Chapter 3: Use Case Notebook for Instructors

Ram Gopal, Dan Philps, and Tillman Weyde

Summer 2022

Table of Contents

library(dplyr)  
library(tidyverse)  
library(readxl)

# Use Case: Filter, Sort, Bin Sales data

df = read\_excel("../../data/supermarket\_sales.xlsx")

Select rows where the “Gender” column equals “Female”:

head(df[df$Gender == 'Female',])

## # A tibble: 6 × 17  
## `Invoice ID` Branch City Custo…¹ Gender Produ…² Unit …³ Quant…⁴ Tax 5…⁵ Total  
## <chr> <chr> <chr> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 750-67-8428 A Yang… Member Female Health… 74.7 7 26.1 549.   
## 2 226-31-3081 C Nayp… Normal Female Electr… 15.3 5 3.82 80.2  
## 3 355-53-5943 A Yang… Member Female Electr… 68.8 6 20.7 434.   
## 4 315-22-5665 C Nayp… Normal Female Home a… 73.6 10 36.8 772.   
## 5 665-32-9167 A Yang… Member Female Health… 36.3 2 3.63 76.1  
## 6 692-92-5582 B Mand… Member Female Food a… 54.8 3 8.23 173.   
## # … with 7 more variables: Date <dttm>, Time <dttm>, Payment <chr>, cogs <dbl>,  
## # `gross margin percentage` <dbl>, `gross income` <dbl>, Rating <dbl>, and  
## # abbreviated variable names ¹​`Customer type`, ²​`Product line`,  
## # ³​`Unit price`, ⁴​Quantity, ⁵​`Tax 5%`

Sort this filtered dataframe by “Quantity” of sales, with a descending sort (i.e., highest values are first):

df\_female = df[df$Gender == 'Female',]  
df\_female = df\_female[order(df\_female$Quantity,decreasing = T),]

We now want to examine whether the hour of the day has a bearing on sales volumes, for instance if most sales happen in the final hour of business of the day, and if so, if we should keep stores open for one hour more to encourage more sales. To do this we can “bin” the rows into equally sized bins.

df\_female$hour = format(as.POSIXct(df\_female$Time, format="%Y-%m-%d %H:%M"), format="%H")  
df\_female$hour = as.numeric(df\_female$hour)  
breaks <- c(0,2,4,6,8,10,12,14,16,18,20,22)  
tags <- c("[0-2)","[2-4)", "[4-6)", "[6-8)", "[8-10)", "[10-12)","[12-14)", "[14-16)","[16-18)", "[18-20)","[20-22)")  
  
df\_female$time\_bin <- cut(df\_female$hour,   
 breaks=breaks,   
 include.lowest=TRUE,   
 right=FALSE,   
 labels=tags)

We can now plot the relationship between Total sales in each time bin:

ggplot(df\_female,aes(x =time\_bin, y = Quantity)) +   
 geom\_col(size = 1)

