Data Transformation 1 (Sept. 29)

The data used is baby names

For Proportion of boys with the name Garrett: You need to do some transformation to the data to get the result you want.

Select function: select(data,column name, column name) or
select(data,first column name: last column name) or
select(data,c(column1, column2, column3))

The select function will select only the columns specified.
If you add a - sign before the c(column names) the function will select all other columns.
If you want to find columns that start a certain way use select(data, start_with(“”))
Also works with end_with()
Also works with contains()
If you have variables like x1,x2,x3 you can use select(data,num_range(“x”,c(1,2,3)))
Another way of extraction is using $, however the output is a vector rather than a tibble
data$column

Filter function extracts rows. Use logical operators because the rows are observations
filter(data, logical criteria)
Say you have a tibble DF with var x=c(1,2,3,4) you can use filter(DF, x >2) to select rows where x>2
Group membership: %in%
filter(Df, x %in% c(1,3)) will select rows where x = 1 or 3
To use more than 1 criteria with filter: filter(data, criteria1, criteria2) or
filter(data, criteria1 & criteria 2) or
| means or filter(data, criteria1 | criteria 2)

If data contains missing values they will appear as NA which won’t allow some functions to work.
To not include NA values use function(variable, na.rm = TRUE)
To find out how many NA values use sum(is.na(variable))
To remove NA values use filter(data, !is.na(variable))

Arrange function is used to arrange/sort data
arrange(data, variable) will sort the data by the variable
arrange(data,desc(variable)) will sort the data by the variable in decreasing order
NA values always come last when using arrange

To arrange data in order with more than one variable, put variable who's order comes first then next variable will be ordered within the first
arrange(data,x,y) will order data by x, then order that by y