\* Tạo một tệp new\_data mới bao gồm các biến cần phân tích

> new\_data = missing\_data[,c("nearest\_warehouse","order\_price","delivery\_charges","customer\_lat","customer\_long","coupon\_discount","order\_total","season","is\_expedited\_delivery","distance\_to\_nearest\_warehouse","is\_happy\_customer")]

\* Trong RStudio, kiểu kí tự không có tác dụng trong thống kê, vì vậy cần chuyển đổi những cột có dữ liệu kiểu ký tự (nearest\_warehouse, season) sang kiểu Factor. Đối với biến chỉ có giá trị False/True như is\_expedited\_delivery và is\_happy\_customer, thay False = 0, True = 1.

> new\_data$is\_expedited\_delivery[new\_data$is\_expedited\_delivery == 'True'] = 1

> new\_data$is\_expedited\_delivery[new\_data$is\_expedited\_delivery == 'False'] = 0

> new\_data$is\_expedited\_delivery = as.integer(new\_data$is\_expedited\_delivery)

> new\_data$is\_happy\_customer[new\_data$is\_happy\_customer == 'True'] = 1

> new\_data$is\_happy\_customer[new\_data$is\_happy\_customer == 'False'] = 0

> new\_data$is\_happy\_customer = as.integer(new\_data$is\_happy\_customer)

> new\_data$is\_happy\_customer = as.integer(new\_data$is\_happy\_customer)

> new\_data$season = factor((new\_data$season), levels = c("Spring","Summer","Autumn","Winter"))

> levels(new\_data$season)

[1] "Spring" "Summer" "Autumn" "Winter"

> new\_data$nearest\_warehouse = factor((new\_data$nearest\_warehouse), levels = c("Bakers","Nickolson","Thompson"))

> levels(new\_data$nearest\_warehouse)

[1] "Bakers" "Nickolson" "Thompson"

> summary(new\_data)

A screenshot of a computer

Description automatically generated