# Exploratory Data Analysis (EDA) Report

**Dataset: cardio_data_processed.csv**
**Records:** 68,205 rows × 14+ columns
**Goal:** To prepare a clean, structured, and insightful dataset for cardiovascular risk prediction.

## 1. Data Loading & Initial Overview

### 1.1 Libraries Used

We used the following Python libraries:

- **NumPy**, **pandas** for data manipulation
- **Matplotlib**, **Seaborn**, **Plotly** for visualizations

### 1.2 Data Import & Structure

The dataset was read from a **tab-delimited CSV file**, ensuring that any merged columns were split so each feature had a dedicated field.
**Data shape: 68,205 rows × 14+ columns**
**Column categories include:**

- Demographics (age, gender, height, weight)
- Clinical measures (blood pressure readings)
- Lifestyle indicators (smoking, alcohol intake, activity)
- Target variable (**cardio**) indicating cardiovascular disease

**Significance:**
The large sample size increases the reliability of statistical estimates, though even small proportions of missing or incorrect data may affect analysis outcomes.

# 2. Data Quality Checks

## 2.1 Missing Values

We computed the percentage of missing values per column. Result:

- **No major missing data.**
- Most columns were **100% populated**, indicating high data integrity.

## 2.2 Duplicate Records

- Initial checks revealed **no exact duplicate rows**.
- **Conclusion:** Full dataset retained for analysis to preserve representativeness

# 3. Data Type Consistency & Conversions

- **Height and Weight** columns were converted to numeric types.
- Binary fields (**smoking**, **alcohol**, **active**, **gender**) cast 0**/1 integers**

# 4. Outlier Detection & Treatment

## 4.1 Height

- **Detected outliers: < 120 cm** or **> 210 cm**
- **Action:** Replaced with **median height (~165 cm)**
- **Result:** More realistic height distribution

## 4.2 Weight

- Non-convertible values dropped
- Flagged outliers: **< 30 kg** or **> 300 kg**
- **Action:** Replaced with **median weight (~75 kg)**

### 4.3 Blood Pressure (ap_hi, ap_lo)

- Verified **ap_hi ≥ ap_lo**
- **Dropped** inconsistent records (**ap_hi < 50**, implausible ranges)