Exploratory Data Analysis

- StudentID: 21800418

- Name: Seonggyun Ahn

- 1st Major: Biology

- 2nd Major: AI

"Exploring which item types can maximize total profit under specific conditions"

1. Data overview

S_data contains 5,000,000 entries and 15 variables.

These are the formats and assumed meaning of the variables

`Region` (String) - The broad area where the sale occurred.
`Country` (String) -The specific nation of the sale.
`Item.Type`: (String) - The category of the sold product.
`Sales.Channel` (String) - How the sale was made (online or offline).
`Order.Priority` (String) - The urgency level of the order.
`Order.Date`: (String) - When the order was placed.
`Order.ID` (Integer) - A unique number identifying the order.
`Ship.Date` (String) - When the order was shipped.
`Units.Sold` (Integer) - Quantity of product sold.
`Unit.Price` (Float) - Sale price per unit of product.
`Unit.Cost` (Float) - Cost per unit to the seller.
`Total.Revenue` (Float) - Total income from a sale.
`Total.Cost` (Float) - Total expense for the sold units.
`Total.Profit` (Float) - Total income minus total cost.

Region	Country	Item Type	Sales Channel	Order Priority	Order Date
Central America and the Caribbean	Saint Lucia	Clothes	Offline	H	11/8/2011

Order ID	Ship Date	Units Sold	Unit Price	Unit Cost	Total Revenue	Total Cost	Total Profit
839619243	12/23/2011	847	109.28	35.84	92560.16	30356.48	62203.68

Table 1. First line of the S_data

2. Univariate analysis

In this project, we are most interested in the maximization of total profits.
This can be expressed as `Total Profits` = `Unit Profit` X `Units Sold`.
Therefore, we want to perform a univariate analysis on `Unit Profit` and `Units Sold`, which are the key variables for the objective function.

2.1 Unit Profit

We created the variable {Unit Profit} through the calculation of an existing variable.
`Unit Profit` = `Unit Price` - `Unit Cost`

The distribution of Unit Profit was clustered in a specific area.
The most heavily clustered section is on 51.44.
(Actually, it is clustered in twice compared to other section.)
The number of clustered sections is 11.

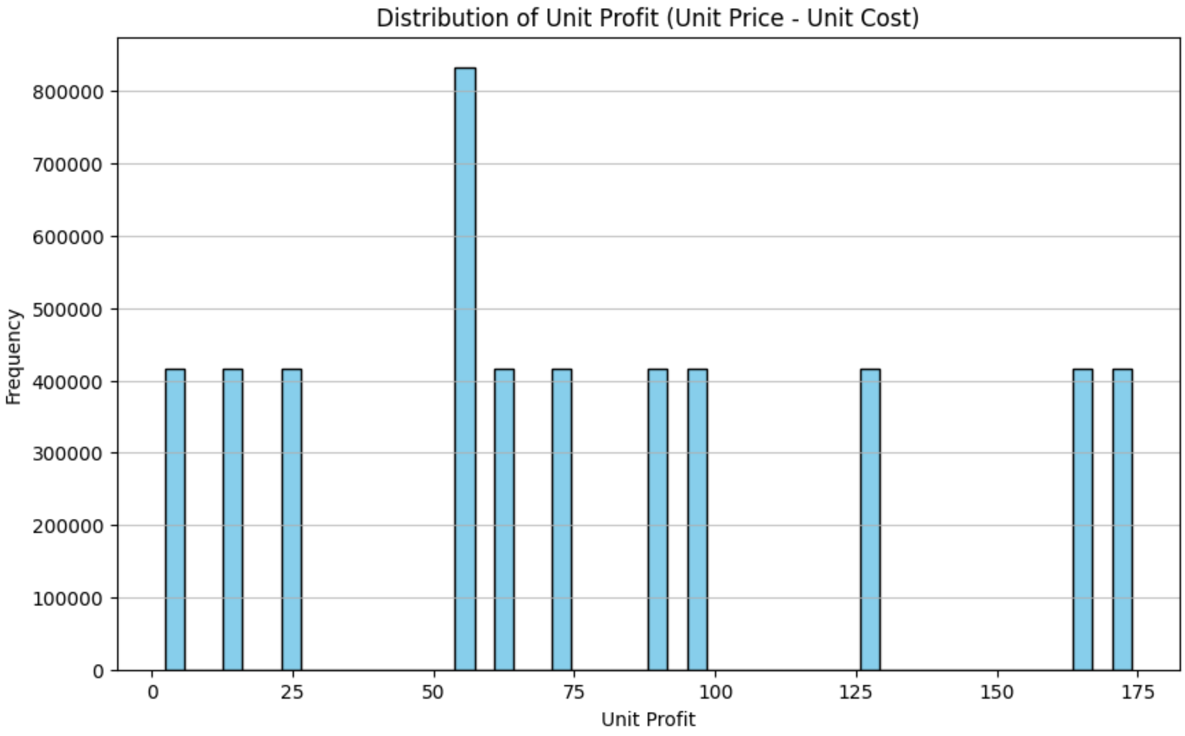

Figure 1. Histogram of Unit Profit

Mean Unit Profit: 78.53462
Variance of Unit Profit: 2770.19704
Min-Max and Quartiles of Unit Profit:
Min 2.41
Q1 55.14

Q2 73.44
Q3 126.25
Max 173.87

2.2 Sold Units

{Units Sold} is one of the main variables along with `Unit Profit` that determines `Total Profit`.

The Units Sold variable is almost uniformly distributed in all areas between max and min.

This ensures that conclusions drawn from that data will be generalizable with a sufficient and equitable sample size.

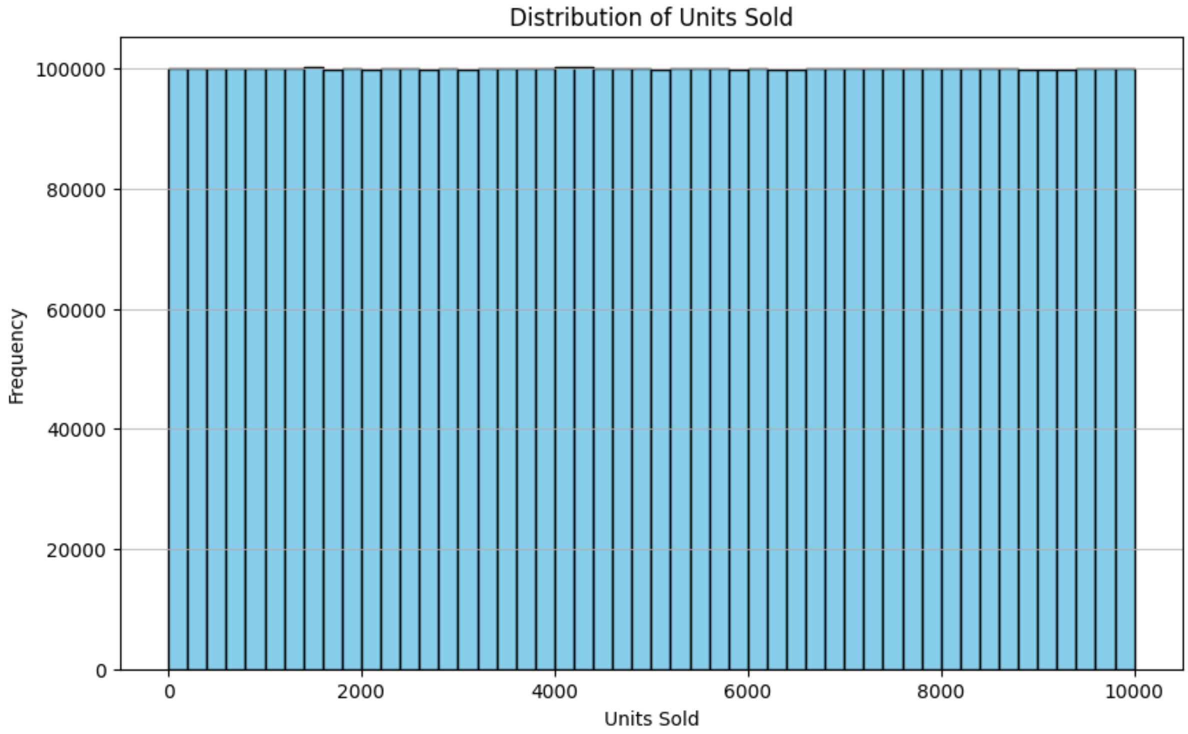

Figure 2. Histogram of Units Sold

Mean Units Sold: 4999.99106
Variance of Units Sold: 8333539.22402
Min-Max and Quartiles of Units Sold:
Min 1
0.25 2500.0
0.50 4999.0
0.75 7500.0
Max 10000

2.3 Item Type

Since the goal of the project is to select "item types" to maximize total profit under given conditions, we need to see the distribution over `Item Type`

There are 12 categories in `Item Type`:
“Office Supplies" "Beverages" "Cereal" "Snacks" "Personal Care" "Cosmetics" "Clothes" "Meat" "Fruits" "Household" "Vegetables" "Baby Food"

The number in all categories is the same, around 416,600

Also, this ensures that conclusions drawn from that data will be generalizable with a sufficient and equitable sample size.

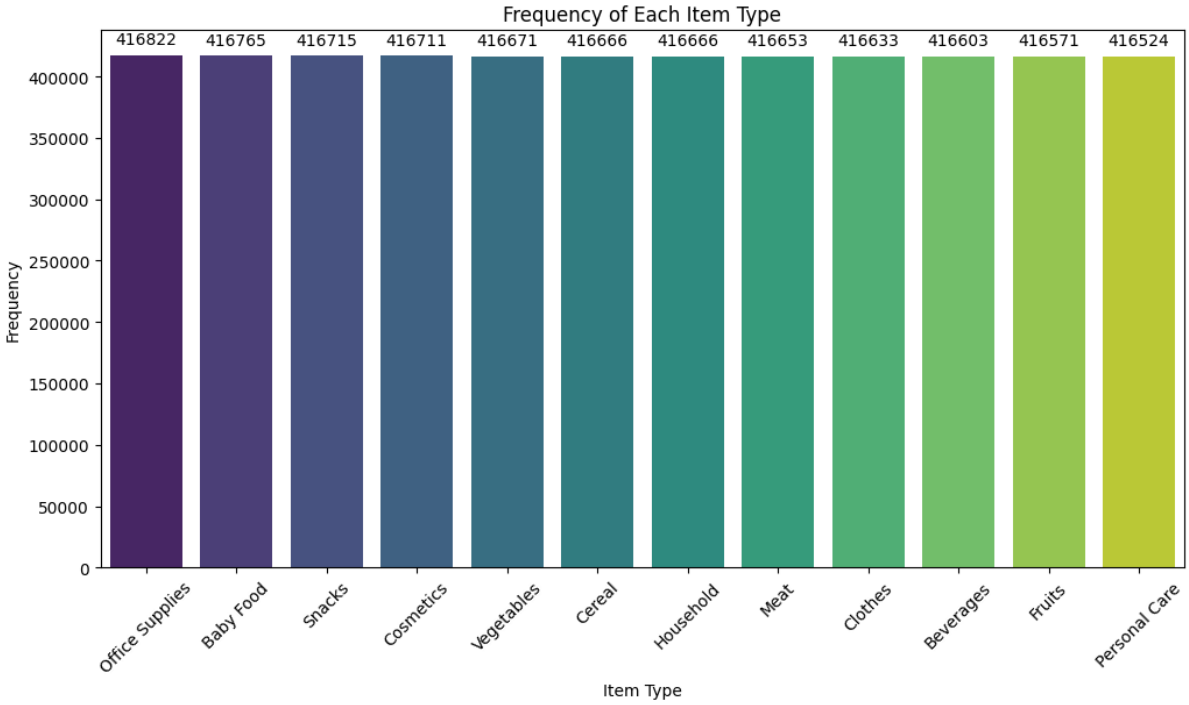

Figure 3. Bar graph of Item Type

3. Multivariate analysis

3.1 The distribution of unit profit by item type

As you can see in the graph, there is no change in the profit ratio for each group of item types.
This means that once `Item Type` is determined, the variable `Unit Profit` can be treated as a constant, and from then on, `Units Sold` is more important in determining `Total Profits`.

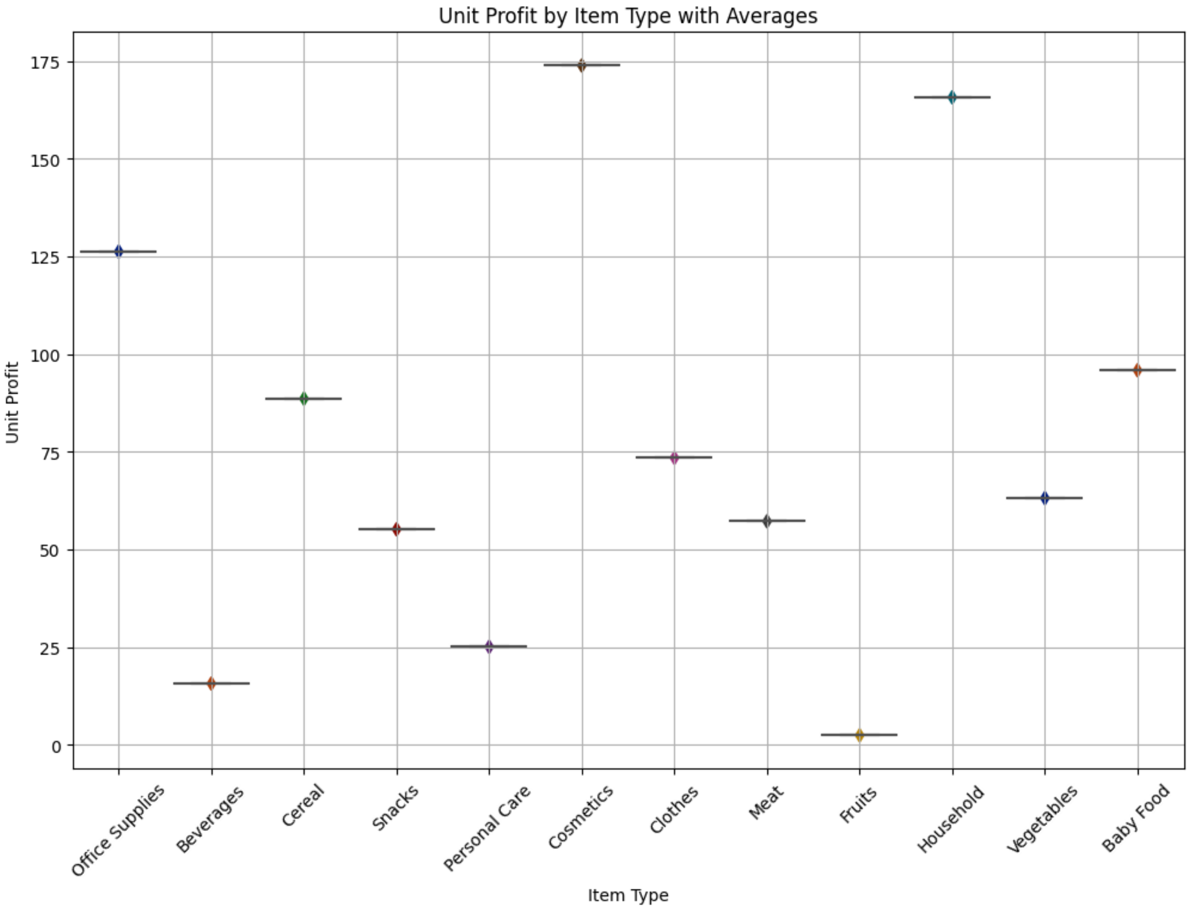

Figure 4. Box plot of Unit Profit by Item Type

3.2 ANOVA to test independence between `Units Sold` and `Item Type`

As you can see in the graph below, the distribution of Sold Units across categories in the Item Type is almost all uniformly distributed.
This suggests that the two variables may be independent.

ANOVA was performed to statistically test the independence of the two variables.
The ols model from the python statsmodels package was used.

In the result of ANOVA, The F-statistic is quite low, and the P-value is very high (close to 1), much higher than any typical alpha level (0.05 or 0.01), which would indicate statistical significance.
Therefore, we can conclude that `Item Type` does not appear to have a significant effect on `Units Sold`.

Based on the visual distribution of the two variables and the statistical significance of the ANOVA results, it is reasonable to assume that the two variables have an independent relationship. This means that there are no quantity constraints by item types.

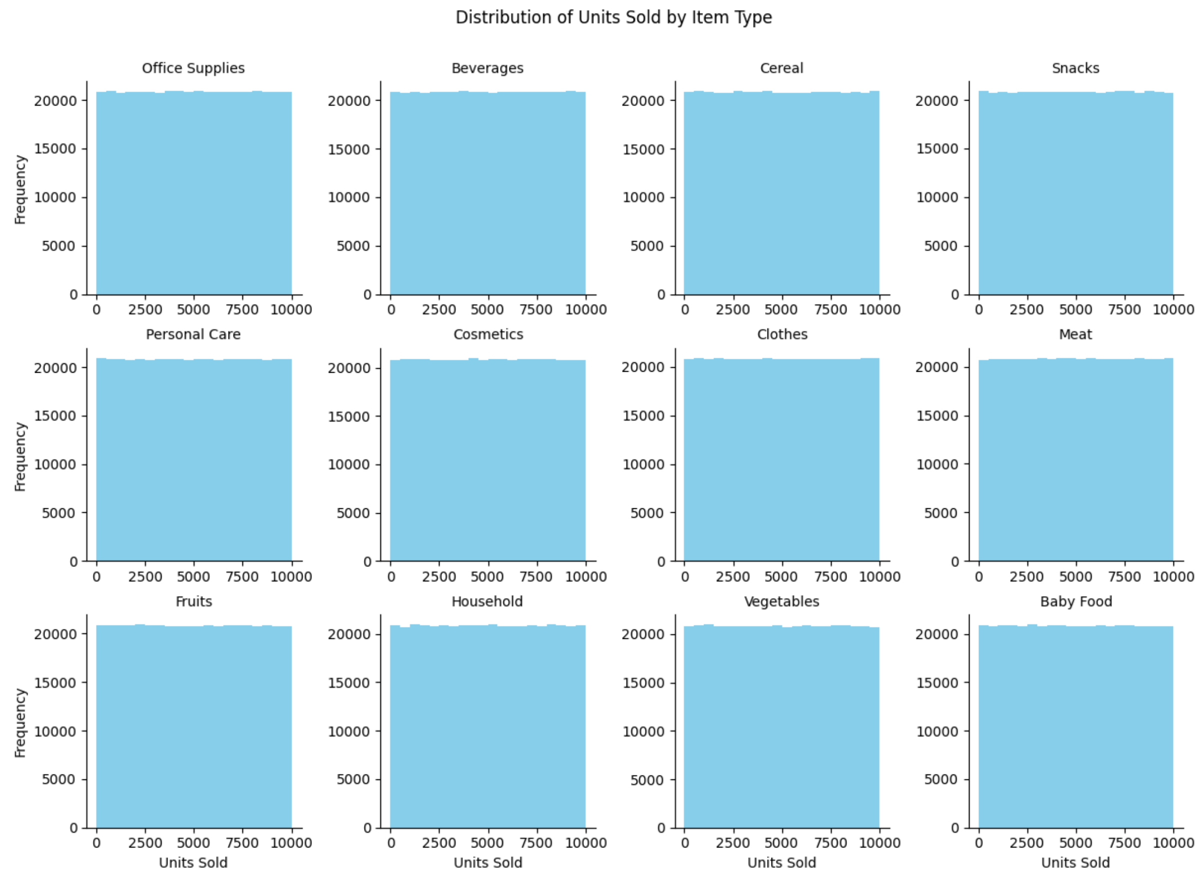

Figure 5. Histogram of Sold Units by Item Type

Source	sum_sq	df	F	PR(>F)
"Item Type"	7.221663e+06	11.0	0.07878	0.999976
Residual	4.166768e+13	4999988.0	NaN	NaN

Table 2. Result of ANOVA between Sold Units and Item Type

3.3 ANOVA to test independence between `Units Sold` and other variables

In the previous result, we found that it is reasonable enough to assume that the two variables `Units Sold` and `Unit Profit` are independent.

However, this still does not guarantee that the two variables `Units Sold` and `Unit Profit` are independent when other conditions intervene.

Given that events `A` and `B` are independent, it does **not** necessarily follow that events `A` and `B` are independent when conditioned on `C`.

Notation:

- `A` and `B` are independent: `A ⫫ B`
- `A` and `B` are **not** necessarily independent given `C`: `A ⫫ B | C` is not assured.

In this dataset, there are additional variables—`Region`, `Country`, `Sales Channel`, and `Order Priority`—that could be correlated with `Units Sold`. Consequently, we conducted an ANOVA, similar to previous analyses, to ascertain the correlation between `Units Sold` and these variables.

Fortunately, as with the previous results, the ANOVA results did not support a correlation between the variable `Units Sold` and the other variables

Source	sum_sq	df	F	PR(>F)
C(Q("Region"))	4.645852e+06	6.0	0.092915	0.997067
Residual	4.166768e+13	4999993.0	NaN	NaN
------------------	--------------	----------	---------	--------
C(Q("Country"))	2.118857e+08	184.0	0.138179	1.0
Residual	4.166748e+13	4999815.0	NaN	NaN
----------------------	--------------	----------	--------	----------
C(Q("Sales Channel"))	5.431802e+05	1.0	0.06518	0.798489
Residual	4.166769e+13	4999998.0	NaN	NaN
-----------------------	--------------	----------	---------	---------
C(Q("Order Priority"))	1.637459e+06	3.0	0.065497	0.978154
Residual	4.166769e+13	4999996.0	NaN	NaN

Table 3. Result of ANOVA between Sold Units and other variables

3.4 ANOVA to test independence between `Units Sold` and other variables

From steps 3.1 through 3.3, we can conclude that the only considerations in the process of selecting an `Item Type` to maximize gross profit are `Unit Profit` and `Unit Cost`, which are determined by the `Item Type` itself.

The scatter plot below can help you decide which Item Type to choose.

If you want to maximize your total profit, you should choose an item type with a **high absolute value** of Unit Profit, and if you have limited costs and prioritize profit over cost, you should choose an item type with a **high slope** of the line connecting the origin and each point.

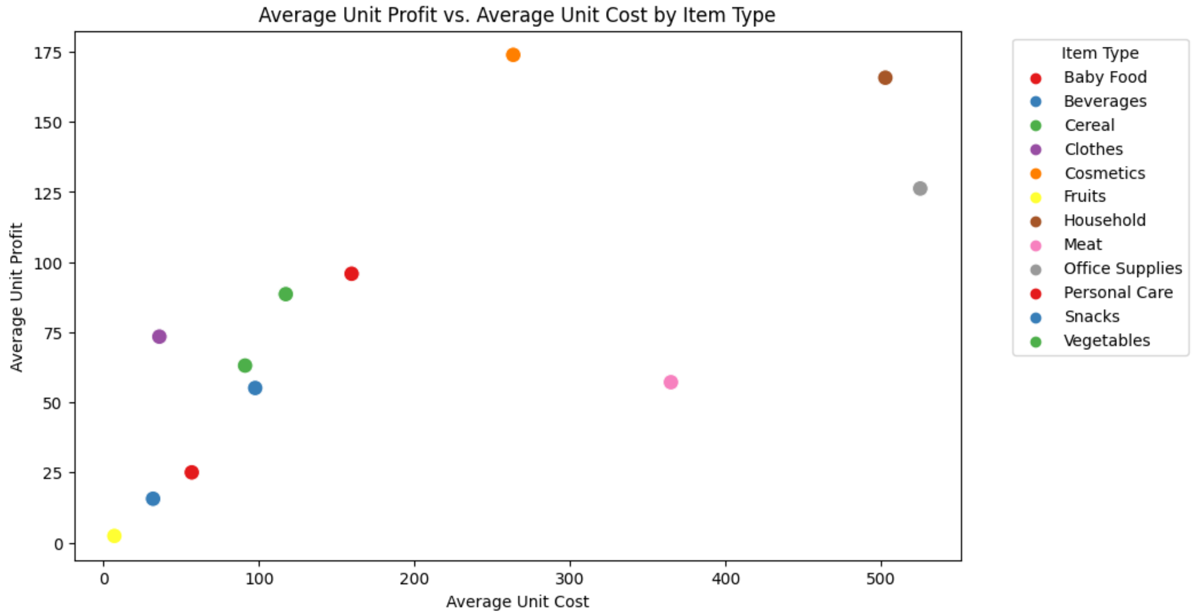

Figure 6. Scatter Plot by avg Unit Profit and avg Unit Cost

4. Suggestion

The goal of the project is to maximize `Total Profit`, which can be expressed as follows: `Total Profits` = `Unit Profit` X `Units Sold`

We have seen in Sections 3.2 and 3.3 that `Units Sold` can be assumed to be independent of `Units Profit` under any condition.
Therefore, increasing `Units Sold` as much as possible will increase `Total Profits`.

As we have seen in 3.1, the `Unit Profit` (and also the `Unit Cost`) is determined by the `Item Type`.
Therefore, if the selected `Unit Profit` is large, selecting `Item Type` will increase the `Total Profits`.

Referring to the results of 3.4
If you want to maximize your profit regardless of the cost of consumption, you should choose **Cosmetics**. If you want to maximize the profits over the costs, you should choose **Clothes**.