Final\_project

Tanmay Khairnar, Sohail Shaikh, Ashish Solanki

2023-10-05

#Problem Statement

Regork Inc. is seeking to enhance its sales and customer satisfaction by gaining a comprehensive understanding of the purchasing behavior of married individuals within various age groups. The primary objective of this data analysis project is to identify buying patterns among this demographic and formulate personalized recommendations and marketing strategies tailored to each age group. By performing a thorough data analysis, we aim to empower Regork to boost sales effectively and cater to the diverse preferences of its customer base.

**The Regork CEO should be highly interested in this problem statement and the accompanying data analysis for several compelling reasons:**

**Sales Growth:** The data analysis project aims to uncover insights into the buying behavior of married individuals across different age groups. This information will provide actionable strategies for boosting sales. Increased sales directly contribute to revenue growth, which is a fundamental goal for any business, including Regork.

**Customer Satisfaction:** By tailoring marketing strategies and product offerings based on the preferences of specific age groups, Regork can enhance customer satisfaction. Satisfied customers are more likely to become repeat buyers and brand advocates, leading to long-term profitability.

**Adaptation to Changing Trends:** Consumer preferences evolve over time. By continuously analyzing data, Regork can stay ahead of changing trends and adapt its strategies accordingly, ensuring long-term relevance and success.

In addressing our business challenge, we have undertaken a comprehensive analysis of the purchasing behaviour exhibited by a specific demographic segment, namely, married individuals.

Our principal objective is to discern patterns and preferences within this group, particularly focusing on different age cohorts. By doing so, we aim to pinpoint frequently purchased products among married individuals of varying age groups and, in turn, identify product combinations that can significantly enhance sales.

*To achieve this, we intend to implement a tailored recommendation system that offers personalized suggestions to customers, further augmenting sales opportunities for frequently bought items.*

**From our analysis we propose following recommendations:**

**Product Combos:** Tailored product combinations that align with the preferences of specific age groups. For instance, combining Soft Drinks with Bags of Chips for the 19 - 24 age group.

**Marketing Strategies:** Strategies aimed at effectively promoting these combos, including discounts and promotional campaigns, to attract and engage customers.

**Strategic Marketing Timing:** To further optimize sales, we have identified specific days of the week when these product combos should be strategically marketed to maximize customer footfall and purchasing behavior. For example, we recommend marketing these combos on Mondays for several age groups.

## Data Sources

*Our analysis hinges on three primary data sources:*

**Transaction Data Exploration:** We delve into transactional data to identify purchasing patterns among married individuals. This step allows us to uncover which products are more likely to be bought by this segment.

**Demographics Data Review:** By examining demographic attributes, we gain insights into the distribution of married individuals across different age groups. This segmentation helps us understand variations in buying behaviour based on age.

**Product Data Overview:** An understanding of our product catalog is essential. We assess which items are available and frequently purchased, forming the basis for our analysis.

*Throughout the following sections of this report, we will present our findings, including visualizations and insights, to guide our subsequent analysis and decision-making processes.*

**Libraries utilized for data ana;ysis**

library(completejourney)  
library(tidyverse)  
library(dplyr)  
library(ggplot2)  
library(scales)

Here’s an explanation of the use of each of these libraries:

The **completejourney** package provides access to data sets characterizing household level transactions over one year from a group of 2,469 households who are frequent shoppers at a grocery store. It contains all of each household’s purchases, not just those from a limited number of categories. For certain households, demographic information as well as direct marketing contact history are captured.

The **tidyverse** is a collection of R packages (including dplyr, ggplot2, and more) that are designed to work together seamlessly for data manipulation, visualization, and analysis.

**dplyr** is used for data manipulation, including filtering, summarizing, and transforming data frames.

**ggplot2** is a powerful package for creating customized data visualizations and plots.

**scales** is a package that provides tools for formatting and scaling axes in plots created with ggplot2.

## Exploratory Data Analysis

In this section, we employ a range of data visualizations to shed light on the critical aspects of our business problem. Our primary focus is on understanding the buying behaviour of a specific demographic segment—married individuals.

To enable a comprehensive analysis of our business problem, we undertake a process of data integration, bringing together essential information from multiple sources. Specifically, we integrate the transactions data and the demographics data by performing an inner join on the shared key ‘household id.’ This step allows us to associate transactional information with various households.

Subsequently, we conduct another inner join, this time on the key ‘product id,’ incorporating the product data into our integrated dataset. This meticulous data integration process results in the creation of our master data frame, which serves as the foundation for our analytical endeavours.

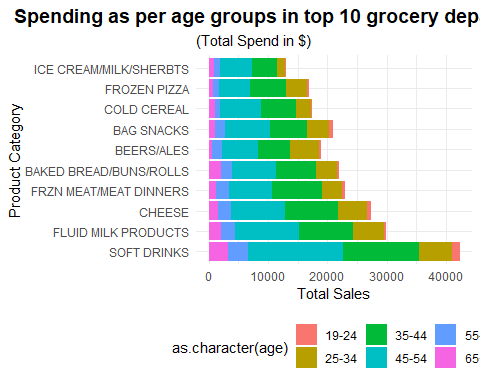
transactions <- get\_transactions()  
  
new\_df1 <- transactions%>%  
 inner\_join(demographics, by = 'household\_id')%>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married") %>%  
 group\_by(basket\_id, product\_category) %>%  
 summarize(Total\_sales = sum(sales\_value))

*Initial Analysis: Identifying the “Most Purchased Product” Across Age Groups*

In the preliminary stages of our analysis, a key point of interest is to determine the product that emerges as the “Most Purchased Product” within each age group. This analysis is pivotal in understanding which products hold the highest appeal across the entire demographic spectrum.

Our approach involves leveraging the master data frame, which amalgamates transactional, demographic, and product information. By utilizing this comprehensive dataset, we identify the product with the highest sales within each distinct age group.

df1<-transactions%>%  
 inner\_join(demographics, by ='household\_id')  
  
df5<-df1%>%  
 inner\_join(products, by= 'product\_id')%>%  
 filter(department=='GROCERY',marital\_status=='Married')%>%  
 group\_by(product\_category,age)%>%  
 summarise(Total\_sales = sum(sales\_value)) %>%  
 arrange(desc(Total\_sales))%>%  
 arrange(product\_category, desc(Total\_sales)) %>%  
 mutate( product\_category= fct\_reorder(product\_category, Total\_sales, .desc = TRUE))  
  
  
desired\_order <- c("SOFT DRINKS", "FLUID MILK PRODUCTS","CHEESE"," BAG SNACKS","FRZN MEAT/MEAT DINNERS","BAKED BREAD/BUNS/ROLLS","BEERS/ALES")  
#df5$product\_category <- factor(df5$product\_category, levels = desired\_order)  
  
  
all\_desired\_levels <- union(desired\_order, df5$product\_category)  
df5$product\_category <- factor(df5$product\_category, levels = all\_desired\_levels)  
  
  
top\_grocery <- df5 %>%  
 group\_by(product\_category) %>%  
 summarise(Total\_sales = sum(Total\_sales)) %>%  
 arrange(desc(Total\_sales)) %>%  
 top\_n(10)  
  
  
#custom\_colors <- c("0" = "blue", "1" = "red", "2" = "green", "3+" = "yellow")  
  
  
df5 %>%  
 filter(product\_category %in% top\_grocery$product\_category) %>%  
 ggplot(aes(x = product\_category, y = Total\_sales, fill = as.character(age))) +  
 geom\_bar(stat = "identity") +  
 #scale\_fill\_manual(name = "Number of Kids", values = custom\_colors) +  
 labs(  
 title = "Spending as per age groups in top 10 grocery department",  
 subtitle = "(Total Spend in $)",  
 x = "Product Category",  
 y = "Total Sales"  
 ) +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(size = 14, face = "bold", hjust = 0.7),  
 legend.position = "bottom" # Adjust the legend position  
 ) +  
coord\_flip()



*A prominent and striking revelation emerges from our analysis, as illustrated by the graph above: “SOFT DRINKS” stands out as the unequivocal front runner in sales across all age groups. This finding underscores the broad appeal and consistent demand for Soft Drinks, transcending demographic boundaries.*

**Identifying Frequently Purchased Products with Soft Drinks by Age Group**

In our pursuit of a deeper understanding of customer behaviour, we aim to uncover not only the popularity of **Soft Drinks** but also the accompanying products frequently bought by customers within specific age groups. To achieve this, we employ a targeted approach by analyzing the data pertaining to the chosen age demographic.

Our methodology entails the following steps:

**Data Selection by Age Group**: We begin by selecting the specific age group under investigation. This segmentation allows us to focus on a particular subset of customers.

**Filtering by Product Category**: We filter the master data frame to include only those transactions where the product category is “Soft Drinks.” This step provides us with a list of basket IDs that contain Soft Drinks.

**Identifying Frequently Purchased Products**: For the basket IDs containing Soft Drinks within the chosen age group, we then proceed to identify the products that are frequently purchased alongside Soft Drinks. This analysis reveals the items most found in shopping baskets along with Soft Drinks for the selected age group.

new\_df2 <- transactions %>%  
 inner\_join(demographics, by = 'household\_id') %>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married") %>%  
 group\_by(basket\_id, product\_category, transaction\_timestamp, sales\_value, age) %>%  
 summarize(sales = sum(sales\_value)) %>%  
 ungroup() %>%  
 filter(str\_detect(product\_category, regex("soft drinks", ignore\_case = TRUE)))  
  
  
new\_df3 <- new\_df1 %>%  
 inner\_join(new\_df2, by = 'basket\_id') %>%  
 filter(!str\_detect(product\_category.x, regex("soft drinks", ignore\_case = TRUE))) %>%  
 group\_by(basket\_id, product\_category.x, age, sales\_value) %>%  
 summarize(unique\_prod = list(unique(product\_category.x))) %>%  
 group\_by(unique\_prod, age) %>%  
 summarize(Final\_sales = sum(sales\_value)) %>%  
 arrange(desc(Final\_sales)) %>%  
 unnest(unique\_prod)  
  
new\_df3

## # A tibble: 1,464 × 3  
## # Groups: unique\_prod [286]  
## unique\_prod age Final\_sales  
## <chr> <ord> <dbl>  
## 1 FLUID MILK PRODUCTS 45-54 6432.  
## 2 BAKED BREAD/BUNS/ROLLS 45-54 5621.  
## 3 FLUID MILK PRODUCTS 35-44 5121.  
## 4 BAG SNACKS 45-54 4950.  
## 5 BAKED BREAD/BUNS/ROLLS 35-44 4465.  
## 6 CHEESE 45-54 4264.  
## 7 TROPICAL FRUIT 45-54 3883.  
## 8 BAG SNACKS 35-44 3633.  
## 9 CHEESE 35-44 3601.  
## 10 BEEF 45-54 3214.  
## # ℹ 1,454 more rows

*Analysis for Age Groups: Soft Drinks and Accompanying Products*

For the age group spanning from 19 to 24 years to 65+ years, our analysis reveals products, frequently purchased alongside **Soft Drinks**. To further refine our marketing strategy, we identify the total sales generated by these products when sold in conjunction with Soft Drinks.

Upon analyzing the data, we find the commonly purchased products with Soft Drinks in diverse age groups, that exhibit the highest total sales figures. Our primary objective is to identify product combinations that make marketing sense and can effectively boost sales.

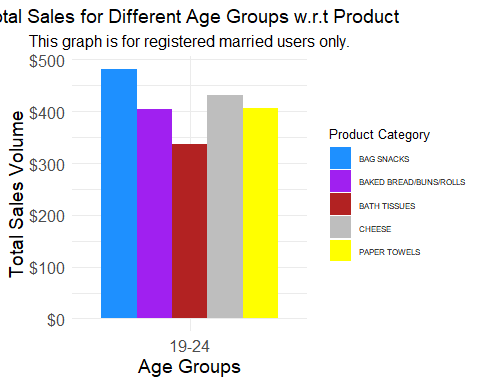
*Customer Purchasing Trends Across the Week*

To optimize the sales of the selected combo, it is imperative to understand customer purchasing trends throughout the week. Our analysis tries to reveal weekdays when, for different age categories, the sales of the identified combo reach their zenith and dip to their lowest point.

This insight offers a valuable marketing strategy. By strategically promoting our **combo** on days when sales are typically lower, we can capitalize on this opportunity to boost sales and potentially encourage customers to make purchases they may not have considered otherwise.

##Age group : 19-24

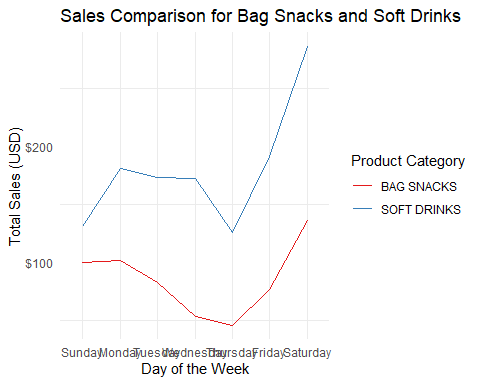
category\_sales\_1 <- new\_df3 %>%  
 filter(age == "19-24") %>%  
 group\_by(unique\_prod, age) %>%  
 summarise(total\_sales = sum(Final\_sales)) %>%  
 arrange(desc(total\_sales))  
  
  
# Select the top 5 product categories  
top\_5\_categories\_1 <- category\_sales\_1 %>%  
 unnest(unique\_prod) %>%  
 top\_n(5, total\_sales) %>%  
 pull(unique\_prod)  
  
  
  
# Filter the data to include only the top 5 categories  
filtered\_data\_1 <- new\_df3 %>%  
 filter(unique\_prod %in% top\_5\_categories\_1, age == "19-24") %>%  
 unnest(unique\_prod) %>%  
 head(5)  
  
  
category\_colors\_1 <- c("BAG SNACKS" = "dodgerblue", "CHEESE" = "grey", "PAPER TOWELS" = "yellow",  
 "BAKED BREAD/BUNS/ROLLS" = "purple", "BATH TISSUES" = "firebrick")  
  
# Create the ggplot visualization with the filtered data  
ggplot(filtered\_data\_1, aes(x = age, y = Final\_sales, fill = as.factor(unique\_prod))) +  
 geom\_bar(stat = "identity", position = "dodge") +   
 scale\_y\_continuous(name = "Total Sales Volume", labels = scales::dollar) +  
 labs(  
 title = "Total Sales for Different Age Groups w.r.t Product",  
 subtitle = "This graph is for registered married users only.",  
 x = "Age Groups",  
 y = "Total Sales",  
 fill = "Product Category"  
 ) +  
 scale\_fill\_manual(values = category\_colors\_1) +  
 theme\_minimal() +  
 theme(  
 legend.position = "right", # Move the legend to the right  
 plot.title = element\_text(size = 14, hjust = 0.5), # Adjust title font size and alignment  
 plot.subtitle = element\_text(size = 12, hjust = 0.5), # Adjust subtitle font size and alignment  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text = element\_text(size = 12),  
 legend.title = element\_text(size = 10),  
 legend.text = element\_text(size = 6)  
   
 )



*Following the analysis of the plot presented above, it becomes evident that the demographic group aged 19 to 25 demonstrates a distinct inclination towards purchasing Chips*

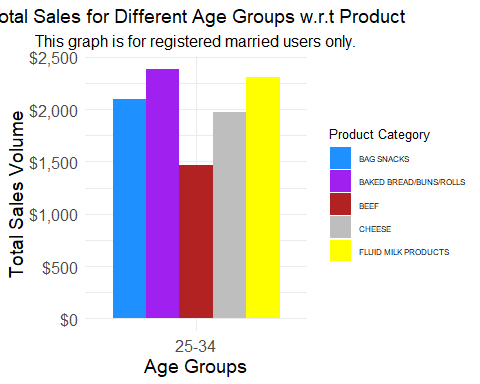
*Day of Week comparison*

prod\_1 <- transactions %>%  
 inner\_join(demographics, by = 'household\_id') %>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married", age == "19-24") %>%  
 filter(  
 str\_detect(product\_category, regex("BAG SNACKS|SOFT DRINKS", ignore\_case = TRUE)),  
 ) %>%  
 group\_by(product\_category, transaction\_timestamp, age) %>%  
 summarize(sales = sum(sales\_value)) %>%  
 arrange(desc(sales))  
  
  
prod\_1\_sales <- prod\_1 %>%  
 mutate(DayofWeek = weekdays(transaction\_timestamp)) %>%  
 group\_by(DayofWeek, product\_category) %>%  
 summarize(Total\_sales = sum(sales)) %>%  
 arrange(desc(DayofWeek))  
  
  
ggplot(prod\_1\_sales, aes(x = factor(DayofWeek, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")),   
 y = Total\_sales, color = product\_category, group = product\_category)) +  
 geom\_line() +  
 labs(title = "Sales Comparison for Bag Snacks and Soft Drinks",  
 x = "Day of the Week",  
 y = "Total Sales (USD)", # Added "USD" as y-axis label  
 color = "Product Category") +  
 theme\_minimal() +  
 scale\_color\_brewer(palette = "Set1") +  
 theme(  
 legend.position = "right",  
 panel.grid.major.y = element\_blank(),  
 panel.border = element\_blank()  
 ) +  
 scale\_y\_continuous(labels = dollar\_format(prefix = "$")) + # Format y-axis labels in thousands  
 guides(color = guide\_legend(title = "Product Category")) # Rename legend title

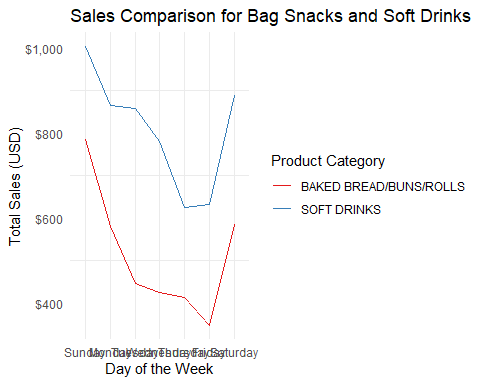
 *In reference to the plot above, it is clear that the highest sales are consistently observed on Saturdays, while the lowest sales occur on Thursdays. This observation underscores the importance of considering these sales trends when planning marketing and promotional activities throughout the week*

*Age group : 25-34*

category\_sales\_2 <- new\_df3 %>%  
 filter(age == "25-34") %>%  
 group\_by(unique\_prod, age) %>%  
 summarise(total\_sales = sum(Final\_sales)) %>%  
 arrange(desc(total\_sales))  
  
  
# Select the top 5 product categories  
top\_5\_categories\_2 <- category\_sales\_2 %>%  
 unnest(unique\_prod) %>%  
 top\_n(5, total\_sales) %>%  
 pull(unique\_prod)  
  
  
# Filter the data to include only the top 5 categories  
filtered\_data\_2 <- new\_df3 %>%  
 filter(unique\_prod %in% top\_5\_categories\_2, age == "25-34") %>%  
 unnest(unique\_prod) %>%  
 head(5)  
  
  
category\_colors\_2 <- c("BAG SNACKS" = "dodgerblue", "CHEESE" = "grey", "FLUID MILK PRODUCTS" = "yellow",  
 "BAKED BREAD/BUNS/ROLLS" = "purple", "BEEF" = "firebrick")  
  
# Create the ggplot visualization with the filtered data  
ggplot(filtered\_data\_2, aes(x = age, y = Final\_sales, fill = as.factor(unique\_prod))) +  
 geom\_bar(stat = "identity", position = "dodge") +   
 scale\_y\_continuous(name = "Total Sales Volume", labels = scales::dollar) +  
 labs(  
 title = "Total Sales for Different Age Groups w.r.t Product",  
 subtitle = "This graph is for registered married users only.",  
 x = "Age Groups",  
 y = "Total Sales",  
 fill = "Product Category"  
 ) +  
 scale\_fill\_manual(values = category\_colors\_2) +  
 theme\_minimal() +  
 theme(  
 legend.position = "right", # Move the legend to the right  
 plot.title = element\_text(size = 14, hjust = 0.5), # Adjust title font size and alignment  
 plot.subtitle = element\_text(size = 12, hjust = 0.5), # Adjust subtitle font size and alignment  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text = element\_text(size = 12),  
 legend.title = element\_text(size = 10),  
 legend.text = element\_text(size = 6)  
 )

 *Based on the insights extracted from the plot above, it’s apparent that customers in the age bracket of 25 to 34 exhibit a notable pattern of frequent purchases of Baked Breads and Rolls.*

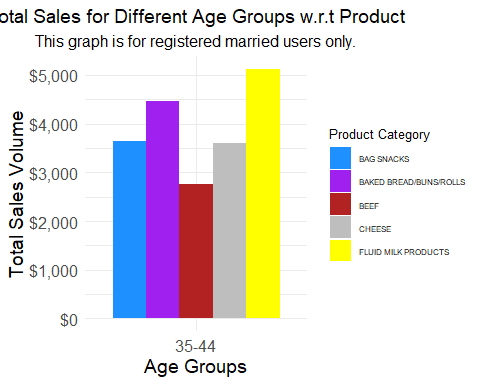
prod\_2 <- transactions %>%  
 inner\_join(demographics, by = 'household\_id') %>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married", age == "25-34") %>%  
 filter(  
 str\_detect(product\_category, regex("BAKED BREAD/BUNS/ROLLS|SOFT DRINKS", ignore\_case = TRUE)),  
 ) %>%  
 group\_by(product\_category, transaction\_timestamp, age) %>%  
 summarize(sales = sum(sales\_value)) %>%  
 arrange(desc(sales))  
  
  
prod\_2\_sales <- prod\_2 %>%  
 mutate(DayofWeek = weekdays(transaction\_timestamp)) %>%  
 group\_by(DayofWeek, product\_category) %>%  
 summarize(Total\_sales = sum(sales)) %>%  
 arrange(desc(DayofWeek))  
  
  
ggplot(prod\_2\_sales, aes(x = factor(DayofWeek, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")),   
 y = Total\_sales, color = product\_category, group = product\_category)) +  
 geom\_line() +  
 labs(title = "Sales Comparison for Bag Snacks and Soft Drinks",  
 x = "Day of the Week",  
 y = "Total Sales (USD)", # Added "USD" as y-axis label  
 color = "Product Category") +  
 theme\_minimal() +  
 scale\_color\_brewer(palette = "Set1") +  
 theme(  
 legend.position = "right",  
 panel.grid.major.y = element\_blank(),  
 panel.border = element\_blank()  
 ) +  
 scale\_y\_continuous(labels = dollar\_format(prefix = "$")) + # Format y-axis labels in thousands  
 guides(color = guide\_legend(title = "Product Category")) # Rename legend title



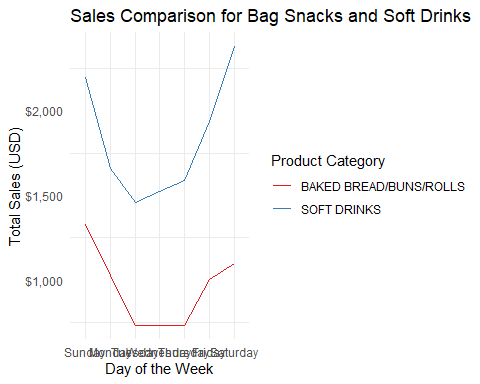
*In reference to the plot above, it is clear that the highest sales are consistently observed on Sunday, while the lowest sales occur on Thursdays. This observation underscores the importance of considering these sales trends when planning marketing and promotional activities throughout the week*

*Age group : 35-44*

category\_sales\_3 <- new\_df3 %>%  
 filter(age == "35-44") %>%  
 group\_by(unique\_prod, age) %>%  
 summarise(total\_sales = sum(Final\_sales)) %>%  
 arrange(desc(total\_sales))  
  
  
# Select the top 5 product categories  
top\_5\_categories\_3 <- category\_sales\_3 %>%  
 unnest(unique\_prod) %>%  
 top\_n(5, total\_sales) %>%  
 pull(unique\_prod)  
  
  
  
# Filter the data to include only the top 5 categories  
filtered\_data\_3 <- new\_df3 %>%  
 filter(unique\_prod %in% top\_5\_categories\_3, age == "35-44") %>%  
 unnest(unique\_prod) %>%  
 head(5)  
  
  
category\_colors\_2 <- c("BAG SNACKS" = "dodgerblue", "CHEESE" = "grey", "FLUID MILK PRODUCTS" = "yellow",  
 "BAKED BREAD/BUNS/ROLLS" = "purple", "BEEF" = "firebrick")  
  
# Create the ggplot visualization with the filtered data  
ggplot(filtered\_data\_3, aes(x = age, y = Final\_sales, fill = as.factor(unique\_prod))) +  
 geom\_bar(stat = "identity", position = "dodge") +   
 scale\_y\_continuous(name = "Total Sales Volume", labels = scales::dollar) +  
 labs(  
 title = "Total Sales for Different Age Groups w.r.t Product",  
 subtitle = "This graph is for registered married users only.",  
 x = "Age Groups",  
 y = "Total Sales",  
 fill = "Product Category"  
 ) +  
 scale\_fill\_manual(values = category\_colors\_2) +  
 theme\_minimal() +  
 theme(  
 legend.position = "right", # Move the legend to the right  
 plot.title = element\_text(size = 14, hjust = 0.5), # Adjust title font size and alignment  
 plot.subtitle = element\_text(size = 12, hjust = 0.5), # Adjust subtitle font size and alignment  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text = element\_text(size = 12),  
 legend.title = element\_text(size = 10),  
 legend.text = element\_text(size = 6)  
 )

 *The plot presented above reveals a noteworthy trend: customers within the age group of 35 to 44 consistently exhibit a high frequency of purchasing Baked Breads and Rolls. This insight highlights a specific buying pattern within this demographic.*

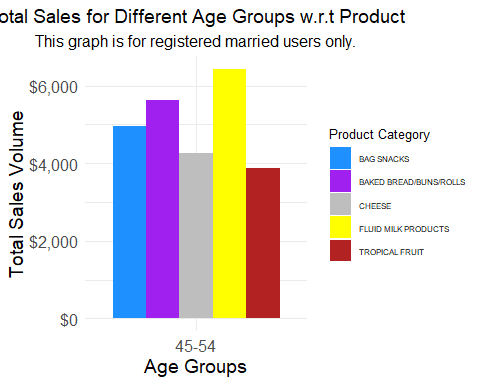
prod\_3 <- transactions %>%  
 inner\_join(demographics, by = 'household\_id') %>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married", age == "35-44") %>%  
 filter(  
 str\_detect(product\_category, regex("BAKED BREAD/BUNS/ROLLS|SOFT DRINKS", ignore\_case = TRUE)),  
 ) %>%  
 group\_by(product\_category, transaction\_timestamp, age) %>%  
 summarize(sales = sum(sales\_value)) %>%  
 arrange(desc(sales))  
  
  
prod\_3\_sales <- prod\_3 %>%  
 mutate(DayofWeek = weekdays(transaction\_timestamp)) %>%  
 group\_by(DayofWeek, product\_category) %>%  
 summarize(Total\_sales = sum(sales)) %>%  
 arrange(desc(DayofWeek))  
  
  
ggplot(prod\_3\_sales, aes(x = factor(DayofWeek, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")),   
 y = Total\_sales, color = product\_category, group = product\_category)) +  
 geom\_line() +  
 labs(title = "Sales Comparison for Bag Snacks and Soft Drinks",  
 x = "Day of the Week",  
 y = "Total Sales (USD)", # Added "USD" as y-axis label  
 color = "Product Category") +  
 theme\_minimal() +  
 scale\_color\_brewer(palette = "Set1") +  
 theme(  
 legend.position = "right",  
 panel.grid.major.y = element\_blank(),  
 panel.border = element\_blank()  
 ) +  
 scale\_y\_continuous(labels = dollar\_format(prefix = "$")) + # Format y-axis labels in thousands  
 guides(color = guide\_legend(title = "Product Category")) # Rename legend title



*With reference to the plot presented above, it is evident that Saturday consistently registers the highest sales, while Thursdays consistently report the lowest sales. Understanding these sales trends can be pivotal in strategically planning marketing efforts and promotions throughout the week.*

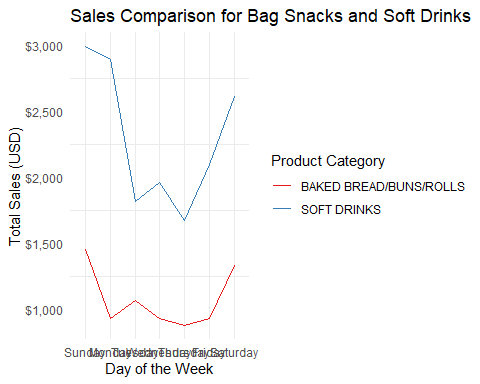
*Age group : 45-54*

category\_sales\_4 <- new\_df3 %>%  
 filter(age == "45-54") %>%  
 group\_by(unique\_prod, age) %>%  
 summarise(total\_sales = sum(Final\_sales)) %>%  
 arrange(desc(total\_sales))  
  
  
# Select the top 5 product categories  
top\_5\_categories\_4 <- category\_sales\_4 %>%  
 unnest(unique\_prod) %>%  
 top\_n(5, total\_sales) %>%  
 pull(unique\_prod)  
  
  
# Filter the data to include only the top 5 categories  
filtered\_data\_4 <- new\_df3 %>%  
 filter(unique\_prod %in% top\_5\_categories\_4, age == "45-54") %>%  
 unnest(unique\_prod) %>%  
 head(5)  
  
  
category\_colors\_3 <- c("BAG SNACKS" = "dodgerblue", "CHEESE" = "grey", "FLUID MILK PRODUCTS" = "yellow",  
 "BAKED BREAD/BUNS/ROLLS" = "purple", "TROPICAL FRUIT" = "firebrick")  
  
# Create the ggplot visualization with the filtered data  
ggplot(filtered\_data\_4, aes(x = age, y = Final\_sales, fill = as.factor(unique\_prod))) +  
 geom\_bar(stat = "identity", position = "dodge") +   
 scale\_y\_continuous(name = "Total Sales Volume", labels = scales::dollar) +  
 labs(  
 title = "Total Sales for Different Age Groups w.r.t Product",  
 subtitle = "This graph is for registered married users only.",  
 x = "Age Groups",  
 y = "Total Sales",  
 fill = "Product Category"  
 ) +  
 scale\_fill\_manual(values = category\_colors\_3) +  
 theme\_minimal() +  
 theme(  
 legend.position = "right", # Move the legend to the right  
 plot.title = element\_text(size = 14, hjust = 0.5), # Adjust title font size and alignment  
 plot.subtitle = element\_text(size = 12, hjust = 0.5), # Adjust subtitle font size and alignment  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text = element\_text(size = 12),  
 legend.title = element\_text(size = 10),  
 legend.text = element\_text(size = 6)  
 )

 *Our analysis of the plot depicted above indicates a clear and consistent trend: customers falling within the age bracket of 45 to 54 display a consistent pattern of frequent purchases of Baked Breads and Rolls. This noteworthy buying behavior is characteristic of this particular demographic group.*

*Day of Week comparison*

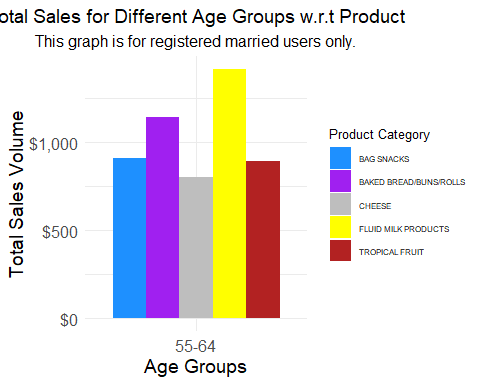
prod\_4 <- transactions %>%  
 inner\_join(demographics, by = 'household\_id') %>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married", age == "45-54") %>%  
 filter(  
 str\_detect(product\_category, regex("BAKED BREAD/BUNS/ROLLS|SOFT DRINKS", ignore\_case = TRUE)),  
 ) %>%  
 group\_by(product\_category, transaction\_timestamp, age) %>%  
 summarize(sales = sum(sales\_value)) %>%  
 arrange(desc(sales))  
  
  
prod\_4\_sales <- prod\_4 %>%  
 mutate(DayofWeek = weekdays(transaction\_timestamp)) %>%  
 group\_by(DayofWeek, product\_category) %>%  
 summarize(Total\_sales = sum(sales)) %>%  
 arrange(desc(DayofWeek))  
  
  
ggplot(prod\_4\_sales, aes(x = factor(DayofWeek, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")),   
 y = Total\_sales, color = product\_category, group = product\_category)) +  
 geom\_line() +  
 labs(title = "Sales Comparison for Bag Snacks and Soft Drinks",  
 x = "Day of the Week",  
 y = "Total Sales (USD)", # Added "USD" as y-axis label  
 color = "Product Category") +  
 theme\_minimal() +  
 scale\_color\_brewer(palette = "Set1") +  
 theme(  
 legend.position = "right",  
 panel.grid.major.y = element\_blank(),  
 panel.border = element\_blank()  
 ) +  
 scale\_y\_continuous(labels = dollar\_format(prefix = "$")) + # Format y-axis labels in thousands  
 guides(color = guide\_legend(title = "Product Category")) # Rename legend title



*Observing the plot above, it is apparent that Sunday consistently records the highest sales, while Thursdays consistently exhibit the lowest sales. Recognizing these sales patterns can inform strategic decisions regarding marketing and promotional activities scheduled throughout the week.*

*Age group : 55-64*

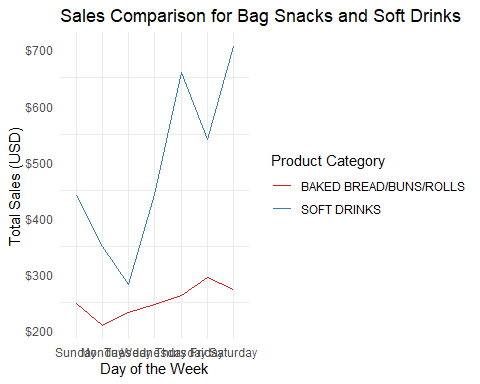
category\_sales\_5 <- new\_df3 %>%  
 filter(age == "55-64") %>%  
 group\_by(unique\_prod, age) %>%  
 summarise(total\_sales = sum(Final\_sales)) %>%  
 arrange(desc(total\_sales))  
  
  
# Select the top 5 product categories  
top\_5\_categories\_5 <- category\_sales\_5 %>%  
 unnest(unique\_prod) %>%  
 top\_n(5, total\_sales) %>%  
 pull(unique\_prod)  
  
  
# Filter the data to include only the top 5 categories  
filtered\_data\_5 <- new\_df3 %>%  
 filter(unique\_prod %in% top\_5\_categories\_5, age == "55-64") %>%  
 unnest(unique\_prod) %>%  
 head(5)  
  
  
category\_colors\_3 <- c("BAG SNACKS" = "dodgerblue", "CHEESE" = "grey", "FLUID MILK PRODUCTS" = "yellow",  
 "BAKED BREAD/BUNS/ROLLS" = "purple", "TROPICAL FRUIT" = "firebrick")  
  
# Create the ggplot visualization with the filtered data  
ggplot(filtered\_data\_5, aes(x = age, y = Final\_sales, fill = as.factor(unique\_prod))) +  
 geom\_bar(stat = "identity", position = "dodge") +   
 scale\_y\_continuous(name = "Total Sales Volume", labels = scales::dollar) +  
 labs(  
 title = "Total Sales for Different Age Groups w.r.t Product",  
 subtitle = "This graph is for registered married users only.",  
 x = "Age Groups",  
 y = "Total Sales",  
 fill = "Product Category"  
 ) +  
 scale\_fill\_manual(values = category\_colors\_3) +  
 theme\_minimal() +  
 theme(  
 legend.position = "right", # Move the legend to the right  
 plot.title = element\_text(size = 14, hjust = 0.5), # Adjust title font size and alignment  
 plot.subtitle = element\_text(size = 12, hjust = 0.5), # Adjust subtitle font size and alignment  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text = element\_text(size = 12),  
 legend.title = element\_text(size = 10),  
 legend.text = element\_text(size = 6)  
 )



*The plot above provides a distinct insight: customers in the age range of 55 to 64 consistently exhibit a high frequency of purchasing Baked Breads and Rolls. This observation underscores a specific purchasing trend within this age group*

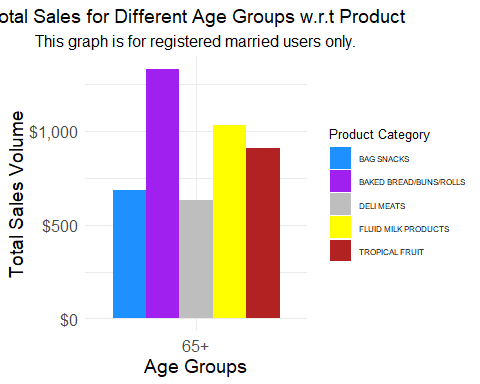
*Day of Week Comparison*

prod\_5 <- transactions %>%  
 inner\_join(demographics, by = 'household\_id') %>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married", age == "55-64") %>%  
 filter(  
 str\_detect(product\_category, regex("BAKED BREAD/BUNS/ROLLS|SOFT DRINKS", ignore\_case = TRUE)),  
 ) %>%  
 group\_by(product\_category, transaction\_timestamp, age) %>%  
 summarize(sales = sum(sales\_value)) %>%  
 arrange(desc(sales))  
  
  
prod\_5\_sales <- prod\_5 %>%  
 mutate(DayofWeek = weekdays(transaction\_timestamp)) %>%  
 group\_by(DayofWeek, product\_category) %>%  
 summarize(Total\_sales = sum(sales)) %>%  
 arrange(desc(DayofWeek))  
  
  
ggplot(prod\_5\_sales, aes(x = factor(DayofWeek, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")),   
 y = Total\_sales, color = product\_category, group = product\_category)) +  
 geom\_line() +  
 labs(title = "Sales Comparison for Bag Snacks and Soft Drinks",  
 x = "Day of the Week",  
 y = "Total Sales (USD)", # Added "USD" as y-axis label  
 color = "Product Category") +  
 theme\_minimal() +  
 scale\_color\_brewer(palette = "Set1") +  
 theme(  
 legend.position = "right",  
 panel.grid.major.y = element\_blank(),  
 panel.border = element\_blank()  
 ) +  
 scale\_y\_continuous(labels = dollar\_format(prefix = "$")) + # Format y-axis labels in thousands  
 guides(color = guide\_legend(title = "Product Category")) # Rename legend title

 *Based on the data presented in the plot above, it is evident that Sunday consistently represents the day with the highest sales, while Tuesday consistently exhibit the lowest sales. Understanding these sales trends is crucial for planning effective marketing and promotional strategies tailored to specific days of the week.*

*Age group : 65+*

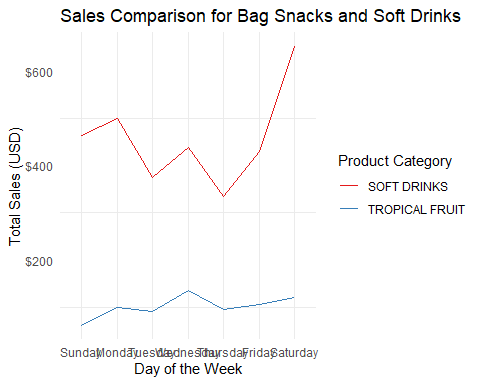
category\_sales\_6 <- new\_df3 %>%  
 filter(age == "65+") %>%  
 group\_by(unique\_prod, age) %>%  
 summarise(total\_sales = sum(Final\_sales)) %>%  
 arrange(desc(total\_sales))  
  
# Select the top 5 product categories  
top\_5\_categories\_6 <- category\_sales\_6 %>%  
 unnest(unique\_prod) %>%  
 top\_n(5, total\_sales) %>%  
 pull(unique\_prod)  
  
  
# Filter the data to include only the top 5 categories  
filtered\_data\_6 <- new\_df3 %>%  
 filter(unique\_prod %in% top\_5\_categories\_6, age == "65+") %>%  
 unnest(unique\_prod) %>%  
 head(5)  
  
  
category\_colors\_4 <- c("BAG SNACKS" = "dodgerblue", "DELI MEATS" = "grey", "FLUID MILK PRODUCTS" = "yellow",  
 "BAKED BREAD/BUNS/ROLLS" = "purple", "TROPICAL FRUIT" = "firebrick")  
  
# Create the ggplot visualization with the filtered data  
ggplot(filtered\_data\_6, aes(x = age, y = Final\_sales, fill = as.factor(unique\_prod))) +  
 geom\_bar(stat = "identity", position = "dodge") +   
 scale\_y\_continuous(name = "Total Sales Volume", labels = scales::dollar) +  
 labs(  
 title = "Total Sales for Different Age Groups w.r.t Product",  
 subtitle = "This graph is for registered married users only.",  
 x = "Age Groups",  
 y = "Total Sales",  
 fill = "Product Category"  
 ) +  
 scale\_fill\_manual(values = category\_colors\_4) +  
 theme\_minimal() +  
 theme(  
 legend.position = "right", # Move the legend to the right  
 plot.title = element\_text(size = 14, hjust = 0.5), # Adjust title font size and alignment  
 plot.subtitle = element\_text(size = 12, hjust = 0.5), # Adjust subtitle font size and alignment  
 axis.title.x = element\_text(size = 14),  
 axis.title.y = element\_text(size = 14),  
 axis.text = element\_text(size = 12),  
 legend.title = element\_text(size = 10),  
 legend.text = element\_text(size = 6)  
 )



*Analyzing the plot presented above, it becomes evident that customers aged 65 and above frequently opt for purchasing tropical fruits. This finding highlights a notable purchasing pattern within this specific age group.*

*Day of Week Comparison*

prod\_6 <- transactions %>%  
 inner\_join(demographics, by = 'household\_id') %>%  
 inner\_join(products, by = 'product\_id') %>%  
 filter(marital\_status == "Married", age == "65+") %>%  
 filter(  
 str\_detect(product\_category, regex("TROPICAL FRUIT|SOFT DRINKS", ignore\_case = TRUE)),  
 ) %>%  
 group\_by(product\_category, transaction\_timestamp, age) %>%  
 summarize(sales = sum(sales\_value)) %>%  
 arrange(desc(sales))  
  
  
prod\_6\_sales <- prod\_6 %>%  
 mutate(DayofWeek = weekdays(transaction\_timestamp)) %>%  
 group\_by(DayofWeek, product\_category) %>%  
 summarize(Total\_sales = sum(sales)) %>%  
 arrange(desc(DayofWeek))  
  
  
ggplot(prod\_6\_sales, aes(x = factor(DayofWeek, levels = c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday")),   
 y = Total\_sales, color = product\_category, group = product\_category)) +  
 geom\_line() +  
 labs(title = "Sales Comparison for Bag Snacks and Soft Drinks",  
 x = "Day of the Week",  
 y = "Total Sales (USD)", # Added "USD" as y-axis label  
 color = "Product Category") +  
 theme\_minimal() +  
 scale\_color\_brewer(palette = "Set1") +  
 theme(  
 legend.position = "right",  
 panel.grid.major.y = element\_blank(),  
 panel.border = element\_blank()  
 ) +  
 scale\_y\_continuous(labels = dollar\_format(prefix = "$")) + # Format y-axis labels in thousands  
 guides(color = guide\_legend(title = "Product Category")) # Rename legend title



*Based on the data depicted in the plot above, it is evident that Saturday consistently registers the highest sales, while Thursdays consistently exhibit the lowest sales. Understanding these sales patterns can be instrumental in formulating effective marketing and promotional strategies tailored to different days of the week.*

*Summarizing the problem statement.*

Through our analysis we aimed to study the buying behavior of married individuals in pursuit of understanding the products frequently brought by them to provide personalized recommendation for every age group to boost sales.

The core objective of our analysis was to gain a profound understanding of the buying behavior exhibited by married individuals. Our primary focus was to identify the products that are frequently purchased by this demographic and leverage this insight to provide personalized recommendations tailored to each age group.

The goal was to enhance sales and customer satisfaction through data-driven strategies.

*Summarize how you addressed this problem statement.*

In addressing the problem statement, we employed a systematic approach leveraging transactions, demographic, and product data to gain insights into customer behavior and formulate strategies for boosting sales. Here is a summary of how we tackled the problem:

1. Data Integration: We started by integrating the transactions, demographic, and product data to create a comprehensive dataset that allowed us to correlate customer information with purchasing behavior.
2. Age Group Analysis: We segmented customers into different age groups, recognizing that buying behavior can vary significantly across these demographics.
3. Frequent Product Analysis: Within each age group, we identified products that were frequently purchased together. This insight allowed us to understand customer preferences and associations between different items.
4. Day-of-Week Analysis: We examined sales patterns for the identified product combos and products. This analysis highlighted specific days of the week when these products experienced high and low sales.
5. Combo Creation: Based on the frequent product associations and day-of-week sales patterns, we proceeded to create product combinations that were well-suited for marketing campaigns. These combos aimed to capitalize on customer preferences and timing.
6. Marketing Recommendations: With the product combos in place, we formulated marketing recommendations to effectively promote these combos on days when sales were lower, maximizing sales potential.

*Summarize the interesting insights that your analysis provided.*

One of the most compelling insights derived from our analysis is the consistent and widespread popularity of “SOFT DRINKS” across all age groups. Regardless of demographic differences, Soft Drinks emerge as the top-selling product, emphasizing its universal appeal and consumer preference.

Furthermore, our analysis has yielded products for different age group which has been frequently purchased alongside Soft Drinks:

• Age Group 19 - 24: Bags of Chips - This demographic exhibits a propensity to purchasing Chips. • Age Group 25 - 34: Baked Breads and Rolls - For customers aged 25 to 34, the data indicates Baked Breads and Rolls are frequently purchased. • Age Groups 35 - 64: Baked Breads and Rolls - Interestingly, the analysis reveals that customers in the age groups of 35 to 64 frequently bought Baked Breads and Rolls. • Age Group 65+: Baked Breads and Rolls - the age groups of 65 and above exhibit a preference for Tropical fruits.

Recommendation to the Regork CEO?

Our overall takeaway is that by analyzing buying patterns of married populous we can identify products which are frequently bought together. Studying this behavior and effectively creating personalized combos of frequently bought items can drive more sales of these products. Furthermore, utilizing our analysis on when these products are purchased the most and when to the sales for these products dip, we can effectively create marketing campaign specifically targeting the days which see a dip in the sales of these products to increase footfalls of customers on this day.

Based on our Analysis we make the following recommendations to the CEO of Regork.

For age group 19 – 24 : Creating a combo of Soft Drinks and Bag of Chips and providing discount to these products on Monday For age group 25 – 34 : Creating a combo of Soft Drinks and Baked Breads and rolls and providing discount to these products on Monday. For age group 35 – 44 : Creating a combo of Soft Drinks and Baked Breads and rolls and providing discount to these products on Monday For age group 45 – 54 : Creating a combo of Soft Drinks and Baked Breads and rolls and providing discount to these products on Monday For age group 55 – 64 : Creating a combo of Soft Drinks and Baked Breads and rolls and providing discount to these products on Monday For age group 65+: Creating a combo of Soft Drinks and Baked Breads and rolls and providing discount to these products on Monday.

*The implications of our analysis on consumer behaviour are as follows:*

1. Personalized Combos: Consumers can expect to see personalized product combinations that align with their preferences and age group. This means they are more likely to find and purchase products they frequently buy together, enhancing their shopping experience.
2. Discount Opportunities: By creating combos and offering discounts, consumers can enjoy cost savings when purchasing frequently paired items. This incentivizes them to buy these items together, thus maximizing the value of their purchases.
3. Optimized Shopping Days: Our analysis identifies specific days when certain product combos are more likely to be on sale. Consumers can plan their shopping trips accordingly to take advantage of discounts and potentially less crowded shopping days.

Recommendations to the Regork CEO:

Based on our analysis, we propose the following recommendations to the CEO of Regork:

• For Age Group 19 - 24: Create a combo of Soft Drinks and Bag of Chips and offer discounts on these products, particularly on Mondays, to attract and cater to the preferences of this age group.

• For Age Group 25 - 34: Develop a combo of Soft Drinks and Baked Breads and Rolls, with discounts available on Mondays, to resonate with the buying behavior of customers in this age bracket.

• For Age Groups 35 - 64: Introduce a combo of Soft Drinks and Baked Breads and Rolls, accompanied by discounts on Mondays, to target a broad range of customers within these age groups.

• For Age Group 65+: Create a combo of Soft Drinks and Baked Breads and Rolls, along with discounts on Mondays, to cater to the preferences of the elderly demographic.

*By implementing these recommendations, Regork can enhance its marketing strategies, boost sales, and provide a more personalized and cost-effective shopping experience for its customers, ultimately improving customer satisfaction and loyalty.*