**Group 1:**

Cleaning data

It is a process of making your data set homogen, compact and usable for data analysis.

**Why?** To remove all the incoherent data

**What?** Can be missing data, not well reported data, insufficient data

**When?** Before data analysis

**How?** Using some codes :

**“Name of cleaned data <-dataset[complete.cases(dataset),]”**

**“Name of cleaned data<-na.omit(dataset)”**

Use the function

**“Is.na(Dataset)”**

To check your data set before cleaning. It will give an overview of data missing

Because missing values are coded as “Na”, to analyze data without removing missing values, you can give a value to all the Na as 99, by “Dataset[dataset == 99]

Use of Dplyr

Dplyr package helps to do so many functions like:

|  |  |  |
| --- | --- | --- |
| **Filter()** | Select a subset of Rows by conditions | *“Filter(data, condition)”* |
| **Arrange()** | Reorders Rows in a data frame | *“Arrange(data,* ***column to sort by****)”* |
| **Select()** | Select the Columns of interest | *“Select(data,* ***colums****)”* |
| **Mutate()** | Create new columns based on existing columns | *“Mutate(data,* ***new\_col****=oldcolumn)”* |
| **Summarise()** | Aggregate values for each group, rediuces to single values | *“Summarize(data)”* |