```r
#Loading Libraries
library(data.table)
library(readr)
library(ggplot2)
library(ggmosaic)
library(readxl)

#Importing Datasets
filepath <- "~/Data Analysis Projects/"
transactionData <- read_excel(paste0(filepath,"QVI_transaction_data.xlsx"))
CustomerData <- fread(paste0(filepath,"QVI_purchase_behaviour.csv"))

#Exploratory Data Analysis
str(transactionData)
head(transactionData)
setDT(transactionData)

str(CustomerData)
head(CustomerData)

#Date Type Conversion
transactionData$DATE <- as.Date(transactionData$DATE,origin = "1899-12-30")
str(transactionData)

#Summary of Product Names in Transaction Data
summary(transactionData$PROD_NAME)
```

```
#Product Names Analysis
productWords <-
data.table(unlist(strsplit(as.character(unique(transactionData[,PROD_NAME])),"\\s+")))
setnames(productWords,'words')
View(productWords)

productWords_withunwanted <- grepl('[0-9]|[&,"]',productWords$words)
productWords_cleaned <- productWords[!productWords_withunwanted]
View(productWords_cleaned)

word_counts <- productWords_cleaned[,.N,by=words][order(-N)]
View(word_counts)

#Removing Salsa Products
transactionData[, Salsa := grepl("salsa",tolower(PROD_NAME))]
transactionData <- transactionData[Salsa==FALSE, ][ ,Salsa:=NULL]
View(transactionData)

#Summary of Transaction Data
summary(transactionData)
#Outlier Detection
transactionData[PROD_QTY == 200]
transactionData[LYLTY_CARD_NBR == 226000]
#Removing Outlier
transactionData <- transactionData[LYLTY_CARD_NBR!=226000]
```

```
#Reexamine the Data
summary(transactionData)
#Transactions over Time
transaction_counts <- transactionData[,.N,by=DATE]
View(transaction_counts[order(DATE)])

#Transactions Distribution
theme_set(theme_bw())
theme_update(plot.title = element_text(hjust=0.5))

ggplot(transaction_counts,aes(x=DATE,y=N)) +
  geom_line() +
  labs(x="Day",y="Transaction Count",title="Transactions over time") +
  scale_x_date(breaks = '1 month') +
  theme(axis.text.x = element_text(angle = 90,vjust = 0.5))

ggplot(subset(transaction_counts,between(DATE,"2018-12-01","2018-12-31")),aes(x=DATE,y=N)) +
  geom_line() +
  labs(x="Day",y="Transaction Count",title="Transactions over December") +
  scale_x_date(breaks = '1 week')
  theme(axis.text.x = element_text(angle = 90,vjust = 0.5))

ggplot(subset(transaction_counts,between(DATE,"2018-12-21","2018-12-31")),aes(x=DATE,y=N)) +
  geom_line() +
  labs(x="Day",y="Transaction Count",title="Transactions over Christmas time") +
  scale_x_date(breaks = '1 day')
  theme(axis.text.x = element_text(angle = 45,hjust = 0.5))

#Chips Product Sizes
transactionData[, PACK_SIZE := parse_number(PROD_NAME)]
productSizes <- transactionData[, .N, by = PACK_SIZE][order(PACK_SIZE)]
View(productSizes)

#Product Sizes Frequency
ggplot(productSizes,aes(x=factor(PACK_SIZE),y=N)) +
  geom_col(fill = "steelblue") +
  labs(x="Sizes",y="Count",title="Product Sizes Distribution") +
  theme(axis.text.x = element_text(vjust = 0.5))

#Chips Product Brands
transactionData[, Brand_Name:= sub(" .*","",PROD_NAME)]
View(transactionData)
transactionData[Brand_Name == "Red", Brand_Name:="RRD"]
transactionData[Brand_Name == "WW", Brand_Name:="Woolworths"]
productBrands <- transactionData[, .N, by = Brand_Name][order(Brand_Name)]
View(productBrands)

#Customer Data analysis
summary(CustomerData)
head(CustomerData)
subscription_dist <- CustomerData[,.N, PREMIUM_CUSTOMER]
```

```
View(subscription_dist)
Lifestage_dift <- CustomerData[,.N, LIFESTAGE]
View(Lifestage_dift)

#Customers Premium Type Distribution
ggplot(subscription_dist,aes(x=reorder(PREMIUM_CUSTOMER,N),y=N)) +
  geom_col(fill = "darkgreen") +
  labs(
    title = "Customers Premium Type Distribution",
    x= "Premium Type",
    y= "No. of customers"
  )

#Families Distribution
ggplot(Lifestage_dift,aes(x=reorder(LIFESTAGE,N),y=N)) +
  geom_col(fill = "yellow") +
  labs(
    title = "Families Distribution",
    x= "Family Type",
    y= "No. of Families"
  ) +
  theme(axis.text.x = element_text(size=6))

#Merging Data
Data <- merge(transactionData,CustomerData,all.x = TRUE)
View(Data)
```

```
#Merge validation
dim(Data)
dim(transactionData)

#Null Check
sum(is.na(Data))

#Saving data in csv file
fwrite(Data,paste0(filepath,"QVI_data.csv"))

#Customer Segment analysis
#Total Sales distribution
sales_by_groups<-Data[,.(Totalsale=sum(TOT_SALES)),by=.(LIFESTAGE,PREMIUM_CUSTOMER)][order(-Totalsale)]
View(sales_by_groups)

ggplot(sales_by_groups,aes(x=LIFESTAGE,y=Totalsale,fill=PREMIUM_CUSTOMER)) +
 geom_bar(stat="identity",position = "dodge") +
 labs(
  title = "Total Chip Sales by Lifestage and Premium Segment",
  x= "Lifestage",
  y= "Total Sale",
  fill="Premium Customer"
 ) +
 theme_minimal()
```

```
#Customers Distribution
customers_by_groups<- Data[,.(TotalCustomers = uniqueN(LYLTY_CARD_NBR)),by=.(LIFESTAGE,PREMIUM_CUSTOMER)][order(-TotalCustomers)]
View(customers_by_groups)

ggplot(customers_by_groups,aes(x=LIFESTAGE,y=TotalCustomers,fill=PREMIUM_CUSTOMER)) +
  geom_bar(stat="identity",position = "dodge") +
  labs(
    title = "Total Customers by Lifestage and Premium Segment",
    x= "Lifestage",
    y= "Total Customers",
    fill="Premium Customer"
  ) +
  theme_minimal()

#Average Units Purchased by Customers
avgunits_by_customers<- Data[,.(AvgUnits = mean(PROD_QTY)),by=.(LIFESTAGE,PREMIUM_CUSTOMER)][order(-AvgUnits)]
View(avgunits_by_customers)

ggplot(avgunits_by_customers,aes(x=reorder(LIFESTAGE,-AvgUnits),y=AvgUnits,fill=PREMIUM_CUSTOMER)) +
  geom_bar(stat="identity",position = "dodge") +
  labs(
  title = "Average Units sold by Lifestage and Premium Segment",
  x= "Lifestage",
  y= "Avg Units",
  fill="Premium Customer"
 ) +
 theme_minimal()

#Average Sales Analysis
avgsale_by_customers<- Data[,.(Avgsale = mean(TOT_SALES)),by=.(LIFESTAGE,PREMIUM_CUSTOMER)][order(-Avgsale)]
View(avgsale_by_customers)

ggplot(avgsale_by_customers,aes(x=LIFESTAGE,y=Avgsale,fill=PREMIUM_CUSTOMER)) +
 geom_bar(stat="identity",position = "dodge") +
 labs(
  title = "Average Sale by Lifestage and Premium Segment",
  x= "Lifestage",
  y= "Avg Sale",
  fill="Premium Customer"
 ) +
 theme_minimal()

#Hypothesis Analysis to check the significance difference in avg unit price
#between Mainstream and Premium,Budget Customers who are Mid age or Young Singles/Couples
#H0: There is no significant difference
#H1: There is significant difference
```

```
Data[,UnitPrice:=TOT_SALES/PROD_QTY]

sum(is.na(Data$UnitPrice))

sum(is.infinite(Data$UnitPrice))

midage_lifestage_name <- "MIDAGE SINGLES/COUPLES"

young_lifestage_name  <- "YOUNG SINGLES/COUPLES"

mainstream_name      <- "Mainstream"

premium_name        <- "Premium"

budget_name         <- "Budget"


midage_main_prem <- Data[LIFESTAGE == midage_lifestage_name & PREMIUM_CUSTOMER %in% c(mainstream_name, premium_name)]

midage_main_budget <- Data[LIFESTAGE == midage_lifestage_name & PREMIUM_CUSTOMER %in% c(mainstream_name, budget_name)]

young_main_prem <- Data[LIFESTAGE == young_lifestage_name & PREMIUM_CUSTOMER %in% c(mainstream_name, premium_name)]

young_main_budget <- Data[LIFESTAGE == young_lifestage_name & PREMIUM_CUSTOMER %in% c(mainstream_name, budget_name)]


#Two sample Test

t_test_1 <- t.test(UnitPrice ~ PREMIUM_CUSTOMER, data = midage_main_prem)

print(t_test_1)

t_test_2 <- t.test(UnitPrice ~ PREMIUM_CUSTOMER, data = midage_main_budget)

print(t_test_2)

t_test_3 <- t.test(UnitPrice ~ PREMIUM_CUSTOMER, data = young_main_prem)

print(t_test_3)

t_test_4 <- t.test(UnitPrice ~ PREMIUM_CUSTOMER, data = young_main_budget)

print(t_test_4)
```

```
#Results

#All the 4 T-Tests p value < 2.2e-16. Since the p value is < 0.05, the null hypothesis

#is rejected. It is statistically proven that there is a significant difference

#in average unite price between Mainstream and Premium,Budget Customers.



#Proportional Analysis

####BRANDS####

#Brands preferred by Mainstream Mid age,Young Singles/Couples Customers

Data[,Istargetsegment := (LIFESTAGE == young_lifestage_name & PREMIUM_CUSTOMER == mainstream_name)]

View(Data)

brand_counts <- Data[,.N,by = .(Brand_Name,Istargetsegment)]

View(brand_counts)

total_counts <- Data[,.(TransactionCounts = .N),.(Istargetsegment)]

View(total_counts)

brand_proportions <- merge(brand_counts,total_counts,by="Istargetsegment")

brand_proportions[,Proportion := N/TransactionCounts]

View(brand_proportions)

proportion_comparison <- dcast(brand_proportions, Brand_Name ~ Istargetsegment, value.var = "Proportion")

View(proportion_comparison)

setnames(proportion_comparison,c(2,3),new=c("OtherProportion","TargetProportion"))

proportion_comparison[, PreferenceRatio := TargetProportion / OtherProportion]

View(proportion_comparison)
```

```
ggplot(proportion_comparison, aes(x = reorder(Brand,PreferenceRatio), y =
PreferenceRatio)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  coord_flip() +
  labs(title = "Brand Preference Ratio for Mainstream Young Singles/Couples",
       subtitle = "Ratio > 1 indicates higher preference by Target Segments",
       x = "Brand",
       y = "Preference Ratio (Target Proportion / Other Proportion)") +
  theme_minimal()
```

```
plot_data_long <- melt(proportion_comparison[, .(Brand, TargetProportion,
OtherProportion)],
                       id.vars = "Brand",
                       variable.name = "SegmentGroup",
                       value.name = "Proportion")
```

```
ggplot(plot_data_long, aes(x = reorder(Brand,-Proportion), y = Proportion, fill =
SegmentGroup)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Brand Purchase Proportion: Target Segment vs. Others",
       x = "Brand",
       y = "Proportion of Transactions") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
##INSIGHTS##

# The brand preference analysis shows that Mainstream Young Singles/Couples have

# a higher inclination towards brands like Tyrrells, Twisties, Doritos, and Tostitos,

# as indicated by their preference ratios being greater than 1. This suggests a stronger

# affinity for these brands compared to other segments. In contrast, brands like

# Smiths, Sunbites, and Woolworths show lower ratios, indicating they are

# less favored by this group. The trend hints at a preference for bold or

# premium-style brands within this segment.



####PACK SIZE####

#Preferred pack size by Mainstream Mid age,Young Singles/Couples Customers


packsize_counts <- Data[,.(Pack_counts = .N),.(PACK_SIZE,Istargetsegment)]

View(packsize_counts)


pack_proportions <- merge(packsize_counts,total_counts,by ="Istargetsegment")

pack_proportions[,Proportion := Pack_counts/TransactionCounts]

View(pack_proportions)


packProportion_comparision <- dcast(pack_proportions,PACK_SIZE ~ Istargetsegment,
value.var = "Proportion")

View(packProportion_comparision)


setnames(packProportion_comparision, c(2,3),c("OtherSegments", "TargetSegments"))
```

```
packProportion_comparision[,PreferenceRatio :=  TargetSegments/OtherSegments]


ggplot(packProportion_comparision, aes(x = reorder(PACK_SIZE,PreferenceRatio), y = PreferenceRatio)) +
 geom_bar(stat = "identity", fill = "skyblue") +
 coord_flip() +
 labs(title = "Pack Size Preference Ratio for Mainstream Young Singles/Couples",
     subtitle = "Ratio > 1 indicates higher preference by Target Segments",
     x = "Pack Size",
     y = "Preference Ratio (Target Proportion / Other Proportion)") +
 theme_minimal()

pack_data_long <- melt(packProportion_comparision[, .(PACK_SIZE, TargetSegments, OtherSegments)],
                 id.vars = "PACK_SIZE",
                 variable.name = "Segment_Type",
                 value.name = "Proportion"
                 )

ggplot(pack_data_long, aes(x = reorder(PACK_SIZE,-Proportion), y = Proportion, fill = Segment_Type)) +
 geom_bar(stat = "identity", position = "dodge") +
 scale_y_continuous(labels = scales::percent) +
 labs(title = "Pack Size Purchase Proportion: Target Segment vs. Others",
     x = "Pack Size",
     y = "Proportion of Transactions") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

##INSIGHTS##
#The pack size preference analysis reveals that this customer segment tends to
# prefer larger pack sizes such as 270g, 380g, and 330g, which might reflect bulk
# buying behavior or social consumption habits. Mid-range pack sizes (e.g., 135g–165g)
# are moderately preferred, while smaller packs (below 125g) are less favored.
# This indicates that Mainstream Young Singles/Couples are likely seeking better
# value or are purchasing for sharing occasions.
```