Final Project

CECS 551 Advanced Artificial Intelligence

Professor: Dr.Mahshid Fardadi

Submitted to

Allen Bolourchi, PhD,MBA

Submitted by

Yashpaul Vallabhapurapu, 029349864

Priyanka Ponnaganti, 029349825

Madhu Kiran Pathuri, 029415761

Deepak Goud Mamidi, 029426928

California State University, Long Beach

March 2022

Deliverables:

1. Drive link to complete project runnable code an necessary files are located here:
2. Report for
    a. Data analysis and its visualization
    b. Model building and predicting
    c. Thought provoking initiatives and Explainer Dashboards
3. Colab Link to complete runnable code
4. Colab Link to Explainer Dashboards
5. PPT for presentation

Team Members' Roles:

Yashpaul V	Part A 5,6, for individual stores and its analysis, Part B, 4,5 and its analysis. Temp data gathering and adding other features to dataset
Priyanka P	Part A 7th,9th and Overall Documentation
Madhu Kiran	Part A 4th and few analysis points
Deepak God	Part A 8th and few analysis points
Mandy He	Part A 1-3 for three individual stores (18, 117, 332(, and its corresponding questions in Part B for all stores)).

Section 1: Data Analysis and Visualization

Part A

Analyze the data of individual stores and draw insights.

A-1: Provide the box plots and statistics of the last 14 products, inventory patterns, stock out patterns and missed sales.

As assigned, we are working on the last 14 products as follows:

['Honey Raisin Bran Muffin', 'Jalapeno Cheese Bagel', 'Lemon Loaf', 'Mixed Berries & Granola Yogurt Parfait', 'Mixed Fruit Snack Pot', 'Muffin - Blueberry Streusel', 'Muffin - Double Chocolate', 'New York Cheesecake','Plain Bagel', 'Pressed Juicery Spicy Greens w Ginger', 'Protein Box', 'Smoked Salmon Sandwich', 'Tasty Tuna Salad Sandwich', 'Vive Juice Shot']

a) *Overall statistics for inventory patterns for store 18, 117, 332.*

For inventory patterns, we first look at the overall statistics for ReceivedQuantity, StockedOut, and MissedSales. Store 332 receives more inventory among the three stores, and also has higher missed sales according to the statistics.

Store 18	Store 117

	ReceivedQuantity	StockedOut	MissedSales
count	4793.000000	4793.000000	4749.000000
mean	3.132068	0.207177	0.245770
std	2.985545	0.405326	0.681091
min	0.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000
50%	2.000000	0.000000	0.000000
75%	5.000000	0.000000	0.000000
max	24.000000	1.000000	6.691424

	ReceivedQuantity	StockedOut	MissedSales
count	4796.000000	4796.000000	4763.000000
mean	2.713720	0.255004	0.355623
std	2.596136	0.435909	0.868604
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	2.000000	0.000000	0.000000
75%	4.000000	1.000000	0.000000
max	12.000000	1.000000	6.261246

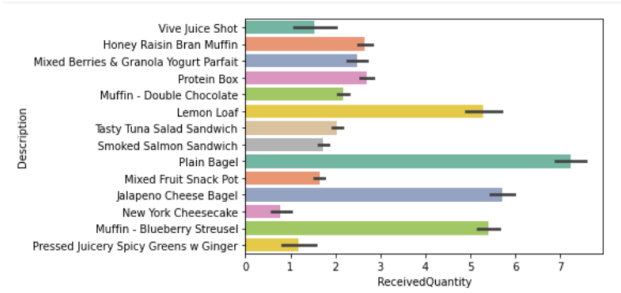
Store 332

	ReceivedQuantity	StockedOut	MissedSales
count	4906.000000	4906.000000	4902.000000
mean	5.227069	0.204647	0.299117
std	4.691003	0.403485	0.819046
min	0.000000	0.000000	0.000000
25%	1.000000	0.000000	0.000000
50%	4.000000	0.000000	0.000000
75%	8.000000	0.000000	0.000000
max	30.000000	1.000000	7.020856

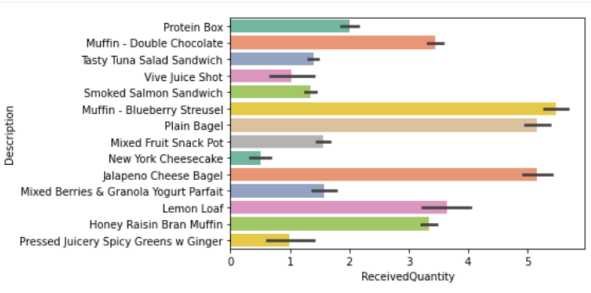
b) *Bar plot for inventory patterns (ReceivedQuantity) for store 18, 117, 332 per products:*

The per product bar chart shows the average daily received quantity per product at each store. The more inventory usually indicates the more popular the products. We can see there are several common products popular across all stores. We also see some products are uniquely popular in one store. We analyze the most popular products in the later part. This chart shows store 332 holding more inventory on the most popular products that may indicate a higher sales of them.

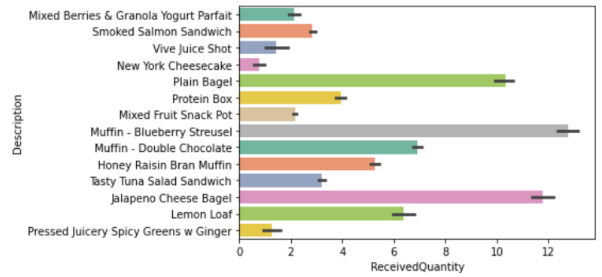
Store 18



Store 117



Store 332



c) *Box plot for inventory patterns (ReceivedQuantity) for store 18, 117, 332 per products:*

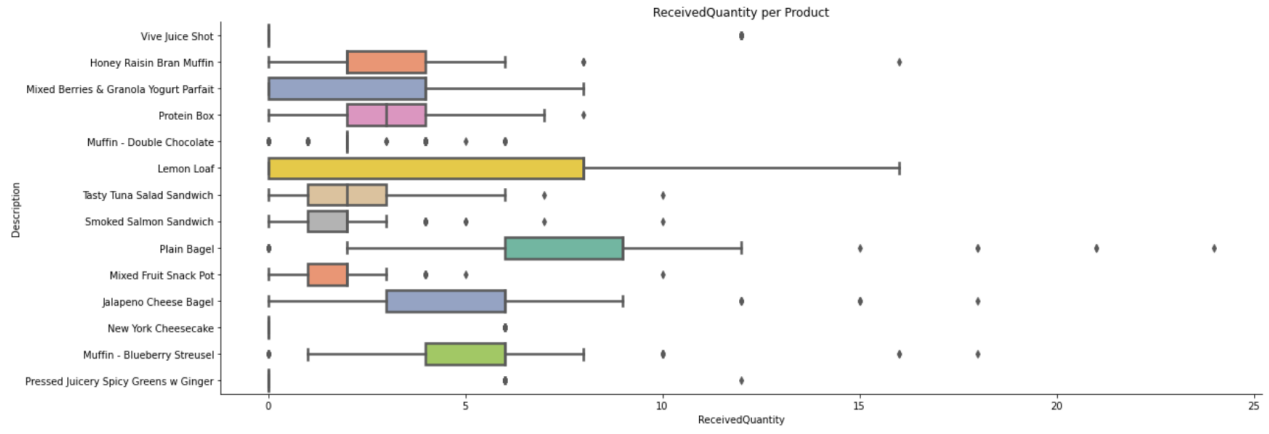
Below box plots show the distribution of inventory (ReceivedQuantity) per product for three stores. Across all three stores, plain bagel, jalapeno cheese bagel, and muffin- blueberry streusel have a stable, relatively higher number of inventory. These products are also sold well with good demand so stores maintain a stable available inventory. On the other hand, there are some products that their inventories are not stable, sometimes stores have more inventory other times less inventory. These pro ki

ducts are:

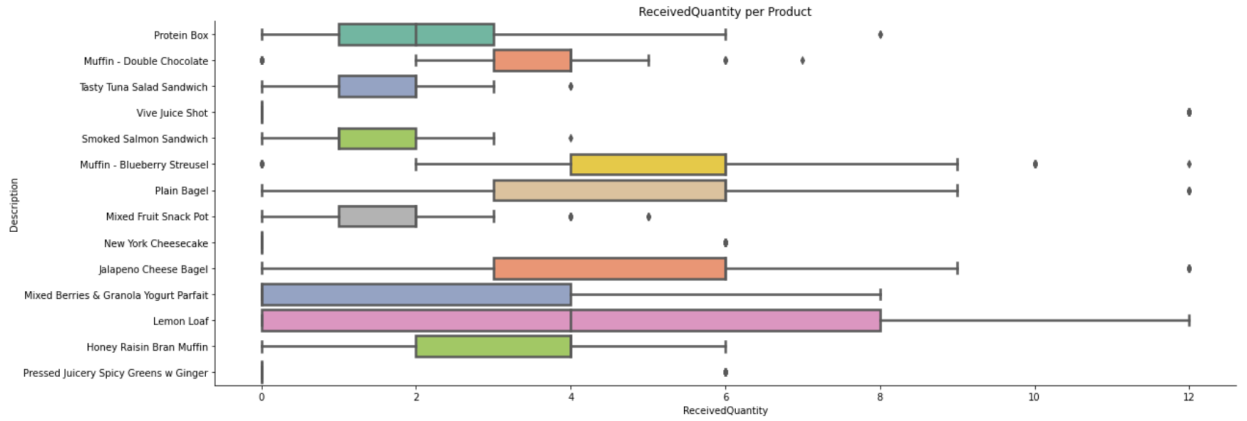
- Store 18: Lemon loaf, mix berries & granola yogurt parfait
- Store 117: Lemon loaf, mix berries & granola yogurt parfait
- Store 332: Mix berries & granola yogurt parfait

Overall, store 332 maintains a higher inventory for the most popular products, which might indicate this store overall has a good sales performance.

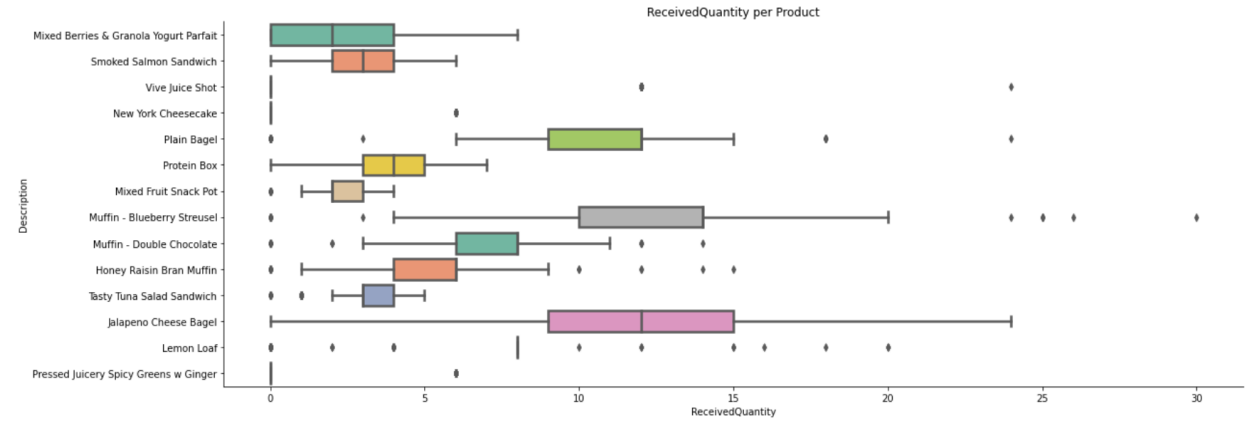
Store 18:



Store 117:



Store 332:



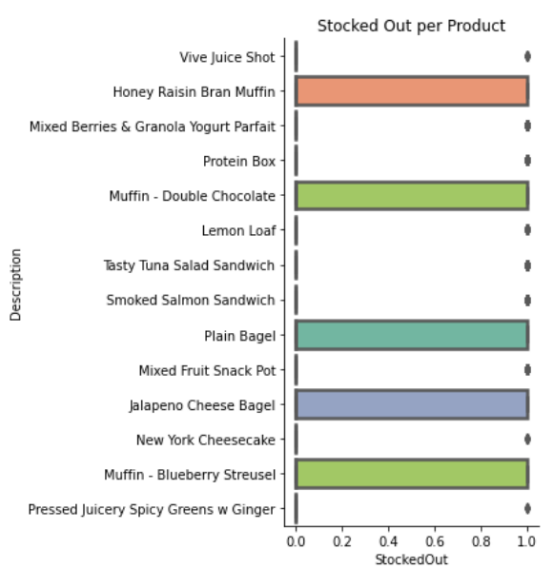
d) <u>Box plot for stock out for store 18, 117, 332</u>

- Below charts plot the stocked out per product for three stores. Stocked out products indicate running out-of-stock and missed sales opportunities/revenue. Overall store 18 regularly has 5 products often running out, while store 117 has 6, store 332 has 4 products. Again, store 332 manages inventory more efficiently than the other two stores, particularly store 117.

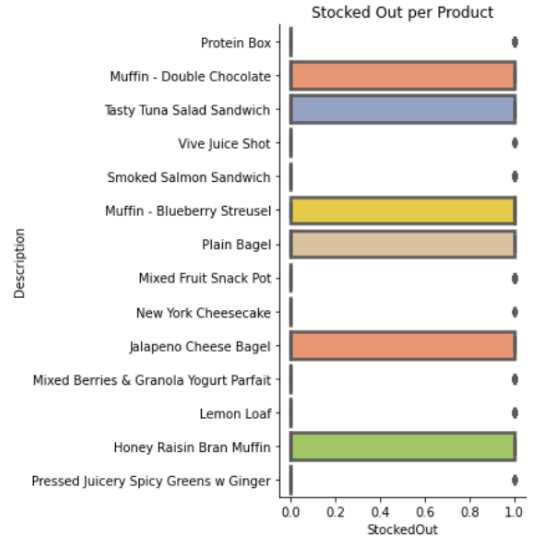
- It is also noticed that the three most popular products, plain bagel, jalapeno cheese bagel, and muffin- blueberry streusel are often out of stock in store 18 and 117, while store 332 manages to have it available except sometimes run out plain bagel. If these products are the best sold ones, maintaining a higher inventory of them will bring more sales and revenue.

- Noticed that store 332's out of stock products often happen to low sales products. So it might not hurt the business very much. But the other two stores shall have more space to improve their sales by maintaining a good inventory for the best sold products.

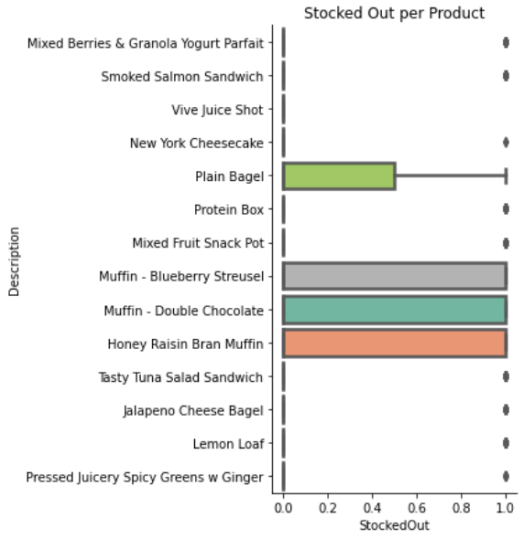
Store 18:



Store 117:



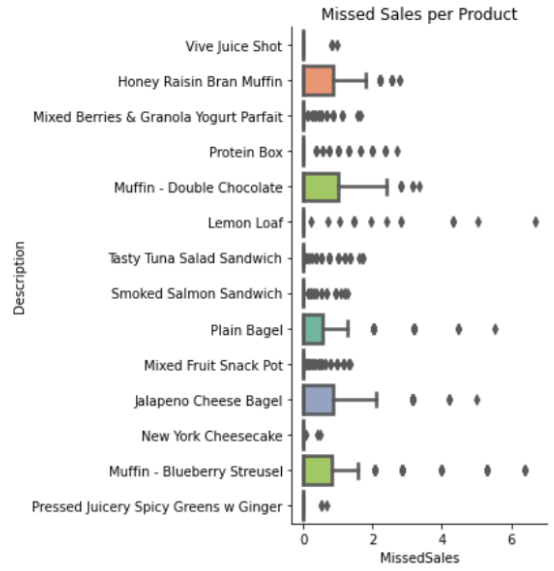
Store 332:



e) <u>Box plot for missed sales:</u>

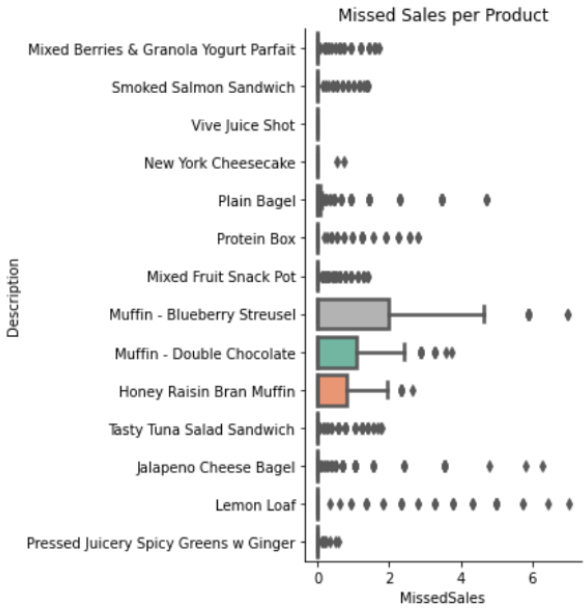
- Even though the document mentions missed sales records might not be reliable, we plot it for more insight. Key findings are:
- Store 332 has a consistent record for stocked-out and missed sales. Meanwhile, its missed sales happened mostly for those un-popular products that usually maintain a low inventory and high out-of-stock records.
- The other two stores (18, 117) have missed sales for best-sold products and medium-popular products. It indicates there is a big space for store 18 and 117 to improve inventory to boost up sales.

Store 18:



Store 117:

Store 332:



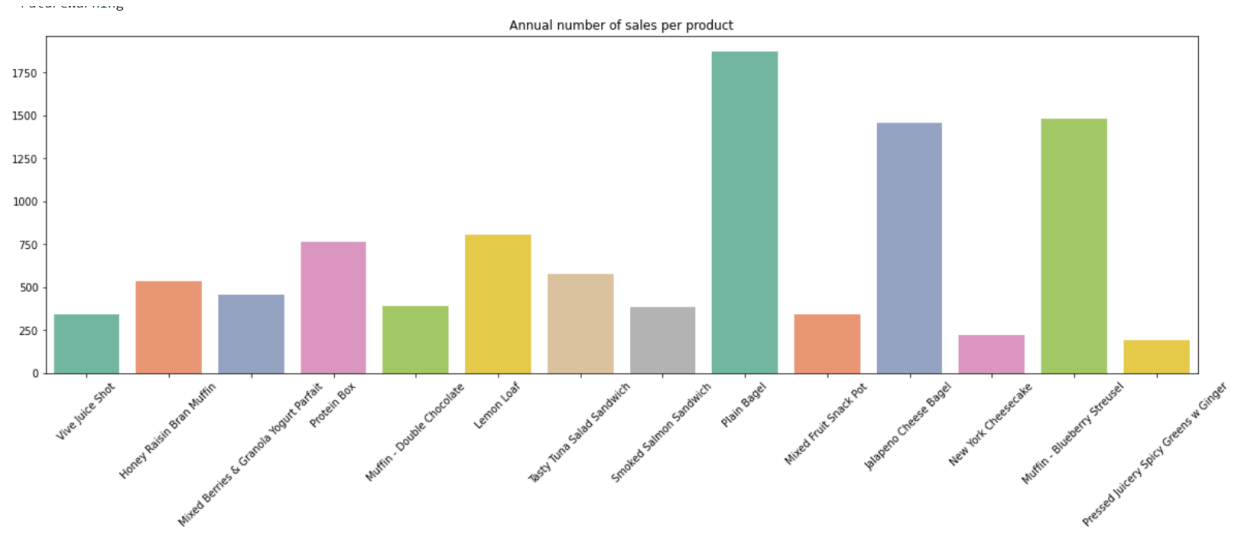
A-2: Show graphs of best seller and worst seller products of top 25% and bottom 25% and provide your insight into data.

To determine the best and worst seller product, let's look at the total annual sales of each product first. It reflects the accumulated selling performance within the data reporting period. In further, we plot the line plot of each product according to the soldQuantity every day. This will display the sales stability during the data reporting period. Notice we will analyze it based on sold quantity rather than revenue due to each product having a different price.

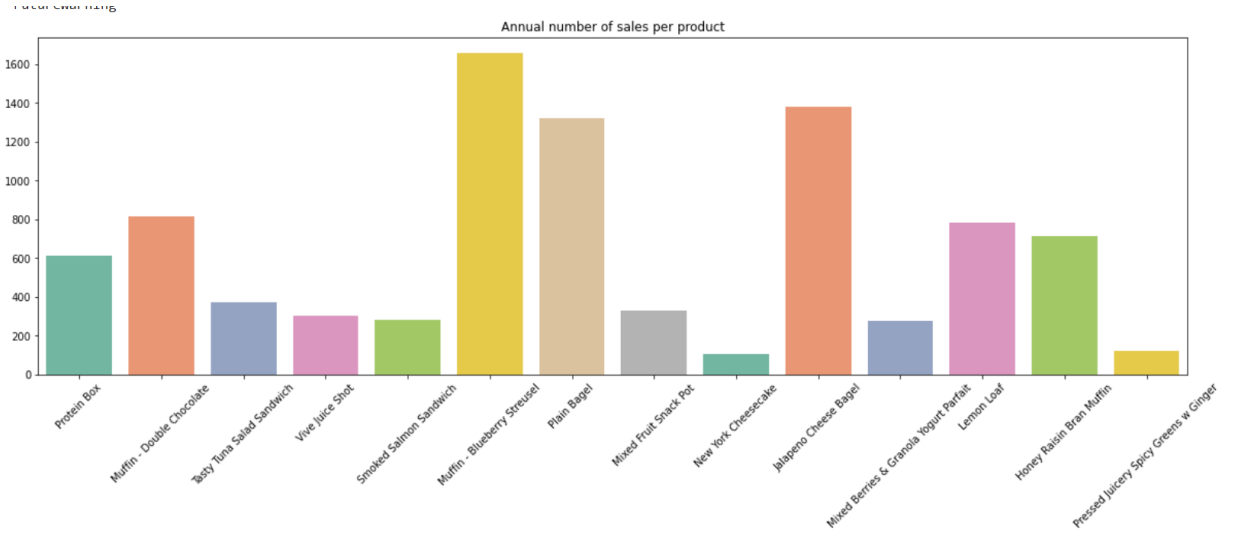
a) *Annual sold quantities for each product with the top 25% best product sales*

Below bar plot shows the total sold quantities per product for each store. The most sold three products are common in three stores, they are: Jalapeno cheese bagel, muffin- blueberry streusel, and plain bagel.

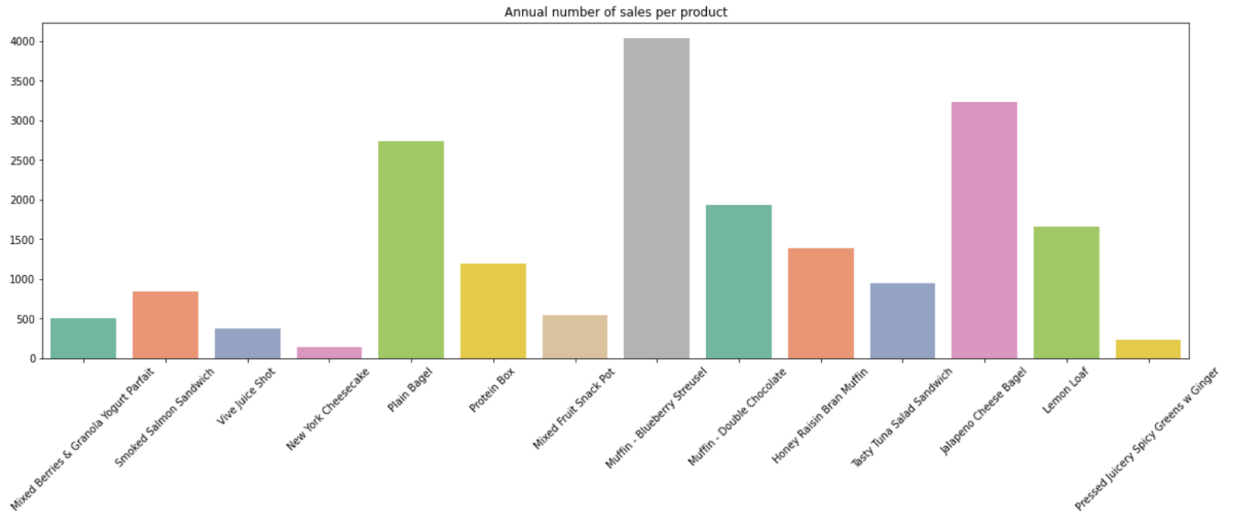
Store 18:



Store 117:



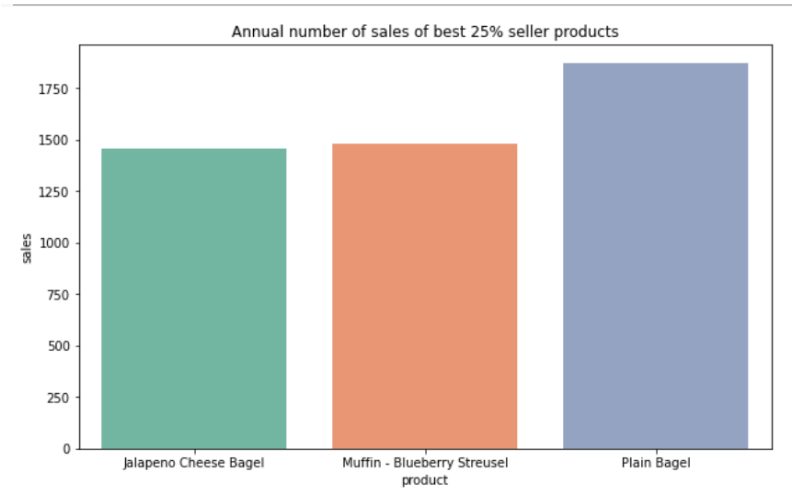
Store 332:



b) *Plot top 25% product total annual sales*

Store 332 has the highest sales to the top best sellers. For example, 4034 muffin-blueberry streusel are sold in store 332, while store 18 and 117 sold only 1480 and 1656.

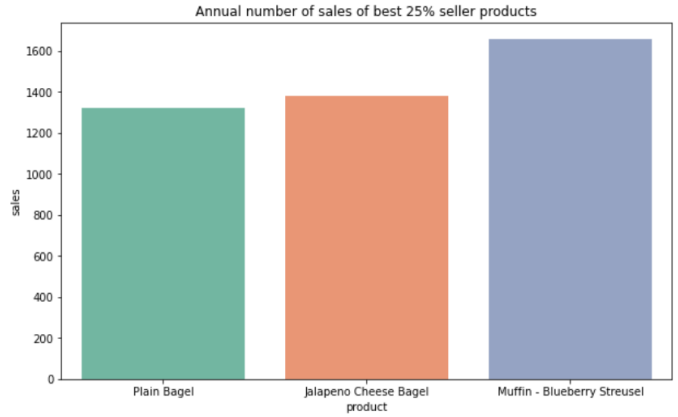
Store 18:



top sales:

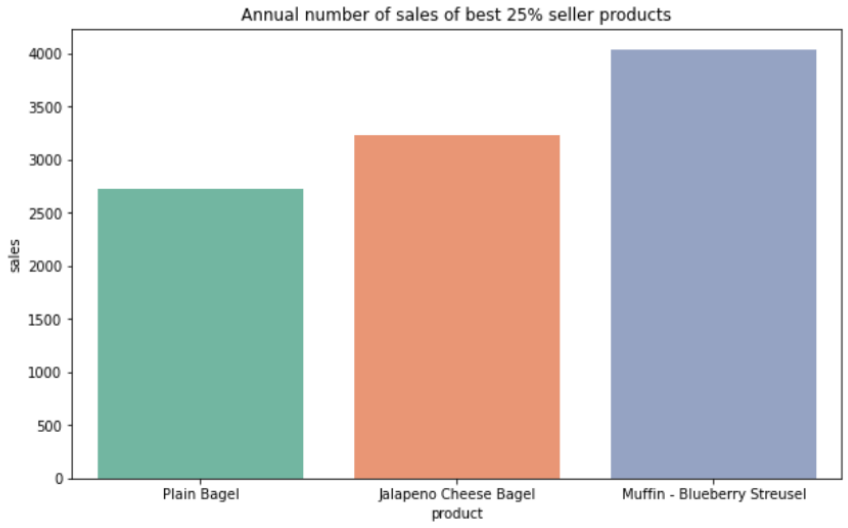
	product	sales
11	Jalapeno Cheese Bagel	1454
12	Muffin - Blueberry Streusel	1480
13	Plain Bagel	1870

Store 117:



```
top sales:
                         product  sales
11                  Plain Bagel   1321
12        Jalapeno Cheese Bagel   1378
13  Muffin - Blueberry Streusel   1656
```

Store 332:



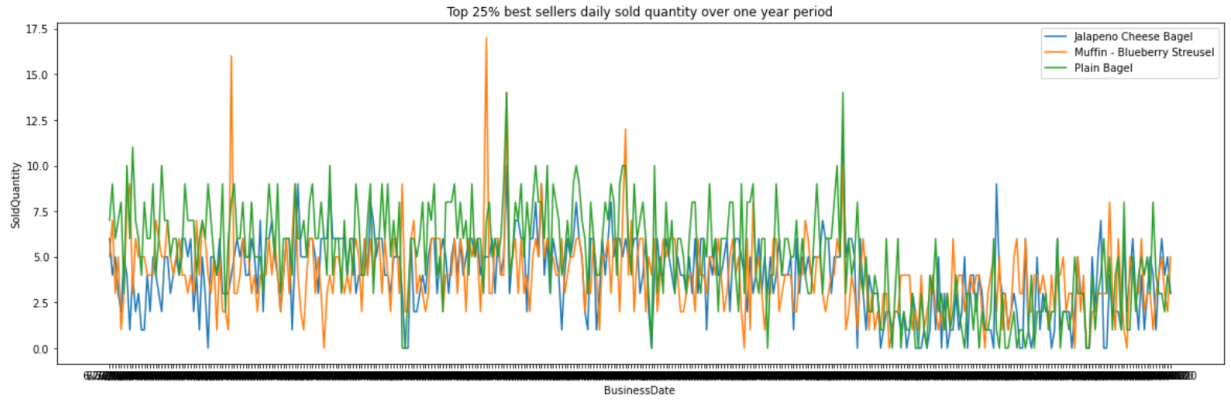

```
top sales:
                         product  sales
11                  Plain Bagel   2731
12        Jalapeno Cheese Bagel   3231
13  Muffin - Blueberry Streusel   4034
```

c) *Top 25% best sellers daily sales*

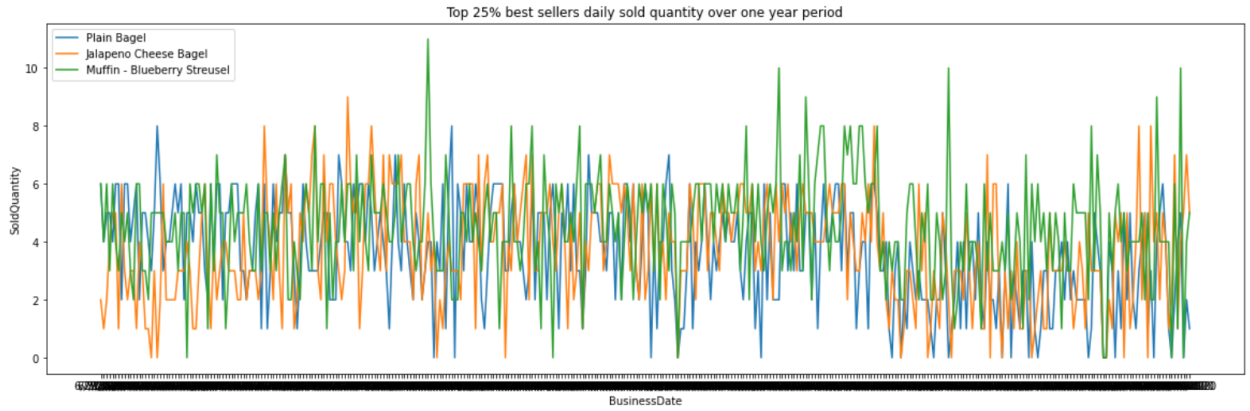
Below graph shows daily sold quantities for the top 25% best sellers. Our key findings are as follows:

- Sales quantities vary in a pattern, which could be due to the day of a week, or weather. This can be analyzed in question 5 with the new features of weather and date.
- Sales have seasonal changes. There is a dip and sudden increase for some stores.
- There are several peak sales days that could be due to holiday. This can be analyzed in further with the holiday, weather, date feature in question 5 and 7.

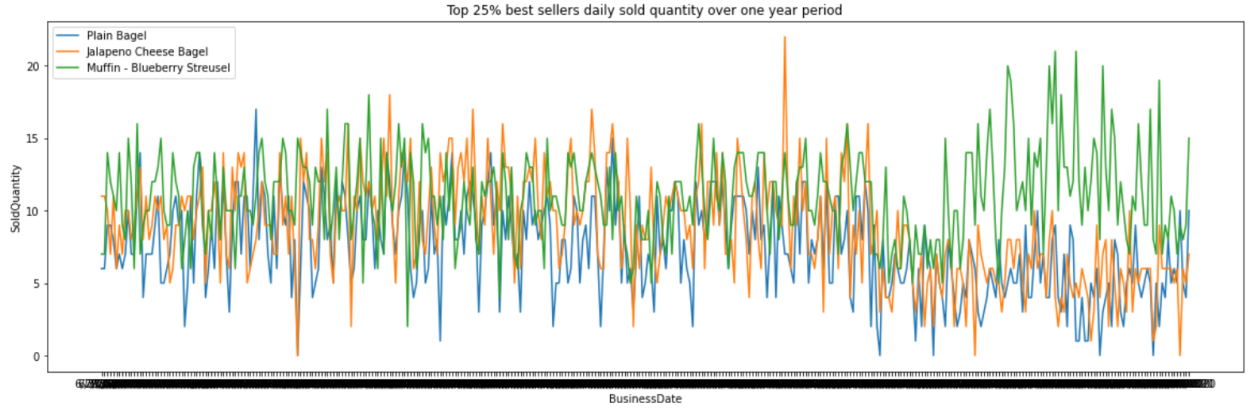
Store 18:



Store 117:



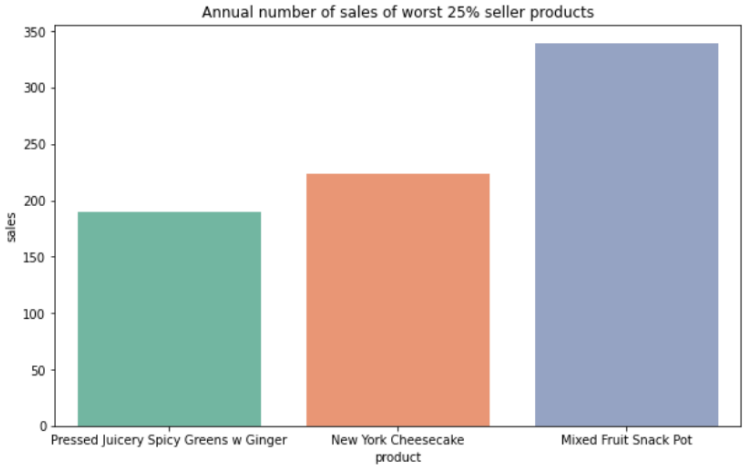
Store 332:



d) *<u>Plot bottom 25% worst products total annual sales</u>*

Below plots show the bottom 25% sellers annual sales. Their sales columns are quite small. But they can provide complementary products that consumers may need.

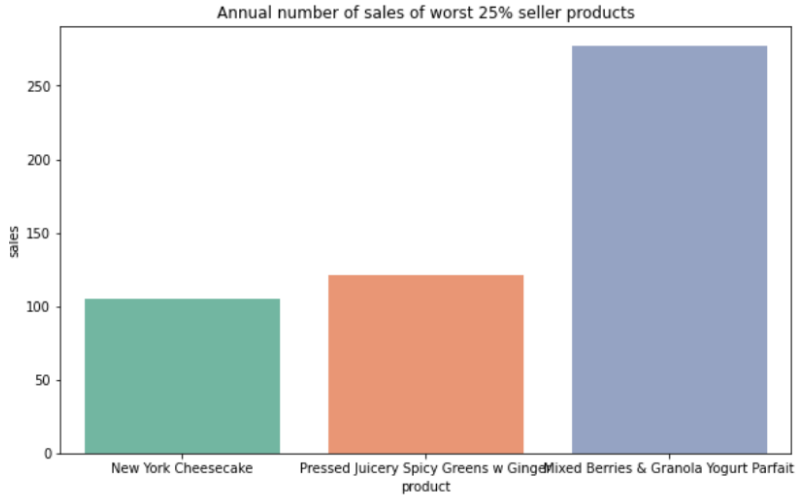
Store 18:



bottom sales:

	product	sales
0	Pressed Juicery Spicy Greens w Ginger	190
1	New York Cheesecake	224
2	Mixed Fruit Snack Pot	339

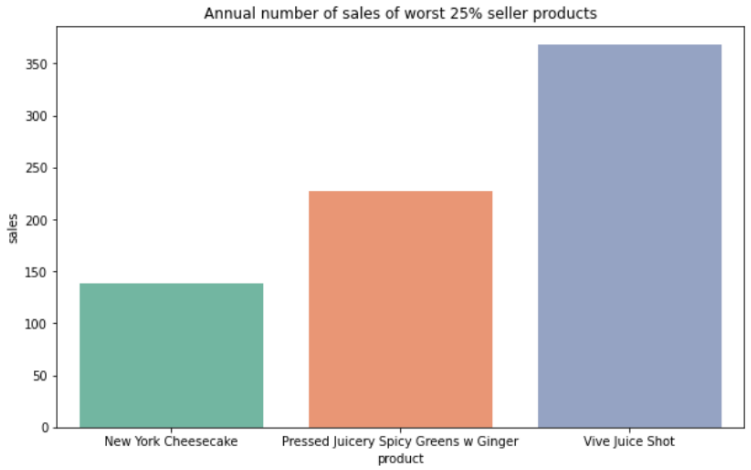
Store 117:



bottom sales:

	product	sales
0	New York Cheesecake	105
1	Pressed Juicery Spicy Greens w Ginger	121
2	Mixed Berries & Granola Yogurt Parfait	277

Store 332:



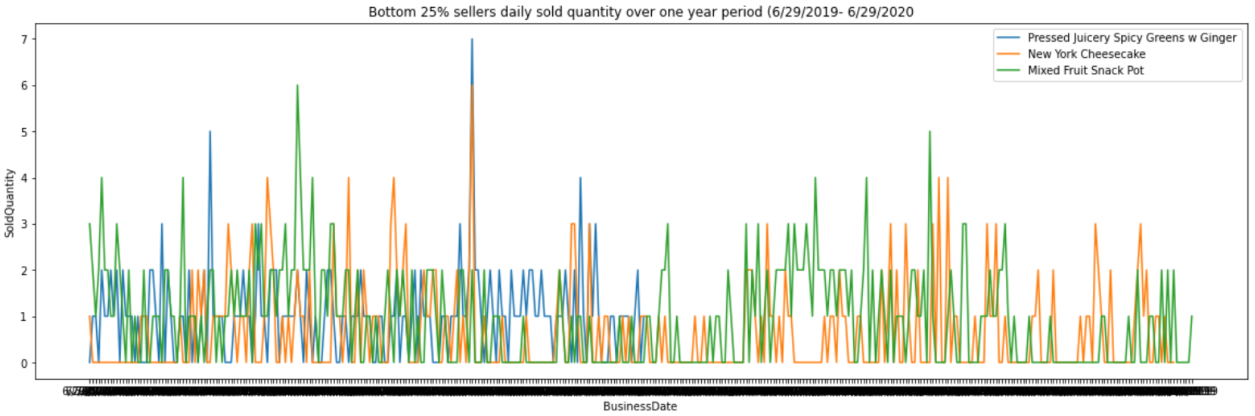
bottom sales:

	product	sales
0	New York Cheesecake	138
1	Pressed Juicery Spicy Greens w Ginger	227
2	Vive Juice Shot	368

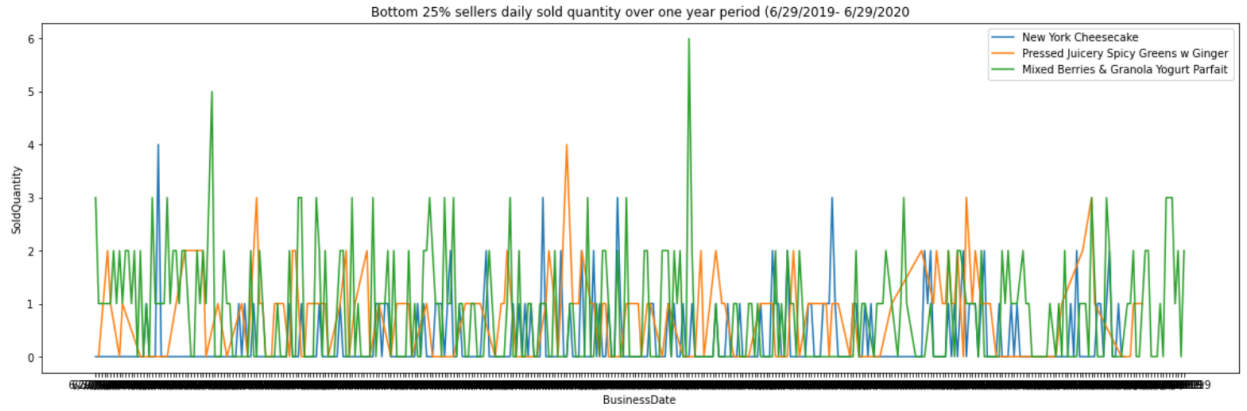
e) *Bottom 25% worst sellers daily sales*

Below graph shows daily sold quantities for the bottom 25% worst sellers. Some products' daily sold quantity can be close to zero. Worst sellers do not boost up total sales and revenue, but as we discussed they may provide a nice complement. Meantime, it is also noticed there are several days these bottom sellers' sales go up, which might be due to seasons or holidays.

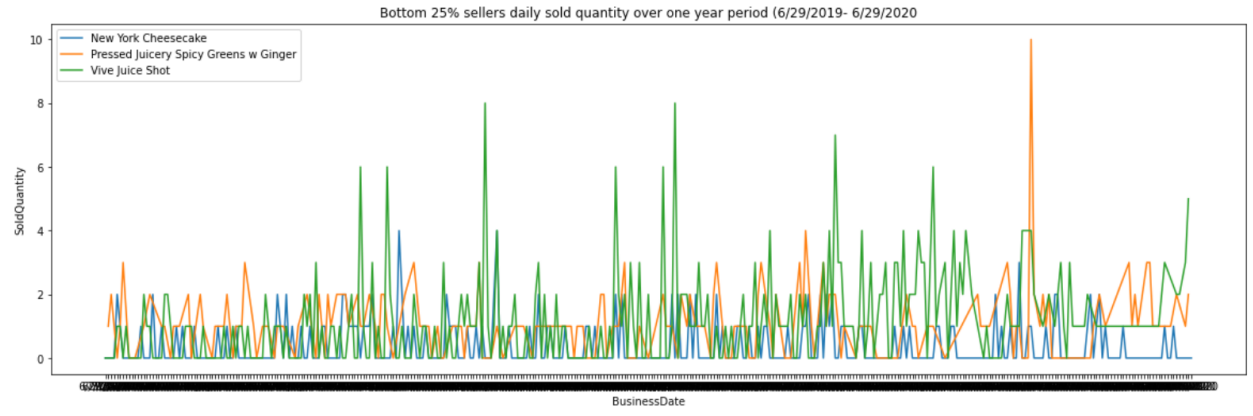
Store 18:



Store 117:



Store 332:



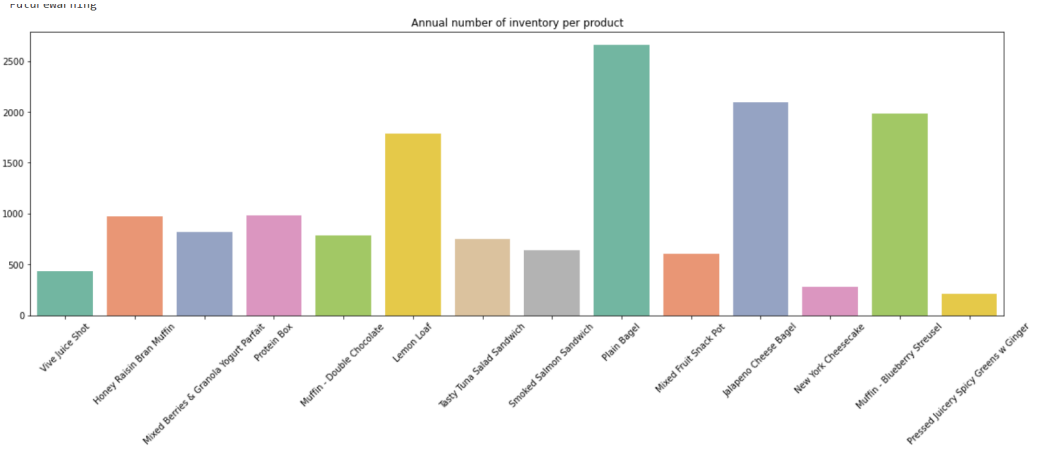
A-3: Show graphs of best and worst products based on their inventory management - Top 25% and bottom 25% and provide your insight into data.

To determine the best and worst seller product, let's look at the total annual sales of each product first. It reflects the accumulated selling performance within the data reporting period. In further, we plot the line plot of each product according to the soldQuantity every day. This will display the sales stability during the data reporting period. Notice we will analyze it based on sold quantity rather than revenue due to each product having a different price.

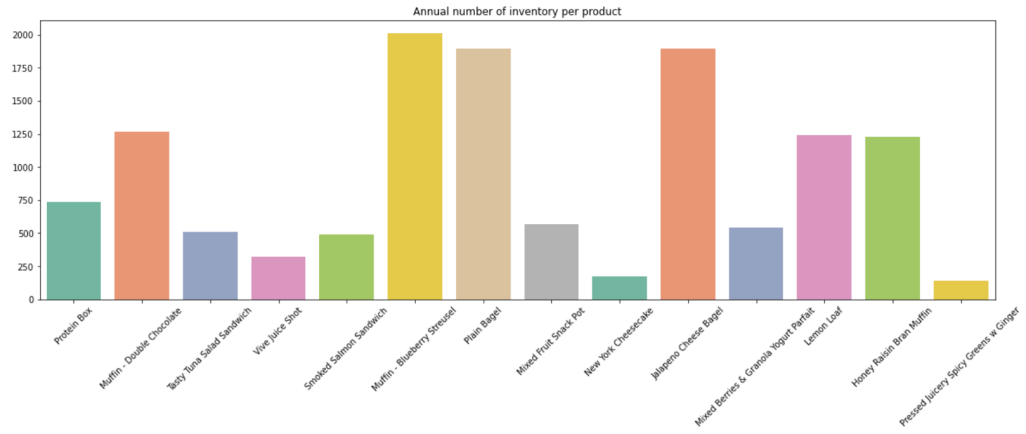
a) *Bar plot for annual inventory (Received Quantity) for all products*

Below bar plot shows the annual received quantity for each product in the three stores. Overall, store 332 has some inventory than store 18 and 117 because they have more sales. The inventory patterns across the three stores are similar.

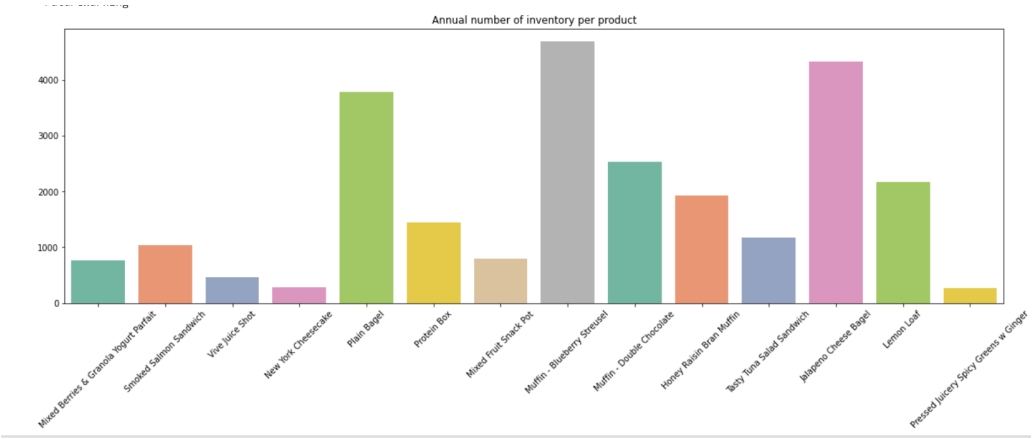
Store 18:



Store 117:



Store 332:



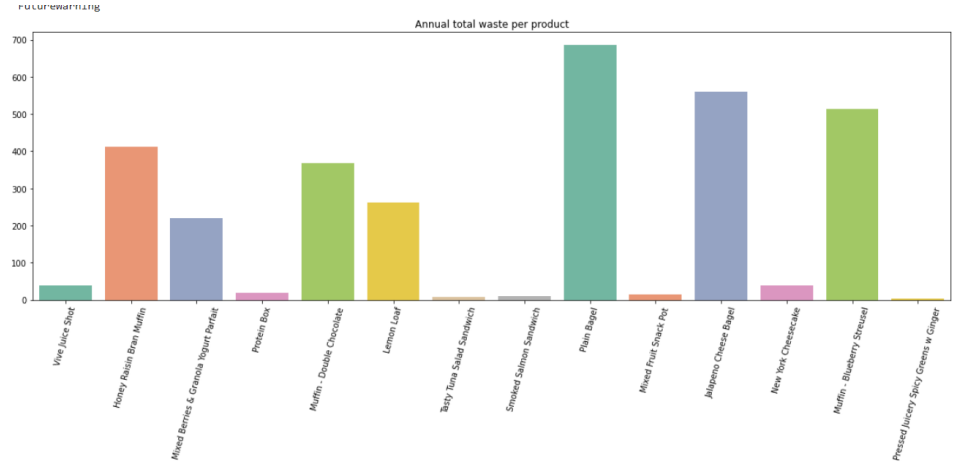
b) *Bar plot for annual waste for all products*

We calculate waste based on the following formula:

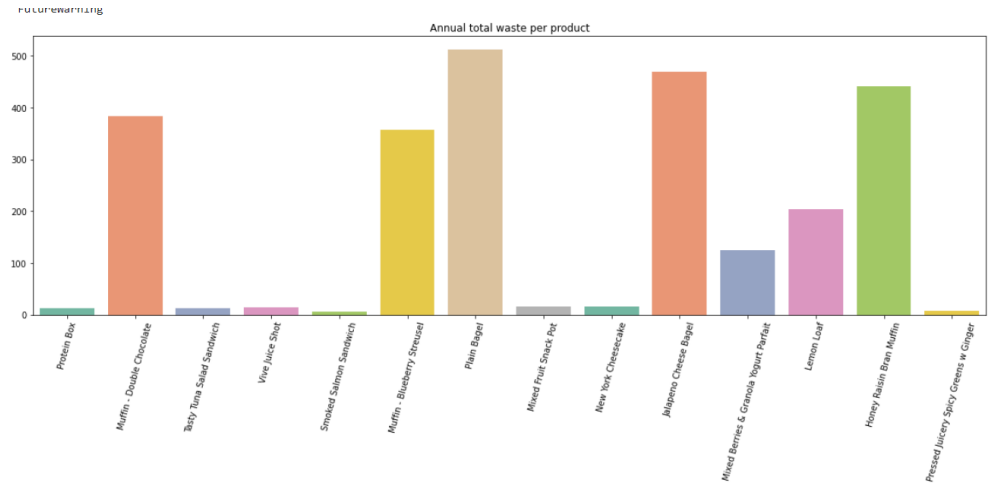
Waste = ReceivedQuantity - SoldQuantity - EndQuantity

We calculated each product's annual waste and benchmark across three stores, as shown in the bar plot below. The higher the bar the more waste of the product. We notice that some best seller products (e.g. plain bagel, jalapeno cheese bagel, etc) can also have many wastes. Meantime notice for those products, there are missed sales as well, which means that a good on-time inventory management is critical to reduce the waste and boost up sales to meet consumers' needs.

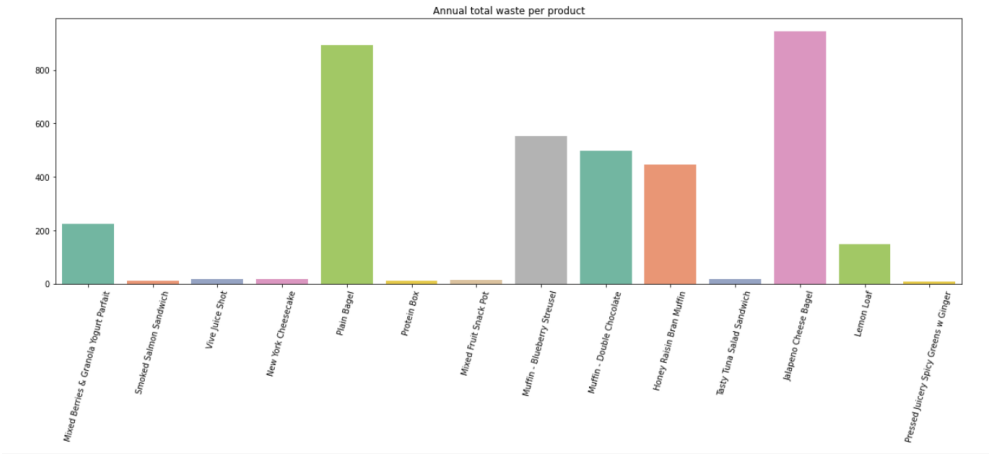
Store 18:



Store 117:



Store 332:



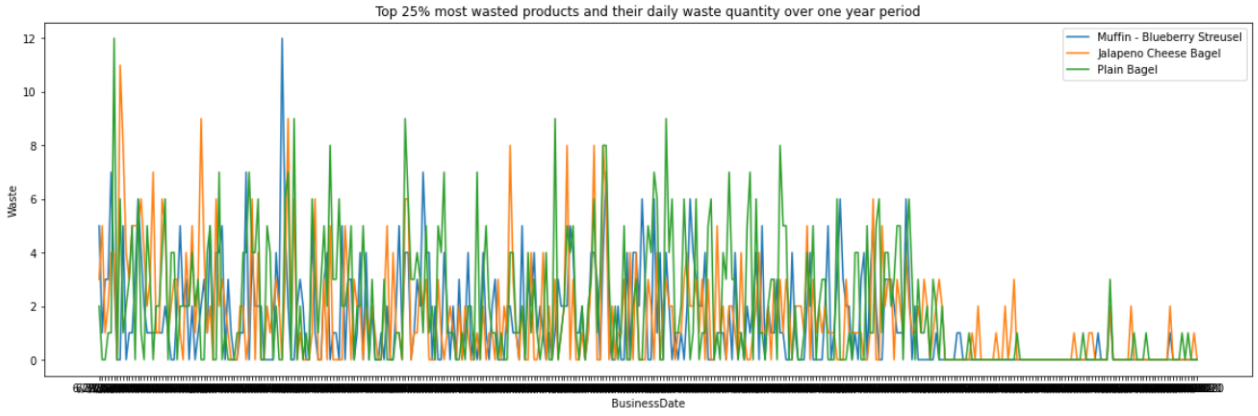
c) *25% Worst Inventory Management*

According to annual wastes per product, we sort out the most wasted 25% products and their annual wastes, as shown in below charts. We also identified their daily wastes so we know when these wastes happened the most. In general, from June 2019 to December 2019, there will be more waste. After that, waste becomes smaller, which could be due to company waste and inventory management improvement.

In general, waste is a big issue to the company, which is proportional to the sales columns. However, store 18 has the worst inventory management, that it has lower sales quantity but higher wastes.

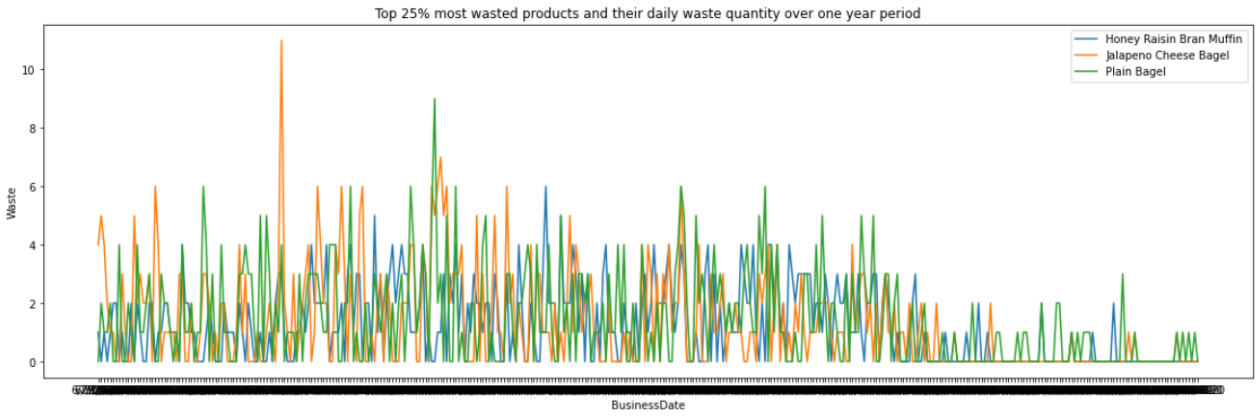
Store 18:

	product	waste
Muffin - Blueberry Streusel		513.0
Jalapeno Cheese Bagel		559.0
Plain Bagel		686.0



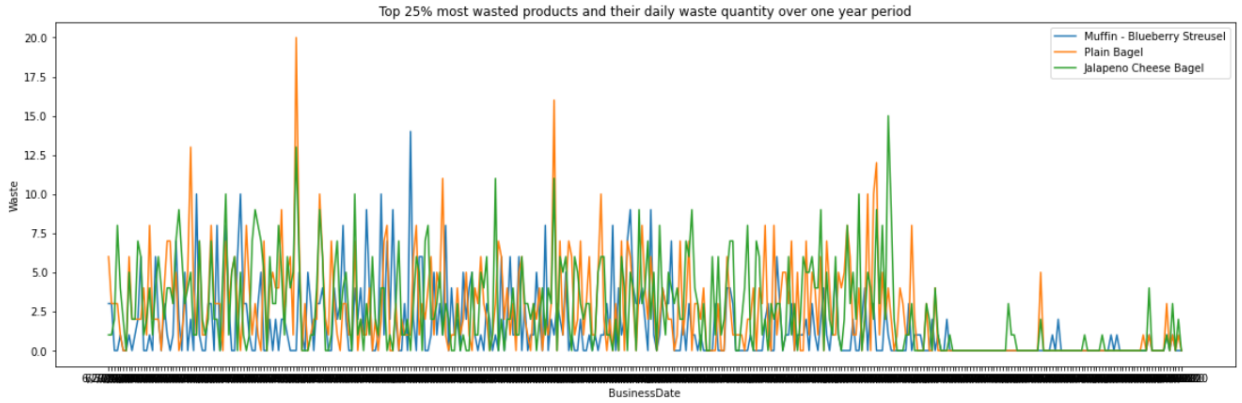
Store 117:

	product	waste
Honey Raisin Bran Muffin		441.0
Jalapeno Cheese Bagel		469.0
Plain Bagel		512.0



Store 332:

product	waste
Muffin - Blueberry Streusel	551.0
Plain Bagel	892.0
Jalapeno Cheese Bagel	945.0



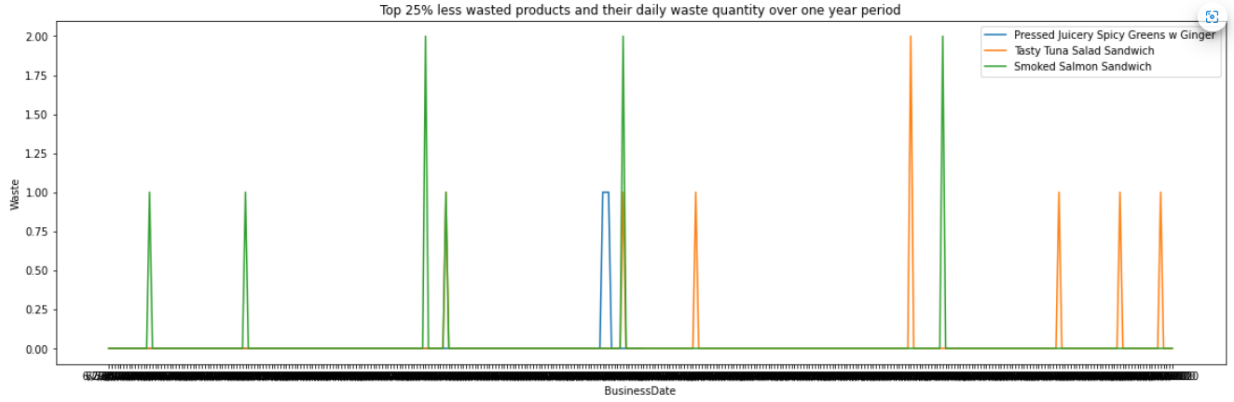
d) *25% Best Inventory Management*

According to annual wastes per product for three stores, we sort out the less wasted 25% products and their annual wastes, as shown in below charts. We also identified their daily wastes so we know when these wastes happened. In general, for the less wasted products, there is not a seasonal pattern, but these products are less popular products so each store might not have much inventory at all.

We observe from below charts that products are thrown periodically when they are not sold. Inventory management is a challenging topic for best store sales and revenue management.

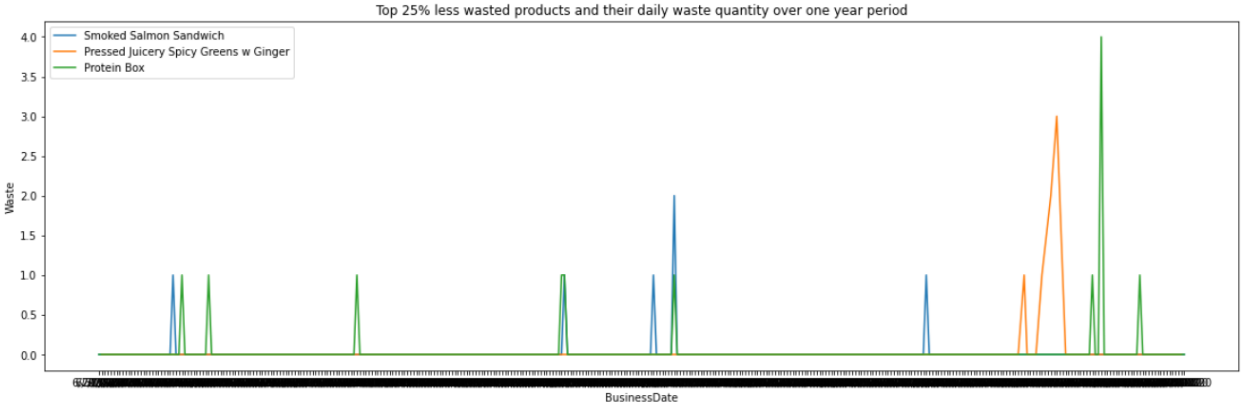
Store 18:

product	waste
Pressed Juicery Spicy Greens w Ginger	3.0
Tasty Tuna Salad Sandwich	8.0
Smoked Salmon Sandwich	9.0



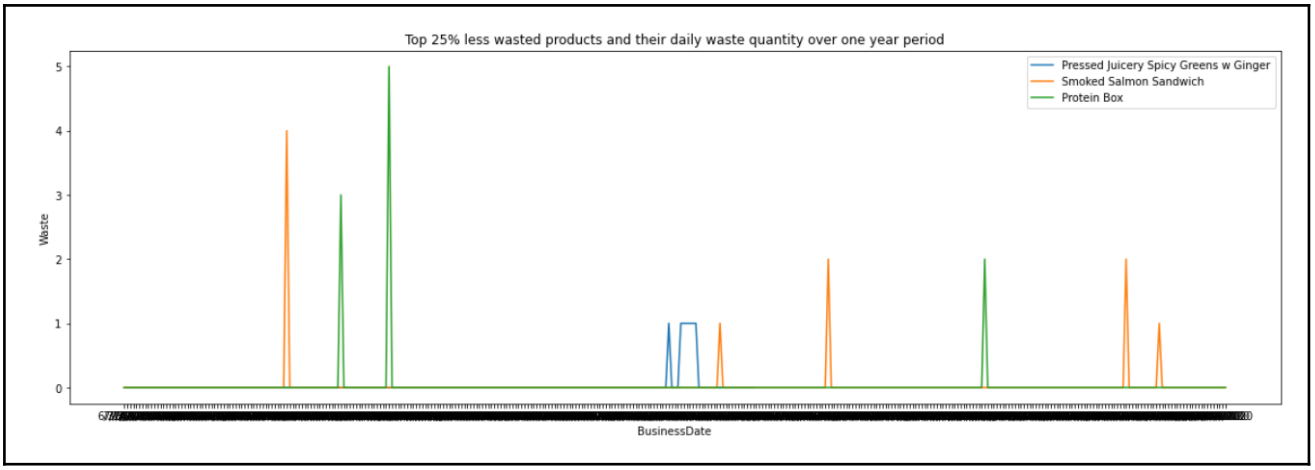
Store 117:

product	waste
Smoked Salmon Sandwich	6.0
Pressed Juicery Spicy Greens w Ginger	7.0
Protein Box	12.0



Store 332:

product	waste
Pressed Juicery Spicy Greens w Ginger	7.0
Smoked Salmon Sandwich	10.0
Protein Box	10.0



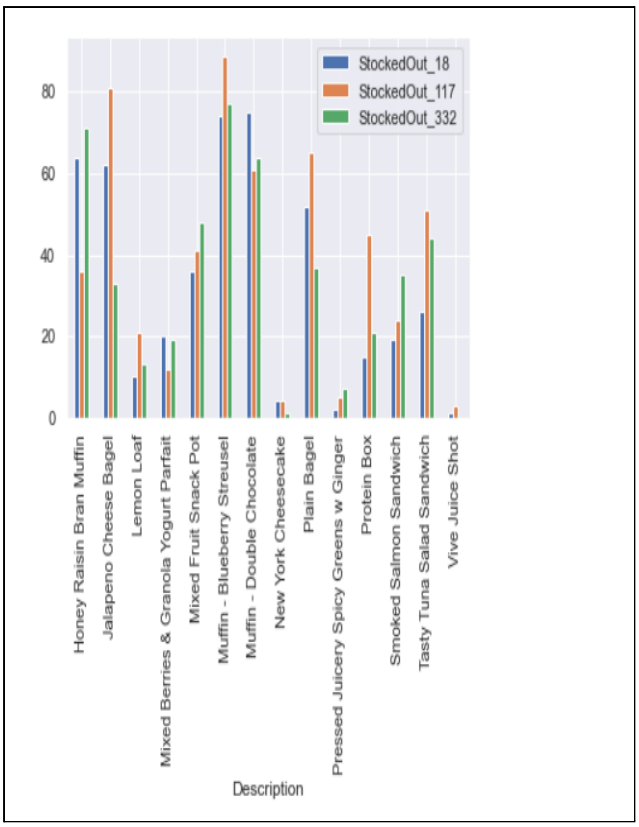
A-4 : Identify stock outs and estimate the loss of sales per year per product.

a) *Grouped bar plot for Loss of sales per year per product.*

As given , the average selling price of the product is $3 and the average cost of the product is $0.5 . Therefore if one product is stoked out, loss per item would be **$2.5 (3-0.5)**.

According to the below graph The highest stock out item in all the three stores is Muffin-Blueberry Streusel,So it would be profitable if we increase the Latest order for Muffin-Blueberry Streusel, then next comes Muffin-double chocolate. By this we can consider that people are more interested in the Muffins and Scones category. So we need to increase our stock to get more profit out of them. Vive juice shot is the lowest stoked out product ,due to which loss is less.

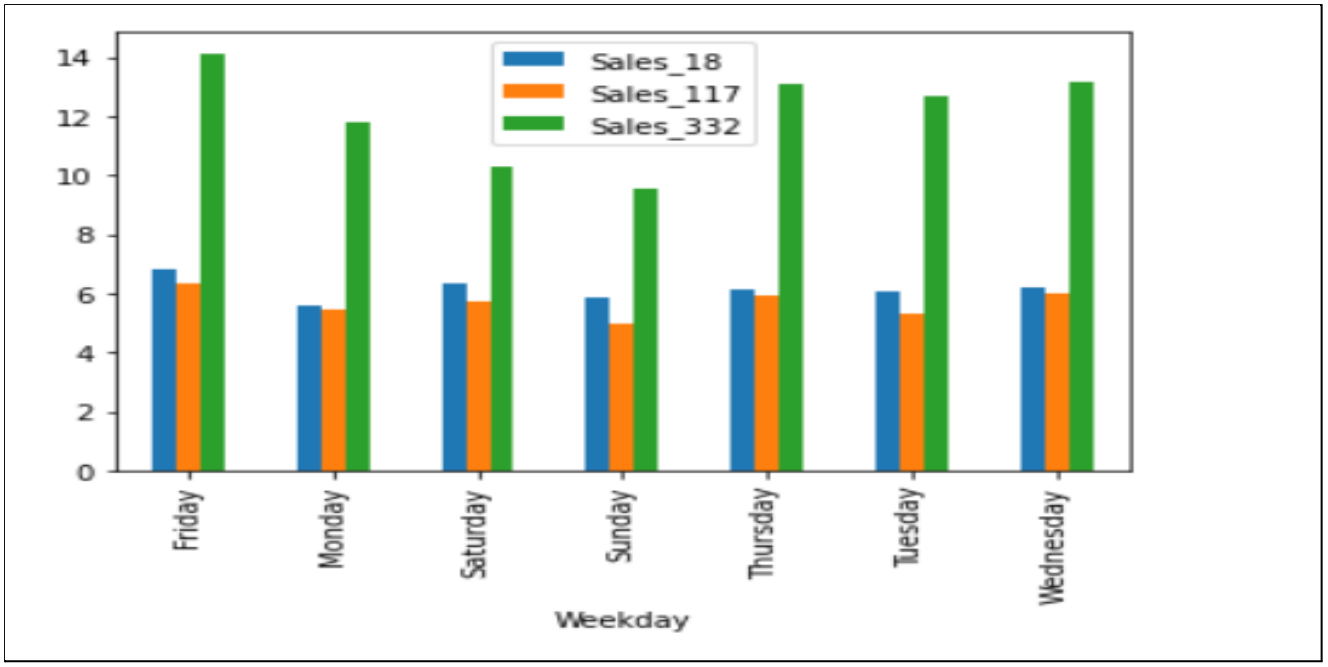
2019 **2020**

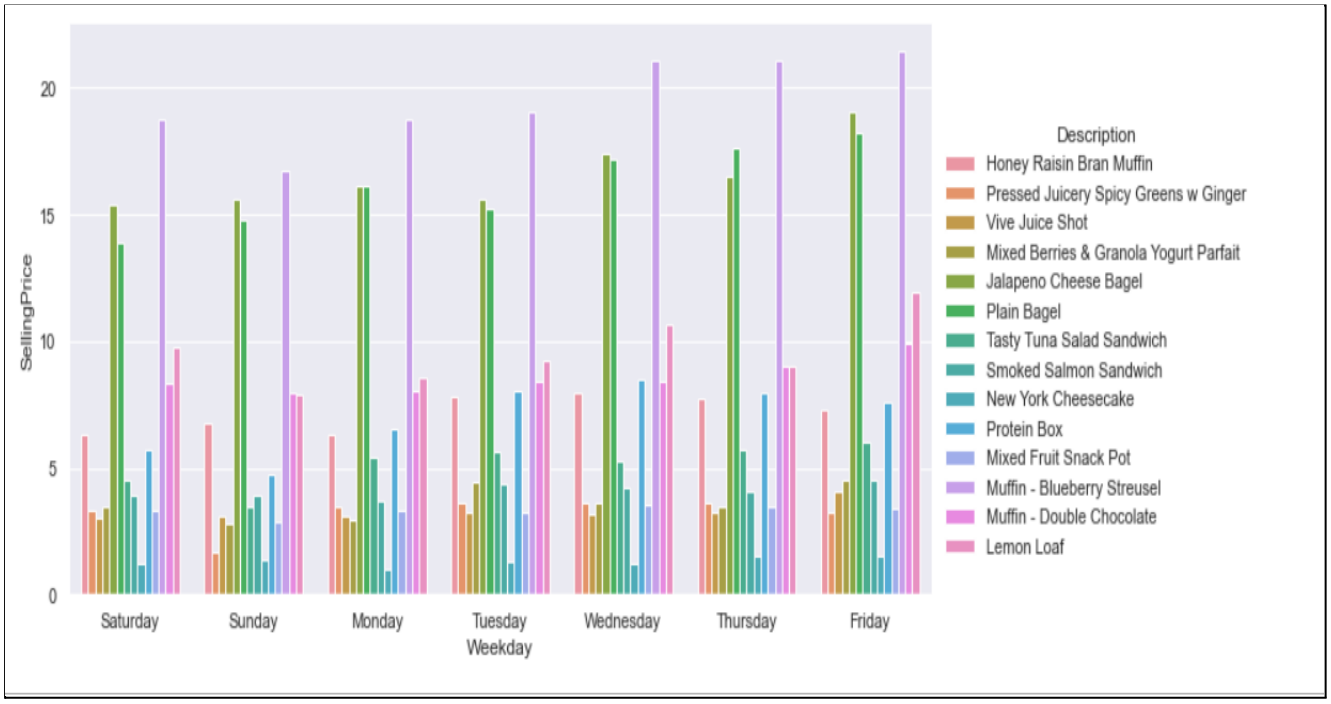


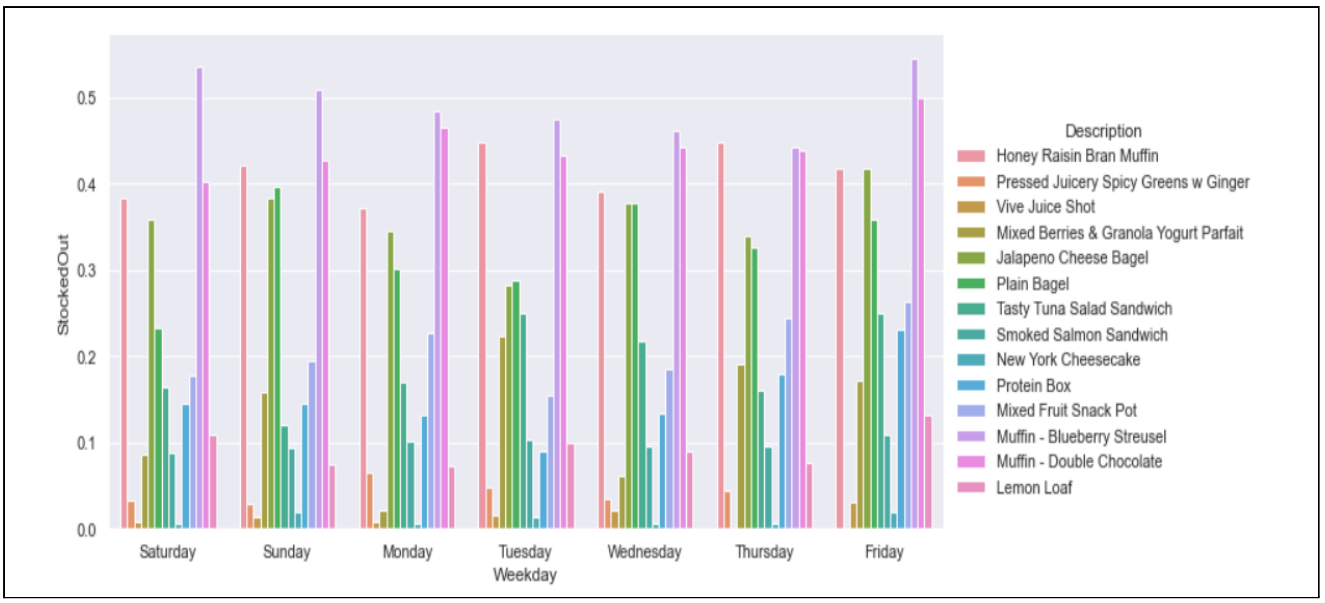
A-5: Show graphically how the product sales and inventory waste change.

a) *Impact of day of the week on sales and stocks*

By the graph we can observe that the sales are high on friday for all the three stores. Clearly store 332 has the highest sales as it has drive through which is not in other two stores . In comparison between store 18 and 117 store 117 has the highest sales. As we can see in the graph on Sunday store 332 has lowest sales this may be due to low drive through orders when compared to other days . By these trends we can clearly say that drive through orders are more when compared to dine in , so introducing drive through options in the other two stores would increase the sales and profits .


Sales_18
Sales_117
Sales_332
14
12
10
8
6
4
2
0
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Weekday

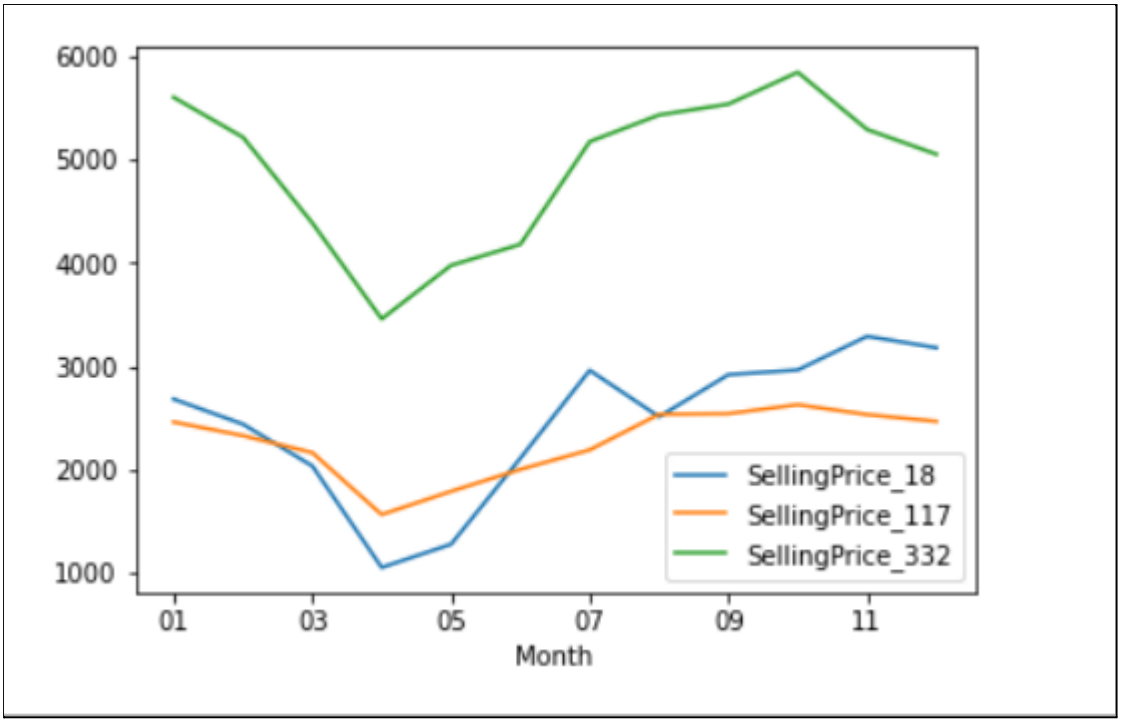

SellingPrice
20
15
10
5
0
Saturday
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Weekday
Description
Honey Raisin Bran Muffin
Pressed Juicery Spicy Greens w Ginger
Vive Juice Shot
Mixed Berries & Granola Yogurt Parfait
Jalapeno Cheese Bagel
Plain Bagel
Tasty Tuna Salad Sandwich
Smoked Salmon Sandwich
New York Cheesecake
Protein Box
Mixed Fruit Snack Pot
Muffin - Blueberry Streusel
Muffin - Double Chocolate
Lemon Loaf



b) *Monthly changes and patterns:*

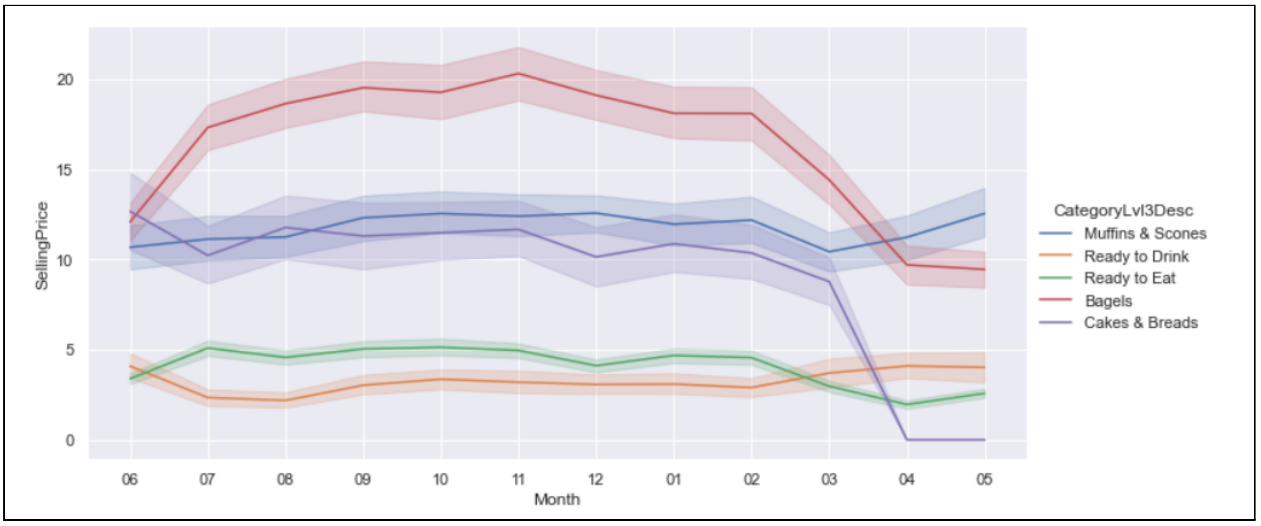
1) *Monthly sales of three individual stores:*

According to the graph we can observe that the sales decreased in the months of March, April and May, this may be due to the effect of covid people preferring not to come out of houses. Again there is an increase in the sales in November and December as it is the holiday season



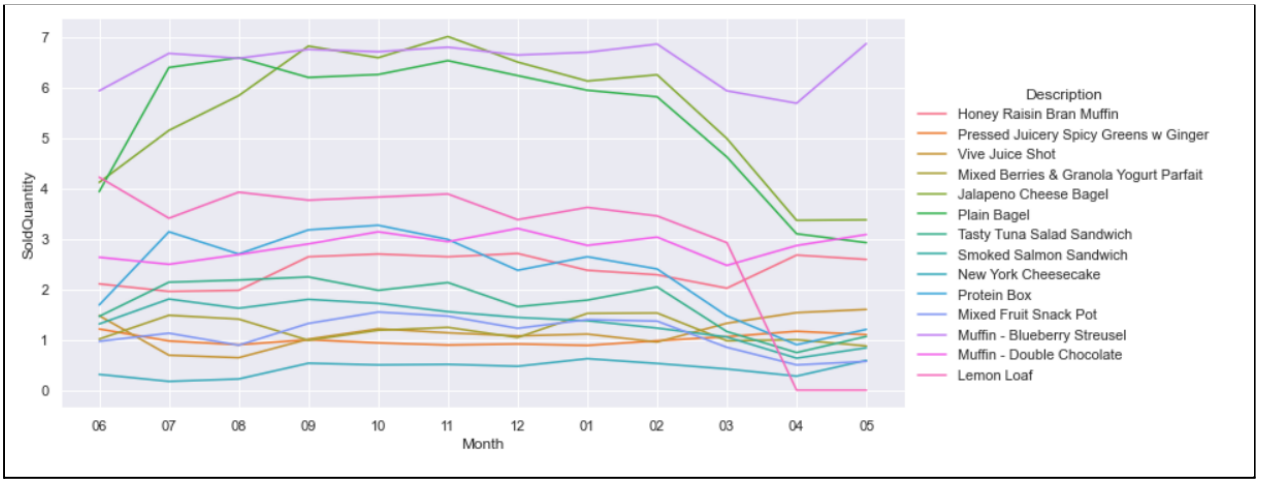
2) *Monthly Revenue for different category products:*

In the below graph we can see a high increase in Ready to Drink in March, April, and May because the season is Summer where everyone will be feeling like having more drinks.



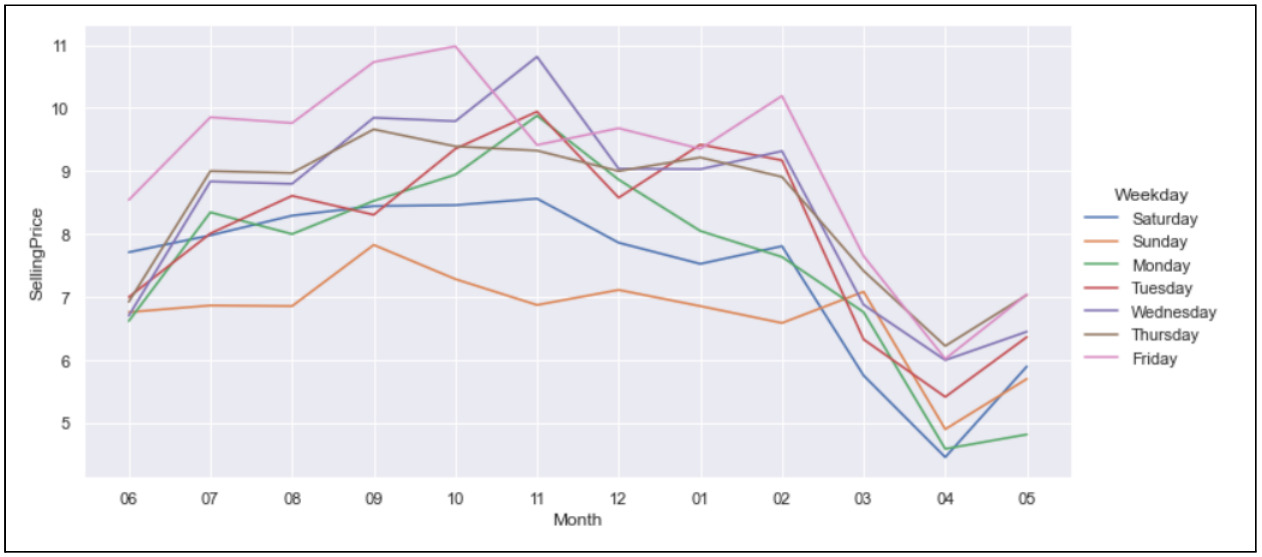
3) *Monthly Revenue for each product for three stores combined:*

We can see there is an increase in peak Smoked Salmon Sandwich in the months of June and July because it is the Commercial Salmon Season in California. The Council establishes management measures for commercial, tribal, and recreational salmon fisheries off the coasts of Washington, Oregon, and California. Tuna do have peak seasons. Peak tuna fishing seasons are from June to October. That's why we can see a significant increase in Tasty Tuna Salad Sandwiches.



4) *Monthly Revenue based on Day of the week:*

In all the months except november december and april the sales are high on friday and less on sunday . clearly we can say that sales are less on sunday when compared to other days . sales are more on weekdays when compared to weekends this may be due to office and schools . people preferred our stores to save their cooking time on their work day .

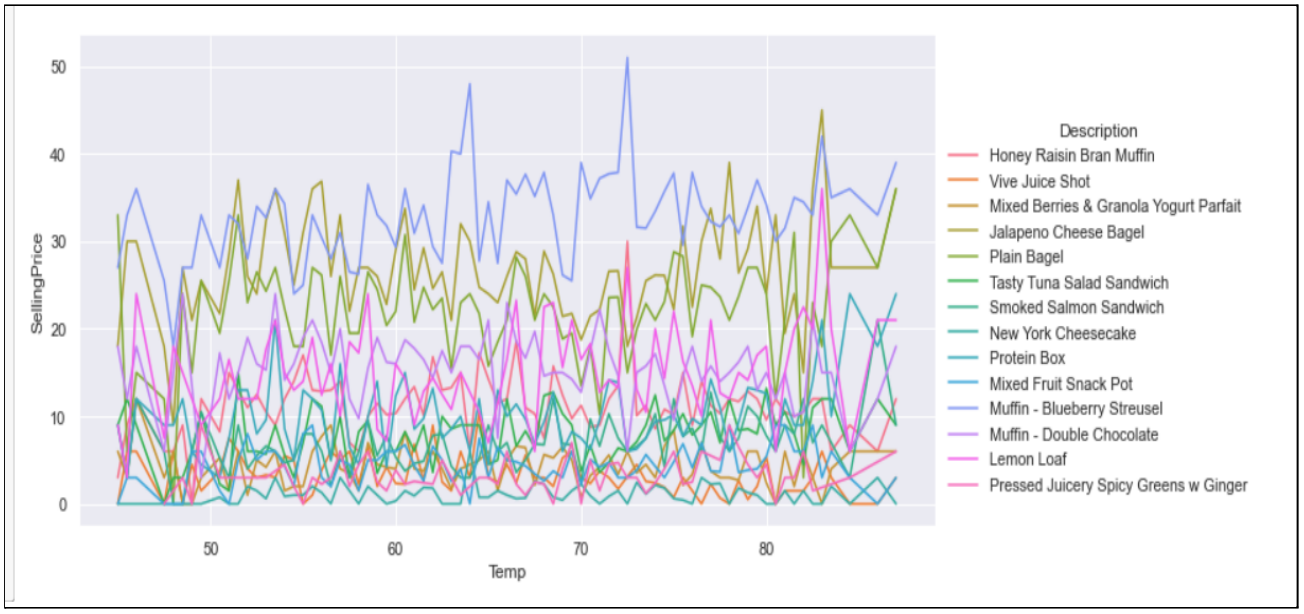


c) *Impact of weather on product sales:*

On the above Graph we can notice that Muffin Blueberry Streusel sales increase rapidly when the temperatures are peak. (Assumption). When the temperatures are higher we will be more active as well as thirsty. We tend to have

more sugar foods where our body produces a quick hit of energy. Especially Muffin Blueberry Streusel contains more sugars compared to other food items.

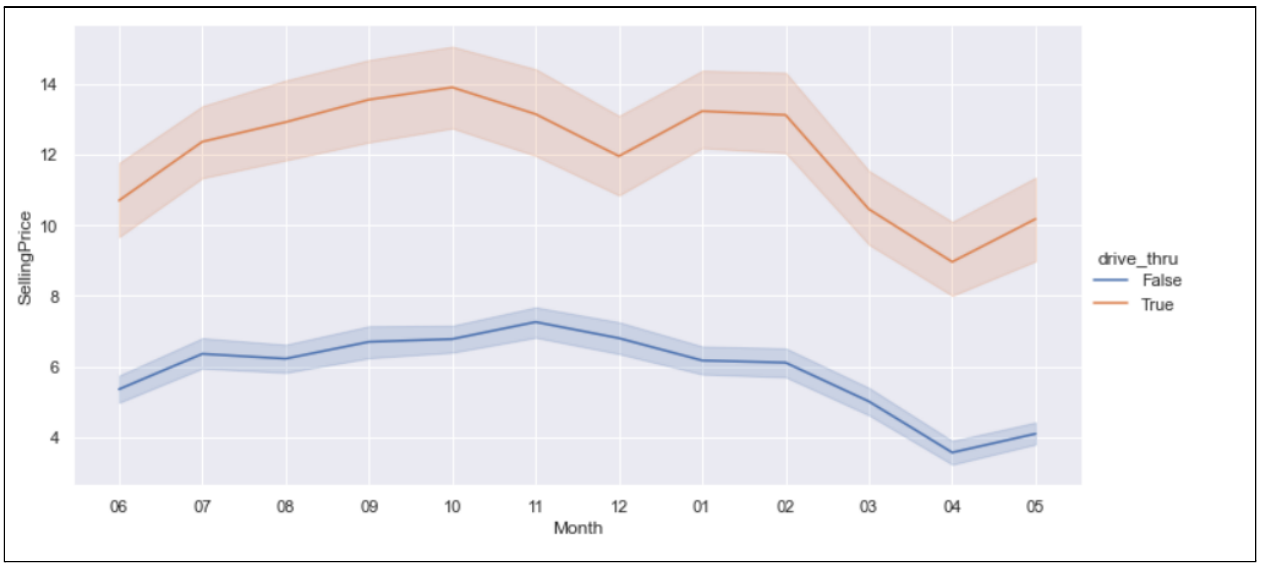
.



A-6: Investigate whether drive thru features cause certain products to sell better or worse.

a) *Revenue Graph showing difference between stores having drive thru and not:*

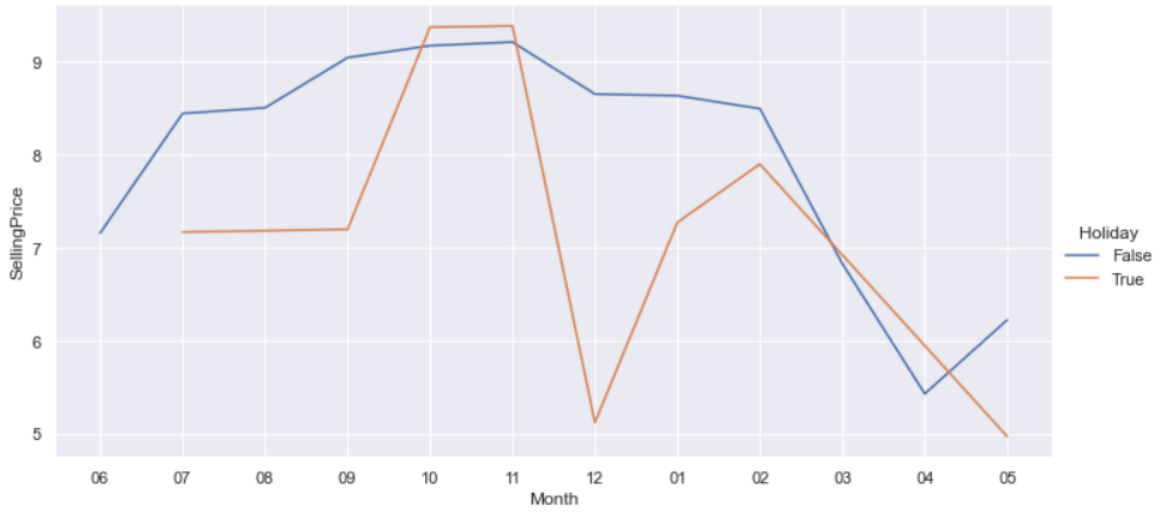
By the graph we can clearly say that people are much more interested in drive-through than dine. This may be due to many reasons , covid may be the main reason .so introducing the drive through option in other two stores would be profitable .



A-7: Investigate the impact of weekday/weekends and National Holidays by adding extra features.

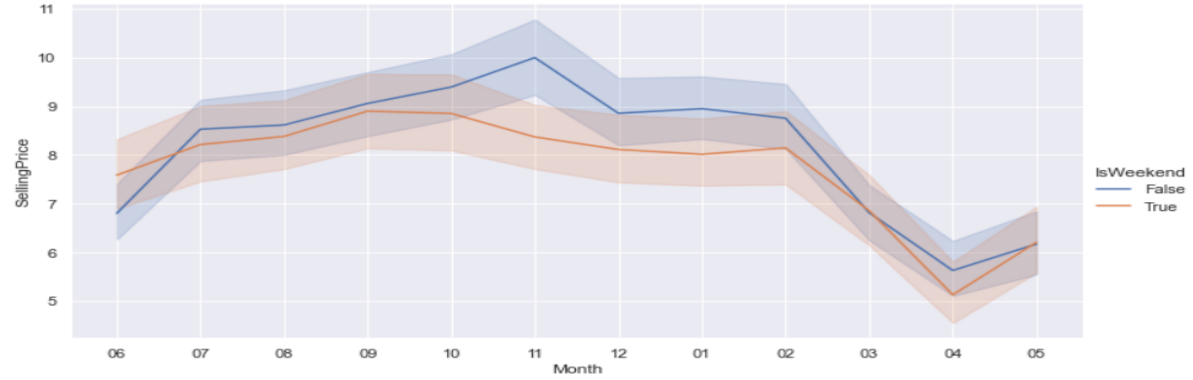
a) .

As per the graph holiday sales are high in november and october .the holiday sales record lowest in december and may ,as these months have only one federal holiday and stores would be closed during christmas The sales dropped from february to may this may be due to covid .



b) *Impact of Weekend on Monthly Sales:*

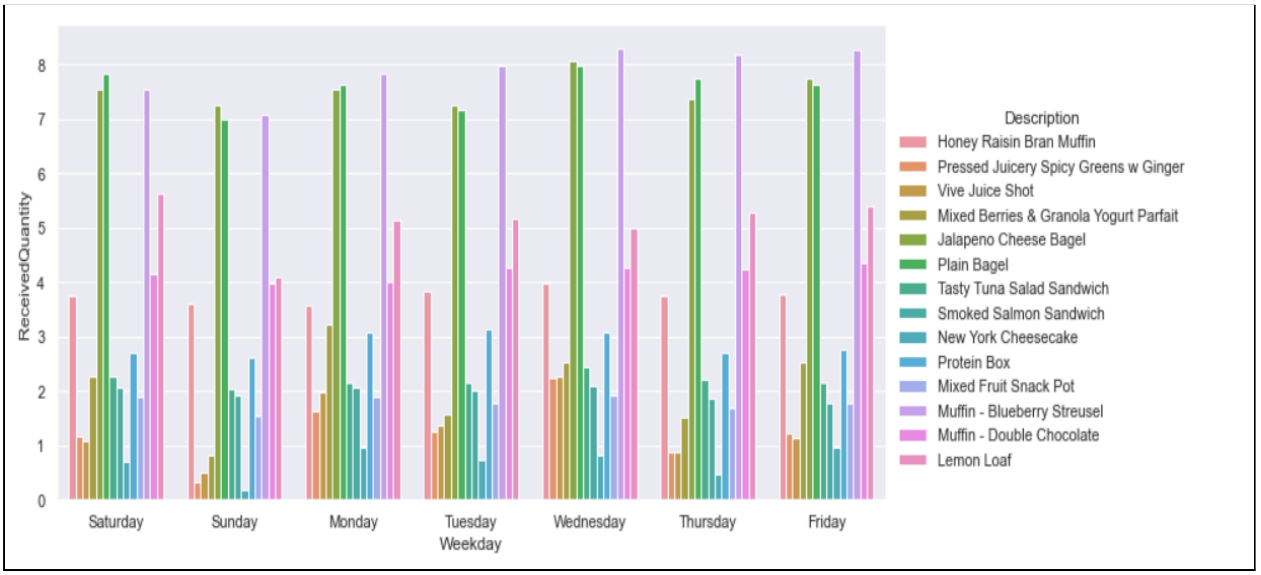
The below graph shows the monthly sales on weekdays vs weekends . As we can see in the month of June the sales on weekends are more than on weekdays . For all the remaining months weekday sales are more when compared to weekdays .and we can clearly see that in the month of march both the sales are equal .



A-8: Based on the store data, identify the stocking patterns across multiple stores.

a) *Received Quantity for each product based on weekday*:

By the below graph we can see that jalapeno cheese bagel , plain bagel and muffin blueberry Streusel are highly stocked on daily bases .pressed juicery spicy greens w ginger and vive juice shot are noticed with low received quantity . As we previously noticed that muffins are highly sold products, stoking up muffin-double chocolate also helps to increase the sales which indeed will be profitable .Lemon loaf remains constant all over the week. Sunday is the least stocking happening over the week. However, we can see the least sales happening on sunday.



A-9: Conclusions and recommendations to optimize the Stocking:

- Managing an optimized stocking for the best-sellers is essential to maintain and grow sales (that when consumers ask if the products are always available) while reducing waste. We notice that the top 25% of best-sellers are also the top missed sales and high waste. This is due to the fact the stores want to maintain high availability; however, the pattern of consumer demands is hard to predict.

- Sales are changing along with many factors, such as weekends, holidays, month-by-month, and weather. These factors can be carefully collected, then used through a predictive model to predict the quantity for the next week. When any factors (such as weather) change, re-input the features and predict again. Through the store manager's professional experience with an AI-driven prediction (multiple predictions) model(s), hopefully, we can optimize the inventory.

- Driving through is an important feature. We observe store 332 with a drive-through feature has two-fold to three-fold more sales than stores 18 and 117 without this feature. This feature is obvious for a coffee store to attract consumers for a cup of coffee or treat. We also notice store 332 has better inventory management overall. When a store has a drive-through, sales are more robust and stable, it is easier to predict the inventory to stock up.

Part B

Analyze the data of ALL Stores

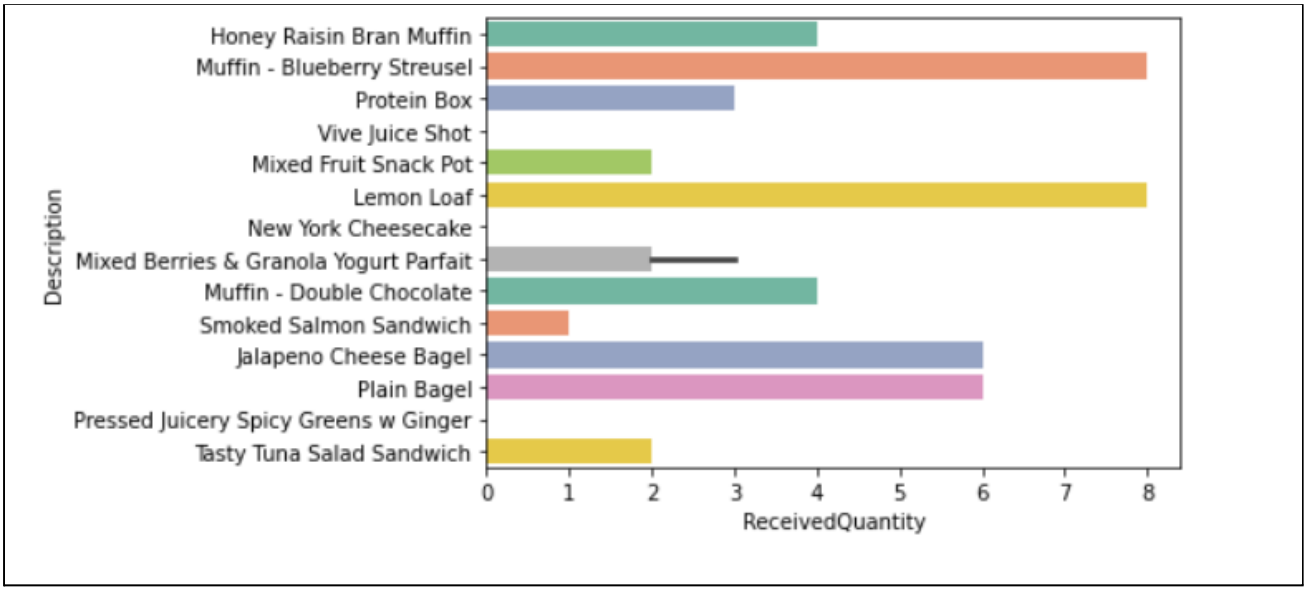
B-1: Provide the box plots and statistics of 14 products, inventory patterns, stock out patterns and missed sales for ALL Stores.

The per product bar chart shows the average daily received quantity per product at each store. The more inventory usually indicates the more popular the products.

a) Bar plot for inventory patterns (ReceivedQuantity) for store all stores for products:

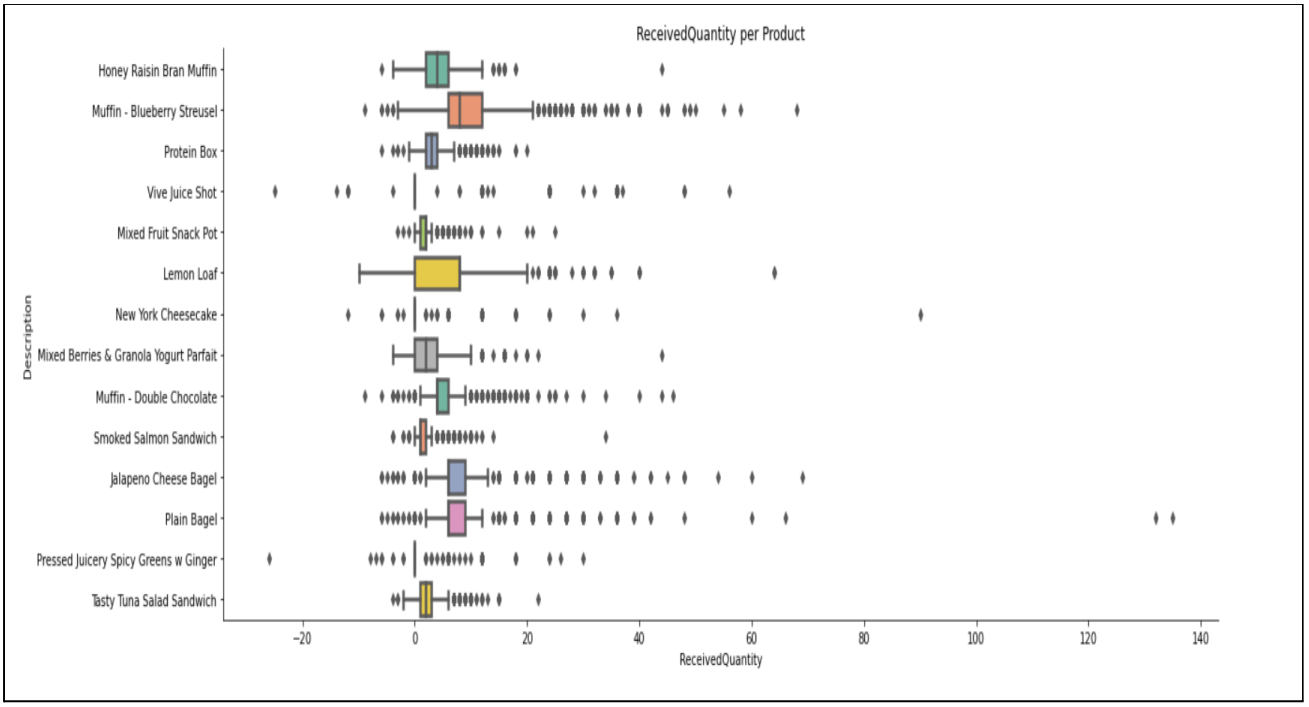
Below chart shows the median volume of ReceivedQuantity for each product across all stores. It reveals the most popular needed products they are:

1. Muffin-Blueberry Streusel
2. Lemon Loaf
3. Jalapeno Cheese Bagel
4. Plain Bagel.



b) *Box plot for inventory patterns (ReceivedQuantity) for store all stores for each products:*

The best sellers are consistently revealed from below box plot, which shows mean ReceivedQuantity and variance. These four products are the most popular products across all stores.



c) *<u>Box plot for stock out for all stores:</u>*

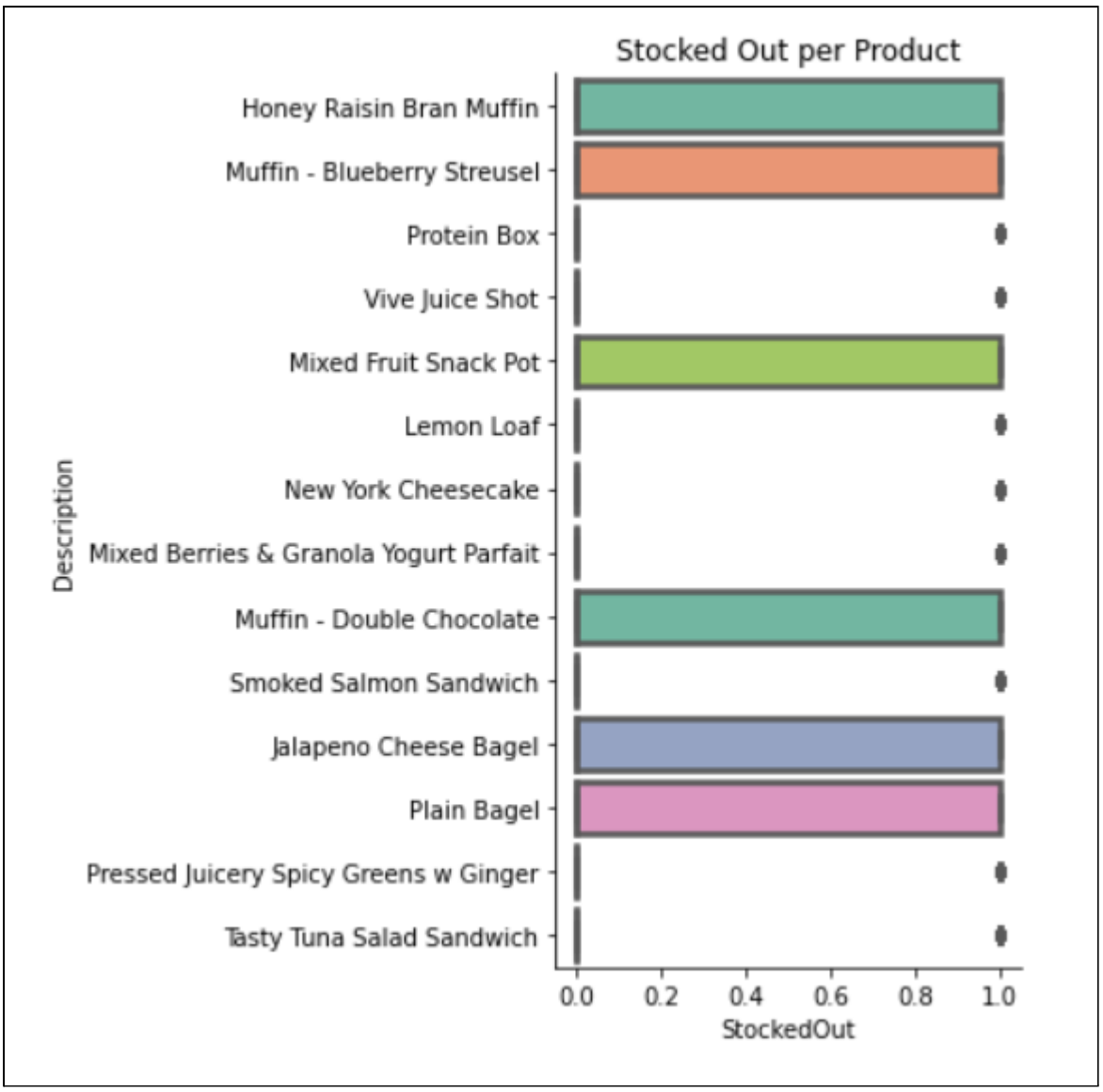
Now let's look at the bar plot and box Plot for stock out for all stores. From the data summary we can see the mean for StockOut is 0.228, the max value is 1. This data is not reliable due to the record method might be casual.

	StockedOut
count	486078.000000
mean	0.228360
std	0.419777
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

Even the StockedOut data might not be so reliable, Below plot shows the following products are often StockedOut but customers asked.

- Honey Raisin Bran Muffin
- Muffin-Blueberry Streusel
- Mixed Fruit Snack Pot
- Muffin-Double Chocolate
- Jalapeno Cheese Bagel
- Plain Bagel.

The reason for StckedOut could be due to inventory management gaps or products sold well. We will look into these data later to find out.

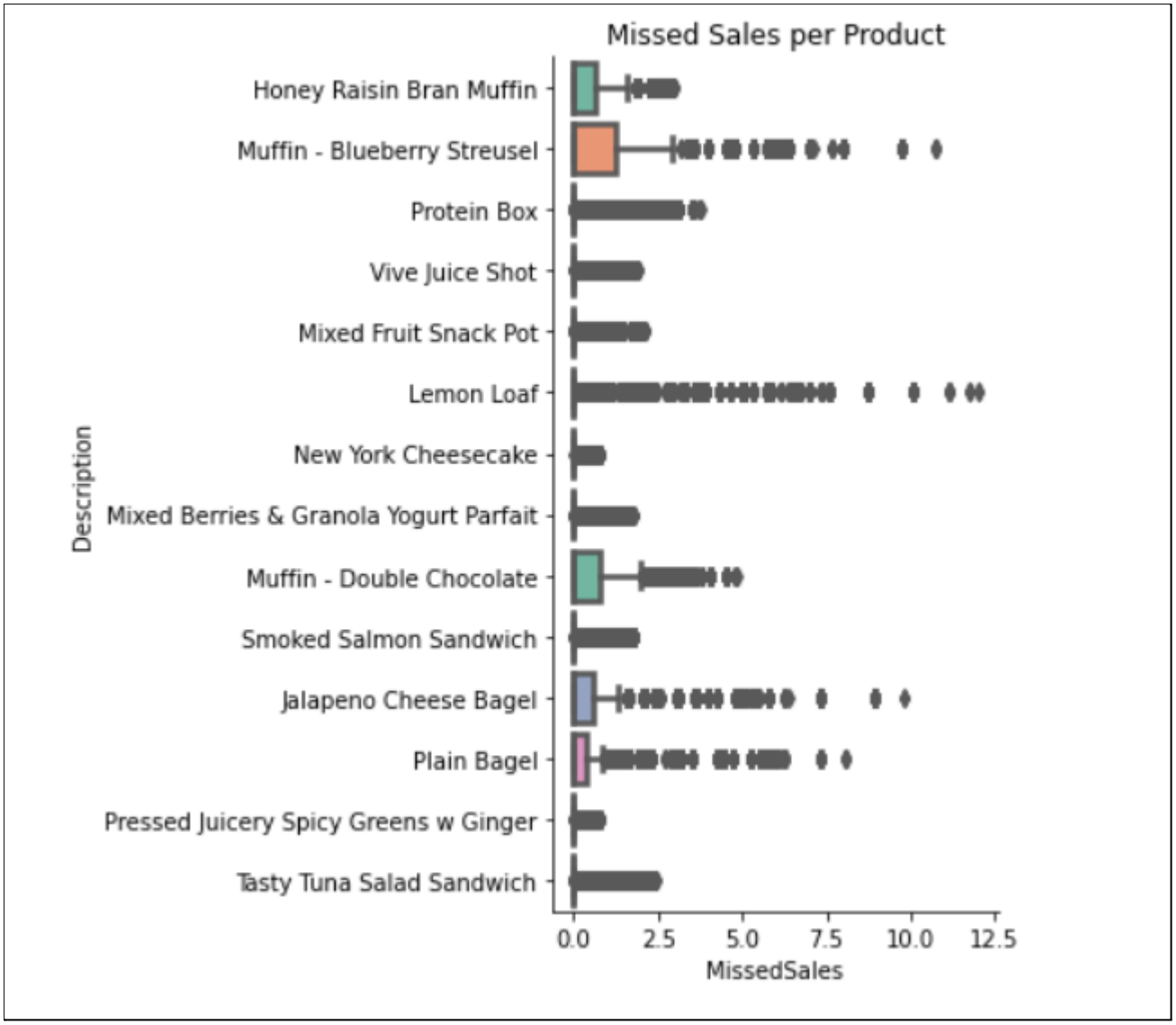

Stocked Out per Product
Honey Raisin Bran Muffin
Muffin - Blueberry Streusel
Protein Box
Vive Juice Shot
Mixed Fruit Snack Pot
Lemon Loaf
New York Cheesecake
Mixed Berries & Granola Yogurt Parfait
Muffin - Double Chocolate
Smoked Salmon Sandwich
Jalapeno Cheese Bagel
Plain Bagel
Pressed Juicery Spicy Greens w Ginger
Tasty Tuna Salad Sandwich
Description
0.0
0.2
0.4
0.6
0.8
1.0
StockedOut

d) *Box Plot for missed sales for all stores:*

Box plot for Missed Sales. The most missed sales products are:

- Muffin-Blueberry Streusel
- Muffin-Double Chocolate
- Jalapeno Cheese Bagel
- Plain Bagel.

In the later part, we know that Jalapeno Cheese Bagel and Plain Bagel are the top sellers, so there exists space to improve inventory management.

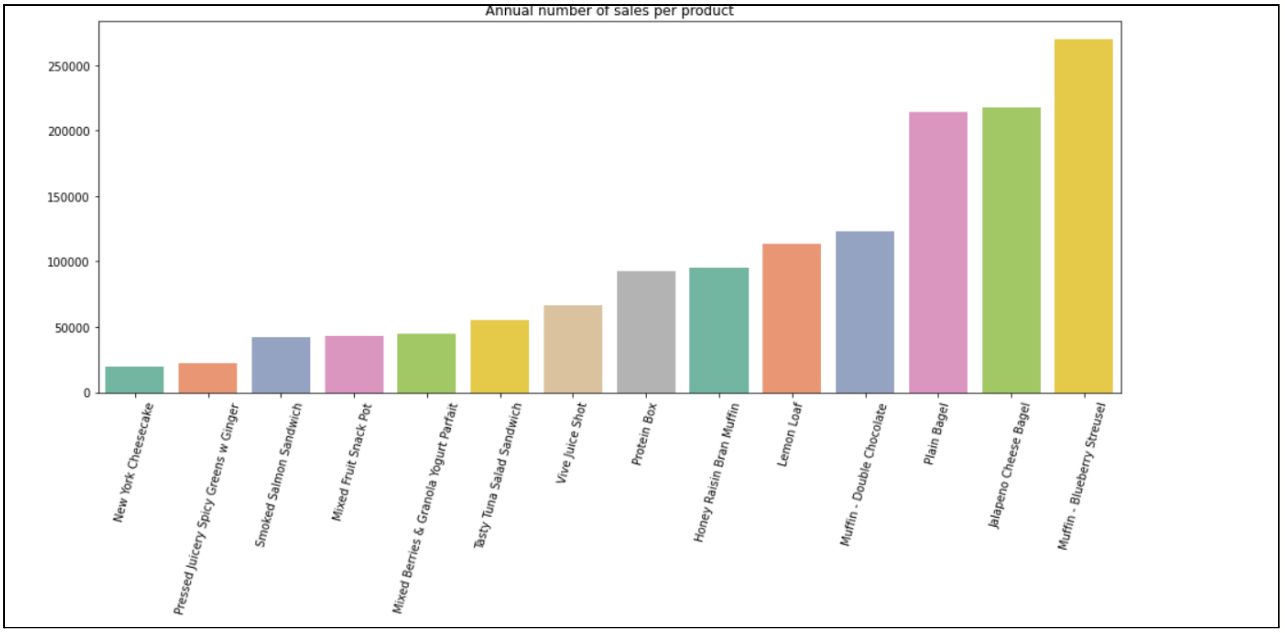


B-2: Show graphs of best seller and worst seller products of top 25% and bottom 25% and provide your insight for all stores data.

To determine the best and worst seller product, let's look at the total annual sales of each product first. It reflects the accumulated selling performance within the data reporting period. In further, we plot the line plot of each product according to the soldQuantity every day.

a) *Annual sold quantities for each product with the top 25% best product sales:*

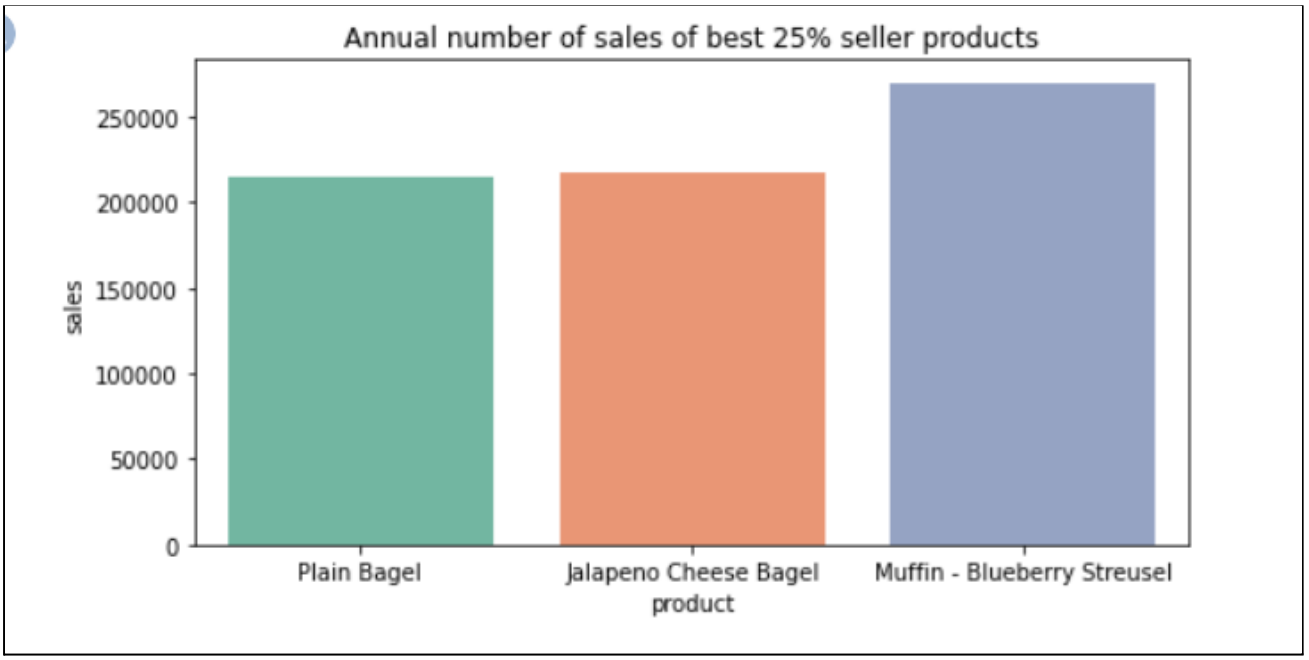
Bar plot for annual sales per product. From the bar chart, we can see the most sold and less sold products across all stores. We also understand from our individual store analysis that the best sellers (worst sellers) contain shared items and unique items. For example, plain bagel and Jalapeno Cheese Bagel are commonly shared best sellers. The third most popular product varies across different store locations.



b) *Plot top 25% product total annual sales:*

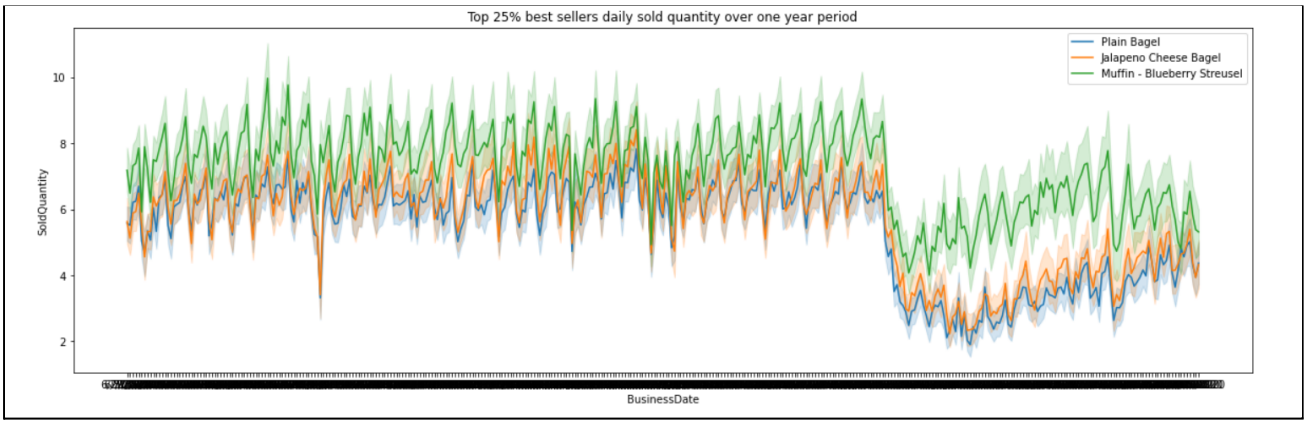
Insight of top 25% best sellers (see below plot):

- Muffin-Blueberry Streusel (green line) has the most sold quantity on a daily basis, followed by Jalapeno Cheese Bagel and Plain Bagel.
- Sales quantities vary periodically. Which could be due to the day of a week, or weather.
- Across all stores, from the chart below, we can see for most months, the coffee stores have stable and good sales, but there is some months that are cold seasons that sales are dipping.



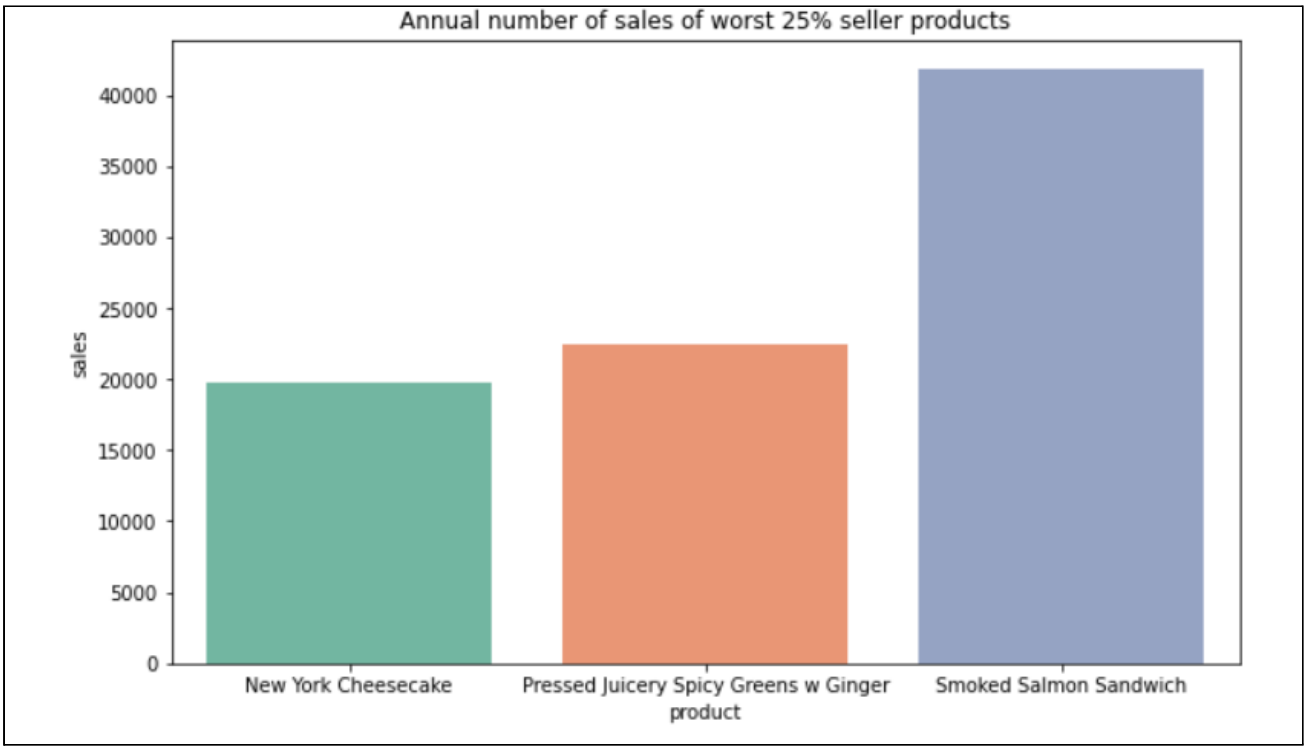
c) *Top 25% best sellers daily sales:*

Below graph shows daily sold quantities for the top 25% best sellers. Sales quantities vary in a pattern, which could be due to the day of a week, or weather.Sales have seasonal changes. There is a dip and sudden increase for some stores.There are several peak sales days that could be due to holiday.



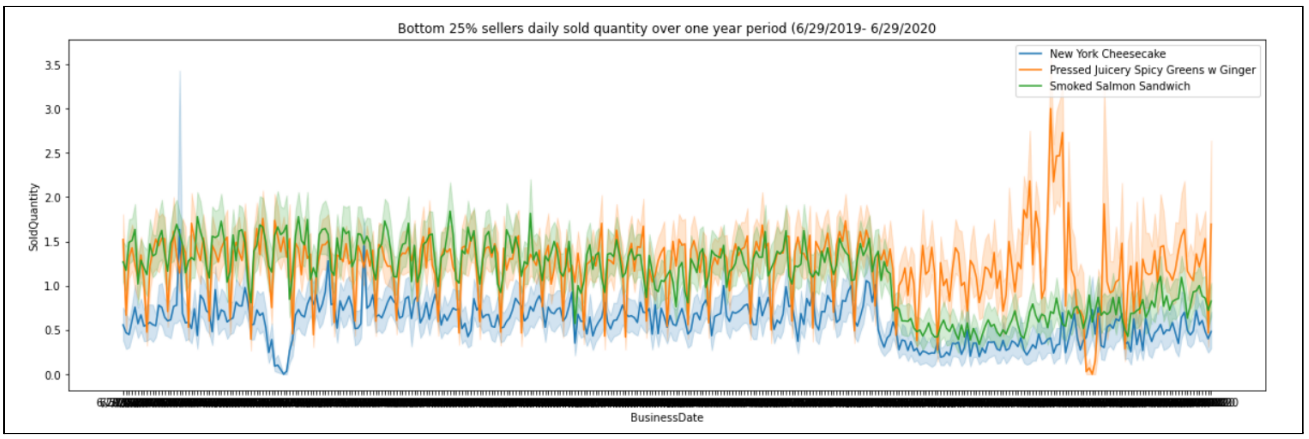
d) *Plot bottom 25% product total annual sales:*

This graph is plotted for the worst 25%seller products .New york cheesecake, pressed juicery spicy greens and ginger and smoked salmon sandwich are the worst 25% seller products .On comparisons new york cheesecake has the least sales and smoked salmon sandwich has highest sales .



e) *Top 25% best sellers daily sales:*

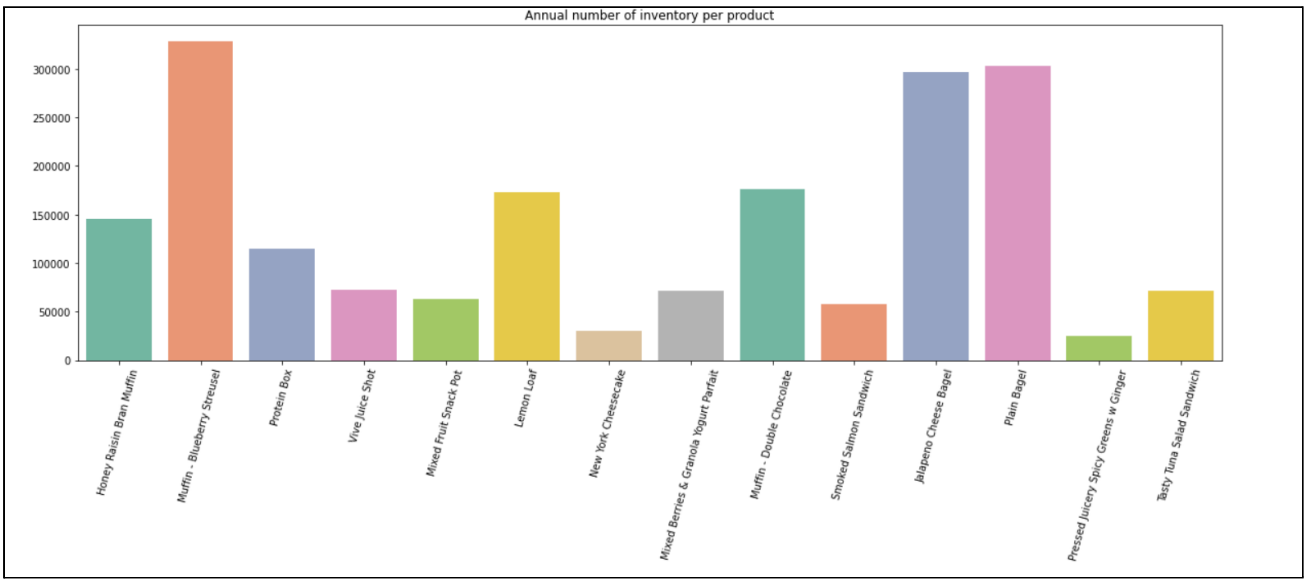
Plot the worst sold products over the year. From our previous analysis, sales overall dips in April. The orange line peaks on the right side shows the pressed juicery spicy greens w Ginger is in demand in the time of May and June 2020. Then move back to normal. It could be because of less customer demand or lack of inventory.



B-3: Show graphs of best and worst products based on their inventory management - Top 25% and bottom 25% and provide your insight into data.

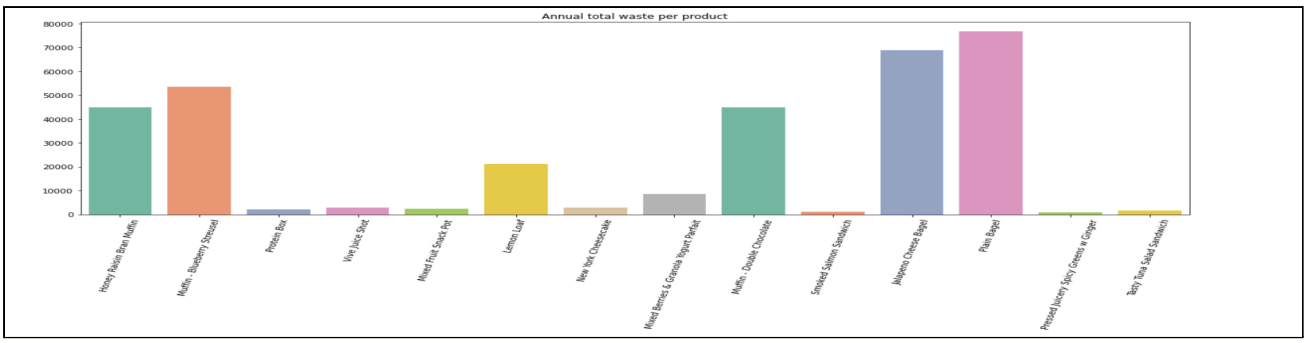
a) *Bar plot for annual inventory (Received Quantity) for all products:*

The below graph shows the best and worst product based on the product inventory management . Muffin-blueberry Streusel is the best product and New York Cheesecake ,pressed juicery, spicy greens and ginger are the worst products based on their inventory management. Moreover, plain bagel and jalapeno cheese bagel are also the best products .



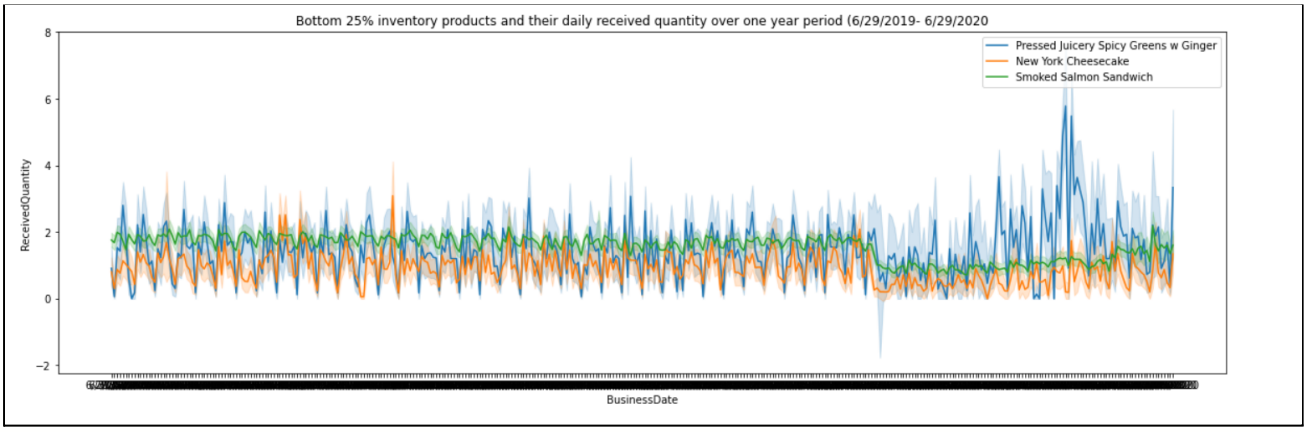
b) *Bar plot for annual waste for all products:*

The below plot shows the annual waste of the products . by the graph we can clearly see that plain bagel has the highest annual waste per year and jalapeno cheese bagel has second highest annual waste per year .Out of 14 products only 6 products wastage sums up to 90 per of total wastage.



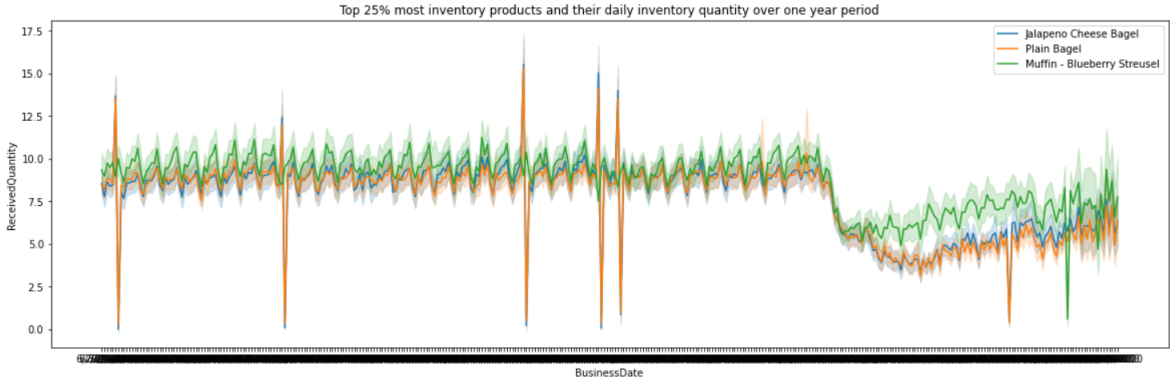
c) *25% Worst Inventory Management:*

- For the low inventory products, the inventory pattern is generally stable over the one year period of time.
- As we analyzed previously, Pressed juicery spicy Greens w Ginger is in demand during May and June 2020, so accordingly we see the inventory is increased in the below plot. We do not know if the product becomes a popular product after the time period of the dataset, but it could be a new product that consumers like.



d) *25% Best Inventory Management:*

- It looks like all stores have the pattern that some days in the one year period will have more/low demands so they store more/low inventories for these three products.
- There are impacts from some special days that have dramatically high or low inventory.From the plot, it looks like the holidays have strong impact to the quantity of holding inventory.

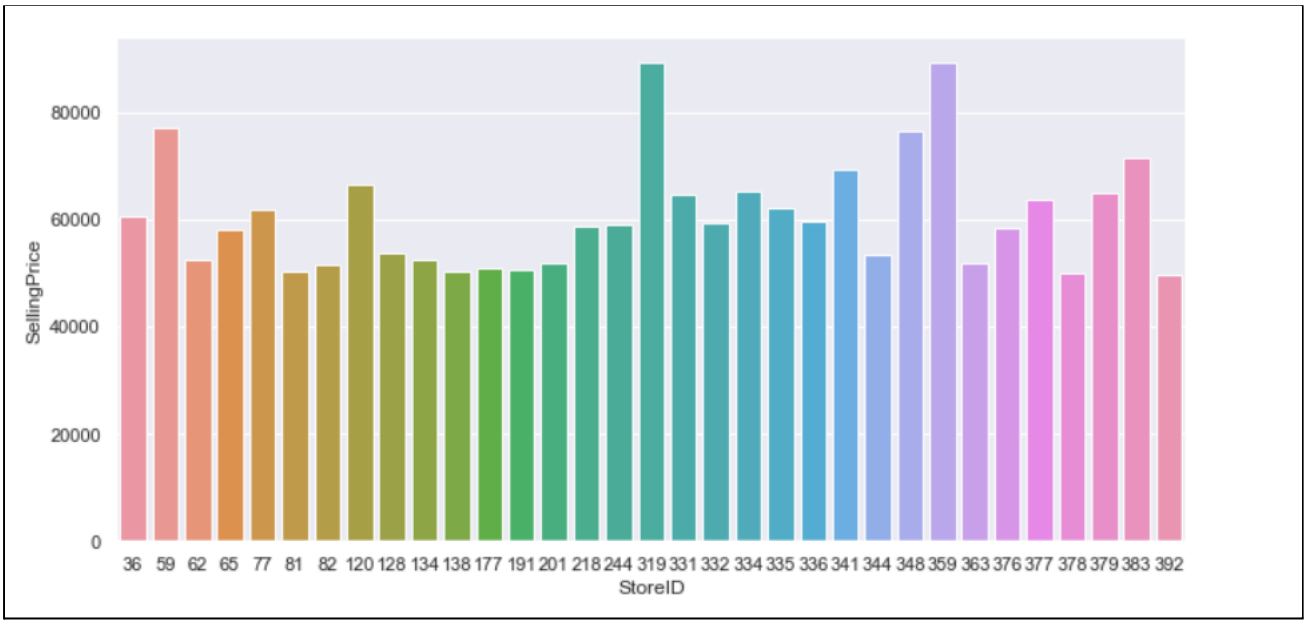


B-4: Identify best and worst stores based on top 25% and bottom 25%:

Here we have calculated best and worst store performance based on revenue generated for the given data (one year).

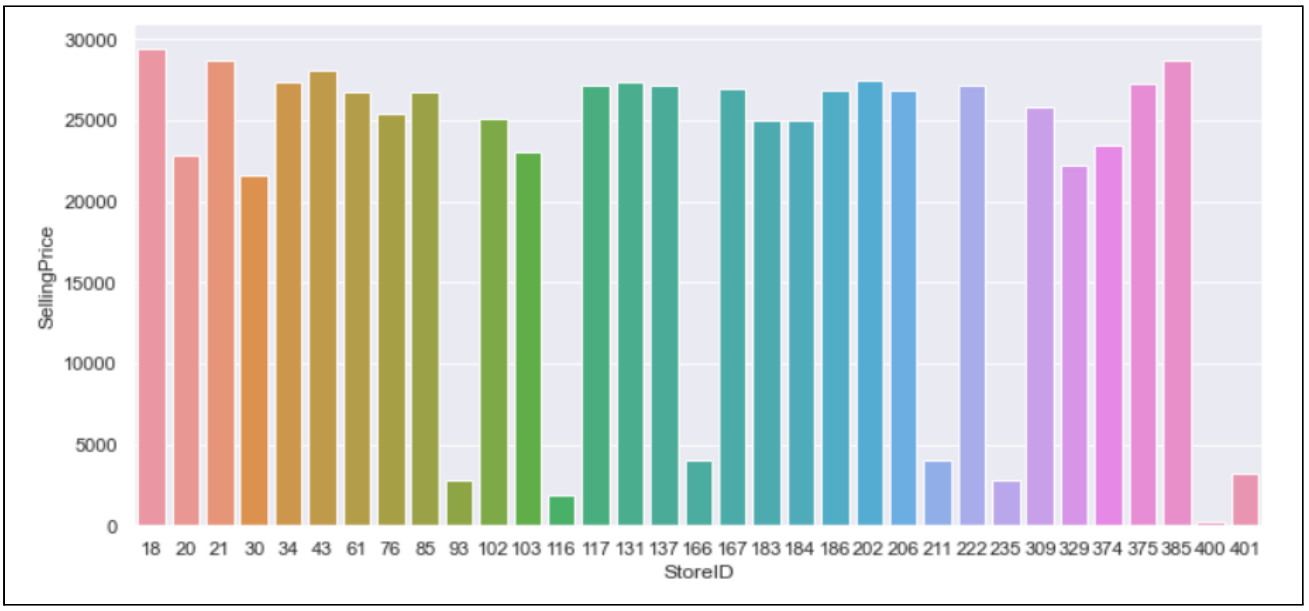
a) *Bar plot showing best stores based on revenue:*

The top store with average revenue more than 80,000 is 319. Only three or four stores stand out when compared to others. There might be a drive_thru option along with other amenities like a good place(area) of store and others.



b) *Bar plot showing worst stores based on Revenue:*

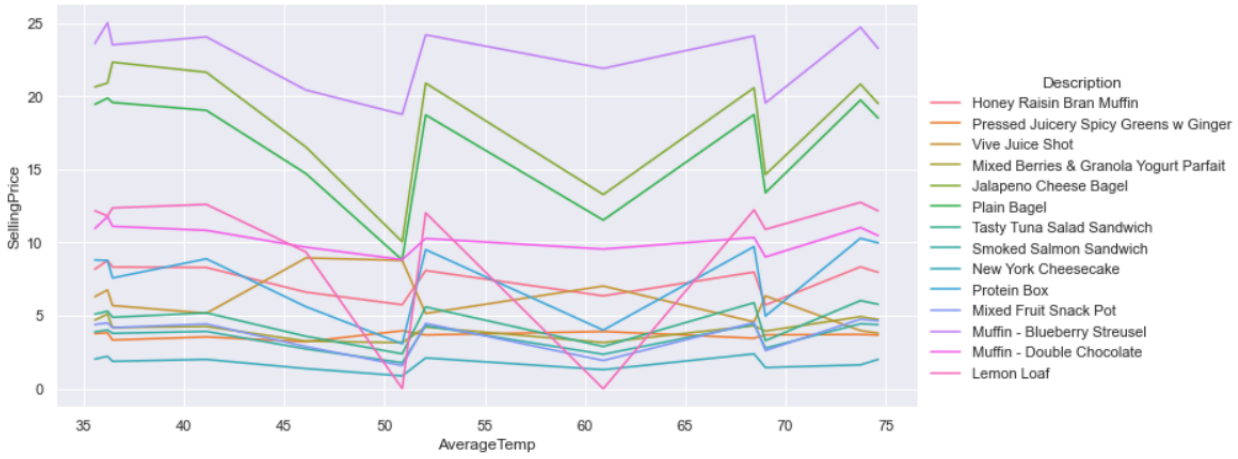
The least store having almost zero sales is 400 and 116. Revenue of the store might depend on many factors like drive thru, area in which store is, how proximity it is to the public and others.



B-5: Seasonal changes based on the Average temperature of US:

a) *Line Graph on Average temp of USA for each month vs Revenue:*

The below graph shows the changing in sales based on seasonal changes . The selling price of all products decreased between 45 degrees to 52 degrees . There is a drastic change in the selling price of lemon loaf between 45 - 52 degrees and 55-62degrees .The products muffin-blueberry streusel ,mixed berries and granola yogurt parfait ,plain bagel and lemon loaf has the same trend (has the same sales according to the temperature changes).



Conclusion for next Steps:

After observing the data patterns and type of data we have. We can decide to continue with Time series analysis for the inventory management and optimization for this coffee store data.

a) *Periodic Review for the data:*
   - With a periodic review system, the inventory is checked and reordering is done only at specified points in time. For example, inventory may be checked and orders placed on a weekly, biweekly, monthly, or some other periodic basis.
   - When a firm or business handles multiple products, the periodic review system offers the advantage of requiring that orders for several items be placed at the same preset periodic review time. With this type of inventory system, the shipping and receiving of orders for multiple products are easily coordinated.

b) *Continuous Review*
   - In the multi period model, the inventory system operates continuously with many repeating periods or cycles; inventory can be carried from one period to the next. Whenever the inventory position reaches the reorder point, an order for Q units is placed. Because demand is probabilistic, the time the reorder point will be reached, the time between orders, and the time the order of Q units will arrive in inventory cannot be determined in advance

References:

1. http://scacis.rcc-acis.org/
2. https://www.statista.com/statistics/513628/monthly-average-temperature-in-the-us-fahrenheit/
3. https://towardsdatascience.com/inventory-management-using-python-17cb7ddf9314
4. Anderson, Sweeney, Williams, Camm, Cochran, Fry, Ohlmann. An Introduction to Management Science: Quantitative approaches to Decision Making. 14th Edition, 2015. Cengage Learning. pp. 457–478

Section 2: Prediction

Yashpaul V	Data Preprocessing and LSTM, Informer
Priyanka P	Data Pre processing, Documentation
Madhu Kiran	Ensemble model analysis
Deepak	Ensemble model analysis

Introduction:

In the previous section we have analyzed the data using different graphs and showcased how each feature is important and effective in sales for the given dataset. Using that analysis and understanding of the data we have moved forward in implementing different machine learning algorithms and tried to fine tune the models in order to increase the accuracy of the model.

Produce synthetic data using Generative models (GANs) to get accurate predictions even with little historical data where necessary.

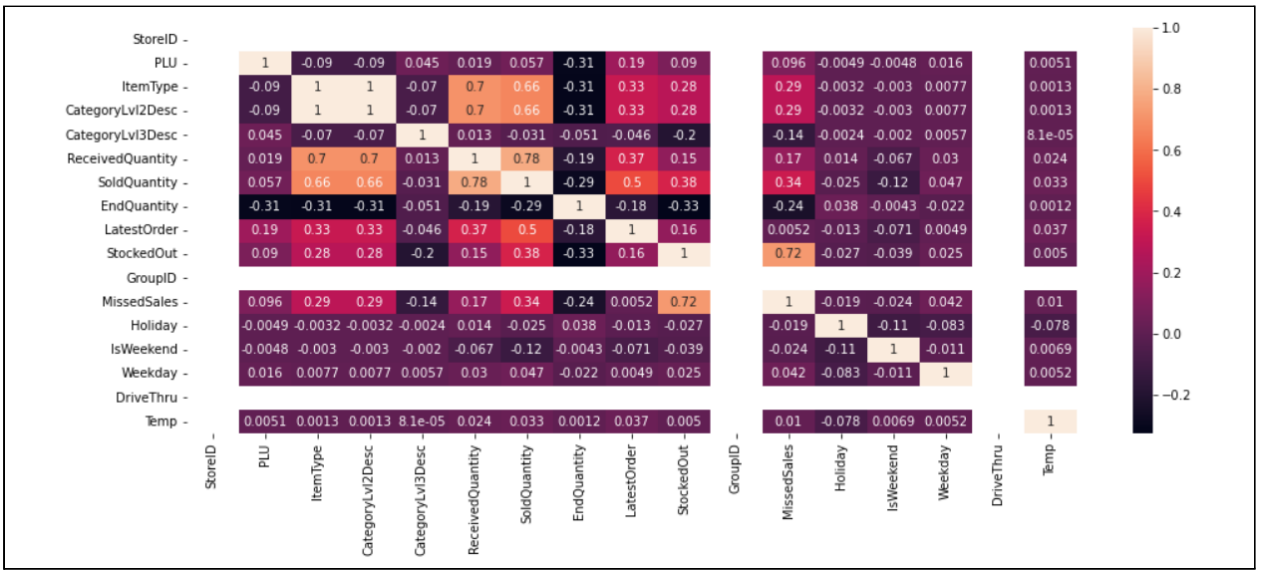
We have produced GANs whenever necessary in the downline of this implementation part. For the purpose of GANs we have used ctgans library, CTGANSynthesizer module which are open source for using. The creation of synthetic data stems from real data, and a good synthetic dataset is able to capture the underlying structure and display the same statistical distributions as the original data, rendering it indistinguishable from the real one.The first major benefit of synthetic data is its ability to support machine learning/deep learning model development.

```
In [6]: from ctgan import CTGANSynthesizer
        discrete_columns = ['BusinessDate','Description','ItemType','CategoryLvl1Desc','CategoryLvl2Desc',
                            'CategoryLvl3Desc','Holiday','IsWeekend','Weekday','Year', 'Month','Date']
        ctgan = CTGANSynthesizer(epochs=10)
        ctgan.fit(dataset, discrete_columns)
        #generate synthetic data, 1000 rows of data
        synthetic_data = ctgan.sample(1000)
        #print(synthetic_data.head(5))
```

Iterate through different combinations of features to identify the optimal features and remove potential correlated features (if any) for your predictions. Add weather, weekdays, holidays and temperature data to your features.

We have added a weather data feature with US average weather taken from some website and converted it into a dataframe and added to the dataset. Federal Holidays list can be generated from python us federal calendar and added to the dataset. Remaining features are derived from the BusinessDate feature.

Initially we have created correlation maps between all the available features and studied them.



We can observe that there are mixed values of both positive and negative. Using this correlation map we have decided to go with a few columns. Before generating this correlation map we have converted all the string and categorical values into numericals which we thought isgood procedure in order to generate a valid correlation map.

After checking this map we have decided to eliminate a few columns and go with the selected column to get good accuracy for the machine learning models.

Down the line of the project we have modified columns based on the correlation map to get more accuracy. The above map is at the initial point.

Initial selected features are :

['BusinessDate','PLU','CategoryLvl3Desc','ReceivedQuantity','SoldQuantity','EndQuantity','StockedOut','Temp','Holiday','Weekday']

Start with a quick linear regression to get a sense of data. Linear regression may not result in a great prediction.

In linear regression R-Squared value is one of the goodness-of-fit measures . R-Square value determines the strength of the relationship between the model and the dependent variable .

Here R-Square value is 0.708. This shows that only 70% of the points lie on the line of fit .This value is not good for a regression model to be considered reliable.

OLS Regression Results

Dep. Variable:	SoldQuantity	**R-squared:**	0.708
Model:	OLS	**Adj. R-squared:**	0.707
Method:	Least Squares	**F-statistic:**	1868.
Date:	Tue, 26 Apr 2022	**Prob (F-statistic):**	0.00
Time:	12:23:47	**Log-Likelihood:**	-21847.
No. Observations:	11596	**AIC:**	4.373e+04
Df Residuals:	11580	**BIC:**	4.384e+04
Df Model:	15		
Covariance Type:	nonrobust		

Use ensemble models. Develop the following models and compare the accuracies by comparing the confusion matrix, true positive rate, true negative rate, precision, and F1-score. -

- Random Forest

  Random forests is a learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by

most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set.

- Gradient Boosting Machine

  Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction in the form of weak prediction models, which are typically decision trees.When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function.

- Light GBM

  Light Gradient Boosting Machine, is based on decision tree algorithms and used for ranking, classification and other machine learning tasks. The development focus is on performance and scalability.

- XGBoost

  XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting that solves many data science problems in a fast and accurate way. The same code runs on major distributed environments (Hadoop, SGE, MPI) and can solve problems beyond billions of examples.

  As we have time series data, to predict the sales based on the weather, weekday, weekend, etc. This is clearly a regression problem. The metrics like true positive, true negative rate, precision and f1 score are used for classification problems. The metrics used for regression problems are Mean square error, Mean absolute error and R square. Below are the comparison of the algorithms

	model	best_score	best_params
0	random_forest	0.764695	{'max_depth': 10, 'n_estimators': 300}
1	gbm_regression	0.760905	{'learning_rate': 0.2, 'max_depth': 5, 'n_esti...
2	lgb	0.772509	{'learning_rate': 0.01, 'max_depth': 20, 'n_es...
3	XGB_regression	0.722736	{'learning_rate': 0.005, 'max_depth': 2, 'n_es...
4	XGBRF_regression	0.421075	{'learning_rate': 0.5, 'max_depth': 10, 'n_est...

From the above we can clearly see the Light gradient boosting machine(lgb) shows the best score among all the other algorithms. We have used the learning rate 0.001, max_depth 20 and n_estimators 2800 to get the best results. After changing the parameters max_depth 12 and n_estimators 1440 the modal shows a slight improvement. Below is the screenshot for the after changing the parameters.

	model	best_score	best_params
0	random_forest	0.763881	{'max_depth': 10}
1	gbm_regression	0.767107	{'learning_rate': 0.2, 'max_depth': 5, 'n_esti...
2	lgb	0.776624	{'learning_rate': 0.01, 'max_depth': 20, 'n_es...
3	XGB_regression	0.752898	{'learning_rate': 0.1, 'max_depth': 2, 'n_esti...
4	XGBRF_regression	0.421075	{'learning_rate': 0.5, 'max_depth': 10, 'n_est...

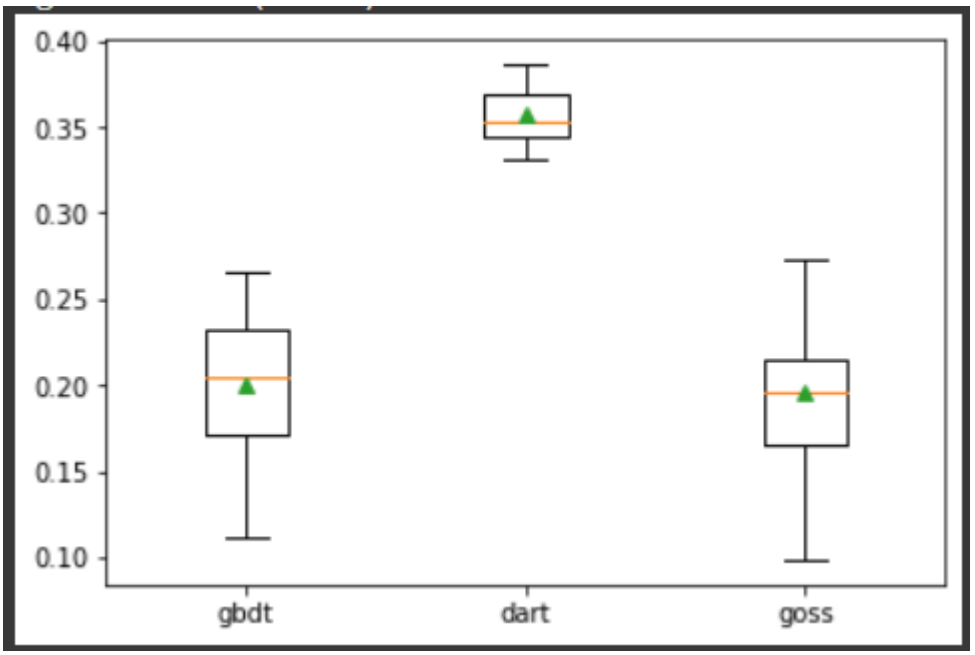
A feature of LightGBM is that it supports a number of different boosting algorithms, referred to as boosting types.

The boosting type can be specified via the “boosting_type” argument and take a string to specify the type. The options include:

- ‘gbdt‘: Gradient Boosting Decision Tree (GDBT).
- ‘dart‘: Dropouts meet Multiple Additive Regression Trees (DART).
- ‘goss‘: Gradient-based One-Side Sampling (GOSS).

The default is GDBT, which is the classical gradient boosting algorithm.

Below are the results when using the different boosting algorithms




```
>gbdt 0.200 (0.042)
>dart 0.357 (0.016)
>goss 0.195 (0.040)
```

From the above results we can see the default boosting method performed better than the other two techniques that were evaluated.

As we are dealing with time-series data, we would like to compare the results of the previous models with the following deep methods:

- LSTM
- Informe**r**

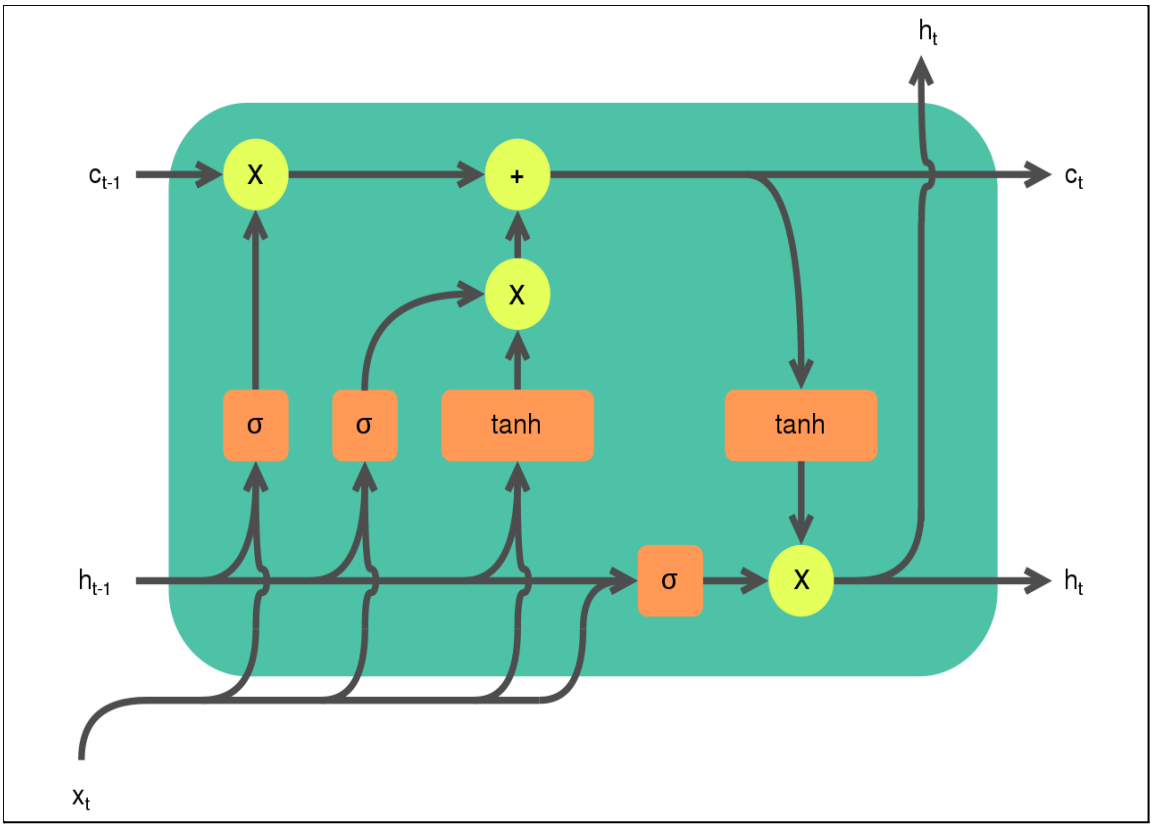
Time series forecasting is one of the major building blocks of Machine Learning. There are many methods in the literature to achieve this like Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving-Average (SARIMA), Vector Autoregression (VAR), and so on.

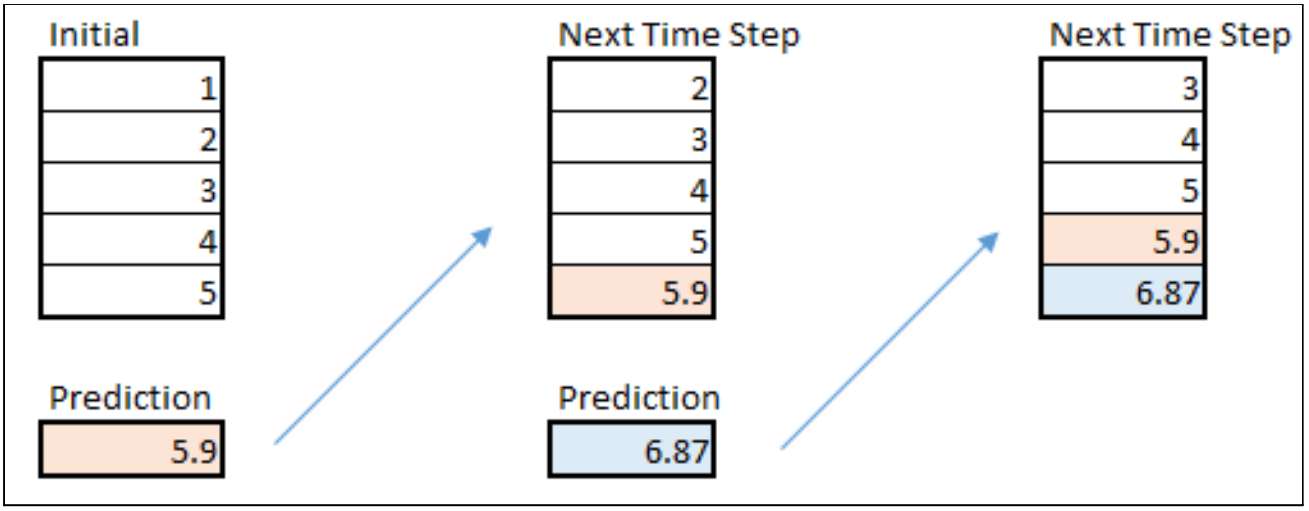
LSTM methodology, while introduced in the late 90's, has only recently become a viable and powerful forecasting technique. Classical forecasting methods like ARIMA and HWES are still popular and powerful but they lack the overall generalizability that memory-based models like LSTM offer.

LSTMs help in solving exploding and vanishing gradient problems. In simple terms, these problems are a result of repeated weight adjustments as a neural network trains.

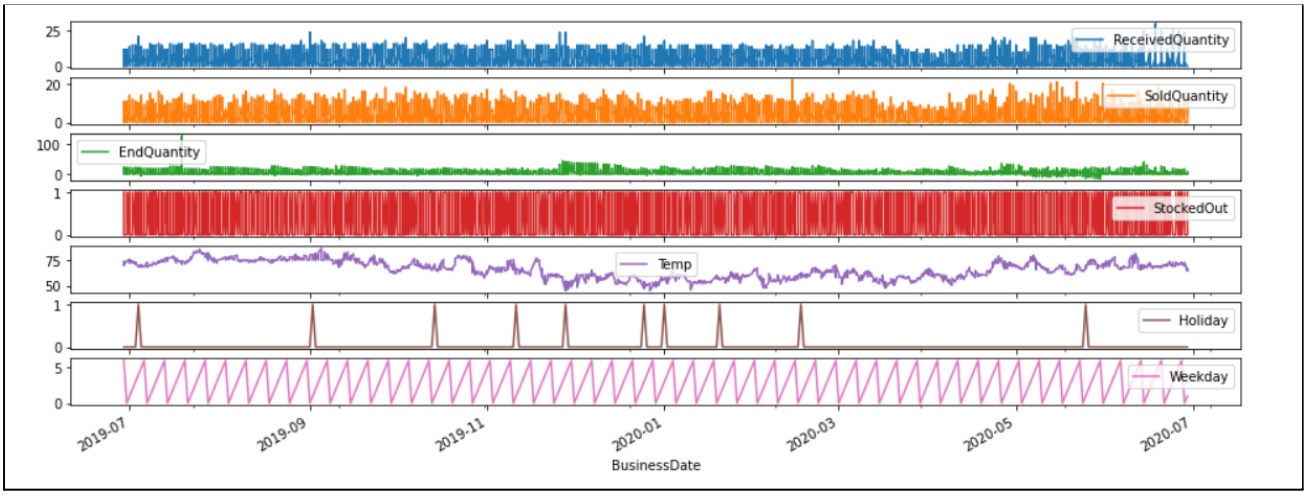
With repeated epochs, gradients become larger or smaller, and with each adjustment, it becomes easier for the network's gradients to compound in either direction. This compounding either makes the gradients way too large or way too small. While exploding and vanishing gradients are huge downsides of using traditional RNN's, LSTM architecture severely mitigates these issues.

After a prediction is made, it is fed back into the model to predict the next value in the sequence. With each prediction, some error is introduced into the model. To avoid exploding gradients, values are 'squashed' via (typically) sigmoid & tanh activation functions prior to gate entrance & output. Below is a diagram of LSTM architecture:





How ever we have created a graph for BusinessDate on X axis align with multiple column value for y axis



We can observe that the pattern of received and sold Quantity are almost same and from this we can infer that soldQuantity which is sales of all products are based on the number of items received.
Holidays and weekends have minimal effect on the overall sales as per the graph shown above.
We have generated the similar kind of graphs for each product separately in order to analyze the trend in sales with effect on the features.

LSTM:

Initially we have scaled the data that we have using the **MinMaxScalar** module found in Sklearn. After that we have generated a time series model using **TimeseriesGenerator** from kera library.As per the parameters required for LSTM, we have given a window size of 7 days for a batch of 30 days with 7 features. This tries to create a LSTM model for 7 upcoming days based on 30 previous dates available

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 lstm_6 (LSTM)                (None, 100, 128)          67072

 leaky_re_lu_4 (LeakyReLU)    (None, 100, 128)          0

 lstm_7 (LSTM)                (None, 100, 128)          131584

 leaky_re_lu_5 (LeakyReLU)    (None, 100, 128)          0

 dropout_4 (Dropout)          (None, 100, 128)          0

 lstm_8 (LSTM)                (None, 64)                49408

 dropout_5 (Dropout)          (None, 64)                0

 dense_2 (Dense)              (None, 1)                 65

=================================================================
Total params: 248,129
Trainable params: 248,129
Non-trainable params: 0
_________________________________________________________________
```

Sales Prediction:

Using LSTM here are the sales forecast accuracy for :

1 day from today :

```
ean_absolute_error: 0.2549

model.evaluate_generator(test_generator, verbose=0)

[0.09795643389225006, 0.25488731265068054]
```

3 days from today :

```
ean_absolute_error: 0.0708

model.evaluate_generator(test_generator, verbose=0)

[0.02678406983613968, 0.07079347968101501]
```

10 days today:

```
ean_absolute_error: 0.0575

model.evaluate_generator(test_generator, verbose=0)

[0.023589937016367912, 0.05753243342041969]
```

The below code is for sales prediction using LSTM :
Where :

win length is number of feature days to predict
Batch size is number of records taken at a time
Num features is number of features in data set

- we generated time series for testing and training
- Early_stopping → We are stopping the training process when value loss is minimum
- we are using mean absolute error as matrix for this LSTM model
- Using this model we have done the prediction
- Changing the win length value we have done predictions for 1 , 3 , and 10 days

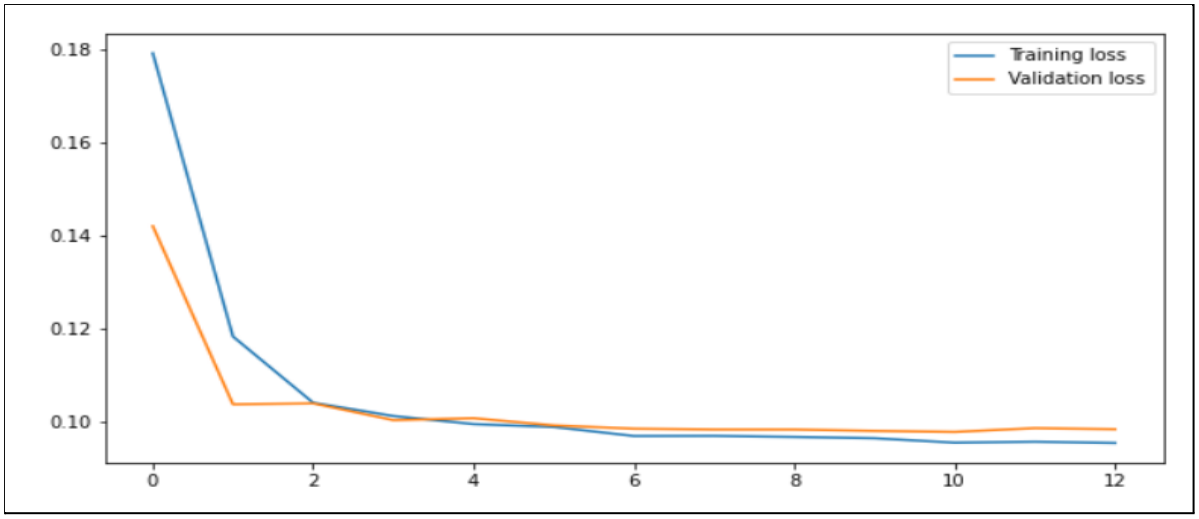
```
win_length = 1
batch_size = 30
num_features=9
train_generator = TimeseriesGenerator(x_train,y_train,length=win_length,sampling_rate=1,batch_size=batch_size)
test_generator = TimeseriesGenerator(x_test,y_test,length=win_length,sampling_rate=1,batch_size=batch_size)

early_stopping = tf.keras.callbacks.EarlyStopping(monitor='val_loss',patience=2,mode='min')
model.compile(loss=tf.losses.MeanSquaredError(),optimizer=tf.optimizers.Adam(),metrics=[tf.metrics.MeanAbsoluteError()])
history = model.fit_generator(train_generator, epochs=50,validation_data=test_generator,shuffle=False,callbacks=[early_stopping])

model.evaluate_generator(test_generator, verbose=0)

predictions=model.predict_generator(test_generator)
predictions.shape

df_pred=pd.concat([pd.DataFrame(predictions), pd.DataFrame(x_test[:,1:][win_length:])],axis=1)
df_pred
```

Predicted values for one day is shown in the below table :

	SoldQuantity	PLU	CategoryLvl3Desc	ReceivedQuantity	EndQuantity	StockedOut	Temp	Holiday	Weekday	Sold_Pred
11597	1.0	6228	1	0.0	8.0	0	61.0	0	0	1.859619
11598	2.0	12216	1	0.0	10.0	0	61.0	0	0	0.876825
11599	0.0	3000277	2	0.0	0.0	0	61.0	0	0	0.799602
11600	8.0	851004	3	12.0	0.0	0	61.0	0	0	4.848939
11601	9.0	810407	3	9.0	0.0	1	61.0	0	0	4.681669
...	...	...	...	...	...	...	...	...	...	...
14490	3.0	3000162	2	0.0	0.0	1	65.0	0	1	2.914027
14491	0.0	3000181	2	0.0	0.0	0	65.0	0	1	5.840012
14492	15.0	3000211	0	0.0	5.0	0	65.0	0	1	4.997616
14493	2.0	3000212	0	0.0	1.0	0	65.0	0	1	4.026688
14494	4.0	3000024	4	0.0	0.0	1	65.0	0	1	4.462466

2898 rows × 10 columns

Graph between training loss and validation loss:



Conclusion:

We found LSTM and LGBM to be best fit models for our data .Although this is for individual stores , we are working on all other stores and for individual store product level .

Section 3: Inventory optimization

Team Members' Roles:

Yashpaul V	Initiatives 1,2,3 and its coding part
Priyanka P	Documentation and PPT
Madhu Kiran	Explainer Dashboards and documentation
Deepak Goud	Explainer Dashboards

Initiative 1:

We have given a great thought process in increasing sales and minimizing end quantity, which directly reduces missed sales. As we are thinking through it we have got an idea here.

We have seen data of one store, one product within one month and drew few insights and from that we understood that almost everyday we are receiving products for the entire month irrespective of weekends and latest order.

Keeping these points in mind. We tried to re-calculate the received quantity for that product every day of that month.

Assumptions:

- Everyday we are receiving items of that product
- Life expectancy of each item received is 3 days
- If the end Quantity is Zero we consider that, store has thrown all the expired items.
- Order is placed two days before (Eg:For Friday we order on Wednesday)

For suppose we are ordering on Wednesday for Friday. We need to know answers to two questions

1. How much do we have to order?
2. How many will be left by thursday?

For the first question, we are taking the mean of last four week sales of the same day (mean of sold quantity of last four Fridays). Expecting this week sales will also be the same and with that amount we are adding one threshold value (miscellaneous). That threshold value is calculated by means of the complete month received quantity.

Now the second question, as we are ordering for Friday, we have the end quantity on Thursday and we are adding the above generated value to the end quantity of the previous day(Thursday).

We have given this newly created value to a column and recalculated end quantity and performed LSTM on this new model.

However we can find a slight increase in the sales which implies that the initiative is applicable. However there might be some seasonality issues here as well.

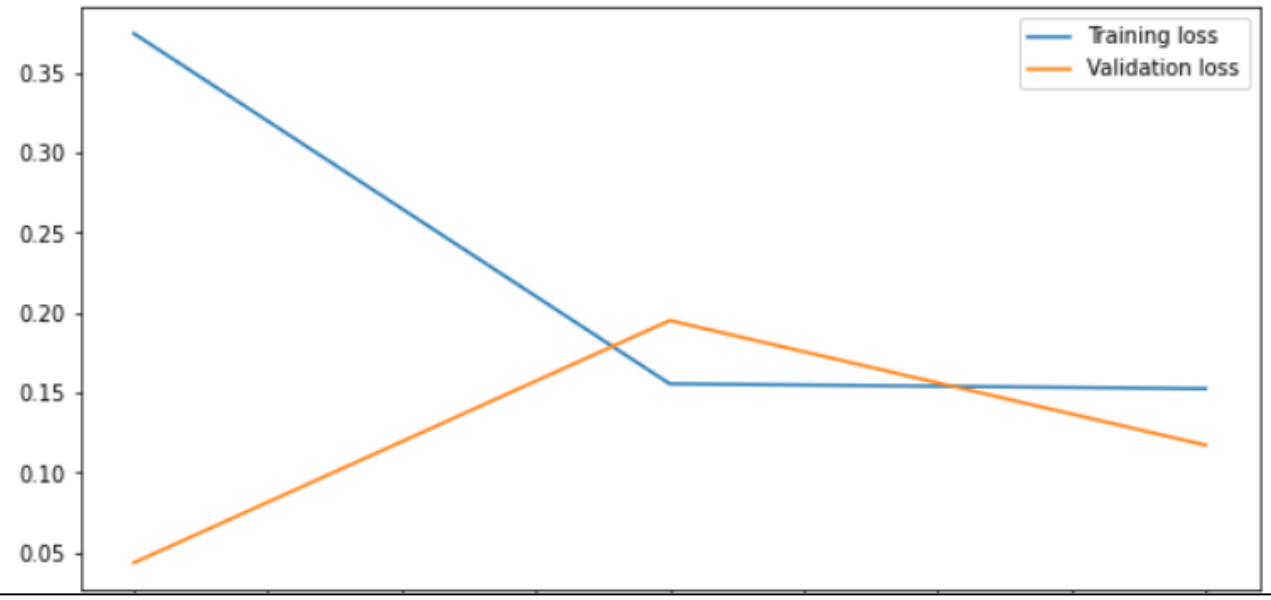
We have given below the prediction results with original value and modified value.

Original Received Quantity

	SoldQuantity	ReceivedQuantity	Sold_Pred
28	1.0	4.0	1.240881
29	0.0	4.0	1.431629

```
fig, ax = plt.subplots(figsize=(10,5))
plt.plot(history.history['loss'], label='Training loss')
plt.plot(history.history['val_loss'], label='Validation loss')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x1e298451df0>
```

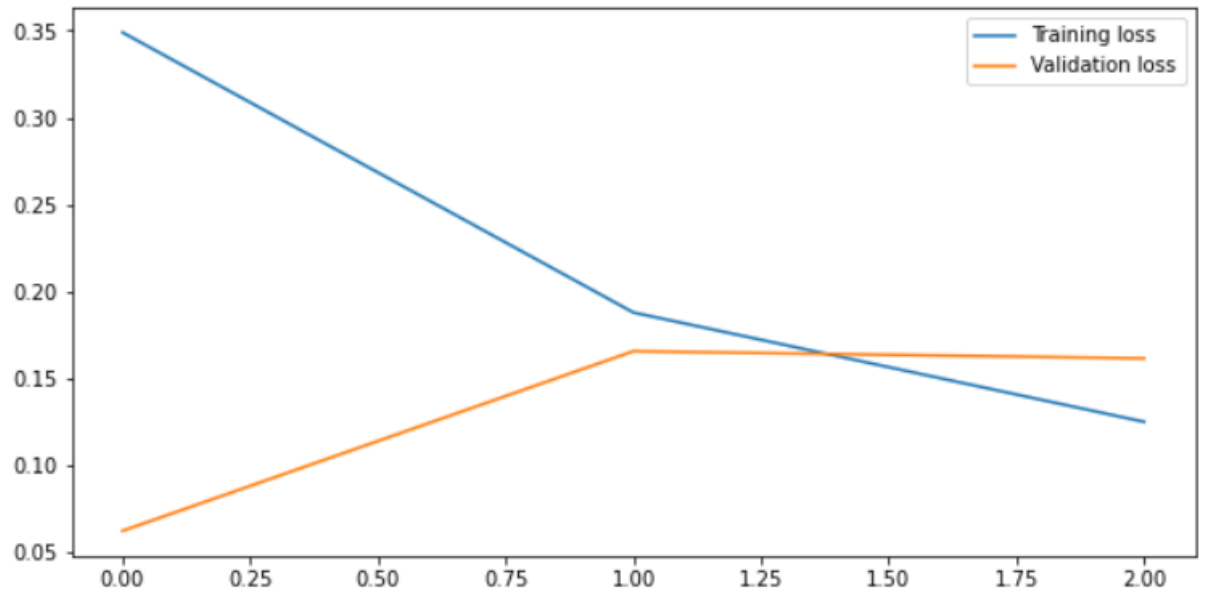


Prediction using Re calculated Received Quantity:

	SoldQuantity	New_RecievedQuantity	Sold_Pred
28	1.0	9.0	1.574611
29	0.0	8.0	1.604431

```
fig, ax = plt.subplots(figsize=(10,5))
plt.plot(history.history['loss'], label='Training loss')
plt.plot(history.history['val_loss'], label='Validation loss')
plt.legend()
```

```
<matplotlib.legend.Legend at 0x239ab94cb20>
```



Initiative 2:

Although we have not implemented this using code, we can follow the FIFO method of arranging stock and delivery of items will help in managing life expectancy of products in store or warehouse.

Initiative 3:

We can work on inventory for some special days like Christmas and New Year's Eve. We don't have much data for that but we believe there will be a good flow of customers on those special seasonalities and we can reduce the maximum of missed sales. However we can sometimes rely on temperature data for inventory management.

Explainer Dashboard

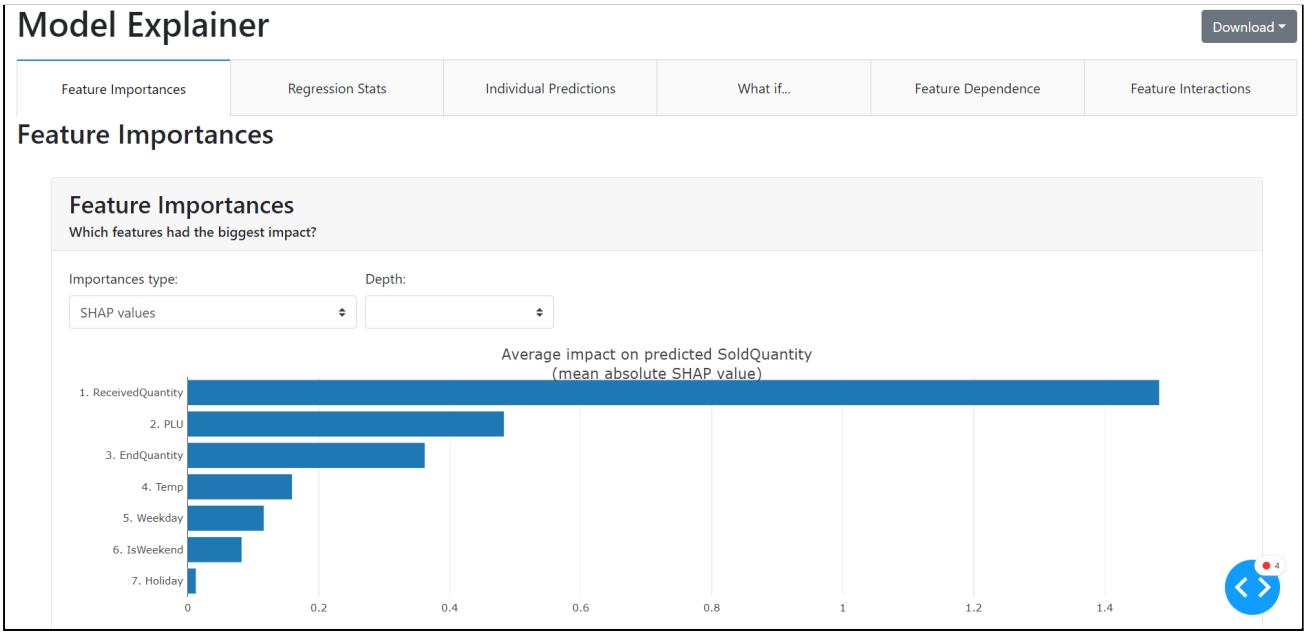
Explainer Dashboard is a library for quickly building interactive dashboards for analyzing and explaining the predictions and workings of (scikit-learn compatible) machine learning models, including xgboost, catboost and lightgbm. This makes your model transparent and explainable with just two lines of code.

Below are the few lines of code to run the explainer dashboard.

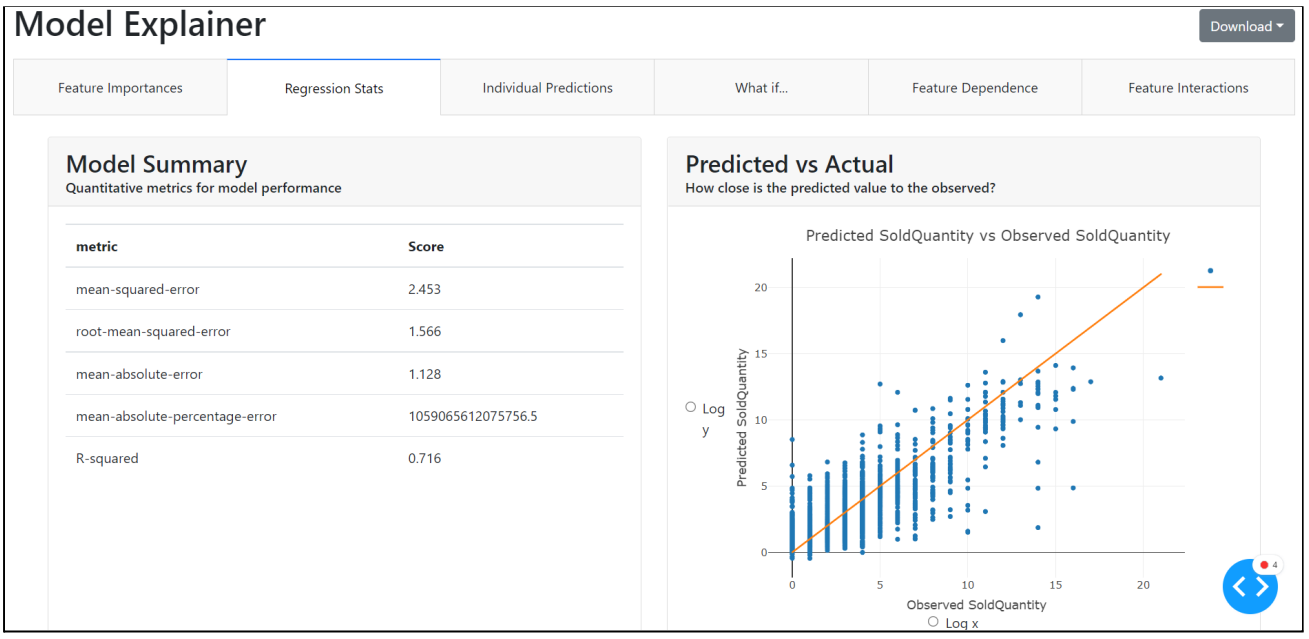
```
e1 = ExplainerDashboard(randomForest_explainer)
e2 = ExplainerDashboard(gradientBoosting_explainer)
e3 = ExplainerDashboard(lgbm_explainer)
e4 = ExplainerDashboard(xgbr_explainer)
e2.run()
```

Gradient boosting

The below diagram shows the model explainer for the gradient boosting dashboard.



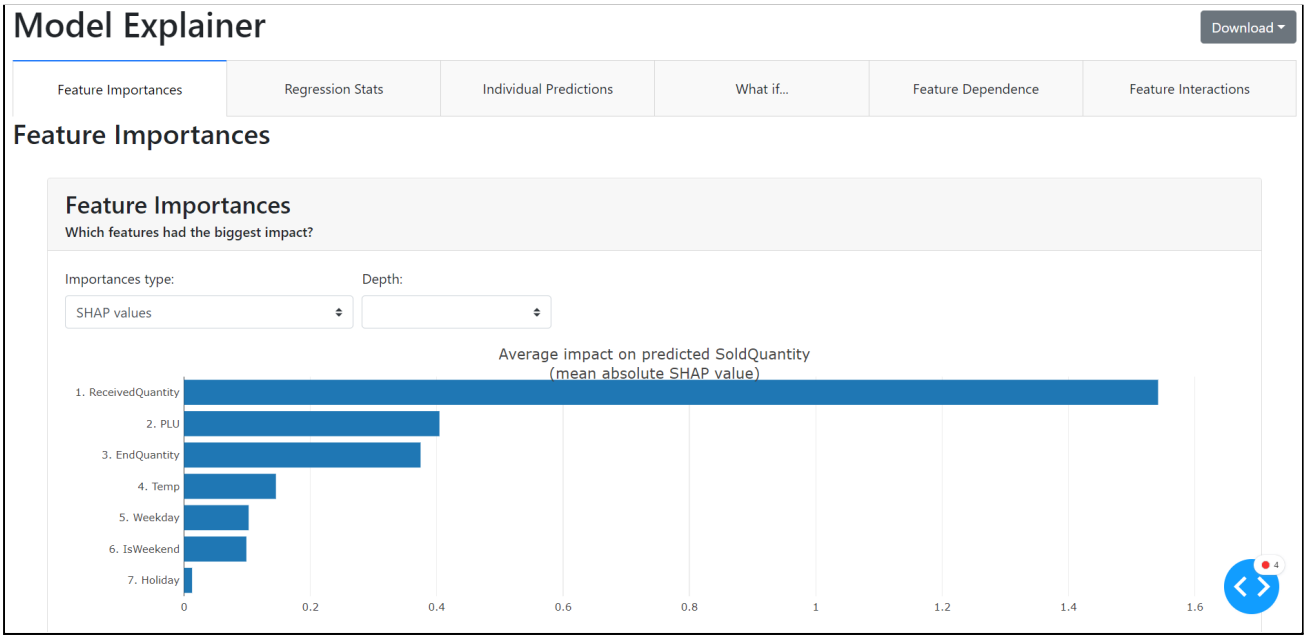
The below diagram illustrates regression stats of the gradient boosting algorithm



.

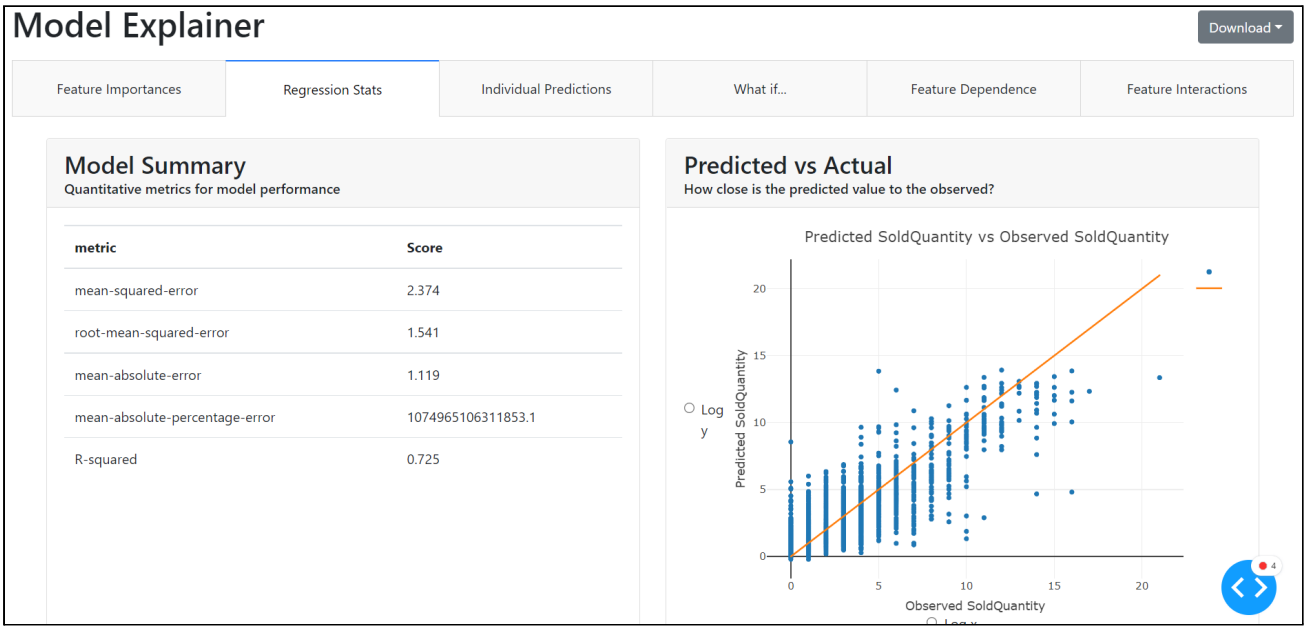
Light Gradient boosting machine

The below diagram shows the model explainer for the light gradient boosting dashboard



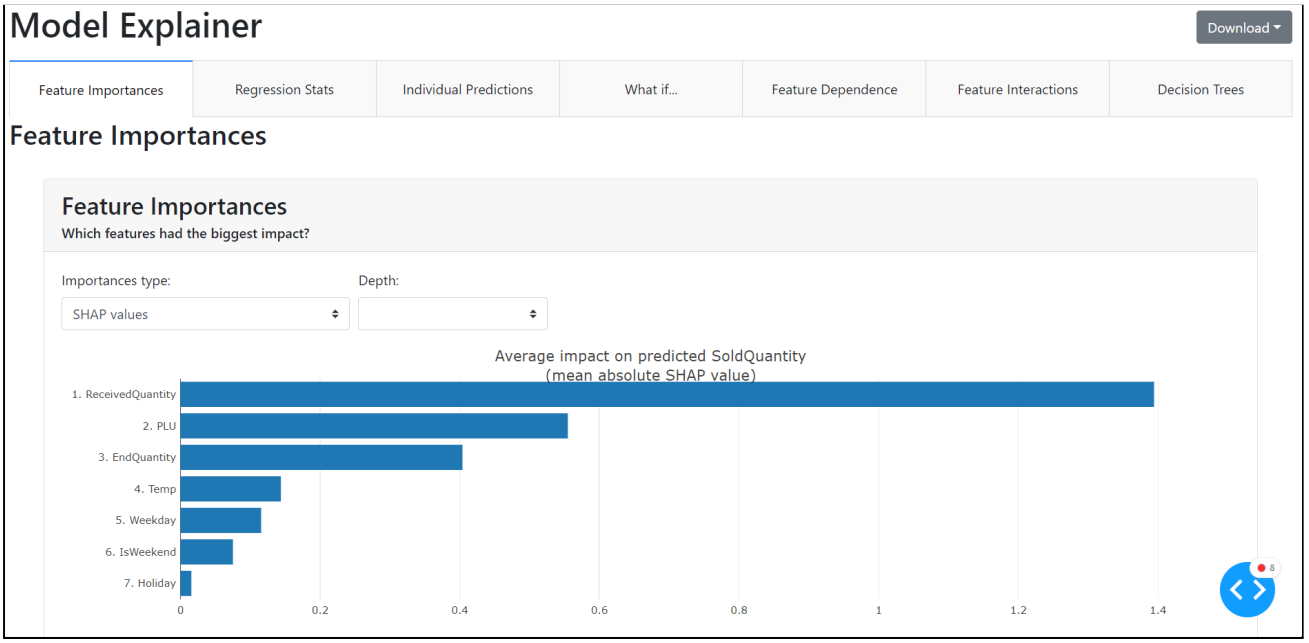
.

The below diagram illustrates regression stats of the lgbm.

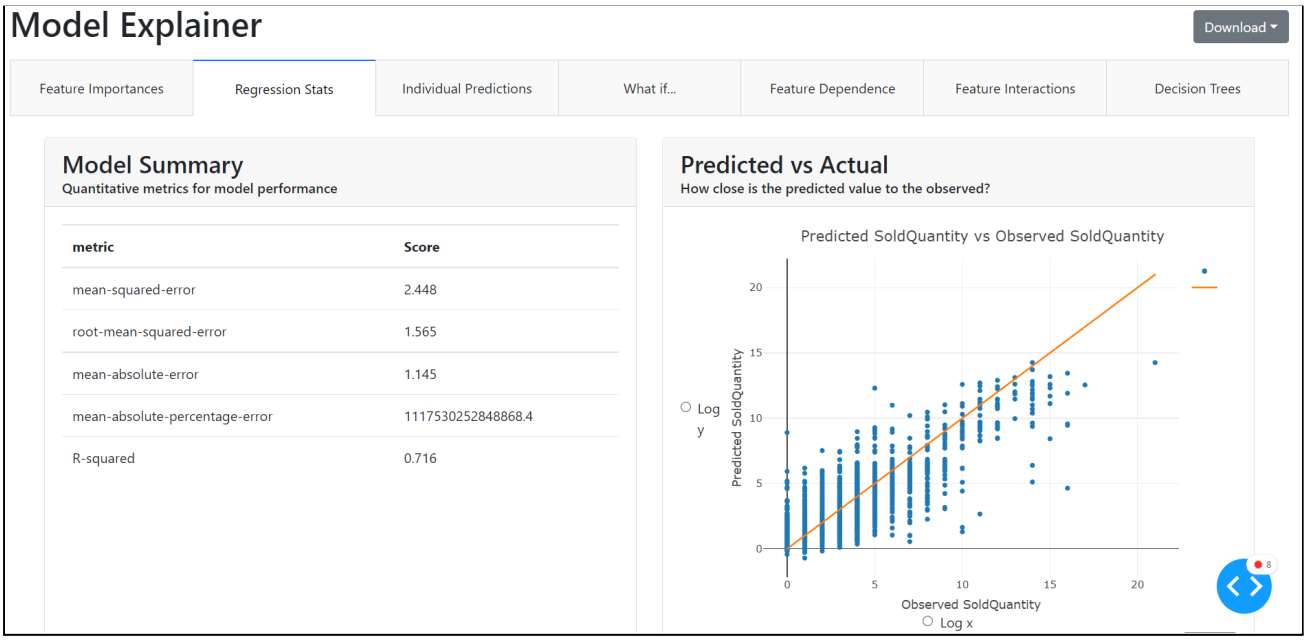


XGBR

The below diagram shows the model explainer for the XGBR dashboard.

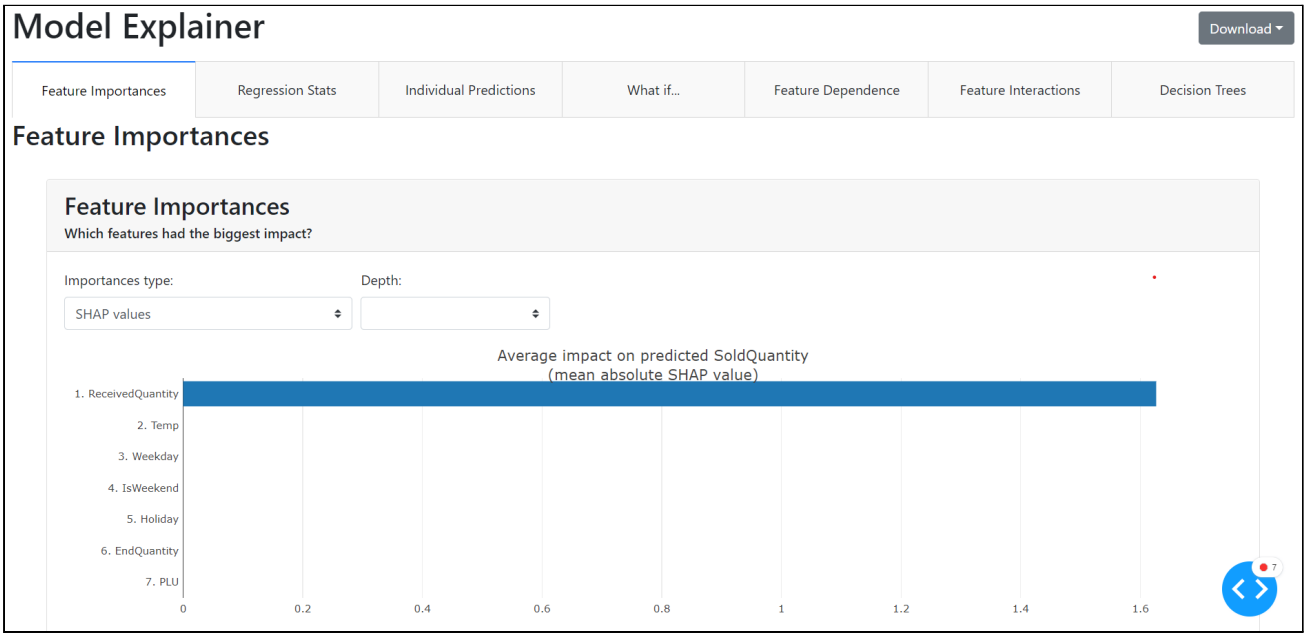


The below diagram illustrates regression stats of the lgbm.

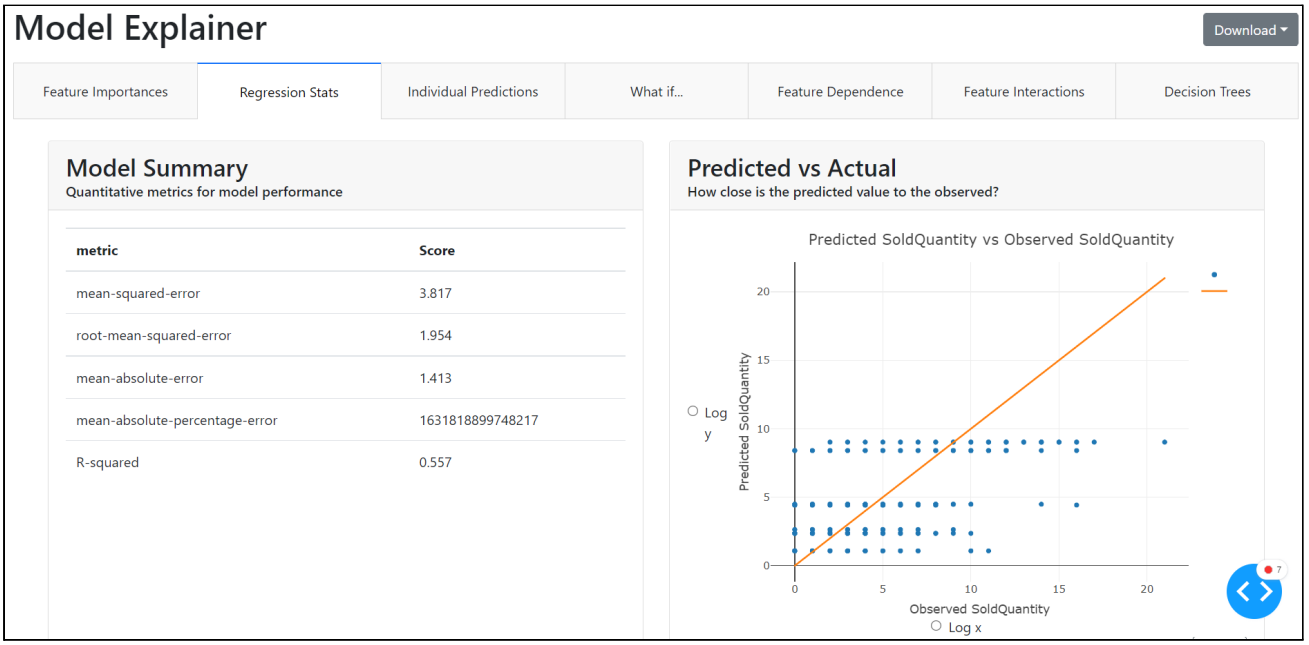


Random Forest

The diagram below describes the random forest dashboard with feature importance.



The below diagram illustrates the regression stats .



Conclusion:

- Overall, **initiative 1** will increase sales upto 10% and will be effective when we apply it to other products and all stores data.
- Explainer Dashboards have given elaborate explanations on the visulas on models which helped us in understanding more about the models we have built in previous sections.