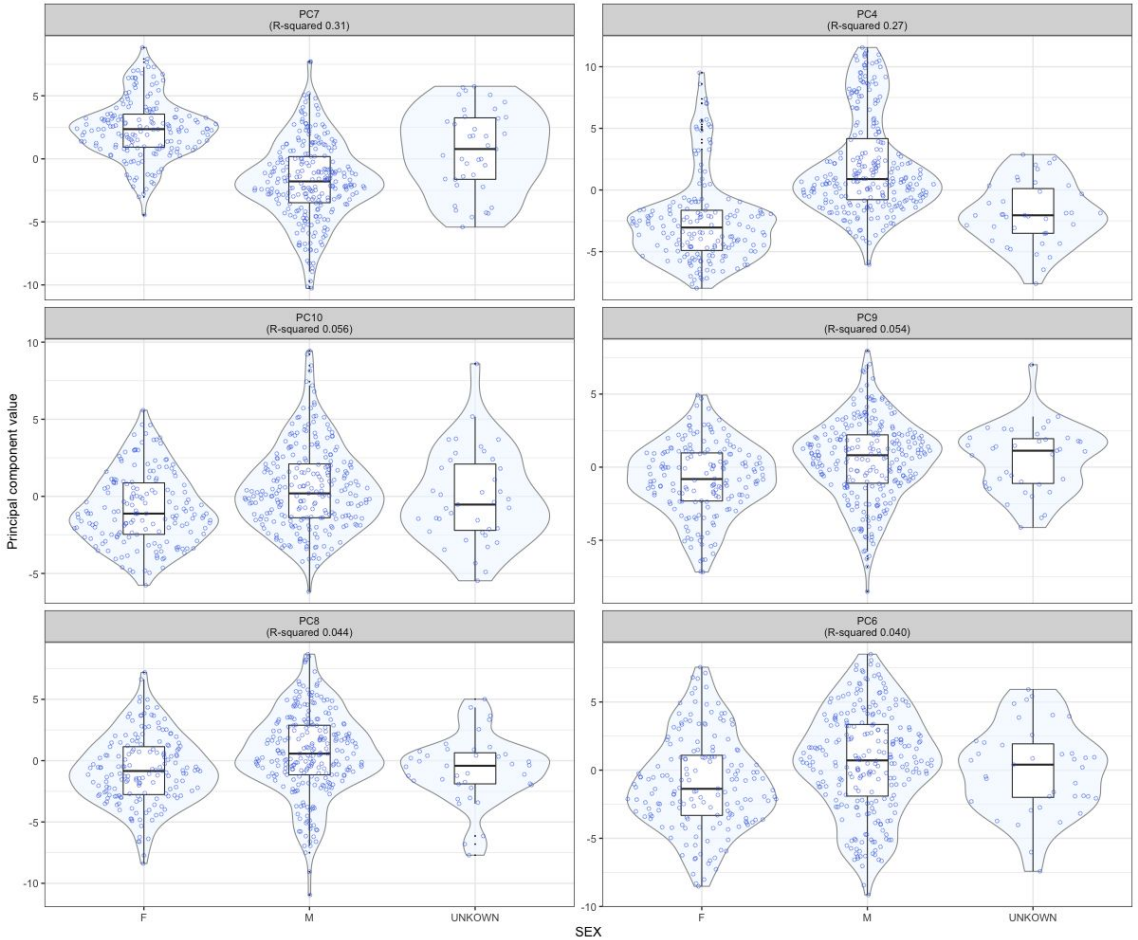
Heatmap of R^2 Matrix showing Correlations of PC scores and Independent Variables



Scatterplot Matrix of PC scores and Independent Variables

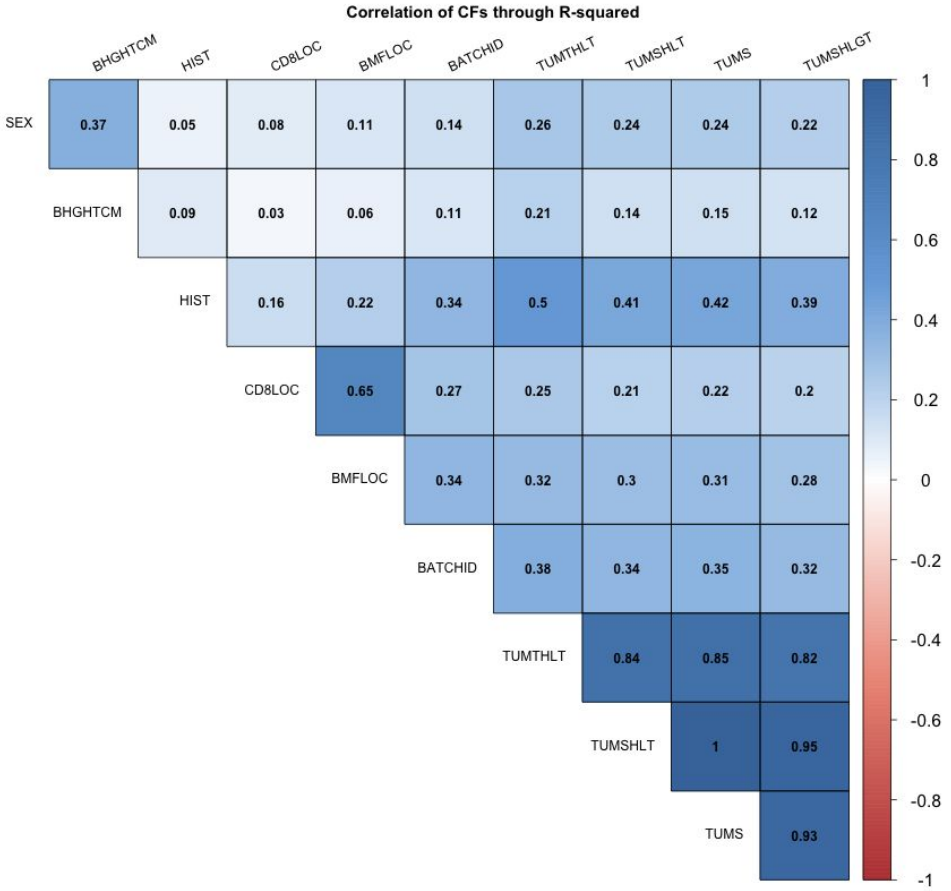


Violin Plot of PC scores and Independent Variables



using ggforce::geom_sina()

R^2 Correlation Matrix comparing pairwise Independent Variables

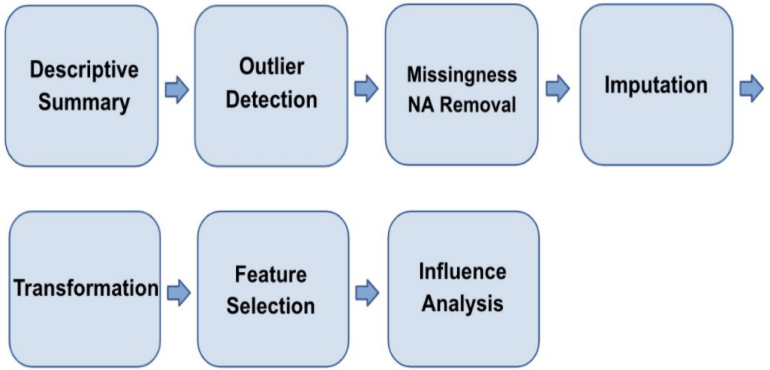


PCA based Influence Analysis

Package	Method	Description
stats	prcomp()	PCA analysis
pcaMethods	pca()	Different PCA methods
FactoMineR	FAMD()	Factor analysis for mixed data
PCAmixdata	PCAmix()	PCA analysis for mixed data
fastICA	fastICA()	ICA analysis
mixOmics	ipca()	Independant Principal Component Analysis
psych	fa()	Exploratory factor analysis

EDA Phase	Package	Method
Descriptive Summary	EDAWB	get_summary_stats(data)
	DataExplorer	introduce(data)
Outlier Detection	EDAWB	get_outlier_summary(continuous_data)
	EDAWB	get_outliers_by_performance_check_outliers(continuous_data)
Missingness	naniar	miss_var_summary(data, order=TRUE, add_cumsum=TRUE)
	EDAWB	get_pct_bins_of_nas_for_rows_and_columns(data, get_standard_pct_missing_bins())
	visdat	vis_miss(data, warn_large_data=FALSE, sort_miss = TRUE, cluster = TRUE)
	DataExplorer	plot_missing(data)
NA Removal	EDAWB	remove_nas_by_using_only_rows(data, threshold)
	EDAWB	remove_nas_by_using_only_columns(data, threshold)
	EDAWB	run_recursive_na_removal_using_columns_and_rows()
Imputation	EDAWB	impute_by_column_mean()
	EDAWB	impute_by_column_median()
	EDAWB	impute_by_column_mode()
	mice	mice()
	VIM	kNN()
	VIM	hotdeck()
Transformation	EDAWB	do_log2_transformation()
	EDAWB	do_zscore_transformation()
	EDAWB	do_minmax_transformation()
	EDAWB	do_softmax_transformation()
	EDAWB	discretize_by_median()
	EDAWB	discretize_by_quartiles()
	EDAWB	discretize_by_deciles()
	arules	discretize()
	varrank	discretization()
Feature Selection	EDAWB	select_top_features_by_variance()
	EDAWB	select_top_features_by_mad()
	EDAWB	select_top_features_by_entropy()
	EDAWB	select_top_features_by_mutual_information()
	EDAWB	select_top_features_by_correlation()
	EDAWB	select_top_features_by_laplacian_score()
Influence Analysis	stats	prcomp()
	pcaMethods	pca()
	FactoMineR	FAMD()
	PCAmixdata	PCAmix()
	fastICA	fastICA()
	mixOmics	ipca()
	psych	fa()

Overview of all Methods



Appendix

Coloring particular cells in a table depending on a condition
https://vincentarelbundock.github.io/modelsummary/articles/datasummary.html

```
library(gt)

datasummary(All(mtcars) ~ Mean + SD,
            data = mtcars,
            fmt = NULL,
            output = 'gt') %>%
    tab_style(style = cell_fill(color = "pink"),
              locations = cells_body(rows = Mean > 10, columns = 2))
```

	Mean	SD
mpg	20.090625	6.0269481
cyl	6.187500	1.7859216
disp	230.721875	123.9386938
hp	146.687500	68.5628685
drat	3.596563	0.5346787
wt	3.217250	0.9784574
qsec	17.848750	1.7869432
vs	0.437500	0.5040161
am	0.406250	0.4989909
gear	3.687500	0.7378041
carb	2.812500	1.6152000

Cheat Sheet – PCA Dimensionality Reduction

What is PCA?

- Based on the dataset find a new set of orthogonal feature vectors in such a way that the data spread is maximum in the direction of the feature vector (or dimension)
- Rates the feature vector in the decreasing order of data spread (or variance)
- The datapoints have maximum variance in the first feature vector, and minimum variance in the last feature vector
- The variance of the datapoints in the direction of feature vector can be termed as a measure of information in that direction.

Steps

1. Standardize the datapoints $X_{new} = \frac{X - mean(X)}{std(X)}$
2. Find the covariance matrix from the given datapoints $C[i,j] = cov(x_i, x_j)$
3. Carry out eigen-value decomposition of the covariance matrix $C = V\Sigma V^{-1}$
4. Sort the eigenvalues and eigenvectors $\Sigma_{sort} = sort(\Sigma)$ $V_{sort} = sort(V, \Sigma_{sort})$

Dimensionality Reduction with PCA

- Keep the first m out of n feature vectors rated by PCA. These m vectors will be the best m vectors preserving the maximum information that could have been preserved with m vectors on the given dataset

Steps:

1. Carry out steps 1-4 from above
2. Keep first m feature vectors from the sorted eigenvector matrix $V_{reduced} = V[:, 0:m]$
3. Transform the data for the new basis (feature vectors) $X_{reduced} = X_{new} \times V_{reduced}$
4. The importance of the feature vector is proportional to the magnitude of the eigen value

Figure 1

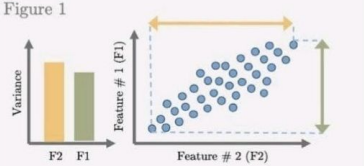


Figure 2

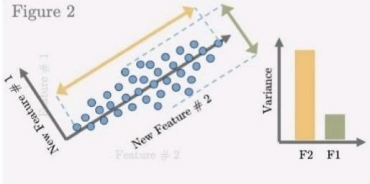


Figure 3

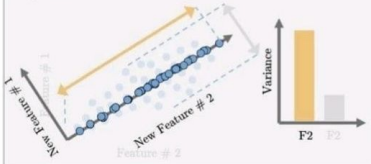
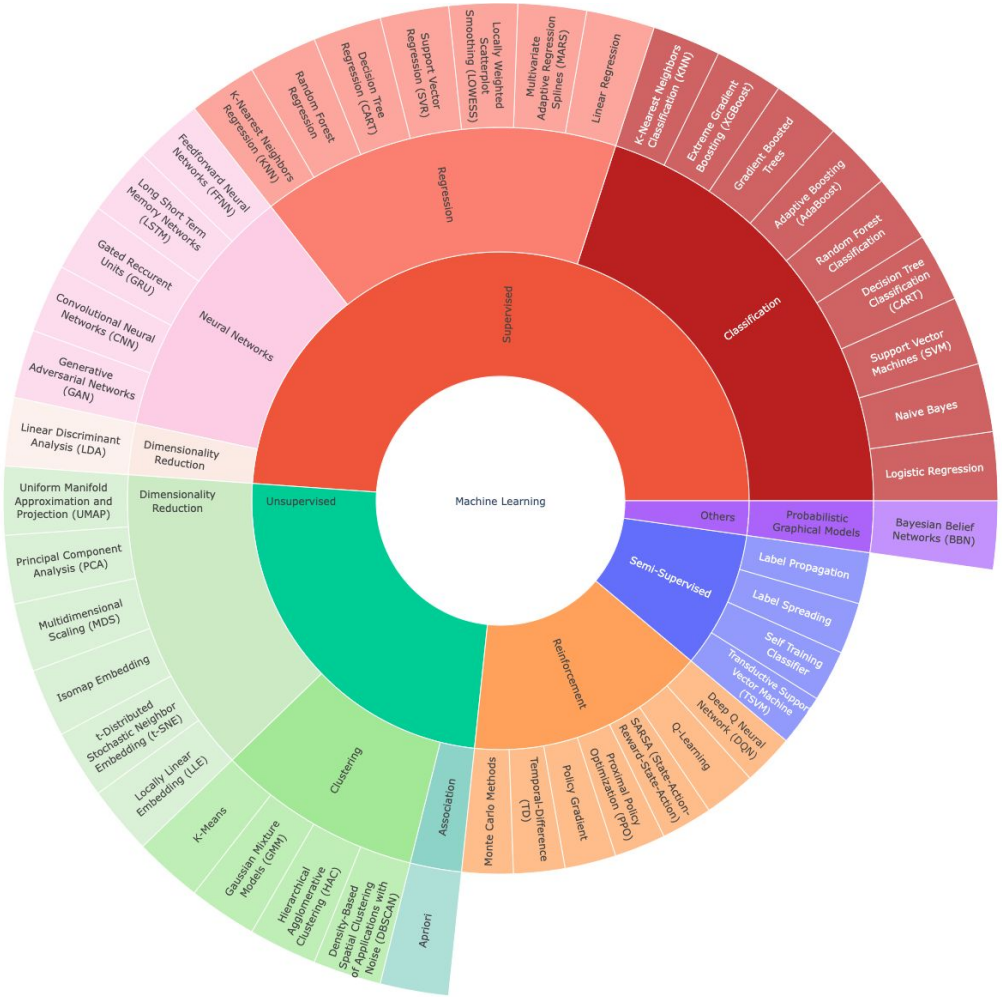


Figure 1: Datapoints with feature vectors as x and y-axis
Figure 2: The cartesian coordinate system is rotated to maximize the standard deviation along any one axis (new feature # 2)
Figure 3: Remove the feature vector with minimum standard deviation of datapoints (new feature # 1) and project the data on new feature # 2

Source: https://twitter.com/datasciencedojo/status/1469776002796724230?s=27



Source: Self-Training Classifier: How to Make Any Algorithm Behave Like a Semi-Supervised One
https://towardsdatascience.com/self-training-classifier-how-to-make-any-algorithm-behave-like-a-semi-supervised-one-2958e7b54ab7

R package sjmisc: *https://strengejacke.github.io/sjmisc/reference/empty_cols.html*

Return or remove variables or observations that are completely missing

Source: `R/is_empty.R`

These functions check which rows or columns of a data frame completely contain missing values, i.e. which observations or variables completely have missing values, and either 1) returns their indices; or 2) removes them from the data frame.

```
empty_cols(x)

empty_rows(x)

remove_empty_cols(x)

remove_empty_rows(x)
```

Check: https://strengejacke.github.io/sjmisc/articles/exploringdatasets.html