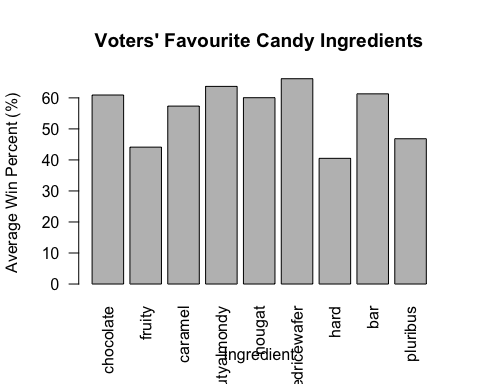
task\_1.R

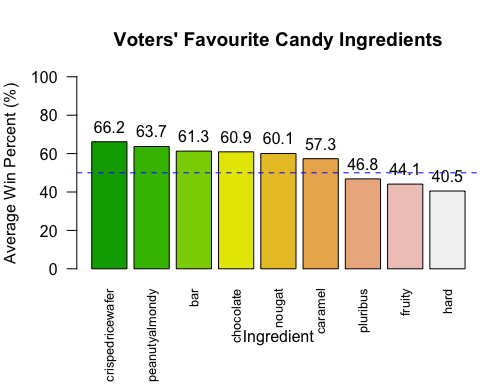
kaushikshamantha

2025-04-16

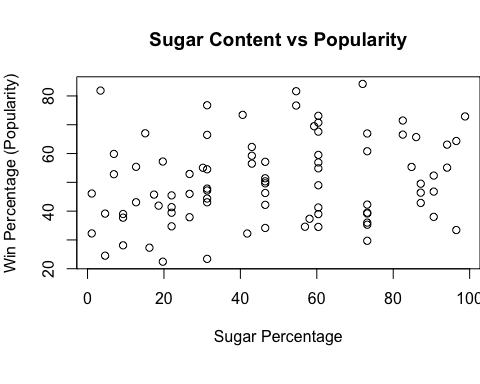
candy\_data = read.csv("CMM703 Candy Data 2024-2025.csv", header = TRUE)  
  
# these are the candy ingredients  
ingredients = c("chocolate", "fruity", "caramel", "peanutyalmondy", "nougat", "crispedricewafer", "hard", "bar", "pluribus")  
  
# lets analysis favorite candy ingredient  
feature\_average = data.frame(  
 feature\_name = character(),  
 average\_win\_percentage = numeric()  
 )  
  
for (ingredient in ingredients) {  
 has\_feature = candy\_data[candy\_data[[ingredient]] == 1, ]  
   
 # calculate the average 'winpercent' where particular ingredient is in the candy  
 average\_win = mean(has\_feature$winpercent, na.rm = TRUE)  
 feature\_average = rbind(  
 feature\_average,  
 data.frame(  
 feature\_name = ingredient,  
 average\_win\_percentage = average\_win  
 )  
 )  
}  
  
barplot(  
 feature\_average$average\_win\_percentage,  
 names.arg = feature\_average$feature\_name,  
 las = 2,  
 main = "Voters' Favourite Candy Ingredients",  
 xlab = "Ingredient",  
 ylab = "Average Win Percent (%)",  
)



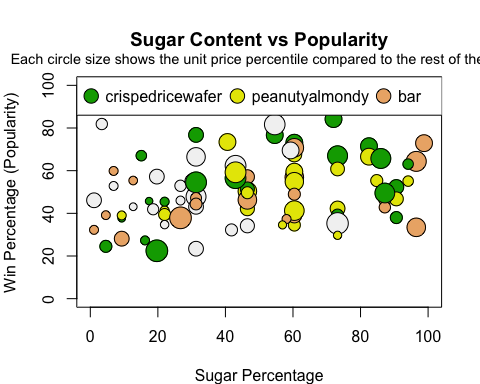
# as you can see here it is bit hard to analysis the information given in the chart.  
# lets improve the chart by adding colors, sort the data and few extra labels.  
  
# sort by popularity  
feature\_average = feature\_average[order(-feature\_average$average\_win\_percentage), ]  
  
par(mar = c(6, 4, 4, 0))  
  
bar\_colors = terrain.colors(length(feature\_average$average\_win\_percentage))  
  
bar\_chart = barplot(  
 feature\_average$average\_win\_percentage,  
 names.arg = feature\_average$feature\_name,  
 las = 2,  
 col = bar\_colors,  
 main = "Voters' Favourite Candy Ingredients",  
 xlab = "Ingredient",  
 ylab = "Average Win Percent (%)",  
 ylim = c(0, 100),  
 cex.names = 0.8  
)  
  
# add values on top of the bars  
text(  
 x = bar\_chart,   
 y = feature\_average$average\_win\_percentage,   
 label = round(feature\_average$average\_win\_percentage, 1),   
 pos = 3  
)  
  
# blue dashed line at 50%  
abline(h = 50, col = "blue", lty = 2)



# according to the chart you can see that most popular candy ingredient is "crispedricewafer" and the least is candy which is "hard".  
# there are 5 candy ingredients achieve more that 50%  
# 'chocolate' and 'nougat' ingredients almost share the same popularity  
  
# lets pick 4 ingredients where the most popular  
candy\_data$candy\_type = "other"  
candy\_data$candy\_type[candy\_data$crispedricewafer == 1] = "crispedricewafer"  
candy\_data$candy\_type[candy\_data$peanutyalmondy == 1] = "peanutyalmondy"  
candy\_data$candy\_type[candy\_data$bar == 1] = "bar"  
  
candy\_data$sugarpercent = candy\_data$sugarpercent \* 100  
  
plot(  
 x = candy\_data$sugarpercent,  
 y = candy\_data$winpercent,  
 main = "Sugar Content vs Popularity",  
 xlab = "Sugar Percentage",  
 ylab = "Win Percentage (Popularity)",  
)



# this scatter plot shows how candy distribute across win percentage against sugar percentage.  
# as you can see it is not easy to get information from the chart.  
# lets improve the chart by adding colors, some info replated to the price.  
  
par(mar = c(4, 4, 4, 2))  
  
scatter\_colors = terrain.colors(length(unique(candy\_data$candy\_type)))  
  
candy\_price\_sizes = 1 + (candy\_data$pricepercent / max(candy\_data$pricepercent)) \* 2  
  
plot(  
 x = candy\_data$sugarpercent,  
 y = candy\_data$winpercent,  
 main = "Sugar Content vs Popularity",  
 xlab = "Sugar Percentage",  
 ylab = "Win Percentage (Popularity)",  
 xlim = c(0, 100),  
 ylim = c(0, 100),  
 col = scatter\_colors,  
 cex = candy\_price\_sizes,  
 pch = 19  
)  
  
mtext(  
 "Each circle size shows the unit price percentile compared to the rest of the set",  
 side = 3,  
 line = 0.5,  
 cex = 0.9  
)  
  
points(  
 x = candy\_data$sugarpercent,  
 y = candy\_data$winpercent,  
 cex = candy\_price\_sizes,  
 pch = 21,  
 bg = scatter\_colors,  
 col = "black"  
)  
  
legend(  
 "topleft",  
 legend = c("crispedricewafer", "peanutyalmondy", "bar", "other"),  
 pt.bg = scatter\_colors,   
 col = "black",  
 pch = 21,  
 pt.cex = 2,  
 horiz = TRUE,  
)



# here in the chart you can see that data is scatter between (20 - 85)% in the y axis the win percentage  
# according to the chart more popular candys have higher sugar content.  
# popular and sugary candys are not cheep according to the chart