Data Manipulation

Data Cleaning

Missing Data

- In R, missing values are represented by the symbol NA (not available).

- Impossible values (e.g., dividing by zero) are represented by the symbol NaN (not a number).

- Unlike SAS, R uses the same symbol for character and numeric data.

Testing for missing values

```
is.na(x)          # returns TRUE of x is missing

y <- c(1,2,3,NA)

is.na(y)        # returns a vector (F F F T)
```

Print the index of NA values

Testing for missing values in a Dataframe

```
df <- data.frame(col1 = c(1:3, NA),
          col2 = c("this", NA,"is", "text"),
          col3 = c(TRUE, FALSE, TRUE, TRUE),
          col4 = c(2.5, 4.2, 3.2, NA),
          stringsAsFactors = FALSE)
# identify NAs in full data frame
>is.na(df)
```

Testing for missing values in a specific column/row of a Dataframe

# identify NAs in specific data frame column

```
> is.na(df$col4)
```

Or

```
> is.na(df[,4])
```

Print NAs in row 3 in given data frame “df”

Location and the number of NAs

# identify location of NAs in vector

```
> which(is.na(df))
```

o/p: [1] 4 6 16

# identify count of NAs in data frame

```
> sum(is.na(df))
```

o/p: [1] 3

- **Print the count of NAs in row 3 in given data frame "df"**
- **Print count of NAs NAs in col 4 in given data frame "df"**

Excluding missing values

- Arithmetic functions on missing values yield missing values.

Ex:

```
x <- c(1,2,NA,3)
mean(x)                    # returns NA
mean(x, na.rm=TRUE)        # returns 2
```

- **Print the median of vector x**
- **Print the median of col1 in dataframe "df"**

Identifying the complete cases

- The function **complete.cases()** returns a logical vector indicating which cases are complete.

Ex:

# list rows of data without missing values

```
> df[complete.cases(df),]
```

Print the list of rows with missing values in given data frame “df”

listwise deletion of missing values

- The functions **na.omit() or na.exclude()** returns the object with listwise deletion of missing values

```
# create new dataset without missing data
> newdata <- na.omit(df)
```

```
# print the records without missing values
> na.exclude(df)
```