Case Studies: Health Data

Descriptive Data Analysis

Muhammad Raafey Tariq Farrukh Ahmed Amirreza Khamehchin Khiabani Aymane Hachcham

Technische Universität Dortmund

April 12, 2023

Presentation Overview

Dataset description

- Dataset has 16386 data points
- Total number of variables is 18
- Table 1 categorizes the given variables into different attribute types:

Dataset description

Variable	Type
id	Ordinal/Discrete
zeit	Ordinal/Discrete
terminal	Cardinal/Discrete
postleitzahl	Cardinal/Discrete
gemeinde	Nominal/Discrete
bezirk	Nominal/Discrete
Bundesland	Nominal/Discrete
befinden	Ordinal/Discrete
geburtsjahr	Cardinal/Discrete
gesclecht	Nominal/Discrete
raucher	Nominal/Discrete
Blutzucker_bekannt	Nominal/Discrete
cholesterin_bekannt	Nominal/Discrete
in_behandlung	Nominal/Discrete
schaetzwert_bp_sys	Cardinal/Continuous
schaetzwert_by_dia	Cardinal/Continuous
messwert_bp_sys	Cardinal/Continuous
messwert_bp_dia	Cardinal/Continuous

Table: A categorization of variables into different attribute types

Missing Values

- In total 382 rows have missing values
- Table 2 summarizes the number of missing values per variable:

Variable	Nr. Missing Values
postleitzahl	334
gemeinde	331
bezirk	331
Bundesland	331
befinden	23
geburtsjahr	23
gesclecht	23
schaetzwert_bp_sys	45
schaetzwert_by_dia	56

Table: Missing values per variable

Missing Values

- About 331 rows were records belonging to foreigners
- 3 Local residents also had missing *postleitzahl*
- 22 records with missing values for *befinden*, *geburtsjahr* and *gesclecht* belonged to foreigners and 1 belonged to a local resident
- All 45 records with missing values for *schaetzwert_bp_sys* also had missing values for *schaetzwert_by_dia* but not vise versa

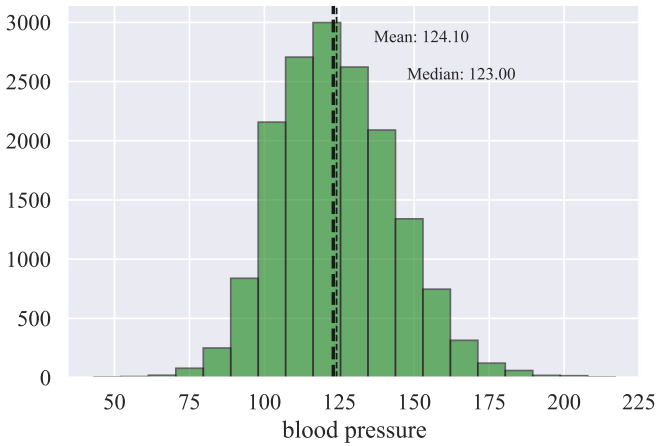
Statistical Measures

	schaetzwert_bp_sys	schaetzwert_by_dia	messwert_bp_sys	messwert_bp_dia
Count	16341	16330	16386	16386
Mean	122.41	79.86	124.10	82.04
Min	34.00	30.00	43.00	27.00
25%	115.00	75.00	110.00	73.00
50%	120.00	80.00	123.00	81.00
75%	130.00	85.00	137.00	90.00
Max	299.00	212.00	217.00	197.00
Std	16.94	9.96	19.68	14.64

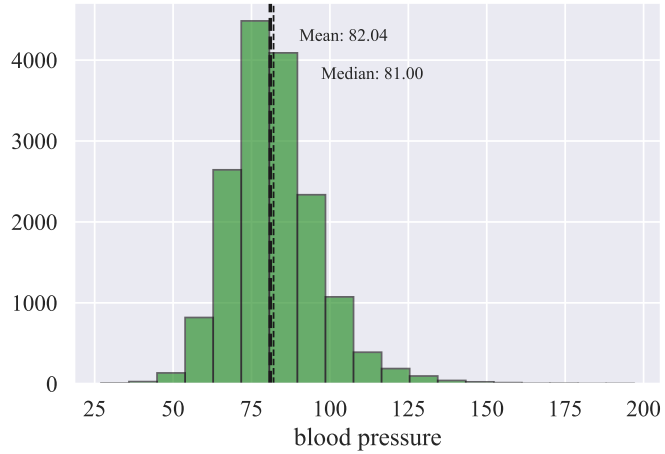
Table: A summary of statistical measures for continuous variables

Plots

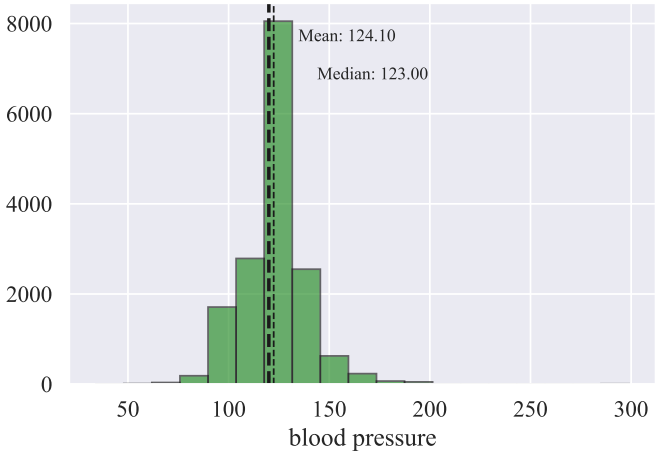
Frequency Distributions



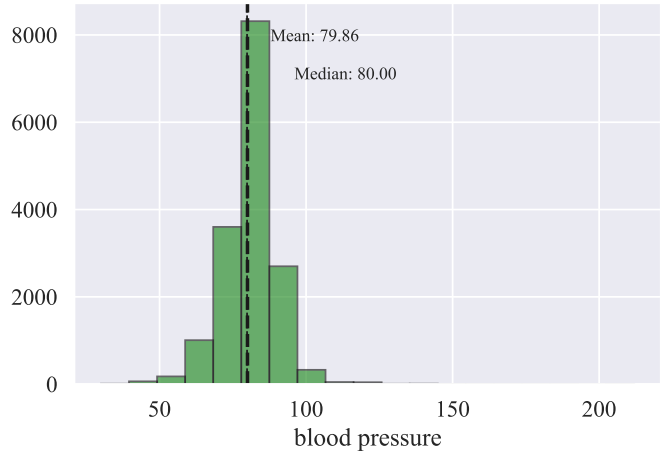
(a) Measured systolic



(b) Measured Diastolic



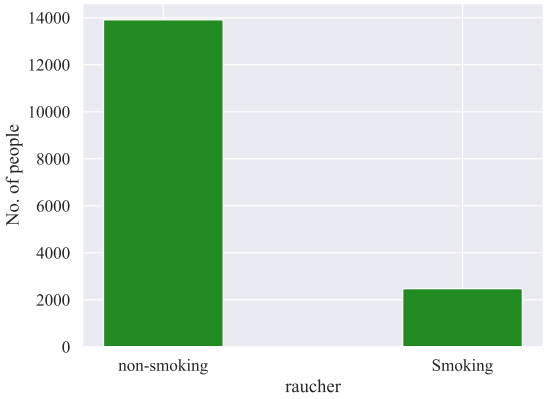
(c) Self-estimated systolic



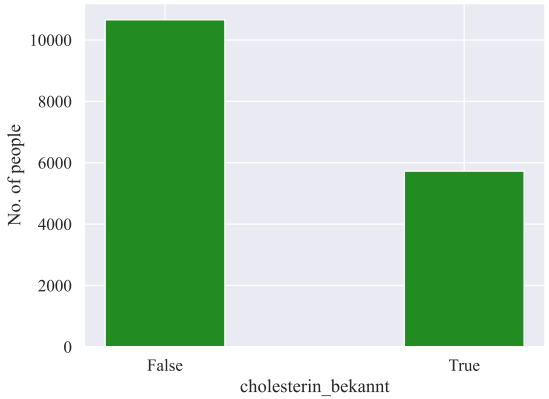
(d) Self-estimated diastolic

Plots

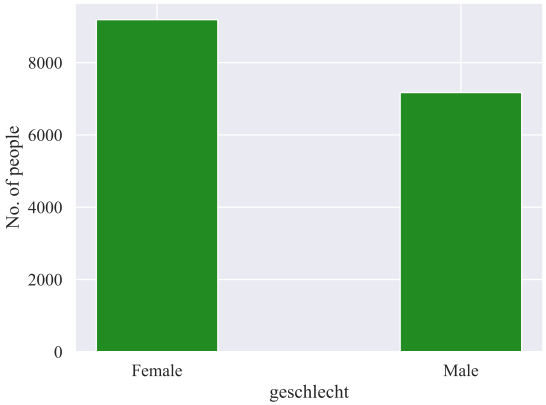
Bar plots



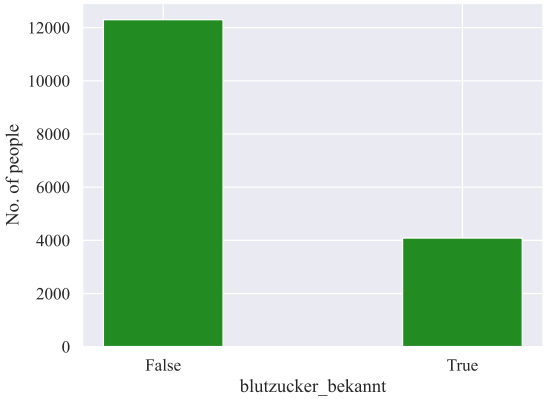
(a) Smokers



(b) Cholesterol patients



(c) Genders



(d) Diabetic patients

Figure: Bar plots for a few key discrete variables

Plots

Correlation

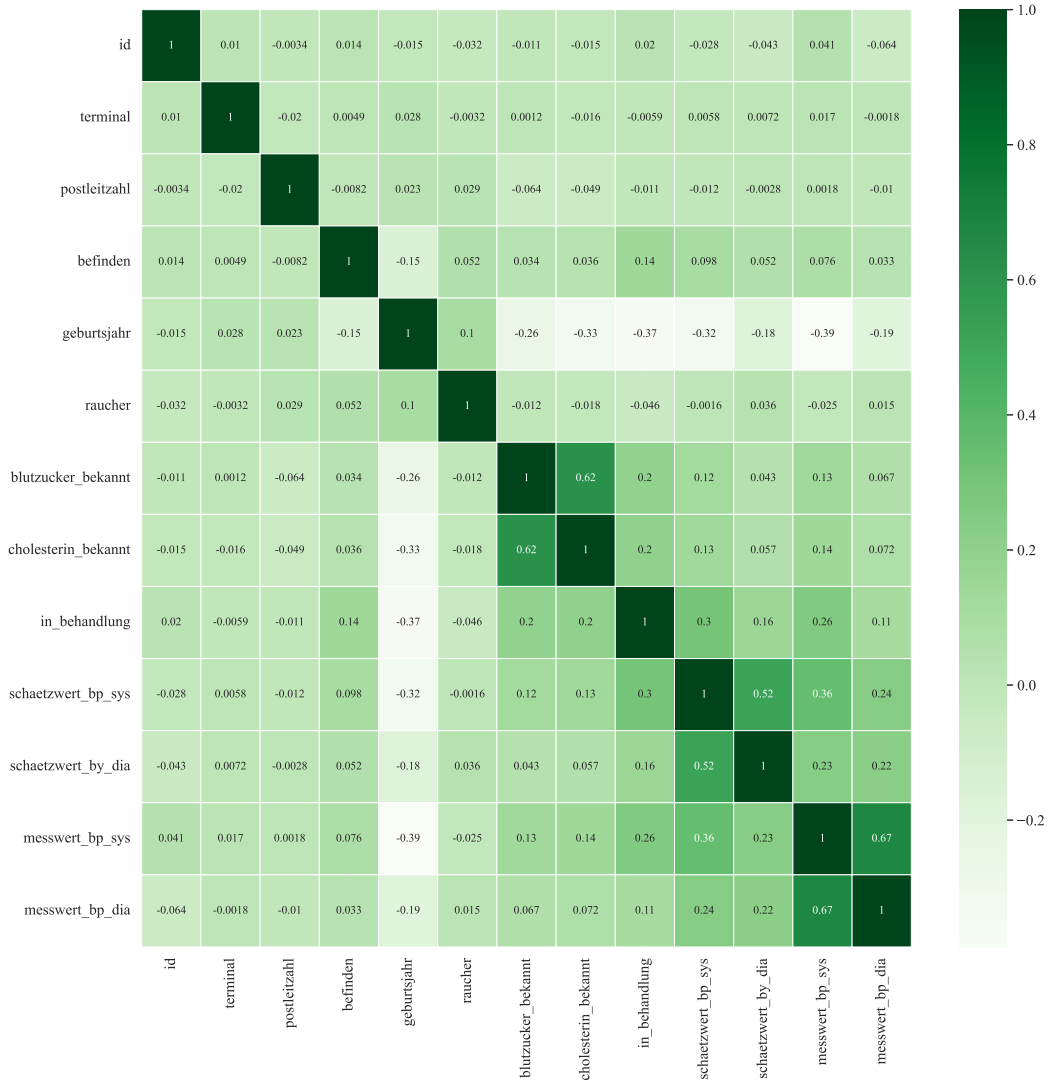
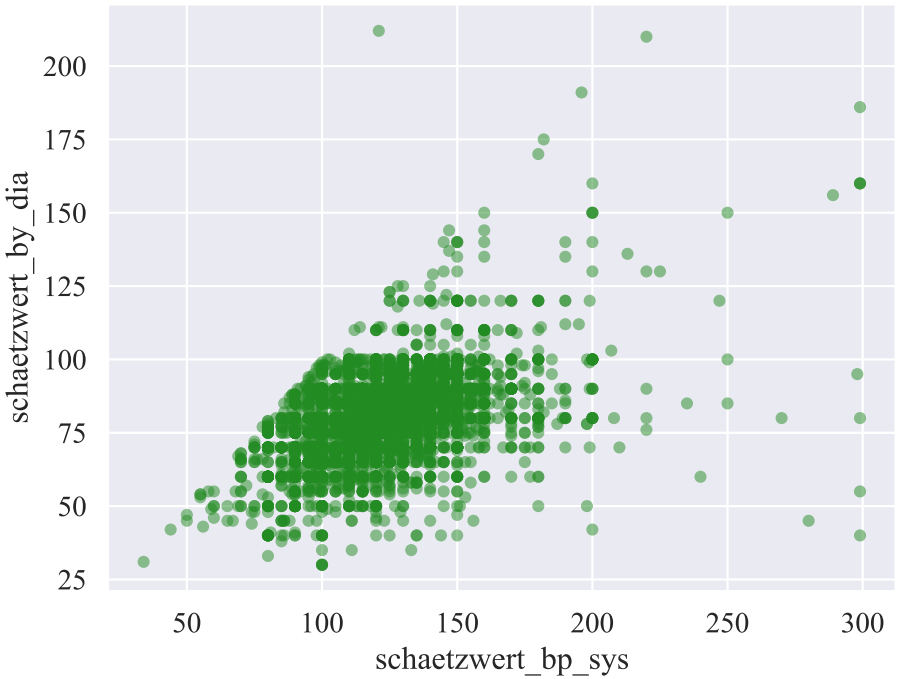


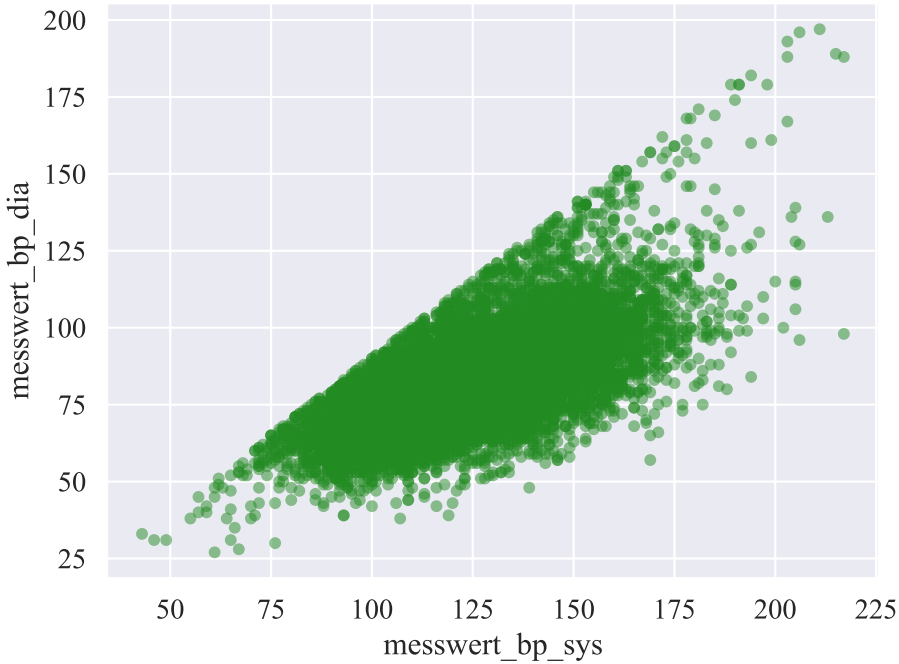
Figure: Correlation matrix using Spearman's correlation

Plots

Correlation



(a) Self-estimated systolic and diastolic blood pressures



(b) Measured systolic and diastolic blood pressures

- The graphs indicate that systolic and diastolic blood pressures are non-linearly correlated.

Open Questions

- New variables could be constructed using existing ones. e.g. *geburtsjahr* to *age*, *zeit* can be split up into *date* and *time*
- Could create box plots for homogeneity/heterogeneity comparisons
- Grouped box plots could be constructed against continuous variables e.g. effect of smoking on blood pressure

References

Van Rossum, G. & Drake Jr, F. Python reference manual. (Centrum voor Wiskunde en Informatica Amsterdam, 1995)

Fernando Pérez, Jupyter Notebook: Open-source platform for interactive programming. (Project Jupyter, 2021)

Illowsky, B. & Dean, S. Introductory Statistics. (OpenStax College, 2013)

Team, T. pandas-dev/pandas: Pandas. (Zenodo, 2020)

Chok, N. Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data. (2010), http://d-scholarship.pitt.edu/8056/

Waskom, M. seaborn: statistical data visualization. *Journal Of Open Source Software*. **6**, 3021 (2021), https://doi.org/10.21105/joss.03021

Thank you!

Questions?