Introduction

This R Markdown document outlines the initial findings from the analysis of customer purchasing trends and behaviors, specifically focusing on chip purchasing behavior. The goal is to provide data-supported insights for the upcoming category review, with a keen interest in customer segments and their chip purchasing behavior. This document includes data loading, high-level data checks, feature engineering, and analysis of key metrics to describe customer purchasing behavior.

Load Required Libraries and Datasets

The following R libraries are loaded for data manipulation, analysis, and visualization. Please ensure these packages are installed in your R environment. If you encounter an error like "there is no package called 'package_name'", please uncomment the corresponding `install.packages()` line and run it in your R console first.

Code snippet
```
# Example code to install packages (uncomment and run if needed)
# install.packages("data.table")
# install.packages("ggplot2")
# install.packages("lubridate")
# install.packages("stringr")
# install.packages("dplyr")

# Load required Libraries
library(data.table)
library(ggplot2)
library(lubridate)
library(stringr)
library(dplyr)

# Point to where you have downloaded the datasets.
# Ensure 'QVI_transaction_data.csv' and 'QVI_purchase_behaviour.csv' are in the same directory as this .Rmd file,
# or provide the full path to the files.
transaction_data <- fread("QVI_transaction_data.csv")
purchase_behaviour <- fread("QVI_purchase_behaviour.csv")
```

Exploratory Data Analysis and Data Preparation

The first step in any analysis is to understand and prepare the data. We will perform high-level data checks, identify and handle outliers, correct data formats, and derive new features.

Examining Transaction Data

Let's check the structure and a sample of the transaction data. We will also convert the 'DATE' column to a proper date format and extract 'PACK_SIZE' and 'BRAND_NAME' from the 'PROD_NAME' column.

Code snippet
```
# Convert DATE to Date format (assuming Excel date format: 1899-12-30 as the epoch)
transaction_data[, DATE := as.Date(DATE, origin = "1899-12-30")]

# Extract pack size *before* cleaning PROD_NAME further
# Extracts the numerical part before 'g' or 'G' (e.g., '175' from '175g')
transaction_data[, PACK_SIZE := as.numeric(str_extract(PROD_NAME, "\\d+(?=gG|g|G)"))]

# Clean PROD_NAME: Remove digits and 'gG', then special characters, and finally trim whitespace
transaction_data[, PROD_NAME := str_replace_all(PROD_NAME, "\\d+[gG]", "")]
transaction_data[, PROD_NAME := str_replace_all(PROD_NAME, "[^[:alnum:] ]", "")]
transaction_data[, PROD_NAME := str_trim(PROD_NAME)]

# Extract brand name (first word of PROD_NAME)
transaction_data[, BRAND_NAME := word(PROD_NAME, 1)]

# Display info and head to verify changes
cat("Transaction Data Structure after initial cleaning and feature engineering:\n")
str(transaction_data)
cat("\nTransaction Data Head after initial cleaning and feature engineering:\n")
print(head(transaction_data))
```

Outlier Detection and Removal

We will check for outliers in the `PROD_QTY` column. A quantity of 200 is highly unusual for a single transaction and might indicate data entry error.

Code snippet
```
# Identify transactions with PROD_QTY = 200
cat("Transactions with PROD_QTY = 200:\n")
print(transaction_data[PROD_QTY == 200])

# Filter out the outlier transaction (PROD_QTY = 200)
transaction_data <- transaction_data[PROD_QTY != 200]
cat("\nOutlier transaction (PROD_QTY = 200) has been removed.\n")
```

Merging Datasets

Now, we merge the transaction data with the customer purchase behavior data using `LYLTY_CARD_NBR` to link transactions to customer segments.

Code snippet

```
# Merge the two dataframes on LYLTY_CARD_NBR
merged_data <- merge(transaction_data, purchase_behaviour, by = "LYLTY_CARD_NBR",
all.x = TRUE)

cat("\n--- Merged Data Structure after preprocessing ---\n")
str(merged_data)
cat("\n--- Merged Data Head after preprocessing ---\n")
print(head(merged_data))
```

Customer Purchasing Behavior Analysis

We will define key metrics to describe customer purchasing behavior and analyze these across different customer segments (`LIFESTAGE` and `PREMIUM_CUSTOMER`).

Calculating Key Metrics

Code snippet

```
# Calculate total sales by LIFESTAGE and PREMIUM_CUSTOMER
sales_by_segment <- merged_data %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(TOT_SALES = sum(TOT_SALES)) %>%
  ungroup()

cat("\nTotal Sales by Segment:\n")
print(sales_by_segment)

# Calculate number of unique customers by LIFESTAGE and PREMIUM_CUSTOMER
customers_by_segment <- merged_data %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(`Number of Customers` = n_distinct(LYLTY_CARD_NBR)) %>%
  ungroup()

cat("\nNumber of Customers by Segment:\n")
print(customers_by_segment)

# Calculate average price per unit (total sales / total quantity)
merged_data[, PRICE_PER_UNIT := TOT_SALES / PROD_QTY]

avg_price_per_unit_segment <- merged_data %>%
  group_by(LIFESTAGE, PREMIUM_CUSTOMER) %>%
  summarise(PRICE_PER_UNIT = mean(PRICE_PER_UNIT, na.rm = TRUE)) %>%
  ungroup()
```

```
cat("\nAverage Price Per Unit by Segment:\n")
print(avg_price_per_unit_segment)
```

Visualizing Purchasing Trends

Visualizations provide a clearer understanding of the purchasing trends across different customer segments.

Total Sales by Segment

Code snippet

```
ggplot(sales_by_segment, aes(x = LIFESTAGE, y = TOT_SALES, fill =
PREMIUM_CUSTOMER)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Total Sales by Lifestage and Premium Customer",
       x = "Lifestage",
       y = "Total Sales ($)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_viridis_d()

# The plot is automatically embedded in the PDF when knitted.
# ggsave() is typically used for saving plots to an external file, not strictly necessary for Rmd
output.
# ggsave("total_sales_by_segment.png", width = 12, height = 6)
```

Number of Customers by Segment

Code snippet

```
ggplot(customers_by_segment, aes(x = LIFESTAGE, y = `Number of Customers`, fill =
PREMIUM_CUSTOMER)) +
  geom_bar(stat = "identity", position = position_dodge()) +
  labs(title = "Number of Customers by Lifestage and Premium Customer",
       x = "Lifestage",
       y = "Number of Customers") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_viridis_d()

# ggsave("num_customers_by_segment.png", width = 12, height = 6)
```

Average Price Per Unit by Segment

Code snippet

```
ggplot(avg_price_per_unit_segment, aes(x = LIFESTAGE, y = PRICE_PER_UNIT, fill =
PREMIUM_CUSTOMER)) +
  geom_bar(stat = "identity", position = position_dodge()) +
```

```
  labs(title = "Average Price Per Unit by Lifestage and Premium Customer",
       x = "Lifestage",
       y = "Average Price Per Unit ($)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_viridis_d()

# ggsave("avg_price_per_unit_by_segment.png", width = 12, height = 6)
```

Conclusion

This section would typically contain a summary of findings and strategic recommendations based on the analysis. For example:

- **Key Segments for Sales:** Older Families (Budget), Older Singles/Couples (Mainstream & Budget), and Retirees (Mainstream) contribute significantly to total sales.
- **Customer Base Leaders:** Retirees (Mainstream) and Young Singles/Couples (Mainstream) have the largest customer bases.
- **Price Sensitivity Insights:** Young Singles/Couples (Mainstream) show a higher average price per unit, suggesting they are less price-sensitive or prefer premium products. Conversely, Budget and Premium segments within Young Singles/Couples have lower average price per unit.
- **Growth Opportunities:** New Families represent a segment with lower sales and customer numbers, indicating a potential area for targeted growth strategies.

A strategic recommendation would leverage these insights to propose actions for Julia, the Category Manager, such as tailored marketing campaigns, product placement adjustments, or new product development initiatives for specific high-value or high-potential customer segments.