# **Retailer Case**

# Problem Statement

**The goal of this project** is to optimize shelf spacing. We are basing our models on three assumptions/hypotheses:

* The more shelf space given to a product, the more it will sell.
* The more of a product type is sold, the more a complementary product will be sold.
* If the final products are next to each other, the complementary effects will be even greater.

**Our first objective** is to confirm the validity of these assumptions.

**Out second objective** is to figure out what model(s) are appropriate in deciding how much shelf should be allotted for each product type to maximize profit.

There are a few restrictions that are imposed upon us:

* Minimum and maximum amount of shelf space for each product
* Total shelf space is a constant amount

And we are told to ignore:

* Promotions for certain products or product types
* The fact that some companies pay for more space

The greatest difficulty in optimizing for sales is how to accurately predict the sales of a product/product type based on the shelf space it is given.

Now that we have the problem defined, let’s dive in!

# Objective 1 – Validate the hypotheses

**First Hypothesis -** To prove the first hypothesis, it requires us to prove that a marginal increase in shelf space will result in a marginal increase in sales. Intuitively, this would not be a linear relationship but might resemble a logarithmic function. That is a small increase in shelf space from 0 to 10 square feet might result in huge sales and at some point, increasing shelf space will not increase the sales by much.

To confirm this, an A/B test could be performed. Seeing whether different shelf space allotted for the same item results in what corresponding change. There are two issues I foresee with this. First, we would have to control for the variable of store location. The resulting question is, can we perform a test so that the variance in change in sales can be 90 to 100 percent predicted by the shelf space? Not only is this relatively difficult to accomplish but it is also difficult to confirm whether we controlled for other variables successfully.

Another approach could be we increase shelf space gradually and track the changes in sales. The disadvantage of this is that we would have to take space away from other products as we increase shelf space for the product we are testing. It also may take some time to see the results if we perform this approach. Seasonality and trend may also be of concern. And the solutions to these concerns are simply not feasible.

Ultimately, a modified version of the first approach could be used to determine the impact on sales from shelf space. We could account for the other effects that was of concern in the first approach. The issue remained however that proving causation – increasing shelf space results in an increase in the sales – remains elusive.

There remains another approach, where a new data source that tracks customers throughout a store would allow us to determine who walked past a product type, who stopped at a product type and who purchased a product type. Using hypothesis testing, one theoretically could confirm the first hypothesis. Unfortunately, this method for various reason does not work in practice.

**Second hypothesis** – Using data that the grocery chain has (containing exact sales data of each product) provides us the information to test which products sold well together and which did not. A correlation test could be done on each product pair and if the correlation is strong enough, then a determination could be made that a relationship exists between the sales of two products. One might be inclined to think that proving a correlation does not necessarily prove causation. In this context, meaning that just because a certain product sells higher, does not mean that it will drive sales for the other product, vice-versa. However, as the grocery store does not need to prove causation but only needs to prove that there is indeed a relationship between certain types of products, a hypothesis test will suffice. The p-value can of course be altered. Increasing it may increase the number of incorrect relationships, where a low p-value might ignore potential relationships. If one wanted to further explore what significance level to set, then one could pick a p value that would optimize for the cost of ignoring a potential relationship between products and validating incorrect relationships. Leaving the p-value as is at 0.05 may also be acceptable as well.

Using a regression model would also enable us to measure the magnitude of the complementary effects. It is important for our regression model to model non-linear relationships as we should not assume that increasing sales of one product will correspond with a linear increase in the other.

**Third Hypothesis –** This hypothesis proves to be the most challenging to confirm. There is relatively limited information on specific product locations from the collected data which makes this much more difficult. But using the limited that the grocery chain does have can prove the third hypothesis for many of the relationships detected in our validation of the first hypothesis.

After exhausting all possible methods, we cannot confirm hypothesis 1. Objective 1 was only partially met. Nevertheless, for the purposes of this assignment we will assume hypothesis 1 to be true and move on to Objective 2.

# Objective 2 – Optimize Shelf Space and Location of Products

**Model Relationship between Shelf Space and Profit for Each Product Type**

The first step would be to try to model the relationship between shelf space and profit for each product type. This is very much akin to the process we used to prove Hypothesis 1. While it may not be practical to try and model the shelf space and sales for every single product type, we can try and model the relationships for a few, disparate product types. We can then extrapolate it for all the other product types as well.

As we found out when attempting to confirm hypothesis 1, there is no perfect method. However, I am personally in favour of an exponential smoothing model that would track sales based on changes in shelf space. While this may indeed require long-term sales data and would take time to fully implement, we would be able to observe changes after small increases in shelf space versus comparing between two values of shelf space in A/B Testing. Of course, we would need prior data so that we could detrend seasonality and trend. However, since we will only be performing this model on select product types, it should reduce the difficulty of finding prior data. We will not need prior data for all product types.

After extrapolating for other product types, we can then alter our model to compare shelf space with profit and not sales. Ultimately it is profit we want to maximize.

Conclusion:

Given past sales data of at least 2 years, we can use triple exponential smoothing to determine corresponding changes in sales from changes in shelf space.

**Model Relationship between Different Product Types**

When confirming hypothesis 2, we used previous sales data to determine which products sold well together and which did not. If we used a simple correlation test to determine whether there existed relationships between certain pairs of products, we could rank which pairs of products were most complementary to least complementary.

We could use a Pearson correlation test, or we could also use the R^2 value from a linear regression on the sales data between two variables. If we use the R^2 value, it will serve us later as well when we end up clustering product types. Either should work for our purposes.

Conclusion:

Given previous sales data, we can use a Pearson correlation test or a linear regression test to determine which pairs of products complement each other and how much they complement each other

**Model Clusters of Product types**

There are several ways we can figure out which products should be clustered together. Given the third hypothesis that complementary products sell well when close to each other, we can use clustering algorithms to determine which products should be next to each other based on the data we acquired in the previous model which tells us the strength of the relationships between pairs of products.

I think the Louvain algorithm will work best in this context, as the arc weights can be defined by either transformed values of the Pearson correlation test from the previous model or R^2 values from the previous model. Doing this will allow us to determine where products should be placed.

Conclusion:

Given the results of the linear regression or the Pearson Correlation test we can use the Louvain algorithm to determine which products should be clustered together.

**Optimize for Shelf Space**

Once we have the locations of where items should be placed, an optimization algorithm would give us the best shelf space square footage to allot for each product. However, as I may have mentioned previously, the relationship between shelf space and profit may not be linear. For a practical optimization algorithm therefore, we would have to assume model linear relationships between shelf space and profit. While the optimization algorithm is designed to give us the optimal shelf space values, it may only get us close but not perfect.

Our constraints would include:

* Minimum and maximum amount of shelf space for each product
* Total shelf space is a constant amount

Using this, we should be able to derive optimal values for shelf space for each product type.

Conclusion:

Given the results from modeling the relationship between shelf space and product type, we can use optimization with linear constraints to determine which how much shelf space should be allotted for each product type.

# Analysis Results

Using the results of the Louvain Algorithm Model and the Optimization Model, we will know how much shelf space to allot for each product type as well as which products should be placed close together so as to maximize the complementary effects of certain pairs of products.