INTRO
Hi everyone, I'm Nate and I'm a Data Scientist. Today I will be talking about the profits of a food distribution company in the San Francisco Bay Area.

You may notice I have the same last name as the one in the company I'll be talking about.
That's no coincidence.
My grandfather started this business in 1967 from his garage.
So, I've been listening to my family talk about this subject for about 30 years.

OUTLINE
Here is a quick overview of what I'll be going over.
I'll present the problem that the business is having.
We will explore some of the data.
I'll share a predictive model I created and some analysis
At the end, I'll provide my recommendations.

First, let's take a look at the structure of the business.

HIERARCHY
Here is a simplified look at the product lines at Dale Cox Distributing.
I'll split the company into two groups.
Pepperidge Farm and not Pepperidge Farm.
Pepperidge Farm is a franchise within Dale Cox Distributing. There are set stores to distribute to and profit margins are set by Pepperidge Farm.
I'll refer to the **not** Pepperidge Farm side of the business as DCD.
DCD contains hundreds of different brands of products. There is a lot more freedom with these brands. Profit margins are more flexible and can be determined by Dale Cox Distributing.

We'll touch on a few things concerning Pepperidge Farm, but ultimately, the main focus will be on DCD and all of the product labels that make it up.

PROBLEM STATEMENT
The Pepperidge Farm franchise has been having supply issues. Less product to distribute has caused the profits of Dale Cox Distributing to decrease.

I plan to answer the question:
How can these lost profits be recovered?

DATA
Gathering the data to try and solve this problem was the first challenge I faced.
Dale Cox Distributing's systems and software are a bit older and the data is stored in a flat-file database. I worked with two different individuals to get the data I needed for this project.

I ended up with 5 years of monthly data on the profits of Pepperidge Farm and DCD, and 5 years of data on orders of DCD products.

In the product data there was information on over 250 stores in 7 different territories.
There is over 50 different product types and around 2000 unique products

DCD AND PF PROFITS
Here is a look at the monthly gross profits from Pepperidge Farm and DCD products. Both product lines saw increases in profits in 2020. The main contributor to that was COVID. With restaurants being closed or reduced to take-out only, people had to rely on grocery stores even more for their meals, snacks, and desserts.
In the middle of 2020, Pepperidge Farm profits began to decrease.

Let's take a closer look.

PF PROFITS
Here is the same data, but with only Pepperidge Farm. The dropoff of monthly profits is more visible now. A decrease of nearly $10,000 a month is huge for a company of this size. Expenses have been increasing across the board from areas such as fuel, utilities, wages, and truck expenses. Before this major drop in gross profits, the company would usually hover around breaking even or operating at a net loss. They aren't profiting large amounts of money, so losses like this hurt a lot.

PF ORDERS
Every week, an order is placed for more cases of Pepperidge Farm products to restock the distribution warehouse.
Not once during this 10 month span of data did the number of cases received match the number that were ordered.
For some weeks, they only received a third of what they needed.
This clearly shows that there is a demand, but the supply is not there.
This is what is causing profits from Pepperidge Farm to decrease.

PF ORDERS TABLE
Here are the numbers for those weekly orders over the 10 months of data. The number of cases and the value of those cases that were received is less than the missing amount. So, the company received less than half of what was needed.
The estimated profits that were lost from this was almost $350,000.

SWITCH TO DCD
So, the issues with profits have everything to do with Pepperidge Farm and not Dale Cox Distributing.
To make up for these losses, I want to focus on DCD.

DCD PROFITS
Here is a look at the monthly gross profits of the product lines that make up DCD. After the spike in profits that came with COVID, the numbers began to decrease.
I want to make sure this trend doesn't continue.

Let's explore what makes up DCD a little closer.

PRODUCT TYPES
These are the top 10 product types based on the number of orders received. Chips and Candy clearly dominate the types of foods being distributed.
Not all products from DCD are food. During the pandemic they picked up some masks to sell. They have some earth friendly house cleaners. For a number of years, they even had umbrellas and rain gear.

This gives you a better idea of what is being distributed.
There is a lot I could talk about with this product data, but
I want to stay on course and jump into my process of how I think DCD's profits can increase.

MARKET BASKET ANALYSIS
One of the first things I looked into was performing market basket analysis.
I wanted to see if there were any items that were frequently ordered together from stores.
Unfortunately, I did not find anything that was significant.
The threshold to identify a relationship between products in an order had to be set very low to have any results.

RFM
I decided to switch gears and try using RFM as an analysis tool.
RFM stands for Recency, Frequency, and Monetary.
I used this tool to create a profile for each store that Dale Cox Distributing sells products to based on the Recency, or the last time they ordered something.
The Frequency of the stores orders in the 5 years of data that I have.
Also, the monetary value, or how much money the stores would spend on orders.

CLUSTERS
With this RFM information, I was able to group the stores together using an unsupervised learning algorithm known as K-means clustering.
I'll spare you the details on how it works and instead show you the results.
I determined the optimal number of clusters for the stores was 3. Here are all three clusters of stores in a 3d scatterplot. The red cluster includes the best stores to distribute to. Followed by the green, and then the blue.

FOCUS
I decided I wanted to focus on the red and green clusters. The stores in these clusters had more recent orders, ordered more frequently, and spent more on orders.

To increase profits from orders from these two clusters of stores, I now had to look at products.
I broke down products into two groups.
Products that Dale Cox Distributing had control over and top selling products.

CONTROLLED PRODUCTS
Some of the items sold by Dale Cox Distributing are products where they own the label, or they are the only distributor.
I felt these products are the best ones to go after because there is no competition from other distributors and they have complete control over the profit margins.

TOP PRODUCTS
Along with the controlled products, I wanted to look at the top selling products.
One issue with these is that some stores might be already carrying these items because another distributor is selling them.
I will look into them, but ultimately someone will have to check individual stores to see if each of these items are already there.

MODEL
So, I created a profile for each store using RFM.
I also had monthly data for the stores that ordered each product.
I knew I wanted to target the stores that weren't ordering these products.

I decided I would treat this as a regression problem and try and predict how much product a store would order in a month.
I created and tested 5 different types of models. I ended up using a XGBoost Regressor for my production model because it performed the best.

The features of the model were recency, frequency, and monetary value of a store's orders.
The target, or thing I was trying to predict, was the quantity of product a store would order in a month.

I trained the model on the stores that ordered the product.
Once the model was trained, I then would take the stores that were not ordering this product with their RFM information and predict how much those stores would order.

Once I had my predictions, I could then sort the stores based on the predicted quantity ordered. This would list the stores in order from best to worst target when determining which stores to attempt to push more products to.

This process was repeated for each of the 7 product labels that Dale Cox Distributing has control over, as well as some top selling products.

RESULTS
I judged my models performance on the RMSE, or root mean squared error.
You can think of these numbers as being the quantity of a product.
The baseline that I was trying to beat would be predicting the average quantity in the dataset for every order.
The lower the number, the better.
So, each of my models beat the baseline, but nothing spectacular. The real value of my model was how it acted as a recommender.
I fed it the information about the stores, and it would give me those predictions on how much product a store would order in a month.
This allowed me to sort all those stores in those two clusters I showed you all previously.

PREDICTIONS
Here is an example of what my model would produce. These are the top 5 stores that should be targeted to sell La Morenita Chips.
It's pretty simple, the stores that have ordered more recently, more frequently, and spend more money on orders are at the top of the list.

POTENTIAL PROFITS

I looked at the potential profits my work could generate.
For each of the 15 products I made predictions on, I took the top 3 stores and calculated the profits they could have generated the previous year.
Selling these products in those stores would have been a $26,000 or 2.2% increase in profits for 2021.
I feel like these are convervative predictions considering there are many more products to look at and stores to sell to.

DELIVERABLES

There are a couple things I created that I want to hand over.
I have created an Excel workbook with multiple sheets. Each sheet is for a product. On each sheet, I have the information from my predictions.
The stores are ordered from best to worst stores to target for distribution of a product.

I have also created a Tableau Dashboard for the company to allow them to further explore the data themselves.

RECOMMENDATIONS

So, by following my recommendations, I believe the profits lost from the Pepperidge Farm supply issues can be recovered.

My first recommendation would be some upgrades for Dale Cox Distributing.
Although I was able to retrieve the data needed, it was not easy.
Upgrading the database, systems, and software would allow for more access to the data for better analysis. With more analysis, I believe more profits can be generated.
My second recommendation is to push the 7 controlled labels into more stores.
I have provided the stores to target for each product.

NEXT STEPS

I have more plans on how I can continue to help the company.
There is a lot more data to explore that I believe I can provide valuable analysis on.
I want to specifically look into the locations of stores and the delivery routes to weigh the cost of distributing to stores based on how much they order

THANKS

Thank you all for your time, I'd be happy to answer any questions.