Irkutsk National Research Technical University
Baikal school of BRICS

09.04.02 Information systems and technology
Program: Information technologies, networks and big data

Course: Data analysis

Atalyan Alina Valerievna
alinaa@mail.ru

Practical class

Data Manipulation and Cleaning

1. Missing Data
   - Identifying and dealing with NA values
   - Imputation techniques
2. Outliers
   - Outlier Detection.

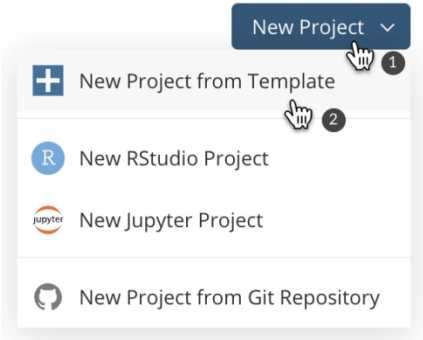
Posit Cloud

Getting Started with Posit Cloud

Sign up at https://posit.cloud/

Create the R project named Data Analysis 2025

Create two folders in the project Data and Result



Start coding – the interface is nearly identical to RStudio!

Data Set

Name: “Hormonal and Biochemical Profile in Patients"
Source: Internal company database
Size: 1148 rows × 41 columns
Format: CSV
Columns:

record_id (int) Unique order identifier
outcome (int) Presence of a tumor
hormone1 (num) result of hormone 1 level assessment
…
hormone4 (num) result of hormone 4 level assessment
h_index_34 (num) index (hormonal ratio)
…
hormone14 (num) result of hormone 4 level assessment
lipids1 (num) result of lipid 1 level assessment
…
lipids5 (num) result of lipid 5 level assessment
carb_metabolism (num) result of carbohydrate metabolism assessment
lipid_pero1 (num) result of lipid peroxidation 1 level assessment
…
lipid_pero5 (num) result of lipid peroxidation 5 level assessment
antioxidant1 (num) result of antioxidant 1 level assessment
…
antioxidant5 (num) result of antioxidant 5 level assessment
factor_eth (int)) Ethnicity
factor_h (int) Increased hormones
factor_pcos (int) Diagnosis
factor_prl (int) Increased hormone

Potential Use for the Practical class 1: Data Manipulation and Cleaning (Missing Data&Outliers)

Notes for the final project

Dataset Description Template

1. General Information

- **Dataset Name**: (e.g., "Customer Purchases 2023")
- **Source**: Where the data was obtained (e.g., company database, Kaggle, API, web scraping).
- **Author/Owner**: (e.g., "Company X," "UCI Machine Learning Repository")
- **Creation/Update Date**: (e.g., "Last updated: March 2024")
- **License**: (e.g., "CC0: Public Domain," "Commercial use restricted")
- **Purpose**: Why the dataset was created (e.g., "To analyze customer behavior").

2. Dataset Structure

- **Format**: (CSV, JSON, SQL, Parquet, etc.)
- **Size**:
  - Number of records (rows)
  - Number of features (columns)
- **Variables/Columns Description (Tab.)**:

Column Name	Data Type	Description	Possible Values	Missing Values?

3. Data Content & Semantics

Domain/Context: (e.g., "E-commerce transactions")
Target Variable (if applicable): (e.g., "price for regression analysis")
Time Coverage: (e.g., "Jan 2020 – Dec 2023")
Geographical Coverage (if applicable): (e.g., "US customers only")

4. Data Quality

Missing Data: Percentage and columns affected.
Duplicates: Are there duplicate entries?
Outliers/Anomalies: Unusual values (e.g., negative prices).
Data Balance: For classification, check class distribution.

5. Usage Examples

Possible Analyses: (e.g., "Customer segmentation, sales forecasting")
Machine Learning Tasks: (e.g., "Classification: Predict product_category")

6. Additional Notes

Collection Method: (e.g., "Web scraping, manual entry")
Preprocessing Done: (e.g., "Normalized prices, removed duplicates")
Limitations: (e.g., "Only includes US data, no refund records")

Missing Data

Strategy:

Identify Missing Data:

Where and how data is missing in the dataset → the nature of the missing

→ hypothesize about potential reasons for its absence

Quantifying Missing Data:

Determine the extent of missing data, both at a *variable level* and *observation level*

Visualizing Missing Data Patterns: Visual patterns of missing data sometimes can reveal insights, indicating whether missing data is random or systematic.

Analyzing the Impact of Missing Data:

Assess how missing values in one variable might relate to another, aiding in understanding if data is Missing Completely At Random (MCAR), Missing At Random (MAR), or Missing Not At Random (MNAR).

Handle Missing Data

Missing Data

up to 15% - can be restored safely,
15-30% - with caution,
>30% - it is often better to exclude or use special methods.

Key factors for decision making:

Missing type:

MCAR (random): easier to impute
MAR/MNAR: require complex methods

Variable importance:

Key predictors can be recovered even at 20-30%
Minor variables with >15% are sometimes better excluded

Sample size:

In large datasets (>10,000 observations) recovery can be attempted even at 20-25%
In small samples (>30%) often not recoverable

Analysis method:

For descriptive statistics - stricter restrictions
For ML models, higher percentages are sometimes acceptable

Missing Data

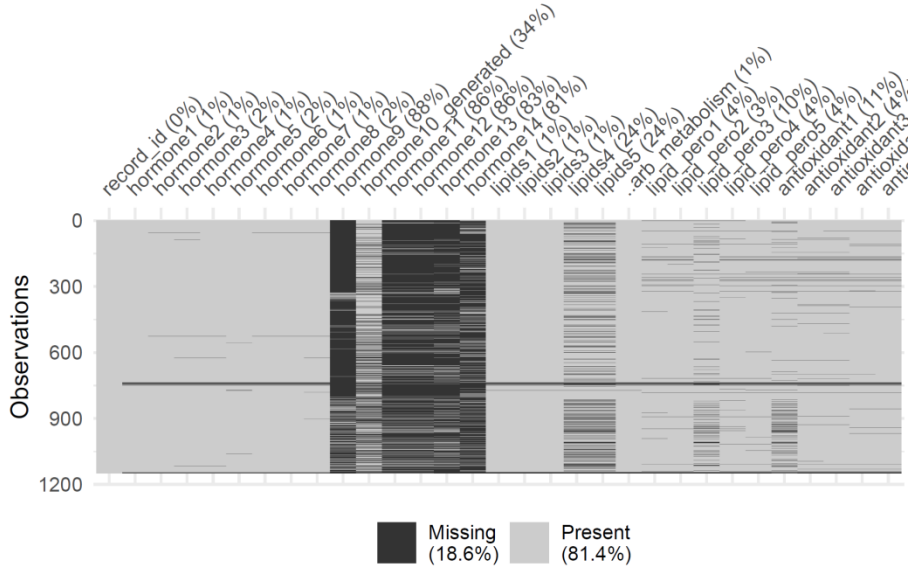
R-package **naniar**

Missing values, plotted with the vis_miss function

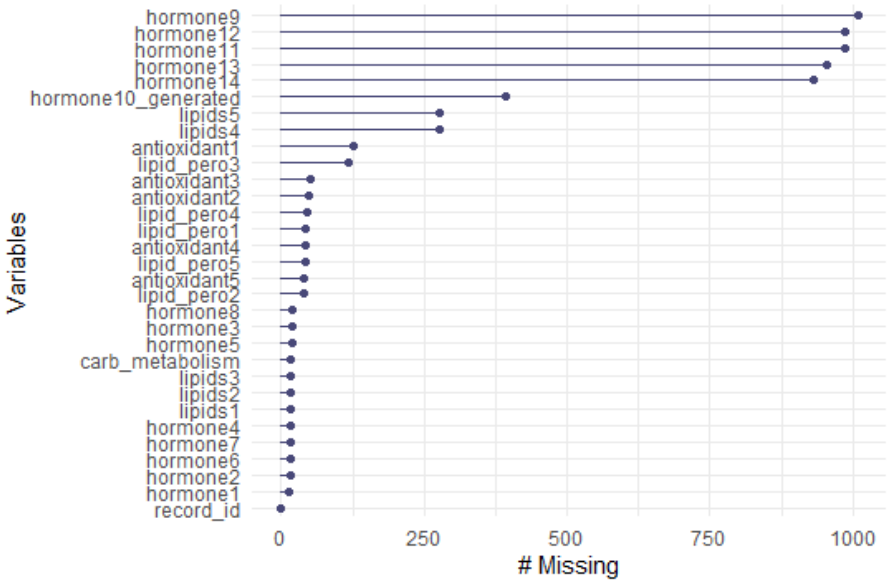
R-package **mice** (Multivariate Imputation via Chained Equations) - uses chained equations starting with the least missing

Missing Data

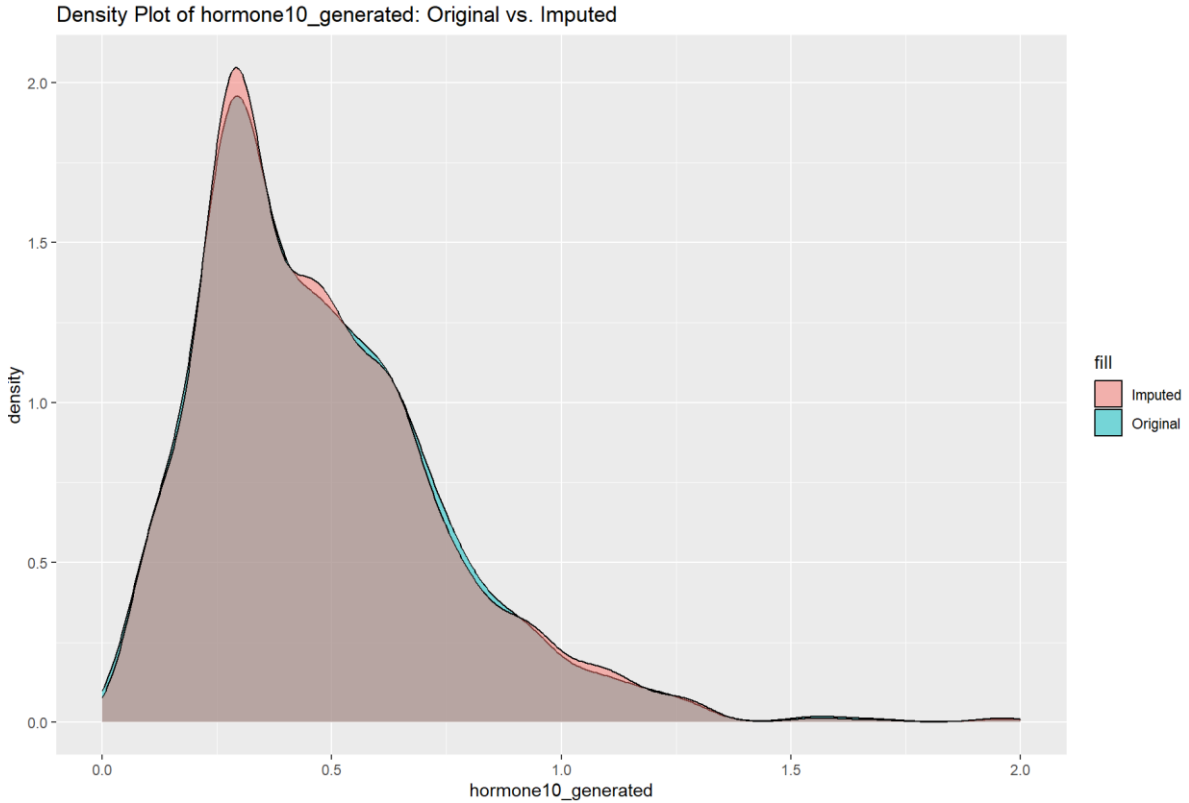
library(visdat)
vis_miss(MD_df)



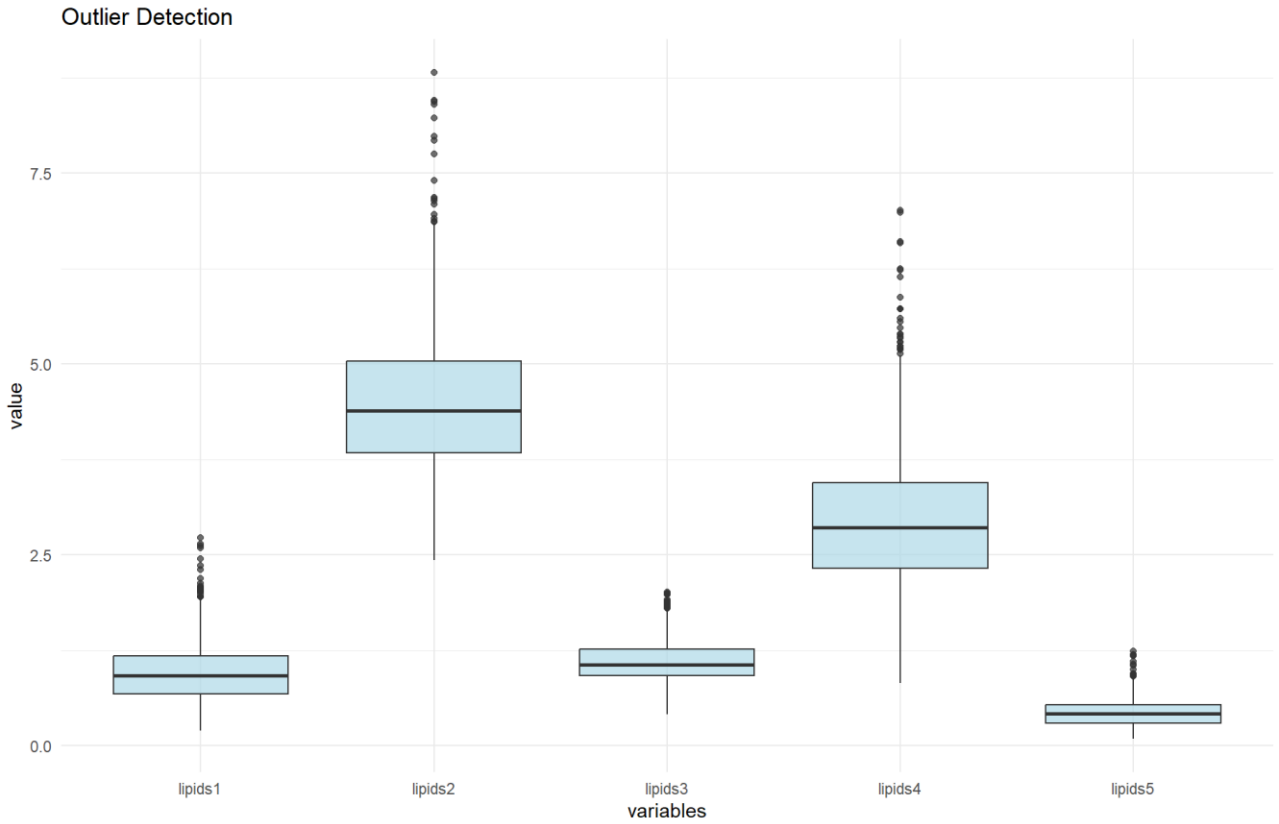
library(naniar)
gg_miss_var(MD_df)



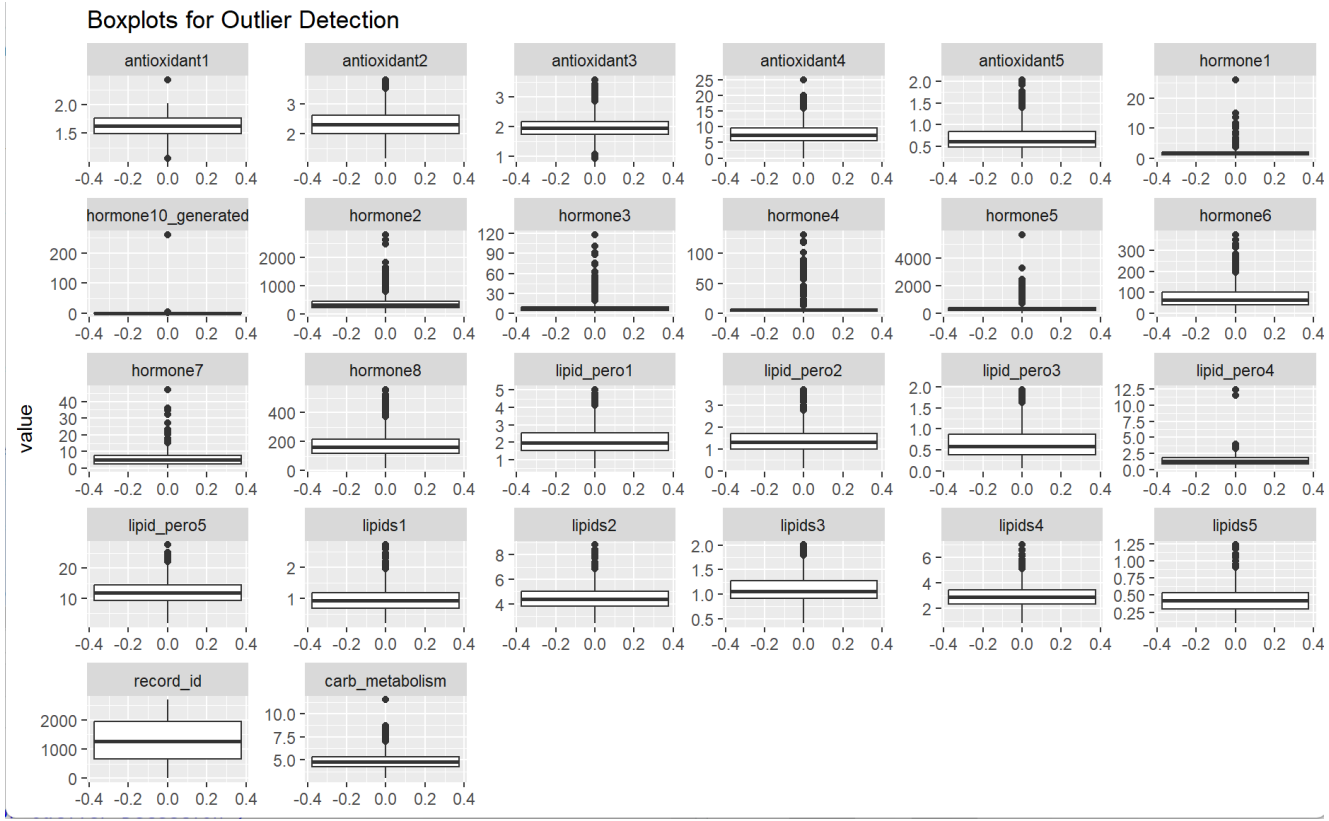
Missing Data



Outliers


Outlier Detection
7.5
5.0
2.5
0.0
value
lipids1
lipids2
lipids3
lipids4
lipids5
variables

Outliers


Boxplots for Outlier Detection
value
antioxidant1
antioxidant2
antioxidant3
antioxidant4
antioxidant5
hormone1
hormone10_generated
hormone2
hormone3
hormone4
hormone5
hormone6
hormone7
hormone8
lipid_pero1
lipid_pero2
lipid_pero3
lipid_pero4
lipid_pero5
lipids1
lipids2
lipids3
lipids4
lipids5
record_id
carb_metabolism

Task 1 (Deadline 13.04.2025 23:59)

Dataset “handle_MD_df”
Perform Little's MCAR Test

Key Outputs to Check:

- **p-value**:
  - *p* > *0.05* → Fail to reject H_0 (**data is MCAR**).
  - *p* ≤ *0.05* → Reject H_0 (**data is not MCAR**, likely MAR/MNAR).

1. For R 3.6.0-3.6.3

```
install.packages("BaylorEdPsych") # For Little's test
install.packages("mice") # For missing data visualization
library(BaylorEdPsych)
library(mice)

# Assuming your data is named 'df'
mcar_test <- LittleMCAR(df)
# View results
print(mcar_test)
```

2. For R 4.0+

```
library(naniar)
library(mice)
```

Task 1 (Deadline 13.04.2025 23:59)

- use default **'pmm'** method and compare with original, with **'rf'** method
- ***make a conclusion***

Imputation Method (method = 'pmm'):
PMM means Predictive Mean Matching and it's a non-parametric approach particularly suited for continuous data. PMM operates by finding observed values with similar predictive characteristics to the missing entries. The missing values are then imputed u, thus preserving the distribution and variance of the original data more effectively than simpler methods, such as mean imputation.

Task 1 (Deadline 13.04.2025 23:59)

for extra points!

For dataset “imputed_handle_MD_df_final”
packages("dbscan")

Local Outlier Factor (LOF) algorithm for detecting outliers in multivariate data.

Calculating LOF factors
Visualizing results (Histogram of LOF factors, Bivariate scatterplot with outliers)