````
---
title: "Technical Appendix for General Mills"
author: "Adrian Alvarez Sanabria"
date: "`r format(Sys.time(), '%B %d, %Y')`"
output:
  html_document:
    toc: true
    toc_depth: 3
    toc_float: true
    number_sections: true
---
```{r echo=FALSE}
#Start date: April 10th, 2019
#Author: Adrian Alvarez
#Course: BUAN 3210
#Purpose: EDA and Technical Appendix, with html output
#Title: Technical Appendix for Office Salary Data
```

I was asked to perform a brief analysis on sales data from several anonymized stores, with the objective
of understanding the effectiveness of General Mills's current ad and promo strategy. The data set
contained 21850 observations with 11 different variables like ad, flavor, brand, etc.

```{r warning=FALSE, message=FALSE}
#I will clear existing environment of its packages, variables and functions
rm(list = ls(all = TRUE))
if(is.null(sessionInfo()$otherPkgs) == FALSE)lapply(paste("package:", names(sessionInfo()$otherPkgs),
sep=""), detach, character.only = TRUE, unload = TRUE)
#I will be using this set of packages, to better modify and visualize the data
library(tidyverse)#ggplot, dplyr, etc.
library(gridExtra)#grid.arrange to combine plots
library(psych)#correlation plots
library(janitor)#tidyverse functions for cross-tables
library(stringr)#functions to manipulate strings
```



#Organizing Data
##Importing Datasets
```{r warning=FALSE}
#Code chunk is importing both datasets
product_data<-read.csv("mtp_product_data.csv")
sales_data<-read.csv("mtp_sales_data.csv")

head(product_data)
head(sales_data)
#UPC number format does not match between data sets. Strings need to be manipulated in order to join
tables
```

- This code chunk imports the two data sets, it also returns a head of each data set

- Here we can notice that UPC numbers' formats do not match so the strings need to be manipulated.

##Manipulating UPCs and Joining Tables
```{r}
pupc<-product_data$UPC%>%#Manipulating the product_data UPC number to make it the same format as the
sales_data UPC number
  str_replace_all("-",".")
prod_upc<-mutate(product_data,UPC=pupc)#Mutating a string, to have the new formated product_data UPC.
Naming the new data set 'prod_upc'

supc<-sales_data$UPC%>%#Manipulating the sales_data UPC number to make it match the format of the new
product_data UPC number
str_replace("0","00.0")
sales_spu<-mutate(sales_data,UPC=supc)#Mutating the string to add the newly formated sales_data UPC into
the data set. Naming the new data set 'sales_spu'

prod_sal<- full_join(prod_upc,sales_spu, by="UPC", copy=FALSE)#Joining both datasets by UPC number

```

- This code chunk manipulates the UPC string in both data sets to match the the format

- Then it creates two new data sets with the matching UPC format

- Finally it joins both data sets to have all variables and observations in one table.


#Basic EDA
##Summarizing "prod_sal" dataset
```{r}
#Using summary(), head(), and string() to summarize the prod_sal data. Code returns count of factor
variables, and most descriptive stats of quantitative variables
summary(prod_sal)
head(prod_sal)
str(prod_sal)
```

Comments

- Iri-key is the store id so we will not use that variable

- Promo, volume, and week are all categorical variables treated as a quantitative so their descriptive
stats do not help us

- We should create a revenue(price*units) variable to look at flavors and brands generating the most
revenue

- Regular is the flavor with the most sales, closely followed by toasted

- Price variable seems to have a symmetrical distribution, there is a pretty small difference between
the median and the mean. Maybe the distribution has a light left skew

- Kelloggs frosted flakes have the most number of sales, followed by Kelloggs fruit loops

- Closest General Mills brand in sales is Cinnamon Toast Crunch

- The most sold flavor is regular, General Mills's Cinnamon Toast doesn't fall in that category


Questions

- What does the price distribution look like?

- How does the revenue or unit distribution look like?

- Does price vary across different flavors of cereal?

- Do ads and promos vary across flavors?

- Are the flavors with most sales the ones with more ads and promos?

##Univariate, Graphical-Categorical
```{r}
flavors1<-prod_sal$flavor%>%#Manipulating flavor string in the prod_sal data set to make a new variable
str_sub(1,2)#to substitute size of the string to two characters to have space in the graph

#Here we are going to use a grid arrange function to the categorical variables in one panel. Code
returns three bargraphs in two different columns.
grid.arrange(
  prod_sal%>%
    ggplot(aes(package)) +
      geom_bar()+
    theme(panel.background = element_blank()),#Theme code to get rid of the grey panel background.

  prod_sal%>%
    mutate(promotions=(promo>0)) %>%#Mutated variable to creat a true or false variable for promo
    ggplot(aes(promotions)) +
      geom_bar()+
    theme(panel.background = element_blank()),

  prod_sal%>%
    ggplot(aes(flavors1),legend(flavor)) +
      geom_bar()+
    coord_flip()+#flip chart coordinates for better view
    theme(legend.position="right")+
    theme(panel.background = element_blank()),

  prod_sal%>%
    ggplot(aes(ad)) +
      geom_bar()+
    theme(panel.background = element_blank()),


  ncol = 2
)
```

Comments

- Most units are packaged in boxes

- Probably no package differentiation regarding ad and promo strategy

- Less than 1/3 of all units have promotions

- Fruit, Cocoa, and cinnamon have very similar sales

- Most units have no ads or promos

- Should run t.tests on both promo, and ad variables to see if we have a representative sample of the
population


Questions

- What flavors and brands generate the most revenue?

- Do all flavors have the same amount of promos and advertising?

- Whats the most common form of advertisement? A or B?

##Univariate, Graphical-Quantitative
```{r}
# Code histograms using grid.arrange so can see the different quant. variables together
grid.arrange(

  prod_sal %>%
    ggplot(aes(price)) +
      geom_histogram(binwidth = .09)+
    theme(panel.background = element_blank()),#adjusted bandwidth to get a better idea of price's
distribution

  prod_sal %>%
    mutate(revenue=(price*units)) %>%#Making a revenue variable of of price and units
    ggplot(aes(revenue), bin) +
      geom_histogram(binwidth =.5)+#adjusting size of binwidth to improve visualization
    theme(panel.background = element_blank()),

  #To make a boxplot of a single continuous variable you have to set x = 1 since the boxplot does not
have a width associated with it so gives alignment.We use the boxplot to look at the median, 1st & 2nd
quartile, inner quartile range, and outliers
  ggplot(prod_sal, aes(x = 1, y = price)) +
    geom_boxplot()+
    theme(panel.background = element_blank()),

  ggplot(prod_sal, aes(x = 1, y = units)) +
    geom_boxplot()+
    theme(panel.background = element_blank()),

  ncol=2
)
```

Comments

- The price histogram does seem to have a normal distribution with some possible outlines. No or very
few cereals cost more than $8

- The box plot allows us to see that there are actually some outlines that are priced at more than $8


- Revenue distribution reminds us that cereal is a product of relatively low cost. So most sales revenue
less than $50

- The unit distribution would be the same as revenue, observations are just divided by price

- Since price of sales is very rarely more than $100, a price oriented promo or ad strategy might not be
effective.

- Want to look at promotions and ads across flavors and brands.


Questions

- Are ads and promos the same across flavors?

- Are revenues and ads or promos correlated?

##Multivariate-nongraphical
```{r}
grid.arrange(
  # Add margins, i.e. row and column totals
tableGrob(
 prod_sal%>%
    mutate(promotions=(promo>0)) %>%
    tabyl(promotions, flavor) %>%
    adorn_totals(where = c("row", "col")) %>%#Code adds a total row
    adorn_percentages(denominator = "all") %>% #Code returns as proportion table
    adorn_rounding(2)#code rounds to two decimal places
 ),

tableGrob(
 prod_sal%>%
    tabyl(ad, flavor) %>%
    adorn_totals(where = c("row", "col")) %>%
    adorn_percentages(denominator = "all") %>%
    adorn_rounding(2)
),

nrow=2
)
```

Comments

- Regular is the most discounted flavor, 20% of regular flavor units are discounted

- Followed by toasted, 21% of toasted flavored units are discounted

- Cinnamon Toast is the least discounted flavor

- 21% of all units are discounted, while only 12% have adds

- Regular and toasted flavors have about the same amount of A and B ads

- Cinnamon, Cocoa and Fruit only have the same amount of ads and only have A ads

- Together they have only 27% of all ads

- Amount and size of ads is not proportional to the number of sales by flavor.


Questions

- Does demand of cereal vary throughout the weeks of the year?

- Should there be there be a season oriented promo and ad strategy?

###Correlations
```{r warning=FALSE}
#Table shows the correlation between the different numeric variables
prod_sal %>%
  select(-iri_key)%>%
  mutate(revenue=(price*units)) %>%#mutates to create a revenue variable
  select_if(is.numeric)%>%#selects numeric variables for correlation
  cor() %>%
  round(2)
```

Comments

- Since units and week variables are correlated, we do not need to graph that the week of the year has
no impact on the units of cereal sold

- So there a season oriented ad or promo strategy might not be effective

- We can see that volume and price have a correlation of 50%, which makes sense because price is pretty
dependent on the amount of product that is being sold

- Similarly units and revenue have a 91% correlation because units=revenue/price

- Promo and units have a positive correlation of 20%, which tells us that the variation in promotions
can explain about 20% of the variation in units sold.


Questions

- Is it useful to look at promos and ads across different packages?

- Are ads and promos the same across brands?

##Multi-variate Graphical
###Mosaico Plots of Cross Tables
```{r}
#In this code chunk we use grid arrange to have both mosaics together. Mosaics allow us to get a better
idea of the number of certain observtions.
grid.arrange(
# Note have to calculate and provide a variable for filling the graph
prod_sal %>%
  filter((ad!="NONE"))%>%
  group_by(ad, flavor) %>%
  summarise(count = n()) %>%
  ggplot(aes(ad, flavor)) +
  geom_tile(aes(fill = -count))+
    theme(panel.background = element_blank()) ,

prod_sal %>%
  mutate(promotions=(promo>0)) %>%
  group_by(promotions, flavor) %>%
  summarise(count = n()) %>%
  ggplot(aes(promotions, flavor)) +
  geom_tile(aes(fill = -count))+
    theme(panel.background = element_blank()), # Put "-" in front of "count" to reverse color order
nrow=2
)

```

Comments

- Mosaic allows us to look at the count of only observations that have either A or B ad.

- Both toasted and regular have the most A ads

- Almost no B ads at all

- Mosaic charts make it clear both that most promotions fall in the toasted and regular flavors while
Fruit, Cocoa and Cinnamon toast have the same amount of advertisement

#Detailed EDA
##Revenue and Promo across Brands with Flavor Fill
```{r echo=FALSE, warning=FALSE}
brands1<-prod_sal$brand%>%#Manipulating brand string in the prod_sal data set to make a new variable
with less characters
str_sub(1,4)
# Code univariate bar graphs using grid.arrange so can see all quant variables together
summary(prod_sal)

grid.arrange(

  prod_sal %>%
 ggplot(aes(x=brands1, y=promo))+
  geom_bar(position="dodge", stat="summary", fun.y="sum")+
coord_flip()+
theme(panel.background = element_blank()),

  prod_sal %>%
 ggplot(aes(x=brands1, y=units))+
  geom_bar(position="dodge", stat="summary", fun.y="sum")+
coord_flip()+
theme(panel.background = element_blank()),

  ncol = 1
)

```

Comments

- Kelloggs is the brand with the most units with discounts and ads

- General Mills follows Kelloggs in number of units with ads and discounts

- General Mills is the most responsive brand to advertisement and promo

- Difference between General Mills's amount of discounts or adds in comparison to Kellogs is not
proportional to the difference in profits.

- The 'prod_sal' summary shows us that the most sold General Mill's products are: Cinnamon Toast Crunch,
Lucky Charms and Cheerios respectively

- None of those cereals fall in the regular category

- Regular is the flavor with the biggest market share

Questions

- Does General Mills keep promos the same across flavors?

- Should there be more promos for regular flavor?


```{r}
grid.arrange(

  prod_sal %>%
  mutate(revenue_div_by10=((price*units)/10)) %>%
 ggplot(aes(x=brands1, y=promo, fill=flavors1))+
  geom_bar(position="dodge", stat="summary", fun.y="sum")+
coord_flip()+
theme(panel.background = element_blank()),

  ncol = 1
)
```


Comments

- Cinnamon is General Mill's most discounted flavor

- Sales with most promotions fall in the regular flavor, followed by toasted. Most sales do not have
promotions.

- All flavors have some type of advertisement

- Post is competing strongly for regular flavor market

- Regular has the largest number of sales

- Toast is the flavor generating the most profit for General Mills

Questions

- Do ads vary across brands?


## Units with Ads Across Brands
Big and Small Ads (A and B)
```{r}
prod_sal %>%
  mutate(units_by10=(units/10)) %>%
  filter(ad=="A")%>%
  ggplot(aes(x=brand, y=units_by10, fill=ad))+
  geom_bar(position="dodge", stat="summary",fun.y="sum")+
coord_flip()+
theme(panel.background = element_blank())

prod_sal %>%
  mutate(units_by10=(units/10)) %>%
  filter(ad=="B")%>%
  ggplot(aes(x=brand, y=units_by10, fill=ad))+
  geom_bar(position="dodge", stat="summary",fun.y="sum")+
coord_flip()+
theme(panel.background = element_blank())

```

Comments

- Kelloggs is also the brand with the most ads, both A and B

- Cinnamon Toast crunch is General Mill's most advertised product

- It is also where they make most of their revenues

- Here we are able to see the units sold of the flavors with and without promo, or with a,b or no ad

- This graphs give us a better idea of the distribution of ads happening outside the regular and toasted
flavors


#Stat EDA
##T-test for 'Ad' Variable
```{r warning=FALSE}

t.test(prod_sal$units[prod_sal$ad=="A"],
       prod_sal$units[prod_sal$ad=="B"])

t.test(prod_sal$price[prod_sal$ad=="A"],
       prod_sal$price[prod_sal$ad=="B"])
```

- Null Hypothesis: units sold are the same across ads

- Alternative Hypothesis: units sold are are different across different ads

- Technical: zero is not in the confidence interval so we reject the null hypothesis in favor of the
alternative, and units sold vary across different ad types.

- We are 95% confident that observations in this sample fall within the normal distribution.


##T-test for Promo Variable
```{r warning=FALSE, message=FALSE}

t.test(prod_sal$units[prod_sal$promo=="1"])

t.test(prod_sal$price[prod_sal$promo=="1"])
```
- Null Hypothesis: units sold are the same across ads

- Alternative Hypothesis: units sold are are different across different ads

- Technical: zero is not in the confidence interval so we reject the null hypothesis in favor of the
alternative and units sold vary across different promos.

- We are 95% confident that observations in this sample fall within the normal distribution.
````