Understanding Shopping Behavior

A Data Science Retail Case Study (DSRCS)

Contents

[Introduction 2](#_Toc494584965)

[Business Interest 2](#_Toc494584966)

[Business Question 2](#_Toc494584967)

[Objectives 2](#_Toc494584968)

[Business Question to Data Question 2](#_Toc494584969)

[Case Study Parameters 2](#_Toc494584970)

[Retail Chain Z 2](#_Toc494584971)

[Outlets X and Y 2](#_Toc494584972)

[Item of Interest 2](#_Toc494584973)

[Data 3](#_Toc494584974)

[Exploratory Data Analysis 4](#_Toc494584975)

[Data Visualization 4](#_Toc494584976)

[Assumptions 4](#_Toc494584977)

[Normalizing figures 4](#_Toc494584978)

[Initial Analysis 4](#_Toc494584979)

[Potential Factors 4](#_Toc494584980)

[Experimental Design 4](#_Toc494584981)

[Hypothesis 1 5](#_Toc494584982)

[Hypothesis 2 5](#_Toc494584983)

[Hypothesis 3 5](#_Toc494584984)

[Data Analysis with Python 5](#_Toc494584985)

# Introduction

Big retail chains in U.S. such as Walmart, Target and Sears each operate hundreds of outlets across the country.

# Business Interest

Inventory Management is an integral part of their operation as they must manage huge inventories of the same item across different outlets. The ability to maintain lean inventories is key to keeping the carrying cost of inventory down.

Most retail chains maintain uniform pricing (excluding local taxes) despite uneven income levels across different states. Choosing the right pricing strategy to maximize total sales is paramount to profit maximization.

# Business Question

One question that is intertwined with the business interests mentioned above relates to the difference in total sales of the same item across outlets in different locations.

# Objectives

The objective of this case study is to formulate the Business Question as a Data Question and to explore a plausible framework to answer the data question.

# Business Question to Data Question

To translate the business question into a data question or problem, the business question must be reframed so that it is specific and answerable.

Concretely, the Business Question can be refined as follows:

***“What causes the difference in the annual total sales of item A between two outlets X and Y of retail chain Z?”***

# Case Study Parameters

## Retail Chain Z

In this case study, Sears is chosen as the Retail Chain Z.

## Outlets X and Y

In this case study, X and Y are located in different states. X is in the state of Indiana and Y is in Florida.

## Item of Interest

In this case study, an Action Figure is chosen to be the item of interest (item A). An Action Figure is a poseable character figurine that is usually based upon characters from a film, comic book or video game[[1]](#footnote-1).

The reason for choosing an Action Figure is due to the following reasons:

* It is very brand specific, and hence it is less likely to suffer from the substitution effect from other similar toys.
* Customers are more likely to want to look at it before buying it, and hence the outlet itself can affect its sales.

Concretely, the specific item of interest is a 16” Toy Story Woody Talking Action Figure from Disney[[2]](#footnote-2). Its retail price is $30.97, which is below the minimum order amount to qualify for free shipping.

## Data

To help us answer this data question, we will need the following types of raw data:

Internal Data (from the retail chain)

* Consolidated Purchase Record (of this item over the financial year)
* Individual Receipts (that includes this item)
* Transaction History (of customers who paid by credit card or members)
* Inventory Record (of this item over the financial year)

External Data (from public sources or business intelligence)

* Age Demographics of neighborhood
* Income Demographics of neighborhood
* Location of competitor stores selling the same figurine

Each type of raw data gives us different information that can help us better understand the profile of our buyers.

**Item Information**

* Item #
* Item Name
* Price
* Brand

**Primary Variables**

* Purchase Date
* Location

**Secondary Variables**

* Credit Card Name
* Similar Purchases from members and/or card members
* Items from the same receipt

## Exploratory Data Analysis

The purpose of the exploratory data analysis is to identify patterns, trends and/or relationships between primary variables. This will greatly help us in the experimental design stage.

## Data Visualization

Often, making a plot helps us to make sense of the data:

* Histogram (Sales Figures at each location)
* Scatterplot (Sales Figures over time)
* Line Graph (Cumulative Purchases)

## Assumptions

In any experiment, we need to make certain underlying assumptions about the environment and the initial data:

* Both stores sell the same Action Figure A at the same price
* Both stores have identical store layouts.

## Normalizing figures

We can attempt to take into account differing patronage by normalizing the sales figures:

* Sales figures per 1000 transactions per location

# Initial Analysis

Depending on the exploratory data analysis, we might consider the following variables as potential contributing causes behind the difference in sales figures.

## Potential Factors

* Demographics (percentage of children living in the vicinity of the outlet)
* Income Level (of households living in the vicinity of the outlet)
* Location (of outlets vis-à-vis potential customers)
* Availability of Substitute (stores that sell the same figurine)

# Experimental Design

The next step is to design an experiment to test out certain hypothesis about the population, based on our initial analysis. The outcome of each hypothesis will lead to different potential business decisions.

## Hypothesis 1

* Explanation: Demand for item is elastic
* Variable: Price
* Customer Factor: Income level
* Business Outcome: Pricing Strategy

## Hypothesis 2

* Explanation: Substitution Effect
* Variable: Number of competing stores selling the same item
* Customer Factor: Convenience
* Business Outcome: Inventory and Sales Projection

## Hypothesis 3

* Explanation: Visual Attraction
* Variable: Prominence of item display
* Customer Factor: Attention Span
* Outcome: Optimal Store Configuration/Layout

# Data Analysis with Python

The exploratory data analysis is carried out with python on simulated data. The source code and data files can be downloaded from GitHub:

<https://github.com/kanechew/Data-Science-Retail>

1. https://en.wikipedia.org/wiki/Action\_figure [↑](#footnote-ref-1)
2. http://www.sears.com/disney-16inch-toy-story-woody-talking-action-figure/p-004W002022527003P?sellerId=KMART&prdNo=2&blockNo=2&blockType=G2 [↑](#footnote-ref-2)