Experiment 3

Objective: Learn data cleaning techniques, including handling missing data, outliers, and data imputation.

1. Exploring Inbuilt Functions for Data Cleaning

- Check missing values in a dataset using is.na(), complete.cases(), and summary().
- Identify outliers using boxplot(), quantile(), and IQR().
- Explore imputation methods like mean, median, and mode replacement using na.omit(), impute(), and mice().
- Learn about tidyverse functions (mutate(), filter(), replace_na()).
- Use summary(df) and str(df) to get dataset insights.
- Read and write cleaned data using read.csv() and write.csv().

2. Handling Missing Data (NA, NaN, Inf, NULL)

Create a sample dataset (dataframe) with missing values for hands-on practice

ID	Name	Age	Salary	Score
1	Alice	25	50000	80
2	Bob	NA	60000	90
3	NA	30	55000	NaN
4	David	29	NA	85
5	Emma	NA	70000	88
6	Frank	35	75000	92
7	NA	40	80000	NA
8	Hannah		NA	65000 77
9	Ian	50	NA	95
10	Jack	27	72000	Inf

Tasks (Handling Missing Data)

i. Identify missing data (is.na(df), sum(is.na(df))).
ii. Remove missing rows (na.omit(df)).
iii. Replace NA with zero (df[is.na(df)] <- 0).
iv. Replace NA with column mean (df$Age[is.na(df$Age)] <- mean(df$Age, na.rm=TRUE)).
v. Remove Inf and NaN (df$Score[is.infinite(df$Score) | is.nan(df$Score)] <- NA).

vi. Use tidyverse's replace_na() for selective column handling.

vii. Drop columns with excessive missing data (df <- df[, colSums(is.na(df)) < nrow(df) * 0.5]).

viii. Fill missing categorical values with the mode.

3. Outlier Detection & Handling

Detect and remove outliers from the dataset after handling missing data.

Tasks:

i. Boxplot Visualization to visualize salary data

ii. Z-Score Method (values outside ±3 standard deviations).

iii. IQR Method: Remove values outside Q1 - 1.5*IQR and Q3 + 1.5*IQR.

iv. Winsorization: Replace extreme values with percentiles (Winsorize()).

v. Detect & Remove Outliers Using tidyverse (filter()).

vi. Detect Outliers in Multiple Columns (apply()).

vii. Create a Clean Dataset After Removing Outliers.

4. Data Imputation

Explore data imputation techniques to fill missing values.

Tasks:

i. Convert **NaN** and **Inf** values to **NA** before applying imputation.

ii. Remove rows with missing values using na.omit(df).

iii. Drop columns where more than 50% of data is missing.

iv. Replace all NA values with 0 for numerical columns.

v. Replace missing values in Age with the mean.

vi. Replace missing values in Salary with the median.

vii. Replace missing **Name** values with the most frequent name (Mode)