## James Machado, Justin Kaplan, Jonny Shakerchi

## BUS256 Marketing Analytics Spring 2017

### Case: SVEDKA Vodka (C): Marketing Mix in the Vodka Industry

**Introduction:**  The SVEDKA case series examines one industry from the perspective of a single firm that wants to position a new product. This assignment provides hands-on experience building a marketing mix model for the vodka industry.

**Objectives**:

* Apply your understanding of multiple regression to a marketing mix problem
* Gain practice in drawing conclusions for a firm, relying on regression models.

**Summary**: The (A) case of SVEDKA gave us a chance to see how a managing director of a firm thinks about introducing a potentially disruptive new product into a fairly mature market. What began simply as an idea developed into something much larger, and required a thorough understanding of the competitive market.

The C case provides us with a large dataset and the challenge of investigating the data to understand the relative importance of the four elements of the Marketing Mix (informally known as the “4 P’s”): Product, Price, Advertising (“Promotion” or Communication), and Distribution (“Place”).

Before starting this assignment, review the A case and read the very short C case closely. Also, familiarize your self with the Data Dictionary tab in the SVEDKA Excel file before downloading the data. All of the data you require is in the spreadsheet already.

Your submission should be created in R Markdown. Please knit the file to Word, and then re-save the Word doc as a pdf. You should upload the pdf.

**Tasks and Questions: FOR EACH model that you run, show the code and the regression summary.**

# James Mac Working Directory

setwd("~/Dropbox/BA MBA/Spring 2017/BUS 256A Mkt Analytics/Assignments/Svedka Analysis/Svedka\_Analysis")

# Import "Svedka data.csv"

library(readr)

Svedka\_data <- read\_csv("Svedka data.csv")

View(Svedka\_data)

# James: My csv is giving me a bunch of unecessary columns at the end, so I need to remove them.

# You may not need to do this if you use your own csv file

Svedka\_data = subset(Svedka\_data, select = -c(X68, X69, X70, X71, X72, X73, X74, X75) )

# There is a typo in the PricePerUnit column -- the column in listed as 'PriceRerUnit'

library(plyr)

Svedka\_data <- rename(Svedka\_data, c("PriceRerUnit"="PricePerUnit"))

1. (5) Create a simple histogram of our target variable, **TotalSales**. Comment on the shape of the distribution that you see.

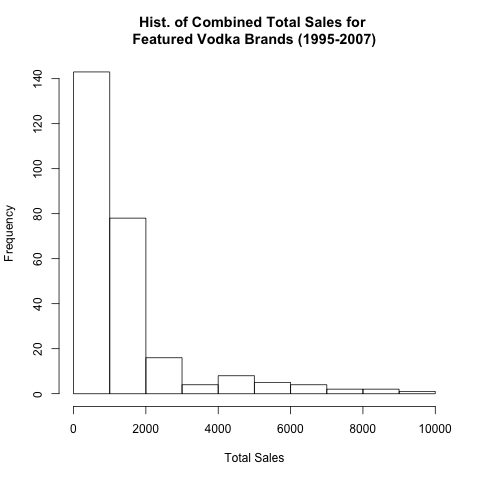
**R Code:**

hist(Svedka\_data$TotalSales,

main = 'Hist. of Combined Total Sales for \nFeatured Vodka Brands (1995-2007)',

xlab = 'Total Sales')

**Histogram:**



**Interpretation:**

The shape of the distribution is extremely downward trending with the majority of Total Sales at $2,000 or less. This means that most brands had Total Sales of $2,000 or less.

1. (10) Run a regression of the natural logarithm of total sales on the the following variables: price, print marketing expenditure, outdoor marketing expenditure, broadcast marketing expenditure, and previous year’s sales. Keeping in mind your answer to #1, explain why it makes sense to use ln(TotalSales) for the dependent variable. Comment on the relative influence of the five variables on sales.

**R Code:**

RegQ2 <- lm(LnSales ~ PricePerUnit + Print + Outdoor + Broad + LagTotalSales, data = Svedka\_data)

summary(RegQ2)

**Regression Summary:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | term | estimate | std.error | statistic | p.value |
| 1 | (Intercept) | 6.470163028 | 0.086154313 | 75.09969994 | 7.19E-177 |
| 2 | PricePerUnit | -0.007026292 | 0.000783984 | -8.96229116 | 6.69E-17 |
| 3 | Print | 5.15E-05 | 7.57E-06 | 6.800452075 | 7.27E-11 |
| 4 | Outdoor | -0.000366311 | 9.81E-05 | -3.732239412 | 0.00023366 |
| 5 | Broad | -4.53E-05 | 4.80E-05 | -0.943017476 | 0.346557595 |
| 6 | LagTotalSales | 0.000533418 | 3.88E-05 | 13.73459674 | 1.46E-32 |

**Interpretation:**

1. (15) Sometimes we can improve model fit by taking logs on independent variables. Run a second regression of the natural logarithm of change in sales on the the natural logarithm of previous period’s prices, and the natural log of marketing expenditures on print, outdoor, and broadcasting. Comment on the comparison of your two models at this point in the analysis.

**R Code:**

RegQ3 <- lm((LnSales-LnLSales) ~ LnPrice + LnPrint + LnOut + LnBroad, data = Svedka\_data)

summary(RegQ3)

**Regression Summary:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | term | estimate | std.error | statistic | p.value |
| 1 | (Intercept) | -0.11736535 | 0.098395868 | -1.192787385 | 0.234048456 |
| 2 | LnPrice | 0.036774462 | 0.025379124 | 1.449004426 | 0.148550674 |
| 3 | LnPrint | 0.014776085 | 0.004600179 | 3.212067579 | 0.001485189 |
| 4 | LnOut | -0.012622055 | 0.005785704 | -2.181593474 | 0.030042341 |
| 5 | LnBroad | -0.005763719 | 0.005158459 | -1.117333398 | 0.2648912 |

**Interpretation:**

1. (15 pts) To understand the influence of vodka quality, expand your regression model from question 2 by adding the tier 1 and tier 2 dummy variables (that indicate whether a vodka brand belongs to the first or second quality tiers) to the set of independent variables named in question 2. How does quality influence sales?

**R Code:**

RegQ4 <- lm(LnSales ~ PricePerUnit + Print + Outdoor + Broad + LagTotalSales + Tier1 + Tier2, data = Svedka\_data)

summary(RegQ4)

**Regression Summary:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | term | estimate | std.error | statistic | p.value |
| 1 | (Intercept) | 6.55222952 | 0.088065848 | 74.40148062 | 6.23E-175 |
| 2 | PricePerUnit | -0.007382457 | 0.001386619 | -5.324070373 | 2.23E-07 |
| 3 | Print | 4.00E-05 | 7.38E-06 | 5.418475066 | 1.39E-07 |
| 4 | Outdoor | -0.000317689 | 8.88E-05 | -3.578430889 | 0.000413617 |
| 5 | Broad | -4.36E-05 | 4.30E-05 | -1.014212187 | 0.311442855 |
| 6 | LagTotalSales | 0.000558629 | 3.51E-05 | 15.91373374 | 4.54E-40 |
| 7 | Tier1 | 0.259127332 | 0.210496189 | 1.231030991 | 0.219445478 |
| 8 | Tier2 | -0.464516646 | 0.11974273 | -3.879288924 | 0.000133405 |

**Interpretation:**

1. (15) To understand the influence of competition and brand power, expand your model again and run a regression by adding the sum of sales of all the competing brands in the previous year (“lagtotalminussales”) to the independent variables in question 3. What additional insight does this model provide?

**R Code:**

RegQ5 <- lm((LnSales-LnLSales) ~ LnPrice + LnPrint + LnOut + LnBroad + LagTotalMinusSales, data = Svedka\_data)

summary(RegQ5)

**Regression Summary:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | term | estimate | std.error | statistic | p.value |
| 1 | (Intercept) | -1.970292823 | 0.545316051 | -3.613120902 | 0.000363852 |
| 2 | LnPrice | -0.005790494 | 0.027748336 | -0.208678968 | 0.834864155 |
| 3 | LnPrint | 0.018269009 | 0.00461803 | 3.956018244 | 9.86E-05 |
| 4 | LnOut | -0.006700901 | 0.005920904 | -1.131736044 | 0.25880031 |
| 5 | LnBroad | 0.003567274 | 0.005730176 | 0.622541726 | 0.534137683 |
| 6 | LagTotalMinusSales | 3.20E-05 | 9.26E-06 | 3.452241505 | 0.000649897 |

**Interpretation:**

1. (10) To measure the sales growth of new brands compared to the existent ones, include the variable “firstintro” to the independent variable set in question 4. “Firstintro” is equal to 1 in the first three years after a brand is introduced, and equals 0 elsewhere. How does it help to include this variable in the model?

**R Code:**

RegQ6 <- lm(LnSales ~ PricePerUnit + Print + Outdoor + Broad + LagTotalSales + Tier1 + Tier2 + Firstintro, data = Svedka\_data)

summary(RegQ6)

**Regression Summary:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | term | estimate | std.error | statistic | p.value |
| 1 | (Intercept) | 6.583002839 | 0.086425622 | 76.16957458 | 6.11E-177 |
| 2 | PricePerUnit | -0.00769211 | 0.001356955 | -5.668653334 | 3.90E-08 |
| 3 | Print | 3.53E-05 | 7.33E-06 | 4.81218626 | 2.56E-06 |
| 4 | Outdoor | -0.000228172 | 9.01E-05 | -2.532070949 | 0.01194256 |
| 5 | Broad | -4.39E-05 | 4.20E-05 | -1.045524953 | 0.296774369 |
| 6 | LagTotalSales | 0.000541916 | 3.46E-05 | 15.66686511 | 3.61E-39 |
| 7 | Tier1 | 0.365394561 | 0.207642175 | 1.759731905 | 0.079657055 |
| 8 | Tier2 | -0.42642806 | 0.11741661 | -3.631752439 | 0.000340486 |
| 9 | Firstintro | -0.94107252 | 0.257890771 | -3.649112825 | 0.000319358 |

**Interpretation:**

1. (10) Examine the coefficients of the price and advertising variables in your last four regressions. Why do the coefficients of price and advertising change in the above regressions?

**Interpretation:**

1. (10) Create a time-series plot with two lines on it: total industry sales units for Tier 1 brands and total industry sales units for Tier 2 brands. NOTE: This will require some aggregation and pre-processing of the data, and is more of a challenge than it might appear.

**R Code:**

Tier1\_Agg <- aggregate(Svedka\_data$TotalSales ~ Svedka\_data$Year + Svedka\_data$Tier1, FUN = sum)

View(Tier1\_Agg)

Tier1\_Agg <- rename(Tier1\_Agg,c("Svedka\_data$Tier1" = "Tier1","Svedka\_data$Year" ="Year",

"Svedka\_data$TotalSales" = "Total Sales"))

Tier2\_Agg <- aggregate(Svedka\_data$TotalSales ~ Svedka\_data$Year + Svedka\_data$Tier2, FUN = sum)

View(Tier2\_Agg)

Tier2\_Agg <- rename(Tier2\_Agg,c("Svedka\_data$Tier2" = "Tier2","Svedka\_data$Year" ="Year",

"Svedka\_data$TotalSales" = "Total Sales"))

Tier1 <- Tier1\_Agg[which(Tier1\_Agg$Tier1==1 & Tier1\_Agg$Year),]

View(Tier1)

Tier1$Tier1 <- NULL

View(Tier1)

Tier2 <- Tier2\_Agg[which(Tier2\_Agg$Tier2==1 & Tier2\_Agg$Year),]

View(Tier2)

Tier2$Tier2 <- NULL

View(Tier2)

library(ggplot2)

ggplot(Tier1,aes(Year,`Total Sales`)) +

geom\_line(aes(color="Tier1")) +

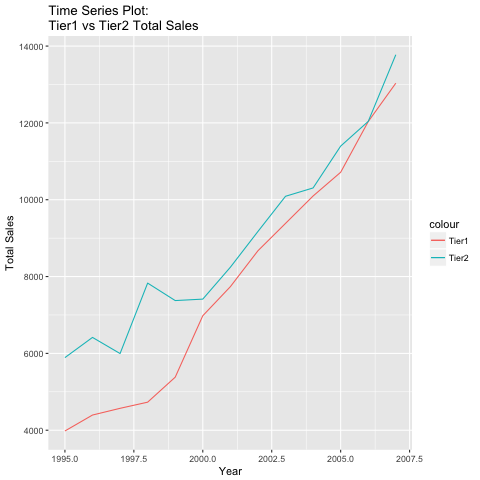
geom\_line(data=Tier2,aes(color="Tier2")) +

ylab("Total Sales") +

xlab("Year") +

ggtitle("Time Series Plot: \nTier1 vs Tier2 Total Sales")

**Time Series Plot:**



**Interpretation:**

1. (10) Conclude with a short summary of your findings. How do the 4 elements of the marketing mix influence unit sales in this industry? What insights should we communicate with M. Cuvelier?

**Interpretation:**

When finished, please save your paper as a pdf. This facilitates the grading and feedback process in LATTE.