Week 01: Descriptive Analysis

* Cleaned data – reformatted some columns to correct data-type
* Identified counts of datapoints and variables
* Translated meaning of variables and categorized the types of variables, i.e., Nominal, Categorical, Discrete etc

|  |  |
| --- | --- |
| Variable name (German) | 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 |

* Identified missing data:
  + Number of missing rows
  + Missing values per column
  + Possible question: Whether to keep or remove some rows?
* Bar plots for discrete and categorical variables
* Frequency Distributions for continuous variables
* Investigate correlation of variables via heatmaps or scatterplots – use some metric for correlation
* Possible Questions:
  + Create further new variables? Split Zeit variable into day, month, year, time perhaps? Create variable for age
  + Create group wise boxplots
  + Boxplots for homogeneity/heterogeneity comparisons