example_data

Autogenerated data summary from dataMaid

2019-03-15 15:44:47

Part 1

Data report overview

The dataset examined has the following dimensions:

Feature	Result
Number of observations	1000
Number of variables	9

Checks performed

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

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

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

Part 2

Summary table

	Variable class	# unique values	Missing observations	Any problems?
ID	character	1000	0.00 %	×
Race	factor	8	10.70 %	×
Age	character	17	12.20 %	×
Sex	factor	2	0.00 %	
Height_cm_	numeric	365	0.00 %	×
IQ	numeric	58	10.20 %	×
Smokes	logical	2	0.00 %	
Income	factor	901	10.00 %	×
Died	logical	2	0.00 %	

Part 3

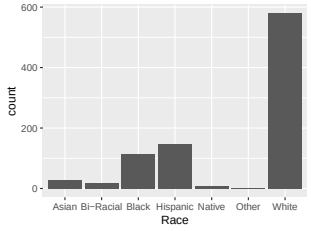
Variable list

ID

- The variable is a key (distinct values for each observation).

Race

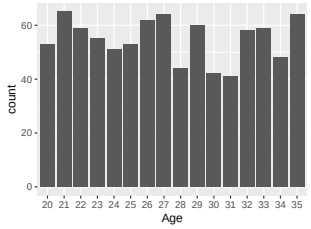
Feature	Result
Variable type	factor
Number of missing obs.	107 (10.7 %)
Number of unique values	7
Mode	“White”
Reference category	White



- Note that the following levels have at most five observations: "Other".

Age

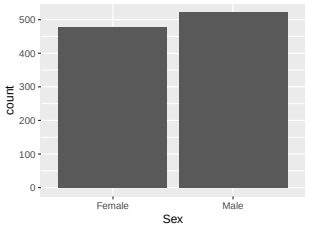
Feature	Result
Variable type	character
Number of missing obs.	122 (12.2 %)
Number of unique values	16
Mode	“21”



- Note: The variable consists exclusively of numbers and takes a lot of different values. Is it perhaps a misclassified numeric variable?

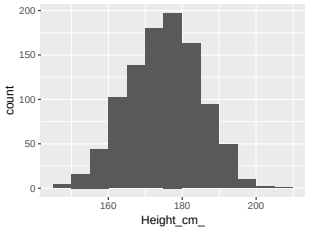
Sex

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



Height_cm_

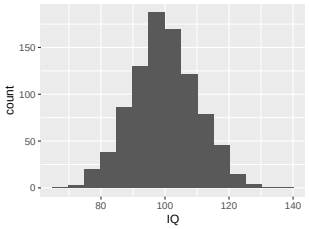
Feature	Result
Variable type	numeric
Number of missing obs.	0 (0 %)
Number of unique values	365
Median	175.3
1st and 3rd quartiles	168.2; 182.03
Min. and max.	146.3; 207.2



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

IQ

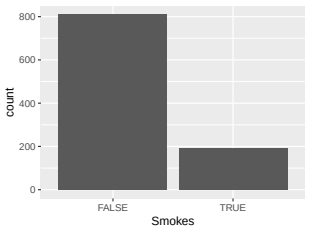
Feature	Result
Variable type	numeric
Number of missing obs.	102 (10.2 %)
Number of unique values	57
Median	100
1st and 3rd quartiles	93; 107
Min. and max.	68; 137



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

Smokes

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



Income

Feature	Result
Variable type	factor
Number of missing obs.	100 (10 %)
Number of unique values	900
Mode	“592.09”
Reference category	330.89

- Note that the following levels have at most five observations: "10093.41", "101416.01", "101711.6", "101922.46", "10209.53", "10309.37", "10360.78", "10368.95", "104430.62", "10611.98" (890 additional values omitted).
- Note: The variable consists exclusively of numbers and takes a lot of different values. Is it perhaps a misclassified numeric variable?

Died

Feature	Result
Variable type	logical
Number of missing obs.	0 (0 %)
Number of unique values	2
Mode	“TRUE”

Report generation information:

- Created by Mateusz Staniak (username: `mstaniak`).
- Report creation time: pt. mar 15 2019 15:44:48
- Report was run from directory: `C:/Users/mstaniak/Projekty/MI2DataLab/autoEDA-resources/autoEDA-paper`
- dataMaid v1.2.0 [Pkg: 2018-10-03 from CRAN (R 3.5.2)]
- R version 3.5.1 (2018-07-02).
- Platform: x86_64-w64-mingw32/x64 (64-bit)(Windows >= 8 x64 (build 9200)).
- Function call: `dataMaid::makeDataReport(data = example_data, output = "pdf", file = "./plots/dataMaid/dataMaid_report.pdf")`